TEMA 1: Bootstrap

José R. Berrendero

Departamento de Matemáticas, Universidad Autónoma de Madrid

Temas a tratar

- La función de distribución empírica
- Contraste de Kolmogorov-Smirnov
- Idea básica del bootstrap: ejemplos
- Consistencia
- Situaciones en las que el bootstrap no funciona
- Intervalos de confianza bootstrap

Métodos de remuestreo

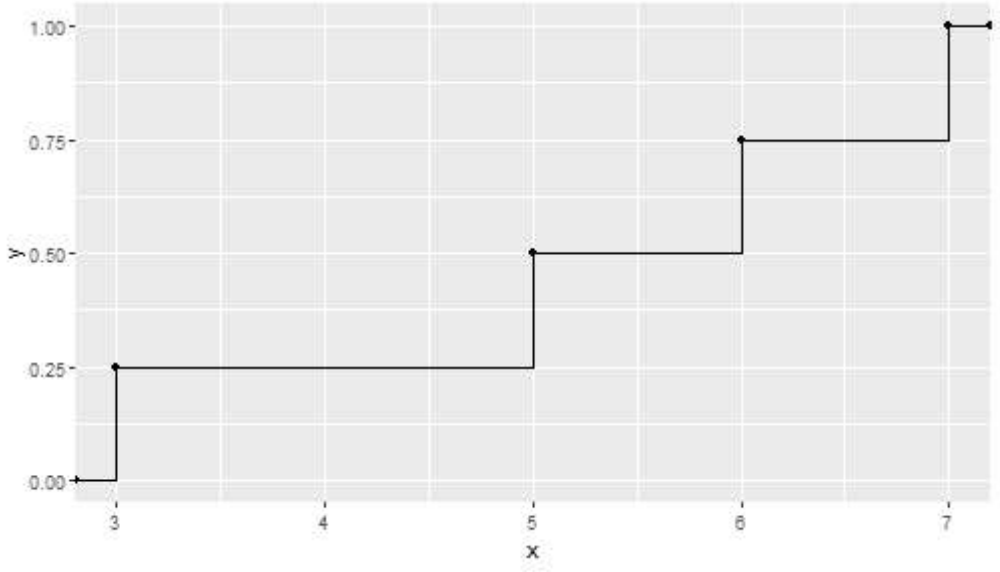
- Extraer repetidamente muestras de los datos de entrenamiento (remuestras) con el fin de:
  - obtener información de las propiedades de un estimador
  - valorar el ajuste de un modelo o su capacidad predictiva
- Métodos de remuestreo que aparecen en esta asignatura:
  - Bootstrap
  - Validación cruzada
- El bootstrap (Efron, 1979) tiene por objetivo aproximar la distribución en el muestreo de un estadístico. Combina dos ideas:
  - **Principio *plug-in*.** Cualquier cantidad desconocida que dependa de F se puede aproximar reemplazando F por un estimador *adecuado*.
  - **Simulación.** Se puede obtener un número grande de réplicas de la aproximación anterior usando métodos de simulación.

Función de distribución empírica

Dada una muestra de vaiid $X_1, \ldots, X_n$, se define su **función de distribución empírica** como

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{X_i \leq x\}}$$

Ejemplo. Si $x_1 = 3, x_2 = 5, x_3 = 6, x_4 = 7$, la función de distribución empírica es

