ESTADÍSTICA
MÓDULO 1

Lic. OSVALDO JORGE CASTILLO

2009

E-mail: castillo@cqc.com.ar Internet: http://www.cqc.com.ar

INDICE

1. ESTADÍSTICA:

Definición: La ciencia que se ocupa de la recopilación, tabulación, análisis, interpretación y presentación de datos.

Clases

- Descriptiva o deductiva : Se ocupa de la descripción y el análisis de un tema o grupo de datos.
- Inferencial o inductiva: A partir de una determinada cantidad de datos (muestra), se obtiene una conclusión acerca de una mayor cantidad de datos (población). Dado que no es posible establecer tales conclusiones o inferencias con total certeza, se utilizan términos del lenguaje de la probabilidad

Recopilación de datos: Se realiza mediante la observación directa o indirecta, a través de preguntas hechas por escrito o verbalmente. A esta recopilación de datos se la clasifica como variables que pueden ser numéricas o por atributos.

Las variables numéricas son medibles, las de atributos no.

Si la variable numérica puede ser dividida infinitamente se la denomina como ***contínua***, las que se producen saltos o vacios se las conocen como ***discretas***. En general las variables contínuas son medibles, las discretas son el resultado de un conteo.

Muchas veces es necesario expersar datos no numéricos como variables numéricas, ejemplo el nivel de calificación A, B, C, D y F se le hace corresponder a valores numéricos 4,3,2,1 y 0, respectivamente; estos valores numéricos discretos se usarán como variables discretas al efectuar cálculos.

Por lo general aquellas características que se evaluan a través de una observación visual se les clasifica como atributos. Esto se da claramente cuando se trata de determinar si se cumple o no alguna especificación.

Otro tema importante a tener en cuenta para el caso de las variables numéricas es la exactitud, ya que por lo general cuanto más cifras haya a la derecha del punto decimal, más complejo deberá ser el instrumento de medición utilizado. Existe la posibilidad que los instrumentos de medición no tengan una lectura fidedigna debido a problemas relacionados con la exactitud y la precisión.

<u>TRABAJO PRÁCTICO Nº 1</u>

1 Según su especialidad, defina algunas variables por atributos.

2. Idem a la anterior pero discretas.

3. Idem a la anterior pero contínuas.

2.- ESTADÍSTICA DESCRIPTIVA O DEDUCTIVA:

Es la rama de la estadística que se utiliza para describir hechos. Consiste en organizar, resumir y simplificar, en términos generales, información que a menudo es bastante compleja. El objeto es hacer que las cosas se comprendan más fácilmente, que sea más sencillo referirse a ellas y analizarlas.

2.1. MEDIDAS DE TENDENCIA CENTRAL:

Estas medidas se utilizan para indicar un valor que tiende a tipificar o a ser el más representativo de un conjunto de números. Las tres medidas que más comúnmente se emplean son la media, la mediana y el modo.

Media

La media aritmética es lo que habitualmente se conoce como "promedio". Se obtiene al sumar los valores de un conjunto y al dividir el producto de esta suma entre el número de valores del mismo.

Ej.

$$\frac{70 + 80 + 120}{3} = 90$$

La media se representa con $\overline{X}$ (se lee X raya), y su cálculo se puede expresar como se observa a continuación:

$$\boxed{\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}}$$

Mediana

La segunda medida de tendencia central de un conjunto de números es la mediana. Su característica principal es que divide un conjunto ordenado en dos grupos iguales; la mitad de los números tendrá valores que son menores que la mediana, y la otra mitad alcanzará valores mayores que esta. Para encontrar la mediana primeramente es necesario ordenar los valores (generalmente de menor a mayor). Posteriormente se deberá separar la mitad de los valores para obtener la mediana.

Por ejemplo, la mediana del grupo 5, 6 y 8 es 6, en el cual 6 está en medio. En términos generales, la mediana ocupa la posición (n + 1) / 2. Por tanto, para tres números, la posición es (3 + 1)/ 2 = 2 o sea, la segunda posición. Otro ejemplo: obtener la mediana de 7, 8, 9 y 10. Según la fórmula , la posición de la mediana es (4 + 1) / 2 = 2,5 que se encuentra a la mitad de los dos valores intermedios, o sea 8,5 en este caso. Esto deja dos valores menores y dos mayores.

El procedimiento para obtener la mediana es como sigue:

1. Ordenar o clasificar los valores
2. Contar para saber si el número de valores es par o impar
3. En caso de que tenga un número impar de valores, la mediana es el valor intermedio. En cambio, para un número par de valores, la mediana es el promedio de los dos valores intermedios.

Ej.

Par	Mediana
2, 3 ,3 ,4	3
1, 18, 19, 20	18.5
5.1, 6.5, 8.1, 9.1, 10.1, 15.5	8.6

Impar	Mediana
1, 2, 3, 3, 3, 4, 7	3
9, 40, 80, 81, 100	80
3.7, 9.2, 10.1, 11.8, 12.8	10.1

En síntesis

La mediana de un conjunto de números es mayor que la mitad de los valores y menor que la otra mitad de los mismos.

Comparación entre la media y la mediana

Elegir el uso de la media o la mediana como medidas de tendencia central de un conjunto de números depende de varios factores. La media se ve afectada o es influida por todo valor del conjunto, incluyendo los extremos. La mediana, por otra parte, es insensible a valores extremos. Ej. la mediana describe en forma más adecuada datos sobre ingresos o gastos domésticos, ya que unos pocos valores muy grandes tienden a *inflar* la media aritmética.

Modo

El modo es el valor que con más frecuencia se presenta en un conjunto. Por ejemplo en el conjunto 10, 10, 8, 6 y 10, el 10 se presenta tres veces, en tanto que cada uno de los otros valores sólo una vez. El valor más frecuente, el modo es 10.

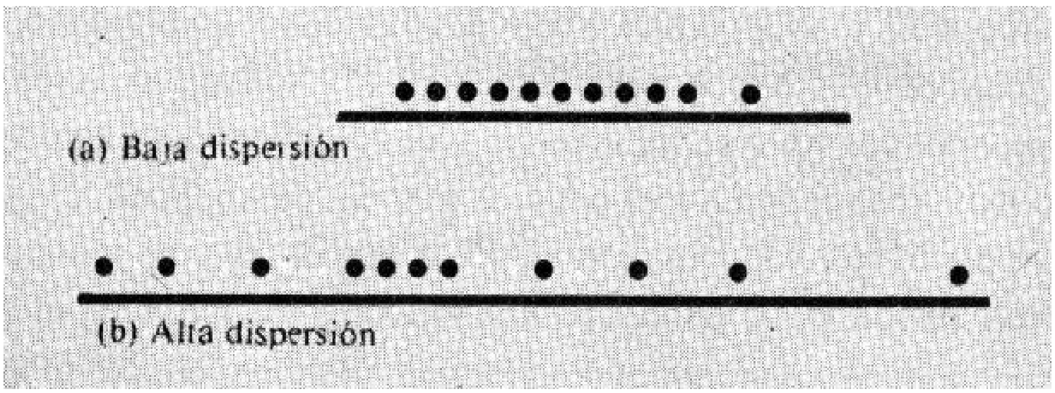
En comparación con la media y la mediana el modo es el menos útil para la mayoría de los problemas estadísticos, ya que no se inclina por un análisis matemático en el mismo sentido que lo hacen la media y la mediana. Sin embargo desde un punto de vista meramente descriptivo, el modo es indicativa del valor típico. El modo es útil cuando uno o dos valores, o un grupo de éstos, ocurren con mayor frecuencia que otros. Por el contrario, cuando la mayoría o todos los valores se presentan casi con la misma frrecuencia, el modo no sirve para describir datos.

En síntesis:
El modo es el valor que ocurre con mayor frecuencia.

2.3. MEDIDAS DE DISPERSIÓN

Para describir en forma adecuada un conjunto de datos, son necesarios dos tipos de medidas resumen. Además, para obtener información respecto a la parte media de un conjunto de números, es conveniente también tener un método para expresar la cantidad de dispersión que hay entre los mismos. Las medidas de dispersión indican si los valores están relativamente cercanos uno del otro o si se encuentran dispersos.

Esquemáticamente:



Es conveniente considerar cuatro variables de dispersión: la amplitud de variación, la desviación media, la varianza y la desviación estándar. Todas estas medidas, excepto la amplitud de variación, toman a la media como punto de referencia.

Amplitud de variación

Es generalmente la medida más sencilla de calcular y comprender. Se concentra en el número mayor y el menor del grupo, es decir los puntos extremos.

Se puede expresar de dos formas:

- La diferencia entre los valores mayor y menor
- Los valores mayor y menor del grupo

Ej.

	Amplitud
1, 10 y 25	25 - 1 = 24

de 1 a 25 (proporciona más información que la anterior)

Desviación absoluta media (DAM)

Mide la desviación promedio de valores con respecto a la media del grupo, sin tomar en cuenta el signo d ella desviación. Se obtiene al restar la media de cada valor del grupo, eliminando el signo (+ o -) de la desviación, hallando después el promedio

$$\boxed{\mathrm{DAM} = \frac{\sum|x_i - \bar{x}|}{n}}$$

Ej.

1º, calculamos la media

2, 4, 6, 8, 10

$$\bar{X} = \frac{2 + 4 + 6 + 8 + 10}{5} = 6$$

2º, calculamos las diferencias entre la media y cada valor

2 - 6 = -4
4 - 6 = -2
6 - 6 = 0
8 - 6 = 2
10-6 = 4

0 (comprobación)

3º, convertimos las diferencias a valores absolutos y sumamos

4 + 2 + 0 + 2 + 4 = 12

4º, hallamos la desviación media

12 / 5 = 2.4

Varianza

La varianza de una muestra se calcula casi en la misma forma que la desviación media, con dos pequeñas diferencias: 1) las desviaciones se elevan al cuadrado antes de ser sumadas y, 2) se obtiene el promedio, utilizando n -1 en lugar de n.

La varianza muestral se puede calcular mediante la fórmula siguiente:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

Si el conjunto de números constituye una población, o bien, si el objeto de resumir los datos es únicamente para describir un conjunto d edatos en lugar de sacar inferencias respecto a una población, entonces se deberá sustituir en el denominador, (n - 1) por n.

Ej. Calcular la varianza de la muestra 2, 4, 6, 8, 10.

X_i	$\bar{X}$	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$
2	6	-4	16
4	6	-2	4
6	6	0	0
8	6	2	4
10	6	4	16
		2	40

$$S^2 = \frac{40}{5 - 1} = 10.0$$

Si tales valores hubieran sido todos los valores de una población, su varianza sería 40/5 = 8

Desvío estándar

El desvío estándar es simplemente la raiz cuadrada positiva de la varianza. De este modo si la varianza es 8.1, la desviación estándar es 9. Para obtener la desviación estándar se puede utilizar la siguiente fórmula:

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$$

El desvío estándar es una de las medidas d eresumen que más se utiliza y desempeña un papel muy importante en la estadística. Es importante observar que las unidades de la desviación estándar son las mismas que las de la media. Por ejemplo, si la media está en unidades monetarias, la desviación estándar también lo estará. Si la media es en metros, lo mismo ocurrirá con la desviación estándar, la varianza se expresa en unidades al cuadrado.

En síntesis:

El desvío estándar de un conjunto de números se define como la raiz cuadrada positiva de la varianza.

TRABAJO PRÁCTICO Nº 2

3 Aplicando los conceptos de media, mediana, modo, amplitud y desvío estándar a las variables que definió en el trabajo práctico nº 1 le ayudarían a mejorar su información, piense en cada variable, ¿qué cree que le resuelve cada medida?.

a. Variable:

la media …

la mediana …

el modo…

la amplitud…

el desvío estándar

b. Variable:

la media …

la mediana …

el modo…

la amplitud…

el desvío estándar

c. Variable:

la media …

la mediana …

el modo…

la amplitud…

el desvío estándar

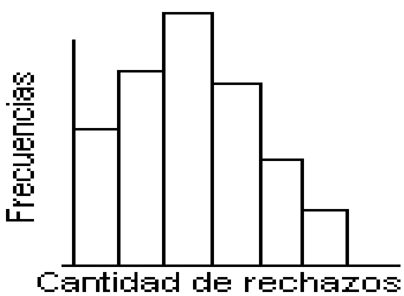
3.0 ANALISIS DE GRANDES CONJUNTOS DE DATOS

Los métodos principales para organizar datos estadísticos es mediante el uso de medidas numéricas como la media, mediana modo, desvío, etc, o en tablas también conocidas como tablas de frecuencias o distribuciones de frecuencias.

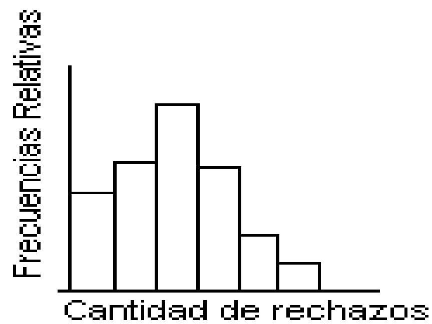
3.1 TABLAS DE FRECUENCIAS:

Las tablas de frecuencias o histogramas están formados por un conjunto de rectángulos que representan la frecuencia de cada categoría, siendo estas las fecuencias de valores observados, las frecuencias relativas de valores observados, las fecuencias acumuladas de valores observados y las fecuencias relativas acumuladas de valores observados.

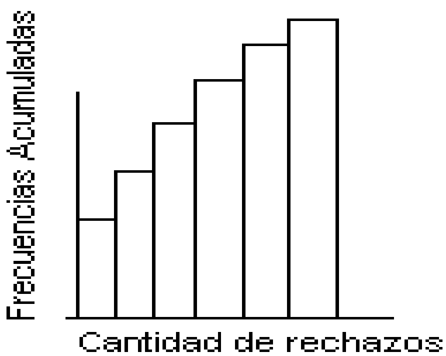
Gráficamente.



