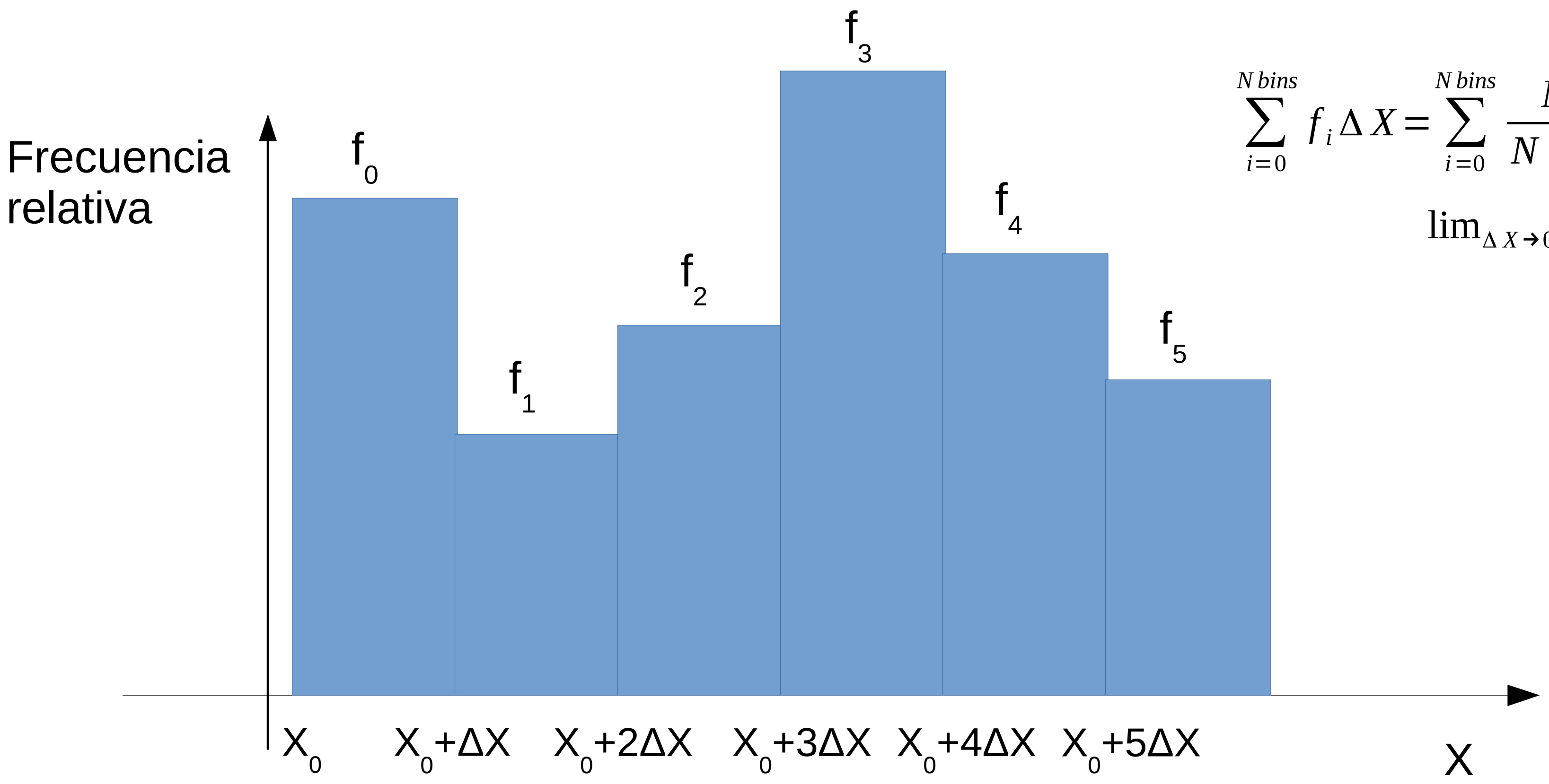
Introducción a las técnicas de Bootstrap

Santander, 2019-2020

Técnicas de remuestreo

- En muchas ocasiones es posible definir estrategias **sobre la muestra obtenida**, que permiten estimar o mejorar los sesgos de nuestros estimadores (o procedimientos).
- **No conocemos de forma exacta la pdf del proceso → la muestra "aproxima" esta pdf.**
- Consideremos una variable aleatoria X con una densidad de probabilidad F(X).
- Consideremos también una muestra de "N" medidas $\{X_1, X_2, \ldots, X_N\}$.
- Podemos colocar estas N medidas en un histograma de frecuencias.

