Estadística aplicada a las Ciencias Geológicas

Autor:
Arnaldo Mangeaud

Facultad de Ciencias Exactas, Físicas y Naturales

Universidad nacional de Córdoba

Córdoba, 2014

Índice

Capítulo 1.

Introducción

Concepto de Estadística.

Para comenzar definiremos a la Estadística como la ciencia que estudia la manera en que se recolecta, se analiza, se interpreta la información proveniente de una población, así como el modo en que se extrapola esos resultados a otros casos similares.
La estadística es una ciencia auxiliar de otras ciencias, ayuda a resolver problemas planeados por otras disciplinas en la toma de decisiones o para explicar el comportamiento de fenómenos determinados.

La Estadística es transversal a diferentes disciplinas, desde las ciencias consideradas más "duras" como la Geología, Biología o Ingeniería hasta las llamadas ciencias "blandas" como la Psicología, todas necesitan de la ayuda de esta herramienta.

La estadística se divide en dos grandes áreas:

La estadística descriptiva se dedica a la descripción y resumen de la información originada a partir de los fenómenos de estudio. Esa información puede ser resumida mediante tablas, gráficos o medidas de resumen.

La estadística inferencial, por su parte extrapola las conclusiones observadas para algunos casos a otros casos de similares características.

En la actualidad son muy escasas empresas, industrias o trabajos de investigación básica o aplicada donde no se aplique estadística. A partir de la última década del siglo pasado con la generalización de computadoras se hicieron muy accesibles cálculos que habían sido demasiado complicados de realizar anteriormente. Eso permitió que no sólo se extienda su uso sino que se desarrollaron nuevas teorías que fueron acompañadas por procedimientos desarrollados directamente en las computadoras y que manualmente son inaccesibles.

Población. Muestra. Elementos.

La Estadística es una ciencia y como tal posee un objeto de estudio. El objeto de estudio de la estadística son las poblaciones.
La definición estadística de población es: conjunto de cosas, o conjunto de elementos, o bien conjunto de unidades. Cada unidad, elemento o cosa será definido por la persona a cargo del trabajo (investigador u operador) y dependerá de los objetivos e hipótesis del trabajo.
Recordemos que definimos como objetivo a aquella meta a la cual se pretende arribar y como hipótesis a una conjetura, a una suposición realizada de cómo "funciona el sistema en estudio". Entonces a partir de objetivos e hipótesis el operador decidirá quién es el elemento al que le tomará la información. El conjunto de todos estos elementos (obviamente definidos en un espacio y tiempo) serán la población en estudio.

A modo de ejemplo, en un estudio de prospección se toman porciones de rocas de un cuerpo geológico. Cada porción es un elemento, el conjunto de todos ellos, es decir el cuerpo geológico es la población.
Cuando accedemos a la información a partir de todos los elementos de la población estamos realizando un censo. Realizar un censo es prácticamente imposible. Sería demasiado costoso en términos económicos o logísticos que se pueda contar con la totalidad de la información. Ante esta imposibilidad, se debe recurrir a la información de algunos de los elementos. Esos elementos a su vez deben representar a la población, por lo que, cuando extraigamos una conclusión para esa porción de la población vamos a estar extrapolando esa información a toda la población.
De este modo estamos definiendo el concepto de muestra: Una muestra es una porción representativa de la población. Y decimos que es representativa para poder inferir sus resultados al resto de la población.

Ejemplo 1.1. En un estudio en la línea de costa se toman 120 porciones de 1 m^2 del fondo rocoso de una playa. Cada elemento es cada cuadrata de 1 m^2, la muestra es el conjunto de 120 cuadratas y la población van a ser todas las posibles cuadratas que componen la playa.

El modo de pensar en que la muestra sea representativa de la población es pensar que cada uno de los elementos de la población debieran tener igual chance de formar parte de la muestra (es decir de ser escogidos en ella). Si por alguna razón algunos de los elementos no pueden formar parte de la muestra, entonces sus características propias no estarán representadas.

Ejemplo 1.2. En el estudio del Ejemplo 1.1 por una cuestión de comodidad o de ineptitud se toman sólo porciones de los 10 primeros metros de la playa, excluyendo a las zonas más cercanas al mar. Esas 120 porciones de 1 m^2 no representan el fondo rocoso de la playa. Sólo representan el fondo rocoso de los 10 primeros metros de playa.

Hay diferentes modos de realizar muestreos. Estos dependen del área de aplicación y de quién es el elemento o unidad. Por ejemplo en estudios de mineralogía una porción de roca de 20cm^3 es el elemento donde el conjunto de éstos forman un cuerpo ígneo (la población). En estudio de geoquímica de aguas, una alícuota de 200ml de agua es el elemento y el conjunto de éstos forman el río en este momento (población).

Hay diversas proposiciones de la clasificación de los tipos de muestreo, aunque en general pueden dividirse en dos grandes grupos: métodos de muestreo probabilísticos y métodos de muestreo no probabilísticos.
A continuación se describen los distintos tipos para estos dos grandes grupos, si bien en Geología, para los estudios que se llevan a cabo, se trabajarán con muestreos probabilísticos.

<u>Métodos de muestreo probabilísticos</u>
Los métodos de muestreo probabilísticos son aquellos que se basan en el principio de igualdad de chances (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 elegidas. Sólo estos métodos de muestreo probabilísticos nos aseguran la representatividad de la muestra extraída y son, por tanto, los más recomendables. Dentro de los métodos de muestreo probabilísticos encontramos los siguientes tipos:

Muestreo aleatorio simple: El procedimiento empleado es el siguiente: Se le asigna un número a cada elemento de la población y a través de algún medio mecánico (bolillas dentro de una bolsa, tablas de números aleatorios, números aleatorios generados con una calculadora o computadora, etc.) se eligen tantos elementos como sea necesario para completar el tamaño de la muestra requerido. Este procedimiento, atractivo por su simpleza, tiene poca o nula utilidad práctica cuando la población que estamos manejando es muy grande o cuando es imposible enumerar a todos los elementos de la población. Los investigadores a campo suelen seleccionar un valor al azar que distinga latitud y otro que distinga longitud para cada punto.

Muestreo aleatorio 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 ocupan 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 sobre 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 de 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. A su vez son muy cómodos en control de calidad de una cinta de producción. Es sencillo pensar que cada 100 productos que salen se tomará uno.

Muestreo aleatorio estratificado: Un estrato se define como un conjunto de unidades que son homogéneas entre sí y son heterogéneas con las unidades de otros estratos. Este tipo de muestreo consiste en definir previamente los estratos que posee una población. A partir de eso se deben tomar muestras aleatorias en cada estrato. 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 muestrales. (Por ejemplo 30 elementos de cada estrato)
Afijación Proporcional: La distribución se hace de acuerdo con el peso (tamaño) de la población en cada estrato (Por ejemplo si un estrato posee el 20% de la población, contribuirá con el 20% de la muestra).
Afijación Óptima: Se tiene en cuenta la dispersión de los resultados. Como veremos más adelante, a mayor variabilidad en la población, se deben tomar mayor número de elementos de ella para que sea representativa la muestra. Una afijación es un concepto teórico interesante, pero tiene poca aplicación ya que no se suele conocerse la variabilidad en la población.

Muestreo aleatorio por conglomerados: En el muestreo por conglomerados la unidad muestral es un grupo de elementos de la población que forman una unidad llamada conglomerado. Las unidades hospitalarias, los departamentos universitarios, una caja de determinado producto 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) y en investigar después todos los elementos pertenecientes a los conglomerados elegidos.

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,** pues no se tiene certeza de que la muestra extraída sea representativa, ya que no todos los elementos de la población tienen la misma probabilidad de ser elegidos. En general se seleccionan a los personas siguiendo determinados criterios procurando que la muestra sea representativa. Los resultados que se obtienen son descripciones de esa muestra. Este tipo de muestreo es muy utilizado en estudios sociales.

Estudios de caso: A los fines descriptivos se toma un grupo de unidades o elementos con ciertas características. El investigador sabe que la población está formada por un grupo grande de unidades y las unidades que tomó no las representan. Es muy común el estudio de todos los pacientes de un hospital. Los alumnos de un curso en particular o un colegio.

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: 30 personas de 30 a 40 años, de sexo femenino y residentes en una determinada región. 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: 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.

Muestreo casual o incidental: Se trata de un proceso en el que el investigador selecciona 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 frecuencia a sus propios alumnos). Un caso particular es el de los voluntarios.

Bola de nieve: Se localiza a algunos individuos, los cuales conducen a otros, y estos a otros, y así hasta conseguir una muestra suficiente. Este tipo se emplea muy frecuentemente

cuando se hacen estudios con poblaciones "marginales", delincuentes, sectas, determinados tipos de enfermos, etc.

Ya hemos visto el modo en que vamos a acceder a la muestra para que ésta sea representativa de la población, ahora la pregunta es: ¿qué información vamos a tomar de los elementos o unidades?
Los elementos tienen muchísimas características que pueden ser tomadas y/o medidas y volvemos a objetivos e hipótesis como los estandartes que van a llevar al operador a tomar sólo las características importantes. A su vez aquellas características que son constantes en la población no serán tomadas, siendo claro que se van a tomar características que son variables.

Variables. Tipos

Las variables son entonces características que pueden variar de elemento en elemento. Existen varias clasificaciones de las variables. Una de ellas combina clasificación y formas de medición quedando entonces:

Variables cualitativas
Las variables cualitativas expresan una cualidad, se representa por una letra, no expresa una cantidad. Pueden ser de dos tipos:

a) **Variables nominales**: Son aquellas que expresan el atributo como tal, no expresan ni orden ni escala de medición.

Ejemplo 1.3. Tipo de roca presente: cuarzo, feldespato, mica.

b) **Variables ordinales**: Son aquellas que expresan el atributo en un ranking u orden. Aunque se las exprese en un número este no es tal.

Ejemplo 1.4. Escala de dureza de Mohs: (1) Talco, (2) Yeso, (3) Calcita,…., (9) Corindón, (10) Diamante.

Variables cuantitativas
Expresan una cantidad, un número, una métrica o medida. Pueden ser de dos tipos:

a) **Variables discretas:** Está representada por números enteros.

Ejemplo 1.5. Número de rocas de cuarzo de más de 1 cm de diámetro en una superficie de $30cm^2$.

b) **Variables continuas:** Posee infinitos valores entre los enteros.

Ejemplo 1.6. Presión a la cual se fractura un bloque de 20 cm^3 de una roca.

La naturaleza de la variable es la que marca a quéel tipo de variable ésta pertenece. Como se ve, las variables cualitativas nominales son sólo un nombre, sin orden preestablecido, mientras que una variable cualitativa ordinal puede ser una letra pero entre estas letras ya hay un orden establecido. Las variables discretas poseen una métrica, una medida verdadera, el salto de 1 a 2 es el mismo que de 4 a 5. A su vez las variables continuas poseen infinitos decimales entre los enteros, las diferencias entre un valor y otro pueden ser mucho más pequeñas, más sutiles y precisas.
El operador puede decidir hacerle perder esta "jerarquía" a la variable, por cuestiones de practicidad, pero no puede inventarle jerarquía. Esto quiere decir que se puede medir como grande, mediano o pequeño a algo que es una variable cuantitativa, pero en una encuesta a personas de una fábrica no existe razón alguna para colocarle el valor 1 si es mujer y 2 si es varón.

Debemos advertir que se suele incurrir en una discrepancia de términos entre la Geología y la Estadística. Para los estadísticos la muestra es el conjunto de elementos. Para los geólogos la muestra es cada uno de los elementos. Entonces debemos ser cautos en las definiciones para no confundir los términos. Por otra parte, a los fines coloquiales en este apunte otorgaremos la definición de dato a aquel valor de la variable, ya sea cuali como cuantitativa. Por ejemplo un dato sería: 32,25 cm; 8 años, Verde, Mujer, 42 grados centígrados, etc.

Capítulo 2.

Análisis descriptivo de una variable

Una vez que se ha tomado la decisión sobre cuál será la población estadística y la muestra a analizar, una vez que se toman o miden las variables en cada elemento, deberá registrarse por escrito esta información. Después de eso el operador puede además trasladar esos datos a una base de datos de modo informático. Lo que nunca debe hacerse es destruir la base de datos en formato papel. En algún libro alguna vez se leyó que es más duradera la más suave de las tintas sobre papel que la mejor de las memorias y esto incluye tanto a memoria neuronal como informática. El papel es siempre el mejor plan B ante la ausencia de un valor, la confirmación de un dato anómalo o un posible error de tipeo.
Ahora, ya sea en papel o en la computadora, se tiene un listado de letras o números, dependiendo si la variable es cualitativa o cuantitativa y debiéramos entonces resumir esa información. El modo en que la resumimos es mediante tablas, gráficos o medidas de resumen.

Tablas de distribución de frecuencia

Variables cualitativas nominales

Se tiene una variable cualitativa nominal. Por ejemplo se tiran 20 agujas al azar en el fondo de un río serrano y se registra la roca en la cual toca la aguja. Los resultados son:

Cuarzo, Cuarzo, Feldespato, Mica, Mica, ..., Cuarzo.
Entonces se resume esa información en la siguiente tabla:

Variable (X)	n_i (FA)	h_i (FR)
Cuarzo	11	0,55
Feldespato	6	0,3
Mica	3	0,15

Tabla 2.1: Distribución de frecuencia del tipo de roca.

Donde: a la variable tipo de roca la denominamos X
n_i : Frecuencia Absoluta: número de elementos que poseen un valor determinado de la variable. En la tabla 2.1: 6 agujas tocaron a Feldespato.
h_i : Frecuencia Relativa: proporción de elementos que poseen un valor determinado de la variable. En la tabla 2.1: una proporción de 0,3 tocaron a Feldespato (dicho de otro modo, el 30,00% del fondo del lecho del río sería Feldespato).

Los gráficos correspondientes a esta tabla son gráficos de tortas y de barras (no confundir con histograma).

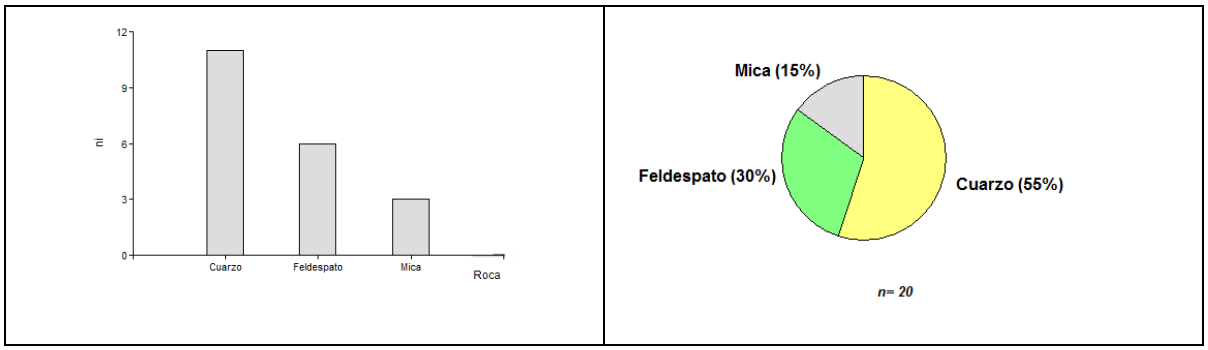


Figura 2.1 Gráfico de barras y de torta de una variable cualitativa

Variables cuantitativas

Si ocurriera que la información que poseemos es escasa, con pocos números diferentes, ni vale la pena agruparla. Entonces se está en presencia de una Distribución simple o de tipo 1. Por ejemplo:

Se estudia la ley de oro en una mina, se tiene una muestra de 5 porciones de roca y los valores son (en g/tn):
12; 17; 11; 13; 08.

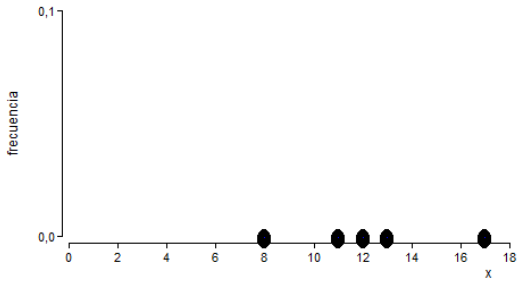


Figura 2.2. Representación gráfica de una distribución de tipo 1

Si se tienen más elementos tomados en la muestra, vale la pena el resumirlos en una tabla.

Distribución de tipo 2

Se tiene una variable cuantitativa discreta con pocos valores diferentes. Por ejemplo en una mina se le pregunta a cada empleado de una muestra de 130, el número de hijos que posee. Los resultados son:

1; 2, 0; 5; 7; 3; 1; 1; 3;…; 4.

Entonces ahora sí podemos resumir en una tabla donde en cada fila colocaremos el valor de la variable del siguiente modo:

Variable (X)	n_i (FA)	h_i (FR)	$N_i \downarrow$ (FAAa)	$F_i \downarrow$ (FRAa)	$N_i \uparrow$ (FAAd)	$F_i \uparrow$ (FRAd)
0	20	0,1538	20	0,1538	130	1
1	31	0,2385	51	0,3923	110	0,8462
2	36	0,2769	87	0,6692	79	0,6077
3	19	0,1462	106	0,8154	43	0,3308
4	11	0,0846	117	0,9000	24	0,1846
5	9	0,0692	126	0,9692	13	0,1000
6	3	0,0231	129	0,9923	4	0,0308
7	1	0,0077	130	1	1	0,0077

Tabla 2.2: Tabla de distribución de frecuencias de la variable número de hijos

Donde: X es la variable número de hijos.
n_i : Frecuencia Absoluta: número de elementos que poseen un valor determinado de la variable. En la tabla 2.2: 31 personas poseen 1 hijo
h_i : Frecuencia Relativa: proporción de elementos que poseen un valor determinado de la variable. En la tabla 2.2: una proporción de 0,2385 poseen 1 hijo (dicho de otro modo, el 23,85% poseen un hijo).
$N_i \downarrow$ Frecuencia Absoluta Acumulada ascendente: número de elementos acumulados que poseen un valor determinado de la variable y sus valores inferiores. En la tabla 2.2: 51 personas poseen 1 hijo o menos.
$F_i \downarrow$ Frecuencia RelativaAcumulada ascendente: proporciónde elementos acumulados que poseen un valor determinado de la variable y sus valores inferiores. En la tabla 2.2: una proporción de 0,3923 poseen 1 hijo o menos.
$N_i \uparrow$ Frecuencia Absoluta Acumulada descendente:número de elementos acumulados que poseen un valor determinado de la variable y sus valores superiores. En la tabla 2.2: 110 personas poseen 1 hijo o más.
$F_i \uparrow$ Frecuencia Relativa Acumulada descendente: proporciónde elementos acumulados que poseen un valor determinado de la variable y sus valores superiores. En la tabla 2.2: una proporción de 0,8426 poseen 1 hijo o más.

Los gráficos que corresponden a esta distribución son los que se presentan en la Figura 2.2.

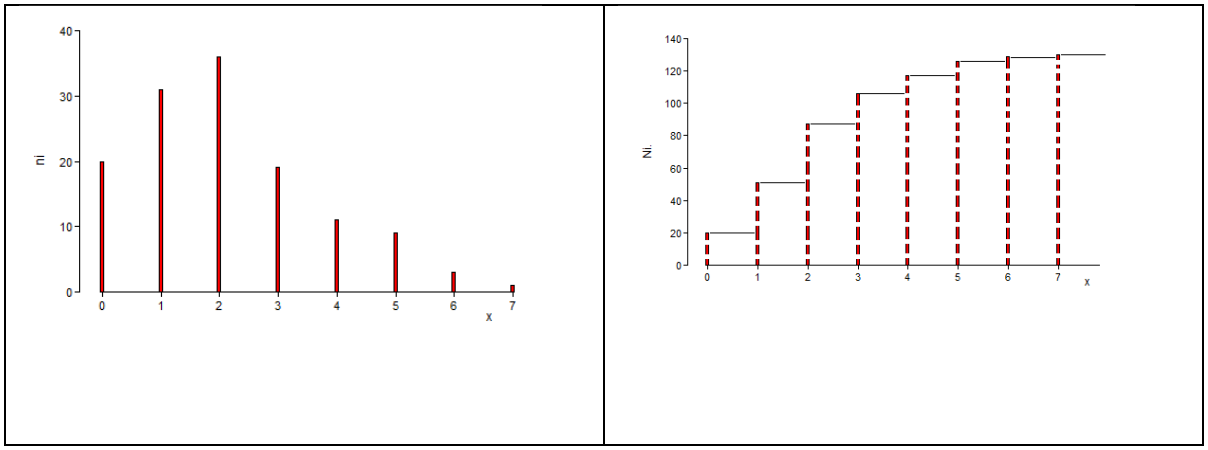


Figura 2.3.Representación gráfica de una distribución de tipo 2. Se presenta un gráfico de bastones y de escalones (frecuencia acumulada)

Distribución de tipo 3

Se tiene una variable cuantitativa continua o una cuantitativa discreta con muchos valores diferentes. Por ejemplo se tomaron 68 porciones de roca para estudiar la distribución del cobre en una mina. Los resultados (en g/tn) son:

423,07; 452,23;.....; 561,34.

Como se puede ver es imposible disponer una tabla en la forma de la tabla 2.2, ya que si una variable es continua es poco probable tener dos valores iguales de la variable. De tal modo, quedaría una tabla con 68 filas donde cada fila poseerá una frecuencia absoluta de 1 (uno) para resumir 68 valores. Por lo tanto es necesario dividir a la variable en distintos intervalos de clase. Cada intervalo a su vez posee una amplitud determinada.
No hay un número de intervalos exactos, pero una aproximación a éste fue propuesta por Sturges en 1936 que se obtiene mediante la fórmula:

k= 1+3,33 Log_{10}(n),
Donde k es el número óptimo de intervalos.

Variable X x_{i-1} - x_{i+1}		x_i MC	n_i (FA)	h_i (FR)	$N_i \downarrow$ (FAAa)	$F_i \downarrow$ (FRAa)	$N_i \uparrow$ (FAAd)	$F_i \uparrow$ (FRAd)
423,07	448,15	435,61	2	0,0294	2	0,0294	68	1
448,15	473,23	460,69	11	0,1618	13	0,1912	66	0,9706
473,23	498,30	485,77	18	0,2647	31	0,4559	55	0,8088
498,30	523,38	510,84	17	0,25	48	0,7059	37	0,5441
523,38	548,46	535,92	15	0,2206	63	0,9265	20	0,2941
548,46	573,54	561,00	5	0,0735	68	1	5	0,0735

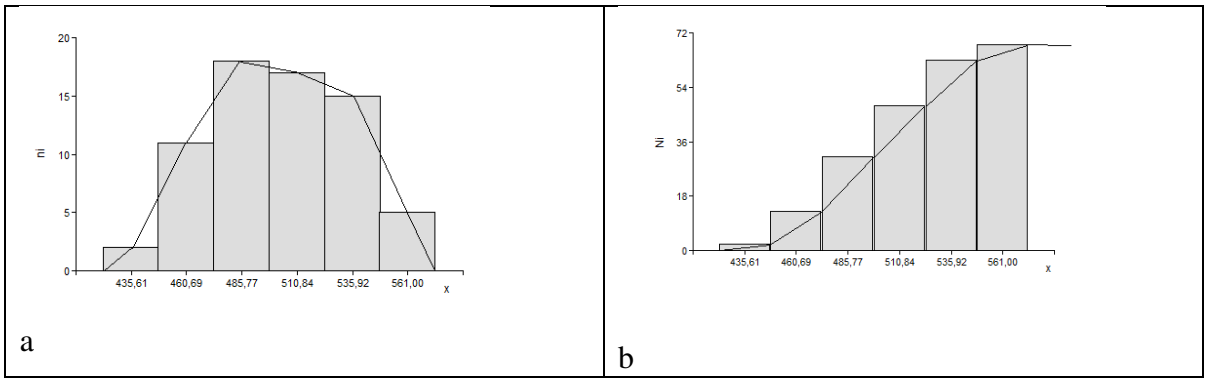
Tabla 2.3. Tabla de distribución de frecuencias de la variable ley de Cobre

Donde:

X: variable ley de Cobre (en g/tn)
MC: Marca de clase (se obtiene promediando los extremos del intervalo)
n_i : Frecuencia Absoluta: número de elementos que se encuentran en un determinado intervalo de valores de la variable. En la tabla 2.3: 11 porciones de roca poseen entre 448,15 y 473,23 g/tn.
h_i : Frecuencia Relativa: proporción de elementos que se encuentran en un determinado intervalo de valores de la variable. En la tabla 2.3: una proporción de 0,1618 poseen entre 448,15 y 473,23 g/tn (dicho de otro modo, el 16,18%).
$N_i \downarrow$ Frecuencia Absoluta Acumulada ascendente: número de elementos acumulados que poseen un valor determinado de la variable y sus valores inferiores. En la tabla 2.3: 13 porciones de roca poseen menos de 473,23 g/tn.

$F_i \downarrow$ Frecuencia Relativa Acumulada ascendente: proporción de elementos acumulados que poseen un valor determinado de la variable y sus valores inferiores. En la tabla 2.3: una proporción 0,1912, poseen menos de 473,23 g/tn.
$N_i \uparrow$ Frecuencia Absoluta Acumulada descendente:número de elementos acumulados que poseen un valor determinado de la variable y sus valores superiores. En la tabla 2.3: 66 porciones de roca poseen más de 448,15 g/tn.
$F_i \uparrow$ Frecuencia Relativa Acumulada descendente: proporciónde elementos acumulados que poseen un valor determinado de la variable y sus valores superiores. En la tabla 2.3: una proporción de 0,9706 poseen más de 448,15 g/tn.

A partir de la tabla 2.3 se pueden construir dos tipos de gráficos: histogramas de frecuencias y su respectivo gráfico acumulado.



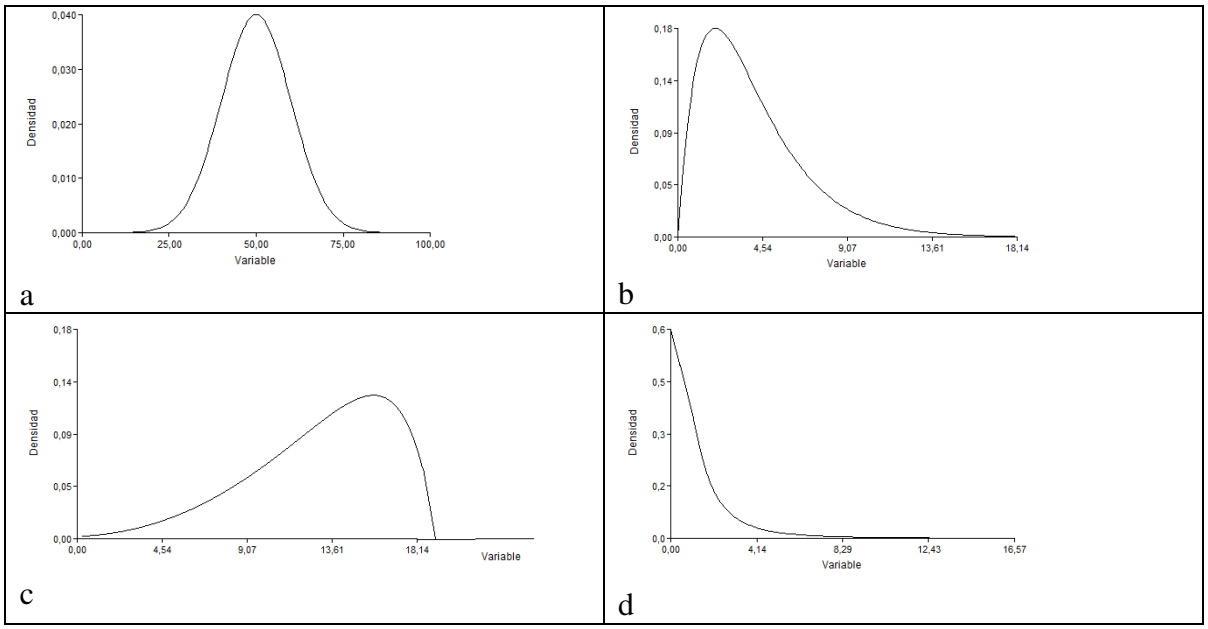
a b

Figura 2.4. Representación gráfica de una distribución de tipo 3. Se presenta un histograma de frecuencias y de frecuencias acumuladas.

Se observa en la Figura 2.4a la presencia de un polígono de frecuencias que pasa por el valor central de cada intervalo y en la Figura 2,4b una ojiva de frecuencias que pasa por el mayor valor del intervalo.

Formas de la distribución.

Es muy común observar gráficos de distribución donde sólo se grafican el eje x y una línea suavizada del polígono de frecuencias. De ese modo podremos observar la forma de la distribución de la variable. Esas formas pueden ser simétricas, asimétricas a la derecha, a la izquierda, entre otros casos. (Figuras 2.5).



Figuras 2.5. Formas de diferentes distribuciones. a: simétrica, b: asimétrica a la derecha, c: asimétrica a la izquierda, d: en forma de J invertida.

Medidas de resumen

Antes que explicar las medidas de resumen de la información, debemos advertir que en el Capítulo 8 se verán las diferencias que hay entre la información tomada de una población (exhaustiva) y la información tomada de una muestra (parcial). Es claro que la información de la población es total y completa, mientras que lo que tomamos en una muestra es incompleta y depende de cuáles fueron los elementos que ingresaron en la muestra. Por esta razón es que vamos a diferenciar algunas medidas de resumen dependiendo si han sido tomadas de una población o de una muestra.

Medidas de posición

Media, media aritmética o promedio.

Su definición es exclusivamente algebraica: es la sumatoria de los valores de la variable, dividido el número de elementos de la muestra o población.

Poblacional

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

Donde:
μ: Media poblacional
x_i: Valores de la variable
N: número de elementos de la población

Muestral

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

Donde:
$\bar{x}$: Media muestral
x_i: Valores de la variable
n: número de elementos de la muestra

Obtención de la media según el tipo de distribuciones.

Distribución de tipo 1:

Volviendo al ejemplo sobre los 5 datos de ley de oro. 12; 17; 11; 13; 08.
Veremos que:

$$\bar{x} = \frac{12 + 17 + 11 + 13 + 8}{5} = \frac{61}{5} = 12{,}2$$

Distribución de tipo 2:
Volviendo al ejemplo descripto en la tabla 2.2, el promedio de hijos por empleado se calcula:

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

$$\bar{x} = \frac{(0 * 20) + (1 * 31) + (2 * 36) + (3 * 19) + (4 * 11) + (5 * 9) + (6 * 3) + (7 * 1)}{130} =$$

$$\bar{x} = \frac{274}{130} = 2{,}1077$$

Distribución de tipo 3:
Volviendo al ejemplo descripto en la tabla 2.3, el promedio de Cobre es de:

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

$$\bar{x} = \frac{(435{,}61 * 2) + (460{,}69 * 11) + (485{,}77 * 18) + (510{,}84 * 17) + (535{,}92 * 15) + (561 * 5)}{68} =$$

$$\bar{x} = \frac{34210{,}75}{68} = 503{,}0993$$

Mediana
Es el valor de la variable que es superado por el 50% de los valores más grandes de la variable y supera al 50% más pequeño. Dicho de otro modo, es el valor de la variable que está (de modo ordenado) en el centro de la distribución. La fórmula que se presentará a continuación es la fórmula para hallar la posición de la mediana y no es la fórmula de la Mediana.

$Pos\ (Me) = \frac{n+1}{2}$,
Donde Pos (Me) es la posición de la mediana y n: número de elementos de la Muestra.
Si se posee un número impar de datos, el valor de la Mediana corresponde a la posición correspondiente, si es un número par, se promedian los dos valores centrales.

Obtención de la mediana según el tipo de distribuciones.

Distribución de tipo 1:
Recurriendo al ejemplo sobre los 5 datos de ley de oro. 12; 17; 11; 13; 08.
Veremos que ordenados ellos son: 08; 11; 12; 13; 17.
Si n= 5, entonces Pos (Me)= 3, por lo que la Mediana= 12

Distribución de tipo 2:
Recurriendo al ejemplo descripto en la tabla 2.2, la Mediana de hijos por empleado se obtiene buscando al valor de la variable donde está el valor que ordenado es (n+1)/2. Como n= 130, la posición sería 131/2= 65,5. En la tabla se observa que el valor acumulado 65,5 corresponde a 2 hijos

Distribución de tipo 3:
Con el ejemplo descripto en la tabla 2.3, la Mediana de Cobre se encuentra en el intervalo que acumula al valor que ordenado está en la posición (68+1)/2= 34,5. Entonces la mediana está en el intervalo entre 498,30 y 523,38. Una manera de encontrar un valor estimado en ese intervalo es mediante la siguiente fórmula:

$$Me\ (x) = x_{j-1} + C_j * \frac{\frac{n}{2} - N_{j-1}}{N_j - N_{j-1}}$$

Donde:
x_{j-1} : Valor inferior de la variable en el intervalo determinado como que posee a la Me (x)
C_j : Amplitud del intervalo
N_j : Frecuencia absoluta acumulada del intervalo determinado como que posee a la Me (x)
N_{j-1}: Frecuencia absoluta acumulada del intervalo anterior al determinado como que posee a la Me (x)

Esto se apoya con el modo gráfico de obtener la mediana (Figura 2.5)

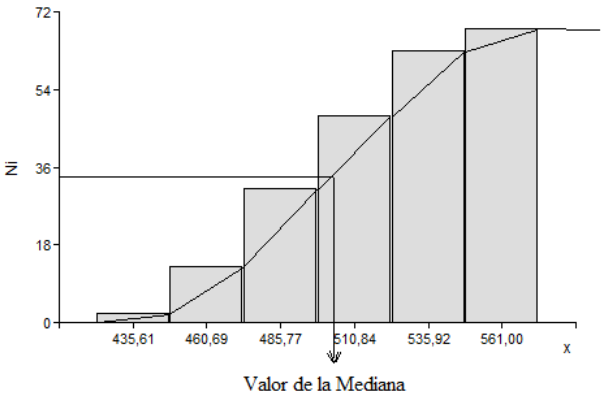


Figura 2.6. Método gráfico de obtención de la Mediana en distribuciones de tipo 3.

Cuartilos: Son valores de la variable que superan a cierto porcentajes de la variable.
Cuartilo 1: Q_1: Supera al 25% más pequeño de la variable y es superado por el 75% más grande.

Cuartilo 2: Q_2=Me: Supera al 50% más pequeño de la variable y es superado por el 50% más grande.
Cuartilo 1: Q_3: Supera al 75% más pequeño de la variable y es superado por el 25% más grande.

Percentiles: Son valores de la variable que superan a un porcentaje en particular según el interés del operador, por ejemplo:
Percentil 5: P_{05}: Supera al 5 % más pequeño y es superado por el 95% más grande.
Percentil 99: P_{99}: Supera al 99% más pequeño y es superado por el 1% más grande.