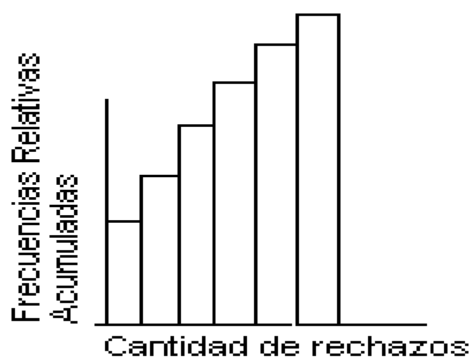
Histograma de Frecuencias



Histograma de Frecuencias Relativas



Histograma de Frecuencias Acumuladas



Histograma de Frecuencias Relativas Acumuladas

Ej. Diversos tipos de distribuciones de frecuencia de los datos

CANTIDAD DE RECHAZOS	FRE CUEN CIA	FRECUENCIA RELATIVA	FRECUENCIA ACUMULADA	FRECUENCIA ACUMULADA RELATIVA
0	9	9/35 = 0,26	9	9/35 = 0,26
1	13	13/35 = 0,37	9 + 13 = 22	22/35 = 0,63
2	5	5/35 = 0,14	22 + 5 = 27	27/35 = 0,77
3	4	4/35 = 0,11	27 + 4 = 31	31/35 = 0,89
4	3	3/35 = 0,09	31 + 3 = 34	34/35 = 0,97
5	1	1/35 = 0,03	34 + 1 = 35	35/35 = 1,00
TOTAL	35	1,00		

Distribución de frecuencias relativas

El término relativo indica la representación de proporciones o fracciones del total. La frecuencia relativa se calcula dividiendo la frecuencia de cada uno de los valores de los datos (en este caso, la cantidad de rechazos) entre el total, que viene a ser la suma de las frecuencias. Estos cálculos corresponden a la tercera columna de la tabla. La frecuencia relativa tiene la ventaja de servir como referencia. Por ejemplo, la proporción correspondiente a nueve unidades rechazadas es 0,26.

Distribución de frecuencias acumuladas

Se calcula sumando a la frecuencia de cada uno de los valores de los datos la suma de las frercuencias de los valores anteriores de los datos. Como puede observarse en la cuarta columna de la tabla, la frercuencia acumulada de 0 unidades rechazadas es 9; para una sola unidad rechazada es 9 + 13 = 22; para 2 unidades rechazadas, 22 + 5, etc. La fecuencia acumulada es la cantidad de puntos de datos igual o inferior al valor de los datos. Por ejemplo, la cantidad de lotes que tienen dos o menos unidades rechazadas es 27.

Distribución de frecuencias relativas acumuladas

Se calcula dividiendo la fecuencia acumulada de cada uno de los valores de los datos entre el total. Este cálculo aparece en la quinta columna de la tabla. La gráfica indica que la tasa de errores que tienen 2 o menos unidades rechazadas es de 0,77 o 77 %.

Este ejemplo se limitó a variables discretas, cuando se trabaja con valores contínuos será necesario agruparlos.

Datos agrupados

Los pasos principales en la elaboración de tablas de datos agrupados es la siguiente:

1. Establecer las clases o intervalos en los que se agruparán los datos.
2. Ordenarlos en clases mediante conteo por marcas
3. Contar el número de cada clase
4. Presentar los resultados en una tabla o gráfica

Ej. Datos contínuos no agrupados.

11.1	12.5	32.4	7.8	21.0	16.4	11.2	22.3
4.4	6.1	27.5	32.8	18.5	16.4	15.1	6.0
10.7	15.8	25.0	18.2	12.2	12.6	4.7	23.5
14.8	22.6	16.0	19.1	7.4	9.2	10.0	26.2
3.5	16.2	14.5	3.2	8.1	12.9	19.1	13.7

1. Establecer la amplitud de variación de los datos. El mayor es 32.8 y el menor es 3.2, por lo tanto la magnitud es 29.6.

2. Decidir el número de intervalos que se va a emplear. Se recomienda entre 5 y 15

3. Dividir la amplitud de variación entre el número de intervalos, para obtener la amplitud del mismo: 29.6 / 6 = 4.93 se elige 5.
4. Calcular los límites del intervalo comenzando por el menor. Se recomienda tomar la parte entera del menor valor (en el ejemplo 3.2, parte entera 3)

   Límite inferior: 3 límite superior: 3 + amplitud = 3 + 5 = 8

   El segundo intervalo es:

   Límite inferior: Límite sup.del intervalo anterior límite superior: 8 + 5 = 13

   Y así sucesivamente

   Es importante que no hayan vacíos en algunos intervalos.

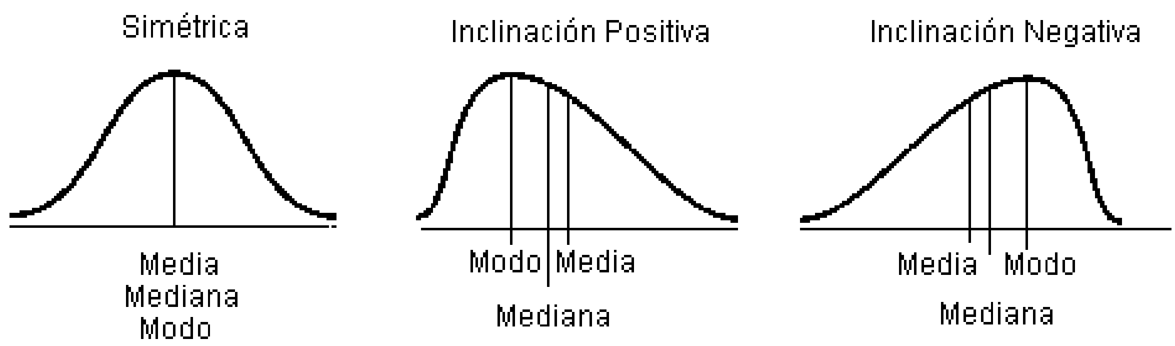
5. Anotación de las frecuencias por intervalos. Se deberá colocar o asignar al intervalo adecuado mediante el conteo por marcas. Por ejemplo, el primer valor es 11.1, que queda en el segundo intervalo.

Intervalo	Marcas	Conteo
3 a 8	~~IIII~~ III	8
8 a 13	~~IIII~~ ~~IIII~~	10
13 a 18	~~IIII~~ IIII	9
18 a 23	~~IIII~~ II	7
23 a 28	IIII	4
28 a 33	II	2

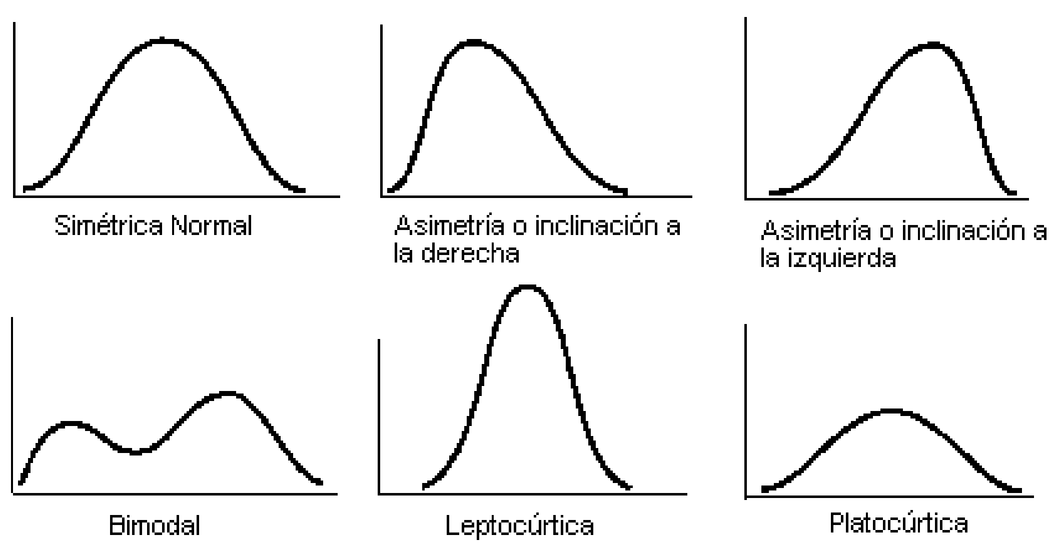
6. Se representa en forma de tabla o gráficamente como se indico para variables discretas.

3.2. ANÁLISIS DE LOS HISTOGRAMAS

Las distribuciones de frrecuencia se utilizan para describir y analizar grandes conjuntos de datos, lo cual se puede hacer mediante gráficas o tablas. Utilizando líneas contínuas en vez de las formas rectangulares características de los histogramas tendremos las curvas de distribución de frecuencias.



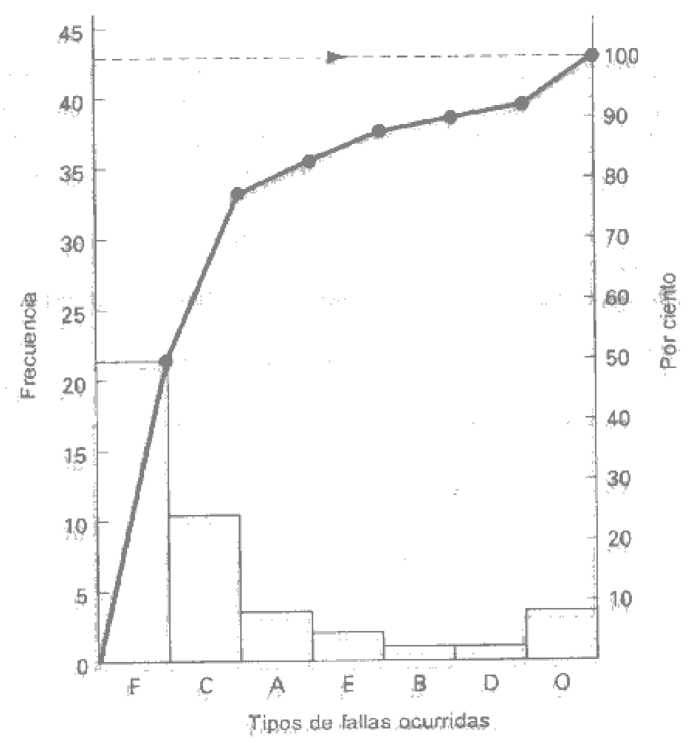
Resulta útil saber si la distribución es simétrica (el lado izquierdo es la imagen del lado derecho) o si la distribución es sesgada u oblicua (orientada) en cierta dirección.



Otra característica es la que se refiere a la cantidad de modos o picos de los datos. Puede haber un modo, dos modos (bimodal) o muchos. Otra característica se refiere a la agudeza de la distribución d ellos datos. Si la forma de la curva es excesivamente aguda se le conoce como leptocúrtica; si está bien aplanada se le denomina platocúrtica. Las gráficas nos permiten tomar decisiones sin necesidad de hacer mayor análisis.

Diagrama de Pareto

Otro tipo de distribución semejante al histograma es el diagrama de Pareto. Este diagrama es una técnica muy eficiente para descubrir dónde residen los principales factores o categorías de problemas vinculados generalmente con la calidad, en la absisa se usan categorías en vez de valores de datos. La otra diferencia es que las categorías o factores se disponen por orden descendente, empezando por la mayor frecuencia relativa, y no por orden numérico.



La gráfica nos permite identificar rápidamente las minorías vitales. Ejemplo.

La minoría de clientes que representan la mayoría de las ventas.

La minoría de productos, procesos, o características de la calidad causantes del grueso de desperdicio o de los costos de reelaboración.
La minoría de rechazos que representa la mayoría de quejas de la clientela.
La minoría de vendedores que está vinculada a la mayoría de partes rechazadas.
La minoría de problemas causante del grueso del retraso de un proceso.
Etc.

TRABAJO PRÁCTICO Nº 3

4 De las variables que definió en el trabajo práctico nº 2, ¿puede imaginar como serían las representaciones gráficas de las tablas de frecuencia de cada variable?. Dibújelas como curvas indicando si son simétricas, asimétricas, bimodales, platocúrticas, leptocúrticas, dibuje donde estaría la media, la mediana y el modo, según el gráfico que haya imaginado y dibujado.

a. Variable :

b. Variable :

c. Variable :

AUTOEVALUACIÓN MÓDULO 1

Nombre y apellido: ....

Curso:

Fecha: / /

PROBLEMA 1

Clasifique las siguientes variables en discretas o contínuas.

	Disc.	Cont.
a. Número de accidentes ocurridos en el año en una Empresa	□	□
b. Producción mensual de heladeras	□	□
c. Sus gastos semanales de supermercado	□	□
d. Cantidad de defectos encontrados en un elemento producido por una máquina	□	□
e. Ingresos mensuales de los empleados	□	□

PROBLEMA 2

a. Calcule la media, mediana y modo de la siguiente serie

2, 3 ,3 ,4, 4, 4, 5, 5, 4, 66

Media	Mediana	Modo

b. Qué medida de tendencia central es más representativa de las tres y por qué?

....

PROBLEMA 3

A su criterio, ¿qué población tiene mayor dispersión?.

a. Desvío estándar 10 cm y media 100 cm
b. Desvío estándar 0,1 mm y media 0,5 mm
c. Varianza 1 mm^2 y media 4 mm

Resp. a. □ b. □ c. □

PROBLEMA 4

En el ejemplo de datos agrupados (utilice el conteo por marcas), ¿si esos datos representarían días de desfasaje en el cobro de facturas, que agrupamiento de datos le interesaría analizar con más detenimiento

a. Si Usted fuese auditor? Intervalo □
Por qué? ....

b. Si fuese comprador? Intervalo □
Por qué? ....

REFERENCIAS BIBLIOGRAFICAS

- **Estadística para Administración y Economía**
  Stevenson, Editorial Harla
- **Muestreo**
  Morris J. Sloni M., Editorial Americana
- **Análisis y Planeación de la Calidad**
  Juran Gryna, Editorial Mc Graw Hill
- **Análisis Estadístico**
  Ya-Lun Chou, Editorial Mc Graw Hill
- **Control de Calidad**
  Besterfield, Editorial Prentic Hall
- **Introducción a la Inferencia Estadística**
  Castello Minolli, Editorial El Coloquio

ESTADÍSTICA APLICADA MÓDULO 2

Lic. OSVALDO CASTILLO

2009

E-mail: castillo@cqc.com.ar Internet: http://www.cqc.com.ar

INDICE

4.0 ESTADÍSTICA INFERENCIAL:

NO HACE FALTA TOMARSE TODO EL VINO PARA SABER SI ES BUENO !!!

