centro panamericano de fiebre aftosa

SERIE DE MANUALES DIDACTICOS

Nº 11

INFERENCIA ESTADISTICA EN SALUD ANIMAL

organización panamericana de la salud
oficina sanitaria panamericana, oficina regional
de la organización mundial de la salud

ORGANIZACION PANAMERICANA DE LA SALUD
Oficina Sanitaria Panamericana, Oficina Regional de la
ORGANIZACION MUNDIAL DE LA SALUD

CENTRO PANAMERICANO DE FIEBRE AFTOSA
CAIXA POSTAL 589 - ZC/00 - RIO DE JANEIRO, BRASIL

INFERENCIA ESTADISTICA EN SALUD ANIMAL

por

Vicente M. Astudillo

Melba Wanderley

1 9 7 7

INFERENCIA ESTADISTICA

Vicente M. Astudillo
Melba E. Wanderley

1977

1. INFERENCIA ESTADISTICA

Uno de los objetivos fundamentales de la Estadística consiste en el estudio de las relaciones existentes entre una población y las muestras extraídas desde ella. Se entiende por población, la totalidad de las mediciones o cuentas obtenidas a partir de todas las unidades o individuos que tienen en común el carácter en estudio. Muestra es un subconjunto de mediciones obtenido de la población de manera que sea representativo de ella. En la práctica, al querer estudiar un determinado carácter, por una u otra razón, no se puede hacerlo en la población, de manera que se recurre a una muestra. El principal objetivo al tomar una muestra es obtener alguna conclusión acerca de la población de la cual fue elegida.

El problema que preocupa ahora, es proyectar aquello que se ha observado en un número limitado de unidades o individuos, a la población. La relación entre la población y las muestras es lo más importante dentro de la teoría estadística y desde el punto de vista práctico. Se necesitan muestras representativas para poder proyectar los resultados desde las observaciones muestrales, particulares, a la población. El procedimiento habitual para obtener una muestra representativa, es la selección aleatoria. Se conoce como muestra aleatoria aquella en que cada unidad poblacional tiene igual probabilidad de ser elegida que otra.

De hecho la información muestral puede ser utilizada para avanzar afirmaciones acerca de las características poblacionales. Este es un proceso de naturaleza inductiva; sin embargo, es conveniente tener presente que estas afirmaciones en ningún caso se traducen en expresiones de certeza, sino que sólo están respaldadas por cierto grado de probabilidad, que permite apreciar su nivel de incertidumbre.

Parámetros y estadísticos

Parámetro es un valor característico de la población, o sea, es una expresión numérica que sintetiza una propiedad de todos los N elementos de la población. Ej.: una media, una tasa. La población se define como el agregado total de elementos, N, que poseen una propiedad que es medible o enumerable. Estadístico es el valor característico correspondiente a una muestra, o sea, es una expresión numérica que sintetiza una propiedad que presentan los n elementos que componen la mues-

tra. Esta es una fracción de la población, cuyos componentes han sido elegidos mediante un proceso al acaso.

Los parámetros poblacionales están basados en todas las unidades de ella, en cambio los estadísticos muestrales se basan sólo en una parte de la población, por lo cual estos estadísticos varían de una muestra a otra y de ahí la incerti - dumbre.

Se llama parámetro a un valor característico en la población. Se llama estadístico al valor característico calculado en la muestra. En síntesis, en la inferencia estadística, se utiliza una muestra, a partir de la cual, por inducción, se proyectan los resultados a la población.

Al conocer los valores característicos de una muestra de n observaciones, pueden surgir ciertas interrogantes acerca de lo que se puede decir de los paráme tros poblacionales, después de haber examinado, solamente, un número limitado de individuos. También puede inquietar el grado de confianza que merecen los estadís ticos obtenidos, como buenas estimaciones de los parámetros.

Para una propiedad dada, el valor del parámetro es uno sólo ya que depen de de los N elementos. Es una constante que no depende de fluctuaciones de selec ción de los elementos, porque son considerados todos. El valor del estadístico se calcula a partir de los n elementos seleccionados. Por lo tanto, la estimación que se hace de un parámetro a través de un estadístico, es solamente una de las tantas estimaciones que pudieron hacerse con el mismo diseño de muestreo.

Teoría elemental de muestreo

La teoría de muestreo estudia las relaciones existentes entre una pobla- ción y las muestras obtenidas a partir de esa población. Esta teoría es útil pa ra:

a) Estimar un valor de la población desconocido (parametros como ser media aritmética, tasa, varianza, etc.) a partir de los correspondientes valo res muestrales conocidos (estadísticos).

b) Dilucidar si las diferencias observadas entre dos estadísticos muestrales son debi- das a la variación aleatoria o si ellas son realmente significativas.

En términos generales, la teoría de muestreo permite indicar la confianza o seguridad que podemos tener en una inferencia estadística. Esto es al hacer generalización acerca de una propiedad de la población a través del uso de muestras obtenidas desde ella.

Para que las aseveraciones de la inferencia estadística sean válidas, las muestras deben ser representativas de la población. Una forma de conseguir una muestra sea representativa de la población de la cual es obtenida es aplicar un proceso aleatório, en el cual cada miembro de una población tiene la misma chance de ser incluido en la muestra. Una técnica para obtener muestras aleatorias es asignar un número a cada componente de la población, escribiendo simultaneamente el número en un pedazo de papel o en una ficha de plástico o de madera (lotería). Estos papeles o fichas son colocados en una urna o bolsa, siendo bien mezcladas y se procede a hacer tantas extracciones como componentes se hayan considerados necesarios en la muestra. Este procedimiento puede ser remplazado por el uso de una tabla de números aleatorios.

En este punto se deben considerar dos posibilidades:

a) que el número elegido sea devuelto a la urna antes de una segunda elección

b) que el número elegido no sea devuelto a la urna antes de una segunda elección

En el primer caso el componente representado por el número seleccionado puede ser incluido varias veces en la muestra. Este procedimiento se llama muestreo aleatorio con reposición. En el segundo caso el componente representado por el número seleccionado no puede ser elegido más que una sola vez para su inclusión en la muestra. Este procedimiento es llamado muestreo aleatorio sin reposición.

A partir de una población finita como la compuesta por N=5 elementos o componentes

A, B, C, D, E

podemos obtener infinita cantidad de muestras de tamaño n=2, si la selección es hecha con restitución, puesto que cualquier cantidad de muestras de n=2 pueden ser obtenidas sin agotar la población. En cambio, si se aplica un procedimiento sin reposición para la selección de los 2 elementos componentes de la muestra, la cantidad de muestras de tamaño n=2 es igual a 10:

AB, AC, AD, AE, BC, BD, BE, CD, CE, DE.

cantidad de muestras que puede ser calculada a través de

$$\frac{N!}{n!\ (N-n)!} = \frac{1x2x3x4x5}{(1x2)\ (1x2x3)} = \frac{120}{12} = 10$$

Distribuciones teóricas de muestreo. Teoría de muestras grandes.

Si consideramos todas las posibles muestras de un tamaño determinado, $n > 30$, que pueden ser obtenidas a partir de una población, utilizando, sea un procedimiento con reposición, sea un procedimiento sin reposición, para cada muestra se puede calcular un estadístico, como una tasa o una media aritmética, estadístico que variará de una muestra a otra.

Dado que los estadísticos provenientes de varias muestras seleccionadas desde una población, presentan variación, es posible pensar que ellos orginan distribuciones de frecuencias, para poder desarrollar procedimientos estadísticos que se quieren estudiar. Las distribuciones de los estadísticos se llaman distribuciones de muestreo, ya que la variación entre èllos es producto del proceso de muestreo. Es comprensible que de una población podríamos obtener una cantidad infinita de muestras, todas del mismo tamaño. De esta manera se puede concebir que para un dado estadístico podemos tener una distribución de probabilidades.

Errores de muestreo

Los estadísticos están sujetos a errores de muestreo y también a errores que no son de muestreo. Los errores de muestreo ocurren porque la muestra continene solo una parte de los N elementos de la población. Para una población de N bovinos podemos tener una tasa de infección por fiebre aftosa, P. Al seleccionar una muestra de n bovinos, a partir de esa población, la tasa muestral p depende de los n bovinos seleccionados, Diferentes muestras de n bovinos, o sea del mismo tamaño, seleccionados con el mismo diseño a partir de la misma población, producirán diferentes valores de p. La desviación de cualquier valor de p desde P es desconocida, puede ser grande o pequeña, positiva o negativa.

El conocimiento de la variabilidad del muestreo se hace en términos de probabilidades. Dado un diseño de muestreo y un tamaño de la muestra podemos preguntar cuál es la probabilidad de que ocurra un cierto valor de p. El conjunto de valores posibles de p, cada cual con su probabilidad de ocurrir (frecuencia relativa) forman una distribución hipotética de muestreo de las posibles tasas de p de infección por fiebre aftosa, para una población determinada, con un tamaño de la muestra y procedimiento de selección de los elementos, fijos.

La media de tal distribución hipotética de muestreo de las posibles tasas p corresponde al valor esperado de estimador, o sea, el valor del parámetro poblacional P.

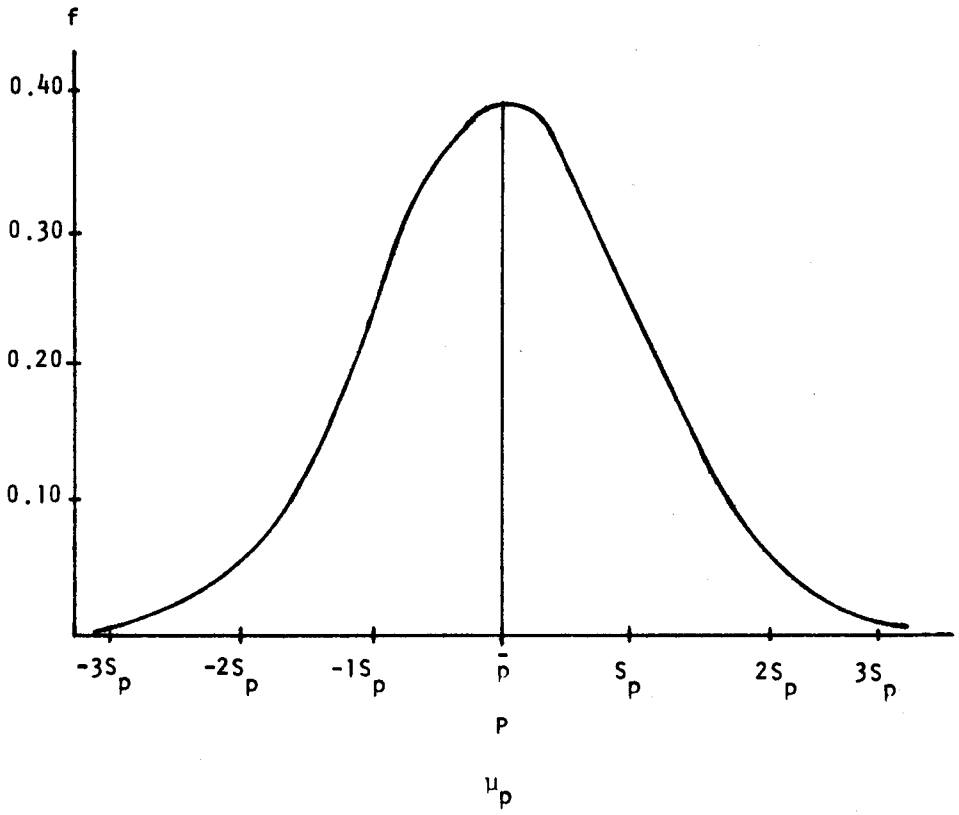
$$E(p) = \Sigma\ f(p) = P$$

tenemos que considerar varias condiciones, como ser que desde el punto de vista estadístico probabilidad corresponde a frecuencia relativa (f) de una distribución; que el procedimiento de selección de los elementos producen estimaciones no viciadas, es decir, $E(p) = P$; que las muestras hipotéticas tienen un tamaño $n \geq 30$ por lo cual la distribución de muestreo de p puede ser considerada como normal.

Por otra parte la desviación estándar de la distribución hipotética de muestreo de p se llama error estándar. En este punto debemos tener en cuenta que la distribución de muestreo referida es hipotética, ya que en la realidad conocemos un solo punto de toda la distribución de muestreo. Cuando se emplea un muestreo simple al acaso con restitución, el error estándar sería $\sigma_p = \sqrt{\frac{PQ}{n}}$

Sin embargo, el valor de error estándar de la población no se conoce, porque su valor depende de toda la población y por tanto no se puede calcular a partir de una muestra. Sin embargo, se ha desarrollado un estadístico para estimar el error estándar a partir de la muestra.

$$S_p = \sqrt{\frac{pq}{n}}$$


f
0.40
0.30
0.20
0.10
-3S_p
-2S_p
-1S_p
p̄
S_p
2S_p
3S_p
P
μ_p

Distribución teórica de muestreo de una tasa (p).

Tomemos como ejemplo una población bovina donde hace algún tiempo fue notificada la ocurrencia de fiebre aftosa. Por estudios serológicos se ha establecido que la tasa de prevalencia de infección por el virus de la fiebre aftosa (presencia de anticuerpos anti VIA) es P = 20%, en un momento dado. Entonces podemos decir que la probabilidad de encontrar un bovino infectado (positivo al VIA) en esa población es P = 20%.

Supongamos que se han elegido 150 muestras, cada una compuesta por 100 bovinos (n), seleccionados al azar desde la población. Para cada muestra se determina el estadístico muestral p, que representa la tasa (%) de infectados (positivos al VIA). Habrá por tanto 150 valores de p: $p_1, p_2, p_3, \ldots, p_{150}$. Este estadístico p representa variaciones, ya que no es igual en todas las 150 muestras.