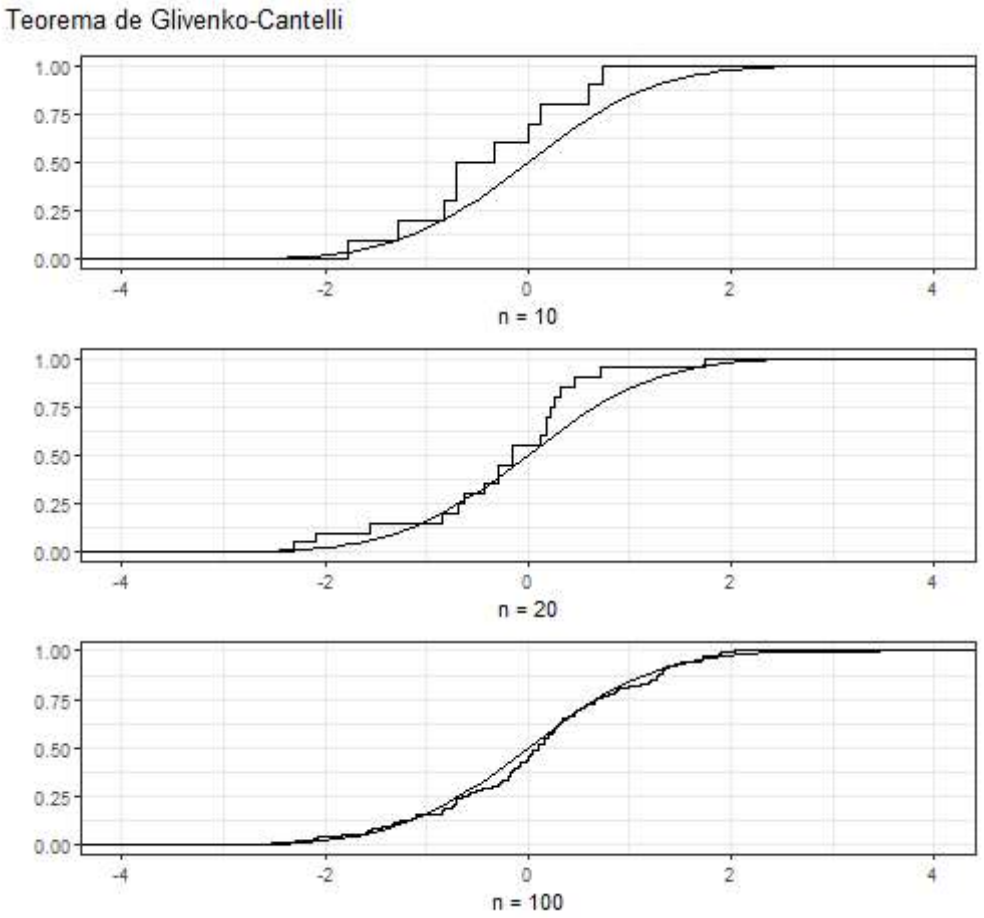


Función de distribución empírica

- ¿Cómo se reparte la probabilidad según esta distribución?
- Si X^* tiene distribución dada por F_n, ¿cuánto vale $\mathrm{E}(X^*)$?
- Fijamos x, ¿cuál es la distribución de la variable aleatoria $nF_n(x)$?
- Determina el valor de $\mathrm{E}[F_n(x)]$ y $\mathrm{Var}[F_n(x)]$.
- Para cada $x \in \mathbb{R}$, $F_n(x) \to_p F(x)$.
- De hecho, se cumple un resultado mucho más fuerte, el teorema de Glivenko-Cantelli:

$$\lim_{n\to\infty} \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| = 0, \quad \text{c.s.}$$

Teorema de Glivenko-Cantelli



Contraste de Kolmogorov-Smirnov

- **Contraste de bondad de ajuste.** Tenemos una muestra $X_1, \ldots, X_n$ procedente de F. Objetivo: contrastar $H_0: F = F_0$, donde F_0 es continua y conocida.

- **Estadístico de Kolmogorov-Smirnov.**

$$D_n := \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)|.$$

- Si H_0 es cierta se cumple que $D_n \to 0$ c.s.
- Idea: rechazar H_0 si D_n es *suficientemente grande*.

- La distribución de D_n bajo H_0 es la misma para cualquier distribución continua F_0 (se dice que D_n es de **distribución libre**).

Contraste de Kolmogorov-Smirnov

El comando en el que está implementado este contraste es `ks.test`

```
set.seed(100)

n <- 100
x <- rnorm(n)
ks.test(x, "pnorm") # H0 verdadera
```

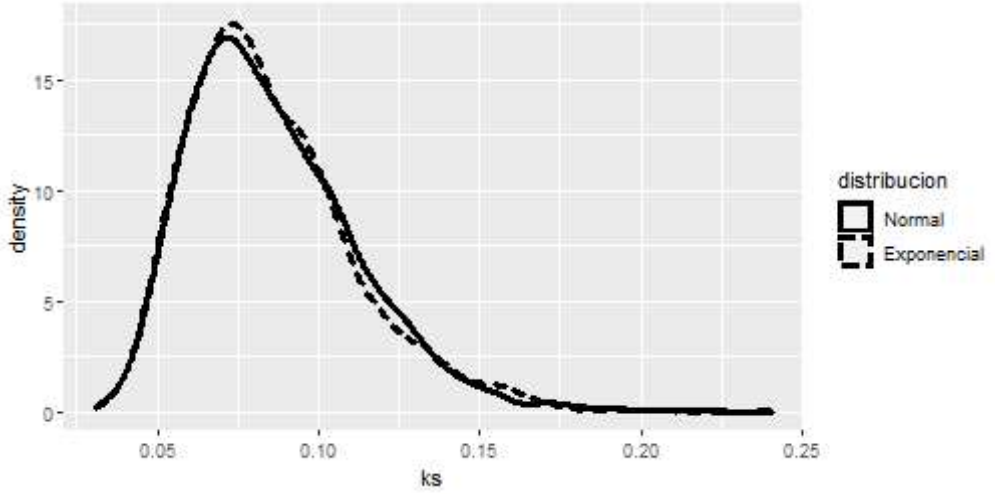
```
##
##     One-sample Kolmogorov-Smirnov test
##
## data:  x
## D = 0.073373, p-value = 0.6546
## alternative hypothesis: two-sided
```

```
x <- rchisq(n, 2)
ks.test(x,"pexp") # H0 falsa
```

```
##
##     One-sample Kolmogorov-Smirnov test
##
## data:  x
## D = 0.25824, p-value = 3.226e-06
## alternative hypothesis: two-sided
```

