Simulación

7. Técnicas de remuestreo: Jackknife, Bootstrap, Validación cruzada

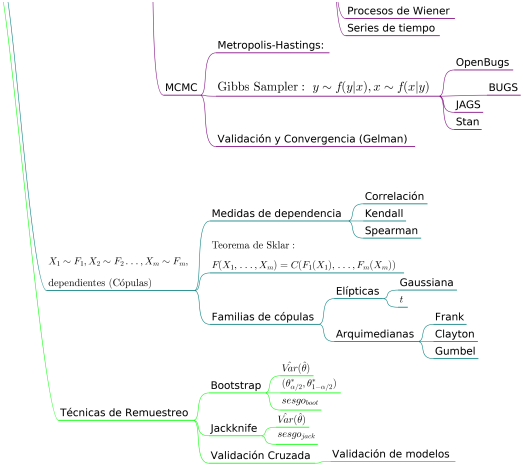
Jorge de la Vega Góngora

Departamento de Estadística,
Instituto Tecnológico Autónomo de México

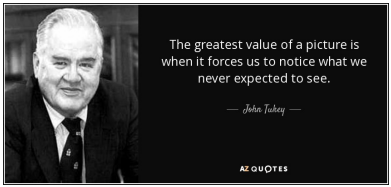
Clase 12

Introducción

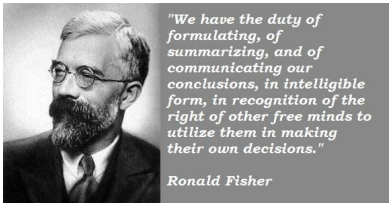
¡Ya (casi) acabamos!



Contexto histórico



John Tukey, 1915-2000.

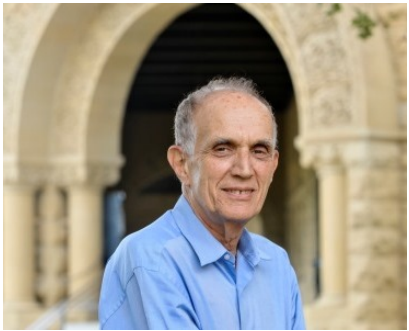


Ronald Avery Fisher, 1890-1962.

- El *remuestreo* se refiere a un conjunto de técnicas estadísticas, computacionalmente intensivas, que estiman la distribución de una población basadas en muestreo aleatorio *con reemplazo*.
- Se considera que una muestra aleatoria $X_1, X_2, \ldots, X_n$ como si fuera una *población* finita y se generan muestras aleatorias de la misma muestra para estimar características poblacionales y hacer inferencia de la población muestreada.
- Las técnicas de remuestreo permiten calcular *medidas de ajuste* (en términos de sesgo, varianza, intervalos de confianza, errores de predicción o de algunas otras medidas) a los estimados basados en muestras.
- Estas técnicas son usualmente *no paramétricas*, y varias son tan antiguas como la estadística misma. Por ejemplo, las técnicas de permutación son de Fisher (1935) y Pitmann (1937); la validación cruzadas fue propuesta por Kurtz en 1948, y el Jackknife fue propuesto por Maurice Quenouille en 1949 aunque John Tukey en 1958 fue quien le dió nombre a la técnica.

Contexto histórico

- Bradley Efron introdujo el Bootstrap en 1979, y sus estudiantes Rob Tibshirani y Trevor Hastie han aportado mucho a la ciencia estadística. Ofrecen un curso en Statistical Learning en la plataforma MOOC de la Universidad de Stanford.
- El término 'bootstrapping' se refiere al concepto de "pulling oneself up by one's bootstraps", frase que aparentemente se usó por primera vez en *The Singular Travels, Campaigns and Adventures of Baron Munchausen*



Bradley Efron en 2014.

Introducción I

- El objetivo del *remuestreo* es estimar alguna característica poblacional, representada por $\boldsymbol{\theta}$ (tal como media, mediana, desviación estándar, coeficientes de regresión, matriz de covarianza, etc.) basado en los datos.
- También interesan las propiedades de la distribución de estimador, sin hacer supuestos restrictivos sobre la forma de la distribución de los datos originales.
- Para una muestra aleatoria $X_1, \ldots, X_n$, la distribución de remuestreo es la distribución empírica $\hat{F}_n$, que asigna probabilidad $1/n$ a cada una de las observaciones de la muestra.
- Se consideran como remuestreo, entre otras, las siguientes técnicas:
  - Jackknife
  - Bootstrap
  - Pruebas de permutación
  - Validación cruzada

  y los usos asociados a estas técnicas son los siguientes:
  - Estimación de sesgo de un estimador (jackknife, boostrap)
  - Reducción/corrección de sesgo de un estimador (jackknife, boostrap)
  - Pruebas estadísticas exactas (pruebas de permutación)
  - Validación de modelos (validación cruzada)

Ejemplo Motivación I

Consideremos una muestra de 6 parejas. La variable de interés es la diferencia del ingreso de los miembros de cada pareja (en miles de pesos, al mes)

i	P_1	P_2	$d_i = P_1 - P_2$
1	24	18	6
2	14	17	-3
3	40	35	5
4	44	41	3
5	35	37	-2
6	45	45	0

- Sea $\theta =$ promedio de las diferencias de ingreso poblacional. Podemos estimar θ con $\hat{\theta} = \bar{d} = \frac{6-3+5+3-2+0}{6} = 1.5$. La desviación estándar de $\hat{\theta}$ es $sd(d) = \sigma/\sqrt{n}$, donde σ^2 es la varianza poblacional.
- Si σ^2 es conocida, y si d_i tuviera distribución normal, un intervalo de confianza del 95 % para μ sería $\bar{d} \pm z_{0.975}\sigma/\sqrt{n}$.

Ejemplo Motivación II

- Si d_i no es normal, el resultado aún se cumple asintóticamente, pero en este problema, $n = 6$. Entonces, ¿cómo podemos concluir sobre $\bar{d}$ sin conocer la distribución de las observaciones?
- Como en la vida real usualmente no conocemos σ, tenemos que estimar también este parámetro a través del estimador de la desviación estándar, que es el *error estándar*:

$$\hat{\sigma} = s = \sqrt{\frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1}}$$

y el intervalo de confianza para μ cambia a $\bar{d} \pm t_{n-1,0.975} s/\sqrt{n}$.

- Ahora, supongamos que queremos estimar un parámetro diferente, por ejemplo, la mediana, $q_{.5} = F^{-1}(1/2)$. Un estimador de $q_{.5}$ es por ejemplo

$$\hat{q}_{.5} = \begin{cases} X_{\left(\frac{n+1}{2}\right)} & \text{si } n \text{ es impar} \\ \frac{X_{(n/2)} + X_{(n/2+1)}}{2} & \text{si } n \text{ es par} \end{cases}$$

En nuestro ejemplo $\hat{q}_{.5} = \frac{0+3}{2} = 1.5$

Ejemplo Motivación III

- Aquí no es evidente cómo obtener un estimador de la variabilidad de $\hat{q}_{.5}$. Para contar con estimadores de variabilidad de parámetros como éste es para lo que se usan las técnicas de remuestreo.
- De hecho, el estimador de la varianza de la mediana dependerá de la distribución de la que vengan los datos, al ser una estadística de orden (o de hecho el promedio de dos estadísticas de orden)

El jackknife y el bootstrap nos ofrecen opciones para hacer la extensión de la medición del error estándar de un estimador.

Bootstrap

Bootstrap

- Sea $\mathbf{x} = \{X_1, \ldots, X_n\}$ la muestra observada de una variable aleatoria $X \sim F$.
- La distribución empírica F_n es un estimador de F (de hecho, $F_n(x)$ es una estadística suficiente para $F(x)$[1]), que pone masa $1/n$ sobre cada observación de la muestra.
- Extraer una muestra al azar de la distribución empírica F_n de tamaño n es equivalente a seleccionar valores $x_1^*, \ldots, x_n^*$ del conjunto $\mathbf{x}$ *con reemplazo*.
- Entonces, en el bootstrap hay dos aproximaciones:
  1. La distribución empírica F_n aproxima a F, y
  2. la función de distribución empírica de las observaciones bootstrap F_n^* aproxima a F_n.

  Las aproximaciones se pueden representar en el diagrama:

$$F \implies X_1, \ldots, X_n \implies F_n \implies X_1^*, \ldots, X_n^* \implies F_n^*$$

  Entonces indirectamente podríamos estimar de F_n^* la distribución F.

[1] Dada una muestra $\mathbf{X} = X_1, \ldots, X_n$, una estadística $T(\mathbf{X})$ es suficiente para un parámetro θ si la distribución condicional de la muestra dado el valor de $T(\mathbf{X})$ no depende de θ.

Algoritmo bootrstrap

El siguiente algoritmo describe cómo obtener una muestra de observaciones bootstrap vía Montecarlo a partir de una muestra observada, para calcular el error estándar del estimador $\hat{\theta}$ de un parámetro θ.

Algoritmo bootstrap

Sea $\boldsymbol{\theta}$ el parámetro de interés de una distribución F, y $\hat{\boldsymbol{\theta}} = T(X_1, \ldots, X_n)$ un estimador de $\boldsymbol{\theta}$ obtenido a partir de una muestra $X_1, \ldots, X_n$. El *estimador bootstrap* de $\boldsymbol{\theta}$ se obtiene de la siguiente manera:

1. Para $b = 1, \ldots, B$:
   (a) Genera una muestra $\mathbf{x}^*_{(b)} = (x^*_1, \ldots, x^*_n)$ *con reemplazo* de la muestra observada $X_1, \ldots, X_n$.
   (b) Calcular $\hat{\boldsymbol{\theta}}^{(b)} = T(\mathbf{x}^*_{(b)})$ con la $b-$ésima muestra bootstrap.
2. El estimado bootstrap de F_n es la distribución empírica $F_{\hat{\boldsymbol{\theta}}^{(\cdot)}}$. de las replicaciones bootstrap $\hat{\boldsymbol{\theta}}^{(1)}, \ldots, \hat{\boldsymbol{\theta}}^{(B)}$.

A partir de la muestra bootstrap se pueden estimar características relevantes de la distribución del parámetro $\boldsymbol{\theta}$.

