3 - Varianza y proporciones

18 July 2020 20:14

Estimación de varianzas

Si bien la varianza muestral s^2 es un *estimador insesgado* de σ^2 , esto **no implica** que la desviación estándar muestral s sea un *estimador insesgado* de σ. No lo es

Para muestras de *gran tamaño*, el sesgo es pequeño y se acostumbra a estimar σ con s.

Rango muestral R:

- *Mayor_valor_de_una_muestra - Menor_valor_de_la_muestra*

Dada una muestra de tamaño **n** de una *población normal*, puede verificarse que la distribución muestral del rango tiene:

Media:

- $d_2\sigma$

Deviación estándar:

- $d_3\sigma$

Donde las **d** son constantes que *dependen del tamaño* de la muestra. Tenemos el siguiente cuadro:

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

Luego $\frac{R}{d_2}$ es un *estimador isesgado* de σ que proporciona una estimacion de σ tan buena como s para muestras *muy pequeñas* (n ≤ 5). Cuando se *incrementa* el tamaño muestral sucede lo **contrario**.

$$\sigma = \frac{R}{d_2}$$

R se emplea fundamentalmente en *Control de Calidad*

Estimación de intervalos para σ (o σ^2):

La mayoría de aplicaciones practicas se basan en la desviación estándar muestral (o en la varianza muestral). En muestras aleatorias de **poblaciones normales**

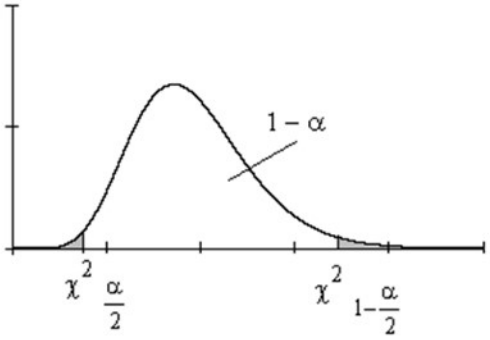
$$\chi^2 = \frac{(n-1)\cdot s^2}{\sigma^2} \qquad \text{con } \nu = n\text{-}1$$ Grados de libertad

Siendo esta la abscisa que deja a la *izquierda* un área α.
Se puede asegurar con una probabilidad de 1-α que se satisface:

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

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

Se puede determinar un *intervalo de confianza* 1-α para la varianza poblacional



Esto se aplica a muestras aleatorias de **poblaciones normales**, pero si el tamaño de la muestra es *grande*, la distribución muestral de la deviación estándar puede *aproximarse* a una distribución normal con media σ y deviación estándar $\frac{\sigma}{\sqrt{2n}}$.

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

Por lo que se puede generar un intervalo de confianza 1-α

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

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

Que es un intervalo de confianza para σ en muestas de *gran tamaño*

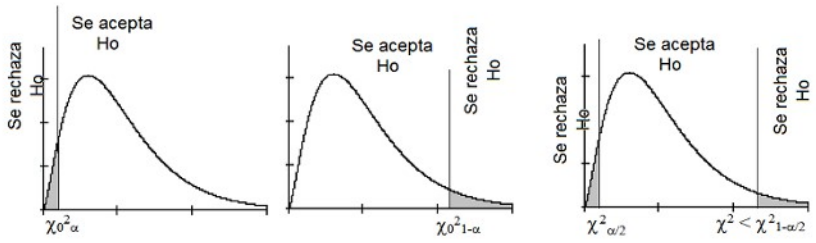
Hipótesis referida a una varianza

Para muestras aleatorias extraídas de una **población normal** con varianza σ^2, se puede usar

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

Para tomar las *regiones criticas* para probar $\sigma^2 = \sigma^2_0$

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



Si la muestra es *grande*, la hipótesis nula se puede probar con:

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

Estimación de proporciones

Usualmente la información que se dispone para estimar la proporción es el **numero de veces** (**x**) que ocurre un evento y el **numero de observaciones** (**n**) que se realizaron. Por lo que la *estimación puntual* suele ser la proporción muestral ($\frac{x}{n}$)
Para los ensayos que satisfacen una *distribución binomial*, se verifica:

$\mu = np$
$\sigma = \sqrt{np(1-p)}$

Que si se los divide por **n**, se encuentra la media y la desviación estándar de la *proporción de éxitos*, es decir, la ***proporción muestral.***

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

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

Ya que x e $\frac{x}{n}$ son *variables discretas* y que para hallar $\frac{\sigma}{n}$ hace falta conocer **p**, es difícil crear un *intervalo de confianza* con un nivel de confianza $1 - \alpha$. Para lograrlo se deben determinar x_0 y x_1 para un *conjunto determinado* de valores de **p.**

Donde:

- x_0 es el **maximo** entero para el que se verifica la desigualdad:

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

- x_1 es el **minimo** entero para el que se verifica la desigualdad:

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

La **distribución binomial** es una distribución de *probabilidad discreta*, aplicable a procesos de Bernoulli.

