Ciencias de Datos con R: Fundamentos Estadísticos

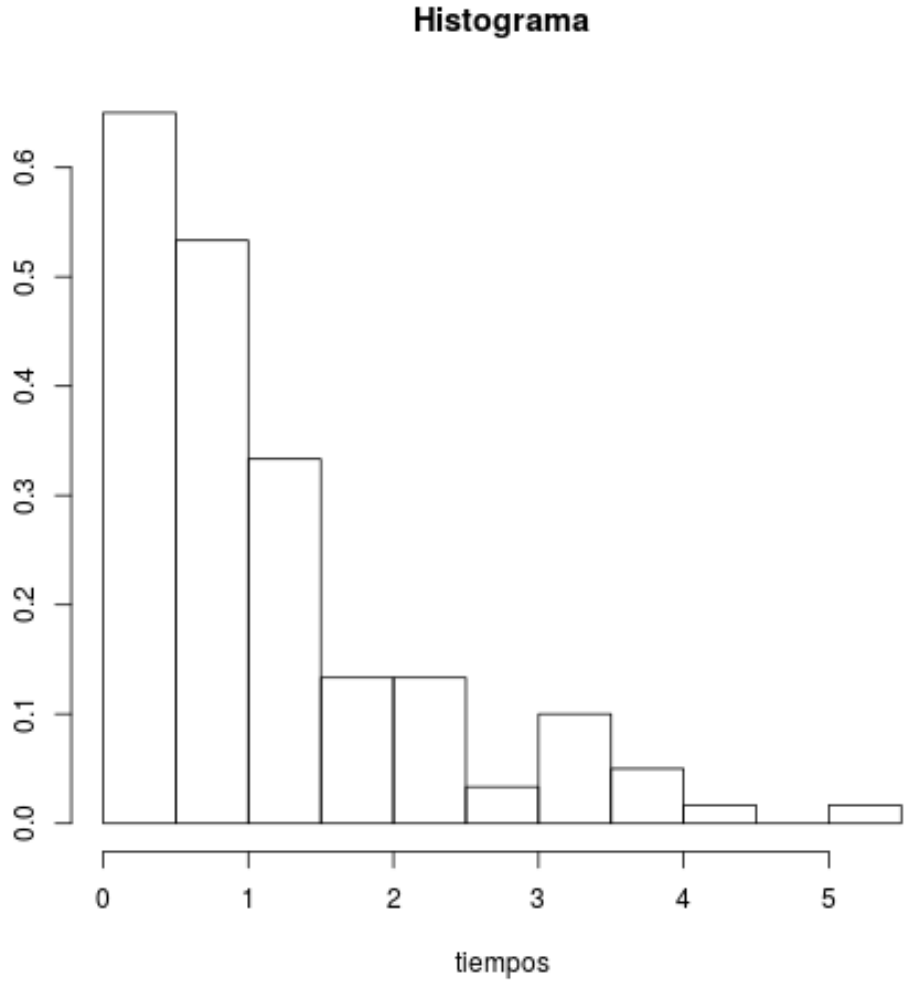
Ana M. Bianco, Jemina García y Mariela Sued.

Error de Estimación – Incertidumbre

Motivación:

La mediana nos dió 0.77 pero el referee nos pide el *error*.

Te traje esto



n=120, mean(tiempos)=1.11, median(tiempos)=0.77,
sd(tiempos)=1.04

¿Qué hacemos?

Una manera de informar: $\pm$

On the other hand, the invariance of retrograde speeds distribution suggests that the endogenous and KHC576-mCherry-Pex26 motors contribute similarly to retrograde transport. In addition, KHC576-mCherry-Pex26 did not affect the median values of plus and minus-end directed run lengths (Table 1) and the number of reversions in the trajectories (1.8 $\pm$ 0.1 reversions/trajectory).

