USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
FACULTADE DE MATEMÁTICAS

Trabajo Fin de Grado

CONTRASTES DE BONDAD DE AJUSTE

Emilio Argibay Alló

Curso 2024/2025

UNIVERSIDAD DE SANTIAGO DE COMPOSTELA

GRADO DE MATEMÁTICAS

Trabajo Fin de Grado

CONTRASTES DE BONDAD DE AJUSTE

Emilio Argibay Alló

Julio, 2025

UNIVERSIDAD DE SANTIAGO DE COMPOSTELA

Trabajo propuesto

Área de Conocimiento: Estadística y Investigación Operativa
Título: Contrastes de bondad de ajuste
Breve descripción del contenido
Los contrastes de bondad de ajuste permiten verificar si es plausible cierta hipótesis paramétrica sobre el modelo de distribución que genera los datos de una muestra. Durante el grado se introduce este problema, presentando los tests tipo Kolmogorov-Smirnov o los contrastes tipo ji-cuadrado. En este TFG se pretende profundizar en este problema, tanto desde el punto de vista teórico como en su vertiente aplicada, comprobando mediante estudios de simulación la idoneidad de algunas de las recomendaciones prácticas habituales para este tipo de contrastes.
Recomendaciones
Otras observaciones

Índice

Resumen

El desarrollo de este trabajo se centra en el estudio teórico y práctico del test $ji-cuadrado$, el cual se utiliza para contrastar cierta hipótesis sobre una muestra de datos. Se presenta, en primer lugar, el contexto histórico del test, incluyendo su origen y evolución.

En el estudio teórico, se formaliza el test, con la definición de su respectivo estadístico de contraste y con los principales resultados relacionados con el mismo, incluyendo aquellos sobre la convergencia asintótica del estadístico. Se diferencian el caso en el que la hipótesis nula es simple y el caso en el que se trata de una familia paramétrica.

En el ámbito práctico, se realizan simulaciones del test en sus distintos casos mediante la herramienta R, con la finalidad de analizar y caracterizar su comportamiento real. Se estudian el calibrado del test (casos en los que se cumple la hipótesis nula) y su potencia (casos en los que no se cumple la hipótesis nula). También se ponen a prueba ciertas recomendaciones prácticas mencionadas durante el grado.

Abstract

The development of this project focuses on the theoretical and practical study of the chi-squared test, which is used to test certain hypotheses on a data sample. Firstly, the historical context of the test is presented, including its origin and evolution.

In the theoretical study, the test is formalized by defining its corresponding test statistic and presenting the main results related to it, including those concerning the asymptotic convergence of the statistic. The case in which the null hypothesis is simple is distinguished from the case in which it belongs to a parametric family.

In the practical part, simulations of the test are carried out in its different cases using the R software, with the aim of analyzing and characterizing its actual behavior. The calibration of the test (cases in which the null hypothesis holds) and its power (cases in which the null hypothesis does not hold) are studied. Certain practical recommendations mentioned during the degree are also tested.

Capítulo 1

Introducción y contexto histórico del test *ji-cuadrado*

En este capítulo introduciremos la motivación alrededor del test *ji-cuadrado* y también nos centraremos en situar históricamente los inicios del test, así como los cambios y correcciones que se fueron llevando a cabo a lo largo del tiempo.

1.1. Introducción y motivación

Cuando analizamos datos, hay situaciones en las que observamos frecuencias asociadas a diferentes categorías, y queremos comprobar si estas se ajustan a lo que esperaríamos bajo cierta hipótesis relacionada con la distribución de los datos. El test *ji-cuadrado* fue creado para dar solución a este tipo de problemas.

Ilustraremos a continuación este tipo de situaciones mediante un ejemplo. Supongamos que lanzamos un dado 60 veces y observamos cuántas veces sale cada una de las 6 caras del mismo. Si el dado es un dado totalmente normal, esperaremos que cada una de las caras salga más o menos 10 veces. Ahora bien, si al realizar los 60 lanzamientos obtenemos que una cara sale unas 20 veces, nos haremos la siguiente pregunta: ¿Es esto parte del azar o podemos sospechar que el dado está trucado?

La clave del test *ji-cuadrado* que trataremos en este trabajo está en cuantificar la discrepancia global entre las frecuencias observadas y las esperadas bajo cierta hipótesis. A través de un estadístico, denominado estadístico de contraste, podemos determinar si la desviación global entre lo esperado y lo observado es suficientemente grande como para rechazar la hipótesis (en el ejemplo, que el dado no está trucado). El valor de este test reside en su generalidad: se puede

aplicar a cualquier situación en la que se contengan elementos en distintas clases o categorías, y se quiera contrastar si la distribución observada difiere significativamente de la esperada.

1.2. Origen del test

El test *ji-cuadrado* fue introducido por Karl Pearson en 1900, cuyo trabajo abordó principalmente la necesidad de establecer criterios objetivos para estudiar si una distribución empírica se ajustaba a una distribución teórica.

En su primer artículo de 1900 (ver [13]), del cual Cochran comenta su contenido en su artículo de 1952 (ver [3]), Pearson definió el estadístico $ji - cuadrado$ como la suma de las diferencias al cuadrado entre las frecuencias observadas y esperadas, dividida por las frecuencias esperadas. Este estadístico era bastante intuitivo, ya que cuanto mayor fuese, mayor sería la discrepancia entre la distribución de los datos y la hipótesis. Lo aplicó a numerosos casos prácticos, incluidos experimentos en biología.

Un claro ejemplo es el estudio que hizo acerca de ciertas muestras de datos publicadas por Sir George Airy y el profesor Merriman en sus libros de 1861 y 1884, respectivamente (ver [1] y [12]), los cuales afirmaban que estas provenían de la distribución normal. Para asegurar esto, solo podían utilizar la inspección visual. Cuando Pearson aplicó su test, mostró que ambos estaban equivocados. Este caso muestra tanto la utilidad práctica del test como la intención de Pearson de dotar a los científicos de una herramienta para evaluar hipótesis sobre distribuciones observadas.

En [3], también se revisa que el procedimiento de Pearson tenía ciertas limitaciones. La principal limitación era la ausencia de una demostración formal de que el estadístico $ji-cuadrado$ seguía la distribución que él asumía. Pearson se basaba en una prueba considerada hoy en día informal, sin entrar en el análisis asintótico que requiere esta demostración, que veremos en la **Sección 2.3**.

Pearson tampoco tuvo en cuenta que, cuando los parámetros de la distribución esperada son estimados a partir de los datos, como ocurre en la familia normal, la distribución del estadístico de contraste cambia. Esto implica que el número de grados de libertad del test debería ser ajustado, cosa que Pearson no consideró. Este error dio lugar a aplicaciones incorrectas del test durante varios años.

1.3. Corrección del test: la importancia de Fisher

Ronald A. Fisher fue quien, en las décadas siguientes, abordó estas limitaciones desde una perspectiva matemática más rigurosa, principalmente en su artículo de 1924 (ver [7]). Fisher

entendió que, para que el test *ji-cuadrado* tuviese una base sólida, era necesario demostrar que su distribución límite era efectivamente la distribución χ^2, bajo ciertas condiciones.

Uno de sus aportes fundamentales fue la introducción del concepto de grados de libertad efectivos, que implica descontar del número de grados de libertad del test el número de parámetros estimados a partir de los datos. Esta corrección es clave para usar correctamente el test, ya que, como dijimos, afecta a la distribución con la que se compara el valor del estadístico observado.

Además, Fisher estableció condiciones para que la aproximación χ^2 sea válida, como que las frecuencias esperadas no sean demasiado pequeñas. Este tema fue objeto de mucha discusión posterior, especialmente cuando hay categorías con frecuencias muy bajas.

También popularizó el uso de los llamados p-valores como medida de evidencia contra la hipótesis nula. En vez de simplemente comparar un estadístico con un umbral crítico, propuso calcular la probabilidad de obtener un resultado tan extremo como el obtenido, suponiendo cierta la hipótesis nula. Esta idea sirvió de gran ayuda a la hora de interpretar el test.

En resumen, la aportación de Fisher no se basó únicamente en corregir los errores de Pearson, sino en reformular el enfoque desde una perspectiva más moderna. Gracias a él, el test dejó de ser una herramienta empírica con una justificación heurística para convertirse en un método estadístico bien fundamentado.

1.4. El aporte de Cochran: revisión crítica y aplicaciones del test

La contribución de Cochran en su artículo de 1952 (ver [3]) consistió en una revisión crítica, histórica y metodológica del test.

Aunque Fisher ya había aportado ciertas condiciones para que la aproximación χ^2 fuese correcta, fue Cochran quien formalizó criterios prácticos para asegurar la validez de esta aproximación. Algunos de estos criterios, como que todas las frecuencias esperadas sean mayores que 5, aún se siguen utilizando hoy en día.

Cochran también fue crítico con el uso excesivamente automático del test. Su revisión deja claro que el test *ji-cuadrado* puede ser mal aplicado si no se interpreta con cuidado.

Otra aportación importante fue su énfasis en el papel que juega el tamaño de la muestra. Cochran insistió en que, en muestras pequeñas, la distribución del estadístico puede desviarse notablemente de la distribución χ^2.

Cochran resume con bastante claridad las etapas por las que pasó el test: su nacimiento como idea intuitiva en Pearson, su formalización teórica con Fisher y su expansión como herramienta

práctica. Su artículo muestra que, incluso medio siglo después de su creación, el test $ji-cuadrado$ seguía requiriendo reflexión crítica y revisión de sus condiciones de aplicación.

Por todo esto, la revisión de Cochran no se centra únicamente en cerrar el recorrido histórico del test *ji-cuadrado*, sino que también ayuda a una comprensión moderna de su uso responsable.

Capítulo 2

Teoría sobre el test *ji-cuadrado*

En este capítulo formalizaremos teóricamente el test *ji-cuadradado* para el contraste de una hipótesis nula simple, con la expresión de su estadístico de contraste y con los principales resultados relacionados con el mismo. Dedicaremos también una sección a la demostración del teorema de convergencia del estadístico de contraste, resultado de gran importancia.

2.1. Preliminares matemáticos

En esta sección nos centraremos en definir los conceptos que nos harán falta para el desarrollo del trabajo.

Primero de todo, procederemos con definiciones básicas que serán utilizadas a lo largo del documento:

Definición 2.1. Sea X una variable aleatoria unidimensional continua con distribución F. Se define su esperanza como:

$$\mathbb{E}[X] = \int_{\mathbb{R}} xF(dx)$$

siempre que $\int_0^{+\infty} xF(dx) < \infty$ y $\int_{-\infty}^0 xF(dx) > -\infty$.

En el caso de que X sea una variable aleatoria unidimensional discreta, la esperanza de X viene dada por:

$$\mathbb{E}[X] = \sum_k x_k P(X = x_k),$$

siendo $\{x_1, x_2, ...\}$ los posibles valores de X.

La esperanza cumple las siguientes propiedades:

1. Sean $a, b \in \mathbb{R}$:
$$\mathbb{E}[a + bX] = a + b\mathbb{E}[X].$$

2. Si $(X_1, X_2)'$ es un vector aleatorio bidimensional:
$$\mathbb{E}[X_1 + X_2] = \mathbb{E}[X_1] + \mathbb{E}[X_2].$$

3. Si $(X_1, X_2)'$ es un vector aleatorio bidimensional y X_1 y X_2 son independientes:
$$\mathbb{E}[X_1 X_2] = \mathbb{E}[X_1]\mathbb{E}[X_2].$$

Definición 2.2. Si X es una variable aleatoria unidimensional, cuya media $\mu = \mathbb{E}[X]$ existe y es finita, se define su varianza como:
$$\mathrm{Var}(X) = \mathbb{E}[(X - \mu)^2] = \mathbb{E}[X^2] - \mu^2.$$

Sean $a, b \in \mathbb{R}$, entonces
$$\mathrm{Var}(a + bX) = b^2\mathrm{Var}(X).$$

Se define la desviación típica como la raíz cuadrada de la varianza, y se denota $\sigma = \sqrt{\mathrm{Var}(X)}$.

Definición 2.3. Se define la covarianza entre 2 variables aleatorias unidimensionales X e Y como $\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$. La covarianza cumple también las siguientes propiedades:

1. $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$.
2. Sean $a, b \in \mathbb{R}$, entonces: $\mathrm{Cov}(a + bX, Y) = b \cdot \mathrm{Cov}(X, Y)$.
3. $\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.
4. Si X e Y son independientes, entonces: $\mathrm{Cov}(X, Y) = 0$.
5. $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$.

