FACULTAD DE MATEMÁTICAS
PONTIFICIA UNIVERSIDAD
CATÓLICA DE CHILE

EYP1113 - Probabilidad y Estadística

Taller R - 08: Funciones de variables aleatorias y Teorema del límte central

Ricardo Aravena C. - Cristian Capetillo C. - Ingrid Guevara R.
Ricardo Olea O. - Bladimir Morales T., Daniel Saavedra M.

Facultad de Matemáticas
Departamento de Estadística
Pontificia Universidad Católica de Chile

Segundo Semestre 2024

Contenido I

Motivación

Las funciones de variables aleatorias y el Teorema del Límite Central (TLC) son herramientas fundamentales en estadística y probabilidad, utilizadas para simplificar problemas complejos y facilitar el análisis de datos en diversas aplicaciones prácticas.

Las transformaciones de variables aleatorias se aplican para ajustar la forma de una distribución o hacer que una variable cumpla con ciertos supuestos.

El Teorema del Límite Central es fundamental en estadística, ya que establece que la suma (o promedio) de un número suficientemente grande de variables aleatorias independientes e idénticamente distribuidas, sin importar su distribución inicial, tiende a seguir una distribución normal.

Funciones de variables aleatorias

Vamos a plantear un ejemplo práctico para entender inicialmente lo que se logra con las funciones de las variables aleatorias, considere el archivo `maquinas.xlsx`.

Si se tiene un sistema de producción de tornillos donde una máquina tiene dos motores para funcionar. Cada motor tiene un tiempo de falla en horas que sigue una distribución exponencial ($X =$ tiempo hasta la falla del motor 1 e $Y =$ tiempo hasta la falla del motor 2).

Supongamos que ambos motores tienen la misma tasa de falla, $\lambda = 1.5$, es decir que se espera que el motor falle en promedio 1.5 veces por hora. En este contexto, la probabilidad de que un motor falle en un tiempo t específico puede modelarse usando la función de densidad de la distribución exponencial.

Vamos a R...

Funciones de variables aleatorias

Cuando uno de los motores falla, el sistema puede seguir funcionando, pero cuando ambos motores han fallado, el sistema completo se detiene. Por lo tanto, un aspecto clave de este análisis es calcular el tiempo total hasta que ambos motores fallen.

Para esto se sumarán las variables aleatorias exponenciales, $Z = X + Y$, entonces Z ¿qué función de distribución de probabilidad tendrá?, para esto se utilizará la suma de variables aleatorias con el método de convolución.

Suma de Variables Aleatorias Continuas

Sea $Z = X + Y$, en este caso, la función de densidad de Z está dada por

$$f_Z(z) = \int_{-\infty}^{\infty} f_{X,Y}(z-y, y)\, dy \quad \text{o} \quad f_Z(z) = \int_{-\infty}^{\infty} f_{X,Y}(x, z-x)\, dx$$

Si X e Y son independientes, entonces

$$f_{X,Y} = f_X(x) \cdot f_Y(y)$$

Como $X, Y \overset{i.i.d.}{\sim} Exp(\lambda = 1.5)$, se puede demostrar mediante convolución que

$$Z = X + Y \sim Gamma(\alpha = 2, \lambda = 1.5)$$

Vamos a verificar esto en R...

Transformación de variables aleatorias

Si queremos facilitar el análisis o transformar la variable para obtener una nueva forma de distribución que sea más manejable analíticamente, se pueden utilizar funciones de variables aleatorias.

Bajo nuestro ejemplo, si queremos aplicar la transformación $W = \log(Z) = \log(X + Y)$ nos da una nueva variable que es menos sesgada y con menor variabilidad. En análisis de confiabilidad, el logaritmo del tiempo de vida total puede ser útil para construir intervalos de confianza o para aplicar técnicas de regresión que modelen el log-tiempo de vida en función de otras covariables (estos tópicos se verán más adelante).

Transformación de variables aleatorias

Sea $W = g(Z) \Rightarrow W = \log(Z)$. Para calcular la densidad usando el método del jacobiano:

1. Inversa de la transformación: $g^{-1}(W) = Z \Rightarrow \exp(W) = Z$.
2. Densidad de $W = \log(Z)$:

$$f_W(w) = f_Z(g^{-1}(W)) \cdot \left| \frac{d}{dw} g^{-1}(W) \right|$$

3. Expresando la densidad en función de f_Z:

$$f_W(w) = f_Z(\exp(W)) \cdot \left| \frac{d}{dw} \exp(W) \right|$$

4. Finalmente:

$$f_W(w) = f_Z(\exp(W)) \cdot \exp(W)$$

Verifiquemos esto en R...

