Manual de Estadística

Nombre de la materia: Probabilidad Y Estadística

CUERPO COLEGIADO DE DIRECTORES Y PROFESORES

abril 2016

PROBABILIDAD Y ESTADISTICA

Competencia de la asignatura: Dirigir el soporte técnico de sistemas mecánicos considerando el diagnóstico y reparación para el óptimo funcionamiento del equipo

1.- Estadística Descriptiva

Competencia del Módulo: Plantear y solucionar problemas con base en los principios y teorías de física, química y matemáticas, a través del método científico para sustentar la toma de decisiones en los ámbitos científico y tecnológico.

Resultado de aprendizaje (Tarea integradora): El alumno elaborará un ensayo que contenga:

- Variable de estudio
- Diseño del muestreo
- Tabla de distribución de frecuencia
- Gráficos
- Medidas de tendencia central, localización y dispersión
- Interpretación de resultados

Unidad 1

1.- Introducción a la estadística.

1.1 Definir los conceptos de estadística, estadística descriptiva e inferencial y sus aplicaciones.

Estadística: Es una ciencia formal y una herramienta que estudia usos y análisis provenientes de una muestra representativa de datos, busca explicar las correlaciones y dependencias de un fenómeno físico o natural, de ocurrencia en forma aleatoria o condicional.
Es transversal a una amplia variedad de disciplinas, desde la física hasta las ciencias sociales, desde las ciencias de la salud hasta el control de calidad.
Estadística descriptiva: Se dedica a la descripción, visualización y resumen de datos originados a partir de los fenómenos de estudio. Los datos pueden ser resumidos numérica o gráficamente. Ejemplos básicos de parámetros son:
la media y la desviación estándar. Algunos ejemplos gráficos son: histograma, pirámide, gráfico circular, entre otros.

La directora de la escuela San Jorge se propone averiguar cuál es el peso y la altura promedio de los niños de séptimo grado de la institución para analizar

qué tipo de alimentos deberían servirse en el comedor.
Para cumplir con su objetivo, comienza a elaborar una tabla con las columnas **alumno, peso y altura**, completando los datos correspondientes.

Carlos Gómez – 35 kilogramos – 1,45 metros.
Damián Ramírez – 38 kilogramos – 1,43 metros.
Juan Lopresti – 46 kilogramos – 1,46 metros.
Marcos Elusorte – 53 kilogramos – 1,52metros.

Estadística inferencial: Se dedica a la generación de los modelos, inferencias y predicciones asociadas a los fenómenos en cuestión teniendo en cuenta la aleatoriedad de las observaciones. Se usa para modelar patrones en los datos y extraer inferencias acerca de la población bajo estudio. Estas inferencias pueden tomar la forma de respuestas a preguntas sí/no (prueba de hipótesis), estimaciones de unas características numéricas (estimación), pronósticos de futuras observaciones, descripciones de asociación (correlación) o modelamiento de relaciones entre variables (análisis de regresión). Otras técnicas de modelamiento incluyen análisis de varianza, series de tiempo y minería de datos.

1.2 Identificar los conceptos de estadística descriptiva:

1.2.1 Variable estadística

Variable estadística: Es una propiedad que puede fluctuar y cuya variación es susceptible de adoptar diferentes valores, los cuales pueden medirse u observarse. Las variables adquieren valor cuando se relacionan con otras variables, es decir, si forman parte de una hipótesis o de una teoría. En este caso se las denomina constructos o construcciones hipotéticas.

EJEMPLOS:

El estado civil, con las siguientes modalidades: soltero, casado, separado, divorciado y viudo.

Variable cualitativa ordinal o variable cuasi cuantitativa la nota en un examen: suspenso, aprobado, notable, sobresaliente.

Puesto conseguido en una prueba deportiva: 1º, 2º, 3º,...

Medallas de una prueba deportiva: oro, plata, bronce.

Variable cuantitativa

1.2.2 Datos: cualitativos, cuantitativos discretos y continuos

Cualitativos: son los que nos dan el enfoque o el punto de vista de la muestra o de una variable hacia lo que nos dice la investigación. Estos son igual de importantes, porque aunque no son numéricos, se pueden medir de ciertas maneras. Estos datos hay que ver como se recolectan, a través de entrevistas, observación, encuestas o algún método que sea válido para saber con certeza la tendencia de una población hacia lo que se quiera saber.

Ejemplos: 1) El restaurante Pasión Roja ofrece a sus clientes una variedad de platos preparados, en las ventas de una semana reportan lo siguiente datos, de una totalidad de 300 platos vendidos, los que se subdividen en:
Lomo a lo pobre: 40 platos.
Salmón a la mantequilla acompañado con puré picante: 35 platos.
Fettucini al huevo con salsa Alfredo: 20 platos.
Ceviche de atún ala peruana: 25 platos.
Carpaccio de res: 15 platos.
Ensalada cesar: 30 platos.
Suprema de ave a la plancha acompañada de papas duquesa: 25 platos
Nuggets de pollo acompañados con papas fritas: 20 platos.
Mousse de frutilla con decoración de chocolate: 30 platos.
Suspiro limeño: 20 platos.
Tiramisú: 25 platos.
Panqueque celestino: 15 platos.
A) defina la variable en estudio. ¿De qué tipo de variable se trata?
X: preferencia de sus clientes. Variable cualitativa.

B) Elabore una tabla de frecuencia que permita resumir los datos.

Plato preferido
Cantidad de platos
Lomo a lo pobre -40
Salmón a la mantequilla acompañado con puré picante-35
Fettucini al huevo con salsa Alfredo-20
Ceviche de atún a la peruana-25
Carpaccio de res-15
Ensalada cesar-30
Suprema de ave a la plancha acompañada de papas duquesa-25
Nuggets de pollo acompañados con papas fritas-20
Mousse de frutilla con decoración de chocolate-30
Suspiro limeño-20
Tiramisú-25
Panqueque celestino-15

C) ¿Cuál es plato más consumido por los clientes y a qué porcentaje corresponde del total de platos vendidos?
El plato más vendido es el lomo a lo pobre, el que representa un 13,3 % del total delas ventas de la semana.
D) ¿Cuántas personas decidieron comer suspiro limeño?
20 personas, las que representan un 7% del total de las ventas de la semana.
2) El restaurante Pasión Roja reunió la siguiente información sobre la asistencia mensual de clientes en el mes de mayo:
15
20
16
30
23
20
50
19
23
35
24
20
35
44
27
33
27
34
45
24
67
18
21
19
30
35
13
80

A)Calcular media o promedio con los datos de la tabla:
15+20+16+30+23+20+50+19+23+35+24+20+35+44+27+33+27+34+45+24
+67+18+21+19+30+35+ 13+80= 847
847/28= 30.25

El promedio de clientes en el restaurante es de 30.25 clientes por día aproximadamente.
B) Calcular la mediana con los datos de la tabla:
13-15-16-18-19-19-20-20-20-21-23-23-24-24-27-27-30-30-33-34-35-35-35-44-45-50-67-80=
24+27=51/2= 25,5
La mediana es 25,5 clientes.
C) Calcular la moda con los datos de la tabla:
13-15-16-18-19-19-20-20-20-21-23-23-24-24-27-27-30-30-33-34-35-35-35-44-45-50-67-80
La moda es 20 y 35 es un conjunto binomial.

Cuantitativos discretos:

Los cuales producen respuestas numéricas, pero en números enteros, generalmente producto de un conteo, no pueden tener valores intermedios en un rango, por ejemplo: número de empleados o número de puestos que ha ocupado una persona en una compañía, los cuales no pueden ser 450.3 empleados o 3 puestos y medio.

Ejemplo: Una variable cuantitativa discreta es aquella que obedece a una cantidad numérica exacta. Por ejemplo: 1.- el número de hijos en una familia, 2.- el número de carros de la familia, 3.-el número de exámenes en el semestre. 4.-el número de ojos en una persona 5.- El número de animales en una granja. Todos son enteros no puedes tener 0,5 hijos 0 4,6 vacas.

Cuantitativos continuo:

Que si puede adoptar cualquier valor numérico intermedio en un rango, generalmente producto de una medición, por ejemplo: edad de los empleados o sueldo de los ejecutivos, que puede ser medido de manera precisa, como una edad de 38 años, 6 meses y 18 días o un sueldo de Q. 4,529.33.

Ejemplo: Una variable Cuantitativa Continua por su parte es una variable numérica con expresiones decimales. Ejemplo: 1.- El peso de una fruta. 2.- El tiempo medido en una carrera 3.-La distancia de una carretera 4.-La intensidad de la corriente eléctrica 5.- El precio de los artículos en un supermercado. Saludos Espero te ayude... ;).

1.2.3 Población finita e infinita

Población finita: Constan de un número determinado de elementos, susceptible a ser contado. Ejemplo: Los empleados de una fábrica, elementos de un lote de producción, etc.

Población infinita: Tienen un número indeterminado de elementos, los cuales no pueden ser contados. Ejemplo: Los números naturales.

1.2.4 Muestra

Muestra: una muestra es un subconjunto de casos o individuos de una población estadística. En diversas aplicaciones interesa que una muestra sea una muestra representativa y para ello debe escogerse una técnica de muestreo adecuada que produzca una muestra aleatoria adecuada (contrariamente se obtiene una muestra sesgada cuyo interés y utilidad es más limitado dependiendo del grado de sesgo que presente).

Ejemplos:

1.- Población mexicana en general; muestra, población de mujeres mexicanas, menores de 35 años.

2.- Población de libros de una biblioteca; muestra, población de libros en la sección de histórica.

1.3 Clasificar datos cualitativos y cuantitativos.

Clasificar datos cualitativos: Datos cuantitativos son aquellos en los que se puede llevar un control estadístico...

Número de personas en una población

Edad promedio de las personas

Cantidad de personas de género femenino

Cantidad de género masculino

Cantidad de personas con trabajo remunerado.

Clasificar datos cuantitativos: Datos cualitativos aquellos más difíciles de manejar estadísticamente pero que pueden decir un poco más acerca de temas con muchas variables...

Porque las personas prefieren el automóvil sobre el transporte público.

Que motivos se tienen para dejar de fumar.
Porque prefieren cierta marca de cereal y no otra
¿Por qué se tiñe el pelo en color rubio?
¿Es más agradable la cerveza o el vino tinto?

2.- Población, muestra y muestreo.

2.1 Identificar los conceptos de:

2.1.1 Censo

- Es un recuento de individuos que conforman una población estadística, definida como un conjunto de elementos de referencia sobre el que se realizan las observaciones.

2.1.2 Parámetro

- Nombre dado a una característica global de una población. En general, un parámetro no es conocido. Por ejemplo, la edad promedio de una población de habitantes de una región.

2.1.3 Muestreo

- Selección de un conjunto de personas o cosas que se consideran representativos del grupo al que pertenecen, con la finalidad de estudiar o determinar las características del grupo.

2.1.4 Estadístico

- es una ciencia formal y una herramienta que estudia usos y análisis provenientes de una muestra representativa de datos, busca explicar las correlaciones y dependencias de un fenómeno físico o natural, de ocurrencia en forma aleatoria o condicional.

2.2 Clasificar las técnicas de muestreo:

2.2.1 Probabilístico:

- Los métodos de muestreo probabilísticos son aquellos que se basan en el principio de equiprobabilidad. Es decir, aquellos en los que todos los individuos tienen la misma probabilidad de ser elegidos para formar parte de una muestra y, consiguientemente, todas las posibles muestras de tamaño n tienen la misma probabilidad de ser seleccionadas. Sólo estos métodos de muestreo probabilísticos nos aseguran la representatividad de la muestra extraída y son, por tanto, los más recomendables.

2.2.1.1 Aleatorio simple

- El procedimiento empleado es el siguiente: 1) se asigna un número a cada individuo de la población y 2) a través de algún medio mecánico (bolas dentro de

una bolsa, tablas de números aleatorios, números aleatorios generados con una calculadora u ordenador, etc.) se eligen tantos sujetos como sea necesario para completar el tamaño de muestra requerido.

Este procedimiento, atractivo por su simpleza, tiene poca o nula utilidad práctica cuando la población que estamos manejando es muy grande.

2.2.1.2 Sistemático

- Este procedimiento exige, como el anterior, numerar todos los elementos de la población, pero en lugar de extraer n números aleatorios sólo se extrae uno. Se parte de ese número aleatorio i, que es un número elegido al azar, y los elementos que integran la muestra son los que ocupa los lugares i, i+k, i+2k, i+3k,...,i+(n-1)k, es decir se toman los individuos de k en k, siendo k el resultado de dividir el tamaño de la población entre el tamaño de la muestra: k= N/n. El número i que empleamos como punto de partida será un número al azar entre 1 y k.

El riesgo este tipo de muestreo está en los casos en que se dan periodicidades en la población ya que al elegir a los miembros de la muestra con una periodicidad constante (k) podemos introducir una homogeneidad que no se da en la población. Imaginemos que estamos seleccionando una muestra sobre listas de 10 individuos en los que los 5 primeros son varones y los 5 últimos mujeres, si empleamos un muestreo aleatorio sistemático con k=10 siempre seleccionaríamos o sólo hombres o sólo mujeres, no podría haber una representación de los dos sexos.

2.2.1.3 Estratificado

- Trata de obviar las dificultades que presentan los anteriores ya que simplifican los procesos y suelen reducir el error maestral para un tamaño dado de la muestra. Consiste en considerar categorías típicas diferentes entre sí (estratos) que poseen gran homogeneidad respecto a alguna característica (se puede estratificar, por ejemplo, según la profesión, el municipio de residencia, el sexo, el estado civil, etc.). Lo que se pretende con este tipo de muestreo es asegurarse de que todos los estratos de interés estarán representados adecuadamente en la muestra. Cada estrato funciona independientemente, pudiendo aplicarse dentro de ellos el muestreo aleatorio simple o el estratificado para elegir los elementos concretos que formarán parte de la muestra. En ocasiones las dificultades que plantean son demasiado grandes, pues exige un conocimiento detallado de la población. (Tamaño geográfico, sexos, edades,...).

La distribución de la muestra en función de los diferentes estratos se denomina afijación, y puede ser de diferentes tipos:

- Afijación Simple: A cada estrato le corresponde igual número de elementos muéstrales.

- Afijación Proporcional: La distribución se hace de acuerdo con el peso (tamaño) de la población en cada estrato.

- Afijación Óptima: Se tiene en cuenta la previsible dispersión de los resultados, de modo que se considera la proporción y la desviación típica. Tiene poca aplicación ya que no se suele conocer la desviación.

2.2.1.4 Conglomerado:

Los métodos presentados hasta ahora están pensados para seleccionar directamente los elementos de la población, es decir, que las unidades muéstrales son los elementos de la población.

En el muestreo por conglomerados la unidad muestral es un grupo de elementos de la población que forman una unidad, a la que llamamos conglomerado. Las unidades hospitalarias, los departamentos universitarios, una caja de determinado producto, etc., son conglomerados naturales. En otras ocasiones se pueden utilizar conglomerados no naturales como, por ejemplo, las urnas electorales. Cuando los conglomerados son áreas geográficas suele hablarse de "muestreo por áreas".

El muestreo por conglomerados consiste en seleccionar aleatoriamente un cierto número de conglomerados (el necesario para alcanzar el tamaño muestral establecido) e investigar después todos los elementos pertenecientes a los conglomerados elegidos.

2.2.2 No probabilístico

- A veces, para estudios exploratorios, el muestreo probabilístico resulta excesivamente costoso y se acude a métodos no probabilísticos, aun siendo conscientes de que no sirven para realizar generalizaciones (estimaciones inferenciales sobre la población), pues no se tiene certeza de que la muestra extraída sea representativa, ya que no todos los sujetos de la población tienen la misma probabilidad de ser elegidos. En general se seleccionan a los sujetos siguiendo determinados criterios procurando, en la medida de lo posible, que la muestra sea representativa.

En algunas circunstancias los métodos estadísticos y epidemiológicos permiten resolver los problemas de representatividad aun en situaciones de muestreo no probabilístico, por ejemplo los estudios de caso-control, donde los casos no son seleccionados aleatoriamente de la población.

2.3 Identificar el proceso del diseño de una muestra:

2.3.1 Tipo de variable

Una variable estadística es una propiedad que puede fluctuar y cuya variación es susceptible de adoptar diferentes valores, los cuales pueden medirse u

observarse. Las variables adquieren valor cuando se relacionan con otras variables, es decir, si forman parte de una hipótesis o de una teoría. En este caso se las denomina constructos o construcciones hipotéticas.

2.3.2 Tamaño de la muestra

El tamaño de la muestra es el número de sujetos que componen la muestra extraída de población.

2.3.3 Técnica de muestreo

Técnicas de muestreo

1. Muestreo probabilístico: es aquel en el que cada muestra tiene la misma probabilidad de ser elegida.

2. Muestreo intencional: en el que la persona que selecciona la muestra es quien procura que sea representativa, dependiendo de su intención u opinión, siendo por tanto la representatividad subjetiva.

3. Muestreo sin norma: se toma la muestra sin norma alguna, de cualquier manera, siendo la muestra representativa si la población es homogénea y no se producen sesgos de selección.

3.- Distribución de frecuencias y su representación gráfica

3.1 Identificar el concepto de datos agrupados y no agrupados.

DATOS AGRUPADOS

1.- su fin es resumir la información.

2.- generalmente, los elementos son de mayor tamaño, por lo cual requieren ser agrupados, esto implica: ordenar, clasificar y expresar los en una tabla de frecuencias.

3.- se agrupa a los datos, si se cuenta con 20 o más elementos. Aunque contemos con más de 20 elementos, debe de verificarse que los datos n sean significativos, Esto es: que la información sea “repetitiva”, también debemos de verificar que los datos puedan clasificarse. Y que dicha clasificación tiene coherencia y lógica (de acuerdo a lo que se nos está pidiendo).

Una vez que ya hemos ordenado y clasificado, presentaremos la información obtenida mediante una “tabla de frecuencias”

4.- la agrupación de los datos puede ser simple o mediante intervalos de clase.

DATOS NO A GRUPADOS.

1.- los datos son brutos(es decir, no se presentan clasificados)

2.- no es necesario clasificar ni generar una tabla de frecuentas, ya que no tiene “mucho sentido”.

3.- elementos que menor tamaño (generalmente menor a 20 elementos).Esto no sucede así siempre.

Aunque contemos con menos de 20 elementos, debe de verificarse que los datos no sean significativos, Esto es: que la información no sea “repetitiva”, de esta forma, sabremos que no se podrá clasificar y por lo tanto ser resumida en una “tabla de frecuencias”.

En caso de que una vez que hayamos ordenado los elementos, se cuente con datos significativos. Procedemos a clasificarlos (si es posible, ya que también debemos de buscar la lógica al clasificar los elementos) para convertirlos en “datos agrupados”.

Por ejemplo:

*si nos pidieran obtener la información del territorio de cada uno de los estados de México. No tiene mucho sentido que “que tratemos de agrupar”, ya que solo nos pide el nombre del estado de la república mexicana y la extensión territorial. ¿Para qué necesitaríamos una tabla de frecuencia de 32 elementos, cuando estos se repiten solo 1 vez?

4.- los datos no agrupados, también pueden ser ordenamos y de la misma forma, también se pueden obtener gráficas, determinar media, desviación estándar, etc. El hecho de que los datos “no agrupados” pueden ordenarse, no significa que se conviertan en “datos agrupados”.

Ejemplos:
Vas a investigar la edad a un grupo de 20 Niños en datos no agrupados (es decir, vienen los 20 niños y así como te dan la edad así la anotas 2,2,1,3,3,3,4,4,5,6,1,2,2,3,3,3,4,4,3,6 (Total 20 niños)

Estos son datos no agrupados por qué no los has clasificado y contado 1,1,2,2,2,2,3,3,3,3,3,3,3,4,4,4,4,5,5,6 (Total 20 niños).

Los datos no agrupados también los puedes ordenar, por ejemplo de la edad menor a la edad mayor, no están contabilizados ni clasificados solamente están ordenados
Para que sean datos agrupados tienes que contarlos y clasificarlos, por ejemplo cuántos niños había de cada año. (Y siguen siendo 20 niños)

Edad.....Frecuencia
1.....2
2.....4
3.....7
4.....4
5.....2
6.....1
Total.....20

o también los puedes agrupar (Serie agrupada) en clases, rangos, grupos o intervalos por ejemplo de 2 años para este caso (y siguen siendo 20)

Edad.....Frecuencia
1-2.....6
3-4.....11
5-6.....3
Total.....20

3.2 Identificar el concepto y los elementos de la distribución de frecuencias:

3.2.1 Clase

Se refiere a una separación de datos agrupados por una característica específica.

Cada clase está delimitada por el límite inferior de la clase y el límite superior de la clase.

3.2.2 Límites de clase

En una distribución de frecuencias agrupadas el límite inferior de una clase pertenece al intervalo, pero el límite superior no pertenece intervalo, se cuenta en el siguiente intervalo.

Ejemplo:

3, 15, 24, 28, 33, 35, 38, 42, 43, 38, 36, 34, 29, 25, 17, 7, 34, 36, 39, 44, 31, 26, 20, 11, 13, 22, 27, 47, 39, 37, 34, 32, 35, 28, 38, 41, 48, 15, 32, 13.

	c_i	f_i	F_i	n_i	N_i
[0, 5)	2.5	1	1	0.025	0.025
[5, 10)	7.5	1	2	0.025	0.050
[10, 15)	12.5	3	5	0.075	0.125
[15, 20)	17.5	3	8	0.075	0.200
[20, 25)	22.5	3	11	0.075	0.2775
[25, 30)	27.5	6	17	0.150	0.425
[30, 35)	32.5	7	24	0.175	0.600
[35, 40)	37.5	10	34	0.250	0.850
[40, 45)	42.5	4	38	0.100	0.950
[45, 50)	47.5	2	40	0.050	1
		40		1	

3.2.3 Amplitud

Es la diferencia entre el límite superior e inferior del intervalo de clase.

La amplitud total (A_T) es la diferencia entre la puntuación de mayor valor y la de menor valor:

$$A_T = X_{max} - X_{min}$$

Ejemplo: 2,5,6,8 $A_T = 8-2 = 6$

La amplitud total es un estadístico muy sencillo y fácil de calcular, pero a menudo esta simplicidad es un inconveniente. Consideremos el siguiente ejemplo:

Grupo	A_T
A: 2,3,3,4,5,5,6,7	5
B: 2,3,3,4,5,5,6,1000	998

Los grupos A y B son bastante semejantes, pero no los coeficientes de amplitud. La diferencia en los coeficientes es ocasionada por la variación introducida por una sola puntuación con valor extremo, el 1000. Por esta razón es conveniente disponer de otras medidas más adecuadas.

Principales características: Además de la ya señalada, el coeficiente de amplitud total no tiene en cuenta los valores entre extremos, que son los que determinan su valor.

3.2.4 Marca de clase

Marca de clase: Es el punto medio de una clase y se obtiene sumando los límites inferiores (LIA) y superiores de una clase (LSA) y dividiendo el resultado entre dos. La marca de clase la denotaremos como MC. =2 LSA+ LIA/2
Donde:
M C = Marca de clase
LIA = Límite inferior aparente
LSA = Límite superior aparente
Ejemplo: De la siguiente tabla obtenga la marca de clase.

2 LSA LIA MC

Ejemplo: De la siguiente tabla obtenga la marca de clase	*Fi*	*(LIA + LSA)/2*	*MC*
5 – 7	*5*	*(5 + 7)/2*	*6*
8 – 10	*10*	*(8 + 10)/2*	*9*
11 – 13	*15*	*(11 + 13)/2*	*12*
14 – 16	*18*	*(14 + 16)/2*	*13*
17 – 19	*11*	*(17 + 19)/2*	*18*
20 – 22	*5*	*(20 + 22)/2*	*21*
Totales		*64*	

3.2.5 Frecuencias:

Frecuencia es una magnitud que mide el número de repeticiones por unidad de tiempo de cualquier fenómeno o suceso periódico.

Para calcular la frecuencia de un suceso, se contabilizan un número de ocurrencias de este teniendo en cuenta un intervalo temporal, luego estas repeticiones se dividen por el tiempo transcurrido. Según el Sistema Internacional (SI), la frecuencia se mide en hercios (Hz), en honor a Heinrich Rudolf Hertz. Un hercio es la frecuencia de un suceso o fenómeno repetido una vez por segundo. Así, un fenómeno con una frecuencia de dos hercios se repite dos veces por segundo. Esta unidad se llamó originalmente «ciclo por segundo» (cps).
Otras unidades para indicar frecuencias son revoluciones por minuto (rpm o r/min según la notación del SI. Las pulsaciones del corazón se miden en latidos por minuto (lat/min) y el *tempo* musical se mide en «pulsos por minuto» (bpm, del inglés “beats per minute”).

$$1\,\mathrm{Hz} = \frac{1}{\mathrm{s}}$$

Un método alternativo para calcular la frecuencia es medir el tiempo entre dos repeticiones (periodo) y luego calcular la frecuencia (f) recíproca de esta manera:

$$f = \frac{1}{T}$$

Donde T es el periodo de la señal.

3.2.5.1 Absoluta

El número de veces que aparece un valor, se representa con *fi* donde el subíndice representa cada uno de los valores.

La suma de las frecuencias absolutas es igual al número total de datos, representado por *N*.

$f1+f2+f3+\ldots+fn=N$

Equivalente a

$\sum i=1nfi=N$

3.2.5.2 Relativa

El resultado de dividir la frecuencia absoluta de un determinado valor entre el número total de datos, se representa por *ni*.

$Ni=fiN$

La suma de las frecuencias relativas es igual a 1. Lo cual puede verse fácilmente si se factoriza *N*.

3.2.5.3 Porcentual

Es su frecuencia dividida entre la frecuencia total de todas las clases y se expresa generalmente como un porcentaje. Por ejemplo, la frecuencia relativa de la clase 66-68 de la tabla de estaturas de estudiantes del tema anterior es 42/100 = 42%. Es claro que la suma de todas las frecuencias relativas de las clases es 1, o sea 100%.

3.2.5.4 Acumulada

La suma de frecuencias absolutas de todos los valores iguales o inferiores al valor considerado, se representa por *Fi*.

Frecuencia relativa acumulada: el resultado de dividir la frecuencia acumulada entre el número total de datos, se representa por *Ni*.

(Nótese que cuando se trata de acumuladas las letras que las representan están en mayúscula)

Ejemplo

15 alumnos contestan a la pregunta de cuantos hermanos tienen. Las respuestas son

1,1,2,0,3,2,1,4,2,3,1,0,0,1,2

A continuación construimos una tabla de frecuencias:

Hermanos	Frecuencia absoluta *fi*	Frecuencia relativa *ni*	Frecuencia acumulada *Fi*	Frecuencia relativa acumulada *Ni*
0	3	315	3	315
1	5	515	3+5=8	315+515=815
2	4	415	3+5+4=12	1215
3	2	215	3+5+4+2=14	1415
4	1	115	3+5+4+2+1=15	1515
∑	15	1		

Nótese que la diferencia entre la frecuencia acumulada y la relativa es solamente que en el caso de la relativa debemos dividir por el número total de observaciones, lo que nos puede ayudar a ahorrar cálculos.

3.3 Explicar la construcción e interpretación de gráficas:

3.3.1 Histograma

Un histograma es una representación gráfica de una variable en forma de barras, donde la superficie de cada barra es proporcional a la frecuencia de los valores representados, ya sea en forma diferencial o acumulada. Sirven para obtener una "primera vista" general, o panorama, de la distribución de la población, o la muestra, respecto a una característica, cuantitativa y continua, de la misma y que es de interés para el observador (como la longitud o la masa). De esta manera ofrece una visión en grupo permitiendo observar una preferencia, o tendencia, por parte de la muestra o población por ubicarse hacia una determinada región de valores dentro del espectro de valores posibles (sean infinitos o no) que pueda adquirir la característica. Así pues, podemos evidenciar comportamientos, observar el grado de homogeneidad, acuerdo o concisión entre los valores de todas las partes que componen la población o la muestra, o, en contraposición, poder observar el grado de variabilidad, y por ende, la dispersión de todos los valores que toman las partes, también es posible no evidenciar ninguna tendencia y obtener que cada miembro de la población toma por su lado y adquiere un valor de la característica aleatoriamente sin mostrar ninguna preferencia o tendencia, entre otras cosas.

Construcción de un histograma

Paso 1

Determinar el rango de los datos. Rango es igual al dato mayor menos el dato menor.

Paso 2

Obtener todos los números de clases, existen 2 criterios para determinar el número de clases (o barras) –por ejemplo, la regla de Sturges. Sin embargo ninguno de ellos es exacto. Algunos autores recomiendan de cinco a quince clases, dependiendo de cómo estén los datos y cuántos sean. Un criterio usado frecuentemente es que el número de clases debe ser aproximadamente a la raíz cuadrada del número de datos. Por ejemplo, la raíz cuadrada de 30 (número de artículos) es mayor que cinco, por lo que se seleccionan seis clases.

Paso 3

Establecer la longitud de clase: es igual al rango dividido por el número de clases.

Paso 4

Construir los intervalos de clases: Los intervalos resultan de dividir el rango de los datos en relación al resultado del PASO 2 en intervalos diferentes

Paso 5

Graficar el histograma: En caso de que las clases sean todas de la misma amplitud, se hace una gráfica de pastel, las bases de las barras son los intervalos de clases y la altura es la frecuencia de las clases. Si se unen los puntos medios de la base superior de los rectángulos se obtiene el polígono de frecuencias.

3.3.2 Polígono de frecuencias

Es el gráfico de línea que se diseña utilizando en el eje horizontal la marca de clase de cada intervalo de una distribución de frecuencias (marca de clase = punto medio)

La información que se grafica en un polígono puede ser absoluta o relativa. Algunas de sus principales características son:

- En el eje horizontal se colocan las marcas de clase de cada intervalo
- En el eje vertical se ubica la frecuencia absoluta o la frecuencia porcentual
- Todos los puntos tienen la misma distancia en el eje X
- Las líneas siempre permanecen unidas
- Ambos extremos deben terminar sobre el eje horizontal
- Sólo funciona con datos numéricos o continuos
- En el cambio de intervalo se puede colocar la frecuencia absoluta o relativa para mayor comprensión de los datos.

3.3.3 Ojiva

Es el polígono frecuencia acumulado, es decir, que permite ver cuántas observaciones se encuentran por encima o debajo de ciertos valores, en lugar de solo exhibir los números asignados a cada intervalo.

La ojiva apropiada para información que presente frecuencias mayores que el dato que se está comparando tendrá una pendiente negativa (hacia abajo y a la derecha) y en cambio la que se asigna a valores menores, tendrá una pendiente positiva. Una gráfica similar al polígono de frecuencias es la ojiva, pero ésta se obtiene de aplicar parcialmente la misma técnica a una distribución acumulativa y de igual manera que éstas, existen las ojivas "mayor que" y las ojivas "menor que".

Existen dos diferencias fundamentales entre las ojivas y los polígonos de frecuencias (y por esto la aplicación de la técnica es parcial):

Un extremo de la ojiva no se toca al eje horizontal, para la ojiva "mayor que" sucede con el extremo izquierdo; para la ojiva "menor que", con el derecho.

En el eje horizontal, en lugar de colocar las marcas de clase, se colocan las fronteras de clase. Para el caso de la ojiva "mayor que" es la frontera menor; para la ojiva menor que, la mayor.

La ojiva "mayor que" se le denomina de esta manera porque viendo el punto que está sobre el límite superior se ven las frecuencias que tienen por encima de ese límite superior. De forma análoga, en la ojiva "menor que" la frecuencia que se representa en cada frontera de clase son el número de observaciones menores que la frontera señalada (en caso de tiempos sería el número de observaciones antes de la hora que señala la frontera)

3.3.4 Pareto

El diagrama de Pareto, también llamado curva cerrada o Distribución A-B-C, es una gráfica para organizar datos de forma que estos queden en orden descendente, de izquierda a derecha y separados por barras. Permite asignar un orden de prioridades.

El diagrama permite mostrar gráficamente el principio de Pareto (pocos vitales, muchos triviales), es decir, que hay muchos problemas sin importancia frente a unos pocos muy importantes. Mediante la gráfica colocamos los "pocos que son vitales" a la izquierda y los "muchos triviales" a la derecha.

El diagrama facilita el estudio de las fallas en las industrias o empresas comerciales, así como fenómenos sociales o naturales psicosomáticos, como se puede ver en el ejemplo de la gráfica al principio del artículo.

Hay que tener en cuenta que tanto la distribución de los efectos como sus posibles causas no es un proceso lineal sino que el 20% de las causas totales hace que sean originados el 80% de los efectos y rebotes internos del pronosticado.

El principal uso que tiene el elaborar este tipo de diagrama es para poder establecer un orden de prioridades en la toma de decisiones dentro de una organización. Evaluar todas las fallas, saber si se pueden resolver o mejor evitarla.

3.3.5 Pastel

Una gráfica circular, también llamada gráfico de pastel, gráfico de tarta o gráfica de 360 grados, es un recurso estadístico que se utiliza para representar porcentajes y proporciones. El número de elementos comparados dentro de una gráfica circular suele ser de más de 4.

Al igual que en la gráfica de barras, el empleo de tonalidades o colores facilita la diferenciación de los porcentajes o proporciones. A diferencia de otros tipos de gráficos, el circular no tiene ejes x o y.

Se utilizan en aquellos casos donde interesa no sólo mostrar el número de veces que se da una característica o atributo de manera tabular sino más bien de manera gráfica, de tal manera que se pueda visualizar mejor la proporción en que aparece esa característica respecto del total.

A pesar de su popularidad, se trata de un tipo de gráfico poco recomendable debido a que nuestra capacidad perceptual para estimar relaciones de proporción o diferencias entre áreas de sectores circulares es mucho menor que, por ejemplo, entre longitudes o posiciones, tal y como sucede en otras gráficas.

El gráfico circular más temprano conocido se atribuye generalmente al escocés William Playfair, en la obra Statistical Breviary de 1801.

3.3.6 Barras

El diagrama de barras (o gráfico de barras) es un gráfico que se utiliza para representar datos de variables cualitativas o discretas. Está formado por barras rectangulares cuya altura es proporcional a la frecuencia de cada uno de los valores de la variable.

Las principales características del diagrama de barras son:

- En el eje de accisas se colocan las cualidades de la variable, si la variable es cualitativa, o los valores de dicha variable, si es discreta.
- En el eje de ordenadas se colocan las barras proporcionales a la frecuencia relativa o absoluta del dato.
- Las barras pueden ser horizontales o verticales, según si los valores de la variable se reflejan en el eje horizontal o vertical.
- Todas las barras deben tener el mismo ancho y no deben superponerse las unas con las otras.

3.3.7 Tallo y hoja

El diagrama de tallo y hojas (Stem-and-Leaf Diagram) es un semigráfico que permite presentar la distribución de una variable cuantitativa. Consiste en separar cada dato en el último dígito (que se denomina hoja) y las cifras delanteras restantes (que forman el tallo).

Dibujo del tallo y la hoja de un dato es especialmente útil para conjuntos de datos de tamaño medio (entre 20 y 50 elementos) y que sus datos no se agrupan alrededor de un único tallo. Con él podemos hacernos la idea de qué distribución tienen los datos, la asimetría, etc.

Para construir el diagrama de tallo y hojas, debemos seguir los siguientes pasos:

- Ordenar los datos.
- Redondear los números (en el caso de que no lo estén) hasta tengan las cifran que queramos. Por ejemplo, si tenemos el número 3,62856 y queremos que tenga 2 dígitos la parte decimal, lo redondeamos a 3,63.
- Dibujar una tabla con dos columnas, la primera columna para el tallo y la segunda para las hojas. Disponer todos los tallos en la primera columna en orden descendente. Cada tallo solo se escribe una vez.
- Registrar en la segunda columna todas las hojas, en orden creciente, junto al tallo correspondiente.

4.- Medidas de tendencia central, localización y dispersión

4.1 Definir los conceptos de medidas de:

Tendencia central: media, mediana y moda

La medidas de centralización nos indican en torno a qué valor (centro) se distribuyen los datos.

Las medidas de centralización son:

Moda

La moda es el valor que tiene mayor frecuencia absoluta. Se representa por Mo.

Se puede hallar la moda para variables cualitativas y cuantitativas.
Hallar la moda de la distribución:

2, 3, 3, 4, 4, 4, 5, 5 Mo= 4

Si en un grupo hay dos o varias puntuaciones con la misma frecuencia y esa frecuencia es la máxima, la distribución es bimodal o multimodal, es decir, tiene varias modas.

1, 1, 1, 4, 4, 5, 5, 5, 7, 8, 9, 9, 9Mo= 1, 5, 9

Cuando todas las puntuaciones de un grupo tienen la misma frecuencia, no hay moda.

2, 2, 3, 3, 6, 6, 9, 9

Si dos puntuaciones adyacentes tienen la frecuencia máxima, la moda es el promedio de las dos puntuaciones adyacentes.

0, 1, 3, 3, 5, 5, 7, 8Mo = 4

Cálculo de la moda para datos agrupados 1º Todos los intervalos tienen la misma amplitud.

$$Mo = L_i + \frac{f_i - f_{i-1}}{(f_i - f_{i-1}) + (f_i - f_{i+1})} \cdot a_i$$

Li es el límite inferior de la clase modal fi es la frecuencia absoluta de la clase modal. fi--1 es la frecuencia absoluta inmediatamente inferior a la en clase modal. fi-+1 es la frecuencia absoluta inmediatamente posterior a la clase modal. ai es la amplitud de la clase.

También se utiliza otra fórmula de la moda que da un valor aproximado de ésta:

$$Mo = L_i + \frac{f_{i+1}}{f_{i-1} + f_{i+1}} \cdot a_i$$

Ejemplo

Calcular la moda de una distribución estadística que viene dada por la siguiente tabla:

	f_i
[60, 63)	5
[63, 66)	18
[66, 69)	42
[69, 72)	27
[72, 75)	8
	100

$$Mo = 66 + \frac{(42 - 18)}{(42 - 18) + (42 - 27)} \cdot 3 = 67.846$$

$$Mo = 66 + \frac{27}{18 + 27} \cdot 3 = 67.8$$

2º Los intervalos tienen amplitudes distintas. En primer lugar tenemos que hallar las alturas.

$$h_i = \frac{f_i}{a_i}$$

La clase modal es la que tiene mayor altura.

$$Mo = L_i + \frac{h_i - h_{i-1}}{(h_i - h_{i-1}) + (h_i - h_{i+1})} \cdot a_i$$

La fórmula de la moda aproximada cuando existen distintas amplitudes es:

$$Mo = L_i + \frac{h_{i+1}}{h_{i-1} + h_{i+1}} \cdot a_i$$

Ejemplo

En la siguiente tabla se muestra las calificaciones (suspenso, aprobado, notable y sobresaliente) obtenidas por un grupo de 50 alumnos. Calcular la moda.

	f_i	h_i
[0, 5)	15	3
[5, 7)	20	10
[7, 9)	12	6
[9, 10)	3	3
	50	

$$Mo = 5 + \frac{10 - 3}{(10 - 3) + (10 - 6)} \cdot 2 = 6.27$$

$$Mo = 5 + \frac{6}{3 + 6} \cdot 2 = 6.33$$

Mediana

Es el valor que ocupa el lugar central de todos los datos cuando éstos están ordenados de menor a mayor.

La mediana se representa por Me.

La mediana se puede hallar sólo para variables cuantitativas.

Cálculo de la mediana

1 Ordenamos los datos de menor a mayor.

2 Si la serie tiene un número impar de medidas la mediana es la puntuación central de la misma.

2, 3, 4, 4, 5, 5, 5, 6, 6Me= 5

3 Si la serie tiene un número par de puntuaciones la mediana es la media entre las dos puntuaciones centrales.

7, 8, 9, 10, 11, 12Me= 9.5

Cálculo de la mediana para datos agrupados

La mediana se encuentra en el intervalo donde la frecuencia acumulada llega hasta la mitad de la suma de las frecuencias absolutas.

Es decir tenemos que buscar el intervalo en el que se encuentre $\frac{N}{2}$.

$$Me = L_i + \frac{\frac{N}{2} - F_{i-1}}{f_i} \cdot a_i$$

Li es el límite inferior de la clase donde se encuentra la mediana.

$\frac{N}{2}$ Es la semisuma de las frecuencias absolutas.

Fi-1 es la frecuencia acumulada anterior a la clase mediana. ai es la amplitud de la clase.

La mediana es independiente de las amplitudes de los intervalos.

Ejemplo

Calcular la mediana de una distribución estadística que viene dada por la siguiente tabla:

	f_i	F_i
[60, 63)	5	5
[63, 66)	18	23
[66, 69)	42	65
[69, 72)	27	92
[72, 75)	8	100
	100	

100 / 2 = 50

Clase modal: [66, 69)

$$Me = 66 + \frac{50 - 23}{42} \cdot 3 = 67.93$$

Media aritmética

La media aritmética es el valor obtenido al sumar todos los datos y dividir el resultado entre el número total de datos.

$\bar{x}$ Es el símbolo de la media aritmética.

