



PRUEBAS DE HIPÓTESIS JI CUADRADO

Contenidos

- 12.1. Introducción
- 12.2. Pruebas paramétricas *versus* Pruebas no paramétricas
- 12.3. Experimentos multinomiales
- 12.4. Distribución de probabilidad Ji cuadrado
- 12.4. Clasificación de las pruebas Ji cuadrado
- 12.5. Estadígrafo Ji cuadrado
- 12.6. Ejemplos de aplicación de las pruebas Ji cuadrado

ANEXO: Tabla de la función de distribución de probabilidad acumulada de Ji cuadrado

12.1. INTRODUCCIÓN

Este capítulo está destinado a presentar un grupo de pruebas estadísticas, que tienen como denominador común la utilización de un estadígrafo de prueba denominado Ji cuadrado, simbolizado por tradición con la letra griega Ji “elevada” al cuadrado, esto es χ^2 . En el muestreo repetitivo, el estadígrafo χ^2 se comporta como una variable aleatoria denominada Ji cuadrado, que por consistencia con lo indicado para variables aleatorias se simbolizará como X^2 . El comportamiento del estadígrafo Ji cuadrado en el muestreo, así como el de una variable aleatoria X^2 es modelado por una distribución continua de probabilidades, denominada distribución Ji cuadrado.

Las pruebas de hipótesis Ji cuadrado son aplicables a variadas situaciones problemáticas, Los cuatro tipos de pruebas Ji cuadrado que se abordarán, en correspondencia a las preguntas que llevarán a su empleo, son los siguientes:

- 1) **Prueba para una varianza:** ¿una varianza poblacional es igual a otra de valor conocido?
- 2) **Prueba de bondad de ajuste:** ¿una distribución de frecuencias empíricas es significativamente diferente de la distribución esperada?
- 3) **Prueba de independencia:** ¿la clasificación de acuerdo a un atributo es independiente de la clasificación con respecto a otro?
- 4) **Prueba de homogeneidad:** ¿se puede considerar que un grupo de k muestras procede de una misma población?

12.2. PRUEBAS NO PARAMÉTRICAS *VERSUS* PRUEBAS PARAMÉTRICAS

Hasta ahora se han aplicado pruebas de hipótesis utilizando los estadígrafos de prueba “z” y “t”, En tales casos, las pruebas han coincidido en las siguientes características:

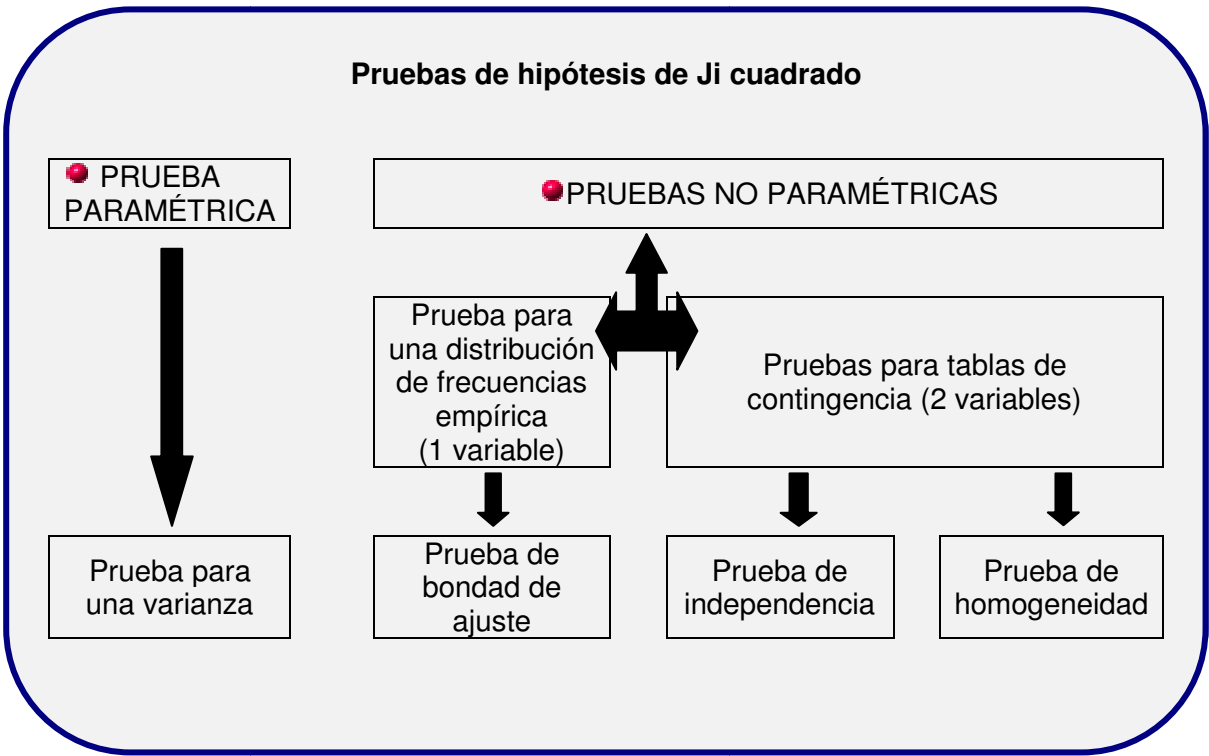
- a) se ha requerido suponer que las muestras eran aleatorias.
- b) se ha tenido datos empíricos de variables medidas en una escala de intervalo o de razones (cuando se trató de una variable cualitativa dicotómica, se usaron proporciones).
- c) se han postulado hipótesis referidas a parámetros: μ , $\mu_1 - \mu_2$, π y π_1 y π_2 .
- d) se ha establecido el cumplimiento de supuestos con relación a las distribuciones de las poblaciones originales de donde se han extraído las muestras (normales o estudentizadas).
- e) a la hora de definir la regla de decisión se presupuso una distribución teórica subyacente para explicar el comportamiento del estadígrafo de prueba en el muestreo repetitivo que fueron: la distribución normal y la distribución T de Student.

En capítulos siguientes se verán otras aplicaciones de la prueba T de Student y de una nueva prueba llamada F de Fisher. Todas ellas pertenecen al grupo de la **pruebas paramétricas**, que en el campo de las ciencias empíricas son las más utilizadas.

Existe por otra parte, el grupo de las denominadas **pruebas no paramétricas**, que se diferencian de las anteriores en lo siguiente:

- a) no necesitan cumplimentar supuestos exigentes sobre las poblaciones de las que se extraen las muestras
- b) además de lo visto para el caso paramétrico, se pueden aplicar a variables cualitativas
- c) las hipótesis no se plantean en relación directa a parámetros

Las pruebas de Ji cuadrado indicadas en la introducción pertenecen a ambos grupos; al paramétrico en el caso de la varianza, y al no paramétrico en el caso restante, que comprende pruebas referidas a un modelo probabilístico (Bondad de ajuste) y pruebas referidas a tablas de contingencia (independencia y homogeneidad).



12.3. EXPERIMENTOS MULTINOMIALES

Al presentar la distribución binomial, se definió el experimento binomial y se analizaron recuentos asociados a variables dicotómicas, con aplicaciones a casos como los siguientes: a) bioensayo para investigar la acción de un insecticida en grupos de 50 pulgones (insecto vivo-insecto muerto), b) relevamiento a campo para evaluar el desarrollo de plantas parásitas sobre una especie nativa de interés en 80 cuadrículas (con parasitismo-sin parasitismo) y c) medición de la calidad de un proceso industrial a través de muestreos con recuento del número de unidades defectuosas. De modo análogo se dan muchas situaciones donde el interés está puesto en clasificar las unidades de análisis en k categorías, por ejemplo: se ha elaborado un producto con cuatro formulaciones diferentes y se realiza un ensayo de evaluación sensorial con consumidores, quienes deben elegir cual es la formulación preferida, en época de elecciones se realiza una encuesta a 1000 personas que deben indicar el partido al que pertenece el candidato que van a votar para gobernador, o se clasifican los alumnos de una muestra según el tipo de estudio con que ingresaron a la universidad y la condición lograda al finalizar el primer año en matemática (aprobado, regular, libre)..Los ejemplos dados tienen, con cierta aproximación, las características que definen a un *experimento multinomial*.

Def. 12.1. Un experimento multinomial es aquél que:

- a) consta de n pruebas idénticas,
- b) el resultado de cada prueba se localiza en una de las k categorías,
- c) la probabilidad p_i de que un resultado de una prueba se localice en la i-ésima categoría, es constante de una prueba a otra.
Nótese lo siguiente: $p_1 + p_2 + p_3 + \dots + p_k = 1$, siendo $i = 1, 2, 3, \dots, k$.
- d) las pruebas son independientes.
- e) interesan las frecuencias $n_1, n_2, n_3, \dots, n_k$, donde n_i ($i=1, 2, \dots, k$) es igual al número de pruebas en las cuales el resultado se clasifica en la i-ésima categoría.
Nótese que $n_1 + n_2 + n_3 + \dots + n_k = n$.

Nótese la similitud entre los experimentos binomial y multinomial. En particular, el experimento binomial representa el caso especial del experimento multinomial donde $k=2$. Mientras que las dos probabilidades, p y q , del experimento binomial están representadas por las k probabilidades (p_1, p_2, \dots, p_k) , asociadas a las k categorías derivadas de un experimento multinomial.