Para valores grandes de tamaño de la muestra ($n \geq 30$), el estadístico p tiene una distribución que se aproxima mucho a la distribución normal con

$$\mu_p = P \qquad \sigma_p^2 = PQ/n \qquad \sigma_p = \sqrt{\frac{PQ}{n}}$$

desarrollemos el ejemplo propuesto

p (%)	Nº de muestras	Nº de positivos al VIA
18	15	270
19	35	665
20	50	1.000
21	30	630
22	20	440
Total	150	3.005

$\mu_p = P = (3.005/150) = 20 = \bar{p}$

$\sigma_p^2 = (20)(80)/100 = 16$

$\sigma_p = 4$

Por lo tanto la variable aleatoria p da origen a una distribución normal estandarizada a través de

$$Z = \frac{p - P}{\sigma_p}$$

supongamos que a partir de una población de vacas se obtiene una muestra al azar de n= 900 animales. Dicha población tiene una tasa de prevalencia para brucelosis de P = 10%, El objetivo del estudio es a través del calculo de p responder cuál es la probabilidad de que p esté comprendida entre 8% y 11%.

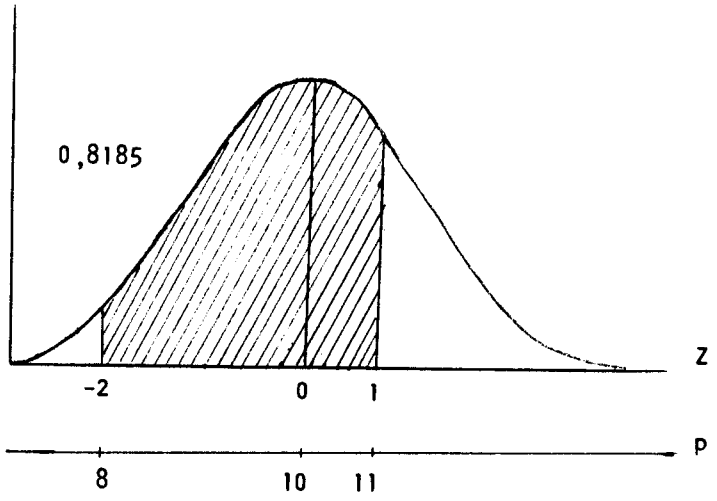
$$P_r\ (8 < p < 11)$$

$$Z = \frac{8 - 10}{\sqrt{\frac{10 \times 90}{900}}} = \frac{-2}{1} = -2,00$$

$$Z = \frac{11 - 10}{\sqrt{\frac{10 \times 90}{900}}} = \frac{1}{1} = 1,00$$

o sea

$$P_r\ (-2,00 < Z < 1,00) = 0,8185$$



Distribución teórica de muestreo de la media aritmética ($\bar{X}$).

Si a partir de una población dada, se obtiene una muestra por algún proceso aleatorio, al describir estadísticamente la muestra se tendrá una media muestral ($\bar{X}$) que es la estimación de la media poblacional (μ_x). Aqui cabe preguntar qué confianza podemos tener en que la diferencia entre μ_x y la $\bar{X}$ obtenida sea tolerablemente pequeña? Para responder esta interrogante debemos necesariamente conocer la distribución de las $\bar{X}$ correspondientes a muestras con el mismo n.

Si a partir de una población de la variable X, se obtienen muchas (r) muestras al azar, todas del mismo tamaño, se tendrá $\bar{X}_1$, $\bar{X}_2$, ... , $\bar{X}_r$, cada una de las cuales es una estimación de la media poblacional, estas estimaciones tienen distintos valores, aunque algunos pueden ser muy parecidos (incluso puede haber dos o más medias iguales), de manera que dan lugar a una nueva variable ($\bar{X}_1$, $\bar{X}_2$, ... $\bar{X}_r$), pudiendo construir con ella una distribución de frecuencias. Para encontrar como esos varios valores de $\bar{X}$ están distribuídos, podemos pensar que $\bar{X}$ es una variable aleatoria. Por tanto la cantidad de muestras del tamaño n establecido se distribuye aleatoriamente de acuerdo al valor de $\bar{X}$.

Tomemos como ejemplo la siguiente población (N = 5) de niveles de anticuerpos de seroprotección en bovinos: 0,0 - 1,0 - 2,0 - 3,0 - 4,0. Si planeamos elegir una muestra de tamaño n = 2 desde esta población, estamos frente a un fenómeno aleatorio y la media muestral $\bar{X}$ se convierte en una variable aleatoria. Para determinar cuales son los valores que posiblemente $\bar{X}$ puede tomar, es necesario listar todas las muestras posibles de n = 2 que pueden ser seleccionadas con reposición ya que un bovino, una vez e escogido, vuelve a la población. El número de arreglos con repetición de N bovinos n a n igual a $5^2 = 25$

Muestra	F	Media	f
0,0 - 0,0	1	0,0	1/25
0,0 - 1,0	2	0,5	2/25
0,0 - 2,0	2	1,0	3/25
1,0 - 1,0	1		
1,0 - 2,0	2	1,5	4/25
0,0 - 3,0	2		
0,0 - 4,0	2		
1,0 - 3,0	2	2,0	5/25
2,0 - 2,0	1		
1,0 - 4,0	2	2,5	4/25
2,0 - 3,0	2		
2,0 - 4,0	2	3,00	3/25
3,0 - 3,0	1		
3,0 - 4,0	2	3,50	2/25
4,0 - 4,0	1	4,0	1/25

MEDIAS

	0,0	1,0	2,0	3,0	4,0
0,0	0,0	0,5	1,0	1,5	2,0
1,0	0,5	1,0	1,5	2,0	2,5
2,0	1,0	1,5	2,0	2,5	3,0
3,0	1,5	2,0	2,5	3,0	3,5
4,0	2,0	2,5	3,0	3,5	4,0

Si existe una variable X que una población de tamaño N con media μ_X y varianza σ^2_X, entonces para la distribución teórica de muestreo de $\overline{X}$ se tiene que

$$\mu_{\overline{X}} = \mu_X \qquad y \qquad \sigma^2_{\overline{X}} = \frac{\sigma^2_X}{n}$$

En este ejemplo

$$\mu_X = 2,0 \qquad \sigma^2_X = 2,0$$

por lo tanto

$$\sigma^2_{\overline{X}} = \frac{2,0}{2} = 1$$

nosotros podemos calcular $\mu_{\overline{X}}$ y $\sigma^2_{\overline{X}}$ directamente a partir de la distribución de muestreo de $\overline{X}$

$\bar{X}$	F	$F\bar{X}$	$F\bar{X}^2$
0,0	1	0,0	0,00
0,5	2	1,0	0,50
1,0	3	3,0	3,00
1,5	4	6,0	9,00
2,0	5	10,0	20,00
2,5	4	10,0	25,00
3,0	3	9,0	27,00
3,5	2	7,0	24,50
4,0	1	4,0	16,00
	25	50,0	125,00

$$\mu_{\bar{X}} = \frac{50,00}{25} = 2,0$$

$$\sigma^2_{\bar{X}} = \frac{125,00 - \frac{(50,00)^2}{25}}{25} = \frac{25}{25} = 1$$

Por tanto se comprueba la definición hecha anteriormente.

La distribución de frecuencias a que da lugar el estadístico $\bar{X}$ (distribución de frecuencias de medias muestrales), tiene las características generales que ya se conocen. Gran concentración en la región central, a partir de la cual hacia ambos lados, las frecuencias son más pequeñas en la medida que se alejan de la porción central. Estas distribuciones de frecuencias tienen una forma semejante a la Curva Normal, especialmente, cuando el número de muestras consideradas es alto y la población original tiene una distribución Normal.

Supóngase que se construye la distribución de muestreo de la media aritmética. Se tiene un número r de muestras, lo que equivale a decir que se tiene un número de r medias aritméticas ($\bar{X}_1$, $\bar{X}_2$, ... , $\bar{X}_r$). La media de esta distribución de muestreo como se sabe es $\mu_{\bar{x}}$, la cual tiende a ser igual a μ_x, la media

poblacional. La desviación estándar de esta distribución de muestreo de $\bar{X}$ se simboliza por $\sigma_{\bar{x}}$, que es igual a la raíz cuadrada de la varianza de la distribución de muestreo ($\sigma_{\bar{x}} = \sqrt{\sigma^2_{\bar{x}}}$). Dicho de otra manera es igual a $\sigma/\sqrt{n}$. Este parámetro es llamado comunmente error estándar. Con la obtención de la media y del error estándar de la distribución queda totalmente definida.

En resumen, para valores grandes de n ($n \geq 30$), la distribución de muestreo de la media aritmética tiende a ser Normal, con media $\mu_{\bar{x}} = \mu_x$ y desviación estándar (llamado error estándar) $\sigma_{\bar{x}}$. A este respecto es necesario recordar un teorema de la Estadística que expresa la seguridad de la tendencia anteriormente descrita a medida que n aumenta.

Teorema del límite central.

Si una población tiene varianza finita (σ^2_x) y media (μ_x), la distribución teórica de muestreo de la media muestral $\bar{X}$, tiende a la distribución Normal con varianza $\sigma^2_{\bar{x}} = \dfrac{\sigma^2_x}{n}$ y media $\mu_{\bar{x}} = \mu_x$ al aumentar el tamaño (n) de la muestra.

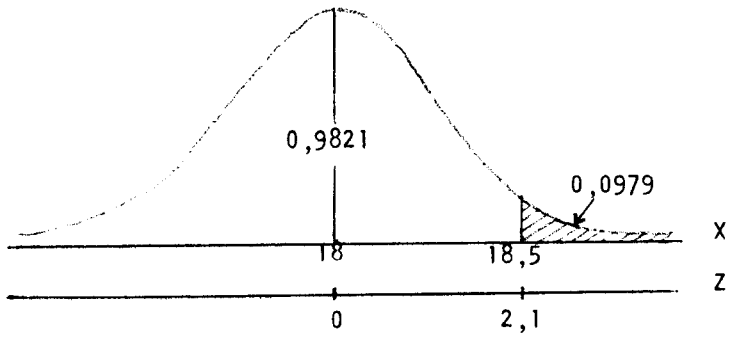
De esta manera la variable aleatoria $\bar{X}$, media de una muestra al azar, de tamaño n, obtenida desde una poblaciôn infinita con media μ_x y varianza σ^2_x se aproxima a una distribución normal con media μ_x y varianca σ^2_x/n, con lo cual el error estándar de $\bar{X}$ es $\sigma_x/\sqrt{n}$. La variable aleatoria $\bar{X}$ da origen a una distribución normal estandarizada (Z) a través de

$$Z = \frac{\bar{X} - \mu_x}{\sigma_x/\sqrt{n}}$$

supongamos que a partir de una población de ratas de 40 días se pretende obtener una muestra al azar de n=100 ratas. Supongamos que esta población tiene una media $\mu_x = 18$ gramos y una varianza $\sigma^2_x = 6$ gramos. El objetivo de este estudio es cal calcular $\bar{X}$. Nos preguntamos ¿Cuál es la probabilidad de que $\bar{X}$ tenga un valor mayor que 18,5 gramos?

$$Pr\ (\bar{X} > 18,5) = Pr\left(Z > \frac{18,5 - 18}{2,4/\sqrt{100}}\right)$$

$$= Pr\ (Z > 2,1)$$
$$= 1 - 0,9821$$
$$= 0,0979$$



Método para muestras grandes

En estadística existen dos formas de trasladar los valores característicos de una muestra a la población, ellas son: la estimación y la prueba de hipótesis.

Estimación. La estimación estadística es un método mediante el cual se obtiene información acerca de los parámetros (población) a partir de los estadísticos (muestra). El problema radica en pasar desde hechos conocidos, dados por la muestra (estadísticos), a generalizaciones acerca de ciertas expresiones características de una población, como son los parámetros.

En estadística existen dos formas de estimación de parámetros: la estimación puntual y la estimación por intervalo de confianza.

1.1. Estimación por puntos: Esta es una forma habitual de hacer estimación. Es una estimación a través de un solo valor. Se usa como aproximación del valor exacto del parámetro , de manera que la estimación de éste, se hace a partir de las observaciones muestrales (estadístico). Es decir, la media aritmética muestral es una estimación puntual del parámetro μ : del mismo modo, p es una estimación puntual de P, en el caso de una tasa. Lo mismo ocurre con otros valores característicos.

a) Estimación no viciada. Dado un cierto estadístico, p, si la media aritmética de la distribución de muestreo de ese estadístico es semejante al respectivo parámetro poblacional (P), se dice que el estadístico es un estimador no viciado del parámetro; se parte del supuesto que los valores del estadístico, que dan lugar a la distribución de muestreo correspondiente, han sido obtenidos a partir de muestras grandes de tamaño fijo.

b) Estimación consistente. Dado un cierto estadístico, $\overline{X}$, se puede afirmar que si él hubiese sido obtenido a partir del total de las observaciones de la población (n=N), su valor coincidiría con el del parámetro Debe tenerse en cuenta que en este criterio el tamaño de la muestra al aumentar hace que $\overline{X} \longrightarrow \mu$.

c) Estimación eficiente. Para que un estadístico (p) sea considerado un estimador eficiente del parámetro (P), debe ir acompañado por el menor error posible, es decir, en este caso del menor valor de $\sqrt{\frac{pq}{n}}$.

1.2. Estimación por invervalo de confianza. Se llama de este modo a aquella estimación de un parámetro, dada por dos números entre los cuales el parámetro debe caer, con cierta probabilidad de error. En la estimación puntual se trataba de determinar un solo valor para estimar el parámetro respectivo, siendo obvio, que es bastante improbable que ese valor coincida con el verdadero valor del parámetro. Como se pudo apreciar, no se hizo ninguna afirmación acerca de la probabilidad de error de la estimación.

La ventaja que presenta la estimación por intervalo de confianza, deriva del hecho de que esta forma de estimación permite indicar la precisión del procedimiento al acompañar al tamaño de la estimación, la probabilidad de error o el grado de incertidumbre de ella. La amplitud del intervalo indica con cuanta precisión ha sido estimado el parámetro. Si el intervalo de confianza es pequeño, se puede decir que la estimación es muy precisa, por el contrario si el intervalo de estimación es amplio, la estimación del parámetro correspondiente es poco precisa, para una misma probabilidad de error.

