Apuntes de Estadística

Recopilación y traducción
Alicia de la Macorra

Temas:

Introducción

1. Estudio descriptivo de datos
1.1 Medidas de tendencia central
1.2 Medidas de variación
1.3 Descripción de datos por tablas
1.4 Descripción gráfica de datos

2. Conceptos básicos de probabilidad
2.1 Espacio muestral y eventos
2.2 La probabilidad de un evento
2.3 Las leyes básicas de la probabilidad

3. Distribuciones de probabilidad
3.1 Variables aleatorias
3.2 Representación gráfica de una distribución de probabilidad
3.3 La esperanza y sus propiedades
3.4 La varianza: una medida de dispersión
3.5 La distribución binomial
3.6 La distribución normal

4. Distribuciones muestrales
4.1 Muestreo aleatorio simple
4.2 Distribución muestral
4.3 Distribución de la media de la muestra y el teorema central dé límite.
4.4 Distribución de la proporción de la muestra

5. Estimación
5.1 Estimación puntual de un parámetro
5.2 Estimación por intervalos de confianza
5.3 Intervalo de confianza para la media de una población

5.4 La distribución t
5.5 Determinación del tamaño de la muestra para estimar medias
5.6 Determinación del tamaño de la muestra para estimar proporciones

6. Prueba de hipótesis

6.1 La hipótesis nula y la hipótesis alternativa
6.2 Los dos tipos de errores y el poder de una prueba
6.3 Prueba sobre la media de una población
6.4 Prueba sobre la proporción de una población

INTRODUCCION

I. Que es la Estadística.

Es un cuerpo de conceptos y métodos usados para colectar e interpretar datos que pertecen a una área particular de investigación así como para formular algunas conclusiones en situaciones con cierto grado de incertitud y variación.

Etimológicamente proviene de "*STATUS*" o estado inicialmente asociado solo a la economía, demografía y situaciones políticas de un país. Concepto que persiste pero se ha ampliado.

La busqueda científica no esta rígidamente estructurada, puede ser descrita como cualquier proceso encaminado a aprender sobre las regularidades escondidas de algunos aspectos de aquello que parece ser mas ó menos caótico en el mundo. Así son postulados teorías y modelos que son contrastados contra hallazgos fácticos, y entonces el modelo o la teoría son modificados, de esta manera continua la busqueda de mejores explicaciones.

Algunos pasos seguidos en la búsqueda científica son:

a) Especificación de Objetivos.- Al plantear claramente un problema o una investigación tendremos avanzado ya gran parte del camino a recorrer.

b) Recopilación de Información.- Información objetiva y pertinente.

c) Análisis de Datos.- Extracción de la información mas importante relacionada con el propósito del estudio o investigación.

d) Definición de Resultados.- La información significativa proporcionada por los datos y obtenida por el análisis, es evaluada en el contexto del cuerpo teórico en que los objetivos fueron especificados. Estableciendo un nivel de relevancia en relación a estos resultados.

La estadística esta presente en cada uno de dichos pasos:

a) Diseño de Experimentos y Diseño Muestral: que trata sobre la planeación de experimentos y la recopilación de datos.

b) Estadísticas Descriptivas: Métodos estadísticos que resumen y describen las características principales de los datos.

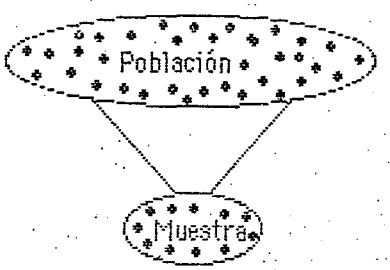
c) Inferencia Estadística: Evaluación de la información presente en los datos, los hechos apoyan o contradicen los postulados del modelo, y sugieren una revisión de la teoría existente o nuevas pistas de investigación.

Estas áreas de las estadísticas no son ajenas entre sí, es decir están integradas en un solo sistema de métodos, interaccionando y retroalimentandose.

<u>Población y Muestras.</u>

Población: Es un conjunto de posibles medidas o registros de algún rasgo cualitativo que comprende a la colección completa de unidades para la cuál la inferencia será realizada. La población representa el objeto de estudio de la investigación y el propósito del proceso de recopilación de datos es encontrar conclusiones sobre dicha población.

Muestreo: Un conjunto de medidas recolectadas de una población durante el curso de una investigación.



El concepto de población no implica necesariamente seres vivos.

Los principales objetivos de la estadística:

a) Hacer inferencia sobre la población a partir de la información contenida en los datos muestrales.

b) Apreciar el grado de incertitud involucrado en la inferencia.

c) Diseñar el proceso y alcance de las muestras para que las observaciones formen una base para bosquejar inferencias válidas y certeras.

ESTUDIO DESCRIPTIVO DE DATOS.

La medida registrada como un conjunto de datos son las piezas básicas de la información sobre el fenómeno estudiado, que es obtenida por el investigador. Esta información habrá que resumirla y resaltar sus aspectos más importantes.
Condensación de datos en forma de tablas, gráficas, cálculo de indicadores numéricos de centralidad y variabilidad.

Principales Aspectos:

a) Resumen y descripción de los patrones promedio de los datos, por medio de:
   * Presentación de tablas y gráficas.
   * Examen de la configuración obtenida, de sus características, incluyendo simetrías y desviaciones.
   * Análisis de los datos graficados incluyendo los datos que se encuentran lejos de conjunto general de datos.

b) Cálculo de medidas numéricas para:
   * Indicador típico o representativo del centro de los datos.
   * La cantidad de variación o dispersión presente en los datos.

Gráfica y Tablas.

Diagrama de puntos: Cuando se tiene un conjunto pequeño de medida, se pueden graficar sobre una línea utilizando un punto para cada medida individual.

Ejemplo: Utilizando las edades de los niños que van a una fiesta:
2, 3, 4, 5, 5, 5, 6, 6, 7, 8

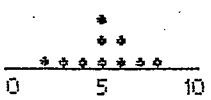


Diagrama de líneas: Uniendo los puntos del diagrama anterior para obtener líneas.

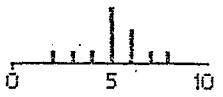
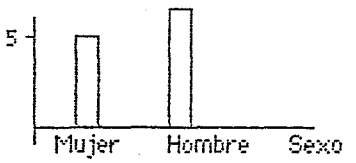


Diagrama de barras: Similar a la anterior utilizando barras para representar cada modalidad de la variable observada. Ejemplo: El sexo de los niños.



NOTA: Hasta aquí no es necesario que exista un continuo entre las modalidades.

Histrograma: Semejante al diagrama de barras pero implicamos un continuo sobre las modalidades observadas, cada modalidad se le denomina clase.

Polígono de frecuencia: La representación de una línea a partir del punto medio de cada clase.

NOTA: Podemos crear diferentes codificaciones al agrupar dos ó mas clases ó intervalo.

Ejemplo: Los niños del cumpleaños:
2, 3, 4, 5, 5, 5, 6, 6, 7, 8

Variable	Frecuencia		Variable	Frecuencia
2-3	2		menos ó 4	3
4-5	4	ó	5-6	5
6-7	3		7 ó mas	2
8-9	1			

Distribución de Frecuencia.

Cuando los datos consisten en una gran cantidad de medidas estos pueden concentrarse en un tabla de frecuencias.

Construcción de la tabla:

a) Busque el mínimo y el máximo valor del conjunto de datos.

b) Escoja una cantidad de intervalos de tamaño constante que cubran desde la modalidad mínima observada hasta la máxima. Cada grupo o intervalo formado se denomina clase o intervalo de clase.

c) Cuente el número de observaciones del conjunto de datos que pertenecen a dicha clase. A cada uno de estos resultados se le denomina frecuencia de clase.

d) Determine la frecuencia relativa de cada clase, al dividir la frecuencia observada sobre el número total de observaciones.

$$\text{Frecuencia relativa de una clase} = \frac{\text{Frecuencias observadas en la clase.}}{\text{Número total de observaciones.}}$$

Esto es la fracción de observaciones (el porcentaje) que perteneca a dicha clase (en relación al total observado).

Símbolos para el Conjunto de Datos y Operaciones de Suma.

Generalmente nuestros datos son una muestra de una población, las mediciones númericas a continuación descritas se refieren solo a los datos muestrales.

Cada observación es representada por $x_1, x_2, \ldots\ldots x_n$.

n = el número de medidas que forman el conjunto de datos.

$\sum_{i=1}^{n} x_i$ = la suma de los n números $x_1, x_2 \ldots x_n$

i varía de uno hasta n

Propiedades básicas de la Sumatoria.

Si a y b son constantes:

$$\sum_{i=1}^{n} bx = b \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} (bx_i + a) = b \sum_{i=1}^{n} x_i + na$$

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

Tipos de Escalas.

Escala:	Características:	
Nominal	categórica	Discreta
Ordinal	"	"
Intervalos	Númerica	Continua
De Razón	"	"
Absoluta	"	Discreta

Nominal: Son expresiones ó etiquetas que nos permiten diferenciar cada categoría.
Ej. Hombre - Mujer (sexo)
Ciencias Sociales - Matemáticas - Biología - Administración (preferencia)
Caminar - Taxi - Coche - Camión (transporte)

Ordinal: Son expresiones ordenables.
Ej. Nunca - Rara vez - Algunas veces - Casí siempre - Siempre (hábito)
Desacuerdo - No muy a menudo - Algo de acuerdo - Acuerdo (opinión)

Intervalos: Asignación de un valor númericos ante ciertas propiedades presentadas por las características. No tiene interpretación aritmética.
Ej. Temperatura (Grados °C ó °F)
I Q

Razón: Expresión numérica con propiedades aritméticas.
Ej. Distancia : Peso (kg, g, tm), Volumen (m^3), Tiempo (hrs, min, seg).

Absoluta: Conteo
Ej. Número de hijos, Número de coches

1.1 Medidas de Tendencia Central

Uno de los aspectos más importantes de la distribución de las medidas de la muestra es la posición del valor central. Existen communmente tres tipos de medida central.

La media muestral o promedio del conjunto de n medidas $x_1, x_2, \ldots\ldots x_n$. y la suma de estas medidas es dividida entre n. La media se denota $\bar{x}$ entonces se expresa:

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

De acuerdo al concepto de promedio, la media representa el centro del conjunto de datos. En sentido físico es el centro de gravedad.

La mediana muestral de un conjunto de n medidas $x_1, x_2, \ldots\ldots x_n$ es el valor de en medio cuando los datos son ordenados del menor al mayor. Si n es número impar, solo existe una única mediana, pero si es par existen dos valores en medio, entonces la mediana será el promedio de ambos.

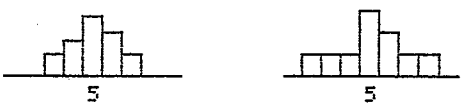
En una plabra la mediana es el valor que divide al conjunto de datos en dos mitades exactas, 50% antes y 50% despues de ella.

