



Universidade de Santiago de Compostela

FACULTAD DE MATEMÁTICAS

EL TEOREMA CENTRAL DEL LÍMITE

Autor: Adrián Gándara Díez

2021-2022

Índice general

Resumen.	II
Introducción.	III
1. Desarrollo histórico.	1
1.1. Un cambio de mentalidad: Jakob Bernoulli.	2
1.2. Contraste de significación del ratio entre nacimientos de hombres y mujeres.	4
1.3. Teorema de de Moivre-Laplace.	5
1.4. Contribuciones posteriores.	8
2. Herramientas matemáticas.	11
2.1. Fundamentos de la Teoría de la Probabilidad.	11
2.2. Distribuciones y valor esperado.	14
2.3. Convergencia en distribución.	21
3. Demostración del Teorema Central del Límite.	25
3.1. Funciones características.	25
3.2. Teorema de Continuidad.	27
3.3. Teorema Central del Límite.	31
4. Aplicaciones del Teorema Central del Límite.	33
4.1. Distribución empírica.	33
4.2. Intervalo de confianza de una proporción.	36

Resumen.

En este trabajo de fin de grado se estudiará en profundidad el Teorema Central del Límite para el caso de la distribución binomial. Desde un punto de vista histórico, se detallará su evolución desde la búsqueda de una buena aproximación de la binomial, dando origen a la curva normal, hasta los desarrollos que dan lugar a su versión clásica. Seguidamente, se presentarán una serie de herramientas matemáticas y resultados con el propósito de probar formalmente el teorema. Por último, se expondrán dos ejemplos de aplicaciones que tiene este resultado en la inferencia estadística.

Abstract.

In this work of the end of degree, we will deeply study the binomial distribution case of the Central Limit Theorem. From an historical point of view, we will show the evolution from the research of an accurate binomial approximation, which it has supposed the first appearance of the normal curve. Then, we will explain the mathematical tools and results which formally proves the theorem. Finally, we will expose two examples of its applications in the statistical inference.

Introducción.

El Teorema Central del Límite (TCL) es uno de los resultados más importantes de la historia de la matemática y, especialmente, de la estadística. Nace de la necesidad de aproximar la suma de variables aleatorias con el fin de simplificar el cálculo de probabilidades. Aunque se llame Teorema Central del Límite, realmente es una colección de resultados bajo distintas condiciones, cuya versión clásica fue formulada por Lindeberg, Feller y Lévy:

Teorema 0.1. (*Teorema de Lindeberg-Feller-Lévy*) Sean $X_1, X_2, \dots$ una sucesión de variables aleatorias independientes e idénticamente distribuidas con media μ y desviación típica σ , entonces:

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} N(0, 1),$$

donde $S_n = X_1 + \dots + X_n$ es la suma de las variables y $N(0, 1)$ denota una distribución normal de media 0 y varianza 1.¹ (véase Billingsley, 1995, Sección 27.)

Este teorema nos asegura que la distribución de S_n es aproximadamente normal², siempre que n sea lo suficientemente grande.

Sin embargo, este resultado es fruto del trabajo de numerosos matemáticos desde comienzo del siglo XVIII. El objetivo central de este trabajo será la versión dada por Abraham de Moivre en 1733 y conocida con el nombre de *Teorema de de Moivre-Laplace*. Se trata de un caso particular del anterior teorema para una sucesión de variables aleatorias independientes con distribución dicotómica o de Bernoulli.

Esta distribución está definida en experimentos aleatorios cuyos posibles resultados son éxito o fracaso, tomando los valores 1 y 0 respectivamente. La suma de n variables Bernoulli independientes con probabilidad p , denotadas como $Ber(p)$, da como resultado una variable con distribución binomial de parámetros n y p , $Bin(n, p)$, la cual designa la probabilidad de

¹Nótese que en esta expresión $n\mu$ es la media de S_n y $\sigma\sqrt{n}$ su desviación típica.

²La distribución normal es una poderosa herramienta matemática que tiene una importancia mayúscula en la estadística. Denotada como $N(\mu, \sigma^2)$, donde los parámetros $\mu \in \mathbb{R}$ y $\sigma^2 \in \mathbb{R}^+$ indican la media y varianza, respectivamente. Su función de densidad, para $x \in \mathbb{R}$ es

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Destacar el caso particular de la normal estándar $N(0, 1)$, presente en el TCL, donde la media y la varianza valen 0 y 1, respectivamente.

obtener k éxitos en n intentos ($0 \leq k \leq n$) independientes con probabilidad p :

$$f(k) = \mathbb{P}(X = k) = \binom{n}{k} p^k q^{n-k}, \quad (1)$$

donde $q = 1 - p$ y $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ (véase Billingsley, 1995, Sección 20).

La búsqueda de una buena aproximación de la expresión (1) para grandes valores de n da como resultado la primera versión del TCL, que coincide además con la primera aparición de la curva normal. En notación moderna, se expresa como sigue:

Teorema 0.2. (*Teorema de de Moivre-Laplace*) Sea $X_n \sim \text{Bin}(n, p)$ una variable aleatoria con distribución binomial, se cumple que su media es np y su varianza npq , con $q=1-p$. Entonces:

$$\frac{X_n - np}{\sqrt{npq}} \xrightarrow{d} N(0, 1).$$

A lo largo de este trabajo, trataremos cuatro aspectos claves de este teorema: su historia, las herramientas matemáticas presentes en la demostración, la propia demostración y sus aplicaciones en la estadística.

En el Capítulo 1 abordaremos cuestiones históricas del TCL: desde los primeros estudios de la binomial realizados por Jakob Bernoulli, que concluyen con su “ley de los grandes números”; pasando por la aplicación de este resultado al primer contraste de significación llevado a cabo por su sobrino Nicolas I Bernoulli; hasta las aproximaciones que realizan de Moivre y Laplace de la binomial y que dan origen al TCL. Además, se verá la influencia que tuvo este resultado en matemáticos posteriores que estudiarían condiciones más generales del teorema, concluyendo en el *Teorema de Lindeberg-Feller-Lévy* visto al principio de esta introducción.

En el Capítulo 2 se expondrán las herramientas matemáticas presentes en el TCL y se asentarán las bases de la Teoría de la Probabilidad. definiendo formalmente conceptos fundamentales de ésta como son el espacio de probabilidad, la variable aleatoria, la independencia y la distribución. Seguidamente, se verá la relación existente entre ésta última y el valor esperado de una variable. Ya en la parte final, se mencionarán los distintos tipos de convergencia para una sucesión de variables aleatorias, centrándonos en la convergencia en distribución presente en el TCL.

El Capítulo 3 es el relativo a la demostración del TCL, para la cual la función característica jugará un papel fundamental. En las primeras dos secciones se definirá esta idea y se probarán una serie de resultados que justifican su uso en la demostración del teorema. En la Sección 3.3 se procederá a la demostración rigurosa del *Teorema de de Moivre-Laplace*.

En el Capítulo 4 se mencionarán un par de aplicaciones que tiene el teorema en la inferencia estadística: la distribución empírica, que es la estimación de una función de distribución desconocida a partir de una muestra aleatoria; y el cálculo de un intervalo de confianza para una proporción.

Capítulo 1

Desarrollo histórico.

El Teorema Central del Límite debe su nombre a Georg Pólka, quien en 1920 denominó la versión dada por Laplace en 1810 como el “teorema central del límite de la teoría de la probabilidad”. Hoy en día, este término se refiere a la versión clásica del teorema, relativa a la convergencia en distribución de la suma de variables aleatorias independientes e idénticamente distribuidas a la distribución normal. Aunque lleve este nombre, el TCL no es solamente un teorema, es una colección de resultados sobre la convergencia de funciones de distribución, de densidad y de probabilidades discretas de la suma de variables aleatorias independientes.



Figura 1.1: De izquierda a derecha, los retratos de Jakob Bernoulli (Bernoulli¹, 1687), de Moivre (Desconocido, 1736) y Laplace (Guérin, 1838) // Wikimedia Commons.

A lo largo de este trabajo nos centraremos en un caso particular del TCL, que además es el primero y el que inicia una andadura de casi 200 años de desarrollo matemático: el *Teorema de de Moivre-Laplace* (véase Teorema 0.2).

¹El autor del retrato es el pintor Nicolas Bernoulli (1662-1716), hermano de Jakob y padre de Nicolas I Bernoulli.

La binomial cobra especial relevancia en la segunda mitad del siglo XVII por su alta implicación en los juegos de azar que dan origen a la Teoría de la Probabilidad. A pesar de que numerosos matemáticos se centran en su estudio, es Jakob Bernoulli (Figura 1.1, izquierda) el precursor de la búsqueda de buenas aproximaciones para grandes valores de n . Además, en su obra póstuma, *Ars Conjectandi*, demuestra el primer teorema límite de la Teoría de la Probabilidad, la hoy conocida como “ley débil de los grandes números”. Tras él, su sobrino Nicolas I Bernoulli, consigue mejorar los resultados de Jakob y aplicarlos a la práctica con su estudio del ratio entre nacimientos de hombres y mujeres.

Claramente influenciado por los Bernoulli, Abraham de Moivre (Figura 1., centro) demuestra que la distribución de una binomial se aproxima a la entonces desconocida distribución normal, en lo que supone la primera aparición de la curva normal. Tres cuartos de siglo más tarde y tras la derivación de la función de densidad de la normal por parte de Gauss, Pierre Simon Laplace (Figura 1.1, derecha) mejora la aproximación dada por de Moivre y extiende su estudio a la suma de variables aleatorias independientes.

1.1. Un cambio de mentalidad: Jakob Bernoulli.

Se considera que el origen de la Teoría de la Probabilidad se remonta al año 1654, cuando Pierre de Fermat (1601-1665) y Blaise Pascal (1623-1662) comienzan a mantener correspondencia con el fin de resolver una serie de problemas relacionados con los juegos de azar (véase Hald, 2003, Capítulo 5). Aunque muchos le atribuyen este mérito a Gerolamo Cardano (1501-1576), lo cierto es que su obra *Liber de Ludo Aleae* (Libro de los Juegos de Azar), escrita durante la década de 1560, vio la luz póstumamente en 1663.

Durante la siguiente mitad de siglo, muchos matemáticos comienzan a publicar trabajos relacionados con los juegos de azar y las apuestas, pero todos ellos tienen la misma base: sabiendo las probabilidades en relación a un experimento aleatorio, calcular la probabilidad de un determinado suceso. Es decir, el proceso de razonamiento iba desde las causas hasta los efectos.

Desde un punto de vista filosófico, en este siglo se produce un cambio en el pensamiento que influiría enormemente en la Teoría de la Probabilidad. El racionalismo que intentaba deducir el comportamiento de los fenómenos naturales a partir de supuestas leyes que constituían la materia era demasiado limitado por la complejidad de la naturaleza y la propia limitación de los sentidos. Esto obligó a emplear un esquema inverso de razonamiento empirista: a partir de los efectos observables, poder deducir las causas ocultas que los provocaban (véase Landro y González, 2013). De esta manera, los matemáticos intentaron mejorar los métodos de predicción a partir de las probabilidades de las hipótesis, ampliando los alcances de la teoría probabilística a la inferencia empírica, y en particular, al contraste de hipótesis.

Uno de los primeros matemáticos en llevar a cabo este cambio de mentalidad fue Jakob (James o Jaques) Bernoulli (1654-1705). Perteneciente a una de las familias más influyentes en las matemáticas, Jakob realizó importantes contribuciones a la teoría de las ecuaciones diferenciales, el cálculo de variaciones, la teoría de series infinitas y, en el caso que nos ocupa, la Teoría de la Probabilidad.

En este último campo, además de las aportaciones que veremos a continuación, da nombre a una de las distribuciones discretas más famosas: la distribución dicotómica o de Bernoulli. Tal y como se ha visto en la introducción de este trabajo, si la probabilidad de éxito es p , la suma de n variables Bernoulli independientes da como resultado una binomial de parámetros n y p , que denotaremos mediante $\text{Bin}(n, p)$.

Precisamente, el considerado por él mismo como su mejor trabajo es el referente a la distribución binomial, perteneciente a la obra *Ars Conjectandi* (El Arte de la Conjetura, 1713). Editado y publicado ocho años después de su muerte por su sobrino Nicolas I, debido a disputas dentro de la familia Bernoulli, supone un avance revolucionario para la Teoría de la Probabilidad.

En las tres primeras partes de su obra, mejora y generaliza los trabajos realizados por Pascal y Huygens en relación a los juegos de azar (véase Hald, 2003, Capítulos 15 y 16). Bernoulli insiste en que la probabilidad de ganar en un determinado juego es constante, es decir, es independiente de los resultados obtenidos en juegos previos. Esto le motiva a estudiar el comportamiento de las probabilidades en un juego con n intentos en el que cada uno tiene la misma probabilidad p de éxito, observando que la probabilidad de obtener k éxitos en un determinado orden será $p^k q^{n-k}$. En caso de que el orden sea indiferente, habría que multiplicar el anterior término por el número de combinaciones posibles $\binom{n}{k}$. Ya en la última parte de su obra, a la que dedicó sus últimos dos años de vida, discute la convergencia de las frecuencias relativas distribuidas binomialmente hasta demostrar su famoso teorema.

En lugar del antiguo concepto de azar basado en la simetría de los posibles resultados, introduce un nuevo concepto de probabilidad, definida como “un grado de certeza”. Esto es un paso revolucionario porque la probabilidad en este sentido se relaciona con proposiciones y no directamente con eventos. Él trata de dar reglas para la medición y combinación de tales probabilidades, aunque no triunfa fuera de los campos ya conocidos de los juegos de azar y frecuencias relativas estables.

Considerando que la frecuencia relativa de un evento estará más cerca de la verdad cuanto mayor sea el número de observaciones, prueba la primera versión de la “ley de grandes números”. En notación moderna, es como sigue:

Teorema 1.1. (Ley débil de los grandes números) *Sean n lanzamientos independientes con probabilidad de éxito p , X_n el número de éxitos distribuidos de forma binomial y $p_n = X_n/n$ la frecuencia relativa, dado $d > 0$ entero se cumple:*

$$\lim_{n \rightarrow \infty} \mathbb{P}(|p_n - p| \leq d) = 1.$$

Este es el primer teorema límite en la Teoría de la Probabilidad y hoy en día se conoce como “ley débil de los grandes números”. Su importancia en la estadística es mayúscula, pues da una justificación teórica al uso de p_n como estimador de la probabilidad p , ya que garantiza que las frecuencias relativas empíricas convergen a un número fijo que se corresponde con la probabilidad del evento, p . Sin embargo, no fue capaz de indicar cómo usar dicho teorema para encontrar un intervalo de confianza para p sabiendo el valor de p_n . Este pudo ser uno de los motivos por los que no publicó este trabajo en vida.

A partir de este resultado y considerando el término np entero, Jakob intenta evaluar la suma

$$P_d = \mathbb{P}(|X_n - np| \leq d) = \sum_{\substack{x \in \{0, \dots, n\} \\ |x - np| \leq d}} \binom{n}{x} p^x q^{n-x} \quad (1.1)$$

pero únicamente consigue una cota inferior bajo algunas restricciones. Dicho resultado fue bastante insatisfactorio, pues además era necesario que el número de observaciones n fuese demasiado grande (véase Hald, 2003, Sección 16.2). Sin embargo, sería el primero en estudiar las sumas binomiales, sentando unas bases que serían clave en los posteriores trabajos de su sobrino Nicolas I y, sobre todo, de de Moivre.

1.2. Contraste de significación del ratio entre nacimientos de hombres y mujeres.

Durante el año 1712, John Arbuthnott (1667-1735), profesor de matemáticas en Londres, publica un artículo en el que estudia el sexo de las personas nacidas en Londres entre los años 1629 y 1710. Para ello, toma los datos de bautizos durante esos años y observa que el número de hombres excedía al de mujeres en todos y cada uno de esos 82 años. Sin comprobar el total de nacimientos de hombres y mujeres ni el ratio de estos valores y comparando dicho evento con el lanzamiento de una moneda, estudia que la probabilidad de que esto suceda durante 82 años es $(1/2)^{82}$. Dándole un sentido teológico a dicho resultado, atribuye este fenómeno a la intervención divina para explicar que nazcan más hombres que mujeres.

