ESTIMACIÓN DE VARIANZAS

Si bien la varianza muestral s es un estimador insesgado de σ^2, esto no implica que la desviación estándar muestral sea un estimador insesgado de σ. En realidad no lo es. Sin embargo, para muestras de gran tamaño el sesgo es pequeño y se acostumbra a estimar σ con s.

Se llama Rango Muestral R, al valor más grande de una muestra menos el más pequeño. Dada una muestra de tamaño n de una población normal, puede verificarse que la distribución muestral del rango tiene la media $d_2.\sigma$ y la desviación estándar $d_3.\sigma$. donde d_1 y d_2 son constantes que dependen del tamaño de la muestra.

Se verifica que:

d_2	1.128	1.693	2.054	2.326	2.534	2.704	2.847	2.970	3.078
d_3	0.853	0.888	0.880	0.864	0.848	0.833	0.830	0.808	0.797
N	2	3	4	5	6	7	8	9	10

Esto se puede comprobar mediante el siguiente segmento de programa en Matlab, donde se genera una población normal estándar (media cero, varianza 1) seleccionando muestras de tamaño n a las cuales se les determina el rango.

Hallando la media de los rangos, se obtiene un número próximo a d_2 y halando la desviación estándar se obtiene un número próximo a d_3.

```
function estima_var(n,m)
% Verificacion de las propiedades del Rango Muestral
% poblacion con distribucion normal estandar
% Entrada: n, entero, numero de elementos de cada muestra
%          m, entero, numero de muestras para la prueba
% Salida: M, real, media de los rangos
%         S, real, desviacion estandar de los rangos

% Generacion de la muestra de varianzas
for j=1:m,
   for k=1:n,
     % Generacion de la muestra normal estandar
     s=0;
     for i=1:12, s=s+rand;end
     T(k)=s-6;
  End
  M(j)=max(T)-min(T);
End
mean(M)
std(M)
```

como se puede de la corrida:

```
>> estima_var(5,1000)
ans =
   2.3180
ans =
   0.8506
```

Luego R/d_2 es un estimador insesgado de σ, y para muestras muy pequeñas

($n \leq 5$), proporciona una estimación de σ tan buena como s. Lo contrario sucede cuando se incrementa el tamaño muestral.

El rango se emplea fundamentalmente en Control de Calidad.

Ejemplo: Las siguiente muestra aleatoria son mediciones de la capacidad de producción de calor (en millones de calorías por tonelada) de especímenes de carbón de una mina:

8260 8130 8350 8070 8340

usar el rango de la muestra para estima σ de la capacidad de producción de calor del carbón de la mina.

Dado que n = 4, $d_2 = 2.326$ y el rango vale: R = 8350-8070= 280

Luego la estimación de σ queda:

$$\sigma = \frac{R}{d_2} = \frac{280}{2.326} = 120.378$$

mientras que calculado directamente da:

$$\bar{x} = \frac{8260 + 8130 + 8350 + 8070 + 8340}{5} = 8230$$

$$s = \sqrt{\frac{8260^2 + 8130^2 + 8350^2 + 8070^2 + 8340^2 - 5\cdot 8230^2}{4}} = 125.499$$

En la mayoría de las aplicaciones prácticas, la estimación de intervalos para σ o σ^2 se basan en la desviación estándar muestral o en la varianza muestral. En muestras aleatorias de poblaciones normales:

$$\chi^2 = \frac{(n-1)\cdot s^2}{\sigma^2} \qquad \text{con } \nu = \text{n-1 grados de liberta}$$

definiendo a χ_α como la abscisa que deja a la izquierda un área α para una distribución chi-cuadrado con n-1 grados de libertad.

Con esto se puede asegurar con probabilidad 1-α que se satisface:

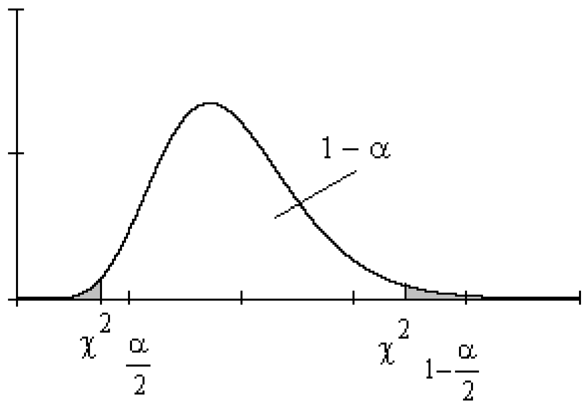
$$\left(\chi^2\right)_{\frac{\alpha}{2}} < \frac{(n-1)\cdot s^2}{\sigma^2} < \left(\chi^2\right)_{1-\frac{\alpha}{2}}$$

de aquí operando matemáticamente:

$$\frac{(n-1)\cdot s^2}{\left(\chi^2\right)_{1-\frac{\alpha}{2}}} < \sigma^2 < \frac{(n-1)\cdot s^2}{\left(\chi^2\right)_{\frac{\alpha}{2}}}$$

se determina el intervalo de confianza (1- α).100% para la varianza poblacional.

Gráficamente:



Ejemplo: La desviación estándar de las duraciones de una muestra de 20 lámpara eléctricas fue de 100 horas. Hallar los límites de confianza del 95% para la desviación estándar de dichas lámparas.

ν = n-1 = 20-1 = 19 grados de libertad

$\alpha/2 = 0.025$ $\chi_{\alpha/2} = 8.907$ $\chi_{1-\alpha/2} = 32.852$

el intervalo de confianza para la varianza será:

$$\frac{(20-1)\cdot 100^2}{32.852} < \sigma^2 < \frac{(20-1)\cdot 100^2}{8.907}$$

y para la desviación estándar:

$$74.02 < \sigma < 146.048$$

Lo que significa que hay un 95% en el nivel de confianza que la desviación estándar estará entre 74.02 y 146.048.

Este método se aplica a muestras aleatorias de poblaciones normales. Si el tamaño de la muestra es grande, la distribución muestral de la desviación estándar puede aproximarse a una distribución normal con media σ y desviación estándar $\sigma/\sqrt{2\cdot n}$. Luego:

$$z = \frac{s-\sigma}{\frac{\sigma}{\sqrt{2\cdot n}}}$$

lo que supone una desigualdad, para un intervalo de confianza (1-α).100%, del tipo:

$$-z_{\frac{\alpha}{2}} < \frac{s-\sigma}{\frac{\sigma}{\sqrt{2\cdot n}}} < z_{\frac{\alpha}{2}}$$

lo que lleva a:

$$\frac{s}{1+\dfrac{z_{\frac{\alpha}{2}}}{\sqrt{2\cdot n}}} < \sigma < \frac{s}{1-\dfrac{z_{\frac{\alpha}{2}}}{\sqrt{2\cdot n}}}$$

que es un intervalo de confianza para σ en muestras de gran tamaño.

Ejemplo: La desviación estándar de la resistencia a la rotura de 100 cables producidos por una compañía fue de 180 libras. Hallar los límites de confianza del 95% para la desviación estándar de todos los cables producidos por la compañía.

