TEMAS DE MATEMÁTICAS
(Oposiciones de Secundaria)

TEMA 67

<u>INFERENCIA ESTADÍSTICA. TEST DE HIPÓTESIS.</u>

1. Intervalos confidenciales.
2. Intervalos confidenciales para la media de una distribución normal.
3. Intervalos confidenciales para la varianza de una distribución normal.
4. Región confidencial para la media y la varianza de una distribución normal.
5. Intervalos confidenciales para el parámetro de una distribución binomial.
6. Intervalos confidenciales múltiples.
7. Intervalos confidenciales para muestras grandes.
8. Intervalos confidenciales múltiples.
9. Introducción al contraste de hipótesis.
10. Contraste de una hipótesis simple contra una alternativa simple.
11. Hipótesis compuestas.
12. Contraste de è è$_1$ contra è>è$_1$ para densidades con un parámetro único è.

Bibliografía Recomendada.

TEMA 67

<u>INFERENCIA ESTADÍSTICA. TEST DE HIPÓTESIS.</u>

1. <u>INTERVALOS CONFIDENCIALES.</u>

La estimación puntual de un parámetro no resulta de mucho valor si no se posee alguna medida del posible error cometido en la estimación. Toda estimación $\hat{q}$ de un parámetro q debería acompañarse de cierto intervalo que incluyera a $\hat{q}$, por ejemplo, de la forma ($\hat{q}$-d, $\hat{q}$+d), junto con alguna medida de seguridad de que el parámetro verdadero q fuera interior a dicho intervalo. A menudo las estimaciones se dan de esta manera. Así la carga electrónica puede estimarse que vale $(4{,}770 \pm 0{,}005)10^{-10}$ unidades electrostáticas, dándose a entender con ello que es muy poco probable que el primer factor sea exterior al intervalo 4,765 a 4,775. Un contable que se ocupe de los costes de una editorial, y que quiera tener en cuenta todos los factores que intervienen en le coste de producción de cierto libro (costes efectivos de producción, proporción de sostenimiento, proporción de sueldos directos, etc.), podrá estimar el coste en $83 \pm 4{,}5$ centavos por volumen, lo que significa que es muy probable que el coste correcto esté comprendido entre 78,5 y 87,5 centavos por volumen. La Oficina de Estadística del Trabajo puede estimar el número de parados en un momento dado en $2{,}4 \pm 0{,}3$ millones, teniendo bastante seguridad en que el número verdadero está comprendido entre 2,1 y 2,7 millones.

A fin de precisar a estas ideas, consideremos un caso particular. Supongamos una muestra (1,2; 3,4; 0,6; 5,6) de cuatro observaciones, extraída de una población normal de media desconocida m y desviación estándar conocida 3. La estimación máximo-verosímil de m es la media de las observaciones maestrales,

$$\bar{x} = 2{,}7$$

Queremos determinar los límites superior e inferior entre los cuales queda comprendido, con bastante seguridad, el valor verdadero del parámetro. En general, para muestras de tamaño cuatro, procedentes de la distribución dada, la cantidad

$$y = \frac{\bar{x} - m}{3/2}$$

tendrá una distribución normal con media cero y varianza unidad; $\bar{x}$ es la media muestral, y 3/2 es $s/\sqrt{n}$. Por tanto, la cantidad g tiene por función de densidad

$$f(y) = \frac{1}{\sqrt{2p}} e^{-\frac{1}{2}y^2} \qquad (3)$$

que es independiente del valor verdadero del parámetro desconocido, y se podrá calcular la probabilidad de que g esté situado entre dos números elegidos arbitrariamente. Así, por ejemplo,

$$P(-1{,}96 < y < 1{,}96) = \int_{-1,96}^{1,96} f(y)dy = 0{,}95 \qquad (4)$$

En esta relación, la desigualdad -1,96< g, o bien

$$-1{,}96 < \frac{\bar{x} - m}{3/2} \qquad (5)$$

equivale a la desigualdad

$$m < \bar{x} + 3/2(1{,}96) = \bar{x} + 2{,}94$$

y la desigualdad

$$g < 1{,}96$$

es equivalente a

$$m > \bar{x} - 2{,}96$$

cabe por tanto, volver a escribir (4) en la forma

$$P(\bar{x} - 2{,}94 < m < \bar{x} + 2{,}94) = 0{,}95 \qquad (6)$$

y sustituyendo $\bar{x}$ por 2,7

$$P(-0{,}24 < m < 5{,}64) = 0{,}95 \qquad (7)$$

Podemos decir que estos límites obtenidos, -0,24 y 5,64, contendrán el valor del parámetro verdadero, con una seguridad del 95%.

Debe examinarse cuidadosamente el significado de (6) y (7). La probabilidad de que el intervalo aleatorio $\bar{x}$ -2,94 a $\bar{x}$ +2,94 cubra a la media verdadera m es 0,95. Esto es, si se extraen repetidamente de la población muestras de tamaño 4, y si se calcula para cada muestra el intervalo aleatorio $\bar{x}$ -2,94 a $\bar{x}$ +2,94, es de esperar que le 95% de estos intervalos contengan la media verdadera m. Tenemos, pues, una gran confianza en que el intervalo -0,24 a 5,64 cubra la media verdadera. La medida de nuestra confianza es 0,95, porque antes de extraer la muestra, la probabilidad de que el intervalo que intentamos construir cubra la media verdadera es 0,95.

El intervalo -0,24 a 5,64 recibe el nombre de intervalo confidencial o, más concretamente, intervalo confidencial del 95%; la probabilidad, en este caso 0,95, se denomina coeficiente confidencial o coeficiente de confianza.

Es posible obtener intervalos con cualquier grado de confianza que se desee. Así, puesto que

$$P(-2{,}58 < y < 2{,}58) = 0{,}99 \qquad (8)$$

se obtiene un intervalo confidencial del 99% para la media verdadera considerando las desigualdades como antes, y sustituyendo $\bar{x}$ =2,7, con lo que resulta

$$P(-1{,}17 < m < 6{,}57) = 0{,}99 \qquad (9)$$

Debe observarse que hay muchos intervalos posibles con la misma probabilidad. Así , por ejemplo, ya que

$$P(-1{,}68 < y < 2{,}70) = 0{,}95 \qquad (10)$$

tenemos otro intervalo confidencial del 95% para m, dado por

$$P(-1{,}35 < m < 5{,}22) = 0{,}95 \qquad (11)$$

Este intervalo es inferior al de antes obtenido, ya que su longitud 6,57 es superior a la longitud 5,88 del intervalo dado en (7), por lo que procura una información menos precisa sobre la situación de m. Dos números cualesquiera a y b, tales que las ordenadas que les corresponden incluyan el 95%. En general, se desea que el intervalo confidencial sea lo más pequeño posible; esto se logra haciendo que a y b estén tan próximos como sea posible, ya que la relación P(a < g < b)=0,95 da lugar a un intervalo confidencial de longitud $(s/\sqrt{n})$ (b-a). La distancia (b-a) se hace mínima para un área dada cuando f(a) = f(b), como se ve claramente en la figura 1. si el punto b se desplaza un poco hacia la izquierda, el a deberá moverse una distancia menor hacia la izquierda, a fin de que el área siga siendo la misma; esta operación disminuye la longitud del intervalo y continua disminuyéndola mientras f(b) < f(a). Como en este ejemplo f(y) es simétrica respecto a y = 0, el valor mínimo de b – a, para un valor prefijado del área, corresponde a b = -a. Por tanto, (7) da el intervalo confidencial más corto del 95%, y (9) da el intervalo confidencial más corto del 99%, ambos para el parámetro m.

En muchos problemas no es posible construir los intervalos confidenciales más cortos para un coeficiente de confianza dado. En estos casos, resultará deseable hallara un intervalo confidencial que tenga las más corta longitud esperada, o que sea tal que haga mínima la probabilidad de que el intervalo confidencial cubra un valor m^*, donde $m^* \neq m$.

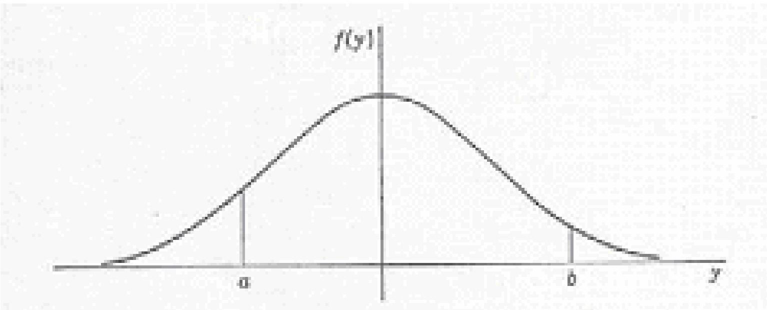

Fig. 1

El método general que aquí exponemos es el siguiente. Se halla, si es posible, una función de las observaciones muestrales y del parámetro a estimar (la función y anterior), cuya distribución sea independiente del parámetro y de otros parámetros cualesquiera. Entonces, cualquier afirmación probabilística de la forma P(a < y < b) = y. en donde g es la función, dará lugar a una afirmación probabilística relativa al parámetro. Esta técnica es aplicable en muchos problemas importantes, pero hay también otros muchos en los que no puede aplicarse, por ser imposible hallar funciones de la forma deseada y cuya distribución no dependa de parámetros. Estos últimos problemas se abordan mediante una técnica más general que describiremos en la sección 5.