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{N}$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{N}$$

Ejemplo

Los pesos de seis amigos son: 84, 91, 72, 68, 87 y 78 kg. Hallar el peso medio.

$$\bar{x} = \frac{84 + 91 + 72 + 68 + 87 + 78}{6} = 80\ Kg$$

Media aritmética para datos agrupados

Si los datos vienen agrupados en una tabla de frecuencias, la expresión de la media es:

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + x_3 f_3 + \ldots + x_n f_n}{N}$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i f_i}{N}$$

Ejercicio de media aritmética

En un test realizado a un grupo de 42 personas se han obtenido las puntuaciones que muestra la tabla. Calcula la puntuación media.

	x_i	f_i	$x_i \cdot f_i$

[10, 20)	15	1	15
[20, 30)	25	8	200
[30,40)	35	10	350
[40, 50)	45	9	405
[50, 60	55	8	440
[60,70)	65	4	260
[70, 80)	75	2	150
		42	1 820

$$\bar{x} = \frac{1820}{42} = 43.33$$

Propiedades de la media aritmética

1 La suma de las desviaciones de todas las puntuaciones de una distribución respecto a la media de la misma igual a cero.

$$\sum(X_i - \bar{X}) = 0$$

Las suma de las desviaciones de los números 8, 3, 5, 12, 10 de su media aritmética 7.6 es igual a 0:

8 − 7.6 + 3 − 7.6 + 5 − 7.6 + 12 − 7.6 + 10 − 7.6 =

= 0. 4 − 4.6 − 2.6 + 4. 4 + 2. 4 = 0

2 La media aritmética de los cuadrados de las desviaciones de los valores de la variable con respecto a un número cualquiera se hace mínima cuando dicho número coincide con la media aritmética.

$$\sum(X_i - \bar{X})^2 \ Minimo$$

3 Si a todos los valores de la variable se les suma un mismo número, la media aritmética queda aumentada en dicho número.

4 Si todos los valores de la variable se multiplican por un mismo número la media aritmética queda multiplicada por dicho número.

Observaciones sobre la media aritmética

1 La media se puede hallar sólo para variables cuantitativas.

2 La media es independiente de las amplitudes de los intervalos.

3 La media es muy sensible a las puntuaciones extremas. Si tenemos una distribución con los siguientes pesos:

65 kg, 69kg, 65 kg, 72 kg, 66 kg, 75 kg, 70 kg, 110 kg.

La media es igual a 74 kg, que es una medida de centralización poco representativa de la distribución.

4 La media no se puede calcular si hay un intervalo con una amplitud indeterminada.

	x_i	f_i
[60, 63)	61.5	5
[63, 66)	64.5	18
[66, 69)	67.5	42
[69, 72)	70.5	27
[72, ∞)		8
		100

En este caso no es posible hallar la media porque no podemos calcular la marca de clase de último intervalo.

4.2.1 Localización:

Cuartiles

Los cuartiles son los tres valores de la variable que dividen a un conjunto de datos ordenados en cuatro partes iguales.

Q1, Q2 y Q3 determinan los valores correspondientes al 25%, al 50% y al 75% de los datos.

Q2 coincide con la mediana.

Cálculo de los cuartiles

1 Ordenamos los datos de menor a mayor.

3 Buscamos el lugar que ocupa cada cuartil mediante la expresión

$$\frac{k \cdot N}{4},\ k = 1, 2, 3$$

.

Número impar de datos

2, 5, 3, 6, 7, 4, 9

$$\begin{array}{ccccccc} 2, & 3, & 4, & 5, & 6, & 7, & 9 \\ & \downarrow & & \downarrow & & \downarrow & \\ & Q_1 & & Q_2 & & Q_3 & \end{array}$$

Número par de datos

2, 5, 3, 4, 6, 7, 1, 9

$$\begin{array}{ccccccccc} 1, & \underline{2,} & 3, & \underline{4,} & 5, & \underline{6,} & 7, & 9 \\ & 2.5 & & 4.5 & & 6.5 & & \\ & \downarrow & & \downarrow & & \downarrow & & \\ & Q_1 & & Q_2 & & Q_3 & & \end{array}$$

Cálculo de los cuartiles para datos agrupados

En primer lugar buscamos la clase donde se encuentra $\frac{k \cdot N}{4},\ k = 1, 2, 3$, en la tabla de las frecuencias acumuladas.

$$Q_k = L_i + \frac{\frac{k \cdot N}{4} - F_{i-1}}{f_i} \cdot a_i \qquad k = 1, 2, 3$$

Li es el límite inferior de la clase donde se encuentra la mediana.

N es la suma de las frecuencias absolutas.

Fi-1 es la frecuencia acumulada anterior a la clase mediana.

ai es la amplitud de la clase.

Ejercicio de cuartiles

Calcular los cuartiles de la distribución de la tabla:

	f_i	F_i
[50, 60)	8	8
[60, 70)	10	18
[70, 80)	16	34
[80, 90)	14	48
[90, 100)	10	58

[100, 110)	5	63
[110, 120)	2	65
	65	

Cálculo del primer cuartil

$$\frac{65 \cdot 1}{4} = 16.25$$

$$Q_1 = 60 + \frac{16.25 - 8}{10} \cdot 10 = 68.25$$

Cálculo del segundo cuartil

$$\frac{65 \cdot 2}{4} = 32.5$$

$$Q_2 = 70 + \frac{32.5 - 18}{16} \cdot 10 = 79.0625$$

Cálculo del tercer cuartil

$$\frac{65 \cdot 3}{4} = 48.75$$

$$Q_3 = 90 + \frac{48.75 - 48}{10} \cdot 10 = 90.75$$

4.2.1.1 Decíles

Los deciles son los nueve valores que dividen la serie de datos en diez partes iguales.

Los deciles dan los valores correspondientes al 10%, al 20%... y al 90% de los datos.

D_5 coincide con la mediana.

Cálculo de los deciles

En primer lugar buscamos la clase donde se encuentra $\frac{k \cdot N}{10}, \; k = 1, 2, ..9$, en la tabla de las frecuencias acumuladas.

$$D_k = L_i + \frac{\frac{k \cdot N}{10} - F_{i-1}}{f_i} \cdot a_i \qquad k = 1, 2, ..9$$

L_i es el límite inferior de la clase donde se encuentra la mediana.

N es la suma de las frecuencias absolutas.

F_{i-1} es la frecuencia acumulada anterior a la clase mediana.

a es la amplitud de la clase.

Ejercicio de deciles

Calcular los deciles de la distribución de la tabla:

	f_i	F_i
[50, 60)	8	8
[60, 70)	10	18
[70, 80)	16	34
[80, 90)	14	48
[90, 100)	10	58
[100, 110)	5	63
[110, 120)	2	65
	65	

Cálculo del primer decil

$$\frac{65 \cdot 1}{10} = 6.5$$

$$D_1 = 50 + \frac{6.5 - 0}{8} \cdot 10 = 58.12$$

Cálculo del segundo decil

$$\frac{65 \cdot 2}{10} = 13$$

$$D_2 = 60 + \frac{13 - 8}{10} \cdot 10 = 65$$

Cálculo del tercer decil

$$\frac{65 \cdot 3}{10} = 19.5$$

$$D_3 = 70 + \frac{19.5 - 18}{16} \cdot 10 = 70.94$$

Cálculo del cuarto decil

$$\frac{65 \cdot 4}{10} = 26$$

$$D_4 = 70 + \frac{26 - 18}{16} \cdot 10 = 75$$

Cálculo del quinto decil

$$\frac{65 \cdot 5}{10} = 32.5$$

$$D_5 = 70 + \frac{32.5 - 18}{16} \cdot 10 = 79.06$$

Cálculo del sexto decil

$$\frac{65 \cdot 6}{10} = 39$$

$$D_6 = 80 + \frac{39 - 34}{14} \cdot 10 = 83.57$$

Cálculo del séptimo decil

$$\frac{65 \cdot 7}{10} = 45.5$$

$$D_7 = 80 + \frac{45.5 - 34}{14} \cdot 10 = 88.21$$

Cálculo del octavo decil

$$\frac{65 \cdot 8}{10} = 52$$

$$D_8 = 90 + \frac{52 - 48}{10} \cdot 10 = 94$$

Cálculo del noveno decil

$$\frac{65 \cdot 9}{10} = 58.5$$

$$D_9 = 100 + \frac{58.5 - 58}{5} \cdot 10 = 101$$

4.2.1.2 Percentiles

Los percentiles son los 99 valores que dividen la serie de datos en 100 partes iguales.

Los percentiles dan los valores correspondientes al 1%, al 2%... y al 99% de los datos.

P_{50} coincide con la mediana.

Cálculo de los percentiles

En primer lugar buscamos la clase donde se encuentra $\frac{k \cdot N}{100}, \quad k = 1, 2, ..99$, en la tabla de las frecuencias acumuladas.

$$P_k = L_i + \frac{\frac{k \cdot N}{100} - F_{i-1}}{f_i} \cdot a_i \qquad k = 1, 2, ..99$$

L_i es el límite inferior de la clase donde se encuentra la mediana.

N es la suma de las frecuencias absolutas.

F_{i-1} es la frecuencia acumulada anterior a la clase mediana.

a_i es la amplitud de la clase.

Ejercicio de percentiles

Calcular el percentil 35 y 60 de la distribución de la tabla:

	f_i	F_i
[50, 60)	8	8
[60, 70)	10	18
[70, 80)	16	34
[80, 90)	14	48
[90, 100)	10	58
[100, 110)	5	63
[110, 120)	2	65
	65	

Percentil 35

$$\frac{65 \cdot 35}{100} = 22.75$$

$$P_{35} = 70 + \frac{22.75 - 18}{16} \cdot 10 = 72.97$$

Percentil 60

$$\frac{65 \cdot 60}{100} = 39$$

$$P_{60} = 80 + \frac{39 - 34}{14} \cdot 10 = 83.57$$

4.2.2 Dispersión:

4.2.2.1 Rango

Rango es el intervalo entre el valor máximo y el valor mínimo; por ello, comparte unidades con los datos. Permite obtener una idea de la dispersión de los datos, cuanto mayor es el rango, más dispersos están los datos de un conjunto.

Por ejemplo, para una serie de datos de carácter cuantitativo, como lo es la estatura medida en centímetros, tendríamos:

$$x_1 = 185, x_2 = 165, x_3 = 170, x_4 = 182, x_5 = 155$$

Es posible ordenar los datos como sigue:

$$x_{(1)} = 155, x_{(2)} = 165, x_{(3)} = 170, x_{(4)} = 182, x_{(5)} = 185$$

Donde la notación $x_{(i)}$ indica que se trata del elemento *i*-ésimo de la serie de datos. De este modo, el rango sería la diferencia entre el valor máximo $x_{(k)}$ y el mínimo $x_{(1)}$; o, lo que es lo mismo:

$$R = x_{(k)} - x_{(1)}$$

En nuestro ejemplo, con cinco valores, nos da que *R* = 185-155 = 30.

4.2.2.2 Varianza

La varianza es la media aritmética del cuadrado de las desviaciones respecto a la media de una distribución estadística.

La varianza se representa por σ^2.

$$\sigma^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{N} \quad \sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{N}$$

Varianza para datos agrupados

$$\sigma^2 = \frac{(x_1 - \bar{x})^2 f_1 + (x_2 - \bar{x})^2 f_2 + \ldots + (x_n - \bar{x})^2 f_n}{N} \quad \sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 f_i}{N}$$

Para simplificar el cálculo de la varianza vamos o utilizar las siguientes expresiones que son equivalentes a las anteriores.

$$\sigma^2 = \frac{x_1^2 + x_2^2 + \ldots + x_n^2}{N} - \bar{x}^2 \quad \sigma^2 = \sum_{i=1}^{n}\frac{x_i^2}{N} - \bar{x}^2$$

Varianza para datos agrupados

$$\sigma^2 = \frac{x_1^2 f_1 + x_2^2 f_2 + \ldots + x_n^2 f_n}{N} - \bar{x}^2 \quad \sigma^2 = \sum_{i=1}^{n}\frac{x_i^2 f_i}{N} - \bar{x}^2$$

Ejercicios de varianza

Calcular la varianza de la distribución:

9, 3, 8, 8, 9, 8, 9, 18

$$\bar{x} = \frac{9+3+8+8+9+8+9+18}{8} = 9$$

$$\sigma^2 = \frac{(9-9)^2 + (3-9)^2 + (8-9)^2 + (8-9)^2 + (9-9)^2 + (8-9)^2 + (9-9)^2 + (18-9)^2}{8} = 15$$

Calcular la varianza de la distribución de la tabla:

	x_i	f_i	$x_i \cdot f_i$	$x_i^2 \cdot f_i$
[10, 20)	15	1	15	225
[20, 30)	25	8	200	5000
[30,40)	35	10	350	12 250
[40, 50)	45	9	405	18 225
[50, 60	55	8	440	24 200
[60,70)	65	4	260	16 900
[70, 80)	75	2	150	11 250
		42	1 820	88 050

$$\bar{x} = \frac{1820}{42} = 43.33$$

$$\sigma^2 = \frac{88050}{42} - 43.33^2 = 218.94$$

1 La varianza será siempre un valor positivo o cero, en el caso de que las puntuaciones sean iguales.

2 Si a todos los valores de la variable se les suma un número la varianza no varía.

Si todos los valores de la variable se multiplican por un número la varianza queda multiplicada por el cuadrado de dicho número.

4 Si tenemos varias distribuciones con la misma media y conocemos sus respectivas varianzas se puede calcular la varianza total.

Si todas las muestras tienen el mismo tamaño:

$$\sigma^2 = \frac{\sigma_1^2 + \sigma_2^2 + \ldots + \sigma_n^2}{n}$$

Si las muestras tienen distinto tamaño:

$$\sigma^2 = \frac{k_1 \cdot \sigma_1^2 + k_2 \cdot \sigma_2^2 + \ldots + k_n \cdot \sigma_n^2}{k_1 + k_2 + \ldots + k_n}$$

Observaciones sobre la varianza

1 La varianza, al igual que la media, es un índice muy sensible a las puntuaciones extremas.

2 En los casos que no se pueda hallar la media tampoco será posible hallar la varianza.

3 La varianza no viene expresada en las mismas unidades que los datos, ya que las desviaciones están elevadas al cuadrado.

4.2.2.3 Desviación Estándar

La desviación típica o desviación estándar (denotada con el símbolo σ o s, dependiendo de la procedencia del conjunto de datos) es una medida de dispersión para variables de razón (variables cuantitativas o cantidades racionales) y de intervalo. Se define como la raíz cuadrada de la varianza de la variable.

Para conocer con detalle un conjunto de datos, no basta con conocer las medidas de tendencia central, sino que necesitamos conocer también la desviación que presentan los datos en su distribución respecto de la media aritmética de dicha distribución, con objeto de tener una visión de los mismos más acorde con la realidad al momento de describirlos e interpretarlos para la toma de decisiones.

a desviación estándar (DS/DE), también llamada desviación típica, es una medida de dispersión usada en estadística que nos dice cuánto tienden a alejarse los valores concretos del promedio en una distribución de datos. De hecho, específicamente, el cuadrado de la desviación estándar es "el promedio del cuadrado de la distancia de cada punto respecto del promedio". Se suele representar por una S o con la letra sigma, σ.

La desviación estándar de un conjunto de datos es una medida de cuánto se desvían los datos de su media. Esta medida es más estable que el recorrido y toma en consideración el valor de cada dato.

Distribución de probabilidad continúa

Es posible calcular la desviación estándar de una variable aleatoria continua como la raíz cuadrada de la integral

$$\sigma^2 = \int (x-\mu)^2 f(x)dx$$

Donde

$$\mu = \int x f(x)dx$$

Distribución de probabilidad discreta

La Desviación Estándar es la raíz cuadrada de la varianza de la distribución de probabilidad discreta:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Cuando los casos tomados son iguales al total de la población se aplica la fórmula de desviación estándar poblacional. Así la varianza es la media de los cuadrados de las diferencias entre cada valor de la variable y la media aritmética de la distribución.

En el caso de que la variable x no distribuya uniforme, si no con una probabilidad p_i para cada x_i, la desviación estándar será:

$$\sigma = \sqrt{\sum_{i=1}^{N} p_i(x_i-\mu)^2}, \quad \text{donde} \quad \mu = \sum_{i=1}^{N} p_i x_i.$$

Aunque esta fórmula es correcta, en la práctica interesa el realizar inferencias poblacionales, por lo que en el denominador en vez de n, se usa $n-1$ según la corrección de Bessel. Esta ocurre cuando la media de muestra se utiliza para centrar los datos, en lugar de la media de la población. Puesto que la media de la muestra es una combinación lineal de los datos, el residual a la muestra media se

extiende más allá del número de grados de libertad por el número de ecuaciones de restricción —en este caso una—. Dado esto a la muestra así obtenida de una muestra sin el total de la población se le aplica esta corrección con la fórmula desviación estándar muestral.

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}$$

Aquí se muestra cómo calcular la desviación estándar de un conjunto de datos. Los datos representan la edad de los miembros de un grupo de niños: {4, 1, 11, 13, 2, 7}

1. Calcular el promedio o media aritmética $\overline{x}$.

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i .$$

En este caso, *N* = 6:

$x_1 = 4$

$x_2 = 1$

$x_3 = 11$

$x_4 = 13$

$x_5 = 2$

$x_6 = 7$

$\overline{x} = \frac{1}{6} \sum_{i=1}^{6} x_i$ Sustituyendo *N* por 6

$\overline{x} = \frac{1}{6} (x_1 + x_2 + x_3 + x_4 + x_5 + x_6)$

$\overline{x} = \frac{1}{6} (4 + 1 + 11 + 13 + 2 + 7)$

$\overline{x} = 6,33$

2. Calcular la desviación estándar σ

$$\sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2}$$

$$\sigma = \sqrt{\frac{1}{6-1}\sum_{i=1}^{6}(x_i - \overline{x})^2}$$

Sustituyendo *N* por 6;

$$\sigma = \sqrt{\frac{1}{5}\sum_{i=1}^{6}(x_i - 6,33)^2}$$

Sustituyendo $\overline{x}$ por 6,33

$$\sigma = \sqrt{\frac{1}{5}\left[(4-6,33)^2 + (1-6,33)^2 + (11-6,33)^2 + (13-6,33)^2 + (2-6,33)^2 + (\right.}$$

$$\sigma = \sqrt{\frac{1}{5}\left[(-2,33)^2 + (-5,33)^2 + 4,67^2 + 6,67^2 + (-4,33)^2 + 0,67^2\right]}$$

$$\sigma = \sqrt{\frac{1}{5}\left(5,43 + 28,4 + 21,8 + 44,5 + 18,7 + 0,449\right)}$$

$$\sigma = \sqrt{\frac{119,28}{5}}$$

$$\sigma = \sqrt{23,856}$$

$$\sigma \approx 4,88.$$

4.2.2.4 Desviación media

La desviación respecto a la media es la diferencia entre cada valor de la variable estadística y la media aritmética.

$D_i = x - x$

La desviación media es la media aritmética de los valores absolutos de las desviaciones respecto a la media.

La desviación media se representa por $D_{\bar{x}}$

$$D_{\bar{x}} = \frac{|x_1 - \bar{x}| + |x_2 - \bar{x}| + \ldots + |x_n - \bar{x}|}{N}$$

$$D_{\bar{x}} = \frac{\sum_{i=1}^{n}|x_i - \bar{x}|}{N}$$

Ejemplo

Calcular la desviación media de la distribución:

9, 3, 8, 8, 9, 8, 9, 18

$$\bar{x} = \frac{9+3+8+8+9+8+9+18}{8} = 9$$

$$D_{\bar{x}} = \frac{|9-9|+|3-9|+|8-9|+|8-9|+|9-9|+|8-9|+|9-9|+|18-9|}{8} = 2.25$$

Si los datos vienen agrupados en una tabla de frecuencias, la expresión de la desviación media es:

$$D_{\bar{x}} = \frac{|x_1-\bar{x}|f_1+|x_2-\bar{x}|f_2+\ldots+|x_n-\bar{x}|f_n}{N}$$

$$D_{\bar{x}} = \frac{\sum_{i=1}^{n}|x_i-\bar{x}|f_i}{N}$$

Ejemplo

Calcular la desviación media de la distribución:

	x_i	f_i	$x_i \cdot f_i$	$\lvert x - \bar{x}\rvert$	$\lvert x - \bar{x}\rvert \cdot f_i$
[10, 15)	12.5	3	37.5	9.286	27.858
[15, 20)	17.5	5	87.5	4.286	21.43
[20, 25)	22.5	7	157.5	0.714	4.998
[25, 30)	27.5	4	110	5.714	22.856
[30, 35)	32.5	2	65	10.174	21.428

		21	457.5		98.57

$$\bar{x} = \frac{457.5}{21} = 21.786$$

$$D_{\bar{x}} = \frac{98.57}{21} = 4.69$$

}

Unidad 2 Probabilidad

Competencia del Módulo: El alumno determinará las probabilidades de datos estadísticos para contribuir a la toma de decisiones.

Resultado de aprendizaje: A partir de un caso práctico el alumno elaborará una hoja de verificación que contenga:

* Compendio de 8 ejercicios:

- Uno de operaciones y uno de representaciones de conjuntos

- Uno de probabilidad clásica y otro de probabilidad condicional

- Uno de cada técnica de conteo

- Estimación de parámetros aplicando el Teorema de Límite Central

- Cálculo de probabilidades con la distribución muestral.

Estadística Descriptiva e Inferencial

Se puede dividir la estadística en dos grandes ramas: la estadística Descriptiva y la estadística Inferencial.

Estadística Descriptiva: procedimientos empleados para organizar y resumir conjuntos de observaciones en forma cuantitativa. El resumen de los puede hacerse mediante tablas, gráficos o valores numéricos. Los conjuntos de datos que contienen observaciones de más de una variable permiten estudiar la relación o asociación que existe entre ellas.

“La Estadística Descriptiva es el estudio que incluye la obtención,
Organización, presentación y descripción de información numérica”.

Ejemplo:
Un director de escuela desea conocer las aptitudes de cinco secretarias que trabajan en dicha institución.
Se aplica una prueba de aptitudes a las cinco secretarias y las calificaciones son 82, 85, 95, 92 y 91. La medida estadística que emplea el Director es la aptitud promedio o media aritmética, la cual es la suma de los valores obtenidos dividida por el número de observaciones. Entonces, la calificación promedio es:

$$\frac{82+85+95+92+91}{5} = \frac{445}{5} = 89$$

Estadística Inferencial: métodos empleados para inferir algo acerca de una población basándose en los datos obtenidos a partir de una muestra. Los datos estadísticos son cálculos aritméticos realizados sobre los valores obtenidos en una porción de la población, seleccionada según criterios rigurosos.

“La inferencia estadística es una técnica mediante la cual se obtienen Generalizaciones o se toman decisiones en base a una información parcial o Completa obtenida mediante técnicas descriptivas”.

Para concluir diremos que existe otra gran división de las técnicas estadísticas:
a) Estadística Paramétrica.
b) Estadística No Paramétrica.
La Estadística Paramétrica es un conjunto de técnicas desarrolladas para niveles altos de medición como el de intervalos.
Los métodos paramétricos permiten hacer inferencias acerca de parámetros Poblacionales de las distribuciones. Estos métodos fueron los primeros en ser Desarrollados por los investigadores de la Estadística.

La Estadística no paramétrica es un conjunto de técnicas diseñadas para niveles de medición menor, por ejemplo, el nominal y ordinal, para efectuar. Estimaciones no habrá parámetros en estricto sentido.
A los procedimientos estadísticos que no dependen para su validez de la forma funcional de la distribución original de la población se les denomina procedimientos no paramétricos o libres de distribución.
Los Procedimientos No Paramétricos disponibles actualmente ofrecen varias ventajas para el investigador y analista de datos; entre ellos se pueden mencionar los que estableció Bradley en 1968:
1) La mayoría de los procedimientos no paramétricos se basan en un conjunto mínimo de suposiciones y esto tiende a reducir la posibilidad de utilizarlos inadecuadamente.
2) Los cálculos aritméticos necesarios para la aplicación de muchos procedimientos no paramétricos son cortos y fáciles, de manera que con su empleo se puede ahorrar tiempo.
3) Los procedimientos no paramétricos son por lo general fácilmente comprensibles para personas no muy formadas matemática o estadísticamente.
4) Se pueden aplicar los procedimientos no paramétricos cuando los datos que se van a analizar consisten más bien en rangos o conteos de frecuencia tales como porcentaje de pruebas, estatura, peso, longitud, entre otras.

Una variable estadística es cada una de las características o cualidades que poseen los individuos de una población.

Tipos de variable estadísticas

Variable cualitativa

Las variables cualitativas se refieren a características o cualidades que no pueden ser medidas con números. Podemos distinguir dos tipos:

Variable cualitativa nominal

Una variable cualitativa nominal presenta modalidades no numéricas que no admiten un criterio de orden.

Ejemplo:

El estado civil, con las siguientes modalidades: soltero, casado, separado, divorciado y viudo.

Variable cualitativa ordinal

Una variable cualitativa ordinal presenta modalidades no numéricas, en las que existe un orden.

Ejemplo

La nota en un examen: suspenso, aprobado, notable, sobresaliente.

Puesto conseguido en una prueba deportiva: 1º, 2º, 3º,...

Medallas de una prueba deportiva: oro, plata, bronce.

Variable cuantitativa

Una variable cuantitativa es la que se expresa mediante un número, por tanto se pueden realizar operaciones aritméticas con ella. Podemos distinguir dos tipos:

Variable discreta

Una variable discreta es aquella que toma valores aislados, es decir no admite valores intermedios entre dos valores específicos.

Ejemplo

El número de hermanos de 5 amigos: 2, 1, 0, 1, 3.

Variable continúa

Una variable continua es aquella que puede tomar valores comprendidos entre dos números.

Ejemplos

La altura de los 5 amigos: 1.73, 1.82, 1.77, 1.69, 1.75.En la práctica medimos la altura con dos decimales, pero también se podría dar con tres decimales.

Población

En estadística, población es el conjunto de cosas, personas, animales o situaciones que tiene una o varias características o atributos comunes, por ejemplo: los habitantes de El Salvador en el presente año, las personas menores de edad en el año 2001; los estudiantes de la Universidad, las reacciones de un nuevo medicamento, las diferencias entre los tratamientos de diferentes formulaciones de insecticidas, entre otras.

Población Finita: es el conjunto compuesto por una cantidad limitada de elementos, como el número de especies, el número de estudiantes, el número de obreros.

Población Infinita: es la que tiene un número extremadamente grande de componentes, como el conjunto de especies que tiene el reino animal o la cantidad de estrellas en el universo.

Muestra

La muestra es una parte, generalmente pequeña, que se toma del conjunto total para analizarla y hacer estudios que le permitan al investigador inferir o estimar las características de un problema.

La persona interesada en resolver un problema no tiene siempre a la mano toda la información, por lo que debe conformarse con pequeños detalles, carentes de precisión, que le ayuden a tomar decisiones bajo riesgo.

A un paciente que debe ser operado quirúrgicamente se le analiza su sangre tomando una muestra pequeña para conocer el grado de coagulación. No es necesario extraerle toda la sangre.

El industrial que desea saber si en alambre que produce tiene la resistencia necesaria a la tensión deseada, toma solamente una muestra de su producción, debido a que el alambre que se destruye con la prueba y de otra manera tendría que destruir toda la existencia.

Generalmente, los resultados obtenidos en una muestra son satisfactorios y permiten al investigador tener un conocimiento aceptable del problema.

La información o características que se encuentran en la muestra se llaman estimadores y sirven para deducir cómo son las características llamadas parámetros de la población.

Cuantitativos

Son aquellos que se pueden medir. Determinan variables estadísticas que pueden ser:

Discretas

Sólo pueden tomar un número finito de valores enteros, los valores posibles de estas variables son aislados.

Ejemplos de variables estadísticas cuantitativas discretas

- Número de hermanos: pueden ser 1, 2, 3…, pero nunca podrá ser 3,45.
- Número de empleados de una fábrica.
- Número de goles marcados por un equipo de futbol en la liga.

Continuas

Pueden tomar cualquier valor real (infinitos) dentro de un intervalo.

Ejemplos de variables estadísticas cuantitativas continuas

- Velocidad de un vehículo: puede ser 20; 54,2; 100; … km/h
- Temperaturas registradas en un observatorio cada hora.
- Peso en kg de los recién nacidos en un día en España.

Cualitativos

No se pueden medir numéricamente.

Ejemplos de variables estadísticas cualitativas

- Color de los ojos.
- Bondad de una persona.
- Profesión de una persona.

Determinan modalidades. Las modalidades del carácter profesión pueden ser: arquitecto, albañil, médico, etc.

2.-Población muestra y muestreo

Censo:

Se denomina censo, en estadística descriptiva, al recuento de individuos que conforman una población estadística, definida como un conjunto de elementos de referencia sobre el que se realizan las observaciones. El censo de una población estadística consiste básicamente, en obtener mediciones del número total de individuos mediante diversas técnicas de recuento, y que se hace cada 10 años

Parámetro

En estadística, un parámetro es un número que resume la gran cantidad de datos que pueden derivarse del estudio de una variable estadística.

El cálculo de este número está bien definido, usualmente mediante una fórmula aritmética obtenida a partir de datos de la población.

Los parámetros estadísticos son una consecuencia inevitable del propósito esencial de la estadística: crear de la realidad.

El estudio de una gran cantidad de datos individuales de una población puede ser farragoso e inoperativo, por lo que se hace necesario realizar un resumen que permita tener una idea global de la población, compararla con otras, comprobar

su ajuste a un modelo ideal, realizar estimaciones sobre datos desconocidos de la misma y, en definitiva, tomar decisiones.

A estas tareas contribuyen de modo esencial los parámetros estadísticos.

Por ejemplo, suele ofrecerse como resumen de la juventud de una población la media aritmética de las edades de sus miembros, esto es, la suma de todas ellas, dividida por el total de individuos que componen tal población

Muestreo

En estadística se conoce como muestreo a la técnica para la selección de una muestra a partir de una población.

Al elegir una muestra aleatoria se espera conseguir que sus propiedades sean extrapolables a la población. Este proceso permite ahorrar recursos, y a la vez obtener resultados parecidos a los que se alcanzarían si se realizase un estudio de toda la población. En las investigaciones llevadas por empresarios y de la medicina se usa muestreo extensivamente en recoger información sobre poblaciones.

Cabe mencionar que para que el muestreo sea válido y se pueda realizar un estudio adecuado (que consienta no solo hacer estimaciones de la población sino estimar también los márgenes de error correspondientes a dichas estimaciones), debe cumplir ciertos requisitos. Nunca podremos estar enteramente seguros de que el resultado sea una muestra representativa, pero sí podemos actuar de manera que esta condición se alcance con una probabilidad alta.

En el muestreo, si el tamaño de la muestra es más pequeño que el tamaño de la población, se puede extraer dos o más muestras de la misma población. Al conjunto de muestras que se pueden obtener de la población se denomina *espacio maestral*. La variable que asocia a cada muestra su probabilidad de extracción, sigue la llamada distribución muestral.

Estadístico

En estadística un estadístico (muestral) es una medida cuantitativa, derivada de un conjunto de datos de una muestra, con el objetivo de estimar o inferir características de una población o modelo estadístico.

Más formalmente un estadístico es una función medible T que, dada una muestra estadística de valores $(X_1, X_2, ..., X_n)$, les asigna un número, $T(X_1, X_2, ..., X_n)$, que sirve para estimar determinado parámetro de la distribución de la que procede la muestra. Así, por ejemplo, la media de los valores de una muestra (media muestral) sirve para estimar la media de la población de la que se ha extraído la misma; la varianza muestral podría usarse para estimar la varianza poblacional, etc. Esto se denomina como realizar una estimación puntual.

Técnicas de muestreo

a) Probabilístico

Muestreo aleatorio simple

Una muestra aleatoria simple es seleccionada de tal manera que cada muestra posible del mismo tamaño tiene igual probabilidad de ser seleccionada de la población. Para obtener una muestra aleatoria simple, cada elemento en la población tenga la misma probabilidad de ser seleccionado, el plan de muestreo puede no conducir a una muestra aleatoria simple. Por conveniencia, este método pude ser reemplazado por una tabla de números aleatorios. Cuando una población es infinita, es obvio que la tarea de numerar cada elemento de la población es infinita, es obvio que la tarea de numerar cada elemento de la población es imposible. Por lo tanto, ciertas modificaciones del muestreo aleatorio simple son necesarias. Los tipos más comunes de muestreo aleatorio modificado son sistemáticos, estratificados y de conglomerados. Las encuestas por muestreo consisten en extraer de una población finita de N unidades, subpoblaciones de un tamaño fijado de antemano. Si todas las unidades son indistinguibles, el número de muestras de tamaño n viene dado por:

Por ejemplo, si la población contiene 5 unidades A, B, C, D, E; existen 10 muestras diferentes de tamaño 3, que son:

ABC, ABD, ABE, ACD, ACE ADE, BCD, BCE, BDE. CDE

Debe notarse que la misma letra no ocurre dos veces en la misma muestra; y, también, que el orden de los elementos no tiene importancia, las seis muestras ABC, ACB, BAC, BCA, CAB, CBA son consideradas como iguales. El muestreo aleatorio simple es un método de selección de n unidades sacadas de N, de tal manera que cada una de las muestras tiene la misma probabilidad de ser elegida. En la práctica una muestra aleatoria simple es extrañada de la siguiente forma: Se numeran las unidades de la población del 1 al N, y por medio de una tabla de números aleatorios o colocando los números 1 a N en una urna, se extraen sucesivamente n números. Las unidades que llevan estos números constituyen la

muestra. El método elegido debe de verificar que en cualquier fase de la obtención de la muestra cada individuo que no ha sido sacado previamente, tiene la misma probabilidad de ser elegido. Es fácil ver que cada una de las NCn muestras tiene igual posibilidad de obtenerse. Cuando un número ha sido sacado de la urna, este no es reemplazado, ya que esto dará lugar a que la misma unidad entrara en la muestra más de una vez. Por esta razón el muestreo es descrito como sin reemplazo. El muestreo con reemplazo, es totalmente factible, aunque rara vez es usado, ya que no se ve la conveniencia de tener el mismo individuo dos veces en la misma muestra

Muestreo estratificado aleatorio

En este tipo de muestreo, la población de N unidades es dividida en subpoblaciones de N1, N2, N3, NL unidades, respectivamente. Estas subpoblaciones no se sobreponen y juntas forman la totalidad de la población, por lo que N1+ N2+ N3+ NL = N Las subpoblaciones son llamados estratos. Para obtener un beneficio completo de la estratificación se debe de conocer Nh. Una vez que han sido determinados los estratos, se saca una muestra de cada uno, la extracción se realiza de forma independiente en cada estrato. Los tamaños de la muestra dentro de los estratos son representados por n1, n2, n3, nL. Si se toma una muestra aleatoria simple en cada estrato, el procedimiento completo es conocido como muestreo estratificado aleatorio. La estratificación es una técnica común. Hay muchas razones para realizarla; las principales son:

Si se desea cierta precisión en alguna subdivisión, es necesario tratarla como si fuera una subpoblación por sí sola.

Las conveniencias de tipo administrativo.

La diversidad de determinados grupos (por ejemplo, hoteles, hospitales, prisiones, etc.) hace necesario un enfoque diferente al de las personas normales. O, por ejemplo, las grandes compañías conviene separarlas en un estrato diferente, para las pequeñas empresas se puede utilizar un tipo de muestreo por áreas.

La estratificación puede dar lugar a una ganancia en precisión de los estimadores de la población. Esto ocurre cuando una población heterogénea es dividida en subpoblaciones cada una de las cuales es internamente homogénea.

Muestreo sistemático

Este método de muestreo consiste en lo siguiente: Supóngase que las N unidades de la población se numeran en algún orden de 1 a N. Para seleccionar una muestra de n unidades tomamos una al azar de las k primeras unidades, a continuación elegimos la que viene k unidades siguientes y así sucesivamente. Por ejemplo, si k =30 y la primera unidad elegida es la 19, las subsiguientes unidades serán los números 49, 79, 109, etc. La selección de la primera unidad

determina la muestra completa. Este tipo de muestreo se llama muestra sistemática de cada k-enésima unidad

Las ventajas de este método sobre el aleatorio simple son:

Es más fácil obtener la muestra y ejecutarlo con menos errores.

Intuitivamente aparece como más preciso que el muestreo simple aleatorio. En efecto, estratifica la población en n sustratos, los cuales consisten en las primeras k unidades, las segundas k unidades, etc. Eligiendo una unidad por estrato. La diferencia está en que en el muestreo estratificado la unidad dentro de cada sustrato se elige al azar, en este siempre está en la misma posición relativa.

Una variante del muestreo sistemático consiste en escoger cada unidad en el centro del estrato; esto es, en lugar de empezar la secuencia con un número al azar escogido del 1 al k, tomamos el número inicial como (k+1)/2 si k es impar y (k+2)/2 si k es par.

Muestreo estratificado

Consiste en la división previa de la población de estudio en grupos o clases que se suponen homogéneos respecto a característica a estudiar. A cada uno de estos estratos se le asignaría una cuota que determinaría el número de miembros del mismo que compondrán la muestra. Dentro de cada estrato el muestreo se realizaría mediante más. Según la cantidad de elementos de la muestra que se han de elegir de cada uno de los estratos, existen dos técnicas de muestreo estratificado:

Asignación proporcional: el tamaño de cada estrato en la muestra es proporcional a su tamaño en la población.

Asignación óptima: la muestra recogerá más individuos de aquellos estratos que tengan más variabilidad. Para ello es necesario un conocimiento previo de la población.

Por ejemplo, para un estudio de opinión, puede resultar interesante estudiar por separado las opiniones de hombres y mujeres pues se estima que, dentro de cada uno de estos grupos, puede haber cierta homogeneidad. Así, si la población está compuesta de un 55% de mujeres y un 45% de hombres, se tomaría una muestra que contenga también esa misma proporción. Consiste en la división previa de la población de estudio en grupos o clases que se suponen homogéneos con respecto a alguna característica de las que se van a estudiar. A cada uno de estos estratos se le asignaría una cuota que determinaría el número de miembros del mismo que compondrán la muestra. Dentro de cada estrato se suele usar la técnica de muestreo sistemático, una de las técnicas de selección más usadas en la práctica.

Según la cantidad de elementos de la muestra que se han de elegir de cada uno de los estratos, existen dos técnicas de muestreo estratificado:

Asignación proporcional: el tamaño de la muestra dentro de cada estrato es proporcional al tamaño del estrato dentro de la población.

Asignación óptima: la muestra recogerá más individuos de aquellos estratos que tengan más variabilidad. Para ello es necesario un conocimiento previo de la población.

Por ejemplo, para un estudio de opinión, puede resultar interesante estudiar por separado las opiniones de hombres y mujeres pues se estima que, dentro de cada uno de estos grupos, puede haber cierta homogeneidad. Así, si la población está compuesta de un 55% de mujeres y un 45% de hombres, se tomaría una muestra que contenga también esos mismos porcentajes de hombres y mujeres.

Para una descripción general del muestreo estratificado y los métodos de inferencia asociados con este procedimiento, suponemos que la población está dividida en h subpoblaciones o estratos de tamaños conocidos N1, N2,..., Nh tal que las unidades en cada estrato sean homogéneas respecto a la característica en cuestión. La media y la varianza desconocidas para el i-ésimo estrato son denotadas por mí y si2, respectivamente.

Muestreo conglomerado

La población está dividida en áreas lo más heterogéneas posibles internamente y lo más homogéneas posibles entre sí. Selecciona al azar un conglomerado que será el que formará la muestra. Para este diseño el procedimiento a seguir sería:

Dividir la población en conglomerados (generalmente zonas geográficas)

Seleccionar un cierto número de conglomerados, por algún procedimiento aleatorio.

Seleccionar los sujetos, dentro de los conglomerados elegidos, según los tamaños de muestras asignadas a cada uno de ellos, empleando MÁS o MS.

b) No probabilístico identificar el proceso del diseño de una muestra:

Tamaño de la Muestra

Si se establece una muestra probabilística y se conoce el tamaño de la población, de ahora en adelante denotado por N se procede a determinar por fórmula el tamaño de la muestra adecuado. No siempre se tiene el dato del tamaño de la población y entonces existe otra fórmula para obtenerlo. ¿Cuál es el menor número de unidades muéstrales (personas, familias, grupos, organizaciones, etc.) que se necesitan para conformar una muestra (n) que asegure un error de muestreo menor de 0.01, 0.03 o 0.05?

Tamaño de la muestra denotado por n.

Métodos de muestreo no probabilísticos

A veces, para estudios exploratorios, el muestreo probabilístico resulta excesivamente costoso y se acude a métodos no probabilísticos, aun siendo conscientes de que no sirven para realizar generalizaciones (estimaciones inferenciales sobre la población), pues no se tiene certeza de que la muestra extraída sea representativa, ya que no todos los sujetos de la población tienen la misma probabilidad de ser elegidos. En general se seleccionan a los sujetos siguiendo determinados criterios procurando, en la medida de lo posible, que la muestra sea representativa.

