

Teoría de Normalidad

Martínez M., Denisse A. Román G., Luis M.
000159780 000117077

Salgado F., Lisett P.
000160314

31 de octubre de 2015

Resumen

En este documento se examinan ciertos temas centrales a la teoría de normalidad. Se exploran algunos resultados teóricos como el teorema central del límite clásico y algunas de sus generalizaciones. Se investigan algunas pruebas de normalidad, su interpretación y sus aplicaciones. Se deduce el teorema de Chebyshev y se dan algunos ejemplos prácticos.

Índice

1. Introducción	4
2. Teorema Central del Límite	9
2.1. TCL Clásico	9
2.2. TCL Variables Dependientes	12
2.3. TCL Variables No Identicamente Distribuidas	12
2.4. TCL Segundo Momento Infinito	13
3. Pruebas de Normalidad	14
3.1. Criterios de Calidad de Pruebas	14
3.2. La Prueba de Normalidad Jarque-Bera	15
3.3. La Prueba de Normalidad D'Agostino	16
3.4. La Prueba de Normalidad χ^2	16
4. Teorema de Chebyshev	17

5. Cotas	21
6. Ejemplos	22

1. Introducción

Antes de comenzar con la exposición de las distintas formas y generalizaciones del Teorema Central del Límite, es conveniente llevar a cabo un breve análisis de las diferentes contribuciones históricas que lo llevaron a su actual concepción.

Hoy en día, el Teorema Central del Límite **TCL** está asociado con toda una serie de resultados que involucran la convergencia débil ¹ de distribuciones de probabilidad formadas con variables aleatorias de una o múltiples dimensiones ² a una distribución normal. Para evitar ambigüedad, en este trabajo nos referiremos al resultado *clásico* que trata el comportamiento asintótico de la suma de un conjunto arbitrario numerable de variables aleatorias independientes o débilmente correlacionadas.

Es importante notar que el **TCL** no es el trabajo de un solo hombre. La concepción actual del mismo, se debe a las contribuciones y al trabajo conjunto de distintos matemáticos a lo largo de casi un siglo. En efecto, la primera vez que se nombró a este resultado como el Teorema Central del Límite fue en 1920 en un artículo de George Pòlya en donde menciona que

*Es generalmente conocido que la aparición de la distribución de probabilidad Gaussiana $\frac{1}{\sqrt{\pi}}e^{-x^2}$ puede ser explicada por el mismo **teorema del límite** que juega un papel **central** en la teoría de la probabilidad.*

¹en distribución

²Se podría hablar de objetos aleatorios mucho más generales con valores en espacios generales de Banach.

No obstante, este resultado comenzó a tomar forma más de cien años antes³ con los trabajos de Pierre-Simon de Laplace. Concretamente, en 1812 con la primera edición de *Théorie analytique des probabilités* fue que Laplace dio su versión más general del **TCL**.

Teorema 1.1. Sean $\epsilon_1, \epsilon_2, \dots, \epsilon_n$ un número **grande** de errores de observaciones independientes. Todos con una misma distribución de media μ y varianza σ^2 . Si $\lambda_1, \lambda_2, \dots, \lambda_n$ son multiplicadores constantes y $a > 0$ entonces

$$P\left(\left\|\sum_{j=1}^n \lambda_j(\epsilon_j - \mu)\right\|_1 \leq a \sqrt{\sum_{j=1}^n \lambda_j^2}\right) \approx \frac{2}{\sigma\sqrt{2\pi}} \int_0^a e^{-\frac{x^2}{2\sigma^2}} dx$$

A Laplace le tomó aproximadamente 40 años de trabajo llegar a esta forma del **TCL**. Este camino comenzó en 1774 con su *Mémoire sur la probabilité des causes*. En este trabajo se comenzó a preguntar sobre métodos adecuados para aproximar integrales de funciones que dependen de parámetros con un gran valor.⁴

A pesar de que este resultado parece bastante general, el método utilizado por Laplace no le permitía obtener cotas absolutas o relativas sobre el error de aproximación y la confiabilidad de este método dependía completamente de la calidad de las aproximaciones generadas por series de potencias. Era tal la confianza de Laplace en su método que en 1814 escribió en la introducción de su *Théorie analytique*

³Incluso más si consideramos las contribuciones de Moivre

⁴Ejemplo $\Gamma(s+1) = \int_0^\infty e^{-x} x^s dx$ con $s \gg 0$. La idea básica es la siguiente: Si el integrando $f(x)$ depende de un gran parámetro tal que la función f tiene un único *pico* de tal forma que la mayor parte del valor de la integral se concentra en una vecindad ϵ alrededor de este máximo, entonces se espera que asintóticamente esta función se comporte de manera similar a una función de la forma $f(a)e^{\alpha(x-a)^{2k} \pm \dots}$ con $\alpha > 0$ si f alcanza su máximo en $x = a$.

Las series convergen más rápido entre más complicada es la función. Por lo que el método se vuelve más preciso cada vez que es más necesario.