Otra concepción práctica para comprender de modo general a los experimentos multinomiales, es hacer una analogía con un experimento de lanzamiento de n bolas donde se dispone de k cajas, de modo que toda pelota lanzada caerá en alguna de las k cajas. El experimento se repite n veces (n tiradas) tal que la probabilidad de que una pelota caiga en una caja varía de una caja a otra, pero permanece constante a lo largo del experimento para cada caja en particular, además los lanzamientos se hacen en forma independiente. En tal caso al finalizar el experimento, resultarán n_1 pelotas en la primera caja, n_2 en la segunda, \dots , y n_k en la k -ésima caja, donde el número total de pelotas es igual a $\sum n_i = n$, siendo $i=1, 2, \dots, n$.

12.4. DISTRIBUCIÓN DE PROBABILIDAD JI CUADRADO

Si se recuerda, al presentar el capítulo destinado a las distribuciones continuas de probabilidad, se estableció que la distribución Ji cuadrado pertenece a tal grupo de distribuciones probabilísticas.

En términos generales su función de densidad $f(x;V)$ se deriva como un caso muy especial de la distribución continua de probabilidades gamma, Γ , y está totalmente definida por un único parámetro, que son los grados de libertad, V .

Def. 12.2. Función de densidad de probabilidad para la distribución Ji cuadrado

$$f(x; \nu) = \begin{cases} \frac{1}{\Gamma\left(\frac{\nu}{2}\right) 2^{\frac{\nu}{2}}} x^{\frac{\nu}{2}-1} e^{-\frac{x}{2}} & ; \quad 0 \leq X^2 \leq \infty \\ 0, \text{ en cualquier otro caso} \end{cases}$$

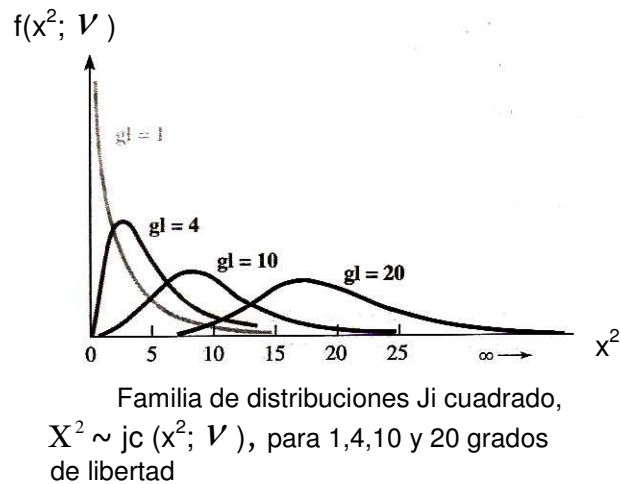
Donde: Γ es la función gamma y V los grados de libertad

La distribución Ji cuadrado es fundamental en inferencia estadística, ya que modela la distribución de la variable aleatoria “suma de los cuadrados de n variables independientes”, lo que permite su utilización en las pruebas de hipótesis que se tratarán en este capítulo. Por tal razón, en este contexto se utilizará la expresión $f(x^2; V)$, en lugar de la más general $f(x; V)$.

Es importante notar que, al igual que lo visto en el caso de la variable T de Student, en realidad se trata de una **familia de distribuciones Ji cuadrado**: existe una distribución Ji cuadrado distinta para cada número de grados de libertad, V , por tanto existen infinitas distribuciones posibles. Pero, a diferencia de la función estudentizada que como la normal tipificada, siempre es simétrica con respecto a su centrado en $\mu_t = 0$, la distribución Ji cuadrado es asimétrica positiva.

Propiedades de la función de densidad Ji cuadrado:

- 1) La variable no toma valores negativos, su campo de variación (R_{x^2}) es igual a $0 \leq X^2 \leq \infty$.
- 2) La función $f(x^2; V)$ es ≥ 0 .
- 3) Por ser una función de densidad, el área bajo una curva Ji cuadrado y sobre el eje horizontal tiene un valor unitario.



Además, como se muestra gráficamente, la función de densidad de probabilidad de una variable aleatoria Ji cuadrado, X^2 , es:

- a) unimodal,
- b) marcadamente asimétrica con sesgo positivo, es decir con cola a la derecha, cuando el número de grados de libertad es muy pequeño. Conforme aumentan los grados de libertad, se hace menos sesgada y para 20 grados de libertad resulta bastante simétrica. A partir de Para $V \geq 30$, la distribución se considera aproximadamente normal.

Tabla de la función de distribución de probabilidad acumulada para una variable Ji cuadrado:

Los valores las áreas de probabilidad acumulada desde $X^2 = 0$, hasta los percentiles x^2_{α} más utilizados en las pruebas de hipótesis se encuentran tabulados (Ver Tabla correspondiente en Anexo).

Mediante la Tabla de la función de distribución acumulada, $F(x^2; V)$, se pueden resolver problemas del tipo siguiente: ¿cuál es la probabilidad de encontrar valores mayores a cierto x^2_i ?; ¿Qué proporción del área de probabilidad se encuentra a la izquierda de cierto x^2_i ?; ¿Qué valor de la variable X^2 es superado solamente por el 10% de los datos posibles?.

12.5. CLASIFICACIÓN DE LAS PRUEBAS DE HIPÓTESIS JI CUADRADO

Hasta ahora se han resuelto problemas de inferencia estadística referidos a medias poblacionales y proporciones. Como se anticipara, las pruebas de Ji cuadrado consideradas en este capítulo, serán de tipo paramétrico para la varianza poblacional y, de tipo no paramétrico con tres tipos de objetivos diferentes para situaciones en las que se tienen disponibles datos de frecuencias. A partir de estos dos tipos de pruebas Ji cuadrado se formularán, correspondientemente, hipótesis en términos de un parámetro, específicamente σ^2 , y otro tipo de hipótesis con un formato diferente como son las siguientes: $H_0 : X^2 \sim N(x; \mu, \sigma^2)$, o bien $H_0 : \pi_{ij} = \pi_i \pi_j$. Ver el Cuadro 12.1. Análisis comparativo de las Pruebas de Ji cuadrado.

Cuadro 12.1. Análisis comparativo de las Pruebas de Ji cuadrado.

Prueba de hipótesis	Objetivo	Hipótesis
P. para una varianza	Interesa determinar si una varianza poblacional es igual a otra conocida.	a) P. unilateral por derecha(izquierda) $H_0 : \sigma^2 = \sigma_0^2$ $H_1 : \sigma^2 > \sigma_0^2$ (o bien $H_1 : \sigma^2 < \sigma_0^2$) b) P. bilateral $H_0 : \sigma^2 = \sigma_0^2$ $H_1 : \sigma^2 \neq \sigma_0^2$
P. de bondad de ajuste	Caso a: Interesa determinar si los datos disponibles de una muestra aleatoria univariada de tamaño n provienen de una población que tiene una distribución de probabilidad conocida .	a) Distribución binomial $H_0 : X \sim B(x; n, p)$ $H_1 : X$ sigue otra distribución b) Distribución Poisson $H_0 : X \sim P(x; \lambda)$ $H_1 : X$ sigue otra distribución c) Distribución normal $H_0 : X \sim N(x; \mu, \sigma^2)$ $H_1 : X$ sigue otra distribución
	Caso b: Interesa determinar si los datos disponibles de una muestra aleatoria univariada de tamaño n provienen de una población que tiene una distribución de probabilidad específica	$H_0 : \pi_1 : \pi_2 : \dots : \pi_k$ Por ejemplo: $H_0 : \frac{9}{16} : \frac{3}{16} : \frac{3}{16} : \frac{1}{16}$ $H_1 : \text{las } k \text{ probabilidades se interrelacionan de otra manera}$
P. de independencia	Interesa determinar para una muestra aleatoria bivariada, de tamaño n, si la clasificación según una de las variables es independiente de la clasificación según la otra variable.	$H_0 : \pi_{ij} = \pi_i \pi_j$; para todo (i,j) $H_1 : \pi_{ij} \neq \pi_i \pi_j$ para al menos un (i,j)
P. de homogeneidad	Interesa determinar si los datos correspondientes a dos o más muestras aleatorias, clasificadas según dos variables, se distribuyen probabilísticamente de la misma manera.	$H_0 : \pi_{1j} = \pi_{2j} = \dots = \pi_{rj}$; para todo j $H_1 : \text{al menos un } \pi_{ij} \text{ diferente}$

12.6. ESTADÍGRAFO JI CUADRADO

En los análisis inferenciales a considerar, surgirán dos formas posibles para el estadígrafo Ji Cuadrado, que se denotará como χ^2 , de acuerdo al tipo de prueba.

a) Prueba para la varianza

Al realizar una prueba referente a la media poblacional, se ha visto que se utiliza como estadígrafo de prueba a la media muestral. Esto es posible porque la distribución del estadígrafo media y el estadígrafo diferencia de medias, tiene en el muestreo repetitivo una distribución probabilística que es conocida. Esto no ocurre en el caso de la varianza.

Si se extrae una muestra aleatoria de una población que tiene distribución normal, con media μ y varianza σ^2 , se conoce que la varianza muestral, s^2 , calculada como $\sum_{i=1}^n \left(x_i - \bar{x} \right)^2 / n-1$, puede ser utilizada como estimador de la varianza poblacional σ^2 . Pero para poder sustentar probabilísticamente inferencias (estimaciones intervalares y pruebas de hipótesis) relacionadas con σ^2 , se requiere conocer la distribución del estadígrafo en el muestreo repetitivo, y resulta que la s^2 sigue una distribución de probabilidad con una media igual a σ^2 , pero su forma es asimétrica y depende del tamaño muestral. Afortunadamente existe un modo de simplificar el problema, que consiste en tipificar el valor de la varianza muestral a través de una transformación (recordar que para la media muestral, la transformación z permitió utilizar las tablas de la normal). La transformación es $\frac{(n-1)s^2}{\sigma^2}$, donde n es el tamaño de una muestra aleatoria, s^2 es la varianza muestral y σ^2 es la varianza hipotética de la población.