Ejemplo de las diferencias I

Siguiendo con el ejemplo de las diferencias, podemos encontrar la distribución de las muestras bootstrap para la mediana $\theta = q_{.5}$:

```
d <- c(3, 5, -3, 6, -2, 0)
quantile(d, 0.5)
```

```
50%
1.5
```

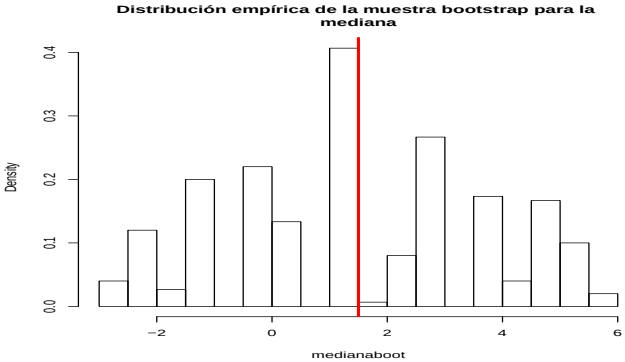
```
Boot <- NULL
B <- 300
Boot <- matrix(0, nrow = B, ncol = length(d))
for(i in 1:B) Boot[i,] <- sample(d, replace = T)
head(Boot)
```

```
     [,1] [,2] [,3] [,4] [,5] [,6]
[1,]    3    3    3    3    5    6
[2,]   -3    6    0    5    6   -3
[3,]   -2    6    0    0   -2    5
[4,]    5    3    5    0    3    5
[5,]   -3    3    6    0    3    6
[6,]    3    0    6    5   -2    6
```

Ejemplo de las diferencias II

```
medianaboot <- apply(Boot, 1, quantile,0.5)
mean(medianaboot)

[1] 1.741667

sd(medianaboot) #un estimador de la dispersión de la mediana

[1] 2.274425

hist(medianaboot, prob = T, main = "Distribución empírica de la muestra bootstrap para la
mediana", breaks = 30)
abline(v = quantile(d, 0.5), lwd = 3, col = "red")
```

Ejemplo de las diferencias III



Observaciones al bootstrap I

- Si F_n es un estimador pobre de F, entonces la distribución F_n^* de las replicas también será un estimador pobre de F, **aunque** F_n^* **estime bien a** F_n.
- Algunos parámetros pueden escribirse como *funcionales*[2] de F. Por ejemplo, si T es una funcional:

$$\boldsymbol{\theta} = T(F).$$

Algunos ejemplos son los siguientes:

1. $\mu = E_F(X) = \int x\, dF$ (la integral de Riemman-Stieltjes. Si F'(x)=f(x), entonces $\int x\, dF = \int x f(x)\, dx$)
2. $\sigma^2 = E_F((X-\mu)^2) = \int (x-\mu)^2 dF$
3. $\eta = P_F(x \in A) = \int_A dF$

- El siguiente principio es de particular importancia para los métodos de remuestreo.

Principio del plug-in:

Estima $\boldsymbol{\theta}$ del siguiente modo: si $\boldsymbol{\theta} = T(F)$, entonces $\hat{\boldsymbol{\theta}} = T(\hat{F})$.

Por ejemplo $\bar{X}$, s^2 y $\hat{\rho}_{XY}$ son estimados tipo plug-in.

Observaciones al bootstrap II

- Cuando hay información adicional sobre F que no viene de la muestra (por ejemplo, si se asume que F viene de una familia paramétrica), el principio *plug-in* es menos efectivo en general. Pero aun en estos casos el principio *plug-in* puede ser adoptado.

[2] Las funcionales son funciones de funciones, es decir su dominio es un espacio de funciones como $\mathcal{L}^p$

Estimación bootstrap del error estándar I

Algoritmo para estimar el error estándar de un estimador

Si $\boldsymbol{\theta}$ es el parámetro de interés y $\hat{\boldsymbol{\theta}} = T(\mathbf{X})$ un estimador, entonces para calcular el error estándar de $\hat{\boldsymbol{\theta}}$:

1. Extraer una muestra bootstrap $X_1^*, \ldots, X_n^* \sim \hat{F}$.
2. Calcular $\hat{\boldsymbol{\theta}}^* = T(X_1^*, \ldots, X_n^*)$.
3. Repetir los pasos 1 y 2, B veces para obtener la muestra bootstrap $\hat{\boldsymbol{\theta}}_1^*, \ldots, \hat{\boldsymbol{\theta}}_B^*$
4. Definir

$$\hat{se}_{boot}(\hat{\boldsymbol{\theta}}^*) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\boldsymbol{\theta}}_b^* - \hat{\boldsymbol{\theta}}^{*\cdot}\right)^2}$$

donde $\hat{\boldsymbol{\theta}}^{*\cdot} = \frac{1}{B}\sum_{r=1}^{B}\hat{\boldsymbol{\theta}}_r^*$ es el promedio de las muestras bootstrap.

- Por la ley de los grandes números, $v_{boot} \xrightarrow{cs} Var_{\hat{F}}(\hat{\theta})$ conforme $B \to \infty$.

Estimación bootstrap del error estándar II

- De acuerdo a Efron, usualmente basta que el número de muestras bootstrap B esté entre 50 y 200, para el estimador del error estándar.

Ejemplo de las diferencias

Siguiendo con el ejemplo de las diferencias, podemos encontrar el error estándar para la mediana $\theta = q_{.5}$:

```
medianaboot <- apply(Boot, 1, quantile,0.5) # recordar que Boot es la matriz donde se guardan las muestras bootstrap
medianaboot
```

```
  [1]  3.0  2.5  0.0  4.0  3.0  4.0  5.5 -2.0 -1.0  3.0  1.5  3.0 -1.0 -2.0  4.0
 [16]  0.5 -2.0  5.5  0.5  4.5  1.5  3.0  1.5  4.0  0.0  1.5  1.5  1.5  1.5  4.0
 [31] -1.0  3.0  4.0  0.5  4.0  5.0 -1.0  1.5 -2.5  2.5  1.5  3.0  0.5  3.0  5.0
 [46]  1.5  4.0  0.5  1.5 -1.0  5.0  0.5 -1.0  0.0  3.0  0.0  5.0 -2.5  0.0  0.5
 [61]  2.5  6.0  1.5 -2.0  5.0  0.0  0.0  4.0  1.5  5.0  3.0  3.0  0.0  0.0 -1.0
 [76]  4.0  5.5  3.0  0.0  3.0  3.0  1.5  0.0  2.5  5.5  0.5 -2.0 -2.5 -1.0  0.5
 [91]  1.5  2.5  0.0 -1.0 -1.0 -1.0  1.5  6.0  5.5  4.5  5.0 -2.0  0.5  4.5  3.0
[106] -1.0  0.0  0.0 -1.0 -1.0  5.0  0.0  1.5 -1.0 -1.0  1.5  5.0  1.5  1.5  4.5
[121] -2.0 -2.0  5.0  4.0  0.5  1.5 -2.0  1.5  5.5  1.5  5.0  4.0 -1.0  1.5  1.5
[136] -2.0 -1.0  2.5  5.0 -1.0  3.0  1.5 -1.0  2.5 -2.0  1.5 -1.0  0.0  3.0 -1.0
[151]  2.5  3.0  0.0  4.0  1.5  1.5 -2.0  1.5  3.0  5.0  3.0  0.5  0.0  3.0  3.0
[166]  5.5 -1.0  1.5  5.0  3.0  1.5 -1.0  0.0  3.0  0.0  3.0 -1.0  2.5  5.5  0.0
[181]  4.0  4.0  1.5  0.5 -3.0  1.5  5.5  3.0 -1.0 -1.0  3.0  1.5 -1.0  3.0  5.5
[196]  0.0  3.0  4.0  1.5 -1.5  1.5  1.5  4.0  4.0 -1.5  1.5  0.5  5.5  4.5  0.5
[211]  1.5  4.5  0.0  4.0  1.5  0.5 -2.0  3.0  0.0  1.5  5.0  1.5 -1.5  2.5  4.0
[226]  5.0  5.5  5.0  5.0  0.5  1.5  4.0  1.5  1.5  1.5  1.5  0.0 -1.0  3.0  5.0
[241]  5.5  5.0 -1.0  0.0  5.0  0.0  5.0  3.0  5.5  3.0  2.5  1.5  3.0 -2.0  1.5
[256]  0.5  3.0  5.5  3.0  1.5  1.5  0.0  0.0  2.5  4.0  1.5  0.0 -1.5  3.0 -2.0
[271]  0.5  4.0 -2.0 -2.0  1.5  1.5 -2.5  2.0  3.0  5.0  5.0  1.5  3.0  1.5  0.0
[286]  4.0  1.5  1.5  4.0  6.0  0.5  1.5  5.0  0.0  0.0 -2.5  4.0  3.0 -2.0  1.5
```

```
#Error estándar de T es simplemente la desviación estándar de las muestras boot de T
sd(medianaboot)
```

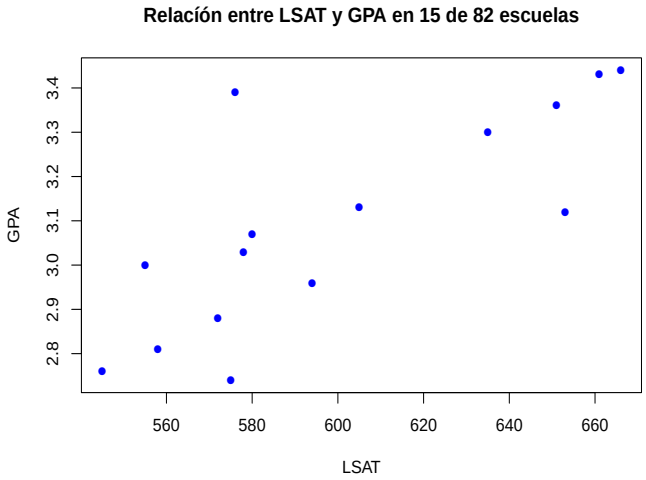
```
[1] 2.274425
```

Ejemplo: Correlación (Efron & Tibshirani) I

- De una población de $N = 82$ escuelas americanas de leyes, se toma una muestra de tamaño $n = 15$. Se miden los resultados promedios de dos scores: LSAT (promedio de evaluaciones de la prueba de admisión a Leyes) y GPA (promedio de licenciatura).
- Los datos están en la variable `law` que se obtiene al cargar el paquete `boostrap`. En este ejercicio el parámetro de interés es el coeficiente de correlación ρ. A partir de la muestra original obtenemos el coeficiente de correlación muestral, $\hat{\rho} = 0.776$.
- Noten también que la correlación muestral es un estimador *plug-in*.

```
library(bootstrap)
data("law")
plot(law, main="Relación entre LSAT y GPA en 15 de 82 escuelas", pch=16, col = "blue")
```