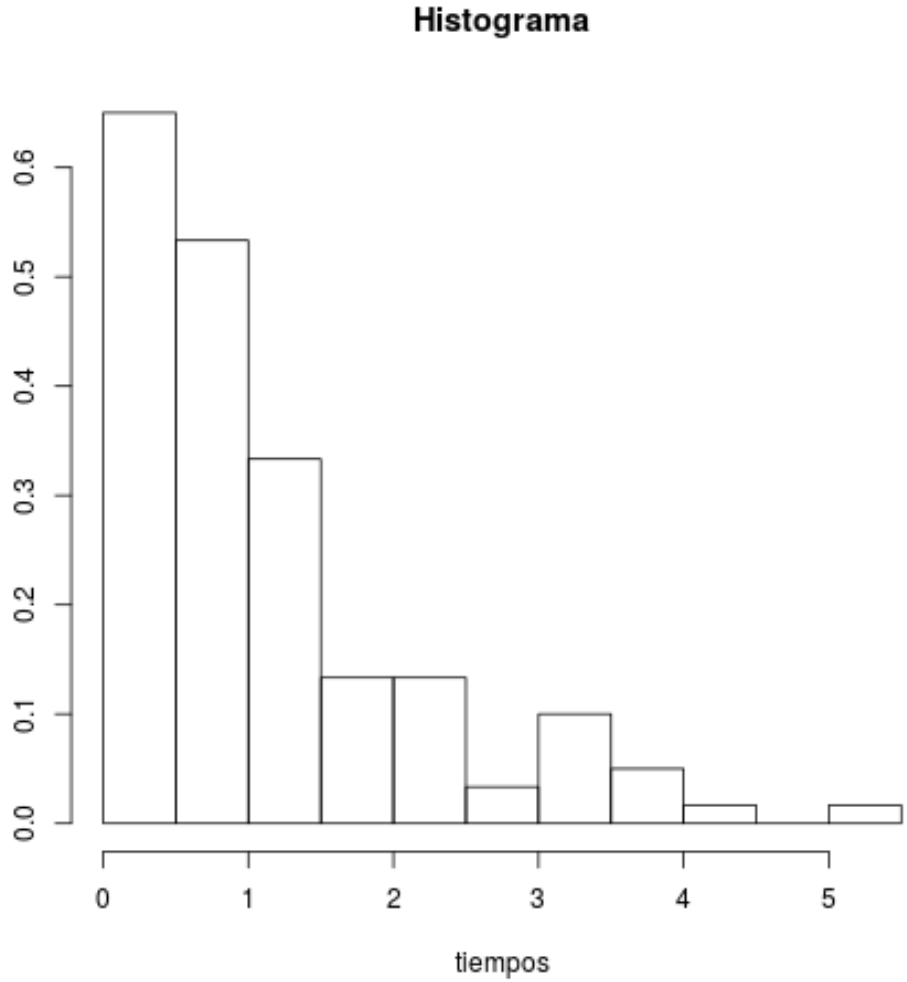
- Se informa poniendo el valor de la estimación y **"su error"** - **uncertainty**.
- ¿A qué llamamos **"error"** (uncertainty)?

Otra manera de informar: p-valor

gated from PD1 to PD7 for each mother. Analyzing this parameter, we found that the parental generation (F0) of LP mothers spent less time providing nurturing behavior than NP mothers ($t_{25} = -2.14$, $P = 0.04$)

El universo Test de Hipótesis.

Te traje esto



n=120, mean(tiempos)=1.11, median(tiempos)=0.77,
sd(tiempos)=1.04
... *median value 0.77 (±?)*

Más generalmente

La **estimación** nos dió xxx pero el referee nos pide el *error*.
¿Qué hacemos?

Más generalmente

La **estimación** nos dió xxx pero el referee nos pide el *error*.
¿Qué hacemos?

"Toda estimación relevante conlleva un error".
Qué es (y qué no es) la estadística: Usos y abusos de una disciplina clave . Walter Sosa Escudero.

Manos a la obra

Utilizando los datos que te hemos asignado

- Realizá un histograma para ver de que manera se distribuyen. ¿Son normales?
- Obtené una estimación de la media poblacional. ¿Sabés informar una medida de error?
- Obtené una estimación de la mediana poblacional. ¿Sabés informar una medida de error?