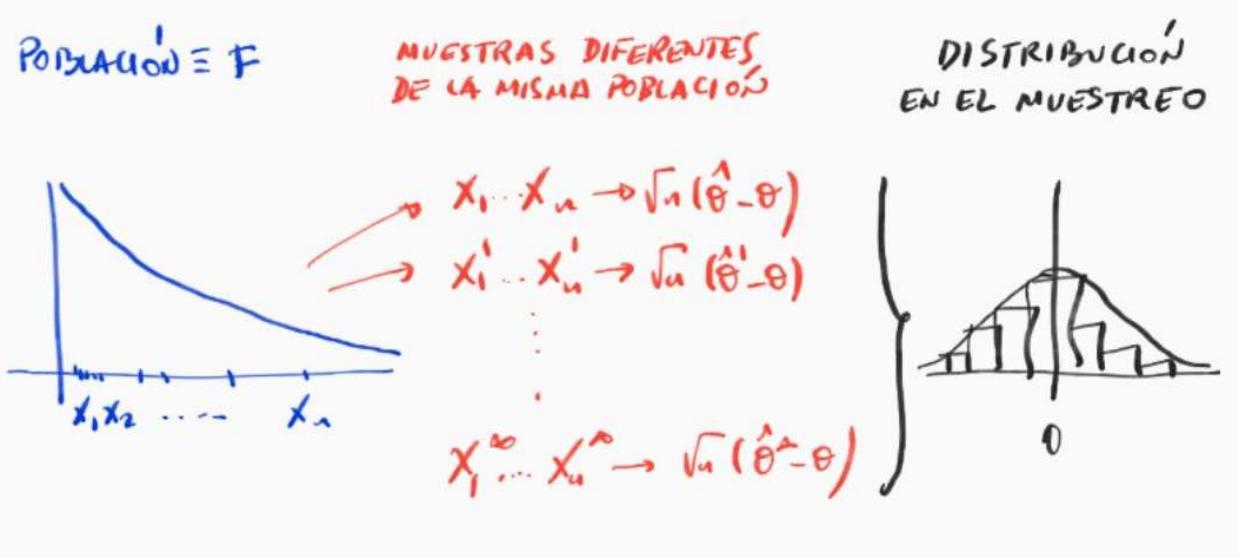
Distribución bajo la nula

```
R <- 2000
n <- 100
ks_norm <- replicate(R, ks.test(rnorm(n), "pnorm")$statistic)
ks_exp <- replicate(R, ks.test(rexp(n), "pexp")$statistic)
distribucion <- gl(2, R, labels = c("Normal", "Exponencial"))

df <- data.frame(ks = c(ks_norm, ks_exp), distribucion = distribucion)
ggplot(df) +
  geom_density(aes(x = ks, y = ..density.., linetype = distribucion), size = 1.1
```



Bootstrap: idea básica

Objetivo. Aproximar la distribución en el muestreo de $T = T(x_1, \ldots, x_n; F)$ (por ejemplo, $T = \sqrt{n}(\hat{\theta} - \theta)$)



$$H_n(x) = \mathrm{P}_F(T(X_1, \ldots, X_n; F) \leq x)$$

Dificultad. Solo tenemos una muestra y F es desconocida.

Bootstrap: idea básica

La idea básica es sustituir F por F_n, con lo que resulta el estimador bootstrap *ideal*:

$$\hat{H}_n(x) = \mathrm{P}_{F_n}(T(X_1^*, \ldots, X_n^*; F_n) \leq x).$$

Casi siempre es imposible obtener una expresión cerrada de $\hat{H}_n$.

Aproximación mediante simulación de $\hat{H}_n$

Se simulan muestras de observaciones $X_1^{*b}, \ldots, X_n^{*b}$ iid de F_n (con $b = 1, \ldots, B$ y B grande).

Para cada muestra artificial se calcula $T^{*(b)} = T(X_1^{*b}, \ldots, X_n^{*b}; F_n)$.