En algunas circunstancias los métodos estadísticos y epidemiológicos permiten resolver los

Problemas de representatividad aun en situaciones de muestreo no probabilístico, por ejemplo los estudios de caso-control, donde los casos no son seleccionados aleatoriamente de la población.

Entre los métodos de muestreo no probabilísticos más utilizados en investigación encontramos:

Muestreo por cuotas:

También denominado en ocasiones "accidental". Se asienta generalmente sobre la base de un buen conocimiento de los estratos de la población y/o de los individuos más "representativos" o "adecuados" para los fines de la investigación. Mantiene, por tanto, semejanzas con el muestreo aleatorio estratificado, pero no tiene el carácter de aleatoriedad de aquél.

En este tipo de muestreo se fijan unas "cuotas" que consisten en un número de individuos que reúnen unas determinadas condiciones, por ejemplo: 20 individuos de 25 a 40 años, de sexo femenino y residentes en Gijón. Una vez determinada la cuota se eligen los primeros que se encuentren que cumplan esas características. Este método se utiliza mucho en las encuestas de opinión.

Muestreo intencional o de conveniencia:

Este tipo de muestreo se caracteriza por un esfuerzo deliberado de obtener muestras "representativas" mediante la inclusión en la muestra de grupos supuestamente típicos. Es muy frecuente su utilización en sondeos preelectorales de zonas que en anteriores votaciones han marcado tendencias de voto.

También puede ser que el investigador seleccione directa e intencionadamente los individuos de la población. El caso más frecuente de este procedimiento el utilizar como muestra los individuos a los que se tiene fácil acceso (los profesores de universidad emplean con mucha

3. Frecuencia y su representación grafica

En estadística, la frecuencia (o frecuencia absoluta) de un evento *x*, es el número de veces n_i que dicho evento se repite durante un experimento o muestra estadística. Comúnmente, la distribución de la frecuencia suele visualizarse con el uso de histogramas.

¿Para qué nos sirven los gráficos y las tablas de datos?

Los gráficos y las tablas representan e interpretan información procedente de diferentes fuentes, de forma clara, precisa y ordenada. Casi todo tipos de información puede organizarse en una tabla de datos y ser representada en algún tipo de gráfico.

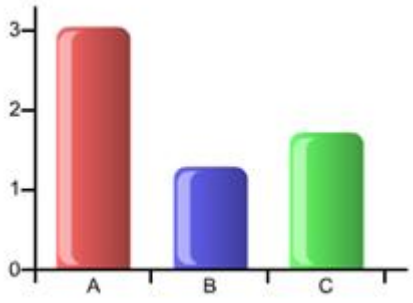
Según las características y la cantidad de datos, conviene utilizar uno u otro gráfico.

Gráficos
Los gráficos permiten visualizar la información contenida en las tablas de manera rápida y sencilla, demostrando con mayor claridad la relación que estos datos tienen entre sí.

Los más conocidos son:

Gráficos de barras

Son aquellos que emplean rectángulos (barras) que se colocan paralelamente. La altura indica la frecuencia de ese dato. Los gráficos de barras, permiten representar información numérica en forma clara y ordenada, para comunicarla a otras personas. Con la información representada en los gráficos puedes interpretar rápidamente y de manera visual la información, facilitando su posterior análisis.



Datos agrupados y no agrupados:

Datos Agrupados

Los datos agrupados son como lo indica su nombre, una cantidad dada de datos que puede clasificarse, ya sea por sus cualidades cualitativas o cuantitativas, y por tal agruparse para su análisis.

Estos datos por lo general son aconsejable agruparles cuando su población cuenta con alrededor de 20 o más elementos que comparten una característica y caben dentro de una categorización (repeticiones de un valor), pues permite un mejor manejo y análisis más profundo de los mismos. Porque al emplear este método podemos manejarlos por clases (una clase es una categoría en la que se agrupan los datos).

Por lo cual pueden organizarse o clasificarse de dos formas: datos agrupados en frecuencia o en intervalos.

Los datos agrupados en frecuencia son los que se distribuyen u organizan en una tabla de frecuencia (La frecuencia es igual al número de veces en que se repite cada valor en una serie de datos.), así, Por medio de ella, es fácil identificar la cantidad de respuestas repetidas.

Los datos agrupados por intervalos son los que se organizan dentro de un rango y se delimita su amplitud por límites establecidos. Así, por medio de esta, es fácil identificar la cantidad de elementos en un determinado rango de valores.

Concluyendo con la distinción de puntos significativos de este tema.

1.- su fin es resumir la información mediante el uso de tablas que organizan sus elementos y agrupan sus valores para ser presentados numérica o gráficamente. Esto implica: ordenar, clasificar y expresar los en una tabla de frecuencias o intervalos.

2.- Se agrupa a los datos, si se cuenta con 20 o más elementos. Aunque contemos con más de 20 elementos, debe de verificarse que los datos n sean significativos, esto es: que la información sea “repetitiva”, también debemos de verificar que los datos puedan clasificarse. Y que dicha clasificación tiene coherencia y lógica (de acuerdo a lo que se nos está pidiendo).

Ejemplos:

Se busca determinar el número de niños en cada uno de los grados escolares de una primaria, (del 1 al 6 grado), por lo que se recolectan los datos y se organizan y agrupan en una tabla de frecuencias.

Edad.....Frecuencia
1.....2
2.....4
3.....7
4.....4
5.....2
6.....1
Total.....20

Agrupación en intervalos, por ejemplo, de 2 años para este caso.

Edad.....Frecuencia
1-2.....6
3-4.....11
5-6.....3
Total.....20

DATOS NO AGRUPADOS

Los datos no agrupados son el conjunto de datos que no se ha clasificado y se es presentada en su forma de aparición en una tabla de datos donde cada valor se representa de forma individual. Por lo general este conjunto comprende una cantidad de elementos menor a 30 (n<30) con poca o nula repetición.

El tratamiento de estos datos sin agrupar. El manejo de estos datos es simple, se recolectan los datos de la población de estudio y dichos datos se distribuyen en una tabla de datos y se analizan sin necesidad de formar clases con ellos.

Estos datos al distribuirse en tabla de frecuencia donde cada dato mantiene su propia identidad después que la distribución de frecuencia se ha elaborado.

Vas a investigar la edad a un grupo de 20 Niños en datos no agrupados (es decir, vienen los 20 niños y así como te dan la edad así la anotas)

2,2,1,3,3,3,4,4,5,6,1,2,2,3,3,3,4,4,3,6 (Total 20 niños)

Estos son datos no agrupados por qué no los has clasificado y contado.
1,1,2,2,2,2,3,3,3,3,3,3,3,4,4,4,4,5,5,6 (Total 20 niños)

Los datos no agrupados también los puedes ordenar, por ejemplo de la edad menor a la edad mayor, no están contabilizados ni clasificados solamente están ordenados.

En una investigación sobre el calentamiento de varios elementos líquidos para determinar en cada uno de ellos el punto, la temperatura, en la cual cambian de estado, los científicos van anotando las temperaturas que van dando efecto.

134°C, 345°C, 234°C, 456°C, 837°C, 456°C, 122°C, 4567°C, 3456°C, 456°C, 190°C, 900°C.

Estas medidas pueden ser apiladas en una tabla de datos, y mantener su independencia como valor único y representativo

Elementos de distribución de frecuencia

Tipos de frecuencia

Frecuencia absoluta

La frecuencia absoluta es el número de veces que aparece un determinado valor en un estudio estadístico.

Se representa por f_i.

La suma de las frecuencias absolutas es igual al número total de datos, que se representa por N.

$$f_1 + f_2 + f_3 + \ldots + f_n = N$$

Para indicar resumidamente estas sumas se utiliza la letra griega Σ (sigma mayúscula) que se lee suma o sumatoria.

$$\sum_{i=1}^{i=n} f_i = N$$

Frecuencia relativa

La frecuencia relativa es el cociente entre la frecuencia absoluta de un determinado valor y el número total de datos.

Se puede expresar en tantos por ciento y se representa por n_i.

$$n_i = \frac{f_i}{N}$$

La suma de las frecuencias relativas es igual a 1.

Frecuencia acumulada

La frecuencia acumulada es la suma de las frecuencias absolutas de todos los valores inferiores o iguales al valor considerado.

Se representa por F_i

Frecuencia relativa acumulada

La frecuencia relativa acumulada es el cociente entre la frecuencia acumulada de un determinado valor y el número total de datos. Se puede expresar en tantos por ciento.

Ejemplo

Durante el mes de julio, en una ciudad se han registrado las siguientes temperaturas máximas:

32, 31, 28, 29, 33, 32, 31, 30, 31, 31, 27, 28, 29, 30, 32, 31, 31, 30, 30, 29, 29, 30, 30, 31, 30, 31, 34, 33, 33, 29, 29.

En la primera columna de la tabla colocamos la variable ordenada de menor a mayor, en la segunda hacemos el recuento y en la tercera anotamos la frecuencia absoluta.

x_i	Recuento	f_i	F_i	n_i	N_i
27	I	1	1	0.032	0.032
28	II	2	3	0.065	0.097
29	𝍸 I	6	9	0.194	0.290
30	𝍸 II	7	16	0.226	0.516
31	𝍸 III	8	24	0.258	0.774
32	III	3	27	0.097	0.871
33	III	3	30	0.097	0.968
34	I	1	31	0.032	1
		31		1	

Este tipo de tablas de frecuencias se utiliza con variables discretas.

Distribución de frecuencias agrupadas

La distribución de frecuencias agrupadas o tabla con datos agrupados se emplea si las variables toman un número grande de valores o la variable es continua.

Se agrupan los valores en intervalos que tengan la misma amplitud denominados clases. A cada clase se le asigna su frecuencia correspondiente.

Límites de la clase:

Cada clase está delimitada por el límite inferior de la clase y el límite superior de la clase.

Amplitud de la clase:

La amplitud de la clase es la diferencia entre el límite superior e inferior de la clase.

Marca de clase:

La marca de clase es el punto medio de cada intervalo y es el valor que representa a todo el intervalo para el cálculo de algunos parámetros.

Construcción de una tabla de datos agrupados

3, 15, 24, 28, 33, 35, 38, 42, 43, 38, 36, 34, 29, 25, 17, 7, 34, 36, 39, 44, 31, 26, 20, 11, 13, 22, 27, 47, 39, 37, 34, 32, 35, 28, 38, 41, 48, 15, 32, 13.

1º se localizan los valores menor y mayor de la distribución. En este caso son 3 y 48.

2º Se restan y se busca un número entero un poco mayor que la diferencia y que sea divisible por el número de intervalos de queramos poner.

Es conveniente que el número de intervalos oscile entre 6 y 15.

En este caso, 48 - 3 = 45, incrementamos el número hasta 50: 5 = 10 intervalos.

Se forman los intervalos teniendo presente que el límite inferior de una clase pertenece al intervalo, pero el límite superior no pertenece intervalo, se cuenta en el siguiente intervalo.

	c_i	f_i	F_i	n_i	N_i
[0, 5)	2.5	1	1	0.025	0.025
[5, 10)	7.5	1	2	0.025	0.050
[10, 15)	12.5	3	5	0.075	0.125
[15, 20)	17.5	3	8	0.075	0.200
[20, 25)	22.5	3	11	0.075	0.2775
[25, 30)	27.5	6	17	0.150	0.425
[30, 35)	32.5	7	24	0.175	0.600
[35, 40)	37.5	10	34	0.250	0.850
[40, 45)	42.5	4	38	0.100	0.950
[45, 50)	47.5	2	40	0.050	1
		40		1	

Construcción e interpretación de graficas:

Histogramas

Histograma El histograma es aquella representación gráfica de estadísticas de diferentes tipos. La utilidad del histograma tiene que ver con la posibilidad de establecer de manera visual, ordenada y fácilmente comprensible todos los datos numéricos estadísticos que pueden tornarse difíciles de entender. Hay muchos tipos de histogramas y cada uno se ajusta a diferentes necesidades como también a diferentes tipos de información.

Los histogramas son utilizados siempre por la ciencia estadística. Su función es exponer gráficamente números, variables y cifras de modo que los resultados se visualicen más clara y ordenadamente. El histograma es siempre una representación en barras y por eso es importante no confundirlo con otro tipo de gráficos como las tortas. Se estima que por el tipo de información brindada y por la manera en que ésta es dispuesta, los histogramas son de especial utilidad y eficacia para las ciencias sociales ya que permiten comparar datos sociales como los resultados de un censo, la cantidad de mujeres y/o hombres en una comunidad, el nivel de analfabetismo o mortandad infantil, etc.

Para un histograma existen dos tipos de informaciones básicas (que pueden ser complementados o no de acuerdo a la complejidad del diseño): la frecuencia de los valores y los valores en sí. Normalmente, las frecuencias son representadas en el eje vertical mientras que en el horizontal se representan los valores de cada una de las variables (que aparecen en el histograma como barras bi o tridimensionales).

Existen diferentes tipos de histogramas. Los histogramas de barras simples son los más comunes y utilizados. También están los histogramas de barras compuestas que permiten introducir información sobre dos variables. Luego están los histogramas de barras agrupadas según información y por último el polígono de frecuencias y la ojiva porcentual, ambos sistemas utilizados normalmente por expertos. Trabajar con histogramas es muy simple y seguramente proveerá con una mejor comprensión de diferente tipo de datos e información.

Polígono de frecuencia

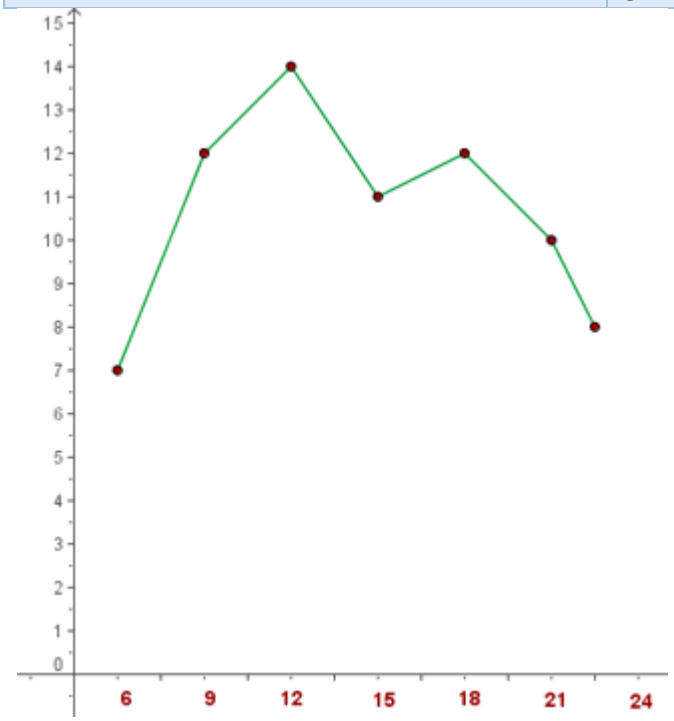
Un polígono de frecuencias se forma uniendo los extremos de las barras de un diagrama de barras mediante segmentos.

También se puede realizar trazando los puntos que representan las frecuencias y uniéndolos mediante segmentos.

Ejemplo

Las temperaturas en un día de otoño de una ciudad han sufrido las siguientes variaciones:

Hora	Temperatura
6	7º
9	12°
12	14°
15	11°
18	12°
21	10°
24	8°



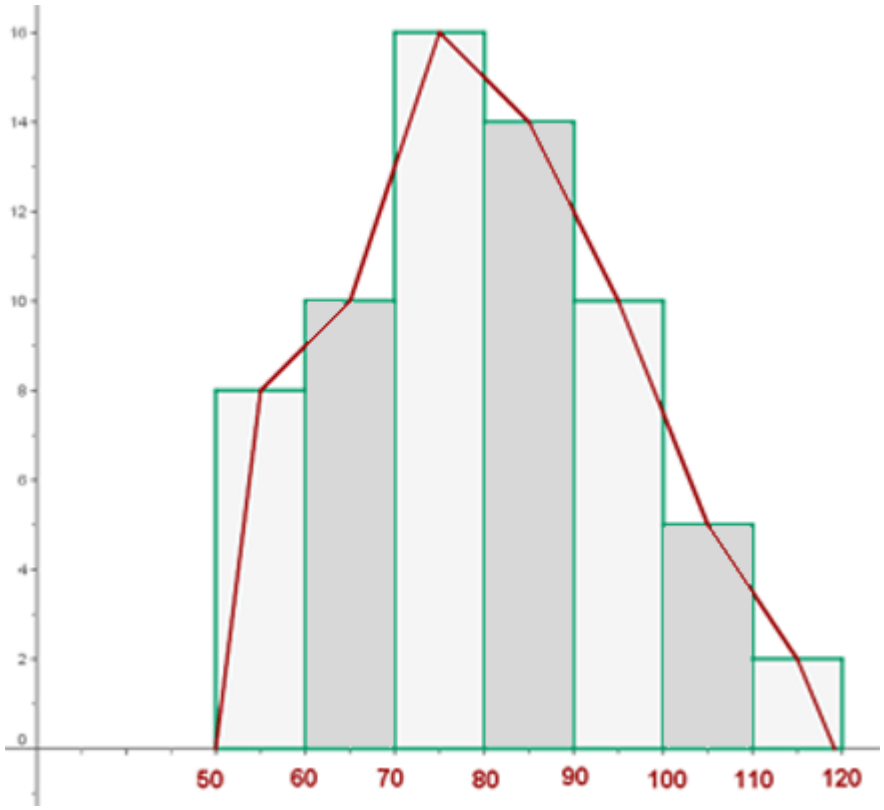
Polígonos de frecuencia para datos agrupados

Para construir el polígono de frecuencia se toma la marca de clase que coincide con el punto medio de cada rectángulo de un histograma.

Ejemplo

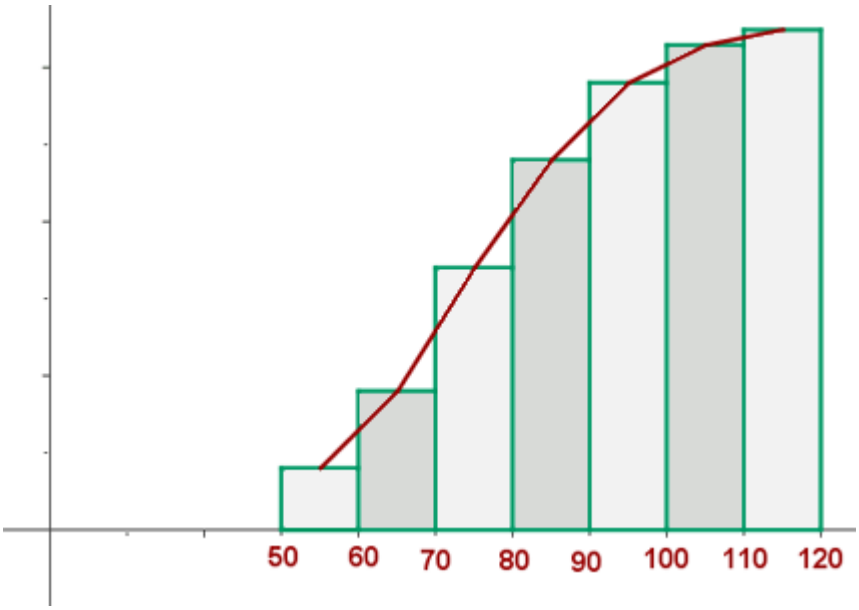
El peso de 65 personas adultas viene dado por la siguiente tabla:

	c_i	f_i	F_i
[50, 60)	55	8	8
[60, 70)	65	10	18
[70, 80)	75	16	34
[80, 90)	85	14	48
[90, 100)	95	10	58
[100, 110)	110	5	63
[110, 120)	115	2	65
		65	



Polígono de frecuencias acumuladas

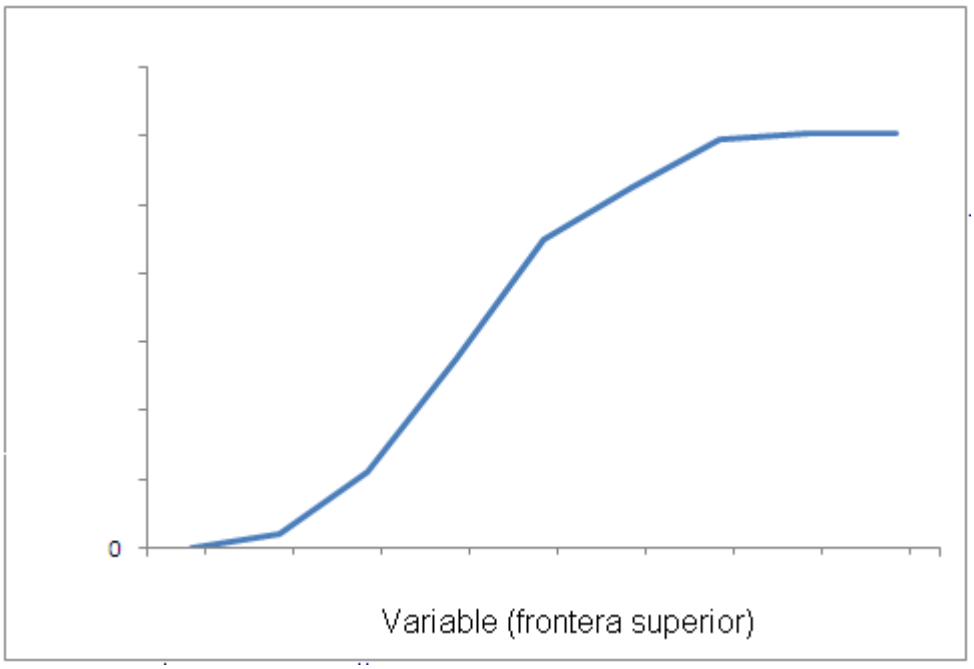
Si se representan las frecuencias acumuladas de una tabla de datos agrupados se obtiene el histograma de frecuencias acumuladas o su correspondiente polígono.



O j i v a

Es un gráfico de línea que se diseña utilizando en el eje horizontal las fronteras superiores de una distribución de frecuencias. La información se obtiene de la columna de frecuencias acumuladas (absoluta o relativa). Las características son las siguientes:

- En el eje horizontal se colocan las fronteras superiores de cada intervalo
- Todos los puntos tienen la misma distancia en el eje X
- Las líneas permanecen unidas
- El primer extremo termina sobre el eje horizontal
- Los datos son numéricos o continuos
- En el cambio de intervalo es posible colocar el valor de la frecuencia absoluta o relativa para una mejor comprensión de los datos.
- La forma general de una ojiva es la siguiente:



EJEMPLO

Se va a utilizar el mismo ejemplo de la muestra de 50 restaurantes dentro de la ciudad.

Objetivo: Trazar la ojiva a partir de la distribución de frecuencias

PRECIO POR PLATO	RESTAURANTES
14 pero menos de 21	1
21 pero menos de 28	5
28 pero menos de 35	7
35 pero menos de 42	16
42 pero menos de 49	10
49 pero menos de 56	9
56 pero menos de 63	1
63 pero menos de 70	1
Total	50

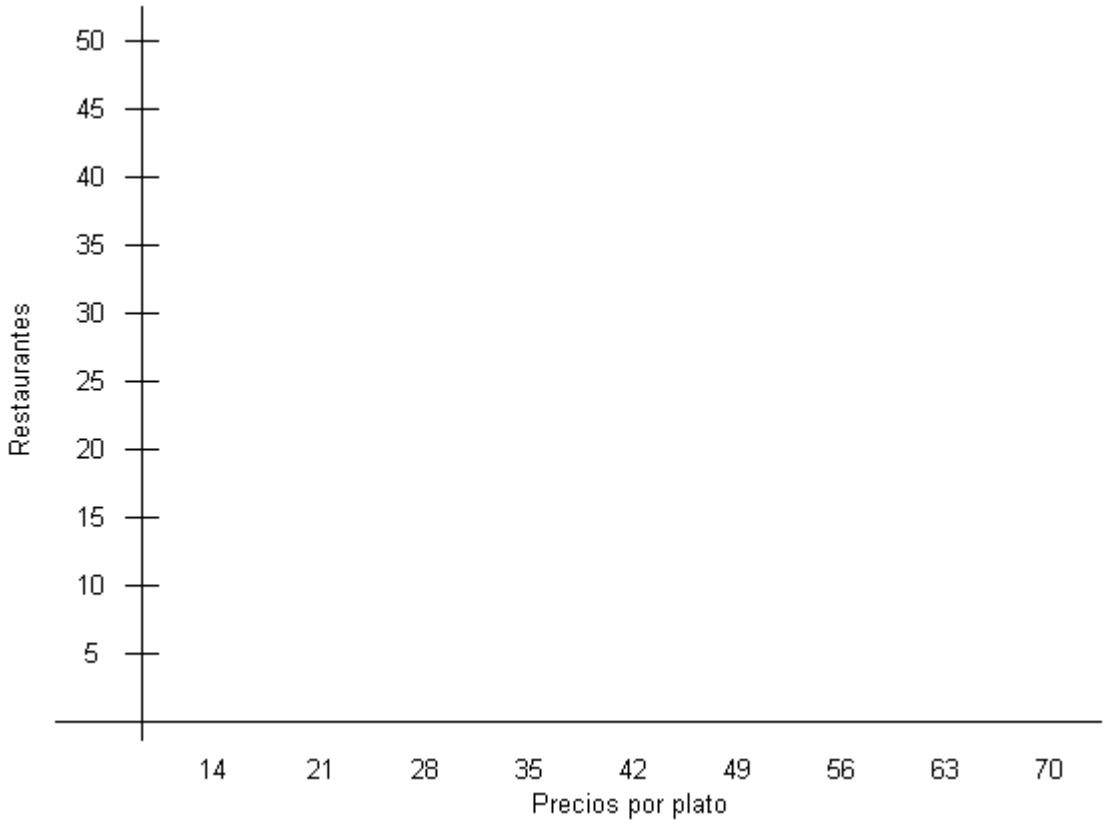
Siga paso a paso las siguientes indicaciones para trazar la ojiva:

1. Calcular la frecuencia acumulada de la distribución agregándole una nueva columna a la derecha, el primer intervalo es el mismo valor de la frecuencia absoluta, a partir del segundo se va acumulando con

todos los intervalos anteriores.

PRECIO POR PLATO	RESTAURANTES	Frecuencia acumulada		
14 pero menos de 21	1	1	=	1
21 pero menos de 28	5	1 + 5	=	6
28 pero menos de 35	7	1 + 5 + 7	=	13
35 pero menos de 42	16	1 + 5 + 7 +16	=	29
42 pero menos de 49	10	1 + 5 + 7 +16 + 10	=	39
49 pero menos de 56	9	1 + 5 + 7 +16 + 10 + 9	=	48
56 pero menos de 63	1	1 + 5 + 7 +16 + 10 + 9 + 1	=	49
63 pero menos de 70	1	1 + 5 + 7 +16 + 10 + 9 + 1 + 1	=	50
Total	50			

2. Trace una línea horizontal
3. Trace una línea vertical junto a la horizontal
4. Haga 9 marcas en la línea horizontal todas a la misma distancia
5. Haga 10 marcas en la línea vertical total a la misma distancia
6. Coloque la primera frontera superior que es 21 , es el primer valor que se marca en el eje horizontal
7. A continuación se colocan las demás fronteras superiores hasta llegar a 70 (ninguna marca debe quedar en blanco)
8. En el eje vertical se coloca la escala para las frecuencias, se empieza en 5 y termina en 50 (50 es el total de los datos)
9. La primera fase tendrá una forma similar a la siguiente:



10. Para la primera frontera (es la primera inferior = 10) no existe ningun dato en la muestra que sea menor que 10, por lo tanto la frecuencia acumulada es 0 y el punto se ubica en 10 sobre el eje horizontal
11. En la segunda frontera (es la primera superior = 21) existe 1dato en la muestra que son menores que 21, se coloca (21 , 1).
12. En la tercera frontera (es la segunda superior = 28) existen 5 datos; pero como en el intervalo anterior hay 1, en total son 6.
13. En la cuarta frontera (es la tercera superior = 35) existen 7 datos, pero como en los intervalos anteriores hay 5 y 11 respectivamente.
14. Se sigue haciendo los cálculos restantes hasta que la gráfica queda de la siguiente manera:

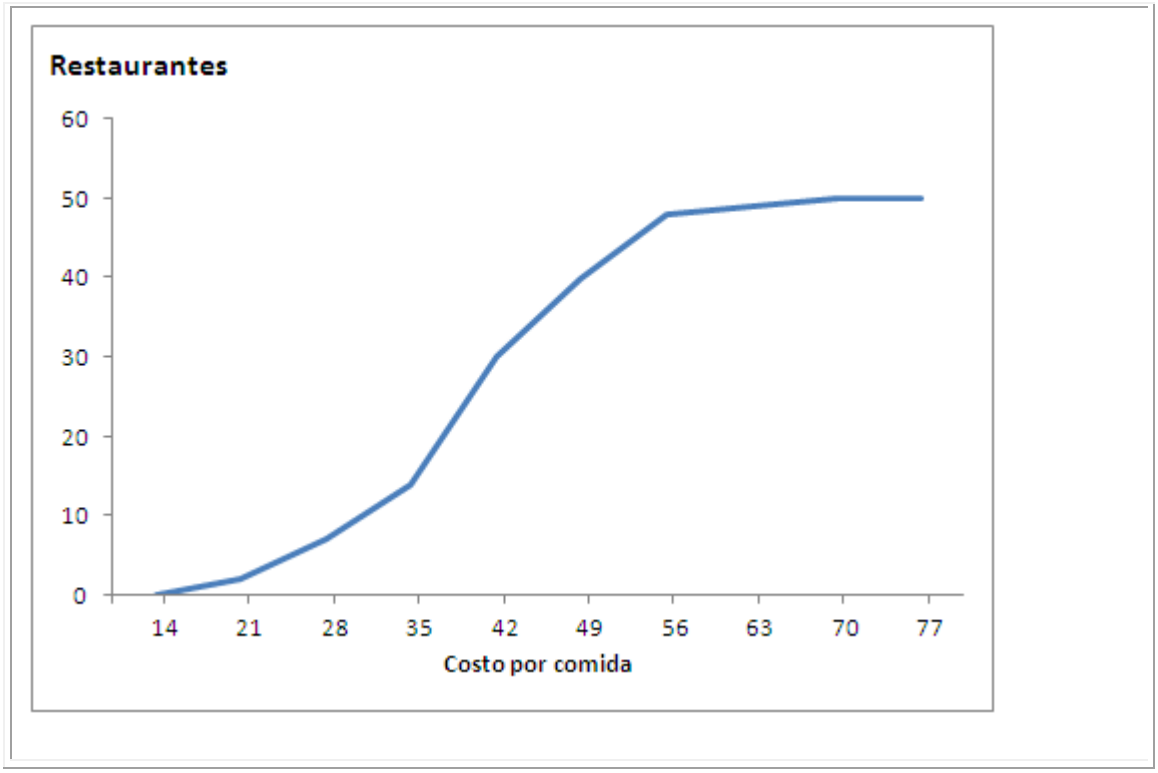


Tabla de Distribución de Frecuencias

Tal como puede leerse en la literatura estadística ésta es una ciencia que se encarga de recoger, organizar y analizar los hechos de naturaleza numérica referente a cualquier tópico. El ejercicio que se desarrolla a continuación aspira mostrar la manera de organizar y representar mediante gráficos las edades de un grupo de 30 estudiantes.

Objetivos

Organizar los datos en una tabla de distribución de frecuencias agrupándolos en clases de igual amplitud

Construir los gráficos que sean necesarios para el análisis de las edades.

Metodología:

Transcriban la etiqueta

Edades en A1 y los datos a partir de la columna

28 34 43 30 47 38 34 40 31 33
42 33 42 39 30 32 47 37 32 35
41 35 37 33 39 34 32 43 40 3

Calcular el Rango (R)

1. En la celda B2 escriban la etiqueta ***Rango***
2. En la celda C2 escriban la fórmula para calcular el Rango. Recuerden que éste se define como la diferencia entre el Valor Máximo y el Valor Mínimo de los datos, es decir: $= Max( ) - Min( )$.

   Procedimiento para calcular el rango en MS Excel:

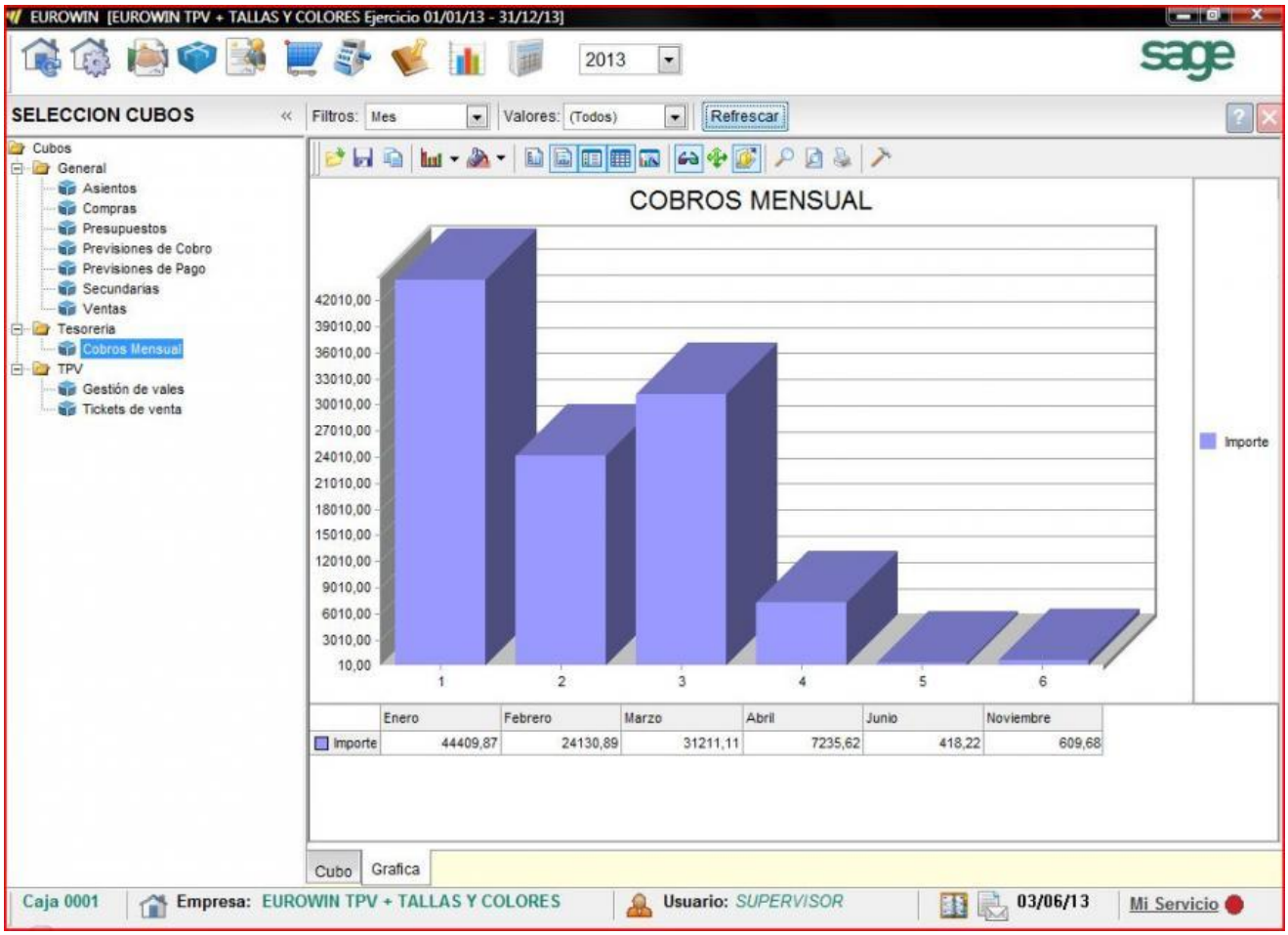
   - Ordenen los datos en forma ascendente. Para ello ejecuten las siguientes comandos: ***Datos/Ordenar*** - ***Ascendente/Aceptar***
   - Escriban y ejecuten la siguiente función: $= Max(A2:A31) - Min(A2:A31)$
   - Anoten el resultado en su cuaderno

Calcular el número de Clases o Intervalos (K)

1. En la celda B3 escriban la etiqueta ***Clases***
2. En la celda C3 determinen el número de Clases o de Intervalos (K). Pueden calcular el número de clases mediante uno de los siguientes métodos:

   No obstante que se puede aplicar cualquiera de los métodos, en este ejercicio solo se presentan los resultados del primero

Graficas con software



Mes	Valor	Fra...	Incide...
Ene	1,034	0	
Feb	1,130	0	
Mar	1,131	0	
Abr	1,284	0	
May	1,315	0	
Jun	1,263	0	
Jul	1,493	0	
Ago	1,569	0	
Sep	1,234	0	
Oct	1,561	0	
Nov	1,828	0	
Dic	340	0	
Total	15,182	0	0

4. Medidas de tendencia central

La medidas de centralización nos indican en torno a qué valor (centro) se distribuyen los datos.

Las medidas de centralización son:

Moda

La moda es el valor que tiene mayor frecuencia absoluta. Se representa por Mo. Se puede hallar la moda para variables cualitativas y cuantitativas.

Hallar la moda de la distribución:

2, 3, 3, 4, 4, 4, 5, 5 Mo= 4

Si en un grupo hay dos o varias puntuaciones con la misma frecuencia esa frecuencia es la máxima, la distribución es bimodal o multimodal, es decir, tiene varias modas.

1, 1, 1, 4, 4, 5, 5, 5, 7, 8, 9, 9, 9Mo= 1, 5, 9

Cuando todas las puntuaciones de un grupo tienen la misma frecuencia, no hay moda.

2, 2, 3, 3, 6, 6, 9, 9

Si dos puntuaciones adyacentes tienen la frecuencia máxima, la moda es el promedio de las dos puntuaciones adyacentes.

0, 1, 3, 3, 5, 5, 7, 8Mo = 4

Cálculo de la moda para datos agrupados

$$Mo = L_i + \frac{f_i - f_{i-1}}{(f_i - f_{i-1}) + (f_i - f_{i+1})} \cdot a_i$$

Localización y dispersión

En el caso de las variables con valores que pueden definirse en términos de alguna escala de medida de igual intervalo, puede usarse un tipo de indicador que permite apreciar el grado de dispersión o variabilidad existente en el grupo de variantes en estudio.

A estos indicadores les llamamos medidas de dispersión, por cuanto que están referidos a la variabilidad que exhiben los valores de las observaciones, ya que si no hubiere variabilidad o dispersión en los datos interés, entonces no habría necesidad de la gran mayoría de las medidas de la estadística descriptiva.

Las medidas de tendencia central tienen como objetivo el sintetizar los datos en un valor representativo, las medidas de dispersión nos dicen hasta qué punto estas medidas de tendencia central son representativas como síntesis de la información. Las medidas de dispersión cuantifican la separación, la dispersión, la variabilidad de los valores de la distribución respecto al valor central. Distinguimos entre medidas de dispersión absolutas, que no son comparables entre diferentes muestras y las relativas que nos permitirán comparar varias muestras.

LA DISPERSIÓN.

Al igual que sucede con cualquier conjunto de datos, la media, la mediana y la moda sólo nos revelan una parte de la información que necesitamos acerca de las características de los datos. Para aumentar nuestro entendimiento del patrón de los datos, debemos medir también su dispersión, extensión o variabilidad.

Definir los conceptos de medidas de: tendencia central:

Media

Este tipo de medidas nos permiten identificar y ubicar el punto (valor) alrededor del cual se tienden a reunir los datos (“Punto central”). Estas medidas aplicadas a las características de las unidades de una muestra se les denomina estimadores o

estadígrafos; mientras que aplicadas a poblaciones se les denomina parámetros o valores estadísticos de la población. Los principales métodos utilizados para ubicar el punto central son la media, la mediana y la moda.

Es la medida de posición central más utilizada, la más conocida y la más sencilla de calcular, debido principalmente a que sus ecuaciones se prestan para el manejo algebraico, lo cual la hace de gran utilidad. Su principal desventaja radica en su sensibilidad al cambio de uno de sus valores o a los valores extremos demasiado grandes o pequeños. La media se define como la suma de todos los valores observados, dividido por el número total de observaciones.

$$Media\ Aritmética = \frac{Suma\ de\ todos\ los\ valores\ observados}{Número\ total\ de\ observaciones}$$

Cuando los valores representan una población la ecuación se define como:

$$\overline{\mu} = \frac{X_1 + X_2 + X_3 + \ldots + X_n}{N} = \frac{\sum_{i=1}^{n} X_i}{N}$$

Donde (m) representa la media, (N) representa el tamaño de la población y (Xi) representa cada uno de los valores de la población. Ya que en la mayoría de los casos se trabajan con muestras de la población todas las ecuaciones que se presenten a continuación serán representativas para las muestras. La media aritmética para una muestra está determinada como

$$\overline{\mathbf{X}} = \frac{X_1 + X_2 + X_3 + \ldots + X_n}{\mathbf{n}} = \frac{\sum_{i=1}^{n} X_i}{\mathbf{n}}$$

Donde (X) representa la Media para la muestra, (n) el tamaño de la muestra y (Xi) representa cada uno de los valores observados. Esta fórmula únicamente es aplicable si los datos se encuentran desagrupados; en caso contrario debemos calcular la media mediante la multiplicación de los diferentes valores por la frecuencia con que se encuentren dentro de la información; es decir,

$$\overline{\mathbf{X}} = \frac{\sum_{i=1}^{n} Y_i\, \mathbf{n}_i}{\mathbf{n}}$$

Donde (Yi) representa el punto medio de cada observación, (ni) es la frecuencia o número de observaciones en cada clase y (n) es el tamaño de la muestra siendo igual a la suma de las frecuencias de cada clase.