A pesar de la simpleza del trabajo de Arbuthnott, sus conclusiones influirían notablemente en los teólogos y matemáticos del momento. Entre los muchos matemáticos que discutieron los trabajos del profesor inglés, destacan el científico holandés Willem J. 'sGravesande (1688-1742) y Nicolas I Bernoulli (1687-1759), quienes tras conocerse comenzaron a discutir en una intensa correspondencia el trabajo de Arbuthnott. Como Arbuthnott solo había usado el hecho de que durante esos 82 años el número de hombres era mayor que el de mujeres, 'sGravesande lo lleva más lejos estudiando la frecuencia relativa de nacimientos de hombres. Observó que ésta variaba entre 0.5027 en 1703 y 0.5362 en 1661. Calculó el número total de hombres multiplicando estas frecuencias relativas por la media de nacimientos esos 82 años, 11429. De esta manera, el valor mínimo de nacimientos sería 5745 y el máximo 6128. Entonces consideró los datos de las 82 observaciones como una binomial de parámetro $n = 11429$ y todas las observaciones contenidas en el intervalo $[5745, 6128]$. Para obtener la probabilidad de este evento, el matemático holandés hizo uso de la recursión

$$\binom{n}{x+1} = \binom{n}{x} \frac{n-x}{x+1}$$

propuesta por Newton para propósitos similares, consiguiendo la siguiente aproximación de la suma de los 384 términos de dicho intervalo²:

$$\mathbb{P}(5745 \leq X \leq 6128) = \sum_{k=5745}^{6128} \binom{11429}{k} \left(\frac{1}{2}\right)^{11429} \approx 0,292. \quad (1.2)$$

²El procedimiento completo se puede ver en la Sección 17.2 de Hald (2003).

Bajo esta hipótesis, que esto ocurra durante los 82 años, se traduce en una probabilidad total de $0,292^{82}$, muy por debajo de lo calculado por Arbuthnott. El trabajo de 'sGravesande, aunque tedioso, da un buen resultado, pues con las herramientas computacionales actuales se sabe que la suma (1.2) es igual a 0.287, mientras que su aproximación normal es 0.284.

Nicolas I Bernoulli estaba en desacuerdo con algunas afirmaciones de carácter matemático de Arbuthnott que contradecían los conocimientos que tenía sobre probabilidad y, en especial, sobre la distribución binomial, cuya fuente principal era la obra de su tío Jakob Bernoulli. Nicolas I buscó obtener una fórmula general para el cálculo de la binomial para evitar el duro trabajo computacional que fue necesario para el cálculo de 'sGravesande. Él no estaba interesado en la visión compartida por Arbuthnott y 'sGravesande de que este fenómeno era causado por una deidad y su actitud fue más pragmática: buscó estimar la probabilidad de nacimiento de los hombres a partir de las observaciones, comparar la distribución de las observaciones con la de la distribución binomial y encontrar una herramienta matemática que le ayudase a encontrar una aproximación para la binomial para valores grandes de n .

Nicolas I no sólo consiguió mejorar teóricamente la ley de los grandes números dada por su tío, sino que además la aplicó al caso de los datos recogidos por Arbuthnott. Nicolas I explicó la regularidad del ratio de nacimientos a través de un modelo binomial, un modelo aleatorio alejado de la intervención divina propugnada por Arbuthnott y 'sGravesande. Los razonamientos efectuados nos retrotraen al inicio de la técnica matemática de los contrastes de significación que siglos después Fisher y Neyman consolidarían. Su razonamiento se basaba en que, dadas las observaciones, lo razonable es cuestionar la hipótesis $p = 1/2$.

Para dar un rango para el valor posible de p , sería deseable mejorar el cálculo probabilístico que aparece en (1.1). Sin embargo, Nicolas I tampoco consiguió una buena aproximación, limitándose a mejorar la cota dada por Jakob. De hecho, si hubiese intentado llevar su aproximación un paso más allá, habría conseguido de manera relativamente sencilla demostrar el mismo resultado que el presentado por de Moivre en 1733, aunque bajo otro método de estudio.

1.3. Teorema de de Moivre-Laplace.

Abraham de Moivre (1667-1754) nace en una familia protestante en Vitri, Francia. Tras el Edicto de Nantes de 1685 que provocó la persecución que sufrieron los protestantes, de Moivre se vio obligado a huir a Inglaterra, donde se desarrolló como matemático compartiendo trabajo con científicos de la talla de Newton. Dedicó la primera parte de su vida al estudio de las ecuaciones algebraicas y las series infinitas consiguiendo demostrar la ecuación epónima $(\cos(x) + i \operatorname{sen}(x))^n = \cos(nx) + i \operatorname{sen}(nx)$ previa a la demostración de la famosa fórmula de Euler $e^{ix} = \cos(x) + i \operatorname{sen}(x)$ que usaremos en próximos capítulos.

Sus estudios sobre la Teoría de la Probabilidad no comienzan hasta 1708, a la edad de 41 años, pero la importancia que adquieren sus trabajos recogidos en *De Mensura Sortis* en 1712 y sobre todo *Doctrine of Chances* con sus tres ediciones de 1718, 1738 y 1756 lo convierten en el mejor probabilista del momento. Sus trabajos probabilísticos se basaron puramente en la deducción teórica, pues nunca analizó observaciones para demostrar la aplicación que

pudiesen tener sus teorías.

En su época, muchos matemáticos buscaron una aproximación simple de la distribución binomial, pues la solución de muchos problemas relacionados con los juegos de azar conducían a fórmulas que contenían estos valores, cuyo cálculo se volvía demasiado engorroso cuando n tomaba grandes valores.

De Moivre comienza sus investigaciones en 1721 con la distribución binomial simétrica ($p = 1/2, q = 1/2$). El enfoque del matemático francés difiere en el de los Bernoulli, pues busca obtener una aproximación del coeficiente $\binom{n}{x}p^xq^{n-x}$ en lugar de hacerlo con el término P_d de (1.1) directamente. En 1730 consigue una primera aproximación del máximo término M de la suma, correspondiente al término central de la binomial, gracias a la ayuda de James Stirling (1692-1770) y su fórmula homónima³, obteniendo la expresión en notación moderna

$$M = \frac{\binom{n}{n/2}}{2^n} \approx \frac{2}{\sqrt{2\pi n}}.$$

Esto no es otra cosa que la probabilidad de que la binomial X_n tome el valor $n/2$ bajo la condición de que n sea un entero par.

Tras esto, de Moivre consiguió obtener una aproximación del cociente M/Q , siendo Q un término situado a una distancia d de M , observando que “*si $\frac{n}{2}$ es una cantidad infinitamente grande, el logaritmo de la razón entre un término que dista del central una distancia d y el término central es $-2d^2/n$* ”⁴. Traducido a lenguaje matemático, esto es

$$\ln \frac{Q}{M} = \ln \left(\frac{\mathbb{P}(X_n = n/2 + d)}{\mathbb{P}(X_n = n/2)} \right) \approx -\frac{2d^2}{n},$$

siendo n un valor grande en relación a d .

Por lo tanto, eliminando logaritmos se tiene que la probabilidad de obtener el valor $n/2 + d$ en una binomial simétrica X_n es

$$\mathbb{P}(X_n = n/2 + d) \approx \frac{2}{\sqrt{2\pi n}} e^{-2d^2/n},$$

lo que le permitió concluir que

$$P_d = \mathbb{P}(|X_n - n/2| \leq d) \approx \frac{2}{\sqrt{2\pi n}} \int_{-d}^d e^{-\frac{2s^2}{n}} ds. \quad (1.3)$$

Esto constituye la primera aparición de la curva normal⁵. El término de la derecha en (1.3) es lo que hoy conocemos como distribución normal de media 0 y varianza $n/4$.

³La “Fórmula de Stirling” se puede escribir de la siguiente manera: $n! \approx \sqrt{2\pi n} \left(\frac{n}{e}\right)^n$

⁴Esta afirmación forma parte del Corolario I de *Aproximatio*, tal y como se muestra en Stigler (1986), Capítulo 2, página 75.

⁵Recordamos que la función de densidad de una distribución normal estándar es $\phi(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$ y su función de distribución $\Phi(x) = \int_{-\infty}^x \phi(t)dt$.

Estos resultados los publica ya en 1733, veinte años después de los trabajos de los Bernoulli, en un artículo llamado *Approximatio ad Summam Terminorum Binomii $(a + b)^n$ in Series Expansi*, incluido en la segunda edición de su *Doctrine of Chances*, donde no solo aporta una aproximación de la binomial simétrica, sino que consigue extenderlo a cualquier valor de p que cumple que np sea entero:

$$P_d = \mathbb{P}(|X - np| \leq d) \approx \frac{1}{\sqrt{2\pi npq}} \int_{-d}^d e^{-\frac{s^2}{2npq}} ds, \quad (1.4)$$

resultando la curva normal de media 0 y varianza npq (las demostraciones de estos resultados se pueden ver en Hald, 2003, Capítulo 24).

Aunque de Moivre no le dio más importancia a la distribución normal que la de aproximar la binomial, sí que pensó en ella como una curva, ya que en palabras del propio de Moivre, “*si se piensa en los términos de la binomial sobre una línea recta en posición vertical igualmente espaciados y formando ángulos rectos con la recta, los extremos de los términos siguen una curva.*” (véase Stigler, 1986, Capítulo 2). Esto se puede comprobar en la Figura 1.2, donde se puede aproximar la binomial de parámetros $n = 25$ y $p = 0,2$ por la normal de media $\mu = np = 5$ y varianza $\sigma^2 = npq = 4$.

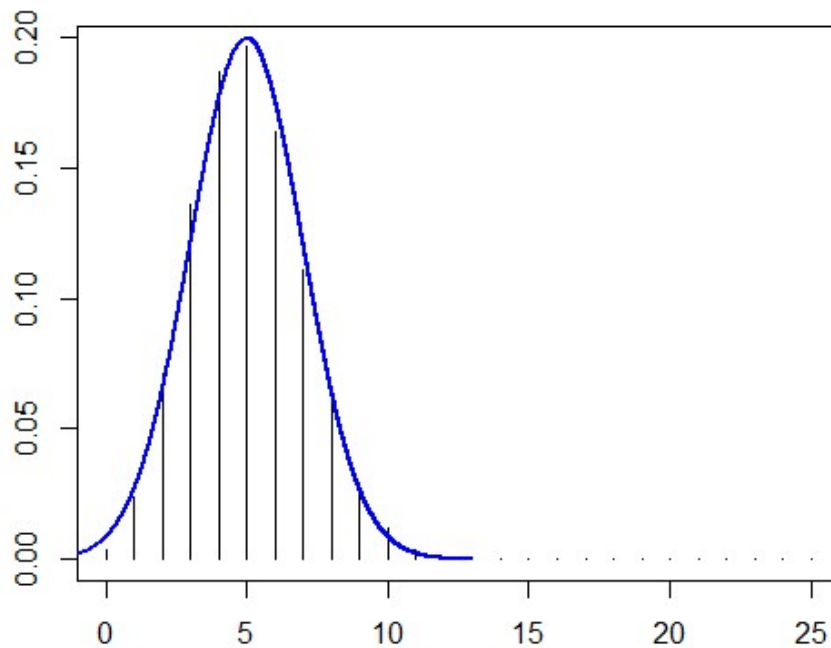


Figura 1.2: Aproximación de una distribución binomial $\text{Bin}(25, 0,2)$ por una normal $N(5, 4)$. En negro, el valor de las probabilidades de obtener x aciertos en la $\text{Bin}(25, 0,2)$; en azul, la función de densidad de $N(5, 4)$.

Históricamente, la distribución normal surge del estudio que realiza Carl Friedrich Gauss (1777-1855)⁶ en 1809 sobre los errores en la medición de las distancias de la Tierra a los

⁶Es por ello que la curva normal se conoce también como “curva gaussiana” o “campana de Gauss”.

cuerpos celestes. Gauss consigue derivar la función de densidad normal usando el método de máxima verosimilitud en una familia de localización y realiza las estimaciones suponiendo que los errores de medición siguen esta distribución (véase Stahl, 2006).

Un año después, llega el trabajo de Gauss a manos de Pierre Simon Laplace (1749-1847), reconocido matemático, físico y astrónomo francés, quien llevaba estudiando desde 1776 el comportamiento de la suma de variables aleatorias independientes. En su trabajo, suponía que los ángulos de inclinación de los meteoros con respecto a la eclíptica estaban uniformemente distribuidos entre 0° y 90° y buscó determinar las probabilidades de la suma de estos ángulos. En sus obras, Laplace no presenta un teorema general que se corresponda con el TCL general, sino que trata problemas particulares relativos a la aproximación de las probabilidades de la suma o combinaciones lineales de un gran número de variables aleatorias. En notación moderna, la versión más general del TCL de Laplace se puede ver en la Sección 2.1. de Fischer (2010):

Teorema 1.2. (*Teorema Central del Límite, Laplace(1810)*). Sean $\varepsilon_1, \dots, \varepsilon_n$ un gran número de errores independientes de observación, cada uno con la misma densidad de media μ y varianza σ^2 . Si $\lambda_1, \dots, \lambda_n$ son constantes multiplicadoras y $a > 0$, entonces:

$$\mathbb{P} \left(\left| \sum_{i=1}^n \lambda_i (\varepsilon_i - \mu) \right| \leq a \sqrt{\sum_{i=1}^n \lambda_i^2} \right) \approx \frac{2}{\sigma \sqrt{2\pi}} \int_0^a e^{-\frac{x^2}{2\sigma^2}} dx. \quad (1.5)$$

Laplace consigue mejorar el trabajo de Gauss, pues justifica la elección que hizo de su curva de errores.

Ya en 1812, Laplace mejora la aproximación dada por de Moivre, aportando un término de corrección a la ecuación (1.4), motivo por el cual el Teorema 0.2. visto en la introducción lleva también su nombre. Basa su estudio en la suposición de que no sea necesario que el valor np sea entero, sino que toma como valor central el valor $m = \lfloor (n+1)p \rfloor = np + z$, con $-q < z \leq p$, pues es el que tiene mayor probabilidad en la binomial. Con razonamientos similares a los de de Moivre, y haciendo uso de las fórmulas de Stirling y Euler-Maclaurin con sus propias demostraciones, establece un término de corrección de la aproximación dada por su predecesor, dando como resultado

$$\mathbb{P}(|X_n - m| \leq d) = \frac{1}{\sqrt{2\pi}} \int_{-t}^t e^{-y^2/2} dy + O(1/\sqrt{n}), \quad (1.6)$$

siendo $t = d/\sqrt{npq}$ (véase Hald, 2003, Sección 24.6).

1.4. Contribuciones posteriores.

Tras Laplace, fueron numerosos los matemáticos que realizaron sus aportaciones al TCL: Poisson, Dirichlet, Cauchy, Chebyshev o Markov realizaron mejoras durante el Siglo XIX, pero sin ser del todo rigurosos; y ya en el Siglo XX, Lyapunov, von Mises, Pólya, Lindeberg, Lévy y Feller continúan con el trabajo de los anteriores, demostrando, esta vez sí, de manera

rigurosa sus versiones. Durante las siguientes líneas se hará un breve resumen de los avances más importantes desde Laplace hasta culminar en el *Teorema de Lindeberg-Feller-Lévy*, siendo recomendable la lectura de Fischer (2010) para una mejor comprensión de los mismos.

De todos los matemáticos que contribuyeron al TCL de Laplace durante el Siglo XIX, ninguno tuvo tanta importancia como Poisson, que en sus artículos de 1824 y 1829 discute el teorema. Aplicando un análisis matemático más exacto, consigue una aportación doble: generaliza la prueba a la suma de variables aleatorias independientes con distribuciones diferentes y discute la validez del TCL mencionando algunos contraejemplos como la distribución de Cauchy, lo que motivaría a buscar condiciones necesarias y suficientes del teorema.

A finales de este siglo, miembros de la escuela de San Petersburgo como Chebyshev y Markov aplican el método de los momentos para sus demostraciones en las que consideran que las variables sean únicamente independientes, estableciendo condiciones explícitas para la validación del teorema, extendiéndolo a cantidades y no solo estudiando los errores como sus predecesores. El uso de la función característica por parte de Lyapunov, discípulo de Chebyshev, da como resultado la primera demostración rigurosa del TCL bajo su “condición de Lyapunov” en 1901.

Durante el Siglo XX y tras la demostración de Lyapunov, el reto consistiría en encontrar condiciones necesarias y suficientes para el teorema: primero Lindeberg, quien en 1922 demuestra de manera rigurosa las condiciones suficientes del TCL; y ya en 1935, Feller y Lévy presentan dos trabajos similares en los que aportan también condiciones necesarias. Todo ello resulta en el *Teorema de Lindeberg-Feller-Lévy* correspondiente al Teorema 0.1 de este trabajo.

Capítulo 2

Herramientas matemáticas.

