Estadística [continuación]

Santander, 2018-2019

Temario y estructura del curso (I)

- Tema 5. Modelos de regresión. Estimación de máxima similitud.
  - [T.5.1] El problema de la estimación a través de una muestra limitada de la distribución.
  - [P.5.2] Estimación y comparación de estadísticos en muestras de una misma distribución.
  - [T.5.3] Concepto de regresión. Función de coste. Algoritmos de minimización.
  - [P.5.4] Práctica: estimación de función de coste y minimización.
  - [T.5.5] Regresión lineal en una dimensión. Ajuste de polinomios. Sesgo y varianza.
  - [P.5.6] Práctica: implementación de un ajuste lineal y predicción. Estudio de sesgo y varianza.
  - [T.5.7] Regresión logística.
  - [P.5.8] Práctica: implementación de una regresión logística.
  - [T.5.9] Uso de cross-validation para entender las propiedades de una regresión.
  - [P.5.10] Práctica: Basado en P.5.10 estudiar las propiedades usando cross-validation
  - [T.5.11] Estimación de máxima similitud. Distribución chi2.
  - [P.5.12] Práctica: Aplicación de máxima similitud para realizar ajuste.
  - [T.5.13] Estadístico G2 statistic y bondad de un ajuste.
  - [P.5.14] Práctica: Bondad de un ajuste para diferentes modelos.

Temario y estructura del curso (I)

- Tema 5. Técnicas de remuestreo (bootstrap)
  - [T.6.1] Introducción a las técnicas de remuestreo. Conceptos básicos.
  - [T.6.2] Algoritmo de remuestreo BootStrap
  - [T.6.3] Algoritmo de remuestreo Jacknkife
  - [T.6.4] Conceptos generales de cross validation.
  - [P.6.5] Utilización de la técnica Bootstrap/Jacknife para mejorar la estimación de estadísticos sencillos.

¿Qué debería saber tras esta parte del curso?

- Trabajar con estimadores basados en muestras finitas y a estimar sus sesgos.
- Realizar ajustes de datos a modelos paramétricos sencillos.
- Clasificar diferentes categorías de datos con modelos sencillos.
- Entender los problemas asociados a los ajustes y clasificaciones, y su mejora.

Tema 5. Modelos de regresión. Estimación de máxima verosimilitud.

"I couldn't claim that I was smarter than sixty-five other guys--but the average of sixty-five other guys, certainly!"

— Richard P. Feynman,
Surely You're Joking, Mr. Feynman!: Adventures of a Curious Character

"If your experiment needs a statistician, you need a better experiment."