El estadígrafo definido, tiene una distribución muestral que sigue la distribución Ji cuadrado con n-1 grados de libertad, por lo que se lo denomina estadígrafo Ji cuadrado, χ^2 .

Estadígrafo Ji cuadrado para la prueba de hipótesis de la varianza	$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$, donde χ^2 se comporta como una variable aleatoria $X^2 \sim \chi^2_c(x^2; V)$
---	---

Los valores críticos para el estadígrafo Ji cuadrado, χ^2_c , se obtienen en la Tabla de la función $F(x^2; V)$, ingresando por filas con los grados de libertad ($V = n-1$) y, por columnas con la probabilidad $1-\alpha$ siendo $\alpha = P(X^2 > \chi^2_c)$.

Es importante advertir acerca de que este tipo de prueba de hipótesis presupone que los datos de la población se distribuyen de forma normal. Lamentablemente, esta prueba es sensible a desviaciones de esta suposición, por lo que si la población no tiene distribución normal, y en especial si las muestras son de tamaño pequeño, la exactitud de la prueba puede resultar seriamente afectada.

b) Pruebas con datos de frecuencias

En pruebas de hipótesis relacionadas con una distribución de frecuencias (P. de bondad de ajuste) o bien con tablas de contingencia (P. de independencia y P. de homogeneidad), el estadígrafo de prueba Ji cuadrado, χ^2_c , responde a la siguiente fórmula:

$\chi^2 = \sum_{i=1}^k \left[\frac{(n_i - \hat{n}_i)^2}{\hat{n}_i} \right]$ siendo $i= 1, 2, \dots ,k$.	<p>El valor muestral de χ^2 es igual a la suma de k cocientes con numerador “el cuadrado de la diferencia entre la i-ésima frecuencia observada y su correspondiente frecuencia teórica o esperada”, y con denominador igual a esta última frecuencia, calculada como:</p> <p>$\hat{n}_i = (\text{probabilidad de la } i - \text{ésima clase}) \times (\text{tamaño muestral})$</p> <p><u>Notar:</u> la suma afecta a k cocientes, esto es $\sum_{i=1}^k (\text{cocientes})$, NO al numerador de un cociente.</p>
--	---

Estadígrafo Ji cuadrado para pruebas de hipótesis con datos de frecuencias de una muestra	$\chi^2 = \sum_{i=1}^k \left[\frac{(n_i - \hat{n}_i)^2}{\hat{n}_i} \right], \text{ donde } \chi^2 \text{ se comporta como una variable aleatoria } \chi^2 \sim \text{jc}(\chi^2; V)$
--	---

La construcción del estadígrafo requiere un raciocinio simple:

- a) el valor de Ji cuadrado calculado proviene de la suma de varios números: uno por cada categoría
- b) el numerador de cada término en la fórmula es igual al cuadrado de la diferencia entre las frecuencias observadas y estimadas para cada una de las categorías o celdas. Cuanto más cercanos estén éstos valores, tanto más pequeño es el valor de $(n_i - \hat{n}_i)^2$; y cuanto más distantes, tanto más grande es su valor. El denominador de cada celda pone en perspectiva el tamaño del denominador. Es decir, una diferencia $(n_i - \hat{n}_i)$ igual a 10, como resultado de la diferencia entre frecuencias de 110 y 100, es muy distinta de una que proviene de la diferencia entre 15 y 5. Estas ideas indican que los valores pequeños del estadígrafo Ji cuadrado χ^2 , señalan concordancia entre los dos conjuntos de frecuencias, mientras que los grandes implican discrepancia. De modo que es común que estas pruebas sean de una sola cola, con la región crítica a la derecha.

El estadígrafo de prueba Ji cuadrado fue propuesto en 1900 por Karl Pearson, como una función de los cuadrados de las desviaciones entre las frecuencias observadas y sus respectivos valores esperados, ponderados por el recíproco de sus valores esperados. La demostración matemática está fuera del alcance de este curso, basta saber que se puede demostrar que *el estadígrafo Ji cuadrado χ^2 , en el muestreo repetitivo sigue una distribución que se puede aproximar con una distribución de probabilidad de la variable aleatoria Ji cuadrado, χ^2 , para n grande ($n \geq 50$) y si las frecuencias esperadas para las k categorías son iguales o mayores a 5.*

Los valores críticos para el estadígrafo Ji cuadrado, χ^2_c , se obtienen como en el caso anterior de la Tabla de la función $F(\chi^2; V)$, ingresando por filas con los grados de libertad (V) y, por columnas con la probabilidad $1-\alpha$ siendo $\alpha = P(X^2 > \chi^2_c)$. La fórmula general para el cálculo de los grados de libertad es la siguiente:

$V = k - p - 1;$

donde k es el número de categorías, y p es el número de parámetros que se necesita estimar.

En el cuadro 12.2. se presenta un resumen para cada caso en particular en el Cuadro 12.2. Análisis comparativo para el cálculo de los grados de libertad en las Pruebas de Ji cuadrado.

Prueba de la varianza	Prueba de bondad de ajuste	Pruebas con datos de tablas de contingencia (rxc)	
		Prueba de independencia	Prueba de homogeneidad
$V = n - 1$	Ajustamiento del modelo binomial, $X \sim b(x;n,\pi)$ a) parámetro π conocido $V = k - p - 1 = k-0-1 = \mathbf{k-1}$ b) parámetro π desconocido $V = k - p - 1 = k-1-1 = \mathbf{k-2}$	$V = (r-1)(c-1)$ r: nº de filas c: nº de columnas	$V = (r-1)(c-1)$ r: nº de filas c: nº de columnas
	Ajustamiento del modelo Poisson, $X \sim p(x;\lambda)$ a) parámetro λ conocido $V = k - p - 1 = k-0-1 = \mathbf{k-1}$ b) parámetro λ desconocido $V = k - p - 1 = k-1-1 = \mathbf{k-2}$		
	Ajustamiento del modelo normal, $X \sim n(x;\mu,\sigma)$ a) parámetros μ y σ conocidos $V = k - p - 1 = k-0-1 = \mathbf{k-1}$ b) un parámetro desconocido (μ o bien σ) $V = k - p - 1 = k-1-1 = \mathbf{k-2}$ c) los dos parámetros desconocidos $V = k - p - 1 = k-2-1 = \mathbf{k-3}$		

12.7. EJEMPLOS DE APLICACIÓN DE LAS PRUEBAS DE JI CUADRADO

12.7.1. Prueba de una hipótesis concerniente a una varianza (o desviación típica) poblacional

Al analizar una variable cuantitativa en una muestra, frecuentemente el interés se centra en estimar el valor de la media y de la varianza o desviación típica, pero también suele interesar comprobar el valor de la evidencia muestral a partir del planteo de alguna hipótesis paramétrica. Las pruebas para la media μ ya fueron consideradas, ahora se presentará el caso para la varianza σ^2 .

Se parte del concepto que la varianza muestral, al igual que la media muestral, es una variable aleatoria. Por muestreo aleatorio repetitivo aplicado a una población normal con media μ y varianza σ^2 , resulta que


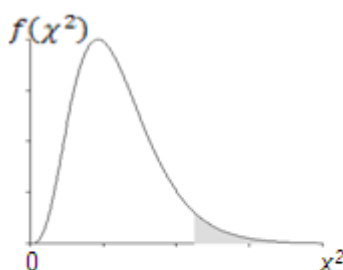
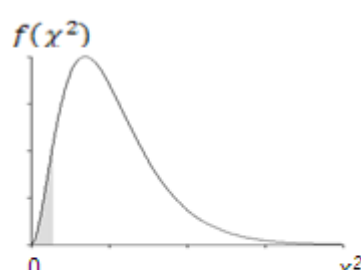
la distribución en el muestreo de la varianza muestral tiene un valor esperado $E[s^2] = \frac{(n-1)}{n} \sigma^2$, o

sea que el valor esperado de la varianza muestral no coincide con el valor deseado que es la varianza poblacional, $E[s^2] \neq \sigma^2$, además tiene asimetría positiva, y depende del tamaño muestral. Pero resulta

que se puede considerar una variable aleatoria tipificada X^2 , definida como $X^2 = \sum_{i=1}^n \left[\frac{(X_i - \bar{X})^2}{\sigma^2} \right]$ que

tiene una distribución conocida, que es tipo Ji cuadrado con $v = n - 1$, esto es $\chi^2_{(n-1)} = \frac{(n-1)S^2}{\sigma^2}$ y que está tabulada para varios valores de áreas en las colas asimétricas de la distribución.

Las pruebas de hipótesis para la varianza poblacional, pueden responder a alguno de los tres siguientes casos

<p>Caso 1: Prueba de dos colas</p>  <p>$H_0: \sigma_x^2 = \sigma_0^2$ $H_1: \sigma_x^2 \neq \sigma_0^2$</p> <p>En ambas colas las áreas de probabilidad son igual a $\alpha/2$</p>	<p>Caso 1: Prueba de cola superior</p>  <p>$H_0: \sigma_x^2 < \sigma_0^2$ $H_1: \sigma_x^2 \geq \sigma_0^2$</p> <p>La región de rechazo se encuentra en la cola derecha y es igual a α.</p>	<p>Caso 1: Prueba de cola inferior</p>  <p>$H_0: \sigma_x^2 > \sigma_0^2$ $H_1: \sigma_x^2 \leq \sigma_0^2$</p> <p>La región de rechazo se encuentra en la cola izquierda y es igual a α.</p>
--	---	---

Ejemplo 12.1. En los procesos industrializados es muy importante obtener conclusiones acerca del valor promedio y de la variabilidad. Con relación al primero, mediante el control estadístico de la calidad, se analiza a través de muestras aleatorias si las variables medidas como por ejemplo del peso neto, la humedad, etc., indican que: 1) el proceso está centrado (la media del proceso coincide con la media especificada o paramétrica) y 2) que la variabilidad es mínima, para reducir el número de productos que resulten defectuosos. En este contexto, antes de analizar el valor medio hay que asegurarse de tener una mínima variabilidad. Por esta razón, una industria que produce cajas de cereales que ha incorporado un nuevo equipamiento y está ajustando su funcionamiento, quiere comprobar si la varianza del proceso actual es mayor a la varianza que debería tener de acuerdo a la especificación del proceso productivo ($\mu \pm 15$ gramos), fijando un $n=25$ y un $\alpha= 0,05$.