Para entender mejor este concepto vamos a suponer que hemos tomado la edad de 5 personas al azar cuyos resultados fueron (22, 33, 35, 38 y 41). Para facilitar su interpretación se han generado tres rangos de edad los cuales se han establecido de 21 a 30 años, de 31 a 40 años y de 41 a 50 años. Si nos fijamos en estos rangos notaremos que los puntos medios son 25, 35 y 45 respectivamente. Los resultados de la organización de estos datos se representan en la tabla [5-1].

RANGO	Y_i	n_i	$Y_i * n_i$
21-30	25	1	25
31-40	35	3	105
41-50	45	1	45

Mediana

Con esta medida podemos identificar el valor que se encuentra en el centro de los datos, es decir, nos permite conocer el valor que se encuentra exactamente en la mitad del conjunto de datos después que las observaciones se han ubicado en serie ordenada. Esta medida nos indica que la mitad de los datos se encuentran por debajo de este valor y la otra mitad por encima del mismo. Para determinar la posición de la mediana se utiliza la fórmula

$$Posición\ de\ la\ mediana = \frac{n+1}{2}$$ Ecuación 5-5

Para comprender este concepto vamos a suponer que tenemos la serie ordenada de valores (2, 5, 8, 10 y 13), la posición de la mediana sería:

$$Posición\ de\ la\ mediana = \frac{n+1}{2} = \frac{5+1}{2} = 3$$

Lo que nos indica que el valor de la mediana corresponde a la tercera posición de la serie, que equivale al número (8). Si por el contrario contamos con un conjunto de datos que contiene un número par de observaciones, es necesario promediar los dos valores medios de la serie. Si en el ejemplo anterior le anexamos el valor 15, tendríamos la serie ordenada (2, 5, 8, 10, 13 y 15) y la posición de la mediana sería,

$$Posición\ de\ la\ mediana = \frac{n+1}{2} = \frac{6+1}{2} = 3{,}5$$

Es decir, la posición tres y medio. Dado que es imposible destacar la posición tres y medio, es necesario promediar los dos valores de la posiciones tercera y cuarta para producir una mediana equivalente, que para el caso corresponden a (8 + 10)/2 =9. Lo que nos indicaría que la mitad de los valores se encuentra por debajo del valor 9 y la otra mitad se encuentra por encima de este valor.

En conclusión la mediana nos indica el valor que separa los datos en dos fracciones iguales con el cincuenta por ciento de los datos cada una. Para las muestras que cuentan con un número impar de observaciones o datos, la mediana dará como resultado una de las posiciones de la serie ordenada; mientras que para las muestras con un número par de observaciones se debe promediar los valores de las dos posiciones centrales.

MODA

La medida modal nos indica el valor que más veces se repite dentro de los datos; es decir, si tenemos la serie ordenada (2, 2, 5 y 7), el valor que más veces se repite es el número 2 quien sería la moda de los datos. Es posible que en algunas ocasiones se presente dos valores con la mayor frecuencia, lo cual se denomina *Bimodal* o en otros casos más de dos valores, lo que se conoce como *multimodal*.

En conclusión las *Medidas de tendencia central*, nos permiten identificar los valores más representativos de los datos, de acuerdo a la manera como se tienden a concentrar. La *Media* nos indica el promedio de los datos; es decir, nos informa el valor que obtendría cada uno de los individuos si se distribuyeran los valores en partes iguales. La *Mediana* por el contrario nos informa el valor que separa los datos en dos partes iguales, cada una de las cuales cuenta con el cincuenta por ciento de los datos. Por último la *Moda* nos indica el valor que más se repite dentro de los datos.

Localización: cuartiles

Las medidas de localización dividen la distribución en partes iguales, sirven para clasificar a un individuo o elemento dentro de una determinada población o muestra.

Q1= valor de la variable que deja a la izquierda el 25%de la distribución

Medida de localización que divide la población o muestra en cuatro partes iguales.

Q2=valor de la variable que deja a la izquierda el 50% de la distribución = mediana.

Q3= valor de la variable que deja a la izquierda el 75% de la distribución. Al igual que le ocurre con el cálculo de la mediana, el cálculo de estos estadísticos, dependen del tipo de variable.

En este caso el cálculo es más simple.

Li-2—Li-1) ni-1 ni-1

(Li-1—Li) ni Ni

Q1= Li-1+n/4-Ni-1/Ni-Ni-1

Q2=Li-1+2n/4-Ni-1/ Ni-Ni-1

Q3=Li-1+3n/4-Ni-1 / Ni-Ni-1

Deciles

Medida de localización que divide la población o muestra en 10 partes iguales

Dk=decil k-simo es aquel valor de la variable que deja a su izquierda el k 10% de la distribución

(Li-2—Li-1) ni-1 Ni-1

(Li-1—Li) ni Ni

Dk= Li-1+ k.n/10-Ni-1, ai / Ni- Ni-1

K=1—9

Percentiles

Medida de localización que divide la población o muestra en 100 partes iguales

PK= percentil k-simo es aquel valor de la variable que deja a su izquierda el k% de la distribución.

(Li-2—Li-1) ni-1 Ni-1

(Li-1—Li) ni Ni

PK= Li-1+ k.n/100-Ni-1 / NI-Ni-1

K= 1……..99

EJEMPLO:

Li_1	Li	ni	Ni
45	55	6	6
55	65	10	16
65	75	19	35
75	85	11	46
85	95	4	50

Q1: Busquemos en le columna de las frecuencias acumuladas el valor que supere al 25% de N= 50 corresponde al 2 intervalo (50/4= 12.5)

Q1= 55 + 50/4-6.10 / 16-6= 61.5

Dispersión: de rango

Las medidas de dispersión, también llamadas medidas de variabilidad, muestran la variabilidad de una distribución, indicando por medio de un número si las diferentes puntuaciones de una variable están muy alejadas de la media. Cuanto mayor sea ese valor, mayor será la variabilidad, y cuanto menor sea, más homogénea será a la media. Así se sabe si todos los casos son parecidos o varían mucho entre ellos.

Para calcular la variabilidad que una distribución tiene respecto de su media, se calcula la media de las desviaciones de las puntuaciones respecto a la media aritmética. Pero la suma de las desviaciones es siempre cero, así que se adoptan dos clases de estrategias para salvar este problema. Una es tomando las desviaciones en valor absoluto (desviación media) y otra es tomando las desviaciones al cuadrado (varianza).

.Para paliar este inconveniente a veces se utilizan otros dos rangos:

- Rango intercuartílico: $Q = Q_3 – Q_1$
- Rango entre percentiles: $P = P_{90} – P_{10}$

Estos rangos son algo más estables, ya que tienden a eliminar aquellos valores extremadamente alejados.

Varianza

Varianza: Mide la distancia existente entre los valores de la serie y la media. Se calcula como sumatorio de las diferencias al cuadrado entre cada valor y la media, multiplicadas por el número de veces que se ha repetido cada valor. El sumatorio obtenido se divide por el tamaño de la muestra.

$$S^2_x = \frac{\Sigma (x_i - xm)^2 * n_i}{n}$$

La varianza siempre será mayor que cero. Mientras más se aproxima a cero, más concentrados están los valores de la serie alrededor de la media. Por el contrario, mientras mayor sea la varianza, más dispersos están.

Desviación estándar

Esta medida nos permite determinar el promedio aritmético de fluctuación de los datos respecto a su punto central o media. La desviación estándar nos da como resultado un valor numérico que representa el promedio de diferencia que hay entre los datos y la media. Para calcular la desviación estándar basta con hallar la raíz cuadrada de la varianza, por lo tanto su ecuación sería:

$S = \sqrt{S^2}$ Ecuación 5-8

Para comprender el concepto de las medidas de distribución vamos a suponer que el gerente de una empresa de alimentos desea saber que tanto varían los pesos de los empaques (en gramos), de uno de sus productos; por lo que opta por seleccionar al azar cinco unidades de ellos para pesarlos. Los productos tienen los siguientes pesos (490, 500, 510, 515 y 520) gramos respectivamente.

Por lo que su media es:

$$\overline{X} = \frac{490+500+510+515+520}{5} = \frac{2535}{5} = 507$$

La varianza sería:

$$S^2 = \frac{(490-507)^2+(500-507)^2+(510-507)^2+(515-507)^2+(520-507)^2}{(5-1)}$$

$$S^2 = \frac{(-17)^2+(-7)^2+(3)^2+(8)^2+(13)^2}{4} = \frac{289+49+9+64+169}{4} = \frac{580}{4} = 145$$

Por lo tanto la desviación estándar sería:

$S = \sqrt{145} = 12.04 \cong 12$

Con lo que concluiríamos que el peso promedio de los empaques es de 507 gramos, con una tendencia a variar por debajo o por encima de dicho peso en 12 gramos. Esta información le permite al gerente determinar cuesta medida nos permite determinar el promedio aritmético de fluctuación de los datos respecto a su punto central o media. La desviación estándar nos da como resultado un valor numérico que representa el promedio de diferencia que hay entre los datos y la media. Para calcular la desviación estándar basta con hallar la raíz cuadrada de la varianza, por lo tanto su ecuación sería:

$S = \sqrt{S^2}$ Ecuación 5-8

Para comprender el concepto de las medidas de distribución vamos a suponer que el gerente de una empresa de alimentos desea saber que tanto varían los pesos de los empaques (en gramos), de uno de sus productos; por lo que opta por seleccionar al azar cinco unidades de ellos para pesarlos. Los productos tienen los siguientes pesos (490, 500, 510, 515 y 520) gramos respectivamente.

Por lo que su media es:

$$\overline{X} = \frac{490+500+510+515+520}{5} = \frac{2535}{5} = 507$$

La varianza sería:

$$S^2 = \frac{(490-507)^2+(500-507)^2+(510-507)^2+(515-507)^2+(520-507)^2}{(5-1)}$$

$$S^2 = \frac{(-17)^2+(-7)^2+(3)^2+(8)^2+(13)^2}{4} = \frac{289+49+9+64+169}{4} = \frac{580}{4} = 145$$

Por lo tanto la desviación estándar sería:

$$S = \sqrt{145} = 12.04 \cong 12$$

Con lo que concluiríamos que el peso promedio de los empaques es de 507 gramos, con una tendencia a variar por debajo o por encima de dicho peso en 12 gramos. Esta información le permite al gerente determinar cuanto es el promedio de perdidas causado por el exceso de peso en los empaques y le da las bases para tomar los correctivos necesarios en el proceso de empacado.

Desviación media

Una medida de dispersión es aquella que por medio de un número nos indica que tanto están variando los datos recolectados con respecto a las medidas de tendencia central (principalmente la media).

En otras palabras, las medidas de dispersión nos indican si los datos recolectados en las encuestas son similares, o por el contrario, son muy diferentes entre ellos.

Las principales medidas de dispersión son: Desviación Media, Varianza y Desviación Estándar.

Ejemplo:

Considera los siguientes datos

7, 9, 8, 10, 9, 8, 9 y 10

Ahora bien, antes de empezar el cálculo de las medidas de dispersión, necesitamos calcular la Media.

Siempre tenemos que ordenar los datos antes de empezar. Resulta ser de gran ayuda al momento de hacer los cálculos.

Ordenamos: 7,8, 8, 9, 9, 9, 10, 10

Sumamos todos los datos y lo dividimos entre el número de datos.

$$\frac{7 + 8 + 8 + 9 + 9 + 9 + 10 + 10}{8} = 8.75$$

La Media Aritmética o Media es igual a 8.75.

$$Media = \bar{X} = 8.75$$

El número de veces que se repite un dato se llama frecuencia absoluta.

¿Cómo se obtiene?

La frecuencia del 7 es 1, porque aparece 1 vez.
La frecuencia del 8 es 2, porque aparece 2 veces.
La frecuencia del 9 es 3, porque aparece 3 veces.
La frecuencia del 10 es 2, porque aparece 2 veces.

La mayoría de las veces se hace uso de tablas para organizar la información:

X	Frecuencia (f)
7	1
8	2
9	3
10	2
Suma	8

Con la información obtenida, ya podemos empezar a calcular las Medidas de Tendencia Central.

- Desviación Media

La desviación media es un promedio de las desviaciones de todos los datos con respecto a la media.

Cabe señalar que dichas desviaciones siempre se van a considerar positivas.

La desviación del 7 es: 7-8.75 = -1.75
En positivo sería 1.75 (Valor absoluto
La desviación del 8 es: 8-8.75 = -0.75
En positivo sería 0.75

La desviación del 9 es: 9-8.75 = 0.25

Y…

La desviación del 10 es: 10-8.75 = 1.25

Segundo paso: Multiplicar por la frecuencia, sumar y dividir entre el número de datos.
Para obtener el promedio de las desviaciones, tenemos que considerar cuantas veces se está presentando la desviación, por lo que es necesario multiplicar por la frecuencia, sumar y por último dividir entre el número de datos:

$$Desviación\ Media = \frac{1(1.75) + 2(0.75) + 3(0.25) + 2(1.25)}{8}$$

Explica el proceso del cálculo de las medidas de tendencia central, localización y dispersión

Las medidas de tendencia central son medidas estadísticas que pretenden resumir en un solo valor a un conjunto de valores. Representan un centro en torno al cual se encuentra ubicado el conjunto de los datos. Las medidas de tendencia central más utilizadas son: media, mediana y moda. Las medidas de dispersión en cambio miden el grado de dispersión de los valores de la variable. Dicho en otros términos las medidas de dispersión pretenden evaluar en qué medida los datos difieren entre sí. De esta forma, ambos tipos de medidas usadas en conjunto permiten describir un conjunto de datos entregando información acerca de su posición y su dispersión.

Los procedimientos para obtener las medidas estadísticas difieren levemente dependiendo de la forma en que se encuentren los datos. Si los datos se encuentran ordenados en una tabla estadística diremos que se encuentran “agrupados” y si los datos no están en una tabla hablaremos de datos “no agrupados”.

Según este criterio, haremos primero el estudio de las medidas estadísticas para datos no agrupados y luego para datos agrupados.

AGRUPADOS:

Agrupados

$$Q_k = L_k + \frac{k\left(\frac{n}{4}\right) - F_k}{f_k} * c$$

Como los cuartiles adquieren su mayor importancia cuando contamos un número grande de datos y tenemos en cuenta que en estos casos generalmente los datos son resumidos en una tabla de frecuencia. La fórmula para el cálculo de los cuartiles cuando se trata de datos agrupados es la siguiente:

k= 1, 2,3
Dónde:
Lk = Límite real inferior de la clase del cuartil *k*
n = Número de datos
Fk = Frecuencia acumulada de la clase que antecede a la clase del cuartil *k*.
Fk = Frecuencia de la clase del cuartil *k*
c = Longitud del intervalo de la clase del cuartil *k*
Si se desea calcular cada cuartil individualmente, mediante otra fórmula se tiene lo siguiente:

- El primer cuartil Q1, es el menor valor que es mayor que una cuarta parte de los datos; es decir, aquel valor de la variable que supera 25% de las observaciones y es superado por el 75% de las observaciones.

  Fórmula de Q1, para series de Datos agrupados:

  $$Q_1 = l_i + \frac{P - f_{a-1}}{f_1} * I_c \quad P = \frac{n}{4}$$

  Dónde:
  L1 = límite inferior de la clase que lo contiene
  P = valor que representa la posición de la medida
  f1 = la frecuencia de la clase que contiene la medida solicitada.
  Fa-1 = frecuencia acumulada anterior a la que contiene la medida solicitada.
  Ic = intervalo de clase
- El segundo cuartil Q2, (coincide, es idéntico o similar a la mediana, Q2 = Md), es el menor valor que es mayor que la mitad de los datos, es decir el 50% de las observaciones son mayores que la mediana y el 50% son menores.

  Fórmula de Q2, para series de Datos agrupados:

  $$Q_1 = l_i + \frac{P - f_{a-1}}{f_1} * I_c \quad P = \frac{2n}{4}$$

  Dónde:
  L1 = límite inferior de la clase que lo contiene
  P = valor que representa la posición de la medida
  f1 = la frecuencia de la clase que contiene la medida solicitada.
  Fa-1 = frecuencia acumulada anterior a la que contiene la medida solicitada.
  Ic = intervalo de clase
- El tercer cuartil Q3, es el menor valor que es mayor que tres cuartas partes de los datos, es decir aquel valor de la variable que supera al 75% y es superado por el 25% de las observaciones.

  Fórmula de Q3, para series de Datos agrupados:

  $$Q_1 = l_i + \frac{P - f_{a-1}}{f_1} * I_c \quad P = \frac{3n}{4}$$

  Dónde:
  L1 = límite inferior de la clase que lo contiene
  P = valor que representa la posición de la medida

f1 = la frecuencia de la clase que contiene la medida solicitada.
Fa-1 = frecuencia acumulada anterior a la que contiene la medida solicitada.
Ic = intervalo de clase.
Otra manera de verlo es partir de que todas las medidas no son sino casos particulares del percentil, ya que el primer cuartil es el 25% percentil y el tercer cuartil 75% percentil.

No agrupados

Si se tienen una serie de valores X1, X2, X3... Xn, se localiza mediante las siguientes fórmulas:
- El primer cuartil:
Cuando n es par:

$$\frac{1*n}{4}$$

Cuando n es impar:

$$\frac{1(n+1)}{4}$$

- Para el tercer cuartil

Cuando n es par:

$$\frac{3*n}{4}$$

Cuando n es impar:

$$\frac{3(n+1)}{4}$$

DECILES
Los decirles son ciertos números que dividen la sucesión de datos ordenados en diez partes porcentualmente iguales. Son los nueve valores que dividen al conjunto de datos ordenados en diez partes iguales, son también un caso particular de los percentiles. Los decirles se denotan D1, D2,..., D9, que se leen primer dócil, segundo dócil, etc.
Los decirles, al igual que los cuartiles, son ampliamente utilizados para fijar el aprovechamiento académico.
Datos Agrupados
Para datos agrupados los decirles se calculan mediante la fórmula.

$$D_k = L_k + \frac{k\left(\frac{n}{10}\right) - F_k}{f_k} * c$$

k= 1, 2,3,... 9
Dónde:
Lk = Límite real inferior de la clase del dócil *k*
n = Número de datos
Fk = Frecuencia acumulada de la clase que antecede a la clase del dócil *k*.
Fe = Frecuencia de la clase del dócil *k*

c = Longitud del intervalo de la clase del dócil *k*

Otra fórmula para calcular los decirles:

- El cuarto dócil, es aquel valor de la variable que supera al 40%, de las observaciones y es superado por el 60% de las observaciones.

$$D_4 = l_i + \frac{P - f_{a-1}}{f_1} * I_c \quad P = \frac{4n}{10}$$

- El quinto dócil corresponde a la mediana.

$$D_5 = l_i + \frac{P - f_{a-1}}{f_1} * I_c \quad P = \frac{5n}{10}$$

- El noveno dócil supera al 90% y es superado por el 10% restante.

$$P = \frac{9n}{10}$$

$$D_9 = l_i + \frac{P - f_{a-1}}{f_1} * I_c$$

Donde (para todos):

L1 = límite inferior de la clase que lo contiene

P = valor que representa la posición de la medida

f1 = la frecuencia de la clase que contiene la medida solicitada.

Fa-1 = frecuencia acumulada anterior a la que contiene la medida solicitada.

Ic = intervalo de clase.

Fórmulas Datos No Agrupados

Si se tienen una serie de valores X1, X2, X3 ... Xn, se localiza mediante las siguientes fórmulas:

$\frac{A*n}{10}$ Cuando n es par:

$\frac{A(n+1)}{10}$ Cuando n es impar:

Siendo Al número del dócil.

CENTILES O PERCENTILES

Los percentiles son, tal vez, las medidas más utilizadas para propósitos de ubicación o clasificación de las personas cuando atienden características tales como peso, estatura, etc.

Los percentiles son ciertos números que dividen la sucesión de datos ordenados en cien partes porcentualmente iguales. Estos son los 99 valores que dividen en cien partes iguales el conjunto de datos ordenados. Los percentiles (P1, P2,... P99), leídos primer percentil,..., percentil 99.

Datos Agrupados

Cuando los datos están agrupados en una tabla de frecuencias, se calculan mediante la fórmula:

$$P_k = L_k + \frac{k\left(\frac{n}{100}\right) - F_k}{f_k} * c$$

k= 1, 2,3,... 99

Dónde:

Lk = Límite real inferior de la clase del dócil *k*

n = Número de datos

Fk = Frecuencia acumulada de la clase que antecede a la clase del decir *k*.

Fk = Frecuencia de la clase del decir *k*

c = Longitud del intervalo de la clase del decir *k*

Otra forma para calcular los percentiles es:

- Primer percentil, que supera al uno por ciento de los valores y es superado por el noventa y nueve por ciento restante.

$$P = \frac{1n}{100}$$

$$P_1 = l_i + \frac{P - f_{a-1}}{f_1} * I_c$$

- El 60 percentil, es aquel valor de la variable que supera al 60% de las observaciones y es superado por el 40% de las observaciones.

$$P_{60} = l_i + \frac{P - f_{a-1}}{f_1} * I_c \; P = \frac{60n}{100}$$

$$P_{99} = l_i + \frac{P - f_{a-1}}{f_1} * I_c \; P = \frac{99n}{100}$$

- El percentil 99 supera 99% de los datos y es superado a su vez por el 1% restante.

Fórmulas Datos No Agrupados

Si se tienen una serie de valores X1, X2, X3... Xn, se localiza mediante las siguientes fórmulas:

Para los percentiles, cuando n es par:

$$\frac{A*n}{10}$$

Explica el cálculo de las medidas de tendencia central, localización y dispersión con software

Las medidas de tendencia central son valores que se ubican al centro de un conjunto de datos ordenados según su magnitud. Generalmente se utilizan 4 de estos valores también conocidos como estadígrafos, la media aritmética, la mediana, la moda y al rango medio.

La media aritmética es la medida de posición utilizada con más frecuencia. Si se tienen n valores de observaciones, la media aritmética es la suma de todos y cada uno de los valores dividida entre el total de valores: Lo que indica que puede ser afectada por los valores extremos, por lo que puede dar una imagen distorsionada de la información de los datos.

La Mediana, es el valor que ocupa la posición central en un conjunto de datos, que deben estar ordenados, de esta manera la mitad de las observaciones es menor que la mediana y la otra mitad es mayor que la mediana, resulta muy apropiada cuando se poseen observaciones extremas.

La Moda es el valor de un conjunto de datos que aparece con mayor frecuencia. No depende de valores extremos, pero es más variables que la media y la mediana.

Rango Medio es la media de las observaciones menor y mayor. Como intervienen solamente estas observaciones, si hay valores extremos, se distorsiona como medida de posición, pero
Ofrece un valor adecuado, rápido y sencillo para resumir al conjunto de datos.

Datos Discretos

No Agrupados

Analicemos para ello las edades que utilizamos cuando se vio la organización y presentación de datos discretos:

12	15	14	15	16
18	19	14	15	17
15	17	18	16	19
16	17	15	15	17
16	18	17	19	17
23	16	17	18	19

Estos fueron los datos mostrados originalmente, no se han ordenado ni agrupado, determinemos ahora los valores de la Media, la Mediana y la moda, para ello recurramos a las fórmulas de estas medidas que resumimos en la siguiente tabla:

Medida	Formula	Observaciones
Media	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$	Donde x_i se refiere a todo y cada uno de los elementos de la muestra y n es el número total de elementos en la muestra.
Mediana	a) p = (n/2)	Es la posición en donde se encuentra la mediana. Si n es impar, entonces es la opción a, en caso contrario, la b. El valor de la mediana se obtiene por observación
	b) p = (n/2) + 1	
Moda		Se obtiene el valor por observación
Rango Medio	(Valor máximo + Valor Mínimo) / 2	

Aplicando, se obtienen los siguientes valores:

Para la media:

$$\bar{X} = \frac{12 + 15 + 14 + 15 + 16 + 18 + 19 + 14 + 15 + 17 + 15 + 17 + 18 + 16 + 19 + 16 + 17 + 15 + 15 + 17 + 16 + 18 + 17 + 19 + 17 + 23 + 16 + 17 + 18 + 19}{30}$$

$$\bar{X} = \frac{500}{30} = 16.6667$$

Para la mediana deberá ordenarse el grupo de datos, como n = 30, utilizaremos la posición p = (30/2) = 15, el primer valor mayor a 15 corresponde a la clase 17.

La moda estaría determinada por observación directa, y correspondería al valor 17, que se presenta hasta 7 veces en la muestra.

El rango medio se determina por la suma entre 23 y 12 dividido entre 2 (23 + 12)/2 = 35/2 = 17.5

Si observamos los valores obtenidos veremos que solo para el cálculo de la mediana se obtiene tuvo que ordenar la información (así lo específica la definición), sin embargo podemos también observar que este ordenamiento no afecta de manera directa ninguno de los cálculos, de esta manera se puede construir la siguiente tabla:

Medida	Valor Calculado	Observaciones
Media	16.6667	
Mediana	17	Se requirió el cálculo de la frecuencia acumulada
Moda	17	
Rango Medio	17.5	

Es de notar lo cercano de todos los valores que se han calculado, que circundan el valor de 17, no se notan cambios en los resultados comparados con los datos originales, sin embargo las formulas si se ven modificadas.

Agrupados

Recurramos ahora al agrupamiento de los datos discretos del ejercicio que hemos estado utilizando:

Clase	Repeticiones	Total de Años de la clase
12	1	12
14	2	28
15	6	90
16	5	80
17	7	119
18	4	72
19	4	76
23	1	23
Total	30	500

En donde podemos observar la suma de las frecuencias y de los años multiplicados por la clase que agrupa a los datos coinciden con los datos utilizados cuando no se agruparon en la sección anterior, utilizando ahora las formulas de la siguiente tabla:

Medida	Formula	Observaciones
Media	$\bar{x} = \frac{\sum_{i=1}^{n} x_i * f_i}{n}$	Donde x_i se refiere a todo y cada uno de los elementos de la muestra y n es el número total de elementos en la muestra y f_i se refiere a la frecuencia de la clase.

Mediana	p = (n/2)	Es la posición en donde se encuentra la mediana. Se ubica en la tabla el primer valor de frecuencia acumulada mayor a la posición calculada, si ese valor es mayor, entonces la mediana es la clase correspondiente al mismo. Si el valor es igual a la posición, entonces se suman el valor anterior más el valor obtenido y se divide entre 2.
Moda		Se obtiene el valor por observación de la mayor frecuencia
Rango Medio	(Valor máximo + Valor Mínimo) / 2	

Aplicando, se obtienen los siguientes valores:

Para la media:

$$\bar{X} = \frac{12 * 1 + 14 * 2 + 15 * 6 + 16 * 5 + 17 * 7 + 18 * 4 + 19 * 4 + 23 * 1}{30} = \frac{12 + 28 + 90 + 80 + 119 + 72 + 76 + 23}{30}$$

$$\bar{X} = \frac{500}{30} = 16.6667$$

UNIDAD 3

III. Estadística Inferencial

El alumno realizará estimaciones de datos estadísticos para contribuir a la toma de decisiones. A partir de un estudio de casos, el alumno elaborará un reporte técnico que contenga:

a) Una estimación puntual

b) Una estimación por intervalos

c) Prueba de hipótesis con:

- Establecimiento de hipótesis

- Criterio de aceptación

- Estadístico de prueba

- Conclusión

* A partir de un caso dado de su entorno profesional, realizar en software:

- Regresión lineal

- Pronóstico

- Prueba ANOVA

- Interpretación

- Conclusión

2. PRUEBA DE HIPOTESIS

Las secciones anteriores han mostrado cómo puede estimarse un parámetro a partir de los datos contenidos en una muestra. Puede encontrarse ya sea un sólo número (estimador puntual) o un intervalo de valores posibles (intervalo de confianza). Sin embargo, muchos problemas de ingeniería, ciencia, y administración, requieren que se tome una decisión entre aceptar o rechazar una proposición sobre algún parámetro. Esta proposición recibe el nombre de hipótesis. Este es uno de los aspectos más útiles de la inferencia estadística, puesto que muchos tipos de problemas de toma de decisiones, pruebas o experimentos en el mundo de la ingeniería, pueden formularse como problemas de prueba de hipótesis.

Una hipótesis estadística es una proposición o supuesto sobre los parámetros de una o más poblaciones.

Suponga que se tiene interés en la rapidez de combustión de un agente propulsor sólido utilizado en los sistemas de salida de emergencia para la tripulación de aeronaves. El interés se centra sobre la rapidez de combustión promedio. De manera específica, el interés recae en decir si la rapidez de combustión promedio es o no 50 cm/s. Esto puede expresarse de manera formal como

Ho; = 50 cm/s

H1; 50 cm/s

La proposición Ho; = 50 cm/s, se conoce como hipótesis nula, mientras que la proposición H1; 50 cm/s, recibe el nombre de hipótesis alternativa. Puesto que la hipótesis alternativa especifica valores de que pueden ser mayores o menores que 50 cm/s, también se conoce como hipótesis alternativa bilateral. En algunas situaciones, lo que se desea es formular una hipótesis alternativa unilateral, como en

Ho; μ= 50 cm/s Ho; μ= 50 cm/s

ó

H1; μ < 50 cm/s H1; μ > 50 cm/s

Es importante recordar que las hipótesis siempre son proposiciones sobre la población o distribución bajo estudio, no proposiciones sobre la muestra. Por lo general, el valor del parámetro de la población especificado en la hipótesis nula se determina en una de tres maneras diferentes:

Puede ser resultado de la experiencia pasada o del conocimiento del proceso, entonces el objetivo de la prueba de hipótesis usualmente es determinar si ha cambiado el valor del parámetro.

Puede obtenerse a partir de alguna teoría o modelo que se relaciona con el proceso bajo estudio. En este caso, el objetivo de la prueba de hipótesis es verificar la teoría o modelo.

Cuando el valor del parámetro proviene de consideraciones externas, tales como las especificaciones de diseño o ingeniería, o de obligaciones contractuales. En esta situación, el objetivo usual de la prueba de hipótesis es probar el cumplimiento de las especificaciones.

Un procedimiento que conduce a una decisión sobre una hipótesis en particular recibe el nombre de prueba de hipótesis. Los procedimientos de prueba de hipótesis dependen del empleo de la información contenida en la muestra aleatoria de la población de interés. Si esta información es consistente con la hipótesis, se concluye que ésta es verdadera; sin embargo si esta información es inconsistente

con la hipótesis, se concluye que esta es falsa. Debe hacerse hincapié en que la verdad o falsedad de una hipótesis en particular nunca puede conocerse con certidumbre, a menos que pueda examinarse a toda la población. Usualmente esto es imposible en muchas situaciones prácticas. Por tanto, es necesario desarrollar un procedimiento de prueba de hipótesis teniendo en cuenta la probabilidad de llegar a una conclusión equivocada.

2.1 Hipótesis

La hipótesis nula

Representada por Ho, es la afirmación sobre una o más características de poblaciones que al inicio se supone cierta (es decir, la "creencia a priori").

La hipótesis alternativa, representada por H1, es la afirmación contradictoria a Ho, y ésta es la hipótesis del investigador.

La hipótesis nula se rechaza en favor de la hipótesis alternativa, sólo si la evidencia muestral sugiere que Ho es falsa. Si la muestra no contradice decididamente a Ho, se continúa creyendo en la validez de la hipótesis nula. Entonces, las dos conclusiones posibles de un análisis por prueba de hipótesis son rechazar Ho o no rechazar Ho.

Prueba de una Hipótesis Estadística

Para ilustrar los conceptos generales, considere el problema de la rapidez de combustión del agente propulsor presentado con anterioridad. La hipótesis nula es que la rapidez promedio de combustión es 50 cm/s, mientras que la hipótesis alternativa es que ésta no es igual a 50 cm/s. Esto es, se desea probar:

Ho; $\mu = 50$ cm/s

H1; $\mu \geq 50$ cm/s

Supóngase que se realiza una prueba sobre una muestra de 10 especímenes, y que se observa cual es la rapidez de combustión promedio muestral. La media muestral es un estimador de la media verdadera de la población. Un valor de la media muestral que este próximo al valor hipotético = 50 cm/s es una evidencia de que el verdadero valor de la media es realmente 50 cm/s; esto es, tal evidencia apoya la hipótesis nula Ho. Por otra parte, una media muestral muy diferente de 50 cm/s constituye una evidencia que apoya la hipótesis alternativa H1. Por tanto, en este caso, la media muestral es el estadístico de prueba.

La media muestral puede tomar muchos valores diferentes. Supóngase que si 48.551.5, entonces no se rechaza la hipótesis nula Ho; $\mu = 50$ cm/s, y que si $\mu < 48.5$ ó $\mu > 51.5$, entonces se acepta la hipótesis alternativa H1; 50 cm/s.

Los valores de que son menores que 48.5 o mayores que 51.5 constituyen la región crítica de la prueba, mientras que todos los valores que están en el intervalo 48.551.5 forman la región de aceptación. Las fronteras entre las regiones críticas y de aceptación reciben el nombre de valores críticos. La costumbre es establecer conclusiones con respecto a la hipótesis nula Ho. Por tanto, se rechaza Ho en favor de H1 si el estadístico de prueba cae en la región crítica, de lo contrario, no se rechaza Ho.

Este procedimiento de decisión puede conducir a una de dos conclusiones erróneas. Por ejemplo, es posible que el valor verdadero de la rapidez promedio de combustión del agente propulsor sea igual a 50 cm/s. Sin embargo, para todos los especímenes bajo prueba, bien puede observarse un valor del estadístico de prueba que cae en la región crítica. En este caso, la hipótesis nula Ho será rechazada en favor de la alternativa H1cuando, de hecho, Ho en realidad es verdadero. Este tipo de conclusión equivocada se conoce como error tipo I.

El error tipo I se define como el rechazo de la hipótesis nula Ho cuando ésta es verdadera. También es conocido como o nivel de significancia.

Si tuviéramos un nivel de confianza del 95% entonces el nivel de significancia sería del 5%. Análogamente si se tiene un nivel de confianza del 90% entonces el nivel de significancia sería del 10%.

Ahora supóngase que la verdadera rapidez promedio de combustión es diferente de 50 cm/s, aunque la media muestral caiga dentro de la región de aceptación. En este caso se acepta Ho cuando ésta es falsa. Este tipo de conclusión recibe el nombre de error tipo II.

El error tipo II o error se define como la aceptación de la hipótesis nula cuando ésta es falsa.

Por tanto, al probar cualquier hipótesis estadística, existen cuatro situaciones diferentes que determinan si la decisión final es correcta o errónea.

Decisión	**H_o es verdadera**	**H_o es falsa**
Aceptar H_o	No hay error	Error tipo II ó β
Rechazar H_o	Error tipo I ó α	No hay error

Los errores tipo I y tipo II están relacionados. Una disminución en la probabilidad de uno por lo general tiene como resultado un aumento en la probabilidad del otro.

El tamaño de la región crítica, y por tanto la probabilidad de cometer un error tipo I, siempre se puede reducir al ajustar el o los valores críticos.

Un aumento en el tamaño muestral n reducirá y de forma simultánea.

Si la hipótesis nula es falsa, es un máximo cuando el valor real del parámetro se aproxima al hipotético. Entre más grande sea la distancia entre el valor real y el valor hipotético, será menor.

PRUEBAS DE HIPÓTESIS
Conceptos básicos para el procedimiento

Etapas básicas en pruebas de hipótesis. Al realizar pruebas de hipótesis, se parte de un valor supuesto (Hipotético) en parámetro poblacional. Después de recolectar una muestra aleatoria, se compara la estadística muestral, así como la media, con el parámetro hipotético, se compara con una supuesta media poblacional. Después se acepta o se rechaza el valor hipotético, según proceda. Se rechaza el valor hipotético sólo si el resultado muestral resulta muy poco probable cuando la hipótesis es cierta.

- Etapa 1. Planear la hipótesis nula y la hipótesis alternativa. La hipótesis nula (H_0) es el valor hipotético del parámetro que se compra con el resultado muestral resulta muy poco probable cuando la hipótesis es cierta.
- Etapa 2. Especificar el nivel de significancia que se va a utilizar. El nivel de significancia del 5%, entonces se rechaza la hipótesis nula solamente si el resultado muestral es tan diferente del valor hipotético que una diferencia de esa magnitud o mayor, pudiera ocurrir aleatoria mente con una probabilidad de 1.05 o menos.
- Etapa 3. Elegir la estadística de prueba. La estadística de prueba puede ser la estadística muestral (el estimador no segado del parámetro que se prueba) o una versión transformada de esa estadística muestral. Por ejemplo, para probar el valor hipotético de una media poblacional, se toma la media de una muestra aleatoria de esa distribución normal, entonces es común que se transforme la media en un valor z el cual, a su vez, sirve como estadística de prueba.

Consecuencias de las Decisiones en Pruebas de Hipótesis.

Decisiones Posibles	Situaciones Posibles	
	La hipótesis nula es verdadera	La hipótesis nula es falsa
Aceptar la Hipótesis Nula	Se acepta correctamente	Error tipo II o Beta
Rechazar la Hipótesis Nula	Error tipo I o Alfa	Se rechaza correctamente

- Etapa 4. Establecer el valor o valores críticos de la estadística de prueba. Habiendo especificado la hipótesis nula, el nivel de significancia y la estadística de prueba que se van a utilizar, se produce a establecer el o los valores críticos de estadística de prueba. Puede haber uno o más de esos valores, dependiendo de si se va a realizar una prueba de uno o dos extremos.

- Etapa 5. Determinar el valor real de la estadística de prueba. Por ejemplo, al probar un valor hipotético de la media poblacional, se toma una muestra aleatoria y se determina el valor de la media muestral. Si el valor crítico que se establece es un valor de z, entonces se transforma la media muestral en un valor de z.

- Etapa 6. Tomar la decisión. Se compara el valor observado de la estadística muestral con el valor (o valores) críticos de la estadística de prueba. Después se acepta o se rechaza la hipótesis nula. Si se rechaza ésta, se acepta la alternativa; a su vez, esta decisión tendrá efecto sobre otras decisiones de los administradores operativos, como por ejemplo, mantener o no un estándar de desempeño o cuál de dos estrategias de mercadotecnia utilizar.

La distribución apropiada de la prueba estadística se divide en dos regiones: una región de rechazo y una de no rechazo. Si la prueba estadística cae en esta última región no se puede rechazar la hipótesis nula y se llega a la conclusión de que el proceso funciona correctamente.

Al tomar la decisión con respecto a la hipótesis nula, se debe determinar el valor crítico en la distribución estadística que divide la región del rechazo (en la cual la hipótesis nula no se puede rechazar) de la región de rechazo. A hora bien el valor crítico depende del tamaño de la región de rechazo.

Pasos de la Prueba de Hipótesis

- Expresar la hipótesis nula
- Expresar la hipótesis alternativa
- Especificar el nivel de significancia
- Determinar el tamaño de la muestra
- Establecer los valores críticos que establecen las regiones de rechazo de las de no rechazo.
- Determinar la prueba estadística.
- Coleccionar los datos y calcular el valor de la muestra de la prueba estadística apropiada.
- Determinar si la prueba estadística ha sido en la zona de rechazo a una de no rechazo.
- Determinar la decisión estadística.
- Expresar la decisión estadística en términos del problema.

Hipótesis Estadística. Al intentar alcanzar una decisión, es útil hacer hipótesis (o conjeturas) sobre la población aplicada. Tales hipótesis, que pueden ser o no ciertas, se llaman hipótesis estadísticas. Son, en general, enunciados acerca de las distribuciones de probabilidad de las poblaciones.

Hipótesis Nula. En muchos casos formulamos una hipótesis estadística con el único propósito de rechazarla o invalidarla. Así, si queremos decidir si una moneda está trucada, formulamos la hipótesis de que la moneda es buena (o sea p=0,5, donde p es la probabilidad de cara). Analógicamente, si deseamos decidir si un procedimiento es mejor que otro, formulamos la hipótesis de que no hay diferencia entre ellos (o sea. Que cualquier diferencia observada se debe simplemente a fluctuaciones en el muestreo de la misma población). Tales hipótesis se suelen llamar hipótesis nula y se denotan por H_o.