Manos a la obra

Utilizando los datos que te hemos asignado

- Realizá un histograma para ver de que manera se distribuyen. ¿Son normales?
- Obtené una estimación de la media poblacional. ¿Sabés informar una medida de error?
- Obtené una estimación de la mediana poblacional. ¿Sabés informar una medida de error?

Puesta en común con los resultados obtenidos.

Manos a la obra

Utilizando los datos que te hemos asignado

- Realizá un histograma para ver de que manera se distribuyen. ¿Son normales?
- Obtené una estimación de la media poblacional. ¿Sabés informar una medida de error?
- Obtené una estimación de la mediana poblacional. ¿Sabés informar una medida de error?

Puesta en común con los resultados obtenidos.

- Histogramas con un conjunto de datos. ¿Son normales?
- Histograma con las estimaciones de la media.
- Histograma con las estimaciones de la mediana.

Pero para la media si lo sabemos hacer!

n=120, mean(tiempos)=1.11, median(tiempos)=0.77,
sd(tiempos)=1.04
... *mean value 1.11* $(\pm 1.04/\sqrt{120})$

Pero para la media si lo sabemos hacer!

n=120, mean(tiempos)=1.11, median(tiempos)=0.77,
sd(tiempos)=1.04
... *mean value 1.11* $(\pm 1.04/\sqrt{120})$

¿Qué está pasando?
¿Qué estamos calculando?

Incertidumbre - Error de Estimación para $\widehat{\mu}_n$

- Parámetro de interés: $\mu = \mathbb{E}(X)$. Estimador: $\widehat{\mu}_n = \overline{X}_n$.
- Estimación: Promedio de mis datos; en R: `mean(datos)`
- Varianza y desvío estandar del estimador:

$$V(\widehat{\mu}_n) = \frac{V(X)}{n}\,, \quad \text{se} = \sqrt{V(\widehat{\mu}_n)} = \sqrt{\frac{V(X)}{n}}$$

- Incertidumbre -Error de Estimación
  1. Si $V(X) = \sigma_0^2$ es un valor conocido, el error de estimación es

     Incertidumbre - Error de estimación : $\text{se} = \sigma_0/\sqrt{n}$

  2. Si $V(X)$ es desconocida, estimamos se:

     Incertidumbre - Error de estimación : $\widehat{\text{se}}_{\text{obs}} = S_{\text{obs}}/\sqrt{n}$

     en R : `sd(datos)/sqrt(length(datos))`

Error de estimación - Uncertainty - Incertidumbre

Definición: llamamos incertidumbre (de la estimación) o error de una estimación a la estimación del desvío (exacto o aproximado) del estimador con el cual estimamos.

- Estimador: $\widehat{\theta}_n$
- Estimación: $\widehat{\theta}_{n,\text{obs}}$
- se $= \sqrt{V(\widehat{\theta}_n)}$ o se $\approx \sqrt{V(\widehat{\theta}_n)}$
- Incertidumbre o Error de Estimación: $\widehat{\text{se}}_{\text{obs}}$

Incertidumbre - Error de Estimación para $\widehat{p}_n$

- $Y_i \sim \mathcal{B}(1, p)$. Parámetro de interés: $p = \mathbb{E}(Y)$.
- Estimador: $\widehat{p}_n = \overline{Y}_n$.
- Estimación: Promedio de mis datos; en R: `mean(datos)`
- Varianza y desvío estandar del estimador:

$$V(\widehat{p}_n) = \frac{V(Y)}{n}, \quad \text{se} = \sqrt{V(\widehat{p}_n)} = \sqrt{\frac{V(Y)}{n}} = \sqrt{\frac{p(1-p)}{n}}$$

Incertidumbre - Error de estimación :

$$\widehat{\text{se}}_{\text{obs}} = \sqrt{\widehat{p}_{n,\text{obs}}(1 - \widehat{p}_{n,\text{obs}})/n}$$

en R : `mean(datos)*(1-mean(datos))/sqrt(length(datos))`

Hipótesis: perros y dueños se parecen – n=25

© Kevin Klöpper/iStockphoto

Escenario 1: A $n = 25$ encuestados se le muestra la foto del dueño y dos mascotas, siendo una de ellas la propia. 15 personas aciertan (identifican la mascota del dueño).