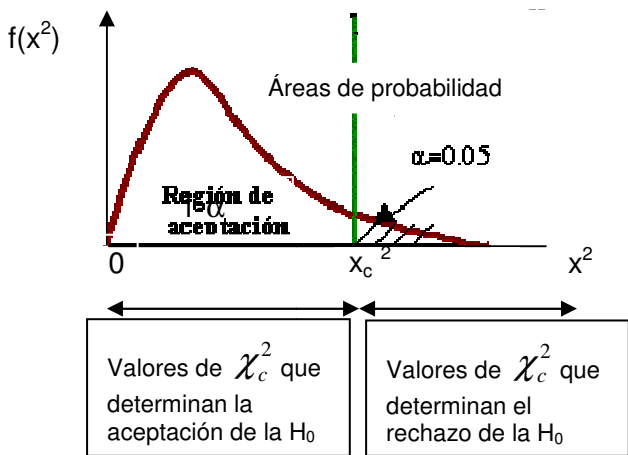
Solución: según la información dada, se puede plantear la siguiente terna de hipótesis,

H_0 : la varianza del peso neto es mayor a 225 gramos.

$$\begin{cases} H_0 : \sigma^2 = \sigma_0^2 ; \text{ donde } \sigma_0^2 = 225 \text{ (gramos)}^2 \\ H_1 : \sigma^2 > \sigma_0^2 \end{cases}$$

Al analizar H_1 queda claro, que se trata de una prueba de cola derecha (el tipo de cola se identifica según la dirección a la que apunta la hipótesis alternativa), y que por lo tanto la hipótesis que se somete a prueba, esto es H_0 , sólo será rechazada cuando los datos muestrales aporten evidencia suficiente, arrojando un valor de estadígrafo de prueba muestral mayor que al valor crítico que indica la distribución Ji cuadrado para el nivel de significancia α fijado.

La regla de decisión se puede graficar como sigue:



Los grados de libertad resultan igual a $n - 1 = 24$, y como se ha fijado un nivel de significancia de 0,05, en la Tabla de la distribución acumulada de probabilidades de Ji cuadrado, se encuentra un $\chi_c^2 = \chi_{\alpha, \nu}^2 = 36, 415$. Luego si resultara $\chi_m^2 < \chi_c^2$, es decir, $\chi_m^2 < 36,415$, se deberá aceptar la H_0 , en tanto que si $\chi_m^2 > \chi_c^2$ se considerará que la muestra aportó suficiente evidencia para rechazar a la hipótesis de igualdad de varianzas.

Dado que el estadígrafo muestral es igual a $\chi_m^2 = \frac{[(25-1)17,3]}{15^2} = 31,92$, el valor muestral es menor al valor crítico, $31,92 < 36,415$. La decisión estadística es: *no corresponde rechazar la H_0 , para un $\alpha = 0,05$* . Luego, en términos del problema se concluye que: no hay evidencia empírica de que haya aumentado la variabilidad del proceso por encima de 15 gramos, para el nivel de significancia fijado.

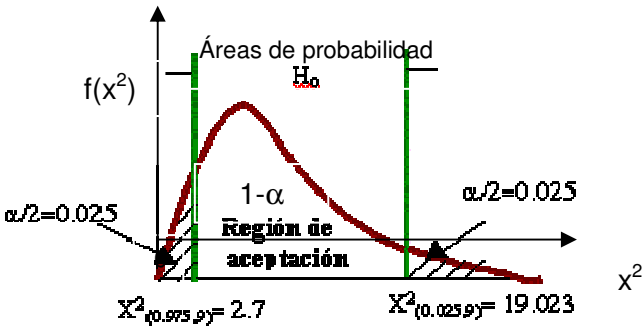
Ejemplo 12.2. Este ejemplo se utilizará para ilustrar un caso de prueba de hipótesis bilateral para la varianza. En una agroindustria, el contenido de azúcar del almíbar de los duraznos enlatados se distribuye normalmente, y se por datos históricos se considera que la varianza es $\sigma^2 = 18 \text{ mg}^2$. Se ha tomado una muestra de 10 latas obteniéndose una desviación típica de 4,8 mg. ¿Muestran estos datos suficiente evidencia para decir que la varianza ha cambiado?. Use $\alpha= 0.05$ y responda teniendo en cuenta el p-valor.

Solución: según la información dada, las hipótesis de interés son

H_0 : la varianza de lo producido tiene un valor numérico diferente al valor histórico.

$$\begin{aligned} H_0 : \sigma^2 &= \sigma_0^2 ; \text{ donde } \sigma_0^2 = 23,0 \text{ mg}^2 \\ H_1 : \sigma^2 &\neq \sigma_0^2 \end{aligned}$$

La regla de decisión es la siguiente:



Usando el criterio tradicional, significa que si $2.70 \leq \chi_m^2 \leq 19.023$ no se debe rechazar la H_0 , y que si contrariamente ocurre que $\chi_m^2 < 2.7$ ó si $\chi_m^2 > 19.023$ no podrá sostenerse la H_0 , pero hay que recordar que se quiere la conclusión en términos de un p-valor.

Al calcular el estadígrafo muestral, resulta:

$$\chi_m^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(10-1)(4.8)^2}{18} = 11.52$$

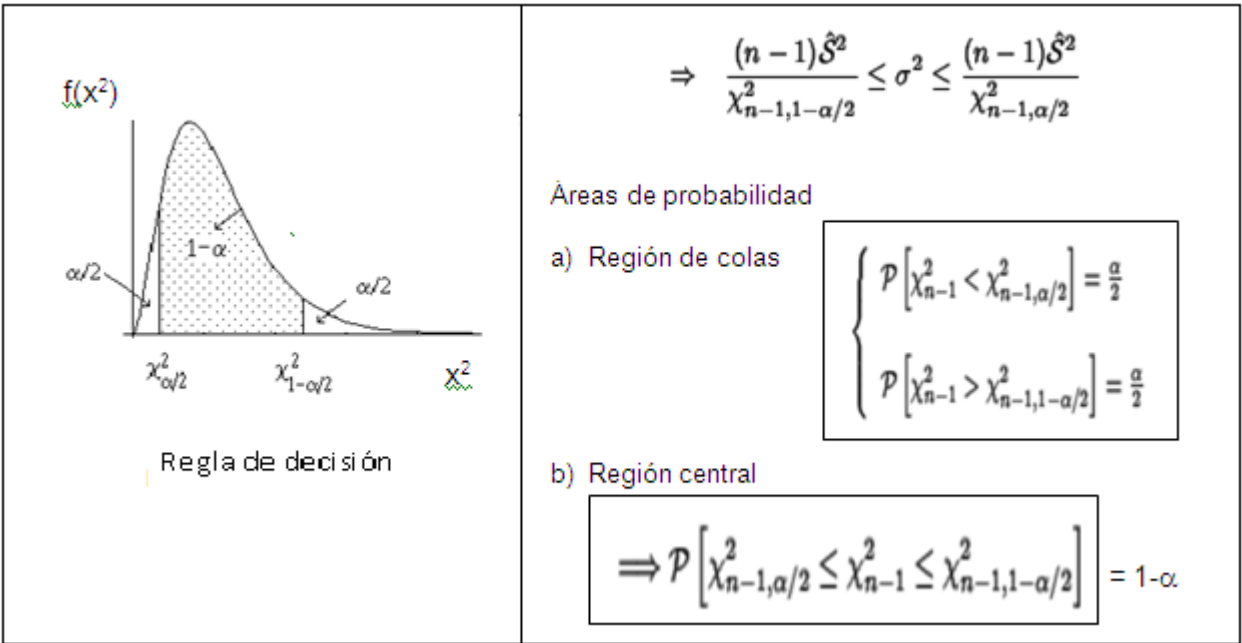
Al entrar con el valor $\chi_m^2 = 11.52$ y $V = 9$, en la Tabla de la distribución de Ji cuadrado, se obtiene un área igual 0.2423, por lo tanto el p-valor resulta igual a $(2)(0.2423) = 0.4846$. Se interpreta entonces que el valor de probabilidad observado es igual a 0,48 de modo que la probabilidad de que se haya presentado un valor de estadígrafo muestral de 11,52 por azar es muy alto, entonces no cabe un rechazo de la hipótesis de que las varianzas son iguales (no ha cambiado la variabilidad).

Ejemplo 12.3. Este ejemplo se utilizará para ilustrar la aplicación de la distribución Ji cuadrado a la estimación intervalar de la varianza. En una semillera el historial de las estadísticas productivas ha mostrado que el peso de las bolsas de semillas para césped contenido se distribuye normalmente. Con motivo de haber realizado una modificación en el sistema de llenado, se han tomado una muestra para conocer la variabilidad del proceso, con el siguiente resultado: 46,4 – 46,1- 45,8 – 47,0 – 46,1 – 45,9 – 45,8 - 46.9 – 45,2 y 46,0. Se quiere estimar la varianza mediante un intervalo de confianza de 95%.

