EYP1113 - Probabilidad y Estadística

Laboratorio 5

Profesoras: Natalia Venegas y Pilar Tello

Facultad de Matemáticas
Departamento de Estadística
Pontificia Universidad Católica de Chile

Segundo Semestre 2019

Contenido I

1 Medidas Descriptivas Teóricas vs Empíricas
- Medidas Centrales
- Medidas de Posición
- Medidas de Dispersión
- Medidas de Asimetría y Forma

Medidas Descriptivas Teóricas vs Empíricas

Medidas Descriptivas Teóricas vs Empíricas

Una variable aleatoria puede ser descrita totalmente por su **función de distribución de probabilidad o de densidad**, o bien por su **función de distribución de probabilidad acumulada**.

Sin embargo, en la práctica la forma exacta puede no ser totalmente conocida.

En tales casos se requieren ciertas “medidas” para tener una idea de la forma de la distribución:

- Medidas Centrales.
- Medidas de Dispersión.
- Medidas de Asimetrías y Forma

Valor esperado (media)

Para una variable aleatoria X se define el valor esperado, μ_X, como:

$$\mu_X = E(X) = \begin{cases} \displaystyle\sum_{x \in \Theta_X} x \cdot p_X(x), & \text{Caso Discreto} \\[2ex] \displaystyle\int_{-\infty}^{\infty} x \cdot f_X(x)\, dx, & \text{Caso Continuo} \end{cases}$$

En **R**, la función **mean(, na.rm = TRUE)** la calcula de manera empírica.

Otras medidas de centro son:

- **La Moda**: Valor más frecuente o con mayor probabilidad.

  ```
  filtro = (table(X) == max(table(X)))
  table(X)[filtro]
  ```

- **La Mediana**: Sea x_{med} el valor que toma la mediana, entonces:

  $$F_X(x_{\text{med}}) = 1/2$$

  En **R**, la función **median(, na.rm = TRUE)** la calcula de manera empírica.

Esperanza matemática

La noción del valor esperado como un promedio ponderado puede ser generalizado para funciones de la variable aleatoria X.

Dada una función $g(X)$, entonces el valor esperado de esta puede ser obtenido como:

$$E[g(X)] = \begin{cases} \displaystyle\sum_{x \in \Theta_X} g(x) \cdot p_X(x), & \text{Caso Discreto} \\ \displaystyle\int_{-\infty}^{\infty} g(x) \cdot f_X(x)\, dx, & \text{Caso Continuo} \end{cases}$$

```
g = function(X){X^2}
mean(g(X))
```

Percentil: Valor en los reales, llamemos x_p, que es superior al $p \times$ 100 % de la información.

$$F_X(x_p) = p$$

En **R** las siguientes funciones entregan percentiles empíricos.

```
"quantile" : Percentil
"min"      : Mínimo
"max"      : Máximo
```

Varianza

$$\sigma_X^2 = \mathrm{Var}(X) = \mathrm{E}[(X-\mu_X)^2] = \begin{cases} \displaystyle\sum_{x \in \Theta_X} (x-\mu_X)^2 \cdot p_X(x), & \text{Caso Discreto} \\ \displaystyle\int_{-\infty}^{\infty} (x-\mu_X)^2 \cdot f_X(x)\, dx, & \text{Caso Continuo} \end{cases}$$

En **R**, la función **var()** la calcula.

Rango: Min - Max

man(X)-min(X)

Rango Intercuantil: $x_{0,75} - x_{0,25}$

IQR(X) = quantile(X, 0.75) - quantile(X, 0.25)

En términos de dimensionalidad, es conveniente utilizar la desviación estándar, es decir,

$$\sigma_X = \sqrt{\mathrm{Var}(X)}$$

En **R**, la función **sd()** la calcula.

Ahora, si $\mu_X > 0$, una medida adimensional de la variabilidad es el coeficiente de variación (c.o.v)

$$\delta_X = \frac{\sigma_X}{\mu_X}$$

$\delta =$ sd(X, na.rm = TRUE)/mean(X, na.rm = TRUE)

Se define una medida de asimetría (skewness) como al tercer momento central:

$$\mathrm{E}[(X-\mu_X)^3] = \begin{cases} \displaystyle\sum_{x_i \in \Theta_X} (x_i - \mu_X)^3 \cdot p_X(x_i), & \text{Caso Discreto} \\ \displaystyle\int_{\infty}^{\infty} (x - \mu_X)^3 \cdot f_X(x)\, dx, & \text{Caso Continuo} \end{cases}$$

Una medida conveniente es el coeficiente de asimetría que se define como:

$$\theta_X = \frac{E[(X-\mu_X)^3]}{\sigma_X^3}$$

El cuarto momento central se conoce como la kurtosis

$$E[(X-\mu_X)^4] = \begin{cases} \displaystyle\sum_{x_i \in \Theta_X} (x_i - \mu_X)^4 \cdot p_X(x_i), & \text{Caso Discreto} \\ \\ \displaystyle\int_{\infty}^{\infty} (x-\mu_X)^4 \cdot f_X(x)\, dx, & \text{Caso Continuo} \end{cases}$$

que es una medida del “apuntamiento” o “achatamiento” de la distribución de probabilidad o de densidad.

Usualmente se prefiere el coeficiente de kurtosis

$$K_X = \frac{E[(X-\mu_X)^4]}{\sigma_X^4} - 3$$

```
Skewness = function(x){
n = length(x)
sum((x - mean(x))^3/n)/sd(x)^3
}

Kurtosis = function(x){
n = length(x)
sum((x - mean(x))^4/n)/sd(x)^4-3
}
```

Cuando hay dos variable aleatoria X e Y, puede haber una relación entre las variables.

En particular, la covarianza definida como

$$\text{Cov}(X, Y) = \text{E}[(X - \mu_X)(Y - \mu_Y)] = \text{E}(X \cdot Y) - \mu_X \cdot \mu_Y$$

mide el grado de asociación lineal entre dos variables, pero es preferible su normalización llamada correlación para poder cuantificar la magnitud de la relación.

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$

Este coeficiente toma valores en el intervalo (-1, 1).

En **R**, las funciones **cov()** y **cor()** entrega ambas medidas.