El valor de $\hat{H}_n(x)$ se puede aproximar por:

$$\hat{H}_n(x) \approx \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}_{\{T^{*(b)} \leq x\}}.$$

Bootstrap: idea básica

Procedimiento

- Se estima F mediante F_n. (Principio de sustitución o *plug-in*.)
- Para $b = 1, \ldots, B$
  - Se obtienen muestras bootstrap $X_1^{*b}, \ldots, X_n^{*b}$ procedentes de F_n, sorteando con reemplazamiento entre los datos originales.
  - Se calcula $T^{*(b)} = T(X_1^{*b}, \ldots, X_n^{*b}; F_n)$ para cada una de las muestras bootstrap.
- Se calcula la proporción $\tilde{H}_B(x) = B^{-1} \sum_{b=1}^{B} \mathbb{I}_{\{T^{*(b)} \leq x\}}$.

Una doble aproximación

En el procedimiento anterior hay una doble aproximación:

$$H_n(x) \approx \hat{H}_n(x) \approx \tilde{H}_B(x).$$

- La primera aproximación requiere n grande y regularidad
- La segunda requiere B grande (no es problema)

Estimador bootstrap de la varianza

Esencialmente el mismo algoritmo sirve para estimar cualquier aspecto de la distribución, en lugar de la función de distribución completa.

Para estimar la varianza (o la desviación típica) de un estimador $\mathrm{Var}_F(\hat{\theta})$:

- El estimador bootstrap *ideal* es $\mathrm{Var}_{F_n}(\hat{\theta}^*)$.
- La correspondiente aproximación basada en B remuestras es

$$\mathrm{Var}_{F_n}(\hat{\theta}^*) \approx \frac{1}{B-1}\sum_{j=1}^{B}(\hat{\theta}_j^* - \bar{\theta}^*)^2,$$

donde $\hat{\theta}_j^*$ es el valor del estimador para la remuestra j, y
$\bar{\theta}^* = B^{-1}\sum_{j=1}^{B}\hat{\theta}_j^*$ es el promedio de todas las versiones bootstrap de $\hat{\theta}$.

Ejemplo: la varianza de la mediana

- Supongamos que $X_1, \ldots, X_n$ son v.a.i.i.d. de una distribución de Cauchy centrada en θ y con parámetro de escala igual a uno.
- La función de densidad es:

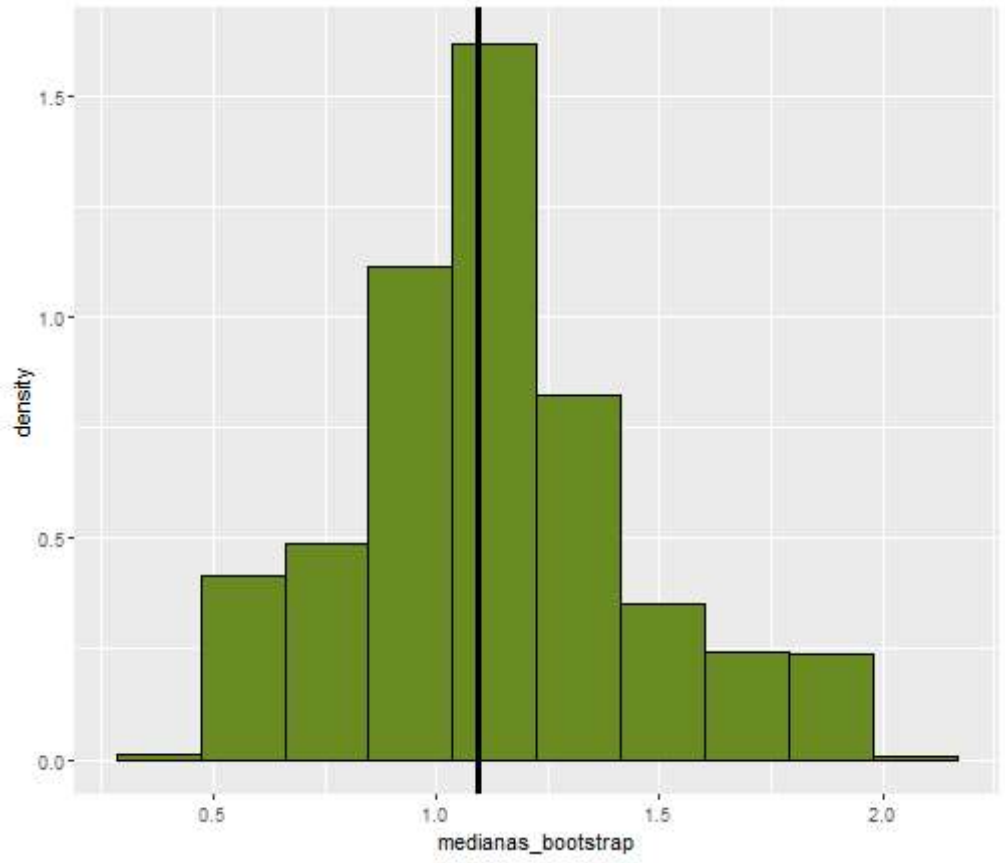
$$f(x) = \frac{1}{\pi} \frac{1}{1 + (x - \theta)^2}, \quad x \in \mathbb{R}.$$

- La esperanza de esta distribución no existe, por lo que para estimar θ se usa la mediana.
- ¿Cuál es la desviación típica de esta mediana? Necesitamos estimar este valor para calcular, por ejemplo, un intervalo de confianza para θ.