Ejemplo: Correlación (Efron & Tibshirani) II



Ejemplo: Correlación (Efron & Tibshirani) III

```
cor(law)
```

```
          LSAT       GPA
LSAT 1.0000000 0.7763745
GPA  0.7763745 1.0000000
```

Ejemplo: Correlación (Efron & Tibshirani) I

- ¿Cómo podemos estimar la varianza de ρ? Obtengamos un estimador bootstrap del error estándar del estimador del coeficiente de correlación.
- Para esto, necesitamos obtener muestras boostrap del estimador. Lo cual implica obtener muestras con reemplazo de los 15 pares de puntos.

```
B <- 400   # Un número suficientemente grande
n <- nrow(law) #tamaño de la muestra, 15
R <- numeric(B)

for (b in 1:B){
        i <- sample(1:n,size=n, replace=T)
        LSAT <- law$LSAT[i]
        GPA <- law$GPA[i]
        R[b] <- cor(LSAT,GPA)
}
head(R)   # Algunas observaciones de rho estimadas
```

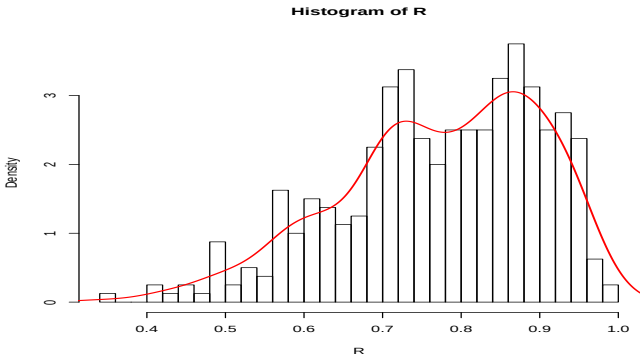
```
[1] 0.7267477 0.8477111 0.8453010 0.8751269 0.7896185 0.8580472
```

```
sd(R)     # error estándar
```

```
[1] 0.1277658
```

```
hist(R, prob = T, breaks = 30)
lines(density(R), col = "red", lwd = 2)
```

Ejemplo: Correlación (Efron & Tibshirani) II



- El estimador bootstrap del error estándar de $\hat{\rho}$ es 0.1277658 y claramente su distribución no es normal. La teoría dice que el estimador basado en normalidad de (X, Y) del error estándar de $\hat{\rho}$ es

$$\hat{\sigma}(\hat{\rho}) = \frac{1 - \hat{\rho}^2}{\sqrt{n-3}} = 0.115$$

Ejemplo: Correlación (Efron & Tibshirani) III

- En este caso, el tamaño de muestra no nos permite evaluar la normalidad.

Ejemplo: Correlación con el paquete `boot` I

- En ejemplos más complejos, donde se requiera hacer el cálculo de una estadística más complicada, la obtención de las muestras Montecarlo puede ser más complicado.
- El paquete de `R` `boot` (desarrollado por Davison y Hinkley) facilita la programación.
- En este paquete se requiere definir a la estadística $\boldsymbol{\theta}$ como función de la muestra.

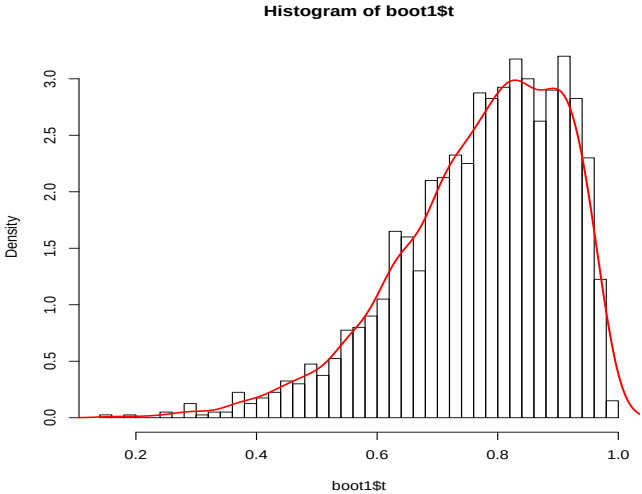
```
library(boot)

rho <- function(x,i){cor(x[i,1],x[i,2])}  # Definición de la estadística como función de la muestra
boot1 <- boot(data = law, statistic = rho, R=2000)
boot1


ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = law, statistic = rho, R = 2000)


Bootstrap Statistics :
     original       bias    std. error
t1* 0.7763745 -0.003054814   0.1363425
names(boot1)
 [1] "t0"        "t"         "R"         "data"      "seed"      "statistic"
 [7] "sim"       "call"      "stype"     "strata"    "weights"
```

Ejemplo: Correlación con el paquete `boot` II

- Las réplicas bootstrap están en la variable `boot1$t`. La notación del paquete `boot` y del paquete `bootstrap` (basado en Efron y Tibshirani) son diferentes.

```
hist(boot1$t, prob = T, breaks = 50)
lines(density(boot1$t), col = "red", lwd = 2)
```

Ejemplo: Correlación con el paquete boot III



Estimación bootstrap del sesgo I

Sesgo

Si $\hat{\boldsymbol{\theta}}$ es un estimador insesgado de $\boldsymbol{\theta}$, entonces $E(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}$. El sesgo de un estimador se define como

$$sesgo(\hat{\boldsymbol{\theta}}) = E(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta}$$

- Por ejemplo, sabemos que si X viene de una población con media μ y varianza σ^2, entonces $E(\bar{X}) = \mu$, por lo que $sesgo(\bar{X}) = 0$, mientras que el estimador plug-in de σ^2, que es $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$, cumple con $E(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2$, por lo que $sesgo(\hat{\sigma}^2) = -\frac{\sigma^2}{n}$.
- La estimación bootstrap del sesgo es el estimador *plug-in* del sesgo.

Estimador bootstrap de sesgo

$$\widehat{sesgo}_{boot}(\hat{\boldsymbol{\theta}}) = \hat{\boldsymbol{\theta}}^{*\cdot} - \hat{\boldsymbol{\theta}}$$

donde $\hat{\boldsymbol{\theta}}$ es el estimador del parámetro con la muestra original y $\hat{\boldsymbol{\theta}}^{*\cdot} = \frac{1}{B}\sum_{r=1}^{B}\hat{\boldsymbol{\theta}}_r^*$ es el promedio de las muestras bootstrap.

Ejemplo: Sesgo en la estimación de la correlación datos `law`

Estimamos el sesgo del estimador del coeficiente de correlación

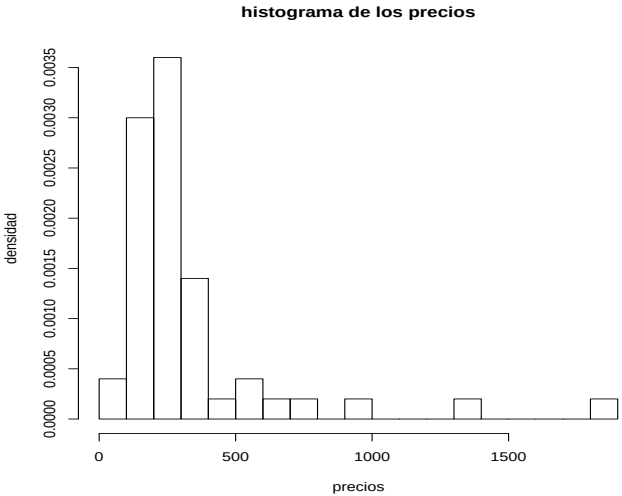
```
        thetahat <- cor(law$LSAT, law$GPA)
    thetahat

[1] 0.7763745

        #estimacion del sesgo:
        B <- 500 #Para la estimación de sesgo usualmente B tiene que ser mayor
        theta.b <- NULL

        for(i in 1:B){
                j <- sample(1:n, size = n, replace = T)
                LSAT <- law$LSAT[j]
                GPA <- law$GPA[j]
                theta.b[i] <- cor(LSAT, GPA)
        }
        sesgo <- mean(theta.b) - thetahat
        sesgo

[1] -0.0132109
```

Ejemplo: precios y distribución de la media recortada I

Los datos siguientes corresponden a una muestra de precios de venta de propiedades en Seattle en 2002. Los datos no distinguen entre propiedades residenciales que son la mayoría, pero hay algunas comerciales en la muestra, lo cual puede incrementar el precio promedio de venta muestral. Una estadística más resistente a valores extremos es la media recortada.

```
precios <- c(142, 175, 197.5, 149.4, 705, 232, 50, 146.5, 155, 1850,
             132.5, 215, 116.7, 244.9, 290, 200, 260, 449.9, 66.407, 164.95,
             362, 307, 266, 166, 375, 244.95, 210.95, 265, 296, 335,
             335, 1370, 256, 148.5, 987.5, 324.5, 215.5, 684.5, 270, 330,
             222, 179.8, 257, 252.95, 149.95, 225, 217, 570, 507, 190)
mean(precios)

[1] 329.2571

mean(precios, trim = 0.25)

[1] 244.0019

hist(precios, probability = T,
     main = "histograma de los precios", ylab = "densidad", breaks = 20)
```

Ejemplo: precios y distribución de la media recortada II

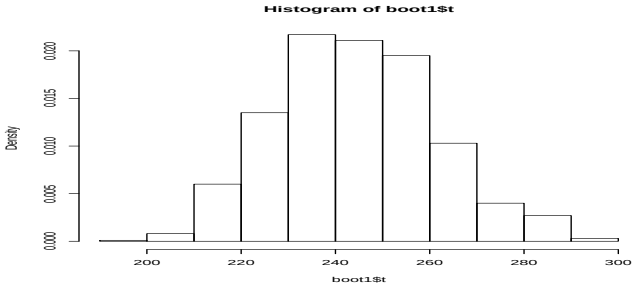


Ejemplo: precios y distribución de la media recortada I

- ¿Qué podemos decir de la distribución muestral de $\bar{x}_{25\,\%}$? No mucho. Pero podemos estimar lo que necesitamos con bootstrap.

```
media.recortada <- function(x,i){mean(x[i],trim=0.25)}
boot1 <- boot(data = precios, statistic = media.recortada, R=1000)
boot1


ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = precios, statistic = media.recortada, R = 1000)


Bootstrap Statistics :
    original    bias    std. error
t1* 244.0019 0.3567558    16.92816
hist(boot1$t, prob = T) #gráfica de las replicas bootstrap
```

Ejemplo: precios y distribución de la media recortada II



- En el reultado, se puede concluir que la forma de la distribución de la media recortada es parecida a la normal.
- La estimación del sesgo es 0.3567558, que es un sesgo pequeño relativo al tamaño de la estadística. La estadística (la media recortada de la muestra) tiene un sesgo pequeño como un estimado del parámetro (que es la media recortada de la población).
- El estimador bootstrap del error estándar es 16.9281574.