Inferencia estadística es el procesamiento de ***datos muestrales*** para estimar las condiciones poblacionales que por diversas causas se hace imposible de controlar en su totalidad. El muestreo es medir una porción pequeña, pero típica, de alguna población, y posteriormente utilizar dicha información para inferir o conjurar inteligentemente, qué característica tiene la población total. Existen muchos ejemplos cotidianos de procesos de inferencia intuitivos, como por ejemplo probar el vino, hojear un libro, hacer zapping, etc. Los que nos interesan a los efectos del estudio son aquellas infinitas aplicaciones en la economía y la industria y que pueden ser resueltos aplicando modelos estadísticos:

.Para controlar la cantidad de materia prima entregada por el proveedor;
.Para controlar la calidad de un producto;
.Para verificar el buen funcionamiento de un proceso productivo;
.Para realizar estudios de mercado;
.Etc..

Se realizan muestreos por muchas razones. El costo suele ser el principal factor. A medida que aumenta la cantidad de datos reunidos aumenta el costo. El muestreo reduce costos. Otra razón para realizar muestreo es la reducción del tiempo. La estimación es el proceso mediante el cual, utilizando los datos muestrales, arribamos aproximadamente a conocer el verdadero valor del parámetro poblacional desconocido. Esencialmente, cualquier característica de la población se puede estimar a partir de una muestra al azar.

TRABAJO PRÁCTICO Nº 4

A qué problema concreto piensa que podría aplicar inferencia estadística, utilizando para ello las variables que definió en el trabajo práctico nº 1.

1.

2.

3.

4.

5.

6.

7.

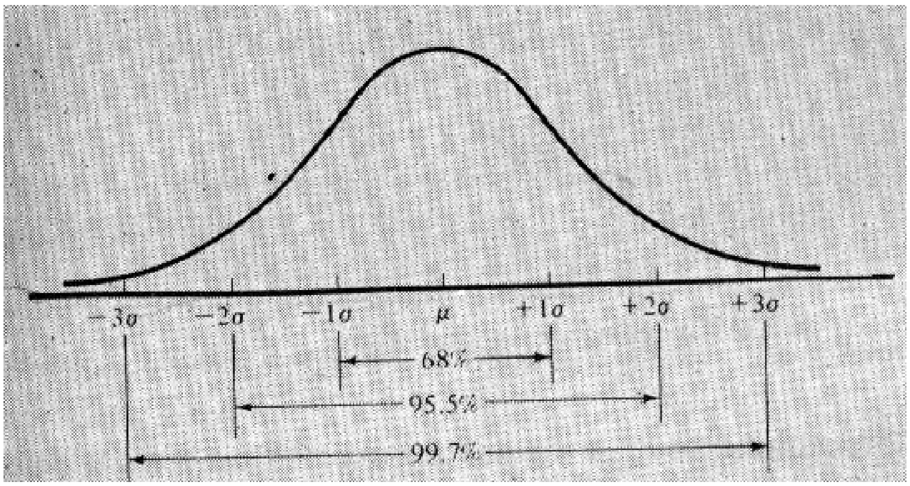
8.

9.

10.

5. La curva normal

La curva normal es una distribución simétrica, unimodal, en forma de campana, y en la que la media, mediana y modo tienen el mismo valor.

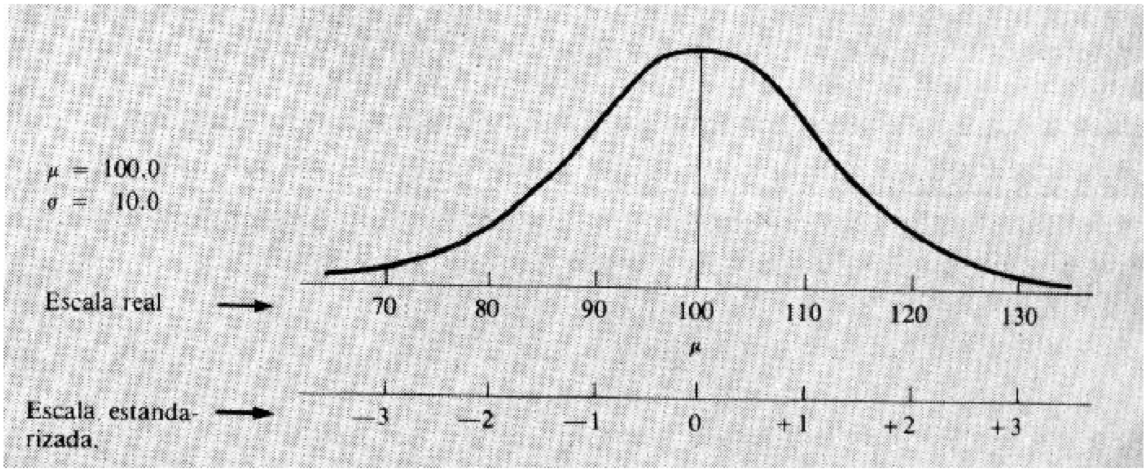


Todas las distribuciones normales se pueden convertir a la distribución normal estandarizada, utilizando el valor normal estandarizado Z.

$$Z = \frac{X_i - \mu}{\sigma} = \frac{110 - 100}{10} = +1$$

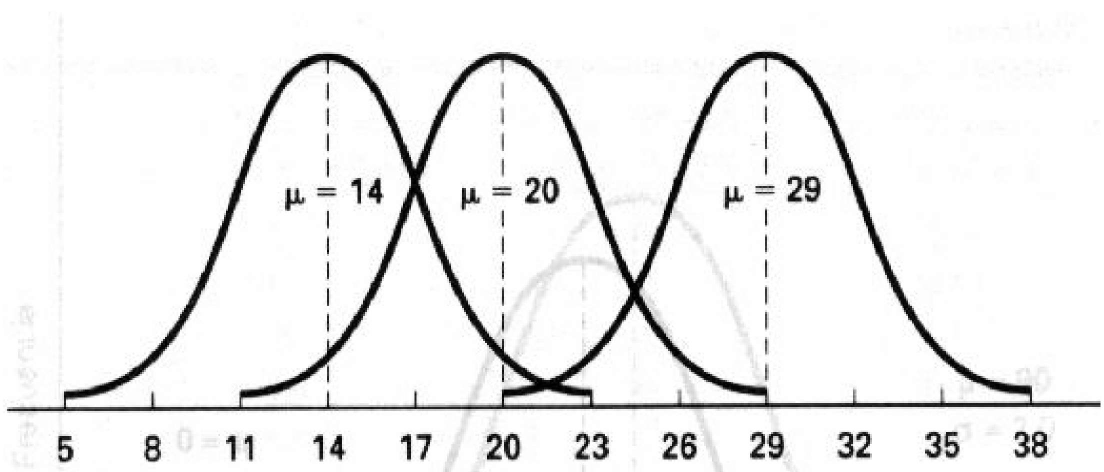
Donde	Z	= valor normal estándar
	Xi	= valor individual
	μ	= media
	σ	= desvío estándar de la población

Ej. Curva normal estandarizada



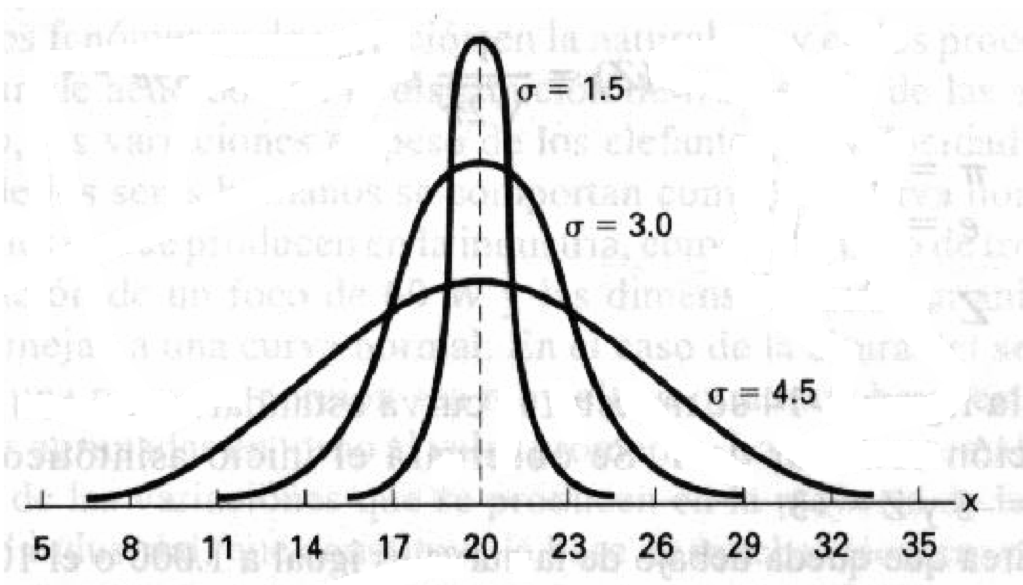
Existe una relación bien definida entre la media, y el desvío estándar en la curva normal. En la figura se muestran tres curvas normales con valores de media distintos; como se podrá ver, en lo único que difieren es en su ubicación.

Ej. Distribuciones con distinta media e igual desvío



En la siguiente figura se muestran tres curvas normales cuya media es la misma, pero tienen distintos devíos estándares. Aquí podemos ver como cuanto mayor sea el desvío estándar más aplastada es la curva (los datos están muy dispersos) y cuando menor sea el desvío estándar, más puntiaguda es la curva (los datos se dispersan en un área muy estrecha). Si el desvío es cero, todos los valores serán idénticos a la media y no habrá curva.

Ej. Distribuciones con misma media y distinto desvío



Aplicaciones

Para definir el porcentaje de elementos que están comprendidos entre dos valores determinados se puede recurrir al siguiente procedimiento:

1.- calcular el valor de Z utilizando la fómula:

$$Z = \frac{X_i - \mu}{\sigma}$$

2.- con el valor obtenido de Z, se encontrará en la tabla el valor del área que está debajo de la curva, a la izquierda de Xi. Por ejemplo, si el valor de Z es de -1.76, el correspondiente valor del área es de 0.0392. Dado que el área total bajo la curva es de 1.0000, el valor de 0.0392 del área se convierte al porcentaje de elementos que están bajo la curva tan solo desplazando el punto decimal dos lugares a la derecha. Por lo tanto, 3.92 % de los elementos tienen valor inferiora valor particular de Xi.

Entonces, suponiendo que los datos están distribuídos normalmente, se puede determinar el porcentaje de elementos de los datos que tienen un valor inferior a un valor determinado, mayor que un valor dado, y comprendido entre dos valores.

TABLA A Areas debajo de la curva normal.*

$\frac{X_i - \mu}{\sigma}$	0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.00
−3.5	0.00017	0.00017	0.00018	0.00019	0.00019	0.00020	0.00021	0.00022	0.00022	0.00023
−3.4	0.00024	0.00025	0.00026	0.00027	0.00028	0.00029	0.00030	0.00031	0.00033	0.00034
−3.3	0.00035	0.00036	0.00038	0.00039	0.00040	0.00042	0.00043	0.00045	0.00047	0.00048
−3.2	0.00050	0.00052	0.00054	0.00056	0.00058	0.00060	0.00062	0.00064	0.00066	0.00069
−3.1	0.00071	0.00074	0.00076	0.00079	0.00082	0.00085	0.00087	0.00090	0.00094	0.00097
−3.0	0.00100	0.00104	0.00107	0.00111	0.00114	0.00118	0.00122	0.00126	0.00131	0.00135
−2.9	0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0017	0.0018	0.0019
−2.8	0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025	0.0026
−2.7	0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034	0.0035
−2.6	0.0036	0.0037	0.0038	0.0039	0.0040	0.0041	0.0043	0.0044	0.0045	0.0047
−2.5	0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060	0.0062
−2.4	0.0064	0.0066	0.0068	0.0069	0.0071	0.0073	0.0075	0.0078	0.0080	0.0082
−2.3	0.0084	0.0087	0.0089	0.0091	0.0094	0.0096	0.0099	0.0102	0.0104	0.0107
−2.2	0.0110	0.0113	0.0116	0.0119	0.0122	0.0125	0.0129	0.0132	0.0136	0.0139
−2.1	0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174	0.0179
−2.0	0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222	0.0228
−1.9	0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281	0.0287
−1.8	0.0294	0.0301	0.0307	0.0314	0.0322	0.0329	0.0336	0.0344	0.0351	0.0359
−1.7	0.0367	0.0375	0.0384	0.0392	0.0401	0.0498	0.0418	0.0427	0.0436	0.0446
−1.6	0.0455	0.0465	0.0475	0.0485	0.0495	0.0505	0.0516	0.0526	0.0537	0.0548
−1.5	0.0559	0.0571	0.0582	0.0594	0.0606	0.0618	0.0630	0.0643	0.0655	0.0668
−1.4	0.0681	0.0694	0.0708	0.0721	0.0735	0.0749	0.0764	0.0778	0.0793	0.0808
−1.3	0.0823	0.0838	0.0853	0.0869	0.0885	0.0901	0.0918	0.0934	0.0951	0.0968
−1.2	0.0895	0.1003	0.1020	0.1038	0.1057	0.1075	0.1093	0.1112	0.1131	0.1151
−1.1	0.1170	0.1190	0.1210	0.1230	0.1251	0.1271	0.1292	0.1314	0.1335	0.1357
−1.0	0.1379	0.1401	0.1423	0.1446	0.1469	0.1492	0.1515	0.1539	0.1562	0.1587
−0.9	0.1611	0.1635	0.1660	0.1685	0.1711	0.1736	0.1762	0.1788	0.1814	0.1841
−0.8	0.1867	0.1894	0.1922	0.1949	0.1977	0.2005	0.2033	0.2061	0.2090	0.2119
−0.7	0.2148	0.2177	0.2207	0.2236	0.2266	0.2297	0.2327	0.2358	0.2389	0.2420
−0.6	0.2451	0.2483	0.2514	0.2546	0.2578	0.2611	0.2643	0.2676	0.2709	0.2743
−0.5	0.2776	0.2810	0.2843	0.2877	0.2912	0.2946	0.2981	0.3015	0.3050	0.3085
−0.4	0.3121	0.3156	0.3192	0.3228	0.3264	0.3300	0.3336	0.3372	0.3409	0.3446
−0.3	0.3483	0.3520	0.3557	0.3594	0.3632	0.3669	0.3707	0.3745	0.3783	0.3821
−0.2	0.3859	0.3897	0.3936	0.3974	0.4013	0.4052	0.4090	0.4129	0.4168	0.4207
−0.1	0.4247	0.4286	0.4325	0.4364	0.4404	0.4443	0.4483	0.4522	0.4562	0.4602
−0.0	0.4641	0.4681	0.4721	0.4761	0.4801	0.4840	0.4880	0.4920	0.4960	0.5000

* Proporción el área total bajo la curva que está debajo de la curva desde $-\infty$ a $(X_i - \mu)/\sigma$ (X_i representa cualquier valor deseado de la variable X).