— Ernest Rutherford

Repaso de estadística: elementos abstractos

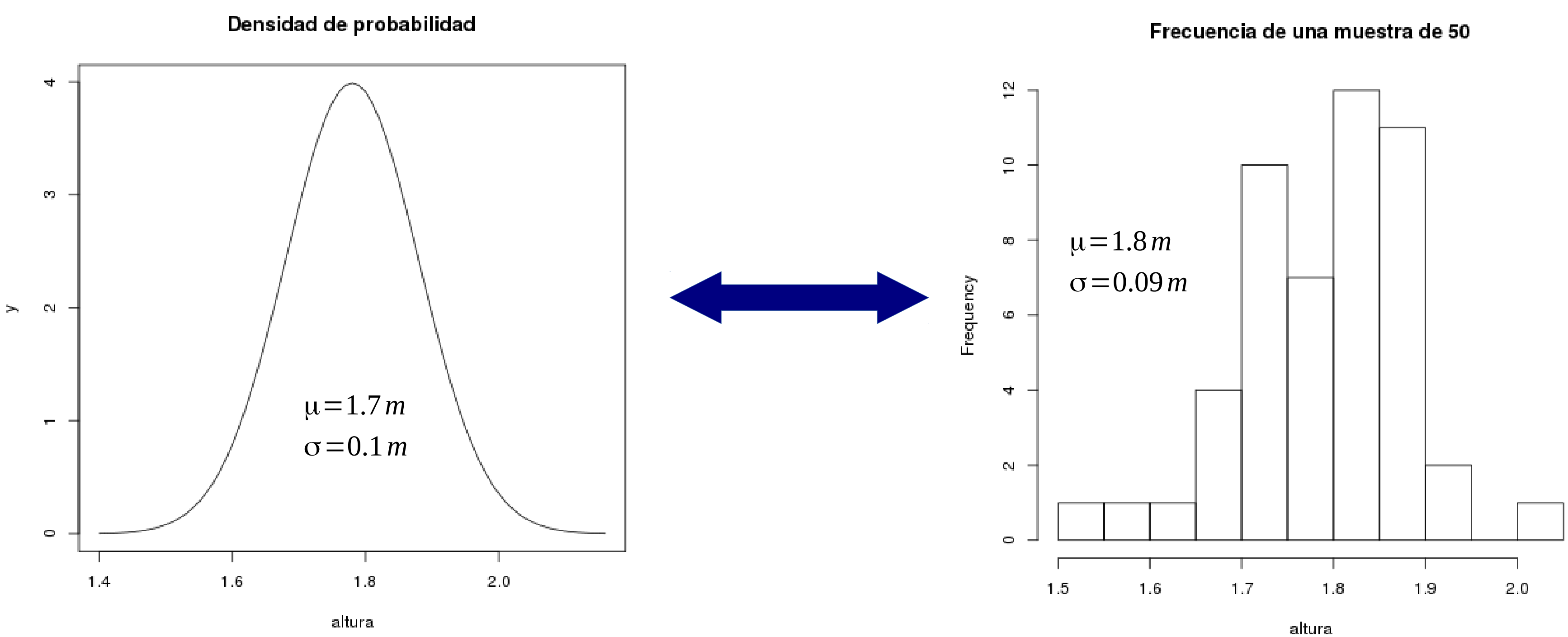
- Población: Conjunto finito o infinito que incluye la totalidad de los elementos de estudio. Ejemplo: las personas con edades mayores de 20 años.
- Variable aleatoria: Una función que asigna un valor numérico al resultado de un experimento aleatorio sobre una población. Ejemplo: La altura de las personas con edades mayores de 20 años.
- Probabilidad: frecuencia o recurrencia de un valor concreto de una variable aleatoria cuando se realizan infinitos experimentos. Ejemplo: La probabilidad de que una persona con edad mayor de 20 años mida 1.80m es 0.1.
- Densidad de probabilidad: Expresión matemática que asigna una probabilidad concreta a un rango infinitesimal de una variable aleatoria. Ejemplo: La variable aleatoria “altura de las personas con edades mayores de 20 años” sigue una distribución gaussiana.
- Parámetro estadístico: Valor o característica numérica representativa de una población. Ejemplo: el promedio de altura para las personas con edad mayor de 20 años es 1.78m.