Sean $\mu_p = P$ y σ_p la media y el error estándar de la distribución de muestreo del estadístico p. Dado que todas las muestras son grandes y de tamaño fijo, la distribución de muestreo de p es normal aproximadamente. De ahí que se puede esperar que se encuentre un estadístico p cualquiera situado entre

i) $\mu_p \pm \sigma_p$ el 68,27% de las veces,

ii) $\mu_p \pm 2\sigma_p$ el 95,45% de las veces,

iii) $\mu_p \pm 3\sigma_p$ el 99,73% de las veces

Del mismo modo se puede esperar que el parámetro $P = \mu_p$ esté incluido.

a) el 68,27% de las veces entre $p \pm \sigma_p$,

b) el 95,45% de las veces entre $p \pm 2\sigma_p$, y

c) el 99,73% de las veces entre $p \pm 3\sigma_p$

Por esta razón esos intervalos son llamados de intervalos de confianza al 68,27%, al 95,45%, al 99,73% para la evaluación de $P=\mu_p$. Los valores extremos de esos intervalos son llamados de límites de confianza al 68,27%, al 95,45%, al 99/73%.

Colocado así el problema

$$p \pm 1{,}96\,\sigma_p$$

son los límites de confianza al 95% para P. El porcentaje de confianza se llama nivel de confianza. Valores numéricos como 1,96 son valores críticos de Z (abscisa de la curva normal) que algunos autores llaman también coeficientes de confianza.

Intervalo de confianza de tasas

Existe una población finita, dicotómica (binomial), como la población de vacas frente a la TBC (positivos y negativos), en la cual la tasa P es positivos (parámetro) es desconocida.

Si en una muestra grande ($n \geq 30$), de tamaño fijo, seleccionada (al azar) a partir de la población citada, la tasa de bovinos positivos a la TBC es p, entonces los límites de confianza de P son dados por

$$p \pm Z\sigma_p$$

Dado que el valor de σ_p se desconoce también, para obtener el límite de confianza antes mencionado, se emplea el valor correspondiente a la muestra

$$S_p = \sqrt{\frac{pq}{n}}$$

que es satisfactório siempre que $n \geq 30$. Cuando $n < 30$, esta aproximación deja de ser satisfactoria y debe utilizarse un procedimiento exacto basado en la distribución binomial.

De una población de vacas lecheras se obtiene una muestra de 1.000 vacas que se someten a la prueba de la tuberculina en la tabla del cuello. Resultan positivas 80 vacas. Los límites de confianza de la tasa población P, al 99%, son dados por

$$\frac{80}{1.000} \pm 2,58\sqrt{\frac{0.08 \times 0.92}{1.000}}$$

$$0,08 \pm 2,58\ (0,01$$

$$0,08 \pm 0,02$$

o sea

$$8\% \pm 2\%$$

Intervalo de confianza de medias aritméticas.

Existe una población finita, de valores de niveles de anticuerpos de bovinos 60 días después de vacunados contra la fiebre aftosa, en la cual la media aritmética μ es desconocida.

A partir de esta población se extrae mediante un procedimiento aleatorio, una muestra grande de bovinos ($n \geq 30$) que son sangrados y en cuyos sueros se mide el titulo de anticuerpos neutralizantes contra la fiebre aftosa. Ahora tenemos una media aritmética muestral $\bar{X}$. Entonces los límites de confianza de μ son dados por

$$\bar{X} \pm Z\sigma_{\bar{X}}$$

donde

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Dado que casi siempre se desconoce el valor poblacional σ, para determinar los límites de confianza de μ, se emplea el valor muestral correspondiente S, cuando $n \geq 30$, esta aproximación resulta poco satisfactoria por lo cual se debe adoptar un método de muestras pequeñas.

Por razones expuestas, el error estándar es

$$S_{\bar{X}} = \frac{S}{\sqrt{n}}$$

con lo cual, los límites de confianza para la media poblacional μ, son dados por

$$\bar{X} \pm ZS_{\bar{X}}$$

Desde una población de bovinos mayores de 3 años se ha obtenido una muestra de n = 225 bovinos, 60 días después de vacunados contra la fiebre aftosa, se han sangrado y determinado el titulo de los anticuerpos contra esta enfermedad mediante la técnica de seroneutralización en tubos (cultivo celular). Los valores muestrales son los siguientes:

$$\bar{X} = 2.5$$

$$S = 0.8$$

$$S_{\bar{X}} = \frac{0.8}{\sqrt{225}} = \frac{0.8}{15} = 0,053$$

El intervalo de confianza de μ al 95% es dado por

$$2.5 \pm (1,96)\ (0,05)$$

$$2.5 \pm 0,1$$

Prueba de hipótesis: Toma de decisiones.

Es común en salud animal tener que tomar decisiones acerca de la población animal que es asistida, basados en informaciones muestrales. Estas decisiones son llamadas estadísticas. Es común tener que decidir, sobre datos muestrales, si una vacuna es eficaz para proteger en buen nivel a la población bovina frente a la fiebre aftosa; si un procedimiento de diagnóstico es mejor que otro, etc.

Para tomar una decisión es necesario formular alguna hipótesis acerca de la población bajo estudio.

Se entiende por hipótesis estadística, un cierto supuesto que se refiere a la distribución de frecuencias de una variable aleatoria. Estos supuestos pueden ser acerca de que un dado parámetro asume un cierto valor; acerca de las frecuencias relativas, de una distribución; acerca de que algunas variables son independientes unas de otras, etc., los llamamos supuestos. Los llamados supuestos son valores hipotéticos que deben ser probados, en relación con los hechos observados Si lo observado es claramente inconsistente, con la hipótesis dada, esta última debe ser rechazada. Si lo observado no es inconsistente con la hipótesis planteada, ella debe ser aceptada. Lo expresado se sintetiza diciendo que, las hipótesis son "cotejadas" contra los datos que proporciona la naturaleza, con el objetivo de verificar un supuesto. Esta es la forma como se toma una decisión estadística acerca de la población, basándose en una información muestral.

Una hipótesis sobre un parámetro, es comparada con un valor que proviene de una muestra, por lo tanto el problema se circunscribe a determinar si una muestra que tiene un estadístico determinado podría haber sido obtenida a partir de una población, cuyo parámetro correspondiente es dado por la hipótesis.

Es conveniente recordar que en Matemáticas, el proceso mediante el cual se acepta o se rechaza un cierto supuesto, indica con absoluta certeza una de estas situaciones alternativas, es decir, si la hipótesis es verdadera o es falsa. Este procedimiento se conoce con el nombre de demostración matemática.

Ejemplo:

Hipótesis: $a \cdot b = 0 \Longrightarrow a = 0$ ó $b = 0$

Demostración: $X \cdot 0 = X(0 + 0) = X \cdot 0 + X \cdot 0$ adicionando $- X \cdot 0$

$$X \cdot 0 - \left(X \cdot 0\right) = X \cdot 0 + X \cdot 0 - \left(X \cdot 0\right)$$

$$0 = X \cdot 0$$

En cambio, en las ciencias observacionales no es posible tener certeza acerca de la veracidad o falsedad de una hipótesis planteada, sino que siempre prevalece cierto grado de incertidumbre en la decisión acerca de lo cierta o falsa que sea una hipótesis estadística. En estas ramas, el procedimiento mediante el cual se toma una decisión de este tipo, se basa en la información muestral. Este procedimiento se denomina prueba de hipótesis, siendo su principal propósito, el conocimiento de la discrepancia entre la información muestral observada (estadístico) y el valor hipotético del parámetro.

El planteamiento de una hipótesis estadística, implica necesariamente el aceptar o rechazar la hipótesis planteada, a través de un mecanismo que recibe el nombre de prueba de hipótesis, decisión estadística o prueba de significación.

En la mayoría de los casos, se formula una hipótesis con el propósito de rechazarla, por ejemplo: Cuando se necesita decidir si una ración alimenticia es mejor que otra, la hipótesis planteada es que no hay diferencia entre las dos raciones: cuando se quiere decidir, si la distribución de una variable es Normal o no, la hipótesis que se plantea para probar la bondad del ajuste, es que la distribución de la variable estudiada es de tipo Normal.

A través de estos ejemplos, se ve claramente, que las hipótesis planteadas suponen la no existencia de discrepancias. En general, a este tipo de hipótesis se les llama hipótesis nula (H_o). A su vez, existen hipótesis que afirman lo contrario, y a ellas se les llama hipótesis alternativas (H_1).

Teoría de la decisión estadística.

Supóngase que se tiene una variable aleatoria X, cuya población se distribuye de acuerdo a una cierta distribución de probabilidades. Se piensa que hay sólo dos distribuciones de probabilidades posibles: A y B. Se plantea la hipótesis nula (H_o) de que la población de X tiene una distribución A; se contrapone a ella, la hipótesis alternativa (H_1), que afirma que dicha población posee una distribución B.

$$H_o : A$$

$$H_1 : B$$

Las proposiciones o hipótesis que se formulan son las mencionadas anteriormente: H_o y H_1 (no H_o). En lógica matemática, los valores de verdad que permiten juzgar una hipótesis, son dos: Verdadero (V) y Falso (F). Con estos elementos se puede construir la siguiente tabla:

H_o	H_1
V	F
F	V

Tomando en cuenta que en Estadística no existe certeza absoluta, el esquema dado permite decir que si H_o es verdadera la población de X debe tener una distribución A, por lo que H_1 se considera falsa. Por otra parte, si H_o es falsa, la población de X tendrá una distribución no A, o sea, una distribución B, ya que H_1 es verdadera. Sin embargo, debe recordarse que en Estadística no se pueden hacer afirmaciones absolutamente ciertas.

Como se ha dicho, una vez establecida la hipótesis nula se hace necesario determinar un procedimiento que permita tomar una decisión, con el objeto de aceptar o rechazar la hipótesis planteada, con cierto grado de incertidumbre.

La estrategia utilizada para llegar a decidir este problema, es la de lograr una información muestral que permita hacerlo, ya que ella proporcionará un valor de X que se puede ubicar en el eje de las abscisas.

Para aplicar algunos conceptos que forman parte de la teoría de decisiones estadísticas se hace necesario recurrir a un ejemplo. Supongamos que la tasa de prevalencia por brucelosis en bovinos, en una región determinada alcanzaba a un 12% según opiniones de veterinarios y ganaderos conocedores del problema. Para llevar adelante un programa de control de la brucelosis se ha hecho un muestreo (n=1.000 vacas) obteniéndose una tasa de 8%. Puede aceptarse la diferencia 12% - 8% = 4% como variación por muestreo?

Para tomar una decisión acerca de esta cuestión tenemos que desarrollar un proceso (secuencia de pasos) que nos conduzca a tal finalidad.

a) formulación de hipótesis

Toda decisión estadística se plantea en términos de una disyuntiva, a lo menos, entre dos hipótesis referente a la o las poblaciones en estudio. Como ya hemos visto anteriormente en situaciones como éstas se plantean dos hipótesis.

- hipótesis nula (H_o): la diferencia observada entre la tasa muestral y la poblacional es consecuencia del error de muestreo, es decir que la tasa de prevalencia poblacional por brucelosis es 12%.

- hipótesis alternativa (H_1): la tasa poblacional es diferente de 12. Es decir, la diferencia observada es efectiva y en consecuencia, la población no tiene una tasa tan alta como la manifestada por los opinantes.

b) verificación de la hipótesis

La verificación de una hipótesis estadística se hace sobre una base observacional (datos), pero teniendo como referencia ciertos elementos de la teoría estadística (distribuciones de probabilidades)

b.1) Distribución teórica de muestreo del estadístico.
El estadístico muestral es la tasa de prevalencia p.
Para muestras grandes el estadístico p tiene una distribución teórica de muestreo simétrica en forma de campana semejante a la curva normal, bajo el supuesto de que H_o es verdadera. La media de esta distribución teórica de muestreo de p es $\mu_p = \bar{p} = P$ y la media de variación es S_p (error estándar).

Dado que, el propósito es someter a prueba si el alejamiento de p del valor poblacional supuesto P, es "relevante", y además que el tamaño de la muestra es grande, se elige el estadístico Z de la curva normal

$$Z = \frac{p - P}{S_p}$$

donde:

p = tasa prevalencia muestral

P = tasa prevalencia poblacional

S_p = error estándar de la tasa

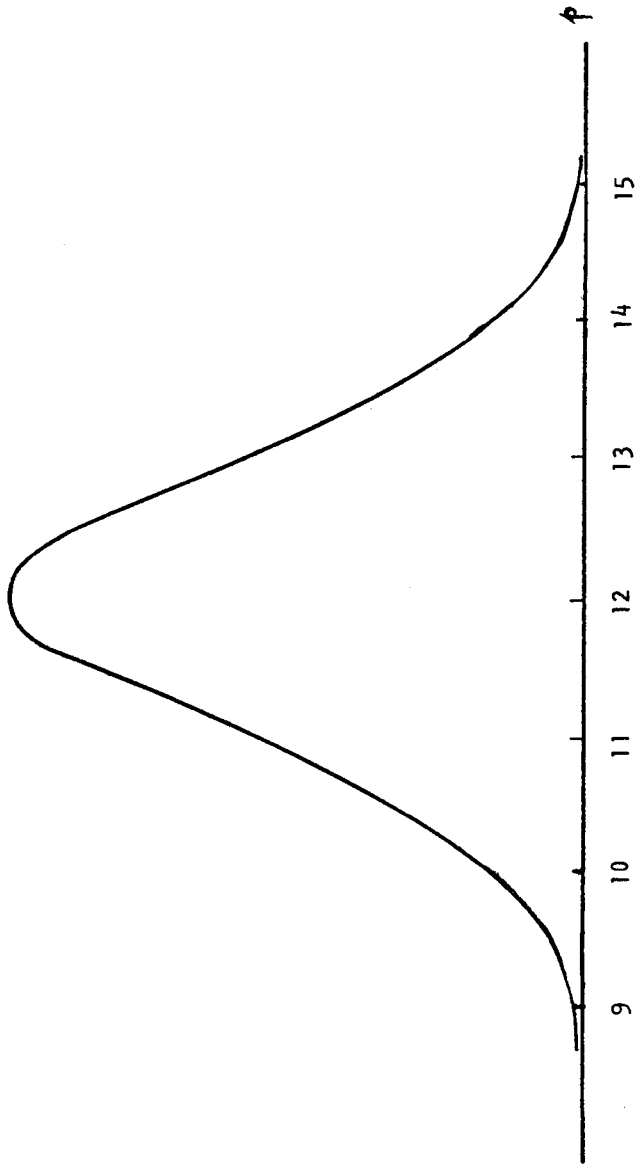
De ahí que estandarizando los valores de p a través de Z, tenemos una distribución de probabilidades normal (frecuencias relativas esperadas) para los diversos valores que puede asumir el estadístico Z.

