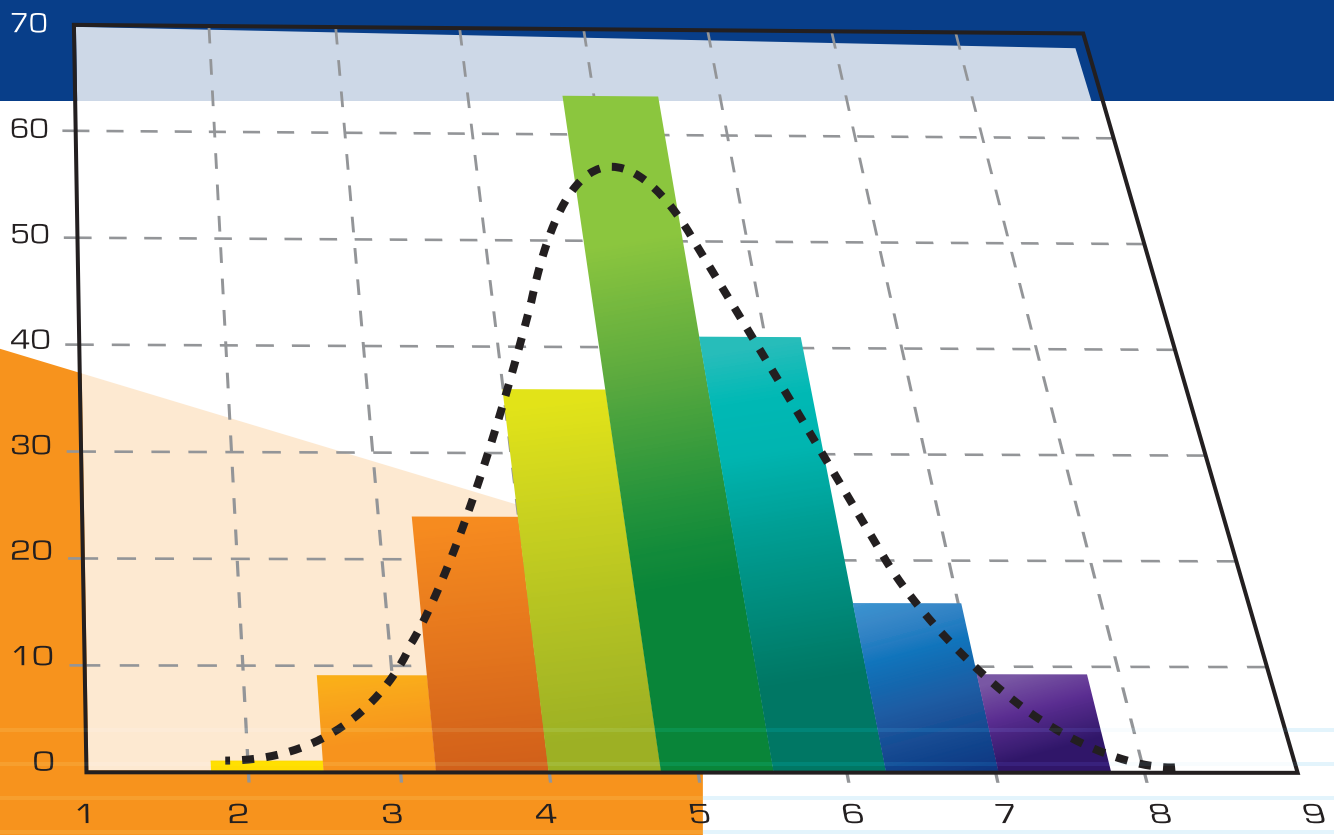
PROBABILIDAD Y ESTADÍSTICA
BÁSICA PARA INGENIEROS

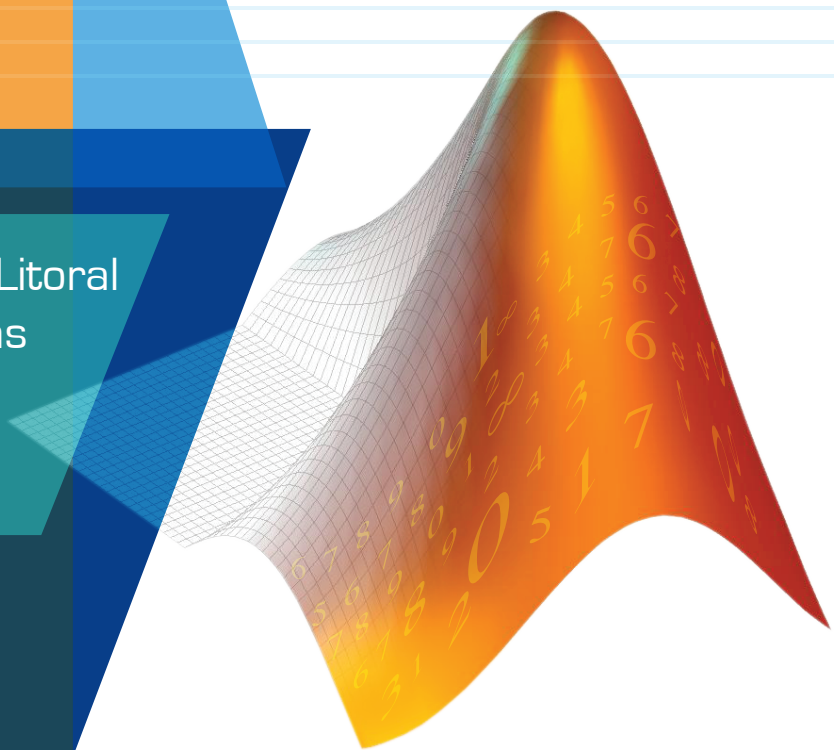
Con el soporte de MATLAB para cálculos y gráficos estadísticos

ISBN: 978-9942-922-02-1



Escuela Superior Politécnica del Litoral
Instituto de Ciencias Matemáticas
Guayaquil - Ecuador
2007

Luis Rodríguez Ojeda, MSc.
lrodrig@espol.edu.ec



CONTENIDO

PROBABILIDAD Y ESTADÍSTICA BÁSICA PARA INGENIEROS

Con el Soporte de MATLAB® para Cálculos y Gráficos Estadísticos

PREFACIO

Esta obra es una contribución bibliográfica para los estudiantes que toman un primer curso de Probabilidad y Estadística a nivel universitario en las carreras de ingeniería. El pre-requisito es el conocimiento del cálculo diferencial e integral y alguna experiencia previa con el programa MATLAB para aprovechar el poder de este instrumento computacional como soporte para los cálculos y gráficos estadísticos.

Este libro se originó en la experiencia desarrollada por el autor en varios años impartiendo el curso de Estadística en forma presencial y a distancia que ofrece el Instituto de Ciencias Matemáticas para estudiantes de ingeniería de la ESPOL y contiene el material del curso con algunos ejemplos basados en temas propuestos en exámenes receptados.

El enfoque de esta obra también tiene como objetivo que los estudiantes aprecien el uso de un instrumento computacional moderno y flexible que en forma integradora puede ser usado como soporte común para los diferentes cursos básicos de matemáticas, incluyendo Probabilidad y Estadística. Este soporte lo proporciona el programa MATLAB que dispone de un amplio repertorio de funciones especializadas para manejo estadístico y de muchas otras áreas de las ciencias y la ingeniería. Todos los cálculos en esta obra, incluyendo el manejo matemático simbólico y gráfico fueron realizados con estas funciones. Al final de este libro se incluye la descripción de dos instrumentos computacionales interactivos para experimentar con modelos de probabilidad y con la generación de muestras aleatorias.

Otro objetivo importante de esta obra se relaciona con el desarrollo de textos virtuales para ser usados interactivamente, reduciendo el consumo de papel y tinta, contribuyendo así con el cuidado del medio ambiente. Una ventaja adicional de los libros virtuales es la facilidad para su actualización y mejoramiento continuo del contenido.

El libro ha sido compilado en formato pdf. El tamaño del texto en pantalla es controlable, contiene un índice electrónico para facilitar la búsqueda de temas y dependiendo de la versión del programa de lectura de este formato, se pueden usar las facilidades disponibles para resaltar digitalmente texto, insertar comentarios, notas, enlaces, revisiones, búsqueda por contenido, lectura, etc.

 Su realización ha sido factible por el apoyo de la Institución a sus profesores en el desarrollo de sus actividades académicas

Luis Rodríguez Ojeda, M.Sc.
lrodrig@espol.edu.ec
Profesor titular
Instituto de Ciencias Matemáticas
Escuela Superior Politécnica del Litoral, ESPOL
Guayaquil, Ecuador
2007

1 FUNDAMENTOS DE LA ESTADÍSTICA

1.1 OBJETIVO

El objetivo fundamental de la Estadística es analizar datos y transformarlos en información útil para tomar decisiones.

El conocimiento de la Estadística se remonta a épocas en las que los gobernantes requerían técnicas para controlar a sus propiedades y a las personas.

Posteriormente, el desarrollo de los juegos de azar propició el estudio de métodos matemáticos para su análisis los cuales con el tiempo dieron origen a la Teoría de la Probabilidad que hoy es el sustento formal de la Estadística.

El advenimiento de la informática ha constituido el complemento adecuado para realizar estudios estadísticos mediante programas especializados que facilitan enormemente el tratamiento y transformación de los datos en información útil.

La Estadística ha alcanzado un nivel de desarrollo muy alto y constituye actualmente el soporte necesario para todas las ciencias y para la investigación científica, siendo el apoyo para tomar decisiones en un entorno de incertidumbre.

Es importante resaltar que las técnicas estadísticas deben usarse apropiadamente para que la información obtenida sea válida.

1.2 DEFINICIONES PRELIMINARES

ESTADÍSTICA
Ciencia inductiva que permite inferir características cualitativas y cuantitativas de un conjunto mediante los datos contenidos en un subconjunto del mismo.

POBLACIÓN OBJETIVO
Conjunto total de individuos u objetos con alguna característica que es de interés estudiar.

PARÁMETRO
Es alguna característica de la población en estudio y que es de interés conocer.

MUESTRA
Es un subconjunto de la población y contiene elementos en los cuales debe estudiarse la característica de interés para la población.

VARIABLE
Representación simbólica de alguna característica observable de los elementos de una población y que puede tomar diferentes valores.

OBSERVACIÓN o DATO
Cada uno de los valores obtenidos para los elementos incluidos en la muestra. Son el resultado de algún tipo de medición.

MODELO
Descripción simbólica o física de una situación o sistema que se desea estudiar

MODELO DETERMINÍSTICO
Representación exacta de un proceso. Permite obtener respuestas precisas si se conocen los valores de las variables incluidas en el modelo.

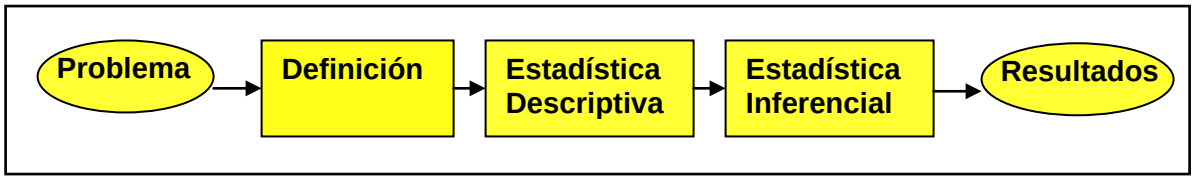
MODELO PROBABILISTA
Representación de un sistema que incluye componentes aleatorios. Las respuestas obtenidas se expresan en términos de probabilidad.

ESTADÍSTICA DESCRIPTIVA
Técnicas para recopilar, organizar, procesar y presentar datos obtenidos en muestras.

ESTADÍSTICA INFERENCIAL
Técnicas para obtención de resultados basados en la información contenida en muestras.

INFERENCIA ESTADÍSTICA
Es la extensión a la población de los resultados obtenidos en una muestra

1.3 DESARROLLO DE UN PROYECTO ESTADÍSTICO



En forma resumida, se describen los pasos para resolver un problema usando las técnicas estadísticas

PROBLEMA
Es una situación planteada para la cual se debe buscar una solución.

DEFINICIÓN
Para el problema propuesto deben establecerse los objetivos y el alcance del estudio a ser realizado considerando los recursos disponibles y definiendo actividades, metas y plazos. Se debe especificar la población a la cual está dirigido el estudio e identificar los parámetros de interés así como las variables que intervienen.

Se deben formular hipótesis y decidir el nivel de precisión que se pretende obtener en los resultados. Deben elegirse el tamaño de la muestra y las técnicas estadísticas y computacionales que serán utilizadas.

ESTADÍSTICA DESCRIPTIVA
Es el uso de las técnicas para obtener y analizar datos, incluyendo el diseño de cuestionarios en caso de ser necesarios. Se debe usar un plan para la obtención de los datos.

ESTADÍSTICA INFERENCIAL
Son las técnicas estadísticas utilizadas para realizar inferencias estadísticas que permiten validar las hipótesis propuestas.

RESULTADOS

Los resultados obtenidos deben usarse para producir información útil en la toma de decisiones.

La metodología de diseño en otros ámbitos de la ciencia e ingeniería usa la retroalimentación para corregir las especificaciones con las que se ejecutan las actividades, hasta que los resultados obtenidos concuerden con las especificaciones y requerimientos iniciales.

Sin embargo, el uso de retroalimentación en la resolución de un problema estadístico podría interpretarse como un artificio para modificar los datos o la aplicación de las técnicas estadísticas para que los resultados obtenidos concuerden con los requerimientos e hipótesis formuladas inicialmente. En este sentido, usar retroalimentación no sería un procedimiento aceptable.

1.4 PREGUNTAS

1) ¿Cual es la relación entre **dato, información y Estadística**?

2) ¿Cual es el aporte de la informática para el uso de las técnicas estadísticas?

3) ¿Por que hay que tener precaución en el uso de los resultados estadísticos?

4) ¿Cual es la diferencia entre **población y muestra**?

5) ¿Cual es la característica principal de un modelo probabilista?

6) ¿Cual es el objetivo de realizar una inferencia estadística?

2 ESTADÍSTICA DESCRIPTIVA

Es el estudio de las técnicas para recopilar, organizar y presentar datos obtenidos en un estudio estadístico para facilitar su análisis y aplicación.

2.1 RECOPILACIÓN DE DATOS

Fuentes de datos

1) Investigación en registros administrativos: INEC, Banco Central, Cámaras de la Producción, Universidades, etc. para obtener índices de empleo, índice de precios, datos de salud, datos de eficiencia, etc.
2) Obtención de datos mediante encuestas de investigación Ej. Estudios de mercado. Estudios de preferencia electoral, etc
3) Realización de experimentos estadísticos

Criterios para diseñar una encuesta de investigación

1) Definir el objetivo del estudio
2) Definir la población de interés
3) Determinar el tamaño de la muestra
4) Seleccionar el tipo de muestreo
5) Elegir temas generales
6) Elaborar el formulario para la encuesta: Preguntas cortas, claras y de opciones.
7) Realizar pruebas
8) Realizar la encuesta

Tipos de datos

Los resultados que se obtiene pueden ser

1) Datos cualitativos: corresponden a respuestas categóricas
   Ej. El estado civil de una persona
2) Datos cuantitativos: corresponden a respuestas numéricas
   Ej. La edad en años.

Los datos cuantitativos pueden ser

1) Discretos: Se obtienen mediante conteos
2) Continuos: Se obtienen mediante mediciones

2.2 SIMBOLOGÍA

Sea **N** el tamaño de la población objetivo y **n** el número de elementos que se incluyen en la muestra, entonces si **X** representa la característica que es de interés estudiar, la muestra es el conjunto de variables:

X: $\{X_1, X_2, ..., X_n\}$

En la notación vectorial, **X** es un vector de **n** variables:

$\mathbf{X^T} = (X_1, X_2, ..., X_n)$

Cada variable puede tomar un valor que se obtiene mediante una medición, y estos valores se los puede representar por

x: $\{x_1, x_2, ..., x_n\}$

Si se escribe $X_1 = x_1$ debe entenderse que al tomar la medición, para la variable X_1 se obtuvo el valor x_1. Entonces el vector de datos se puede escribir

$\mathbf{x^T} = (x_1, x_2, ..., x_n)$

Ejemplo. Una bodega contiene **N** = 50 artículos. Cada uno puede estar en tres estados: aceptable (**a**), regular (**r**), o defectuoso (**d**). Para una inspección se decide tomar una muestra **X** de **n** = 4 artículos elegidos al azar. Entonces,

X: $\{X_1, X_2, X_3, X_4\}$, representa cada muestra que se puede obtener

Supongamos que los valores obtenidos son respectivamente: d, a, a, r. Entonces

$X_1 = d$, $X_2 = a$, $X_3 = a$, $X_1 = r$

x: {d, a, a, r} son los datos que se obtuvieron en esta muestra

Es útil ordenar los datos de la muestra. Para representar una muestra de tal manera que los valores de las **n** variables estén en forma ordenada creciente se usa la siguiente notación:

$X_{(1)}, X_{(2)}, ..., X_{(n)}$

Esto implica que $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n-1)} \leq X_{(n)}$

Las variables $X_{(1)}, X_{(2)}, ..., X_{(n)}$ se denominan **estadísticos de orden** 1, 2, ..., n respectivamente.

Ejemplo. Una muestra de tamaño **n** = 4 contiene los valores

$X_1 = 7, X_2 = 8, X_3 = 5, X_4 = 2$

Entonces

$X_{(1)} = 2, X_{(2)} = 5, X_{(3)} = 7, X_{(4)} = 8$

2.2.1 PREGUNTAS

a) En las fuentes de recopilación de datos no se ha mencionado el uso de Internet. ¿Cuales son las ventajas y peligros de su uso?
b) Al diseñar el formulario de una encuesta de investigación. ¿Porqué se prefieren preguntas con opciones para elegir?
c) El número telefónico de una persona. ¿Es un dato cualitativo o cuantitativo?
d) El dinero es un dato cuantitativo, ¿Discreto o continuo?

2.3 DESCRIPCIÓN DE CONJUNTOS DE DATOS

Los datos obtenidos se los puede representar de diferentes formas:

1) Tabularmente
2) Gráficamente
3) Mediante números que caracterizan al grupo de datos

Si la muestra contiene pocos datos, estos se pueden representar directamente. Pero si el número de datos es grande conviene agruparlos para facilitar su análisis

2.4 TABLA DE FRECUENCIAS

Es un dispositivo para agrupación de datos y facilitar su interpretación.

Recomendaciones para construir la Tabla de Frecuencias

Sea **X** una muestra de tamaño **n**

1) Identificar la unidad de medida de los datos

2) Obtener el rango de los datos: distancia entre el mayor y el menor valor de los datos

R = $X_{(n)} - X_{(1)}$ (Rango de los datos)

3) Seleccionar el numero de clases (o intervalos) **k**, para agrupar los datos.

Sugerencia para elegir **k**

Sean **n:** número de datos

k: Número de clases

n	k
Menos de 50	**5** a **7**
Entre 50 y 100	**6** a **10**
Entre 100 y 250	**7** a **12**
Mas de 250	**10** a **20**

4) Obtener la longitud de las clases,

L = R/k (Longitud)

Se puede redefinir la longitud, el número de clases y los extremos de cada clase de tal manera que las clases tengan la misma longitud y los intervalos de cada clase incluyan a todos los datos, sean excluyentes y los valores en los extremos de cada clase sean simples.

Si $\mathbf{a_i}$, $\mathbf{b_i}$ son los extremos de la clase **i**, entonces el intervalo de la clase **i** es $\mathbf{[a_i, b_i)}$

5) Realizar el conteo de datos para obtener la frecuencia en cada clase

Notación

- **n:** número de datos
- **k:** número de clases
- $\mathbf{f_i}$**:** frecuencia de la clase **i, i=1, 2, 3, …, k**
- $\mathbf{f_i/n}$**:** frecuencia relativa de la clase **i**
- $\mathbf{F_i}$**:** frecuencia acumulada de la clase **i:** $\mathbf{F_i = f_1+f_2+f_3+\ldots+f_i}$
- $\mathbf{F_i/n}$**:** frecuencia acumulada relativa de la clase **i**
- $\mathbf{m_i}$ **:** marca de la clase **i** (es el valor central del intervalo de la clase **i**)

Los resultados se los organiza en un cuadro denominado Tabla de Frecuencia

Ejemplo.- Obtenga la Tabla de Frecuencias para los siguientes 40 datos de una muestra, correspondientes al tiempo que se utilizó para atender a las personas en una estación de servicio:

3.1	4.9	2.8	3.6
4.5	3.5	2.8	4.1
2.9	2.1	3.7	4.1
2.7	4.2	3.5	3.7
3.8	2.2	4.4	2.9
5.1	1.8	2.5	6.2
2.5	3.6	5.6	4.8
3.6	6.1	5.1	3.9
4.3	5.7	4.7	4.6
5.1	4.9	4.2	3.1

Solución

1) Precisión: un decimal
2) Rango: R = 6.2 – 1.8 = 4.4
3) Número de clases: k=6
4) Longitud: R/k = 0.7333...
   Por simplicidad se redefine la longitud como 1 y se usan números enteros para los extremos de las clases.
5) Conteo de los datos (puede hacerse en un solo recorrido), **n=40**

Número	**Clase (Intervalo)**	**Frecuencia absoluta**
1	[1, 2)	1
2	[2, 3)	9
3	[3, 4)	11
4	[4, 5)	12
5	[5, 6)	5
6	[6, 7)	2

Tabla de Frecuencias

Número **i**	**Clase (Intervalo)** **[a, b)**	**Marca de clase** **m**	**Frecuencia absoluta** **f**	**Frecuencia relativa** **f/n**	**Frecuencia absoluta acumulada** **F**	**Frecuencia relativa acumulada** **F/n**
1	[1, 2)	1.5	1	0.025	1	0.025
2	[2, 3)	2.5	9	0.225	10	0.250
3	[3, 4)	3.5	11	0.275	21	0.525
4	[4, 5)	4.5	12	0.300	33	0.825
5	[5, 6)	5.5	5	0.125	38	0.950
6	[6, 7)	6.5	2	0.050	40	1.000

2.4.1 EJERCICIOS

1) Suponga que una población objetivo consta de 5 personas y que es de interés para un estudio la edad en años. Los valores incluidos en esta población son: 25, 30, 40, 25, 20. De esta población se toma una muestra de tamaño 3

Si representamos la muestra con **X**: {X_1, X_2, X_3}

a) ¿Cuantas muestras diferentes pueden obtenerse? (Las muestras son combinaciones)
b) Liste todas las muestras diferentes que se pueden tomar de esta población.
(Debe considerar todos los valores que pueden tomar las variables X_1, X_2, X_3)

Sugerencia: Revise la fórmula de combinaciones de las Técnicas de Conteo

2) Con los resultados obtenidos y descritos en la Tabla de Frecuencias del ejemplo desarrollado en la Sección 1.4.4 conteste las siguientes preguntas

a) ¿Cuántas personas requirieron no más de 4 minutos para ser atendidas?
b) ¿Cuántas personas requirieron entre 2 y 5 minutos?
c) ¿Cuántas personas requirieron al menos 4 minutos?
d) ¿Cuál es la duración que ocurre con mayor frecuencia?

3) Suponga que se desean analizar los siguientes datos correspondientes al costo de electricidad durante un mes y que se obtuvieron en una muestra de 50 casas en una zona residencial de Guayaquil:

96	171	202	178	147	102	153	129	127	82
157	185	90	116	172	111	148	213	130	165
141	149	206	175	123	128	144	168	109	167
95	163	150	154	130	143	187	166	139	149
108	119	183	151	114	135	191	137	129	158

Procedimiento para decidir el número de clases para la Tabla de Frecuencias

Rango:
Número de clases:
Longitud:

Conteo de Frecuencias

Número	Clase (intervalo)	Conteo	Frecuencia
1			
2			
3			
4			
5			
6			
7			
8			

Tabla de Frecuencias

Número	Clase (Intervalo)	Marca de clase	Frecuencia absoluta	Frecuencia relativa	Frec. abs. acum.	Frec. rel. acum.
1						
2						
3						
4						
5						
6						
7						
8						

MATLAB

Construcción de la Tabla de Frecuencias

Vector con los datos

```
>> x=[3.1 4.9 2.8 3.6 4.5 3.5 2.8 4.1 2.9 2.1 3.7 4.1 2.7 4.2 3.5 3.7 3.8 2.2 4.4 2.9...
      5.1 1.8 2.5 6.2 2.5 3.6 5.6 4.8 3.6 6.1 5.1 3.9 4.3 5.7 4.7 4.6 5.1 4.9 4.2 3.1];
```

```
>> a = min(x)
   a =
       1.8000
```
El menor valor

```
>> b = max(x)
   b =
       6.2000
```
El mayor valor

```
>> m=[1.5 2.5 3.5 4.5 5.5 6.5];
```
Para definir 6 clases 1-2, 2-3, . . ., 6-7 se crea un vector con las marcas de clase

```
>> f=hist(x,m)
   f =
     1    9    11    12    5    2
```
Obtención de las frecuencias en las marcas de clase

```
>> fr=f/40
   fr =
       0.0250    0.2250    0.2750    0.3000    0.1250    0.0500
```
Frecuencias relativas

```
>> F=cumsum(f)
   F =
     1    10    21    33    38    40
```
Frecuencias acumuladas

```
>> Fr=F/40
   Fr =
       0.0250    0.2500    0.5250    0.8250    0.9500    1.0000
```
Frecuencias acumuladas relativas

2.5 REPRESENTACIÓN GRÁFICA DE CONJUNTOS DE DATOS

En esta sección revisamos algunos dispositivos frecuentemente usados para resaltar visualmente las características de grupos de datos.

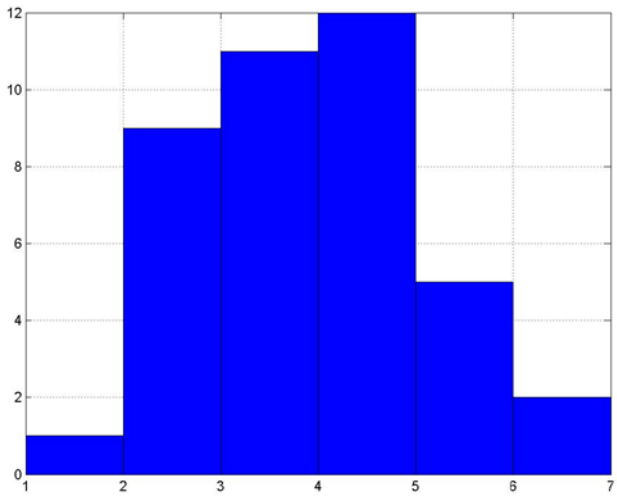
2.5.1 HISTOGRAMA DE FRECUENCIAS

Es la manera más común de representar gráficamente la distribución de frecuencia de los datos. Se lo construye dibujando rectángulos cuya base corresponde a cada intervalo de clase, y su altura según el valor de la frecuencia. Puede ser la frecuencia absoluta o la frecuencia relativa.

Ejemplo. Construya el histograma para el ejemplo de la unidad anterior. Use los valores de la frecuencia absoluta
:

Tabla de Frecuencia

Número	Clase (Intervalo)	Marca de clase	Frecuencia absoluta	Frecuencia relativa	Frecuencia absoluta acumulada	Frecuencia relativa acumulada
1	[1, 2)	1.5	1	0.025	1	0.025
2	[2, 3)	2.5	9	0.225	10	0.250
3	[3, 4)	3.5	11	0.275	21	0.525
4	[4, 5)	4.5	12	0.300	33	0.825
5	[5, 6)	5.5	5	0.125	38	0.950
6	[6, 7)	6.5	2	0.050	40	1.000



Histograma

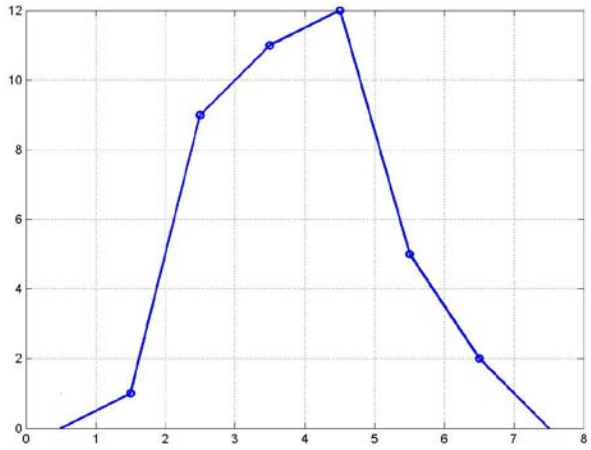
El histograma permite dar una primera mirada al tipo de distribución de los datos:

1) Si las alturas de las barras son similares se dice que tiene distribución tipo “uniforme”
2) Si las alturas son mayores en la zona central se dice que tiene forma tipo “campana” y puede ser simétrica o asimétrica, con sesgo hacia el lado positivo o al lado negativo
3) Si hay barras muy alejadas del grupo, se dice que son **datos atípicos.** Probablemente estos datos se pueden atribuir a errores de medición y se los puede descartar pues no pertenecen al grupo que se desea caracterizar.

2.5.2 POLÍGONO DE FRECUENCIAS

Es una manera de representar el perfil de la distribución de los datos. Se obtiene uniendo mediante segmentos de recta los puntos **(marca de clase, frecuencia)**
Para cerrar el polígono se puede agregar un punto a cada lado con frecuencia 0.

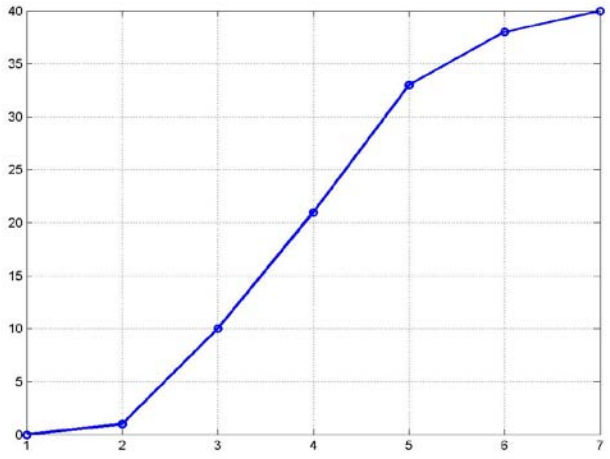
Polígono de frecuencia para el ejemplo dado:



2.5.3 OJIVA

Este gráfico se usa para representar la frecuencia acumulada, absoluta o relativa. Se lo obtiene uniendo segmentos de recta que se extienden entre los extremos de las clases y usando los valores de la frecuencia acumulada.

Ojiva para el ejemplo dado:



La ojiva permite responder preguntas tipo “cuantos datos son menores que”

Ejemplo. ¿Cuantos datos tienen un valor menor a 4.5?

Respuesta: aproximadamente 27 datos

2.5.4 GRÁFICOS DE FRECUENCIAS CON FORMAS ESPECIALES

Los gráficos pueden tomar otros aspectos usando barras, colores, efectos tridimensionales, sombreado, etc. o usando una representación tipo pastel. Como ilustración se muestran algunos:

Diagrama de barras

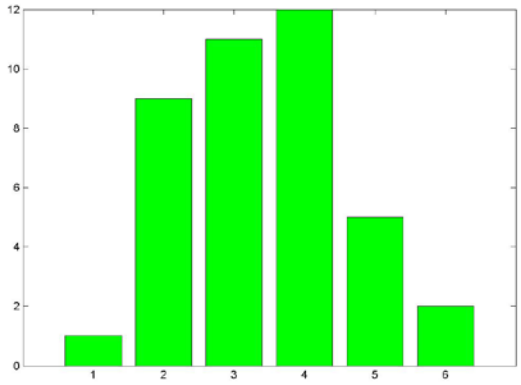


Diagrama de barras con efecto tridimensional

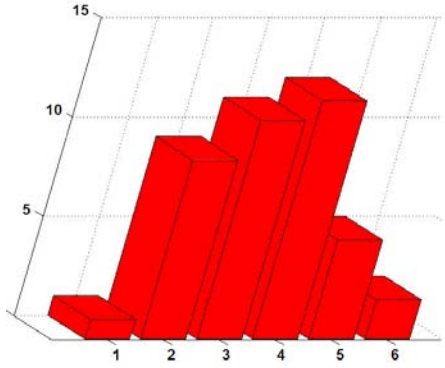
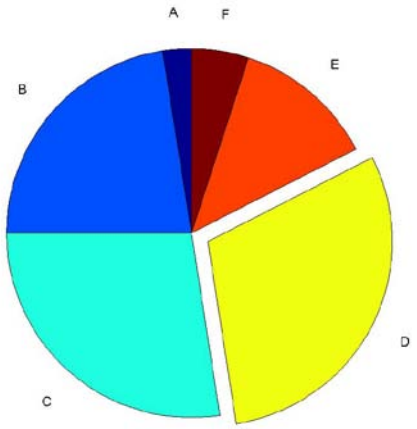


Diagrama tipo pastel

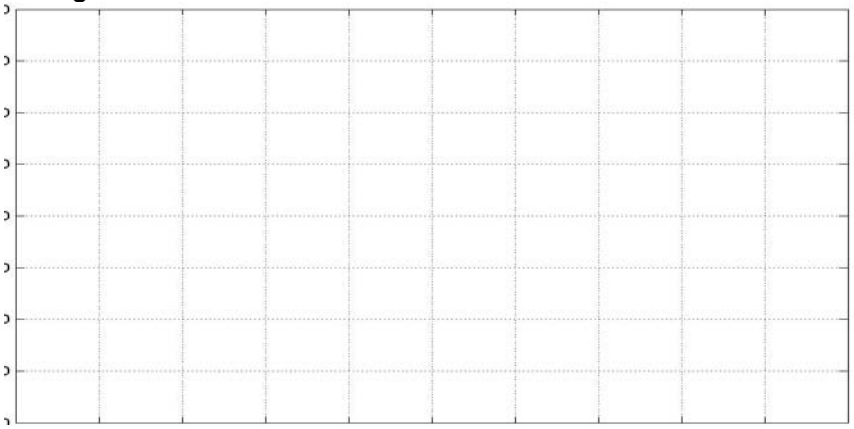


El ángulo de cada sector circular es proporcional al valor de la frecuencia respectiva.
Se puede resaltar algún valor particular separándolo del dibujo.

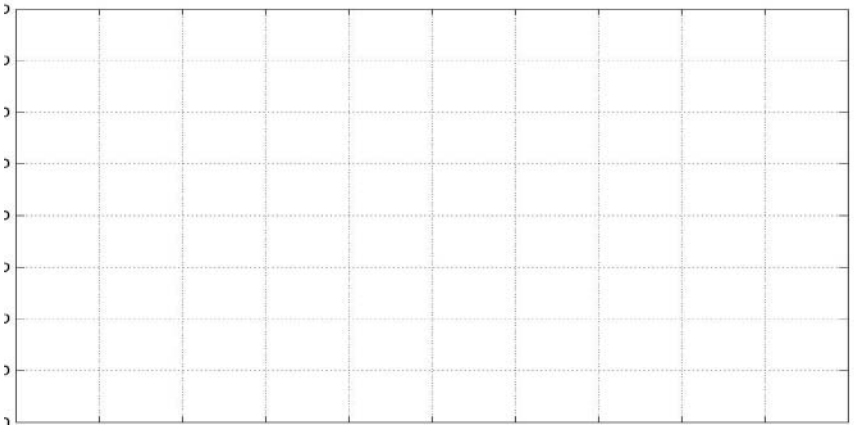
2.5.5 EJERCICIOS

Dibuje los siguientes gráficos con los resultados del ejercicio 3 de la Sección 1.4.5

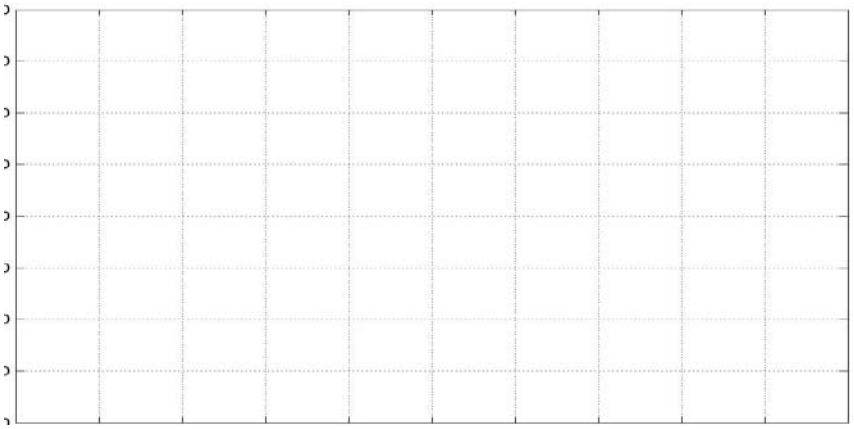
Histograma



Polígono de Frecuencia



Ojiva



MATLAB

Obtención de gráficos. Los dibujos obtenidos se muestran en las páginas anteriores

Vector con los datos

```
>> x = [3.1 4.9 2.8 3.6 4.5 3.5 2.8 4.1 2.9 2.1 3.7 4.1 2.7 4.2 3.5 3.7 3.8 2.2 4.4 2.9...
        5.1 1.8 2.5 6.2 2.5 3.6 5.6 4.8 3.6 6.1 5.1 3.9 4.3 5.7 4.7 4.6 5.1 4.9 4.2 3.1];
```

Vector con las marcas de clase

```
>> m=[1.5 2.5 3.5 4.5 5.5 6.5];
```

Graficación del histograma

```
>> hist(x, m);          Dibujar el histograma con barras sobre las marcas
>> grid on              Dibujar cuadrículas
```

Graficación del polígono de frecuencias

```
>> mp=[0.5 m 7.5];      Se agrega un punto con frecuencia cero a los lados
>> f = hist(x, m);      Obtención de las frecuencias en las marcas de clase
>> fp=[0 f 0];

>> clf                  Borrar el gráfico anterior
>> plot(mp,fp,'o')      Dibujar los puntos del polígono

>> hold on              Mantener el gráfico anterior para superponer otro
>> plot(mp,fp)          Trazado de las líneas del polígono
>> grid on              Cuadrículas
```

Graficación de la ojiva

```
>> c=[1 2 3 4 5 6 7];   Vector con los extremos de las seis clases
>> F=cumsum(f);         Vector con las frecuencias acumuladas
>> Fo=[0 F];            Se agrega un punto a la izquierda con frecuencia cero

>> clf
>> plot(c,Fo,'o')       Dibujo de los puntos en un nuevo gráfico
>> hold on              Para superponer el siguiente gráfico
>> plot(c, Fo)          Trazado de las líneas de la ojiva
>> grid on
```

Gráfico de diagrama de barras, color verde

```
>> clf
>> bar(f,'g')
```

Gráfico de diagrama de barras, horizontal con efecto tridimensional, color rojo

```
>> clf
>> bar3h(f,'r')
```

Gráfico tipo pastel, con rótulos y extracción de porciones

```
>> sacar = [0 0 0 1 0 0];                  Sacar la cuarta porción
>> nombres = {'A','B','C','D','E','F'};    Rótulos para las porciones
>> pie(f, sacar, nombres)                  Dibujar el pastel con rótulos
```

2.6 MEDIDAS DE TENDENCIA CENTRAL

Son números que definen cual es el valor alrededor del que se concentran los datos. Se indican a continuación los más utilizados.

2.6.1 MEDIA MUESTRAL

Si **X:** $X_1, X_2, \ldots, X_n$ es una muestra de **n** datos, entonces la media muestral es el promedio aritmético simple de los datos:

Definición: Media Muestral

$$\overline{X} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Ejemplo. Si los datos son **2, 6, 11, 8, 11, 4, 7, 5**
Entonces $\overline{X}$ = **(2+6+11+8+11+4+7+5)/8 = 6.75**

La media muestral es una medida de uso común. En el cálculo intervienen todos los datos, sin embargo, algunos datos pueden hacer cambiar significativamente el valor de la media muestral.

Ejemplo. Si los datos son **2, 6, 11, 8, 11, 4, 7, 5, 90**
Entonces $\overline{X}$ = **(2+6+11+8+11+4+7+5 + 90)/9 = 16**

Un sólo dato cambió significativamente el valor de la media con respecto al ejemplo anterior

Para evitar esta distorsión, una estrategia consiste en descartar algún porcentaje de los datos más grandes y más pequeños antes de calcular la media muestral. Este porcentaje puede ser por ejemplo **5%** o **10%**. Cuando se usa este criterio la media se denomina **media cortada**.

2.6.2 MODA MUESTRAL

Es el dato que ocurre con mayor frecuencia en una muestra. Puede ser que no exista la moda y también es posible que exista más de una moda.

Definición: Moda Muestral

Moda muestral: **Mo** es el valor que más veces se repite

Ejemplo. Si los datos son **2, 6, 11, 8, 11, 4, 7, 5**
Entonces **Mo = 11**

2.6.3 MEDIANA MUESTRAL

Es el valor ubicado en el centro de los datos ordenados
Sean **X:** $X_1, X_2, \ldots, X_n$ una muestra de tamaño **n**
$X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ los elementos de la muestra ordenados en forma creciente

Definición: Mediana Muestral

$$\tilde{X} = \begin{cases} X_{(\frac{n+1}{2})}, & \text{si } n \text{ es impar} \\ \frac{1}{2}(X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}), & \text{si } n \text{ es par} \end{cases}$$

Ejemplo: Si los datos son **2, 6, 11, 8, 11, 4, 7, 5**

Los datos ordenados: **2, 4, 5, 6, 7, 8, 11, 11,** entonces $\tilde{X} = \frac{1}{2}(6+7) = 6.5$

Las medidas de tendencia central no son suficientes para describir de manera completa el comportamiento de los datos de una muestra. Se necesitan otras medidas.

2.7 MEDIDAS DE DISPERSIÓN

Son números que proveen información adicional acerca del comportamiento de los datos, describiendo numéricamente su dispersión.

2.7.1 RANGO

Es la diferencia entre el mayor valor y el menor valor de los datos de la muestra.

Definición: Rango

$$R = X_{(n)} - X_{(1)}$$

Ejemplo. Si los datos son **2, 6, 11, 8, 11, 4, 7, 5**

Entonces el rango es: **R = 11 – 2 = 9**

2.7.2 VARIANZA MUESTRAL

Esta medida cuantifica las distancias de los datos con respecto al valor de la media muestral

Definición: Varianza Muestral

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$ Fórmula para calcular la varianza

$$S^2 = \frac{n\sum_{i=1}^{n}X_i^2 - (\sum_{i=1}^{n}X_i)^2}{n(n-1)}$$ Fórmula alternativa para calcular la varianza

El motivo que en el denominador se escriba **n – 1** en lugar de **n** (que parece natural), se justificará formalmente en el estudio de la Estadística Inferencial.

Ambas fórmulas son equivalentes. Se puede demostrar mediante desarrollo de las sumatorias

Ejemplo. Si los datos son **2, 6, 11, 8, 11, 4, 7, 5** y se ha calculado que $\bar{X} = 6.75$

Entonces la varianza es

$$S^2 = \frac{(2-6.75)^2 + (6-6.75)^2 + \ldots + (5-6.75)^2}{7} = 10.2143$$

2.7.3 DESVIACIÓN ESTÁNDAR MUESTRAL

Es la raíz cuadrada positiva de la variancia. La desviación estándar muestral o desviación típica está expresada en las mismas unidades de medida que los datos de la muestra

Definición: Desviación Estándar Muestral

$$S = +\sqrt{S^2}$$

Ejemplo. Calcule la desviación estándar para el ejemplo anterior.

Si la varianza es $S^2 = \mathbf{10.2143}$, entonces, la desviación estándar es

$$S = \sqrt{S^2} = \sqrt{10.2143} = 3.196$$

2.8 MEDIDAS DE POSICIÓN

Son números que distribuyen los datos ordenados de la muestra en grupos de aproximadamente tamaño con el propósito de resaltar su ubicación relativa. Estos números se denominan **cuantiles** en forma genérica.

2.8.1 CUARTILES

Son números que dividen a los datos de la muestra en grupos de tamaño aproximado de 25%.

Primer Cuartil (Q_1)
A la izquierda de Q_1 están incluidos 25% de los datos (aproximadamente)
A la derecha de Q_1 están el 75% de los datos (aproximadamente)

Segundo Cuartil (Q_2)
Igual que la mediana divide al grupo de datos en dos partes, cada una con el 50% de los datos (aproximadamente)

Tercer Cuartil (Q_3)
A la izquierda de Q_3 están incluidos 75% de los datos (aproximadamente)
A la derecha de Q_3 están el 25% de los datos (aproximadamente)

Ejemplo. Suponer que una muestra contiene 40 datos ordenados:
$X_{(1)}, X_{(2)}, \ldots, X_{(40)}$. Calcular $\mathbf{Q_1, Q_2, Q_3}$

$\mathbf{Q_1}$: 25% de 40 = 10
Por lo tanto: $\mathbf{Q_1} = (X_{(10)} + X_{(11)})/2$

$\mathbf{Q_2}$: 50% de 40 = 20 es igual a la mediana
$\mathbf{Q_2} = (X_{(20)} + X_{(21)})/2$

$\mathbf{Q_3}$: 75% de 40 = 30
$\mathbf{Q_3} = (X_{(30)} + X_{(31)})/2$

2.8.2 DECILES

Son números que dividen a los datos de la muestra en grupos de tamaño aproximado de 10%.

Primer Decil (D_1)
A la izquierda de D_1 están incluidos 10% de los datos (aproximadamente)
A la derecha de D_1 están el 90% de los datos (aproximadamente)

Segundo Decil (D_2)
A la izquierda de D_2 están incluidos 20% de los datos (aproximadamente)
A la derecha de D_2 están el 80% de los datos (aproximadamente)
Etc.

Ejemplo. Suponer que una muestra contiene 40 datos ordenados:
$X_{(1)}, X_{(2)}, \ldots, X_{(40)}$. Calcular $\mathbf{D_1}$

$\mathbf{D_1}$: 10% de 40 = 4
Por lo tanto: $\mathbf{D_1} = (X_{(4)} + X_{(5)})/2$

2.8.3 PERCENTILES (O PORCENTILES)

Son números que dividen a los datos de la muestra en grupos de tamaño aproximado de 1%.

Primer Percentil (P_1)
A la izquierda de P_1 están incluidos 1% de los datos (aproximadamente)
A la derecha de P_1 están el 99% de los datos (aproximadamente)

Segundo Percentil (P_2)
A la izquierda de P_2 están incluidos 2% de los datos (aproximadamente)
A la derecha de P_2 están el 98% de los datos (aproximadamente)
Etc.

Ejemplo. Suponer que una muestra contiene 400 datos ordenados:
$X_{(1)}, X_{(2)}, \ldots, X_{(400)}$. Calcular P_1, P_{82}

$\mathbf{P_1}$: 1% de 400 = 4
Por lo tanto: $\mathbf{P_1} = (X_{(4)} + X_{(5)})/2$ (Percentil 1)

$\mathbf{P_{82}}$: 82% de 400 = 328 (Percentil 82)
$\mathbf{P_{82}} = (X_{(328)} + X_{(329)})/2$

2.9 COEFICIENTE DE VARIACIÓN

Es un número que se usa para comparar la variabilidad de los datos de diferentes grupos. Es una medida adimensional.

Definición: Coeficiente de Variación

$$\mathbf{V = \frac{S}{\overline{X}}}$$

Ejemplo: Para un grupo de datos $\overline{X}$ **= 20, S = 4**, entonces **v = 4/20 = 0.2 = 20%**
Para un segundo grupo $\overline{X}$ **= 48, S = 6**, entonces **v = 6/48 = 0.125 = 12.5%**
Se concluye que el primer grupo tiene mayor variabilidad relativa con respecto a su media.

2.9.1 EJERCICIOS

1) Demuestre mediante propiedades de las sumatoria que $\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}x_i^2-\frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}$

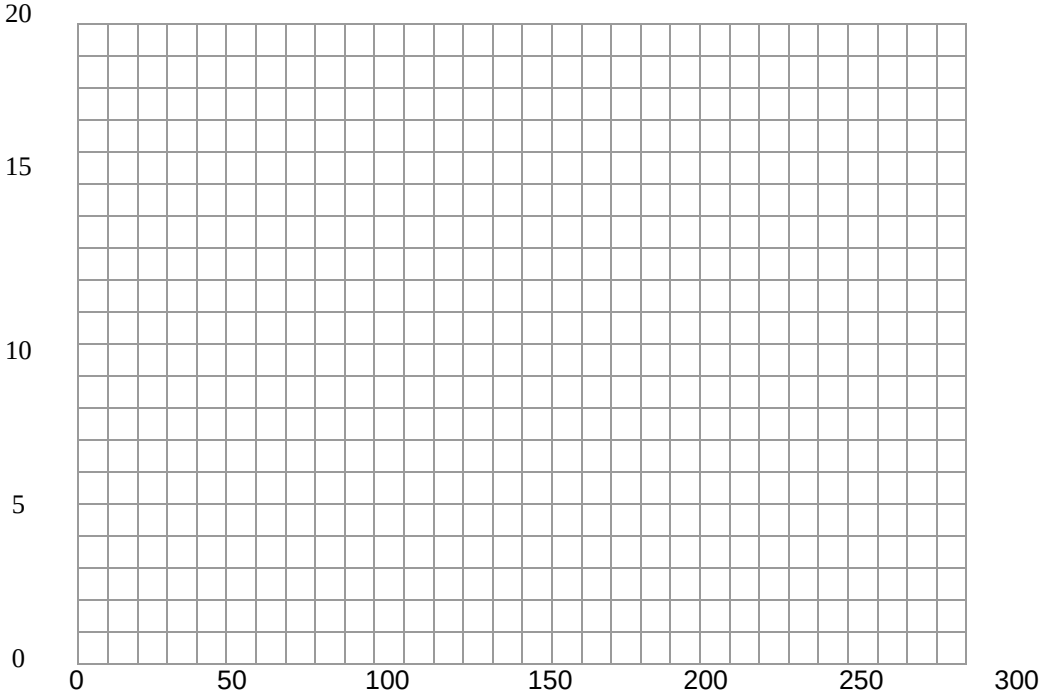
Esto demuestra la equivalencia entre las dos fórmulas definidas para calcular la varianza.

2) Se tiene una muestra aleatoria con datos del costo por consumo de electricidad en una zona residencial de cierta ciudad.

96	171	202	178	147
157	185	90	116	172
141	149	206	175	123
95	163	150	154	130
108	119	183	151	114

Calcule $\bar{X}$, $\tilde{X}$, S^2, S, Q_1, Q_3, R, D_1, D_5

3) Se tienen los siguientes datos de la cantidad de barriles por día que producen 20 pozos petroleros en un campo: cantidad mínima: 45; cantidad máxima 265; primer cuartil 85; mediana 160; tercer cuartil 205. Grafique la Ojiva con la mayor precisión que le sea posible.



4) Respecto al problema anterior. Una compañía está interesada en comprar solamente los pozos que produzcan mas de 100 barriles por día y pagará $150000 por cada uno. ¿Cuanto le costaría la inversión aproximadamente?

MATLAB

Fórmulas para estadística descriptiva

```
>> x=[2 6 11 8 11 4 7 5];          Vector con los datos de una muestra
>> xb=mean(x)                      Media aritmética
   xb =
        6.7500
>> m=median(x)                     Mediana
   m =
        6.5000
>> x=0:1:100;                      Vector con los primeros 100 números naturales
>> xb=mean(x)                      Media aritmética
   xb =
        50
>> x=[x 1000];                     Vector con un valor grande agregado al final
>> xb=mean(x)                      Media aritmética
   xb =
        59.3137
>> xb=trimmean(x,10)               Media aritmética omitiendo 5% de datos en cada lado
   xb =
        50.5000
>> x=[2 6 11 8 11 4 7 5];          Vector con los datos de una muestra
>> r=range(x)                      Rango de los datos
   r =
        9
>> a=min(x)                        El menor valor
   a =
        2
>> b=max(x)                        El mayor valor
   b =
        11
>> s2=var(x)                       Varianza muestral
   s2 =
        10.2143
>> s=std(x)                        Desviación estándar muestral
   s =
        3.1960
>> rq=iqr(x)                       Rango intercuartil
   rq =
        5
>> q1=prctile(x,25)                Primer cuartil (percentil 25)
   q1 =
      4.5000
>> q3=prctile(x,75)                Tercer cuartil (percentil 75)
   q3 =
      9.5000
>> y=sort(x)                       Datos ordenados en forma creciente
   y =
     2    4    5    6    7    8   11   11
>> x=rand(1,400);                  Vector con una fila de 400 números aleatorios
>> d7=prctile(x,70)                Decil 7  (percentil 70)
   d7 =
        0.7013
>> p82=prctile(x,82)               Percentil 82
   p82 =  0.8335
```

2.10 FÓRMULAS PARA DATOS AGRUPADOS

Si los datos de una muestra están disponibles únicamente en una Tabla de Frecuencias, se pueden usar fórmulas para calcular las medidas estadísticas descriptivas, en forma aproximada

Suponer que se dispone de la Tabla de Frecuencias con los valores que se indican en forma simbólica:

Número	Clase	Marca	f	F	f/n	F/n
1	$[a_1, b_1]$	m_1	f_1	F_1	f_1/n	F_1/n
2	$[a_2, b_2]$	m_2	f_2	F_2	f_2/n	F_2/n
. . .	. . .	. . .	. . .	. . .	. . .	. . .
k	$[a_k, b_k]$	m_k	f_k	F_k	f_k/n	F_k/n

Definición: Media de datos agrupados

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{k} m_i f_i$$

n número de datos
k número de clases
m_i marca de la clase i (es el valor central del intervalo de la clase)
f_i frecuencia de la clase i

Definición: Varianza de datos agrupados

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{k} f_i (m_i - \overline{X})^2$$

n número de datos
k número de clases
m_i marca de la clase i (es el centro del intervalo de la clase)
f_i frecuencia de la clase i

Ejemplo: La Tabla de Frecuencias siguiente contiene los datos agrupados en 6 clases del número de artículos vendidos por un almacén en 50 días. Calcule la media y varianza

Número	Clase	Marca	f	F	f/n	F/n
1	[10, 20)	15	2	2	0.04	0.04
2	[20, 30)	25	10	12	0.2	0.24
3	[30, 40)	35	12	24	0.24	0.48
4	[40, 50)	45	14	38	0.28	0.76
5	[50, 60)	55	9	47	0.18	0.94
6	[60, 70)	65	3	50	0.06	1

Media

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{k} m_i f_i = \frac{1}{50}[(15)(2) + (25)(10) + \ldots + (65)(3)] = 40.4$$

Varianza

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{k} f_i (m_i - \overline{X})^2$$

$$= \frac{1}{49}[2(15 - 40.4)^2 + 10(25 - 40.4)^2 + \ldots + 3(65 - 40.4)^2] = 164.12$$

Para comparar, se tienen los datos originales de los cuales se obtuvo la Tabla de Frecuencias:

37	48	48	57	32	63	55	34	48	36
32	47	50	46	28	19	29	33	53	68
49	26	20	63	20	41	35	38	35	25
23	38	43	43	45	54	58	53	49	32
36	45	43	12	21	55	50	27	24	42

Con estos datos, los resultados calculados son:

$\overline{X} = 40.16$

$S^2 = 169.81$

Hay una diferencia, aunque no muy grande, por el uso de las fórmulas con datos agrupados

Ejemplo. Se dispone de los siguientes datos incompletos en una Tabla de Frecuencias

Número	Clase	Marca	f	F	f/n	F/n
1	[1, 2)		1			
2				6		
3					0.25	
4						0.7
5			8			0.9
6					0.05	
7						

Completar la Tabla de Frecuencias

Solución

Se escriben directamente los intervalos, marcas de clase y algunos valores de frecuencia que se pueden determinar observando los datos dados y con las definiciones establecidas

Numero	Clase	Marca	f	F	f/n	F/n
1	[1, 2)	1.5	1	1		
2	[2, 3)	2.5	5	6		
3	[3, 4)	3.5			0.25	
4	[4, 5)	4.5				0.7
5	[5, 6)	5.5	8		0.2	0.9
6	[6, 7)	6.5			0.05	0.95
7	[7, 8)	7.5			0.05	1

Para continuar usamos la siguiente relación contenida en la tabla: **8/n = 0.2**
De donde se obtiene que **n = 40**. Conocido el valor de **n,** se puede continuar desde arriba

Número	Clase	Marca	f	F	f/n	F/n
1	**[1, 2)**	1.5	**1**	1	0.025	0.025
2	[2, 3)	2.5	5	**6**	0.125	0.15
3	[3, 4)	3.5			**0.25**	0.40
4	[4, 5)	4.5			0.3	**0.7**
5	[5, 6)	5.5	**8**		0.2	**0.9**
6	[6, 7)	6.5			**0.05**	0.95
7	[7, 8)	7.5			0.05	1

Finalmente, con la definición de frecuencia relativa **F = f/n** se puede completar la tabla

Número	Clase	Marca	f	F	f/n	F/n
1	**[1, 2)**	1.5	**1**	1	0.025	0.025
2	[2, 3)	2.5	5	**6**	0.125	0.15
3	[3, 4)	3.5	10	16	**0.25**	0.40
4	[4, 5)	4.5	12	28	0.3	**0.7**
5	[5, 6)	5.5	**8**	36	0.2	**0.9**
6	[6, 7)	6.5	2	38	**0.05**	0.95
7	[7, 8)	7.5	2	40	0.05	1

Calcular la media y varianza

Con las fórmulas correspondientes se pueden calcular las medidas descriptivas indicadas igual que en el ejemplo anterior

2.10.1 EJERCICIOS

Se dispone de los siguientes datos incompletos en una Tabla de Frecuencias

Número	Clase	Marca	f	F	f/n	F/n
1				**2**		
2						**0.25**
3	**[15, 20)**		**14**			**0.6**
4						
5				**36**		
6						**0.975**
7						

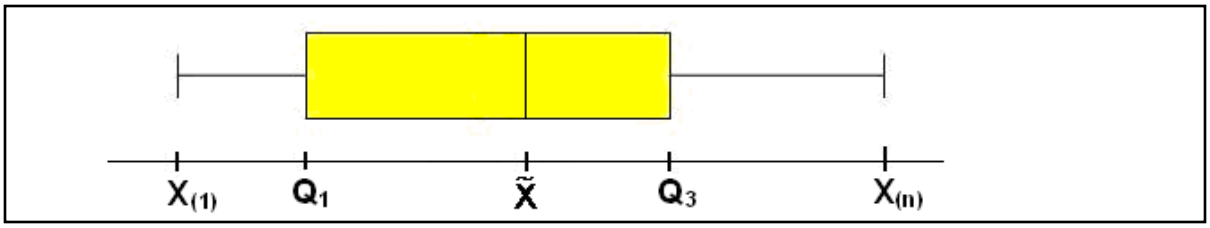
Se conoce además que la media calculada con los datos agrupados es **19.7**

a) Complete la Tabla de Frecuencias
b) Calcule la media y varianza

2.11 INSTRUMENTOS GRÁFICOS ADICIONALES

2.11.1 DIAGRAMA DE CAJA

Es un dispositivo gráfico que se usa para expresar en forma resumida, algunas medidas estadísticas de posición:



El diagrama de caja describe gráficamente el rango de los datos, el rango intercuartílico ($Q_3 - Q_1$) los valores extremos y la ubicación de los cuartiles. Es una representación útil para comparar grupos de datos. Por ejemplo se resalta el hecho que el 50% de los datos está en la región central entre los valores de los cuartiles Q_1 y Q_3

2.11.2 DIAGRAMA DE PUNTOS

Si la cantidad de datos es pequeña, (alrededor de 20 o menos), se los puede representar mediante puntos directamente sin agruparlos en intervalos.

2.11.3 DIAGRAMA DE PARETO

Es un gráfico útil para identificar las causas principales que producen cierto tipo de resultados. La **Ley de Pareto** dice que de cualquier conjunto de eventos que pueden asociarse a un suceso, solamente unos pocos contribuyen en forma significativa mientras que los demás son secundarios. Generalmente hay únicamente 2 o 3 causas que explican mas de la mitad de las ocurrencias del suceso.

Procedimiento para construir el diagrama de Pareto

1) Categorice los datos por tipo de problema
2) Determine la frecuencia y ordene en forma decreciente
3) Represente la frecuencia relativa con barras
4) Superponga la ojiva de la frecuencia relativa acumulada
5) Analice cuales son las causas mas importantes que inciden en el suceso de interés

Ejemplo

Un fabricante ha realizado un conteo de los tipos de defectos de sus productos y ha registrado su frecuencia. Se desea analizar su incidencia en la producción con un Diagrama de Pareto.

Los resultados, tabulados según el procedimiento anterior son:

Tipo de Defecto	Frecuencia	Frecuencia relativa (%)	Frecuencia acumulada	Frecuencia acumulada relativa (%)
A	66	0.33	66	0.33
B	44	0.22	110	0.55
C	34	0.17	144	0.72
D	20	0.10	164	0.82
E	14	0.07	178	0.89
F	12	0.06	190	0.95
G	10	0.05	200	1.00

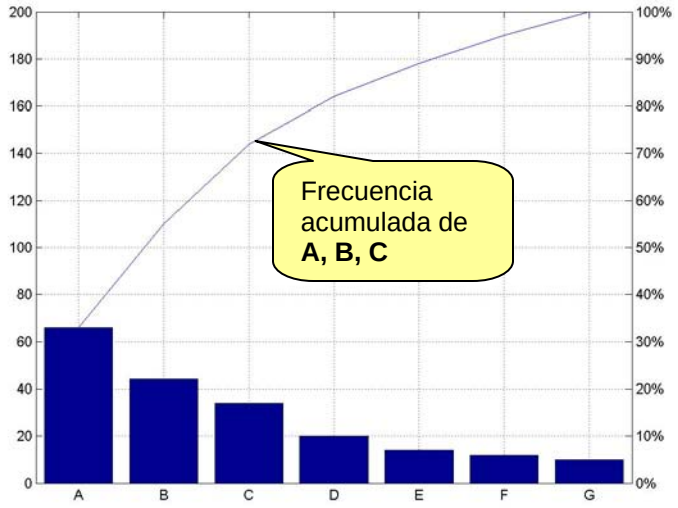


Diagrama de Pareto

Se puede observar que más del 70% de los defectos de producción corresponden a los tipos **A**, **B** y **C**. Con esta información, una decisión adecuada sería asignar recursos para solucionar estos tipos de problemas pues son los que tienen mayor incidencia en la producción.

2.11.4 DIAGRAMA DE TALLO Y HOJAS

Es un dispositivo utilizado cuando la cantidad de datos es pequeña. Permite describir la distribución de frecuencia de los datos agrupados pero sin perder la información individual de los datos.

La longitud de cada fila ayuda a visualizar la frecuencia, en forma parecida a un histograma pero al mismo tiempo se pueden observar individualmente los datos.

Se construye escribiendo verticalmente las primera(s) cifra(s) de los datos (**tallo**) y escribiendo las restantes cifras horizontalmente (**hojas**).

Ejemplo. Los siguientes datos corresponden a la cantidad de artículos defectuosos producidos en una fábrica en 20 días:

65, 36, 59, 84, 79, 56, 28, 43, 67, 36, 43, 78, 37, 40, 68, 72, 55, 62, 22, 82

Dibuje el diagrama de tallo y hojas

Se elige la cifra de las decenas como tallo y la cifra de las unidades como las hojas:

Tallo	Hojas			
2	**2**	**8**		
3	**6**	**6**	**7**	
4	**0**	**3**	**3**	
5	**5**	**6**	**9**	
6	**2**	**5**	**7**	**8**
7	**2**	**8**	**9**	
8	**2**	**4**		

2.11.5 EJERCICIOS

1) Dibuje un Diagrama de Caja para los siguientes datos

1.42	**1.26**	**1.10**	**1.33**	**1.41**
1.00	**1.34**	**1.18**	**1.41**	**1.25**
1.35	**1.21**	**1.81**	**1.65**	**1.18**

2) Dibuje un Diagrama de Pareto con los siguientes datos

46 4 26 15 52 2 5

Tipo	Frecuencia	Frecuencia relativa (%)	Frecuencia acumulada	Frecuencia acumulada relativa (%)
A				
B				
C				
D				
E				
F				
G				

3) Realice un Diagrama de Tallo y Hojas con los siguientes datos

8.3 4.5 9.5 1.4 8.6 7.6 4.4 6.2 9.5 6.4 2.4 3.5 1.8 4.9 4.0
4.6 6.1 8.7 3.1 6.0 1.7 6.2 2.4 5.8 5.0 4.6 5.4 9.4 3.4 4.0
3.0 4.1 2.8 3.9 5.0 7.2 3.0 1.1 4.4 4.6 7.1 6.6 7.2 2.8 2.6

Tallo	Hojas									

4) Un fabricante de cierto componente electrónico se interesa en determinar el tiempo de vida (en horas) de estos dispositivos, para lo cual ha tomado una muestra de **12** observaciones:

123, 116, 120, 130, 122, 110, 175, 126, 125, 110, 119, ?

Uno de los datos se ha extraviado pero se conoce que la media de los **12** datos es **124** horas.

a) Encuentre el dato faltante
b) Calcule la mediana, primer y tercer cuartil
c) Encuentre el rango, varianza y desviación estándar
d) Dibuje el diagrama de caja

MATLAB

Dibujar un diagrama de Pareto para los siguientes datos

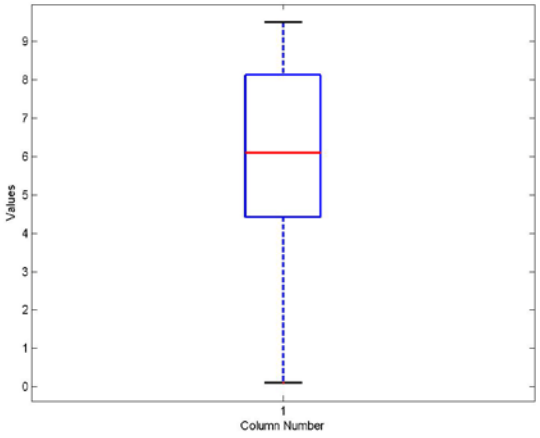
>> x = [66 44 34 20 14 12 10];	Vector con los datos
>> nombres = {'A' 'B' 'C' 'D' 'E' 'F','G'};	Nombres para los componentes en el diagrama
>> pareto(x, nombres)	Dibujar el diagrama de Pareto
>> grid on	Agregar cuadrículas

El dibujo resultante se muestra en la página anterior

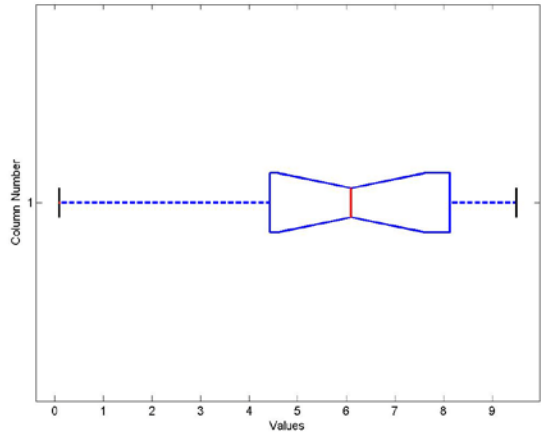
Dibujar un diagrama de caja

>> x = [0.1 1.7 2.3 4.4 4.5 4.8 6.0 6.1 7.3 7.6 7.9 8.2 8.9 9.2 9.5]; Vector con datos

>> boxplot(x) Diagrama de caja



>> boxplot(x, 1, '', 0) Diagrama de caja horizontal, con muesca



2.12 MUESTRAS BIVARIADAS

Es común tener que estudiar muestras con datos que miden dos características, siendo de interés determinar si hay alguna relación entre ellas.

Para visualizar la relación entre las variables de una muestra bivariada, es útil graficar los datos en una representación que se denomina Diagrama de Dispersión.

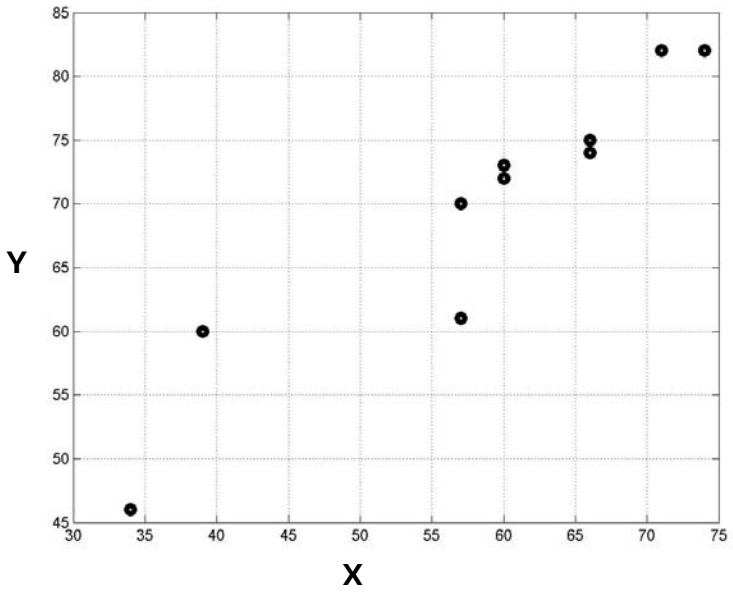
Introducimos este importante concepto mediante un ejemplo

Ejemplo 2.1
Se tiene una muestra con las calificaciones de 10 estudiantes de sus exámenes parcial y final.

Examen Parcial	**60**	**74**	**66**	**34**	**60**	**66**	**57**	**71**	**39**	**57**
Examen Final	**72**	**82**	**75**	**46**	**73**	**74**	**70**	**82**	**60**	**61**

Dibuje el Diagrama de Dispersión.

Sean **X**: Calificación del primer parcial (variable independiente)
Y: Calificación del examen final (variable dependiente)



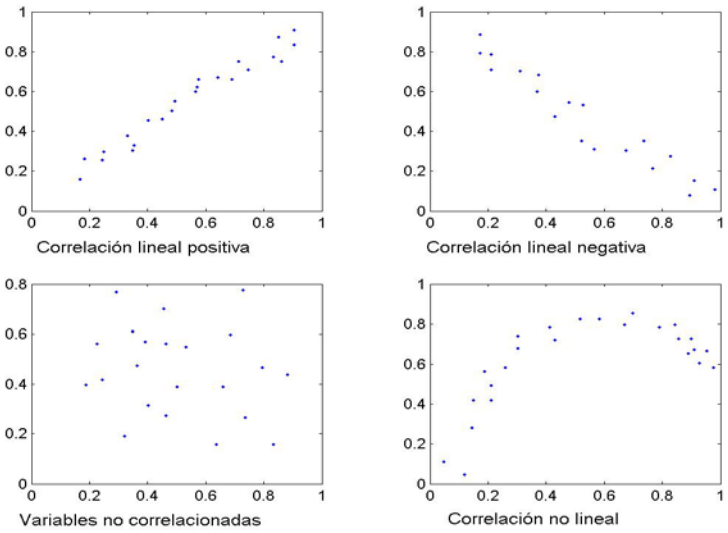
Se observa que los datos están relacionados con una **tendencia lineal** con **pendiente positiva**

En la siguiente sección se definen los instrumentos matemáticos para cuantificar el nivel y el tipo de correlación.

2.12.1 CORRELACIÓN

Se usa el término **correlación** para describir la relación entre los datos de muestras bivariadas.

Los siguientes gráficos son casos típicos para observar la correlación entre dos variables:



Se puede decir que los datos en el **Ejemplo 2.1** tienen **correlación lineal positiva**

2.12.2 COVARIANZA MUESTRAL

Esta definición permite cuantificar el nivel de correlación lineal que existe entre dos variables. Primero anotamos algunas definiciones conocidas para muestras univariadas:

Sean **X, Y**: Variables muestrales
n: Tamaño de la muestra
$\overline{X}$, $\overline{Y}$: Medias aritméticas de **X**, **Y**, respectivamente
S_X^2 , S_Y^2 : Varianzas muestrales de **X**, **Y**, respectivamente
$S_X = \sqrt{S_X^2}$, $S_Y = \sqrt{S_Y^2}$: Desviaciones estándar muestrales de **X, Y** respectivamente

Medias aritméticas muestrales

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \,, \qquad \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

Varianzas muestrales

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \,, \quad S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$$

Ahora se proporciona una definición de variablidad conjunta para muestras con dos variables. Note que si la variable **X** es igual a **Y**, esta fórmula se reduce a la fórmula de varianza:

Definición: Covarianza muestral

S_{XY}: Covarianza muestral

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

2.12.3 SIGNOS DE LA COVARIANZA MUESTRAL

La covarianza es una medida del nivel de correlación entre las variables muestrales **X, Y**.
La covarianza tiene significado si la relación entre las variables es **lineal**.

Si valores grandes de **X** están asociados con valores grandes de **Y**, y si valores pequeños de **X** están asociados con valores pequeños de **Y** entonces **la covarianza tiene signo positivo.** En este caso los datos tienen una tendencia lineal con pendiente positiva.

Si valores grandes de **X** están asociados con valores pequeños de **Y**, y si valores pequeños de **X** están asociados con valores grandes de **Y** entonces **la covarianza tiene signo negativo.** En este caso los datos tienen una tendencia lineal con pendiente negativa

Para entender este comportamiento debemos referirnos a la definición de covarianza:

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Si en las parejas x_i, y_i ambos valores son mayores que su media o ambos valores son menores que su media respectiva, entonces el producto de las diferencias $(x_i - \bar{x})(y_i - \bar{y})$ tendrá signo positivo, y la suma tendrá signo positivo. Pero si en las parejas x_i, y_i, un valor es mayor que su media y el otro valor es menor que su media, entonces el producto de las diferencias $(x_i - \bar{x})(y_i - \bar{y})$ tendrá signo negativo y por lo tanto la suma tendrá signo negativo.

Es importante que se mida la correlación entre variables cuya asociación tenga algún significado de interés. Asimismo, si las variables no están correlacionadas linealmente, pudiera ser que tengan algún otro tipo de correlación, pero no lineal

Es necesario distinguir entre correlación y causalidad. Si dos variables están correlacionadas, esto no implica necesariamente que una sea causa de la otra pues ambas pueden depender de una tercera variable. Aún en el caso de que la correlación represente una causalidad, la estadística solamente permite detectarla y medirla, pero no demostrarla pues esto cae en el ámbito de la ciencia en la que se aplica la estadística

2.12.4 COEFICIENTE DE CORRELACION LINEAL MUESTRAL

Es una definición para cuantificar el grado de correlación lineal entre dos variables en forma adimensional y normalizada.

Definición: Coeficiente de Correlación Lineal

$$r = \frac{S_{XY}}{S_X S_Y}, \quad -1 \le r \le 1$$

Valores referenciales

Valor de **r**	**X** y **Y**
Cercano a **1**	Tienen correlación lineal positiva fuerte
Cercano a **-1**	Tienen correlación lineal negativa fuerte
Cercano a **0**	Tienen correlación lineal muy débil o no están correlacionadas linealmente.

El valor que puede tomar **r**, matemáticamente representa la pendiente de la tendencia de los puntos en el Diagrama de Dispersión.

Consideremos el caso en el que **X**, **Y** son variables con componentes idénticos, tales que: **X = Y**

$$\Rightarrow S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = S_{XX} = S_X^2$$

$$\Rightarrow r = \frac{S_{XY}}{S_X S_Y} = \frac{S_{XX}}{S_X S_X} = \frac{S_X^2}{S_X^2} = 1$$

2.12.5 MATRIZ DE VARIANZAS Y COVARIANZAS

Es una matriz simétrica con la que se pueden representar ordenadamente las varianzas y las covarianzas entre las variables.

Para definirla se puede usar la notación:

$X_1 = X, \quad S_{X_1} = S_X$

$X_2 = Y, \quad S_{X_2} = S_Y$

Definición: Matriz de Varianzas y Covarianzas

$$\left[S_{X_i X_j}\right] = \begin{bmatrix} S_{X_1}^2 & S_{X_1 X_2} \\ S_{X_2 X_1} & S_{X_2}^2 \end{bmatrix}$$

2.12.6 MATRIZ DE CORRELACION

Es una representación ordenada de los coeficientes de correlación de cada variable con la otra variable y consigo misma.

Para definirla se puede usar la notación:

$X_1 = X, \quad S_{X_1} = S_X$

$X_2 = Y, \quad S_{X_2} = S_Y$

Coeficiente de Correlación lineal entre X_i y X_j

$$r_{ij} = \frac{S_{X_i X_j}}{S_{X_i} S_{X_j}}$$

Definición: Matriz de Correlación

$$\left[r_{ij}\right] = \begin{bmatrix} r_{1,1} & r_{1,2} \\ r_{2,1} & r_{2,2} \end{bmatrix}$$

Es una matriz simétrica. Los valores en la diagonal principal son iguales a 1

Las definiciones establecidas para la Matriz de Varianzas-Covarianzas y Matriz de Correlación con dos variables, pueden extenderse directamente a más variables

Ejemplo 2.2
Se tiene una muestra con las calificaciones de 10 estudiantes del primer parcial y del segundo parcial.

Primer Parcial	**60**	**74**	**66**	**34**	**60**	**66**	**57**	**71**	**39**	**57**
Segundo Parcial	**72**	**82**	**75**	**46**	**73**	**74**	**70**	**82**	**60**	**61**

Encuentre el Coeficiente de Correlación Lineal e interprete el resultado

Solución

Sean: **X:** Calificación del primer parcial
Y: Calificación del segundo parcial

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{10}(60+74+66+34+60+66+57+71+39+57) = 58.4$$

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{1}{9}[(60-58.4)^2+(74-58.4)^2+\ldots+(57-58.4)^2] = 166.4889$$

$$s_x = \sqrt{s_X^2} = \sqrt{166.4889} = 12.9031$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{10}(72+82+75+46+73+74+70+82+60+61) = 69.5$$

$$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2 = \frac{1}{9}[(72-69.5)^2+(82-69.5)^2+\ldots+(61-69.5)^2] = 121.8333$$

$$s_Y = \sqrt{s_Y^2} = \sqrt{121.8333} = 11.0378$$

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

$$= \frac{1}{9}[(60-58.4)(72-69.5)+(74-58.4)(82-69.5)+\ldots$$

$$+(57-58.4)(61-69.5)] = 134.1111$$

Coeficiente de Correlación

$$r = \frac{S_{XY}}{S_X S_Y} = \frac{134.1111}{(12.9031)(11.0378)} = 0.9416$$

El resultado indica que la correlación es fuertemente positiva

Escriba las matrices de Varianzas-Covarianzas y de Correlación.

Sean $X_1 = X, \quad S_{X_1} = S_X$

$X_2 = Y, \quad S_{X_2} = S_Y$

Matriz de Varianzas-Covarianzas

$$\left[S_{X_i X_j}\right] = \begin{bmatrix} S^2_{X_1} & S_{X_1 X_2} \\ S_{X_2 X_1} & S^2_{X_2} \end{bmatrix} = \begin{bmatrix} 166.4889 & 134.1111 \\ 134.1111 & 121.8333 \end{bmatrix}$$

Matriz de Correlación

Con la definición:

$$r_{ij} = \frac{S_{X_i X_j}}{S_{X_i} S_{X_j}}$$

Sustituyendo los valores calculados respectivos se obtiene

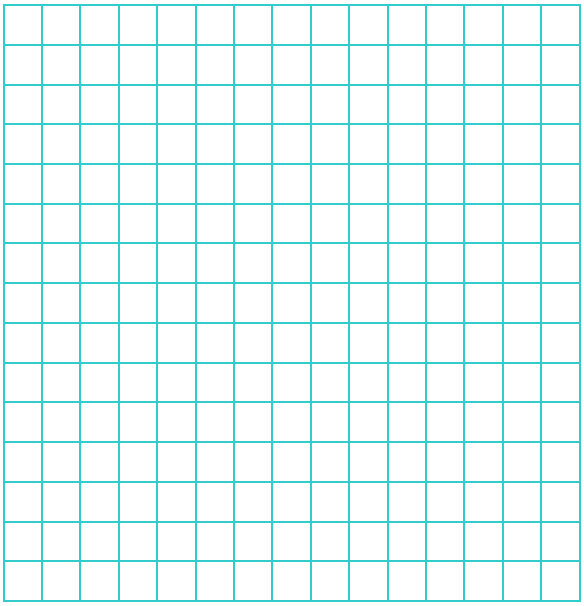
$$\left[r_{ij}\right] = \begin{bmatrix} r_{1,1} & r_{1,2} \\ r_{2,1} & r_{2,2} \end{bmatrix} = \begin{bmatrix} 1 & 0.9416 \\ 0.9416 & 1 \end{bmatrix}$$

2.12.7 EJERCICIOS

Los siguientes datos representan el tiempo de entrenamiento, en horas, que recibieron los trabajadores de una empresa, y el tiempo, en minutos, que posteriormente tardaron en realizar la actividad encomendada

Tiempo de entrenamiento	**10**	**5**	**12**	**8**	**6**	**8**	**4**	**10**
Tiempo que tardaron en la actividad	**9**	**12**	**8**	**10**	**13**	**11**	**12**	**8**

a) Dibuje el Diagrama de Dispersión e indique que tipo de correlación parecen tener las variables **X** y **Y**



b) Escriba la Matriz de Varianzas y Covarianzas

$$\left[S_{X_i X_j}\right] = \begin{bmatrix} S^2_{X_1} & S_{X_1 X_2} \\ S_{X_2 X_1} & S^2_{X_2} \end{bmatrix} =$$

c) Escriba la Matriz de Correlación

$$\left[r_{ij}\right] = \begin{bmatrix} r_{1,1} & r_{1,2} \\ r_{2,1} & r_{2,2} \end{bmatrix} =$$

d) Calcule el Coeficiente de Correlación e interprete el resultado

MATLAB

Vectores con datos de dos variables

```
>> x=[60 74 66 34 60 66 57 71 39 57];
>> y=[72 82 75 46 73 74 70 82 60 61];
```

Diagrama de dispersión. El gráfico aparece en la primera página de esta sección

```
>> scatter(x,y,'k')
>> grid on
```

Matriz de varianzas y covarianzas

```
>> v=cov(x,y)
   v =
      166.4889  134.1111
      134.1111  121.8333
```

Matriz de correlación

```
>> r=corrcoef(x,y)
   r =
      1.0000   0.9416
      0.9416   1.0000
```

Varianza, covarianza y coeficiente de correlación:

```
>> vx = v(1,1)
   vx =
        166.4889
```
Varianza de X

```
>> vy = v(2,2)
   vy =
        121.8333
```
Varianza de Y

```
>> vxy = v(2,1)
   vxy =
        134.1111
```
Covarianza de X, Y

```
>> rxy = r(2,1)
   rxy =
        0.9416
```
Coeficiente de correlación de X, Y

```
>> v=diag(cov(x,y))
   v =
        166.4889
        121.8333
```
Vector con las varianzas (es la diagonal de la matriz)

```
>> s=sqrt(diag(cov(x,y)))
   s =
        12.9031
        11.0378
```
Vector con las desviaciones estándar

3 FUNDAMENTOS DE LA TEORÍA DE LA PROBABILIDAD

En esta unidad se revisan algunas definiciones necesarias para fundamentar el estudio de la Teoría de la Probabilidad.

3.1 FÓRMULAS DE CONTEO

En esta sección revisamos algunas fórmulas básicas para conteo de los elementos de grupos.

Definición: Principio Básico del Conteo

> Si un grupo tiene **m** elementos y otro grupo tiene **n** elementos, entonces existen **mxn** formas diferentes de tomar un elemento del primer grupo y otro elemento del segundo grupo.

Ejemplo. Se lanzan un dado y una moneda. ¿Cuantos resultados diferentes se obtienen en este experimento?

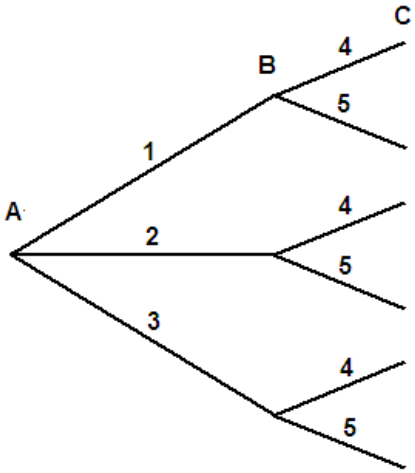
Respuesta: Al lanzar el dado se pueden tener **m** = **6** resultados diferentes, mientras que al lanzar la moneda se obtienen **n** = **2** resultados diferentes. Por lo tanto, el número total de resultados del experimento es **mxn** = **6x2** = **12**. El conjunto de resultados posibles es:

{(1, c), (1, s), (2, c), (2, s), (3, c), (3, s), (4, c), (4, s), (5, c), (5, s), (6, c), (6, s)}, **c**: cara, **s**: sello

Ejemplo: Para ir de su casa a la universidad un estudiante debe ir primero a una estación intermedia de transferencia:
Sean **A**: Casa del estudiante
B: Estación intermedia de transferencia
C: Universidad

Suponga que para ir de **A** hasta **B** hay tres líneas de buses y que para ir desde **B** hasta **C,** puede usar el bus de la universidad o el carro de un amigo. ¿De cuantas formas diferentes puede ir de su casa a la universidad?

Respuesta: Sean **1, 2, 3** las líneas de buses de **A** a **B**, y **4, 5** las formas de ir de **B** a **C**.
Representemos las diferentes opciones mediante un **diagrama de árbol.**



Para ir de **A** a **B** hay 3 formas diferentes. Para ir de **B** a **C** hay 2 formas diferentes.

Por lo tanto, para ir de **A** a **C** hay en total **3x2 = 6**, formas diferentes.

El conjunto de resultados posibles es: **{(1, 4), (1, 5), (2, 4), (2, 5), (3, 4), (3, 5)}**

La fórmula de conteo puede extenderse directamente a más grupos

Ejemplo. Un club de 10 personas debe elegir a su directiva; presidente, secretario, tesorero. Todos pueden ser elegidos, pero una persona no puede tener más de un cargo.
¿De cuantas maneras diferentes puede realizarse la elección?

Respuesta:
Para elegir presidente en el grupo existen 10 opciones distintas.
Para elegir secretario queda un grupo con 9 opciones distintas
Para elegir tesorero queda un grupo con 8 opciones distintas

Por el Principio Básico del Conteo, hay **10** x **9** x **8** = **720** formas diferentes de realizar la elección.

Ejemplo. ¿Cuantos números de placas diferentes pueden existir en la provincia del Guayas?

Respuesta: Cada número de placa tiene la siguiente estructura:
G (letra) (letra) (dígito) (dígito) (dígito)

Hay **26** letras diferentes (sin incluir **ñ**) y **10** dígitos diferentes. Si no importa repetir letras o dígitos en cada placa, el total es: 26 x 26 x 10 x 10 x 10 = **676000**

3.1.1 PERMUTACIONES

Son los arreglos diferentes que se pueden hacer con los elementos de un grupo.

En estos arreglos **se debe considerar el orden** de los elementos incluidos.

Suponga un conjunto de **n** elementos diferentes, del cual se toma un arreglo de **r** elementos.

Si cada arreglo incluye un elemento (**r**=**1**), la cantidad de arreglos diferentes que se obtienen es:
n (Cualquiera de los **n** elementos puede ser elegido)

Si cada arreglo incluye **2** elementos (**r**=**2**), la cantidad de arreglos diferentes que se obtienen es:
n(n-1) (Para elegir el segundo elemento quedan **n – 1** disponibles)

Si cada arreglo incluye **3** elementos (**r**=**3**), la cantidad de arreglos diferentes que se obtienen es:
n(n-1)(n-2) (Para elegir el tercer elemento quedan **n – 2** disponibles)
. . .

Si cada arreglo incluye **r** elementos, entonces la cantidad de arreglos diferentes obtenidos es:
n(n-1)(n-2). . .(n-r+1) (Para elegir el elemento **r** quedan **n – r + 1** disponibles)

Con eso se puede escribir la fórmula general para la cantidad de permutaciones:

Definición: Número de permutaciones

Número de permutaciones con **n** elementos diferentes de un conjunto del cual se toman arreglos conteniendo **r** elementos

$$_nP_r = n(n-1)(n-2)\ldots(n-r+1)$$

Ejemplo. Un grupo de 10 personas debe elegir a su directiva; presidente, secretario, tesorero. Todos pueden ser elegidos, pero una persona no puede tener más de un cargo. ¿De cuantas maneras diferentes puede realizarse la elección? (Use la fórmula de permutaciones)

Respuesta: Los arreglos posibles son permutaciones pues el orden en cada uno si es de interés. Por lo tanto

n = 10, r = 3, $_{10}P_3 = 10x9x8 = 720$

La fórmula de permutaciones se puede expresar en notación factorial completando el producto:

Definición: Fórmula alterna para calcular el número de permutaciones

$$_nP_r = n(n-1)(n-2)\ldots(n-r+1) = \frac{n(n-1)(n-2)\ldots(n-r+1)(n-r)(n-r-1)\ldots(2)(1)}{(n-r)(n-r-1)\ldots(2)(1)} = \frac{n!}{(n-r)!}$$

CASOS ESPECIALES

3.1.2 PERMUTACIONES CON TODOS LOS ELEMENTOS

Definición: Permutaciones con todos los elementos de un conjunto

$$_nP_n = \frac{n!}{(n-n)!} = \frac{n!}{0!} = n!,$$ **n** es la cantidad de elementos del conjunto

Ejemplo: Una máquina desarmada tiene cinco componentes. Para ensamblarla se pueden colocar sus cinco componentes en cualquier orden. ¿Cuantas pruebas diferentes de ensamblaje pueden realizarse?

Respuesta: Son permutaciones con todos los elementos: $_5P_5 = 5! = 120$

3.1.3 ARREGLO CIRCULAR

Suponga un grupo conteniendo **n** elementos diferentes**.** Un arreglo circular es una permutación con todos los elementos del grupo, tal que el primero y el último elemento están conectados. Para que los arreglos sean diferentes, se debe fijar un elemento, mientras que los otros pueden ser intercambiados.

Definición: Número de permutaciones en un arreglo circular

(n-1)!	**n** es el número total de elementos

Ejemplo: ¿De cuantas formas diferentes pueden colocarse 5 personas alrededor de una mesa?

Respuesta: 4! = 24

3.1.4 PERMUTACIONES CON ELEMENTOS REPETIDOS

Si del total de **n** elementos, n_1 fuesen repetidos, entonces los arreglos tendrían formas idénticas cuando se considera el orden de los n_1 elementos repetidos. Existen $n_1!$ formas de tomar los n_1 elementos repetidos, por lo tanto, la cantidad de permutaciones se reduciría por el factor $n_1!$

Definición: Cantidad de permutaciones con elementos repetidos

$\frac{n!}{n_1!}$,	**n** elementos, de los cuales n_1 son repetidos

Este razonamiento, puede extenderse cuando hay más grupos de elementos repetidos

Sean: **n:** Cantidad total de elementos
n_1**:** Cantidad de elementos repetidos de un primer tipo
n_2**:** Cantidad de elementos repetidos de un segundo tipo
Se debe cumplir que $n_1 + n_2 = n$

Definición: Permutaciones con dos tipos de elementos repetidos

$\frac{n!}{n_1!\ n_2!}$,	**n** elementos, de los cuales n_1 son de un tipo y n_2 son de otro tipo

Ejemplo: En una caja hay 3 botellas de vino tinto y 2 de vino blanco. Las botellas de cada uno de los dos tipos de vino tienen la misma marca y forma. ¿De cuantas formas diferentes pueden colocarse en una hilera las 5 botellas?

Respuesta: Son permutaciones con elementos repetidos con $n=5, n_1=3, n_2=2,$

$$\frac{5!}{2!\ 3!} = 10$$

La fórmula se puede generalizar a más grupos con elementos repetidos

Definición: Permutaciones con n elementos y k grupos con elementos repetidos

Sean **n:** Total de elementos, distribuidos en **k** grupos
n_1: Número de elementos repetidos de tipo 1
n_2: Número de elementos repetidos de tipo 2
.
.
n_k: Número de elementos repetidos de tipo **k**

Siendo $n_1 + n_2 + \ldots + n_k = n$
Cantidad de arreglos diferentes que se pueden obtener

$$\frac{n!}{n_1!\,n_2!\ldots n_k!}$$

.
Ejemplo. ¿Cuántos arreglos diferentes pueden hacerse con las letras de la palabra **MATEMÀTICA**?

n=10.

n_1=2 (repeticiones de la letra M)

n_2=3 (repeticiones de la letra A)

n_3=2 (repeticiones de la letra T)

las otras letras ocurren una sola vez

Respuesta: $\frac{10!}{2!\ 3!\ 2!\ 1!\ 1!\ 1!} = 151200$

3.1.5 COMBINACIONES

Son los arreglos que se pueden hacer con los elementos de un conjunto considerando que **el orden de los elementos en cada arreglo no es de interés**.

Cada arreglo se diferencia únicamente por los elementos que contiene, sin importar su ubicación

Sean **n:** Cantidad de elementos del conjunto
r: Cantidad de elementos en cada arreglo

Se usa la notación ${}_nC_r$, o C_r^n, o $\binom{n}{r}$ para denotar la cantidad de combinaciones de tamaño **r** que se pueden realizar con los **n** elementos distintos de un conjunto

Para obtener la fórmula del número de combinaciones, consideremos la fórmula de las permutaciones.

Debido a que en las combinaciones no interesa el orden de los elementos en cada arreglo, es equivalente a tener permutaciones con elementos repetidos. Así se obtiene la fórmula.

Definición: Número de combinaciones

n elementos con los cuales se forman arreglos conteniendo **r** elementos

$$_nC_r = \frac{_nP_r}{r!} = \frac{n!}{(n-r)!\ r!} = \frac{n(n-1)(n-1)...(n-r+1)}{r!}$$

Ejemplo. Un bar dispone de 10 frutas diferentes de las cuales pueden elegirse tres para un batido. ¿De cuantas maneras diferentes puede hacerse la elección?

Respuesta: Son combinaciones pues el orden de las frutas no es de interés.

n=10, r=3, $\Rightarrow$ $_{10}C_3 = \frac{10!}{7!\ 3!} = 120$

Ejemplo. Para probar un test de aptitud debe elegirse una muestra de cinco estudiantes de un curso que contiene 20 estudiantes. ¿De cuantas formas puede tomarse la muestra?

Respuesta: En la muestra no interesa el orden de los estudiantes

n=20, r=5, $\Rightarrow$ $_{20}C_5 = \frac{20!}{15!\ 5!} = 15504$

Ejemplo. De una caja que contiene 6 baterías de las cuales 4 están en buen estado, se extrae una muestra de dos baterías

a) ¿De cuantas formas diferentes se puede tomar la muestra?

Respuesta: n=6, r=2, $\Rightarrow$ $_6C_2 = \frac{6!}{4!\ 2!} = 15$

b) ¿En cuantas de estas muestras, las dos baterías están en buen estado?

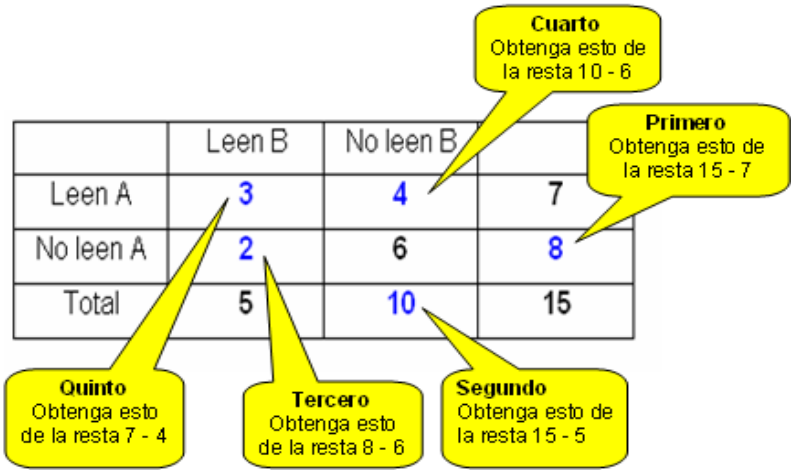
Respuesta: n=4, r=2, $\Rightarrow$ $_4C_2 = \frac{4!}{2!\ 2!} = 6$

Es la cantidad de formas de sacar 2 baterías en buen estado de las 4 existentes

Ejemplo. En un grupo de 15 personas, 7 leen la revista A, 5 leen la revista B y 6 ninguna revista. Encuentre la cantidad de personas que leen al menos una revista

Respuesta. Para el cálculo puede usarse una representación gráfica de conjuntos, pero una representación tabular facilita hallar el número de elementos de cada evento.

Primero colocamos en el cuadro los datos (color negro). y luego completamos el cuadro con los valores faltantes (color azul). Para los cálculos se ha seguido el orden indicado en el dibujo.



	Leen B	No leen B	
Leen A	3	4	7
No leen A	2	6	8
Total	5	10	15

Del cuadro se obtiene directamente que

4 leen A, únicamente
2 leen B, únicamente
3 leen A y B

Por lo tanto, 9 personas leen al menos una revista

Encuentre la cantidad de formas diferentes de elegir cuatro personas que al menos lean una revista

Respuesta: $_9C_4 = \frac{9!}{5!\ 4!} = 126$

Encuentre la cantidad de formas diferentes de elegir cuatro personas de tal manera que dos lean solamente A, una lea solamente B, y una no lea revistas.