Como hemos visto en el anterior capítulo, la Teoría de la Probabilidad surge de la necesidad de entender qué hay detrás del azar y de los procesos aleatorios. Para ello, se asientan unas bases matemáticas sólidas que permiten su estudio pormenorizado. En este capítulo, detallaremos los conceptos fundamentales de la Teoría de la Probabilidad y los resultados que posteriormente nos permitirán demostrar el TCL.

2.1. Fundamentos de la Teoría de la Probabilidad.

El primero de los conceptos fundamentales a tratar es el de espacio de probabilidad. Está definido como una terna $(\Omega, \mathcal{F}, \mathbb{P})$ formada por el espacio muestral Ω , la σ -álgebra $\mathcal{F}$ y la probabilidad $\mathbb{P}$.

El espacio muestral Ω es un conjunto no vacío que contiene todos los posibles resultados de un experimento aleatorio.

La σ -álgebra $\mathcal{F}$ es una familia de subconjuntos de Ω , denominados sucesos. Por tratarse de una σ -álgebra, se tiene que $\mathcal{F}$ es cerrada bajo la formación del complementario y la unión e intersecciones no vacías numerables. Esto es, dado un conjunto $A \in \mathcal{F}$, su complementario A^c también pertenecerá a $\mathcal{F}$; y dada una colección numerable $\{A_i, i = 1, 2, \dots\} \in \mathcal{F}$, tanto su unión $\bigcup_{i=1}^{\infty} A_i$ como su intersección $\bigcap_{i=1}^{\infty} A_i$ pertenecerán a $\mathcal{F}$.

Es obvio que $\mathcal{F}$ contendrá siempre a Ω y al conjunto vacío $\emptyset$, pues $\Omega = A \cup A^c$ y $\emptyset = A \cap A^c$ para cualquier conjunto $A \in \mathcal{F}$, que existe por suponer $\mathcal{F}$ no vacío.

No es necesario que $\mathcal{F}$ incluya todos los subconjuntos de Ω (en muchos casos ni siquiera es deseable). De hecho, se puede definir la σ -álgebra generada por una familia de subconjuntos de Ω , $\mathcal{S}$, como la menor σ -álgebra que contiene a $\mathcal{S}$, denotándose $\sigma(\mathcal{S})$. Particularmente importante es la σ -álgebra de Borel $\mathcal{B}$ de un espacio topológico, que es la σ -álgebra generada por los conjuntos abiertos. A los elementos de $\mathcal{B}$ se les denomina conjuntos de Borel o borelianos.

La probabilidad $\mathbb{P}$ es una aplicación que asigna un valor en $[0,1]$ a cada suceso de $\mathcal{F}$ tal que $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(\Omega) = 1$ y

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i), \quad (2.1)$$

si $\{A_i, i = 1, 2, \dots\} \in \mathcal{F}$, con $A_i \cap A_j = \emptyset$ ($i \neq j$).

Como consecuencia de la anterior definición, se cumplen las siguientes propiedades de $\mathbb{P}$, cuyas demostraciones son inmediatas y se pueden ver en las páginas 24 y 25 de Billingsley:

P1. Si $A, B \in \mathcal{F}$ y $A \subset B$, entonces $\mathbb{P}(A) \leq \mathbb{P}(B)$.

P2. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A) \quad \forall A \in \mathcal{F}$.

P3. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) \quad \forall A, B \in \mathcal{F}$.

P4. $\mathbb{P}\left(\bigcup_{i=1}^n A_i\right) \leq \sum_{i=1}^n \mathbb{P}(A_i)$.

P5. (Continuidad desde abajo) Si $\{A_n\} \subset \mathcal{F}$, $A \in \mathcal{F}$ y $\{A_n\} \uparrow A$ (es decir, $A_1 \subset A_2 \subset \dots$ y $A = \bigcup_{i=1}^{\infty} A_i$), entonces:

$$\mathbb{P}(A_n) \uparrow \mathbb{P}(A).$$

P6. (Continuidad desde arriba) Si $\{A_n\} \subset \mathcal{F}$, $A \in \mathcal{F}$ y $\{A_n\} \downarrow A$ (es decir, $A_1 \supset A_2 \supset \dots$ y $A = \bigcap_{i=1}^{\infty} A_i$), entonces:

$$\mathbb{P}(A_n) \downarrow \mathbb{P}(A).$$

Así definido, el espacio de probabilidad no es más que el instrumento que modela la realización del sorteo en un experimento aleatorio. Pero esta primera etapa aleatoria no tiene sentido si no llevamos a cabo la medición de los resultados que se obtienen. Es en esta segunda etapa, denominada determinista, cuando entra en juego el concepto de variable aleatoria: una función que asigna un valor real a cada elemento del espacio muestral.

Formalmente, dado un espacio de probabilidad $(\Omega, \mathcal{F}, \mathbb{P})$ una variable aleatoria X es una función medible¹ de Ω en $\mathbb{R}$ tal que para todo B conjunto de Borel,

$$X^{-1}(B) = \{\omega \in \Omega, X(\omega) \in B\} \in \mathcal{F}.$$
²

Por lo tanto, $\mathbb{P}$ asignará una probabilidad a cada suceso $\{X \in B\}$. A la función $\mathbb{P}_X$ que a cada conjunto de Borel B de la recta real le asigna el valor $\mathbb{P}_X(B) = \mathbb{P}(X \in B)$ se le denomina distribución de X , concepto en el que se ahondará en la Sección 2.2 de este capítulo.

Se expondrán un par de ejemplos sencillos a fin de facilitar la comprensión de lo definido hasta ahora:

¹Se dirá que una función $f : \Omega \rightarrow \Omega'$ es $\mathcal{F}/\mathcal{F}'$ -medible si para todo $A' \in \mathcal{F}'$ se cumple

$$f^{-1}(A') = \{\omega \in \Omega : f(\omega) \in A'\} \in \mathcal{F}.$$

En el caso de una variable aleatoria, $\mathcal{F}'$ es la σ -álgebra de Borel del espacio $\mathbb{R}$ dotado de la topología euclídea.

²Habitualmente, se denotará el conjunto $X^{-1}(B)$ como $\{X \in B\}$.

Ejemplo 2.1: Supongamos que queremos observar la cantidad de coches defectuosos al fabricar tres. El espacio muestral en este experimento será

$$\Omega = \{NNN, NND, NDN, NDD, DNN, DND, DDN, DDD\}$$

siendo N “no defectuoso” y D “defectuoso”. Considerando $\mathcal{F}$ como la σ -álgebra generada por los conjuntos $\{NNN\}$, $\{DNN, NDN, NND\}$, $\{DDN, DND, NDD\}$ y $\{DDD\}$, la variable aleatoria X = “número de coches defectuosos” asignará los valores 0, 1, 2 y 3 a los elementos de cada conjunto, respectivamente. En este caso, dado que X toma un número finito de valores, diremos que X es una variable discreta.

Ejemplo 2.2: Si quisiéramos medir con exactitud el tiempo que tarda un alumno en resolver un determinado problema, el espacio muestral será $\mathbb{R}^+$. La variable aleatoria Y = “tiempo” será la función identidad y obtendrá un número infinito no numerable de posibles valores, por lo que hablamos de variable aleatoria continua. Aquí la σ -álgebra natural es la generada por los conjuntos abiertos de $\mathbb{R}$.

Existe una cualidad fundamental asociada a los conceptos de suceso y de variable aleatoria y que es clave en la Teoría de la Probabilidad: la independencia.

Dos sucesos A y B son independientes si $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. De forma general, una colección de sucesos $\{A_\alpha\}_{\alpha \in \mathcal{I}}$ se dice que son independientes si para cada $j \in \mathbb{N}$ y para cada elección finita $\alpha_1, \dots, \alpha_j \in \mathcal{I}$ se tiene:

$$\mathbb{P}(A_{\alpha_1} \cap \dots \cap A_{\alpha_j}) = \mathbb{P}(A_{\alpha_1}) \dots \mathbb{P}(A_{\alpha_j}).$$

Conceptualmente, las variables aleatorias X e Y serán independientes si los resultados de una no influyen sobre la otra. Formalmente, se cumplirá si para todo par de conjuntos borelianos, B_1 y B_2 , se cumple que $X^{-1}(B_1)$ y $Y^{-1}(B_2)$ son sucesos independientes. Es decir, $\mathbb{P}(X \in B_1, Y \in B_2) = \mathbb{P}(X \in B_1)\mathbb{P}(Y \in B_2)$. De forma análoga, una familia de variables aleatorias $\{X_\alpha\}_{\alpha \in \mathcal{I}}$ será independiente si para cada $j \in \mathbb{N}$, para cada elección finita $\alpha_1, \dots, \alpha_j \in \mathcal{I}$ y cualquier $B_1, \dots, B_j$ conjuntos de Borel se tiene:

$$\mathbb{P}(X_{\alpha_1} \in B_1, \dots, X_{\alpha_j} \in B_j) = \mathbb{P}(X_{\alpha_1} \in B_1) \dots \mathbb{P}(X_{\alpha_j} \in B_j).$$

Siguiendo con el caso de la fabricación de coches expuesto en el Ejemplo 2.1, si suponemos que la fabricación de cada uno de los coches es independiente, calcular la probabilidad de que al fabricar tres unidades el resultado sea NDN , sucederá con una probabilidad $\mathbb{P}(NDN) = \mathbb{P}(N)\mathbb{P}(D)\mathbb{P}(N)$.

La siguiente proposición será útil para demostrar ciertos resultados que veremos en la próxima sección:

Proposición 2.1. Sean X e Y variables aleatorias independientes y $f, g : \mathbb{R} \rightarrow \mathbb{R}$ funciones medibles con la σ -álgebra de Borel, entonces $f(X)$ y $g(Y)$ son también independientes.

Demostración: El objetivo es probar que para todo B_1, B_2 conjuntos borelianos, $\mathbb{P}((f \circ X)^{-1}(B_1), (g \circ Y)^{-1}(B_2)) = \mathbb{P}((f \circ X)^{-1}(B_1))\mathbb{P}((g \circ Y)^{-1}(B_2))$.

Consideremos el espacio de probabilidad $(\Omega, \mathcal{F}, \mathbb{P})$ de X e Y . Usando la definición que vimos en la Página 12 de función medible, sabemos que dados $B_1, B_2 \in \mathcal{B}$, $f^{-1}(B_1)$ y

$g^{-1}(B_2)$ también serán conjuntos de Borel. Ahora bien, por la independencia de X e Y , se cumplirá que $\mathbb{P}(X^{-1}(f^{-1}(B_1)), Y^{-1}(g^{-1}(B_2))) = \mathbb{P}(X^{-1}(f^{-1}(B_1)))\mathbb{P}(Y^{-1}(g^{-1}(B_2)))$, que es exactamente lo que queremos demostrar. $\square$

2.2. Distribuciones y valor esperado.

Una vez asentadas las bases de la Teoría de la Probabilidad, es momento de profundizar sobre los conceptos de distribución y función de distribución para centrarnos en las que de verdad nos importan en este trabajo: la distribución binomial y la normal.

Como se ha mencionado anteriormente, la distribución de una variable aleatoria X es una medida de probabilidad $\mathbb{P}_X$ ³, es decir, una función que asigna a cada posible suceso una probabilidad de que ocurra. Formalmente, sea $(\Omega, \mathcal{F}, \mathbb{P})$ un espacio de probabilidad, a cada conjunto de Borel B , $\mathbb{P}_X$ le asigna un valor entre 0 y 1:

$$\mathbb{P}_X(B) = \mathbb{P}(X \in B) = \mathbb{P}(\omega \in \Omega : X(\omega) \in B).$$

Ligado a este concepto está el de función de distribución. Se define como la probabilidad de que la variable obtenga un valor menor o igual a x :

$$F_X(x) = \mathbb{P}_X((-\infty, x]) = \mathbb{P}(X \leq x) \quad \forall x \in \mathbb{R}.$$

Esta función cumple las propiedades de continuidad por la derecha, es monótonamente creciente y los límites cuando x tiende a $-\infty$ y $+\infty$ son 0 y 1, respectivamente (véase Billingsley, 1995, Sección 14).

Que la función F_X lleve el nombre de función de distribución no es fruto de la casualidad o de la mala fe de los matemáticos, sino porque caracteriza por completo la distribución de la variable X , cumpliéndose lo siguiente:

Proposición 2.2. *Si X e Y son dos variables aleatorias, entonces $\mathbb{P}_X = \mathbb{P}_Y$ si y solo si $F_X(x) = F_Y(x)$ para todo $x \in \mathbb{R}$.*

Este resultado es un caso particular de un teorema fundamental de la Teoría de la Medida, denominado *Teorema de Extensión*. Éste dice que cualquier medida de probabilidad definida en un álgebra⁴ $\mathcal{C}$ puede ser extendida, de manera única, a la σ -álgebra generada por $\mathcal{C}$. Las demostraciones de ambos resultados, así como los conceptos básicos de la Teoría de la Medida, se pueden ver en el Capítulo 2 de Billingsley (1995).

Por lo anterior, a lo largo de este trabajo se hablará indistintamente de distribución o de función de distribución.

La distribución discreta por excelencia y que tiene una enorme importancia, incluso a nivel histórico como ya se vio en el Capítulo 1, es la binomial. Ésta designa la probabilidad

³Si X tiene la distribución $\mathbb{P}_X$, lo denotaremos $X \sim \mathbb{P}_X$.

⁴Un álgebra es una clase no vacía $\mathcal{C}$ de subconjuntos de Ω si es cerrada para la formación de complementarios y uniones e intersecciones finitas.

de obtener k éxitos en n intentos ($0 \leq k \leq n$) independientes con probabilidad p fija, y lo denotamos por $Bin(n, p)$. Considerando $X_n \sim Bin(n, p)$ como la suma de n variables Bernoulli independientes con probabilidad p de éxito, y mediante el uso de combinatoria, se llega a la siguiente expresión, con la que podemos calcular la probabilidad de que X sea igual a k :

$$f(k) = \mathbb{P}(X = k) = \binom{n}{k} p^k q^{n-k}, \quad (2.2)$$

donde $k \in \{0, 1, \dots, n\}$, $q = 1 - p$ y $\binom{n}{k} = \frac{n!}{k!(n-k)!}$.

Por otro lado tenemos la distribución normal, una herramienta fundamental de la estadística pues permite modelar numerosos fenómenos naturales y sociales. Denotaremos $N(\mu, \sigma^2)$ a una variable aleatoria normal con media $\mu \in \mathbb{R}$ y varianza σ^2 . Su función de densidad⁵ para todo $x \in \mathbb{R}$ es

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}.$$

En el caso particular de que μ y σ^2 tomen los valores 0 y 1, respectivamente, estaremos ante una normal estándar, cuya función de densidad en este caso es

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

En la Figura 2.1 se aprecia la gráfica de esta función, también conocida por el nombre de “campana de Gauss”.

Es muy sencillo convertir una normal no estandarizada en una que sí lo es y viceversa. Si X es una variable estándar, $X \sim N(0, 1)$, la variable $Y = \mu + \sigma X$ seguirá una distribución $N(\mu, \sigma^2)$.

Para el caso de la normal $N(\mu, \sigma^2)$, se obtiene su función distribución a partir de la función de densidad, obteniéndose la siguiente expresión

$$F(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^x e^{-(t-\mu)^2/2\sigma^2} dt.$$

El caso particular de la normal estándar tiene la función de distribución

$$\Phi(x) = \int_{-\infty}^x e^{-t^2/2} dt.$$

A continuación veremos que las distribuciones y las funciones de distribución son muy útiles a la hora de calcular, por ejemplo, la media y la varianza de una variable aleatoria.