La idea de la estimación por intervalos puede generalizarse de modo que incluya la estimación simultánea de varios parámetros. Así, los dos parámetros de la distribución normal se estiman mediante una cierta región plana R, en el llamado espacio paramétrico, espacio de todas las combinaciones posibles de los valores de m y s^2. Una región confidencial del 95% es una región que se puede construir a partir de la muestra, de tal forma que, si se extraen muestras repetidamente, construyendo una región para cada una de ellas, el 95% (por término medio) de estas regiones incluirán el punto paramétrico verdadero (m_0, s_0^2) . (véase figura 2)

Los intervalos y regiones confidenciales ilustran adecuadamente acera ce la incertidumbre de las inferencias. En (7) se hizo la inferencia de que el intervalo -0,24 a5,64 cubre el valor verdadero del parámetro, pero no se estableció de forma categórica. La medida 0,05 de la incertidumbre de esta inferencia constituye parte esencial de la afirmación.

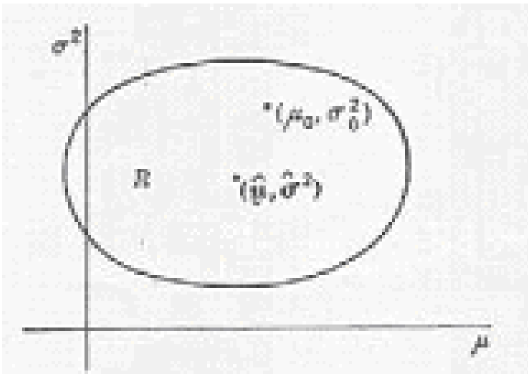

Fig. 2

2. INTERVALOS CONFIDENCIALES PARA LA MEDIA DE UNA DISTRIBUCIÓN NORMAL.

El método utilizado en la sección anterior no suele ser de posible utilización para estimar la media de una población normal, pues lo corriente es que se desconozca la varianza s^2. La función g toma la forma (para muestras de tamaño n)

$$y = \frac{\overline{x} - m}{s / \sqrt{n}} \qquad (1)$$

y transformando las desigualdades:

$$P(-1{,}96 < y < 1{,}96) = 0{,}95 \qquad (2)$$

se tiene

$$P\left(\bar{x} - 1{,}96\frac{s}{\sqrt{n}} < m < \bar{x} + 1{,}96\frac{s}{\sqrt{n}} \right) = 0{,}95 \qquad (3)$$

Para una muestra dada se conocen $\bar{x}$ y n, pero no s, de modo que no será posible calcular límites para m. Claro es que puede sustituirse s por una estimación $\hat{s}$; pero entonces la afirmación probabilística ya no sería exacta, y para muestras pequeñas podría ser muy errónea.

W. S. Gossett (que utilizó el seudónimo de Student) indicó el camino para resolver esta dificultad en una publicación clásica en que introdujo la distribución t. Se le considera como fundador de la teoría de la inferencia estadística exacta. La cantidad

$$t = \frac{\bar{x} - m}{\sqrt{\sum (x_i - \bar{x})^2 / n(n-1)}} \qquad (4)$$

comprende solo el parámetro m y tiene la distribución t con n -1 grados de libertad, sin incluir parámetros desconocidos. Por tanto, será posible hallar un número $t_{0,05}$ tal que

$$P(-t_{0,05} < t < t_{0,05}) = \int_{-t_{0,05}}^{t_{0,05}} f(t; n-1)dt = 0{,}90 \qquad (5)$$

convirtiendo después las desigualdades para obtener

$$P\left[\bar{x} - t_{0,05}\sqrt{\frac{\sum (x_i - \bar{x})^2}{n(n-1)}} < m < \bar{x} + t_{0,05}\sqrt{\frac{\sum (x_i - \bar{x})^2}{n(n-1)}} \right] = 0{,}90 \quad (6)$$

donde los limites se calculan para cada muestra dada, obteniendo así un intervalo confidencial del 90%.

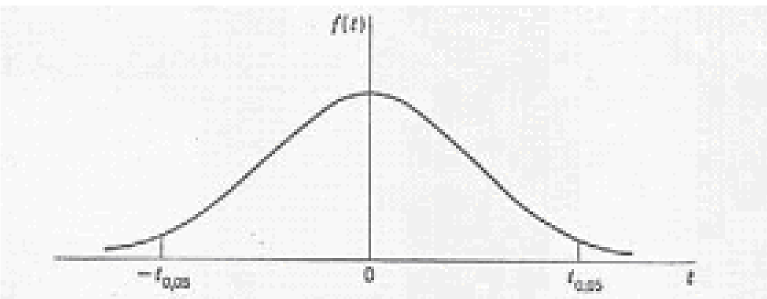

Fig. 3

El número $t_{0,05}$ recibe el nombre de nivel del 5% de t, y sitúa a los puntos que separa un 5% del área limitada por f(t) en cada rama de la curva. Cabe obtener otros intervalos confidenciales, empleando distintos niveles de t. Así, se puede hallar un intervalo confidencial del 98% usando el número $t_{0,01}$, que separa 0,01 del área en cada rama de la distribución. (véase fig. 3)

En este ejemplo la longitud del intervalo confidencial es

$$w = \overline{x} + t_{0,05}\sqrt{\frac{\sum (x_i - \overline{x})^2}{n(n-1)}} - \overline{x} + t_{0,05}\sqrt{\frac{\sum (x_i - \overline{x})^2}{n(n-1)}} = 2t_{0,05}\sqrt{\frac{\sum (x_i - \overline{x})^2}{n(n-1)}}$$

La longitud es una variable aleatoria, ya que es función de las variables aleatorias x_i. Es también función del tamaño n de la muestra en que se basa el intervalo confidencial. Si este es muy amplio, quizá resulta poco útil aunque sea alta la probabilidad de que cubra al parámetro desconocido. Así, es preciso que el tamaño n de la muestra sea suficientemente grande para que siendo la probabilidad alta, la longitud resulte lo bastante pequeña para ser útil.

3. INTERVALOS CONFIDENCIALES PARA LA VARIANZA DE UNA DISTRIBUCIÓN NORMAL.

Para muestras de tamaño n de una población normal la cantidad

$$u = \frac{\sum (x_i - \overline{x})^2}{s^2} \qquad (1)$$

donde $\overline{x}$ es la media muestral, tiene la distribución ji cuadrado con n – 1 grados de libertad. Por tanto, puede construirse un intervalo confidencial con coeficiente confidencial gamma, hallando dos números a y b tales que

$$P(a < u < b) = \int_a^b f(c^2)\,dc^2 = g \qquad (2)$$

Transformando las desigualdades, obtenemos

$$P\left[\frac{\sum (x_i - \overline{x})^2}{b} < s^2 < \frac{\sum (x_i - \overline{x})^2}{a}\right] = g \qquad (3)$$

que proporciona un intervalo confidencial para s^2.

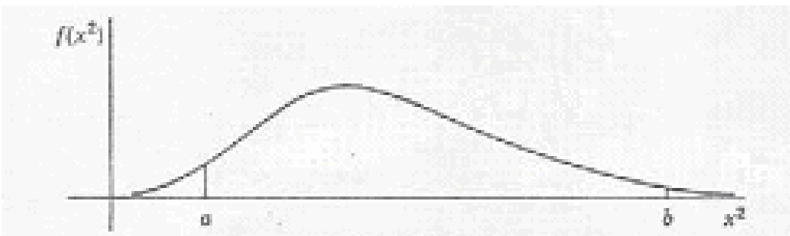

Fig. 5

Puesto que la longitud de este es

$$\left(\frac{1}{a}-\frac{1}{b}\right)\sum(x_i-\bar{x})^2 \qquad (4)$$

el intervalo confidencial más pequeño para una muestra dada se obtendría eligiendo a de modo que [(1/a)-(1/b)] resultase mínimo para el valor elegido de γ. El cálculo necesario resulta muy laborioso. Las tablas ordinarias de la distribución ji cuadrado proporcionan números χ^2 tales que

$$P(u > \chi_e^2) = \int_{\chi_e^2}^{\infty} f(\chi^2)\,d\chi^2 = \in \qquad (5)$$

para valores elegidos de $\in$. Al construir, por ejemplo, un intervalo confidencial del 95%, se suele elegir a = $\chi^2_{0,975}$ y b = $\chi^2_{0,025}$, esto es, se eligen a y b de modo que quede separado 0,025 del área en cada rama de la distribución. Esto da aproximadamente la longitud mínima del intervalo confidencial, a menos que el número de grados de libertad sea muy pequeño (véase fig. 5).