Jackknife

Jackknife

- El *Jackknife* es un método de remuestreo propuesto por Quenouille (1949) para estimar sesgo, y por Tukey (1958) para estimar error estándar.
- *Jackknife* es un tipo de *validación cruzada* en donde se deja un dato fuera a la vez: si $\mathbf{x} = (x_1, \ldots, x_n)$ es una muestra aleatoria, hay n muestras jackknife $\mathbf{x}_{-i}$ que es la muestra original sin la i-ésima observación.
- Si $\hat{\boldsymbol{\theta}} = T_n(\mathbf{x})$, entonces la réplica i-ésima jackknife es $\hat{\boldsymbol{\theta}}_{-i} = T_{n-1}(\mathbf{x}_{-i})$.
- Se puede generalizar a considerar g grupos de tamaño h tales que $g \cdot h = n$ y se calcula la i-ésima réplica jacknife dejando uno de los g grupos afuera.

Estimador jackknife del sesgo I

Estimador Jackknife del sesgo

El estimador jackknife del sesgo es

$$sesgo_{jack} = (n-1)\left(\frac{\sum_{i=1}^{n}\hat{\boldsymbol{\theta}}_{-i}}{n} - \hat{\boldsymbol{\theta}}\right)$$

y el estimador corregido por sesgo es:

$$\hat{\boldsymbol{\theta}}_{jack} = \hat{\boldsymbol{\theta}} - sesgo_{jack} = n\hat{\boldsymbol{\theta}} - \frac{n-1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{\theta}}_{-i}$$

Ejemplo jackknife

En nuestro ejemplo inicial de diferencias de sueldos, $\hat{\theta} = \bar{d} = 1.5$, y el sesgo está dado por:

```
d <- c(6, -3, 5, 3, 0, -2)
dajus <- NULL #guarda los valores ajustados
for(i in 1:length(d)) dajus[i] <- mean(d[-i])
mean(dajus)

[1] 1.5

sesgo <- 3*(mean(dajus)-mean(d))
sesgo

[1] 0

dajus

[1] 0.6 2.4 0.8 1.2 1.8 2.2
```

Ejemplo sesgo Jackknife I

Consideren una muestra aleatoria $X_1, \ldots, X_n \sim$ **Bernoulli** (θ). Queremos estimar θ^2.

- El estimador máximo verosimil de θ^2 es $\bar{x}^2$. Sin embargo, este estimador de θ^2 tiene sesgo, el cuál es fácil de calcular considerando que $Y = \sum_{i=1}^{n} X_i \sim$ **Bin** (n, θ) y $E(Y^2) = n\theta(1-\theta) + n^2\theta^2$:

$$E(\hat{\theta}^2) = \theta^2 + \frac{\theta(1-\theta)}{n}$$

- De acuerdo a las definiciones previas, el estimador jackknife está dado por:

$$\begin{aligned} \hat{\theta}^2_{jack} = \bar{x}^2 - sesgo_{jack} &= \bar{x}^2 - (n-1)\left(\frac{1}{n}\sum_{i=1}^{n} \bar{x}^2_{-i} - \bar{x}^2\right) \\ &= n\bar{x}^2 - \frac{n-1}{n}\sum_{i=1}^{n} \bar{x}^2_{-i} \end{aligned}$$

Ejemplo sesgo Jackknife II

- numéricamente:

```
set.seed(1)
n <- 100  #tamaño de muestra
x <- rbinom(n,size=1,p=0.3)
x
```

```
  [1] 0 0 0 1 0 1 1 0 0 0 0 0 0 0 1 0 1 1 0 1 1 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 1
 [38] 0 1 0 1 0 1 0 0 1 0 0 1 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 1 0 1 0 0
 [75] 0 1 1 0 1 1 0 1 0 0 1 0 1 0 0 0 0 0 0 1 1 1 0 0 1 0
```

```
thetahat2 <- mean(x)^2  #estimador máximo verosímil
thetahat2
```

```
[1] 0.1024
```

```
xj <- NULL #calcula los estimados jack
for(i in 1:n) xj[i] <- mean(x[-i])^2
sesgo <- (n-1)*(mean(xj)-thetahat2)
sesgo
```

```
[1] 0.00219798
```

```
thetahat.jack <- n*thetahat2-(n-1)*sum(xj)/n #Estimador jack corregido
thetahat.jack
```

```
[1] 0.100202
```

Notas al Jackknife I

- Si definimos $\tilde{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}_{-i}$ como *pseudovalores*, entonces $\hat{\theta}_{jack}$ es el promedio de estos pseudovalores:

$$\hat{\theta}_{jack} = \frac{1}{n}\sum_{i=1}^{n}\tilde{\theta}_i$$

  En el ejemplo considerado, los pseudovalores son iguales a los valores originales.

- El estimador jackknife tiene un factor $n-1$. Para ver su origen, consideren el caso donde $\theta = \sigma^2$ la varianza de la población. El estimador plug-in de la varianza es:

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

  Este estimador es sesgado para σ^2 con sesgo:

$$sesgo(\hat{\theta}) = E(\hat{\theta} - \sigma^2) = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}.$$

Notas al Jackknife II

Cada replicación jackknife calcula el estimado $\hat{\theta}_{(i)}$ sobre una muestra de tamaño $n-1$, así que el sesgo en la replicación jackknife es $-\sigma^2/(n-1)$. Entonces

$$\begin{aligned} E(\hat{\theta}_{(i)} - \hat{\theta}) &= E(\hat{\theta}_{-i} - \theta) - E(\hat{\theta} - \theta) \\ &= sesgo(\hat{\theta}_{-i}) - sesgo(\hat{\theta}) \\ &= -\frac{\sigma^2}{n-1} - (-\frac{\sigma^2}{n}) \\ &= -\frac{\sigma^2}{n(n-1)} = \frac{sesgo(\hat{\theta})}{n-1} \end{aligned}$$

Entonces el estimador jackknife con factor $n-1$ da un estimado correcto de sesgo en el estimador plugin de la varianza.

- El estimador jackknife puede fallar cuando la estadística $\hat{\boldsymbol{\theta}}$ no es una función "suave" (en el sentido usual del cálculo). Por ejemplo, la mediana $q_{.5}$ no es una función suave de los datos.

Estimador jackknife de error estándar

Estimador Jackknife de varianza

El estimador jackknife del error estándar del estimador es:

$$\hat{se}_{jack} = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\boldsymbol{\theta}}_{-i} - \frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{\theta}}_{-i}\right)^2}$$

Ejemplo sesgo y error estándar jackknife I

Las siguientes corresponden a mediciones de cierta hormona en el torrente sanguíneo de ocho sujetos después de ponerse un parche médico (Efron y Tibshirani). El parámetro de interés es:

$$\theta = \frac{E(new) - E(old)}{E(old) - E(placebo)}$$

Si $|\theta| \leq 0.20$, esto indica bioequivalencia entre los parches viejo y nuevo. La estadística es $\bar{Y}/\bar{Z}$.

```
data(patch, package = "bootstrap")
patch
```

```
  subject placebo oldpatch newpatch     z     y
1       1    9243    17649    16449  8406 -1200
2       2    9671    12013    14614  2342  2601
3       3   11792    19979    17274  8187 -2705
4       4   13357    21816    23798  8459  1982
5       5    9055    13850    12560  4795 -1290
6       6    6290     9806    10157  3516   351
7       7   12412    17208    16570  4796  -638
8       8   18806    29044    26325 10238 -2719
```

A continuación calculamos el sesgo y el error estándar jackknife.

Ejemplo sesgo y error estándar jackknife II

```r
n <- nrow(patch)
y <- patch$y
z <- patch$z
theta.hat <- mean(y)/mean(z)
theta.hat
```

```
[1] -0.0713061
```

```r
#Calculamos las replicaciones jackknife
theta.jack <- numeric(n)
for (i in 1:n) theta.jack[i] <- mean(y[-i])/mean(z[-i])
sesgo <- (n-1)*(mean(theta.jack)-theta.hat)
sesgo
```

```
[1] 0.008002488
```

```r
#error estándar:
se <- sqrt((n-1)*mean((theta.jack - mean(theta.jack))^2))
se
```

```
[1] 0.1055278
```

Jackknife after bootstrap I

- Los estimadores bootstrap de error estándar y de sesgo son variables aleatorias. Por lo tanto, podemos calcular la variabilidad de estos estimadores. Podemos usar el jackknife para estimarlo
- Para obtener un estimador jackknife de la desviación estándar de esas variables aleatorias, aplicamos el siguiente procedimiento: Considerando que x_i es la observación que se deja afuera,
  1. Sea $J(i)$ el conjunto de índices de las muestras bootstrap que no contienen x_i y sea $B(i)$ el número de muestras bootstrap que no contienen x_i.
  2. Se calcula la réplica jackknife dejando fuera las $B - B(i)$ muestras que contienen a x_i con la fórmula

$$\hat{se}_{jack}(\hat{se}_{B(1)}, \ldots, \hat{se}_{B(n)})$$

donde

$$\hat{se}_{B(i)} = \sqrt{\frac{1}{B(i)} \sum_{j \in J(i)} \left[\hat{\theta}_{-j} - \frac{1}{B(i)} \sum_{j \in J(i)} \hat{\theta}_{-j} \right]^2}$$

Ejemplo I

Para los datos de los parches, estimamos el error estándar del error estándar para $\hat{\theta}$.

```
data(patch, package = "bootstrap")
n <- nrow(patch)
y <- patch$y
z <- patch$z
B <- 2000
theta.b <- numeric(B)

#almacenamiento de los índices muestreados
indices <- matrix(0,nrow=B, ncol=n)

#jackknife-after-bootstrap paso 1: corre el bootstrap
for (b in 1:B) {
        i <- sample(1:n, size = n, replace = T)
        y <- patch$y[i]
        z <- patch$z[i]
        theta.b[b] <- mean(y)/mean(z)
        #se guardan los índices para el jackknife
        indices[b, ] <- i
}
```

Ejemplo II

```r
#jackknife-after-bootrstrap para estimar se(se)
se.jack <- numeric(n)
for (i in 1:n) {
        # en la i-ésima réplica omitir todas las muestras con x[i]
        usar <-(1:B)[apply(indices, MARGIN = 1, FUN      = function(k) {!any(k==i)})]
        se.jack[i] <- sd(theta.b[usar])
}

sd(theta.b)

[1] 0.09848677

sqrt((n-1)*mean((se.jack-mean(se.jack))^2))

[1] 0.02841287
```

El estimador bootstrap del error estándar es 0.0984868 y el estimador jackknife-after-bootstrap de su error estándar es 0.0284129.