$\alpha/2 = 0.025 \quad z_{\alpha/2} = 1.96$

$$\frac{180}{1+\dfrac{1.96}{\sqrt{2\cdot 100}}} < \sigma < \frac{180}{1-\dfrac{1.96}{\sqrt{2\cdot 100}}} \qquad 158.09 < \sigma < 208.96$$

HIPÓTESIS REFERIDA A UNA VARIANZA

Se considera el problema de probar la hipótesis nula de que la varianza de una población es igual a una determinada, contra una alternativa unilateral o bilateral adecuada.

Para muestras aleatorias extraídas de una población normal con varianza σ^2:

$$\chi^2 = \frac{(n-1)\cdot s^2}{(\sigma_0)^2}$$

con lo que se pueden resumir las regiones críticas para probar $\sigma^2 = \sigma_0^2$ (para poblaciones normales) con el siguiente cuadro:

Hipótesis Alterna	Se rechaza la Hipótesis Nula si
$\sigma^2 < \sigma_0^2$	$\chi^2 < \chi_0{}^2{}_{1-\alpha}$
$\sigma^2 > \sigma_0^2$	$\chi^2 < \chi_0{}^2{}_{\alpha}$
$\sigma^2 <> \sigma_0^2$	$\chi^2 < \chi_0{}^2{}_{1-\alpha/2}$ $\chi^2 < \chi_0{}^2{}_{\alpha/2}$

Ejemplo: El proceso de bruñido se usa para esmerilar ciertos discos de silicio al grueso apropiado es aceptable sólo si σ (desviación estándar de la población de grosores de los cubitos cortados de dichos discos) es a lo sumo 0.50 mil. Emplear el nivel de significación de 0.05 para probar la hipótesis nula $\sigma = 0.50$ contra la hipótesis alterna $\sigma > 0.5$ si el grosor de 15 cubitos cortados de tales discos tienen una desviación estándar de 0.64 mil.

1 – Hipótesis Nula $\sigma = 0.50$
Hipótesis Alternativa $\sigma > 0.50$ (unilateral)

2 - Nivel de significancia: α =0.05. ν=14 $\chi^2_\alpha = \chi^2_{0.05} = 23.685$

3 - Para trabajar con tablas normalizadas

$$\chi^2 = \frac{(n-1)\cdot s^2}{(\sigma_0)^2}$$

4 - Cálculos:

$$\chi^2 = \frac{(15-1)\cdot 0.64^2}{0.5^2} = 22.938$$

5 - Dado que $22.938 < \chi^2_{0.05}$ se Acepta la Hipótesis Nula σ = 0.5. Vale decir, el grosor de los cubitos tienen una desviación estándar significativamente igual a 0.5 mil.

Si la población que se muestrea tiene un tamaño relativo grande ($n \geq 30$) la hipótesis nula se puede probar con:

$$z = \frac{s - \sigma_0}{\frac{\sigma_0}{\sqrt{2\cdot n}}}$$

ESTIMACIÓN DE PROPORCIONES

La información que suele disponerse al estimar la proporción es el número de veces x, que un evento ocurre en n ensayos, ocasiones u observaciones. La estimación puntual suele ser x/n (proporción muestral). Para n ensayos que satisfacen la distribución binomial, se verifica $\mu = n.p$ y $\sigma = \sqrt{n\cdot p\cdot(1-p)}$, si se divide por n, se encuentra la media y la desviación estándar de la proporción de éxitos (es decir la proporción muestral):

$$\frac{\mu}{n} = n\cdot\frac{p}{n} = p \qquad \text{y} \qquad \frac{\sigma}{n} = \frac{\sqrt{n\cdot p\cdot(1-p)}}{n} = \sqrt{\frac{p\cdot(1-p)}{n}}$$

La proporción muestral es un estimador insegado del parámetro binomial p (proporción real que se desea estimar a partir de la muestra).

Ya que x y x/n son variables discretas, un intervalo de confianza de exactamente (1- α).100% es difícil de lograr y para hallar σ/n hace falta conocer p.

Para construir un intervalo de confianza para p que tenga aproximadamente un nivel de confianza (1- α).100%, se deben determinar x_0 y x_1 para un conjunto determinado de valores de p.

Donde:

$$\sum_{k=0}^{x_0} b(k,n,p) \leq \frac{\alpha}{2}$$

siendo x_0 el máximo entero para el que se verifica la desigualdad. Y:

$$\sum_{k=x_1}^{n} b(k,n,p) \le \frac{\alpha}{2}$$

siendo x_1 el <u>mínimo</u> entero para el que se verifica la desigualdad.

En este punto se habrá de hacer una revisión de conceptos de la Distribución Binomial. Dicha distribución es de características Discretas, aplicable cada vez que un proceso de muestreo conforma un proceso de Bernoulli:

1- Hay dos resultados posibles mutuamente excluyentes en cada ensayo. Para mayor conveniencia estos se denominan éxito y fracaso.
2- La serie de ensayos u observaciones constituyen eventos independientes.
3- La probabilidad de éxito, designada por p, permanece constante de ensayo a ensayo. Es decir el proceso es estacionario.

Para el cálculo se requieren tres valores: el número designado de éxitos (x); el número de ensayos (n) y la probabilidad de éxito en cada ensayo (p):

$$b(x,n,p) = C_{n,x} \cdot p^x \cdot (1-p)^{n-x} = \frac{n!}{x! \cdot (n-x)!} \cdot \left[p^x \cdot (1-p)^{n-x}\right]$$

Con la tablas se utiliza el valor acumulado, por ejemplo:

$$B(3,5,0.2) = b(0,5,0.2) + b(1,5,0.2) + b(2,5,0.2) + b(3,5,0.2)$$
$$B(3,5,0.2) = 0.99328$$

En general:

$$B(x,n,p) = \sum_{k=0}^{x} C_{n,x} \cdot p^k \cdot (1-p)^{n-k}$$

si se quiere calcular b(3,5,0.2) se lo hace así:

b(3,5,0.2) = b(3,5,0.2) - b(2,5,0.2) = 0.9933 – 0.9421 = 0.0512

Para calcular con la tabla valores de $p > 0.5$, se aplica la siguiente propiedad:

B(x,n,p) = 1 - B(n – x –1 , n, 1 - p)

Retomando el tema principal, se puede asegurar con probabilidad de aproximadamente (1- α).100% y al menos (1- α).100% que se cumple la siguiente desigualdad:

$$x_0(p) < x < x_1(p)$$

Para transformar esta desigualdad en un intervalo de confianza para p, se recurre a un método gráfico.

Ejemplo: Se desean hallar intervalos con un nivel de confianza del 95%, aproximadamente, para p en muestras de tamaño n=20.

Por medio de las tablas se calculan x_0 y x_1 para distintos valores de p, tales que x_0 sea el máximo entero de modo que:

$B(x_0, 20, p) \leq 0.025$ (para $p<0.5$)
ó $B(n-x_0-1, 20,1- p) \geq 0.975$ (para $p >0.5$)
ó $B(n-x_0, 20,1- p) \leq 0.025$ (para $p >0.5$)

mientras que x_1 sea el mínimo entero de modo que:

$1 - B(x_1-1, 20, p) \leq 0.025$ (para $p<0.5$)
ó $B(n-x_1, 20,1- p) \leq 0.025$ (para $p >0.5$)

más sencillamente, los valores correspondientes a los p del primer caso, se encuentran restando el valor obtenido de n.