Ejemplo: la varianza de la mediana

```
set.seed(100)

# Parámetros
R <- 1000
n <- 30
theta <- 1

# Generamos los datos
muestra_original <- rt(n, 1) + theta   # Cauchy con theta = 0 coincide con t Student con
mediana_original <- median(muestra_original)

# Generamos las remuestras (matriz n x R, cada columna una remuestra)
muestras_bootstrap <- sample(muestra_original, n*R, rep = TRUE)
muestras_bootstrap <- matrix(muestras_bootstrap, nrow = n)

# Medianas de las remuestras
medianas_bootstrap <- apply(muestras_bootstrap, 2, median)

# Histograma de las medianas bootstrap
df <- data.frame(medianas_bootstrap = medianas_bootstrap)
ggplot(df) +
  geom_histogram(aes(x = medianas_bootstrap, y = ..density..),
                 bins = 10, fill = 'olivedrab4', col = 'black') +
  geom_vline(xintercept = mediana_original, size = 1.1)

# Estimador bootstrap de la desviación típica de la mediana
sd_mediana <- sd(medianas_bootstrap)
sd_mediana
```

Ejemplo: la varianza de la mediana



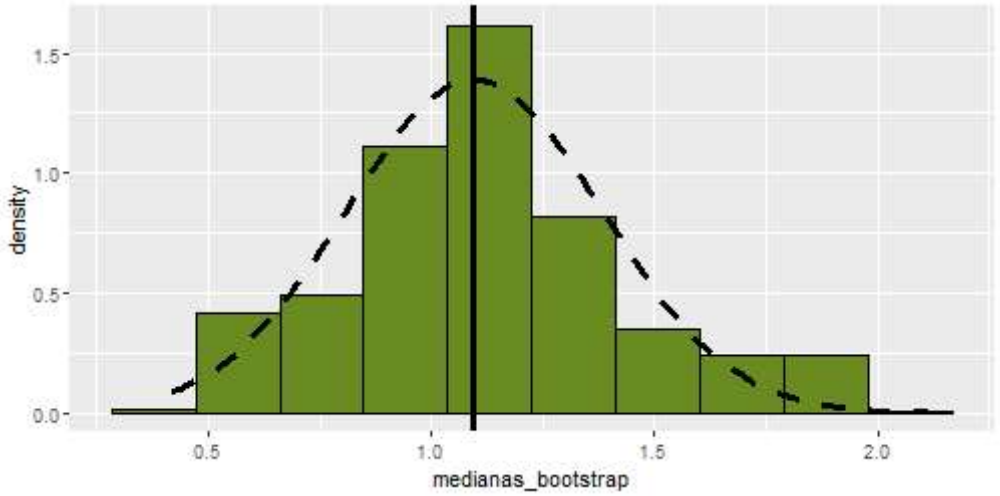

```
## [1] 0.3135752
```

Comparación con la distribución asintótica

Si M_n es la mediana muestral, $m = F^{-1}(1/2)$ es la única mediana poblacional, y la distribución tiene densidad f, continua en un entorno de m tal que $f(m) > 0$,

$$\sqrt{n}(M_n - m) \to_d \mathrm{N}\left(0, \frac{1}{4f(m)^2}\right).$$

¿Qué aproximación se obtiene para la distribución de Cauchy?



Consistencia del bootstrap

Sea ρ una distancia entre distribuciones de probabilidad.

Ejemplos:

- Kolmogorov: $\rho(F,G) = \|F-G\|_\infty = \sup_x |F(x)-G(x)|$
- Mallows

El bootstrap es fuertemente consistente cuando $\rho(H_n, \hat{H}_n) \to 0$ c.s. si $n \to \infty$.

Es débilmente consistente si $\rho(H_n, \hat{H}_n) \to_p 0$.

Un resultado clásico:

Teorema (Singh, 1981). Supongamos $\mathrm{E}_F(X^2) < \infty$ y denotemos $\mu = \mathrm{E}_F(X)$, $H_n(x) = \mathrm{P}_F(\sqrt{n}(\bar{X}-\mu) \leq x)$ y $\hat{H}_n(x) = \mathrm{P}_{F_n}(\sqrt{n}(\bar{X}^*-\bar{X}) \leq x)$. Entonces $\|H_n - \hat{H}_n\|_\infty \to 0$, con probabilidad 1.

