INTRODUCCIÓN A LA ESTADÍSTICA

DIRECCIÓN DE LA PRODUCCIÓN

Por:

- LUIS ARENCIBIA SÁNCHEZ

Índice.

1. Control estadístico de calidad.

Los conocimientos que proporciona la Estadística permiten analizar los datos que se obtienen de un suceso, experimento o prueba e interpretarlos adecuadamente. Al mismo tiempo permitirá deducir una mayor información a partir de los datos existentes, enfocándolos hacia el objetivo que se persiga en cada caso.

Toman una "muestra" de individuos de una "población", obtienen sus opiniones (datos) y de ellos deducen o infieren el "probable" comportamiento de esa población.

La estadística se puede clasificar, en función de los objetivos a conseguir, en: **descriptiva** e **inductiva.**

- La ***Estadística Descriptiva*** consiste en el conjunto de instrumentos y de técnicas relacionados con la descripción de las características de una población.
- La ***Estadística Inductiva*** se ocupa de la lógica y de los procedimientos para la obtención de información o inducción de las propiedades de una población teniendo en cuenta los resultados obtenidos de una muestra representativa de dicha población.

En la estadística aplicada a la función de calidad tiene una relevancia especial, ya que la mayor parte de las decisiones se toman basadas en la recopilación, el análisis y la interpretación de los datos obtenidos de una muestra de la población. Por tanto, la Estadística se convierte en un instrumento de ayuda en la resolución de problemas.

En un proceso productivo la perfección es imposible debido al gran número de partes incontrolables que influyen en el mismo y que determinan una cierta variabilidad en los resultados. Por ejemplo, la diferente dureza del material utilizado, la incertidumbre de los aparatos de medida, el factor humano, etc. hacen que el diámetro interior de unos cojinetes fabricados por una maquina oscilen dentro de ciertos límites, por cuidadoso que sea el operario que la maneje. Esto lleva implícito la aceptación de unas ***"tolerancias de fabricación"***, es decir, unos límites entre los cuales puede variar la característica considerada sin perjudicar la utilización del producto.

Aún con la existencia de tolerancias, es imposible evitar que en una producción en serie aparezcan unidades fuera de los límites marcados por dichas tolerancias.

Podría pensarse que mediante una inspección 100% del producto terminado evitaríamos la aparición de defectuosos. Sin embargo, este procedimiento puede resultar inviable por su coste o por la destrucción del producto.

Este obstáculo conduce a admitir que en todo proceso de fabricación (o en todo lote o partida) pueden aparecer un número de unidades defectuosas y a intentar controlar la calidad mediante el examen de unas pocas unidades del producto, es decir, de una muestra.

Al conjunto de técnicas que pretenden obtener conclusiones respecto a la calidad de un producto mediante la extracción y posterior análisis de muestras del mismo es a lo que se llama **"Control Estadístico de Calidad"**.

2. Datos.

- **Fenómenos aleatorios**

En la vida real existen procesos cuyo resultado final puede predecirse con exactitud si se conocen las condiciones en que se desarrollan. Por ejemplo, si calentamos agua pura manteniendo la presión a una atmósfera comenzará a hervir cuando llegue a una temperatura de 100 º C. Este tipo de fenómenos se denominan ***determinísticos.***

Por el contrario, existen otros en los cuales no se puede predecir con exactitud su resultado aunque podamos afirmar algo respecto a la frecuencia con que se presentan éstos. Por ejemplo, si en un lote de tornillos que contiene algunas unidades defectuosas se elige uno al azar, no sabremos hasta haberlo comprobado si es defectuoso o no lo es; si arrojamos una moneda al aire no sabremos predecir si saldrá cara o cruz.

Ahora bien, lo que sí sabemos es que si arrojamos una moneda un gran número de veces, aproximadamente la mitad se obtendría cara y la otra mitad cruz. Y cuanto mayor sea el número de lanzamientos, más próxima a 0,5 será la relación de caras o cruces obtenidas respecto al número de lanzamientos efectuados.

Este tipo de fenómenos caracterizados por la impredecibilidad de sus resultados y por la tendencia a la estabilidad de la frecuencia con que ocurre cada uno de ellos se conocen como ***aleatorios***.

El hecho de que todos los procesos productivos sean de carácter aleatorio justifica la existencia del Control Estadístico de la Calidad.

- **Tipos de datos.**

En el control estadístico de la calidad vamos a manejar datos. Datos que son, por lo general, resultados de la observación de las características del producto o servicio en las que estemos interesados.

Este resultado será función de la naturaleza de la característica: longitud en milímetros de una pieza, peso en gramos de una bolsa de infusión o en miligramos de dosificación de un medicamento, tiempo medio en minutos de espera en la caja de un gran almacén o en horas de gestión de un documento en una entidad bancaria, número de defectos por metro cuadrado de moqueta o por cada 100 metros de cable eléctrico, el elemento a considerar es defectuoso o no, etc.

El tratamiento que daremos a estos datos, será distinto según sea su naturaleza:

Atributos: Características cualitativas que sólo pueden evaluarse en casos de conformidad o disconformidad con un criterio predeterminado (pasa - no pasa, bueno

1.2. Presentación de datos.

1.2.1. Series estadísticas.

Los datos procedentes de una muestra o experimento se presentan de forma desordenada. Si ordenamos estos datos obtendremos lo que se llama una ***serie estadística***, que es un primer paso, para conocer la evolución o comportamiento del proceso que se controla.

La forma de registrar y contabilizar (ordenar) los valores obtenidos en un aspecto muy importante para poder obtener condiciones sobre la característica medida, como veremos a continuación.

- ***Ejemplo:***

Se han pesado 50 unidades de una misma pieza y se han obtenido los valores siguientes expresados en gramos:

5,45	*5,80*	*5,71*	*5,90*	*5,80*	*5,59*	*5,69*	*5,90*	*5,90*	*5,70*
6,10	*5,80*	*6,00*	*5,80*	*5,65*	*5,80*	*5,60*	*5,90*	*5,80*	*5,80*
5,61	*5,70*	*5,80*	*6,15*	*5,57*	*6,00*	*5,80*	*5,73*	*5,90*	*5,68*
6,00	*5,90*	*5,64*	*5,90*	*5,80*	*5,67*	*5,60*	*5,80*	*5,80*	*6,00*
5,74	*5,80*	*5,71*	*5,80*	*5,62*	*5,90*	*5,80*	*6,00*	*5,90*	*5,69*

La presentación de estos datos apenas dice nada sobre la pieza; puede pensarse que el campo de variación de los datos está entre 5'50 y 6'15.

Si ahora ordenamos estos datos de menor a mayor y los agrupamos en pequeños grupos dividiendo la variación total (5,45 - 6,15) en intervalos (en este caso siete) de igual magnitud obtenemos lo que se denomina una serie estadística.

Intervalo	**Valores**	***Intervalo***	***Valores***	***Intervalo***	***Valores***
5,45 - 5,55	5,50	5,75 - 5,85	5,76 5,76	5,85 - 5,95	5,85 5,85
5,55 - 5,65	5,57 5,59 5,60 5,60 5,61 5,62 5,64		5,78 5,79 5,79 5,79 5,80 5,80 5,81		5,88 5,89 5,90 5,90 5,92 5,93 5,93
5,65 - 5,75	5,65 5,67 5,68 5,69 5,69 5,70 5,70 5,71 5,71 5,73 5,74		5,81 5,82 5,82 5,82 5,84 5,84	5,95 - 6,05	5,98 5,99 6,00 6,00 6,02
				6,05 - 6,15	6,10 6,15

En esta serie ya podemos comprobar que el mayor número de datos corresponden al intervalo (5'75 - 5'85) y que esté número de datos va descendiendo hacia el máximo 6'15 y hacia el mínimo 5'45.

Para determinar la serie estadística de un modo simple se utilizan unos conceptos que definimos a continuación:

- ***Intervalos de clase.***

  Cada una de las partes en que divididos el campo de variación total de los datos obtenidos. Como regla general, los valores extremos de un intervalo se agrupan en el intervalo mayor siguiente.

♦ ***Ejemplo:***

El valor 5,65 podría pertenecer a los intervalos 5,55 - 5,65 y 5,65 - 5,75. De acuerdo con la regla anterior le hemos agrupado en el intervalo 5,65 - 5,75.

- ***Longitud del intervalo.***

  Es la diferencia entre los valores máximo y mínimo de cada intervalo.

♦ ***Ejemplo:***

Para el intervalo 5,45 - 5,55

Longitud del intervalo = 5,55 - 5,45 = 0,1 g.

En el ejemplo (ver tabla anterior) se han fijado, como suele ser más usual, todos los intervalos de igual longitud.

- ***Marca de clase de un intervalo.***

  Es valor del punto medio del intervalo.

  En estadística, cuando los datos se encuentran agrupados en intervalos, se realiza la aproximación de que todos los datos pertenecientes a un mismo intervalo tienen el mismo valor, igual a la marca de clase de dicho intervalo.

♦ ***Ejemplo:***

Intervalo 5,95 - 6,05

Marca de clase: 6,00

Aproximación : las cinco piezas de pesos 5,98, 5,99, 6,00, 6,00 y 6,02 g suponemos que pesan todas 6,00 g.

- ***Frecuencia absoluta y relativa.***

  - **Frecuencia absoluta:** La frecuencia absoluta o simplemente la frecuencia de un intervalo, es el número de datos que se ubican en dicho intervalo.
  - **Frecuencia acumulada:** La frecuencia acumulada hasta un determinado valor es el número total de datos con valores iguales o inferiores al valor considerado.
  - **Frecuencia relativa:** La frecuencia relativa de un intervalo es el resultado de dividir la frecuencia absoluta de dicho intervalo por el número total de datos del

experimento. Indica la proporción de datos que hay en dicho intervalo en relación al total de datos.

- ***Ejemplo*:**

  En el intervalo 5,95 - 6,05

  Frecuencia absoluta = 5

  Número total de datos = 50

  Frecuencia relativa = 5 / 50 = 0,1

  - ***Distribución de frecuencias.***

    Es la relación expresada conjuntamente entre unos valores y sus frecuencias absolutas o relativas correspondientes.

2.2. Técnicas de presentación de datos.

Las técnicas de presentación de datos varían según el método utilizado: tabular o gráfico.

- **Tabular.**

  Son tablas en las que se representan los valores de una característica y el número de veces que estos datos se dan.

  La tabla de frecuencias es la representación de la relación entre los valores de la característica controlada, agrupada en sus correspondientes intervalos o marcas de clase, y sus correspondientes frecuencias.

  Los pasos para construir una tabla de frecuencias son los siguientes:

  1. ***Decidir el número de intervalos.***

     Lo habitual es no utilizar nunca un número de intervalos inferior a 6 y no superior a 20. Se recomienda elegir los valores de acuerdo a la tabla siguiente.

Nº de observaciones	***Nº de intervalos***
20 - 50	6
51 - 100	7
101 - 200	8
201 - 500	9
501 - 1000	10
Más de 1000	11 - 20

  2. ***Calcular el intervalo de clase*** (observación mayor, menos observación menor, dividiendo el resultado por el número de clases y redondeando el resultado).
  3. ***Construir las clases***, indicando los extremos de las mismas y teniendo en cuenta que:
     - Los extremos de clase tendrán un decimal más que los datos reales y terminarán en 5.
     - El intervalo de clase debe ser constante para toda la distribución.
  4. ***Marcar cada observación en la clase que le corresponda y, a continuación determinar la frecuencia de cada clase.***

♦ ***Ejemplo:***

La distribución de frecuencias del experimento que venimos analizando es:

Tabla de Frecuencias

Intervalos	***Marca de Clase***	**Tabulación**	***Frecuencia***	***Frecuencia Acumulada***	***Frecuencia Relativa***
5,45 - 5,55	5,50	I	1	1	0,02
5,55 - 5,65	5,60	IIIII II	7	8	0,14
5,65 - 5,75	5,70	IIIII IIIII I	11	19	0,22
5,75 - 5,85	5,80	IIIII IIIII IIIII	15	34	0,30
5,85 - 5,95	5,90	IIIII IIII	9	43	0,18
5,95 - 6,05	6,00	IIIII	5	48	0,10
6,05 - 6,15	6,10	II	2	50	0,04
TOTAL			50		1,00

Los datos ordenados de esta manera proporcionan mayor información sobre la muestra tomada, indicando claramente que la mayoría de los datos se encuentran alrededor de 5'80 y que el número de datos o frecuencia disminuye al alejarse de ese valor hacia ambos extremos y de forma aproximadamente sistemática.

- **Gráfica.**

  Las representaciones gráficas permiten obtener de un vistazo una idea general de la forma y, por tanto, de algunas de sus características de una distribución de frecuencias.

  - ***A. Histograma y Diagrama de Barras.***

  El histograma y el diagrama de barras son las representaciones gráficas más habituales de las distribuciones de frecuencia.

  El ***histograma*** se construye tomando en abcisas los extremos de los intervalos de la clase y levantando sobre cada uno de ellos un rectángulo, cuya base es el intervalo y cuya altura es, de acuerdo con una determinada escala, la frecuencia de cada intervalo.

  Denominamos histograma de frecuencias a la figura formada por el contorno exterior de estos rectángulos.

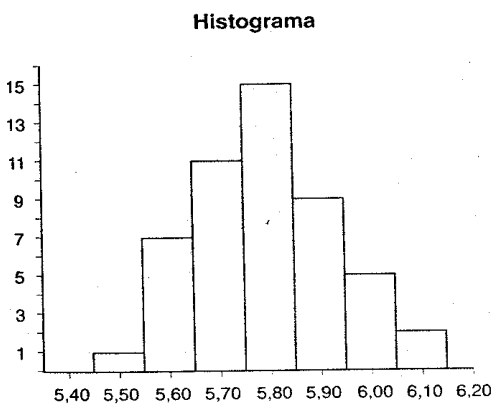
  El ***diagrama de barras*** se construye de manera similar, levantando una barra sobre cada una de las marcas de clase, cuya altura es proporcional a la correspondiente frecuencia según una determinada escala.

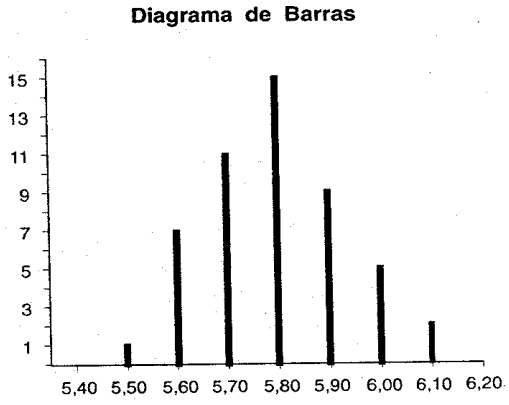
♦ ***Ejemplo*:**

El Histograma y el Diagrama de barras del ejemplo que vemos analizando es:

Intervalos	*Frecuencia*
5,45 - 5,55	1
5,55 - 5,65	7
5,65 - 5,75	11
5,75 - 5,85	15
5,85 - 5,95	9
5,95 - 6,05	5
6,05 - 6,15	2

Marcas de Clase	*Frecuencia*
5,50	1
5,60	7
5,70	11
5,80	15
5,90	9
6,00	5
6,10	2





En general, para analizar un histograma o un diagrama de barras y extraer conclusiones que puedan ser representativas, se necesita disponer de 50 observaciones como mínimo.

El histograma es una herramienta sencilla pero eficaz para un primer análisis de los datos e identificación de problemas ya que:

- Cualquier proceso que se mida presenta siempre una variación. Esta variación se debe a innumerables pequeños factores que continuamente están afectando al proceso, bien sea un proceso de fabricación, de servicios o administrativo. La variación es inevitable.
- Esta variación muestra siempre un patrón determinado, que se representa por una distribución de frecuencias.
- Los patrones de variación o distribuciones son difíciles de ver con simples tablas de números. Son más fáciles de ver cuando los datos se representan gráficamente.

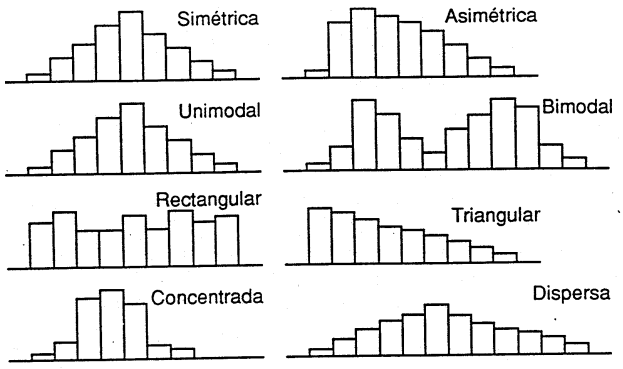
Para ello es necesario analizar tres aspectos del mismo:

- *situación del centro del histograma,*
- *el ancho del mismo,*
- *su forma.*

Del último, es decir de la "*forma del histograma*", nos vamos a servir para intentar conocer alguna característica del proceso que ha dado lugar a los datos que estamos estudiando, por ejemplo:

- ***Distribuciones simétricas unimodales*** - Son el ejemplo típico de la mayoría de los procesos industriales, y se caracterizan porque las observaciones equidistantes del máximo central tienen aproximadamente la misma frecuencia.
- ***Asimétrica*** - Distribuciones típicas de datos económicos, distribuciones de renta, tamaño empresas etc.
- ***Triangular*** - Distribución típica de fabricaciones con imposibilidad física de superar un valor o bien sometidas a una selección 100% en una de las características.
- ***Bimodal*** - Suele representar conjuntos de valores obtenidos a partir de dos procesos distintos (característica de una pieza suministrada por dos fabricantes y mezcladas en un mismo contenedor, mezcla de la fabricación de los elementos iguales por dos máquinas distintas, distintos turnos de operarios, etc.)
- ***Rectangular o uniforme*** - En el caso de que la mezcla de productos de distintas fabricaciones que hemos visto en la distribución bimodal fuera de más de dos, llegaría a dar lugar a una distribución con forma uniforme.
- ***Concentrada o truncada*** - Muestra poblaciones obtenidas de procesos no capaces de cumplir las especificaciones sobrepasando tanto el límite superior como el inferior y que han sido seleccionadas. También puede ser síntoma de una mala elección del número de clases (menor de lo adecuado).

En la siguiente figura se muestran estos diferentes tipos de formas.



- ***B. Polígono de Frecuencias.***

Un polígono de frecuencias es un gráfico de línea trazado sobre las marcas de clase. Puede obtenerse uniendo los puntos medios de los lados menores superiores de los rectángulos del histograma.

- ***C. Polígono de Frecuencias Acumuladas***

Es un gráfico de línea que representa la frecuencia acumulada para cada intervalo, representado por su marca de clase.

3. Estadísticos.

3.1. Introducción.

La observación visual de las representaciones gráficas de las distribuciones de frecuencias es sin duda alguna, un método elemental y aproximado para el análisis de sus propiedades. Por otro lado, en algunos casos donde el número de los datos es muy elevado puede ser necesario reducir los resultados todavía más, incluso a uno sólo.

Si analizamos la Tabla de Distribución de Frecuencias del ejemplo anterior, vemos en primer lugar, que el peso se distribuye entre 5,5 gramos (marca de clase intervalo más bajo) y 6,1 gramos (marca de clase del intervalo más alto), con una diferencia entre uno y otro de 0,6 gramos, que proporciona una idea de la dispersión existente.

Se observa también que los datos se distribuyen alrededor de un valor central (5,8), que es el que presenta una frecuencia mayor. Conforme los valores se alejan de el, la frecuencia de los mismos disminuye. Es, por tanto, una referencia de la tendencia central.

Por ello, para estudiar las características más sobresalientes de las distribuciones de frecuencia, se utilizan las medidas de tendencia central y las medidas de dispersión. A estas medidas, que estudiaremos a continuación, se las denomina:

- ***Parámetros*** - cuando están calculados de todo el colectivo o población.
- ***Estadísticos*** (o estimadores) - cuando están calculados a partir de una muestra de la misma.

3.2. Medidas de la tendencia central.

La posición o tendencia central de una distribución se refiere al lugar donde se concentra una mayor cantidad de valores.