Definición 2.4. Si $X = (X_1, ..., X_m)'$ es un vector aleatorio de dimensión m y existe la media de cada uno de sus componentes, se define el vector de medias de X como
$$\mathbb{E}(X) = \begin{pmatrix} \mathbb{E}(X_1) \\ \vdots \\ \mathbb{E}(X_m) \end{pmatrix}.$$

El vector de medias cumple las siguientes propiedades:

1. Sea X un vector aleatorio de dimensión m, $\alpha \in \mathbb{R}^p$ y A una matriz de dimensión $p \times m$. Entonces:
$$\mathbb{E}(\alpha + AX) = \alpha + A\mathbb{E}(X).$$

2. Si X_1 y X_2 son 2 vectores aleatorios de dimesnión m, y definidos sobre el mismo espacio de probabilidad, entonces:
$$\mathbb{E}(X_1 + X_2) = \mathbb{E}(X_1) + \mathbb{E}(X_2).$$

Definición 2.5. Sea $X = (X_1, ..., X_m)'$ un vector aleatorio de dimensión m de forma que cada componente tiene su media y varianza, se define la matriz de covarianzas de X como una matriz de dimensión $m \times m$:
$$\Sigma = \mathrm{Cov}(X, X) = \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_m) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \cdots & \mathrm{Cov}(_2, X_m) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_m, X_1) & \mathrm{Cov}(X_m, X_2) & \cdots & \mathrm{Var}(X_m) \end{pmatrix}.$$

Si el vector aleatorio X tiene como vector de medias (ver **Definición 2.4**) $\mu = \mathbb{E}(X)$, entonces se puede afirmar lo siguiente:
$$\mathrm{Cov}(X, X) = \mathbb{E}[(X - \mu)(X - \mu)'].$$

También se cumple que dado $v \in \mathbb{R}^m$:
$$\mathrm{Var}(v'X) = v'\Sigma v.$$

A continuación, definiremos la distribución de Bernoulli y la Binomial, las cuales son necesarias para la definición de la distribución multinomial que se dará posteriormente:

Definición 2.6. Se dice que una variable aleatoria X pertenece a una distribución de Bernoulli con parámetro $p \in (0, 1)$ ($X \in Bernoulli(p)$) si solo tiene 2 posibles valores: 1 (éxito), con probabilidad p, y 0 (fracaso), con probabilidad $1 - p$. Tendrá como esperanza $\mathbb{E}[X] = p$ y, como varianza, $\mathrm{Var}(X) = p(1 - p)$.

Definición 2.7. Se dice que una variable aleatoria X pertenece a una distribución binomial con parámetros n y p si representa el número de éxitos en n intentos independientes de una $Bernoulli(p)$ (ver **Definición 2.6**). Su soporte es $k \in \mathbb{Z}^+ \cup \{0\}$, su función de masa de probabilidad $f(k) = \binom{n}{k} p^k (1 - p)^{n-k}$, su esperanza $\mathbb{E}[X] = np$ y su varianza $\mathrm{Var}(X) = np(1 - p)$. Se denota $X \in Binomial(n, p)$.

También es esencial definir la distribución χ^2, que es fundamental para el procedimiento del test $ji - cuadrado$:

Definición 2.8. Se dice que una variable aleatoria $X \in \chi^2_k$ (χ^2 con k grados de libertad) si su función de densidad de probabilidad viene dada por la siguiente expresión:

$$f(x) = \begin{cases} \dfrac{1}{2^{k/2}\,\Gamma\left(\frac{k}{2}\right)}\, x^{\frac{k}{2}-1} e^{-x/2}, & x > 0, \\ 0, & x \leq 0 \end{cases}$$

donde $\Gamma(\cdot)$ es la función Gamma de Euler, definida por:

$$\Gamma(s) = \int_0^\infty t^{s-1} e^{-t}\, dt, \quad s > 0.$$

La esperanza de X será $\mathbb{E}[X] = k$ y la varianza será $\mathrm{Var}(X) = 2k$.

Definamos ahora en qué consiste la convergencia en distribución, que será muy utilizada en la **Sección 2.3**:

Definición 2.9. Se dice que $\{X_n\}$ (sucesión de vectores aleatorios de dimensión m) converge en distribución al vector aleatorio X de dimensión m si:

$$X_n \xrightarrow{d} X \Leftrightarrow \lim_{n\to\infty} F_n(x) = F(x), \forall x \in C_F,$$

siendo F_n la distribución de X_n, F la distribución de X y C_F el conjunto de puntos de continuidad de F.

También enunciaremos el siguiente teorema que será usado igualmente en la **Sección 2.3**:

Teorema 2.10 (de la aplicación continua)**.** *Sea $g : \mathbb{R}^m \to \mathbb{R}^k$ una aplicación continua, $\{X_n\}$ una sucesión de vectores aleatorios de dimensión m y X un vector aleatorio de dimensión m:*

$$\textit{Si } X_n \xrightarrow{d} X, \textit{ entonces } g(X_n) \xrightarrow{d} g(X).$$

Mencionemos ahora un resultado acerca de las matrices diagonales:

Proposición 2.11. *Sea D una matriz diagonal $n \times n$ con elementos en la diagonal $d_1, ..., d_n$, se denota $D = diag(d_1, ..., d_n) \in \mathbb{R}^{n\times n}$, de forma que $d_i \neq 0, \forall i \in \{1, ..., n\}$, entonces existe D^{-1}, que vendrá dada por:*

$$D^{-1} = diag(1/d_1, ..., 1/d_n) \in \mathbb{R}^{n\times n}.$$

Este resultado acerca de la inversión de cierto tipo de matrices, que podemos encontrar en la Sección 2.1 del libro de Golub y Van Loan (ver [10]), será también de uso en la **Sección 2.3**:

Teorema 2.12 (de Sherman-Morrison)**.** *Sea $A \in \mathbb{R}^{n\times n}$ una matriz invertible y $u \in \mathbb{R}^n$ de forma que $1 - u'A^{-1}u \neq 0$, entonces la matriz $A - uu'$ es invertible, y su inversa está dada por:*

$$(A - uu')^{-1} = A^{-1} + \frac{A^{-1}uu'A^{-1}}{1 - u'A^{-1}u}.$$

2.2. Formalización del test

El test de *ji-cuadrado* fue diseñado con la finalidad de poder probar si una muestra de datos proviene de una distribución específica F_0, lo que hace que la hipótesis nula sea $H_0 : F = F_0$ (contraste de hipótesis nula simple).

Consideraremos una muestra aleatoria simple $X_1, ...X_n$, donde cada observación X_i tiene la misma distribución F y todas n las observaciones son independientes.

Supongamos que la distribución F es discreta y denotemos por $A_1, ..., A_k$ una partición en conjuntos de F. Sea N_j el número de observaciones de la muestra pertenecientes a A_j. Entonces, se puede definir $(N_1, ..., N_k)'$ como el vector de frecuencias observadas.

La probabilidad de que una observación cualquiera de la muestra dada pertenezca a A_j bajo la distribución F_0 la denotaremos por p_j^0. Así, si se cumple H_0, cada $N_j \in Binomial(n, p_j^0)$ (teniendo en cuenta la **Definición 2.7**) y, por tanto, se cumpliría $\mathbb{E}[N_j] = np_j^0$. Esto indica que, bajo la hipótesis nula, el número de observaciones de la muestra esperadas en A_j sería np_j^0. Se puede definir entonces el vector de frecuencias esperadas: $(np_1^0, ..., np_k^0)'$.

Con todo esto, se puede definir el estadístico de contraste del test $ji - cuadrado$ como sigue:

$$D = \sum_{j=1}^{k} \frac{(N_j - np_j^0)^2}{np_j^0} \tag{2.1}$$

Estudiaremos ahora el caso en el que F es continua en $\mathbb{R}$:
Al ser F continua en todo $\mathbb{R}$, sabemos que su dominio será $\mathbb{R}$ y, por tanto, una partición de F en k conjuntos $A_1, ..., A_k$ será equivalente a dividir $\mathbb{R}$ en k intervalos $A_1, ...A_k$.

A continuación, se calculan las frecuencias observadas en cada intervalo, que se pueden denotar de igual forma que en el caso discreto por N_j (número de observaciones de la muestra en el intervalo A_j), $\forall j \in \{1, ..., k\}$. Utilizando el mismo razonamiento que en el anterior caso, se puede definir el vector de frecuencias esperadas por $(np_1^0, ..., np_k^0)'$, donde p_j^0 es la probabilidad bajo H_0 de que cierta observación de la muestra pertenezca al intervalo $A_j, \forall j \in \{1, ..., k\}$. Con todo esto, se puede definir el estadístico para el caso en el que F es continua mediante la Ecuación (2.1), utilizada para el caso discreto.

El estadístico D mide la discrepancia entre lo observado y lo esperado bajo H_0, como dijimos. Este test nos llevará a rechazar la hipótesis nula cuando D toma un valor grande, es decir, rechazaremos $H_0 : F = F_0$ cuando $D > c$ para cierta constante c que depende del nivel de significación y se obtiene de la distribución de D. Para la expresión de la distribución de D tomaremos una aproximación asintótica, la cual veremos a continuación después de proceder con

ciertos enunciados.

Pero antes, debemos definir la distribución multinomial, que se obtiene a partir de la **Definición 2.7** y tiene gran relación con el test *ji-cuadrado*:

Definición 2.13. Si cierto experimento aleatorio presenta k resultados posibles, incompatibles dos a dos, con $p_1, ..., p_k$ probabilidades respectivas de cada uno de los posibles resultados, y se realiza n veces de forma independiente, entonces se puede afirmar que el vector aleatorio $(Y_1, ..., Y_k)'$, el cual contiene las frecuencias observadas de cada posible resultado, tiene distribución multinomial con parámetros n como número de intentos, y $p_1, ..., p_k$ como probabilidades. Se denota en este caso $(Y_1, ..., Y_k)' \in Multinomial(n; p_1, ..., p_k)$

Teniendo en cuenta la **Definición 2.13** y considerando el ejemplo mencionado en la **Sección 1.1**, podemos afirmar que $(C_1, ..., C_6)' \in Multinomial(60; 1/6, ..., 1/6)$ suponiendo que el dado no está trucado, siendo $(C_1, ..., C_6)'$ el vector que acumula el número de veces que sale cada cara en los 60 lanzamientos.

Una vez definida la distribución multinomial, debemos presentar ciertas propiedades que cumple esta distribución, las cuales podemos encontrar en [5]:

Proposición 2.14. *Supongamos que $(Y_1, ..., Y_k)' \in Multinomial(n; p_1, ..., p_k)$. Entonces se verifican las siguientes propiedades:*

1. $Y_j \in Binomial(n, p_j)$, $\forall j \in \{1, ..., k\}$.
2. $\mathbb{E}(Y_j) = np_j, \forall j \in \{1, ..., k\}$.
3. $Var(Y_j) = np_j(1 - p_j), \forall j \in \{1, ..., k\}$.
4. $Cov(Y_j, Y_l) = -np_jp_l, \forall j, l \in \{1, ..., k\}$ *con* $j \neq l$.
5. *El vector aleatorio $(Y_1, ..., Y_k)'$ es discreto y su función de masa de probabilidad viene dada por $P(Y_1 = y_1, ..., Y_k = y_k) = \frac{n!}{y_1!...y_k!} p_1^{y_1}...p_k^{y_k}$, donde los valores posibles de la distribución son los $y_1, ...y_k \in \{0, 1, ..., n\}$ tales que $\sum_{j=1}^{k} y_j = n$.*

Podemos proceder, por tanto, con el resultado que expresa la distribución asintótica del estadístico de *ji-cuadrado* D y el cual será probado en la **Sección 2.3**:

Teorema 2.15. *Sea $X_1, ..., X_n$ una muestra aleatoria simple de datos donde todas las observaciones pertenecen a una distribución F, sea $H_0 : F = F_0$, $A_1, ..., A_k$ una partición en conjuntos de F, $(N_1, ...N_k)'$ el vector de frecuencias observadas, de forma que N_j representa el número de observaciones pertenecientes al conjunto A_j, y $(np_1^0, ..., np_k^0)'$ el vector de frecuencias esperadas,*

donde p_j^0 *representa la probabilidad bajo* F_0 *de que una observación pertenezca a* A_j*. Bajo* H_0 *y suponiendo* $0 < p_i^0 < 1, \forall i \in \{1, ..., k\}, \sum_{i=1}^k p_i^0 = 1$*,*

$$D = \sum_{j=1}^{k} \frac{(N_j - np_j^0)^2}{np_j^0} \xrightarrow{d} \chi^2_{k-1},$$

cuando $n \to \infty$*.*

Utilizando el **Teorema 2.15** se puede tomar $c = \chi^2_{k-1,\beta}$ de forma que $\lim_{n\to+\infty} \mathbb{P}(D > \chi^2_{k-1}) = \beta$. Teniendo esto en cuenta, el test consistirá en:

$$\text{Rechazar } H_0 : F = F_0 \text{ si } D = \sum_{j=1}^{k} \frac{(N_j - np_j^0)^2}{np_j^0} > \chi^2_{k-1,\beta}.$$

2.3. Teorema de convergencia del estadístico *ji-cuadrado*

En esta sección procederemos con una demostración del teorema enunciado previamente (ver **Teorema 2.15**) que podemos encontrar en la Sección 2 del Capítulo 10 de [9], la cual se centra en el caso $k \geq 2$ y a la cual añadiremos ciertas puntualizaciones, como referencias a enunciados de gran importancia, con la finalidad de que sea una demostración más formalmente correcta.

La demostración basará gran parte de su contenido en el **Teorema Central del Límite Multivariante**, del cual podemos encontrar una versión en la Sección 3 de [8], y cuyo enunciado es el que veremos a continuación:

Teorema. *Sea* $\{X_n\}$ *una sucesión de vectores aleatorios mutuamente independientes con la misma distribución de cierto vector* X*, con vector de medias* $\mu = \mathbb{E}(X)$ *y matriz de covarianzas* $\Sigma = Cov(X, X)$*. Entonces se tiene lo siguiente:*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu) \xrightarrow{d} N_m(0, \Sigma).$$

Procedamos, por tanto, con la prueba del teorema en cuestión:
Sea $\xi_r = (I_{\{X_r \in A_1\}}, ..., I_{\{X_r \in A_{k-1}\}})'$ el vector aleatorio de dimensión $k-1$ que indica a qué conjunto $A_1, ..., A_{k-1}$ pertenece la observación X_r.