Relación entre pdf e histograma de frecuencias

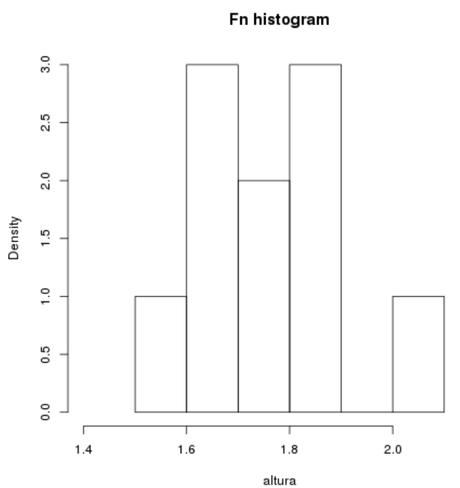


$$\sum_{i=0}^{N\,bins} f_i \Delta X = \sum_{i=0}^{N\,bins} \frac{N_i}{N \Delta X} \Delta X = 1$$

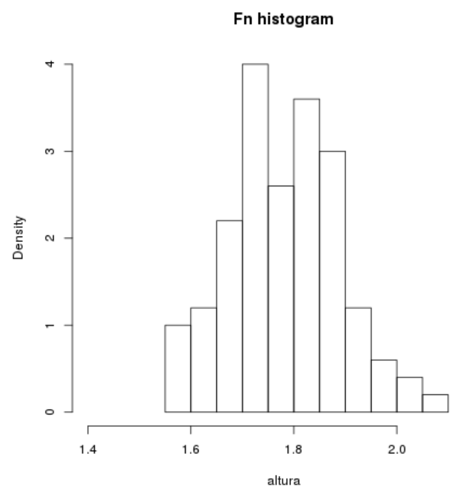
$$\lim_{\Delta X \to 0, N \to \infty} F_N = F$$

- Este histograma nos dice la densidad de sucesos que cayeron en un bin determinado.
- Si el tamaño del bin fuese cero y hubiese infinitas medidas este histograma sería la pdf.

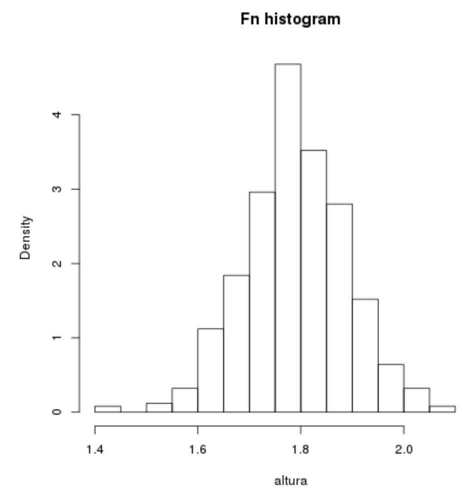
Relación entre pdf e histograma de frecuencias



