2.2. Bootstrap

Este método es un poco más sencillo de implementar que Jacknife y es igualmente de eficaz propuesto por [4].

Primero recordemos que estamos estimando una estadístico a partir de una muestra de modo que $T_n = g(X_1, \ldots, X_n)$ donde g es cualquier función (media, varianza, quantiles, etc).

Supongamos que conocemos la distribución real de los X's, llamada $F(x)$. Si uno quisiera estimar la varianza de X basta con hacer

Real

$$\boxed{\mathrm{Var}_F(T_n) = \frac{\sigma^2}{n} = \frac{\int x^2 dF(x) - \left(\int x dF(x)\right)^2}{n}}$$

donde $\sigma^2 = \mathrm{Var}(X)$ y el subindice F es solo para indicar la dependencia con la distribución real.

?

Ahora dado que no tenemos la distribución real $F(x)$, una opción es encontrar un estimador de esta llamado $\hat{F}_n$.

La técnica de boostrap se basa en extraer muchas muestras iid de la distribución $\hat{F}_n$ de modo que se pueda conocer su varianza.

$X_1, \ldots, X_n \sim \cancel{F(x)}$

En simple pasos la técnica es

$\hat{F}_n$ → Muestras

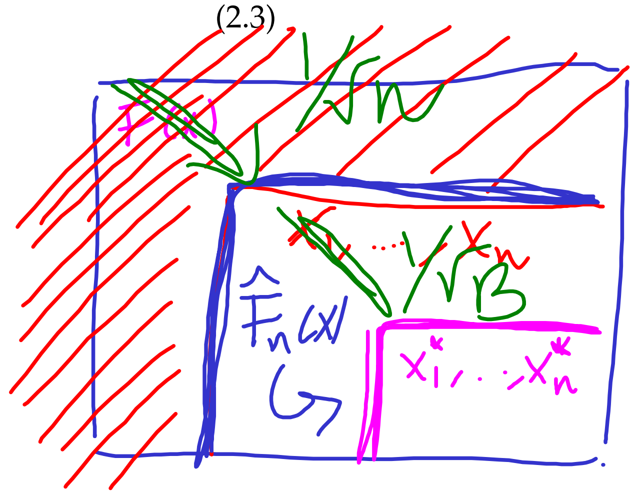
1. Seleccione $X_1^*, \ldots, X_n^* \sim \widehat{F}_n$
2. Estime $T_n^* = g\left(X_1^*, \ldots, X_n^*\right)$
3. Repita los Pasos 1 y 2, B veces para obtener $T_{n,1}^*, \ldots, T_{n,B}^*$

4. Estime

$$v_{\text{boot}} = \frac{1}{B}\sum_{b=1}^{B}\left(T^*_{n,b} - \frac{1}{B}\sum_{r=1}^{B} T^*_{n,r}\right)^2$$

Por la ley de los grandes números tenemos que

$$v_{\text{boot}} \xrightarrow{\text{a.s.}} \mathbb{V}_{\widehat{F}_n}(T_n), \quad \text{si } B \to \infty. \tag{2.3}$$



además llamaremos,

$$\boxed{\widehat{\text{se}}_{\text{boot}} = \sqrt{v_{\text{boot}}}}$$

En pocas palabras lo que tenemos es que

$$\begin{aligned} &\text{Mundo Real: } F && \Longrightarrow X_1,\ldots,X_n \Longrightarrow && T_n = g\left(X_1,\ldots,X_n\right)\\ &\text{Mundo Bootstrap: } \widehat{F}_n && \Longrightarrow X_1^*,\ldots,X_n^* \Longrightarrow && T_n^* = g\left(X_1^*,\ldots,X_n^*\right)\end{aligned}$$

En términos de convergencia lo que se tiene es que

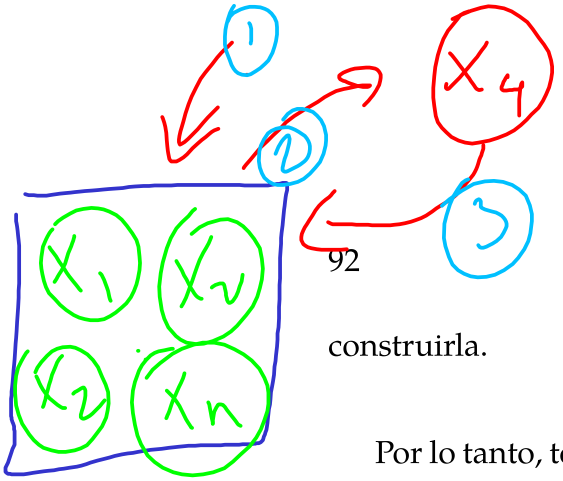
$$\text{Var}_F(T_n) \overbrace{\approx}^{O(1/\sqrt{n})} \text{Var}_{\widehat{F}_n}(T_n) \overbrace{\approx}^{O(1/\sqrt{B})} v_{boot}$$

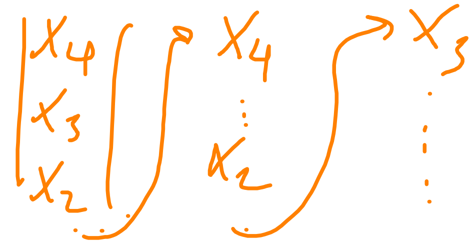
Pregunta 2.2.1

¿Cómo extraemos una muestra de $\hat{F}_n$?