La medidas de la tendencia central más utilizadas son tres: la media aritmética, la mediana y la moda. Existen otras medidas de la tendencia central que son apropiadas para situaciones especiales que, sin embargo, no son de uso común en la estadística utilizada en la gestión de calidad.

a. Media aritmética.

La media aritmética de una serie de valores se obtiene sumando esos valores y dividiendo su suma por el número de valores (también se la conoce como Promedio).

Se representa por una $\bar{x}$ (equis barra)

$$\bar{x} = \frac{\text{suma de valores}}{\text{nº de valores}} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5} = \frac{\sum_i^n x_i}{n}$$

$\sum_i^n x_i$ significa: suma de los "n" valores $x_1, x_2, \ldots, x_n$

En el caso de contar con un número elevado de datos y/o que estos se presenten como tabla de frecuencias, se puede suponer que todos los datos de un intervalo tienen un mismo valor que es igual a la marca de clase.

$$\bar{x} = \frac{\text{suma valores}}{\text{nº valores}} = \frac{x_1 f_1 + - + x_n f_n}{f_1 + - + f_n} = \frac{\sum_{i=1}^n x_i f_i}{\sum_{i=1}^n f_i}$$

donde x_i es la marca de clase del intervalo i y f_i es el número de valores del intervalo y.

También, teniendo en cuenta que la frecuencia relativa f_i , es $f_i = \frac{n_i}{n}$; donde n_i es la frecuencia absoluta y, n es el número total de valores observados: la media aritmética puede calcularse como:

$$\bar{x} = \frac{x_1 n_1}{n} + \frac{x_2 n_2}{n} + \ldots + \frac{x_n n_n}{n} = x_1 f_1 + x_2 f_2 + \ldots + x_n f_n$$

♦ ***Ejemplo:***

Calcular la media aritmética de los valores obtenidos en la medida de la longitud de una tabla de madera expresadas en metros:

1,86	1,89	1,92	1,93	1,90	1,86	1,84	1,82	1,88	1,87
1,86	1,81	1,85	1,88	1,88	1,90	1,88	1,90	1,87	1,89
1,89	1,88	1,88	1,85	1,86	1,84	1,86	1,88	1,92	1,91
1,87	1,90	1,86	1,85	1,80	1,87	1,91	1,83	1,88	1,82
1,85	1,84	1,83	1,82	1,87	1,90	1,87	1,86	1,83	1,88
1,91	1,82	1,85	1,86	1,82	1,92	1,89	1,90	1,87	1,86
1,87	1,89	1,91	1,89	1,91	1,83	1,84	1,84	1,85	1,89
1,92	1,88	1,87	1,85	1,84	1,86	1,87	1,82	1,81	1,85

Estos valores se pueden agrupar como sigue:

Intervalo	**Marca de clase x**	**Frecuencia absoluta f**	**Frecuencia relativa**	**f.x**
1,795 - 1,815	1,805	3	3	5,415
1,815 - 1,835	1,825	10	13	18,25
1,835 - 1,855	1,845	14	27	25,83
1,855 - 1,875	1,865	20	47	37,3
1,875 - 1,895	1,885	17	64	23,045
1,895 - 1,915	1,905	11	75	20,955
1,915 - 1,935	1,925	5	80	9,625
		$\sum f = 80$		$\sum f.x = 149{,}42$

Y la media se puede calcular como:

$$\bar{x} = \frac{\text{suma de los 80 valores}}{80} = \frac{149{,}39}{80} = 1{,}8674$$

$\bar{x}$ = 1,8674

La aproximación de considerar que los valores de un intervalo tienen un mismo valor e igual a su marca de clase es suficiente a efectos de cálculo.

$$\bar{x} = \frac{5{,}415 + 18{,}25 + 25{,}83 + 37{,}3 + 32{,}045 + 20{,}955 + 9{,}625}{3 + 10 + 14 + 20 + 17 + 11 + 5} = \frac{\sum fx}{\sum f} = \frac{149{,}42}{80} = 1{,}8677$$ El

error respecto al calculado con los valores reales es del 0'0003, que no es significativo.

b. Mediana.

La mediana de una distribución es el punto o valor numérico que deja por debajo (y por encima) a la mitad de los valores de dicha distribución.

Si se ordenan las variables estadísticas en orden de magnitud, la mediana es el valor situado en el medio.

Cuando el número de datos es muy elevado, la mediana se calcula fácilmente si los tenemos representados en una tabla de frecuencias, ya que corresponde al valor cuya frecuencia acumulada contiene la mitad del número de datos estadísticos de la muestra y su cálculo es más preciso utilizando la distribución de frecuencias relativas acumuladas.

En general, la mediana se usa para reducir el efecto de los valores extremos o para datos que pueden ordenarse, pero que no se pueden medir en términos numéricos, tales como las tonalidades de color, la apariencia visual o los olores.

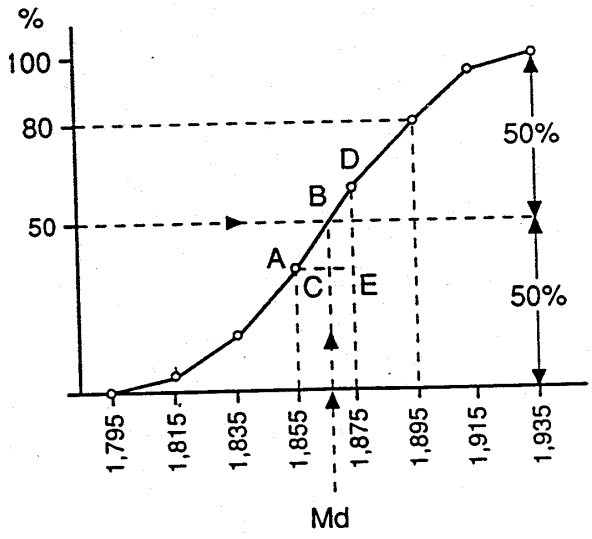
- ***Ejemplo*:**

El valor de la mediana se puede obtener de este modo.

Marca de Clase		frecuencia	
1,805		3	
1,825		10	39 datos
1,845		14	
		12	
1,865	20	2	Mediana = 1,865
		6	
1,885		17	
1,905		11	39 datos
1,925		5	

Los valores 13º y 14º de la marca de clase 1,865 dividen la distribución de frecuencias en 2 grupos de 39 datos cada uno. La mediana será ½ (1,865+1,865)=1,865.

El cálculo más preciso de la mediana se obtiene utilizando la distribución de frecuencias relativas acumuladas.



c. Moda.

En una distribución de frecuencias de tipo discreto, la moda es el valor de la variable al que corresponde la mayor frecuencia, es decir el que más se repite.

La moda se usa para distribuciones muy sesgadas, para eliminar el efecto de los valores o para describir una situación irregular en la que aparecen dos picos.

♦ ***Ejemplo*:**

En este caso consideramos que la moda es la "marca de clase" correspondiente al intervalo en el que sitúan más datos (o sea el intervalo que "más se repite").

Marca de Clase	frecuencia	
1,805	3	
1,825	10	
1,845	14	
1,865	20	Moda = 1,865
1,885	17	
1,905	11	
1,925	5	

En esta distribución de frecuencias la moda sería 1,865 puesto que a este valor le corresponde mayor número de datos (20).

3.3. Medidas de dispersión.

Las medidas de dispersión determinan el grado de variabilidad o de concentración de los valores de la característica estudiada en torno al valor central.

Las medidas de dispersión más comunes son el recorrido, la varianza y la desviación típica.

a. Recorrido.

Es la diferencia entre el valor más alto y el más bajo de la característica estudiada.

Dado que el recorrido se basa sólo en dos valores, es útil únicamente cuando el número de observaciones es pequeño.

Cuando los datos se dan en forma de tabla de frecuencias entre las marcas de clase máxima y mínima, el recorrido se calcula como la diferencia entre ambas.

♦ ***Ejemplo:***

Marca de clase máxima = 1,925 m.

Marca de clase mínima = 1,805 m.

Recorrido = 1,925 - 1,805 = 0,120 **R=0,120 m.**

b. Varianza.

Se define como la suma de los cuadrados de las distancias de cada valor observando a la media y dividido por el número de observaciones.

Se denomina σ^2 o S^2, según se trate de la varianza de toda la población o de la muestra. La formula es:

$$s^2 = \frac{(x_1 - m)^2 + (x_2 - m)^2 + \ldots + (x_n - m)^2}{N} = \frac{\sum (x_i - m)^2}{N}$$ (varianza de la poblacion)

donde x_1, x_2 ….x_n son los valores observados, μ es la media de la población y N es el número de valores de la población.

$$S^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1} = \frac{\sum (x_i - \bar{x})^2}{n-1}$$ (varianza de una muestra)

donde x_1, x_2 ….x_n son los valores observados, $\bar{x}$ es la media de la muestra y n es el número de valores de la muestra.

Valores altos de la varianza indican que los valores de las observaciones están muy dispersos. Valores bajos de la varianza indican que los valores de las observaciones están más concentrados alrededor de la media.

c. Desviación Típica.

Un inconveniente de la varianza es que las unidades en que viene expresada no coinciden con las unidades de los valores observados, por lo que la medida de dispersión más usada es la desviación típica, que se obtiene extrayendo la raíz cuadrada de la varianza.
Por tanto la desviación típica representa también una medida de la desviación o "dispersión" de un grupo de valores alrededor de la media y viene expresada por la formula:

$$s=\sqrt{\frac{(x_1-m)^2+(x_2-m)^2+....+(x_n-m)^2}{N}}=\sqrt{\frac{\sum(x_i-m)^2}{N}}$$ (para el total de la poblacion)

donde x_i son los valores de la población, N es el número total de elementos de la población y μ es la media aritmética de la población.

$$S=\sqrt{\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+....+(x_n+\bar{x})^2}{n-1}}=\sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}$$(para muestra de la poblacion)

donde x_i son los valores de la muestra, n es el número de elementos de la muestra y $\bar{x}$ es la media aritmética de la muestra.

Cuando los datos se dan en forma de tabla de frecuencias:

$$s=\sqrt{\frac{f_1(x_1-\bar{x})^2+f_2(x_2-\bar{x})^2+....+f_n(x_n-\bar{x})^2}{f_1+f_2+....+f_n}}=\sqrt{\frac{\sum f_i(x_i-\bar{x})^2}{\sum f_i}}$$(para el total de la poblacion)

donde x_i es la marca de clase del intervalo f_i es la frecuencia en el intervalo i y $\bar{x}$ es la media aritmética de la población.

Existe también una forma más sencilla para estimar la desviación típica a partir de los recorridos. Para ello, se deben seguir los siguientes pasos:

1. Suponer que los datos se han obtenido de forma aleatoria.
2. Agrupar los datos en subgrupos de tamaño n.
3. Calcular el recorrido de cada subgrupo.
4. Calcular el recorrido medio.
5. Obtener el valor d_2 de la tabla siguiente:

Constante d_2									
n	2	3	4	5	6	7	8	9	10
d_2	1,128	1,693	2,059	2,326	2,534	2,704	2,847	2,970	3,078

6. Calcular la desviación típica $S=\frac{\bar{R}}{d_2}$

- ***Ejemplo:***

Marca de Clase x	Media $\bar{x}$	x - $\bar{x}$	(x - $\bar{x}$)^2	frac. f	f (x - $\bar{x}$)^2
1,805	|	-0,627	0,003931	3	0,011793
1,825	|	-0,0427	0,001823	10	0,018231
1,845	|	-0,0227	0,000515	14	0,007213
1,865	1,8677	0,0027	0,000001	20	0,000144
1,885	|	0,0173	0,000299	17	0,005083
1,905	|	0,0373	0,001391	11	0,015301
1,925	|	0,0573	0,003283	5	0,016425

$$\sum f = 80 \quad \sum_{i}^{m} f\left(x - \bar{x}\right)^2 = 0'074190$$

$$s = \sqrt{\frac{\sum f_i\left(x_i - \bar{x}\right)^2}{\sum f_i}} = \sqrt{\frac{0'07418}{80}} = 0'3045$$

4. Distribuciones de probabilidad.

4.1. Introducción.

Cuando hablábamos en el apartado de los fenómenos aleatorios, comentamos que una de sus características fundamentales es la tendencia a la estabilidad con que se presenta cada uno de los posibles resultados.

Esto nos va a permitir medir la incertidumbre asociada a un suceso como la probabilidad de su ocurrencia.

Así por ejemplo si decimos que la puntualidad de los aviones de una determinada línea aérea expresada en términos de puntualidad es del 0,8%, queremos decir que tales vuelos llegarán a tiempo el 80% de las ocasiones.

Por ello, para el análisis de los problemas que se plantean en el Control de Calidad y que tienen el carácter de procesos aleatorios (como el de aceptación o rechazo de lotes de productos sin realizar inspecciones cien por cien 100%), es necesario estudiar las variables que definen dichos procesos y sus correspondientes distribuciones de probabilidad.

Estas distribuciones nos servirán también como conjuntos de referencia en aplicaciones más avanzadas de la estadística como pueden ser las técnicas de contraste de hipótesis, intervalos de confianza, etc., que veremos en los apartados 5 y 6 siguientes

4.2. Probabilidad y sus propiedades.

a. Concepto.

Un hecho comprobable empíricamente, es que la frecuencia relativa de aparición de ciertos sucesos en experiencias similares, se aproxima a un valor fijo constante al aumentar el número de experiencias.

Esta propiedad fue inicialmente descubierta en los juegos de azar: al tirar una moneda, la frecuencia relativa del suceso cara tiende, al aumentar el número de tiradas, hacia el valor constante ½ si la moneda está bien hecha.

Estas experiencias condujeron en el siglo XIX a definir la probabilidad de un suceso como el valor límite de su frecuencia relativa al repetir infinitamente la experimentación.

Esta definición presenta problemas importantes: desde el punto de vista teórico el límite anterior no puede interpretarse en el sentido del análisis, ya que no es posible fijar a priori un número de repeticiones n tal que, a partir de él, la diferencia entre la frecuencia relativa y la probabilidad sea menor que una cantidad prefijada.

Estrictamente pues, la probabilidad depende del grado de información disponible y la probabilidad de un suceso A debería indicarse como P(A/I), donde I representa un conjunto de información definida que contiene:

a. Los sucesos posibles al realizar el experimento. Se denomina espacio muestral a este conjunto de todos los sucesos posibles que es definido por el experimentador.

b. La evidencia empírica existente respecto a la ocurrencia de estos sucesos.

Para simplificar, supondremos en adelante que el conjunto I está perfectamente definido y escribiremos P(A) para indicar la probabilidad de un suceso cualquiera.

b. Propiedades.

Las propiedades de la probabilidad están basadas en su similitud con la frecuencia relativa, y en adelante seguiremos este enfoque.

1. La frecuencia relativa de un suceso *A*, *fr(A),* es un valor entre cero y uno, por tanto

$$0 \leq P(A) \leq 1$$

2. Llamaremos suceso seguro, *E*, al que ocurre siempre. Entonces:

$$P(E)=1$$

3. Si *A* y *B* son categorías mutuamente excluyentes y las unimos en una nueva *C=A+B*, que ocurre cuando se da o bien *A*, o bien *B*; la frecuencia relativa de C es la suma de las frecuencias relativas de *A* y *B*. Por tanto, para sucesos mutuamente excluyentes:

$$P(A+B)=P(A)+P(B)$$

4. Si *A* y *B* no son mutuamente excluyentes y llamamos $n_{AB}, n_{A\bar{B}}$ y $n_{\bar{A}B}$ al número de veces que aparecen los sucesos mutuamente excluyentes: (*A* y *B*), (*A* y no *B*), (no *A* y *B*), tendremos:

$$n_A = n_{AB} + n_{A\bar{B}}$$

$$n_B = n_{AB} + n_{\bar{A}B}$$

$$n_{A+B} = n_{AB} + n_{A\bar{B}} + n_{\bar{A}B}$$

en consecuencia:

$$n_{A+B} = n_A + n_B - n_{AB}$$

que conduce a la relación entre probabilidades:

$$P(A+B)=P(A)+P(B)-P(AB)$$

La frecuencia relativa de A condicionada a la ocurrencia de B se define considerando únicamente los casos en los que aparece B, y viendo en cuántos de estos casos ocurre el suceso A; es, por tanto, igual a la frecuencia de ocurrencia conjunta de A y B, partida por el número de veces que ha ocurrido B. Escribiremos:

$$fr(A/B) = \frac{n_{AB}}{n_B}$$

entonces, como $fr(A) = n_A / n; fr(B) = n_B / n; fr(AB) = n_{AB} / n$, se tiene:

$$\text{fr(A/B)} = \frac{\text{fr(AB)}}{\text{fr(B)}}$$

o, lo que es lo mismo,

$$fr(AB)=fr(A/B)fr(B)=fr(B/A)fr(A)$$

En consecuencia, exigiremos esta misma propiedad a la probabilidad y definiremos probabilidad de un suceso *A* condicionada a otro *B* por:

$$P(A/B)=\frac{P(AB)}{P(B)}$$

donde *AB* representa el suceso ocurrencia conjunta de *A* y *B*.

- ***Ejemplo*:**

Sea el experimento lanzar una dado y observar un número por un suceso E. Este suceso ocurrirá si, y sólo si, ocurre uno de los tres siguientes sucesos simples:

- observar un 2 (suceso A)
- observar un 4 (suceso B)
- observar un 6 (suceso C)

Entonces el suceso E sucede si suceden cualquiera de los sucesos simples. Puesto que no pueden suceder dos o más sucesos al mismo tiempo. La probabilidad de ocurrencia del suceso E se calcula como:

$$p(E)=p(A)+p(B)+p(C)=1/6+1/6+1/6=\frac{3}{6}=\frac{1}{2}$$

Supongamos que el suceso además de un número par requiera que éste sea menor o igual que 4 (suceso D):

$$P(D/E)=\frac{P(ED)}{P(E)}=\frac{2/6}{1/2}=\frac{2}{3}$$

c. Independencia de sucesos.

Diremos que dos sucesos *A* y *B* son independientes si el conocimiento de la ocurrencia de uno no modifica la probabilidad de aparición del otro. Por tanto, *A* y *B* son independientes entre si:

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

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

por (b.4), una definición equivalente de independencia de dos sucesos es:

$$P(AB)=P(A)P(B)$$

Esta definición se generaliza para cualquier número de sucesos: diremos que los sucesos A_1, ..., A_n son independientes si la probabilidad conjunta de todos los subconjuntos que pueden formarse con ellos es el producto de las probabilidades individuales.

La independencia entre sucesos puede en algunos casos preverse, pero en general debe determinarse experimentalmente. Por ejemplo, las averías en dos talleres contiguos pueden ser independientes si estos no guardan relación, y dependientes si las averías van ligadas al tipo de producto fabricado y ambos talleres producen el mismo.

- ***Ejemplo*:**

Sea el experimento extracción de piezas con reposición, los sucesos extraer la pieza correcta en la primera extracción (suceso A) y extraer la pieza correcta en la segunda extracción (suceso B) son independientes dado que el mismo número de piezas buenas/malas no ha variado.

d. Teorema de Bayes.

Consideremos un experimento que se realiza en dos etapas:

En la primera, los sucesos posibles, A_1, ..., A_n, son mutuamente excluyentes, con probabilidades conocidas, $P(A_i)$, y tales que:

$$\sum P(A_i) = 1$$

En la segunda etapa, los resultados posibles, B_j, dependen de los de la primera, y se conocen las probabilidades condicionadas $P(B_j/A_i)$ de obtener cada posible resultado B_j cuando aparece en la primera etapa el A_i.

Se efectúa ahora el experimento, pero el resultado de la primera fase, A_i, no se conoce, aunque sí el de la segunda, que resulta ser B_j. El teorema de Bayes permite calcular las probabilidades $P(A_i/B_j)$ de los sucesos no observados de la primera etapa, dado el resultado de la segunda.