TABLA A (Continuación)

$\frac{X_i - \mu}{\sigma}$	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
+0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
+0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
+0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
+0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
+0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
+0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
+0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
+0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
+0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8079	0.8106	0.8133
+0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
+1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
+1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
+1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
+1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
+1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
+1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
+1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
+1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
+1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
+1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
+2.0	0.9773	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
+2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
+2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
+2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
+2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
+2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
+2.6	0.9553	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
+2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
+2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
+2.9	0.9981	0.9982	0.9983	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
+3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.99900
+3.1	0.99903	0.99906	0.99910	0.99913	0.99915	0.99918	0.99921	0.99924	0.99926	0.99929
+3.2	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950
+3.3	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965
+3.4	0.99966	0.99967	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976
+3.5	0.99977	0.99978	0.99978	0.99979	0.99980	0.99981	0.99981	0.99982	0.99983	0.99983

TRABAJO PRÁCTICO Nº 5

De las variables con que viene trabajando, cuales piensa que distribuyen normalmente.

1.

2.

3.

4.

5.

6.

7.

8.

9.

10.

6. Estimación de la media poblacional :

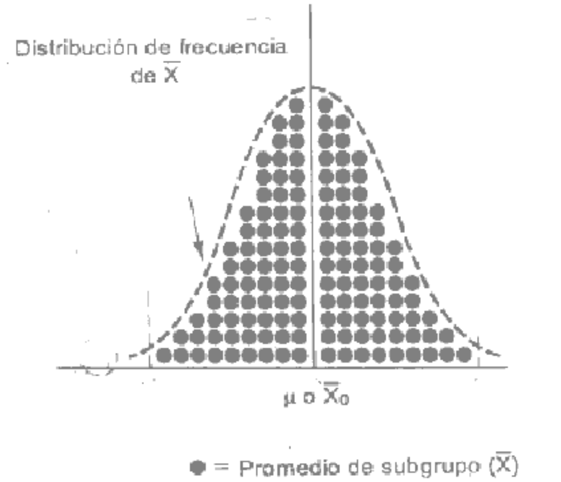
Los valores estadísticos muestrales se utilizan como estimadores de los parámetros de población. Así, la media de una muestra se utiliza como una estimación del valor medio de la población; una desviación estándar muestral se emplea como una estimación de la desviación estándar d ella población; y la proporción de elementos de una muestra se utiliza como estimación de la proporción de la población. Tales estimaciones reciben el nombre de estimaciones de punto, o puntuales, ya que proporcionan la estimación d eun solo valor de un parámetro. Sin embargo, se sabe que las muestras aleatorias tienden a producir muestras en las que, por ejemplo, la media de la muestra no es igual a la de la población, aunque ambos valores generalmente están muy cercanos entre sí. Debido a la variabilidad en el muestreo, suele ser deseable incluir una "estimación de intervalo" para acompañar la estimación puntual. Esta estimación proporciona un intervalo de los valores posibles para el parámetro de la población.

Explicación de la estimación por intervalo

Por ejemplo, supongamos que se toma una muestra de la edad de los profesionales, en la que se observa que la edad promedio es 35.3 años. Se sabe que este es uno de los valores de la distribución en el muestreo, pero ¿cuán cercano está 35.3 a la media de la población?.

La estimación de intervalo de la media poblacional se basa en el supuesto de que la distribución en el muestreo de los valores medios de la muestra es normal. En caso de tamaños de muestras considerables, esto no constituye una gran dificultad, puesto que se aplica el Teorema del Límite Central. Sin embargo, para muestras de 30 observaciones o menos, es importante saber que la población que se muestrea está distribuida normalmente, o por lo menos, casi normalmente. De otra manera, no pueden utilizarse estas técnicas.

Graficamente



Con esta premisa, y suponiendo distribución normal de nuestra media muestral, se sabe que alrededor del 68 % de los valores estadísticos de la muestra están comprendidos dentro de una desviación estándar de la media muestral, y que casi el 95 % de los valores medios de la muestra estarán comprendidos dentro de 1.96 desviaciones estándar de la media. Por la misma razón se sabe que el 32 % de los posibles valores medios de la muestra estarán mas allá de 1 desviación estándar (es decir, 1.00 - 0.68), y que aproximadamente el 5 % de los valores medios de la muestra estarán a más de 1.96 desviaciones estándar de la media.

En consecuencia, si se establece la proposición de que la media verdadera de la población estará dentro de 1.96 desviaciones estándar de la media muestral, es posible esperar estar en lo cierto en un 95 % de las veces, y estar equivocado el 5 % restante. De esta manera decir que 35.3 queda dentro de 1.96 desviaciones estándar de la media muestral conlleva un riesgo de error del 5 %. A esta estimación probabilística del intervalo en el que puede estar comprendido el verdadero valor de la media poblacional se lo conoce como *intervalo de confianza*, y la *confianza* es 1 - P(error). Por tanto, un intervalo de confianza de 95 % implicaría un 5 % de riesgo de error; el 5 % de los intervalos así designados no incluirían la media de la población. Cabe observar que el riesgo disminuye a medida que aumenta el valor de Z; un intervalo de 2.33 presenta menor riesgo que uno con límite 1.96.

El intervalo de confianza presenta la forma $\overline{X} \pm Z\, \sigma_{\overline{x}}$; un intervalo de confianza de 95 % para la media de 35.3 sería $35.3 \pm 1.96\, \sigma_{\overline{x}}$

Hay tres factores determinantes del grado de error: 1) la confianza deseada, (representada en la fórmula como Z); 2) la dispersión en la población y 3) el tamaño de la muestra (representados ambos conceptos en el valor de $\sigma_{\overline{x}}$, desviación estándar de la distribución de muestreo). Se debe recordar que a medida que aumenta el tamaño de la muestra, la desviación estándar de la muestra disminuye. Por tanto muestras grandes tenderán a producir valores medios muestrales más cercanos a la media poblacional que los de las muestras pequeñas. Además, la variabilidad en la población total también es un factor importante; cuanto mayor sea la variabilidad en la población, mayor será la variabilidad en la muestra y por lo tanto mayor el intervalo de confianza.

Desviación estándar poblacional conocido

El método empleado en la estimación de la media de una población depende, entre otras cosas, de si se conoce o no la desviación estándar de la misma o si ésta se debe estimar a partir de los datos muestrales.

Si se conoce, entonces $\sigma_{\bar{x}} = \sigma_x / \sqrt{n}$

Ej. Intervalo de confianza para la media poblacional, conociendo la $\sigma_x = 3$ y siendo el tamaño de la muestra n = 36, media muestral = 35.3 años.

CONFIANZA DESEADA	Z	CÁLCULOS	Error "e"	INTERVALO
90 %	1.65	$35.3 \pm 1.65\ (3 / \sqrt{36})$	$\pm$ 0.825	34.475 a 36.125
95 %	1.96	$35.3 \pm 1.96\ (3 / \sqrt{36})$	$\pm$ 0.980	34.320 a 36.280
99 %	2.58	$35.3 \pm 2.58\ (3 / \sqrt{36})$	$\pm$ 1.290	34.010 a 36.590

Al ser la población mayor a 30 se aplica el teorema del límite central por lo que no tiene sentido preguntar si la población es o no normal.

Error de estimación

El error en una estimación se refiere a la desviación (diferencia) entre el valor medio de la muestra y la media real de la población. Como el intervalo de confianza está centrado con respecto al valor medio muestral, el error máximo probable equivale a la mitad de la amplitud del intervalo.

Siendo el error:

$$\text{Error} = Z\ (\sigma_{x} / \sqrt{n}\ )$$

Determinación del tamaño de muestra :

Una de las preguntas que con mayor frecuencia se plantean en estadística es: De que tamaño debe ser la muestra?.

El mismo dependerá de 1) el grado de confianza deseado, 2) la cantidad de dispersión entre los valores individuales de la población y 3) cierta cantidad especificada de error tolerable.

$$\text{Error} = Z\ (\sigma_{x} / \sqrt{n}\ )$$

$$\sqrt{n} = Z\ (\sigma_{x} / \text{Error}\ )$$

$$n = (\ Z\ (\sigma_{x} / \text{Error}\ )\ )^{2}$$

Estimación de la media cuando no se conoce el desvío estándar poblacional (σ_x) :

Cuando no se conoce el valor de la desviación estándar de la población (lo cual generalmente ocurre), la desviación estándar de la muestra se utiliza como una estimación de la desviación estándar poblacional. Además, por el teorema del límite central se sabe que, cuando el tamaño d ella muestra es mayor que 30, la distribución de muestreo de las medias será casi normal. Sin embargo para muestras de 30 o menos observaciones, la aproximación normal resulta inadecuada. En lugar de ello, los cálculos de los intervalos de confianza se deben basar en la distribución t de student.

La forma de la distribución t es muy parecida a la de la distribución normal, la diferencia principal entre las dos consiste en que la distribución t presenta un área (probabilidad) mayor en sus extremos (colas). Esto significa que, para un nivel de confianza dado, el valor t será un poco mayor que el correspondiente a Z.. Como existe una distribución t diferente para cada tamaño de muestra, sería poco práctico intentar proporcionar tablas completas de las distribuciones. En lugar de esto, en la tabla solamente se incluyen los valores que se utilizan más comunmente.

Para utilizar una tabla de valores t, se deben conocer dos cosas: el nivel de confianza deseado y los grados de libertad. Estos últimos se relacionan con la forma como se calcula la desviación estándar de la muestra:

$$\boxed{s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}}$$

Donde Sx = desvío estándar de la muestra
n- 1 = grados de libertad

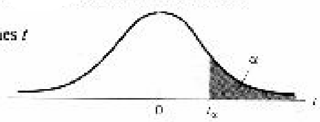
Así, los grados de libertad son igual a n-1, o una unidad menor que el tamaño muestral.
Para utilizar la tabla se debe especificar el área en los extremos (colas) de la distribución (riesgo) y los grados de libertad (ver tabla).

El intervalo de confianza para una media muestral cuando se usa Sx, es muy semejante al intervalo con σx.

De modo, el intervalo es:

$$\bar{X} \pm t \frac{Sx}{\sqrt{n}}$$

TABLA Distribuciones t



La siguiente tabla proporciona los valores de t_α que corresponden a una determinada área α de cola superior y un número especificado de grados de libertad.

Grados de libertad	Area α de cola superior							
	.1	.05	.025	.01	.005	.0025	.001	.0005
1	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.598
3	1.638	2.353	3.182	4.541	5.841	7.453	10.214	12.924
4	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	1.371	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.767
24	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

Fuente: De Ronald A. Fisher: *Statistical Methods for Research Workers*, 14th ed., copyright © 1970, University of Adelaide.

Ej. Cálculo de intervalos de confianza utilizando valores de t

Media muestral 20
Desviación estándar de la muestra Sx 1.5
Tamaño de la muestra 25 (grados de libertad n- 1 = 24)

CONFIANZA DESEADA	t	CÁLCULOS	INTERVALO
90 %	1.711	20.0 ± 1.711 (1.5/√ 25)	20.00 ± 0.5133
95 %	2.064	20.0 ± 2.064 (1.5/√ 25)	20.00 ± 0.6192
99 %	2.797	20.0 ± (1.5/√ 25)	20.00 ± 0.8391

Muestreo a partir de poblaciones pequeñas: El factor finito de corrección.

Cuando la población es finita y el tamaño de la muestra constituye más del 5 % de la población, se usa el factor finito de corrección para modificar las desviaciones estándar de las fórmulas.

	Intervalo de Confianza	*Error*
con σ_x conocida:	$\bar{x} \pm z \frac{\sigma_x}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$	$z \frac{\sigma_x}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$
con σ_x no conocida:	$\bar{x} \pm t \frac{s_x}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$	$t \frac{s_x}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$

Estimación de la proporción en una población :

¿Qué Porcentaje de los productos de un gran cargamento está defectuoso? ¿Qué porcentaje de los votantes aprobarán un decreto?. Estas preguntas y otras semejantes se pueden contestar utilizando datos muestrales para estimar el parámetro de la población. Como se hizo anteriormente, las estimaciones suelen establecerse, en términos de estimaciones depunto y de intervalo.

La estimación de proporciones de la población es similar a la de medias de la población. Por ejemplo, un intervalo de confianza de una muestra grande se basa en una distribución de muestreo que es aproximadamente normal, y el valor estadístico de la muestra (en este caso, a proporción de la muestra) se utiliza como estimación de punto del parámetro verdadero (proporción de la población). La excepción la constituye que no se utiliza la distribución t de Student.

La estimación de intervalo del parámetro de la población (para tamaños grandes de muestra) es simétrica respecto de la proporción de la muestra, del mismo modo que el intervalo para una media de la población es simétrico respecto a la media muestral. La principal diferencia entre la estimación de medias y la de proporciones radica en las desviaciones estándar de las distribuciones de muestreo. La desviación estándar de una proporción se basa en la distribución binomial.