Procesos de Bernoulli:
1. Hay **dos** resultados posibles *mutuamente excluyentes* en cada ensayo (éxito y fracaso).
2. La serie de ensayos constituyen *eventos independientes.*
3. La probabilidad de éxito **p** permanece *constante* en todos los ensayos, es decir, el proceso es *estocástico.*

Para el calculo de la distribución binomial se necesitan *3 valores*
- **x**: el numero de éxitos
- **n**: el numero de ensayos
- **p**: la probabilidad de éxito

Para el calculo de la distribución binomial se necesitan *3 valores*

- **x**: el numero de éxitos
- **n**: el numero de ensayos
- **p**: la probabilidad de éxito

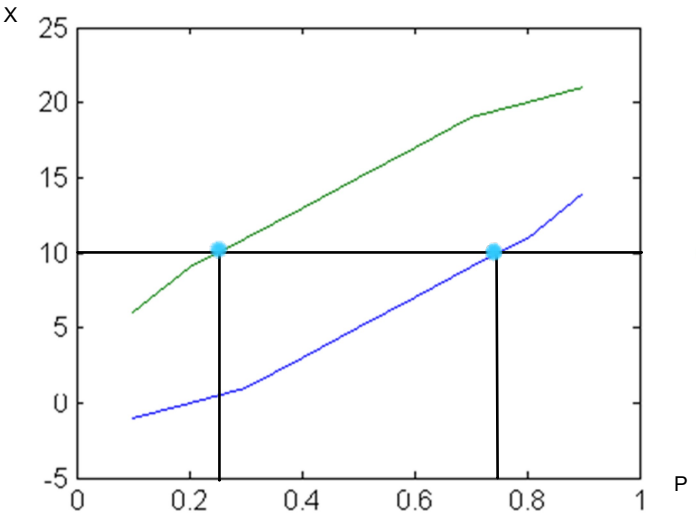
$$b(x,n,p) = C_{n,x}\cdot p^{x}\cdot(1-p)^{n-x} = \frac{n!}{x!\cdot(n-x)!}\cdot\left[p^{x}\cdot(1-p)^{n-x}\right]$$

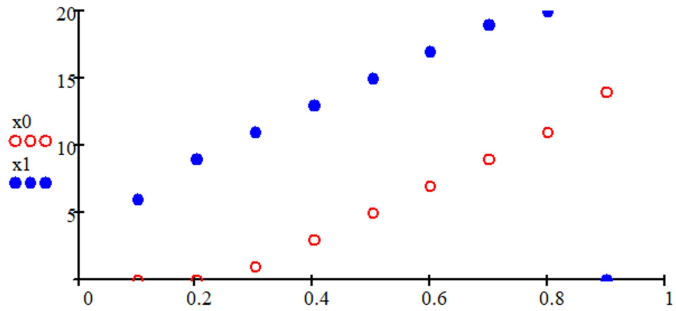
Entonces se puede asegurar que se cumple la siguiente *desigualdad* con una probabilidad de aproximadamente 1-α:

$$x_{0(p)} < x < x_{1(p)}$$

Para transformar esta desigualdad en *intervalo de confianza* se recurre a un método grafico:

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	-





Para distintos valores de **n** se obtienen distintas ramas

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

Por lo que dado un **x**, se pueden obtener *cotas* para **p** con un nivel de significación de 1-α.
Existen gráficos de este tipo para *niveles de confianza* de 95 y 99%, para varios valores de n. En los mismos se emplea la proporción muestral ($\frac{x}{n}$) en lugar de **x**.

Para poder *aproximar* la **distribución binomial** a la **normal**, se debe cumplir:

- $np > 5$
- $(1-p) > 5$

Y se puede asegurar con una *probabilidad* 1-α que se cumple:

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

Para evitar cálculos complicados, se *aproxima* p con $\frac{x}{n}$

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

Para muestras de *gran tamaño*
La magnitud del **error** cometido cuando se usa esta aproximación esta dada por: $\left|\frac{x}{n} - p\right|$
En base a la *distribución normal*, se puede asegurar con *probabilidad* 1-α que se cumplirá:

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

Por lo que este será el *Error Máximo de estimación* cuando se sustituye p con $\frac{x}{n}$

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

Este error puede ser usado para determinar el *tamaño muestral* que es necesario para llegar a un *grado* deseado de *precisión*.

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

Formula que solo se puede usar si se tiene información sobre **p** (en base a datos auxiliares, como una muestra previa). Si no se conoce, se sabe que $p(1-p)$ es a lo sumo $\frac{1}{4}$, correspondiente a $p = \frac{1}{2}$

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

Entonces, se puede asegurar con una *probabilidad* de 1-α que el error de utilizar $\frac{x}{n}$ como estimacion de p, **no** va a exceder E
Una vez obtenidos los datos, se puede *asegurar* con una *confianza* de al menos 1-α que el error **no** sobrepasa E.