Así, $N = (N_1, ..., N_{k-1})' = \sum_{r=1}^n \xi_r$. Cada ξ_r tendrá como vector de medias $p = (p_1^0, ..., p_{k-1}^0)'$ y $\Sigma = (\sigma_{ij})$ como matriz de covarianzas, donde cada σ_{ij} proviene de la siguiente expresión:

$$\sigma_{ij} = \mathbb{E}[I_{\{X_r \in A_i\}} I_{\{X_r \in A_j\}}] - p_i^0 p_j^0 = \begin{cases} -p_i^0 p_j^0, & \text{si } i \neq j \\ p_i^0(1 - p_i^0), & \text{si } i = j \end{cases} \tag{2.2}$$

Esta expresión proviene de utilizar el tercer apartado de la **Definición 2.3**. Utilizando lo mencionado, obtenemos $\sigma_{ij} = \mathbb{E}[I_{\{X_r \in A_i\}} I_{\{X_r \in A_j\}}] - \mathbb{E}[I_{\{X_r \in A_i\}}]\mathbb{E}[I_{\{X_r \in A_j\}}]$. Empleando también que $\mathbb{E}[I_{\{X_r \in A_i\}}] = p_i^0, \forall i \in \{1, ..., k-1\}$ bajo H_0 y que $I_{\{X_r \in A_i\}} I_{\{X_r \in A_j\}} = 0, i \neq j$, obtenemos la expresión vista en la Ecuación (2.2).

Probaremos a continuación que la matriz Σ es invertible. Llevaremos a cabo esta prueba para poder hacer uso de su inversa Σ^{-1}, la cual usaremos más adelante.

Teniendo en cuenta que $\mathbb{E}[\xi_r] = p$, Σ se puede expresar como $\Sigma = \mathbb{E}[(\xi_r - p)(\xi_r - p)']$ (siguiendo la **Definición 2.5**). Considerando ahora la expresión vista en la Ecuación (2.2), se puede afirmar que $\Sigma = \text{diag}\ (p_1^0, ..., p_{k-1}^0) - pp'$.

Denotando por D a $\text{diag}(p_1^0, ..., p_{k-1}^0)$ y teniendo en cuenta que está en las hipótesis de la **Proposición 2.11**, podemos afirmar que es invertible y que su inversa es $D^{-1} = \text{diag}(1/p_1^0, ..., 1/p_{k-1}^0)$.

Empleando la expresión de Σ dada previamente, podemos asegurar que $\Sigma = D - pp'$. Esta matriz claramente cumple las hipótesis del **Teorema 2.12**, ya que vimos que D es invertible y sabemos que $1 - p'D^{-1}p = 1 - \sum_{i=1}^{k-1} p_i^0 = p_k \neq 0$. Con todo esto, podemos garantizar que la matriz Σ es invertible.

Definamos ahora $\eta_r = \Sigma^{-1/2}(\xi_r - p)$. Cada η_r será un vector aleatorio con I como matriz de covarianzas (matriz identidad $(k-1) \times (k-1)$) y 0 como vector de medias. Este vector se define con la forma dada para facilitar el uso del **Teorema Central del Límite Multivariante**, que será empleado a continuación.

Aplicando el teorema en cuestión a la sucesión de vectores aleatorios $\{\eta_r\}$, que claramente están en las hipótesis del teorema ya que cada η_r sigue la misma distribución con vector de medias 0 y matriz de covarianzas I, obtenemos:

$$\Sigma^{-1/2} \frac{N - p}{\sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{r=1}^{n} \eta_r \xrightarrow{d} N_{k-1}(0, I)$$

Para la primera igualdad se utiliza que $N = \sum_{r=1}^{n} \xi_r$ y la definición de cada η_r.

Utilizando a continuación el **Teorema de la aplicación continua** (ver **Teorema 2.10**), podemos obtener:

$$\left(\frac{N - p}{\sqrt{n}}\right)' \Sigma^{-1} \left(\frac{N - p}{\sqrt{n}}\right) \xrightarrow{d} \chi_{k-1}^2, \tag{2.3}$$

donde aplicamos el teorema mencionado a la expresión $\Sigma^{-1/2} \frac{N-p}{\sqrt{n}}$ con $g : \mathbb{R}^{k-1} \to \mathbb{R} \mid g(c) = ||c||^2$, que claramente es una aplicación continua.

El final de la demostración consiste en dar una expresión explícita de la matriz Σ^{-1}, la cual es la siguiente:

$$\Sigma^{-1} = \begin{pmatrix} 1/p_1^0 + 1/p_k^0 & 1/p_k^0 & \cdots & 1/p_k^0 \\ 1/p_k^0 & 1/p_2^0 + 1/p_k^0 & \cdots & 1/p_k^0 \\ \vdots & \vdots & \ddots & \vdots \\ 1/p_k^0 & 1/p_k^0 & \cdots & 1/p_{k-1}^0 + 1/p_k^0 \end{pmatrix}$$

Esta expresión proviene de tener en cuenta la hipótesis inicial de que $\sum_{i=1}^{k} p_i^0 = 1$ y $0 < p_i^0 < 1, \forall i \in \{1, ..., k\}$, la cual implica que $p_k^0 = 1 - \sum_{i=1}^{k-1} p_i^0$, y también el **Teorema 2.12** considerando $A = D = \text{diag}(p_1^0, ..., p_{k-1}^0)$ y $u = p$. Estamos ante las hipótesis del teorema, como comprobamos previamente. Así, obtenemos que $\Sigma^{-1} = D^{-1} + 1/p_k^0 \cdot I$ (teniendo en cuenta que $D^{-1}pp'D^{-1} = (I)_{k-1\times k-1}$).

Para llegar finalmente a la convergencia del estadístico de $ji-cuadrado\ D$, la prueba se basa en desarrollar la expresión la cual vimos que converge a una χ^2_{k-1} (ver Ecuación (2.3)):

$$\left(\frac{N-p}{\sqrt{n}}\right)' \Sigma^{-1} \left(\frac{N-p}{\sqrt{n}}\right) =$$

$$\begin{aligned} &= \sum_{i=1}^{k-1} \left(\frac{1}{p_i^0} + \frac{1}{p_k^0}\right) \frac{(N_i - np_i^0)^2}{n} + \sum_{i\neq j} \frac{1}{p_k^0} \frac{(N_i - np_i^0)(N_j - np_j^0)}{n} \\ &= \sum_{i=1}^{k-1} \frac{(N_i - np_i^0)^2}{np_i^0} + \frac{1}{np_k^0} \left(\sum_{i=1}^{k-1} (N_i - np_i^0)\right)^2 \\ &= \sum_{i=1}^{k} \frac{(N_i - np_i^0)^2}{np_i^0} \end{aligned}$$

donde la última igualdad se obtiene teniendo en cuenta que $N_k = n - \sum_{i=1}^{k-1} N_i$ y $p_k^0 = 1 - \sum_{i=1}^{k-1} p_i^0$, lo que da lugar a que $N_k - np_k^0 = n - \sum_{i=1}^{k-1} N_i - n(1 - \sum_{i=1}^{k-1} p_i^0) = \sum_{i=1}^{k-1}(np_i^0 - N_i)$.

Así llegamos a probar que, en efecto, el estadístico del test $ji - cuadrado\ D$ (ver Ecuación (2.1)) converge a la distribución χ^2_{k-1}.

Capítulo 3

Test *ji-cuadrado* para familias paramétricas

Para desarrollar el contenido de este capítulo, nos basaremos en lo expuesto en el Capítulo 2 del libro de 1996 de Greenwood y Nikulin (ver [11]).

El test *ji-cuadrado* también se puede emplear para contrastar una hipótesis nula del tipo:

$$H_0 : F \in \{F_\theta \mid \theta = (\theta_1, ..., \theta_q)' \in \Theta \subset \mathbb{R}^q\}, q \geq 1,$$

en el que se comprueba si la distribución de cierta muestra de datos $X_1, ..., X_n$ pertenece al modelo paramétrico F_θ, donde θ representa el parámetro del modelo. Como se comenta en la **Sección 1.2**, Pearson ya consideraba este problema en su artículo de 1900 ([13]) cuando analizó el carácter normal de las muestras de datos de Airy y Merriman.

Por ejemplo, en el problema de contrastar si una muestra de datos pertenece a la distribución normal, se denotaría la hipótesis nula por $H_0 : F \in \{N(\mu, \sigma^2) \mid (\mu, \sigma^2)' \in \Theta = \{(x, y) \in \mathbb{R}^2 : y > 0\} \subset \mathbb{R}^2\}$, donde como vemos q sería 2 ya que tenemos 2 parámetros μ y σ^2.

Para resolver este tipo de contrastes se debe estimar el parámetro θ. Dividiremos esta sección en 3 casos distintos según cuándo se estime el parámetro y según cómo se agrupen los datos:

1. Caso en el que se estima el parámero con los datos ya agrupados.

2. Caso en el que se estima el parámetro con los datos sin agrupar.

3. Caso en el que se escojen las categorías a partir de los datos.

3.1. Parámetro estimado con los datos ya agrupados

De forma similar a en la **Sección 2.2**, se considera una partición del soporte de la variable aleatoria $A_1, ..., A_k$ y se denota N_j a la frecuencia observada en A_j. Se definirá también, para un cierto θ, $p_j(\theta)$ como la probabilidad bajo F_θ de que una observación pertenezca a A_j.

Para conocer las frecuencias esperadas se debe estimar θ, ya que es desconocido. Esta estimación se realizará a partir de los datos agrupados, es decir, a partir del vector de frecuencias observadas $(N_1, ..., N_k)'$. Se puede usar alguno de los siguientes métodos, los cuales son asintóticamente equivalentes, para ello:

- El estimador de mínimos cuadrados, que busca minimizar la diferencia entre las frecuencias observadas y las esperadas bajo H_0, donde se estima θ mediante $\hat{\theta}$, el cual se obtiene de $\sum_{j=1}^{k} \frac{(N_j - np_j(\hat{\theta}))^2}{np_j(\hat{\theta})} = \text{mín}_{\theta\in\Theta} \sum_{j=1}^{k} \frac{(N_j - np_j(\theta))^2}{np_j(\theta)}$.
- El estimador de máxima verosimilitud, que busca maximizar la combinación total de las probabilidades de los conjuntos $A_1, ..., A_k$, donde $\hat{\theta}$ (estimador de θ) maximiza la función $\frac{n!}{\prod_{j=1}^{k} N_j} \prod_{j=1}^{k} p_j^{N_j}(\theta)$.
- El estimador de mínimos cuadrados modificado, similar al estimador de mínimos cuadrados pero con la diferencia de que en el denominador se utilizan las frecuencias observadas, que obtiene $\hat{\theta}$ de forma que $\sum_{j=1}^{k} \frac{(N_j - np_j(\hat{\theta}))^2}{N_j} = \text{mín}_{\theta\in\Theta} \sum_{j=1}^{k} \frac{(N_j - np_j(\theta))^2}{N_j}$.

Suponiendo θ estimado con alguno de los métodos comentados previamente, $\hat{\theta}$, se calculan las probabilidades $p_j(\hat{\theta})$ de cada conjunto A_j bajo la distribución $F_{\hat{\theta}}$. Así, el vector de frecuencias esperadas será $(np_1(\hat{\theta}), ..., np_k(\hat{\theta}))'$. El estadístico $ji - cuadrado$ se calcula de la siguiente forma:

$$D(\hat{\theta}) = \sum_{j=1}^{k} \frac{(N_j - np_j(\hat{\theta}))^2}{np_j(\hat{\theta})}.$$

Debemos definir las siguientes condiciones, dadas por Cramer en 1946 (ver [6]), que se tendrán en cuenta para estudiar la distribución asintótica del estadístico en este caso:

Criterio 3.1. *Las funciones $p_1(\theta), \ldots, p_k(\theta)$ verifican lo siguiente:*

1. $\sum_{j=1}^{k} p_j(\theta) = 1, \theta \in \Theta$.
2. $\exists c > 0 \mid p_i(\theta) > c > 0, i = 1, ..., k$.
3. $\frac{\partial^2 p_i(\theta)}{\partial\theta_l \partial\theta_j}, l, j = 1, ..., q$, *son funciones continuas* $\forall i \in \{1, ..., k\}$.
4. *El rango de la matriz* $T = T(\theta) = (\frac{\partial p_i(\theta)}{\partial\theta_j})_{ij}$ *es* q.

5. *El rango de* $B = B(\theta)$ *es* q*, donde* $B = \begin{pmatrix} 1/\sqrt{p_1(\theta)} \cdot \frac{\partial p_1(\theta)}{\partial \theta_1} & \cdots & 1/\sqrt{p_1(\theta)} \cdot \frac{\partial p_1(\theta)}{\partial \theta_q} \\ \vdots & \ddots & \vdots \\ 1/\sqrt{p_k(\theta)} \cdot \frac{\partial p_k(\theta)}{\partial \theta_1} & \cdots & 1/\sqrt{p_k(\theta)} \cdot \frac{\partial p_k(\theta)}{\partial \theta_q} \end{pmatrix}$.

