INFERENCIA ESTADISTICA

TEORÍA ELEMENTAL DE MUESTREO

La teoría de muestreo se refiere al estudio de las relaciones que existen entre un colectivo o población y las muestras que se extraen de las mismas. El estudio de las muestras permite hacer estimaciones de características desconocidas de la población (tales como media, desviación típica, proporciones, etc). Estas estimaciones se hacen a partir del conocimiento de las características de las muestras (media, desviación típica, proporción, etc).

Las características o medidas obtenidas de una muestra se llaman estadísticos; y las medidas correspondientes a la población parámetros. Cuando una medida muestral o estadístico es utilizada como representante de una característica poblacional o parámetro se denomina estimador.

Ventajas de la utilización de las muestras

1) El costo es menor y se puede obtener un mejor rendimiento del dinero invertido.
2) Se obtiene una disminución notable del tiempo necesario para alcanzar la información

Cuando una muestra posee 30 o más datos se denomina grandes muestras y si la muestra tiene menos de 30 observaciones se denomina pequeñas muestras.

Al procedimiento utilizado para elegir una muestra se denomina Muestreo.

Necesidad del Muestreo.

1. Población Infinita
2. Población uniforme
3. Proceso de investigación destructiva
4. Economía de costos
5. Calidad

Muestreo con o sin reemplazamiento:

- Con reemplazamiento cuando un elemento de la población puede ser escogido varias veces para formar parte de la muestra

- Sin reemplazamiento cuando un elemento de la población solo puede ser seleccionado una sola vez para formar parte de la muestra.

Población: es una colección de todos los elementos que estamos estudiando y acerca de los cuales se intenta extraer conclusiones. Puede ser infinita o finita.

Muestra: Una parte de la población o un subconjunto del conjunto de unidades obtenidas con el objeto de investigar las propiedades de la población.

Muestreo estadístico: Es un enfoque sistemático para seleccionar unos cuantos elementos (una muestra) de un grupo de datos (población) a fin de hacer algunas inferencias sobre el grupo total. Desde el punto de vista matemático, podemos describir las muestras y las poblaciones mediante medidas como la media, la moda, la desviación estándar, etc. No es mas que el procedimiento a través del cual se obtienen las muestras.

Tipos de muestreo

Muestreo de juicio o no probabilístico. (opinático). Se basa en el conocimiento de la población por parte de alguien, quien hace a la muestra representativa, dependiendo de su intención, por lo tanto es subjetiva.

Probabilístico(Errático): Todos los elementos de la población tienen la posibilidad de pertenecer a la muestra.

Muestreo Aleatorio:

1. Muestreo aleatorio simple
2. Muestreo Sistemático.
3. Muestreo Estratificado
4. Muestreo por Conglomerado

Muestreo de juicio: A través del conocimiento y la opinión personal, basada en la experiencia del investigador, se identifican los elementos de la población que van a formar parte de la muestra. Una muestra seleccionada por muestreo de juicio se basa en el conocimiento de la población por parte de alguien. Por ejemplo, un guardabosques tomará una muestra de juicio si decide con antelación que parte de una gran zona reforestada deberá recorrer para estimar el total de metros de madera que pueden cortarse. En ocasiones el muestreo de juicio sirve de muestra piloto para decidir cómo seleccionar después una muestra aleatoria.

INFERENCIA ESTADISTICA

Muestreo aleatorio: Cuando se conoce la probabilidad de que un elemento de la población figure o no en la muestra, puede ser:

Muestreo Aleatorio Simple (Irrestrictamente Aleatorio):

Un muestreo es aleatorio cuando cada elemento de la población tiene la misma probabilidad de ser escogido para formar parte de la muestra. Este tipo de muestreo evita que la muestra sea sesgada evitando por lo tanto que se realice una mala inferencia estadística. Por ejemplo, supóngase que un investigador quiera estimar el módulo de ruptura promedio de un material determinado formado por una población de tamaño

N = 500; por ser ensayos destructivos este quiere seleccionar una muestra de tamaño

n = 10 que le permita realizar la inferencia, ahora bien el criterio que usó el investigador para seleccionar dicha muestra fue el de tomar 10 materiales que estaban más próximos a él; evidentemente esta muestra no es representativa de la población, se dice que esta sesgada, por lo que la inferencia estadística que se realice será errónea. Por lo tanto, una muestra se dice que esta sesgada cuando los elementos seleccionados tenían mayor probabilidad de pertenecer a la misma.

Cómo hacer el muestreo aleatorio

La forma más fácil de realizarlo es usando números aleatorios, para esto se puede recurrir a una tabla o a un generador de números aleatorios. Actualmente, se recurre a computadora.

Muestreo Sistemático o Secuencial.

Los elementos se seleccionan de la población con un intervalo uniforme en el tiempo, en el orden o en el espacio. Por ejemplo, supongamos que se quiere estudiar una determinada característica de un producto fabricado en serie y se decide seleccionar a cada veinte producto hasta formar la muestra, para esto se escoge un punto aleatorio de arranque en los primeros veinte productos y luego se escoge cada vigésimo producto hasta completar la muestra. Una de las ventajas de este muestreo es cuando los elementos presentan un patrón secuencial, tal vez requiera menos tiempo y algunas veces cuesta menos que el método de muestreo aleatorio.

INFERENCIA ESTADISTICA

Muestreo Estratificado.

Para aplicar el muestreo estratificado, se divide la población en grupos homogéneos, llamados estratos, los cuales son heterógeneos entre si. Después se recurre a uno de dos métodos posibles:

a) Se selecciona al azar en cada estrato un número especificado de elementos correspondientes a la proporción del estrato de la población total
b) Se extrae al azar un número igual de elementos de cada estrato y damos un peso a los resultados de acuerdo a la proporción del estrato en la población total

El muestreo estratificado es adecuado cuando la población ya está dividida en grupos de diferentes tamaños y queremos reconocer este hecho. La ventaja de las muestras estratificadas, es que cuando se diseñan bien, reflejan más exactamente las características de la población de donde se extrajeron que otras clases de muestreo.

Muestreo por Conglomerado.

En el muestreo por conglomerados, se divide la población en grupos o conglomerados de elementos heterogéneos, pero homogéneos con respecto a los grupos entre si. Un procedimiento bien diseñado, de muestreo por conglomerados, puede producir una muestra más precisa a un costo mucho menor que el de un simple muestreo aleatorio. Se usa el muestreo estratificado cuando cada grupo presenta una pequeña variación en su interior, pero existe una amplia variación entre ellos. Se usa el muestreo por conglomerado en el caso contrario, cuando hay considerable variación dentro de cada grupo pero los grupos son esencialmente semejantes entre sí.

INFERENCIA ESTADISTICA

DISTRIBUCIONES MUESTRALES

1. DISTRIBUCIÓN MUESTRAL DE MEDIAS
2. DISTRIBUCIÓN MUESTRAL PARA DIFERENCIAS DE MEDIAS
3. DISTRIBUCIÓN MUESTRAL DE PROPORCIONES Y DIFERENCIAS
4. DISTRIBUCIÓN MUESTRAL DE VARIANZAS

Se define la distribución muestral de un estadístico (distribución de muestreo) en una población, como la distribución de probabilidad de todos los posibles valores que un estadístico puede asumir para cierto tamaño de la muestra. Específicamente, se trabajará con las distribuciones muestrales para: medias, proporciones y varianzas.