La distribución teórica de muestreo de un estadístico, en algunos casos depende del tamaño de la muestra. Además la teoría supone que los individuos incluidos en las muestras serían elegidos en forma aleatoria.

Generalmente, las probabilidades asociadas a diversos valores posibles de un estadístico, en este caso Z, están ya tabuladas.

La distribución teórica de muestreo, bajo el supuesto de que H_o es verdadera, que corresponde a este caso a la distribución normal de probabilidades, presenta las siguientes propiedades:

- el 68.26% de las veces P quedará incluida en el intervalo $p \pm Sp$
- el 95.45% de las veces P quedará incluida en el intervalo $p \pm 2Sp$
- el 99.73% de las veces P quedara incluida en el intervalo $p \pm 3Sp$


p
9
10
11
12
13
14
15

b.2) Definición del nivel de significación.

Es definido arbitrariamente por los encargados del estudio.
Al someter a prueba una hipótesis planteada, la probabilidad máxima asociada a correr el riesgo de cometer un error de tipo α o I es llamada nivel de significación de la prueba.

Con frecuencia se utilizan niveles de significación 5% y 1%. Si se elige un α = 5% para verificar una hipótesis, quiere decir que hay 5 probabilidades en 100 de rechazar la hipótesis, cuando debería haber sido aceptada. O sea, existe una confianza de 95% de tomar una decisión correcta. La probabilidad de error en la decisión es del 5%.
En este caso hemos elegido arbitrariamente un α (probabilidad de cometer un error de tipo I) = 5%.

b.3) Definición de la región de rechazo de la hipótesis nula.

La región de rechazo es una región de la distribución teórica de muestreo del estadístico. La distribución de muestreo incluye todos los posibles valores que el estadístico elegido puede tomar bajo la hipótesis nula. La región de rechazo es un subconjunto de todos esos posibles valores, el cual es tan extremo que cuando la hipótesis nula es verdadera, la probabilidad es muy pequeña de que la muestra seleccionada produzca un valor del estadístico que caiga entre los valores extremos que forman esa región. La probabilidad asociada con cualquier valor en la región de rechazo es igual o menor que α.

En este caso la región de rechazo de H_o expresada en términos estandarizados de la curva normal es:

$$Z \leq -1.96 \quad y \quad Z \geq 1.96$$

por lo tanto la región de aceptación de H_o está formada por:

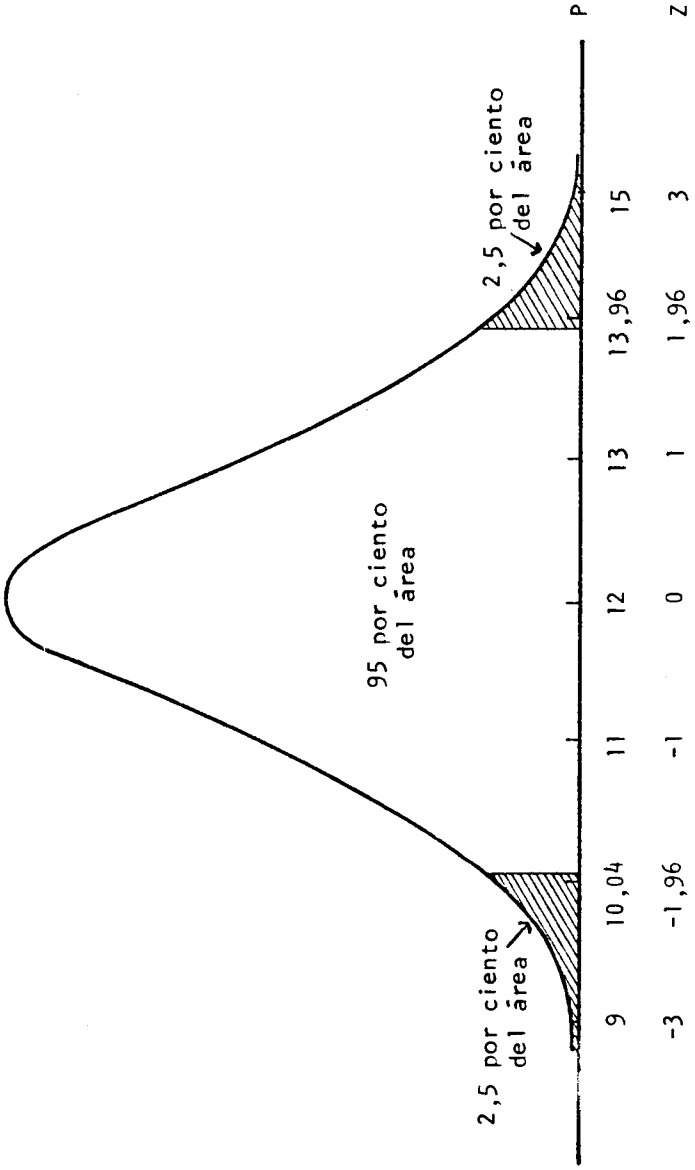
$$-1.96 < Z < 1.96$$

Ahora en términos de prevalencia de brucelosis, la región de rechazo de H_o es

$$p \leq -1.96\ Sp \quad y \quad p \geq 1.96Sp \quad \text{o sea} \quad p \leq 10\% \quad y\ p \geq 14\%$$

por lo tanto la región de aceptación de H_o es

$$10\% < p < 14\%$$


2,5 por ciento
del área
95 por ciento
del área
2,5 por ciento
del área
P
9
10,04
11
12
13
13,96
15
Z
-3
-1,96
-1
0
1
1,96
3

b.4) Cálculo del valor del estadístico a partir de las observaciones. Con la finalidad de someter a prueba de hipótesis planteada (H_o) se hacen observaciones sea a través de encuestas o de experimentación. En el campo de las ciencias naturales y sociales, la verificación de las hipótesis se hace sobre bases observacionales (hechos). A partir de los datos obtenidos se calculan los estadísticos correspondientes. En este caso se ha seleccionado a partir de la población de vacas de la región, una muestra de n = 1.000 vacas. De estos, 90 vacas resultaron positivas a brucelosis, por tanto

tasa prevalencia: $p = \frac{80}{1.000} \times 100 = 8\%$ $\quad q = 1-p = 92\%$

error estándar: $S_p = \frac{8 \times 92}{1.000} = 0,8\%$

$$Z = \frac{8 - 12}{0,8} = \frac{-4}{0,8} = -5$$

c) Toma de una decisión estadística

En este tipo de proceso la decisión que se toma no es con certeza, sino que hay un cierto margen de incertidumbre al hacerlo (probabilidad), debido a que la toma de decisión está expuesta a error.

Si el estadístico observado cae en la región de rechazo, se decide rechazar la hipótesis nula. La razón de este procedimiento radica en el hecho de que si la probabilidad asociada con la ocurrencia bajo la hipótesis nula de un valor cualquiera es pequeña, tenemos dos caminos para explicar la ocurrencia de ese valor del estadístico. El primero camino a tomar será considerar la H_o como falsa. El otro camino sería considerar ese valor como raro e infortunado y por ende muy poco probable de ocurrir bajo el supuesto de H_o ser verdadera. En el proceso de decisión estadística, lo habitual es elegir el primer camino. Ocasionalmente el segundo camino puede ser correcto, ya que la probabilidad de tal evento es α, es decir de rechazar H_o cuando de becho ella fuese verdadera (error tipo I o tipo α) Los límites de este tipo de error han sido ya fijados en 5%, en el punto anterior.

El criterio para rechazar la H_o, como se ha indicado ya debe ser establecido previamente al examen de los datos y en ningún caso debe subordinarse a los hallazgos del estudio.

Podría pensarse que el procedimiento más seguro es reducir a un mínimo este tipo de error (I). Sin embargo, esta posición produciría un aumento de la probabilidad de cometer un error de tipo II (β), que es el error de aceptar una hipótesis nula siendo que ella fuese falsa. El esquema de las situaciones posibles es el siguiente:

Decisión sobre H_o	Realidad sobre H_o	
	Verdadera	Falsa
Aceptación	Decisión correcta	Error II (tipo β)
Rechazo	Error I (tipo α)	Decisión correcta

Para un tamaño fijo de muestra (n), elegido α, queda determinado β. Una disminución de α, en estas condiciones hace aumentar β. Si se desea controlar ambos α y β, se debe aumentar n.

En resumen se decide que H_o es falsa, cuando la probabilidad asociada al valor del estadístico observado es igual o menor que el valor de α ya establecido. En tal caso se dice que el valor observado del estadístico es significativo.

Pruebas de hipótesis envolviendo medias

1. Una media: prueba de hipótesis que en una población el parámetro μ tiene un cierto valor.

Queremos tomar una decisión sobre el supuesto de que la media de duración (en días) de los episódios sobre fiebre aftosa en los rebaños bovinos en una cierta región, considerando los últimos años con buenos registros, es de 25 días. Se ha programado tomar una muestra al azar de 196 episódios, desde el archivo de tarjetas cada una de las cuales se refiere a un episódio de la enfermedad en un rebaño bovino

a) Planteamiento de la hipótesis

- Hipótesis nula (H_o): $\mu = 25$
- Hipótesis alternativa (H_1): $\mu \neq 25$

b) Distribución teórica de muestreo del estadístico correspondiente.

El estadístico muestral es la media aritmetica ($\bar{x}$). Si hipotéticamente se obtuvieran desde esa población muchas muestras todas del mismo tamaño (n=196) el conjunto de medias aritméticas por estas muestras generado tomaria una distribución de frecuencias simétrica, semejante a la curva normal, con media $\mu_{\bar{x}} = \mu = 25$ días y una media de variación, $S_{\bar{x}}$.

Por esta razón el modelo de la teoría estadística que aplicamos para someter a prueba la H_o, es la curva normal de probabilidades, en la cual el estadístico se transforma en

$$Z = \frac{\bar{x} - \mu}{S_{\bar{x}}}$$

Además se aplican a la toma de decisión, las propiedades de la curva normal.

c) Definición del nivel de significación

$$\alpha = 5\%$$

d) Establecer la región de rechazo H_o

De acuerdo con el nivel de significación indicado la región de rechazo de H_o es:

$$Z \leq -1,96 \quad y \quad Z \geq 1,96$$

la región de aceptación es:

$$-1,96 < Z < 1,96$$

e) Calcular el valor del estadístico correspondiente a partir de los valores observados en la muestra. En este caso

Días	F	X	FX	X^2	FX^2
6 - 10	6	8	48	64	384
11 - 15	40	13	520	169	6.760
16 - 20	81	18	1.458	324	26.244
21 - 25	59	23	1.357	529	31.211
26 - 30	10	28	280	784	7.840
Total	196		3.663		72.439

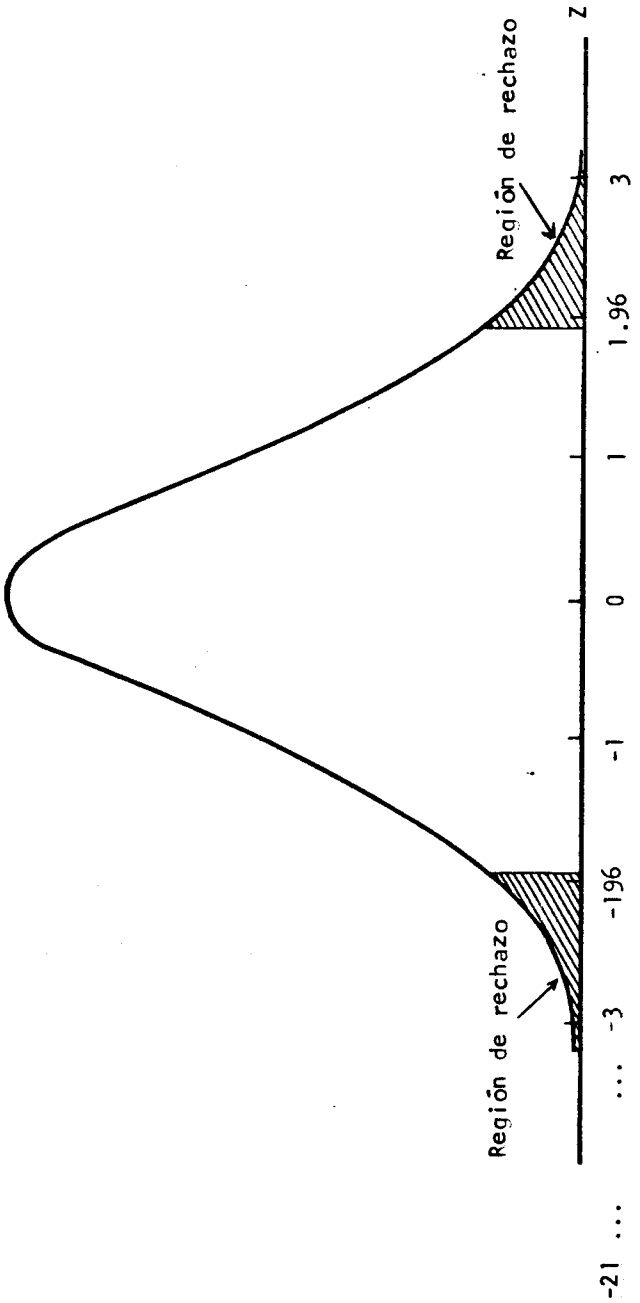
$$\bar{x} = \frac{3.663}{196} = 18,7$$

$$s^2 = \frac{72.439 - (3.663)^2/196}{196} = 20,3$$

$$s = 4,5$$

$$s_{\bar{x}} = \frac{4,5}{\sqrt{196}} = 0,3$$

$$Z = \frac{18,7 - 25}{0,3} = \frac{-6,3}{0,3} = -21$$


Región de rechazo
Región de rechazo
Z
3
1.96
1
0
-1
-196
-3
... ...
-21 ...

f) Toma de decisión

El valor observado del estadístico Z = -21 cae en la región de rechazo de H_o ya que -21 < -1,96. En el supuesto que H_o fuese verdadera es un valor muy poco probable, razón por la cual se toma la decisión de rechazar H_o, aceptando por tanto H_1. El riesgo de rechazarla siendo H_o verdadera es muy bajo, prácticamente despreciable.