Con una confianza de 95% y un error de a lo sumo 0.04:

- *No se sabe* como podría ser la proporción muestral

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

- *Se sabe* que la proporción real no excede de 0.127

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

Esto ilustra que conocer alguna información auxiliar de p reduce en gran medida el tamaño de muestra requerida

Cuando **p** es *próximo* a 0 (alta confiabilidad) y cuando **p** es la *probabilidad de fracaso*, se necesitan *intervalos de confianza unilaterales*. Para **p** *chicos* y **n** *grandes*, la distribución de Poisson se *aproxima* a la binomial, por lo que se puede mostrar que:

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

Con $\nu = 2(x+1)$ grados de libertad

Hipótesis relativa a una proporción

Se verán los casos para *muestras grandes*, donde se probarán:

Hipótesis Nula:
- $p = p_0$

Hipótesis Alternativas:
- $p \neq p_0$
- $p < p_0$
- $p > p_0$

Aplicando el estadístico:

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

Hipótesis relativa a varias proporciones

Cuando interesa probar si dos poblaciones binomiales tienen el *mismo parámetro* **p**.

Hipótesis Nula:
- $p_1 = p_2 = \cdots = p_k = p$

Hipótesis Alterna:
- Al menos 1 es *significativamente* distinto

Para poder saberlo, se necesitan **k** muestras aleatorias de **k** poblaciones de tamaño $n_1, n_2, \ldots, n_k$, y el numero de éxitos $x_1, x_2, \ldots, x_k$.

La prueba se fundamenta en:
- Para **muestras grandes** la distribución muestral de

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

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

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

$\nu = k$ grados de libertad

Como todas las p_i son *iguales por la hipótesis*,

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

Se las puede *sustituir* por $\hat{p}$.

La Hipótesis Nula se rechaza si las diferencias entre x_i y $n_i\hat{p}$ son grandes. Por lo que la región crítica es $\chi^2 > \chi^2_{\alpha}$, tomando $\nu = k - 1$ grados de libertad (por haber reemplazado **p** con su estimacion $\hat{p}$)

Para el calculo en conveniente tener los datos en un cuadro similar al siguiente

	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

- Renglón: $i = 1,2,\dots,k$
- Columna: $j = 1,2,\dots,k$
- Frecuencia observada en la celda: $o_{i,j}$
- $\hat{p} = \frac{x}{n}$
- Éxitos para la *j-ésima* muestra (frecuencia esperada):

$$e_{1j} = n_j \cdot \hat{p} = \frac{n_j \cdot x}{n}$$

- Fracasos para la *j-ésima* muestra (frecuencia esperada):

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

- En la misma notación, el estadístico χ^2

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

$\nu = k - 1$ grados de libertad

$p_i = \hat{p}$

Deducción:

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

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

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

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

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

Habrá casos en los que $k = 2$, por lo que la Hipótesis Alternativa podría ser:

- $p_1 \neq p_2$
- $p_1 > p_2$
- $p_1 < p_2$

Por lo que se puede fundamentar la prueba en el siguiente estadístico:

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

Donde $p = \frac{x_1 + x_2}{n_1 + n_2}$

Deducción:

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

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

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

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

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

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

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

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

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

*"El cuadrado de una variable aleatoria con función de densidad **normal estándar**, es otra variable aleatoria con distribución **chi-cuadrado** con 1 grado de libertad."*

Tablas r x c ó Tablas de Contingencia

Los datos están dispuestos en dos criterios de clasificación:

- **r:** renglones
- **c:** columnas

Como en los casos anteriores, pero con *mas resultados*

Para su análisis, se calculan las frecuencias esperadas en $(r-1)(c-1)$ celdas $e_{i,j}$, el resto se calculan por *sustraccion de totales* en los renglones o columnas apropiadas.

$$e_{i,j} = \frac{n_j x_i}{n}$$

El estadístico para el análisis de la tabla es:

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

Donde se rechaza la Hipótesis Nula si $\chi^2 > \chi_\alpha{}^2$ para $\nu = (r-1)(c-1)$ grados de libertad

Bondad de ajuste

Se habla de ella cuando se quiere *comparar* una distribución de frecuencias observadas con los valores correspondientes a una distribución de frecuencias esperadas o teóricas.

Ejemplo:
400 intervalos de 5 minutos en el control de trafico aéreo en cuanto a la recepción de mensajes de radio, comparándola con una distribución de Poisson $\lambda = 4.6$.

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: "Probabilidad de Poisson", se obtiene de tablas o de la expresión

$$\frac{\lambda^k}{k!} \cdot e^{-\lambda}$$ *Densidad de probabilidad*

Además, en el cuadro se han *combinado* datos, de manera que **ninguna** de las frecuencias esperadas sea menor a 5 (líneas verticales).

Para probar que las discrepancias entre las frecuencias observados y las esperadas pueden *atribuirse al azar*, se usa otra vez el estadístico:

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

$\nu = k - m - 1$ grados de libertad

Siendo:

- **k:** el *número de términos* de la formula (renglones del cuadro después de la combinación)
- **m:** el *número de parámetros* de la distribución supuesta (en nuestro ejemplo Poisson)

Hipótesis:
Hipótesis Nula:
- La variable aleatoria **tiene** una distribución x con parámetros p_1 ...

Hipótesis Alternativa:
- **No** la tiene