Repaso de estadística: elementos medibles

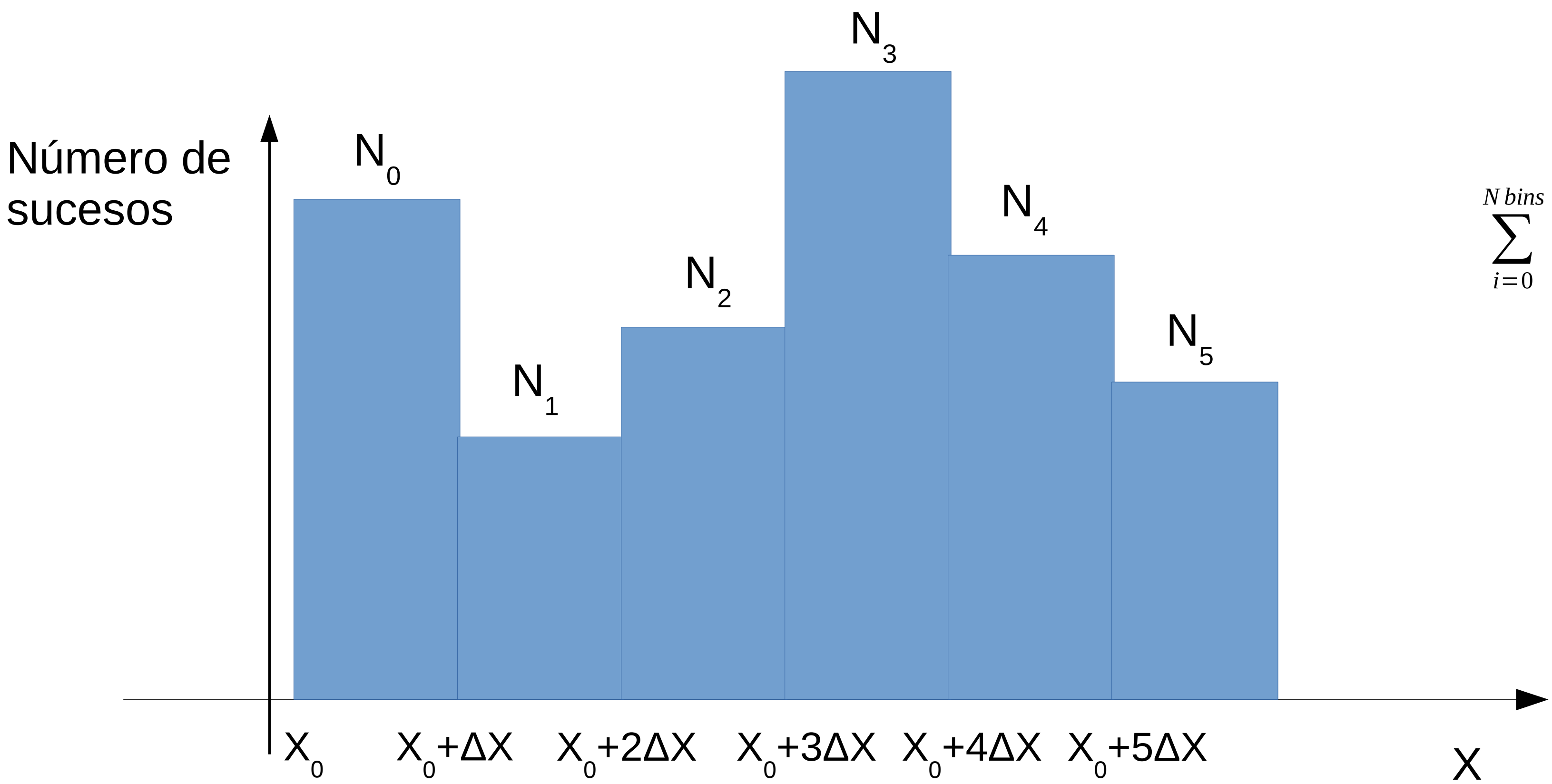
- Muestra: Subconjunto finito y concreto de la Población. Ejemplo: un grupo de personas con edades mayores de 20 años formado por 4 individuos concretos: {María, Luis, Pepe, Diana}.
- Realización de una variable: Valor numérico concreto obtenido X_i como resultado de un experimento concreto. Ejemplo: La altura de María es 1.78m, la de Luis 1.78m, la de Pepe 1.92m, la de Diana 1.75m.
- Frecuencia: frecuencia o recurrencia de un valor concreto de una variable aleatoria sobre una muestra concreta. Ejemplo: La frecuencia de la altura 1.78m en la muestra anterior es: 2.
- Distribución de frecuencias: Función que asigna la frecuencia observada para cada uno de los elementos de la muestra. Ejemplo: La distribución de frecuencias en la muestra anterior es: 1 para 1.75m, 2 para 1.78m y 1 para 1.92m.
- Estadístico/estimador: Valor o característica numérica representativa de una muestra. Ejemplo: el promedio de altura para la muestra anterior es ~ 1.81m.

Repaso de estadística: muestreo

- La estadística pretende obtener información acerca de algún parámetro de una población.
- Sin embargo típicamente no tenemos acceso a toda la población ni tampoco a un número infinito de experimentos → tan sólo podemos acceder a una muestra finita.