Una distribución muestral es una distribución de probabilidad de un estadístico muestral calculado a partir de todas las muestras posibles de tamaño n, elegidas al azar en una población determinada. Si la población es infinita, tenemos que concebir la distribución muestral como una distribución muestral teórica, ya que es imposible sacar todas las muestras aleatorias posibles de tamaño **n** de una población infinita. Si la población es finita y moderada se puede construir una distribución muestral experimental, sacando todas las muestras posibles de un tamaño dado, calculando para cada muestra el valor del estadístico que nos interesa. Ejemplo, supongamos que se tiene una población de tamaño N = 10 y queremos extraer con reemplazamiento todas las muestras posibles de tamaño n = 5, para esto se utiliza la relación $\mathbf{N^n}$, es decir,

$10^5 = 100000$ muestras de tamaño n = 5.

En cambio, si el muestreo es sin reemplazamiento, el número de muestras de tamaño N = 5 viene dado por la combinatoria:

$$\left(\frac{N}{n}\right) = \frac{N!}{n!(N-n)!} = \frac{10!}{5!(10-5)1} = \frac{10.9.8.7.6.5!}{5!.5.4.3.2.1} = 252 \text{ muestras.}$$

Por lo que considerando este caso, la distribución muestral para un estadístico determinado, por ejemplo, la media $\bar{X}$ viene dado por:

$$\begin{array}{l}\text{muestra } 1 \rightarrow \overline{X}_1 \\ \text{muestra } 2 \rightarrow \overline{X}_2 \\ \vdots \\ \text{muestra } 252 \rightarrow \overline{X}_{252}\end{array}$$

Esto es, $\overline{X}_1, \overline{X}_2, \overline{X}_3, \ldots, \overline{X}_{252}$ o sea, la distibución muestral de medias.

Se puede hacer una aproximación experimental de distribuciones muestrales basadas en poblaciones infinitas o finitas grandes, sacando un número de muestras aleatorias y siguiendo el mismo procedimiento anterior.

1) DISTRIBUCIÓN MUESTRAL DE MEDIAS:

Es la distribución de probabilidad de todas las medias posibles de las muestras, para un tamaño **n** determinado. Ver ejemplo, anterior. Esta distribución de probabilidad tiene asociados (parámetros) tales como la media $\mu_{\bar{X}}$ y desviación estándar $\sigma_{\bar{X}}$. Para calcular, estos parámetros de la distribución muestral de medias se utilizan las siguientes relaciones:

$$\mu_{\bar{X}} = \mu$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \quad \text{para poblaciones finitas}$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \quad \text{para poblaciones infinitas}$$

La expresión es la desviación estándar de la distribución muestral de medias, se le llama error típico o estándar de la media y nos indica la diferencia promedio entre los diversos valores de $\overline{X}$ y μ. Como se observa, a medida que el tamaño de la muestra aumenta este error dismunuye, las diversas medias muestrales se hacen más uniforme en su valor, y en consecuencia, cualquier media muestral es una buena estimación de la media poblacional **μ.**

INFERENCIA ESTADISTICA

Distribuciones Muestrales

Construcción

- De una población discreta, finita, de tamaño N, extraer todas las muestras posibles de tamaño n
- Calcular el valor del estadístico de interés de cada muestra
- Hacer una tabla con dos columnas: en la primera los posibles valores diferentes del estadístico y en la segunda, la frecuencia de ocurrencia.

Distribución Muestral de la Media

Una población consiste de 10 vendedores de una compañía. La variable de interés, X, es la antigüedad.

X = {1, 2, 3, 4, 5, 6, 7, 8, 9,10} Podemos calcular los siguientes

$$\mu = \frac{\sum x_i}{N} = \frac{55}{10} = 5{,}5$$

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} = 8{,}25$$

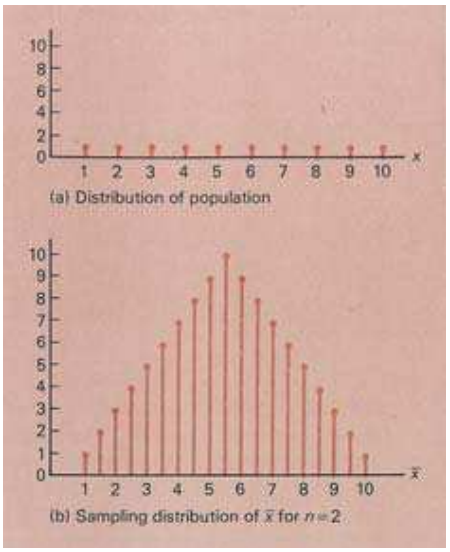
Distribución Muestral de la Media

1. Extraemos todas las posibles muestras. Supongamos n=2 (100 muestras).
2. Calculamos la media para cada una de esas muestra $\bar{x}$
3. Listar los valores diferentes del estadístico y sus frecuencias.

First draw	Second draw									
	1	2	3	4	5	6	7	8	9	10
1	1,1 (1)	1,2 (1.5)	1,3 (2)	1,4 (2.5)	1,5 (3)	1,6 (3.5)	1,7 (4)	1,8 (4.5)	1,9 (5)	1,10 (5.5)
2	2,1 (1.5)	2,2 (2)	2,3 (2.5)	2,4 (3)	2,5 (3.5)	2,6 (4)	2,7 (4.5)	2,8 (5)	2,9 (5.5)	2,10 (6)
3	3,1 (2)	3,2 (2.5)	3,3 (3)	3,4 (3.5)	3,5 (4)	3,6 (4.5)	3,7 (5)	3,8 (5.5)	3,9 (6)	3,10 (6.5)
4	4,1 (2.5)	4,2 (3)	4,3 (3.5)	4,4 (4)	4,5 (4.5)	4,6 (5)	4,7 (5.5)	4,8 (6)	4,9 (6.5)	4,10 (7)
5	5,1 (3)	5,2 (3.5)	5,3 (4)	5,4 (4.5)	5,5 (5)	5,6 (5.5)	5,7 (6)	5,8 (6.5)	5,9 (7)	5,10 (7.5)
6	6,1 (3.5)	6,2 (4)	6,3 (4.5)	6,4 (5)	6,5 (5.5)	6,6 (6)	6,7 (6.5)	6,8 (7)	6,9 (7.5)	6,10 (8)
7	7,1 (4)	7,2 (4.5)	7,3 (5)	7,4 (5.5)	7,5 (6)	7,6 (6.5)	7,7 (7)	7,8 (7.5)	7,9 (8)	7,10 (8.5)
8	8,1 (4.5)	8,2 (5)	8,3 (5.5)	8,4 (6)	8,5 (6.5)	8,6 (7)	8,7 (7.5)	8,8 (8)	8,9 (8.5)	8,10 (9)
9	9,1 (5)	9,2 (5.5)	9,3 (6)	9,4 (6.5)	9,5 (7)	9,6 (7.5)	9,7 (8)	9,8 (8.5)	9,9 (9)	9,10 (9.5)
10	10,1 (5.5)	10,2 (6)	10,3 (6.5)	10,4 (7)	10,5 (7.5)	10,6 (8)	10,7 (8.5)	10,8 (9)	10,9 (9.5)	10,10 (10)

INFERENCIA ESTADISTICA

$\bar{x}$	Frequency	Relative frequency	$\bar{x}$	Frequency	Relative frequency
1	1	1/100	6	9	9/100
1.5	2	2/100	6.5	8	8/100
2	3	3/100	7	7	7/100
2.5	4	4/100	7.5	6	6/100
3	5	5/100	8	5	5/100
3.5	6	6/100	8.5	4	4/100
4	7	7/100	9	3	3/100
4.5	8	8/100	9.5	2	2/100
5	9	9/100	10	1	1/100
5.5	10	10/100	Total	100	100/100



Calculamos la media de la distribución muestral con reemplazamiento

$$\mu_{\bar{x}} = \frac{\sum \bar{x}_i}{N^n} = \frac{550}{100} = 5{,}5$$

¿Calcular la media muestral sin reemplazamiento?

