CAPITULO I

DISTRIBUCIÓN CHI CUADRADA

La distribución de Chi cuadrada tiene muchas aplicaciones especialmente en las ciencias biológicas y sociales, en donde se estudia una conducta (lo esperado) en función de una respuesta (lo observado). Si el conjunto de valores observados sigue el mismo comportamiento de lo esperado, entonces, estadísticamente, se acepta la hipótesis que lo observado sigue el comportamiento de lo esperado.

Esta metodología puede ser utilizada para una prueba de:

Frecuencias y bondad de Ajuste.
Independencia entre variables
Homogeneidad de muestras
Homogeneidad de variancias.

Casos de frecuencias y bondad de Ajuste, probar estadísticamente si:

- la relación de ingresantes a la UNALM de colegios particulares a nacionales es de 2 a 1.
- un juego al azar (Ruleta) es realmente al azar.
- el número de accidentes que ocurre en un determinado lugar sigue una ley de Poisson,
- el número de tubérculos dañados en plantas sigue una ley Poisson
- la longitud de una cola de espera en un lugar de atención al publico sigue una Poisson,
- el tiempo de respuesta de una transacción en un banco sigue una ley exponencial.
- el numero de artículos defectuosos en cajas de 10, sigue una ley Binomial.
- el numero de plantas germinadas de paquetes de 10 semillas sigue una ley Binomial.
- el numero de bolsas de leche defectuosas producidas en una hora sigue una ley Poisson.

Casos de Independencia

- Preferencias a ciertos productos y localidades,
- Procedencia de colegio nacional y privado y el rendimiento en la Universidad
- Relación talla, sexo, peso, situación económica y el rendimiento en la Universidad
- El nivel de pobreza y estudio en la zona rural y urbana

Casos de Homogeneidad de muestra

- La distribución del consumo de tipo de carne en distritos de la provincia Lima
- La preferencia o popularidad de candidatos por distritos
- La distribución de estudiantes por procedencia de lugar en las Universidades de Lima.

Todas estas pruebas y otras que involucren la comparación de lo observado frente a lo esperado pueden ser analizado estadísticamente mediante la prueba de Chi Cuadrada.

La distribución Chi cuadrada X^2, permite resolver tal inferencia, bajo el supuesto que la variable aleatoria W definida por:

$$W = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{E_i}$$

Esta variable (W) sigue una distribución de probabilidades Chi cuadrada, con cierto grado de libertad, que es su parámetro, donde:

O_i = Frecuencia observada en una clase o categoría de estudio.
E_i = Frecuencia esperada en la misma clase o categoría.

Cuando el numero de grados de libertad es igual a 1, se utiliza la corrección de Yates (corrección por continuidad)

$$W = \sum_{i=1}^{k} \frac{(|O_i - e_i| - 0.5)^2}{E_i}$$

En algunos casos cuando el tamaño de la frecuencia es mayor de 50, se puede obviar la corrección.

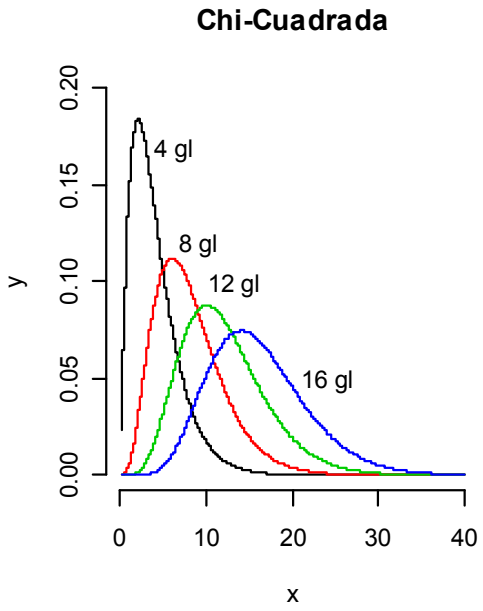
La Distribución Chi Cuadrada tiende a ser simétrica, a medida que los grados de libertad aumentan. (Ver grafico).

Programa en R para el grafico:

```
x<-seq(0.1,40,0.1);
y1<-dchisq(x,4);
y2<-dchisq(x,8);
y3<-dchisq(x,12);
y4<-dchisq(x,16);
y<-cbind(y1,y2,y3,y4);

matplot(x,y,type=rep("s",4),
lty=rep(7,4),
ylim=c(0,0.2),xlim=c(0,40),
frame=F,main="Chi-Cuadrada");

text(6,0.17,"4 gl");
text(9,0.12,"8 gl");
text(13,0.10,"12 gl");
text(24,0.05,"16 gl")
```



PRUEBA DE FRECUENCIAS

Es útil en el estudio de la distribución de frecuencias de una variable. El numero de clases o categorías debe ser al menos 2, lo suficiente como para no tener frecuencias menores del 5%. Muchas o pocas categorías dispersa o concentra la frecuencia en las categorías.

Para las pruebas estadísticas de frecuencia se requiere hallar los grados de libertad.

Para el caso de frecuencias, los grados de libertad es igual a “k-1”, donde “k” es el numero de clases o categorías.

En el estudio, se supone una distribución teórica, el cual se prueba con la muestra observada.