En resumen, la muestra de 196 episódios de fiebre aftosa no proviene de la población con media de duración de los episódios de 25 días.

2. Dos medias: prueba de la hipótesis que la diferencia entre μ_1 y μ_2 es igual a cero.

Supongamos que se quiera tomar una decisión sobre el supuesto de que en un lugar A en un año B la edad media (días) al primer parto en vaquillas nacidas en épocas de "seca" (μ_1) es igual que la media correspondiente a vaquillas nacidas en épocas de "lluvia" (μ_2).

Para tomar esta decisión es necesario que se tenga una muestra de vaquillas en primer parto para cada una de las 2 épocas.

a) Planteamiento de la hipótesis

$H_0 : \mu_1 = \mu_2$ ó $\mu_1 - \mu_2 = 0$

$H_1 : \mu_1 \neq \mu_2$ ó $\mu_1 - \mu_2 \neq 0$

b) Distribución teórica de muestreo del estadístico correspondiente.

El estadístico muestral es la diferencia entre dos medias aritméticas $(\bar{x}_1 - \bar{x}_2)$. De acuerdo con la hipótesis nula de que no hay diferencias en la edad al primer parto si construyéramos una distribución de frecuencias de diferencias entre medias aritméticas $(\bar{x}_1 - \bar{x}_2)$, siendo las muestras del mismo tamaño, ella tendría una curva cimétrica semejante a la curva normal con media aritmética $\mu_{\bar{x}_1 - \bar{x}_2} = 0$ y una medida de variación estimada por $\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}$

Por esta razón el modelo de la teoría estadística que aplicamos para someter a prueba H_0 es la distribución normal de probabilidades en la cual el estadístico estandarizado es

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{S_{\bar{x}_1 - \bar{x}_2}}$$

$$= \frac{x_1 - x_2}{S_{\bar{x}_1 - \bar{x}_2}}$$

De esta manera se aplican a la toma de decisión, las propiedades de la curva normal.

c) Definición del nivel de significación

$$\alpha = 5\%$$

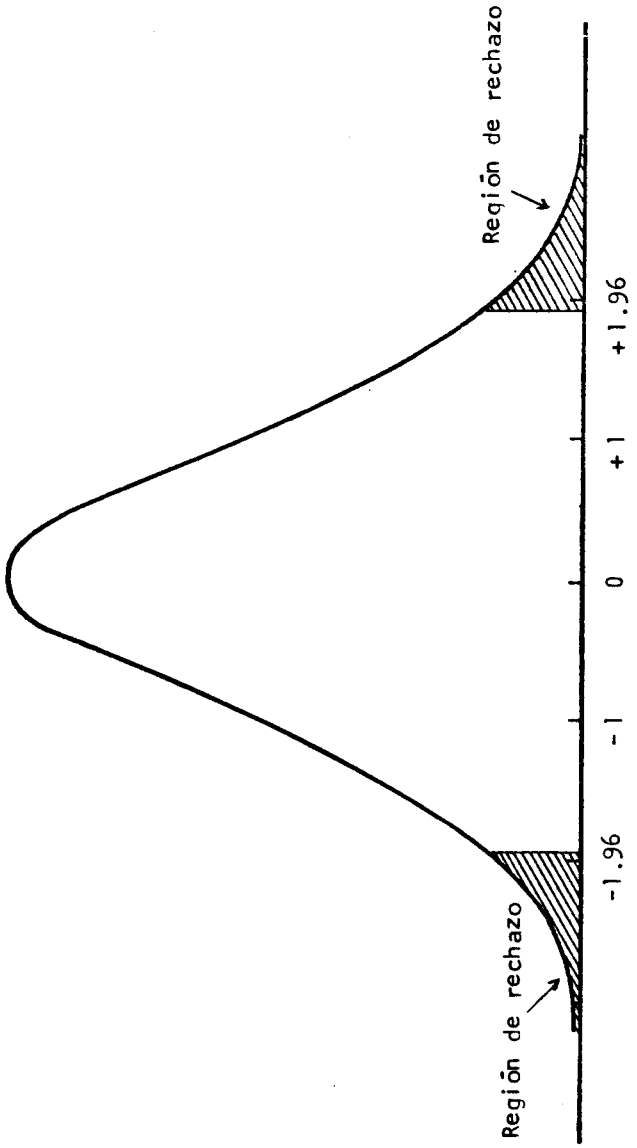
d) Establecer la región de rechazo de H_o de acuerdo con el nivel de significación establecido la región de rechazo de H_o es:

$$Z \leqslant -1.96 \quad y \quad Z \geqslant 1.96$$

y la región de aceptación es:

$$-1.96 < Z > 1.96$$

e) Calcular el valor del estadístico correspondiente a partir de los valores observados en la muestra


Región de rechazo
+1.96
+1
0
-1
-1.96
Región de rechazo

EPOCAS			
SECA		LLUVIOSA	
1.190	1.187	1.330	1.210
1.135	1.160	1.335	1.200
1.230	1.109	1.246	1.195
1.195	1.190	1.048	1.213
1.205	1.199	1.285	1.233
1.190	1.240	1.181	1.201
1.201	1.180	1.218	1.160
1.195	1.200	1.233	1.260
1.221	1.139	1.269	1.020
1.159	1.198	1.290	1.217
1.238	1.160	1.123	1.295
1.204	1.260	1.156	1.321
1.248	1.190	1.237	1.157
1.234	1.300	1.203	1.273
1.166	1.200	1.240	1.226
1.300	1.230	1.128	1.266
1.196	1.240	1.144	1.172
$n = 34$		$n = 34$	
$\Sigma x = 40.889$		$\Sigma x = 41.285$	
$\bar{x} = 1.202,62$		$\bar{x} = 1.214,26$	
$\Sigma (x-\bar{x})^2 = 57.058,03$		$\Sigma (x-\bar{x})^2 = 177.129,62$	

$$\bar{x}_1 - \bar{x}_2 = 11.64$$

$$S_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{2\hat{S}^2}{n}}$$

$$= \sqrt{\frac{(2)(3.458,30)}{34}}$$

$$= 14,45$$

$$\hat{S}^2 = \frac{\Sigma (x - \bar{x}_1)^2 + \Sigma (x - \bar{x}_2)^2}{n_1 + n_2 - 2}$$

$$= \frac{234.187,65}{66}$$

$$= 3.548,30$$

$$Z = \frac{11.64}{14.45} = 0.81$$

f) Toma de decisión

El valor observado de Z = 0,81 cae en la región de aceptación de la H_0 ya que la diferencia observada entre las medias muestrales es producto de la variación de muestreo y consideramos que ambas muestras provienen de una misma población.

La decisión es aceptar la H_0. Esto significa que la época de nacimientos (seca y lluviosa) de las vaquillas no influye sobre la media aritmética de edad al primer parto.

Pruebas de hipótesis envolviendo tasas

1. Una tasa: prueba de la hipótesis que en una población el parámetro P tiene un cierto valor.

Un investigador propone un nuevo método indirecto de diagnóstico para una enfermedad E, afirmando que es efectivo en más de 90% (valor considerado aceptable por los expertos) para identificar individuos portadores de la enfermedad E. Las autoridades médicas, con la esperanza de verificar lo expuesto, nombran una comisión de cientificos para evaluar la eficiencia del método de diagnóstico propuesto. Algunos miembros de la comisión consideran que lo dicho por el investigador es confiable, otros miembros de la comisión manifiestan algunas dudas, por lo cual se resuelve que es necesario verificar lo planteado por el investigador considerando como hipótesis el planteamiento hecho.

a) Planteamiento de la hipótesis

- Hipótesis nula (H_0): $P = 90\%$
- Hipótesis alternativa (H_1): $P \neq 90\%$

b) Distribución teórica de muestreo del estadístico correspondiente

El estadístico muestral es p, la tasa de posicitivos en la muestra.

Por hipótesis, si, de una población de individuos afectados por la enfermedad E se seleccionasen muchas muestras, con el mismo tamaño, siendo $n > 30$ para cada una de esas muestras se calcularía la tasa de positivos p, através del nuevo método. Con el conjunto resultante de valores de p (tantos como muestras

se hayan tomado) se confeccionaría una distribución de frecuencias de tasas, que tomaría una forma acampanada, simétrica, semejante a la curva normal con una tasa media $\mu_p = P = 90\%$ y una medida de variación (error estándar de una tasa) σ_p, de acuerdo con la hipótesis nula.

Por esta razón

$$Z = \frac{p - P}{\sigma_p}$$

es una variable estandarizada a la curva normal bajo el supuesto que H_0 es verdadera.

Entonces, se aplican a esta toma de decisión las propiedades de la curva normal.

c) Definición del nivel de significación

$$\alpha = 5\%$$

d) Establecer la región de rechazo de H_0

Considerando el nivel de significación elegido, la región de rechazo de H_0 es:

$$Z \leq -1.96 \quad y \quad Z \geq 1.96$$

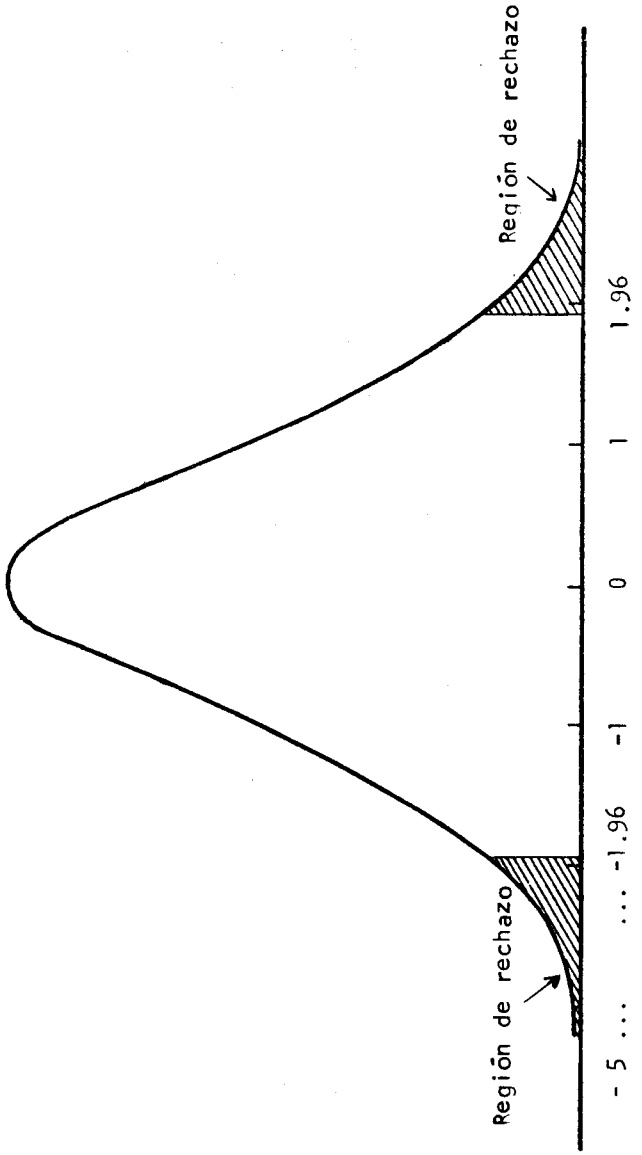
e) Calcular el valor del estadístico correspondiente a partir de observaciones muestrales.

Se seleccionó al azar desde la población de enfermos de la enfermedad E, una muestra de 200 individuos que fueron sometidos al diagnóstico a través del nuevo método propuesto. De éstos 150 resultaron positivos, entonces

$$p = \frac{150}{200} = 75\%$$

$$\sigma_p = \sqrt{\frac{(75)(25)}{200}} = \sqrt{9,38} \cong 3\%$$

$$Z = \frac{75 - 90}{3} = \frac{-15}{3} = -5$$


Región de rechazo
Región de rechazo
- 5 ...
... -1.96
-1
0
1
1.96

g) Toma de decisión

El valor observado del estadístico es Z = -5 el cual es menor que -1.96 y por tanto cae en la región de rechazo de H_0 . En el supuesto que H_0 fuese verdadera este valor es muy improbable, razón por la cual se decide rechazar H_0, aceptando H_1. El riesgo de estar tomando una decisión errada (rechazar H_0 si ella fuese verdadera) es prácticamente despreciable. Esto quiere decir que el método indirecto propuesto para el diagnóstico de la enfermedad E no alcanza el nivel mínimo de eficiencia.

2. Dos tasas: prueba de la hipótesis que la diferencia entre dos tasas, P_1 y P_2, es cero.

En una región de cría de suinos, existen dos tipos de manejo de las camadas de lechones.

En uno de los tipos de establecimientos las porquerizas de cría no tienen ningún tipo de protección de los lechones para no ser aplastados por la madre.

En el otro tipo de establecimientos, las porquerizas han sido dotadas de unas guardas de protección metálica en las junturas formadas por el piso y las paredes.