De esta manera podemos establecer otros indicadores, por ejemplo los cuartiles muestrales, que ordenando el conjunto de medidas del menor al mayor, el primer cuartil se encontrará en donde se tiene el 25% de los datos, el segundo cuartil coincide con la mediana, el tercer cuartil abarca el 75% de los datos y el cuarto cuartil el 100%.

La moda es el valor con mayor frecuencia entre los datos. Si existen dos valores se dice que la distribución es bimodal, y así sucesivamente.

1.2 Medidas de Variación.

Aunada a las medidas de tendencia central nos interesa saber cuál es la variabilidad que existe alrededor de ellas. Ya que dos conjuntos de datos pueden tener exactamente las mismas medidas de tendencia central y diferente variabilidad, por ejemplo:



Considerando la media muestral, podemos construir una medida de variabilidad, al obtener las diferencias de cada una de las observaciones $x_1, x_2, \ldots . x_n$., con respecto a la media. Es decir $(x_1-\bar{x}), (x_2-\bar{x}), \ldots (x_n-\bar{x})$ que llamamos desviaciones en relación a la media. Si tomamos el promedio de estas desviaciones obtendremos el valor cero, ya que la suma de las desviaciones es siempre cero (Por construcción ya que la media es la medida situada al centro de la distribución).

Entonces elevemos al cuadrado cada diferencia o desviación en relación a la media, y sacamos el promedio de dicha suma de cuadrados al dividirla sobre n-1 (el número de desviaciones no relacionadas entre sí, pues sabemos que la suma de las n desviaciones es cero). Así obtenemos la varianza muestral que denotamos como sigue :

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

que es igual a:

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

de ahí obtenemos la desviación estandar muestral que está definida como la raíz cuadrada de la varianza muestral. Es decir :

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

Regla de Chebyshev.

Para todo conjunto de datos:

El intervalo $\bar{x} \pm 2s$ contiene al menos $(1-\frac{1}{2^2}) = 3/4$ de los datos.

El intervalo $\bar{x} \pm 3s$ contiene al menos $(1-\frac{1}{3^2}) = 8/9$ de los datos.

En general, para cualquier multiplicador $k > 1$, el intervalo contiene al menos $(1 - \frac{1}{k^2})$ fracción de los datos.

Otra medida de dispersión es el rango que se define como la diferencia entre la medida observada de mayor valor menos la de menor valor. El rango indica la longitud en la que se expanden los datos.

Análogamente podemos obtener el rango intercuatil definido como la diferencia entre el tercer y primer cuartil.

Selección de la Medida Numérica.

Para escoger las medidas apropiadas para reportar los datos debemos considerar :

a). Propósito en el cuál el resumen descriptivo es citado.
b). Fácil de interpretar.
c). Grado de protección sobre observaciones aberrantes.
d). Uso potencial en la estadística descriptiva.

Conceptos básicos de Probabilidad.

Para estudiar un fénomeno que no esta regido por alguna ley precisa, realizamos varios experimentos y observamos su desarrollo social, económico, psicológico según el caso. Hemos visto como condensar y resumir los datos. Pero en general nuestro interés está centrado en hacer inferencias sobre la población de donde fueron extraídos esos datos (la muestra). La realización de dichas inferencias se basa en conceptos probabilísticos. Por ejemplo no será el mismo procedimiento si el método de muestreo empleado fue diferente, ó si las suposiciones sobre la población son tales o cuales.

En este capítulo consideraremos (excepcionalmente) que la estructura de la población es conocida, es decir que conocemos todos los datos de ella.

2.1 Espacio muestral y evento.

Al conjunto de todos los posibles resultados, o muestras observables en un experimento se denomina espacio muestral. Cada uno de estos posibles resultados se denominará elemento del espacio muestral. El espacio muestral se denota por S . Y cada elemento del espacio se denota por e_1, e_2, e_3, e_4,...

Ej.:
a) Anote le sexo de las dos primeras ratas que nacieron el bioterio de la Universidad.

$$S= \{ MM,FF,MF,FM\}$$

b) En un grupo de 15 animales al primer ensayo.

$$S= \{0,1,2,3,...15\}$$

c) Se mide la cantidad de una droga ingerida por chimpances, el alimento contenia 125mg.

$$S= \{ x: 0<=x<=125\}$$

d) Se anota la reacción ante un estímulo negativo 0 si no hay cambio conductual, 1 si lo hay. El registro se lleva ha cabo hasta encontrar un sujeto que no tenga cambio conductual.

$$S=\{0,10,110,1110,.....\}$$

El grupo de posibles resultados determinados por una característica descriptiva es llamado *evento*. Un evento es un subconjunto del espacio muestral S. El evento se denota por las primeras letras mayúsculas del alfabeto.

Ej:
a) Consideremos el espacio muestral

$$S = \{MM, FF, MF, FM\}$$

Sea A el evento de - Una rata macho únicamente -
Esta característica descriptiva identifica al evento como:

$$A = \{MF, FM\}$$

Un espacio muestral compuesto de un número finito o incluso infinito pero numerable de elementos se denomina espacio muestral discreto. Cuando el espacio muestral incluye todos los números en algun intervalo de los Reales, es llamado espacio muestral continuo.

Los ej. a) b) y d) son espacios muestrales discretos y c) es un espacio muestral continuo.

2.2 La probabilidad de un evento.

Intuitivamente probabilidad es una medida numérica de la posibilidad de ocurrencia de un evento. Es decir la proporción de veces que se espera que el evento ocurra cuando el experimento es repetido en condiciones idénticas.

El símbolo P(A) es usado para designar la probabilidad del evento A.

Equiprobabilidad de elementos (en el espacio muestral).

Cuando cada uno de los elementos del espacio muestral tiene la misma

posibilidad de ocurrencia, se dice que existe equiprobabilidad.
Ej.
Sea un dado perfecto
$S = \{e_1, e_2, e_3, e_4, e_5, e_6\}$ donde e_i corresponde a un lado del dado.

La probabilidad de cada elemento esta dada por:
$P(e_1) = P(e_2) = P(e_3) = P(e_4) = P(e_5) = P(e_6) = 1/6$

La probabilidad del evento A obtener un numero par es entonces:
$A = \{e_2, e_4, e_6\}$
$P(A) = P(e_2) + P(e_4) + P(e_6)$
$P(A) = 1/6 + 1/6 + 1/6 = 3/6 = 1/2$

Si un espacio muestral S esta formado de k elementos, todos equiprobables, la probabilidad de ocurrencia de cada elemento es 1/k. Y un evento A compuesto por m de estos elementos, tendra entonces la probabilidad :

$$P(A) = m/k$$

Un espacio muestral que posee equiprobabilidad en sus elementos es denominado un modelo uniforme de probabilidad.

Estabilidad de la frecuencia relativa.

Existen otros experimentos donde la suposición de equiprobabilidad no es posible, por ej. el coeficiente intelectual (que variará entre otros segun la edad del sujeto). En estas circunstancias, la probabilidad de un evento no puede calcularse a partir de un modelo uniforme. Para construir su probabilidad consideremos entonces la frecuencia de aparición del evento, cuando el experimento es repetido en condiciones idénticas.

Frecuencia relativa del evento A en N ensayos

$$r_N(A) = \frac{\text{Numero de veces que A ocurre en N ensayos}}{N}$$

donde $r_N(A)$ es simplemente la proporción de veces en que A se presenta en N repeticiones del experimento. Esta proporción tiende a estabilizarse hacia un valor numérico a medida que N aumenta.

La evidencia empírica de asignar un valor numérico a P(A) (a la probabilidad del evento A) esta derivada de la estabilización observada de la frecuencia relativa de A despues de muchas repeticiones.

Considerando estos dos conceptos: probabilidad uniforme y estabilidad de la frecuencia relativa definamos la probabilidad para espacios muestrales discretos.

Condiciones de un evento de la probabilidad para un espacio muestral discreto

i) Para todo evento A, $0 <= P(A) <= 1$

ii) P(A) es la suma de las probabilidades de los elementos del espacio muestral que pertenecen a A

$$P(A) = \sum_{\text{todo } e_i \text{ en } A} P(e_i)$$

iii) $P(S) = \sum_{\text{todo } e_i \text{ en } S} P(e_i)$

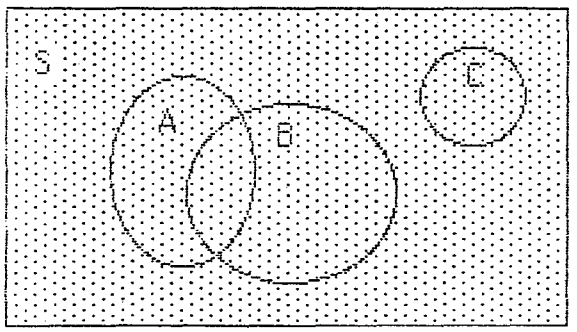
Entonces para asignar la probabilidad de un evento es necesario y suficiente asignar probabilidades a cada uno de los elementos de S.

2.3 Leyes básicas de Probabilidad.

A veces resulta tedioso o difícil enumerar todos los elementos de S que pertenecen al evento A, entonces tratamos de expresar A como integrada de eventos cuyas probabilidades se determinen mas fácilmente. Como estos eventos mas sencillos pueden estar combinados de diferentes maneras necesitamos saber como se comporta su probabilidad correspondiente. Por ello analicemos las propiedades básicas de la probabilidad.

Representemos S, en el siguiente esquema, utilizando un diagrama de Venn en donde cada uno de los elementos del espacio muestral lo representamos por un punto. El espacio que contendra esos elementos lo representamos por un rectángulo, o sea S (el espacio muestral). Entonces cada uno de los eventos puede ser representado por un subconjunto de este rectángulo, que

incluye los puntos de los elementos que pertenecen al evento de interés.



Las operaciones básicas son:

a) Unión del evento A y B que es el conjunto de los elementos pertenecientes a A como a B, se denota como A U B que es igual a B U A . La ocurrencia de AUB significa que al menos uno de los dos eventos ocurre.

b) Intersección de dos eventos, es decir A ∩ B = B ∩ A, que estará compuesta por los elementos de S que pertenecen simúltaneamente a A y a B. La podemos anotar también como AB que es igual a BA. La ocurrencia de AB significa que ambos A y B ocurren.

c) Complemento de un evento A, es el conjunto de todos los elementos de S que no pertenecen a A. Lo denotamos como $\bar{A}$. Si A no ocurre, significa que $\bar{A}$ ocurre.

Operación	Ocurrencia
AUB	Al menos uno, A o B
AB	Ambos, A y B
$\bar{A}$	No A

Dos eventos A y B serán considerados ajenos si la intersección AB es un conjunto vacío.

Ley de la complementación: $P(A)= 1-P(\bar{A})$

Ley de la adición: $P(AUB) = P(A) + P(B) - P(AB)$

Reglas de conteo y su uso en Modelos de Probabilidad Uniforme

Hemos especificado que la probabilidad de un evento es:

$$P(A) = \frac{\text{Número de elementos en } A.}{\text{Número de elementos en } S.}$$

Entonces el cálculo de la probabilidad de un evento puede ser reducido esencialmente al conteo de los elementos del eventos y del número de elementos del espacio muestral.