Intervalos de confianza :

El valor esperado de una proporción muestral (es decir, la media de una distribución de muestreo de proporciones muestrales) siempre es igual a la proporción de la población verdadera. Por tanto, a roporción de la muestra se utiliza como la estimación de punto de la proporción verdadera:

$$\text{Estimación de punto de p: } P = \frac{X}{n}$$

La estimación de intervalo del parámetro de la población (para tamaños grandes de muestra) es simétrica respecto de la proporción de la muestra, del mismo modo que el intervalo para la media poblacional es simétrico respecto de la media muestral. La principal diferencia entre la estimación de medias y la de proporciones radica en los desvíos estándar de las distribuciones de muestreo. La desviación estándar de una proporción se basa en la distribución binomial. La estimación de σ_p se presenta a continuación.

$$\sigma_{x/n} = \sqrt{\frac{(x/n)[1-(x/n)]}{n}}$$

Donde

$\sigma_p = \sigma_{x/n} =$ Desvío estándar de la proporción

x = Número de elementos de la muestra

n = Tamaño de la muestra

Estimación de Intervalo de *p:*

$$p: \frac{x}{n} \pm z\sqrt{\frac{(x/n)[1-(x/n)]}{n}}$$

Ejemplo de cálculo de intervalos de confianza para proporciones

n	x	Conf. deseada	z	x/n	Error	Intervalo de confianza
40	8	90 %	1.65	8/40 = 0.20	1.65√0.20(0.80)/40=0.104	0.096 a 0.304
80	20	95 %	1.96	20/80=0.25	1.96√0.25(0.75)/80=0.095	0.155 a 0.345
100	30	98 %	2.33	30/100=0.30	2.33√0.30(0.70)/100=0.107	0.193 a 0.417

Determinación del tamaño de la muestra :

La determinación del tamaño muestral es necesario para obtener un grado deseado de precisión en la estimación de proporciones.

Para calcular el error posible en la estimación del intervalo de confianza, en la ecuación interviene el valor de la proporción muestral (p), que justamente que queremos estimar. Para salvar este inconveniente en condiciones de completa incertidumbre, inicialmente se puede suponer que la proporción muestral (p) es igual al 50 %, y esto producirá el mayor error posible en la estimación de intervalo, ya que:

estimación de intervalo = $p \pm$ error

$$\text{error} = z * \sqrt{p\,(1-p)} \,/\, \sqrt{n}$$

$$\text{error}^2 = z^2 * \{\, p\,(1-p) / n \,\}$$

$$n = z^2 * \{\, p\,(1-p) / \text{error}^2 \,\}$$

Por otra parte, si se dispone de alguna información con respecto al tamaño de la proporción muestral -digamos una pequeña muestra piloto- entonces es posible reducir el tamaño del intervalo, o bien, el tamaño muestral necesario. De los contrario se utiliza p = 0.50 para obtener el intervalo más amplio posible.

Ej.

Confianza deseada = 95 %
z = 1.96
error = 0.08

$$n = 1.96^2 \{\, (0.50)(0.5) / 0.08^2 \,\} = 3.84 \{0.25 / 0.0064 \} = 149.9 \text{ o bien, } 150$$

Estimación de la Varianza Poblacional :

Si se desea estimar la varianza poblacional de una variable que se distribuye normalmente, la construcción del intervalo de confianza ofrece como particularidad que la función pivotal no es simétrica. La función de distribución es una Ji cuadrado con n-1 grados de libertad, si no se conoce la media poblacional. Si se conoce la media poblacional la función de distribución sigue también una Ji cuadrado pero con n grados de libertad.

TRABAJO PRÁCTICO Nº 6

De las variables con las que viene trabajando, como expresaria el intervalo de confianza, si el mismo es del 95 %?, tenga presente aquellos casos donde se trata de proporciones o de medias.

1.

2.

3.

4.

5.

6.

7.

8.

9.

10.

7 Pruebas de significación o Test de hipótesis :

La finalidad de la prueba de significación es decidir si una afirmación acerca de un parámetro de población es verdadera. Por ejemplo es posible desear determinar si afirmaciones como las siguientes son ciertas:

- El 5% de la producción es defectuosa.
- El peso promedio de las latas de tomates es 200 gramos
- El 45% de los votantes prefieren a tal candidato

El **objeto de la prueba de significación** es evaluar proposiciones o afirmaciones acerca de los valores de los parámetros poblacionales.

El primer paso de la prueba de significación es formular dos hipótesis con respecto a un aserto. Las hipótesis son teorías que intentan informar acerca de hechos observados en situaciones en las que existen algunos factores desconocidos. En el caso del ejemplo 1, la incógnita es el verdadero porcentaje de producción defectuosa. Si como resultado de una muestra determinamos que el porcentaje es mayor, digamos 8%. Cuál será la causa del hecho de que una muestra señalara una diferencia del 3% con la proposición de que la producción defectuosa es del 5%?. Una posibilidad es la variabilidad del muestreo.

Estas dos proposiciones reciben distintos nombres formales. La que señala que la proposición es verdadera recibe el nombre de **hipótesis nula** y se representa mediante el símbolo **Ho**; y la segunda, que afirma que la proposición es falsa se denomina **hipótesis alternativa** y se designa mediante el signo **H1**.

En el ejemplo (1) la hipótesis nula es la que establece que el porcentaje de producción defectuosa es el 6%, lo que representaría mediante:

Ho: p = 5%

La alternativa es que el porcentaje de producción defectuosa p es mayor del 5%, lo cual se representaría como:

H1: p > 5%

Si el sistema, después de efectuar el análisis, acepta la hipótesis nula, significa que la discrepancia entre el porcentaje de productos defectuosos observados en la muestra (en el ejemplo 8%) y el porcentaje de elementos defectuosos de la población o sea el propuesto (en el ejemplo 5%) se debe probablemente a la variación casual del muestreo. Por el contrario, la decisión de rechazar la hipótesis nula significaría que la variación entre el valor observado y el propuesto es demasiado grande como para deberse únicamente a la casualidad.

Los valores críticos que generalmente se seleccionan en las pruebas de significación son los que comprenden riesgos del 1% o 5% de rechazar Ho cuando sea verdadera. La probabilidad de rechazar una hipótesis nula verdadera recibe el nombre de nivel de significación de una prueba.

El **nivel de significación** de una prueba es la probabilidad de rechazar una hipótesis nula que sea verdadera.

La esencia de una prueba de significación es dividir una distribución de muestreo con base en el supuesto de que Ho es verdadera, en regiones de aceptación y rechazo respecto de Ho. Un valor crítico se selecciona con base en una probabilidad especificada de que el decididor esté dispuesto a aceptar o rechazar una Ho verdadera. Un valor estadístico de prueba se calcula a partir de datos de la muestra y del valor esperado o propuesto, el cual es comparado con el valor crítico. Un valor estadístico de prueba superior al valor crítico señala que se debe rechazar la hipótesis nula, mientras que un valor de prueba menor que el valor crítico indica que se debe aceptar la hipótesis nula.

Pruebas uni y bilaterales :

El interés por identificar desviaciones no aleatorias, es decir significativas, a partir de un parámetro especificado, puede comprender desviaciones en ambas direcciones o en una sola. De ese modo y siguiendo con el ejemplo podemos querer verificar la proposición de que los defectos de producción son mayores al 5%, cuando debemos aceptar la entrega de un proveedor. Que el porcentaje es menor al 5%, cuando queremos comprobar una mejora en el proceso productivo. O saber si el porcentaje es distinto al 5%, en la práctica esta prueba se utiliza cuando la divergencia en ambas direcciones es crítica, por ejemplo diferencias en medidas de elementos que deben ajustarse entre sí. A esta última prueba se la conoce como **bilateral** y a las dos primeras como **unilateral**. La prueba unilateral es de cola izquierda cuando se comprueba si se ha cumplido con un mínimo. Y de cola derecha cuando se comprueba máximos que no deben ser excedidos.

Errores de Tipo I y de Tipo II :

Existen dos tipos de errores inherentes al proceso de test de hipótesis. Cuando se cree que la hipótesis nula es falsa siendo realmente es verdadera, también conocido como **error de tipo I**. O Cuando se cree que la hipótesis nula es verdadera siendo realmente falsa o **error de tipo II**.

Se comete un error de tipo I si se rechaza Ho cuando es verdadera. La probabilidad de cometer un error de tipo I es igual al nivel de significación del test de hipótesis.

La probabilidad de un error de tipo I es empleado en el sistema para determinar valores críticos que separan los resultados casuales de los no casuales.

Test de hipótesis para la Media y Proporción :

Las reglas de decisión para efectuar dócimas referentes a la media y la proporción guardan similitud entre si.

Si se trata de una dócima de extremo izquierdo, la regla de decisión se formula de la siguiente manera:
Si el valor del pivote es menor que el valor crítico, siendo el valor crítico aquel valor de abscisa de la función de distribución de la función pivotal hasta donde se acumula una probabilidad igual a alfa, entonces se rechaza la hipótesis nula, pues el valor del pivote cae en la zona de rechazo.

Si se trata de una dócima de extremo derecho, la regla de decisión se formula de la siguiente manera:

Si el valor del pivote es mayor que el valor crítico, siendo el valor crítico aquel valor de abscisa de la función de distribución hasta donde se acumula una probabilidad igual a (1 - alfa), entonces se rechaza la hipótesis nula, pues el valor del pivote cae en la zona de rechazo.

Si se trata de una dócima bilateral (distinto a), la regla de decisión se formula de la siguiente manera:

Si el valor absoluto del pivote es mayor que el valor crítico, siendo éste positivo y el valor de abscisa de la función de distribución de la función pivotal hasta donde se acumula una probabilidad de (1 - alfa), entonces se rechaza la hipótesis nula, pues el valor del pivote cae en zona de rechazo.

Ej. Supónga que se desea evaluar la afirmación que hace un fabricante sobre la vida útil de su producto, el cual él asegura que es de 40.000 horas. La hipótesis nula se convierte en:

Ho = 40.000 horas

Las hipótesis alternativas podrían ser:

H1 # 40.000 H1> 40.000 H1< 40.000

La evaluación d ecualquiera de las anteriores propuestas debe tener en cuenta el grado en el que el valor estadístico de la muestra puede variar o desviarse del parámetro propuesto, debido a la variación casual en el muestreo. La distribución en el muestreo será normal para muestras que se obtengan a partir de una población normal con una desviación estándar conocida, y presentará una distribución t cuando la desviación estándar de la población sea estimada a partir de la desviación estándar muestral. Cuando el tamaño de la muestra es mayor que 30, se puede hacer a un lado la necesidad de suponer que se está muestreando a partir de una población normal, ya que se aplica el teorema del límite central.

Conceptualmente, los valores críticos se pueden establecer en valores que se relacionen específicamente con un problema dado. Por ejemplo, en este caso los valores críticos pueden ser 39.000 horas y 41.000 horas. Sin embargo, es mucho más sencillo trabajar con valores estadísticos de pruebas y valores críticos estandarizados, ya que la mayoría de las tablas de probabilidad se establecen en términos de valores estandarizados.

Ahora se establece un nivel de significación, el cual, a su vez, dará lugar a un valor crítico, previo al muestreo. Con la muestra y el nivel de significación calculamos el valor estadístico de la prueba.

$$\text{Valor estadístico de prueba} = \frac{\text{Valor medio de la muestra - media propuesta}}{\text{Desvío estándar de la distribución de muestreo}}$$

Si se tiene como dato la desviación estándar de la población, el valor estadístico de prueba es:

$$Zprueba = \frac{\bar{X} - \mu_o}{\sigma_x \ / \ \sqrt{n}}$$

Si no se conoce σ_x, el valor estadístico de prueba es

$$T\ prueba = \frac{\bar{X} - \mu_o}{S_x \ / \ \sqrt{n}}$$

Donde μ_o es la media propuesta

Ej.

El fabricante asegura una vida útil de 40.000 horas con un desvío conocido de 3.500 horas, se toma una muestra de n = 49 piezas, el valor medio muestral es de 38.000 horas. Realizar la prueba de significación.

1. Se establece la hipótesis nula y alternativa:

   Ho = 40.000 horas

   Si la prueba la realiza un consumidor deseará saber si la vida útil es menor a 40.000 horas.

   H1 < 40.000 horas

2. Se selecciona el nivel de significación. Por ejemplo 5 % de riesgo.

3. Como se conoce σ_x = 3.500 horas y la muestra es grande, mayor a 30 elementos se utiliza la distribución normal.

$$Zprueba = \frac{\bar{X} - \mu_o}{\sigma_x / \sqrt{n}} = \frac{38.000 - 40.000}{3.500 / \sqrt{49}} = -4.0$$

4. Comparar el valor de prueba con el valor crítico. Como -4.0 excede a -1.65, Ho es rechazada. De modo que se concluye que el promedio de vida útil es menor que 40.000 horas.

Test de Hipótesis para la Varianza :

Las reglas de decisión para efectuar dócimas referentes a la varianza guardan similitud a las de medias o proporciones.

Si se trata de una dócima de extremo izquierdo, la regla de decisión se formula de la siguiente manera:

Si el valor del pivote es menor que el valor crítico, siendo el valor crítico aquel valor de abscisa de una función de distribución Ji cuadrado, hasta donde se acumula una probabilidad igual a alfa, entonces se rechaza la hipótesis nula, pues el valor del pivote cae en la zona de rechazo.

Si se trata de una dócima de extremo derecho, la regla de decisión se formula de la siguiente manera:

Si el valor del pivote es mayor que el valor crítico, siendo el valor crítico aquel valor de abscisa de una función de distribución Ji cuadrado hasta donde se acumula una probabilidad igual a (1 - alfa), entonces se rechaza la hipótesis nula, pues el valor del pivote cae en la zona de rechazo.

Si se trata de una dócima bilateral (distinto a), la regla de decisión se formula de la siguiente manera:

Si el valor absoluto del pivote es menor que el valor de absisa de una función de distribución Ji cuadrado con (n - 1) grados de libertad hasta donde se acumula una probabilidad igual a alfa/2 o si el valor del pivote es mayor que el valor de abscisa de igual distribución hasta donde se acumula una probabilidad de (1 - alfa/2), entonces se rechaza la hipótesis nula, pues el valor del pivote cae en zona de rechazo.

TRABAJO PRÁCTICO Nº 7

De las variables con las que viene trabajando, como expresaria una dósima o test de hipótesis.

1.

2.

3.

4.

5.

6.

7.

8.

9.

10.

8. Glosario estadístico

Estimación :

Es el proceso mediante el cual, utilizando los datos muestrales, arribamos aproximadamente a conocer el verdadero valor del parámetro poblacional desconocido. Esencialmente, cualquier característica de la población se puede estimar a partir de una muestra **al azar**.