No obstante la confianza que Laplace tenía en su método, muchos de sus contemporáneos lo consideraron poco riguroso. Más aún, una vez que Adrien Marie Legendre probó la divergencia de la serie:

$$E(n) = \sum_{k=1}^m \frac{B_{2k}}{2k(2k-1)n^{2k-1}}$$

con

$$B_m(n) = n^m - \sum_{k=0}^{m-1} \binom{m}{k} \frac{B_k(n)}{m-k+1}$$

$$B_0(n) = 1$$

Cómo esta serie y la utilizada por Laplace para aproximar el valor de la integral arriba mencionada son equivalentes, pareció evidente que el método de Laplace podría llegar a ser, a lo más *semi-convergente*.

El siguiente gran paso en dirección hacia la generalización del **TCL** lo dio Siméon Denis Poisson con su aportación del concepto de *variable aleatoria* ⁵.

Teorema 1.2. Sean X_1, X_2, \dots, X_n un **gran** número de variables aleatorias con funciones de densidad de probabilidad con decrecimiento **suficientemente rápido**⁶ mientras sus argumentos tienden a $\pm\infty$. Entonces para los valores $\rho_n(\alpha)$ de las funciones características de X_n existe una función $r(\alpha)$ independiente de n con $0 \leq r(\alpha) \leq 1$; $\forall \alpha \neq 0$ tal que

$$\rho_n(\alpha) \leq r(\alpha)$$

⁵Todavía muy lejano del concepto formal de variable aleatoria explicado por Kolmogorov en 1933.

⁶Poisson no especifica que tan rápido.

Entonces para cualquier γ, γ' se tiene que

$$P\left(\gamma \leq \frac{\sum_{n=1}^s (X_n - E(X_n))}{\sqrt{\sum_{n=1}^s Var(X_n)}} \leq \gamma'\right) \approx \frac{1}{\sqrt{\pi}} \int_{\gamma}^{\gamma'} e^{-u^2} du$$

Donde la aproximación es más precisa conforme s se vuelve más grande. Poisson enfatizó particularmente este punto pues representaba el paso más delicado de su prueba por la intervención de cantidades *infinitas* y su representación probabilística. En 1824 escribió:

Queremos considerar cantidades de s infinitamente grandes, tal que la fórmula sea rigurosamente cierta en su límite. Ahora, con respecto a la cantidad ρ , siendo menor que 1 si la variable $\alpha \neq 0$ se sigue que en el límite $s = \infty$ la potencia ρ^s alcanza valores finitos sólo para valores infinitamente pequeños de esta variable y se vuelve infinitamente pequeña si α tiene un valor finito.

Es notorio que a diferencia de Laplace, Poisson notó que sus versión del **TCL** requería supuestos adicionales (cómo la condición de que la función característica cumplía con la condición expuesta). El mismo dio el notable contraejemplo de funciones idénticamente distribuidas bajo la distribución:

$$f(x) = \frac{1}{\pi(1+x^2)}$$

Finalmente en 1853 Cauchy y Dirichlet de manera casi simultánea e inde-

pendiente le dieron forma a la primera versión equivalente a la actual del **TCL**. En aras de mantener la introducción lo más breve posible, sólo expondremos el resultado obtenido por el primero.

El resultado obtenido por Cauchy no sólo proporciona un límite asintótico sino que también proporciona cotas superiores para la velocidad de convergencia. Cauchy considera la distribución límite de combinaciones lineales $\sum_{j=1}^n \lambda_j \epsilon_j$ de errores ϵ_j independientes, idénticamente distribuidos bajo una distribución f tal que $\|f(x)\| \approx 0$ para valores de $x \geq |k|$. Adicionalmente, requirió que $\|\lambda_j\| \approx \frac{1}{n}$; $\forall j$ y que $\sum_{j=1}^{\infty} \lambda_j^2 = \Lambda$ donde $\|\Lambda\| \approx \frac{1}{n}$. Dadas estas condiciones, la versión del **TCL** de Cauchy dice que

$$\left\| P\left(-v \leq \sum_{j=1}^n \lambda_j \epsilon_j \leq v\right) - \frac{2}{\sqrt{\pi}} \int_0^{\frac{v}{2\sqrt{c\Lambda}}} e^{-\theta^2} d\theta \right\| \leq C_1(n) + C_2(n, v) + C_3(n)$$

donde $c = \int_0^k x^2 f(x) dx$.

Cauchy sólo esbozo la prueba de este resultado, y la deducción formal de los valores C_1, C_2 y C_3 fue establecida hasta 1892 por Ivan Vladislavovich Sleshinski. Si $\Theta = n^{\frac{1}{2}+\delta}$; $(0 < \delta < \frac{1}{4})$ entonces

$$C_1 = \frac{e^{-N}}{\pi N}; \quad N = \frac{r\Delta\Theta^2}{2(1+r\lambda^2\Theta^2)}$$

$$C_2(n, v) = \frac{2h\sqrt{3}}{\pi} \log \left(\frac{v\Theta}{\sqrt{3}} + \sqrt{1 + \frac{\Theta^2 v^2}{3}} \right)$$

$$C_3(n) = \frac{e^{c\Lambda\Theta^2}}{c\pi\Lambda\Theta^2}$$

No resultaría entonces una sorpresa que Cauchy, siendo Cauchy, escribiera en 1853:

El análisis por medio del cual él (Laplace) estableció las propiedades del método para el cual usó series de potencias cuya convergencia no fue probada. M. Cauchy lo ha reemplazado por una fórmula exacta y rigurosa.