Respuesta:

Cantidad de formas diferentes de elegir 2 de las que leen solamente A: $_4C_2 = 6$

Cantidad de formas diferentes de elegir 1 de las que leen solamente B: $_2C_1 = 2$

Cantidad de formas diferentes de elegir 1 de las que no leen revistas: $_6C_1 = 6$

Por el Principio Básico del Conteo el resultado final es: 6 x 2 x 6 = 72

3.1.6 EJERCICIOS

1) Un taller de mantenimiento tiene tres técnicos: A, B, C. Cierto día, dos empresas X, Y requieren un técnico cada una. Describa el conjunto de posibles asignaciones si cada técnico puede ir solamente a una empresa.

2) En el ejercicio anterior, suponga que el mismo técnico debe ir primero a la empresa X y luego a la empresa Y. Describa el conjunto de posibles asignaciones.

3) Hay tres paralelos para el curso de Cálculo Diferencial y tres paralelos para Algebra Lineal. Un estudiante desea tomar ambos cursos. Escriba el conjunto de posibles asignaciones.

4) En un curso preuniversitario los exámenes solían contener 20 preguntas y cada una con cinco opciones. ¿De cuantas formas diferentes se podía contestar el examen?

5) Una caja contiene cinco libros de Matemáticas y una segunda caja contiene 4 libros de Física. ¿De cuantas maneras diferentes se puede tomar un libro para materia? a) si todos los libros son diferentes, b) si los libros de cada materia son iguales

6) Una caja contiene 3 bolas azules y 2 rojas. Una segunda caja contiene dos bolas rojas. De la primera caja se extrae una bola y se la coloca en la segunda caja. Finalmente, de la segunda caja se extraen dos bolas. ¿Cuantos resultados diferentes se pueden obtener al tomar las dos bolas de la segunda caja? ¿En cuantos de estos resultados se obtendrían dos bolas de diferente color?

7) Para un proyecto se requiere dos ingenieros y tres técnicos. Si hay cuatro ingenieros y cinco técnicos disponibles. ¿De cuantas maneras se puede hacer la elección?

8) Una caja contiene 6 baterías de las cuales 2 son defectuosas. ¿De cuantas maneras se pueden tomar tres baterías de tal manera que solamente haya una defectuosa?

9) En un grupo de 60 estudiantes, 42 están registrados en Análisis Numérico, 38 en Estadística y 10 no están registrados en ninguna de estas dos materias. ¿Cuantos están registrados únicamente en Estadística? ¿Cuantos están registrados en Estadística pero no en Análisis Numérico?

10) El cable de seguridad de una bicicleta tiene un candado que contiene 4 discos. Cada disco tiene seis números. Si probar cada combinación toma cinco segundos, determine el tiempo máximo que le tomará a una persona encontrar la clave para quitar el cable de seguridad que sujeta a la bicicleta

MATLAB

```
>> c = nchoosek(9,4)                    Cálculo de 9C4
c  =  126
>> r = factorial(5)                     Factorial de 5
r =   120
>> x=[2 3 5 7];                         Conjunto de 4 elementos
>> lista=combnk(x,3)                    Lista de combinaciones de 3 elementos
lista =
    2    3    5
    2    3    7
    2    5    7
    3    5    7
>> n=length(lista)                      Número de combinaciones
n =  4
>> x=[3 5 7];                           Conjunto de tres elementos
>> lista=perms(x)                       Lista de permutaciones
lista =
    7    5    3
    7    3    5
    5    7    3
    5    3    7
    3    5    7
    3    7    5
>> x = {'Juan', 'Pedro', 'Pablo'};      Conjunto con tres elementos
>> lista=combnk(x,2)                    Lista de combinaciones de 2 elementos
lista =
   'Juan'    'Pedro'
   'Juan'    'Pablo'
   'Pedro'   'Pablo'
```

3.2 EXPERIMENTO ESTADÍSTICO

Es un procedimiento que se realiza con el propósito de obtener observaciones para algún estudio de interés. Un experimento requiere realizar pruebas o ensayos para obtener resultados.

Un Experimento Estadístico tiene las siguientes características:

1. Se conocen **todos los resultados** posibles antes de realizar el experimento.
2. No se puede predecir el resultado de **cada ensayo** realizado (propiedad de aleatoriedad)
3. Debe poderse reproducir o repetir el experimento en condiciones similares.
4. Se puede establecer un patrón predecible a lo largo de muchas ejecuciones del experimento. Esta propiedad se denomina regularidad estadística.

Ejemplos

1) Lanzar un dado y observar el resultado obtenido.
2) Medir la altura de una persona
3) Observar el tipo de defecto de un artículo producido por una fábrica

3.3 ESPACIO MUESTRAL

El Espacio Muestral, representado con la letra **S**, es el conjunto de todos los resultados posibles de un experimento. Cada elemento de **S** se denomina **Punto Muestral**.

Según la naturaleza del experimento, los puntos muestrales pueden determinar que **S** sea **discreto** o **continuo**.

S es **discreto** si sus elementos pueden ponerse en correspondencia con los números naturales. En este caso **S** puede se **finito** o **infinito**.

S es **continuo** si los resultados corresponden a algún intervalo de los números reales. En este caso **S** es **infinito** por definición.

Ejemplos

Experimento:	Lanzar un dado y observar el resultado
Espacio Muestral:	**S={1, 2, 3, 4, 5, 6]**
Propiedades de S:	Discreto y finito
Experimento:	Elegir al azar dos artículos de un lote y observar la cantidad de artículos defectuosos
Espacio Muestral:	**S={0, 1, 2}**
Propiedades de S:	Discreto y finito
Experimento:	Lanzar un dado y contar la cantidad de intentos hasta obtener como resultado el **6**
Espacio Muestral:	**S={1, 2, 3, . . .}**
Propiedades de S:	Discreto e infinito
Experimento:	Medir el peso en gramos de un artículo elegido al azar
Espacio Muestral:	**S={x | x>0, x∈R}**
Propiedades de S:	Continuo (infinito por definición)

3.4 EVENTOS

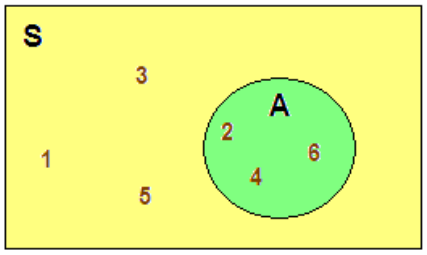
Un evento es algún subconjunto del Espacio Muestral **S**. Se pueden usar letras mayúsculas para denotar eventos: **A, B, . . .** También se pueden usar índices **E_1, E_2, . . .**

Ejemplo:
Experimento: Lanzar un dado y observar el resultado
Espacio Muestral: **S** = {1, 2, 3, 4, 5, 6]

Describa el evento de interés: **A:** el resultado es un número par

Respuesta: **A** = {2, 4, 6}

Representación gráfica con un Diagrama de Venn



Definiciones:

Evento nulo:	No contiene resultados (puntos muestrales)
Evento simple:	Contiene un solo resultado (punto muestral)
Eventos excluyentes:	Eventos que no contienen resultados comunes

3.5 σ-ALGEBRA

El soporte matemático natural para el estudio de las propiedades de los eventos es la Teoría de Conjuntos. Pero existe un álgebra formal específica para su estudio denominada **σ-Algebra** (sigma álgebra).

σ-Algebra $\mathcal{A}$ es una colección no vacía de subconjuntos de **S** tales que

1) $S \in \mathcal{A}$
2) Si $A \in \mathcal{A}$, entonces $A^c \in \mathcal{A}$
3) Si $A_1, A_2, \ldots \in \mathcal{A}$, entonces $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$

En resumen una **σ-Algebra** $\mathcal{A}$ incluye a **S,** a sus subconjuntos y es cerrada con respecto a la operación de unión de conjuntos.

3.6 PROBABILIDAD DE EVENTOS

El valor de la probabilidad de un evento es una medida de la certeza de su realización

Sea **A** un evento, entonces **P(A)** mide la probabilidad de que el evento **A** se realice

P(A)=0	es la certeza de que no se realizará
P(A)=1	es la certeza de que si se realizará
P(A)=0.5	indica igual posibilidad de que se realice o no se realice

3.6.1 Asignación de valores de probabilidad a eventos

1) Empírica

Es la proporción de veces que un evento tuvo el resultado esperado respecto al total de intentos realizados.

Ejemplo. Se han realizado 20 ensayos en un experimento en condiciones similares. Cuatro ensayos tuvieron el resultado esperado. Entonces, la probabilidad que en el siguiente ensayo se obtenga el resultado esperado tiene un valor aproximadamente: **4/20 = 0.2 = 20%**

2) Mediante modelos matemáticos

Para muchas situaciones de interés puede construirse modelos matemáticos con los cuales se puede determinar la probabilidad de eventos. Algunos de estos modelos son estudiados en este curso, tanto para variables discretas como continuas.

3) Asignación clásica

Su origen es la Teoría de Juegos. El valor de probabilidad de un evento es la relación entre la cantidad de resultados que se consideran favorables para el evento de interés, respecto al total de resultados posibles (Espacio Muestral).

Definición: Asignación Clásica de Probabilidad a Eventos

Sean **S**: Espacio muestral
A: Evento de interés .
Si **N(S)** y **N(A)** representan la cardinalidad (número de elementos)
Entonces la probabilidad del evento **A** es: $\mathbf{P(A) = \frac{N(A)}{N(S)}}$

.

Ejemplo. Calcule la probabilidad que al lanzar una vez un dado y una moneda se obtenga un número impar y sello

Si **c, s** representan los valores **cara** y **sello** de la moneda, entonces el espacio muestral es:
S = {(1,c),(2,c),(3,c),(4,c),(5,c),(6,c),(1,s),(2,s),(3,s),(4,s),(5,s),(6,s)}

Mientras que el evento de interés es: **A** = {(1,s),(3,s),(5,s)}

Repuesta: P(A) = N(A)/N(S) = 3/12 = 1/4 = 0.25 = 25%

Ejemplo. En un grupo de **15** personas, **7** leen la revista A, **5** leen la revista B y **6** ninguna revista.

a) Encuentre la probabilidad que al elegir al azar una persona, ésta lea al menos una revista

Respuesta: Representación tabular de datos:

	Leen B	No leen B	
Leen A	**3**	**4**	**7**
No leen A	**2**	**6**	**8**
	5	**10**	**15**

Del cuadro se obtiene que:
4 únicamente leen A
2 únicamente leen B
3 leen A y B
9 personas leen al menos una revista

Sean

E: Evento que la persona elegida al azar lea al menos una revista
S: Conjunto de todas las personas entre las que se puede elegir una.

Entonces

P(E) = N(E)/N(S) = 9/15 = 0.6

b) Encuentre la probabilidad que al elegir al azar tres personas, dos lean ambas revistas y una no lea revistas.

Respuesta:
Sean

E: Evento que dos personas lean ambas revistas y una no lea revistas
S: Incluye todas las formas diferentes de elegir tres personas
N(S) = $_{15}C_3$ = 455

Cantidad de formas diferentes de elegir 2 de las 3 que leen ambas

$_3C_2$ = 3

Cantidad de formas diferentes de elegir 1 de las 6 que no leen revistas

$_6C_1$ = 6

Por el Principio Básico del Conteo, la cantidad de elementos en el evento E

N(E) = 3 x **6** = **18**

Por lo tanto

P(E) = N(E)/N(S) = 18/455 = 0.0396 = 3.96%

Ejemplo. Suponga que se ha vendido una serie completa de las tablas del Peso Millonario. Cada tabla es diferente y contiene 15 números diferentes elegidos al azar entre los enteros del 1 al 25. Calcule la probabilidad que al comprar una tabla esta sea la tabla ganadora.

Respuesta:

Sea **S**: conjunto de tablas del Peso Millonario

N(S) = $_{25}C_{15}$ = 3268760 (Cantidad de tablas diferentes que se generan)

Sea **E**: evento de tener la tabla premiada (solamente hay una tabla premiada)

P(E) = N(E)/N(S) = 1/3268760 ≅ 0.0000003 (Cercano a cero)

Para tomar una idea de lo pequeño que es este número imagine cual sería su chance de sacar el premio si en una caja hubiesen **1000** tablas entre las que está la tabla ganadora.

Si usted debe elegir al azar una tabla y obtener la tabla ganadora, es muy poco probable que acierte.

Ahora suponga que en una bodega hay **3268** cajas, cada una con **1000** tablas. Primero usted debe elegir al azar la caja que contiene la tabla ganadora, y luego de esta caja elegir al azar una tabla esperando que esta sea la tabla ganadora.

Se puede concluir que la probabilidad del evento de obtener el premio es insignificante.

3.6.2 Probabilidad de Eventos Simples

Un Evento Simple incluye un solo punto muestral. Un evento cualquiera **A** de **S** puede considerarse entonces como la unión de sus eventos simples.

Definición: Probabilidad de Eventos Simples

Sean **S**: Espacio muestral, con **n** puntos muestrales
A: Evento cualquiera de **S** con **k** puntos muestrales
$\mathbf{E_1, E_2, \ldots, E_k}$: Eventos simples incluidos en **A**

Entonces

$$\mathbf{P(A) = P(E_1 \cup E_2 \cup \ldots \cup E_k) = P(E_1) + P(E_2) + \ldots + P(E_k)}$$

Si cada evento simple tiene igual probabilidad, entonces

$$\mathbf{P(A) = k\,(1/n)}$$

Ejemplo. ¿Cual es la probabilidad que al lanzar un dado se obtenga un número par?

Respuesta: **S** = {1, 2, 3, 4, 5, 6}
A = {2, 4, 6} Evento de interés
$\mathbf{E_1 = \{2\},\ E_2 = \{4\},\ E_3 = \{6\}}$: Eventos simples incluidos en el evento **A**

Entonces:

$$\mathbf{P(A) = P(E_1 \cup E_2 \cup E_3) = P(E_1) + P(E_2) + P(E_3)} = 3\,(1/6) = 0.5$$

Ejemplo. Suponga que un dado está desbalanceado de tal manera que se conoce que la probabilidad que salga el número 6 es el doble que los otros números. ¿Cual es la probabilidad que al lanzarlo salga un número par?

Respuesta: En este ejemplo los puntos muestrales no tienen la misma probabilidad (1/6).

Sea **x** la probabilidad que salga alguno de los números 1, 2, 3, 4, 5. Por lo tanto, la probabilidad que salga el número 6 es el doble, **2x**

Entonces $\mathbf{x + x + x + x + x + 2x = 1 \Rightarrow x = 1/7}$

Sean **A** = {2, 4, 6}: Evento que salga un número par
$\mathbf{E_1 = \{2\},\ E_2 = \{4\},\ E_3 = \{6\}}$: Eventos simples incluidos en **A**

$$\mathbf{P(A) = P(E_1) + P(E_2) + P(E_3) = 1/7 + 1/7 + 2/7 = 4/7}$$

Ejemplo. De una caja que contiene 6 baterías de las cuales 4 están en buen estado, se extrae una muestra de dos baterías

Calcule la probabilidad que ambas baterías en la muestra estén en buen estado.

Respuesta:
Cantidad total de muestras que se pueden obtener:

$$\mathbf{N(S) = {}_6C_2 = \frac{6!}{4!\ 2!} = 15}$$

Sea **E:** Evento correspondiente a la obtención de una muestra con ambas baterías buenas

Cantidad total de muestras en las que ambas baterías están en buen estado

$$\mathbf{N(E) = {}_4C_2 = \frac{4!}{2!\ 2!} = 6}$$

Entonces

$$\mathbf{P(E) = N(E)/N(S) = 6/15 = 0.4}$$

3.7 AXIOMAS DE PROBABILIDAD DE EVENTOS

En esta sección se introduce la formalidad matemática necesaria para fundamentar la Teoría de la Probabilidad de Eventos.

Sea **S:** Espacio muestral
E: Evento de **S**
P(E): Probabilidad del evento **E**
$\mathfrak{R}$: Conjunto de los reales

Sea **P** una función que asocia a cada evento **E** de **S** un número real:

Definición: Función de Probabilidad de un Evento

$$\mathbf{P:\ S \to \mathfrak{R}}$$
$$\mathbf{E \to P(E)},\quad \text{dom } \mathbf{P} = \mathbf{S},\ \text{rg } \mathbf{P} = [0, 1]$$

P se denomina Función de Probabilidad de un Evento y cumple los siguientes axiomas

Axiomas de Probabilidad de Eventos

1) $P(E) \geq 0$
2) $P(S) = 1$
3) $E_1, E_2 \in S \ \wedge\ E_1 \cap E_2 = \varnothing \ \Rightarrow\ P(E_1 \cup E_2) = P(E_1) + P(E_2)$

El primer axioma indica que la probabilidad de un evento no puede tener valores negativos. El segundo axioma establece que la probabilidad de que un resultado pertenezca al espacio muestral es 1, lo cual es evidente pues **S** contiene todos los resultados posibles.

El tercer axioma establece que si dos eventos son **mutuamente excluyentes** entonces la probabilidad del evento que resulta de la unión de estos eventos, es la suma de las probabilidades de ambos eventos.

3.8 PROPIEDADES DE LA PROBABILIDAD DE EVENTOS

Con los axiomas establecidos se pueden demostrar algunas propiedades de interés, para los eventos de un espacio muestral **S**.

3.8.1 Demostraciones basadas en axiomas de probabilidad

a) Probabilidad de un Evento Nulo: $P(\varnothing) = 0$

Demostración:

$S = S \cup \varnothing$	eventos excluyentes
$\Rightarrow P(S) = P(S) + P(\varnothing)$	por el Axioma 3
$\Rightarrow 1 = 1 + P(\varnothing)$	por el Axioma 2
$\Rightarrow P(\varnothing) = 0$	

b) Probabilidad del Evento Complemento: $P(E^c) = 1 - P(E)$

Demostración:

$S = E \cup E^c$	eventos excluyentes
$\Rightarrow P(S) = P(E) + P(E^c)$	por el Axioma 3
$\Rightarrow 1 = P(E) + P(E^c)$	por el Axioma 2
$\Rightarrow P(E^c) = 1 - P(E)$	

c) Probabilidad de Eventos Incluidos: Si $A \subset B$, entonces $P(A) \leq P(B)$

Demostración:

Sean A, B eventos de S
Si A está incluido en B se puede escribir

$B = A \cup (A^C \cap B)$	eventos excluyentes
$P(B) = P(A) + P(A^C \cap B)$	por el Axioma 3
$P(B) \geq P(A)$	por el Axioma 1

d) La probabilidad de un Evento está entre 0 y 1: $\mathbf{0 \le P(E) \le 1}$

Demostración Sea **E** un evento cualquiera de S, entonces

$\varnothing \subset E \subset S$

$P(\varnothing) \le P(E) \le P(S)$ por la Propiedad 3

$0 \le P(E) \le 1$ por la Propiedad 1 y Axioma 2

e) Probabilidad de la Diferencia de Eventos:

$$\mathbf{P(A - B) = P(A) - P(A \cap B) = P(A \cap B^c)}$$

Demostración: $A = (A - B) \cup (A \cap B)$ eventos excluyentes

$\Rightarrow P(A) = P(A - B) + P(A \cap B)$ por el Axioma 3

$\Rightarrow P(A - B) = P(A) - P(A \cap B) = P(A \cap B^c)$

f) Regla Aditiva de Probabilidad de Eventos:

$$\mathbf{P(A \cup B) = P(A) + P(B) - P(A \cap B)}$$

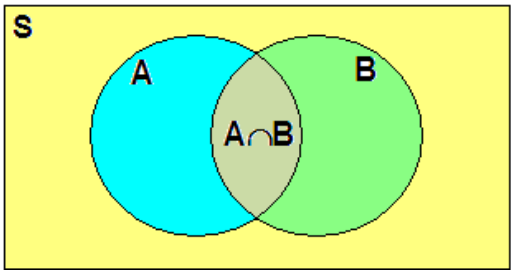
Demostración: $A \cup B = (A - B) \cup (A \cap B) \cup (B - A)$ eventos excluyentes

$\Rightarrow P(A \cup B) = P(A - B) + P(A \cap B) + P(B - A)$ por el Axioma 3

$\Rightarrow P(A \cup B) = P(A - B) + P(A \cap B) + P(B - A) + P(A \cap B) - P(A \cap B$

$\Rightarrow P(A \cup B) = P(A) + P(B) - P(A \cap B)$ con la Propiedad 5

Representación gráfica con un Diagrama de Venn de la Regla Aditiva de Probabilidad



Ejemplo. Si la probabilidad que un estudiante apruebe Álgebra Lineal es 0.7, la probabilidad que apruebe Ingles es 0.8 y la probabilidad que apruebe ambas materias es 0.6, ¿cual es la probabilidad que el estudiante apruebe al menos una de estas dos materias?

Respuesta:

Sean los eventos

A: El estudiante aprueba Álgebra Lineal

B: El estudiantes aprueba Ingles

A∩B: El estudiante aprueba ambas materias

S: Conjunto de todos los estudiantes

A y **B no son eventos excluyentes**, entonces, por la Regla Aditiva de Probabilidad se tiene

$\mathbf{P(A \cup B)} = \mathbf{P(A) + P(B) - P(A \cap B)} = 0.7 + 0.8 - 0.6 = 0.9$

Ejemplo. Sean **A, B** eventos de **S**, tales que $\mathbf{P(A) = 0.35,\ P(B^c) = 0.27,\ P(A^c \cap B) = 0.59}$

Calcule:

a) $\mathbf{P(A \cap B)}$

b) $\mathbf{P(A \cup B)}$

c) $\mathbf{P(A \cup B^c)}$

d) $\mathbf{P(A^c \cup B^c)}$

Respuesta:
Una representación tabular de los valores de probabilidad facilita los cálculos.

	B	**B^c**	
A	**0.14**	**0.21**	**0.35**
A^c	**0.59**	**0.06**	**0.65**
	0.73	**0.27**	**1**

Cada respuesta se la obtiene directamente de la tabla:

a) **$P(A \cap B)$** = 0.14
b) **$P(A \cup B)$** = P(A) + P(B) - $P(A \cap B)$ = 0.35 + 0.73 - 0.14 = 0.94
c) **$P(A \cup B^c)$** = P(A) + $P(B^c)$ – $P(A \cap B^c)$ = 0.35 + 0.27 – 0.21 = 0.41
d) **$P(A^c \cup B^c)$** = $P(A^c)$ + $P(B^c)$ – $P(A^c \cap B^c)$= 0.65 + 0.27 – 0.06 = 0.86

Ejemplo. En un grupo de **15** personas, **7** leen la revista A, **5** leen la revista B y **6** ninguna revista. Encuentre la probabilidad que al elegir al azar una persona, ésta lea al menos una revista

Respuesta: Representación tabular para los datos:

	Leen B	No leen B	
Leen A	**3**	**4**	**7**
No leen A	**2**	**6**	**8**
	5	**10**	**15**

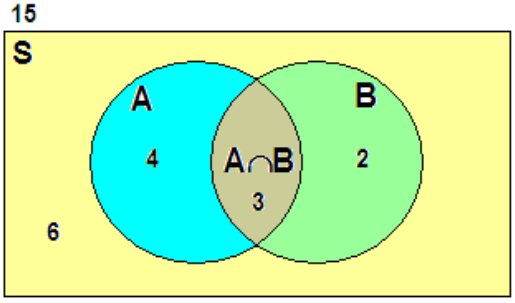
4 únicamente leen A
2 únicamente leen B
3 leen A y B

Entonces, **9** personas leen al menos una revista

Sean los eventos
A: La persona elegida al azar lee la revista A
B: La persona elegida al azar lee la revista B

$A \cup B$: La persona elegida al azar lee al menos una revista
$A \cap B$: La persona elegida al azar no lee ni la revista A ni la revista B
S: Conjunto de las 15 personas

Representación gráfica con un Diagrama de Venn



Por lo tanto, con la Regla Aditiva de Probabilidad,

P(AUB) = **P(A) + P(B) – $P(A \cap B)$** = 7/15 + 5/15 – 3/15 = 9/15 = 0.6

Las propiedades pueden extenderse a más eventos

Sean **A, B, C,** tres eventos del espacio muestral **S**

Definición: Regla Aditiva de Probabilidad para tres Eventos

Si **A, B, C** son eventos mutuamente excluyentes,

$P(A \cup B \cup C) = P(A) + P(B) + P(C)$

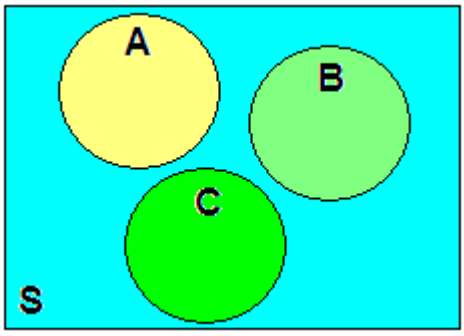
Si **A, B, C** son eventos cualesquiera,

$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$

Ejemplo. Si la probabilidad que Juan vaya al estadio, al cine o a estudiar son respectivamente 0.3, 0.2, 0.4, ¿cual es la probabilidad de que no haga alguna de estas tres actividades?

Respuesta:

Sean **A, B, C** los eventos de que vaya al estadio, al cine o a estudiar



S: Conjunto de todas las actividades que puede realizar Juan

Siendo estos **eventos mutuamente excluyentes**, la probabilidad es

$P(A \cup B \cup C) = P(A) + P(B) + P(C)$ = 0.3 + 0. 2 + 0.4 = 0.9

Por lo tanto, la probabilidad de que no haga alguna de estas tres actividades es

$P(A \cup B \cup C)^C$ = 1 – **$P(A \cup B \cup C)$** = 1 – 0.9 = 0.1

3.8.2 EJERCICIOS

1) En una fábrica hay cinco motores, de los cuales tres están defectuosos. Calcule la probabilidad que al elegir dos motores al azar,

a) Ambos estén en buen estado
b) Solamente uno esté en buen estado
c) Al menos uno esté en buen estado

2) En un grupo de 60 estudiantes, 42 están registrados en Análisis Numérico, 38 en Estadística y 10 no están registrados en ninguna de estas dos materias. Calcule la probabilidad que al elegir entre los 60 algún estudiante al azar,

a) Esté registrado únicamente en Estadística
b) Esté registrado en ambas materias

3) Sean A, B eventos cualesquiera de un espacio muestral.
Si $P(A)=0.34$, $P(B)=0.68$, $P(A\cap B)=0.15$, calcule

a) $P(A\cup B)$
b) $P(A\cap B^c)$
c) $P(A^c\cup B^c)$

4) En una encuesta en la ciudad se ha hallado que

La probabilidad que una familia tenga TV es 0.7
La probabilidad que una familia tenga reproductor de DVD es 0.4
La probabilidad que una familia tenga TV pero no tenga reproductor de DVD es 0.36
Calcule la probabilidad que una familia tenga ni TV ni reproductor de DVD

a) Use una representación tabular
b) Use únicamente reglas de probabilidad

3.9 PROBABILIDAD CONDICIONAL

La probabilidad de un evento puede depender o estar condicionada al valor de probabilidad de otro evento. Introducimos este concepto con un ejemplo:

Ejemplo. Un experimento consiste en lanzar una vez un dado y una moneda. Calcule la probabilidad de obtener como resultados el número 5 y sello

Sean **c**, **s** los valores **cara** y **sello** de la moneda, entonces el espacio muestral **S** es:

S = {(1,**c**),(2,**c**),(3,**c**),(4,**c**),(5,**c**),(6,**c**),(1,**s**),(2,**s**),(3,**s**),(4,**s**),(5,**s**),(6,**s**)}

Sea el evento de interés,

A: obtener como resultados el **número 5** y **sello**
A = {(5, **s**)}

El evento **A** contiene un punto muestral. Entonces la probabilidad del evento **A** es 1 entre 12:

P(A) = 1/12 ≅ 0.0833

Suponga ahora que luego de lanzar el dado y la moneda, nos informan que el número del dado fue **impar**. ¿Cual es la probabilidad del evento **A** dado el evento indicado?

Sea **B** este evento conocido: **B** = {(1,c),(3,c),(5,c),(1,s),(3,s),(5,s)}

Entonces, la probabilidad del evento **A** dado el evento **B,** es 1 entre 6:

P(A) dado B = **1/6** ≅ **0.1667**

Definición: Probabilidad Condicional

Sean **A, B** eventos de **S**
La **Probabilidad Condicional** del evento **A** dado el evento **B** se escribe **P(A|B)** y es:

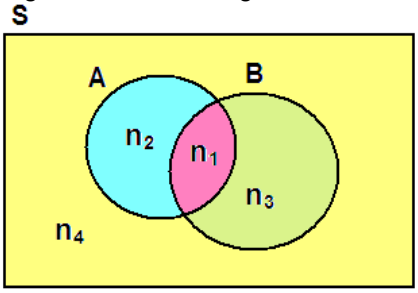
$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \quad P(B) \neq 0$$

Para justificar esta importante fórmula, suponga que **S** contiene solo dos eventos, **A** y **B**.
En la siguiente tabla se ha escrito simbólicamente el número de elementos de cada evento, siendo **N** el total de elementos del espacio muestral:

	B	**B^c**	
A	n_1	n_2	
A^c	n_3	n_4	
			N

Entonces, $P(A \mid B) = \frac{n_1}{n_1 + n_3} = \frac{\frac{n_1}{N}}{\frac{n_1 + n_3}{N}} = \frac{P(A \cap B)}{P(B)}$, resultado igual a la fórmula anterior

Interpretación gráfica con un Diagrama de Venn



P(A | B) es una función de probabilidad y cumple los axiomas anteriormente escritos.

Ejemplo.- Use la fórmula de la Probabilidad Condicional para resolver el ejemplo anterior,

S = {(1,**c**),(2,**c**),(3,**c**),(4,**c**),(5,**c**),(6,**c**),(1,**s**),(2,**s**),(3,**s**),(4,**s**),(5,**s**),(6,**s**)}
B = {(1,c),(3,c),(5,c),(1,s),(3,s),(5,s)}

$\mathbf{A \cap B} = \{(5, s)\} \Rightarrow \mathbf{P(A \mid B)} = \dfrac{\mathbf{P(A \cap B)}}{\mathbf{P(B)}} = \dfrac{\mathbf{1/12}}{\mathbf{6/12}} = \mathbf{1/6}$

Ejemplo. En una empresa hay 200 empleados, de los cuales 150 son graduados. 60 empleados realizan trabajo administrativo. De estos últimos, 40 son graduados. Si se toma al azar un empleado, encuentre la probabilidad que,

a) Sea graduado y no realiza trabajo administrativo.
b) Sea graduado dado que no realiza trabajo administrativo.
c) No sea graduado dado que realiza trabajo administrativo

Solución:

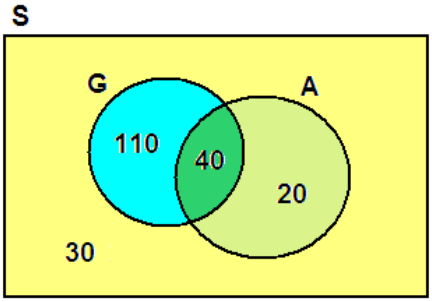
Sean estos eventos:

G: el empleado es graduado
A: el empleado realiza trabajo administrativo

Para facilitar el cálculo completamos el cuadro con la cantidad de elementos de cada evento. Los datos faltantes se los ha escrito con color negro:

	A	**A^c**	
G	**40**	**110**	**150**
G^c	**20**	**30**	**50**
	60	**140**	**200**

Representación gráfica con un Diagrama de Venn



Respuestas

a) $\mathbf{P(G \cap A^c)}$ = 110/200 = 0.55
b) $\mathbf{P(G|A^c)} = \mathbf{P(G \cap A^c)/P(A^c)}$ = (110/200) / (140/200) = 110/140 = 0.7857
c) $\mathbf{P(G^c|A)} = \mathbf{P(G^c \cap A)/P(A)}$ = (20/200) / (60/200) = 20/60 = 0.3333

Ejemplo. Las enfermedades **A** y **B** son comunes entre las personas de una región. Suponga conocido que 10% de la población contraerá la enfermedad **A,** 5% la enfermedad **B,** y 2% ambas enfermedades.

Encuentre la probabilidad que cualquier persona

a) Contraiga al menos una enfermedad
b) Contraiga la enfermedad **A** pero no **B**
c) Contraiga la enfermedad **A** dado que ya contrajo **B**
d) Contraiga la enfermedad **B** dado que no contrajo **A**
e) Contraiga ambas enfermedades dado que ya contrajo al menos una.

Solución:

Para facilitar el cálculo completamos el cuadro de probabilidades, siendo **A** y **B** los eventos que corresponden a contraer las enfermedades **A** y **B**, respectivamente

	B	**B^c**	
A	**0.02**	**0.08**	**0.10**
A^c	**0.03**	**0.87**	**0.90**
	0.05	**0.95**	**1**

Ahora se puede expresar cada pregunta en forma simbólica y obtener cada respuesta directamente del cuadro:

Respuestas

a) $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.1 + 0.05 - 0.02 = 0.13 = 13\%$
b) $P(A \cap B^c) = 0.08 = 8\%$
c) $P(A \mid B) = P(A \cap B) \,/\, P(B) = 0.02 / 0.05 = 0.4 = 40\%$
d) $P(B \mid A^c) = P(B \cap A^c) \,/\, P(A^c) = 0.03 / 0.9 = 0.3 = 30\%$
e) $P[(A \cap B) \mid (A \cup B)] = P[(A \cap B) \cap (A \cup B) \mid P(A \cup B)] = P(A \cap B) \,/\, P(A \cup B) = 0.02/0.13 = 0.1538$

3.9.1 EJERCICIOS

1) En un club de amigos, 10 practican tenis, 7 practican fútbol, 4 practican ambos deportes y los restantes 5 no practican algún deporte. Si se elige una de estas personas al azar, calcule la probabilidad que,

a) Al menos practique un deporte
b) No practique tenis
c) Practique tenis y no practique fútbol
d) Practique tenis dado que no practica fútbol

2) Sean los eventos A, B tales que $P(A)=0.4$, $P(B)=0.3$, $P(A \cap B)=0.1$, encuentre

a) $P(A|B)$
b) $P(B|A)$
c) $P(A|A \cup B)$
d) $P(A|A \cap B)$
e) $P(A \cap B|A \cup B)$

3) En una granja se tiene que la probabilidad que un animal tenga la gripe aviar es 0.3. La probabilidad que la reacción a una prueba sea negativa para un animal sano es 0.9, y que sea positiva para un animal enfermo es 0.8

a) Calcule la probabilidad que para un animal elegido al azar, el examen sea positivo
b) Calcule la probabilidad que el animal elegido al azar esté enfermo, dado que el examen fue positivo

3.10 EVENTOS INDEPENDIENTES

Sean **A** y **B** eventos cualesquiera de un espacio muestral **S**. Se dice que **A** y **B** son independientes si **P(A|B) = P(A)** y **P(B|A) = P(B)**, es decir que el evento **A** no depende del evento **B** y el evento **B** no depende del evento **A**

Lo anterior es equivalente a la siguiente definición:

Definición: Eventos Independientes

> **A y B** son eventos independientes si **$P(A \cap B) = P(A)\,P(B)$**

Demostración:
De la definición de probabilidad condicional,

$$P(A|B) = P(A \cap B)/P(B), \quad P(B) \neq 0$$

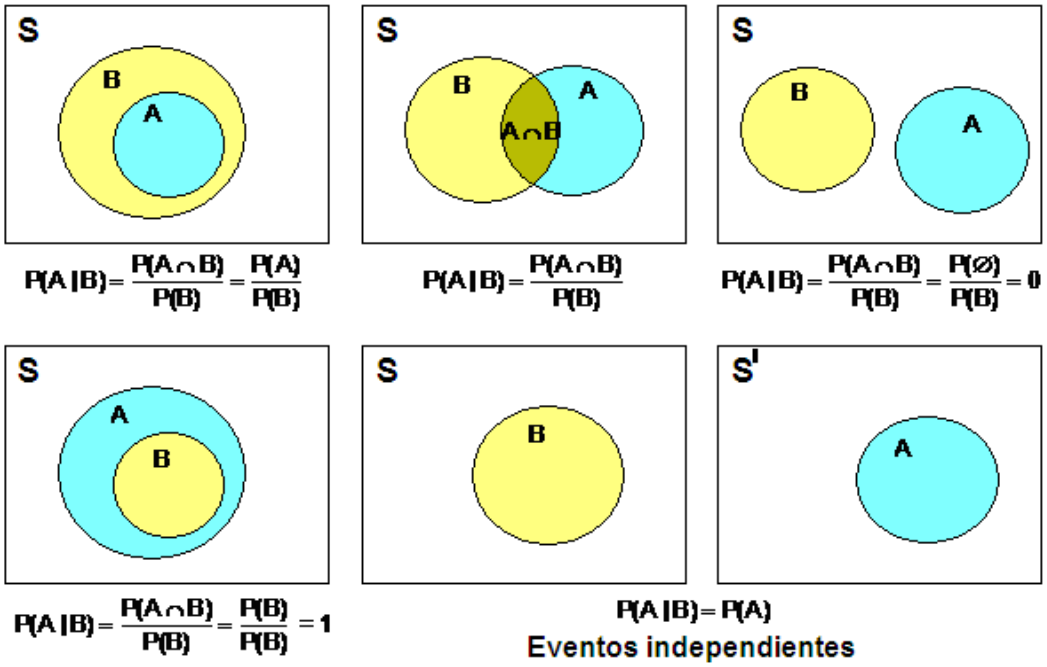
Si **A** y **B** son independientes: **$P(A|B) = P(A)$**.

Si se sustituye en la fórmula de probabilidad condicional:

$$P(A) = P(A \cap B)/P(B)$$

Se obtiene el la fórmula en la definición

INTERPRETACIÓN GRÁFICA



Ejemplo. Calcule la probabilidad que el último dígito del número de una placa de carro elegida al azar sea **par** y el penúltimo dígito sea **impar**

Sean los eventos

A: El último digito es **par**

B: El penúltimo dígito es **impar**

Cada evento no está relacionado con el otro evento, entonces son independientes.

Por lo tanto,

$$P(A \cap B) = P(A)\,P(B) = 0.5 \times 0.5 = 0.25$$

Ejemplo. En una caja hay 10 baterías de las cuales 4 están en buen estado. Se repite dos veces el siguiente ensayo: **extraer una batería al azar, revisar su estado y devolverla a la caja**.

a) Encuentre la probabilidad que en ambos intentos se obtenga una batería en buen estado.

Sean los eventos

A: La primera batería que se toma de la caja está en buen estado
B: La segunda batería que se toma de la caja está en buen estado

Este tipo de experimento se denomina: **Muestreo con Reemplazo**.

La primera batería se toma de la caja y se la devuelve, entonces el evento **B** no es afectado por el resultado que se obtuvo en el evento **A**, por lo tanto **son eventos independientes**.

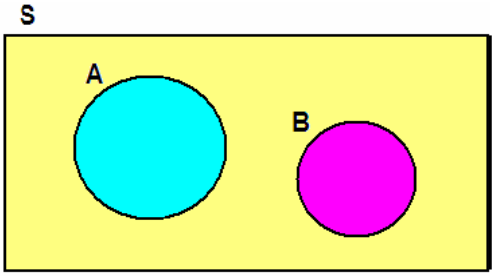
$$P(A \cap B) = P(A)\,P(B) = 0.4 \times 0.4 = 0.16$$

b) Calcule la probabilidad que en los dos intentos se obtenga al menos una batería en buen estado

Con la conocida **Fórmula Aditiva de Probabilidad**,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.4 + 0.4 - 0.16 = 0.64$$

Pregunta. Si **A, B** son eventos no nulos, mutuamente excluyentes, de un espacio muestral **S**. ¿Son **A** y **B** independientes?



Respuesta:

Nuestra intuición nos puede hacer pensar que **A** y **B** son eventos independientes, sin embargo no es verdad, como se demuestra:

Si **A, B** son eventos no nulos: $P(A) > 0, P(B) > 0 \Rightarrow P(A)\,P(B) > 0$
Pero si **A** y **B** son excluyentes: $A \cap B = \varnothing \Rightarrow P(A \cap B) = 0$

Por lo tanto: $P(A \cap B) \neq P(A)\,P(B) \Rightarrow$ **A y B no son independientes**

Pregunta. Si **A, B** son eventos no nulos e independientes ¿son **A, B** mutuamente excluyentes?

Si **A, B** son eventos independientes y no nulos: $P(A \cap B) = P(A)\,P(B) > 0$
Pero $P(A \cap B) > 0 \Rightarrow A \cap B \neq \varnothing$

Por lo tanto **A, B** no pueden ser mutuamente excluyentes

NOTA: Ambos razonamientos son lógicamente equivalentes como se muestra a continuación:

Sean las proposiciones:
p: **A** y **B** son eventos no nulos mutuamente excluyentes,
q: **A** y **B** son eventos no nulos e independientes
Entonces, por la conocida equivalencia lógica:
$p \Rightarrow \neg q \equiv q \Rightarrow \neg p$

La definición de independencia entre dos eventos puede extenderse a más eventos

Definición: Eventos Independientes con tres Eventos

Si A, B, C son eventos mutuamente independientes, entonces
$P(A \cap B \cap C) = P(A)\,P(B)\,P(C)$.

3.11 REGLA MULTIPLICATIVA DE LA PROBABILIDAD

Sean **A**, **B** eventos no nulos cualquiera de **S**, entonces

Definición: Regla Multiplicativa de la Probabilidad

$$P(A \cap B) = P(A)\,P(B|A)$$

Esta fórmula se la obtiene directamente despejando **$P(A \cap B)$** de la fórmula de Probabilidad Condicional

Ejemplo. En una caja hay 10 baterías de las cuales 4 están en buen estado. Se extraen al azar dos baterías sin devolverlas a la caja. Calcule la probabilidad que,

a) Ambas baterías estén en buen estado
b) Solamente una batería esté en buen estado
c) Al menos una batería esté en buen estado
d) Ninguna batería esté en buen estado

Solución:

Sean los eventos

A: La primera batería que se toma de la caja está en buen estado
B: La segunda batería que se toma de la caja está en buen estado

Este tipo de experimento se denomina: **Muestreo sin Reemplazo**.

Al tomar la primera batería de la caja y no devolverla, el evento **B** es afectado por el resultado que se obtuvo en el evento **A**, por lo tanto **no son eventos independientes**.

a) La probabilidad que ambas baterías estén en buen estado es **$P(A \cap B)$**, pero los eventos **A** y **B** no son independientes. Entonces con la fórmula anterior

$$P(A \cap B) = P(A)\,P(B|A) = \left(\frac{4}{10}\right)\left(\frac{3}{9}\right) = 2/15 = 0.1333$$

La probabilidad de éxito del evento **A** es 4/10. Para el evento **B** la probabilidad de éxito es 3/9, dado que **A** es favorable (quedan 3 baterías en buen estado del total de 9 baterías)

b) La probabilidad que una batería esté en buen estado y la otra en mal estado:

$$P(A \cap B^c) + P(A^c \cap B) = P(A)P(B^c|A) + P(A^c)P(B|A^c)$$
$$= (4/10)(6/9) + (6/10)(4/9) = 12/15 = 0.5333$$

Los eventos que solamente la primera batería esté en buen estado y que solamente la segunda batería esté en buen estado son excluyentes, por lo tanto sus probabilidades se suman:

c) La probabilidad que al menos una esté en buen estado. Con los resultados de en **a)** y **b)**:

$$P(A\cup B) = P(A\cap B)\cup P(A\cap B^c)\cup P(A^c\cap B) = 2/15 + 8/15 = 2/3 = 0.6666$$

Los eventos que ambas estén en buen estado o que solamente una esté en buen estado son mutuamente excluyentes, por lo tanto sus probabilidades se suman.

d) La probabilidad que ninguna esté en buen estado

$$P((A\cup B)^c) = 1 - P(A\cup B) = 1 - 2/3 = 1/3 = 0.3333$$

Es el complemento del evento que al menos una esté en buen estado.

El ejemplo anterior también puede resolverse con las conocidas fórmulas de conteo:

a) Probabilidad que ambas baterías estén en buen estado

Sea **A:** evento que ambas baterías están en buen estado
N(A): cantidad de formas de sacar 2 baterías en buen estado de las 4 existentes:
N(S): cantidad de formas de sacar 2 baterías del total de 10 baterías

$$P(A) = N(A) / N(S) = {}_4C_2 / {}_{10}C_2 = 2/15$$

b) Probabilidad que solamente una batería esté en buen estado

Sea **A:** Evento que una batería está en buen estado y la otra esté en mal estado.
Este evento incluye las formas de sacar una batería en buen estado de las 4 existente: ${}_4C_1$, y una en mal estado de las 6 existentes: ${}_6C_1$

$$P(A) = {}_4C_1\ {}_6C_1 / {}_{10}C_2 = 8/15$$

c) Probabilidad que al menos una batería esté en buen estado

Sean los eventos
A: Ambas baterías están en buen estado
B: Solamente una batería está en buen estado
A y **B** son eventos excluyentes, por lo tanto

$$P(A\cup B) = P(A) + P(B) = 2/15 + 8/15 = 10/15 = 2/3$$

La Regla Multiplicativa de Probabilidad puede extenderse a más eventos.

Definición: Regla Multiplicativa de Probabilidad para tres Eventos

Sean **A, B, C** eventos cualesquiera de **S**, entonces
$$P(A\cap B\cap C) = P(A)\ P(B|A)\ P(C|A\cap B)$$

.

Ejemplo. En el ejemplo anterior de las 10 baterías con 4 en buen estado, encuentre la probabilidad que al extraer tres sin devolverlas, las tres estén en buen estado

Solución:

Sean **A, B, C** eventos correspondientes a que la primera, segunda y tercera batería estén en buen estado,

Con la Regla Multiplicativa de Probabilidad:

$$P(A\cap B\cap C) = P(A)\ P(B|A)\ P(C|A\cap B) = (4/10)\ (3/9)\ (2/8) = 1/30 = 0.0333 = 3.33\%$$

Este ejemplo también se puede resolver usando fórmulas de conteo:

Sean **A**: Evento que las tres baterías están en buen estado
N(A): cantidad de formas de tomar 3 baterías en buen estado de las 4 existentes
N(S): cantidad de formas de tomar 3 baterías del total de 10

$P(A) = N(A) / N(S) = {}_4C_3 / {}_{10}C_3 = 1/30$

3.11.1 EJERCICIOS

1) Dos jugadores de fútbol realizan un disparo cada uno. Se conoce que la probabilidad de éxito del primero es 0.7 mientras que la probabilidad de éxito del segundo jugador es 0.6. Calcule la probabilidad que

a) Ambos jugadores tengan éxito.
b) Ninguno tenga éxito.
c) Al menos uno tenga éxito

2) Dos alarmas contra incendio funcionan independientemente. La probabilidad de éxito de detección de la primera es 0.95, mientras que para la segunda es 0.9. Calcule la probabilidad que:

a) Al menos una alarma tenga éxito.
c) Solamente una alarma tenga éxito.

3) Sean A, B eventos independientes. Demuestre que los eventos A^c, B^c también son eventos independientes.

3.12 PROBABILIDAD TOTAL

Existen situaciones en las cuales varios eventos intervienen en la realización de algún otro evento del mismo espacio muestral.

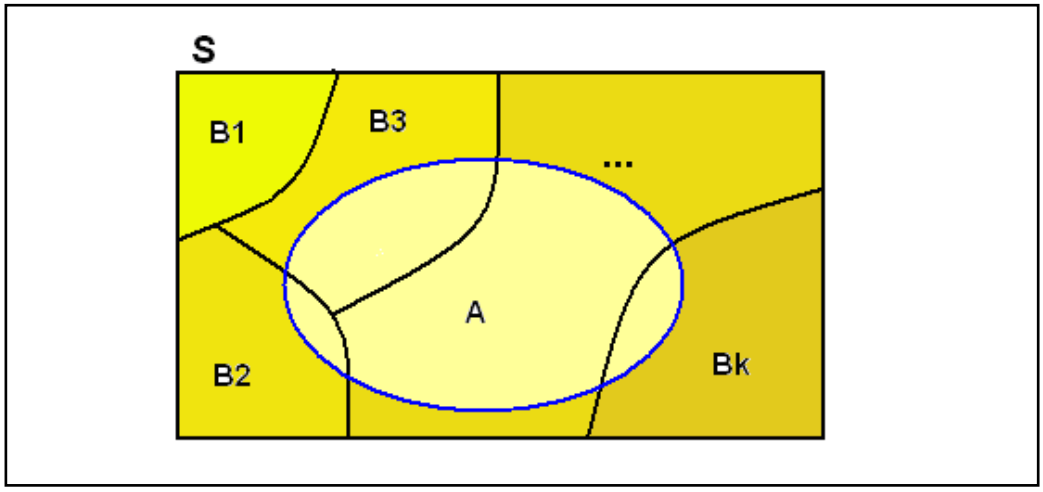
Sean $\mathbf{B_1, B_2, \ldots, B_K}$ eventos mutuamente excluyentes en **S** y que constituyen una **partición de S**, es decir, cumplen las siguientes propiedades:

a) $\forall i,j\ (B_i \cap B_j = \varnothing, i \neq j)$ (Los eventos son mutuamente excluyentes)

b) $B_1 \cup B_2 \cup \ldots \cup B_K = S$ (La unión de todos estos eventos es **S**)

Sea **A** un evento cualquiera de **S** La realización de **A** depende de los eventos $\mathbf{B_1, B_2, \ldots, B_K}$

El siguiente gráfico permite visualizar esta relación entre los eventos descritos:



La siguiente fórmula permite calcular la probabilidad del evento **A** conocidos los valores de probabilidad de los eventos $\mathbf{B_1, B_2, \ldots, B_K}$

Definición:

Fórmula de la Probabilidad Total

$$P(A) = P(B_1)\,P(A|B_1)+P(B_2)\,P(A|B_2)+\ldots+P(B_k)\,P(A|B_k) = \sum_{i=1}^{k} P(B_i)P(A \mid B_i)$$

Demostración

$A = (A \cap B_1) \cup (A \cap B_2) \cup \ldots \cup (A \cap B_K)$ **A** es la unión de eventos excluyentes

$P(A) = P(A \cap B_1) + P(A \cap B_2) + \ldots + P(A \cap B_K)$ Por el Axioma 3 de probabilidad

Entonces, con la definición de Probabilidad Condicional

$$P(A) = P(B_1)\,P(A|B_1)+P(B_2)\,P(A|B_2)+\ldots+P(B_k)\,P(A|B_k)$$

Ejemplo. Una empresa tiene tres personas para atender a sus clientes: María, Carmen y Beatriz. Se dispone de un registro histórico del porcentaje de quejas de los clientes atendidos por estas tres personas: 1%, 3%, 2% respectivamente. Cierto día acudieron 50 clientes a la empresa de los cuales 15 fueron atendidos por María, 10 por Carmen y 25 por Beatriz.

Calcule la probabilidad que un cliente elegido al azar de entre los que fueron atendidos ese día se queje por la atención recibida.

Solución.

Los datos disponibles son

Persona	Clientes atendidos	Probabilidad de queja
María	**15**	**1%**
Carmen	**10**	**3%**
Beatriz	**25**	**2%**

Si se definen los siguientes eventos:

A: El cliente elegido al azar presenta una queja
B_1: El cliente fue atendido por María
B_2: El cliente fue atendido por Carmen
B_3: El cliente fue atendido por Beatriz

B_1, B_2, y **B_3** son eventos que **conforman una partición**, y contribuyen a la realización de otro evento, **A**. Por lo tanto es un problema de Probabilidad Total:

$$P(A) = P(B_1)\,P(A|B_1) + P(B_2)\,P(A|B_2) + P(B_3)\,P(A|B_3)$$
$$= (15/50)0.01 + (10/50)0.03 + (25/50)0.02 = 0.019 = 1.9\%$$

Ejemplo. Una fábrica tiene tres máquinas **M_1, M_2, M_3** para la producción de sus artículos. El siguiente cuadro describe el porcentaje de producción diaria de cada una y la frecuencia de artículos defectuosos que producen cada una.

Máquina	Producción	Artículos defectuosos
M_1	**50%**	**4%**
M_2	**30%**	**3%**
M_3	**20%**	**2%**

Determine la probabilidad que un artículo elegido al azar de la producción total de un día, sea defectuoso.

Solución

Sea **A**: Evento que el artículo elegido al azar sea defectuoso

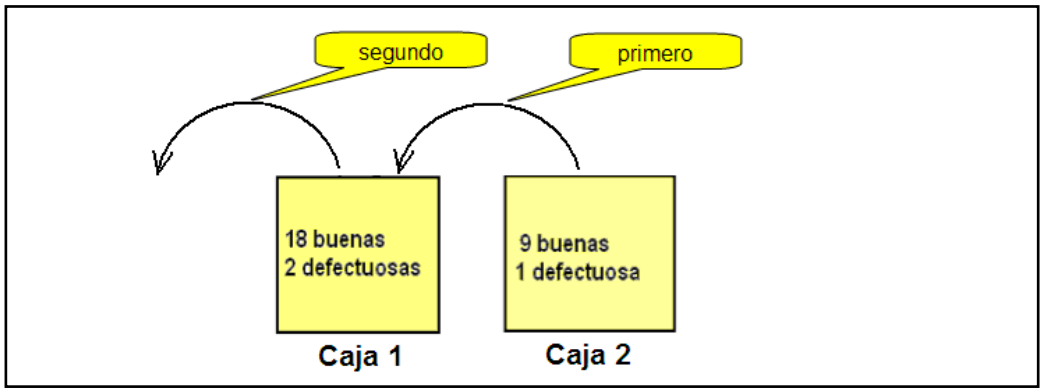
El evento **A** depende de **B_1, B_2, B_3** que representan los eventos de que un artículo sea producido por las máquinas: **M_1, M_2, M_3** respectivamente. Estos eventos forman una partición por lo que con la **fórmula de la Probabilidad Total**

$$P(A) = P(B_1)\,P(A|B_1) + P(B_2)\,P(A|B_2) + P(B_3)\,P(A|B_3)$$
$$= (0.5)(0.04) + (0.3)(0.02) + (0.2)(0.03) = 0.032 = 3.2\%$$

Ejemplo. En la caja **1** hay 20 baterías de las cuales 18 están en buen estado. En la caja **2** hay 10 baterías de las cuales 9 están en buen estado. Se realiza un experimento que consiste en las siguientes dos acciones:

Primero se toma al azar de la caja **2** una batería y sin examinarla se la coloca en la caja **1**.
Segundo, se toma al azar una batería de la caja **1** y se la examina.
Encuentre la probabilidad que esta última batería esté en buen estado.

Respuesta
El siguiente gráfico describe el experimento:



Sean los eventos

B: La batería tomada de la caja 2 y colocada en la caja 1 está en buen estado
B^c: La batería tomada de la caja 2 y colocada en la caja 1 **no** está en buen estado
A: La batería tomada de la caja 1 está en buen estado

El evento **A** depende de los eventos **B** y **B^c**, los cuales son excluyentes y forman una partición. De estos eventos depende el evento **A.** Entonces con la fórmula de la Probabilidad Total:

$$P(A) = P(B)\,P(A|B) + P(B^c)\,P(A|B^c) = (9/10)(19/21) + (1/10)(18/21) = 0.9$$

3.13 TEOREMA DE BAYES

Sean $B_1, B_2, \ldots, B_K$ eventos no nulos mutuamente excluyentes de **S** y que constituyen una partición de **S**, y sea **A** un evento no nulo cualquiera de **S**

La siguiente fórmula se denomina **Fórmula de Bayes** y permite calcular la probabilidad correspondiente a cada uno de los eventos que contribuyen a la realización de otro evento, dado que se conoce la probabilidad de este evento.

Definición: Fórmula de Bayes

$$P(B_i|A) = \frac{P(B_i)\ P(A\,|\,B_i)}{P(A)} = \frac{P(B_i)\ P(A\,|\,B_i)}{\sum_{I=1}^{K} P(B_i)\ P(A\,|\,B_i)},\ i=1, 2, \ldots, k$$

Demostración. Se obtiene directamente de la definición de Probabilidad Condicional y la fórmula de Probabilidad Total:

$$P(B_i|A) = \frac{P(B_i \cap A)}{P(A)} = \frac{P(B_i)\ P(A\,|\,B_i)}{P(A)},\ i = 1, 2, \ldots, k$$

Ejemplo. En el problema de la fábrica del ejemplo anterior, suponga que el artículo elegido al azar fue defectuoso. Calcule la probabilidad que haya sido producido por la máquina $\mathbf{M_1}$:

Solución:

$$P(B_1|A) = \frac{P(B_1)\,P(A\,|\,B_1)}{P(A)} = \frac{(0.50)(0.04)}{0.032} = 0.625 = 62.5\%$$

Ejemplo. Sean **A, B** eventos de algún espacio muestral **S**. Se conoce que

P(B) = 0.4
P(A|B) = 0.3
P(A|B^c) = 0.8

Encuentre

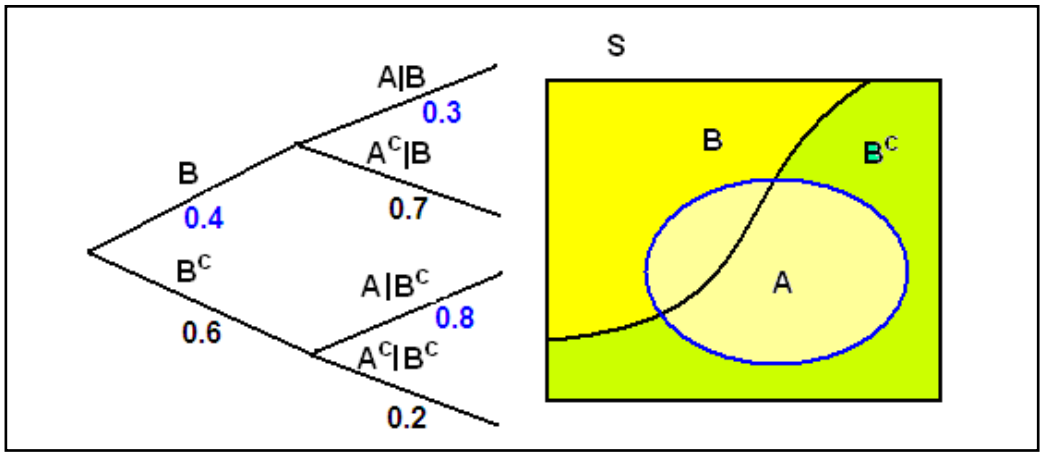
a) **P(A)** b) **P(B|A)** c) **P(B|A^c)**

Solución:

Para facilitar la interpretación de este problema colocamos los datos en un diagrama de árbol y con un Diagrama de Venn visualizamos los eventos.

Los datos los escribimos en color **azul.** Los valores faltantes los completamos en **negro.**

Los eventos **B** y **B^c** constituyen una partición y contribuyen a la realización del evento **A.**



Con los valores indicados en el diagrama y las fórmulas de Probabilidad Total y el Teorema de Bayes se obtienen las respuestas:

a) $P(A) = P(B)\,P(A|B) + P(B^c)\,P(A|B^c) = (0.4)(0.3) + ((0.6)(0.8) = 0.6$

b) $P(B|A) = P(B \cap A)/P(A) = P(B)\,P(A|B) / P(A) = (0.4)\,(0.3) / 0.6 = 0.2$

c) $P(B|A^c) = P(B \cap A^c)/P(A^c) = P(B)\,P(A^c|B) / P(A^c) = (0.4)\,(0.7) / 0.4 = 0.7$

3.14 EJERCICIOS

1) La Comisión de Tránsito del Guayas ha implantado un sistema de control de velocidad mediante un radar colocado en cuatro puntos de la ciudad: **X_1, X_2, X_3, X_4**. Cada día, estos aparatos están activos en los sitios indicados, 16 horas, 10 horas, 12 horas y 15 horas respectivamente en horarios al azar. Una persona maneja a su trabajo diariamente y lo hace con exceso de velocidad y la probabilidad de que pase por alguno de estos sitios es respectivamente **0.3, 0.1, 0.4** y **0.2**

a) Calcule la probabilidad que en algún día reciba una multa por exceso de velocidad.
b) Cierto día, la persona recibió una multa por exceso de velocidad. Determine el sitio en que hay la mayor probabilidad de haber sido multado.

2) Para concursar por una beca de estudio en el exterior se han presentado a rendir un examen **10** estudiantes de la universidad **X_1**, **20** de la universidad **X_2** y **5** de la universidad **X_3**. De experiencias anteriores, se conoce que las probabilidades de éxito en el examen son respectivamente: **0.9, 0.6, 0.7**

a) Calcule la probabilidad que un estudiante elegido al azar apruebe el examen
b) Calcule la probabilidad condicional de que un estudiante elegido al azar y que haya aprobado el examen, sea de la universidad **X_1**.

MATLAB

Ejemplo

Eventos	P(B)	P(A|B)
B_1	50%	4%
B_2	30%	3%
B_3	20%	2%

Probabilidad total

```
>> pb = [.5 .3 .2];
>> pab = [.04 .03 .02];
>> pa = sum(pb.*pab)
   pa =    0.0330
```

$P(B_i)$
$P(A|B_i)$
$P(A) = P(B_1)P(A|B_1) + ...$

Fórmula de Bayes

```
>> pba = pb.*pab/pa
   pba =   0.6061    0.2727    0.1212
```

$P(B_i|A)$

4 VARIABLES ALEATORIAS DISCRETAS

En el material estudiado anteriormente aprendimos a calcular la probabilidad de **eventos** de un espacio muestral **S.** En esta unidad estudiaremos **reglas** para establecer correspondencias de los **elementos** de **S** con los números reales, para luego asignarles un valor de probabilidad.

Ejemplo.
En un experimento se lanzan tres monedas y se observa el resultado (c**:** cara o **s:** sello).
El conjunto de posibles resultados (espacio muestral) para este experimento, es el siguiente:

S = {(c, c, c),(c, c, **s**),(c, **s**, c),(**s**, c, c),(c, **s**, **s**),(**s**, c, **s**),(**s**, **s**, c),(**s**, **s**, **s**)}

Describa con una variable, el **número de sellos que se obtienen**.

Los posibles resultados se los puede representar con una variable. Si **X** es ésta variable, entonces se dice que **X** es una variable aleatoria:

X: Variable aleatoria (**número de sellos que se obtienen**)

Al realizar el experimento, se obtendrá cualquier elemento del espacio muestral **S**.
Por lo tanto, la variable aleatoria **X** puede tomar alguno de los números: **x = 0, 1, 2, 3**.

Las **Variables Aleatorias** establecen correspondencia del espacio muestral **S** al conjunto de los números reales. Esta correspondencia es funcional y se la puede definir formalmente.

Definición: Variable aleatoria

Sean **X:** Variable aleatoria
S: Espacio muestral
e: Cualquier elemento de **S**
x: Valor que puede tomar **X**
$\mathfrak{R}$**:** Conjunto de los números reales

Entonces

X: $S \rightarrow \mathfrak{R}$ Es la correspondencia que establece la variable aleatoria **X**

$e \rightarrow x$, **dom X = S,** **rg X** $\subset \mathfrak{R}$.

Ejemplo: Tabule la correspondencia que establece la variable aleatoria **X** del ejemplo anterior:

S = {(c, c, c),(c, c, **s**),(c, **s**, c),(**s**, c, c),(c, **s**, **s**),(**s**, c, **s**),(**s**, **s**, c),(**s**, **s**, **s**)}
X: Variable aleatoria (número de sellos que se obtienen)
x = **0, 1, 2, 3**

e (elemento de **S**)	**x** (valor de **X**)
(c, c, c)	**0**
(c, c, **s**)	**1**
(c, **s**, c)	**1**
(**s**, c, c)	**1**
(c, **s**, **s**)	**2**
(**s**, c, **s**)	**2**
(**s**, **s**, c)	**2**
(**s**, **s**, **s**)	**3**

dom **X = S**, **rg X = {0, 1, 2, 3}**

Las variables aleatorias pueden representarse con las letras mayúsculas **X, Y, ...**

Para un mismo espacio muestral **S** pueden definirse muchas variables aleatorias.
Para el ejemplo de las 3 monedas, algunas otras variables aleatorias sobre **S** pueden ser:

Y: Diferencia entre el número de caras y sellos
Z: El número de caras al cubo, mas el doble del número de sellos, etc.

Para cada variable aleatoria el rango es un subconjunto de los reales. Según el tipo de correspondencia establecida, las variables aleatorias pueden ser **discretas** o **continuas**.

En el ejemplo de las monedas, **X** es una **variable aleatoria discreta** pues su **rango** es un subconjunto de los enteros. Además es **finita**.

Ejemplo. En un experimento se lanza repetidamente una moneda. Determine el rango y tipo de la variable aleatoria discreta siguiente:

X: Cantidad de lanzamientos realizados hasta que sale un sello

S={(s), (c, s), (c, c, s), (c, c, c, s), ...}, resultados posibles

rg **X** = {1, 2, 3, 4, ...}

X es una **variable aleatoria discreta infinita**

4.1 DISTRIBUCIÓN DE PROBABILIDAD DE UNA VARIABLE ALEATORIA DISCRETA

Cada valor de una variable aleatoria discreta puede asociarse a un valor de probabilidad

Definición: Probabilidad de una Variable Aleatoria Discreta

Sea **X:** Variable aleatoria discreta
Entonces, **P(X=x)** representa la probabilidad que la variable **X** tome el valor **x**

La correspondencia que define **P(X=x)** es **una función** y se denomina **Distribución de Probabilidad** de la variable aleatoria **X**. Esta correspondencia puede definirse formalmente y ser designarda con la notación **f:**:

Definición: Distribución de Probabilidad de una Variable Aleatoria Discreta X

Sean **X:** Variable aleatoria discreta
f(x) = P(X=x): Probabilidad que **X** tome el valor **x**

Entonces, la correspondencia

$$f: X \to \Re,$$
$$x \to f(x) = P(X=x), \quad \text{dom } f = X, \quad \text{rg } f \subset [0, 1]$$

Es la Distribución de Probabilidad de la Variable Aleatoria Discreta **X**

f es una **función de probabilidad,** por lo tanto su rango está en el intervalo **[0, 1]**

Definición: Propiedades de la Distribución de Probabilidad de una Variable Aleatoria Discreta

Sean **X:** Variable aleatoria discreta
f(x): Distribución de Probabilidad de **X**

Propiedades de f(x)

1) $\forall x\ [\ f(x) \geq 0\ ]$ Los valores de probabilidad no pueden ser negativos

2) $\sum_{x} f(x) = 1$ La suma de todos los valores de probabilidad es **1**

La correspondencia que establece **f** puede describirse en **forma tabular** como en el ejemplo de las tres monedas. También puede describirse **gráficamente**, y en algunos casos mediante una **fórmula matemática** como se verá en los siguientes capítulos.

Ejemplo. En el experimento de lanzar tres monedas y observar el resultado de cada una: cara(c), o sello(**s**). Encuentre la Distribución de Probabilidad en forma tabular, de la variable aleatoria **X:** cantidad de sellos que se obtienen

Espacio muestral: **S** = {(c, c, c),(c, c, **s**),(c, **s**, c),(**s**, c, c),(c, **s**, **s**),(**s**, c, **s**),(**s**, **s**, c),(**s**, **s**, **s**)}

e (elemento de **S**)	**x** (valor de **X**)
(c, c, c)	**0**
(c, c, **s**)	**1**
(c, **s**, c)	**1**
(**s**, c, c)	**1**
(c, **s**, **s**)	**2**
(**s**, c, **s**)	**2**
(**s**, **s**, c)	**2**
(**s**, **s**, **s**)	**3**

Los valores de probabilidad para este ejemplo se pueden obtener del conteo de valores **x:**

El valor 0 ocurre 1 vez entre 8, el valor 1 ocurre 3 veces entre 8, etc

x	**f(x)=P(X=x)**
0	**1/8**
1	**3/8**
2	**3/8**
3	**1/8**

Ejemplo. En un lote de 5 artículos, 3 son defectuosos y 2 aceptables. Se toma una **muestra aleatoria sin reemplazo** de 2 artículos. Encuentre la Distribución de Probabilidad de la variable aleatoria: **cantidad de artículos defectuosos** que se obtienen en la muestra.

Respuesta

Sean: **a, b, c**: artículos defectuosos
d, e: artículos aceptables

Cantidad de formas diferentes de obtener la muestra de 2 artículos cualesquiera

N(**S**) = $_5\mathbf{C}_2$ =10

S = {(a, b), (a, c), (a, d), (a, e), (b, c), (b, d), (b, e), (c, d), (c, e), (d, e)}

Sea **X**: Variable aleatoria discreta (cantidad de artículos defectuosos)

x = 0, 1, 2

La distribución de probabilidad de **X** en forma tabular se obtiene mediante un conteo directo

x	**f(x)=P(X=x)**
0	**1/10**
1	**6/10**
2	**3/10**

Ejemplo. Sea **X** una variable aleatoria discreta cuya distribución de probabilidad está dada por

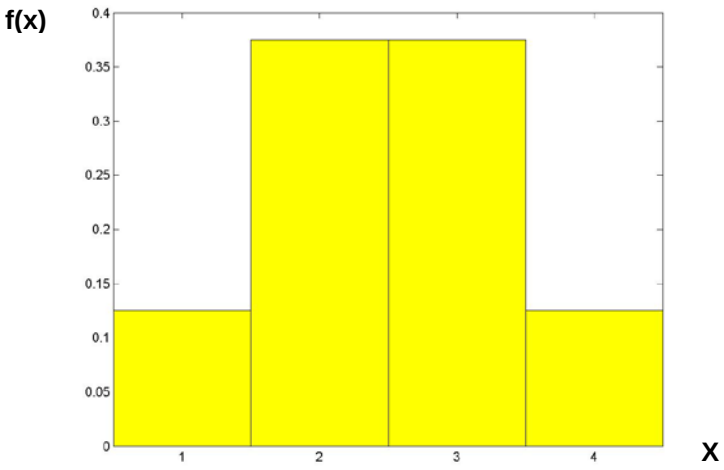
$$\mathbf{f(x) = P(X=x)} = \begin{cases} \mathbf{kx^2, x = 0,1,2,3} \\ \mathbf{0,} \quad \mathbf{otro} \quad x \end{cases}$$ Encuentre **P(X=2)**

Respuesta. Por la propiedad **2)** $\sum_x \mathbf{f(x) = 1}$

$$\sum_{x=0}^{3} \mathbf{kx^2 = k(0)^2 + k(1)^2 + k(2)^2 + k(3)^2 = 1} \Rightarrow \mathbf{k = 1/14} \Rightarrow \mathbf{f(x) = P(X=x)} = \begin{cases} \mathbf{\frac{1}{14}x^2, x = 0,1,2,3} \\ \mathbf{0,} \quad \mathbf{otro} \quad \mathbf{x} \end{cases}$$

Por lo tanto, $\mathbf{P(X=2) = (1/14)(2)^2 = 2/7}$

Ejemplo. Grafique un histograma de la distribución de probabilidad para el ejemplo de las tres monedas



Ejemplo. Para ensamblar una máquina se usan dos componentes electrónicos. Suponga que la probabilidad que el primer componente cumpla las especificaciones es **0.95**, y para el segundo es **0.98**. Además, los componentes funcionan independientemente.
Encuentre la distribución de probabilidad del número de componentes que cumplen las especificaciones, **x = 0, 1, 2**

Sea **X:** Variable aleatoria discreta (número de componentes que cumplen las especificaciones)
x = 0, 1, 2

Sean los eventos:

A: el primer componente cumple las especificaciones
B: el segundo componente cumple las especificaciones
A^C: el primer componente no cumple las especificaciones
B^C: el segundo componente no cumple las especificaciones

Entonces

$P(X=0) = P(A^C)P(B^C) = (1 - 0.05)(1 - 0.98) = 0.001$ (Eventos independientes)
$P(X=1) = P(A \cap B^C) + P(B \cap A^C) = 0.95(1-0.98) + 0.98(1-0.95) = 0.068$
$P(X=2) = P(A)P(B) = (0.95)(0.98) = 0.931$ (Eventos independientes)

Por lo tanto, Distribución de Probabilidad de la Variable Aleatoria **X** es:

x	f(x) = P(X=x)
0	0.001
1	0.068
2	0.931

Estos resultados se fundamentan en la propiedad de que si **A, B** son eventos independientes, entonces también **A^C, B^C** son eventos independientes.

4.2 DISTRIBUCIÓN DE PROBABILIDAD ACUMULADA DE UNA VARIABLE ALEATORIA DISCRETA

También es importante conocer la probabilidad que la variable aleatoria tome algún valor menor o igual que un valor dado. Esta función se denomina **Distribución de Probabilidad Acumulada** y su dominio incluye a todos los números reales

Definición: Distribución de Probabilidad Acumulada de la variable aleatoria X

Sean **X:** Variable aleatoria discreta,
f: Distribución de Probabilidad de la variable aleatoria discreta **X**
F: Distribución de Probabilidad Acumulada de la variable aleatoria discreta **X**

Entonces

$$\mathbf{F(x) = P(X \le x) = \sum_{t \le x} f(t)}$$ es la Distribución de Probabilidad Acumulada de **X**

Correspondencia funcional de la distribución de probabilidad acumulada

$$\mathbf{F: \Re \to \Re, \quad dom\ F = \Re, \quad rg\ F \subset [0, 1]}$$

Ejemplo. Encuentre la distribución de probabilidad acumulada para el ejemplo de las tres monedas

Respuesta:

Sea **X:** variable aleatoria discreta (cantidad de sellos que se obtienen)
Su distribución de probabilidad es:

x	f(x)=P(X=x)
0	1/8
1	3/8
2	3/8
3	1/8

Entonces,

$$F(0) = P(X \le 0) = \sum_{t \le 0} \mathbf{f(t)} = f(0) = 1/8$$

$$F(1) = P(X \le 1) = \sum_{t \le 1} \mathbf{f(t)} = f(0) + f(1) = 1/8 + 3/8 = 1/2$$

$$F(2) = P(X \le 2) = \sum_{t \le 2} \mathbf{f(t)} = f(0) + f(1) + f(2) = 1/8 + 3/8 + 3/8 = 7/8$$

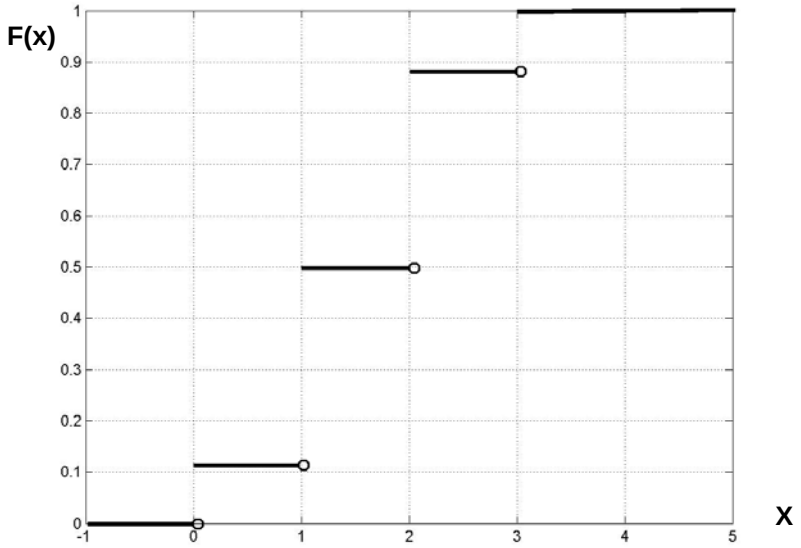
$$F(3) = P(X \le 3) = \sum_{t \le 3} \mathbf{f(t)} = f(0) + f(1) + f(2) + f(3) = 1$$

Distribución de Probabilidad Acumulada de la vaiable aleatoria **X**:

$$\mathbf{F(x) = \begin{cases} 0, & x < 0 \\ 1/8, & 0 \le x < 1 \\ 1/2, & 1 \le x < 2 \\ 7/8, & 2 \le x < 3 \\ 1, & x \ge 3 \end{cases}}$$

La Distribución Acumulada puede graficarse

Ejemplo. Grafique la Distribución Acumulada obtenida para el ejemplo anterior



Definición: Propiedades de la Distribución Acumulada para Variables Aleatorias Discretas

1)	**$0 \le F(x) \le 1$**	**F** es una función de probabilidad
2)	**$a \le b \Rightarrow F(a) \le F(b)$**	**F** es creciente
3)	**$P(X>a) = 1 - P(X \le a) = 1 - F(a)$**	Complemento de Probabilidad .

El dominio de **F** es el conjunto de los números reales, por lo tanto es válido evaluar **F(x)** para cualquier valor real **x**.

Ejemplo. Calcule algunos valores de **F(x)** para el ejemplo anterior.

Con la Distribución Acumulada **F(x)** previamente calculada:

$F(2.5) = P(X \le 2.5) = 7/8$
$F(-3.4) = 0$
$F(24.7) = 1$

4.2.1 EJERCICIOS

1) Sea **X** una variable aleatoria discreta y su función de distribución de probabilidad**:**

$$f(x) = \frac{2x+1}{25}, x = 0,\ 1,\ 2,\ 3,\ 4$$

a) Verifique que **f** satisface las propiedades de las distribuciones de probabilidad
b) Grafique **f** mediante un histograma
c) Calcule **P(X=3), P(2≤X<4)**

2) Para ensamblar una máquina se usan dos componentes mecánicos. Suponga que la probabilidad que el primer componente cumpla las especificaciones es **0.95**, y para el segundo es **0.98**. Además, los componentes funcionan independientemente.
Encuentre la función de distribución de probabilidad del número de componentes que cumplen las especificaciones, **X = 0, 1, 2**

3) Respecto al ejercicio **1)**

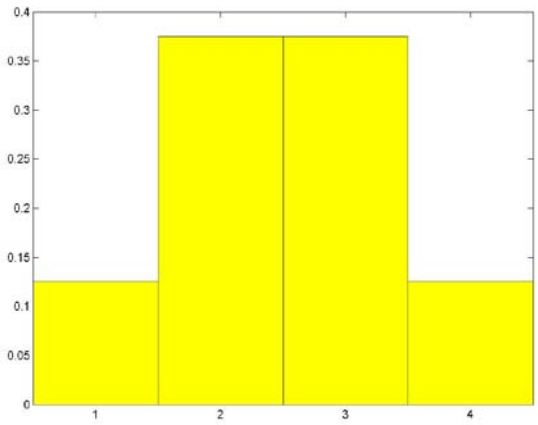
a) Encuentre y grafique la función de distribución acumulada **F**
c) Usando **F** calcule P(X<1.25), P(1.5<X≤3), P(X<2.5 ∨ X>3.2)

MATLAB

Probabilidad con variables aleatorias discretas

```
>> x = [0 1 2 3];
>> f = [1/8 3/8 3/8 1/8];
>> bar(f, 1, 'y'), grid on
```

Valores de una variable aleatoria **X**
Distribución de probabilidad **f(x)**
Histograma de probabilidad, color amarillo




```
>> F=cumsum(f)
   F =
       1/8    1/2     7/8     1
```

Probabilidad acumulada **F(x)**

4.3 VALOR ESPERADO DE UNA VARIABLE ALEATORIA DISCRETA

El Valor Esperado o Media es una medida estadística que describe la tendencia central de una variable aleatoria. Podemos pensar que representa el valor promedio que tomaría la variable aleatoria si el experimento se realizara un gran número de veces en condiciones similares.

Definición: Valor Esperado o Media de una Variable Aleatoria Discreta

Sean **X:** Variable aleatoria discreta
f(x): Distribución de probabilidad de **X**
μ o E(X): Media o Valor Esperado de la Variable Aleatoria **X**
Entonces:

$$\mu = E(X) = \sum_{x} xf(x)$$ es la Media o Valor Esperado de **X**

La definición representa la suma de los valores de **X** ponderados con su valor de probabilidad

Ejemplo. Calcule el valor esperado de la variable aleatoria **X** en el experimento de lanzar tres monedas, siendo **X:** Número de sellos que se obtienen

Respuesta: De un ejemplo anterior, se tiene la Distribución de Probabilidad de **X**

x	f(x)=P(X=x)
0	1/8
1	3/8
2	3/8
3	1/8

Entonces, el valor esperado de **X** es:

$$\mu = E(X) = \sum_{x=0}^{3} xf(x) = 0(1/8) + 1(3/8) + 2(3/8) + 3(1/8) = 1.5$$

Significa que si se realizaran un gran número de ensayos, en promedio se obtendrán **1.5** sellos.

En el ejemplo anterior, el valor esperado se ubica en el centro de la distribución de los valores de **X**. Esto se debe a que la distribución de probabilidad de **X** es **simétrica** alrededor de la media.

Ejemplo. En el experimento de obtención de muestras del lote de 5 artículos, encuentre el valor esperado de la variable aleatoria **X:** Número de artículos defectuosos.

Respuesta: Se tiene la Distribución de Probabilidad de **X**:

x	f(x)=P(X=x)
0	1/10
1	6/10
2	3/10

Entonces, el valor esperado de **X** es:

$$\mu = E(X) = \sum_{x=0}^{2} xf(x) = 0(1/10) + 1(6/10) + 2(3/10) = \mathbf{1.2},$$ (media de artículos defectuosos)

En este ejemplo, el Valor Esperado no está en el centro de la distribución de los valores de **X**. Esto se debe a que la Distribución de Probabilidad de **X** no es simétrica alrededor de la media. Es natural que el Valor Esperado se ubique cercano a la región en la que se encuentran los valores de **X** que tienen mayor probabilidad de ocurrir.

La media μ de una variable aleatoria es una medida estadística referida al espacio muestral; mientras que la media muestral $\bar{X}$ se refiere a un subconjunto de la población (espacio muestral)

4.3.1 VALOR ESPERADO DE EXPRESIONES CON UNA VARIABLE ALEATORIA

Se pueden construir expresiones con variables aleatorias. Estas expresiones también son variables aleatorias y su dominio generalmente es el mismo que el dominio de las variables aleatorias, mientras que el rango puede ser diferente.

Definición: Valor Esperado de Expresiones con una Variable Aleatoria

Sea **X:** Variable aleatoria discreta
f(x): Distribución de probabilidad de **X**
G(X): Alguna expresión con la variable aleatoria **X**
Entonces

$$\mu_{G(X)} = E[G(X)] = \sum_{x} G(x)f(x)$$ es la Media o Valor Esperado de **G(X)**

Ejemplo. Sea **X** una variable aleatoria discreta con distribución de probabilidad:

x	f(x)
1	0.1
2	0.4
3	0.3
4	0.2

Sea **G(X) = 2X + 1**. Encuentre **E[G(X)]**

Respuesta.

$$\mu_{G(X)} = E[G(X)] = \sum_{x=1}^{4} G(x)f(x) = (2(1)+1)(0.1) + (2(2)+1)(0.4) + (2(3)+1)(0.3) + (2(4)+1)(0.2) = 6.2$$

Ejemplo. Un almacén vende diariamente **0, 1, 2, 3**, o **4** artículos con probabilidad **10%, 40%, 30%, 15%**, y **5%** respectivamente. Mantener el local le cuesta diariamente **$40** a la empresa. Por cada artículo que vende, tiene una ganancia de **$50**.
Encuentre el valor esperado de la ganancia diaria.

Respuesta:
Sea **X:** Variable aleatoria discreta (número de artículos que vende cada día). Se tiene:

x	f(x)=P(X=x)
0	0.1
1	0.4
2	0.3
3	0.15
4	0.05

Distribución de Probabilidad de **X**

Sea **G(X) = 50X – 40**, variable aleatoria que representa la **ganancia diaria**

Entonces: **E[G(X)]=** $\sum_{x=0}^{4} G(x)f(x)$ = (50(0)-40)(0.1) + (50(1)-40)(0.4) + ... = **42.5**

Definición: Juego Justo

Se dice que "**un juego es justo**" si el valor esperado de la ganancia es cero: μ = **E(X)** = **0**

Ejemplo. Un juego consiste en lanzar tres monedas. Si salen 1 o 2 sellos, se pierde $2. ¿Cuanto se debe ganar en los otros casos para que sea un "juego justo"?

Respuesta:

Sea **X**: Número de sellos (variable aleatoria discreta)
f(x): Distribución de probabilidad de **X**
G(X): Ganancia (variable aleatoria)

Se tiene la Distribución de Probabilidad de **X**:

x	f(x)=P(X=x)	G(x)
0	1/8	k
1	3/8	-2
2	3/8	-2
3	1/8	k

k es la cantidad que se debe ganar cuando salen 0 o 3 sellos.

Entonces **E[G(X)]** = $\sum_{x=0}^{3} G(x)f(x)$ = k(1/8) + (-2)(3/8) + (-2)(3/8) + k(1/8) = 0

Pues el valor esperado debe ser 0. De donde se obtiene **k = 6** dólares.

4.3.2 PROPIEDADES DEL VALOR ESPERADO

Propiedades del Valor Esperado

Sean **X:** Variable aleatoria discreta
f(x): Distribución de probabilidad de X
a, b $\in \Re$**:** Números reales cualesquiera

Entonces **E(aX + b) = aE(X) + b** .

Demostración

$$E(aX+b) = \sum_x (ax+b)f(x) = \sum_x axf(x) + \sum_x bf(x) = a\sum_x xf(x) + b\sum_x f(x)$$

Se tiene $E(X) = \sum_x xf(x)$, además $\sum_x f(x) = 1$, con lo que se completa la demostración.

4.3.3 COROLARIOS

1) **E(aX) = a E(X),** 2) **E(b)=b** .

El segundo corolario establece que si el resultado de un experimento es un valor constante, el valor esperado o media, también debe ser constante.

Ejemplo. Calcule el valor esperado para el ejemplo del almacén usando la nueva fórmula

Respuesta:

$G(X) = 50X - 40$

$E[G(X)] = E(50X - 40) = 50\,E(X) - 40$

$E(X) = \sum_{x=0}^{4} xf(x) = 0(0.1) + 1(0.4) + 2(0.3) + 3(0.15) + 4(0.05) = 1.65$

$\Rightarrow$ $E[G(X)] = 50(1.65) - 40 = \mathbf{42.5}$

4.4 VARIANZA DE UNA VARIABLE ALEATORIA DISCRETA

La Varianza o Variancia es una medida estadística que cuantifica el nivel de dispersión o variabilidad de los valores la variable aleatoria alrededor de la media.

Definición: Varianza de una Variable Aleatoria Discreta

Sea **X:** Variable aleatoria discreta
f(x): Distribución de probabilidad
μ, o **E(X):** Media o Valor Esperado de la variable aleatoria **X**

Entonces

$\sigma^2 = V(X) = E[(X - \mu)^2] = \sum_{x} (x - \mu)^2 f(x)$ es la Varianza de la variable aleatoria **X**

En la definición de la Varianza se suman las diferencias de cada valor **x** con respecto a la media ponderadas con los valores de probabilidad. Elevar al cuadrado puede interpretarse que se suman las magnitudes de las diferencias. El verdadero motivo pertenece a la Teoría Estadística.

Ejemplo. En el experimento de lanzar tres monedas, se definió la variable aleatoria
X: Número de sellos que se obtienen. Calcule la varianza de esta variable aleatoria.

Respuesta: Se tiene la distribución de probabilidad de **X**

x	f(x)=P(X=x)
0	1/8
1	3/8
2	3/8
3	1/8

Previamente se ha calculado el valor esperado de **X:** $\mu = E(X) = \sum_{x=0}^{3} xf(x) = 1.5$

Entonces, usando la definición de la varianza de **X**,

$\sigma^2 = V(X) = E[(X-\mu)^2] = \sum_{x=0}^{3} (x - \mu)^2 f(x) =$

$(0-1.5)^2(1/8) + (1-1.5)^2(3/8) + \ldots + (2-1.5)^2(3/8) + (3-1.5)^2(1/8) = 0.75$

4.4.1 FÓRMULA PARA CALCULAR LA VARIANZA

La siguiente fórmula es equivalente a la anterior. Es importante recordarla

Definición: Fórmula alterna para calcular la Varianza de una Variable Aleatoria Discreta

$$\sigma^2 = V(X) = E[(X-\mu)^2] = E(X^2) - \mu^2$$.

Demostración. Usando las propiedades del valor esperado:

$$V(X) = E[(X-\mu)^2] = E(X^2 - 2\mu X + \mu^2) = E(X^2) - E(2\mu X) + E(\mu^2) =$$
$$= E(X^2) - 2\mu E(X) + \mu^2 = E(X^2) - 2\mu^2 + \mu^2 = E(X^2) - \mu^2$$

Ejemplo. Calcule la varianza en el ejemplo anterior usando la fórmula alterna

$$E(X^2) = \sum_{x=0}^{3} x^2 f(x) = 0^2(1/8) + 1^2(3/8) + 2^2(3/8) + 3^2(1/8) = 3$$

$$\sigma^2 = V(X) = E(X^2) - \mu^2 = 3 - 1.5^2 = 0.75$$

4.4.2 PROPIEDADES DE LA VARIANZA

Definición: Propiedades de la Varianza

Sean **X:** Variable aleatoria discreta
f(x): Distribución de probabilidad de **X**
a, b $\in \Re$**:** Números reales cualesquiera

Entonces $V(aX + b) = a^2V(X)$.

Demostración
Usando la fórmula alterna de varianza y las propiedades del valor esperado:

$$V(aX+b) = E[(aX + b)^2] - E^2(aX + b) = E(a^2X^2 + 2abX + b^2) - [aE(X) + b]^2$$
$$= a^2E(X^2) + 2abE(X) + b^2 - [a^2E^2(X) + 2abE(X) + b^2]$$
$$= a^2[E(X^2) - E^2(X)] = a^2\, V(X)$$

4.4.3 COROLARIOS

1) $V(aX) = a^2\, V(X)$ 2) $V(b)=0$.

El segundo corolario muestra que si el resultado de un experimento es un valor constante entonces la varianza (o variabilidad), es nula.

4.4.4 EJERCICIOS

1) Sea X una variable aleatoria discreta y f su función de distribución de probabilidad:

$$f(x) = \frac{2x+1}{25}, \quad x = 0,\ 1,\ 2,\ 3,\ 4$$

a) Calcule la media de X
b) Sea G(X) = 2X+1. Calcule la media de G(X)
c) Calcule la varianza de X

2) Para ensamblar una máquina se usan dos componentes mecánicos. Suponga que la probabilidad que el primer componente cumpla las especificaciones es 0.95, y para el segundo es 0.98. Además, los componentes funcionan independientemente.
Usando función de distribución de probabilidad de la variable aleatoria X que representa al número de componentes que cumplen las especificaciones, x = 0, 1, 2, obtenida en la unidad anterior.

a) Encuentre la media y la varianza de la variable aleatoria X
b) Suponga que el costo asociado con los componentes instalados que no cumplen las especificaciones es $G(X)=\$5000X^2$. Encuentre el valor esperado de este costo.

MATLAB

Cálculo del valor esperado de una variable aleatoria discreta

```
>> x = [1 2 3 4];
>> f = [0.1 0.4 0.3 0.2];
>> mu = sum(x.*f)
   mu =
        2.6000
```

Valores de la variable aleatoria **X**
Distribución de probabilidad de la variable **X**
Media de **X**

Valor esperado de una expresión

```
>> g = 2*x+1;
>> mug=sum(g .*f)
   mug =
         6.2000
```

Una expresión con **X:** **g(X) = 2x + 1**
Media de **g(X)**

Cálculo de la varianza de una variable aleatoria discreta

```
>> sigma2 = var(x, f)
   sigma2 =
              0.8400
```

4.5 MOMENTOS DE UNA VARIABLE ALEATORIA DISCRETA

La media de una variable aleatoria discreta describe su tendencia central y la variancia mide su variabilidad, pero estas medidas no son suficientes para describir completamente la forma de la distribución de probabilidad.

Los momentos de una variable aleatoria son los valores esperados de algunas funciones de la variable aleatoria y constituyen una colección de medidas descriptivas con las que se puede caracterizar de manera única a su distribución de probabilidad. Usualmente estas definiciones se las hace usando como referencia el origen o la media de la variable aleatoria.

4.5.1 MOMENTOS ALREDEDOR DEL ORIGEN

Definición: Momentos alrededor del Origen

Sea **X:** Variable aleatoria discreta
f(x): Distribución de probabilidad de **X**
Entonces, el r-ésimo momento de **X** alrededor del origen es:

$$\mu'_r = E(X^r) = \sum_x x^r f(x) \quad .$$

r=1: $\mu'_1 = E(X) = \sum_x x f(x) = \mu$ (Primer Momento alrededor del origen. Es la **media**)

r=2: $\mu'_2 = E(X^2) = \sum_x x^2 f(x)$ (Segundo Momento alrededor del origen)

etc.

4.5.2 MOMENTOS ALREDEDOR DE LA MEDIA

Definición: Momentos alrededor de la Media

Sea **X:** Variable aleatoria discreta
f(x): Distribución de probabilidad de **X**
Entonces, el r-ésimo momento de **X** alrededor de la media o r-ésimo momento central, es:

$$\mu_r = E[(X-\mu)^r] = \sum_x (x-\mu)^r f(x) \quad .$$

r=1: $\mu_1 = E[(X-\mu)] = E(X) - \mu = 0$ (Primer Momento Central)
r=2: $\mu_2 = E[(X-\mu)^2] = \sigma^2$ (Segundo Momento Central. Es la **varianza**)
r=3: $\mu_3 = E[(X-\mu)^3]$ (Tercer Momento Central)
r=4: $\mu_4 = E[(X-\mu)^4]$ (Cuarto Momento Central)

El Segundo Momento Central o Varianza, mide la **dispersión**
El Tercer Momento Central, mide la **asimetría** o sesgo
El Cuarto Momento Central, mide la **curtosis** o "puntiagudez".

Se definen coeficientes para expresar los momentos en forma adimensional para que no dependan de la escala de medición y puedan usarse para comparar la distribución entre variables aleatorias. Para los tres momentos centrales indicados arriba, son respectivamente:

4.5.3 COEFICIENTES PARA COMPARAR DISTRIBUCIONES

Definiciones

Coeficiente de Variación: σ/μ
Coeficiente de Asimetría: $\mu_3/(\mu_2)^{3/2}$
Coeficiente de Curtosis $\mu_4/(\mu_2)^2$

VALORES REFERENCIALES

Valores referenciales y significado de algunos coeficientes

Coeficiente de Asimetría

Positivo:	La distribución tiene sesgo positivo (se extiende a la derecha)
Cero:	La distribución es simétrica.
Negativo:	La distribución tiene sesgo negativo (se extiende a la izquierda)

Coeficiente de Curtosis

Mayor a 3:	La distribución es “puntiaguda” o “leptocúrtica”
Igual a 3:	La distribución es “regular”
Menor a 3:	La distribución es “plana” o “platicúrtica”

4.5.4 EQUIVALENCIA ENTRE MOMENTOS

Los momentos centrales pueden expresarse mediante los momentos alrededor del origen usando la definición de valor esperado:

$\mu_2 = E[(X-\mu)^2] = E(X^2) - \mu^2 = \mu'_2 - \mu^2$ (Es la definición de varianza)

$\mu_3 = E[(X-\mu)^3] = \mu'_3 - 3\mu\mu'_2 + 2\mu^3$

$\mu_4 = E[(X-\mu)^4] = \mu'_4 - 4\mu\mu'_3 + 6\mu^2\mu'_2 - 3\mu^4$

4.6 FUNCIÓN GENERADORA DE MOMENTOS

Es una función especial que puede usarse para obtener todos los momentos de una variable aleatoria discreta

Definición: Función Generadora de Momentos de una Variable Aleatoria Discreta

Sea **X:** Variable aleatoria discreta
f(x): Distribución de probabilidad de **X**
Entonces la función generadora de momentos de **X** es:

$$M(t) = E(e^{tX}) = \sum_x e^{tx} f(x)$$

4.6.1 OBTENCIÓN DE MOMENTOS

La definición matemática de la función generadora de momentos se fundamenta en el desarrollo de e^{tX} en serie de potencias:

$$e^{tX} = 1 + tX + t^2X^2/2! + t^3X^3/3! + \ldots$$

Con la definición de valor esperado se obtiene:

$$\begin{aligned} M(t) = E(e^{tX}) &= E(1) + E(tX) + E(t^2X^2/2!) + E(t^3X^3/3!) + \ldots \\ &= 1 + t\,E(X) + t^2/2!\,E(X^2) + t^3/3!\,E(X^3) + \ldots \\ &= 1 + (t)\,\mu'_1 + (t^2/2!)\,\mu'_2 + (t^3/3!)\,\mu'_3 + \ldots \end{aligned}$$

Este desarrollo justifica el uso de la siguiente fórmula como un dispositivo matemático para obtener cualquier momento alrededor del origen, de una variable aleatoria discreta:

Definición: Fórmula para obtención de Momentos alrededor del origen

$$\mu'_r = \frac{d^r}{dt^r} M(t)\,|_{t=0}$$

Aplicación

1) Primer momento alrededor del origen:

$$\frac{d}{dt}M(t)|_{t=0} = \frac{d}{dt}E[e^{tX}]|_{t=0} = E[\frac{d}{dt}e^{tX}]|_{t=0} = E[Xe^{tX}]_{t=0} = E(X) = \mu'_1$$

2) Segundo momento alrededor del origen:

$$\frac{d^2}{dt^2}M(t)|_{t=0} = \frac{d^2}{dt^2}E[e^{tX}]|_{t=0} = E[\frac{d^2}{dt^2}e^{tX}]|_{t=0} = E[X^2 e^{tX}]_{t=0} = E(X^2) = \mu'_2$$

Ejemplo.
Suponga una variable aleatoria discreta **X** con la siguiente distribución de probabilidad:

x	f(x)
1	0.2
2	0.3
3	0.4
4	0.1

a) Encuentre el Coeficiente de Variación

$$\mu = \mu'_1 = E(X) = \sum_{x=1}^{4} xf(x) = 1(0.2) + 2(0.3) + 3(0.4) + 4(0.1) = 2.4$$

$$\mu'_2 = E(X^2) = \sum_{x=1}^{4} x^2 f(x) = 1^2(0.2) + 2^2(0.3) + 3^2(0.4) + 4^2(0.1) = 6.6$$

$$\mu_2 = \sigma^2 = E[(X-\mu)^2] = E(X^2) - \mu^2 = \mu'_2 - (\mu'_1)^2 = 6.6 - (2.4)^2 = 0.84$$

$$v = \sigma/\mu = \sqrt{0.84}/2.4 = 0.3819$$

b) Encuentre el Coeficiente de Asimetría

$$\mu'_3 = E(X^3) = \sum_{x=1}^{4} x^3 f(x) = 1^3(0.2) + 2^3(0.3) + 3^3(0.4) + 4^3(0.1) = 19.8$$

$$\mu_3 = E[(X-\mu)^3] = \mu'_3 - 3\mu\mu'_2 + 2\mu^3 = 19.8 - 3(2.4)(6.6) + 2(2.4)^3 = -0.072$$

Coeficiente de asimetría: $\mu_3/(\mu_2)^{3/2} = -0.072/(0.84)^{3/2} = -0.0935$

Siendo este valor negativo, se concluye que la distribución es asimétrica con sesgo hacia la izquierda.

c) Encuentre la Función Generadora de Momentos

$$M(t) = E(e^{tX}) = \sum_{x} e^{tx} f(x) = \sum_{x=1}^{4} e^{tx} f(x) = 0.2e^t + 0.3e^{2t} + 0.4e^{3t} + 0.1e^{4t}$$

d) Encuentre la Media de X usando la Función Generadora de Momentos

$$\frac{d}{dt}M(t)|_{t=0} = \frac{d}{dt}(0.2e^t + 0.3e^{2t} + 0.4e^{3t} + 0.1e^{4t})|_{t=0}$$

$$= [0.2(e^t) + 0.3(2e^{2t}) + 0.4(3e^{3t}) + 0.1(4e^{4t})]|_{t=0}$$

$$= 0.2(1) + 0.3(2) + 0.4(3) + 0.1(4) = 2.4$$

Con la función generadora de momentos se pueden obtener todos los momentos de la variable aleatoria. Los momentos son las medidas descriptivas de la variable aleatoria, con los cuales se puede caracterizar a su función de probabilidad.