Hipótesis: perros y dueños se parecen – n=250

© Kevin Klöpper/iStockphoto

Escenario 2: A $n = 250$ encuestados se le muestra la foto del dueño y dos mascotas, siendo una de ellas la propia. 150 personas aciertan (identifican la mascota del dueño).

Mundo Bernoulli- la importancia del n

- Escenario 1: estimación: 0.6 y $n = 25$.

  Error de estimación:

- Escenario 2: estimación: 0.6 y $n = 250$.

  Error de estimación:

Mismo valor de la estimación pero calculada con otro n.
Puede cambiar la conclusión.

Intervalo

¿Y para los percentiles?

- ¿Cuál es la **"incertidumbre** (o **"error"**) de la mediana?
- ¿Cómo hacemos para asingar incertirudumbre a un percentil muestral?

$$V(\widehat{q}_{n,\alpha})?$$

Error de estimación para la mediana?

- Desvio del Estimador:

$$\text{se} = \text{se}(\text{med}(X_1, \ldots, X_n)) = \sqrt{V\{\text{med}(X_1, \ldots, X_n)\}} = ???$$

- $\widehat{\text{se}} = ??$
- ¿Qué nos gustaria poder hacer?
- ¿Podemos aprovechar algo de lo que hicimos en esta clase?

Error de estimación para la mediana?

- Desvio del Estimador:

$$\text{se} = \text{se}(\text{med}(X_1, \ldots, X_n)) = \sqrt{V\{\text{med}(X_1, \ldots, X_n)\}} = ???$$

- $\widehat{\text{se}} = ??$
- Y si no tenemos a ayuda de los compañeros, ¿qué hacemos?

Error de estimación para la mediana?

- Desvio del Estimador:

$$\text{se} = \text{se}(\text{med}(X_1, \ldots, X_n)) = \sqrt{V\{\text{med}(X_1, \ldots, X_n)\}} = ???$$

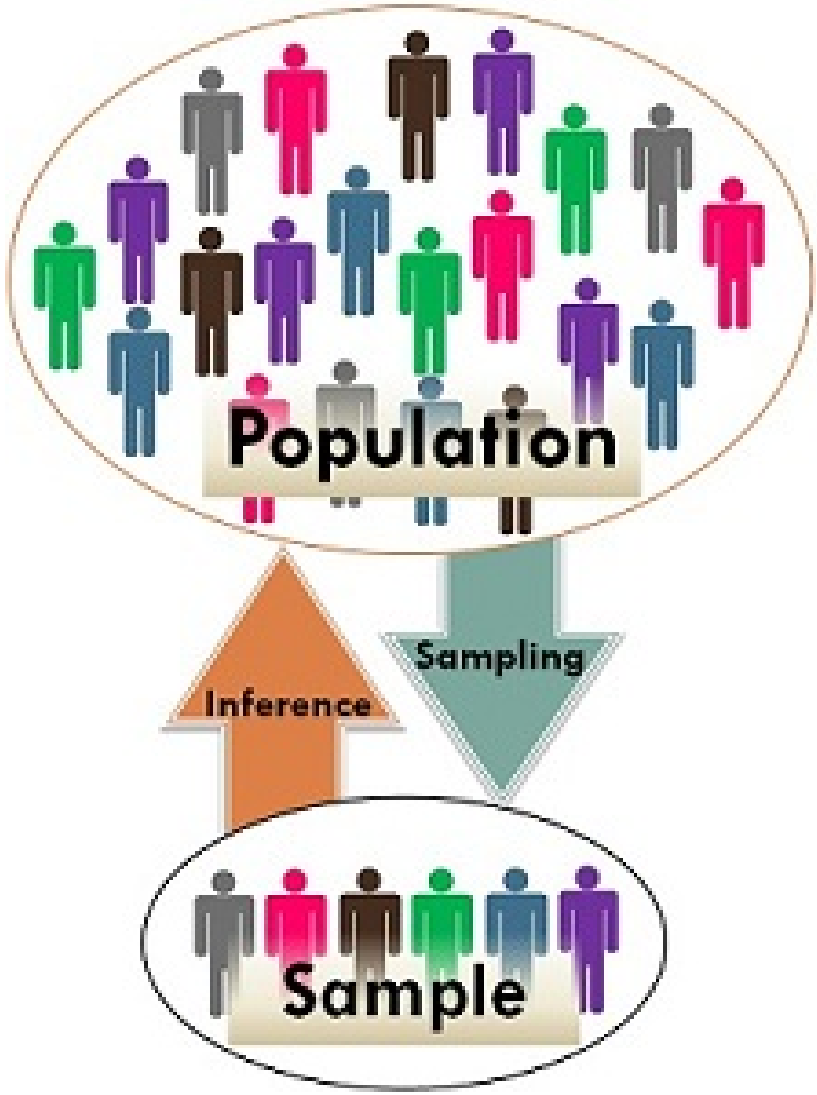
- $\widehat{\text{se}} = ??$
- Y si no tenemos a ayuda de los compañeros, ¿qué hacemos?
- Bootstrap! $\widehat{\text{se}}_{boot}$