A partir de ahora, en esta sección supondremos ciertas las condiciones 1-4 del **Criterio 3.1**. Debemos enunciar también un resultado esencial para la distribución asintótica:

Teorema 3.2. *Sea* $\{\hat{\theta_n}\}$ *la secuencia de estimadores de* θ_0*, siendo* θ_0 *el valor verdadero de* θ *que generó la muestra, obtenida mediante alguno de los métodos comentados. Entonces, si* $X_1, \ldots, X_n$ *siguen la distribución* F_{θ_0}*:*

$$\lim_{n\to\infty} \sqrt{n}(\hat{\theta_n} - \theta_0) = J^{-1}(\theta_0)B'(\theta_0)X_n(\theta_0) + o_p(1), \tag{3.1}$$

donde

$$J(\theta) = B'(\theta)B(\theta) \tag{3.2}$$

es la matriz de información y $o_p(1)$ *es un vector aleatorio que converge a* 0 *en probabilidad.*

Podemos, por tanto, enunciar el teorema de convergencia que nos muestra la distribución asintótica del estadístico para este primer caso:

Teorema 3.3. *Si se cumple* H_0 *y* $\hat{\theta}$ *es un estimador que cumple la Ecuación (3.1), la distribución asintótica del estadístico* $D(\hat{\theta})$ *es:*

$$D(\hat{\theta}) = \sum_{j=1}^{k} \frac{(N_j - np_j(\hat{\theta}))^2}{np_j(\hat{\theta})} \xrightarrow{d} \chi^2_{k-q-1}.$$

En el **Teorema 3.3** q representa el número de parámetros estimados. Como mencionamos antes, en el caso de la distribución normal q sería 2. Para una $Poisson(\lambda)$, por ejemplo, q sería 1 ya que sólo se estima el parámetro λ.

3.2. Parámetro estimado con los datos sin agrupar

¿Qué diferencia este caso con el estudiado en la sección anterior? Para aclararlo, tomaremos como ejemplo la distribución normal. En este caso, sabemos que la media y la cuasi-varianza muestrales son estimadores consistentes de μ y σ^2. Entonces, parece lógico contrastar normalidad construyendo el estadístico *ji-cuadrado* a partir de estos dos estimadores. A pesar de que esta idea parece bastante intuitiva, es bastante más difícil de tratar teóricamente que la situación anterior.

Definamos, en adición a las condiciones 1-4 del **Criterio 3.1**, las siguientes condiciones que serán de uso para poder definir la distribución asintótica del estadístico en este caso:

Criterio 3.4. *1. La función de distribución $F_\theta(x) = F(x;\theta)$ tiene como función de densidad de probabilidad $f(x;\theta) = \frac{d}{dx}F(x;\theta), x \in \mathbb{R}, \theta \in \Theta$, de forma que las derivadas $\frac{\partial^2 f(x;\theta)}{\partial\theta_i \partial\theta_j}$ existen y son continuas $\forall i, j \in \{1, ..., q\}$.*

2. La matriz de información de Fisher

$$I(\theta) = \mathbb{E}_\theta[\Lambda_j \Lambda_j'], \tag{3.3}$$

donde $\mathbb{E}_\theta$ se refiere a la esperanza bajo la distribución F_θ, correspondiente a cualquier observación X_j, es finita y definida positiva para cualquier $\theta \in \Theta$, donde $\Lambda_j = \mathbf{grad}_\theta \log f(X_j;\theta) = (\frac{\partial \log f(X_j;\theta)}{\partial\theta_1}, ..., \frac{\partial \log f(X_j;\theta)}{\partial\theta_q})'$.

3. $\int_{-\infty}^{\infty} \frac{\partial}{\partial\theta_i} f(x;\theta)dx = 0, i = 1, ..., q$.

Utilizaremos, como estimador de θ, el estimador de máxima verosimilitud basado en la muestra de datos sin agrupar, el cual necesita las condiciones del **Criterio 3.4** para poder garantizar su existencia y consistencia.

Sea $L(\theta) = \prod_{j=1}^n f(X_j;\theta)$ la función de verosimilitud de la muestra. Denotemos, a continuación, $\Lambda(\theta) = \mathbf{grad}_\theta \log L(\theta) = \sum_{j=1}^n \mathbf{grad}_\theta \log f(X_j;\theta) = \sum_{j=1}^n \Lambda_j(\theta)$. Supongamos $\{\hat{\theta_n}\}$ una sucesión de estimadores de máxima verosimilitud de θ. Entonces, está claro que $\Lambda(\hat{\theta_n}) = 0, \forall n$, ya que el método de máxima verosimilitud busca, en este caso, aquellos parámetros que maximizan $L(\theta)$.

De la misma forma que en el caso anterior, denotamos $A_1, ..., A_k$ una partición del soporte de la variable aleatoria. Se define entonces N_j como el número de observaciones de la muestra pertenecientes a A_j.

Si $\hat{\theta}$ es el estimador de máxima verosimilitud de θ, se puede definir el estadístico del test, en este caso, como sigue:

$$D(\hat{\theta}) = \sum_{j=1}^{k} \frac{(N_j - np_j(\hat{\theta}))^2}{np_j(\hat{\theta})},$$

donde $p_j(\hat{\theta})$ es la probabilidad bajo $F_{\hat{\theta}}$ de que una observación pertenezca a A_j.

Enunciaremos ahora el teorema de convergencia del estadístico en este caso:

Teorema 3.5 (de Chernoff y Lehmann, 1954 ([2]))**.** *Sea $\hat{\theta}$ el estimador de máxima verosimilitud de θ_0, valor real de θ. Si se cumplen las condiciones 1-4 del **Criterio 3.1**, las condiciones del **Criterio 3.4** y $X_1, ..., X_n$ siguen la distribución F_{θ_0}, entonces:*

$$\lim_{n\to\infty} D(\hat{\theta}) = \chi^2_{k-q-1} + \sum_{j=1}^{q} \lambda_j \xi_j^2,$$

donde $\chi^2_{k-q-1}, \xi_1, ..., \xi_q$ son independientes, $\xi_i \in N(0,1), \forall i \in \{1, ..., q\}$, y $\lambda_1, ..., \lambda_q$ son raíces de $|(1-\lambda)I(\theta_0) - J(\theta_0)| = 0$ ($J(\theta)$ definida en la Ecuación (3.2) y $I(\theta)$ en la Ecuación (3.3)).

Como vemos en el **Teorema 3.5**, la distribución límite de $D(\hat{\theta})$ en este caso depende de θ_0, que generalmente es desconocido, lo que hace que calibrar este test bajo la hipótesis nula sea imposible.

3.3. Categorías a partir de los datos

Estudiaremos ahora el tercer caso mencionado para el modelo paramétrico, en el que las categorías se escogerán a partir de los datos.

Para ilustrar este caso haremos uso de la distribución normal, como antes. En esta situación, con las probabilidades prefijadas (por ejemplo $1/k$), a diferencia de en los casos anteriores, las categorías dependen de la muestra, es decir, si se toman como estimadores de μ y σ^2 la media y varianza muestrales $\bar{X}_n$ y S^2, respectivamente, las categorías se podrán tomar como cuantiles de $N(\bar{X}_n, S^2)$.

Nos centraremos en el estudio para el caso en el que F sea continua, lo que convierte las categorías en intervalos. Se define primero un vector de probabilidades $(p_1, ..., p_k)'$ tal que $p_i > 0, \forall i \in \{1, ..., k\}, \sum_{j=1}^{k} p_j = 1$. Denotamos, a partir de esta definición, $u_i = p_1 + ... + p_i, i = 1, ..., k$.

Considerando $F^{-1}(u;\theta) = \text{ínf}\{x : F(x;\theta) \geq u\}$, siendo $F(x;\theta) = F_\theta(x)$, podemos tomar $a_i(\theta) = F^{-1}(u_i;\theta), i = 1, ..., k$. Denotemos también $p_i(\theta', \theta) = \int_{a_{i-1}(\theta')}^{a_i(\theta')} dF(x;\theta), i = 1, ..., k$ y $a_0(\theta) = -\infty, a_k(\theta) = +\infty, \forall \theta, \theta' \in \Theta$. Sea θ_0 el valor real del parámetro θ, si definimos $p_i(\theta) = p_i(\theta_0, \theta)$, entonces se tiene: $p_i(\theta_0) = p_i, i = 1, ..., k,$.

Si suponemos θ^* un estimador de θ obtenido a partir de la muestra de datos, podemos definir los siguientes intervalos en los que dividir $\mathbb{R}$:

$$(a_0(\theta^*) = -\infty, a_1(\theta^*)), [a_1(\theta^*), a_2(\theta^*)), ..., [a_{k-1}(\theta^*), a_k(\theta^*) = +\infty).$$

Con esta división, se define $(N_1^*, ...N_k^*)'$ como el vector de frecuencias observadas, de forma que N_j^* recoge el número de observaciones de la muestra $X_1, ...X_n$ que caen en el intervalo $[a_{j-1}(\theta^*), a_j(\theta^*))$. A continuación, podemos tomar $(np_1(\theta^*, \theta^*), ..., np_k(\theta^*, \theta^*))'$ como el vector de frecuencias esperadas. Con todo esto, se puede definir el estadístico de $ji - cuadrado$ como sigue:

$$D_R(\theta^*) = \sum_{j=1}^{k} \frac{(N_j^* - np_j(\theta^*, \theta^*))^2}{np_j(\theta^*, \theta^*)}.$$

Será imprescindible que las funciones $a_i(\theta), \theta \in \Theta$, sean continuas con derivadas parciales continuas para la convergencia asintótica del estadístico en este caso, por lo que supondremos esto cierto para poder enunciar el teorema de convergencia sin impedimentos.

Podemos, entonces, enunciar el teorema de convergencia asintótica de $D_R(\theta^*)$, siendo θ^* un estimador de θ_0 que cumple lo siguiente, considerando que $X_1, ..., X_n$ siguen la distribución F_{θ_0}:

$$\theta^* - \theta_0 = \frac{1}{n}\sum_{j=1}^{n} v(X_j) + o_p(1),$$

donde $v = (v_1, ..., v_q)'$ es un vector de influencias asociadas con θ^* de forma que $\mathbb{E}_{\theta_0}(v(X_j)) = 0$,

$$\mathrm{Var}_{\theta_0} v(X_j) = V \tag{3.4}$$

es finita y $o_p(1)$ es un vector aleatorio que converge a 0 en probabilidad. En el caso de la distribución normal ($N(\mu, \sigma^2)$), si estimamos μ y σ^2 mediante la media y varianza muestrales $\bar{X_n}$ y S^2, el vector de influencias para una observación X_j seguirá la siguiente expresión: $v(X_j) = (X_j - \mu, (X_j - \mu)^2 - \sigma^2)'$.

Enunciemos el teorema de convergencia en cuestión:

Teorema 3.6. *Si θ^* es un estimador de θ_0 obtenido a partir de la muestra de datos $X_1, ..., X_n$, de la que todas las observaciones siguen la distribución F_{θ_0}, se tiene:*

$$\lim_{n\to\infty} D_R(\theta^*) = \lambda_1\xi_1^2 + ... + \lambda_k\xi_k^2,$$

donde $\xi_1, ..., \xi_k$ son variables aleatorias mutuamente independientes, $\xi_i \in N(0,1), \forall i \in \{1, ..., k\}$, y $\lambda_1, ..., \lambda_k$ son los autovalores de la matriz $P^{-1}\Sigma$, donde $P = P(\theta_0)$ es la matriz diagonal con elementos en la diagonal principal $p_1(\theta_0), ..., p_k(\theta_0)$, $\Sigma = P - pp' - UU' - W'U + U'VU$, donde

$$p = p(\theta_0) = (p_1(\theta_0), ..., p_k(\theta_0))' = (p_1, ..., p_k)',$$

$$U' = (U_{ij})_{k\times q}, U_{ij} = U_{ij}(\theta) = \int_{a_{i-1}(\theta)}^{a_i(\theta)} \frac{\partial f(x;\theta)}{\partial \theta_j} dx,$$

$$W' = (W_{ij})_{k\times q}, W_{ij} = W_{ij}(\theta) = \int_{a_{i-1}(\theta)}^{a_i(\theta)} v_j(x) f(x;\theta) dx,$$

y V viene dada por la Ecuación (3.4).

Como pasaba en el **Teorema 3.5**, en el **Teorema 3.6** la distribución asintótica del estadístico de contraste $D_R(\theta^*)$ depende, de forma general, de θ_0, lo que hace que no sea posible calibrar el test en esta situación bajo H_0.

Sin embargo, hay una excepción para la que la distribución asintótica de $D_R(\theta^*)$ no depende de θ_0 cuando $q = 2$. Este es el caso de las familias de escala y localización.