Implementando un programa Matlab como el siguiente:

```
function estima_prop(n,alfa)

% Tabla de valores de x0 para determinacion de intervalo
% de confianza en proporciones
%
% Entrada: n, entero, tamaño de la muestra
%          m, real, nivel de significancia
% Salida: C, juego de valores de x1
%         A, juego de valores de x0
i=1;
for p=0.1:0.1:0.99,
   B=0;
   if p<=0.5,
      k=0;
      while B<alfa/2,
        B=B+factorial(n)*p^k*(1-p)^(n-k)/(factorial(k)*factorial(n-k));
        k=k+1;
     end
     A(i)=k-2; i=i+1;
  end
  if p>0.5,
     k=0;
     while B<alfa/2,
        k1=n-k;p1=1-p;
        B=B+factorial(n)*p1^k1*(1-p1)^(n-k1)/(factorial(k1)*factorial(n-k1));
        k=k+1;
     end
     A(i)=k-2; i=i+1;
  end
end
A
for k=1:i-1,C(k)=n-A(i-k);end
C
k=0.1:0.1:0.9;plot(k,A,k,C)
```

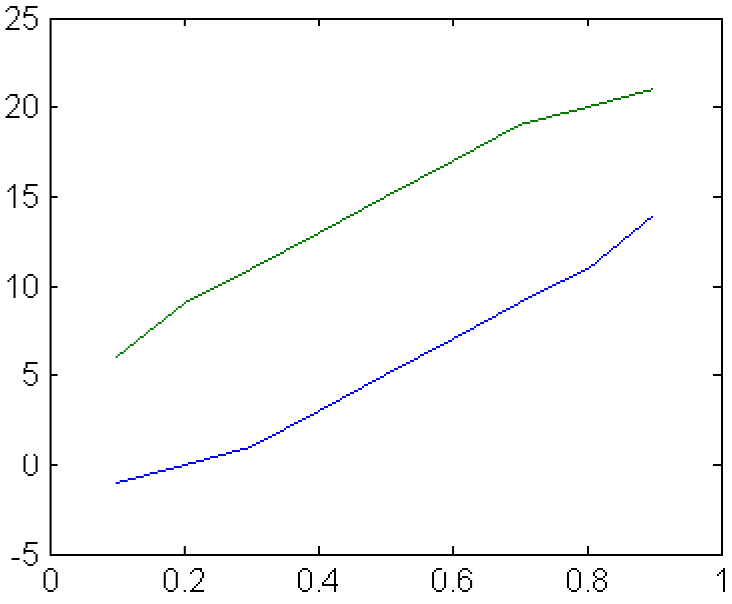
se puede calcular lo siguiente cuando se ejecuta:

```
>> estima_prop(20,0.05)
A =
   -1    0    1    3    5    7    9   11   14
C =
    6    9   11   13   15   17   19   20   21
```

que interpretado como un cuadro de valores queda:

p	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
x_0	-	0	1	3	5	7	9	11	14
x_1	6	9	11	13	15	17	19	20	-

Con guiones se han indicado los valores que no tienen sentido.



Para distintos valores de n se obtienen distintas ramas. Esto se grafica en forma normalizada (Biometrike Tables for Statisciens, Vol. 1, New York).

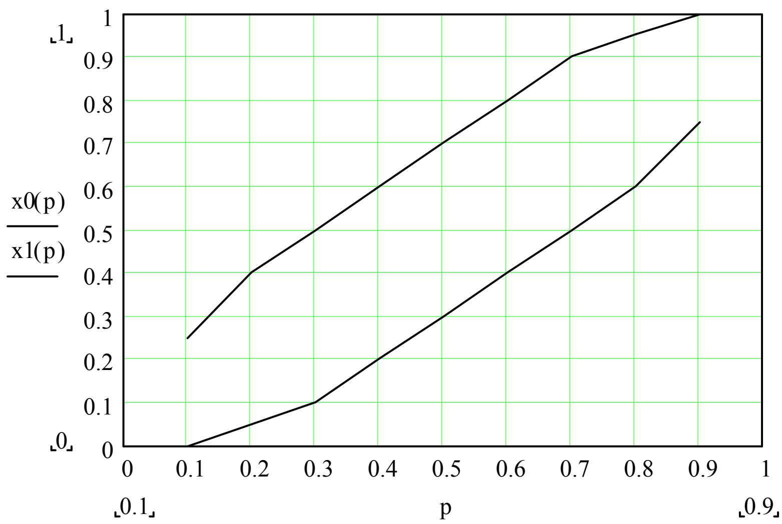
Estas gráficas se pueden obtener muy fácilmente en Mathcad a partir de las facilidades que dispone para el cálculo de distribuciones, en este caso para la binomial.

$n := 20$ Tamaño de la muestra

$\alpha := 0.05$ Nivel de significación de la prueba

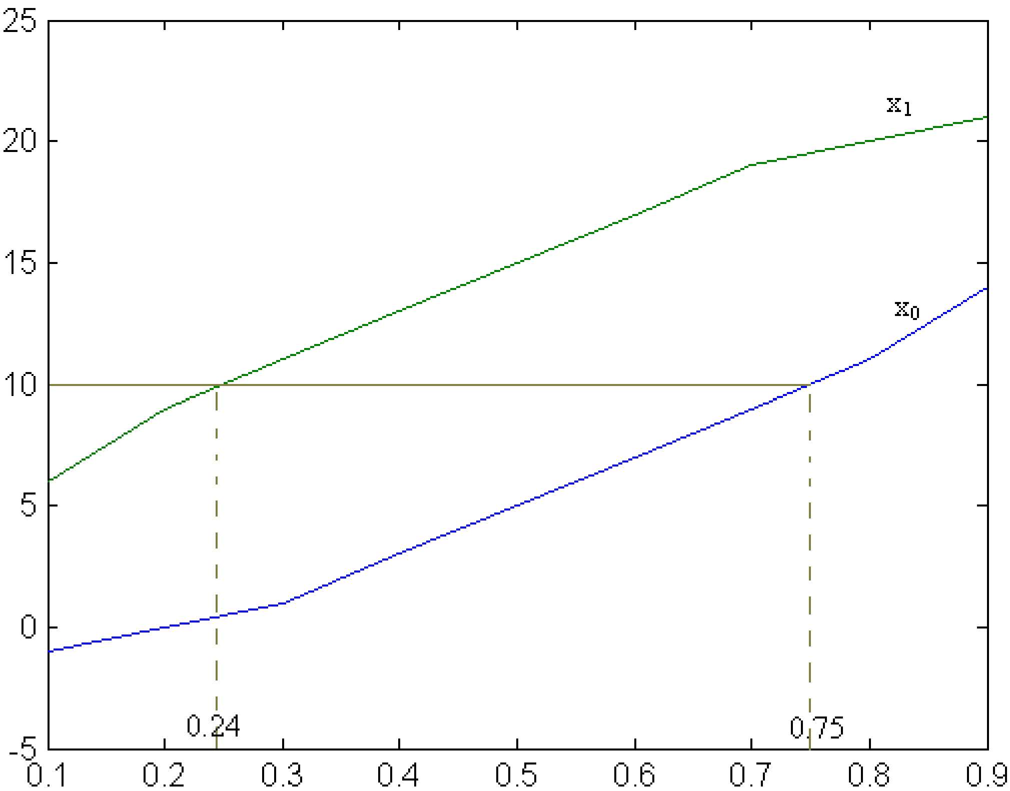
$p := 0.1, 0.2 .. 0.9$ Rango para p

$$x0(p) := \frac{\text{qbinom}\left(\frac{\alpha}{2}, n, p\right)}{n} \qquad x1(p) := \frac{n - \text{qbinom}\left(\frac{\alpha}{2}, n, 1-p\right)}{n}$$



Para un valor dado de x, es posible obtener cuotas de p, con un nivel de significación del 95%, trazando una línea horizontal de una curva a otra y marcando los valores correspondientes.