Modo o moda
Es el valor de la variable más frecuente, el valor que posee mayor frecuencia absoluta o relativa (n_i ó h_i).

Obtención del modo según el tipo de distribuciones.

No se puede obtener el Modo en distribuciones de tipo 1, ya que ninguno de los valores es el que más se repite.

Distribución de tipo 2:
Recurriendo al ejemplo descripto en la tabla 2.2, el Modo de hijos por empleado se obtiene buscando al valor de la variable con mayor frecuencia. Esto es que con una frecuencia de 36 se encuentra el valor 2. Por lo tanto la moda es 2 hijos.

Distribución de tipo 3:
Con el ejemplo descripto en la tabla 2.3, la Moda de Cobre se encuentra en el intervalo que acumula al n_i= 18. Entonces la Moda está en el intervalo entre 473,23 y 498,30. Una manera de encontrar un valor estimado en ese intervalo es mediante la siguiente fórmula:

$$Mo\ (x) = x_{j-1} + C_j * \frac{n_j - n_{j-1}}{(n_j - n_{j-1}) + (n_j - n_{j+1})}$$

Donde:
x_{j-1} : Valor inferior de la variable en el intervalo determinado como que posee a la Mo(x)
C_j : Amplitud del intervalo
n_j : Frecuencia absoluta del intervalo determinado como que posee a la Mo(x)
n_{j-1}: Frecuencia absoluta del intervalo anterior al determinado como que posee a la Mo(x)
n_{j+1}: Frecuencia absoluta del intervalo posterior al determinado como que posee a la Mo(x)

Esto se apoya con el modo gráfico de obtener la Moda (Figura 2.7)

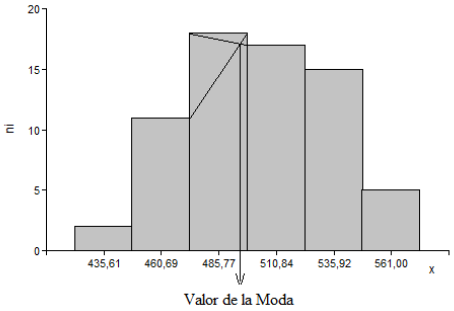
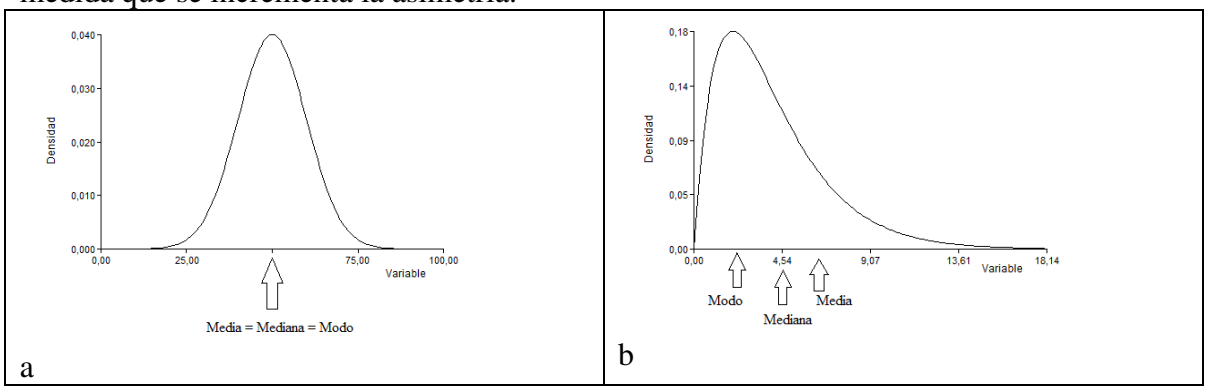


Figura 2.7. Método gráfico de obtención de la Moda en distribuciones de tipo 3.

Existen otras medidas de posición como Media geométrica, Media armónica, entre otras que poseen otras propiedades y las cuales escapan a los objetivos de este apunte.

Relaciones entre las distintas medidas de posición.

En variables cuya distribución es simétrica se observa que los valores de Media, Mediana y Modo coinciden. Mientras que en las distribuciones asimétricas se van distanciando a medida que se incrementa la asimetría.



Figuras 2.8.a: Distribución simétrica donde coinciden Media, Mediana y Modo. b: Distribución asimétrica a la derecha, donde Media > Mediana> Modo.

El rango de valores de las medidas de posición es de menos infinito a infinito ó bien depende del rango de valores de cada variable en particular. La unidad en que se expresan las medidas de posición corresponde a la misma unidad de la variable. Dicho de otro modo si la variable es hijos, la Media, Mediana y Modo se expresarán en hijos y si es g de Cu, será g de Cu.

Medidas de dispersión.

Rango o recorrido.
Es la diferencia entre el mayor y el menor valor de la variable, es decir, que es la distancia que recorre la variable desde el valor más pequeño al más grande:

$$R_x = x_{max} - x_{min}$$

Donde:
R_x = Rango o recorrido
x_{max} = mayor valor de la variable
x_{min} = menor valor de la variable

Recorrido intercuartílico
Es la diferencia entre el Cuartilo 3 y el Cuartilo 1 de la variable. Es decir es la distancia en la que se encuentran el 50% central de los valores de loa variable

$$R_{IQx} = Q_3 - Q_1$$

Donde:
R_{IQx} =Recorrido intercuartílico
Q_1 = Cuartilo 1
Q_3 =Cuartilo 3

Desviación cuartílica
Es el promedio de las distancias entre los Cuartilos 1 y 3 con la Mediana.

$$D_{Qx} = \frac{Q_3 - Q_1}{2}$$

Donde:
D_{Qx} = Desviación cuartílica
Q_1 = Cuartilo 1
Q_3 =Cuartilo 3

Varianza.

Al igual que la Media la definición de Varianza es algebraica: es la sumatoria de los cuadrados de las distancias de cada valor de la variable con respecto a la media, dividido el número de elementos de la muestra o población. Dicho de otro modo es el promedio de los cuadrados de las distancias de los datos con respecto a su media.

Varianzas

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

Donde:
σ^2: Varianza poblacional
μ: Media poblacional
x_i:Valores de la variable
N: número de elementos de la población

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

Donde:
S^2: Varianza muestral
$\bar{x}$: Media muestral
x_i: Valores de la variable
n: número de elementos de la muestra

Desviación estándar o Desvío estándar

El desvío estándar no posee una fórmula propia, sino que es la raíz cuadrada de la Varianza.

$\sigma = \sqrt[2]{\sigma^2}$ $S = \sqrt[2]{S^2}$

Fórmulas de trabajo
Para hacer más sencillo el cálculo manual de estas dos fórmulas se puede trabajar con las siguientes fórmulas:

$\sigma^2 = \frac{1}{N}\sum(x-\mu)^2$	$S^2 = \frac{1}{n-1}\sum(x-\bar{x})^2$
$\sigma^2 = \frac{1}{N}\sum(x^2 - 2x\mu + \mu^2)$	$S^2 = \frac{1}{n-1}\sum(x^2 - 2x\bar{x} + \bar{x}^2)$
$\sigma^2 = \frac{1}{N}(\sum x^2 - \sum 2x\mu + \sum \mu^2)$	$S^2 = \frac{1}{n-1}(\sum x^2 - \sum 2x\bar{x} + \sum \bar{x}^2)$
$\sigma^2 = \frac{1}{N}(\sum x^2 - 2\mu \sum x + N\,\mu^2)$	$S^2 = \frac{1}{n-1}(\sum x^2 - 2\bar{x} \sum x + n\,\bar{x}^2)$
$\sigma^2 = \frac{1}{N}(\sum x^2 - 2\mu N\mu + N\mu^2)$	$S^2 = \frac{1}{n-1}(\sum x^2 - 2\bar{x} n\bar{x} + n\bar{x}^2)$
$\sigma^2 = \frac{1}{N}(\sum x^2 - 2N\mu^2 + N\mu^2)$	$S^2 = \frac{1}{n-1}(\sum x^2 - 2n\bar{x}^2 + n\bar{x}^2)$
$\sigma^2 = \frac{1}{N}(\sum x^2 - N\mu^2)$	$S^2 = \frac{1}{n-1}(\sum x^2 - n\bar{x}^2)$

De este modo se observa que:

$\sigma^2 = \frac{\sum_{i=1}^{N} x_i^2}{N} - \mu^2$ $S^2 = \frac{\sum_{i=1}^{n} x_i^2}{n-1} - \frac{n\bar{x}^2}{n-1}$

Coeficiente de variación.
Para realizar comparaciones entre variables con diferentes métricas o escalas se utiliza el Coeficiente de variación. Este es una medida adimensional (se simplifican las unidades) y se lo suele multiplicar x 100 para expresarlo en porcentajes. Expresa cuán grande es el desvío estándar con respecto a la media.

$CV_x = \frac{\sigma}{\mu}$ $CV_x = \frac{S}{\bar{x}}$

Donde: Donde:

CV_x:Coeficiente de variación
σ : Desvío poblacional
μ: Media poblacional

CV_x: Coeficiente de variación
S : Desvío muestral
$\bar{x}$: Media muestral

El rango de valores de las medidas de dispersión descriptas es de 0 a ∞.
La varianza posee como unidad al cuadrado de la unidad de la variable, mientras que el Coeficiente de variación no posee unidad, en todos los casos restantes que se han visto, la unidad en la que se expresa cada medida de dispersión es la misma de su variable.

Medidas de Asimetría

Con la finalidad de darle un valor a la forma de las diferentes distribuciones se han propuestos distintas fórmulas. Estas intentan cuantificar la forma que se produce la asimetría (a la derecha como positiva y a la izquierda como negativa). Si es simétrica la distribución su coeficiente será cero.

Asimetría de Pearson:

$$A_{p1} = \frac{\bar{x} - Mo}{s}$$

$$A_{p2} = \frac{3(\bar{x} - Me)}{s}$$

Asimetría de Fisher

$$A_F = \frac{\mu_3}{s^3} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3}$$

Medidas de Curtosis (Apuntalamiento)
Otorgan un valor numérico a la puntiagudez. Se dice que es mesocúrtica una distribución con valor cero, mientras que es platicúrtica una distribución "aplastada" y leptocúrtica una distribución puntiaguda.

$$K_F = \frac{\mu_4}{s^4} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{s^4} - 3$$

Momentos de orden r
Como se puede ver en las fórmulas de Media, Varianza, Simetría (de Fisher) y Curtosis se está ante fórmulas muy similares que cambian sólo a qué valor están elevadas.

Entonces definimos momento de orden r a:

$$\overline{Mr} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^r}{n-1}$$

Y como se ha visto:
El momento de orden 1 corresponde a la Media, el de orden 2 a la Varianza, el de orden 3 a la Asimetría y el de orden 4 a la Curtosis.

Sólo a los fines de terminar el capítulo presentaremos un gráfico denominado Gráfico de cajas, donde se puede observar Media, Mediana, Cuartilos y Percentiles. A su vez se observa la distancia del recorrido intercuartílico, si la variable es o no simétrica y si se presentan datos anómalos o extremos (Figura 2.9).

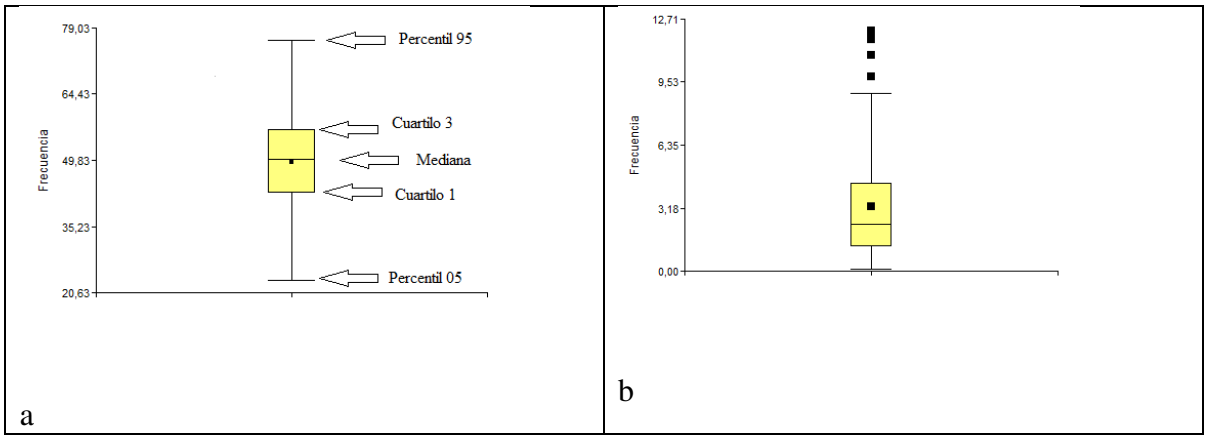


Figura 2.9. Gráfico de cajas de una distribución simétrica (a) y asimétrica (b). Se observan Percentiles, Cuartilos y Mediana, así como el punto representa Media y datos extremos.

Capítulo 3.

Análisis descriptivo de dos variables conjuntas.

Introducción.

Muchas veces ocurre que el investigador está interesado en estudiar no sólo las variables por separado sino que además cómo se comportan conjuntamente dos variables. En este capítulo realizaremos la descripción de esos casos, tal como lo hicimos en el capítulo 2 con sólo una variable, mediante tablas, gráficos y medidas de resumen.

Tablas de distribución bivariadas

Del mismo modo que describimos el modo en que se realizaban tablas para conocer la distribución de frecuencias de una variable, también ocurre que con dos variables se pueden construir tablas que combinen distribuciones de tipo II o III de ambas variables.

Ejemplo 3.1. Se tienen 44 muestras de suelo en una zona contaminada y se desea saber cómo se comportan conjuntamente el Cadmio y Níquel.

A partir de esto se construye una tabla conjunta, teniendo en cuenta que ambas distribuciones son de tipo III.

Variable				**Níquel**			
			8 a 20	**20 a 32**	**32 a 44**	**44 a 56**	$\mathbf{n_{.j}}$
		MC	14	26	38	50	
	1 a 3	**2**	4	2	0	0	6
Cadmio	**3 a 5**	**4**	6	7	1	0	14
	5 a 7	**6**	2	3	9	2	16
	7 a 9	**8**	0	1	2	3	6
	9 a 11	**10**	0	0	1	1	2
	$\mathbf{n_{i.}}$		12	13	13	6	44

Tabla 3.1. Tabla de distribución de frecuencias bidimensional de Cadmio y Níquel

En la Tabla 3.1 se observa la variable Níquel cuyos intervalos se presentan en columnas, mientras que en filas se presentan los intervalos de la variable Cadmio. Se presentan las marcas de clases (MC) de los respectivos intervalos y las frecuencias absolutas marginales. Podremos decir que 6 muestras presentaron entre 1 a 3 g de Cadmio, que 13 muestras presentaron entre 20 a 32 g de Níquel. La distribución conjunta de ambas variables nos lleva a concluir que 9 muestras presentaron entre 32 a 44 g de Níquel y 5 a 7 g de Cadmio.

A partir de las marcas de clases respectivas y los n marginales podremos calcular todas las medidas de posición y dispersión vistas en el Capítulo 2.

Covarianza.

La Covarianza es una medida de variación conjunta de ambas variables, a saber:

Covarianzas

$$\sigma^2_{x_1x_2} = \frac{\sum_{i=1}^{N}(x_{1i}-\mu_1)\ (x_{2i}-\mu_2)}{N} \qquad S^2_{x_1x_2} = \frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)\ (x_{2i}-\bar{x}_2)}{n-2}$$

Donde:
$\sigma^2_{x_1x_2}$: Covarianza poblacional
μ_1: Media poblacional de la variable x_1
μ_2: Media poblacional de la variable x_2
x_i Valores de la variable
N: número de elementos de la población

Donde:
$S^2_{x_1x_2}$: Covarianza muestral
$\overline{x_1}$: Media muestral de la variable x_1
$\overline{x_2}$: Media muestral de la variable x_2
$x_i$valores de la variable
n: número de elementos de la muestra

El recorrido de la Covarianza es:

$$-\infty < \sigma^2_{x_1x_2} < \infty$$

Cuando la Covarianza obtiene valores cero se dice que las variables x_1 y x_2 no poseen asociación lineal, cuando adquiere valore positivos se dice que x_1 y x_2 poseen una asociación lineal positiva (x_1 y x_2 incrementan juntos), cuando la covarianza es negativa se dice que la asociación lineal es inversa: cuando x_1 sube, x_2 baja. La Figura 3.1 muestra tres casos de asociaciones, donde se interpreta porqué la Covarianza arroja diferentes valores.

Coeficiente de Correlación lineal de Pearson.

A partir de un teorema que dice que: $\left|\sigma^2_{x_1x_2}\right| \leq \sigma_{x_1}\,\sigma_{x_2}$el investigador Karl Pearson(1857-1936) estandarizó la Covarianza en lo que se denomina el Coeficiente de Correlación Lineal, donde:

Coeficiente de Correlación lineal Poblacional

$$\rho_{x_1x_2} = \frac{\sigma^2_{x_1x_2}}{\sigma_{x_1}\,\sigma_{x_2}}$$

Donde:
$\rho_{x_1x_2}$: Coef. de correlación lineal poblacional
$\sigma^2_{x_1x_2}$: Covarianza poblacional
σ_{x_1} Desvío poblacional de x_1
σ_{x_2} Desvío poblacional de x_2

Coeficiente de Correlación lineal Muestral

$$r_{x_1x_2} = \frac{S^2_{x_1x_2}}{S_{x_1}\,S_{x_2}}$$

Donde:
$r_{x_1x_2}$: Coef. de correlación lineal muestral
$S^2_{x_1x_2}$: Covarianza muestral
S_{x_1} Desvío muestral de x_1
S_{x_2} desvío muestral de x_2

El recorrido del Coeficiente de correlación lineal es:

$$-1 < \rho < 1$$

Cuando ρ obtiene valores cero se dice que las variables x_1 y x_2 no están correlacionadas, cuando adquiere valores positivos se dice que x_1 y x_2 poseen una correlación lineal positiva (x_1 y x_2 incrementan juntos), cuandoρ es negativo se dice que la correlación lineal es inversa: cuando x_1 sube, x_2 baja.

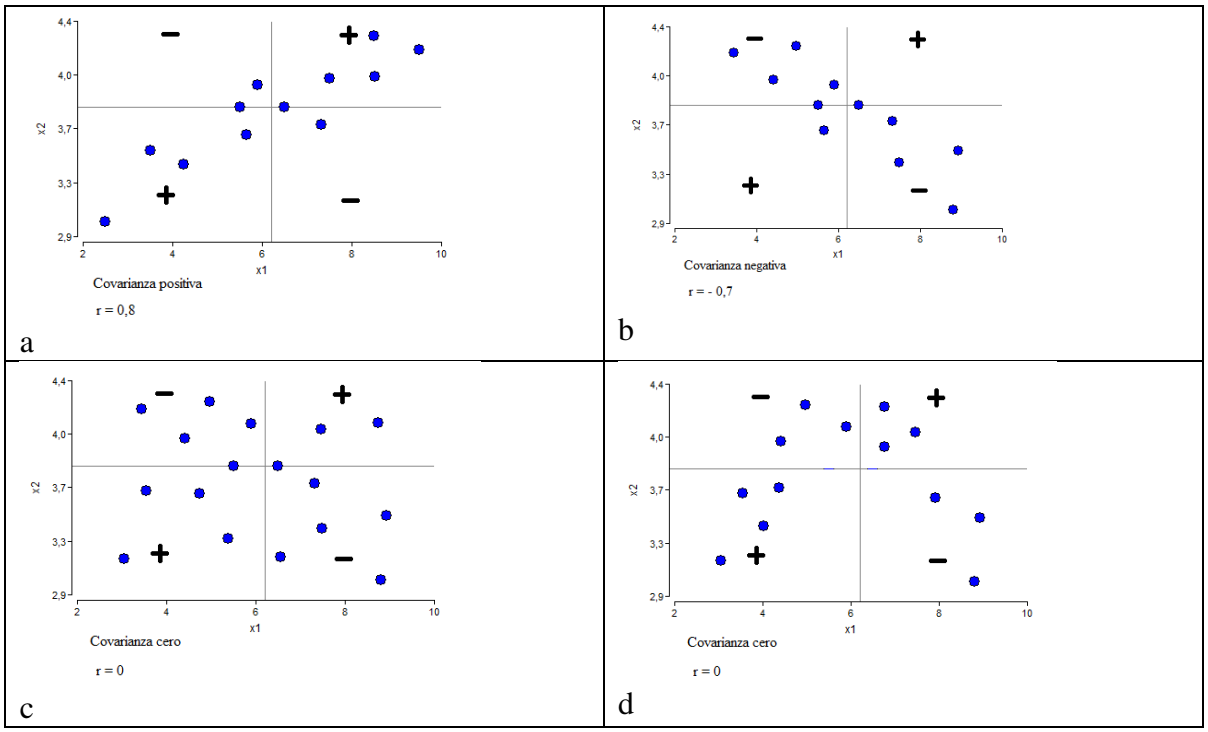


Figura 3.1.Gráficos con cuatro casos con valores de correlación y covarianza. a: asociación lineal positiva, b: asociación lineal negativa, c: ausencia de asociación, d: presencia de una asociación pero no lineal.

Capítulo 4.

Probabilidad.

Introducción.

Cuando sin conocimientos previos pensamos en la idea de una probabilidad, surgen como sinónimos palabras como posibilidad o chance. En principio continuaremos con la idea de "chance" hasta que tengamos las herramientas para definir matemáticamente a las probabilidades.

Experimento aleatorio

Definiremos como experimento aleatorio a toda aquella experiencia que trae aparejado un resultado que no se sabe a ciencia cierta cuál será. Es decir no hay certeza del resultado. Aquí resulta sencillo pensar como ejemplos en los juegos de azar. Si una persona tiene en sus manos una moneda y decide "tirarla" no hay certeza del resultado.

Ejemplo 4.1: Se tira una moneda y se observa el resultado.
Ejemplo 4.2: Se tiran dos dados y se cuenta la suma de sus caras.
Ejemplo 4.3: Se analiza una población de niños de un colegio y se realiza un análisis para observar la presencia o ausencia de un determinado virus.

Espacio muestral.

Definiremos al espacio muestral como el conjunto de todos los resultados posibles arrojados por un experimento aleatorio. Para continuar con el ejemplo de la tirada de una moneda los dos resultados posibles son cara y cruz. Distintos autores definen al espacio muestral con la letra **S** o bien Omega (**Ω**).
Entonces para el caso anterior:

Ejemplo 4.1: Ω= {C, X}
Ejemplo 4.2: Ω= {2, 3, 4, 5,..., 11, 12}
Ejemplo 4.3: Ω= {presencia de virus, ausencia de virus}

Evento: Cada uno de los resultados posibles de una experiencia aleatoria se denominará evento. En realidad se define como evento a cualquier subconjunto del espacio muestral (pero a los fines prácticos en este apunte utilizaremos como sinónimos lo que en algunos libros denominan eventos y sucesos).

Ahora, una vez definidos experimento aleatorio, espacio muestral y evento, definiremos Probabilidad.

Sea E un experimento aleatorio, que genera un espacio muestral Ω, compuesto por eventos, una **probabilidad** es un número real que se le asigna a cada uno de los eventos, tal que cumplen con los siguientes axiomas:

1) $P(A) \geq 0$
   La probabilidad de cualquier evento (por ejemplo A) es mayor o igual a cero (dicho de otro modo, no existen las probabilidades negativas).

2) $P(\Omega) = 1$
   La probabilidad del espacio muestral es igual a uno (no existen probabilidades mayores a uno).

3) Si $A \cap B = \emptyset$, entonces $P(A \cup B) = P(A) + P(B)$
   Si no existe la intersección entre los conjuntos A y B (es un espacio vacío) entonces la probabilidad de la unión entre A y B será igual a la suma de sus probabilidades.

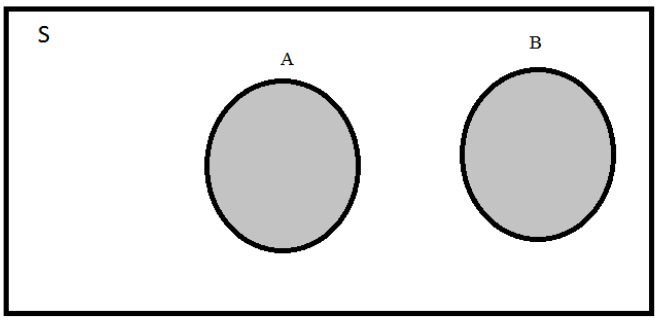


Figura 4.1. Diagrama de Venn, donde se observa que no existe intersección entre A y B.

A partir de estos axiomas postulados por Kolmogorov en 1937, se desprenden varias propiedades, las más importantes a los fines de este curso son:

Propiedades

1) $P(\emptyset) = 0$
   La probabilidad del espacio vacío es igual a cero. Lo que no existe, tiene probabilidad cero.

2) Si $A \cap B \neq \emptyset$, entonces $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
   Si la intersección entre los conjuntos A y B existe, entonces la probabilidad de la unión entre A y B será igual a la suma de sus probabilidades menos la probabilidad de la intersección.

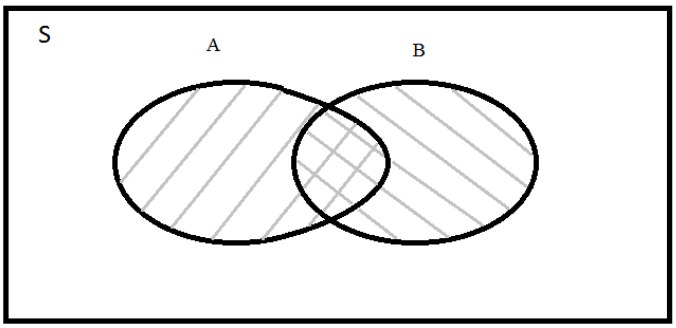


Figura 4.2. Diagrama de Venn donde se observa la intersección (a cuadros) entre A y B.

3) $P\ (\overline{A}) = 1 - P(A)$
Podríamos decir que el complemento del evento A es el conjunto de todos los eventos que no son A, dicho de otro modo la probabilidad que no ocurra A. De ese modo la probabilidad de que no ocurra A es la diferencia entre la probabilidad del espacio muestral menos la del evento A.

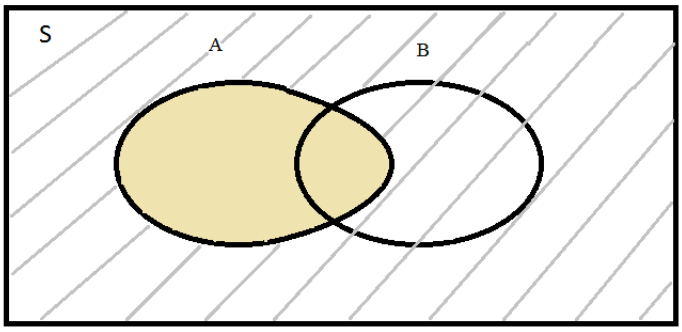


Figura 4.3 Diagrama de Venn donde se destaca (rayado) el complemento de A.

4) Independencia. Se dice que dos eventos A y B son independiente si la probabilidad de la intersección entre ellos es igual al producto de sus probabilidades, es decir, si

P (A ∩ B)= P (A) * P (B), entonces A y B son independiente

5) Condicionalidad. Se define la condicionalidad de un evento a otro del siguiente modo:
$P(A/B) = \frac{P(A \cap B)}{P(B)}$, lo cual se lee: La probabilidad de A condicionado a B es igual a la probabilidad de la intersección entre A y B dividido la probabilidad de B.

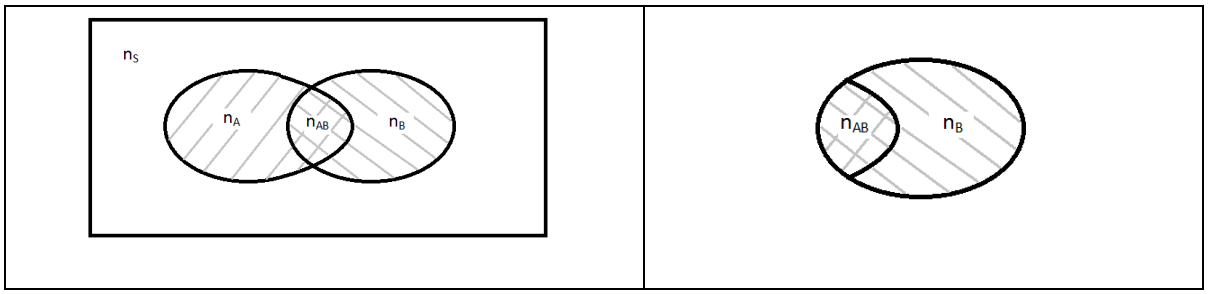


Figura 4.4: Diagrama de Venn donde se destaca que la probabilidad de que ocurra A dado que ya ocurrió B es el cociente entre el n_{AB} y n_B

6) Dos eventos A y B son mutuamente excluyentes si la ocurrencia de A excluye a B y viceversa.
P(A ∩ B) =∅,

Asignación de probabilidad.

1) Asignación Clásica:
Para aplicar esta asignación es necesario que todos los eventos del espacio muestral tengan igual probabilidad de ocurrencia:

$$P(A) = \frac{número\ de\ eventos\ de\ A\ (favorables)}{número\ de\ eventos\ igualmente\ posibles}$$

Ejemplo 4.4: Para la experiencia aleatoria tirada de una moneda perfecta, cada una de los dos resultados posibles posee una probabilidad de ½.

Ejemplo 4.5: Para la experiencia aleatoria tirada de un dado perfecto, cada uno de los resultados posibles posee una probabilidad de 1/6.

Como se puede observar la asignación clásica está restringida a escasos experimentos aleatorios (en general a juegos de azar). Para los otros casos (la gran mayoría) para conocer las probabilidades se necesita aplicar otra asignación.

2) Asignación frecuencista:

$$P(A)_{n\to\infty} = \frac{número\ de\ oportunidades\ que\ se\ produjo\ el\ evento\ \ A\ (favorables)}{número\ de\ oportunidades\ que\ se\ repitió\ el\ experimento\ aleatorio}$$

Ejemplo 4.6: En el ejemplo 4.3, se realiza un muestreo y se divide el número de niños con virus sobre el número de niños muestreados, eso arroja la probabilidad de encontrar niños con virus.
Ejemplo 4.7: La Tabla 2.1 del capítulo 2 presenta la distribución de frecuencias sobre las rocas encontradas en el fondo de un río serrano, a partir de eso decimos que existe una probabilidad de 0,55 de encontrar cuarzo (11/20).

3) Asignación Bayesiana.

Se basa en el teorema de Bayes (o de las probabilidades condicionadas). Los investigadores pueden intervenir subjetivamente en la asignación de las probabilidades. Dada la complejidad de esta asignación sólo la nombraremos, aunque en algunas áreas como mejoramiento animal y genética es muy utilizada.

Capítulo 5.

Variables Aleatorias I.

Introducción.

En el Capítulo 1 definimos a una variable como un carácter, una característica de las unidades que tomábamos con el objetivo de estudiarlas. Repetimos este concepto para enfatizar que NO es esa la definición de Variable Aleatoria que veremos a continuación por lo que se debe prestar especial atención ya que no son sinónimos variable y variable aleatoria.

Variable aleatoria

Dado un experimento aleatorio E y un espacio muestral Ω asociado, una variable aleatoria es una función X que le hace corresponder a cada uno de los eventos del espacio muestral un número real. Si el número real sólo se puede expresar en enteros, a la variable aleatoria se le denomina variable aleatoria discreta, pueden tener decimales, es una variable aleatoria continua.

Variables aleatorias discretas

Ejemplo 5.1 Se tiran dos monedas y se anota el número de caras que salen.

El espacio muestral original de la experiencia anterior es:

Ω= {XX, XC, CX, CC}, y esto trae aparejado:
0 1 1 2

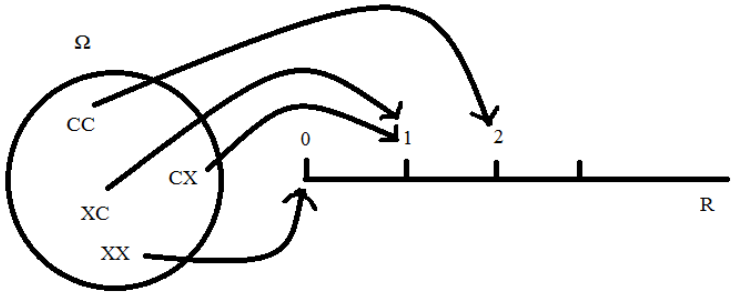


Figura 5.1. Diagrama de la aplicación de una variable aleatoria.

Entonces si tabulamos los resultados tendremos

X	p (x)	P (x)
0	1/4	1/4
1	2/4	3/4
2	1/4	1

Tabla 5.1. Valores de probabilidad basados en la asignación clásica.

Función de probabilidad.

Sea X una variable aleatoria discreta, se denomina función de probabilidad a la función p que le hace corresponder a cada uno de los resultados posibles x, un número real p(x), tal que se cumplan los axiomas de probabilidad.

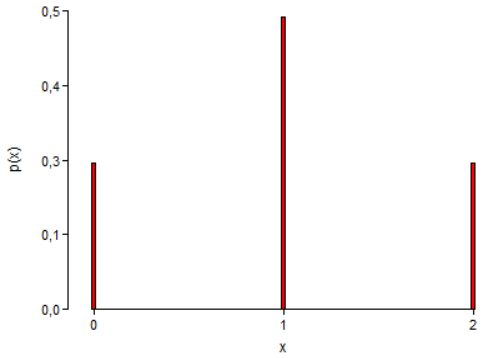


Figura 5.2. Gráfico de la función de probabilidad en el Ejemplo 5.1

Función de distribución.

Sea X una variable aleatoria discreta, se denomina función de distribución a la función P, tal que:

$$P_{x_j} = \sum_{x=x_1}^{x_j} p_{(x)}$$

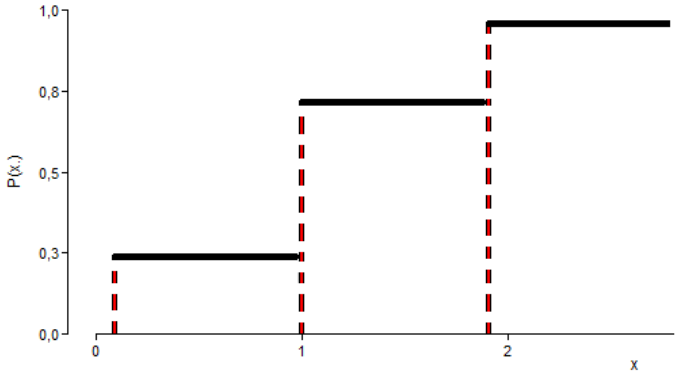


Figura 5.3. Gráfico de la función de distribución en el Ejemplo 5.1.