Si la función generadora de momentos existe, entonces esta es única. Por lo tanto permite describir completamente a la distribución de probabilidad de una variable aleatoria. Una consecuencia de este argumento es la siguiente propiedad

4.6.2 UNICIDAD DE FUNCIONES DE DISTRIBUCIÓN DE PROBABILIDAD

Definición: Unicidad de Funciones de Distribución de Probabilidad

Sean

X, Y	Variables aleatorias discretas
f(x), f(y)	Distribuciones de Probabilidad de **X**, **Y** respectivamente
$M_X(t)$, $M_y(t)$	Funciones Generadoras de Momentos de **X**, **Y** respectivamente

Si $M_X(t) = M_y(t)$ entonces $f(x) = f(y)$

Si dos variables aleatorias tienen funciones generadoras de momentos idénticas, entonces tienen idénticas funciones de distribución de probabilidad. Esta propiedad usa el hecho de que una función generadora de momentos describe la forma de la distribución de probabilidad

4.7 TEOREMA DE CHEBYSHEV

Este teorema establece un valor mínimo para la probabilidad de una variable aleatoria en un intervalo alrededor de la media, independientemente de su función de probabilidad. El valor que se obtiene es únicamente una referencia.

Definición: Teorema de Chebyshev

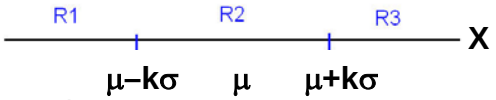
Sea **X** una variable aleatoria discreta con media μ y varianza σ^2, entonces, la probabilidad que **X** tome un valor dentro de **k** desviaciones estándar σ de su media μ, es al menos $1 - 1/k^2$

$$P(\mu - k\sigma < X < \mu + k\sigma) \geq 1 - 1/k^2 \ , \ k \in \Re^+, k \geq 1$$

Demostración:

En esta demostración se incluye una variable aleatoria discreta, pero también se puede demostrar para una variable aleatoria continua.

Separamos el dominio de la variable aleatoria **X** en tres regiones R_1, R_2, R_3:



Con la definición de varianza:

$$\sigma^2 = E[(X-\mu)^2] = \sum_x (x-\mu)^2 f(x)$$

$$= \sum_{R1} (x-\mu)^2 f(x) + \sum_{R2} (x-\mu)^2 f(x) + \sum_{R3} (x-\mu)^2 f(x)$$

$$\sigma^2 > \sum_{R1} (x-\mu)^2 f(x) + \sum_{R3} (x-\mu)^2 f(x),$$ se suprime un término positivo

En **R1**: $x \leq \mu - k\sigma \Rightarrow -x \geq -\mu + k\sigma \Rightarrow -(x-\mu) \geq k\sigma \Rightarrow (x-\mu)^2 \geq k^2\sigma^2$

En **R3**: $x \geq \mu + k\sigma \Rightarrow x-\mu \geq k\sigma \Rightarrow (x-\mu)^2 \geq k^2\sigma^2$

Al sustituir en las sumatorias, se mantiene la desigualdad:

$$\sigma^2 > \sum_{R1} k^2\sigma^2 f(x) + \sum_{R3} k^2\sigma^2 f(x),$$

De donde se obtiene, simplificando,

$$1/k^2 > \sum_{R1} f(x) + \sum_{R3} f(x),$$

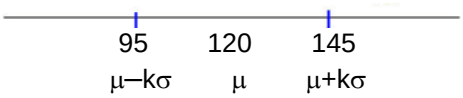
Las sumas son valores de probabilidad

$$1/k^2 > P(X \leq \mu - k\sigma \ \vee \ X \geq \mu + k\sigma),$$

Finalmente, con el complemento de probabilidad: $1 - 1/k^2 \leq P(\mu - k\sigma < X < \mu + k\sigma)$

Ejemplo.
La producción diaria de una fábrica es una variable aleatoria discreta con media 120 artículos, y desviación estándar de 10 artículos. Calcule la probabilidad que en cualquier día la producción esté entre 95 y 145 artículos.

Respuesta



Por lo tanto, $k\sigma = 25 \Rightarrow k(10) = 25 \Rightarrow k = 2.5$

$P(95 < X < 145) \geq 1 - 1/2.5^2 \Rightarrow P(95 < X < 145) \geq 0.84$

4.8 EJERCICIOS

1) Suponga una variable aleatoria discreta **X** con la siguiente distribución de probabilidad:

x	f(x)
1	0.10
2	0.20
3	0.50
4	0.15
5	0.05

a) Encuentre el coeficiente de variación
b) Encuentre el coeficiente de asimetría e interprete el resultado
c) Encuentre el coeficiente de curtosis e interprete el resultado
d) Encuentre la función generadora de momentos
e) Encuentre la media de la variable aleatoria usando la función generadora de momentos

2) Encuentre el menor valor de **k** en el teorema de Chebyshev para el cual la probabilidad de que una variable aleatoria tome un valor entre $\mu - k\sigma$ y $\mu + k\sigma$ sea
a) cuando menos 0.95
b) cuando menos 0.99

MATLAB

```
>> x = [1 2 3 4];                    Valores de la variable aleatoria X
>> f = [0.2 0.3 0.4 0.1];            Distribución de probabilidad de la variable X
>> mu=sum(x.*f)                      Media
   mu =
        2.4000
>> mu2=sum((x-mu).^2.*f)             Varianza
   mu2 =
         0.8400
>> mu3=sum((x-mu).^3.*f)             Asimetría
   mu3 =
        -0.0720
>> mu4=sum((x-mu).^4.*f)             Curtosis
   mu4 =
         1.4832
>> syms t
>> fgm=sum(exp(x*t).*f)              Función generadora de momentos
   fgm =
        1/5*exp(t)+3/10*exp(2*t)+2/5*exp(3*t)+1/10*exp(4*t)
>> t=0;
>> mu=eval(diff(fgm))                Media usando la función generadora de momentos
   mu =
        2.4000
```

5 DISTRIBUCIONES DE PROBABILIDAD DISCRETAS

En este capítulo se estudian los modelos matemáticos para calcular la probabilidad en algunos problemas típicos en los que intervienen variables aleatorias discretas.

El objetivo es obtener una fórmula matemática **f(x)** para determinar los valores de probabilidad de la variable aleatoria **X**.

5.1 DISTRIBUCIÓN DISCRETA UNIFORME

Una variable aleatoria tiene distribución discreta uniforme si cada uno de los resultados de su espacio muestral tiene puede obtenerse con igual probabilidad.

Definición: Distribución discreta uniforme

Sean **X:** Variable aleatoria discreta
x = x_1, x_2, x_3, ..., x_n Son los **n** valores que puede tomar **X** con igual probabilidad
Entonces la distribución de probabilidad de **X** es:

$$f(x) = \begin{cases} \frac{1}{n}, & x = x_1, x_2, \ldots, x_n \\ 0, & \text{otro } x \end{cases}$$

.

Ejemplo. Un experimento consiste en lanzar un dado y observar el resultado.
Si **X** es la variable aleatoria correspondiente a los seis resultados posibles, encuentre su distribución de probabilidad.

Respuesta
Cada resultado tiene igual probabilidad, por lo tanto la distribución de probabilidad de **X** es discreta uniforme:

$$P(X = x) = f(x) = \begin{cases} 1/6, & x=1, 2, \ldots, 6 \\ 0, & \text{para otro } x \end{cases}$$

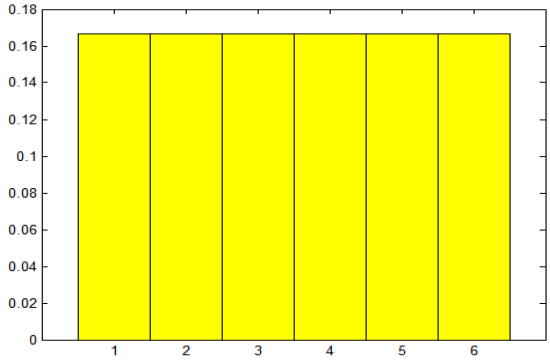
Calcule la probabilidad que **X** tome el valor **3**

P(X = 3) = f(3) = 1/6

Gráfico de la distribución discreta uniforme

El gráfico de la distribución discreta uniforme tiene forma regular

Ejemplo. Graficar la distribución de probabilidad para el ejemplo anterior



5.1.1 MEDIA Y VARIANZA DE LA DISTRIBUCIÓN DISCRETA UNIFORME

Se obtienen directamente de las definiciones correspondientes

Definición: Media y Varianza de una variable con Distribución Discreta Uniforme

Sea **X**: Variable aleatoria con Distribución Discreta Uniforme

Media: $$\mu = E(X) = \sum_x xf(x) = \sum_{i=1}^{n} x_i f(x_i) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Varianza: $$\sigma^2 = E[(X-\mu)^2] = \sum_x (x_i-\mu)^2 f(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

Ejemplo. Un almacén vende diariamente 0, 1, 2, 3, o 4 artículos con igual probabilidad. Calcule la probabilidad que en algún día venda al menos 2 artículos

Respuesta

Sea **X**: Cantidad de artículos que vende cada día (variable aleatoria discreta)
x = 0, 1, 2, 3, 4
X tiene distribución uniforme con probabilidad **1/5**
P(X = x) = f(x) = 0.2, x = 0, 1, 2, 3, 4
P(X≥2) = f(2) + f(3) + f(4) = 3(0.2) = 0.6

5.2 DISTRIBUCIÓN DE BERNOULLI

Es un experimento estadístico en el que pueden haber únicamente dos resultados posibles. Es costumbre designarlos como **"éxito"** y **"fracaso"** aunque pueden tener otra representación y estar asociados a algún otro significado de interés.

Si la probabilidad de obtener **"éxito"** en cada ensayo es un valor que lo representamos con **p**, entonces, la probabilidad de obtener **"fracaso"** será el complemento **q = 1 – p**.

Definición: Distribución de Bernoulli

Sean **X:** Variable aleatoria cuyos valores pueden ser **1:** "éxito", **0:** "fracaso"
p: Valor de probabilidad de que el resultado del ensayo sea "éxito"
Entonces, la distribución de probabilidad de **X** es

$$f(x) = \begin{cases} p, & x = 1 \\ 1-p, & x = 0 \end{cases}$$

El experimento puede repetirse y en cada ensayo el valor de probabilidad **p** se mantiene **constante**. Se supondrá también que los ensayos son **independientes,** es decir el resultado de un ensayo no afecta a los resultados de los otros ensayos.

Suponer que se desean obtener los siguientes resultados: **1 1 0 0 1 0 ...**, en donde **1** es **"exito", 0** es **"fracaso"**

Sean **p** Probabilidad que el resultado sea éxito
q = 1 – p Probabilidad que el resultado sea fracaso

Entonces la probabilidad de obtener esta secuencia de resultados es:

P(X=1,X=1,X=0,X=0,X=1,X=0, ...) = f(1) f(1) f(0 f(0) f(1) f(0) ... = pp(1-p)(1-p)pq...

Ejemplo. Suponer que la probabilidad de éxito de un experimento es **0.2** y se realizan cinco ensayos independientes. Calcule la probabilidad que el **primero y el último ensayo sean éxitos**, y los **tres ensayos intermedios sean fracasos**.

Sean **1**: El ensayo es éxito (con probabilidad **0.2**)
0: El ensayo es fracaso (con probabilidad **0.8**)
Entonces

P(X=1,X=0,X=0,X=0,X=1) = f(1)f(0)f(0)f(0)f(1) = (0.2)(0.8)(0.8)(0.8)(0.2) = 0.0205 = 2.05%

5.3 DISTRIBUCIÓN BINOMIAL

Esta distribución es muy importante y de uso frecuente. Corresponde a experimentos con características similares a un experimento de Bernoulli, pero ahora es de interés la variable aleatoria relacionada con la **cantidad de "éxitos"** que se obtienen en el experimento.

Características de un Experimento Binomial

a) La cantidad de ensayos **n**, que se realizan es finita.
b) Cada ensayo tiene únicamente **dos** resultados posibles: **"éxito"** o **"fracaso"**
c) Todos los ensayos realizados son **independientes**
d) La probabilidad **p**, de obtener **"éxito"** en cada ensayo permanece constante.

Algunos ejemplos de problemas con estas características

1) Determinar la probabilidad de la cantidad de artículos que son defectuosos en una muestra tomada al azar de la producción de una fábrica, suponiendo conocida la probabilidad de que un artículo sea defectuoso
2) Determinar la probabilidad de la cantidad de personas que están a favor de un candidato, en una muestra de personas elegidas al azar de una población grande. Suponiendo conocida la probabilidad de que una persona esté a favor del candidato.

Definición: Distribución Binomial

Sean **X:** Variable aleatoria discreta (representa la cantidad de ensayos considerados **"éxitos"** en una serie de **n** ensayos realizados).
x = **0, 1, 2, ..., n** valores que puede tomar **X**
p: Probabilidad de que el resultado de cada ensayo sea **"éxito"**
Entonces, la distribución de probabilidad de **X** es

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, \ldots, n$$

Demostración

Al realizar **n** ensayos se obtienen **x** éxitos y **n - x** fracasos, por lo tanto siendo ensayos independientes la probabilidad de obtener estos resultados es $p^x (1-p)^{n-x}$

Pero, en los **n** ensayos realizados hay $\binom{n}{x}$ formas diferentes de obtener los **x** éxitos y los **n - x** fracasos. Este número es entonces un factor para el valor de probabilidad anterior.

Los símbolos $\binom{n}{x}$, $_nC_x$, C_x^n representan el número de combinaciones o arreglos diferentes que se obtienen con **n** elementos de los cuales se toman **x** elementos.

Ejemplo. Se realizan 8 lanzamientos de un dado. Calcule la probabilidad de obtener 4 veces el número 5.

Respuesta. Este experimento tiene las características de un experimento binomial con:

n = 8:	Cantidad de ensayos realizados (se suponen independientes)
p = 1/6	Probabilidad que cada ensayo sea "éxito" (se obtiene el 5)
X:	Variable aleatoria discreta (cantidad de veces que sale el 5)
x = 0, 1, 2, ..., 8	Valores que puede tomar **X**

Es un problema cuyo modelo de probabilidad es Binomial. Sustituyendo los datos:

$$P(X=x) = f(x) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{8}{x} (1/6)^x (5/6)^{8-x}, \; x = 0, 1, 2, \ldots, 8$$

De donde se obtiene

$$P(X=4) = f(4) = \binom{8}{4} (1/6)^4 (5/6)^{8-4} = (70)(1/6)^4 (5/6)^4 = 0.026 = 2.6\%$$

Ejemplo Una fábrica tiene una norma de control de calidad consistente en elegir al azar diariamente 20 artículos producidos y determinar el número de unidades defectuosas. Si hay dos o más artículos defectuosos la fabricación se detiene para inspección de los equipos. Se conoce por experiencia que la probabilidad de que un artículo producido sea defectuoso es 5%. Encuentre la probabilidad de que en cualquier día la producción se detenga al aplicar esta norma de control de calidad.

Respuesta

Esta situación corresponde a un experimento binomial

n = 20 Cantidad de ensayos (independientes)
p = **0.05** Probabilidad de éxito (constante)
X: Variable aleatoria discreta (cantidad de artículos defectuosos)
x = **0, 1, ..., 20** Valores que puede tomar **X**

$$P(X=x) = f(x) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{20}{x} 0.05^x (0.95)^{20-x}, \quad x = 0,1,2, \ldots,20$$

Entonces

$$P(X \geq 2) = 1 - P(X \leq 1) \qquad \text{(conviene usar esta propiedad)}$$
$$= 1 - (P(X=0) + P(X=1)) = 1 - (f(0) + f(1))$$
$$f(0) = \binom{20}{0} 0.05^0 (0.95)^{20} = 0.3585$$
$$f(1) = \binom{20}{1} 0.05^1 (0.95)^{19} = 0.3774$$
$$P(X \geq 2) = 1 - 0.3585 - 0.3774 = 0.2641 = 26.41\%$$

5.3.1 PARÁMETROS Y VARIABLES

Los **parámetros** de un modelo de distribución de probabilidad se refieren a valores con los que se describe un problema particular. Para la Distribución Binomial los parámetros son **n** y **p**.

Una vez que está definido el problema, se especifica la **variable aleatoria** de interés y se procede a calcular la probabilidad correspondiente a los valores que puede tomar esta variable.

Se puede usar la siguiente notación para distinguir entre **variables** y **parámetros**:

$$f(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, \ldots, n$$

Variable — Parámetros

En el ejemplo anterior, el modelo de distribución de probabilidad se puede escribir:

$$f(x; 20, 0.05) = \binom{20}{x} 0.05^x (0.95)^{20-x}, \quad x = 0, 1, \ldots, 20$$

5.3.2 DISTRIBUCIÓN DE PROBABILIDAD BINOMIAL ACUMULADA

Definición: Distribución de Probabilidad Binomial Acumulada

Sea **X:** Variable aleatoria discreta con Distribución Binomial con parámetros **n, p**

Entonces, la Distribución de Probabilidad Acumulada **F** de la variable **X** es

$$F(x) = P(X \leq x) = \sum_{t \leq x} \binom{n}{t} p^t (1-p)^{n-t}, \quad x \geq 0$$

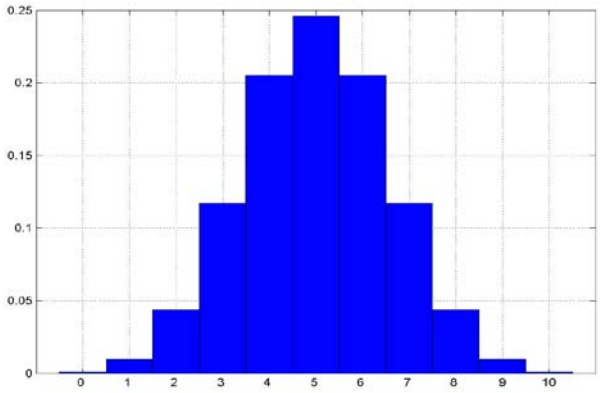
5.3.3 GRAFICO DE LA DISTRIBUCIÓN BINOMIAL

La distribución binomial tiene su gráfico con forma simétrica cuando **p=0.5**

Ejemplo. Grafique la distribución binomial con **n=10**, **p=0.5**

$$f(x) = \binom{10}{x}(0.5)^x(0.5)^{10-x} = \binom{10}{x}(0.5)^{10}, x = 0,1,...,10$$

f(0) = 0.0010
f(1) = 0.0098
f(2) = 0.0439
. . .
f(8) = 0.0439
f(9) = 0.0098
f(10) = 0.0010



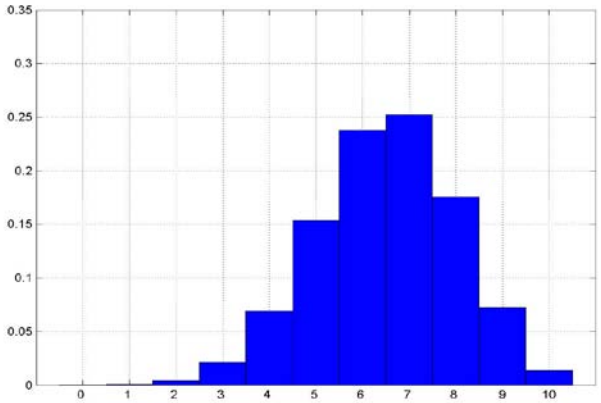
Distribución binomial con p=0.5

Si **p>0.5**, la forma de la distribución binomial tiene sesgo negativo.
Si **p<0.5**, la forma de la distribución binomial tiene sesgo positivo.

Ejemplo
Grafique la distribución binomial con **n=10, p=0.65**

$$f(x) = \binom{10}{x}(0.65)^x(0.35)^{10-x}, x = 0,1,...,10$$

f(0) = 0.0000
f(1) = 0.0005
. . .
f(9) = 0.0725
f(10) = 0.0135



Distribución binomial con p>0.5

5.3.4 MEDIA Y VARIANZA DE LA DISTRIBUCIÓN BINOMIAL

Definición: Media y Varianza de la Distribución Binomial

Sea **X**: variable aleatoria discreta con Distribución Binomial con parámetros **n, p**
Entonces

$\mu = E(X) = np$ Media de **X**

$\sigma^2 = V(X) = np(1-p)$ Varianza de **X**

Demostración
Esta demostración usa la definición de función generadora de momentos para variables aleatorias discretas

Distribución de probabilidad de la distribución binomial:

$$f(x) = \binom{n}{x} p^x q^{n-x}, \; x = 0, 1, ..., n, \qquad \text{siendo } q = 1 - p$$

Los términos de la distribución binomial coinciden con el desarrollo del binomio: $(q + p)^n$

$$(q + p)^n = \binom{n}{0} p^0 q^n + \binom{n}{1} p^1 q^{n-1} + ... + \binom{n}{n} p^n q^0 = \sum_{x=0}^{n} \binom{n}{x} p^x q^{n-x}$$

La función generadora de momentos para la distribución binomial:

$$m(t) = E(e^{tX}) = \sum_{x=0}^{n} e^{tx} f(x) = \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x q^{n-x} = \sum_{x=0}^{n} \binom{n}{x} (e^t p)^x q^{n-x}$$

Se puede observar que la última expresión tiene la misma forma que la fórmula del binomio sustituyendo **p** por $e^t p$. Entonces se tiene

Definición: Función Generadora de Momentos de la Distribución Binomial

$$m(t) = (q + e^t p)^n$$

Con la definición correspondiente se pueden obtener los momentos alrededor del origen:

$$\mu = \mu'_1 = \frac{d}{dt} m(t) \Big|_{t=0} = \frac{d}{dt} (e^t p + q)^n \Big|_{t=0} = n(e^t p + q)^{n-1} e^t p \Big|_{t=0} = n(p + q)^{n-1} p,$$

Pero $p + q = 1$, entonces: $\mu = np$. Esto completa la demostración.

La demostración de la varianza sigue un camino similar. Primero se encuentra μ'_2 con la definición:

$\mu'_2 = \frac{d^2}{dt^2} m(t)|_{t=0}$, y después se usa la definición: $\sigma^2 = V(X) = E(X^2) - \mu^2 = \mu'_2 - \mu^2$

Ejemplo.
Encuentre la media y la varianza para el ejemplo del control de calidad en la fábrica.

Respuesta:

$\mu = np = 20\,(0.05) = 1$

$\sigma^2 = npq = 20(0.05)(0.95) = 0.95$

μ representa la cantidad promedio de artículos defectuosos que se obtienen cada día
σ^2 es una medida de la variabilidad o dispersión de los valores de **X**

5.3.5 EJERCICIOS

1) La variable aleatoria **X** tiene distribución discreta uniforme para **x=1, 2, 3, . . . , 50**
a) Determine la media y varianza de **X**
b) Calcule **P(5<X≤10)**
c) Calcule la media y varianza de la variable aleatoria **Y=5X**

2) La variable aleatoria **X** tiene distribución binomial con **n=8, p=0.4**.
a) Defina la función de distribución de probabilidad de **X**
b) Grafique la función de distribución de probabilidad
c) Grafique la función de distribución de probabilidad acumulada
d) Cuales son los valores de **X** mas factibles que ocurran
e) Cuales son los valores de **X** menos factibles
f) Calcule **P(X=5)**
g) Calcule **P(X≤2)**

3) Un ingeniero que labora en el departamento de control de calidad de una empresa eléctrica, inspecciona una muestra al azar de tres motores de la producción. Se sabe que 15% de los motores salen defectuosos. Calcule la probabilidad que en la muestra
a) Ninguno sea defectuoso,
b) Uno sea defectuosos,
c) Al menos dos sean defectuosos?
d) Obtenga la media y la varianza de la variable aleatoria del problema

4) La probabilidad de que disco compacto dure al menos un año sin que falle es de 0.95. Calcule la probabilidad de que en 15 de estos aparatos elegidos al azar,
a) 12 duren menos de un año,
b) A lo más 5 duren menos de un año,
c) Al menos 2 duren menos de un año.
d) Obtenga la media y la varianza de la variable aleatoria del problema

5) Un examen de opciones múltiples tiene 20 preguntas y cada pregunta tiene cuatro posibles respuestas entre las cuales se debe elegir la correcta. Un estudiante decide usar una moneda para contestar el examen de la siguiente manera:

Para cada pregunta lanza dos veces la moneda.
Si el resultado es (cara, cara) marca la primera opción
Si el resultado es (cara, sello) marca la segunda opción
Si el resultado es (sello, cara) marca la tercera opción
Si el resultado es (sello, sello) marca la cuarta opción
Para aprobar el examen se necesita marcar al menos 60% de las respuestas correctas.

Calcule la probabilidad que este estudiante (**?**) apruebe el examen

MATLAB

Probabilidad con la distribución binomial

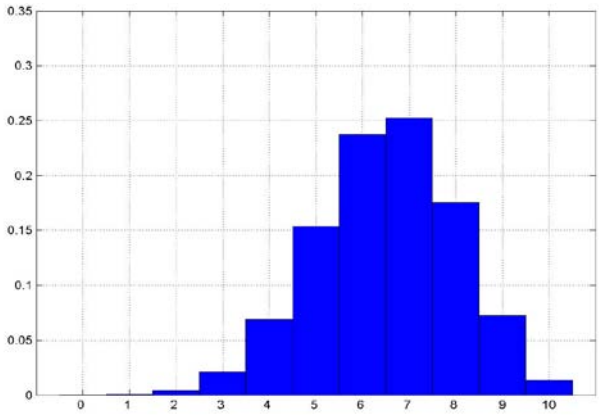
```
>> f = binopdf(0, 20, 0.05)
    f =
        0.3585
```
Probabilidad con la distribución binomial: **x=0, n=20, p=0.05**

```
>> f = binopdf(1, 20, 0.05)
    f =
        0.3774
```
Probabilidad con la distribución binomial: **x=1, n=20, p=0.05**

```
>> f = binocdf(3, 10, 0.2)
    f =
        0.8791
```
Probabilidad con la distribución binomial **acumulada**
P(X≤3), n = 10, p = 0.2

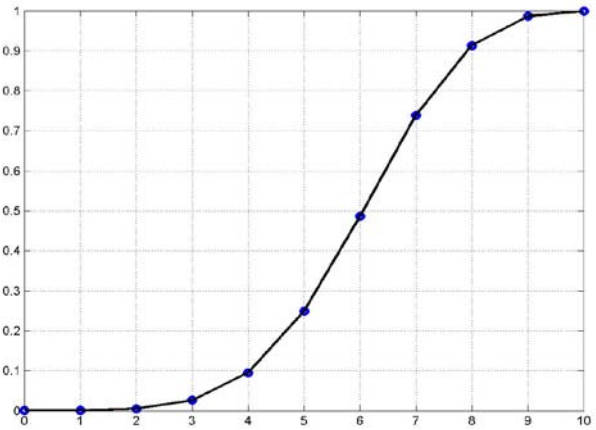
```
>> x = 0:10;
>> f = binopdf(x, 10, 0.65)
    f =
    0.0000  0.0005  0.0043  0.0212  0.0689  0.1536  0.2377  0.2522  0.1757  0.0725  0.0135
```
Valores para evaluar la distribución binomial, **x=0, 1, 2, . . ., 10**
Distribución binomial, **x=0, 1, 2, . . ., 10; n=10, p=0.65**

```
>> bar(f, 1, 'b'), grid on
```
Gráfico de la distribución de probabilidad en color azul




```
>> f = binocdf(x, 10, 0.65);
   f =
    0.0000  0.0005  0.0048  0.0260  0.0949  0.2485  0.4862  0.7384  0.9140  0.9865  1.0000
```
Distribución binomial acumulada, **x=0, 1, 2, . . ., 10**
n=10, p=0.65

```
>> plot(x, f, 'ob')
>> hold on
>> plot(x,f,'k'), grid on
```
Gráfico de los puntos de la distribución acumulada, en azul
Gráfico superpuesto de la distribución acumulada, en negro



5.4 DISTRIBUCIÓN BINOMIAL NEGATIVA

Este modelo de probabilidad tienen características similares al modelo binomial: los ensayos son independientes, cada ensayo tiene únicamente dos resultados posibles, y la probabilidad que cada ensayo tenga un resultado favorable es constante. Pero, en este modelo la variable aleatoria es diferente:

En la Distribución Binomial Negativa, la variable de interés es la cantidad de ensayos que se realizan hasta obtener un número requerido de éxitos, **k**

Definición: Distribución Binomial Negativa

Sea **X**: Variable aleatoria discreta con Distribución Binomial Negativa
(cantidad de ensayos realizados hasta obtener **k** "éxitos")
p: Probabilidad de "éxito". Es un valor constante en cada ensayo
x = k, k+1, k+2, ... (valores que puede tomar la variable **X**)
Entonces la distribución de probabilidad de **X** es:

$$P(X=x) = f(x) = \binom{x-1}{k-1} p^k (1-p)^{x-k}, \quad x = k, k+1, k+2, \ldots$$

Demostración

Cada **"éxito"** ocurre con probabilidad **p** y cada **"fracaso"** con probabilidad **1 – p.**

En algún ensayo **x** se tendrán finalmente **k** éxitos. Por lo tanto siendo ensayos independientes la probabilidad de obtener los **k** "éxitos" y los **x – k** "fracasos" es el producto: $p^k (1-p)^{x-k}$

Pero, antes de obtener el **k-ésimo** "éxito" se realizaron **x–1** ensayos con los previos **k – 1** "éxitos". Esto puede ocurrir en $\binom{x-1}{k-1}$ formas diferentes, por lo que este número es un factor para la fórmula. Esto se completa la demostración

Está claro que la cantidad de ensayos que deben realizarse es al menos **k.**

Ejemplo.
Suponiendo que la probabilidad de que una persona contraiga cierta enfermedad a la que está expuesta es **30%**, calcule la probabilidad que la **décima** persona expuesta a la enfermedad sea la **cuarta** en contraerla.

Respuesta
Cada persona expuesta a la enfermedad constituye un ensayo. Estos ensayos son independientes y la probabilidad de "éxito" es constante: **0.3**. (Note que "éxito" no siempre tiene una connotación favorable)

Por la pregunta concluimos que la variable de interés **X** tiene Distribución Binomial Negativa con **k=4**, **p=0.3**.

Sean **X:** Cantidad de ensayos realizados hasta obtener **k** "éxitos" (variable aleatoria discreta)
x = 4, 5, 6, . . .

$$P(X=x) = f(x) = \binom{x-1}{4-1} 0.3^4 (1-0.3)^{x-4}, \quad x=4, 5, 6, \ldots$$

Por lo tanto

$$P(X=10) = f(10) = \binom{10-1}{4-1} 0.3^4\, 0.7^{10-4} = 0.08$$

5.4.1 MEDIA Y VARIANZA DE LA DISTRIBUCIÓN BINOMIAL NEGATIVA

Definición: Media y Varianza para la Distribución Binomial Negativa

Media: $\mu = E[X] = \frac{k}{p}$, **Varianza:** $\sigma^2 = V[X] = \frac{k}{p}(\frac{1}{p} - 1)$

5.5 DISTRIBUCIÓN GEOMÉTRICA

Es un caso especial de la distribución binomial negativa, cuando **k=1**. Es decir interesa conocer la probabilidad respecto a la cantidad de ensayos que se realizan hasta obtener el primer "éxito"

Definición: Distribución Geométrica

Sean **X**: Variable aleatoria discreta con Distribución Geométrica
(cantidad de ensayos realizados hasta obtener el primer 'éxito')
x = 1, 2, 3, ... (valores factibles para la variable **X**)
p: probabilidad constante de "éxito" en cada ensayo

Entonces la distribución de probabilidad de **X** es:

$$P(X=x) = f(x) = p(1-p)^{x-1}, \quad x = 1, 2, 3, \ldots$$

Demostración
Se obtiene directamente haciendo **k=1** en el modelo de la distribución binomial negativa.

5.5.1 MEDIA Y VARIANZA DE LA DISTRIBUCIÓN GEOMÉTRICA

Definición: Media y Varianza para la Distribución Geométrica

Media: $\mu = E[X] = \frac{1}{p}$, **Varianza:** $\sigma^2 = V[X] = \frac{1}{p}(\frac{1}{p} - 1)$

Ejemplo.
Calcule la probabilidad que en el quinto lanzamiento de tres monedas se obtengan tres sellos por primera vez.

Respuesta:
En el experimento de lanzar tres monedas hay 8 resultados posibles.
En cada ensayo la probabilidad que salgan tres sellos es constante e igual a 1/8 y la probabilidad que no salgan tres sellos es 7/8.

Estos ensayos son independientes, y por la pregunta concluimos que la variable de interés **X** tiene distribución geométrica con **p=1/8,**

Sea **X:** Cantidad de ensayos hasta obtener el primer "éxito" (variable aleatoria discreta)
x = 1, 2, 3, . . .
$P(X=x) = f(x) = (1/8)(7/8)^{x-1}$, **x=1, 2, 3,** ...
Por lo tanto
$P(X=5) = f(5) = (1/8)(7/8)^{5-1} = 0.0733$

5.6 DISTRIBUCIÓN HIPERGEOMÉTRICA

Esta distribución se refiere a los experimentos estadísticos que consisten en tomar una **muestra sin reemplazo**, de un conjunto finito el cual contiene algunos elementos considerados "éxitos" y los restantes son considerados "fracasos".

Tomar una **muestra sin reemplazo** significa que los elementos son tomados uno a uno, **sin devolverlos.** Podemos concluir entonces que los ensayos ya no pueden ser considerados independientes porque la probabilidad de "éxito" al tomar cada nuevo elemento es afectada por el resultado de los ensayos anteriores debido a que la cantidad de elementos de la población está cambiando.

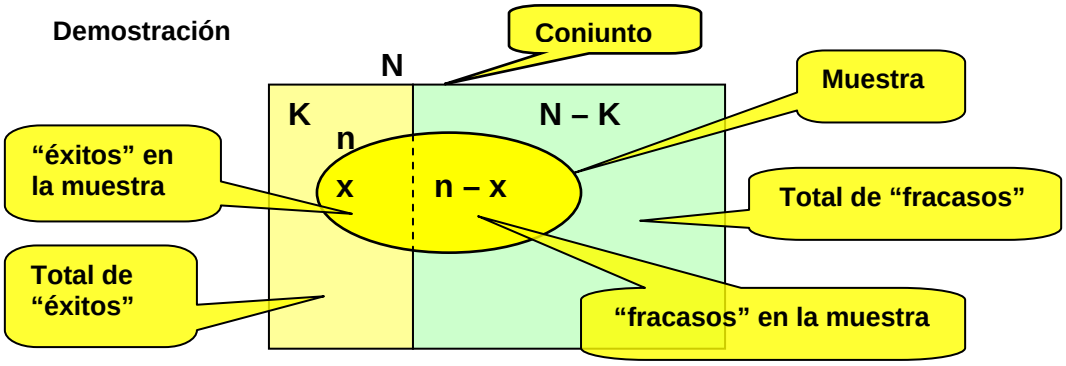
Definición: Distribución Hipergeométrica

Sean **N:** Cantidad de elementos del conjunto del que se toma la muestra
K: Cantidad de elementos existentes que se consideran "éxitos"
n: Tamaño de la muestra
X: Variable aleatoria discreta (es la cantidad de resultados considerados "éxitos" que se obtienen en la muestra)
x = 0, 1, 2, ..., n (son los valores que puede tomar **X**)

Entonces, la distribución de probabilidad de **X** es

$$f(x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}, \quad x = 0,1,2,\ldots,n$$

Demostración



Con referencia al gráfico:

$\binom{K}{x}$ es la cantidad total de formas de tomar **x** "éxitos" en la muestra de los **K** existentes

$\binom{N-K}{n-x}$ es la cantidad total de formas de tomar **n – x** "fracasos" de los **N – K** existentes.

$\binom{K}{x}\binom{N-K}{n-x}$ es la cantidad total de formas de tomar **x** "éxitos" y **n**–**x** "fracasos" en la muestra

$\binom{N}{n}$.cantidad total de formas de tomar la muestra de **n** elementos del conjunto de **N** elementos

Finalmente, mediante la asignación clásica de probabilidad a eventos obtenemos la fórmula para la distribución hipergeométrica. Esto completa la demostración

Se observa que **x** no puede exceder a **K**. La cantidad de **"éxitos"** que se obtienen en la muestra no puede exceder a la cantidad de **"éxitos"** disponibles en el conjunto. Igualmente, la cantidad de **n - x** **"fracasos"** no puede exceder a los **N - K** disponibles.

Ejemplo. Una caja contiene 9 baterías de las cuales 4 están en buen estado y las restantes defectuosas. Se toma una muestra eligiendo al azar tres baterías. Calcule la probabilidad que en la muestra se obtengan,

a) Ninguna batería en buen estado
b) Al menos una batería en buen estado
c) No mas de dos baterías en buen estado

Respuesta. Este es un experimento de **muestreo sin reemplazo**, por lo tanto es un experimento hipergeométrico con

N=9 (Total de elementos del conjunto)
K=4 (Total de elementos considerados 'éxitos')
n=3 (Tamaño de la muestra)
X: Cantidad de baterías en buen estado en la muestra (Variable aleatoria discreta)

Entonces la distribución de probabilidad de **X** es:

$$f(x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}} = \frac{\binom{4}{x}\binom{9-4}{3-x}}{\binom{9}{3}}, \quad x = 0,1,2,3$$

a) $P(X=0) = f(0) = \dfrac{\binom{4}{0}\binom{9-4}{3-0}}{\binom{9}{3}} = 0.119$

b) $P(X \geq 1) = 1 - P(X<1) = 1 - f(0) = 1 - 0.119 = 0.881$

c) $P(X \leq 2) = P(X=0) + P(X=1) + P(X=2) = f(0) + f(1) + f(2)$

$$= \frac{\binom{4}{0}\binom{9-4}{3-0}}{\binom{9}{3}} + \frac{\binom{4}{1}\binom{9-4}{3-1}}{\binom{9}{3}} + \frac{\binom{4}{2}\binom{9-4}{3-2}}{\binom{9}{3}} = 0.119 + 0.4762 + 0.3571 = 0.9523$$

También se puede calcular **c)** considerando que

$P(X \leq 2) = 1 - P(X>2) = 1 - f(3)$

5.6.1 MEDIA Y VARIANZA DE LA DISTRIBUCIÓN HIPERGEOMÉTRICA

Definición: Media y Varianza para la Distribución Hipergeométrica

Media: $\mu = E[X] = n\frac{K}{N}$, **Varianza:** $\sigma^2 = V[X] = \frac{nK}{N}(1-\frac{K}{N})(\frac{N-n}{N-1})$

Las demostraciones se las puede encontrar en textos de Estadística Matemática. En el desarrollo se usa la definición de valor esperado y las propiedades de las sumatorias.

Ejemplo. Calcule la media y la varianza para el ejemplo anterior

Respuesta:

$\mu = 3(4/9) = 1.333$ (es la cantidad promedio de baterías en buen estado que se obtienen al tomar muestras)

$\sigma^2 = \frac{3(4)}{9}(1-\frac{4}{9})(\frac{9-3}{9-1}) = 0.555$

5.6.2 APROXIMACIÓN DE LA DISTRIBUCIÓN IPERGEOMÉTRICA CON LA DISTRIBUCIÓN BINOMIAL

Si el tamaño de la muestra **n** es muy pequeño respecto a **N**, entonces se puede aceptar que la probabilidad de "éxito" en cada ensayo no cambia significativamente, es decir podemos considerar que los ensayos son "aproximadamente independientes".

Por ejemplo, si **N=1000** y **n=10**, y hay **200** elementos considerados "éxitos", entonces, la probabilidad de "éxito" del primer ensayo será **200/1000=0.2**, la probabilidad de "éxito" del segundo ensayo podrá ser **199/999=0.1992** o **200/999=0.2002**, dependiendo si el primer resultado fue o no "éxito". Ambos resultados son muy cercanos.

En esta situación, se puede considerar que el Modelo Hipergeométrico es **'aproximadamente binomial'** y se puede usar la fórmula de la Distribución Binomial con **p=K/N**

La bibliografía estadística establece que esta aproximación es aceptable si **n < 5% N**.

Sea **h:** Distribución Hipergeométrica
b: Distribución Binomial

Si **n<5%N**, entonces **h(x; N, K, n)** $\cong$ **b(x; n, K/N)**

5.6.3 EJERCICIOS

1) La probabilidad que una persona expuesta a cierta enfermedad la contraiga es 0.3.
Calcule la probabilidad que la quinta persona expuesta a esta enfermedad sea la segunda en contraerla.

2) Suponga que en dos de cada diez intentos, un vendedor realiza una venta.
Calcule la probabilidad que en el sexto intento realice la primera venta.

3) Suponga que la probabilidad de tener un hijo varón o mujer son iguales a 0.5.
Calcule la probabilidad que en una familia

a) El cuarto hijo sea el primer varón
b) El tercer hijo sea la segunda mujer
c) El quinto hijo sea el tercer varón o sea la cuarta mujer

4) Un caja de 10 alarmas contra robo contiene 4 defectuosas. Si se seleccionan al azar 3 de ellas y se envían a un cliente.
Calcule la probabilidad que el cliente reciba

a) Ninguna defectuosa;
b) No más de una defectuosa;
c) Al menos una defectuosa

5) Si probabilidad de que un estudiante en una escuela de conducción obtenga su licencia de conducir es 0.8, encuentre la probabilidad que uno de estos estudiantes apruebe el examen

a) En el segundo intento.
b) En el tercer intento.

MATLAB

Distribución binomial negativa

```
>> f =nbinpdf(6, 4, 0.3)
    f =
       0.0800
```

Probabilidad con la distrib. binomial negativa: **x=6, k=4, p=0.3**
x es el número de **"fracasos"** hasta obtener **k** "éxitos"

```
>> f=nbincdf(6, 4, 0.3)
    f =
       0.3504
```

Probabilidad con la distrib. binomial negativa **acumulada**
P(x≤6), k=4, p=0.3, x = 0, 1, 2, ..., 6

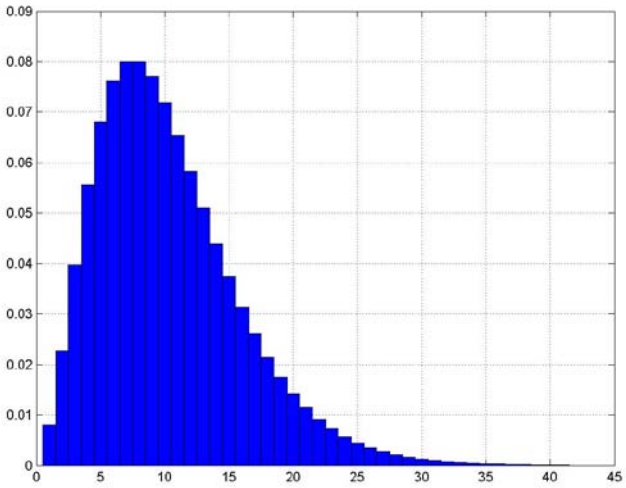
```
>> x = 0:40;
>> f = nbinpdf(x, 4, 0.3)
```

x=0, 1, 2, ..., 40
Distribución binomial negativa: **k=4, p=0.3, x=0, 1, 2, ..., 40**

```
f =  0.0081  0.0227  0.0397  0.0556  0.0681  0.0762  0.0800  0.0800  0.0770  0.0719  0.0654
     0.0583  0.0510  0.0439  0.0374  0.0314  0.0261  0.0215  0.0175  0.0142  0.0114  0.0092
     0.0073  0.0058  0.0045  0.0036  0.0028  0.0022  0.0017  0.0013  0.0010  0.0008  0.0006
     0.0004  0.0003  0.0003  0.0002  0.0001  0.0001  0.0001  0.0001
```

```
>> bar(f, 1, 'b'), grid on
```

Gráfico de la distribución binomial negativa, color azul



Distribución geométrica

```
>> f = geopdf(4, 1/8)
    f =
       0.0733
```

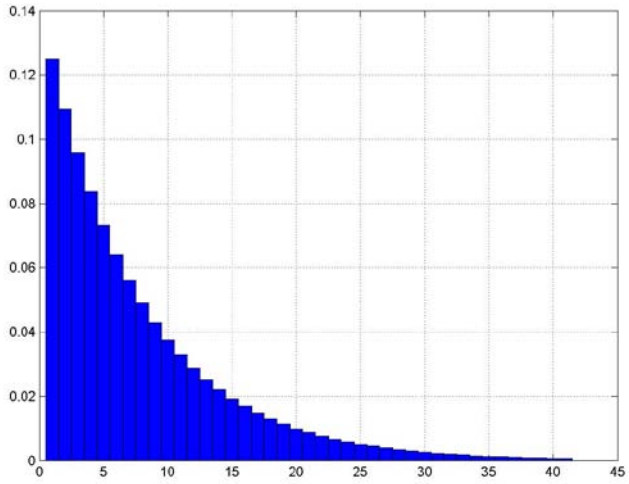
Probabilidad con la distribución geométrica: **x=4, p=1/8**
x es el número de fracasos hasta obtener el primer "éxito"

```
>> x = 0:40;
>> f = geopdf(x, 1/8)
```

x=0, 1, 2, ..., 40
Distribución geométrica: **p=1/8, x=0, 1, 2, ..., 40**

```
f =  0.1250  0.1094  0.0957  0.0837  0.0733  0.0641  0.0561  0.0491  0.0430  0.0376  0.0329
     0.0288  0.0252  0.0220  0.0193  0.0169  0.0148  0.0129  0.0113  0.0099  0.0087  0.0076
     0.0066  0.0058  0.0051  0.0044  0.0039  0.0034  0.0030  0.0026  0.0023  0.0020  0.0017
     0.0015  0.0013  0.0012  0.0010  0.0009  0.0008  0.0007  0.0006
```

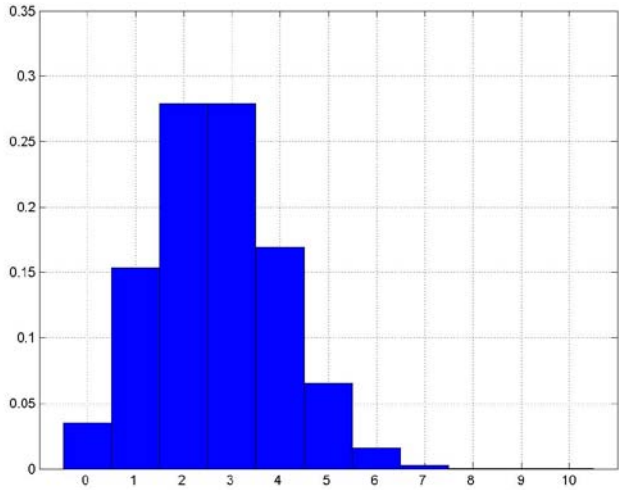
>> bar(f,1,'b'), grid on Gráfico de la distribución geométrica, en color azul



Distribución hipergeométrica

>> f = hygepdf(0, 9, 4, 3) Distribución hipergeométrica **x=0, N=9, K=4, n=3**
f = Cálculo de **P(X = 0)**
0.1190
>> f = hygecdf(2, 9, 4, 3) Distribución hipergeométrica **acumulada**
f = **P(X≤2), N=9, K=4, n=3, x=0, 1, 2**
0.9524
>> [mu, var]=hygestat(9, 4, 3) Media y varianza de la distr. hipergeométrica: **N=9, K=4, n=3**
mu = 1.3333
var = 0.5556
>> x = 0:10;
>> f = hygepdf(x, 75, 20, 10)
f = 0.0353 0.1534 0.2791 0.2791 0.1694 0.0651
0.0159 0.0025 0.0002 0.0000 0.0000

>> bar(f, 1, 'b'),grid on Gráfico de la distribución hipergeométrica, en color azul



>> f = hygepdf(6, 1000, 200, 10) Distrib. hipergeométrica **x=6, N=1000, K=200, n=10**
f =
0.0053
>> f = binopdf(6, 10, 200/1000) Distribución binomial **x=6, n=10, p=K/N**
f =
0.0055 Los resultados son cercanos pues **n < 5%N**

5.7 DISTRIBUCIÓN DE POISSON

La distribución de Poisson es un modelo que puede usarse para calcular la probabilidad correspondiente al número de "éxitos" que se obtendrían en una región o en intervalo de tiempo especificados, si se conoce el número promedio de "éxitos" que ocurren.

Este modelo requiere que se cumplan las siguientes suposiciones:

a) El número de "éxitos" que ocurren en la región o intervalo es independiente de lo que ocurre en otra región o intervalo
b) La probabilidad de que un resultado ocurra en una región o intervalo muy pequeño, es igual para todos los intervalos o regiones de igual tamaño y es proporcional al tamaño de la región o intervalo.
c) La probabilidad de que más de un resultado ocurra en una región o intervalo muy pequeño no es significativa.

Algunas situaciones que se pueden analizar con este modelo:

Número de defectos por unidad de área en piezas similares de un material.
Número de personas que llegan a una estación en un intervalo de tiempo especificado.
Número de errores de transmisión de datos en un intervalo de tiempo dado.
Número de llamadas telefónicas que entran a una central por minuto.
Número de accidentes automovilísticos producidos en una intersección, en una semana.

Definición: Distribución de Poisson

Sea **X**: Variable aleatoria discreta con distribución de Poisson
(cantidad de "éxitos" en una región o intervalo especificados)
x = 0, 1, 2, . . . (valores posibles para la variable **X**)
λ: Cantidad promedio de "éxitos" en la región o intervalo especificados
Entonces la distribución de probabilidad de **X** es:

$$\mathbf{f(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}, \; x=0, 1, 2, ...,} \qquad \mathbf{e = 2.71828...}$$

Ejemplo.
La cantidad de errores de transmisión de datos en una hora es 5 en promedio. Suponiendo que es una variable con distribución de Poisson, determine la probabilidad que:

a) En cualquier hora ocurra solamente 1 error.
b) En cualquier hora ocurran al menos 3 errores
c) En dos horas cualesquiera ocurran no más de 2 errores.

Respuesta:

Sea **X**: Variable aleatoria discreta (cantidad de errores por hora)
λ = 5 (promedio de errores de transmisión en 1 hora)

a) $\mathbf{P(X=1)} = f(1) = \frac{e^{-5}5^{1}}{1!} = 0.0337$

b) $\mathbf{P(X\geq 3)} = 1 - P(X\leq 2) = 1 - (f(0) + f(1) + f(2)) = 1 - 0.1247 = 0.8743$

c) Sea **X**: variable aleatoria discreta (cantidad de errores en 2 horas)
λ = 10 (promedio de errores de transmisión en dos horas)

$$\mathbf{P(X\leq 2)} = f(0) + f(1) + f(2) = \frac{e^{-10}10^{0}}{0!} + \frac{e^{-10}10^{1}}{1!} + \frac{e^{-10}10^{2}}{2!} = 0.0028$$

5.7.1 MEDIA Y VARIANZA DE LA DISTRIBUCIÓN DE POISSON

Definición: Media y Varianza de la Distribución de Poisson

Media: $\mu = E[X] = \lambda$, **Varianza:** $V[X] = \lambda$

Demostración

Primero se obtiene la función generadora de momentos de la Distribución de Poisson.

$$m(t) = E[e^{tX}] = \sum_{x=0}^{\infty} e^{tx} f(x) = \sum_{x=0}^{\infty} e^{tx} \frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} e^{tx} \frac{\lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{(e^t\lambda)^x}{x!}$$

Se tiene el desarrollo de la función exponencial:

$$e^y = 1 + y + \frac{y^2}{2!} + \frac{y^3}{3!} + \ldots$$

Haciendo $y = e^t\lambda$ se obtiene

Definición: Función Generadora de Momentos de la Distribución de Poisson

$$m(t) = e^{-\lambda} e^{e^t\lambda}$$

Entonces con la definición conocida:

$$\mu = \mu'_1 = \frac{d}{dt} m(t)\,|_{t=0} = \frac{d}{dt} e^{-\lambda} e^{e^t\lambda}\,|_{t=0} = e^{-\lambda} e^{e^t\lambda} e^t\lambda\,|_{t=0} = \lambda$$

Con esto se completa la demostración

La demostración de la varianza sigue un camino similar.

5.7.2 APROXIMACIÓN DE LA DISTRIBUCIÓN BINOMIAL CON LA DISTRIBUCIÓN DE POISSON

En la Distribución Binomial cuando **n** es grande no es práctico el uso de la fórmula. Para entender esto, suponga que **n=100, p=0.05** y se quiere calcular la probabilidad que la variable aleatoria **X** tome el valor 4:

$$P(X = 4) = f(4) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{100}{4} 0.05^4\, 0.95^{100-4}$$

En esta situación se puede calcular aproximadamente la probabilidad mediante otro modelo, en este caso, con la Distribución de Poisson.

Del desarrollo algebraico, que lo omitimos, se obtiene el siguiente resultado para la Distribución Binomial:

$$f(x; n, p) \to \frac{e^{-\lambda}\lambda^x}{x!}, \; x = 0, 1, 2, \ldots, ,$$

cuando $n \to \infty$ y $p \to 0$.

Este modelo corresponde a la distribución de Poisson, siendo $\lambda = np$

Las referencias bibliográficas indican que esta aproximación es aceptable si $n \geq 20$ y $p \leq 0.05$.

Otro criterio utilizado establece que la aproximación es aceptable si $n \geq 100$ y $np \leq 10$

Ejemplo.
Calcular con la Distribución Binomial **x=4, n=100, p=0.05**.

$$P(X=4) = f(4) = \binom{100}{4} 0.05^4\ 0.95^{100-4} = \frac{100!}{4!\ 96!} 0.05^4 0.95^{96} = 0.1781$$

Calcular un valor aproximado con la Distribución de Poisson **x=4, $\lambda = np = 100*0.05 = 5$**

$$P(X=4) = f(4) \cong \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-5}5^4}{4!} = 0.1755$$

Valor cercano al resultado anterior pues **$n \geq 20$** y **$p \leq 0.05$**

5.7.3 EJERCICIOS

1) Cierto tipo de tela usada en tapicería tiene, en promedio, dos defectos por metro cuadrado. Si se supone una distribución de Poisson, calcule la probabilidad que

a) Un rollo de 30 m^2 tenga no más de 5 defectos
b) Un rollo de 30 m^2 tenga al menos 6 defectos
c) Un rollo de 60 m^2 tenga exactamente 10 defectos

2) Un cargamento grande de libros contiene 3% de ellos con encuadernación defectuosa. Utilice la aproximación de Poisson para determinar la probabilidad que entre 400 libros seleccionados al azar del cargamento,

a) Exactamente 10 libros estén defectuosos
b) Al menos 10 tengan defectos

3) Un bar prepara un batido especial que contiene en promedio 4 frutas diferentes, encuentre la probabilidad de que el batido contenga más de 4 frutas:
a) En un determinado día, b)En tres de los siguientes 5 días,

MATLAB

Probabilidad con la Distribución de Poisson

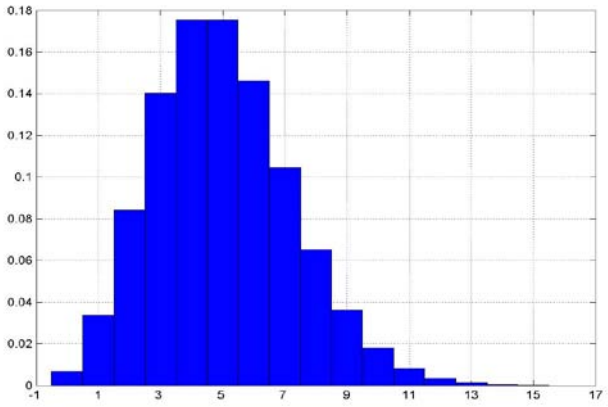
```
>> f=poisspdf(1,5)
   f =
       0.0337
```
Probabilidad con la distribución de Poisson, **x=1**, **λ=5**

```
>> f=poisscdf(2,5)
   f =
       0.1247
```
Probabilidad con la distribución de Poisson **acumulada** **P(X≤5), λ=5**

```
>> x=0:15;
```
x = 0, 1, 2, . . ., 15

```
>> f=poisspdf(x,5)
   f =  0.0067  0.0337  0.0842  0.1404  0.1755  0.1755  0.1462  0.1044
        0.0653  0.0363  0.0181  0.0082  0.0034  0.0013  0.0005  0.0002
```
Probabilidad con la distribución de Poisson, **λ=5, x=0, 1, 2, ...15**

```
>> bar(f,1,'b')
```
Gráfico de la distribución de Poisson **λ=5, x=0, 1, 2, ...15**



6 VARIABLES ALEATORIAS CONTINUAS

Las variables aleatorias continuas definen reglas de correspondencia entre los resultados obtenidos en experimentos cuyos valores se miden en una escala continua y el conjunto de los números reales.

6.1 FUNCIÓN DE DENSIDAD DE PROBABILIDAD

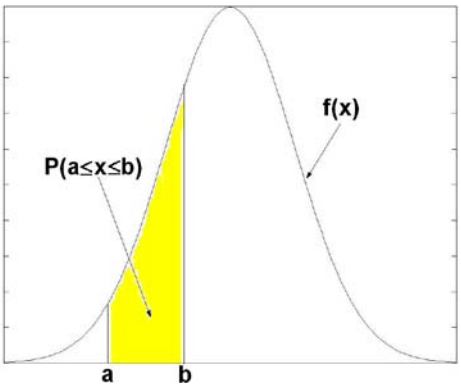
La probabilidad de una variable aleatoria continua puede especificarse si existe una función denominada **función de densidad de probabilidad (**o simplemente **función de densidad)**, tal que el área debajo del gráfico de esta función cumpla los requisitos para que sea una medida del valor de probabilidad. Para variables aleatorias discretas, la probabilidad se obtiene de la sumatoria de **f(x).** En el límite, esta sumatoria se transforma en un integral.

Definición: Función de Densidad de Probabilidad

Sea **X** una variable aleatoria continua.
Se dice que **f** es una **función de densidad de probabilidad** si y solo si,

$$P(a \leq X \leq b) = \int_a^b f(x)dx \,, \qquad \text{siendo } a,b \in \Re$$

Representación gráfica



Cada función de densidad de probabilidad debe cumplir las siguientes propiedades:

Definición: Propiedades de una Función de Densidad de Probabilidad

1) $f(x) \geq 0, \quad -\infty < x < +\infty$ — **f(x)** no puede tomar valores negativos

2) $\int_{-\infty}^{+\infty} f(x)dx = 1$ — El área total debajo de **f(x)** debe ser igual a **1**

Esta propiedad implica que la probabilidad para variables aleatorias continuas solamente puede calcularse para intervalos de la variable. La probabilidad que la variable aleatoria tome un valor real específico es cero. Este resultado debe entenderse de la siguiente definición:

$$\lim_{a \to b} P(a \leq X \leq b) = P(b \leq X \leq b) = P(X = b) = \int_b^b f(x)dx = 0$$

Por lo tanto, en el cálculo de probabilidad para variables aleatorias continuas, es igual incluir o no incluir los extremos del intervalo:

$$P(a \leq X \leq b) = P(a < X < b)$$

.

Ejemplo
Suponga que el tiempo de atención de cada cliente en una estación de servicio es una variable aleatoria continua con la siguiente función de densidad de probabilidad:

$$f(x) = \begin{cases} \frac{2}{5}(x+2), & 0 \le x \le 1 \\ 0, & \text{otro } x \end{cases}$$

a) Verifique que cumple las propiedades de una función de densidad

Sea **X**: variable aleatoria continua (duración en horas)

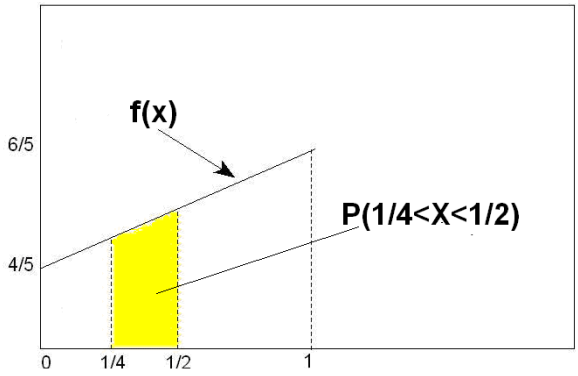
1) $f(x) \ge 0$, $-\infty < x < +\infty$: evidente para **f(x)** especificada

2) $\int_{-\infty}^{+\infty} f(x)dx = 1$: $\int_0^1 \frac{2}{5}(x+2)dx = \frac{2}{5}(\frac{x^2}{2}+2x)\Big|_0^1 = 1$

b) Calcule la probabilidad que el tiempo de atención esté entre 15 y 30 minutos

$$P(1/4<X<1/2) = \int_{1/4}^{1/2} \frac{2}{5}(x+2)dx = \frac{2}{5}(\frac{x^2}{2}+2x)\Big|_{1/4}^{1/2} = 19/80 = 0.2375$$

Representación gráfica



6.2 FUNCIÓN DE DISTRIBUCIÓN

Al igual que en el caso discreto se puede definir una función de probabilidad acumulada, la cual en el caso continuo se denomina función de distribución

Definición: Función de Distribución

Sea **X** una variable aleatoria contínua con función de densidad **f(x)**
Entonces, la función

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(t)dt, \quad \text{para } -\infty < x < +\infty$$

se denomina **función de distribución** de la variable aleatoria **X**

Definición: Propiedades de la Función de Distribución

1) $\frac{d}{dx}F(x) = f(x)$ La derivada de la función de distribución es la densidad
2) $a < b \Rightarrow F(a) < F(b)$, F es una función creciente
3) $P(a \le x \le b) = F(b) - F(a)$

La propiedad 3) es útil para calcular valores de probabilidad de la variable X

Ejemplo.
Encuentre la función de distribución para el ejemplo anterior

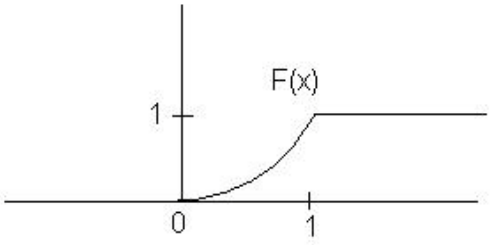
Respuesta

$$F(x) = \int_{-\infty}^{x} f(t)dt = \int_{0}^{x} \frac{2}{5}(t+2)dt = \frac{2}{5}\left(\frac{t^2}{2}+2t\right)\Big|_0^x = \frac{2}{5}\left(\frac{x^2}{2}+2x\right)$$

Esta es una función cuyo dominio es el conjunto de los números reales::

$$F(x) = \begin{cases} 0, & x < 0 \\ \frac{2}{5}\left(\frac{x^2}{2}+2x\right), & 0 \le x < 1 \\ 1, & x \ge 1 \end{cases}$$

Gráfico de la función de distribución



Use la Función de Distribución para calcular **P(1/4<X<1/2)** en el ejemplo anterior

Respuesta

$$P(1/4<X<1/2) = F(1/2) - F(1/4) = \frac{2}{5}\left(\frac{(1/2)^2}{2}+2(1/2)\right) - \frac{2}{5}\left(\frac{(1/4)^2}{2}+2(1/4)\right) = 19/80$$

6.2.1 EJERCICIOS

1) La densidad de probabilidad de una variable aleatoria **X** está dada por

$$f(x) = \begin{cases} 630x^4(1-x)^4, 0 < x < 1 \\ 0, \quad \text{otro } x \end{cases}$$

a) Verifique que satisface las propiedades de una función de densidad
c) Calcule la probabilidad que **X** tenga un valor mayor a 0.75.
e) Determine la probabilidad que **X** tome un valor dentro del intervalo de dos desviaciones estándar alrededor de la media y compare con el valor proporcionado por el Teorema de Chebyshev.

2) El tiempo que tardan en atender a un individuo en una cafetería es una variable aleatoria con densidad de probabilidad

$$f(x) = \begin{cases} 0.25e^{-0.25X}, x > 0 \\ 0, \quad \text{otro } x \end{cases}, \quad \mathbf{x} \text{ en minutos}$$

Calcule la probabilidad que el tiempo que tardan en atenderlo sea más de 5 minutos

MATLAB

Probabilidad con variables aleatorias continuas

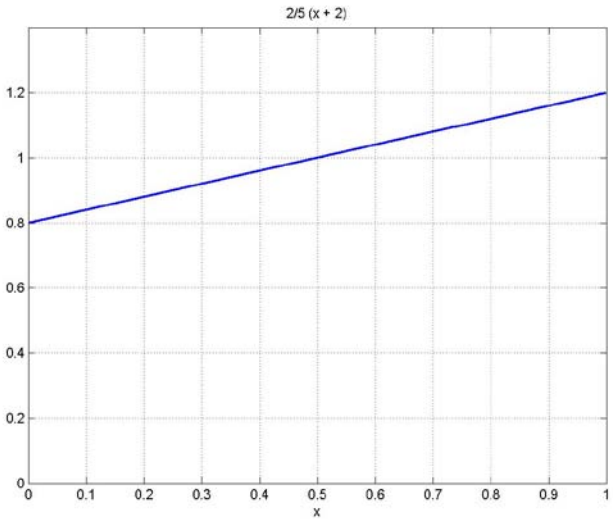
```
>> syms x
```
Para manejo simbólico de la variable **x**

```
>> f = 2/5*(x + 2);
```
Definición de una función de densidad

```
>> p = int(f, 1/4, 1/2)
```
Cálculo de la probabilidad **P(1/4 < X < 1/2)**

```
    p =
        19/80
```

```
>> ezplot(f, 0, 1), grid on
```
Gráfico de la función de densidad



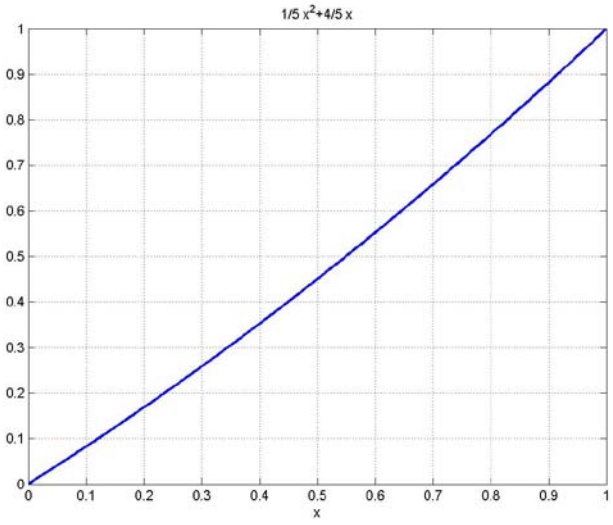

```
>> F = int(f)
```
Obtención de la función de distribución

```
    F =
        1/5*x^2+4/5*x
```

```
>> p=eval(subs(F,'1/2')) - eval(subs(F,'1/4'))
```
Cálculo de la probabilidad **P(1/4 < X < 1/2)** con la función de distribución: **F(1/2) – F(1/4)**

```
    p =
        19/80
```

```
>> ezplot(F, 0, 1), grid on
```
Gráfico de la función de distribución



6.3 MEDIA Y VARIANZA DE VARIABLES ALEATORIAS CONTINUAS

Definición: Media y Varianza de Variables Aleatorias Continuas

Sean **X:** Variable aleatoria continua
f(x): Función de densidad de probabilidad

Media de **X:** $\mu = E(X) = \int_{-\infty}^{+\infty} x f(x) dx$

Varianza de **X:** $\sigma^2 = V(X) = E[(X - \mu)^2] = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx$

Ejemplo

Calcule la media y la varianza para el ejemplo de la estación de servicio en donde **X** es una variable aleatoria continua que representa tiempo de atención en horas, siendo su densidad de probabilidad:

$$f(x) = \begin{cases} \frac{2}{5}(x+2), & 0 \le x \le 1 \\ 0, & \text{otro } x \end{cases}$$

Respuesta:

$$\mu = E(X) = \int_0^1 x \frac{2}{5}(x+2)dx = \frac{2}{5}\left[\frac{x^3}{3} + x^2\right]\Big|_0^1 = 8/15 = 0.533$$

Es el tiempo de atención promedio para los clientes

$$\sigma^2 = V(X) = E[(X-\mu)^2] = E(X^2) - \mu^2 = \int_0^1 x^2 \frac{2}{5}(x+2)dx - (8/15)^2 = 0.0822$$

6.3.1 PROPIEDADES DE LA MEDIA Y LA VARIANZA

Definiciones: Propiedades de la Media y la Varianza

Sea **X:** una variable aleatoria continua con densidad de probabilidad **f(x)**
a, b $\in \Re$

Media: $E(aX + b) = aE(X) + b$
Corolarios: $E(aX) = aE(X)$; $E(b) = b$

Varianza: $V(aX + b) = a^2V(X)$
Corolarios: $V(aX) = a^2V(X)$; $V(b) = 0$

Las demostraciones y los corolarios son similares al caso de las variables aleatorias discretas, pero usando integrales en lugar de sumas.

6.3.2 VALOR ESPERADO DE EXPRESIONES CON UNA VARIABLE ALEATORIA CONTINUA

Estas expresiones también son variables aleatorias y su dominio generalmente es el mismo que el dominio de la variable aleatoria original. El rango puede ser diferente.

Definición: Valor Esperado de Expresiones con una Variable Aleatoria Continua

Sea **X:** Variable aleatoria continua
f(x): Densidad de probabilidad de**X**
G(X): Alguna expresión con la variable aleatoria X
Entonces

$$\mu_{G(X)} = E[G(X)] = \int_{-\infty}^{+\infty} G(x)f(x)dx\,,$$ es la media o valor esperado de **G(X)**

Ejemplo
Suponga que en ejemplo de la estación de servicio, el costo de atención a cada cliente está dado por la siguiente variable aleatoria:
G(X) = 10 + 5X en dólares
Calcule la media del costo de atención

Respuesta
E[G(X)] = E(10 + 5X) = 10 + 5E(X) = 10 + 5(8/15) = 12.667 dólares

6.4 MOMENTOS Y FUNCIÓN GENERADORA DE MOMENTOS PARA VARIABLES ALEATORIAS CONTÍNUAS

Las definiciones que fueron establecidas para las variables aleatorias discretas se extienden al caso discreto sustituyendo sumatorias por integrales

Definiciones: Momentos y Funciones Generadoras de Momentos

Sean **X:** Variable aleatoria continua
f(x): Densidad de probabilidad

r-ésimo momento de X alrededor del origen

$$\mu'_r = E(X^r) = \int_{-\infty}^{+\infty} x^r f(x)dx$$

r-ésimo momento de X alrededor de la media, o r-ésimo momento central

$$\mu_r = E[(X-\mu)^r] = \int_{-\infty}^{+\infty} (x-\mu)^r f(x)dx$$

Función generadora de momentos

$$M(t) = E(e^{tX}) = \int_{-\infty}^{+\infty} e^{tx} f(x)dx$$

Obtención de momentos alrededor del origen

$$\mu'_r = \frac{d^r}{dt^r} M(t)\,|_{t=0}$$

6.5 TEOREMA DE CHEBYSHEV

El Teorema de Chebyshev es aplicable también a variables aleatorias contínuas.
La demostración usa integrales en lugar de sumatorias

Definición: Teorema de Chebyshev (Variables Aleatorias Continuas)

Sea **X** una variable aleatoria continua con media μ y varianza σ^2, entonces la probabilidad que **X** tome algún valor que no se desvíe de su media μ en más de $k\sigma$, es al menos $1 - 1/k^2$:

$$P(\mu - k\sigma < x < \mu - k\sigma) \geq 1 - 1/k^2 \ , \ k \in \Re^+$$

6.6 EJERCICIOS

1) La densidad de probabilidad de una variable aleatoria **X** está dada por

$$f(x) = \begin{cases} 630x^4(1-x)^4, 0 < x < 1 \\ 0, \quad \text{otro } x \end{cases}$$

a) Calcule la media y varianza de **X**
b) Calcule la media y varianza de la variable **Y=2X+1**.

2) El tiempo que tardan en atender a una persona en una cafetería es una variable aleatoria con densidad de probabilidad

$$f(x) = \begin{cases} 0.25e^{-0.25x}, x > 0 \\ 0, \quad \text{otro } x \end{cases}, \quad \mathbf{x} \text{ en minutos}$$

Calcule la media y varianza de **X**

3) Demuestre que si **X** es una variable aleatoria con media μ tal que f(x)=0, para x<0, entonces para una constante positiva **k** cualquiera, se tiene:

$$P(x \geq k) \leq \frac{\mu}{k}$$

Esta desigualdad se conoce como **desigualdad de Markov** y es utilizada también para acotar el valor de probabilidad de una variable aleatoria.

MATLAB

Media y varianza de variables aleatorias continuas

```
>> syms x                                  Definir X para manejo simbólico
>> f = 2/5*(x + 2);                        Función de densidad de X

>> mu = int(x*f, 0,1)                      Media de X
   mu =
        8/15

>> sigma2 = int(x^2*f,0,1)-mu^2            Varianza de X
   sigma2 =
            37/450

>> sigma = eval(sqrt(sigma2))              Desviación estándar
  sigma =
           0.5355
```

7 DISTRIBUCIONES DE PROBABILIDAD CONTINUAS

En este capítulo se estudian los modelos matemáticos para calcular la probabilidad en algunos problemas típicos en los que intervienen variables aleatorias continuas.

El objetivo es obtener una fórmula matemática f(x) para determinar los valores de probabilidad de la variable aleatoria X.

7.1 DISTRIBUCIÓN UNIFORME CONTINUA

Este modelo corresponde a una variable aleatoria continua cuyos valores tienen igual valor de probabilidad en un intervalo especificado para la variable

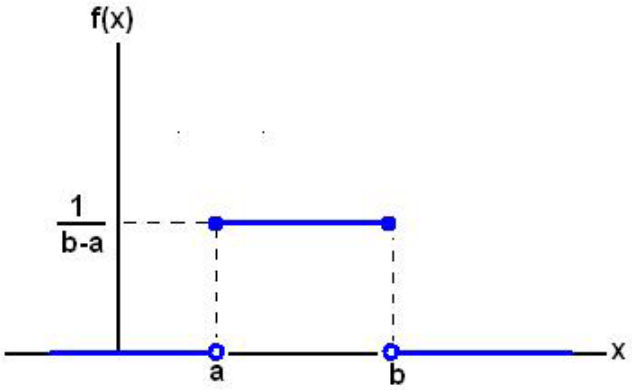
Definición: Distribución Uniforme Continua

Sea **X:** Variable aleatoria continua.
X tiene distribución Uniforme si su densidad de probabilidad está dada por,

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & a \le x \le b \\ 0, & \text{para otro } x \end{cases}$$

a, b son los parámetros para este modelo

Representación gráfica de la distribución Uniforme Continua



Se puede observar que f(x) cumple las propiedades de las funciones de densidad

7.1.1 MEDIA Y VARIANZA DE LA DISTRIBUCIÓN UNIFORME CONTINUA

Definición: Media y Varianza de la Distribución Uniforme Continua

Sea **X**: Variable aleatoria con distribución Uniforme Continua

Media: $\mu = E(X) = \dfrac{1}{2}(a+b)$

Varianza: $\sigma^2 = V(X) = \dfrac{1}{12}(b-a)^2$

Se obtienen directamente de las definiciones respectivas

Demostración para la media

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx = \int_{a}^{b} x\frac{1}{b-a}dx = \frac{1}{b-a}\left[\frac{1}{2}\left(b^2 - a^2\right)\right] = \frac{1}{2}(a+b)$$

7.1.2 FUNCIÓN DE DISTRIBUCIÓN DE PROBABILIDAD

De acuerdo a la definición establecida:

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(t)dt \text{ , para } -\infty < x < +\infty$$

Para la Distribución Uniforme Continua:

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(t)dt = \int_{a}^{x} \frac{1}{b-a}dx = \frac{x-a}{b-a} \Rightarrow F(x) = \begin{cases} 0, & x < a \\ \dfrac{x-a}{b-a}, & a \le x < b \\ 1, & x \ge b \end{cases}$$

Ejemplo

Cuando falla cierto componente de una máquina, esta debe detenerse hasta que sea reparado. Suponiendo que el tiempo de reparación puede tomar cualquier valor entre 1 y 5 horas.

a) Calcule la probabilidad que la duración tome al menos 2 horas

Solución

X: Variable aleatoria Continua (duración de la reparación)
Tiene distribución Uniforme, por lo tanto, su función de densidad es

$$f(x) = \frac{1}{b-a} = \frac{1}{5-1} = 1/4 \text{ , } 1 \le x \le 5$$

$$P(X \ge 2) = \int_{2}^{5} \frac{1}{4}dx = 3/4 = 75\%$$

b) Calcule el valor esperado de la duración de la reparación

Solución

$$E(X) = \frac{1}{2}(a+b) = \frac{1}{2}(1+5) = 3 \text{ horas}$$

b) Suponga que la reparación tiene un costo fijo de $100 y un costo variable de $10, el cual se incrementa cuadráticamente dependiendo de la duración. Calcule el valor esperado del costo de la reparación.

Solución

C: Costo de la reparación (es una variable aleatoria continua)

$$C = 100 + 10\,x^2$$

$$E(C) = E(100 + 10\,x^2) = 100 + 10\,E(X^2)$$

$$E(x^2) = \int_{1}^{5} x^2 \frac{1}{4}dx = \frac{1}{4}\left[\frac{x^3}{3}\right]_1^5 = 31/3$$

$$E(C) = 100 + 10(31/3) = \$203.3$$

7.1.3 EJERCICIOS

1) Se elige un punto **C** sobre una recta **AB** cuya longitud es **k**. Si la distancia entre **C** y **A** es una variable aleatoria **X** con distribución uniforme continua, calcule la probabilidad que la diferencia de longitud entre los segmentos **AC** y **BC** no exceda en mas de 10% de **k.**

2) En un negocio de hamburguesas se despacha el refresco en vasos. La cantidad es una variables aleatoria con una distribución uniforme entre 130 y 160 ml. (mililitros)

a) Calcule la probabilidad de obtener un vaso que contenga a lo más 140 ml.
b) ¿Cuántos ml. contiene en promedio un vaso?
c) Obtenga la varianza para la variable aleatoria

3) Una resistencia eléctrica se comporta de acuerdo a una distribución continua con valores entre 900 y 1100 ohms. Encuentre la probabilidad que la resistencia,

a) Aguante a lo más 950 ohms antes de quemarse
b) Tenga un valor entre 950 y 1050 ohms.

7.2 DISTRIBUCIÓN NORMAL

La Distribución Normal es la piedra angular de la teoría estadística moderna. Conocida y estudiada desde hace mucho tiempo, es utilizada para describir el comportamiento aleatorio de muchos procesos que ocurren en la naturaleza y también realizados por los humanos.

Definición: Función de Densidad de la Distribución Normal

Sea **X:** Variable aleatoria continua con media μ y varianza σ^2
X tiene distribución Normal si su función de densidad es:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < +\infty$$

Se puede demostrar que **f** cumple las propiedades de una función de densidad:

1) $f(x) \geq 0$, $-\infty<x<+\infty$:

2) $\int_{-\infty}^{+\infty} f(x)dx = 1$

La gráfica de **f** es similar al perfil del corte vertical de una campana y tiene las siguientes características:

1) Es simétrica alrededor de μ
2) Su asíntota es el eje horizontal
3) Sus puntos de inflexión están ubicados en $\mu - \sigma$ y $\mu + \sigma$

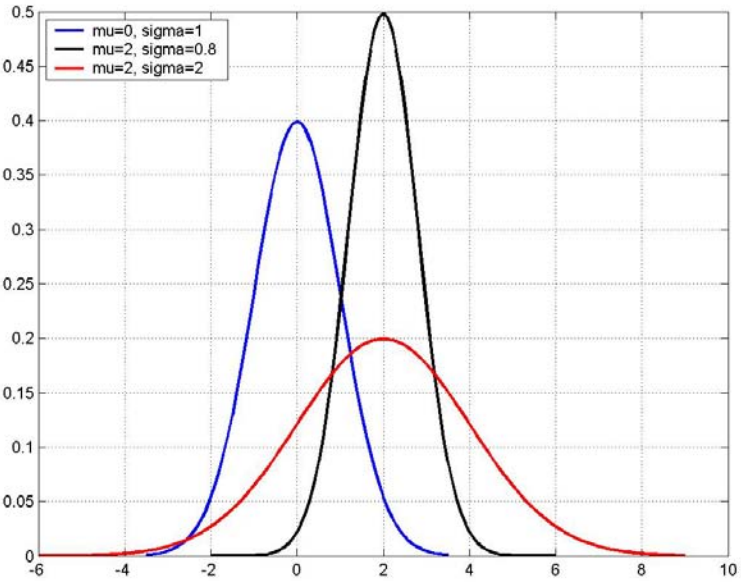


Gráfico de la distribución Normal para varios valores de μ *y* σ

Para calcular probabilidad se tiene la definición

$$P(a \leq X \leq b) = \int_a^b f(x)dx, \quad \text{siendo } a,b \in \Re$$

También se puede usar la definición de distribución acumulada o función de distribución:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt, \quad \text{para } -\infty < x < +\infty$$

Esta definición es útil para calcular probabilidad con la propiedad: $P(a \leq X \leq b) = F(b) - F(a)$

7.2.1 DISTRIBUCIÓN NORMAL ESTÁNDAR

Para generalizar y facilitar el cálculo de probabilidad con la distribución Normal, es conveniente definir la **Distribución Normal Estándar** que se obtiene haciendo $\mu = 0$, y $\sigma^2 = 1$ en la función de densidad de la Distribución Normal

Definición: Función de densidad de la distribución Normal Estándar

Sea **Z:** Variable aleatoria continua con media $\mu = 0$ y varianza $\sigma^2 = 1$
Z tiene distribución Normal Estándar si su función de densidad es:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \quad -\infty < z < +\infty$$

Para calcular probabilidad con la distribución Normal Estándar se puede usar la definición de la **distribución acumulada** o **función de distribución:**

$$F(z) = P(Z \le z) = \int_{-\infty}^{z} f(t)dt = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt, \quad -\infty < z < +\infty$$

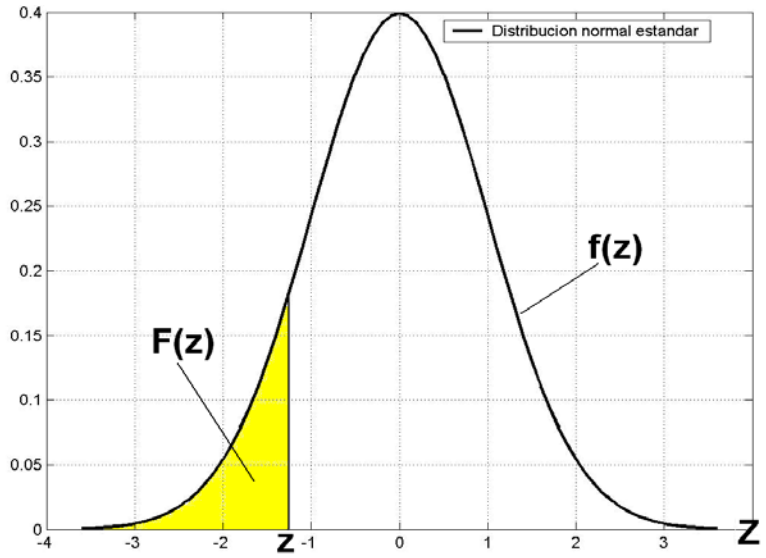


Gráfico de la distribución Normal Estándar

Para el cálculo manual se pueden usar tablas con valores de **F(z)** para algunos valores de **z**

En un anexo se incluye una **Tabla de la Distribución Normal Estándar**. Esta tabla contiene los valores de **F(z)** con 6 decimales para valores de **z** en el intervalo de **–3.59** a **3.59** con incrementos de **0.01**. Los valores de **F(z)** fuera de este intervalo ya no son significativamente diferentes.

Para aplicaciones comunes es suficiente usar sólo los cuatro primeros decimales de **F(z)** redondeando el último dígito.

Algunas tablas de la distribución Normal Estándar no incluyen valores de **F(z)** para valores negativos de **z**, por lo cual y por la simetría de **f(z)**, se puede usar la siguiente relación:

$$F(-z) = P(Z \le -z) = P(Z \ge z) = 1 - P(Z \le z) = 1 - F(z) \Rightarrow F(-z) = 1 - F(z)$$

Ejemplos

Usando la Tabla de la Distribución Normal Estándar calcule:

a) P(Z ≤ 1.45)

P(Z ≤ 1.45) = F(1.45) = 0.9265

El resultado se toma directamente de la Tabla de la Distribución Normal Estándar:

Z	0.00	0.01	0.02	0.03	0.04	**0.05**	0.06	0.07	0.08	0.09
0.0	0.500000	0.503989	0.507978	0.511967	0.515953	0.519939	0.532922	0.527903	0.531881	0.535856
0.1	0.539828	0.543795	0.547758	0.551717	0.555760	0.559618	0.563559	0.567495	0.571424	0.575345
0.2	0.579260	0.583166	0.587064	0.590954	0.594835	0.598706	0.602568	0.606420	0.610261	0.614092
0.3	0.617911	0.621719	0.625516	0.629300	0.633072	0.636831	0.640576	0.644309	0.648027	0.651732
0.4	0.655422	0.659097	0.662757	0.666402	0.670031	0.673645	0.677242	0.680822	0.684386	0.687933
0.5	0.691462	0.694974	0.698468	0.701944	0.705401	0.708840	0.712260	0.715661	0.719043	0.722405
0.6	0.725747	0.729069	0.732371	0.735653	0.738914	0.742154	0.745373	0.748571	0.751748	0.754903
0.7	0.758036	0.761148	0.764238	0.767305	0.770350	0.773373	0.776373	0.779350	0.782305	0.785236
0.8	0.788145	0.791030	0.793892	0.796731	0.799546	0.802338	0.805106	0.807850	0.810570	0.813267
0.9	0.815940	0.818589	0.821214	0.823815	0.826391	0.828944	0.831472	0.833977	0.836457	0.838913
1.0	0.841345	0.843752	0.846136	0.848495	0.850830	0.853141	0.855428	0.857690	0.859929	0.862143
1.1	0.864334	0.866500	0.868643	0.870762	0.872857	0.874928	0.876976	0.878999	0.881000	0.882977
1.2	0.884930	0.886860	0.888767	0.890651	0.892512	0.894350	0.896165	0.897958	0.899727	0.901475
1.3	0.903199	0.904902	0.906582	0.908241	0.909877	0.911492	0.913085	0.914657	0.916207	0.917736
1.4	0.919243	0.920730	0.922196	0.923641	0.925066	**0.926471**	0.927855	0.929219	0.930563	0.931888
1.5	0.933193	0.934478	0.935744	0.936992	0.938220	0.939429	0.940620	0.941792	0.942947	0.944083
1.6	0.945201	0.946301	0.947384	0.948449	0.949497	0.950529	0.951543	0.952540	0.953521	0.954486

F(1.45) = 0.9265 redondeando al cuarto decimal

b) P(Z ≥ 1.45)

P(Z ≥ 1.45) = 1 – P(Z<1.45) = 1 – F(1.45) = 1 – 0.9264 = 0.0735

c) P(Z ≤ –1.45)

P(Z ≤ –1.45) = F(–1.45) =0.0735 (Directamente de la Tabla)

F(–1.45) = 1 – F(1.45) = 1 – 0.9265 = 0.0735 (Usando la relación para valores negativos)

d) P(1.25 ≤ Z ≤ 1.45)

P(1.25 ≤ Z ≤ 1.45) = F(1.45) – F(1.25) = 0.9265 – 0.8944 = 0.0321

e) Encuentre z tal que P(Z ≤ z) = 0.64

P(Z ≤ z) = F(z) = 0.64

En la Tabla, el valor de **z** más cercano a **F(z) = 0.64** corresponde a **z = 0.36**

Z	0.00	0.01	0.02	0.03	0.04	0.05	**0.06**	0.07	0.08	0.09
0.0	0.500000	0.503989	0.507978	0.511967	0.515953	0.519939	0.532922	0.527903	0.531881	0.535856
0.1	0.539828	0.543795	0.547758	0.551717	0.555760	0.559618	0.563559	0.567495	0.571424	0.575345
0.2	0.579260	0.583166	0.587064	0.590954	0.594835	0.598706	0.602568	0.606420	0.610261	0.614092
0.3	0.617911	0.621719	0.625516	0.629300	0.633072	0.636831	**0.640576**	0.644309	0.648027	0.651732
0.4	0.655422	0.659097	0.662757	0.666402	0.670031	0.673645	0.677242	0.680822	0.684386	0.687933
0.5	0.691462	0.694974	0.698468	0.701944	0.705401	0.708840	0.712260	0.715661	0.719043	0.722405
0.6	0.725747	0.729069	0.732371	0.735653	0.738914	0.742154	0.745373	0.748571	0.751748	0.754903
0.7	0.758036	0.761148	0.764238	0.767305	0.770350	0.773373	0.776373	0.779350	0.782305	0.785236
0.8	0.788145	0.791030	0.793892	0.796731	0.799546	0.802338	0.805106	0.807850	0.810570	0.813267

Este es el valor más cercano a **F(z) = 0.64**

7.2.2 ESTANDARIZACIÓN DE LA DISTRIBUCIÓN NORMAL

Si una variable tiene distribución Normal, mediante una sustitución se la puede transformar a otra variable con distribución Normal Estándar. Este cambio de variable facilita el cálculo de probabilidad y se denomina estandarización de la distribución de la variable.

Notación

$X \sim N(\mu, \sigma)$ Define a **X** como una variable con **distribución Normal** con media μ y desviación estándar σ

$Z \sim N(0, 1)$ Define a **Z** como una variable con **distribución Normal Estándar** con media **0** y desviación estándar **1**

Definición: Estandarización de la distribución Normal

Sea **X** una variable aleatoria con distribución Normal: $X \sim N(\mu, \sigma)$,

Entonces, la variable aleatoria $Z = \dfrac{X - \mu}{\sigma}$

Tiene distribución Normal Estándar: $Z \sim N(0, 1)$

Representación gráfica

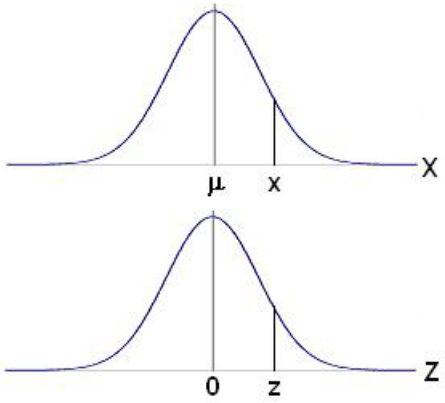


Gráfico de la distribución Normal y la distribución Normal Estándar

La relación entre **X** y **Z** es lineal, por lo tanto la distribución de **Z** debe tener una forma similar a la distribución Normal**.** Mediante las definiciones de valor esperado y varianza:

$$E(Z) = E\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma}[E(X) - E(\mu)] = \frac{1}{\sigma}(\mu - \mu) = 0$$

$$V(Z) = V\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma^2}[V(X) - V(\mu)] = \frac{1}{\sigma^2}(\sigma^2 - 0) = 1$$

Se puede probar que **Z** tiene distribución Normal Estándar: $Z \sim N(0, 1)$

Ejemplo.

La duración de un evento tiene distribución Normal con media 10 y varianza 4.
Encuentre la probabilidad que el evento dure,

a) Menos de 9 horas
b) Entre 11 y 12 horas

Solución

Sea **X:** Variable aleatoria continua (duración en horas) con distribución Normal:

$$X \sim N(10, 2)$$

Entonces $Z = \dfrac{X-10}{2}$ tiene distribución Normal Estándar: $Z \sim N(0, 1)$

a) $P(X \le 9) = P(Z \le \dfrac{9-10}{2}) = P(Z \le -0.5) = F(-0.5) = 0.3085 = 30.85\%$

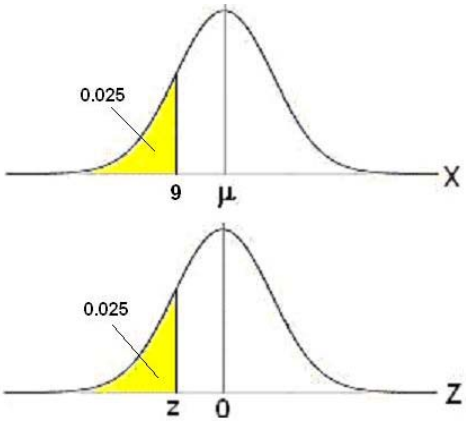
b) $P(11 \le X \le 12) = P(\dfrac{11-10}{2} \le Z \le \dfrac{12-10}{2}) = P(0.5 \le Z \le 1) = F(1) - F(0.5)$

$= 0.8413 - 0.6915 = 0.1498$

Ejemplo

Sea $X \sim N(10, \sigma)$. Encuentre σ tal que $P(X \le 9) = 0.025$

Solución



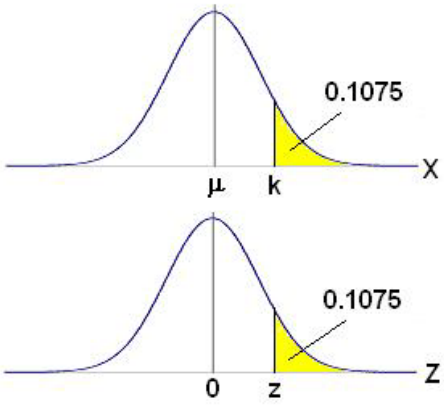
$P(X \le 9) = P(Z \le z) = F(z) = 0.025 \Rightarrow z = -1.96$

$$z = \frac{X-\mu}{\sigma} \Rightarrow -1.96 = \frac{9-10}{\sigma} \Rightarrow \sigma = 0.5102$$

Ejercicio

Sea $X \sim N(300, 50)$. Encuentre el valor de **k** tal que $P(X>k) = 0.1075$

Solución



$P(X > k) = 0.1075 \Rightarrow P(Z > z) = 0.1075 \Rightarrow P(Z \le z) = 1 - 0.1075 = 0.8925$

$P(Z \le z) = F(z) = 0.8925 \Rightarrow z = 1.24$

Pero, $\mathbf{z = \frac{k-\mu}{\sigma}}$ **,** por lo tanto $\mathbf{1.24 = \frac{k-300}{50} \Rightarrow k = 362}$

7.2.3 VALORES REFERENCIALES DE LA DISTRIBUCIÓN NORMAL

Hay ciertos valores de la distribución Normal de uso frecuente.

Si **X** es una variable aleatoria con distribución Normal, la probabilidad que tome valores en un intervalo centrado en μ, hasta una distancia de una desviación estándar σ es aproximadamente **68%,** hasta una distancia de 2σ es aproximadamente **95%** y hasta una distancia de 3σ es cercano a **100%** como se demuestra a continuación:

1) $P(\mu - \sigma \le X < \mu + \sigma) = P\left(\frac{(\mu-\sigma)-\mu}{\sigma} \le Z \le \frac{(\mu+\sigma)-\mu}{\sigma}\right) = P(-1 \le Z \le 1)$

$= F(1) - F(-1) = 0.8413 - 0.1587 = 0.6826 =$ **68.26%**

2) $P(\mu - 2\sigma \le X < \mu + 2\sigma) = P\left(\frac{(\mu-2\sigma)-\mu}{\sigma} \le Z \le \frac{(\mu+2\sigma)-\mu}{\sigma}\right) = P(-2 \le Z \le 2)$

$= F(2) - F(-2) = 0.9773 - 0.0228 = 0.9545 =$ **95.45%**

3) $P(\mu - 3\sigma \le X < \mu + 3\sigma) = P\left(\frac{(\mu-3\sigma)-\mu}{\sigma} \le Z \le \frac{(\mu+3\sigma)-\mu}{\sigma}\right) = P(-3 \le Z \le 3)$

$= F(3) - F(-3) = 0.9987 - 0.0014 = 0.9973 =$ **99.73%**

7.2.4 APROXIMACIÓN DE LA DISTRIBUCIÓN BINOMIAL CON LA DISTRIBUCIÓN NORMAL ESTÁNDAR

Sea **X** una variable aleatoria discreta con **distribución Binomial** con media $\mu = \mathbf{np}$, y varianza $\sigma^2 = \mathbf{np(1-p)}$

Entonces, el límite de la distribución de la variable aleatoria

$$\mathbf{Z = \frac{X-\mu}{\sigma} = \frac{X-np}{\sqrt{np(1-p)}}}\text{, cuando } n \to \infty,$$

Es la d**istribución Normal Estándar: N(0,1)**

La demostración es una aplicación del Teorema del Límite Central, uno de los teoremas fundamentales de la estadística y que será enunciado posteriormente

La bibliografía estadística establece que la aproximación es aceptable aún con valores pequeños de **n**, siempre que **p** esté cerca de **0.5**, o si simultáneamente:

$\mathbf{np > 5}$ y $\mathbf{n(1-p) > 5}$

Ejemplo
En una fábrica, el 20% de los artículos salen defectuosos. Calcule la probabilidad que en un lote de 100 artículos elegidos al azar, 15 sean defectuosos.