Algunas maneras de conocer el número de elementos de S:

a) Regla del producto:
Cuando un experimento consiste en dos partes, y la primera parte tiene n posibles resultados, y la segunda parte m posibles resultados. El espacio muestral del experimento tendra $n \times m$ elementos.
(Esto es generalizable a cualquier cantidad de partes de un experimento)

b) Regla de permutaciones
El número de arreglos diferentes que pueden formarse con r objetos tomados de n objetos distintos, se denota por P_n^r.
Y es:

$$P_n^r = n(n-1)(n-2)\ldots\ldots\ldots(n-r+1)$$

Así el número de arreglos distintos de n objetos tomados de n elementos es :

$$P_n^n = n(n-1)(n-2)\ldots\ldots\ldots(n-n)$$

c) Regla de combinaciones
El número de grupos ó colecciones de r objetos tomados de un conjunto de n objetos distintos se denota por :

$$\binom{n}{r}$$

que se lee el número de combinaciones posibles de r objetos tomados de n,

esto es :

$$\binom{n}{r} = \frac{n(n-1)(n-2)\ldots\ldots(n-r+1)}{r!}$$

$$= \frac{n!}{r!\,(n-r)!}$$

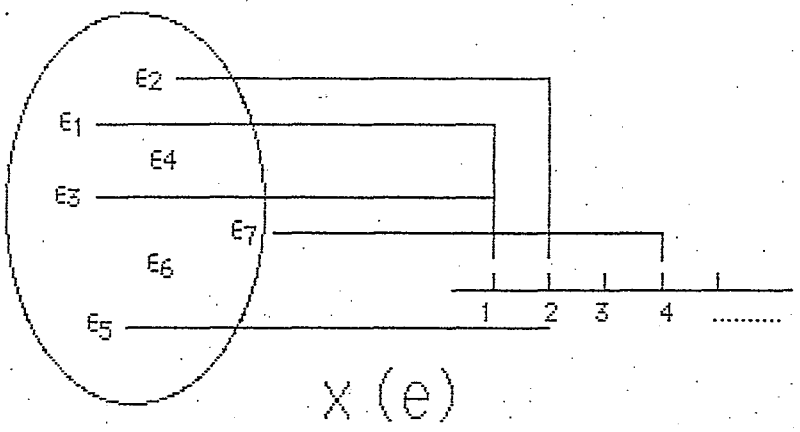
Por definición $0!=1$ entonces $\binom{n}{n} = 1 = \binom{n}{0}$.

Como regla general : la permutación es utilizada cuando el evento es caracterizado por un orden específico. Y cuando el orden en que se produzcan las observaciones no es importante utilizamos la regla de las combinaciones.

Distribuciones de Probabilidad

3.1 Variable Aleatoria.

Sea un espacio muestral con k eventos, la v.a. es una función numérica de ese espacio muestral sobre **R**.



Prácticamente se denota solamente como X y se nombra X, pues tendremos otras funciones de X, que designamos f(x) que son en realidad f(X(e)).

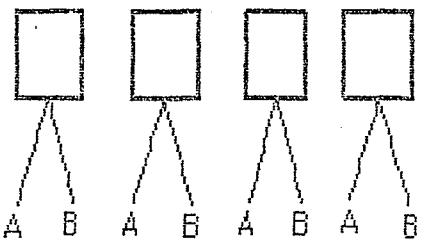
El adjetivo "aleatoria" hace referencia a:
- No sabemos de antemano que evento particular ocurrirá y
- Cuál valor en **R** tomará.

Distribución de probabilidad de una v.a.

Para denotar la variable aleatoria utilizamos las mayúsculas, como X y las minusculas x o x_i para denotar los valores particulares que puede tomar una v.a. (sus modalidades).

Al conjunto de probabilidades para cada valor de "x" de una v.a. X se llama distribución de probabilidad de la v.a. X

ejemplo: Aplicación de alguna de dos pruebas al azar a cuatro alumnos



2 x 2 x 2 x 2 = 16 Posibilidades ó patrones de respuesta.

El espacio muestral S, es:

AAAA	AAAB	AABB	BBBA	BBBB
	AABA	ABAB	BBAB	
	ABAA	ABBA	BABB	
	BAAA	BABA	ABBB	
		BAAB		
		BBAA		

Si consideramos equiprobabilidad para cada elemento (1/k):

1/16	1/16	1/16	1/16	1/16
	1/16	1/16	1/16	
	1/16	1/16	1/16	
	1/16	1/16	1/16	
		1/16		
		1/16		

Si escogemos la v.a. X como las veces que aparece o es aplicada A tenemos:

X = 4 3 2 1 0 veces.

La probabilidad de cada modalidad será la suma de los elementos del espacio muestral que corresponden a dicha modalidad, en este caso:

1/16 4/16 6/16 4/16 1/16

La distribución de probabilidad de la v.a. X es la lista de cada x_i en donde asociamos a cada una su probabilidad

	x_1	x_2	x_3	x_4	x_5
X=	0	1	2	3	4
Prob	1/16	4/16	6/16	4/16	1/16
f(x)	$f(x_1)$	$f(x_2)$	$f(x_3)$	$f(x_4)$	$f(x_5)$

En general:
Para cada valor X_i de X, asociamos $f(X_i) = P(X= X_i)$ para todo i= 1,...,k

<u>Propiedades de la función de probabilidad.</u>

$f(x_i) <= 1$ ó $P(X=x_i) <=1$ $p(x_i)<=1$

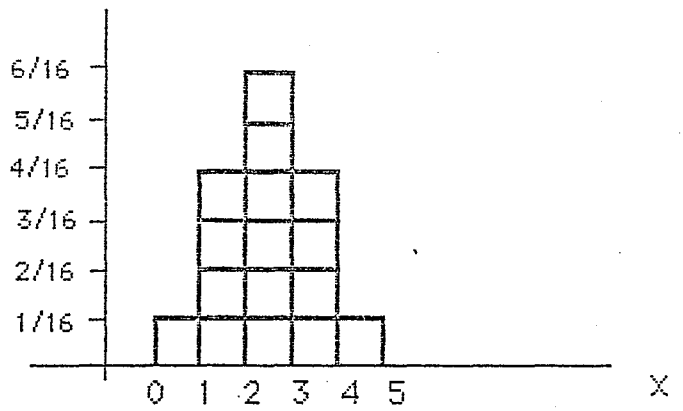
$$\sum_{i=1}^{k} f(x_i)=1 \qquad \sum_{i=1}^{k} p(x_i)=1$$

Conclusión:

La distribución de probabilidad de una v.a. es un modelo *teórico* que asigna probabilidades a cada una de las modalidades de dicha v.a.

3.2 Representación gráfica de una distribución de probabilidad

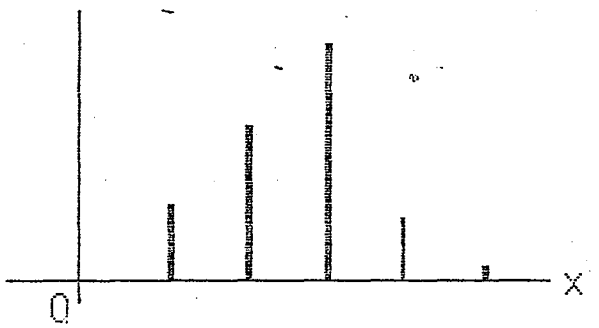
Construyamos el histograma a partir de la distribución obtenida teóricamente.



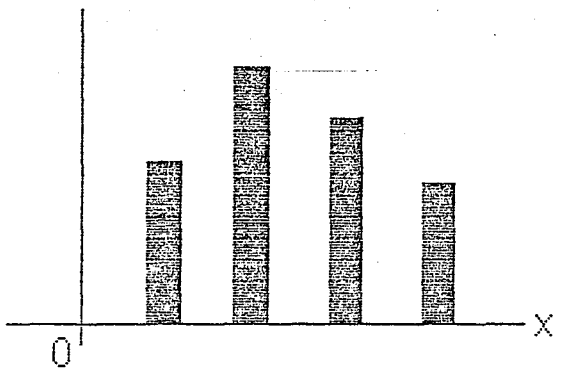
El área de cada barra es la probabilidad asignada a dicha x_i.

Se recomienda utilizar el histograma si X es continua y los valores de x_i estan igualmente espaciados sino utilice:

a) Diagrama de líneas:



b) Diagrama de barras:



Nota:
Si comparamos la representación gráfica de datos obtenidos en una muestra, habrá mas posiblidades que este se asemeje a la distribución de probabilidad de la v.a. si el tamaño de la muestra es grande.

3.3 La Esperanza y sus propiedades.

Sea X una v.a. con función de probabilidad f(x). Entonces el valor esperado de X, que anotamos como **E(X)**, esta definido por:

$$E(X)=\sum_i x_i\, f(x)$$

En términos físicos la esperanza es el centro de gravedad de la distribución de probabilidad de la v.a.

Característica:

$E(X)=\mu$ μ=media poblacional

Propiedades:
Sea a,b,c, constantes entonces

$E(c) = c$
$E(cX)= cE(X)$
$E(X+c)= E(X)+c$
$E(a+bX)= a+bE(X)$
$E(a+bX+cX^2)= a+bE(X)+cE(X^2)$

Así la esperanza de una suma es igual a la suma de las esperanzas

3.4 Varianza.

Medida de dispersión de la distribución de probabilidad de una v.a., está definida por:

$Var\,(X) = E[(X- E(X))^2]$ ó $\sigma^2(X) = E(X)-\mu^2$

Propiedades:

$\sigma^2(X)>=0$ (siempre positiva)
$\sigma^2(X+a)=\sigma^2(X)$
$\sigma^2(bX)=b^2\sigma^2(X)$
$\sigma^2(a+bX)=b^2\sigma^2(X)$

Nota: La varianza es muy sensible a las unidades de medida utilizadas.
La desviacion estandar estará expresada en las mismas unidades que la v.a.

Construcción de un modelo de probabilidad.

Como generalmente no conocemos la distribución de probabilidad de la v.a. construimos entonces un modelo de probabilidad.

Características:
. Adecuado.
. Simple.
. Parsimonia en los parámetros.

Ensayo de Bernoulli

a. Consideremos el caso en que para cada ensayo solo pueden darse dos resultados:

éxito (s)
fracaso (f)

b. Para cada ensayo la probabilidad de éxito es p(s) ó p, y como solo hay dos posibles resultados, la probabilidad de fracaso es;

$p(f)=1-p(s)$
$q=1-p$ $\quad p+q=1$

c. Cada ensayo es independiente o sea que cada uno no da más o menos información de los anteriores o futuros.