Calculando la varianza de la distribución muestral:

$$\sigma_{\bar{x}}^{\ 2} = \frac{\sum (\bar{x}_i - \mu_{\bar{x}})^2}{N^n} = \frac{412{,}5}{100} = 4{,}125$$

$$\sigma_{\bar{x}}^{\ 2} = \frac{\sigma^2}{n} = \frac{8{,}25}{2} = 4{,}125$$

Error estándar de la media: $\varepsilon = \frac{\sigma}{\sqrt{n}}$

INFERENCIA ESTADISTICA

Distribuciones Muestrales

Cuando el muestreo se extrae de una población distribuida normalmente, la distribución muestral de la media muestral tiene las siguientes propiedades:

1. La distribución de la media es normal, independientemente del tamaño de la muestra.
2. La Media de la distribución de las medias es igual a la media de la población.
3. La varianza de la distribución de las medias es igual a la varianza de la población, dividida entre n.

Teorema del Límite Central

Dada una población con media μ y varianza finita σ^2, con cualquier distribución, la distribución muestral de la media, calculada de muestras aleatorias de tamaño n, está distribuida normalmente con media μ y varianza finita σ^2/n, cuando n es grande.
La regla de oro dice que $n \geq 30$.

Ejemplo

La vida promedia de cierta herramienta es de 41.5 horas, con una desviación estándar de 2.5 horas. ¿Cuál es la probabilidad de que una muestra aleatoria de tamaño 50 extraída de esta población tenga una media entre 40.5 y 42 horas?

$$P(40{,}5 \leq \bar{x} \leq 42) = P(-2{,}86 \leq z \leq 1{,}43) = P(0 \leq z \leq 2{,}86) + P(0 \leq z \leq 1{,}43) = 0{,}9215$$

DISTRIBUCIÓN MUESTRAL PARA LA DIFERENCIA DE MEDIAS ($\overline{X}_1 - \overline{X}_2$).-

A veces interesa hacer inferencias sobre la diferencia poblacional de medias $\mu_1 - \mu_2$, o saber si es razonable concluir que dos medias poblacionales no son iguales, considerando que se tienen sendas muestras para las poblaciones 1 y 2, respectivamente, donde:

n_1 = tamaño de la muestra de la población 1
$\overline{X}_1$ = media de la muestra 1
σ_1^2 = varianza de la población 1
n_2 = tamaño de le muestra de la población 2
$\overline{X}_2$ = media de la muestra 2
σ_2^2 = varianza de la población 2

Entonces, la diferencia de las medias muestrales $\overline{X}_1 - \overline{X}_2$, estima a $\mu_1 - \mu_2$. La forma funcional de la distribución muestral de $\overline{X}_1 - \overline{X}_2$ depende de la forma funcional de las poblaciones donde se extraen las muestras tomando en cuenta:

- *Si ambas poblaciones son normales la distribución muestral de la diferencia de medias es normal.*
- *Si una o ambas de las poblaciones no es normal, la distribución muestral de las diferencias de medias $\overline{X}_1 - \overline{X}_2$ es normal si $n_1 + n_2 - 2 > 30$ (grandes muestras), este resultado se deduce del teorema del límite central*

En estos casos, los parámetros que definen esta distribución muestral de las diferencias de medias vienen dados por:

$$\mu_{\bar{X}_1-\bar{X}_2} = \mu_1 - \mu_2$$

$$\sigma_{\bar{X}_1-\bar{X}} = \sqrt{\frac{\sigma_1^{\,2}}{n_1} + \frac{\sigma_2^{\,2}}{n_2}}$$

El cual se aplica para dos casos específicos dependiendo de la muestra:

a) Para grandes muestras, cuando **$v = n_1+n_2 - 2 > 30$,** se trabaja con la distribución normal. En estos casos, estandarizando la diferencia de medias muestrales, se tiene:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^{\,2}}{n_1} + \frac{\sigma_2^{\,2}}{n_2}}}$$

b) Para pequeñas muestras, Cuando **$v = n_1 + n_2 - 2 < 30$, se trabaja con la Distribución t de Student.** Por lo tanto, el valor viene dado por:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_p^{\,2}}{n_1} + \frac{S_p^{\,2}}{n_2}}}$$

donde:

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

Ejemplo

Dos compañías fabrican lubricantes de alta temperatura, para el mismo mercado. La compañía A anuncia que en promedio, su lubricante deja de ser efectivo a 505 °F, con una desv. est. de 10 °F. La compañía B anuncia que su producto tiene una media de 475 °F, con una desv. est. de 7 °F. Suponga que una muestra de tamaño 20 para la primera compañía y otra independiente de tamaño 25 para la segunda son extraídas aleatoriamente. ¿ Cuál es la probabilidad de que la diferencia en temperatura promedio de falla para las dos muestras esté entre 25 y 35 °F?

Ejemplo

$$z = \frac{(x_1 - x_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$z_1 = \frac{25 - (505 - 475)}{\sqrt{\frac{10^2}{20} + \frac{7^2}{25}}} = -1.89$$

$$z_2 = \frac{35 - (505 - 475)}{\sqrt{\frac{10^2}{20} + \frac{7^2}{25}}} = 1.89$$

$$P(-1.89 \le z \le 1.89) = 0.9706 - 0.0294 = 0.9412$$

DISTRIBUCIÓN DE UNA PROPORCION MUESTRAL ($\hat{P}$).-

Se define una proporción muestral como el cociente:

$$p = \frac{\text{número de casos favorables}}{\text{total de casos}}$$

Por ejemplo: si de una población de N = 50, empleados de una empresa, 15 de ellos no cumplen con su horario de trabajo, la proporción de empleados que no cumplen horario con relación al total, viene dado:

P = 15/50 = 0,3; es decir, el 30 % de los empleados no cumplen su horario.

La proporción muestral ($\hat{p}$), se define como:

$$\hat{p} = \frac{\text{número de casos favorables}}{\text{tamaño de la muestra}}$$

Ejemplo:

Si se toma una muestra aleatoria de tamaño n = 1000 y 425 personas satisfacen un evento, entonces p = 425 / 1000 = 0,425. Esto significa que el 42,5 % de las personas satisfacen dicho evento.

INFERENCIA ESTADISTICA

La distribución de una proporción muestral, se define de una manera análoga a a la distribución de media, o sea:

Muestra 1---- $\hat{p}_1$

Muestra 2---- $\hat{p}_2$

Muestra 3---- $\hat{p}_3$

Muestra 252---- $\hat{p}_{252}$

De esta forma: $\hat{p}_1$, $\hat{p}_2$, $\hat{p}_3$,..., $\hat{p}_{252}$ corresponden a la distribución de una proporción muestral.