De este modo se puede observar que la tabla 5.1 posee tres columnas: la columna de los valores de la variable aleatoria discreta, la columna de la función de probabilidad y la de la función de distribución. Relacionemos esto con lo que se ha visto en el capítulo 2, Tabla 2.1 y en el capítulo 4: asignación frecuencista de las probabilidades. Podemos observar que lo que en forma descriptiva llamábamos frecuencia relativa, y asumiendo que esa tabla fue fruto de un experimento aleatorio, se tienen que esa columna coincide con la idea de la función de probabilidad, estableciendo probabilidades de un modo frecuencista. La diferencia fundamental es que lo desarrollado en el capítulo 2 ocurre después de haber tomado la muestra, en este capítulo se desarrolla de forma teórica, antes de tomar la muestra.

Esperanza y varianza de una variable aleatoria discreta. (v.a.d.)

Sea X una v.a.d., la Esperanza matemática o valor esperado de la variable aleatoria X se define como:

$$E(X) = \sum_{R_x} x\, p(x)$$

Para la Tabla 5.1 se obtiene:

X	p (x)	x *p (x)
0	1/4	0
1	2/4	2/4
2	1/4	2/4
∑		4/4

Tabla 5.2. Procedimiento para el cálculo de la E(x) en el ejemplo 5.1.

$$E(X) = 1$$

Resulta lógico pensar que si tiramos muchas veces dos monedas el valor esperado de número de caras que sacaremos será 1.

Sea X una v.a.d., la Varianza de X se define como:

$$V(X) = E(X^2) - (E(X))^2$$

Veamos el ejemplo anterior:

X	p (x)	x *p (x)	x^2	$x^2 * p(x)$
0	1/4	0	0	0
1	2/4	2/4	1	2/4
2	1/4	2/4	4	4/4

∑		4/4		6/4

Tabla 5.3.Procedimiento para el cálculo de la V(x) en el ejemplo 5.1.

$$V(X) = (6/4)^2 - (4/4)^2$$

$V(X) = 2{,}25 - 1 = 1{,}25$

Variables aleatorias continuas (v.a.c.)

Ejemplo 5.2: Supongamos que se toma de una población de alícuotas de agua de un lago (es decir todo el lago) una variable aleatoria continua, como por ejemplo pH. Entonces definimos:

Función de densidad.

Sea X una v.a.c., se dice que f (x) es una función de densidad si satisface las siguientes condiciones (recordemos que esto viene de los axiomas de probabilidad):

a) F(x)>0 para todo X en el R(x)
b) $\int_{Rx} f(x)dx = 1$

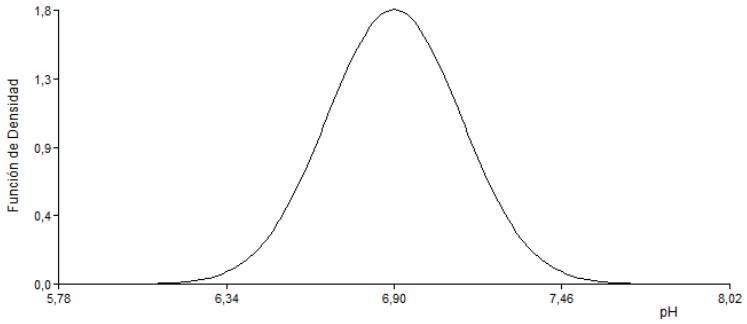


Figura 5.4 Gráfico de la función de densidad del ejemplo 5.2.

Función de distribución.
Sea X una v.a.c., se denomina función de distribución a la función F, tal que:

$$F_x = \int_{-\infty}^{x} f(x)dx$$

Ejemplo 5.3: Se tiene una función de las siguientes características:
Función de densidad

f(x)=	2x,	0 < x < 1
	0	para cualquier otro valor.

Entonces tendremos una función de distribución:

$$F(x)= \begin{cases} 0 & \text{si } x<0 \\ \int_0^x 2x\,dx \quad ;= |x^2|_0^X = x^2 & \text{si } 0 < x < 1 \\ 1 \quad \text{si } x > 1 & \end{cases}$$

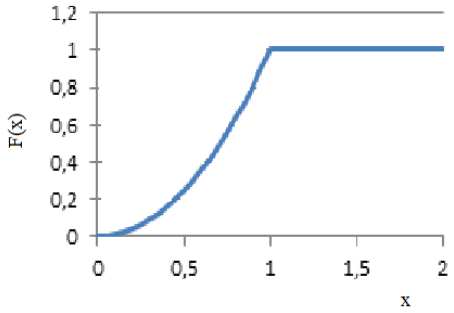


Figura 5.5a Gráfico de la función de distribución del ejemplo 5.3.

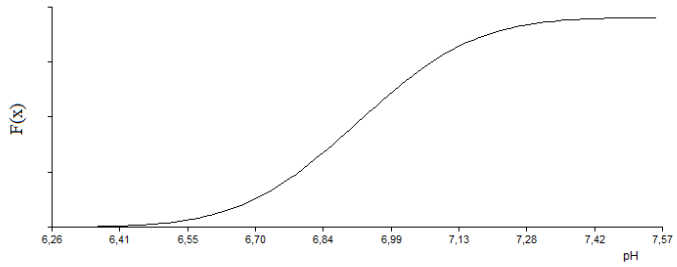


Figura 5.5b. Gráfico de la función de distribución del ejemplo 5.2.

Esperanza y varianza de una variable aleatoria continua. (v.a.c.)

Esperanza

Sea X una v.a.c., la Esperanza matemática o valor esperado de la variable aleatoria X se define como:

$$E(X) = \int_R x\, f(x)dx$$

Volviendo al ejemplo 5.3:

$$E(X) = \int_R^1 2x^2\, dx = \left|\frac{2}{3}x^3\right|_0^1 = \frac{2}{3}$$

Varianza

Sea X una v.a.c., la Varianza de X se define como:

$$V(X) = E(X^2) - (E(X))^2$$

Del ejemplo 5.3 se obtiene:

$$V(X) = E(x^2) - (E(x))^2 = \left|\frac{2}{4}x^4\right|_0^1 - (\frac{2}{3})^2 = \frac{1}{18}$$

Propiedades de la Esperanza y la Varianza de variables aleatorias

Existen numerosas propiedades de éstas, veremos sólo las relevantes:

Sea X una v.a., y E(X) y V(X) sus Esperanza y Varianza, y c= constante

1) E(c) = c
2) E (X + c) = E (X) + c
3) E (X * c) = E (X) * c
4) V(c) = 0
5) V (X + c) = V (X)
6) V (X * c) = V (X) * c^2

Distribuciones especiales de Probabilidad

1) Distribuciones de variables aleatoria discretas

a) Distribución Binaria.

Esta distribución no es aplicable a situaciones reales, sino que es teórica y sirve como una introducción a aquellas distribuciones más complejas:

Sea E un experimento aleatorio con un espacio muestral que posee sólo dos resultados posibles: Éxito o Fracaso (definido el éxito por el operador, dependiendo de la situación estudiada). Se define a la variable aleatoria X como 1, si el resultado es éxito y 0 si el resultado es fracaso

Ω= {Éxito, Fracaso}, y esto trae aparejado:
X 0 1
Y el recorrido de la Variable será entonces $R_x = [0, 1]$

Función de probabilidad

Se le asignará **p** como la probabilidad de x=1 y **q=1-p** a la probabilidad de x=0

Esto se suele expresar del siguiente modo:

$$X \sim B(p) = p^x q^{1-x}$$

Lo que significa que la variable aleatoria X posee distribución Binaria, con parámetro p. Luego se expresa la función de probabilidad.

Como se puede observar si el evento es un éxito, x=1, por lo que la probabilidad será p, mientras que si el evento es un fracaso, x=0, por lo que su probabilidad será q.

Esperanza y varianza de la distribución Binaria.

E(X)= p V(X)=p q

Ejemplo 5.4. El ejemplo es trivial. Un suelo de un campo de cultivo posee el 85% contaminado por herbicidas: ¿Cuál es la probabilidad de tomar una porción del suelo al azar y que esté contaminado?

Asumamos que el éxito a los fines de esta investigación estará dad por una muestra contaminada, por lo tanto:

$$X \sim B(0{,}85) = 0{,}85^x 0{,}15^{1-x}$$

Entonces

$P(x = 1) = 0{,}85^1 0{,}15^0 = 0{,}85$

Podríamos preguntarnos ¿Cuál es la probabilidad de tomar una porción del suelo al azar y que no esté contaminada?

$P(x = 0) = 0{,}85^0 0{,}15^1 = 0{,}15$

Aunque todas las distribuciones que veremos a continuación poseen una definición teórica, las veremos con el desarrollo de un ejemplo para que resulte más sencilla su interpretación.

b) Distribución Binomial.

Surge de repetir n veces una experiencia aleatoria binaria.

Ejemplo 5.5: Retomando el ejemplo anterior (Ejemplo 5.4) se pide ahora que un operador tome 5 muestras en el suelo y la variable aleatoria X será el número de muestras de suelo contaminadas (Nótese que hemos utilizado el término muestra que en la jerga de la geología implica una porción de suelo, aunque sabemos que estadísticamente nos estamos refiriendo a unidad).

Siendo N no contaminada y C contaminada, el espacio muestral para esta experiencia está dado por:

Ω= {NNNNN; CNNNN; NCNNN; NNCNN; NNNCN; NNNNC; CCNNN; CNCNN; CNNCN; CNNNC; ..... ; NCCCC; CCCCC}

Mientras que los posibles valores de X serán 0; 1; 2; 3; 4 y 5.

a) Empezaremos por pensar qué probabilidad hay de que ninguna muestra esté contaminada. Lo que debiera ocurrir para que el valor de x=0, es que cada una de las 5 muestras independientemente estén sin contaminar, lo que arrojaría una probabilidad de:

0,15*0,15*0,15*0,15*0,15= $0{,}15^5 = 0{,}0000759375$

Veamos que

$$P(x = 0) = q^{1-x}$$

b) ¿Cuál será la probabilidad de encontrar una muestra contaminada? Lo que debiera ocurrir para que el valor de x=1, es que sólo una de las 5 muestras esté contaminada, mientras que 4 estarán sin contaminar, lo que arrojaría una probabilidad de:

0,85*0,15*0,15*0,15*0,15= $0{,}85^1 * 0{,}15^4 = 0{,}0004303125$

Veamos que

$$P(x = 1) = p^x q^{1-x}$$

Pero no tenemos en cuenta que sólo hemos tenido en cuenta que la primera muestra esté contaminada. Debemos darle la posibilidad a que la segunda muestra lo esté, así como la tercera, cuarta o quinta muestra. Para eso debiéramos representar en la fórmula a la cantidad de combinaciones en que se puede "acomodar" una muestra contaminada en grupos de 5 muestras. Para ello utilizaremos la fórmula de combinatoria:

Cuando se expresa: $\binom{n}{x} = \frac{n!}{x!(n-x)!}$

Se denota el número de combinaciones de x en un grupo de n, en el ejemplo:

$$\binom{5}{1} = \frac{5!}{1!\,(5-1)!} = \frac{5!}{(4)!} = \frac{5 * 4!}{4!} = 5$$

Lo que vimos cuando desarrollamos el espacio muestral, que una muestra contaminada se puede "acomodar" de 5 maneras diferentes en un grupo de 5 muestras.

Volvamos entonces a la función de Probabilidad ahora completa:

$$X \sim Bi(n;\ p) = p^x q^{1-x} \binom{n}{x}$$

Donde:
X: variable aleatoria de distribución Binomial
n: número de muestras (en realidad elementos de la muestra)
p: probabilidad del éxito en la población
q= 1-p: probabilidad de fracaso en la población

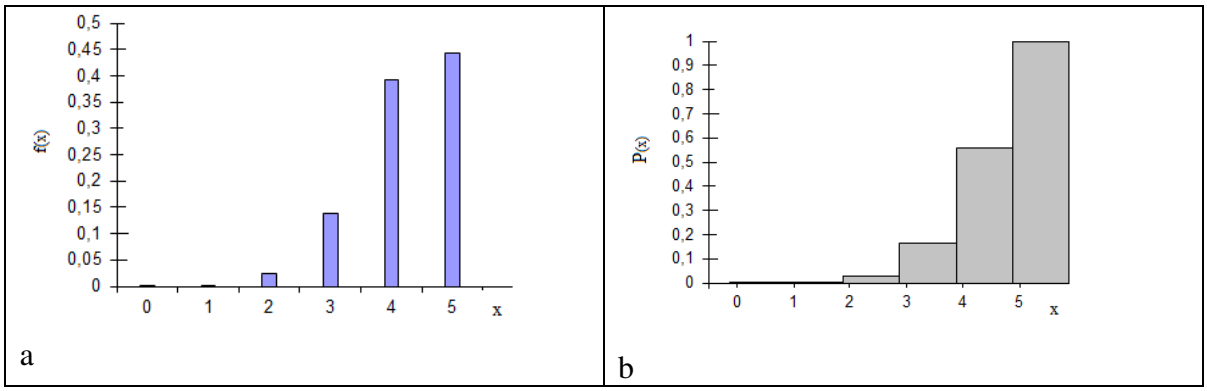
Parámetros, Esperanza y Varianza de la distribución Binomial

Parámetros: n y p
E(X)= n p
V(X)= n p q

Para nuestro ejemplo (Ejemplo 5.5)

E(X)= 5 * 0,85 = 4,25
V(X)= 5 * 0,85 * 0,15= 0,6375

Dicho de otro modo el número esperado de muestras contaminadas en grupos de 5 muestras será de 4, 24, con una varianza de 0,6375. La Figura 5.6 representa la función de probabilidad del Ejemplo 5.5.



a b

Figura 5.6a. Función de Probabilidad de la distribución Binomial del Ejemplo 5.5 y b: Función de distribución.

c) Distribución Poisson.

Vamos a considerar dos formas diferentes de abordaje a esta distribución (aunque hay más de dos formas).

I) Poisson como aproximación a la Binomial

Imaginemos que el evento que se está estudiando tiene una probabilidad de ocurrencia de 0,5 (p=0,5), si esto ocurre, con un n de 4 estamos frente a un valor esperado de 2. Dicho de otro modo con tomar una muestra de 4 tenemos altas chances de encontrar 2 unidades con

la característica éxito. Pero si el evento tiene una probabilidad de 0,1, debiéramos tomar 10 para tener un valor esperado de 1. Del mismo modo, si la probabilidad del éxito es de 0,01, el n debiera ser 100. Así se va observando que a menor probabilidad del éxito el operador deberá hacer más esfuerzo (aumentar el n).
Dicho de otro modo, cuando p→ 0; n→ ∞, de modo tal que n*p es constante.
Se puede demostrar matemáticamente que:

$$\lim_{n\to\infty} p^x q^{1-x} \binom{n}{x} = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

Donde:
$\lambda = n * p$

II) Proceso de Poisson

En estos casos la variable en estudio no es sólo una aproximación de la Binomial. Son ejemplos de esta aplicación eventos que se miden en conteos y distribución espacial de algunos eventos que siguen una distribución Poisson. En estos casos la distribución se refiere a una variable discreta donde como se verá en algunos párrafos la varianza crece a medida que se incrementa la esperanza. La función es la misma que se mencionó anteriormente, pero aquí λ = número de casos por unidad de tiempo, espacio o superficie.

Ejemplo 5.6. La probabilidad de ocurrencia de eventos sísmicos sigue una distribución de Poisson, con probabilidades muy pequeñas de ocurrencia en un determinado momento y con la necesidad de monitorear períodos largos de tiempo.

Ejemplo 5.7. En un estudio paleontológico sobre *Ammonia tepida* en estuarios se observó que la distribución de esta especie en los sedimentos seguía un proceso de Poisson con un λ de 3,2 individuos por m^2.

Parámetro, Esperanza y Varianza de la Distribución de Poisson

Parámetro: λ
$E(X)= \lambda$
$V(X)= \lambda$

2) Distribuciones de variables aleatorias continuas

a) Distribución Uniforme.

Del mismo modo que la distribución Binaria es a las variables discretas, la Uniforme es a las continuas. Sólo es una distribución teórica en la cual es difícil encontrar ejemplos de aplicación.

La variable aleatoria X posee un rango de valores: $[a; b]$

Todos los valores de X son igualmente posibles, por lo que:

$f(x) = \frac{1}{b-a}$ para a < x < b

F(x) $\begin{cases} 0 & \text{para } x < a \\ \frac{x-a}{b-a} & \text{para } a < x < b \\ 1 & \text{para } x > b \end{cases}$

Parámetros, Esperanza y varianza de la distribución Uniforme

Parámetros: a y b
E(X)= $(a+b)\ /2$
V(X)= $(b-a)^2/12$

Ejemplo 5.8. Una variable x posee distribución Uniforme, con un rango de valores que van de 1 a 3. La figura 5.7 representa su función de densidad.

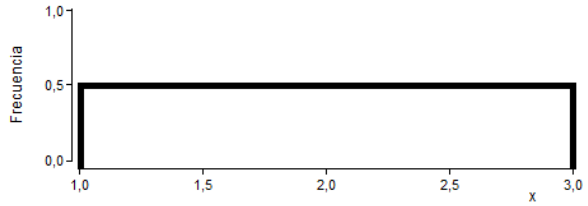


Figura 5.7. Función de densidad de la variable Uniforme, donde a=1, b= 3)

b) Distribución Normal.

A partir de numerosas observaciones sobre los errores de medición, el matemático Gauss observó que la distribución de esos errores era simétrica y eran muy comunes los errores de poca magnitud, mientras que eran poco frecuentes los errores grandes. A partir de eso adaptó una función desarrollada por De Moivre y la transformó en una distribución probabilística agregándole un factor de corrección (es decir que sea una variable aleatoria donde se cumplen los axiomas de probabilidad, dicho de otro modo la probabilidad del espacio muestral debe ser 1). Ver Figuras 5.4 y 5.5b

La función de densidad de esta distribución es (Figura 5.4):

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma}}$$

La forma de esta distribución es la denominada campana de Gauss o campana de la distribución normal

Por su parte la función de distribución está dada por (Figura 5.5b):

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma}} dx \qquad -\infty < x < \infty$$

Parámetros, Esperanza y Varianza de la distribución Normal

Parámetros:μ $y\ \sigma^2$
E(X)= μ
V(X)= σ^2

Por lo tanto se dice que:

$$X \sim N(\mu;\ \sigma^2)$$

Ejemplo 5.9.Supongamos que la variable aleatoria concentración de Monóxido de Carbono dentro de una empresa durante el período de máxima producción sigue la distribución Normal con Media 8 ppm y varianza 2 ppm^2.

De este modo la forma de la distribución será:

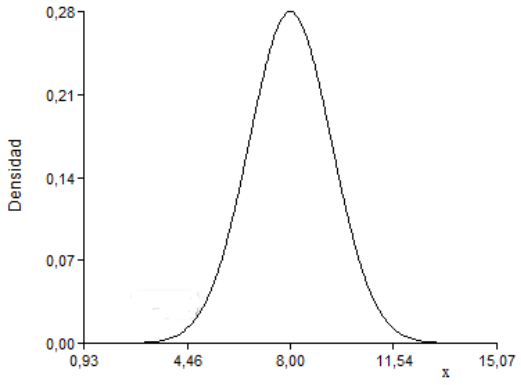


Figura 5.8 Distribución de la Concentración de Monóxido de Carbono dentro de una empresa

Una de las preguntas que pueden surgir es ¿qué probabilidad hay que los valores superen los 10 ppm?

Como bien se vio anteriormente se debiera resolver la integral para calcular el espacio bajo la curva entre el valor 10 e infinito. Como bien se supone, resolver una integral tan compleja, cada vez que es necesario encontrar una probabilidad, es una tarea muy tediosa. Como se vio anteriormente en distribuciones de Poisson y Binomial podría estar tabuladas, pero recordemos que los parámetros de la distribución normal (que son la media y la varianza) pueden tener infinitos valores diferentes, lo que implicaría poseer una tabla para cada valor de μ y para cada de σ^2.

Por esta primera razón se creó una tabla denominada Normal estandarizada o tipificada que tiene características muy particulares. Es una distribución Normal pero que posee Esperanza cero (0) y Varianza uno (1), es decir es una variable llamada Z que:

$$Z \sim N(0;\ 1)$$

Esta distribución Z está tabulada y mediante una sencilla operación se pueden transformar a todas las variables X en la variable Z, esta operación se denomina Estandarización

Estandarización

Definimos como Estandarización a la operación mediante la cual se transforma a una variable X: $X \sim N(\mu;\ \sigma^2)$ a una Z: $Z \sim N(0;\ 1)$

La fórmula general de la Estandarización es:

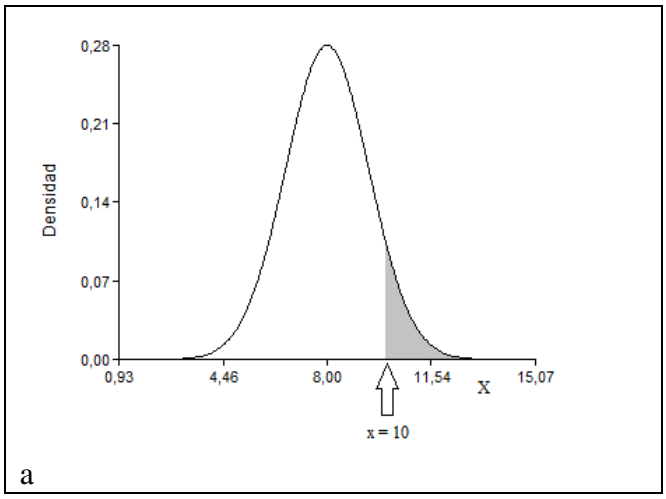
$$Z = \frac{Variable\ aleatoria - Esperanza\ de\ la\ v.a.}{Raíz\ cuadrada\ de\ la\ varianza\ de\ la\ v.a.}$$

Para este caso en particular (Ejemplo de una variable X) entonces: E(X)= μ y V(X)= σ^2, la fórmula es:

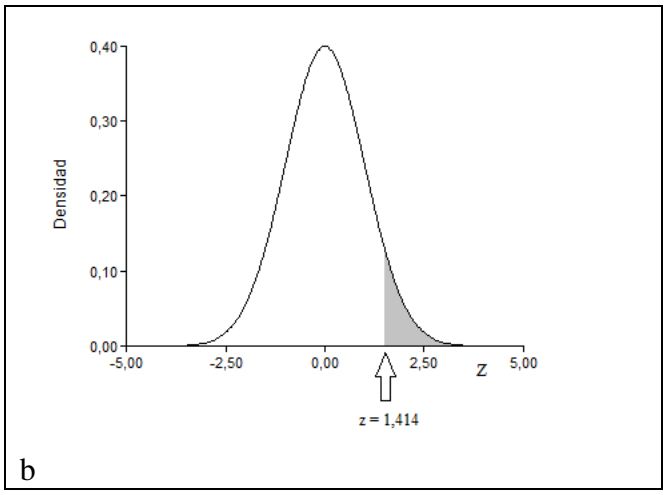
$$Z = \frac{x - \mu}{\sqrt{\sigma^2}}$$

Observemos para el ejemplo 5.9 qué ocurre con la estandarización del valor 10:

$$Z = \frac{10 - 8}{\sqrt{2}} = 1{,}414$$



a

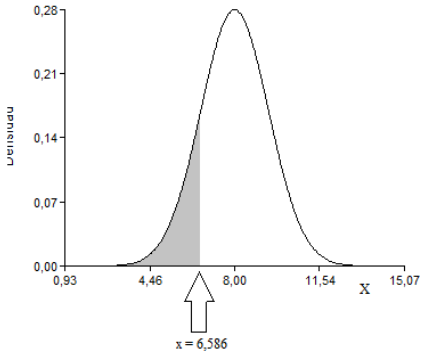
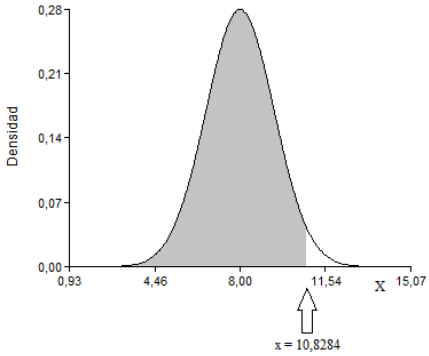
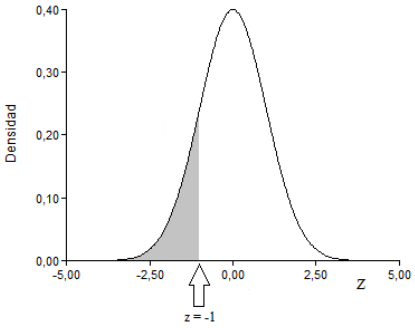
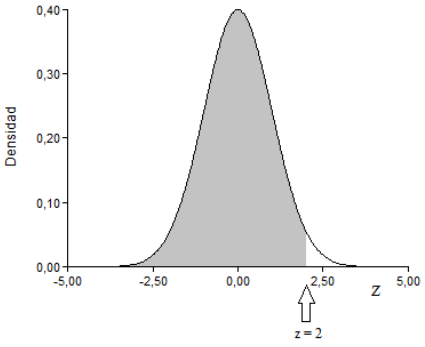


b

Figura 5.9a. Distribución Normal y la ubicación del valor x=10 y del mismo valor estandarizado z=1,414 (b)

Ante la pregunta cuál es la probabilidad de que x>10: P (x >10)= P (z >1,414)= 0,0787, (de la búsqueda de la tabla Normal estándar surge el valor de la probabilidad).

Para diversos casos ver Figuras 5.10

P (x < 6,586)= P (z < -1) = 0,1587	P (x < 10,8284)= P (z < 3) =0,9772
 a	 c
 b	 d

Figuras 5.10. Valores de x graficados en la distribución de la variable x (a y c) y sus respectivas estandarizaciones (b y d).

Ahora nótese que se puede visualizar que los valores z están mostrando cuántos desvíos antes o después de la media se encuentra el valor de la variable, es decir si un valor está alejado o no de la media. El valor de x= 6,583 se encuentra ubicado un desvío antes d ela media, por eso su valor z es -1.

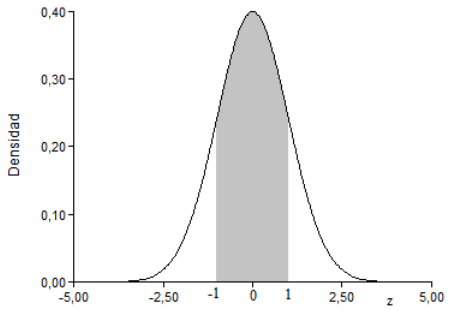
Entonces los valores z están relativizando a los valores de la variable. ¿Una persona de 1,85 m es alta o baja? La respuesta debiera ser: depende. Si esa persona nació en Argentina, es una persona relativamente alta, para la media de este país y su desvío esa persona tendrá un valor z de aproximadamente 2. Pero si nació en Finlandia, es una persona promedio, pues su valor z es 0, ya que los finlandeses en promedio miden 1,85 m.

Observando la superficie bajo la curva se constata que:

P (-1 < z < 1) = 0,68; es decir que entre 1 desvío a la derecha y uno a la izquierda se encuentran aproximadamente el 68% central de los datos.

P (-1,96 < z < 1,96)= 0,95; es decir que casi 2 desvíos a la derecha y a la izquierda contienen aproximadamente el 95% central de los datos.

P (-3 < z < 3)= 0,99; es decir que 3 desvíos a la derecha y a la izquierda contienen algo más del 99% central de los datos

a. P(-1<z<1)= 0,6827	

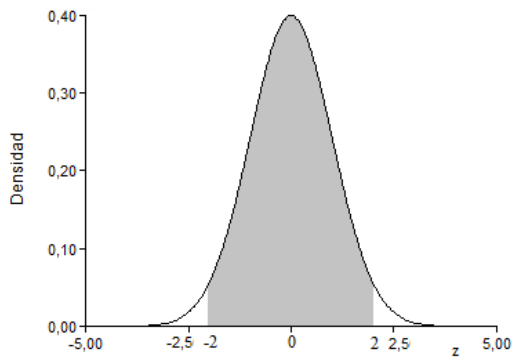
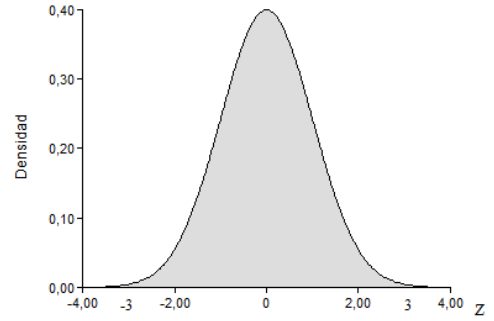
b. P(-2 <z< 2)= 0,9545	
c. P(-3 <z< 3)= 0,9973	

Figura.5.11. Funciones de densidad y valores estandarizados: 1, 2 y 3.

Capítulo 6.

Variables Aleatorias II.

Distribución de funciones de variables aleatorias

Introducción.

En el capítulo anterior se desarrolló la distribución de variables aleatorias. A su vez en el Capítulo 2 se desarrollaron conceptos como la media, la varianza entendiendo que son funciones de variables aleatorias. Ahora vamos a ver de qué modo se distribuyen esas funciones de variables aleatorias.

Si "X" es una variable aleatoria y "H(X)" es una función de ésta, "Y" es una variable aleatoria de la función de "X" con:

$$E(Y) = E(H(X))$$

De tal modo que en caso discreto sería:

$$E(Y) = \sum_{Rx} H(x)p(x)$$

Y en el caso continuo:

$$E(Y) = \int_{Rx} H(x)\, f(x)dx$$

Y para ambos casos, tanto discretos como continuos:

$$V(Y) = E(Y^2) - (E(Y))^2$$

Propiedades de Esperanza y varianza de funciones

1) Sean X e Y dos variables aleatorias cualquiera (no necesariamente independientes), entonces
   E(X + Y)= E (X) + E (Y)
2) Sean X e Y dos variables aleatorias cualquiera e independientes
   V(X + Y)= V (X) + V (Y)
3) Sean X e Y dos variables aleatorias cualquiera y no independientes
   V(X + Y)= V (X) + V (Y) ± 2 Covarianza(X,Y)

Distribución del estadístico "χ^2".

Dadas Z_1, Z_2, Z_3; .... Z_n, n variables aleatorias independientes, con distribución Normal estándar ($Z \sim N(0;\ 1)$), entonces definimos a la distribución Chi cuadrado como:

$$\chi^2_{n-1} = Z_1^2 + Z_2^2 + Z_3^2 + \cdots + Z_n^2$$

Es decir una Chi cuadrado es la sumatoria de Normales estandarizadas al cuadrado.

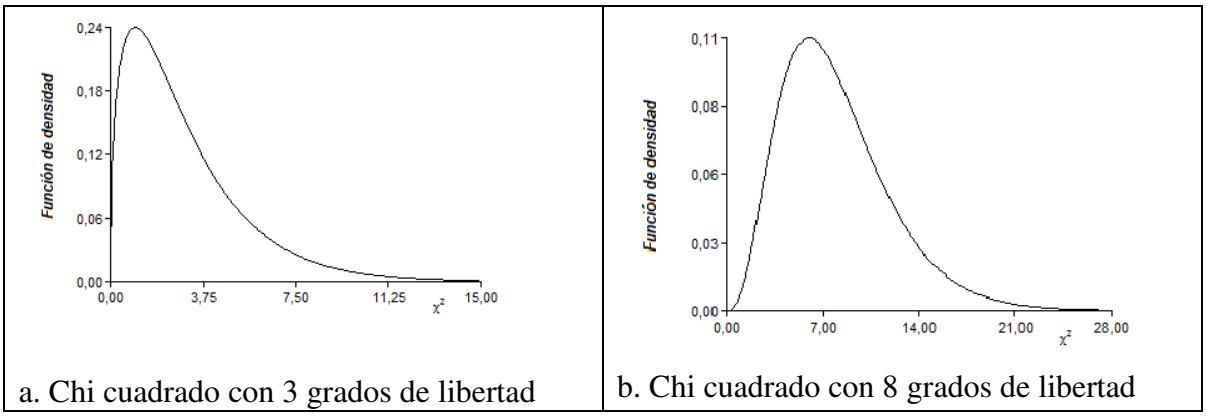
Función de densidad de la distribución Chi Cuadrado

$f(n-1) = \frac{(\chi^2)^{\frac{n-1}{2}-1} e^{-\frac{\chi^2}{2}}}{2^{\frac{n-1}{2}}(\frac{n-1}{2}-1)}$, para $\chi^2 > 0$

Parámetro, Esperanza y varianza de la Distribución Chi cuadrado

Parámetro: n-1 (grados de libertad)
E (χ^2)= n-1
V (χ^2)= 2 n-1

En las figuras 6.1 se observa que la distribución Chi cuadrado es asimétrica a la derecha y que va haciéndose más simétrica a medida que se incrementan los grados de libertad



a. Chi cuadrado con 3 grados de libertad

b. Chi cuadrado con 8 grados de libertad

Figura 6.1. Función de densidad de la distribución Chi Cuadrado

Distribución del estadístico "t".

Dada Z una variable aleatoria con distribución normal estándar ($Z \sim N(0;\ 1)$), y dada Y una variable aleatoria con distribución Chi cuadrado, entonces definimos a la distribución "t" de Student como:

$$t_{n-1} = \frac{X}{\sqrt{\frac{Y}{n-1}}}$$

Es decir una distribución "t" es una normal dividido la raíz de una chi cuadrado dividida sus grados de libertad

Función de densidad de la distribución "t"

$$f(t) = \frac{\left[\frac{(n)}{2}\right]!}{\left[\frac{(n-1)}{2}\right]!\sqrt{\pi\, n-1}}\left[1 + \frac{t^2}{n-1}\right]^{\frac{n}{2}}, \qquad -\infty < t < \infty$$

Parámetro, Esperanza y varianza de la Distribución t
Parámetro: n-1 (grados de libertad)
E (t)= 0
V (t)= $\frac{n-1}{n-3}$

Tiene forma de campana, es simétrica, a medida que se incrementa el n, arroja probabilidades similares a la distribución normal, cuando n>30 se dice que la t es significativamente similar a la normal

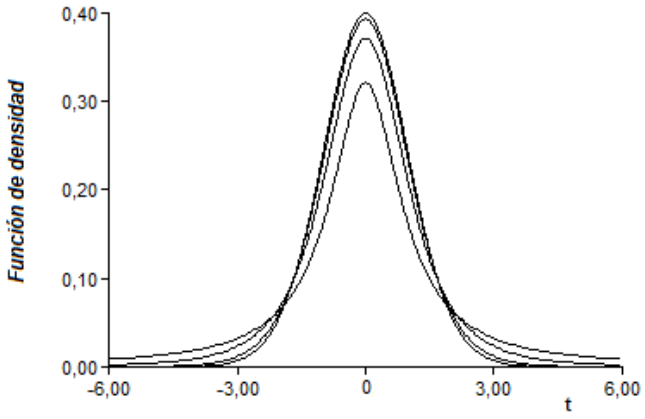


Figura 6.2. Funciones de densidad de diferentes distribuciones "t". A medida que se incrementa el n, aumenta su puntiagudez.