Ejemplo:

4 Candidatos postulan para la Presidencia de la Republica. Según los sondeos se tiene la siguiente distribución:

Candidatos A= 34 %, B = 28%, C=14% D=8% Otros 16%

Las categorías son (A, B, C, D y Otros)

Solución:

Hp: La preferencia a los candidatos se mantiene.
Ha: Hay cambios en la preferencia.
Nivel de riesgo: 0.10

Se hizo el estudio de una muestra en donde arrojo los siguientes resultados:

Total de encuestados: 120

Preferencias a los candidatos:

A = 45, B=30, C=18, D=6 y otros = 21

Candidato	Observado	Esperado	% Teórico
A	45	40.8	34
B	30	33.6	28
C	18	16.8	14
D	6	9.6	8
Otros	21	19.2	16
Total	120	120	100

Los valores esperados se determina por el producto del total (120) multiplicado por cada frecuencia relativa de cada clase:

40.8 = 120 x 34/100

con estos resultados se calcula el valor de Chi-cuadrado:

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

$$\frac{(45-40.8)^2}{40.8} + \frac{(30-33.6)^2}{33.6} \frac{(18-16.8)^2}{16.8} \frac{(6-9.6)^2}{9.6} \frac{(21-19.2)^2}{19.2} = 2.4225$$

El valor critico se busca en la tabla de Chi cuadrada con (k-1) = 5-1=4 grados de libertad y con nivel de significación de 0.10

Chi-Cuadrado tabular = 7.77.

El valor calculado es inferior al tabular, por lo tanto se acepta la hipótesis que las tendencias se mantienen.

Solución mediante el programa R.

```
# Valores teóricos
y<-c(A=34 , B=28, C=14,  D=8,  Otros= 16)
# Valores de la muestra
x<- c(A = 45, B=30, C=18, D=6, otros = 21)
# Prueba estadistica
chi.calculado <- chisq.test(x, p=y/sum(y) )
# Valor critico para 0.10 y 4 gl.
qchisq(1- 0.10 , 4)
```

Resultados: X-squared = 2.4225, df = 4, p-value = 0.6586
Valor critico: 7.77944

Ejemplo en proporciones:

Las frecuencias esperadas de un cruce genético entre la prole están en una proporción fenotipo de 3:1 de normal a mutante. Las frecuencias observadas fueron:

Fenotipo	Observadas
Normal	80
Mutante	10
----→ Total	90

Realice la prueba estadística para la prueba de la proporción planteada.

Hipótesis:

Hp : La proporción fenotipo normal y mutante es de 3:1
Ha : La proporción no es 3:1
Nivel de riesgo $\alpha = 0.10$

El número esperado por fenotipo es:

Fenotipo normal = (ΣO_i) P(normal) = 90 (3/4) = 67.5
Fenotipo mutante= (ΣO_i) P(mutante) = 90 (1/4) = 22.5

Fenotipo	Esperados
Normal	67.5
Mutante	22.5
----→ Total	90

Grados de libertad = 2 - 1 = 1

Los grados de libertad es igual a 1, no es necesario la corrección de Yates porque la muestra es mayor de 50. El valor de Chi cuadrado calculado es:

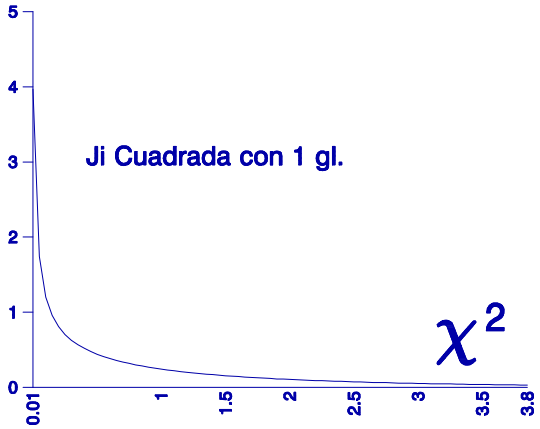
$$\chi^2(calculado) = \frac{(80 - 67.5)^2}{67.5} + \frac{(10 - 22.5)^2}{22.5}$$

X^2(calculado) = 2.3148 + 6.9444 = 9.2592

El valor crítico para gl = 1 y $\alpha = 0.10$

$$\chi^2(1) = 2.705$$

Se observa que el valor calculado es mayor que el tabular, entonces se rechaza la hipótesis planteada; por lo tanto se concluye que no hay suficiente razón estadística para tal afirmación sobre la proporción planteada.



Solución mediante el programa R.

```
> prop.test(80,90,0.75,correct=F)

1-sample proportions test without continuity correction

data:  80 out of 90, null probability 0.75
X-squared = 9.2593, df = 1, p-value = 0.002343
alternative hypothesis: true p is not equal to 0.75
95 percent confidence interval:
 0.80742 0.93852
sample estimates:
      p
0.88889

> qchisq(1- 0.10 , 1)
[1] 2.7055
```

Valor critico: 2.7055

Aplicación por Yates (caso de dos categorías y total de observaciones menos de 50).

Una moneda supuestamente balanceada, se somete a una prueba para certificar si es correcta para ser utilizada en ciertos ejercicios de selección aleatoria.

Ho: Moneda correctamente balanceada.
Ha: no es balanceada.

La moneda es lanzada 25 veces, resultado cara 10 veces.

Con esta respuesta, ¿ podemos aceptar la hipótesis?

Tabla de información:

	Observado	Esperado
Cara	10	12.5
Sello	15	12.5
	25	25

Utilizamos la corrección de Yates.

$$\chi^2_{calc} = \frac{\left(|10-12.5|-0.5\right)^2}{12.5} + \frac{\left(|15-12.5|-0.5\right)^2}{12.5} = 0.64$$

el valor de la tabla con un grado de libertad es 2.7055

Por lo tanto se acepta la hipótesis que la moneda es balanceada.

Mediante el programa R

En 25 ensayos, se observo 10 caras

```
> prop.test(10,25,0.5, correct=T)

1-sample proportions test with continuity correction

data:  10 out of 25, null probability 0.5
X-squared = 0.64, df = 1, p-value = 0.4237
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.2181 0.6111
sample estimates:
  p
0.4

> qchisq(1- 0.10 , 1)
[1] 2.7055
```

Valor critico: 2.7055

PRUEBA DE BONDAD DE AJUSTE

La prueba de bondad de ajuste consiste en determinar en que medida una distribución experimental sigue una determinada distribución de probabilidades teórica. El objetivo de esta prueba es la posibilidad de utilizar la distribución teórica en el tratamiento de los datos experimentales para un análisis estadístico.