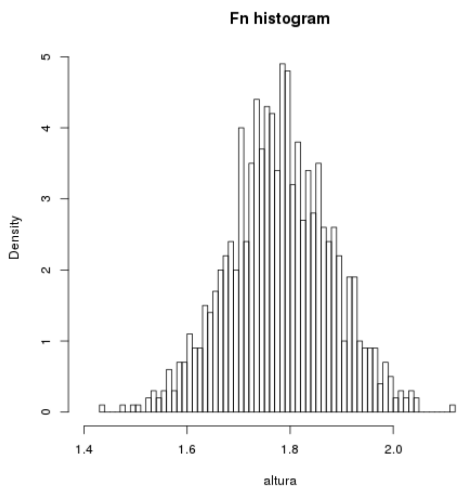
N=10, Bin=0.15m



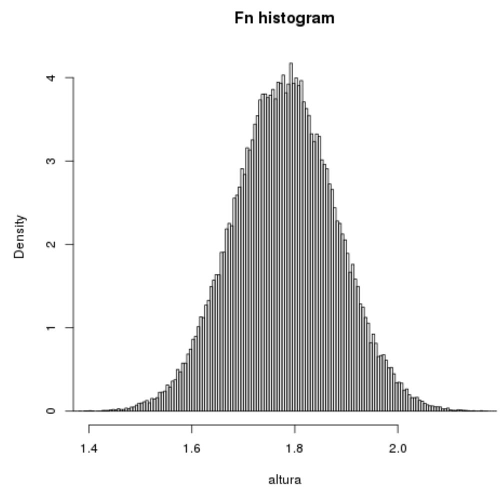
N=100, Bin=0.08m



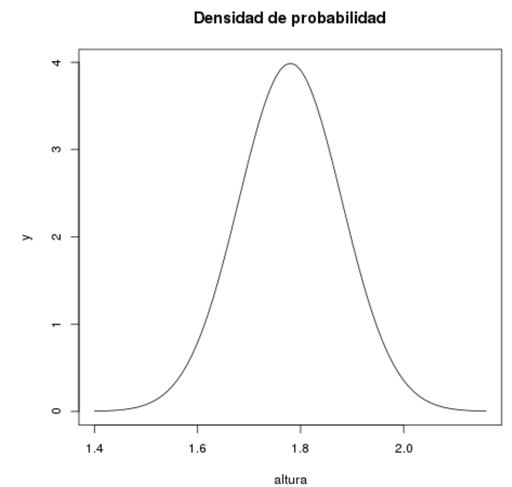
N=500, Bin=0.08m



N=1000, Bin=0.02m



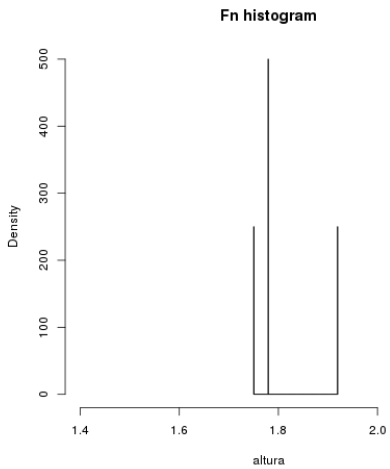
N=100000, Bin=0.004m

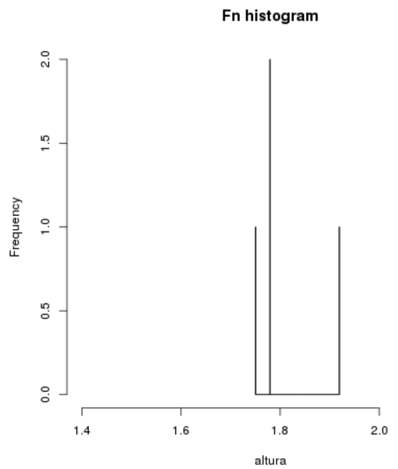


Relación entre pdf e histograma de frecuencias

- Es posible aproximar la pdf de un proceso por el histograma de densidad de frecuencias.
- Para ello es necesario construir Fn con un tamaño de bin infinitesimalmente pequeño.
- Cuando hacemos esto básicamente cada valor medido diferente $\{X_1,X_2,\ldots,X_N\}$ ocupa un bin.
- Ejemplo: volviendo al ejemplo de las alturas de los primeros capítulos en donde teníamos {1.78, 1.78, 1.92, 1.75}.
- Fn no es una aproximación perfecta de la pdf → pero permite simular pseudo-muestras (toys)

Es posible aplicar técnicas de MC sobre el histograma de frecuencias/fracción de frecuencias para generar pseudo-muestras.





Técnica bootstrap

- Supongamos un parámetro Θ medido sobre una muestra obteniéndose un estadístico $\overline{\Theta}$
- Si fuésemos capaces de conseguir M nuevas muestras y computar en cada una $\overline{\Theta}_i$, entonces

$$\overline{\overline{\Theta}} = \frac{1}{M} \sum_{i=1}^{M} \overline{\Theta}_i$$

- Este estimador contiene mucha más información que $\overline{\Theta}$ que sólo involucra una muestra y su covarianza es menor.
- Tomar M nuevas muestras resulta costoso, sin embargo obtenido Fn de la muestra tomada, es posible generar M nuevas pseudomuestras (bootstrap) sampleando dicha distribución.
- Siendo posible demostrar que la distribución de la variable de bootstrap generada así se distribuye de modo que

$$\overline{\Theta}_B - \overline{\Theta} \approx \overline{\Theta} - \Theta$$

Efron, B. (1979). "Bootstrap methods: Another look at the jackknife". The Annals of Statistics. 7 (1): 1–26. doi:10.1214/aos/1176344552.

¿Cómo hacer el resampling en bootstrap?

