Paul Pierre Lévy (1886-1971, Francia)

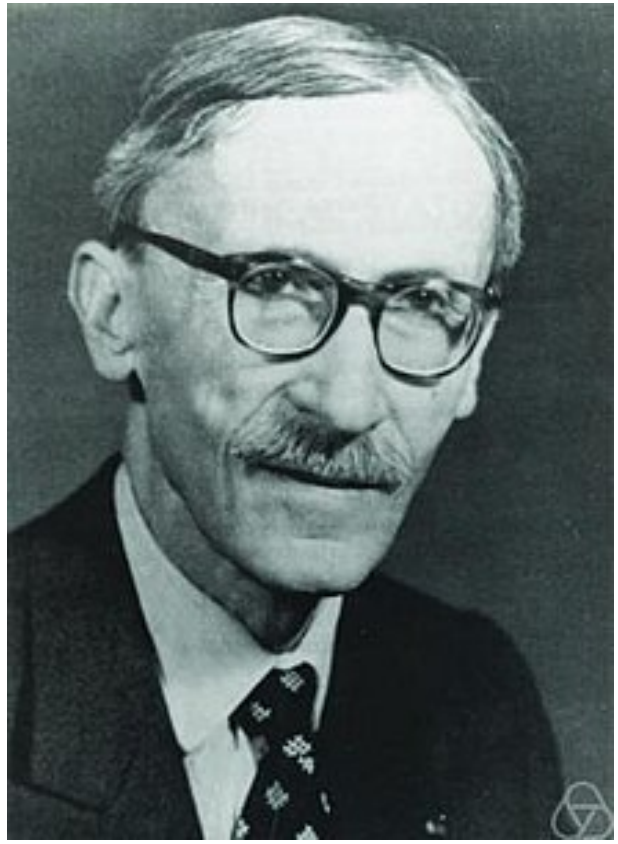


Figura: Demostró el teorema que veremos hoy

Clase 13 de Bioestadística

Teorema Central del Límite

Ernesto Mordecki

CMAT, Facultad de Ciencias, Universidad de la República.

Uruguay

Contenidos de la clase

Esperanza y varianza (repaso)

Esperanza y varianza de una suma de variables aleatorias

Covarianza

El problema de la velocidad

Teorema Central del Límite

Aplicación

Esperanza y varianza (repaso)

Llamamos esperanza matemática de una variable aleatoria X al número $E(X)$ definido como

$$E(X) = \sum_k x_k P(X = x_k), \qquad \int_{-\infty}^{\infty} x f(x) dx$$

si X es una variable discreta o continua, y llamamos varianza al número

$$Var(X) = E\left[(X - E(X))^2\right]$$

Sucesión de v.a.i.i.d.

- El objeto central de la clase de hoy es una sucesión de variables aleatorias

$$X_1, X_2, \ldots, X_n, \ldots$$

que suponemos

  - Independientes: las probabilidades de sucesos de n variables son el producto de las probabilidades individuales. Por ejemplo

$$P(X_1 \leq a, X_2 \leq b, X_3 \leq c) = P(X_1 \leq a) \times P(X_2 \leq b) \times P(X_3 \leq c)$$

  - Idénticamente distribuídas: si bien los resultados son distintos, tienen las mismas probabilidades de ocurrir:

$$P(X_k \leq a) = P(X_1 \leq a) = F(a) \qquad \text{para todo } k = 1, 2, \ldots$$

- Las sucesiones de variables aleatorias independientes e idénticamente distribuídas

  v.a.i.i.d.

  son una idealización teórica de un experimento que se repite una gran cantidad de veces bajo las mismas condiciones.

- El ejemplo que consideramos es tirar un  una gran cantidad de veces

- Podríamos modelar la vacunación masiva de una población y medir la respuesta en anticuerpos al año de aplicar la vacuna

- ¿Ejemplos?