Se quiere tomar una decisión sobre el supuesto de que no existen diferencias de mortalidad le lechones por aplastamiento materno, en ambos tipos de explotación suina.

Para resolver este problema, se decide someter a prueba la hipótesis estadística planteada

a) Formulación de la hipótesis

$H_0 : P_1 = P_2$ ó $P_1 - P_2 = 0$

$H_1 : P_1 \neq P_2$ ó $P_1 - P_2 \neq 0$

b) Distribución teórica de muestreo del estadístico correspondiente

El estadístico muestral es la diferencia entre dos tasas muestrales $(p_1 - p_2)$. De acuerdo con la hipótesis nula, de que no hay diferencias en la mortalidad de lechones por aplastamiento materno entre ambos tipos de explotaciones, si se construyese una distribución de frecuencias de diferencias entre tasas muestrales $(p_1 - p_2)$, siendo todas las muestras grandes y siempre del mismo tamaño, esta distribución de frecuencais tendría una forma semejante a la curva normal con media aritmética $\mu_{p_1 - p_2} = 0$ y una medida de variación $\sigma_{p_1 - p_2}$.

Al estandarizar a la curva normal la variable de diferencias de tasa muestrales (p_1-p_2) queda convertida en el estadístico

$$Z = \frac{(p_1 - p_2) - 0}{\sigma_{p_1 - p_2}}$$

bajo el supuesto de que H_0 es verdadera.

c) Definición del nivel de significación

$$\alpha = 1\%$$

d) Establecer la región de rechazo de H_0

De acuerdo con el nivel de significación elegido, la región de rechazo de H_0 es

$$Z \leq -2.58 \qquad y \qquad Z \geq 2.58$$

e) Calcular el valor del estadístico correspondiente a partir de observaciones muestrales.

Se eligen al azar dos muestras de 400 camadas, una a apartir del subuniverso de establecimientos con porquerizas, sin protección y otra muestra del subuniverso de planteles con porquerizas con protección específica para evitar el aplastamiento de lechones.

Para el sistema de cría sin protección, considerando la media de 8 lechones nacidos vivos, la muestra incluye 3.200 lechones, de los cuales murieron por aplastamiento materno 352.

Para el sistema de cría con protección metálica, teniendo en cuenta una media de 9 lechones nacidos vivos, la muestra incluye 3.600 lechones, de los cuales murieron por aplastamiento 288.

$$n_1 = 3.200 \qquad n_2 = 3.600$$

$$p_1 = \frac{352}{3.200} \times 100 = 11\% \qquad p_2 = \frac{288}{3.600} \times 100 = 8\%$$

El error estándar de la diferencia de tasas es:

$$\sigma_{p_1 - p_2} = \sqrt{\frac{PQ}{n_1} + \frac{PQ}{n_2}} \qquad Q = 1 - P$$

pero las tasas poblacionales no las conocemos, entonces tomamos la media ponderada de las dos tasas muestrales como la estimación de P

Si

r_1 = número de lechones muertos por aplastamiento en el grupo 1

r_2 = número de lechones muertos por aplastamiento en el grupo 2

$$\hat{P} = \bar{p} = \frac{r_1 + r_2}{n_1 + n_2} \times 100$$

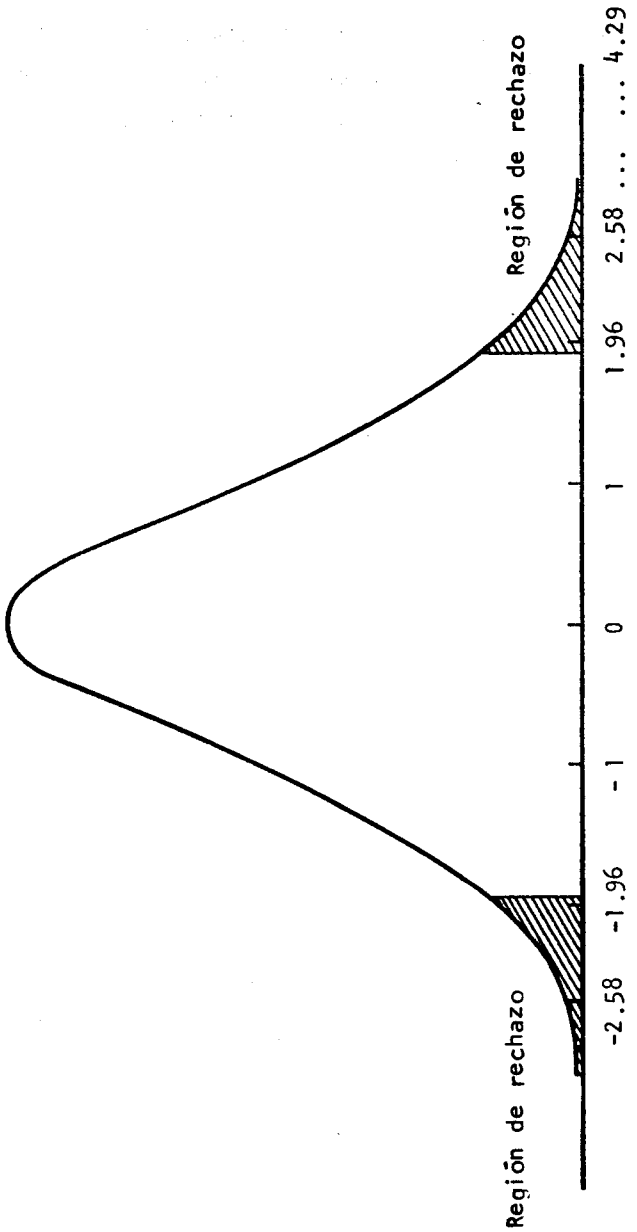
$$\hat{\bar{P}} = \bar{p} = \frac{352 + 288}{3.200 + 3.600} \times 100 = 9\%$$

$$\sigma_{p_1 - p_2} = \sqrt{\frac{(9)(91)}{3.200} + \frac{(9)(91)}{3.600}} = \sqrt{0,26 + 0,23}$$

$$= \sqrt{0,48} = 0,7$$

Ahora resolvemos Z

$$Z = \frac{11 - 8}{0,7} = 4,29$$


Región de rechazo
Región de rechazo
-2.58
-1.96
-1
0
1
1.96
2.58 ... ... 4.29

f) Zona de decisión

El valor del estadístico Z = 4,29 el cual es mayor que 2,58 y por lo tanto cae en la región de rechazo de H_0.

En el supuesto de que H_0 fuera verdadera este es un valor extremamente improbable, hecho que nos lleva a tomar la decisión de rechazar H_0 y aceptar la hipótesis alternativa H_1. Esto quiere decir que las tasas de mortalidad bajo ambos sistemas presentan diferencias marcadas.

Métodos para muestras pequeñas

En los capítulos anteriores nos hemos referido a muestras grandes cuyo tamaño es $n > 30$.

Con tal tipo de muestras las distribuciones teóricas de muestreo de varios estadísticos se asemejan a la curva normal siendo tanto mayor la semejanza cuanto mayor sea n.

De lo anterior se deduce que esas propiedades no son compatibles con el uso de muestras pequeñas ($n < 30$) de manera que hay que adoptar otros métodos.

El estudio de las distribuciones teóricas de muestreo de estadísticos obtenidos a partir de muestas de este tipo corresponde a lo que se podría llamar teoría exacta de muestreo.

En este capítulo presentaremos dos distribuciones: la de "t" de Student y la de X^2 (ji cuadrado).

1. Distribución de "t" de Student.

Al utilizar muestras pequeñas la desviación estándar S (DE) no es una buena aproximación de la desviación estándar de la población σ. Para pequeñas muestras una buena estimación de σ sería

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

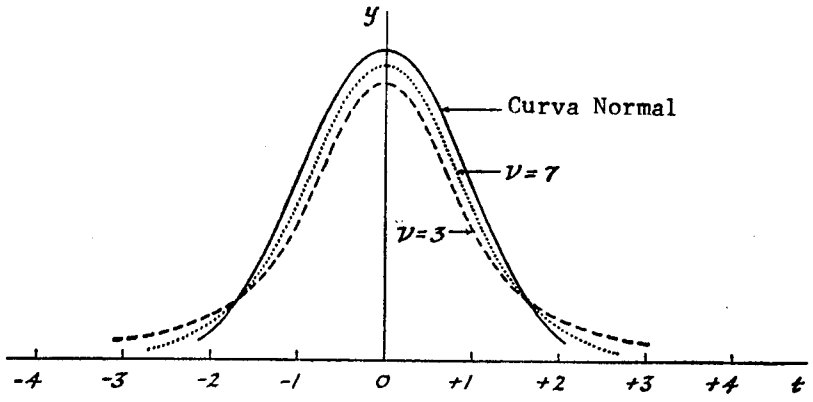
donde $n - 1$ son los grados de libertad (ν), que corresponden al número de observaciones independientes de la muestra. También se puede definir como el número total de observaciones de la muestra menos el número de parámetros a ser estimados a través de las observaciones muestrales.

El estadístico "t" se define

$$t = \frac{x - \mu}{S_{\bar{x}}}$$

donde $S_{\bar{x}} = \dfrac{\hat{S}}{\sqrt{n}}$ $\qquad$ $\hat{S} = \sqrt{\dfrac{\Sigma (x - \bar{x})^2}{n - 1}}$

de lo anterior se desprende que la distribución de probabilidades de "t" de student depende del grado de libertad (ν), dicho de otra manera habría tantas distribuciones de "t" de student cuantos diversos grados de libertad se consideren. Para grandes valores de grados de libertad (ν) (correspondientes a muestras de tamaño $n > 30$) las curvas de "t" sob muy semejantes a la curva normal cuyo estadístico conocemos como Z.



Distribuciones de "t" para

$\nu = 3$ y $\nu = 7$

Intervalo de confianza de una media.

Al igual que lo que se hace con las muestras grandes con la curva normal utilizando la tabla de las distribuciones de "t" se pueden establecer intervalos de confianza de una media poblacional al 95%, al 99% o a otros niveles de confianza.

Por límites de confianza para las medias poblacionales se pueden determinar por:

$$\mu = \bar{x} \pm \tau S_{\bar{x}}$$

donde el valor de t es el correspondiente al nivel de confianza utilizado y al número de grados de libertad de la muestra (ν). con estas condiciones debe ser leído el valor de t de la tabla.

Supongamos que hemos obtenido una muestra de tamaño n = 6 personas para estudiar la edad promedio de una población

x	$x - \bar{x}$	$(x - \bar{x})^2$
19	- 1	1
18	- 2	4
22	2	4
20	0	0
16	- 4	16
25	5	25
$\Sigma x = 120$	0	50

$$\bar{x} = \frac{120}{6} = 20 \qquad \hat{s} = \frac{50}{5}$$

$$= 10$$

$$= 3.16$$

$$S_{\bar{x}} = \frac{3.16}{6} = 1.29$$

Como $\nu = 5$ (grados de libertad) si deseamos determinar el intervalo de confianza de μ al 95% entonces

$$\mu = 20 \pm 2.57\ (1.29)$$
$$= 20 \pm 3.32$$
$$= 16.68 \text{ a } 23.32$$

El valor 2.57 fue leído en la tabla bajo la columna de $A = 0.05$ que corresponde a un 95% de confianza y $\nu = 5$.

Prueba de hipótesis de la diferencia entre dos medias.

Si se tienen dos muestras de una variable, obtenidas a partir de una población normalmente distribuida, la diferencia entre las dos medias muestrales se puede someter a prueba mediante la distribución de "t", con la siguiente expresión:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{S_{\bar{x}_1 - \bar{x}_2}}$$

donde $S_{\bar{x}_1 - \bar{x}_2} = S^2 \left(\frac{n_2 + n_1}{n_1 \cdot n_2}\right)$ y

$$S^2 = \frac{\Sigma (x_1 - \bar{x}_1)^2 + \Sigma (x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}$$

que satisface la distribución de "t" de Student con $n_1 + n_2 - 2$ grados de libertad

En este tipo de pruebas de hipótesis, el número de individuos que forman cada muestra puede ser igual o desigual.

En el numerador aparecen consignadas las diferencias entre dos medias muestrales y las respectivas medias poblacionales.

En el denominador de la expresión "t", se encuentra el error estándar de la diferencia, el cual se obtiene considerando una varianza común a ambas muestras (S^2), razón por la cual se combinan las sumas de cuadrados de desvíos y los grados de libertad respectivos. Todo esto descansa sobre el supuesto de que las varianzas de las dos muestras son iguales y por lo tanto tienen una varianza poblacional común σ^2, cuya mejor estimación es S^2.

Ejemplo: se tiene el peso del vellón (kg) de ovinos de seis dientes, de la raza Corriedale, de dos estancias vecinas en Tierra del Fuego.

Los pesos de los vellones sob los siguientes:

Peso del vellón (kg) de ovinos Corriedale de seis dientes. Tierra del Fuego.

Estancia 1	Estancia 2
x_1	x_2
4,3	3,9
3,8	4,3
3,9	3,8
4,0	4,0
4,4	4,1
4,5	4,0
3,9	3,6
4,6	3,9
4,2	
3,8	$\Sigma x_2 = 31,6$
$\Sigma x_1 = 41,4$	
$\Sigma x_1^2 = 172,20$	$\Sigma x_2^2 = 125,12$
$\bar{x}_1 = 4,14$	$\bar{x}_2 = 3,95$
$n_1 = 10$	$n_2 = 8$

Se desea saber si la diferencia observada entre las medias aritméticas muestrales de las dos estancias, para peso del vellón, es real o si ella puede atribuirse al azar.

El procedimiento a seguir para tomar una decisión, es el mismo usado en la sección anterior.

a) Planteamiento de hipótesis

Hipótesis nula (H_o): existe diferencia entre el promedio de la muestra de la estancia 1 y el promedio de la muestra de la estancia 2. Es decir, ambas muestras provienen de una misma población ($\mu_1 = \mu_2$). En otros términos, la diferencia entre $\bar{x}_1 = 4,14$ kg y $\bar{x}_2 = 3,95$ kg, ha sido obtenida a partir de una población de diferencias, con media igual a cero: $\mu_D = \mu_1 - \mu_2 = 0$. De manera que la diferencia que se observa, a simple vista, se debería a variaciones

de tipo aleatorio.