Fundamentación: la estimación de la varianza poblacional mediante un intervalo de confianza se basa en lo siguiente

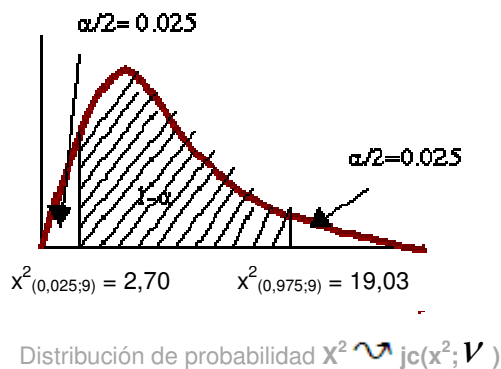
$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \quad \text{que se distribuye como una ji}(\chi^2; V) \quad ; \quad \sigma^2 \in \left[\frac{(n-1)\hat{S}^2}{\chi_{n-1,1-\alpha/2}^2}, \frac{(n-1)\hat{S}^2}{\chi_{n-1,\alpha/2}^2} \right]$$

La primera expresión indica que el estadígrafo a utilizar para construir el intervalo de interés, en el muestreo repetitivo sigue la distribución probabilística de la variable aleatoria Ji cuadrado con grados de libertad $V = n-1$. La segunda explica que se trata de construir un intervalo que contenga a la varianza poblacional, basado en lo anterior. En otros términos, se trata entonces de identificar los dos percentiles de la distribución Ji cuadrado que definen un área central de probabilidad igual a $1-\alpha$, y dejan a ambos lados áreas igual a $\alpha/2$, como muestra el siguiente gráfico y lo indica la siguiente simbología



Solución: a partir de los datos muestrales se obtiene una estimación puntual de la varianza. Para esto, primeramente se calculan los estadígrafos media y varianza:

$$\bar{x} = 46,12 \text{ kg} \quad \text{y,} \quad s^2 = 0,286 \text{ kg}^2 \quad \text{Por tanto} \quad \bar{x} \pm s^2 ; 46,12 \pm 0,535 \text{ kg}$$



Al fijar un intervalo de confianza de 95%, significa que en términos probabilísticos, el nivel de confianza es igual a $1-\alpha$, por tanto el área central en la distribución X^2 (ji cuadrado) limitada por la curva y el eje de abscisas es igual a 0,95, y que a ambas colas les corresponde un área igual a $\alpha/2$, o sea de 0,025 (notar que por la asimetría de la distribución, las áreas de las colas no son simétricas). Con los valores de probabilidad acumulada de 0,025 y 0,975, para grados de libertad $V = n-1=9$, se ingresa a la Tabla de Ji cuadrado para obtener los valores de los percentiles $X^2_{(0,025;9)}$ y $X^2_{(0,975;9)}$, que resultan igual a 2,70 y 19,03 respectivamente.

Por lo tanto, reemplazando en

$$\left[\frac{(n-1)\hat{S}^2}{X^2_{n-1,1-\alpha/2}}, \frac{(n-1)\hat{S}^2}{X^2_{n-1,\alpha/2}} \right]$$

los límites del intervalo de confianza de 95% para la varianza resultan igual a:

Lím inf (int. para σ^2)	$= \frac{(10-1)(0.286)}{19.023} = 0.135$
Lím sup (int. para σ^2)	$= \frac{(10-1)(0.286)}{2.7} = 0.953$

Finalmente, se construye el intervalo de confianza de interés

$$P(0,135 \text{ kg}^2 < \sigma^2 < 0,953 \text{ kg}^2) = 0,95$$

que se interpreta de forma análoga a como se vio para el caso de la estimación intervalar de μ : se tiene una confianza a nivel del 95% que el intervalo construido contenga a la verdadera varianza poblacional de los pesos de las bolsas de semilla, σ^2 .

12.7.2. Pruebas de hipótesis para datos de frecuencias

Según se ha anticipado en la clasificación de las pruebas de hipótesis de Ji cuadrado, en general puede decirse que hay dos situaciones de problemas relacionados con datos de frecuencia:

- a) **Prueba de hipótesis para la bondad de ajuste:** la situación problema se refiere a que se tiene una distribución de frecuencias empíricas (n_i) para una muestra aleatoria univariada, y se ha procedido a ajustar un modelo probabilístico que se elige pensando que explica el comportamiento de la variable de interés en la población, obteniéndose una distribución de frecuencias teóricas calculadas como:

$$\hat{n}_i = (\text{probabilidad de la clase}) \times (\text{tamaño muestral})$$

Luego, el objetivo al aplicar este tipo de prueba es comprobar que no existen discrepancias importantes entre ambas frecuencias, o sea, que se trata de probar que el ajustamiento realizado resulta apropiado.

- b) **Prueba de hipótesis para tablas de contingencia:** las tablas de contingencia muestran datos empíricos de frecuencia (frecuencias observadas), referidos a la clasificación de acuerdo a atributos (variables medidas en escala nominal), o bien a categorías (clases derivadas de la medición en escala ordinal), o de variables cuantitativas originalmente transformadas en variables cualitativas como sería medir rendimientos parcelarios en kg/ha y posteriormente dar los resultados como categorías: rendimiento alto, normal y bajo. Los datos de estas tablas de contingencia dan lugar a dos tipos de análisis, según sea la situación problema.

b.1. Prueba de independencia: se parte de una distribución conjunta de frecuencias empíricas (n_{ij}), obtenida a partir de **una muestra aleatoria de tamaño n**, en la que cada unidad de análisis se clasifica de acuerdo a dos criterios. Esto lleva a un tipo de análisis estadístico para probar que la clasificación de las unidades de análisis según las categorías o clases de una de las variables, es independiente de la clasificación según la otra variable; probabilísticamente para cada celda ij:

$$\hat{p}_{ij} = (\text{probabilidad de ser clasificado en la fila } i - \text{ésima}) \times (\text{probabilidad de ser clasificado en la columna } j - \text{ésima})$$

Bajo el supuesto de una clasificación bivariada independiente, o sea, bajo el supuesto de independencia estadística, se obtienen las frecuencias teóricas (\hat{n}_{ij}) como:

$$\hat{n}_{ij} = \frac{(n_{\bullet j}) \times (n_{i \bullet})}{n_{\bullet \bullet}}$$

b.2. Prueba de homogeneidad: se dispone de datos de frecuencias empíricas para una variable de carácter cualitativo, medida en **r muestras aleatorias de tamaño fijo n para cada caso**, que se consideran proceden de una misma población. En este caso interesa conocer si los datos muestrales aportan evidencia suficiente para comprobar que las r muestras aleatorias clasifican en las j categorías (j conjuntos disyuntos) de forma homogénea, lo que permite inferir para las sendas poblaciones que las mismas son homogéneas entre sí, y por tanto concluir estadísticamente que las muestras proceden de una misma población. La tabla de contingencia en este caso presenta el siguiente aspecto:

Muestra	Atributo A						Total fila
	A ₁	A ₂	...	A _j	...	A _k	
1	n ₁₁	n ₁₁	...	n _{1j}	...	n _{1k}	n _{1.}
2	n ₂₁	n ₂₂	...	n _{2j}	...	n _{2k}	n _{2.}
...
i	n _{k1}	n _{k2}		n _{ij}		n _{jk}	n _{i.}
...
r	n _{k1}	n _{k2}		n _{ij}		n _{rk}	n _{r.}
Total columna	n _{.1}	n _{.2}	...	n _{.j}	...	n _{.k}	n _{..}

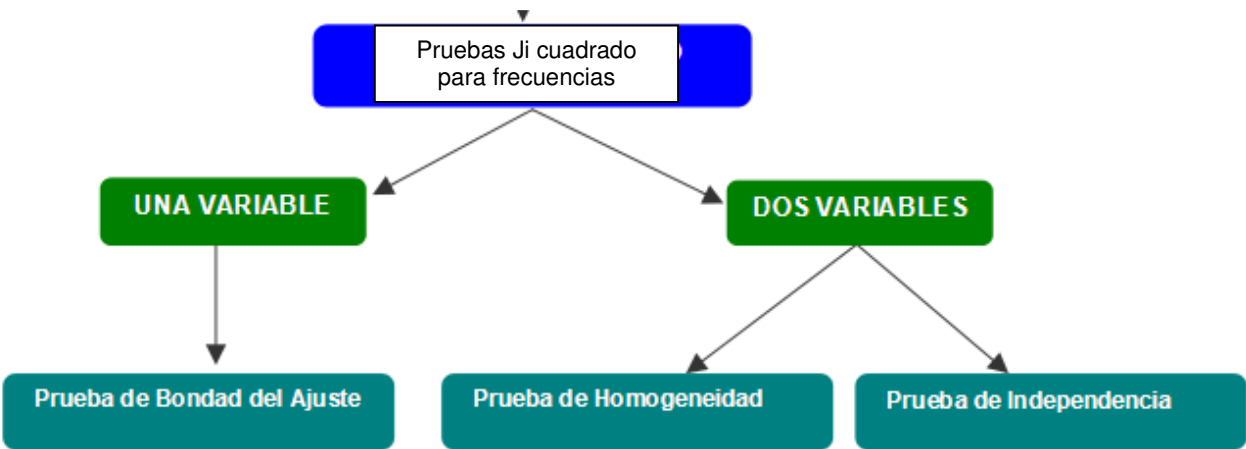
La hipótesis de que las r poblaciones son homogéneas, se traduce en que cada conjunto o categoría A_j debe tener una probabilidad teórica desconocida, que no varía de población a población (las categorías son homogéneas en las r poblaciones). El estadígrafo de prueba, se calcula en forma análoga a una prueba de bondad de ajuste, esto es, para cada una de las muestras se compara la frecuencia observada en cada categoría, con la correspondiente esperada. La frecuencia esperada de que en la muestra i se den observaciones para la categoría j , bajo el supuesto de homogeneidad, se

expresa como $\hat{n}_{ij} = n_i \times (\text{probabilidad de ser clasificado en la categoría } A_j)$

es decir, el número de individuos que tiene la muestra i por la probabilidad de que ocurra la característica j en la población:

$$\hat{n}_{ij} = n_{i.} \times \frac{(n_{\bullet j})}{n_{\bullet \bullet}}$$

El siguiente diagrama sintetiza los casos expuestos



12.7.3. Aplicaciones del Ji cuadrado a Pruebas de hipótesis para datos de frecuencias

12.7.3.1. Prueba de bondad de ajuste

En la Unidad de Probabilidad se presentó el concepto de ajustamiento de una distribución de probabilidades, y se cumplieron los primeros pasos de una prueba de bondad de ajuste:

1º) A partir del análisis de la distribución de frecuencias observadas en una muestra, se eligió una distribución de probabilidad para modelar la distribución de la correspondiente variable aleatoria (distribución poblacional de donde se supone fue extraída la muestra aleatoria).