Partiendo de la definición de probabilidad condicionada:

$$P(A_i / B_j) = \frac{P(A_i B_j)}{P(B_j)} = \frac{P(B_j / A_i)P(A_i)}{P(B_j)}$$

y, por otro lado:

$$P(B_j)=P(B_jA_1+B_jA_2+...+B_jA_n)$$

ya que B_j debe ocurrir con alguno de los n posibles sucesos A_i. Como los sucesos B_jA_1, B_jA_2, ... son mutuamente excluyentes, al serlo los A_i, tenemos:

$$P(B_j) = \sum_i P(B_j A_i) = \sum_i P(B_j / A_i)P(A_i)$$

y sustituyendo en la expresión de $P(A_i/B_j)$:

$$P(A_i / B_j) = \frac{P(B_j / A_i)P(A_i)}{\sum_i P(B_j / A_i)P(A_i)}$$

que se conoce como Teorema de Bayes.

- ***Ejemplo*:**

Supongamos que tenemos tres instalaciones I_1, I_2, I_3, produciendo un mismo elemento y que disponemos de datos de los últimos meses que nos indican que la instalación I_1 produce un 7% de elementos defectuosos, la I_2 un a 4% y la I_3 un 2%.

Si en un día determinado las tres instalaciones producen la misma cantidad de elementos, la probabilidad de obtener un elemento defectuoso condicionado a que lo haya fabricado la instalación nº1 se expresa mediante $P(D|I_1) = 0.07$.

En las mismas condiciones $P(D|I_2) = 0.04$ y $P(D|I_3) = 0.02$

Vemos pues que es muy sencillo determinar estas probabilidades condicionadas. Sin embargo en ocasiones, las probabilidades condicionadas de interés son las inversas, por ejemplo: ¿cuál es la probabilidad de que de la producción diaria de las tres instalaciones, que se encuentra mezclada, al extraer un elemento que resulta defectuoso haya sido producido por la instalación I_1?

Esta probabilidad la expresaríamos mediante: $P(I_1|D)$

Ahora ya podemos contestar a la pregunta anterior:

$$P(I_1|D)=P(D|I_1)\ P(I_1) / [P(D|I_1)\ P(I_1) + P(D|I_2)\ P(I_2) + P(D|I_3)\ P(I_3)]$$

En el caso de que las producciones de las tres instalaciones fueran iguales, el resultado sería:

$P(D|I_1) = 0.07$
(probabilidad de que un elemento sea defectuoso si lo ha producido la instalación nº 1)

$P(D|I_2) = 0.04$
(probabilidad de que un elemento sea defectuoso si lo ha producido la instalación nº 2)

$P(D|I_3) = 0.02$
(probabilidad de que un elemento sea defectuoso si lo ha producido la instalación nº 3)

$P(I_1) = 1/3$
(probabilidad de que el elemento lo haya producido la instalación nº 1)

$P(I_2) = 1/3$
(probabilidad de que el elemento lo haya producido la instalación nº 2)

$P(I_3) = 1/3$
(probabilidad de que el elemento lo haya producido la instalación nº 3)

$$P(I1|D) = (0.07)\ (1/3) / [(0.07)\ (1/3) + (0.04)\ (1/3) + (0.02)\ (1/3)] = 0.54$$

e. La estimación de probabilidades en la práctica.

- ***Sucesos elementales y compuestos.***

Llamaremos ***sucesos elementales*** de un experimento a un conjunto de resultados posibles (*a, b, c,* ...) que verifican:

1. Siempre ocurre alguno de ellos.

2. Son mutuamente excluyentes: la ocurrencia de uno implica la no ocurrencia de los demás.

Llamaremos **sucesos compuestos** a los construidos a partir de uniones de resultados elementales. Por ejemplo:

Experimento	***Sucesos elementales***	***Sucesos compuestos***
Tirar un dado	(1, 2, 3, 4, 5, 6)	Número par; número impar; menor que 4; múltiplo de 3.
Contar los varones en familias con tres hijos	(0,1,2,3)	Más de uno; menos de tres.
Número de días que una máquina está averiada en un mes	(0, 1, ...30)	Más de 10; menos de 20; entre 5 y 15 inclusive.

- **Métodos para determinar probabilidades.**

La determinación de probabilidades para sucesos compuestos requiere conocer las de los sucesos elementales. Estas probabilidades se determinan:

1. *Estudiando la frecuencia relativa al repetir el experimento en condiciones similares*. Este método sólo es factible en ocasiones en que es posible una experimentación continuada.

2. *Encontrando, a partir de la naturaleza del experimento, relaciones que liguen a sus probabilidades elementales y determinen sus valores*. El caso más simple es el de equiprobabilidad, que estudiaremos a continuación.

3. *Combinando la experimentación con la teoría sobre la naturaleza del experimento*. Este es el método más utilizado en la práctica y más fructífero. Lo utilizaremos en la sección 4 para construir los modelos de distribución de probabilidad más importantes.

- **El caso de equiprobabilidad.**

  En ocasiones, la simetría de los sucesos elementales sugiere considerarlos equiprobables.

  Este razonamiento se ha aplicado repetidamente en los juegos de azar a problemas como tirar dados o monedas, extraer naipes de barajas, etc. A veces, el mecanismo generador de los resultados está diseñado para intentar asegurar esta equiprobabilidad, como en la lotería o la ruleta.

  En estos casos, si existen *n* sucesos elementales equiprobables, la probabilidad de cada uno de ellos debe ser 1/*n*, para asegurar que la suma total sea uno.

  La probabilidad de un suceso compuesto *A* que contiene f sucesos elementales será *f/n*, lo que da lugar a la regla:

$$P(A) = \frac{\text{casos favorables } (f)}{\text{casos posibles } (n)}$$

  Esta regla sólo debe utilizarse cuando la simetría esté confirmada por el mecanismo generador (como en la lotería) o por la evidencia empírica.

4.3. Variables aleatorias.

El cálculo de probabilidades utiliza variables numéricas que se denominan aleatorias, porque sus valores vienen determinados por el azar.

En todo proceso de observación o experimento podemos definir una variable aleatoria asignando a cada resultado del experimento un número:

a. Si el resultado del experimento es numérico porque contamos o medimos, los posibles valores de la variable coinciden con los resultados del experimento.
b. Si el resultado del experimento es cualitativo, hacemos corresponder a cada resultado un número arbitrariamente; por ejemplo, 0, si una pieza es buena, y 1, si es defectuosa.

Diremos que se ha definido una variable aleatoria o que se ha construido un modelo de distribución de probabilidad, cuando se especifican los posibles valores de la variable con sus probabilidades respectivas.

a. Variables aleatorias discretas.

Diremos que una variable aleatoria es discreta cuando toma un número de valores finito, o infinito numerable. Estas variables corresponden a experimentos en los que se cuenta el número de veces que ha ocurrido un suceso.

- ***Función de distribución.***

Un modelo de distribución de probabilidad es la representación idealizada de un experimento aleatorio y se construye indicando los valores posibles de la variable aleatoria asociada al experimento y sus probabilidades respectivas.

La forma más general de caracterizar estos modelos es mediante la función de distribución, *F(x)*, definida en cada punto x_0 como la probabilidad de que la variable aleatoria *x* tome un valor menor o igual que x_0. Escribiremos:

$$F(x_0) = P(x \leq x_0)$$

La función de distribución, se define para todo punto del eje real, es siempre no decreciente, y por convenio:

$$F(-\infty) = 0$$
$$F(+\infty) = 1$$

Suponiendo que la variable *x* toma los valores posibles (x_1, ..., x_n), siendo

$$x_1 \leq x_2 \leq x_3 \ldots \leq x_n$$

y

$$\sum P(x_i) = 1$$

entonces, la función de distribución vendrá definida por:

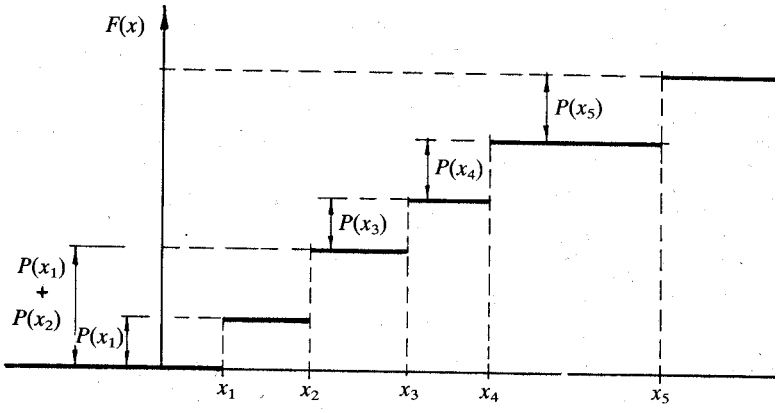
$$F(x_1) = P(x \le x_1) = P(x_1)$$
$$F(x_2) = P(x \le x_2) = P(x_1) + P(x_2)$$

.....

$$F(x_n) = P(x \le x_n) = \sum_{i=1}^{n} P(x_i) = 1$$

Por tanto, la función de distribución, *F(x),* tendrá saltos en los puntos (x_1, ...,x_n) iguales a la probabilidad de dicho punto, siendo constante en los intervalos entre los puntos de salto.

La figura representa gráficamente *F(x)* para una variable discreta.



Función de distribución para una variable discreta con valores posibles x1,x2,x3,x4,x5.

b. Variables aleatorias continuas.

Diremos que una variable aleatoria es continua cuando puede tomar cualquier valor en un intervalo.

Por ejemplo, el peso de una persona, el tiempo de duración de un suceso, etc., corresponden a variables aleatorias continuas.

No es posible conocer el valor exacto de una variable continua, ya que medir su valor consiste en clasificarlo dentro de un intervalo: si el resultado de medir una longitud es 23 mm, todo lo que podemos afirmar es que la longitud real, no observable, está en el intervalo 22,5 mm a 23,5 mm. Los modelos descriptivos de variables aleatorias continuas se basan en este principio.

– ***Función de densidad.***

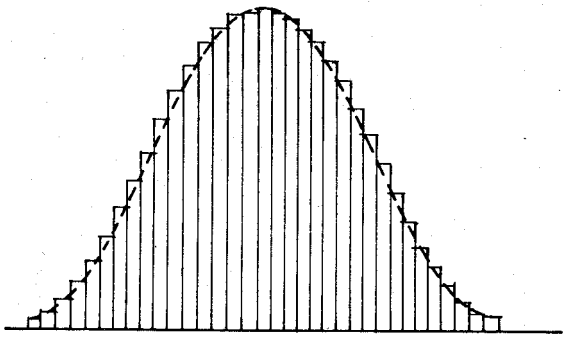
Supongamos, para concretar, que medimos una magnitud (longitud de piezas, tiempo de funcionamiento, etc.), y representamos las medidas obtenidas en un histograma; es razonable admitir (y se ha comprobado repetidamente en la práctica) que, tomando más y más observaciones y haciendo clases cada vez más finas, el histograma tenderá a una curva suave que describirá el comportamiento a largo plazo de la variable estudiada.

El conocimiento de la función de densidad *f(x)* permite calcular cualquier probabilidad por integración. Por ejemplo, la probabilidad de que la variable *x* sea menor que x_0 corresponde a sumar las frecuencias relativas de todas las clases que contienen valores menores que x_0; este resultado se obtiene fácilmente calculando el área bajo la función de densidad hasta el punto x_0 mediante:

$$P(x < x_0) = \int_{-\infty}^{x_0} f(x)dx$$

Análogamente, la probabilidad de que la variable x tome un valor entre x_0 y x_1 se calculará como:

$$P(x_0 < x < x_1) = \int_{x_0}^{x_1} f(x)dx$$



Histograma y función de densidad

- ***Función de distribución.***

La función de distribución para una variable aleatoria continua se define como en el caso discreto por:

$$F(x_0) = P(x \le x_0) = \int_{-\infty}^{x_0} f(x)dx$$

4.4. Distribuciones de probabilidad discretas.

Las distribuciones de probabilidad discretas más importantes son: la binomial, la geométrica, la binomial negativa, la hipergeométrica y la de Poîsson.

a. El proceso de Bernouilli y sus distribuciones asociadas.

Supongamos que observamos elementos de una población con las siguientes características:

- En la población existen elementos clasificados en dos categorías (p.e. correcto/incorrecto, blanco/negro, etc.).
- La proporción de elementos en ambas categorías se mantiene constante a lo largo de todo el proceso (p.e. cuando se realiza una extracción de la población para realizar una observación, se devuelven los elementos extraídos a la población, es decir el muestreo se realiza con reemplazamiento. O bien el tamaño de la población es tan grande en comparación con la muestra extraída que aunque no se realice reemplazamiento, la proporción no se ve afectada).

Los sucesos de obtener elementos de una característica o de la otra son independientes entre sí. (p.e. La probabilidad de que al extraer un elemento sea defectuoso de un lote con una proporción determinada de estos es siempre la misma, no viéndose afectada por haber extraído una combinación anterior de elementos defectuosos y/o conformes).

Dependiendo de como definamos la variable aleatoria, respecto al proceso de Bernouilli, obtendremos distintas distribuciones de probabilidad.

– ***Distribución Binomial.***

La variable aleatoria la definimos como:

X = número de elementos de una de las dos categorías que aparecen al observar una muestra de tamaño n de la población.

En estas condiciones si llamamos:

p = proporción de elementos de una categoría

q = proporción de elementos de la otra categoría

$$q = 1 - p$$

se cumple que la probabilidad de obtener a elementos de la primera categoría al observar una muestra de tamaño "n" obtenida de la población, viene dada por la expresión:

$$P(X=a)=\binom{n}{a}p^a q^{(n-a)}=\frac{n!}{a!(n-a)!}p^a q^{(n-a)}$$

Las medidas de tendencia central y dispersión de esta distribución son:

Media $\Rightarrow \bar{x} = np$

Varianza $\Rightarrow \sigma^2 = npq$

- ***Ejemplo*:**

Si de una población consistente en la producción diaria de tuercas en una fábrica (10.000 unidades), que contiene un 2.5% de defectuosas, tomamos una muestra de cuatro, ¿cuál es la probabilidad de obtener una defectuosa? ¿y de obtener 3 defectuosas?

Tenemos que p=0,025 y 1-p=0,975 por lo que:

- La probabilidad de que sea defectuosa una de las cuatro tuercas de la muestra es:

$$p\binom{1}{4} = \frac{4!}{1!(4-1)} 0{,}025^1 0{,}975^{(4-1)} = 0{,}093$$

- La probabilidad de que sean defectuosas tres de las cuatro de la muestra es:

$$p\binom{3}{4} = \frac{4!}{3!(4-3)} 0{,}025^3 0{,}975^{(4-3)} = 0{,}00006$$

– ***Distribución geométrica o de Pascal.***

La variable aleatoria se define como:

X = número de elementos de una categoría obtenidos antes de que aparezca el primero de la segunda categoría.

Se cumple que la probabilidad de obtener "r" elementos de una categoría antes de aparecer el primer elemento de la otra categoría es:

$$P(X = r) = (1-p)^r p$$

siendo p la proporción de elementos de la segunda categoría.

Las medidas de tendencia central y dispersión de esta distribución son:

Media $\Rightarrow \bar{x} = q / p$

Varianza $\Rightarrow \sigma^2 = q / p^2$

- ***Ejemplo*:**

Si en la misma población del ejemplo anterior extraemos tuercas de una en una reemplazándolas después, la probabilidad de extraer 5 tuercas buenas seguidas antes de que aparezca la primera defectuosa es:

$P(X = 5) = (1 - 0.025)^5 = 0.022$

La probabilidad pedida es de un 2.2%

– ***Distribución binomial negativa.***

La variable aleatoria se define como:

X = número de elementos de una categoría obtenidos antes de que aparezca el número k de la segunda categoría.

Se cumple que la probabilidad de obtener "r" elementos de una categoría antes de aparecer el elemento k de la otra categoría es:

$$P(X=r)=\binom{k+r-1}{r}p^k q^r$$

siendo p la proporción de elementos de la segunda categoría.

Las medidas de tendencia central y dispersión de esta distribución son:

Media $\Rightarrow \bar{x} = kq / p$

Varianza $\Rightarrow \sigma^2 = kq / p^2$

♦ ***Ejemplo*:**

Si en la misma población del ejemplo anterior extraemos tuercas de una en una reemplazándolas después, la probabilidad de extraer 50 tuercas buenas antes de que aparezca la tercera defectuosa es:

$$P(X=50)=\binom{3+50-1}{50}(0.025)^3(0.975)^{50}=0.0058$$

La probabilidad pedida es de un 0.6%

– ***Distribución hipergeométrica.***

En el caso de encontrarnos con un fenómeno en una población, definida exactamente igual que el caso de la distribución binomial, salvo en lo que respecta a constancia de la proporción de elementos en cada una de las dos categorías (p.e. cuando las extracciones no pueden ser con reposición debido a ensayos o pruebas de tipo destructivo y además el tamaño de la muestra como para poder suponer despreciable la variación en las proporciones), se utiliza las distribución hipergeométrica.

La probabilidad de obtener "a" elementos de una categoría al extraer una muestra de tamaño "n" de un lote de tamaño "N", sin reposición viene dada por la expresión:

$$P(X=a)=\binom{Np}{a}\binom{Nq}{(n-a)}/\binom{N}{n}$$

Las medidas de tendencia central y dispersión de esta distribución son:

Media $\bar{x} = np$

Varianza $\sigma^2 = npq\ (N\text{-}n)/(N\text{-}1)$

- ***Ejemplo*:**

  Dado un lote de 50 bengalas pirotécnicas de señalización, con un porcentaje de defectuosos del 6%, se extrae una muestra de 5 unidades para realizar un ensayo de encendido. ¿Cuál es la probabilidad de que dos de ellas sean defectuosas?

  En este caso es necesario utilizar la distribución hipergeométrica debido a que el ensayo es destructivo y no puede hacerse con reposición, siendo además el lote muy pequeño con respecto a la muestra como para suponer constante la proporción de elementos defectuosos.

$$P(x=2)=\binom{50x0,06}{2}\binom{50x0,94}{5-2}/\binom{50}{5}=\binom{3}{2}\binom{47}{3}/\binom{50}{5}=0,023$$

b. El proceso de Poisson y sus distribuciones asociadas.

Cuando observamos sucesos puntuales sobre un soporte continuo (averías de una instalación en el tiempo, defectos de una plancha de metal, etc.) y el proceso se caracteriza por:

- Es estable, es decir, se produce un número medio "δ" constante de sucesos en un intervalo dado del soporte continuo:

  δ =dos picaduras por cada 10 metros de tubería

  δ =5 defectos en la trama por cada 25_m^2 de tejido

- Los sucesos son independientes (el proceso no tiene memoria).

  En estos casos diremos que el proceso es del tipo de Poisson. Según definamos distintas variables aleatorias para este proceso, obtendremos distintas distribuciones de probabilidad.

– ***Distribución de Poisson.***

  La variable aleatoria la definimos como:

  X = número de sucesos en un intervalo de longitud fijo.

  Se cumple que la probabilidad de obtener "a" sucesos en un intervalo de longitud igual al que define el número de sucesos medio por longitud de intervalo es:

$$P(X = a) = (d^a / a!)e^{-d}$$

Las medidas de tendencia central y dispersión de esta distribución son:

Media $\Rightarrow \bar{x} = \delta$

Varianza $\Rightarrow \sigma^2 = \delta$

- ***Ejemplo*:**

A una empresa llegan una media de 10 llamadas cada cinco minutos. Suponiendo que las llamadas telefónicas siguen una distribución de Poisson, calcular la probabilidad de que lleguen 20 llamadas cada cinco minutos.

Puesto que en la distribución de Poisson, la media es δ y nos dicen que el número medio de llamadas es de 10 cada cinco minutos, consideraremos:

$$\delta = 10$$

Si la expresión:

$$P(X = a) = (d^a / a!)\, e^{-d}$$

sustituimos

a = 20

$\delta = 10$

$$P(X = 20) = (10)^{20} e^{-10} / 20! = 0.002$$

– ***Aproximación de la distribución binomial a la de Poisson.***

Cuando en una población, el porcentaje "p" de elementos de una de las dos categorías es muy pequeño con relación al tamaño de la muestra n, la distribución Binomial puede ser aproximada a la distribución de Poisson. (Para n>50 y p<0.1, o bien np<5).