- Puesto que nuestra distribución de frecuencias Fn tiene un bin para cada valor diferente...
- ...generar una nueva pseudomuestra consiste en sustituir cada uno de los elementos de la muestra por cualquier otro elegido aleatoriamente de la propia muestra hasta completar N

$$X = \{X1, X2, X3, X4, X5, X6, X7, X8, X9, X10, X11, X12, X13\}$$

$$X_B^{1} = \{X3, X13, X8, X1, X13, X4, X12, X10, X11, X8, X3, X2, X1\}$$

- Cada una de las nuevas psuedomuestras tiene el mismo número de elementos que la original
- No todos los elementos de la muestra original tienen por qué aparecer en una pseudomuestra
- Algunos elementos de la muestra orginal pueden aparecer repetidos en una pseudomuestra

Aplicación 1: Aproximación del error standard de un estadístico

- Supongamos un parámetro Θ del que hemos obtenido el estimador $\overline{\Theta}$ y queremos conocer su desviación standard

$$SD=\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\overline{\Theta}_i-\Theta)^2}$$

- Cada uno de los $\overline{\Theta}_i$ se correspondería con el valor obtenido en un experimento diferente.

- Hacer diferentes experimentos resulta costoso, así que podemos estimarlo usando bootstrapping ya que:

$$\overline{\Theta}_B-\overline{\Theta}\approx\overline{\Theta}-\Theta$$

- Utilizando la Fn de la muestra que ha dado lugar a $\overline{\Theta}_i$ podemos generar M muestras de tipo bootstrap.

- Y podemos sustituir en la fórmula $\overline{\Theta}_i$ por cada una de las muestras bootstrap y Θ por $\overline{\Theta}$

$$SD=\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\overline{\Theta}_i^B-\overline{\Theta})^2}$$

Aplicación 2: Aproximación del sesgo de un estadístico

- Tal y como hemos definido previamente el sesgo de un estimador viene dada por la cantidad

$$Sesgo(\overline{\Theta}) = E[\overline{\Theta} - \Theta] = E[\overline{\Theta}] - \Theta$$

- Es posible utilizar una vez más la igualdad vista anteriorment entre las distribuciones anteriores

$$\overline{\Theta_B} - \overline{\Theta} \approx \overline{\Theta} - \Theta$$

- Para aproximar el sesgo sustiyendo por las variables de bootstrap y dando el siguiente resultado

$$Sesgo(\overline{\Theta}) = E[\overline{\Theta} - \Theta] \approx \frac{1}{N} \sum_{i=0}^{M} \Theta_i^B - \overline{\Theta}$$

El método jackknife

- La técnica de bootstraping es la más extendida aunque existió otro algoritmo muy popular previo.
- Se trata del **jackknife,** muy utilizado cuando los recursos de cálculo eran limitados.
- Las ideas detrás de esta técnica son las mismas que las asociadas al bootstrap.
- La diferencia consiste en la manera de construir los nuevo datasets.
- En Jackknife cada uno se corresponde simplemente con la muestra formada tras quitar 1 elemento.
- Jackknife es una opción menos potente que Bootstrap y en general se aconseja el otro.

$$SD=\sqrt{\frac{\boxed{N}}{N-1}\sum_{i=1}^{N}\left(\overline{\Theta}^{JK_i}-\overline{\Theta}\right)^2}$$