Para x=10:



Vale decir:

$0.24 < p < 0.75$ con un nivel de significación de 0.05

Hay gráficos de este tipo para niveles de confianza de 95 y 99% para varios valores de n. Se emplea la proporción muestral (x/n) en lugar de x.

Vale decir, para el ejemplo n=20, x=10 y x/n=10/20=0.5. Con estos valores se puede verificar lo correcto de lo realizado.

Como regla práctica para que una distribución binomial se acerque a la normal basta con que n.p y (1-p).p sean a la vez mayores que 5. Por ejemplo, para n=200 se puede usar la aproximación si p está entre 0.025 y 0.975. Esta es la significación de hacer "n suficientemente grande".

Se puede asegurar con una probabilidad 1-α que se cumpla:

$$-z_{\frac{\alpha}{2}} < \frac{x - n\cdot p}{\sqrt{n\cdot p\cdot(1-p)}} < z_{\frac{\alpha}{2}}$$

para evitar cálculos complicados, se hace una aproximación más al sustituir x/n por p en el denominador del segundo miembro:

$$\frac{x}{n} - z_{\frac{\alpha}{2}}\cdot\sqrt{\frac{\frac{x}{n}\cdot\left(1-\frac{x}{n}\right)}{n}} < p < \frac{x}{n} + z_{\frac{\alpha}{2}}\cdot\sqrt{\frac{\frac{x}{n}\cdot\left(1-\frac{x}{n}\right)}{n}}$$

lo que permite lograr este intervalo de confianza para p en muestras de gran tamaño, con un nivel de confianza (1-α).100%.

Ejemplo: si x=36 de n=100 entrevistados están familiarizados con los incentivos en los impuestos que se ofrecen por instalar ciertos dispositivos para ahorrar energía. Construir un intervalo de confianza de 95% para la correspondiente proporción real.

x/p=36/100 $z_{\alpha/2}$=1.96

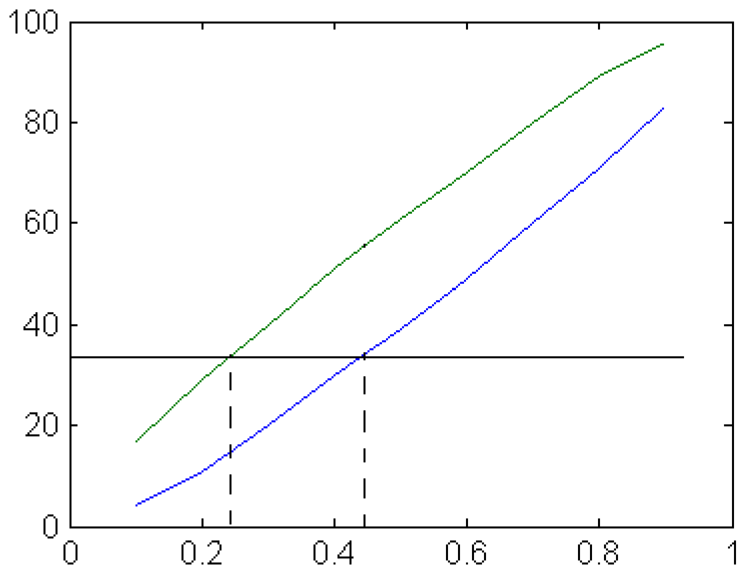
$$1.96\cdot\sqrt{\frac{0.36\cdot(1-0.36)}{100}} = 0.094$$

$$0.36 - 0.094 < p < 0.36 + 0.094 \qquad 0.266 < p < 0.454$$

ejecutando la función Matlab:

```
>> estima_prop(100,0.05)
A =
   4   11   20   30   39   49   60   71   83
C =
  17   29   40   51   61   70   80   89   96
```

Gráficamente:



La magnitud del error cometido cuando se usa x/n como un estimador de p está dada por |x/n – p|. En base a la distribución normal, se puede asegurar con probabilidad 1-α que se cumplirá:

$$\left|\frac{x}{n}\right| - p \leq z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p\cdot(1-p)}{n}}$$

es decir que el error será a lo sumo:

$$z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p\cdot(1-p)}{n}}$$

Con x/n sustituyendo a p, esto produce el Error Máximo de Estimación:

$$E = z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\frac{x}{n}\cdot\left(1-\frac{x}{n}\right)}{n}}$$

Ejemplo: En una encuesta levantada en una gran ciudad, 136 de 400 personas respondieron afirmativamente a la pregunta de si el sistema de transporte público es adecuado. Con una confianza del 99% ¿Qué se puede decir acerca del error máximo si x/n =136/400=0.34 se emplea como una estimación de la correspondiente población real?.

x/p=136/400 $z_{\alpha/2}= z_{0.99}=2.575$

$$E = z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\frac{x}{n}\cdot\left(1-\frac{x}{n}\right)}{n}} = 2.575\cdot\sqrt{\frac{0.34\cdot(1-0.34)}{400}} = 0.061$$

con lo que se puede construir el intervalo:

$$0.34 - 0.061 < p < 0.34 + 0.061 \qquad 0.279 < p < 0.401$$

Para verificar gráficamente, se hace una pequeña modificación al programa de Matlab, ya que en el original se calculan factoriales. Al ser, en este caso, muy grande el valor de n, los números combinatorios se calculan por productos sucesivos.

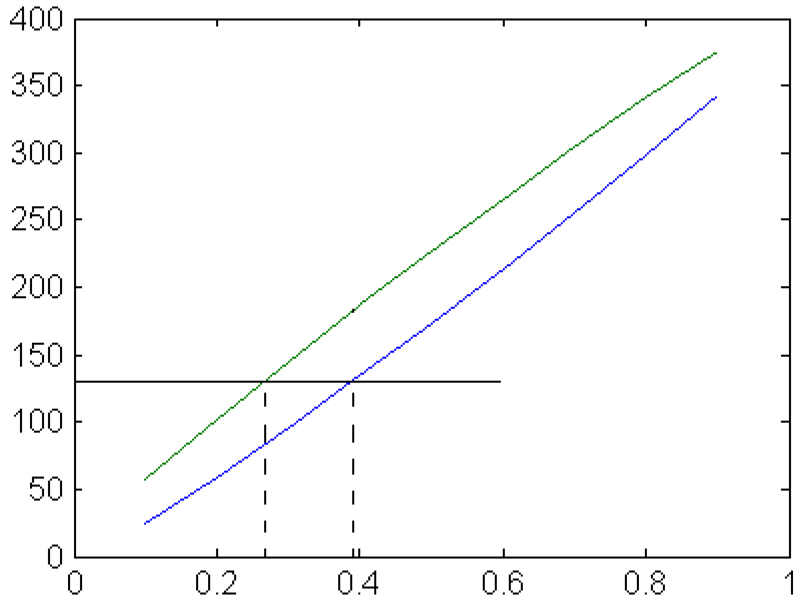
```
function estima_prop1(n,alfa)
% Tabla de valores de x0 para determinacion de intervalo
% de confianza en proporciones
%
% Entrada: n, entero, tamaño de la muestra
%          m, rel, nivel de significancia
% Salida: C, juego de valores de x1
%         A, juego de valores de x0
i=1;
for p=0.1:0.1:0.99,
   B=0;
   if p<=0.5,
      k=0;
      while B<alfa/2,
        n1=n;m1=k;F=1;
        while m1>=1
            F=F*n1/m1;m1=m1-1;n1=n1-1;
        end
        B=B+F*p^k*(1-p)^(n-k); k=k+1;
     end
     A(i)=k-2; i=i+1;
  end
  if p>0.5,
     k=0;
     while B<alfa/2,
        n1=n;m1=k;F=1;
        while m1>=1
            F=F*n1/m1;m1=m1-1;n1=n1-1;
        end
        k1=n-k;p1=1-p;
        B=B+F*p1^k1*(1-p1)^(n-k1);
        k=k+1;
     end
     A(i)=k-2; i=i+1;
  end

end
A
for k=1:i-1,C(k)=n-A(i-k);end
C
k=0.1:0.1:0.9;plot(k,A,k,C)
```