Distribución del estadístico "F".

Dada una variable aleatoria X con distribución chi cuadrado y n-1 grados de libertad, dada otra variable aleatoria Y con distribución chi cuadrado y m-1 grados de libertad, definimos a la variable aleatoria F (de Fisher) como:

$$F_{n-1;m-1} = \frac{X}{Y}$$

Es decir una F es un cociente entre dos variables aleatorias Chi cuadrado.

Función de densidad de la Variable F

$$f(F) = \frac{\left[\frac{(m+n-2)}{2}\right]!}{\left[\frac{(m-2)}{2}\right]!\left[\frac{(n-2)}{2}\right]!}\left[\frac{m}{n}\right]^{\frac{m}{2}} \frac{F^{\frac{m-2}{2}}}{\left[1+\frac{mF}{n}\right]^{\frac{m+n}{2}}}, \qquad F > 0$$

Parámetro, Esperanza y varianza de la Distribución F

Parámetros: n-1 y m-1 (grados de libertad de ambas Chi cuadrado (X e Y)
E (F)= 1
V (F)= $\frac{2\,n^2(m+n-2)}{m(n-2)^2(n-4)}$

F siempre es positiva, asimétrica a la derecha,

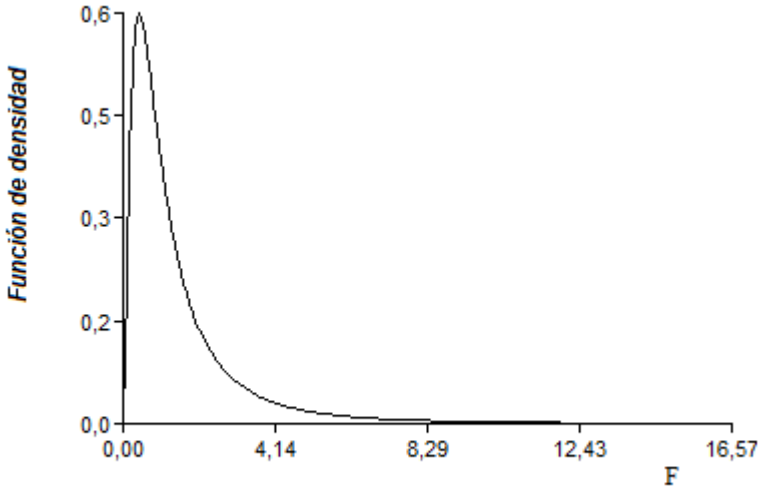


Figura 6.3. Función de densidad de una variable F con 5 y 6 grados de libertad.

Teorema Central del Límite.

Este teorema es muy importante en la práctica de la estadística. Posee varios abordajes, pero lo simplificaremos a los fines didácticos:

Dadas X_1, X_2, X_3; .... X_n, n variables aleatorias independientes e idénticamente distribuidas, es decir todas con igual Esperanza (μ) y con igual varianza (σ^2). La sumatoria de estas variables se distribuirá con distribución Normal, si el n es suficientemente grande, sin importar la distribución original de aquellas.

La aplicación más importante de este teorema tiene que ver con la distribución de la media muestral. Como se observa en la fórmula de la media muestral:

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

Entonces si $X_1.X_2$, X_n son variables aleatorias, la media es la sumatoria de variables aleatorias que multiplican a una constante:($\frac{1}{n}$), que no cambia la distribución.

$$\bar{X} = \sum X\frac{1}{n}$$

Capítulo 7.

Distribuciones en el muestreo

Introducción.

Si hacemos un pequeño repaso de lo que se ha visto hasta ahora podríamos destacar que en los primeros capítulos definimos estadística, población y muestra. Luego se describieron la distribución de las variables, medidas de posición y dispersión. Pasamos luego a los conceptos básicos de probabilidad, de variables aleatorias con la distribución de variables y la distribución de funciones de variables.
En este capítulo uniremos muchos de esos conceptos y desarrollaremos las bases que constituyen el puntapié inicial para comprender la inferencia estadística.

Razones para utilizar muestras. Muestreo aleatorio.

En el capítulo 1 se desarrollaron los conceptos de muestreo desde el punto de vista de la capacidad de conseguir una muestra que represente a la población. En capítulos subsiguientes hemos visto que es necesario que la variable en estudio sea una variable aleatoria, producto de un experimento aleatorio y no cualquier variable tomada subjetivamente por el operador.

Distribución de los estimadores. Parámetros poblacionales.

Si recordamos el Capítulo 2, veremos que se describieron dos formas distintas de cálculo de la media y de la varianza, dependiendo si estamos hablando de una población o de una muestra:

	Poblacional	Muestral
Media	$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
Varianza	$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$	$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$

A partir de esto podremos definir a un **parámetro** como a la función tomada de la población, mientras que llamaremos **estimador** a aquellas tomadas en la muestra. De este modo podremos decir que el parámetro poblacional es único pues se obtiene del censo realizado en la totalidad de la población, mientras que los estimadores muestrales pueden obtener tantos valores diferentes como la combinación de unidades que puedan "entrar" en las distintas muestras. Es decir un estimador va a variar de muestra en muestra, por lo tanto y como se ha visto en el capítulo anterior, es una variable aleatoria con una distribución determinada.

Distribución de la media muestral.

Sea $X \sim N(\mu;\ \sigma^2)$, ¿qué distribución tendrá $\bar{X}$?

Como la media muestral es la sumatoria de variables aleatorias con distribución normal, por lo tanto va a tener distribución normal como resultado de la suma de normales.
¿Qué pasaría si las variables aleatorias no tienen distribución normal? Si el n es suficientemente grande, por teorema central del límite, la media muestral tendrá distribución normal.

¿Cuál será la esperanza de la media muestral, sabiendo que la E (X)=μ?

$$E(\bar{x}) = E(\frac{\sum x}{n}) \; ; E(\bar{x}) = E(\sum x \frac{1}{n}) \; ; E(\bar{x}) = \frac{1}{n} E(\sum x)$$
$$E(\bar{x}) = \frac{1}{n} \sum E(x) \; ; E(\bar{x}) = \frac{1}{n} \sum \mu \; ; E(\bar{x}) = \frac{1}{n} \, n \, \mu$$
$$E(\bar{x}) = \mu$$

Entonces el valor esperado de $\bar{x}$, es el mismo de x: μ.

¿Cuál será la varianza de la media muestral, sabiendo que la varianza de X es σ^2?

$$V(\bar{x}) = V(\frac{\sum x}{n}) \; ; \; V(\bar{x}) = V(\sum x \frac{1}{n}); V(\bar{x}) = \frac{1}{n^2} V(\sum x)$$
$$V(\bar{x}) = \frac{1}{n^2} \sum V(x) \; ; V(\bar{x}) = \frac{1}{n^2} \sum \sigma^2 \; ; V(\bar{x}) = \frac{1}{n^2} \, n \, \sigma^2$$
$$V(\bar{x}) = \frac{1}{n} \sigma^2 \; ; \; V(\bar{x}) = \frac{\sigma^2}{n}$$

Entonces la varianza de $\bar{x}$, no es la misma varianza de x, es: $\frac{\sigma^2}{n}$, dicho de otro modo, a mayor cantidad de unidades que forman parte de la muestra, la media muestral varía cada vez menos. Entonces para resumir:

$$\bar{X} \sim N\left(\mu; \frac{\sigma^2}{n}\right)$$

De este modo el desvío de la variable media muestral recibe el nombre de error estándar y está definido por:

$$EE = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

Estandarización de la media muestral

Si recordamos del capítulo 6 la fórmula general de la Estandarización, vimos que:

$$Z = \frac{Variable\ aleatoria - Esperanza\ de\ la\ v.a.}{Raíz\ cuadrada\ de\ la\ varianza\ de\ la\ v.a.}$$

Para este caso en particular entonces donde $E(\bar{x}) = \mu$ y $V(\bar{x}) = \frac{\sigma^2}{n}$, la fórmula es:

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

Distribución de la diferencia de medias muestrales.

Sea $X_1 \sim N(\mu_1;\ \sigma_1^2)$ y $\overline{X_1} \sim N\left(\mu_1;\ \frac{\sigma_1^2}{n_1}\right)$, y sea $X_2 \sim N(\mu_2;\ \sigma_2^2)$ y $\overline{X_2} \sim N\left(\mu_2;\ \frac{\sigma_2^2}{n_2}\right)$

La diferencia de medias muestrales se distribuirá como:

$$\overline{X_1} - \overline{X_2} \sim N\left(\mu_1 - \mu_2;\ \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

La estandarización de la diferencia de medias muestrales será:

$$Z = \frac{\overline{(X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Distribución del estadístico $\frac{(n-1)S^2}{\sigma^2}$. Distribución de la varianza muestral.

Sea $X \sim N(\mu;\ \sigma^2)$ y S^2 la varianza muestral obtenida de una muestra de tamaño n, ésta se distribuye como una Chi cuadrado bajo estas condiciones:

$$\chi^2_{n-1} = \frac{(n-1)S^2}{\sigma^2}$$

Distribución del estadístico $\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}$. Distribución de la media con varianza poblacionaldesconocida

Sea $X \sim N(\mu;\ \sigma^2)$; $\bar{X}$ la media muestral, S^2 la varianza muestral, ambas obtenidas de una muestra de tamaño n, la media muestral tendrá distribución t bajo estas condiciones:

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

Distribución del estadístico $\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$

Sea $X_1 \sim N(\mu_1;\ \sigma_1^2)$ y $X_2 \sim N(\mu_2;\ \sigma_2^2)$;S_1^2la varianza muestral de la población 1, obtenida de una muestra de tamaño n1; S_2^2la varianza muestral de la población 2 obtenida de una

muestra de tamaño n2. El cociente entre las varianzas muestrales de las muestras 1 y 2 se distribuye como una F de Fisher bajo estas condiciones:

$$F_{(n_1-1;n_2-1)} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

Capítulo 8.

Estimación.

Introducción.

En capítulos anteriores ya hemos visto la diferencia entre estimador y parámetro, entendiendo que un parámetro es una característica que está tomada sobre la población, mientras que el estimador es una característica tomada a partir de la muestra. El parámetro es único, mientras que el estimador difiere de muestra en muestra y posee una distribución, ya que es una variable aleatoria.
En este capítulo ya entraremos de lleno a la parte de la Estadística llamada Inferencial, ya que a partir de resultados particulares de una muestra inferimos sobre la población.

Estimación puntual.

Vamos a denominar en forma genérica aθ como el parámetro y a $\hat{\theta}$como el estimador. El estimador $\hat{\theta}$ es un estimador puntualdeθ ya que lo estima en la recta de los números reales y debemos comprender que no cualquier $\hat{\theta}$ es un buen estimador deθ para ello tiene que cumplir ciertos requisitos:

Propiedades de los buenos estimadores

1) Insesgabilidad

Sea $\hat{\theta}$ un estimador de θ, se dice que $\hat{\theta}$ es un estimador insesgado de θ si se cumple que:

$E(\hat{\theta}) = \theta$;
entendemos por sesgo entonces a la diferencia entre la esperanza del estimador y el parámetro:
Sesgo $\hat{\theta} = \theta - E(\hat{\theta})$

Ejemplo 8.1: Ya se ha visto en el capítulo 7 que $E(\bar{x}) = \mu$, por lo tanto $\bar{x}$ es un estimador insesgado de μ.

Ejemplo 8.2: En este ejemplo demostraremos la razón por la cual la fórmula de la varianza muestral es diferente del de la varianza poblacional.

Recordemos que en el capítulo 2 se vieron dos fórmulas que son semejantes, pero difieren en el denominador:

$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$ $\qquad$ $S^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$

Donde:
σ^2 : Varianza poblacional y S^2: Varianza muestral.

El motivo por el cual estas fórmulas son diferentes es el siguiente: si calculamos la esperanza de la varianza muestral sólo dividiendo por n, se demuestra que el estimador es sesgado, mientras que si se corrige con n-1, el estimador es insesgado:

Denominaremos S_S^2 al estimador sesgado de σ^2:

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

Entonces:
$E(S_S^2) = \frac{1}{n}\sum E((x_i - \bar{x})^2)$; recordemos la fórmula de trabajo desarrollada en el Capítulo 2, y la rescribiremos:
$S_S^2 = \frac{1}{n}(\sum x^2 - n\bar{y}^2)$; de modo tal que:

$$E(S_S^2) = E(\frac{1}{n}(\sum x^2 - n\bar{x}^2)) \; ; \; E(S_S^2) = \frac{1}{n}E(\sum x^2) - E(n\bar{x}^2))$$
$$E(S_S^2) = \frac{1}{n}\left(\sum E(x^2) - nE(\bar{x}^2)\right)$$

Realizaremos un cálculo auxiliar:

$Var\,(x) = \sigma^2 = E(x-\mu)^2$;

$$\sigma^2 = E(x^2 - 2x\mu + \mu^2)$$
$$\sigma^2 = E(x^2) - E(2x\mu) + E(\mu^2)$$
$$\sigma^2 = E(x^2) - 2\mu E(x) + E(\mu^2)$$
$$\sigma^2 = E(x^2) - 2\mu\mu + \mu^2$$
$$\sigma^2 = E(x^2) - 2\mu^2 + \mu^2$$
$$\sigma^2 = E(x^2) - \mu^2$$
$$\sigma^2 + \mu^2 = E(x^2)$$

Retomando la demostración anterior:

$E(S_S^2) = \frac{1}{n}(\sum E(x^2) - nE(\bar{x}^2))$, reemplazando e identificando si se trata de la variable o la media:

$$E(S_S^2) = \frac{1}{n}\left(\sum(\mu_x^2 + \sigma_x^2) - n(\mu_{\bar{x}}^2 + \sigma_{\bar{x}}^2)\right)$$
$$E(S_S^2) = \frac{1}{n}\left(\sum \mu_x^2 + \sum \sigma_x^2 - n\mu_{\bar{x}}^2 + n\sigma_{\bar{x}}^2\right)$$
$$E(S_S^2) = \frac{1}{n}(n\mu_x^2 + n\sigma_x^2 - n\mu_{\bar{x}}^2 + n\sigma_{\bar{x}}^2)$$

$E(S_S^2) = \mu_x^2 + \sigma_x^2 - \mu_{\bar{x}}^2 + \sigma_{\bar{x}}^2$, como ya se explicó en el Capítulo 7, por Esperanza y varianza de la media:

$$E(S_S^2) = \mu_x^2 + \sigma_x^2 - \mu_x^2 + \frac{\sigma_x^2}{n}$$

$E(S_S^2) = \sigma_x^2 - \frac{\sigma_x^2}{n}$,

$$E(S_S^2) = \sigma_x^2(1 - \frac{1}{n})$$

$E(S_S^2) = \sigma_x^2(\frac{n-1}{n})$; por lo que se nota que el valor esperado de la varianza sesgada es menor a la varianza poblacional, siendo $\frac{n-1}{n}$ el sesgo.

Entonces para corregir el sesgo se debe multiplicar a la varianza sesgada por $\frac{n}{n-1}$ de modo tal que:

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} * \frac{n}{n-1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

2) Consistencia

Sea $\hat{\theta}$ un estimador de θ , n el número de unidades de la muestra y k un número positivo cualquiera, se dice que $\hat{\theta}$ es un estimador consistente de θ si se cumple que:

$$P_{\lim_{n\to\infty}} \quad (|\hat{\theta} - \theta| < k) = 1$$

Dicho de otro modo, un estimador es consistente cuando a medida que se incrementa el n se hace más similar al parámetro.

Ejemplo 8.3: Supongamos que se desconoce el valor de la media poblacional de la concentración de Cadmio en una mina. Se utiliza a $\bar{x}$ como un estimador insesgado de μ, y a medida que se incrementa el n, cada vez a $\bar{x}$ se hace más parecido a μ. Cuando n pasa a ser N (es decir se realizó un censo, la diferencia entre $\bar{x}$ y μ pasa a ser cero.

3) Eficiencia

Sean $\hat{\theta}_1$ y $\hat{\theta}_2$, ambos estimadores insesgados y consistentes de θ, se dice que $\hat{\theta}_1$ es un estimador eficiente con respecto a $\hat{\theta}_2$ si se cumple que:

$$Var\ (\hat{\theta}_1) < Var\ (\hat{\theta}_2)$$

Ejemplo 8.4: Está demostrado ya que la media muestral es un estimador insesgado de la media poblacional. También bajo condiciones de distribución normal se ha demostrado que la Mediana muestral es un estimador insesgado y consistente de la media poblacional. La pregunta es cuál es mejor? La respuesta es: la media muestral, ¿pero por qué?, porque ya demostramos que la varianza de la media muestral es $\frac{\sigma^2}{n}$, mientras que en trabajos estadísticos se ha demostrado que la varianza de la Mediana es $1{,}57 * \frac{\sigma^2}{n}$.Esto significa que a un n determinado, la media muestral varía menos que la mediana, por lo que es más eficiente.

Estimación por intervalos.

Un estimador puntual no necesariamente tendrá exactamente el mismo valor que el parámetro que está estimando. Es más, lo más probable es que no tenga ese valor. El valor que se obtenga de $\hat{\theta}$, estará más o menos alejado de θ, dependiendo de la varianza del estimador y del n de la muestra. Al realizar una estimación puntual no se tiene ningún indicio de cuán alejado está el estimador del parámetro. Por esa razón suele ser necesaria realizar también una estimación por intervalos, es decir plantear un intervalo que posea cierta probabilidad definida de antemano de contener el valor del parámetro.

$P(LI < \theta < LS) = 1 - \alpha$,

A continuación vamos a ver varios casos desde los más sencillos a los más complejos:

Caso 1) Intervalos para la media poblacional, con σ^2 conocida.

Parámetro a estimar: μ

Estimador: $\bar{x}$, y $\bar{x} \sim N\left(\mu; \frac{\sigma^2}{n}\right)$

σ^2conocida.

Definimos el estadístico como la función que relaciona al estimador con el parámetro, en este caso:

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

Definimos a una cierta probabilidad que llamaremos confianza (probabilidad del intervalo) y denotaremos como 1-α, así una doble igualdad queda:

$P(Z_1 < Z < Z_2) = 1 - \alpha$, para no confundir con los subíndices de Z, los obviaremos.

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

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

$P\left(\bar{x} - Z\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + Z\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$,

Nótese que por comodidad y debido a que la distribución Z es simétrica no se trabajó con el subíndice de la probabilidad acumulada en los valores Z.

Nótese también que de este intervalo planteado se obtiene un límite inferior (LI) y uno superior (LS) que conforman un intervalo con una probabilidad 1- α de contener al valor del parámetro μ. La construcción se realiza sumando y restándole al estimador puntual la confianza por su error estándar:

$$LI\ (1 - \alpha) = \bar{x} - Z * \frac{\sigma}{\sqrt{n}}$$

$$LS\ (1-\alpha) = \bar{x} + Z * \frac{\sigma}{\sqrt{n}}$$

Desde el punto de vista práctico, una vez obtenidos los valores, ya deja de ser una probabilidad y pasa a denominarse confianza

Caso 2) Intervalos para la media poblacional, con σ^2 desconocida.

Parámetro a estimar: μ
Estimador: $\bar{x}$
σ^2desconocida.
Estadístico: $t = \frac{\bar{x}-\mu}{\frac{S}{\sqrt{n}}}$
Confianza: 1-α,
El intervalo será:
$P(t_1 < t < t_2) = 1-\alpha$, para no confundir con los subíndices de t, los obviaremos.

$$P\left(t < \frac{\bar{x}-\mu}{\frac{S}{\sqrt{n}}} < t\right) = 1-\alpha\ ,$$

$$P\left(t\frac{S}{\sqrt{n}} < \bar{x}-\mu < t\frac{S}{\sqrt{n}}\right) = 1-\alpha\ ,$$

$$P\left(\bar{x}-t\frac{S}{\sqrt{n}} < \mu < \bar{x}+t\frac{S}{\sqrt{n}}\right) = 1-\alpha\ ,$$

Entonces se obtiene un límite inferior (LI) y uno superior (LS) que conforman un intervalo con una probabilidad 1-α de contener al valor del parámetro μ

$$LI\ (1-\alpha) = \bar{x} - t * \frac{S}{\sqrt{n}}$$

$$LS\ (1-\alpha) = \bar{x} + t * \frac{S}{\sqrt{n}}$$

Caso 3) Intervalo para la diferencia de dos medias poblacionales, con σ_1^2 y σ_2^2 conocidas

Parámetro a estimar: $\mu_1 - \mu_2$
Estimador: $\overline{X_1} - \overline{X_2}$
σ_1^2yσ_2^2 conocidas
Estadístico: $Z = \frac{(\overline{X_1}-\overline{X_2})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$
Confianza: 1-α,
El intervalo será:

$P(Z_1 < Z < Z_2) = 1-\alpha$, para no confundir con los subíndices de Z, los obviaremos.

$$P\left(Z < \frac{(\overline{X_1}-\overline{X_2})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} < Z\right) = 1-\alpha\,,$$

$$P\left(Z\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} < \overline{(X_1-X_2)} - (\mu_1-\mu_2) < Z\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right) = 1-\alpha\,,$$

$$P\left(\overline{(X_1-X_2)} - Z\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} < \mu_1-\mu_2 < \overline{(X_1-X_2)} + Z\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right) = 1-\alpha\,,$$

Los límites del intervalo quedan:

$$LI\ (1-\alpha) = \overline{(X_1-X_2)} - Z * \sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

$$LS\ (1-\alpha) = \overline{(X_1-X_2)} + Z * \sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

Caso 4) Intervalo para la diferencia de dos medias poblacionales, con σ_1^2 y σ_2^2 desconocidas, pero supuestas iguales y $n_1 = n_2$

Parámetro a estimar: $\mu_1 - \mu_2$
Estimador: $\overline{x_1} - \overline{x_2}$
σ_1^2yσ_2^2 desconocidasy $n_1 = n_2$
Estadístico: $t = \frac{(\overline{X_1}-\overline{X_2})-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$
Confianza: 1-α,
El intervalo será:

$P(t_1 < t < t_2) = 1-\alpha$, para no confundir con los subíndices de t, los obviaremos.

$$P\left(t < \frac{(\overline{X_1}-\overline{X_2})-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} < t\right) = 1-\alpha\,,$$

$$P\left(t\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}} < \overline{(X_1-X_2)} - (\mu_1-\mu_2) < t\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}\right) = 1-\alpha\,,$$

$$P\left(\overline{(X_1-X_2)} - t\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}} < \mu_1-\mu_2 < \overline{(X_1-X_2)} + t\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}\right) = 1-\alpha\,,$$

Los límites del intervalo quedan:

$$LI\ (1-\alpha) = \overline{(X_1} - \overline{X_2)} - t * \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

$$LS\ (1-\alpha) = \overline{(X_1} - \overline{X_2)} + t * \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

Caso 5) Intervalo para la diferencia de dos medias poblacionales, con σ_1^2 y σ_2^2 desconocidas, pero supuestas iguales y $n_1 \neq n_2$

Parámetro a estimar: $\mu_1 - \mu_2$
Estimador: $\overline{x_1} - \overline{x_2}$
σ_1^2 y σ_2^2 desconocidas, pero supuestas iguales y $n_1 \neq n_2$, por lo que es necesario definir una varianza ponderada o varianza amalgamada (S_A^2), que es una varianza promedio entre las dos varianzas muestrales, pero ponderadas por los grados de libertad de cada muestra:

$$S_A^2 = \frac{(n_1 - 1) * S_1^2 + (n_2 - 1) * S_2^2}{(n_1 + n_2 - 2)}$$

Estadístico: $t = \frac{\overline{(X_1} - \overline{X_2)} - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1-1)*S_1^2+(n_2-1)*S_2^2}{(n_1+n_2-2)} * (\frac{1}{n_1} + \frac{1}{n_2})}}$

Confianza: 1-α,
El intervalo será:

$P(t_1 < t < t_2) = 1 - \alpha$, para no confundir con los subíndices de t, los obviaremos.

$$P\left(t < \frac{\overline{(X_1} - \overline{X_2)} - (\mu_1 - \mu_2)}{\sqrt{S_A^2\ (\frac{1}{n_1} + \frac{1}{n_2})}} < t\right) = 1 - \alpha\ ,$$

$$P\left(t\sqrt{S_A^2\ (\frac{1}{n_1} + \frac{1}{n_2})} < \overline{(X_1} - \overline{X_2)} - (\mu_1 - \mu_2) < t\sqrt{S_A^2\ (\frac{1}{n_1} + \frac{1}{n_2})}\right) = 1 - \alpha\ ,$$

$$P\left(\overline{(X_1} - \overline{X_2)} - t\sqrt{S_A^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} < \mu_1 - \mu_2 < \overline{(X_1} - \overline{X_2)} + t\sqrt{S_A^2\ (\frac{1}{n_1} + \frac{1}{n_2})}\right) = 1 - \alpha\ ,$$

Los límites del intervalo quedan:

$$LI\ (1-\alpha) = \overline{(X_1} - \overline{X_2)} - t * \sqrt{\frac{(n_1-1)*S_1^2+(n_2-1)*S_2^2}{(n_1+n_2-2)} * (\frac{1}{n_1} + \frac{1}{n_2})}$$

$$LS\ (1-\alpha) = \overline{(X_1} - \overline{X_2)} + t * \sqrt{\frac{(n_1-1)*S_1^2+(n_2-1)*S_2^2}{(n_1+n_2-2)} * (\frac{1}{n_1} + \frac{1}{n_2})}$$

Caso 6) Intervalo para la varianza poblacional.

Parámetro a estimar: σ^2
Estimador: $\overline{S^2}$
Estadístico: $\chi^2_{n-1} = \frac{(n-1)S^2}{\sigma^2}$

Confianza: 1-α,
El intervalo será:
$P(\chi_1^2 < \chi^2 < \chi_2^2) = 1 - \alpha$, para no confundir con los subíndices de Chi, los obviaremos.

$$P\left(\chi^2 < \frac{(n-1)S^2}{\sigma^2} < \chi^2\right) = 1 - \alpha ,$$

$$P\left(\chi^2 < \frac{(n-1)S^2}{\sigma^2} < \chi^2\right) = 1 - \alpha ,$$

$$P\left(\frac{\chi^2}{(n-1)S^2} < \frac{1}{\sigma^2} < \frac{\chi^2}{(n-1)S^2}\right) = 1 - \alpha ,$$

$$P\left(\frac{(n-1)S^2}{\chi^2} > \sigma^2 > \frac{(n-1)S^2}{\chi^2}\right) = 1 - \alpha ,$$

Los límites del intervalo quedan:
$LI\ (1-\alpha) = \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}$
$LS\ (1-\alpha) = \frac{(n-1)S^2}{\chi^2_{\alpha/2}}$

Caso 7) Intervalo para el cociente de dos varianzas poblacionales.

Parámetro a estimar: $\frac{\sigma_1^2}{\sigma_2^2}$
Estimador: $\frac{S_1^2}{S_2^2}$
Estadístico: $F_{(n_1-1;n_2-1)} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$

Confianza: 1-α,

El intervalo será:
$P(F_1 < F < F_2) = 1 - \alpha$,

$$P\left(F < \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} < F\right) = 1 - \alpha ,$$

$$P\left(F\frac{S_1^2}{S_2^2} < \frac{\sigma_1^2}{\sigma_2^2} < F\frac{S_1^2}{S_2^2}\right) = 1 - \alpha\ ,$$

Los límites del intervalo quedan:

$$LI\ (1-\alpha) = F_{(\frac{\alpha}{2};n_1-1;n_2-1)}\frac{S_1^2}{S_2^2}$$

$$LS\ (1-\alpha) = F_{(1-\frac{\alpha}{2};n_1-1;n_2-1)}\frac{S_1^2}{S_2^2}$$

Cálculo de tamaño de la muestra.

En el capítulo se explicó que para obtener muestras representativas de la población éstas debían ser tomadas de un modo probabilístico, es decir se explicó el cómo. Aquí se comentará el cuánto. Cuántas unidades debemos tener para que la muestra represente la variabilidad de la población.
Está claro que si todas las unidades fuesen exactamente iguales sólo debiéramos tomar una unidad o elemento. A mayor variabilidad, mayor número de unidades serán necesarias.

1) **Variables cuantitativas normales**

A partir del Estadístico: $t = \frac{\bar{x}-\mu}{\frac{S}{\sqrt{n}}}$
Y despejando n, se puede obtener:

$$n_{min} = \frac{t^2 S^2}{e^2}$$

Donde:

S^2: varianza muestral
t: valor del estadístico t para n-1 y 1-α.
e: error máximo que se está dispuesto a cometer. Se expresa en la unidad de la variable, por ejemplo un error máximo del 10 % corresponde al valor e= 0.1* $\bar{x}$

Ejemplo 8.5. Se desea estimar a la media poblacional del contenido de sales minerales (en %) del agua en el mar en una zona cercana a la Costa de la provincia de Chubut. Se cuenta con una muestra piloto de 20 muestras con los siguientes valores: $\bar{x}$ = 3,5 %; S = 0, 05%. Se desea trabajar con una confianza del 95% y un error no mayor al 1%.

Si e<1%→ 0,01 * 3,5 = 0,035.
Si 1-α= 0,95 y n-1= 19 → valor t de tabla: 2,09, entonces:

$$n = \frac{2{,}09^2 0{,}05^2}{0{,}035^2} = 8{,}91 \cong 9$$

Por lo que para estimar a la media poblacional se deben tomar 9 elementos en la muestra.

A partir de la fórmula anterior, se observa que a mayor precisión (menor error) el número aumenta, que a mayor confianza el número aumenta y a mayor varianza, el número aumenta.

2) Variables nominales (aproximación de la Binomial a la Normal)

A partir del Estadístico: $t = \frac{np-NP}{\sqrt{npq}}$
Y despejando n, se puede obtener:

$$n_{min} = \frac{t^2\, pq}{e^2}$$

Donde:

p: proporción de éxitos de la muestra
q= (1-p): proporción de fracasos en la muestra
t: valor del estadístico t para n-1 y 1-α., si se trabaja con n>30, entonces es el valor Z de la distribución normal estándar.
e: error máximo que se está dispuesto a cometer. Se expresa en la unidad de la variable, por ejemplo un error máximo del 10 % corresponde al valor e= 0.1.

Nótese que como la máxima varianza se obtiene con p=0,5, si se desconoce el valor de la varianza poblacional o no existe una muestra piloto, se puede estimar a p=0,5 de un modo conservador para trabajar con la máxima varianza posible.

Ejemplo 8.6.
Se desea saber qué número de personas se deberá encuestar para conocer la intención de voto de una población. Si se desconoce p, entonces
p=0,5; q= 0,5; e= 0,1 y 1-α= 0,95 y → valor Z de tabla: 1,96, entonces:

$$n_{min} = \frac{1{,}96^2\ 0{,}5\ 0{,}5}{0{,}1^2} = 96{,}04 \cong 96$$

Por lo que para estimar la proporción poblacional con una confianza del 95% y un error máximo del 10% se deben encuestar 96 personas.

Capítulo 9.

Pruebas de Hipótesis

Introducción.

Antes que avanzar en este capítulo debemos definir qué es para nosotros una hipótesis. Vamos a definirla como una conjetura, una aseveración de cómo funciona el sistema, es una frase afirmativa.

Ejemplo 9.1: Esta noche no llueve sobre mi casa.
Ejemplo 9.2: En el próximo examen de Estadística me sacaré un 10.

En estos ejemplos se observa que transcurrido cierto tiempo, se pueden poner a prueba esas hipótesis. Transcurrida la noche o después del examen, se puede constatar si era o no cierta la hipótesis. Cuando decimos si era o no cierta es porque existe la posibilidad de que la hipótesis sea falsa.
Vamos a plantear de nuevo los ejemplos con las dos posibilidades, llamándole Hipótesis nula (H_0) a la aseveración a los resultados posibles donde ésta es cierta e Hipótesis alternativa (H_1) al complemento de ésta, es decir a todos los resultados cuando demuestren que la Hipótesis nula sea falsa

Ejemplo 9.1:
H_0) Esta noche no llueve sobre mi casa.
H_1) Esta noche llueve sobre mi casa.
Esta noche puede llover mucho o poco, en cualquier caso la hipótesis nula será falsa

Ejemplo 9.2:
H_0) En el próximo examen de Estadística me sacaré un 10.
H_1) En el próximo examen de Estadística no me sacaré un 10.
Si en el examen me saco un 9,75 o un 3, en ambos casos la hipótesis nula será falsa.

Si se tuviese a priori certeza del resultado, no sería una hipótesis, la prueba de hipótesis es una experiencia aleatoria que arroja un espacio muestral, espacio que dividimos en 2, una parte donde corresponde a la hipótesis nula y la otra a la alternativa.

Ya planteamos lo que se denomina la batería de hipótesis (planteo de hipótesis nula y alternativa, pero en los ejemplos 9.1 y 9.2 era sencilla la prueba porque se podía contar con un censo para la prueba, la noche determinada o el examen determinado arrojan sólo un resultado posible.
Qué pasa ahora cuando planteamos otras dos hipótesis:

Ejemplo 9.3: Todos los gatos tienen 4 patas.
Ejemplo 9.4: En toda una formación determinada el cuarzo se encuentra en un 45%.

¿Cómo hago para contar las patas de todos los gatos del mundo o para revisar toda la formación que puede tener millones de toneladas de roca? La respuesta es **No lo voy a poder hacer,** voy a tener que tomar una muestra que represente la población y extrapolar las conclusiones de la muestra a la población, pero como no tengo al TODO, voy a tener que contemplar una tasa de errores (error en este caso entendido como equivocación).

Hay muchísimas pruebas de hipótesis, y todos los días se inventan más, vamos a ver sólo algunos casos, el primer caso lo vamos a desarrollar con un ejemplo y los conceptos generales los trataremos después de este caso para que resulte más didáctico.

Caso 1) Prueba de hipótesis para la media poblacional, con σ^2 conocida.