Bajo la hipótesis nula, la función densidad de probabilidad $f(x;\theta)$ debe pertenecer a la familia de escala y localización

$$\{f(x;\theta) = \frac{1}{\sqrt{\theta_2}} f\left(\frac{x-\theta_1}{\sqrt{\theta_2}}\right) \mid \theta = (\theta_1, \theta_2)' \in \Theta = \{(\theta_1, \theta_2)' \mid \theta_1 \in \mathbb{R}, \theta_2 > 0\}\}, \tag{3.5}$$

donde θ_1 y $\sqrt{\theta_2}$ son parámetros de localización y escala, respectivamente, y $f(x;\theta)$ es continua en x y diferenciable con respecto a los parámetros. Además, se debe cumplir también

$$\mathbb{E}[X_i] = \theta_1, \mathrm{Var}(X_i) = \theta_2, \tag{3.6}$$

para cualquier observación X_i.

Aplicando el **Teorema 3.6** a una familia que siga la Ecuación (3.5), cumpliendo también la Ecuación (3.6), los autovalores $\lambda_1, ..., \lambda_k$ no dependen de θ_0 si se toma, a diferencia de hasta ahora, $a_i(\theta) = \theta_1 + c_i\sqrt{\theta_2}, c_i \in \mathbb{R}, i = 1, ..., k$. Esto hará que en este caso el test sí se pueda calibrar bajo la hipótesis nula.

Por ejemplo, la familia de distribuciones $N(\theta_1, \theta_2)$ claramente cumple la Ecuación (3.6) y la Ecuación (3.5). Así, si tomamos $\theta^* = (\theta_1^*, \theta_2^*)'$ como estimador de $\theta = (\theta_1, \theta_2)'$, de forma que $\theta_1^* = \bar{X_n} = \frac{1}{n}\sum_{i=1}^n X_i$ y $\theta_2^* = S^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X_n})^2$, y definimos $a_i(\theta^*) = \bar{X_n} + c_i \cdot S, c_i \in \mathbb{R}, i = 1, ..., k$, entonces los autovalores $\lambda_1, ..., \lambda_k$ de la matriz $P^{-1}\Sigma$ (**Teorema 3.6**) no dependerán de θ_0 y, por tanto, se podrá calibrar el test bajo H_0.

Capítulo 4

Simulación en *R* del test

Este capítulo estará centrado en la simulación en R del test $ji - cuadrado$. El **Teorema 2.15** garantiza que la distribución del estadístico D de la Ecuación (2.1) es una χ^2 cuando n es suficientemente grande. El problema está en que no sabemos cuánto es suficientemente grande y, por tanto, no sabemos a partir de qué n la aproximación es útil en la práctica. Es por eso que es necesario comprobar su funcionamiento empírico. Nos ayudaremos para ello de la siguiente condición, de la que veremos si realmente tiene algún fundamento para su existencia:

Criterio 4.1. *Para que la aproximación de la distribución por la ji-cuadrado sea buena, es conveniente que las frecuencias esperadas sean todas suficientemente grandes. Como criterio, se aconseja lo siguiente:*

$np_j^0 \geq 5, \forall j \in \{1, ..., k\}$.

Esta condición, también considerada una recomendación, la podemos encontrar en la Sección 10 de [3].

También analizaremos en este capítulo, por medio de la simulación, la potencia del test, que nos indica la capacidad del test de detectar los casos en los que H_0 no es cierta, es decir, cuando F y F_0 son distintas.

Aparte de esto, dedicaremos una sección a la simulación del test para el modelo paramétrico, el cual estudiamos teóricamente en el **Capítulo 3**.

4.1. Calibrado del test

Para comprobar la certeza del criterio en cuestión (**Criterio 4.1**) consideraremos el caso en el que se cumple H_0 ($F = F_0$), simularemos muchas muestras bajo H_0 (tomaremos 1000 muestras)

con varios tamaños muestrales n distintos (probaremos con 5, 10, 20, 30, 40 y 50), aplicaremos el test con $k = 5$ categorías y comprobaremos si el porcentaje de muestras rechazadas es próximo al nivel de significación (trataremos con los habituales niveles de significación: 0.01, 0.05 y 0.1).

Para estudiar tanto el caso continuo como el discreto para F_0, llevaremos a cabo la simulación para una distribución continua, una normal estándar ($N(0,1)$), y también para una distribución discreta, una Poisson de media 2 ($Poisson(2)$).

4.1.1. Caso continuo: *N(0,1)*

Para que la recomendación mencionada (**Criterio 4.1**) funcione bien en esta simulación, el test deberá estar bien calibrado para tamaños muestrales de $n = 30$ en adelante. Esto se debe a que si cogemos, por ejemplo, A_j de forma que $p_j^0 = \frac{1}{k} = 1/5, \forall j \in \{1, ..., 5\}$, obtendríamos $n \cdot p_j^0 > 5, \forall j \in \{1, ..., 5\}$, que claramente es la hipótesis del criterio.

Con la finalidad de comprobar si el porcentaje de muestras de una $N(0,1)$ rechazadas es próximo al nivel de significación en cada caso para cada n seleccionada, ejecutamos el siguiente código en R:

Lo primero es definir los parámetros que entrarán en juego:

```
M <- 1000  # Número de simulaciones
n <- 5        # Tamaño de cada muestra (5, 10, 20, 30, 40 o 50)
alpha <- 0.01  # Nivel de significación (0.01, 0.05 o 0.1)
rechazos <- 0  # Contador de rechazos de H0
k <- 5         # Número de categorias
```

Una vez definidos los parámetros que se tendrán en cuenta, debemos calcular las categorías $A_1, ...A_k$ que utilizaremos para proseguir con el test, las cuales estarán basadas en cuantiles de la normal para facilitar así la simulación, ya que serán equiprobables :

```
breaks <- qnorm(seq(0, 1, length.out = k+1))  # 5 intervalos
```

Posteriormente, se realiza el test $ji-cuadrado$ para cada muestra y se comprueba si se acepta o no H_0:

```
set.seed(123)  # Semilla
for (i in 1:M) {
datos <- rnorm(n, mean = 0, sd = 1)  # Muestra de N(0,1)
observed_freq <- table(cut(datos, breaks = breaks))  # Frec. obs
```

```
# Frecuencias esperadas asumiendo N(0,1)
expected_freq <- rep(n / k, k)

# Test de Ji-cuadrado
test <- sum((observed_freq-expected_freq)^2/expected_freq)
  pvalor<-1-pchisq(test,df=k-1)
  if (pvalor < alpha) {
    rechazos <- rechazos + 1
    }
}
```

Vemos que el p-valor del test se obtiene empleando la distribución asíntótica del test estudiada en la **Sección 2.3**. Como para cada muestra de datos consideramos la hipótesis nula H_0 de que los datos pertenezcan a una $N(0,1)$ y un nivel de significación α, se rechazará H_0 en cada caso cuando el p-valor sea menor que α. Así, por cada muestra en la que se rechaza H_0 se suma una unidad a la variable *rechazos* (contabilizará todos los rechazos de H_0 tras las M simulaciones).

Por último, tan solo quedaría comprobar si en cada caso el test está bien calibrado, para lo cual debemos calcular el porcentaje de rechazos y ver si es próximo al nivel de significación empleado. El porcentaje de rechazos fue calculado como sigue:

```
# Proporción de rechazos de H0
prop_rechazos<- rechazos / M
```

El porcentaje obtenido es un porcentaje muestral sujeto a variabilidad aleatoria, por lo que debemos definir un intervalo para el cual se considera que el porcentaje de rechazos obtenido es próximo al nivel de significación. Podemos emplear un intervalo de confianza del 95 % para el nivel de significación utilizando una $Binomial(M, \alpha)$, siendo M el número de simulaciones y α el nivel de significación, el cual seguirá la siguiente expresión:

$$(\alpha - 1.96 \cdot \sigma, \alpha + 1.96 \cdot \sigma)$$

siendo $\sigma = \sqrt{\frac{\alpha \cdot (1-\alpha)}{M}}$ en este caso.

Una vez que tenemos calculado el porcentaje de rechazos, veamos cuál es la forma del intervalo de confianza de cada nivel de significación utilizando la expresión mencionada previamente:

Tabla 4.1: Intervalos de confianza al 95 % para cada α considerando M=1000 simulaciones y redondeando al cuarto decimal.

	Lim. Inf	**Lim. Sup**
α=**0.01**	0.0038	0.0162
α=**0.05**	0.0365	0.0635
α=**0.1**	0.0814	0.1186

Teniendo esto en cuenta y tras realizar los respectivos cálculos en R podemos obtener la Tabla 4.2, que muestra para cada α y para cada tamaño muestral n la proporción de rechazos:

Tabla 4.2: Proporciones de rechazo del test $ji-cuadrado$ a la hora de contrastar $H_0 : F_0 = N(0,1)$ para distintos niveles de significación α y tamaños muestrales n. Todas las proporciones han sido calculadas a partir de $M = 1000$ muestras de una $N(0,1)$. Se señalan en negrita las proporciones que están fuera del intervalo de confianza al 95 % para su nivel de significación (ver Tabla 4.1).

	n=**5**	n=**10**	n=**20**	n=**30**	n=**40**	n=**50**
α=**0.01**	**0.003**	0.007	0.011	0.015	0.009	0.014
α=**0.05**	**0.036**	0.041	0.043	0.046	0.051	0.044
α=**0.1**	0.084	0.094	0.094	0.088	0.104	0.092

Si nos fijamos en la anterior tabla (Tabla 4.2), vemos que los únicos casos en los que la proporción de rechazos cae fuera del intervalo de confianza al 95 % de α ocurren para tamaños muestrales $n = 5$, lo que no afecta al cumplimiento de la recomendación mencionada. Así, podemos decir que en esta simulación del test $ji - cuadrado$ para una $N(0,1)$ no tenemos pruebas en contra del cumplimiento del **Criterio 4.1**.

4.1.2. Caso discreto: *Poisson(2)*

Simularemos en R, de la misma forma que en el anterior caso, 1000 muestras de datos de distintos tamaños muestrales n (mismos valores que antes) con cada uno de los principales niveles de significación (0.01 , 0.05 y 0.1), con la hipótesis nula H_0 de que los datos pertenecen a una $Poisson(\lambda = 2)$. Para cada una de las 1000 muestras de tamaño n, se definen las k=5 categorías en las que se divide la muestra: $X = 0, X = 1, X = 2, X = 3$ y $X \geq 4$.

El código a usar es similar, aunque, como es entendible, tendrá ciertas variaciones. En primer lugar, como previamente, debemos definir los parámetros que serán considerados en nuestra simulación y también podemos calcular las frecuencias esperadas (serán las mismas para todas las muestras):

```
lambda <- 2        # Parámetro de la distribución Poisson(2)
n <- 5           # Tamaño muestral
M <- 1000          # Número de simulaciones
alpha <- 0.01      # Nivel de significación
rechazos <- 0      # Contador de rechazos
k<- 5              # Número de categorías
# Calcular las frecuencias esperadas según F_0=Poisson(2):
probs_teoricas <- c(dpois(0, lambda), dpois(1, lambda), dpois(2, lambda),
dpois(3, lambda), 1 - ppois(3, lambda))
frec_teoricas<-probs_teoricas*n
```

En este caso, con las probabilidades teóricas calculadas, podemos afirmar que el mínimo tamaño muestral n que asegura el cumplimiento del **Criterio 4.1** es $n = 40$.

Una vez definidos los parámetros de la simulación, podemos ejecutar el bucle que nos realizará la simulación de 1000 muestras mediante el test $ji - cuadrado$:

```
set.seed(123456)
for (i in 1:M) {
  muestra <- rpois(n, lambda = lambda)

  # Contar las frecuencias observadas en cada conjunto
  freq_0  <- sum(muestra == 0)
  freq_1  <- sum(muestra == 1)
  freq_2  <- sum(muestra == 2)
  freq_3  <- sum(muestra == 3)
  freq_4p <- sum(muestra >= 4)

  # Mostrar resultados
  frecuencias <-  c(freq_0, freq_1, freq_2, freq_3, freq_4p)

  # Test de Chi-cuadrado
  test <- sum((frecuencias-frec_teoricas)^2/frec_teoricas)
  pvalor<-1-pchisq(test,df=k-1)
  # Contar rechazos si p-valor < alpha
  if (pvalor < alpha) {
    rechazos <- rechazos + 1
  }      }
```

Si en una muestra concreta el p-valor obtenido considerando la distribución asintótica del test es menor que el nivel de significación α, se suma 1 al número de rechazos de H_0, que se acumula en la variable *rechazos*.

Lo siguiente sería, de la misma manera que en la anterior simulación, calcular la proporción de rechazos, dividiendo el número de rechazos (variable *rechazos*) entre el número de simulaciones ($M = 1000$). Cuando tenemos este valor, para tomar conclusiones simplemente debemos ver en qué casos este valor se aleja del nivel de significación (no está dentro del intervalo de confianza del 95 %), casos en los cuales diremos que el test no está bien calibrado.