Si repetimos varios ensayos de Bernoulli n veces con una probabilidad p para cada uno de los ensayos, podremos representar una v.a. X que representa el número de éxitos en esos ensayos.

3.5 Distribucion Binomial.

Modelo de probabilidad de n ensayos de Bernoulli con p (es decir probabilidad de éxito).

ej:	Nacimientos:	Hombre	Mujer
	Control calidad:	Defectuoso	Aprobado
	Moneda :	Aguila	Sol
	Enfermedad:	Presencia	Ausencia

Sea n=4 hay 16 patrones o posibles resultados, ya que aplicando la ley del producto dividimos el experimento en 4 partes en cada una pueden aparecer 2 posibles resultados, entonces:

$$2 \times 2 \times 2 \times 2 = 16$$

(como vimos en el ejemplo de las pruebas aplicadas a los alumnos)

$$\binom{n}{x} = \frac{n!}{x!\,(n-x)!}$$

n: número de repeticiones de Bernoulli.
x: frecuencia de la modalidad que nos interesa.

Verifiquemos:

$$\binom{4}{0} = \frac{4!}{0! \times 4!} = 1 \qquad \binom{4}{1} = \frac{4!}{1! \times 3!} = 4 \qquad \binom{4}{2} = \frac{4!}{2! \times 2!} = 6 \qquad \binom{4}{3} = \frac{4!}{3! \times 1!} = 4 \qquad \binom{4}{4} = \frac{4!}{4! \times 0!} = 1$$

Para obtener la probabilidad para cada patrón consideremos los 4 ensayos de Bernoulli, pues n = 4,cada uno es independiente, entonces el producto de sus probabilidades nos dará la probabilidad del patrón es decir:

$$p^x\ (1-p)^{n-x} \qquad \text{ó} \qquad p^x\, q^{n-x}$$

obtenemos así la distribución de cualquier v.a. dicotómica.

$$X \sim B(n,p)$$

Con $p + q = 1$ definida por:

$$\sum_{x=0}^{n} \binom{n}{x} p^x q^{n-x} = (p+q)^n = 1$$

Así por ejemplo la probabilidad de $x = 2$ de nuestro ejemplo será:

$$\binom{4}{2} (1/2)^2 (1/2)^{4-2} = 6\ (1/4)\ (1/4) = 6/16$$

que es el mismo resultado que obtuvimos.

Nota: Se llama distribución binomial puesva generando los coeficientes binomiales - relación con triángulo de Pascal -.

Esperanza y Varianza de una distribución binomial.

Esperanza:

a) Consideremos el caso de un solo ensayo de Bernoulli

$$E(X_1) = p \times 1 + q \times 0 = p$$

X	1	0
probabilidad	p	q

b) Si son n ensayos:

$$E(X) = E(X1) + E(X2) + \ldots\ldots\ldots\ldots + E(Xn)$$
$$= p + p + \ldots\ldots\ldots\ldots + p .$$
$$E(X) = np.$$

Varianza:

a) Consideremos un solo ensayo:

$$\text{Var}(X_1) = E(x_1^2) - [E(X_1)]^2$$

como $E(X_1) = p \times 1^2 + q \times 0^2 = p$ y $E(X_1)$

entonces

$$\text{Var}(X_1) = p - p^2$$
$$= p\,(1-p)$$

$$\text{Var}(X_1) = p\,q$$

b) Como son n ensayos independientes:

$$\text{Var}(X) = \text{Var}(x_1) + \ldots\ldots\ldots + \text{Var}(x_n)$$
$$= p\,q + \ldots\ldots\ldots + p\,q$$

$$\text{Var}(X) = n\,p\,q$$

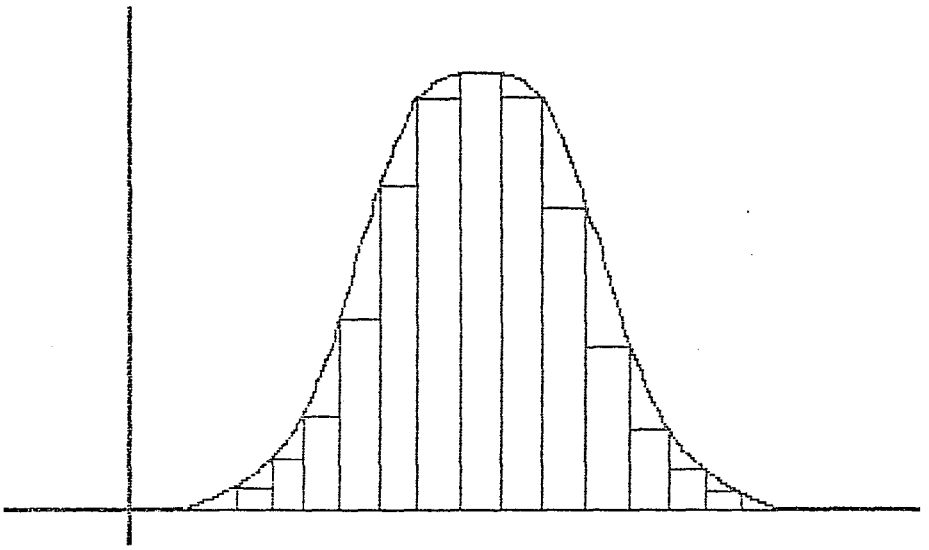
3.6 Distribucion Normal o Laplace - Gauss.

Modelos de probabilidad para una variable aleatoria X continua es decir que X puede asumir cualquier valor en un intervalo dado por **R** .

Ejemplo:

- Peso.
- Tiempo - edad,duración -
- Talla.
- Distancia.
- Temperatura.

Si hacemos un histograma de un gran número de observaciones, de X_1, vamos perdiendo el el escalonamiento para obtener una curva.



Propiedades:

a) El área bajo la curva es 1.
b) La probabilidad entre dos puntos a y b es el área de la curva perteneciente a este intervalo [a,b].
c) La probabilidad en un punto "a" es cero.
d) f(x) >= 0 lo llamamos función de densidad.
e) Simetría.

Entonces

p[a <= x <= b] = p[a < x <= b] = p[a <= x < b] = p[a < x <b]
p[a < x < b]= área a la izquierda de b) -(área a la izquierda de a).
p[x > b] = área a la izquierda de b.

Modelo

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

con media μ y desviación estandar σ

$$X \sim N(\mu,\sigma)$$

características

$$p[\mu - \sigma < x < \mu + \sigma] = .683$$
$$p[\mu - 2\sigma < x < \mu + 2\sigma] = .954$$
$$p[\mu - 3\sigma < x < \mu + 3\sigma] = .997$$

Distribución Normal Estandar

$$\mathcal{N}(0,1)$$

Para estandarizar nuestras distribuciones y poder comparar con otras distribuciones normales usamos para ello el cambio de variable definida por:

$$Z = \frac{X - \mu}{\sigma} \sim N(0,1)$$

Curva símetrica, con $\mu = 0$ y $\sigma = 1$, entonces

a) $p[Z <= 0] = .5$
b) $p[Z <= -z] = 1 - p[Z <= z]$
c) $p[Z <= z] = .5 + p[0 < Z <= z]$
$p[Z >= -z] = .5 - p[0 > Z >= z]$

Propiedades derivadas de la linealidad de la esperanza.

a) Si $Y = a + bX$ y X esta en $\mathcal{N}(\mu,\sigma)$
entonces Y esta $\mathcal{N}(a+b\mu, |b|\sigma)$

b) La suma de v.a independiente con distribución normal es támbien una distribución normal.

osea

$$X \sim N(\mu_1, \sigma_1)$$
$$Y \sim N(\mu_2, \sigma_2)$$

$$X+Y \sim (\mu_1+\mu_2, \sqrt{\sigma_1^2+\sigma_1^2})$$

c) Cuando una B(n,p) con n grande y p no cerca de 0 ó 1 :

$$B(n,p) \quad N(n\,p, \sqrt{npq})$$

$$Z = \frac{X - np}{\sqrt{npq}} \approx N(0,1)$$

DISTRIBUCIONES MUESTRALES

4.1 Muestreo aleatorio simple

Generalmente no podemos observar todos los resultados posibles de una v.a., es decir que rara vez observamos todos lo eventos de $. Sino que solo observamos algunos, entonces a partir de una muestra tratamos de conocer el todo. Por ejemplo: no necesitamos tomarnos toda la sopa para saber si esta caliente.

El procedimiento de tomar una muestra es muy útil en muchas cosas: economía, investigación de mercado, sociología, control de calidad, etc. La población a estudiar en cada caso son el conjunto de individuos u objetos en los cuáles el investigador esta interesado.

Definición: Sea N y n respectivamente el número de elementos en la población y en la muestra. El muestreo aleatorio simple implica que todas las muestras tienen la misma probabilidad de ser escogidas.

Tenemos $X_1,.....,X_n$ v.a. observables y cuando tomamos la muestra obtenemos $x_1,.....,x_n$ datos observados. Para describir estos datos utilizamos la media $\bar{x}$, la mediana, la varianza s^2, etc.

A cada una de estas medidas en función de la muestra les llamamos estadísticos. El estadístico es entonces, en sí mismo, una v.a.

4.2 Distribución muestral

Siendo un estadístico una función de las v.a. $X_1.....X_n$, tiene a su vez una distribución de probabilidad, es decir que el comportamiento del estadístico puede ser descrito por alguna distribución.

Definición: Distribución muestral es la distribución de probabilidad de un estadístico.

Para conocer la distribución de probabilidad de $\bar{X}$:
Supongamos una muestra de 2 unidades de observación que han sido observadas de acuerdo al ejemplo del tema anterior, es decir tenemos X_1 y X_2, y calculamos $\bar{X}$; ¿ Cuál es la distribución de probabilidad del estadístico $\bar{X}$?

Nota: X_1 y X_2 son variables aleatorias independientes entonces:

$$P[X_1=x_1,\ X_2=x_2] = P[X_1=x_1] \times P[X_2=x_2]$$

La distribución de probabilidad de X_1 y X_2 es:

X_2 \ X_1	0	1	2	3	4	
0	.003962	.015625	.0234375	.015625	.0039062	.0625
1	.015625	.0625	.09375	.0625	.015625	.25
2	.0234375	.09375	.140625	.09375	.0234375	.375
3	.015625	.0625	.09375	.0625	.015625	.25
4	.0039062	.015625	.0234375	.015625	.0039062	.0625
	.0625	.25	.375	.25	.0625	

El estadístico que nos interesa es:

$$\bar{X} = \frac{X_1 + X_2}{2}$$

Para obtener su distribución de probabilidad:
a) Enunciemos todos los posibles resultados.