2. Teorema Central del Límite

Como se mencionó en la introducción, dentro de la teoría de la probabilidad existen pocos resultados tan importantes como el Teorema Central del Límite **TCL**. En esta sección analizaremos y discutiremos diferentes representaciones de este importante resultado, exploraremos los supuestos sobre los cuales se sustentan y los relajaremos paulatinamente.

2.1. TCL Clásico

Teorema 2.1. Sea $(X_n)_{n=1}^{\infty}$ una sucesión de variables aleatorias independientes o débilmente correlacionadas,⁷ idénticamente distribuidas bajo una distribución F tal que $\int_{-\infty}^{\infty} X_i^2 dF < \infty$; $\forall i$ Si definimos a $(Z_n)_{n=0}^{\infty}$:

$$Z_n = \frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}}$$

⁷ $\forall \epsilon > 0 \exists N \in \mathbb{N}$ tal que $\|\rho(X_i, X_j)\| < \epsilon \quad \forall i, j \geq N$

Entonces tenemos que

$$\lim_{n \rightarrow \infty} \left| P(z_n \leq x) - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt \right| \rightarrow 0$$

Cómo siempre sucede con este tipo de resultados asintóticos, para fines prácticos, nos interesaría más bien cuantificar la velocidad o error de convergencia. Inmediatamente surgen dos interrogantes:

1. ¿Es posible caracterizar el error de convergencia en general?
2. ¿Cómo se comporta el error asintótico para ciertos casos de interés?

La primera interrogante encuentra una respuesta satisfactoria en el teorema de Berry–Esseen.

Teorema 2.2. *Usando los mismos supuestos y notación de (2.1), sean $E(X_i) = \mu$, $\sigma^2 = E(\|X_i - \mu\|^2)$ y $\rho = E(\|X_i - \mu\|^3)$ tenemos que:*

$$\left| P(z_n \leq x) - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt \right| \leq \frac{C\rho}{2\sigma^3\sqrt{n}}$$

Donde C es una constante independiente de n .

Es interesante notar que para acotar el error de aproximación basta tomar en consideración los tres primeros momentos centrales. Distintas mejoras se han hecho a esta cota a lo largo de los años.

- **Feller 1969** $C = 6$
- **Beek 1972** $C = 0.7975$
- **Shiganov 1986** $C = 0.7915$

Para responder a la segunda interrogante hay un gran número de resultados que envuelven distintas distribuciones particulares de las X_i . En específico son interesantes los resultados obtenidos para distribuciones Poisson, Beta, Binomial, Student-t y Gamma. Aquí profundizaremos en el caso de $X_i \sim \text{Binomial}(n, p)$. Por (2.2) sabemos que:

$$\begin{aligned} \left| P(z_n \leq x) - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt \right| &\leq \frac{C\rho}{2\sigma^3\sqrt{n}} \\ &= \frac{C(p^2 + (1-p)^2)}{\sqrt{np(1-p)}} \end{aligned}$$

Esta función alcanza su mínimo cuando $p = \frac{1}{2}$. En cualquier otro caso, supongamos sin pérdida de generalidad que $p < \frac{1}{2}$ ⁸. Entonces tenemos que $\frac{(p^2 + (1-p)^2)}{\sqrt{p(1-p)}} \leq \frac{1}{\sqrt{p}}$. Por ende una cota razonable para el error sería $\frac{0.7915}{\sqrt{np}}$.

Ahora, procedamos a explorar algunas generalizaciones del **TCL**. Cómo puede notarse en el enunciado del teorema, existen tres hipótesis que se pueden relajar:

1. Independencia de variables.
2. Variables idénticamente distribuidas.
3. Segundo momento finito.

⁸Si este no es el caso, hacer el cambio de variable $p' = 1 - p$.

2.2. TCL Variables Dependientes

Este resultado es principalmente el producto del trabajo de Liapounov, Feller, Liendeberg y Lévy. Antes de anunciarlo, es necesaria la siguiente definición.

Definición 2.1. Una sucesión $(X_n)_{n=1}^{\infty}$ de variables aleatorias es F intercambiable si para cualquier permutación $s(i)$ finita de sus índices, se cumple que

$$F_{x_1, x_2, \dots, x_k, \dots}(X_0) = F_{x_{s(1)}, x_{s(2)}, \dots, x_{s(k)}, \dots}(X_0); \quad \forall X_0$$

Ahora bien, el enunciado del teorema es como sigue

Teorema 2.3. Si $(X_n)_{n=1}^{\infty}$ son F intercambiables con F una distribución de probabilidad determinada de manera *única* por sus momentos $\alpha_k; k \geq 1$ y $E(X_1 \cdot X_2 \cdot X_3 \dots X_k) = \alpha_k \forall k \geq 1$. Entonces

$$\lim_{n \rightarrow \infty} \left| P(z_n \leq x) - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt \right| \rightarrow 0 \Leftrightarrow E(X_1 \cdot X_2 \cdot X_3 \dots X_k) = \alpha_k \forall k \geq 1$$

2.3. TCL Variables No Identicamente Distribuidas

Para esta sección, supongamos sin pérdida de generalidad que cada variable X_i pertenece a una distribución de probabilidad con media cero y varianza σ_n^2 . Definamos también $S_n = \sum_{i=1}^n X_i$ y $s_n^2 = \sum_{i=1}^n \sigma_i^2$. Entonces tenemos el siguiente resultado, debido también a Liapounov.