Para todo tipo de investigación en la que tenemos dos o más grupos, se establecerá una hipótesis nula. La hipótesis nula es aquella que nos dice que no existen diferencias significativas entre los grupos. Por ejemplo, supongamos que un investigador cree que si un grupo de jóvenes se somete a un entrenamiento intensivo de natación, éstos serán mejores nadadores que aquellos que no recibieron entrenamiento. Para demostrar su hipótesis toma al azar una muestra de jóvenes, y también al azar los distribuye en dos grupos: uno que llamaremos experimental, el cual recibirá entrenamiento, y otro que no recibirá entrenamiento alguno, al que llamaremos control. La hipótesis nula señalará que no hay diferencia en el desempeño de la natación entre el grupo de jóvenes que recibió el entrenamiento y el que no lo recibió.

Una hipótesis nula es importante por varias razones:

- Es una hipótesis que se acepta o se rechaza según el resultado de la investigación.
- El hecho de contar con una hipótesis nula ayuda a determinar si existe una diferencia entre los grupos, si esta diferencia es significativa, y si no se debió al azar.
- No toda investigación precisa de formular hipótesis nula. Se recomienda que la hipótesis nula es aquella por la cual indicamos que la información a obtener es contraria a la hipótesis de trabajo.

Al formular esta hipótesis, se pretende negar la variable independiente. Es decir, se enuncia que la causa determinada como origen del problema fluctúa, por tanto, debe rechazarse como tal.

Hipótesis Alternativa: Toda hipótesis que difiere de una dada se llamará una hipótesis alternativa. Una hipótesis alternativa a la hipótesis nula se denotará por H_1.

Al responder a un problema, es muy conveniente proponer otras hipótesis en que aparezcan variables independientes distintas de las primeras que formulamos. Por tanto, para no perder tiempo en búsquedas inútiles, es necesario hallar diferentes hipótesis alternativas como respuesta a un mismo problema y elegir entre ellas cuáles y en qué orden vamos a tratar su comprobación.

Las hipótesis, naturalmente, serán diferentes según el tipo de investigación que se esté realizando. En los estudios exploratorios, a veces, el objetivo de la investigación podrá ser simplemente el de obtener los mínimos conocimientos que permitan formular una hipótesis. También es aceptable que, en este caso, resulten poco precisas, como cuando afirmamos que "existe algún tipo de problema social en tal grupo", o que los planetas poseen algún tipo de atmósfera, sin especificar de qué elementos está compuesto.

Los trabajos de índole descriptiva generalmente presentan hipótesis del tipo "todos los X poseen, en alguna medida, las característica Y". Por ejemplo, podemos decir que todas las naciones poseen algún comercio internacional, y dedicarnos a describir, cuantificando, las relaciones comerciales entre ellas. También podemos hacer afirmaciones del tipo "X pertenece al tipo Y", como cuando decimos que una tecnología es capital - intensiva. En estos casos, describimos, clasificándolo, el objeto de nuestro interés, incluyéndolo en un tipo ideal complejo de orden superior.

Por último, podemos construir hipótesis del tipo "X produce (o afecta) a Y", donde estaremos en presencia de una relación entre variables.

Errores de tipo I y de tipo II. Si rechazamos una hipótesis cuando debiera ser aceptada, diremos que se ha cometido un error de tipo I. Por otra parte, si aceptamos una hipótesis que debiera ser rechazada, diremos que se cometió un error de tipo II.

En ambos casos, se ha producido un juicio erróneo. Para que las reglas de decisión (o no contraste de hipótesis) sean buenos, deben diseñarse de modo que minimicen los errores de la decisión; y no es una cuestión sencilla, porque para cualquier tamaño de la muestra, un intento de disminuir un tipo de error suele ir acompañado de un crecimiento del otro tipo. En la práctica, un tipo de error puede ser más grave que el otro, y debe alcanzarse un compromiso que disminuya el error más grave. La única forma de disminuir ambos a la vez es aumentar el tamaño de la muestra que no siempre es posible.

Niveles de Significación. Al contrastar una cierta hipótesis, la máxima probabilidad con la que estamos dispuesto a correr el riesgo de cometerán error de tipo I, se llama nivel de significación. Esta probabilidad, denota a menudo por se, suele especificar antes de tomar la muestra, de manera que los resultados obtenidos no influyan en nuestra elección.

En la práctica, es frecuente un nivel de significación de 0,05 ó 0,01, si bien se une otros valores. Si por ejemplo se escoge el nivel de significación 0,05 (ó 5%) al diseñar una regla de decisión, entonces hay unas cinco (05) oportunidades entre 100 de rechazar la hipótesis cuando debiera haberse aceptado; Es decir, tenemos un 95% de confianza de que hemos adoptado la decisión correcta. En tal caso decimos que la hipótesis ha sido rechazada al nivel de significación 0,05, lo cual quiere decir que tal hipótesis tiene una probabilidad 0,05 de ser falsa.

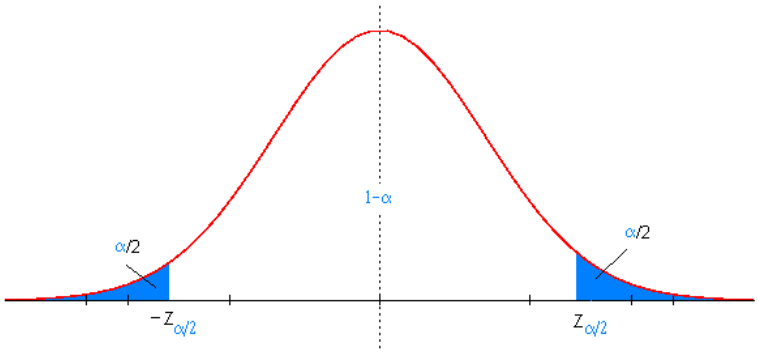
Prueba de 1 o 2 Extremos. Cuando estudiamos ambos valores estadísticos es decir, ambos lados de la media lo llamamos prueba de uno y dos extremos o contraste de una y dos colas. Con frecuencia no obstante, estaremos interesados tan sólo en valores extremos a un lado de la media (o sea, en uno de los extremos de la distribución), tal como sucede cuando se contrasta la hipótesis de que un proceso es mejor que otro (lo cual no es lo mismo que contrastar si un proceso es mejor o peor que el otro) tales contrastes se llaman unilaterales, o de un extremo. En tales situaciones, la región crítica es una región situada a un lado de la distribución, con área igual al nivel de significación.

Curva Característica Operativa y Curva de Potencia. Podemos limitar un error de tipo I eligiendo adecuadamente el nivel de significancia. Es posible evitar el riesgo de cometer el error tipo II simplemente no aceptando nunca la hipótesis, pero en muchas aplicaciones prácticas esto es inviable. En tales casos, se suele recurrir a curvas características de operación o curvas de potencia que son gráficos que muestran las probabilidades de error de tipo II bajo diversas hipótesis. Proporcionan indicaciones de hasta qué punto un test dado nos permitirá evitar un error de tipo II; es decir, nos indicarán la potencia de un test a la hora de prevenir decisiones erróneas. Son útiles en el diseño de experimentos por que sugieren entre otras cosas el tamaño de muestra a manejar.

Inferencias acerca de la Media Poblacional (varianza conocida). Supongamos que de una población normal con media desconocida μ. Y varianza conocida σ^2 se extrae una muestra de tamaño n, entonces de la distribución de la media muestral $\bar{x}$ se obtiene que:

$$Z = \frac{\bar{x} - \mu_o}{\sigma / \sqrt{n}}$$

Se distribuye como una normal estándar. Luego, $P(-Z_{\alpha/2} < Z < Z_{\alpha/2}) = 1 - \alpha$



Donde $Z_{\alpha/2}$ es un valor de la normal estándar tal que el área a la derecha de dicho valor es $\alpha/2$, como se muestra en la figura

Sustituyendo la fórmula de z se obtiene:

$$P\left(-Z_{\alpha/2} < \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} < Z_{\alpha/2}\right) = 1-\alpha$$

Haciendo un despeje algebraico, se obtiene

$$P\left(\mu - \frac{Z_{\alpha/2}*\sigma}{\sqrt{n}} < \bar{x} < \mu + \frac{Z_{\alpha/2}*\sigma}{\sqrt{n}}\right) = 1-\alpha$$

De lo anterior se puede concluir que un Intervalo de Confianza del 100(1-α)% para la media poblacional μ, es de la forma:

$$\left(\bar{x} - \frac{Z_{\alpha/2}*\sigma}{\sqrt{n}}, \bar{x} + \frac{Z_{\alpha/2}*\sigma}{\sqrt{n}}\right)$$

Usualmente α=0.1, 0.05 ó 0.01, que corresponden a intervalos de confianza del 90, 95 y 99 por ciento respectivamente. La siguiente tabla muestra los $Z_{\alpha/2}$ más usados.

Nivel de Confianza	$Z_{\alpha/2}$
90	1.645

95	1.96
99	2.58

En la práctica si la media poblacional es desconocida entonces, es bien probable que la varianza también lo sea puesto que en el cálculo de σ^2 interviene μ. Si ésta es la situación, y si el tamaño de muestra es grande (n>30, parece ser lo más usado), entonces σ^2 es estimada por la varianza muestral s^2 y se puede usar la siguiente fórmula para el intervalo de confianza de la media poblacional:

$$\left(\bar{x} - \frac{Z_{\alpha/2} * s}{\sqrt{n}}, \bar{x} + \frac{Z_{\alpha/2} * s}{\sqrt{n}} \right)$$

Por otro lado, también se pueden hacer pruebas de hipótesis con respecto a la media poblacional μ. Por conveniencia, en la hipótesis nula siempre se asume que la media es igual a un valor dado. La hipótesis alterna en cambio, puede ser de un sólo lado: menor ó mayor que el número dado, ó de dos lados: distinto a un número dado.
Existen dos métodos de hacer la prueba de hipótesis: el método clásico y el método del P-Value.

a. En el método clásico, se evalúa la prueba estadística de Z y al valor obtenido se le llama Z calculado (Z_{calc}). Por otro lado el nivel de significación α dado determina una región de rechazo y una de aceptación. Si Z_{calc} cae en la región de rechazo, entonces se concluye que hay suficiente evidencia estadística para rechazar la hipótesis nula con base en los resultados de la muestra tomada. Las fórmulas están resumidas en la siguiente tabla:

Caso I	Caso II	Caso III
H_o: $\mu=\mu_0$	H_o : $\mu=\mu_0$	H_o : $\mu=\mu_0$
H_a: $\mu<\mu_0$	H_a : $\mu \neq \mu_0$	H_a : $\mu>\mu_0$

Prueba Estadística: $Z = \dfrac{\bar{x} - \mu_o}{\sigma / \sqrt{n}}$

Aquí Z_α es el valor de la normal estándar tal que el área a la derecha de dicho valor es α. Recordar también que σ puede ser sustituido por *s,* cuando la muestra es relativamente grande (n>30). Los valores de α más usados son 0.01 y 0.05. Si se rechaza la hipótesis nula al .01 se dice que la hipótesis alterna es altamente significativa y al .05 que es significativa.

b. Trabajar sólo con esos dos valores de α simplificaba mucho el aspecto computacional, pero por otro lado creaba restricciones. En la manera moderna de probar hipótesis se usa una cantidad llamada P-Valué. El P-Valué llamado el nivel de significación observado, es el valor de α al cual se rechazaría la hipótesis nula si se usa el valor calculado de la prueba estadística. En la práctica un P-Value cercano a 0 indica un rechazo de la hipótesis nula. Así un P-Value menor que .05 indicará que se rechaza la prueba estadística.

   Fórmulas para calcular P-Value:

- Si H_o: $\mu > \mu_o$, entonces P-value = 1*Prob($Z > Z_{calc}$).
- Si H_o: $\mu < \mu_o$, entonces P-value = 1*Prob($Z < Z_{calc}$).
- Si H_o: $\mu \neq \mu_o$, entonces P-value = 2*Prob($Z > |Z_{calc}|$).

   Los principales programas estadísticos dan los P-Value para la mayoría de las pruebas estadísticas. A través de todo el texto usamos el método del P-Value para probar hipótesis.

Concepto. Afirmación acerca de los parámetros de la población.

PRUEBAS DE HIPÓTESIS PARA LA MEDIA Y PROPORCIONES

Debido a la dificultad de explicar este tema se enfocará un problema basado en un estudio en una fábrica de llantas. En este problema la fábrica de llantas tiene dos turnos de operarios, turno de día y turno mixto. Se selecciona una muestra aleatoria de 100 llantas producidas por cada turno para ayudar al gerente a sacar conclusiones de cada una de las siguientes preguntas

- ¿Es la duración promedio de las llantas producidas en el turno de día igual a 25 000 millas?
- ¿Es la duración promedio de las llantas producidas en el turno mixto menor de 25 000 millas?
- ¿Se revienta más de un 8% de las llantas producidas por el turno de día antes de las 10 000 millas?

Prueba de Hipótesis para la media. En la fábrica de llantas la hipótesis nula y alternativa para el problema se plantearon como,

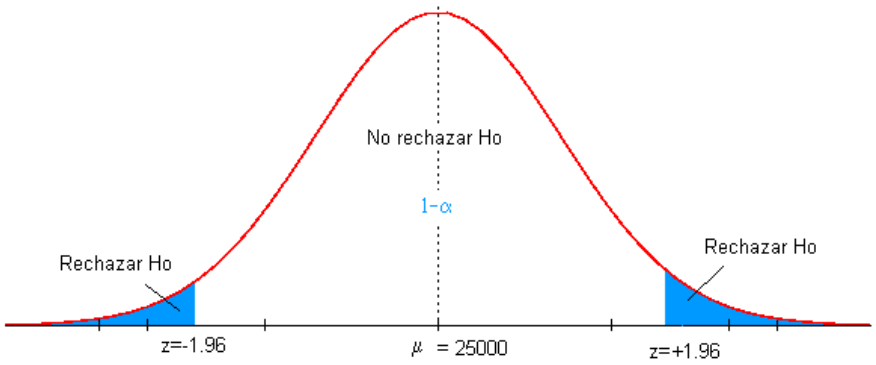
H_o: $\mu = 25\,000$ H_1: $\mu \neq 25\,000$

Si se considera la desviación estándar σ las llantas producidas en el turno de día, entonces, con base en el teorema de limite central, la distribución en el muestreo de la media seguiría la distribución normal, y la prueba estadística que esta basada en la diferencia entre la media $\overline{X}$ de la muestra y la media μ hipotética se encontrara como

$$Z = \frac{\bar{x} - \mu_o}{\sigma / \sqrt{n}}$$

Si el tamaño de la región α de rechazo se estableciera en 5% entonces se podrían determinar los valores críticos de la distribución. Dado que la región de rechazo está dividida en las dos colas de la distribución, el 5% se divide en dos partes iguales de 2.5%.

Dado que ya se tiene la distribución normal, los valores críticos se pueden expresar en unidades de desviación. Una región de rechazo de 0.25 en cada cola de la distribución normal, da por resultado un área de .475 entre la media hipotética y el valor crítico. Si se busca está área en la distribución normal, se encuentra que los valores críticos que dividen las regiones de rechazo y no rechazo son + 1.96 y - 1.96



Por tanto, la regla para decisión sería rechazar H_o si Z > +1.96 o sí z < -1.96, de lo contrario, no rechazar H_o. No obstante, en la mayor parte de los casos se desconoce la desviación estándar σ de la población. La desviación estándar se estima al calcular S, la desviación estándar de la muestra. Si se supone que la población es normal la distribución en el muestreo de la media seguiría una distribución t con n-1 grados de libertad.

En la práctica, se ha encontrado que siempre y cuando el tamaño de la muestra no sea muy pequeño y la población no esté muy sesgada, la distribución t da una buena aproximación a la distribución de muestra de la media. La prueba estadística para determinar la diferencia entre la media $\bar{X}$ de la muestra y la media μ de la población cuando se utiliza la desviación estándar S de la muestra, se expresa

$$t_{n-1} = \frac{\bar{X} - \mu}{s / \sqrt{n}}$$

Para una muestra de 100, si se selecciona un nivel de significancia de 0.05, los valores críticos de la distribución t con 100-1= 99 grados de libertad se puede obtener como se indica en la siguiente tabla tenemos el valor de 1.9842. Como esta prueba de dos colas, la región de rechazo de 0.05 se vuelve a dividir en dos partes iguales de 0.025 cada una. Con el uso de las tablas para t, los valores críticos son –1.984 y +1.984. La regla para la decisión es,

Rechazar H_o si $t_{99} > +1.9842$ o $t_{99} < -1.9842$ de lo contrario, no rechazar H_o

Los resultados de la muestra para el turno de día (en millas) fueron $\overline{X}_{dìa} = 25.430$, $S_{dìa} = 4.000$ y $n_{día} = 100$ millas. Puesto que se está probando si la media es diferente a 25 000 millas, se tiene con la ecuación

$$t_{n-1} = \frac{\overline{X} - \mu}{S/\sqrt{n}} \qquad t_{100-1} = \frac{25.430 - 25.00}{4.000/\sqrt{100}} = 1.075$$

Dado que t_{100-1}=1.075, se ve que -1.984 < +1.075 < + 1.984, entonces no se rechaza Ho.

Por ello, la decisión de no rechazar la hipótesis nula Ho. En conclusión es que la duración promedio de las llantas es 25 000 millas. A fin de tener en cuenta la posibilidad de un error de tipo II, este enunciado se puede redactar como no hay pruebas de que la duración promedio de las llantas sea diferente a 25 000 millas en las llantas producidas en el turno de día.

PRUEBA DE HIPÓTESIS PARA PROPORCIONES

El concepto de prueba de hipótesis se puede utilizar para probar hipótesis en relación con datos cualitativos. Por ejemplo, en el problema anterior el gerente de la fábrica de llantas quería determinar la proporción de llantas que se reventaban antes de 10.000 millas. Este es un ejemplo de una variable cualitativa, dado que se desea llegar a conclusiones en cuanto a la proporción de los valores que tienen una característica particular.

El gerente de la fábrica de llantas quiere que la calidad de llantas producidas, sea lo bastante alta para que muy pocas se revienten antes de las 10.000 millas. Si más de un 8% de las llantas se revientan antes de las 10.000 millas, se llegaría a concluir que el proceso no funciona correctamente. La hipótesis nula y alternativa se pueden expresar como sigue:

$H_0 = P \leq 0.08$ (Funciona correctamente)

$H_1 = P > 0.08$ (No funciona correctamente)

La prueba estadística se puede expresar en términos de la proporción de éxitos como sigue:

$$Z = \frac{P_s - P}{\sqrt{\frac{Pq}{n}}} \rightarrow P_s = \frac{X}{n}$$

Siendo X y N el número de éxitos de la muestra y n el tamaño de la muestra, P la proporción de éxitos de la hipótesis nula. Ahora se determinará si el proceso funciona correctamente para las llantas producidas para el turno de día. Los resultados del turno de día indican que cinco llantas en una muestra de 100 se reventaron antes de 10,000 millas para este problema, si se selecciona un nivel de significancia $\alpha = 0.05$, las regiones de rechazo y no rechazo se establecerían como a continuación se muestra. Y la regla de decisión sería: Rechazar Ho si z> + 1.645; de lo contrario no rechazar H_o. Con los datos que se tienen,

$$P_s = 0.05 \Rightarrow Z = \frac{P_s - P}{\sqrt{\frac{Pq}{n}}} = -1.107$$

Una vez reemplazado, recuerde p+q=1
Z=-1.107 +1.645; por tanto no rechazar H_o.

La hipótesis nula no se rechazaría por que la prueba estadística no ha caído en la región de rechazo. Se llegaría a la conclusión de que no hay pruebas de que más del 8% de las llantas producidas en el turno de día se revienten antes de 10,000 millas. El gerente no ha encontrado ninguna prueba de que ocurra un número excesivo de reventones en las llantas producidas en el turno de día.

Una hipótesis estadística es una suposición hecha con respecto a la función de distribución de una variable aleatoria. Para establecer la verdad o falsedad de una hipótesis estadística con certeza total, será necesario examinar toda la población. En la mayoría de las situaciones reales no es posible o practico efectuar este examen, y el camino más aconsejable es tomar una muestra aleatoria de la población y en base a ella, decidir si la hipótesis es verdadera o falsa.

En la prueba de una hipótesis estadística, es costumbre declarar la hipótesis como verdadera si la probabilidad calculada excede el valor tabular llamado el nivel de significación y se declara falsa si la probabilidad calculada es menor que el valor tabular. La prueba a realizar dependerá del tamaño de las muestras, de la homogeneidad de las varianzas y de la dependencia o no de las variables. Si las muestras a probar involucran a más de 30 observaciones, se aplicará la prueba de Z, si las muestras a evaluar involucran un número de observaciones menor o igual que 30 se emplea la prueba de t de student. La fórmula de cálculo depende de si las varianzas son homogéneas o heterogéneas, si el número de observaciones es igual o diferente, o si son variables dependientes.

Para determinar la homogeneidad de las varianzas se toma la varianza mayor y se divide por la menor, este resultado es un estimado de la F de Fisher. Luego se busca en la tabla de F usando como numerador los grados de libertad (n-1) de la varianza mayor y como denominador (n-1) de la varianza menor para encontrar la F de Fisher tabular. Si la F estimada es menor que la F tabular se declara que las varianzas son homogéneas. Si por el contrario, se declaran las varianzas heterogéneas. Cuando son variables dependientes (el valor de una depende del valor de la otra), se emplea la técnica de pruebas pareadas.

Como en general estas pruebas se aplican a dos muestras, se denominarán a y b para referirse a ellas, así entenderemos por:

- n_a al número de elementos de la muestra a
- n_b al número de elementos de la muestra b
- x_b al promedio de la muestra b
- s^2_a la varianza de la muestra a
- Y así sucesivamente

Entonces se pueden distinguir 6 casos a saber:

- Caso de muestras grandes (n>30)
- Caso de $n_a = n_b$ y $s^2_a = s^2_b$
- Caso de $n_a = n_b$ y $s^2_a \neq s^2_b$
- Caso de $n_a \neq n_b$ y $s^2_a = s^2_b$
- Caso de $n_a \neq n_b$ y $s^2_a \neq s^2_b$
- Caso de variables dependientes

1.-Cuando las muestras a probar involucran observaciones y a más de 30 observaciones homogéneas $$z_c = \frac{a\overline{X} - b\overline{X}}{\sqrt{\frac{as^2}{na} + \frac{bs^2}{nb}}}$$	2.-Caso de número igual de varianzas $$t_c = \frac{a\overline{X} - b\overline{X}}{\sqrt{2\frac{as^2 + bs^2}{(2/n)}}}$$
3.-Caso de igual número de observaciones observación- Y varianzas heterogéneas. $$t_c = \frac{a\overline{X} - b\overline{X}}{\sqrt{\frac{as^2 + bs^2}{n}}}$$	4.-Caso de diferente número de Nes y varianzas homogéneas $$t_c = \frac{a\overline{X} - b\overline{X}}{\sqrt{\frac{cs^2}{an} + \frac{cs^2}{bn}}}$$

5.- Caso de diferente número de observaciones y varianzas heterogéneas. En este caso, la t_c es comparada con la t_g (t generada), que a diferencia de los casos anteriores, hay que calcularla.

$$t_g = \frac{t_a \frac{s_a^2}{n_a} + t_b \frac{s_b^2}{n_b}}{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_n}} \qquad t_c = \frac{\overline{X}_a - \overline{X}_b}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_n}}}$$

Dónde: t_a y t_b son los valores de la tabla con n-1 grados de libertad para a y b respectivamente

6.- Caso de muestras pareadas (de variables dependientes). En este caso, se asume que las muestras han sido distribuidas por pares.

$$t_c = \frac{\sum D / n}{\sqrt{\frac{\sum (D - \overline{D})^2}{n-1}} \Big/ \sqrt{n}}$$

TEST DE HIPÓTESIS ESTADÍSTICA

En la sección anterior tratamos la estimación y precisión de los estimadores, que conforman una de las dos áreas principales de la Inferencia estadística. En esta sección presentaremos una forma diferente de obtener inferencia acerca de parámetros poblacionales, probando hipótesis respecto a sus valores. Un test de hipótesis es una metodología o procedimiento que permite cuantificar la probabilidad del error que se cometería cuando se hace una afirmación sobre la población bajo estudio, es decir, nos permite medir la fuerza de la evidencia que tienen los datos a favor o en contra de alguna hipótesis de interés sobre la población.

Ejemplo. Una industria usa como uno de los componentes de las máquinas de producción una lámpara especial importada que debe satisfacer algunas exigencias. Una de esas exigencias está relacionada a su vida útil en horas. Esas lámparas son fabricadas por dos países y las especificaciones técnicas varían de país a país. Por ejemplo el catálogo del producto americano afirma que la vida útil media de sus lámparas es de 15500 horas, con un SD de 1200. Mientras que para el producto europeo la media es de 16500, y el SD es de 2000.

Un lote de esas lámparas de origen desconocido es ofrecido a un precio muy conveniente. Para que la industria sepa si hace o no una oferta ella necesita saber

cuál es el país que produjo tales lámparas. El comercio que ofrece tales lámparas afirma que será divulgada la vida útil media de una muestra de 25 lámparas del lote antes de la oferta. ¿Qué regla de decisión deben usar los responsables de la industria para decir que las lámparas son de procedencia americana o europea?. Una respuesta que surge inmediatamente es la de considerar como país productor aquel en la cual la media de la muestra se aproxima más a la media de la población. Así, la decisión sería si $\bar{x} \leq 16000$ (el punto medio entre 15500 y 16500) diremos que es de procedencia americana; en caso contrario diremos que es de procedencia europea.

Suponga que en el día de la licitación se informó que, de acuerdo con la regla de decisión diríamos que las lámparas son de origen americano. ¿Podemos estar herrados en esa conclusión? O en otras palabras, ¿es posible que una muestra de 25 lámparas de origen europeo presente una media de 15800? Sí, es posible. Entonces, para un mejor entendimiento de la regla de decisión adoptada, es interesante estudiar los tipos de errores que podemos cometer y las respectivas probabilidades de cometer esos errores.

Los test de hipótesis consisten en confrontar dos hipótesis, una llamada hipótesis nula que denotamos con H_o y otra llamada hipótesis alternativa denotada con H_1. En el ejemplo las hipótesis que se plantean son:

En el ejemplo las hipótesis consideradas son
H_o Las lámparas son de origen europeo, esto equivale a decir que la vida útil X de cada lámpara sigue una distribución con media μ=16500 horas y un SD=2000 horas.
H_1; Las lámparas son de origen americano, es decir la media poblacional μ= 15500 horas con un SD=1200 horas.

Bajo este planteo un test de hipótesis estadística no es otra cosa que un procedimiento para tomar una decisión, bajo incertidumbre, sobre la validez de la hipótesis nula usando la evidencia de los datos. Puesto que trabajamos bajo incertidumbre es claro que cualquiera sea la decisión que tomemos siempre existe una probabilidad de cometer error. A fin de clarificar esto podemos presentar el siguiente esquema:

Esquema del procedimiento

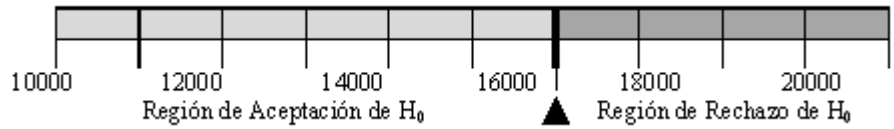
Decisión	Realidad sobre H_o	
	Cierta	Falsa
Rechazar H_o	Error Tipo I	Decisión correcta
No rechazar H_o	Decisión correcta	Error Tipo II

Como se puede ver en el esquema, con cada tipo de decisión que se tome hay asociado una posibilidad de cometer un error. Un procedimiento de este tipo sería óptimo cuando las probabilidades de cometer un error, cualquiera sea la decisión que se adopte, sean pequeñas. Lamentablemente, en la mayoría de los tests de

hipótesis sólo es posible controlar una de ellas, con la circunstancia agravante de que estos errores son competitivos, es decir, cuando se disminuye mucho la probabilidad de uno aumenta la probabilidad del otro.

Puesto que, el interés generalmente es "rechazar H_o" la probabilidad de error que se controla durante este procedimiento, es justamente el error asociado a esta decisión (Probabilidad del Error Tipo I), es decir, la probabilidad de rechazar H_0 cuando es cierta. La máxima probabilidad de error tipo I se denota con α y recibe el nombre de nivel de significación del test y él debe ser prefijado de antemano. La probabilidad de Error Tipo II se denota con β y es útil para encontrar la bondad del test que se mide en términos de la cantidad 1-β denominada Poder del Test.

El nivel de significación que se usa generalmente es α=0.05 lo que corresponde a un 5% en término de porcentaje.
Retomando el ejemplo vamos a indicar por RC una región determinada por los valores de X menores que 16000, es decir RC= {X $\leq$ 16000}. El valor 16000 se denomina punto crítico y se denotará como x_c.



Con las notaciones indicadas arriba, la probabilidad de cometer cada uno de los errores puede ser escrito del siguiente modo:

- P[Error Tipo I] = P[$\overline{X}$ pertenezca a RC | H_0 es verdadera] = α.
- P[Error Tipo II] = P[$\overline{X}$ no pertenezca a RC | H_0 es falsa] = β

Ejemplo. En el ejemplo anterior, cuando H_0 es verdadera, es decir, las lámparas son de origen europea, sabemos del teorema central del límite que $\overline{x}$, o sea la media de las muestras de tamaño 25, tendrán distribución aproximadamente normal con media 16500 y $\sigma = \frac{2000}{\sqrt{25}} = 400$, es decir $\overline{X} \sim$ N(16500, 1600).

Entonces,

P[Error Tipo I] = P[$\overline{X} \in$ RC | H_0 es verdadera] =
= P [$\overline{X} \leq$ 16000 | $\overline{X} \sim$ N(16500, 1600)] = P[Z $\leq$ (16000 – 16500)/ 400]
= P[Z $\leq$ -1.25] = 0.106 = 10.6%.

Para cada regla de decisión adoptada, es decir, para cada valor crítico x_c se obtiene un valor de probabilidad de error tipo 1. Por otra parte, si x_c se elige menor que 15000 α disminuye pero β aumenta.
Sin embargo, se puede proceder de manera inversa, es decir, fijado α encontramos la regla de decisión que corresponderá a una probabilidad de error 1 igual a α.

Ejemplo. Si se toma α = 5%, y se procede a encontrar la regla de decisión correspondiente:

5%= P[Error Tipo I]= P[$\overline{X} \leq$ x_c | $\overline{X}$ ~ N(16500, 1600)] = P[Z < -1.645],

Pero se sabe que, para una distribución normal estándar

$$-1.645 = \frac{x_c - 16500}{400}$$

De donde x_c = 15842 horas. Entonces, la regla de decisión será

"Si $\overline{X}$ fuera inferior a 15842 se dice que el lote es americano, en caso contrario se dice que es europeo".

Con esta regla la probabilidad de error tipo II será

P[Error Tipo II] = P[$\overline{X}$ > 15842 | $\overline{X}$ ~ N(16500, 1600)] = P[Z > 1.425] = 7.93%

Procedimiento general de un test de hipótesis basado en la región de rechazo. Se da ahora una secuencia de pasos que puede ser usada sistemáticamente para cualquier test de hipótesis.

- Iniciar el procedimiento estableciendo, de manera clara y explícita, cuál es la hipótesis nula, es decir, H_0.
- Usar la teoría estadística para construir un indicador de concordancia entre los datos y la hipótesis nula. Este indicador denominado estadístico del test será usado para juzgar la hipótesis H_0.
- Fijar el nivel de significación deseadoα, que es el máximo error aceptable cuando se rechaza H_0, y usar este valor para construir la región crítica.
- Calcular el valor del estadístico a partir de la muestra.
- Si el valor del estadístico pertenece a la región crítica, entonces rechazar H_0. En caso contrario, lo que se puede afirmar es que no hay suficiente evidencia para rechazar H_0.
- Si se dispone de una hipótesis alternativa y de la distribución del estadístico del test bajo la suposición que vale la hipótesis alternativa, se puede calcular la probabilidad de error Tipo II.

Procedimiento general de un test de hipótesis basado en el P-value. Otro procedimiento general de un test de hipótesis más usado en la actualidad debido a la disponibilidad de paquetes de programas estadísticos, consiste en tomar la decisión a partir de la probabilidad del error Tipo I que brindan las salidas de tales paquetes de programas, denominado P-value o simplemente P. Este procedimiento lo podemos resumir en los siguientes pasos:

- Suponer que H_o es cierta.
- Para confrontar esta suposición con la información (parcial) que proveen los datos sobre la realidad de H_o, se forma "una especie de indicador" de concordancia, denominado estadístico del test, el cual es función del de los datos.
- Como el estadístico depende de la información de los datos, con cada muestra posible hay asociado un valor de este estadístico y en consecuencia se genera una nueva variable aleatoria. Asociada a esta variable hay una cierta

distribución de probabilidad, a partir de la cual se determina la probabilidad de que la información de los datos concuerde con la hipótesis nula, denominado P-Value. De esta manera, el P-Value representaría la probabilidad de cometer un error cuando se toma la decisión de rechazar H_o.

- Es claro que si de antemano se fija que la máxima probabilidad de error al rechazar H_o debe ser igual a α, otra manera de tomar la decisión es comparar el valor del P- valué con α. Así
- Si $P \leq \alpha$ entonces la decisión es Rechazamos H_o
- Si $P > \alpha$ la decisión es No hay evidencia suficiente para rechazar H_o

PRUEBAS DE HIPÓTESIS UNILATERALES Y BILATERALES

Las pruebas o test de hipótesis se relacionan con los parámetros poblacionales (medias o proporciones, etc.). Se puede utilizar los estimadores puntuales de los parámetros poblacionales como estadístico del test en cuestión. Supongamos, como ilustración que se utiliza el símbolo θ para denotar el parámetro poblacional de interés, por ejemplo, θ puede serμ, (μ_1- μ_2), p ó (p_1-p_2), y el símbolo $\hat{\theta}$ para denotar el estimador puntual encestado correspondiente.

Desde el punto de vista práctico se puede tener interés en contrastar la hipótesis nula H_0: $\theta = \theta_0$, contra la alternativa de que el parámetro poblacional es mayor que θ_0, o sea H_1: $\theta > \theta_0$. En esta situación, se rechazará H_0 cuando $\hat{\theta}$ sea grande, o sea cuando el estadístico del test sea mayor que un cierto valor llamado valor crítico, que separa las regiones de rechazo y no rechazo del test.

La probabilidad de rechazar la hipótesis nula cuando es cierta será igual al área bajo la curva de la distribución muestral del estadístico del test sobre la región de rechazo. En el caso que estemos trabajando con una distribución normal, y un α = 0,05, se rechaza la hipótesis nula cuando $\hat{\theta}$ se encuentre a más de $1{,}645\,\sigma_{\hat{\theta}}$ a la derecha de θ_0. De esta manera, se puede definir como

Una prueba estadística de una cola o unilateral es aquella en la que la región de rechazo se localiza solamente en una cola o extremo de la distribución muestral del estadístico del test.
Para detectar $\theta > \theta_0$, se sitúa la región de rechazo en la extremidad de valores superiores a $\hat{\theta}$. Para detectar $\theta < \theta_0$ se ubica la región de rechazo en la extremidad izquierda de la distribución de $\hat{\theta}$, o sea para valores inferiores a $\hat{\theta}$. Si hay que detectar diferencias mayores o menores de θ_0, la hipótesis alternativa será
H_1: $\theta \neq \theta_0$
es decir
$\theta > \theta_0$ o bien $\theta < \theta_0$

En este caso la probabilidad de error Tipo I α se repartirá entre las dos colas de la distribución muestral del estadístico, y se rechazará H_0 para valores de $\hat{\theta}$ mayores que un valor crítico (θ_0 + C) o menor que (θ_0- C). Esta prueba se llama prueba estadística bilateral o de dos colas.

CONTRASTES DE HIPÓTESIS

Pueden presentarse en la práctica, situaciones en las que exista una teoría preconcebida relativa a la característica de la población sometida a estudio. Tal sería el caso, por ejemplo si pensamos que un tratamiento nuevo puede tener un porcentaje de mejoría mayor que otro estándar, o cuando nos planteamos si los niños de las distintas comunidades españolas tienen la misma altura. Este tipo de circunstancias son las que nos llevan al estudio de la parcela de la Estadística Inferencial que se recoge bajo el título genérico de Contraste de Hipótesis. Implica, en cualquier investigación, la existencia de dos teorías o hipótesis implícitas, que denominaremos hipótesis nula e hipótesis alternativa, que de alguna manera reflejarán esa idea a priori que tenemos y que pretendemos contrastar con la realidad.
De la misma manera aparecen, implícitamente, diferentes tipos de errores que podemos cometer durante el procedimiento. No podemos olvidar que, habitualmente, el estudio y las conclusiones que obtengamos para una población cualquiera, se habrán apoyado exclusivamente en el análisis de sólo una parte de ésta. De la probabilidad con la que estemos dispuestos a asumir estos errores, dependerá, por ejemplo, el tamaño de la muestra requerida. Desarrollamos en este capítulo los contrastes de hipótesis para los parámetros más usuales que venimos estudiando en los capítulos anteriores: medias, varianzas y proporciones, para una o dos poblaciones. Los contrastes desarrollados en este capítulo se apoyan en que los datos de partida siguen una distribución normal.

Los contrastes de significación se realizan:

- suponiendo a priori que la ley de distribución de la población es conocida.
- Se extrae una muestra aleatoria de dicha población.
- Si la distribución de la muestra es diferente de la distribución de probabilidad que hemos asignado a priori a la población, concluimos que probablemente sea errónea la suposición inicial.

Ejemplo, Supongamos que debemos realizar un estudio sobre la altura media de los habitantes de cierto pueblo. Antes de tomar una muestra, lo lógico es hacer la siguiente suposición a priori, (hipótesis que se desea contrastar y que denotamos H_0):
H_0: la altura media no difiere del resto del país

Al obtener una muestra de tamaño n=8, podríamos encontrarnos ante uno de los siguientes casos:

1. Muestra = {1,50; 1,52; 1,48; 1,55; 1,60; 1,49; 1,55; 1,63}

2. Muestra = {1,65; 1,80; 1,73; 1,52; 1,75; 1,65; 1,75; 1,78}

Sistema recreado para rechara la hipótesis nula o inicial H_0

Nombre	H_0	Prueba	$H_1:\neq$	$H_1:>$	$H_1:<$
Media con varianza desconocida	$\mu-\mu_0$	$z=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$	$z\le z_{\alpha/2}$ $z\ge z_{1-\alpha/2}$	$z\ge z_{1-\alpha}$	$z\le z_{\alpha}$
Media para varianza desconocida	$\mu-\mu_0$	$t=\frac{\bar{x}-\mu_0}{s/\sqrt{n}}$	$t\le t_{\alpha/2,n-1}$ $t\ge t_{1-\alpha/2,n-1}$	$t\ge t_{1-\alpha,n-1}$	$t\le t_{\alpha,n-1}$
Dos medias Normales con varianzas conocidas	$\mu_x-\mu_y=\delta_0$	$z=\frac{\bar{x}-\bar{y}-\delta_0}{\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}}$	$z\le z_{\alpha/2}$ $z\ge z_{1-\alpha/2}$	$z\ge z_{1-\alpha}$	$z\le z_{\alpha}$
Dos medias Normales con varianzas desconocidas *	$\mu_x-\mu_y=\delta_0$	$t=\frac{\bar{x}-\bar{y}-\delta_o}{s_p\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}}$	$t\le t_{\alpha/2,n_x+n_y-2}$ $t\ge t_{1-\alpha/2,n_x+n_y-2}$	$t\ge t_{1-\alpha,n_x+n_y-2}$	$t\le t_{\alpha,n_x+n_y-2}$
Observaciones pareadas	$\mu_d=\delta_0$	$t=\frac{\bar{d}-\delta_0}{s_d/\sqrt{n}}$	$t\le t_{\alpha/2,n-1}$ $t\ge t_{1-\alpha/2,n-1}$	$t\ge t_{1-\alpha,n-1}$	$t\le t_{\alpha,n-1}$
Varianza Normal con media desconocida	$\sigma^2=\sigma_0^2$	$\chi^2=\frac{(n-1)s^2}{\sigma_0^2}$	$\chi^2\le\chi^2_{\alpha/2,n-1}$ $\chi^2\ge\chi^2_{1-\alpha/2,n-1}$	$\chi^2\ge\chi^2_{1-\alpha,n-1}$	$\chi^2\le\chi_{\alpha,n-1}$
Dos distribuciones con varianza Normal	$\sigma_x^2=\sigma_y^2$	$f=\frac{s_x^2}{s_y^2}$	$f\le f_{1-\alpha/2,n_x-1,n_y-1}$ $f\ge f^{-1}_{1-\alpha/2,n_y-1,n_x-1}$	$f\ge f_{1-\alpha,n_x-1,n_y-1}$	$f\le f^{-1}_{1-a,n_x-1,n_y-1}$
Probabilidad de p éxitos	$p=p_0$	$z=\sqrt{\frac{pq}{n}}$	$z\le z_{\alpha/2}$ $z\ge z_{1-\alpha/2}$	$z\ge z_{1-\alpha}$	$z\le z_{\alpha}$

* $s_p^2=\frac{\sigma^2(z_x+z_y)^2}{(\mu_x-\mu_y)^2}$

Intuitivamente, en el caso a sería lógico suponer que salvo que la muestra obtenida sobre los habitantes del pueblo sea muy poco representativa, la hipótesis H_0 debe ser rechazada. En el caso b tal vez no podamos afirmar con rotundidad que la hipótesis H_0 sea cierta, sin embargo no podríamos descartarla y la admitimos por una cuestión de simplicidad.