⁵La función de densidad de una variable aleatoria continua X se define como

$$f(x) = \lim_{h \rightarrow 0} \frac{\mathbb{P}(x-h \leq X \leq x+h)}{2h}.$$

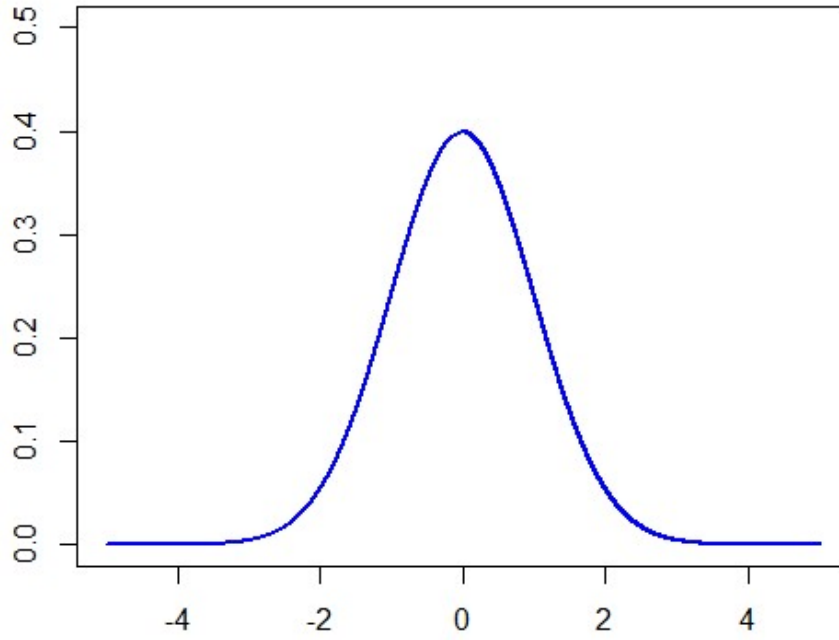


Figura 2.1: Función de densidad de una normal estándar $N(0, 1)$.

A pesar de que estos conceptos han sido mencionados desde el principio de este trabajo, aún no han sido definidos de manera formal. Para ello, jugará un papel fundamental el valor esperado o esperanza, que no es otra cosa que el valor medio teórico, es decir, el valor al que se aproximará la media de los valores obtenidos al realizar un número suficiente de experimentos aleatorios independientes⁶. Veremos su definición formal distinguiendo tres tipos de variables en función de los posibles valores que pueden tomar: comenzaremos con las variables simples o discretas, aquellas cuyo número de posibles valores es finito; seguiremos con las continuas no negativas, que tienen infinitos posibles resultados y todos son no negativos; y finalizaremos con las continuas generales, con infinitos posibles resultados reales.

Suponiendo que X es una variable simple que toma los valores $x_1, \dots, x_k$ y siendo $A_1, \dots, A_k$ los subconjuntos de Ω definidos como $A_i = \{\omega \in \Omega : X(\omega) = x_i\}$, podemos escribir $X = \sum_{i=1}^k x_i \mathbb{I}_{A_i}$, donde $\mathbb{I}$ es la función indicatriz⁷. De esta manera, definimos el valor esperado de X ,

$$\mathbb{E}(X) = \sum_{i=1}^k x_i \mathbb{P}(A_i).$$

⁶Esto es posible afirmarlo gracias a la versión general de la ley débil de los grandes números vista en el Capítulo 1 de este trabajo.

⁷La función indicatriz $\mathbb{I}_A$ sobre un conjunto A se define como

$$\mathbb{I}_A(x) = \begin{cases} 1 & \text{si } x \in A \\ 0 & \text{si } x \notin A \end{cases}$$

Nótese que $\{A_i\}_{i=1}^k$ define una partición de Ω con elementos de $\mathcal{F}$ y que $\mathbb{E}(\mathbb{I}_{A_i}) = \mathbb{P}(A_i)$. Sabiendo esto, se tienen las siguientes propiedades de $\mathbb{E}$, cuyas demostraciones son inmediatas y se pueden encontrar en las páginas 77 y 78 de Billingsley:

- $\mathbb{E}$ es un operador lineal. Es decir, sean $X = \sum_{i=1}^k x_i \mathbb{I}_{A_i}$ e $Y = \sum_{j=1}^h y_j \mathbb{I}_{B_j}$ para dos particiones finitas $\{A_i\}_{i=1}^k$ $\{B_j\}_{j=1}^h$ de Ω ,

$$\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y) \quad \forall a, b \in \mathbb{R}.$$

- Conserva el orden: si $X(\omega) \leq Y(\omega) \quad \forall \omega \in \Omega$ ($X \leq Y$), entonces $\mathbb{E}(X) \leq \mathbb{E}(Y)$.
- Si X e Y son independientes, $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.

Considerando la variable binomial $X_n \sim \text{Bin}(n, p)$ como suma de n variables Bernoulli con probabilidad p , es sencillo comprobar que su valor esperado será np :

$$\mathbb{E}(\text{Bin}(n, p)) = n\mathbb{E}(\text{Ber}(p)) = np.$$

Una vez entendido el valor esperado para variables aleatorias simples, es momento de extender la noción a las variables no negativas. Sea X una variable aleatoria no negativa, su esperanza se define como sigue:

$$\mathbb{E}(X) = \sup\{\mathbb{E}(Y), \quad Y \text{ simple}, \quad Y \leq X\}. \quad (2.3)$$

En este caso, al contrario de lo que pasaba con las variables aleatorias simples, es posible que la esperanza valga infinito (si es finito, se dirá que X es integrable). Un ejemplo de esto es la variable $X = 2^n$ donde n es el número de lanzamientos de una moneda necesarios para obtener la primera cara⁸.

Sea Y una variable aleatoria simple tal que $Y = X$ si $X \leq 2^N$, su valor esperado será $\mathbb{E}(Y) = \sum_{i=1}^N 2^i \left(\frac{1}{2^i}\right) = N$. Por lo tanto, por definición, $\mathbb{E}(X) \geq \mathbb{E}(Y) = N$ para todo N y, en consecuencia, $\mathbb{E}(X) = \infty$.

La conservación del orden es consecuencia inmediata de la definición vista en (2.3). A pesar de que, como veremos, las otras dos propiedades anteriormente mencionadas para variables aleatorias simples también se cumplen para las variables no negativas, sus demostraciones no son tan sencillas e intuitivas. Será esencial el Teorema de Convergencia Monótona que se presenta a continuación:

⁸Como curiosidad, este ejemplo fue propuesto por Nicolas I Bernoulli en 1713 a modo de juego de apuestas y se conoce como “la paradoja de San Petersburgo”. En él, X representa la cantidad a percibir por el jugador y la decisión está en cuánto estaría dispuesto a apostar para jugarlo sabiendo que la esperanza es infinita. Debido a que sus estudios no obtuvieron resultados, presentó el problema a su primo Daniel Bernoulli, quien, tras años de investigación, publica en 1738 un nuevo sistema de medición en su obra *Exposición de una nueva teoría de la medición del riesgo*, sentando las bases de la teoría de la utilidad esperada (véase Bernoulli, 1738).

Teorema 2.1. (*Teorema de convergencia monótona*) Sea $\{X_n\}_{n \in \mathbb{N}}$ una sucesión de variables aleatorias no negativas que convergen monótonamente a X , esto es $0 \leq X_1 \leq X_2 \leq \dots$ y $\lim_{n \rightarrow \infty} X_n(\omega) = X(\omega) \quad \forall \omega \in \Omega$ y se denota por $\{X_n\} \uparrow X$. Entonces, $\mathbb{E}(X) = \lim_{n \rightarrow \infty} \mathbb{E}(X_n)$.

Demostración: Por la conservación del orden de la esperanza, tenemos que $\mathbb{E}(X_1) \leq \mathbb{E}(X_2) \leq \dots \leq \lim_{n \rightarrow \infty} \mathbb{E}(X_n) \leq \mathbb{E}(X)$. Por tanto, bastará probar la otra desigualdad $\mathbb{E}(X) \leq \lim_{n \rightarrow \infty} \mathbb{E}(X_n)$. Para ello, probaremos que $\forall Y \leq X, \lim_{n \rightarrow \infty} \mathbb{E}(X_n) \geq \mathbb{E}(Y)$, siendo Y una variable aleatoria simple.

Suponiendo $Y = \sum_{j=1}^k y_j \mathbb{I}_{A_j}$ para una partición finita $\{A_j\}_{j=1}^k$ de Ω , se cumple que $y_j \leq X(\omega) \quad \forall \omega \in A_j$. Ahora, fijado $\varepsilon > 0$ arbitrario y definiendo $A_{nj} = \{\omega \in A_j : X_n(\omega) \geq y_j - \varepsilon\}$, se tiene que $\{A_{nj}\} \uparrow A_j$. Además, por la continuidad de $\mathbb{P}$, $\lim_{n \rightarrow \infty} \mathbb{P}(A_{nj}) = \mathbb{P}(A_j)$; y por la definición de A_{nj} , $\mathbb{E}(X_n) \geq \sum_{j=1}^k (y_j - \varepsilon) \mathbb{P}(A_{nj})$. Esto implica que $\lim_{n \rightarrow \infty} \sum_{j=1}^k (y_j - \varepsilon) \mathbb{P}(A_{nj}) = \sum_{j=1}^k (y_j - \varepsilon) \mathbb{P}(A_j) = \mathbb{E}(Y) - \varepsilon$ para todo $\varepsilon > 0$, con lo que se concluye la prueba de la segunda desigualdad. $\square$

Este resultado garantiza la convergencia de las esperanzas si la sucesión de variables aleatorias converge monótonamente. ¿Pero dicha sucesión existirá dada cualquier variable aleatoria no negativa X ? La respuesta es sí, y además se puede conseguir que la sucesión sea de variables simples.

Antes de probarlo, debemos introducir un poco de notación. Consideremos para cualquier $x \geq 0$ la función $\Psi_n(x) = \min\{n, 2^{-n} \lfloor 2^n x \rfloor\}$, donde $\lfloor y \rfloor$ denota la parte entera de y . Veamos que dado un $x \geq 0$ fijo, se tiene que $\{\Psi_n(x)\} \uparrow x$.

Fijado entonces $x \geq 0$, como se cumple $2 \lfloor 2^n x \rfloor \leq \lfloor 2^{n+1} x \rfloor$, se tiene $\Psi_n(x) \leq \Psi_{n+1}(x)$ para todo $n \in \mathbb{N}$. Faltaría únicamente probar que $\lim_{n \rightarrow \infty} \Psi_n(x) = x$. Para ello, sabiendo que la desigualdad $2^{-n}(2^n x - 1) \leq 2^{-n} \lfloor 2^n x \rfloor \leq x$ se cumple para todo $n \in \mathbb{N}$, concluyendo que $x = \lim_{n \rightarrow \infty} 2^{-n}(2^n x - 1) \leq \lim_{n \rightarrow \infty} 2^{-n} \lfloor 2^n x \rfloor \leq x$.

Nótese que $\Psi_n(x)$ tiene un número de $n2^n + 1$ posibles valores, es decir, un número finito.

Ahora sí estamos en condiciones de enunciar el resultado:

Teorema 2.2. Sea X una variable aleatoria no negativa y sea $X_n = \Psi_n(X)$ para todo n , se cumple que $\{X_n\}$ es una sucesión de variables aleatorias simples que convergen monótonamente a X .

Estos dos teoremas nos permiten extender la primera de las propiedades que vimos en las variables simples a las no negativas de manera inmediata. Sean X y Y variables aleatorias no negativas y sean $X_n = \Psi_n(X)$ e $Y_n = \Psi_n(Y)$, usando la linealidad de la esperanza para variables simples, para cualquier $a \geq 0$ y $b \geq 0$ se cumple

$$\mathbb{E}(aX + bY) = \lim_{n \rightarrow \infty} \mathbb{E}(aX_n + bY_n) = \lim_{n \rightarrow \infty} (a\mathbb{E}(X_n) + b\mathbb{E}(Y_n)) = a\mathbb{E}(X) + b\mathbb{E}(Y).$$

Para la esperanza de la multiplicación de variables independientes, debemos usar la Proposición 2.1. Sean X e Y variables aleatorias independientes, lo serán también $X_n = \Psi_n(X)$ e $Y_n = \Psi_n(Y)$. Por lo tanto $\mathbb{E}(XY) = \lim_{n \rightarrow \infty} \mathbb{E}(X_n Y_n) = \lim_{n \rightarrow \infty} \mathbb{E}(X_n) \mathbb{E}(Y_n) = \mathbb{E}(X) \mathbb{E}(Y)$.

Como última propiedad interesante es la aditividad numerable. Sea $\{X_n\}$ una sucesión de variables aleatorias no negativas, usando teorema de convergencia monótona para $S_n = \sum_{i=1}^n X_i$ se garantiza que $\mathbb{E}(\sum_{n \in \mathbb{N}} X_i) = \sum_{n \in \mathbb{N}} \mathbb{E}(X_i)$.

Ahora ya estamos en condiciones de extender todos estos conceptos a una variable aleatoria general X . Considerando $X^+(\omega) = \max\{X(\omega), 0\}$ y $X^-(\omega) = \max\{-X(\omega), 0\}$, podemos escribir $X = X^+ - X^-$, siendo X^+ y X^- variables no negativas. Por lo tanto, la esperanza de X será $\mathbb{E}(X) = \mathbb{E}(X^+) - \mathbb{E}(X^-)$, supuesto que $\mathbb{E}(X^+)$ o $\mathbb{E}(X^-)$ es finita. En caso de que ambos valores sean finitos, se dirá que X es integrable.

Salvo la aditividad numerable, la esperanza para variables generales cumple las propiedades anteriormente mencionadas (véase Billingsley, 1995, Sección 21).

Como ya se ha mencionado anteriormente, los conceptos de distribución y de valor esperado están relacionados. El siguiente resultado, denominado *Teorema de cambio de variable*, garantiza que la distribución de una variable aleatoria determina por completo su valor esperado, así como la esperanza de cualquier función de la misma. Además, permite expresar cualquier esperanza en el espacio general Ω como una esperanza en la recta real.

Se utilizará la siguiente notación para el valor esperado de una variable aleatoria X definida en un espacio de probabilidad $(\Omega, \mathcal{F}, \mathbb{P})$:

$$\mathbb{E}(X) = \int X d\mathbb{P} = \int_{\Omega} X(\omega) \mathbb{P}(d\omega). \quad (2.4)$$

Así, para una función f medible,

$$\mathbb{E}(f(X)) = \int f(X) d\mathbb{P} = \int_{\Omega} f(X(\omega)) \mathbb{P}(d\omega).$$

En el caso de integrales en $\mathbb{R}$, siendo $\mathbb{P}_X$ la distribución de X y F_X la distribución inducida por $\mathbb{P}_X$ (es decir, $F_X(x) = \mathbb{P}_X(-\infty, x]$) se utilizará indistintamente

$$\int_{-\infty}^{\infty} f(x) \mathbb{P}_X(dx) = \int_{-\infty}^{\infty} f(x) F_X(dx).$$

Teorema 2.3. (*Teorema de cambio de variable*) Sea $(\Omega, \mathcal{F}, \mathbb{P})$ un espacio de probabilidad y X una variable aleatoria definida en él con distribución $\mathbb{P}_X$, entonces para cualquier función medible $f : \mathbb{R} \rightarrow \mathbb{R}$ se tiene que

$$\mathbb{E}(f(X)) = \int_{\Omega} f(X(\omega)) \mathbb{P}(d\omega) = \int_{-\infty}^{\infty} f(x) \mathbb{P}_X(dx). \quad (2.5)$$

Es decir, $\mathbb{E}_{\mathbb{P}}(f(X)) = \mathbb{E}_{\mathbb{P}_X}(f)$ (suponiendo que la esperanza esté bien definida).

Demostración: El método a seguir en la demostración es similar al que utilizamos en las anteriores propiedades de la esperanza: desde el estudio de las variables simples hasta extenderlo a variables generales.

Primero supongamos que $f = \mathbb{I}_B$ es la función indicatriz para un conjunto $B \subset \mathbb{R}$ de Borel. Se cumple que $\int_{\Omega} f(X(\omega))\mathbb{P}(d\omega) = \int_{\Omega} \mathbb{I}_B(\omega)\mathbb{P}(d\omega) = \int_{\{\omega \in \Omega / X(\omega) \in B\}} \mathbb{P}(d\omega) = \mathbb{P}(X^{-1}(B))$ mientras que $\int_{\mathbb{R}} f(x)\mathbb{P}_X(dx) = \int_B \mathbb{P}_X(dx) = \mathbb{P}_X(B)$, que coinciden por definición.

Ahora bien, si consideramos que f es una función simple no negativa, entonces es una combinación lineal positiva finita de funciones indicadoras. Solamente aplicando la linealidad de la integral, que no deja de ser una notación para la esperanza en este contexto, se verifica que (2.5) también se cumple.

Si se trata de un caso en el que f es una función general no negativa, por el Teorema 2.2. existe una sucesión de funciones $\{f_n\}$ simples que converge monótonamente a f . Como (2.5) se cumple cuando f es sustituida por f_n para todo n , el Teorema de Convergencia Monótona (Teorema 2.1.) permite asegurar que también se cumplirá para f .