Los pasos a seguir en una prueba de bondad son:

1. Plantear la hipótesis nula.- Los datos siguen la distribución teórica, frente a la hipótesis alterna que no siguen dicha distribución.

2. Se fija el nivel el nivel de significación de la prueba (α). Para estos casos, por lo general se utiliza 0.10 ó 0.05

3. Se calcula el valor de Chi cuadrada (X^2) según la fórmula para el caso de estudio y se determina los grados de libertad según la fórmula:

   gl = (k-p-1), donde:

   k = número de clases o categorías
   p = número de parámetros que fueron estimados a partir de la muestra.

4. Buscar el valor de Chi cuadrado tabular (valor crítico) con los grados de libertad y el nivel de significación dado.
5. Comparar los dos valores de Chi cuadrado.

   Si X^2(calculado) $\leq X^2$(tabular), se acepta la hipótesis planteada con una determinada probabilidad de error de aceptar algo falso, a este error se conoce como error tipo II y su probabilidad es representado por la letra griega beta (ß).

   Si X^2(calculado) > X^2(tabular), se rechaza la hipótesis, significa que la distribución de la muestra no sigue la distribución teórica planteada con una probabilidad de error tipo I o "α".

Pruebas de Normalidad (test normality)

Para las pruebas estadísticas que requiere la normalidad de los datos, es necesario probar si los datos están normalmente distribuidas.

Hp: Los datos están normalmente distribuidos
Ha: No hay normalidad.
Riesgo 0.10

Algunas pruebas útiles que podemos usar con el programa R son:

Shapiro-Wilk test for normality

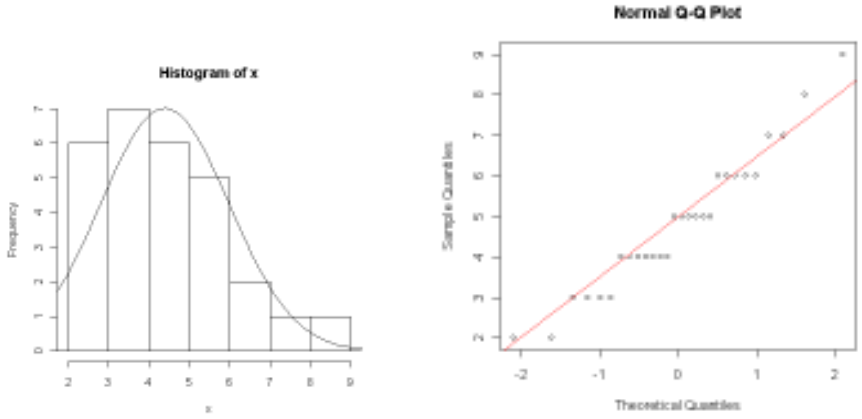
```
shapiro.test{base}
normal.frec{taller}

Ejemplo:

x<- c(3,2,5,4,7,6,5,4,3,4,5,6,5,3,4,6,7,2,3,4,5,6,4,6,5,4,8,9)
shapiro.test(x)
h<-hist(x)
normal.frec(h)
qqnorm(x); qqline(x, col = 2)

        Shapiro-Wilk normality test

data:  x
W = 0.9554, p-value = 0.2693
```



Según los resultados, nos fijamos en el valor P-Value, como es mayor del riego de 0.10, entonces aceptamos la hipótesis planteada, es decir hay normalidad en los datos.

Anderson-Darling test for normality
ad.test {nortest}

Ampliamente utilizada.

```
> library(nortest)
> ad.test(x)

        Anderson-Darling normality test

data:  x
A = 0.5173, p-value = 0.1732
```

Con esta prueba tambien los datos cumplen la normalidad

Cramer-von Mises test for normal
cvm.test {nortest}

```
> library(nortest)
> cvm.test(x)

       Cramer-von Mises normality test

data:  x
W = 0.0928, p-value = 0.1343
```

De igual forma se acepta la normalidad de los datos.
Como conclusión de estos resultados se puede decir, que las pruebas nos dan diferente respuesta en medir el riesgo de rechazar la hipótesis siendo esta verdadera. Como vemos entre una prueba menos exigente como es la de Shapiro y otra más exigente como es la Cramer-von, tenemos una intermedia que es la prueba de Anderson que la podemos utilizar como una buena opción.

Aplicación en la Binomial.

En la Binomial son dos parámetros a estimar son “n” y “p”.

La función de probabilidad esta dado por:

$$P(X=k)=\binom{n}{k}p^{k}(1-p)^{n-k};\quad k=0,1,2,..,n$$

p = Probabilidad de éxito de un ensayo de bernoulli,
n = total de ensayos de benoulli,
X = Variable aleatoria numero de éxitos,
K = es el numero de éxitos que se desea encontrar en “n” experimentos de bernoulli.
P(X=k) = Probabilidad de k Exitos.

Ejemplo. Considere el siguiente caso. En la comercialización de manzanas, una empresa exportadora envía semanalmente lotes de 50 cajas al exterior, cada caja tiene un peso aproximado de 20 kilos. Las cajas son previamente almacenadas. Para el control de calidad se examinan 5 cajas al azar, si en alguna caja encuentran por lo menos una manzana malograda, esta es calificada como mala. Para que pase el control mediante la inspección de la muestra no debe haber caja malograda, si solo existe una caja esta será cambiada, si hay mas de 1 en las 5 inspeccionadas, se inspeccionaran las cincuenta cajas. Según las estadísticas pasadas de un total de 40 envíos, se registro lo siguiente:

X	0	1	2	3	4	5
Oi	6	13	10	7	3	1

Se puede afirmar que la variable numero de cajas malogras en la muestra de 5 sigue una distribución binomial?.

Solución:

Hp: La variable numero de cajas sigue una distribución Binomial.
Ha: No siguen una binomial.
Riesgo 0.10

Estimación de parámetros.