Respuesta
Sea **X:** variable aleatoria discreta con distribución Binomial, con **n=20, p=0.2**

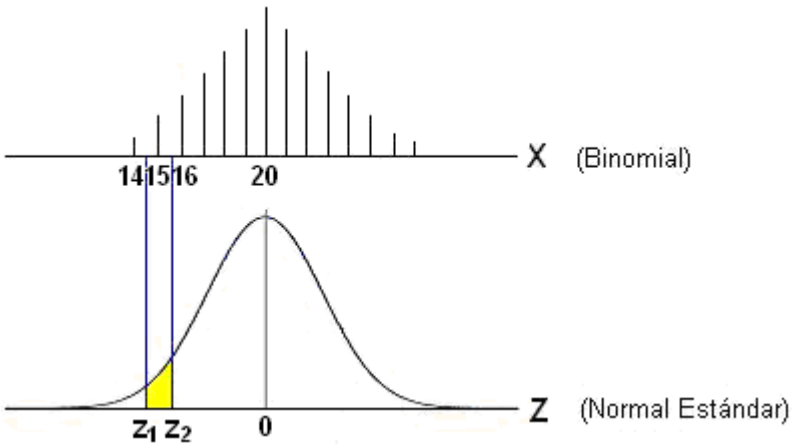
El cálculo con el modelo de la distribución Binomial puede ser impráctico:

$$P(X=x) = \binom{n}{x} p^x (1-p)^{n-x} \Rightarrow P(X=15) = \binom{100}{15} (0.2)^{15} (0.8)^{85}$$

Se observa que **np = 100(0.2) = 20, n(1–p) = 100(0.8) = 80**.

Siendo ambos productos mayores a **5**, según el criterio dado, la distribución Normal Estándar será una aproximación aceptable:

$$Z = \frac{X-\mu}{\sigma} = \frac{X-np}{\sqrt{np(1-p)}} = \frac{X-100(0.20)}{\sqrt{100(0.20)(0.80)}} = \frac{X-20}{4}$$



$$P(X=15) \cong P\left(\frac{14.5-\mu}{\sigma} \le Z \le \frac{15.5-\mu}{\sigma}\right) = P\left(\frac{14.5-20}{4} \le Z \le \frac{15.5-20}{4}\right)$$

$$= P(-1.375 \le Z \le -1.125) = F(-1.125) - F(-1.375)$$

$$= 0.130 - 0.084 = 0.046 = 4.6\%$$

Observe la corrección que se realiza al tomar el valor discreto para usarlo en la distribución Normal. Para la distribución Normal se considera que un valor discreto se extiende entre las mitades de los valores adyacentes: el valor **15** de la distribución Binomial corresponde al intervalo (**14.5, 15.5**) para la distribución Normal.

7.2.5 EJERCICIOS

1) Suponga que **Z** es una variable aleatoria con distribución Normal Estándar. Use la tabla para calcular:

a) P(Z<1.45)
b) P(Z>2.01)
c) P(Z<-1.24)
d) P(Z>1.78)
e) P(-1.25<Z<2.31)

2) Suponga que **X** es una variable aleatoria con distribución Normal, con media 25 y desviación estándar 5. Use la tabla para calcular

a) P(X<18)
b) P(X>30)
c) P(24<X<27)

3) Si $\mathbf{X \sim N(10, \sigma^2)}$ determine el valor de la varianza si **P(X<9)=0.025**

4) El peso de los artículos producidos por una fábrica tiene distribución Normal con una media de 50 gr. y una desviación estándar de 5 gr.

a) Calcule la probabilidad que un artículo elegido al azar tenga un peso de mas de 60 gr.
b) Calcule la proporción de los paquetes que tendrían un peso entre 46 y 54 gr.

5) El tiempo necesario para llenar un frasco de un producto es una variable aleatoria que sigue una distribución Normal con una media de 10 segundos y una desviación estándar de dos segundos.

a) Calcule la probabilidad que el tiempo de llenado exceda a 11 segundos
b) Encuentre el tiempo de llenado del frasco tal que la probabilidad de excederlo tenga una probabilidad de 3%

6) Una fábrica de tornillos produce un tipo de tornillo con un diámetro promedio de 6.5 mm. y una desviación estándar de 1.5 mm. Suponiendo que la distribución es Normal calcule la probabilidad de encontrar tornillos con diámetro,

a) mayor que 7mm.
b) entre 6 y 7 mm.

7) El pH de un químico tiene una distribución $\mathbf{N(\mu, 0.10^2)}$. Durante la elaboración del producto se ordena suspender la producción si el pH supera el valor 7.20 o es inferior a 6.80.

a) Calcule la probabilidad que la producción no sea suspendida si μ=7.0
b) Calcule la probabilidad que la producción no sea suspendida si μ=7.05
c) Cual debe ser μ para que la probabilidad de que se suspenda la producción sea 0.85

8) La tolerancia especificada para aceptar los ejes producidos por una fábrica es que el diámetro sea 0.45 ± 0.005 cm. Si los ejes producidos por la fábrica tienen distribución Normal con media 0.452 y desviación estándar 0.003 cm., determine cuantos ejes serán rechazados de cada lote de 500 ejes producidos.

MATLAB

Probabilidad con la distribución Normal

```
>> p=normcdf(-1.45)
    p =
        0.0735
```
Distribución Normal Estándar acumulada, **P(Z ≤ -1.45)**

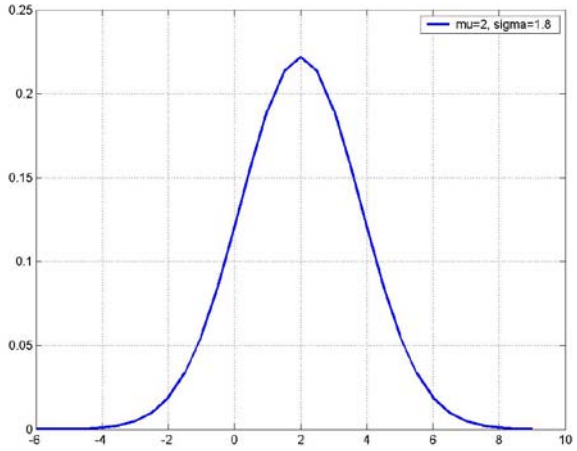
```
>> p=normcdf(1.45)-normcdf(1.25)
    p =
        0.0321
```
Calcular **P(1.25 ≤ Z ≤ 1.45)**

```
>> p=normcdf(9, 10, 2)
    p =
        0.3085
```
Distribución Normal: calcular **P(X ≤ 9)**, **μ =10**, **σ = 2**

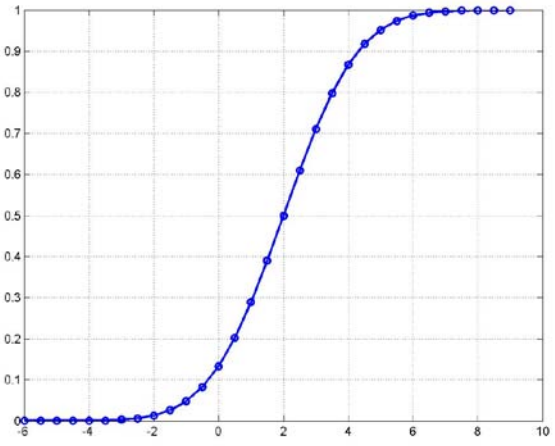
```
>> x=norminv(0.3085, 10, 2)
    x =
      8.9998
```
Función inversa**:** calcular **x** tal que **F(x) = 0.3085**
μ =10, **σ = 2**

```
>> x=-6: 0.5: 9;
```
x = -6, -5.5, -5.0, . . ., 9

```
>> f=normpdf(x, 2, 1.8);
>> plot(x,f,'b'), grid on
>> legend('mu=2, sigma=1.8')
```
Valores de densidad Normal **f(x)**, **μ = 2**, **σ = 1.8**
Gráfico de la función de densidad Normal




```
>> f=normcdf(x, 2, 1.8);
>> plot(x,f,'ob'), grid on
>> hold on
>> plot(x,f,'b')
```
Valores de la distribución acumulada **μ = 2**, **σ = 1.8**
Gráfico de puntos de **F(x)**
Superponer gráfico
Gráfico de la distribución acumulada **F(x)**



7.3 DISTRIBUCIÓN GAMMA

Es un modelo básico en la Teoría de la Probabilidad y corresponde a la siguiente definición

Definición: Distribución Gamma

Sea **X:** Variable aleatoria continua

X tiene **distribución Gamma** si su función de densidad es

$$f(x) = \begin{cases} \dfrac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, & x > 0 \\ 0, & \text{para otro } x \end{cases}$$

$\alpha>0$, $\beta>0$ son los parámetros para este modelo

$\Gamma(\alpha)$ es la **función Gamma** que está definida de la siguiente forma:

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx$$

Si α es un entero positivo, entonces

$$\Gamma(\alpha) = (\alpha - 1)!$$

Demostración

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx$$

$u = x^{\alpha-1} \Rightarrow du = (\alpha-1)x^{\alpha-2}\,dx$ (integración por partes)

$dv = e^{-x}\,dx \Rightarrow v = -e^{-x}$

Se obtiene

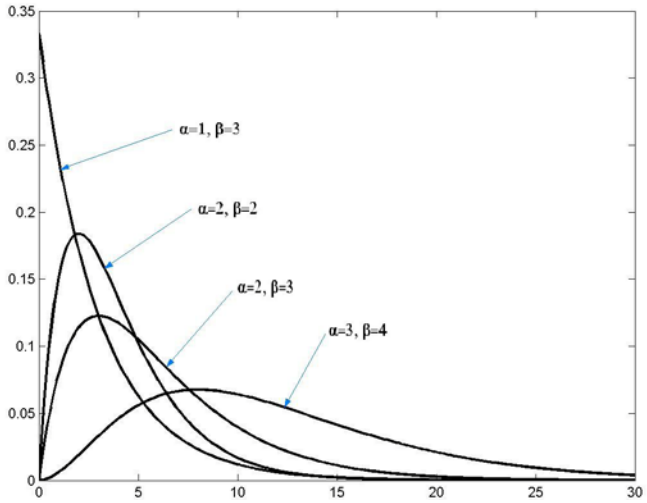
$$\Gamma(\alpha) = (\alpha - 1)\int_0^{\infty} x^{\alpha-2} e^{-x} dx = (\alpha - 1)\Gamma(\alpha - 1)$$

Sucesivamente

$\Gamma(\alpha) = (\alpha -1)(\alpha-2)(\alpha-3)...\Gamma(1)$. Finalmente, $\Gamma(1) = 1$ por integración directa.

Gráfico de la Distribución Gamma

Son gráficos asimétricos con sesgo positivo y su dominio es $\Re^+$



La distribución Gamma para algunos valores de α, β

7.3.1 MEDIA Y VARIANZA PARA LA DISTRIBUCIÓN GAMMA

Definición: Media y Varianza para la Distribución Gamma

Sea **X** una variable aleatoria continua con distribución Gamma, entonces
Media: $\mu = E(X) = \alpha\beta$ Varianza: $\sigma^2 = V(X) = \alpha\beta^2$

Demostración

$$\mu = \int_{-\infty}^{\infty} xf(x)dx = \int_{0}^{\infty} x\frac{1}{\beta^{\alpha}\Gamma(\alpha)}x^{\alpha-1}e^{-x/\beta}dx = \frac{1}{\beta^{\alpha}\Gamma(\alpha)}\int_{0}^{\infty} x^{\alpha}e^{-x/\beta}dx$$

Mediante la sustitución $y = x/\beta$

$$\mu = \frac{1}{\beta^{\alpha}\Gamma(\alpha)}\int_{0}^{\infty}(\beta y)^{\alpha}e^{-y}\beta dy$$

$$= \frac{\beta}{\Gamma(\alpha)}\int_{0}^{\infty} y^{\alpha}e^{-y}dy$$

Con la definición de la función Gamma:

$$= \frac{\beta}{\Gamma(\alpha)}\Gamma(\alpha+1) = \frac{\beta}{\Gamma(\alpha)}\alpha\Gamma(\alpha) = \alpha\beta$$

Ejemplo

El tiempo en horas que semanalmente requiere una máquina para mantenimiento es una variable aleatoria con distribución gamma con parámetros $\alpha=3, \beta=2$

a) Encuentre la probabilidad que en alguna semana el tiempo de mantenimiento sea mayor a **8** horas
b) Si el costo de mantenimiento en dólares es $C = 30X + 2X^2$, siendo **X** el tiempo de mantenimiento, encuentre el costo promedio de mantenimiento.

Solución

Sea **X:** duración del mantenimiento en horas (variable aleatoria)
Su densidad de probabilidad es:

$$f(x) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)}x^{\alpha-1}e^{-x/\beta} = \frac{1}{2^3\Gamma(3)}x^{3-1}e^{-x/2} = \frac{1}{16}x^2e^{-x/2}$$

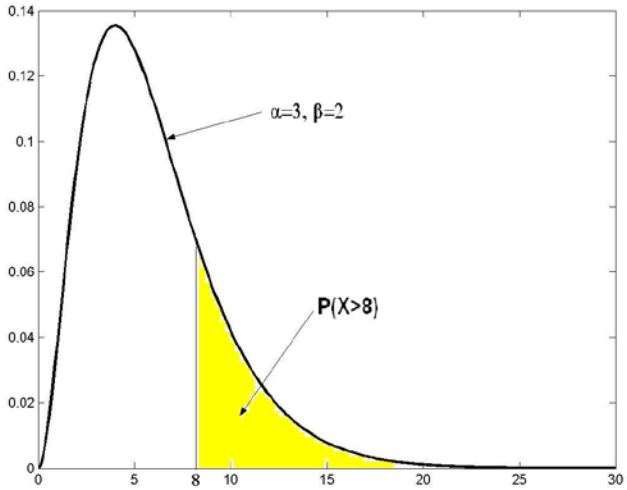


Gráfico de la función de densidad para el ejemplo

a) **P(X>8)** es el área resaltada en el gráfico

$$P(X>8) = 1 - P(X\le 8) = 1 - \frac{1}{16}\int_0^8 x^2 e^{-x/2}dx$$

Para integrar se pueden aplicar dos veces la técnica de integración por partes:

$$\int x^2 e^{-x/2}dx\ ,$$

$u = x^2 \Rightarrow du = 2x\,dx$
$dv = e^{-x/2}dx \Rightarrow v = -2\,e^{-x/2}$

$$= -2x^2 e^{-x/2} + 4\int x\ e^{-x/2}dx$$

$$\int x\ e^{-x/2}dx$$

$u = x \Rightarrow du = dx$
$dv = e^{-x/2}dx \Rightarrow v = -2\,e^{-x/2}$

$$= -2x\,e^{-x/2} + 2\int e^{-x/2}dx$$

Sustituyendo los resultados intermedios,

$$P(X>8) = 1 - \frac{1}{16}\left[-2x^2e^{-x/2} + 4(-2x\,e^{-x/2} + 2(-2\,e^{-x/2}))\right]_0^8 = 0.2381$$

b) $E(C) = E(30X + 2X^2) = 30\,E(X) + 2\,E(X^2)$

$E(X) = \alpha\beta = 3(2) = 6$

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)dx = \int_0^{\infty} x^2\,\frac{1}{16}x^2e^{-x/2}dx = \frac{1}{16}\int_0^{\infty} x^4e^{-x/2}dx$$

Sustituya $y = x/2$ para usar la función Gamma

$$= \frac{1}{16}\int_0^{\infty}(2y)^4e^{-y}(2dy) = 2\int_0^{\infty} y^4e^{-y}dy = 2\Gamma(5) = 2(4!) = 48$$

Finalmente se obtiene

E(C) = 30(6) + 2(48) = 276 dólares

7.4 DISTRIBUCIÓN EXPONENCIAL

Es un caso particular de la distribución Gamma y tiene aplicaciones de interés práctico.
Se obtiene con $\alpha = 1$ en la distribución Gamma

Definición: Distribución Exponencial

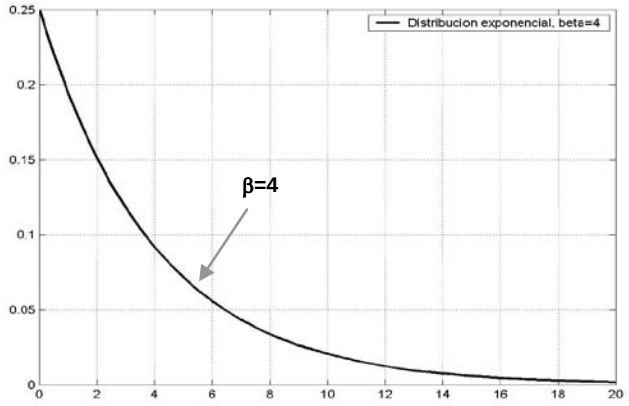
Sea **X:** Variable aleatoria continua
X tiene **distribución Exponencial** si su densidad de probabilidad está dada por

$$f(x) = \begin{cases} \frac{1}{\beta}e^{-x/\beta}, & x > 0 \\ 0, \text{ para otro } x \end{cases}$$

En donde $\beta > 0$, es el parámetro para este modelo

Gráfico de la Distribución Exponencial

El gráfico de la densidad de probabilidad de la distribución Exponencial tiene la forma típica decreciente y su dominio es $\Re^+$



7.4.1 MEDIA Y VARIANZA PARA LA DISTRIBUCIÓN EXPONENCIAL

Definición: Media y Varianza de la Distribución Exponencial

Sea **X:** Variable aleatoria continua con distribución Exponencial, entonces

Media: $\mu = E(X) = \beta$ Varianza: $\sigma^2 = V(X) = \beta^2$

Se obtienen directamente de la distribución Gamma con $\alpha = 1$

Problema

Un sistema usa un componente cuya duración en años es una variable aleatoria con distribución Exponencial con media de 4 años. Si se instalan 3 de estos componentes y trabajan independientemente, determine la probabilidad que al cabo de 6 años, dos de ellos sigan funcionando.

Solución

Sea **Y:** Variable aleatoria **continua** (duración de un componente en años)

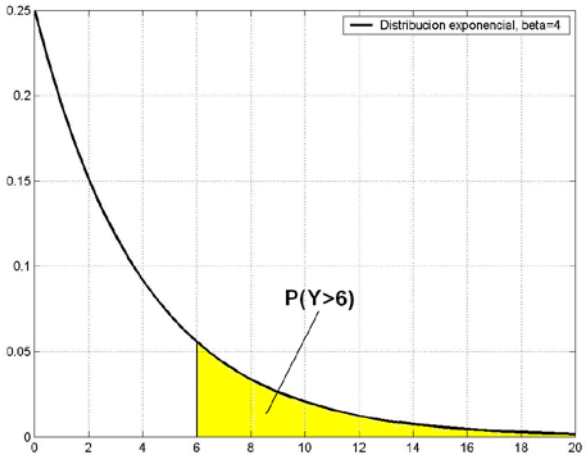
Y tiene distribución Exponencial con $\mu = \beta = 4$

Su densidad de probabilidad es

$$f(y) = \frac{1}{4}e^{-y/4}, y > 0$$

La probabilidad que un componente siga funcionando al cabo de 6 años:

$$P(Y \geq 6) = 1 - P(Y < 6) = 1 - \int_0^6 \frac{1}{4}e^{-y/4}dy = 0.2231$$



Sea **X:** Variable aleatoria **discreta** (cantidad de componentes que siguen funcionando luego de 6 años)

X tiene distribución Binomial con **n=3, p=0.2231**

Su función de distribución de probabilidad es:

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{3}{x} 0.2231^x 0.7769^{3-x}$$

Entonces,

$$P(X=2) = f(2) = \binom{3}{2} 0.2231^2 0.7769^{3-2} = 0.1160 = 11.6\%$$

7.4.2 UNA APLICACIÓN DE LA DISTRIBUCIÓN EXPONENCIAL

Puede demostrarse que si una variable aleatoria tiene **distribución de Poisson** con parámetro λ, entonces el tiempo de espera entre dos "éxitos" consecutivos es una variable aleatoria con **distribución Exponencial** con parámetro $\beta = 1/\lambda$

Ejemplo

La llegada de los barcos a un puerto tiene distribución de Poisson con media de 4 por día. Calcule la probabilidad que el tiempo transcurrido entre la llegada de dos barcos consecutivos en algún día sea menor a 4 horas.

Solución

Sea **X** el tiempo transcurrido entre dos llegadas consecutivas (en días)
X es una variable aleatoria continua con distribución Exponencial con parámetro
$\beta = 1/\lambda = 1/4$

Su función de probabilidad es

$$f(x) = \frac{1}{\beta} e^{-x/\beta} = \lambda e^{-\lambda x} = 4e^{-4x},\ x>0$$

Por lo tanto, $P(X<1/6) = \int_0^{1/6} 4e^{-4x} dx = 0.4866 = 48.66\%$

7.4.3 EJERCICIOS

1) En cierta ciudad, el consumo diario de energía eléctrica en millones de Kw-hora puede considerarse como una variable aleatoria con distribución Gamma con α**=3** y β**=2**. Si la planta de energía tiene una capacidad de producción diaria de doce millones de Kw-hora, calcule la probabilidad que en un día cualquiera, el suministro de energía sea insuficiente.

2) La duración en miles de Km. de cierto tipo de llantas, es una variable aleatoria con distribución exponencial con media 40 mil Km. Calcule la probabilidad que una de estas llantas dure
a) Al menos 20 mil Km.
b) No más de 30 mil Km.

3) El tiempo que transcurre antes de que una persona sea atendida en un bar es una variable aleatoria que se puede modelar col la distribución exponencial con una media de 5 minutos. Calcule la probabilidad de que una persona sea atendida antes de que transcurran 3 minutos en al menos 4 de los 7 días siguientes.

4) Se conoce que la cantidad de reparaciones que cierto tipo de electrodoméstico necesita, tiene distribución de Poisson con una media de una vez cada dos años. Suponiendo que los intervalos entre reparaciones tienen distribución exponencial. Calcule la probabilidad que este artículo funcione por lo menos tres años sin requerir reparación.

MATLAB

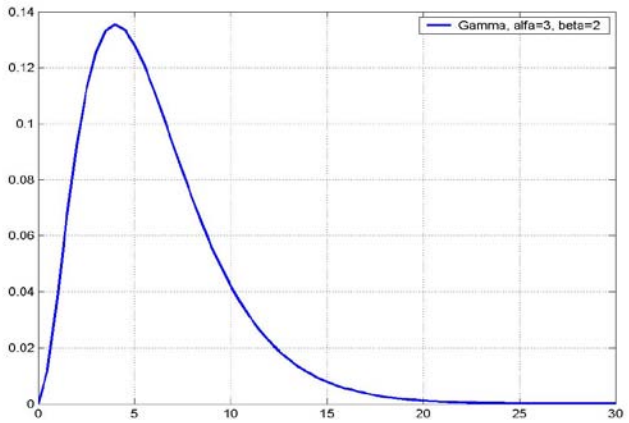
Probabilidad con la distribución gamma

```
>> x=0:0.5:30;
>> f=gampdf(x, 3, 2);
>> plot(x, f, 'b'), grid on
>> legend('Gamma, alfa=3, beta=2');
```

x = 0, 0.5, 1, . . ., 30

Valores de densidad de probabilidad gamma, $\alpha = 3,\ \beta = 2$

Gráfico de la densidad de probabilidad gamma, $\alpha = 3,\ \beta = 2$




```
>> p=gamcdf(8, 3, 2)
   p =
       0.7619
```

Distribución gamma acumulada: **F(8) = P(X≤8)**, $\alpha = 3,\ \beta = 2$

```
>> x=gaminv(0.7619, 3, 2)
   x =
       8.0000
```

Distribución gamma acumulada inversa:
Encontrar **x** tal que **P(X≤x) = 0.7619**, $\alpha = 3,\ \beta = 2$

```
>> [mu, var]=gamstat(3, 2)
    mu =  6
    var = 12
```

Media y varianza de la distribución gamma, $\alpha = 3,\ \beta = 2$

Probabilidad con la distribución exponencial

```
>> x=0:0.5:20;
>> f=exppdf(x,4);
```

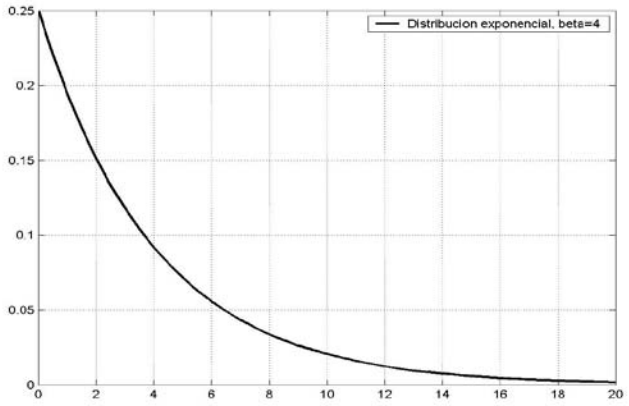
x = 0, 0.5, 1.0, ..., 20

Valores de densidad de probabilidad exponencial, $\beta = 4$

Distribucion exponencial, beta=4

```
>> plot(x,f,'k'),grid on
>> legend('')
```

Gráfico de la densidad de probabilidad exponencial $\beta = 4$



```
>> p=expcdf(6, 4)
   p =
       0.7769
>> x=expinv(0.7769, 4)
   x =
       6.000
```

Distribución exponencial acumulada: **F(6) = P(X≤6)**, **β = 4**

Distribución exponencial acumulada inversa:
Encontrar **x** tal que **P(X≤x) = 0.7769, β = 4**

7.5 DISTRIBUCIÓN DE WEIBULL

Este modelo propuesto por Weibull se usa en problemas relacionados con fallas de materiales y estudios de confiabilidad. Para estas aplicaciones es más flexible que el modelo exponencial.

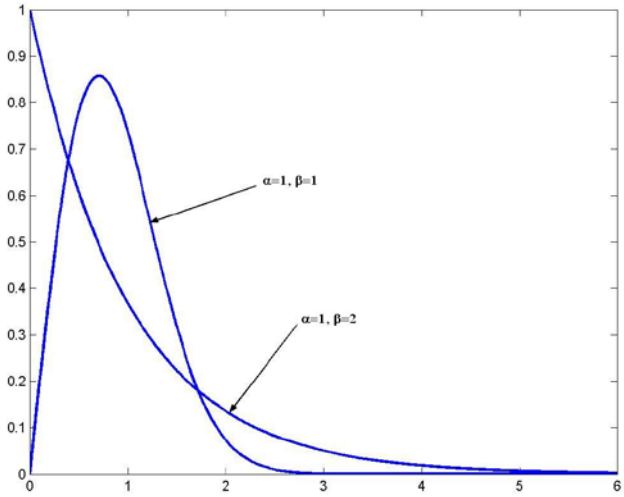
Definición: Distribución de Weibull

Una variable aleatoria continua **X** tiene **distribución de Weibull** si su densidad de probabilidad está dada por

$$f(x) = \begin{cases} \alpha\beta x^{\beta-1} e^{-\alpha x^{\beta}}, & x > 0 \\ 0, & \text{para otro } x \end{cases}$$

En donde $\alpha>0$, $\beta>0$ son los parámetros para este modelo

Si $\beta = 1$, este modelo se reduce a la distribución Exponencial.
Si $\beta > 1$, el modelo tiene forma tipo campana con sesgo positivo



Gráficos de la distribución de Weibull

7.5.1 MEDIA Y VARIANZA PARA LA DISTRIBUCIÓN DE WEIBULL

Definición: Media y Varianza para la distribución de Weibull

Si **X** es una variable aleatoria continua con distribución de Weibull, entonces

Media $\mu = E(X) = \alpha^{-1/\beta}\Gamma(1+1/\beta)$

Varianza $\sigma^2 = V(X) = \alpha^{-2/\beta}[\Gamma(1+2/\beta) - (\Gamma(1+1/\beta))^2]$

Demostración

Con la definición

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx = \int_{0}^{\infty} x\alpha\beta x^{\beta-1} e^{-\alpha x^{\beta}} dx$$

Usando la sustitución

$$y = \alpha x^{\beta} \Rightarrow dy = \alpha\beta x^{\beta-1}dx = \beta y x^{-1}dx = \beta y (y/\alpha)^{-1/\beta}dx$$

Se obtiene

$$\mu = \alpha^{-1/\beta}\int_{0}^{\infty} y^{1/\beta} e^{-y} dy$$

Finalmente, se compara con la función Gamma

$$\mu = \alpha^{-1/\beta}\Gamma(1+1/\beta)$$

Ejemplo
Suponga que la vida útil en horas de un componente electrónico tiene distribución de Weibull con $\alpha=0.1,\ \beta=0.5$

a) Calcule la vida útil promedio
b) Calcule la probabilidad que dure mas de 300 horas

Solución

Sea **X**: Vida útil en horas (variable aleatoria continua)
su densidad de probabilidad:

$$f(x)= \alpha\beta x^{\beta-1}e^{-\alpha x^{\beta}} = 0.05x^{-0.5}e^{-0.1x^{0.5}}$$

a) $\mu = \alpha^{-1/\beta}\Gamma(1+1/\beta) = (0.1)^{-1/0.5}\Gamma(1+1/0.5) = 0.1^{-2}\Gamma(3) = 200$ **horas**

b) $P(X>300) = \int_{300}^{\infty} 0.05x^{-0.5}e^{-0.1x^{0.5}}\,dx$

Mediante la sustitución $y=x^{0.5} \Rightarrow dy = 0.5x^{-0.5}dx = 0.5(\frac{1}{y})dx \Rightarrow dx = \frac{y}{0.5}dy$

se obtiene

$$P(X>300) = 0.05\int_{\sqrt{300}}^{\infty}\frac{1}{y}e^{-0.1y}\frac{y}{0.5}dy = 0.1\int_{\sqrt{300}}^{\infty}e^{-0.1}dy$$

$$= 1 - P(X\leq 300) = 1 - 0.1\int_{0}^{\sqrt{300}}e^{-0.1}dy = 0.177$$

7.6 RAZÓN DE FALLA

Si la variable aleatoria es el tiempo **t** en que falla un equipo, el índice o razón de falla en el instante **t** es la función de densidad de falla al tiempo **t**, dado que la falla no ocurre antes de **t**.

Definición: Razón de Falla

Sean **t:** Variable aleatoria continua (tiempo)
f(t): Función de densidad de probabilidad
F(t): Función de distribución (función de probabilidad acumulada)
Entonces

$$r(t) = \frac{f(t)}{1-F(t)}$$ es la razón de falla

7.7 DISTRIBUCIÓN BETA

Este modelo tiene aplicaciones importantes por la variedad de formas diferentes que puede tomar su función de densidad eligiendo valores para sus parámetros.

Definición: Distribución Beta

Una variable aleatoria continua **X** tiene **distribución Beta** si su densidad de probabilidad está dada por

$$f(x) = \begin{cases} \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,x^{\alpha-1}(1-x)^{\beta-1}, & 0<x<1 \\ 0, & \text{para otro } x \end{cases}$$

En donde $\alpha>0,\ \beta>0$ son los parámetros para este modelo. $\Gamma(\ )$ es la función Gamma

El dominio de la distribución Beta es el intervalo **(0, 1)**, pero puede adaptarse a otros intervalos finitos mediante una sustitución de la variable aleatoria.

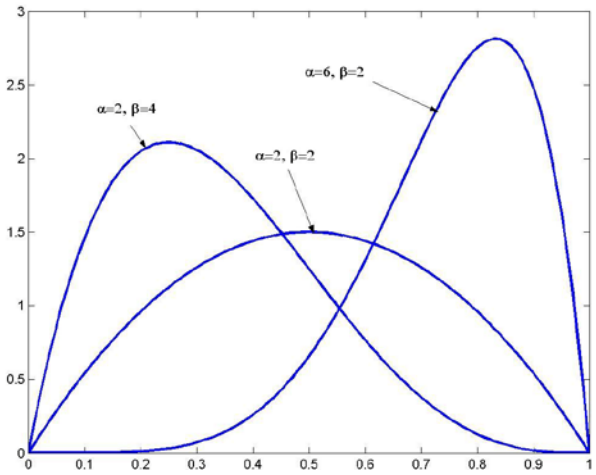


Gráfico de la distribución Beta para algunos valores α, β

7.7.1 MEDIA Y VARIANZA PARA LA DISTRIBUCIÓN BETA

Definición: Media y Varianza para la Distribución Beta

Si **X** es una variable aleatoria continua con distribución Beta, entonces

Media: $\mu = E(X) = \dfrac{\alpha}{\alpha+\beta}$

Varianza: $\sigma^2 = V(X) = \dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Se omite la demostración, la cual se fundamenta en la definición de la **función Beta.**

Ejemplo

Un distribuidor de cierto producto llena su bodega al inicio de cada semana. La proporción del artículo que vende semanalmente se puede modelar con la **distribución Beta con α=4, β=2**

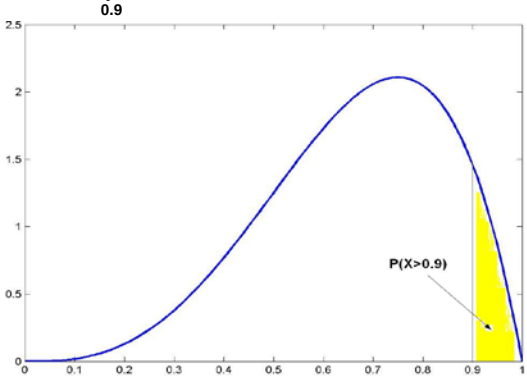
a) Encuentre la probabilidad que en alguna semana venda al menos 90%
b) Encuentre el valor esperado de la proporción de venta semanal

Solución

Sea **X**: Proporción del artículo que vende semanalmente (variable aleatoria continua)
Su densidad de probabilidad es

$$f(x) = \frac{\Gamma(4+2)}{\Gamma(4)\Gamma(2)} x^{4-1}(1-x)^{2-1} = 20x^3(1-x),\ \ 0<x<1$$

a) $P(x \geq 0.9) = 20\int_{0.9}^{1} x^3(1-x)dx = 0.082 = 8.2\%.$ Es el área marcada en el siguiente gráfico



b) $\mu = E(X) = \dfrac{\alpha}{\alpha+\beta} = \dfrac{4}{4+2} = 2/3$ (vende en promedio 2/3 del producto cada semana)

7.8 DISTRIBUCIÓN DE ERLANG

La función de densidad de la distribución de Erlang es igual a la distribución gamma, pero el parámetro α debe ser entero positivo.

Definición: Distribución de Erlang

Una variable aleatoria continua X tiene distribución de **Erlang** si su densidad de probabilidad está dada por

$$f(x) = \begin{cases} \dfrac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, & x > 0 \\ 0, & \text{para otro } x \end{cases}$$

α**>0,** β**>0** son los parámetros para este modelo, α entero positivo

7.8.1 MEDIA Y VARIANZA PARA LA DISTRIBUCIÓN DE ERLANG

Definición: Media y Varianza para la Distribución de Erlang

Si **X** es una variable aleatoria continua con distribución de Erlang, entonces

Media: $\mu = E(X) = \alpha\beta$, **Varianza:** $\sigma^2 = V(X) = \alpha\beta^2$

7.9 DISTRIBUCIÓN JI-CUADRADO

Este modelo es importante en el estudio de la Estadística Inferencial. Se obtiene de la distribución Gamma con $\alpha = \nu/2,\ \beta = 2$

Definición: Distribución Ji-cuadrado

Una variable aleatoria continua **X** tiene distribución **Ji-cuadrado** si su densidad de probabilidad está dada por

$$f(x) = \begin{cases} \dfrac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\frac{\nu}{2}-1} e^{-x/2}, & x > 0 \\ 0, & \text{para otro } x \end{cases}$$

Esta distribución tiene un parámetro: $\nu > 0$ y se denomina **número de grados de libertad**.

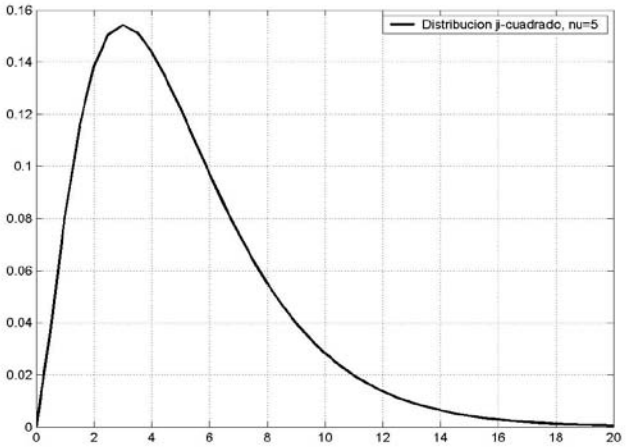


Gráfico de la distribución Ji-cuadrado con $\nu = 5$

7.9.1 MEDIA Y VARIANZA DE LA DISTRIBUCIÓN JI-CUADRADO

Definición: Media y Varianza para la Distribución Ji-cuadrado

Si **X** es una variable aleatoria continua con distribución Ji-cuadrado, entonces

Media $\mu = E(X) = \nu$, **Varianza:** $\sigma^2 = V(X) = 2\nu$

Se obtienen directamente de la distribución Gamma con $\alpha = \nu/2,\ \beta = 2$

7.9.2 EJERCICIOS

1) Si la proporción anual de declaraciones incorrectas del impuesto sobre la renta entregadas al fisco puede considerarse como una variable aleatoria que tiene una distribución Beta con **α=2 y β=9**.

a) Calcule la probabilidad que en un año cualquiera haya mas de 40% de declaraciones incorrectas
b) Encuentre la media de esta distribución, es decir, la proporción de declaraciones que en promedio serán incorrectas

2) Suponga que el tiempo de servicio en horas de un semiconductor es una variable aleatoria que tiene distribución de Weibull con **α=0.025, β=0.5**

a) Calcule el tiempo esperado de duración del semiconductor
b) Calcule la probabilidad que este semiconductor esté funcionando después de 4000 horas de uso

3) Sea **t** una variable aleatoria continua que representa el tiempo de falla de un equipo. Demuestre que si **t** tiene distribución exponencial, la razón de falla es constante.

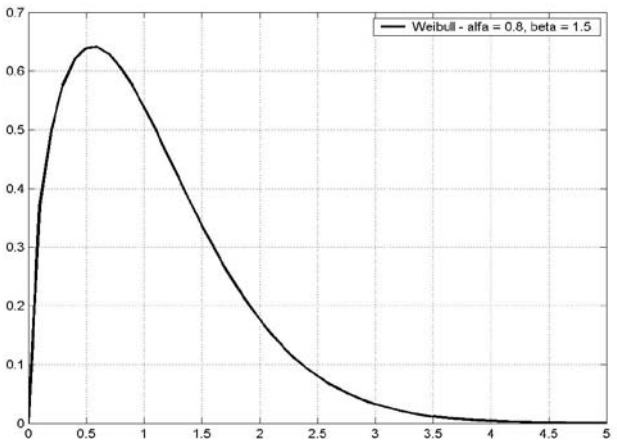
4) Durante cada turno de trabajo de 8 horas, la proporción de tiempo que una máquina está en reparación tiene distribución Beta con **α=1** y **β=2**.

a) Determine la probabilidad que la proporción del turno que la máquina está en reparación se menor que 2 horas
b) Si el costo de reparación es $100 más $10 por la duración al cuadrado, encuentre el valor esperado del costo de reparación

MATLAB

Distribución de Weibull

```
>> p=weibcdf(300,0.1,0.5)        Distribución acumulada Weibull , α = 0.1, β = 0.5
   p =                           Calcular P(X≤300)
       0.8231
>> [mu, var]=weibstat(0.1, 0.5)  Media y varianza distr. Weibull,  α = 0.1, β = 0.5
   mu =  200.0000
   var =  2.0000e+005
>> x=0:0.1:5;
>> f=weibpdf(x,0.8,1.5);         Puntos de la distr. Weibull,  α = 0.8, β = 1.5
>> plot(x,f,'k'), grid on        Gráfico de la distribución Weibull
>> legend('Weibull - alfa = 0.8, beta = 1.5')
```



Distribución beta

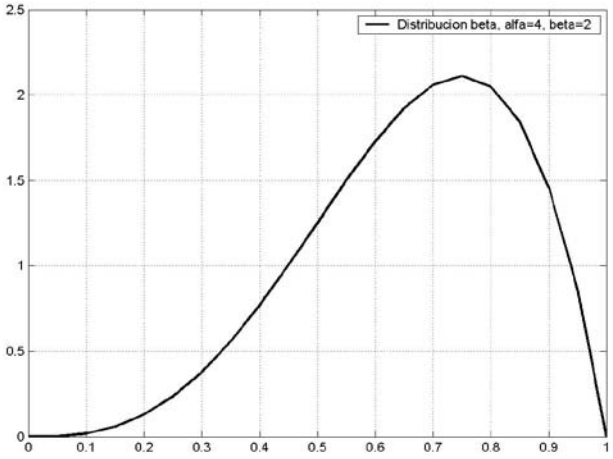
```
>> p=betacdf(0.9, 4, 2)
    p =
        0.9185
>> x=betainv(0.9185, 4, 2)
    x =
        0.9000
>> [mu, var] = betastat(4, 2)
     mu = 0.6667
     var =  0.0317
>> x=0: 0.05: 1;
>> f=betapdf(x, 4, 2);
>> plot(x, f, 'k'), grid on
>> legend('Distribucion beta, alfa=4, beta=2')
```

Distribución acumulada beta, $\alpha = 4, \beta = 2$
Calcular **P(X≤0.9)**

Distribución beta inversa
Calcular **x** tal que **F(x) = 0.9185**, $\alpha = 4, \beta = 2$

Media y varianza distr. beta, $\alpha = 4, \beta = 2$

Puntos de la distr. beta, $\alpha = 4, \beta = 2$
Gráfico de la distribución beta



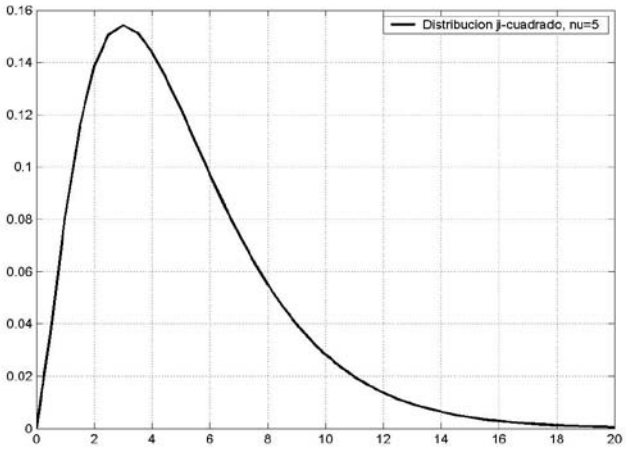
Distribución ji-cuadrado

```
>> p=chi2cdf(2,5)
    p =
        0.1509
>> [mu, var]=chi2stat(5)
     mu =  5
     var = 10
>> x=0:0.5:20;
>> f=chi2pdf(x,5);
>> plot(x,f,'k'), grid on
>> legend('Distribucion ji-cuadrado, nu=5')
```

Distribución acumulada ji-cuadrado, $\nu = 5$
Calcular **P(X≤2)**

Media y varianza distr. ji-cuadrado, $\nu = 5$

Puntos de la distr. ji-cuadrado, $\nu = 5$
Gráfico de la distribución ji-cuadrado



7.10 DISTRIBUCIÓN EMPÍRICA ACUMULADA

Esta distribución es un modelo matemático que se asigna a un conjunto de datos cuando se desconoce si pertenecen a un modelo de probabilidad específico. La Distribución Empírica Acumulada es una función de probabilidad que asocia cada valor de la variable **x** con la proporción de datos menores que el valor de **x** dado

Definición: Distribución Empírica Acumulada

Sean

$x_1, x_2, \ldots, x_n$, datos obtenidos en una muestra.

Si se escriben estos datos en orden creciente:

$x_{(1)}, x_{(2)}, \ldots, x_{(n)}$

Se define la Distribución Empírica Acumulada

$$F(x) = \begin{cases} 0, & x < x_{(1)} \\ \frac{i}{n}, & x_{(i)} \le x < x_{(i+1)} \\ 1, & x \ge x_{(n)} \end{cases}, \quad x \in \Re$$

Ejemplo. Dados los siguientes datos de una muestra: **4, 3, 8, 6, 5**
Encuentre y grafique la distribución Empírica

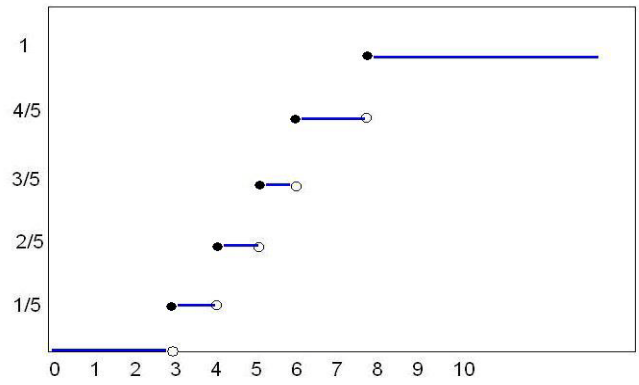
Solución

Datos ordenados: **3, 4, 5, 6, 8** **(n=5)**

Su distribución Empírica Acumulada es:

$$F(x) = \begin{cases} 0, & x < 3 \\ 1/5, & 3 \le x < 4 \\ 2/5, & 4 \le x < 5 \\ 3/5, & 5 \le x < 6 \\ 4/5, & 6 \le x < 8 \\ 1, & x \ge 8 \end{cases}$$

Gráfico de la distribución Empírica Acumulada



7.10.1 EJERCICIOS

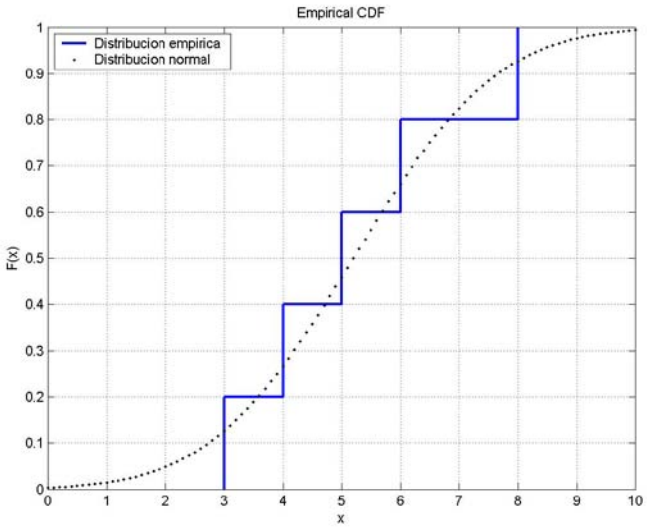
1) Grafique la distribución empírica correspondiente a los siguientes datos
14, 5, 8, 3, 8, 7, 11, 13, 14, 3

2) Calcule la media aritmética, mediana, varianza, y distribución empírica de la siguiente muestra: **4, 8, 2, 7, 10, 8, 4, 9, 7**

MATLAB

Gráfico de la distribución empírica y la distribución normal acumuladas

>> x=[3 4 5 6 8];	Vector con datos de una muestra
>> cdfplot(x)	Gráfico de la distribución empírica acumulada
>> m=mean(x);	Media muestral
>> s=std(x);	Desviación estándar muestral
>> z=0: 0.1: 10;	Puntos para la distribución normal acumulada
>> hold on	Para superponer gráficos
>> f=normcdf(z, m, s);	Valores de la distribución normal acumulada para los puntos
>> plot(z, f, '.k')	Gráfico de la distribución normal acumulada, puntos en negro
>> legend('Distribucion empirica','Distribucion normal',2)	Colocar rótulos arriba izquierda



El número 2 indica que los rótulos se coloquen arriba a la izquierda

8 DISTRIBUCIONES DE PROBABILIDAD CONJUNTA

Algunos experimentos estadísticos pueden incluir más de una variable aleatoria las cuales actúan en forma conjunta, y es de interés determinar la probabilidad correspondiente a los diferentes valores que estas variables puedan tomar.

8.1 CASO DISCRETO BIVARIADO

8.1.1 DISTRIBUCIÓN DE PROBABILIDAD CONJUNTA

Definición: Distribución de Probabilidad Conjunta

Sean **X, Y:** variables aleatorias discretas.
x, y: valores que pueden tomar **X, Y**
Su función de distribución de probabilidad conjunta se escribe **f(x,y)**
y describe el valor de probabilidad en cada punto **P(X=x, Y=y)**

Esta función establece correspondencia de **(x,y)** a **(0,1)** y satisface las siguientes propiedades

1) $\forall x \forall y\ f(x,y) \geq 0$ — **f** no puede tomar valores negativos

2) $\sum_x \sum_y f(x,y) = 1$ — La suma de todos los valores de **f** debe ser **1**

3) $P(X=x, Y=y) = f(x,y)$ — **f** debe ser un modelo para calcular probabilidad

8.1.2 DISTRIBUCIÓN DE PROBABILIDAD ACUMULADA CONJUNTA

Definición: Distribución de Probabilidad Acumulada Conjunta

$$F(x,y) = P(X \leq x, Y \leq y) = \sum_{s \leq x} \sum_{t \leq y} f(s,t)\ ,\ -\infty < x, y < \infty$$

Ejemplo
Suponga que **X, Y** son variables aleatorias discretas cuya función de distribución de probabilidad está descrita en el siguiente cuadro:

		X = 0	X = 1	X = 2
Y	1	0.1	0.2	0.05
	2	0.3	0.1	0.25

Valores de f(x, y)

a) Verifique que **f(x, y)** cumple las propiedades **1) y 2)**

Por simple observación en el cuadro con los valores de **f(x,y)**

b) Determine la probabilidad que **X=0** y que **Y=2**

P(X=0, Y=2) = f(0, 2) = 0.3

c) Calcule la probabilidad que **X>0** y que **Y=1**

P(X>0, Y=1) = f(1,1) + f(2,1) = 0.2 + 0.05 = 0.25

Ejemplo
Determine el valor de **k** para que la función
f(x,y) = **kxy**, **x** = **1, 2, 3**; **y** = **1, 2**
Pueda usarse como una función de probabilidad conjunta con las variables **X, Y**

Si es una función de probabilidad debe cumplir la propiedad $\sum_x \sum_y f(x,y) = 1$

Tabulación de los valores de **f(x, y)**

x	y	f(x,y)
1	1	k
1	2	2k
2	1	2k
2	2	4k
3	1	3k
3	2	6k

Entonces: $\sum_{x=1}^{3} \sum_{y=1}^{2} f(x,y) = k + 2k + 2k + 4k + 3k + 6k = 18k = 1 \Rightarrow \mathbf{k = 1/18}$

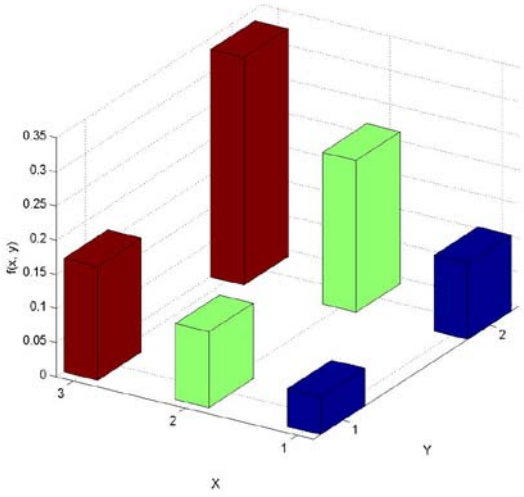
Así, la función de distribución de probabilidad conjunta es

$$\mathbf{f(x, y)} = \frac{\mathbf{1}}{\mathbf{18}}\mathbf{xy}, \quad \mathbf{x=1, 2, 3; \quad y=1, 2; \quad cero} \text{ para otros } \mathbf{(x, y)}$$

Se puede expresar en forma tabular

		X = 1	X = 2	X = 3
Y	1	1/18	2/18	3/18
	2	2/18	4/18	6/18

Una representación gráfica en tres dimensiones:



8.1.3 DISTRIBUCIONES DE PROBABILIDAD MARGINAL

Cuando se estudian más de una variable aleatoria en forma conjunta, puede ser de interés conocer la distribución de probabilidad de las variables aleatorias individualmente. Estas funciones se denominan distribuciones marginales.

Definiciones: Distribuciones de Probabilidad Marginal con Variables Aleatorias Discretas

Sean **X,Y:** Variables aleatorias discretas y
f(x,y): Función de probabilidad conjunta.
Entonces

$\mathbf{g(x)} = \sum_y \mathbf{f(x,y)}$ Distribución marginal de **X**

$\mathbf{h(y)} = \sum_x \mathbf{f(x,y)}$ Distribución marginal de **Y**

Las distribuciones marginales **g(x), h(y)** son funciones de probabilidad de las variables aleatorias **X, Y** separadamente. Estas funciones deben cumplir las propiedades de una función de probabilidad y pueden ser usadas para calcular probabilidad para cada variable.

1) $g(x) \geq 0,\ h(y) \geq 0,\ x, y \in \Re$
2) $\sum_y g(x) = 1,\ \sum_x h(y) = 1$
3) $P(X=x) = g(x),\ P(Y=y) = h(y)$

Ejemplo.
Suponga que **X, Y** son variables aleatorias discretas cuya función de distribución de probabilidad conjunta está descrita en el siguiente cuadro

Y \ X	0	1	2
1	0.1	0.2	0.05
2	0.3	0.1	0.25

a) Encuentre las distribuciones marginales tabularmente

Se suman los valores de filas y columnas y se escriben en los márgenes. Estos valores representan la probabilidad de una variable, incluyendo todos los valores de la otra variable.

Y \ X	0	1	2	h(y)
1	0.1	0.2	0.05	**0.35**
2	0.3	0.1	0.25	**0.65**
g(x)	**0.4**	**0.3**	**0.3**	**1**

b) Calcule **P(X=1)**

$P(X=1) = g(1) = 0.3$

c) Calcule **P(Y=2)**

$P(Y=2) = h(2) = 0.65$

Ejemplo
Sean **X, Y** variables aleatorias con la siguiente función de probabilidad conjunta

$$f(x,y) = \frac{1}{18}xy,\ x = 1, 2, 3;\ \ y = 1, 2$$

a) Encuentre las distribuciones marginales analíticamente

$$g(x) = \sum_y f(x,y) = \sum_{y=1}^{2} \frac{1}{18}xy = \frac{x}{18}\sum_{y=1}^{2} y = \frac{x}{18}(1+2) = \frac{x}{6}, \qquad x = 1, 2, 3$$

$$h(y) = \sum_x f(x,y) = \sum_{x=1}^{3} \frac{1}{18}xy = \frac{y}{18}\sum_{x=1}^{3} x = \frac{y}{18}(1+2+3) = \frac{y}{3}, \qquad y = 1, 2$$

b) Calcule **P(X=3), P(Y=1)**

$P(X=3) = g(3) = 1/2$

$P(Y=1) = h(1) = 1/3$

En los ejemplos anteriores se puede verificar que las distribuciones marginales **g(x)** y **h(y)** cumplen las propiedades **1)**, **2),** tabularmente o analíticamente.

8.1.4 DISTRIBUCIONES DE PROBABILIDAD CONDICIONAL

Cuando se estudian más de una variable aleatoria en forma conjunta, puede ser de interés conocer la distribución de probabilidad de cada variable aleatoria dado que la otra variable aleatoria toma un valor específico. Estas funciones se denominan Distribuciones Condicionales

Recordemos la fórmula de probabilidad condicional para eventos,

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad P(B) \neq 0$$

Definamos los eventos **A, B** de la siguiente manera

A: X=x
B: Y=y

Siendo **X, Y** variables aleatorias discretas con distribución de probabilidad conjunta **f(x,y),** Entonces,

$$P(X=x|Y=y) = \frac{P(X = x, Y = y)}{P(Y = y)}$$

Que se puede expresar con la notación establecida para las distribuciones conjuntas:

$$f(x|y) = \frac{f(x,y)}{h(y)}$$

La función **f(x|y)** también satisface las propiedades de las funciones de probabilidad

Definiciones: Distribuciones de Probabilidad Condicional

Sean **X, Y:** Variables aleatorias discretas
f(x, y): Distribución de probabilidad conjunta
Entonces,

$$f(x|y) = \frac{f(x,y)}{h(y)}$$ Es la distribución condicional de **X** dado que **Y=y**

$$f(y|x) = \frac{f(x,y)}{g(x)}$$ Es la distribución condicional de **Y** dado que **X=x**

Las distribuciones condicionales **f(x|y), f(y|x)** son funciones de probabilidad de **X, Y**. Estas funciones cumplen las propiedades establecidas y pueden usarse para calcular probabilidad condicional.

1) $f(x|y) \geq 0, x \in \Re, \quad f(y|x) \geq 0, \ y \in \Re$
2) $\sum_x f(x \mid y) = 1, \quad \sum_y f(y \mid x) = 1$

Ejemplo.
Suponga que **X, Y** son variables aleatorias discretas cuya función de distribución de probabilidad está descrita en el siguiente cuadro:

Y \ X	0	1	2	h(y)
1	0.1	0.2	0.05	0.35
2	0.3	0.1	0.25	0.65
g(x)	0.4	0.3	0.3	1

Calcule la probabilidad condicional **P(X=2 | Y=1)**

$$P(X=2 \mid Y=1) = f(2 \mid 1) = \frac{f(2,1)}{h(1)} = \frac{0.05}{0.35} = 0.1429$$

Ejemplo

Sean **X, Y** variables aleatorias con la siguiente función de probabilidad conjunta

$$f(x,y) = \frac{1}{18}xy, \quad x = 1, 2, 3; \quad y = 1, 2; \quad \textbf{cero} \text{ para otro } (x, y)$$

a) Encuentre las distribuciones condicionales, analíticamente

Previamente se obtuvieron las distribuciones marginales:

$$g(x) = x/6, \quad x = 1, 2, 3$$

$$h(y) = y/3, \quad y = 1, 2$$

Por lo tanto, para este problema:

$$f(x \mid y) = \frac{f(x,y)}{h(y)} = \frac{\frac{1}{18}xy}{\frac{y}{3}} = \frac{x}{6}$$

Significa que **X** no depende de **Y**

$$f(y \mid x) = \frac{f(x,y)}{g(x)} = \frac{\frac{1}{18}xy}{\frac{x}{6}} = \frac{y}{3}$$

Significa que **Y** no depende de **X**

b) Calcule la probabilidad condicional **P(X=1 | Y=2)**

$$P(X=x \mid Y=y) = f(x \mid y) = \frac{x}{6} \Rightarrow P(X=1 \mid Y=2) = f(1 \mid 2) = 1/6$$

8.1.5 VARIABLES ALEATORIAS DISCRETAS INDEPENDIENTES

Definición: Variables Aleatorias Discretas Independientes

Se dice que **X, Y** son variables aleatorias discretas estadísticamente independientes si y solo si **f(x,y)** = **g(x) h(y),** en cada punto **(x, y)**.

Demostración

Sean **X,Y** variables aleatorias discretas y **f(x,y)** su distribución de probabilidad conjunta.

Su distribución condicional **f(x|y)** es:

$$f(x|y) = \frac{f(x,y)}{h(y)}$$

Su distribución marginal **g(x)** es:

$$g(x) = \sum_y f(x,y)$$

Sustituimos la distribución condicional en la distribución marginal:

$$g(x) = \sum_y f(x \mid y)h(y)$$

Supongamos que **f(x|y)** no depende de **y**. Esto significa que la expresión **f(x|y)** no contendrá a la variable **y**. Por lo tanto, puede salir de la sumatoria:

$$g(x) = f(x|y)\sum_y h(y)$$

Pero $\sum_y h(y) = 1$, pues **h(y)** es también una función de distribución de probabilidad .

Entonces **f(x|y) = g(x)**

Sustituyendo en la distribución condicional en el inicio, se obtiene **f(x,y) = g(x) h(y)**

Ejemplo

Sean **X, Y** variables aleatorias discretas cuya función de distribución de probabilidad conjunta es $f(x,y) = \frac{1}{18}xy,\ x=1, 2, 3;\ \ y=1, 2$

Pruebe que **X, Y** son variables aleatorias estadísticamente independientes

Solución

Se tienen las distribuciones marginales

$$g(x) = \frac{x}{6}, \quad x = 1, 2, 3$$

$$h(y) = \frac{y}{3}, \quad y = 1, 2$$

Entonces

$$g(x)h(y) = (\frac{x}{6})(\frac{y}{3}) = \frac{1}{18}xy = f(x,y),\ x = 1, 2, 3;\ y = 1, 2$$

Por lo tanto, **X, Y** son variables aleatorias estadísticamente independientes.

Siendo **X, Y** variables aleatorias estadísticamente independientes se cumple también que

$$f(x \mid y) = \frac{f(x,y)}{h(y)} = \frac{\frac{1}{18}xy}{\frac{y}{3}} = \frac{x}{6} = g(x), \qquad f(y \mid x) = \frac{f(x,y)}{g(x)} = \frac{\frac{1}{18}xy}{\frac{x}{6}} = \frac{y}{3} = h(y)$$

8.2 CASO DISCRETO TRIVARIADO

Las definiciones para distribuciones bivariadas pueden extenderse a más variables.

El siguiente ejemplo es una referencia para los conceptos relacionados

Ejemplo

Sea **V** un vector aleatorio discreto cuyos componentes son las variables aleatorias **X, Y, Z** con distribución de probabilidad conjunta

$$f(x,y,z) = \begin{cases} kx^2(y-z); & x = 1,2,3; y = 3,4; z = 1,2 \\ 0; & \text{para el resto de } x,y,z \end{cases}$$

a) Tabule **f(x,y,z)** para cada valor de los componentes

Primero debe determinarse **k** con la propiedad de las funciones de probabilidad:

$$\sum_{x=1}^{3}\sum_{y=3}^{4}\sum_{z=1}^{2} f(x,y,z) = 1$$

$$\sum_{x=1}^{3}\sum_{y=3}^{4}\sum_{z=1}^{2} kx^2(y-z) = k\sum_{x=1}^{3}x^2\sum_{y=3}^{4}\sum_{z=1}^{2}(y-z) =$$

$$k[(1^2+2^2+3^2)((3-1)+(3-2)+(4-1)+(4-2))] = k(14)(8) = 1 \Rightarrow k = \frac{1}{112}$$

Entonces la distribución conjunta es:

$$f(x,y,z) = \begin{cases} \frac{1}{112}x^2(y-z); & x = 1,2,3; y = 3,4; z = 1,2 \\ 0; & \text{para el resto de } x,y,z \end{cases}$$

Tabulación

x	y	z	f(x,y,z)
1	3	1	2/112
1	3	2	1/112
1	4	1	3/112
1	4	2	2/112
2	3	1	8/112
2	3	2	4/112
2	4	1	12/112
2	4	2	8/112
3	3	1	18/112
3	3	2	9/112
3	4	1	27/112
3	4	2	18/112

b) Encuentre las distribuciones marginales univariadas

Analíticamente, dando por entendido el dominio de cada función

$$f(x)=\sum_{y=3}^{4}\sum_{z=1}^{2}f(x,y,z)=\sum_{y=3}^{4}\sum_{z=1}^{2}\frac{1}{112}x^2(y-z)=\frac{1}{112}x^2\sum_{y=3}^{4}\sum_{z=1}^{2}(y-z)=\frac{8}{112}x^2$$

$$f(y)=\sum_{x=1}^{3}\sum_{z=1}^{2}\frac{1}{112}x^2(y-z)=\frac{1}{112}\sum_{x=1}^{3}x^2\sum_{z=1}^{2}(y-z)=\frac{1}{112}(14)(y-1+y-2)=\frac{14}{112}(2y-3)$$

$$f(z)=\sum_{x=1}^{3}\sum_{y=3}^{4}\frac{1}{112}x^2(y-z)=\frac{1}{112}\sum_{x=1}^{3}x^2\sum_{y=3}^{4}(y-z)=\frac{14}{112}(-2z+7)$$

Tabularmente, sumando el contenido de la tabla de la distribución conjunta

x	1	2	3
f(x)	8/112	32/112	72/112

y	3	4
f(y)	42/112	70/112

z	1	2
f(z)	70/112	42/112

c) Encuentre las distribuciones marginales bivariadas

Analíticamente, dando por entendido el dominio de cada función

$$f(x,y)=\sum_{z=1}^{2}f(x,y,z)=\sum_{z=1}^{2}\frac{1}{112}x^2(y-z)=\frac{1}{112}x^2\sum_{z=1}^{2}(y-z)=\frac{x^2}{112}(2y-3)$$

$$f(x,z)=\sum_{y=3}^{4}f(x,y,z)=\sum_{y=3}^{4}\frac{1}{112}x^2(y-z)=\frac{1}{112}x^2\sum_{y=3}^{4}(y-z)=\frac{x^2}{112}(-2z+7)$$

$$f(y,z)=\sum_{x=1}^{3}f(x,y,z)=\sum_{x=1}^{3}\frac{1}{112}x^2(y-z)=\frac{1}{112}(y-z)\sum_{x=1}^{3}x^2=\frac{14}{112}(y-z)$$

Tabularmente, sumando el contenido de la tabla de la distribución conjunta

f(x,y)

y \ x	1	2	3
3	3/112	12/112	27/112
4	5/112	20/112	45/112

f(x,z)

z \ x	1	2	3
1	5/112	20/112	45/112
2	3/112	12/112	27/112

f(y,z)

z \ y	3	4
1	28/112	42/112
2	14/112	28/112

Se puede observar, analítica o tabularmente, que

$$f(x,y) = f(x)\, f(y)$$
$$f(x,z) = f(x)\, f(z)$$
$$f(y,z) \neq f(y)\, f(z)$$

Entonces,

X, Y son variables aleatorias estadísticamente independientes
X, Z son variables aleatorias estadísticamente independientes
Y, Z son variables aleatorias estadísticamente **no** independientes

d) Encuentre las distribuciones condicionales

Analíticamente, dando por entendido el dominio de cada función

$$f(X = x \mid Y = y) = f(x \mid y) = \frac{f(x,y)}{f(y)} = \frac{\frac{x^2}{112}(2y-3)}{\frac{14}{112}(2y-3)} = \frac{x^2}{14} = \frac{8x^2}{112}$$

$\Rightarrow$ $f(x|y) = f(x)$ pues **X, Y** son estadísticamente independientes

También se puede verificar que

$f(x|z) = f(x)$ pues **X, Z** son estadísticamente independientes

Mientras que para $f(y|\, z)$, se debe encontrar la relación

$$f(y \mid z) = \frac{f(y,z)}{f(z)} = \frac{\frac{14}{112}(y-z)}{\frac{14}{112}(-2z+7)} = \frac{y-z}{-2z+7}$$

Tabularmente:

y	f(y|z=1)	f(y|z=2)
3	2/5	1/3
4	3/5	2/3

8.2.1 EJERCICIOS

Si la distribución de probabilidad conjunta de las variables aleatorias discretas **X, Y** está dada por

$$f(x,y) = \frac{1}{30}(x+y)\ ,\ x=0,\ 1,\ 2,\ 3; \quad y=0,\ 1,\ 2$$

a) Verifique que es una función de probabilidad
b) Construya una tabla con todos los valores de probabilidad
c) Obtenga tabularmente la distribución marginal de **X**
d) Exprese mediante una fórmula la distribución marginal de **Y**
e) Obtenga la distribución condicional de **X** dado que **Y=1**
f) Obtenga la distribución condicional de **Y** dado que **X=2**
g) Determine si las dos variables aleatorias son estadísticamente independientes

MATLAB

Manejo simbólico de una distribución trivariada continua (comparar con el ejemplo)

```
>> syms x y z                                  Definición de variables simbólicas X, Y, Z
>> f=x*(y+z);                                  Función de densidad trivariada f(x,y,z)
>> p=int(int(int(f,y,0,z),z,0,1), x,0,2)       Verificar que f es función de densidad
   p =
   1
>> p=int(int(int(f,y,0.1,0.4),z,0.5,0.8), x,1.2,1.8)   Calcular P(0.1<Y<0.4, 0.5<Z<0.8,1.2<Z<1.8)
   p =
   729/10000

>> fx=int(int(f,y,0,z),z,0,1)                  Densidad marginal f(x)
   fx =
   1/2*x
>> fy=int(int(f,x,0,2),z,y,1)                  Densidad marginal f(y)
   fy =
   2*y*(1-y)+1-y^2
>> fy=expand(fy)                               Expansión algebraica
   fy =
   2*y-3*y^2+1
>> fz=int(int(f,y,0,z),x,0,2)                  Densidad marginal f(z)
   fz =
   3*z^2
>> fxy=int(f,z,y,1)                            Densidad marginal f(x,y)
   fxy =
   x*y*(1-y)+1/2*x*(1-y^2)
>> fxy=expand(fxy)
   fxy =
   x*y-3/2*x*y^2+1/2*x
>> fxz=int(f,y,0,z)                            Densidad marginal f(x,z)
   fxz =
   3/2*x*z^2
>> fyz=int(f,x,0,2)                            Densidad marginal f(y,z)
   fyz =
   2*y+2*z
>> r=expand(fxy)==expand(fx*fy)                Verificar que X, Y son variables independientes
   r =
        1
>> r=expand(fxz)==expand(fx*fz)                Verificar que X, Z son variables independientes
   r =
        1
>> r=expand(fyz)==expand(fy*fz)                Verificar que Y, Z no son var. independientes
   r =
        0
>> p=int(fx, 1.2, 1.8)                         Calcular la marginal P(1.2<X<1.8)
   p =
   9/20
>> p=int(int(fxy, x, 1.2, 1.8),y,0.2, 0.8)     Calcular la marginal P(1.2<X<1.8, 0.2<Y<0.8)
   p =
   783/2500
```

8.3 CASO CONTINUO BIVARIADO

Algunos experimentos estadísticos pueden incluir más de una variable aleatoria contínua, las cuales pueden actuar en forma conjunta, y es de interés determinar la probabilidad correspondiente a los valores que estas variables puedan tomar.

8.3.1 DENSIDAD DE PROBABILIDAD CONJUNTA

Definición: Función de Densidad de Probabilidad Conjunta

Sean **X, Y:** Variables aleatorias continuas.
Su función de densidad de probabilidad conjunta se escribe **f(x,y)**

Esta función debe satisfacer las siguientes propiedades

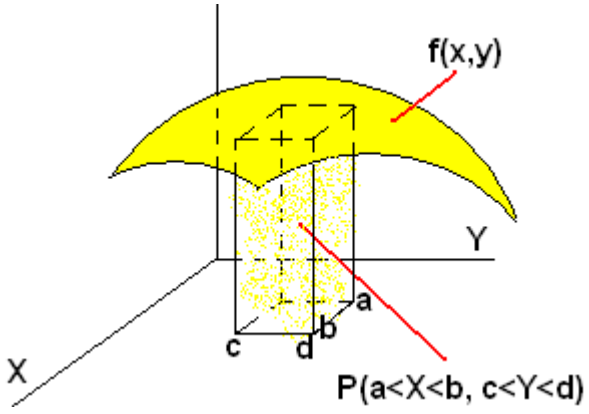
1) $f(x,y) \geq 0,\ x \in \Re,\ y \in \Re$
2) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1$

La función de densidad de probabilidad conjunta puede usarse para calcular probabilidad

3) $P(a \leq X \leq b,\ c \leq Y \leq d) = \int_{c}^{d}\int_{a}^{b} f(x,y)dxdy$

La función de densidad de probabilidad de dos variables aleatorias continuas **X, Y** es una superficie en el espacio. El volumen debajo de esta superficie sobre el plano **X-Y** es igual a **1**.

La probabilidad **P(a≤X≤b, c≤Y≤d)** es igual a la porción del volumen debajo de la superficie **f(x,y)** y sobre el rectángulo **a≤X≤b, c ≤Y≤d**



8.3.2 DISTRIBUCIÓN DE PROBABILIDAD ACUMULADA CONJUNTA

Definición: Distribución de Probabilidad Acumulada Conjunta

$$P(X \leq x,\ Y \leq y) = F(x,y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f(u,v)dudv \quad -\infty < x, y < +\infty$$

Ejemplo.
Suponga que el tiempo semanal de mantenimiento de una máquina depende de dos variables aleatorias continuas medidas en horas:

X: duración del mantenimiento mecánico
Y: duración el mantenimiento eléctrico

Suponga que la densidad de probabilidad conjunta es

$$f(x,y) = \begin{cases} \frac{2}{3}(x+2y), & 0 \le x, y \le 1 \\ 0, & \text{otros } x, y \end{cases}$$

a) Verifique que **f(x, y)** es una función de densidad de probabilidad

1) $f(x,y) \ge 0,\ x \in \Re,\ y \in \Re.$

2) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = \int_0^1\int_0^1 \frac{2}{3}(x+2y)dxdy = \frac{2}{3}\int_0^1\int_0^1 (x+2y)dxdy$$

$$= \frac{2}{3}\int_0^1 [\frac{x^2}{2} + 2xy]_0^1 dy = \frac{2}{3}\int_0^1 (\frac{1}{2} + 2y)dy = \frac{2}{3}[\frac{y}{2} + y^2]_0^1 = 1$$

b) Calcule la probabilidad que en alguna semana, el mantenimiento mecánico dure menos de **15** minutos y el mantenimiento eléctrico dure más de **30** minutos

$$P(X \le 1/4,\ Y \ge 1/2) = \int_{1/2}^{1}\int_{0}^{1/4} \frac{2}{3}(x+2y)dxdy = 13/96$$

8.3.3 DENSIDAD DE PROBABILIDAD MARGINAL

Cuando se estudian más de una variable aleatoria en forma conjunta, puede ser de interés conocer la distribución de probabilidad de las variables aleatorias individualmente. Estas funciones se denominan densidades marginales

Definiciones: Densidades de Probabilidad Marginal

Sean **X,Y:** Variables aleatorias continuas
f(x,y): Función de densidad de probabilidad conjunta.
Entonces,

$g(x) = f(x) = \int_{-\infty}^{\infty} f(x,y)dy$ Densidad de probabilidad marginal de **X**

$h(y) = f(y) = \int_{-\infty}^{\infty} f(x,y)dx$ Densidad de probabilidad marginal de **Y**

Para cada variable la densidad marginal se obtiene integrando la función de probabilidad sobre la otra variable.

Las densidades marginales **g(x), h(y)** son funciones de probabilidad de **X, Y** en forma separada. Estas funciones deben cumplir las propiedades respectivas:

1) $g(x) \ge 0,\ x \in \Re, \quad h(y) \ge 0,\ y \in \Re$

2) $\int_{-\infty}^{\infty} g(x)dx = 1, \quad \int_{-\infty}^{\infty} h(y)dy = 1$

Las densidades marginales pueden usarse para calcular probabilidad de cada variable.

Ejemplo.
En el problema del mantenimiento de la máquina, en donde **X, Y** es tiempo en horas

a) Encuentre las densidades marginales

$$g(x) = \int_{-\infty}^{\infty} f(x,y)dy = \int_0^1 \frac{2}{3}(x+2y)dy = \frac{2}{3}\left[xy + y^2\right]_0^1 = \frac{2}{3}(x+1), \quad 0 \le x \le 1$$

$$h(y) = \int_{-\infty}^{\infty} f(x,y)dx = \int_0^1 \frac{2}{3}(x+2y)dx = \frac{2}{3}\left[\frac{x^2}{2} + 2yx\right]_0^1 = \frac{1}{3} + \frac{4y}{3}, \quad 0 \le y \le 1$$

b) Calcule **P(0.25 ≤ X ≤ 0.75)**

$$P(0.25 \le X \le 0.75) = \int_{0.25}^{0.75} g(x)dx = \int_{0.25}^{0.75} \frac{2}{3}(x+1)dx = 0.5$$

8.3.4 DENSIDAD DE PROBABILIDAD CONDICIONAL

Cuando se estudian más de una variable aleatoria en forma conjunta, puede ser de interés conocer la distribución de probabilidad de cada variable aleatoria dado que la otra variable aleatoria tiene un valor específico. Estas funciones se denominan densidades condicionales.

Recordemos la fórmula de probabilidad condicional para eventos

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad P(B) \neq 0$$

Definamos los eventos **A, B** de la siguiente manera

A: $X \in R_x$

B: $Y \in R_y$

Siendo **X, Y** variables aleatorias continuas con distribución de probabilidad conjunta **f(x,y)**, mientras que R_x, R_y son regiones arbitrarias.
Entonces,

$$P(X \in R_x | Y \in R_y) = \frac{P(X \in R_x, Y \in R_y)}{P(Y \in R_y)}$$

Que se puede expresar con la notación establecida para las distribuciones conjuntas:

$$f(x|y) = \frac{f(x,y)}{h(y)}$$

La función **f(x|y)** también satisface las propiedades de las funciones de probabilidad

Definiciones: Densidades de Probabilidad Condicional

Sean **X, Y:** Variables aleatorias continuas
f(x, y): Densidad de probabilidad conjunta
Entonces,

$f(x|y) = \frac{f(x,y)}{h(y)}$ Es la densidad condicional de **X** dada **h(y)**

$f(y|x) = \frac{f(x,y)}{g(x)}$ Es la densidad condicional de **Y** dada **g(x)**

Las densidades condicionales **f(x|y), f(y|x)** son funciones de probabilidad. Estas funciones cumplen las propiedades establecidas y pueden usarse para calcular probabilidad condicional.

1) $f(x|y) \ge 0,\ x \in \Re, \quad f(y|x) \ge 0,\ y \in \Re$

2) $\int_{-\infty}^{\infty} f(x \mid y)dx = 1, \quad \int_{-\infty}^{\infty} f(y \mid x)dy = 1$

Ejemplo.
En el problema del mantenimiento de la máquina, en donde **X, Y** es tiempo en horas

a) Encuentre la densidad condicional **f(y|x)**

$$f(y|x) = \frac{f(x,y)}{g(x)} = \frac{\frac{2}{3}(x+2y)}{\frac{2}{3}(x+1)} = \frac{x+2y}{x+1}, \quad 0 \le x, y \le 1$$

b) Calcule la probabilidad que el mantenimiento eléctrico **Y** dure menos de 15 minutos dado que el mantenimiento mecánico **X** duró 30 minutos

$$P(Y \le 0.25|X = 0.5) = \int_0^{0.25} f(y \mid 0.5)dy = \int_0^{0.25} \frac{0.5+2y}{0.5+1}dy = 0.125$$

8.3.5 VARIABLES ALEATORIAS CONTINUAS INDEPENDIENTES

Definición: Variables Aleatorias Continuas Independientes

Se dice que **X, Y** son variables aleatorias continuas estadísticamente independientes si y solo si **f(x,y) = g(x) h(y)** en el dominio de **X, Y**

Demostración

Sean **X,Y** variables aleatorias continuas y **f(x,y)** su densidad de probabilidad conjunta.
La densidad condicional **f(x|y)** es:

$$f(x|y) = \frac{f(x,y)}{h(y)}$$

Y la densidad marginal **g(x)** es:

$$g(x) = \int_{-\infty}^{\infty} f(x,y)dy$$

Sustituyendo la densidad condicional en la densidad marginal:

$$g(x) = \int_{-\infty}^{\infty} f(x \mid y)h(y)dy$$

Supongamos que **f(x|y)** no depende de **y**. Esto significa que la expresión **f(x|y)** no contendría a la variable **y**. Por lo tanto, puede salir del integral:

$$g(x) = f(x|y) \int_{-\infty}^{\infty} h(y)dy$$

Pero $\int_{-\infty}^{\infty} h(y)dy = 1$, pues **h(y)** es también una función de densidad de probabilidad .

Entonces **g(x) = f(x|y).**

Sustituyendo en la densidad condicional inicial se obtiene **f(x,y) = g(x) h(y)**

Ejemplo

Sea **[X, Y]** un vector aleatorio bivariado cuya densidad de probabilidad conjunta es:
f(x,y) = **kxy**, **0 ≤ x, y ≤ 1**, **cero** para otro **(x,y)**

a) Encuentre el valor de **k** para que sea una función de probabilidad

El dominio: **0 ≤ x, y ≤ 1** es equivalente a: **0 ≤ x ≤ 1, 0 ≤ y ≤1**

Se debe cumplir que $\int_0^1\int_0^1 kxydxdy = 1$

$$k\int_0^1\int_0^1 kxydxdy = k\int_0^1[\frac{x^2}{2}]_0^1 ydy = \frac{k}{2}\int_0^1 ydy = \frac{k}{2}[\frac{y^2}{2}]_0^1 = \frac{k}{4} = 1 \Rightarrow k = 4$$

$\Rightarrow$ **f(x,y)** = **4xy**, **0 ≤ x, y ≤ 1**, **cero** para otro **(x,y)**

b) Calcule la probabilidad **P(X < 0.5, Y > 0.75)**

$$P(X < 0.5, Y > 0.75) = \int_{0.75}^1\int_0^{0.5} 4xydxdy = 4\int_{0.75}^1[\frac{x^2}{2}]_0^{0.5} ydy = \frac{1}{2}\int_{0.75}^1 ydy = 0.1094$$

c) Encuentre las densidades marginales

$$f(x) = \int_0^1 f(x,y)dy = \int_0^1 4xydy = 4x[\frac{y^2}{2}]_0^1 = 2x, \quad 0 \le x \le 1$$

$$f(y) = \int_0^1 f(x,y)dx = \int_0^1 4xydx = 4y[\frac{x^2}{2}]_0^1 = 2y, \quad 0 \le y \le 1$$

d) Determine si **X, Y** son variables aleatorias independientes

Se debe cumplir que **f(x,y)** = **f(x)f(y)** para todo **(x,y)**

f(x,y) = **4xy**, **f(x)f(y)** = **(2x)(2y) = 4xy = f(x, y)** $\Rightarrow$ **X, Y** son independientes

e) Encuentre las densidades condicionales

f(x|y) = **f(x,y)/f(y)** = **4xy/2y=2x** = **f(x)** Resultado previsto pues **X,Y** son independientes
0 ≤ x ≤ 1

f(y|x) = **f(x,y)/f(x)=4xy/2x=2y** = **f(y)** Resultado previsto pues **X,Y** son independientes
0 ≤ y ≤1

Ejemplo

Sea **[X, Y]** un vector aleatorio bivariado cuya densidad de probabilidad conjunta es:
f(x,y) = **kxy**, **0 ≤ x ≤ y ≤ 1**, **cero** para otro **(x,y)**

a) Encuentre el valor de **k** para que **f(x, y)** sea una función de probabilidad

El dominio: **0 ≤ x ≤ y ≤ 1** es equivalente a: **0 ≤ x ≤ y, 0 ≤ y ≤ 1**

Se debe cumplir que $\int_0^1\int_0^y kxydxdy = 1$

$$\int_0^1\int_0^y kxydxdy = k\int_0^1[\frac{x^2}{2}]_0^y ydy = \frac{k}{2}\int_0^1 y^3dy = \frac{k}{2}[\frac{y^4}{4}]_0^1 = \frac{k}{8} = 1 \Rightarrow k = 8$$

$\Rightarrow$ **f(x,y)** = **8xy**, **0 ≤ x ≤ y ≤ 1**, **cero** para otro **(x,y)**

b) Encuentre las densidades marginales

$$f(x) = \int_x^1 f(x,y)dy = \int_x^1 8xydy = 8x[\frac{y^2}{2}]_x^1 = 4(x - x^3), \quad 0 \le x \le 1$$

$$f(y) = \int_0^y f(x,y)dx = \int_0^y 8xydx = 8y[\frac{x^2}{2}]_0^y = 4y^3, \quad 0 \le y \le 1$$

c) Determine si **X, Y** son variables aleatorias independientes

Se debe cumplir que **f(x,y)** = **f(x)f(y)** para todo **(x,y)**

f(x,y) = **8xy**, **f(x)f(y)** = $4(x - x^3)(4y^3) \neq 8xy \Rightarrow$ **X, Y** no son independientes

c) Encuentre las densidades condicionales

$$f(x|y) = f(x,y)/f(y) = \frac{8xy}{4y^3} = \frac{2x}{y^2}, \qquad 0 \le x \le y \le 1,\ f(y) \neq 0$$

$$f(y|x) = f(x,y)/f(x) = \frac{8xy}{4(x - x^3)} = \frac{2y}{1 - x^2}, \quad 0 \le x \le y \le 1,\ f(x) \neq 0$$

8.4 CASO CONTINUO TRIVARIADO

Las definiciones para distribuciones bivariadas pueden extenderse a más variables.

El siguiente ejemplo es una referencia para revisar los conceptos relacionados

Ejemplo

Sea **[X, Y, Z]** un vector aleatorio trivariado cuya distribución de probabilidad conjunta es:

f(x,y,z) = **kx(y+z), 0 < x < 2, 0 < y < z < 1,** cero para otro **(x,y,z)**

a) Encuentre el valor de **k** para que **f(x, y, z)** sea una función de probabilidad

El dominio: **0 < y < z < 1** es equivalente a: **0 < y < z, 0 < z < 1**

Se debe cumplir que $\int_0^2\int_0^1\int_0^z kx(y+z)\,dydzdx = 1$

$$\int_0^2\int_0^1\int_0^z kx(y+z)\,dydzdx = k\int_0^2 x\int_0^1\int_0^z (y+z)\,dydzdx = k\int_0^2 x\int_0^1 [\frac{y^2}{2} + yz]_0^z\,dzdx$$

$$= k\int_0^2 x\int_0^1 [\frac{y^2}{2} + yz]_0^z\,dzdx = k\int_0^2 x\int_0^1 (\frac{z^2}{2} + z^2)\,dzdx = \frac{3k}{2}\int_0^2 x\int_0^1 z^2\,dzdx$$

$$= \frac{3k}{2}\int_0^2 [\frac{z^3}{3}]_0^1 x\,dx = \frac{k}{2}\int_0^2 x\,dx = \frac{k}{2}[\frac{x^2}{2}]_0^2 = \frac{k}{2}(\frac{4}{2}) = 1 \quad \Rightarrow \quad k = 1$$

$\Rightarrow$ **f(x,y,z)** = **x(y+z), 0 < x < 2, 0 < y < z < 1,** cero para otro **(x,y,z)**

b) Encuentre las distribuciones marginales univariadas

$$f(x) = \int_0^1\int_0^z x(y+z)\,dydz = x\int_0^1\int_0^z (y+z)\,dydz = x\int_0^1 [\frac{y^2}{2} + yz]_0^z\,dz$$

$$= x\int_0^1 (\frac{z^2}{2} + z^2)\,dz = \frac{3x}{2}\int_0^1 z^2\,dz = \frac{3x}{2}[\frac{z^3}{3}]_0^1 = \frac{3x}{2}(\frac{1}{3}) = \frac{x}{2}, \quad 0 < x < 2$$

$$f(y) = \int_y^1\int_0^2 x(y+z)\,dxdz = \int_y^1 (y+z)\int_0^2 x\,dxdz = \int_y^1 (y+z)[\frac{x^2}{2}]_0^2\,dz$$

$$= 2\int_y^1 (y+z)\,dz = 2[yz + \frac{z^2}{2}]_y^1 = 1 + 2y - 3y^2, \qquad 0 < y < 1$$

$$f(z) = \int_0^2\int_0^z x(y+z)\,dydx = \int_0^2 x\int_0^z (y+z)\,dydx = \int_0^2 x[\frac{y^2}{2} + zy]_0^z\,dx$$

$$= \frac{3}{2}\int_0^2 xz^2\,dx = \frac{3}{2}z^2[\frac{x^2}{2}]_0^2 = 3z^2, \qquad 0 < z < 1$$

b) Encuentre las distribuciones marginales bivariadas

$$f(x,y)=\int_y^1 x(y+z)dz = x[yz+\frac{z^2}{2}]_y^1 = x(y+\frac{1}{2}-y^2-\frac{y^2}{2}=\frac{x}{2}(1+2y-3y^2)$$

$$0<x<2,\quad 0<y<1$$

$$f(x,z)=\int_0^z x(y+z)dy = x[\frac{y^2}{2}+zy]_0^z = x(\frac{z^2}{2}+z^2)=\frac{3xz^2}{2},\quad 0<x<2,\quad 0<z<1$$

$$f(y,z)=\int_0^2 x(y+z)dx = (y+z)[\frac{x^2}{2}]_0^2 = 2(y+z),\qquad 0<y<z<1$$

c) Determine si **X, Y, Z** son variables aleatorias estadísticamente independientes

$f(x,y)=\frac{x}{2}(1+2y-3y^2)=f(x)f(y)$ $\Rightarrow$ **X, Y** son independientes

$f(x,z)=\frac{3xz^2}{2}=f(x)f(z)$ $\Rightarrow$ **X, Z** son independientes

$f(y,z)=2(y+z)$

$f(y)f(z)=(1+2y-3y^2)(3z^2)\neq f(y,z)$ $\Rightarrow$ **Y, Z** no son independientes

d) Verifique que **f(x)** es una función de densidad de probabilidad

$$\int_0^2 \frac{x}{2}dx=\frac{1}{2}[\frac{x^2}{2}]_0^2=\frac{1}{2}(\frac{2^2}{2})=1$$

e) Verifique que **f(x, z)** es una función de densidad de probabilidad

$$\int_0^2\int_0^1 \frac{3xz^2}{2}dzdx=\frac{3}{2}\int_0^2 x\int_0^1 z^2dzdx=\frac{3}{2}\int_0^2 x[\frac{z^3}{3}]_0^1dx=\frac{1}{2}\int_0^2 xdx=\frac{1}{2}[\frac{x^2}{2}]_0^2=1$$

8.4.1 EJERCICIOS

1) X_1 y X_2 tienen la función de densidad de probabilidad conjunta dada por

$$f(x_1,x_2)=\begin{cases} kx_1x_2, & 0\le x_1\le 1,\ 0\le x_2\le 1 \\ 0, & \text{para otros puntos}\end{cases}$$

a) Calcule el valor de **k** que hace que **f** sea una función de densidad de probabilidad
b) Calcule $P(X_1\le 0.75, X_2\ge 0.5)$

2) X_1 y X_2 tienen la función de densidad de probabilidad conjunta dada por

$$f(x_1,x_2)=\begin{cases} k(1-x_2), & 0\le x_1\le x_2\le 1 \\ 0, & \text{para otros puntos}\end{cases}$$

a) Calcule el valor de **k** que hace que **f** sea una función de densidad de probabilidad
b) Calcule $P(X_1\le 0.75, X_2\ge 0.5)$

3) Si la densidad de probabilidad conjunta de las variables aleatorias continuas **X, Y** está dada por

$$f(x,y)=\begin{cases} \frac{1}{4}(2x+y), & 0<x<1,\ 0<y<2 \\ 0, & \text{para otros valores}\end{cases}$$

a) Verifique que es una función de densidad de probabilidad
b) Obtenga la densidad marginal de **X**
c) Obtenga la densidad marginal de **Y**
d) Obtenga la densidad condicional de **X** dado que **Y=1**
e) Obtenga la densidad condicional de **Y** dado que **X=1/4**
f) Determine si **X, Y** son variables aleatorias estadísticamente independientes

8.5 MEDIA PARA VARIABLES ALEATORIAS CONJUNTAS CASO BIVARIADO

Definición: Media para Variables Aleatorias Conjuntas Caso Bivariado

Sean **X, Y:** Variables aleatorias discretas (o continuas)
f(x, y): Distribución (o densidad) de probabilidad conjunta

Sea **G(X,Y)**, alguna expresión con **X, Y**.

Si **X, Y** son variables aleatorias discretas, entonces

$$\mu_{G(X,Y)} = E[G(X,Y)] = \sum_X \sum_Y G(x,y)f(x,y)$$, es la media o valor esperado de **G(X,Y)**

Si **X, Y** son variables aleatorias continuas, entonces

$$\mu_{G(X,Y)} = E[G(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} G(x,y)f(x,y)dxdy$$, es la media o valor esperado de **G(X,Y)**

Ejemplo

Sean **X, Y** variables aleatorias discretas cuya función de distribución de probabilidad conjunta es $f(x,y) = \begin{cases} \frac{1}{18}xy, & x = 1,2,3;\ y = 1,2 \\ 0, & \text{otro } (x,y) \end{cases}$

Calcule la media de la suma **X + Y**

G(X,Y) = X + Y

$$E[G(X,Y)] = E(X+Y) = \sum_{x=1}^{3}\sum_{y=1}^{2}(x+y)\frac{1}{18}xy = \frac{1}{18}\sum_{x=1}^{3}\sum_{y=1}^{2}(x+y)xy$$

$$= \frac{1}{18}[(1+1)1+(1+2)2+(2+1)2+(2+2)4+(3+1)3+(3+2)6 = 4$$

Ejemplo

Sean **X, Y** variables aleatorias continuas cuya función de densidad de probabilidad conjunta es

$$f(x,y) = \begin{cases} \frac{2}{3}(x+2y), & 0 \le x,y \le 1 \\ 0, & \text{otro } (x,y) \end{cases}$$

Calcule la media de la suma **X + Y**:

G(X,Y) = X + Y

$$E[G(X,Y)] = E(X+Y) = \int_0^1\int_0^1 (x+y)f(x,y)dxdy = \int_0^1\int_0^1 (x+y)\frac{2}{3}(x+2y)dxdy = 7/6$$

8.5.1 MEDIA PARA VARIABLES ALEATORIAS CONJUNTAS CASOS ESPECIALES

Definición: Media para Variables Aleatorias Conjuntas Casos Especiales

Sean **X, Y** Variables aleatorias discretas (o continuas)
f(x, y) Distribución (o densidad) de probabilidad conjunta
g(x), h(y) Distribuciones (o densidades) marginales de **X** y **Y** respectivamente

Si **X, Y** son variables aleatorias discretas

Si **G(X,Y) = X**, entonces su media es

$$\mu_X = \mathbf{E(X)} = \sum_X \sum_Y \mathbf{x f(x,y)} = \sum_x \mathbf{x} \sum_y \mathbf{f(x,y)} = \sum_x \mathbf{x\ g(x)}$$

Si **G(X,Y) = Y**, entonces su media es

$$\mu_Y = \mathbf{E(Y)} = \sum_X \sum_Y \mathbf{y f(x,y)} = \sum_y \mathbf{y} \sum_x \mathbf{f(x,y)} = \sum_y \mathbf{y\ h(y)}$$

Si **X, Y** son variables aleatorias continuas

Si **G(X,Y) = X**, entonces su media es

$$\mu_X = \mathbf{E(X)} = \int_{-\infty}^{\infty} \mathbf{x\ g(x)dx}$$

Si **G(X,Y) = Y**, entonces su media es

$$\mu_Y = \mathbf{E(Y)} = \int_{-\infty}^{\infty} \mathbf{y\ h(y)dy}$$

8.6 COVARIANZA PARA VARIABLES ALEATORIAS CONJUNTAS: CASO BIVARIADO

La definición de varianza se extiende a variables aleatorias conjuntas y se denomina covarianza. Es una medida de la dispersión combinada de ambas variables.