Este ejemplo sirve como introducción de los siguientes conceptos: En un contraste de hipótesis (también denominado test de hipótesis *o* Contraste de significación)

se decide si cierta hipótesis H_0 que denominamos hipótesis nula puede ser rechazada o no a la vista de los datos suministrados por una muestra de la población. Para realizar el contraste es necesario establecer previamente una hipótesis alternativa (H_1) que será admitida cuando H_0 sea rechazada. Normalmente H_1 es la negación de H_0, aunque esto no es necesariamente así.
El procedimiento general consiste en definir un estadístico *T* relacionado con la hipótesis que deseamos contrastar. A éste lo denominamos estadístico del contraste. A continuación suponiendo que H_0 es verdadera se calcula un intervalo de denominado intervalo de aceptación de la hipótesis nula, (T_i,T_s) de manera que al calcular sobre la muestra $T=T_{exp}$ el criterio a seguir sea:

Si $T_{exp} \in (T_i,T_s)$, entonces aceptamos H_o o rechazamos H_1, y si $T_{exp} \notin (T_i,T_s)$, entonces rechazamos H_o o aceptamos H_1

El intervalo de aceptación o más precisamente, de no rechazo de la hipótesis nula, se establece fijando una cantidad suficientemente pequeña denominada nivel de significación, de modo que la probabilidad de que el estadístico del contraste tome un valor fuera del mismo - región crítica- cuando la hipótesis nula es cierta sea inferior o al 100%; Esto se ha de entender como sigue:

Si H_0 es correcta el criterio de rechazo sólo se equivoca con probabilidad, que es la probabilidad de que una muestra dé un valor del estadístico del contraste extraño (fuera del intervalo de aceptación). La decisión de rechazar o no la hipótesis nula están al fin y al cabo basado en la elección de una muestra tomada al azar, y por tanto es posible cometer decisiones erróneas. Los errores que se pueden cometer se clasifican como sigue:

Error de tipo I: Es el error que consiste en rechazar H_0 cuando es cierta. La probabilidad de cometer este error es lo que anteriormente hemos denominado nivel de significación. Es una costumbre establecida el denotarlo siempre con la letra=P (Rechazar H_0/H_0 es cierta)=P (Aceptar H_1/H_0 es cierta)

Error de tipo II: Es el error que consiste en no rechazar H_0 cuando es falsa. La probabilidad de cometer este error la denotamos con la letra =P (Rechazar H_1/H_1 es cierta)=P (Aceptar H_0/H_1 es cierta)

1. Los errores de tipo I y II no están relacionados más que del siguiente modo: Cuando decrece o crece. Por tanto no es posible encontrar test que hagan tan pequeños como queramos ambos errores simultáneamente. De este modo es siempre necesario privilegiar a una de las hipótesis, de manera que no será rechazada, a menos que su falsedad se haga muy evidente. En los contrastes, la hipótesis privilegiada es H_0 que sólo será rechazada cuando la evidencia de su falsedad supere el umbral del 100*(1-Ho) %.
2. Al tomar Ho muy pequeño tendremos que se puede aproximar a uno. Lo ideal a la hora de definir un test es encontrar un compromiso satisfactorio entre 10 y

15 (aunque siempre a favor de H_0). Denominamos potencia de un contraste a la cantidad 1-5, es decir

Potencia=1-□=P (Rechazar H_0/H_0 es falsa)

	Aceptar Ho	Rechazar Ho
Ho es Cierta	Correcto Probabilidad 1- Ho	Error tipo I Probabilidad Ho
Ho es Falsa	Error Tipo II Probabilidad Ho	Correcto Probabilidad 1- Ho

En el momento de elegir una hipótesis privilegiada podemos en principio dudar entre si elegir una dada o bien su contraria. Criterios a tener en cuenta en estos casos son los siguientes:

Simplicidad científica: A la hora de elegir entre dos hipótesis científicamente razonables, tomaremos como H_0 aquella que sea más simple.

Las consecuencias de equivocarnos: Por ejemplo al juzgar el efecto que puede causar cierto tratamiento médico que está en fase de experimentación, en principio se ha de tomar como hipótesis nula aquella cuyas consecuencias por no rechazarla siendo falsa son menos graves, y como hipótesis alternativa aquella en la que el aceptarla siendo falsa trae peores consecuencias.

Volviendo al ejemplo de la estatura de los habitantes de un pueblo, un estadístico de contraste adecuado es $\overline{X}$. Si la hipótesis H_0 fuese cierta se tendría que X~N (Ho ^{2}/n) (suponiendo claro está que la distribución de las alturas de los españoles siga una distribución normal de parámetros conocidos, por ejemplo N(1.74,100)

Denotemos mediante Ho el verdadero valor de la media en el pueblo que estudiamos. Como la varianza de $\overline{X}$ esa pequeña para grandes valores de n, lo lógico es pensar que si el valor obtenido con la muestra $\overline{X} = \overline{x}$ está muy alejado de Ho =1.74 (región crítica), entonces

- o bien la muestra es muy extraña si H_0 es cierta (probabilidad Ho);
- o bien la hipótesis H_0 no es cierta.

Concretamente en el caso a, donde la muestra es (1.50, 1.52, 1.48, 1.55, 1.60, 1.49, 1.55, 1.63)

El contraste de hipótesis conveniente es:

H_0: Ho = 1 H_1: Ho > 0

En este caso H_1 no es estrictamente la negación de H_0. Esto dará lugar a un contraste unilateral, que son aquellos en los que la región crítica está formada por un sólo intervalo: Intervalo re rechazo de H0: (T_i, ∞). Región crítica: (∞, Ti) Contrastes paramétricos en una población normal. Supongamos que la característica X que estudiamos sobre la población sigue una distribución normal y tomamos una muestra de tamaño n: X_1,.., Xn mediante muestreo aleatorio simple. Vamos a ver cuáles son las técnicas para contrastar hipótesis sobre los parámetros que rigen X. Vamos a comenzar haciendo diferentes tipos de contrastes para medias y después sobre las varianzas y desviaciones típicas.

CONTRASTES PARA LA MEDIA

Test de dos colas con varianza conocida. Suponemos que X~N(Ho) donde es conocido y queremos contrastar si es posible que (desconocida) sea en realidad cierto valor Ho fijado. Esto es un supuesto teórico que nunca se dará en la realidad pero servirá para introducir la teoría sobre contrastes. El test se escribe entonces como:
H_0: Ho =0 H_1: Ho ≠1

Como hemos mencionado anteriormente, la técnica para hacer el contraste consiste en suponer que H_0 es cierta, y averiguar con esta hipótesis quien es la distribución del estadístico del contraste que este caso es lógico que deba estar muy relacionado con $\overline{X}$. Si al obtener una muestra concreta se tiene que $\overline{X} = \overline{x}$ es un valor muy alejado de Ho, se debe rechazar H_0. Veamos esto con más detalle:

H_0 cierta → X~N (Ho, 1) entonces, $Z_{exp} = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \approx N(0,1)$

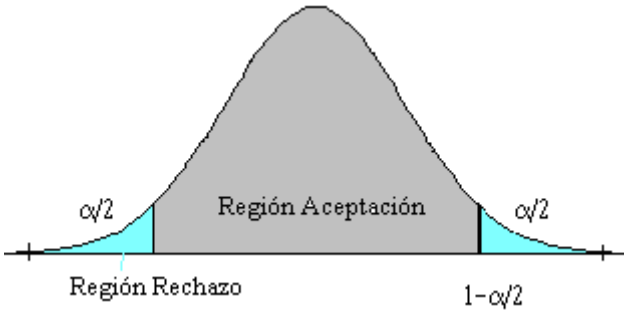
Para poder acceder a las probabilidades de la normal, hemos tipificado (ya que los valores para hacer la tipificación son conocidos). Si H_0 es cierta, entonces esperamos que el valor z_{exp} obtenido sobre la muestra esté cercano a cero con una gran probabilidad. Esto se expresa fijando un nivel de significación y tomando como región crítica C, a los valores que son muy extremados y con probabilidad en total, o sea,

$P(Z_{exo} \le z_{\alpha/2}) = \alpha/2$ y $P(Z_{exo} \ge z_{1-\alpha/2}) = \alpha/2 \Rightarrow P(-z_{1-\alpha/2} \le Z \le z_{1-\alpha/2}) = 1-\alpha$

Entonces la región crítica consiste en $C = \left\{ z_{exp} : \left| z_{exp} \right| \le z_{1-\alpha/2} \right\}$

Luego rechazaremos la hipótesis nula si $\left| z_{exp} \right| > z_{1-\alpha/2}$, aceptando en consecuencia la hipótesis alternativa.

La región de rechazo de la hipótesis nula es la sombreada. Se rechaza H_0 cuando el estadístico z_{exp} toma un valor comprendido en la zona sombreada de la gráfica pequeña, N (0,1), o equivalentemente, cuando el estadístico $\overline{X}$ toma un valor en la zona sombreada de la gráfica grande.



Test de una cola con varianza conocida. Consideremos un contraste de hipótesis donde ahora la hipótesis alternativa es compuesta:
H_0: Ho =0 H_1: Ho <1

Bajo la hipótesis nula la distribución de la media muestral es

H_0 cierta → X~N (Ho,1) entonces, $Z_{exp} = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \approx N(0,1)$

Y como región crítica consideraremos aquella formada por los valores extremadamente bajos de Z_{exp}, con probabilidad Ho, es decir
P ($Z_{exp} \leq z_a$)= Ho, entonces, P ($z_a \leq z_{exp}$)=1-0

Entonces la región de aceptación, o de modo más correcto, de no rechazo de la hipótesis nula es: $z_{exp} > z_a$ Se rechaza la hipótesis nula, cuando uno de los estadístico Z o $\overline{X}$ toma un valor en la zona sombreada (similar a la gráfica anteriormente mostrada).

Al no conocer T va a ser necesario estimarlo a partir de su estimador indeseado: la cuasi varianza muestral, $\hat{s}^2$, Por ello la distribución del estimador del contraste será una t-Student, que ha perdido un grado de libertad, según el teorema de Cochran, y la definición de la distribución de t-Student:

H_0 cierta → $T_{exp} = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} \approx t_{n-1}$

Consideramos como región crítica C, a las observaciones de T_{exp} extremas
P ($T_{exp} \leq t_{\square/2,n-1}$)=T, y P($T_{exp} \geq t_{1-\square/2,n-1}$)=T entonces, P($-t_{1T/2,n-1} \leq zT_{exp} \leq t_{1-\square/2,n-1}$)=1-T

Entonces la región crítica consiste en $C = \{T_{exp} < -t_{1-\alpha/2,n-1} \quad \text{ó} \quad T_{1-\alpha/2,n-1} < T_{exp}\}$
Para dar una forma homogénea a todos los contrastes de hipótesis es costumbre denominar al valor del estadístico del contraste calculado sobre la muestra como valor experimental y a los extremos de la región crítica, como valores teóricos. Definiendo entonces

$$T_{exp} = \frac{\overline{X} - \mu_0}{\hat{s}/\sqrt{n}} \qquad T_{teo} = t_{1-\alpha/2,n-1}$$

El resultado del contraste es el siguiente: Si $|T_{exp} \leq T_{teo}|$ no rechazamos H_0, de contrario sí.

Definimos T_{exp} y T_{teo} como anteriormente y el criterio a aplicar es:
Si $T_{exp} \leq T_{teo}$ no rechazamos H_0, de contrario sí.

Ejemplo. Conocemos que las alturas *X* de los individuos de una ciudad, se distribuyen de modo gaussiano. Deseamos contrastar con un nivel de significación de X=5% si la altura media es diferente de 174 cm. Para ello nos basamos en un estudio en el que con una muestra de n=25 personas se obtuvo: media 170 y desviación 10

Solución: El contraste que se plantea es:
H_0:X=1 H_1:□≠174

La técnica a utilizar consiste en suponer que H_0 es cierta y ver si el valor que toma el estadístico

$$T_{exp} = \frac{\overline{X} - 174}{\hat{s}/\sqrt{n}} \approx t_{24}$$

Es razonable o no bajo esta hipótesis, para el nivel de significación dado. Aceptaremos la hipótesis alternativa (y en consecuencia se rechazará la hipótesis nula) si no lo es, es decir, si
$|T_{exp}| \geq t_{1-\alpha/2,24} = t_{0.975,24} = 2.06$

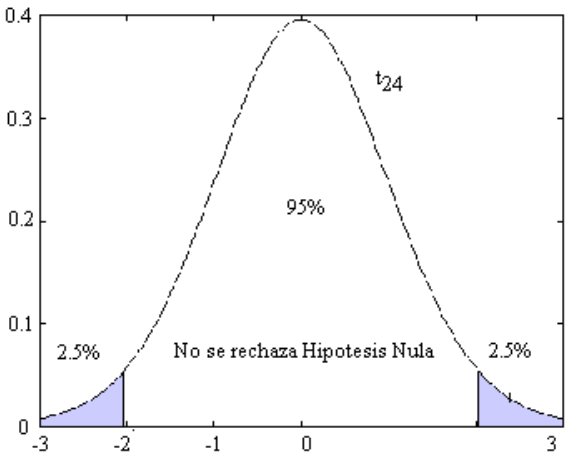
Para ello procedemos al cálculo de T_{exp} con s=10, y n=25

$$\hat{s} = s\sqrt{\frac{n}{n-1}} = 10\sqrt{\frac{25}{24}} = 10.206 \qquad \text{entonces}$$

$$|T_{exp}| = \frac{|170 - 174|}{10.206/\sqrt{25}} = |1.959| \leq 2.06$$

Luego, aunque podamos pensar que ciertamente el verdadero valor de □ no es 174, no hay una evidencia suficiente para rechazar esta hipótesis al nivel de confianza del 95%. Es decir, no se rechaza H_0.

El valor de T_{exp} no está en la región crítica (aunque ha quedado muy cerca), por tanto al no ser la evidencia en contra de H_0 suficientemente significativa, ésta hipótesis no se rechaza.



CONTRASTES PARA LA VARIANZA

Consideremos que el carácter que estudiamos sobre la población sea una variable aleatoria normal cuya media y varianza son desconocidas. Vamos a contrastar la hipótesis

H_0: $X^2=X_0^2$,

Donde X_0^2 es un valor prefijado frente a otras hipótesis alternativas que podrán dar lugar a contrastes bilaterales o unilaterales. La técnica consiste en utilizar el teorema de Cochran, para observar que el siguiente estadístico experimental que utiliza el estimador encestado de la varianza, posee una distribución χ^2, con n-1 grados de libertad:

$$H_0\text{: cierta} \rightarrow \chi^2_{exp} = (n-1)\cdot\frac{\hat{s}^2}{\sigma_0^2} \approx \chi^2_{n-1}$$

Entonces construimos las regiones críticas que correspondan a las hipótesis alternativas que se formulen en cada caso atendiendo a la ley de distribución χ^2.

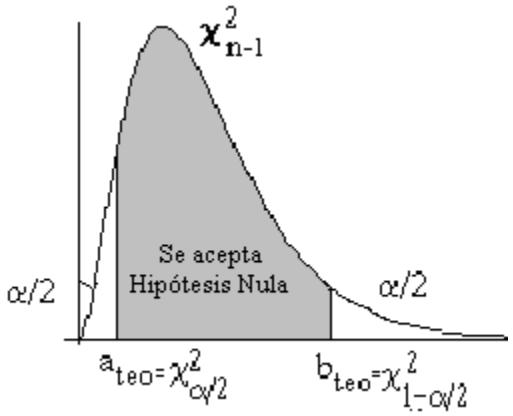
Contraste bilateral. Cuando el contraste a realizar es

$$H_0 : \sigma^2 = \sigma_0^2 \qquad H_1 : \sigma^2 \neq \sigma_0^2$$

Entonces, definimos

$$\chi^2_{exp} = (n-1)\cdot\frac{\hat{s}^2}{\sigma_0^2} \qquad a_{teo} = \chi^2_{\alpha/2,n-1} \qquad b_{teo} = \chi^2_{1-\alpha/2,n-1}$$

Y el criterio que suministra el contraste es el expresado en la figura:



Si $a_{teo} \le \chi^2_{exp} \le b_{teo}$ aceptamos a H_0, de contrario lo rechazamos

Contrastes unilaterales. Para un contraste de significación al nivel Ho del tipo
$H_0 : \sigma^2 = \sigma_0^2 \qquad H_1 : \sigma^2 < \sigma_0^2$

Entonces, $a_{teo} = \chi^2_{\alpha,n-1}$

Si $a_{teo} \le \chi^2_{exp}$ aceptamos a H_0, de contrario lo rechazamos

Para el contraste contrario tenemos la formulación análoga:
$H_0 : \sigma^2 = \sigma_0^2 \qquad H_1 : \sigma^2 > \sigma_0^2$

Entonces, $b_{teo} = \chi^2_{1-\alpha,n-1}$

Si $a_{teo} > \chi^2_{exp}$ aceptamos a H_0, de contrario lo rechazamos

CONTRASTES DE UNA PROPORCIÓN

Supongamos que poseemos una sucesión de observaciones independientes, de modo que cada una de ellas se comporta como una distribución de Bernoulli de parámetro p: $X_1,\ldots X_n$~Binomial de parámetro p.

La variable aleatoria $X=X_1+X_2+\ldots+X_n \sim B(n,p)$. La proporción muestral (estimador del verdadero parámetro *p* a partir de la muestra) es $\hat{p} = X/n$

Nos interesamos en el contraste de significación de H_0: $p=p_0$, siendo p un valor prefijado frente a otras hipótesis alternativas. Para ello nos basamos en un estadístico (de contraste) que ya fue considerado anteriormente en la construcción

de intervalos de confianza para proporciones y que sigue una distribución aproximadamente normal para tamaños muéstrales suficientemente grandes:

$$\hat{P} = \frac{X}{n} \approx N\left(p, \frac{pq}{n}\right)$$

Si la hipótesis H_0 es cierta se tiene

$$\hat{P} = \frac{X}{n} \approx N\left(p_0, \frac{p_0 q_0}{n}\right) \Leftrightarrow \frac{\hat{P} - p_0}{\sqrt{p_0 q_0 / n}} = Z_{exp} \approx N(0,1)$$

Contraste bilateral. Para el contraste

$H_0 : p = p_0 \qquad H_1 : p \neq p_0$

Extraemos una muestra y observamos el valor X=x, entonces $\hat{p} = x/n$. Entonces se define

$$Z_{exp} = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}} \qquad Z_{teo} = z_{1-\alpha/2}$$

Siendo el criterio de aceptación o rechazo de la hipótesis nula
Si $|z_{exp}| \leq Z_{teo}$ aceptamos a H_0, de contrario lo rechazamos

Contrastes unilaterales. Consideremos un contraste del tipo

$H_0 : p = p_0 \qquad H_1 : p < p_0$

Definiendo a

$$Z_{exp} = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}} \qquad Z_{teo} = z_{\alpha}$$

Siendo el criterio de aceptación o rechazo de la hipótesis nula
Si $z_{exp} \leq Z_{teo}$ rechazamos a H_0, de contrario lo aceptamos

Para el test unilateral contrario, se tiene la expresión simétrica:

$H_0 : p = p_0 \qquad H_1 : p > p_0$

Definiendo a

$$Z_{exp} = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}} \qquad Z_{teo} = z_{1-\alpha}$$

Siendo el criterio de aceptación o rechazo de la hipótesis nula
Si $z_{exp} \leq Z_{teo}$ aceptamos a H_0, de contrario lo rechazamos

CONTRASTES PARA LA DIFERENCIA DE MEDIAS APAREADAS

Las muestras apareadas aparecen como distintas observaciones realizadas sobre los mismos individuos. Un ejemplo de observaciones apareadas: Medir a un conjunto de n personas el nivel de insulina en la sangre antes (X) y después (Y) del tratamiento

Paciente	x_i	y_i	Diferencia d_i
1	150	120	30
…	…	…	…
N	140	90	50

No es posible considerar a X y Y como variables independientes ya que va a existir una dependencia clara entre las dos variables. Si queremos contrastar el que los pacientes han experimentado o no una mejoría con el tratamiento, llamemos d_i a la diferencia entre las observaciones antes y después del tratamiento $d_i=x_i-y_i$. Supongamos que la variable aleatoria que define la diferencia entre el antes y después del tratamiento es una variable aleatoria *d* que se distribuye normalmente, pero cuyas media y varianza son desconocidas
$d \sim N(D_d, D_d)$
Si queremos contrastar la hipótesis de que el tratamiento ha producido cierto efecto
H_0: $d_d=1$

En el caso en que H_0 fuese cierta tendríamos que el estadístico de contraste que nos conviene es

$$T_{exp} = \frac{\bar{d} - \Delta}{\hat{s}_d / \sqrt{n}} \approx t_{n-1}$$

Donde $\bar{d}$ es la media muestral de las diferencias d_i y $\hat{s}_d$ es la cuasi varianza muestral de las mismas. El tipo de contraste sería entonces del mismo tipo que el realizado para la media con varianza desconocida.

Contraste bilateral. Consideramos el contraste de tipo
H_0: $d_d=1$ $\quad d_d \neq 0$

Entonces se define

$$T_{exp} = \frac{\bar{d} - \Delta}{\hat{s}_d / \sqrt{n}} \approx t_{n-1}$$

Y se rechaza la hipótesis nula cuando $T_{exp} < -t_{1-\alpha/2,n-1}$ ó $T_{exp} > t_{1-\alpha/2,n-1}$

Contrastes unilaterales. Si el contraste es
H_0: μ=1 μ<10

Se rechaza la hipótesis nula cuando $T_{exp} < -t_{1-\alpha,n-1}$. Para el test contrario
H_0: μ=0 μ>0

Se rechaza la hipótesis nula cuando $T_{exp} > t_{1-\alpha,n-1}$

No supone ninguna dificultad el haber realizado el contraste con d_d^2 conocida, ya que entonces el estadístico del contraste es

$Z = \frac{\bar{d} - \Delta}{\hat{s}_d / \sqrt{n}} \approx N(0,1)$ Y el tratamiento sería análogo.

CONTRASTES DE DOS DISTRIBUCIONES NORMALES INDEPENDIENTES

Consideramos a lo largo de toda esta sección a dos poblaciones normales que representamos mediante $X_1 \sim N(Z_1,Z_1)$ y $X_2 \sim N(Z_2,Z_2)$

De las que de modo independiente se extraen muestras de tamaño respectivo n_1 y n_2. Los test que vamos a realizar están relacionados con las diferencias existentes entre ambas medias o los cocientes de sus varianzas.

CONTRASTE DE MEDIAS CON VARIANZAS CONOCIDAS

De manera similar al caso del contraste para una media, queremos en esta ocasión contrastar la hipótesis de que las dos poblaciones (cuyas varianzas suponemos conocidas) sólo difieren en una cantidad
H_0: Z_1-Z_2=1

Frente a hipótesis alternativas que darán lugar a contrastes unilaterales o bilaterales como veremos más tarde. Para ello nos basamos en la distribución del siguiente estadístico de contraste:

H_0 es cierta $\rightarrow \bar{X}_1 \approx N(\mu_1, \sigma_1/\sqrt{n_1})$ y $\bar{X}_2 \approx N(\mu_2, \sigma_2/\sqrt{n_2})$

$\bar{X}_1 - \bar{X}_2 \approx N(\Delta, s_1/\sqrt{n_1}, s_2/\sqrt{n_2})$ entonces,

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - \Delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \approx N(0,1)$$

Define-se entones

$$Z_{exp} = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - \Delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \qquad Z_{teo} = z_{1-\alpha/2}$$

y el test consiste en

$\left|Z_{exp}\right| \le Z_{teo} \rightarrow$ Aceptamos H_0 y rechazamos H_1

y el test consiste en

$Z_{teo} = z_{\alpha} = -z_{1-\alpha}$ entonces, si $Z_{exp} \ge Z_{teo} \rightarrow$ Aceptamos H_0 y rechazamos H_1

y el test consiste en

$Z_{teo} = z_{1-\alpha}$ entonces, si $Z_{exp} \le Z_{teo} \rightarrow$ Aceptamos H_0 y rechazamos H_1

CONTRASTE DE MEDIAS HOMOCEDÁTICAS

Cuando sólo conocemos que las varianzas de ambas poblaciones son iguales, pero desconocidas. El estadístico que usaremos para el contraste fue ya introducido en la relación, pues si suponemos que H_0 es cierta se tiene

$$T_{exp} = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\hat{s}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \approx t_{n_1+n_2-2}$$

Donde $\hat{s}^2$ es la cuasi varianza muestral ponderada de $\hat{s}_1^2$ y de $\hat{s}_2^2$ donde

$$\hat{s}^2 = \frac{(n_1 - 1)\hat{s}_1^2 + (n_2 - 1)\hat{s}_2^2}{n_1 + n_2 - 2}$$

Obsérvese que se han perdido dos grados de libertad a causa de la estimación de $\sigma_1^2 = \sigma_2^2$ mediante $\hat{s}_1^2$ y de $\hat{s}_2^2$.

se tiene como en casos anteriores que el contraste adecuado consiste en definir

$$T_{exp} = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\hat{s}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad T_{teo} = t_{1-\alpha/2,n_1+n_2-2}$$

y rechazar o admitir la hipótesis nula siguiendo el criterio

$\left|T_{exp}\right| \le T_{teo} \rightarrow$ Aceptar H_0

y rechazar o admitir la hipótesis nula siguiendo el criterio

$T_{teo} = -t_{1-\alpha/2, n_1+n_2-2} \Rightarrow T_{exp} \geq T_{teo} \rightarrow$ Aceptar H_0

y rechazar o admitir la hipótesis nula siguiendo el criterio

$T_{teo} = t_{1-\alpha, n_1+n_2-2} \Rightarrow T_{exp} \leq T_{teo} \rightarrow$ Aceptor H_0

CONTRASTE DE MEDIAS NO HOMOCEDÁTICAS

En el caso más problemático, es decir cuando sólo conocemos de las dos poblaciones que su distribución es normal, y que sus varianzas no son conocidas y significativamente diferentes. En este caso el estadístico de contraste tendrá una ley de distribución muy particular. Consistirá en una distribución t-Student, con un número de grados de libertad que en lugar de depender de modo determinista de la muestra (a través de su tamaño), depende de un modo aleatorio mediante las varianzas muéstrales. Concretamente, el estadístico que nos interesa es

$$T = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - \Delta}{\sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}} \approx t_\nu$$

Donde V es el número de grados de libertad que se calcula mediante la fórmula de Welch:

$$\nu = \frac{\left(\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}\right)}{\frac{1}{n_1+1}\left(\frac{\hat{s}_1^2}{n_1}\right)^2 + \frac{1}{n_2+1}\left(\frac{\hat{s}_2^2}{n_2}\right)^2} - 2$$

No desarrollamos en detalle los cálculos a realizar, pues la técnica para efectuar los contrastes es análoga a los vistos anteriormente cuando las varianzas son desconocidas e iguales. Si lo que pretendemos contrastar es si las medias poblacionales de dos muestras independientes obtenidas de poblaciones normales son idénticas, esto se reduce a los casos anteriores tomando R=0

CONTRASTES DE LA RAZÓN DE VARIANZAS

Consideramos dos muestras independientes de dos poblaciones que se distribuyen normalmente (cuyas medias y varianzas son desconocidas). Vamos a abordar cuestiones relacionadas con saber si las varianzas de ambas poblaciones son las mismas, o si la razón (cociente) entre ambas es una cantidad conocida, R. La igualdad entre las dos varianzas puede escribirse R_1^2-R_2^2=0 o bien, la existencia de una diferencia entre ambas (R), del modo R_1^2-R_2^2=1. Este modo de escribir la diferencia entre varianzas (que era el adecuado para las medias) no es sin embargo fácil de utilizar para las varianzas, de modo que nos será más fácil sacarle partido a las expresiones de las relaciones entre varianzas como

$$\frac{\sigma_1^2}{\sigma_2^2} = R$$

Por ejemplo, si R=1 tenemos que ambas varianzas son iguales. Consideramos entonces la hipótesis nula

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = R$$

la cual vamos a contrastar teniendo en cuenta que:

$\frac{(n_1 - 1)\hat{s}_1^2}{\sigma_1^2} \approx \chi^2_{n_1 - 1}$ $\frac{(n_2 - 1)\hat{s}_2^2}{\sigma_2^2} \approx \chi^2_{n_2 - 1}$ Que conlleva

$$\frac{\frac{1}{n_1 - 1}\frac{(n_1 - 1)\hat{s}_1^2}{\sigma_1^2}}{\frac{1}{n_2 - 1}\frac{(n_2 - 1)\hat{s}_2^2}{\sigma_2^2}} = \frac{\hat{s}_1^2 / \sigma_1^2}{\hat{s}_2^2 / \sigma_2^2} = \frac{\sigma_2^2 \hat{s}_1^2}{\sigma_1^2 \hat{s}_2^2} \approx F_{n_1 - 1, n_2 - 1}$$

Por tanto el estadístico del contraste que nos conviene tiene una distribución conocida cuando H_0 es cierta. Véase la definición de la distribución de F-Snedecor:

$$F = \frac{1}{R}\frac{\hat{s}_1^2}{\hat{s}_2^2} \approx F_{n_1 - 1, n_2 - 1}$$

Contraste bilateral. El contraste bilateral para el cociente de varianzas se escribe como:

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = R \qquad H_0 : \frac{\sigma_1^2}{\sigma_2^2} \neq R$$

Habida cuenta que la distribución F-Snedecor no es simétrica sino que sólo toma valores positivos, se rechazará la hipótesis nula cuando el valor que tome el estadístico del contraste al aplicarlo sobre una muestra sea muy cercano a cero, o bien, muy grande. Es decir, se define el estadístico experimental y los límites de la región crítica como:

$$F_{exp} = \frac{1}{R}\frac{\hat{s}_1^2}{\hat{s}_2^2} \qquad a_{teo} = F_{\alpha/2, n_1 - 1, n_2 - 1} \qquad b_{teo} = F_{1 - \alpha/2, n_1 - 1, n_2 - 1}$$

Y el criterio de aceptación o rechazo es:

si $a_{teo} \le F_{exp} \le b_{teo} \rightarrow$ Aceptamos a H_0

No se debe olvidar que para la función F-Snedecor, $F_{\alpha/2,n_1-1,n_2-1} \neq -F_{1-\alpha/2,n_1-1,n_2-1}$ dada la no simetría de F. A la hora de usar una tabla de la distribución podemos tal vez encontrar que no está tabulada para los valores pequeños, pero si para 1-□. Una regla que es de bastante utilidad para estos casos es la siguiente (ojo, se invierten los órdenes de los grados de libertad),

$$F_{\alpha,n,m} = \frac{1}{F_{1-\alpha,m,n}}$$

Contrastes unilaterales. El primer contraste unilateral que consideramos es:

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = R \qquad H_0 : \frac{\sigma_1^2}{\sigma_2^2} < R$$

Para el cual se tiene $a_{teo} = F_{\alpha,n_1-1,n_2-1}$, si $a_{teo} \le F_{exp}$ aceptamos a H_0

El test unilateral opuesto es:

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = R \qquad H_0 : \frac{\sigma_1^2}{\sigma_2^2} > R$$

Para el cual se tiene $b_{teo} = F_{1-\alpha,n_1-1,n_2-1}$, si $F_{exp} \le b_{teo}$ aceptamos a H_0

Caso particular: Contraste de homocedasticidad. En la práctica un contraste de gran interés es el de la homocedasticidad o igualdad de varianzas. Decimos que dos poblaciones son homocedáticas si tienen la misma varianza. El test de homocedasticidad sería entonces el mismo que el de un cociente de varianzas, donde R=1, es decir:

$$\sigma_1^2 = \sigma_2^2 \Rightarrow H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1 \qquad H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

Una de las razones de la importancia de este contraste es la siguiente: Si queremos estudiar la diferencia entre las medias de dos poblaciones normales, el caso más realista es considerar un contraste donde las varianzas de las poblaciones son desconocidas. Ante esta situación podemos encontrarnos dos situaciones:

1. Las dos varianzas son iguales. Este es el caso más favorable pues utilizamos la distribución de Student para el contraste con un número de grados de libertad que sólo depende del tamaño de la muestra.
2. Las varianzas son distintas. En este caso el número de grados de libertad es una variable aleatoria (fórmula de Welch) y por tanto al realizar el contraste se pierde cierta precisión.

En esta situación lo recomendable es

- En primer lugar realizar un test de homocedasticidad.
- Si la igualdad de varianzas no puede ser rechazada de modo significativo, aplicamos un test de diferencia de medias suponiendo que las varianzas son desconocidas pero iguales.

  En otro caso se utiliza la aproximación de Welch.

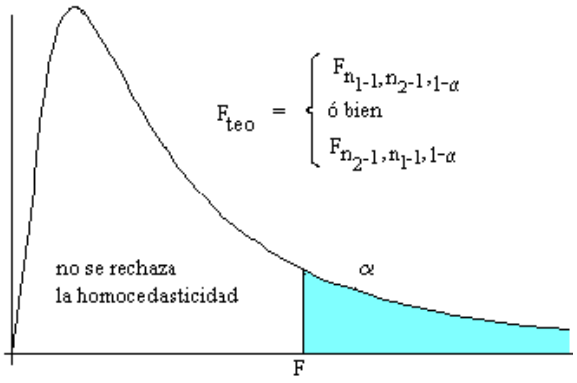
Al realizar el contraste bilateral sobre la igualdad de varianzas podemos también economizar parte de trabajo definiendo F_{exp} como el cociente entre la mayor varianza muestral y la menor

$$F_{exp} = \begin{cases} \dfrac{\hat{s}_1^2}{\hat{s}_2^2} \rightarrow \hat{s}_1^2 \geq \hat{s}_2^2 \\ \\ \dfrac{\hat{s}_2^2}{\hat{s}_1^2} \rightarrow \hat{s}_2^2 > \hat{s}_2^2 \end{cases} \Rightarrow F_{exp} \geq 1$$

Ya que así no es necesario calcular el extremo inferior para la región donde no se rechaza H_0, pues F_{exp} nunca estará próxima a 0. Con esta definición de F_{exp} el criterio a seguir frente al contraste de significación para un valor F dado es,

Criterio para el rechazo de la hipótesis nula sobre la homocedasticidad. Aunque en realidad el test a realizar es bilateral, al elegir el estadístico del contraste de modo que el numerador sea mayor que el numerador, podemos concentrar toda la probabilidad del error de tipo I,F, en la cola derecha de la distribución.

$$F_{teo} = \begin{cases} F_{1-\alpha,n_1-1,n_2-1} \rightarrow \hat{s}_1^2 \geq \hat{s}_2^2 \\ F_{1-\alpha,n_2-1,n_1-1} \rightarrow \hat{s}_2^2 > \hat{s}_1^2 \end{cases} \Rightarrow \begin{cases} F_{exp} \leq b_{teo} \quad \text{aprobar} \quad H_0 \\ F_{exp} > b_{teo} \quad \text{rechazar} \quad H_0 \end{cases}$$



Ejemplo. Se desea comparar la actividad motora espontánea de un grupo de 25 ratas control y otro de 36 ratas desnutridas. Se midió el número de veces que pasaban delante de una célula fotoeléctrica durante 24 horas. Los datos obtenidos fueron los siguientes:

Ratas de control	$n_1=25$	$\bar{x}_1=869.8$	$S_1=106.7$
Ratas Desnutridas	$n_2=36$	$\bar{x}_2=465$	$S_2=153.7$

¿Se observan diferencias significativas entre el grupo control y el grupo desnutrido?

Solución: En primer lugar, por tratarse de un problema de inferencia estadística, nos serán más útiles las cuasi varianzas que las varianzas. Por ello calculamos:

$$\hat{s}_1^2=\frac{n_1}{n_1-1}s_1^2=\frac{25}{24}(106.7)^2=11.859 \qquad \hat{s}_2^2=\frac{n_2}{n_2-1}s_2^2=\frac{36}{35}(153.7)^2=24.298$$

El contraste que debemos realizar está basado en el de la t-Student para la diferencia de medias de dos poblaciones. Para ello conocemos dos estadísticos posibles, según que las varianzas poblacionales de ambos grupos de ratas puedan ser supuestas iguales (homocedasticidad) o distintas (heterocedasticidad). Para ello realizamos previamente el contraste:

$$H_0:\frac{\sigma_1^2}{\sigma_2^2}=1 \qquad H_1:\frac{\sigma_1^2}{\sigma_2^2}\neq 1$$

Suponiendo H_0 cierta, tenemos que el estadístico del contraste conveniente es

$$F_{exp} = \begin{cases} \dfrac{\hat{s}_1^2}{\hat{s}_2^2} \rightarrow \hat{s}_1^2 \geq \hat{s}_2^2 \\ \\ \dfrac{\hat{s}_2^2}{\hat{s}_1^2} \rightarrow \hat{s}_2^2 > \hat{s}_2^2 \end{cases} \Rightarrow F_{exp} \geq 1$$

ya que así no es necesario calcular el extremo inferior para la región donde no se rechaza H_0. En este caso:

$$F_{exp} = \frac{\hat{s}_2^2}{\hat{s}_1^2} = 2.049 \approx F_{n_2-1,n_1-1} \qquad F_{teo} = 2.97$$

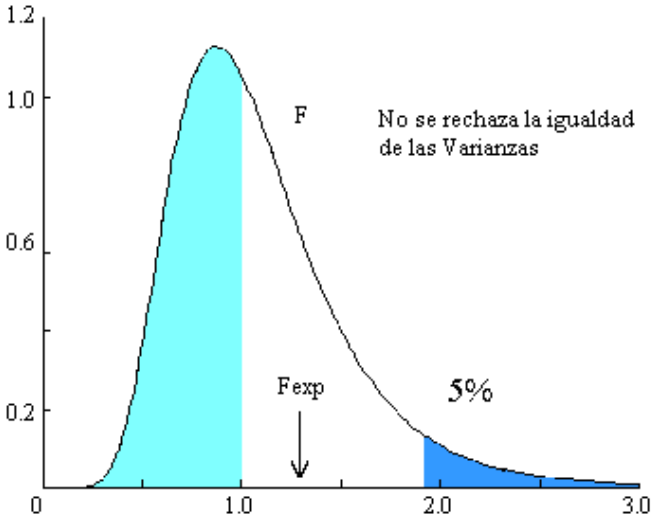
Como $F_{exp} \leq T_{teo}$, no podemos concluir (al menos al nivel de significación α=5%) que H_0 deba ser rechazada.

No hay evidencia significativa para rechazar la homocedasticidad. El estadístico del contraste ha sido elegido modo que el numerador de F_{exp} sea mayor que el denominador, es decir, $F_{exp}>1$.

Utilizando el estadístico más sencillo (el que no necesita aproximar los grados de libertad mediante la fórmula de Welch). Para ello calculamos en primer lugar la cuasi varianza muestral ponderada y los valores del test:

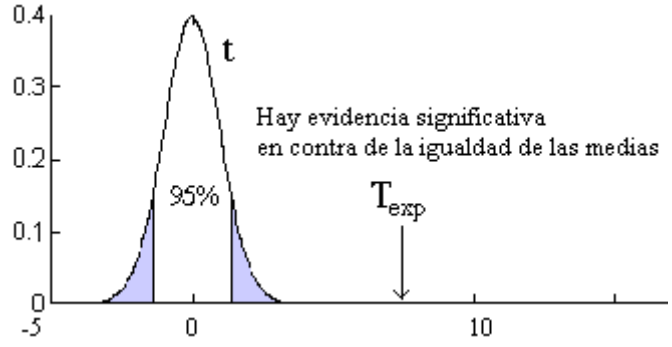
$$\hat{s}^2 = \frac{(n_1-1)\hat{s}_1^2 + (n_2-1)\hat{s}_2^2}{n_1+n_2-2} = 19.238$$

$$T_{exp} = \frac{\bar{x}_1 - \bar{x}_2}{\hat{s}\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} = 11.210 \approx t_{n_1+n_2-2} = t_{59}$$



Como $\left|T_{teo}\right| \leq T_{exp}$ concluimos que se ha de rechazar la hipótesis de igualdad de las medias, y por tanto aceptamos que las medias son diferentes. Además, como se aprecia en la figura, la evidencia a favor de la hipótesis alternativa es muy alta, y se puede afirmar que con gran probabilidad la media poblacional de las ratas de control es mayor que la de las ratas desnutridas.

Hay una gran evidencia en contra de la hipótesis de que ambas medias poblacionales coincidan, y a favor de que la de la primera población es mayor que la de la segunda.