Ejemplo 9.5. Un Geólogo necesita saber si el promedio poblacional de la ley de Oro en una zona es mayor o igual a 12 g/Tn. Para eso realiza una serie de excavaciones y muestreando sistemáticamente toma una muestra de tamaño n (como este es un ejemplo práctico y el caso 1 es un caso teórico, necesitamos asumir inocentemente que se conoce el valor de la varianza poblacional).

$$H_0) \mu \geq 12$$
$$H_1) \mu < 12$$

Ahora bien, esto no sería un libro de estadística si corroborásemos matemáticamente el resultado y si la media muestral es exactamente menor que 12 rechazamos la hipótesis nula. Esta acción la podríamos realizar sólo si tuviésemos el verdadero valor deμ. Como no lo tenemos y sólo poseemos un estimador, debemos razonar de forma diferente:
Pensemos que por más que la hipótesis nula sea cierta, por más que $\mu \geq 12$,la media muestral sólo por azar podría ser más pequeña que 12 (por el azar de cuáles fueron los elementos que ingresaron en la muestra), entonces debiéramos permitirle a la media muestral que se "mueva" sólo por azar alrededor del 12, dicho de otro modo, si la media muestral está "cerquita" del 12, debiéramos considerarlo que no es "significativamente" diferente de 12, debiéramos considerar que se diferenció de 12 sólo por azar. Pero si la media muestral está "muy" alejada del 12, ya debiéramos decir que no es sólo producto del azar, sino que debiéramos decir que la media muestral está alejada de 12, porque la hipótesis nula es falsa.
¿Cuán alejada debiéramos considerar a la media muestral del valor 12 para que sea lejos?

Esa pregunta ya la trataron muchísimas reuniones entre especialistas, y hace casi 100 años un grupo de ellos decidieron por consenso propiciar que si un evento se desarrollaba con una probabilidad menor a 1/20 podía ser considerado un evento raro, pero si estaba en más del 1/20 de los caso era un evento común. De allí y por consenso surgió la idea del 5%. Un evento con una probabilidad menor a 0,05 sería considerado raro y si está dentro del 95% (p=0,95) es un evento común.

Volviendo al tema de la media muestral, si está entonces más alejada del 95% de los casos, entonces es una media rara!

Veamos en la Figura 9.1a, allí se plantea la distribución de las medias muestrales si la hipótesis nula es cierta, entonces ¿cuál sería la zona donde no se va a rechazar la hipótesis nula con una probabilidad de 1- α, (que por consenso sería 0,95) y la zona donde se la rechazaría con una probabilidad (α) (que en este caso sería de 0,05)?,
Recordemos que estamos hablando en la Figura 9.1a de la distribución de la media muestral y no la de la variable x) que se presenta en la Figura 9.1b. La Figura 9.1c representa exactamente la figura 9.1a, pero con las zonas destacadas y la Figura 9.1d representa la misma, pero ya estandarizada.

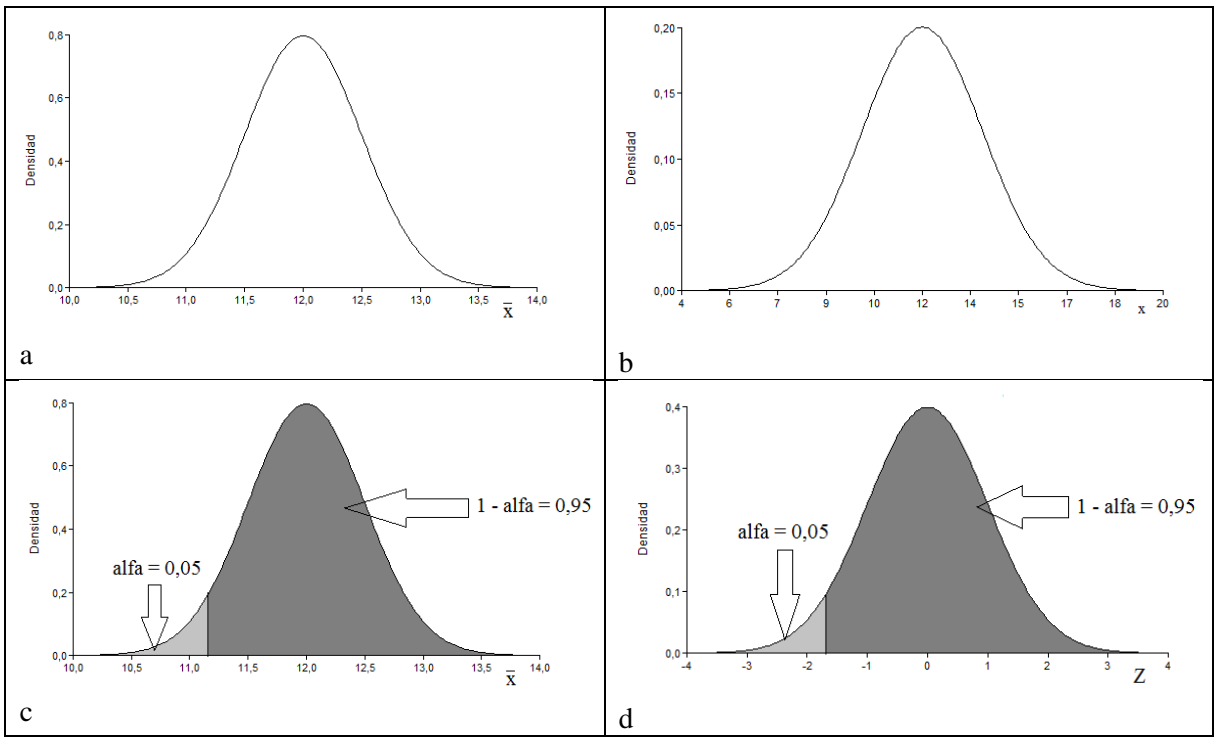


Figura 9.1. Distribución de la ley de oro del Ejemplo 9.5.a: Distribución de la media muestral si la H_0 es cierta, b: Distribución de la variable x si la H_0 es cierta, c: Distribución de la media muestral con las zonas de rechazo y no rechazo, d: Distribución de z con las zonas de rechazo y no rechazo

Resultados posibles:
Cuatro tipos de resultados diferentes pueden ocurrir y vamos a describirlos:

I) Para los dos primeros supongamos que la hipótesis nula sea cierta, que aunque nosotros no lo conozcamos, la $H_0)\mu \geq 12$ es cierta:

Ejemplo 9.5.a. Se toma la muestra, se calcula le media muestral y ésta arroja un valor de 11,5 g/Tn, lo que se ubica en la Figura 9.2 del siguiente modo:

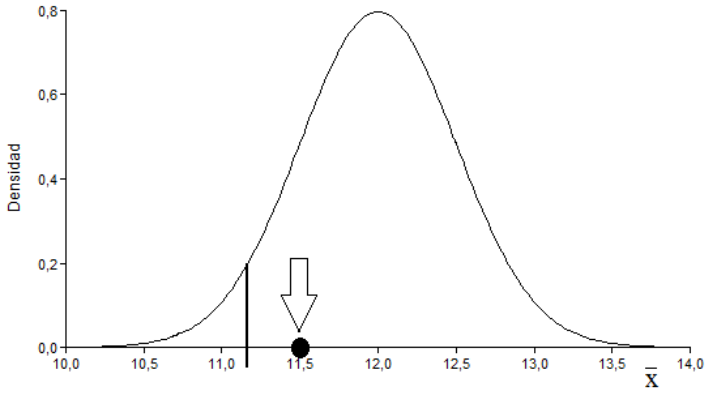


Figura 9.2. Ubicación de un valor de media muestral de 11,5 en el caso de Ejemplo 9.5, cuando la Hipótesis nula es cierta.

Entonces la decisión sería no rechazar H_0. No rechazar la H_0 cuando ésta es cierta es algo bueno, se denomina confianza y su probabilidad es de 1-α.

Ejemplo 9.5.b. Se toma la muestra, se calcula le media muestral y ésta arroja un valor de 11,02 g/Tn, lo que se ubica en la Figura 9.3 del siguiente modo:

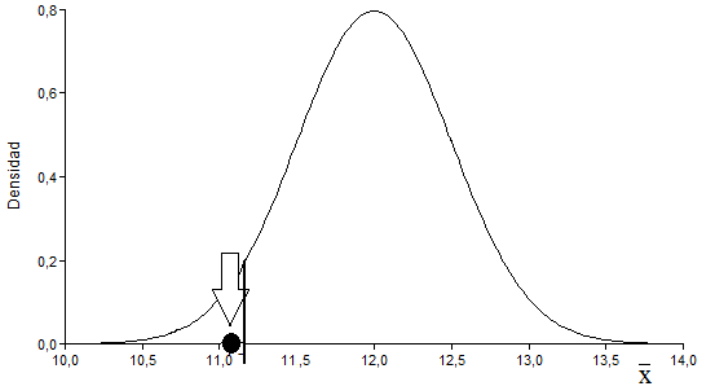


Figura 9.3.Ubicación de un valor de media muestral de 11,02 en el caso de Ejemplo 9.5, cuando la Hipótesis nula es cierta.

Entonces la decisión sería rechazar H_0. Rechazar la H_0 cuando ésta es cierta es algo malo, es una equivocación, se denomina Error de Tipo I y su probabilidad es de α.

II) Para los dos restantes supongamos que la hipótesis nula sea falsa, que aunque nosotros no lo conozcamos, la $H_0)\mu \geq 12$ es falsa. Pero necesitamos para graficar el ejemplo colocarle el valor de media a la verdadera μ. Entonces colocaremos a la verdadera $\mu = 10$

Ejemplo 9.5.c. Se toma la muestra y ésta arroja el valor de la media muestral de 9,5, lo que se ubica en la figura 9.4 del siguiente modo:

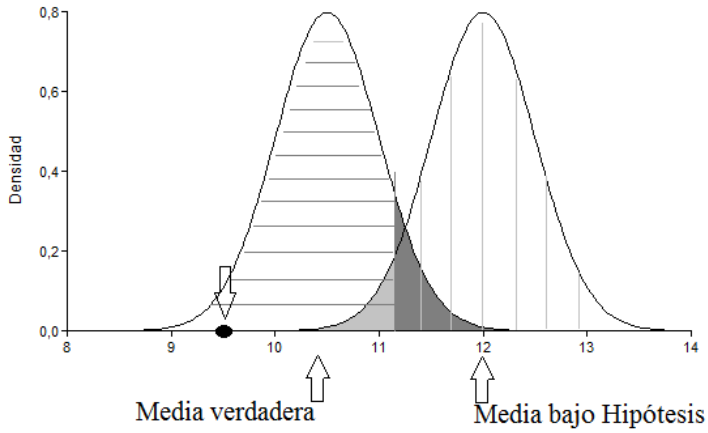


Figura 9.4. Ubicación de un valor de media muestral de 9,5 en el caso de Ejemplo 9.5, cuando la Hipótesis nula es falsa.

Entonces la decisión sería rechazar H_0. Rechazar la H_0 cuando ésta es falsa es algo bueno, se denomina potencia y su probabilidad es de 1-β. Nótese que toda la superficie con rayas horizontales constituye la potencia.

Ejemplo 9.5.d. Se toma la muestra y ésta arroja el valor de la media muestral de 11,9, lo que se ubica en la figura 9.5 del siguiente modo:

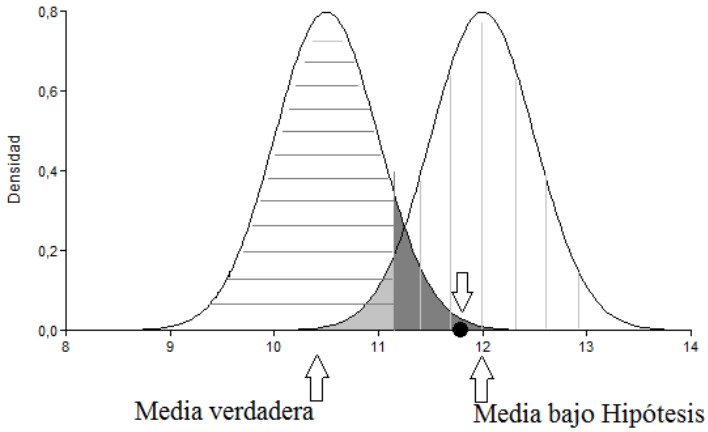


Figura 9.5.Ubicación de un valor de media muestral de 11,9 en el caso de Ejemplo 9.5, cuando la Hipótesis nula es falsa.

Entonces la decisión sería no rechazar H_0. No rechazar la H_0 cuando ésta es falsa es algo malo, se denomina Error de tipo II y su probabilidad es deβ. Nótese que la superficie pintada en gris oscuro constituye el Error de tipo II.

Resumiendo entonces en una tabla los cuatro resultados posibles:

	Si la H_0 es cierta	Si la H_0 es falsa
No se rechaza H_0	Confianza 1-α.	Error de tipo II β
Se rechaza H_0	Error de tipo I α.	Potencia 1-β

Tabla 9.1. Cuadro de las cuatro posibles resultados en una prueba de errores.

Entonces el Geólogo antes de tomar la muestra posee cuatro resultados posibles. Una vez que ya tomó la muestra, realizó la prueba y concluyó, va a tener dos posibilidades:

I) Si rechazó H_0, no sabe si está bien rechazado y está en zona de Potencia o si se trata de un error y está en zona de Error de tipo I.

II) Si no rechazó H_0, no sabe si está bien no rechazarlo porque es cierta la Hipótesis nula y está en zona de confianza o si se trata de un error y está en la zona del error de tipo II.

La pregunta es ¿cuándo se tendrá la certeza del resultado? La única certeza se consigue sólo cuando se accede al valor de la media poblacional, es decir cuando se realiza un censo. Si no se realiza un censo, la respuesta es: NUNCA, la única certeza que tienen los investigadores es que la estadística no trabaja con certezas, sino en que asevera una verdad sólo con un alto grado de certidumbre.

Nótese que nunca nos referimos a que se acepta la Hipótesis nula, las acciones posibles son dos: o Rechazamos o No rechazamos H_0.

Volvamos al ejemplo 9.5 y pongamos valores para realizar finalmente la prueba:

$$H_0) \mu \geq 12$$
$$H_1) \mu < 12$$

Parámetro a poner a prueba: μ

Estimador: $\bar{x}$= 11,95 y $\bar{x} \sim N\left(\mu; \frac{\sigma^2}{n}\right)$

σ^2se conoce que =4

Estadístico: $Z_0 = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$

Confianza: 0,95.

n: 64

Regla de decisión:

Si $(Z_\alpha < Z_0 < \infty)$, entonces no se rechaza $H_0) \mu \geq 12$

Si $(-\infty < Z_0 < Z_\alpha]$, entonces se rechaza $H_0) \mu \geq 12$

Análisis:

$$Z_0 = \frac{11{,}95 - 12}{\frac{2}{\sqrt{64}}} = -0.2$$

En la búsqueda de la tabla de la distribución normal estandarizada se observa que el valor de Z que acumula desde menos infinito hasta él un 0,05 de probabilidades es el valor: -1,6449, por lo tanto:
Zona de No rechazo: $(-1{,}6449 < Z_0 < \infty)$,
Zona de Rechazo: $(-\infty < Z_0 < -1{,}6449]$, entonces como Z_0 es: -0,2, no se rechaza $H_0)\ \mu \geq 12$

En las Figuras 9.6 se presentan los valores de la media muestral (11,95) y su valor estandarizado (-0,2) en sus respectivas distribuciones.

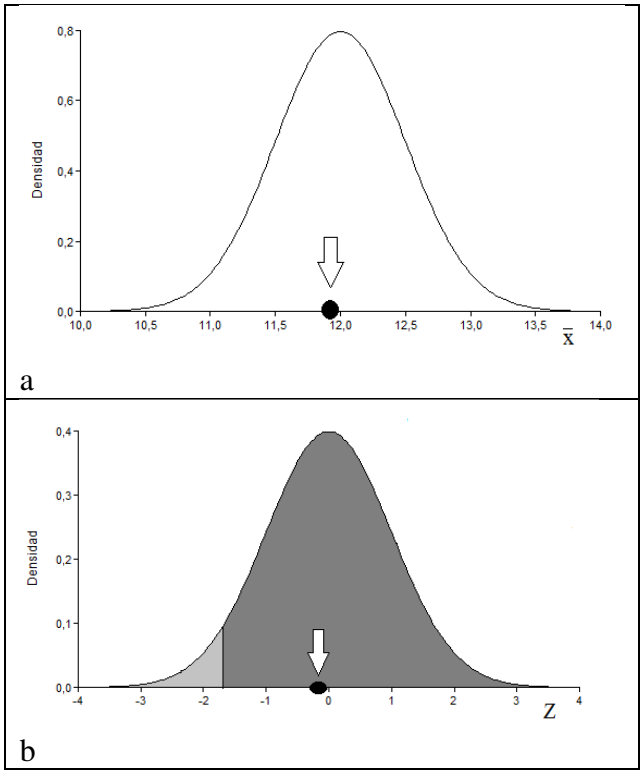


Figura 9.6.Ubicación del valor de media muestral de 11,95 (a) y su valor estandarizado (b) para la resolución del Caso 1 con el Ejemplo 9.5.

Baterías de posibles Hipótesis

Dependiendo de la hipótesis planteada, se puede recurrir a tres tipos diferentes de baterías de hipótesis (Figuras 9.7).

$H_0)\mu \geq 12$	$H_0)\mu \leq 12$	$H_0)\mu = 12$
$H_1)\ \mu < 12$	$H_1)\ \mu > 12$	$H_1)\ \mu \neq 12$
Prueba Unilateral Izquierda	Prueba unilateral derecha	Prueba bilateral

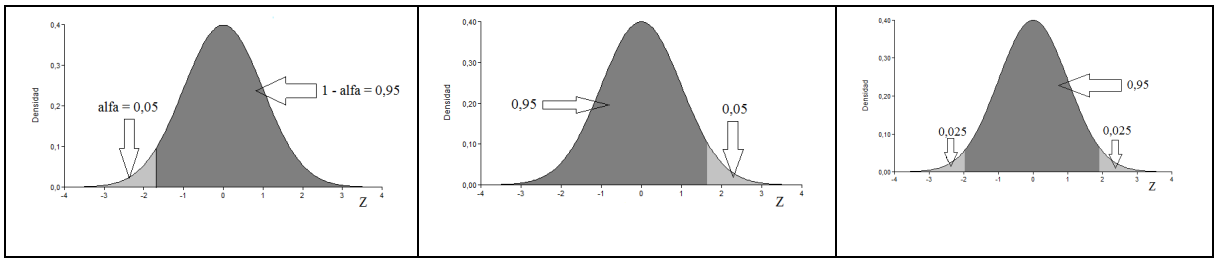


Figura 9.7. Casos posibles de baterías de hipótesis.

Caso 2) Prueba de hipótesis para la media poblacional, con σ^2 desconocida.

Ejemplo 9.6. Se quiere saber si cambió el pH de un Lago que hasta la década anterior poseía un pH de 7,3. Se tomó una muestra de 27 alícuotas de agua, arrojando una media muestral de 7.25 y una varianza muestral de 0,02.

$$H_0)\ \mu = 7{,}3$$
$$H_1)\ \mu \neq 7{,}3$$

Parámetro a poner a prueba: μ

Estimador: $\bar{x}$= 7,25 y $\bar{x} \sim N\left(\mu;\ \frac{\sigma^2}{n}\right)$

σ^2 = desconocido, estimado con S^2= 0,02

Estadístico: $t_0 = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$

Confianza: 1-α: 0,95

n: 27

Regla de decisión:

Si $\left(t_{n-1;\alpha/2} < t_0 < t_{n-1;1-\alpha/2}\right)$, entonces no se rechaza $H_0)\ \mu = 7{,}3$

Si $(-\infty < t_0 < t_{n-1;\alpha/2}]$, ó $[t_{n-1;1-\alpha/2} < t_0 < \infty$, entonces se rechaza $H_0)\ \mu = 7{,}3$

Para este caso entonces:

$$t_0 = \frac{7{,}25 - 7{,}3}{\frac{0{,}1414}{\sqrt{27}}} = -1.8373$$

Como (-2,055 < 1.8373 < 2,055) entonces no se rechaza la Hipótesis nula, el lago no cambió significativamente su pH.

Caso 3) Pruebas para la igualdad entre dos medias poblacionales, muestras independientes y σ_1^2 y σ_2^2 conocidas.

$$H_0)\ \mu_1 = \mu_2 \rightarrow \mu_1 - \mu_2 = 0$$
$$H_1)\ \mu_1 - \mu_2 \neq 0$$

Parámetro a poner a prueba: $\mu_1 - \mu_2$
Estimador: $\overline{x_1} - \overline{x_2}; \bar{x} - \overline{x_2} \sim N\left(\mu_1 - \mu_2;\ \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$
$\sigma_1^2 y \sigma_2^2$ conocidas
Estadístico: $Z = \frac{\overline{(X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$
Confianza: 1-α,
$n_1 y\ n_2$

Si $\left(Z_{\alpha/2} < Z_0 < Z_{1-\alpha/2}\right)$, entonces no se rechaza H_0)
Si $(-\infty < Z_0 < Z_{\alpha/2}]$, ó $(Z_{1-\alpha/2} < Z_0 < \infty)$, entonces se rechaza H_0)

Caso 4) Pruebas para la igualdad entre dos medias poblacionales, muestras independientes, σ_1^2 y σ_2^2 desconocidas y $n_1 = n_2$.

$$H_0)\ \mu_1 - \mu_2 = 0$$
$$H_1)\ \mu_1 - \mu_2 \neq 0$$

Parámetro a poner a prueba: $\mu_1 - \mu_2$
Estimador: $\overline{x_1} - \overline{x_2}; \bar{x} - \overline{x_2} \sim N\left(\mu_1 - \mu_2;\ \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$
$\sigma_1^2 y \sigma_2^2$ desconocidas
Estadístico: $t = \frac{\overline{(X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$
Confianza: 1-α,
$n_1 = n_2$

Regla de decisión:
Si $\left(t_{n-1;\alpha/2} < t_0 < t_{n-1;1-\alpha/2}\right)$, entonces no se rechaza H_0)
Si $(-\infty < t_0 < t_{n-1;\alpha/2}]$, ó $[t_{n-1;1-\alpha/2} < t_0 < \infty)$, entonces se rechaza H_0)

Caso 5) Pruebas para la igualdad entre dos medias poblacionales, muestras independientes, σ_1^2 y σ_2^2 desconocidas y $n_1 \neq n_2$

$$H_0)\ \mu_1 - \mu_2 = 0$$
$$H_1)\ \mu_1 - \mu_2 \neq 0$$

Parámetro a poner a prueba: $\mu_1 - \mu_2$
Estimador: $\overline{x_1} - \overline{x_2}; \bar{x} - \overline{x_2} \sim N\left(\mu_1 - \mu_2;\ \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$
$\sigma_1^2 y \sigma_2^2$ desconocidas

Estadístico: $t = \frac{\overline{(X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1-1)*S_1^2+(n_2-1)*S_2^2}{(n_1+n_2-2)} * (\frac{1}{n_1}+\frac{1}{n_2})}}$

Confianza: 1-α,

$n_1 \neq n_2$

Regla de decisión:

Si $\left(t_{n-1;\alpha/2} < t_0 < t_{n-1;1-\alpha/2}\right)$, entonces no se rechaza H_0)

Si $(-\infty < t_0 < t_{n-1;\alpha/2}]$, ó $[t_{n-1;1-\alpha/2} < t_0 < \infty)$, entonces se rechaza H_0)

Caso 6) Pruebas para la igualdad entre dos medias poblacionales, muestras dependientes (o apareadas).

$$H_0)\ \mu_d = 0$$
$$H_1)\ \mu_d \neq 0$$

Parámetro a poner a prueba: μ_d

Estimador = $\bar{d}$ y $\bar{d} \sim N\left(\mu_d; \frac{\sigma_d^2}{n}\right)$

σ_d^2= desconocido, estimado con S_d^2

Estadístico: $t_0 = \frac{\bar{d}-\mu}{\frac{S_d}{\sqrt{n}}}$

Confianza: 1-α

n: n_0

Regla de decisión:

Si $\left(t_{n-1;\alpha/2} < t_0 < t_{n-1;1-\alpha/2}\right)$, entonces no se rechaza H_0)

Si $(-\infty < t_0 < t_{n-1;\alpha/2}]$, ó $[t_{n-1;1-\alpha/2} < t_0 < \infty)$, entonces se rechaza H_0)

Caso 7) Prueba para la varianza poblacional.

$$H_0)\ \sigma^2 = \sigma_0^2$$
$$H_1)\ \sigma^2 \neq \sigma_0^2$$

Parámetro a poner a prueba: σ^2

Estimador: $\overline{S^2}$

Estadístico: $\chi_0^2 = \frac{(n-1)S^2}{\sigma^2}$

Confianza: 1-α,

n; n_0

Regla de decisión:

Si ($\chi_{\alpha/2}^2 < \chi_0^2 < \chi_{1-\alpha/2}^2$), entonces no se rechaza H_0)

Si ($0 < \chi_0^2 < \chi_{\alpha/2}^2$)Si ($\chi_{1-\alpha/2}^2 < \chi_0^2 < \infty$)entonces se rechaza H_0)

Caso 8) Prueba para la igualdad de dos varianzas poblacionales.

$$H_0)\ \frac{\sigma_1^2}{\sigma_2^2} = 1$$
$$H_1)\ \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

Parámetro a poner a prueba: $\frac{\sigma_1^2}{\sigma_2^2}$

Estimador: $\frac{S_1^2}{S_2^2}$

Estadístico: $F_0 = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$

Confianza: 1-α,
$n_1 y\ n_2$

Regla de decisión:
Si ($F_{n_1-1;\, n_2-1;\, \alpha/2} < F_0 < F_{n_1-1;\, n_2-1;\, 1-\alpha/2}$), entonces no se rechaza H_0)

Si ($0 < F_0 < F_{n_1-1;\, n_2-1;\, \alpha/2}$)Si ($F_{n_1-1;\, n_2-1;\, 1-\alpha/2} < F_0 < \infty$)entonces se rechaza H_0)

Caso 9) Prueba de Bondad de ajuste.

Ejemplo 9.9. A los fines de observar si la litología se distribuía uniformemente en el terreno, se realizaron 30 perfiles observando si correspondía a: A, B o C. Se tomó la litología (si era A, B ó C) y se realizó una tabla de frecuencias arrojando:

Variable	Frecuencia absoluta
A	9
B	16
C	5

Tabla 9.2. Frecuencia de los tres tipos de litología encontrados

La pregunta que tiene el investigador es si la distribución es uniforme o las diferencias que se ven en sus frecuencias son sólo por azar. Para esto se aplica una prueba denominada Prueba Chi cuadrado de Bondad de Ajuste, que en este caso ajustaremos a una distribución Uniforme.

$$H_0)\ X \sim Uniforme$$
$$H_1)\ X \sim / \ Uniforme$$

Parámetro a poner a prueba: Función poblacional de la Distribución Uniforme
Estimador: Función muestral de la Distribución Uniforme
Estadístico: $\chi_0^2 = \sum_{i=1}^{k} \frac{(O-e)^2}{e}$

Donde
k: Número de categorías de la variable
O: frecuencia Observada
e: Frecuencia esperada de la variable bajo H_0) cierta
Confianza: 1-α,

Regla de decisión:
Si ($0 < \chi_0^2 < \chi^2_{k-p-1;\,1-\alpha}$), entonces no se rechaza H_0)
Si ($\chi^2_{k-p-1;\,1-\alpha} < \chi_0^2 < \infty$)entonces se rechaza H_0)

Nótese que el valor de Chi cuadrado implica la discrepancia entre la frecuencia observada y la frecuencia esperada, por lo tanto un valor de Chi cuadrado de cero implicaría igualdad absoluta. A medida que lo observado es más disímil a lo esperado el valor del estadístico va incrementándose. Por esa razón esta prueba es sólo unilateral derecha.

Variable	Frecuencia absoluta	Frecuencia Esperada	$(O-e)$	$(O-e)^2$	$\frac{(O-e)^2}{e}$
A	9	10	-1	1	0,1
B	16	10	6	36	3,6
C	5	10	-5	25	2,5

Tabla 9.3. Desarrollo del valor chi cuadrado para el ejemplo 9.9.

Por lo que $\chi_0^2 = 6{,}2$

Los grados de libertad del $\chi^2_{0,95}$ con k-p-1, donde k es la cantidad de categorías (en este caso 3), p son los parámetros estimados (en este caso 0), por lo que$\chi^2_{alfa} = 5{,}995$, por lo tanto se rechaza H_0, se concluye que las diferencias observadas no pueden ser sólo por azar.

Así como se desarrolló el ejemplo para una variable cualitativa, se puede aplicar la prueba de Bondad de Ajuste Chi cuadrado para variables cuantitativas, pero posee menor potencia que otras pruebas más precisas.

Para tales casos, cuando se desea poner a prueba una distribución Normal, la prueba de Shapiro Wilks es la más recomendada. Debido a que ésta última es sólo una prueba para Normalidad, para el ajuste de otras distribuciones se recomienda la prueba de Kolmogorov Smirnov. En ambas pruebas la hipótesis que se prueba es la de la distribución de la variable, utilizando cada una de ellas estadísticos específicos.

Caso 10) Prueba de independencia entre dos variables cualitativas.

Ejemplo 9.10.Se quiere conocer si en una empresa internacional minera está asociada el género de los empleados a su estado de complacencia por la remuneración recibida, dicho de otro modo, alguno de los géneros se percibe con peor remuneración que el otro. Se realiza un muestreo de 78 personas preguntándole a cada uno de ellos su género y si creían que estaban cobrando lo que se merecían o menos, arrojando los siguientes resultados:

	Cobra lo que se

	merece	
Género	SI	NO
Femenino	3	15
Masculino	21	39

Tabla 9.4. Frecuencia de personas según género y su complacencia por la remuneración recibida.

$H_0)\ X_1\ es\ independiente\ de\ X_2$
$H_1)\ X_1\ no\ es\ independiente\ de\ X_2$

Donde:
X_1 = Variable merecimiento de la remuneración
X_2 = Género

Parámetro a poner a prueba: Función poblacional conjunta de las variables $X_1\ y\ X_2$
Estimador: Función muestral conjunta de las variables $X_1\ y\ X_2$
Estadístico: $\chi_0^2 = \sum_{i=1}^{k} \frac{(O-e)^2}{e}$

Donde
k: Número de combinaciones de categorías de ambas variables
O: frecuencia Observada
e: Frecuencia esperada de la variable bajo $H_0)$ cierta
Confianza: 1-α,

Regla de decisión:
Si $(0 < \chi_0^2 < \chi^2_{(k1-1)*(k2-1);\,1-\alpha})$, entonces no se rechaza $H_0)$
Si $(\chi^2_{(k1-1)*(k2-1);\,1-\alpha} < \chi_0^2 < \infty)$entonces se rechaza $H_0)$

Para el cálculo de la frecuencia esperada bajo independencia, se debe recurrir a lo que hemos visto en el capítulo 4 como la propiedad 4 de las probabilidades: Para ser independientes dos eventos A y B la probabilidad conjunta debe ser igual al producto de sus probabilidades marginales. De ese modo obtenemos las probabilidades marginales y las conjuntas:

	SI	NO	Total
Fem.	3	15	18
Masc.	21	39	60
Total	26	52	78

a

	SI	NO	Marginal
Fem.			0,2308
Masc.			0,7692
Marginal	0,3333	0,6667	1

b

	SI	NO	Marginal
Fem.	0,0769	0,1538	0,2308
Masc.	0,2564	0,5128	0,7692
Marginal	0,3333	0,6667	1

	SI	NO	Total
Fem.	5,54	12,46	18
Masc.	18,46	41,54	60
Total	26	52	78

c d

Tablas 9.5. a: Frecuencias absolutas, b: probabilidades marginales, c: probabilidades esperadas por celda, d: frecuencias esperadas por celda.

$$\chi_0^2 = \frac{(3-5{,}54)^2}{5{,}54} + \frac{(21-18{,}46)^2}{18{,}46} + \frac{(15-12{,}46)^2}{12{,}46} + \frac{(39-41{,}54)^2}{41{,}54}$$

$$\chi_0^2 = 2{,}185$$

El valor χ^2_{tabla} se distribuye con $(k_1 - 1) * (k_2 - 1)$ grados de libertad, en este caso el valor para 0,95 y 1 grado de libertad es: 3,845, por lo tanto:

2,185 < 3,845 por lo que no se rechaza la H_0) X_1 $es\ independiente\ de\ X_2$.

Transformaciones de variables

Todas las pruebas descriptas anteriormente que se basan en hipotetizar parámetros necesitan como supuesto que la variable aleatoria posea distribución Normal. En el caso que la variable no posea distribución normal el autor, a los fines de realizar una prueba transforma a la variable original con alguna transformación no lineal y que no le impida después inferir sobre la variable original.
Hay familias enteras de transformaciones, siendo las más relevantes:

1) Transformación logarítmica

Sea X una variable aleatoria con distribución No normal se dice que X´ es su transformada cuando:

$X' = Log_{10}\ (X)$, o bien: $X' = Ln\ (X)$

Si la variable X posee valores cero (0), se recomienda antes de transformar sumarle un valor ya que el Log de 0 no está definido:

$X' = Log\ (X + 1)$, o bien: $X' = Ln\ (X + 1)$

Se recomienda esta transformación para distribuciones asimétricas a la derecha.

Las propiedades de la transformación hacen que la forma de la distribución resultante no cambie si utilizamos logaritmos con diferente base, lo que ocurre es que cuando la base es mayor, disminuye la varianza de la X´.

2) Transformación de Potencia

Sea X una variable aleatoria con distribución No normal se dice que X´ es su transformada cuando:

$X' = X^r$, siendo r un número positivo > 0

a) Cuando r < 1

Constituyen lo que normalmente llamamos raíces, recordemos que, por ejemplo:

$$X^{0,5} = \sqrt[2]{X}$$

Así estas transformaciones se obtienen:

$X' = \sqrt[2]{X}$; $X' = \sqrt[3]{X}$; $X' = \sqrt[4]{X}$

Al igual que las transformaciones logarítmicas, e recomienda esta transformación para distribuciones asimétricas a la derecha. Obteniendo diferentes resultados a las anteriores. Estas transformaciones hacen que la forma de la distribución resultante cambie si utilizamos potencia con diferente índice.

b) Cuando r > 1

Constituyen lo que acostumbramos a llamar potencia Así estas transformaciones se obtienen:

$X' = X^2$; $X' = X^3$; $X' = X^4$

A estas transformaciones se las recomienda para distribuciones asimétricas a la izquierda. Es necesario ser muy cautos en estas transformaciones ya que la X´posee una varianza mucho mayor que la X.

Pruebas para variables con distribuciones desconocidas

Si las transformaciones no consiguen normalizar a la variable respuesta, o bien si el investigador no desea transformar la variable, o bien si se tiene una variable no cuantitativa (una cualitativa ordinal) se pueden utilizar pruebas alternativas a las que se han visto. Esta rama de la Estadística se denomina Estadística No paramétrica o Estadística de distribución libre.