2º) Se estimaron los parámetros de la distribución de probabilidad elegida, esto es $\hat{\pi}; \hat{\lambda}; \hat{\mu}; \hat{\sigma}$, etc. a partir de información real o de un conocimiento completo disponible sobre la población.

3º) Se usó la distribución de probabilidad teórica para determinar la probabilidad de ocurrencia de los valores (puntuales o intervalares) de la variable aleatoria en el muestreo, para calcular las correspondientes frecuencias teóricas \hat{n}_i .

4º) Por último, se calcularon las diferencias $\hat{n}_i - n_i$, y se estableció si sus magnitudes indicaban una discrepancia grande o pequeña entre lo observado y lo modelado, como para sospechar que la muestra, respectivamente, no provenía de la población supuesta (mal ajustamiento) o sí (buen ajustamiento).

Mediante la prueba Ji cuadrado para bondad de ajuste, se dispondrá de una herramienta que permitirá justificar en términos probabilísticos, la decisión de considerar que el modelo fue “adecuado” para explicar el comportamiento de los datos muestrales, o en otras palabras si el modelo se ajusta a lo observado (“ajustamiento bueno”), o bien si no resultó un modelo apropiado y lo observado requiere otro modelo explicativo (“ajustamiento malo”).

En pruebas de bondad de ajuste existen dos casos posibles con relación al modelo probabilístico a utilizar para estimar las \hat{n}_i :

- Modelos probabilísticos conocidos de aplicación generalizada como el normal, binomial o Poisson.
- Modelos que especifican interrelaciones de interés particular en determinados campos del saber, tal el caso de los modelos probabilísticos referidos a las leyes de Mendel que explican la segregación de los caracteres genéticos.

12.7.3.1.1. Prueba de bondad de ajuste con modelos probabilísticos conocidos

Caso 1. Distribución uniforme

Situación problema: un jugador compró un dado corriente de seis caras y quiere comprobar si está bien construido. Para esto se realiza 120 lanzamientos y registra las frecuencias correspondientes a los seis resultados posibles. ¿Indican los datos experimentales que el dado es legal ($\alpha=0,05$)?

Hipótesis:

$$\begin{cases} H_0: \text{el dado no es legal} \\ H_0: P(X = x_i) = \frac{1}{6} \quad \text{para } i=1, 2, \dots, 6 \\ H_1: P(X = x_i) \neq \frac{1}{6} \quad \text{para al menos un } i \end{cases}$$

Regla de decisión:

$$\chi_m^2 = \sum_{i=1}^k \left[\frac{(n_i - \hat{n}_i)^2}{\hat{n}_i} \right] \sim \chi^2(x^2; \nu), \text{ donde } \nu = k - 1 = 6 - 1 = 5; \quad \chi_{(1-\alpha)}^2; \nu = 11,07$$

Cálculo del estadígrafo de prueba:

$$\text{Datos de 120 lanzamientos de un dado legal } \left(\pi_i = \frac{1}{6} \right)$$

Puntaje	Nº de ocurrencias n_i	Frecuencia teórica $\hat{n}_i = P(X = x_i) n$ $= (1/6) 120$	$(n_i - \hat{n}_i)$	$\frac{(n_i - \hat{n}_i)^2}{\hat{n}_i}$
1	18	20,00	-2,00	0,20
2	21	20,00	0,00	0,00
3	25	20,00	7,00	2,45
4	13	20,00	-5,00	1,25
5	23	20,00	3,00	0,45
6	17	20,00	-3,00	0,45
-	120	120,00	0,00	$\chi_m^2 = 4,80$

donde
$$\chi_m^2 = \frac{(18-20,00)^2}{20,00} + \frac{(21-20,00)^2}{20,00} + \frac{(25-20,00)^2}{20,00} + \frac{(13-20,00)^2}{20,00} + \frac{(23-20,00)^2}{20,00} + \frac{(17-20,00)^2}{20,00}$$

Conclusiones:

- a) *Conclusión estadística:* dado que $\chi_m^2 < \chi_{(1-\alpha);v}^2$, esto es el valor muestral de Ji cuadrado es menor al valor que indica la distribución de probabilidades Ji cuadrado ($4,80 < 11,07$), o sea que pertenece al intervalo de valores de la variable en correspondencia a la región de aceptación, se decide aceptar la hipótesis nula, al nivel de significancia de 0,05.
- b) *Conclusión en términos del problema:* dada la conclusión estadística que antecede, hay que aceptar que se trata de un dado legal, es decir, que en una larga serie de tiradas hay que esperar que todas las caras del dado (1 al 6) se presenten con similar número de ocurrencias (frecuencia real).

Caso 2. Distribución de Poisson

Situación problema: de un monte de cerezos atacado por pulgón verde, un técnico fruticultor ha extraído una muestra aleatoria de 100 hojas. Examinado el material recolectado, se han encontrado los siguientes resultados:

Nº de pulgones/hoja	0	1	2	3	4	5	6	≥ 7
Nº de hojas	39	21	18	9	5	4	3	1

El técnico postula que $X \sim p(x; \lambda)$. Pruebe la bondad del ajuste para un $(\alpha=0,05)$.

Hipótesis:

H_0 : el número de pulgones verdes por hoja, sigue una distribución de Poisson

$$\begin{cases} H_0 : X \sim p(x; \lambda); \\ H_1 : X \text{ sigue otra distribución} \end{cases}$$

Regla de decisión:

$$\chi_m^2 = \sum_{i=1}^k \left[\frac{(n_i - \hat{n}_i)^2}{\hat{n}_i} \right] \sim \chi^2(x^2; \nu) \text{ , donde } \nu = k - 2 = 7 - 1 - 1 = 5^1; \chi_{(1-\alpha); \nu}^2 = 11,07$$

Cálculo del estadígrafo de prueba:

Resulta conveniente notar que para calcular las probabilidades $P(X = x_i)$ se requiere conocer el parámetro de la distribución de Poisson, $P(X = x_i) = p(x_i) = \frac{e^{-\lambda} \lambda^x}{x!}$; donde $\lambda = n.p$ pero p , la probabilidad p no es conocida: el planteo de la situación problema no incluye un valor especificado de λ , ni tampoco se informa en la hipótesis, por tanto habrá que estimar su valor a partir de los datos muestrales para poder realizar el ajustamiento, como

$$\hat{\lambda} = \bar{x} = \frac{\sum_{i=1}^k x_i n_i}{n} = 1,49$$

¹ Recordar que $\nu = k - s - 1$, y en este caso $s = 1$, se pierde un grado de libertad al estimar a λ . Estos grados de libertad luego serán corregidos por agrupamiento de clases

Datos de recuento de pulgones en 100 hojas de cerezo.					
Nº de pulgon es /hoja	Nº de hojas n_i	Probabilidad $p(x_i) = \frac{e^{-\lambda} \lambda^x}{x!}$	Frecuencia teórica $\hat{n}_i = p(x_i) n$	$(n_i - \hat{n}_i)$	$\frac{(n_i - \hat{n}_i)^2}{\hat{n}_i}$
0	39	0,225	22,54	16,46	12,020
1	21	0,336	33,58	-12,58	4,713
2	18	0,250	25,02	-7,02	1,970
3	9	0,124	12,43	-8,43	0,946
4	5	0,046	4,63	6,43	6,713
5	4	0,014	1,38		
6	3	0,003	0,34		
7	1	0,001	0,08		
-	100	1,000	100,00	0,00	$\chi_m^2 = 26,362$

donde las clases cuyas frecuencias esperadas han sido menores a 5 en correspondencia a la cola superior de la distribución, esto es \hat{n}_i para $x_i = 5,6,7$, se han agrupado hasta cumplir con el requisito $\hat{n}_i \geq 5$ obteniéndose un valor grupal de 6,43. Lo propio se ha hecho luego con las respectivas n_i dando 13. Esto lleva a recalcular los grados de libertad iniciales, $v=n-k=7-2=5$, resultando como $v=n-k=5-2=3$.

