Capítulo 2

Jacknife y Bootstrap

Suponga que se quiere estimar un intervalo de confianza para la media μ desconocida de un conjunto de datos $X_1, \ldots, X_n$ que tiene distribución $\mathcal{N}(\mu, \sigma^2)$.

Primero se conoce que

$$\sqrt{n}\left(\hat{\mu} - \mu\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \sigma^2\right),$$

y esto nos permite escribir el intervalo de confianza como

$$\left[\hat{\mu} - \hat{\sigma} z_{1-\frac{\alpha}{2}}, \hat{\mu} + \hat{\sigma} z_{1-\frac{\alpha}{2}}\right]$$

donde $z_{1-\frac{\alpha}{2}}$ es el cuantil $1 - \frac{\alpha}{2}$ de una normal estándar.

La expresión anterior es posible ya que el supuesto es que la distribución de $\hat{\theta}$ es normal.

Pregunta 2.0.1

¿Qué pasaría si este supuesto es falso o al menos no conocemos la distribución de $\hat{\theta}$?

¿Cómo podemos encontrar ese intervalo de confianza?

Cuidado 2.0.2

Para una muestra fija, el estimador anterior $\hat{\mu}$ solamente un valor. No se conoce la distribución de $\hat{\mu}$. Lo único que se puede estimar son valores puntuales como la media, varianza, mediana, etc, pero no sabemos nada de su distribución.

2.0.1. Caso concreto

Suponga que tenemos la siguiente tabla de datos, que representa una muestra de tiempos y distancias de viajes en Atlanta.

Cargamos la base de la siguiente forma:

```
CommuteAtlanta <- read.csv2("data/CommuteAtlanta.csv")
kable(head(CommuteAtlanta))
```

City	Age	Distance	Time	Sex
Atlanta	19	10	15	M
Atlanta	55	45	60	M
Atlanta	48	12	45	M
Atlanta	45	4	10	F
Atlanta	48	15	30	F
Atlanta	43	33	60	M

Para este ejemplo tomaremos la variable `Time` que la llamaremos `x` para ser más breves. En este caso note que

```
x <- CommuteAtlanta$Time
```

```
mean(x)
```

```
## [1] 29.11
```

y su varianza es

```
(Tn <- var(x))
```

```
## [1] 429.2484
```

A partir de estos dos valores, ¿Cuál sería un intervalo de confianza para la media?

Note que esta pregunta es difícil ya que no tenemos ningún tipo de información adicional.

Las dos técnicas que veremos a continuación nos permitirán extraer *información adicional* de la muestra.

Nota 2.0.3

Para efectos de este capítulo, llamaremos $T_n = T(X_1, \ldots, X_n)$ al estadístico formado por la muestra de los X_i's.