En estas condiciones, la probabilidad de que al extraer una muestra de tamaño n, de una población con un porcentaje p de elementos de una categoría, aparezcan "a" elementos de esa categoría viene dada por la expresión resultante de sustituir el valor del producto np por el de δ en la expresión de Poisson:

$$P(X=a)=(np)^a e^{-np}/a!$$

♦ ***Ejemplo*:**

En un libro, la probabilidad de que una palabra esté mal escrita es de 1/50.000. Calcular, en el caso de que un libro tenga 200.000 palabras, la probabilidad de que:

- *No haya errores:*

  Suponemos una distribución binomial donde:

  n=200.000

  p=1/50.000

  q=49.999/50.000

$$P(X=0)=\binom{200.000}{0}(1/50.000)(49.999/50.000)^{200.000}=0.018$$

  Si hubiéramos utilizado Poisson (n>5, p<0.1)

  $d=np=200.000/50.000=4$

  $P(X=0)=(4^{0e-4})/0!=0.018$ coincide como era de esperar.

- *Haya más de seis errores:*

$$P(X>6)=1-P(X\le 6)=1-\sum_{i=1}^{6}(4^i e^{-4})(i!)$$

  entrando en la Tabla del Ejemplo 1 con: k=6 y δ=4 obtenemos $P(X\le 6)$=0.889

  Luego P(X>6)=1-0.889=0.111 del orden de un 11%

4.5. Distribuciones de probabilidad continuas.

a. La distribución normal.

El modelo de distribución de probabilidad para variables continuas más importante es la distribución normal, cuya función de densidad es:

$$f(x) = \frac{1}{s\sqrt{2p}} \exp\left[-\frac{1}{2s^2}(x-m)^2\right]$$

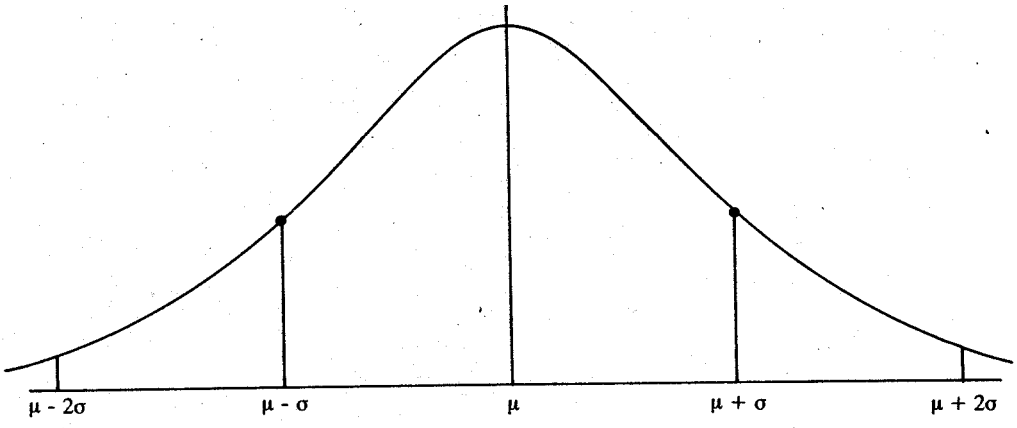
$$-\infty < x < \infty$$

que aparece dibujada en la siguiente figura.

La función *f* depende de dos parámetros:

- μ , que es al mismo tiempo la media, la mediana y la moda de la distribución,
- σ , que es la desviación típica.

Diremos que una variable es *N(μ , σ)* cuando sigue la función de densidad (4.6). La mediana de las desviaciones absolutas para una variable *N(μ , σ)* es 0,68 σ .



La distribución normal

La curva de distribución normal es una curva conforma de campana extendida indefinidamente en ambas direcciones.

La probabilidad de que la variable se encuentre entre dos valores determinados viene dada por el área encerrada debajo de la curva entre los dos valores.

Por ejemplo, si tenemos un conjunto de elementos cuyo peso sigue una distribución de media 10 gr. y desviación típica 2, y queremos saber cual es la probabilidad de tener elementos cuyo peso esté entre 8 y 12 gr. nos bastará con calcular el área encerrada bajo la curva (que está definida por su media=10 y su desviación típica=2) y los valores 8 y 12.

La distribución normal aproxima lo observado en muchos procesos de medición sin errores sistemáticos.

Por ejemplo, las medidas físicas del cuerpo humano en una población, las características psíquicas medidas por test de inteligencia o personalidad, las medidas de calidad en muchos procesos industriales o los errores de las observaciones astronómicas siguen distribuciones normales.

Una justificación de la frecuente aparición de la distribución normal es el teorema central del límite, que establece que cuando los resultados de una experimento son debidos a un conjunto muy grande de causas independientes, que actúan sumando sus efectos, siendo cada efecto individual de poca importancia respecto al conjunto, es esperable que los resultados sigan una distribución normal.

La variable normal con $\mu = 0$ y $\sigma = 1$ se denomina ***normal estándar***, N(0,1), y su función de distribución está tabulada. Para calcular probabilidades en el caso general, transformaremos la variable aleatoria normal x en la variable normal estándar z, mediante:

$$z = \frac{x - m}{s}$$

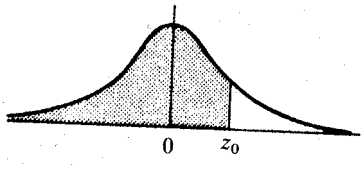
que convierte una variable x con media μ y desviación típica σ en la normal estándar z. La función densidad para el cambio de variable:

$$f(z) = \frac{1}{s\sqrt{2p}} e^{-\frac{z^2}{2}} \qquad s = \frac{1}{\sqrt{2p}} e^{-\frac{z^2}{2}}$$

que es la normal estándar. El cálculo de probabilidades de x se efectúa utilizando la expresión:

$$F(x_0) = P(x \le x_0) = P(m + sz \le x_0) = P\left(z \le \frac{x_0 - m}{s}\right) = F\left(\frac{x_0 - m}{s}\right)$$

donde $\Phi(.)$ representa la función de distribución de la normal estándar y que corresponde al valor del área rayada en la figura.



Para calcular ese área utilizaremos las tablas destinadas a estos efectos. En dichas tablas aparecen los siguientes campos:

- A la izquierda con valores desde 0 a 4 están los posibles valores de la variable normal reducida z.
- En la parte superior aparece una fila de valores comprendidos entre 0.09 y 0.00 que se utilizan para expresar las centésimas de la variable normal estándar z.
- El resto de valores que aparecen en la zona central de la tabla son probabilidad de que la variable z tome un valor inferior o igual al determinado por la columna y fila anteriormente comentadas.

Como ejemplo, ¿cuál es la probabilidad de que una variable normalizada z tome valores inferiores a 1.53?

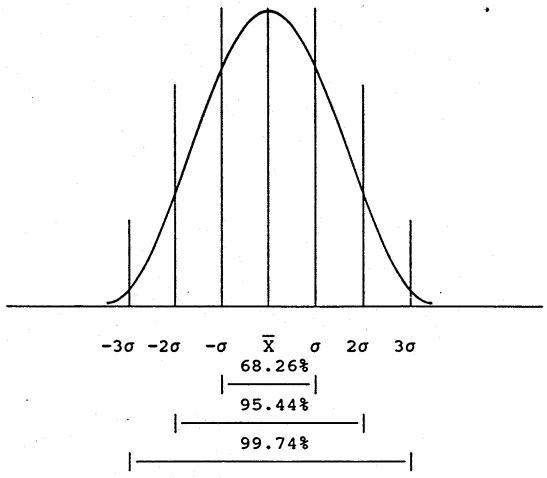
$$P(Z<1.53)$$

Entraríamos en la tabla con el valor 1.5 en la columna de la izquierda y el valor 0.03 de la fila superior (1.53=1.50+0.03) dándonos en el cruce de ambos valores el valor 0.9406.

Por lo tanto P(Z<1.53)=0,93699

Además para el cálculo rápido de probabilidades con la utilización de las tablas de la variable normal estándar hay que tener en cuenta las siguientes propiedades:

- La curva N(0,1) es simétrica respecto a su media.
- El área total bajo la curva es igual a la unidad.
- El 99.74% de los valores de X se encuentran en un intervalo que centrado en la media se extiende 3σ a ambos lados de la misma, el 95.44% a 2σ y el 68.26% a σ.



b. Aproximaciones a la Normal.

Una propiedad importante de la normal es que puede utilizarse para aproximar probabilidades de variables binomiales y de Poisson.

En 1733, De Moivre demostró que si x es una variable binomial de parámetro p, la distribución de:

$$\frac{x - np}{\sqrt{npq}}$$

converge hacia la distribución de la normal cero, uno.

En la práctica, esto se traduce en que si n es grande (mayor que 30), y p no muy cercano a cero o uno, podemos calcular la probabilidad de que la variable binomial x esté en (a, b) considerando a x como una variable normal, de $\mu = np$ y $\sigma = \sqrt{npq}$, y buscando el área encerrada entre a y b.

La aproximación mejora tomando el intervalo (*a*-0,5*; b*+0,5), que tiene en cuenta que el número entero n equivale al intervalo continuo (*n*-0,5; *n*+0,5). Por tanto, la condición para una variable discreta:

$$a \leq x \leq b$$

equivale, para una variable continua, a:

$$a-0,5\le x\le b+0,5$$

En general esta aproximación se utiliza para $npq>5$.

En la siguiente tabla se muestran las probabilidades binomiales acumuladas.

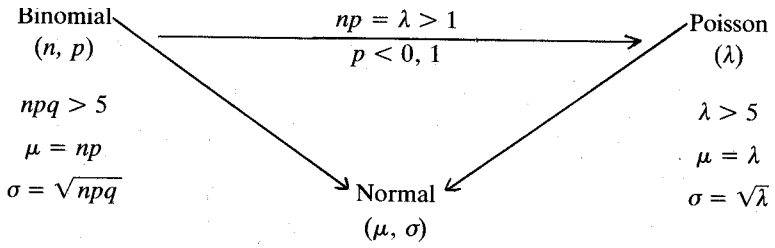
La distribución normal también puede aproximar la distribución de Poisson cuando $\lambda>5$. El procedimiento es:

$$p(a\le x_p\le b)\simeq p(a-0,5\le x_n\le b+0,5)$$

donde x_p es una variable de Poisson de parámetro λ y x_n es una variable normal de parámetros $m=l,\ s=\sqrt{l}$.

En la siguiente tabla se muestran las probabilidades de Poisson acumuladas.

A continuación mostramos la relación existente entre las comentadas distribuciones.



Relación entre distribuciones

c. La Distribución χ^2 de Pearson.

Si consideramos las variables z_1, z_2, ..., z_n que se caracterizan por:

a. Ser independientes.
b. Cada una de ellas es una N(0,1).

A la nueva variable aleatoria χ^2 se la define como:

$$\chi^2 = z_1^2, z_2^2, \ldots z_n^2$$

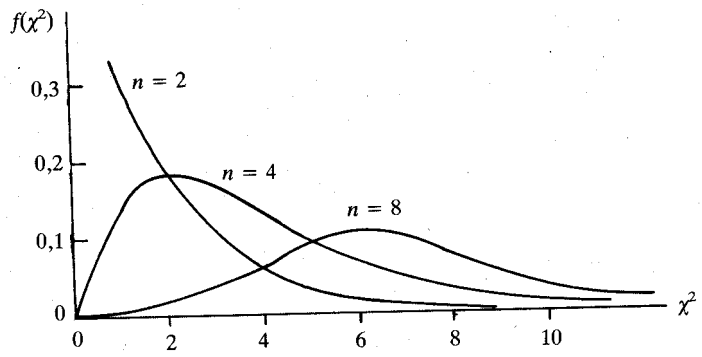
con n grados de libertad.

La distribución χ^2 es asimétrica y está tabulada en función de n.

Sus parámetros son:

Media $\Rightarrow \mu = n$

Varianza $\Rightarrow \sigma^2 = 2n$



En la siguiente tabla se muestran los valores estandarizados de esta distribución.

d. La Distribución t de Student.

La distribución t con n grados de libertad se define como:

$$t = \frac{n}{\left[(1/n)\chi^2\right]^{0.5}}$$

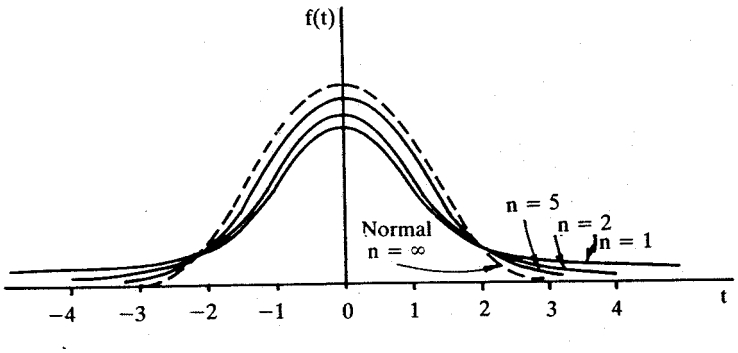
siendo χ^2 una distribución de Pearson con n grados de libertad.

La variable t es simétrica, con mayor dispersión que la distribución N(0,1). Cuando el número de grados de libertad crece, esta distribución se aproxima a la normal N(0,1).

Sus parámetros son:

Media $\Rightarrow \mu = 0$

Varianza $\Rightarrow \sigma^2 = n/(n-2)$ para n>2



En la siguiente tabla se muestran los valores estandarizados de esta distribución.

e. La Distribución F de Fisher.

Si tenemos dos distribuciones χ^2 de Pearson de n y m grados de libertad, se define la variable $F_{(n,m)}$ como:

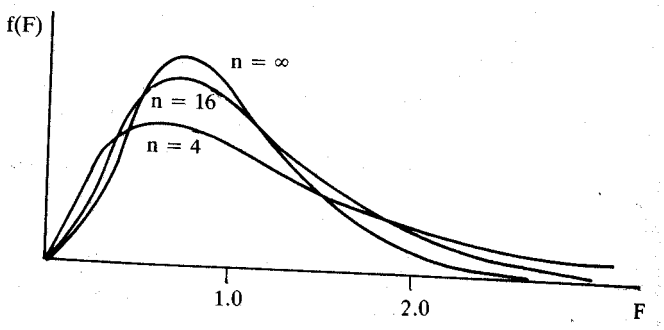
$$F_{(n,m)} = \frac{(1/n)\chi^2{}_n}{(1-m)\chi^2{}_m}$$

cumpliéndose que: $F_{(n,m)} = 1/F_{(m,n)}$

sus parámetros son:

Media $\Rightarrow \mu = m/(m-2) \quad m > 2$

Varianza $\Rightarrow \sigma^2 = \dfrac{2m^2(n+m-2)}{n(m-2)^2(m-4)}$



Se requiere de la tabla que muestra los valores estandarizados de esta distribución.

5. La estimación del modelo.

5.1. Introducción a la inferencia estadística.

Cuando estudiamos las distribuciones de probabilidad, ya vimos como establecidas unas hipótesis del mecanismo generador de datos, (proceso de Bernouilli, Poisson, etc.) se deducía cuales eran las probabilidades asociadas a cada valor de la variable.

$$P(x = x_i) = f(x_i)$$

La inferencia estadística realiza el proceso inverso: mediante la observación se obtienen datos de las probabilidades asociadas a cada valor de la variable (frecuencias) y a partir de éstos ha de deducirse (inferir) el modelo de probabilidad que los ha generado.

Existen fundamentalmente dos procedimientos para realizar la inferencia estadística: el método paramétrico y el no-paramétrico.

a. ***Método paramétrico***: Se supone que los datos provienen de un tipo de distribución conocida (Normal, Binomial, Poisson, etc.) siendo la incógnita los parámetros de la distribución.

b. ***Método no-paramétrico***: No se supone conocida la distribución que siguen las observaciones siendo las hipótesis muy generales (continua/discreta, simétrica/asimétrica, etc.).

Independientemente del método utilizado, es necesario estudiar la toma de muestra de donde vamos a extraer la información necesaria para analizar la población.

5.2. Muestreo.

Llamaremos ***población*** a un conjunto de elementos en los que se estudia una característica dada. Por lo general no es posible estudiar esta característica en todos los elementos de la población debido a:

- Experimentos destructivos.
- Los elementos no existen físicamente (poblaciones virtuales).
- Problemas económicos (tiempo y/o dinero).

En estas ocasiones se selecciona un subconjunto representativo de la población para su estudio que llamaremos **muestra**.

a. Muestreo aleatorio simple.

El muestreo se denomina aleatorio simple cuando:

- Para todos los elementos es igual su probabilidad de ser elegido.
- La población es idéntica en cada extracción (p.e. muestreo con reemplazamiento).

La muestra se selecciona mediante un mecanismo aleatorio como puede ser una tabla de números aleatorios como la mostrada en la siguiente figura. Para ello se numeran los elementos de la población del 1 al N y se toman de la tabla números aleatorios de tantas cifras como tenga N. El valor del número aleatorio indicará el elemento seleccionado.

Si llamamos:

$$x=(x_1, x_2, x_3, ..., x_n)$$

al conjunto de valores que componen una muestra de tamaño n, la probabilidad de obtener ese conjunto de valores y no otro, es:

$$P(x)=P(x_1, x_2, x_3, ..., x_n)$$

Con las condiciones del muestreo aleatorio simple, esa probabilidad será:

$$P(x)=P(x_1) \times P(x_2) \times P(x_3) ..., \times P(x_n)$$

que es la condición matemática de la muestra aleatoria simple.

b. Método de Montecarlo.

Se utiliza para general muestras de una población de la que no se disponen elementos pero sin embargo sí se conocen sus características. (Función de densidad o de distribución).

– ***Distribuciones Discretas.***

Supongamos que tenemos una variable aleatoria con la siguiente función de distribución:

x	P(x)	F(x)
0	0.1	0.1
1	0.2	0.3
2	0.5	0.8
3	0.1	0.9
4	0.1	1.0

1. Extraemos un número aleatorio de dos dígitos (N.A.).

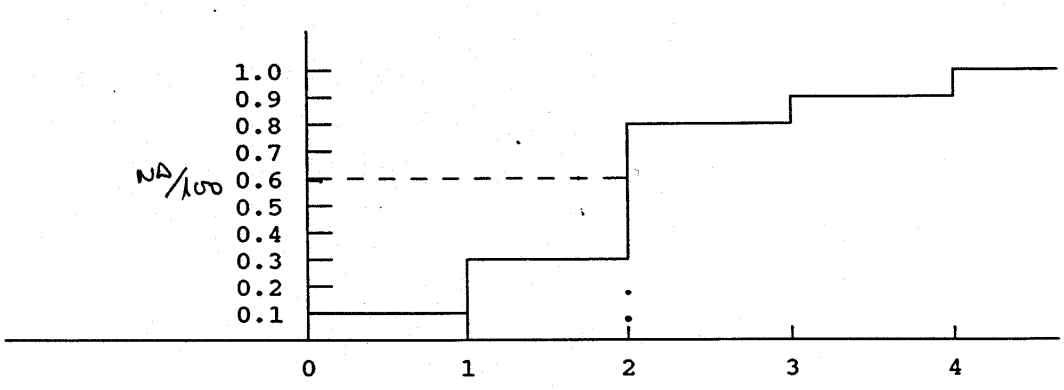
2. El número aleatorio lo dividimos por 100 y lo convertimos en decimal.

3. Establecemos la siguiente correspondencia entre el número aleatorio dividido por cien y el valor de x de la muestra, tal que x es el valor más pequeño que verifica F(x)>N.A./100

	x
N.A./100 < 0.1	0
0.1 ≤ N.A./100 < 0.3	1
0.3 ≤ N.A./100 < 0.8	2
0.8 ≤ N.A./100 < 0.9	3
0.9 ≤ N.A./100 < 1.0	4

4. Lo repetimos tantas veces como el tamaño de la muestra deseada.

De una forma gráfica: Para N.A./100 = 0.60 X=2



- ***Distribuciones continuas.***

  1. Se toma un número aleatorio de tantas cifras como precisión se desee.
  2. Se convierte en número decimal dividiendo el número aleatorio por 10^n, siendo n el número de cifras del número aleatorio.
  3. Si la función de distribución es F(X), se toma X=F-1(N.A.)
  4. Repetirlo tantas veces como el tamaño de muestra deseado.