Remuestreo - Bootstrap

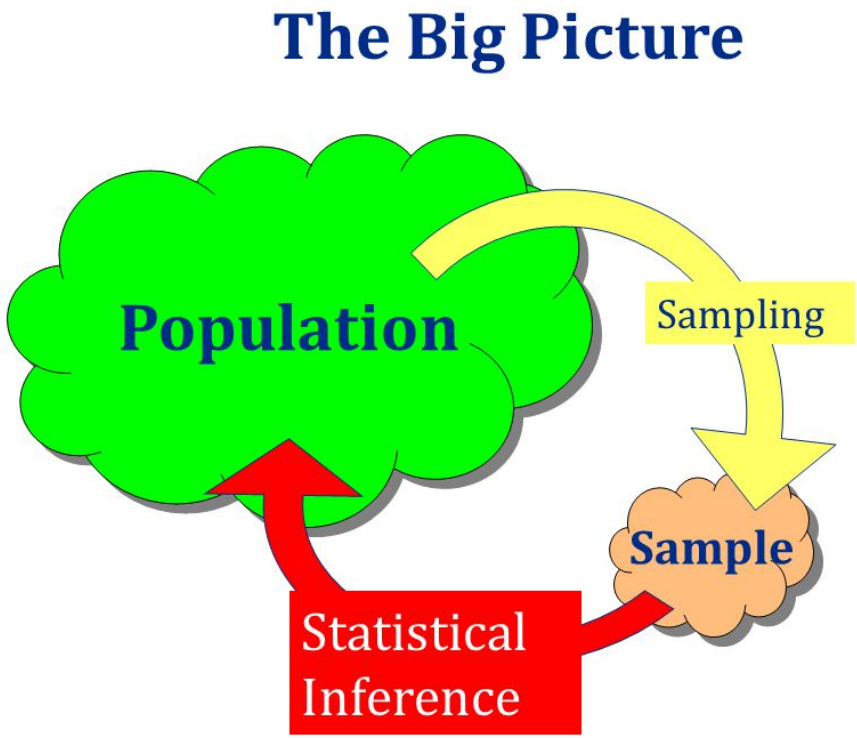
Definiciones

- Población Biológica
- Población Estadística: Distribución F que *dá* origen a los datos.
- La Población Estadística es un modelo matemático para la Población Biológica.
- Si conocemos la *POBLACION* F, no necesitamos hacer estadística.
- Se conocemos la *POBLACION* F, podemos caclular cualquier parámetro de interés asociado a F, haciendo cuentas de proba.
- Muestra $X_1, \ldots, X_n$ iid, $X_i \sim F$.