Distribuciones Muestrales

Resultados

Si $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, entonces:

1. $Z_i = \dfrac{X_i - \mu}{\sigma} \overset{iid}{\sim} \mathcal{N}(0,1)$ y $\displaystyle\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2} \sim \chi_n^2$
2. $\overline{X} \sim \mathcal{N}(\mu, \sigma^2/n)$, y $W_i = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$.
3. $\displaystyle\sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\sigma^2} \sim \chi_{n-1}^2$.
4. $\dfrac{\overline{X} - \mu}{S_x/\sqrt{n}} \sim t_{n-1}$, donde $S_x^2 = \frac{1}{n-1} \sum (X_i - \overline{X})^2$.
5. Si $Y_1, \ldots, Y_m \overset{iid}{\sim} \mathcal{N}(\theta, \tau^2)$, entonces $\dfrac{S_x^2/\sigma^2}{S_y^2/\tau^2} \sim F_{n-1,m-1}$

Teorema del Limite Central (TLC)

El teorema dice que si $X_1, \ldots, X_n$ son variables aleatorias independientes e idénticamente distribuidas (iid) con

$$E(X_i) = \mu \quad \text{y} \quad \mathrm{Var}(X_i) = \sigma^2, \;\; \forall\, i = 1, \ldots, n.$$

Entonces, cuando $n \to \infty$,

$$Z_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\,\sigma} = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \to Z \sim \mathrm{Normal}(0, 1)$$

En otras palabras,

$$\sum_{i=1}^{n} X_i \overset{\text{aprox}}{\sim} \mathrm{Normal}\left(n\,\mu,\; n\sigma^2\right) \quad \text{o} \quad \overline{X}_n \overset{\text{aprox}}{\sim} \mathrm{Normal}\left(\mu,\; \frac{\sigma^2}{n}\right)$$

Resultados TLC

- Sean $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$, entonces

$$S_n = \sum_{i=1}^{n} X_i \sim \text{Binomial}(n, p) \overset{\text{aprox.}}{\sim} \text{Normal}\left(np,\ np(1-p)\right)$$

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \overset{\text{aprox.}}{\sim} \text{Normal}\left(p,\ \frac{p(1-p)}{n}\right)$$

- Sean $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda)$, entonces

$$S_n = \sum_{i=1}^{n} X_i \sim \text{Poisson}(n\lambda)$$

$$\overset{\text{aprox.}}{\sim} \text{Normal}\left(n\lambda, n\lambda\right)$$

Aplicación TLC

En nuestro contexto de las máquinas tenemos solo 2 motores de la máquina que tienen distribución exponencial, ¿si aplicamos el siguiente resultado que obtenemos?

- Sean $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Exponencial}(\lambda)$, entonces

$$S_n = \sum_{i=1}^{n} X_i \sim \text{Gamma}(n, \lambda)$$

$$\overset{\text{aprox.}}{\sim} \text{Normal}\left(\frac{n}{\lambda}, \frac{n}{\lambda^2}\right)$$

recuerde que para nuestro ejemplo $n = 2$.

Aplicación TLC

Ahora suponga que se tiene otra máquina que tiene 25 motores para su funcionamiento bajo el mismo contexto de producción de tornillos, para esto simule 25 variables aleatorias exponenciales i.i.d con parámetro $\lambda = 1.5$, es decir $X_1, \ldots, X_{25} \overset{\text{i.i.d.}}{\sim}$ Exponencial(λ), entonces pruebe que,

$$
\begin{aligned}
S_n = \sum_{i=1}^{n} X_i &\sim \text{Gamma}(n, \lambda) \\
&\overset{\text{aprox.}}{\sim} \text{Normal}\left(\frac{n}{\lambda}, \frac{n}{\lambda^2}\right).
\end{aligned}
$$

Luego realice el mismo procedimiento pero ahora simulando 200 variables aleatorias exponenciales i.i.d con parámetro $\lambda = 1.5$, ¿qué diferencias encuentra?