Puesto que estas clases se encuentran en las secciones del extremo inferior y del superior de la distribución, se tienen que combinar con categorías adyacentes respectivas para el propósito de realizar el análisis. Luego el valor crítico del estadígrafo de prueba resulta igual a $\chi_m^2 = \chi_{(3;0,95)}^2 = 7,82$; es decir que los valores que determinarán el rechazo de la H_0 , al nivel $\alpha=0,05$, pertenecen al intervalo $[7,82; +\infty]$

Conclusiones:

- a) *Conclusión estadística:* dado que $\chi_m^2 > \chi_{(1-\alpha);v}^2$, esto es el valor muestral de Ji cuadrado es mayor al valor que indica la distribución de probabilidades Ji cuadrado ($26,36 > 7,82$), o sea que pertenece al intervalo de valores de la variable en correspondencia a la región de rechazo, se decide rechazar la hipótesis nula, al nivel de significancia de 0,05.
- b) *Conclusión en términos del problema:* dada la conclusión estadística que antecede, resulta que los datos sobre el número de pulgones/hoja no siguen una distribución Poisson con tasa media igual a 1,49.

Caso 3. Distribución Binomial

Para resolver situaciones problema relacionadas con la distribución binomial se debe seguir un camino análogo al indicado para la distribución binomial, recordando que $H_0 : X \sim b(x; n, p)$ y, que

$$v = n - s - 1 \begin{cases} p \text{ conocido} & v = k - 1 \\ p \text{ desconocido (se estima a través de la muestra)} & v = k - 2 \end{cases}$$

Caso 4. Distribución normal

Situación problema: de la base del censo provincial del arbolado público viario o “arbolado de calle”, se conoce para la variable circunferencia de tronco de los plátanos lo siguiente; $\mu \pm \sigma$ es igual a $190,85 \pm 34,54$ en cm. Para una ciudad donde todavía no se ha llevado a cabo este censo, se ha extraído una muestra aleatoria de $n=228$ plátanos. Interesa modelar la distribución teórica de la variable aleatoria circunferencia de tronco, suponiendo que la muestra procede de la población conocida $X \sim n(x; 190,85, 34,54)$. Pruebe la bondad del ajuste para un $(\alpha=0,05)$.

Hipótesis:

$$\begin{cases} H_c: \text{la circunferencia de tronco de los plátanos de la ciudad considerada, se distribuye normalmente} \\ H_0 : X \sim n(x; 190,85 \text{ cm}, 34,54 \text{ cm}) \\ H_1 : X \text{ sigue otra distribución} \end{cases}$$

Regla de decisión:

$$\chi_m^2 = \sum_{i=1}^k \left[\frac{(n_i - \hat{n}_i)^2}{\hat{n}_i} \right] \sim \chi^2(x^2; \nu) \text{ , donde } \nu = k - 1 = 13 - 1 = 12 \text{ ; } \chi_{(1-\alpha); \nu}^2 = 21,03$$

En este caso los dos parámetros de la distribución normal, μ y σ , son conocidos por tanto $s=0$. Pero habrá que ver si resulta necesario agrupar clases para determinar la necesidad de corregir los grados de libertad.

Cálculo del estadígrafo de prueba

Se observará que en este caso ha resultado necesario agrupar clases en ambos extremos de la distribución, con lo cual se tienen que recalcular los grados de libertad iniciales $\nu = k - 1 = 13 - 1 = 12$ como $\nu = k - 1 = 8 - 1 = 7$, según ocurrió en el caso del ajustamiento con la distribución de Poisson. De este modo el valor crítico del estadígrafo de prueba resulta igual a $\chi_m^2 = \chi_{(7;0,95)}^2 = 14,07$; es decir que los valores que determinarán el rechazo de la H_0 , al nivel $\alpha=0,05$, pertenecen al intervalo $[14,07; +\infty]$

Datos de circunferencia de tronco de plátanos del arbolado viario para una ciudad.						
Intervalos de clase	Punto medio x_i	Probabilidad del intervalo	Frecuencia absoluta n_i	Frecuencia teórica $\hat{n}_i = (prob \text{ int})n$	$(n_i - \hat{n}_i)$	$\frac{(n_i - \hat{n}_i)^2}{\hat{n}_i}$
Menos de 100	-	0,00430	2	1,0002		
100 < 120	110	0,01590	5	3,7047	-0,4947	0,015
120 < 140	130	0,05060	9	11,7898		
140 < 160	150	0,11590	15	22,0047	-7,0047	2,230
160 < 180	170	0,19160	40	44,6428	-4,6428	0,483
180 < 200	190	0,22430	59	52,2619	6,7381	0,869
200 < 220	210	0,19690	37	45,8777	-8,8777	1,718
220 < 240	230	0,12270	21	28,5891	-7,5891	2,015
240 < 260	250	0,05500	16	12,8150	3,1850	0,792
260 < 280	270	0,01790	10	4,1707		
280 < 300	290	0,00411	7	0,9580	18,6872	65,730
300 < 320	310	0,00070	6	0,1631		
320 ó más	-	0,00009	1	0,0210		
-	-	1,00000	228	227,9987	0,0013	$\chi_m^2 = 73,850$

Conclusiones:

- a) *Conclusión estadística:* dado que $\chi_m^2 > \chi_{(1-\alpha); \nu}^2$, esto es el valor muestral de Ji cuadrado es mayor al valor que indica la distribución de probabilidades Ji cuadrado (73,85 > 14,07), o sea que pertenece al intervalo de valores de la variable en correspondencia a la región de rechazo, se decide rechazar la hipótesis nula ($\alpha=0,05$).
- b) *Conclusión en términos del problema:* dada la conclusión estadística que antecede, la muestra no aporta suficiente evidencia a favor de H_0 . No puede decirse que los datos sobre la circunferencia no siguen la distribución normal propuesta al nivel de significancia de 0,05. Es importante destacar que ha existido una gran discrepancia entre lo observado y lo teórico en la cola superior de la distribución, la muestra presenta considerablemente mayor número de árboles con circunferencia grande que lo que puede esperarse en una muestra extraída de la población censal.

12.7.3.1.2. Prueba de bondad de ajuste con modelos específicos

Hay campos de aplicación donde se utilizan leyes probabilísticas que establecen interrelaciones particulares entre las probabilidades multinomiales que corresponden a k clases. Esto es lo que ocurre con las leyes de Mendel que son un conjunto de reglas básicas acerca de cómo se transmiten por herencia, las características de los padres a sus hijos. Este conocimiento es muy utilizado en agronomía en la genética vegetal y animal, para lograr ejemplares con características deseables.


Situación problema: de acuerdo a la teoría mendeliana de la herencia, cuando se cruzan plantas de arveja (*Pisum sativum*), puede esperarse con relación a la herencia de las características textura del tegumento y el color del grano que se presente una interrelación de 9:3:3:1.


Genotipos parentales AaLl x AaLl			
Donde para color, A es amarillo (dominante) y, a es verde (recesivo), y para forma, L es lisa (dominante) y l rugosa (recesivo)			
Segregación fenotípica esperada para la descendencia			
Granos arveja amarillos y lisos	Granos de arveja amarillos y rugosos	Granos de arveja verdes y lisos	Granos de arveja verdes y rugosos
9	3	3	1
Proporciones mendelianas			
9/16	3/16	3/16	1/16

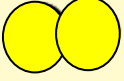
Características del grano


Forma

Color


lisa


rugosa


Amarillo


Verde

Las proporciones establecidas se pueden interpretar como estimaciones empíricas de las correspondientes probabilidades, y entonces se tiene una ley específica de probabilidad para el caso.

Un genetista ha realizado un experimento genético y quiere comprobar si los datos obtenidos están de acuerdo con las proporciones mendelianas dadas para el nivel de significación 0,05.

Hipótesis:

H_c : los caracteres color y forma de las semillas de arvejas segregan en proporción 9:3:3:1, de este modo se está suponiendo que los 4 tipos de semilla, en la población se presentan de acuerdo a la siguiente proporción $\frac{9}{16} : \frac{3}{16} : \frac{3}{16} : \frac{1}{16}$

$$\begin{cases} H_o : \pi_1 = \frac{9}{16} ; \pi_2 = \frac{3}{16} ; \pi_3 = \frac{3}{16} , \pi_4 = \frac{1}{16} \\ H_1 : \text{al menos una } \pi_i \text{ se presenta con una probabilidad diferente a la especificada} \end{cases}$$

Regla de decisión:

$$\chi_m^2 = \sum_{i=1}^k \left[\frac{(n_i - \hat{n}_i)^2}{\hat{n}_i} \right] \sim \chi^2 (x^2; \nu) \text{ , donde } \nu = k - 1 = 4 - 1 = 3 ; \chi_{(1-\alpha); \nu}^2 = 7,82$$

Notar que en este caso los valores de probabilidad están especificados por el modelo, de modo que para el cálculo de los grados de libertad, s=0.

Cálculo del estadígrafo de prueba

Para calcular las frecuencias esperadas, se aplica lo siguiente $\frac{9}{16}n ; \frac{3}{16}n ; \frac{3}{16}n ; \frac{1}{16}n$ donde n= 556 semillas, por ejemplo $\hat{n}_1 = \pi_1 \cdot n = \left(\frac{9}{16}\right)596 = 312,75$. El ajustamiento de acuerdo a la ley mendeliana propuesta resultó:

Datos de un cruzamiento de arveja.					
Fenotipo	n_i	\hat{n}_i	$(n_i - \hat{n}_i)$	$(n_i - \hat{n}_i)^2$	$\frac{(n_i - \hat{n}_i)^2}{\hat{n}_i}$
Amarilla-Lisa	315	312,75	2,25	5,0625	0,0162
Amarilla-rugosa	101	104.25	-3,25	10,5625	0,1013
verde-lisa	108	104.25	3,75	10,5625	0,1349
Verde-rugosa	32	34.75	-2,75	7,5625	0,2176
Total	556	556,00	0,00	-----	$\chi_m^2 = 0,47$

Conclusiones:

- a) *Conclusión estadística:* dado que $\chi_m^2 < \chi_{(1-\alpha); \nu}^2$, esto es el valor muestral de Ji cuadrado es menor al valor que indica la distribución de probabilidades de la variable aleatoria Ji cuadrado

(0,47 < 7,82), o sea que pertenece al intervalo de valores de la variable en correspondencia a la región de aceptación, se decide no rechazar la hipótesis nula ($\alpha=0,05$).