4. REGIÓN CONFIDENCIAL PARA LA MEDIA Y LA VARIANZA DE UNA DISTRIBUCIÓN NORMAL.

Al construir una región para la distribución conjunta de la media μ_0 y la varianza μ_0^2 de una distribución normal, cabría inclinarse a primera vista a utilizar las estimaciones individuales dadas por las distribuciones t y χ^2. Así, por ejemplo, podría construirse una región 0,9025 (= 0,95^2), como en la figura 6, haciendo uso de las dos relaciones:

$$P\left[\bar{x}-t_{0,025}\sqrt{\frac{\sum(x_i-\bar{x})^2}{n(n-1)}} < \mu_0 < \bar{x}+t_{0,025}\sqrt{\frac{\sum(x_i-\bar{x})^2}{n(n-1)}}\right] = 0{,}95 \qquad (1)$$

$$P\left[\frac{\sum(x_i-\bar{x})^2}{\chi^2_{0,025}} < \sigma_0^2 < \frac{\sum(x_i-\bar{x})^2}{\chi^2_{0,975}}\right] = 0{,}95 \qquad (2)$$

y suponiendo que la probabilidad de los dos sucesos fuera el producto de las dos probabilidades de cada uno. Esto no es correcto, puesto que las distribuciones de t y χ^2

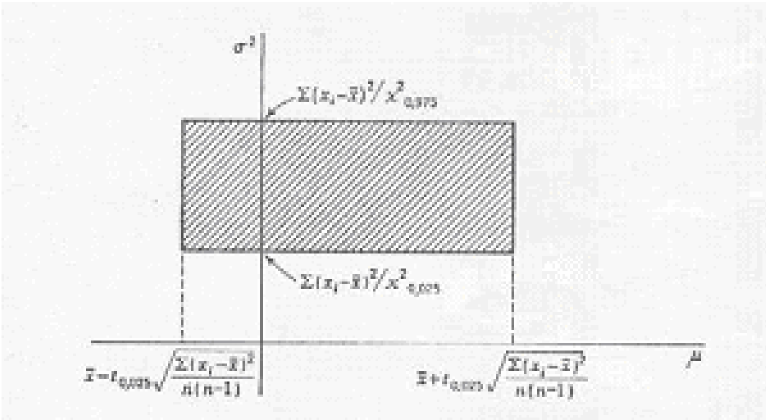

Fig. 6

no son independientes. La probabilidad conjunta de que ambos parámetros cubran los valores del parámetro verdadero no es igual al producto de las probabilidades correspondientes. Por tanto, la probabilidad de que la región rectangular de la figura 6 cubra al punto paramétrico verdadero (m_0, s_0^2) no es 0,9025.

Sin embargo es posible construir una región confidencial utilizando las distribuciones de $\bar{x}$ y $\sum (x_i - \bar{x})^2$, que son independientes. Si, por ejemplo, se desea una región confidencial del 95%, pueden hallarse números a, a’ y b’ tales que

$$P\left(-a < \frac{\bar{x} - m_0}{s_0 / \sqrt{n}} < a\right) = \sqrt{0{,}95} \cong 0{,}975 \qquad (3)$$

$$P\left[a' < \frac{\sum (x_i - \bar{x})^2}{s_0^2} < b'\right] = \sqrt{0{,}95} \qquad (4)$$

La probabilidad conjunta es

$$P\left[-a < \frac{\bar{x} - m_0}{s_0 / \sqrt{n}} < a,\ a' < \frac{\sum (x_i - \bar{x})^2}{s_0^2} < b'\right] = 0{,}95 \qquad (5)$$

debido a la independencia de las distribuciones. Las cuatro desigualdades de (5) determinan una región en el espacio paramétrico, fácil de determinar trazando las líneas que la limitan.

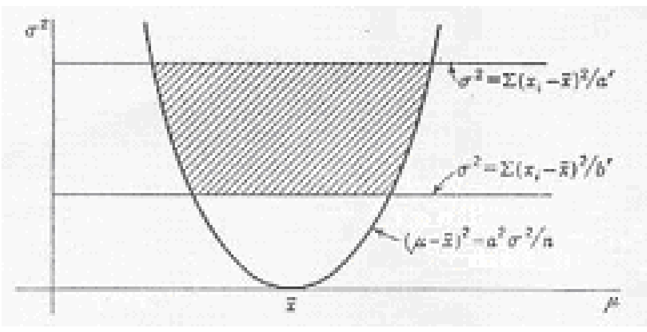

Fig. 7

Basta reemplazar los signos de desigualdad por otros de igualdad y representar cada una de las cuatro relaciones resultantes como funciones de m y s^2 en el espacio paramétrico. Resultará así una región como la que aparece rayada en la figura 7. Exactamente del mismo modo se obtendría una región confidencial para (m_0, s_0); la relaciones se representarían como funciones de s en lugar de s^2, y la parábola de la figura 7 se transformaría en un par de rectas

$$m = \bar{x} \pm \frac{a\,s}{\sqrt{n}}$$

que se cortarían en $\bar{x}$ sobre el eje de las m.

La región que hemos construido no es la de área mínima, pero se construye fácilmente a partir de las tablas y difiere poco de la región de área mínima, a menos que sea pequeño el tamaño de la muestra. La región mínima es, aproximadamente, de forma elíptica y difícil de construir.

5. MÉTODO GENERAL PARA LA OBTENCIÓN DE INTERVALOS CONFIDENCIALES.

El método utilizado en las secciones anteriores para la determinación de intervalos y regiones confidenciales obliga a encontrar funciones de la muestra y de los parámetros, distribuidas independientemente de estos. No obstante, es posible establecer intervalos confidenciales sin tener en cuenta la existencia previa de tales funciones.

Dada una población por f(x ; q) y un estimador $(x_1, x_2, ..., x_n)$ para muestras de tamaño n (generalmente, se usará el estimador de máxima verosimilitud), determinaremos la distribución del estimador, que vendrá dada por g($\hat{q}$;q). Supongamos, para fijar ideas, que se desea un intervalo confidencial del 95%. Si se sustituye q , en g($\hat{q}$;q), por el número arbitrario q', la distribución de $\hat{q}$ quedará completamente especificada, y será posible dar enunciados probabilísticos relativos a $\hat{q}$. En particular, será posible hallar dos números h_1 y h_2 tales que

$$P\left(\hat{q} < h_1\right) = \int_{-\infty}^{h_1} g(\hat{q};q')\,d\hat{q} = 0{,}025 \qquad (1)$$

$$P\left(\hat{q} > h_2\right) = \int_{-\infty}^{h_2} g(\hat{q};q')\,d\hat{q} = 0{,}025 \qquad (2)$$

Claro es que los números h_1 y h_2 dependerán del número que sustituye a $\hat{q}$ en g($\hat{q}$;q). En efecto, h_1 y h_2 son ciertas funciones de q, esto es $h_1(q)$ y $h_2(q)$. Los valores de estas funciones para cualquier valor de q vienen determinados por las dos ecuaciones anteriores. Evidentemente,

$$P\left[h_1(q) < \hat{q} < h_2(q)\right] = \int_{h_1(q)}^{h_2(q)} g(\hat{q};q)\,d\hat{q} = 0{,}95 \qquad (3)$$

Las funciones $h_1(q)$ y $h_2(q)$ pueden representarse en función de q, como se ha hecho en la figura 8. Trazando una vertical por cualquier valor q' de q, esta cortará a ambas curvas en puntos que, proyectados sobre el eje de las $\hat{q}$, darán limites entre los cuales caerá $\hat{q}$, con probabilidad de 0,95.

Construidas las dos curvas $\hat{q} = h_1(q)$ y $\hat{q} = h_2(q)$, cabe obtener un intervalo confidencial para q del siguiente modo: Se extrae una muestra de tamaño n y se calcula el valor del estimador $\hat{q}'$. La horizontal trazada por el punto $\hat{q}'$ del eje $\hat{q}$ (fig. 8) cortará a ambas curvas en puntos que pueden proyectarse sobre el eje q y que llamaremos q_1 y q_2, según se indica en la figura. Estos dos números definen el intervalo confidencial, pues se ve fácilmente que

$$P(q_2 < q < q_1) = 0{,}95 \qquad (4)$$

Supongamos que estuviésemos extrayendo muestras de una población en que el valor de q fuese q'. La probabilidad de que la estimación $\hat{q}$ quede comprendida entre $h_1(q')$ y $h_2(q')$ es 0,95. Si la estimación cae entre estos dos limites, dicho horizontal cortará a la vertical trazada por q' en cierto punto situado entre las curvas, y el intervalo correspondiente (q_2, q_1) cubrirá a q'. Se deduce, por tanto, que la probabilidad de que un intervalo (q_2, q_1), construido por este método, cubra a q', es exactamente 0,95. Esta afirmación es cierta cualquiera que sea el valor de q en la población. A veces, es posible determinar los límites q_2 y q_1 para una estimación dada, si necesidad de hallar efectivamente las funciones $h_1(q)$ y $h_2(q)$.