En este caso n=5 y “p” es la probabilidad de encontrar una caja malograda que es desconocida, pero se supone constante a través del proceso de control de calidad.

Estimación de p.

Promedio (x) = np

Promedio ponderado = (0x6+…+5x1) /40 = 1.775

“p” estimado es: 1.775/ 5 = 0.355

Con estos resultados se procede a los cálculos de los valores esperados,

Bajo la hipótesis planteada, que la variable X es binomial, los valores observados y esperados serian:

X	0	1	2	3	4	5	Suma
Oi	6	13	10	7	3	1	40
Ei	4.465381	12.28845	13.52682	7.444996	2.048817	0.225529	40

|Oi-Ei|²/Ei	0.527403	0.041201	0.919542	0.026598	0.441596	2.659555	4.615895
Probab.	0.111635	0.307211	0.338171	0.186125	0.05122	0.005638	1

Los calculos de la tabla se realizan utilizando la distribucion binomial:

Para x=0, P(x=0) = (1-0.355)^5 = 0.1116

y el valor esperado es 40x0.1116 = 4.465

El valor Chi cuadrada calculado es: 4.6158

El valor de la tabla para α=0.10 y gl=(6-1-1), el val;or es: 7.78

Por lo tanto la variable en estudio sigue una distribución binomial con p=0.355

Con este resultado, estime cuanto esperan gastar los productores por inspección semanal (por envío de 50 cajas) y cuanto es la perdida esperada al año (48 envíos), si solo se considera 2.5 dolares por inspeccion como costo fijo por envio (es decir $0.50 por inspeccion de una caja).

Solución:

Tabla de ocurrencias y el costos semanal por inspeccion.

X	0	1	2	3	4	5	Esperado
Costo	2.5	2.5	25	25	25	25	Semanal
P(x)	0.111635	0.307211	0.338171	0.186125	0.05122	0.005638	Dolares
Esperado	0.279086	0.768028	8.454265	4.653123	1.28051	0.140955	15.43501

El valor esperado es: 15.43 Dólares por envío, al Año serián =48*15.42 = 740.16 Dólares. Si el costo fijo anual es 2.5*48 = $ 120, entonces, la perdida esperada anual es de $ 620.16.

Solucion mediante el programa en R

```
> # Pruebas de bondad de ajuste Binomial
> x<-c(0,1,2,3,4,5)
> obs<-c(6,13,10,7,3,1)
> media<- weighted.mean(x,obs)
> n<-5
> parametro=media/n
> k<-length(x)
> gl<-k-1-1
> prob <- dbinom(x[-k], size=n,prob=parametro)
> prob<-c(prob,1-sum(prob))
> esperado <- sum(obs) * prob
> chi<-sum((obs - esperado)^2/esperado)
> chi.tabular<-qchisq(1-0.10,gl)
> tabla<-cbind(x,obs,esperado,prob)
> p.value<-1-pchisq(chi,gl)
> # resultados
> tabla
     x obs esperado      prob
[1,] 0   6  4.46538 0.1116345
[2,] 1  13 12.28845 0.3072113
[3,] 2  10 13.52682 0.3381706
```

```
[4,] 3   7  7.44500 0.1861249
[5,] 4   3  2.04882 0.0512204
[6,] 5   1  0.22553 0.0056382
> chi
[1] 4.6159
> chi.tabular
[1] 7.7794
> p.value
[1] 0.32903
```

Aplicación en la Poisson.

La distribución probabilistica tiene un solo parámetro (λ), estimado este valor, se puede hallar las probabilidades para cada caso.

El valor esperado de X = E[X] = λ, esto significa que se puede estimar lambda, a partir del promedio de la variable X observada. También se cumple que la Varianza de X es Lambda, esto implica que:

$V(X) = E[(x-E[X])^2] = E[X^2] - 2E[X\,E[X]] + (E[X])^2$, Como el esperado de una constante es igual a la constante, entonces:

$V(X) = E[X^2] - \lambda^2 = \lambda$; implica que

$E[X^2] = \lambda\,(\lambda+1)$

La función de probabilidad esta dado por:

$$P(X = k) = \frac{e^{-\lambda}\,\lambda^{k}}{k!}, \quad k=0,1,2,\ldots$$

Ejemplo 4. Una maquina, cuando funciona perfectamente puede producir una utilidad de D dólares por hora (D>2), sin embargo, esta maquina tiene una tendencia a fallar a horas inesperadas e impredecibles. La Variable numero de fallas durante un periodo de tiempo t horas por teoría se supone que sigue una POISSON, sin embargo para poder utilizar esta información , es necesario que la empresa que utiliza el equipo realice el estudio apropiado. Para tal efecto se considero el tiempo de una hora como longitud fija de tiempo, y se observo el numero de fallas ocurridas. El experimento se realizo durante 50 horas, dando los siguiente resultados.

X=Nro de Fallas	0	1	2	3	4	5 ó más
Horas Observadas	10	18	11	6	3	2

Solución:

Hp: la variable número de fallas es una variable Poisson.
Ha: No sigue una Poisson.
Riesgo 0.10

Con los datos observados de la experimentación, se obtiene el promedio ponderado que resulta 1.6 fallas por hora, se considero 5 para la clase (5 ó más):

En Poisson, solo se tiene un parámetro que es lambda (λ) y su valor esperado E[X] = λ, por lo tanto una estimación de este parámetro es el promedio de los datos, que es 1.6.

Con el valor estimado de λ, se determina las probabilidades:

y los valores esperados para las 50 horas.