CONTRASTES SOBRE LA DIFERENCIA DE PROPORCIONES

Supongamos que tenemos dos muestras independientes tomadas sobre dos poblaciones, en la que estudiamos una variable de tipo dicotómico (Bernoulli):

$$\vec{X}_1 = X_{11},...,X_{1n_1} \qquad \vec{X}_2 = X_{21},..,X_{2n_2}$$

Si X_1 y X_2 contabilizan en cada caso el número de éxitos en cada muestra se tiene que cada una de ellas se distribuye como una variable aleatoria binomial:

$$X_1 = \sum_{i=1}^{n_1} X_{1i} \approx B(n_1, p_1) \qquad X_2 = \sum_{i=1}^{n_1} X_{2i} \approx B(n_2, p_2)$$

De modo que los estimadores de las proporciones en cada población tienen distribuciones que de un modo aproximado son normales (cuando n_1 y n_2 son bastante grandes)

$$\hat{P}_1 = \frac{X_1}{n_1} \approx N\left(p_1, \frac{p_1 q_1}{n_1}\right) \qquad \hat{P}_2 = \frac{X_2}{n_2} \approx N\left(p_2, \frac{p_2 q_2}{n_2}\right)$$

El contraste que nos interesa realizar es el de si la diferencia entre las proporciones en cada población es una cantidad conocida □
H_0: p_1-p_2=P

Si H_0 fuese cierta se tendría que

$$\hat{P}_1 - \hat{P}_2 \approx N\left(p_1 - p_2, \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}\right)$$

Desafortunadamente ni p_1 ni p_2 son conocidos de antemano y utilizamos sus estimadores, lo que da lugar a un error que es pequeño cuando los tamaños muéstrales son importantes:

$$\frac{(\hat{p}_1 - \hat{p}_2) - \Delta}{\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}} = Z_{exp} \approx N(0,1)$$

Contraste bilateral. El contraste bilateral sobre la diferencia de proporciones es
H_0: p_1-p_2=1 H_1: p_1-p_2≠0

Entonces se define

$$Z_{exp} = \frac{(\hat{p}_1 - \hat{p}_2) - \Delta}{\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}}$$

Y se rechaza la hipótesis nula si $Z_{exp} < -z_{1-\alpha/2}$ o si $Z_{exp} > z_{1-\alpha/2}$

Contrastes unilaterales. En el contraste
H_0: p_1-p_2=0 H_1: p_1-p_2<1

Y se rechaza la hipótesis nula si $Z_{exp} > z_{1-\alpha/2}$

Hipótesis Estadísticas

Conceptos Generales (cont.)

- Una hipótesis estadística es una aseveración o conjetura con respecto a una o más poblaciones.
- La verdad o falsedad de una hipótesis estadística nunca se sabe con absoluta certidumbre, a menos que examinemos toda la población; lo cual es poco práctico en la mayoría de las situaciones.
- Por lo cual, se toma una muestra aleatoria de la población de interés, y se utilizan los datos contenidos en esta muestra para proporcionar evidencia que apoye o no la hipótesis.
- La evidencia de la muestra que sea inconsistente con la hipótesis que se establece conduce al rechazo de la misma.

Hipótesis Estadísticas

Conceptos Generales (cont.)

1. Un procedimiento de decisión debe hacerse con la noción de la probabilidad de una conclusión errónea.
2. El rechazo de una hipótesis simplemente implica que la evidencia de la muestra la refuta.
3. El rechazo significa que hay una pequeña probabilidad de obtener la información muestral observada cuando, de hecho, la hipótesis es verdadera.
4. Como resultado, el analista de los datos establece una conclusión firme cuando se rechaza una hipótesis.
5. Cuando el análisis de datos formaliza la evidencia experimental con base en la prueba de hipótesis, es muy importante la declaración o el establecimiento formal de la hipótesis.

- Un procedimiento de decisión debe hacerse con la noción de la probabilidad de una conclusión errónea.
- El rechazo de una hipótesis simplemente implica que la evidencia de la muestra la refuta.
- El rechazo significa que hay una pequeña probabilidad de obtener la información muestral observada cuando, de hecho, la hipótesis es verdadera.
- Como resultado, el analista de los datos establece una conclusión firme cuando se rechaza una hipótesis.

- Cuando el análisis de datos formaliza la evidencia experimental con base en la prueba de hipótesis, es muy importante la declaración o el establecimiento formal de la hipótesis.

Hipótesis Estadísticas

Conceptos Generales (cont.)

1. La estructura de la prueba de hipótesis se formula usando el término hipótesis nula, el cual se refiere a cualquier hipótesis que se desea probar y se denota con H0.
2. El rechazo H0 de conduce a la aceptación de la hipótesis alternativa, que se denota con H1.
3. La hipótesis alternativa H1 representa la pregunta que se debe responder o la teoría que se debe probar y, por ello, su especificación es muy importante.
4. La hipótesis nula H0 anula o se opone a H1, y a menudo es el complemento lógico para H1.

Prueba de una Hipótesis Estadística

- Una prueba o contraste de una hipótesis estadística es una regla o procedimiento que conduce a una decisión de aceptar o rechazar cierta hipótesis con base en los resultados de una muestra.
- Los procedimientos de prueba de hipótesis dependen del empleo de la información contenida en una muestra aleatoria de la población de interés.
- Si esta información es consistente con la hipótesis se concluye que es verdadera.
- En caso contrario, la información es inconsistente con la hipótesis, se concluye que es falsa.

- Estadístico de prueba. Valor obtenido a partir de la información muestral, se utiliza para determinar si se rechaza o no la hipótesis.

- Toda prueba de hipótesis lleva implícita un estadístico para probarla.

- Este estadístico depende del problema planteado y de la codificación de las variables.

- Región crítica. Región de rechazo de la hipótesis nula H0.

- Se denomina C (subconjunto del espacio muestral) a la región crítica de una prueba dada si nos lleva a rechazar la hipótesis nula H0 cuando la muestra cae en la región C.

- Valor crítico. El punto que divide la región de aceptación y la región de rechazo de la hipótesis nula H0.

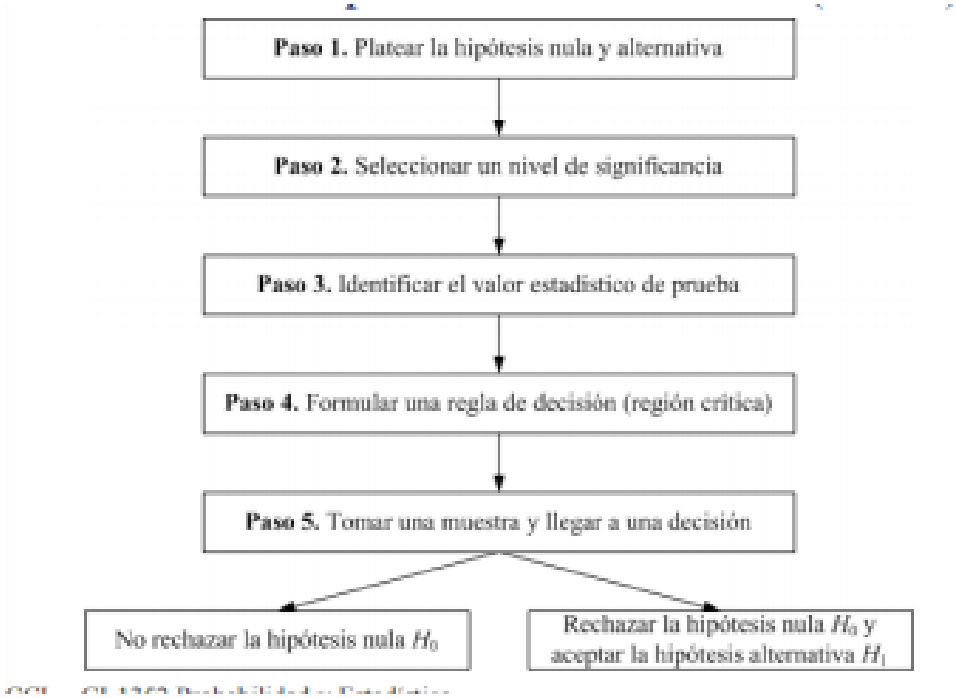
    - Nivel de significancia. Probabilidad de rechazar la hipótesis nula cuando es verdadera. Se denota con α.
    - Error Tipo I. Rechazar la hipótesis nula cuando en realidad es verdadera. La probabilidad de cometer un error tipo I se le llama nivel de significancia.
    - Error Tipo II. Aceptar la hipótesis nula cuando en realidad es falsa.
    - La probabilidad de cometer el error tipo II es imposible de calcular a menos que tengamos una hipótesis alternativa específica. Se denota con β.

        - Al probar cualquier hipótesis estadística, hay cuatro situaciones posibles que determinan si nuestra decisión es correcta o errónea (ver la siguiente tabla).

		Situaciones Posibles	
		H_0 verdadera	**H_0 falsa**
Decisiones Posibles	**No rechazar H_0**	Decisión correcta	Error tipo II
	Rechazar H_0	Error tipo I	Decisión correcta

$$\alpha = P(\text{error tipo I}) = P(\text{Rechazar } H_0 | H_0 \text{ verdadera}) = P(\text{Muestra dentro de } C | H_0)$$
$$\beta = P(\text{error tipo II}) = P(\text{No rechazar } H_0 | H_0 \text{ falsa}) = P(\text{Muestra fuera de } C | H_1)$$

- Los valores para la significancia de una prueba de uso más común son 0.01, 0.05 y 0.10; o sea, el investigador está dispuesto a permitir 1%, 5% o 10% de cometer un error tipo I.
- Idealmente se desearía que la probabilidad de error tipo I fuera igual a cero. Sin embargo, si se desea α = 0, nunca se podría tomar la decisión de rechazar la hipótesis nula.
- La decisión de rechazar la hipótesis nula es importante, ya que la decisión se basa en una muestra y no en la población; por lo cual existe la posibilidad de cometer un error tipo I



ESTABLECIMIENTO DE HIPOTESIS Y VARIABLES

2.3Establecimiento De Hipótesis Y Variables

Este establecimiento estará dentro del marco teórico donde se inserta el fenómeno de nuestro interés.

Hipótesis:

Es el intento de explicación a una respuesta “provisional” a un fenómeno. Su función consiste en delimitar el problema que se va a investigar según algunos elementos tales como el tiempo, el lugar, las características de los sujetos, etc. Es decir, es una respuesta tentativa a un problema, pero que no está comprobada. Llegar a comprobar o rechazar la hipótesis que se ha elaborado previamente, confrontando su enunciado teórico con los hechos empíricos, es el objetivo primordial de todo estudio que pretenda explicar algún campo de la realidad. Existen dos variables que se toman como variables principales que son: las variables independientes y las variables dependientes.

Variable:

Es cualquier característica o cualidad que es susceptible de asumir diferente valores, ya sea cuantitativamente o cualitativamente.

Es decir, que puede variar aunque para un objeto determinado pueda tener un valor fijo. Por ejemplo si hablamos de un celular, no puede ser en si una variable

pero si nos referimos a las opciones que maneja cada celular, estamos en presencia de una variable. O sea que esa cualidad de la mesa puede asumir diferentes valores.

FORMULACION DE LA HIPOTESIS Y VARIABLES

Establecimiento de hipótesis

Las hipótesis son suposiciones conjeturales, en transición hacia su confirmación. Se desprenden del análisis teórico para plantear supuestos con alto grado de certeza.

Las hipótesis son el vínculo entre la teoría y la práctica; se construyen con tres elementos:

El objeto de estudio, al cual se denomina unidad de análisis.

Las variables, que se conocen como propiedades de las unidades del análisis.

La relación, que se describe como los términos lógicos que unen los objetos con sus propiedades.

Engels dice: "hipótesis es una forma de desarrollo de las ciencias naturales, por cuanto son pensamientos..."

Algunos autores conciben la hipótesis como una proposición que puede ser puesta a prueba para determinar su validez.

"La hipótesis es una afirmación tentativa, más que definitiva. Debe ser formulada de tal manera que pueda ser potencialmente aceptada o rechazada por medio de los hallazgos. La teoría sirve de base a la hipótesis y a su vez es modificada por ésta. La hipótesis requiere de la investigación, para la comprobación de los postulados que contiene".

Requisitos para elaborar una hipótesis

Construirla con base en la realidad que se pretende explicar.

Fundamentarla en la teoría referente al hecho que se pretende explicar.

Establecer relaciones entre variables.

Ser susceptible de ponerse a prueba, para verificar su validez.

Dar la mejor respuesta al problema de investigación, con un alto grado de probabilidad.

No incurrir en nada superfluo en su construcción.

Clasificación de las hipótesis

Sustantivas.

Se refieren a la realidad social.

De generalización. Se refieren a los datos.

Generales. Relación entre variables básicas.

Particulares. Derivan de una hipótesis básica.

Alternativas. Misma variable independiente, con otras dependientes.

Descriptivas. Señalan la existencia de regularidades empíricas.

Tipos ideales complejos. Ponen a prueba la existencia de relaciones entre un tipo ideal y la realidad.

Analíticas. Formulan relaciones entre variables y explican la relación entre diversos factores.

Postfacto. Se deducen de la observación de un fenómeno.

Antefacto. Inducen a una explicación antes de la observación.

Nulas. Se diseñan para reafirmar que no se ha rechazado una hipótesis verdadera por una falsa.

De trabajo. Provisional y previa a la investigación definitiva, a efecto de hallar otras más sugestivas.

Función de las hipótesis

Indicar el camino para la búsqueda de la verdad objetiva.

Impulsar el trabajo científico.

Sistematizar el conocimiento.

Permiten explicar el objeto de estudio.

Sirven de enlace entre el conocimiento ya obtenido y el que se busca.

Las hipótesis son intentos de explicación mediante una suposición verosímil que requiere comprobarse.

Variables
Son discusiones que pueden darse entre individuos y conjuntos. El término variable significa características, aspecto, propiedad o dimensión de un fenómeno puede asumir distintos valores.

Para operatividad variable, se requiere precisar su valor, traduciéndolas a conceptos susceptibles de medir, Por tanto, conviene considerar su definición nominal, real, operativa: lo que significa el término, la realidad y la práctica.
Clasificación de variables

En términos generales, las variables se clasifican según el nivel de medición que representan:
Variables cualitativas. Son aquéllas que se refieren a cualidades o atributos no medibles en números. Por ejemplo, organización, personal y funciones.
Variables cuantitativas. Son las susceptibles de medirse en términos numéricos. Se subdividen a su vez en:

Cuantitativas continuas. Pueden asumir cualquier valor. Por ejemplo: peso, edad y talla.
Cuantitativas discontinuas. Asumen sólo valores enteros. Por ejemplo, número de hijos.
Variables independientes. Expresan las causas del fenómeno. Por ejemplo, organización deficiente.

Variables dependientes. Expresan las consecuencias del fenómeno. Por ejemplo, calidad de la enseñanza.

Prueba de hipótesis.

El propósito central de la investigación lo constituye la prueba de hipótesis. Se pretende comprobar si los hechos observados concuerdan con las hipótesis planteadas. En general, comprende dos pasos, que son:

Selección de la técnica.

Recolección de la información.

Selección de la técnica

Para comprobar o refutar las hipótesis es necesario elegir por lo menos dos o tres técnicas de investigación, y diferentes tipos de observación de fenómenos. En ciencias sociales, deben aplicarse la técnica documental y la de campo. Es importante hacer las siguientes consideraciones:

La técnica será acorde al tipo de hipótesis que se desea comprobar. Diseñar los instrumentos según la técnica elegida.

Probar los instrumentos.

Determinar la muestra.

Recolección de la Información

La manera más formal de proceder a la búsqueda de información es seguir los lineamientos del método científico. La estadística resulta de gran utilidad en el manejo de información. El proceso consiste en:

Recoger la información.

Tabularla.
Presentarla.
Analizarla.
El aspecto medular del manejo de información es la recolección, ya que el procesamiento de datos depende de la confiabilidad que aquélla pueda tener.

Métodos de recolección de datos:

Encuestas: La información se recoge por muestras, por lo que no se aplica a la población total.

Censos: La información se recoge en forma general a toda la población. Registros: La información es continua. Se recoge a medida que se va produciendo.

APLICACIONES DEL MUESTREO DE ACEPTACIÓN

El concepto de muestreo de aceptación va asociado a inspección, por lo que acarrea todos los problemas que supone confiar la calidad en la inspección. Sin embargo, esto no es achacable al muestreo en sí, ya que este mismo inconveniente lo tiene la inspección 100%. La primera cuestión que se plantea ante una inspección de recepción es si se realiza un muestreo o si es preciso una inspección al 100%. Deming demuestra que la situación óptima (mínimo coste esperado) es:

Si $p < k1 / k2$ Aceptar sin inspección.

Si $p > k1 / k2$ Realizar inspección 100%. dónde:

p: Peor fracción defectuosa esperada del lote.

k1 : Coste de inspeccionar una pieza.

k2 : Coste de aceptar una pieza defectuosa.

TIPOS DE MUESTREOS DE ACEPTACIÓN

Los planes de muestreo se pueden clasificar de diversas formas:

♦ De acuerdo con la naturaleza de la población base. Pueden ser:

Lote aislado.

Lote a lote (producción uniforme de lotes).

Fabricaciones continuas (por ejemplo industria química, plantas embotelladoras, etc.).

♦ De acuerdo con la naturaleza de la característica inspeccionada. Pueden ser:

Por atributos. La característica es de tipo cualitativo (pasa /nopasa). Una variante es la que considera “el número de defectos”, de modo que un pieza puede estar penalizada por varios defectos.

Por variables. La característica es de tipo cuantitativo (por ejemplo longitud, peso, etc.).

♦ De acuerdo con el número de muestras a tomar. Pueden ser:

Simples. Se toma una muestra con la que hay que decidir la aceptación o el rechazo.

Dobles. Se toman hasta dos muestras con las que hay que decidir la aceptación o el rechazo. Es posible aceptar o rechazar solo con la primera muestra si el resultado es muy bueno o muy malo. Si es un resultado intermedio, se extrae una segunda muestra. En principio el tamaño de las dos muestras puede ser diferente.

Múltiple. Conceptualmente es igual al muestreo doble pero en este caso se extrae hasta n muestras diferentes.

Secuencial. En este caso se van extrayendo los elementos uno a uno y según los resultados que se van acumulando de elementos aceptados y rechazados, llega un momento en el que se tiene información suficiente para aceptar o rechazar el lote.

Regresión Lineal y Correlación

Introducción:

La regresión y la correlación son dos técnicas estrechamente relacionadas y comprenden una forma de estimación.

En forma más especifica el análisis de correlación y regresión comprende el análisis de los datos muéstrales para saber qué es y cómo se relacionan entre sí dos o más variables en una población. El análisis de correlación produce un número que resume el grado de la correlación entre dos variables; y el análisis de regresión da lugar a una ecuación matemática que describe dicha relación.

El análisis de correlación generalmente resulta útil para un trabajo de exploración cuando un investigador o analista trata de determinar que variables son potenciales importantes, el interés radica básicamente en la fuerza de la relación. La correlación mide la fuerza de una entre variables; la regresión da lugar a una ecuación que describe dicha relación en términos matemáticos

Los datos necesarios para análisis de regresión y correlación provienen de observaciones de variables relacionadas.

Lineal

La regresión lineal es una técnica que permite cuantificar la relación que puede ser observada cuando se grafica un diagrama de puntos dispersos correspondientes a dos variables, cuya tendencia general es rectilínea; relación que cabe compendiar mediante una ecuación “del mejor ajuste” de la forma:

$y = a + bx$

Dos características importantes de una ecuación línea:

- La independencia de la recta
- La localización de la recta en algún punto.

En esta ecuación, “y” representa los valores de la coordenada a lo largo del eje vertical en el gráfico (ordenada); en tanto que “x” indica la magnitud de la coordenada sobre el eje horizontal (abscisa). El valor de “a” (que puede ser negativo, positivo o igual a cero) es llamado el intercepto; en tanto que el valor de “b” (el cual puede ser negativo o positivo) se denomina la pendiente o coeficiente de regresión.

Serie de datos para el cálculo de una regresión (“a” y “b”) y del coeficiente de correlación (“r”)

Número	Valores de x	Valores de y	Número	Valores de x	Valores de y
1	9,0	0,50	7	6,7	1,00
2	9,4	0,50	8	8,4	0,50
3	7,4	1,23	9	8,0	0,50
4	9,7	1,00	10	10,0	0,50
5	10,4	0,30	11	9,2	0,50
6	5,0	1,50	12	6,2	1,00
			13	7,7	0,50

El procedimiento para obtener valores de “a” y “b” para una serie de pares de datos de “x” y de “y” (tal como la presentada en la Figura 1 y/o en la Tabla 1) es como sigue:

- Calcule, para cada par de valores de “x” e “y”, las cantidades “x^2”, “y^2”, y “x.y”.
- Obtenga las sumas (∑) de estos valores para todos los pares de datos de “x” e “y”, así como las sumas del total de los valores de “x” e “y”. Los resultados de los Pasos 1 y 2 aparecerán en forma similar a la siguiente:

Número de pares de datos	x	x^2	y	y^2	x.y
1	…	…	…	…	…

2 3 . . . n	… … …	… … …	… … …	… … …	… … …
Monto de las sumas	∑x	∑x²	∑y	∑y²	∑x·y

- Estime la pendiente (b) por medio de la relación:

$$b = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}$$

- Estime el intercepto (a) por medio de la relación:

$$a = \left[\frac{\Sigma y}{n} - \left(b \cdot \frac{\Sigma x}{n}\right)\right]$$

Correlación

El análisis de correlación se encuentra estrechamente vinculado con el análisis de regresión y ambos pueden ser considerados de hecho como dos aspectos de un mismo problema.

La correlación entre dos variables es el grado de asociación entre las mismas. Este es expresado por un único valor llamado coeficiente de correlación (r), el cual puede tener valores que oscilan entre -1 y +1. Cuando “r” es negativo, ello significa que una variable (ya sea “x” o “y”) tiende a decrecer cuando la otra aumenta (se trata entonces de una “correlación negativa”, correspondiente a un valor negativo de “b” en el análisis de regresión). Cuando “r” es positivo, en cambio, esto significa que una variable se incrementa al hacerse mayor la otra (lo cual corresponde a un valor positivo de “b” en el análisis de regresión).

Los valores de “r” pueden calcularse fácilmente en base a una serie de pares de datos de “x” e “y”. De este modo “r” puede ser obtenido indirectamente a partir de la relación:

$$r^2 = \frac{\Sigma xy - \left[\frac{(\Sigma x)(\Sigma y)}{n}\right]^2}{\left[\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right] \cdot \left[\Sigma y^2 - \frac{(\Sigma y)^2}{n}\right]}$$

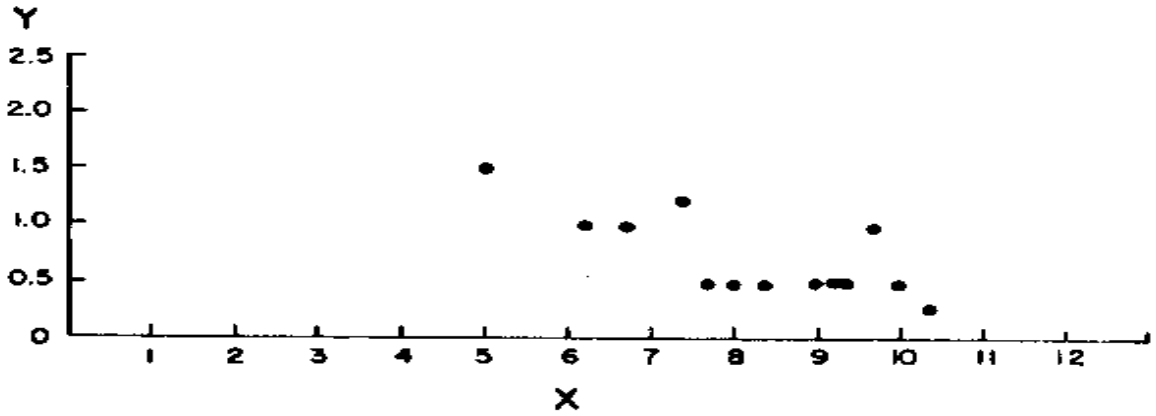
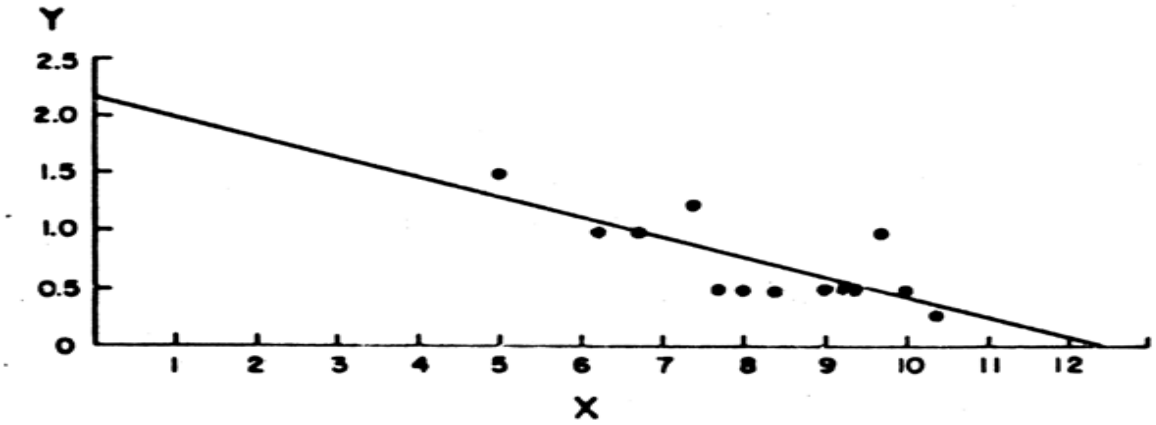


Diagrama de puntos dispersos correspondientes a pares de valores de “x” y de “y”. Nótese que “y” tiende a decrecer con el aumento de “x”, lo cual sugiere coeficientes de regresión y de correlación negativos (basado en la Tabla 1).



Los mismos datos que en 1a Fig. 1a, pero ajustados en base a la regresión y = 2,16

- 0,173x, con r = 0,75

La cual proporciona el valor del “coeficiente de determinación” (r^2). Entonces, lo único necesario es calcular

$$|r| = \sqrt{r^2}$$

Es decir, tomar la raíz indicada del coeficiente de determinación a los fines de obtener el valor absoluto de “r”, y luego agregar el signo (+ o -) de acuerdo a que la correlación sea positiva o negativa (lo cual puede ser establecido visualmente a partir del gráfico, o bien en base al cálculo del valor de “b” de la correspondiente regresión y utilizando para “r” el mismo signo).

Cuando se calculan los valores de “r” se querrá saber, sin embargo, hasta qué punto la correlación identificada pudiera haber surgido únicamente por casualidad. Esto puede ser establecido verificando si el valor estimado de “r” es “significativo”, es decir si el valor absoluto de “r” es mayor o igual que un valor “crítico” de “r” indicado en las tablas estadísticas.

Identificar el proceso de construcción del diagrama de dispersión.

Diagrama de dispersión:

Es el proceso a seguir para analizar la existencia de una relación lógica entre dos variables.
Describe la construcción de los Diagramas de Dispersión a partir de la recogida de datos acerca de dichas variables y el análisis, ya que ésta no implica la existencia de una relación lógica

Pasos previos a la construcción de un diagrama de dispersión

Paso 1: Elaborar una teoría admisible y relevante sobre la supuesta relación entre dos variables

Este paso previo es de gran importancia, puesto que el análisis de un diagrama de dispersión permite obtener conclusiones sobre la existencia de una relación entre dos variables, no sobre la naturaleza de dicha relación.

Paso 2: Obtener los pares de datos correspondientes a las dos variables

Al igual que en cualquier otra herramienta de análisis de datos, estos son la base de las conclusiones obtenidas, por lo tanto cumplirán las siguientes condiciones:

- En cantidad suficiente: Se consideran necesarios al menos 40 pares de datos para construir un diagrama de dispersión.
- Datos correctamente emparejados: Se estudiará la relación entre ambos.
- Datos exactos: Las inexactitudes afectan a su situación en el diagrama desvirtuando su apariencia visual.
- Datos representativos: Asegúrese de que cubren todas las condiciones operativas del proceso.
- Información completa: Anotar las condiciones en que han sido obtenidos los datos.

Pasos en la construcción de un diagrama de dispersión

Paso 1: Determinar los valores máximo y mínimo para cada una de las variables

Ejemplo: Tabla de los datos recogidos

Teoría: La fatiga es causa de los errores de tecleo Número de errores de tecleo según la hora del día							
Hora	Error	Hora	Error	Hora	Error	Hora	Error
11:00	25	13:30	38	09:30	15	12:45	33
14:15	45	12:15	14	15:45	72	10:45	17
10:00	7	16:30	56	10:30	35	11:45	8
13:45	26	14:45	60	11:30	18	15:30	30
09:15	22	12:30	11	16:15	63	09:45	22
16:00	50	13:30	55	15:30	62	12:45	41
12:30	60	15:15	40	09:45	31	09:15	22
13:45	19	14:45	25	14:30	32	15:15	80
14:30	78	10:45	10	14:45	56	12:00	30
13:00	22	13:45	19	12:15	45	10:15	22

Paso 2: Decidir sobre qué eje se representará a cada una de las variables

Si se está estudiando una posible relación causa-efecto, el eje horizontal representará la supuesta causa.

Paso 3: Trazar y rotular los ejes horizontal y vertical

La construcción de los ejes afecta el aspecto y la consiguiente interpretación del diagrama.

a. Los ejes han de ser aproximadamente de la misma longitud, determinando un área cuadrada.
b. La numeración de los ejes ha de ir desde un valor ligeramente menor que el valor mínimo de cada variable hasta un valor ligeramente superior al valor máximo de las mismas. Esto permite que los puntos abarquen toda el área de registro de los datos.
c. Numerar los ejes a intervalos iguales y con incrementos de la variable constantes.
d. Los valores crecientes han de ir de abajo hacia arriba y de izquierda a derecha en los ejes vertical y horizontal respectivamente.
e. Rotular cada eje con la descripción de la variable correspondiente y con su unidad de medida.

Paso 4: Marcar sobre el diagrama los pares de datos

Para cada par de datos localizar la intersección de las lecturas de los ejes correspondientes y señalarlo con un punto o símbolo.

Cuando coinciden muchos pares de puntos, el diagrama de dispersión puede hacerse confuso. En este caso es recomendable utilizar una “Tabla de Correlación” para representar la correlación.

Paso 5: Rotular el gráfico

Se rotula el título del gráfico y toda aquella información necesaria para su correcta comprensión.

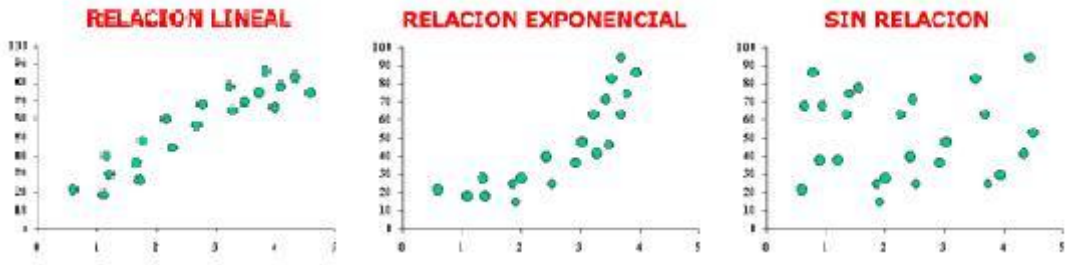
En general, es conveniente incluir una descripción adicional del objeto de las medidas y de las condiciones en que se han realizado, ya que esta información puede ayudar en la interpretación del diagrama.

Identificar el concepto de coeficiente de correlación.

En una distribución bidimensional puede ocurrir que las dos variables guarden algún tipo de relación entre sí.

Por ejemplo, si se analiza la estatura y el peso de los alumnos de una clase es muy posible que exista relación entre ambas variables: mientras más alto sea el alumno, mayor será su peso.

Mide el grado de intensidad de esta posible relación entre las variables. Este coeficiente se aplica cuando la relación que puede existir entre las variables es lineal (es decir, si representáramos en un gráfico los pares de valores de las dos variables la nube de puntos se aproximaría a una recta).



No obstante, puede que exista una relación que no sea lineal, sino exponencial, parabólica, etc. En estos casos, el coeficiente de correlación lineal mediría mal la intensidad de la relación las variables, por lo que convendría utilizar otro tipo de coeficiente más apropiado.
Para ver, por tanto, si se puede utilizar el coeficiente de correlación lineal, lo mejor es representar los pares de valores en un gráfico y ver qué forma describe.
El coeficiente de correlación lineal se calcula aplicando la siguiente fórmula:

$$r = \frac{1/n * \Sigma (x_i - xm) * (y_i - ym)}{\left((1/n * \Sigma (x_i - xm)^2) * (1/n * \Sigma (y_i - ym)^2) \right)^{1/2}}$$

Es decir:
Numerador: se denomina covarianza y se calcula de la siguiente manera:

En cada par de valores (x,y) se multiplica la "x" menos su media, por la "y" menos su media. Se suma el resultado obtenido de todos los pares de valores y este resultado se divide por el tamaño de la muestra.
Denominador se calcula el producto de las varianzas de "x" y de "y", y a este producto se le calcula la raíz cuadrada.
Los valores que puede tomar el coeficiente de correlación "r" son:
-1 < r < 1 Si "r" > 0, la correlación lineal es positiva (si sube el valor de una variable sube el de la otra). La correlación es tanto más fuerte cuanto más se aproxime a 1.
Por ejemplo: altura y peso: los alumnos más altos suelen pesar más.
Si "r" < 0, la correlación lineal es negativa (si sube el valor de una variable disminuye el de la otra). La correlación negativa es tanto más fuerte cuanto más se aproxime a -1.
Por ejemplo: peso y velocidad: los alumnos más gordos suelen correr menos.
Si "r" = 0, no existe correlación lineal entre las variables. Aunque podría existir otro tipo de correlación (parabólica, exponencial, etc.)
De todos modos, aunque el valor de "r" fuera próximo a 1 o -1, tampoco esto quiere decir obligatoriamente que existe una relación de causa-efecto entre las dos variables, ya que este resultado podría haberse debido al puro azar.

Explicar el proceso de regresión lineal y su interpretación:
En estadística la regresión lineal o ajuste lineal es un modelo matemático usado para aproximar la relación de dependencia entre una variable dependiente Y, las variables independientes X_i y un término aleatorio ε. Este modelo puede ser expresado como:

$$Y_t = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

Dónde:

Y_t: Variable dependiente, explicada o regresando.

$X_1, X_2, \cdots, X_p$: Variables explicativas, independientes o regresares.

$\beta_0, \beta_1, \beta_2, \cdots, \beta_p$: Parámetros, miden la influencia que las variables explicativas tienen sobre el reprendiendo.

Donde $\beta_0$es la intersección o término "constante", las $\beta_i \ (i > 0)$son los parámetros respectivos a cada variable independiente, y $p$es el número de parámetros independientes a tener en cuenta en la regresión. La regresión lineal puede ser contrastada con la regresión no lineal.

La correlación ("r") de las rectas determinará la calidad del ajuste. Si r es cercano o igual a 1, el ajuste será bueno y las predicciones realizadas a partir del modelo obtenido serán muy fiables (el modelo obtenido resulta verdaderamente representativo); si r es cercano o igual a 0, se tratará de un ajuste malo en el que las predicciones que se realicen a partir del modelo obtenido no serán fiables (el

modelo obtenido no resulta representativo de la realidad). Ambas rectas de regresión se intersecan en un punto llamado centro de gravedad de la distribución.

- Diagrama de dispersión

Es el primer paso para determinar si existe o no una relación entre dos variables es observar la gráfica de datos observados. Esta grafica se llama diagrama de dispersión.

Un diagrama nos puede dar dos tipos de información, visualmente podemos buscar patrones que nos indiquen que las variables están relacionadas. Entonces si esto sucede, podemos ver qué tipo de línea, o ecuación de estimación, describe esta relación.

Primero tomamos los datos de la tabla que deseamos analizar y dependiendo de que se desea averiguar se construye la gráfica colocando la variable dependiente en el eje Y y la independiente en el eje X, Cuando vemos todos estos puntos juntos, podemos visualizar la relación que existe entre estas dos variables. Como resultado, también podemos trazar, “o ajustar” una línea recta a través de nuestro diagrama de dispersión para representar la relación. Es común intentar trazar estas líneas de forma tal que un número igual de puntos caiga a cada lado de la línea.

Para calcular una ecuación para una línea dibujada en medio de un conjunto de puntos en un diagrama de dispersión. Para esto debemos minimizar el error entre los puntos estimados en la línea y los verdaderos puntos observados que se utilizaron para trazarla.

Para esto debemos introducir un nuevo símbolo, para simbolizar los valores individuales de los puntos estimados, esto es, aquellos puntos que caen en la línea de estimación. En consecuencia escribiremos la ecuación para la línea de estimación como

Una forma en que podemos medir el error de nuestra línea de estimación es sumando todas las diferencias, o errores, individuales entre los puntos observados y los puntos estimados.

La suma de las diferencias individuales para calcular el error no es una forma confiable de juzgar la bondad de ajuste de una línea de estimación.

El problema al añadir los errores individuales es el efecto de cancelación de los valores positivos y negativos, por eso usamos valores absolutos en esta diferencia a modo de cancelar la anulación de los signos positivos y negativos, pero ya que estamos buscando el menor error debemos buscar un método que nos muestre la magnitud del error, decimos que la suma de los valores absolutos no pone énfasis en la magnitud del error.

Parece razonable que mientras más lejos este un punto de la línea e estimación, más serio seria el error, preferiríamos tener varios errores pequeños que uno grande. En efecto, deseamos encontrar una forma de “penalizar” errores absolutos grandes, de tal forma que podamos evitarlos. Puede lograr esto si cuadramos los errores individuales antes de sumarlos. Con estos se logran dos objetivos:

- penaliza los errores más grandes
- cancela el efecto de valores positivos y negativos

Como estamos buscando la línea de estimación que minimiza la suma de los cuadrados de los errores a esto llamamos método de mínimos cuadrados.

Si usamos el método de mínimos cuadrados, podemos determinar si una línea de estimación tiene un mejor ajuste que otro. Pero para un conjunto de puntos de datos a través de los cuales podríamos trazar un número infinito de líneas de estimación, ¿cómo podemos saber cuándo hemos encontrado la mejor línea de ajuste?

Los estadísticos han derivado dos ecuaciones que podemos utilizar para encontrar la pendiente y la intersección Y de la línea de regresión del mejor ajuste. La primera fórmula calcula la pendiente.