En General, a las pruebas denominadas no paramétricas clásicas, realizan dos cambios fundamentales con respecto a las pruebas llamadas paramétricas:

1) Realizan una transformación cambiando la métrica original de la variable por el número de orden, llamado rango como traducción de término inglés Rank.
2) Buscan los valores exactos de probabilidad de rechazo de la prueba y no los compara con una tabla estandarizada (como Z, t, χ^2 o F).

A continuación daremos un listado de pruebas que permiten cumplir objetivos similares a los propuestos en los casos estudiados:

Alternativa al caso 2: Prueba de hipótesis para la media poblacional, con σ^2 desconocida. Se presenta como alternativa la prueba de Wilcoxon (para una Mediana).
En realidad la Hipótesis nula está referida al orden promedio (promedio de Rank), pero como en el orden promedio se encuentra ubicada la Mediana, se puede entender como:

$$H_0)\ Med_x = Med_0$$
$$H_1)\ Med_x \neq Med_0$$

Esta prueba utiliza un estadístico llamado W_1.

Alternativa a los casos 4 y 5: Pruebas para la igualdad entre dos medias poblacionales, muestras independientes, σ_1^2 y σ_2^2 desconocidas. Se presentan como alternativa dos pruebas que son exactamente iguales, sólo cambian en la construcción del estadístico: Prueba de Wilcoxon para muestras independientes y prueba U de Mann Whitney. En ambos casos las hipótesis nula y alternativas son:

$$H_0)\ Med_1 = Med_2$$
$$H_1)\ Med_1 \neq Med_2$$

Wilcoxon utiliza un estadístico llamado W, y Mann Whitney uno denominado U

Alternativa al caso 6: Pruebas para la igualdad entre dos medias poblacionales, muestras dependientes (o apareadas). Se presenta como alternativa la Prueba de Wilcoxon para muestras apareadas, con las hipótesis:

$$H_0)\ Med_{dif} = 0$$
$$H_1)\ Med_{dif} \neq 0$$

Wilcoxon utiliza un estadístico llamado W,

Valor p

En todo este capítulo hemos estado presentando diferentes pruebas de hipótesis, donde la regla de decisión para poner a prueba la hipótesis presenta los intervalos con las zonas de rechazo y de no rechazo. Si el estadístico está comprendido en alguna de las zonas se concluye como consecuencia de esto.
En esta sección veremos otra manera de expresar el resultado de la prueba de hipótesis.

Retomaremos el Ejemplo 9.5, recordemos que las hipótesis eran:

$$H_0)\ \mu \geq 12$$
$$H_1)\ \mu < 12$$

Con esa batería de hipótesis corresponde aplicar lo que denominamos como caso 1 o como caso 2, según corresponda y dependiendo si es conocida o no la varianza poblacional (σ^2).

Se toma la decisión de cuál caso corresponde aplicar, se toma la muestra, se calcula la media muestral y se obtiene un resultado. En realidad podrían ocurrir infinitos diferentes resultados, es decir se podrían obtener infinitos valores de la media muestral e infinitos gráficos que corresponda a cada valor. En las Figuras 9.8 se presentan 4 ejemplos de posibles valores de la media muestral y su ubicación con respecto a la media poblacional bajo hipótesis (estandarizada a Z=0).

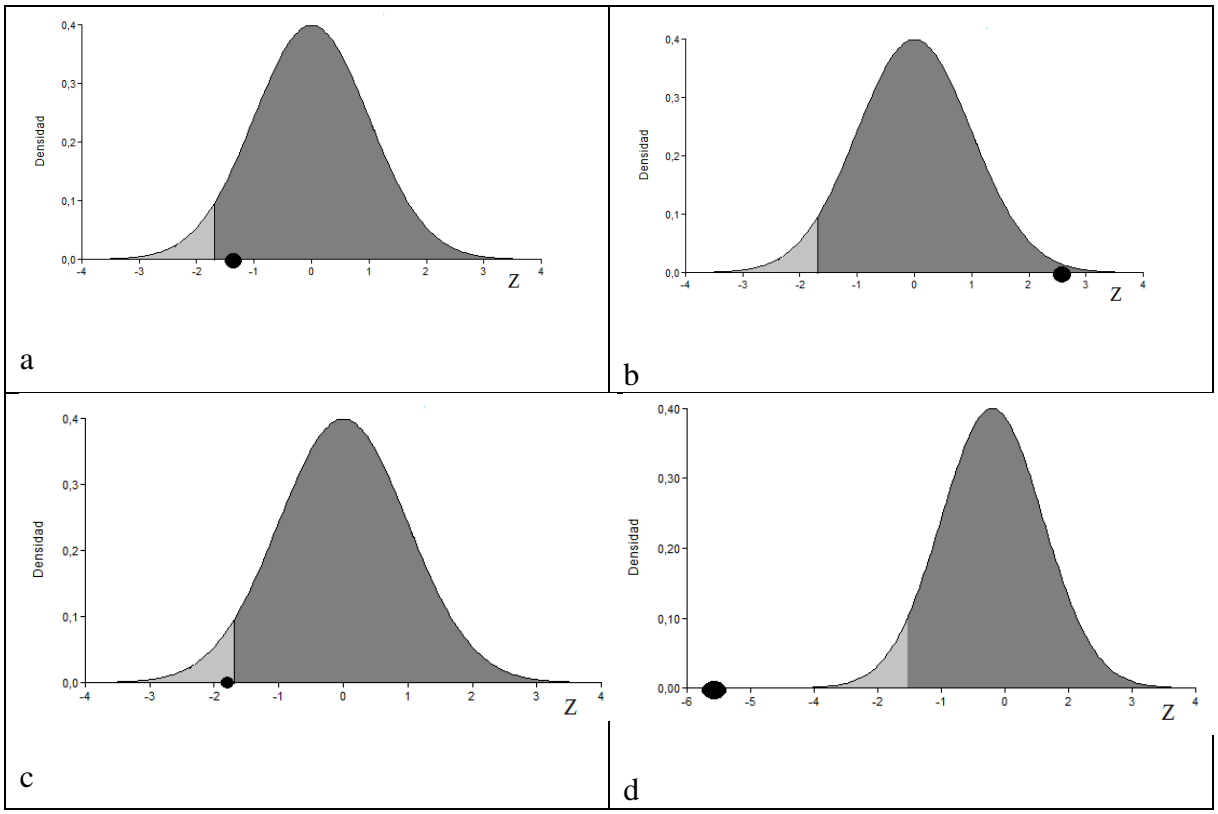


Figura 9.8. Cuatro diferentes posibilidades de la ubicación de la media muestral con respecto a la media poblacional bajo hipótesis.

Tanto en la Figura 9.8a como en la 9.8b, la conclusión será que no se rechaza la hipótesis nula y quien informa los resultados acota: No se rechaza Ho, sin informar cuán alejada se presenta la media muestral de la μ_0.
Tanto en la Figura 9.8c como 9.8d la conclusión será que se rechaza la hipótesis nula y quien informa los resultados acota: Se rechaza Ho, sin decir si la media muestral estaba apenas rechazada o se encontraba muy lejos de la μ_0.

Miremos ahora si sombreamos toda la superficie desde el valor de la media hacia menos infinito (Figuras 9.9).

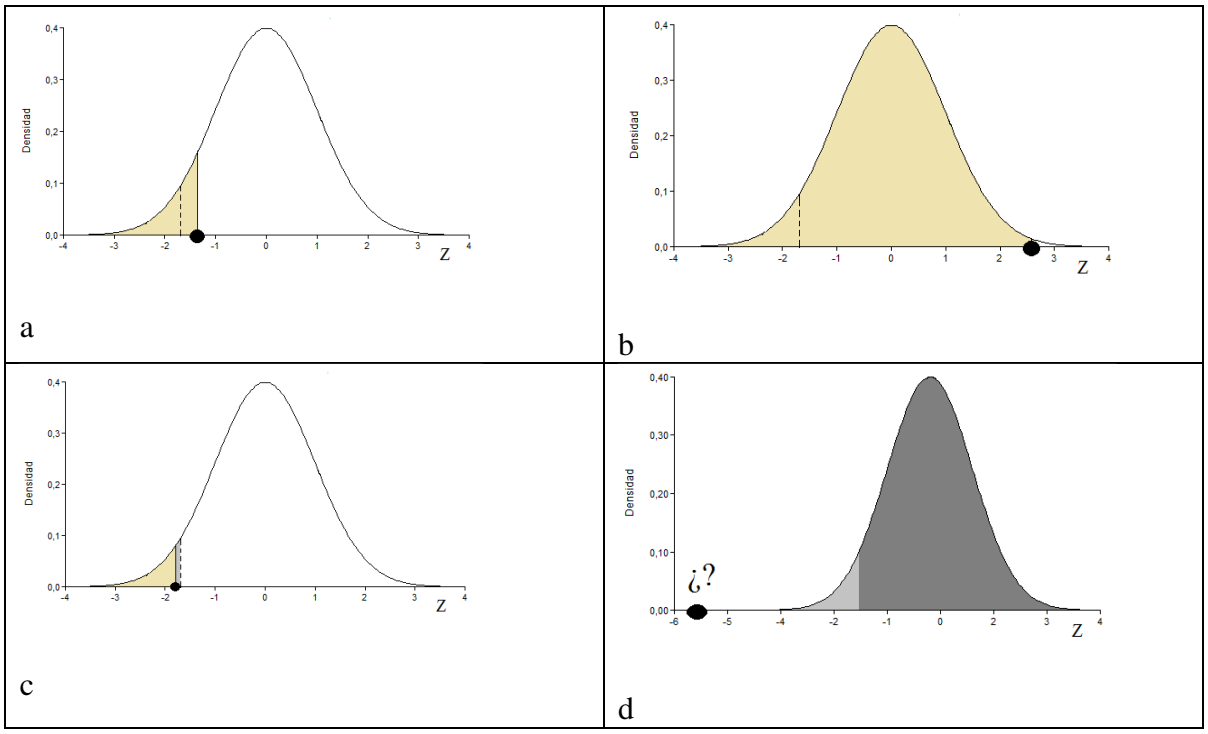


Figura 9.9. Gráficos de las Figuras 9.8 donde a las Cuatro diferentes posibilidades de la ubicación de la media muestral con respecto a la media poblacional bajo hipótesis muestra la probabilidad desde el valor del estimador a menos infinito.

Esa superficie nos da la probabilidad de obtener medias muestrales al azar que estén igual o más alejadas que la nuestra del valor bajo hipótesis. Dicho de otro modo, el valor p nos da la probabilidad de obtener estimadores igual o más alejados que el estimador de la muestra con respecto al parámetro bajo hipótesis.

De esta manera, una vez obtenido el valor p, se lo compara con α y la regla de decisión es:

Si $p < \alpha$, Rechazamos la Hipótesis nula
Si $p \geq \alpha$, No rechazamos la Hipótesis nula

Cuando el operador realiza los cálculo manualmente resulta sencillo establecer las zonas de rechazo y no rechazo de la manera clásica ya que para algunas distribuciones es muy complejo la búsqueda de los valores p. Sin embargo la totalidad de los programas estadísticos responden a las pruebas de hipótesis exclusivamente expresando el resultado con el valor p.

Capítulo 10.

Diseños Estadísticos.

Introducción.

El término de Diseño Estadístico es poco utilizado en ambientes académicos. Necesariamente debemos utilizarlo para poder diferenciar a aquellos Diseños donde el operador manipula la experiencia (Diseños Experimentales) de los otros donde el operador está limitado a medir los efectos que se han producido en algún sistema (Diseños observacionales).

En principio vamos a definir a Diseño como un procedimiento, un proceso, una sucesión de pasos ordenados a los fines de poner a prueba hipótesis o cumplir objetivos.
Por su parte los Diseños estadísticos los definiremos como todos los pasos realizados a los fines de poner a prueba hipótesis.

Como destacamos anteriormente los Diseño estadístico pueden ser de dos tipos:

1) **Diseño Experimental**. El investigador opera sobre las variables causales, las controla, luego de lo cual mide la variable objeto del estudio. Son también llamados diseños manipulativos.

Ejemplo 10.1. Un Geólogo desea saber cuál de 4 procedimientos conseguirá extraer más petróleo de la pizarra bituminosa (shale) ubicada en cierta zona. Para ello traslada desde el campo al laboratorio varios kilogramos de pizarra de una zona homogénea dentro del lugar en estudio. Luego procede a realizarle los respectivos procedimientos de extracción fijando ciertas condiciones y sólo cambiando entre ellas el método. Para finalizar procede a analiza cuántos litros por m^3 de petróleo extraen cada uno de los cuatro tipos de métodos.

2) **Diseño Observacional.** El investigador está imposibilitado de operar sobre las variables causales, sólo se limita a tomar o medir esas variables causales y mide la variable objeto de su interés. Son también llamados experimentos mensurativos.

Ejemplo 10.2. Un Geólogo desea saber si antes y después de una planta depuradora de líquidos cloacales se mantienen constantes las variables que indican contaminación ambiental, es decir conocer si la planta está cumpliendo con los requisitos de calidad del efluente. Para ello toma muestras de agua antes y después de la planta y luego in situ o en laboratorio mide la variable de interés (Concentración de Nitrito).

Necesidades y propósitos de un diseño estadístico.

El panorama ideal para un diseño sería aquel en que si se repite una experiencia, todos los resultados obtenidos fuesen exactamente iguales. Si eso ocurriese estaríamos en presencia

de una variable respuesta constante y no tiene sentido la estadística. Tanto en presencia de una constante o bien si se dispusiera del conocimiento de los parámetros poblacionales, no harían falta ni la estadística inferencial ni las complejas pruebas de hipótesis que hemos visto y aún nos quedan por describir. Como no conocemos los valores de los parámetros, la estadística inferencial es fundamental, ya que nos arroja un cierto grado de certidumbre de que el fenómeno que estamos observando sea causado sólo por el azar.

Se considera que el padre del Diseño Experimental fue Ronald Fisher (1890-1962), quien, trabajando para una estación agronómica Británica desde 1919 desarrolló las bases y los principios que quedaron plasmados en su libro "Diseño Experimental", publicado en 1935. Cabe destacar que Fisher era Matemático, dictó clases de Genética toda su vida y nunca estuvo en ninguna Cátedra de Estadística, pero ha sido el referente principal de la Estadística moderna.

Vamos a definir ciertos términos y conceptos:

Variable respuesta:
Constituye la variable aleatoria objeto del estudio, a la cual están apuntando las hipótesis, en el ejemplo 10.1, la cantidad (en l/m^3) de petróleo, en el ejemplo 10.2, la concentración de Nitritos. Nótese que estamos observando que necesariamente la variable respuesta debe ser cuantitativa.

Factor y niveles del factor:
La variable independiente constituye el factor, la variable que en el párrafo anterior denominamos variable causal. Es la variable que el operador considera que causa las diferencias posibles en la variable dependiente. Es una variable cualitativa nominal, los niveles del factor son cada una de las categorías en las que está dividido el factor. En el ejemplo 10.1 se presentan 4 niveles, mientras que en el ejemplo 10.2 se presentan 2 niveles. Cuando Fisher estableció estos términos lo hizo en experimentos de Agronomía, donde un nivel del factor coincidía con lo que era un tratamiento, por lo que el término tratamiento es utilizado como un sinónimo de niveles del factor en el marco exclusivo de ciertos diseños experimentales.

Ejemplo 10.3. Se deseaba saber si la adición de nitrógeno mejora el rendimiento del trigo. Entonces el Factor es Nitrógeno y se establecen 3 niveles: Control, Concentración de N_1, Concentración de N_2. En ese caso se presentan 3 tratamientos, siendo el control (cero Nitrógeno) también un tratamiento. La variable respuesta es la producción.

Unidad estadística:
Es la unidad en la que se va a medir la variable respuesta. Sería un sinónimo de lo que en la definición de población constituye un elemento. La unidad debe estar correctamente elegida, pues el conjunto de unidades representan a una población estadística específica.

Error estadístico

Se denomina error estadístico a la **variabilidad no explicada**. Cuando se someten varias unidades experimentales a un nivel del factor, se observan diferencias entre ellas, dicho de otro modo no todas las unidades responden igual a los niveles del factor.
Es muy común creer que este error experimental es sinónimo de equivocación, eso no es así, constituye la varianza entre todas las unidades que pertenecen a la misma población. En él puede estar mezclado algún tipo de error de medición, pero generalmente se deben a factores no tenidos en cuenta en la investigación.

Principios básicos del Diseño Estadístico

Reproducción.

Si la aplicación de un nivel del factor arrojara siempre una constante como respuesta, con sólo una unidad en cada nivel bastaría. Pero como la respuesta posee variabilidad (error experimental) es necesario repetir la experiencia en unidades independientes tantas veces como las necesarias para que ese grupo de unidades represente la variabilidad de la población a inferir. Además de esto, como se vio en el capítulo 8, por la propiedad de la consistencia, a medida que el n se incrementa se obtienen mejores estimaciones de los parámetros.
Volviendo al ejemplo 10.1: Será necesario repetir con cada tipo de extracción n veces la experiencia
Volviendo al ejemplo 10.2: Será necesario tomar más de una alícuota de agua, para representar la variabilidad de la variable en cada una de las dos zonas muestreadas.

Aleatorización.

Ya se ha hablado del término aleatorización cuando en el Capítulo 1 se desarrolló el concepto de la toma al azar de las unidades o elementos de la población con el objeto de formar la muestra.
El concepto en sí de aleatorización sigue siendo el mismo, pero debemos explicitarlo por separado para los dos tipos de diseños:
En Diseños experimentales: Se debe asignar las unidades a los niveles del factor en forma aleatoria o bien asignar los niveles a las unidades, de este modo cualquier unidad tiene la chance de "entrar" en cualquier nivel.
En diseños observacionales: Se deben seleccionar al azar las unidades en cada uno de los niveles del factor. En este caso la unidad ya viene con el nivel del factor impuesto, y debemos tomar muestras al azar de ellos para que, como se desarrolló en muestreo, representen la población a la que se va a inferir.

Recordemos que para conseguir una variable aleatoria, y poder gozar de las propiedades de éstas, se debe realizar un experimento aleatorio. Si las unidades son elegidas subjetivamente no estaremos en presencia de una variable aleatoria.

En el ejemplo 10.1 el operador decide separar en 24 porciones de 15 dm^3, que constituye la unidad. Al azar se asignan cada una de esas 24 porciones a los 4 niveles, quedando entonces aleatorizadas 6 porciones a cada nivel.

En el ejemplo 10.2 el operador decide tomar 3 alícuotas de agua en cada uno de los 2 sitios. La selección de los lugares debe necesariamente ser aleatoria.
Recordemos como se vio en los finales del capítulo 8, cuanto mayor es la variabilidad entre las unidades mayor es el n que represente esa variabilidad.

Control Local.
Definimos como Control Local a todos los procedimientos que realiza el investigador a los fines de reducir la variabilidad no explicada o error experimental. Algunos autores lo llaman precisión, pero para no confundir el término con las propiedades de los buenos estimadores aquí será llamado Control Local.
Obviamente el investigador siempre utilizará la prueba estadística que le reporte mayor potencia, así como el n más grande posible y realizar, en la medida de lo posible, un balanceo entre niveles (que los niveles del factor tengan igual n) a los fines de rechazar la hipótesis nula

Siempre pensando como protagonista principal a la variable respuesta, es más que obvio que no sólo una variable independiente (Factor) puede influir sobre ésta. Se puede pensar que el investigador conoce que son innumerables los factores que pueden modificarla. Ante tantos factores a tener en cuenta el operador puede decidir dos caminos:

Ignorar variables: El investigador decide conscientemente o por simple desconocimiento no tener en cuenta esas variables o factores en el análisis. En el ejemplo 10.2, la profundidad donde toma la muestra podría afectar la respuesta, pero en la porción central del cauce se toma la porción de agua a nivel de superficie y hacia las orillas la toma a 5 cm de profundidad porque asume que la variable profundidad no afectará a la respuesta, dicho de otro modo: decide ignorarla.

Controlar las variables. Implica realizar un control estadístico de las variables que pueden afectar la investigación y para esto hay dos formas:

Control mediante el diseño
El investigador va a ir descartando grupos de unidades que forman subpoblaciones particulares y se va reduciendo la población estadística a la cual muestrea o experimenta. Por lo tanto acota la inferencia. En la medicina se utilizan para tales fines los llamados criterios de inclusión y exclusión. En el ejemplo 10.2: el investigador decide tomar muestras subsuperficiales, a 5 cm debajo de la superficie, sólo en la porción central del río y en lugares donde la velocidad de la corriente no produzca burbujeo. Eso implica tomar todas porciones de agua que son lo más homogéneas posibles, descartando el "efecto orilla", el "efecto profundidad", el "efecto turbulencia", entre otros.

Control mediante el modelo estadístico
El investigador decide incorporar al modelo estadístico los factores que pueden influir en la variable respuesta, es decir va anexando parámetros al modelo, éstos pueden tener asociadas hipótesis, como la incorporación de un factor o una covariable o no tenerla, como es en el caso de la incorporación de bloques en el modelo.

Cuando se vea el modelo estadístico se comprenderá la importancia de ir incorporando factores al mismo. Sólo destacaremos que cuando se van construyendo modelos más complejos, se incorporan más parámetros.

Modelos Estadísticos

Análisis de Varianza a un factor (Anava o Anova).

En el Capítulo 9, específicamente en el caso 4 se vio una prueba a los fines de observar si dos medias eran o no tomadas de la misma población. Es decir, bajo el supuesto de que las varianzas poblacionales eran iguales, se toman dos muestras y se hipotetiza: $\mu_1 = \mu_2$. Esta prueba es la denominada prueba t de diferencia de medias, desarrollada por el estadístico Gosset (1876-1937). Para el caso que el investigador desease saber si tres medias poblacionales son iguales (un factor con tres niveles) se plantea: $\mu_1 = \mu_2 = \mu_3$.
Está demostrado que no debe utilizarse la prueba t de diferencia de medias para contrastar esa hipótesis, pues para probarla se debieran realizar tres pruebas: $\mu_1 = \mu_2$; $\mu_1 = \mu_3$ y $\mu_2 = \mu_3$.De ese modo, en cada una de las tres pruebas se tiene un error de tipo I (cuyo valor es α) y se concluye la prueba con un error de tipo I que casi triplica las expectativas iniciales.

A partir del trabajo de Ronald Fisher, a los fines de poner a prueba la hipótesis de la igualdad de tres o más medias ($H_0: \mu_1 = \mu_2 = \mu_3$), crea el denominado Análisis de la Varianza (Anova o Anava) que con un solo análisis...contrasta a k medias con un solo valor fijo de probabilidad de error de tipo I.

Del mismo modo que en el Capítulo 9 se explicaron los conceptos de las pruebas de hipótesis con un ejemplo, en este capítulo realizaremos el mismo procedimiento con la explicación del Análisis de la Varianza.

Ejemplo 10.3. Un Geólogo que realiza estudios en volcanes extrae los valores de CO_2 cada 1 hora en 4 chimeneas. Necesita saber si hay diferencias significativas en las emisiones de estos 4 respiraderos. Los datos obtenidos fueron:

Vent 1	Vent 2	Vent 3	Vent 4
27	31	30	16
28	34	38	20
31	35	42	21
32	36	43	26
33	39		27
	40		29
			35

Media	30,20	35,83	38,25	24,86
n	5	6	4	7
Desvío est.	2,59	3,31	5,91	6,36

Tabla 10.1. Valores de CO_2 en 4 chimeneas, valor medio, n y desvío estándar del ejemplo 10.3.

Ahora primero recordemos los tres principios del Diseño Estadístico. ¿Se tomaron repeticiones? ¿Se tomaron o asignaron al azar las muestras?, ¿Se realizó control local? Asumamos que las tres respuestas son positivas, prosigamos con el análisis que pondrá a prueba la hipótesis del investigador. En definitiva la hipótesis geológica es que en alguna chimenea se produce más CO_2. Eso se lo debe traducir a las baterías de hipótesis estadística, las que serán:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

H_1: al menos una $\mu_i$es diferente a las otras .

Nótese que como hipótesis alternativa no se está pidiendo que sean todas diferentes de todas, se pide que para rechazar la hipótesis nula, al menos una media poblacional sea diferente a las otras.
Es cierto que las cuatro medias muestrales obtenidas son matemáticamente diferentes, ellas son: 30,2; 35,83; 38,25 y 24,86. Pero ¿estas diferencias representarán diferencias entre las 4 poblaciones? o ¿estas diferencias serán sólo por azar y realmente sólo hay una población estadística con sólo una media poblacional?

A partir de estas preguntas desarrollaremos la idea de cuántas varianzas se pueden obtener de la tabla 10.1.
Como sabemos una varianza es una distancia entre un valor de la variable aleatoria y un estimador. Entonces denominaremos como "y" a la variable respuesta, siendo y_{ij} cada valor en particular. De cada grupo o nivel del factor, obtendremos una media: $\overline{y}_i$ para cada uno de los cuatro grupos. Además obtendremos la media general de todo el análisis que denominaremos $\bar{\bar{y}}$.

Una varianza total, expresa la diferencia entre cada valor de la variable y la media general, es decir la varianza total expresa la distancia entre: $y - \bar{\bar{y}}$

Una varianza del factor es la que es causada por el efector del factor, entonces es la distancia entre cada media de nivel del factor con la media general: $\bar{y}_i - \bar{\bar{y}}$. Se la conoce como **varianza entre grupos**.

Una varianza del error es la distancia entre cada valor de la variable y la media de su propio grupo: $y_{ij} - \bar{y}_i$.Se la conoce como **varianza dentro de los grupos**.

Definiremos además a k como el número de niveles del factor:

	Nivel 1	Nivel 2	…	Nivel k
1	y_{11}	y_{21}	…	y_{k1}
2	y_{12}	y_{21}	…	…
3	y_{13}	…		
…	…			

n_i	y_{1n_1}	y_{2n_2}	...	y_{kn_k}	
$\sum y_{i.}$	$\sum y_{1.}$	$\sum y_{2.}$	$\sum y_{i.}$	$\sum y_{k.}$	
$\bar{y}_i$	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_i$	$\bar{y}_k$	$\bar{\bar{y}}$

Tabla 10.2. Tabla General de la nomenclatura utilizada en el Anova a un factor.

Recordemos que cuando vimos en los Capítulos 2 y 7 el concepto de la varianza muestral lo hicimos dividiendo el cuadrado de las diferencias entre los valores de la variable y un estimador, con los grados de libertad. Podríamos pensar en el numerador de la varianza como una suma de cuadrados y a esta la dividimos por los grados de libertad, eso nos daría un promedio cuadrado o cuadrado medio:

$$S^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = \frac{Suma\ de\ cuadrados}{Grados\ de\ Libertad} = Cuadrados\ medios$$

La siguiente tabla mostrará ahora cómo se deben realizar los cálculos para obtener esas tres varianzas o cuadrados medios:

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios
Factor	$\sum \frac{(\sum y_{i.})^2}{n_i} - \frac{(\sum\sum y_{ij})^2}{n}$	k-1	$\frac{SCF}{GLF}$
Error	$\sum\sum y_{ij}^2 - \sum \frac{(\sum y_{i.})^2}{n_i}$	n - k	$\frac{SCE}{GLE}$
Total	$\sum\sum y_{ij}^2 - \frac{(\sum\sum y_{ij})^2}{n}$	n - 1	

Tabla 10.3. Fórmulas de Suma de cuadrados, grados de libertad y cuadrados medios para cada fuente de variación.

Debido a que las sumas de cuadrados del factor y del error son ortogonales (independientes que no comparten información) se da que: SCF + SCE= SCT
Del mismo modo: GLF + GLE = GLT

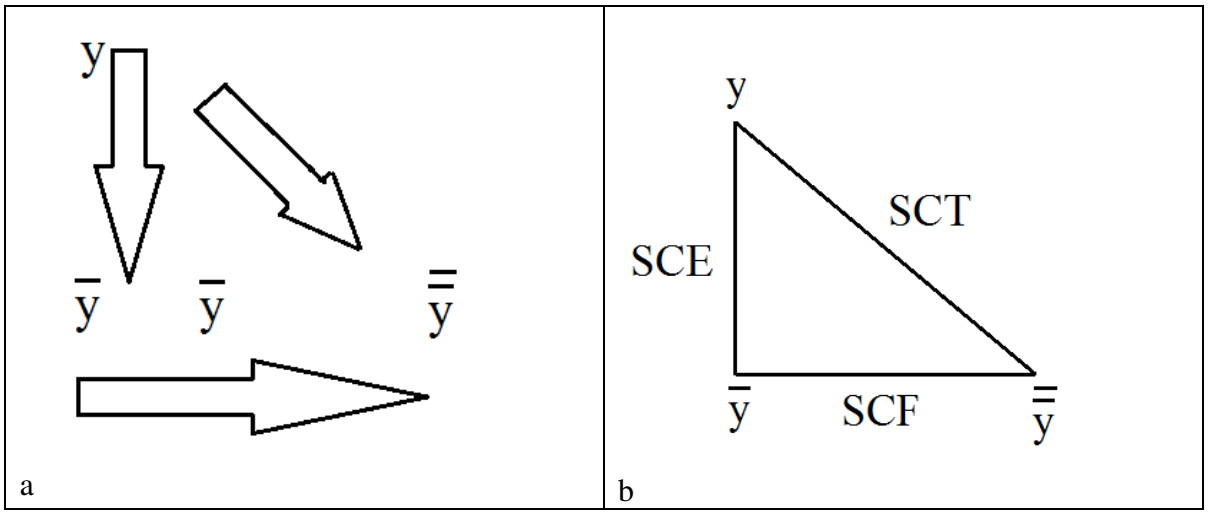


Figura 10.1. Triángulo de tres varianzas independientes en el Anova. a: Sentido en que se estiman las diferencias entre variable y estimador. b: Ortogonalidad del triángulo entre las tres sumas de cuadrados.

Como sabemos, si queremos saber si dos varianzas son iguales o difieren, podremos construir el estadístico F.

F_0= CMF/CME

Para comprender dónde se establecen las zonas de rechazo y de no rechazo de este estadístico en el Anova debemos ver que:

E(CME)=σ^2, ya que se puede entender al Cuadrado medio de error como un promedio de las varianzas de todos los niveles del factor. Es una ampliación de la fórmula de la varianza amalgamada que se vio en el caso 5 del capítulo 9.

E(CMF)= σ^2+d, donde d >0.
Para esta explicación es más sencillo pensar en lo siguiente. En cada uno de los k grupos se elige un representante del grupo, que es la media muestral de cada grupo. Esta media representa un valor promedio, pero también representa la propia variabilidad del grupo que la generó. Cuando esa media va a "sentarse en la mesa de los representantes", mesa en la cual se encuentran sólo las medias para ver si son diferentes entre ellas, cada media no sólo aporta el valor que posee, sino que además aporta a la varianza de cada grupo, varianza que lleva sobre sus espaldas. Dicho de otro modo, mientras que un valor cualquiera de la variable sólo da su opinión, la media debe dar su opinión más la variabilidad de la opinión de sus representados.
Por esa razón la variabilidad entre las medias será un estimador de la varianza poblacional más la distancia (d) que hay entre medias.

De este modo si $F = \frac{CMF}{CME} \rightarrow \frac{\sigma^2 + d}{\sigma^2}$

Es así que si la Hipótesis nula es cierta ($H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$), el valor esperado de F será= 1, ya que no hay distancia entre medias muestrales (d=0). Si la hipótesis nula es falsa, entonces d es un "número grande" y el valor de F tiende a infinito. Por esa razón la prueba de Anova es exclusivamente unilateral derecha, teniendo como criterio de decisión:

Zona de no rechazo: Si ($0 < F_0 <$ F), entonces no se rechaza H_0
Zona de rechazo. Si (F $< F_0 < \infty$)entonces se rechaza H_0

Volvamos al ejemplo 10.3:

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	F_0	$F_{t\,(0,95)}$
Factor	612,26	3	204,09	8,56	3,16
Error	429,24	18	23,85		
Total	1041,50	21			

Tabla 10.4. Valores de las diferentes variancias según ejemplo 10.3.

Como $F_0 > F_t$ entonces rechazamos la Hipótesis nula, es decir: al menos una μ_i es diferente a las otras.

Modelo estadístico y supuestos

El **modelo Estadístico** que responde el Anova a un factor es:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

Donde:
y_{ij}es la variable respuesta
μes la media poblacional
$\tau_i$es el efecto de cada nivel del factor
ε_{ij}es el error (variabilidad no explicada después de ajustar el modelo)

Supuestos

$\varepsilon_{ij} \sim NI(0, \sigma^2)$, es decir que los errores posean distribución normal, sean independientes, com media cero y homogeneidad de varianzas (una varianza común y no una varianza para cada nivel del factor).

Para que el modelo sea válido y lo que se expresó como conclusión de la tabla 10.4 sea cierto, debemos corroborar el cumplimiento de los supuestos.
Para ello se presentan métodos no formales y pruebas formales para hacerlo.

1) **Normalidad.**

El método no formal para observar si los errores poseen distribución normal es un gráfico de QQplot. Este es un gráfico que contrasta los percentiles de una normal con la

distribución de los errores del análisis. Si los errores poseen distribución normal seguirán una recta de 45 grados.

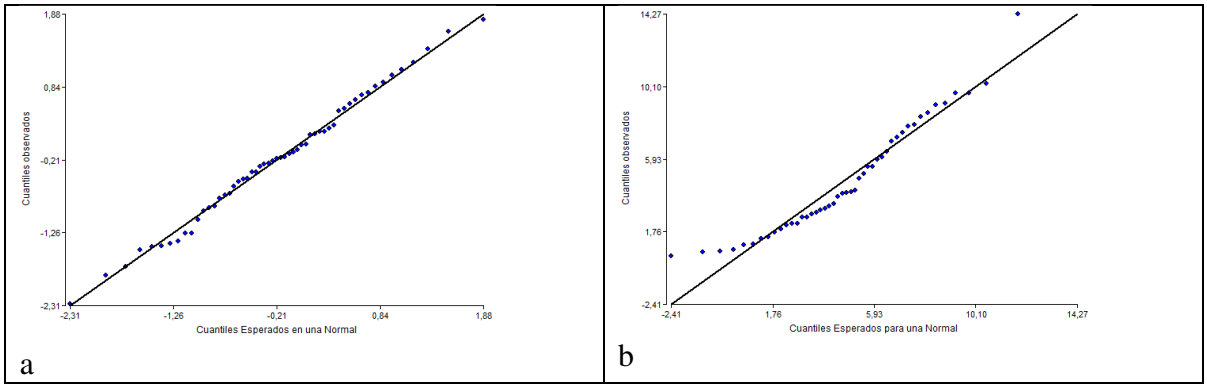


Figura 10.2. Gráficos de QQplot para el ajuste de la normalidad de los errores.

En la figura 10.2a se observa que los errores siguen una distribución normal, mientras que en la Figura 10.2b se observa que se salen del patrón de una normal.