Definición: Covarianza

Sean **X, Y** variables aleatorias discretas con distribución conjunta **f(x,y)**
Entonces, la **covarianza de X, Y** es

$$\sigma_{XY} = \mathbf{Cov(X,Y)} = \mathbf{E}\left[(X-\mu_X)(Y-\mu_Y)\right] = \sum_x \sum_y (x-\mu_X)(y-\mu_Y) f(x,y)$$

Sean **X, Y** variables aleatorias continuas con densidad conjunta **f(x,y)**
Entonces, la **covarianza de X, Y** es

$$\sigma_{XY} = \mathbf{Cov(X,Y)} = \mathbf{E}\left[(X-\mu_X)(Y-\mu_Y)\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\mu_X)(y-\mu_Y) f(x,y)dxdy$$

La siguiente fórmula es equivalente a la anterior y es de uso común para calcular la covarianza:

Definición: Fórmula alterna para la Covarianza

$$\sigma_{XY} = \mathbf{Cov(X,Y)} = \mathbf{E(XY)} - \mu_X\mu_Y$$ Variables aleatorias Discretas o Continuas

Demostración

$$\begin{aligned} \text{Cov}(X,Y) &= E[(X-\mu_X)(Y-\mu_Y)] = E[XY - X\mu_Y - Y\mu_X + \mu_X\mu_Y] \\ &= E(XY) - \mu_Y E(X) - \mu_X E(Y) + \mu_X\mu_Y \\ &= E(XY) - \mu_Y\mu_X - \mu_X\mu_Y + \mu_X\mu_Y = E(XY) - \mu_X\mu_Y \end{aligned}$$

Si **X** = **Y**, la covarianza se reduce a la varianza

$$\sigma_X^2 = \mathbf{V(X)} = \mathbf{E[(X-\mu_X)^2]} = \mathbf{E(X^2) - \mu_X^2}$$

Ejemplo
Sean **X, Y** variables aleatorias discretas cuya función de distribución de probabilidad conjunta es $f(x,y) = \frac{1}{18}xy$, $x = 1, 2, 3$; $y = 1, 2$, **cero** para otro **(x,y)**

Encuentre la covarianza entre X, Y

Para usar la fórmula de la covarianza: $\sigma_{XY} = Cov(X,Y) = E(XY) - \mu_X\mu_Y$

Se necesitan las distribuciones marginales

$$g(x) = \sum_y f(x,y) = \sum_{y=1}^{2} \frac{1}{18}xy = \frac{x}{18}\sum_{y=1}^{2} y = \frac{x}{18}(1+2) = \frac{x}{6}, \qquad x = 1, 2, 3$$

$$h(y) = \sum_x f(x,y) = \sum_{x=1}^{3} \frac{1}{18}xy = \frac{y}{18}\sum_{x=1}^{3} x = \frac{y}{18}(1+2+3) = \frac{y}{3}, \qquad y = 1, 2$$

Entonces

$$\mu_X = E(X) = \sum_{x=1}^{3} xg(x) = \sum_{x=1}^{3} x\frac{x}{6} = \frac{1}{6}\sum_{x=1}^{3} x^2 = \frac{1}{6}(1^2+2^2+3^2) = \frac{7}{3}$$

$$\mu_Y = E(Y) = \sum_{y=1}^{2} yh(y) = \sum_{y=1}^{2} y\frac{y}{3} = \frac{1}{3}\sum_{y=1}^{2} y^2 = \frac{1}{3}(1^2+2^2) = \frac{5}{3}$$

Además

$$E(XY) = \sum_{x=1}^{3}\sum_{y=1}^{2} xy\frac{1}{18}xy = \frac{1}{18}\sum_{x=1}^{3}\sum_{y=1}^{2} x^2y^2 = \frac{1}{18}[1^21^2 + 1^22^2 + 2^21^2 + 2^22^2 + 3^21^2 + 3^22^2] = \frac{70}{18}$$

Sustituyendo

$$\sigma_{XY} = Cov(X,Y) = E(XY) - \mu_X\mu_Y = 70/18 - (7/3)(5/3) = 0$$

Ejemplo
Sean **X, Y** variables aleatorias continuas cuya función de densidad de probabilidad conjunta es

$f(x,y) = \frac{2}{3}(x+2y)$, $0 \le x, y \le 1$, **cero** para otro **(x,y)**

Encuentre la covarianza entre X, Y

Para usar la fórmula de la covarianza: $\sigma_{XY} = Cov(X,Y) = E(XY) - \mu_X\mu_Y$

Se necesitan las distribuciones marginales

$$g(x) = \int_0^1 \frac{2}{3}(x+2y)dy = \frac{2}{3}(x+1), \qquad h(y) = \int_0^1 \frac{2}{3}(x+2y)dx = \frac{1}{3}(1+4y)$$

Entonces

$$\mu_X = E(X) = \int_0^1 xg(x)dx = \int_0^1 x\frac{2}{3}(x+1)dx = \frac{5}{9}$$

$$\mu_Y = E(Y) = \int_0^1 yh(y)dy = \int_0^1 y\frac{1}{3}(1+4y)dy = \frac{11}{18}$$

Además

$$E(XY) = \int_0^1\int_0^1 xyf(x,y)dxdy = \int_0^1\int_0^1 xy\frac{2}{3}(x+2y)dxdy = \frac{1}{3}$$

Sustituyendo

$$\sigma_{XY} = Cov(X,Y) = E(XY) - \mu_X\mu_Y = 1/3 - (5/9)(11/18) = -1/162$$

8.6.1 SIGNOS DE LA COVARIANZA CASO BIVARIADO

La covarianza es una medida del nivel de relación entre las variables aleatorias **X, Y**.
La covarianza tiene significado si la relación entre las variables aleatorias es **lineal.**

a) Si valores grandes de **X** están asociados probabilísticamente con valores grandes de **Y**, o si valores pequeños de **X** están asociados probabilísticamente con valores pequeños de **Y** entonces **la covarianza tiene signo positivo.**

b) Si valores grandes de **X** están asociados probabilísticamente con valores pequeños de **Y**, o si valores pequeños de **X** están asociados probabilísticamente con valores grandes de **Y** entonces **la covarianza tiene signo negativo.**

Para entender este comportamiento debemos referirnos a la definición de covarianza:

$$\mathbf{Cov(X,Y) = E\left[(X-\mu_X)(Y-\mu_Y)\right] = \sum_x \sum_y (x-\mu_X)(y-\mu_Y) f(x,y)}$$

Si los valores de **X** y **Y** son ambos mayores o ambos menores con respecto a su media, cada producto de las diferencias $\mathbf{(x-\mu_X)(y-\mu_Y)}$ tendrá signo positivo. Si en la sumatoria, estos términos tienen valores de probabilidad altos, entonces el resultado final tendrá signo positivo. En los casos contrarios la suma tendrá signo negativo.

Esta relación se puede visualizar como la pendiente de una recta que relaciona **X** y **Y**.

c) Si X, Y son variables aleatorias estadísticamente independientes, entonces Cov(X,Y)=0

Demostración

Si **X, Y** son variables aleatorias estadísticamente independientes, se tiene que

$$\mathbf{f(x,y) = g(x)\, h(y)}.$$

Esto permite separar las sumatorias:

$$\mathbf{E(XY) = \sum_x \sum_y xyf(x,y) = \sum_x \sum_y xyg(x)h(y) = \sum_x xg(x) \sum_y yh(y) = E(X)\, E(Y)}$$

Este resultado se sustituye en la fórmula de la covarianza:

$$\mathbf{Cov(X,Y) = E(XY) - \mu_X\mu_Y = E(X)E(Y) - \mu_X\mu_Y = \mu_X\mu_Y - \mu_X\mu_Y = 0}$$

NOTA:
Si **Cov(X,Y)** = **0** esto no implica necesariamente que **X, Y** sean variables aleatorias independientes.

Ejemplo

Sean **X, Y** variables aleatorias discretas cuya función de distribución de probabilidad conjunta es $\mathbf{f(x,y) = \frac{1}{18}xy}$, $\mathbf{x = 1, 2, 3;}$ $\mathbf{y = 1, 2,}$ **cero** para otro **(x,y)**

Demuestre con la propiedad anterior que **Cov(X,Y) = 0**

Solución

Se obtuvieron previamente las distribuciones marginales

$$\mathbf{g(x) = \sum_{y} f(x,y) = \sum_{y=1}^{2} \frac{1}{18}xy = \frac{x}{18}\sum_{y=1}^{2} y = \frac{x}{18}(1+2) = \frac{x}{6}, \qquad x = 1, 2, 3}$$

$$\mathbf{h(y) = \sum_{x} f(x,y) = \sum_{x=1}^{3} \frac{1}{18}xy = \frac{y}{18}\sum_{x=1}^{3} x = \frac{y}{18}(1+2+3) = \frac{y}{3}, \qquad y = 1, 2}$$

Se tiene que

$$\mathbf{f(x,y) = \frac{1}{18}xy,\ x=1, 2, 3;\ y = 1, 2.}$$

$$\mathbf{g(x)h(y) = (\frac{x}{6})(\frac{y}{3}) = \frac{1}{18}xy,\ x = 1, 2, 3;\ y = 1, 2.}$$

Se cumple que

$$\mathbf{f(x,y) = g(x)h(y),\ x = 1, 2, 3;\ y = 1, 2.}$$

Por lo tanto, **X, Y** son variables aleatorias estadísticamente independientes

En consecuencia,

$$\mathbf{Cov(X,Y) = 0}$$

8.6.2 MATRIZ DE VARIANZAS Y COVARIANZAS

Es una representación ordenada de las varianzas y covarianzas entre las variables aleatorias.

Definición: Matriz de Varianzas y Covarianzas Caso Bivariado

Sean **X** y **Y** variables aleatorias conjuntas (discretas o continuas)

$\sigma_X^2 = V(X), \quad \sigma_Y^2 = V(Y)$ Varianzas

$\sigma_{XY} = \sigma_{YX} = Cov(X, Y) = Cov(Y, X)$ Covarianzas

Entonces la matriz de varianzas y covarianzas es

$$[\sigma_{XY}] = \begin{bmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{YX} & \sigma_Y^2 \end{bmatrix}$$

Esta matriz es simétrica y contiene en la diagonal las varianzas de cada variable. Los otros componentes son las covarianzas entre las dos variables: $\sigma_{XY} = \sigma_{YX}$

Ejemplo

Sean **X, Y** variables aleatorias discretas cuya función de distribución de probabilidad conjunta es $f(x,y) = \frac{1}{18}xy$, $x = 1, 2, 3$; $y = 1, 2$, **cero** para otro **(x,y)**

Encuentre la matriz de varianzas y covarianzas

Se obtuvieron previamente las distribuciones marginales

$$g(x) = \sum_y f(x,y) = \sum_{y=1}^{2} \frac{1}{18}xy = \frac{x}{18}\sum_{y=1}^{2} y = \frac{x}{18}(1+2) = \frac{x}{6}, \qquad x = 1, 2, 3$$

$$h(y) = \sum_x f(x,y) = \sum_{x=1}^{3} \frac{1}{18}xy = \frac{y}{18}\sum_{x=1}^{3} x = \frac{y}{18}(1+2+3) = \frac{y}{3}, \qquad y = 1, 2$$

Medias, varianzas y covarianza

$$\mu_X = E(X) = \sum_{x=1}^{3} x\,g(x) = \sum_{x=1}^{3} x\frac{x}{6} = \frac{1}{6}\sum_{x=1}^{3} x^2 = \frac{1}{6}(1^2+2^2+3^2) = \frac{7}{3}$$

$$\mu_Y = E(Y) = \sum_{y=1}^{2} y\,h(y) = \sum_{y=1}^{2} y\frac{y}{3} = \frac{1}{3}\sum_{y=1}^{2} y^2 = \frac{1}{3}(1^2+2^2) = \frac{5}{3}$$

$$E(X^2) = \sum_{x=1}^{3} x^2 g(x) = \sum_{x=1}^{3} x^2\frac{x}{6} = \frac{1}{6}\sum_{x=1}^{3} x^3 = \frac{1}{6}(1^3+2^3+3^3) = 6$$

$$E(Y^2) = \sum_{y=1}^{2} y^2 h(y) = \sum_{y=1}^{2} y^2\frac{y}{3} = \frac{1}{3}\sum_{y=1}^{2} y^3 = \frac{1}{3}(1^3+2^3) = 3$$

$$E(XY) = E(YX) = \sum_{x=1}^{3}\sum_{y=1}^{2} xy\frac{1}{18}xy = \frac{70}{18}$$

$$g(x)\,h(y) = \left(\frac{x}{6}\right)\left(\frac{y}{3}\right) = \frac{1}{18}xy = f(x,y), \quad x = 1, 2, 3; \quad y = 1, 2$$

$\Rightarrow$ **X, Y** son variables aleatorias independiente $\Rightarrow$ $\sigma_{XY} = Cov(X, Y) = 0$

$\sigma_X^2 = V(X) = E(X^2) - E^2(X) = 6 - (7/3)^2 = 5/9$

$\sigma_Y^2 = V(Y) = E(Y^2) - E^2(Y) = 3 - (5/3)^2 = 2/9$

Matriz de varianzas - covarianzas

$$[\sigma_{XY}] = \begin{bmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{YX} & \sigma_Y^2 \end{bmatrix} = \begin{bmatrix} 5/9 & 0 \\ 0 & 2/9 \end{bmatrix}$$

8.6.3 COEFICIENTE DE CORRELACIÓN LINEAL

Es una medida normalizada de la relación lineal entre dos variables aleatorias. Se puede demostrar que el coeficiente de correlación reduce el rango de la covarianza al intervalo **[-1, 1]**

Definición: Coeficiente de Correlación Lineal Caso Bivariado

Sean **X, Y** variables aleatorias conjuntas (discretas o continuas)

entonces, el coeficiente de correlación lineal de **X, Y** es:

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{V(X)}\sqrt{V(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}, \quad -1 \le \rho_{XY} \le 1$$

Valores referenciales

Valor de ρ_{XY}	**X** y **Y**
Cercano a **1**	Tienen correlación lineal positiva fuerte
Cercano a **-1**	Tienen correlación lineal negativa fuerte
Cercano a **0**	Tienen correlación lineal muy débil o no están correlacionadas linealmente.

Es importante que se mida la correlación entre variables cuya asociación tenga algún significado de interés. Asimismo, si las variables no están correlacionadas linealmente, pudiera ser que tengan algún otro tipo de correlación, pero no lineal

Es necesario distinguir entre correlación y causalidad. Si dos variables están correlacionadas, esto no implica necesariamente que una sea causa de la otra pues ambas pueden depender de una tercera variable. Aún en el caso de que la correlación represente una causalidad, la estadística solamente permite detectarla y medirla, pero no demostrarla pues esto cae en el ámbito de la ciencia en la que se aplica la estadística

8.6.4 MATRIZ DE CORRELACIÓN

Es una representación ordenada de los valores de correlación entre las variables aleatorias.

Definición: Matriz de Correlación Caso Bivariado

Sean **X** y **Y** variables aleatorias conjuntas (discretas o continuas)

Entonces la matriz de correlación es

$$[\rho_{XY}] = \begin{bmatrix} 1 & \rho_{XY} \\ \rho_{YX} & 1 \end{bmatrix}$$

Esta matriz es simétrica y contiene el valor **1** en la diagonal. Los otros componentes son valores de correlación entre las dos variables tales que $\rho_{XY} = \rho_{YX}$

Las definiciones anteriores pueden extenderse a más variables aleatorias conjuntas

Definiciones: Matriz de Varianzas y Covarianzas y Matriz de Correlación Caso k-variado

Sean: $X_1, X_2, \ldots X_k$ Variables aleatorias conjuntas (discretas o continuas)

$\sigma_{ii} = V(X_i)$ Varianza de la variable X_i

$\sigma_{ij} = Cov(X_i, X_j)$ Covarianza de las variables X_i, X_j

ρ_{ij} Coeficiente de correlación lineal entre las variables X_i, X_j

Matriz de Varianzas-Covarianzas

$$\left[\sigma_{ij}\right] = \begin{bmatrix} \sigma_{11} & \sigma_{12} & . & . & \sigma_{1k} \\ \sigma_{21} & \sigma_{22} & . & . & \sigma_{2k} \\ . & . & . & . & . \\ . & . & . & . & . \\ \sigma_{k1} & \sigma_{k2} & . & . & \sigma_{kk} \end{bmatrix}$$

Matriz de Correlación

$$\left[\rho_{ij}\right] = \begin{bmatrix} 1 & \rho_{12} & . & . & \rho_{1k} \\ \rho_{21} & 1 & . & . & \rho_{2k} \\ . & . & . & . & . \\ . & . & . & . & . \\ \rho_{k1} & \rho_{k2} & . & . & 1 \end{bmatrix}$$

8.7 MEDIA Y VARIANZA PARA VARIABLES ALEATORIAS CONJUNTAS TRIVARIADAS

Las definiciones para distribuciones bivariadas pueden extenderse a más variables.

Los siguientes son ejemplos referenciales

Ejemplo con tres variables aleatorias discretas

Sea V un vector aleatorio discreto cuyos componentes son las variables aleatorias X, Y, Z, con distribución de probabilidad conjunta

$$f(x,y,z)=\begin{cases}\dfrac{1}{112}x^2(y-z); x=1,2,3; y=3,4; z=1,2\\ 0; \quad \text{para el resto de } x,y,z\end{cases}$$

Encuentre la matriz de varianzas y covarianzas y la matriz de correlación

Distribuciones marginales (se da por entendido el dominio de cada una)

$$f(x)=\sum_{y=3}^{4}\sum_{z=1}^{2}f(x,y,z)=\sum_{y=3}^{4}\sum_{z=1}^{2}\frac{1}{112}x^2(y-z)=\frac{1}{112}x^2\sum_{y=3}^{4}\sum_{z=1}^{2}(y-z)=\frac{8}{112}x^2$$

$$f(y)=\sum_{x=1}^{3}\sum_{z=1}^{2}\frac{1}{112}x^2(y-z)=\frac{1}{112}\sum_{x=1}^{3}x^2\sum_{z=1}^{2}(y-z)=\frac{1}{112}(14)(y-1+y-2)=\frac{14}{112}(2y-3)$$

$$f(z)=\sum_{x=1}^{3}\sum_{y=3}^{4}\frac{1}{112}x^2(y-z)=\frac{1}{112}\sum_{x=1}^{3}x^2\sum_{y=3}^{4}(y-z)=\frac{14}{112}(-2z+7)$$

$$f(x,y)=\sum_{z=1}^{2}f(x,y,z)=\sum_{z=1}^{2}\frac{1}{112}x^2(y-z)=\frac{1}{112}x^2\sum_{z=1}^{2}(y-z)=\frac{x^2}{112}(2y-3)$$

$$f(x,z)=\sum_{y=3}^{4}f(x,y,z)=\sum_{y=3}^{4}\frac{1}{112}x^2(y-z)=\frac{1}{112}x^2\sum_{y=3}^{4}(y-z)=\frac{x^2}{112}(-2z+7)$$

$$f(y,z)=\sum_{x=1}^{3}f(x,y,z)=\sum_{x=1}^{3}\frac{1}{112}x^2(y-z)=\frac{1}{112}(y-z)\sum_{x=1}^{3}x^2=\frac{14}{112}(y-z)$$

$f(x,y) = f(x)\,f(y)$ $\Rightarrow$ **X, Y** son variables aleatorias independientes

$f(x,z) = f(x)\,f(z)$ $\Rightarrow$ **X, Z** son variables aleatorias independientes

$f(y,z) \neq f(y)\,f(z)$ $\Rightarrow$ **Y, Z** son variables aleatorias **no** independientes

Medias, varianzas y covarianzas

$$E(X)=\sum_{x=1}^{3}xf(x)=\sum_{x=1}^{3}x\frac{8}{112}x^2=\frac{8}{112}\sum_{x=1}^{3}x^3=\frac{288}{112}$$

$$E(X^2)=\sum_{x=1}^{3}x^2f(x)=\sum_{x=1}^{3}x^2\frac{8}{112}x^2=\frac{8}{112}\sum_{x=1}^{3}x^4=\frac{784}{112}$$

$$E(Y)=\sum_{y=3}^{4}yf(y)=\sum_{y=3}^{4}y\frac{14}{112}(2y-3)=\frac{14}{112}\sum_{y=3}^{4}y(2y-3)=\frac{29}{8}$$

$$E(Y^2)=\sum_{y=3}^{4}y^2f(y)=\sum_{y=3}^{4}y^2\frac{14}{112}(2y-3)=\frac{14}{112}\sum_{y=3}^{4}y^2(2y-3)=\frac{107}{8}$$

$$E(Z)=\sum_{z=1}^{2}zf(z)=\sum_{z=1}^{2}z\frac{14}{112}(-2z+7)=\frac{14}{112}\sum_{z=1}^{2}z(-2z+7)=\frac{11}{8}$$

$$E(Z^2)=\sum_{z=1}^{2}z^2f(z)=\sum_{z=1}^{2}z^2\frac{14}{112}(-2z+7)=\frac{14}{112}\sum_{z=1}^{2}z^2(-2z+7)=\frac{17}{8}$$

$$E(YZ)=\sum_{y=3}^{4}\sum_{z=1}^{2}(yz)f(y,z)=\frac{14}{112}\sum_{y=3}^{4}\sum_{z=1}^{2}yz(y-z)=\frac{560}{112}$$

$$\sigma_X^2 = V(X) = E(X^2) - E^2(X) = \frac{784}{112} - (\frac{288}{112})^2 = 83/196$$

$$\sigma_Y^2 = V(Y) = E(Y^2) - E^2(Y) = \frac{107}{8} - (\frac{29}{8})^2 = \frac{15}{64}$$

$$\sigma_Z^2 = V(Z) = E(Z^2) - E^2(Z) = \frac{17}{8} - (\frac{11}{8})^2 = \frac{15}{64}$$

$\sigma_{XY} = \sigma_{YX} = Cov(XY) = 0$ Por ser variables aleatorias independientes

$\sigma_{XZ} = \sigma_{ZX} = Cov(XZ) = 0$ Por ser variables aleatorias independientes

$$\sigma_{YZ} = \sigma_{ZY} = Cov(YZ) = E(YZ) - E(Y)E(Z) = \frac{560}{112} - \frac{29}{8}\frac{11}{8} = \frac{1}{64}$$

Matriz de varianzas y covarianzas

$$[\sigma_{ij}] = \begin{bmatrix} \sigma_X^2 & \sigma_{XY} & \sigma_{XZ} \\ \sigma_{YX} & \sigma_Y^2 & \sigma_{YZ} \\ \sigma_{ZX} & \sigma_{ZY} & \sigma_Z^2 \end{bmatrix} = \begin{bmatrix} 83/196 & 0 & 0 \\ 0 & 15/64 & 1/64 \\ 0 & 1/64 & 15/64 \end{bmatrix}$$

Coeficientes de correlación

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{V(X)}\sqrt{V(Y)}} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y} = 0$$

$$\rho_{XZ} = \frac{Cov(X,Z)}{\sqrt{V(X)}\sqrt{V(Z)}} = \frac{\sigma_{XZ}}{\sigma_X\sigma_Y} = 0$$

$$\rho_{YZ} = \frac{Cov(Y,Z)}{\sqrt{V(Y)}\sqrt{V(Z)}} = \frac{\sigma_{YZ}}{\sigma_Y\sigma_Z} = \frac{1/64}{\sqrt{15/64}\sqrt{15/64}} = \frac{1}{15}$$

Matriz de correlación

$$[\rho_{i,j}] = \begin{bmatrix} 1 & \rho_{XY} & \rho_{XZ} \\ \rho_{YX} & 1 & \rho_{YZ} \\ \rho_{ZX} & \rho_{ZY} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1/15 \\ 0 & 1/15 & 1 \end{bmatrix}$$

Ejemplo con tres variables aleatorias continuas

Sea **[X, Y, Z]** un vector aleatorio trivariado cuya distribución de probabilidad conjunta es:
f(x,y,z) = x(y+z), 0 < x < 2, 0 < y < z < 1, cero para otro **(x,y,z)**

Encuentre la matriz de varianzas y covarianzas

Distribuciones marginales

$$f(x) = \int_0^1\int_0^z x(y+z)dydz = x\int_0^1\int_0^z (y+z)dydz = x\int_0^1 [\frac{y^2}{2} + yz]_0^z dz$$

$$= x\int_0^1 (\frac{z^2}{2} + z^2)dz = \frac{3x}{2}\int_0^1 z^2 dz = \frac{3x}{2}[\frac{z^3}{3}]_0^1 = \frac{3x}{2}(\frac{1}{3}) = \frac{x}{2}, \quad 0 < x < 2$$

$$f(y) = \int_y^1\int_0^2 x(y+z)dxdz = \int_y^1 (y+z)\int_0^2 xdxdz = \int_y^1 (y+z)[\frac{x^2}{2}]_0^2 dz$$

$$= 2\int_y^1 (y+z)dz = 2[yz + \frac{z^2}{2}]_y^1 = 1 + 2y - 3y^2, \qquad 0 < y < 1$$

$$f(z) = \int_0^2\int_0^z x(y+z)dydx = \int_0^2 x\int_0^z (y+z)dydx = \int_0^2 x[\frac{y^2}{2} + zy]_0^z dx$$

$$= \frac{3}{2}\int_0^2 xz^2 dx = \frac{3}{2}z^2[\frac{x^2}{2}]_0^2 = 3z^2, \qquad 0 < z < 1$$

$$f(x,y)=\int_y^1 x(y+z)dz = x[yz+\frac{z^2}{2}]_y^1 = x(y+\frac{1}{2}-y^2-\frac{y^2}{2}=\frac{x}{2}(1+2y-3y^2)$$

$$0<x<2,\quad 0<y<1$$

$$f(x,z)=\int_0^z x(y+z)dy = x[\frac{y^2}{2}+zy]_0^z = x(\frac{z^2}{2}+z^2)=\frac{3xz^2}{2},\quad 0<x<2,\quad 0<z<1$$

$$f(y,z)=\int_0^2 x(y+z)dx = (y+z)[\frac{x^2}{2}]_0^2 = 2(y+z),\qquad 0<y<z<1$$

$$f(x,y)=\frac{x}{2}(1+2y-3y^2)=f(x)f(y) \quad \Rightarrow \textbf{X, Y} \text{ son independientes}$$

$$f(x,z)=\frac{3xz^2}{2}=f(x)f(z) \quad \Rightarrow \textbf{X, Z} \text{ son independientes}$$

$$f(y,z)=2(y+z)$$

$$f(y)f(z)=(1+2y-3y^2)(3z^2)\neq f(y,z) \quad \Rightarrow \textbf{Y, Z no} \text{ son independientes}$$

Medias, varianzas y covarianzas

$$E(X)=\int_0^2 xf(x)dx=\int_0^2 x(\frac{x}{2})dx=\frac{4}{3}$$

$$E(X^2)=\int_0^2 x^2f(x)dx=\int_0^2 x^2(\frac{x}{2})dx=2$$

$$E(Y)=\int_0^1 yf(y)dy=\int_0^1 y(1+2y-3y^2)dy=\frac{5}{12}$$

$$E(Y^2)=\int_0^1 y^2f(y)dy=\int_0^1 y^2(1+2y-3y^2)dy=\frac{7}{30}$$

$$E(Z)=\int_0^1 zf(z)dz=\int_0^1 z(3z^2)dz=\frac{3}{4}$$

$$E(Z^2)=\int_0^1 z^2f(z)dz=\int_0^1 z^2(3z^2)dz=\frac{3}{5}$$

$$E(YZ)=\int_0^1\int_0^z yz(2(y+z))dydz=2\int_0^1\int_0^z (y^2z+yz^2)dydz=2\int_0^1[\frac{y^3z}{3}+\frac{y^2}{2}z^2]_0^z dz$$

$$=\frac{5}{3}\int_0^1 z^4dz=1/3$$

$$\sigma_X^2=V(X)=E(X^2)-E^2(X)=2-(4/3)^2=2/9$$

$$\sigma_Y^2=V(Y)=E(Y^2)-E^2(Y)=7/30-(5/12)^2=43/720$$

$$\sigma_Z^2=V(Z)=E(Z^2)-E^2(Z)=3/5-(3/4)^2=3/80$$

$\sigma_{XY}=Cov(XY)=0$ Por ser variables aleatorias independientes

$\sigma_{XZ}=Cov(XZ)=0$ Por ser variables aleatorias independientes

$$\sigma_{YZ}=Cov(Y,Z)=E(YZ)-E(Y)E(Z)=1/3-(5/12)(3/4)=1/48$$

$$[\sigma_{ij}]=\begin{bmatrix}\sigma_X^2 & \sigma_{XY} & \sigma_{XZ}\\ \sigma_{YX} & \sigma_Y^2 & \sigma_{YZ}\\ \sigma_{ZX} & \sigma_{ZY} & \sigma_Z^2\end{bmatrix}=\begin{bmatrix}2/9 & 0 & 0\\ 0 & 43/720 & 1/48\\ 0 & 1/48 & 3/80\end{bmatrix}$$

8.7.1 EJERCICIOS

1) Si la distribución de probabilidad conjunta de las variables aleatorias discretas **X, Y** está dada por

$$\mathbf{f(x,y)} = \frac{\mathbf{1}}{\mathbf{30}}\mathbf{(x+y)}\ ,\ \mathbf{x=0,\ 1,\ 2,\ 3;}\quad \mathbf{y=0,\ 1,\ 2}$$

Encuentre la matriz de varianzas y covarianzas

2) Si la distribución de probabilidad conjunta de las variables aleatorias continuas **X, Y** está dada por

$$\mathbf{f(x,y)} = \begin{cases} \frac{\mathbf{1}}{\mathbf{4}}\mathbf{(2x+y),} & \mathbf{0<x<1,\ 0<y<1} \\ \mathbf{0,} & \mathbf{\text{para otros valores}} \end{cases}$$

Encuentre la matriz de correlación

MATLAB

Manejo simbólico de media y varianza para distribuciones conjuntas

Variables aleatorias discretas

Sean **X, Y** variables aleatorias discretas cuya función de distribución de probabilidad conjunta es $f(x,y) = \begin{cases} \frac{1}{18}xy, & x = 1,2,3;\ y = 1,2 \\ 0, & \text{otro } (x,y) \end{cases}$

```
>> syms x y                          Definición de variables simbólicas
>> f=x*y/18;                         Distribución de probabilidad conjunta (discreta)
>> g=0;                              Obtención de la distribución marginal g(x)
>> for y=1:2
       g=g+eval(subs(f,'y',y));
   end
>> g
   g =
      1/6*x
>> syms x y
>> h=0;                              Obtención de la distribución marginal h(y)
>> for x=1:3
       h=h+eval(subs(f,'x',x));
   end
>> h
   h =
      1/3*y
>> EX=0;                             Obtención de E(X)
>> for x=1:3
       EX=EX+eval(x*g);
   end
>> EX
   EX =
       7/3
>> EY=0;                             Obtención de E(Y)
>> for y=1:2
       EY=EY+eval(y*h);
   end
>> EY
   EY =
       5/3
>> EX2=0;                            Obtención de E(X^2)
>> for x=1:3
       EX2=EX2+eval(x^2*g);
   end
>> EX2
   EX2 =
        6
```

```
>> EY2=0;                                  Obtención de E(Y^2)
>> for y=1:2
       EY2=EY2+eval(y^2*h);
   end
>> EY2
   EY2 =
        3
>> EXY=0;                                  Obtención de E(XY)
>> for x=1:3
       for y=1:2
          EXY=EXY+eval(x*y*f);
       end
   end
>> EXY
   EXY =
      35/9
>> sigma2X=EX2-EX^2                        Varianza de X
   sigma2X =
        5/9
>> sigma2Y=EY2-EY^2                        Varianza de Y
   sigma2Y =
        2/9
>> CovXY=EXY-EX*EY                         Covarianza de X, Y
   CovXY =
    8.8818e-016                            El resultado es aproximadamente cero
```

Variables aleatorias continuas

Sean **X, Y** variables aleatorias continuas cuya función de densidad de probabilidad conjunta es

$$f(x,y) = \begin{cases} \frac{2}{3}(x+2y), & 0 \le x, y \le 1 \\ 0, & \text{otro } (x,y) \end{cases}$$

```
>> syms x y                                Definición de variables simbólicas
>> f=2/3*(x + 2*y);                        Función de densidad conjunta f(x,y)

>> g=int(f,y,0,1)                          Densidad marginal  g(x)
   g =
      2/3*x+2/3
>> h=int(f,x,0,1)                          Densidad marginal h(y)
   h =
      1/3+4/3*y
>> EX=int(x*g,0,1)                         Obtención de  E(X)
   EX =
      5/9
>> EY=int(y*h,0,1)                         Obtención de  E(Y)
   EY =
      11/18
>> EXY=int(int(x*y*f,x,0,1),y,0,1)         Obtención de  E(XY)
   EXY =
        1/3
>> CovXY=EXY-EX*EY                         Covarianza de  X,Y
   CovXY =
          -1/162                           X, Y no son independientes
>> r=expand(f)==expand(g*h)                Verificar que  f(x,y) = g(x) h(y)
   r =
        0                                  No es verdad
```

8.8 DISTRIBUCIÓN MULTINOMIAL

Es una generalización de la distribución Binomial. Se presenta cuando los resultados de cada ensayo tienen más de dos resultados posibles. Se supondrá que los ensayos son independientes y que la probabilidad se mantiene constante para cada tipo de resultado.

Definición: Distribución Multinomial

Sean **n**: cantidad de ensayos realizados
k: cantidad de resultados diferentes que se pueden obtener en cada ensayo

Sean las variables aleatorias discretas:
$\mathbf{X_1}$: Cantidad de resultados de tipo 1
$\mathbf{X_2}$: Cantidad de resultados de tipo 2
. . .
$\mathbf{X_k}$: Cantidad de resultados de tipo k
Tales que $\mathbf{x_1 + x_2 + \ldots + x_k = n}$

Sean las probabilidades correspondientes a cada tipo de resultado
$\mathbf{p_1}$: Probabilidad que el resultado sea de tipo 1
$\mathbf{p_2}$: Probabilidad que el resultado sea de tipo 2
. . .
$\mathbf{p_k}$: Probabilidad que el resultado sea de tipo k
Tales que $\mathbf{p_1 + p_2 + \ldots + p_k = 1}$

Entonces, las variables aleatorias $\mathbf{X_1, X_2, \ldots X_k}$ tienen distribución Multinomial y la distribución de probabilidad está dada por:

$$f(x_1,x_2,\ldots,x_k) = \binom{n}{x_1,x_2,\ldots,x_k} p_1^{x_1}p_2^{x_2}\ldots p_k^{x_k} = \frac{n!}{x_1!x_1!\ldots x_k!}p_1^{x_1}p_2^{x_2}\ldots p_k^{x_k}$$

$$x_1, x_2, \ldots, x_k = 0, 1, 2, \ldots, n; \quad x_1 + x_2 + \ldots + x_k = n$$

Demostración

Siendo ensayos independientes, la probabilidad de tener $\mathbf{x_1}$ resultados de tipo **1,** $\mathbf{x_2}$ resultados de tipo **2, ...,** $\mathbf{x_k}$ resultados de tipo **k**, es $p_1^{x_1}p_2^{x_2}\ldots p_k^{x_k}$. Pero existen $\binom{n}{x_1,x_2,\ldots,x_k}$ formas diferentes de obtener estos resultados, por lo tanto, esta cantidad es un factor.

En donde $\binom{n}{x_1,x_2,\ldots,x_k} = \frac{n!}{x_1!x_2!\ldots x_k!}$

8.8.1 MEDIA Y VARIANZA PARA LA DISTRIBUCIÓN MULTINOMIAL

Se puede calcular la media y varianza de cada variable aleatoria considerando a las demás variables aleatorias como otra variable:

Definición: Media y Varianza para la Distribución Multinomial

Sea $\mathbf{X_i}$ cualquiera de las variables discretas de la distribución binomal
Entonces

Media de $\mathbf{X_i}$ $\quad \mu_{X_i} = E(X_i) = np_i$

Varianza de $\mathbf{X_i}$ $\quad \sigma^2_{X_i} = V(X_i) = np_i(1-p_i),\ i = 1, 2, \ldots, k$

Ejemplo
Cada artículo producido por una fábrica puede ser de tipo aceptable, regular o defectuoso, con probabilidad 0.85, 0.10, y 0.05 respectivamente. Si se toman 5 artículos para examinarlos, calcule la probabilidad que 4 sean aceptables, 1 sea regular y ninguno defectuoso

Es un experimento multinomial con

$n = 5$	Cantidad de artículos tomados para examinar
X_1:	Cantidad de artículos aceptables
X_2:	Cantidad de artículos regulares
X_3:	Cantidad de artículos defectuosos
$p_1 = 0.85$	Probabilidad que un artículo sea aceptable
$p_2 = 0.10$	Probabilidad que un artículo sea regular
$p_3 = 0.05$	Probabilidad que un artículo sea defectuoso

La distribución de probabilidad para este experimento es:

$$f(x_1,x_2,x_3) = \binom{5}{x_1,x_2,x_3} p_1^{x_1} p_2^{x_2} p_3^{x_3} = \frac{5!}{x_1!x_1!x_3!} p_1^{x_1} p_2^{x_2} p_3^{x_3}$$

$x_1, x_2, x_3 = 0, 1, 2, 4, 5; \quad x_1 + x_2 + x_3 = 5$

Entonces

$$P(X_1=4, X_2=1, X_3=0) = f(4,1,0) = \binom{5}{4,1,0} 0.85^4 0.10^1 0.05^0 = \frac{5!}{4!1!0!} 0.85^4 0.10^1 0.05^0$$

$$= 0.261$$

NOTA. Este problema puede reducirse a dos variables definiendo $X_3 = 5 - X_1 - X_2$ mientras que $p_3 = 1 - (p_1 + p_2)$ con lo cual, la distribución de probabilidad es:

$$f(x_1,x_2) = \binom{5}{x_1,x_2,\ 5-x_1-x_2} p_1^{x_1} p_2^{x_2} (1-p_1-p_2)^{5-x_1-x_2}$$

$x_1, x_2 = 0, 1, 2, 3, 4, 5; \quad x_1 + x_2 \le 5; \quad x_3 = 5 - x_1 - x_2$

8.9 DISTRIBUCIÓN HIPERGEOMÉTRICA MULTIVARIADA

Esta distribución es una generalización de la distribución Hipergeométrica. Se aplica a experimentos de muestreo sin reemplazo de una población finita en la que hay objetos de más de dos tipos diferentes. Los objetos tomados no son devueltos a la población. Por lo tanto la cantidad de objetos en el conjunto cambia y los valores de probabilidad cambian.

Definición: Distribución Hipergeométrica Multivariada

Sean **N**: Cantidad de objetos en un conjunto en el que existen **k** diferentes tipos.
C_1: Cantidad de objetos de tipo 1 en el conjunto
C_2: Cantidad de objetos de tipo 2 en el conjunto
. . .
C_k: Cantidad de objetos de tipo k en el conjunto

Tales que $C_1 + C_2 + \ldots + C_k = N$

Sea **n:** Cantidad de objetos que se toman en la muestra, sin devolverlos.

Sean las variables aleatorias discretas:
X_1: Cantidad de objetos de tipo 1 que se obtienen en la muestra.
X_2: Cantidad de objetos de tipo 2 que se obtienen en la muestra.
. . .
X_k : Cantidad de objetos de tipo k que se obtienen en la muestra.

Tales que $x_1 + x_2 + \ldots + x_k = n$

Entonces, la distribución de probabilidad de $X_1, X_2, \ldots X_k$ está dada por la función:

$$f(x_1,x_2,\ldots,x_k) = \frac{\binom{C_1}{x_1}\binom{C_2}{x_2}\cdots\binom{C_k}{x_k}}{\binom{N}{n}}$$

$x_1, x_2, \ldots, x_k = 0, 1, \ldots, n; \quad x_1 + x_2 + \ldots + x_k = n; \quad C_1 + C_2 + \ldots + C_k = N$

Demostración

Se tienen $\binom{C_1}{x_1}$ formas diferentes de tomar x_1 objetos de tipo **1** de los C_1 disponibles

Se tienen $\binom{C_2}{x_2}$ formas diferentes de tomar x_2 objetos de tipo **2** de los C_2 disponibles

. . .

Se tienen $\binom{C_k}{x_k}$ formas diferentes de tomar x_k objetos de tipo **k** de los C_k disponibles

Además hay $\binom{N}{n}$ formas diferentes de tomar **n** objetos de los **N** existentes en la población

La fórmula se obtiene aplicando el principio fundamental del conteo y la asignacíón clásica de probabilidad

Ejemplo

Una caja contiene 4 baterías en buen estado, 3 baterías en regular estado, y 2 baterías defectuosas. De esta caja se toma una muestra aleatoria de dos baterías.

a) Encuentre la distribución de probabilidad conjunta.

Sean las variables aleatorias discretas

X: Cantidad de baterías aceptables en la muestra
Y: Cantidad de baterías en regular estado en la muestra
Z: Cantidad de baterías defectuosas en la muestra.

Es un experimento hipergeométrico. Entonces, la distribución de probabilidad conjunta es

$$P(X=x, Y=y, Z=z) = f(x,y,z) = \frac{\binom{4}{x}\binom{3}{y}\binom{2}{z}}{\binom{9}{2}}, \quad x, y, z = 0,1,2; \quad x+y+z=2$$

b) Calcule la probabilidad de obtener una en buen estado y una defectuosa

$$P(X=1, Y=0, Z=1) = f(1,0,1) = \frac{\binom{4}{1}\binom{3}{0}\binom{2}{1}}{\binom{9}{2}} = 0.2222$$

NOTA. Este problema puede reducirse a dos variables definiendo $Z = 2 - X - Y$

Con esta sustitución, la distribución de probabilidad es:

$$P(X=x, Y=y) = f(x,y) = \frac{\binom{4}{x}\binom{3}{y}\binom{2}{2-x-y}}{\binom{9}{2}}, \quad x, y = 0,1,2; \quad x+y \le 2$$

c) Calcule la probabilidad de obtener una en buen estado y una defectuosa

$$P(X=1, Y=0) = f(1,0) = \frac{\binom{4}{1}\binom{3}{0}\binom{2}{2-1-0}}{\binom{9}{2}} = 0.2222$$

d) Calcule **P(X=0)**

La probabilidad de una variable es la distribución marginal

$$P(X=0) = g(0) = \sum_{y=0}^{2} f(0,y) = f(0,0)+f(0,1)+f(0,2) = 0.2778$$

e) Obtenga una fórmula para la distribución marginal **g(x)**
Separamos las variables en dos grupos: **X** y las demás: **2 – x**

$$g(x) = \frac{\binom{4}{x}\binom{5}{2-x}}{\binom{9}{2}}, \quad x = 0, 1, 2$$

f) Calcule **P(X=0)** con la distribución marginal **g(x)**

$$P(X=0) = g(0) = \frac{\binom{4}{0}\binom{5}{2-0}}{\binom{9}{2}} = 0.2778$$

g) Encuentre la distribución condicional de **X** dado que **Y = 1**

$$f(x|1) = f(x,1)/h(1)$$

$$h(1) = \sum_{x=0}^{2} f(x,1) = f(0,1)+f(1,1)+f(2,1) = 0.5$$

$$f(x|1) = \frac{f(x,1)}{0.5}$$

h) Calcule la probabilidad que al tomar la segunda batería, ésta sea aceptable dado que la primera fue una batería en estado regular **Y = 1**

$$P(X=1|Y=1) = \frac{f(1,1)}{0.5} = \frac{0.3333}{0.5} = 0.6667$$

8.9.1 EJERCICIOS

1) En una ciudad, 60% de los empleados viaja a su trabajo en bus, 25% lo hace en su auto, 10% usa bicicleta y 5% camina. Encuentre la probabilidad que en una muestra de 8 empleados, 5 usen bus, 2 usen su auto, 1 camine y ninguno use bicicleta.

2) De acuerdo con la teoría de la genética, un cierto cruce de conejillos de indias resultara en una descendencia roja, negra y blanca en la relación 8:4:4. Encuentre la probabilidad de que de 10 descendientes, 6 sean rojos, 3 negros y 1 blanco.

3) Un frasco contiene 25 pastillas de igual forma y color. 15 son laxantes, siete son calmantes y tres son vitaminas. Si se eligen al azar cinco de estas pastillas, calcule la probabilidad de obtener

a) Cuatro laxantes y un calmante
b) Dos laxantes, un calmante y dos vitaminas.

4) Un club de estudiantes tiene en su lista a 3 serranos, 2 amazónicos, 5 costeños y 2 insulares. Si se selecciona aleatoriamente un comité de 4 estudiantes encuentre la probabilidad de que:

a) Estén representadas todas las regiones del país.
b) Estén representadas todas las nacionalidades excepto la amazonía.

8.10 PROPIEDADES DE LAS VARIABLES ALEATORIAS CONJUNTAS

En esta sección se establecen algunas propiedades útiles que serán usadas posteriormente en el tema principal de esta unidad que es el estudio de las **Distribuciones de Muestreo**.

PROPIEDAD 1

Sean X_1, X_2 variables aleatorias (discretas o continuas)

$a_1, a_2 \in \Re$

$Y = a_1 X_1 + a_2 X_2$, variable aleatoria definida con las variables X_1 y X_2

Entonces la media, o valor esperado de la variable **Y** es

$$\mu_Y = E(Y) = E(a_1 X_1 + a_2 X_2) = a_1 E(X_1) + a_2 E(X_2) = a_1 \mu_{X_1} + a_2 \mu_{X_2}$$

Esta definición se puede extender a expresiones con más variables aleatorias:

Sea	$Y = a_1 X_1 + a_2 X_2 + \ldots + a_n X_n$,	(X_i: variables aleatorias)
Entonces	$\mu_Y = a_1 \mu_{X_1} + a_2 \mu_{X_2} + \ldots + a_n \mu_{X_n}$	

PROPIEDAD 2

Sean X_1, X_2 variables aleatorias (discretas o continuas)

$a_1, a_2 \in \Re$

$Y = a_1 X_1 + a_2 X_2$, variable aleatoria definida con las variables X_1 y X_2

Entonces la varianza de la variable aleatoria **Y** es

$$\sigma^2_Y = V(Y) = a^2_1 V(X_1) + a^2_2 V(X_2) + 2 a_1 a_2 Cov(X_1, X_2)$$

Demostración

$$\begin{aligned} V(Y) &= V(a_1 X_1 + a_2 X_2) = E[(a_1 X_1 + a_2 X_2)^2] - E^2(a_1 X_1 + a_2 X_2) \\ &= E(a^2_1 X^2_1 + 2 a_1 a_2 X_1 X_2 + a^2_2 X^2_2) - [a_1 E(X_1) + a_2 E(X_2)]^2 \\ &= a^2_1 E(X^2_1) + 2 a_1 a_2 E(X_1 X_2) + a^2_2 E(X^2_2) - a^2_1 E^2(X_1) - 2 a_1 a_2 E(X_1)E(X_2) - a^2_2 E^2(X^2_2) \\ &= a^2_1 E(X^2_1) - a^2_1 E^2(X_1) + a^2_2 E(X^2_2) - a^2_2 E^2(X^2_2) + 2 a_1 a_2 E(X_1 X_2) - 2 a_1 a_2 E(X_1)E(X_2) \\ &= a^2_1 [E(X^2_1) - E^2(X_1)] + a^2_2 [E(X^2_2) - E^2(X^2_2)] + 2 a_1 a_2 [E(X_1 X_2) - E(X_1)E(X_2)] \\ &= a^2_1 V(X_1) + a^2_2 V(X_2) + 2 a_1 a_2 Cov(X_1, X_2) \end{aligned}$$

Si X_1, X_2 son variables aleatorias estadísticamente independientes

$Cov(X_1, X_2) = 0$

Entonces

$$\sigma^2_Y = a^2_1 V(X_1) + a^2_2 V(X_2) = a^2_1 \sigma^2_{X_1} + a^2_2 \sigma^2_{X_2}$$

Esta propiedad se puede extender a expresiones con más variables aleatorias:

Sea	$Y = a_1 X_1 + a_2 X_2 + \ldots + a_n X_n$,	(X_i: variables aleatorias **independientes**)
Entonces	$\sigma^2_Y = a^2_1 \sigma^2_{X_1} + a^2_2 \sigma^2_{X_2} + \ldots + a^2_n \sigma^2_{Xn}$	

COROLARIO

Sean $\mathbf{X_1}$, $\mathbf{X_2}$ variables aleatorias (discretas o continuas)

$\mathbf{a_1, a_2 \in \Re}$

Entonces, la covarianza entre $\mathbf{a_1X_1}$ y $\mathbf{a_2X_2}$ es:

$$\mathbf{Cov(a_1X_1, a_2X_2) = a_1\, a_2\, Cov(X_1, X_2)}$$

Demostración

$$Cov(a_1X_1, a_2X_2) = E(a_1X_1\, a_2X_2) - E(a_1X_1)E(a_2\, X_2)$$

$$= a_1\, a_2\, E(X_1\, X_2) - a_1\, E(X_1)\, a_2\, E(X_2)$$

$$= a_1\, a_2\, [E(X_1\, X_2) - E(X_1)\, E(X_2)]$$

$$= a_1\, a_2\, Cov(X_1, X_2)$$

PROPIEDAD 3

Sean $\mathbf{X_1}$, $\mathbf{X_2}$ variables aleatorias estadísticamente independientes (discretas o continuas)

$\mathbf{a_1, a_2 \in \Re}$

$\mathbf{Y = a_1\, X_1 + a_2\, X_2}$, variable aleatoria definida con las variables $\mathbf{X_1}$ y $\mathbf{X_2}$

Entonces la función generadora de momentos de la variable aleatoria **Y** es

$$\mathbf{m_y(t) = m_{a_1 x_1}(t)\, m_{a_2 x_2}(t)}$$

Demostración

$$\mathbf{m_Y(t) = E(e^{Yt}) = E[e^{(a_1 X_1 + a_2 X_2)t}] = E(e^{a_1 X_1 t} e^{a_2 X_2 t})}$$

Si $\mathbf{X_1, X_2}$ son variables aleatorias estadísticamente independientes ($E(X_1X_2) = E(X_1)E(X_2)$)

$$\mathbf{E(e^{a_1 X_1 t} e^{a_2 X_2 t}) = E(e^{a_1 X_1 t})\, E(e^{a_2 X_2 t})}$$

Por lo tanto

$$\mathbf{m_y(t) = m_{a_1 x_1}(t)\, m_{a_2 x_2}(t)}$$

Esta propiedad se puede extender a expresiones con más variables aleatorias:

Sea	$\mathbf{Y = a_1\, X_1 + a_2\, X_2 + ... + a_n\, X_n}$, ($\mathbf{X_i}$: variables aleatorias **independientes**)
Entonces	$\mathbf{m_Y(t) = m_{a_1 x_1}(t)\, m_{a_2 x_2}(t)\ \ldots\ m_{a_n x_n}(t)}$

9 MUESTREO ESTADÍSTICO

El muestreo estadístico es un procedimiento para obtener datos de una población con la finalidad de usar esta información para realizar inferencias acerca de dicha población mediante las técnicas que se estudian en **Estadística Inferencial**.

Las muestras son subconjuntos de los datos. El conjunto de todas las muestras que se pueden obtener de la población se denomina espacio muestral.

El muestreo estadístico se basa en el principio de equiprobabilidad, es decir que cada individuo de la población tiene la misma probabilidad de ser elegido. Consecuentemente, cada muestra también tendrá la misma probabilidad de ser seleccionada.

Para obtener conclusiones y evidencias comprobatorias suficientes, el investigador no está obligado a examinar todos y cada uno de los individuos o muestras de una población. Solamente debe examinar una muestra representativa de dicha población. El tamaño de la muestra, el tipo de muestreo, la escala de medición, el procedimiento de recolección de datos y otros aspectos relacionados, forman parte del diseño estadístico previo que debe concordar con el objetivo del estudio y con el nivel de confiabilidad que se pretende obtener.

Técnicas de selección de muestras

Muestreo probabilístico
Todas las muestras de la población tienen la misma probabilidad de ser elegidas.

Muestreo no probabilístico
La selección de la muestra está influenciada por la persona que la realiza o por otros factores no estadísticos.

Muestreo aleatorio simple
Una muestra aleatoria simple es aquel en el que cada elemento de la población tiene la misma probabilidad de ser seleccionado.

Muestreo con reemplazo
Cada elemento que se extrae de la población es observado y luego devuelto a la población, por lo que puede ser elegido más de una vez. Esto permite tomar infinidad de muestras de una población finita

Muestreo sin reemplazo
Los elementos elegidos en la muestra no son devueltos a la población

Muestreo simple, doble, múltiple
En estos tipos de muestreo se toman una, dos o más muestras de la población para analizar resultados y llegar a conclusiones definitivas.

Muestreo estratificado
La población previamente es dividida en grupos o clases a los que se les asigna una cuota de individuos de la población que comparten la característica que se estudia.

Muestreo por conglomerados
Las muestras son elegidas de grupos en los que se divide de manera natural la población y que representan la variabilidad de la población. El muestreo puede concentrarse únicamente en estos grupos.

Muestreo sistemático
Las muestras son elegidas recorriendo los elementos en un orden previamente determinado.

Muestreo errático o asistemático
El muestreo se realiza priorizando algún aspecto no estadístico conveniencia, reducción de costo o tiempo, etc.

Escalas de medición

Escala nominal
Es la asignación arbitraria de números o símbolos a cada una de las diferentes categorías en las que se puede dividir la característica que se estudia. No permite establecer relaciones entre categorías, solamente distinguirlas.

Escala ordinal
Se utiliza para establecer diversos grados de la característica que se observa en los individuos con lo que se puede establecer una relación de orden entre ellos. La asignación de los números o símbolos debe reflejar este orden.

Escala de intervalos
Utiliza una unidad de medida común y constante en la valoración de la característica que se observa en los individuos. Esta escala permite determinar la distancia entre los elementos y utilizar las medidas estadísticas cuantitativas.

Escala de coeficientes
Es similar a la escala de intervalos pero posee adicionalmente un punto de origen o cero para la escala, el cual representa la ausencia de la característica que se estudia. Esta escala permite medir la variabilidad de la característica que se mide.

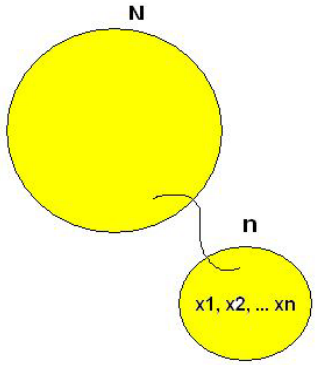
9.1 DISTRIBUCIONES DE MUESTREO

En esta sección se establecen algunas definiciones y términos relacionados con el estudio de la **Estadística Inferencial** que constituye el componente fundamental del estudio de la Estadística.

Una inferencia estadística es una afirmación que se hace acerca de algún parámetro de la población utilizando la información contenida en una muestra tomada de esta población.

Debemos aceptar que por la naturaleza aleatoria de los datos obtenidos en la muestra, hay un riesgo en la certeza de la afirmación propuesta, y es necesario establecer una medida para determinar la magnitud de este riesgo.

Supongamos una población de tamaño **N** de la cual se toma una muestra de tamaño **n**, obteniéndose los siguientes resultados: $x_1, x_2, ..., x_n$



Los **n** resultados obtenidos $x_1, x_2, ..., x_n$ son algunos de los posibles valores que se extraen de la población cada vez que se toma una muestra de tamaño **n**. Por lo tanto, podemos representarlos mediante **n** variables aleatorias: $X_1, X_2, ..., X_n$

Definición: Muestra Aleatoria

> Es un conjunto de **n** variables aleatorias $X_1, X_2, ..., X_n$ tales que son **independientes** y provienen de la misma población, es decir que tienen la misma función de probabilidad.

Para que esta definición sea válida, **N** debe ser muy grande respecto a **n**, o debe realizarse muestreo con reemplazo. Adicionalmente, cada elemento de la población debe tener la misma probabilidad de ser elegido.

Definiciones relacionadas:

> **Parámetro:** Es una medida estadística poblacional, cuyo valor es de interés conocer Por ejemplo, la media poblacional μ es un parámetro.
>
> **Estadístico o Estimador:** Es una variable aleatoria definida con las variables de la muestra aleatoria. Por ejemplo, la media muestral $\bar{X}$ es un estadístico.
>
> **Distribución de Muestreo de un Estadístico:** Es la distribución de probabilidad del estadístico

9.2 DISTRIBUCIÓN DE MUESTREO DE LA MEDIA MUESTRAL

En esta sección se estudian las propiedades de la distribución de probabilidad de la Media Muestral.

Definición: Media y Varianza de la Media Muestral

Sean $X_1, X_2, ..., X_n$ una muestra aleatoria tomada de una población con media μ y varianza σ^2, entonces, la media muestral es una variable aleatoria que se define con la siguiente fórmula:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$ y su media y varianza son:

Media de $\bar{X}$: $\mu_{\bar{X}} = E(\bar{X}) = \mu$

Varianza de $\bar{X}$: $\sigma^2_{\bar{X}} = V(\bar{X}) = \frac{\sigma^2}{n}$

Demostración

Media muestral: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}X_1 + \frac{1}{n}X_2 + ... + \frac{1}{n}X_n$

Por las propiedades estudiadas anteriormente, si $X_1, X_2, ..., X_n$ son **variables aleatorias independientes**, entonces

$$\mu_{\bar{X}} = \frac{1}{n}\mu_{X_1} + \frac{1}{n}\mu_{X_2} + ... + \frac{1}{n}\mu_{X_n}$$

$$\sigma^2_{\bar{X}} = (\frac{1}{n})^2\sigma^2_{X_1} + (\frac{1}{n})^2\sigma^2_{X_2} + ... + (\frac{1}{n})^2\sigma^2_{X_n}$$

Además, como las variables aleatorias provienen de la misma población:

$\mu_{X_i} = E(X_i) = \mu, \quad i = 1, 2, 3, . . ., n$

$\sigma^2_{X_i} = V(X_i) = \sigma^2, \quad i = 1, 2, 3, . . ., n$

Al sustituir en las fórmulas anteriores y simplificar, se completa la demostración.

La media o valor esperado $\mu_{\bar{X}}$ de la media muestral $\bar{X}$ debe entenderse como el valor que tomaría la variable aleatoria $\bar{X}$ si se tomase una cantidad muy grande de muestras y se calculara su promedio. Entonces el resultado se acercaría cada vez más al valor de μ

9.2.1 CORRECCIÓN DE LA VARIANZA

Si el tamaño **N** de la población es finito y este número no es muy grande con respecto al tamaño **n** de la muestra, se debe usar la siguiente fórmula para corregir la varianza muestral, la cual se aplica si el tamaño de la muestra es mayor al 5% del tamaño de la población.

Definición: Corrección de la Varianza

$$\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right), \quad \text{si} \quad n > 5\%N.$$

9.2.2 MEDIA MUESTRAL DE UNA POBLACIÓN NORMAL

Definición: Media Muestral de una Población Normal

Si la muestra proviene de una población con **Distribución Normal** con media μ y varianza σ^2, entonces la media muestral $\bar{X}$ tiene también **Distribución Normal** y su media y varianza son:

Media: $\mu_{\bar{X}} = \mu$

Varianza: $\sigma^2_{\bar{X}} = \dfrac{\sigma^2}{n}$

Demostración

Para demostrar que la media muestral tiene Distribución Normal, se compara la función generadora de momentos de una variable aleatoria **X** con Distribución Normal y la función generadora de momentos de la media muestral $\bar{X}$ definida con el producto de las funciones generadoras de momentos de las variables aleatorias. Se omite el desarrollo.

Ejemplo

Un fabricante especifica que la duración de sus baterías tiene Distribución Normal con media **36** meses y desviación estándar **8** meses. Calcule la probabilidad que una muestra aleatoria de **9** baterías tenga una duración media menor o igual que **30** meses.

Especificaciones para la población

X: Variable aleatoria continua (duración en meses de cada batería)

μ: Parámetro de interés (media poblacional)

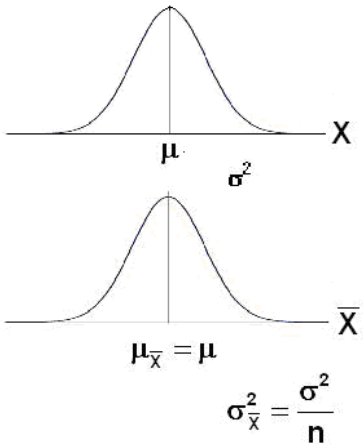
X: Tiene **Distribución Normal** con μ = **36**, $\sigma^2 = 8^2$

Datos de la muestra

Media muestral: $\bar{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$, tamaño de la muestra: **n** = **9**

Por la propiedad anterior:

$\bar{X}$ tiene aproximadamente **Distribución Normal** con $\mu_{\bar{X}} = \mu$ = **36** y $\sigma^2_{\bar{X}} = \dfrac{\sigma^2}{n} = \dfrac{8^2}{9}$ = **7.1**



La variable aleatoria y la media muestral tienen Distribución Normal aproximadamente

$$P(\bar{X} \le 30) = P\left(Z \le \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}}\right) = P\left(Z \le \frac{30 - 36}{\sqrt{7.1}}\right) = P(Z \le -2.6) = F(-2.6) = 0.0122 = 1.22\%$$

La media o valor esperado de $\bar{X}$ es igual a la media poblacional μ, por lo tanto, cualquier valor de $\bar{X}$, aunque aleatorio, debería estar razonablemente cerca de μ.

El resultado obtenido indica que la probabilidad de que la media muestral obtenida con los datos sea menor o igual al valor propuesto de 30 meses, tiene un valor muy pequeño. Esto podría interpretarse como un indicio de que la muestra no apoya a lo afirmado por el fabricante.

9.3 TEOREMA DEL LÍMITE CENTRAL

El siguiente enunciado es uno de los más importantes teoremas de la Estadística Inferencial

Definición: Teorema del Límite Central

Si $\bar{X}$ es la media de una muestra aleatoria de tamaño n extraída de una población que tiene media μ y varianza σ^2, entonces:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}},$$

es una variable aleatoria cuya función de probabilidad se aproxima a la **Distribución Normal Estándar** a medida que n aumenta

La demostración formal de este teorema requiere el manejo del límite de la función generadora de momentos de la variable aleatoria $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$.

Se puede experimentar mediante simulaciones con el computador observándose que, sin importar la distribución de probabilidad de una variable aleatoria discreta o continua **X** de la cual se muestrea, el límite de la variable aleatoria **Z** tiende a la forma tipo campana de la Distribución Normal Estándar cuando **n** crece.

Con carácter general, o al menos en los modelos de probabilidad clásicos, se admite como una aproximación aceptable al modelo Normal siempre que $n \ge 30$, y se dice que la muestra es **"grande".** Adicionalmente en este caso, si se desconoce la varianza de la población se puede usar como aproximación la varianza muestral: $\sigma^2 \cong S^2$

NOTA: El Teorema del Límite Central no implica que la distribución de la variable $\bar{X}$ tiende a la Distribución Normal a medida que **n** crece. El teorema establece que la distribución de la variable **Z** tiende a la Distribución Normal Estándar cuando **n** crece.

Ejemplo

Un fabricante especifica que cada paquete de su producto tiene un peso promedio **22.5** gr. con una desviación estándar de **2.5** gr. Calcule la probabilidad que una muestra aleatoria de **40** paquetes de este producto tenga un peso promedio menor o igual que **20** gr.

Especificaciones para la población

X: Variable aleatoria continua (peso en gr. de cada paquete)

Tiene media igual a μ**=22.5,** y varianza $\sigma^2 = 2.5^2$**. No se especifica su distribución**

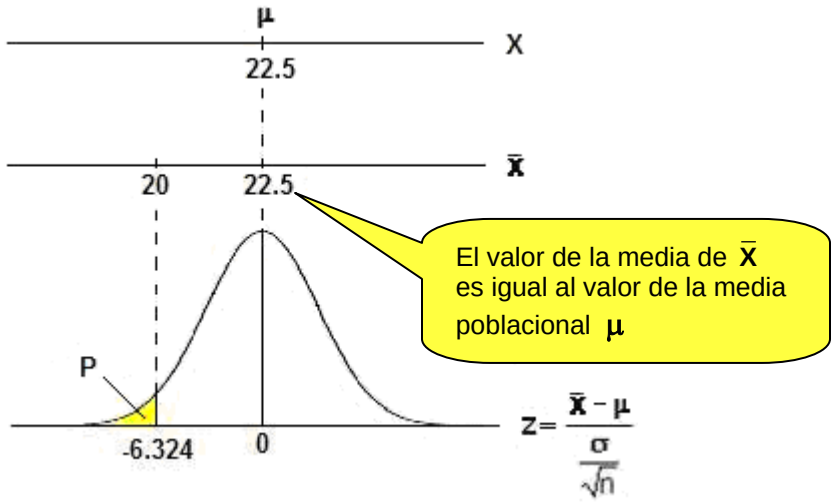
μ: Parámetro (media poblacional)

Datos de la muestra

Media muestral: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, tamaño de la muestra **n = 40,** (muestra grande),

Por el Teorema del Límite Central, la variable aleatoria $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ tiene distribución de

probabilidad aproximadamente Normal Estándar



$$P = P(\bar{X} \le 20) \cong P\left(Z \le \frac{20-\mu}{\frac{\sigma}{\sqrt{n}}}\right) = P\left(Z \le \frac{20-22.5}{\frac{2.5}{\sqrt{40}}}\right) = P(Z \le -6.3246) = F(-6.3246) \cong 0$$

Conclusión

Se observa que la probabilidad de que la media muestral tenga un valor menor o igual a 20 es aproximadamente cero, por lo tanto inferimos que lo especificado por el fabricante no es verdad.

Ejemplo

Si **X** es una variable aleatoria exponencial con parámetro β = **4** y de esta población se toma una muestra aleatoria de tamaño **36**, determine la probabilidad de que la media muestral tome algún valor entre **3.60** y **4.11**

Si la variable **X** tiene distribución exponencial, entonces su media y varianza son:

$$\mu = E(X) = \beta = 4, \qquad \sigma^2 = V(X) = \beta^2 = 16 \quad \Rightarrow \quad \sigma = 4$$

Si la muestra es grande, entonces por el Teorema del Límite Central

$$z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$ tiene Distribución Normal Estándar aproximadamente

Entonces

$$P(3.60 < \bar{X} < 4.11) = P\left(\frac{3.60 - 4}{4/\sqrt{36}} < Z < \frac{4.11 - 4}{4/\sqrt{36}}\right) = P(-0.6 < Z < 0.165) = 0.2913$$

9.3.1 EJERCICIOS

1) Una máquina envasadora de refrescos está programada para que la cantidad de líquido sea una variable aleatoria con distribución normal, con media 200 mililitros y una desviación estándar de 10 mililitros. Calcule la probabilidad que una muestra aleatoria de 20 envases tenga una media menor que 185 mililitros

2) La altura media de los alumnos de un plantel secundario es 1.50 mts. con una desviación estándar de 0.25 mts. Calcule la probabilidad que en una muestra aleatoria de 36 alumnos, la media sea superior a 1.60 mts.

9.4 LA DISTRIBUCIÓN T

La distribución **T** o de Student es una función de probabilidad con forma tipo campana simétrica Su aplicación más importante se describe a continuación.

Suponer que se toma una muestra aleatoria de tamaño **n<30** de una población con **distribución normal** con media μ y varianza **desconocida**. En este caso ya no se puede usar la variable aleatoria **Z**. En su lugar debe usarse otro estadístico denominado **T** o de Student

Este estadístico es útil cuando por consideraciones prácticas no se puede tomar una muestra aleatoria grande y se desconoce la varianza poblacional. Pero es necesario que la población tenga distribución normal.

Definición: Distribución T

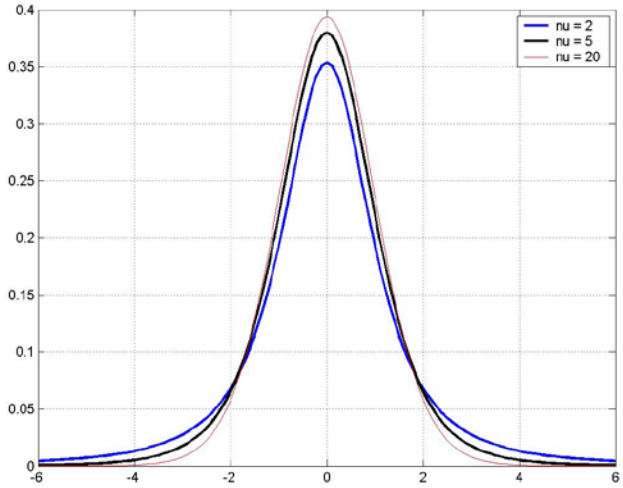
Sean $\bar{X}$ y S^2 la media y varianza de una muestra aleatoria de tamaño **n<30** tomada de una población **normal** con media μ y varianza **desconocida**, entonces la variable aleatoria

$$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}},$$

tiene distribución **T** con $\nu = n - 1$ grados de libertad

9.4.1 GRAFICO DE LA DISTRIBUCIÓN T

La forma específica de la distribución **T** depende del valor de ν, el cual es el parámetro para este modelo con la definición: $\nu = n - 1$ y se denomina "grados de libertad".

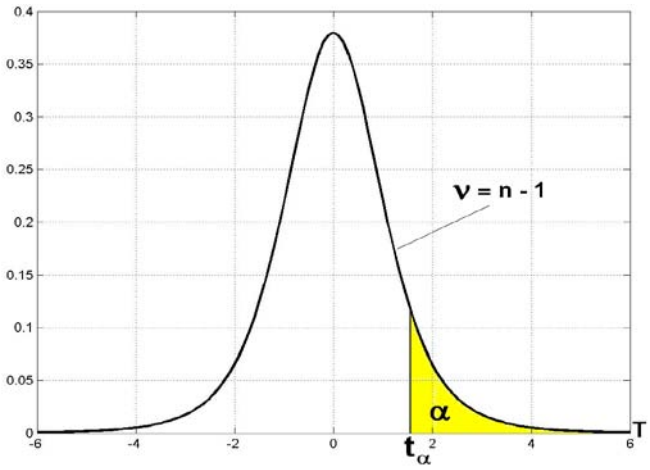


Distribución T para ν = 2, 5, 20 grados de libertad.

Para calcular probabilidad con la distribución **T**, si no se dispone de una calculadora estadística o un programa computacional estadístico, se pueden usar tablas que contienen algunos valores de esta distribución para diferentes grados de libertad con la siguiente definición:

Definición: t_α

t_α es el valor de **T** tal que el área a la derecha es igual a α: $P(T \geq t_\alpha) = \alpha$



*Uso de la distribución **T***

Ejemplo

Una población con distribución aproximadamente normal tiene una media especificada de **5.5** siendo su varianza desconocida.

Calcule la probabilidad que una muestra aleatoria de tamaño **6** tenga una media mayor o igual a **6.5** con una desviación estándar de **0.5**.

Los datos especificados corresponden a la distribución **T** con n = 6

$$\mathbf{T = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}}}$$, con $\nu = \mathbf{n - 1 = 5}$ grados de libertad

$$\mathbf{P(\bar{X} \geq 6.5) = P\left(T \geq \frac{6.5 - 5.5}{\frac{0.5}{\sqrt{6}}}\right) = P(T \geq 4.9)}$$

En la **Tabla T**, se puede observar en la fila $\nu = \mathbf{n - 1 = 5}$,

	α	.40	.25	.10	.05	.025	.01	.005	**.0025**	**.001**	.0005
ν	1	.325	1.000	3.078	6.314	12.706	31.821	63.657	127.320	318.310	636.620
	2	.289	.816	1.886	2.920	4.303	6.965	9.925	14.089	23.326	31.598
	3	.277	.765	1.638	2.353	3.182	4.541	5.841	7.453	10.213	12.924
	4	.271	.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
	5	.267	.727	1.476	2.015	2.571	3.365	4.032	**4.773**	**5.893**	6.869
	6	.265	.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
	7	.263	.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408

Aquí se ubica t = 4.9

$\mathbf{t_{0.0025} = 4.773}$: $\mathbf{P(T \geq 4.773) = 0.0025}$

$\mathbf{t_{0.001} = 5.893}$: $\mathbf{P(T \geq 5.893) = 0.001}$

Por lo tanto $\mathbf{0.001 \leq P(T \geq 4.9) \leq 0.0025}$

Se puede concluir que $\mathbf{0.001 \leq P(\bar{X} \geq 6.5) \leq 0.0025}$

Mediante una interpolación lineal se puede obtener una aproximación más precisa.

9.5 LA DISTRIBUCIÓN JI-CUADRADO

Esta distribución se la obtiene de la distribución gamma. Tiene forma tipo campana con sesgo positivo. Se puede demostrar que si **X** es una variable aleatoria con distribución normal, entonces $\mathbf{X^2}$ es una variable aleatoria con distribución **ji-cuadrado**. Este hecho explica la importancia de la distribución **ji-cuadrado** en problemas de muestreo de poblaciones con distribución normal. Una aplicación práctica es la estimación de la varianza poblacional.

Definición: Distribución Ji-Cuadrado

Sean $\bar{X}$ y S^2 la media y varianza de una muestra aleatoria de tamaño **n** tomada de una población normal con media μ y varianza σ^2, entonces la variable aleatoria

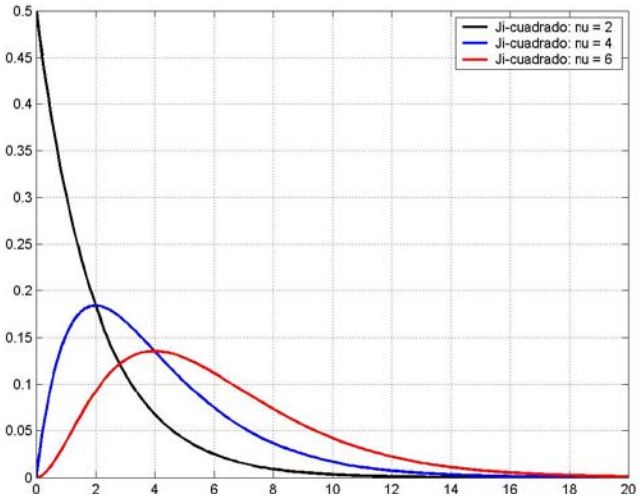
$$\chi^2 = (n-1)\frac{S^2}{\sigma^2},$$

tiene distribución **Ji-cuadrado** con $\nu = n - 1$ grados de libertad

El valor esperado de la variable χ^2 es $E(\chi^2) = n - 1$

9.5.1 GRÁFICO DE LA DISTRIBUCIÓN JI-CUADRADO

La forma específica de esta distribución de probabilidad depende del valor de ν, el cual es el parámetro para este modelo con la definición $\nu = n - 1$ y se denomina "grados de libertad"

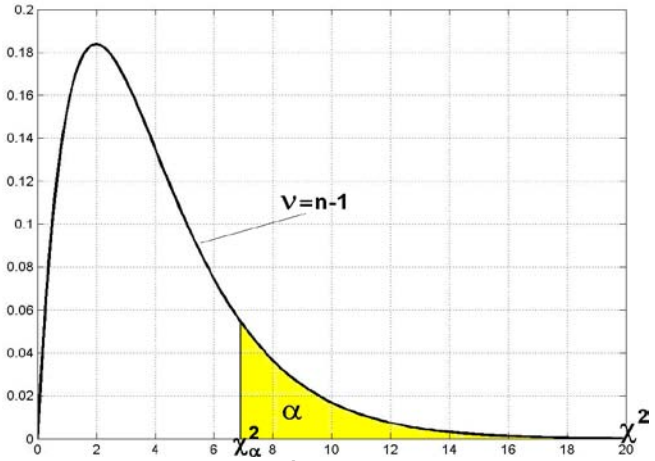


La distribución ji-cuadrado con $\nu = 2, 4, 6$

Algunos valores de la distribución ji-cuadrado están tabulados para ciertos valores de ν y para valores típicos de α con la siguiente definición

Definición: χ^2_α

χ^2_α es el valor de χ^2 tal que el área a la derecha es igual a α: $P(\chi^2 \geq \chi^2_\alpha) = \alpha$



Uso de la distribución ji-cuadrado

Ejemplo

Una población con distribución aproximadamente normal tiene varianza especificada de **0.8**. Calcule la probabilidad que una muestra aleatoria de tamaño **6** tenga una varianza mayor o igual a **1.2**.

Los datos especificados corresponden al uso de la distribución ji-cuadrado:

$$\chi^2 = (n-1)\frac{S^2}{\sigma^2}, \text{ con } \nu = n-1 \text{ grados de libertad}$$

$$P(S^2 > 1.2) = P\left(\chi^2 > (n-1)\frac{S^2}{\sigma^2}\right) = P\left(\chi^2 > (6-1)\frac{1.2}{0.8}\right) = P(\chi^2 > 7.5)$$

En la **Tabla_ji-cuadrado** se puede observar en la fila $\nu = n-1 = 5$

ν \ α	.995	.990	.975	.950	.900	**.500**	**.100**	.050	.025	.010	.005
1	.00003	.0001	.0009	.0039	.02	.45	2.71	3.84	5.02	6.63	7.88
2	.01	.02	.05	.10	.21	1.39	4.61	5.99	7.38	9.21	10.60
3	.07	.11	.22	.35	.58	2.37	6.25	7.81	9.35	11.34	12.84
4	.21	.30	.48	.71	1.06	3.36	7.78	9.49	11.14	13.28	14.86
5	.41	.55	.83	1.15	1.61	**4.35**	**9.24**	11.07	12.83	15.09	16.75
6	.68	.87	.24	1.64	2.20	5.35	10.65	12.59	14.45	16.81	18.55
7	.99	.24	.69	2.17	2.83	6.35	12.02	14.07	16.01	18.48	20.28

Aquí se ubica χ^2=7.5

$\chi^2_{0.5}$ **= 4.36:** $P(\chi^2 \geq 4.35) = 0.5$

$\chi^2_{0.1}$ **= 9.24:** $P(\chi^2 \geq 9.24) = 0.1$

Por lo tanto $0.1 \leq P(\chi^2 \geq 7.5) \leq 0.5$

Con lo cual se puede concluir que $0.1 \leq P(S^2 \geq 1.2) \leq 0.5$

Mediante una interpolación lineal se puede obtener una aproximación mas precisa.

9.6 DISTRIBUCIÓN F

Esta distribución es útil para realizar inferencias con las varianzas de dos poblaciones normales usando los datos de las varianzas de dos muestras aleatorias independientes con la siguiente definición.

Definición: Distribución F

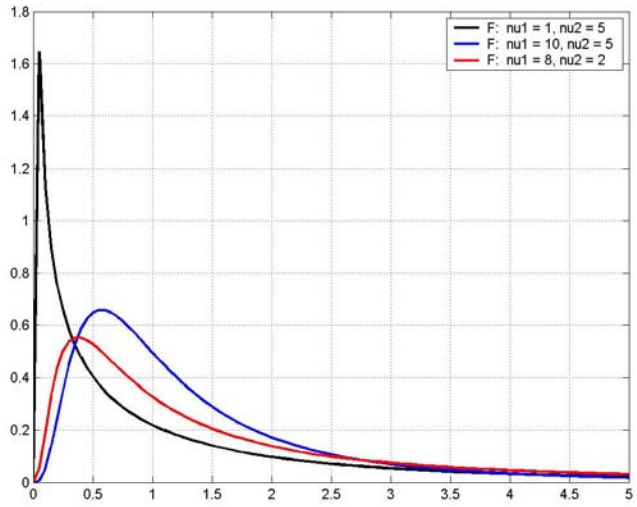
Sean S_1^2 y S_2^2 las varianzas de dos muestras aleatorias independientes de tamaño n_1 y n_2 tomadas de poblaciones **normales** con varianzas σ_1^2, σ_2^2, entonces la variable aleatoria

$$F = \frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2}$$

tiene distribución **F** con $\nu_1 = n_1 - 1$, $\nu_2 = n_2 - 1$ grados de libertad

9.6.1 GRÁFICO DE LA DISTRIBUCIÓN F

La distribución **F** tiene forma tipo campana con sesgo positivo y depende de dos parámetros para este modelo: ν_1 , ν_2 los cuales se denominan "grados de libertad"

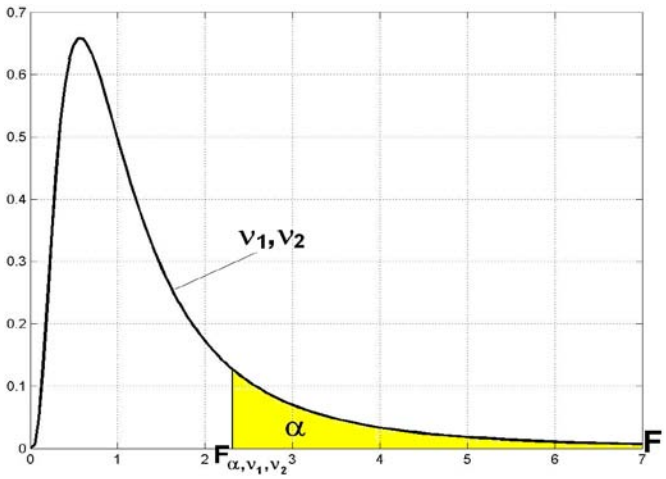


*La distribución **F** para varios* ν_1 , ν_2

Algunos valores de esta distribución están tabulados para valores específicos de α, ν_1, ν_2 de acuerdo a la siguiente definición:

Definición: F_{α,ν_1,ν_2}

F_{α,ν_1,ν_2} es el valor de **F** tal que el área a la derecha es igual a α: $P(F \geq F_{\alpha,\nu_1,\nu_2}) = \alpha$



Uso de la distribución **F**

La siguiente es una relación útil para obtener otros valores de la distribución **F**:

$$F_{1-\alpha,\nu_1,\nu_2} = \frac{1}{F_{\alpha,\nu_2,\nu_1}}$$

Ejemplo

Calcule **F** con $\alpha = 0.05$ y $\alpha = 0.95$ si $\nu_1 = 9$, $\nu_2 = 7$

Tabla **F** para $\alpha = 0.05$

ν_2 \ ν_1	1	2	3	4	5	6	7	**8**	**9**	10	12	15	20	24	30	40	60	120	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	**3.68**	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	**3.29**	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71

$F_{0.05,\,9,\,7} = 3.68$

$$F_{0.95,\,9,\,7} = \frac{1}{F_{0.05,\,7,\,9}} = \frac{1}{3.29} = 0.304$$

9.7 ESTADÍSTICAS DE ORDEN

Sea una población infinita con densidad de probabilidad continua de la que se toma una muestra aleatoria de tamaño **n** y se obtienen los valores:

$$x_1, x_2, x_3, \ldots, x_n.$$

Los datos se los escribe en orden creciente:

$$x_{(1)}, x_{(2)}, x_{(3)}, \ldots, x_{(n)}$$

Estos valores son instancias de las variables aleatorias

$$X_{(1)}, X_{(2)}, X_{(3)}, \ldots, X_{(n)}$$

Las variables definidas se denominan estadísticas de orden

Definición: Estadísticas de Orden para una Muestra Aleatoria de Tamaño n

$$X_{(1)}, X_{(2)}, X_{(3)}, \ldots, X_{(n)}$$

9.7.1 DENSIDAD DE PROBABILIDAD DE LAS ESTADÍSTICAS DE ORDEN

Se puede probar que si **f** y **F** son respectivamente la densidad y la distribución acumulada de **X**, entonces la densidad f_r del estadístico de orden **r** es

Definición: Densidad de Probabilidad de la Estadística de Orden r

$$f_r(x_{(r)}) = \frac{n!}{(r-1)!(n-r)!}[F(x_{(r)})]^{r-1}[1-F(x_{(r)})]^{n-1}f(x_{(r)}), \quad x_{(r)} \in \Re$$

Ejemplo. Se tiene una población cuyos elementos están definidos por una variable aleatoria contínua **X** con densidad de probabilidad:

$$f(x) = \begin{cases} kx, & 0 < x < 1 \\ 0, & \text{para otro } x \end{cases}$$

De esta población se toma una muestra aleatoria de tamaño $n = 5$

Encuentre las estadísticas de orden 1, 2, 3, 4, 5

Solución

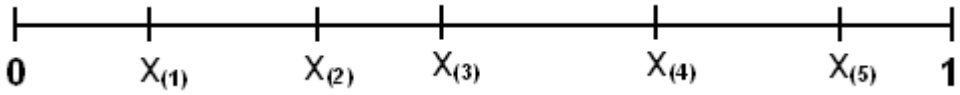
Primero determinamos el valor de **k** con la propiedad:

$$\int_0^1 f(x)dx = \int_0^1 kxdx = k\left[\frac{x^2}{2}\right]_0^1 = \frac{k}{2} = 1 \quad \Rightarrow \quad k = 2 \quad \Rightarrow \quad f(x) = \begin{cases} 2x, & 0 < x < 1 \\ 0, & \text{para otro } x \end{cases}$$

Densidad de la variable poblacional: $f(x) = 2x, \ 0 < x < 1$

Su distribución acumulada: $F(x) = \int_0^x 2xdx = x^2, \quad -\infty < x < \infty$

Estadísticas de orden para la muestra aleatoria de tamaño **n = 5**



Densidad del estadístico de orden **r** para **n = 5, r = 1, 2, 3, 4, 5**

$$f_r(x_{(r)}) = \frac{5!}{(r-1)!(5-r)!}[F(x_{(r)})]^{r-1}[1-F(x_{(r)})]^{5-r} f(x_{(r)}), \quad x_{(r)} \in \Re$$

Densidad del estadístico de orden uno

r = 1, n = 5, f(x) = 2x, 0 < x < 1, F(x) = x^2, con la notación: **x = $x_{(r)}$**

$$f_1(x) = \frac{5!}{(1-1)!(5-1)!}[x^2]^{1-1}[1-x^2)]^{5-1}(2x), \quad x \in (0,1)$$

Simpificando se obtiene

$$f_1(x) = 10(x)(1-x^2)^4, \qquad 0 < x < 1$$

Sucesivamente se obtienen las densidades de los otros estadísticos de orden

r = 2, n = 5, f(x) = 2x, F(x) = x^2, con la notación: **x = $x_{(r)}$**

$$f_2(x) = 40x^3(1-x^2)^3, \qquad 0 < x < 1$$

r = 3, n = 5, f(x) = 2x, F(x) = x^2, con la notación: **x = $x_{(r)}$**

$$f_3(x) = 60x^5(1-x^2)^2, \qquad 0 < x < 1$$

r = 4, n = 5, f(x) = 2x, F(x) = x^2, con la notación: **x = $x_{(r)}$**

$$f_4(x) = 40x^7(1-x^2), \qquad 0 < x < 1$$

r = 5, n = 5, f(x) = 2x, F(x) = x^2, con la notación: **x = $x_{(r)}$**

$$f_5(x) = 10x^9, \qquad 0 < x < 1$$

Determine la probabilidad que la estadística de orden cuatro tome un valor menor que 1/2

$$P(X_{(4)} < 1/2) = \int_0^{1/2} 40x^7(1-x^2)dx = 1/64$$

Graficar las densidades de las estadísticas de orden obtenidas

Gráfico de $f_1(x)$, $0 < x < 1$

Extremos $f_1(0) = 0$, $f_1(1) = 0$

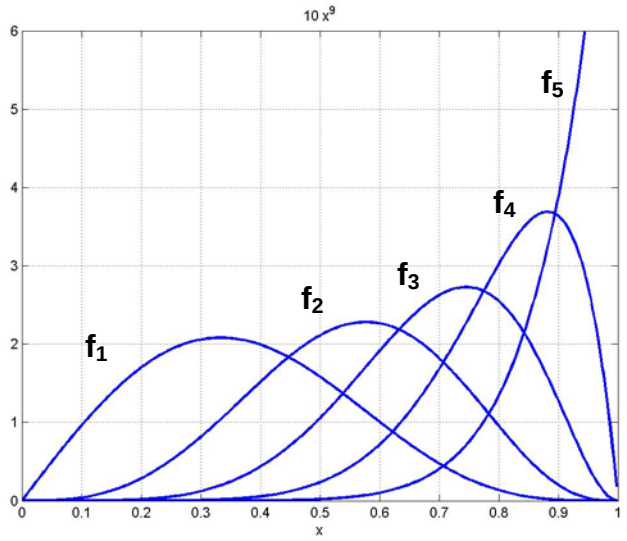
Máximo: $f_1'(x) = 10(1-x^2)^4 - 80x^2(1-x^2)^3 = 10(1-x^2)^3[(1-x^2) - 8x^2] = 0$

$\Rightarrow \quad (1-x^2)^3 = 0 \quad \Rightarrow \quad x = \pm 1$

$(1-x^2) - 8x^2 = 1 - 9x^2 = 0 \Rightarrow x = \pm\frac{1}{3}$

Máximo: **(1/3, 2.081)**

Gráficos de las densidades de las estadísticas de orden



9.7.2 EJERCICIOS

1) a) Encuentre $t_{0.1}$ **con** ν**=18**.
b) Encuentre t_α dado que **P(t>t_α) = 0.05,** ν**=16**

2) Una población normal tiene especificada su media con el valor **5**. Calcule la probabilidad que una muestra de **6** observaciones tenga una media menor que 4 con varianza de **1.2**

3) Una población con distribución aproximadamente normal tiene varianza especificada de **1.4**. Calcule la probabilidad que una muestra aleatoria de tamaño 8 tenga una varianza menor que **0.8**

4) Calcule **F** con $\alpha = 0.05$ y $\alpha = 0.95$ si $\nu_1 = 15$, $\nu_2 = 20$

5) Se tiene una población cuya variable aleatoria **X** tiene la siguiente densidad de probabilidad:

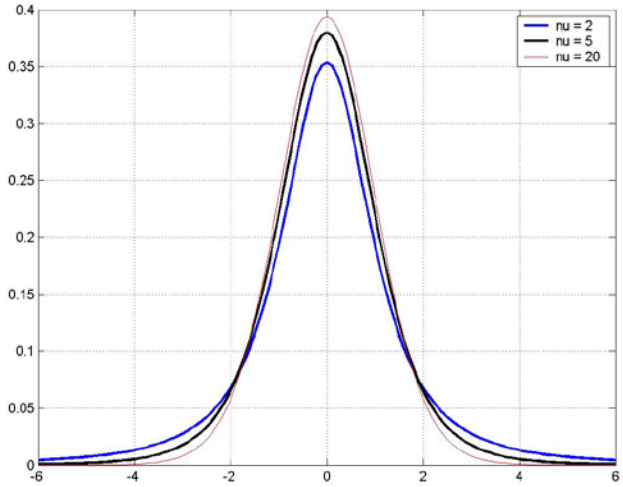
$$f(x) = \begin{cases} \frac{2}{5}(x+1), & 1 < x < 2 \\ 0, & \text{para otro } x \end{cases}$$

Si se toma una muestra aleatoria de tamaño **n=**4, calcule la probabilidad que la estadística de orden dos tome un valor mayor que **1.5**

MATLAB

Graficar la densidad de la distribución T

```
>> t=-6:0.1:6;                          Puntos para evaluar la distribución T
>> f1=tpdf(t, 2);                       Puntos de la distribución T
>> f2=tpdf(t, 5);
>> f3=tpdf(t, 20);
>> plot(t,f1,'b'), grid on, hold on     Graficación
>> plot(t,f2,'k')
>> plot(t,f3,'r')
>> legend('nu=2','nu=5','nu=20')        Rótulos
```



Obtener y graficar las estadísticas de orden 1, 2, 3, 4, 5 para f(x) = 2x, 0<x<1

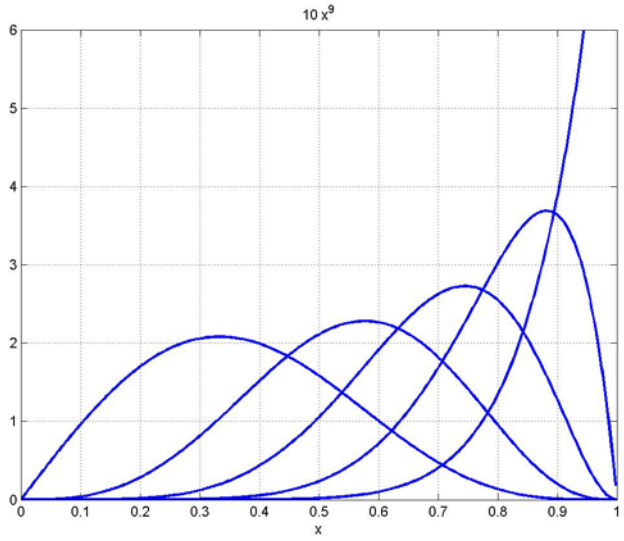
```
>> syms x r
>> f=2*x;                               Definición de la densidad  f(x)
>> F=int(f)                             Obtención de la distribución  F(x)
    F =  x^2
```

Obtención de estadísticas de orden

```
>> r=1;f1=factorial(5)/(factorial(r-1)*factorial(5-r))*F^(r-1)*(1-F)^(5-r)*f
    f1 =  10*(1-x^2)^4*x

>> r=2;f2=factorial(5)/(factorial(r-1)*factorial(5-r))*F^(r-1)*(1-F)^(5-r)*f
    f2 =  40*x^3*(1-x^2)^3

>> r=3;f3=factorial(5)/(factorial(r-1)*factorial(5-r))*F^(r-1)*(1-F)^(5-r)*f
    f3 =  60*x^5*(1-x^2)^2

>> r=4;f4=factorial(5)/(factorial(r-1)*factorial(5-r))*F^(r-1)*(1-F)^(5-r)*f
    f4 =  40*x^7*(1-x^2)

>> r=5;f5=factorial(5)/(factorial(r-1)*factorial(5-r))*F^(r-1)*(1-F)^(5-r)*f
    f5 =  10*x^9
```

Gráficación de las estadísticas de orden.

```
>> ezplot(f1,[0,1]), grid on,hold on
>> ezplot(f2,[0,1])
>> ezplot(f3,[0,1])
>> ezplot(f4,[0,1])
>> ezplot(f5,[0,1])
```



Calcular para la estadística de orden 4: $P(X_{(4)} < 1/2)$

```
>> p = int(f4, 0, 1/2)
    p =  1/64
```

10 ESTADÍSTICA INFERENCIAL

La Estadística Inferencial proporciona las técnicas para formular proposiciones acerca de la población, incluyendo una medida para determinar el riesgo de la afirmación.

10.1 INFERENCIA ESTADÍSTICA

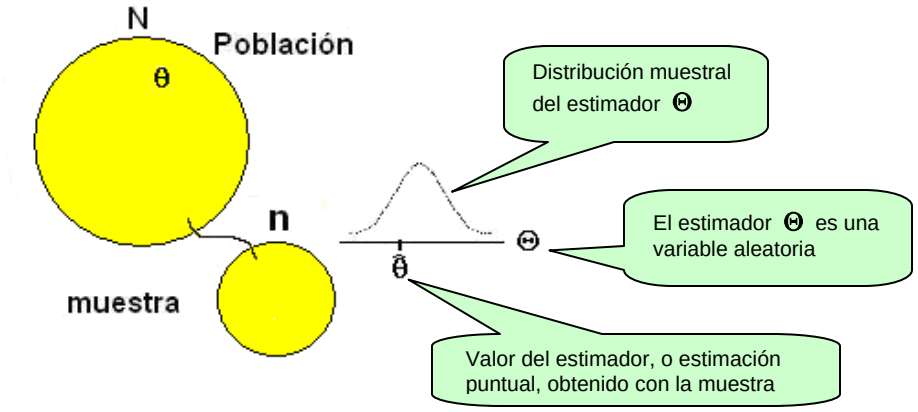
Una inferencia estadística es una afirmación que se hace acerca de la población en base a la información contenida en una muestra aleatoria tomada de esta población.

Debido a la naturaleza aleatoria de los datos obtenidos en la muestra, hay un riesgo en la certeza de la afirmación propuesta, y es necesario cuantificar el valor de este riesgo.

Un **estimador** es una variable aleatoria cuyas propiedades permiten estimar el valor del parámetro poblacional de interés. La muestra aleatoria proporciona únicamente un valor de esta variable y se denomina estimación puntual.

Para estimar al parámetro poblacional, es posible definir más de un estimador, por ejemplo para a la media poblacional μ pueden elegirse la mediana muestral $\tilde{X}$ o la media muestral $\bar{X}$. Cada uno tiene sus propias características, por lo tanto, es necesario establecer criterios para elegirlo.

Sean θ **:** Parámetro poblacional de interés (Ej. μ) (Valor desconocido)
Θ **:** Estimador (Ej. $\bar{X}$) (Variable aleatoria)
$\hat{\theta}$ **:** Estimación puntual de Θ (Ej. $\bar{x}$) (Un valor del estimador)



La intuición sugiere que el estimador debe tener una distribución muestral concentrada alrededor del parámetro y que la varianza del estimador debe ser la menor posible. De esta manera, el valor que se obtiene en la muestra será cercano al valor del parámetro y será útil para estimarlo.

10.2 MÉTODOS DE INFERENCIA ESTADÍSTICA

Sean θ : Parámetro poblacional de interés (Ej. μ) (Valor desconocido)
Θ : Estimador (Ej. $\bar{X}$) (Variable aleatoria)
$\hat{\theta}$: Estimación puntual de Θ (Ej. $\bar{x}$) (Un valor del estimador)

10.2.1 ESTIMACIÓN PUNTUAL

Se trata de determinar la distancia, o error máximo entre la estimación puntual $\hat{\theta}$ y el valor del parámetro θ que se desea estimar, con algún nivel de certeza especificado.

$$|\hat{\theta} - \theta|$$

10.2.2 ESTIMACIÓN POR INTERVALO

Con el valor $\hat{\theta}$ del estimador Θ se construye un intervalo que contenga al valor del parámetro θ que se desea estimar, con algún nivel de certeza especificado.

$$\mathbf{Li} \leq \theta \leq \mathbf{Ls}$$

En donde **Li** y **Ls** son los límites inferior y superior del intervalo

10.2.3 PRUEBA DE HIPÓTESIS

Se formula una hipótesis acerca del parámetro θ asignándole un valor supuesto θ_0 y con el valor $\hat{\theta}$ del estimador Θ se realiza una prueba para aceptar o rechazar la hipótesis propuesta con algún nivel de certeza especificado.

Hipótesis propuesta: $\theta = \theta_0$

10.3 PROPIEDADES DE LOS ESTIMADORES

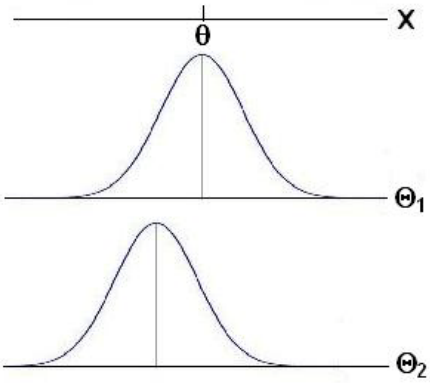
Las siguientes definiciones establecen las características deseables de los estimadores

Sean θ: Parámetro poblacional que se desea estimar.
Θ: Estimador

Definición 1: Estimador insesgado

> Se dice que el estimador Θ es un estimador insesgado del parámetro θ si $\mathbf{E(\Theta) = \theta}$

Un estimador insesgado es aquel cuya media o valor esperado coincide con el parámetro que se quiere estimar.