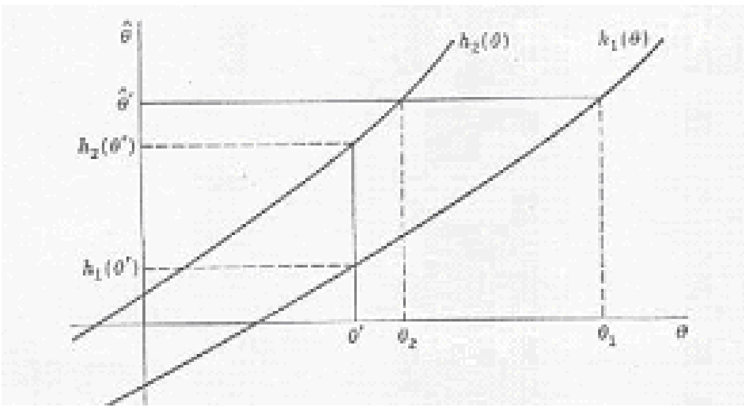

Fig. 8

Con referencia a la fig 8, los limites para q son los puntos q_2 y q_1, tales que $h_1(q_1) = \hat{q}'$ y $h_2(q_2) = \hat{q}'$. Basándonos en la definición de h_1 y h_2, diremos que q_1 es el valor de q para el cual

$$\int_{-\infty}^{\hat{q}'} g(\hat{q};q)\,d\hat{q} = 0{,}025 \qquad (5)$$

y q_2 es el valor de q para el cual

$$\int_{\hat{q}'}^{\infty} g(\hat{q};q)\,d\hat{q} = 0{,}025 \qquad (6)$$

Si es posible expresar los primeros miembros de estas dos ecuaciones explícitamente en función de q, y si las ecuaciones pueden resolverse unívocamente respecto a q, las raíces son los límites confidenciales del 95%, para q.

Si $h_1(q)$ y $h_2(q)$ no son funciones monótonas de q, el intervalo confidencial puede ser, en realidad, un conjunto de intervalos. Así, por ejemplo, supongamos que las curvas de la figura 8 se inclinaran mas hacia la derecha de modo que la horizontal trazada por q' volviera a cortarlas, por ejemplo, en los puntos q_3 y q_4. El intervalo confidencial consistiría en dos intervalos (q_2, q_1) y (q_3, q_4). La afirmación sobre q sería de la forma

$$P(q_2 < q < q_1, ó, q_3 < q < q_4) = 0{,}95 \qquad (7)$$

Sin embargo, en la mayoría de las situaciones que se plantean en la práctica habrá un intervalo único, o será posible elegir un intervalo único basándose en otros datos disponibles relativos al experimento que dio lugar a las observaciones maestrales.

El método aquí descrito para la obtención de intervalos confidenciales se extiende al caso de varios parámetros; pero la representación geométrica ya no es posible, ni siquiera para dos parámetros. Supongamos una distribución que dependa de dos parámetros q_1 y q_2; podemos hallar una región plana R en el plano $\hat{q}_1, \hat{q}_2$ tal que

$$P(\hat{q}_1, \hat{q}_2 en R) = \iint_R g(\hat{q}_1, \hat{q}_2; q_1, q_2)\,d\hat{q}_1 d\hat{q}_2 = 0{,}95 \qquad (8)$$

Considerando todos los pares posibles de valores q_1 y q_2 limitaremos una región cuatridimensional en el espacio, q_1, q_2, $\hat{q}_1$, $\hat{q}_2$, que es análoga a la región bidimensional entre las curvas de la figura 8. Supongamos ahora que se extrae una muestra y se calculan las estimaciones $\hat{q}_1$' y $\hat{q}_2$'. La intersección de los dos hiperplanos $\hat{q}_1 = \hat{q}_1$' y $\hat{q}_2 = \hat{q}_2$' con la región cuatridimensional determinará una región bidimensional que, proyectada sobre el plano q_1, q_2, será una región confidencial del 95% para q_1, q_2.

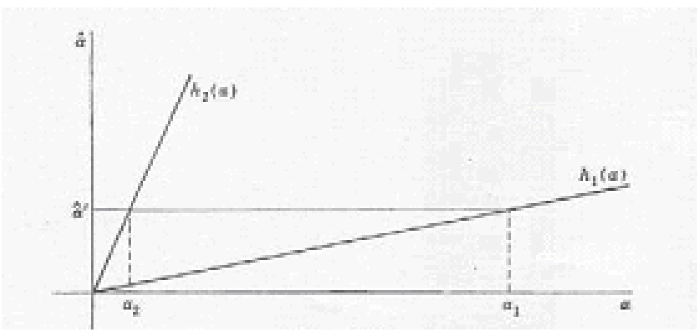

Fig. 9

Este razonamiento se generaliza para abarcar el caso de K parámetros. El método determinará una región confidencial para todos los parámetros de una distribución. Si se desea estimar algunos, pero no todos los parámetros de un conjunto de ellos, dicho método no podrá usarse en general, pero en determinadas circunstancias si puede modificarse para adaptarse al problema en cuestión. Por ahora, no hay solución general del problema de construir regiones confidenciales para una parte del conjunto de K parámetros de una función de distribución, excepto en el caso de muestras grandes.

6. INTERVALOS CONFIDENCIALES PARA EL PARÁMETRO DE UNA DISTRIBUCIÓN BINOMIAL.

Aplicaremos el método general descrito en la sección precedente a un problema que exige su empleo. Si una muestra $x_1, x_2, ..., x_n$ procede de una población binomial con

$$f(x;p) = p^x (1-p)^{1-x} \qquad x = 0{,}1; \quad 0 \le p \le 1 \qquad (1)$$

el estimador máximo-verosímil de p es

$$\hat{p} = \frac{y}{n} \qquad (2)$$

en donde $y = \sum x_i$ puede tomar los valores 0, 1, 2, …, n. La distribución de $\hat{p}$ viene dada por

$$g(\hat{p};p) = \binom{n}{n\hat{p}} p^{n\hat{p}} (1-p)^{n(1-\hat{p})} \qquad \hat{p} = 0, \frac{1}{n}, \frac{2}{n}, ..., 1 \qquad (3)$$

y no es posible hallar una función de $\hat{p}$ y p, cuya distribución sea independiente de p.

Volveremos a suponer, para fijar ideas, que el intervalo confidencial a construir es del 95%. El primer paso consiste en determinar las funciones $h_1(p)$ y $h_2(p)$. Así, para p = 0,4, y de acuerdo con la sección anterior, buscaríamos un número $h_1(0{,}4)$, tal que

$$P[\hat{p} < h_1(0{,}4)] = \sum_{y=0}^{nh_1} \binom{n}{y} (0{,}4)^y (0{,}6)^{n-y} = 0{,}025 \qquad (4)$$

No obstante, por tratarse de una distribución discreta, nh_1 deberá ser un entero, y será imposible lograr que la suma valga exactamente 0,025 para todo valor de p. Sin embargo, no nos preocuparemos por esto, ya que no necesitamos una curva $h_1(p)$ definida para todo valor de p. Los únicos puntos de interés son los que corresponden a valores posibles de $\hat{p}$. En efecto, es posible utilizar la técnica indicada por las ecuaciones (5-5) y (5-6), por disponerse inmediatamente de una expresión explicita por las probabilidades que figuran en el primer miembro de dichas ecuaciones. Suponiendo que tenemos una estimación

$$\hat{p}' = \frac{k}{n} \qquad (5)$$

puede determinarse el límite superior confidencial de p_1, del 95%, hallando el valor de p para el cual

$$\sum_{y=0}^{k}\binom{n}{y}p^y(1-p)^{n-y} = 0{,}025 \qquad (6)$$

siendo el límite inferior p_2 el valor de p para el cual

$$\sum_{y=k}^{n}\binom{n}{y}p^y(1-p)^{n-y} = 0{,}025 \qquad (7)$$

Si es k = 0, se toma cero como límite inferior, y si k = n, se toma 1 como límite superior.

Para valores pequeños de n, las ecuaciones (6) y (7) pueden resolverse por tanteos, a fin de obtener las raíces p_1 y p_2; pero este cálculo se hace más prolijo a medida que aumenta n. Un método sencillo consiste en utilizar las tablas de Pearson para la función beta incompleta. La forma acumulativa de la distribución beta es

$$F(x;\boldsymbol{a},\boldsymbol{b}) = \frac{(\boldsymbol{a}+\boldsymbol{b}+1)!}{\boldsymbol{a}!\,\boldsymbol{b}!}\int_0^x t^a(1-t)^b\,dt \qquad (8)$$

y por integración reiterada por partes se obtiene

$$F(x;\boldsymbol{a},\boldsymbol{b}) = -\sum_{i=0}^{a}\binom{\boldsymbol{a}+\boldsymbol{b}+1}{i}x^i(1-x)^{a+b+1-i} + 1 \qquad (9)$$

Se deduce que las sumas binomiales parciales vienen dadas por la tabla de F(x; $\boldsymbol{ab}$). Podemos escribir la ecuación (6) del siguiente modo:

$$\sum_{y=0}^{k}\binom{n}{y}p^y(y-p)^{n-y} = 1 - F(p;k,n-k-1) = 0{,}025 \qquad (10)$$

hallando inmediatamente en la tabla el valor de p que corresponde a F =0,975 para los valores dados de k y n – k -1. Análogamente, puesto que

$$\sum_{k}^{n}\binom{n}{y}p^y(1-p)^{n-y} = 1 - \sum_{0}^{k-1}\binom{n}{y}p^y(1-p)^{n-y}$$

se obtendrá el límite confidencial inferior escribiendo (7) en la forma

$$\sum_{n}^{k}\binom{n}{y}p^y(1-p)^{n-y} = F(p;k-1,n-k) = 0{,}025 \qquad (11)$$

Para valores de n que excedan de los tabulados, puede emplearse la aproximación normal a la distribución binomial, y obtener intervalos confidenciales de p, tal como se indica en la sección siguiente, o bien utilizar las *Tables of the Binomial Probability Distribution (National Boreau of Standards, Applied Mathematics Series 6, Washington DC, 1950).*