X_2 \ X_1	0	1	2	3	4
0	0	.5	1	1.5	2
1	.5	1	1.5	2	2.5
2	1	1.5	2	2.5	3
3	1.5	2	2.5	3	3.5
4	2	2.5	3	3.5	4

Distribución Muestral de $\overline{X}$

$\overline{x}$	0	.5	1	2	2.5	3	3.5	4
probabilidad	.0039062	.03125	.109375	.21875	.2734347	.21875	.109375	.0039062

ya que

$$P[\overline{X} = 0] = P[X_1=0, X_2=0] = .0039062$$

$$P(\overline{X} = 1] = P[X_1=2, X_2=0] + P[X_1=0, X_2=2] + P[X_1=1, X_2=1]$$
$$= .0234375 + .0234375 + .0625 = .109375$$

y así, para cada modalidad de $\overline{X}$

4.3. Distribución de Probabilidad de la Media de la Muestra y Teorema Central del Límite

Hemos visto el comportamiento de un estadístico, es decir de una v.a. función de la muestra, pero lo que nos interesa es hacer inferencia sobre la población.

En una muestra aleatoria, las v.a. $X_1, X_2, \ldots\ldots\ldots\ldots, X_n$,son independientes, todas con la misma distribución, que incluso es similar a la de la población es decir:

$$E(X_1) = E(X_2) = \ldots\ldots\ldots = E(X_n) = \mu$$
$$Var(X_1) = \ldots\ldots\ldots = Var(X_n) = \sigma^2$$

Como: $\overline{X} = \frac{1}{n}(X_1 + X_2 + \ldots\ldots\ldots + X_n)$

$$E(\overline{X}) = \frac{1}{n} E(X_1 + X_2 + \ldots\ldots\ldots + X_n)$$

$$= \frac{1}{n}(E(X_1) + \ldots\ldots\ldots + E(X_n))$$

$$= \frac{1}{n}(\mu + \ldots\ldots\ldots + \mu)$$

$$= \frac{n\mu}{n}$$

$$= \mu$$

$$\text{Var}(\overline{X}) = \frac{1}{n^2}\text{Var}(X_1 + X_2 + \ldots\ldots\ldots + X_n) \quad \text{por la independencia entonces:}$$

$$= \frac{1}{n^2}(\text{Var}(X_1) + \ldots\ldots\ldots\ldots + \text{Var}(X_n))$$

$$= \frac{1}{n^2}(\sigma^2 + \ldots\ldots\ldots\ldots + \sigma^2)$$

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

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

O sea que la distribución de X tendrá como esperanza

$$E(\overline{X}) = \mu$$

$$\text{Var}(\overline{X}) = \frac{\sigma^2}{n}$$

Atención: aqui relacionamos estadísticos (de la muestra) con parámetros (de la población).
Resumen: el centro de la distribución de probabilidad de X es μ y la desviación estandar es $\sigma/\sqrt{n}$.

La desviación decrementará a medida que n aumenta. Así podemos establecer dos importantes resultados para la estadística:

Resultado 1: A partir de una muestra aleatoria de tamaño n de una población normal con media μ y desviación estandar σ, la media de la distribución de $\overline{X}$ tienen una distribución normal con media μ y desviación $\sigma/\sqrt{n}$

Resultado 2: Teorema central del limite. Más aun en un muestreo aleatorio de una población **arbitraria** o cualquiera con media μ y desviación estandar σ, la distribución de $\overline{X}$ cuando n es grande (n>30), es aproximadamente normal con media μ y desviación $\sigma/\sqrt{n}$ es decir:

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

es por ello que la distribución normal tiene un lugar primordial en los procedimientos estadísticos.

4.4 Distribución de la proporción de la muestra.

La razón por la cual la binomial (con n>30 y p no extrema) es aproximada por una normal, entonces el Teorema Central del Límite.
Supongamos para cada ensayo de Bernoulli

Sea $X_i = 1$ si el ensayo es éxito.

$X_i = 0$ si el ensayo es fracaso.

cada v.a. $X_1, \ldots\ldots X_n$, de los n ensayos tendra distribución

X	0	1
f(X)	q	p

Con media p y desviación estandar $\sqrt{pq}$ (ver capítulo distribución de probabilidad).
Como $X_1 \ldots\ldots X_n$, pueden ser consideradas como una muestra aleatoria, la consecuencia del teorema central del límite es entonces que la distribución es aproximadamente normal $\mathcal{N}(p, (\sqrt{pq}/\sqrt{n})$ cuando n>30.
Ahora llamemos T al total de éxitos

$$T = \sum_{i=1}^{n} X_i$$

$$\overline{X} = \frac{T}{n}$$

osea que $\overline{X}$ es la proporción o la frecuencia relativa de los n ensayos exitosos. Prácticamente para designar a esta proporción la anotamos como p en lugar de $\overline{X}$.

Resumen: Cuando n>30, la distribución muestral de p (la proporción de éxitos en los n ensayos de Bernoulli) se distribuye aproximadamente

$$Z = \frac{\hat{p} - p}{\sqrt{pq / n}} \sim N(0,1)$$

que tambien podemos encontrar como:

$$Z = \frac{T - np}{\sqrt{npq}} \sim N(0,1)$$

ya que multiplicamos el numerador y el denominador por n y consideramos np=T .

Estimación

Introducción:

Recordando que la distribución de probabilidad puede ser utilizada como un modelo para ajustar diferentes mediciones, y asociando el concepto de distribución muestral, estamos ahora preparados para introducir técnicas de inferencia en el análisis de datos. Cuando los datos son recopilados en una muestra a partir de una población, el objetivo principal del análisis estadístico es de establecer algunas inferencias ó generalidades sobre la población a partir de la información parcial contenida en los datos de la muestra. Generalemente, nuestro interés está centrado en algunas características típicas de la población, tales como establecer una proporción de cierta característica, la media y la desviación sobre la distribución de la población, o cualquier otra medida numérica de tendencia central o variabilidad. Cualquier característica de la distribución de la población puede ser expresada en términos de parámetros.

Parámetro : Es cualquier característica de la distribución de la población. Y la estadística inferencial trata de definir generalizaciones sobre los parámetros poblacionales a partir del análisis de los datos muestrales.

Antes de recopilar el conjunto de datos para realizar inferencia, deberemos considerar los siguientes puntos:
a) Tamaño de la muestra y manera de muestrear.
b) Naturaleza de la inferencia deseada.
c) Presición y aseveridad de nuestras conclusiones.

Introduciremos solo técnicas inferenciales que estan basadas en muestras aleatorias.

Haremos énfasis solo en dos tipos inferencia:
a) Estimación de un parámetro.
b) Prueba de hipotesis.

Utilizando cualquiera de estos tipos de inferencia, el valor verdadero de un

parámetro es una constante desconocida que puede ser correctamente reconocida solo haciendo un estudio exhaustivo de la población, en el caso de que esto fuese posible.

Un objetivo realístico es el de estimar el valor verdadero del parámetro o un intervalo plausible donde se encuentre el parámetro, estos calculados a partir de los datos de la muestra; tratando incluso de determinar la precisión del procedimiento. Este tipo de inferencia es la llamada estimación de parámetros.

Un objetivo alternativo puede ser el tratar de decidir si un parámetro pertenece o no (dentro de un rango específico de valores) a alguna conjetura científica. Es decir la prueba de hipotesis.

5.1 Estimación puntual de un parámetro.

Su objetivo principal es de producir un solo número a partir de los datos muestrales que sea lo más parecido al valor desconcido del parámetro.

Para este propósito utilizamos el símbolo θ para denotar el parámetro que puede ser por ejemplo una proporción, una media, mediana o cualquier medida de variabilidad. Es importante recordar que θ tiene un valor numérico constante, sin embargo este valor es desconocido para nosotros.

La información disponible esta dada en la muestra aleatoria, en donde se ha observado una variable aleatoria X, de forma $X_1, \ldots X_n$, en donde se denotan las modalidades observadas $x_1, \ldots x_n$ en una muestra de tamaño n tomada de una población. Queremos formular una función de la observaciones muestrales ($X_1, \ldots X_n$) , es decir un estadístico tal que el valor calculado a partir de los datos de la muestra pueda reflejar el parámetro de la población lo mas cercanamente posible.

Un intento estadístico para estimar el parámetro θ es llamado estimador puntual o estimador de θ y se denota comunmente por $\hat{\theta}$. Este estimador es solamente una función de la muestra, entonces su valor podrá ser calculado una vez observada la muestra.
Un estimador $\hat{\theta}$ es una función de las observaciones muestrales y es usado para estimar el valor desconocido de un parámetro θ. Este estimador $\hat{\theta}$ es

una variable aleatoria con una distribución de probabilidad. Cuando una muestra aleatoria es factible para una población y $\hat{\theta}$ es calculado a partir de un conjunto de datos, el valor numérico obtenido es llamada una estimación de θ a partir de una muestra específica.

Como seleccionar el estimador apropiado ?

Un estimador es insesgado en relación a un parámetro si $E(\hat{\theta})=\theta$.

Seleccionar un estimador insesgado de θ, se escogerá aquel que tenga la varianza mínima. Si existe se denomina estimador insesgado de mínima varianza de θ.

No olvidemos sin embargo que hay estimadores sesgados que pueden tener una mayor concentración de probabilidad para encontrarse cerca del parámetro y que tienen menor varianza que algún otro insesgado. Sin embargo en este curso nos remitiremos a los estimadores insesgados.

Una vez determinado el estimador a utilizar (su procedimiento de cálculo y sus características), calculamos un solo número a partir del conjunto de datos. --estimacion puntual-- .

Pero necesitamos establecer que precisión tiene esta estimación sino no podremos utilizarla. Para ello determinamos la variabilidad de la distribución del estimador.

La desviación estandar del estimador $\hat{\theta}$ es llamada error estandar y se designa como $s.d.(\hat{\theta})$.

a) Estimación puntual de la media de la población.

Para estimar una media de una población a partir de una muestra aleatoria, por intuición el estimador a utilizar es la media de la muestra $\bar{X}$. Analicemos las propiedades de este estimador para sustentar la intuición. El parámetro θ en este caso es la media de la población μ, y el estimador $\hat{\theta}$ es la media muestral $\bar{X}$.

De acuerdo a los resultados del teorema del límite central $E(\bar{X})=\mu$ y $s.d.(\bar{X})=\sigma/\sqrt{n}$ Osea con n grande $\bar{X}$ se distribuye como una Normal de media μ y desviación estandar $\sigma/\sqrt{n}$.

Más aún, entonces, utilizando las tablas de la distribución normal, el estimador $\bar{X}$ estará a una distancia máxima de $2\sigma/\sqrt{n}$ con una probabilidad de .954 del valor verdadero del parámetro μ. Y a 3 (error estandar) con una probabilidad de .997. En otras palabras podemos decir que cuando estamos estimando μ a partir de $\bar{X}$, la cota de error al 95.4% es de $2\sigma/\sqrt{n}$ y la cota de error al 99.7% es de $3\sigma/\sqrt{n}$.

Nótese que el cálculo implica un parámetro de la población σ. Como desconocemos el valor de σ entonces podemos estimar la desviación estandar de la población en función de la desviación estandar de la muestra.