X=Nro de Fallas	0	1	2	3	4	5 ó más
P(X=k)	0.2019	0.323	0.2584	0.1378	0.0551	0.0238
Esperados	10.095	16.15	12.92	6.89	2.755	1.19

El valor de Chi-cuadrada resulta:

$$\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i} = 1.1971$$

El valor tabular para (6-1-1=4) grados de libertad y alpha de 0.10 es: 7.779434

Por lo tanto, la variable aleatoria número de fallas de ala maquina, tiene una distribución Poisson con parámetro $\lambda = 1.6$

Observación. En algunos casos, en el cual la variable va al infinito, es conveniente juntar las clases o categorías tal que el valor esperado alcance un valor mayor al 5% del total observado, en nuestro caso, podría sumar las dos ultimas como una sola categoría, sin alterar los parámetros estimados previamente. Debe comprender que este proceso es para contrarrestar el efecto de los grados de libertad, a medida que los grados aumentan, el valor de la tabla disminuye y pueda incrementar un error, por otro lado no es posible tener pocas categorías, para una prueba de bondad, esta debe tener por lo menos 5 clases. Pocos grados de libertad, también incrementan un error y tiene mayor efecto que el error anterior, porque los cambios de la tabla son bruscos en los pequeños valores, no así en los valores grandes. El criterio es asumido por el investigador, lo más recomendable para los análisis y las características de los datos. Debe entender que este no es un problema matemático, sino mas bien un problema de decisión estadística.

Solución mediante el programa R.

```
> # Pruebas de bondad de ajuste Poisson
> x<-c(0,1,2,3,4,5)
> obs<-c(10,18,11,6,3,2)
> parametro <- weighted.mean(x,obs)
> k<-length(x)
> gl<-k-1-1
> prob <- dpois(x[-k], lambda=parametro)
> prob<-c(prob,1-sum(prob))
> esperado <- sum(obs) * prob
> chi<-sum((obs - esperado)^2/esperado)
> chi.tabular<-qchisq(1-0.10,gl)
> tabla<-cbind(x,obs,esperado,prob)
> p.value<-1-pchisq(chi,gl)
> # resultados

> tabla
     x obs esperado     prob
[1,] 0  10  10.0948 0.201897
[2,] 1  18  16.1517 0.323034
[3,] 2  11  12.9214 0.258428
[4,] 3   6   6.8914 0.137828
[5,] 4   3   2.7566 0.055131
[6,] 5   2   1.1841 0.023682
```

```
> chi
[1] 1.1971
> chi.tabular
[1] 7.7794
> p.value
[1] 0.32903
```

PRUEBA DE INDEPENDENCIA

La independencia está referida a la independencia estadística entre dos variables aleatorias (atributos), de una población. Para las pruebas, la muestra se tabula en un cuadro de doble entrada y se establece la frecuencia en cada celda. El cuadro así establecido se conoce como TABLA DE CONTINGENCIA.

Por ejemplo, se puede plantear la siguiente interrogante: ¿La clase de jabón que usa una persona es independiente de los ingresos que tiene? ?

En este caso se plantea la hipótesis que los ingresos de una persona es totalmente independiente de la clase de jabón que usa. La información procesada para tal fin, se presenta en el siguiente cuadro.

Ingreso	Jabón Marca 1 B1	Jabón Marca 2 B2	Total
Elevado (A1)	O11	O12	O1.
Bajo (A2)	O21	O22	O2.
Total	O.1	O.2	O..

Partiendo de la hipótesis de independencia, las frecuencias esperadas en cada celda se calcula como sigue:

Por ejemplo (2,1) que corresponde a ingreso bajo y marca de jabón 1, se obtiene de la forma siguiente:

$$(O_{2.}\ O_{.1})/O_{..}$$

Este es el resultado de multiplicar el total de observaciones por la probabilidad de la celda respectiva.

Si A y B son independientes, la probabilidad de la ocurrencia de ambos sucesos es dada por el producto de sus probabilidades marginales.

$$P(A2 \cap B1) = P(A2)\ P(B1)$$

Así, bajo el supuesto de independencia, los valores espera- dos para cada celda serían:

$$(A2, B1):\ O..(O_{2.}/O..)(O._1/O..)$$

Entonces, para la prueba respecto a la independencia, se debe plantear:

H_P : Ambos atributos son independientes
H_a : No son independientes.

Nivel de significación (riesgo) $\alpha = 0.10$

Estadístico $\chi^2 = \Sigma\Sigma \dfrac{(O_{ij} - E_{ij})^2}{E_{iJ}}$

El valor tabular (crítico) se localiza con grados de libertad igual al producto (filas-1) por (columnas-1) y el nivel α.

Para aceptar la independencia, es necesario que el valor calculado sea menor o igual al valor tabular.

Si utiliza el valor de la probabilidad de un programa de computadora, llamado **p-value**, este debe ser mayor que el nivel de riesgo (alpha)

Si rechaza la hipótesis planteada, entonces existe un grado de asociación entre las variables de estudio, para ello debe calcular este grado de asociación (Correlación de Cramer's V)

$$V = \sqrt{\frac{\chi^2 calculado}{N * \min((columnas - 1),(filas - 1))}}$$

Cuando los grados de libertad es igual a uno, se realiza la corrección de Yates.

$$\chi^2 = \Sigma\Sigma \frac{(|O_{ij} - E_{ij}| - 0.5)^2}{E_{iJ}}$$

sin embargo cuando el total de frecuencia es mayor de 50, la corrección puede obviarse

En tablas de 2X2, el cálculo se simplifica de la siguiente forma:

A	b	a + b
C	d	c + d
A+c	b + d	a + b +c + d

$$\chi^2 = \frac{(ad - bc)^2 N}{(a+b)(c+d)(a+c)(b+d)}$$

En el caso de corrección por continuidad,

$$\chi^2 = \frac{(|ad - bc| - \frac{N}{2})^2 N}{(a+b)(c+d)(a+c)(b+d)}$$

N= a+b+c+d

En el programa R la corrección es automática par grados de libertad = 1. La corrección se puede obviar si se escribe:

```
> Chisq(x,correct=FALSE)
```

Ejemplo Se realizo un experimento con la inoculación de una bacteria a 111 ratones, a 57 de ellos se aplico un antisuero. Al final del experimento se observo un total de 38 ratones muertos de los cuales 13 habían recibido el antisuero.