7. INTERVALOS CONFIDENCIALES PARA MUESTRAS GRANDES.

Para muestras grandes, el estimulador $\hat{q}$ máximo-verosímil para el parámetro q de una distribución dada por f(x; q) tiene, bajo condiciones bastante generales, una distribución aproximadamente normal respecto de q. Cuando se satisfacen tales condiciones, se obtienen fácilmente intervalos confidenciales aproximados. La varianza del estimulador en las muestras grandes es

$$s^2(q) = \frac{-1}{nE\left[\partial^2 \log f(x;q) / \partial q^2\right]} \qquad (1)$$

en donde $s^2(q)$ indica que es una función de q, porque ordinariamente dependerá de este parámetro. Para muestras grandes, por tanto, puede determinarse un intervalo confidencial con probabilidad g, convirtiendo las desigualdades en

$$P\left[-d_g < \frac{\hat{q}-q}{s(q)} < d_g\right] \cong 2g \qquad (2)$$

en donde d_g se ha elegido de modo que

$$\int_{-d_l}^{d_g} \frac{1}{\sqrt{2p}} e^{-\frac{1}{2}t^2}\, dt = 2g$$

8. INTERVALOS CONFIDENCIALES MÚLTIPLES.

En la secciones anteriores hemos indicado que la interpretación frecuencial-probabilística de los intervalos confidenciales es la siguiente: En repetidos muestreos, 100(1-a)% de los intervalos confidenciales construidos contendrán el parámetro desconocido q, donde 1 - a es el coeficiente confidencial. Para ilustrar esta interpretación con mayor precisión, supongamos que se extrae una muestra aleatoria de tamaño k de cada una de 3 poblaciones normales de medias m_1, m_2 y m_3, respectivamente y varianza común s^2.

Construiremos un intervalo confidencial del 95% para $m_1 - m_2$, $m_2 - m_3$ y $m_1 - m_3$. Para hallar un intervalo confidencial para $m_1 - m_2$ tenemos en cuenta que $\bar{x} - m_1$ es normal, con media 0 y varianza s^2/k; $\bar{y} - m_2$ es normal con media 0 y varianza s^2/k; $\bar{y} - m_2$ y $\bar{x} - m_1$ son independientes luego

$$w = (\bar{x} - m_1) - (\bar{y} - m_2) = (\bar{x} - \bar{y}) - (m_1 - m_2)$$

es normal con media 0 y varianza 2 $\boldsymbol{s}^2$ /k y , por tanto,

$$\frac{w}{\sqrt{2\boldsymbol{s}^2/k}}$$

es también normal, con media 0 y varianza 1. Si hacemos

$$s_1^2 = \frac{1}{k-1}\sum(x_i - \overline{x})^2$$

$$s_2^2 = \frac{1}{k-1}\sum(y_i - \overline{y})^2$$

$$s_3^2 = \frac{1}{k-1}\sum(z_i - \overline{z})^2$$

entonces

$$\frac{3(k-1)s^2}{\boldsymbol{s}^2} = \frac{(k-1)s_1^2 + (k-1)s_2^2 + (k-1)s_3^2}{\boldsymbol{s}^2} \qquad (2)$$

se distribuye según una ji cuadrado con 3k -3 grados de libertad, y s^2 es independiente de w. Por tanto,

$$t = \frac{w(\sqrt{k}\boldsymbol{s})}{\sqrt{2}s\boldsymbol{s}} = \frac{w\sqrt{k}}{s\sqrt{2}}$$

se distribuye según una t de Student con 3(k -1) grados de libertad. Un intervalo confidencial del 95% para m_1 - m_2 es

$$P\left[(\overline{x}-\overline{y}) - t_{0,025}\sqrt{\frac{2s^2}{k}} < m_1 - m_2 < (\overline{x}-\overline{y}) + t_{0,025}\sqrt{\frac{2s^2}{k}}\right] = 0{,}95 \qquad (3)$$

Por un proceso semejante se deduce que un intervalo confidencial del 95% para m_1 - m_3 es

$$P\left[(\overline{x}-\overline{z}) - t_{0,025}\sqrt{\frac{2s^2}{k}} < m_1 - m_3 < (\overline{x}-\overline{z}) + t_{0,025}\sqrt{\frac{2s^2}{k}}\right] = 0{,}95 \qquad (4)$$

y, análogamente, un intervalo confidencial del 95% para m_2 - m_3 es

$$P\left[(\overline{y}-\overline{z}) - t_{0,025}\sqrt{\frac{2s^2}{k}} < m_2 - m_3 < (\overline{y}-\overline{z}) + t_{0,025}\sqrt{\frac{2s^2}{k}}\right] = 0{,}95 \qquad (5)$$

Si se toman repetidos conjuntos de observaciones (1), y se calcula (3) para cada conjunto de 3k observaciones, entonces, para un número grande de repeticiones, el 95% de los intervalos confidenciales cubrirán a m_1 - m_2 .

Si para cada conjunto de 3k observaciones se calcula el intervalo confidencial (4), para un número grande de repeticiones el 95% de estos intervalos cubrirán a $m_1 - m_3$. Análogamente, si para cada conjunto se calcula el intervalo confidencial (5), en un número grande de repeticiones, el 95% de los intervalos contendrán a $m_2 - m_3$. Deseamos calcular intervalos confidenciales para $m_1 - m_2$, $m_1 - m_3$ y $m_2 - m_3$, tales que la probabilidad de que los tres intervalos confidenciales resulten simultáneamente verdaderos sea, por ejemplo, el 95%. Si los tres intervalos dados por (3) a (5) fuesen independientes, en un número grande de repeticiones, para el $(0,95)^3$ de los conjuntos, (3) cubriría a $m_1 - m_2$, (4) cubriría a $m_1 - m_3$, y (5) cubriría a $m_2 - m_3$. Sin embargo, puesto que (3), (4) y (5) no son independientes, esta probabilidad no es $(0,95)^3$. Para resolver este problema definiremos el coeficiente confidencial experimentativo. Un conjunto de observaciones tales como (1) recibirá el nombre de experimento; puede haber t poblaciones en lugar de 3. En cada experimento, se calculan intervalos confidenciales para las t(t-1) diferencias $m_i - m_j$. Si en el 95% de los experimentos la totalidad de los t(t-1) intervalos confidenciales cubren a sus diferencias respectivas ($m_i - m_j$), diremos que el coeficiente confidencial experimentativo es 0,95.

Enunciaremos el siguiente teorema aunque no daremos su demostración.

Teorema.

Sea $v_1, v_2, ..., v_n$ una muestra aleatoria de tamaño n de una población normal de media 0 y varianza $\boldsymbol{s}^2$, y designaremos por R el recorrido de estas variables aleatorias; es decir, R = máx v_i -mínv_i. Supongamos que $vs^2/\boldsymbol{s}^2$ es independiente de las v_i y esta distribuida según una ji cuadrado con v grados de libertad. La variable aleatoria

$$q = \frac{R}{s}$$

se distribuye como el recorrido studentizado, con n y v grados de libertad en el numerador y en el denominador, respectivamente.

La función frecuencial de q es bastante complicada y no se dará aquí, pero la cantidad q_a, definida por P(q< q_a)=1-$\boldsymbol{a}$, puede obtenerse en para varios valores de n, v y $\boldsymbol{a}$ = 0,01, 0,05 y 0,10.