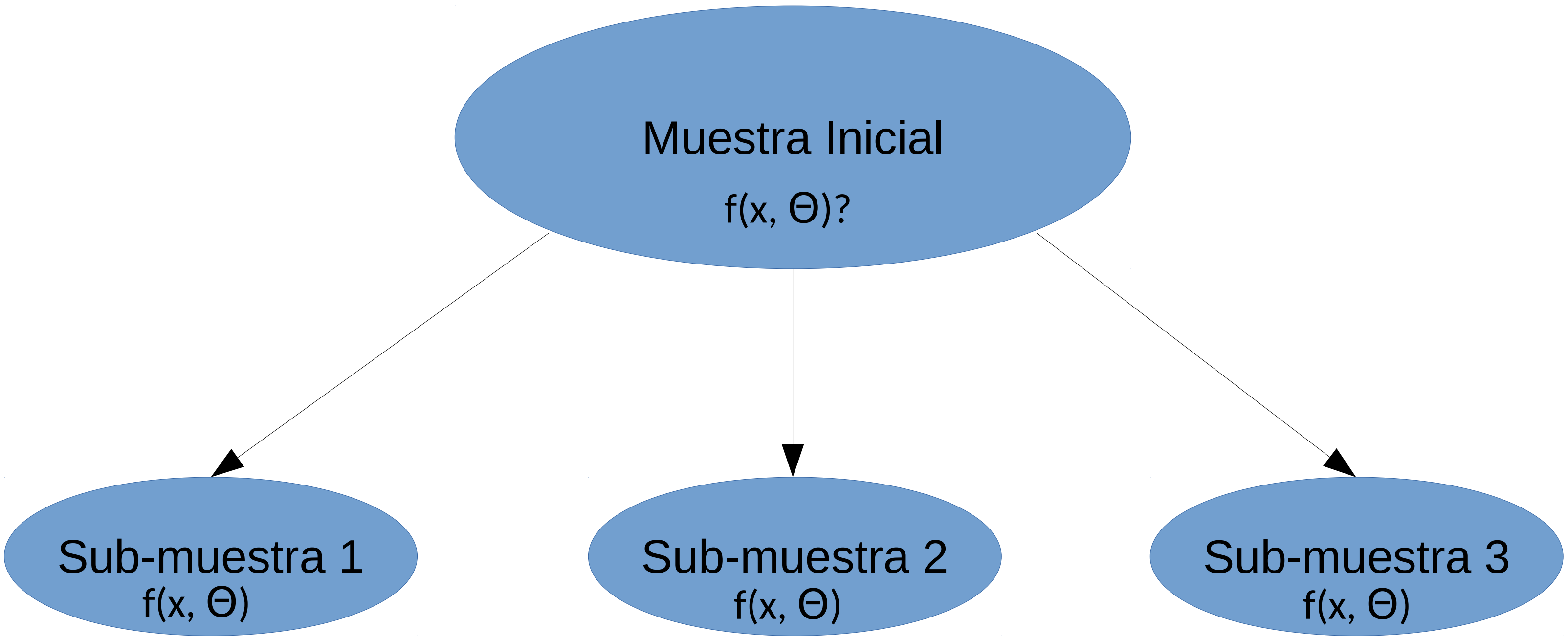
Las muestras tienen tamaño N-1

$$\{X1, X2, X3, X4, X5, X6, X7, X8\} \left\{ \begin{array}{l} \{X1, X2, X3, X4, X5, X6, X7\} \\ \{X1, X2, X3, X4, X5, X6, X8\} \\ \{X1, X2, X3, X4, X5, X7, X8\} \\ \{X1, X2, X3, X4, X6, X7, X8\} \end{array} \right.$$

Cross validation

- Cross validation es una técnica que consiste en la partición de una muestra en **N bloques**...
- ...si la muestra se rige por una pdf determinada cada uno de los N bloques también lo hace.
- Y esto permite comprobar los resultados obtenidos en un bloque con los de los otros bloques.
- Supongamos que "creemos" que una muestra se distribuye según una pdf f(x, Θ).
- Supongamos también que contamos con algún tipo de estimador del parámetro Θ.
- ¿Cómo estar seguros de que el estimador tiene sentido si en realidad no conocemos f?
- Podemos estimarlo en el bloque 1, para luego compararlo con el resto de bloques.
- Si las hipótesis son correctas, los resultados deberían ser consistentes entre bloques.
- Es exactamente lo mismo que hemos hecho con la regresión lineal y el modelado.

Cross validation



- El concepto de cross-validation se utiliza en problemas de modelado, como ya vimos.
- Normalmente una de las muestras es llamada de "training" ya que sobre ella se estiman los parámetros, mientras que otra se suele llamar de "test" ya que sobre ella se comprueba la calidad del modelado. En ocasiones se habla también de muestra de "validación".

Ejercicio 2

- Escribe una función de R que reciba un vector de números x (la muestra), y genere **una** muestra bootstrap de ese vector.
- Utilizando la función anterior, escribe una función que reciba un vector de números x (la muestra original) y un número natural N, y que genere una matrix que tenga N columnas, siendo cada una una de las muestras de bootstrap.
- Escribe una función de R que reciba un vector de números x (la muestra) y genere una matriz que contenta TODAS las muestras jacknife con el mismo formato del ejercicio anterior.
- Utilizando las funciones anteriores considera la desviación estándar de la media muestral para una muestra de N=10000 que se distribuya como en el ejercicio 1 (gaussiana centrada en 1.70m y sigma=1.7). Compara la desviación estándar obtenida, con la obtenida con bootstrap.
- Repite el ejercicio anterior utilizando la técnica jackknife. ¿Cuál da mejor resultado?