Los valores estadísticos muestrales se utilizan como estimaciones de los parámetros de la población. Así, la media de una muestra se utiliza como una estimación de la media poblacional; la proporción de elementos de una muestra con cierta característica en común se utiliza como estimación de la población. A éstas estimaciones se las conoce como estimaciones puntuales. Pero debido a que la media de una muestra aleatoria no es necesariamente igual a la media poblacional, aunque ambos valores generalmente se encuentran muy cercanos entre sí. Debido a la variabilidad en el muestreo, suele ser deseable incluir una estimación de intervalo para acompañar a la estimación puntual.

Estimación por intervalo :

Incluye un rango de valores posibles entre un máximo y un mínimo en el que se considera que está comprendido el parámetro de la población desconocido.

Un **intervalo de confianza** o estimación por intervalo, proporciona un intervalo de valores, centrado en el valor estadístico de la muestra, en el cual supuestamente se ubica el parámetro de la población, con un riesgo de error conocido.

Test de hipótesis :

La otra rama de la estadística inferencial la comprende la prueba de significación o **test de hipótesis**. La finalidad del test de hipótesis es decidir si una afirmación acerca de un parámetro de población es verdadera

El aspecto principal de la prueba de significación es determinar si la diferencia entre un valor propuesto de un parámetro de población y el valor estadístico de la muestra se debe razonablemente a la variabilidad del muestreo, o si la discrepancia es demasiado grande para ser considerada por esa causa.

El procedimiento general para efectuar la prueba de significación es el siguiente:

1. Formular una hipótesis nula, o enunciado que expresa que el parámetro de la población es como se especificó (es decir, que la proposición es verdadera);
2. Formular una hipótesis alternativa, es decir una alternativa a la proposición (por ejemplo, el parámetro es distinto al valor propuesto);

3. Seleccionar un nivel de significación, o probabilidad de rechazar una hipótesis nula que sea verdadera.

AUTOEVALUACIÓN MÓDULO 2

Nombre y apellido: ....

Fecha: / /

PROBLEMA 1

En el Teorema del límite central, qué es lo que distribuye normalmente?

a. El desvío poblacional □
b. La varianza muestral □
c. La media muestral □

PROBLEMA 2

a. Cuándo una muestra es grande?. Indique la cantidad de elementos muestrales Mínimos que debe tener la muestra para cumplir esa condición.

n = □

b. Cuál es la ventaja de contar con una muestra grande?

....

PROBLEMA 3

A su criterio, ¿qué población produciría menor error en el muestreo?. En todas se eligen 45 elementos muestrales.

a. Desvío estándar 10 cm y media 100 cm
b. Desvío estándar 0,1 mm y media 0,5 mm
c. Varianza 1 mm² y media 4 mm

Resp. a. □ b. □ c. □

PROBLEMA 4

Si el intervalo de confianza de la edad media de los profesionales de la Empresa está comprendido entre un valor mínimo de 31,4 años y un valor máximo de 35,4, siendo estos datos obtenidos de una muestra aleatoria de 50 elementos tomados entre todos los profesionales de la Empresa.

1.- Cuál es el error de muestreo cometido?. □

2.- Cúal es la media muestral? □

Enviar antes de las 48 horas de recibido - muchas gracias -

REFERENCIAS BIBLIOGRAFICAS

- **Estadística para Administración y Economía**
  Stevenson, Editorial Harla
- **Muestreo**
  Morris J. Sloni M., Editorial Americana
- **Análisis y Planeación de la Calidad**
  Juran Gryna, Editorial Mc Graw Hill
- **Análisis Estadístico**
  Ya-Lun Chou, Editorial Mc Graw Hill
- **Control de Calidad**
  Besterfield, Editorial Prentic Hall
- **Introducción a la Inferencia Estadística**
  Castello Minolli, Editorial El Coloquio

ESTADÍSTICA APLICADA MÓDULO 3

Lic. OSVALDO CASTILLO

2009

E-mail: castillo@cqc.com.ar Internet: http://www.cqc.com.ar

INDICE

9.0 GRÁFICAS DE CONTROL

Gráficas de Control
Por qué?

- Excelente técnica en la resolución de problemas
- Consecuente mejora de la calidad

Mejora de la calidad.
Cuándo?

- Al principio un proceso, por lo general es inestable
- Prueba y evaluación de ideas
- Control del proceso en régimen normal

AXIOMA DE LA PRODUCCIÓN DE BIENES O SERVICIOS

o

Ley inherente a la naturaleza

Nunca se producen dos bienes o servicios que sean exactamente iguales

Clases de Variaciones

- Variación en el mismo producto o servicio
- Variación de un producto o servicio a otro
- Variación en el tiempo

Variación en el mismo producto o servicio: Ejemplo diferente tonalidad en la pintura de un elemento, una zona más aspera que otra, diferente calidad de atención en una misma caja, diferentes problemas en la impresión de una factura.

Variación de un producto o servicio a otro: Ejemplo diferente tonalidad en la pintura de un elemento respecto de otro, un producto más áspero que otro, diferente calidad de atención de una caja a otra, diferentes problemas en la impresión de varias facturas.

Variación en el tiempo: Son las diferencias que se producen de una hora a otra, a lo largo del tiempo.

Causas de Variaciones

- Equipo
- Material
- Entorno
- Operario

Variaciones debidas al Equipo

- Desgaste de herramientas
- Vibraciones
- Posicionamientos
- Fluctuaciones eléctricas
- Fluctuaciones hidráulicas
- etc.

Variaciones debidas al Material

- Resistencias
- Ductilidad
- Grosor
- Porosidad
- Contenido de humedad
- Etcétera

Variaciones debidas al Entorno

- La temperatura
- La luz
- La radiación
- El tamaño de las partículas
- La presión
- Etcétera

Variaciones debidas al Operario

- Método que emplea
- Bienestar emocional
- Bienestar físico
- Etcétera

Las variaciones naturales o previstas de estas cuatro causas, producirán un patrón estable de diversas causas fortuitas (causas aleatorias), dando un proceso de producción en estado de control estadístico predecible

Aquellas causas de variación cuyas magnitudes sean grandes e identificables fácilmente, se las clasifica como *causas atribuibles*

Magnitudes de Variación

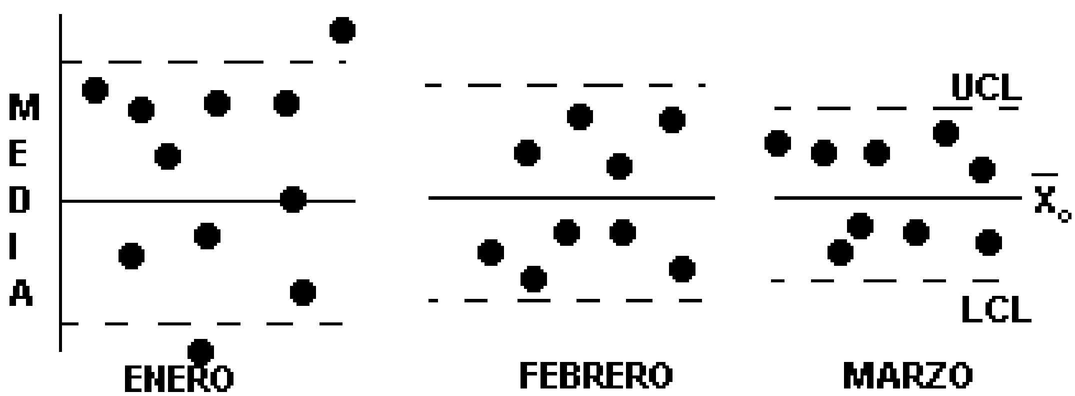
- Aleatorias, naturales y previstas de poca importancia, no asignables
- Atribuibles, se las puede identificar fácilmente, de gran magnitud o asignables

El método de la gráfica de control

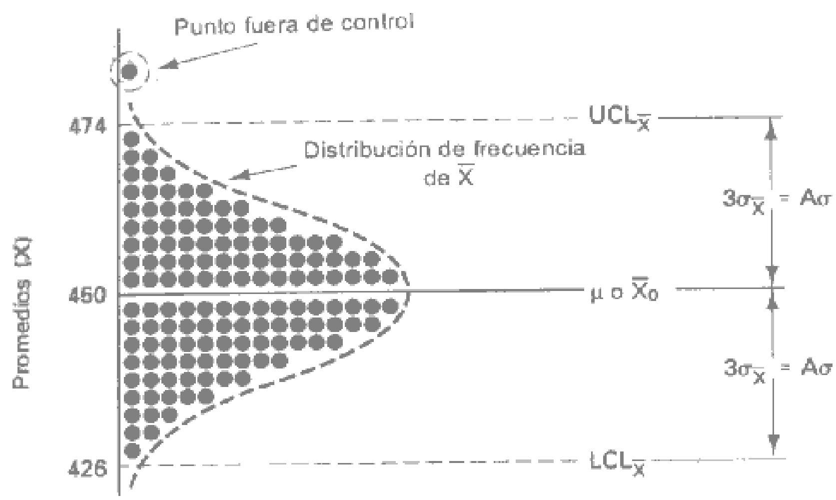
- Permite identificar rápidamente las variaciones debidas a causas asignables de las que no lo son

Las gráficas muestran el comportamiento del proceso

- **Las gráficas de control son recursos excelentes al facilitar el mejoramiento de la calidad**

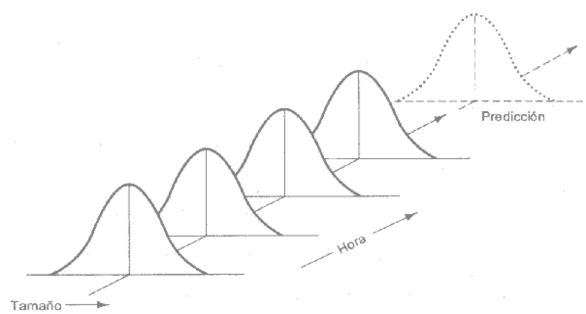


Explicación Gráfica
Distribución de medias muestrales



Como se vio en el módulo dos, según el Teorema del Límite Central cuando la muestra es grande la media muestral distribuye normalmente, si aplicamos este teorema podemos conjeturar que sí la media de una conjunto de observaciones supera el valor de la media de las medias más 3 desvíos estándar, entonces esa variación no será aleatoria, será producto de una causa atribuíble al equipo y/o el material y/o el entorno y/o el operario.

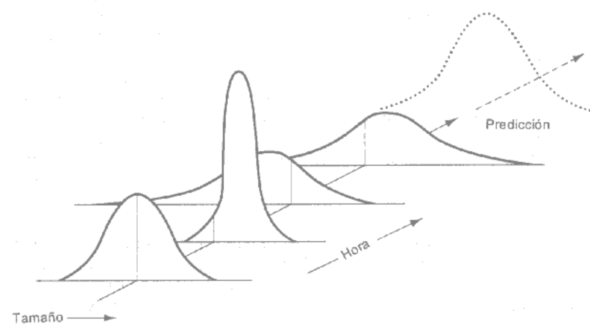
Variaciones no asignables



(a) Unicas causas fortuitas de variación que se presentan

Si sólo existen causas de variación fortuitas, el proceso es estable y predecible permanentemente, como se muestra en la figura. Sabemos que las variaciones que se produzcan en el futuro, tal como se indica en la figura de puntos será la misma a menos que se introduzca un cambio en el proceso debido a una causa atribuible.

Variaciones Asignables



(b) Causas atribuibles de variación se presentan

El efecto a lo largo del tiempo de las causas de variación atribuibles. La naturaleza no natural, inestable de la variación, impide predecir las variaciones futuras. Es necesario determinar las causas atribuibles y corregirlas si es que se desea que continúe en proceso natural, estable.

Cálculo de las líneas centrales

$$\overline{\overline{X}} = \frac{\sum_{i=1}^{g} \overline{X}_i}{g} \qquad \overline{R} = \frac{\sum_{i=1}^{g} R_i}{g} \qquad \overline{S} = \frac{\sum_{i=1}^{g} \bar{s}_i}{g}$$

$\overline{\overline{X}}$ = promedio de los promedios de los subgrupos
$\overline{X}_i$ = promedio del subgrupo i
g = cantidad de subgrupos
$\overline{R}$ = promedio de los rangos de los subgrupos
R_i = rango del subgrupo i
$\overline{S}$ = promedio de las desviaciones estándar de las muestras de los subgrupos
$\bar{s}_i$ = desviación estándar de la muestra de los valores del subgrupo del subgrupo i

Cálculo de los límites $\overline{X}$ & R

$$UCL_{\overline{X}} = \overline{\overline{X}} + A_2 R \qquad UCL_R = D_4 R$$

$$LCL_{\overline{X}} = \overline{\overline{X}} - A_2 R \qquad LCL_R = D_3 R$$

UCL = límite de control superior

LCL = límite de control inferior

A2, D4, D3 = son factores que dependen del tamaño del subgrupo, se obtienen de la Tabla 2 (Factores de cálculos para las gráficas de control) ISO 8258: 1991(E)

Cálculo de los límites $\overline{X}$ & S

$$UCL_{\overline{X}} = \overline{\overline{X}} + A_3 \overline{S} \qquad UCL_S = B_4 \overline{S}$$

$$LCL_{\overline{X}} = \overline{\overline{X}} - A_3 \overline{S} \qquad LCL_S = B_3 \overline{S}$$

UCL = límite de control superior

LCL = límite de control inferior

A3, B4, B3 = son factores que dependen del tamaño del subgrupo, se obtienen de la Tabla 2 (Factores de cálculos para las gráficas de control) ISO 8258: 1991(E)