Teorema 2.4. Sea $(X_n)_{n=1}^{\infty}$ una sucesión de variables aleatorias distribuidas bajo $F_i, i = 1, 2, \dots$. Si se cumple que $\exists \delta > 0$ tal que $E(|X_i|^{2+\delta}) < \infty; \quad \forall i$ y además $\lim_{n \rightarrow \infty} s_n^{-(1+\frac{\delta}{2})} \sum_{1 \leq k \leq n} E(|X_n|^{2+\delta}) \rightarrow 0$. Entonces

$$\lim_{n \rightarrow \infty} \frac{S_n}{s_n} \rightarrow \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt$$

2.4. TCL Segundo Momento Infinito

En esta sección relajaremos la hipótesis de varianza finita y requeriremos que las variables sean independientes e idénticamente distribuidas. Adicionalmente, requeriremos que existan a, b tales que la distribución a la que converge la

sucesión $\sum_{i=1}^{\infty} (X_i - a)$ sea estable.

Definición 2.2. Sean X_0, X_1, X_2 variables aleatorias iid. Decimos que la distribución de estas variables es **estable** si $\forall a, b \in \mathbb{R}^+ \exists c \in \mathbb{R}^+, d \in \mathbb{R}$ tales que

$$F_{cX_0+d}(x) = F_{aX_1+bX_2}(x) \quad \forall x$$

Una de las grandes ventajas de las distribuciones estables es que estas se pueden especificar completamente por medio de tres parámetros. Uno de los parámetros es el exponente $0 \leq \alpha \leq 2$. Este parámetro controla el ancho de las colas de la distribución. Si $\alpha = 2$ tenemos que la distribución es normal. Por otro lado si $\alpha < 2$ la distribución de probabilidad es asintóticamente proporcional a $|x|^{-\alpha-1}$.

Todas las distribuciones estables tienen colas anchas (excepto para $\alpha = 2$) y por ende, segundo momento infinito. Un ejemplo de una distribución estable es la función **Cauchy**⁹.

La importancia de las distribuciones estables, se ve reflejada en el siguiente teorema debido a Gnedenko y Kolmogorov.

Teorema 2.5. Si $(X_n)_{i=1}^{\infty}$ es una sucesión de variables aleatorias tales que si tomamos $S_n = \sum_{i=1}^n X_i$ tenemos que S_n sigue una distribución cuyas colas

⁹ $\frac{1}{\pi(1+x^2)}$

decrecen de manera proporcional a $|x|^{-\alpha-1}$; $0 < \alpha < 2$ para toda n . Entonces S_n tenderá hacia una distribución estable conforme $n \rightarrow \infty$. Si $\alpha > 2$ la sucesión convergerá a una distribución estable con parámetro de estabilidad igual a 2 i.e. a una distribución normal.

3. Pruebas de Normalidad

Cómo vimos en la sección anterior, sin importar la distribución de la cual provenga una muestra de variables aleatorias X_n , estandarizarlas y sumarlas nos dará en el límite una distribución normal. Por esta razón, la mayor parte de las pruebas de hipótesis paramétricas traen intrínseco un fuerte supuesto de normalidad. Es conveniente entonces desarrollar métodos o pruebas que nos permitan dar un diagnóstico del nivel de *confianza* que se puede tener en este supuesto.

Antes de hablar de pruebas específicamente diseñadas para determinar normalidad, exploraremos los distintos criterios que determinan la eficiencia de una prueba.

3.1. Criterios de Calidad de Pruebas

La calidad de una sucesión de pruebas puede ser determinada por su **poder** ante alternativas arbitrariamente cercanas a la hipótesis nula. La eficiencia relativa entre dos pruebas es entonces el cociente entre el número de observaciones requeridas por cada una de las pruebas para alcanzar el mismo **poder**.

Definición 3.1. El **poder** de una prueba que rechaza la hipótesis nula H_0 :

$\theta \in \Theta_0$ si un estadístico de prueba cae en una región crítica K_n es la función $\theta \rightarrow \pi_n(\theta) = P_\theta(T_n \in K_n)$.

Definición 3.2. Una prueba es de **tamaño** α si $\sup_{\theta \in \Theta_0} \{\pi_n(\theta)\} \leq \alpha$.

Es lógico pensar que un estadístico con **poder** π es mejor que uno con poder $\bar{\pi}$ si y sólo si

$$\pi_n(\theta) \leq \bar{\pi}_n(\theta), \quad \theta \in \Theta_0$$

$$\pi_n(\theta) \geq \bar{\pi}_n(\theta), \quad \theta \in \Theta_1$$

Por lo regular nos interesa la comparación en cuanto al comportamiento asintótico de las pruebas, de aquí surge siguiente definición.¹⁰

Definición 3.3. Una prueba $\theta \rightarrow \pi_n(\theta)$ es **consistente** a nivel α si es asintóticamente de nivel α y $\pi_n(\theta) \rightarrow 1$.