  Por ejemplo, si consideramos una variable distribuida como exponencial:

$$F(x) = 1 - e^{-ax}$$

$$N.A./10^n = 1 - e^{-ax}$$

$$-ax = Log.(1 - N.A./10^n)$$

$$x = (-1/a)Log.(1 - N.A./10^n)$$

c. Muestreo estratificado.

En ocasiones interesa generar muestras que tengan una cierta estructura. Esto sucede cuando los elementos de la población no son homogéneos respecto a la muestra a estudiar y además disponemos de información sobre ello.

Por ejemplo los sondeos de opinión donde los elementos son heterogéneos en razón a sexo, edad, profesión, etc. La muestra deberá recoger elementos en proporción igual a los de la población para tener una estructura análoga a ésta. A partir de este momento, dentro de cada grupo o estrato, la elección se hace por muestreo aleatorio simple.

d. Muestreo sistemático.

Se utiliza cuando los elementos de la población están ordenados en listas.

Tamaño de la Población = N

Tamaño deseado de la muestra = n

K = entero más cercano a N/n

Se toma un primer elemento de la población como muestra, por ejemplo el que tiene el número de orden n_1. El resto de los elementos de la muestra se toman a intervalos constantes k.

$$n_1, n_1+k, n_1+2k, n_1+3k$$

Si el orden en la lista es al azar, este muestreo es equivalente al aleatorio simple. Si existe algún orden en el cual tienden a ser más semejantes los elementos de la población en función de su proximidad, este muestreo es más preciso que el aleatorio simple.

e. Muestreo polietápico.

Para poblaciones muy heterogéneas. Por ejemplo para seleccionar una muestra de personas de una ciudad, seleccionaríamos los barrios mediante muestreo aleatorio simple, luego calles dentro de cada barrio, a continuación viviendas en cada calle y por último un piso en cada vivienda.

5.3. La estimación puntual.

Suponemos que tenemos una población que sigue una distribución con forma conocida aunque con parámetros desconocidos. De esta población extraemos una muestra aleatoria simple y el problema consiste en estimar los parámetros de la población a partir de los datos muestrales.

a. El método de los momentos.

El primer método para obtener estimadores es el método de los momentos que consiste en tomar como estimador del momento de orden k de la población, el momento de orden k calculado de la muestra.

La idea es simple: tomar como estimador de la varianza de la población la varianza de la muestra; de la media de la población, la media muestral, y así sucesivamente.

Para juzgar la bondad de los estimadores obtenidos por el método de los momentos, necesitamos establecer las propiedades deseables de los estimadores.

La principal dificultad a la hora de definir la bondad o no de un estimador es que este estimador es una variable aleatoria que varía de muestra en muestra. Si consideramos una población de la que se toman muestras con reemplazamiento de tamaño n, y calculamos el valor medio de cada muestra, N muestras darán lugar a N valores de medias muestrales:

$$\bar{x}_1, \bar{x}_2, \bar{x}_3, \bar{x}_4, \bar{x}_5, \bar{x}_6, \bar{x}_n$$

Si N es muy grande, estos valores seguirán una distribución que denominaremos distribución muestral de la media.

En estas condiciones, se cumple que la distribución muestral de la media sigue una distribución normal (cuando n>30) cuyos parámetros son:

Media muestral = Media poblacional

Varianza muestral = Varianza poblacional/n

5.4. Propiedades de los estimadores.

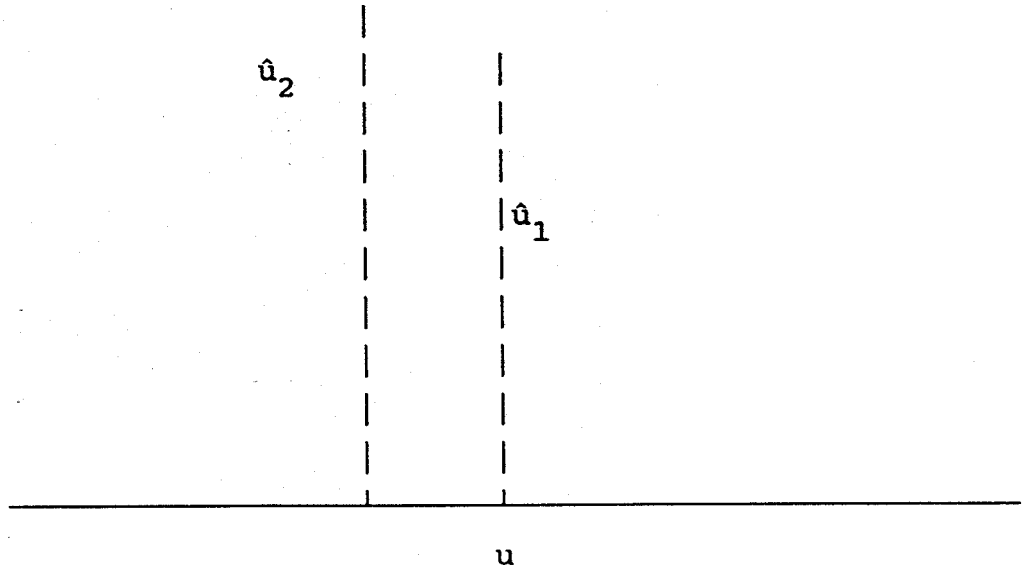
a. Centrado e insesgado.

Decimos que un estimador es centrado o insesgado cuando la media de la distribución del estimador coincide con el valor del parámetro, denominándose sesgo a la diferencia entre el verdadero valor del parámetro menos la media de la distribución del estimador.

sesgo $\hat{u} = u - E[\hat{u}]$

Un estimador centrado se dice que es insesgado o que su sesgo es cero.

La propiedad de ser centrado no es por sí sola concluyente a la hora de decidir si un estimador es bueno o no. En la figura se ve un estimador centrado que sin embargo no parece ser más adecuado que el estimador sesgado que se representa a su lado.



b. Eficacia y Precisión.

Diremos que un estimador $\hat{u}_2$ es más eficiente que otro estimador $\hat{u}_1$ si se cumple que para cualquier tamaño muestral se cumple que:

$$\text{Var}(\hat{u}_2) \leq \text{Var}(\hat{u}_1)$$

llamándose precisión o eficacia de un estimador a la inversa de la varianza de su distribución muestral.

$$\text{Efi}(\hat{u}_2) \leq \text{Efi}(\hat{u}_1)$$

denominándose eficacia relativa a:

$$ER(\hat{u}_2/\hat{u}_1)=\text{Efi}(\hat{u}_2)/\text{Efi}(\hat{u}_1)$$

Entre dos estimadores centrados se elegirá el más eficiente.

c. Error cuadrático medio.

Cuando tenemos dos estimadores con propiedades contrapuestas, se hace difícil elegir entre ellos. Para ello se utiliza el error cuadrático medio eligiendo aquel estimador que minimiza este valor.

Error Cuadrático Medio = ECM = $E\,[(u - \hat{u})^2]$

ECM (u) = $[\text{sesgo}\,(u)]^2$ + Var ($\hat{u}$).

5.5. Estimación estadística por intervalos de confianza.

a. Introducción.

En la práctica interesa no solamente dar una estimación de un parámetro sino, además, un intervalo que permita precisar la incertidumbre existente en la estimación.

A ese intervalo se le denomina ***intervalo de confianza*** y es un conjunto de valores en el que se incluye, con una probabilidad preasignada, llamada nivel de confianza, el verdadero valor del parámetro de la población.

A los límite inferior y superior del intervalo de confianza se les denomina límites de confianza.

Llamaremos ***nivel de confianza*** (1 - α) a la probabilidad de que la afirmación que se hace sobre el verdadero valor del parámetro (que se encuentre entre los límites de confianza) sea cierto.

Por ejemplo, sea un estadístico m y μ_m la media y σ_m la desviación típica (error típico) de la distribución muestral de dicho estadístico. Si la distribución muestral de m es aproximadamente normal es de esperar que al extraer muestras, el estadístico m se encuentre en los intervalos de confianza:

$\mu_m \pm \sigma_m$ el 68,27% de las veces

$\mu_m \pm 2\sigma_m$ el 95,45% de las veces

$\mu_m \pm 3\sigma_m$ el 99,73% de las veces

Análogamente puede confiarse en encontrar μ_m en los intervalos m $\pm$ $z_c\sigma_m$ (z_c =1, 2 y 3) en las mismas circunstancias anteriores.

Por otra parte m $\pm$ 1,96 σ_m y m $\pm$ 2,58 σ_m son los límites del 95% y 99% de nivel de confianza para μ_m.

Los valores de z_c se llaman ***coeficientes de confianza*** y una gama significativa de los mismos es la siguiente:

Nivel confianza	99,73	99	98	96	95	90	50
z_c	3	2,58	2,33	2,05	1,96	1,64	0,67

b. Límites de confianza para la media poblacional.

Si el estadístico bajo estudio es la media muestral $\bar{x}$ los límites de confianza del 95% y 99% para la estimación de la media de la población μ vienen dados por $\bar{x} \pm 1{,}96 s_{\bar{x}}$ y $\bar{x} \pm 2{,}58 s_{\bar{x}}$. Los límites de confianza generales son:

$$\bar{x} \pm z_c \sigma_{\bar{x}} = \bar{x} \pm z_c \frac{\sigma}{\sqrt{n}}$$

En general σ es desconocido por lo que se utiliza el estimador muestral $\hat{s}$ o s. Esta aproximación es válida para n>30.

Para n<30 la aproximación no es buena y debe aplicarse los límites de confianza dados por la distribución "t" de Student.

Para el caso de la distribución "t" si $-t_{1-\alpha/2}$ y $t_{1-\alpha/2}$ son los valores de t para los que el $(\alpha/2)$% del área total se encuentra en cada cola de la distribución, un intervalo de confianza bilateral del (1-α)% para t será:

$$-t_{1-\frac{a}{2}} < \frac{\bar{x} - \mu}{s} \sqrt{n-1} < t_{1-\frac{\alpha}{2}}$$

por lo que μ se encuentra en el intervalo:

$$\bar{x} - t_{1-\frac{a}{2}} \frac{s}{\sqrt{n-1}} < \mu < \bar{x} + t_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n-1}}$$

El valor $t_{1-\alpha/2}$ representa el valor del percentil 1-$\alpha/2$ mientras que $t_{\alpha/2} = -t_{1-\alpha/2}$ tratándose de una distribución simétrica, representa el valor del percentil $\alpha/2$.

En general los límites de confianza para la media de la población μ (cuando n<30) vienen dados por:

$$\bar{x} \pm t_c \frac{s}{\sqrt{n-1}}$$

siendo $\pm t_c$ los valores críticos que dependen del nivel de confianza utilizado y del número de grados de libertad v=n-1.

♦ ***Ejemplo*:**

Las medidas de los diámetros de una muestra de 100 ejes dieron una media de 0,8 mm y una desviación típica de 0,03 mm. Hallar los límites de confianza del 95% y 99% para el diámetro medio de los ejes de la producción.

Los límites de confianza del 95% son:

$$\bar{x} \pm 1{,}96\frac{\sigma}{\sqrt{n}} = 0{,}8 \pm 1{,}96 \times 0{,}03 \times \frac{1}{\sqrt{100}} = 0{,}8 \pm 0{,}0058 mm.$$

Los límites de confianza del 99% son:

$$\bar{x} \pm 2{,}58\frac{\sigma}{\sqrt{n}} = 0{,}8 \pm 2{,}58 \times 0{,}03 \times \frac{1}{\sqrt{100}} = 0{,}8 \pm 0{,}0077 mm.$$

c. Límites de confianza para la desviación típica poblacional.

En el caso de que el estadístico muestral fuera la desviación típica s, hemos de aplicar, en general la distribución χ^2 para conocer el verdadero valor σ de la población.

Si $\chi^2_{\frac{\alpha}{2}}$ y $\chi^2_{1-\frac{\alpha}{2}}$ son los valores para los que el $\frac{\alpha}{2}$ %

del área se encuentra en cada cola de la distribución, el intervalo de confianza 1-α viene dado por:

$$\chi^2_{\frac{\alpha}{2}} < \frac{ns^2}{\sigma^2} < \chi^2_{1-\frac{\alpha}{2}}$$

de donde se deduce que σ se encuentra en el intervalo:

$$\frac{s\sqrt{n}}{\sqrt{\chi^2_{1-\frac{\alpha}{2}}}} < \sigma < \frac{s\sqrt{n}}{\sqrt{\chi^2_{\frac{\alpha}{2}}}}$$

con un (1-α)% de nivel de confianza.

En el caso particular de que el tamaño de muestra n sea superior a 100 podemos basarnos en el hecho de que la distribución de las desviaciones típicas muestrales $s_1 s_2 ... s_n$ es normal con desviación típica $s/\sqrt{2n}$ por lo que la verdadera desviación típica de la población se encuentra en el intervalo:

$$s \pm z_c \frac{s}{\sqrt{2n}}$$

siendo Z_c el valor crítico utilizado anteriormente para diferentes niveles de confianza.

♦ ***Ejemplo*:**

La desviación típica de las duraciones de una muestra de 200 tubos es de 100 h.

Hallar los límites de confianza del 95% para la desviación típica de la población.

Dichos límites vienen dados, para el estadístico s, por:

$$s \pm z_c \frac{s}{2n}$$

Por tanto los límites de confianza son:

$$100 \pm 1{,}96 \frac{100}{\sqrt{400}} = 100 \pm 9{,}8$$

por lo que se puede esperar con el 95% de confianza que la desviación típica de la población se encuentre entre 90,2 y 109,8 horas.

d. Límites de confianza para proporciones poblacionales.

Si el estadístico m es la proporción P (por ejemplo porcentaje defectuoso) los límites de confianza para p (proporción defectuosa poblacional) vienen dados por:

$$P \pm z_c \sqrt{\frac{pq}{n}}$$

siendo P el porcentaje defectuoso de la muestra de tamaño n. (El valor de p utilizado es su estimador P).

e. Límites de confianza para las distribuciones diferencia de dos medias o proporciones.

Si m_1 y m_2 son dos estadísticos con distribuciones muestrales aproximadamente normales, los límites de confianza para la distribución diferencia de los parámetros correspondientes a m_1 y m_2 vienen dados por:

$$m_1 - m_2 \pm z_c \sigma_{m_1 - m_2} = m_1 - m_2 \pm z_c \sqrt{\sigma_{m_1}^2 + \sigma_{m_2}^2}$$

Los límites de confianza para la distribución suma de los parámetros poblacionales son:

$$m_1 + m_2 \pm z_c \sigma_{m_1 + m_2} = m_1 + m_2 \pm z_c \sqrt{\sigma_{m_1}^2 + \sigma_{m_2}^2}$$

Si el estadístico m es la media muestral $\bar{x}_1$, los límites para la distribución diferencia de dos medias poblacionales vienen dados por:

$$\bar{x}_1 - \bar{x}_2 \pm z_c \sigma_{\bar{x}_1 - \bar{x}_2} = \bar{x}_1 - \bar{x}_2 \pm z_c \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

(Los valores de σ_1 y σ_2 se estiman a partir de s_1 y s_2 para n>30).

Para el caso en que el estadístico sea el porcentaje defectuoso muestral P, los límites para la distribución diferencia de dos medias (porcentajes defectuosos) correspondientes a sendas poblaciones distribuidas, como de costumbre, binomialmente vienen dados por:

$$P_1 - P_2 \pm z_c \sigma_{P_1 - P_2} = P_1 - P_2 \pm z_c \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

donde P_1 y P_2 son los porcentajes defectuosos muestrales de dos muestras extraídas de poblaciones cuyos porcentajes defectuosos son p_1 y p_2 (cuyos estimadores son P_1 y P_2 para n>30).

- ***Ejemplo*:**

Una muestra de 100 bombillas del fabricante F_1 dieron una vida media de 1.500 horas y una desviación típica de 70 horas. Otra muestra de 200 bombillas del fabricante F_2 dieron una vida media de 1.400 horas con una desviación típica de 90 horas. Hallar los límites de confianza del 95% para la distribución diferencia de las vidas medias de las fabricaciones F_1 y F_2.

Los límites del 95% vienen dados por:

$$\bar{x}_{F_1} - \bar{x}_{F_2} \pm z_c \sqrt{\frac{\sigma_{F_1}^2}{n_1} + \frac{\sigma_{F_2}^2}{n_2}} = 1500 - 1400 \pm 1{,}96 \sqrt{\frac{70^2}{100} + \frac{90^2}{200}} = 100 \pm 1{,}96 \times 9{,}46 = 100 \pm 18{,}54$$

f. Error probable.

Los límites de confianza del 50% de los parámetros de la población correspondientes a un estadístico m son $m \pm 0{,}674\sigma_m$. La cantidad $0{,}674\,\sigma_m$ se conoce como error probable.

6. Contraste de hipótesis.

6.1. Introducción.

Contrastar una hipótesis estadísticamente es juzgar si cierta propiedad supuesta para una población es compatible con lo observado en una muestra de ella.

Por ejemplo, consideremos un proceso de fabricación que en condiciones correctas produce componentes cuya resistencia eléctrica se distribuye normalmente con media 20 Ohm y desviación típica 0,5 Ohm.

A veces, y de forma imprevisible, el proceso se desajusta, produciendo un aumento o disminución de la resistencia media de los componentes, pero sin variar la desviación típica. Para contrastar si el proceso funciona correctamente se toma una muestra de cinco unidades y se mide su resistencia resultando 22,2; 21; 18,8; 21,5; 20,5. ¿Podríamos concluir con estos datos que el proceso está desajustado?

Para responder a esta pregunta podemos razonar como sigue: si el proceso está bien ($\mu = 20$) haciendo un cambio de variable:

$$z = \frac{\bar{x} - 20}{0{,}5 / \sqrt{5}}$$

z es una variable N(0,1).

Por tanto, si al introducir en esta expresión el valor de $\bar{x}$ se obtiene un valor "razonable" (por ejemplo, entre $\pm$ 2) concluiremos que no hay evidencia de que el proceso está desajustado.

Por el contrario, si este valor es muy extremo, la diferencia observada entre $\bar{x}$ y 20 será demasiado grande para atribuirla al azar, por lo que concluiremos que el proceso está desajustado.

En este caso, $\bar{x}$ = 20,8 conduce a z = 3,58, que es un valor muy extremo. Por tanto, es razonable pensar que el proceso está efectivamente desajustado.

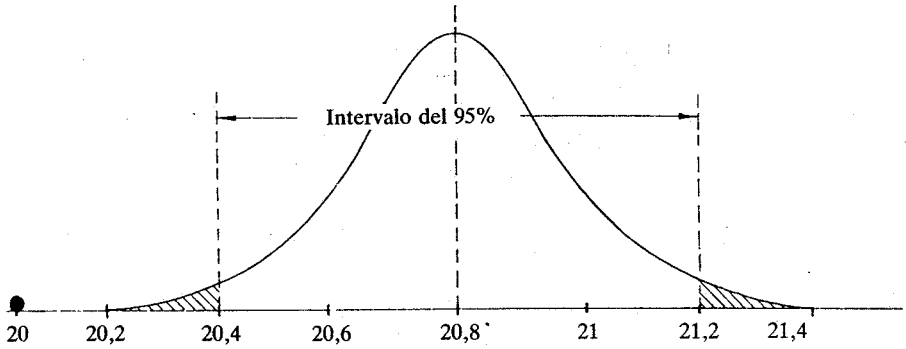
Un enfoque alternativo, que conduce al mismo resultado, es construir un intervalo de confianza para la media del proceso. En este caso:

$$m \in \left(20{,}8 \pm z_{a/2} \cdot \frac{0{,}5}{\sqrt{5}}\right)$$

$$m \in \left(20{,}8 - z_{a/2} \cdot 0{,}22; 20{,}8 + z_{a/2} \cdot 0{,}22\right)$$

La distribución de confianza que genera todos estos intervalos es N(20,8; 0,22) representada en la siguiente figura.

Al representar en ella el valor 20, que corresponde a las condiciones normales de funcionamiento, observamos que este valor no estará incluido en los intervalos de confianza habituales con α = 0,05 ó α = 0,01. Por tanto, tenderemos a pensar que el proceso se ha desajustado y que la media de la distribución que ha generado la muestra es 20 + δ, siendo 0,8 la mejor estimación de δ.