Ejecutando:

```
>> estima_prop1(400,0.01)
A =
   24   59   96   134   173   214   255   298   343

C =
   57   102   145   186   227   266   304   341   376
```



La fórmula anterior de E puede servir para determinar el tamaño muestral que es necesario para alcanzar un grado deseado de precisión.

$$n = p\cdot(1-p)\cdot\left(\frac{z_{\frac{\alpha}{2}}}{E}\right)^2$$

esta fórmula no se puede utilizar a menos que se tenga información de p (en base a datos auxiliares, digamos una muestra previa). Si no se conoce, se sabe que p.(1-p) es a lo sumo ¼, correspondiente a p=1/2. Luego:

$$n = \frac{1}{4}\cdot\left(\frac{z_{\frac{\alpha}{2}}}{E}\right)^2$$

se puede asegurar con una probabilidad de al menos 1-α que el error de servirse de x/n como una estimación de p, no excede E. Una vez obtenidos los datos, se puede asegurar con una confianza de al menos (1-α).100% que el error no sobrepasa E.

Ejemplo: Suponer que se desea estimar la proporción real de unidades defectuosas en un cargamento de ladrillos y que se necesita una confianza de al menos 95% de que el error sea a lo sumo de 0.04 ¿De qué tamaño se necesita la muestra si:

a) No se tiene idea de cómo podría ser la proporción muestral

b) Se sabe que la proporción real no excede de 0.127

a)

$$n = \frac{1}{4}\cdot\left(\frac{z_{\frac{\alpha}{2}}}{E}\right)^2 = \frac{1}{4}\cdot\left(\frac{1.96}{0.04}\right)^2 = 600.25$$

b)

$$n = p\cdot(1-p)\cdot\left(\frac{z_{\frac{\alpha}{2}}}{E}\right)^2 = 0.127\cdot(1-0.127)\cdot\left(\frac{1.96}{0.04}\right)^2 = 266.201$$

Esto ilustra cómo conociendo alguna información auxiliar de p, se puede reducir en gran medida el tamaño de la muestra requerida.

Cuando p es próximo a 0 (alto confiabilidad) y cuando p es la probabilidad de fracaso, se necesitan intervalos de confianza unilaterales para p chico y n grande, la distribución de Poisson se aproxima a la binomial. Se puede mostrar que:

$$p < \frac{1}{2\cdot n}\cdot \chi_\alpha{}^2$$

con ν=2.(x+1) grados de libertad.

Ejemplo: Si hay x=4 fallas en n=2000 partes utilizadas continuamente durante un mes, construir un intervalo unilateral con un nivel de confianza del 99% para la probabilidad de que una de tales partes falle en las condiciones establecidas.

$1-\alpha = 0.99 \qquad \alpha = 0.01 \qquad \nu = 2.(4+1) = 10$

$$p < \frac{1}{2\cdot n}\cdot \chi_\alpha{}^2 = \frac{1}{2\cdot 2000}\cdot 23.209 = 0.058 \qquad\qquad p < 0.058$$

HIPÓTESIS RELATIVA A UNA PROPORCIÓN

Se considerarán casos de grandes muestras. Se probará la hipótesis nula $p = p_0$ contra las alternativas $p < p_0$, $p > p_0$ ó $p <> p_0$ mediante al aplicación del siguiente estadístico:

$$z = \frac{x - n\cdot p_0}{\sqrt{n\cdot p_o\cdot(1-p_0)}}$$

Ejemplo: En un estudio diseñado para investigar si ciertos detonadores empleados en explosivos en una mina de carbón cumplen con los requerimientos de que al menos el 90% encenderá el explosivo a ser detonado, se encontró que 174 de 200 funcionaron adecuadamente. Probar la hipótesis nula de que p=090 contra la hipótesis alterna p<0.90 con un nivel de significancia de 0.05.

1 – Hipótesis Nula p = 0.90,
Hipótesis Alternativa p < 0.90 (unilateral),
2 - Nivel de significancia: α =0.05. z_α = -1.645 (cola izquierda)
3 - Para trabajar con tablas normalizadas:

$$z = \frac{x - n\cdot p_0}{\sqrt{n\cdot p_o\cdot(1-p_0)}}$$

4 – Cálculos:

$$z = \frac{174 - 200\cdot 0.9}{\sqrt{200\cdot 0.9\cdot(1 - 0.9)}} = -1.414$$

5 – Dado que $-1.414 > z_{0.05}$ ($z_{0.05}$ = -1.645) no se Rechaza la Hipótesis Nula. No hay suficiente evidencia que la clase determinada de detonador no cumpla con las normas.

HIPÓTESIS RELATIVA A VARIAS PROPORCIONES

Cuando se compara la respuesta del consumidor (porcentaje favorable y desfavorable) con respecto a diferentes productos, cuando se decide que un proceso permanece constante día a día, etc. interesa probar si dos o más poblaciones binomiales tienen el mismo parámetro p. Si estos parámetros son p_1, p_2,.., p_k , interesa probar la Hipótesis Nula:

$$p_1 = p_2 = \ldots = p_k = p$$

contra la Hipótesis Alterna de que no lo sean. Bastaría con que uno de ellos fuese significativamente distinto para que la Hipótesis Nula no se cumpla.

Para aplicar esta técnica se necesitan muestras aleatorias de tamaño n_1, n_2,.., n_k de las k poblaciones. Luego, si los números correspondientes de éxitos son x_1, x_2,.., x_k respectivamente, la prueba que se utiliza se fundamenta en el hecho que:

- Para grandes muestras la distribución muestral de:

  $$z_i = \frac{x_i - n_i\cdot p_i}{\sqrt{n_i\cdot p_i\cdot\left(1 - p_i\right)}}$$

  es aproximadamente la distribución normal estándar.
- El cuadrado de una variable aleatoria, con función densidad normal estándar, es otra variable aleatoria con distribución chi-cuadrado con 1 grado de libertad.
- La suma de k variables aleatorias independientes cuyas distribuciones son chi-cuadrado con 1 grado de libertad, es otra variable aleatoria chi-cuadrado con k grados de libertad.

en consecuencia:

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(x_i - n_i\cdot p_i\right)^2}{n_i\cdot p_i\cdot\left(1 - p_i\right)}$$

es una variable con distribución chi-cuadrado con ν=k grados de libertad.