Resultados de validez para la mediana en Ghosh et al (1984).

¿Cuándo falla el bootstrap?

El método bootstrap no siempre es consistente. Suele fallar cuando $T(X_1, \ldots, X_n; F)$ no admite un teorema central del límite.

- $T(X_1, \ldots, X_n; F) = \sqrt{n}(\bar{X} - \mu)$, pero $\mathrm{Var}(X) = \infty$.
- $T(X_1, \ldots, X_n; F) = \sqrt{n}(g(\bar{X}) - g(\mu))$, pero g no es derivable en μ.
- $T(X_1, \ldots, X_n; F) = \sqrt{n}(F_n^{-1}(p) - F^{-1}(p))$, pero $f(F^{-1}(p)) = 0$.
- La distribución de los datos es F_θ y el soporte de F_θ depende del parámetro.

Ejemplo. Sea $X_1, \ldots, X_n$ vaiid de una distribución uniforme en el intervalo $(0, \theta)$. El estimador de máxima verosimilitud de θ es $\hat{\theta} = X_{(n)}$.

El bootstrap no sirve para aproximar la distribución de $\hat{\theta}$.

Intervalos de confianza bootstrap

¿Qué haríamos si $H_n(x)$, la distribución de $\sqrt{n}(\hat{\theta} - \theta)$, fuese conocida?

Despejar θ en la ecuación

$$1 - \alpha = \mathrm{P}_F\{H_n^{-1}(\alpha/2) \leq \sqrt{n}(\hat{\theta} - \theta) \leq H_n^{-1}(1 - \alpha/2)\},$$

lo que da lugar al intervalo

$$\left[\hat{\theta} - n^{-1/2} H_n^{-1}(1 - \alpha/2),\ \hat{\theta} - n^{-1/2} H_n^{-1}(\alpha/2)\right].$$

Como H_n no es conocida, la sustituimos por el estimador bootstrap $\hat{H}_n$.

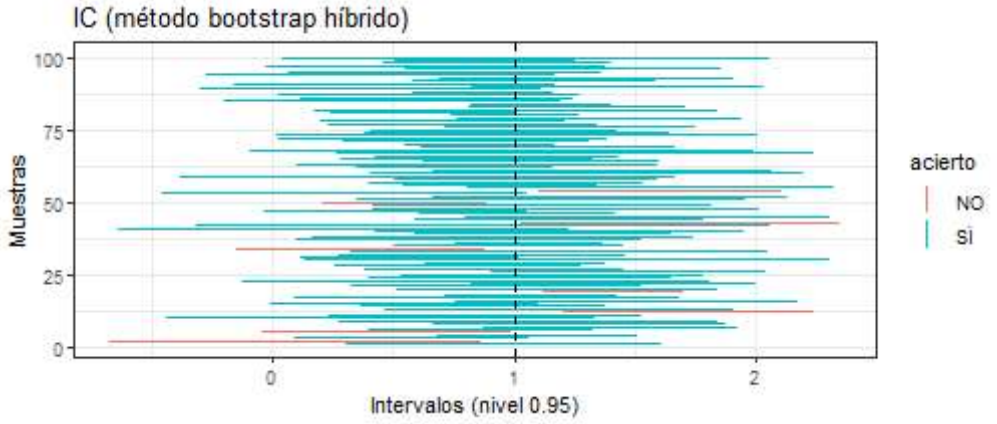
Hay muchos otros métodos, este es el llamado *método híbrido*.

Intervalos de confianza bootstrap

```
set.seed(100)

# Parámetros
R <- 1000
n <- 30
theta <- 1
m <- 100
alfa <- 0.05

# Cálculo de los intervalos y % de aciertos
acierto <- NULL
intervalo <- NULL
for (i in 1:m){
  muestra_original <- rt(n, 1) + theta
  mediana_original <- median(muestra_original)

  muestras_bootstrap <- sample(muestra_original, n*R, rep = TRUE)
  muestras_bootstrap <- matrix(muestras_bootstrap, nrow = n)
  medianas_bootstrap <- apply(muestras_bootstrap, 2, median)
  T_bootstrap <- sqrt(n) * (medianas_bootstrap - mediana_original)
  ic_min <- mediana_original -  quantile(T_bootstrap, 1-alfa/2)/sqrt(n)
  ic_max  <- mediana_original -  quantile(T_bootstrap, alfa/2)/sqrt(n)
  intervalo <- rbind(intervalo, c(ic_min, ic_max))
  acierto <- c(acierto, ic_min < theta & ic_max > theta)
}
```