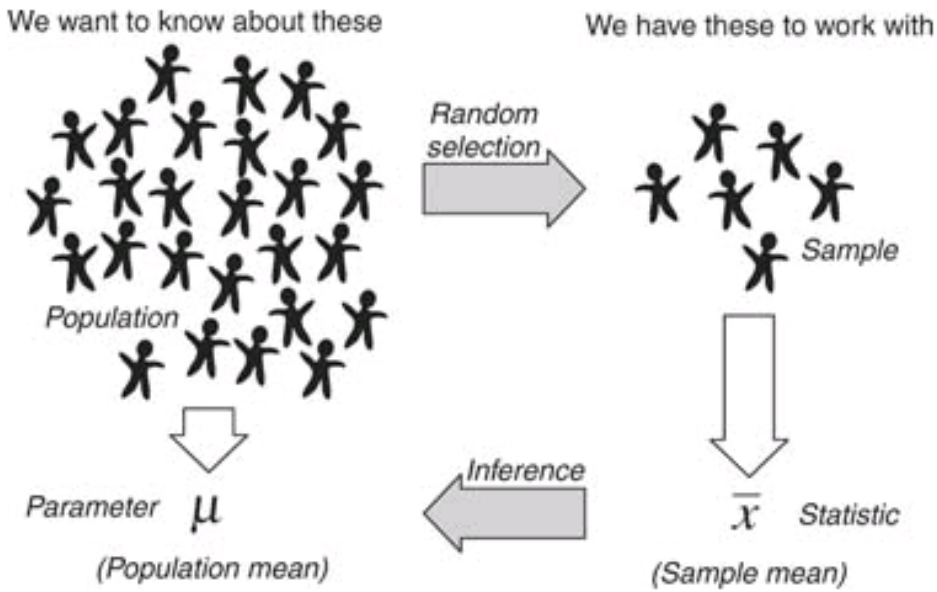
Población vs. Muestra



Población vs. Muestra - En forma abstracta...



Población vs. Muestra -



Error de estimación

Definición: llamamos error de una estimación a la estimación del desvío (exacto o aproximado) del estimador con el cual estimamos.

- Estimador: $\widehat{\theta}_n$
- Estimación: $\widehat{\theta}_{n,\text{obs}}$
- se $= \sqrt{V(\widehat{\theta}_n)}$ o se $\approx \sqrt{V(\widehat{\theta}_n)}$
- Error de estimación: $\widehat{\text{se}}_{\text{obs}}$

Errores de estimación

$$\widehat{\mu}_n\ , \qquad \mathsf{se}^2 = V(\widehat{\mu}_n) = \sigma^2/n\ , \quad \widehat{\mathsf{se}} = \sqrt{S^2/n}\ ,$$

Errores de estimación

$$\widehat{\mu}_n\ , \qquad \mathsf{se}^2 = V(\widehat{\mu}_n) = \sigma^2/n\ , \ \ \widehat{\mathsf{se}} = \sqrt{S^2/n}\ ,$$

$$\widehat{p}_n\ , \ \ \mathsf{se}^2 = V(\widehat{p}_n) = p(1-p)/n\ , \ \widehat{\mathsf{se}} = \sqrt{\widehat{p}_n(1-\widehat{p}_n)/n}\ ,$$

Error de estimación para la mediana?

- Desvio del Estimador:

$$\text{se} = \text{se}(\text{med}(X_1, \ldots, X_n)) = \sqrt{V\{\text{med}(X_1, \ldots, X_n)\}} = ???$$