De acuerdo a lo expuesto, la distribución muestral de proporciones corresponde a una distribución de probabilidad de todas las proporciones posibles de las muestras, para un tamaño n determinado.

Los parámetros que definen esta distribución vienen dados por:

$$\mu_{\hat{p}} = \mu_p = P$$

$$\sigma_{\bar{X}} = \sqrt{\frac{p.q}{n}}\sqrt{\frac{N-n}{N-1}} \quad \text{para poblaciones finitas}$$

$$\sigma_{\bar{X}} = \sqrt{\frac{p.q}{n}} \quad \text{para poblaciones infinitas}$$

Para el cálculo de probabilidades relativa a proporciones, se trabaja de manera análoga al caso de la distribución muestral de medias.

Ejemplo: Un encuestador sabe que en cierta área el 20 % está a favor de las emisiones en bonos. Considerando una muestra de 64 personas, hallar la probabilidad de que la proporción muestral difiera de la proporción real a lo sumo en un 0,06.

Solución:

p = 0.20 proporción de personas de la población que están a favor de la emisión

$\hat{p}$ = proporción de personas de la muestra que están a favor de la emisión

entonces nos están pidiendo la siguiente probabilidad:

$$P(|\hat{p}-p| \le 0{,}06) = P\left(-\frac{0{,}06}{\sqrt{\frac{0{,}2.0{,}8}{64}}} \le \frac{\hat{p}-p}{\sqrt{\frac{p.q}{n}}} \le \frac{0{,}06}{\sqrt{\frac{0{,}2.0{,}8}{64}}}\right) = P(-0{,}27 \le Z \le 0{,}27) = 0{,}204$$

INFERENCIA ESTADISTICA

ESTIMACION DE PARAMETROS

a) *ESTIMACIÓN PUNTUAL*

Para estimar un parámetro **θ** de una población se toma una muestra representativa de la misma y se calcula el estadístico $\hat{\theta}$, el valor del estadístico se conoce como la estimación puntual del parámetro **θ**. Por ejemplo,

Parámetro	Estimación puntual
θ = μ	$\hat{\theta} = \overline{X}$ (media muestral)
θ = σ	$\hat{\theta} = S$ (varianza muestral)
θ = p	$\hat{\theta} = \hat{p}$ (proporción muestral)
θ = $\mu_1 - \mu_2$	$\overline{X}_1 - \overline{X}_2$ (diferencia muestral de medias)

b) *ESTIMACIÓN POR INTERVALOS DE CONFIANZA*

En la sección anterior se habló sobre la estimación puntual, una de sus desventajas es el hecho de no saber que tan próxima está del parámetro, es decir, cuando se obtiene una estimación $\hat{\theta}$, a partir de una muestra aleatoria de tamaño n, se desconoce que tan cerca (por defecto o exceso) está del parámetro a estimar θ. Por eso se utiliza frecuentemente otro tipo de estimación, la estimación por intervalos, la cual nos permite de acuerdo a un nivel de confianza especificado obtener una información más precisa sobre el parámetro a estimar.

1. Intervalo de confianza para medias con n › 30 (grandes muestras):

$\mu \in \left(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$ es una estimación por intervalo de la media de la población para un nivel de confianza del (1-α)%; por ejemplo, si se define un nivel de confianza del 95 %, esto significa que por cada 100 muestras de tamaño n › 30 en 95 de ellas la media de la población cae dentro de este intervalo.

INFERENCIA ESTADISTICA

Intervalo de confianza para medias con n < 30 (pequeñas muestras):

Se utiliza la t de Student para estos casos y cuando se desconoce la desviación de la población, utilizando la siguiente expresión:

$$\mu \in \left(\bar{x} - t_{\alpha/2} \frac{S}{\sqrt{n}}, \bar{x} + t_{\alpha/2} \frac{S}{\sqrt{n}} \right)$$

Es una estimación por intervalo de la media de la población para un nivel de confianza del $(1-\alpha)\%$.

Intervalo de confianza para diferencias de medias ($\mu_1 - \mu_2$):

a) si n >30 (grandes muestras) se usa la distribución normal:

$$(\mu_1 - \mu_2) \in \left((\bar{X}_1 - \bar{X}_2) - Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \ , \ (\bar{X}_1 - \bar{X}_2) + Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)$$

b) si n < 30 (pequeñas muestras) se usa la t de Student:

$$(\mu_1 - \mu_2) \in \left((\bar{X}_1 - \bar{X}_2) - t_{\frac{\alpha}{2}} \cdot \sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}} \ , \ (\bar{X}_1 - \bar{X}_2) + t_{\frac{\alpha}{2}} \cdot \sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}} \right)$$

para un nivel de confianza del $(1-\alpha)\%$ y $v = n_1 + n_2 - 2$ g.l.

donde $S_p^{\ 2} = \dfrac{(n_1 - 1)S_1^{\ 2} + (n_2 - 1)S_2^{\ 2}}{n_1 + n_2 - 2}$

Intervalo de confianza para proporciones ($\hat{p}$):

a) grandes muestras:

$$p \in \left(\hat{p} - Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p.q}{n}} \ , \ \hat{p} + Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p.q}{n}} \right)$$

b) pequeñas muestras:

$$p \in \left(\hat{p} - t_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p.q}{n}} \, , \, \hat{p} + t_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p.q}{n}} \right)$$

Intervalo de confianza para varianzas:

$$\sigma^2 \in \left(\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}} , \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}} \right)$$

Ejemplos:

1. Un fabricante de monitores prueba dos diseños de microcircuitos para determinar si producen un flujo de corriente equivalente:

Diseño 1 $n_1 = 18$ $\overline{X}_1 = 24.2$ $S_1^2 = 10$

Diseño 2 $n_2 = 20$ $\overline{X}_2 = 23.9$ $S_2^2 = 20$

Hallar un intervalo de confianza con un nivel de confianza del 98 % para:

a) el flujo medio del diseño 1 y diseño 2.

b) La diferencia media del flujo entre los dos diseños.

c) La variabilidad del diseño 1 y diseño 2.

Solución:

a) para un nivel de confianza del 98 % se tiene que $\alpha = 1 - 0{,}98 = 0{,}02$ **y** $\alpha/2 = 0{,}01$**.**

- diseño1**, como n = 18 < 30, entonces para v = 17 grados de libertad se tiene que** $t_{\frac{\alpha}{2}} = t_{0.01} = 2.567$ **y se emplea la fórmula**

$$\mu \in \left(\bar{x} - t_{\alpha/2} \frac{S}{\sqrt{n}}, \bar{x} + t_{\alpha/2} \frac{S}{\sqrt{n}} \right) = \left(24.2 - 2.567 . \frac{3.16}{\sqrt{18}}, \; 24.2 + 2.567 \frac{3{,}16}{\sqrt{18}} \right) = (20.38 \, , \, 26.11)$$

Esto significa, que por cada 100 muestras de tamaño n =18 en 98 de ellas la media poblacional μ cae dentro de este intervalo.