Intervalos de confianza bootstrap

```
# Gráfico
df <- data.frame(ic_min <- intervalo[,1],
                 ic_max <- intervalo[, 2],
                 ind = 1:m,
                 acierto = acierto)
ggplot(df) +
  geom_linerange(aes(xmin = ic_min, xmax = ic_max, y = ind, col = acierto)) +
  scale_color_hue(labels = c("NO", "SÍ")) +
  geom_vline(aes(xintercept = theta), linetype = 2) +
  theme_bw() +
  labs(y = 'Muestras', x = 'Intervalos (nivel 0.95)',
       title = 'IC (método bootstrap híbrido)')
```



Otros métodos

Aproximar la distribución del estimador estandarizado

Si $\text{et}(\hat{\theta})$ denota el error típico de $\hat{\theta}$ Se pueden también aproximar los percentiles de la distribución de

$$(\hat{\theta} - \theta)/\text{et}(\hat{\theta}).$$

Suponer que la distribución de $\hat{\theta}$ es aproximadamente normal

Si la distribución de $\hat{\theta}$ es aproximadamente normal, un posible IC es

$$\text{IC}_{1-\alpha}(\theta) = [\hat{\theta} \mp z_{\alpha/2}\ \text{et}_{boot}(\hat{\theta})],$$

donde $\text{et}_{boot}(\hat{\theta})$ es un estimador de la desviación típica (error típico) de $\hat{\theta}$.

Método del percentil bootstrap

Sea $\hat{H}_n(x) := \mathrm{P}_{F_n}(\hat{\theta}^* \leq x)$

El intervalo basado en el percentil bootstrap es

$$\mathrm{IC}_{1-\alpha}(\theta) = [\hat{H}_n^{-1}(\alpha/2),\ \hat{H}_n^{-1}(1-\alpha/2)]$$

En la práctica se usan los percentiles de los valores bootstrap generados por simulación:

- Generamos $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_B$
- Sea $\hat{\theta}^*(\alpha)$ tal que $\#\{b:\ \hat{\theta}^*_b \leq \hat{\theta}^*(\alpha)\}/B = \alpha$
- El intervalo es $[\hat{\theta}^*(\alpha/2),\ \hat{\theta}^*(1-\alpha/2)]$

Método del percentil bootstrap

- Supongamos que existe una transformación monótona g que normaliza y estabiliza la varianza, es decir, tal que

$$\hat{\phi} := g(\hat{\theta}) \cong \mathrm{N}(\phi, c^2), \qquad \phi = g(\theta), \;\; c \in \mathbb{R}.$$

- Un IC en este caso es:

$$IC_{1-\alpha}(\theta) = [g^{-1}(\hat{\phi} - cz_{\alpha/2}),\; g^{-1}(\hat{\phi} + cz_{\alpha/2})]$$

- El método del bootstrap percentil es una aproximación de este intervalo **pero no requiere conocer ni g ni c**:
  - Si aplicamos el método percentil a ϕ resulta

$$[g(\hat{H}_n^{-1}(\alpha/2)),\;\; g(\hat{H}_n^{-1}(1-\alpha/2))] \approx [\hat{\phi} - cz_{\alpha/2},\; \hat{\phi} + cz_{\alpha/2}]$$

  - Por lo tanto

$$[\hat{H}_n^{-1}(\alpha/2),\; \hat{H}_n^{-1}(1-\alpha/2)] \approx [g^{-1}(\hat{\phi} - cz_{\alpha/2}),\; g^{-1}(\hat{\phi} + cz_{\alpha/2})]$$

Ejemplo: correlaciones entre notas

Notas en 2009 y 2010 de una prueba al final de primaria en 100 colegios de la Comunidad de Madrid:

```
set.seed(100)
n <- 100

colegios <- read_table("http://verso.mat.uam.es/~joser.berrendero/datos/notas.tx
                       locale = locale(decimal_mark = ",")) %>%
  mutate(tipo = factor(tipo)) %>%
  slice_sample(n = n)   # selecciona n colegios aleatoriamente

head(colegios)
```

```
## # A tibble: 6 x 3
##   tipo       nota09 nota10
##   <fct>       <dbl>  <dbl>
## 1 concertado   6.9    5.27
## 2 publico      5.78   4.11
## 3 concertado   5.71   4.1
## 4 publico      5.97   5.78
## 5 concertado   6.98   5.48
## 6 concertado   6.21   6.26
```

Ejemplo

Transformación z de Fisher del coeficiente de correlación

$$\hat{\phi} = g(\hat{\rho}) = \frac{1}{2}\log\frac{1+\hat{\rho}}{1-\hat{\rho}}, \quad \phi = g(\rho), \quad \hat{\phi} \cong \mathrm{N}\left(\phi, \sigma^2 = \frac{1}{n-3}\right)$$

```
alpha <- 0.05  # 1 - nivel de confianza

datos_xy <- cbind(colegios$nota09, colegios$nota10)
correlacion <- cor(datos_xy)[1,2]
correlacion_fisherz <- 0.5 * log ((1+correlacion)/(1-correlacion))

round(c(correlacion, correlacion_fisherz), 2)
```

```
## [1] 0.56 0.63
```