Dado que las p_i son todas iguales por hipótesis, se pueden sustituir por:

$$\hat{p} = \frac{x_1 + x_2 + \cdots + x_k}{n_1 + n_2 + \cdots + n_k}$$

ya que la Hipótesis Nula se debe rechazar si las diferencias entre las x_i y las $n_i\cdot\hat{p}$ son grandes, la región crítica es $\chi^2 > \chi^2_\alpha$ y el número de grados de libertad es $k-1$. La pérdida de un grado de libertada surge de reemplazar a p por su estimación $\hat{p}$.

Para calcular χ^2 conviene disponer los datos de la siguiente manera:

	Muestra 1	Muestra 2	- - -	Muestra k	Total
éxitos	x_1	x_2	- - -	x_k	x
fallas	$n_1 - x_1$	$n_2 - x_2$	- - -	$n_k - n_k$	$n - x$
Total	n_1	n_2		n_k	n

El elemento correspondiente al renglón i y a la columna j se denomina frecuencia observada en la celda **$o_{i,j}$** con i = 1, 2 y j = 1, 2, …, k.

Como antes, $\hat{p} = x/n$. Luego, el número esperado de éxitos y fracasos para la j-ésima muestra se estima con:

$$e_{1j} = n_j \cdot \hat{p} = \frac{n_j \cdot x}{n} \qquad e_{2j} = n_j \cdot (1 - \hat{p}) = \frac{n_j \cdot (n - x)}{n}$$

$e_{1,j}$ y **$e_{2,j}$** se denominan frecuencias esperadas en celdas para j = 1, 2, …, k.

Nótese que la frecuencia esperada para cualquier celda puede obtenerse multiplicando los totales de la columna y del renglón a los cuales pertenece y dividiendo después por el total general.

En esta notación el estadístico χ^2 (con $p_i = \hat{p}$)

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{(o_{i,j} - e_{i,j})^2}{e_{i,j}} \qquad \text{con } \nu = k-1 \text{ grados de libertad}$$

Deducción: A partir de:

$$\chi^2 = \sum_{j=1}^{k} \frac{(x_j - n_j \cdot p_j)^2}{n_j \cdot p_j \cdot (1 - p_j)} = \sum_{j=1}^{k} \frac{(x_j - n_j \cdot \hat{p})^2}{n_j} \cdot \left(\frac{1}{\hat{p}} + \frac{1}{1 - \hat{p}}\right)$$

$$\chi^2 = \sum_{j=1}^{k} \left[\frac{(x_j - n_j \cdot \hat{p})^2}{n_j \cdot \hat{p}} + \frac{(x_j - n_j \cdot \hat{p})^2}{n_j \cdot (1 - \hat{p})}\right]$$

$$\chi^2 = \sum_{j=1}^{k} \left[\frac{(x_j - n_j \cdot \hat{p})^2}{n_j \cdot \hat{p}} + \frac{(n_j - x_j - n_j + n_j \cdot \hat{p})^2}{n_j \cdot (1 - \hat{p})}\right]$$

$$\chi^2 = \sum_{j=1}^{k} \left[\frac{(o_{1,j} - e_{1,j})^2}{e_{1,j}} + \frac{(o_{2,j} - e_{2,j})^2}{e_{2,j}}\right]$$

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{(o_{i,j} - e_{i,j})^2}{e_{i,j}}$$

Ejemplo: Muestras de tres tipos de materiales sujetas a cambios extremos de temperatura producen los resultados que aparecen en la siguiente tabla:

	material A	material B	material C
Desmoronado	41	27	22
Resto intacto	79	53	78

utilizar el nivel de significación 0.05 para probar si, bajo las condiciones establecidas, la probabilidad de desmoronamiento es la misma en los tres tipos de materiales.

1. Hipótesis Nula: $p_1 = p_2 = p_3$
   Hipótesis Alternativa: $p_1 <> p_2 <> p_3$ (bilateral)
2. Nivel de significancia: α =0.05. $\chi^2_\alpha = \chi^2_{0.05} = 5.991$ con $\nu = k - 1 = 2$ g.d.l
3. Criterio: Se rechaza la Hipótesis Nula si $\chi^2 > 5.991$
4. Cálculos: Las frecuencias esperadas para las dos primeras celdas del primer renglón:

$$e_{1,1} = \frac{n_1 \cdot x}{n} = 120 \cdot \frac{90}{300} = 36 \qquad e_{1,2} = \frac{n_2 \cdot x}{n} = 80 \cdot \frac{90}{300} = 24$$

puede comprobarse que la suma de las frecuencias esperadas para cualquier renglón o columna es igual a las frecuencias observadas correspondientes, luego:

$$e_{1,3} = 90 - (36 + 24) = 30$$

$$e_{2,1} = 120 - 36 = 84 \qquad e_{2,2} = 80 - 24 = 56 \qquad e_{2,3} = 100 - 30 = 70$$

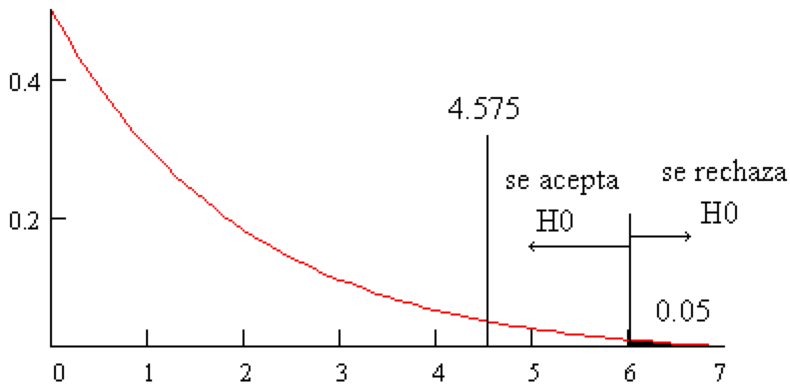
luego:

$$\chi^2 = \frac{(41-36)^2}{36} + \frac{(27-24)^2}{24} + \frac{(22-30)^2}{30} + \dots = 4.575$$

$$+ \frac{(79-84)^2}{84} + \frac{(53-56)^2}{56} + \frac{(78-70)^2}{70}$$

5- Decisión: dado que $4.575 < \chi^2$ no se Rechaza la Hipótesis Nula. Luego, la probabilidad de desmoronamiento es la misma en los tres tipos de materiales

Gráficamente:



Una función Matlab que resuelve este problema es la siguiente:

```
function hipo_prop(alfa)

obser =[41    27    22
   79    53    78];

% Prueba de Hipotesis para varias proporciones
%
% Entrada: obser, matriz con el cuadro de observaciones
% Salida: chi, estadistico chi-cuadrado
%
j=size(obser,1);  %calcula numero de filas de la matriz de observaciones
k=size(obser,2);  %calcula numero de columnas de la matriz de observaciones
suma_filas=sum(obser);suma_cols=sum(obser');
gt=sum(suma_filas);   %gran total
for i=1:k, esper(1,i)=suma_filas(i)*suma_cols(1)/gt; end
for i=1:k, esper(2,i)=suma_filas(i)*(gt-suma_cols(1))/gt; end
chi=0;
for i=1:j,for h=1:k,
      chi=chi+(obser(i,h)-esper(i,h))^2/esper(i,h);
end,end
chi
```

Ejecutando:

```
>> hipo_prop
chi =
   4.5754
```

Se acostumbra a redondear las frecuencias esperadas en celda al entero más cercano o a un decimal.