- **diseño 2**, como n = 20 < 30, para v = 19 g.l. se tiene que $t_{\frac{\alpha}{2}} = t_{0.01} = 2.539$ por lo tanto

$$\mu \in \left(\bar{x} - t_{\alpha/2} \frac{S}{\sqrt{n}}, \bar{x} + t_{\alpha/2} \frac{S}{\sqrt{n}} \right) = \left(24.2 - 2.539 . \frac{4.47}{\sqrt{20}},\ 24.2 + 2.539 \frac{4,47}{\sqrt{20}} \right) = (21.661,\ 26.739)$$

b) para la diferencia de flujo medio $(\mu_1 - \mu_2)$, como n = 18+20 – 2 = 36 > 30 , se tiene que $Z_{\frac{\alpha}{2}} = 2.33$ y se utiliza la fórmula

$$(\mu_1 - \mu_2) \in \left((\bar{X}_1 - \bar{X}_2) - Z_{\frac{\alpha}{2}} . \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\ ,\ (\bar{X}_1 - \bar{X}_2) + Z_{\frac{\alpha}{2}} . \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)$$

$$\left((24.2 - 23.9\ ) - 2.33 . \sqrt{\frac{10}{18} + \frac{20}{20}}\ ,\ (24.2 - 23.9\ ) + 2.33 . \sqrt{\frac{10}{18} + \frac{20}{20}} \right) = (0.3 - 2.90\ ,\ 0.3 + 2.90) = (-2.6\ ,\ 3.2)$$

Se concluye que por cada 100 muestras de tamaño $n_1 = 18$ y $n_2 = 20$ en 98 de ellas la diferencia de medias poblacionales $(\mu_1 - \mu_2)$ está dentro de este intervalo.

c) **diseño 1**, con v = 17 g.l. se tiene $\chi^2_{\frac{\alpha}{2}} = \chi^2_{0.01} = 33,41$ y $\chi^2_{1-\frac{\alpha}{2}} = \chi^2_{0.99} = 6,41$

usando la relación:

$$\sigma^2 \in \left(\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}},\ \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}} \right) = \left(\frac{17.10}{33,41}, \frac{17.10}{6,41} \right) = (5.09\ ,\ 26.52)$$

Por cada 100 intervalos de tamaño n = 18 en 98 de ellos la varianza poblacional σ^2 cae dentro de este intervalo.

INFERENCIA ESTADISTICA

PRUEBA DE HIPÓTESIS

Para probar una hipótesis relativa a un parámetro θ se debe proceder de la siguiente manera:

Definir la hipótesis nula H_0 : (se considera la aseveración del fabricante)

$H_0: \theta = \theta_0$

2. **Establecer la hipótesis alternativa:**

$H_a: \theta \neq \theta_0$ (prueba de dos colas o bilaterales)

$H_a: \theta > \theta_0$ (prueba de cola derecha o unilateral)

$H_a: \theta < \theta_0$ (prueba de cola izquierda o unilateral)

Nota: La hipótesis alternativa se escoge de acuerdo a cada problema en particular.

Por ejemplo: supongamos que un fabricante de bombillos asegura que su producto tiene una duración promedio de 2000 horas. Por lo que el dueño de una ferretería quiere contrastar esta aseveración. Para esto se deben definir la hipótesis nula y alternativa:

$H_0: \theta = \mu_0 = 2000$ h

$H_a: \theta < \mu_0 = 2000$ (prueba de cola izquierda)

Se selecciona esta hipótesis alternativa ya que si la duración promedio es mayor que 2000 h, entonces esta hipótesis no es antagónica con H_0, es decir, es mejor para el dueño de la ferretería.

3. Definir el nivel de significación:

Para realizar una prueba de hipótesis relativa a un parámetro, se debe fijar el nivel de confianza $(1-\alpha)$ % , de aquí definimos el nivel de significación como el valor de α. Si el nivel de confianza es del 95 %, $1-\alpha = 0{,}95$ de donde $\alpha = 0{,}05$.

INFERENCIA ESTADISTICA

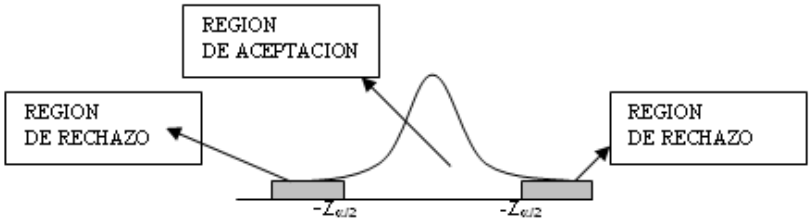
Calcular el Estadístico de Prueba:

HIPOTESIS NULA	ESTADISTICO DE PRUEBA
H_0: $\mu = \mu_0$ (para medias)	Grandes muestras Pequeñas muestras $z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$ $t = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}}$
H_0: $\mu_1 - \mu_2 = 0$ (diferencia de medias)	Grandes muestras $z = \frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ Pequeñas muestras $t = \frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}}$
H_0 : $\sigma^2 = \sigma_0^2$ (para varianzas)	$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$
H_0 : $\sigma_1^2 = \sigma_2^2$ (igualdad de varianzas)	$F = \frac{(n_M - 1)S_M^2}{(n_m - 1)S_m}$ S_M^2 : var*ianza* mayor S_m^2 : varianza menor
H_0 : $P = P_0$ (para proporciones)	$z = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}}$ $\mu_{\hat{p}} = p \quad \sigma_{\hat{p}} = \sqrt{\frac{p.q}{n}}$
H_0 : $P_1 - P_2 = 0$ (diferencia de proporciones)	

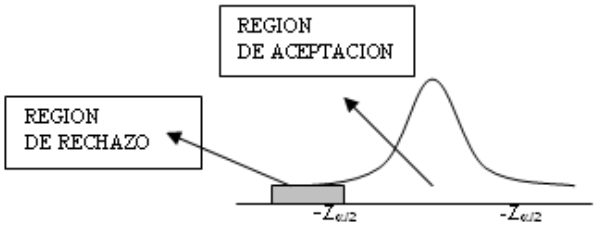
5. Establecer la región de rechazo.

Relativas a medias:

a) para la prueba de dos colas $H_a: \theta \neq \theta_0$, la región de rechazo viene dada por: estadístico $< -Z_{\alpha/2}$ y estadístico $> Z_{\alpha/2}$

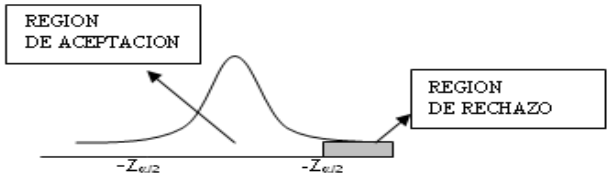


b) para la prueba de cola izquierda $H_a: \theta < \theta_0$, la región de rechazo viene dada por: estadístico $< -Z_{\alpha/2}$.



c) para la prueba de cola derecha $H_a: \theta > \theta_0$, la región de rechazo viene dada por: estadístico $> Z_{\alpha/2}$.

c) para la prueba de cola derecha $H_a: \theta > \theta_0$, la región de rechazo viene dada por: estadístico $> Z_{\alpha/2}$.