En el gráfico se observa que Θ_1 es un estimador insesgado del parámetro θ pues $\mathbf{E(\Theta_1) = \theta}$.
En cambio, Θ_2 no es un estimador insesgado del parámetro θ pues $\mathbf{E(\Theta_2) \neq \theta}$.

Debido a lo anterior, es mas probable que una estimación puntual de Θ_1 esté más cercana al parámetro θ, que una estimación puntual de Θ_2

Ejemplo. La media muestral $\bar{X}$ es un estimador insesgado del parámetro μ (media poblacional)

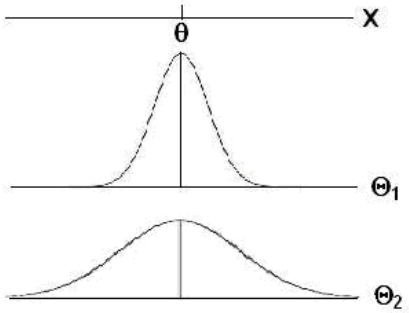
Demostración:

$$E(\bar{X}) = E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\left[\sum_{i=1}^{n} E[x_i]\right] = \frac{1}{n}\left[\sum_{i=1}^{n} \mu\right] = \frac{1}{n} n\mu = \mu$$

Definición 2: Estimador más eficiente

Se dice que un estimador Θ_1 es más eficiente que otro estimador Θ_2 si ambos son insesgados y además $V(\Theta_1) < V(\Theta_2)$

Un estimador es más eficiente si tiene menor varianza.



En el gráfico se observa que Θ_1 es un estimador más eficiente del parámetro θ, que el estimador Θ_2 pues ambos son insesgados pero la varianza de Θ_1 es menor que la varianza de Θ_2. Por lo tanto, es mas probable que una estimación puntual de Θ_1 esté más cercana al valor de θ, que una estimación puntual de Θ_2

Definición 3: Estimador consistente

Se dice que un estimador Θ es un estimador consistente del parámetro θ si Θ es un estimador insesgado de θ y $\lim_{n\to\infty} V(\Theta) = 0$

Ejemplo. La media muestral $\bar{X}$ es un estimador consistente de μ

Demostración:

$$V(\bar{X}) = \frac{\sigma^2}{n} \Rightarrow \lim_{n\to\infty} \frac{\sigma^2}{n} = 0 \Rightarrow \bar{X} \to \mu$$

Definición 4: Sesgo de un estimador

El sesgo **B** de un estimador Θ está dado por

$B = E(\Theta) - \theta$

Es la diferencia entre el valor esperado del estadístico y el valor del parámetro.

De acuerdo con la definición anterior, el sesgo de un estimador insesgado es cero pues $E(\Theta_1) = \theta$.

Definición 5: Error cuadrático medio (ECM)

Es el valor esperado del cuadrado de la diferencia entre el estimador Θ y el parámetro θ:

$$ECM(\Theta) = E[\Theta - \theta]^2$$

Si se desarrolla el cuadrado y se sustituye la definición de varianza y de sesgo se obtiene:

$$ECM(\Theta) = V(\Theta) + [E(\Theta) - \theta]^2 = V(\Theta) + B^2$$

Esta definición resume las características deseables de un estimador: su varianza debe ser mínima y su distribución de muestreo debe estar concentrada alrededor del parámetro que es estimado, es decir el sesgo debe ser mínimo.

Ejemplo

Pruebe que la varianza muestral es un estimador insesgado de la varianza poblacional si se toma una muestra de tamaño n de una población normal con media μ y varianza σ^2

Sea $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Se tiene que probar que $E(S^2) = \sigma^2$

Primero expresamos la varianza muestral en una forma conveniente

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \frac{1}{n-1}(\sum_{i=1}^{n}x_i^2 - 2\bar{x}\sum_{i=1}^{n}x_i + \sum_{i=1}^{n}\bar{x}^2)$$

$$= \frac{1}{n-1}(\sum_{i=1}^{n}x_i^2 - 2\bar{x}(n\bar{x}) + n\bar{x}^2) = \frac{1}{n-1}(\sum_{i=1}^{n}x_i^2 - 2n\bar{x}^2 + n\bar{x}^2) = \frac{1}{n-1}(\sum_{i=1}^{n}x_i^2 - n\bar{x}^2)$$

Con la definición de valor esperado

$$E(S^2) = E[\frac{1}{n-1}(\sum_{i=1}^{n}X_i^2 - n\bar{X}^2)] = \frac{1}{n-1}[\sum_{i=1}^{n}E(X_i^2) - nE(\bar{X}^2)]$$

Cada variable X_i proviene de la misma población con varianza σ^2 y media μ

$$\sigma_{x_i}^2 = \sigma^2 = E(X_i^2) - E^2(X_i) = E(X_i^2) - \mu^2 \quad \Rightarrow \quad E(X_i^2) = \sigma^2 + \mu^2$$

La media muestral es una variable aleatoria con media μ y varianza σ^2/n

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} = E(\bar{X}^2) - E^2(\bar{X}) = E(\bar{X}^2) - \mu^2 \quad \Rightarrow \quad E(\bar{X}^2) = \frac{\sigma^2}{n} + \mu^2$$

Se sustituyen en la definición anterior con lo cual se completa la demostración

$$E(S^2) = \frac{1}{n-1}(\sum_{i=1}^{n}(\sigma^2 + \mu^2) - n(\frac{\sigma^2}{n} + \mu^2)) = \frac{1}{n-1}(n\sigma^2 + n\mu - \sigma^2 - n\mu)$$

$$= \frac{\sigma^2}{n-1}(n-1) = \sigma^2$$

Ejemplo

Se tiene una población de tamaño **N = 6** definida por: **{1, 2, 3, 3, 4, 5}**

a) Calcule la media de la población μ
b) Calcule la varianza de la población σ^2
c) Especifique cuales son todas las muestras de tamaño **n = 3** que se pueden obtener
d) Determine la distribución de la media muestral
e) Determine la distribución de la mediana muestral
f) Verifique que la media muestral es un estimador insesgado
g) Verifique si la mediana muestral es un estimador insesgado
h) Verifique que la media muestral es un estimador mas eficiente que la mediana muestral

Solución

a) Calcule la media de la población μ

De la población especificada se deduce que la distribución de probabilidad es:

$$f(x) = P(X = x) = \begin{cases} 1/6, & x = 1,\ 2,\ 4,\ 5 \\ 2/6, & x = 3 \\ 0, & \text{otro } x \end{cases}$$

$$\mu = \sum_x xf(x) = 1(1/6) + 2(1/6) + 3(2/6) + 4(1/6) + 5(1/6) = 3$$

b) Calcule la varianza de la población σ^2

$$\sigma^2 = E(X^2) - E^2(X)$$

$$E(X^2) = \sum_x x^2 f(x) = 1^2\,(1/6) + 2^2\,(1/6) + 3^2\,(2/6) + 4^2\,(1/6) + 5^2\,(1/6) = 32/3$$

$$\sigma^2 = 32/3 - 3^2 = 5/3$$

c) Especifique cuales son todas las muestras de tamaño n = 3 que se pueden obtener

Cantidad de muestras de tamaño 3

$$\binom{N}{n} = \binom{6}{3} = \frac{6!}{3!\ 3!} = 20$$ (Las muestras son combinaciones)

Muestras	Cantidad	Media muestral $\bar{x}$	Mediana muestral $\tilde{x}$
(1, 2, 3)	2 (*)	6/3	2
(1, 2, 4)	1	7/3	2
(1, 2, 5)	1	8/3	2
(1, 3, 3)	1	7/3	3
(1, 3, 4)	2	8/3	3
(1, 3, 5)	2	9/3	3
(1, 4, 5)	1	10/3	4
(2, 3, 3)	1	8/3	3
(2, 3, 4)	2	9/3	3
(2, 3, 5)	2	10/3	3
(2, 4, 5)	1	11/3	4
(3, 3, 4)	1	10/3	3
(3, 3, 5)	1	11/3	3
(3, 4, 5)	2	12/3	4
Total	20		

(*) La cantidad de formas diferentes de tomar el elemento 1, existiendo solamente uno en la población, el elemento 2, existiendo solamente uno en la población, y el elemento 3, del cual existen dos en la población es:

$$\binom{1}{1}\binom{1}{1}\binom{2}{1} = 2 \text{ , etc}$$

Las muestras son combinaciones, por lo tanto el orden de los elementos no es de interés.

d) Determine la distribución de probabilidad de la media muestral $\bar{X}$

Media muestral $\bar{x}$	$f(\bar{x}) = P(\bar{X} = \bar{x})$
6/3	2/20
7/3	2/20
8/3	4/20
9/3	4/20
10/3	4/20
11/3	2/20
12/3	2/20
Total	1

e) Determine la distribución de probabilidad de la mediana muestral $\tilde{X}$

Mediana muestral $\tilde{x}$	$f(\tilde{x}) = P(\tilde{X} = \tilde{x})$
2	4/20
3	12/20
4	4/20
Total	1

f) Verifique que la media muestral es un estimador insesgado de μ

$$\mu_{\bar{X}} = E(\bar{X}) = \sum_{\bar{x}} \bar{x}\, f(\bar{x}) = (6/3)(2/20) + (7/3)(2/20) + \ldots + (12/3)(2/20) = 3$$

$E(\bar{X}) = 3 = \mu \quad \Rightarrow \quad \bar{X}$ es un estimador insesgado de μ

g) Verifique en este ejemplo si la mediana muestral es un estimador insesgado de μ

$$\mu_{\tilde{X}} = E(\tilde{X}) = \sum_{\tilde{x}} \tilde{x}\, f(\tilde{x}) = 2(4/20) + 3(12/20) + 4(4/20) = 3$$

$E(\tilde{X}) = 3 = \mu \quad \Rightarrow \quad \tilde{X}$ es un estimador insesgado de μ

Nota: La media muestral es un estimador insesgado de μ. En cambio, la mediana es un estimador insesgado de μ únicamente cuando la distribución de probabilidad de la variable **X** es simétrica alrededor de μ

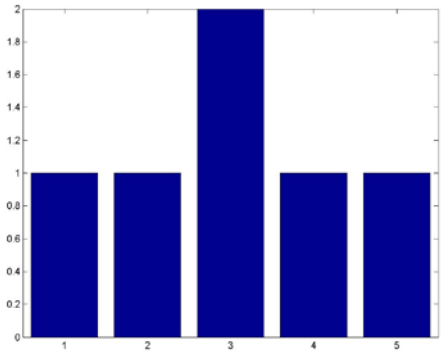


Diagrama de barras de la variable aleatoria **X**

h) Verifique que la media muestral es un estimador más eficiente que la mediana muestral

Se deben comparar las varianzas de los estimadores $\bar{X}$ y $\tilde{X}$

$$\sigma^2_{\bar{X}} = V(\bar{X}) = E(\bar{X}^2) - E^2(\bar{X})$$

$$E(\bar{X}^2) = \sum_{\bar{x}} \bar{x}^2 f(\bar{x}) = (6/3)^2\,(2/20) + (7/3)^2\,(2/20) + \ldots + (12/3)^2\,(2/20) = 9.333$$

$$V(\bar{X}) = 9.333 - 3^2 = 0.333$$

$$\sigma^2_{\tilde{X}} = V(\tilde{X}) = E(\tilde{X}^2) - E^2(\tilde{X})$$

$$E(\tilde{X}^2) = \sum_{\tilde{x}} \tilde{x}^2 f(\tilde{x}) = 2^2\,(4/20) + 3^2\,(12/20) + 4^2\,(4/20) = 47/5$$

$$V(\tilde{X}) = 47/5 - 3^2 = 0.4$$

$V(\bar{X}) < V(\tilde{X}) \Rightarrow$ La media muestral $\bar{X}$ es un estimador más eficiente que la mediana muestral $\tilde{X}$ para estimar a la media poblacional μ

10.3.1 EJERCICIOS

1) Suponga que se tiene una población cuyos elementos son: **{ 3, 4, 4, 6}** de la cual se toman muestras de tamaño **2**.

- **a)** Escriba el conjunto de todas las muestras de tamaño **2** que se pueden obtener con los elementos de la población dada.
- **b)** Grafique el histograma de frecuencias de la media muestral
- **c)** Determine la distribución de probabilidad de la media muestral
- **d)** Demuestre que la media muestral es un estimador insesgado de la media poblacional.

2) Si se toma una muestra de tamaño **n = 3** de una población cuya distribución de probabilidades está dada por

$$f(x) = \begin{cases} \frac{1}{10}x, & x = 1,2,3,4 \\ 0, & \text{otro } x \end{cases}$$

Determine si la mediana muestral es un estimador más eficiente que la media muestral para estimar a la media poblacional.

Sugerencia: Asocie la distribución de probabilidad de la variable aleatoria **X** a la siguiente población: **{ 1, 2, 2, 3, 3, 3, 4, 4, 4, 4** } y liste todas las muestras de tamaño **3**

MATLAB

Estudio de estimadores de la media poblacional

```
>> x=[1 2 3 3 4 5];                    Población
>> format rat                          Formato para ver números racionales
>> mu = mean(x)                        Media poblacional
    mu =
         3
>> sigma2 = var(x, 1)                  Varianza poblacional.  (Se escribe var(x)
    sigma2 =                                para varianza de una muestra)
          5/3
>> muestras=combnk(x,3)                Lista de las muestras de tamaño 3
   muestras =
    3    4    5
    3    4    5
    3    3    5
    3    3    4
    2    4    5
    2    3    5
    2    3    4
    2    3    5
    2    3    4
    2    3    3
    1    4    5
    1    3    5
    1    3    4
    1    3    5
    1    3    4
    1    3    3
    1    2    5
    1    2    4
    1    2    3
    1    2    3

>> n=length(muestras)                  Cantidad de muestras de tamaño 3
n =
   20
>> medias = mean(muestras' )           Lista de las medias de las 20 muestras
   medias =
     4        4       11/3     10/3     11/3     10/3       3       10/3       3       8/3
   10/3       3        8/3       3        8/3      7/3      8/3      7/3        2        2
```

```
>> medianas = median(muestras' )
    medianas =
      4     4     3     3     4     3     3     3     3     3
      4     3     3     3     3     3     2     2     2     2
```

Lista de las medianas de las 20 muestras

```
>> mmedias = mean(medias)
    mmedias =
          3
```

Media de las medias muestrales

(estimador insesgado)

```
>> mmedianas=mean(medianas)
    mmedianas =
     3
```

Media de las medianas muestrales

Coincide con la media poblacional

```
>> vmedias =var(medias', 1)
    vmedias =
              1/3
```

varianza de la media muestral

```
>> vmedianas=var(medianas', 1)
   vmedianas =
              2/5
```

Varianza de la mediana muestral

La varianza de la mediana muestral es mayor a la varianza de la media muestral, por lo tanto la media muestral es un estimador más eficiente de la media poblacional

10.4 INFERENCIAS RELACIONADAS CON LA MEDIA

10.4.1 ESTIMACIÓN PUNTUAL DE LA MEDIA

Caso: Muestras grandes (n ≥ 30)

Parámetro: μ (Es la medida poblacional cuyo valor se desea estimar)

Población con distribución desconocida, varianza σ^2

Estimador: $\bar{X}$ (Media muestral)

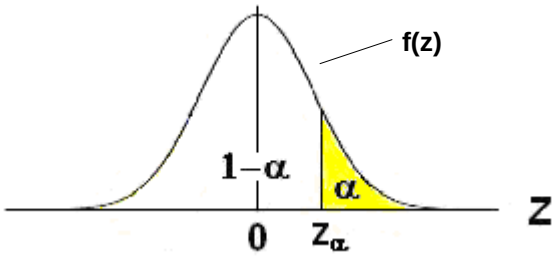
$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i\,, \quad \text{Media: } \mu_{\bar{X}} = \mu\,, \quad \text{Varianza: } \sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$$

Siendo la muestra grande, por el Teorema del Límite Central, el estadístico

$Z = \dfrac{\bar{X} - \mu}{\sigma / \sqrt{n}}$, es una variable con distribución normal estándar, aproximadamente

Definición: Z_α

Z_α es el valor de la variable Z en la distribución normal estándar tal que el área a la derecha debajo de $f(z)$ es igual a un valor especificado α: $P(Z \geq z_\alpha) = \alpha$.



Ejemplo

Encuentre $z_{0.01}$

$P(Z \geq z_{0.01}) = 0.01 \Rightarrow P(Z \leq z_{0.01}) = 0.99 \Rightarrow F(z_{0.01}) = 0.99$

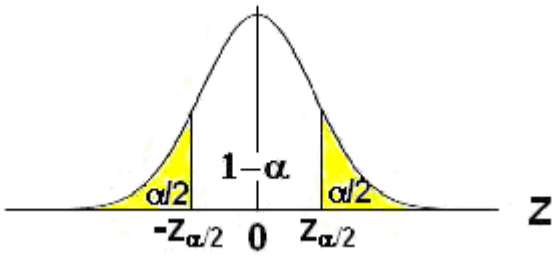
$\Rightarrow z_{0.01} = 2.33$ (Con la tabla de la distribución normal estándar)

ALGUNOS VALORES DE USO FRECUENTE PARA RECORDAR

$z_{0.1} = 1.28$
$z_{0.05} = 1.645$
$z_{0.025} = 1.96$
$z_{0.01} = 2.33$
$z_{0.005} = 2.575$

Fórmula para Estimación Puntual de la Media

Consideremos la distribución normal estándar separando el área en tres partes. La porción central con área o probabilidad **1 - α**, y dos porciones simétricas a los lados con área o probabilidad **α/2** cada una, siendo **α** un valor especificado



Por la definición de probabilidad, se puede escribir:

$$P(-z_{\alpha/2} \le Z \le z_{\alpha/2}) = 1 - \alpha$$

Es equivalente a decir que la desigualdad

$-z_{\alpha/2} \le Z \le z_{\alpha/2}$ se satisface con probabilidad **1 - α**

O equivalentemente:

$|Z| \le z_{\alpha/2}$ se satisface con probabilidad **1 - α**

Como se supone que la muestra es grande, por el Teorema del Límite Central

$Z = \dfrac{\bar{X} - \mu}{\sigma / \sqrt{n}}$, tiene distribución normal estándar aproximadamente

Sustituyendo en la desigualdad se obtiene:

$\left|\dfrac{\bar{X} - \mu}{\sigma / \sqrt{n}}\right| \le z_{\alpha/2}$ con probabilidad **1 - α**

De donde $|\bar{X} - \mu| \le z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$ con probabilidad **1 - α.**

$|\bar{X} - \mu|$ es el error en la estimación del parámetro μ mediante $\bar{X}$. Su máximo valor establece una cota para este error

Definición: Estimación puntual de la media, $n \ge 30$

$E = z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$ es el máximo error en la estimación con probabilidad **1 - α**

Es decir que si se estima μ mediante $\bar{X}$ con una muestra de tamaño **n≥30,** entonces se puede afirmar con una confianza de **1 - α** que el máximo error no excederá de $z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$

NOTA: Si se desconoce la varianza poblacional σ^2 se puede usar como aproximación la varianza muestral S^2, siempre que $n \ge 30$

Ejemplo

Se ha tomado una muestra aleatoria de **50** artículos producidos por una industria y se obtuvo que el peso de la media muestral fue **165** gr. con una desviación estándar de **40** gr. Encuentre el mayor error en la estimación de la media poblacional, con una confianza de **95%.**

Parámetro: μ

Estimador: $\bar{X}$

n≥30: muestra grande

$1 - \alpha = 0.95 \Rightarrow \alpha/2 = 0.025 \Rightarrow Z_{0.025} = 1.96$

$\sigma^2 \cong S^2 \Rightarrow \sigma \approx S = 40$

$E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 1.96\left(\frac{40}{\sqrt{50}}\right) = 11.08$ gr.

Conclusión

Se puede afirmar con una confianza de **95%** que al usar la media muestral para estimar a la media poblacional el error no excederá en más de **11.08** gr.

.

10.4.2 TAMAÑO DE LA MUESTRA

La fórmula anterior también se puede usar para estimar el tamaño de la muestra para que el error en la estimación no exceda a cierto valor con una probabilidad especificada

Definición: Tamaño de la muestra, $n \geq 30$

Tamaño de la muestra para que con probabilidad $1 - \alpha$ el máximo error en la estimación no exceda al valor especificado **E**

$$n = \left[Z_{\alpha/2}\frac{\sigma}{E}\right]^2$$

Se obtiene directamente de la fórmula anterior:

Ejemplo

Se conoce que la varianza de una población es 20. Determine cual debe ser el tamaño de la muestra para que el error máximo en la estimación de la media poblacional mediante la media muestral no exceda de 1 con una probabilidad de 99%

Solución

$1 - \alpha = 0.99 \Rightarrow Z_{\alpha/2} = Z_{0.005} = 2.575$

$\sigma = \sqrt{20} = 4.4721$

$E = 1$

$$n = \left[Z_{\alpha/2}\frac{\sigma}{E}\right]^2 = \left[2.575\frac{4.4721}{1}\right]^2 = 132.6 \Rightarrow n \cong 133$$

Conclusión

Debe usarse una muestra de tamaño 133

10.4.3 ESTIMACIÓN POR INTERVALO

Caso: Muestras grandes (n ≥ 30)

Parámetro: μ (Es la medida poblacional cuyo valor se desea estimar)
Población con distribución desconocida, varianza σ^2
Estimador: $\bar{X}$ (Media muestral)

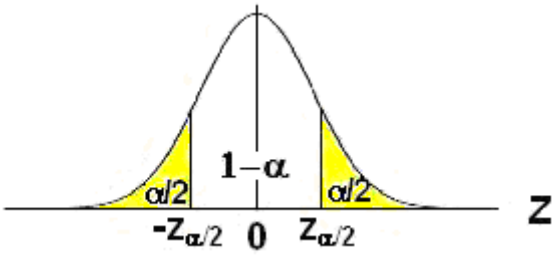
$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i\,, \quad \text{Media: } \mu_{\bar{X}} = \mu \quad \text{Varianza: } \sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$$

Siendo la muestra grande, por el Teorema del Límite Central, el estadístico

$Z = \dfrac{\bar{X} - \mu}{\sigma / \sqrt{n}}$, es una variable con distribución normal estándar aproximadamente

Fórmula para Estimación por Intervalo para la Media

Consideremos la distribución normal estándar separando el área en tres partes. La porción central con área o probabilidad **1 - α**, y dos porciones simétricas a los lados con área o probabilidad $\alpha/2$ cada una, siendo α un valor especificado:



Por la definición de probabilidad, se puede escribir:

$$P(-z_{\alpha/2} \le Z \le z_{\alpha/2}) = 1 - \alpha$$

Es equivalente a decir que la desigualdad

$-z_{\alpha/2} \le Z \le z_{\alpha/2}$ se satisface con probabilidad **1 - α**

Como se supone que la muestra es grande, por el Teorema del Límite Central

$Z = \dfrac{\bar{X} - \mu}{\sigma / \sqrt{n}}$, tiene distribución normal estándar aproximadamente

Sustituyendo se obtiene:

$-z_{\alpha/2} \le \dfrac{\bar{X} - \mu}{\sigma / \sqrt{n}} \le z_{\alpha/2}$ con probabilidad **1 - α**

De donde al despejar el parámetro de interés μ se tiene,

$\bar{X} - z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$, con probabilidad **1 - α**

Definición: Estimación por Intervalo para la Media con Muestras Grandes

Intervalo de confianza para μ con nivel **1 - α,** con una muestra de tamaño **$n \ge 30$,**

$$\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Los valores extremos se denominan **límites de confianza (inferior y superior)**

Ejemplo
Se ha tomado una muestra aleatoria de 50 artículos producidos por una industria y se obtuvo que la media muestral del peso de los artículos fue 165 gr. con una desviación estándar de 40 gr. Encuentre un intervalo para la media poblacional, con un nivel de confianza de 98%.

Parámetro: μ (población con distribución desconocida)
Estimador: $\bar{X}$

$n \geq 30$: muestra grande

$1 - \alpha = 0.98 \Rightarrow \alpha/2 = 0.01 \Rightarrow Z_{0.01} = 2.33$
$\sigma^2 \cong S^2 \Rightarrow \sigma \cong S = 40$

$$\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Sustituimos los datos

$$165 - 2.33\frac{40}{\sqrt{50}} \leq \mu \leq 165 + 2.33\frac{40}{\sqrt{50}}$$

$$151.8 \leq \mu \leq 178.1$$

Conclusión
Se puede afirmar con una confianza de 98% que la media poblacional se encuentra entre 151.8 y 178.1 gr.

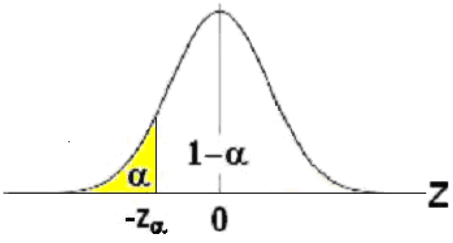
10.4.4 INTERVALOS DE CONFIANZA UNILATERALES

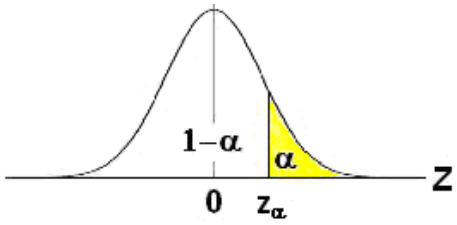
Caso: Muestras grandes (n ≥ 30)

Parámetro: μ (Es la medida poblacional cuyo valor se desea estimar)
Población con distribución desconocida, varianza σ^2
Estimador: $\bar{X}$ (Media muestral, se usa para estimar al parámetro)

Fórmula para Estimación por Intervalos Unilaterales

Con referencia a la distribución normal estándar:





En forma similar al caso considerado para el intervalo de confianza bilateral, se pueden obtener fórmulas para intervalos de confianza unilaterales que, con una probabilidad especificada, contengan a la media poblacional.

Definición: Estimación por intervalo para la media

Intervalo de confianza para μ con nivel $1 - \alpha$, con una muestra de tamaño $n \geq 30$,

$\mu \leq \bar{X} + z_{\alpha}\frac{\sigma}{\sqrt{n}}$ Intervalo de confianza unilateral inferior

$\mu \geq \bar{X} - z_{\alpha}\frac{\sigma}{\sqrt{n}}$ Intervalo de confianza unilateral superior

10.4.5 EJERCICIOS

1) Calcule $Z_{0.025}$

2) La media de la presión sanguínea de 40 mujeres de edad avanzada es 140. Si estos datos se pueden considerar como una muestra aleatoria de una población cuya desviación estándar es 10, encuentre, con una confianza de 95%, el mayor error en la estimación de la media poblacional.

3) De una población con distribución desconocida se tomó una muestra aleatoria de tamaño 40 y se obtuvo una media de 65.2 y una desviación estándar de 16. Construya un intervalo de confianza de 90% para la media poblacional.

4) Un fabricante de pinturas desea determinar el tiempo promedio de secado de una nueva pintura. En 36 pruebas realizadas obtuvo un tiempo de secado medio de 64.2 minutos con una desviación estándar de 8.5 minutos. Construya un intervalo de confianza unilateral inferior de 95% para la media del tiempo de secado de la nueva pintura.

MATLAB

Obtención de intervalos de confianza para la media, $n \geq 30$

Se pueden calcular intervalos de confianza usando la función inversa de la distribución normal

```
>> p = [0.01, 0.99];
>> x = norminv(p, 165, 40/sqrt(50))
   x =
      151.8402  178.1598
```

Intervalo de confianza bilateral
$1 - \alpha = 98\%$, $\bar{X} = 165$, $S = 40$, $n = 50$

```
>> p = [0, 0.98];
>> x = norminv(p, 165, 40/sqrt(50))
   x =
      -Inf  176.6178
```

Intervalo de confianza unilateral inferior
$1 - \alpha = 98\%$, $\bar{X} = 165$, $S = 40$, $n = 50$

```
>> p = [0.02, 1];
>> x = norminv(p, 165, 40/sqrt(50))
   x =
      153.3822       Inf
```

Intervalo de confianza unilateral superior
$1 - \alpha = 98\%$, $\bar{X} = 165$, $S = 40$, $n = 50$

10.4.6 ESTIMACIÓN PUNTUAL DE LA MEDIA

Caso: Muestras pequeñas (n<30)

Parámetro: μ (Es la medida poblacional cuyo valor se desea estimar)
Población con distribución **normal**, varianza σ^2 **desconocida**
Estimador: $\bar{X}$ (Media muestral, se usa para estimar al parámetro)

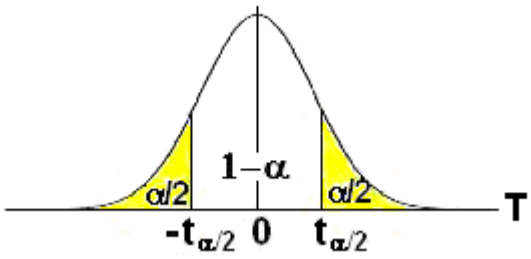
Para realizar inferencias se usa una variable aleatoria con distribución **T**

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}}, \text{ con } \nu = n - 1 \text{ grados de libertad}$$

NOTA: Si la población tuviese distribución **normal** y la varianza poblacional σ^2 fuese **conocida**, la variable aleatoria para realizar inferencias tendría distribución normal estándar **Z,** sin importar el tamaño de la muestra.

Fórmula para Estimación Puntual de la Media

Consideremos la distribución **T** separando el área en tres partes. La porción central con área o probabilidad **1 - α**, y dos porciones simétricas a los lados con área o probabilidad $\alpha/2$ cada una, siendo α un valor especificado



Por la definición de probabilidad, se puede escribir:

$$P(-t_{\alpha/2} \le T \le t_{\alpha/2}) = 1 - \alpha$$

Es equivalente a decir que la desigualdad

$-t_{\alpha/2} \le T \le t_{\alpha/2}$ se satisface con probabilidad **1 - α**

O equivalentemente:

$|T| \le t_{\alpha/2}$ se satisface con probabilidad **1 - α**

Como se supone que la muestra es grande, por el **teorema del límite central**

$T = \frac{\bar{X} - \mu}{s/\sqrt{n}}$, tiene distribución normal estándar aproximadamente

Sustituyendo en la desigualdad se obtiene:

$\left|\frac{\bar{X} - \mu}{s/\sqrt{n}}\right| \le t_{\alpha/2}$ con probabilidad **1 - α**

De donde $|\bar{X} - \mu| \le t_{\alpha/2}\frac{s}{\sqrt{n}}$ con probabilidad **1 - α.**

$|\bar{X} - \mu|$ es el error en la estimación del parámetro μ mediante $\bar{X}$

Definición: Estimación puntual de la media, n < 30

$$E = t_{\alpha/2}\frac{s}{\sqrt{n}}$$ es el máximo error en la estimación con probabilidad **1 - α**

Es decir que si se estima μ mediante $\bar{X}$ con una muestra de tamaño $n < 30$, entonces se puede afirmar con una confianza de $1 - \alpha$ que el máximo error no excederá a $t_{\alpha/2}\frac{s}{\sqrt{n}}$

Ejemplo

Se ha tomado una muestra aleatoria de 20 artículos producidos por una industria y se obtuvo que el peso de la media muestral fue 165 gr. con una desviación estándar de 40 gr. Encuentre el mayor error en la estimación de la media poblacional, con una confianza de 95%. Suponga que la población tiene distribución **normal.**

Solución

Parámetro: μ, población **normal**, varianza **desconocida**

Estimador: $\bar{X}$

n <30: muestra pequeña

$1 - \alpha = 0.95 \Rightarrow \alpha/2 = 0.025 \Rightarrow t_{0.025} = 2.093$, con la tabla T

$\nu = 20 - 1 = 19$ grados de libertad

$E = t_{\alpha/2}\frac{s}{\sqrt{n}} = 2.093\left(\frac{40}{\sqrt{20}}\right) = 18.72$ gr.

Conclusión

Se puede afirmar con una confianza de 95% que al usar la media muestral para estimar a la media poblacional, el error no excederá a 18.72 gr.

10.4.7 ESTIMACIÓN POR INTERVALO PARA LA MEDIA

Caso n < 30 (Muestras pequeñas)

Parámetro: μ (Es la medida poblacional cuyo valor se desea estimar)
Población con distribución **normal**, varianza σ^2 **desconocida**
Estimador: $\bar{X}$ (Media muestral, se usa para estimar al parámetro)

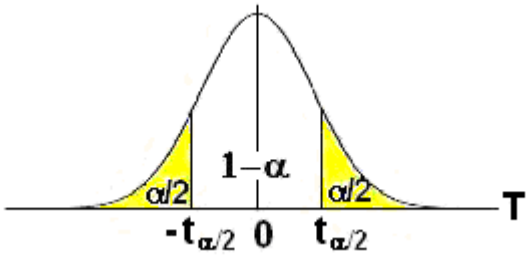
Para realizar inferencias se usa una variable aleatoria con distribución **T**

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}},$$ con $\nu = n-1$ grados de libertad

NOTA: Si la población tuviese distribución **normal** y la varianza poblacional σ^2 fuese **conocida**, la variable aleatoria para realizar inferencias tendría distribución normal estándar **Z,** sin importar el tamaño de la muestra.

Fórmula para Estimación por Intervalo para la Media

Consideremos la distribución **T** separando el área en tres partes. La porción central con área o probabilidad $1 - \alpha$, y dos porciones simétricas a los lados con área o probabilidad $\alpha/2$ cada una, siendo α un valor especificado



Por la definición de probabilidad, se puede escribir:

$$P(-t_{\alpha/2} \le T \le t_{\alpha/2}) = 1 - \alpha$$

Es equivalente a decir que la desigualdad

$-t_{\alpha/2} \le T \le t_{\alpha/2}$ se satisface con probabilidad $1 - \alpha$

Sustituyendo: $T = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$ en la desigualdad

Se obtiene:

$$-t_{\alpha/2} \le \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \le t_{\alpha/2}$$ con probabilidad $1 - \alpha$

De donde al despejar el parámetro de interés μ se tiene,

$$\bar{X} - t_{\alpha/2}\frac{s}{\sqrt{n}} \le \mu \le \bar{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}$$, con probabilidad $1 - \alpha$

Definición: Estimación por Intervalo para la Media con Muestras Pequeñas

Intervalo de confianza para μ con nivel $1 - \alpha$, con $n < 30$, (muestras pequeñas) población **normal** y varianza **desconocida**,

$$\bar{X} - t_{\alpha/2}\frac{s}{\sqrt{n}} \le \mu \le \bar{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}$$

Los valores extremos son los límites de confianza

Ejemplo
De una población con distribución normal se tomó una muestra aleatoria de **4** observaciones obteniéndose: **9.4, 12.2, 10.7, 11.6**. Encuentre un intervalo para la media poblacional, con un nivel de confianza de **90%**

Parámetro: μ, población normal, varianza desconocida

Estimador: $\bar{X}$

n<30: muestra pequeña

Calculamos la media y varianza muestrales:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{4}\sum_{i=1}^{4} x_i = \frac{1}{4}(9.4 + 12.2 + 10.7 + 11.6) = 10.975$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{3}[(9.4 - 10.975)^2 + (12.2 - 10.975)^2 + \ \ldots\ ] = 1.4825$$

$$S = \sqrt{S^2} = \sqrt{1.4825} = 1.2176$$

$1 - \alpha = 0.90 \Rightarrow \alpha/2 = 0.05 \Rightarrow t_{\alpha/2} = t_{0.05} = 2.353,$ (Tabla T)

$\nu = 4 - 1 = 3$ grados de libertad

Sustituímos los valores en la desigualdad

$$\bar{X} - t_{\alpha/2}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}$$

Se obtiene

$$10.975 - 2.353\frac{1.2176}{\sqrt{4}} \leq \mu \leq 10.975 + 2.353\frac{1.2176}{\sqrt{4}}$$

$$9.5425 \leq \mu \leq 12.4075$$

Conclusión
Se puede afirmar con una confianza de 90% que la media poblacional se encuentra entre 9.5425 y 12.4075

10.4.8 EJERCICIOS

1) Un inspector de alimentos examina una muestra aleatoria de 10 artículos producidos por una fábrica y obtuvo los siguientes porcentajes de impurezas: 2.3, 1.9, 2.1, 2.8, 2.3, 3.6, 1.8, 3.2, 2.0, 2.1. Suponiendo que la población tiene distribución normal, encuentre el mayor error en la estimación de la media poblacional, con una confianza de 95%.

2) De una población con distribución normal y varianza 225 se tomó una muestra aleatoria de tamaño 20 y se obtuvo una media de 64.5. Construya un intervalo de confianza de 95% para la media poblacional.

3) Un fabricante de pinturas desea determinar el tiempo promedio de secado de una nueva pintura. En diez pruebas realizadas obtuvo un tiempo de secado medio de 65.2 minutos con una desviación estándar de 9.4 minutos. Construya un intervalo de confianza de 95% para la media del tiempo de secado de la nueva pintura. Suponga que la población es normal.

4) El peso de seis artículos de una muestra aleatoria tomada de la producción de una fábrica fueron: 0.51, 0.59, 0.52, 0.47, 0.53, 0.49 kg. Encuentre un intervalo de confianza de 98% para la media del peso de todos los artículos producidos. Suponga distribución normal.

MATLAB

Obtención de intervalos de confianza para la media, n < 30

```
>> u = [9.4 12.2 10.7 11.6];
>> m = mean(u)
    m =
        10.9750
>> s = std(u)
    s =
        1.2176
>> ta = tinv(0.95,3)
   ta =
        2.3534
>> x =[m - ta*s/sqrt(4), m+ta*s/sqrt(4)]
    x =
        9.5423   12.4077
```

Vector conteniendo una muestra de cuatro datos

Media muestral

Desviación estándar muestral

Valor del estadístico t para $\alpha = 0.05$, $\nu = 3$

Intervalo de confianza bilateral para μ

10.5 PRUEBA DE HIPÓTESIS

Esta técnica estadística es muy utilizada como soporte a la investigación sistemática y científica. Consiste en suponer algún valor para el parámetro de interés y usar los datos de la muestra para aceptar o rechazar esta afirmación.

Es importante entender las diferentes situaciones que pueden ocurrir al probar estadísticamente una hipótesis.

Sea **Ho**: alguna hipótesis que se propone para el parámetro de interés

Suponer que se dispone de datos y que se realiza una prueba estadística para verificar la hipótesis. Entonces pueden ocurrir las siguientes situaciones al tomar una decisión:

Posibles decisiones que pueden tomarse

	Aceptar Ho	Rechazar Ho
Ho es verdadera	Decisión correcta	Error tipo I
Ho es falsa	Error tipo II	Decisión correcta

Suponer que la hipótesis propuesta **Ho es verdadera**, pero la prueba estadística dice que **Ho es falsa,** entonces al **rechazar** la hipótesis propuesta cometemos el **Error Tipo I**

Suponer que la hipótesis propuesta **Ho es falsa**, pero la prueba estadística dice que **Ho es verdadera**, entonces al **aceptar** la hipótesis propuesta cometemos el **Error Tipo II.**

Ambos errores pueden tener consecuencias importantes al tomar una decisión en una situación real. Por lo tanto es necesario cuantificar la probabilidad de cometer cada tipo de error.

Definiciones: Medición de los errores Tipo I y Tipo II

Error Tipo I:

α = **P**(Rechazar **Ho** dado que **Ho** es verdadera)

Error Tipo II:

β = **P**(Aceptar **Ho** dado que otra hipótesis es verdadera)

El valor α se denomina **nivel de significancia** de la prueba y **puede darse como un dato** para realizar la prueba.

Algunos valores típicos para α son: **10%, 5%, 2%, 1%**

Terminología

Ho: Hipótesis nula.	Es la hipótesis propuesta para el parámetro de interés.
Ha: Hipótesis alterna.	Es la hipótesis que se plantea en oposición a **Ho** y que es aceptada en caso de que **Ho** sea rechazada

Generalmente es de interés probar **Ha**, por lo que se plantea **Ho** con la esperanza de que sea rechazada mediante la información contenida en la muestra.

Ejemplo
Suponer que se desea probar que la media poblacional **no es igual a 5**

Entonces se pueden plantear:

Ho: $\mu = 5$ (Hipótesis nula)
Ha: $\mu \neq 5$ (Hipótesis alterna)

Si con los datos de la muestra se puede rechazar **Ho**, entonces habremos probado **Ha**, caso contrario, se dice que no hay evidencia suficiente para rechazar **Ho**

TIPOS DE PRUEBAS

Sean θ: Parámetro de interés
θ_0: Algún valor que se propone para el parámetro

Pruebas de una cola

1) **Ho:** $\theta = \theta_0$: (Hipótesis nula)
Ha: $\theta < \theta_0$: (Hipótesis alterna)

2) **Ho:** $\theta = \theta_0$: (Hipótesis nula)
Ha: $\theta > \theta_0$: (Hipótesis alterna)

Prueba de dos colas

3) **Ho:** $\theta = \theta_0$: (Hipótesis nula)
Ha: $\theta < \theta_0 \vee \theta > \theta_0$: (Hipótesis alterna)

PROCEDIMIENTO PARA REALIZAR UNA PRUEBA DE HIPÓTESIS

Para conocer el procedimiento para la Prueba de Hipótesis, se describe la prueba relacionada con la media, sin embargo la técnica es aplicable para pruebas con otros parámetros

10.5.1 Prueba de Hipótesis relacionada con la media

Caso n ≥ 30 (Muestra grande)

Parámetro: μ (media poblacional)
Población con varianza σ^2, distribución desconocida,
Estimador: $\bar{X}$ (media muestral)
Valor propuesto para el parámetro: μ_0

Procedimiento

Paso 1. Formular la hipótesis nula: **Ho:** $\mu = \mu_0$

Paso 2. Formular una hipótesis alterna, la cual es de interés probar. Elegir una entre:

1) **Ha:** $\mu > \mu_0$
2) **Ha:** $\mu < \mu_0$
3) **Ha:** $\mu < \mu_0 \vee \mu > \mu_0$

Paso 3. Especificar el nivel de significancia de la prueba: α

Paso 4. Seleccionar el estadístico de prueba y definir la región de rechazo de **Ho**

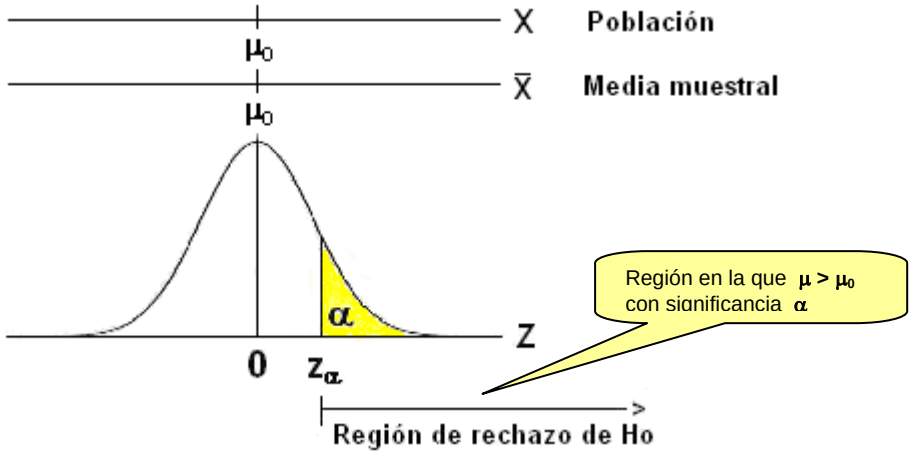
En este caso, por el Teorema del Límite Central, el estadístico

$Z = \dfrac{\bar{X} - \mu_o}{\sigma / \sqrt{n}}$, tiene Distribución Normal Estándar aproximadamente

La región de rechazo **depende de la hipótesis alterna elegida Ha** y está determinada por el valor de α especificado**.** Se analiza la primera situación: **1) Ha:** $\mu > \mu_0$

Ho: $\mu = \mu_0$

Ha: $\mu > \mu_0$



Con el valor especificado α se obtiene un valor para Z_α el cual delimita la región de rechazo.

La media muestral $\bar{X}$ es un estimador insesgado del parámetro μ, por lo tanto su valor esperado coincide con el valor propuesto μ_0 para el parámetro**.**

Según lo anterior, el valor obtenido para la media muestral $\bar{X}$ debería estar cerca de μ_0, y por lo tanto, el valor de $Z = \dfrac{\bar{X} - \mu_o}{\sigma / \sqrt{n}}$ deberá estar cercano a **0**.

Pero si el valor obtenido para la media muestral $\bar{X}$ es significativamente mas grande que μ_0, entonces **Z** caerá en la región de rechazo definida. Esto debe entenderse como una evidencia de que la media μ_0 propuesta para el parámetro μ no es verdad y que debería ser algún valor más grande, es decir se puede concluir que: $\mu > \mu_0$

Con esta interpretación rechazamos Ho en favor de Ha con un nivel de significancia α

Sin embargo, siendo $\bar{X}$ una variable aleatoria, existe la probabilidad de que pueda tomar cualquier valor. Por lo tanto, es posible que se obtenga un valor para $\bar{X}$ que caiga en la región de rechazo aún siendo verdad que su valor esperado sea μ_0**.**

La posibilidad de que se produzca esta situación constituye el **Error Tipo I,** y la probabilidad que esto ocurra es también α

Paso 5. Calcular el valor del estadístico de prueba con los datos de la muestra

Paso 6. Tomar una decisión

Si el valor del estadístico de prueba cae en la región de rechazo, la decisión es rechazar **Ho** en favor de **Ha**. Pero, si el valor no cae en esta región crítica, se dice que no hay evidencia suficiente para rechazar **Ho**. En este caso es preferible abstenerse de aceptar como verdadera **Ho** pues esto puede introducir el **Error tipo II**

Ejemplo

Una muestra aleatoria de 100 paquetes mostró un peso promedio de 71.8 gr. con una desviación estándar de 8.9 gr.

Pruebe, con un nivel de significancia de 5%, que el peso promedio de todos los paquetes (población) es mayor a 70 gr.

Seguimos los pasos indicados en el procedimiento básico indicado anteriormente:

1. Hipótesis nula

Ho: $\mu = 70$

2. Hipótesis alterna

Ha: $\mu > 70$

3. Nivel de significancia

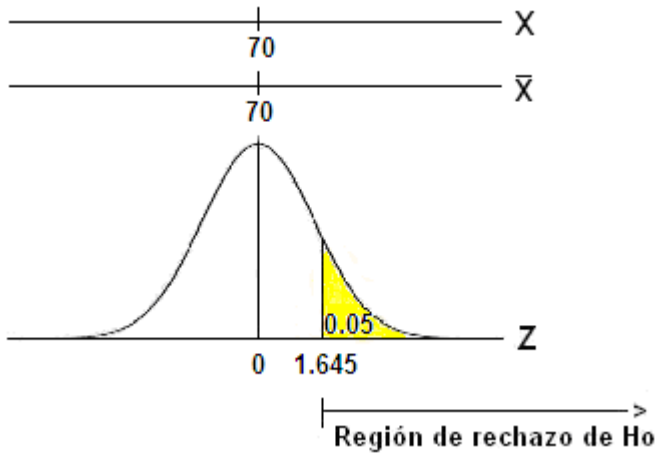
$\alpha = 0.05$

4. Estadístico de prueba

$Z = \dfrac{\bar{X} - \mu_o}{\sigma / \sqrt{n}}$ por el Teorema del Límite Central. Además $\sigma^2 \cong S^2$

Región de rechazo

$Z_\alpha = Z_{0.05} = 1.645 \Rightarrow$ Rechazar **Ho** en favor de **Ha,** si **z > 1.645**



5. Valor del estadístico

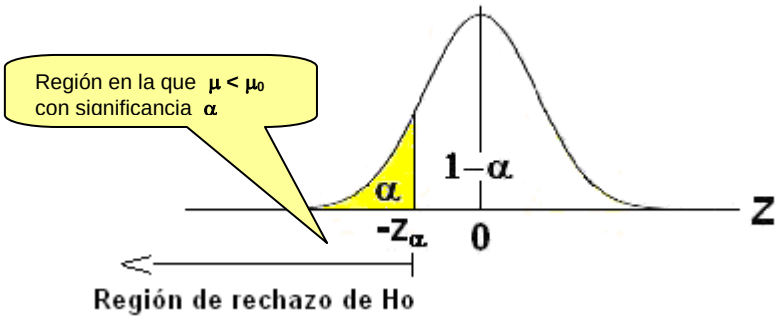
$Z = \dfrac{\bar{X} - \mu_o}{\sigma / \sqrt{n}} = \dfrac{71.8 - 70}{8.9 / \sqrt{100}} = 2.02 \Rightarrow 2.02$ cae en la región de rechazo

6. Decisión

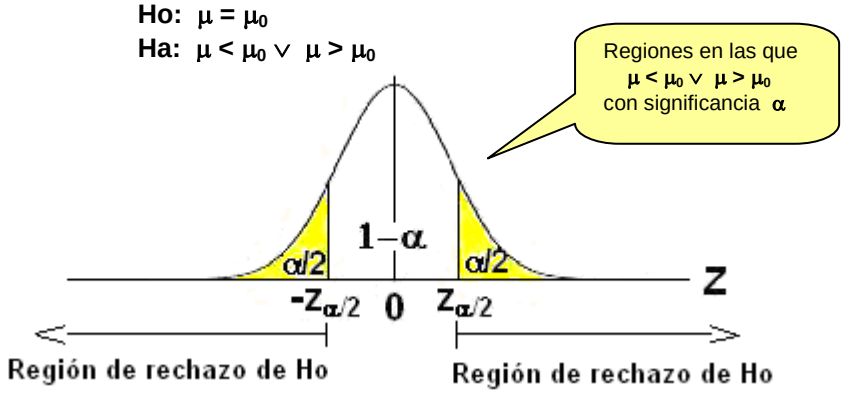
Se rechaza que la media poblacional es 70 y se concluye, con una significancia de **5%** que el peso promedio de la población es mayor a 70 gr,

El análisis anterior permite interpretar las otras dos situaciones para la hipótesis alterna:

2) **Ho:** $\mu = \mu_0$
Ha: $\mu < \mu_0$



3) **Ho:** $\mu = \mu_0$
Ha: $\mu < \mu_0 \vee \mu > \mu_0$



10.5.2 EJERCICIOS

1) Una muestra aleatoria de **n=40** observaciones tomada de una población en estudio, produjo una media $\bar{X}$ **=2.4** y una desviación estándar **S=0.28**. Suponga que se desea demostrar que la media poblacional μ es mayor a **2.3**

a) Enuncie la hipótesis nula para la prueba
b) Enuncie la hipótesis alterna para la prueba
c) Use su intuición para predecir si el valor de la media muestral $\bar{X}$ **= 2.4** es suficiente evidencia para afirmar que la media poblacional μ es mayor que el valor propuesto **2.3**
d) Realice la prueba de hipótesis con un nivel de significancia de α**=0.05** y determine si los datos son evidencia suficiente para rechazar la hipótesis nula en favor de la hipótesis alterna.

2) Repita el ejercicio **1)** con los mismos datos, pero suponiendo que se desea demostrar que la media poblacional es menor que **2.7**

3) Repita el ejercicio **1)** con los mismos datos, pero suponiendo que se desea demostrar que la media poblacional es diferente que **2.7**

MATLAB

Prueba de hipótesis relacionada con la media, $n \geq 30$

Vector con los datos de una muestra

```
>> x = [71.76  69.34  83.16  88.38  67.15  72.72  64.61  77.86  50.76  80.61  73.75  74.13 ...
        82.60  69.36  70.62  60.49  56.99  65.54  74.30  66.98  59.93  81.35  65.46  71.70 ...
        71.79  69.58  75.33  69.45  56.99  62.64  73.96  60.62  68.71  63.42  61.35  62.71 ...
        68.23  73.35  70.77  81.27];
```

```
>> m=mean(x)
   m =
       69.7430
```
media muestral

```
>> s=std(x)
   s =
       8.0490
```
desviación estándar muestral

```
>> [h,p,ci,z]=ztest(x, 67, 8.049, 0.05, 1)
```
Prueba **Ho: $\mu = 67$** vs. **Ha: $\mu > 67$,**
$\sigma \cong S = 8.049$, α = **0.05**. Prueba unilateral derecha

Prueba unilateral derecha

```
   h = 1
   p =  0.0156
   ci = 67.6497      Inf
   z =  2.1553
```

h=1 $\Rightarrow$ La evidencia es suficiente para rechazar **Ho**
Valor **p** de la prueba
Intervalo de confianza con nivel **$1 - \alpha$**
Valor del estadístico de prueba **Z**

10.5.3 PRUEBA DE HIPÓTESIS RELACIONADA CON LA MEDIA

Caso n<30 (Muestras pequeñas)

Parámetro: μ (Es la medida poblacional cuyo valor se desea estimar)

Población con distribución **normal**, varianza σ^2 **desconocida**

Estimador **T** (Variable aleatoria con distribución **T**, con $\nu = n-1$)

Valor propuesto para el parámetro: μ_0

Para realizar inferencias se usa una variable aleatoria con distribución **T**

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}}, \text{ con } \nu = n - 1 \text{ grados de libertad}$$

PROCEDIMIENTO BÁSICO

PASOS

1. Formular la hipótesis nula: **Ho:** $\mu = \mu_0$

2. Formular una hipótesis alterna. Elegir una entre:

Ha: $\mu < \mu_0$

Ha: $\mu > \mu_0$

Ha: $\mu \neq \mu_0$

3. Especificar el nivel de significancia de la prueba: α

4. Seleccionar el estadístico de prueba y definir la región de rechazo de **Ho**

$$t = \frac{\bar{X} - \mu_o}{s/\sqrt{n}}$$, tiene distribución **t** con $\nu = n-1$ grados de libertad

Ha	**Región de rechazo de Ho en favor de Ha**
$\mu < \mu_0$	$t < -t_\alpha$
$\mu > \mu_0$	$t > t_\alpha$
$\mu \neq\!< \mu_0$	$t < -t_{\alpha/2} \vee t > t_{\alpha/2}$

5. Con los datos de la muestra calcular el valor del estadístico

6. Si el valor del estadístico de prueba cae en la región de rechazo, la decisión es rechazar Ho en favor de Ha. Pero, si el valor no cae en esta región crítica, se dice que no hay evidencia suficiente para rechazar Ho. En este caso es preferible abstenerse de aceptar Ho como verdadera pues esto puede introducir el error tipo II

Ejemplo

De una población normal se tomó una muestra aleatoria y se obtuvieron los siguientes resultados: **15, 17, 23, 18, 20.** Probar con una significancia de 10% que la media de la población es mayor a **18**

Solución

1. **Ho:** $\mu = 18$

2. **Ha:** $\mu > 18$

3. Nivel de significancia de la prueba $\alpha = 0.10$

4. Estadístico de prueba

$T = \dfrac{\bar{X} - \mu_o}{s/\sqrt{n}}$, tiene distribución **T** con $\nu = n - 1$ grados de libertad

$\bar{x} = \dfrac{1}{5}(15+17+23+18+20)=18.6$

$S^2 = \dfrac{1}{4}((15-18.6)^2 + (17-18.6)^2 + \ldots) = 9.3 \Rightarrow S = 3.05$

Región de rechazo de Ho

$\alpha = 0.1$, $\nu = 5 - 1 = 4 \Rightarrow t_{0.1} = 1.53$ con la Tabla **T**

Rechazar **Ho** si $t > 1.53$

5. $t = \dfrac{18.6 - 18}{3.05/\sqrt{5}} = 0.44 \Rightarrow$ **0.44** no cae en la región de rechazo

6. Decisión

No hay evidencia suficiente para rechazar que la media poblacional es **18**.

10.5.4 EJERCICIOS

1) Una muestra aleatoria de **10** observaciones tomada de una población con distribución normal produjo una media **2.5** y una desviación estándar **0.28**. Suponga que se desea demostrar que la media poblacional es mayor a **2.3**

a) Enuncie la hipótesis nula para la prueba

b) Enuncie la hipótesis alterna para la prueba

c) Use su intuición para predecir si el valor de la media muestral es suficiente evidencia para afirmar que la media poblacional es mayor que el valor propuesto

d) Realice la prueba de hipótesis con un nivel de significancia de **5%** y determine si los datos son evidencia suficiente para rechazar la hipótesis nula en favor de la hipótesis alterna.

2) El peso de seis artículos de una muestra aleatoria tomada de la producción de una fábrica fueron: **0.51, 0.59, 0.52, 0.47, 0.53, 0.49** kg. Pruebe si estos datos constituyen una evidencia suficiente para afirmar que el peso promedio de todos los artículos producidos por la fábrica es mayor a **0.5** Kg. Encuentre el valor **p** o nivel de significancia de la prueba. Suponga distribución normal.

MATLAB

Prueba de hipótesis relacionada con la media, n < 30

```
>> x = [15 17 23 18 20];
>> [h, p, ci, t] = ttest(x, 18, 0.1, 1)

    h =
         0
    p =
         0.3414
    ci =
        16.5090       Inf
     t =
        tstat: 0.4399
           df: 4
```

Vector con los datos de la muestra

Prueba **Ho: μ =18** vs. **Ho: μ >18**
α = **0.1**. Prueba unilateral derecha

h=0 $\Rightarrow$ La evidencia no es suficiente para rechazar **Ho**

Valor **p** de la prueba

Intervalo de confianza con nivel **1 – α**

Valor del estadístico de prueba
Grados de libertad

10.5.5 VALOR P DE UNA PRUEBA DE HIPÓTESIS

El **Valor–p** de una prueba de hipótesis, o **Probabilidad de Cola**, es el valor de probabilidad correspondiente al área de la cola (o colas), a partir del valor observado y representa el nivel de significancia obtenido con la muestra.

Si esta probabilidad es pequeña, es un indicativo de que los datos de la muestra no apoyan a la hipótesis nula propuesta pues el valor del estadístico de prueba se ubica lejos del valor propuesto para el parámetro. Pero si esta probabilidad es grande, significa que los datos de la muestra favorecen a la hipótesis nula pues el valor del estadístico se ubica cerca del valor especificado para el parámetro

Ejemplo

Una muestra aleatoria de 100 paquetes mostró un peso promedio de 71.8 gr. con una desviación estándar de 8.9 gr. Pruebe que el peso promedio de todos los paquetes (población) es mayor a 70 gr. Exprese la respuesta mediante el Valor **p** de la prueba

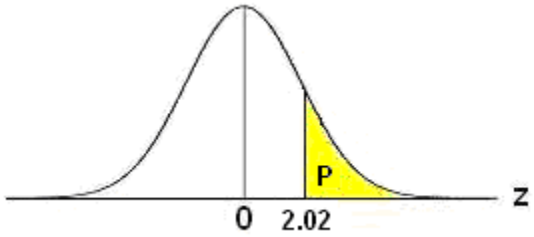
El nivel de significancia α no está especificado, por lo tanto se lo puede definir mediante los datos de la muestra

Hipótesis nula **Ho: $\mu = 70$**

Hipótesis alterna **Ha: $\mu > 70$**

Valor del estadístico de prueba

$$Z = \frac{\bar{X} - \mu_o}{\sigma / \sqrt{n}} \approx \frac{71.8 - 70}{8.9 / \sqrt{100}} = 2.02$$



Probabilidad de cola

$$P = P(Z \geq 2.02) = 1 - F(2.02) = 1 - 0.9783 = 0.0217 = 2.17\%$$

Se puede concluir que la prueba tiene una significancia de **2.17%**

Este valor de probabilidad se denomina **Valor p** de la prueba o **Probabilidad de Cola.**

10.5.6 CÁLCULO DEL ERROR TIPO I

El **Error Tipo I** tiene el mismo valor de probabilidad que el nivel de significancia α de la prueba y representa el error en que se incurrirá si la evidencia de la muestra nos hace rechazar **Ho**, sin conocer que **Ho** es verdadera.

Suponga que se define la siguiente hipótesis relacionada con la media, con una muestra grande.

Ho:	$\mu = \mu_0$	(Hipótesis nula)
Ha:	$\mu > \mu_0$	(Hipótesis alterna)
α**:**		(Nivel de significancia o **Error Tipo I**)
$Z > z_\alpha$		(Región de rechazo)

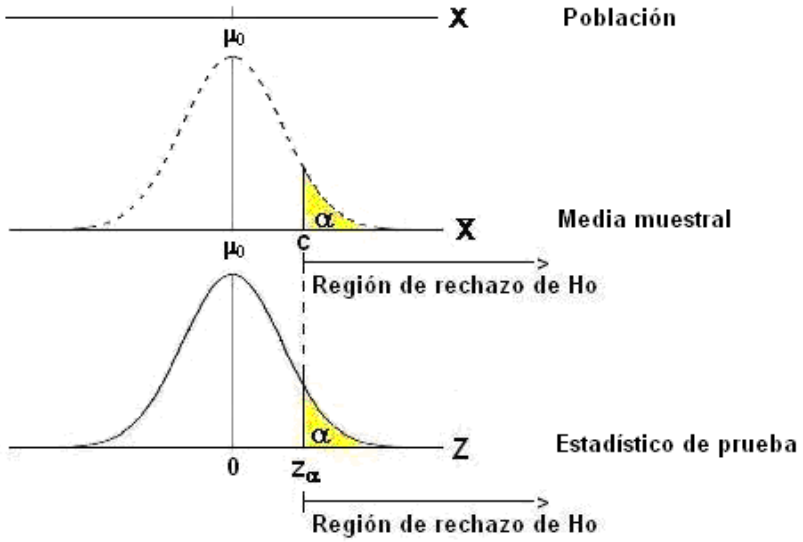
La región de rechazo está definida con el valor crítico z_α que se obtiene del valor especificado α.

La región de rechazo también puede definirse proponiendo un valor crítico **c** para $\bar{X}$, entonces el **Error Tipo I** de la prueba es

Definición: Error Tipo I con Ha: $\mu > \mu_0$ siendo $\mu = \mu_0$

$$\alpha = P(\text{Rechazar } Ho \mid Ho \text{ es verdadera}) = P(\bar{X} > c) = P\left(Z > \frac{c - \mu_0}{\sigma / \sqrt{n}}\right)$$

Los valores z_α y **c** están relacionados directamente: $z_\alpha = \dfrac{c - \mu_0}{\sigma / \sqrt{n}} \Rightarrow c = \mu_0 + z_\alpha (\sigma / \sqrt{n})$



Para facilitar la comprensión del concepto se ha graficado también $\bar{X}$ con distribución normal

Ejemplo. X es una variable aleatoria con distribución normal y varianza **49**. Se plantea el siguiente contraste de hipótesis H_o**: μ = 15** vs H_a**: $\mu > 15$** y se ha especificado como **región de rechazo de H_o** que la media $\bar{X}$ de todas las muestras con **n = 40** tengan un valor mayor a **17**

Encuentre la medida del E**rror Tipo I**

Error Tipo I: $\alpha = P(\bar{X} > c) = P(\bar{X} > 17) = P\left(Z > \dfrac{c - \mu_0}{\sigma / \sqrt{n}}\right) = P\left(Z > \dfrac{17 - 15}{7 / \sqrt{40}}\right) = P(Z > 1.807) \cong 0.04$

10.5.7 CÁLCULO DEL ERROR TIPO II

El E**rror Tipo II** se representa por β, y se usa para cuantificar el error en que se incurrirá al aceptar **Ho:** $\mu=\mu_0$ cuando la evidencia de la muestra no es suficiente para rechazarla, sin saber que el verdadero valor de la media μ es algún otro valor μ_1. Para entender el concepto usamos un caso particular

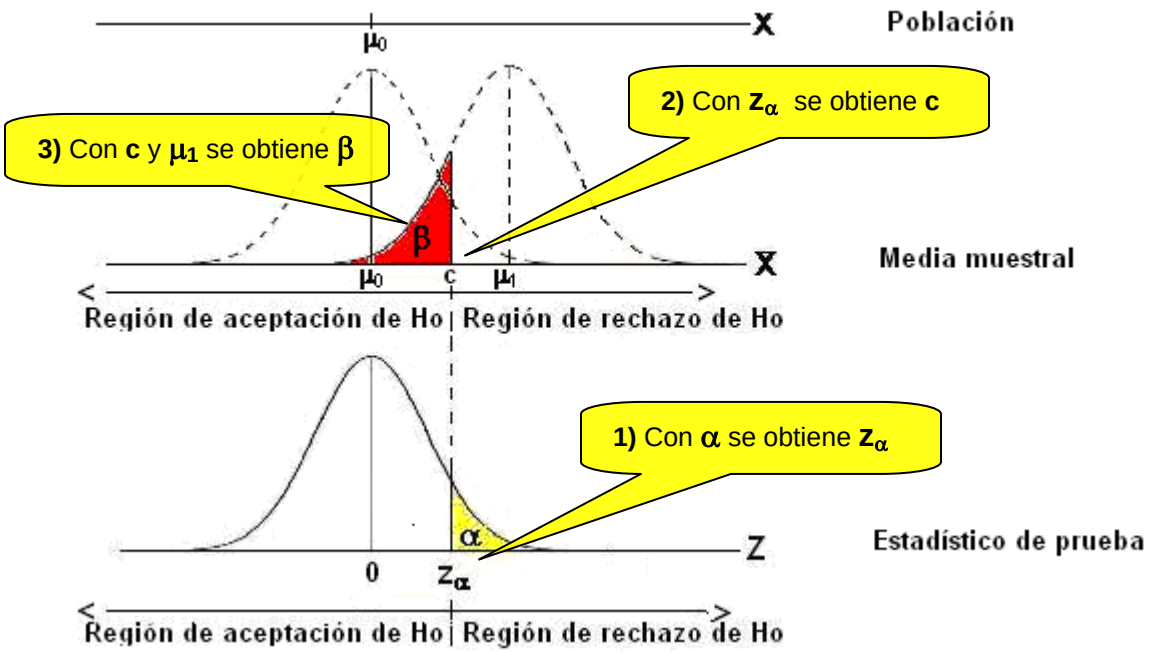
Caso

Ho:	$\mu = \mu_0$	(Hipótesis nula)
Ha:	$\mu > \mu_0$	(Hipótesis alterna)
α**:**		(Nvel de significancia)

Para calcular el valor de β debemos suponer que hay otro valor verdadero para el parámetro μ. Sea μ_1 el valor que suponemos verdadero. Entonces β es la probabilidad calculada con este valor μ_1 (área a la izquierda del valor crítico **c**).

Definición: Error Tipo II con Ha: $\mu > \mu_0$ siendo $\mu = \mu_1$

$$\beta = P(\bar{X} < c)_{\mu=\mu_1} = P\left(Z < \frac{c - \mu_1}{\sigma / \sqrt{n}}\right)$$



Ejemplo.-
Suponga que se define la siguiente hipótesis relacionada con la media.
Muestra: **n = 100, $\bar{X}$ =71.8, S = 8.9**

Ho:	**μ = 70**	(Hipótesis Nula)
Ha:	**μ > 70**	(Hipótesis alterna)
α:	**5%**	(Nivel de significancia)

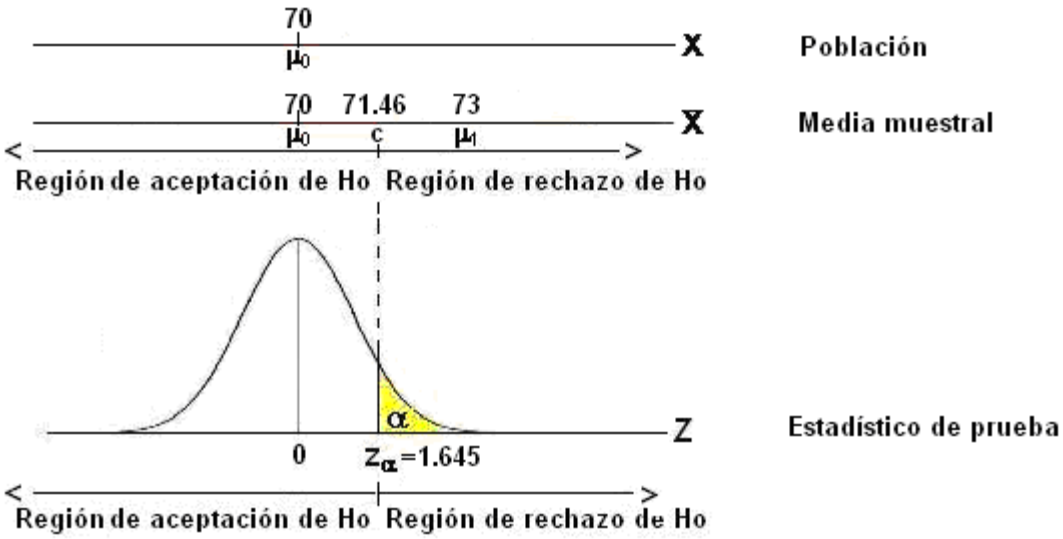
Calcule la magnitud del E**rror Tipo II** suponiendo que la media poblacional verdadera es **μ = 73**

Solución
Región de rechazo de H_0

$$\alpha = 0.05 \Rightarrow z_\alpha = z_{0.05} = 1.645 \Rightarrow z > 1.645$$

Calculemos el valor crítico **c** de $\bar{X}$ para la región de rechazo:

$$z_\alpha = \frac{c - \mu_0}{\sigma / \sqrt{n}} \Rightarrow c = \mu_0 + z_\alpha (\sigma / \sqrt{n}) = 70 + 1.645 (8.9 / \sqrt{100}) = 71.46$$



β = **P(**Aceptar **Ho** dado que la hipótesis verdadera es: $\mu = \mu_1$ **)**

β = **P(** $\bar{X} < c$ **)** con $\mu = \mu_1$

= **P(Z <** $\frac{c - \mu_1}{s / \sqrt{n}}$ **)** = **P(Z <** $\frac{71.46 - 73}{8.9 / \sqrt{100}}$ **)** = **P(Z < –1.73)** = **4.18%** (**Error Tipo II** con μ = **73)**

Se concluye que la probabilidad de aceptar μ = **70** siendo falsa es **4.18%** si μ = **73** es verdadera

10.5.8 CURVA CARACTERÍSTICA DE OPERACIÓN

Si se grafican los puntos de β para algunos valores de μ y se traza una curva, el gráfico resultante se denomina **Curva Característica de Operación**. Esta curva es utilizada como criterio en estudios de Control de Calidad.

10.5.9 POTENCIA DE LA PRUEBA

La **Potencia de una Prueba** estadística es un concepto relacionado con el E**rror Tipo II**.

Suponga que se define la siguiente hipótesis relacionada con la media:

Ho: $\mu = \mu_0$

Ha: $\mu > \mu_0$

Cálculo del Error Tipo II: β **= P(**Aceptar **Ho** dado que otra hipótesis es verdadera: $\mu = \mu_1$ **)**

Si la muestra es grande, entonces:

$$\beta = P(\bar{X} < c)_{\mu=\mu_1} = P(Z < \frac{c - \mu_1}{\sigma / \sqrt{n}})$$

En donde **c** es el valor crítico de $\bar{X}$ con el que se acepta o rechaza Ho:
Es posible calcular β para otros valores $\mu = \mu_1, \mu_2, \mu_3, \ldots$ por lo tanto, β es una función de μ.
El complemento de $\beta(\mu)$ es otra función de μ y se denomina Potencia de la Prueba $K(\mu)$:

Definición: Potencia de la Prueba

$$K(\mu) = 1 - \beta(\mu)$$

Si β mide la probabilidad de aceptar una hipótesis falsa, entonces la Potencia de la Prueba **K** mide la **probabilidad de rechazar una hipótesis falsa.**

El gráfico de $K(\mu)$ representa la probabilidad de rechazar la hipótesis nula dado que es falsa, para diferentes valores de μ

Ejemplo

Un modelo para la describir el error en la calibración de una máquina es que sea $N(\mu, 4^2)$
Se postula el siguiente contraste de hipótesis

$$\text{Ho: } \mu = 250 \text{ vs. } H_1: \mu > 250$$

Determine el tamaño de la muestra **n** y la cantidad **c** para que la región crítica **R** de la muestra sea

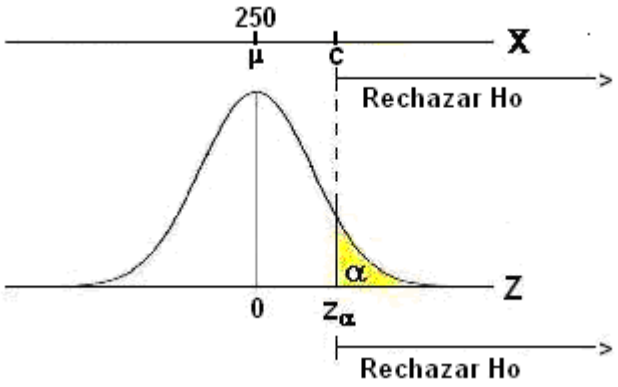
$$R = \{(X_1, X_2, \ldots, X_n) \mid \bar{X} > c\}$$

Se requiere que el nivel de significancia α o **Error Tipo I** de la prueba sea **0.0329**, y que el **Error Tipo II** sea **0.0228** cuando μ valga **252**.

Solución

Modelo poblacional: $X \sim N(\mu, 4^2)$, $\sigma^2 = 4^2 \Rightarrow \sigma = 4$
Hipótesis nula $H_o: \mu = 250$
Hipótesis alterna $H_1: \mu > 250$
La región crítica $R = \{(X_1, X_2, \ldots, X_n) \mid \bar{X} > c\}$ establece que todas las muestras de tamaño **n** deben cumplir que su media aritmética $\bar{X}$ sea mayor a **c**

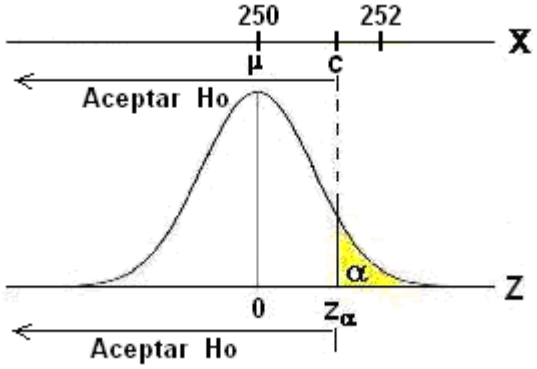
Nivel de significancia de la prueba o Error Tipo I: $\alpha = 0.0329$



$$\alpha = 0.0329 \Rightarrow \alpha = P(Z > Z_\alpha)$$
$$= P(\bar{X} > c)$$
$$= P\left(Z > \frac{c-\mu}{\sigma/\sqrt{n}}\right) \quad \text{con } \mu = 250$$
$$0.0329 = P\left(Z > \frac{c-250}{4/\sqrt{n}}\right)$$
$$\Rightarrow \quad 0.9671 = P\left(Z \le \frac{c-250}{4/\sqrt{n}}\right) \Rightarrow \frac{c-250}{4/\sqrt{n}} = 1.84 \quad (1)$$

Con la Tabla Z

Error Tipo II: $\beta = 0.0228$ **con** $\mu = 252$



$$\beta = 0.0228 \Rightarrow \beta = P(Z \le Z_\alpha) \quad \text{con } \mu = 252$$
$$= P(\bar{X} \le c) \quad \text{con } \mu = 252$$
$$= P\left(Z \le \frac{c-\mu}{\sigma/\sqrt{n}}\right) \quad \text{con } \mu = 252$$
$$0.0228 = P\left(Z \le \frac{c-252}{4/\sqrt{n}}\right) \Rightarrow \frac{c-252}{4/\sqrt{n}} = -2.09 \quad (2)$$

Con la tabla Z

Al resolver las ecuaciones

(1) $\frac{c-250}{4/\sqrt{n}} = 1.84,$ (2) $\frac{c-252}{4/\sqrt{n}} = -2.09$

Se obtienen $c = 250.936,\ n = 61.78 \cong 62$

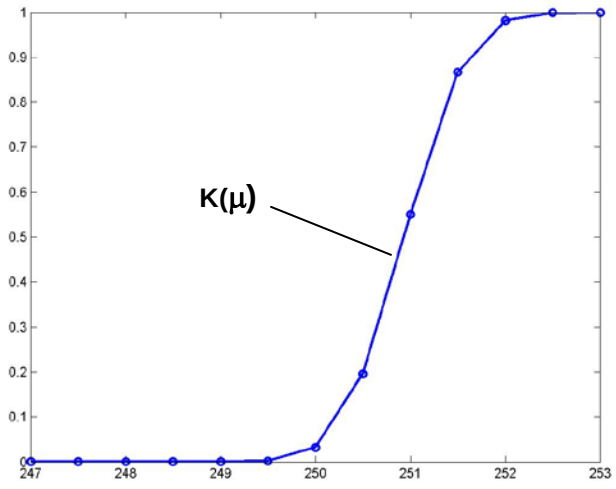
Calcule la Potencia de la Prueba para μ entre 247 y 253. Calcule al menos diez valores

$$\beta = P(Z \le z_\alpha) \quad \text{con } \mu = 247$$
$$= P(\bar{X} \le c) \quad \text{con } \mu = 247$$
$$= P(\bar{X} \le 250.936) \quad \text{con } \mu = 247$$
$$= P\left(Z \le \frac{c-\mu}{\sigma/\sqrt{n}}\right) \quad \text{con } \mu = 247$$
$$= P\left(Z \le \frac{250.936-247}{4/\sqrt{62}}\right)$$
$$= P(Z \le 7.748)$$
$$= F(7.748) \cong 1$$
$$K = 1 - \beta = 1 - 1 = 0$$

Siguiendo este procedimiento con los otros valores de μ, se obtienen los resultados que se muestran en el cuadro más abajo

μ	z	β	$K = 1 - \beta$
247.0	7.7480	1.000	0.000
247.5	6.7638	1.000	0.000
248.0	5.7795	1.000	0.000
248.5	4.7953	1.000	0.000
249.0	3.8110	1.000	0.000
249.5	2.8268	0.997	0.003
250.0	1.8425	0.967	0.033
250.5	0.8583	0.804	0.196
251.0	-0.1260	0.450	0.550
251.5	-1.1102	0.133	0.867
252.0	-2.0945	0.018	0.982
252.5	-3.0787	0.001	0.999
253.0	-4.0630	0.000	1.000

Gráfico de la Potencia de la Prueba



Ejemplo

De una población **$X \sim N(\mu, 7^2)$,** (significa que la variable **X** tiene distribución **normal** con media **μ** y varianza **7^2**), se ha tomado una muestra aleatoria de tamaño **n** para realizar la prueba de hipótesis:

Ho: **$\mu = 15$**

Ha: **$\mu > 15$**

Siendo la región crítica **$\bar{X} > c$**

Se requiere que la **Potencia de la Prueba** tome el valor **0.8** cuando **μ** = 17, y que **la Potencia de la Prueba** tome el valor **0.95** cuando **μ** = 18.

Determine los valores de n, c

Solución

Primero obtenemos los valores respectivos de β

μ	K	β = 1 - K
17	0.8	0.2
18	0.95	0.05

Usamos la fórmula para calcular β: $\beta = P(\bar{X} < c)_{\mu=\mu_1} = P(Z < \frac{c-\mu_1}{\sigma/\sqrt{n}})$

Con μ = 17: $\beta = P(\bar{X} < c)_{\mu=17} = P(Z < \frac{c-17}{7/\sqrt{n}}) = F(\frac{c-17}{7/\sqrt{n}})$

$0.2 = F(\frac{c-17}{7/\sqrt{n}}) \Rightarrow \frac{c-17}{7/\sqrt{n}} = -0.84$ (1) Con la tabla Z

Con μ = 18: $\beta = P(\bar{X} < c)_{\mu=18} = P(Z < \frac{c-18}{7/\sqrt{n}}) = F(\frac{c-18}{7/\sqrt{n}})$

$0.05 = F(\frac{c-18}{7/\sqrt{n}}) \Rightarrow \frac{c-18}{7/\sqrt{n}} = -1.65$ (2) Con la tabla Z

Resolviendo estas dos ecuaciones:

(1) $\frac{c-17}{7/\sqrt{n}} = -0.84$ (2) $\frac{c-18}{7/\sqrt{n}} = -1.65$

Se obtiene **$n \cong 32$, $c = 15.96$**

Calcule el nivel de significancia de la prueba α, o Error Tipo I

Solución

$$\alpha = P(\bar{X} > c)_{\mu=\mu_0} = P(Z > \frac{c-\mu_0}{7/\sqrt{n}}) = P(Z > \frac{15.96-15}{7/\sqrt{32}}) = 1 - F(0.7758) \cong 0.22$$

Calcule la Potencia de la Prueba con $\mu = 12, 13, ..., 19$

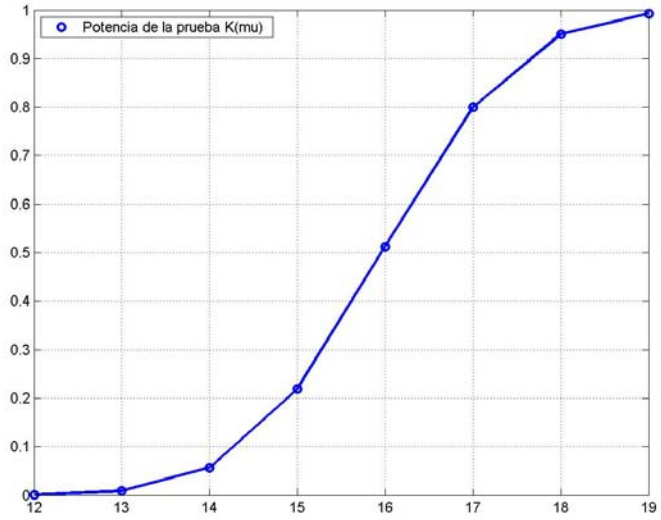
Solución

$$K(\mu) = 1 - \beta(\mu) = 1 - P(\bar{X} < c)_{\mu=\mu_1} = 1 - P(Z < \frac{15.96 - \mu_1}{7/\sqrt{32}})$$

Valores calculados:

μ	β	K=1- β
12	0.999	0.001
13	0.991	0.009
14	0.943	0.057
15	0.781	0.219
16	0.487	0.513
17	0.200	0.800
18	0.049	0.951
19	0.007	0.993

Gráfico de la Potencia de la Prueba



Ejemplo

Se conoce que la estatura de la población en cierto país puede ser modelada como una variable aleatoria **normal** con media μ desconocida y desviación estándar $\sigma = 0.04$ m. Para inferir el valor desconocido de la media se plantea el siguiente contraste de hipótesis:

$H_o: \mu = 1.7$ vs. $H_1: \mu < 1.7$, y se define la región crítica como:

$R = \{(x_1, x_2, \ldots, x_n) \in \Re^n \mid x_1 + x_2 + \ldots + x_n < k\}$

Determine **k** y **n** si se requiere que el nivel de significancia α o **Error Tipo I** sea **0.01**, y que la **Potencia de la Prueba** sea igual a **0.98** cuando $\mu = 1.67$

Solución

Modelo poblacional: $X \sim N(\mu, 0.04^2)$

Hipótesis nula $H_o: \mu = 1.7$

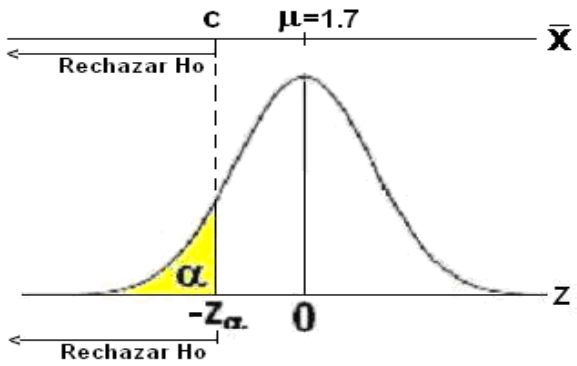
Hipótesis alterna $H_1: \mu < 1.7$ (H_1, es la hipótesis alterna)

La región crítica $R = \{(x_1, x_2, \ldots, x_n) \in \Re^n \mid x_1 + x_2 + \ldots + x_n < k\}$ establece que todas las muestras de tamaño **n** deben cumplir que $x_1 + x_2 + \ldots + x_n < k$

La especificación: $x_1 + x_2 + \ldots + x_n < k$ si se divide para **n** es equivalente a especificar que la región crítica o de rechazo es: $\bar{x} < k/n$. Sea $c = k/n$

Los cálculos se describen a continuación:

Error Tipo I: $\alpha = 0.01$



$$\alpha = 0.01 \Rightarrow \alpha = P(Z < -z_\alpha)$$

$$= P(\bar{X} < c)$$

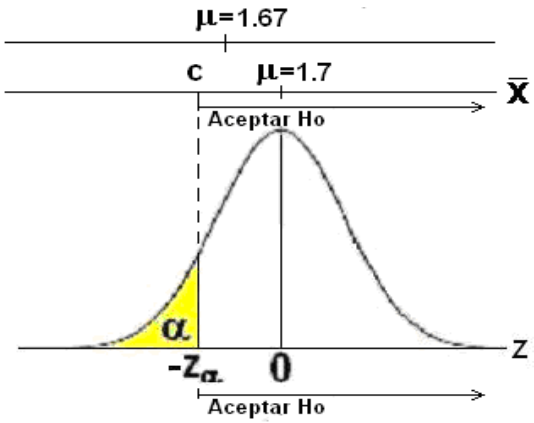
$$= P\left(Z < \frac{c - \mu}{\sigma / \sqrt{n}}\right) \quad \text{con } \mu = 1.7$$

$$0.01 = P\left(Z < \frac{c - 1.7}{0.04 / \sqrt{n}}\right) \Rightarrow \frac{c - 1.7}{0.04 / \sqrt{n}} = -2.33 \quad (1)$$

Con la tabla **Z**

En donde $c = k/n$ es el valor crítico de $\bar{X}$ que define a la región de rechazo de **Ho**

Potencia de la Prueba: $K = 0.98$ cuando $\mu = 1.67$



$K = 0.98$, con $\mu = 1.67 \Rightarrow$ **Error Tipo II:** $\beta = 1 - K = 1 - 0.98 = 0.02$, con $\mu = 1.67$

$$\beta = 0.02 \Rightarrow \beta = P(Z > -Z_\alpha) \quad \text{con } \mu = 1.67$$
$$= P(\bar{X} > c) \quad \text{con } \mu = 1.67$$
$$= P\left(Z > \frac{c-\mu}{\sigma/\sqrt{n}}\right) \quad \text{con } \mu = 1.67$$
$$0.02 = P\left(Z > \frac{c-1.67}{0.04/\sqrt{n}}\right)$$
$$\Rightarrow \quad 0.98 = P\left(Z \le \frac{c-1.67}{0.04/\sqrt{n}}\right) \quad \Rightarrow \quad \frac{c-1.67}{0.04/\sqrt{n}} = 2.055 \qquad (2) \qquad \text{Con la tabla } \mathbf{Z}$$

Al resolver las dos ecuaciones:

$$(1) \qquad \frac{c-1.7}{0.04/\sqrt{n}} = -2.33$$

$$(1) \qquad \frac{c-1.67}{0.04/\sqrt{n}} = 2.055$$

Se obtiene $c = 1.684$, $n = 34.3 \cong 35 \Rightarrow k = nc = 58.94$

10.5.10 EJERCICIOS

1) Una variable aleatoria **X** tiene distribución normal con varianza **49**. Se plantea el siguiente contraste de hipótesis:

$H_o: \mu = 15$ vs $H_a: \mu > 15$

La región crítica para rechazar **Ho** es $R = \{(X_1, X_2, \ldots, X_n) \in \Re^n \mid \bar{X} > c\}$. Esto significa que la media muestral $\bar{X}$ debe ser mayor a **c** para todas las muestras aleatorias reales de tamaño **n** tomadas de la población.

Se desea que el **error tipo I** sea **0.05**, y que el **error tipo II** sea **0.04** cuando $\mu = 17$

a) Determine **c** y **n**

b) Calcule y grafique la potencia de la prueba con $\mu = 13.0, 13.5, 14.0, 14.5, 15.0, 15.5, 16.0$

MATLAB

Potencia de la prueba

Resolver el sistema de ecuaciones del último ejemplo

(1) $\frac{c-17}{7/\sqrt{n}} = -0.84$ (2) $\frac{c-18}{7/\sqrt{n}} = -1.65$

```
>> [c,n]=solve('(c-17)/(7/sqrt(n))=-0.84','(c-18)/(7/sqrt(n))=-1.65')
   c =
        15.9629
    n =
        32.1489
```

Graficar la curva de la potencia de la prueba para el último ejemplo

```
>> mu = 12:19
```
Valores de μ

```
    mu =
         12    13    14    15    16    17    18    19
```

```
>> beta = normcdf((15.96 - mu)/(7/sqrt(32)))
```
Valores de $\beta(\mu)$

```
   beta =
         0.9993    0.9916    0.9434    0.7811    0.4871    0.2003    0.0496    0.0070
```

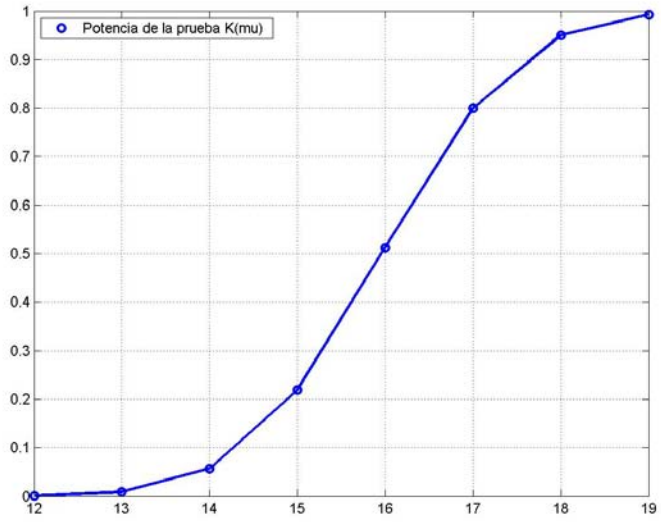
```
>> k = 1- beta
```
Valores de $k(\mu) = 1 - \beta(\mu)$

```
    k =
         0.0007    0.0084    0.0566    0.2189    0.5129    0.7997    0.9504    0.9930
```

```
>> plot(mu,k,'ob'),grid on,hold on
```
Gráfico de los puntos $k(\mu)$

```
>> plot(mu,k,'b')
```
Gráfico de las líneas de $k(\mu)$

```
>> legend('Potencia de la prueba K(mu)',2)
```



10.6 INFERENCIAS RELACIONADAS CON LA PROPORCIÓN

En muchas aplicaciones interesa conocer el valor de un índice, tasa, etc., la cual representa la proporción de datos que consideramos “favorables” del total de datos en la población.

En estas situaciones el modelo de probabilidad es la distribución binomial. Este modelo requiere conocer el valor de probabilidad de "exito" **p** en cada ensayo. Por lo tanto, es de interés práctico determinar o al menos estimar el valor de este parámetro poblacional **p**.

Sea una variable aleatoria **X** con distribución binomial, con media $\mu=np$ y varianza $\sigma^2 = npq$.

De esta población se toma una muestra de tamaño **n** y se obtienen **x** datos favorables. La relación **x/n** se denomina proporción muestral $\bar{p}$ y es un estimador para el parámetro **p.**

Caso $n \geq 30$ (Muestras grandes)

La variable aleatoria $\bar{p}=x/n$ es la media muestral. Esta variable es un estimador insesgado del parámetro **p**, es decir $E(\bar{p}) = p$

Demostración

Media de $\bar{p}$: $\mu_{\bar{p}} = E(\bar{p}) = E(X/n) = 1/n\ E(X) = 1/n\ (np) = p$

Además:

Varianza de $\bar{p}$: $\sigma^2_{\bar{p}} = V(\bar{p}) = V(X/n) = 1/n^2\ V(X) = 1/n^2\ (npq) = \dfrac{pq}{n}$

10.6.1 ESTIMACIÓN PUNTUAL

Parámetro: **p** (Es la proporción poblacional y cuyo valor se desea estimar)
Población con distribución **binomial** con media μ y varianza σ^2 desconocidas
Estimador: $\bar{p} = x/n$ (Proporción muestral, se usa para estimar al parámetro)
Muestras grandes ($n \geq 30$).

Por el Teorema del Límite Central, el estadístico

$$Z = \frac{\bar{p} - \mu_{\bar{p}}}{\sigma_{\bar{p}}} = \frac{\bar{p} - p}{\sqrt{pq/n}}$$ tendrá aproximadamente distribución normal estándar.

Fórmula para la Estimación Puntual de la Proporción

Este análisis es similar al realizado para la estimación de la media muestral cuando $n \geq 30$.
Suponer especificado un valor de probabilidad $1 - \alpha$ ubicado en la parte central del dominio de **Z**

La desigualdad $-z_{\alpha/2} \leq Z \leq z_{\alpha/2}$ se satisface con probabilidad $1 - \alpha$

Equivale a decir que $|Z| \leq z_{\alpha/2}$ tiene probabilidad $1 - \alpha$

Sustituyendo **Z** se obtiene: $\left|\dfrac{\bar{p} - p}{\sqrt{pq/n}}\right| \leq z_{\alpha/2}$, con probabilidad $1 - \alpha$

De donde $|\bar{p} - p| \leq z_{\alpha/2}\sqrt{\frac{pq}{n}}$ con probabilidad $1 - \alpha$

$|\bar{p} - p|$ es el error en la estimación de p mediante $\bar{p}$

Definición: Estimación Puntual de la Proporción con Probabilidad 1 - α, $n \geq 30$

$$E = z_{\alpha/2}\sqrt{\frac{pq}{n}}$$ Es el máximo error en la estimación de p

Para evaluarlo, se usa la varianza muestral como aproximación para la varianza poblacional:

$$\frac{pq}{n} \cong \frac{\bar{p}\bar{q}}{n}$$

10.6.2 ESTIMACIÓN POR INTERVALO

Parámetro: p (Es la medida poblacional cuyo valor se desea estimar)
Población con distribución **binomial** con media μ y varianza σ^2 desconocidas
Estimador: $\bar{p} = x/n$ (Proporción muestral)
Muestras grandes ($n \geq 30$).

Por el Teorema del Límite Central, el estadístico

$$Z = \frac{\bar{p} - \mu_{\bar{p}}}{\sigma_{\bar{p}}} = \frac{\bar{p} - p}{\sqrt{pq/n}}$$ tendrá aproximadamente distribución normal estándar.

Fórmula para Estimación por Intervalo de la Proporción

En la misma desigualdad anterior:

$-z_{\alpha/2} \leq Z \leq z_{\alpha/2}$, con probabilidad $1 - \alpha$

Sustituimos $Z = \frac{\bar{p} - p}{\sqrt{\frac{pq}{n}}}$ y despejamos del numerador el parámetro p

Definición: Intervalo de Confianza para la Proporción con Nivel $1 - \alpha$, $n \geq 30$

$$\bar{p} - z_{\alpha/2}\sqrt{\frac{pq}{n}} \leq p \leq \bar{p} + z_{\alpha/2}\sqrt{\frac{pq}{n}}$$

Para evaluarlo, se usa la varianza muestral como aproximación para la varianza poblacional:

$$\frac{pq}{n} \cong \frac{\bar{p}\bar{q}}{n}$$

Ejemplo
En un estudio de mercado para un producto se tomó una muestra aleatoria de 400 personas de las cuales 140 respondieron favorablemente.

Encuentre el error máximo en la estimación con probabilidad de 95%

$$1 - \alpha = 0.95 \Rightarrow z_{\alpha/2} = z_{0.025} = 1.96$$

$$\bar{p} = 140/400 = 0.35$$

$$E = z_{\alpha/2}\sqrt{\frac{pq}{n}} \cong 1.96\sqrt{\frac{(0.35)(0.65)}{400}} = 4.67\%$$

Encuentre un intervalo de confianza para p con un nivel de 95%

$$\bar{p} - z_{\alpha/2}\sqrt{\frac{pq}{n}} \le p \le \bar{p} + z_{\alpha/2}\sqrt{\frac{pq}{n}}$$

$$0.35 - 1.96\sqrt{\frac{(0.35)(0.65)}{400}} \le p \le 0.35 + 1.96\sqrt{\frac{(0.35)(0.65)}{400}}$$

$$0.303 \le p \le 0.397$$

Se puede afirmar con una confianza del 95% que la proporción de personas en la población que favorecen al producto está entre 30.3% y 39.7%

10.6.3 PRUEBA DE HIPÓTESIS

Parámetro: **p** (Es la medida poblacional cuyo valor se desea estimar)
Población con distribución **binomial** con media μ y varianza σ^2 desconocidas
Estimador: $\bar{p} = x/n$ (Proporción muestral)
Muestras grandes ($n \ge 30$).
Valor propuesto para el parámetro: p_0
Por el Teorema del Límite Central, el estadístico

$$Z = \frac{\bar{p} - \mu_{\bar{p}}}{\sigma_{\bar{p}}} = \frac{\bar{p} - p_0}{\sqrt{p_0 q_0 / n}}$$ tendrá aproximadamente distribución normal estándar.

Procedimiento Básico

1) Formular la hipótesis nula: **Ho:** $p = p_0$ (algún valor específico para **p**)

2) Formular una hipótesis alterna, elegir una entre:

Ha: $p < p_0$
Ha: $p > p_0$
Ha: $p \ne p_0$

3) Especificar el nivel de significancia α para la prueba

4) Seleccionar el estadístico de prueba y definir la región de rechazo

$$z = \frac{\bar{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$$ por el teorema del límite central tiene distribución normal estándar

Ha	Región de rechazo de Ho en favor de Ha
$p < p_0$	$z < -z_\alpha$
$p > p_0$	$z > z_\alpha$
$p \neq p_0$	$z < -z_{\alpha/2} \vee z > z_{\alpha/2}$

5) Con los datos de la muestra calcule el valor del estadístico

6) Si el valor del estadístico de prueba cae en la región de rechazo, la decisión es rechazar Ho en favor de Ha. Caso contrario, se dice que no hay evidencia suficiente para rechazar Ho.