Cuando n es grande, se puede hacer caso omiso del efecto de aproximar $\sigma/\sqrt{n}$ con $s/\sqrt{n}$. El estadístico $s/\sqrt{n}$ es un estimador del error estandar.

Resumen: Estimación puntual de la Media.
Parámetro: media de la población μ.
Datos: $X_1, \ldots, X_n$ (una muestra aleatoria de tamaño n).
Estimador: $\bar{X}$ (media de la muestra).
S.E. $(\bar{X}) = \sigma/\sqrt{n}$, estimada por S.E. $(\bar{X}) = s/\sqrt{n}$.
Para n grande, con una aproximación de 95.4% la cota de error es $\pm 2s/\sqrt{n}$.

b) Estimacion de una proporcion binomial

Cuando n elementos constituyen una muestra aleatoria de una población, y los datos consisten en contar X, el número de elemntos de la muestra que poseen la característica. La proporción de la muestra es $\hat{p}=X/n$, por sentido común quisieramos utilizarla como estimador de p (de la población).

Cuando el tamaño de la muestra es n, un pequeña fracción de la población, las n observaciones de los n elementos pueden ser consideradas como n ensayos independientes de Bernoulli con una probabilidad de éxito de p.

De la muestra se contará X, que tiene una distribución $b(n,p)$ con media np y varianza npq. Entonces las propiedades del estimador son:

$$E(\hat{p}) = 1/nE(X) = np/n = p$$

$$\text{Var}(\hat{p}) = 1/n^2 \text{ var}(X) = npq/n^2 = pq/n$$

Entonces $\hat{p}$ es un estimador insesgado de p. Ademas de que $\hat{p}$ tienen la varianza mas pequeña de todos los estimadores insesgados.

El error estandar esta dado por S.E. $(\hat{p}) = \sqrt{pq/n}$ que será estimado con $\hat{p}$ y $\hat{q}$. Notemos que cuando n es grande $\hat{p} = X/n$ se distribuye normalmente con media p y desviación estandar $\sqrt{pq/n}$. Esta distribución normal nos permite asegurar que con una probabilidad aproximada de .954 el error de estimación $|p - \hat{p}|$ será menor a 2 S.E. estimada.

Resumen de Estimación puntual de un parámetro binomial.
Parámetro: proporción de la población p.
Datos : X= Número de elementos que tienen la característica estudiada entre los n elementos de la muestra aleatoria.
Estimador: $\hat{p} = X/n$
S.E. $(\hat{p}) = \sqrt{pq/n}$ y S.E. estimada $(\hat{p}) = \sqrt{\hat{p}\hat{q}/n}$
Para n grande, y una aproximación de 95.4% la cota de error es $\pm 2\sqrt{pq/n}$.

5.2 Estimación por intervalo de confianza.

Un estimador puntual calculado a partir de una muestra nos provee de un único número como estimador de un parámetro. Este único número es impreciso en cierta manera debido a la desviación estandar asociada. Una opción alternativa es utilizar el concepto de cota de error para producir un intervalo de valores que presumiblemente contengan el verdadero valor del parámetro.

Formalmente, sea $X_1, \ldots X_n$ una muestra aleatoria y θ un parámetro desconocido de la población. Un intervalo para θ es un intervalo (inf, sup) calculado a partir de las $X_1, \ldots X_n$ observaciones muestrales, que debidamente muestreados, incluye el verdadero valor desconocido de θ con una probabilidad específica. Esta probabilidad se denota por $(1-\alpha)$, que generalmente es tomado como .90, .95 o .99.

Entonces :
Sea $(1-\alpha)$ la especificación de una probabilidad alta, y inf y sup funciones

de $X_1,...X_n$ tal que:

$$P(Inf<\theta<Sup)=(1-\alpha).$$

Entonces, el intervalo (inf,sup) es llamado intervalo de confianza al $100(1-\alpha)\%$ para el parámetro, y $(1-\alpha)$ es llamado nivel de confianza asociada al intervalo.

a) Intervalo de confianza para μ con σ conocida.
Para clarificar los conceptos enunciados, construyamos un intervalo de confianza para una media poblacional μ cuando el tamaño de muestra es grande y la desviación estandar σ es conocida (esta última suposición es artificial, pero en una primera instancia nos facilita el entendimiento del procedimiento).

De acuerdo con el teorema del limite central, la distribución de $\bar{X}$ puede ser aproximada por $\mathcal{N}(\mu,\sigma/\sqrt{n})$, donde $\sigma/\sqrt{n}$ es un número conocido. Esta aproximación es buena para muestras grandes provenientes de poblaciones normales o no, pero se sostiene igualmente para muestras pequeñas cuando la distribución de la poblacion es normal.

Las tablas de la normal muestran que con variable aleatoria normal $(\bar{X})$ caerá con una probabilidad del .95 a 1.96 desviaciones estandar de la media. Es decir:

$$P(\mu-1.96\,\sigma/\sqrt{n}<\bar{X}<\mu+1.96\,\sigma/\sqrt{n})=.95$$

El evento $(\mu - 1.96\,\sigma/\sqrt{n} < \bar{X})$ es equivalente a $(\mu< \bar{X} + 1.96\,\sigma/\sqrt{n})$, que como si agregamos la constante $1.96\,\sigma/\sqrt{n}$ a cada lado de la inegalidad. Así de la primera expresión tenemos que $(\bar{X}< \mu+1.96\,\sigma/\sqrt{n})$ es equivalente a $(\bar{X}-1.96\,\sigma/\sqrt{n}<\mu)$, así la probabilidad anterior puede expresarse como:

$$P(\bar{X}-1.96\,\sigma/\sqrt{n}<\mu<\bar{X}+1.96\,\sigma/\sqrt{n})=.95$$

Así podemos identificar:

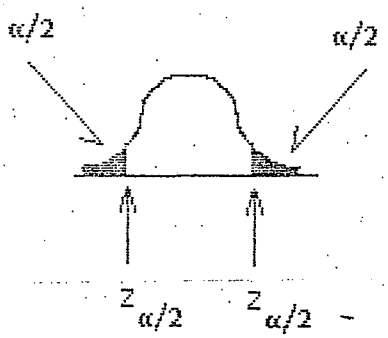
$$Inf=\bar{X}-1.96\,\sigma/\sqrt{n} \quad <\mu< \quad Sup=\bar{X}+1.96\,\sigma/\sqrt{n}$$

Esto implica entonces que en el intervalo generado por la muestra se incluye el parámetro μ con una probabilidad de .95. Como en este caso hemos supuesto σ conocida entonces este intervalo puede ser calculado tan pronto se tengan los datos muestrales.

De manera general, cuando n es grande y σ es conocida, el intervalo de confianza al $100(1-\alpha)\%$ para μ esta dado por:

$$(\bar{X}-z_{\alpha/2}\,\sigma/\sqrt{n},\ \bar{X}+z_{\alpha/2}\,\sigma/\sqrt{n})$$

donde $\alpha/2$ denota el punto del la distribución normal estandar cuya área hacia las orillas es la probabilidad no considerada. Ilustremos esto:



Así por ejemplo para:

$(1-\alpha)$	.80	.85	.90	.95	.99
$z_{\alpha/2}$	1.28	1.44	1.64	1.96	2.58

Remarquemos las características de este intervalo:

a) Un intervalo de confianza $(\bar{X}-z_{\alpha/2}\,\sigma/\sqrt{n},\ \bar{X}+z_{\alpha/2}\,\sigma/\sqrt{n})$ es un intervalo aleatorio que trata de cubrir el valor verdadero de μ.

b) La probabilidad

$$P(\bar{X}-1.96\,\sigma/\sqrt{n}<\mu<\bar{X}+1.96\,\sigma/\sqrt{n})=.95$$

interpretada como una frecuencia relativa sobre muchas repeticiones de

muestreo, establece que aproximadamente el 95% de los intervalos generados por cada muestra cubren μ.

c) Cuando $\bar{x}$ es calculado a partir de una muestra. El intervalo ($\bar{x}$-1.96 $\sigma/\sqrt{n}$, $\bar{x}$+1.96 $\sigma/\sqrt{n}$), que es una realización de un intervalo aleatorio, es presentado como un intervalo de confianza del 95% para μ. Habiendo determinado un intervalo numérico, no se debe hablar de el como una probabilidad.

Recuerdese que lo que se establece es que el 95% de los posibles intervalos cubren el valor del verdadero parámetro.
Por otro lado es necesario recalcar que estos intervalos no implican que el verdadero valor del parámetro se localice en algún punto específico del intervalo, sino que solamente el valor buscado se encontrará entre los valores generados por el intervalo.

b) Intervalos de confianza para μ con muestras grandes y σ desconocida.
Cuando n es grande y la σ de la población es desconocida, el intervalo de confianza para μ al 100(1-α)% esta dado por:

$$(\bar{x}-z_{\alpha/2}\ s/\sqrt{n},\ \bar{x}+z_{\alpha/2}\ s/\sqrt{n})$$

donde s es la desviación estandar de la muestra.

Y ninguna consideración sobre la varianza de la población es necesario hacer. Que como veremos más adelante para las muestras pequeñas será necesario.

c) Intervalo de Confianza para proporciones:

De manera similar a la media con muestras grandes, el intervalo de confianza en muestras grandes para el parámetro p, se basa en el estimador $\hat{p}$.
Como ya vimos por el teorema del límite central $\hat{p}$ tienen una distribución aproximadamente normal con media p y desviación estandar $\sqrt{pq/n}$ cuando n es grande y $\hat{p}$ no es cerana a 0 ó 1.

Note que aquí se tiene un doble problema ya que p, el parámetro desconocido, interviene tambien en el cálculo de la desviación estandar.
En ambos casos utilizamos a $\hat{p}$ como estimador (esto es posible siempre y

cuando el tamaño de muestra sea grande).

Así el intervalo de confianza para p es:

$$(\hat{p}-z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n},\ \hat{p}+z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}).$$

5.4 Distribucion t

Intervalo de confianza para μ con muestras pequeñas

Anteriormente hemos calculado los intervalos de confianza para μ basados en que las $\bar{X}$ se distribuyen aproximadamente $\mathcal{N}(\mu,\sigma/\sqrt{n})$ cuando n es grande. En estos casos no hemos tenido que hacer ninguna suposición adicional sobre la distribución de la población, debido al teorema central del límite. Pero en mucho campos, especialmente en aquellos en que la investigación es costosa o cuyos sujetos de estudio son escasos, es necesario limitar el tamaño de muestra.

Cuando n es pequeña ($n<30$) no siempre se cumple la suposición sobre la distribucion normal de $\bar{X}$. El tipo de inferencia que se haga dependerá de las suposiciones sobre la distribución de la población. En este capítulo presentamos un método de construcción de intervalos de confianza para μ considerando que la población original se distribuye aproximadamente de forma normal. Por lo tanto es importante señalar que este procedimiento no funcionará si la distribución de la poblacion difiere mucho de una normal.