Para ilustrar como puede emplearse este teorema, hallaremos un conjunto de intervalos confidenciales con un coeficiente confidencial experimentativo del 0,95. Consideremos las variables aleatorias (nos limitaremos al caso especial de 3)

$$\frac{3(k-1)s^2}{\boldsymbol{s}^2}, u_1, u_2, u_3$$

en donde s^2 está dada por (2), y u_1, u_2, u_3 son los estadísticos ordinales de las tres variables aleatorias v_1, v_2, v_3 con

$$v_1 = (\bar{x} - m_1)\sqrt{k} \quad v_2 = (\bar{y} - m_2)\sqrt{k} \quad v_3 = (\bar{z} - m_3)\sqrt{k}$$

Puesto que las v_i son variables normales independientes, de medias 0 y varianzas $\boldsymbol{s}^2$, y dado que $3(k-1)s^2/\boldsymbol{s}^2$ es una variable de ji cuadrado independiente, con v = 3(k -1) g. de l., utilizaremos el teorema 1 para demostrar que q se distribuye como el recorrido studentizado, con n =3 g. de l. en el numerador y v =3(k -1) g. de l. en el denominador, siendo

$$q = \frac{R}{s} = \frac{u_3 - u_1}{s} = \frac{\max v_i - \min v_i}{s}$$

También

$$1 - \boldsymbol{a} = P(q < q_a) = P\left(\frac{u_3 - u_1}{s} < q_a\right) =$$

$$= P\left(\frac{\max v_i - \min v_i}{s} < q_a\right) =$$

$$= P(\max v_i - \min v_i < sq_a) \qquad (6)$$

Pero si $\max v_i - \min v_i < \mathrm{s}q_a$, se tienen las tres desigualdades siguientes:

$$\left|(\bar{x} - m_1) - (\bar{y} - m_2)\right| < \frac{sq_a}{\sqrt{k}}$$

$$\left|(\bar{x} - m_1) - (\bar{z} - m_3)\right| < \frac{sq_a}{\sqrt{k}}$$

y

$$\left|(\bar{y} - m_2) - (\bar{z} - m_3)\right| < \frac{sq_a}{\sqrt{k}}$$

lo que implica

$$-\frac{sq_a}{\sqrt{k}} < (\bar{x} - \bar{y}) - (m_1 - m_2) < \frac{sq_a}{\sqrt{k}}$$

$$-\frac{sq_a}{\sqrt{k}} < (\bar{x} - \bar{z}) - (m_1 - m_3) < \frac{sq_a}{\sqrt{k}} \qquad (7)$$

$$-\frac{sq_a}{\sqrt{k}} < (\bar{y} - \bar{z}) - (m_2 - m_3) < \frac{sq_a}{\sqrt{k}}$$

Si utilizamos (7) con (6), la probabilidad de que las seis desigualdades (8) sean verdaderas es 1 - $\boldsymbol{a}$:

$$(\bar{x}-\bar{y})-\frac{sq_a}{\sqrt{k}}< m_1-m_2 <(\bar{x}-\bar{y})+\frac{sq_a}{\sqrt{k}}$$

$$(\bar{y}-\bar{x})-\frac{sq_a}{\sqrt{k}}< m_2-m_1 <(\bar{y}-\bar{x})+\frac{sq_a}{\sqrt{k}}$$

$$(\bar{x}-\bar{z})-\frac{sq_a}{\sqrt{k}}< m_1-m_3 <(\bar{x}-\bar{z})+\frac{sq_a}{\sqrt{k}}$$

$$(\bar{z}-\bar{x})-\frac{sq_a}{\sqrt{k}}< m_3-m_1 <(\bar{z}-\bar{x})+\frac{sq_a}{\sqrt{k}}$$

$$(\bar{y}-\bar{z})-\frac{sq_a}{\sqrt{k}}< m_2-m_3 <(\bar{y}-\bar{z})+\frac{sq_a}{\sqrt{k}}$$

$$(\bar{z}-\bar{y})-\frac{sq_a}{\sqrt{k}}< m_3-m_2 <(\bar{z}-\bar{y})+\frac{sq_a}{\sqrt{k}}$$

En el caso de haber más de tres poblaciones, serían válidas las mismas fórmulas, salvo que variarían los grados de libertad para q_a y que existirían t(t-1) intervalos confidenciales.

9. INTRODUCCIÓN AL CONTRASTE DE HIPÓTESIS.

La inferencia estadística comprende dos partes principales, a saber: la estimación de parámetros y los contrastes de hipótesis. En este capítulo estudiaremos la segunda de ellas, con el objetivo de desarrollar métodos generales para los contrastes de hipótesis y su aplicación a algunos problemas corrientes. Estos métodos también se utilizarán en capítulos posteriores.

En la investigación experimental se pretende a veces simplemente estimar un parámetro; por ejemplo, puede que interese estimar la producción de un nuevo híbrido de maíz. Muchas veces, el objetivo final es la utilización de dicha estimación. Así ocurre cuando se quiere comparar la producción del nuevo híbrido con la correspondiente a una variedad conocida, a fin de recomendar la sustitución de esta por aquel, en caso de que parezca superior. Esto sucede corrientemente en la investigación; puede ocurrir que interese determinar si un método nuevo para cerrar lámparas aumenta la vida de éstas; si un nuevo germicida resulta más efectivo en el tratamiento de cierta infección; si un método de conservación de alimentos es preferible a otros, en lo que se refiere a la conservación de vitaminas, etc.

Utilizando como ejemplo el caso de las lámparas, supongamos que la vida media de las fabricadas por medio de un proceso conocido es de 1400 h. Se desea contrastar un nuevo procedimiento para la fabricación de lámparas. En este caso, el modelo estadístico es el siguiente: se trata de dos poblaciones de lámparas, la constituida por las correspondientes al proceso que se propone. Sabemos (en virtud de numerosas investigaciones anteriormente realizadas) que la media de la primera población es aproximadamente 1400. Se desea averiguar si la media de la segunda población es superior o inferior a 1400. Tradicionalmente, para resolver este problema, se establece la hipótesis de que una medida es mayor que la otra. Basándose en una muestra de las poblaciones se aceptará o rechazará la hipótesis. (Naturalmente, se confía en que el nuevo proceso es mejor y que la hipótesis será rechazada).

Para contrastar la hipótesis se fabrica cierto número de lámparas mediante el nuevo procedimiento, midiendo después su duración. Supongamos que la media de esta muestra de observaciones es de 1550 h. Esto parece indicar que el nuevo proceso es mejor; pero supongamos que la estimación de la desviación estándar de la media es $d/\sqrt{n}$, igual a 125 (siendo n el tamaño de la muestra). Por tanto, el intervalo confidencial del 95% para la media de la segunda población (suponiendo la población normal) es aproximadamente de 1300 h a 1800 h. La media muestral 1550 podría proceder fácilmente de una población cuya media fuese 1400. No tenemos, pues, motivos suficientes para rechazar la hipótesis. Por otra parte, si $d/\sqrt{n}$ fuese igual a 25, podríamos rechazar la hipótesis con gran confianza y afirmar la superioridad del nuevo proceso de fabricación.

Se ve, pues, que los contrastes de hipótesis está relacionada íntimamente con el problema de la estimación. No obstante resulta instructivo desarrollar la teoría de los contrastes independientemente de la de la estimación al menos en principio.

Los contrastes de hipótesis puede integrarse en la estructura del problema general de decisión de la siguiente forma: existen dos acciones finales posibles, a_1 y a_2. La acción apropiada a tomar depende del valor del parámetro desconocido q, llamado algunas veces estado de la naturaleza, que es un elemento del espacio paramétrico Ω. El conjunto Ω puede descomponerse en dos conjuntos, v_1 y v_2, tales que se elige la acción a_1 si q pertenece a v_1, y la acción a_2 si q pertenece a v_2. La pérdida asociada a la acción a y al estado de la naturaleza q viene dada por $l(a;q)$, donde $l(a;q) \geq 0$ y

$$\begin{aligned} l(a_1;q) &= 0 \quad \text{Si } q \text{ está en } v_1 \\ l(a_2;q) &= 0 \quad \text{Si } q \text{ está en } v_2 \end{aligned} \qquad (1)$$

Sea $s = (x_1, x_2, ..., x_n)$ una muestra aleatoria procedente de $f(x;q)$, y S, el espacio muestral n-dimensional. Una estrategia (función de decisión) es una función d que asigna a cada posible muestra una acción de A, donde

$$A = \{a : a = a_1 \text{ o } a_2\}.$$

La acción que se toma es

$$a = d(x_1, x_2, ..., x_n)$$

En este problema en el que existen sólo dos acciones, cada estrategia d (función de decisión) puede representarse por una partición del espacio muestral n-dimensional en dos conjuntos disjuntos, S_1 y S_2, siendo

$$S_2 = \overline{S}_1 = S - S_1$$

tales que se toma la acción a_1 si el punto muestral s cae en S_1, y la a_2 si s cae en S_2. El riesgo (pérdida esperada) correspondiente a la estrategia d está dado por

$$R(d;\theta) = l(a_1;\theta)P(s \in S_1|\theta) + l(a_2;\theta)P(s \in S_2|\theta) \qquad (2)$$

donde $P(s \in S_1|\theta)$ denota la probabilidad de que el punto muestral s caiga en S_1 cuando el valor del parámetro (estado de la naturaleza) es θ, y análogamente para $P(s \in S_2|\theta)$.

Puesto que se toma la acción a_1 si s cae en S_1 y la a_2 si cae en S_2, las probabilidades en la ecuación anterior son las correspondientes a adoptar las acciones a_1 y a_2, respectivamente, cuando θ es el estado de la naturaleza. Se denominan probabilidades de acción.

DEF Sea S un espacio muestral n-dimensional, y S_1 y S_2, una partición del espacio muestral, tal que si un punto muestral

$$s = (x_1, x_2, ..., x_n)$$

cae en S_1, se toma la acción a_1, y si s cae en S_2 se adopta la acción a_2. Las siguientes probabilidades se denominan probabilidades de acción:

$$P(s \in S_1|\theta) \qquad P(s \in S_2|\theta)$$

donde $P(s \in S_i|\theta)$ es la probabilidad de que s caiga en S_i (probabilidad de que se tome la acción a_i) cuando el verdadero estado de la naturaleza es θ.