Construir una tabla de contingencia y realizar la prueba que la supervivencia es independiente del factor bacteria y antisuero y del factor bacteria:

Solución:

Hp: La supervivencia es independiente del Antisuero.
Ha: No hay independencia entre el antisuero y la supervivencia de los ratones

Nivel de riesgo: $\alpha = 0.10$

Tabla de valores observados.

Antisuero	Muertos	Vivos	Total
Se aplico	13	44	57
No se aplico	25	29	54
Total	38	73	111

Tabla de valores esperados.

Bacteria	Muertos	Vivos	Total
y antisuero	38*57/111	73*57/111	57
únicamente	38*54/111	73*54/111	54
Total	38	73	111

Como la suma de frecuencias es mayor de 50 se obvia la corrección:

Para el ejemplo planteado, el valor Chi cuadrado calculado es:

$$\chi^2 = \frac{(13x29 - 44x25)^2 x111}{(13+44)(25+29)(13+25)(44+29)} = 6.79$$

Completar el análisis estadístico y dar la conclusión respectiva.

Si realiza con el programa R, escriba las siguientes instrucciones en la consola.

```
> x<-matrix(c(13,25,44,29),nc=2)
> x
     [,1] [,2]
[1,]   13   44
[2,]   25   29

> sol<-chisq.test(x,correct=F)
> sol

        Pearson's Chi-squared test

data:  x
X-squared = 6.7955, df = 1, p-value = 0.009139


> sol$"observed"
     [,1] [,2]
[1,]   13   44
[2,]   25   29

> sol$"expected"
       [,1]   [,2]
[1,] 19.514 37.486
[2,] 18.486 35.514

> sol$"statistic"
```

```
X-squared
   6.7955

> chi.tabular<-qchisq(1-0.10,1)
> chi.tabular
[1] 2.7055

> sol$"p.value"
[1] 0.0091385
```

```
> V<-sqrt(sol$"statistic"/(sum(x)*min((ncol(x)-1),(nrow(x)-1))))
> names(V)<-"Cramer's V"
> V
Cramer's V
  0.24743
```

Conclusión la supervivencia de los ratones no es independiente del ansiuero aplicado, el grado se asociación es 0.24

Problema. Un ecólogo estudió 100 árboles pertenecientes a una especie no conocida, en un área de 400 millas cuadradas. Se registró si cada árbol estaba enraizado en el suelo en forma de serpentina o no, y si sus hojas tenían pelusa o son lisas.

Suelo en …	Con Pelusa	Lisa
Serpentina	12	22
No serpentina	16	50

Probar la hipótesis que el tipo de hoja es independiente de la forma de enraizado en el suelo.

Ejercicios:

1. Frecuencia de ejemplares de una especie de escarabajo (Cicindela fulgida) recogidos en diversas estaciones.

	COLOR DEL EJEMPLAR	
ESTACION	Rojo brillante	Rojo no brillante
Principios de primavera	29	11
Finales de primavera	273	191
Principios de verano	8	31
Finales de verano	64	64

   Probar si son independientes el color del ejemplar y la estación.

2. Estudiar la asociación entre las reacciones a la Lepromina y a la tuberculina, consecutivas a la vacunación con BCG, en nativos.

Reacción a la Lepromina	Reacción a la Tuberculina	
	Positivo	Negativo
Positivo	95	10
Negativo	48	24

PRUEBA DE HOMOGENEIDAD DE MUESTRAS

Está referido a la semejanza en la distribución de muestras, es decir si las muestras pertenecientes a dos poblaciones o más, tienen una distribución similar, es decir si son homogéneas según ciertas categorías o clases. Por ejemplo:

Se hace un test de rendimiento a estudiantes de dos colegios distintos. Se toma, muestras de n_1 = 100 y n_2 = 200 estudiantes de cada colegio y se contabiliza según nota y colegio:

Nota		Colegio 1	Colegio 2
P_1	A	n_{11}= 10	n_{12}= 20
P_2	B	n_{21}= 20	n_{22}= 20
P_3	C	n_{31}= 30	n_{32}= 80
P_4	D	n_{41}= 20	n_{42}= 60
P_5	E	n_{51}= 20	n_{52}= 20

Se formula la siguiente pregunta:

¿ Existe alguna diferencia en la distribución de notas entre los dos colegios o ambas muestras provienen de la misma población?.

Se plantea la siguiente hipótesis planteada (nula):

Hp: Las muestras son homogéneas
Ha: No hay homogeneidad de muestras.
Riesgo 0.05

La palabra homogéneo se utiliza en estadística para indicar "lo mismo" ó "igual". Así pues, estamos interesados en comprobar si la distribución de los datos se presenta en forma similar (homogéneo). A esta prueba se le denomina Prueba de Homogeneidad de muestras.

Según la hipótesis planteada, se supone que hay una sola distribución de probabilidades del atributo de la población en cuestión, cuyas probabilidades serían:

$$p_1 = \frac{(n_{11}+n_{12})}{n_{..}} \quad p_2 = \frac{(n_{21}+n_{22})}{n_{..}} \quad p_3 = \frac{(n_{31}+n_{32})}{n_{..}} \quad ....$$

La Prueba Estadística.

Según la hipótesis planteada, si es verdad que las muestras proceden de la misma población, significa que ambas muestras tienen la misma distribución de probabilidades dada por: $p_1, p_2,...,p_s$.