Distribución de confianza para la media de una población normal.

Este ejemplo pone de manifiesto los rasgos principales de un contraste de hipótesis, así como su estrecha relación con la estimación por intervalos.

6.2. Contrastes de significación.

a. Tipos de hipótesis.

Llamaremos ***hipótesis estadística*** a una suposición que determina, parcial o totalmente, la distribución de probabilidad de una variable aleatoria. Estas hipótesis pueden clasificarse en dos grupos, según que:

1. Especifiquen un valor concreto o un intervalo para los parámetros del modelo.

2. Determinen el tipo de distribución de probabilidad que han generado los datos.

Un ejemplo del primer grupo es la hipótesis de que la media de una variable es 10, y del segundo que la distribución es normal.

Aunque la metodología para realizar el contraste de hipótesis es análoga en ambos casos, distinguir ambos tipos de hipótesis es importante, porque muchos problemas de contraste de hipótesis respecto a un parámetro son en realidad problemas de estimación, que tienen una respuesta más clara dando un intervalo de confianza (o conjunto de intervalos de confianza) para dicho parámetro.

Sin embargo, las hipótesis respecto a la forma de la distribución pertenecen a la fase de diagnóstico y validación del modelo y serán estudiadas en el capítulo siguiente.

Centrándonos en hipótesis del primer tipo, llamaremos ***hipótesis simples*** a aquellas que especifican un único valor para el parámetro, ***e hipótesis compuestas*** a las que especifican varios.

Llamaremos ***hipótesis nula***, H_0, a la hipótesis que se contrasta. El nombre de "nula" proviene de que H_0 representa la hipótesis que mantendremos a no ser que los datos indiquen su falsedad, y debe entenderse, por tanto, en el sentido de "neutra". La hipótesis H_0 nunca se considera probada, aunque puede ser rechazada por los datos. Por ejemplo, la hipótesis de que todos los elementos de las poblaciones A y B son idénticos, puede ser rechazada encontrando elementos de A y B distintos, pero no puede ser "demostrada" más que estudiando todos los elementos de ambas poblaciones, tarea que puede ser imposible.

Análogamente, la hipótesis de que dos poblaciones tienen la misma media puede ser rechazada fácilmente cuando ambas difieran mucho, analizando muestras suficientemente grandes de ambas poblaciones, pero no puede ser "demostrada" mediante muestreo (es posible que las medias difieran en δ, siendo δ un valor pequeño imperceptible en el muestreo).

b. Nivel de significación. Errores tipo I y II.

A la hora de evaluar una hipótesis, podemos cometer dos tipos de error:

1. ***Rechazar la hipótesis cuando es cierta***. Se denomina ***error de tipo I*** o ***"nivel de significación"***. La probabilidad del error de tipo I se designa por α y su complemento a 1 es decir $1-\alpha$, nivel de confianza o probabilidad de aceptar la hipótesis verdadera.

100α suele valer para los diferentes tipos de ensayos 10%, 5%, 1%, valores a los que corresponden niveles de confianza del 90, 95 y 99% de tomar la decisión adecuada.

2. ***No rechazar la hipótesis cuando es falsa***. Se denomina ***error de tipo II*** y su probabilidad es β.

Estos errores están definidos en términos de probabilidad y se pueden controlar a los valores que se deseen. Los posibles resultados son los siguientes:

	H es verdadera	***H es falsa***
Aceptamos H	Decisión Correcta $P = 1-\alpha$	Decisión Equivocada $P = \beta$
No aceptamos H	Decisión Equivocada $P = \alpha$	Decisión Correcta $P = 1-\beta$

La interpretación estadística del error tipo I es la siguiente: si el procedimiento se repitiera muchas veces sobre una población en la que $\mu = 20$ en 100 $(1-\alpha)$% de los casos se llegaría a la conclusión verdadera y en 100α% de las veces se concluiría con la decisión falsa de que $\mu \neq 20$.

La interpretación estadística del error tipo II es como sigue: si el procedimiento se repitiera numerosas veces sobre una población en la que ciertamente $\mu \neq 20$ en 100β% de las veces se llegaría a la conclusión de que $\mu = 20$ y en 100 $(1-\beta)$% se tomaría la decisión verdadera de que $\mu \neq 20$.

6.3. Aplicación a la distribución normal. Ensayos de una y dos colas.

Supongamos que con una hipótesis dada, la distribución muestral de un estadístico s es una distribución normal con media μ_s y desviación típica σ_s. Entonces la distribución de la variable tipificada dada por:

$$z = \frac{s - \mu_s}{\sigma_s}$$

es una distribución normal de media 0 y varianza 1.

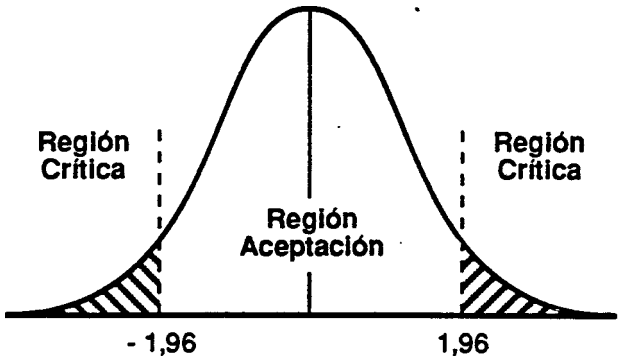
Con un 95% de confianza podemos esperar que el valor de z de una muestra obtenida se encontrará entre -1,96 y +1,96.

Si al elegir la muestra objeto del estudio el valor de z encontrado estuviera fuera de este margen, la probabilidad de que la hipótesis planificada fuera verdadera sería de un 5%.

Se puede decir que este valor de z difiere significativamente del que cabía esperar bajo la hipótesis dada y rechazaríamos la hipótesis.

El área de 0,05 representa la probabilidad de rechazar la hipótesis siendo verdadera.

El conjunto de los valores de z que se encuentran fuera del rango -1,96 a 1,96 se llama región crítica y los que se encuentran dentro, región de aceptación de la hipótesis o región no significativa.



Como resumen de lo expuesto se actuará de la siguiente forma:

- Si la z obtenida en el ensayo para el estadístico de que se trate (media o desviación típica) está fuera de la zona -1,96 a 1,96 se rechaza la hipótesis al nivel de significación del 5%.

- Se acepta la hipótesis o no se toma decisión alguna en caso contrario.

Lo que hasta ahora se ha citado se refería a los valores de z correspondientes a los dos extremos de la distribución (ensayos bilaterales o de dos colas). Existe pues una situación de indiferencia que hace necesario planificar el ensayo en estas condiciones. Si se desea conocer si un proceso se ha deteriorado se planteará la hipótesis nula de que no existe variación actual respecto a la situación histórica y los resultados del muestreo confirmarán si el proceso ha

sufrido o no variación que a su vez puede ser positiva (mejora del proceso) o negativa (degradación del proceso).

Existen, sin embargo, situaciones en las que se ensayan hipótesis sobre si un proceso nuevo es mejor que el existente. Es evidente que no perseguimos la condición disyuntiva sino la confirmación o no de que el proceso actual es mejor que el histórico lo que quiere decir que realizaremos un ensayo unilateral o de una cola en cuyo caso la región crítica se encuentra a un lado de la distribución con un área igual al nivel de significación con que se establezca el ensayo.

En la tabla siguiente se muestran los valores de z para ensayos de una y dos colas y niveles de significación del 0,1; 0,05 y 0,01.

Nivel de significación	0,10	0,05	0,01
Valores críticos de z para ensayos de una cola	-1,28 ó 1,28	-1,645 ó 1,645	-2,33 ó 2,33
Valores críticos de z para ensayos de dos colas	-1,645 y 1,645	-1,96 y 1,96	-2,58 y 2,58

6.4. Curva característica de operación (oc). Curva de potencia.

Es evidente que en los ensayos de hipótesis se tiende a tomar decisiones con un mínimo de riesgos o errores tipos I y II a que ya nos hemos referido.

Minimizar el error tipo I de rechazar una hipótesis que debiéramos aceptar por ser cierta, se consigue eligiendo adecuadamente el nivel de significación del ensayo. en cuanto a reducir al máximo el riesgo de aceptar hipótesis que debieran rechazarse por ser falsas, error tipo II, se conseguirá no aceptando nunca las hipótesis.

En la práctica el juego de aceptación-rechazo se resuelve mediante el empleo de las curvas características de operación o curvas OC cuya construcción y utilización vamos a exponer partiendo de un ejemplo práctico.

Supongamos que en un proceso de fabricación de lámparas se consigue una vida media de estas de 2.300 horas con una desviación típica de 45 horas. Estudios hechos en la competencia nos mueven a incrementar la vida media con un nuevo proceso de fabricación.

Deberemos realizar, una vez en marcha el nuevo proceso, un estudio que permita:

- Diseñar una regla por la que se pueda decidir rechazar el proceso clásico con un nivel de significación de 0,01 después de realizar un ensayo con 81 lámparas.
- En las condiciones anteriores, conocer la probabilidad con que se aceptaría el proceso histórico cuando en realidad el nuevo proceso incrementa significativamente la vida media a 2.320 horas, en el supuesto de que no ha variado la desviación típica.

Planteado el problema, fijemos las premisas del mismo.

$H_0: \mu$ = 2.300 horas y el nuevo proceso no representa mejora alguna.

$H_1: \mu$ > 2.300 horas y el nuevo proceso mejora significativamente el histórico.

Se trata según se citó anteriormente de un ensayo unilateral y por tanto, al nivel de significación de 0,01, se rechaza la hipótesis nula si el valor de z obtenido en el ensayo supera el valor z = 2,33, aceptándose en caso contrario.

El valor z de la distribución normal tipificada vale:

$$z = \frac{\bar{x} - m}{\frac{s}{\sqrt{n}}} = \frac{\bar{x} - 2.300}{\frac{45}{\sqrt{81}}}$$

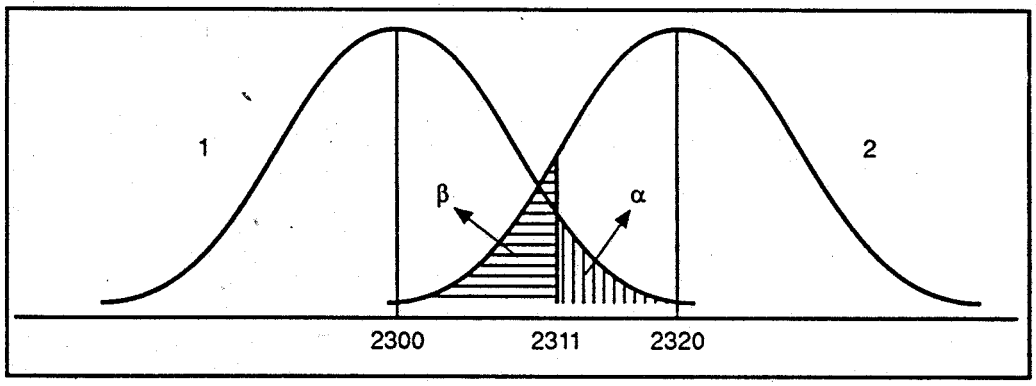
$$\bar{x} = 2.300 + 5z$$

Si z > 2,33 para que exista una mejora significativa (rechazo de la hipótesis nula)

$$\bar{x} > 2.300 + 5 \times 2,33 = 2311,65$$

Así pues el ensayo se configura como sigue:

- Se rechaza H_0 si la vida media de 81 lámparas supera las 2.311 h.
- Se acepta H_0 (o no se toma decisión) en caso contrario.



En la figura anterior se han representado las distribuciones correspondientes a las hipótesis:

$$H_0: \mu = 2.300 \text{ h.}$$

$$H_1: \mu = 2.320 \text{ h.}$$

La probabilidad de aceptar el proceso histórico cuando realmente la nueva vida media es 2.320 h., viene representada por el área β. Para calcular el valor de dicha probabilidad calcularemos el valor 2.311 en unidades tipificadas.

$$z = \frac{2.311 - 2.320}{5} = -\frac{9}{5} = -1,8$$

Por tanto β vendrá dada por el área de la curva 2 a la izquierda de z=-1,8 o sea $\beta = 0,0359$ o sea existe un 3,59% de probabilidades de aceptar el proceso histórico en el que $\mu = 2.300$ y un 96,41% de probabilidades de aceptar el nuevo proceso.

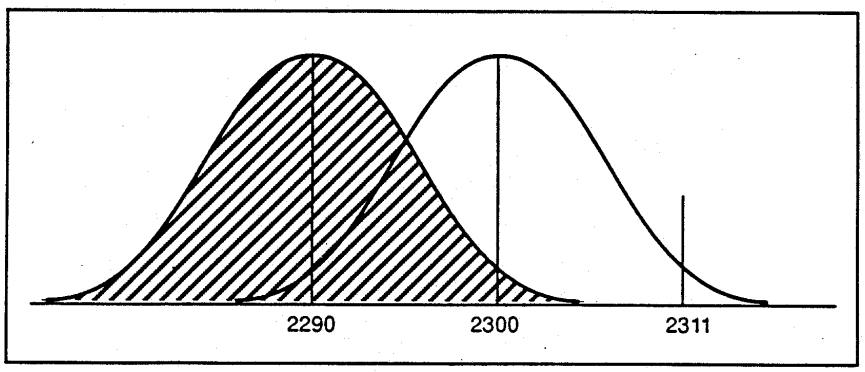
De forma análoga a como hemos actuado para una duración media de 2.320 h., podemos hacer extensivo el proceso para otras hipótesis de vidas medias correspondientes al nuevo proceso con lo que calcularemos y representaremos una curva β, μ que llamaremos curva característica de operación o curva OC.

μ	β	**1-** β
2290	1	0
2300	0,9890	0,0139
2310	0,5793	0,4207
2320	0,0359	0,9641
2330	0	1
2340	0	1

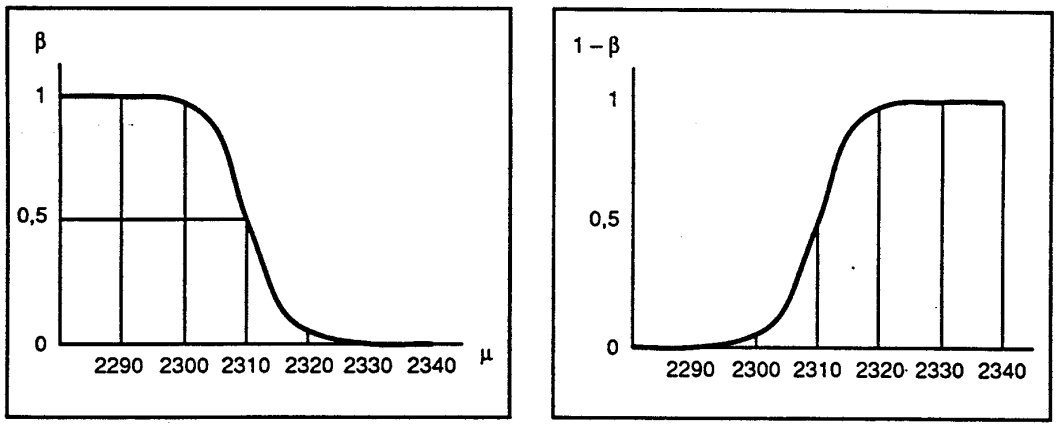
Por ejemplo (ver siguiente figura) si $\mu = 2290$ h., el valor calculado 2311 en unidades tipificadas vale:

$$z = \frac{2.311 - 2.290}{5} = 4,2$$

por lo que β (área a la izda. de z = 4,2) vale 1.



De acuerdo con los valores tabulados β y μ la representación de la curva OC se refleja en la gráfica izquierda de la siguiente figura. En esta curva puede observarse que la probabilidad de seguir con el proceso histórico si el nuevo ofrece una vida media inferior a 2.300 h., es prácticamente 1. Después la curva cae a valores de un 3% para una vida media de 2.320 h.



El gráfico (1- β), se denomina curva de potencia de la decisión indicando la aptitud para rechazar hipótesis falsas. En la gráfica derecha de la figura anterior se aprecia esta característica.

Podríamos desarrollar las curvas características operacionales para contrastes a una cola siguiendo el mismo método que para el de dos colas. Estas serían distintas puesto que a pesar de que la probabilidad de error de tipo I es la misma, la probabilidad β del error de tipo II varía dependiendo de que se use un contraste a una o dos colas.

En algunos problemas, tenemos información para poder decir que en caso de ser la media verdadera de la población distintas del valor de la hipótesis, este valor estará con toda seguridad por encima (o por debajo) de él.

Por ejemplo, un nuevo material con supuesta mayor resistencia, tendrá una media igual o mayor que la del material actual. Tal información nos ayudará a seleccionar un contraste a una o dos colas de tal forma que haga β lo más pequeño posible.

Si analizamos las curvas operativas sacaremos las siguientes conclusiones:

- Se utilizará un contraste a dos colas si:
  - No existe información previa de la situación de la verdadera media poblacional.
  - Queremos detectar una media poblacional < ó > que la establecida en la hipótesis (μ_0).
- Se utilizará un contraste de una sola cola con el riesgo α a la derecha si se sospecha que:
  - En caso de H_0 no ser cierto, la verdadera media $> \mu_0$.
  - Valores de media poblacional $< \mu_0$ son aceptables y sólo queremos detectar si la media poblacional $> \mu_0$.
- Se utilizará un contraste de una sola cola con el riesgo α a la izquierda si se sospecha que:
  - En caso de H_0 no ser cierta, la verdadera media $< \mu_0$.
  - Valores de media poblacional $> \mu_0$ son aceptables y sólo queremos detectar si la media poblacional $< \mu_0$.

Cuando tenemos un contraste a dos colas

- Hipótesis Nula $H_0: \mu_0 = 30.0$
- Hipótesis Alternativa $H_1: \mu_0 \neq 30.0$

y cuando tenemos un contraste a una cola

- Hipótesis Nula $H_0: \mu_0 = 30.0$
- Hipótesis Alternativa $H_1: \mu_0 < 30.0$ α a la izquierda

  $H_1: \mu_0 > 30.0$ α a la derecha

6.5. Diferentes tipos de ensayos.

a. Contraste de hipótesis con una población.

El objetivo de los contrastes de hipótesis con una población es comparar los parámetros estadísticos de la población con otros dados y ver si existe evidencia para rechazar la hipótesis H_0.

Si no se puede ver la evidencia para rechazar H0, existen dos opciones:

- obtener más datos y repetir el contraste.
- asumir que H0 es verdadera.

Los pasos genéricos a seguir para realizar un contraste de hipótesis, son:

1. Enunciar la hipótesis.
2. Elegir el coeficiente de distribución a usar.
3. Definir un coeficiente de significación α.
4. Calcular la zona de aceptación de la prueba que dé como resultado la aceptación de la hipótesis.
5. Con los valores muestrales, calcular el coeficiente de distribución elegido.
6. Comparar este coeficiente con la zona de aceptación con objeto de aceptar o rechazar la hipótesis.

Algunos de los casos más comunes son:

- *Comparación de la media de una población, cuya desviación típica se conoce, con una valor dado.*

  Se utiliza en el estudio de procesos estables, en los que se dispone de información histórica sobre la población. El objetivo es determinar si la media del proceso ha cambiado, o no , mediante el análisis de una muestra.

- *Comparación de la media de una población, cuya desviación típica no es conocida, con un valor dado.*

  Se utiliza en el estudio de procesos inestables, o de procesos en los que no se dispone de información histórica de la población. El objetivo es determinar si la media del proceso es igual, mayor o menor que un valor establecido, mediante el análisis de una muestra.

- *Comparación de la desviación típica de una población, con un valor dado.*

  Se utiliza como medida de la variabilidad de un proceso, independientemente del valor de la media. El objetivo es determinar si la variabilidad de un proceso ha cambiado, o no, mediante el análisis de una muestra.

- *Comparación de la proporción de unidades defectuosas en la población, con un valor dado.*

Se utiliza en procesos con características a medir de tipo discreto. El objetivo es determinar si el proceso mejora, o no, mediante el análisis de una muestra.