OBSERVACIONES EN LA MUESTRA, n	GRAFICA PARA PROMEDIOS — FACTORES PARA LOS LIMITES DE CONTROL			GRAFICA PARA LAS DESVIACIONES ESTANDAR — FACTOR PARA LINEA CENTRAL	GRAFICA PARA LAS DESVIACIONES ESTANDAR — FACTORES PARA LIMITES DE CONTROL				GRAFICA DE LOS RANGOS — FACTOR PARA LINEA CENTRAL	GRAFICA DE LOS RANGOS — FACTORES PARA LOS LIMITES DE CONTROL			
	A	A_2	A_3	c_4	B_3	B_4	B_5	B_6	d_2	d_3	D_1	D_2	D_3
2	2.121	1.880	2.659	0.7979	0	3.267	0	2.606	1.128	0.853	0	3.686	0
3	1.732	1.023	1.954	0.8862	0	2.568	0	2.276	1.693	0.888	0	4.358	0
4	1.500	0.729	1.628	0.9213	0	2.266	0	2.088	2.059	0.880	0	4.698	0
5	1.342	0.577	1.427	0.9400	0	2.089	0	1.964	2.326	0.864	0	4.918	0
6	1.225	0.483	1.287	0.9515	0.030	1.970	0.029	1.874	2.534	0.848	0	5.078	0
7	1.134	0.419	1.182	0.9594	0.118	1.882	0.113	1.806	2.704	0.833	0.204	5.204	0.076
8	1.061	0.373	1.099	0.9650	0.185	1.815	0.179	1.751	2.847	0.820	0.388	5.306	0.136
9	1.000	0.337	1.032	0.9693	0.239	1.761	0.232	1.707	2.970	0.808	0.547	5.393	0.184
10	0.949	0.308	0.975	0.9727	0.284	1.716	0.276	1.669	3.078	0.797	0.687	5.469	0.223
11	0.905	0.285	0.927	0.9754	0.321	1.679	0.313	1.637	3.173	0.787	0.811	5.535	0.256
12	0.866	0.266	0.886	0.9776	0.354	1.646	0.346	1.610	3.258	0.778	0.922	5.594	0.283
13	0.832	0.249	0.850	0.9794	0.382	1.618	0.374	1.585	3.336	0.770	1.025	5.647	0.307
14	0.802	0.235	0.817	0.9810	0.406	1.594	0.399	1.563	3.407	0.763	1.118	5.696	0.328
15	0.775	0.223	0.789	0.9823	0.428	1.572	0.421	1.544	3.472	0.756	1.203	5.741	0.347
16	0.750	0.212	0.763	0.9835	0.448	1.552	0.440	1.526	3.532	0.750	1.282	5.782	0.363
17	0.728	0.203	0.739	0.9845	0.466	1.534	0.458	1.511	3.588	0.744	1.356	5.820	0.378
18	0.707	0.194	0.718	0.9854	0.482	1.518	0.475	1.496	3.640	0.739	1.424	5.856	0.391
19	0.688	0.187	0.698	0.9862	0.497	1.503	0.490	1.483	3.689	0.734	1.487	5.891	0.403
20	0.671	0.180	0.680	0.9869	0.510	1.490	0.504	1.470	3.735	0.729	1.549	5.921	0.415

GRÁFICAS DE CONTROL POR ATRIBUTOS

Tipos de Atributos

- Cuando NO es posible hacer mediciones
  - color
  - rayaduras
  - partes faltantes
  - etc..
- Cuando SI es posible hacer mediciones pero es muy costoso o no hay tiempo
  - pasa / no pasa

Terminología de atributos que no cumplen con las especificaciones

- No conformidad
- Defecto
- Unidad no conforme

Tipos de Gráficas por Atributos mas completas

- Proporción de unidades no conformes presentes en una muestra (p)
- Cantidad de unidades no conformes presentes en un muestra (np)

Gráficas para el Control por Numero de **unidades no conformes**

$$p = np / n$$

donde:
p = proporción o fracción de no conformidad
np = cantidad de elementos no conforme de la muestra o subgrupo
n = cantidad de elementos de la muestra o subgrupo

Como construir una gráfica *p*

- Definir las características de calidad
- Estimación del subgrupo
- Recopilación de datos
- Calculo de línea central y limites de control
- Calculo de línea central y limites de control corregidos

Calculo de limites en gráfica *p*

$$UCL = \bar{P} + 3\sqrt{\frac{\bar{P}(1-\bar{P})}{n}}$$

$$LCL = \bar{P} - 3\sqrt{\frac{\bar{P}(1-\bar{P})}{n}}$$

UCL = Limite de control Superior

LCL = Limite de control Inferior

Gráfica de **cantidad de no conformes** *np*

- Características -

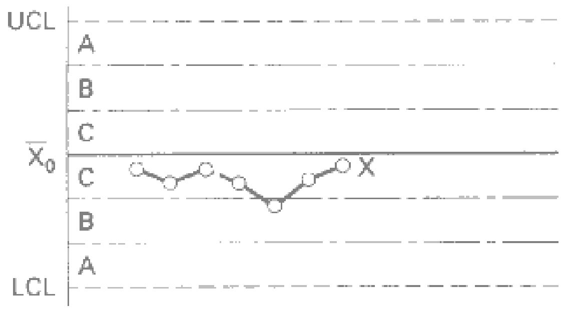
- Los subgrupos deben ser constantes
- Equivale a la gráfica *p* multiplicada por el *n*
- Es mas fácil de comprender por el personal ya que se muestran las cantidades diarias de no conformidades sin hacer cálculos

Calculo de los limites en gráfica *np*

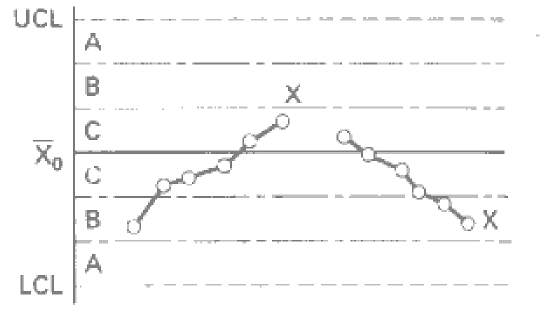
$$\text{Línea Central} = n\,p_0$$

$$\text{Límites de Control} = n\,p_0 \pm 3\sqrt{n\,p_0\,(1-p_0)}$$

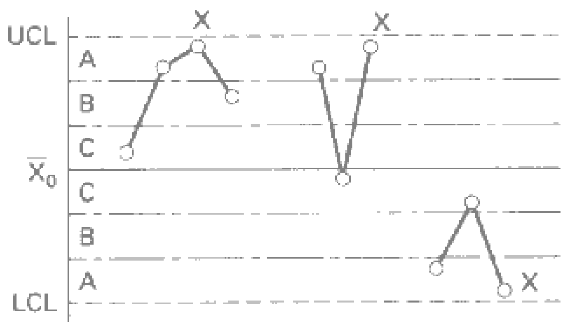
ANÁLISIS DE LAS GRÁFICAS DE CONTROL



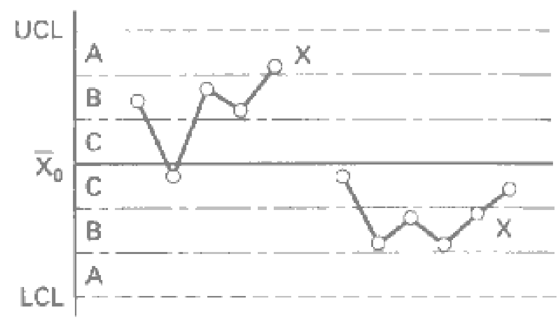
(a) Siete puntos en una hilera en la zona C o más allá



(b) Dos puntos en una hilera que se incrementan o disminuyen en forma constante



(c) Dos de tres puntos en una hilera de la zona A



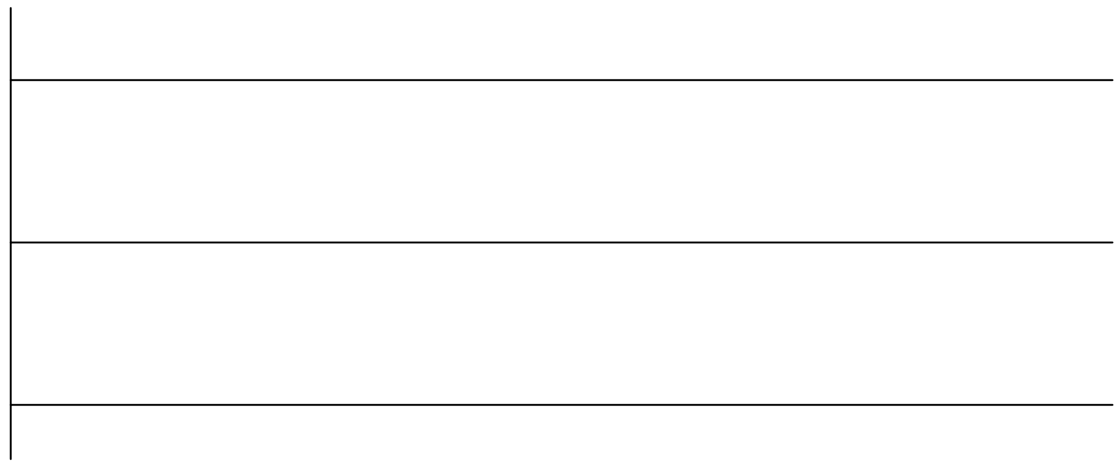
(d) Cuatro de cinco puntos en una hilera en la zona B o más allá

También se considera que un proceso está fuera de control incluso si los puntos están dentro de los límites de 3 desvíos. Ejemplo, no es natural que siete o más puntos estén por arriba o por debajo de la línea central (fig a). Lo mismo cuando seis o más puntos consecutivos son crecientes (fig b). Otro caso cuando dos de tres puntos caen en la zona A cerca del límite de 3 desvíos (fig c). Otro caso donde cuatro de cinco puntos en la zona B (fig d).

TRABAJO PRÁCTICO Nº 9

Elija una variable de las que definió en el trabajo práctico nº 1 y dibuje en la gráfica de control diez puntos indicando una zona bajo control y otra fuera de control. Además coloque que significa $\bar{X}$. Y que significa cada punto de la gráfica.

$\bar{X}$ =



Cada punto de la gráfica representa:

AUTOEVALUACIÓN MÓDULO 3

Nombre y apellido: ....
Fecha: / /

PROBLEMA 1
Qué tipo de causa de variación existe cuando la media de un subconjunto de datos supera los tres desvíos estándares respecto de la media de las medias?

a. Asignable
b. No aleatoria
c. No asignable
d. Aleatoria

Resp. a. □ b. □ c. □ d. □

PROBLEMA 2
Diga si la siguiente afirmaciones son verdaderas o falsas:

a. Un proceso es estadísticamente estable cuando no tiene variaciones asignables?

Resp. verdadero □ Falso □

b. Un proceso que produce dentro de los tres desvíos siempre es estable?
Resp. verdadero □ Falso □

Enviar antes de las 48 horas de recibido - muchas gracias -

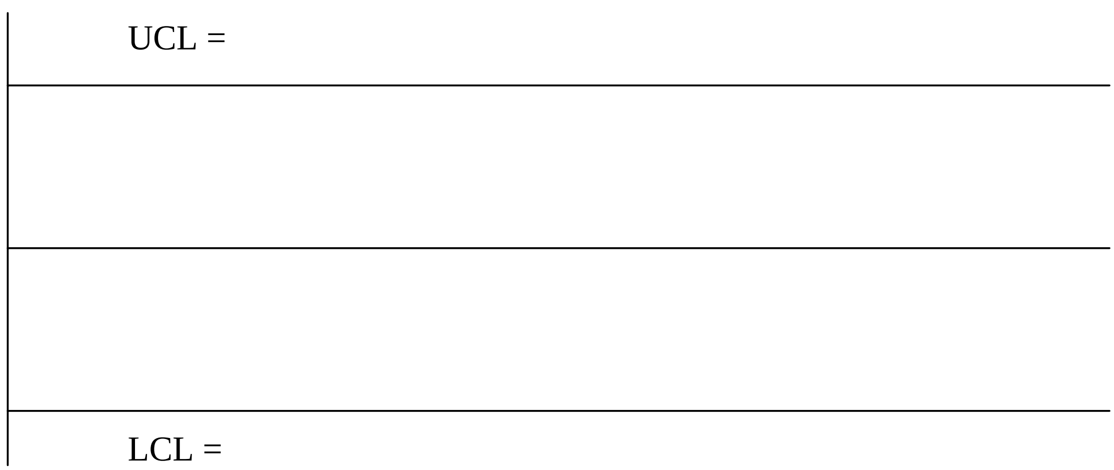
AUTOEVALUACIÓN MÓDULO 3 - continuación

PROBLEMA 3

En el laboratorio de control de calidad de una importante empresa elaboradora de productos alimenticios se utiliza el control estadístico periódico de los insumos. Uno de los controles consiste en sacar cinco muestras diarias de la línea de producción y efectuar una gráfica de control para verificar si el contenido de humedad se encuentra dentro de los límites aceptables según la norma ISO 8248, los datos botenidos son los siguientes :

Número de		Mediciones (% Hum)				Promedio	Desvio estándard
Subgrupo	Fecha	X1	X2	X3	X4		de la muestra
1	20/11	6,35	6,40	6,32	6,37	6,36	0,034
2	21/11	6,46	6,37	6,36	6,41	6,40	0,045
3	22/11	6,34	6,40	6,34	6,36	6,36	0,028
4	23/11	6,69	6,64	6,68	6,59	6,65	0,045
5	24/11	6,38	6,34	6,44	6,40	6,39	0,042

Pregunta. Coloque los límites de la gráfica de control.



REFERENCIAS BIBLIOGRAFICAS

- **Estadística para Administración y Economía**
  Stevenson, Editorial Harla
- **Muestreo**
  Morris J. Sloni M., Editorial Americana
- **Análisis y Planeación de la Calidad**
  Juran Gryna, Editorial Mc Graw Hill
- **Análisis Estadístico**
  Ya-Lun Chou, Editorial Mc Graw Hill
- **Control de Calidad**
  Besterfield, Editorial Prentic Hall
- **Introducción a la Inferencia Estadística**
  Castello Minolli, Editorial El Coloquio

ESTADÍSTICA APLICADA MÓDULO 4

Lic. OSVALDO CASTILLO

2009

E-mail: castillo@cqc.com.ar Internet: http://www.cqc.com.ar

INDICE

SERIES CRONOLÓGICAS

Definición

Una serie cronológica es un conjunto de observaciones (ordenado en términos de tiempo). Algunos ejemplos de series cronológicas serían aspectos tales como los registros de precipitación pluvial diaria, las ventas semanales, el producto bruto trimestral, etc.

Ej.