La prueba para f general se realiza descomponiendo f en su parte positiva y su parte negativa: $f(x) = f^+(x) - f^-(x)$, siendo $f^+(x) = \max\{0, f(x)\}$ y $f^-(x) = \max\{0, -f(x)\}$ dos funciones generales no negativas. De nuevo, aplicando la linealidad de la integral y el resultado ya probado para funciones no negativas, queda demostrado. $\square$

Además de la media, la esperanza es un instrumento útil para definir otros conceptos claves en la Teoría de la Probabilidad tales como la varianza o la función característica.

La varianza es una medida de dispersión que nos indica cuan alejados están los posibles valores del valor esperado. Su fórmula es $\sigma_X^2 = \text{Var}(X) = \mathbb{E}((X - \mathbb{E}(X))^2)$. Desarrollando esta expresión, se obtiene equivalentemente $\text{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$. Cuanto mayor sea la varianza, mayor será la dispersión de los valores. Su raíz cuadrada, σ_X , se denomina desviación típica.

Es inmediato comprobar que en el caso de una variable dicotómica con probabilidad p se cumplirá que su varianza es igual a $p(1-p)$. Para calcular, entre otras cosas, la varianza de la binomial, será necesario mencionar un par de propiedades importantes de la varianza.

La primera de ellas es que $\text{Var}(aX + b) = a^2\text{Var}(X)$ para todo $a, b \in \mathbb{R}$. Esto se puede desarrollar fácilmente a partir de la definición: $\text{Var}(aX + b) = \mathbb{E}((aX + b - \mathbb{E}(aX + b))^2)$. Desarrollando esta expresión se llega a que esto es igual a $a^2\mathbb{E}((X - \mathbb{E}(X))^2) = a^2\text{Var}(X)$.

Por otro lado, si X e Y son independientes, entonces la varianza de la suma será la suma de las varianzas: $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$. También, a partir de la definición, $\text{Var}(X + Y) = \mathbb{E}((X + Y - \mathbb{E}(X + Y))^2) = \mathbb{E}((X - \mathbb{E}(X) + Y - \mathbb{E}(Y))^2) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$. Como la covarianza entre X e Y se define como $\text{Cov}(X, Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$ y será nula siempre que X e Y sean independientes, queda demostrada la propiedad.

Por lo anterior, se demuestra que la varianza de la binomial será $np(1-p)$.

2.3. Convergencia en distribución.

Cuando se trata de una sucesión de variables aleatorias $X_1, X_2, \dots$ definidas en el mismo espacio de probabilidad, es posible que converjan a otra variable X . Al contrario de lo que sucede en otros casos de las matemáticas como en el de las series o las funciones, en las variables aleatorias existen varias nociones de convergencia. En el TCL está presente la convergencia en distribución, concepto en el que ahondaremos en esta sección, pero será necesario mencionar los otros tipos de convergencia para comprender algunas definiciones y resultados de este trabajo.

Ya hemos mencionado la convergencia puntual previamente cuando tratamos el teorema de convergencia monótona. En ese caso, se cumple cuando para todo $\omega \in \Omega$, $\lim_{n \rightarrow \infty} X_n(\omega) = X(\omega)$. Para la convergencia casi segura (c.s. o con probabilidad uno), denotada $X_n \xrightarrow{c.s.} X$, será suficiente que converja para todo ω salvo un conjunto de probabilidad nula, esto es

$$\mathbb{P}(\lim_{n \rightarrow \infty} X_n = X) = \mathbb{P}(\omega \in \Omega : \lim_{n \rightarrow \infty} X_n(\omega) = X(\omega)) = 1.$$

Por otra parte, diremos que converge en probabilidad ($X_n \xrightarrow{p} X$) si, y solo si

$$\lim_{n \rightarrow \infty} \mathbb{P}(|X_n - X| \geq \varepsilon) = 0 \quad \forall \varepsilon > 0.$$

Es decir, para n suficientemente grande, la probabilidad de obtener un valor alejado de X es infinitamente baja, o lo que es lo mismo, la probabilidad de que X_n esté en un determinado intervalo vendrá determinado por la probabilidad de que X esté en él.

La relación entre ambos tipos de convergencia es que la casi segura implica la convergencia en probabilidad. La demostración de esto se puede ver en la página 70 de Billingsley. Como ejemplo que demuestra que la otra implicación no se cumple es el que considera la sucesión de variables aleatorias independientes $\{X_n\}_{n \in \mathbb{N}}$ tales que $\mathbb{P}(X_n = 1) = 1/n = 1 - \mathbb{P}(X_n = 0)$. Es obvio X_n converge en probabilidad, pues la probabilidad de obtener valores alejados del 0 con n grande es casi nula; pero no lo hace de forma casi segura, ya que el límite superior de X_n será 1, cumpliéndose $\mathbb{P}(X_n \rightarrow 0) = 0$.

Ligado a la convergencia en probabilidad, existen los conceptos de O grande en probabilidad (O_P) y o pequeña en probabilidad (o_P). La notación $o_P(1)$ representa una sucesión de variables aleatorias que convergen en probabilidad a cero, es decir, $X_n \xrightarrow{p} 0 \Leftrightarrow X_n = o_P(1)$. Por otra parte, O_P representa una sucesión de variables aleatorias acotada en probabilidad. Esto es, para todo $\varepsilon > 0$, existe $M > 0$ tal que $\sup_{n \in \mathbb{N}} \mathbb{P}(|X_n| > M) < \varepsilon$, que no es otra cosa que decir que la probabilidad de que X_n esté dentro del conjunto $[-M, M]$ es casi uno, para todo $n \in \mathbb{N}$.

Por comodidad, los símbolos O_P y o_P se reducen a los habituales O y o , respectivamente. Simplemente usando las definiciones, es simple comprobar que si $X_n = o(1)$ entonces $X_n = O(1)$.

De forma general, se dice que $X_n = o(R_n)$ si $X_n = R_n o(1)$. Es decir, la tasa de convergencia en probabilidad de X_n es R_n . Esto ocurre de forma análoga para la O grande en probabilidad.

Por ejemplo, cuando decimos que una sucesión X_n es $O(1/\sqrt{n})$, significa que $\sqrt{n}X_n$ está acotada en probabilidad.

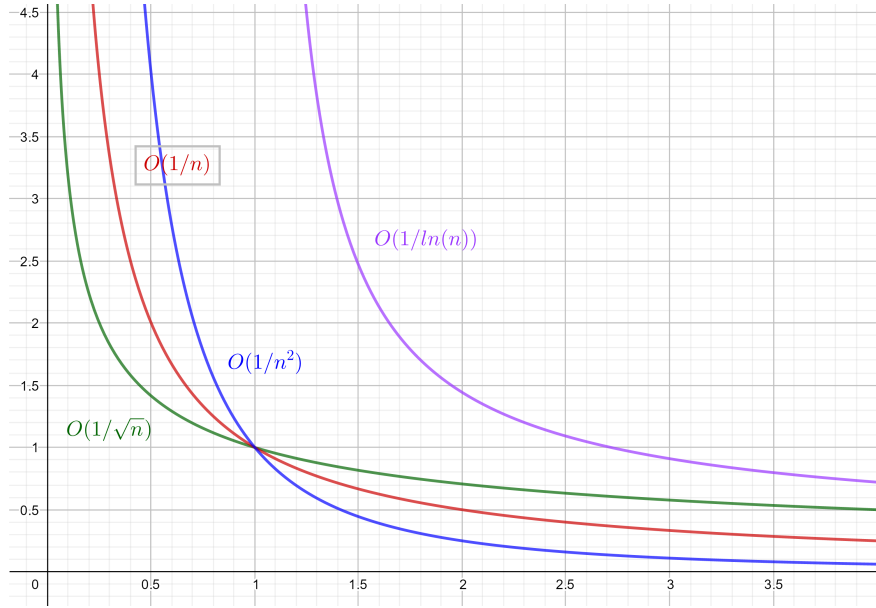


Figura 2.2: Orden de convergencia de sucesiones $O(1/\ln(n))$, en lila; $O(1/\sqrt{n})$, en verde; $O(1/n)$, en rojo; y $O(1/n^2)$, en azul.

En la Figura 2.2 se aprecian sucesiones con distinta tasa de convergencia, observando que cuanto menor es la tasa R_n para grandes valores de n , más rápido converge al cero.

Por último y más importante, por ser el tipo de convergencia presente en el TCL, será necesario profundizar sobre el concepto de convergencia en distribución o convergencia débil.

La convergencia en distribución de una sucesión de variables aleatorias $\{X_n\}$ a X se producirá siempre que

$$\lim_n \mathbb{P}(X_n \leq x) = \mathbb{P}(X \leq x) \quad \forall x \in \mathbb{R} / \mathbb{P}(X = x) = 0$$

y se denotará como $X_n \xrightarrow{d} X$.

Esta definición se podrá escribir en términos de las funciones de distribución, de manera que si $\{F_n\}$ y F son las funciones de distribución de dichas variables aleatorias, sucederá que

$$\lim_n F_n(x) = F(x) \quad \text{para todo punto de continuidad } x \text{ de } F,$$

puesto que es obvio que los puntos de continuidad x de F son aquellos que cumplen que $\mathbb{P}(X = x) = 0$. Se denotará $F_n \Rightarrow F$.

Como se ha visto en la Sección 2.2, el *Teorema de Extensión* garantiza que la función de distribución F_X de una variable X induce una única medida de probabilidad. Por ello, se podrá definir equivalentemente la convergencia en distribución usando la notación con distribuciones. A partir de este momento, por comodidad, se usará la notación μ_X para hablar de la distribución de la variable X .

Por lo tanto, se dirá que la sucesión de distribuciones $\{\mu_n\}$ converge débilmente a μ , y se denota $\mu_n \Rightarrow \mu$, si se cumple

$$\lim_n \mu_n(A) = \mu(A)$$

para todo conjunto A de la forma $A = (-\infty, x]$ tal que $\mu\{x\} = 0$.

Para probar el teorema de caracterización de la convergencia en distribución, será útil mencionar un par de resultados conocidos con los nombres de Teorema de Convergencia Dominada y el Teorema de Skorohod, correspondientes a los Teoremas 16.5. y 25.6. de Billingsley, respectivamente.

Teorema 2.4. (Teorema de Convergencia Dominada) Sea μ una medida de probabilidad, si la sucesión de funciones $\{f_n\}$ está uniformemente acotada, es decir, $\sup_{n,x} \{|f_n(x)|\} < \infty$, y además f_n converge a f en casi todo punto (el conjunto de puntos que no convergen a f es de medida nula), entonces se cumple que $\int f_n d\mu \rightarrow \int f d\mu$.

Teorema 2.5. (Teorema de Skorohod) Sean $\{\mu_n\}$ y μ distribuciones de probabilidad que cumplen $\mu_n \Rightarrow \mu$, entonces existen variables aleatorias $\{Y_n\}$, Y en algún espacio de probabilidad $(\Omega, \mathcal{F}, \mathbb{P})$ tales que Y_n tiene distribución μ_n para cada $n \in \mathbb{N}$, Y tiene distribución μ y $Y_n(\omega) \rightarrow Y(\omega)$ para todo $\omega \in \Omega$.

Teorema 2.6. (Teorema de Caracterización) Las siguientes afirmaciones son equivalentes:

- (1) $\mu_n \Rightarrow \mu$.
- (2) $\lim_n \int f d\mu_n = \int f d\mu$ para toda función f real, continua y acotada.
- (3) $\lim_n \mu_n(A) = \mu(A)$ para todo conjunto de Borel A tal que $\mu(\partial A) = 0$.

Demostración: “(1) $\implies$ (2)” Suponiendo $\mu_n \Rightarrow \mu$ y considerando las variables Y_n , Y del Teorema de Skorohod. Sea f una función real que cumple $\mu(D_f) = 0$ siendo D_f el conjunto de discontinuidades de f , entonces $f(Y_n)$ converge de forma casi segura a $f(Y)$. Ahora, por el Teorema de Convergencia Dominada se tiene que $\int f d\mu_n = \mathbb{E}[f(Y_n)] \rightarrow \mathbb{E}[f(Y)] = \int f d\mu$.

“(1) $\implies$ (3)” Sea A conjunto de Borel tal que $\mu(A) = 0$ y $f = \mathbb{I}_A$, entonces $D_f = \partial A$, de donde se sigue a partir del mismo razonamiento que en la anterior implicación que $\mu_n(A) = \int f d\mu_n \rightarrow \int f d\mu = \mu(A)$.

“(2) $\implies$ (1)” Esta implicación es obvia, ya que $\partial(-\infty, x] = \{x\}$.

“(3) $\implies$ (1)” Sea $\varepsilon > 0$, y sea f la función definida como $f(t) = 1$ si $t \leq x$, $f(t) = 0$ si $t \geq x + \varepsilon$ y lineal en $(x, x + \varepsilon)$, es decir, $f(t) = (\varepsilon + x - t)/\varepsilon$ si $x < t < x + \varepsilon$. Así definida, f es continua y se cumple que $\mathbb{I}_{(-\infty, x]} \leq f \leq \mathbb{I}_{(-\infty, x+\varepsilon]}$. Esto implica que

$$\limsup_n \mu_n((-\infty, x]) \leq \limsup_n \int f d\mu_n = \int f d\mu \leq \mu((-\infty, x + \varepsilon]).$$

Como esto se cumple $\forall \varepsilon > 0$, la continuidad de μ implica que $\limsup_n \mu_n((-\infty, x]) \leq \mu((-\infty, x])$.

De forma análoga, sea g la función definida como $g(t) = 1$ si $t \leq x - \varepsilon$, $g(t) = 0$ si $t \geq x$ y lineal en $(x - \varepsilon, x)$, es decir, $g(t) = (x - t)/\varepsilon$ en dicho intervalo. Así definida, se cumple

$$\liminf_n \mu_n((-\infty, x]) \geq \liminf_n \int g d\mu_n = \int g d\mu \geq \mu((-\infty, x - \varepsilon])$$

para todo $\varepsilon > 0$, y por la continuidad de μ , $\liminf_n \mu_n((-\infty, x]) \geq \mu((-\infty, x))$.

Finalmente, como $\mu\{x\} = 0$ se tiene que $\mu((-\infty, x)) = \mu((-\infty, x])$, y por lo tanto $\liminf_n \mu_n((-\infty, x]) = \limsup_n \mu_n((-\infty, x]) = \mu((-\infty, x])$. $\square$

Se concluye por tanto que una sucesión de variables aleatorias X_n convergerá en distribución si para todo suceso A tal que $\mu(\partial A) = 0$ se cumple $\lim_n \mu_n(A) = \mu(A)$. A pesar de que la condición de tipo topológico $\mu(\partial A) = 0$ pueda resultar un poco extraña, es necesaria para que la convergencia en distribución no sea demasiado restrictiva.

A modo de ejemplo, se considera la sucesión de variables aleatorias $X_n \sim \text{Bin}(n, p)$. Suponiendo cierto el TCL, la variable $Z_n = \frac{X_n - np}{\sqrt{npq}}$ converge en distribución a $Z \sim N(0, 1)$. Sin embargo, esta convergencia no se produciría si se pidiese que $\mathbb{P}(Z_n \in A) \rightarrow \mathbb{P}(Z \in A)$ para todo suceso A . Considerando $A_n = \{(k - np)/\sqrt{npq} : k = 0, 1, \dots, n\}$, está claro que $\mathbb{P}(Z_n \in A_n) = 1$ y que $\mathbb{P}(Z \in A_n) = 0$, por ser A un conjunto finito. Es más, si se define $A = \bigcup_n A_n$, se tendría que para todo $n \in \mathbb{N}$, $\mathbb{P}(Z_n \in A) = 1$ por ser $A_n \subset A$ y que $\mathbb{P}(Z \in A) \leq \sum_n \mathbb{P}(Z_n \in A) = 0$. Y por lo tanto, $\mu_n(A) = 1$ no converge a $\mu(A) = 0$.

Capítulo 3

Demostración del Teorema Central del Límite.

Una poderosa herramienta que facilita enormemente la demostración del TCL es la función característica. Como su propio nombre indica, ésta permite determinar completamente a la distribución, esto es, cada distribución tendrá su propia función característica y dos distribuciones distintas no pueden tener la misma función característica. Además, también se cumple que la convergencia puntual de una sucesión de funciones características implicará la convergencia débil de sus correspondientes distribuciones.

A lo largo de este capítulo se demostrarán estos resultados que permitirán trabajar con las funciones características en lugar de hacerlo directamente con las distribuciones.