- b) *Conclusión en términos del problema:* dada la conclusión estadística que antecede, la muestra no aporta suficiente evidencia para rechazar H_0 . Por tanto se considera que la ley mendeliana $\frac{9}{16} : \frac{3}{16} : \frac{3}{16} : \frac{1}{16}$ puede utilizarse como modelo probabilístico para explicar los resultados experimentales obtenidos en experimentos similares al realizado por el genetista, para un nivel de significancia de 0,05.

12.7.3.2. Prueba para tablas de contingencia

Conviene recordar el concepto y la notación de una **tabla de contingencia**: 1) es una disposición de datos de frecuencias observadas, que puede corresponder a una clasificación de doble entrada (bivariada) o de orden superior. Los datos se registran en las celdas de la tabla, identificados según la notación matricial: esto es mediante una notación sub (i,j) donde i se refiere a la fila y, j a la columna, de modo que i= 1,2, ..., i, ...r y, j= 1,2, ..., i, ...c.

Frecuencias observadas		Columnas						Total marginal de fila	Frecuencias teóricas
		1	2	...	j	...	c		
Filas	1	n_{11}	n_{12}	...	n_{1j}	...	n_{1c}	$n_{1\bullet}$	$\hat{n}_{ij} = \frac{(n_{\bullet j}) \times (n_{i\bullet})}{n_{\bullet\bullet}}$
	2	n_{21}	n_{22}	...	n_{2j}	...	n_{2c}	$n_{2\bullet}$	
	
	i	n_{i1}	n_{i2}	...	n_{ij}	$n_{i\bullet}$	
	
Total marginal de columna	r	n_{r1}	n_{r2}	n_{rc}	$n_{r\bullet}$	
		$n_{\bullet 1}$	$n_{\bullet 2}$...	$n_{\bullet j}$...	$n_{\bullet k}$	$n_{\bullet\bullet}$	

En forma abreviada a una tabla de contingencia con r filas y c columnas se le conoce como tabla r x c (se lee r por c).

12.7.3.2.1 Prueba de independencia

Situación problema: se desea probar si la decisión de los votantes respecto a la reforma de la Constitución Provincial de Mendoza es independiente del nivel de ingresos de los mismos para un $\alpha=0,05$. A tal efecto se ha tomado una muestra aleatoria de 1000 votantes² del padrón electoral, resultando lo siguiente:

Tabla de contingencia 2 x 3; n=1000

Reforma de la Constitución	Nivel de ingreso			Total
	Bajo	Medio	Alto	
A favor	182	213	203	598
En contra	154	138	110	402
Total	336	351	313	1000

12.7.2.2.2 Prueba de homogeneidad

A diferencia de la prueba de independencia, en la prueba de homogeneidad interesa determinar si los datos correspondientes a dos o más muestras aleatorias provienen de la misma población.

² Notar que al tomar solo una muestra aleatoria los totales marginales de filas y columnas son aleatorios

Nuevamente el conjunto de posibles valores de las observaciones se divide en k conjuntos disyuntos, A_1, A_2, \dots, A_k , clasificando en ellos las n_k observaciones de cada muestra. Es importante notar que al tener definidos los tamaños muestrales, resulta que los **totales de filas (o bien de columnas según se haya ordenado) resultan fijos**.

Situación problema: en una región minera se están explotando dos minas. Entre los pobladores de la zona de influencia corre la opinión de que la cantidad de enfermos por contaminación ambiental es mayor en una de las minas. Se quiere probar si realmente los registros sanitarios no son homogéneos, a tal fin se toma una muestra de trabajadores de cada mina y se contabilizan los casos con patologías asociadas a las condiciones ambientales. Los resultados han sido:

Tabla de contingencia 2 x 2; $n_1=160$ y $n_2 = 100$.

Explotación minera	Patologías		Total
	Sin	Con	
A	35	125	160
B	37	63	100
C	25	35	60
Total	97	223	320

Hipótesis:

H_0 : para cada mina las proporciones de registros sanitarios, sin y con patologías asociadas a la contaminación ambiental, son las mismas.

$$\left\{ \begin{array}{l} H_0 : \pi_{11} = \pi_{21} = \pi_{31}; \quad \pi_{12} = \pi_{22} = \pi_{32} \quad ; \quad \text{donde} \quad \begin{array}{l} i = 1,2,3 \\ j = 1,2 \end{array} \\ H_1 : \text{las poblaciones no son homogéneas} \end{array} \right.$$

En este contexto, homogénea se interpreta como igual. Las dos poblaciones en estudio serán homogéneas cuando la interrelación entre los casos sin patología y con patología sea igual en ambos, esto es, en ambas minas las proporciones entre los dos tipos de registros son iguales. En esencia, interesa determinar si en las dos explotaciones mineras se dan de forma similar los casos de patologías positivos y casos negativos atribuibles al ambiente laboral.

Regla de decisión:

$$\chi_m^2 = \sum_{i=1}^r \sum_{j=1}^c \left[\frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}} \right] \sim \chi^2(x^2; \nu) \quad , \text{ donde } \nu = (r-1)(c-1) = (3-1)(2-1) = 2 ;$$

$\chi_{(1-\alpha); \nu}^2 = 5,99$

El número de grados de libertad asociado a este tipo de prueba está dado por el número de frecuencias de celdas que pueden llenarse libremente cuando se dan los totales marginales y el total general.

Cálculo del estadígrafo de prueba

Al suponer homogeneidad también los datos se ordenan en una tabla de contingencia, por tanto las frecuencias esperadas de cada celda nuevamente pueden obtenerse multiplicando las frecuencias marginales de la fila y la columna de la celda en cuestión, y dividiendo por el total general ($n_1 + n_2 = n$). Pero en este caso al tratarse de una tabla 2x2 basta calcular la frecuencia teórica para la celda (1,1). que las restantes se obtienen por diferencia, esto es: $\hat{n}_{12} = n_{1.} - \hat{n}_{11}$; $\hat{n}_{21} = n_{.1} - \hat{n}_{11}$; $\hat{n}_{22} = n_{..} - \hat{n}_{11}$.

Tabla de contingencia 2 x 2 con frecuencias observadas y calculadas; $n_1=160$, $n_2 = 100$ y $n_3 = 60$.

Explotación minera	Patologías		Total
	Sin	Con	
A	35 (48,500)	125(111,500)	160
B	37(30,313)	63(69,688)	100
C	25 818,188)	35(41,813)	60
Total	97	223	320

Luego

$$\chi_m^2 = \frac{(35 - 48,500)^2}{48,500} + \frac{(125 - 111,500)^2}{111,500} + \dots + \frac{(35 - 41,813)^2}{41,813} = 11,171$$

Conclusiones:

- a) *Conclusión estadística:* dado que $\chi_m^2 > \chi_{(1-\alpha);v}^2$, esto es el valor muestral de Ji cuadrado es mayor al valor que indica la distribución de probabilidades Ji cuadrado (11,171 > 5,99), o sea que pertenece al intervalo de valores de la variable en correspondencia a la región de rechazo, se decide rechazar la hipótesis nula ($\alpha=0,05$).
- b) *Conclusión en términos del problema:* dada la conclusión estadística que antecede, no hay evidencia para concluir que la proporción de trabajadores con patología y sin patología difiere entre las explotaciones mineras al nivel de significancia de 0,05.

12.7. CORRECCIÓN DE YATES

Es importante recordar que el estadígrafo sobre el cual se basa la decisión en las pruebas de Ji cuadrado no paramétricas, tiene una distribución sólo aproximadas a la distribución de una variable aleatoria Ji cuadrado (la distribución de X^2 es una distribución continua de probabilidades, y el estadígrafo se calcula a partir de datos de frecuencia o conteo). En consecuencia se requiere tomar algunas precauciones:

- 1) que la muestra sea grande (no menor a 50)
- 2) que las frecuencias teóricas no sean menor a 5 (caso contrario agrupar clases)
- 3) que los grados de libertad sean mayores a 1.

Con relación a esto último caso, las tablas de contingencia 2x2 siempre son violatorias de este requisito ya que $v = (r - 1)(c - 1) = (2 - 1)(2 - 1) = 1$. Esto lleva a una corrección en el cálculo del estadígrafo para mejorar la aproximación de su distribución de probabilidades con la distribución continua Ji cuadrado. Tal corrección recibe el nombre de **corrección de Yates para continuidad**.

La corrección consiste en aplicar la fórmula que se da a continuación, en lugar de la utilizada hasta ahora:

$$\chi^2 = \sum \frac{(|n_i - \hat{n}_i| - 0,5)^2}{\hat{n}_i}$$

Por último si se tuvieran frecuencias esperadas menores a 5, se debería aplicar la prueba exacta de Fisher-Irwin.

REQUISITOS PARA APLICAR EL ESTADÍGRAFO

Ji cuadrado, χ^2

- 1º) La o las muestras deben ser aleatorias
- 2º) El tamaño muestral n debe ser grande ($n > 30$)
- 3º) Todas las frecuencias esperadas deben ser iguales o mayores a 5 (en caso de que no sea así agrupar varias categorías hasta tener valores ≥ 5)