El intervalo de confianza al 95 % para α tendrá la misma forma que en el anterior caso (ver Tabla 4.1). Después de simular en *R* se obtiene la Tabla 4.3, que muestra la proporción de rechazos para cada tamaño muestral n y para cada nivel de significación α:

Tabla 4.3: Proporciones de rechazo del test $ji - cuadrado$ a la hora de contrastar $H_0 : F_0 = Poisson(2)$ para distintos niveles de significación α y tamaños muestrales n. Todas las proporciones han sido calculadas a partir de $M = 1000$ muestras de una $Poisson(2)$. Se señalan en negrita las proporciones que están fuera del intervalo de confianza al 95 % para su nivel de significación (ver Tabla 4.1).

	n=**5**	n=**10**	n=**20**	n=**30**	n=**40**	n=**50**
α=**0.01**	**0.022**	0.014	0.011	0.005	0.011	0.011
α=**0.05**	0.058	0.05	0.037	0.045	0.047	0.048
α=**0.1**	0.11	0.092	0.092	0.109	0.096	0.114

De manera muy similar a la anterior simulación, podemos decir que tan solo obtenemos una proporción de rechazos alejada del nivel de significación α respectivo para el caso en el que el tamaño muestral n vale 5. Así, al igual que en el anterior caso, se puede afirmar que no tenemos pruebas en contra del cumplimiento de la recomendación (**Criterio 4.1**), que parece incluso algo conservadora, ya que a partir de cierto tamaño muestral n menor que 40 dejamos de obtener problemas con respecto a la proporción de rechazos.

Como conclusión a estas simulaciones se puede comentar que las simulaciones realizadas concuerdan con la recomendación del **Criterio 4.1**, por lo que las sospechas de que podía no tener suficiente fundamento no fueron respaldadas con esta simulación en *R*. Para haber obtenido algún problema con respecto a la recomendación deberíamos haber tenido alguna mala calibración del test $ji - cuadrado$ para tamaños muestrales $n = 30$ en adelante, ya que estos serían los candidatos a cumplir $np_j^0 > 5, \forall j \in \{1, ..., 5\}$, casos en los que, según el criterio, el test debería estar bien calibrado.

4.2. Potencia del test

En la anterior sección comprobamos que el test funcionaba bien bajo H_0, es decir, cuando la distribución F coincide con F_0. Basaremos ahora nuestro estudio en la potencia del test $ji - cuadrado$.

Para estudiarla, consideraremos el caso en el que $F \neq F_0$ como comentamos previamente, simularemos muchas muestras bajo la distribución F (1000 muestras como en la **Sección 4.1**), fijaremos un nivel de significación α ($\alpha = 0.05$) y $k = 5$ categorías como antes, probaremos para varios valores del tamaño muestral n (tomaremos los mismos valores para n que en **Sección 4.1**) y analizaremos la proporción de rechazos en cada caso para tomar conclusiones.

Primero de todo, probaremos con distribuciones F y F_0 pertenecientes a la misma familia, con un caso discreto y otro continuo (Poisson y Normal), para las que procederemos con simulaciones para distintos valores de los parámetros asociados a F.

4.2.1. Caso discreto: Poisson(2)

Analizaremos primero la potencia del test para $F_0 = Poisson(2)$. Consideraremos los siguientes valores de λ asociados a F de forma que $F = Poisson(\lambda)$: $\lambda = 1$, 1.5, 2.5 y 3.

El código de R usado tendrá la misma forma que el utilizado en la **Sección 4.1** para el caso de la $Poisson(2)$ (con la misma semilla), pero con las siguientes diferencias:

```
# Parámetros
lambda <- 1.5  # Parámetro de la distribución F (tomaremos 1, 1.5, 2.5, 3)
# 4. Calcular las frecuencias esperadas según Poisson(2)
probs_teoricas <- c(dpois(0, 2), dpois(1, 2), dpois(2, 2),
                    dpois(3, 2), 1 - ppois(3, 2))
frec_teoricas<-probs_teoricas*n
```

Como vemos, los cambios se basan en que en este caso el valor de λ asociado a F ($Poisson(\lambda)$) no será siempre el mismo (λ=2) y en que para el cálculo de las frecuencias bajo H_0 no utilizamos el respectivo valor de λ, sino que utilizamos directamente 2, ya que nuestra F_0 será siempre $Poisson(2)$.

Tras realizar las simulaciones para los distintos valores de λ asociados a F y para los distintos valores de n comentados ($n = 5$, $n = 10$, $n = 20$, $n = 30$, $n = 40$ y $n = 50$), obtenemos las proporciones de rechazos asociadas a cada caso que vemos a continuación:

Tabla 4.4: Proporciones de rechazo del test $ji-cuadrado$ a la hora de contrastar $H_0 : F_0 = Poisson(2)$ para $\alpha = 0.05$ fijado y para distintos tamaños muestrales n. Todas las proporciones han sido calculadas a partir de $M = 1000$ muestras de una distribución $F = Poisson(\lambda)$, con distintos valores de λ.

	λ=**1**	λ=**1.5**	λ=**2**	λ=**2.5**	λ=**3**
n=**5**	0.277	0.104	0.058	0.078	0.17
n=**10**	0.413	0.1	0.05	0.119	0.336
n=**20**	0.801	0.19	0.037	0.2	0.591
n=**30**	0.954	0.281	0.045	0.286	0.799
n=**40**	0.992	0.381	0.047	0.358	0.915
n=**50**	0.998	0.493	0.048	0.457	0.953

Como podemos observar en la Tabla 4.4 fijándonos en λ, la proporción de rechazos aumenta a medida que la distribución F ($Poisson(\lambda)$) difiere de la distribución $F_0 = Poisson(2)$. También podemos afirmar que la proporción de rechazos no es simétrica en la distancia de λ a 2, ya que es más fácil detectar $\lambda < 2$ que $\lambda > 2$.

Si nos centramos ahora en qué le ocurre a la proporción de rechazos en función de los distintos valores de n, podemos decir que, en general, la proporción aumenta a medida que aumenta n. La única ocasión donde no observamos un aumento de la potencia al aumentar n es para distribuciones próximas a la nula y tamaños muy bajos ($\lambda = 1.5$, $n = 5$ y $n = 10$), lo que se puede asociar a la variabilidad de la simulación.

Para el caso en el que λ=2 esta tendencia claramente no se cumple, ya que este caso se corresponde con la hipótesis nula H_0, donde se espera que la proporción de rechazos oscile alrededor del nivel de significación, en este caso 0.05.

Con todo esto, podemos decir que efectivamente en esta simulación el test detecta los casos en los que H_0 no es cierta, considerando una misma familia de distribuciones, la $Poisson$.

4.2.2. Caso continuo: *N(0,1)*

Veamos ahora qué ocurre para el caso de la distribución continua, donde consideraremos $F_0 = N(0, 1)$ y probaremos para F con distintos valores de μ y σ. Utilizaremos, como en el anterior caso, un código de R prácticamente idéntico al utilizado en la **Sección 4.1** para el caso de la $N(0, 1)$ (con la misma semilla), en el que consideraremos $\alpha = 0.05$ fijado, $k = 5$ intervalos y los mismos valores de n que en el caso anterior.

Consideraremos primero las siguientes distribuciones para F: $N(1/4, 5/4)$, $N(1/2, 3/2)$ y

$N(1, 2)$.

Las diferencias se basarán en el cálculo de las frecuencias observadas. Se definen, como en la **Sección 4.1**, los 5 intervalos equiprobables asociados a la $N(0, 1)$, y posteriormente se calculan las frecuencias observadas para cada muestra de $N(\mu, \sigma^2)$ a partir de estos intervalos, a diferencia de en la sección anterior, en la que las frecuencias observadas se obtenían a partir de muestras de datos pertenecientes a F_0.

Una vez simuladas $M = 1000$ muestras para cada F distinta y para cada valor de n, obtenemos las siguientes proporciones de rechazos:

Tabla 4.5: Proporciones de rechazo del test $ji-cuadrado$ a la hora de contrastar $H_0 : F_0 = N(0, 1)$ para $\alpha = 0.05$ y distintos tamaños muestrales n. Todas las proporciones han sido calculadas a partir de $M = 1000$ muestras de una distribución $F = N(\mu, \sigma^2)$, con distintos valores de μ y σ.

	$N(0,1)$	$N(1/4,5/4)$	$N(1/2,3/2)$	$N(1,2)$
n=**5**	0.036	0.037	0.085	0.232
n=**10**	0.041	0.072	0.169	0.495
n=**20**	0.043	0.125	0.36	0.825
n=**30**	0.046	0.153	0.466	0.944
n=**40**	0.051	0.211	0.624	0.982
n=**50**	0.044	0.252	0.73	0.999

En la Tabla 4.5 se observa claramente cómo, a medida que la distribución F se aleja de la hipótesis nula $H_0 : F = F_0 = N(0, 1)$ (nos movemos en cada fila de izquierda a derecha), la proporción de rechazos aumenta. También vemos que, sin considerar el caso de $N(0, 1)$ que coincide con la hipótesis nula y, por tanto, no lo consideramos en este estudio de potencia (nos centramos en $F \neq F_0$), la proporción de rechazos aumenta a medida que aumenta el tamaño muestral n.

Todo esto nos indica que para el caso que consideramos ($F_0 = N(0, 1)$ y misma familia de distribuciones para F), el test $ji - cuadrado$ detecta correctamente los casos en los que F difiere de F_0, cambiando tanto μ como σ.

Realizamos, a continuación, simulaciones similares a las realizadas con distintas distribuciones $N(\mu, \sigma^2)$ asociadas a F, pero ahora lo haremos cambiando los parámetros por separado, es decir, primero simularemos con distintos valores de μ fijando $\sigma = 1$ y luego fijaremos $\mu = 0$ y probaremos con distintos valores de σ. Con esto, podremos ver el efecto de cada uno de los parámetros en la potencia del test.

Simulemos muestras de $F = N(\mu, 1)$, con $\mu \in \{$-1, -0.5, 0.5, 1$\}$, tomando, como previamente, $F_0 = N(0, 1)$. Probaremos con los mismos valores para el tamaño muestral n, fijaremos tam-

bién el nivel de significación $\alpha = 0.05$ y $k = 5$ categorías correspondientes a cuantiles de la $N(0,1)$. El código de R será idéntico al anterior, salvo que en este caso las muestras provienen de $N(\mu,1)$, para los valores de μ comentados. Tras realizar las respectivas simulaciones, obtenemos la siguiente tabla que muestra las distintas proporciones de rechazos obtenidas en cada caso:

Tabla 4.6: Proporciones de rechazo del test $ji-cuadrado$ a la hora de contrastar $H_0 : F_0 = N(0,1)$ para $\alpha = 0.05$ y distintos tamaños muestrales n. Todas las proporciones han sido calculadas a partir de $M = 1000$ muestras de una distribución $F = N(\mu,1)$, con distintos valores de μ.

	N**(-1,1)**	N**(-0.5,1)**	N**(0.5,1)**	N**(1,1)**
n=**5**	0.053	0.07	0.085	0.056
n=**10**	0.389	0.05	0.059	0.397
n=**20**	0.828	0.145	0.153	0.828
n=**30**	0.979	0.317	0.27	0.971
n=**40**	0.997	0.457	0.432	0.996
n=**50**	0.999	0.601	0.576	1

Como se puede ver en la Tabla 4.6, las proporciones de rechazos son mayores cuando $|\mu| = 1$, lo que tiene sentido ya que en estos casos la distribución de F se aleja más de $F_0 = N(0,1)$. Cuando $|\mu| < 1$, las proporciones de rechazos se reducen principalmente para tamaños muestrales pequeños ($n \leq 10$), lo que también se respalda comprobando que en estos casos F se acerca más a F_0.

Con todo esto, podemos decir que, de forma general, el test *ji-cuadrado* es capaz de detectar cambios en μ cuando la hipótesis nula es $H_0 : F = N(0,1)$, sobre todo cuando estos cambios son considerables o cuando las muestras no son demasiado pequeñas.

Haremos ahora la misma simulación pero con $F = N(0,\sigma^2)$, es decir, fijando $\mu = 0$ y probando con los siguientes valores de σ: $\sigma \in \{0.75, 0.9, 1.1, 1.25\}$. Tras simular en R el mismo código pero adaptado a esta casuística, bajo los mismos tamaños muestrales n, el mismo nivel de significación $\alpha = 0.05$ y $k = 5$ categorías asociadas a cuantiles de $N(0,1)$, obtenemos las proporciones de rechazos que vemos en la Tabla 4.7.

Estas proporciones dejan claro que al test le cuesta mucho más detectar diferencias con respecto a F_0 cuando los cambios se realizan en σ, ya que si comparamos estos resultados con los obtenidos en la Tabla 4.6 para tamaños muestrales no demasiado pequeños, vemos que las proporciones de rechazos obtenidas en la Tabla 4.7 son, generalmente, mucho menores que las proporciones de la Tabla 4.6.

Tabla 4.7: Proporciones de rechazo del test $ji-cuadrado$ a la hora de contrastar $H_0 : F_0 = N(0,1)$ para $\alpha = 0.05$ y distintos tamaños muestrales n. Todas las proporciones han sido calculadas a partir de $M = 1000$ muestras de una distribución $F = N(0, \sigma^2)$, con distintos valores de σ.

	$N(\mathbf{0,0.75^2})$	$N(\mathbf{0,0.9^2})$	$N(\mathbf{0,1.1^2})$	$N(\mathbf{0,1.25^2})$
n=**5**	0.06	0.04	0.003	0.004
n=**10**	0.024	0.015	0.009	0.014
n=**20**	0.038	0.009	0.017	0.027
n=**30**	0.064	0.02	0.016	0.044
n=**40**	0.114	0.014	0.021	0.067
n=**50**	0.164	0.021	0.022	0.077

4.2.3. Caso continuo con distinta familia de distribuciones entre F y F_0

A continuación, seguiremos realizando simulaciones para analizar la potencia del test $ji-cuadrado$ pero, a diferencia de hasta ahora, tomaremos F de forma que F y F_0 no pertenecen a la misma familia de distribuciones.