Por otra parte, como ya se ha visto en el Capítulo 9, mediante una prueba de bondad de ajuste (Shapiro Wilks) se podrá contrastar la distribución normal de los errores.

2) Homogeneidad de varianzas.

El método no formal para observar si las varianzas son heterogéneas es un gráfico de valores predichos vs errores, allí se observa si la variabilidad de los residuos en cada nivel del factor es similar.

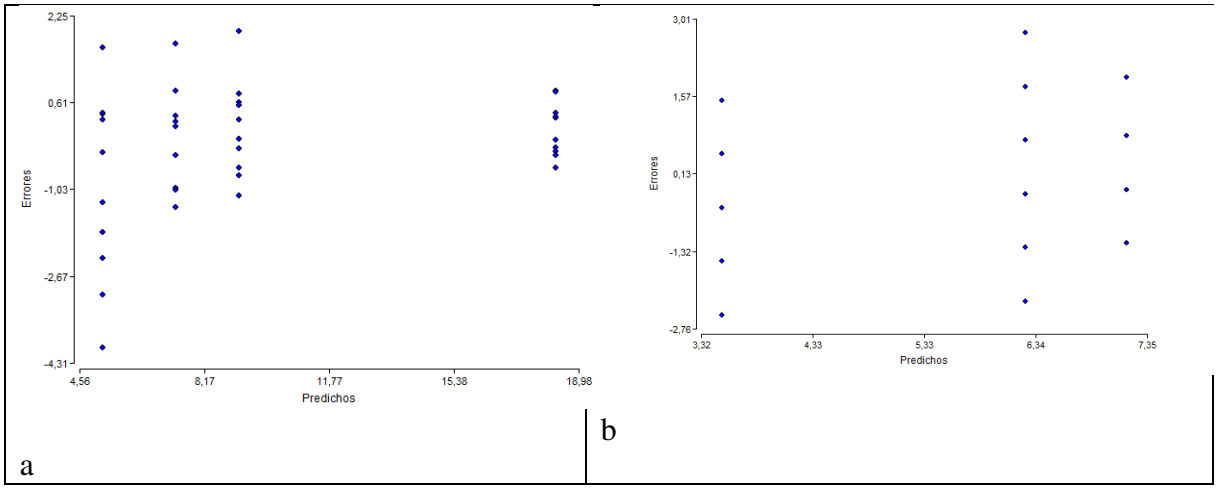


Figura 10.3. Gráficos donde se observa la variabilidad de los residuos ante los diferentes valores predichos. A. Ausencia de homogeneidad de varianzas, b: presencia de tal.

Diferentes métodos formales se han propuesto para poner a prueba la homogeneidad de tres o más varianzas: la prueba de Levene, de Bartley, entre otras.

Ya hemos desarrollado la idea que el Anova utiliza las varianzas para comparar medias. También desarrollamos los supuestos estadísticos del modelo. El Anova, como lo vimos anteriormente, concluye exclusivamente rechazando o no rechazando la hipótesis nula. Si el modelo es válido y no se rechaza la H0, el análisis termina allí. Pero como en el caso del ejemplo 10.3, cuando el Anova dice que al menos una es diferente, la pregunta es cuál. Para ello se han creado un número extenso de pruebas que intentan dilucidar cuáles medias son diferentes de cuáles.

Pruebas de comparaciones múltiples

Son llamadas también pruebas post hoc o pruebas a posteriori, se realizan después que el Anova ha rechazado la hipótesis nula, entre ellas por su simplicidad explicaremos la prueba de Tukey.
Esta prueba busca una diferencia mínima significativa (DMS), que es una distancia a partir de la cual dos medias son significativamente diferentes. Si la distancia entre medias no llega a esa DMS se dice que las medias no son significativamente diferentes.

$$DMS = Q_{k;gle;1-\alpha} * \sqrt{\frac{CME}{n_i}}$$

Donde Q es un valor de la tabla de rangos estudientizados de Tukey.

Volvemos al ejemplo 10.4:

DMS= 3.997* $\sqrt{\frac{23,85}{5,5}}$

DMS= 8,412

Por lo tanto si dos medias están alejadas en más de 8,412 se las considera estadísticamente diferentes. A partir de esto se observan si entre dos medias existen diferencias y se construye la siguiente tabla:

Vent 4	24,86	7	1,85	A	
Vent 1	30,2	5	2,18	A	B
Vent 2	35,83	6	1,99		B
Vent 3	38,25	4	2,44		B

Tabla 10.5 Valores de medias y se consignan letras iguales que implican que las medias son iguales (según prueba de Tukey del ejemplo 10.3).

De este modo se concluye que la chimenea 4 es diferente a las chimeneas 2 y 3.

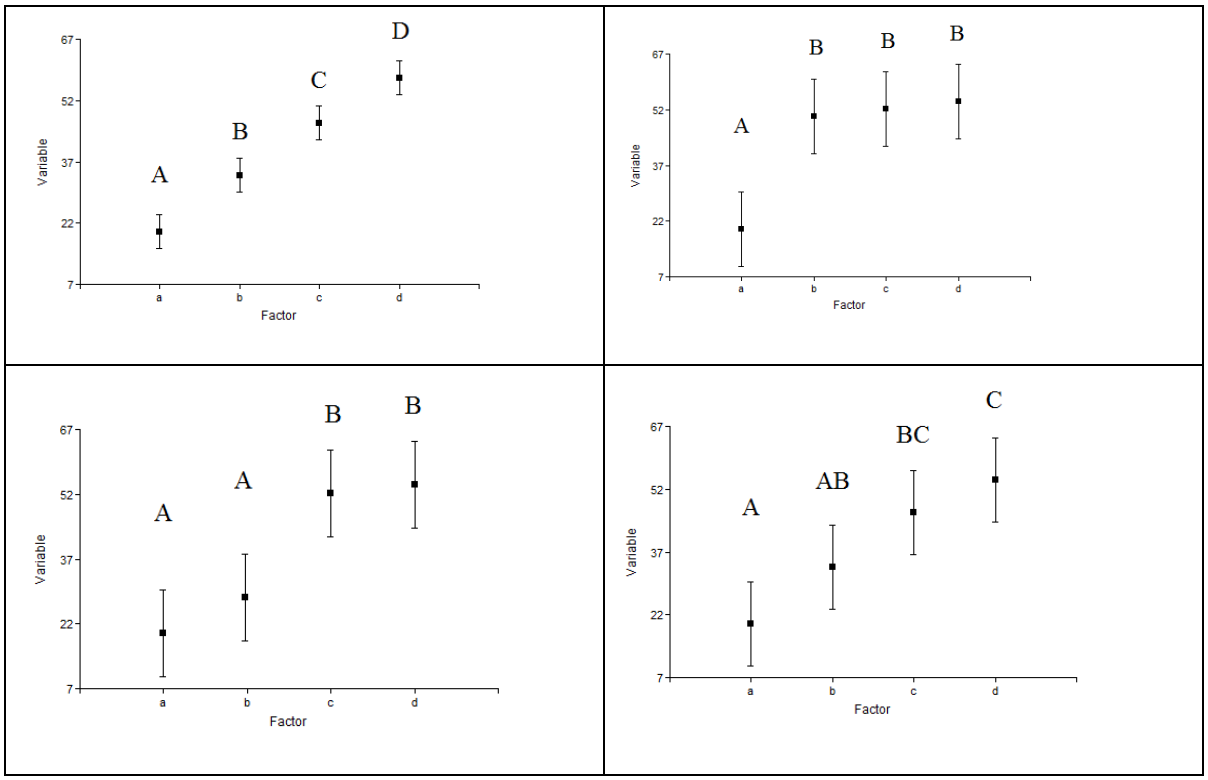


Figura 10.4. Gráfico de puntos de cuatro posibles diferencias entre cuatro medias. Figura la media y el error estándar. Letras iguales muestran diferencias no significativas.

Corolario: Como se puede observar, las diferencias entre medias serán diferentes si el CME es lo suficientemente pequeño para que se destaque la magnitud de la distancia (d) entre medias. Lo que está mostrando el valor F es si la distancia entre las medias es mucho más grande que la distancia entre los datos de grupo de cada media. Por esa razón el Control Local intenta disminuir el error, para hacer que si hay diferencias entre medias, el F lo pueda encontrar. En definitiva y comparando el Anova a un factor con los casos 4 y 5 del Capítulo 9 en las tres pruebas se está mirando la misma tasa de distancias, mientras que la prueba t de diferencia de medias realiza una tasa entre la distancia entre medias dividido el desvío promedio entre ambos grupos, el Anova realiza una tasa entre la varianza entre medias dividido la varianza interna dentro de los grupos.
Está probado que para el caso en que se realiza un análisis de la varianza a un factor con dos niveles, el resultado es exactamente el mismo que si se hiciera la prueba t de diferencia de medias, pues $t^2=F$.

Análisis de la varianza a un factor con bloques (Anova a un factor con bloques).

Ejemplo 10.5. Modificaremos a los fines prácticos, el Ejemplo 10.2. del siguiente modo. A los fines de realizar control local, el Geólogo que está estudiando los efectos de la planta depuradora decide tomar muestras en las zonas, dividiendo a cada una en 4 partes: orilla derecha, orilla izquierda, centro en la superficie y centro 1 metro bajo superficie.

El cambio en la modalidad del diseño y del muestreo repercute en la necesidad de incorporar al modelo un nuevo término que exprese la variabilidad entre estas cuatro partes que podríamos llamarlos microambientes. El término incorporado al modelo se denomina Bloque. Se define un bloque como un grupo de unidades que son homogéneas entre sí y heterogéneas con respecto a las unidades de otros bloques. Es importante recordar que el investigador no plantea hipótesis para el bloque, él ya sabe que hay diferencias entre ellos y lo incorpora al modelo sólo a los fines de realizar control local, es decir reducir la variabilidad no explicada.

Modelo estadístico

$$y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$$

Donde:
y_{ij}:es la variable respuesta
μ:es la media poblacional
τ_i:es el efecto de cada nivel del factor
β_j:es el efecto de cada bloque
ε_{ij}:es el error (variabilidad no explicada después de ajustar el modelo)

Supuestos

$\varepsilon_{ij} \sim NI(0, \sigma^2)$, y además no existe interacción bloque- factor.

Hipótesis

$H_0: \mu_1 = \mu_2$
H_1: al menos una $\mu_i$es diferente a la otra .

Tabla de Anova

Donde:

	Nivel 1	Nivel 2	...	Nivel k	$\sum y_{.j.}$
Bloque 1	y_{11}	y_{21}	...	y_{k1}	$\sum y_{.1}$
Bloque 2	y_{12}	y_{21}	...	...	$\sum y_{.2}$
Bloque 3	y_{13}	...			
...	...				
Bloque n_i	y_{1n_1}	y_{2n_2}	...	y_{kn_k}	$\sum y_{.k}$
$\sum y_{i.}$	$\sum y_{1.}$	$\sum y_{2.}$	$\sum y_{i.}$	$\sum y_{k.}$	
$\bar{y}_i$	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_i$	$\bar{y}_k$	$\bar{\bar{y}}$

Tabla 10.6Tabla General de la nomenclatura utilizada en el Anova a un factor con bloques.

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	F_0	F_t
Factor	$\sum \frac{(\sum y_{i.})^2}{n_i} - \frac{(\sum\sum y_{ij})^2}{n}$	k-1	$\frac{SCF}{GLF}$	$\frac{CMF}{CME}$	$F_{(k-1;\ n-k;\ 1-\alpha)}$
Bloque	$\sum \frac{(\sum y_{.j})^2}{n_i} - \frac{(\sum\sum y_{ij})^2}{n}$	n_i - 1	------		
Error	$\sum\sum y_{ij}^2 - \sum \frac{(\sum y_{i.})^2}{n_i} - \sum \frac{(\sum y_{.j})^2}{n_i} + \frac{(\sum\sum y_{ij})^2}{n}$	(n_i – 1) * (k -1)	$\frac{SCE}{GLE}$		
Total	$\sum\sum y_{ij}^2 - \frac{(\sum\sum y_{ij})^2}{n}$	n - 1			

Tabla 10.7.Fórmulas de Suma de cuadrados, grados de libertad y cuadrados medios para cada fuente de variación.

Como se puede observar:

- Las fórmulas de SCF y SCT son exactamente iguales que en el análisis de la varianza a un factor, entonces el valor de la SCB se extrae de reducir la SCE.
- Se obtiene sólo una F_0, ya que se presenta sólo una hipótesis nula.

De igual modo al Anova a un factor, se deben probar si se cumplen los supuestos del modelo. Una vez puestos a prueba si se rechazara la hipótesis nulase procederá, si es de interés, a las pertinentes pruebas de comparaciones múltiples.

Para poner a prueba los supuestos se realizarán pruebas formales y se graficaran los errores en QQplot o en un gráfico de errores vs predichos. Además de esto el supuesto de no interacción implica que no se comporta de una manera no lineal ningún nivel del factor en ningún bloque. Dicho de otro modo, si hay un bloque "propicio" para la variable, lo será en cualquiera de los niveles del factor.
La figura 10.5 muestra cómo se observa una no interacción y una interacción factor bloque.

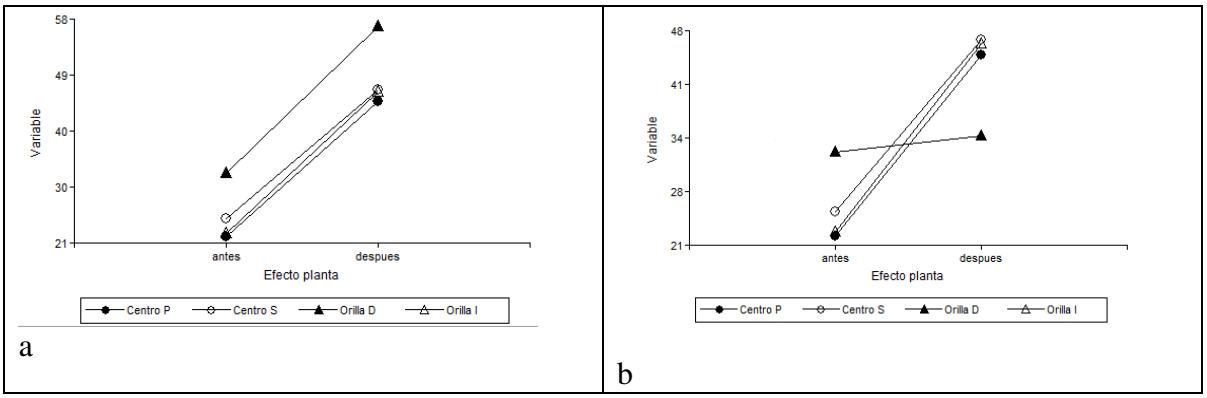


Figura 10.5. Gráfico donde se observan los valores de la variable en los niveles del factor, pero particionados por bloque. A. No se observa interacción. B. Se observa interacción.

Análisis de la varianza a dos factores con interacción.

Del mismo modo en que se planteó la idea de que un factor afecta a la variable respuesta, se puede pensar que dos factores la afectan y lo que es más importante no sólo la afectan por separado sino que juntos podrán tener una acción sinérgica y presentar una interacción sobre ésta.

De este modo se presentará el **modelo estadístico**:

$$y_{ijk} = \mu + \tau_{1i} + \tau_{2j} + (\tau_1\tau_2)_{ij} + \varepsilon_{ijk}$$

Donde:
y_{ijl} :es la variable respuesta
μ :es la media poblacional
τ_{1i} :es el efecto de cada nivel del factor principal 1
τ_{2j} :es el efecto de cada nivel del factor principal 2
$(\tau_1\tau_2)_{ij}$:es la interacción de los efectos de los factores principales
ε_{ijl} :es el error (variabilidad no explicada después de ajustar el modelo)

Supuestos

$\varepsilon_{ijk} \sim NI(0, \sigma^2)$,

Hipótesis
Se presentan tres baterías de hipótesis:

$H_0: \mu_{11} = \mu_{12} = \cdots = \mu_{1k1}$ H_1: al menos una μ_{1i}es diferente a la otra	$H_0 : \tau_{1i} = 0$ $H_1: \tau_{1i} \neq 0$
$H_0: \mu_{21} = \mu_{22} = \cdots = \mu_{2k2}$ H_1: al menos una μ_{2j}es diferente a la otra	$H_0 : \tau_{2j} = 0$ $H_1: \tau_{2j} \neq 0$
...	$H_0 : (\tau_1\tau_2)_{ij} = 0$ $H_1(\tau_1\tau_2)_{ij} \neq 0$

Nótese que las hipótesis referidas a los factores principales siguen siendo igualdades entre medias, pero la hipótesis de la interacción no busca esto mismo, ya que si los factores principales poseen efectos significativos se podrán tener diferencias en las medias de las celdas que constituyen la combinación de niveles de ambos factores. Entonces lo que busca la interacción es saber si las diferencias que se observan entre medias de las celdas son sólo producto de un efecto aditivo entre los niveles del factor o se da es un patrón diferente. Es lo que llamamos interacción entre los factores.

Tabla de Anova, donde:

	Nivel 11	Nivel 12	...	Nivel 1k1	$\sum y_{.j.}$	$\bar{\bar{y}}_{.j}$
Nivel 21	$y_{111} y_{112}$ $y_{11\ldots} y_{11\,l}$ ---------- $\sum y_{ij.}$ $\bar{y}_{ij.}$	$y_{21\ldots}$	...	$y_{k1\,2\ldots}$	$\sum y_{.1}$	$\bar{\bar{y}}_{.1}$
Nivel 22	$y_{12\ldots}$	y_{22} ...	...	...	$\sum y_{.2}$	$\bar{\bar{y}}_{.2}$
...	...					
Nivel 2k2	y_{1k_1}	y_{2n_2}	...	y_{k1k2}	$\sum y_{.k}$	$\bar{\bar{y}}_{.k2}$
$\sum y_{i.}$	$\sum y_{1.}$	$\sum y_{2.}$		$\sum y_{k.}$	$\sum\sum\sum y_{ij}$	
$\bar{\bar{y}}_i$	$\bar{\bar{y}}_{1.}$	$\bar{\bar{y}}_{2.}$		$\bar{\bar{y}}_{k1.}$		$\bar{\bar{y}}$

Tabla 10.8. Tabla General de la nomenclatura utilizada en el Anova a dos factores con interacción.

Fuente de variación	Suma de cuadrados	Grados de libertad
Factor 1	$\sum \frac{(\sum y_{i.})^2}{n_i} - \frac{(\sum\sum\sum y_{ij})^2}{n}$	k_1-1
Factor 2	$\sum \frac{(\sum y_{.j})^2}{n_j} - \frac{(\sum\sum\sum y_{ij})^2}{n}$	k_2 - 1
Factor1*Factor2	$\sum\sum \frac{(\sum y_{ij})^2}{n_{ij}} - \sum \frac{(\sum y_{i.})^2}{n_i} - \sum \frac{(\sum y_{.j})^2}{n_j} + \frac{(\sum\sum\sum y_{ij})^2}{n}$	$(k_1-1)(k_2-1)$
Error	$\sum\sum\sum y_{ij}^2 - \frac{\sum\sum y_{ij.}^2}{n_{ij}}$	GLT-(GLF1+GLF2+GLF1F2)
Total	$\sum\sum\sum y_{ij}^2 - \frac{(\sum\sum\sum y_{ij})^2}{n}$	n - 1

Tabla 10.9.Fórmulas de Suma de cuadrados, grados de libertad y cuadrados medios para cada fuente de variación.

Cuadrados medios	F_0	F_t
$\frac{SCF1}{GLF1}$	$\frac{CMF1}{CME}$	$F_{(k1-1;\ gle;\ 1-\alpha)}$
$\frac{SCF2}{GLF2}$	$\frac{CMF2}{CME}$	$F_{(k2-2;\ gle;\ 1-\alpha)}$
$\frac{SCF1F2}{GLF1F2}$	$\frac{CMF1F2}{CME}$	$F_{((k1-1)(k2-1);\ gle;\ 1-\alpha)}$
$\frac{SCE}{GLE}$		

Tabla 10.9´. Fórmulas de Suma de cuadrados, grados de libertad y cuadrados medios para cada fuente de variación.

En la tablas 10.9 la lectura de los valores F debe hacerse desde abajo hacia arriba:
La ausencia de interacción permite mirar libremente los efectos de los factores principales.
Si se presenta interacción deberemos mirar cuidadosamente a los efectos principales pues la interacción los puede enmascarar y tergiversar los resultados.
Si un factor principal no arroja diferencias significativas no tiene razón de ser el aplicar una prueba de comparaciones múltiples para ese factor. Pero si al menos una de las tres hipótesis ha sido rechazada es importante realizar una prueba para la combinación de los factores.

Ejemplo 10.6. Un Geólogo observa en un área cuatro tipos de suelos diferentes y desea conocer si es diferente la percolación de agua de éstos. Para ello traslada a laboratorio porciones de los cuatro tipos de suelos (Factor 1: Suelo, con 4 niveles).
Por otra parte en condiciones controladas de laboratorio simula 3 caudales de lluvias diferentes (Factor 2: Lluvias, con 3 niveles). Para terminar mide el porcentaje del volumen de agua que percola a través de cierta profundidad en un lapso determinado de tiempo.

El **modelo estadístico** es:

$$y_{ijk} = \mu + \tau_{1i} + \tau_{2j} + (\tau_1\tau_2)_{ij} + \varepsilon_{ijk}$$

Donde:
y_{ijl} :es la variable respuesta: volumen de agua que percola
μ :es la media poblacional
τ_{1i} :es el efecto de cada nivel del factor principal 1: suelo
τ_{2j} :es el efecto de cada nivel del factor principal 2: caudal de lluvia
$(\tau_1\tau_2)_{ij}$:es la interacción de los efectos de los factores suelo y caudal
ε_{ijl} :es el error (variabilidad no explicada después de ajustar el modelo)

Supuestos

$\varepsilon_{ijk} \sim NI(0, \sigma^2)$,

Se van a plantear tres baterías de hipótesis:

$H_0: \tau_{1i} = 0$, es decir que no existe efecto suelo $H_1: \tau_{1i} \neq 0$
$H_0: \tau_{2j} = 0$, es decir que no existe efecto caudal $H_1: \tau_{2j} \neq 0$
$H_0: (\tau_1\tau_2)_{ij} = 0$, es decir no hay interacción suelo caudal $H_1(\tau_1\tau_2)_{ij} \neq 0$

Los resultados de la experiencia son:

		Caudal		
		Poco	Medio	Mucho
Suelo	A	75	78	85
		72	80	87
	B	62	70	78
		64	72	81
	C	85	92	98
		88	93	100
	D	62	75	92
		65	78	96

Tabla 10.10. Valores de la variable respuesta según factores del ejemplo 10.6

De este modo:

Fte. de Var	SC	GL	CM	F_o	F_t	Valor p
Suelo	1453,67	3	484,56	141,82	3,49	< 0,001
Caudal	1300,08	2	650,04	190,26	3,88	< 0,001
Interacción	222,58	6	37,10	10,86	2,99	< 0,001
Error	41,00	12	3,42			
Total	3017,33	23				

Tabla 10.11. Resultados del análisis de la varianza del ejemplo 10.6.

Ya sea comparando a los valores de F_o con los F_t o por la comparación del valor p con alfa, en los tres casos rechazamos las tres hipótesis nulas, entonces observaremos con una prueba de Tukey:

Suelo	Medias	
B	71,17	A
D	78,00	B
A	79,50	B
C	92,67	C

Tabla 10.12. Valores de medias y prueba de Tukey del factor Suelo en el ejemplo 10.6.

Caudal	Medias	
Poco	71,30	A

Medio	79,75	B
Mucho	89,63	C

Tabla 10.13. Valores de medias y prueba de Tukey del factor Caudal en el ejemplo 10.6.

Suelo	Caudal	Medias						
B	Poco	63,00	A					
D	Poco	63,50	A					
B	Medio	71,00		B				
A	Poco	73,50		B	C			
D	Medio	76,50		B	C			
A	Medio	79,00			C	D		
B	Mucho	79,50			C	D	E	
A	Mucho	86,00				D	E	F
C	Poco	86,50					E	F
C	Medio	92,50						F
D	Mucho	94,00						G
C	Mucho	99,00						G

Tabla 10.12. Valores de medias y prueba de Tukey de la combinación de los factores Suelo y Caudal en el ejemplo 10.6.

A partir del análisis de estas tres últimas tablas se pueden sacar diversas conclusiones tanto para cada factor en particular como en la interacción de éstos.

Otros modelos

Si los diseños han permitido tener en cuenta otros factores y el investigador posee hipótesis para los mismos, entonces es posible seguir incorporando factores al modelo, de tal forma que se podrán realizar anovas a tres o más factores, con la incorporación de dos o más bloques y otras variables independientes cuantitativas denominadas covariables.

Capítulo 11.

Correlación y Regresión

Correlación lineal

Recordemos que en el Capítulo 3 se vieron los conceptos de Covarianza y de Correlación lineal de Pearson. En este capítulo retomaremos esos conceptos, y ampliaremos este último con la prueba de hipótesis correspondiente, por lo que a modo de recordatorio veamos que:

Coeficiente de Correlación lineal

$$\rho_{x_1x_2} = \frac{\sigma^2_{x_1x_2}}{\sigma_{x_1}\,\sigma_{x_2}} \qquad\qquad r_{x_1x_2} = \frac{S^2_{x_1x_2}}{S_{x_1}\,S_{x_2}}$$

Donde:
$\rho_{x_1x_2}$: Coef. de correlación lineal poblacional
$\sigma^2_{x_1x_2}$: Covarianza poblacional
σ_{x_1} Desvío poblacional de x_1
σ_{x_2} Desvío poblacional de x_2

Donde:
$r_{x_1x_2}$: Coef. de correlación lineal muestral
$S^2_{x_1x_2}$: Covarianza muestral
S_{x_1} Desvío muestral de x_1
S_{x_2} desvío muestral de x_2

También recordemos que se aplica la correlación lineal cuando se desea saber si dos variables están asociadas linealmente.

Prueba de hipótesis para ρ

Hipótesis:
$H_0 : \rho_{x_1x_2} = 0$
$H_1 : \rho_{x_1x_2} \neq 0$

Parámetro a poner a prueba: $\rho_{x_1x_2}$

Estimador: $r_{x_1x_2}$

Estadístico: $t_0 = \frac{\bar{r}-\rho}{\sqrt{S^2_r}}$, siendo $S^2_r = \frac{1-r^2}{n-2}$

Confianza: 1-α

$$n = n_0$$

Regla de decisión:
Si $\left(t_{n-1;\,\alpha/2} < t_0 < t_{n-1;1-\alpha/2}\right)$, entonces no se rechaza H_0
Si $(-\infty < t_0 < t_{n-1;\,\alpha/2}]$, ó $[t_{n-1;1-\alpha/2} < t_0 < \infty$, entonces se rechaza H_0

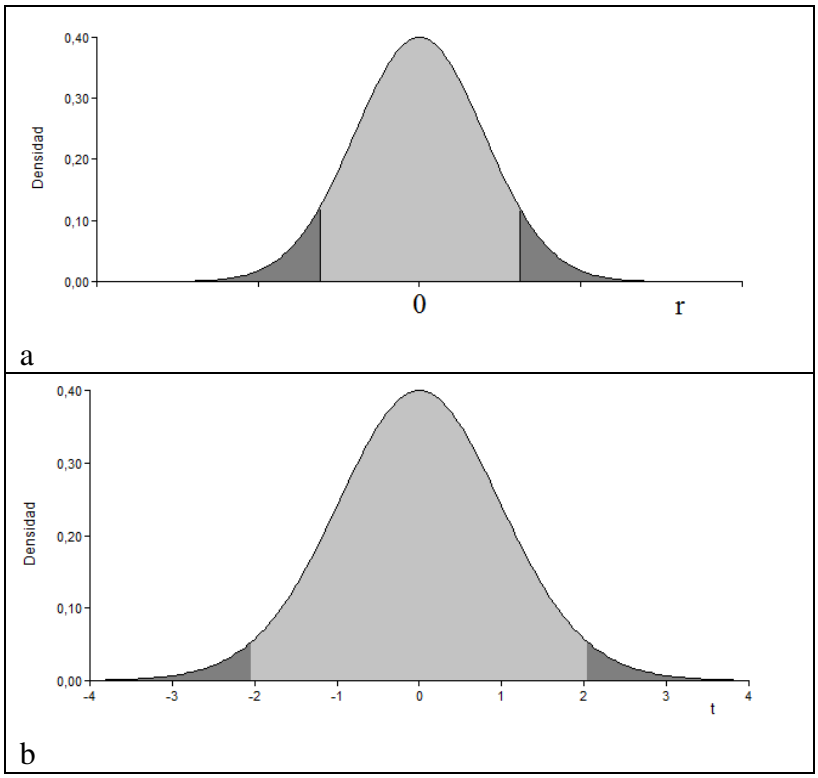


Figura 11.1a: Distribución de r si es cierta la hipótesis nula de la prueba de hipótesis para rho. b: Distribución t de la misma.

Esta prueba posee como supuestos que cada una de las dos variables (x_1 y x_2) posean distribución Normal y que entre x_1 y x_2 posean homogeneidad de varianzas. En el caso que algunos de estos dos supuestos no se cumplieran, el investigador tiene dos opciones: transformar una o las dos variables o aplicar el coeficiente de correlación lineal de Spearman, que se lo conoce como ρ_s, que no posee estos supuestos y también posee un recorrido de: $-1 < \rho_s < 1$ y una prueba de hipótesis correspondiente.

Ejemplo 11.1. Se quiere saber si están correlacionados linealmente las concentraciones de Bario y de Manganeso en afloramientos epitermales. Para ello se toman n unidades, se calculan la Covarianza y el r, luego se realiza la prueba de hipótesis para asegurar que r difiera significativamente de cero.

Regresión lineal

Debido a que los conceptos de correlación y de regresión poseen muchas fórmulas en común, existen confusiones sobre su utilización. Recordemos que los análisis estadísticos responden a hipótesis. El investigador debe tener en claro la hipótesis para decidir cuál es el modelo adecuado. Es absolutamente inadmisible la idea que un investigador primero "pruebe" hacer correlaciones y luego regresiones, ya que sus objetivos son diferentes.

Vamos a llamar Regresión lineal a aquel método estadístico que propone un modelo matemático lineal y que observa la dependencia de una variable "Y" con respecto a otra

variable "X" llamada independiente o regresora. Quien decide cuál de las variables es la X y cuál es la Y es el área de aplicación, en este caso la Geología.
A modo de ejemplo: En el caso de la relación entre el peso y la altura de las personas adultas, desde el punto de vista estadístico cualquiera de las dos podría ser la variable dependiente, pero la diferencia es la interpretación médica:

- Si la variable x es el peso, mientras que la y es la altura, una persona de 1,70 m y 200 kg es una persona "petisa o baja para su peso", le falta crecer para tener la altura que le corresponde por pesar 200 kg.
- Si la variable x es la altura, una persona de 1,70 m y 200 kg es una persona "gorda o pesada para su altura", necesita adelgazar para pesar lo que le corresponde por medir 1,70.

En cualquiera de los dos casos anteriores se puede hacer el ajuste estadístico. La lógica del campo de aplicación es la que impone si es factible que una variable dependa o no de otra ya que las regresiones no prueban causalidad.

Vamos a desarrollar el análisis de regresión lineal con un ejemplo.

Ejemplo 11.2. Un Geólogo desea predecir el caudal máximo de los ríos de una región de glaciares, en función la superficie de la cuenca de drenaje. Por lo tanto el investigador decide que la variable independiente o regresora será superficie de la cuenca y la dependiente será caudal. Los datos son:

x	128	140	161	186	199	216	231	244	250
y	33	41	42	49	60	59	75	76	81

Tabla 11.1. Valores de x e y del ejemplo 11.2.

El gráfico que corresponde se presenta en la Figura 11.2.

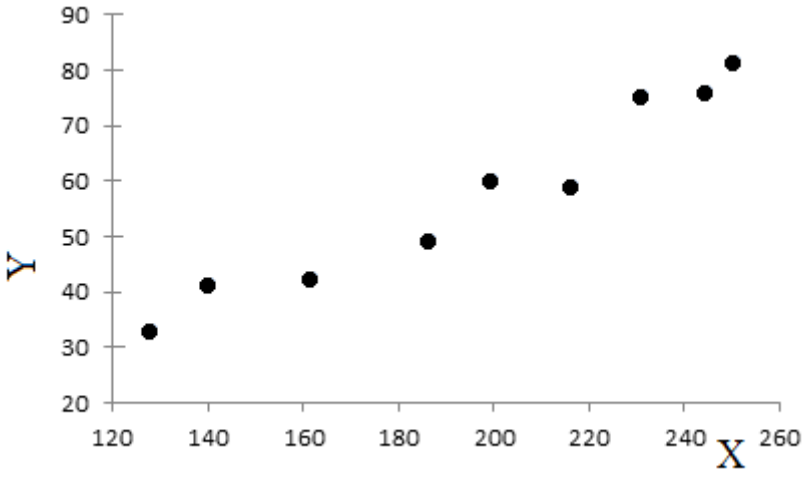


Figura 11.2. Relación lineal entre una variable X (independiente) y una variable Y (dependiente).

Ahora pensemos en que el modelo más sencillo para ajustar a esta nube de puntos es la ecuación de la recta, la cual vamos a definir con el **Modelo estadístico**:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Donde:

y_i : Variable respuesta, en este caso, caudal.
β_0 : Ordenada al origen, valor de y cuando x=0
β_1 : Pendiente, cambios ocurridos en y por cada incremento en una unidad de x
x_i : Variable regresora o independiente, en este caso, superficie
ε_i : Error o variabilidad no explicada, cada valor en particular constituye un residuo

Supuestos

$\varepsilon_i \sim NI(0;\ \sigma^2)$, Errores con distribución normal, independientes y homogeneidad de varianzas.

Decíamos que podríamos ajustar una recta a la nube de puntos establecida en la Figura 11.2, pero como se puede ver podrían presentarse infinitas rectas. Debemos ajustar una recta única, la mejor recta con algún criterio. Por esa razón el método más utilizado de ajuste se denomina método de los mínimos cuadrados.

Método de los mínimos cuadrados.