Si en la ecuación (2) calculamos el riesgo cuando θ pertenece a ω_1, el cual designaremos por $R(d;\theta \in \omega_1)$, se obtiene:

$$R(d;\theta \in \omega_1) = l(a_1;\theta \in \omega_1)P(s \in S_1|\theta \in \omega_1) + l(a_2;\theta \in \omega_1)P(s \in S_2|\theta \in \omega_1) \qquad (3)$$

Utilizando la ecuación (1), resulta

$$R(d;\theta \in \omega_1) = l(a_2;\theta \in \omega_1)P(s \in S_2|\theta \in \omega_1) \qquad (4)$$

Por un procedimiento análogo, calcularemos el riesgo cuando θ está en ω_2, obteniendo

$$R(d;\theta \in \omega_2) = l(a_1;\theta \in \omega_2)P(s \in S_1|\theta \in \omega_2) \qquad (5)$$

Es decir, puesto que una de las dos pérdidas $l(a_1;\theta)$ y $l(a_2;\theta)$ es igual a 0, escribiremos el riesgo en la ecuación (2) en la forma

$$R(d;\theta) = l(\theta)\varepsilon(d;\theta) \qquad (6)$$

donde

$$
\begin{aligned}
&l(q) = l(a_1; q) = 0 \quad \text{Si } q \text{ está en } v_2 \\
&l(a_2; q) = 0 \quad \text{Si } q \text{ está en } v_1
\end{aligned}
\qquad (7)
$$

siendo $l(q)$ la perdida asociada con la acción incorrecta cuando el estado de la naturaleza es q, y $e(d;q)$ en la ecuación (6), es la probabilidad de error definida a continuación.

DEF *Probabilidades de error.*- La probabilidad de error, designada por $e(d;q)$ en la ecuación (6), es la probabilidad de adoptar la acción incorrecta. Es decir, es la probabilidad de tomar la acción a_1 si q está en v_2, o bien tomar la acción a_2 si q está en v_1.

Si $q \in v_1$, esta probabilidad se expresará así:

$$e_1(d;q) = P\left[(x_1,...,x_n) \in S_2 \middle| q \in v_1\right] = P(s \in S_2 | q \in v_1)$$

que es la correspondiente a tomar la acción a_2 erróneamente; y si $q \in v_2$, la probabilidad de error puede escribirse:

$$e_2(d;q) = P\left[(x_1,...,x_n) \in S_1 \middle| q \in v_2\right] = P(s \in S_1 | q \in v_2)$$

que es la probabilidad de adoptar la acción a_1 erróneamente.

DEF *Contrastes de hipótesis.*- Los conjuntos v_1 y v_2 en la formulación anterior del problema de decisión pueden asociarse a la hipótesis o afirmación H_1: "q está en v_1" y a la hipótesis alternativa H_2: "q está en v_2"**,** respectivamente. La acción a_1 consiste en aceptar la hipótesis (aceptar H_1) y la acción a_2 en rechazar la hipótesis (rechazar H_1). La función de decisión d que, aplicada a los datos, conduce a la aceptación o rechazo de la hipótesis se denomina contraste de la hipótesis.

El objetivo es encontrar el contraste (la función de decisión d) que hace mínimo el riesgo para cada valor de q en Ω. Sin embargo, esto no es generalmente posible, sino que una función de decisión puede dar un riesgo mínimo para ciertos valores de q, mientras que otra función de decisión puede hacer mínimo el riesgo para otros valores de q, etc. Por tanto, puesto que q es desconocido, hay que contar con la posibilidad de que no exista un método definido para determinar qué función da riesgo mínimo en un problema particular.

Otra dificultad inherente a la utilización de las ecuaciones (4) y (5) se debe a que en gran parte de los problemas de aplicación, donde un experimentador desea utilizar contrastes de hipótesis, la función de pérdida es totalmente desconocida, o bien no se conoce con la función acuracidad para garantizar su empleo. Si la función de pérdida no es conocida, parece que un procedimiento razonable consistirá en utilizar una función de decisión que, en cierto sentido minimice las probabilidades de error. El procedimiento tradicional es elegir una probabilidad a, usualmente en el entorno de 0,01, 0,05, 0,10, 0,20 y hallar la clase de funciones de decisión (o sea, determinar los conjuntos S_2) tales que se satisfaga

$$P(s \in S_2 | \theta \in \omega_1) \leq \alpha \qquad (8)$$

Entonces, de la clase de contrastes que satisfacen a (8) se considera como “mejor” contraste aquel para la cual

$$P(s \in S_1 | \theta \in \omega_2) \qquad (9)$$

es mínimo. En esta formulación, la cantidad $P(s \in S_2 | \theta \in \omega_1)$ de (8) se llama probabilidad de rechazar una hipótesis verdadera (rechazar la hipótesis H_1 cuando de hecho es cierta), y a veces se la denomina probabilidad de un error de tipo I, y (8) se escribe en la forma $P(I) \leq \alpha$. La cantidad $P(s \in S_1 | \theta \in \omega_2)$ de (9) se llama probabilidad de aceptar una hipótesis falsa (aceptar H_1 cuando no es cierta), pero algunas veces se denomina también probabilidad de un error de tipo II, se escribe P(II). Obsérvese que

$$\varepsilon_1(d;\theta) = P(I) \qquad \text{y} \qquad \varepsilon_2(d;\theta) = P(II)$$

La región S_2 recibe el nombre de región de rechazo o de región crítica, y S_1, región de aceptación. Si la afirmación de (8) es verdadera, se dice que la extensión de el contraste es α. En lugar de la cantidad $P(s \in S_1 | \theta \in \omega_2)$ de (9) es a menudo más conveniente utilizar $P(s \in S_2 | \theta \in \omega_2)$, donde, evidentemente,

$$1 - P(s \in S_1 | \theta \in \omega_2) = P(s \in S_2 | \theta \in \omega_2) \qquad (10)$$

que es la probabilidad de rechazar la hipótesis H_1 cuando de hecho es falsa. La cantidad $P(s \in S_2 | \theta)$ se denomina potencia de el contraste, designándose por $\beta(\theta)$, y es función de θ. Obsérvese que $\beta(\theta)$ = P(I) cuando $\theta \in \omega_1$. También $\beta(\theta)$ = 1 – P(II) si $\theta \in \omega_2$.

A primera vista puede parecer que esta formulación del problema de los contrastes de hipótesis no tiene en cuenta la función de pérdida. En realidad, no prescinde de ella completamente, puesto que llegar a un valor razonable para α requiere que el experimentador sopese las consecuencias de cometer errores de los tipos I y II. La anterior formulación del problema ha recibido una atención preferente por parte de los estadísticos matemáticos y se emplea extensamente por los experimentadores.

10. CONTRASTE DE UNA HIPÓTESIS SIMPLE CONTRA UNA ALTERNATIVA SIMPLE.

Una hipótesis $H : \theta \in \omega$ se llama simple si ω contiene un punto único. Así, si ω_1 consta del punto θ_1 y si ω_2 es el punto θ_2, el problema se denomina contrastar una hipótesis simple contra una alternativa simple.

Aquí la función de riesgo para una estrategia d toma dos valores

$$R(d;\theta_1) = l(\theta_1)P(I) \qquad \text{y} \qquad R(d;\theta_2) = l(\theta_2)P(II);$$

por tanto, para cada función de decisión d, el riesgo $R(d;q_2)$ puede representarse por un punto en un gráfico cuyas coordenadas sean $R(d;q_1)$ y $R(d;q_2)$. Análogamente, $e(d;q)$ podrá representarse en un gráfico cuyas coordenadas son las probabilidades de error P(I) y P(II). Este último gráfico no implica la función de pérdida y es útil en aquellas aplicaciones donde esta función no se conoce perfectamente y P(I) y P(II) pueden utilizarse como se explicó en la sección anterior.

DEF Una estrategia (función de decisión o contraste) d es admisible si no existe otra estrategia d* tal que

$R(d^*;q) \leq R(d;q)$ para todo q de Ω

y

$R(d^*;q) < R(d;q)$ para algún q de Ω

Como se indicó anteriormente, no hay en general, una función de decisión que dé riesgo mínimo para todos los valores de q en Ω; por tanto, se comprende que lo más razonable consiste en hallar la clase de las funciones de decisión admisibles y seleccionar una de ellas.

Para ayudar a encontrar la clase de estrategias admisibles, probaremos que toda estrategia admisible es una estrategia de Bayes, y que toda estrategia de Bayes es una contraste de la razón de verosimilitud. Por tanto, toda estrategia admisible es un contraste de la razón de verosimilitud. En consecuencia, si es posible hallar la clase de contrastes de la razón de verosimilitud, está incluirá todas las estrategias admisibles; la obtención de la clase de contrastes de la razón de verosimilitud es, frecuentemente, bastante fácil dedicaremos el resto de esta sección al desarrollo de estas ideas. Recordemos que nos limitamos a considerar una hipótesis simple y una alternativa simple.

DEF *Estrategia de Bayes.*- Una estrategia d es una estrategia de Bayes correspondiente a probabilidades "a priori h_1 y $h_2 = 1 - h_1$ $(h_i \geq 0)$ si hace mínimo B(d), donde

$$B(d) = E\left[R(d;q)\right] = h_1 R(d;q_1) + h_2 R(d;q_2)$$

Esbozaremos la demostración con el siguiente teorema.