Hasta ahora la Hipótesis Alterna ha sido $p_1 <> p_2 <> \ldots <> p_k$, y para k=2 la Hipótesis Alternativa se reduce a $p_1 <> p_2$. Habrá problemas en que $p_1 < p_2$ o $p_1 > p_2$, se puede fundamentar la prueba en el estadístico:

$$z = \frac{\frac{x_1}{n_1} - \frac{x_2}{n_2}}{\sqrt{\hat{p}\cdot(1-\hat{p})\cdot\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \qquad \text{donde} \qquad p = \frac{x_1 + x_2}{n_1 + n_2}$$

Verificación:

$$\chi^2 = \frac{(x_1 - n_1\cdot p)^2}{n_1\cdot p\cdot(1-p)} + \frac{(x_2 - n_2\cdot p)^2}{n_2\cdot p\cdot(1-p)}$$

$$\chi^2 = \frac{\left(x_1 - n_1\cdot\frac{x_1+x_2}{n_1+n_2}\right)^2}{n_1\cdot p\cdot(1-p)} + \frac{\left(x_2 - n_2\cdot\frac{x_1+x_2}{n_1+n_2}\right)^2}{n_2\cdot p\cdot(1-p)}$$

$$x_1 - n_1\cdot\frac{x_1+x_2}{n_1+n_2} = \frac{x_1\cdot n_1 + x_1\cdot n_2 - n_1\cdot x_1 - n_1\cdot x_2}{n_1+n_2} = \frac{(x_1\cdot n_2 - n_1\cdot x_2)}{(n_1+n_2)}$$

$$x_2 - n_2 \cdot \frac{x_1 + x_2}{n_1 + n_2} = \frac{x_2 \cdot n_1 + x_2 \cdot n_2 - n_2 \cdot x_1 - n_2 \cdot x_2}{n_1 + n_2} = \frac{(x_2 \cdot n_1 - n_2 \cdot x_1)}{(n_1 + n_2)}$$

$$\chi^2 = \frac{\left[\frac{(x_1 \cdot n_2 - n_1 \cdot x_2)}{(n_1 + n_2)}\right]^2}{n_1 \cdot p \cdot (1-p)} + \frac{\left[\frac{(x_2 \cdot n_1 - n_2 \cdot x_1)}{(n_1 + n_2)}\right]^2}{n_2 \cdot p \cdot (1-p)}$$

$$\chi^2 = \left(\frac{1}{n_1} + \frac{1}{n_2}\right) \cdot \frac{(x_2 \cdot n_1 - n_2 \cdot x_1)^2}{(n_1 + n_2)^2 \cdot p \cdot (1-p)} = \frac{n_1 + n_2}{n_1 \cdot n_2} \cdot \frac{(x_2 \cdot n_1 - n_2 \cdot x_1)^2}{(n_1 + n_2)^2 \cdot p \cdot (1-p)} \cdot \frac{(n_1 \cdot n_2)}{n_1 \cdot n_2}$$

$$\chi^2 = (n_1 + n_2) \cdot \frac{\left[\frac{(x_2 \cdot n_1 - n_2 \cdot x_1)}{n_1 \cdot n_2}\right]^2}{(n_1 + n_2)^2 \cdot p \cdot (1-p)} \cdot (n_1 \cdot n_2) = \frac{\left(\frac{x_2}{n_2} - \frac{x_1}{n_1}\right)^2}{p \cdot (1-p)} \cdot \frac{n_1 \cdot n_2}{n_1 + n_2}$$

$$\chi^2 = \frac{\left(\frac{x_1}{n_1} - \frac{x_2}{n_2}\right)^2}{p \cdot (1-p) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Ejemplo: Un estudio señala que 16 de 200 tractores producidos en una línea de ensamblado requieren ajustes minuciosos y lo mismo sucede con 14 de 400 producidos en otra línea. Con un nivel de significación de 0.01 ¿Apoya esto la afirmación de que la segunda línea efectúa un trabajo superior?.

1. Hipótesis Nula: $p_1 = p_2$
   Hipótesis Alternativa: $p_1 >$ (unilateral)
2. Nivel de significancia: $\alpha = 0.01$. $z_\alpha = z_{0.01} = 2.33$
3. Criterio: Se rechaza la Hipótesis Nula si $z > 2.33$
4. Cálculos: Las frecuencias esperadas para las dos primeras celdas del primer renglón:

$$p = \frac{16 + 14}{200 + 400} = 0.05$$

$$z = \frac{\frac{16}{200} - \frac{14}{400}}{\sqrt{0.05 \cdot (1 - 0.05) \cdot \left(\frac{1}{200} + \frac{1}{400}\right)}} = 2.384$$

5. Decisión: dado que $2.384 < z$ se Rechaza la Hipótesis Nula. Luego, la segunda línea es mejor que la primera

TABLAS r x c

Se llaman así a aquellas en que los datos se disponen en dos criterios de clasificación que tiene r renglones y c columnas. Es igual que el caso anterior, pero con más de dos resultados posibles. También se las llama Tablas de Contingencia.

Para el análisis de una tabla r x c, se calculan primero las frecuencias esperadas en cada celda $e_{i,j}$, es decir multiplicando los totales de los renglones y columnas respectivos y dividiendo por el gran total. Sólo es necesario calcular (r-1).(c-1) de las $e_{i,j}$, las restantes salen por sustracción de totales en los renglones o columnas apropiadas. Luego, el estadístico para el análisis de la tabla r x c es:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(o_{i,j} - e_{i,j}\right)^2}{e_{i,j}}$$

Se rechaza la Hipótesis Nula si el valor de χ^2 excede a χ^2_α para $\nu = (r-1).(c-1)$ grados de libertad.

Ejemplo: Para determinar si en realidad existe una relación entre el aprovechamiento de un empleado en el programa de capacitación y su rendimiento real en el trabajo, considerada una muestra de 400 casos de archivos muy detallados de los que se obtuvieron los resultados que se ven en la tabla:

Aprovechamiento en el programa de entrenamiento

		Debajo del Promedio	Promedio	Sobre el Promedio	Total
Exito en el puesto (Clasificación en el empleo)	Deficiente	23	60	29	112
	Promedio	28	79	60	60
	Muy buena	9	49	63	63
	Total	60	188	152	400

con el nivel de significancia 0.01 probar la Hipótesis Nula de que el aprovechamiento en el programa de capacitación y el éxito en el trabajo son independientes.

5. Hipótesis Nula: el aprovechamiento en el programa de capacitación y el éxito en el trabajo son independientes.
   Hipótesis Alternativa: no son independientes
6. Nivel de significancia: α =0.01. $\chi^2_\alpha = \chi^2_{0.01}$ = 13.277 con ν=(c-1).(r-1)=2*2=4
7. Criterio: Se rechaza la Hipótesis Nula si $\chi^2 > 13.277$
8. Cálculos: Las frecuencias esperadas son:

$$e_{1,1} = 60\cdot\frac{112}{400} = 16.8 \qquad e_{1,2} = 188\cdot\frac{112}{400} = 52.6$$

$$e_{2,1} = 60\cdot\frac{167}{400} = 25 \qquad e_{2,2} = 188\cdot\frac{167}{400} = 78.5$$

Por sustracción o cálculo (por ejemplo: $e_{1,3}$ = 112-(16.8+52.6) = 42.6, $e_{2,3}$ =63.5, $e_{3,1}$ = 18.2, $e_{3,2}$ =52.9, $e_{3,3}$ =45.9)

$$\chi^2 = \frac{(23-16.8)^2}{16.8} + \ldots + \frac{(63-45.9)^2}{45.9} = 20.18$$

5- Decisión: dado que 20.18 > 13.277 Se Rechaza la Hipótesis Nula. Luego, existe diferencia entre el aprovechamiento de un empleado en el programa de capacitación y su éxito en el empleo.