Recuerden que $\hat{F}_n$ asigna la probabilidad de $\frac{1}{n}$ a cada valor usado para

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i \le x)$$





construirla.

Por lo tanto, todos los puntos originales $X_1, \ldots, X_n$ tienen probabilidad $\frac{1}{n}$ de ser escogidos, que resulta ser equivalente a un muestreo con remplazo n-veces.

Así que basta cambiar el punto 1. del algoritmo mencionando anteriormente con

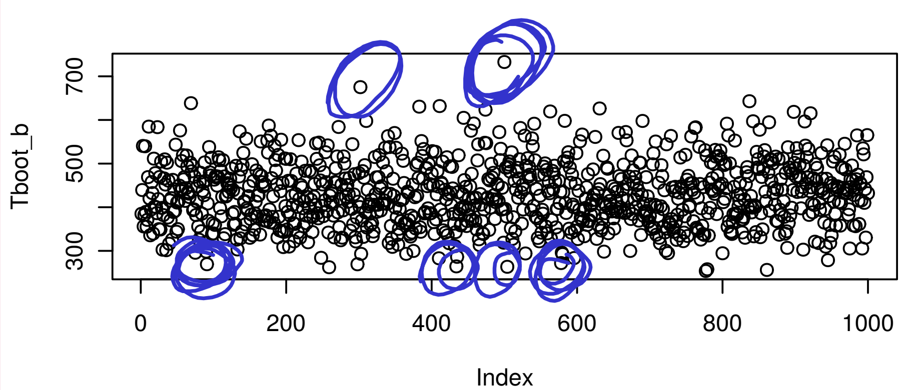
1. Seleccione una muestra con remplazo $X_1^*, \ldots, X_n^*$ de $X_1, \ldots, X_n$.

Laboratorio 2.2.2

En este ejemplo podemos tomar $B = 1000$ y construir esa cantidad de veces nuestro estimador.

```
B <- 1000

Tboot_b <- NULL


for(b in 1:B) {

xb <- sample(x, size = n, replace = TRUE)

Tboot_b[b] <- var(xb)

}




Tboot_b[1:10]

##  [1] 384.8959 439.3992 540.4429 356.0363 382.5001 539.7398 389.2075 355.4033

##  [9] 371.0053 469.0120
```

```
plot(Tboot_b)
```



Por supuesto podemos encontrar los estadísticos usuales para esta nueva muestra

```
(Tboot <- mean(Tboot_b))

## [1] 428.3197
```

```
(Vboot <- var(Tboot_b))

## [1] 5345.401
```

```
(sdboot <- sqrt(Vboot))

## [1] 73.11225
```

2.2.1. Intervalos de confianza

Intervalo Normal

Este es el más sencillo y se escribe como

$$T_n \pm z_{\alpha/2}\widehat{\mathrm{Se}}_{\mathrm{boot}} \tag{2.4}$$

Cuidado 2.2.3

Este intervalo solo funciona si la distribución de T_n es normal.

Laboratorio 2.2.4

El cálculo de este intervalo es

```
c(Tn - z * sdboot,
Tn + z * sdboot)

## [1] 285.9510 572.5458
```

Intervalo pivotal

Sea $\boxed{\theta = T(F)}$ y $\boxed{\widehat{\theta}_n = T\left(\widehat{F}_n\right)}$ y defina la cantidad pivotal $R_n = \widehat{\theta}_n - \theta$.

Sea $H(r)$ la función de distribución del pivote:

$$\boxed{H(r) = \mathbb{P}_F\left(R_n \leq r\right).}$$

Además considere $C_n^{\star} = (a, b)$ donde

$$\boxed{a = \widehat{\theta}_n - H^{-1}\left(1 - \frac{\alpha}{2}\right)} \quad \text{y} \quad \boxed{b = \widehat{\theta}_n - H^{-1}\left(\frac{\alpha}{2}\right).}$$

$\hat{\theta}_n - a = H^{-1}(1 - \alpha/2)$

Se sigue que

$\hat{\theta}_n - b = H^{-1}(\alpha/2)$

$$
\begin{aligned}
\mathbb{P}(a \leq \theta \leq b) &= \mathbb{P}\left(\widehat{\theta}_n - b \leq R_n \leq \widehat{\theta}_n - a\right) \\
&= H\left(\widehat{\theta}_n - a\right) - H\left(\widehat{\theta}_n - b\right) \\
&= H\left(H^{-1}\left(1 - \frac{\alpha}{2}\right)\right) - H\left(H^{-1}\left(\frac{\alpha}{2}\right)\right) \\
&= 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha
\end{aligned}
$$

Nota 2.2.5

$C_n^\star = (a, b)$ es un intervalo de confianza al $1 - \alpha$ de confianza.