Propiedades

- Como tenemos la misma distribución, tienen la misma esperanza y la misma varianza. Les llamamos

$$\mu = E(X_1), \qquad \sigma^2 = var(X_1).$$

- Ahora vamos a quedarnos con las n primeras variables i.i.d:

$$X_1, X_2, \ldots, X_n$$

- Nos interesa en realidad el promedio

$$\overline{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

- Queremos calcular la esperanza y la varianza del promedio en términos de μ y σ^2

Propiedades[1]

Teorema

(a) Sean $Y_1, \ldots, Y_n$ variables aleatorias cualesquiera. Tenemos

$$E(Y_1 + \cdots + Y_n) = E(Y_1) + \cdots + E(Y_n)$$

(b) Sean X e Y variables independientes. Entonces

$$E(X \times Y) = E(X) \times E(Y)$$

[1] Suponemos que las esperanzas y varianzas que aparecen son finitas

Demostraciones

- La propiedad (a) vamos a asumirla verdadera.
- Veamos (b) en el caso que ambas variables sean discretas y finitas.
- X toma valores $x_1, \ldots x_n$ con probabilidades $p_1, \ldots, p_n$, Y toma valores $y_1, \ldots, y_k$ con probabilidades $q_1, \ldots, q_k$
- Entonces XY toma los valores $x_i y_\ell$ con probabilidades[2] $p_i q_\ell$
- Tenemos

$$E(XY) = \sum_{i,\ell} x_i y_\ell \times p_i q_\ell = \sum_{i,\ell} (x_i p_i)(y_\ell q_\ell)$$
$$= \sum_i x_i p_i \sum_\ell y_\ell q_\ell = E(X)E(Y)$$

[2] Suponemos que todos los productos dan distinto resultado

Covarianza

- Definimos la covarianza de dos v.a. X e Y mediante

$$cov(X, Y) = E\left[(X - E(X))(Y - E(Y))\right]$$

- Notemos que si $X = Y$ entonces

$$\begin{aligned} cov(X, X) &= E\left[(X - E(X))(X - E(X))\right] \\ &= E\left[(X - E(X))^2\right] = var(X) \end{aligned}$$

- Operemos:

$$\begin{aligned} cov(X,Y) &= E\left[(X - E(X))(Y - E(Y))\right] \\ &= E\left[XY - YE(X) - XE(Y) + E(X)E(Y)\right] \\ &= E\left[XY\right] - E\left[YE(X)\right] - E\left[XE(Y)\right] + E\left[E(X)E(Y)\right] \\ &= E\left[XY\right] - E[Y]E(X) - E[X]E(Y) + E(X)E(Y) \\ &= E(XY) - E(X)E(Y) \end{aligned}$$

- Conclusión: si X, Y son independientes, entonces

$$\boxed{cov(X,Y) = 0}$$

Prop. Sean X e Y variables independientes. Entonces

$$var(X+Y) = var(X) + var(Y)$$

Dem. Es una cuenta:

$$\begin{aligned} var(X+Y) &= E\left[(X+Y-E(X+Y))^2\right] \\ &= E\left[(X-E(X)+Y-E(Y))^2\right] \\ &= E\Big[(X-E(X))^2 + (Y-E(Y))^2 \\ &\qquad + 2\,(X-E(X))(Y-E(Y))\Big] \\ &= var(X) + var(Y) + 2cov(X,Y) \\ &= var(X) + var(Y). \end{aligned}$$

Generalización

Prop. Sean $X_1, \ldots, X_n$ variables independientes. Entonces

$$var(X_1 + \cdots + X_n) = var(X_1) + \cdots + var(X_n)$$

Dem. Sea $X = X_1$, $Y = X_2 + \cdots + X_n$. Las variables X e Y son independientes, entonces

$$var(\underbrace{X_1}_{=X} + \underbrace{X_2 + \cdots + X_n}_{=Y}) = var(X_1) + var(X_2 + \cdots + X_n)$$

- Un paso mas igual nos da

$$var(X_1+X_2+\cdots+X_n) = var(X_1)+var(X_2)+var(X_3+\cdots+X_n)$$

- Lo hacemos tantas veces como sea necesario. □

- Estamos en condiciones de calcular la esperanza y la varianza de

$$\overline{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

para $X_1, \ldots, X_n$, i.i.d.

- Recordemos que

$$E(aX + b) = aE(X) + b, \qquad var(aX + b) = a^2 var(X).$$

- Entonces

$$\begin{aligned} E(\overline{X}_n) &= E\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{1}{n} E\left(X_1 + \cdots + X_n\right) \\ &= \frac{1}{n}\left[E(X_1) + \cdots + E(X_n)\right] = \frac{1}{n}[nE(X_1)] = \mu. \end{aligned}$$

- La esperanza del promedio es igual a la esperanza de los sumandos.