3.1. Funciones características.

Dada una variable aleatoria X , se define su función característica como

$$\varphi_X(t) = \mathbb{E}(e^{itX}) = \mathbb{E}(\cos(tX)) + i\mathbb{E}(\sin(tX)), \quad t \in \mathbb{R}. \quad (3.1)$$

Nótese que φ estará definida para cualquier variable aleatoria X , ya que las funciones seno y coseno son funciones acotadas y, en consecuencia, la esperanza de su valor absoluto es finita.

Usando dicha definición, si $X \sim \text{Ber}(p)$ es una variable Bernoulli con probabilidad p , su función característica será

$$\varphi_X(t) = e^{it0}\mathbb{P}(X=0) + e^{it1}\mathbb{P}(X=1) = q + pe^{it}. \quad (3.2)$$

Para el cálculo de la función característica de la normal estándar, además de la definición se hará uso del Teorema de Cambio de Variable (véase Teorema 2.3.) y de la conocida Integral de Gauss¹. El procedimiento es el siguiente:

¹La igualdad $\int_{-\infty}^{\infty} e^{-u^2} du = \sqrt{\pi}$ se conoce como integral de Gauss o gaussiana y su demostración se puede ver en Weisstein.

$$\varphi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{itx} e^{-x^2/2} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(\frac{x-it}{\sqrt{2}})^2 - \frac{t^2}{2}} dx.$$

Realizando el cambio de variable $u = \frac{x-it}{\sqrt{2}}$, la anterior expresión queda como sigue:

$$\varphi(t) = \frac{1}{\sqrt{\pi}} e^{-t^2/2} \int_{-\infty}^{\infty} e^{-u^2} du = e^{-t^2/2}. \quad (3.3)$$

Para el cálculo de la función característica de la binomial bastará con probar que la función característica de la suma de variables independientes da como resultado el producto de las funciones características.

Supongamos X_1 y X_2 dos variables aleatorias independientes con funciones características φ_1 y φ_2 , respectivamente. Si $Y_j = \cos(tX_j)$ y $Z_j = \sin(tX_j)$ con $j = 1, 2$, entonces, por la Proposición 2.1, (Y_1, Z_1) , (Y_2, Z_2) , (Y_1, Z_2) y (Y_2, Z_1) serán independientes. Por lo tanto, haciendo uso de las identidades trigonométricas $\cos(\alpha + \beta) = \cos(\alpha)\cos(\beta) - \sin(\alpha)\sin(\beta)$ y $\sin(\alpha + \beta) = \sin(\alpha)\cos(\beta) + \cos(\alpha)\sin(\beta)$, se obtiene la igualdad:

$$\begin{aligned} \varphi_1(t)\varphi_2(t) &= (\mathbb{E}(Y_1) + i\mathbb{E}(Z_1))(\mathbb{E}(Y_2) + i\mathbb{E}(Z_2)) = \\ &= \mathbb{E}(Y_1)\mathbb{E}(Y_2) - \mathbb{E}(Z_1)\mathbb{E}(Z_2) + i(\mathbb{E}(Y_1)\mathbb{E}(Z_2) + \mathbb{E}(Y_2)\mathbb{E}(Z_1)) = \\ &= \mathbb{E}(Y_1Y_2 - Z_1Z_2 + i(Y_1Z_2 + Y_2Z_1)) = \mathbb{E}(e^{it(X_1+X_2)}) = \varphi_{X_1+X_2}(t). \end{aligned}$$

Esto se puede generalizar a la suma de tres o más variables aleatorias por inducción, obteniendo la siguiente fórmula:

$$\varphi_{\sum_{k=1}^n X_k}(t) = \mathbb{E}(e^{it\sum_{k=1}^n X_k}) = \prod_{k=1}^n \mathbb{E}(e^{itX_k}) = \prod_{k=1}^n \varphi_{X_k}(t). \quad (3.4)$$

Por lo tanto, si se trata la binomial $X_n \sim \text{Bin}(n, p)$ como la suma de n variables Bernoulli independientes, por (3.4) y sabiendo que la función característica de una Bernoulli es $q + pe^{it}$, se tiene

$$\varphi_{X_n}(t) = \mathbb{E}(e^{itX_n}) = (q + pe^{it})^n. \quad (3.5)$$

Otra propiedad interesante y que se usará en la demostración del TCL es que si X tiene como función característica φ , entonces la variable $aX + b$ tendrá por función característica

$$\mathbb{E}(e^{it(aX+b)}) = e^{itb}\varphi(at). \quad (3.6)$$

Algo que es importante destacar sobre las funciones características es que determinan de forma única a la distribución μ de la variable aleatoria X . Es decir, dos distribuciones distintas no pueden tener la misma función característica. Esto es esencial para que se pueda trabajar con ellas en lugar de hacerlo con la función de distribución directamente. El resultado que demuestra esto se conoce como Teorema de Unicidad y su enunciado es el siguiente:

Teorema 3.1. (*Teorema de Unicidad*) Si una distribución tiene función característica φ y para todo $a < b$ tal que $\mu\{a\} = \mu\{b\} = 0$ se tiene que

$$\mu(a, b] = \lim_{T \rightarrow \infty} \frac{1}{2\pi} \int_{-T}^T \frac{e^{-ita} - e^{-itb}}{it} \varphi(t) dt. \quad (3.7)$$

Por lo tanto, la función característica caracteriza completamente a la distribución, es decir, dos distribuciones distintas no pueden tener la misma función característica.

Observación 3.1. La ecuación (3.7) se conoce por el nombre de fórmula de inversión de Lévy. Así definida, para valores fijos de a y b , el integrando es una función continua y acotada. En caso de que $\int_{-\infty}^{\infty} |\varphi(t)| dt < \infty$, entonces el *Teorema de Convergencia Dominada* (véase Teorema 2.4) garantiza que (3.7) se puede extender a toda la recta real:

$$\lim_{T \rightarrow \infty} \frac{1}{2\pi} \int_{-T}^T \frac{e^{-ita} - e^{-itb}}{it} \varphi(t) dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-ita} - e^{-itb}}{it} \varphi(t) dt.$$

Si F denota la función de distribución de μ , $a = x$ y $b = x + h$ con $h = b - a > 0$, entonces

$$\frac{F(x + h) - F(x)}{h} = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-itx} - e^{-it(x+h)}}{ith} \varphi(t) dt.$$

Por lo tanto, la derivada de F es

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \varphi(t) dt,$$

con lo cual μ tiene densidad continua que se calcula aplicando la transformada de Fourier inversa a la función característica.

Demostración: A continuación se probará únicamente que la fórmula de inversión implica la unicidad, pudiéndose ver el resto de la demostración en la Sección 26 de Billingsley. Supongamos μ y ν dos distribuciones de probabilidad con la misma función característica. Entonces, por la fórmula de inversión, $\mu(a, b] = \nu(a, b]$ para todo $a < b$ tal que $\mu\{a\} = \mu\{b\} = \nu\{a\} = \nu\{b\} = 0$.

Esto no se trata de una restricción, pues si μ y ν coinciden en todos los intervalos de este tipo, necesariamente deben coincidir para cualquier intervalo. Si se piensa en dos funciones de distribución F y G que coinciden en toda la parte continua, en caso de tener saltos, estos deben coincidir. Además, como la σ -álgebra generada por la familia de conjuntos $\{(a, b]\}_{a < b}$ es la σ -álgebra de Borel, entonces $\mu = \nu$. $\square$

3.2. Teorema de Continuidad.

Tras ver que la función característica es única para cada distribución, se tratará en esta sección el Teorema de Continuidad, que da una condición necesaria y suficiente para la convergencia en distribución de una sucesión de distribuciones $\{\mu_n\}$. Considerando dicha

sucesión y la distribución μ , y sus respectivas funciones características $\{\varphi_n\}$ y φ , se cumplirá que $\{\mu_n\} \Rightarrow \mu$ si y solo si $\varphi_n(t) \rightarrow \varphi(t)$ para todo t . La primera implicación es consecuencia directa de la definición de convergencia débil. Sin embargo, la demostración del recíproco no es tan sencilla. A lo largo de esta sección se expondrán varios resultados, algunos de tipo general, necesarios para completar la misma. La idea es la siguiente:

Sabiendo que la sucesión de funciones características $\{\varphi_n\}$ converge puntualmente a φ para todo punto, si la sucesión de distribuciones $\{\mu_n\}$ es convergente, el *Teorema de Unicidad* garantiza que lo será a μ . Este razonamiento puede ser aplicado a cualquier subsucesión $\{\mu_{n_k}\}$ convergente. Sin embargo, ¿bajo qué condiciones existe dicha subsucesión? Si se piensa en una sucesión de números reales $\{a_n\}$, el Teorema de Bolzano-Weierstrass² dice que existirá una subsucesión $\{a_{n_k}\}$ convergente si la sucesión está acotada. Además, si todas las subsucesiones convergentes lo hacen al mismo límite, entonces la sucesión también lo hará. Esto se puede extender para la sucesión de distribuciones, como veremos en el Teorema 3.3, garantizando la existencia de una subsucesión convergente si la sucesión está acotada en probabilidad. Por la convergencia de las funciones características, toda subsucesión $\{\mu_{n_k}\}$ convergente lo hará a μ , lo que supone una condición suficiente para la convergencia débil de $\{\mu_n\}$, cuestión que se verá en el Corolario 3.1. Finalmente, el Teorema 3.4 prueba que la convergencia de las funciones características implica que $\{\mu_n\}$ está acotada en probabilidad.

Antes de proceder a la demostración de cada uno de los resultados mencionados anteriormente, será útil probar el Teorema de selección de Helly, cuyo enunciado es el siguiente:

Teorema 3.2. (*Teorema de selección de Helly*) *Para cualquier sucesión de funciones de distribución $\{F_n\}$ existe una subsucesión F_{n_k} y una función monótona creciente y continua por la derecha F tales que $\lim_k F_{n_k}(x) = F(x)$ para cualquier punto de continuidad de F .*

Demostración: Como $\mathbb{Q}$ es numerable, se puede considerar $\mathbb{Q} = \{q_1, q_2, \dots\}$. Por ser $\{F_n(q_1)\}$ una sucesión acotada de números reales, el Teorema de Bolzano-Weierstrass permite afirmar que existirá una subsucesión $\{F_{n_k^{(1)}}(q_1)\}$ tal que $\lim_k \{F_{n_k^{(1)}}(q_1)\}$ existe. Análogamente, $\{F_{n_k^{(1)}}(q_2)\}$ es una sucesión acotada de números reales, por lo que existirá una subsucesión convergente $\{F_{n_k^{(2)}}(q_2)\}$. Además, por ser subsucesión de la primera, se cumplirá que $\{F_{n_k^{(2)}}(q_1)\}$ es convergente. Continuando este proceso se consigue que para cualquier $m \in \mathbb{N}$, existirá una subsucesión $\{n_k^{(m)}\}$ tal que $\{F_{n_k^{(m)}}(q_j)\}$ es convergente, con $j = 1, \dots, m$.

Denotando $n_k = n_k^{(k)}$ al elemento k -ésimo de la subsucesión k -ésima, es obvio que n_k es monótona creciente para k y que $\{F_{n_k}\}$ es una subsucesión de $\{F_n\}$. Se cumple además que para todo $q \in \mathbb{Q}$ existe un $m \in \mathbb{N}$ tal que $\{F_{n_k^{(m)}}(q)\}$ es convergente y, como n_k es una subsucesión de $n_k^{(m)}$, la sucesión $\{F_{n_k}(q)\}$ también lo será. Siendo $G(q)$ este límite, se puede ver que G será monótona creciente por serlo F_{n_k} . Definiendo $F(x) = \inf\{G(q) : q > x\}$, F está definida en toda la recta real y es monótona creciente.

Para probar que es continua por la derecha, tomamos un $x \in \mathbb{R}$ y $\varepsilon > 0$. Por la definición de ínfimo, existe $q > x$ tal que $G(q) < F(x) + \varepsilon$. Si $x \leq y < q$, entonces $F(x) \leq F(y) \leq G(q) < F(x) + \varepsilon$.

²Su demostración se puede ver en la Sección 3.4 de Bartle y Sherbert (2004).

Un vez visto que F cumple las condiciones descritas en el enunciado, solo faltaría probar que efectivamente F es el límite de la subsucesión $\{F_{n_k}\}$. Para ello, suponemos $x \in \mathbb{R}$ punto de continuidad de F . Sea $y < x$ tal que $F(x) - \varepsilon < F(y) \leq F(x)$ y sean q_1 y q_2 dos números racionales que cumplen $y < q_1 < x < q_2$ y $G(q_2) < F(x) + \varepsilon$, se verifica que $F(x) - \varepsilon < G(q_1) \leq G(q_2) < F(x) + \varepsilon$. Finalmente, a partir de que $F_n(q_1) \leq F_n(x) \leq F_n(q_2)$, se sabe que $F_{n_k}(x)$ está comprendido entre $F_n(x) - \varepsilon$ y $F_n(x) + \varepsilon$, de lo que se deduce que los límites inferior y superior de $F_{n_k}(x)$ también lo están, concluyendo que ambos límites coinciden y valen $F(x)$. $\square$

El anterior teorema no permite afirmar que F sea una función de distribución, pues no tiene por qué cumplirse las condiciones de $\lim_{x \rightarrow -\infty} F(x) = 0$ y $\lim_{x \rightarrow \infty} F(x) = 1$. Si se considera la sucesión $\{F_n\}$ tal que $F_n(x) = \mathbb{I}_{[n, \infty)}$, su límite será la función $F(x) = 0$.

Para que F sea una función de distribución se debe cumplir que la sucesión de distribuciones $\{\mu_n\}$ estén acotadas en probabilidad (véase la Página 21). En términos de funciones de distribución, $\{F_n\}$ estará acotada en probabilidad si para todo $n \in \mathbb{N}$ existen $x < y$ tales que $F_n(x) < \varepsilon$ y $F_n(y) > 1 - \varepsilon$.

El siguiente teorema se considera una extensión del Teorema de Bolzano-Weierstrass para distribuciones y permite demostrar lo afirmado anteriormente:

Teorema 3.3. *Si $\{\mu_n\}$ es una sucesión de distribuciones acotadas en probabilidad, entonces existirá una subsucesión $\{\mu_{n_k}\}$ y una distribución μ que cumplen $\mu_{n_k} \Rightarrow \mu$.*

Demostración: Se consideran F_n las funciones de distribución de μ_n . La acotación en probabilidad garantiza que para todo $\varepsilon > 0$ existen x e y tales que $F_n(y) - F_n(x) \geq 1 - \varepsilon$ para todo $n \in \mathbb{N}$. Por el Teorema de selección de Helly, existe una subsucesión $\{F_{n_k}\}$ tal que $\lim_k F_{n_k}(x) \rightarrow F(x)$ para todo punto de continuidad de F , siendo esta función monótona creciente y continua por la derecha. Para ver que es una función de distribución únicamente será necesario ver que $\lim_{z \rightarrow -\infty} F(z) = 0$ y $\lim_{z \rightarrow \infty} F(z) = 1$. Por construcción, $F(z) \in [0, 1]$, entonces $1 \geq \lim_{z \rightarrow \infty} (F(-z) - F(z)) = \lim_k (F_{n_k}(y) - F_{n_k}(x)) \geq 1 - \varepsilon$. $\square$

La existencia de dicha subsucesión no nos permite afirmar que la sucesión de distribuciones convergerá a ella. El siguiente resultado demuestra que si cada subsucesión tiene la misma distribución límite, ésta será también el límite de la sucesión inicial.

Corolario 3.1. *Sea $\{\mu_n\}$ una sucesión de distribuciones acotadas en probabilidad, si cada subsucesión $\{\mu_{n_k}\}$ convergente lo hace a μ , entonces $\mu_n \Rightarrow \mu$.*

Demostración: Cada subsucesión $\{\mu_{n_k}\}$ es una sucesión de distribuciones acotadas en probabilidad, entonces por el Teorema 3.3 tendrá a su vez una subsucesión $\{\mu_{n_{k(j)}}\}$ convergente a algún límite cuando j tiende a ∞ , y dicho límite debe ser μ por hipótesis.

Supongamos que $\mu_n \Rightarrow \mu$ es falso. Entonces existirá un $x \in \mathbb{R}$ tal que $\mu\{x\} = 0$ que haga que $|\mu_n(-\infty, x]|$ no converja a $\mu(-\infty, x]$. Pero entonces existirá un $\varepsilon > 0$ para el que cualquier sucesión infinita de enteros $\{n_k\}$ cumpla que $|\mu_{n_k}(-\infty, x] - \mu(-\infty, x]| \geq \varepsilon$. Es decir, no existiría una subsucesión convergente a μ , algo que contradice la hipótesis inicial y permite concluir que $\mu_n \Rightarrow \mu$. $\square$

Finalmente, el siguiente teorema relaciona los conceptos de función característica y las sucesiones de distribuciones, aportando una condición suficiente para que la acotación en probabilidad se produzca.