A continuación expondremos las formas de distintas pruebas de normalidad y daremos una breve idea intuitiva de su funcionamiento.

3.2. La Prueba de Normalidad Jarque-Bera

Esta prueba hace uso del hecho de que si una muestra proviene de una distribución normal, esta deberá ser razonablemente simétrica y deberá concentrar buena parte de su *masa* en una vecindad relativamente pequeña de la media.

¹⁰Existen otros criterios de comparación entre pruebas que no serán discutidos en este documento tales como eficiencia relativa.

El estadístico de prueba se obtiene de la siguiente manera:

$$JB = n \left(\frac{k_3^2}{6} + \frac{k_4^2}{24} \right)$$

Donde

$$k_3 = \frac{\sum_{i=1}^n (X_i - \bar{X})^3}{nS^3}$$

$$k_4 = \frac{\sum_{i=1}^n (X_i - \bar{X})^4}{nS^4} - 3$$

3.3. La Prueba de Normalidad D'Agostino

Esta prueba se basa en el estadístico D y se define como sigue

$$D = \frac{T}{\sqrt{n^3 SS}}$$

Donde

$$T = \sum_{i=1}^n \left(i - \frac{n+1}{2} \right) X_i$$

y SS es la suma de cuadrados de los datos.

3.4. La Prueba de Normalidad χ^2

Supongamos una muestra de variables aleatorias X_1, X_2, \dots, X_n provenientes de una distribución P . Quisiéramos probar la hipótesis $H_0 : P \in \mathbb{P}_0$ i.e. probar si P pertenece a una familia de distribuciones \mathbb{P}_0 . Definimos enton-

ces una partición $S = \cup_j X_j$ Para un determinado conjunto A denotamos $P_n(A) = n^{-1}(1 \leq i \leq n : X_i \in A)$ i.e. la fracción de observaciones que caen dentro de A. Entonces $(P_n(S_1), P_n(S_2), \dots, P_n(S_k))$ tiene una distribución multinomial y el estadístico de prueba está dado por

$$\sum_{j=1}^k \frac{n(P_n(S_j) - \hat{P}(S_j))^2}{\hat{P}(S_j)}$$

donde $\hat{P}(S_j)$ es un estimador de $P_n(S_j)$.

4. Teorema de Chebyshev

Si μ y σ son la media y la desviación estandar de una variable aleatoria entonces para cualquier constante positiva k la probabilidad es al menos $1 - \frac{1}{k^2}$ que X asumirá un valor dentro de k desviaciones estándar de la media; simbólicamente

$$P(|X - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}$$

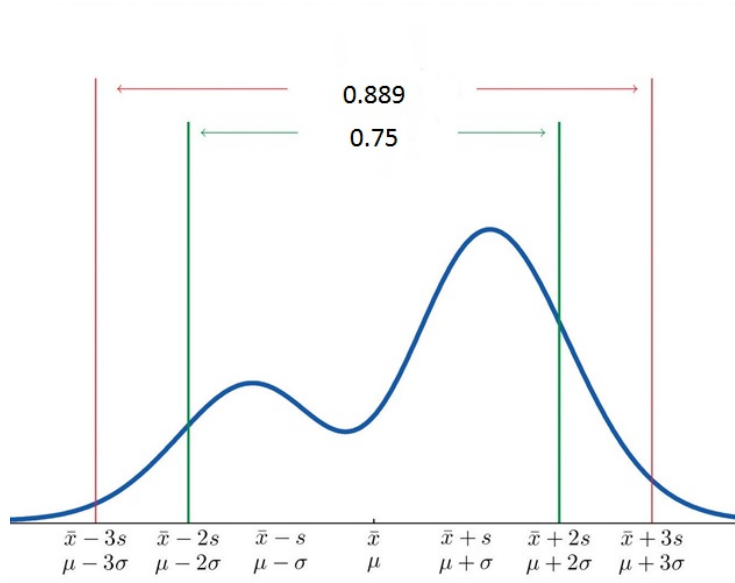


Figura 1: Teorema de Chebyshev

Demostración. Sea:

$$\sigma^2 = E[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx \quad (1)$$

Entonces al dividir la integral en tres partes obtenemos

$$\sigma^2 = \int_{-\infty}^{\mu - k\sigma} (x - \mu)^2 f(x) dx + \int_{\mu - k\sigma}^{\mu + k\sigma} (x - \mu)^2 f(x) dx + \int_{\mu + k\sigma}^{\infty} (x - \mu)^2 f(x) dx \quad (2)$$

Puesto que el integrando $(x - \mu)^2 f(x)$ es no negativo, podemos formar la desigualdad

$$\sigma^2 \geq \int_{-\infty}^{\mu-k\sigma} (x-\mu)^2 f(x) dx + \int_{\mu+k\sigma}^{\infty} (x-\mu)^2 f(x) dx \quad (3)$$

al borrar la segunda integral. Por consiguiente, puesto que $(x-\mu)^2 \geq k^2 \sigma^2$ para $x \leq \mu - k\sigma$ o $x \geq \mu + k\sigma$ se sigue que

$$\frac{1}{k^2} \geq \int_{-\infty}^{\mu-k\sigma} f(x) dx + \int_{\mu+k\sigma}^{\infty} f(x) dx \quad (4)$$

y de ahí que

$$\sigma^2 \geq \int_{-\infty}^{\mu-k\sigma} k^2 \sigma^2 f(x) dx + \int_{\mu+k\sigma}^{\infty} k^2 \sigma^2 f(x) dx \quad (5)$$

siempre y cuando $\sigma^2 \neq 0$. Puesto que la suma de las dos integrales en el lado derecho es la probabilidad de que X asumirá un valor menor que o igual a $\mu - k\sigma$ o mayor que o igual a $\mu + k\sigma$, hemos demostrado así que