Nota: para hallar la región de rechazo para las distribuciones t de Student, χ^2 chi-cuadrado y F de fisher se procede de manera análoga al caso anterior. Es importante destacar, que usualmente los investigadores prefieren hallar el **valor p,** en vez del valor de significación, ya que este valor nos da una información más completa en cuanto a la aceptación o rechazo de la hipótesis nula.

INFERENCIA ESTADISTICA

Determinación del valor p:

Este valor se define como el intervalo formado por los valores de alfa (α) que permiten rechazar la hipótesis nula, es decir: $valor\ p = \alpha > estadístico$.

Ejemplos:

1. Se desea probar con base en 49 observaciones, que la conductividad térmica de cierto tipo de ladrillo es 0,36. La conductividad térmica de las observaciones fue de 0,365 con una desviación de 0,01. ¿Qué puede decir usted acerca de lo que se asegura?. Hallar el valor p.

Solución:

- En primer lugar se plantean las hipótesis:

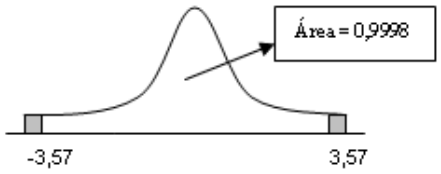
H_0: $\mu = 0,36$

H_a: $\mu \neq 0,36$ (prueba de dos colas)

- Calculo del estadístico de prueba (para grandes muestras n = 49):

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} = \frac{0,365 - 0,36}{\frac{0,01}{\sqrt{49}}} = \frac{0,005}{0,0014} = 3,57$$

- Cálculo del valor p y determinación de la región de rechazo:



ya que el área de las colas está muy cercana a cero (0,0000892649) entonces los valores de α < 0,0000892649 permiten aceptar la hipótesis nula; por lo que el valor p = α > 0,0000892649 permiten rechazar H_0. Por lo tanto, es evidente que para niveles de significación del 1% (α =0,01), 5% (α=0,05),10% (α=0,1) se rechaza H_0. En conclusión se rechaza la hipótesis nula de que la conductividad térmica del ladrillo es igual a 0,36; es decir, se acepta la alternativa de que es diferente.

2. Se analizan dos catalizadores para determinar la forma en que se afectan el rendimiento medio de un proceso químico. De manera específica el catalizador 1 es el que se está empleando en este momento, pero el catalizador 2 es más económico. Los datos de rendimiento de un catalizador se muestran a continuación:

Catalizador 1	91.5	94.18	92.18	95.39	91.79	89.07	94.72	89.21
Catalizador 2	89.19	90.95	90.46	93.21	97.19	97.04	91.07	92.75

Existe alguna diferencia entre los rendimientos medios. Hallar el valor p. Explique sus conclusiones.

Solución:

- formulación de hipótesis

H_0: $\mu_1 - \mu_2 = 0$

H_a: $\mu_1 - \mu_2 \neq 0$

- cálculo del estadístico de prueba

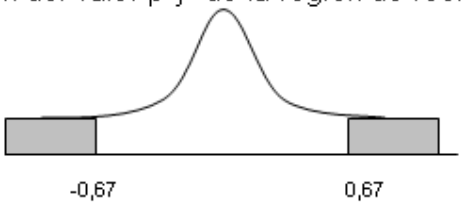
primero se calculan los estadísticos para la población 1 y 2 respectivamente: $\hat{x}_1 = 92{,}255 \quad S_1^2 = 2{,}39$ $\hat{x}_2 = 92{,}7325 \quad S_2^2 = 2{,}98$

luego $S_p^2 = \frac{7\,.\,2{,}39 + 7\,.\,2{,}98}{14} = 2{,}69$, entonces se tiene que:

$$t = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}} = \frac{-0{,}1775}{\sqrt{\frac{2{,}69}{8} + \frac{2{,}69}{8}}} = -0{,}6725$$

INFERENCIA ESTADISTICA

- determinación del valor p y de la región de rechazo



El valor p nos define los α para el cual se rechaza H_0. Considerando v = 14 g.l. y el estadístico = -0,67 como valor crítico, se tiene que el área a la izquierda de –0,67 y a la derecha de 0,67 es igual a 0,255. Por lo que, el valor p = α > 0,255, en consecuencia, la hipótesis nula se rechaza para α > 0,255. De manera particular, se tiene que para los valores usuales de α = 0,01, α=0,05 y α = 0,1. La hipótesis nula H_0 no se rechaza (se acepta). Por tanto, se concluye que para estos niveles de significación la diferencia del rendimiento medio no es estadísticamente significativa.

Tamaño de la Muestra.

La clave del problema estriba en escoger una muestra cuyo selección garantice la representatividad de la población objeto de estudio. En los estudios socio-económicos, una muestra de un 30% de la población, tiene un elevado nivel de representatividad (Ramírez 1995); sin embargo, esta representatividad depende mayormente, del tipo de muestreo. Obviamente, que el trabajar con muestras, por muy confiables que sean, no se obtiene el 100% de exactitud, sin embargo, ese pequeño error que acompaña siempre a los estudios por muestreo, es compensado con el tiempo y costo ahorrado al trabajar con grupos pequeños en vez de toda la población.

- *Determinación del Tamaño de la Muestra en una población*, cuando se utilizan proporciones:

$$n = \left(\frac{Z_{\frac{\alpha}{2}}}{\in} \right)^2 .p.q$$

Donde:

n: Tamaño de la muestra

$Z_{\alpha/2}$: Valor teórico en función del nivel de confianza, para 99 %,

$Z_{\alpha/2}$ **=** 2,56 y para el 95%, **$Z_{\alpha/2}$** =1,96

ε: error de muestreo

P: Número de veces que se produce un evento en %

Q: Es el porcentaje complementario de P

Ejemplo:

Opinión de los electores sobre gestión de gobierno.

Se realizó un estudio piloto de 150 electores donde 60 opinan favorablemente. A cuantas personas es necesario encuestar si se desea un nivel de confiabilidad de 99 % y un error de muestreo +/- 1.5%.

Entonces se tiene:

$n = \left(\frac{Z_{\frac{\alpha}{2}}}{\in} \right)^2 .p.q$ El valor de P viene dado por:

P = 60 / 150 X 100 = 40%, por lo tanto Q = 100 - 40 = 60%.

De esta forma se tiene: $n = \left(\frac{2,56}{0,015} \right)^2 .0,4.0,6 = 6.991$ **.** Es necesario encuestar a 6.991 personas para alcanzar cierta confiabilidad en los resultados.

- *En el caso de una Población Infinita con 95 % de Confiabilidad.*

Utilizando el ejemplo anterior, se tiene:

$n = \left(\frac{1,96}{0,015} \right)^2 .0,4.0,6 = 4098$

Al bajar el coeficiente o el nivel de confiabilidad, también baja el tamaño de la muestra.

- *En el caso de que no exista un Estudio Piloto.*

A los valores de P y Q se les asigna el valor de 50% a cada uno y es lo que se denomina Condiciones desfavorables de muestreo. En el caso del ejemplo citado el tamaño de la muestra viene determinado de la siguiente manera:

$$n = \left(\frac{1{,}96}{0{,}015}\right)^2 . 0{,}5 . 0{,}5 = 4.268$$

Esto quiere decir que habrá que encuestar a 4.268 personas.