La hipótesis alternativa (H_1) por su parte, afirma que la diferencia que se observa entre $\bar{x}_1$ y $\bar{x}_2$ es significativa. En otras palabras, establece que la diferencia observada entre ambas medias aritméticas muestrales, se debe a efectos ajenos al azar. Esto indicaría que las dos muestras, en lo que respecta al promedio, no provienen de la misma población, por lo cual $\mu_1 \neq \mu_2$.

b) Elección del nivel de significación

Como es costumbre, se trabaja con una probabilidad de error de tipo α de 0,05.

c) Se utiliza la distribución de probabilidades de "t", con la siguiente expresión, por hipótesis:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S_{\bar{x}_1 - \bar{x}_2}}$$

d) Determinación de la zona de rechazo

En el ejemplo, los valores críticos de "t", que determinan la zona de rechazo de H_o son $t \leqslant -2,120$ y $t \geqslant 2,120$, valores que tienen una probabilidad de ocurrir menor que 0,05, para $n_1 + n_2 - 2$ grados de libertad.

e) Cálculo de "t"

$$\Sigma (x_1 - \bar{x}_1)^2 = 172,20 - \frac{1.713,96}{10} = 172,20 - 171,40 = 0,80$$

$$\Sigma (x_2 - \bar{x}_2)^2 = 125,12 - \frac{998,56}{8} = 125,12 - 124,82 = 0,30$$

Resumen de los datos. Peso del vellón.

Muestra	n	Grados de libertad	$\bar{x}$	$\Sigma (x_i - \bar{x})^2$
Estancia 1	10	9	$\bar{x}_1 = 4,14$	0,80
Estancia 2	8	7	$\bar{x}_2 = 3,95$	0,30
		16	$\bar{x}_1 - \bar{x}_2 = 0,19$	1,10

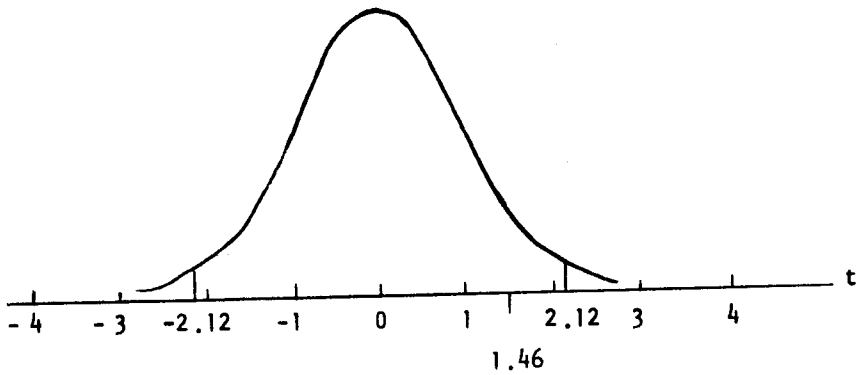
$$s^2 = \frac{1,10}{16} = 0,069$$

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{0,069\ (\frac{18}{80})} = \sqrt{(0,069)(0,225)} = \sqrt{0,016} = 0,127$$

$$t = \frac{4,14 - 3,95}{0,127} = \frac{0,19}{0,127} = 1,496$$

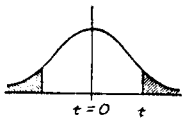
f) Decisión

El valor de "t" observado (1,496) cae en la zona de aceptación de H_o, puesto que la probabilidad de ocurrencia bajo H_o, asociada a él, está entre 0,10 y 0,20. Por esta razón se le considera un valor común, por lo cual se acepta la hipótesis nula. Es decir, la diferencai observada entre $\bar{x}_1 = 4,14$ y $\bar{x}_2$ " 3,95 se debe a variaciones de tipo aleatorio, no significativas. Por lo tanto, se puede afirmar que las muestras de peso del vellón, de las dos estancias, en lo relativo a su promedio, se consideran obtenidas a partir de una misma población $(\mu_1 = \mu_2)$.



Distribución "t" de Student

A - es la suma de las áreas de las dos colas para los valores de t dados a seguir.



ν - denota el número de grados de libertad.

ν o gl	A = 0.1	A = 0.05	A = 0.02	A = 0.01	A = 0.001
1	6.314	12.706	31.821	63.657	636.619
2	2.920	4.303	6.965	9.925	31.598
3	2.353	3.182	4.541	5.841	12.941
4	2.132	2.776	3.747	4.604	8.610
5	2.015	2.571	3.365	4.032	6.859
6	1.943	2.447	3.143	3.707	5.959
7	1.895	2.365	2.998	3.499	5.405
8	1.860	2.306	2.896	3.355	5.041
9	1.833	2.262	2.821	3.250	4.781
10	1.812	2.228	2.764	3.169	4.587
11	1.796	2.201	2.718	3.106	4.437
12	1.782	2.179	2.681	3.055	4.318
13	1.771	2.160	2.650	3.012	4.221
14	1.761	2.145	2.624	2.977	4.140
15	1.753	2.131	2.602	2.947	4.073
16	1.746	2.120	2.583	2.921	4.015
17	1.740	2.110	2.567	2.898	3.965
18	1.734	2.101	2.552	2.878	3.922
19	1.729	2.093	2.539	2.861	3.883
20	1.725	2.086	2.528	2.845	3.850
21	1.721	2.080	2.518	2.831	3.819
22	1.717	2.074	2.508	2.819	3.792
23	1.714	2.069	2.500	2.807	3.767
24	1.711	2.064	2.492	2.797	3.745
25	1.708	2.060	2.485	2.787	3.725
26	1.706	2.056	2.479	2.779	3.707
27	1.703	2.052	2.473	2.771	3.690
28	1.701	2.048	2.467	2.763	3.674
29	1.699	2.045	2.462	2.756	3.659
30	1.697	2.042	2.457	2.750	3.646
40	1.684	2.021	2.423	2.704	3.551
60	1.671	2.000	2.390	2.660	3.460
120	1.658	1.980	2.358	2.617	3.373
∞	1.645	1.960	2.326	2.576	3.291

PRUEBAS DE HIPOTESIS BASADAS EN LA DISTRIBUCION DE JI QUADRADO

Es frecuente que investigadores de distintas áreas como biología, genética, botánica y otras, muestran interés en conocer si los resultados de sus experimentaciones reflejan ciertas características en lo que se refiere a su distribución numérica. Por ejemplo, un genetista que trabaja con distintas cepas de animales de laboratorio, al realizar ciertos cruzamientos, podría esperar que los descendientes presentaran alguna proporción conocida. Por otra parte, otro investigador podría querer determinar si una característica, dentro de un grupo de individuos, presenta una distribución que él supone corresponde a la Normal. Situaciones similares a estas, existen en gran cantidad en distintos campos de la ciencia.

Otro problema, un tanto diferente al anterior, es aquel que suele presentarse cuando los individuos que componen la muestra o población estudiada, son clasificados de acuerdo a dos características cualitativas. Interesa, en este caso, saber si existe alguna relación entre éstas. Supóngase que un clínico estudie la presentación de cierta enfermedad infecciosa entre individuos de distintas edades, o bien que, un zootecnista analice la calidad de la leche en lecherías con diferente tipo de alimentación. En ambos casos, el profesional querrá saber si la variación de una de las características tendría relación con la variación de la otra.

Los métodos estadísticos que permiten resolver estas incógnitas están basados en la distribución de probabilidades de ji cuadrado. El estadígrafo ji cuadrado, simbolizado por χ^2, es una medida de la diferencia entre las frecuencias observadas (O_j) y las correspondientes frecuencias calculadas (C_j), teóricas o esperadas, definida por la siguiente expresión:

$$\chi^2 = \frac{(O_1-C_1)2}{C_1} + \frac{(O_2-C_2)2}{C_2} + \ldots\ldots + \frac{(O_k-C_k)2}{C_k}$$

$$\chi^2 = \sum_{j=1}^{k} \frac{(O_j - C_j)^2}{C_j}$$

donde :

O_j es la frecuencia observada para la $j^{ésima}$ clase.

C_j es la frecuencia calculada par la $j^{ésima}$ clase.

Clases : j = 1, 2, ..., k.

La prueba de χ^2 es aplicada a diferentes tipos de problemas. En este curso sólo interesan los siguientes :

1) Probar el ajuste de la distribución de una variable en la naturaleza a una distribución teórica.

2) Probar la existencia de independencia entre variables cualitativas.

1. Prueba de la hipótesis de bondad del ajuste

Una de las aplicaciones más corrientes de χ^2 es la que permite comprobar la concordancia existente entre una distribución teórica (Normal, Binomial, etc.) y la distribución empírica que se tiene. Esta es la razón por la cual se le ha denominado prueba de la bondad del ajuste.

Si se realiza un experimento con un determinado carácter, que se puede dividir en clases, al hacer un recuento de los individuos, cada clase presenta su correspondiente frecuencia observada. Es posible presumir que la distribución de este carácter siga un cierto modelo matemático, de acuerdo al cual se obtendrán las frecuencias calculadas.

Ejemplo : Se está interesado en determinar la "perfección"

de un dado, para lo cual es lanzado 180 veces. Si el dado carece de " vicio " cada una de las caras deberá aproximadamente aparecer 30 veces.

a) Planteamiento de hipótesis : como se supone que el dado es perfecto, la probabilidad de ocurrencia de cada cara es 1/6 (hipótesis nula : H_0). De manera que la frecuencia absoluta calculada, para cada cara, es 1/6 de 180, es decir, 30.

   Por oposición, existe una hipótesis alternativa (H_1) que supone la existencia de un "vicio" en el dado lo que significa que la probabilidad de ocurrencia de cada cara es cualquiera distinta de 1/6.

b) Se elige el nivel de significación, es decir, la probabilidad de cometer un error de tipo α. En este caso 0,05. La elección de este valor es un tanto arbitraria.

c) La distribución de χ^2 permite tomar una decisión, acerca de si las frecuencias observadas en la distribución empírica estudiada, discrepan en forma significativa, de las frecuencias calculadas de acuerdo a la hipótesis enunciada.

$$\chi^2 = \sum_{j=1}^{k} \frac{(O_j - C_j)^2}{C_j}$$

d) La región de rechazo de la hipótesis nula está formada por todos los valores de χ^2, cuya probabilidad de ocurrencia es igual o menor que 0,05. En el ejemplo en estudio, esta región está formada por todos los valores de $\chi^2 \geq 11,07$ (valor tabular de $\chi^2_{0,05}$ para 5 grados de libertad).

e) A continuación se debe calcular el valor de χ^2 para el estudio en cuestión y determinar si el valor de χ^2 observado cae o no en la región de rechazo.

   Los resultados son los siguientes :

Caras	Observadas
1	24
2	35
3	40
4	32
5	13
6	36
	180

CUADRO Nº 8

Cómputo de χ^2.

Prueba de bondad del ajuste

O	C	(O - C)	$(O - C)^2$	$\frac{(O - C)^2}{C}$
24	30	-6	36	1,20
35	30	5	25	0,83
40	30	10	100	3,33
32	30	2	4	0,13
13	30	-17	289	9,63
36	30	6	36	1,20
180	180	0		χ^2=16,32

f) Decisión. Como el valor de χ^2 obtenido es 16,32, el cual es mayor que 11,07 valor crítico para una probabilidad de cometer

un error de tipo α de 0,05, el cae en la región de rechazo de H_o. Por lo cual se rechaza la hipótesis nula de bondad del ajuste entre los datos observados y los correspondientes al modelo teórico y se acepta la hipótesis alternativa. Estableciéndose así que la distribución de los datos observados en la naturaleza discrepa de la planteada por el modelo teórico. Se concluye entonces que el dado tiene "vicio".

En el ejemplo estudiado el valor de $X^2 = 16,32$ (con 5 grados de libertad), tiene una probabilidad de ocurrencia menor que 0,01, en el supuesto que H_o es verdadera, lo que indica que se trata de un valor muy raro, poco común.

El número de grados de libertad, en la prueba de bondad del ajuste, no está dado por el número de observaciones, sino por el número de clases (k) cuyas frecuencias es necesario calcular, o sea, por el número de clases menos uno (k-1).

2. Prueba de la hipótesis de independencia.

Se considera el caso de una muestra de observaciones en cuya clasificación intervienen simultáneamente dos caracteres o variables de clasificación. Por medio de la distribución de probabilidades de χ^2, es posible someter a prueba la hipótesis de que ambas variables son independientes.

Las tablas de contingencia constituyen el medio más adecuado para estudiar la relación entre dos caracteres o variables de clasificación. Una tabla de contingencia es una tabla de doble entrada donde c son las columnas y f las filas. Existen por lo tanto tablas de 2x2, 2x3, ..., fxc.

El problema en esta prueba de hipótesis difiere de aquel estudiado en la sección 1, que en este caso a cada frecuencia observada le corresponde una frecuencia calculada, que es determinada de acuerdo a la hipótesis nula de independencia. Se designa como frecuencia celular a aquella que ocupa una casilla y como frecuencia marginal a la frecuencia total de cada fila o columna.

VARIABLE A

Variable B	+	−
+	Frecuencia celular a	Frecuencia celular b	Frecuencia marginal a+b
−	Frecuencia celular c	Frecuencia celular d	Frecuencia marginal c+d
	Frecuencia marginal a+c	Frecuencia marginal b+d	Frecuencia total n

Supóngase que se estudia en un grupo de vacunos de lechería el efecto de la inoculación de una nueva vacuna en la prevención de la fiebre aftosa y se analiza la existencia de relación entre estos dos caracteres.