Años	Ventas (miles de toneladas)
1986	2
1987	5
1988	6
1989	8
1990	9
1991	12
1992	15
1993	21
1994	24
1995	23
1996	26

El objeto de analizar tales datos es determinar si se presentan ciertos patrones o pautas no aleatorias. Algunas veces se trata de descubrir patrones no aleatorios que se puedan utilizar para predecir el futuro. Por ejemplo, pronósticos de ventas es un caso en el que se analizan los datos del pasado, con la esperanza de encontrar algo que sea útil para predecir la demanda futura. En otras ocasiones, el objetivo es asegurarse d eque no haya patrones no aleatorios. En estos casos, dichos patrones son considerados como una señal de que un sistema está "fuera de control".

Por ejemplo, el departamento de control de calidad de una fábrica utiliza en gran medida datos de series cronológicas en la revisión y ajuste del proceso de fabricación. Una máquina que está funcionando de manera adecuada producirá piezas con dimensiones promedio estadísticamente independientes entre sí (es decir, no existirá una relación histórica entre las observaciones). En lugar de ello, los valores parecen ser observaciones al azar de alguna distribución de probabilidad, como la distribución normal. En tanto que las pruebas de significación de medias y proporciones son bastante útiles para evaluar las desviaciones respecto de una norma (los valores críticos de estas pruebas se consideran como "límites de control"), las pruebas de series cronológicas se concentran en los valores extremos. Si en los datos se observan algunas pautas que no sean al azar, esto se considerará como prueba de que la máquina o el proceso están fuera de control, y se deberán tomar medidas correctivas para regresar el proceso a un estado de control estadístico.

MODELO CLÁSICO

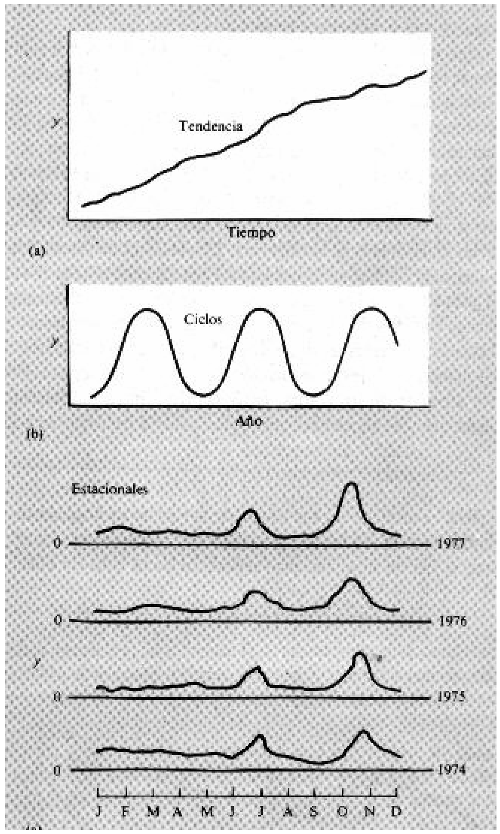
El modelo clásico o de descomposición, considera que los datos de series cronológicas están compuestos de cuatro patrones básicos:

1.- La tendencia (Tt).
2.- Las variaciones cíclicas (Ct).
3.- Las variaciones estacionales (Et).
4.- Las variaciones irregulares (It)

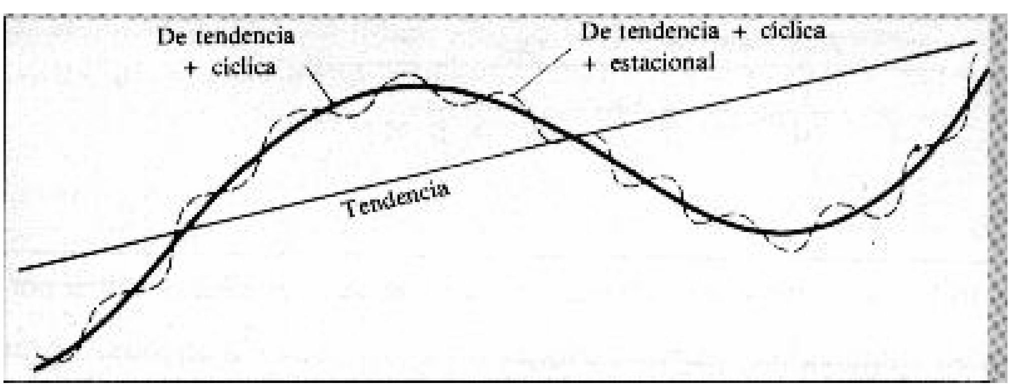
El término tendencia se refiere a un desplazamiento de los datos de modo uniforme y suave, a largo plazo, hacia arriba o hacia abajo. Las tendencias se pueden relacionar con cambios en la población, cambios en las preferencias del consumidor, etc.

Existe un patrón cíclico cuando las fluctuaciones muestran cierto grado de regularidad. Ejemplo demanda de productos duraderos, inventarios, precios de las acciones, la prosperidad, las manchas solares, la lluvia, poblaciones animales, etc.
Las variaciones estacionales son cíclicas y de plazo relativamente corto (un año o menos), las cuales se relacionan a menudo con los cambios estacionales (el clima). Por ejemplo, hay modelos estacionales en las ventas de artículos deportivos como esquís, trineos, lanchas, mallas, las tarjetas de felicitación, helados, libros de texto, la ropa, etc.

Por último, las variaciones irregulares o aleatorias, se componen de causas tales como desastres, huelgas, etc.



En el modelo clásico, el método consiste en descomponer una serie cronológica en cada uno de estos componentes básicos de variación, analizar cada componente en forma separada y combinar después las series a fin de describir las variaciones observadas en la variable en estudio. El proceso de descomposición comprende la separación sistemática de cada componente a partir de los datos, empezando con la tendencia.



En el modelo agregativo el valor de la variable en estudio para un momento t es

$$Y(t) = T(t) + E(t) + C(t) + I(t)$$

En el modelo multiplicativo

$$Y(t) = T(t) * E(t) * C(t) * I(t)$$

Aislamiento de la tendencia

El procedimiento que más se utiliza para adaptar una recta a un conjunto de puntos se conoce como método de los ***mínimos cuadrados.*** La recta resultante presenta dos características importantes: 1) es nula la suma de las desviaciones verticales de los puntos respecto de la recta y 2) es mínima la suma de los cuadrados de dichas desviaciones (es decir, ninguna otras recta daría una suma menor de las desviaciones elevadas al cuadrado).

$$Y(t) = a + b * t$$

Donde :

Y(t) es el valor de la variable dependiente de t (tiempo)
b es la pendiente de la recta o tendencia
a es la ordenada al origen

$$b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

$$a = \frac{\sum y - b\sum x}{n}$$

Aislamiento de la estacionalidad

Para esto se utiliza el método de los promedios móviles. Este método produce índices semanales, mensuales o trimestrales, que establecen observaciones de series cronológicas, en términos de porcentaje del total anual (es decir como relativos estacionales). Por ejemplo, si el mes de junio tiene un índice estacional de 0.80, esto indica que las ventas medias en junio son 80 % del promedio mensual. Si un trimestre presenta índice estacional de 2.00, esto quiere decir que las ventas para un trimestre son el doble de la cantidad promedio para todos los trimestres.

A continuación se presenta el desarrollo del método paso por paso.

Supongamos que tenemos datos mensuales de un fenómeno y simbolicemos con yi el valor mensual (variando yi de 1 a n, tomando n el último mes del que se posee información).

y1; y2; y3;……..y12; y13;……y24;….y36

Para hallar una media movil de un número impar de meses, por ejemplo una media movil de 3 meses hacemos

$$\frac{y1 + y2 + y3}{3}$$

O sea hallamos el promedio de los tr9es primeros meses del año; es fácil determinar a qué mes de los tr9es le asignamos el valor de este promedio. Simplemente al central (febrero), luego

$$\bar{y}2 = \frac{y1 + y2 + y3}{3}$$

la próxima media movil será:

$$\bar{y}3 = \frac{y2 + y3 + y4}{3}$$

y así sucesivamente

Si hacemos una media movil de base 5 tendremos:

$$\bar{y}3 = \frac{y1 + y2 + y3 + y4 + y5}{5}$$

$$\bar{y}4 = \frac{y2 + y3 + y4 + y5 + y6}{5}$$

y así sucesivamente

Cuando hacemos una media movil de base 3 quedaron sin ajustar el primer y último dato; al hacer una media movil d ebase 5 quedan sin ajustar el primero, segundo, penúltimo y último dato. Esto nos indica que a medida que hacemos un promedio movil con mayor cantidad de períodos, perdemos más datos ajustados al principio y final de la serie.

El problema de adjudicar un promedio móvil a un determinado período se plantea cuando se toma un número par de períodos. Supongamos que se quiere hallar una media móvil de 12 meses:

$$\bar{y}_{6,7} = \frac{y_1 + y_2 + y_3 + y_4 + y_5 + y_6 + y_7 + y_8 + y_9 + y_{10} + y_{11} + y_{12}}{12}$$

La primera media móvil debemos adjudicarla al mes que está entre el 6º y 7º (15 de julio).

$$\bar{y}_{7,8} = \frac{y_2 + y_3 + y_4 + y_5 + y_6 + y_7 + y_8 + y_9 + y_{10} + y_{11} + y_{12} + y_{13}}{12}$$

El inconveniente que se plantea al tener un valor ajustado que no corresponde exactamente al mismo período es que no permite comparar el valor ajustado con el dato original o empírico.

Surge entonces la necesidad de usar una media movil centrada, que es un promedio móvil desplazado de forma tal que se puede adjudicar al período correspondiente.

Con datos mensuales centrar la media móvil significa correrla medio mes hacia delante para que se la pueda asignar a un mes determinado y permita entonces la comparación con el dato original.

Para determinar la media movil centrada cuando se toma una base igual a 12 (por estar los datos expresados en forma mensual) debemos tomas la mitad del valor correspondiente al primer mes que interviene en el cálculo y la mitad del mismo mes del año siguiente

El cálculo que a primera vista parece engorroso se simplifica bastante puesto que cualquier media móvil centrada puede ser obtenida como promedio de dos medios móviles no centrados sucesivos, por ejemplo:

$$\bar{y}_{10} = \frac{\bar{y}_{9,10} + \bar{y}_{10,11}}{2}$$

Un ejemplo facilitará la explicación del método de promedios móviles

Período	yi	sumas móviles de base b	sumas de 2 sumas Móviles sucesivas	Promedios móviles $\bar{y}$	Promedios Estacionales Relativos
Meses Trimestres Etc	y1 y2 . . . yn	b= 12 si los datos están expresados en meses b= 4 si los datos están expresados en trimestres		se determinan dividiendo la suma de 2 sumas moviles sucesivas por 2b	Se obtiene haciendo $\frac{yi}{\bar{yi}} \cdot 100$

Ej.

Año	Trimestres	Ventas (miles de unidades)
1994	Iº	101
	IIº	103
	IIIº	98
	IVº	98
1995	Iº	105
	IIº	112
	IIIº	111
	IVº	113
1996	Iº	115
	IIº	126
	IIIº	121
	IVº	138

Trimestre	Datos (Yi)	Suma móvil de BASE 4 (*)	Suma de 2 sumas móviles sucesivas	Promedio móvil de 2 medias móviles (B)	Promedios estacionales relativos
I	101				
II	103	400			
III	98	404	804	100,5	(98/100,5)*100=97,51
IV	98	413	817	102,125	(98/102,125)*100=95,96
I	105	426	839	104,875	(105/104,875)*100=100,12
II	112	441	867	108,375	(112/108,375)*100=103,34
III	111	451	892	111,5	(111/111,5)*100=99,55
IV	113	465	916	114,5	(113/114,5)*100=98,69
I	115	475	940	117,5	(115/117,5)*100=97,87
II	126	500	975	121,875	(126/121,875)*100=103,38
III	121				
IV	138				

(*) Deberían adjudicarse al punto medio entre 2 trimestres, p ejemplo 400 correspondería adjudicarlo al punto medio entre el IIº y el IIIº trimestre (por comodidad en la lectura se asigna de la forma que se hizo).

Ahora realizamos un cuadro con los promedios estacionales relativos.

Trimestres / Años	Iº	IIº	IIIº	IVº
1994	-	-	97,51	95,96
1995	100,12	103,34	99,55	98,61
1996	97,87	103,38	-	-
Sumas	197,99	206,72	197,06	194,65
Promedio	98,995	103,36	98,53	97,325

98,995
103,360
98,530
97,325

398,210 / 4 = 99,5525

El índice estacional para cada período se calcula:

$$\text{IE er.} = \frac{98{,}995}{99{,}5525} = 0{,}99439993973$$

$$\text{IE do.} = \frac{103{,}360}{99{,}5525} = 1{,}03824615152$$

$$\text{IE er.} = \frac{98{,}530}{99{,}5525} = 0{,}98972903744$$

$$\text{IE to.} = \frac{97{,}325}{99{,}5525} = 0{,}97762487129$$

4

Una forma de verificar que los índices estacionales están bien hallados es que su suma da siempre el número de períodos en que venía dividido el año. En este caso 4 (trimestre)

AUTOEVALUACIÓN MÓDULO 4

Nombre y apellido: ....

Fecha: / /

PROBLEMA 1

En una serie cronológica, indique cuatro causas probables de variacion aleatoria

....

PROBLEMA 2

A su criterio, ¿qué serie cronológica tiene mayor tendencia?.

a. y = 0,45 m/hora * (t) + 34 m
b. y = 0,50 puntos/trimestre * (t) + 5,5 puntos
c. y = 0,15 u$s/día * (t) + 20 u$s

Resp. a. ☐ b. ☐ c. ☐

PROBLEMA 3

¿qué valor tendrá la variable ***evaluaciones promedio*** representada en el punto b del problema 4, en el trimestre 5, utilizando para ello los índices estacionales determinados en el trabajo practico del módulo 3 ?.

Resp. ☐

REFERENCIAS BIBLIOGRAFICAS

- **Estadística para Administración y Economía**
  Stevenson, Editorial Harla
- **Muestreo**
  Morris J. Sloni M., Editorial Americana
- **Análisis y Planeación de la Calidad**
  Juran Gryna, Editorial Mc Graw Hill
- **Análisis Estadístico**
  Ya-Lun Chou, Editorial Mc Graw Hill
- **Control de Calidad**
  Besterfield, Editorial Prentic Hall
- **Introducción a la Inferencia Estadística**
  Castello Minolli, Editorial El Coloquio