Teorema 3.4. *Sea $\{\mu_n\}$ una sucesión de distribuciones, si existe una función continua en cero g tal que $\lim_n \varphi_n(t) = g(t)$ para todo t siendo φ_n la función característica de μ_n , entonces $\{\mu_n\}$ está acotada en probabilidad.*

Demostración: Para $u > 0$, por el Teorema de Fubini³,

$$\begin{aligned} \frac{1}{u} \int_{-u}^u (1 - \varphi_n(t)) dt &= \frac{1}{u} \int_{-u}^u \int_{-\infty}^{\infty} \mu_n(dx) dt - \frac{1}{u} \int_{-u}^u \int_{-\infty}^{\infty} e^{itx} \mu_n(dx) dt = \\ &= \int_{-\infty}^{\infty} \left[\frac{1}{u} \int_{-u}^u (1 - e^{itx}) dt \right] \mu_n(dx) = 2 \int_{-\infty}^{\infty} \left(1 - \frac{\sin(ux)}{ux} \right) \mu_n(dx) \geq \\ &\geq 2 \int_{|x| \geq 2/u} \left(1 - \frac{1}{|ux|} \right) \mu_n(dx) \geq \mu_n(\{x : |x| \geq 2/u\}). \end{aligned}$$

Como g es continua en el cero y $g(0) = 1$ por ser límite de la sucesión de funciones características, para cada $\varepsilon > 0$ existe $u > 0$ tal que $\frac{1}{u} \int_{-u}^u (1 - g(t)) dt < \varepsilon$. Como φ_n converge a g , el teorema de convergencia dominada (véase Teorema 2.4.) implica que existe un n_0 tal que si $n \geq n_0$, entonces $\frac{1}{u} \int_{-u}^u (1 - \varphi_n(t)) dt < 2\varepsilon$. Si $a = 2/u$ se tiene que $\mu_n(\{x : |x| \geq a\}) \leq 2\varepsilon$ para los n mayores o iguales a n_0 . Incrementando el valor de a lo que sea necesario, se cumplirá la condición para el número finito de valores menores que n_0 . Por lo tanto, queda demostrado que $\{\mu_n\}$ está acotado en probabilidad. $\square$

Ahora sí, enunciar y demostrar el Teorema de Continuidad será sencillo a partir de estos resultados.

Teorema 3.5. *(Teorema de Continuidad) Sean $\{\mu_n\}$ y μ distribuciones de probabilidad con funciones características φ_n y φ respectivamente, $\mu_n \Rightarrow \mu$ equivale a que $\lim_n \varphi_n(t) = \varphi(t)$ para todo t .*

Demostración: “ $\Rightarrow$ ” Para cada t se cumple que la función e^{itx} tiene módulo acotado y es continua en x , por lo tanto, por consecuencia inmediata del teorema de caracterización de la convergencia en distribución (véase Teorema 2.6), convergerán las funciones características.

“ $\Leftarrow$ ” Suponiendo que $\varphi_n(t) \rightarrow \varphi(t)$ para todo t , por el Teorema 3.4. la sucesión de distribuciones $\{\mu_n\}$ está acotada en probabilidad, por lo tanto existirá una subsucesión de distribuciones $\{\mu_{n_k}\}$ y una distribución ν tales que $\mu_{n_k} \Rightarrow \nu$. Por la primera implicación de este teorema, que esto suceda implica que ν tiene por función característica $\lim_k \varphi_{n_k}(t) = \varphi(t)$. Finalmente, el Teorema de Unicidad garantiza que $\mu = \nu$. $\square$

³Este teorema fundamental del cálculo integral garantiza que, bajo condiciones poco restrictivas, el cálculo de una integral doble se puede realizar integrando primero respecto a x y luego respecto a y o viceversa. Su enunciado y demostración se puede ver en la Sección 18 de Billingsley (1995).

3.3. Teorema Central del Límite.

Finalmente, ya demostrado el *Teorema de Continuidad*, que justifica el uso de la función característica en lugar de la función de distribución, y habiendo sido mencionadas todas las herramientas matemáticas presentes en el Teorema Central del Límite para el caso de la binomial, se procederá en esta sección a su demostración. Como se ha visto en la Sección 1.3 de este trabajo, este resultado es conocido con el nombre de *Teorema de de Moivre-Laplace* por las brillantes aportaciones de estos matemáticos franceses. Sin embargo, sus demostraciones no fueron del todo formales y rigurosas. El desarrollo de la función característica a finales del Siglo XIX simplifica enormemente su prueba.

Únicamente faltaría por mencionar y demostrar el siguiente lema que se usará en la parte final de la demostración del TCL:

Lema 3.1. Sean $z_1, \dots, z_m, z'_1, \dots, z'_m$ números complejos de a lo sumo módulo 1, entonces se cumple

$$|z_1 \cdots z_m - z'_1 \cdots z'_m| \leq \sum_{i=1}^m |z_i - z'_i|.$$

Demostración. La prueba del lema se llevará a cabo por inducción. Veamos que se cumple para $m = 2$.

$$|z_1 \cdot z_2 - z'_1 \cdot z'_2| = |(z_1 - z'_1)z_2 + z'_1(z_2 - z'_2)| \leq |z_1 - z'_1||z_2| + |z'_1||z_2 - z'_2| \leq |z_1 - z'_1| + |z_2 - z'_2|.$$

Ahora se supone cierto para $m - 1$, es decir, $|z_1 \cdots z_{m-1} - z'_1 \cdots z'_{m-1}| \leq \sum_{i=1}^{m-1} |z_i - z'_i|$ y se prueba para m :

$$\begin{aligned} |z_1 \cdots z_m - z'_1 \cdots z'_m| &= |z_m(z_1 \cdots z_{m-1} - z'_1 \cdots z'_{m-1}) + (z_m - z'_m)(z'_1 \cdots z'_{m-1})| \leq \\ &\leq |z_m|(|z_1 \cdots z_{m-1} - z'_1 \cdots z'_{m-1}|) + |(z_m - z'_m)|(|z'_1 \cdots z'_{m-1}|) \leq \\ &\leq |z_1 \cdots z_{m-1} - z'_1 \cdots z'_{m-1}| + |z_m - z'_m| \leq \sum_{i=1}^m |z_i - z'_i|. \end{aligned}$$

Las desigualdades se llevan a cabo por la desigualdad triangular y por la hipótesis de que los valores complejos tienen módulo de a lo sumo 1. $\square$

Teorema 3.6. Sea $X_n \sim \text{Bin}(n, p)$ y Z una variable aleatoria normal estándar, se cumple:

$$Z_n = \frac{X_n - np}{\sqrt{npq}} \xrightarrow{d} Z.$$

Demostración: La idea de la prueba es la siguiente: sea φ_n la función característica de Z_n y φ la función característica de la normal estándar, veremos que la primera converge a la segunda cuando n tiende a infinito.

Recordar que la función característica de la normal estándar Z es $\varphi(t) = e^{-t^2/2}$ y la de la binomial $X_n \sim \text{Bin}(n, p)$ es $\varphi_{X_n}(t) = (q + pe^{it})^n$ (véanse las ecuaciones (3.3) y (3.5), respectivamente).

Obtener la función característica de Z_n es sencillo mediante el uso de (3.6), con $a = \frac{1}{\sqrt{npq}}$ y $b = \frac{-np}{\sqrt{npq}}$:

$$\varphi_n(t) = e^{-itnp/\sqrt{npq}}(q + pe^{it/\sqrt{npq}})^n = (qe^{-itp/\sqrt{npq}} + pe^{itq/\sqrt{npq}})^n$$

Faltaría probar que $\varphi_n(t) = (qe^{-itp/\sqrt{npq}} + pe^{itq/\sqrt{npq}})^n \rightarrow e^{-t^2/2} = \varphi(t)$. Para ello, teniendo en cuenta que $e^{ix} = \cos(x) + i\sin(x)$ y aplicando el teorema de Taylor⁴ para las funciones coseno y seno en $x_0 = 0$, se obtienen las siguientes expresiones con $\sigma_n = \sqrt{npq}$:

$$qe^{-itp/\sigma_n} = q \left(1 - \frac{itp}{\sigma_n} - \frac{t^2 p^2}{\sigma_n^2 2!} + \frac{it^3 p^3}{\sigma_n^3 3!} + Y_{1_n} \right), \quad (3.8)$$

$$pe^{itq/\sigma_n} = p \left(1 + \frac{itq}{\sigma_n} - \frac{t^2 q^2}{\sigma_n^2 2!} - \frac{it^3 q^3}{\sigma_n^3 3!} + Y_{2_n} \right), \quad (3.9)$$

para $Y_{1_n} = \frac{t^4 p^4}{\sigma_n^4 4!} \cos(C_1(n)) - \frac{it^5 p^5}{\sigma_n^5 5!} \sin(C_2(n))$ e $Y_{2_n} = \frac{t^4 q^4}{\sigma_n^4 4!} \cos(C_3(n)) + \frac{it^5 p^5}{\sigma_n^5 5!} \sin(C_4(n))$, donde $C_1(n)$, $C_2(n)$ representan valores comprendidos entre 0 y pt/σ_n ; y $C_3(n)$, $C_4(n)$ son valores entre 0 y qt/σ_n , tendiendo todos ellos a 0 cuando n tiende a infinito.

Ahora bien, considerando $Y_n = \frac{it^3}{3!\sigma_n}(p^2 - q^2) + nqY_{1_n} + npY_{2_n}$, es obvio que $Y_n \rightarrow 0$ y, por lo tanto, $\frac{1}{n}Y_n = o\left(\frac{1}{n}\right)$.

Sumando las expresiones (3.8) y (3.9) se obtiene el siguiente valor para φ_n

$$\varphi_n(t) = \left(1 - \frac{t^2 pq}{2\sigma_n^2} + \frac{it^3 pq}{3!\sigma_n^3}(p^2 - q^2) + qY_{1_n} + pY_{2_n} \right)^n = \left(1 - \frac{t^2}{2n} + o\left(\frac{1}{n}\right) \right)^n.$$

Finalmente, como $qe^{-itp/\sqrt{npq}} + pe^{itq/\sqrt{npq}}$ y $1 - t^2/2n$ tienen módulo a lo sumo 1 si n es suficientemente grande, entonces podremos hacer uso del Lema 3.1 y concluir $|\varphi_n(t) - (1 - t^2/2n)^n| \leq n|(qe^{-itp/\sqrt{npq}} + pe^{itq/\sqrt{npq}}) - (1 - t^2/2n)| = no(1/n) = o(1) \rightarrow 0$. Es decir, $\lim_{n \rightarrow \infty} \varphi_n(t) = \lim_{n \rightarrow \infty} (1 - t^2/2n)^n = e^{-t^2/2}$ con lo que quedaría probado el Teorema Central del Límite para el caso de la binomial.

⁴Este teorema dice que siendo $n \in \mathbb{N}$, el intervalo $I = [a, b]$, $x_0 \in I$ y la función $f : I \rightarrow \mathbb{R}$ tal que f y sus derivadas $f', f'', \dots, f^{(n)}$ son continuas en I y que $f^{(n+1)}$ existe en (a, b) , entonces para cualquier $x \in I$ existe un punto c entre x y x_0 tal que

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \dots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n + \frac{f^{(n+1)}(c)}{(n+1)!}(x - x_0)^{n+1}.$$

Su demostración se puede ver en la Sección 6.4 de Bartle y Sherbert (2004).

Capítulo 4

Aplicaciones del Teorema Central del Límite.

La importancia del TCL en la estadística es enorme. Su versión clásica (véase Teorema 0.1) indica que si se toma una muestra aleatoria lo suficientemente grande, la distribución de la media de dicha muestra se puede aproximar mediante la distribución normal. Además, se pueden obtener las distribuciones asintóticas de otras distribuciones tales como la Poisson, la chi cuadrado o la binomial.

Centrándonos en la versión para la binomial, en este último capítulo se verán un par de aplicaciones que tiene el teorema en la inferencia estadística. En la Sección 4.1 se definirá la distribución empírica, esencial para estimar la distribución de una variable a partir de una muestra aleatoria. Además, se expondrá el procedimiento que habrá que llevar a cabo para encontrar una banda de confianza para la función de distribución desconocida. Ya en la Sección 4.2, calcularemos un intervalo de confianza para la probabilidad de nacer hombre en Londres entre los años 1629 y 1710 a partir de los datos de los que disponía Arbuthnott.

4.1. Distribución empírica.

A la hora estudiar un fenómeno, es necesario realizar observaciones a fin de poder sacar conclusiones de ellas. Esto se traduce en coger una muestra aleatoria de una variable. Lo habitual es que la distribución de esta variable sea desconocida, siendo necesario obtener una aproximación de ella. Esto es lo que hace la distribución empírica, también conocida como distribución muestral.

Como hemos visto en el Capítulo 2, la distribución de una variable aleatoria es caracterizada por su función de distribución. Supongamos una variable X con función de distribución desconocida F , que recordemos que se define para todo $x \in \mathbb{R}$ como $F(x) = \mathbb{P}(X \leq x)$. Si tomamos una muestra aleatoria $X_1, \dots, X_n$ de tamaño n de X , se define la función de

distribución empírica como

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{(-\infty, x]}(X_i). \quad (4.1)$$

Esta función asigna a cada x real la proporción de los valores observados que son menores o iguales que x .

Es inmediato comprobar que la función F_n es una función de distribución: no decreciente, continua por la derecha, con límites cuando tiende a menos infinito y a más infinito iguales a cero y uno, respectivamente. Además, es una distribución discreta que otorga la misma probabilidad $1/n$ a cada observación.

Como ya sabemos, cada término $Y_i = \mathbb{I}_{(-\infty, x]}(X_i)$ es una variable aleatoria de Bernoulli con probabilidad de éxito $p = \mathbb{P}(X_i \leq x) = F(x)$. Por lo tanto, $nF_n(x)$ tiene distribución binomial con parámetros n y $p = F(x)$. Aplicando el TCL, esto equivale a decir que

$$\frac{\sqrt{n}(F_n(x) - F(x))}{\sqrt{F(x)(1 - F(x))}} \xrightarrow{d} N(0, 1).$$

O lo que es lo mismo, F_n se distribuye asintóticamente como una $N(F(x), F(x)(1 - F(x))/n)$.

Por lo tanto, con un tamaño de muestra n grande, es muy probable que el valor de $F_n(x)$ sea muy próximo al de $F(x)$. Si este proceso de muestreo se prolonga indefinidamente, $\{Y_i\}_{i \in \mathbb{N}}$ es una sucesión de variable aleatorias idénticamente distribuidas con media $F(x)$, y por la ley fuerte de los grandes números se cumple que cuando $n \rightarrow \infty$,

$$\frac{1}{n} \sum_{i=1}^n Y_i = F_n(x) \rightarrow F(x) \text{ casi seguro.} \quad (4.2)$$

De este modo, para cada $x \in \mathbb{R}$ fijado, se garantiza que es posible conocer a través de $F_n(x)$ el valor de $F(x)$, pero lo interesante será ver que esto ocurre simultáneamente para todos los valores de x , algo que no es inmediato ya que el número de posibles valores de x no es numerable. El *Teorema de Glivenko-Cantelli* prueba que $F_n(x)$ converge a $F(x)$ uniformemente en x con probabilidad uno, y su demostración se puede ver en el Capítulo 2 de Vélez y García (1993).

Teorema 4.1. (*Teorema de Glivenko-Cantelli*) Sea $\{X_i\}_{i \in \mathbb{N}}$ una sucesión de variables aleatorias independientes con distribución común F , si F_n es la función de distribución empírica de los primeros n valores muestrales y sea $D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|$, se cumple

$$\lim_{n \rightarrow \infty} D_n = 0 \text{ casi seguro.}$$

En la Figura 4.1 se observa claramente que cuanto mayor es el número de observaciones, más se parecerá la gráfica de las funciones empíricas a la función de distribución teórica.