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

y se sigue que

$$P(|X - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}$$

□

Por ejemplo, la probabilidad es al menos $1 - \frac{1}{2^2} = \frac{3}{4}$ que una variable aleatoria X asumirá un valor dentro de dos desviaciones estándar de la media, la probabilidad es de al menos $1 - \frac{1}{3^2} = \frac{8}{9}$ que asumirá un valor dentro de tres desviaciones estándar de la media y la probabilidad es al menos $1 - \frac{1}{5^2} = \frac{24}{25}$ que asumirá un valor dentro de cinco desviaciones estándar de la media. Es en este sentido que σ controla la amplitud o la dispersión de la distribución de una variable aleatoria.

Ejemplos:

Si la densidad de probabilidad de X está dada por

$$f(x) = \begin{cases} 630x^4(1-x)^4 & \text{para } 0 < x < 1 \\ 0 & \text{en cualquier otra parte} \end{cases} \quad (6)$$

entonces queremos encontrar la probabilidad de la variable aleatoria asumirá un valor dentro de dos desviaciones estándar de la media y comparar esta probabilidad con el límite inferior proporcionado por el Teorema de Chebyshev.

Considerando que $\mu = E(X)$ se encuentra definida como $\mu = \int_0^1 xf(x)dx$ tenemos que $\mu = \frac{1}{2}$ y tenemos que $\sigma^2 = E[(X - \mu)^2] = \int_0^1 (x - \mu)^2 f(x)dx$ entonces $\sigma^2 = \frac{1}{44}$, por lo que $\sigma = \sqrt{\frac{1}{44}}$ aproximadamente 0.15. Por lo que la probabilidad de que X tome un valor dentro de dos desviaciones estándar de la media es la probabilidad de que tomará un valor entre 0.20 y 0.80:

$$P(0,20 < X < 0,80) = \int_{0,20}^{0,80} 630x^4(1-x)^4 dx = 0,96$$

Si comparamos el resultado anterior con la probabilidad proporcionada por el Teorema de Chebyshev donde la probabilidad es al menos 0.75, podemos observar que el resultado obtenido es una aseveración más precisa.

5. Cotas

Considerando el resultado anterior, podemos derivar que dado un conjunto de datos que tiene una distribución aproximada de campana entonces:

$$P(|X - \mu| \leq \sigma) \approx 0,68$$

$$P(|X - \mu| \leq 2\sigma) \approx 0,95$$

$$P(|X - \mu| \leq 3\sigma) \approx 0,97$$

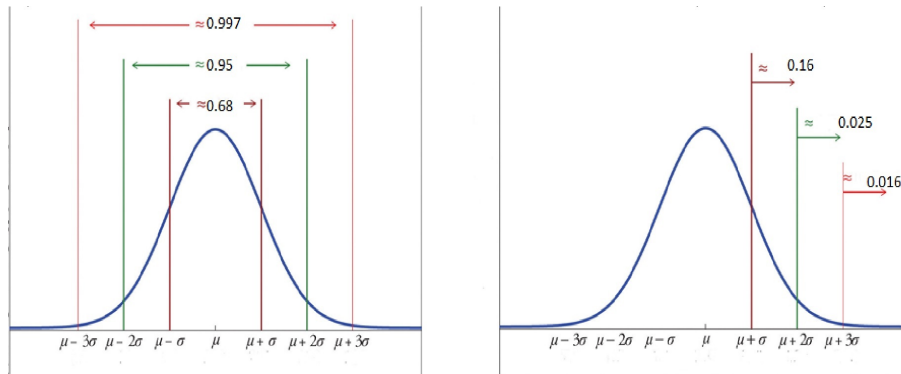


Figura 2: Regla empírica

6. Ejemplos

En esta sección se desarrollan distintos ejemplos en donde se aplicarán algunos de los conceptos mencionados. En la práctica nos encontramos con datos de los cuáles se desconoce su distribución, por lo que si deseamos inferir algo sobre estos, resulta conveniente apoyarnos en el Teorema Central del Límite (TCL) y en el Teorema de Chebyshev.

Comenzamos con un ejemplo en el cual se simulan 1 millón de números aleatorios que tienen una distribución beta con parámetros alfa igual a 1 y beta igual a 5, y se determinó una semilla de 10.