En el caso de poblaciones finitas, el modelo matemático difiere con el de las poblaciones infinitas:

$$n = \frac{Z_{\alpha/2}.p.q.N}{\in^2 (n-1) + Z_{\alpha/2}.p.q}$$

Donde: **N** es el tamaño de la población y **n** el tamaño de la muestra.

Se puede aplicar en el siguiente caso: Conocer la opinión de los miembros de un sindicato, ante un nuevo contrato colectivo. Compuesto por 3.257 obreros. Cuántas obreros se deben entrevistar para obtener un nivel de confianza de 99 % y un error de muestreo de +/- 3%, en condiciones desfavorables?

$$n = \frac{2{,}56^2 . 0{,}5 . 0{,}5 . 3257}{0{,}03^2 (3257-1) + 2{,}56^2 . 0{,}5 . 0{,}5} = 1.168$$

Se requieren encuestar a 1.168 obreros, para lograr cierto grado de Confianza.

Determinación del Tamaño de la Muestra en una población para medias.

En este caso se utiliza la relación:

$$n = \left(\frac{Z_{\frac{\alpha}{2}} . \sigma}{\in} \right)^2$$

Ejemplo: Se quiere estudiar la vida útil media de una marca de neumáticos. Si sabe por estudios anteriores que la desviación estándar es de 800 Km . Determinar el tamaño de la muestra requerido para un nivel de confianza del 95 %, fijando un error de 40.

Sustituyendo los valores se tiene

$$n = \left(\frac{1{,}96 . 800}{40} \right)^2 = \left(\frac{1568}{40} \right)^2 = 1536{,}64 \approx 1537 \text{ neumáticos}$$

En conclusión, la validez en la investigaciones de negocios, está muy relacionada con la confiabilidad del muestreo y una muestra confiable está en función del tipo de población a estudiar (finitas o infinitas); asi mismo, en cuanto al nivel de confiabilidad, ésta será mayor si la muestra es mayor y en relación al error de muestreo, éste será menor cuando la muestra es mayor. Para determinar el tamaño de la muestra de una forma mas rápida y práctica, se han diseñado las Tablas de Harvard, las cuales permiten calcular, rapidamante el tamaño de la muestra a tomar, en función del error de muestreo, niveles de confiabilidad y posibles valores de P y Q.

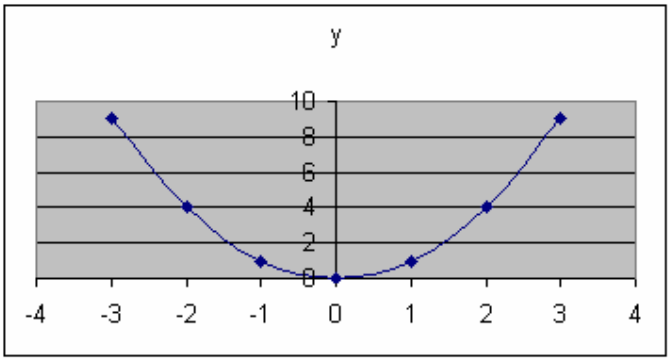
Para profundizar en este aspecto de muestreo, se recomienda consultar los textos especializados en estas áreas. Pues una vez determinado el tamaño de la muestra el paso siguiente que se plantea es lo relacionado al tipo de muestreo que se va a utilizar para escoger los elementos que integran a la muestra y ésto es un amplio e interesante tema a tratar.

INFERENCIA ESTADISTICA

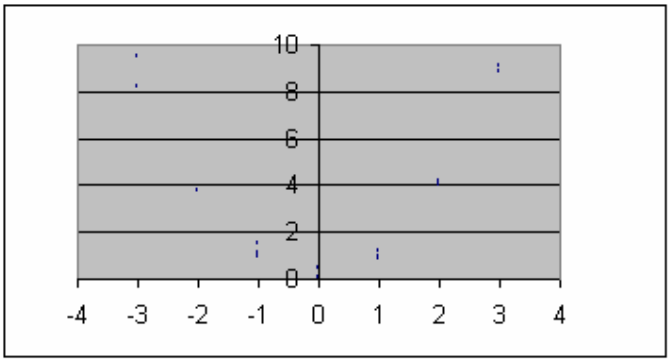
AJUSTES DE CURVAS.

Cuando se quiere estudiar la relación entre variables se puede recurrir a dos tipos de modelos:

a) **modelo determinístico**, la relación viene definida a través de una fórmula. Por ejemplo, sea $y = x^2$, entonces se dice que y está en función de x, donde y se conoce como variable dependiente y x variable independiente. La característica fundamental de este modelo es que para un valor particular de x siempre obtenemos el mismo resultado en y, esto significa que la relación entre las variables es perfecta. Ver gráfica.



b) **modelo probabilistico**, la relación entre las variables no es perfecta, ya que debido a una perturbación aleatoría (ruido) a veces para un mismo valor de la variable independiente x se obtienen valores diferentes para y. En este caso, no se obtiene una curva sino un diagrama de dispersión. Considerando el ejemplo anterior, $y = x^2 + \varepsilon$ donde ε es un ruido. Ver gráfica.



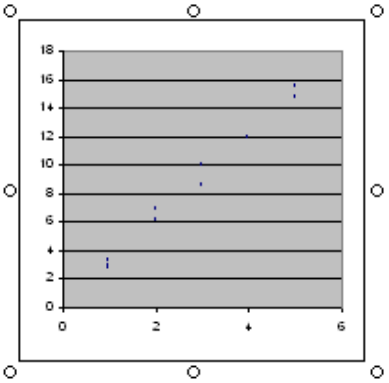
Por tanto, los modelos probabilísticos son útiles cuando se realizan investigaciones del tipo experimental donde a pesar de mantener fijo los valores de la variable independiente ocurren fluctuaciones debido fundamentalmente a errores de medición, de los equipos, etc. En el presente trabajo estamos interesados en este tipo de modelos. A continuación mencionamos los modelos de ajustes más usados:

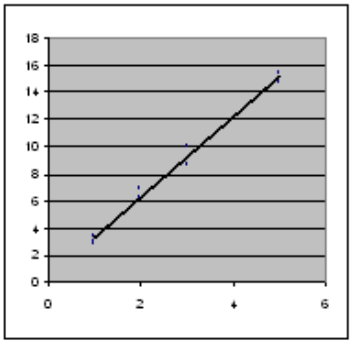
INFERENCIA ESTADISTICA

Regresión simple: Se define como la curva que optimiza (minimiza), mediante el método de los mínimos cuadrados, los saltos o fluctuaciones de los datos. Es decir, es la curva que mejor ajusta los valores del diagrama de dispersión convirtiendo el modelo probabilístico en un modelo determinístico con la finalidad de realizar predicciones. De igual forma, la curva de regresión permite modelar la tendencia de los valores. Los modelos de regresión simple vienen definidos por y = f(x)+ε. A continuación veamos los distintos modelos con su respectivo ajuste o curva de regresión:

Modelos Probabilísticos **Curva de Regresión**

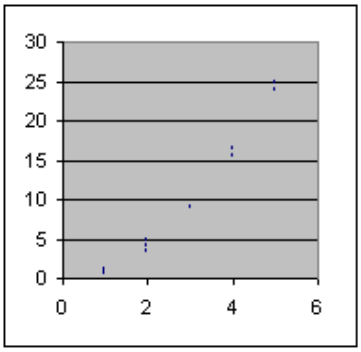
a) Lineal: $y = ax + b + \epsilon$ $\hat{y} = \hat{a}x + \hat{b}$ (línea recta)