Intervalos de confianza bootstrap I

Hay diferentes maneras de calcular intervalos de confianza bootstrap. Varian en su facilidad de cómputo y en su exactitud. Estos se basan en el error estándar bootstrap $\hat{se}_{boot} = \sqrt{v_{boot}}$. Los métodos de intervalos de confianza incluyen:

- bootstrap normal estándar
- bootstrap básico
- bootstrap percentil
- bootstrap t

A continuación revisaremos los supuestos asociados con cada uno de estos casos:

Bootstrap normal estándar

Usualmente este intervalo no es adecuado a menos que $\hat{\theta}$ tenga una distribución parecida a la normal.

Intervalo de confianza bootstrap normal estándar

Si $\hat{\theta}$ es un estimador del parámetro θ y se supone que el error estándar del estimador es $se(\hat{\theta})$, entonces si $\hat{\theta}$ es una media muestral y el tamaño de muestra es grande, el CLT implica que $Z = \frac{\hat{\theta} - E(\hat{\theta})}{se(\hat{\theta})}$ es aproximadamente normal. Un intervalo aproximado de $100(1-\alpha)$ % de confianza para θ es

$$\hat{\theta} \pm z_{\alpha/2} \cdot \hat{se}(\hat{\theta})$$

Intervalo de confianza bootstrap básico I

Este intervalo transforma la distribución de las replicaciones sustrayendo la estadística observada. Los cuantiles de la muestra transformada se utilizan para determinar los intervalos de confianza:

Confianza bootstrap básico

El intervalo de $100(1-\alpha)$ % de confianza está dado por:

$$(2\hat{\theta} - \hat{\theta}_{1-\alpha/2}, 2\hat{\theta} - \hat{\theta}_{\alpha/2})$$

En el caso paramétrico, se tiene que si T es un estimador de θ,

$$P(T - \theta > q_\alpha) = 1 - \alpha \implies P(T - q_\alpha > \theta) = 1 - \alpha$$

Entonces, un intervalo de confianza con $100(1-2\alpha)$ % con errores α en las colas inferior y superior iguales está dado por $(t - q_{1-\alpha}, t - q_\alpha)$.
En el boostrap la distribución de T es generalmente desconocida, pero los cuantiles se pueden estimar por un método aproximado:

Intervalo de confianza bootstrap básico II

1. Calcular los cuantiles muestrales $\hat{\theta}_\alpha$ de la función de distribución empírica de las replicaciones $\hat{\theta}^*$.
2. Denotar los cuantiles α de $\hat{\theta}^* - \hat{\theta}$ por b_α. Entonces $\hat{b}_\alpha = \hat{\theta}_\alpha - \hat{\theta}$ es un estimador de b_α
3. Un límite de confianza superior para el intervalo de $100(1-\alpha)$ % está dado por

$$\hat{\theta} - \hat{b}_{\alpha/2} = \hat{\theta} - (\hat{\theta}_{\alpha/2} - \hat{\theta}) = 2\hat{\theta} - \hat{\theta}_{\alpha/2}.$$

4. Similarmente, un límite de confianza inferior para el intervalo de $100(1-\alpha)$ % está dado por

$$\hat{\theta} - \hat{b}_{1-\alpha/2} = \hat{\theta} - (\hat{\theta}_{1-\alpha/2} - \hat{\theta}) = 2\hat{\theta} - \hat{\theta}_{1-\alpha/2}.$$

Intervalo percentil bootstrap

Los intervalos percentíl se basan en la distribución empírica de las réplicas bootstrap como la distribución de referencia .

Intervalo percentil

A partir de la distribución bootstrap de $\hat{\theta}^*$ se puede calcular un intervalo de confianza no paramétrico. El $(1-\alpha)\times 100$ % intervalo basado en percentiles es

$$(\hat{\theta}^*_{(B\cdot\alpha/2)}, \hat{\theta}^*_{(B\cdot(1-\alpha/2))})$$

basado en los cuantiles $B\cdot\alpha/2$ y $B\cdot(1-\alpha/2)$ de la muestra bootstrap.

- Efron y Tibshirani muestran que los intervalos percentil tienen ventajas teóricas sobre los intervalos estándar normales y un mejor desempeño de cobertura.

Ejemplo de intervalos de confianza I

Para los datos del experimento del parche, obtendremos intervalos de confianza normal, básico, y percentil usando las funciones de `R boot` y `boot.ci` en el paquete `boot`

```
library(boot)
data(patch,package="bootstrap")
theta.boot <- function(datos,indice){
        #función para calcular la estadística
        y <- datos[indice,1]
        z <- datos[indice,2]
        mean(y)/mean(z)
}
y <- patch$y #lee los datos
z <- patch$z
datos <- cbind(y,z)
boot.obj <- boot(datos,statistic=theta.boot,R=2000)
```

Los resultados de los intervalos de confianza se dan a continuación:

Ejemplo de intervalos de confianza II

```
boot.obj


ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = datos, statistic = theta.boot, R = 2000)


Bootstrap Statistics :
     original      bias    std. error
t1* -0.0713061 0.007534835   0.1023291

boot.ci(boot.obj,type=c("basic","norm","perc"))

BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 2000 bootstrap replicates

CALL :
boot.ci(boot.out = boot.obj, type = c("basic", "norm", "perc"))

Intervals :
Level      Normal              Basic              Percentile
95%   (-0.2794,  0.1217 )   (-0.3038,  0.0897 )   (-0.2323,  0.1612 )
Calculations and Intervals on Original Scale
```

Intervalos BC_a

- Los intervalos percentil pueden ser mejorados con ciertos métodos de ajuste de percentiles. El más popular es el método acelerado con corrección de sesgo, BC_a (*bias-corrected, accelerated* en inglés)
- Los intervalos BC_a tienen mejores propiedades teóricas y mejor desempeño en la práctica.
- Para un intervalo de $100(1-\alpha)$ % de confianza, los cuantiles $\alpha/2$ y $1-\alpha/2$ se ajustan por dos factores: una corrección por sesgo y una por asimetría (skewness).
- La corrección por sesgo es z_0 y el ajuste de la asimetría o "aceleración" es a.
- El factor de aceleración recibe su nombre porque estima la tasa de cambio del error estándar de $\hat{\theta}$ con respecto al parámetro objetivo θ (en una escala normalizada).

Intervalos BC_a

Para calcular un intervalo de confianza de $100(1-\alpha)$ % para θ, realizamos los siguientes ajustes:

1. Calcular $z_0 = \Phi^{-1}(\frac{\#\{\hat{\theta}_i^* < \hat{\theta}\}}{B})$. Es el efecto del sesgo, mide el sesgo de la mediana de las replicaciones bootstrap.
2. Calcular
$$a = \frac{\sum_{b=1}^{B}(\hat{\theta}_{(i)} - \bar{\theta}_{(\cdot)})^3}{6\left[\sum_{b=1}^{B}(\hat{\theta}_{(i)} - \bar{\theta}_{(\cdot)})^2\right]^{3/2}}$$
Este cálculo corresponde al cociente del estimador de aceleración.
3. Define $\alpha_1 = \Phi\left[z_0 + \frac{z_0 - z_{\alpha/2}}{1-a(z_0-z_{\alpha/2})}\right]$ y $\alpha_2 = \Phi\left[z_0 + \frac{z_0 + z_{\alpha/2}}{1-a(z_0+z_{\alpha/2})}\right]$. El intervalo está dado por $(\hat{\theta}^*_{(l^*)}, \hat{\theta}^*_{(u^*)})$, donde $l^* = B\alpha_1$ y $u^* = B\alpha_2$.