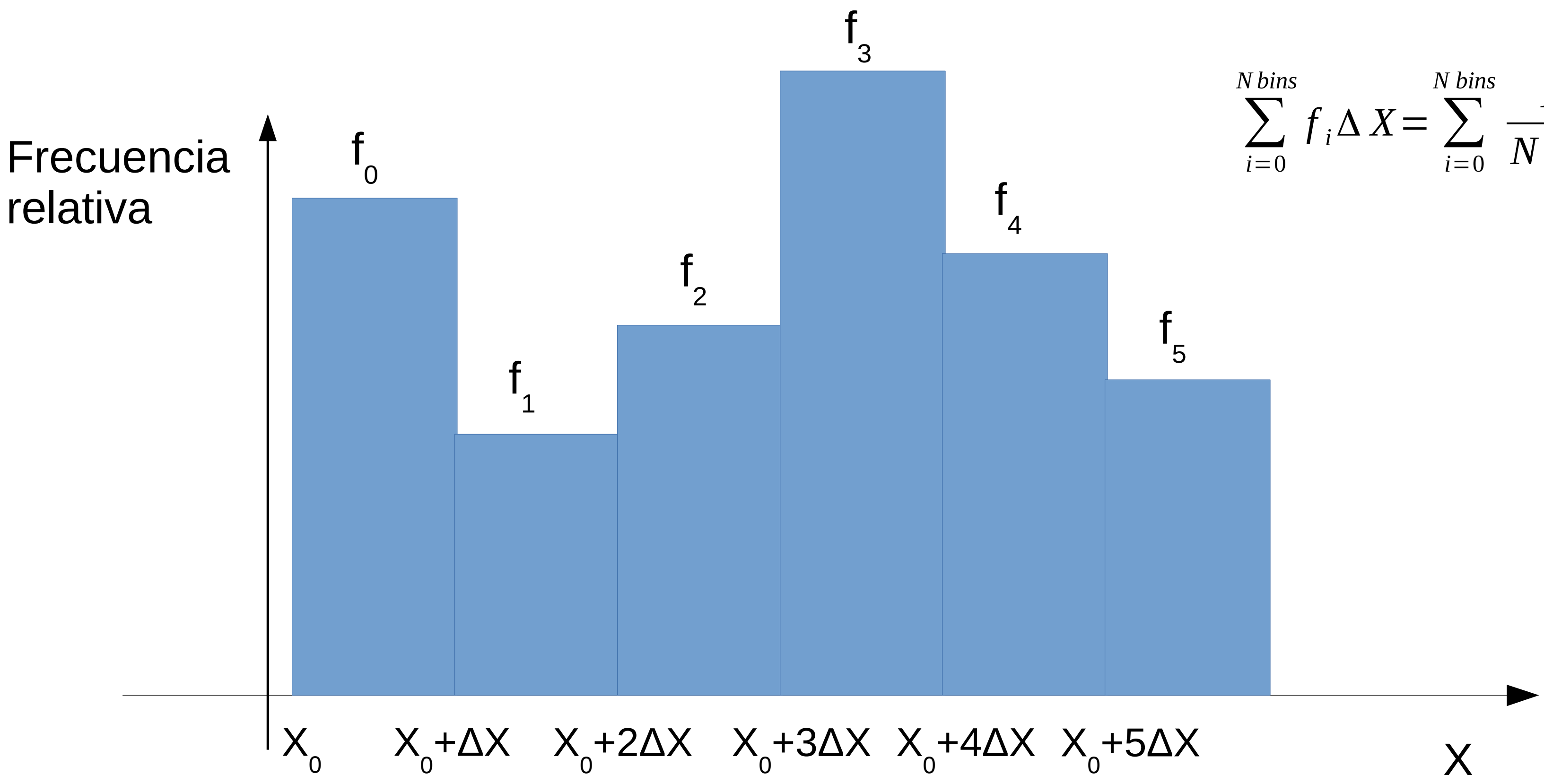
Relación entre pdf e histograma de frecuencias (I)



$$\sum_{i=0}^{N\,bins} N_i = N$$

- En este histograma ΔX es la anchura del bin y N_i es el número de medidas en cada bin.
- Cada bin nos dice cuántas medidas están dentro del intervalo $X_i + \Delta X$.