El problema es que este intervalo depende de H desconocido.

Para resolver este problema, se puede construir una versión *bootstrap* de H usando lo que sabemos hasta ahora.

Bootstrap

$$\boxed{\widehat{H}(r) = \frac{1}{B}\sum_{b=1}^{B} I\left(R_{n,b}^* \leq r\right)}$$

donde $R_{n,b}^* = \widehat{\theta}_{n,b}^* - \widehat{\theta}_n$. ⇒ DATOS ORIGINALES.

Sea r_β^* el cuantil muestral de tamaño β de $\left(R_{n,1}^*, \ldots, R_{n,B}^*\right)$ y sea θ_β^* el cuantil muestral de tamaño β de $\left(\theta_{n,1}^*, \ldots, \theta_{n,B}^*\right)$.

Nota 2.2.6

Según la notación anterior note que

$$\boxed{r_\beta^* = \theta_\beta^* - \widehat{\theta}_n}$$

Con estas observaciones It follows that an approximate $1-\alpha$ confidence interval is $C_n = (\widehat{a}, \widehat{b})$ where

$$\begin{aligned}
\widehat{a} &= \widehat{\theta}_n - \widehat{H}^{-1}\left(1-\frac{\alpha}{2}\right) = \widehat{\theta}_n - r^*_{1-\alpha/2} = \widehat{\theta}_n - \theta^*_{1-\alpha/2} + \widehat{\theta}_n = 2\widehat{\theta}_n - \theta^*_{1-\alpha/2}\\
\widehat{b} &= \widehat{\theta}_n - \widehat{H}^{-1}\left(\frac{\alpha}{2}\right) = \widehat{\theta}_n - r^*_{\alpha/2} = \widehat{\theta}_n - \theta^*_{\alpha/2} + \widehat{\theta}_n = 2\widehat{\theta}_n - \theta^*_{\alpha/2}
\end{aligned}$$

Nota 2.2.7

El intervalo de confianza pivotal de tamaño $1-\alpha$ es

$$C_n = \left(2\widehat{\theta}_n - \widehat{\theta}^*_{((1-\alpha/2)B)}, 2\widehat{\theta}_n - \widehat{\theta}^*_{((\alpha/2)B)}\right)$$

Laboratorio 2.2.8

El intervalo anterior para un nivel de 95 % se estima de la siguiente forma

var(α)

1 - α/2

```
c(2 * Tn - quantile(Tboot_b, 1 - 0.05 / 2) ,
2 * Tn - quantile(Tboot_b, 0.05 / 2))
```

α/2

```
##    97.5%     2.5%
## 275.3768 558.6741
```

200
250
300
α/2
350
360
1-α/2

Intervalo pivotal studentizado

Una mejora del intervalo anterior sería normalizar los estimadores previamente

$$Z_n = \frac{T_n - \theta}{\widehat{\mathrm{se}}_{\mathrm{boot}}}$$

Como θ es desconocido, entonces la versión a estimar es

$$Z^*_{n,b} = \frac{T^*_{n,b} - T_n}{\widehat{\mathrm{se}}^*_b}$$

donde $\widehat{\mathrm{se}}^*_b$ es un estimador del error estándar de $T^*_{n,b}$ no de T_n.

> **Cuidado 2.2.9**
>
> Esto requerira estimar la varianza de $T^*_{n,b}$ para cada b.

⚠ 2 veces Bootstrap.

Con esto se puede obtener cantidades $Z^*_{n,1}, \ldots, Z^*_{n,B}$ que debería ser próximos a Z_n.

Sea z^*_α del α cuantiĺ de $Z^*_{n,1}, \ldots, Z^*_{n,B}$, entonces $\mathbb{P}(Z_n \leq z^*_\alpha) \approx \alpha$.

Define el intervalo

$$\boxed{C_n = \left(T_n - z^*_{1-\alpha/2}\widehat{\mathrm{se}}_{\mathrm{boot}}, T_n - z^*_{\alpha/2}\widehat{\mathrm{se}}_{\mathrm{boot}}\right)}$$

Justificado por el siguiente cálculo:

$$\begin{aligned}
\mathbb{P}\left(\theta \in C_n\right) &= \mathbb{P}\left(T_n - z^*_{1-\alpha/2}\widehat{\text{se}}_{\text{boot}} \leq \theta \leq T_n - z^*_{\alpha/2}\widehat{\text{se}}_{\text{boot}}\right) \\
&= \mathbb{P}\left(z^*_{\alpha/2} \leq \frac{T_n - \theta}{\text{se}_{\text{boot}}} \leq z^*_{1-\alpha/2}\right) \\
&= \mathbb{P}\left(z^*_{\alpha/2} \leq Z_n \leq z^*_{1-\alpha/2}\right) \\
&\approx 1 - \alpha
\end{aligned}$$

Laboratorio 2.2.10

Note que para este caso tenemos que hacer bootstrap para cada estimador bootstrap calculado.