Ejemplo intervalo BC_a

Para los datos del parche,

```
boot.ci(boot.obj,type=c("bca"))

BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 2000 bootstrap replicates

CALL :
boot.ci(boot.out = boot.obj, type = c("bca"))

Intervals :
Level       BCa
95%   (-0.2204,  0.1823 )
Calculations and Intervals on Original Scale
```

Intervalo bootstrap t. I

- Aun si la distribución de un estimador $\hat{\theta}$ es normal y $\hat{\theta}$ es un estimador insesgado de θ, la distribución normal no es exactamente correcta para la estadística Z porque estamos estimando $se(\hat{\theta})$. Tampoco se puede decir que sea t porque no conocemos la distribución de $se(\hat{\theta})$
- El intervalo bootstrap t no usa una distribución t como la distribución de referencia. En su lugar, se utiliza una distribución muestral "de tipo t" por remuestreo

Intervalo t

$$(\hat{\theta} - t^*_{1-\alpha/2}\hat{se}(\hat{\theta}), \hat{\theta} - t^*_{\alpha/2}\hat{se}(\hat{\theta}))$$

donde los valores se calculan con el siguiente procedimiento:

1. Se calcula la estadística observada $\hat{\theta}$
2. Para cada replicación b
   - se obtiene una muestra bootstrap $x^{(b)}$
   - se calcula $\hat{\theta}^{(b)}$

Intervalo bootstrap t. II

- se calcula o estima el error estándar $\hat{se}(\hat{\theta}^{(b)})$ (un estimado separado para cada muestra bootstrap: esto es, se obtienen boostraps de la muestra actual $x^{(b)}$, no de x.
- se calcula la estadística $t^{(b)} = \frac{\hat{\theta}^{(b)} - \hat{\theta}}{se(\hat{\theta}^{(b)})}$

3. La muestra de replicadas $t^{(1)}, \ldots, t^{(B)}$ es la distribución de referencia para bootstrap t. Encontrar los cuantiles $t^*_{\alpha/2}$ y $t^*_{1-\alpha/2}$.
4. Calcular $\hat{se}(\hat{\theta})$, la desviación estándar muestral de las replicadas $\hat{\theta}^{(b)}$.
5. Calcular $(\hat{\theta} - t^*_{1-\alpha/2}\hat{se}(\hat{\theta}), \hat{\theta} - t^*_{\alpha/2}\hat{se}(\hat{\theta}))$

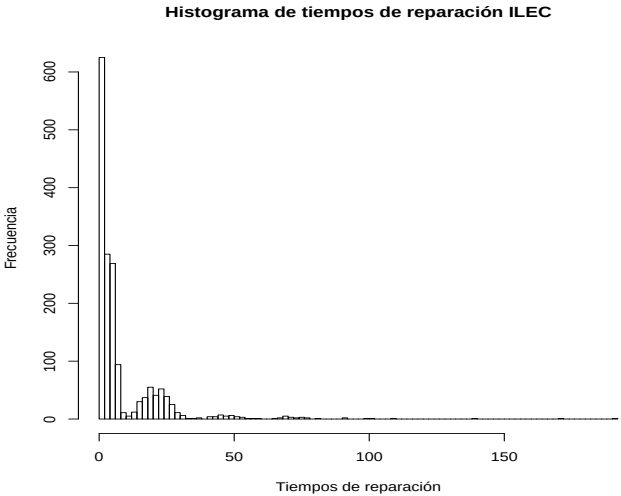
- La principal desventaja de este intervalo es que se tiene bootstrap anidado, lo que puede demandar tiempo de cómputo alto.

Ejemplos adicionales. Media poblacional I

- A continuación veremos la aplicación de Bootstrap, Jackknife, con varios ejemplos.
- Los siguientes datos corresponden a los tiempos de reparación de Verizon, compañía telefónica que actua en dos modos: como compañia primaria local telefónica (Incumbent Local Exchange Carrier ILEC) o como competidor en otras regiones (Competing Local Exchange Carrier, CLEC). Como ILEC Verizon debe proveer servicio de reparación para los clientes de las CLEC en su región. está sujeta a multas y la autoridad requiere el uso de pruebas de significancia para comparar los tiempos de reparación de los dos grupos de clientes.

```
library(resample) #paquete con los datos.


Attaching package: 'resample'
The following objects are masked from 'package:bootstrap':
    bootstrap, jackknife

data(Verizon)
ilec <- Verizon$Time[Verizon$Group=="ILEC"] #Toma los tiempos de reparación para la competencia
hist(ilec,breaks = 100, main = "Histograma de tiempos de reparación ILEC",
             ylab = "Frecuencia",
                 xlab = "Tiempos de reparación")
```

Ejemplos adicionales. Media poblacional II



Ejemplos adicionales. Media poblacional III

Los datos no son normales. Sin embargo, sabemos que la media tiende a distribuirse como una normal.

```
n <- length(ilec)
summary(ilec)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.730   3.590   8.412   7.080 191.600

#Intervalo para la media:
mean(ilec) + c(-1,1)*qt(.975,df = n-1 ,lower.tail = T)*sd(ilec)/sqrt(n)

[1] 7.705276 9.117945
```

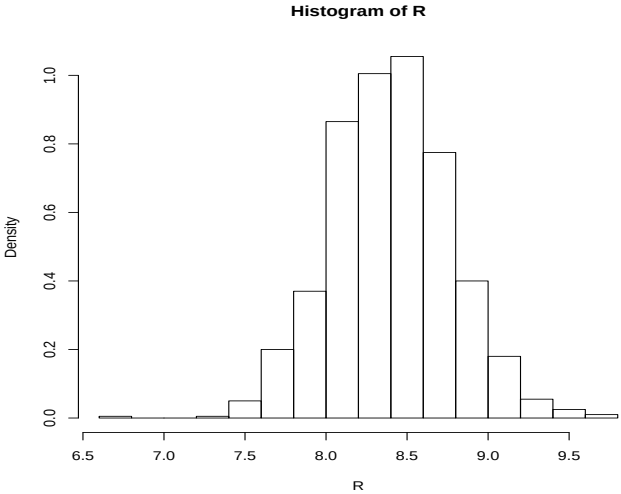
- Con una muestra bootstrap, se puede reproducir la forma y dispersión de la distribución. En el caso de la media es obvio, pero esto no es evidente para otras estadísticas. Ese es justamente la fortaleza del bootstrap, que nos permite obtener una estimación de la distribución de la estadística, no sólo para la media, sino para estadísticas mucho más complejas.

Ejemplos adicionales. Media poblacional IV

- La media de la distribución boostrap está centrada cerca de la media de la muestra obtenida, no la media poblacional. Así que la media de muestra boostrap tiene sesgo como estimador de la media poblacional. Sin embargo, el tamaño del sesgo de la estimación es parecido al sesgo que puede tener la media muestral respecto a la media poblacional.

```
B <- 1000
n <- length(ilec)
z <- numeric(B)
for(b in 1:B){
 i <- sample(1:n,replace=T)
 R[b] <- mean(ilec[i])
}
hist(R,prob=T)
```

Ejemplos adicionales. Media poblacional V



Ejemplos adicionales. Media poblacional VI

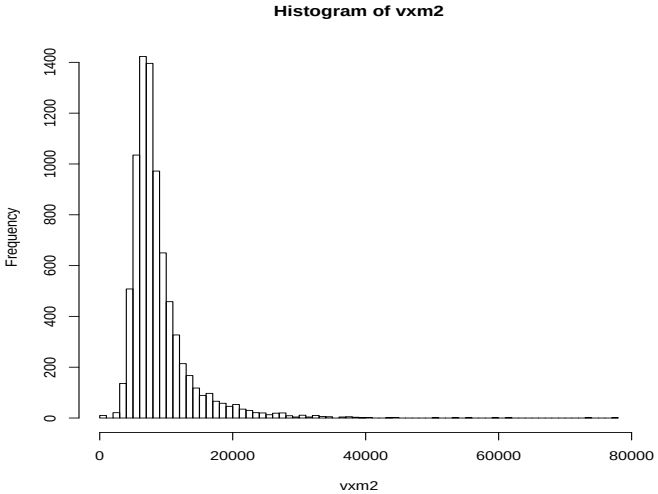
```
mean(R)

[1] 8.401446

se.boot <- sqrt(var(R))
se.boot

[1] 0.3671854
```

Ejemplo 2: Intervalos de confianza I

- Consideren los costos por metro cuadrado de vivienda en México. Los datos corresponden al Promedio de valor del mercado por metro cuadrado de construcción, que se puede obtener de la Sociedad Hipotecaria Federal (avaluos) a partir de avalúos.Los datos corresponden a 8,076 registros de todos los municipios de México registrados en 2015, y los datos consideran todo tipo de vivienda, desde aquella de interés social como la residencial.
- La variable de interés es `vxm2`, el valor de la vivienda por metro cuadrado de construcción, en pesos. Los datos están en el archivo `SHF-PromedioValorMercadoxMetroCuadradoConstruccion.csv` y se encuentran en el blog.

```
vxm2  <-
read.csv("~/Dropbox/Academia/ITAM/SimS19-II/data/SHF-PromedioValorMercadoxMetroCuadradoConstruccion.csv")$vxm2
hist(vxm2,breaks=100)
```

Ejemplo 2: Intervalos de confianza II



Ejemplo 2: Intervalos de confianza III

```
summary(vxm2)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
      0    6242    7616    8861    9848   77161 
```

- La distribución de los precios no es normal, y está 'contaminada' por diferentes tipos de propiedades. Para intentar corregir este efecto, consideremos un estimador robusto del parámetro de localización, que puede ser la media recortada del 25 % que es la media del 50 % de los datos que están en el centro de la distribución.Con bootstrap estimemos su distribución.

```
summary(vxm2)
```

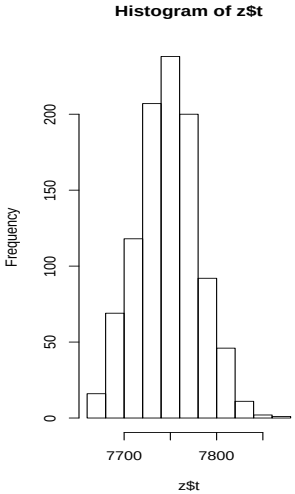
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
      0    6242    7616    8861    9848   77161 
```

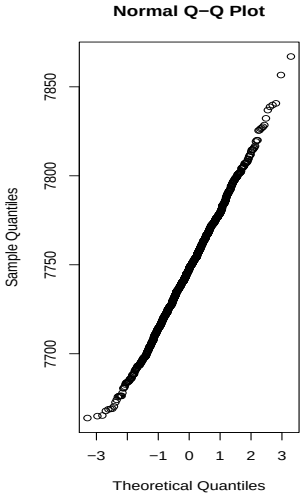
```
n <- length(vxm2)
mean(vxm2,trim = .25)
```

```
[1] 7747.663
```

```
z <- boot::boot(data = vxm2, statistic = function(x,i)mean(x[i],trim=.25), R = 1000)
par(mfrow=c(1,2))
hist(z$t)
qqnorm(z$t)
```

Ejemplo 2: Intervalos de confianza IV





Ejemplo 2: Intervalos de confianza V

```
z


ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot::boot(data = vxm2, statistic = function(x, i) mean(x[i],
    trim = 0.25), R = 1000)


Bootstrap Statistics :
    original     bias    std. error
t1* 7747.663 -0.1287377     33.0734
```

- La distribución boostrap es aproximadamente normal, tiene sesgo pequeño relativo al valor de la media, y tiene un error estándar de 33.0733961. Entonces podemos calcular un intervalo de confianza.