El investigador que realiza el ensayo puede preguntarse si aquellos animales que han sido vacunados estarán mejor protegidos contra la enfermedad que aquellos animales que no han sido sometidos a vacunación.

a) Se parte de la hipótesis nula (H_o) de independencia, que plantea la no existencia de relación entre inoculación y prevención de la enfermedad. Esto quiere decir que se afirma que enferman de fiebre aftosa una proporción similar de animales inoculados con la vacuna y de animales no inoculados. Ambos caracteres son independientes. Por otra parte, la hipótesis alternativa (H_1) plantea la existencia de relación entre las variables. Es decir, que la probabilidad de enfermar de fiebre aftosa no es la misma en los animales inoculados que en los no inoculados.

b) El valor de la probabilidad de cometer un error de tipo α es 0,05.

c) Se utiliza la distribución de χ^2 para establecer si las frecuencias observadas discrepan en forma manifiesta, desde el punto de vista estadístico, de las frecuencias calculadas, de acuerdo a la hipótesis nula planteada.

d) La región de rechazo está constituida por aquellos valores de χ^2, cuya probabilidad de ocurrir es igual o inferior a 0,05. En este caso, por aquellos valores de $\chi^2 \geq 3{,}841$ (valor tabular $\chi^2_{0,05}$ con 1 grado de libertad).

e) Se calcula el valor de χ^2 para determinar si cae o no en la región de rechazo.

Los resultados de la experiencia son los siguientes :

CUADRO Nº 9

Prueba de la hipótesis de Independencia

Frecuencias Observadas

Inoculación de vacuna	Prevención de la Enfermedas		TOTAL
	Enferman de Fiebre Aftosa	No enferman de Fiebre Aftosa	
Inoculados	20	83	103
No Inoculados	45	30	75
TOTAL	65	113	178

Para obtener las frecuencias calculadas, se parte del hecho, que si existe independencia, la relación que se cumple para el total de bovinos que enferman 65/178=36,5 debe cumplirse para cualquiera de los subtotales. En otras palabras, la frecuencia calculada de cualquiera de las casillas se computa mediante una regla de tres simple o de una proporción.

De un total de 178 vacunos, enferman 65 de fiebre aftosa.

De un total de 103 vacunos inoculados, enferman X de fiebre aftosa.

$$\frac{65}{178} = \frac{X}{103}$$

$$X = \frac{(65)(103)}{178} = 38$$

Una vez obtenida la frecuencia calculada para una cierta casilla (a) de la tabla de contingencia, se pueden obtener las restantes por diferencias, puesto que se tienen los totales marginales.

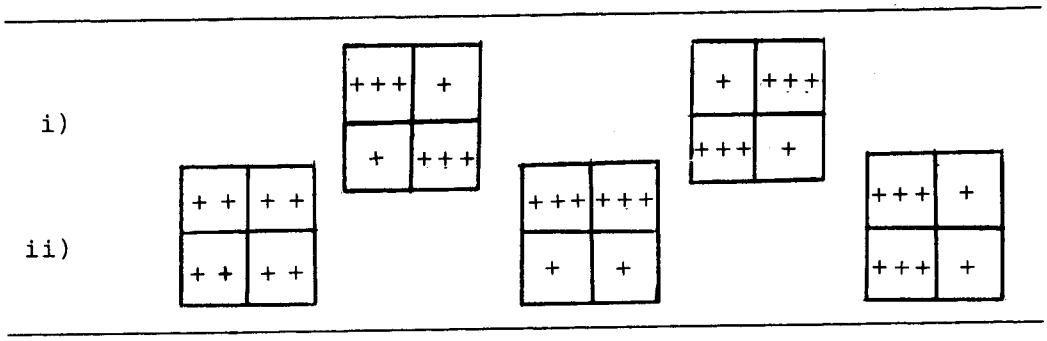
CUADRO Nº 10

Cómputo de χ^2. Prueba de la hipótesis de independencia

O	C	(O-C)	$(O-C)^2$	$\frac{(O-C)^2}{C}$
20	38	-18	324	8,53
83	65	18	324	4,98
30	48	-18	324	6,75
45	27	18	324	12,00
178	178	0		χ^2=32,26

Antes de hacer un comentario acerca del valor de χ^2 obtenido, se debe llamar la atención sobre un esquema, que en general , permite determinar fácilmente si existe o no independencia entre dos caracteres.

Fig. .1. Esquemas de relación entre dos variables de clasificación.



Si las observaciones se encuentran repartidas de tal manera que se observa una frecuencia mayor en las casillas a y d con respecto a las b y c, o si las casillas b y c presentan frecuencias más altas que las a y d, es posible suponer que existe relación entre ambos criterios de clasificación (Fig. 1.i.).

Ahora bien, si las observaciones se encuentran repartidas en forma más o menos homogénea en las cuatro casillas, o bien, si las frecuencias más altas se encuentran ya sea en la misma fila o en la misma columna, es dable suponer que existe independencia (Fig. 1.ii.).

El número de grados de libertad en una tabla de contingencia se obtiene de la siguiente manera : se sabe que en cada fila hay un valor que no es independiente y que en cada columna también existe un valor que no es independiente, por lo tanto,el número de grados de libertad, para cualquier tabla se puede calcular así.

$$G.L. = (f-1)\ (c-1)$$

También los grados de libertad se pueden obtener considerando que ellos corresponden al número de frecuencias que es necesario calcular.

f) Decisión : volviendo al ejemplo, se puede apreciar que la probabilidad de ocurrencia de un valor de χ^2 como el obtenido, considerando la H_o como verdadera, es menor que 0,001. Es decir, dado que 32,36 es mayor que 3,84, nuestro valor cae en la región de rechazo de H_o. Esto permite rechazar la hipótesis nula de independencia, no sólo al nivel 0,05 sino que al 0,001 y aceptar la hipótesis que ambos caracteres o variables de clasificación están relacionados.

El investigador que ha realizado la experiencia podría afirmar, con estos resultados, que la inoculación de la vacuna antiaftosa que él ensaya, protege a los bovinos de la enfermedad.

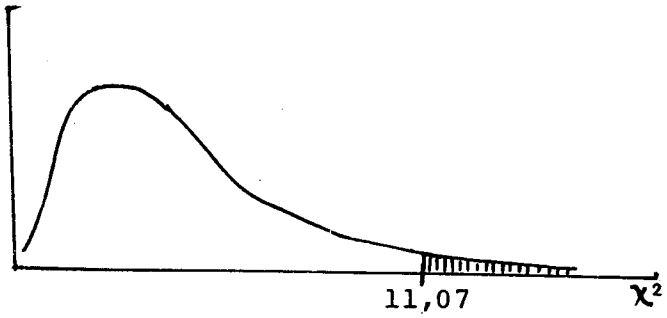
Ahora, conocidas las dos pruebas de hipótesis en que se ha utilizado la distribución de χ^2, se puede hacer un enfoque general al método.

Este tipo de pruebas permite comparar los datos observados, en un ensayo, con datos teóricos o calculados. Mientras mayor sea la semejanza entre ellos, más pequeño será el valor de χ^2, llegando a ser igual a cero, cuando existe un acuerdo perfecto entre lo observado y lo calculado.

A diferencia de otras distribuciones como la representada por la curva Normal, que es única y simétrica, la distribución de χ^2 toma diferentes formas de acuerdo al número de grados de libertad (Fig. 2.). Esta distribución, partiendo de curvas asimétricas, a medida que el número de grados de libertad aumenta, tiende a asemejarse a la Normal.

Si se ilustra gráficamente (Fig. 3.) el problema presentado en la sección 1., se ve que el valor de χ^2 que delimita la región de rechazo es 11,07. La ordenada levantada en este punto corta la cola derecha de la distribución, de manera que todo valor de $\chi^2 \geq 11,07$ tiene una probabilidad de ocurrir igual o menor que 0,05. Es decir, los valores iguales o superiores a 11,07 caen dentro de la llamada región crítica. Por lo tanto, los valores de χ^2 que caigan en esta área, determinan que la hipótesis que se ha planteado sea rechazada.

Fig. 3. Región de rechazo de H_o. 5 grados de libertad.



Consideraciones generales en las pruebas de χ^2.

1. - Las clases deben ser mutuamente excluyentes, es decir, un individuo u observación debe caer solamente dentro de una de ellas.

2. - Las clases deben ser exhaustivas, es decir, todos los individuos estudiados deben quedar incluidos en alguna de las clases.

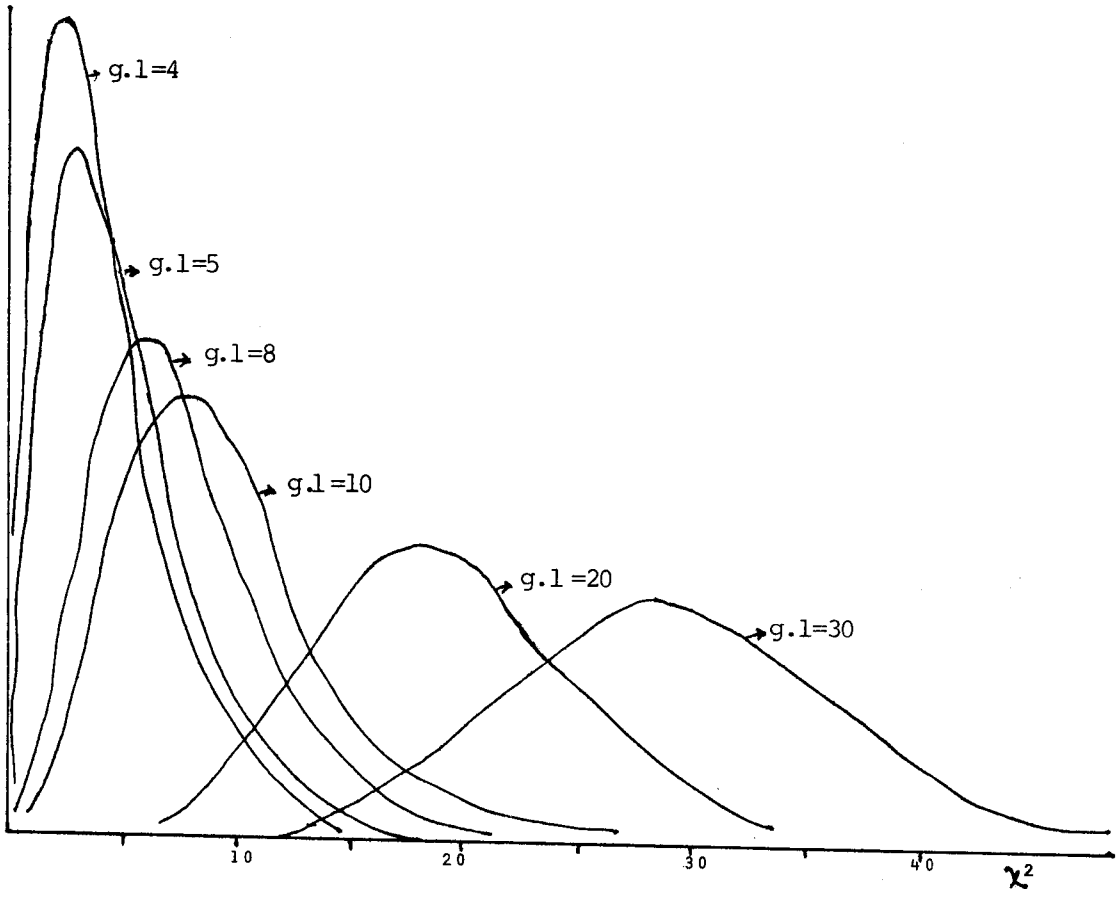
3. - La suma de las diferencias entre las frecuencias absolutas observadas (O_j) y las frecuencias absolutas calculadas (C_j), debe ser igual a cero.

4. - Se debe usar χ^2 solamente para cifras absolutas y no para porcentajes o proporciones. Esto se subsana transformando los porcentajes en frecuencias absolutas.

5. - Es importante también considerar el tamaño de la muestra (n), pues a medida que esta disminuye, la eficiencia de χ^2, como una medida del ajuste entre distribuciones de frecuencias, también disminuye. Se ha sugerido un valor mínimo de n=50.

En resumen, la prueba de χ^2 permite conocer la probabilidad de encontrar, en una muestra tomada al azar, un valor de χ^2 como el obtenido. Si la probabilidad de error es pequeña (inferior a 0,05) hay motivo para sospechar que hay una diferencia significativa entre la teoría y la experiencia. Hay tablas que relacionan el valor de χ^2, para distintos grados de libertad, con diferente probabilidad de cometer un error de tipo α.

Fig. 2. Distribución de χ^2.



6. - Debe evitarse la presencia de frecuencias absolutas calculadas muy pequeñas en las casillas individuales.

		VARIABLE A: +	VARIABLE A: −	Total
Variable B	+	Casilla a	Casilla b	a + b
	−	Casilla c	Casilla d	c + d
	Total	a + c	b + d	n

Si alguna de las frecuencias teóricas es menor que 1, la prueba de χ^2 no puede ser usada. El mínimo debe ser 5, o lo que es mejor 10.

Existen correcciones para frecuencias pequeñas, entre ellas, se menciona la de Yates, que consiste en restar o sumar a cada diferencia entre la frecuencia observada y la teórica, la cantidad de 0,5 antes de elevar al cuadrado, según si la diferencia es positiva o negativa. En otras palabras, al valor absoluto de la diferencia se le resta 0,5. Esto hace que χ^2 se defina de la siguiente manera :

$$\chi^2_{\text{Corregido}} = \sum_{j=1}^{k} \frac{(|O_j - C_j| - 0,5)^2}{C_j}$$

En muestras grandes χ^2 corregido y no corregido presentan prácticamente los mismos resultados. En muestras pequeñas es conveniente comparar los resultados que entregan ambos χ^2. Si

conducen a diferentes conclusiones se puede aumentar el tamaño de la muestra, o bien emplear otro método.

Esta correción se realiza cuando se trabaja con un grado de libertad.

7. - Cuando χ^2 se aplica como una prueba de independencia, entre dos variables de clasificación, el resultado indica si estas variables se encuentran o no relacionadas, sin hacer referencia al grado o al sentido de la relación.