Los valores esperados son calculados como: $n_{.j}\, p_i$

y el valor calculado de Chi cuadrado para la prueba dado por:

$$\chi^2 = \Sigma\Sigma \frac{(n_{ij} - n_{.j}\, p_i)^2}{n_{.j}\, p_i}$$

Para la comprobación, el valor de X^2 (tabular) se busca con α = 0.05 según el problema y grados de libertad es igual a: (r-1)(s-1).

r = número de muestras
s = número de Categorías

Como los p_i son desconocidos y estimados mediante la relación:

$p_i = n_i. /n..$ y constituye la distribución de la muestra que estima a la distribución teórica.

Ejercicio. Para el ejemplo, hallar la distribución estimada de las muestras y comprobar que X^2(calculado)=14.3181 y X^2(tabular)=9.48. Con estos resultados se concluye que no hay evidencia estadística para afirmar que son homogéneas.

Solución mediante el programa R.

```
> colegio1 <- c(10,20,30,20,20)
> colegio2 <- c(20,20,80,60,20)
> categorias<- c("A","B","C","D","E")
> colegio <- colegio1 + colegio2
> p<-colegio/sum(colegio)
> esperado1 <- p*sum(colegio1)
> esperado2 <- p*sum(colegio2)
> obs<-c(colegio1,colegio2)
> esp <-c(esperado1,esperado2)
> gl<- (5-1)*(2-1)
> chi<-sum((obs - esp)^2/esp)
> chi.tabular<-qchisq(1-0.05,gl)
> tabla.obs<-cbind(colegio1,colegio2)
> tabla.esp<-cbind(esperado1,esperado2)
> names(p)<-categorias
> dimnames(tabla.obs)<-list(categorias,c("col 1", "col 2"))
> dimnames(tabla.esp)<-list(categorias,c("col 1", "col 2"))
> p.value<-1-pchisq(chi,gl)
> # resultados
> tabla.obs
  col 1 col 2
A    10    20
B    20    20
C    30    80
D    20    60
E    20    20
> tabla.esp
   col 1  col 2
A 10.000 20.000
B 13.333 26.667
C 36.667 73.333
D 26.667 53.333
E 13.333 26.667

> p
      A       B       C       D       E
0.10000 0.13333 0.36667 0.26667 0.13333
> chi
[1] 14.318
> chi.tabular
[1] 9.4877
> p.value
[1] 0.0063458
```

PRUEBA DE HOMOGENEIDAD DE VARIANCIA

Es útil para estudio de varias poblaciones, si presentan variancias homogéneas, de ser así, permite el estudio comparativo de promedios mediante el análisis de variancia.

Las hipótesis:

Hp: Los grupos tienen variancias homogéneas.
Ha: No presentan homogeneidad de variancia.
Riesgo de 0.10

Uno de los métodos es el TEST de BARTLETT que consiste en:

Las variancias S^2_i son las variancias del error en cada grupo (Tratamiento), se tiene "t" tratamientos con n_i repeticiones.

Se calcula la variancia común S^2 como un promedio ponderado de las variancia S^2_i.

$$S^2 = \frac{\Sigma(n_i - 1)S^2_i}{\Sigma(n_i - 1)}$$

Calcular el valor Chi cuadrado

$$\chi^2 = \frac{\Sigma(n_i - 1)\mathrm{Ln}(S^2) - \Sigma(n_i - 1)\mathrm{Ln}(S^2_i)}{C}$$

donde C es igual a:

$$C = 1 + \frac{1}{3(t-1)}\left[\Sigma\frac{1}{n_i - 1} - \frac{1}{\Sigma(n_i - 1)}\right]$$

El valor de χ^2 se distribuye como una Chi cuadrada con (t-1) grados de libertad.

Donde “t” es el numero de grupos.
Para la aceptación de la hipótesis planteada, el valor de χ^2 Calculado debe ser menor o igual al valor tabular χ^2 o, el p-value mayor o igual al nivel de riesgo planteado, por lo general es 0.10.

Aplicación: 4 grupos y 4 muestras por grupo (t=4 y n=4) probar la homogeneidad:

$$S^2_1 = 0.06\ ;\ S^2_2 = 0.5257\ ;\ S^2_3 = 0.48\ ;\ S^2_4 = 0.74$$

S^2 = 0.4517

C = 1.1388

χ^2 = 3.93 / 1.1388 = 3.45

χ^2 (α=0.05, gl=3) = 7.81

Se acepta la hipótesis que las variancias son homogéneas.

Para realizar la homogeneidad de variancia con el programa R necesita tener los valores originales. Para que el sistema calcule las variancias y realice la prueba.

Por ejemplo se tiene tres grupos A, B y C con 8 muestras cada uno.

A	A	A	A	A	A	A	A	B	B	B	B	B	B	B	B	C	C	C	C	C	C	C	C
4	7	2	9	0	3	5	9	5	6	1	8	6	3	6	8	4	8	3	7	4	4	4	7

Mediante procesos manuales:

	Var	LN(var)
A	10.69643	2.36991
B	5.696429	1.739839
C	3.553571	1.267953
Promedio	6.64881	
Suma		5.377702

Estadisticas para la prueba:

ChiSq	2.139277
C =	1.063492
Chi.corregido	2.011559
Chi.tabular	4.605176
p.value	0.365759

Con esta información se concluye que hay homogeneidad de variancia.
El p.value es mayor de 0.10, también puede usar el valor tabular para la toma de decisiones, Chi.corregido es menor que Chi.tabular al nivel de riesgo de 0.10

Compara estos resultados con el programa R.