```
boot::boot.ci(z,type=c("norm","basic","perc"))

BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot::boot.ci(boot.out = z, type = c("norm", "basic", "perc"))

Intervals :
Level      Normal              Basic              Percentile
95%   (7683, 7813 )   (7684, 7811 )   (7684, 7812 )
Calculations and Intervals on Original Scale
```

Pruebas de permutación

Pruebas de Permutación

o pruebas de aleatorización

- Enfoque no paramétrico a pruebas de hipótesis estadísticas
- El paradigma usual:

Pruebas de hipótesis

- Define la hipótesis a probar: $H_0 : \theta \in \Theta$ vs. $\theta \notin \Theta$
- Observar datos: $x_1, \ldots, x_n$
- Define la estadística de prueba: $T_0 \sim F_0$ bajo H_0
- Determina si el valor de la estadística observada evaluada en los datos observados da un valor dentro de la región de rechazo: $T \in R_\alpha \implies$ rechaza H_0.

- Cuando la hipótesis nula es cierta, induce una distribución de probabilidad sobre la estadística de prueba. Esta distribución usualmente hace que las muestras sean *indistinguibles e intercambiables*.
- Bajo esta distribución nula, se puede construir la **distribución de permutación**.

Casos donde aplica la prueba de permutación I

- Las pruebas de permutación usualmente se aplican a el siguiente tipo de pruebas de hipótesis, con cualquiera de las estadísticas de prueba que se quieran usar:
  - Igualdad de poblaciones: $H_0 : F = G \quad$ vs $\quad H_a : F \neq G.$
  - Independencia: $H_0 : F_{X,Y} = F_X F_y \quad$ vs $\quad H_a : F_{X,Y} \neq F_X F_y$
  - Igualdad de más poblaciones: $H_0 : F_2 = \cdots = F_k \quad$ vs $\quad H_a : F_i \neq F_j$ para algun par de índices i, j.

Ejemplo Pruebas de permutación I

- Se tienen dos muestras y se quieren comparar sus poblaciones. La hipótesis nula supone igualdad de medias, $H_0 : \mu_1 = \mu_2$. Los datos observados son los siguientes:

```
x <- c(4714,4601,4696,4896,4905,4870,4987,5144,3962,4066,4561,4626,4924,5096,4321)
y <- c(4295,4271,4326,4530,4618,4779,4752,4744,3764,3797,4401,4339,4700)
```

- La estadística de prueba se basa en la distribución t (usualmente suponiendo que las muestras provienen de una distribución normal) y se consideran dos casos:
  - Varianzas iguales pero desconocidas: $\sigma_1 = \sigma_2$. En este caso, la estadística de prueba es de la forma:

$$t_0 = \frac{\bar{x} - \bar{y}}{S_p \sqrt{1/n_1 + 1/n_2}} \sim t_{(n_1+n_2-2)}$$

donde $S_p = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$ es la varianza combinada.

Ejemplo Pruebas de permutación II

- Varianzas diferentes y desconocidas: $\sigma_1 \neq \sigma_2$. En este caso, la estadística de prueba es de la forma (aproximada):

$$t_0 = \frac{\bar{x} - \bar{y}}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} \sim t_{(\nu)}$$

  y los grados de libertad se obtienen usando la aproximación de Satterthwaite:

$$\nu = \frac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{\frac{(S_1^2/n_1)^2}{n_1+1} + \frac{(S_2^2/n_2)^2}{n_2+1}} - 2$$

- En los dos casos anteriores, se tienen aproximaciones que pueden fallar si los datos no siguen una distribución normal.
- Las pruebas de permutación simplifican mucho el problema y se obtienen resultados más exactos.

Ejemplo de pruebas de permutación I

- Bajo la hipótesis nula, esperaríamos que las diferencias de medias muestrales $\hat{\theta} = \bar{x} - \bar{y}$ fueran cercanas a 0. Si H_0 no es cierta, entonces tenderíamos a observar valores "grandes" de $\hat{\theta}$. Habiendo observado $\hat{\theta}$, definimos el p-value como $pv = P(\hat{\theta}^* \geq \hat{\theta} | H_0)$ (también se denota como ASL, *achieved significance level*).
- Si H_0 es cierta, cualquiera de las observaciones pudo haber venido de cualquiera de las poblaciones. Podemos entonces combinar *todas* las $n_1 + n_2$ observaciones de las dos muestras[3], y extraer *sin reemplazo* n_1 de ellas para crear el primer grupo, y las n_2 observaciones restantes forman el otro grupo. Con estas nuevas muestras calculamos la estadística de prueba (cualquiera que sea para probar esta hipótesis) y repetimos el proceso B veces.
- En el siguiente ejemplo, escogemos la estadística t_0 que definimos previamente.

```
estadistica <- function(x,n1,n2){
s1 <- sample(x,n1,replace=F)
s2 <- x[-which(s1 %in% x)]
return((mean(s1)-mean(s2))/sqrt(var(s1)/n1+var(s2)/n2))
}
theta <- NULL
B <- 1000
for(i in 1:B) theta[i] <- estadistica(c(x,y),length(x),length(y))
ASL <-sum(theta > (mean(x)-mean(y))/sqrt(var(x)/length(x)+var(y)/length(y)))/B   #p-value
ASL
```

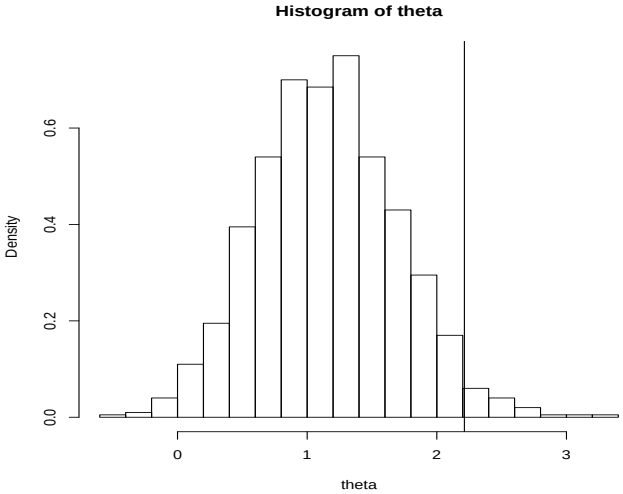
```
[1] 0.028
```

Ejemplo de pruebas de permutación II

Con este procedimiento ya no tenemos que aproximar los grados de libertad.

```
hist(theta,breaks=20,probability=T)
abline(v=quantile(theta,.975)) #agrega una linea en el quantile del p-value.
```

Ejemplo de pruebas de permutación III



[3] En este sentido son indistinguibles e intercambiables, como se comentó anteriormente

Conceptos detrás de las pruebas de permutación I

- Al considerar las permutaciones de las observaciones de los dos grupos hay $\binom{n_1+n_2}{n_1}$ posibles arreglos.
- Bajo la hipótesis nula, cada posible permutación tiene la misma probabilidad $1/\binom{n_1+n_2}{n_1}$ de ser seleccionada.
- Por lo tanto hay $\binom{n_1+n_2}{n_1}$ replicas posibles de permutación $\hat{\theta}^*$. El p-value de permutación se define como la probabilidad de que una réplica $\hat{\theta}*$ exceda el valor observado $\hat{\theta}$:

$$p\text{-val} = \#\{\hat{\theta}^* \geq \hat{\theta}\}/\binom{n_1+n_2}{n_1}$$

Conceptos detrás de las pruebas de permutación II

- En la práctica, el número de posibles permutaciones puede ser muy grande y computacionalmente demandante. Por eso, usualmente el valor del p-value se aproxima por muestreo, extrayendo un número grande de muestras al azar sin reemplazo.

Procedimiento aproximado para pruebas de permutación

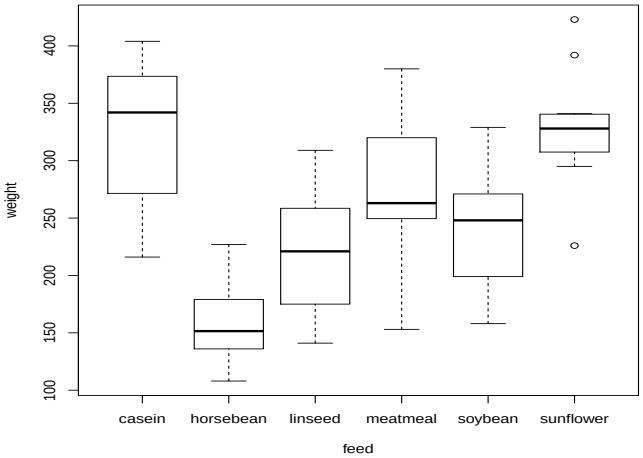
1. Calcular la estadística de prueba observada $t(\mathbf{x}, \mathbf{y})$
2. para las réplicas $b = 1, \ldots, B$:
   (a) Generar una permutación de índices π_b
   (b) Calcular la estadística de prueba $t^{(b)}$ basada en esa permutación de índices.
3. Calcular el p-value como $\hat{p} = \frac{1+\#\{t^{(b)} \geq t\}}{B+1}$
4. Rechaza H_0 al nivel de significancia α si $\hat{p} \leq \alpha$.

Ejemplos I

Los siguientes datos son datos de `R`, corresponden a los pesos, en gramos, para 6 grupos de pollitos alimentados en cada grupo con un tipo de suplemento alimenticio. En el boxplot que compara las medias de los suplementos se puede ver que los grupos `soybean` y `linseed` son similares.