Para esto, consideraremos $F_0 = N(0,1)$ y para F tomaremos una T de Student con distintos grados de libertad. Estas distribuciones tienen parecido ya que ambas son simétricas y, a medida que aumentan los grados de libertad de la T de Student, más se asemeja a una $N(0,1)$, como se puede ver en la Figura 4.1, por lo que deberemos comprobar que, al realizar las simulaciones, la proporción de rechazos debe disminuir a medida que aumenten los grados de libertad.

Simularemos $M = 1000$ muestras de datos pertenecientes a una T de Student con los siguientes grados de libertad: 2, 3 y 5. Tomaremos para el tamaño muestral n los mismos valores utilizados anteriormente, seguiremos fijando $\alpha = 0.05$ y considerando $k = 5$ intervalos equiprobables bajo H_0 (cuantiles de la $N(0,1)$).

El bucle que nos ayudará con la simulación en R tendrá la siguiente forma:

```
set.seed(123)
for (i in 1:M) {
  datos <- rt(n,df=df)     #Grados de libertad -> df= 2, 3 o 5
  # Frecuencias observadas
  observed_freq <- table(cut(datos, breaks = qnorm(seq(0, 1, length.out = 6))))
  # Frecuencias esperadas asumiendo N(0,1)
  expected_freq <- rep(n / 5, 5)
  # Test de Ji-cuadrado
  test <- sum((observed_freq-expected_freq)^2/expected_freq)
```

```
  pvalor<-1-pchisq(test,df=k-1)
  if (pvalor < alpha) {
    rechazos <- rechazos + 1
  }
}
```

Este código sólo se diferencia de los anteriores códigos para la $N(0,1)$ en que, en este caso, las muestras de datos provienen de una t_{df} (con $df \in \{2,3,5\}$) y, por tanto, las frecuencias observadas se obtienen analizando la cantidad de datos de cada una de estas muestras que caen en cada intervalo asociado a los respectivos cuantiles de la normal estándar.

Calculando, para cada respectivo valor de df y de n, la proporción de rechazos, obtenemos la siguiente tabla:

Tabla 4.8: Proporciones de rechazo del test $ji-cuadrado$ a la hora de contrastar $H_0 : F_0 = N(0,1)$ para $\alpha = 0.05$ y distintos tamaños muestrales n. Todas las proporciones han sido calculadas a partir de $M = 1000$ muestras de una distribución $F = t_{df}$, con distintos valores de df (grados de libertad).

	t_2	t_3	t_5
n=**5**	0.046	0.026	0.046
n=**10**	0.047	0.044	0.049
n=**20**	0.076	0.08	0.06
n=**30**	0.097	0.078	0.059
n=**40**	0.124	0.089	0.069
n=**50**	0.137	0.085	0.055

Analizando las proporciones rechazadas que se muestran en la Tabla 4.8, podemos decir primero que el test $ji - cuadrado$ detecta de mejor forma las diferencias entre F y F_0 cuando estas están más alejadas, es decir, cuando trabajamos con menos grados de libertad (t_2) y cuando el tamaño muestral es razonablemente grande ($n \geq 30$). Esto se observa viendo que en estos casos obtenemos los valores más elevados para la proporción de rechazos en la Tabla 4.8, comparados con los obtenidos para los otros 2 grados de libertad estudiados.

También podemos afirmar que para tamaños muestrales pequeños ($n \leq 10$), el test tiene más dificultades para detectar desviaciones de la hipótesis nula, ya que en estos casos la proporción de rechazos es realmente próxima (o incluso inferior) al nivel de significación α utilizado, lo que hará que la potencia del test sea baja.

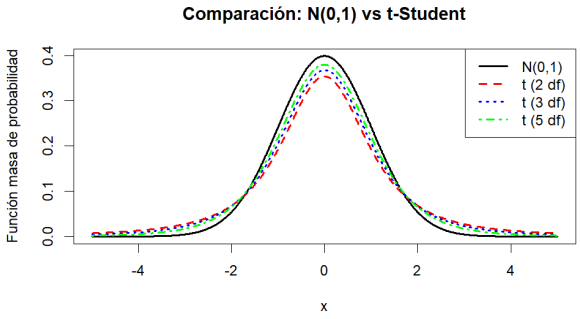

Figura 4.1: Comparación entre la $N(0,1)$ y la t de Student con 2, 3 y 5 grados de libertad.

4.2.4. Comparación con el test de *Kolmogorov-Smirnov*

Para seguir con nuestro estudio sobre la potencia del test $ji-cuadrado$, analizaremos a continuación, mediante simulación, la potencia del test de $Kolmogorv-Smirnov$ (definido en la Sección 3 de [4]), el cual únicamente se puede aplicar en los casos en los que F_0 es continua, y compararemos los resultados con los obtenidos previamente mediante simulación para el test $ji-cuadrado$.

Como hemos mencionado, el test $Kolmogorov-Smirnov$ solo se puede utilizar para distribuciones F_0 continuas, por lo que realizaremos simulaciones en R mediante este test de 2 de los casos estudiados previamente para la potencia del test $ji-cuadrado$ cuando F_0 es continua:

1. El caso en el que consideramos $F_0 = N(0,1)$ y $F = N(\mu, \sigma^2)$ con los respectivos valores empleados anteriormente para μ y σ.

2. El caso en el que tomamos también $F_0 = N(0,1)$, pero F pasa a ser una T de Student con los grados de libertad que ya fueron usados antes ($df \in \{2,3,5\}$).

La estructura de las simulaciones será la misma que hasta ahora:
Simular $M = 1000$ muestras de datos pertenecientes a cada distribución F, tomar $\alpha = 0.05$ y analizar las proporciones de rechazos para los distintos tamaños muestrales n considerados hasta ahora y para los distintos valores de los parámetros asociados a F utilizados en cada uno de los 2 casos estudiados mencionados. Usaremos la misma semilla que la utilizada para la simulación de estos casos mediante el test $ji-cuadrado$.

Para el primero de los 2 casos a estudiar a través de $Kolmogorov-Smirnov$, consideraremos $F_0 = N(0,1)$ y $F = N(\mu, \sigma^2)$, donde $F \in \{N(1/4, 5/4), N(1/2, 3/2), N(1,2)\}$, y utilizaremos el siguiente código de R:

```
for (i in 1:M) {
  x <- rnorm(n,mean=mu,sd=sqrt(sigma2))
  kstest<-ks.test(x, "pnorm", mean = 0, sd = 1)
  pvalor<-kstest$p.value
  if (pvalor < alpha) {
    rechazos <- rechazos + 1
  }
}
```

El bucle que vemos genera muestras de $N(\mu, \sigma^2)$, con los distintos valores de μ y σ comentados, realiza el test de $Kolmogorov - Smirnov$ considerando $F_0 = N(0,1)$ y suma 1 unidad a la variable *rechazos* cuando el p-valor obtenido sea menor que el nivel de significación fijado ($\alpha = 0.05$). La proporción de rechazos se calcula en cada caso, como hasta ahora, dividiendo el número de rechazos acumulado en la variable *rechazos* por el número de simulaciones $M = 1000$.

Tras ejecutarlo para cada respectivo caso, podemos representar en la siguiente tabla las distintas proporciones de rechazos obtenidas:

Tabla 4.9: Proporciones de rechazo del test $Kolmogorov - Smirnov$ a la hora de contrastar $H_0 : F_0 = N(0,1)$ para $\alpha = 0.05$ y distintos tamaños muestrales n. Todas las proporciones han sido calculadas a partir de $M = 1000$ muestras de una distribución $F = N(\mu, \sigma^2)$, con distintos valores de μ y σ.

	$N(\mathbf{0,1})$	$N(\mathbf{1/4,5/4})$	$N(\mathbf{1/2,3/2})$	$N(\mathbf{1,2})$
n=**5**	0.043	0.076	0.162	0.392
n=**10**	0.046	0.107	0.25	0.629
n=**20**	0.047	0.154	0.443	0.903
n=**30**	0.049	0.199	0.584	0.973
n=**40**	0.042	0.271	0.734	0.996
n=**50**	0.038	0.324	0.826	1

Podemos decir, fijándonos en la Tabla 4.9, que en esta simulación mediante el test de $Kolmogorov - Smirnov$ la proporción de rechazos aumenta a medida que aumenta el tamaño muestral n (sin tener en cuenta $N(0,1)$, que corresponde con H_0) y también a medida que la distribución F se aleja de F_0.

Para comparar ahora los resultados mediante el test $Kolmogorov-Smirnov$ con los obtenidos previamente a través del test $ji - cuadrado$, debemos comparar los valores de la Tabla 4.9 con los ya analizados de la Tabla 4.5.

Si nos fijamos en las proporciones obtenidas en ambas tablas para los casos en los que $F \neq F_0$ (columnas 2, 3 y 4), vemos que las correspondientes al test de $Kolmogorov - Smirnov$ son mayores o iguales que las obtenidas para el test $ji - cuadrado$ a cualquier tamaño muestral n. Esto nos indica, en esta simulación, que el test $Kolmogorov - Smirnov$ detecta de mejor forma que el $ji - cuadrado$ los casos en los que no se cumple la hipótesis nula y, por tanto, su potencia será más alta. Esto se debe en gran parte a que el test $Kolmogorov - Smirnov$, a diferencia del $ji - cuadrado$, no agrupa los datos.

El segundo caso que estudiaremos para analizar la potencia del test $Kolmogorov - Smirnov$ será, como comentamos, aquel en el que consideramos $F_0 = N(0,1)$ y $F = t_{df}$, con $df \in \{2,3,5\}$. El bucle de R que nos ayudará a obtener resultados será muy similar al empleado en el caso anterior. El cambio con respecto al anterior bucle se basa en la forma de obtener las respectivas muestras de datos, ya que en este caso cada una de ellas pertenece a una T de Student con df grados de libertad.

Calculando, de la misma forma que hasta ahora, la proporción de rechazos para cada caso, llegamos a obtener la siguiente tabla:

Tabla 4.10: Proporciones de rechazo del test $Kolmogorov - Smirnov$ a la hora de contrastar $H_0 : F_0 = N(0,1)$ para $\alpha = 0.05$ y distintos tamaños muestrales n. Todas las proporciones han sido calculadas a partir de $M = 1000$ muestras de una distribución $F = t_{df}$, con distintos valores de df (grados de libertad).

	t_2	t_3	t_5
n=**5**	0.09	0.054	0.057
n=**10**	0.068	0.074	0.062
n=**20**	0.092	0.071	0.061
n=**30**	0.102	0.089	0.063
n=**40**	0.116	0.086	0.068
n=**50**	0.139	0.089	0.052

Para comparar la potencia del test $ji - cuadrado$ con la potencia del test $Kolmogorov - Smirnov$ en este caso, deberemos comparar la Tabla 4.8 con la Tabla 4.10.

Comparando lo obtenido en ambas tablas, podemos afirmar que las proporciones de rechazos obtenidas mediante el test $Kolmogorov - Smirnov$ son generalmente mayores, salvo en algún caso concreto. En particular, para t_5 y para muestras grandes, el parecido entre F y F_0 reduce considerablemente la potencia, lo que puede beneficiar al test $ji - cuadrado$ debido a la aleatoriedad de la simulación en cuestión. Pese a esto, la tendencia global sigue siendo que para el test $Kolmogorov - Smirnov$ las proporciones de rechazo son mayores, aunque en menor proporción que en la familia normal.

Con todo esto, podemos afirmar que, generalmente, el test $Kolmogorov - Smirnov$ es más eficiente detectando diferencias entre F y F_0 cuando no se cumple la hipótesis nula, siempre que la distribución de los datos sea continua. Esto se debe en gran parte, como ya comentamos, a que no agrupa los datos, manteniendo así toda la información acerca de los mismos.

4.3. Simulación para el modelo paramétrico

En esta sección realizamos la simulación en R del test $ji - cuadrado$ para el caso en el que la hipótesis nula tiene la siguiente forma: $H_0 : F \in \{F_\theta : \theta \in \Theta\}$. Separaremos el estudio en los 3 casos vistos en teoría en el **Capítulo 3**.

Para este estudio, consideraremos en cada caso $H_0 : F \in \{N(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$, realizaremos $M = 1000$ simulaciones de muestras pertenecientes a una distribución $F = N(0, 1)$ con distintos tamaños muestrales n y calcularemos las respectivas proporciones de rechazos. Con esto, como claramente se cumplirá la hipótesis nula ya que $F = N(0, 1) \in \{N(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$, podremos analizar el calibrado del test para cada caso distinto. Consideraremos en las 3 simulaciones la siguiente semilla: $set.seed(123456)$.

4.3.1. Parámetro estimado con los datos ya agrupados

Nos centraremos primero en el caso en el que el parámetro θ se estima con los datos ya agrupados. En este caso, como ya mencionamos, se separa la muestra en k categorías $A_1, ..., A_k$, se estima el parámetro con los datos agrupados y, posteriormente, se calcula el estadístico del test.