Ejemplo

La norma para la cantidad de artículos de artículos aceptables producidos por una fábrica es ≥90%. Se ha tomado una muestra aleatoria de 175 artículos y se encontraron 150 artículos aceptables. Pruebe con una significancia de 5% que no se está cumpliendo con la norma

Solución

Sea **p**: proporción de artículos aceptables que produce la fábrica

$\bar{p} = \mathbf{x/n} = \mathbf{150/175} = \mathbf{0.857} = \mathbf{85.7\%}$

¿Es esto una evidencia de que $\mathbf{p < 90\%}$ o puede atribuirse únicamente a la aleatoriedad de los datos, con **5%** de probabilidad de equivocarnos?

1) Ho: $\mathbf{p = 0.9}$

2) Ha: $\mathbf{p < 0.9}$

3) Nivel de significancia de la prueba $\alpha = \mathbf{0.05}$

4) Estadístico de prueba

$$z = \frac{\bar{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$$

Región de rechazo de Ho

$\alpha = \mathbf{0.5}$, $z_\alpha = z_{0.05} = \mathbf{1.645}$

Rrechazar **Ho** si $\mathbf{z < -1.645}$

5) $$z = \frac{\bar{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} = \frac{0.857 - 0.9}{\sqrt{\frac{(0.9)(0.1)}{175}}} = -1.869 \Rightarrow z < -1.645$$

6) **Decisión:** Hay evidencia suficiente para afirmar que, con una significancia de 5%, no se cumple la norma.

10.6.4 EJERCICIOS

1) Se ha tomado una muestra aleatoria de 200 artículos producidos por una empresa y se observó que 175 fueron aceptables. Encuentre un intervalo de confianza de 95% para la proporción de artículos aceptables.

2) Una muestra aleatoria de 400 observaciones produjo 150 resultados considerados éxitos. Es de interés para una investigación probar que la proporción de éxitos difiere de 0.4
a) Proponga la hipótesis nula y la hipótesis alterna
b) Realice una prueba para determinar si hay evidencia suficiente para rechazar la hipótesis nula en favor de la hipótesis alterna, con 10% de significancia.

3) Una empresa realizó un estudio de mercado de su producto para lo cual consultó a 200 consumidores. 28 expresaron su preferencia por el producto de la empresa. El fabricante cree, con este resultado que tiene el 10% del mercado para su producto. Pruebe con 5% de significancia si esta afirmación es correcta.

10.7 INFERENCIAS RELACIONADAS CON LA VARIANZA

Para algunas pruebas y aplicaciones estadísticas, es importante estimar el valor de la varianza poblacional σ^2.

Suponer una población con distribución normal o aproximadamente normal de la cual se toma una muestra aleatoria de tamaño **n** y se obtiene la varianza muestral S^2**:**

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 , \quad \bar{X} = \frac{1}{n}\sum_{i=1}^{n}X_i$$

El estadístico S^2 es un estimador insesgado del parámetro σ^2 puesto que:

$$E(S^2) = \sigma^2$$

También se puede probar la siguiente fórmula para la varianza muestral:

$$V(S^2) = \frac{2\sigma^4}{n-1}, \quad n > 1$$

Características

Parámetro**:** σ^2 (Es la medida poblacional cuyo valor se desea estimar)

Población con distribución normal

Estimador**:** S^2 (Varianza muestral, se usa para estimar al parámetro

El estadístico de prueba para realizar inferencias es $\chi^2 = (n-1)\frac{S^2}{\sigma^2}$

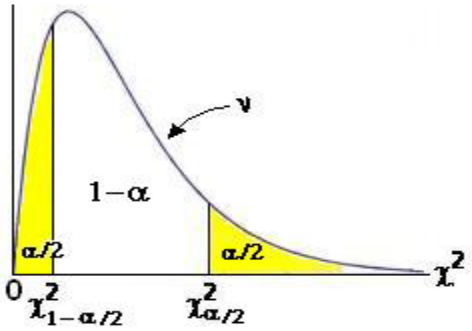
que tiene distribución Ji-cuadrado con ν = **n – 1** grados de libertad

10.7.1 INTERVALO DE CONFIANZA

Para definir un intervalo de confianza, se sigue un procedimiento similar a otros parámetros.

Definimos un intervalo central para la variable χ^2 con área o probabilidad $1 - \alpha$, y la diferencia α se reparte a ambos lados en dos áreas iguales con valor $\alpha/2$.

Debido a que la distribución de χ^2 es asimétrica, los valores de esta variable no tienen la misma distancia desde el centro y se los representa con $\chi^2_{1-\alpha/2}$ y $\chi^2_{\alpha/2}$ de acuerdo a la definición establecida para uso de la Tabla Ji-cuadrado.



Entonces, con probabilidad $1 - \alpha$ se puede construir un intervalo para χ^2**:**

$$\chi^2_{1-\alpha/2} \le \chi^2 \le \chi^2_{\alpha/2}$$

Si se sustituye la definición de la variable aleatoria $\chi^2 = (n-1)\frac{S^2}{\sigma^2}$ y se despeja el parámetro de interés σ^2 se obtiene

Definición: Intervalo de confianza para la Varianza con Nivel $1 - \alpha$

$$(n-1)\frac{S^2}{\chi^2_{\alpha/2}} \le \sigma^2 \le (n-1)\frac{S^2}{\chi^2_{1-\alpha/2}}$$

Ejemplo

En una muestra aleatoria se registró el peso de 10 paquetes y se obtuvieron los siguientes resultados en gramos: 46.4, 46.1, 45.8, 47.0, 46.1, 45.9, 45.8, 41.9, 45.2, 46.0

Encuentre un intervalo de confianza para la varianza del peso de toda la producción, con un nivel de 95%. Suponga que la población tiene distribución normal

$$n = 10, \quad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{10}[46.4 + 46.1 + \dots] = 45.62$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{9}[(46.4 - 45.62)^2 + (46.1 - 45.62)^2 + \dots] = 1.919$$

$1 - \alpha = 0.95, \ \nu = n - 1 = 9 \quad \Rightarrow \chi^2_{\alpha/2} = \chi^2_{0.025} = 19.02$ (Tabla χ^2)

$\Rightarrow \chi^2_{1-\alpha/2} = \chi^2_{0.975} = 2.7$ (Tabla χ^2)

Se sustituye en la definición del intervalo de confianza:

$$(n-1)\frac{S^2}{\chi^2_{\alpha/2}} \le \sigma^2 \le (n-1)\frac{S^2}{\chi^2_{1-\alpha/2}}$$

$9\,(1.919/19.02) \le \sigma^2 \le 9\,(1.919/2.7) \Rightarrow 0.908 \le \sigma^2 \le 6.398$

Se puede afirmar con una confianza de 95% que la varianza poblacional se encuentra en el intervalo **[0,908, 6.398]**

10.7.2 PRUEBA DE HIPÓTESIS

Se usa el mismo procedimiento básico para los parámetros estudiados anteriormente:

1) Definir la hipótesis nula **Ho:** $\sigma^2 = \sigma^2_o$ (algún valor especificado)

2) Elegir una hipótesis alterna: **Ha:** $\sigma^2 < \sigma^2_o$

Ha: $\sigma^2 > \sigma^2_o$

Ha: $\sigma^2 \neq \sigma^2_o$

3) Seleccionar el nivel de significancia α

4) Estadístico de prueba

$\chi^2 = (n-1)\frac{S^2}{\sigma_o^2}$, distribución ji-cuadrado con $\nu = n-1$ grados de libertad

Región crítica

Ha	**Región de rechazo de Ho en favor de Ha**
$\sigma^2 < \sigma_o^2$	$\chi^2 < \chi^2_{1-\alpha}$
$\sigma^2 > \sigma_o^2$	$\chi^2 > \chi^2_{\alpha}$
$\sigma^2 \neq \sigma_o^2$	$\chi^2 < \chi^2_{1-\alpha/2} \vee \chi^2 > \chi^2_{\alpha/2}$

5) Calcular el valor del estadístico de prueba con los datos de la muestra

6) Tomar una decisión.

Ejemplo

Un fabricante afirma que la duración de su producto tiene distribución aproximadamente normal con una **desviación estándar de 0.9** años.

Una muestra aleatoria de **10** productos tuvo una desviación estándar de **1.2** años. Pruebe, con una significancia de **5%**, si esta evidencia es suficiente para afirmar que la desviación estándar poblacional es mayor a la especificada

La prueba es aplicable a la varianza σ^2 por lo tanto $\sigma^2 = (0.9)^2 = 0.81$

1) Ho: $\sigma^2 = 0.81$
2) Ha: $\sigma^2 > 0.81$
3) $\alpha = 0.05$
4) Estadístico de prueba

$\chi^2 = (n-1)\frac{S^2}{\sigma_o^2}$, distribución ji-cuadrado con $\nu = n-1$ grados de libertad

Región de rechazo

$\alpha = 0.05$, $\nu = n - 1 = 9 \Rightarrow \chi^2_{0.05} = 16.91$

Rechazar **Ho** si $\chi^2 > 16.91$

5) $\chi^2 = (n-1)\frac{S^2}{\sigma_o^2} = 9\frac{(1.2)^2}{0.81} = 16.0$

6) Con 5% de significancia se puede concluir que no hay evidencia suficiente para rechazar la afirmación del fabricante

10.7.3 EJERCICIOS

1) Se tomó una muestra aleatoria de 15 observaciones de una población normal y se obtuvo que la media y la varianza muestrales fueron respectivamente 3.92 y 0.325. Encuentre un intervalo de confianza de 90 para varianza de la población.

2) Una muestra aleatoria de 20 observaciones tomada de una población normal produjo una varianza muestral igual a 18.2. Determine si los datos proporcionan suficiente evidencia para afirmar que ñla varianza poblacional es mayor a 15. Haga la prueba con 5% de significancia.

3) El fabricante de un artículo afirma que la resistencia media de su artículo tiene distribución normal con una desviación estándar de 0.5. Una muestra aleatoria 4 observaciones produjo los siguientes resultados de su resistencia: 5.2 4.3 3.7 3.9 5.7. Realice una prueba con 5% de sigificancia para determinar si la desviación estándar especificada por el fabricante es cierta.

4) Un fabricante de cables de cobre afirma que la resistencia de su producto tiene distribución normal con varianza de 100.
Al probar la resistencia de cuatro artículos de una muestra aleatoria se obtuvieron los siguientes resultados: 130, 152, 128, 145.
Pruebe con una significancia de 5% que la varianza excede a la especificación.

MATLAB

Obtención de un intervalo de confianza para la varianza σ^2

Vector conteniendo una muestra de diez datos

```
>> u=[46.4 46.1 45.8 47.0 46.1 45.9 45.8 41.9 45.2 46.0];
>> v=var(u)                    Varianza muestral
    v =
        1.9196
>> ja=chi2inv(0.975,9)         Valor del estadístico χ² para α = 0.025, ν = 9
   ja =
       19.0228
>> j1a=chi2inv(0.025,9)        Valor del estadístico χ² para α = 0.975, ν = 9
  j1a =
        2.7004
>> x=[9*v/ja, 9*v/j1a]         Intervalo de confianza bilateral para σ²
    x =
        0.9082    6.3976
```

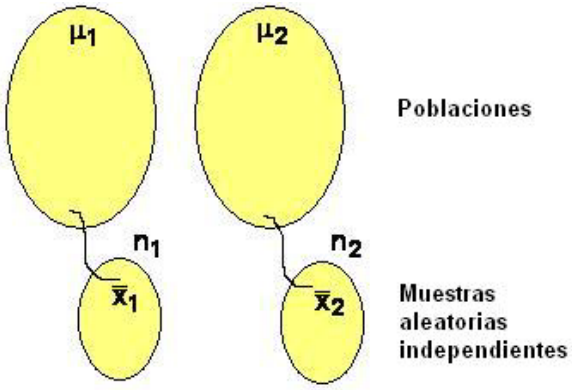
10.8 INFERENCIAS RELACIONADAS CON LA DIFERENCIA ENTRE DOS MEDIAS

10.8.1 ESTIMACIÓN PUNTUAL E INTERVALO DE CONFIANZA

CASO: Muestras grandes (n≥30)

En esta sección se desarrolla la técnica para comparar las medias de dos poblaciones.

Supongamos dos poblaciones de las cuales se toman **muestras aleatorias independientes** y se usa la diferencia de las medias muestrales para estimar la diferencia de las medias poblacionales.



Parámetro: $\mu_1 - \mu_2$ Diferencia de medias poblacionales

Poblaciones con distribuciones desconocidas, con varianzas σ_1^2, σ_2^2

Estimador: $\bar{X}_1 - \bar{X}_2$ Diferencia de medias muestrales

Muestras aleatorias **independientes** de tamaños n_1 y n_2 mayores o iguales a **30**

Media y varianza del estimador:

$$\mu_{\bar{x}_1-\bar{x}_2} = E(\bar{X}_1 - \bar{X}_2) = E(\bar{X}_1) - E(\bar{X}_2) = \mu_1 - \mu_2$$ (Es un estimador insesgado)

$$\sigma^2_{\bar{X}_1-\bar{X}_2} = V(\bar{X}_1 - \bar{X}_2) = V[(1)\bar{X}_1 + (-1)\bar{X}_2] = (1)^2V(\bar{X}_1) + (-1)^2V(\bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

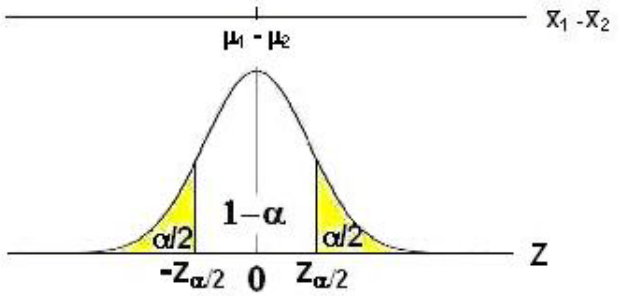
Adicionalmente, pueden aproximarse las varianzas poblacionales con las varianzas muestrales: $\sigma_1^2 \cong S_1^2$, $\sigma_2^2 \cong S_2^2$

Siendo las muestras grandes, por el Teorema del Límite Central, el estadístico

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - \mu_{\bar{x}_1-\bar{x}_2}}{\sigma_{\bar{x}_1-\bar{x}_2}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}},$$

tiene distribución normal estándar aproximadamente,

Con un planteamiento similar al realizado en casos anteriores se tiene



$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Con probabilidad **1 - α**, se cumple la desigualdad: $-z_{\alpha/2} \le Z \le z_{\alpha/2}$

Sustituyendo **Z** y con la definición de error en la estimación se obtiene:

Definición: Error máximo en la estimación de $\mu_1 - \mu_2$ con probabilidad 1 - α

$$E = z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Sustituyendo **Z** y despejando el parámetro de interés $\mu_1 - \mu_2$ se obtiene:

Definición: Intervalo de confianza para $\mu_1 - \mu_2$ con nivel 1 - α

$$(\bar{X}_1 - \bar{X}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le \mu_1 - \mu_2 \le (\bar{X}_1 - \bar{X}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Ejemplo

De dos poblaciones, 1 y 2, se tomaron muestras aleatorias independientes y se obtuvieron los siguientes resultados:

Muestra	n	$\bar{x}$	S^2
1	36	12.7	1.38
2	49	7.4	4.14

Encuentre el mayor error en la estimación puntual de $\mu_1 - \mu_2$ con probabilidad **95%**

$1-\alpha = 0.95 \Rightarrow z_{\alpha/2} = z_{0.025} = 1.96.$ Sustituimos en la fórmula:

$$E = z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \cong 1.96\sqrt{\frac{1.38}{36} + \frac{4.14}{49}} = 0.687$$

Encuentre un intervalo de confianza para $\mu_1 - \mu_2$ con nivel 95%

Sustituimos en la fórmula respectiva:

$$(12.7 - 7.4) - 1.96\sqrt{\frac{1.38}{36} + \frac{4.14}{49}} \leq \mu_1 - \mu_2 \leq (12.7 - 7.4) + 1.96\sqrt{\frac{1.38}{36} + \frac{4.14}{49}}$$

$$4.613 \leq \mu_1 - \mu_2 \leq 5.987$$

Con los datos de las muestras se puede afirmar con una confianza de 95% que μ_1 es mayor a μ_2 en un valor que puede ir desde **4.613** hasta **5.987**

10.8.2 PRUEBA DE HIPÓTESIS

CASO: Muestras grandes (n≥30)

PROCEDIMIENTO BÁSICO

1) Formular la hipótesis nula: Ho: $\mu_1 - \mu_2 = d_0$ (usualmente $d_0=0$ para probar Ho: $\mu_1 = \mu_2$)

2) Formular una hipótesis alterna. Elegir una entre:

Ha: $\mu_1 - \mu_2 < d_0$
Ha: $\mu_1 - \mu_2 > d_0$
Ha: $\mu_1 - \mu_2 \neq d_0$

3) Especificar el nivel de significancia para la prueba α

4) Seleccionar el estadístico de prueba y definir la región de rechazo de Ho

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$ tiene distribución normal estándar aproximadamente

Adicionalmente: $\sigma_1^2 \cong S_1^2$, $\sigma_2^2 \cong S_2^2$

Ha	Región de rechazo de Ho en favor de Ha
$\mu_1 - \mu_2 < d_0$	$z < -z_\alpha$
$\mu_1 - \mu_2 > d_0$	$z > z_\alpha$
$\mu_1 - \mu_2 \neq d_0$	$z<-z_{\alpha/2} \vee z > z_{\alpha/2}$

5) Con los datos de la muestra calcule el valor del estadístico

6) Si el valor del estadístico de prueba cae en la región de rechazo, la decisión es rechazar Ho en favor de Ha. Caso contrario, se dice que no hay evidencia suficiente para rechazar Ho.

Ejemplo. Suponga los siguientes datos correspondientes a dos muestras aleatorias independientes tomadas de dos poblaciones cuyas medias se desea estudiar

Muestra	n	$\bar{x}$	S^2
1	75	82	64
2	50	76	36

Pruebe la hipótesis $\mu_1 > \mu_2$ con un nivel de significancia de 10%

Solución

1) **Ho:** $\mu_1 - \mu_2 = 0$

2) **Ha:** $\mu_1 - \mu_2 > 0$

3) $\alpha = 0.1$

4) $$Z = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$z_\alpha = 1.28$: Rechazar Ho si $z > 1.28$

5) $$Z = \frac{(82 - 76) - 0}{\sqrt{\frac{64}{75} + \frac{36}{50}}} = 4.78$$

6) Con una significancia de 10% se acepta que $\mu_1 > \mu_2$

10.8.3 INTERVALO DE CONFIANZA

Muestras pequeñas (n<30)

En esta sección se desarrolla la técnica para comparar las medias de dos poblaciones. Supongamos dos poblaciones de las cuales se toman **muestras aleatorias independientes** para usar la diferencia de las medias muestrales como una estimación de las medias poblacionales.

Parámetro: $\mu_1 - \mu_2$ Diferencia de medias poblacionales

Poblaciones con distribuciones **normales**, con varianzas σ_1^2, σ_2^2 **desconocidas**

Estimador: $\bar{X}_1 - \bar{X}_2$ Diferencia de medias muestrales

Muestras aleatorias **independientes** de tamaños n_1 y n_2 **menores** a **30**

Media del estimador

$$\mu_{\bar{X}_1-\bar{X}_2} = E[\bar{X}_1 - \bar{X}_2] = E[\bar{X}_1] - E[\bar{X}_2] = \mu_1 - \mu_2$$ (Estimador insesgado)

Estadístico de prueba

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_{\bar{X}_1-\bar{X}_2}}, \quad \text{distribución } T$$

Nota: Si las varianzas poblacionales σ_1^2, σ_2^2 fuesen conocidas teniendo las poblaciones distribución normal el estadístico tendría distribución normal estándar**,** sin importar el tamaño de las muestras

La teoría estadística provee adicionalmente una prueba para verificar estas suposiciones acerca de las varianzas, la misma que se estudiará posteriormente.

Se analizan dos situaciones acerca de las varianzas: $\sigma_1^2 = \sigma_2^2$ y $\sigma_1^2 \neq \sigma_2^2$.

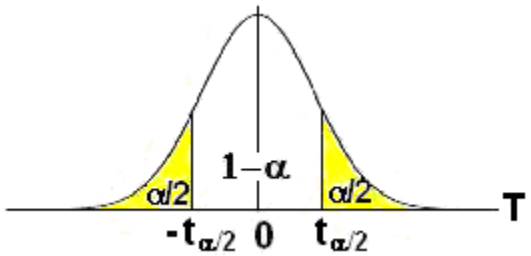
a) Caso: $\sigma_1^2 = \sigma_2^2$

Estadístico de prueba

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_{\bar{X}_1-\bar{X}_2}}$$, distribución T con $\nu = n_1 + n_2 - 2$ grados de libertad

$$S_{\bar{X}_1-\bar{X}_2} = S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \quad S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2}$$

Con un planteamiento similar al realizado en casos anteriores:



$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_{\bar{X}_1 - \bar{X}_2}}$$

Con probabilidad $1 - \alpha$, se tiene la desigualdad: $-t_{\alpha/2} \leq T \leq t_{\alpha/2}$

Sustituyendo **T** y despejando el parámetro de interés $\mu_1 - \mu_2$ se obtiene:

Definición: Intervalo de confianza para $\mu_1 - \mu_2$ con nivel $1 - \alpha$, con $\sigma_1^2 = \sigma_2^2$

$$(\bar{X}_1 - \bar{X}_2) - t_{\alpha/2} S_{\bar{X}_1 - \bar{X}_2} \leq \mu_1 - \mu_2 \leq (\bar{X}_1 - \bar{X}_2) + t_{\alpha/2} S_{\bar{X}_1 - \bar{X}_2}$$

b) Caso: $\sigma_1^2 \neq \sigma_2^2$

Estadístico de prueba

$T = \dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_{\bar{X}_1 - \bar{X}_2}}$, distribución **T** con $\nu = \dfrac{\left(\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}\right)^2}{\dfrac{\left(\dfrac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \dfrac{\left(\dfrac{S_2^2}{n_2}\right)^2}{n_2 - 1}}$ grados de libertad

$S_{\bar{X}_1 - \bar{X}_2} = \sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}$,

Definición: Intervalo de confianza para $\mu_1 - \mu_2$ con nivel $1 - \alpha$, $\sigma_1^2 \neq \sigma_2^2$

$$(\bar{X}_1 - \bar{X}_2) - t_{\alpha/2} S_{\bar{X}_1 - \bar{X}_2} \leq \mu_1 - \mu_2 \leq (\bar{X}_1 - \bar{X}_2) + t_{\alpha/2} S_{\bar{X}_1 - \bar{X}_2}$$

10.8.4 PRUEBA DE HIPÓTESIS

Muestras pequeñas (n<30)

a) Caso: $\sigma_1^2 = \sigma_2^2$

1) Ho: $\mu_1 - \mu_2 = d_0$ (usualmente $d_0 = 0$)

2) Ha: $\mu_1 - \mu_2 < d_0$

$\mu_1 - \mu_2 > d_0$

$\mu_1 - \mu_2 \neq d_0$

3) α**:** nivel de significancia

4) Estadístico de prueba y región de rechazo

$t = \dfrac{(\bar{X}_1 - \bar{X}_2) - d_0}{S_{\bar{X}_1 - \bar{X}_2}}$, distribución **T** con $\nu = n_1 + n_2 - 2$ grados de libertad

$$S_{\bar{X}_1 - \bar{X}_2} = S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \qquad S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Ha	Región de rechazo de Ho
$\mu_1 - \mu_2 < d_0$	$t < -t_\alpha$
$\mu_1 - \mu_2 > d_0$	$t > t_\alpha$
$\mu_1 - \mu_2 \neq d_0$	$t < -t_{\alpha/2} \vee t > t_{\alpha/2}$

b) Caso: $\sigma_1^2 \neq \sigma_2^2$

1) Ho: $\mu_1 - \mu_2 = d_0$ (usualmente $d_0 = 0$)

2) Ha: $\mu_1 - \mu_2 < d_0$

$\mu_1 - \mu_2 > d_0$

$\mu_1 - \mu_2 \neq d_0$

3) α**:** nivel de significancia

4) Estadístico de prueba y región de rechazo

$T = \dfrac{(\bar{X}_1 - \bar{X}_2) - d_0}{S_{\bar{X}_1 - \bar{X}_2}}$, distribución **T** con $\nu = \dfrac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}$ grados de libertad

$$S_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

Ha	Región de rechazo de Ho
$\mu_1 - \mu_2 < d_0$	$t < -t_\alpha$
$\mu_1 - \mu_2 > d_0$	$t > t_\alpha$
$\mu_1 - \mu_2 \neq d_0$	$t < -t_{\alpha/2} \vee t > t_{\alpha/2}$

Ejemplo.

(Caso: $\sigma_1^2 = \sigma_2^2$)

Se realizó un experimento para comparar la resistencia de dos materiales, obteniéndose los siguientes resultados:

Material	n	$\bar{X}$	S
1	12	85	4
2	10	81	5

Suponga que son muestras aleatorias independientes y que provienen de poblaciones normales con varianzas desconocidas pero que se pueden considerar iguales.

Pruebe con 5% de significancia que la resistencia del material uno excede a la resistencia del material dos en dos unidades.

Solución

1) **Ho:** $\mu_1 - \mu_2 = 2$

2) **Ha:** $\mu_1 - \mu_2 > 2$

3) $\alpha = 0.05$

4) Estadístico de prueba

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - d_0}{S_{\bar{X}_1 - \bar{X}_2}},$$ distribución **T** con $\nu = n_1 + n_2 - 2$ grados de libertad

Región de rechazo de Ho

$\alpha = 0.05$, $\nu = n_1 + n_2 - 2 = 12 + 10 - 2 = 20 \Rightarrow t_{0.05} = 1.725$ (Tabla T)

$t > 1.725$

5) Cálculo del valor del estadístico de prueba

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = \frac{(12-1)4^2 + (10-1)5^2}{12 + 10 - 2} = 20.05$$

$$S_{\bar{X}_1 - \bar{X}_2} = S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = \sqrt{20.05}\sqrt{\frac{1}{12} + \frac{1}{10}} = 1.917$$

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - d_0}{S_{\bar{X}_1 - \bar{X}_2}} = \frac{(85 - 81) - 2}{1.917} = 1.043$$

6) **t** no cae en la región de rechazo de **Ho** por lo tanto, con 5% de significancia, no hay evidencia suficiente para rechazar que el material 1 excede al material 2 en mas de unidades.

Ejemplo.

(Caso: $\sigma_1^2 \neq \sigma_2^2$)

Se realizó un experimento para comparar la resistencia de dos materiales, obteniéndose los siguientes resultados:

Material	n	$\bar{X}$	S^2
1	15	3.84	3.07
2	12	1.49	0.80

Suponga que son muestras aleatorias independientes y que provienen de poblaciones normales con varianzas desconocidas, suponer diferentes.

Encuentre un intervalo de confianza de 95% para la diferencia de las medias poblacionales $\mu_1 - \mu_2$.

Solución

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}} = \frac{\left(\frac{3.07}{15} + \frac{0.80}{12}\right)^2}{\frac{\left(\frac{3.07}{15}\right)^2}{15 - 1} + \frac{\left(\frac{0.80}{12}\right)^2}{12 - 1}} \cong 21$$

$1 - \alpha = 0.95 \Rightarrow \alpha/2 = 0.025,\ \nu = 21, \Rightarrow t_{\alpha/2} = t_{0.025} = 2.08$ (Tabla T)

$$S_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} = \sqrt{\frac{3.07}{15} + \frac{0.80}{12}} = 0.521$$

Sustituimos en la fórmula respectiva:

$$(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2} S_{\bar{X}_1 - \bar{X}_2} \leq \mu_1 - \mu_2 \leq (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2} S_{\bar{X}_1 - \bar{X}_2}$$

$$(3.84 - 1.49) - 2.08(0.521) \leq \mu_1 - \mu_2 \leq (3.84 - 1.49) + 2.08(0.521)$$

$$1.266 \leq \mu_1 - \mu_2 \leq 3.434$$

Por lo tanto, se puede afirmar con una confianza de **95%** que la diferencia de las medias de la resistencia de los dos materiales está entre **1.266** y **3.434**

10.8.5 EJERCICIOS

1) De dos poblaciones se tomaron muestras aleatorias independientes y se obtuvieron los siguientes resultados:

Muestra	n	$\bar{x}$	S^2
1	36	1.24	0.056
2	45	1.31	0.054

a) Encuentre un intervalo de confianza para $\mu_1 - \mu_2$ con nivel 90%.
b) Con una significancia de 5% realice una prueba para determinar si la evidencia de las muestras es suficiente para afirmar que las medias poblacionales son diferentes.

2) De dos procesos de producción 1 y 2, se tomaron dos muestras aleatorias independientes y se obtuvieron los siguientes resultados del tiempo de producción de los artículos.

Muestra 1: 14, 10, 8, 12
Muestra 2: 12, 9, 7, 10, 6

Suponga que las poblaciones tienen distribución normal con varianzas aproximadamente iguales

a) Encuentre un intervalo de confianza de 95% para $\mu_1 - \mu_2$
b) Pruebe con 5% de significancia que $\mu_1 > \mu_2$

MATLAB

Inferencias relacionadas con dos medias. Muestras pequeñas. Varianzas iguales

```
>> x=normrnd(22,3,1,10)
   x =
     20.3213   23.3310   19.1503   24.3435   23.7069
     19.5349   21.2032   18.4367   15.3930   24.9590
```

Muestra aleatoria **X:** una fila con 10 cols. **X ~ N(22, 3)**

```
>> y=normrnd(20,3,1,15)
   y =
     18.4441   20.9821   20.7022   20.0644   16.9882
     17.1586   18.8767   16.4423   16.8323   24.4174
     20.1672   16.3480   19.8763   16.6150   15.9522
```

Muestra aleatoria **Y:** una fila con 15 cols. **Y ~ N(20, 3)**

```
>> [h, p, ci, stats]=ttest2(x, y, 0.05, 1)
   h = 1
   p = 0.0193
   ci = 0.5211      Inf
   stats =  tstat: 2.1943
            df: 23
```

Prueba **Ho:** $\mu_X = \mu_Y$ vs. **Ha:** $\mu_X > \mu_Y$, $\sigma_X^2 = \sigma_Y^2$, $\alpha = 0.05$. Prueba unilateral derecha

h =1 $\Rightarrow$ La evidencia es suficiente para rechazar **Ho**

Valor **p** de la prueba

Intervalo de confianza con nivel $1 - \alpha$

Valor del estadístico de prueba **T**

grados de libertad

10.9 INFERENCIAS PARA LA DIFERENCIA ENTRE DOS PROPORCIONES

CASO: Muestras grandes

Esta inferencia se utiliza para relacionar las proporciones entre dos poblaciones.

Sean dos poblaciones con **distribución binomial** de las cuales se toman **muestras aleatorias independientes** para usar su diferencia como una estimación de la diferencia entre las proporciones poblacionales.

Parámetro: $p_1 - p_2$ Diferencia entre proporciones poblacionales
Poblaciones con distribución binomial y parámetros p_1, p_2 desconocidos
Muestras aleatorias **independientes** de tamaños n_1 y n_2 mayores o iguales a **30**
Estimador: $\bar{p}_1 - \bar{p}_2$ Diferencia entre proporciones muestrales
en donde $\bar{p}_1 = x_1/n_1$, $\bar{p}_2 = x_2/n_2$

Media y varianza del estimador

$$\mu_{\bar{p}_1-\bar{p}_2} = E(\bar{p}_1 - \bar{p}_2) = E(\bar{p}_1) - E(\bar{p}_2) = E(x_1/n_1) - E(x_2/n_2) =$$

$$= 1/n_1E(x_1) - 1/n_2E(x_2) = (1/n_1)n_1p_1 - (1/n_2)n_2p_2 = p_1 - p_2 \quad \text{(estimador insesgado)}$$

$$\sigma^2_{\bar{p}_1-\bar{p}_2} = V(\bar{p}_1 - \bar{p}_2) = V[(1)\,\bar{p}_1 + (-1)\,\bar{p}_2] = (1)^2V(\bar{p}_1) + (-1)^2V(\bar{p}_2)$$

$$= V(x_1/n_1) + V(x_2/n_2) = \frac{1}{n_1^2}\,V(x_1) + \frac{1}{n_2^2}\,V(x_2)$$

$$= \frac{1}{n_1^2}(n_1p_1q_1) + \frac{1}{n_2^2}(n_2p_2q_2) = \frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}$$

Estadístico de Prueba

$$Z = \frac{(\bar{p}_1 - \bar{p}_2) - \mu_{\bar{p}_1-\bar{p}_2}}{\sigma_{\bar{p}_1-\bar{p}_2}} = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}}$$

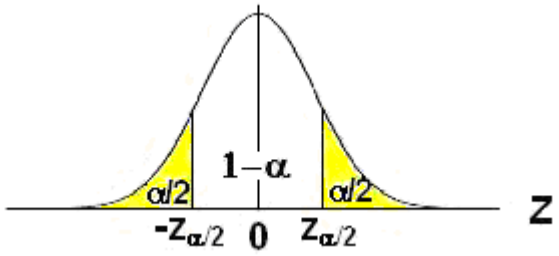
Por el Teorema del Límite Central tiene distribución normal estándar aproximadamente.

Con un criterio similar al usado anteriormente para muestras grandes, se puede aproximar la varianza poblacional mediante la varianza muestral.

$$\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2} \cong \frac{\bar{p}_1\bar{q}_1}{n_1} + \frac{\bar{p}_2\bar{q}_2}{n_2}$$

10.9.1 INTERVALO DE CONFIANZA

Con un planteamiento similar al realizado en casos anteriores para muestras grandes:



$$Z = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}}$$

Con probabilidad **1 - α**, se cumple la desigualdad: $-z_{\alpha/2} \le Z \le z_{\alpha/2}$

Sustituyendo **Z** y despejando el parámetro de interés $p_1 - p_2$ se obtiene:

Definición: Intervalo de confianza para $p_1 - p_2$ con nivel $1 - \alpha$

$$(\bar{p}_1 - \bar{p}_2) - z_{\alpha/2}\sqrt{\frac{\bar{p}_1\bar{q}_1}{n_1} + \frac{\bar{p}_2\bar{q}_2}{n_2}} \le p_1 - p_2 \le (\bar{p}_1 - \bar{p}_2) + z_{\alpha/2}\sqrt{\frac{\bar{p}_1\bar{q}_1}{n_1} + \frac{\bar{p}_2\bar{q}_2}{n_2}}$$

Ejemplo

132 de 200 electores de la región uno favorecen a un candidato, mientras que le son favorables 90 de 150 electores de la región dos. Suponiendo que las muestras son aleatorias e independientes encuentre un intervalo de confianza de 99% para la diferencia entre las proporciones de electores que le son favorables en estas dos regiones.

Solución

$1 - \alpha = 0.99 \Rightarrow z_{\alpha/2} = z_{0.005} = 2.575$

Sustituimos en la fórmula anterior: $\bar{p}_1 = x_1/n_1 = 132/200 = 0.66$, $\bar{p}_2 = x_2/n_2 = 90/150 = 0.6$

$$(0.66 - 0.6) - 2.575\sqrt{\frac{(0.66)(0.34)}{200} + \frac{(0.6)(0.4)}{150}} \le p_1 - p_2 \le$$

$$(0.66 - 0.6) + 2.575\sqrt{\frac{(0.66)(0.34)}{200} + \frac{(0.6)(0.4)}{150}}$$

$\Rightarrow \quad -0.074 \le p_1 - p_2 \le 0.194$

Con una confianza de 99%, se puede afirmar que la proporción de votantes que favorecen al candidato va de 7.74% con una proporción mayor en la región 2, hasta un valor de 19.4% en la que la proporción es mayor en la región 1.

10.9.2 PRUEBA DE HIPÓTESIS

1) Formular la hipótesis nula: **Ho: $p_1 - p_2 = d_0$** (Algún valor especificado. Usualmente: d_0=0)

2) Formular una hipótesis alterna. Elegir una entre:

Ha: $p_1 - p_2 < d_0$
Ha: $p_1 - p_2 > d_0$
Ha: $p_1 - p_2 \neq d_0$

3) Especificar el nivel de significancia α para la prueba

4) Seleccionar el estadístico de prueba y definir la región de rechazo de Ho

$$Z = \frac{(\bar{p}_1 - \bar{p}_2) - d_0}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}}$$, con distribución normal estándar aproximadamente

En donde: $$\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2} \cong \frac{\bar{p}_1 \bar{q}_1}{n_1} + \frac{\bar{p}_2 \bar{q}_2}{n_2}$$

Ha	Región de rechazo de Ho en favor de Ha
$p_1 - p_2 < d_0$	$z < -z_\alpha$
$p_1 - p_2 > d_0$	$z > z_\alpha$
$p_1 - p_2 \neq d_0$	$z < -z_{\alpha/2} \vee z > z_{\alpha/2}$

5) Con los datos de la muestra calcular el valor del estadístico

6) Si el valor del estadístico de prueba cae en la región de rechazo, la decisión es rechazar Ho en favor de Ha. Caso contrario, se dice que no hay evidencia suficiente para rechazar Ho.

10.9.3 EJERCICIOS

Un fabricante modificó el proceso de producción de sus artículos para reducir la proporción de artículos defectuosos. Para determinar si la modificación fue efectiva el fabricante tomó una muestra aleatoria de 200 artículos antes de la modificación y otra muestra aleatoria independiente, de 300 artículos después de la modificación, obteniendo respectivamente 108 y 96 artículos defectuosos.

a) Encuentre un intervalo de confianza de 98% para la diferencia entre las proporciones de artículos defectuosos en ambas poblaciones (antes y después de la modificación)

b) Realice una prueba de hipótesis de 1% de significancia para probar que la modificación realizada en el proceso de producción reduce la proporción de artículos defectuosos.

10.10 INFERENCIAS PARA DOS VARIANZAS

Parámetros: σ_1^2 , σ_2^2 (varianzas poblacionales)

Poblaciones con distribución **normal**

Estimadores: S_1^2 y S_2^2 (varianzas muestrales)

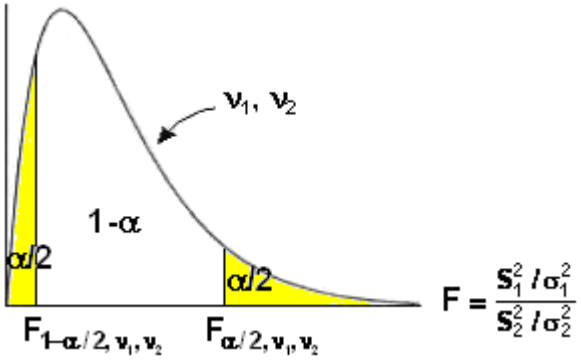
muestras aleatorias independientes de tamaño n_1 y n_2

Estadístico de prueba: $F = \dfrac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2}$

tiene distribución **F,** con $\nu_1 = n_1 - 1$, $\nu_2 = n_2 - 1$ grados de libertad

10.10.1 INTERVALO DE CONFIANZA

Se especifica un valor de probabilidad $1-\alpha$ en la distribución **F** como se muestra en el gráfico



Se tiene

$$F_{1-\alpha/2,\,\nu_1,\,\nu_2} \le F \le F_{\alpha/2,\,\nu_1,\,\nu_2} \quad \text{con probabilidad } 1-\alpha$$

Si se sustituye **F** y se despeja el parámetro de interés se obtiene

$$\frac{S_1^2}{S_2^2}\frac{1}{F_{\alpha/2,\,\upsilon_1,\,\upsilon_2}} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{S_1^2}{S_2^2}\frac{1}{F_{1-\alpha/2,\,\upsilon_1,\,\upsilon_2}}$$

Con la definición $F_{1-\alpha,\,\nu_1,\,\nu_2} = \dfrac{1}{F_{\alpha,\,\nu_2,\,\nu_1}}$ se puede escribir:

Definición: Intervalo de Confianza para σ_1^2/σ_2^2 con nivel $1-\alpha$

$$\frac{S_1^2}{S_2^2}\frac{1}{F_{\alpha/2,\,\upsilon_1,\,\upsilon_2}} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{S_1^2}{S_2^2}F_{\alpha/2,\,\upsilon_2,\,\upsilon_1}$$

Con $\nu_1 = n_1 - 1$, $\nu_2 = n_2 - 1$ grados de libertad

Ejemplo
De dos poblaciones con distribuciones normales se han tomado dos muestras aleatorias independientes y se obtuvieron:

Muestra	n	$\bar{X}$	S^2
1	10	5.9	4
2	8	7.1	5

Encuentre un intervalo para σ_1^2/σ_2^2 con un nivel de confianza de **90%**

Solución

1- $\alpha = 0.9 \Rightarrow \alpha/2 = 0.05,\ \nu_1 = 10 - 1 = 9,\ \nu_2 = 8 - 1 = 7$

$F_{\alpha/2,\nu_1,\nu_2} = F_{0.05,\ 9,\ 7} = 3.68$ (Tabla **F**)

$F_{\alpha/2,\nu_2,\nu_1} = F_{0.05,\ 7,\ 9} = 3.29$

Sustituyendo

$$\frac{4}{5}\frac{1}{3.68} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{4}{5}3.29 \Rightarrow 0.2222 \le \frac{\sigma_1^2}{\sigma_2^2} \le 2.6320$$

10.10.2 PRUEBA DE HIPÓTESIS

1) Definir la hipótesis nula **Ho:** $\sigma_1^2 = \sigma_2^2$

2) Elegir una Hipótesis alterna: **Ha:** $\sigma_1^2 < \sigma_2^2$

Ha: $\sigma_1^2 > \sigma_2^2$

Ha: $\sigma_1^2 \neq \sigma_2^2$

3) Seleccionar el nivel de significancia α

4) Estadístico de prueba. Se obtiene simplificando $\sigma_1^2 = \sigma_2^2$

$F = \frac{S_1^2}{S_2^2}$, distribución **F** con $\nu_1 = n_1 - 1, \nu_2 = n_2 - 1$ grados de libertad

Región crítica

Ha	Región de rechazo de Ho en favor de Ha
$\sigma_1^2 < \sigma_2^2$	$F < F_{1-\alpha}$
$\sigma_1^2 > \sigma_2^2$	$F > F_{\alpha}$
$\sigma_1^2 \neq \sigma_2^2$	$F < F_{1-\alpha/2} \vee F > F_{\alpha/2}$

5) Calcular el valor del estadístico de prueba con los datos de la muestra

6) Decidir

Ejemplo
De dos poblaciones con distribuciones normales se han tomado dos muestras aleatorias independientes y se obtuvieron:

Muestra	n	$\bar{X}$	S^2
1	10	5.9	4
2	8	7.1	5

Pruebe con 10% de significancia que las poblaciones tienen varianzas diferentes

Solución

1) **Ho:** $\sigma_1^2 = \sigma_2^2$

2) **Ha:** $\sigma_1^2 \neq \sigma_2^2$

3) α = **0.1**

4) Estadístico de prueba

F = $\frac{S_1^2}{S_2^2}$, distribución **F** con $\nu_1 = n_1 - 1, \nu_2 = n_2 - 1$ grados de libertad

Región crítica

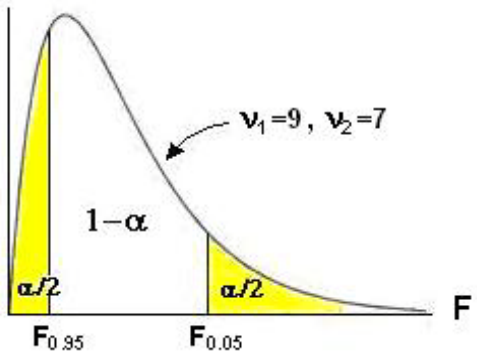
$\alpha = 0.1 \Rightarrow \alpha/2 = 0.05,\ \nu_1 = 10 - 1 = 9,\ \nu_2 = 8 - 1 = 7$

$F_{\alpha/2,\,\nu_1,\,\nu_2} = F_{0.05,\,9,\,7} = 3.68$ (Tabla **F**)

$$F_{1-\alpha/2,\,\nu_1,\,\nu_2} = F_{0.95,\,9,\,7} = \frac{1}{F_{\alpha/2,\,\nu_2,\,\nu_1}} = \frac{1}{F_{0.05,\,7,\,9}} = \frac{1}{3.29} = 0.304$$

Región de rechazo de Ho en favor de Ha

F < 0.304 ∨ F > 3.68



5) Cálculo del estadístico de prueba

F = $\frac{S_1^2}{S_2^2}$ = **4/5** = **0.8**

6) Decisión: No hay evidencia suficiente en la muestra para rechazar la hipótesis que las varianzas poblacionales son iguales

10.10.3 EJERCICIOS

Las siguientes son las calificaciones obtenidas en el examen final de una materia por dos grupos de 8 mujeres y 8 hombres:

Hombres	55	68	70	66	91	78	81
Mujeres	73	65	74	80	76	63	82

Suponiendo que los datos pueden considerarse como muestras aleatorias independientes tomadas de poblaciones con distribución normal, pruebe con 5% de significancia que la varianza de las calificaciones de los hombres es mayor a la de las mujeres.

MATLAB

```
>> alfa=0.1;
>> F1=finv(1-alfa/2,9,7)                Valores de la distribución F
   F1 =
        3.6767
>> F2=finv(1-alfa/2,7,9)
   F2 =
        3.2927
>> IC = [4/5*1/F1, 4/5*F2]              Intervalo de confianza
   IC =
        0.2176    2.6342
```

10.11 PRUEBA PARA LA DIFERENCIA DE MEDIAS CON MUESTRAS PAREADAS

Esta prueba permite comparar las medias de dos poblaciones usando dos muestras aleatorias que **no son independientes**. Esto significa que las observaciones de una muestra influyen en los resultados de la otra.

Suponga que se quiere conocer la opinión acerca de la calidad de dos marcas de cierto producto. Si se eligiera una muestra aleatoria del producto de la una marca y se la probara con un grupo de personas, y se eligiera una muestra aleatoria del producto de la otra marca y se las probara con otro grupo de personas, entonces las muestras serían independientes.

Pero, si se las muestras aleatorias de las dos marcas del producto se las probase con el mismo grupo de personas, entonces los resultados obtenidos ya no son independientes pues la opinión de cada persona respecto a la una marca, afecta a su opinión acerca de la otra marca. Este es un caso de muestras pareadas.

Supongamos dos poblaciones acerca de las cuales es de de interés estimar el valor de la diferencia entre estas medias poblacionales. De estas poblaciones se toman muestras aleatorias pareadas. Al no ser muestras independientes, no se puede usar como estimador la diferencia de las medias muestrales, siendo necesario definir otro estadístico.

$\mu_1 - \mu_2$: Parámetro de interés

n: Tamaño de la muestra pareada

X_1**:** Observaciones obtenidas en la muestra tomada de la población 1

X_2**:** Observaciones obtenidas en la muestra tomada de la población 2

$D_i = X_{1,i} - X_{2,i}$ **, i=1, 2, ..., n**: Diferencias entre observaciones

D_i son variables aleatorias independientes.

Estimador $\bar{D}$: media de las diferencias entre las observaciones

$$\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i = \frac{1}{n}\sum_{i=1}^{n}(X_{1,i} - X_{2,i}) \quad \text{con varianza} \quad S_{\bar{D}}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2$$

$\bar{D}$ es un estimador insesgado del parámetro $\mu_1 - \mu_2$**:**

10.11.1 PRUEBA DE HIPÓTESIS

1) Ho: $\mu_1 - \mu_2 = d_0$ (algún valor especificado, por ejemplo 0)

2) Ha: $\mu_1 - \mu_2 < d_0$
$\mu_1 - \mu_2 > d_0$
$\mu_1 - \mu_2 \neq d_0$

3) α**:** nivel de significancia

4) Estadístico de prueba

Caso: $n \geq 30$

$$Z = \frac{\bar{D} - d_0}{\frac{S_{\bar{D}}}{\sqrt{n}}}$$

Con distribución aproximadamente normal estándar por el Teorema del Límite Central

Caso: $n < 30$. Suponer poblaciones con distribución normal aproximadamente

$$T = \frac{\bar{D} - d_0}{\frac{S_{\bar{D}}}{\sqrt{n}}}$$

Con distribución **T** con $\nu = n - 1$ grados de libertad

Ejemplo

Los siguientes datos corresponden a un estudio de las horas perdidas mensualmente por accidentes de trabajo en 6 fábricas **antes** y **después** de implantar un programa de seguridad industrial.

Fábrica	Antes (horas perdidas)	Después (horas perdidas)
1	45	36
2	73	60
3	46	44
4	39	29
5	17	11
6	30	32

Suponiendo que la población es normal, probar con 5% de significancia que el programa es eficaz.

Solución

Sean μ_1 media de las horas perdidas antes del programa

μ_2 media de las horas perdidas después del programa

Se desea probar que $\mu_1 > \mu_2 \Rightarrow \mu_1 - \mu_2 > 0$

1) Ho: $\mu_1 - \mu_2 = 0$

2) Ha: $\mu_1 - \mu_2 > 0$

3) $\alpha = 0.05$

4) Estadístico de prueba, $n < 30$

$$T = \frac{\bar{D} - d_0}{\frac{S_{\bar{D}}}{\sqrt{n}}}$$

Distribución **T** con $\nu = n - 1$ grados de libertad

$t_\alpha = t_{0.05} = 2.015$, con $\nu = n - 1 = 5$ grados de libertad

Región de rechazo para Ho: $t > 2.015$

5) $\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i = \frac{1}{6}[(45\text{-}36) + (73\text{-}60) + \ldots] = 6.335$

$s^2_{\bar{D}} = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2 = \frac{1}{5}[(9\text{-}5.5)^2 + (13\text{-}5.5)^2 + \ldots] = 30.6666$

$s_{\bar{D}} = \sqrt{30.6666} = 5.5377$

$t = \frac{6.335 - 0}{\frac{5.5377}{\sqrt{6}}} = 2.8022 > 2.015$

6) **Decisión:**

Se rechaza Ho en favor de Ha, es decir, con una significancia de 5% se puede afirmar que el programa si es eficaz

10.11.2 EJERCICIOS

1) Los siguientes datos corresponden a la frecuencia cardiaca de un grupo de 6 personas medida antes y después de haberse sometido a un tratamiento:

Antes: **83, 78, 91, 87, 85, 84**
Después: **76, 81, 88, 86, 83, 87**

Pruebe con 5% de significancia que este tratamiento no varia la frecuencia cardiaca de las personas que lo toman. Suponga que la población es normal

2) Se eligieron 6 trabajadores para realizar una tarea, antes y después de aplicar una nueva técnica, obteniéndose los siguientes resultados en horas:

8 y 6, 10 y 7, 8 y 8, 10 y 8, 8 y 7, 9 y 7

Con un nivel de significancia de 5% pruebe si la nueva técnica es eficaz

MATLAB

Prueba de hipótesis relacionada con muestras pareadas, n < 30

```
>> antes = [45 73 46 39 17 30];
>> despues = [36 60 44 29 11 32];
>> d=antes - despues
d =
    9    13     2    10     6    -2
>>  [h, p, ci, t] = ttest(d, 0, 0.05, 1)
h =
     1
p =
    0.0190
ci =
    1.7778       Inf
t =
    tstat: 2.8014
       df: 5
```

Datos “antes”
Datos “después”
Vector de diferencias

Prueba Ho: $\mu_1 - \mu_2 = 0$ vs. Ho: $\mu_1 - \mu_2 > 0$
$\alpha = 0.1$. Prueba unilateral derecha

h=0 $\Rightarrow$ La evidencia no es suficiente para rechazar **Ho**

Valor **p** de la prueba

Intervalo de confianza para **d**

Valor del estadístico de prueba
Grados de libertad

10.12 TABLAS DE CONTINGENCIA

Esta prueba se puede usar para **determinar la independencia** entre dos métodos o factores involucrados en la obtención de datos.

Para aplicar esta prueba se organiza una tabla, colocando en las filas y columnas los resultados obtenidos con ambos factores.

Terminología

n: Cantidad de observaciones en la muestra
r: Cantidad de filas
c: Cantidad de columnas
r_i: Total de resultados en la fila **i**
c_j: Total de resultados en la columna **j**
$n_{i,j}$: Total de resultados observados en la fila **i**, columna **j** (son los datos muestrales)
$e_{i,j}$: Total de resultados esperados en la fila **i**, columna **j** (se obtienen con la hipótesis)

Obtención de la frecuencia esperada $e_{i,j}$

Definiciones

p_i: Probabilidad que un resultado pertenezca a la fila **i**
$p_i = r_i / n$
p_j: Probabilidad que un resultado pertenezca a la columna **j**
$p_j = c_j / n$
$p_{i,j}$: Probabilidad que un resultado pertenezca a la fila **i**, columna **j**

Hipótesis que se debe probar

Que los resultados son independientes de entre filas y columnas

Ho: $p_{i,j} = p_i\, p_j$

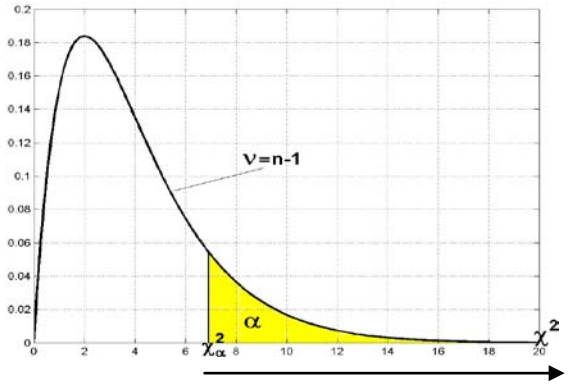
Si esta hipótesis fuese cierta se tendría que la frecuencia esperada sería

$$e_{i,j} = p_{i,j}\, n = p_i\, p_j\, n = \left(\frac{r_i}{n}\right)\left(\frac{c_j}{n}\right)n = \frac{r_i\, c_j}{n}$$

Definición: Estadístico de Prueba para Tablas de Contingencia

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{i,j} - e_{i,j})^2}{e_{i,j}},$$ tiene distribución Ji-cuadrado con ν = **(r–1)(c–1)** grados de libertad

Dado el nivel de significancia α para la prueba, si las diferencias entre la frecuencia observada $n_{i,j}$ y la frecuencia esperada $e_{i,j}$ son significativas, entonces el estadístico de prueba caerá en la región de rechazo de la hipótesis nula H_o la cual propone independencia entre resultados.



Región de rechazo de Ho

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{i,j} - e_{i,j})^2}{e_{i,j}}$$

Si $\chi^2 > \chi^2_\alpha$ se rechaza **Ho** $\Rightarrow$ Los resultados **no** son independientes entre filas y columnas

10.12.1 PRUEBA DE HIPÓTESIS

1) **Ho:** $\forall_{i,j}\ (\ p_{i,j} = p_i\ p_j\ )$ (los resultados son independientes entre filas y columnas)

2) **Ha:** $\urcorner$ **Ho** (los resultados no son independientes)

3) α**:** Nivel de significancia de la prueba

4) Con los valores de α y ν = **(r-1)(c-1)** se define la región de rechazo de **Ho**

$\chi^2 > \chi^2_\alpha$

5) Calcular el valor del estadístico de prueba

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{i,j} - e_{i,j})^2}{e_{i,j}}$$, distribución Ji-cuadrado con ν = **(r-1)(c-1)** grados de libertad

Ejemplo

Los siguientes datos corresponden a la cantidad de errores de producción de artículos en una empresa, organizados por tipo de error (**columnas 1, 2, 3, 4**) y por el equipo de obreros que los fabricó (**filas 1, 2, 3**)

	1	2	3	4
1	15	21	45	13
2	26	31	34	5
3	33	17	49	20

Pruebe con 5% de significancia que la cantidad de errores en la producción de los artículos es independiente del tipo de error y del equipo que los fabricó

Solución

Completamos el cuadro colocando en los bordes las sumas de filas r_i y la suma de columnas c_j y en la parte inferior de cada celda la frecuencia esperada $e_{i,j}$ calculada con la fórmula:

$$e_{i,j} = \frac{r_i\ c_j}{n}$$

$e_{1,1} = r_1\ c_1\ /\ n = (94)(74)/309 = 22.51$
$e_{1,2} = r_1\ c_2\ /\ n = (94)(69)/309 = 20.99$
$e_{1,3} = r_1\ c_3\ /\ n = (94)(128)/309 = 38.94$
$e_{1,4} = r_1\ c_4\ /\ n = (94)(38)/309 = 11.56$
$e_{2,1} = r_2\ c_1\ /\ n = (96)(74)/309 = 22.99$
... etc

Tabulación

	1	2	3	4	r_i
1	15 22.51	21 20.99	45 38.94	13 11.56	94
2	26 22.99	31 21.44	34 39.77	5 11.81	96
3	33 28.50	17 26.57	49 49.29	20 14.63	119
c_j	74	69	128	38	n = 309

Definimos la región de rechazo

$\alpha = 0.05,\ \nu = (r-1)(c-1) = (3)(2) = 6 \Rightarrow \chi^2_\alpha = \chi^2_{0.05} = 12.54$ (Tabla χ^2)

Rechazar **Ho** si $\chi^2 > 12.54$

Cálculo del estadístico de prueba

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{i,j}-e_{i,j})^2}{e_{i,j}} = \frac{(15-22.51)^2}{22.51} + \frac{(21-20.99)^2}{20.99} + \frac{(45-38.94)^2}{38.94} + \ldots = 19.18$$

Decisión

El valor del estadístico de prueba cae en la región de rechazo de **Ho**, por lo tanto se concluye que **no hay independencia** entre el tipo de error en los artículos producidos y el equipo de obreros que los fabricó.

10.12.2 EJERCICIOS

1) Los siguientes datos corresponden a las calificaciones en tres materias (**columnas 1, 2, 3**) obtenidas por cuatro estudiantes (**filas 1, 2, 3, 4**)

	1	**2**	**3**
1	73	68	56
2	65	70	50
3	70	73	55
4	68	71	54

Pruebe con **5%** de significancia que no hay dependencia entre las calificaciones obtenidas en las materias y los estudiantes

2) En una muestra aleatoria de 100 ciudadanos de Guayaquil, se los clasificó por su ocupación: obrero, estudiante, profesional, y se les consultó si están a favor o en contra de la integración de un organismo de justicia, propuesto por el Congreso.
Se obtuvieron los siguientes datos:

	Obrero	**Estudiante**	**Profesional**
A favor	10	16	14
En contra	12	26	22

Proponga y pruebe una hipótesis para demostrar, con 5% de significancia, que la opinión de los ciudadanos es independiente de su ocupación.

MATLAB

Prueba con tablas de contingencia

```
>> n=[15 21 45 13; 26 31 34 5; 33 17 49 20]
   n =
      15    21    45    13
      26    31    34     5
      33    17    49    20
```
Frecuencias observadas

```
>> r=sum(t)
   r =
      74    69   128    38
```
Suma de filas

```
>> c=sum(t' )
   c =
      94    96   119
```
Suma de columnas

```
>> e=(c' *(r))/(sum(sum(t)))
   e =
    22.5113   20.9903   38.9385   11.5599
    22.9903   21.4369   39.7670   11.8058
    28.4984   26.5728   49.2945   14.6343
```
Frecuencias esperadas

```
>> ji2=sum(sum((n-e).^2./e))
   ji2 =
        19.1780
```
Valor del estadístico de prueba

```
>> vc=chi2inv(0.95,6)
   vc =
        12.5916
```
Valor crítico de rechazo

Conclusión: El valor del estadístico cae en la región de rechazo de Ho

10.13 PRUEBAS DE BONDAD DE AJUSTE

Estas pruebas permiten verificar que la población de la cual proviene una muestra tiene una distribución especificada o supuesta.

Sean **X:** Variable aleatoria poblacional
$f_0(x)$: Distribución (o densidad) de probabilidad especificada o supuesta para **X**

Se desea probar la hipótesis: **Ho:** $f(x) = f_0(x)$

En contraste con la hipótesis alterna: **Ha:** $\neg H_0$ (negación de **Ho**)

10.13.1 PRUEBA JI-CUADRADO

Esta prueba es aplicable para variables aleatorias discretas o continuas

Sea una muestra aleatoria de tamaño **n** tomada de una población con una distribución especificada $f_0(x)$ que es de interés verificar.

Suponer que las observaciones de la muestra están agrupadas en **k** clases, siendo n_i la cantidad de observaciones en cada clase **i** = **1, 2, ..., k**

Con el modelo especificado $f_0(x)$ se puede calcular la probabilidad p_i que un dato cualquiera pertenezca a una clase **i.**

Con este valor de probabilidad se puede encontrar la frecuencia esperada e_i para la clase **i,** es decir, la cantidad de datos que según el modelo propuesto deberían estar incluidos en la clase **i:**

$$e_i = p_i\, n, \qquad i = 1, 2, ..., k$$

Tenemos entonces dos valores de frecuencia para cada clase **i**

n_i: frecuencia observada (corresponde a los datos de la muestra)
e_i: frecuencia esperada (corresponde al modelo propuesto)

La teoría estadística demuestra que la siguiente variable es apropiada para realizar una prueba de bondad de ajuste:

Definición: Estadístico para la Prueba de Bondad de Ajuste Ji-Cuadrado

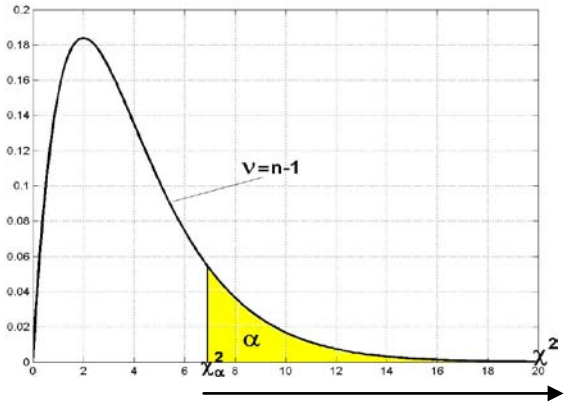
$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i}$$, distribución **Ji-cuadrado** con $\nu = k-1$ grados de libertad

Una condición necesaria para aplicar esta prueba es que: $\forall i (e_i \geq 5)$

Dado el nivel de significancia α se define el valor crítico χ^2_α para el rechazo de la hipótesis propuesta **Ho: f(x)** = $f_0(x)$.

Si las frecuencias observadas no difieren significativamente de las frecuencias esperadas calculadas con el modelo propuesto, entonces el valor de estadístico de prueba χ^2 será cercano a cero. Pero si estas diferencias son significativas, entonces el valor del estadístico χ^2 estará en la región de rechazo de **Ho**:

$$\chi^2 > \chi^2_\alpha$$



Región de rechazo de Ho

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i}$$

Ejemplo.
Se ha tomado una muestra aleatoria de **40** baterías y se ha registrado su duración en años. Estos resultados se los ha agrupado en **7** clases, como se muestra en el siguiente cuadro

i	Clase (duración)	Frecuencia observada (n_i)
1	1.45 – 1.95	2
2	1.95 – 2.45	1
3	2.45 – 2.95	4
4	2.95 – 3.45	15
5	3.45 – 3.95	10
6	3.95 – 4.45	5
7	4.45 – 4.95	3

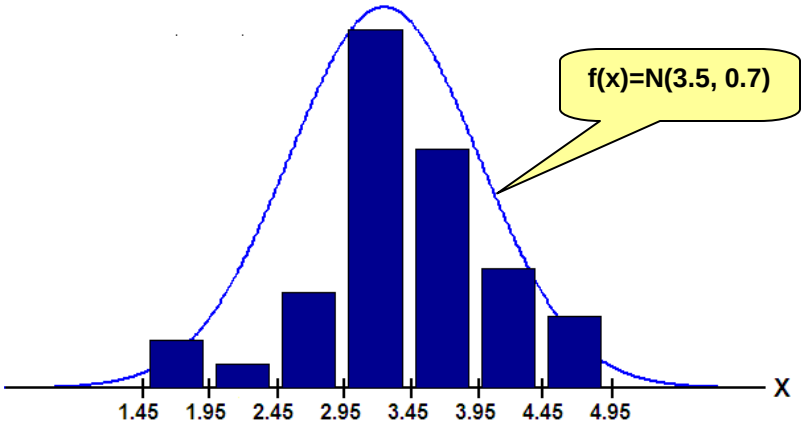
Verificar con 5% de significancia que la duración en años de las baterías producidas por este fabricante tiene duración distribuida normalmente con media **3.5** y desviación estándar **0.7**

Solución

Sea **X**: duración en años (variable aleatoria contínua)

1) **Ho:** **f(x) = N(3.5, 0.7)** (Distribución normal, μ = **3.5**, σ = **0.7**)
2) **Ha:** $\neg H_0$
3) α = **0.05**

Cálculo de la probabilidad correspondiente a cada intervalo con el modelo propuesto



$\mathbf{p_1} = P(X \le 1.95) = \mathbf{P\left(Z \le \frac{1.95 - 3.5}{0.7}\right)} = 0.0136$

$\mathbf{p_2} = P(1.95 \le X \le 2.45) = \mathbf{P\left(\frac{1.95 - 3.5}{0.7} \le Z \le \frac{2.45 - 3.5}{0.7}\right)} = 0.0532$

$\mathbf{p_3} = P(2.45 \le X \le 2.95) = \mathbf{P\left(\frac{2.45 - 3.5}{0.7} \le Z \le \frac{2.95 - 3.5}{0.7}\right)} = 0.135$

... (etc)

Cálculo de las frecuencias esperadas

$\mathbf{e_1} = p_1\, n = 0.0136\,(40) \approx 0.5$

$\mathbf{e_2} = p_2\, n = 0.0532\,(40) \approx 2.1$

$\mathbf{e_3} = p_3\, n = 0.135\,(40) \approx 5.4$

... (etc)

Resumen de resultados

Duración (años)	Frecuencia observada (n_i)	Frecuencia esperada (e_i)
1.45 – 1.95	2	0.5
1.95 – 2.45	1	2.1
2.45 – 2.95	4	5.4
2.95 – 3.45	15	10.3
3.45 – 3.95	10	10.7
3.95 – 4.45	5	7
4.45 – 4.95	3	3.5

Es necesario que se cumpla la condición $\forall \mathbf{i}(\mathbf{e_i} \ge \mathbf{5})$ por lo que se deben agrupar clases adyacentes. Como resultado se tienen cuatro clases: $\mathbf{k = 4}$

Duración (años)	Frecuencia observada (n_i)	Frecuencia esperada (e_i)
1.45 – 2.95	7	8.5
2.95 – 3.45	15	10.3
3.45 – 3.95	10	10.7
3.95 – 4.95	8	10.5

Ahora se puede definir la región de rechazo de Ho

$\alpha = 0.05, \nu = k - 1 = 3, \Rightarrow \chi^2_{0.05} = 7.815$ (Tabla χ^2)

Rechazar Ho si $\chi^2 > 7.815$

5) Cálculo del estadístico de prueba

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i} = \left[\frac{(7-8.5)^2}{8.5} + \frac{(15-10.3)^2}{10.3} + \frac{(10-10.7)^2}{10.7} + \frac{(8-10.5)^2}{10.5} \right] = 3.05$$

6) Decisión

Como **3.05** no es mayor a **7.815**, se dice que no hay evidencia suficiente para rechazar el modelo propuesto para la población.

NOTA IMPORTANTE: En general, si no se especifican los parámetros para el modelo propuesto, pueden estimarse con los datos de la muestra.

10.13.2 EJERCICIOS

1) El siguiente cuadro muestra el registro del tiempo en horas que duran encendidos hasta que fallan una muestra de 200 focos de cierta marca

Tiempo en horas	Cantidad de focos
0 – 250	82
250 – 500	45
500 – 750	34
750 – 1000	15
1000 – 1250	10
1250 – 1500	6
1500 – 1750	4
1750 – 2000	3
2000 – 2250	1

Con 10% de significancia verifique la hipótesis que el tiempo de duración de los focos tiene distribución exponencial.

Debido a que no se especifica el parámetro del modelo propuesto, debe estimarlo a partir de los datos de la muestra (calcule la media muestral con la fórmula para datos agrupados)

MATLAB

Colocar la densidad normal sobre el histograma de la muestra

Datos de la muestra

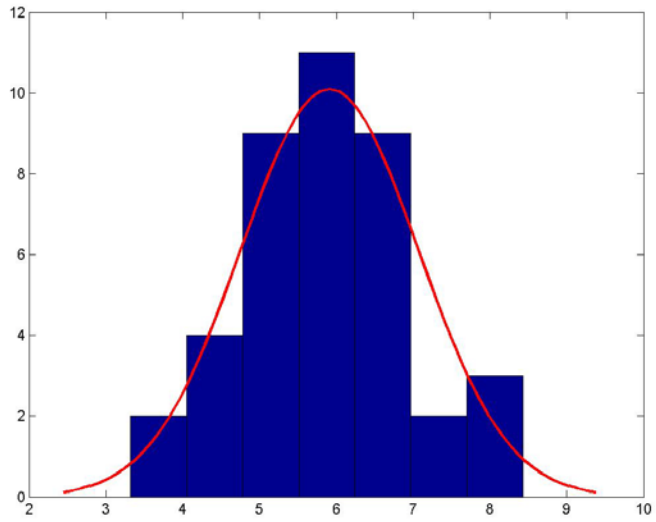
```
>> x = [ 5.73  5.01  6.89  8.28  5.43  5.01  5.85  7.12  5.00  4.51  6.03  6.10  6.87 ...
         5.36  5.99  5.59  6.08  8.34  5.35  4.31  6.85  4.93  6.25  5.32  6.94  6.97 ...
         5.91  3.32  6.38  8.43  7.62  3.98  6.08  5.24  4.76 4.47 6.60  5.59  6.27  5.68];
```

Tabulación de frecuencia en siete clases

```
>> f = hist(x,7)
    f =
        2        4        9       11        9        2        3
```

Graficar el histograma y la distribución normal

```
>> histfit(x, 7)
```



10.13.3 PRUEBA DE KOLMOGOROV - SMIRNOV (K-S)

Esta prueba se usa para probar modelos de probabilidad con variables aleatorias continuas. Es de especial interés para muestras pequeñas. Si la prueba se usa con variables aleatorias discretas, la decisión tiene confianza aceptable cuando se rechaza la hipótesis nula.

Sea **X:** Variable aleatoria continua
$f_0(x)$: Función de densidad de probabilidad especificada o supuesta para **X**

Se desea probar la hipótesis: **Ho:** $f(x) = f_0(x)$

En contraste con la hipótesis alterna: **Ha:** $\neg H_0$ (Negación de Ho)

Sea una muestra aleatoria de tamaño **n** tomada de una población con una distribución especificada $f_0(x)$ que es de interés verificar:

$X_1, X_2, \ldots, X_n$

Las observaciones se las ordenadas en forma creciente:

$X_{(1)}, X_{(2)}, \ldots, X_{(n)}$

Con los valores de **x** se obtienen valores de la siguiente función

Definición: Función de Distribución Empírica de la Muestra

$$S_n(x) = \begin{cases} 0, & x < x_{(1)} \\ i/n, & x_{(i)} \le x < x_{(i+1)} \\ 1, & x \ge x_{(n)} \end{cases}, \qquad i=1,2,\ldots, n-1$$

Sea $F_0(x)$ la función de distribución acumulada correspondiente al modelo propuesto $f_0(x)$:

$F_0(x) = P(X \le x)$

Con los valores de **x** se obtienen valores de la función $F_0(x)$.

Se tabulan los valores calculados de $S_n(x)$ y $F_0(x)$. Entonces se utiliza el estadístico para esta prueba definido de la siguiente forma:

Definición: Estadístico de prueba K-S (Kolmogorov-Smirnov)

$$D_n = \max |S_n(x_i) - F_0(x_i)| \text{ , } i=1, 2, \ldots, n$$

Si se especifica el nivel de significancia α se puede construir la región de rechazo para la prueba

Región de rechazo de Ho

Sea: D_α valor crítico para la prueba K-S

Rechazar **Ho** si $D_n > D_\alpha$

Algunos valores para el estadístico **D** están registrados en la Tabla K-S que se incluye al final de este documento. Si no se especifica α se puede expresar la decisión mediante el valor de significancia obtenido con los datos de la muestra.

Ejemplo

Suponga los siguientes datos obtenidos en una muestra aleatoria:

7.2, 7.5, 8.1 9.6, 9.1, 8.1, 7.6, 6.8

Pruebe con 5% de significancia que provienen de una población con distribución normal, con media 8 y varianza 1: **$X \sim N(8, 1)$**

Solución

Ho: **$f(x) = N(8, 1)$** (Hipótesis que interesa probar)

Ha: $\neg H_0$

$\alpha = 0.05$

Región de rechazo de Ho

$\alpha = 0.05,\ n = 8 \Rightarrow D_{0.05} = 0.457$ (Tabla K-S)

Rechazar **Ho** si **$D_n > 0.457$**

Valores de la Distribución Empírica:

$$S_n(x) = \begin{cases} 0, & x < 6.8 \\ 1/8, & 6.8 \le x < 7.2 \\ 2/8, & 7.2 \le x < 7.5 \\ 3/8, & 7.5 \le x < 7.6 \\ 4/8, & 7.6 \le x < 8.1 \\ 6/8, & 8.1 \le x < 9.1 \\ 7/8, & 9.1 \le x < 9.6 \\ 1, & x \ge 9.6 \end{cases}$$

Cálculo de los valores de $F_0(X)$ según el modelo propuesto

$$F_0(X) = P(X \le X) = P\left(Z \le \frac{x-8}{1}\right)$$ (Distribución Normal Estándar acumulada)

$$F_0(6,8) = P(X \le 6.8) = P\left(Z \le \frac{6.8-8}{1}\right) = F(-1.2) = 0.1151$$ (Tabla Z)

$$F_0(7.2) = P(X \le 7.2) = P\left(Z \le \frac{7.2-8}{1}\right) = F(-0.8) = 0.2119$$

... etc.

Tabulación de los resultados y obtención de **Dn**

x	$S_n(x)$	$F_0(x)$	$\lvert S_n(x) - F_0(x) \rvert$
6.8	1/8	0.1151	0.0099
7.2	2/8	0.2119	0.0381
7.5	3/8	0.3085	0.0665
7.6	4/8	0.3446	0.1554
8.1	6/8	0.5398	**0.2102**
9.1	7/8	0.8643	0.0107
9.6	1	0.9452	0.0548

Valor del estadístico de prueba

$D_n = \max |S_n(x_i) - F_0(x_i)|$, $i=1, 2, ..., n$

$D_n = 0.2102$

Decisión

D_n no cae en la región de rechazo, por lo tanto los datos de la muestra no proporcionan evidencia suficiente para rechazar el modelo propuesto para la población

10.13.4 EJERCICIOS

1) El fabricante de un artículo afirma que la resistencia media de su producto tiene distribución normal con media **4.5** y con desviación estándar de **0.7**. Una muestra aleatoria **6** observaciones produjo los siguientes resultados: **5.2 4.3 3.7 3.9 5.4 4.9**

Realice la prueba de bondad de ajuste **K-S**, con **5%** de significancia para determinar si los datos obtenidos en la muestra provienen de la población especificada.

2) La siguiente es una muestra del tiempo en horas que funciona un dispositivo electrónico de control hasta que se presenta una falla y recibe mantenimiento:

199.4 73.2 40.5 39.2 36.0 24.9 13.5 9.8 5.7 2.5

Realice la prueba de bondad de ajuste **K-S**, con **5%** de significancia para determinar si los datos obtenidos en la muestra provienen de una población con distribución exponencial.

MATLAB

Prueba de bondad de ajuste K - S

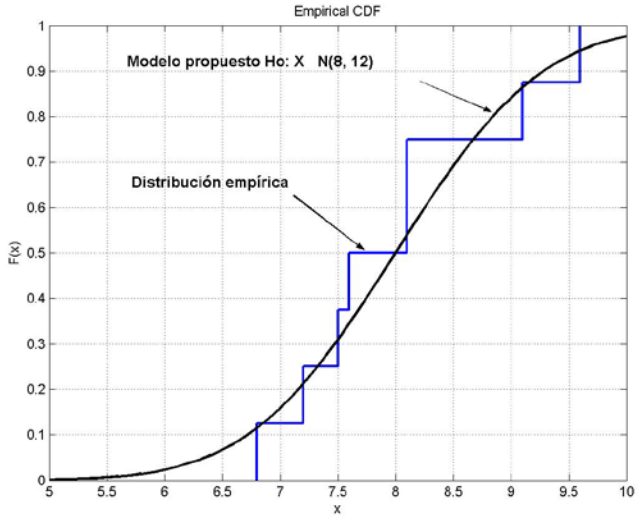
```
>> x=[7.2 7.5 8.1 9.6 9.1 8.1 7.6 6.8];
```
Vector con los datos de una muestra

```
>> cdfplot(x)
```
Gráfico de la distribución empírica acumulada

```
>> z=5: 0.1: 10;
```
Puntos para la distribución normal acumulada

```
>> f=normcdf(z, 8, 1);
```
Valores de la distribución normal acumulada con el modelo propuesto **Ho: X ~ N(8, 1^2)**

```
>> hold on, plot(z, f, 'k')
```
Superponer el gráfico del modelo propuesto




```
>> x = sort(x)
```
Ordenamiento de los datos de la muestra

```
   x =
    6.8000   7.2000   7.5000   7.6000   8.1000   8.1000   9.1000   9.6000
```

```
>> sn = 1/8: 1/8: 1
```
Distribución acumulada empírica

```
  sn =
    0.1250   0.2500   0.3750   0.5000   0.6250   0.7500   0.8750   1.0000
```

```
>> f = normcdf(x,8,1)
```
Distribución acumulada normal **Ho: X ~ N(8, 1^2)**

```
  f =
    0.1151   0.2119   0.3085   0.3446   0.5398   0.5398   0.8643   0.9452
```

```
>> dn = max(sn - f)
```
Valor del estadístico **Dn**: la mayor diferencia

```
   dn =
    0.2102
```

Prueba de bondad de ajuste usando directamente una función especializada de MATLAB

```
>> x=[7.2 7.5 8.1 9.6 9.1 8.1 7.6 6.8];
```
Vector con datos de la muestra

```
>> x=sort(x);
```
Datos ordenados

```
>> f=normcdf(x,8,1);
```
Valores con el modelo propuesto: **Ho: X ~ N(8, 1^2)**

```
>> [h,p,ksstat,vc]=kstest(x,[x' f' ], 0.05,0)
```
Prueba de bondad de ajusta K-S
x' f' son dos columnas con el modelo

```
   h = 0
```
h=0: No se rechaza el modelo

```
   p = 0.8254
```
Valor **p** de la prueba

```
   ksstat = 0.2102
```
Valor del estadístico de prueba

```
   vc =   0.4543
```
Valor crítico para la región de rechazo

10.14 ANÁLISIS DE VARIANZA

Esta prueba se utiliza para determinar si las medias muestrales provienen de poblaciones con medias iguales, cuando hay más de dos poblaciones en estudio.

El análisis de varianza (**ANOVA**) permite comparar simultáneamente todas las medias, evitando tener que realizar pruebas en grupos de dos con las técnicas vistas anteriormente.

La comparación de las medias muestrales se basa en las varianzas muestrales

Suposiciones necesarias para el análisis de varianza

1) Las poblaciones tienen distribución normal
2) Las poblaciones tienen varianzas iguales
3) Las muestras son independientes

Definiciones:

Tratamiento: Es la fuente de datos cuya variación proporciona las observaciones.

Sean. $\mathbf{k}$: Número de tratamientos

$\mathbf{n}$: Número total de observaciones en todos los tratamientos combinados

$\mathbf{n_j}$: Número total de observaciones en cada tratamiento **j = 1, 2, ..., k**

$\mathbf{X_{i,j}}$: Es la i-esima observación del tratamiento **j**

$\overline{\mathbf{X}}_\mathbf{j}$: Media muestral del tratamiento **j** (incluye las observaciones de cada tratamiento)

$\overline{\overline{\mathbf{X}}}$: Media muestral general (incluye a todas las observaciones de todos los tratamientos)

Variación Total: Es la variación total combinada de las observaciones de todos los tratamientos con respecto a la media general

Media muestral general: $$\overline{\overline{\mathbf{X}}} = \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{j=1}}^{\mathbf{k}} \sum_{\mathbf{i=1}}^{\mathbf{n_j}} \mathbf{X_{i,j}}$$

Variación total: $$\mathbf{SCT} = \sum_{\mathbf{j=1}}^{\mathbf{k}} \sum_{\mathbf{i=1}}^{\mathbf{n_j}} \mathbf{(X_{i,j} - \overline{\overline{X}})^2}$$ (Suma cuadrática total)

Variación de tratamientos: Es la variación atribuida a los efectos de los tratamientos

Media muestral del tratamiento **j**: $$\overline{\mathbf{X}}_\mathbf{j} = \frac{\mathbf{1}}{\mathbf{n_j}} \sum_{\mathbf{i=1}}^{\mathbf{n_j}} \mathbf{X_{i,j}}$$

Variación de tratamientos: $$\mathbf{SCTr} = \sum_{\mathbf{j=1}}^{\mathbf{k}} \mathbf{n_j (\overline{X}_j - \overline{\overline{X}})^2}$$ (Suma cuadrática de tratamientos)

Variación aleatoria o error: Es la variación dentro de cada tratamiento debido a errores en el experimento.

Variación aleatoria o error: **SCE = SCT – SCTr** (Suma cuadrática del error)

La ecuación **SCT = SCTr + SCE** separa la variación total en dos componentes: el primero corresponde a la variación atribuida a los tratamientos y el segundo es la variación atribuida a la aleatoriedad o errores del experimento

SCTr tiene **k – 1** grados de libertad (varianza ponderada con k tratamientos)
SCE tiene **n – k** grados de libertad (existen n datos y k tratamientos)
SCT tiene **n – 1** grados de libertad (suma de grados de libertad de **SCTr** y **SCE**)

Si cada uno se divide por el número de grados de libertad se obtienen los **cuadrados medios**

Todos estos resultados se los ordena en un cuadro denominado **tabla de análisis de varianza**

10.14.1 TABLA ANOVA (ANÁLISIS DE VARIANZA)

Fuente de variación	Grados de libertad	Suma de cuadrados	Cuadrados medios	F_0
Tratamiento	**k – 1**	**SCTr**	**SCTr/(k – 1)**	**(SCTr/(k – 1))/(SCE/(n–k))**
Error	**n – k**	**SCE**	**SCE/(n – k)**	
Total	**n – 1**	**SCT**		

El último cociente es el valor de una variable que tiene distribución **F**. Este estadístico se usa para la prueba de hipótesis

10.14.2 PRUEBA DE HIPÓTESIS

1) Hipotesis nula **Ho:** $\mu_1 = \mu_2 = \ldots = \mu_k$ (las medias poblacionales son iguales)

2) Hipótesis alterna: **Ha:** $\neg$ **Ho** (al menos dos medias son iguales)

3) Definir el nivel de significancia de la prueba α

4) Elegir el estadístico de prueba: Distribución **F** con $\nu_1 = k - 1$, $\nu_2 = n - k$ g. l.
Definir la región de rechazo de **Ho**

5) Calcular **Fo**

6) Decidir

Ejemplo

Para comparar las calificaciones promedio que obtienen los estudiantes en cierta materia que la imparten cuatro profesores, se eligieron 32 estudiantes que deben tomar esta materia y se los distribuyó aleatoriamente en los cuatro paralelos asignados a los cuatro profesores.

Al finalizar el semestre los 32 estudiantes obtuvieron las siguientes calificaciones

Profesor A	Profesor B	Profesor C	Profesor D
68	80	87	56
90	73	82	80
67	68	92	71
85	67	72	91
86	49	45	80
53	67	74	56
64	63	85	67
71	60	93	53

Con una significancia de 5% determine si existe evidencia de que hay diferencia en las calificaciones promedio entre los cuatro paralelos.

1) Hipotesis nula **Ho:** $\mu_1 = \mu_2 = \mu_3 = \mu_4$ (Las 4 medias de las notas son iguales)

2) Hipótesis alterna: **Ha:** $\neg$ **Ho** (Al menos en dos paralelos son diferentes)

3) Nivel de significancia $\alpha = 0.05$

4) Estadístico de prueba

F con $\nu_1 = 4 - 1 = 3$, $\nu_2 = 32 - 4 = 28$ g. l.

Región de rechazo

$F_{\alpha,\nu_1,\nu_2} = F_{0.05,\,3,\,28} = 2.95$ (tabla F)

Rechazar **Ho** si **Fo > 2.95**

5) Calcular **Fo**

$$\bar{\bar{X}} = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j} X_{i,j} = \frac{1}{32}\sum_{j=1}^{4}\sum_{i=1}^{n_j} X_{i,j} = \frac{1}{32}(68 + 90 + \ldots + 67 + 53) = 71.7188$$

$$SCT = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{i,j} - \bar{\bar{X}})^2 = (68 - 71.7188)^2 + (90 - 71.7188)^2 + \ldots = 5494.5$$

$$\bar{X}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{i,1} = \frac{1}{8}(68 + 90 + ... + 64 + 71) = 73$$

$$\bar{X}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} X_{i,2} = \frac{1}{8}(80 + 73 + ... + 63 + 60) = 65.875$$

$$\bar{X}_3 = \frac{1}{n_3}\sum_{i=1}^{n_3} X_{i,3} = \frac{1}{8}(87 + 82 + ... + 85 + 93) = 78.75$$

$$\bar{X}_4 = \frac{1}{n_4}\sum_{i=1}^{n_4} X_{i,4} = \frac{1}{8}(56 + 80 + ... + 67 + 53) = 69.25$$

$$SCTr = \sum_{j=1}^{k} n_j(\bar{X}_j - \bar{\bar{X}})^2 = 8(73 - 71.7188)^2 + 8(65.875 - 71.7188)^2 + ... = 730.6$$

$$SCE = SCT - SCTr = 5494.5 - 730.6 = 4763.9$$

$$Fo = \frac{\frac{SCTr}{k-1}}{\frac{SCE}{n-k}} = \frac{\frac{730.6}{3}}{\frac{4763.9}{28}} = 1.4314$$

6) **Decisión: Fo** no cae en la región de rechazo. Por lo tanto no se puede rechazar la hipótesis de que las medias de las calificaciones de los cuatro paralelos son iguales

10.14.3 EJERCICIOS

Para comparar la efectividad de cuatro tipos de fertilizantes para cierto tipo de producto, se dividió una zona de cultivo en veinte parcelas de igual tamaño y se administraron cada uno de los fertilizantes en cinco parcelas elegidas aleatoriamente.

Al finalizar el periodo de cultivo se registraron las cantidades del producto obtenidas en las parcelas asignadas a cada tipo de fertilizante con los siguientes resultados, en las unidades de medida que corresponda:

Fertilizante A	Fertilizante B	Fertilizante C	Fertilizante D
27	26	24	23
21	23	26	27
24	20	27	26
23	26	22	23
28	23	24	25

Con una significancia de 5% determine si existe evidencia de que hay diferencia en las cantidades promedio del producto que se obtuvieron con los cuatro tipos de fertilizante.

MATLAB

Análisis de varianza

Definición de la matriz de datos. Cada columna es un tratamiento (compare con el ejemplo)

```
>> notas=[ 68 80 87 56; 90 73 82 80; 67 68 92 71;85 67 72 91; ...
          86 49 45 80; 53 67 74 56; 64 63 85 67;71 60 93 53]
notas =
    68   80   87   56
    90   73   82   80
    67   68   92   71
    85   67   72   91
    86   49   45   80
    53   67   74   56
    64   63   85   67
    71   60   93   53
```

```
>> [p, tabla, stats] =anova1(notas, {'A','B','C','D'})
```

Análisis de varianza con rótulos

```
p =
   0.2546
```

Valor **p** de la prueba con **F**

```
tabla =
 'Source'     'SS'             'df'    'MS'          'F'     'Prob>F'
 'Columns'  [  730.5938]    [ 3]  [243.5313]  [1.4314]  [0.2546]
 'Error'    [4.7639e+003]  [28]  [170.1384]      []        []
 'Total'    [5.4945e+003]  [31]      []          []        []
```

Tabla ANOVA

```
stats =
  means:   [73 65.8750 78.7500 69.2500]
     df:  28
      s: 13.0437
```

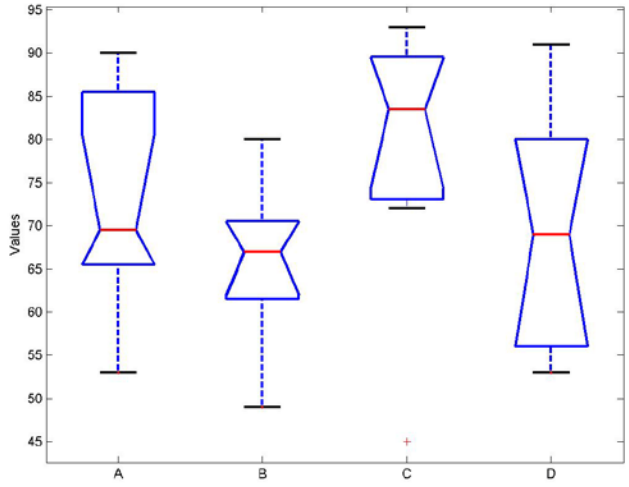
Medias de los tratamientos
Grados de libertad
Error estándar

Adicionalmente MATLAB muestra la **tabla ANOVA** en un formato estándar

ANOVA Table

Source	SS	df	MS	F	Prob>F
Columns	730.59	3	243.531	1.43	0.2546
Error	4763.88	28	170.138		
Total	5494.47	31			

MATLAB también proporciona los **diagramas de caja** de los tratamientos



11 REGRESIÓN LINEAL SIMPLE

El propósito de este estudio es proporcionar los conceptos y técnicas para construir modelos matemáticos que describan de manera apropiada a un conjunto de datos, cuando la relación es de tipo lineal. Estos modelos son útiles para realizar pronósticos.

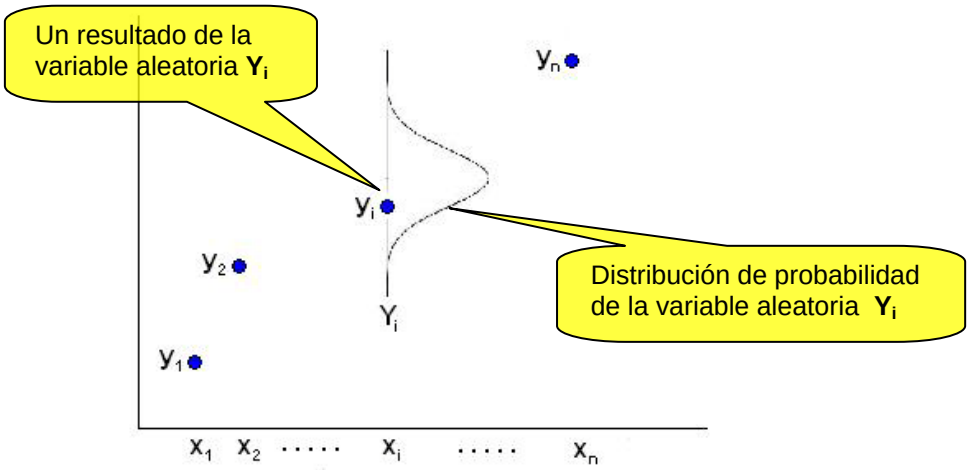
Este estudio se denomina **análisis de regresión** y el objetivo es estimar la **ecuación de regresión** la cual es la recta teórica poblacional (desconocida) de la cual provienen los datos.

Suponer que se tiene un conjunto de **n** mediciones u observaciones $(x_1, y_1), (x_2, y_2),...,(x_n, y_n)$

Estas observaciones provienen de las variables **X** y **Y**. La variable **X** se denomina **variable de predicción** mientras que la variable **Y** se denomina **variable de respuesta.**

Se supondrá que existe una correspondencia de **X** a **Y** y el objetivo es modelar esta relación.

Cada valor y_i es una observación o el resultado de una medición, por lo tanto pudiesen haber otros valores y_i para el mismo valor de x_i. Esto permite entender que y_i es uno de los posibles resultados de la variable aleatoria Y_i. Una variable aleatoria debe tener una distribución de probabilidad. El siguiente gráfico permite visualizar esta suposición:



Si la relación entre **X** y **Y** tiene "tendencia lineal", lo cual puede reconocerse graficando los puntos en una representación que se denomina **gráfico de dispersión,** entonces es razonable proponer un modelo lineal para describir la relación y que tome en cuenta la aleatoriedad de **Y**

Definición: Modelo de regresión lineal probabilista (modelo poblacional desconocido)

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

En donde β_0 y β_1 son los parámetros del modelo y ε es el componente aleatorio de **Y**

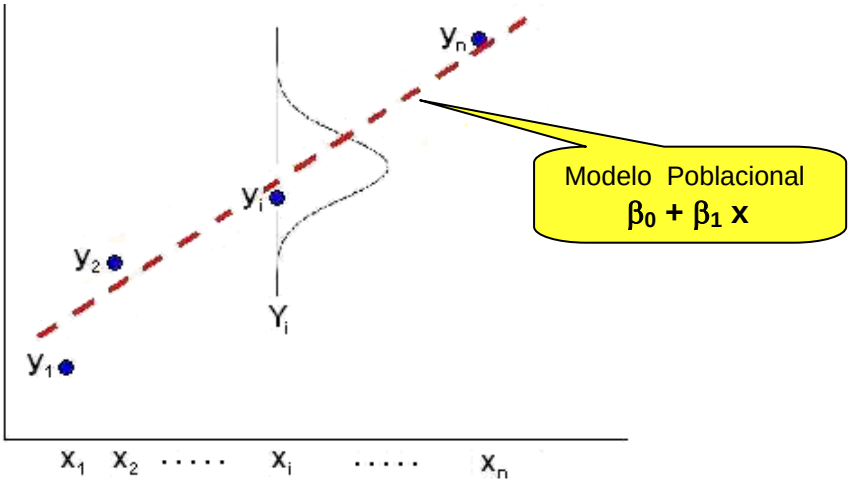
Se supondrá que para cada variable aleatoria Y_i el componente aleatorio ε_i tiene la misma distribución de probabilidad y que además estos componentes son variables independientes:

$\varepsilon_i \sim N(0, \sigma^2)$ (distribución normal con media **0** y varianza desconocida σ^2)

Con este planteamiento, el valor esperado de este modelo constituye la recta teórica que describe al **modelo poblacional** desconocido.

$$E[Y] = \beta_0 + \beta_1 x$$

El modelo poblacional teórico tiene dos parámetros β_0 (intercepción) y β_1 (pendiente)



Para comprensión de conceptos se desarrolla paralelamente un ejemplo

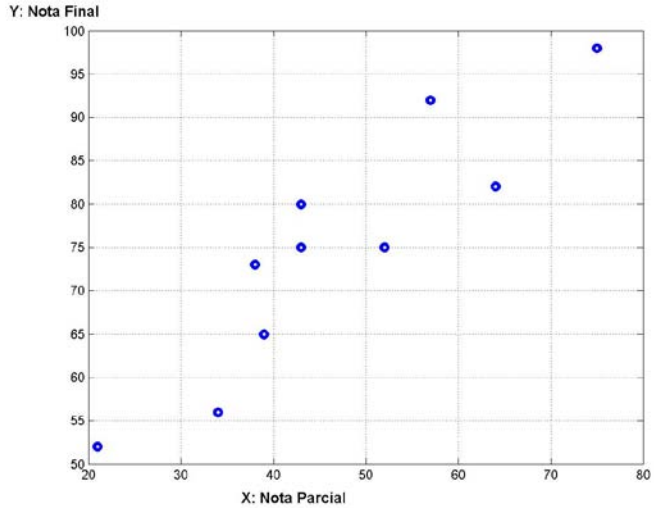
Ejemplo

Se desea construir un modelo de regresión para relacionar las calificaciones parcial y final en cierta materia, utilizando una muestra aleatoria de 10 estudiantes que han tomado esta materia:

Estudiante	1	2	3	4	5	6	7	8	9	10
Nota Parcial	39	43	21	64	57	43	38	75	34	52
Nota Final	65	75	52	82	92	80	73	98	56	75

Diagrama de dispersión

X: calificación parcial
Y: calificación final



Se observa que al incrementar **x** (variable de predicción) también se incrementa **y** (respuesta) con una tendencia aproximadamente lineal

Modelo de regresión lineal poblacional propuesto

$$Y = \beta_0 + \beta_1 x + \varepsilon, \quad \varepsilon_i \sim N(0, \sigma^2), \text{ para cada } Y_i$$

11.1 RECTA DE MÍNIMOS CUADRADOS

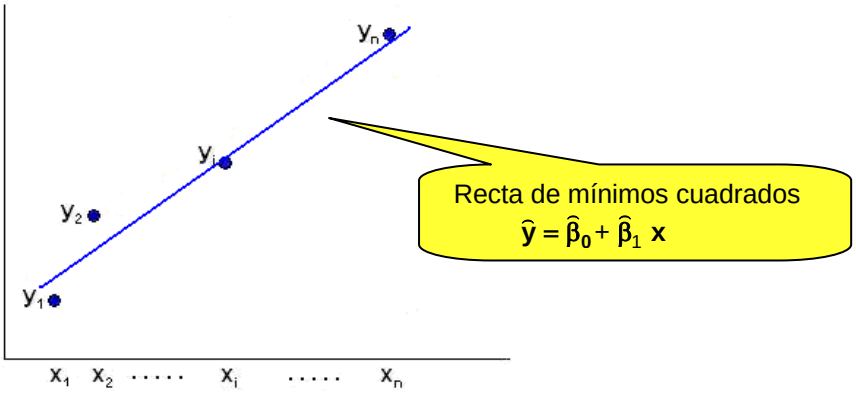
El siguiente procedimiento matemático permite usar los datos dados para construir una recta de la cual se obtienen estimadores para los parámetros β_0 y β_1 de la recta de regresión poblacional $\beta_0 + \beta_1 x$,

Se trata de colocar una recta entre los puntos dados, de la forma mejor balanceada con el criterio de hacer que la suma de las distancias de la recta a los puntos sea la menor posible. Esta recta se denomina recta de mínimos cuadrados.