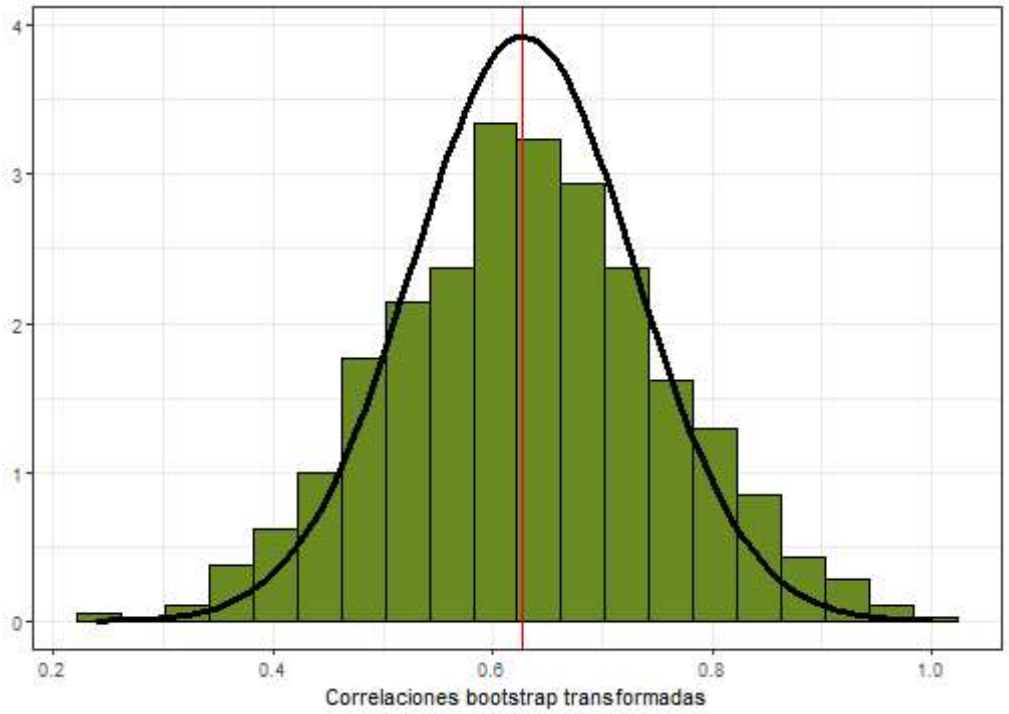
Ejemplo

Representamos las distribuciones bootstrap de las correlaciones transformadas junto con la aproximación normal:

```
R <- 1000     # número de remuestras

corr_bootstrap <- replicate(R, cor(datos_xy[sample(1:n, n, rep=TRUE),])[1,2])
corr_bootstrap_fisherz <- 0.5*log((1+corr_bootstrap)/(1-corr_bootstrap))

df <- data.frame(corr_bootstrap, corr_bootstrap_fisherz)

ggplot(df) +
  geom_histogram(aes(x=corr_bootstrap_fisherz, y=..density..),
                 fill='olivedrab4',
                 col='black',
                 bins = 20) +
  labs(x = 'Correlaciones bootstrap transformadas', y = NULL) +
  geom_vline(xintercept = correlacion_fisherz, col = 'red') +
  geom_function(fun = dnorm,
                args = list(mean=correlacion_fisherz, sd = 1/sqrt(n-3)),
                size = 1.2) +
  theme_bw()
```

Ejemplo

Representamos las distribuciones bootstrap de las correlaciones transformadas junto con la aproximación normal para comprobar si $\hat{\phi}^* \cong \mathrm{N}(\hat{\phi}, \sigma^2 = 1/(n-3))$



Ejemplo

- Correlación entre la nota de 2009 y 2010 en los colegios de Madrid: 0.56
- IC basado en el percentil bootstrap

```
round(c(quantile(corr_bootstrap, alpha/2), quantile(corr_bootstrap, 1-alpha/2)),
```

```
##  2.5% 97.5%
##  0.37  0.70
```

- IC basado en la transformación z de Fisher

```
IC_phi <- c(correlacion_fisherz - qnorm(1-alpha/2)/sqrt(n-3),
            correlacion_fisherz + qnorm(1-alpha/2)/sqrt(n-3))
IC_rho <- (exp(2*IC_phi) - 1) / (exp(2*IC_phi) + 1)
round(IC_rho, 2)
```

```
## [1] 0.40 0.68
```