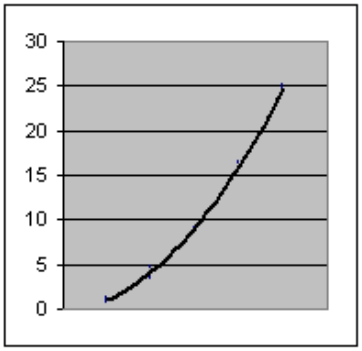




b) Polinómico:

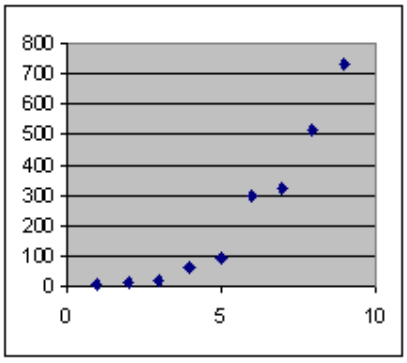
orden dos: $y = ax^2 + bx + c + \epsilon$ $\hat{y} = \hat{a}x^2 + \hat{b}x + \hat{c}$ (parábola)

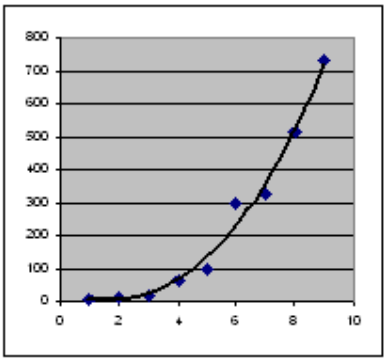




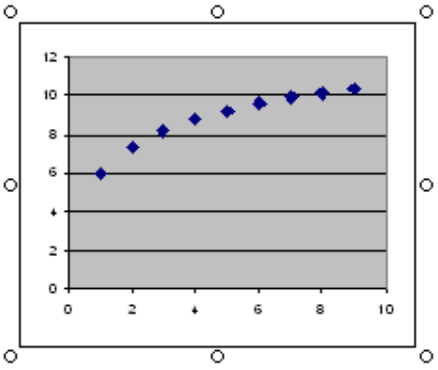
INFERENCIA ESTADISTICA

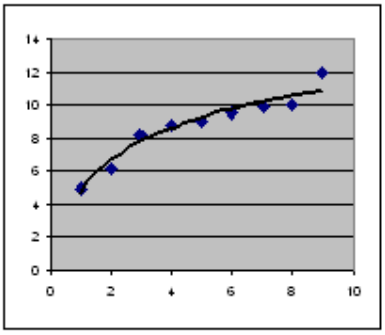
orden tres: $y = ax^3 + bx^2 + cx + d + \epsilon$ $\hat{y} = \hat{a}x^3 + \hat{b}x^2 + \hat{c}x + \hat{d}$



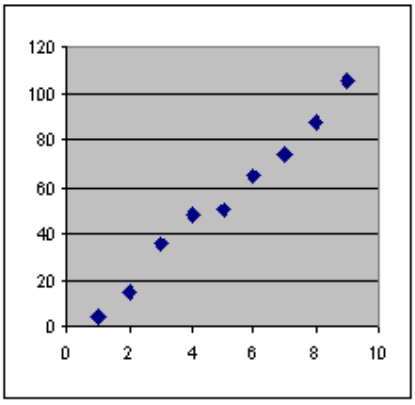


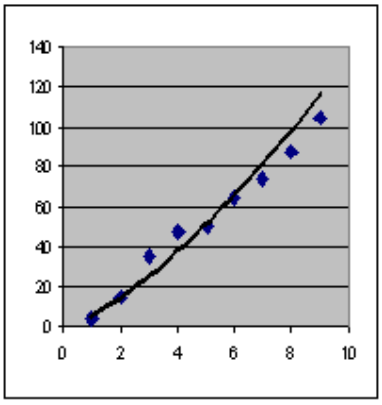
c) Logarítmico: $y = aLn(x) + b + \epsilon$ $\hat{y} = \hat{a}Ln(x) + \hat{b}$



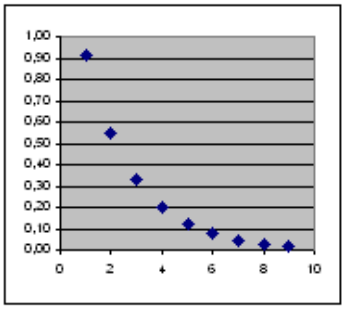


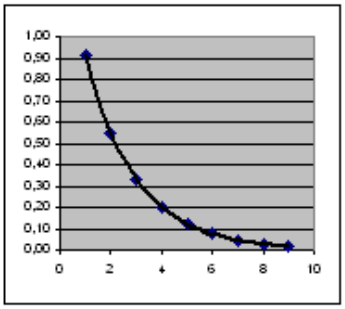
d) Potencial: $y = ax^b + \epsilon$ $\hat{y} = \hat{a}x^{\hat{b}}$





e) Exponencial: $y = ae^{bx} + \in$ $\hat{y} = \hat{a}e^{\hat{b}x}$





Nota: El procedimiento en un análisis de regresión consiste en calcular los estimadores ($\hat{a}, \hat{b}, \hat{c}$ y $\hat{d}$) que definen la curva que mejor ajusta los datos.

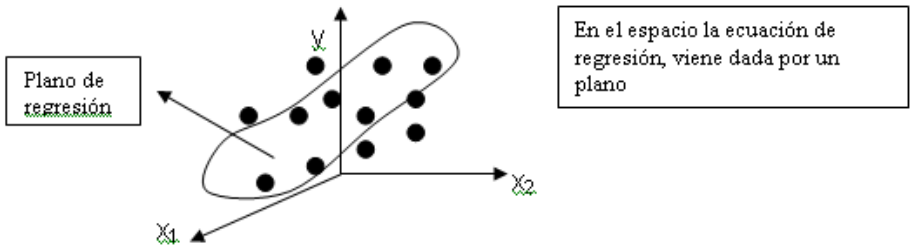
En la actualidad, existen muchos paquetes estadísticos que permiten calcular los estimadores y la curva de regresión sin necesidad de realizar los engorrosos cálculos de manera manual. De manera particular, se mostrará los ajustes de curvas mediante ejemplos y utilizando Excell.

Regresión múltiple lineal.

Modelo probabilístico: $y = a_1x_1 + a_2x_2 + \ldots + a_nx_n + b + \in$

Modelo determinístico: $\hat{y} = \hat{a}_1x_1 + \hat{a}_2x_2 + \ldots + \hat{a}_nx_n$

Nota: todo modelo de regresión simple puede representarse en el plano cartessiano (bidimensional) ya que se requiere de un eje para representar la variable independiente (x) y otro para las observaciones (y). Para el caso de regresión múltiple solamente hay una representación espacial, cuando se tienen dos ejes para las variables independientes (x_1 y x_2) y otro para las observaciones (y), es decir, modelos de la forma y = f(x_1, x_2). Ver gráfica.



Para modelos donde el número de variables independientes es igual o mayor que 3, es imposible realizar una representación gráfica, No obstante la ecuación de regresión se le llama hiperplano.