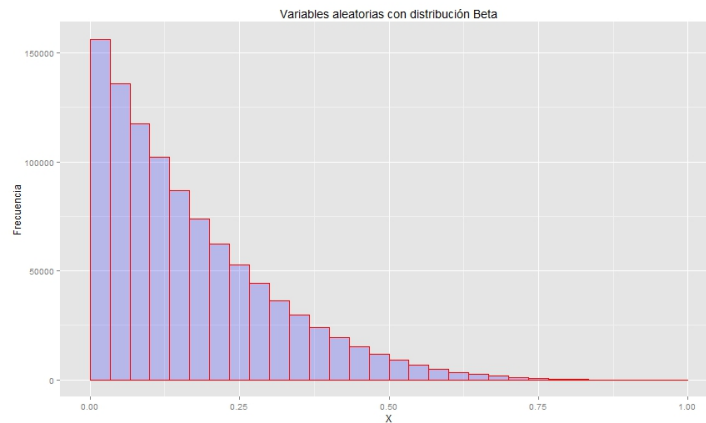


Figura 3: Histograma de 1 millón de números aleatorios simulados

Dado el TCL, se realizaron 1,000 muestras aleatorias con reemplazo de tamaño 500. A cada muestra se le calculó su media con lo cual generamos una lista de 1,000 medias, que se distribuyen de la siguiente forma

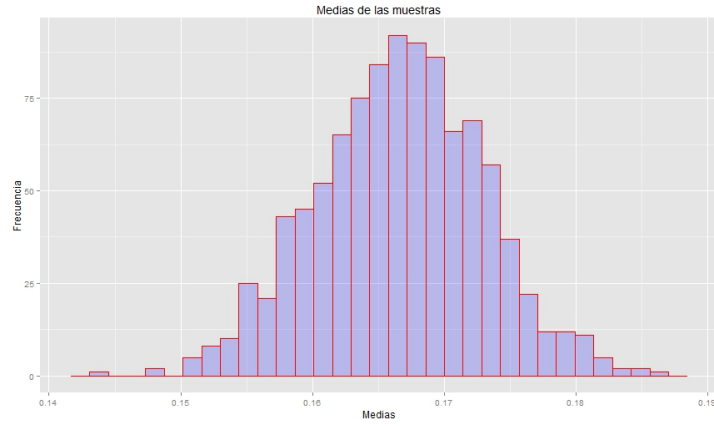


Figura 4: Histograma de las 1,000 medias

Lo siguiente es hacer una prueba de bondad de ajuste que en este caso consiste en probar si las 1,000 medias tienen una distribución normal. Se genera una prueba de χ^2 , la prueba consiste en comparar las frecuencias de los datos en determinado rango según lo esperado de acuerdo a la distribución teórica elegida (en este caso la normal) y las frecuencias observadas. Se establecen la hipótesis nula que los datos se ajustan a una distribución normal y la hipótesis alternativa que los datos no se ajustan a una distribución normal.

Para comparar los datos, se obtienen los déciles de la distribución de medias y se calcula la cantidad de datos que corresponden a cada uno. Según el valor Z de una distribución de probabilidad acumulada normal, un décimo de sus observaciones caen en cada uno de los intervalos: $(\mu - 5\sigma - \mu - 1.29\sigma)$; $(\mu - 1.29\sigma - \mu - 0.85\sigma)$; $(\mu - 0.85\sigma - \mu - 0.53\sigma)$; $(\mu - 0.53\sigma - \mu - 0.26\sigma)$; $(\mu - 0.26\sigma - \mu)$ y simétricamente $(\mu + \mu + 0.26\sigma)$ y así sucesivamente.

déciles	1	2	3	4	5	6	7	8	9	10
esperados	100	100	100	100	100	100	100	100	100	100
observados	108	95	93	98	94	103	110	94	113	92
diferencia	+8	-5	-7	-2	-6	+3	+10	-6	+13	-8

Cuadro 1: Prueba χ^2

Cada diferencia es elevada al cuadrado y dividida entre su valor esperado y sumadas, dando como resultado 5.56. Dicho estadístico se compara con el valor de la χ^2 con $v-1$ grados de libertad, es decir 9, v son el número de clases, en este caso 10. El resultado 5.56 es muy pequeño en relación al valor correspondiente, por lo que no se rechaza la hipótesis nula que establece que los datos se ajustan a una distribución normal.

Dado que se comportan como una normal, calculamos la media, la varianza y su desviación estándar que corresponden a 0.1665908, 4.073302e-05 y 0.006382242, respectivamente. De nuestros datos iniciales que tienen una distribución beta, podemos inferir que tienen una media de 0.1665908 y una desviación estándar de $\sigma\sqrt{n}$, es decir $0.006382242\sqrt{500}$ equivalente a 0.1427113. Mediante al Teorema de Chebyshev podemos inferir que los datos están en el intervalo 0 a 0.45 con una probabilidad de al menos 75 por ciento.