Si $X_1,\ldots,X_n$ son una muestra aleatoria de una población distribuida normalmente con $\mathcal{N}(\mu,\sigma)$, la distribución muestral de $\bar{X}$ estará exactamente distribuida como $\mathcal{N}(\mu,\sigma/\sqrt{n})$. Si s es conocida, un intervalo de confianza al $100(1-\alpha)\%$ para μ, esta dado por $\bar{X}\pm z_{\alpha/2}\ \sigma/\sqrt{n}$. Este intervalo es obtenido a partir de la distribución normal estandarizada, es decir de:

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

Para el cual:

$$P[-z_{\alpha/2} < Z < z_{\alpha/2}] = 1-\alpha.$$

Cuando σ es desconocida, como es frecuente, intuitivamente lo que quisieramos es reemplazar σ por s (desviación estandar de la muestra), consideremos la relación:

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

Notese que aquí reemplazamos σ por s, y z por t, esto último es requerido porque la variable aletoria en el denominador incrementa la varianza en un valor mayor a uno, y la relación deja de ser una variable estandarizada.

Esta distribución t es conocida como prueba de " t de student", que es una contribución de W.S. Gosset en 1908, publicada bajo el pseudónimo de "Student", esta distribución la obtuvo de manera empírica y mas tarde fue verificada de manera analítica.

Si $X_1,.....,X_n$ son una muestra aleatoria de una población normal $\mathcal{N}(\mu,\sigma)$, y

$$\bar{X} = 1/n \sum X_i \qquad \text{y} \qquad s = \sqrt{(\Sigma(X_i-\bar{X})^2)/(n-1)}$$

entonces la distribución de

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

es llamada distribucion *"t de student" con n-1 grados de libertad*

Los (n-1) grados de libertad son necesarios, debido a que para cada tamaño de muestra la distribución t difiere. Y recuerde que para un tamaño de muestra n el estimador s^2 de σ^2 esta basado en (n-1) grados de libertad. Las distribuciones t son simétricas alrededor de 0 pero tiene las colas mas levantadas que la $\mathcal{N}(0,1)$. Pero entre mas grados de libertad tenga las distribuciones t se van aproximando a la $\mathcal{N}(0,1)$. Esto es congruente con nuestros anteriores enunciados, a medida que n es grande la relación

$\frac{\bar{X}-\mu}{s/\sqrt{n}}$ se aproxima mas a la normal estandarizada.

Para determinar el intervalo de confianza para μ, será:

$$P[-t_{\alpha/2} < ((\bar{X}-\mu)/(s/\sqrt{n})) < t_{\alpha/2}] = 1-\alpha$$

donde $t_{\alpha/2}$ esta tabulado en la distribución t con (n-1) grados de libertad.
Rearreglando los términos como anteriormente:

$$P[\bar{X}-t_{\alpha/2}\, s/\sqrt{n} < \mu < \bar{X}+t_{\alpha/2}\, s/\sqrt{n}] = 1-\alpha$$

Resumen:
Si la distribución de la población es normal y si σ es desconocida, el intervalo de confianza para μ a un $100(1-\alpha)\%$ es:

$$\bar{X} \pm t_{\alpha/2}\, s/\sqrt{n}$$

donde $t_{\alpha/2}$ esta tabulado en tablas con n-1 grados de libertad.
Es decir que al obtener un intervalo de confianza para μ tendremos una probabilidad de $(1-\alpha)\%$ de cubrir dicho parámetro μ.

5.5 Determinacion del tamaño de muestra

Hemos visto que en la determinación del error estandar $\sigma/\sqrt{n}$ tiene un rol fundamental, ya que al aumentar o disminuir el tamaño de la muestra disminuimos o aumentamos el error. Pero si bien lo que quisieramos es aumentar siempre el tamaño de muestra esto es imposible porque resulta muy costoso tanto en términos económicos como de procesamiento de datos. En la práctica lo que buscamos es optimizar este tamaño, es decir encontrar un valor para n lo suficientemente grande para disminuir la imprecisión de la estimación, y lo suficientemente pequeño para no elevar los costos.

a) Estimación para medias.

Al estimar una media poblacional el investigador deberá tomar la decisión de cuanto es el error tolerable, así como determinar la probabilidad en la que este error puede ser presentado.

Supongamos que la desviación estandar de la población σ es conocida.

Recordemos que la fórmula para una cota de error al $100(1-\alpha)\%$ la estimación de μ por $\bar{X}$ esta dada por :

$$z_{\alpha/2}\,\sigma/\sqrt{n}$$

Esta relación da como resultado una cantidad d esta razón nos permitirá asegurar que el error al $100(1-\alpha)\%$ no excede a dicha cantidad d es decir que $z_{\alpha/2}\,\sigma/\sqrt{n}=d$ de aquí podemos despejar n, así nos queda:

$$n=\left[\frac{z_{\alpha/2}\,\sigma}{d}\right]^2$$

que nos permitirá determinar el tamaño de la muestra. Claro que tendremos que redondear este valor al próximo mas alto ya que un tamaño de muestra no puede ser fraccional.

Otra manera similar para determinar el tamaño es a partir del intervalo mismo. El investigador requiere un tamaño de n tal que para un intervalo de confianza al $100(1-\alpha)\%$ para μ tenemos una longitud específica c. Así tenemos que la longitud del intervalo $(\bar{X}-z_{\alpha/2}\,\sigma/\sqrt{n},\ \bar{X}+z_{\alpha/2}\,\sigma/\sqrt{n})$ es $2z_{\alpha/2}\,\sigma/\sqrt{n}$, el tamaño de muestra n es determinable por:

$$2z_{\alpha/2}\,\sigma/\sqrt{n}=c \quad \text{ó} \quad z_{\alpha/2}\,\sigma/\sqrt{n}=c/2$$

Observe que estos procedimientos son equivalentes, ya que c/2 es lo mismo que d.

Esto esta basado en la distribución normal de $\bar{X}$, es decir que esto es válido solamente para muestras grandes o siempre y cuando la población tenga una distribución normal. Sino podremos hacer uso del desigualdad de Chebyshev que establece:

$$P[|\bar{x}-\mu|>k\sigma]<1/k^2$$

$$P[|\bar{x}-\mu|<k\sigma]>1-1/k^2$$

Para la media de la muestra, utlizando el teorema del límite central reemplazamos σ por $\sigma/\sqrt{n}$ y $\bar{x}$ por $\bar{X}$ así nos queda:

$$P[|\bar{X}-\mu|<k\sigma/\sqrt{n})>1-k^2$$

y, haciendo $k\sigma/\sqrt{n}=d$, obtenemos:

$$P[|\bar{X}-\mu|<d])>1-\sigma^2/nd^2$$

Si esta segunda parte de la desigualdad la identificamos con $1-\alpha$, el tamaño de muestra esta dado por:

$$n=\sigma^2/\alpha d^2$$

Resumen:
Para estar seguros a un $100(1-\alpha)\%$ de que el error entre $|\bar{X}-\mu|$ no es mayor que d, el tamaño de muestra requerido esta dado por:

$$n=\left[\frac{z_{\alpha/2}\,\sigma}{d}\right]^2$$

Si n es pequeña y la población no es normal entonces emplee:

$$n=\sigma^2/\alpha d^2$$

Notese que en ambos casos el conocimiento de σ o al menos una aproximación de ella es necesario. Si σ es completamente desconocida, debera realizarse un muestreo preliminar para tener un estimador de σ y poder calcular n.

b) Estimación para proporciones:

La cota de error al $100(1-\alpha)\%$ para la estimación de p esta dada por $z_{\alpha/2}\sqrt{pq}/\sqrt{n}$, reemplazando σ^2 por pq, tenemos:

$$n=pq[\,z_{\alpha/2}\,/d\,]^2$$

Aunque en realidad esta solución no es muy utilizable ya que involucra p el parámetro que tratamos de estimar.
Sabemos que p fluctua entre 0 y 1, el valor máximo de pq será 1/4 y decrece hasta cero, de ahí deducimos que:

$$n<1/4\,[\,z_{\alpha/2}\,/d\,]^2$$

Prueba de hipótesis

6.1 La hipótesis nula y la hipótesis alternativa

Otro aspecto de la estadística inferencial, ademas de la estimación de parámetros, es la prueba de hipótesis que trata de evaluar la validez de una conjetura. Generalmente la conjetura es realizada sobre alguna característica de la población, y es a partir de la información de la muestra que trataremos de comprobar su validez. Así una hipótesis estadística es cualquier suposición sobre la población, que puede ser evaluada en base a la información obtenida al muestrar la población.

Una suposición puede ser falsa o verdadera, siendo complementarias entre sí:

hipótesis H : esta suposición es verdadera.
hipótesis H' : esta suposición es falsa.

Usando la información de las observaciones muestrales, el tomador de decisiones debe elegir entre dos decisiones o inferencias:

Sea: Rechazar H' y concluir que H esta fuertemente apoyada por los datos.

O : No rechazar H' y concluir que H no esta apoyada por los datos.

El proceso mediante el cual se hace la selección entre estas dos acciones se denomina Prueba de Hipótesis.

En estadística denominamos hipótesis alternativa H_1 aquella que es postulada por el investigador, e hipótesis nula H_0 aquella que es el complemento de la primera (aquella que no aporta nada nuevo).

La hipótesis nula H_0 deberá ser considerada como verdadera y soló podrá ser rechazada si los datos testifican fuertemente en contra de ella. Es como en el juzgado, la hipótesis nula es "inocente" a menos que haya suficiente evidencia para declararlo "culpable". Entonces los roles de H_0 y de H_1 no son simétricos.

El término de hipótesis nula fue originado a partir de la comparación experimental de nuevos productos ó técnicas al ser comparados con los usuales, y esta hipótesis postula que la diferencia entre lo nuevo y lo estandar es nula o cero.

Una prueba o test sobre la hipótesis nula consistirá en especificar el conjunto de valores de la variable $\bar{X}$ para los cuales H_0 es rechazada. La variable aletoria cuyos valores son determinados se denomina estadístico de prueba, y el grupo de valores en los cuales H_0 es rechazada se denominá región de rechazo de la prueba. Entonces una prueba de hipótesis estará completamente especificada si se determina el estadístico de prueba y la región de rechazo.

Los pasos principales para la prueba de hipótesis son:

a) Identificar el modelo de probabilidad adecuado y el (ó los) parámetro (s) involucrados en la hipótesis formulada.

b) Fomular la hipótesis nula H_0 y la hipótesis alterna H_1.

c) Seleccionar el estadístico de prueba y determinar la estructura de la región de rechazo.

d) Especificar la distribución muestral del estadístico de prueba bajo la hipótesis nula, y determinar la región de rechazo, según el a específico.

e) Implementar la prueba y elaborar las conclusiones.