Definición: Recta de mínimos cuadrados

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

En donde $\hat{\beta}_0$, $\hat{\beta}_1$ son los estimadores de β_0 y β_1 del modelo poblacional $\beta_0 + \beta_1 x$



Para cada valor x_i se tiene el dato observado y_i, mientras que al evaluar la recta de mínimos cuadrados $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ con este mismo valor x_i se obtiene el valor $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

Sea $e_i = y_i - \hat{y}_i$, la diferencia entre estos dos valores. Esta diferencia se denomina el **residual**.

Entonces, el criterio de mínimos cuadrados consiste en minimizar e_i^2 para todos los puntos.

El cuadrado puede interpretarse como una manera de cuantificar las diferencias sin importar el signo. La verdadera razón es formal y corresponde a la teoría de la estimación estadística.

Definición: Suma de los cuadrados del error

$$SCE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

SCE es una función con dos variables: $\hat{\beta}_0$, $\hat{\beta}_1$

Con el procedimiento matemático usual para encontrar su mínimo:

$$\frac{\partial SCE}{\partial \beta_0} = 0, \qquad \frac{\partial SCE}{\partial \beta_1} = 0$$

Después de derivar **SCE**, igualar a cero y simplificar se llega al sistema de ecuaciones lineales:

$$\hat{\beta}_0 n + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

De donde se obtienen finalmente $\hat{\beta}_0$, $\hat{\beta}_1$ para el modelo de mínimos cuadrados**:**

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. Este modelo puede usarse para realizar pronósticos

Obtener la recta de mínimos cuadrados para el ejemplo

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

i	x_i	y_i	x^2_i	$x_i y_i$
1	39	65	1521	2535
2	43	75	1849	3225
3	21	52	441	1092
4	64	82	4096	5248
5	57	92	3249	5244
6	43	80	1849	3440
7	38	73	1444	2774
8	75	98	5625	7350
9	34	56	1156	1904
10	52	75	2704	3900
$\sum_{i=1}^{10}$	**466**	**748**	**23934**	**36712**

$$\begin{aligned} \hat{\beta}_0 n + \hat{\beta}_1 \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i \\ \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i y_i \end{aligned} \quad \Rightarrow \quad \begin{aligned} 10\,\hat{\beta}_0 + 466\,\hat{\beta}_1 &= 748 \\ 466\,\hat{\beta}_0 + 23934\,\hat{\beta}_1 &= 36712 \end{aligned}$$

De donde se obtienen $\hat{\beta}_0 = 35.83, \quad \hat{\beta}_1 = 0.836$

Recta de mínimos cuadrados: $\hat{y} = 35.83 + 0.836\,x$

Pronosticar la calificación final si la calificación parcial es 50

$\hat{y} = 35.83 + 0.836\,(50) = 77.63$

11.2 COEFICIENTE DE CORRELACIÓN

Para determinar el tipo de relación lineal entre las variables **x** y **y** del modelo de regresión lineal se usa el **coeficiente de correlación lineal** que se define a continuación:

Para simplificar la escritura se establecen las siguientes definiciones

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

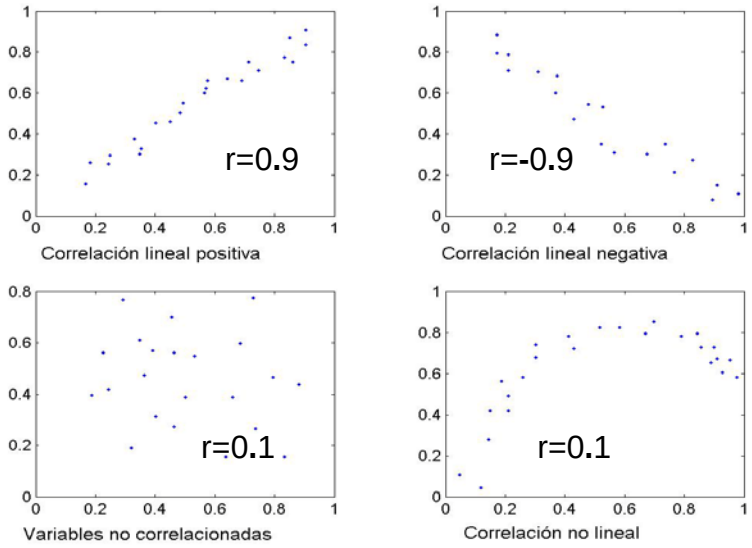
$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Definición: Coeficiente de correlación

$$r = \frac{Sxy}{\sqrt{Sxx\ Syy}}, \quad -1 \le r \le 1$$

El signo de **r** es igual al signo de la pendiente $\hat{\beta}_1$ de la recta de regresión lineal

Si el valor de **r** es cercano a **1** significa que hay una fuerte relación lineal positiva ente **x** y **y**
Si el valor de **r** es cercano a **-1** significa que hay una fuerte relación lineal negativa ente **x** y **y**
Si el valor de **r** es cercano a **0** significa que hay poca relación lineal ente **x** y **y**



Ejemplos de correlación entre dos variables

Calcular el coeficiente de correlación para el ejemplo

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{10}(39 + 43 + \ldots + 52) = 46.6$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{10}(65 + 75 + \ldots + 75) = 74.8$$

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = [(39 - 46.6)^2 + (43 - 46.6)^2 + \ldots + (52 - 46.6)^2] = 2218.4$$

$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = [(65 - 74.8)^2 + (75 - 74.8)^2 + \ldots + (75 - 74.8)^2] = 1885.6$$

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = [(39 - 46.6)(65 - 74.8) + \ldots] = 1855.2$$

$$r = \frac{Sxy}{\sqrt{Sxx\ Syy}} = \frac{1855.6}{\sqrt{(2218.4)(1885.6)}} = 0.9071$$

El resultado indica una fuerte correlación lineal positiva

11.3 ANÁLISIS DEL MODELO DE REGRESIÓN LINEAL SIMPLE

Para simplificar la escritura de algunas expresiones de interés, se definen las siguientes fórmulas equivalentes que pueden demostrarse algebraicamente desarrollando las sumatorias.

(1) $$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$$

(2) $$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$$

(3) $$\mathbf{SCT} = \mathbf{S_{yy}} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2$$

(4) $$\mathbf{SCE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = S_{yy} - \frac{S_{xy}^2}{Sxx}$$

(5) $$\mathbf{SCR} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \frac{S_{xy}^2}{S_{xx}}$$

Demostración de (1)

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2$$

$$= \sum_{i=1}^{n} x_i^2 - 2\bar{x}n\frac{1}{n}\sum_{i=1}^{n} x_i + n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - 2\bar{x}n\bar{x} + n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

$$= \sum_{i=1}^{n} x_i^2 - n\left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$$ esto completa la demostración

11.4 ANÁLISIS DE VARIANZA

El análisis de varianza es un método estadístico para conocer si los valores de un grupo de datos son significativamente diferentes de otro(s) grupo(s) de datos. Este método se puede aplicar al modelo de regresión lineal.

Algunos supuestos son necesarios para su aplicación, entre estos, que las observaciones sean independientes y que la distribución de la variable dependiente sea normal.

Consideremos la fórmula (4):

$$\mathbf{SCE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = S_{yy} - \frac{S_{xy}^2}{Sxx}$$

Se puede escribir

$$\mathbf{S_{yy}} = \frac{S_{xy}^2}{S_{xx}} + \mathbf{SCE}$$

Sustituyendo la fórmula (5)

$$\mathbf{S_{yy}} = \mathbf{SCR} + \mathbf{SCE}$$

Sustituyendo la definición de la fórmula (3)

$$\mathbf{SCT} = \mathbf{SCR} + \mathbf{SCE}$$

Con la sustitución de las equivalencias de las fórmulas (3), (4) y (5) se obtiene

Definición: Descomposición de la variabilidad para el modelo de regresión lineal

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Esta fórmula permite descomponer la variabilidad total **SCT** de la variable de respuesta (**y**) en dos componentes: la variabilidad **SCR** correspondiente a la recta de regresión de mínimos cuadrados, y la variación residual **SCE** que no se ha incluido en la recta de mínimos cuadrados obtenida

SCT: Suma de cuadrados total
SCR: Suma de cuadrados de regresión
SCE: Suma de cuadrados del error

Mientras menor es el valor de **SCE**, mayor es la eficacia del modelo de mínimos cuadrados obtenido, pues su variabilidad se ajusta o explica muy bien a la variabilidad de los datos **y**..

Encontrar los componentes de variación para el modelo del ejemplo

$$\mathbf{SCT = SCR + SCE}$$

$$\mathbf{SCT} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = (65 - 74.8)^2 + (75 - 74.8)^2 + \ldots + (75 - 74.8)^2 = 1885.6$$

$\hat{y} = 35.83 + 0.836\,x$ (Recta de mínimos cuadrados obtenida)

x=39: $\hat{y} = 35.83 + 0.836\,(39) = 68.434$

x=43: $\hat{y} = 35.83 + 0.836\,(43) = 71.778$

. . .

x=52: $\hat{y} = 35.83 + 0.836\,(52) = 79.302$

$$\mathbf{SCR} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

$$= (68.434 - 74.8)^2 + (71.778 - 74.8)^2 + \ldots + (79.302 - 74.8)^2 = 1550.4$$

$$\mathbf{SCE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$= (65 - 68.434)^2 + (75 - 71.778)^2 + \ldots + (75 - 79.302)^2 = 334.138$$

También se puede usar la definición para obtener directamente uno de los tres componentes:

$$\mathbf{SCT = SCR + SCE}$$

11.5 COEFICIENTE DE DETERMINACIÓN

El coeficiente de determinación es otra medida de la relación lineal entre las variables **x** y **y** Es útil para interpretar la eficiencia de la recta de mínimos cuadrados para explicar la variación de la variable de respuesta (**y**)

Definición: Coeficiente de determinación

$$r^2 = \frac{SCR}{SCT}, \quad 0 \le r^2 \le 1$$

El valor de r^2 mide el poder de explicación del modelo de mínimos cuadrados. Si r^2 es cercano a **1** significa que la recta de mínimos cuadrados se ajusta muy bien a los datos.

Calcular el coeficiente de determinación para el ejemplo

$$r^2 = \frac{SCR}{SCT} = \frac{1550.4}{1885.6} = 0.8222$$

El poder de explicación del modelo de mínimos cuadrados es **82.22%**

11.6 TABLA DE ANÁLISIS DE VARIANZA

En la ecuación

$$\mathbf{SCT = SCR + SCE}$$

SCR tiene 1 grado de libertad (varianza ponderada con el modelo con dos parámetros)
SCE tiene n – 2 grados de libertad (existen n datos y dos parámetros en el modelo)
SCT tiene n – 1 grados de libertad (suma de grados de libertad de **SCR** y **SCT**)

Si cada uno se divide por el número de grados de libertad se obtienen los **cuadrados medios**

Todos estos resultados se los ordena en un cuadro denominado **Tabla de Análisis de Varianza** o **Tabla ANOVA**

Tabla ANOVA

Fuente de variación	Grados de libertad	Suma de cuadrados	Cuadrados medios	F_0
Regresión	1	SCR	SCR/1	(SCR/1)/(SCE/(n-2))
Error	n – 2	SCE	S^2 = SCE/(n – 2)	
Total	n – 1	SCT		

El último cociente es el valor de una variable que tiene distribución **F**. Este estadístico se usa para una prueba del modelo propuesto

Escribir la tabla de análisis de varianza para el ejemplo

Fuente de variación	Grados de libertad	Suma de cuadrados	Cuadrados medios	F_0
Regresión	**1**	**1550.4**	**1550.4**	**37.00**
Error	**8**	**335.2**	**41.9**	
Total	**9**	**1885.6**		

11.7 PRUEBA DE DEPENDENCIA LINEAL DEL MODELO

Puede demostrarse que el estadístico

$$\mathbf{F_0 = \frac{SCR}{SCE/(n-2)}}$$ tiene distribución **F** con $\nu_1 = 1$, $\nu_2 = n - 2$ grados de libertad

Este estadístico se puede usar para realizar una prueba de hipótesis para la pendiente del modelo de regresión lineal

H_0: $\beta_1 = 0$, Hipótesis nula para probar que no hay **dependencia lineal** entre **x** y **y**
H_a: $\neg H_0$

Si se especifica el nivel de significancia α de la prueba, entonces la región crítica es

Rechazar H_0 si $f_0 > f_\alpha$ con $\nu_1 = 1$, $\nu_2 = n - 2$ grados de libertad

Probar con 5% de significancia de dependencia lineal para el ejemplo anterior

H_0: $\beta_1 = 0$

Región de rechazo de H_0:

$f_0 > f_{0.05}$ con $\nu_1 = 1$, $\nu_2 = 8$

$f_{0.05,\,1,\,8} = 5.32$ (Tabla **F**)

Conclusión

Debido a que $f_0 > 5.32$, se rechaza H_0, es decir **x** y **y** si están relacionadas linealmente

11.8 ESTIMACIÓN DE LA VARIANZA

La varianza de los errores del modelo σ^2 es desconocida. Para poder hacer inferencias acerca de los parámetros β_0 , β_1 es necesario un estimador.

Definición: Varianza muestral

$$S^2 = \frac{SCE}{n-2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$$

Es un estimador insesgado de la varianza del modelo teórico: $E[S^2] = \sigma^2$.

La variable aleatoria $\chi^2 = (n-2)\frac{S^2}{\sigma^2}$ tiene distribución **ji–cuadrado** con **n – 2** g. de libertad.

Estimación de la varianza para el ejemplo

$$S^2 = \frac{SCE}{n-2} = \frac{334.138}{8} = 41.7673$$

11.9 INFERENCIAS CON EL MODELO DE REGRESIÓN LINEAL

En el modelo probabilista propuesto:

$Y = \beta_0 + \beta_1 x + \varepsilon, \quad \varepsilon_i \sim N(0, \sigma^2)$ para cada variable aleatoria Y_i

El valor esperado de este modelo, es una recta desconocida con parámetros β_0 y β_1

$E[Y] = \beta_0 + \beta_1 x$

El modelo obtenido con el método de mínimos cuadrados es

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

En donde $\hat{\beta}_0$, $\hat{\beta}_1$ son los estimadores de los parámetros β_0 , β_1

Los estimadores son variables aleatorias pues dependen de los valores **y** observados.

Si los componentes ε_i del error son independientes, puede demostrarse que $\hat{\beta}_0$, $\hat{\beta}_1$ son estimadores insesgados, con distribución normal y con las siguientes varianzas:

$$E[\hat{\beta}_0] = \beta_0, \qquad V[\hat{\beta}_0] = \sigma^2_{\hat{\beta}_0} = \sigma^2 \left[\frac{\sum_{i=1}^{n} x_i^2}{nS_{xx}}\right]$$

$$E[\hat{\beta}_1] = \beta_1, \qquad V[\hat{\beta}_1] = \sigma^2_{\hat{\beta}_1} = \frac{\sigma^2}{S_{xx}}$$

Para definir estadísticos con los estimadores $\hat{\beta}_0$, $\hat{\beta}_1$ se sustituye la varianza desconocida σ^2 por el estimador S^2

$$S^2_{\hat{\beta}_0} = S^2\left[\frac{\sum_{i=1}^{n} x_i^2}{nS_{xx}}\right] \qquad S^2_{\hat{\beta}_1} = \frac{S^2}{S_{xx}}$$

Definición: Estadísticos para Estimación de los Parámetros β_0 y β_1

$$t = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{S^2_{\hat{\beta}_0}}}, \qquad t = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{S^2_{\hat{\beta}_1}}}$$

Tienen **distribución t** con $\nu = n - 2$ grados de libertad.

Varianza de los estimadores de mínimos cuadrados en el ejemplo

$$S^2_{\hat{\beta}_0} = S^2\left[\frac{\sum_{i=1}^{n} x_i^2}{nS_{xx}}\right] = 41.7673\left(\frac{39^2 + 43^2 + \ldots + 52^2}{10(2218.4)}\right) = 45.0575$$

$$S^2_{\hat{\beta}_1} = \frac{S^2}{S_{xx}} = \frac{41.7673}{2218.4} = 0.0188$$

11.10 INFERENCIAS ACERCA DE LA PENDIENTE DE LA RECTA

Es importante determinar si existe una relación entre las variables **x** y **y**. Esta relación está determinada por la pendiente β_1 de la recta.

11.10.1 INTERVALO DE CONFIANZA

Parámetro: β_1 (Pendiente de la recta de regresión lineal teórica)

Estimador: $\hat{\beta}_1$ (Pendiente de la recta de mínimos cuadrados)

El estadístico $t = \dfrac{\hat{\beta}_1 - \beta_1}{\sqrt{S^2_{\hat{\beta}_1}}}$, tiene distribución **t** con $\nu = n - 2$ grados de libertad

Como es usual, la desigualdad $-t_{\alpha/2} \le t \le t_{\alpha/2}$ tiene probabilidad $1 - \alpha$, de donde:

Definición: Intervalo de Confianza para la Pendiente β_1 con nivel $1 - \alpha$

$$\hat{\beta}_1 - t_{\alpha/2}\sqrt{S^2_{\hat{\beta}_1}} < \beta_1 < \hat{\beta}_1 + t_{\alpha/2}\sqrt{S^2_{\hat{\beta}_1}}$$

Intervalo de confianza para β_1 con nivel 95% para el ejemplo

$1 - \alpha = 0.95 \Rightarrow t_{\alpha/2} = t_{0.025} = 2.306$, $\nu = 8$ grados de libertad

$$\hat{\beta}_1 - t_{\alpha/2}\sqrt{S^2_{\hat{\beta}_1}} < \beta_1 < \hat{\beta}_1 + t_{\alpha/2}\sqrt{S^2_{\hat{\beta}_1}}$$

$$0.836 - 2.306\sqrt{0.0188} < \beta_1 < 0.836 + 2.306\sqrt{0.0188}$$

$$0.5196 < \beta_1 < 1.1524$$

Es el intervalo para la pendiente de la recta de regresión lineal

11.10.2 PRUEBA DE HIPÓTESIS

Parámetro: β_1 (Pendiente de la recta de regresión lineal teórica)

Estimador: $\hat{\beta}_1$ (Pendiente de la recta de mínimos cuadrados)

H_0: $\beta_1 = b_1$ ($b_1 = 0$, para probar que no hay **relación lineal** entre **x** y **y**)

H_a: $\beta_1 \neq b_1$

$\beta_1 < b_1$

$\beta_1 > b_1$

Estadístico de prueba

$t = \dfrac{\hat{\beta}_1 - b_1}{\sqrt{S^2_{\hat{\beta}_1}}}$, tiene distribución **t** con $\nu = n - 2$ grados de libertad

Si se especifica el nivel de significancia α se puede definir la región crítica

$\beta_1 < b_1$: $t < -t_\alpha$

$\beta_1 > b_1$: $t > t_\alpha$

$\beta_1 \neq b_1$: $t < -t_{\alpha/2} \vee t > t_{\alpha/2}$

Prueba de hipótesis con 5% de significancia que $\beta_1 < 1$ para el ejemplo

H_0: $\beta_1 = 1$

H_a: $\beta_1 < 1$

$\alpha = 0.05$

Región de rechazo de H_0: $t < -t_{0.05}$, $\nu = 8 \Rightarrow t < -1.86$

Cálculo del estadístico de prueba

$$t = \frac{\hat{\beta}_1 - b_1}{\sqrt{S^2_{\hat{\beta}_1}}} = \frac{0.836 - 1}{\sqrt{0.0188}} = -1.196$$

Conclusión

La evidencia no es suficiente para rechazar que la pendiente del modelo es **1**

11.11 INFERENCIAS PARA LA INTERCEPCIÓN DE LA RECTA

También puede ser de interés probar si la intercepción de la recta de regresión es igual a algún valor especificado

11.11.1 INTERVALO DE CONFIANZA

Parámetro: β_0 (Intercepción de la recta de regresión lineal teórica)

Estimador: $\hat{\beta}_0$ (Intercepción de la recta de mínimos cuadrados)

El estadístico $t = \dfrac{\hat{\beta}_0 - \beta_0}{\sqrt{S^2_{\hat{\beta}_0}}}$ tiene distribución t con $\nu = n - 2$ grados de libertad

La desigualdad $-t_{\alpha/2} \leq t \leq t_{\alpha/2}$ se satisface con probabilidad $1 - \alpha$, de donde se obtiene

Definición: Intervalo de Confianza para la Intercepción β_0 con nivel $1 - \alpha$

$$\hat{\beta}_0 - t_{\alpha/2}\sqrt{S^2_{\hat{\beta}_0}} < \beta_0 < \hat{\beta}_0 + t_{\alpha/2}\sqrt{S^2_{\hat{\beta}_0}}$$

Intervalo de confianza para β_0 con nivel 95% para el ejemplo

$1 - \alpha = 0.95 \Rightarrow t_{\alpha/2} = t_{0.025} = 2.306$, $\nu = 8$ grados de libertad

$$\hat{\beta}_0 - t_{\alpha/2}\sqrt{S^2_{\hat{\beta}_0}} < \beta_0 < \hat{\beta}_0 + t_{\alpha/2}\sqrt{S^2_{\hat{\beta}_0}}$$

$$35.83 - 2.306\sqrt{45.0575} < \beta_0 < 35.83 + 2.306\sqrt{45.0575}$$

$$20.351 < \beta_0 < 51.309$$

Es el intervalo para la intercepción de la recta de regresión lineal

11.11.2 PRUEBA DE HIPÓTESIS

Parámetro: β_0 (Intercepción de la recta de regresión lineal teórica)

Estimador: $\hat{\beta}_0$ (Intercepción de la recta de mínimos cuadrados)

H_0: $\beta_0 = b_0$ (**b_0:** algún valor especificado para la intercepción)

H_a: $\beta_0 \neq b_0$

$\beta_0 < b_0$

$\beta_0 > b_0$

Estadístico de prueba

$$t = \frac{\hat{\beta}_0 - b_0}{\sqrt{S^2_{\hat{\beta}_0}}}$$, tiene distribución t con $\nu = n - 2$ grados de libertad

Si se especifica el nivel de significancia α se puede definir la región crítica

$\beta_0 < b_0$: $t < -t_\alpha$

$\beta_0 > b_0$: $t > t_\alpha$

$\beta_0 \neq b_0$: $t < -t_{\alpha/2} \vee t > t_{\alpha/2}$

Prueba de hipótesis con 5% de significancia que $\beta_0 > 30$ para el ejemplo

H_0: $\beta_0 = 30$

H_a: $\beta_0 > 30$

$\alpha = 0.05$

Región de rechazo de **H_0:** $t > t_{0.05}$, $\nu = 8 \Rightarrow t > 1.86$

Cálculo del estadístico de prueba

$$t = \frac{\hat{\beta}_0 - b_0}{\sqrt{S^2_{\hat{\beta}_0}}} = \frac{35.83 - 30}{\sqrt{45.0575}} = 0.8685$$

Conclusión

La evidencia no es suficiente para rechazar que la intercepción del modelo es 30

11.12 PRUEBA DE LA NORMALIDAD DEL ERROR

Se puede usar la prueba **K-S** para probar la suposición de normalidad de los errores

Prueba de Kolmogorov-Smirnov con 5% de significancia para la normalidad del error con los datos del ejemplo

Ho: $\varepsilon \sim N(0, \sigma^2)$ (Distribución normal con media **0** y varianza σ^2)

Ha: $\neg$ **Ho**

$\alpha = 0.05$

Estadístico de prueba

$D_n = \max| S_n(x_i) - F_0(x_i)|$ (Para este ejemplo x_i son los valores e_i)

Región de rechazo de Ho

$\alpha = 0.05$, $n = 10 \Rightarrow D_{0.05} = 0.410$ (Tabla K-S)

Rechazar H_0 si $D_n > 0.410$

$\varepsilon_i \cong e_i = y_i - \hat{y}_i$, $i = 1, 2, .., 10$

$\hat{y} = 35.83 + 0.836\,x$ (Recta de mínimos cuadrados obtenida)

$x_1 = 39 \Rightarrow \hat{y}_1 = 35.83 + 0.836\,(39) = 68.434$

$e_1 = y_1 - \hat{y}_1 = 65 - 68.434 = -3.434$, etc.

$$\begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \\ e_8 \\ e_9 \\ e_{10} \end{bmatrix} = \begin{bmatrix} -3.434 \\ 3.222 \\ -1.386 \\ -7.334 \\ 8.518 \\ 8.222 \\ 5.4020 \\ -0.530 \\ -8.254 \\ -4.302 \end{bmatrix}$$

Modelo propuesto $e_i \sim N(0, \sigma^2)$ (Aproximadamente)

$F_0(x_i) = F_0(e_i) = P\left(Z < \dfrac{e_i - 0}{\sigma}\right)$ Distribución normal estándar acumulada

$\sigma^2 \cong S^2 = 41.7673 \Rightarrow \sigma \cong S = 6.4627$

$F_0(x_1) = F_0(e_1) = P\left(Z < \dfrac{-8.254 - 0}{6.4627}\right) = 0.1008$, etc. (Datos **e** ordenados)

Tabulación de resultados. Se utiliza la notación $x_i = e_i$

i	x_i (ordenados)	$S_n(x_i)$	$F_0(x_i)$	$\lvert S_n(x_i) - F_0(x_i) \rvert$
1	-8.254	0.1	0.1008	0.0008
2	-7.334	0.2	0.1282	0.0718
3	-4.302	0.3	0.2528	0.0472
4	-3.434	0.4	0.2976	0.1024
5	-1.386	0.5	0.4151	0.0849
6	-0.530	0.6	0.4673	**0.1327**
7	3.222	0.7	0.6910	0.0090
8	5.402	0.8	0.7984	0.0016
9	8.222	0.9	0.8984	0.0984
10	8.518	1.0	0.9063	0.0937

$D_n = \max\lvert S_n(x_i) - F_0(x_i) \rvert = 0.1327$

Conclusión: D_n no cae en la región de rechazo, por lo tanto no se puede rechazar **Ho**

11.13 EJERCICIOS

Los siguientes datos, en miles de dólares, representan los ingresos por ventas vs. los gastos de promoción de un producto:

Gastos:	**0.5**	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**
Ingresos:	**3.5**	**4.1**	**5.5**	**7.2**	**8.7**	**9.5**

Suponga que la variable de predicción (**X**) corresponde a los gastos, y la variable de respuesta (**Y**) se refiere a los ingresos.

a) Construya un diagrama de dispersión de los datos
b) Encuentre la recta de mínimos cuadrados
c) Calcule el coeficiente de correlación e interprete el resultado
d) Construya la tabla ANOVA
e) Calcule el coeficiente de determinación e interprete el resultado
f) Encuentre una estimación para la varianza de los errores del modelo
g) Encuentre la varianza de los estimadores del modelo de mínimos cuadrados
h) Construya un intervalo de confianza de 95% para la pendiente del modelo
i) Pruebe con 5% de significancia que la pendiente del modelo es mayor a 2
j) Pruebe la normalidad del error con la prueba K-S

MATLAB

Regresión lineal simple usando notación matricial

```
>> x=[1 39 ; 1 43 ; 1  21;  1 64; 1  57;  1 43; 1  38; 1  75;1  34;1  52]
```
Matriz de diseño **X**

```
   x =
        1    39
        1    43
        1    21
        1    64
        1    57
        1    43
        1    38
        1    75
        1    34
        1    52
>> y=[ 65; 75;  52;  82;  92;  80; 73; 98; 56; 75]
```
Vector de observaciones

```
   y =
       65
       75
       52
       82
       92
       80
       73
       98
       56
       75
>> [b, bint, e, eint, stats] = regress(y,x, 0.05)
```
Regresión lineal simple $\alpha = 0.05$

```
    b =
         35.8294
          0.8363
```
Coeficientes β_0 , β_1 del modelo de mínimos cuadrados

```
    bint =
         20.3497   51.3092
          0.5199    1.1527
```
Intervalos de confianza para β_0 , β_1

```
    e =
        -3.4443
         3.2106
        -1.3913
        -7.3512
         8.5027
         8.2106
         5.3920
        -0.5503
        -8.2629
        -4.3159

   stats =
            0.8228   37.1456    0.0003
```
Coeficiente de determinación R^2**,** valor del estadístico **F,** valor **p** de la prueba **F**

Uso del modelo de mínimos cuadrados

```
>> yp=b(1) + b(2)*50
```
Evaluar el modelo con **x = 50**

```
   yp =
         77.6433
```

Matriz de correlación de los datos de la muestra

```
>> mc = corrcoef(x(:,2),y)
   mc =
        1.0000    0.9071
        0.9071    1.0000
```

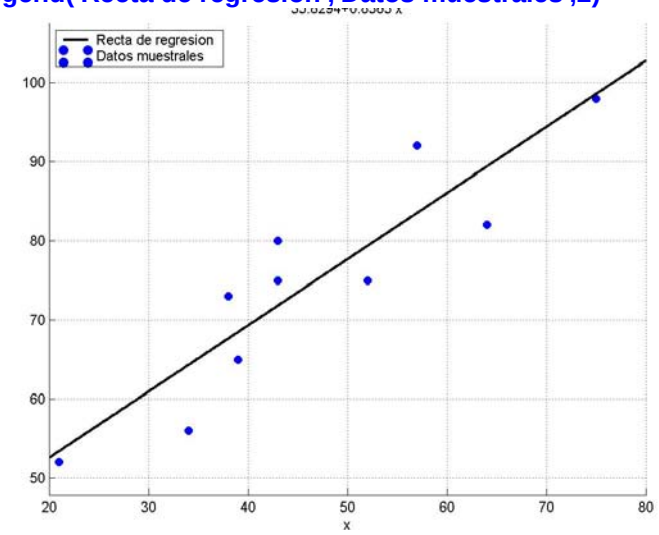
Vectores columnas **X, Y**

Coeficiente de correlación lineal
r = 0.9071

Gráfico de los puntos muestrales y la recta de regresión

```
>> clf
>> scatter(x(:,2),y,'filled'),grid on
>> hold on, ezplot('35.8294+0.8363*x',[20, 80])
>> legend('Recta de regresion','Datos muestrales',2)
```

Gráfico de dispersión
Gráfico de la recta de regresión
Rótulos



Prueba de la normalidad del error de los residuales

```
>> sce=sum(e.^2)
   sce =
        334.1363
>> s2=sce/8
   s2 =
         41.7670
>> t=sort(e);
>> f=normcdf(t, 0, sqrt(s2));
>> [h,p,ksstat,vc]=kstest(t, [t f ], 0.05,0)
   h =
      0
   p =
     0.9891
   ksstat =
           0.1339
   vc =
        0.4093
```

Suma de los cuadrados de residuales

Estimación de la varianza S^2

Residuales ordenados

Modelo a probar $e_i \sim N(0, \sigma^2)$r

Prueba **K-S,** $\alpha = 0.05$

No se puede rechazar el modelo

Valor **p** de la prueba

Valor del estadístico de prueba

Valor crítico de la región de rechazo

Matriz de varianzas y covarianzas de los estimadores β_i

```
>> mvc = inv(x' *x)*s2
   mvc =
        45.0619   -0.8774
        -0.8774    0.0188
```

Usando notación matricial

$V(\beta_0) = 45.0619$, $V(\beta_1) = 0.0188$
$Cov(\beta_0, \beta_1) = -0.8774$

12 REGRESIÓN LINEAL MÚLTIPLE

Consideramos el caso de una variable **Y** que suponemos depende linealmente de otras **k** variables $x_1, x_2, \ldots, x_k$. Para describir esta relación se propone un modelo de regresión lineal múltiple poblacional

Definición: Modelo de regresión lineal múltiple

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

En donde $\beta_0, \beta_2, \ldots, \beta_k$ son los parámetros que deben estimarse para el modelo, mientras que ε es el componente aleatorio de **Y**.

Cuando **k = 1**, se obtiene el modelo de regresión lineal simple previamente estudiado.

Suponer que se tiene una muestra aleatoria $(x_{1,i}, x_{2,i}, \ldots, x_{k,i}, y_i)$, **i = 1, 2, ..., n**

Para cada grupo de **k** valores $x_{1,i}, x_{2,i}, \ldots, x_{k,i}$ se tiene un resultado u observación y_i. Este es uno de los posibles valores de la variable aleatoria Y_i. Una variable aleatoria debe tener una distribución de probabilidad. La aleatoriedad de Y_i está dada por ε_i. Se supondrá que para cada variable aleatoria Y_i el componente aleatorio ε_i es una variable con la misma distribución de probabilidad, y que además son variables independientes.

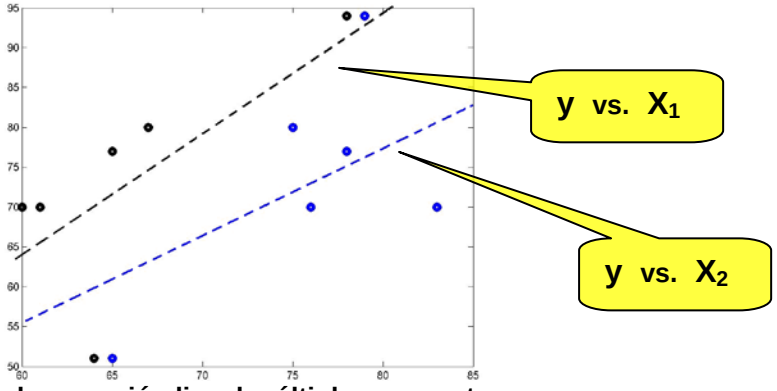
Para comprensión de conceptos se desarrolla paralelamente un ejemplo

Ejemplo

Se desea definir un modelo de regresión relacionando la calificación final en cierta materia con la calificación parcial y el porcentaje de asistencia a clases. Para el análisis se usará una muestra aleatoria de **6** estudiantes que han tomado esta materia.

Estudiante	1	2	3	4	5	6
Nota Parcial X_1	67	65	78	60	64	61
% Asistencia X_2	75	78	79	83	65	76
Nota Final Y	80	77	94	70	51	70

Diagramas de dispersión: y vs. x_1, y vs. x_2



Modelo teórico de regresión lineal múltiple propuesto

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \quad .$$

12.1 MÉTODO DE MÍNIMOS CUADRADOS

El siguiente procedimiento matemático permite usar los datos dados para construir un modelo con el cual se obtienen $\hat{\beta}_0$, $\hat{\beta}_1$, ..., $\hat{\beta}_k$ que serán los estimadores de los parámetros β_0 , β_1 , ..., β_k del modelo teórico de regresión lineal múltiple propuesto.

Definición: Modelo de mínimos cuadrados

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_k x_k$$

En donde $\hat{\beta}_0$, $\hat{\beta}_1$, ..., $\hat{\beta}_k$ son los **k+1** estimadores para los **k+1** parámetros β_0 , β_1 , ..., β_k

Para cada valor x_i se tiene el dato observado y_i, mientras que al evaluar el modelo de mínimos cuadrados con este mismo valor x_i se obtiene el valor $\hat{y}_i$

Sea $e_i = y_i - \hat{y}_i$, la diferencia entre estos dos valores. Esta diferencia se denomina el **residual**.

Definición: Suma de los cuadrados del error

$$SCE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1,i} - \hat{\beta}_2 x_{2,i} - ... - \hat{\beta}_k x_{k,i})^2$$

SCE es una función con **k + 1** variables: $\hat{\beta}_0$, $\hat{\beta}_1$, ..., $\hat{\beta}_k$

Usando el conocido procedimiento matemático para minimizar **SCE**:

$$\frac{\partial SCE}{\partial \hat{\beta}_i} = 0, \quad i=0, 1, 2, ... , k$$

Resulta un sistema de **k+1** ecuaciones lineales de donde se obtienen los **k+1** estimadores

$\hat{\beta}_0$, $\hat{\beta}_1$, ..., $\hat{\beta}_k$

12.2 MÉTODO DE MÍNIMOS CUADRADOS PARA k = 2

Supongamos que **Y** depende de dos variables x_1, x_2

Modelo teórico de regresión lineal múltiple propuesto:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \quad .$$

Modelo de mínimos cuadrados;

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

Para encontrar $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, derivar **SCE** e igualar a cero: $\frac{\partial SCE}{\partial \hat{\beta}_i} = 0$, **i = 0, 1, 2** .

Luego de la aplicación y simplificación algebraica se obtiene

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{1,i} + \hat{\beta}_2 \sum_{i=1}^{n} x_{2,i} = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_{1,i} + \hat{\beta}_1 \sum_{i=1}^{n} x_{1,i}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{1,i} x_{2,i} = \sum_{i=1}^{n} x_{1,i} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_{2,i} + \hat{\beta}_1 \sum_{i=1}^{n} x_{2,i} x_{1,i} + \hat{\beta}_2 \sum_{i=1}^{n} x_{2,i}^2 = \sum_{i=1}^{n} x_{2,i} y_i$$

Al resolver este sistema lineal se obtienen los estimadores $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$

El sistema de ecuaciones se puede expresar en notación matricial

$$\mathbf{A}\,\hat{\beta} = \mathbf{C}$$

Siendo

$$\mathbf{A} = \begin{bmatrix} n & \sum_{i=1}^{n} x_{1,i} & \sum_{i=1}^{n} x_{2,i} \\ \sum_{i=1}^{n} x_{1,i} & \sum_{i=1}^{n} x_{1,i}^2 & \sum_{i=1}^{n} x_{1,i}x_{2,i} \\ \sum_{i=1}^{n} x_{2,i} & \sum_{i=1}^{n} x_{2,i}x_{1,i} & \sum_{i=1}^{n} x_{2,i}^2 \end{bmatrix}, \quad \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{1,i}y_i \\ \sum_{i=1}^{n} x_{2,i}y_i \end{bmatrix}$$

12.3 REGRESIÓN LINEAL MÚLTIPLE EN NOTACIÓN MATRICIAL

En esta sección se describe la notación matricial para expresar el modelo de regresión lineal múltiple. Esta notación es usada después para el modelo de regresión de mínimos cuadrados.

Consideramos el caso específico **k = 2** en donde **Y** depende de dos variables x_1, x_2

Modelo de regresión lineal poblacional propuesto:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon, \qquad \varepsilon_i \sim N(0, \sigma^2)$$

Datos de la muestra:

$$(x_{1,i}, x_{2,i}, y_i),\ i = 1, 2, ..., n$$

Cada observación y_i es un valor de la variable aleatoria Y_i, **i = 1, 2, ..., n**

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i \quad ,i= 1, 2, ..., n$$

En forma desarrollada,

$$Y_1 = \beta_0 + \beta_1 x_{1,1} + \beta_2 x_{2,1} + \varepsilon_1$$
$$Y_2 = \beta_0 + \beta_1 x_{1,2} + \beta_2 x_{2,2} + \varepsilon_2$$
$$.$$
$$.$$
$$.$$
$$Y_n = \beta_0 + \beta_1 x_{1,n} + \beta_2 x_{2,n} + \varepsilon_n$$

El modelo teórico expresado en notación matricial es

$$\begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} \\ 1 & x_{1,2} & x_{2,2} \\ . & . & . \\ . & . & . \\ 1 & x_{1,n} & x_{2,n} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ \varepsilon_n \end{bmatrix}$$

En forma simbólica

$$Y = X\,\beta + \varepsilon$$

En donde

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ Y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} \\ 1 & x_{1,2} & x_{2,2} \\ . & . & . \\ . & . & . \\ 1 & x_{1,n} & x_{2,n} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ \varepsilon_n \end{bmatrix}$$

La matriz **X** se denomina **matriz de diseño**

El sistema de ecuaciones del modelo de regresión lineal múltiple de mínimos cuadrados, **k=2**

$$A\,\hat{\beta} = C$$

puede entonces expresarse con la notación matricial desarrollada para el modelo teórico:

La matriz de coeficientes **A** se puede construir con la **matriz de diseño X**

$$A = \begin{bmatrix} n & \sum_{i=1}^{n} x_{1,i} & \sum_{i=1}^{n} x_{2,i} \\ \sum_{i=1}^{n} x_{1,i} & \sum_{i=1}^{n} x_{1,i}^2 & \sum_{i=1}^{n} x_{1,i}x_{2,i} \\ \sum_{i=1}^{n} x_{2,i} & \sum_{i=1}^{n} x_{2,i}x_{1,i} & \sum_{i=1}^{n} x_{2,i}^2 \end{bmatrix} = \begin{bmatrix} 1 & 1 & . & . & 1 \\ x_{1,1} & x_{1,2} & . & . & x_{1,n} \\ x_{2,1} & x_{2,2} & . & . & x_{2,n} \end{bmatrix} . \begin{bmatrix} 1 & x_{1,1} & x_{2,1} \\ 1 & x_{1,2} & x_{2,2} \\ . & . & . \\ . & . & . \\ 1 & x_{1,n} & x_{2,n} \end{bmatrix}$$

En forma simbólica: $A = X^T X$

El vector **C** puede expresarse también con la **matriz de diseño X**

$$C = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{1,i}y_i \\ \sum_{i=1}^{n} x_{2,i}y_i \end{bmatrix} = \begin{bmatrix} 1 & 1 & . & . & 1 \\ x_{1,1} & x_{1,2} & . & . & x_{1,n} \\ x_{2,1} & x_{2,2} & . & . & x_{2,n} \end{bmatrix} . \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix}$$

En forma simbólica: $C = X^T y$

Con esta notación el modelo de mínimos cuadrados se puede escribir

$$A\,\hat{\beta} = C \Rightarrow X^T X\,\hat{\beta} = X^T y$$

En donde $\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix}$, $y = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix}$

Finalmente, con la inversa de $X^T X$ se pueden obtener los estimadores de mínimos cuadrados:

$$\hat{\beta} = (X^T X)^{-1} (X^T y)$$

Siendo $\hat{\beta}$: Vector con los estimadores de mínimos cuadrados
X: Matriz de diseño (construida con los datos de la muestra)
y: Vector de observaciones obtenidas en la muestra

La extensión de la notación matricial para k > 2 es directa

Modelo de regresión lineal en notación matricial para el ejemplo

Modelo de regresión lineal poblacional propuesto:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

En notación matricial

$$Y = X\beta + \varepsilon$$

En forma desarrollada, **n = 6**

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} \\ 1 & x_{1,2} & x_{2,2} \\ 1 & x_{1,3} & x_{2,3} \\ 1 & x_{1,4} & x_{2,4} \\ 1 & x_{1,5} & x_{2,5} \\ 1 & x_{1,6} & x_{2,6} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{bmatrix}$$

Matriz de diseño con los datos

$$X = \begin{bmatrix} 1 & 67 & 75 \\ 1 & 65 & 78 \\ 1 & 78 & 79 \\ 1 & 60 & 83 \\ 1 & 64 & 65 \\ 1 & 61 & 76 \end{bmatrix}$$

Obtener el modelo de mínimos cuadrados para el ejemplo (usar la matriz de diseño)

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

$$\hat{\beta} = (X^T X)^{-1} (X^T y)$$

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \left(\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 67 & 65 & 78 & 60 & 64 & 61 \\ 75 & 78 & 79 & 83 & 65 & 76 \end{bmatrix} \begin{bmatrix} 1 & 67 & 75 \\ 1 & 65 & 78 \\ 1 & 78 & 79 \\ 1 & 60 & 83 \\ 1 & 64 & 65 \\ 1 & 61 & 76 \end{bmatrix} \right)^{-1} \left(\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 67 & 65 & 78 & 60 & 64 & 61 \\ 75 & 78 & 79 & 83 & 65 & 76 \end{bmatrix} \begin{bmatrix} 80 \\ 77 \\ 94 \\ 70 \\ 51 \\ 70 \end{bmatrix} \right)$$

$$= \begin{bmatrix} 6 & 395 & 456 \\ 395 & 26215 & 30033 \\ 456 & 30033 & 34840 \end{bmatrix}^{-1} \begin{bmatrix} 442 \\ 29431 \\ 33877 \end{bmatrix}$$

$$= \begin{bmatrix} 48.974 & -02880 & -0.3927 \\ -0.2880 & 4.760 \times 10^{-3} & -3.360 \times 10^{-4} \\ -0.3927 & -3.366 \times 10^{-4} & 5.458 \times 10^{-3} \end{bmatrix} \begin{bmatrix} 442 \\ 29431 \\ 33877 \end{bmatrix} = \begin{bmatrix} -134.07 \\ 1.4888 \\ 1.4437 \end{bmatrix}$$

Modelo de mínimos cuadrados para el ejemplo

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 = -134.07 + 1.4888\, x_1 + 1.4437\, x_2$$

Pronosticar la calificación final de un estudiante si la calificación parcial es 75 y el porcentaje de asistencia a clases es 80

$$\hat{y} = -134.07 + 1.4888\,(75) + 1.4437(80) \;\; = 93.08$$

12.4 ANÁLISIS DE VARIANZA

Para este modelo también se aplica la misma interpretación de las fuentes de variación con las siguientes definiciones, similares al modelo de regresión lineal simple:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$\text{SCT} = \sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad \text{SCE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad \text{SCR} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

Se obtiene la relación entre las fuentes de error del modelo de regresión lineal múltiple

$$\text{SCT} = \text{SCR} + \text{SCE} \qquad .$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Esta fórmula permite descomponer la variabilidad total **SCT** de la variable de respuesta (**y**) en dos componentes: la variabilidad **SCR** correspondiente al modelo de regresión de mínimos cuadrados, y la variación residual **SCE** que no se ha incluido en el modelo calculado

SCT: Suma de cuadrados total
SCR: Suma de cuadrados de regresión
SCE: Suma de cuadrados del error

Mientras menor es el valor de **SCE**, mejor es la eficacia del modelo de mínimos cuadrados propuesto.

Análisis de varianza para el ejemplo

$$\text{SCT} = \text{SCR} + \text{SCE}$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{6}(80 + 77 + 94 + 70 + 51 + 70) = 73.6666$$

$$\hat{y} = -134.07 + 1.4888\,x_1 + 1.4437 x_2 \quad \text{(Modelo de mínimos cuadrados obtenido)}$$

$x_1 = 67, x_2 = 75$: $\hat{y} = -134.07 + 14888(67) + 1.4437(75) = 73.9571$
$x_1 = 65, x_2 = 78$: $\hat{y} = -134.07 + 14888(65) + 1.4437(78) = 75.3106$
. . .
$x_1 = 61, x_2 = 76$: $\hat{y} = -134.07 + 14888(61) + 1.4437(76) = 66.4680$

$$\text{SCT} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = (80 - 73.6666)^2 + (77 - 73.6666) + \ldots + (70 - 73.6666)^2 = 1005.3$$

$$SCR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$
$$= (73.9571 - 73.6666)^2 + (75.3106 - 73.6666)^2 + \ldots + (66.4680 - 73.6666)^2$$
$$= 906.7070$$
$$SCE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$
$$= (80 - 73.9571)^2 + (77 - 75.3106)^2 + \ldots + (70 - 66.4680)^2 = 98.5831$$

También se puede usar la definición para obtener directamente uno de los tres componentes:

$$SCT = SCR + SCE$$

12.5 COEFICIENTE DE DETERMINACIÓN

El coeficiente de determinación es otra medida de la relación lineal entre las variables **x** y **y** Es útil para interpretar la eficiencia del modelo de mínimos cuadrados para explicar la variación de la variable de respuesta

Definición: Coeficiente de Determinación

$$r^2 = \frac{SCR}{SCT}, \quad 0 \le r^2 \le 1$$

El valor de r^2 mide el poder de explicación del modelo de mínimos cuadrados. Si r^2 es cercano a **1** significa que el modelo de mínimos cuadrados se ajusta muy bien a los datos.

Coeficiente de determinación para el ejemplo

$$r^2 = \frac{SCR}{SCT} = \frac{906.707}{1005.3} = 0.9019 = 90.19\%$$

El poder de explicación del modelo de mínimos cuadrados es **90.19%**

12.6 TABLA DE ANÁLISIS DE VARIANZA

En la ecuación

$$SCT = SCR + SCE$$

SCR tiene k grados de libertad (varianza ponderada con el modelo con k+1 parámetros)
SCE tiene n – k – 1 grados de libertad (existen n datos y k parámetros en el modelo)
SCT tiene n – 1 grados de libertad

Si cada uno se divide por el número de grados de libertad se obtienen los **cuadrados medios**

Todos estos resultados se los ordena en un cuadro denominado **Tabla de Análisis de Varianza** o **Tabla ANOVA**

Tabla ANOVA

Fuente de variación	Grados de libertad	Suma de cuadrados	Cuadrados medios	F_0
Regresión	k	SCR	SCR/k	(SCR/k)/(SCE/(n–k–1))
Error	n – k – 1	SCE	SCE/(n – k – 1)	
Total	n – 1	SCT		

El último cociente es el valor de una variable que tiene distribución **F**. Este estadístico se usa para una prueba del modelo propuesto

Tabla de Análisis de Varianza para el ejemplo

Fuente de variación	Grados de libertad	Suma de cuadrados	Cuadrados medios	F_0
Regresión	2	906.707	453.3535	13.7961
Error	3	98.5831	32.8610	
Total	5	1005.3		

12.7 PRUEBA DE DEPENDENCIA LINEAL DEL MODELO

Puede demostrarse que el estadístico

$$F_0 = \frac{SCR/k}{SCE/(n-k-1)}$$ tiene distribución **F** con $\nu_1 = k$, $\nu_2 = n-k-1$ grados de libertad

Este estadístico se puede usar para realizar una prueba de hipótesis para determinar la dependencia lineal del modelo de regresión lineal propuesto

H_0: $\beta_1 = \ldots = \beta_k = 0$, No hay **dependencia lineal** de **y** con las X_i

H_a: $\neg H_0$ La respuesta **Y depende linealmente** de **al menos una** variable X_i

Si se especifica el nivel de significancia α de la prueba, entonces la región crítica es

Rechazar H_0 si $f_0 > f_\alpha$ con $\nu_1 = k$, $\nu_2 = n-k-1$ grados de libertad

Pruebe con 5% de significancia la dependencia lineal para el ejemplo anterior

$H_0: \beta_1 = \beta_2 = 0$

Región de rechazo de H_0:

$f_{0.05}$ con $\nu_1 = 2$, $\nu_2 = 3 \Rightarrow f_{0.05,\,2,\,3} = 9.55$ (Tabla **F**)

Rechazar H_0 si $f_0 > 9.55$

Conclusión: Debido a que $f_0 = 13.7961$ es mayor a **9.55,** se rechaza H_0, es decir que **al menos una** de las variables independientes X_1, X_2 contribuyen significativamente al modelo

12.8 ESTIMACIÓN DE LA VARIANZA

La varianza de los errores del modelo σ^2 es desconocida. Para poder hacer inferencias acerca de los parámetros $\beta_0, \beta_1, \ldots, \beta_k$ es necesario un estimador.

Definición: Varianza Muestral

$$S^2 = \frac{SCE}{n-k-1} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k-1}$$

Es un estimador insesgado de la varianza del modelo teórico: $E[S^2] = \sigma^2$

Estimación de la varianza muestral para el ejemplo

$$S^2 = \frac{SCE}{n-k-1} = \frac{98.583}{6-2-1} = 32.861$$

12.9 MATRIZ DE VARIANZAS Y COVARIANZAS

Es una forma ordenada de expresar las varianzas y covarianzas de los estimadores del modelo de regresión lineal

La estadística matemática demuestra la siguiente expresión matricial denominada matriz de varianzas y covarianzas, con la cual se pueden definir los estadísticos de prueba

Definición: Matriz de varianzas y covarianzas

$$[\sigma_{ij}] = (X^T X)^{-1} \sigma^2 \cong (X^T X)^{-1} S^2 = \begin{bmatrix} \sigma_{00} & \sigma_{01} & \cdot & \cdot & \sigma_{0k} \\ \sigma_{10} & \sigma_{11} & \cdot & \cdot & \sigma_{1k} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \sigma_{k0} & \sigma_{k1} & \cdot & \cdot & \sigma_{kk} \end{bmatrix}$$

En donde $\mathbf{X}$ es la matriz de diseño del modelo de regresión lineal múltiple

Las varianzas y covarianzas de los estimadores se definen de la siguiente forma:

$$V[\hat{\beta}_i] = \sigma^2_{\hat{\beta}_i} = \sigma_{ii}, \qquad i = 0, 1, ..., k \qquad \text{(Varianza de } \hat{\beta}_i \text{)}$$

$$Cov[\hat{\beta}_i, \hat{\beta}_j] = \sigma_{\hat{\beta}_i \hat{\beta}_j} = \sigma_{ij} \qquad i = 0, 1, ..., k \qquad \text{(Covarianza de } \hat{\beta}_i, \hat{\beta}_j \text{)}$$

Matriz de varianzas y covarianzas para el ejemplo

$$[\sigma_{i,j}] = (X^T X)^{-1} \sigma^2 \cong (X^T X)^{-1} S^2 = \begin{bmatrix} 48.974 & -02880 & -0.3927 \\ -0.2880 & 4.760x10^{-3} & -3.360x10^{-4} \\ -0.3927 & -3.366x10^{-4} & 5.458x10^{-3} \end{bmatrix} (32.861)$$

$$= \begin{bmatrix} 1609.33 & -9.4653 & -12.904 \\ -9.4653 & 0.15654 & -0.01106 \\ -12.904 & -0.01106 & 0.17937 \end{bmatrix}$$

Varianza de los estimadores de mínimos cuadrados para el ejemplo

$$V[\hat{\beta}_i] = \sigma^2_{\hat{\beta}_i} = \sigma_{ii}, \quad i = 0, 1, 2$$

$$V[\hat{\beta}_0] = \sigma^2_{\hat{\beta}_0} = \sigma_{00} = 1609.33$$

$$V[\hat{\beta}_1] = \sigma^2_{\hat{\beta}_1} = \sigma_{11} = 0.15654$$

$$V[\hat{\beta}_2] = \sigma^2_{\hat{\beta}_2} = \sigma_{22} = 0.17937$$

12.10 INFERENCIAS CON EL MODELO DE REGRESIÓN LINEAL

El modelo teórico probabilista propuesto es:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2)$$

El modelo obtenido con el método de mínimos cuadrados es:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$$

Del cual se obtienen los estimadores $\hat{\beta}_0$, $\hat{\beta}_1$, ..., $\hat{\beta}_k$ para los parámetros β_0 , β_1 , ..., β_k

Los estimadores son variables aleatorias pues dependen de valores aleatorios observados **y**.

Si los componentes ε_i del error son independientes, puede demostrarse que los estimadores son insesgados

$$E[\hat{\beta}_i] = \beta_i, \qquad i = 0, 1, \ldots, k$$

Cada estimador $\hat{\beta}_i$ tiene distribución normal

$$\hat{\beta}_i \sim N(\beta_i, \sigma^2_{\hat{\beta}_i}), \qquad i = 0, 1, \ldots, k$$

12.10.1 ESTADÍSTICOS PARA ESTIMACIÓN DE PARÁMETROS

Se establecen los estadísticos para realizar inferencias

Definición: Estadísticos para estimación de los parámetros β_0 , β_1 , ..., β_k

$$t = \frac{\hat{\beta}_i - \beta_i}{\sqrt{\sigma^2_{\hat{\beta}_i}}},$$ tienen distribución t con $\nu = n - k - 1$ grados de libertad

$i = 0, 1, \ldots, k$

12.10.2 INTERVALO DE CONFIANZA

Parámetro: β_i , $i = 0, 1, \ldots, k$

Estimador: $\hat{\beta}_i$, $i = 0, 1, \ldots, k$

El estadístico

$$t = \frac{\hat{\beta}_i - \beta_i}{\sqrt{\sigma^2_{\hat{\beta}_i}}},$$ tiene distribución t con $\nu = n - k - 1$ grados de libertad

$i = 0, 1, \ldots, k$

Como es usual, la desigualdad $-t_{\alpha/2} \le t \le t_{\alpha/2}$ tiene probabilidad $1 - \alpha$. De donde se obtiene

Definición: Intervalo de confianza para β_i con nivel $1 - \alpha$

$$\hat{\beta}_i - t_{\alpha/2} \sqrt{\sigma^2_{\hat{\beta}_i}} \le \beta_i \le \hat{\beta}_i + t_{\alpha/2} \sqrt{\sigma^2_{\hat{\beta}_i}}, \quad i=0, 1, \ldots, k$$

Intervalo de confianza para β_0 con nivel 95% para el ejemplo

$1 - \alpha = 0.95,\ \nu = n-k-1 = 6-2-1 = 3 \Rightarrow t_{\alpha/2} = t_{0.025} = 3.182$ (Tabla T)

$$\hat{\beta}_0 - t_{\alpha/2}\sqrt{\sigma^2_{\hat{\beta}_0}} \le \beta_0 \le \hat{\beta}_0 + t_{\alpha/2}\sqrt{\sigma^2_{\hat{\beta}_0}}$$

$$-134.071 - 3.188\sqrt{1609.33} \le \beta_0 \le -134.071 + 3.188\sqrt{1609.33}$$

$$-261.72 \le \beta_0 \le -6.4204$$

12.10.3 PRUEBA DE HIPÓTESIS

Parámetro: β_i , i = 0, 1, ..., k

Estimador: $\hat{\beta}_i$, i = 0, 1, ..., k

1) **Ho**: $\beta_i = b_0$ (Algún valor especificado para el parámetro β_i)

2) **Ha:** $\beta_i < b_0$ ó $\beta_i > b_0$ ó $\beta_i \neq b_0$

3) α nivel de significancia de la prueba

4) Estadístico de prueba

$$t = \frac{\hat{\beta}_i - b_0}{\sqrt{\sigma^2_{\hat{\beta}_i}}}$$, tiene distribución **t** con $\nu = n - k - 1$ grados de libertad

i = 0, 1, ..., k

Si se especifica el nivel de significancia α se define la región de rechazo de H_0

Ha: $\beta_i < b_0$	$t < -t_\alpha$
Ha: $\beta_i > b_0$	$t > t_\alpha$
Ha: $\beta_i \neq b_0$	$t<-t_{\alpha/2} \vee t > t_{\alpha/2}$

Es importante probar la hipótesis **Ho:** $\beta_i = 0$ individualmente con cada parámetro β_i.
En caso de que se pueda rechazar **Ho,** se puede concluir que la variable contribuye significativamente a la respuesta. Caso contrario, la variable es redundante y puede eliminarse del modelo.

Prueba con 5% de significancia que $\beta_2 \neq 0$. (En el ejemplo se prueba si la variable X_2, porcentaje de asistencia, contribuye significativamente al modelo)

Ho: $\beta_2 = 0$

Ha: $\beta_2 \neq 0$

$\alpha = 0.05$

$\nu = n - k - 1 = 3,\quad t_{\alpha/2} = t_{0.025} = 3.182$ (Tabla T)

Región de rechazo de **Ho:** $t < -3.182$ **o** $t > 3.182$

Cálculo del estadístico de prueba

$$t = \frac{\hat{\beta}_2 - 0}{\sqrt{\sigma^2_{\hat{\beta}_2}}} = \frac{1.4437 - 0}{\sqrt{0.17937}} = 3.4088,$$ **t** cae en la región de rechazo

Decisión: Se rechaza **Ho** $\Rightarrow$ el aporte de X_2 al modelo si es significativo

12.11 PRUEBA DE LA NORMALIDAD DEL ERROR

Se puede usar la prueba **K-S** para probar la suposición de normalidad de los errores

Prueba de Kolmogorov-Smirnov con 5% de significancia para la normalidad del error con los datos del ejemplo

Ho: $\varepsilon \sim N(0, \sigma^2)$ (Distribución normal con media **0** y varianza σ^2)
Ha: ⌉**Ho**
$\alpha = 0.05$

Estadístico de prueba
$D_n = \max| S_n(x_i) - F_0(x_i)|$ (Para este ejemplo x_i son los valores e_i)

Región de rechazo de Ho
$\alpha = 0.05$, $n = 6 \Rightarrow D_{0.05} = 0.521$ (Tabla K-S)

Rechazar H_0 si $D_n > 0.521$
$\varepsilon_i \cong e_i = y_i - \hat{y}_i$, $i = 1, 2, .., 6$

$\hat{y} = -134.07 + 1.4888\, x_1 + 1.4437 x_2$ (Modelo de mínimos cuadrados obtenido)
$x_1 = 67$, $x_2 = 75 \Rightarrow \hat{y}_1 = -134.07 + 14888(67) + 1.4437(75) = 73.9571$
$e_1 = y_1 - \hat{y}_1 = 80 - 73.9571 = 6.0429$, etc.

$$\begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \end{bmatrix} = \begin{bmatrix} 6.0429 \\ 1.6866 \\ -2.1121 \\ -5.0878 \\ -4.0562 \\ 3.5294 \end{bmatrix}$$

Modelo propuesto $e \sim N(0, \sigma^2)$ (Aproximadamente)

$F_0(x_i) = F_0(e_i) = P\left(Z < \frac{e_i - 0}{\sigma}\right)$ Distribución normal estándar acumulada

$\sigma^2 \cong S^2 = 32.861 \Rightarrow S = 5.7325$

$F_0(x_1) = F_0(-5.0878) = P\left(Z < \frac{-5.0878 - 0}{5.7325}\right) = 0.1874$, etc (Datos **e** ordenados)

Tabulación de resultados con la notación $x_i = e_i$

i	x_i (ordenados)	$S_n(x_i)$	$F_0(x_i)$	$\|S_n(x_i) - F_0(x_i)\|$
1	-5.0878	1/6 = 0.1666	0.1874	0.0207
2	-4.0562	2/6 = 0.3333	0.2396	0.0937
3	-2.1121	3/6 = 0.5	0.3563	0.1437
4	1.6866	4/6 = 0.6666	0.6157	0.0510
5	3.5294	5/6 = 0.8333	0.7310	0.1023
6	6.0401	6/6 = 1	0.8540	**0.1460**

$D_n = \max| S_n(x_i) - F_0(x_i)| = 0.1460$

Conclusión: D_n no cae en la región de rechazo, por lo tanto no se puede rechazar **Ho**

12.12 EJERCICIOS

Se realizó un estudio del desgaste de un rodamiento (**Y**), y su relación con la viscosidad del aceite ($\mathbf{X_1}$) y la carga que soporta ($\mathbf{X_2}$), obteniéndose los siguientes datos, en las unidades que correspondan:

X_1	X_2	Y
1.6	**8.51**	**19.3**
15.5	**8.16**	**23.0**
22.0	**10.58**	**17.2**
43.0	**12.01**	**91.0**

Analice el modelo de regresión lineal múltiple propuesto:

$$\mathbf{Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon, \qquad \varepsilon_i \sim N(0, \sigma^2)}$$

a) Dibuje un diagrama de dispersión **Y vs.** $\mathbf{X_1}$ y **Y vs.** $\mathbf{X_2}$
b) Escriba la matriz de diseño y con ella escriba el modelo propuesto en notación matricial
c) Use el modelo de mínimos cuadrados para encontrar los estimadores del modelo propuesto. Use la matriz de diseño en sus cálculos
d) Use el modelo para pronosticar el desgaste cuando la viscosidad sea **25** y la carga **10.0**
e) Calcule **SCT, SCR, SCE** y escriba la Tabla ANOVA
f) Pruebe con 5% de significancia la dependencia lineal del modelo propuesto
g) Encuentre el coeficiente de determinación e interprete su significado.
h) Calcule una estimación de la variancia
i) Encuentre la matriz de variancia-covariancia
j) Calcule la varianza de los estimadores del modelo de mínimos cuadrados
k) Encuentre un intervalo de confianza de **95%** para cada parámetro
l) Pruebe con **5%** de significancia si el aporte de cada variable $\mathbf{X_1}$, $\mathbf{X_2}$ al modelo es significativo
m) Pruebe la normalidad del error con **5%** de significancia mediante la prueba de Kolmogorov-Smirnov

MATLAB

Regresión lineal múltiple usando notación matricial

```
>> x=[1 67 75; 1 65 78; 1 78 79; 1 60 83; 1 64 65; 1 61 76]
x =
    1   67   75
    1   65   78
    1   78   79
    1   60   83
    1   64   65
    1   61   76
```

Matriz de diseño **X**

```
>> y=[ 80; 77;  94;  70;  51;  70]
  y =
   80
   77
   94
   70
   51
   70
```

Vector de observaciones

```
>> [b, bint, e, eint, stats] = regress(y, x, 0.05)
```

Regresión lineal simple $\alpha = 0.05$

```
b =
 -134.0719
    1.4888
    1.4437
```

Coeficientes β_0 , β_1 , β_2 del modelo de mínimos cuadrados

```
bint =
 -261.7405   -6.4034
    0.2297    2.7480
    0.0959    2.7916
```

Intervalos de confianza para β_0 , β_1 , β_2

```
e =
    6.0401
    1.6866
   -2.1121
   -5.0878
   -4.0562
    3.5294
```

Vector de residuales

```
 stats =
   0.9019   13.7968    0.0307
```

Coeficiente de determinación R^2, valor del estadístico **F,** valor **p** de la prueba **F**

Uso del modelo de mínimos cuadrados

```
>> yp=b(1)+b(2)*75+b(3)*80
   yp =
        93.0893
```

Evaluar el modelo con $x_1 = 75$, $x_2 = 80$

Matriz de correlación lineal de los datos de la muestra

```
>> cx1y =corrcoef(x(:,2),y)
   cx1y =
         1.0000    0.7226
         0.7226    1.0000
```

Correlación lineal entre x_1 y **y**

r = 0.7226 (correlación positiva débil)

```
>> cx2y=corrcoef(x(:,3),y)
   cx2y =
         1.0000    0.6626
         0.6626    1.0000
```

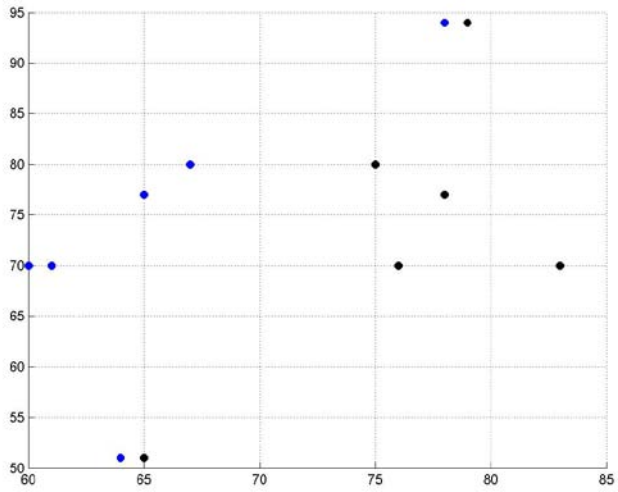
Correlación lineal entre x_2 y **y**

r = 0.6626 (correlación positiva débil)

Gráficos de dispersión recta de regresión

```
>> clf
>> scatter(x(:,2),y,'b','filled'),grid on
>> scatter(x(:,3),y,'k','filled'),grid on
```

Gráfico de dispersión x_1 y **y**
Gráfico de dispersión x_1 y **y**



Prueba de la normalidad del error de los residuales

```
>> sce =sum(e.^2)
    sce =
          98.5830
>> s2 =sce/3
    s2 =
         32.8610
>> t=sort(e);
>> f=normcdf(t, 0, sqrt(s2));
>> [h,p,ksstat,vc]=kstest(t,[t f ], 0.05,0)
   h =
        0
   p =
        0.9700
   ksstat =
             0.1874
   vc =
         0.5193
```

Suma de los cuadrados de residuales

Estimación de la varianza S^2

Residuales ordenados

Modelo a probar $e_i \sim N(0, \sigma^2)r$

Prueba **K-S**, $\alpha = 0.05$

No se puede rechazar el modelo

Valor **p** de la prueba

Valor del estadístico de prueba

Valor crítico de la región de rechazo

Matriz de varianzas y covarianzas de los estimadores β_i

```
>> format long
>> mvc = inv(x' *x)*s2
   mvc =
 1.0e+003 *
  1.60933261666704  -0.00946526866468  -0.01290428413874
 -0.00946526866468   0.00015654447216  -0.00001106020727
 -0.01290428413874  -0.00001106020727   0.00017937387435
```

Para visualizar con mayor precisión
MVC Usando notación matricial
La diagonal contiene los valores $V(\beta_i)$

$V(\beta_0)$ = 1609.3
$V(\beta_1)$ = 0.1565
$V(\beta_2)$ = 0.1793

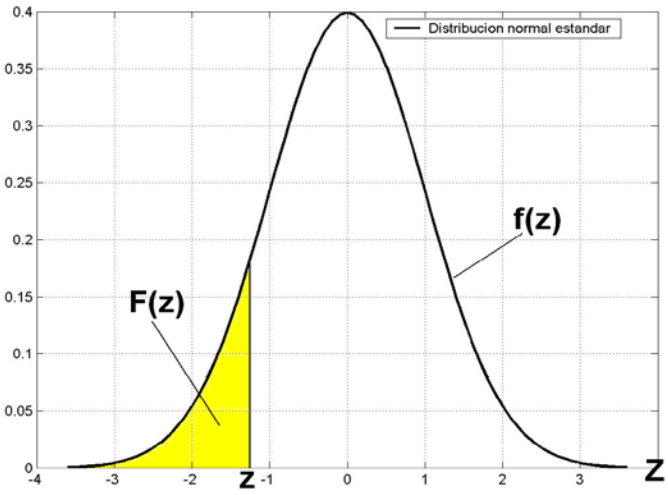
ALFABETO GRIEGO

En la primera columna está el símbolo griego en tipo de letra mayúscula
En la columna central está el símbolo griego en tipo de letra minúscula
En la tercera columna está el nombre en español del símbolo griego

Α	α	alfa
Β	β	beta
Γ	γ	gama
Δ	δ	delta
Ε	ε	épsilon
Ζ	ζ	zeta
Η	η	eta
Θ	θ	theta
Ι	ι	iota
Κ	κ	kappa
Λ	λ	lambda
Μ	μ	mu
Ν	ν	nu
Ξ	ξ	xi
Ο	ο	ómicron
Π	π	pi
Ρ	ρ	rho
Σ	σ	sigma
Τ	τ	tau
Ψ	υ	úpsilon
Φ	ϕ	fi
Χ	χ	ji
Ψ	ψ	psi
Ω	ω	omega

DISTRIBUCIÓN NORMAL ESTÁNDAR

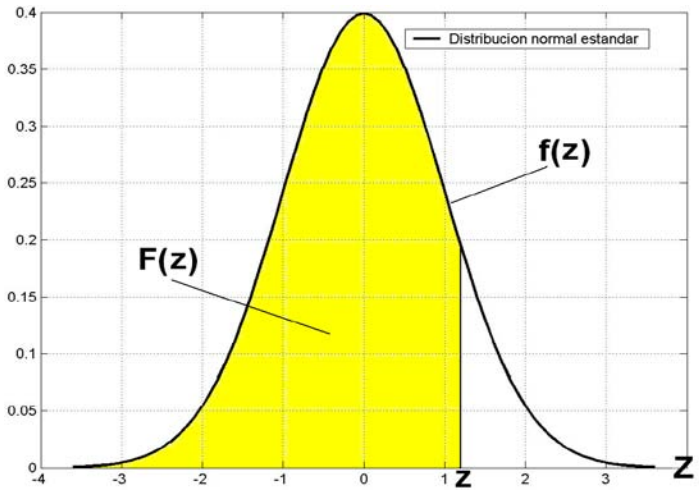
PROBABILIDAD ACUMULADA F(Z), Z ≤ 0



Z	0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.00
-3.5	0.000165	0.000172	0.000179	0.000185	0.000193	0.000200	0.000208	0.000216	0.000224	0.000233
-3.4	0.000242	0.000251	0.000260	0.000270	0.000280	0.000291	0.000302	0.000313	0.000325	0.000337
-3.3	0.000350	0.000362	0.000376	0.000390	0.000404	0.000419	0.000434	0.000450	0.000467	0.000483
-3.2	0.000501	0.000519	0.000538	0.000557	0.000577	0.000598	0.000619	0.000641	0.000664	0.000687
-3.1	0.000711	0.000736	0.000762	0.000789	0.000816	0.000845	0.000874	0.000904	0.000935	0.000968
-3.0	0.001001	0.001035	0.001070	0.001107	0.001144	0.001183	0.001223	0.001264	0.001306	0.001350
-2.9	0.001395	0.001441	0.001489	0.001538	0.001589	0.001641	0.001695	0.001750	0.001807	0.001866
-2.8	0.001926	0.001988	0.002052	0.002118	0.002186	0.002256	0.002327	0.002401	0.002477	0.002555
-2.7	0.002635	0.002718	0.002803	0.002890	0.002980	0.003072	0.003167	0.003264	0.003364	0.003467
-2.6	0.003573	0.003681	0.003793	0.003907	0.004025	0.004145	0.004269	0.004396	0.004527	0.004661
-2.5	0.004799	0.004940	0.005085	0.005234	0.005386	0.005543	0.005703	0.005868	0.006037	0.006210
-2.4	0.006387	0.006569	0.006756	0.006947	0.007143	0.007344	0.007549	0.007760	0.007976	0.008198
-2.3	0.008424	0.008656	0.008894	0.009137	0.009387	0.009642	0.009903	0.010170	0.010444	0.010724
-2.2	0.011011	0.011304	0.011604	0.011911	0.012224	0.012545	0.012874	0.013209	0.013553	0.013903
-2.1	0.014262	0.014629	0.015003	0.015386	0.015778	0.016177	0.016586	0.017003	0.017429	0.017864
-2.0	0.018309	0.018763	0.019226	0.019699	0.020182	0.020675	0.021178	0.021692	0.022216	0.022750
-1.9	0.023295	0.023852	0.024419	0.024998	0.025588	0.026190	0.026803	0.027429	0.028067	0.028717
-1.8	0.029379	0.030054	0.030742	0.031443	0.032157	0.032884	0.033625	0.034379	0.035148	0.035930
-1.7	0.036727	0.037538	0.038364	0.039204	0.040059	0.040929	0.041815	0.042716	0.043633	0.044565
-1.6	0.045514	0.046479	0.047460	0.048457	0.049471	0.050503	0.051551	0.052616	0.053699	0.054799
-1.5	0.055917	0.057053	0.058208	0.059380	0.060571	0.061780	0.063008	0.064256	0.065522	0.066807
-1.4	0.068112	0.069437	0.070781	0.072145	0.073529	0.074934	0.076359	0.077804	0.079270	0.080757
-1.3	0.082264	0.083793	0.085343	0.086915	0.088508	0.090123	0.091759	0.093418	0.095098	0.096801
-1.2	0.098525	0.100273	0.102042	0.103835	0.105650	0.107488	0.109349	0.111233	0.113140	0.115070
-1.1	0.117023	0.119000	0.121001	0.123024	0.125072	0.127143	0.129238	0.131357	0.133500	0.135666
-1.0	0.137857	0.140071	0.142310	0.144572	0.146859	0.149170	0.151505	0.153864	0.156248	0.158655
-0.9	0.161087	0.163543	0.166023	0.168528	0.171056	0.173609	0.176185	0.178786	0.181411	0.184060
-0.8	0.186733	0.189430	0.192150	0.194894	0.197662	0.200454	0.203269	0.206108	0.208970	0.211855
-0.7	0.214764	0.217695	0.220650	0.223627	0.226627	0.229650	0.232695	0.235762	0.238852	0.241964
-0.6	0.245097	0.248252	0.251429	0.254627	0.257846	0.261086	0.264347	0.267629	0.270931	0.274253
-0.5	0.277595	0.280957	0.284339	0.287740	0.291160	0.294599	0.298056	0.301532	0.305026	0.308538
-0.4	0.312067	0.315614	0.319178	0.322758	0.326355	0.329969	0.333598	0.337243	0.340903	0.344578
-0.3	0.348268	0.351973	0.355691	0.359424	0.363169	0.366928	0.370700	0.374484	0.378281	0.382089
-0.2	0.385908	0.389739	0.393580	0.397432	0.401294	0.405165	0.409046	0.412936	0.416834	0.420740
-0.1	0.424655	0.428576	0.432505	0.436441	0.440382	0.444330	0.448283	0.452242	0.456205	0.460172
-0.0	0.464144	0.468119	0.472097	0.476078	0.480061	0.484047	0.488033	0.492022	0.496011	0.500000

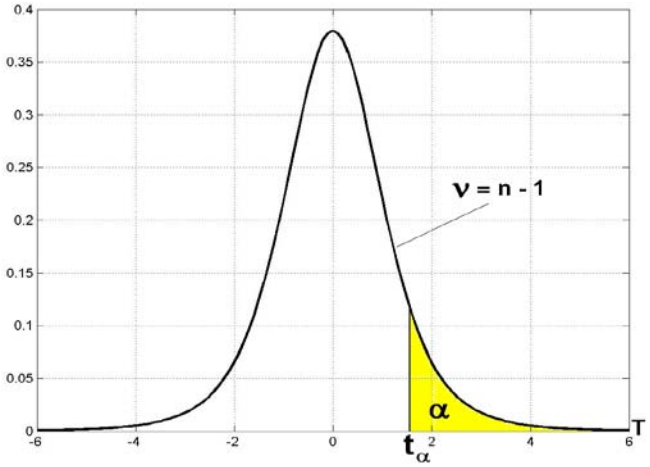
DISTRIBUCIÓN NORMAL ESTÁNDAR

PROBABILIDAD ACUMULADA F(Z), Z ≥ 0



Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.500000	0.503989	0.507978	0.511967	0.515953	0.519939	0.532922	0.527903	0.531881	0.535856
0.1	0.539828	0.543795	0.547758	0.551717	0.555760	0.559618	0.563559	0.567495	0.571424	0.575345
0.2	0.579260	0.583166	0.587064	0.590954	0.594835	0.598706	0.602568	0.606420	0.610261	0.614092
0.3	0.617911	0.621719	0.625516	0.629300	0.633072	0.636831	0.640576	0.644309	0.648027	0.651732
0.4	0.655422	0.659097	0.662757	0.666402	0.670031	0.673645	0.677242	0.680822	0.684386	0.687933
0.5	0.691462	0.694974	0.698468	0.701944	0.705401	0.708840	0.712260	0.715661	0.719043	0.722405
0.6	0.725747	0.729069	0.732371	0.735653	0.738914	0.742154	0.745373	0.748571	0.751748	0.754903
0.7	0.758036	0.761148	0.764238	0.767305	0.770350	0.773373	0.776373	0.779350	0.782305	0.785236
0.8	0.788145	0.791030	0.793892	0.796731	0.799546	0.802338	0.805106	0.807850	0.810570	0.813267
0.9	0.815940	0.818589	0.821214	0.823815	0.826391	0.828944	0.831472	0.833977	0.836457	0.838913
1.0	0.841345	0.843752	0.846136	0.848495	0.850830	0.853141	0.855428	0.857690	0.859929	0.862143
1.1	0.864334	0.866500	0.868643	0.870762	0.872857	0.874928	0.876976	0.878999	0.881000	0.882977
1.2	0.884930	0.886860	0.888767	0.890651	0.892512	0.894350	0.896165	0.897958	0.899727	0.901475
1.3	0.903199	0.904902	0.906582	0.908241	0.909877	0.911492	0.913085	0.914657	0.916207	0.917736
1.4	0.919243	0.920730	0.922196	0.923641	0.925066	0.926471	0.927855	0.929219	0.930563	0.931888
1.5	0.933193	0.934478	0.935744	0.936992	0.938220	0.939429	0.940620	0.941792	0.942947	0.944083
1.6	0.945201	0.946301	0.947384	0.948449	0.949497	0.950529	0.951543	0.952540	0.953521	0.954486
1.7	0.955435	0.956367	0.957284	0.958185	0.959071	0.959941	0.960796	0.961636	0.962462	0.963273
1.8	0.964070	0.964852	0.965621	0.966375	0.967116	0.967843	0.968557	0.969258	0.969946	0.970621
1.9	0.971283	0.971933	0.972571	0.973197	0.973810	0.974412	0.975002	0.975581	0.976148	0.976705
2.0	0.977250	0.977784	0.978308	0.978822	0.979325	0.979818	0.980301	0.980774	0.981237	0.981691
2.1	0.982136	0.982571	0.982997	0.983414	0.983823	0.984222	0.984614	0.984997	0.985371	0.985738
2.2	0.986097	0.986447	0.986791	0.987126	0.987455	0.987776	0.988089	0.988396	0.988696	0.988989
2.3	0.989276	0.989556	0.989830	0.990097	0.990358	0.990613	0.990863	0.991106	0.991344	0.991576
2.4	0.991802	0.992024	0.992240	0.992451	0.992656	0.992857	0.993053	0.993244	0.993431	0.993613
2.5	0.993790	0.993963	0.994132	0.994297	0.994457	0.994614	0.994766	0.994915	0.995060	0.995201
2.6	0.995339	0.995473	0.995604	0.995731	0.995855	0.995975	0.996093	0.996207	0.996319	0.996427
2.7	0.996533	0.996636	0.996736	0.996833	0.996928	0.997020	0.997110	0.997197	0.997282	0.997365
2.8	0.997445	0.997523	0.997599	0.997673	0.997744	0.997814	0.997882	0.997948	0.998012	0.998074
2.9	0.998134	0.998193	0.998250	0.998305	0.998359	0.998411	0.998462	0.998511	0.998559	0.998605
3.0	0.998650	0.998694	0.998736	0.998777	0.998817	0.998856	0.998893	0.998930	0.998965	0.998999
3.1	0.999032	0.999065	0.999096	0.999126	0.999155	0.999184	0.999211	0.999238	0.999264	0.999289
3.2	0.999313	0.999336	0.999359	0.999381	0.999402	0.999423	0.999443	0.999462	0.999481	0.999499
3.3	0.999517	0.999533	0.999550	0.999566	0.999581	0.999596	0.999610	0.999624	0.999638	0.999650
3.4	0.999663	0.999675	0.999687	0.999698	0.999709	0.999720	0.999730	0.999740	0.999749	0.999758
3.5	0.999767	0.999776	0.999784	0.999792	0.999800	0.999807	0.999815	0.999821	0.999828	0.999835

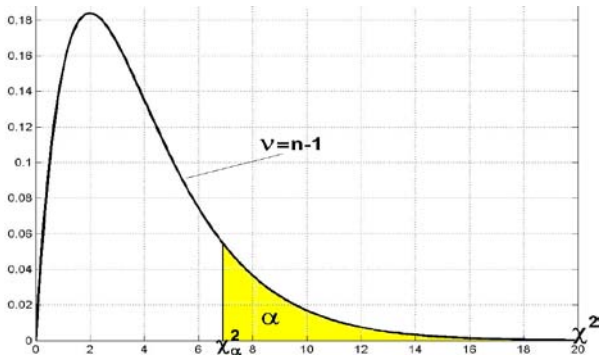
TABLA DE LA DISTRIBUCIÓN T



$t_\alpha \Rightarrow P(T \geq t_\alpha) = \alpha$

ν \ α	.40	.25	.10	.05	.025	.01	.005	.0025	.001	.0005
1	.325	1.000	3.078	6.314	12.706	31.821	63.657	127.320	318.310	636.620
2	.289	.816	1.886	2.920	4.303	6.965	9.925	14.089	23.326	31.598
3	.277	.765	1.638	2.353	3.182	4.541	5.841	7.453	10.213	12.924
4	.271	.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	.267	.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	.265	.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	.263	.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	.262	.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	.261	.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	.260	.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	.260	.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	.259	.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	.259	.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	.258	.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	.258	.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	.258	.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	.257	.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	.257	.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	.257	.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	.257	.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	.257	.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	.256	.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	.256	.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.767
24	.256	.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	.256	.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	.256	.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	.256	.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	.256	.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	.256	.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	.256	.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
∞	.253	.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

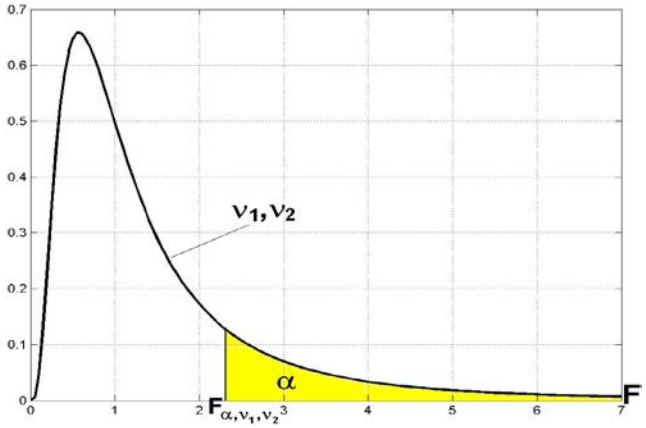
TABLA DE LA DISTRIBUCIÓN JI-CUADRADO



$$\chi^2_\alpha \Rightarrow P(\chi^2 \geq \chi^2_\alpha) = \alpha$$

ν \ α	.995	.990	.975	.950	.900	.500	.100	.050	.025	.010	.005
1	.00003	.0001	.0009	.0039	.02	.45	2.71	3.84	5.02	6.63	7.88
2	.01	.02	.05	.10	.21	1.39	4.61	5.99	7.38	9.21	10.60
3	.07	.11	.22	.35	.58	2.37	6.25	7.81	9.35	11.34	12.84
4	.21	.30	.48	.71	1.06	3.36	7.78	9.49	11.14	13.28	14.86
5	.41	.55	.83	1.15	1.61	4.35	9.24	11.07	12.83	15.09	16.75
6	.68	.87	.24	1.64	2.20	5.35	10.65	12.59	14.45	16.81	18.55
7	.99	.24	.69	2.17	2.83	6.35	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	7.34	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	8.34	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	9.34	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	10.34	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	11.34	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	12.34	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	13.34	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.27	7.26	8.55	14.34	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	15.34	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	16.34	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.87	17.34	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	18.34	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	19.34	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	20.34	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	21.34	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	22.34	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	23.34	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	24.34	34.28	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	25.34	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	26.34	36.74	40.11	43.19	46.96	49.65
28	12.46	13.57	15.31	16.93	18.94	27.34	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	28.34	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	29.34	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	39.34	51.81	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	49.33	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	59.33	74.40	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	55.33	69.33	85.53	90.53	95.02	100.42	104.22
80	51.17	53.54	57.15	60.39	64.28	79.33	96.58	101.88	106.63	112.33	116.32
90	59.20	61.75	65.65	69.13	73.29	89.33	107.57	113.14	118.14	124.12	128.30
100	67.33	70.06	74.22	77.93	82.36	99.33	118.50	124.34	129.56	135.81	140.17

TABLA DE LA DISTRIBUCIÓN F



$$F_{\alpha,\nu_1,\nu_2} \Rightarrow P(F > F_{\alpha,\nu_1,\nu_2}) = \alpha \qquad \alpha = 0.05$$

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.55	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

TABLA PARA LA PRUEBA KOLMOGOROV - SMIRNOV (K-S)

α: Nivel de significancia
n: Tamaño de la muestra

Valores críticos $D_{\alpha, n}$

n \ α	0.20	0.15	0.10	0.05	0.01
1	0.900	0.925	0.950	0.875	0.995
2	0.684	0.726	0.776	0.842	0.929
3	0.565	0.597	0.642	0.708	0.828
4	0.494	0.525	0.564	0.624	0.733
5	0.446	0.474	0.510	0.565	0.669
6	0.410	0.436	0.470	0.521	0.618
7	0.381	0.405	0.438	0.486	0.577
8	0.358	0.381	0.411	0.457	0.543
9	0.339	0.360	0.388	0.432	0.514
10	0.322	0.342	0.368	0.410	0.490
11	0.307	0.326	0.352	0.391	0.468
12	0.295	0.313	0.338	0.375	0.450
13	0.284	0.302	0.325	0.361	0.433
14	0.274	0.292	0.314	0.349	0.418
15	0.266	0.283	0.304	0.338	0.404
16	0.258	0.274	0.295	0.328	0.392
17	0.250	0.266	0.286	0.318	0.381
18	0.244	0.259	0.278	0.309	0.371
19	0.237	0.252	0.272	0.301	0.363
20	0.231	0.246	0.264	0.294	0.356
25	0.210	0.220	0.240	0.270	0.320
30	0.190	0.200	0.220	0.240	0.290
35	0.180	0.190	0.201	0.230	0.270
Mayor a 35	$\frac{1.07}{\sqrt{n}}$	$\frac{1.14}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$

DISTTOOL

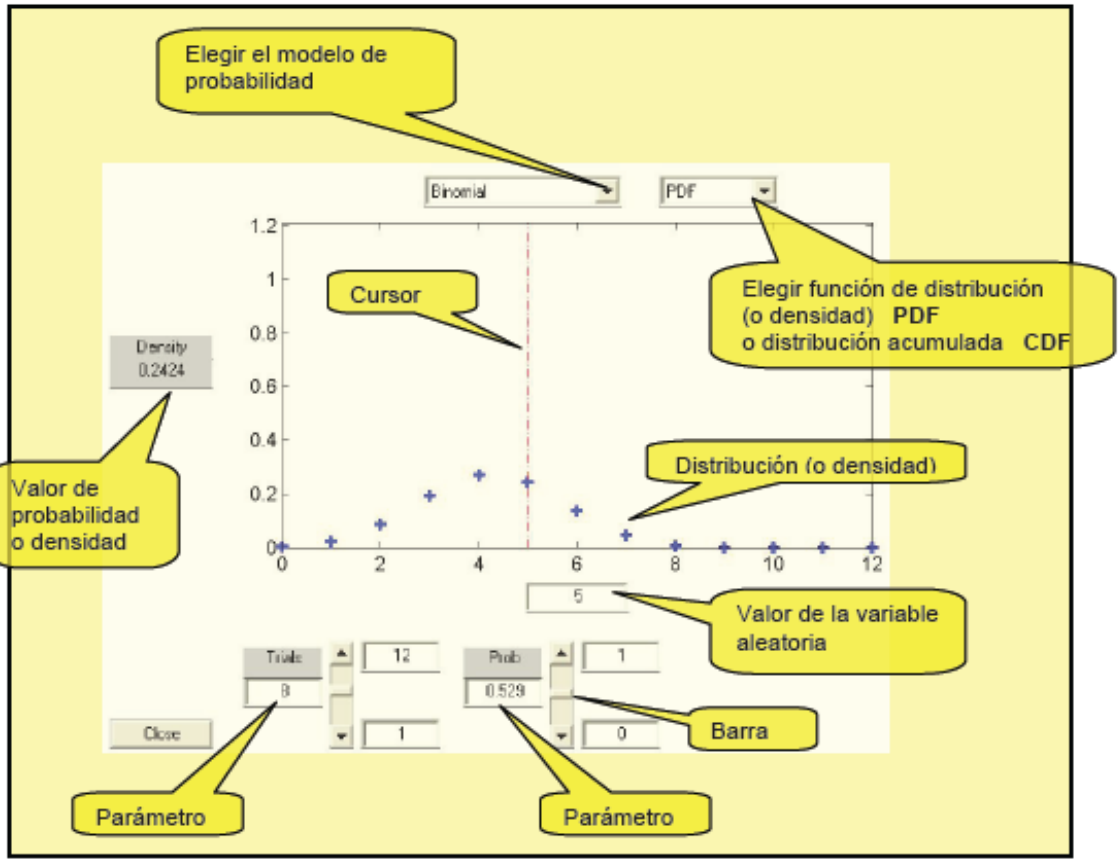
Instrumento computacional gráfico interactivo disponible en **MATLAB** para entender visualmente algunas propiedades de las distribuciones de probabilidad más importantes.

DISTTOOL crea interactivamente el gráfico de la distribución de probabilidad, o densidad de Probabilidad, y la distribución acumulada para los siguientes modelos:

Beta	Binomial	Ji-cuadrado
Uniforme discreta	Exponencial	F
Gamma	Geométrica	Lognormal
Binomial Negativa	F no centrada	T no centrada
Ji-cuadrado no centrada	Normal	Poisson
Rayleigh	T	Uniforme continua
Weibull		

Se pueden cambiar los parámetros escribiendo sus valores o moviendo un cursor sobre el gráfico o barras de desplazamiento. Se pueden obtener valores de la distribución o de probabilidad moviendo una línea de referencia sobre el gráfico

Para activar este utilitario digite **disttool** en la ventana de comandos de **MATLAB**



RANDTOOL

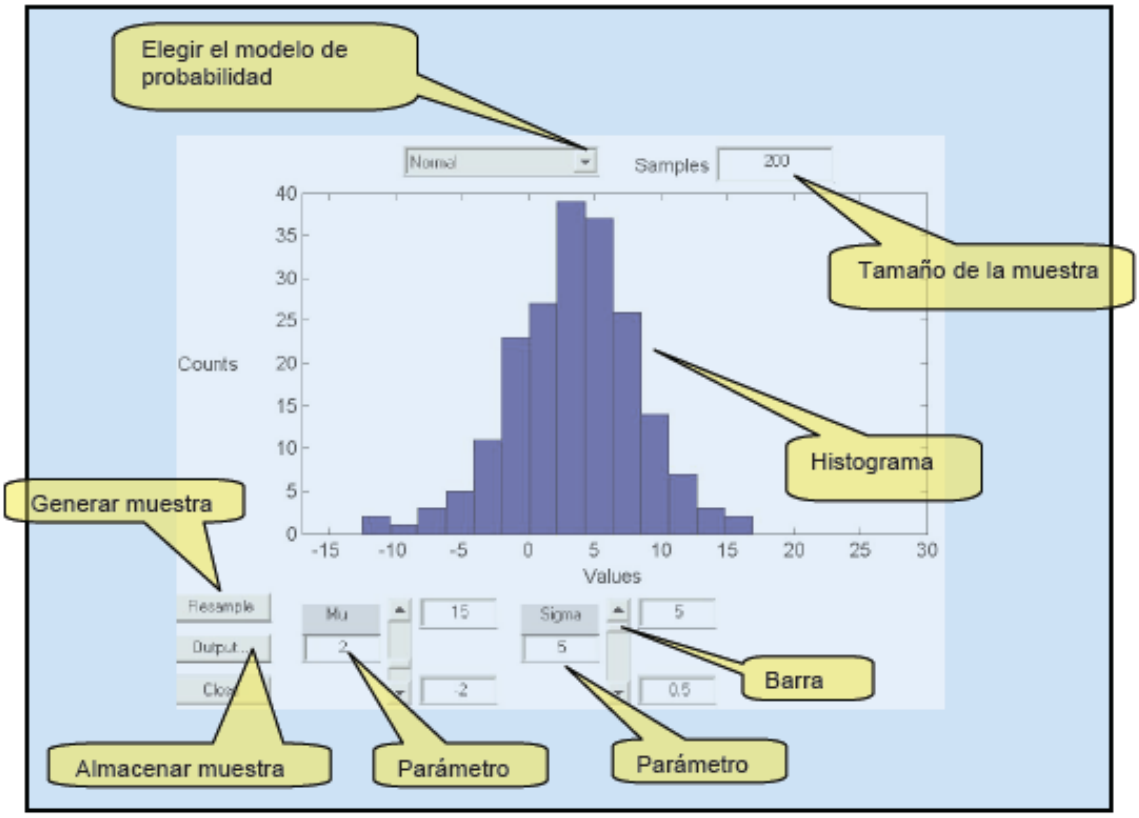
Instrumento computacional gráfico interactivo disponible en **MATLAB** para obtener muestras aleatorias de las distribuciones de probabilidad más importantes.

RANDTOOL crea un histograma con los datos de las muestras aleatorias generadas para los siguientes modelos.

Beta	**Binomial**	**Ji-cuadrado**
Uniforme discreta	**Exponencial**	**F**
Gamma	**Geométrica**	**Lognormal**
Binomial Negativa	**F no centrada**	**T no centrada**
Ji-cuadrado no centrada	**Normal**	**Poisson**
Rayleigh	**T**	**Uniforme continua**
Weibull		

Se pueden cambiar los parámetros escribiendo sus valores o moviendo barras de desplazamiento. Se puede especificar el tamaño de la muestra y se puede almacenar la muestra escribiendo una variable para ser usada desde la ventana de comandos de MATLAB.

Para activar este utilitario digite **randtool** en la ventana de comandos de **MATLAB**



BIBLIOGRAFÍA

ESTADÍSTICA

Canavos, G. C. *Probabilidad y Estadística Aplicaciones y Métodos*, México: McGraw-Hill Interamericana de México, S. A.

Castro A. B. *Probabilidades y Estadística Básicas*, Quito: Escuela Politécnica Nacional

Freund, J. E. y Walpole R. E. *Estadística Matemática con Aplicaciones*, 4a. ed. México: Prentice-Hall Hispanoamericana, S. A.

Hines, W. W. y Montgomery D. C. *Probabilidad y Estadística para Ingeniería*, 3a. ed. México: Compañia Editorial Continental

Mendenhall W. *Introduction to Probability and Statistics*, 3d. ed. California: Duxbury Press

Miller, I. R., Freund J. E. y Johnson R. *Probabilidad y Estadística para Ingenieros* 4a. ed. México: Prentice-Hall Hispanoamericana, S. A.

Montgomery D. C. y Runger G. C. *Probabilidad y Estadística Aplicadas a la Ingeniería*, 2a. ed. México: Editorial Limusa S. A.

Walpole, R. E. y Myers, R. H. *Probabilidad y Estadística para Ingenieros*, 3a. ed. México: McGraw-Hill Interamericana de México, S. A.

COMPUTACIÓN

The MathWorks, Inc. *Statistics Toolbox for use with MATLAB User`s Guide*

The MathWorks, Inc. *Using MATLAB Computation, Visualization, Programming*

Pérez López C. *MATLAB y sus Aplicaciones en las Ciencias y la Ingeniería*, Madrid: Pearson Educación, S. A.

Rodríguez Ojeda L. *MATLAB Conceptos Básicos y Programación*, Tutorial, ICM ESPOL