```
data(chickwts)
boxplot(weight ~ feed,data= chickwts)
```

Ejemplos II


weight
100
150
200
250
300
350
400
casein
horsebean
linseed
meatmeal
soybean
sunflower
feed

Ejemplos III

Los pesos ordenados para esas dos muestras son los siguientes:

```
x <- with(chickwts,sort(weight[feed=="soybean"]))
y <- with(chickwts,sort(weight[feed=="linseed"]))
x
```

```
 [1] 158 171 193 199 230 243 248 248 250 267 271 316 327 329
```

```
y
```

```
 [1] 141 148 169 181 203 213 229 244 257 260 271 309
```

Ambos grupos se pueden comparar utilizando alguna estadística de prueba, basada en la media, mediana, media recortada, o cualquier otra medida de tendencia central. Por ejemplo, considerando las medias, y usando la estadística t bajo el supuesto de normalidad

Ejemplos IV

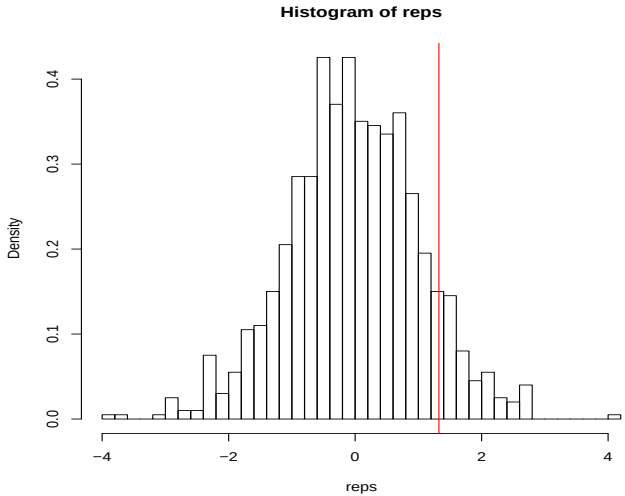
```
t.test(x=x,y=y)


Welch Two Sample t-test

data:  x and y
t = 1.3246, df = 23.63, p-value = 0.198
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -15.48547  70.84262
sample estimates:
mean of x mean of y
 246.4286  218.7500
```

Como las distribuciones de los pesos son desconocidas, sería mejor aplicar una prueba de permutación. Como los tamaños de muestra son $n_1 = 14$ y $n_2 = 12$, hay un total de $\binom{26}{14} = 9,657,700$ posibles particiones de la muestra combinada, por lo que es inadecuado enumerar todos los casos, y por lo que resulta conveniente tomar una muestra de tamaño razonable (entre 99 y 999 observaciones, regla de dedo):

Ejemplos V

```
B <- 999
z <- c(x,y) #muestra combinada
ind <- 1:26  #indices de la muestra combinada
reps <- NULL #vector donde se guardan las replicas
t0 <- t.test(x,y)$statistic

for (i in 1:B){
#genera índices para la primera muestra
k <- sample(ind,size=14, replace=F)
x1 <- z[k]
y1 <- z[-k]
reps[i] <- t.test(x1,y1)$statistic
}
p <- mean(c(t0,reps) >= t0)
p

[1] 0.095

hist(reps, prob=T,breaks=30)
abline(v=t0,col="red")
```

Ejemplos VI



Ejemplos VII

Para una prueba de dos colas, el ASL es $2\hat{p}$ si $\hat{p} \leq 0.5$ (y es $2(1-\hat{p})$ cuando $\hat{p} > 0.5$). En este ejemplo, $ASL = 0.19$. Por lo tanto, no se tiene evidencia para rechazar la hipótesis nula de igualdad de medias.

Validación Cruzada

Validación cruzada: ideas básicas I

- La validación cruzada (*cross-validation*) es una forma de medir el *desempeño predictivo* de un modelo estadístico.
  - Las estadísticas de ajuste de un modelo no son guía adecuada del poder predictivo del modelo. Por ejemplo, en regresión una R^2 alta no necesariamente indican que el modelo es bueno para predecir (se pueden incluir más términos para mejorar R^2 pero su poder predictivo empeora con el número de términos)
- Con la validación cruzada, podemos evaluar:
  1. la estabilidad de los parámetros estimados
  2. la exactitud de un problema de clasificación
  3. la adecuación de un modelo ajustado, etc.
- El jackknife es un caso particular de la validación cruzada.
- El enfoque de la validación cruzada es dividir los datos disponibles en dos conjuntos: un conjunto de *entrenamiento*, que se usa para estimar el modelo y un conjunto de *prueba*, en el que se evalúa el modelo y se obtiene un estimador del error de ajuste del modelo.
- Hay diversas maneras de hacer este procedimiento:

Validación cruzada: ideas básicas II

- **uno-afuera**. se usan $n-1$ datos para estimar el modelo. El modelo se prueba en el dato que se dejó afuera. Esto se puede realizar n veces se utilizan los errores $e_i^* = y_i - \hat{y}_i$ para calcular el error cuadrático medio de validación cruzada: $MSE_{cv} = \frac{\sum_{i=1}^{n} e_i^2}{n}$
- **k-afuera**.
- **Muestreo aleatorio**.

Ejemplo de validación cruzada I

- En el siguiente ejemplo, los datos x y y están relacionados, tienen correlación, pero la relación posiblemente no es lineal.

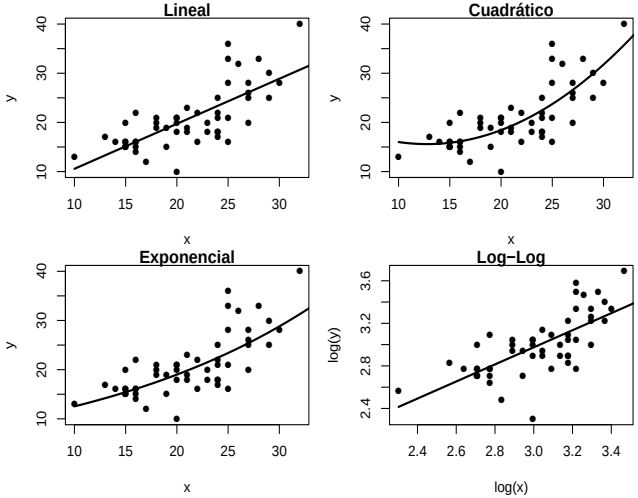
```
x
 [1] 24 16 24 18 18 10 14 16 18 20 21 20 21 15 16 15 17 19 16 15 15 13 24 22 21
[26] 24 15 20 20 25 27 22 20 24 24 23 29 27 23 19 25 15 16 27 27 30 29 26 25 25
[51] 32 28 25

y
 [1] 25 22 17 21 20 13 16 14 19 10 23 20 19 15 16 16 12 15 15 15 15 17 18 16 18
[26] 22 20 21 21 21 25 22 18 21 18 20 25 20 18 19 16 16 16 26 28 28 30 32 28 36
[51] 40 33 33
```

- En este ejemplo, nos concentraremos en el error de predicción, que puede ser estimado por validación cruzada, sin hacer supuestos fuertes acerca del error de la variable.
- Los modelos que se propondrán para la relación son los siguientes:
  1. Lineal: $y = \beta_0 + \beta_1 x + \epsilon$
  2. Cuadrático: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$
  3. Exponencial: $log(y) = log(\beta_0) + \beta_1 x + \epsilon$
  4. Log-Log: $log(y) = \beta_0 + \beta_1 log(x) + \epsilon$

Modelos I

```
par(mfrow=c(2,2))
par(oma=c(1,1,1,1),mar=c(4,4,1,1))
a <- seq(10,40,0.1) #sucesión para graficar los ajustes

L1 <- lm(y ~ x)
plot(x,y,main="Lineal",pch=16)
yhat1 <- L1$coef[1] + L1$coef[2]*a
lines(a,yhat1,lwd=2)

L2 <- lm(y ~ x + I(x^2))
plot(x,y,main="Cuadrático",pch=16)
yhat2 <- L2$coef[1] + L2$coef[2]*a +L2$coef[3]*a^2
lines(a,yhat2,lwd=2)

L3 <- lm(log(y) ~ x)
plot(x,y,main="Exponencial",pch=16)
logyhat3 <- L3$coef[1] + L3$coef[2]*a
yhat3 <- exp(logyhat3)
lines(a,yhat3,lwd=2)

L4 <- lm(log(y) ~ log(x))
plot(log(x),log(y),main="Log-Log",pch=16)
logyhat4 <- L4$coef[1] + L4$coef[2]*log(a)
lines(log(a),logyhat4,lwd=2)
```

Modelos II


Lineal
Cuadrático
Exponencial
Log−Log
y
x
log(y)
log(x)
10
15
20
25
30
40
2.4
2.6
2.8
3.0
3.2
3.4
3.6

Ejemplo Validación Cruzada I

- Una vez que el modelo es ajustado, se evalúa el ajuste.

Procedimiento para estimar el error de predicción usando validación cruzada (uno afuera)

1. Para $k = 1, \ldots, n$ dejar la observación (x_k, y_k) para ser el punto de prueba y usar las observaciones restantes para ajustar el modelo.
   a. Ajusta el modelo usando sólo $n - 1$ observaciones en el conjunto de entrenamiento.
   b. Calcular la respuesta predictiva $\hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_k$ para el punto de prueba
   c. Calcula el error de predicción $e_k = y_k - \hat{y}_k$.
2. Estima la media de los errores de predicción al cuadrado $\hat{\sigma}_\epsilon^2 = \frac{1}{n} \sum_{k=1}^{n} e_k^2$.

Ejemplo Validación Cruzada I

```
n <- length(x)
e1 <- e2 <- e3 <- e4 <- numeric(n)

for(k in 1:n){
        yy <- y[-k]
        xx <- x[-k]
        J1 <- lm(yy ~ xx)
        e1[k] <- y[k] - (J1$coef[1] + J1$coef[2]*x[k])
        J2 <- lm(yy ~ xx + I(xx^2))
        e2[k] <- y[k] - (J2$coef[1] + J2$coef[2]*x[k] + J2$coef[3]*x[k]^2)
        J3 <- lm(log(yy) ~ xx)
        yhat3 <- exp(J3$coef[1] + J3$coef[2]*x[k])
        e3[k] <- y[k] - yhat3
        J4 <- lm(log(yy) ~ log(xx))
        yhat4 <- exp(J4$coef[1] + J4$coef[2]*log(x[k]))
        e4[k] <- y[k] - yhat4
}
```

Los siguientes son los estimados de los errores de predicción

```
c(mean(e1^2), mean(e2^2), mean(e3^2), mean(e4^2))

[1] 19.55644 17.85248 18.44188 20.45424
```

Entonces el mejor modelo es el modelo cuadrático que tiene el menor error cuadrático medio de predicción.