6.2 Los dos tipos de errores y el poder de una prueba

En este tipo de prueba podremos cometer dos tipos de errores, que consisten en:

	H_0	
	Verd.	Falso
No rechazo H_0	✓	Error II
Rechazo H_0	Error I	✓

ó

	H_1	
	Verd.	Falso
No rechazo H_1	✓	Error I
Rechazo H_1	Error II	✓

(ambos cuadros son equivalentes solamente que estan presentados ya sea bajo H_0 o bajo H_1)

Así el :

Error I consiste en : Cuando la hipótesis nula es verdadera la rechazamos (cuando no rechazamos la hipótesis alterna siendo falsa).

Error II consiste en : No rechazar la hipótesis nula cuando es falsa (o cuando rechazamos la hipótesis alterna siendo verdadera).

Las probabilidades de ambos errores es:

α= P(error tipo I) = P(rechazar hipótesis nula cuando verdadera).

β= P(error tipo II)= P(no rechazar la hipótesis nula cuando falsa).

Nótese : la probabilidad α depende de los valores particulares que el parámetro puede tomar bajo H_0 , sin embargo β depende del rango de valores cubiertos por H_1.

Al aumentar una probabilidad disminuimos la otra y viceversa, la única manera de disminuir ambas simultáneamente es aumentando el tamaño de la muestra, o sea incrementando n.

Ilustremos α y β, osea calculemos la probabilidad del error de tipo I y II para un caso dado, supongamos el caso de rechazar H_0: X >= 15 de una distribución binomial para n=20, entonces H_0 sera aceptada si X=<14. Los valores específicos de p los podemos obtener en tablas.

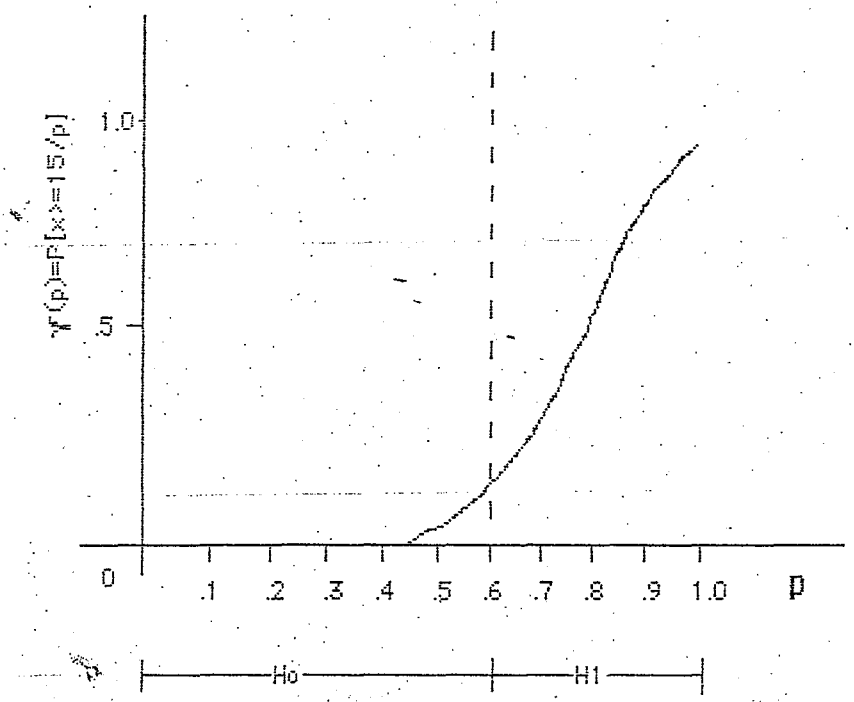
Denotemos la función de probabilidad de rechazo de una prueba por $\gamma(p)$, osea:

$\gamma(p)$= P[la prueba rechaza H_0 cuando el valor verdadero del parámetro es p]

En el caso que nos ocupa, obtendremos de tablas los valores específicos de $\gamma(p)$ para cada p correspondiente

p	.3	.4	.5	.6	.7	.8	.9
$\gamma(p)$=P[x>=15/p]	.00	.002	.021	.126	.416	.804	.989

Es decir que bajo H_0, p esta restringida al rango p=<.6, osea que $\gamma(p)$ en este rango equivale a $\alpha(p)$ probabilidad de error I. Bajo H_1, el rango p es p>.6, osea $1-\gamma(p)$= P[error tipo II]=$\beta(p)$, al graficar esta información obtenemos:



Probabilidad de error tipo I $\alpha(p)=\gamma(p)$ para p=<.6

Probabilidad de error tipo II $\beta(p)= 1-\gamma(p)$ para p>.6

La porción p>.6 cuando H_1 es verdadera, osea la cantidad $\gamma(p)=1-\beta(p)$ es llamada la potencia de la prueba. Esta es la probabilidad de rechazar H_0 cuando H_1 es realmente verdadera, y es en este sentido que esta potencia indica la fortaleza o bondad de la prueba.

Prueba de una o dos colas

Podemos identificar dos grandes tipos de prueba de hipótesis dependiendo de la hipótesis alterna formulada. Hemos establecido que el complemento de ella es la hipótesis nula y en función de esta hipótesis nula. Establecemos los valores particulares de la región de rechazo (en H_0).

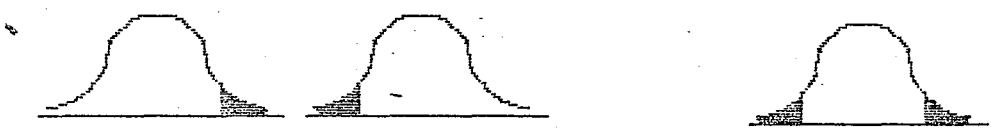
Dos grandes tipos de hipótesis alterna pueden ser establecidas:

a) La comparación Ej: H_1: p<.6 H_1: X>Y H_1:$\theta<\theta_0$

b) La diferencia Ej: H_1:p=.6 H_1:X=Y H_1:$\theta=\theta_0$

La comparación establecerá entonces una distribución de la hipótesis nula cuya región de rechazo se encontrará solamente en algunas de las colas de la distribución muestral del parámetro en estudio. Mientras que en la comparación al no saber hacia donde se situa la distribución de la hipótesis alterna, entonces la región de rechazo en la distribución de la hipótesis nula será en ambas colas de ella.

Prueba de una cola Prueba de dos colas



6.3 Prueba de hipótesis sobre la media de una población.

Primeramente consideremos el caso de muestras pequeñas. Utilizaremos la suposición de que la distribución de la población es normal, así como que la muestra fue tomada alteatoriamente. La distribución de la muestra es $\mathcal{N}(\mu,\sigma)$ y la distribución de $\bar{X}$ es $\mathcal{N}(\mu,\sigma/\sqrt{n})$, como vimos en la estimación el estadístico de prueba apropiado será $\bar{X}$.

Supongamos la hipótesis:

$$H_0: \mu \leq \mu_0 \qquad H_1: \mu > \mu_0$$

donde μ_0 es un número específico, que representa un parametro dado. La distribución muestral de $\bar{X}$ suponemos (como investigador) que está centrada en μ, así los valores de $\bar{X}$ tienen mas probabilidad bajo H_1 que bajo H_0, segun nuestros supuestos teóricos que tratamos de comprobar. La

región de rechazo de H_0 en favor de H_1 será para los valores observados de $\bar{X}$ bastante grandes, es decir estamos hablando de prueba de una cola y la región de rechazo está estructurada:

$$R : \bar{X} >= c$$

La frontera c de la región de rechazo debe ser determinada en función de la tolerancia de un error de tipo I, osea de la probabilidad α. Osea que $P_{\mu_0}[\bar{X} >= c] = \alpha$. Bajo la hipótesis nula $\mu = \mu_0$, la distribución muestral de $\bar{X}$ será una $N(\mu_0, \sigma/\sqrt{n})$. Al estandarizarla nos queda entonces

$$Z = \frac{(\bar{X} - \mu_0)}{\sigma/\sqrt{n}}$$

Entonces el evento $\bar{X} >= c$ será:

$$\frac{(\bar{X} - \mu_0)}{\sigma/\sqrt{n}} >= \frac{(c - \mu_0)}{\sigma/\sqrt{n}}$$

probabilísticamente obtenemos que:

$$P_{\mu_0}[\bar{X} >= c] = P\left[Z >= \frac{(c - \mu_0)}{\sigma/\sqrt{n}}\right]$$

Recordemos que $P[Z >= z_\alpha] = \alpha$, que es la probabilidad de error tolerado de tipo I. Relacionando con la ecuación anterior tenemos que :

$$\frac{(c - \mu_0)}{\sigma/\sqrt{n}} = z_\alpha$$

Despejando c de ella obtenemos $c = \mu_0 + z_\alpha(\sigma/\sqrt{n})$ la región de rechazo

de la hipótesis nula quedará especificada por:

$$R: \bar{X} > = \mu_0 + z_\alpha (\sigma / \sqrt{n})$$

En este caso la prueba de hipótesis es una comparación, osea de una sola cola, cuando se trata de una prueba de dos colas la región de rechazo será entonces :

$$R: \bar{X} > = \mu_0 + z_{\alpha/2} (\sigma / \sqrt{n})$$

6.4 Prueba sobre la proporción de una población.

Es decir para prueba de hipótesis del tipo $H_0 : p=p_0$ vs. $H_1: p \neq p_0$ para un gran número de ensayos ($n>30$). Si consideramos la proporción de éxitos de la muestra $\hat{p}=X/n$, como hemos visto su distribución es normal alrededor de p y con desviación estandar $\sqrt{pq/n}$. Entonces bajo la hipótesis nula tendremos una distribuición de p aproximadamente $\mathcal{N}(p_0, \sqrt{p_0 q_0 / n})$. Consecuentemente el estadístico estandarizado es:

$$Z = (p - p_0) / (\sqrt{p_0 q_0 / n}) \text{ se distribuye } \mathcal{N}(0,1)$$

En el caso que hemos postulado la hipótesis alternativa es de dos colas por lo tanto la región de rechazo a un nivel α de tolerancia esta dada por:

$$R: |Z| > = z_{\alpha/2}$$

Para el caso de una cola utilizamos entonces:

$$R: |Z| > = z_\alpha$$

Obsérvese alternativamente podemos cambiar la proporción a su cantidad correspondiente. Ya que $\hat{p}=X/n$, tenemos que :

$$Z = (X/n - p_0) / (\sqrt{p_0 q_0 / n}) = (X - np_0) / (\sqrt{n p_0 q_0})$$

BIBLIOGRAFIA

BHATTACHARYYA, G. K.; JOHNSON, R. A.
Statistical concepts and methods. Wiley, 1977.

EHRENBERG, A. S. C.
Data reduction. Wiley, 1981.

KREYSZIG, E.
Introducción a la estadística matemática. Limusa, 1979.

MENDENHALL, W.; Sheaffer, R. L.; Wackerly, D. D.
Estadística matemática con aplicaciones. Grupo Editorial Iberoamérica, 1986.

MILLER, J.; Freund, J. E.
Probabilidad y estadística para ingenieros. Reverté, 1980.

TUFTE, Edward R.
The visual display of quantitative information. Graphics Press, 1983.