Se revisan, a continuación, algunos ejemplos de los casos más indicados.

- ***Comparar la media de la población con un valor dado. Se conoce la desviación típica de la población.***

  Hipótesis H_0: $\mu = \mu_0$ ***con*** σ ***conocido =*** σ_0

  El estadístico de ensayo a utilizar:

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma_0}{\sqrt{n}}}$$

  y la distribución de z es N(0,1).

- ***Ejemplo*:**

Un suministrador fabrica tubos fluorescentes con una vida media $\mu = 1.500$ *h. y* $\sigma = 115$ *h. Se propone mejorar el proceso para aumentar la vida media y envía una muestra de 100 tubos en los que la vida media resulta de 1.550 h.*

¿Ha mejorado significativamente el proceso?

H_0: $\mu = 1.500$ h. y el proceso no ha mejorado

H_1: $\mu > 1.500$ h. y el proceso ha mejorado significativamente

$$z = \frac{\bar{x} - m}{\frac{s}{\sqrt{n}}} = \frac{1550 - 1500}{\frac{115}{\sqrt{100}}} = \frac{50}{11,5} = 4,34$$

Como el valor crítico de z_c para un nivel de confianza del 99% y ensayo de una cola es z=2,33 al ser z=4,34 se encuentra en la región crítica, rechazándose la hipótesis nula.

Conclusión.- El fabricante ha mejorado significativamente el proceso.

- ***Comparar la media de la población con un valor dado. No se reconoce la desviación típica de la población.***

  Hipótesis H_0: $\mu = \mu_0$ ***con*** σ ***desconocida.***

  El estadístico de ensayo es la variable t de "Student" con n - 1 grados de libertad

$$t = \frac{\bar{x} - \mu_0}{s_0}\sqrt{n}$$

  ya conocida y aplicable a la teoría exacta del muestreo (pequeñas o grandes muestras). Por otra parte:

$$s_0 = \sqrt{\frac{\sum (x - \bar{x})^2}{(n-1)}}$$

- ***Ejemplo*:**

Una máquina produce piezas con un espesor medio de 0,5 mm. Para determinar si la máquina continúa trabajando normalmente, se toma una muestra de N=15 piezas en las que se encuentra un espesor medio de 0,54 mm. con una desviación típica de 0,035. Ensayar la hipótesis de que la máquina funciona correctamente al nivel de significación del 0,01.

H_0: $\mu = 0{,}5$ h. y la máquina continúa funcionando bien

H_1: $\mu \neq 0{,}5$ h. y la máquina se ha desajustado.

$$t = \frac{\bar{x} - \mu}{s}\sqrt{n-1} = \frac{0{,}54 - 0{,}5}{0{,}035}\sqrt{14} = 4{,}27$$

Para un ensayo bilateral al nivel de significación del 0,01 se sigue la siguiente regla de decisión:

Se acepta H_0 si t se encuentra dentro del intervalo $-t_{0,995}$ a $t_{0,995}$ que con 14 grados de libertad corresponde a los valores -2,98 a 2,98, rechazándose en caso contrario.

Al ser 4,27>|2,98|, t se encuentra en la región crítica, rechazándose la hipótesis nula.

Conclusión: La máquina se ha desajustado.

- ***Comparar la proporción de unidades defectuosas de una población con un valor dado.***

  Hipótesis: H_0:p=p_0 donde p_0 es la proporción defectuosa de la población.

  Si P es el estadístico a tratar, sabemos que:

$$\mu_P = p_0 \quad \text{y} \quad \sigma_P = \sqrt{\frac{p_0(1-p_0)}{n}}$$

  y el valor de z es:

$$z = \frac{P - p_0}{\sqrt{\frac{p_0(1-p)_0}{n}}} \qquad \text{z se distribuye N (0,1)}$$

y si $P = \frac{x}{n}$

siendo x el número de piezas defectuosas de la muestra, queda:

$$z = \frac{x - np_0}{\sqrt{np_0(1-p)_0}}$$

- ***Ejemplo*:**

Un fabricante de productos electrónicos asegura que su producción está libre de fallos en su 90% en un período de 8 horas. En una muestra de 200 equipos, 160 permanecieron 8 horas sin fallo. ¿Es cierta la propaganda del productor?

Sea p la probabilidad de "equipo libre de fallos". Se debe decidir entre las hipótesis:

H_0: p=0,9 y la propaganda es correcta

H_1: p<0,9 y la propaganda es falsa

Tomando un nivel de significación de 0,01, z_1=-2,33 (ensayo de una cola), planteándose la siguiente alternativa:

a. la hipótesis es verdadera si la z encontrada es mayor de z_1

b. la hipótesis es falsa si el valor de z es menor que -2,33

Sabemos que:

$$\mu = np = 200 \times 0{,}9 = 180$$

$$s = \sqrt{npq} = \sqrt{200 \times 0{,}9 \times 0{,}1} = 4{,}2$$

El valor 160 de equipos buenos, en unidades tipificadas vale:

$$z = \frac{160 - 180}{4,2} = -4,73$$

menor que -2,33 luego se deduce, casi con certeza (99% de probabilidades de acertar) que existe una diferencia muy significativa entre lo que anuncia y sus resultados.

Conclusión: La propaganda del fabricante es falsa.

b. Contraste de hipótesis con dos poblaciones.

El objetivo de los contrastes de hipótesis con dos poblaciones es comparar los parámetros estadísticos de ambas poblaciones, mediante el análisis de las muestras, y ver si existe evidencia para rechazar la hipótesis propuesta H_0.

Si no existe evidencia para rechazar H_0 existen dos opciones:

- obtener más datos y repetir el contraste.
- asumir que H_0 es verdadera.

Los pasos genéricos a seguir para realizar un contraste de hipótesis con dos poblaciones son iguales a los indicados para el contraste de hipótesis con una población.

Algunos de los casos más comunes de contraste de hipótesis con dos poblaciones son:

- *Comparación de las medias de dos poblaciones, cuyas desviaciones típicas son conocidas.*

  Se utiliza, en general, para determinar si existen diferencias en las medias de las poblaciones obtenidas de dos procesos independientes y estables, en los que se dispone de información histórica.

- *Comparación de las medias de dos poblaciones, cuyas desviaciones típicas no son conocidas, pero se supone que son iguales.*

  Se utiliza, en general, para determinar si existen diferencias en las medias de las poblaciones obtenidas de dos procesos nuevos, en las que no se conocen sus desviaciones típicas, pero se cree que deben ser similares.

- *Comparación de las medias de dos poblaciones, cuyas desviaciones típicas no son conocidas, y no se supone que sean iguales.*

  Se utiliza, en general, en casos similares al anterior, cuando no se desea asumir el riesgo de que las desviaciones típicas sean iguales.

- *Comparación de las medias de dos poblaciones, cuyas muestras no son independientes.*

  Se utiliza, en general, cuando se desea establecer diferencias en las observaciones realizadas sobre una misma muestra, por dos individuos diferentes.

- *Comparación de las desviaciones típicas de dos poblaciones.*

  Se utiliza, en general, para determinar si existen diferencias en la variabilidad de dos procesos diferentes, o en un proceso a lo largo del tiempo.

– *Comparación de las proporciones de una característica determinada, de dos poblaciones.*

Se utiliza, en general, para determinar si un proceso es igual, mejor o peor que otro.

- ***Comparar medias, conocida la desviación típica de la población***

Hipótesis H₀: $\mu_1 = \mu_2$

Sean $\overline{x_1}$ y $\overline{x_2}$ las medias muestrales de dos muestras n_1 y n_2 extraídas de poblaciones con medias μ_1 y μ_2 y desviación típica conocida σ.

Como ya sabemos la distribución muestral de la diferencia de medias presenta los siguientes estadísticos:

$$\mu_{\bar{x}_1-\bar{x}_2} = \mu_1 - \mu_2 = 0$$

$$\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sigma\sqrt{\frac{n_1+n_2}{n_1 n_2}}$$

La variable tipificada, que sabemos se distribuye como una N(0,1), viene dada por:

$$z = \frac{\overline{x}_1 - \overline{x}_2 - \mu_{\bar{x}_1-\bar{x}_2}}{\sigma_{\bar{x}_1-\bar{x}_2}} = \frac{\overline{x}_1 - \overline{x}_2}{\sigma\sqrt{\frac{n_1+n_2}{n_1 n_2}}}$$

- ***Ejemplo*:**

El espesor medio de 50 piezas de una producción es de 50 mm., con desviación típica histórica de 1,5 mm. Del mismo tipo de producción se extraen al día siguiente 36 piezas con media 51 mm. ensayar la hipótesis de que la producción de donde proceden ambas muestras han sufrido un cambio.

H_0: $\mu_1 = \mu_2$ La diferencia de espesores no es significativa.

H_1: $\mu_2 > \mu_1$ Hay diferencia significativa entre los espesores.

Bajo la hipótesis H_0

$$\mu_{\bar{x}_1-\bar{x}_2} = 0 \qquad \sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{1{,}5^2}{50} + \frac{1{,}5^2}{36}} = 0{,}327$$

El valor de la variable tipificada z será:

$$z = \frac{\overline{x}_1 - \overline{x}_2}{\sigma_{\bar{x}_1-\bar{x}_2}} = \frac{50-51}{0{,}327} = -3{,}05$$

Con un ensayo unilateral al nivel 0,01 se rechaza la hipótesis pues el valor z = -3,05 es menor que el correspondiente -2,33.

Conclusión: Hay cambio significativo en la producción.

- ***Comparar medias. No se conocen las desviaciones típicas pero se suponen iguales.***

  Hipótesis $\mu_1 = \mu_2$ ***con*** $\sigma_1 = \sigma_2$

  Sean dos muestras de tamaños n_1 y n_2 con medias $\bar{x}_1$ y $\bar{x}_2$ y desviaciones típicas s_1 y s_2 respectivamente. Para ensayar la hipótesis H_0 de que provienen de la misma población se utiliza el valor de t dado por

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad \text{siendo} \qquad \sigma = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}}$$

  La distribución de "t" sabemos que es una "Student" con $\nu = n_1 + n_2 - 2$ grados de libertad.

- ***Ejemplo*:**

En una planta de acabados electrolíticos se desea conocer el efecto de un tratamiento especial del que se espera un mayor espesor del acabado. Para ello se eligieron 26 bandejas, la mitad tratadas con el nuevo procedimiento y la otra mitad con el procedimiento tradicional. En el primer caso se obtuvo un espesor de 5,2 micras con una desviación típica de 0,38. ¿Ha supuesto una mejora el procedimiento nuevo, sobre el tradicional?

Si μ_1 y μ_2 son las medias poblacionales, se planteará el ensayo en los siguientes términos

H_0: $\mu_1 = \mu_2$ y la diferencia de espesores se debe a causas aleatorias.

H_1: $\mu_2 > \mu_1$ y hay un cambio significativo en el proceso.

Bajo la hipótesis H_0

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad \sigma = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}}$$

$$= \sqrt{\frac{13 \times 0{,}39^2 + 13 \times 0{,}38^2}{13 + 13 - 2}} = 0{,}4$$

$$t = \frac{5{,}2 - 5}{0{,}4\sqrt{\frac{1}{13} + \frac{1}{13}}} = \frac{0{,}2}{0{,}16} = 1{,}27$$

Con un ensayo unilateral al nivel de significación 0,01, H_0 se acepta al ser $t_{0,99}$ (con 24 grados de libertad) igual a 2,49.

El "t" calculado 1,27 está pues dentro de la región de aceptación.

Conclusión: El nuevo tratamiento no ha supuesto un cambio significativo en el proceso.

- ***Comparar proporciones de una característica determinada.***

  Hipótesis: $p_1=p_2$

  Sean P_1 y P_2 las proporciones defectuosas de dos muestras de tamaños n_1 y n_2 extraídas de poblaciones con proporciones defectuosas p_1 y p_2.

  En el supuesto de que $p_1=p_2$, la distribución diferencia de las proporciones muestrales presenta los siguientes estadísticos:

$$\mu_{P_1-P_2} = P_1 - P_2 = 0$$

$$\sigma_{P_1-P_2} = \sqrt{p(1-p)\times\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

  donde p desconocido puede estimarse como la media ponderada:

$$P = \frac{n_1 P_1 + n_2 P_2}{n_1 + n_2}$$

  La variable tipificada z cuya distribución es N (0,1) vale

$$z = \frac{P_1 - P_2}{\sigma_{P_1-P_2}}$$

- ***Ejemplo*:**

Escogidas dos muestras de 300 y 100 piezas fabricadas por dos máquinas, con el mismo proceso de fabricación, se encuentran 15 y 4 piezas defectuosas respectivamente.

a. ¿es diferente la calidad de fabricación de ambas máquinas?

b. ¿es la máquina B mejor que la A?

H_0: $p_1=p_2$ y las dos máquinas producen la misma calidad.

H_1: $p_1>p_2$ y hay diferencia significativa entre las calidades de ambas máquinas.

$$P_1 = \frac{15}{300} = 0{,}05 \qquad P_2 = \frac{4}{100} = 0{,}04$$

La distribución diferencia tendrá los siguientes estadísticos

$$\mu_{P_1-P_2} = 0 \qquad p = \frac{300\frac{15}{300} + 100\frac{4}{100}}{300+100} = 0{,}0475$$

$$\sigma_{P_1-P_2} = \sqrt{0{,}0475 \times 0{,}9525 \times \left(\frac{1}{300} + \frac{1}{100}\right)} = 0{,}0245$$

de donde:

$$z = \frac{0{,}05 - 0{,}04}{0{,}0245} = 0{,}40$$

En un ensayo unilateral al nivel de significación 0,01, se acepta la hipótesis nula al ser el z calculado inferior a z_c cuyo valor sabemos que es 2,33.

Conclusión: No hay diferencia significativa entre las producciones de ambas máquinas.

- ***Comparar desviaciones típicas***

  Hipótesis H_0: $\sigma_1 = \sigma_2 = \sigma$

  Variación de s_1 y s_2 desconocida σ

  Variable aleatoria: F de SNEDECOR con $\begin{Bmatrix} n_1 - 1 \\ n_2 - 1 \end{Bmatrix}$ grados de libertad

$$F = \frac{(n_2 - 1)n_1}{(n_1 - 1)n_2}\frac{s_1^2}{s_2^2}$$

- ***Ejemplo*:**

De dos formulaciones de PVC (Plástico) se desea saber cual de las dos proporciona una mayor homogeneidad en el alargamiento del plástico una vez extraído. Para lo cual se realiza el siguiente ensayo:

Formulación "A"

- *Nº de ensayos realizados (tamaño de la muestra) n_1=13.*
- *Desviación típica de los alargamientos del plástico una vez extraído: s_1=41,32%.*

Formulación "B"

- *Nº de ensayos realizados (tamaño de la muestra) n_2=11.*
- *Desviación típica de los alargamientos del plástico una vez extraído: s_2=27,13%.*

¿puede el azar explicar estas diferencias?

Hacemos la "hipótesis nula" de que la homogeneidad del alargamiento obtenida a través de las dos formulaciones es la misma.

Utilizando la fórmula:

$$F = \frac{(n_2 - 1)}{(n_1 - 1)} \times \frac{n_1}{n_2} \times \frac{s_1^2}{s_2^2}$$

Distribución "F" de Snedecor con (n_1 - 1) grados de libertad para el numerador y (n_2 - 1) grados de libertad para el denominador

Sustituyendo datos, queda:

$$F = \frac{10}{12} \times \frac{13}{11} \times \frac{41{,}32^2}{27{,}13^2} = 2{,}284$$

Mirando en la tabla para la distribución "F" de Snedecor con 12 grados de libertad en el numerador y 10 grados de libertad en el denominador, se obtiene un valor de 2,91 (95%) y de 4,71 (99%).

Conclusión.- A pesar de la aparente diferencia entre las dos desviaciones típicas se obtiene una probabilidad elevada de que dichas diferencias puedan ser debidas al azar, debiéndose aceptar, en principio, la hipótesis inicial.

Por tanto, no tenemos evidencia de que la formulación "B" mejore significativamente la homogeneidad del alargamiento.

NOTA: Debe tenerse en la cuenta que las muestras han sido muy reducidas, pudiendo ser ésta la causa de no encontrar unas diferencias evidentes. Es aconsejable repetir las pruebas aumentando el tamaño de la muestra.

7. Análisis de regresión.

7.1. Introducción.

Definimos un ***análisis de regresión*** como un conjunto de técnicas que se utilizan para estudiar la relación entre una variable y otra u otras x_1, ..., x_n, y la predicción o pronóstico entre una variable, conocidos los resultados de otra.

Se denominará a como la variable dependiente y x como la variable independiente.

Ejemplos:

- Se quiere estudiar la demanda media de cierto artículo (var y) en función del tiempo (var x).

- Queremos estudiar el crecimiento de una planta en cm. (Var. Y) en relación con la cantidad de agua en riego (var x_1) y la cantidad de fertilizante (var x_n).

Antes de proceder al problema del ajuste, deben dibujarse los puntos representativos de los pares de valores, obteniéndose la denominada nube de puntos. El problema del ajuste consiste en la obtención de una curva que pase cerca de los puntos de la nube, y que se adapte lo mejor posible al conjunto de los mismos, por lo que deberá cumplir determinadas condiciones. Lo primero que deberá hacerse es elegir el tipo de curva que mejor se adapte a los datos disponibles.

La forma de la nube de puntos puede sugerir el ajuste de una recta, de una parábola de 2º grado, de una exponencial, de una hipérbola, etc.

Mediante el ajuste se consigue:

- Llegar a una curva de ecuación conocida.

- Facilitar las descripciones y comparaciones, sustituyendo curvas complicadas e inexpresivas por otras más sencillas.

- Calcular, mediante relaciones de tipo matemático, magnitudes que no pueden calcularse mediante la observación directa, y que constituyen elementos característicos del fenómeno en estudio.

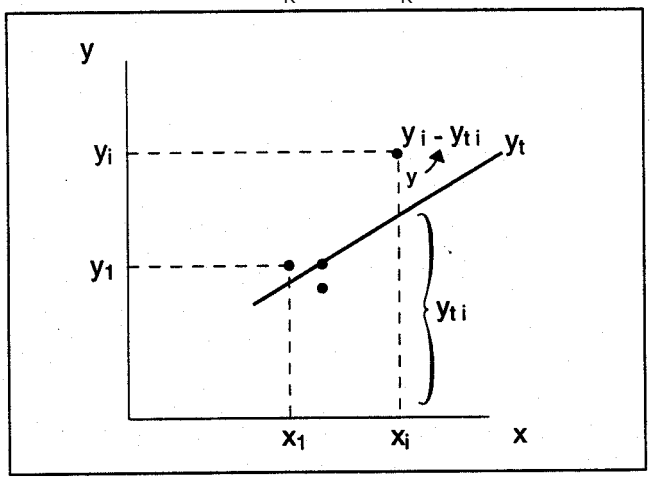
- Permitir las operaciones de interpolar y extrapolar.

Existen diversos métodos de ajuste pero nos limitaremos al método de los mínimos cuadrados.

7.2. Ajuste de una recta por el método de los mínimos cuadrados.

Supongamos conocidas las coordenadas de la nuble de puntos, la cual representamos gráficamente en la siguiente figura.

X_i	Y_i
X_1	Y_1
X_2	Y_2
.	
.	.
.	.
X_R	Y_R



Representación de la recta

La ecuación de la recta es:

$$y_{ti} = a + bx$$

Donde:

y_{ti} = ordenada teórica de la ecuación de la recta

y_i = ordenada observada, dada por la tabla

Para hallar la ordenada teórica de la recta, basta reemplazar x por x_i

$$y_{ti} = a + bx_i$$

Llamaremos desviación entre la ordenada teórica y la observada a la diferencia y_{ti} - y.

Si $y_{ti} > y_i$ la desviación será positiva

Si $y_{ti} < y_i$ la desviación será negativa

La ***condición mínima cuadrática*** consiste en encontrar los parámetros a y b de la ecuación de la recta de forma que la suma de los cuadrados de las desviaciones entre las ordenadas teóricas y las observadas sea mínima.