La varianza de un promedio

- Tenemos

$$var(\overline{X}_n) = var\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{1}{n^2} var\left(X_1 + \cdots + X_n\right)$$
$$= \frac{1}{n^2}\left[var(X_1) + \cdots + var(X_n)\right] = \frac{1}{n^2}[nvar(X_1)] = \frac{\sigma^2}{n}.$$

- La varianza del promedio es la n-ésima parte de la varianza de cada sumando
- Cuantos más sumandos menos varianza

El problema de la velocidad

- Según la ley fuerte de los grandes números sabemos que el promedio se aproxima a la esperanza
- Es decir

$$\overline{X}_n \to \mu \qquad \text{si } n \text{ tiende a infinito}$$

- Pero . . . , cuánto hay que esperar?
- ¿Cuántos experimentos hay que hacer?
- Cuán pequeño es el error

$$\overline{X}_n - \mu$$

- En primer lugar sabemos que

$$E\left(\overline{X}_n - \mu\right) = E\left(\overline{X}_n\right) - E(\mu) = \mu - \mu = 0.$$

- En cuanto a la varianza

$$var\left(\overline{X}_n - \mu\right) = var\left(\overline{X}_n\right) = \frac{1}{n}\sigma^2.$$

- Multiplico por σ^2 y divido por n, y aplico la propiedad,

$$\frac{\sigma^2}{n} var\left(\overline{X}_n\right) = 1, \qquad var\left(\frac{\sqrt{n}}{\sigma}\left[\overline{X}_n - \mu\right]\right) = 1$$

- Consideremos la variable

$$Z_n = \frac{\sqrt{n}}{\sigma}\left[\overline{X}_n - \mu\right]$$

- Tiene esperanza cero y varianza uno.

Teorema Central del Límite

- Paul Lévy demostró que

$$Z_n \to Z \sim \mathcal{N}(0,1) \qquad \text{cuando } n \text{ tiende a infinito}$$

- Eso quiere decir que

$$P(a \leq Z_n \leq b) \to P(a \leq Z \leq b) = \int_a^b \varphi(x)dx$$

- Podemos escribir entonces, si n es grande

$$Z_n = \frac{\sqrt{n}}{\sigma}\left[\overline{X}_n - \mu\right] \approx Z$$

- Podemos escribir entonces, si n es grande

$$Z_n = \frac{\sqrt{n}}{\sigma}\left[\overline{X}_n - \mu\right] \approx Z$$

- Despejando

$$\left[\overline{X}_n - \mu\right] \approx \frac{\sigma}{\sqrt{n}} Z$$

- Despejando

$$\overline{X}_n \approx \mu + \frac{\sigma}{\sqrt{n}} Z$$

- Es decir, si sustituímos μ (desconocido) por $\overline{X}_n$ cometemos un error de la forma

$$\frac{\sigma}{\sqrt{n}} Z$$

- Cuanto más grande es n menor es el error
- El error decrece como la raíz de n.

Aplicación

- Queremos simular Z
- Sabemos simular uniformes
- Resulta que 12 uniformes aproximan bien a Z

$$Z_{12} = \frac{\sqrt{12}}{\sigma}\left[\overline{X}_{12} - \frac{1}{2}\right] \approx Z$$

- Como la varianza es $\sigma^2 = 1/12$, tenemos $\sigma = \sqrt{1/12}$ se multiplica con $\sqrt{n}$
- En conclusión