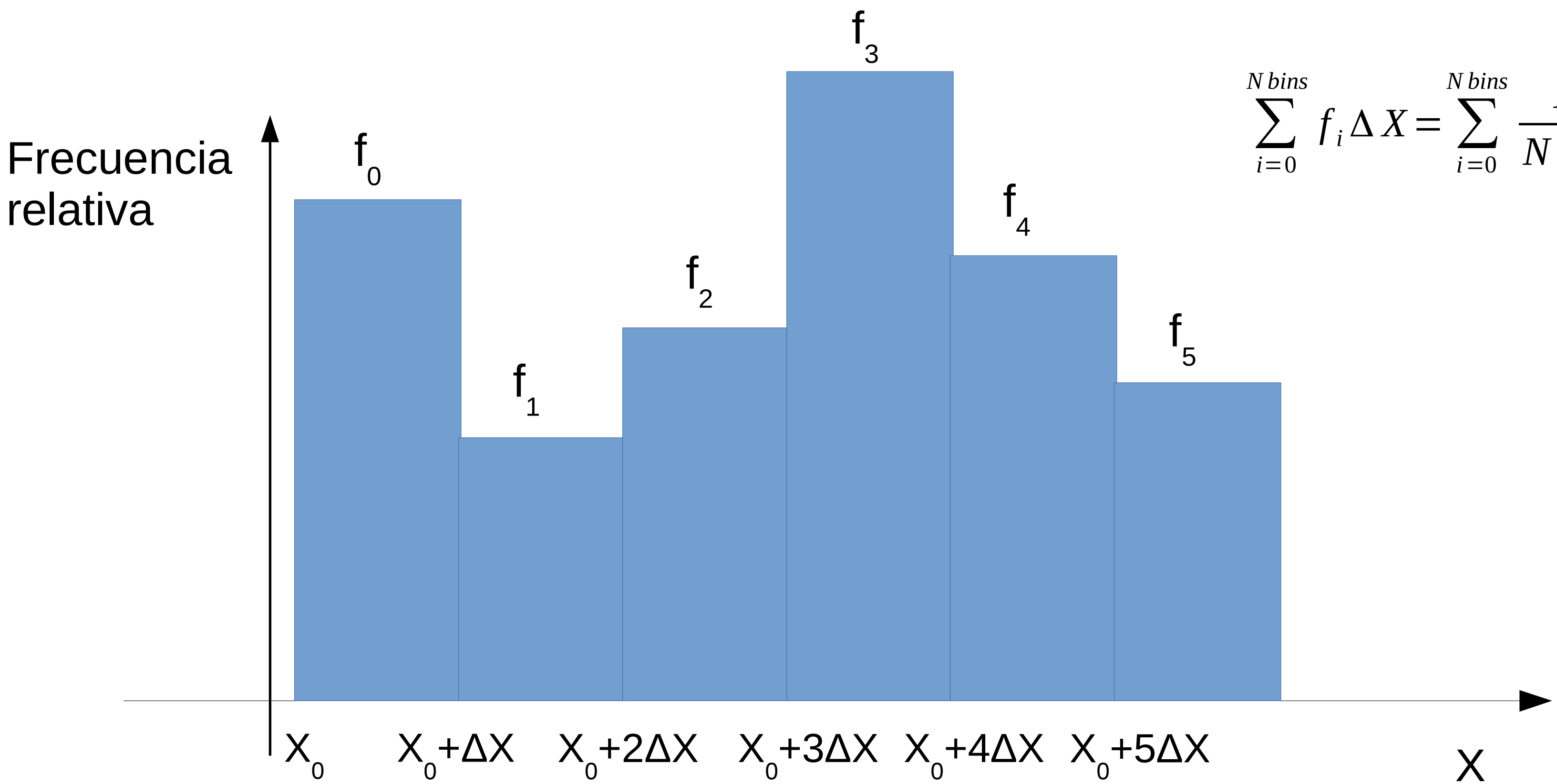
Relación entre pdf e histograma de frecuencias (II)



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

- Podemos obtener la densidad de frecuencia relativa en cada bin dividiendo por el número total y ΔX.
- Esta distribución comienza a guardar similitud con una pdf: en particular su integral es 1.