- $\widehat{\text{se}} = ??$

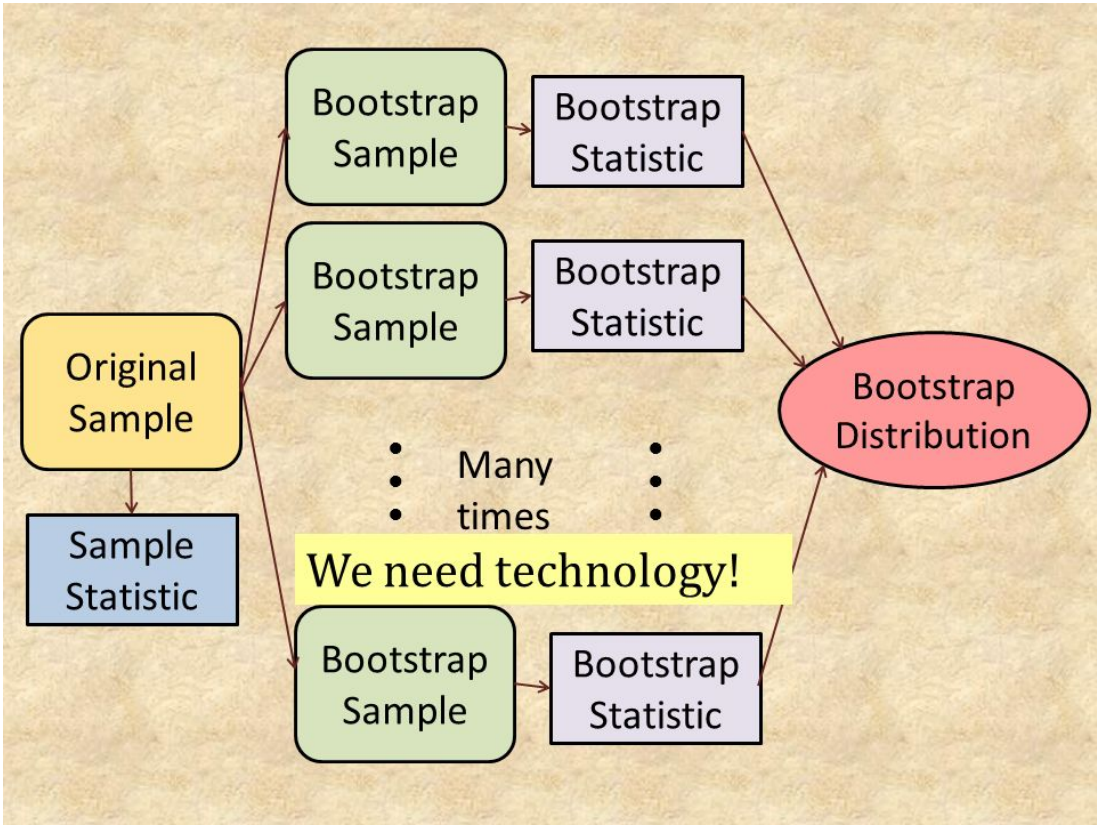
Error de estimación para la mediana?

- Desvio del Estimador:

$$\text{se} = \text{se}(\text{med}(X_1, \ldots, X_n)) = \sqrt{V\{\text{med}(X_1, \ldots, X_n)\}} = ???$$

- $\widehat{\text{se}} = ??$
- Bootstrap! $\widehat{\text{se}}_{boot}$

Esquema Bootstrap



Esquema Bootstrap. All of Statistics. Wasserman. Cap 8.

Bootstrap Variance Estimation

1. Draw $X_1^*, \ldots, X_n^* \sim \widehat{F}_n$.
2. Compute $T_n^* = g(X_1^*, \ldots, X_n^*)$.
3. Repeat steps 1 and 2, B times, to get $T_{n,1}^*, \ldots, T_{n,B}^*$.
4. Let

$$v_{\text{boot}} = \frac{1}{B}\sum_{b=1}^{B}\left(T_{n,b}^* - \frac{1}{B}\sum_{r=1}^{B}T_{n,r}^*\right)^2. \tag{8.1}$$

Bootstrap for The Median

```
Given data X = (X(1), ..., X(n)):

T <- median(X)
Tboot <- vector of length B
for(i in 1:B){
    Xstar <- sample of size n from X (with replacement)
    Tboot[i] <- median(Xstar)
    }
se <- sqrt(variance(Tboot))
```

Dejamos aca una linda referencia.