Este método se define como aquel que consigue sólo una recta que tenga como condición que la sumatoria de los errores al cuadrado sea un mínimo, y del mismo modo se consigue que la sumatoria de los errores sea cero:

$$\sum e^2 = mínimo \text{ y } \sum e \quad = 0$$

Se pueden resolver en una demostración donde se destaca que si $\hat{y}$ es el valor estimado por la recta, entonces:

$$\sum e^2 = \ mín;\ \sum (y - \hat{y})^2 = mín\ ,$$

$$\sum (y - (b_0 + b_1 x))^2 = mín$$

Siendo b_0 estimador de β_0, mientras que b_1 es el estimador de β_1
Finalmente, realizando las derivadas parciales para b_0 y b_1 se obtiene:

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$b_1 = \frac{S_{xy}^2}{S_x^2}$$

De este modo vamos a obtener una única recta que graficada será:

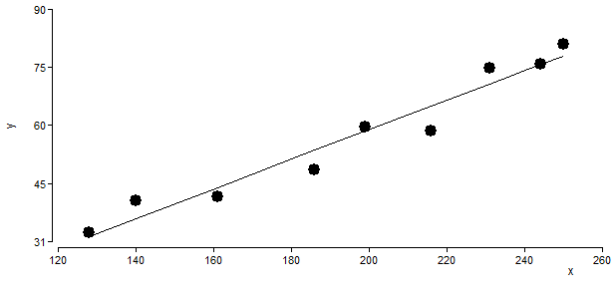


Figura 11.3. Ajuste de una función lineal entre X e Y.

Donde la suma de los errores será cero y la suma de los cuadrados de los mismos será un mínimo (Figura 11.4).

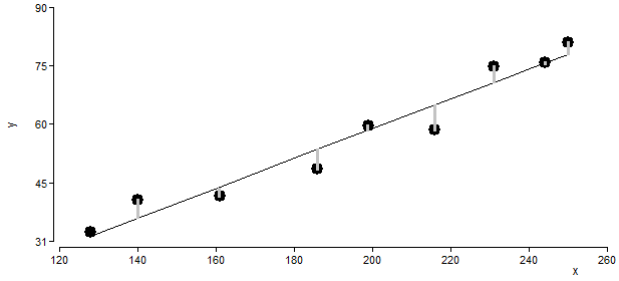


Figura 11.4. Ajuste de una función lineal entre X e Y, donde se expresan los errores para cada valor de y.

Entonces, aplicando las fórmulas descriptas anteriormente, los resultados en el ejemplo 11.2 serán:

b_0= -16,39
b_1= 0,3781

Ahora debiéramos corroborar si este valor de la pendiente muestral no proviene de una población cuya pendiente poblacional sea cero. Para ello debiéramos realizar las pruebas de hipótesis correspondientes.

Prueba de hipótesis para $\boldsymbol{\beta_1}$

Hipótesis:

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

Parámetro a poner a prueba: β_1

Estimador: b_1

Estadístico: $t_0 = \frac{b_1 - \beta_1}{\sqrt{S^2_{b_1}}}$, siendo $S^2_{b_1} = \frac{S^2_e}{\sum x^2 - n\bar{x}^2}$ y $S^2_e = \frac{\sum(y-\hat{y})^2}{n-2}$

Confianza: 1-α
n:
Regla de decisión:
Si $\left(t_{alfa} < t_0 < t_{alfa}\right)$, entonces no se rechaza H_0
Si $(-\infty < t_0 < t_{alfa}]$, ó [$t_{alfa} < t_0 < \infty$, entonces se rechaza H_0

En todas las regresiones es necesario poner a prueba la pendiente, pero sólo cuando los objetivos lo requieren se puede poner a prueba también la ordenada al origen:

Prueba de hipótesis para β_0

Hipótesis:

$$H_0: \beta_0 = 0$$
$$H_1: \beta_0 \neq 0$$

Parámetro a poner a prueba: β_0

Estimador: b_0

Estadístico: $t_0 = \frac{b_0 - \beta_0}{\sqrt{S^2_{b_0}}}$, siendo $S^2_{b_0} = \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum x^2 - n\bar{x}^2}\right) S^2_e$

Confianza: 1-α
n:
Regla de decisión:
Si $\left(t_{alfa} < t_0 < t_{alfa}\right)$, entonces no se rechaza H_0
Si $(-\infty < t_0 < t_{alfa}]$, ó [$t_{alfa} < t_0 < \infty$, entonces se rechaza H_0

Para el caso del ejemplo:

b_0= -16,39, t_0= 11,68; t_t= -2,36; 2,36. Por lo tanto se rechaza H0 y $\beta_0 \neq 0$
b_1= 0,3781 t_0= -2,53; t_t = -2,36; 2,36. Por lo tanto se rechaza H0 y $\beta_1 \neq 0$

Una vez que se ha probado que existe pendiente, es decir que la x explica a la y, podemos inferir los valores esperados de "y" para ciertos valores de "x". Es imprescindible notar que sólo los podemos hacer en el rango de valores que se utilizó la x. En el modelo del ejemplo entre 128 y 250. Para valores de caudales menores a 128 o mayores a 250 no se puede predecir ya que no sabemos si la dependencia entre las variables sigue siendo lineal y con ese mismo modelo.
La función ha quedado:

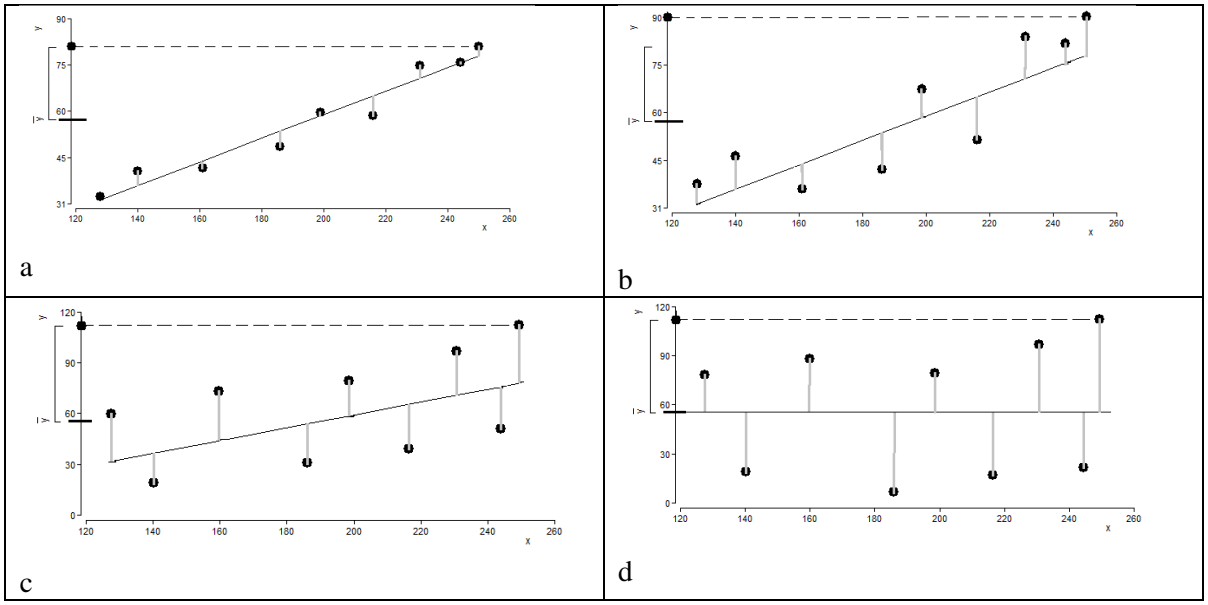
Caudal= -16,39+ 0,3781 superficie. De ese modo entre los valores mínimos y máximos de superficie (entre 128 km^2 y 250 km^2) podremos inferir el caudal. ¿Qué caudal traerá un río de una cuenca de 200 km^2)?

Caudal= -16,39+ 0,3781 *200.= 59,23 m^3/s

Vale recordar que la ordenada al origen nos estaría diciendo que un río con una superficie de 0 km^2 arrastraría un caudal de -16 m^3, lo cual es cierto sólo a los fines del modelo estadístico y no tiene ninguna interpretación ni lógica ni geológica.

Ajuste del modelo.

Hasta el momento hemos explicado el modelo, las pruebas de hipótesis del mismo, pero no hablamos sobre cuánto está siendo explicado por el modelo, para ello veremos tres gráficos:



Figuras 11.5 Representación del valor relativo del error con respecto a la distancia de cada punto a la media de y. Desde a hasta c se representa la misma pendiente, pero con incremento del error. En d se representa una pendiente cero, donde x no explica a y.

En las Figuras 11.5 vemos que desde 11.5a hacia 11.5c se va reduciendo el error, es decir la variabilidad no explicada. El error de sólo un punto es comparado con la distancia desde ese punto a la media de y.

Coeficiente de Determinación

Definimos a R^2 (Coeficiente de determinación) a:

$R^2 = 1 - \frac{SCE}{SCT}$, siendo SCE el denominador de la S^2_e, mientras que $SCT = \sum(y - \bar{y})^2$

Cuando el error es muy pequeño (la distancia entre los puntos y la recta), el R^2 se va acercando al valor uno, mientras que si el error es muy grande, el valor de R^2 se va acercando a cero.

Entonces podemos decir que el Coeficiente de determinación mide el grado de variabilidad de "y" que está siendo explicado por "x" mediante el modelo.
El recorrido del este coeficiente es: $0 \leq R^2 \leq 1$ y para el caso de nuestro ejemplo:

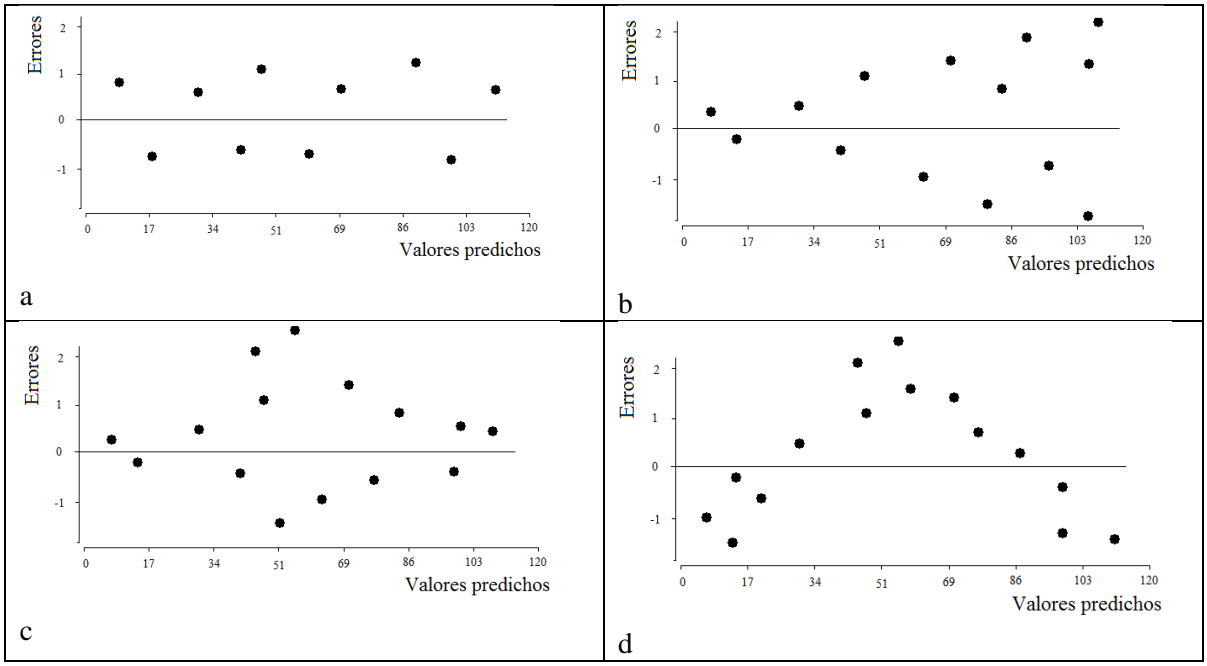
$$R^2 = 1 - \frac{116{,}75}{2394{,}0} = 0{,}9512$$

Esto significa que el 95,12% de la variabilidad de y está siendo explicada por x

Validez del modelo

El modelo será válido sólo para el caso en que se cumplan los supuestos.
Se deberá poner a prueba el supuesto de normalidad y se podrá realizar un QQplot para realizar una verificación visual.
No existe una prueba de homogeneidad de varianzas, por lo que se debe recurrir exclusivamente al gráfico de errores vs predichos. En él se deben observar:

- Por un lado corroborar que las varianzas al comienzo, al centro y al final de la línea de predicción resulten similares.
- Por otro lado se debe observar que no exista ningún tipo de patrón. A modo de ejemplo se presentan las Figuras 11.6.



Figuras 11.6. Gráficos de predichos vs errores. a: No se observa ningún patrón. b: se observa que a medida que aumenta el valor predicho, se incrementa la varianza. c: se observan valores centrales con mucha varianza y los extremos con poca varianza. d: se observa un patrón que implica que el modelo no es adecuado, ya que los valores centrales son todos mayor al predicho, mientas que los extremos son siempre menores.

Ejemplo 11.3. A los fines de predecir la concentración de ciertas variables en agua subterránea, se realiza una regresión lineal simple de modo que se pueda predecir ésta (y) mediante un método eléctrico (x).

Regresión lineal múltiple.

Como una ampliación del modelo de regresión lineal simple se pueden ir incorporando variables independientes que expliquen a la dependiente, tal que:

$$y \quad = \beta_0 + \beta_1 x_1 + \ \beta_2 x_2 + \beta_3 x_3 + \cdots \beta_p x_p + \varepsilon_i$$

Donde:

y_i : Variable respuesta, en este caso, caudal.
β_0 : Ordenada al origen, valor de y cuando x=0
β_1 : Pendiente de la variable x_1.
x_1 : Variable regresora o independiente, x_1.
β_2 : Pendiente de la variable x_2.
x_2 : Variable regresora o independiente, x_2.
ε_i : Error o variabilidad no explicada.

Supuestos

$\varepsilon_i \sim NI(0;\ \sigma^2)$, Errores con distribución normal, independientes y homogeneidad de varianzas. No existe correlación entre cada una de las x_i.

Al ajustar el modelo, se van a encontrar una pendiente para cada variable independiente y se realizará una prueba de hipótesis para corroborar que sea significativa su influencia en y.

Ejemplo 11.4. Se quiere predecir el nivel de agua que tuvo cierta laguna en los últimos 500 años. Para ello se dispone como variable dependiente el nivel de agua y como independientes: x_1: isótopos de Carbono y x_2: Concentración de sólidos disueltos en los sedimentos.

Modelos no lineales.

Existen dos tipos de modelos no lineales:

a) Modelos linealizables.
Se trata de relaciones entre variables donde una sencilla transformación permite linealizar el modelo y someterlo a un ajuste por método de mínimos cuadrados. Son ejemplos el modelo exponencial, potencial, entre otros.

Ejemplo 11.5. A los fines de una datación se estudia la relación entre el C_{14} y el tiempo transcurrido la que posee un ajuste no lineal, se transforma a logaritmo a una de las dos variables para linealizar la función.

Ejemplo 11.6. El modelo:
$y = b_0 * b_1^X$ se linealiza mediante: $\log y = \log b_0 + \log b_1 * X$

b) Modelos no linealizables
Se trata de modelo que necesitan otros métodos de estimación, como por ejemplo algoritmos iterativos, que son computacionalmente complejos.
Se puede dar como ejemplo el modelo logístico, donde

$y = \frac{\text{alfa}}{1+\text{beta}*e^{(-\text{gamma}*X)}}$, siendo alfa, beta y gamma, tres parámetros de la ecuación

Criterios de Ajuste del Modelo

Hemos visto al Coeficiente de Determinación R^2 como un criterio de ajuste del modelo. Este no posee inconvenientes cuando la ecuación es lineal y cuando es una regresión simple. Cuando incorporamos en una regresión lineal más variables X, el R^2 comienza a sobreestimar la varianza estimada, es decir si se le agregan al modelo variables x que no explican mucho, igual se incrementará el valor de R2.
Por esa razón se creó el R2 ajustado, que penaliza al modelo por colocar en él a variables que "gastan" más grados de libertad que la magnitud de la variabilidad que explican.

$R^2_{Ajust} = 1 - (1 - R^2) * (\frac{n-1}{n-p})$; donde:
p: es el número de parámetros dentro del modelo.

En modelos no lineales para la elección del modelo se presentan otros criterios como el criterio de Akaike o el criterio de Schwartz (BIC) en los que también la idea es cuantificar la variabilidad que queda sin explicar mediante los grados de libertad consumidos por el modelo.

Capítulo 12.

Introducción a la Geoestadística

Introducción.

A partir de los años ´50 y principios de los ´60 los las investigaciones relacionadas a la Geología intentaron suplir una problemática que tenía la estadística clásica cuando se la aplicaba a estimaciones de recursos, minería y petróleo entre otras. El problema estaba relacionado con la falta de independencia de las muestras. ¿Cuán separadas debían estar dos muestras geológicas para ser consideradas independientes? Si se tomaban dos muestras geológicas demasiado cercanas no eran independientes y si estaban demasiado lejanas no se podía inferir a los casos que ocurrían entre ellas. Por esa razón, Danie Krige (1919-2913) un Ingeniero en Minas sudafricano y Georges Matheron (1930-2000) un matemático francés y un grupo importante de estadísticos y geólogos desarrollaron las bases de lo que hoy se conoce como Geoestadística.

La Geoestadística es una rama de la estadística que trata fenómenos con distribuciones espaciales o temporales. Estudia particularmente variables distribuidas en el espacio que poseen cierta dependencia entre ellas. El interés principal suele ser la estimación, predicción y simulación de esas variables a las que se las denominan **variables regionalizadas**.

Autocorrelación

Definimos a autocorrelación como la correlación entre elementos de una serie y otros de la misma serie, separados por intervalos definidos de tiempo o distancia.

Ejemplo 12.1. Se ha tomado una transecta longitudinal. Cada 10 m se realizó un análisis de porosidad de la roca base.

metros	10	20	30	40	50	60	70	80	90	...
porosidad	88,07	94,21	97,39	100,27	104,23	106,86	108,89	111,6	118,01	...

Entonces a la serie (lista de datos en el orden tomados) que el investigador posee se la llama “cabeza” (head), es decir en el ejemplo 12.1: 88,07; 94,21; 97,39;....
Si se la desea correlacionar con un “paso o salto” (lag) de 10 m entonces la serie “cola” (tail) será: 94,21; 97,39; 100,27.... Con un paso de 20 m será: 97,39; 100,27; y así sucesivamente.
De este modo construimos covarianzas y coeficientes de correlación lineal del siguiente modo:

Covarianza

$$\sigma^2_{x_1 x_{1+h}} = \frac{\sum_{i=1}^{N}(x_{1i}-\mu_1)\ (x_{1+hi}-\mu_{1+h})}{N}$$

Donde:

Coeficiente de correlación lineal

$$\rho_{x_1 x_{1+h}} = \frac{\sigma^2_{x_1 x_{1+h}}}{\sigma_{x_1}\ \sigma_{x_{1+h}}}$$

Donde:

$\sigma^2_{x_1 x_{1+h}}$ Covarianza $\rho_{x_1 x_{1+h}}$ Correlación lineal
h: período de espacio o tiempo (en este caso 10 m)
x_1: variable cabeza
x_{1+h}: variable cola (cabeza desfasada por período h)

Se obtendrán tantas correlaciones como a tanta distancia de tiempo o espacio se desee correlacionar. Nótese que del n original de la variable cabeza, en cada paso se pierden 2 valores de la variable para la correlación siguiente. Es decir si originalmente la variable tiene 50 valores, en el paso 1 la correlación se realizará con 48, en la siguiente con 46 y así sucesivamente.

Funciones de correlación espacial.

Además de las ya conocidas covarianza y correlación para mensurar la correlación espacial o temporal se utiliza la semivarianza, la que se define como:

Semivarianza

$$\gamma_h = \frac{\sum_{i=1}^{N}(x_{1+hi} - x_{1i})^2}{2N_{p(h)}}$$

Donde:
γ_h : Es la semivarianza
$N_{p(h)}$: Número de pares a la distancia h

Entonces la semivarianza es una medida de la disimilaridad o discrepancia. Se espera que a medida que se incrementa la distancia entre dos puntos, ese valor vaya incrementándose, hasta estabilizarse. El gráfico del cambio en semivarianza a lo largo de la distancia se denomina semivariograma (Figura 12.1a), mientras que el de la correlación se denomina correlograma.

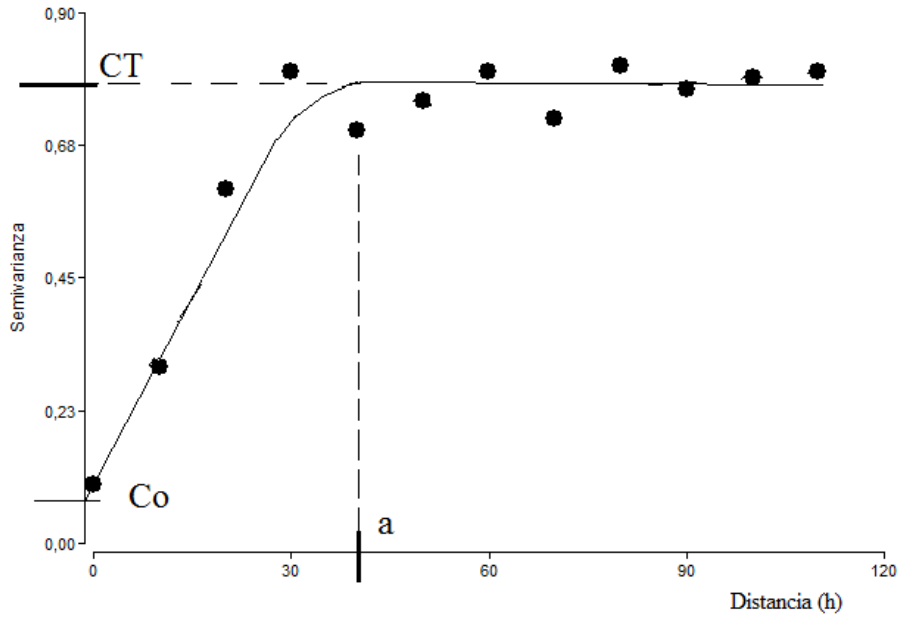


Figura 12.1. Semivariograma del Ejemplo 12.1. Se observan Co: efecto pepita, CT: meseta y a: alcance.

Interpretación del semivariograma.

De la figura 12.1 se puede destacar:

1) **Efecto Pepita (Nugget):** El semivariograma por definición es nulo en el origen, pero en la práctica las funciones obtenidas pueden presentar discontinuidad en el origen, a esta discontinuidad se le llama efecto de pepita (Nugget effect). Puede ser obtenido trazando una línea recta entre los primeros puntos del semivariograma empírico y extender ésta hasta que se intercepte con el eje Y. Si esta intersección ocurre por debajo de cero, el valor asumido por este efecto es cero, pues valores negativos no tienen significado y no es común. El efecto pepita se representa como Co.

2) **Meseta (Sill):** Es el valor de γ_h para el cual con el aumento de h, su valor permanece constante, se representa como (CT = C + Co) y se denomina meseta. Puede obtenerse trazando una línea paralela al eje X de modo que ajuste a los puntos de mayor valor del semivariograma.

3) **Alcance (Range):** La distancia h para la cual las variables x_1 y x_{1+h} son independientes, se denomina alcance y se representa con la letra "a". Nos dice la distancia en la cual los valores de la variable dejan de estar correlacionados, o lo que es lo mismo, la distancia en la cual el semivariograma alcanza su meseta.

Modelando el semivariograma

En diversos semivariogramas la forma o función pueden ser diferentes. Así es el caso que pueden o no tener efecto pepita, que la forma en que se incrementa la semivarianza sea lineal, esférica o adquiera diferentes formas.

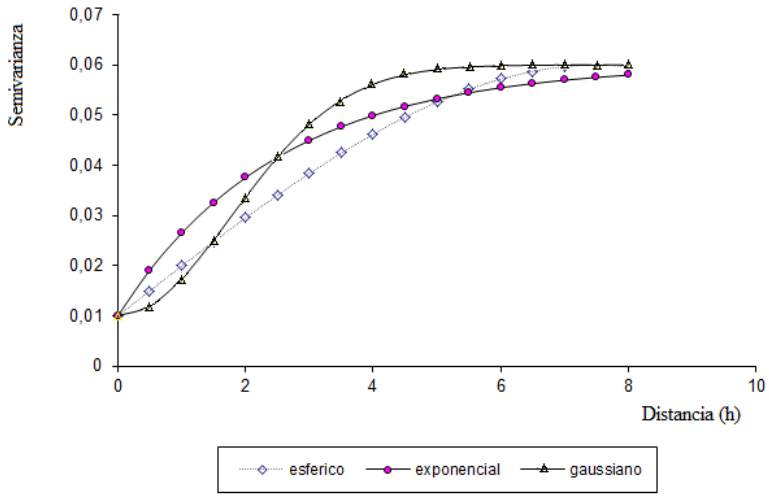


Figura 12.2: Comparación de modelos esférico, exponencial y gaussiano, con valores de Co=0.01, C=0,05 y a= 7,5; siendo h la variable independiente.

Modelo esférico

Llamado también como matheroniano, el valor de semivarianza se modela:

$$\gamma_h = C_0 + C * 1{,}5\frac{h}{a} - 0{,}5 * \left(\frac{h}{a}\right)^3$$

a = alcance
C_0 = efecto pepita
C= diferencia entre efecto pepita y meseta. CT= C_o + C
h= paso (lag)

Modelo exponencial

Llamado también como Formeryano, el valor de semivarianza se modela:

$\gamma_h = C_0 + C\ (1 - e^{-\left(\frac{3h}{a}\right)})$, para el caso en que h ≤ a

$\gamma_h = C_0 + C$, para el caso en que h > a

Modelo Gaussiano (parabólico).

La función puede ser encontrada como:

$\gamma_h = C_0 + C\ (1 - e^{-\left(\frac{3h}{a}\right)^2})$,

O bien:

$$\gamma_h = C_0 + C\,(1 - e^{-\left(\frac{h}{a}\right)^2}),$$

Modelo Wijsiano (sin efecto pepita)

$$\gamma_h = 3\ \mathrm{Ln}\ (h)$$

Otros modelos:

También se han desarrollado el Modelo lineal, el Modelo logarítmico, el Modelo con efecto de agujero, el Modelo transitivo, entre otros.

En el ejemplo 12.1 aplicamos la idea de una correlación en una línea recta, pero en la realidad, un geólogo deberá trabajar con datos espaciales, 2D o 3D. La figura 13.3 presenta la información recabada en un muestreo donde se debe realizar un semivariograma entre los puntos cercanos tanto en latitud como en longitud. Es obvio que en canteras o minas además se debe trabajar en 3 dimensiones por lo que las correlaciones se deberán calcular tanto en latitud como en longitud y en profundidad.

0 m	44		40	42	40	39	37	36	
100 m	42		43	42	39	39	41	40	38
200 m	37	37	37	35	38	37	37	33	34
300 m	35	38		35	37	36	36	35	
400 m	36	35	36	35	34	33	32	29	28
500 m	38	37	35		30		29	30	32
	0 m	100 m	200 m	300 m	400 m	500 m	600 m	700 m	800 m

Figura 12.3. Diagrama de los valores de ley de un muestreo en forma de grilla. Cada punto se encuentra a 100 metros del vecino más cercano

Inferencia

Una vez estudiado de qué forma se diferencian dos puntos de muestreo a lo largo de la distancia, nos haremos una pregunta: Si observamos la Figura 12. 3, ¿Qué concentración (ley) de mineral pensaríamos que va a tener el punto x*, si los otros dos x_1 y x_2 poseen valores de 50 y 10?

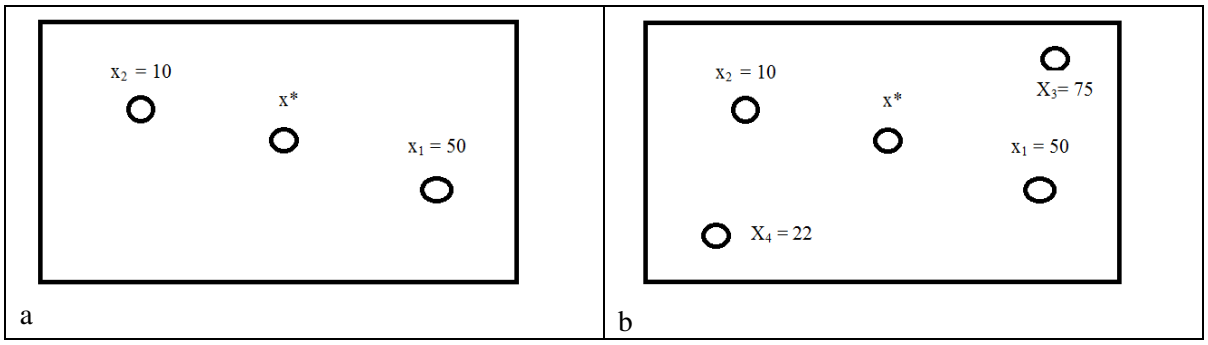


Figura 12.4. Ejemplo de puntos de muestreo y la necesidad de inferir sobre una incógnita.

Pensando en inferir el valor de x* en la figura 12.4a, si el decrecimiento de la ley de este mineral fuese lineal, pensaríamos que debiera tener un valor de 30, es decir el promedio entre los dos valores. Pero si en el valor 50 correspondiese a una pepita, muy cerca de 50 serían todos valores cercanos a 10 y el x* debiera ser 10. Y así se podrían pensar en muchas situaciones diferentes. Del importantísimo estudio sobre los semivariogramas depende de qué valores estimará el Geólogo para x*. Del mismo modo y en el caso de la figura 12.4b, ¿cuál de los puntos debiera tener mayor peso sobre x*?, el x_1 está más cerca y vale 50, el x_3 está más lejos y vale 75. Debiera pensarse en estimar x* con un aporte proporcional de cada uno de los vecinos, para ello se han propuesto diversas formas de estimación ponderada, el más relevante es el Kriging.

Kriging.

Entendemos por Kriging a un método de interpolación óptima basado en regresiones entre los valores observados de puntos vecinos, pesados de acuerdo a sus valores de correlaciones espaciales. El término Kriging se colocó en homenaje a Danie Krige, uno de los padres de la Geoestadística y en definitiva se entiende como el primer método de estimación de datos intermedios que se inventó, hecho para variables autocorrelacionadas, y que se basa en funciones de semivariogramas.

El valor x* estimado por el Kriging va a tener un error o residuo, es decir la diferencia entre ese x* y $\bar{x}$ (la media de la muestra)
Ese error se va a calcular pesando la suma de los errores de todos los datos vecinos de modo tal que si un dato está más lejos influye menos en el error de x* que si está más cerca. Ese peso lo imprime en la función un término llamado lambda (λ) que deriva de la función del semivariograma.

$$x^* - \bar{x} = \sum \lambda_i \, (x_i - \bar{x})$$

Donde λ es el factor de ponderación que por definición $\sum \lambda_i = 1$

La solución para obtener el valor de λ de cada punto para estimar a x* es muy compleja. A partir de la premisa que se obtiene un x* que cumple con la propiedad de ser un valor tal que minimice la varianza del error de predicción se plantean ecuaciones lineales que optimizan ese valor.

Existen diversas variaciones del método de Kriging. Mientras que el Kriging ordinario exige que los estimadores sean insesgados, el Kriging simple relaja esa condición y el Kriging por bloques requiere que un Geólogo conocedor de la zona defina bloque con diferentes características geológicas dentro de la región.

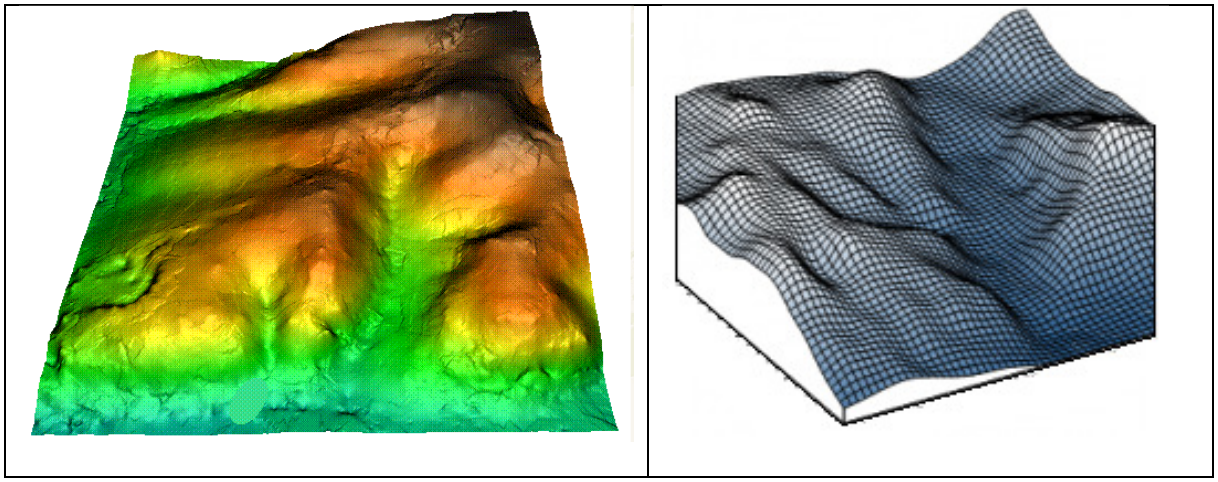


Figura 12.5. Estimaciones de ley de dos minerales diferentes mediante el método de Kriging. En los tres ejes se representa latitud, longitud y concentración.

Validación del Kriging

El más empleado de los métodos de validación es el de "validación cruzada". Este método consiste en excluir uno a uno a todos los n puntos muestreados y con los n-1 valores restantes, predecir el valor de la variable respuesta del dato excluido.
Se realiza este procedimiento para cada uno de los modelos de semivariograma que pueden ser óptimos para la inferencia y se elige el que menores errores de predicción posea entre lo observado y lo esperado.
Una forma descriptiva de hacer la validación cruzada es mediante un gráfico de dispersión de los valores observados contra los valores predichos. En la medida en que la nube de puntos se ajuste más a una línea recta que pase por el origen, mejor será el modelo de semivariograma utilizado para realizar el Kriging.

Intervalos de confianza para la estimación puntual.

Del mismo modo en que estudiamos en el Capítulo 8 a los estimadores puntuales y sus intervalos de confianza, aquí el método de Kriging también otorgará intervalos para cada punto estimado. Por ejemplo el punto estimado de la Figura 12.4.b. se modela con un semivariograma de función esférica, el Kriging arroja un valor de 31 ± 4 g/Tn, lo que significa que la ley de ese mineral estimada es 31 g/Tn, pero si se desea ser conservador el valor será 27 g/Tn y para una estimación liberal se pensará en 35 g/Tn.

La estadística es sólo una herramienta que no puede reemplazar el conocimiento a campo como del génesis de la roca que el Geólogo posee. Como puede observarse dependiendo si

se utiliza un modelo de semivariograma u otro, las estimaciones resultarán distintas. Del mismo modo si se estima con los límites menores o mayores las diferencias son sustanciales. Esas decisiones las deberá tomar el Geólogo basándose en su conocimiento del terreno y no sólo de las estimaciones estadísticas.