TEOREMA

Para contrastar una hipótesis simple contra una alternativa simple, toda estrategia admisible es una estrategia de Bayes.

<u>Dem.</u>

En primer lugar, observamos que la estrategia de Bayes correspondiente a h_1 y h_2 puede representarse geométricamente dibujando la recta

$$h_1 R_1 + h_2 R_2 = c$$

y desplazándola mediante la variación de c, paralelamente a sí misma, hasta que toque a T. El punto (o puntos) donde toca a T corresponde a la estrategia de Bayes. Como h_1 varía desde 0 hasta 1, la pendiente de la recta lo hace, desde 0 hasta $-\infty$. Una propiedad de los conjuntos convexos es que, dado cualquier punto del contorno, existe una recta que pasa por ese punto en la que se apoya el conjunto. Luego para toda estrategia admisible, es decir, para cualquier punto de contorno inferior a T, existe una recta de apoyo que pasa por dicho punto. Por tanto, puede trazarse esta recta con pendiente no positiva, y expresarse en la forma

$$h_1 R_1 + h_2 R_2 = c$$

donde h_1 y h_2 son probabilidades posibles a priori (o sea, $0 \le h_i \le 1$). Por consiguiente, la estrategia admisible es una estrategia de Bayes.

El caso especial de contrastar una hipótesis simple contra una alternativa simple nos lleva a un resultado interesante; es decir, toda estrategia de Bayes es un contraste de la razón de verosimilitud.

DEF *Contraste de la razón de verosimilitud.*- Un contraste basada en una muestra aleatoria $x_1,\dots,x_n$ de la densidad $f(x;\theta)$ para contrastar $H_1: \theta = \theta_1$ contra $H_2: \theta = \theta_2$ es un contraste de la razón de verosimilitud, si existe un número k tal que el contraste permite

Aceptar H_1	(acción a_1)	si $\lambda > k$
Rechazar H_1	(acción a_2)	si $\lambda < k$

y

una de las dos acciones si $\lambda = k$

donde λ es la razón de verosimilitud dada por

$$\lambda = t(x_1,\dots,x_n) = \frac{f(x_1;\theta_1)\, f(x_2;\theta_1) \dots f(x_n;\theta_1)}{f(x_1;\theta_2)\, f(x_2;\theta_2) \dots f(x_n;\theta_2)} \qquad (1)$$

TEOREMA

Para contrastara la hipótesis simple $H_1: \theta = \theta_1$ contra la alternativa simple $H_2: \theta = \theta_2$, toda estrategia de Bayes es un contraste de la razón de verosimilitud.

Cabe interpretar que la razón de verosimilitud λ es una medida de cómo la evidencia confirma H_1. Así, es razonable aceptar H_1 cuando λ es suficientemente grande. Obsérvese que el ser "suficientemente grande" puede depender de factores tales como las pérdidas debidas al error y el grado de confianza previa, si la hay, en la hipótesis.

11. HIPÓTESIS COMPUESTAS.

En la práctica, la mayor parte de los problemas de contrastes implican hipótesis compuestas. Estas hipótesis son de la forma $H_1 : \theta \in \omega_1$, con la alternativa $H_2 : \theta \in \omega_2$, en donde ω_1 y/o ω_2 contienen más de un elemento.

En los contrastes de hipótesis compuesta la situación resulta mucho más compleja que cuando las hipótesis son simples. En el caso compuesto, los contrastes admisibles son difíciles o imposibles de obtener. En este caso, nos contentaremos, en general, con un análisis de las probabilidades de error P(I) y P(II), e intentaremos hallar contrastes que de cierta manera las controlen.

TEOREMA

La región crítica R_k de extensión α que hace máxima la potencia de el contraste de la hipótesis $H_1 : \theta = \theta_1$, contra la alternativa $H_2 : \theta = \theta_2$ donde $x_1, \ldots, x_n$ es una muestra aleatoria de tamaño n de $f(x;\theta)$, se obtiene hallando la región R_k (si existe) que satisface a

$$\lambda = t(x_1,\ldots,x_n) = \frac{f(x_1;\theta_1)f(x_2;\theta_1)\ldots f(x_n;\theta_1)}{f(x_1;\theta_2)f(x_2;\theta_2)\ldots f(x_n;\theta_2)} < k \qquad (2)$$

para un número fijo k y tal que

$$\int\int\underset{R_k}{\ldots}\int f(x_1;\theta_1)f(x_2;\theta_1)\ldots f(x_n;\theta_1)dx_1dx_2\ldots dx_n = \alpha \qquad (3)$$

Esto, evidentemente, constituye una aplicación de la razón de verosimilitud.

A primera vista no parece claro cómo (3) implica k, pero la región en que se verifica (2) cambia al variar k, y cuando esto ocurre puede haber una región (un valor de k) que satisface a (3). Es importante insistir en que este teorema proporciona una región crítica más potente (de extensión α) para contrastar solo que θ es también un único punto. El teorema no da necesariamente un método para hallar una región crítica más potente de extensión α cuando ω_1 o ω_2 contienen más de un punto. Veremos más adelante que algunas veces puede utilizarse en tales situaciones, que, evidentemente, son los casos más útiles. Es decir, un experimentador puede desear contrastar que la diferencia de rendimientos medios de dos variedades de trigo es cero contra la alternativa de que es positiva. O un fabricante deseará quizá contrastar la hipótesis $H_1 : \mu \leq 0$ contra la alternativa $H_2 : \mu > 0$, donde μ es la diferencia de eficacia media de dos medicamentos. En estos casos, ω_1 o ω_2 (o ambos) contienen más de un punto. Existen cuatro casos distintos:

1) ω_1 contiene un punto y ω_2 contiene un punto.
2) ω_1 contiene un punto y ω_2 contiene más de un punto.
3) ω_1 contiene más de un punto y ω_2 contiene un punto.
4) ω_1 y ω_2 contienen más de un punto.

En general, el lema de Neyman-Pearson se aplica únicamente al caso 1, pero veremos que algunas veces es también útil en otros casos.

12. CONTRASTE DE θ θ_1 CONTRA $\theta>\theta_1$ PARA DENSIDADES CON UN PARÉMETRO ÚNICO θ.

En estadística aplicada existen muchas densidades que contienen un parámetro único desconocido, tales como la binomial, la de Poisson, la normal de media conocida, la normal de varianza conocida, la exponencial, etc. Muchas veces un experimentador desea contrastar la hipótesis $H_1:\theta \leq \theta_1$ con la hipótesis alternativa $H_2:\theta > \theta_1$, siendo θ_1 conocido donde la densidad es $f(x;\theta)$.

DEF Un contraste de la hipótesis $H_1:\theta \in \omega_1$, contra la alternativa $H_2:\theta \in \omega_2$, se dice que es un contraste UMP de extensión α si su región critica R es tal que

$$\begin{aligned} &P(I) \leq \alpha && \text{para todo } \theta \text{ de } \omega_1 \\ &\beta(\theta) = 1 - P(II) \quad \text{es máximo} && \text{para cada } \theta \text{ de } \omega_2 \end{aligned} \qquad (1)$$

En la formulación de los contrastes de hipótesis dada en (12-3-1), un contraste UMP es la "mejor" contraste.

A continuación daremos un teorema bastante útil para determinar un contraste UMP de $H_1:\theta \leq \theta_1$ contra la hipótesis alternativa $H_2:\theta > \theta_1$.

TEOREMA

Sea x=(x_1, …, x_n) una muestra aleatoria de una densidad con un único parámetro θ en un intervalo Ω, y sea $f(x;\theta)$ la densidad conjunta de las variables aleatorias. Supongamos que $f(x;\theta)$ puede escribirse así:

$$f(x;\;) = s(\theta)U(x)e^{v(x)t(\theta)} \qquad (2)$$

donde t(θ) es una función estrictamente creciente de θ en Ω. Si existe una constante c tal que $P[v(x) > c|\theta_1] = \alpha$ para un α dado y comprendido entre 0 y 1, R es entonces una región crítica UMP de extensión α para contrastar $H_1:\theta \leq \theta_1$ contra $H_2:\theta \leq \theta_1$, donde $R = \{x : v(x) > c\}$. Si t(θ) es una función estrictamente decreciente de θ en Ω y si existe una constante c tal que $P[v(x) < c|\theta_1] = \alpha$ para un α dado y comprendido entre 0 y 1, R es una región crítica UMP de extensión α para contrastar $H_1:\theta \leq \theta_1$ contra $H_2:\theta \leq \theta_1$, donde $R = \{x : v(x) < c\}$.

BIBLIOGRAFÍA RECOMENDADA.

Estadística Teórica. Aut. J.M.Doblado y M.C. Nieto. Edit. UNED

Introducción a la Estadística Teórica. Aut.: G Arnáiz. Edit.: Lex Nova

Estadística Teórica y Aplicada. Aut.: A. Nortes. Edit.: S. Rodríguez.

Introducción a la Probabilidad y la Medida (I). Aut.: P Zoroa y N. Zoroa. Edit.: Maior DM.