Tomaremos $k = 5$ y las categorías $A_1, ...A_5$ serán los intervalos asociados a los respectivos cuantiles equiprobables de $N(0, 1)$. Estimaremos μ y σ mediante el estimador de máxima verosimilitud para datos agrupados visto en la **Sección 3.1** y realizaremos el test para los niveles de significación α habituales (0.01, 0.05 y 0.1) y tamaños muestrales $n \in \{20, 30, 40, 50\}$.

El código de R que nos permite obtener los estimadores de máxima verosimilitud con datos agrupados de μ y σ es el siguiente:

```
breaks <- qnorm(seq(0, 1, length.out = k+1))  # 5 intervalos basados en
cuantiles de la N(0,1)
# Función de log-verosimilitud con datos agrupados
logveros <- function(param) {
  mu <- param[1]
  sigma <- param[2]
```

```
  if (sigma <= 0) return(-Inf)
  probs_estimadas <- pnorm(breaks[-1], mu, sigma) - pnorm(breaks[-length(breaks)],
  mu, sigma)
  # Evitar log(0)
  probs_estimadas <- pmax(probs_estimadas, 1e-10)
  sum(observed_freq * log(probs_estimadas))
}
resultado <- optim(par = c(0, 1),
                   fn = function(par) -logveros(par),method="L-BFGS-B",
                   lower = c(-10, 1e-5), upper = c(10, 10))
mu_hat <- resultado$par[1];sigma_hat <- resultado$par[2]
```

, donde $observed\_freq$ se refiere al vector de frecuencias observadas de cada muestra de $N(0,1)$ de tamaño n y $probs\_estimadas$ al vector de probabilidades bajo la distribución $N(\mu, \sigma^2)$, con μ y σ arbitrarios, de cada intervalo ya definido .

La función de *R* *logveros* define la función log-verosimilitud, es decir, el logaritmo de la función de verosimilitud para datos agrupados. A continuación, se maximiza a través de *optim*, que por defecto calcula el mínimo (por eso el signo -), tomando como valores iniciales $\mu = 0$ y $\sigma = 1$.

Una vez tenemos calculados $\hat{\mu}$ y $\hat{\sigma}$, deberemos calcular el estadístico de contraste de $ji - cuadrado$ como se define en la **Sección 3.1** y su respectivo p-valor utilizando la distribución asintótica del estadístico dada en el **Teorema 3.3** (en este caso $\chi^2_{5-2-1=2}$).

Lo siguiente será, como en el resto de simulaciones realizadas, sumar una unidad a la variable *rechazos* (la cual empieza en 0) cuando se obtenga un p-valor menor que el nivel de significación α utilizado y dividir el número de rechazos recogidos en la variable *rechazos* por las $M = 1000$ simulaciones para obtener las respectivas proporciones de rechazos:

Tabla 4.11: Proporciones de rechazo del test $ji - cuadrado$ a la hora de contrastar $H_0 : F \in \{N(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$ para los habituales niveles de significación α y distintos tamaños muestrales n. Todas las proporciones han sido calculadas a partir de $M = 1000$ muestras de una distribución $F = N(0,1)$, estimando los parámetros con los datos ya agrupados. Se señalan en negrita las proporciones que están fuera del intervalo de confianza al 95 % para su nivel de significación (ver Tabla 4.1).

	n =**20**	n =**30**	n =**40**	n =**50**
α=**0.01**	**0.002**	0.009	0.004	0.007
α=**0.05**	**0.035**	0.043	0.05	0.045
α=**0.1**	**0.081**	0.095	0.094	0.089

Como vemos en la Tabla 4.11, las proporciones de rechazos son, de forma general, próximas al nivel de significación usado. La única excepción ocurre con $n = 20$, lo que se puede deber a la aleatoriedad de la simulación o a que no es un tamaño muestral suficientemente grande.

Así, podemos decir que para este caso concreto del modelo paramétrico el test $ji - cuadrado$ está bien calibrado bajo ciertas restricciones (tamaño muestral n no demasiado pequeño) en la situación estudiada.

4.3.2. Parámetro estimado con los datos sin agrupar

En este estudio de simulación estimaremos μ y σ^2 con los datos sin agrupar. Se tomarán a continuación k categorías $A_1, ...A_k$ que particionan el soporte y se calculará el estadístico como vimos en la **Sección 3.2**.

Tomaremos $k = 5$ y las categorías serán, como antes, los intervalos asociados a los respectivos cuantiles equiprobables de la $N(0, 1)$. Tomaremos como estimadores de μ y σ^2 los estimadores de máxima verosimilitud. Siguiendo el procedimiento explicado en la **Sección 3.2** para el cálculo del estimador de máxima verosimilitud, obtenemos $\hat{\mu} = \bar{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$ y $\hat{\sigma^2} = S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X_n})^2$. Realizaremos la simulación para los niveles de significación $\alpha = 0.01$, 0.05 y 0.1 y tamaños muestrales $n \in \{20, 30, 40, 50\}$.

Tras calcular para los distintos valores de n y α el estadístico de contraste $D(\hat{\theta})$ y sus respectivos p-valores tomando la distribución asintótica dada en el **Teorema 3.3**, que no depende de θ_0, se obtiene el número de rechazos totales de H_0, que se recoge en la variable *rechazos*. Dividiendo este valor por el número de simulaciones realizadas ($M = 1000$), obtenemos las respectivas proporciones de rechazos:

Tabla 4.12: Proporciones de rechazo del test $ji - cuadrado$ a la hora de contrastar $H_0 : F \in \{N(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$ para los habituales niveles de significación α y distintos tamaños muestrales n. Todas las proporciones han sido calculadas a partir de $M = 1000$ muestras de una distribución $F = N(0, 1)$ y los parámetros han sido estimados con los datos sin agrupar. Se señalan en negrita las proporciones que están fuera del intervalo de confianza al 95 % para su nivel de significación (ver Tabla 4.1).

	n =**20**	n =**30**	n =**40**	n =**50**
α=**0.01**	**0.055**	**0.053**	**0.036**	**0.039**
α=**0.05**	**0.159**	**0.159**	**0.143**	**0.134**
α=**0.1**	**0.249**	**0.25**	**0.251**	**0.223**

La Tabla 4.12 deja claro que en este subcaso del modelo paramétrico el test $ji - cuadrado$

no está bien calibrado, ya que las proporciones de rechazos obtenidas no son próximas a los respectivos niveles de significación α en ningún caso.

Esto concuerda con lo que estudiamos en teoría en la **Sección 3.2**, en la que vimos, mediante el **Teorema 3.5**, que no era posible calibrar el test.

4.3.3. Categorías a partir de los datos

Esta simulación se diferenciará de las 2 anteriores en la forma de escoger las categorías. En este caso, a diferencia de en los anteriores, las categorías dependerán de los datos, mediante el procedimiento explicado en la **Sección 3.3**.

Tomaremos también $k = 5$ y fijaremos inicialmente las probabilidades $p_i = 1/5, i = 1, ..,5$. Los intervalos $(a_{i-1}(\theta_0), a_i(\theta_0)]$ estarán asociados a los respectivos cuantiles de la $N(0,1)$ de forma que $p_i(\theta_0) = 1/5, \forall i = 1, ..., 5$ ($\theta_0 = (0,1)'$ en este caso).

Se estimarán μ y σ^2, de la misma forma que en la anterior simulación, mediante $\hat{\mu} = \bar{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$ y $\hat{\sigma^2} = S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X_n})^2$, de forma que $\hat{\theta} = (\hat{\mu}, \hat{\sigma^2})'$. Las categorías pasarán entonces a ser intervalos de la forma $(a_{i-1}(\hat{\theta}), a_i(\hat{\theta})]$. Se calcularán después las respectivas probabilidades $p_i(\hat{\theta})$ (probabilidad bajo $N(\bar{X_n}, S^2)$ de que una observación pertenezca al intervalo $(a_{i-1}(\hat{\theta}), a_i(\hat{\theta})]$).

El código de R a usar será el siguiente:

```
breaks <- qnorm(seq(0, 1, length.out = k+1))
for (i in 1:M) {
  datos <- rnorm(n, mean = 0, sd = 1)
  mu_hat<-mean(datos);sigma_hat<-sum((datos-mean(datos))^2)/n
  breaks_hat<-breaks*sigma_hat+mu_hat
  probs_hat <- pnorm(breaks_hat[-1], mu_hat, sigma_hat) - pnorm(
  breaks_hat[-length(breaks)], mu_hat, sigma_hat)
  esp <- n * probs_hat
  observed_freq<-table(cut(datos,breaks=breaks_hat))
  test<-sum((observed_freq-esp)^2/esp)
  pvalor<-1-pchisq(test,df=k-2-1)

  if (pvalor < alpha) {
    rechazos <- rechazos + 1
  }
}
```

La variable *breaks_hat* calcula los respectivos $a_i(\hat{\theta})$, haciendo que los intervalos pasen a depender de la estimación de los parámetros a partir de los datos. Mediante *probs_hat* se obtienen los $p_i(\hat{\theta})$, probabilidad de cada intervalo bajo $N(\bar{X}_n, S^2)$.

A continuación, se calculan las frecuencias observadas en los respectivos intervalos transformados, se calcula el estadístico $D_R(\hat{\theta})$ y el p-valor asociado en cada caso mediante la distribución asintótica estudiada en el **Teorema 3.3**.

Con todo esto, tan solo faltaría dividir el número de rechazos de H_0 obtenidos en la simulación, casos en los que el p-valor es menor que el nivel de significación α, por el número de simulaciones ($M = 1000$) para obtener las respectivas proporciones de rechazos que vemos a continuación:

Tabla 4.13: Proporciones de rechazo del test $ji - cuadrado$ a la hora de contrastar $H_0 : F \in \{N(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$ para los habituales niveles de significación α y distintos tamaños muestrales n. Todas las proporciones han sido calculadas a partir de $M = 1000$ muestras de una distribución $F = N(0, 1)$ y las categorías a partir de los datos. Se señalan en negrita las proporciones que están fuera del intervalo de confianza al 95 % para su nivel de significación (ver Tabla 4.1).

	n =**20**	n =**30**	n =**40**	n =**50**
α=**0.01**	0.014	**0.019**	**0.018**	**0.028**
α=**0.05**	**0.113**	**0.1**	**0.13**	**0.122**
α=**0.1**	**0.189**	**0.186**	**0.221**	**0.207**

Los resultados obtenidos en la Tabla 4.13 muestran que claramente el test $ji - cuadrado$ no está bien calibrado en esta situación concreta del modelo paramétrico, ya que de forma general se puede decir que las proporciones de rechazos no son próximas a los respectivos niveles de significación α, tomando como referencia los intervalos vistos en la Tabla 4.1.

Estas afirmaciones se respaldan con las conclusiones generales del **Teorema 3.6**, que aseguran que en este caso el test $ji - cuadrado$ no se puede calibrar, salvo para algunas excepciones.

Como conclusión a esta sección, podemos decir que únicamente obtuvimos una buena calibración del test $ji - cuadrado$ para el primero de los 3 casos del modelo paramétrico, lo que tiene sentido ya que, como vimos en teoría en el **Capítulo 3**, en los otros 2 casos la distribución asintótica del estadístico de contraste depende del parámetro θ_0, lo que hace que de forma general no se puedan calibrar.

Bibliografía

[1] Airy, G. B. (1861). *On the algebraical and numerical theory of errors of observations and the combination of observations.* Macmillan.

[2] Chernoff, H., Lehmann, E. L. (1954). The use of maximum likelihood estimates in χ^2 tests for goodness of fit. *The Annals of Mathematical Statistics.* Vol. **25**, No. 3, pp. 579-586.

[3] Cochran, W. G. (1952). The χ^2 test of goodnes of fit. *The Annals of Mathematical Statistics.* Vol. **23**, No. 3, pp. 315-345.

[4] Conde Amboage, M., Rodríguez Casal, A. (2024). *Inferencia Estadística.* Tema 1. Inferencia no paramétrica no paramétrica basada en la distribución empírica. Grado en Matemáticas, Universidad de Santiago de Compostela.

[5] Conde Amboage, M., Rodríguez Casal, A. (2024). *Inferencia Estadística.* Tema 2. Contrastes ji-cuadrado. Grado en Matemáticas, Universidad de Santiago de Compostela.

[6] Cramer, H. (1946). *Mathematical methods of Statistics.* Princeton University Press.

[7] Fisher, R. A. (1924). The condition under which chi square measures the discrepancy between observation and hypothesis. *Journal of the Royal Statistical Society.* Vol. **87**, No. 3, pp. 442-450.

[8] Freijeiro González, L., González Manteiga, W., Sánchez Sellero, C. (2020). *Probabilidad y Estadística.* Tema 9. Ley de los grandes números y teorema central del límite. Grado en Matemáticas, Universidad de Santiago de Compostela.

[9] García Pérez, A. , Vélez Ibarrola, R. (1997). *Principios de inferencia estadística.* Universidad Nacional de Educación a Distancia (España).

[10] Golub, G. H., Van Loan, C. F. (2013). *Matrix Computations* ($4^{\underline{a}}$ edición). Johns Hopkins University.

[11] Greenwood, P. E., Nikulin, M. S. (1996). *A guide to Chi-Squared Testing.* Wiley Series in Probability and Statistics.

[12] Merriman, M. (1884). *A Text-Book on the Method of Least Squares*. John Wiley & Sons.

[13] Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine, Series 5*. Vol. **50**, No. 302, pp. 157–175.