En un segundo ejemplo trabajamos con la base de datos 'Current Employee Names, Salaries, and Position Titles' de los empleados de la ciudad de Chicago, que incluye Nombre, Posición, Departamento y Salario Anual con 32,181 tuplas. La base puede ser consultada en <https://data.cityofchicago.org/Administration-Finance/Current-Employee-Names-Salaries-and-Position-Title/xzkq-xp2w>. Se seleccionó la variable salario anual, la cual se comporta de la siguiente manera

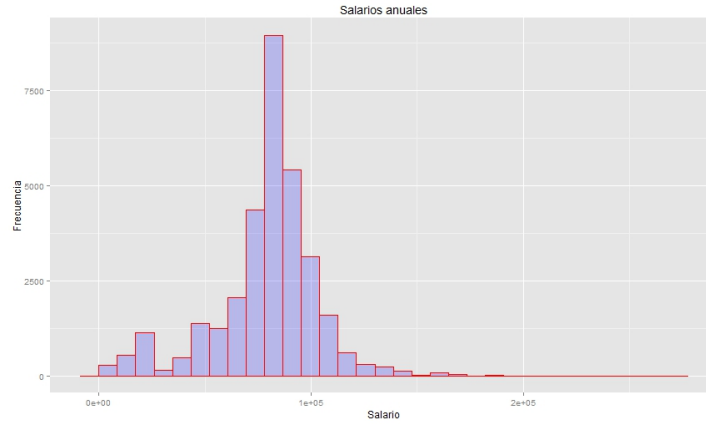


Figura 5: Histograma de los salarios anuales

Se realizaron 1,000 muestras con reemplazo de tamaño 1,000. Siguiendo los mismos pasos que en el ejemplo anterior, se realizó la prueba de χ^2 separando por déciles la frecuencia de datos obtenida en cada intervalo y se obtuvo lo siguiente:

déciles	1	2	3	4	5	6	7	8	9	10
esperados	100	100	100	100	100	100	100	100	100	100
observados	101	105	93	90	101	106	111	98	90	105
diferencia	+1	+5	-7	-10	+1	+6	+11	-2	-10	+5

Cuadro 2: Prueba χ^2

Nuevamente, se plantea en la hipótesis nula que los datos se ajustan a una distribución normal y una hipótesis alternativa que los datos no se ajustan a una distribución normal. El estimador de la χ^2 fue de 4.62, valor que resulta pequeño de acuerdo al estadístico que corresponde a los 9 grados de libertad, por lo que no se rechaza la hipótesis nula que los datos se ajustan a una distribución

normal.

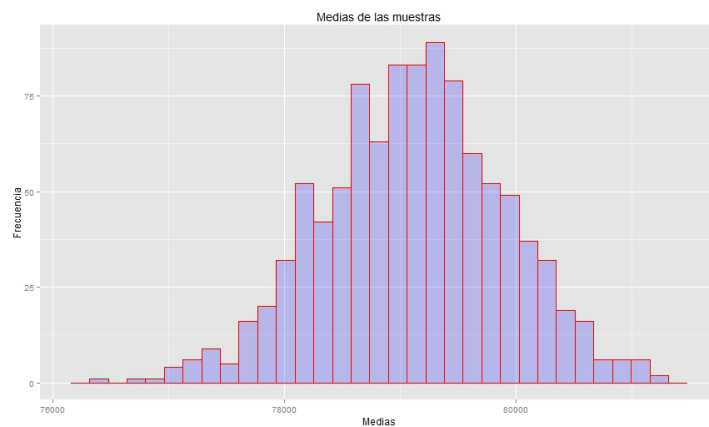


Figura 6: Histograma de las 1,000 medias

Ahora se puede concluir que la media, la varianza y la desviación estándar que corresponden a la lista de 1,000 medias es de 79,111.26, 612,052.8 y 782.338, respectivamente. Además, podemos inferir que la media y la desviación estándar de los datos de los salarios de los empleados de la ciudad de Chicago son 79,111.26 dólares anuales y 24,739.7, respectivamente. Además, mediante el Teorema de Chebyshev concluimos que los salarios están en un intervalo de 29,631 y 128,590 dólares anuales con una probabilidad de al menos 75 por ciento.

Referencias

- [1] A.W. Van Der Vaart, *Asymptotic Statistics*, Cambridge press 1998.
- [2] Hans Fischer, *A history of the Central Limit Theorem From Classical to Modern Probability Theory*, Springer 2010.

- [3] Bradly Efron and Carl Morris, *Stein's Paradox in Statistics*. Scientific American, Volume 236, Issue 5.
- [4] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar *Foundations of Machine Learning*, MIT Press, USA, 2012.
- [5] Shai Shalev-Shwartz, Shai Ben-David. *Understanding Machine Learning, From Theory to Algorithms*, Cambridge University Press, New York, 2014.
- [6] Charles Stein. *Inadmissibility of the usual estimator for the mean of a multivariate normal distribution*, Proceedings of the Third Berkley Symposium on Mathematical Statistics and Probability, Berkley and Los Angeles, University of California Press 1956 Vol 1. pp 197-206.
- [7] W. James and Charles Stein. *Estimation With Quadratic Loss*, Proceedings of the Fourth Berkley Symposium on Mathematical Statistics and Probability, Berkley and Los Angeles, University of California Press 1961. pp 361-379.
- [8] Tukey, J.W., *A survey of sampling from contaminated distributions*. (1960a), Chapter 39 in: *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling* (ed. I. Olkin et al.), Stanford University Press, Stanford, California, 448-485.
- [9] Vladimir N. Vapnik *Statistical Learning Theory*, JOHN WILEY & SONS, INC., USA, 1998.
- [10] Vladimir N. Vapnik *The Nature of Statistical Learning Theory*, Springer, Second Edition, 1999.
- [11] Kuri, Angel *Appendix D: Statistical Determination of the Best Algorithm in U* .
- [12] E. Freund, John *Estadística Matemática con Aplicaciones*.