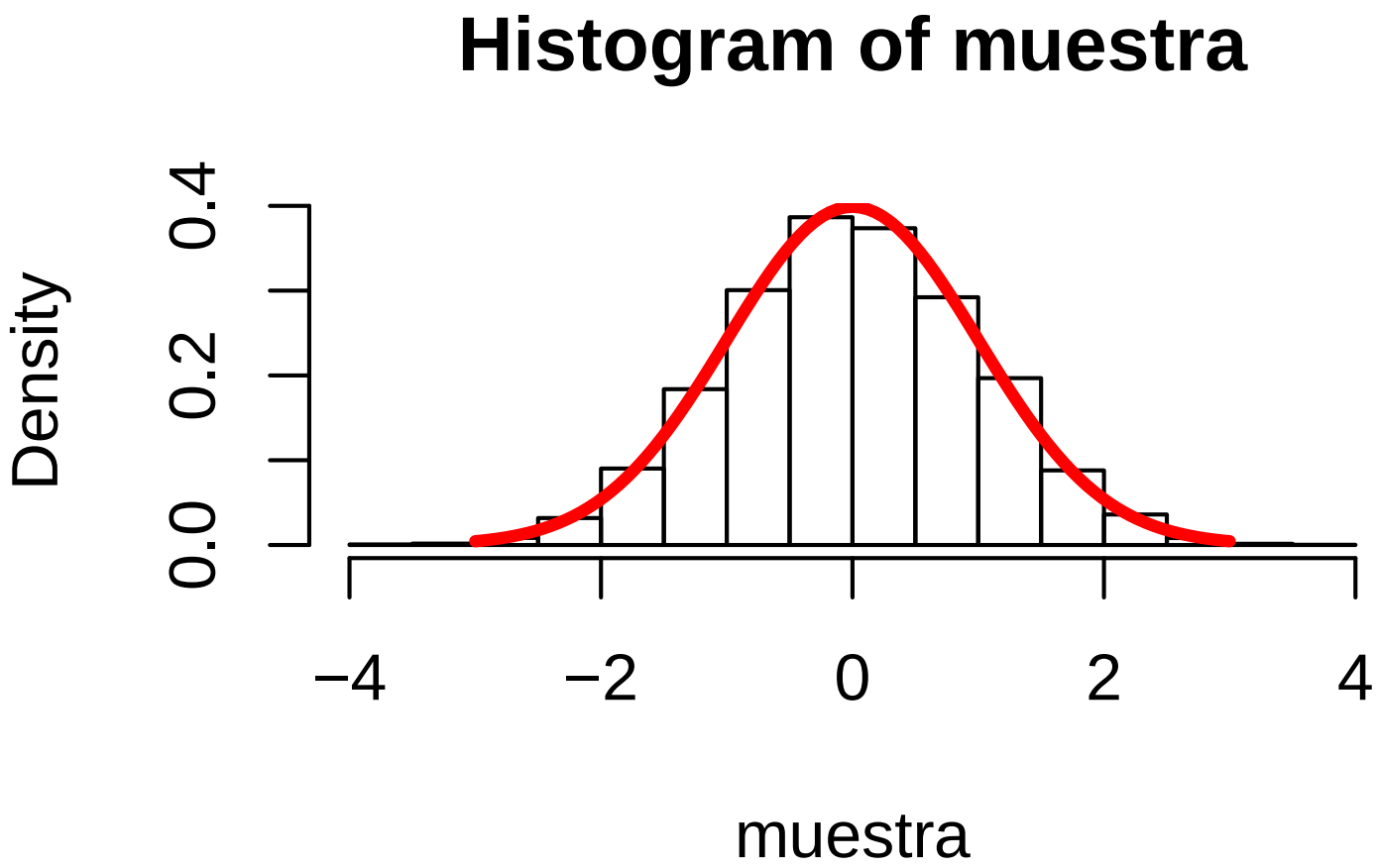
$$Z_{12} = 12\left[\overline{X}_{12} - \frac{1}{2}\right] \approx Z$$

Escribo el siguiente código:

```r
# simula una normal Z:
12*mean(runif(12)-0.5)
# preparo un vector para una muestra
muestra<-c()
# tamaño de la muestra
tamano<-10000
# repito la simulacion "tamano" veces
# y guardo el resultado en muestra
for(i in 1:tamano){
  valor<-12*mean(runif(12)-0.5)
  muestra<-c(muestra,valor)
}
# calculo el histograma de mi muestra
hist(muestra,freq = F)
# lo comparo con la densidad normal estándar
curve(dnorm,-3,3,add=T,lwd=3,col="red")
```

Simulo 10000 muestras con 12 uniformes

El gráfico compara el histograma de la muestra con la densidad normal estándar:



La aproximación indica que los valores simulados son aproximadamente normales estándar.