Haciendo una pequeña modificación en la función Matlab anterior, se puede resolver mediante dicha herramienta este problema:

```
function hipo_rc

obser =[23   60   29
        28   79   60
         9   49   63];
% Prueba de Hipotesis para Tabla de Contingencia
%
% Entrada: obser, matriz con el cuadro de observaciones (externo)
% Salida: chi, estadistico chi-cuadrado
%
j=size(obser,1);  %calcula numero de filas de la matriz de observaciones
k=size(obser,2);  %calcula numero de columnas de la matriz de observaciones
suma_filas=sum(obser);suma_cols=sum(obser');
gt=sum(suma_filas);   %gran total
for t=1:j,
   for i=1:k, esper(t,i)=suma_filas(i)*suma_cols(t)/gt; end
end
chi=0;
for i=1:j,for h=1:k,
      chi=chi+(obser(i,h)-esper(i,h))^2/esper(i,h);
end,end
chi
```

BONDAD DE AJUSTE

Se habla de Bondad de Ajuste cuando se trata de comparar una distribución de frecuencia observada con los valores correspondientes a una distribución de esperada o teórica.

Como ejemplo se analizan 400 intervalos de 5 minutos en el control de tráfico aéreo en cuanto a la recepción de mensajes de radio, comparándola con una distribución de Poisson con λ=4.6.

Número de mensajes en interv. de 5 min.	Frecuencias observadas		Probabilidad de Poisson	Frecuencias esperadas	
0	3	18	0.010	4.0	22.4
1	15		0.046	18.4	
2	47		0.107	42.8	
3	76		0.163	65.2	
4	68		0.187	74.8	
5	74		0.173	69.2	
6	46		0.132	52.8	
7	39		0.087	34.8	
8	15		0.050	20.0	
9	9		0.025	10.0	
10	5	10	0.012	4.8	8.0
11	2		0.005	2.0	
12	2		0.002	0.8	
13	1		0.001	0.4	

La columna correspondiente a Probabilidad de Poisson se obtiene o bien de Tablas (generalmente acumuladas, luego se debe hacer una resta para calcular este valor) o bien resolviendo la siguiente expresión (como ejemplo, para k=3):

$$\frac{\lambda^k}{k!}\cdot e^{-\lambda} = \frac{4.6^3}{3!}\cdot e^{-4.6} = 0.163$$

Se han combinado algunos datos de forma que ninguna de las frecuencias esperadas se menor que 5.

Para probar las discrepancias entre las frecuencias observadas y las esperadas pueden atribuirse al azar, se usa el estadístico:

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

con $k - m - 1$ grados de libertad. Siendo k el número de términos de la fórmula y m el número de parámetros en la distribución supuesta (en este caso Poisson, λ).

Problema I: En relación con el caso anterior, probar con un nivel de significación 0.05 si los datos pueden considerarse como una variable aleatoria que tiene distribución de Poisson con λ=4.6.

1. Hipótesis Nula: variable aleatoria que tiene distribución de Poisson con λ=4.6. Hipótesis Alternativa: No la tiene.
2. Nivel de significancia: α =0.05. $\chi^2_\alpha = \chi^2_{0.05} = 15.507$ con ν=k-m-1=10-1-1=8
3. Criterio: Se rechaza la Hipótesis Nula si $\chi^2 > 15.507$
4. Cálculos:

$$\chi^2 = \frac{(18-22.4)^2}{22.4} + \ldots + \frac{(47-42.8)^2}{42.8} = 6.749$$

5- Decisión: dado que 6.749 < 15.507 No se Rechaza la Hipótesis Nula. Luego, la distribución de Poisson λ=4.6 proporciona un buen ajuste.

Problema II: La siguiente tabla indica la cifra promedio de accidentes por mil horas-hombre de la muestra de 50 firmas , obtenidas de una industria específica. Probar la hipótesis nula de que las frecuencias observadas en esta muestra siguen una distribución normal, utilizando un nivel de significancia de 0.05.

Cifra promedio de accidentes por mil horas-hombre	Número de firmas
1.5 – 1.7	3
1.8 – 2.0	12
2.1 – 2.3	14
2.4 – 2.6	9
2.7 – 2.9	7
3.0 – 3.2	5

Total	50

1. Hipótesis Nula: variable aleatoria que tiene distribución Normal. Hipótesis Alternativa: No la tiene.
2. Nivel de significancia: $\alpha = 0.05$. $\chi^2_\alpha = \chi^2_{0.05} = 3.841$ con $\nu = k-m-1 = 4-2-1 = 1$
3. Criterio: Se rechaza la Hipótesis Nula si $\chi^2 > 3.841$
4. Cálculos: Es necesario determinar dos parámetros muestrales, la media y la desviación estándar:

$$\bar{x} = \frac{3 \cdot 1.6 + 12 \cdot 1.9 + 14 \cdot 2.2 + 9 \cdot 2.5 + 7 \cdot 2.8 + 5 \cdot 3.1}{50} = 2.32$$

$$s = \sqrt{\frac{3 \cdot 1.6^2 + 12 \cdot 1.9^2 + 14 \cdot 2.2^2 + 9 \cdot 2.5^2 + 7 \cdot 2.8^2 + 5 \cdot 3.1^2 - 50 \cdot 2.32^2}{50}} = 0.42$$

Para determinar las frecuencias esperadas se arma el siguiente cuadro:

Cifra promedio de acciden-tes por mil horas-hombre (fronteras de clase)	fronteras de clase (en unidades normales estándar)	Probabilidad de estar en cada categoría	Frecuencia esperada =50.p
1.45 – 1.75	-2.07 a –1.35	0.07	3.5
1.75 – 2.05	-1.36 a –0.64	0.17	8.5
2.05 – 2.35	-0.64 a 0.07	0.27	13.5
2.35 – 2.65	0.07 a 078	0.26	13.0
2.65 – 2.95	0.78 a 1.50	0.15	7.5
2.95 – 3.25	1.50 a 2.21	0.05	2.5

En base a esto se puede generar el siguiente cuadro:

Cifra promedio de accidentes por mil horas-hombre	Frecuencias observadas, f_0		Frecuencias esperadas, f_e		$(f_0 - f_e)^2 / f_e$
1.5 – 1.7	3	15	3.5	12	0.75
1.8 – 2.0	12		8.5		
2.1 – 2.3	14		13.5		0.02
2.4 – 2.6	9		13.0		1.23
2.7 – 2.9	7	12	7.5	10	0.40
3.0 – 3.2	5		2.5		
				$\chi^2 =$	2.40

5- Decisión: dado que $2.40 < 3.841$ No se Rechaza la Hipótesis Nula. Luego, la distribución Normal proporciona un buen ajuste.