La condición mínimo cuadrática se expresará así:

$$\sum_{i=1}^{n} (y_{ti} - y_i)^2$$

que sustituyendo y_{ti} quedará:

$$\sum_{i=1}^{n} (a + bx_i - y_i)^2$$

El cálculo de los parámetros a y b requiere resolver un sistema de dos ecuaciones con dos incógnitas que recibe el nombre de ecuaciones normales.

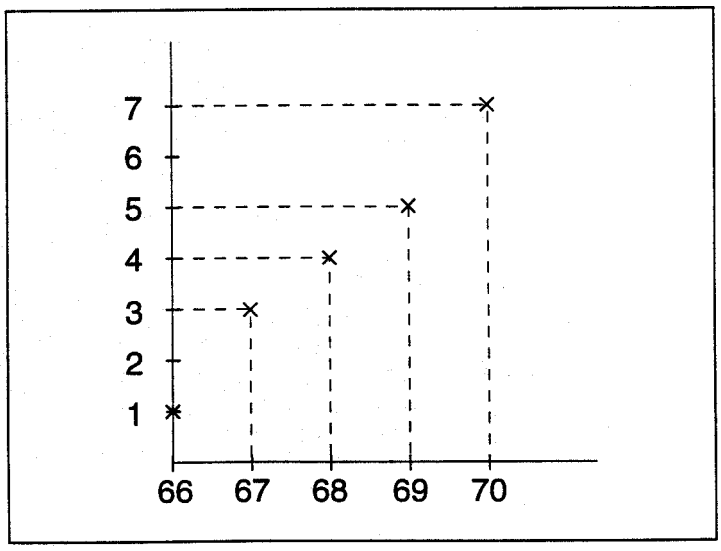
$$na + b\sum x_i = \sum y_i$$

$$a\sum x_i + b\sum x_i^2 = \sum x_i y_i$$

- ***Ejemplo*:**

De la siguiente serie cronológica ajustar una recta por el método de los mínimos cuadrados.

AÑOS	Y_i
1966	1
1967	3
1968	4
1969	5
1970	7



Nube de puntos

Elegimos un sistema de abscisas convencionales, al año 1966 le asignamos la abscisa $x_1=0$, al 1967 será $x_2=1$ ... para 1968 será $x_3=2$, para 1969 será $x_3=4$ y para 1970 será $x_4=5$.

Se tiene la siguiente tabla con los cálculos necesarios para el ajuste.

Años	y_i	x_i	$x_i y_i$	x_i^2
1966	1	0	0	0
1967	3	1	3	1
1968	4	2	8	4
1969	5	3	15	9
1970	7	4	28	16
	20	10	54	30

El sistema de ecuaciones normales es:

$$na + b\sum_1^n x_i = \sum_1^n y_i$$

$$a\sum_1^n x_i + b\sum_1^n x_1^2 = \sum_1^n x_i y_i$$

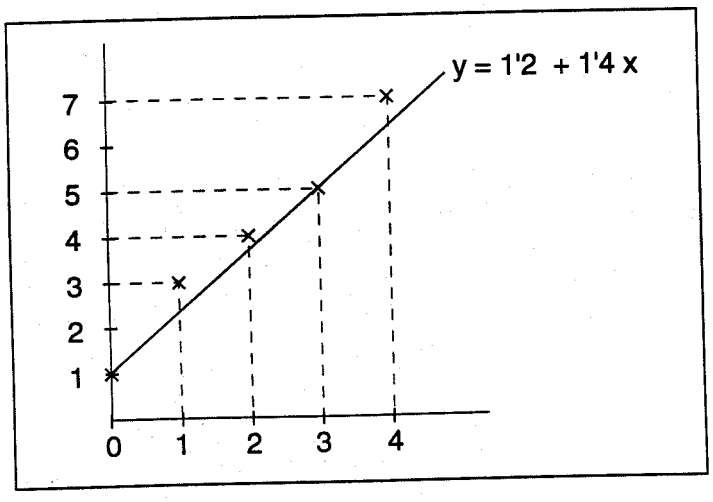
Para n=5

5a + 10b = 20 b = 1,4

10a + 20b = 54 a = 1,2

La ecuación de la recta queda:

y = 1,2 + 1,4 x



Ecuación de la recta sobre la nube de puntos

En la siguiente tabla aparecen las coordenadas teóricas, es decir, las ordenadas de la recta ajustada.

x_i	y_i	$y_{it} = 1{,}2 + 1{,}4\, x_i$	
0	1	$1{,}2 + 1{,}4 \cdot 0 =$	1,2
1	3	$1{,}2 + 1{,}4 \cdot 1 =$	2,6
2	4	$1{,}2 + 1{,}4 \cdot 2 =$	4
3	5	$1{,}2 + 1{,}4 \cdot 3 =$	5,1
4	7	$1{,}2 + 1{,}4 \cdot 4 =$	6,8
	20		20

7.3. Ajuste por mínimos cuadrados. Método abreviado.

Con objeto de abreviar la resolución del sistema de ecuaciones se traslada el eje de ordenadas, eligiendo un nuevo origen de abscisas O_x. La nueva abscisa se designa por "u" con lo que la ecuación de trasformación de abscisas será:

$$u = x - O_x$$

Para la abscisa convencional x_i obtendremos una abscisa en el nuevo sistema de ejes que designaremos por u_i.

$$u_i = x_i - O_x$$

El origen O_x se elige de forma tal que

$$\Sigma u_i = 0$$

a. Procedimiento abreviado para un número impar de años.

Consideramos una serie cronológica de la que tenemos información de un número impar de años.

Años	y_i	x_i	u_i	u_i^2	u_i^3	u_i^4	u_i^5
1965	2	0	-2	4	-8	16	-32
1966	3	1	-1	1	-1	1	-1
1967	5	2	0	0	0	0	0
1968	6	3	1	1	1	1	1
1969	4	4	2	4	8	16	32
	20		0	10	0	34	0

Hemos elegido un sistema de abscisas convencionales, así:

Al año 1965 le asignamos la abscisa $x_i = 0$

Al año 1966 le asignamos la abscisa $x_i = 1$

Si el origen de trabajo se elige en el año 1967 al que corresponde la abscisa convencional $x_i = 2$, hacemos:

$$O_x = 2$$

con lo que la ecuación de trasformación de abscisas es:

$$u = x - O_x = x - 2$$

Para $x_i = 0$ tenemos $u_i = 0 - 2 = -2$

Para $x_i = 1$ tenemos $u_i = 1 - 2 = -1$ etc.

obsérvese que:

$\Sigma u_i = 0$, $\Sigma u_i^3 = 0$ y $\Sigma u_i^5 = 0$

son siempre cero la suma de todas las potencias impares, pero la suma de las potencias pares es distinta de cero.

La ecuación de la recta en el nuevo sistema de coordenadas será:

$$y = a' + b' x$$

Las ecuaciones del ajuste serán:

$$n\, a' + b' \Sigma u_i = \Sigma\ y_i$$

$$a' \Sigma u_i + b' \Sigma u_i^2 = \Sigma\ u_i y_i$$

Al ser $\Sigma u_i = 0$ las ecuaciones normales se reducen a :

$$n\, a' = \Sigma\ y_i$$

$$b' \Sigma u_i^2 = \Sigma\ u_i y_i$$

que despejando obtenemos:

$$a' = \frac{\Sigma y_i}{n}$$

$$b' = \frac{\Sigma u_i y_i}{\Sigma u_i^2}$$

Si se quiere volver al antiguo sistema de coordenadas, tendríamos:

$$y = a' + b' (x - O_x)$$

- ***Ejemplo***:

Ajustar una recta por el método de mínimos cuadrados a la siguiente serie cronológica (utilícese el procedimiento abreviado).

Años	y_i	x_i	u_i	$u_i y_i$	u_i^2
1977	2	0	-2	-4	4
1978	3	1	-1	-3	1
1979	5	2	0	0	0
1980	6	3	1	6	1
1981	4	4	2	8	4
	20			7	10

$$a' = \frac{\Sigma y_i}{n} = \frac{20}{5} = 4$$

$$b' = \frac{\Sigma u_i y_i}{\Sigma u_i^2} = \frac{7}{10} = 0{,}7$$

La ecuación de la recta en el nuevo sistema de coordenadas es:

$y = 4 + 0{,}7\ u$

Teniendo en cuenta la ecuación de transformación de abscisas:

$u = x - 2$ con $O_x = 2$

La ecuación de la recta en el primitivo sistema de coordenadas es:

$y = 4 + 0{,}7\ (x-2) = 4 + 0{,}7\ x - 1{,}4$

$y = 2{,}6 + 0{,}7x$

b. Procedimiento abreviado para un número par de años.

Si el número de años fuera par, se elige la siguiente transformación de abscisas:

$u = 2\ (x - O_x)$

siendo O_x la abscisa promedio de las dos centrales.

♦ ***Ejemplo*:**

Ajustar una recta a la siguiente serie cronológica:

Años	y_i
1966	1
1967	2
1968	3
1969	4
	10

dispondremos los cálculos de la siguiente forma:

Años	y_i	x_i	u_i	$u_i y_i$
1966	1	0	-3	-3
1967	2	1	-1	-2
1968	4	2	1	4
1969	3	3	3	9
	10		0	8

Elegimos como origen $O_x = \frac{1+2}{2}$ (media aritmética de las dos abscisas centrales) $O_x = 1{,}5$

Mediante el cambio de variable:

$u = 2 (x - O_x)$

Para el valor:

$x_i = 0$ tenemos $u_i = 2 (0 - 1,5) = -3$

$x_i = 1$ tenemos $u_i = 2 (1 - 1,5) = -1$

Se consigue $\Sigma\ u_i = 0$

Aplicando las formulas ya conocidas tenemos:

$$b' = \frac{\Sigma u_i y_i}{\Sigma u_i^2} = \frac{8}{20} = 0,4$$

$$a' = \frac{\Sigma y_i}{n} = \frac{10}{4} = 2,5$$

La ecuación de la recta en el nuevo sistema de coordenadas es:

$y = 2,5 + 0,4\ u$

Volviendo al primitivo sistema de coordenadas teniendo en cuenta que $u = 2 (x - 1,5)$

$y = 2,5 + 0,4 \cdot 2 (x - 1,5) = 1,9 + 0,8\ x$

7.4. Medida de la precisión del ajuste. Bondad de ajuste.

Cuando se han ajustado varias funciones por el mismo o distintos procedimientos se presenta el problema de decidir cuál de ellos se adapta mejor a los datos observados. Para poder juzgar objetivamente se han ideado diversas medidas estadísticas.

Pero debe advertirse que no hay que dejarse influir demasiado por los resultados que ofrezcan estas medidas, pues algunas adaptaciones que parecen muy buenas tal vez no hayan eliminado las intervenciones casuísticas.

En los ajustes de rectas se suele utilizar como medida de la bondad de ajuste la denominada varianza residual, cuya fórmula es:

$$S_r^2 = \frac{\sum_{1}^{n} (y_i - yub_{ti})^2}{n}$$

Su raíz cuadrada es el error típico del ajuste o precisión del ajuste.

$$S_r = \sqrt{\frac{\sum_{1}^{n} (y_i - y_{ti})^2}{n}}$$

Sabiendo que:

y_i = ordenada observada

y_{ti} = ordenada teórica de la curva de ajuste

Esta medida sirve para en el caso de tener dos rectas de ajuste se extrapolará en aquella que su error típico sea menor.

- Formulas prácticas para calcular la varianza residual para la recta

$$S_r = \frac{\Sigma y_i^2 - a\Sigma y_i - b\Sigma x_i y_i}{n}$$

♦ ***Ejemplo*:** (continuación del primer ejemplo)

$$S_r = \frac{100 - 1{,}2 \cdot 20 - 1{,}4 \cdot 54}{5} = 0{,}08$$

7.5. Correlación.

La correlación estudia los problemas referentes a la variación conjunta de dos variables, su intensidad y su sentido (positivo o negativo).

a. Coeficiente de correlación de Pearson.

Tenemos:

- La varianza de los y_i explicada por la regresión se representa por S_{yk}^2 y es:

$$S_{yk}^2 = \frac{\sum_1^n (y_{ti} - \bar{y})^2}{n}$$

- La varianza residual, representada por S_r^2 y es:

$$S_r^2 = \frac{\sum_1^n (y_i - y_{ti})^2}{n}$$

- La varianza total, representada por S_y^2 y es la suma de la varianza residual y la varianza de los y_i

$$S_Y^2 = S_r^2 + S_Y^2 = \sum_1^N (y_I - \bar{y})^2$$

El coeficiente de correlación será:

$$r = \sqrt{1 - \frac{S_r^2}{S_Y^2}}$$

Para el caso de la ecuación de la recta de regresión se puede utilizar otra fórmula que resume mucho los cálculos, no será necesario calcular los valores estimados y_{ti}.

Por ser

$$S_r^2 = S_y^2 - \frac{S_{x^2y}}{S_{x^2}}$$

el coeficiente de correlación queda:

$$r = \frac{S_{xy}}{S_x S_Y}$$

donde:

S_{xy} = Covarianza de x e y, $S_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n}$

S_x^2 = Varianza de x, $S_x^2 = \frac{\sum_1^n x_i^2 - n\bar{x}^2}{n}$

S_y^2 = Varianza de y, $S_y^2 = \frac{\sum_1^n y_i^2 - n\bar{y}^2}{n}$

- ***Ejemplo*:** con los datos del ejemplo inicial.

$$\bar{x} = \frac{10}{5} = 2\ , \qquad \bar{y} = \frac{20}{5} = 4$$

$$S_{xy} = \frac{14}{5}$$

Años	y_i	x_i	y_i^2	x_i^2	$(x_i - \bar{x})(y_i - \bar{y})$
1966	1	0	1	0	6
1967	3	1	9	1	1
1968	4	2	16	4	0
1969	5	3	25	9	1
1970	7	4	49	16	6
	20	10	100	30	14

$$S_x^2 = 30 - 5 \bullet 2^2 = \frac{10}{5}$$

$$S_y^2 = 100 - 5 \bullet 4^2 = \frac{20}{5}$$

$$r = \frac{\frac{14}{5}}{\sqrt{\frac{10}{5}}\sqrt{\frac{20}{5}}} = 0{,}9899$$ => Correlación perfecta y directa

7.6. Interpretación del coeficiente de correlación.

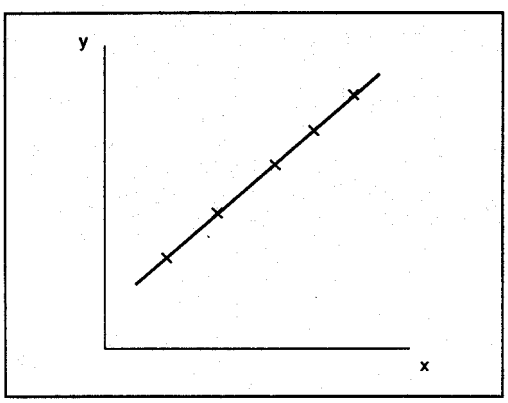
a. Caso 1º. Extremos.

Cuando $S_r^2 = 0$ todos los puntos están sobre la recta de regresión. Los valores observados coinciden con los teóricos, indica dependencia.

En este caso el coeficiente de correlación de Pearson varía entre $-1 \leq r \leq 1$

$$r = \sqrt{1 - \frac{S_r^2}{S_y^2}} = \sqrt{1 - \frac{0}{S_y^2}} = \sqrt{1} = \pm 1$$

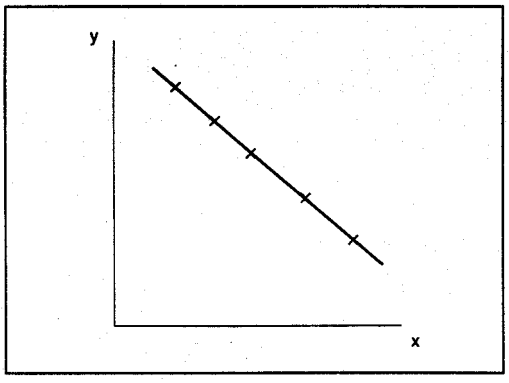
Gráficamente



Coeficiente de correlación r=1

Correlación perfecta y directa. A valores crecientes de x, le corresponde valores crecientes de y.

r = 1



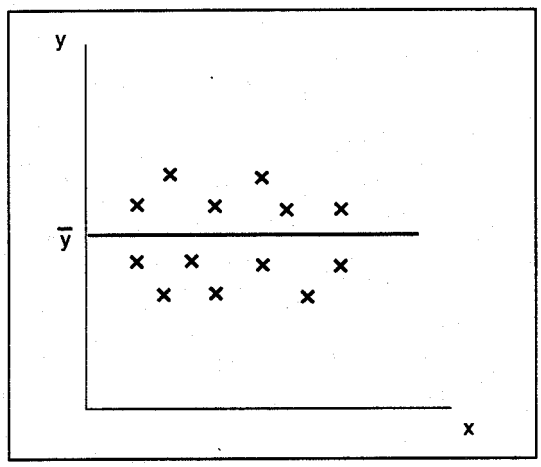
Coeficiente de correlación r =-1

Correlación perfecta e inversa. A valores crecientes de x le corresponden valores decrecientes de y.

r = 1

b. Caso 2º.

Cuando Sxy = 0 la recta de regresión de y sobre x es y = $\bar{y}$



Coeficiente de correlación r=0

A cualquier valor de x corresponde el mismo valor de y, lo que indica falta absoluta de dependencia entre las variables

En este caso $$r = \sqrt{1 - \frac{S_r^2}{S_y^2}} = \sqrt{1-1} = 0$$

r = 0 = No existe dependencia.

c. Caso 3º.

Para valores intermedios entre ± 1.

Cuanto más se acerque a 1 ó -1, mejor será la recta de regresión y mayor la dependencia. El signo nos da el tipo de relación de las variables:

r positiva, existe relación directa.

r negativa, existe relación inversa.

7.7. Coeficiente de determinación R^2.

Es otra medida de ajuste. Nos da la proporción de variabilidad total explicada de la variable y por la variable x. Mide el grado de dependencia.

Se define por: $R^2 = 1 - \frac{S_r^2}{S_y^2}$

Propiedades

a) R^2 varia entre $0 \leq R^2 \leq 1$. Cuanto más cercano a 1 mejor será el ajuste.

b) No se puede aplicar cuando 2 variables no están relacionadas.

c) Cuando: $R^2=1$ es porque $S_r^2 = 0$; dependencia total. Están todos los puntos de la distribución sobre la función teórica ajustada. Correlación perfecta.

Cuando: $R^2 = 0$ es porque $S_r^2 = S_y^2$; no hay dependencia, variables incorreladas.

- ***Ejemplo*:** datos del primer ejemplo.

$R^2 = \frac{S_y^2}{S_x^2 S_y^2}$ por ser regresión lineal.

$R^2 = \frac{14^2}{10 \bullet 20} = 0{,}98$ esto indica que el 98% de la variable y viene explicada por la variable x.

7.8. La predicción.

Los objetivos fundamentales de la teoría de regresión son tres:

1- Describir la dependencia casual entre las variables.

2- Tratar de expresar esta dependencia mediante una función matemática.

3- Predecir valores de la variable dependiente en función de valores de la variable independiente.

Para efectuar predicciones basta con utilizar la ecuación de la función teórica ajustada.

Así pues, si suponemos que la dependencia entre las variables es lineal, considerando la ecuación de la recta de regresión de y sobre x.

$$Y_{ti} = a + b\, x_i$$

para obtener predicciones de los valores de y para valores dados de x no tenemos más que sustituir estos valores de x en la ecuación anterior.

La predicción será más fiable cuanto más pequeños sean los residuos, o sea, cuanto menor sea la varianza residual, y por tato, cuanto más próximo a 1 esté el coeficiente de determinación.

- ***Ejemplo***: del primer ejemplo

Queremos saber el valor de y para el año 1974. A este año le corresponde el valor $x_i = 5$ y sustituyendo en la ecuación:

$Y_{ti} = 1{,}2 + 1{,}4x_i$ => $y_{ti} = 1{,}2 + 1{,}4 \cdot 5 = 8{,}2$ para el año 1974.