Mediante el programa R.

```
>resp<-c(4,7,2,9,0,3,5,9,5,6,1,8,6,3,6,8,4,8,3,7,4,4,4,7)
>grupos<-c(rep("A",8),rep("B",8),rep("C",8))
>bartlett.test(resp, grupos)

        Bartlett test for homogeneity of variances

data:  resp and grupos
Bartlett's K-squared = 2.0116, df = 2, p-value = 0.3658
>
># o tambien:
>grupos <-as.factor(grupos)
>bartlett.test(resp ~ grupos)

        Bartlett test for homogeneity of variances
```

```
data:  resp by grupos
Bartlett's K-squared = 2.0116, df = 2, p-value = 0.3658
```

Según estos resultados, se acepta la hipotesis planteada, es decir se confirma la homogenidad de variancia

PROBLEMAS

1. Se eligió al azar una muestra aleatoria de 375 amas de casa para averiguar su opinión con respecto a las bondades que ofrece un nuevo producto de cocina, los resultados se expresan a continuación:

Opinión	Muy bueno	Bueno	Regular	Malo	Muy Malo
Frecuencia	20	60	180	105	10

   Probar estadísticamente si la opinión de las amas de casa con respecto al nuevo producto se distribuye según la relación 1:3:5:4:1 . Usar α = 0.05

2. El número de llamadas telefónicas a una central, por lo general, sigue una ley Poisson. En un estudio anterior para esta central telefónica se determinó que efectivamente sigue una distribución de Poisson con una tasa de 2 llamadas por hora. Pasado un buen tiempo de ese estudió, se observó en 100 horas el número de llamadas por hora y se estableció la frecuencia de la siguiente forma:

Nro. de llamadas	0	1	2	3	4 ó más
Observadas	14	28	35	15	8

   ¿Puede afirmar qué el número de llamadas sigue una distribución de probabilidad Poisson ?. De aceptar la hipótesis, ¿cuál es el estimado de la media y la variancia del número de llamadas ?.

   Use un riesgo de 0.10

3. En un estudio de Biología se observó la variación de planta a planta en un experimento sobre el color del albumen en las semillas. Los resultados para las primeras diez plantas fueron:

Planta	1	2	3	4	5	6	7	8	9	10
Albumen amarillo	25	32	14	70	24	20	32	44	50	44
Albumen verde	11	7	5	27	13	6	13	9	14	18

   ¿ Existe suficiente evidencia estadística para afirmar que la razón entre el albumen amarillo a albumen verde es de 3:1 ?. Use α = 0.10

4. En un examen de ingreso a la universidad se observó que las variancias muestrales de las notas de muestras de postulantes en diferentes materias fueron: 4.02, 3.75, 5.08, 2.5, 7.6 si es tamaño de muestra fue de 15, 18, 14, 26, 13 respectivamente, ¿Puede afirmar que las variancias son homogéneas ?.

Información para la instalación y uso del programa R.

El programa R se encuentra en la siguiente dirección:

http://www.r-project.org/

Ingrese a la opción: Download CRAN

seleccione un país y una universidad, por ejemplo:

http://cran.us.r-project.org

luego:

Windows (95 and later)

Finalmente ejecute el lugar donde esta la base de R, que es:

base

luego seleccionar el archivo ejecutable de instalacion de R

Una vez copiado a su disco duro, proceda a la instalación.

Sugerencia.

En la pagina de R encontrara toda la documentación que necesita.

El la siguiente paguina encontrara algunos ejemplos de R para uso de los estudiantes de la Universidad Agraria La Molina.

http://tarwi.lamolina.edu.pe/~fmendiburu/

Ejercicios de Poisson.

Caso 1: Parametro desconocido
Conteo de nematodos vistos en una demostración de 60 observadores con el microscopio.

No. nematodes (x):	0	1	2	3	4	5	6
Freqs. Observada(60):	3	12	17	13	9	3	3

```
# Pruebas de bondad de ajuste Poisson
x<-0:6
obs<-c(3,12,17,13,9,3,3)
parametro <- weighted.mean(x,obs)
k<-length(x)
gl<-k-1-1
prob <- dpois(x[-k], lambda=parametro)
prob<-c(prob,1-sum(prob))
esperado <- sum(obs) * prob
chi<-sum((obs - esperado)^2/esperado)
chi.tabular<-qchisq(1-0.10,gl)
tabla<-cbind(x,obs,esperado,prob)
p.value<-1-pchisq(chi,gl)
# resultados

> tabla
     x obs  esperado       prob
[1,] 0   3  4.607465 0.07679109
[2,] 1  12 11.825828 0.19709713
[3,] 2  17 15.176479 0.25294132
[4,] 3  13 12.984321 0.21640535
[5,] 4   9  8.331606 0.13886010
[6,] 5   3  4.276891 0.07128152
[7,] 6   3  2.797410 0.04662350
> chi
[1] 1.232022
> chi.tabular
[1] 9.236357
> p.value
[1] 0.9417687
```

Caso 2: Parametro conocido.
Probar que la distribucion del numero de nematodos visto bajo el microscopio sigue una Poisson con un numero esperado de 3 nematodos.

Se hicieron 60 observaciones bajo el microscopio

Hp Nro. De nematodos sigue una Poisson con Lambda = 3

No. nematodes (x):	0	1	2	3	4	5	6
Freqs. Observada(60):	3	12	17	13	9	3	3

```
# Pruebas de bondad de ajuste Poisson
x<-0:6
obs<-c(3,12,17,13,9,3,3)
parametro <- 3
```

```
k<-length(x)
gl<-k-1
prob <- dpois(x[-k], lambda=parametro)
prob<-c(prob,1-sum(prob))
esperado <- sum(obs) * prob
chi<-sum((obs - esperado)^2/esperado)
chi.tabular<-qchisq(1-0.10,gl)
tabla<-cbind(x,obs,esperado,prob)
p.value<-1-pchisq(chi,gl)
# resultados
> tabla
     x obs  esperado       prob
[1,] 0   3  2.987224 0.04978707
[2,] 1  12  8.961672 0.14936121
[3,] 2  17 13.442508 0.22404181
[4,] 3  13 13.442508 0.22404181
[5,] 4   9 10.081881 0.16803136
[6,] 5   3  6.049129 0.10081881
[7,] 6   3  5.035077 0.08391794
> chi
[1] 4.461775
> chi.tabular
[1] 10.64464
> p.value
[1] 0.6144437
```