- b = pendiente de la línea de estimación de mejor ajuste
- X = valores de la variable independiente
- Y = valores de la variable dependiente
- = media de los valores de la variable independiente
- = media de los valores de la variable dependiente
- n = número de puntos de datos

La segunda ecuación calcula la intersección en Y

- a = intersección en Y
- b = pendiente de la ecuación anterior
- = media de los valores de la variable dependiente
- = media de los valores de la variable independiente

- Coeficiente de correlación

El coeficiente de correlación es la segunda medida que podemos usar para describir que también una variable es explicada por la otra. Cuando tratamos con

muestras, el coeficiente de variación de muestra se denomina como r y es la raíz cuadrada del coeficiente de determinación de muestra:

Cuando la pendiente de estimación de la muestra es positiva, r es la raíz cuadrada positiva, pero si b es negativa, r es la raíz cuadrada negativa. Por lo tanto, el signo de indica la dirección de la relación entre las dos variables X y Y. Si existe una relación inversa, esto es, si y disminuye Y, X

Intersección Y

Variable dependiente

Pendiente de la línea

Variable independiente

$$\hat{Y} = a + bX$$

$$b = \frac{XY - n\bar{X}\bar{Y}}{X^2 - n\bar{X}^2}$$

- Ecuación de regresión

Consiste en determinar los valores de "a" y "b " a partir de la muestra, es decir, encontrar los valores de a y b con los datos observados de la muestra. El método de estimación es el de Mínimos Cuadrados, mediante el cual se obtiene:

$$a = \overline{Y} - b\ \overline{X}$$

$$b = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2} = \frac{\sum X_i\ Y_i - n\ \overline{X}\ \overline{Y}}{\sum X_i^2 - \overline{X}^2}$$

Luego, la ecuación de regresión muestral estimada es

$$\hat{Y} = a + b\ X$$

Que se interpreta como:

a es el estimador de a

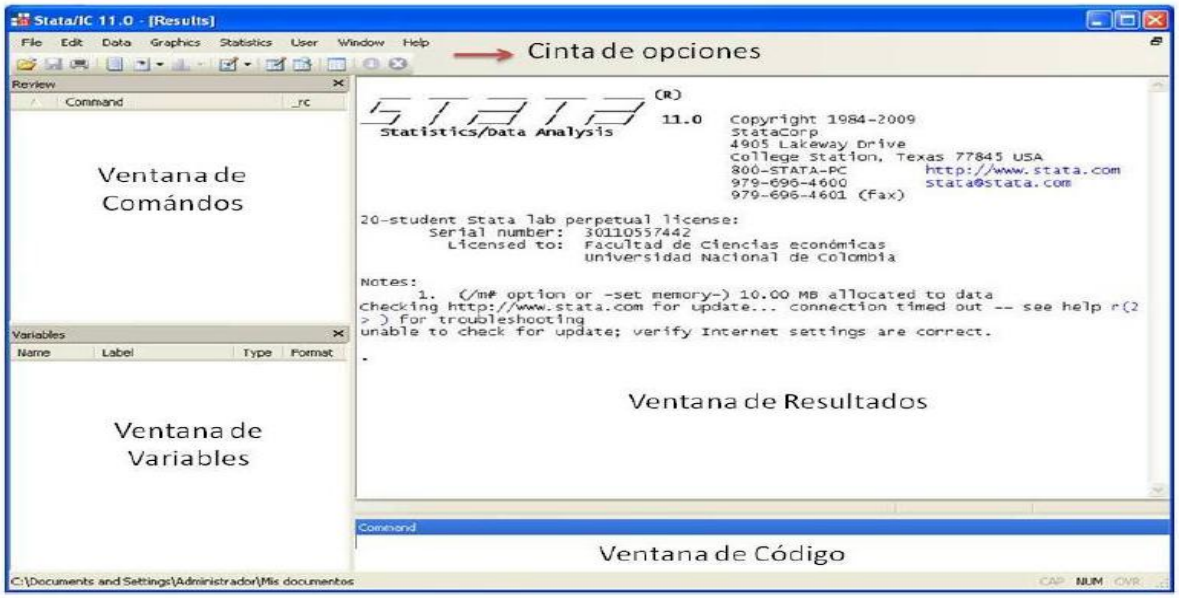
Es el valor estimado de la variable Y cuando la variable X = 0

b es el estimador de b , es el coeficiente de regresión

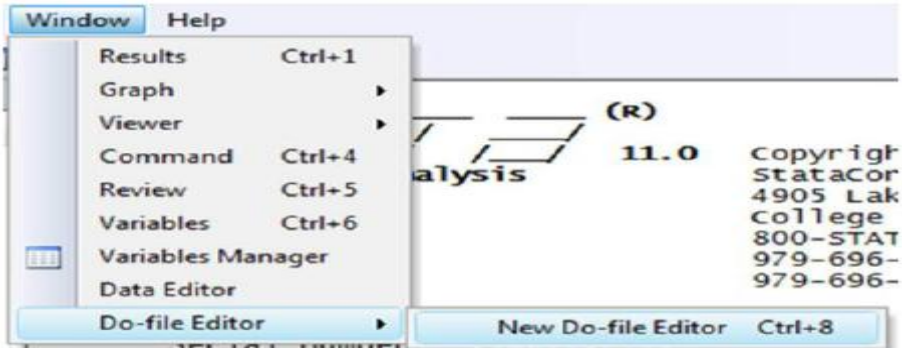
Está expresado en las mismas unidades de Y por cada unidad de X. Indica el número de unidades en que varía Y cuando se produce un cambio, en una unidad, en X (pendiente de la recta de regresión).

Un valor negativo de b sería interpretado como la magnitud del decremento en Y por cada unidad de aumento en X.

Explicar el proceso de regresión lineal en software.



Statu permite trabajar de igual forma por medio de un Script, en el cual se digitan instrucciones sin ser ejecutadas de forma inmediata. Para abrir un nuevo Script debe dirigirse a la pestaña Windows en la cinta de opciones y seleccionar la última opción: Do-file Editor6. En el momento en que se desee ejecutar los comandos, puede: seleccionar el ícono el cual hará que se ejecute todo o seleccionar el comando que desea y presionar Ctrl+d.



En el editor es posible crear un archivo .log en el cual se va a guardar todo lo que se realice. Por medio del siguiente código log using “nombre del archivo”.log, remplace. Cuando termine de trabajar deberá escribir log close para completar el proceso.

Para cargar los datos se debe utilizar el comando insheet using ‚nombre del archivo‘.txt. Para visualizar los datos ya cargados debe dirigirse en la ventana principal de Stata a la pestaña Windows>Data Editor.

Para introducir al usuario al funcionamiento del software, se va a trabajar sobre una base de datos de ejemplo que contiene Stata. Para abrirla debe dirigirse a File(Archivo) Example Datasets (bases de datos de ejemplo), seleccionar las que están instaladas de forma predeterminada (Example Datasets installed with Stata) y finalmente seleccionar la opción que le permitirá utilizarla, use.

Inmediatamente aparecerá en la ventana de resultados un comando que le indica que la base de datos fue cargada exitosamente. Igualmente el contenido de la base de datos aparecerá en la ventana de variables. Igualmente se hubiera podido cargar la base de datos por medio de la ventana de códigos: para esto deberá escribir el comando sysuse seguido del nombre del archivo que desea abrir, en este caso auto.dta y presionando Enter.

```
. sysuse auto.dta
(1978 Automobile Data)
```

En muchos casos es de suma importancia para los investigadores conocer información más detallada sobre el contenido de la base de datos y especialmente en la econometría las estadísticas básicas sobre las variables son necesarias para realizar el análisis. Para esto, Stata tiene en la cinta de opciones un ícono llamado Statistics, el cual además de contener la opción para visualizar las estadísticas básicas de las variables, va a permitir más adelante la aplicación de las diferentes metodologías (regresión lineal, métodos de análisis de series de tiempo, modelos no lineales, etc.).

La obtención de las estadísticas básicas puede ser de dos formas:

1. Seleccionando en orden los íconos Statistics >Summaries, tables, and tests > Summary and descriptive statistics > Summary statistics y finalmente validando la instrucción a través del botón OK.

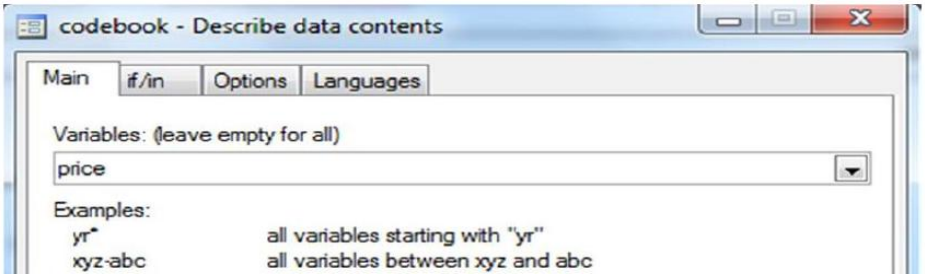
2. Escribiendo en la ventana de códigos el comando summarize y presionando Enter.

Independientemente de la forma en que se realice, en la ventana de resultados aparecerá un cuadro que muestra: 1. El número de observaciones, 2. La media, 3. La desviación estándar, 4. El mínimo y 5. El máximo de cada variable.

```
. summarize
```

	1.	2.	3.	4.	5.
Variable	Obs	Mean	Std. Dev.	Min	Max
make	0				
price	74	6165.257	2949.496	3291	15906
mpg	74	21.2973	5.785503	12	41
rep78	69	3.405797	.9899323	1	5
headroom	74	2.993243	.8459948	1.5	5
trunk	74	13.75676	4.277404	5	23
weight	74	3019.459	777.1936	1760	4840
length	74	187.9324	22.26634	142	233
turn	74	39.64865	4.399354	31	51
displacement	74	197.2973	91.83722	79	425
gear_ratio	74	3.014865	.4562871	2.19	3.89
foreign	74	.2972973	.4601885	0	1

Esta es una vista general de las estadísticas de todas las variables, sin embargo y para más exactitud, Stata ofrece otro comando para analizar cada variable por separado, permitiendo que ningún detalle se escape. Este comando recibe el nombre de codebook y puede ser escrito así en la ventana de códigos u obtenerse el mismo resultado a través de la cinta de opciones: Data > Describe data > Describe data contents (codebook); inmediatamente aparecerá una ventana que le indicará si desea realizar la operación para una sola variable, caso para el cual deberá poner en el espacio Variables el nombre de la variable; o para todas las variables, caso en el cual deberá dejar el espacio vacío.

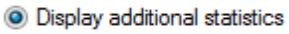


En la ventana de resultados aparecerá:

```
price                                                         Price
-------------------------------------------------------------------

                  type:  numeric (int)

                 range:  [3291,15906]                 units:  1
         unique values:  74                       missing .:  0/74

                  mean:   6165.26
             std. dev:    2949.5

            percentiles:        10%       25%       50%       75%       90%
                               3895      4195    5006.5      6342     11385
--more--
```

Nota: La palabra en azul more indica que más información se encuentra disponible. Para visualizarla solo debe seleccionar la palabra.

A partir del comando que resume las estadísticas básicas de las variables podemos obtener más información esencial. Escribiendo el comando summarize ‚el nombre de la variable', detail (en este caso summarize price, detail) o bien volviendo al menú Statistics >Summaries, tables, and tests > Summary and descriptive statistics > Summary statistics y eligiendo en la ventana que aparece la opción.

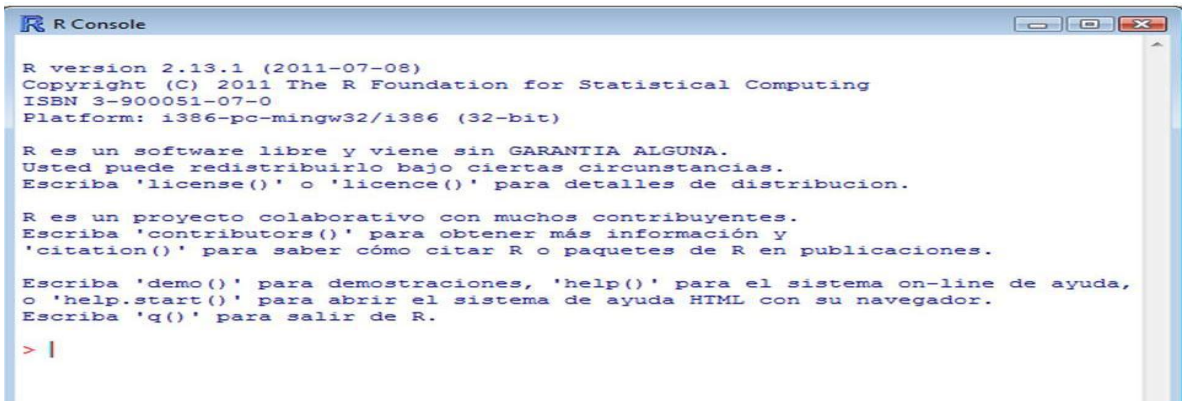


Esta acción mostrará en más detalle los percentiles, la varianza de la variable y su simetría.

R es un lenguaje y entorno (sistema) que provee gran variedad de técnicas estadísticas y gráficas para la aplicación de modelos lineales o no lineales, la realización de pruebas estadísticas básicas, análisis de series de tiempo, entre muchos otros. R se encuentra disponible como software libre7 y es altamente extensible, ofreciendo la posibilidad de incluir cuando sea necesario nuevos paquetes desarrollados por la comunidad que satisfagan la necesidad del usuario.

La interfaz de R es muy sencilla: en la parte superior se encuentra ubicada la cinta de herramientas y algunos botones de uso común tales como abrir, guardar, copiar, etc. La demás parte está compuesta por la consola principal, en la cual todos los comandos van a ser ejecutados (presionando Enter) y los resultados visualizados. Adicionalmente, a medida que se vaya digitando el código, algunas ventanas complementarias irán apareciendo.

De manera alterna, R permite trabajar en un editor o script en donde se digitan los comandos pero no se ejecutan inmediatamente, lo cual brinda al usuario mayor comodidad. Gran cantidad de códigos pueden ser copiados pero únicamente se van a ejecutar en la consola presionando la tecla F5.



Para comenzar a utilizar el software debe cargarse inicialmente una base de datos; para este efecto, es importante mencionar que en R se puede trabajar con varios tipos de archivo: .csv,.txt y .xls. Esta es una gran facilidad porque permite que el usuario maneje sus bases de datos en Excel y luego las importe para trabajar directamente en ellas. En este caso es necesario que en Excel el archivo quede guardado bajo el formato .csv (delimitado por comas).

Para comenzar a utilizar el programa y cargar las bases de datos sin problema alguno, es necesario dirigirse a la opción Archivo > Cambiar dir…, y escoger el destino en el cual se encuentran localizados los archivos sobre los cuales se va a trabajar, esto con el fin de que R los encuentre rápidamente. Igualmente antes de comenzar a insertar las órdenes, es recomendable que se limpie la memoria del software para evitar posibles incongruencias; para esto se utiliza el comando rm(list=ls()). Si el usuario necesita conocer más información acerca de las funciones de un comando, podrá digitar help(), metiendo dentro del paréntesis el código.

Para cargar los archivos, el código que debe transcribirse es read.csv2. La formación del código comienza por el nombre que se le asigna al nuevo objeto formado (esto con el fin de que el software lo identifique fácilmente), seguido de la instrucción formal read.csv2 y un paréntesis en donde debe especificarse el nombre del archivo original y si este tiene etiquetas o títulos. Luego de que el objeto ha sido creado, para visualizarlo es necesario llamarlo escribiendo nuevamente su nombre.

```
> Base1 <- read.csv2("Base1.csv", header=T)
> Base1
```

Nota: El símbolo <- representa la asignación de la orden al objeto Base1. Este procedimiento debe hacerse para todos los comandos debido a que R va a reconocer únicamente los objetos creados a los cuales se les asignó una instrucción. Puede ser sustituido por el símbolo =.

R permite transformar la forma de los datos haciendo que puedan ser leídos y entendidos como matrices, siguiendo el código rh1=as.matrix(Base1)8 . Si desea visualizar el objeto en una ventana adicional y de forma más organizada debe escribir View (), metiendo entre paréntesis el nombre del objeto que desee. En este caso la instrucción View (rh1) genera el siguiente resultado:

Dentro del paréntesis debe ir especificado el nombre del objeto con el cual se reconocen los datos.

Archivo

	CIUDADES	RH_1	RH_2	RH_3	RH_4	RH_5	RH_6	RH_7	RH_8	RH_9	RH_10	RH_11
1	Armenia	284120	0.00565	0.4640147	0.5373766	0.0880	1.095	0.266	22.4	0.354	3037.5	0.6282
2	Barranquilla	1163007	0.00721	0.4883377	0.5554151	0.0922	1.083	0.306	22.5	0.218	1054.1	0.8594
3	Bogotá, D.C.	7050228	0.01513	0.5631922	0.6293213	0.0973	0.999	0.538	22.6	0.447	1042.5	0.7052
4	Bucaramanga	520080	0.00344	0.5108233	0.5666209	0.0760	1.192	0.326	22.2	0.378	1068.0	1.0000
5	Cali	2169801	0.01163	0.5528726	0.6205193	0.0831	1.019	0.223	22.0	0.313	626.8	0.8590
6	Cartage	912674	0.01115	0.4839377	0.5469014	0.1213	1.207	0.147	24.4	0.215	1866.8	0.5235
7	Cúcuta	600049	0.01042	0.5193251	0.5768769	0.1133	1.147	0.217	22.4	0.181	1245.2	0.8040
8	Ibagué	509796	0.01130	0.5166996	0.5919223	0.1073	0.983	0.181	23.4	0.252	1677.9	0.4559
9	Manizales	383483	0.00460	0.4609772	0.5199243	0.0769	1.143	0.230	22.2	0.326	2233.2	0.7433
10	Medellín	2264776	0.01123	0.4963598	0.5592503	0.1256	1.219	0.300	26.3	0.310	1167.4	1.0000
11	Montería	390996	0.01562	0.4797224	0.5548727	0.1587	1.109	0.151	26.1	0.160	836.2	0.6002
12	Neiva	322098	0.00950	0.4980671	0.5510142	0.1131	1.100	0.179	23.3	0.282	1229.8	0.8032
13	Pasto	394074	0.01475	0.4974571	0.5674950	0.1115	1.052	0.128	21.8	0.397	807.1	1.0000
14	Pereira	448971	0.00607	0.4839616	0.5496611	0.1007	1.124	0.283	22.6	0.303	1557.0	0.6990
15	Popayán	261694	0.00805	0.5230494	0.5767888	0.0915	1.135	0.163	21.5	0.395	1994.4	0.6892
16	Riohacha	184847	0.04818	0.4279488	0.5072428	0.2713	1.033	0.152	26.1	0.188	1693.9	0.8537
17	Santa Marta	428374	0.01553	0.4724664	0.5337993	0.1244	1.005	0.116	22.9	0.228	1238.5	0.7396
18	Sincelejo	245180	0.01566	0.4713240	0.5370866	0.1612	1.228	0.107	24.9	0.159	948.7	0.6234
19	Tunja	161209	0.02256	0.5285888	0.5821245	0.1023	1.036	0.245	22.1	0.314	1221.9	0.9228
20	Valledupar	373872	0.02667	0.4363452	0.5069274	0.1587	1.017	0.147	24.5	0.258	757.3	0.9383
21	Villavicencio	400475	0.02595	0.5434616	0.6030498	0.1017	1.193	0.174	25.4	0.281	661.8	0.6042

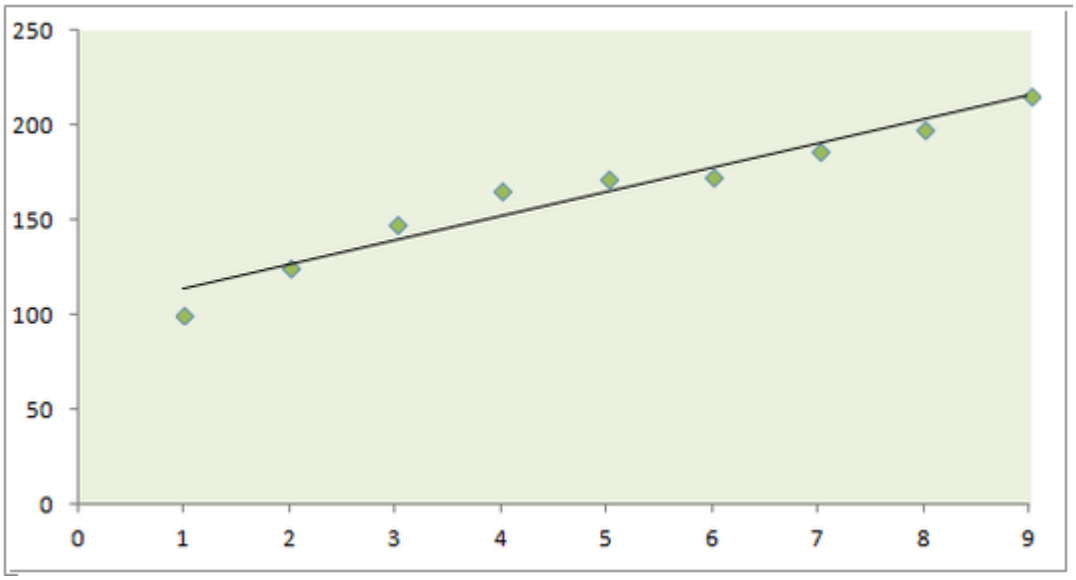
Dado que no todas las variables de la matriz están en formato numérico, para la manipulación de los datos es mejor eliminar la primera columna que contiene el nombre de las ciudades. Para esto, utilice el comando datosrh=Base1[,], en donde en la primera posición indica el número de filas y en la segunda el de columnas. Como desea eliminar una columna, el paréntesis debe contener los elementos de esta forma [,-1]. En cuanto al tema central de la estadística descriptiva, R ofrece una serie de comandos simples y fáciles de recordar a través de los cuales podemos visualizar los principales estadísticos de medición y los cuales se especifican de la siguiente forma:

 Para visualizar los estadísticos más básicos como el mínimo, máximo, la mediana y la media de cada uno de los datos del conjunto, el comando será summary(), colocando entre paréntesis el nombre otorgado a la matriz sin la variable texto.

```
> estadisticas=summary(datosrh)
> estadisticas
      RH_1               RH_2               RH_3               RH_4
 Min.   : 161209   Min.   :0.00344   Min.   :0.4279   Min.   :0.5069
 1st Qu.: 322098   1st Qu.:0.00805   1st Qu.:0.4725   1st Qu.:0.5374
 Median : 400475   Median :0.01130   Median :0.4964   Median :0.5554
 Mean   : 927134   Mean   :0.01430   Mean   :0.4961   Mean   :0.5607
 3rd Qu.: 600049   3rd Qu.:0.01562   3rd Qu.:0.5193   3rd Qu.:0.5769
 Max.   :7050228   Max.   :0.04818   Max.   :0.5632   Max.   :0.6293
```

Explicar el concepto de pronóstico en regresión lineal.
El modelo de pronóstico de regresión lineal permite hallar el valor esperado de una variable aleatoria *a* cuando *b* toma un valor específico. La aplicación de este método implica un supuesto de linealidad cuando la demanda presenta un comportamiento creciente o decreciente, por tal razón, se hace indispensable que previo a la selección de este método exista un análisis de regresión que determine la intensidad de las relaciones entre las variables que componen el modelo.

¿Cuándo utilizar un pronóstico de regresión lineal?
El pronóstico de regresión lineal simple es un modelo óptimo para patrones de demanda con tendencia (Creciente o decreciente), es decir, patrones que presenten una relación de linealidad entre la demanda y el tiempo.



Existen medidas de la intensidad de la relación que presentan las variables que son fundamentales para determinar en qué momento es conveniente utilizar regresión lineal.

Diseño de experimentos

INTRODUCCIÓN:
Los modelos de diseño de experimentos son modelos estadísticos clásicos cuyo objetivo es averiguar si unos determinados factores influyen en una variable de interés y, si existe influencia de algún factor, cuantificar dicha influencia.
Unos ejemplos donde habría que utilizar estos modelos son los siguientes:

— En el rendimiento de un determinado tipo de máquina (unidades producidas por día): se desea estudiar la influencia del trabajador que la maneja y la marca de la máquina.
— Se quiere estudiar la influencia de un tipo de pila eléctrica y de la marca, en la duración de las pilas.

— Una compañía telefónica está interesada en conocer la influencia de varios factores en la variable duración de una llamada telefónica. Los factores que se consideran son los siguientes: hora a la que se produce la llamada; día de la semana en que se realiza la llamada; zona de la ciudad desde la que se hace la llamada; sexo del que realiza la llamada; tipo de teléfono (público o privado) desde el que se realiza la llamada.
— Una compañía de software está interesada en estudiar la variable porcentaje en que se comprime un fichero, al utilizar un programa de compresión teniendo en cuenta el tipo de programa utilizado y el tipo de fichero que se comprime.
— Se quiere estudiar el rendimiento de los alumnos en una asignatura y, para ello, se desean controlar diferentes factores: profesor que imparte la asignatura; método de enseñanza; sexo del alumno.

La metodología del diseño de experimentos se basa en la experimentación. Es sabido que si se repite un experimento, en condiciones indistinguibles, los resultados presentan una cierta variabilidad. Si la experimentación se realiza en un laboratorio donde la mayoría de las causas de variabilidad están muy controladas, el error experimental será pequeño y habrá poca variación en los resultados del experimento. Pero si se experimenta en procesos industriales o administrativos la variabilidad será mayor en la mayoría de los casos.

Explicar el concepto de diseño de experimentos:

El objetivo del diseño de experimentos es estudiar si cuando se utiliza un determinado tratamiento se produce una mejora en el proceso o no. Para ello se debe experimentar aplicando el tratamiento y no aplicándolo. Si la variabilidad experimental es grande, sólo se detectará la influencia del uso del tratamiento cuando éste produzca grandes cambios en relación con el error de observación. La metodología del diseño de experimentos estudia cómo variar las condiciones habituales de realización de un proceso empírico para aumentar la probabilidad de detectar cambios significativos en la respuesta; de esta forma se obtiene un mayor conocimiento del comportamiento del proceso de interés.
Para que la metodología de diseño de experimentos sea eficaz es fundamental que el experimento esté bien diseñado.
Un experimento se realiza por alguno de los siguientes motivos:
— Determinar las principales causas de variación en la respuesta.
— Encontrar las condiciones experimentales con las que se consigue un valor extremo en la variable de interés o respuesta.
— Comparar las respuestas en diferentes niveles de observación de variables controladas.
— Obtener un modelo estadístico-matemático que permita hacer predicciones de respuestas futuras.

Identificar los elementos de ANOVA (Análisis de varianza):
El análisis de la varianza permite contrastar la hipótesis nula de que las medias de K poblaciones (K >2) son iguales, frente a la hipótesis alternativa de que por lo

menos una de las poblaciones difiere de las demás en cuanto a su valor esperado. Este contraste es fundamental en el análisis de resultados experimentales, en los que interesa comparar los resultados de K 'tratamientos' o 'factores' con respecto a la variable dependiente o de interés.

H_0: $\mu_1 = \mu_2 = ... = \mu_K = \mu$
H_1: $\exists \mu_j \neq \mu \quad j = 1, 2, ..., K$

El Anova requiere el cumplimiento los siguientes supuestos:

- Las poblaciones (distribuciones de probabilidad de la variable dependiente correspondiente a cada factor) son normales.
- Las K muestras sobre las que se aplican los tratamientos son independientes.
- Las poblaciones tienen todas igual varianza (homoscedasticidad).

Fuentes de variación
El ANOVA se basa en la descomposición de la variación total de los datos con respecto a la media global (SCT), que bajo el supuesto de que H0 es cierta es una estimación de σ^2 obtenida a partir de toda la información muestral, en dos partes:

- Variación dentro de las muestras (SCD) o Intra-grupos, cuantifica la dispersión de los valores de cada muestra con respecto a sus correspondientes medias.
- Variación entre muestras (SCE) o Inter-grupos, cuantifica la dispersión de las medias de las muestras con respecto a la media global.

Las expresiones para el cálculo de los elementos que intervienen en el Anova son las siguientes:

Media Global: $\overline{X} = \frac{\sum_{j=1}^{K} \sum_{i=1}^{nj} x_{ij}}{n}$.

Variación Total: $\mathrm{SCT} = \sum_{j=1}^{K} \sum_{i=1}^{nj} (x_{ij} - \overline{X})^2$

Variación Intra-grupos: $\mathrm{SCD} = \sum_{j=1}^{K} \sum_{i=1}^{nj} (x_{ij} - \overline{X_j})^2$

Variación Inter-grupos: $\mathrm{SCE} = \sum_{j=1}^{K} (\overline{X_j} - \overline{X})^2 n_j$

Siendo *xij* el i-ésimo valor de la muestra j-ésima; *nj* el tamaño de dicha muestra y $\overline{X_j}$ su media.

Cuando la hipótesis nula es cierta SCE/K-1 y SCD/n-K son dos estimadores insesgados de la varianza poblacional y el cociente entre ambos se distribuye

según una F de Snedecor con K-1 grados de libertad en el numerador y N-K grados de libertad en el denominador. Por lo tanto, si H0 es cierta es de esperar que el cociente entre ambas estimaciones será aproximadamente igual a 1, de forma que se rechazará H0 si dicho cociente difiere significativamente de 1.

La secuencia para realizar un ANOVA es:

Analizar

Comparar medias

ANOVA de un factor

Suma de cuadrados
La técnica fundamental consiste en la separación de la suma de cuadrados (SS, 'sum of squares') en componentes relativos a los factores contemplados en el modelo. Como ejemplo, mostramos el modelo para un ANOVA simplificado con un tipo de factores en diferentes niveles. (Si los niveles son cuantitativos y los efectos son lineales, puede resultar apropiado un análisis de regresión lineal)

$$SS_{\text{Total}} = SS_{\text{Error}} + SS_{\text{Factores}}$$

El número de grados de libertad (gl) puede separarse de forma similar y corresponde con la forma en que la distribución chi-cuadrado (χ^2 o Ji-cuadrada) describe la suma de cuadrados asociada.

$$gl_{\text{Total}} = gl_{\text{Error}} + gl_{\text{Factores}}$$

Cuadrados medios
Los cuadrados medios representan una estimación de la varianza de la población. Se calculan dividiendo la suma correspondiente de los cuadrados entre los grados de libertad.

En ANOVA, los cuadrados medios se utilizan para determinar si los factores (tratamientos) son significativos.

- El cuadrado medio del tratamiento se obtiene dividiendo la suma de los cuadrados del tratamiento entre los grados de libertad. El cuadrado medio del tratamiento representa la variación entre las medias de las muestras.
- El cuadrado medio del error (MSE) se obtiene dividiendo la suma de los cuadrados del error residual entre los grados de libertad. El MSE representa la variación dentro de las muestras.

Por ejemplo, usted hace un experimento para probar la efectividad de tres detergentes para ropa. Recolecta 20 observaciones para cada detergente. La variación entre las medias del Detergente 1, Detergente 2 y Detergente 3 se representa mediante el cuadrado medio del tratamiento. La variación dentro de las muestras se representa mediante el cuadrado medio del error.

Estadístico de prueba

Un estadístico de prueba es un valor estandarizado que se calcula a partir de los datos de la muestra durante una prueba de hipótesis. Puede utilizar los estadísticos de prueba para determinar si puede rechazar la hipótesis nula. El estadístico de prueba compara sus datos con lo que se espera según la hipótesis nula. El estadístico de prueba se utiliza para calcular el valor p. Cuando los datos muestran una clara evidencia en contra de los supuestos de la hipótesis nula, la magnitud del estadístico de prueba será grande y el valor p de la prueba puede ser lo suficientemente pequeño como para rechazar la hipótesis nula.

Por ejemplo, el estadístico de prueba para una prueba Z es el valor Z. Supongamos que usted realiza una prueba Z de dos colas con un nivel de significancia (α) de 0.05 y obtiene un valor Z de 2.5. Este valor Z corresponde a un valor p de 0.0124. Debido a que este valor p es menor que α, usted declara significancia estadística y rechaza la hipótesis nula.

Las diferentes pruebas de hipótesis utilizan diferentes estadísticos de prueba según el modelo de probabilidad asumido en la hipótesis nula. Las pruebas comunes y sus respectivos estadísticos de prueba incluyen:

Prueba de hipótesis	Estadístico de prueba
Prueba Z	Valor Z
Pruebas t	Valor t
ANOVA	Valor F
Pruebas de chi-cuadrada	Chi-cuadrada

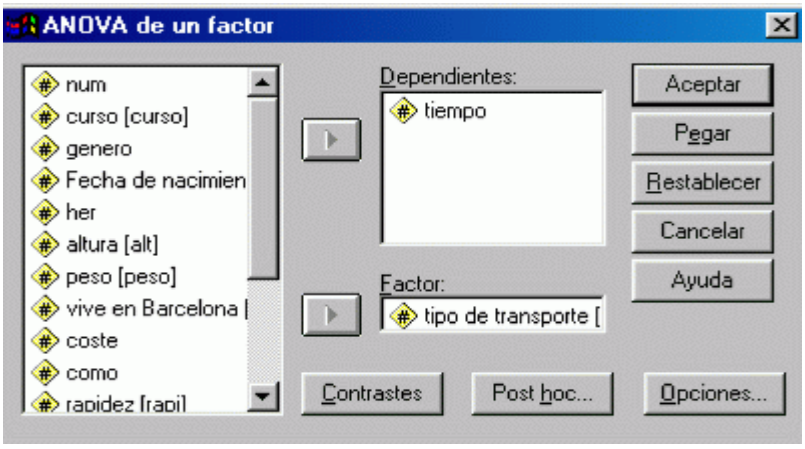
Explicar el proceso de construcción e interpretación de la tabla ANOVA.
Expresiones para el cálculo del ANOVA de un factor.

Fuente	Suma de cuadrados	Grados de libertad	Varianza	F_{cal}
	$SS_{lab} = \sum_{k=1}^{K} n_k(\bar{x}_k - \bar{\bar{x}})^2$	$K-1$	$MS_{lab} = \frac{SS_{lab}}{K-1}$	$F = \frac{MS_{lab}}{MS_R}$
	$SS_R = \sum_{k=1}^{K}\sum_{j=1}^{n_k}(x_{kj} - \bar{x}_k)^2$	$N-K$	$MS_R = \frac{SS_R}{N-K}$	
Total	$SS_T = \sum_{k=1}^{K}\sum_{j=1}^{n_k}(x_{kj} - \bar{\bar{x}})^2$	$N-1$	$MS_T = \frac{SS_T}{N-1}$	

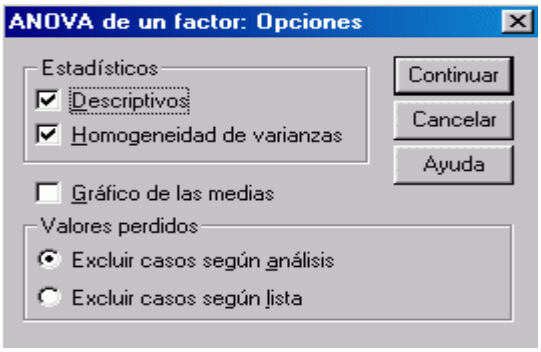
Se calculan, por tanto, *MS*lab y *MS*R como una medida de las dispersiones comentadas y se comparan mediante una prueba de hipótesis *F*. Si no existe diferencia estadísticamente significativa entre ellas, la presencia de errores aleatorios será la causa predominante de la discrepancia entre los valores medios. Si, por el contrario, existe algún error sistemático, *MS*lab será mucho mayor que *MS*R, con lo cual el valor calculado de *F* será mayor que el valor tabulado *F*tab para el nivel de significación a escogido y los grados de libertad mencionados.

Explicar la prueba ANOVA con software.

Se abre el siguiente cuadro de diálogo:



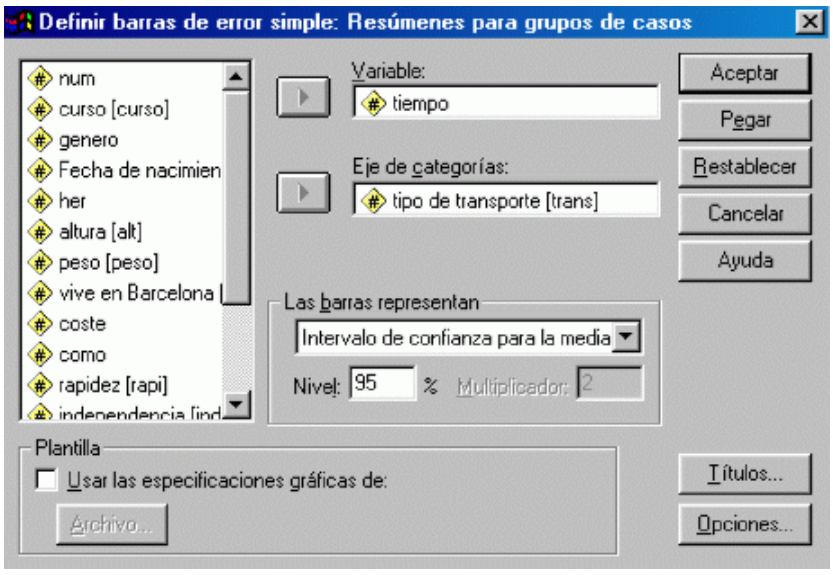
Se selecciona la variable que se considera Dependiente y la variable Factor y con el botón Opciones se activan Estadísticos Descriptivos y Homogeneidad de varianzas.



Al aceptar en el visor de resultados aparecen los siguientes cuadros:

- *Descriptivos.* Recoge la media, la desviación típica, el intervalo de confianza del 95% (por defecto) para la media correspondiente a la variable dependiente para cada uno de los grupos definidos por el factor.
- *Prueba de homogeneidad de varianzas*. Contiene el valor del estadístico de Levene del contraste de la hipótesis de homoscedasticidad con el nivel de significación crítico.
- *ANOVA*. Contiene las sumas de cuadrados inter-grupos, intra-grupos y total, sus correspondientes grados de libertad y el valor del estadístico de prueba F junto con el nivel de significación crítico.

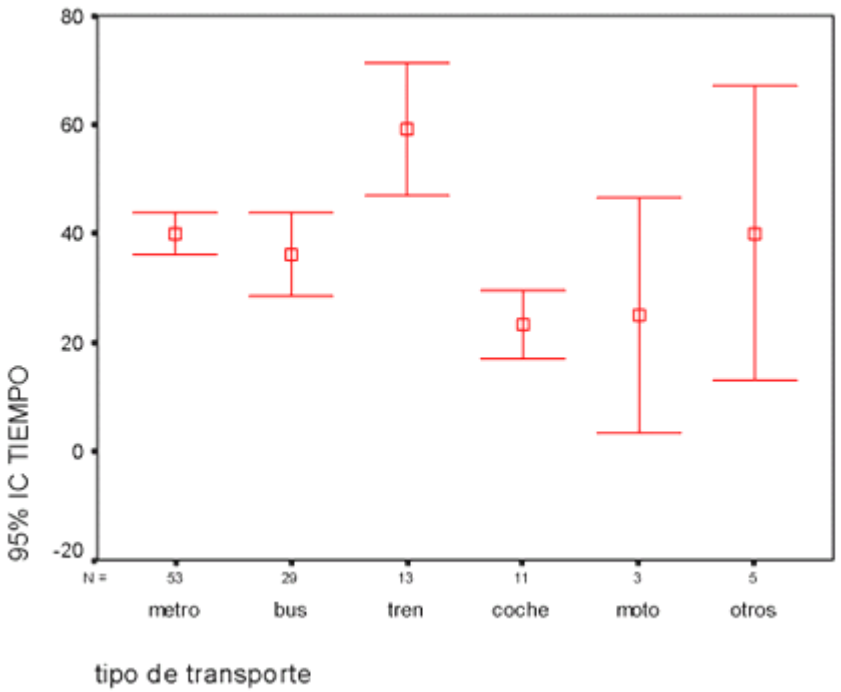
Como complemento gráfico de este análisis, para obtener una primera aproximación acerca de si es razonable o no la hipótesis nula, se selecciona *Gráficos* > *Barras de error* y se activa la opción Simple. Con el botón *Definir* se abre el siguiente cuadro de diálogo:



Se selecciona en *Variable* la variable dependiente del ANOVA y en el *Eje de categorías* la variable factor. El intervalo de confianza para la media se calcula por defecto al 95% de confianza. Al aceptar aparece en el visor de resultados los puntos que representan a la media de cada grupo junto con los límites del correspondiente intervalo de confianza para la media poblacional. Si los puntos que representan las medias están desigualmente distribuidos en el gráfico se tiene un indicio de que a nivel poblacional no puede sostenerse la hipótesis de igualdad de medias; es decir, por lo menos uno de los niveles del factor influye significativamente sobre la variable dependiente.

Con los datos de la encuesta sobre transporte, Enctrans.sav, razonar si puede aceptarse que el tipo de transporte utilizado, Trans, influye sobre la variable tiempo.

Con la opción de menú Gráficos > Barras de error > Simple y con el botón Definir se selecciona como Variable Tiempo y en Eje de categorías la variable Trans; al aceptar se obtiene la siguiente representación gráfica:



Como puede observarse, los puntos que representan a las medias de cada grupo aparecen dispersos a diferentes niveles; sobre todo la media del grupo definido por el factor Tren. El intervalo de confianza para la media correspondiente al grupo definido por el factor Metro está contenido dentro del intervalo correspondiente al grupo definido por el factor Bus, así como, el intervalo correspondiente al factor Coche está contenido dentro de los intervalos correspondientes definidos por los

factores Metro y Otros. El gráfico, por tanto, parece sugerir no una única población sino tres poblaciones con distintas medias.

Para realizar el análisis de la varianza propiamente dicho la secuencia es Analizar > Comparar medias > ANOVA de un factor. En el cuadro de diálogo se selecciona Tiempo como variable Dependiente y Trans como Factor. Para contrastar la hipótesis de igualdad de varianzas se abre con el botón correspondiente el cuadro de diálogo ANOVA de un factor: Opciones y se activa Homogeneidad de varianzas. Si se desea un análisis descriptivo del comportamiento de la variable dependiente dentro de cada grupo se activa también la opción Descriptivos. Al aceptar se obtienen los siguientes cuadros de resultados:

Descriptivos

TIEMPO

	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%	
					Límite inferior	Límite superior
metro	53	39,94	14,20	1,95	36,03	43,86
bus	29	36,17	20,18	3,75	28,50	43,85
tren	13	59,15	20,33	5,64	46,87	71,44
coche	11	23,18	9,56	2,88	16,76	29,60
moto	3	25,00	8,66	5,00	3,49	46,51
otros	5	40,00	21,79	9,75	12,94	67,06
Total	114	39,17	18,51	1,73	35,73	42,60

Este cuadro contiene un análisis descriptivo de la variable dependiente por grupos, así como, los límites superior e inferior para la media de cada grupo al 95% de confianza.

Prueba de homogeneidad de varianzas

TIEMPO

Estadístico de Levene	gl1	gl2	Sig.
1,514	5	108	,191

El estadístico de Levene toma un valor lo suficientemente pequeño para no rechazar las hipótesis de homocesdaticidad a los niveles de significación habituales.

ANOVA

TIEMPO

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	8901,537	5	1780,307	6,450	,000
Intra-grupos	29810,297	108	276,021		
Total	38711,833	113			

En el cuadro de resultados del ANOVA, el valor del estadístico de prueba, F=6,450, es significativamente distinto de 1 para cualquier nivel de significación y, por lo tanto, se rechaza la hipótesis nula de igualdad de medias y queda confirmada la primera impresión proporcionada por el gráfico de barras de error.