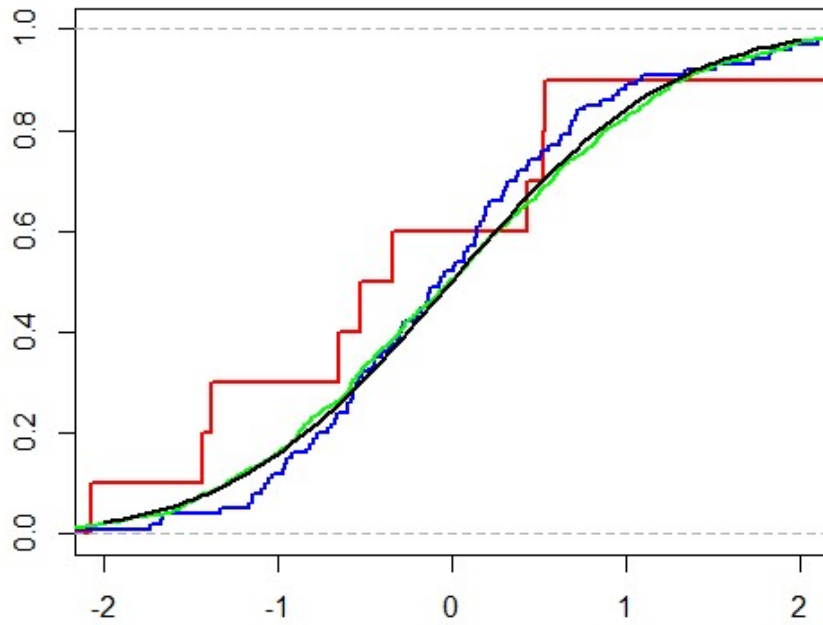


Figura 4.1: Comparación de la función de distribución normal estándar (negro) con funciones de distribución empíricas de 10 (rojo), 100 (azul) y 1000 (verde) observaciones.

Sin embargo, aunque el Teorema de Glivenko-Cantelli garantiza la convergencia uniforme de F_n , no especifica a qué velocidad lo hace. Esta cuestión se aborda en el siguiente teorema, conocida como desigualdad de Dvoretzky-Kiefer-Wolfowitz (DKW), cuya demostración se puede ver en Massart (1990):

Teorema 4.2. (*Desigualdad de DKW*) Sea $X_1, \dots, X_n$ una muestra aleatoria de una variable con función de distribución desconocida F , para todo $\varepsilon > 0$ se cumple

$$\mathbb{P}(D_n > \varepsilon) \leq 2e^{-2n\varepsilon^2}. \quad (4.3)$$

La desigualdad de DKW se puede usar para construir una banda de confianza para F . Igualando la parte de la derecha de (4.3) a α se obtiene que si $\varepsilon_n^2 = \frac{1}{2n} \ln(\frac{2}{\alpha})$, entonces la función de distribución F estará comprendida entre $F_n - \varepsilon_n$ y $F_n + \varepsilon_n$. Por lo tanto, si se definen

$$\begin{aligned} L(x) &= \max\{F_n(x) - \varepsilon_n, 0\}, \\ U(x) &= \min\{F_n(x) + \varepsilon_n, 1\}, \end{aligned}$$

se cumplirá que

$$\mathbb{P}(L(x) \leq F(x) \leq U(x), \forall x \in \mathbb{R}) \geq 1 - \alpha.$$

En la Figura 4.2 se comprueba que a partir de una muestra de 100 observaciones de una normal estándar y tras el cálculo de la función de distribución empírica, las bandas de confianza al 95 % contienen a la función de distribución.

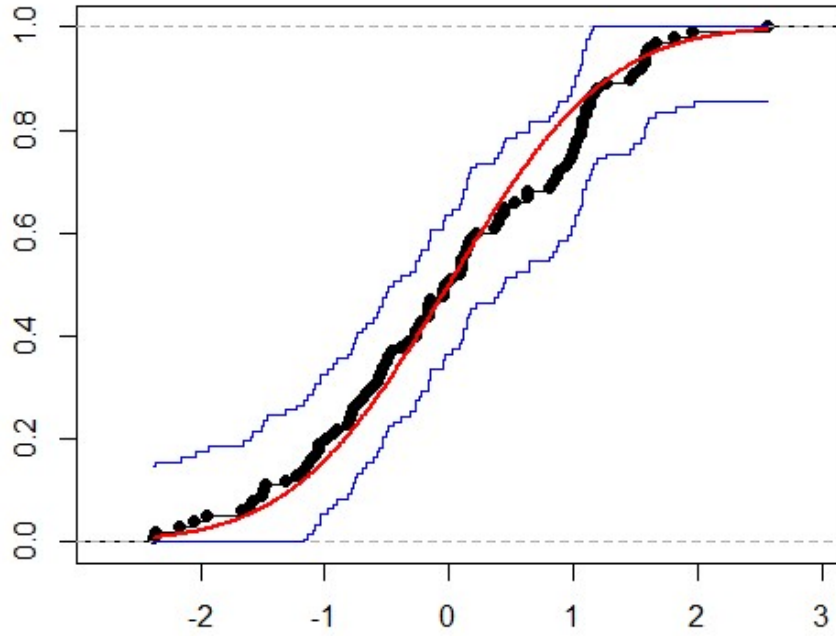


Figura 4.2: Estimación de las bandas de confianza al 95 %, en azul, de la función de distribución de una normal estándar, en rojo, a partir de 100 observaciones, en negro.

4.2. Intervalo de confianza de una proporción.

El TCL también es la justificación teórica de la aproximación binomial en la inferencia sobre proporciones. Si queremos saber la probabilidad con la que ocurre un determinado suceso, es lógico pensar que debemos realizar varios intentos independientes con el fin de observar la proporción de éxitos. Esta proporción muestral $\hat{p}$ no es más que dividir el valor de la binomial entre el número de pruebas realizadas. Decimos por tanto que $\hat{p}$ es el estimador de la probabilidad teórica p .

Es sencillo comprobar que la esperanza de $\hat{p}$ es p y que su varianza es pq/n a partir de las propiedades vistas en el Capítulo 2, lo que nos permite afirmar gracias al TCL que

$$\frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \xrightarrow{d} N(0, 1). \quad (4.4)$$

Más que la estimación puntual del parámetro p , lo interesante es ver una región en la que se encuentre con una probabilidad alta. Lo que buscaremos serán valores L y U para los que la probabilidad de que p esté entre ellos sea $1 - \alpha$, con $\alpha \in [0, 1]$. Esto se conoce como intervalo de confianza de nivel de confianza $1 - \alpha$.

Se denota $z_{\alpha/2}$ al número real tal que $\mathbb{P}(Z > z_{\alpha/2}) = \alpha/2$, siendo $Z \stackrel{d}{=} N(0, 1)$ ¹. Entonces,

¹Este valor se puede obtener en las tablas de la distribución normal estándar o con la función `qnorm` del software estadístico R

se cumplirá

$$1 - \alpha = \lim_n \mathbb{P} \left(\frac{|\hat{p} - p|}{\sqrt{\frac{pq}{n}}} \leq z_{\alpha/2} \right) = \lim_n \mathbb{P} \left(\hat{p} - z_{\alpha/2} \sqrt{\frac{pq}{n}} \leq p \leq \hat{p} + z_{\alpha/2} \sqrt{\frac{pq}{n}} \right).$$

El intervalo de confianza para p con nivel de confianza $1 - \alpha$ estará centrado en $\hat{p}$ y tendrá radio $z_{\alpha/2} \sqrt{\frac{pq}{n}}$. Sin embargo, el error típico $\sqrt{\frac{pq}{n}}$ no se conoce, pues depende del parámetro desconocido p . El Lema de Slutsky que enunciaremos a continuación y cuya demostración se puede ver en la Sección A.14 de Bickel y Doksum (1977), además de la ley fuerte de los grandes números permiten solucionar este problema.

Lema 4.1. (*Lema de Slutsky*) Sean $\{X_n\}$ e $\{Y_n\}$ sucesiones de variables aleatorias, sea X una variable aleatoria y c una constante. Si $X_n \xrightarrow{d} X$ e $Y_n \xrightarrow{p} c$, entonces

$$X_n Y_n \xrightarrow{d} cX.$$

La ley fuerte de los grandes números garantiza la convergencia casi segura de $\hat{p}$ a p , por lo tanto, supuesto $0 < \hat{p} < 1$,

$$\frac{\sqrt{pq}}{\sqrt{\hat{p}\hat{q}}} \xrightarrow{p} 1. \quad (4.5)$$

La aplicación del Lema de Slutsky a (4.4) y (4.5) da como resultado que

$$\frac{\hat{p} - p}{\sqrt{\frac{\hat{p}\hat{q}}{n}}} \xrightarrow{d} N(0, 1), \quad (4.6)$$

lo que permite tomar como intervalo de confianza $1 - \alpha$ a:

$$IC = \left(\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}, \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}} \right). \quad (4.7)$$

Veamos todo esto con el ejemplo de los nacimientos en Londres que nombramos en el primer capítulo. Tomaremos los datos de los que disponía Arbuthnott y que se exponen en la Tabla 4.1 de la Página 39. Debemos suponer que la probabilidad de nacer hombre es igual a lo largo de estos 82 años, supuesto que esta probabilidad se haya mantenido constante a lo largo de este periodo.

El total de bautizos realizados en este tiempo es de 938223, siendo 484382 del sexo masculino, un 51.6276 % del total. Con estos datos, la estimación del error típico $\sqrt{\frac{\hat{p}\hat{q}}{n}}$ de la proporción muestral es 0.000515925. Para un intervalo de confianza de p del 99 %, necesitamos saber el valor $z_{0,005}$ tal que $\mathbb{P}(Z < z_{0,005}) = 0,995$, que será 2.576. Solo faltaría sustituir en (4.7) para obtener el intervalo de confianza (0,51495, 0,51761).

En este caso, al tener una muestra enorme, de casi un millón de datos, la longitud del intervalo es notablemente pequeña, pues se reduce a medida que aumenta el número de

observaciones. Incluso aumentando el nivel de confianza mucho más de lo habitual, por ejemplo, al 99.999 %, el valor de $z_{\alpha/2}$ será 4.417, resultando en el intervalo de confianza (0.5140, 0.5186). Lo que permite asegurar con total seguridad que $p > 0,51$. De hecho, si quisiéramos encontrar un nivel de confianza con el que 0.51 esté incluido en el intervalo de confianza, sería necesario que $z_{\alpha/2} > 12,165$, valor para el cual $\mathbb{P}(Z > z_{\alpha/2}) < 10^{-20}$.

Como vemos en este ejemplo, el TCL es una herramienta muy útil que facilita enormemente los cálculos, pues la alternativa de calcular las probabilidades con una muestra tan grande es extremadamente laboriosa, incluso con las herramientas computacionales actuales. Sin embargo, hay situaciones en las que el uso del teorema nos lleva a resultados imposibles.

Esto ocurre, por ejemplo, cuando tenemos una proporción muestral muy cercana al 0 o al 1, pudiendo obtener intervalos de confianza que no estén contenidos en $[0, 1]$. Supongamos que queremos saber la probabilidad de fabricar un coche defectuoso en una cadena de montaje. Para ello observamos que al producir 100 vehículos, solamente uno ha resultado defectuoso. Si quisiéramos obtener un intervalo de confianza del 95 % para la probabilidad², obtendríamos el intervalo (-0.0095, 0.0295).

Como posible solución a este problema, se buscará un valor U para el cual p sea menor con una probabilidad del 95 %, es decir, $\mathbb{P}(p < U) = 0,95$. Desarrollando esta expresión y aplicando (4.6) se tiene que

$$\mathbb{P}\left(\frac{\hat{p} - p}{\sqrt{\frac{\hat{p}\hat{q}}{n}}} > \frac{\hat{p} - U}{\sqrt{\frac{\hat{p}\hat{q}}{n}}}\right) = \lim_{n \rightarrow \infty} \mathbb{P}\left(\frac{\hat{p} - p}{\sqrt{\frac{\hat{p}\hat{q}}{n}}} > z_{0,05}\right) = 0,95.$$

Por lo tanto, para n grande, el valor de U puede aproximarse por $U = \hat{p} - z_{0,05}\sqrt{\frac{\hat{p}\hat{q}}{n}} = 0,026$ y se concluye que $[0, 0.026]$ será el intervalo de confianza asintótico de p al 95 %.

²En este caso, $z_{\alpha/2} = z_{0,025} = 1,96$

Año	Hombre	Mujer	Año	Hombre	Mujer
1629	5218	4683	1670	6278	5719
1630	4858	4457	1671	6449	6061
1631	4422	4102	1672	6443	6120
1632	4994	4590	1673	6073	5822
1633	5158	4839	1674	6113	5738
1634	5035	4820	1675	6058	5717
1635	5106	4928	1676	6552	5847
1636	4917	4605	1677	6423	6203
1637	4703	4457	1678	6568	6033
1638	5359	4952	1679	6247	6041
1639	5366	4784	1680	6548	6299
1640	5518	5332	1681	6822	6533
1641	5470	5200	1682	6909	6744
1642	5460	4910	1683	7577	7158
1643	4793	4617	1684	7575	7127
1644	4107	3997	1685	7484	7246
1645	4047	3919	1686	7575	7119
1646	3768	3395	1687	7737	7214
1647	3796	3536	1688	7487	7101
1648	3363	3181	1689	7604	7167
1649	3079	2746	1690	7909	7302
1650	2890	2722	1691	7662	7392
1651	3231	2840	1692	7602	7316
1652	3220	2908	1693	7676	7483
1653	3196	2959	1694	6985	6647
1654	3441	3179	1695	7263	6713
1655	3655	3349	1696	7632	7229
1656	3668	3382	1697	8062	7767
1657	3396	3289	1698	8426	7626
1658	3157	3013	1699	7911	7452
1659	3209	2781	1700	7578	7061
1660	3724	3247	1701	8102	7514
1661	4748	4107	1702	8031	7656
1662	5216	4803	1703	7765	7683
1663	5411	4881	1704	6113	5738
1664	6041	5681	1705	8366	7779
1665	5114	4858	1706	7952	7417
1666	4678	4319	1707	8379	7687
1667	5616	5322	1708	8239	7623
1668	6073	5560	1709	7840	7380
1669	6506	5829	1710	7640	7288

Tabla 4.1: Bautizos en Londres entre los años 1629 y 1710 (tabla recogida de Hald (2003), Capítulo 17).

Bibliografía.

- R. G. Bartle, D. R. Sherbert (2004): *Introducción al análisis matemático de una variable*. John Wiley & Sons, Inc.
- D. Bernoulli (1738): *Exposition of a New Theory on the Measurement of Risk*, *Econometrica*, **vol 22**, pp. 23-36.
- N. Bernoulli (1687): *Retrato de Jakob Bernoulli*. [Figura]. Recuperado el 18-06-2022 de https://commons.wikimedia.org/wiki/File:Jakob_Bernoulli.jpg
- P. J. Bickel, K. A. Doksum (1977): *Mathematical statistics : basic ideas and selected topics*. Prentice Hall.
- P. Billingsley (1995): *Probability and measure*, A Wiley-Interscience Publication, John Wiley & Sons, New York.
- Desconocido (1736): *Retrato de Abraham de Moivre*. [Figura]. Recuperado el 18-06-2022 de https://commons.wikimedia.org/wiki/File:Abraham_de_moivre.jpg
- H. Fischer (2010): *A history of the central limit theorem. From clasical to modern probability theory*, Springer Science & Business Media.
- J. B. P. Guérin (1838): *Retrato de Pierre Simon de Laplace*. [Figura]. Recuperado el 18-06- 2022 de [https://commons.wikimedia.org/wiki/File:Pierre-Simon,_marquis_de_Laplace_\(1745-1827\)_-_Guerin.jpg](https://commons.wikimedia.org/wiki/File:Pierre-Simon,_marquis_de_Laplace_(1745-1827)_-_Guerin.jpg)
- A. Hald (2003): *A history of probability and statistics and their applications before 1750*. John Wiley & Sons, Inc.
- A. H. Landro, M. L. González (2013) *Bernoulli, de Moivre, Bayes, Price y los fundamentos de la inferencia inductiva.. Cuadernos del CIMBAGE*, 33-56. Recuperado el 21-01-2022 de <https://www.redalyc.org/articulo.oa?id=46226414003>
- P. Massart (1990). *The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality*. *The Annals of Probability*, **vol 18**, pp. 1269-1283.
- S. Stahl (2006): *The evolution of the normal distribution*. *Mathematics magazine*, **vol 79**, pp. 96-113.
- S. M. Stigler (1986): *The history of statistics: The measurement of uncertainty before 1900*. Harvard University Press.

- R. Vélez y A. García (1993): *Principios de Inferencia Estadística*. UNED.
- E. W. Weisstein: *Gaussian Integral*. MathWorld—A Wolfram Web Resource. Disponible en: <https://mathworld.wolfram.com/GaussianIntegral.html>