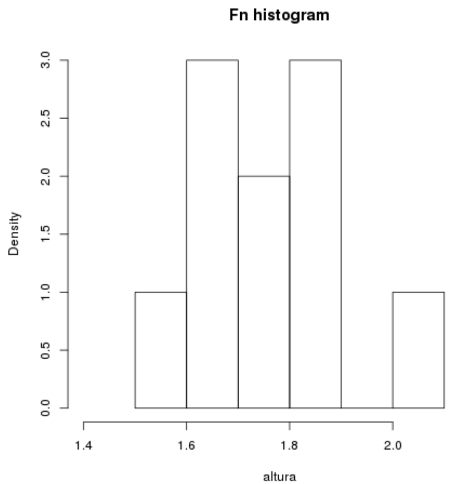
Relación entre pdf e histograma de frecuencias (III)



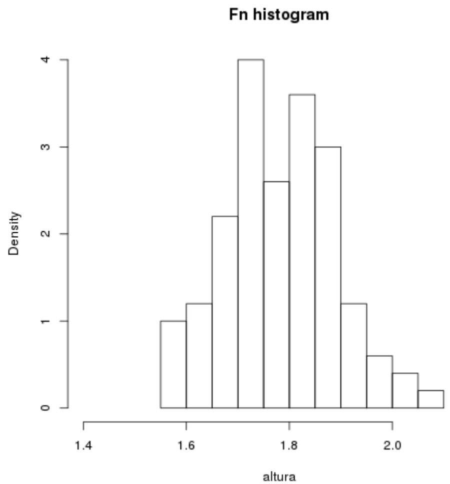
$$\sum_{i=0}^{N\,bins} f_i \Delta X = \sum_{i=0}^{N\,bins} \frac{N_i}{N \Delta X} \Delta X = 1$$

- 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.

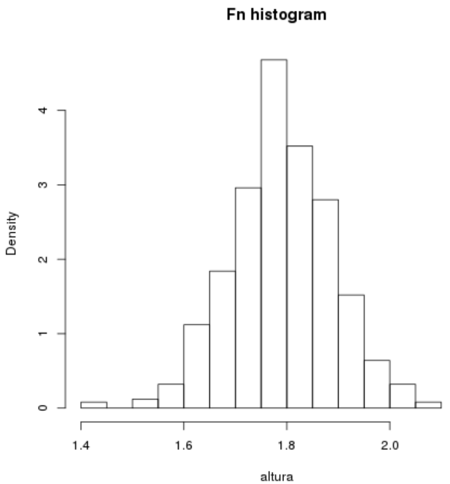
Relación entre pdf e histograma de frecuencias (IV)



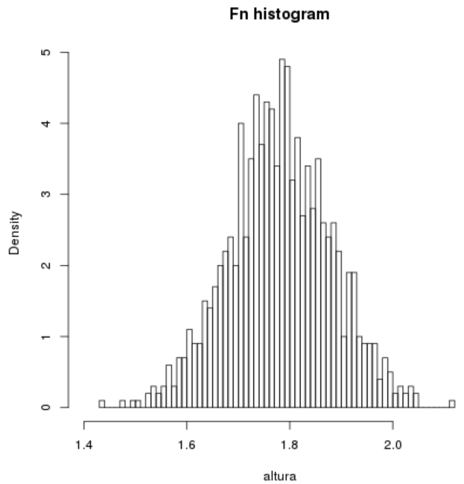
N=10, Bin=0.15m



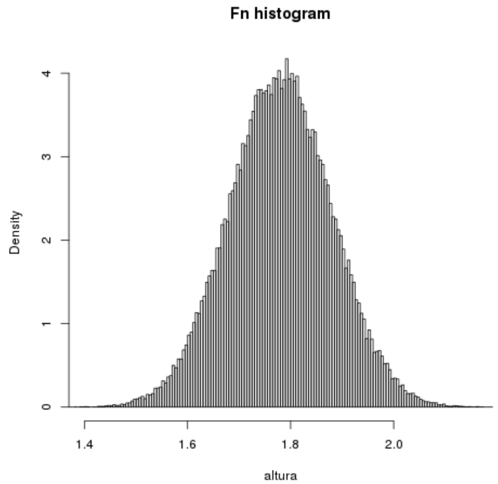
N=100, Bin=0.08m



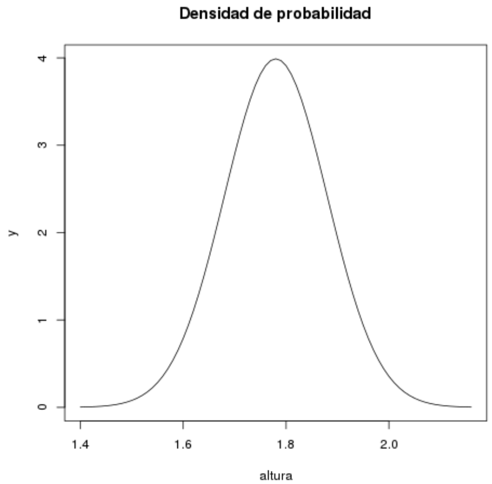
N=500, Bin=0.08m



N=1000, Bin=0.02m



N=100000, Bin=0.004m



Estimadores y sus propiedades (I)

- Los estimadores son una función de la muestra que nos permiten aproximar los parámetros.
- Obviamente estamos interesados en que el estimador sea lo más parecido posible a ese parámetro.
- Supongamos: estimar el parámetro Θ a través de un estimador T_n sobre una muestra de n elementos
- El error cuadrático medio es una medida de cuánto se diferencian estas dos cantidades

$$E[(\theta-T_n)^2]=E[\theta^2+T_n^2-2\theta T_n]=E[\theta]^2+E[T_n^2]-2E[\theta]E[Tn]$$

- Si sumamos y restamos a esa expresión el valor esperado de T_n al cuadrado E[T_n]^2

$$E[\theta]^2+E[T_n^2]-2E[\theta]E[Tn]+E[T_n]^2-E[T_n]^2=E[T_n^2]-E[T_n]^2+E[\theta-T_n]^2=Var(T_n)+Sesgo^2$$

- Si bien el parámetro es un valor fijo, el estimador dará un resultado diferente para cada experimento
- Por lo tanto podemos hablar de una varianza de dicho estimador.
- Por otra parte el estimador podría dar un valor esperado diferente al valor del parámetro.

Estimadores y sus propiedades (II)

- Cuando decidimos utilizar un estimador debemos tener en cuenta diferentes propiedades.
- El **sesgo** del estimador se define como el valor esperado de la diferencia entre T_n y el parámetro

$$E[\theta - T_n] = E[\theta] - E[T_n] = \theta - E[T_n]$$

- La **eficiencia** del estimador está relacionada con la varianza del mismo

$$Var[T_n] > Var[S_n] \rightarrow S_n \, mas \, eficiente$$

- La **consistencia** consiste en que cuando el tamaño de la muestra se aproxima a infinito:

$$\lim n \rightarrow \infty \, E[T_n] = \theta$$
$$\lim n \rightarrow \infty \, Var[T_n] = 0$$

- La **robustez** consiste en que el estimador toma valores similares al margen de la pdf del parámetro
- La **suficiencia** es la propiedad de un estimador que contiene toda la información sobre el parámetro

Ejemplo: la media muestral

$$\bar{X}=\frac{1}{N}\sum_{i=1}^{N} X_i$$

- El sesgo asociado a la media muestral para estimar el valor promedio es cero

$$E[\bar{X}]=E[\frac{1}{N}\sum_{i=1}^{N} X_i]=\frac{1}{N}\sum_{i=1}^{N} E[X_i]=\frac{1}{N}N\theta=\theta$$

- Sin embargo, la media muestral tiene una varianza distinta de cero

$$Var[\bar{X}]=Var[\frac{1}{N}\sum_{i=1}^{N} X_i]=\frac{1}{N^2}\sum_{i=1}^{N} Var[X_i]=\frac{1}{N^2}NVar[X]=\frac{Var[X]}{N}$$

- Puede verse también que la media es consistente ya que su valor esperado es el parámetro y…

$$\lim_{N\to\infty} Var[\bar{X}]=0$$

Ejemplo: la varianza muestral

$$S_n[X]=\frac{1}{N}\sum_{i=1}^{N}(X_i-\bar{X})^2$$

- El sesgo asociado a la varianza muestral para estimar la varianza es…

$$S_n[X]=\frac{1}{N}\sum_{i=1}^{N}(X_i-\bar{X})^2=\frac{1}{N}\sum_{i=1}^{N}((X_i-\mu)-(\bar{X}-\mu))^2=\frac{1}{N}\sum_{i=1}^{N}(X_i-\mu)^2+\frac{1}{N}\sum_{i=1}^{N}(\bar{X}-\mu)^2-\frac{2}{N}(\bar{X}-\mu)\sum_{i=1}^{N}(X_i-\mu)$$

$$E[\frac{1}{N}\sum_{i=1}^{N}(X_i-\mu)^2]=\frac{1}{N}\sum_{i=1}^{N}E[(X_i-\mu)^2]=\frac{1}{N}NVar[X]=Var[X]$$

$$E[\frac{1}{N}\sum_{i=1}^{N}(\bar{X}-\mu)^2]=\frac{1}{N}\sum_{i=1}^{N}E[(\bar{X}-\mu)^2]=\frac{1}{N}\sum_{i=1}^{N}\frac{Var[X]}{N}=\frac{Var[X]}{N}$$

$$E[\frac{2}{N}(\bar{X}-\mu)\frac{1}{N}\sum_{i=1}(X_i-\mu)]=2E[(\bar{X}-\mu)(\bar{X}-\mu)]=\frac{2\,Var[X]}{N}$$

$$E[S_n[X]]=Var[X]-\frac{Var[X]}{N}=\frac{N-1}{N}Var[X]$$

$$\boxed{Sesgo=-\frac{1}{N}Var[X]}$$

Ejercicio 1

1) Generar una muestra de tamaño N = 10000 correspondiente a la altura de personas adultas, asumiendo que su densidad de probabilidad es una función normal/gaussiana con μ= 1.78m y σ= 0.1 m. Dibuja la densidad de frecuencia y la densidad de probabilidad por separado. Compara μ y σ con la media muestral y la varianza muestral.
2) Considerar la distribución de probabilidad anterior y el estimador media muestral para una muestra de tamaño N (T_N). Generar un número alto M = 10000 de pseudo-muestras y estudiar la distribución (μ-T_N), para N = 10, 100, 1000, 10000, 100000. Calcular el valor esperado en cada caso (considerando el valor esperado como el promedio a los M = 10000 psuedo-experimentos) y dibujar el resultado en función de N. Repetir el mismo procedimiento pero calcular el valor esperado de la varianza (considerando de nuevo las M = 10000 psedo-muestras).
3) Considerar la distribución de probabilidad anterior y la fórmula sesgada de la varianza. Generar un número alto de M = 10000 de muestras y estudiar la distribución (σ-S_N), para N = 10, 100, 1000, 10000, 100000. Calcular el valor esperado en cada caso (usando de nuevo la media muestral) y dibujar el resultado en función de N. ¿Se trata de un estimador consistente?
4) Repetir 2) utilizando la mediana en lugar de la media. ¿Cuál de los dos estimadores es más eficiente?