

Estadística [continuación]

Santander, 2017-2018

Temario y estructura del curso (I)

- Tema 5. Técnicas de remuestreo (bootstrap)
 - [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] Introducción a las técnicas de remuestreo. Conceptos básicos.
 - [T.5.4] Algoritmo de remuestreo BootStrap
 - [P.5.5] Utilización de la técnica BootStrap para mejorar la estimación de estadísticos sencillos.
 - [T.5.6] Algoritmo de remuestreo Jackknife
 - [P.5.7] Utilización de la técnica Jackknife para mejorar la estimación de estadísticos sencillos.
 - [T.5.8] Conceptos generales de cross validation

Temario y estructura del curso (II)

- Tema 6. Modelos de regresión. Estimación de máxima similitud
 - [T.6.1] Concepto de regresión. Función de coste. Algoritmos de minimización.
 - [P.6.2] Práctica: estimación de función de coste y minimización.
 - [T.6.3] Regresión lineal en una dimensión. Solución analítica.
 - [P.6.4] Práctica: implementación de un ajuste lineal y predicción.
 - [T.6.5] Regresión lineal en varias dimensiones. Ajuste de polinomios. Sesgo y varianza.
 - [P.6.6] Práctica: ajuste de un polinomio y estudio del sesgo y la varianza.
 - [T.6.7] Regresión logística.
 - [P.6.8] Práctica: implementación de una regresión logística.
 - [T.6.9] Uso de cross-validation para entender las propiedades de una regresión.
 - [P.6.10] Práctica: Basado en P.6.8 estudiar las propiedades usando cross-validation
 - [T.6.11] Maximum likelihood. G2 statistic and goodness of fit.
 - [P.6.12] Práctica: Goodness of fit para diferentes modelos.
 - [T.5.1] El problema de la estimación a través de una muestra limitada de la distribución.

Temario y estructura del curso (III)

- Tema 7. Regularización. Regresión contraída.
 - [T.7.1] Repaso de problemas de sesgo y varianza.
 - [T.7.2] Concepto de regularización. Ridge Regression.
 - [P.7.3] Práctica: Uso de regularización en un problema de regresión.
 - [T.7.3] Propiedades y selección del parámetro "lambda".
 - [P.7.4] Práctica: Comparación de diferentes parámetros lambda.
- Práctica colectiva

¿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. Técnicas de remuestreo (bootstrap)

“ An infinite number of mathematicians walk into a bar. The first one orders a beer. The second orders half a beer. The third, a quarter of a beer. The bartender says "You're all idiots", and pours two beers.

— Randgruppenhumor cont'd ”

Repaso de estadística: **elementos abstractos**

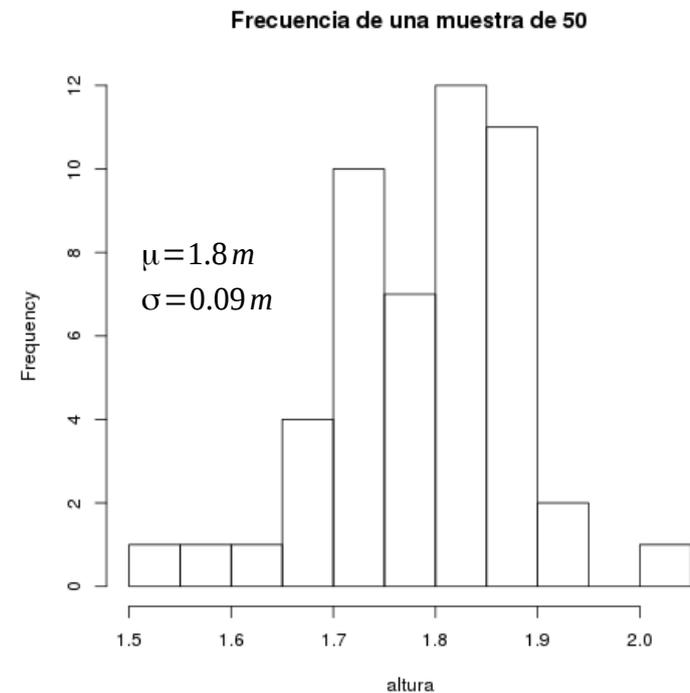
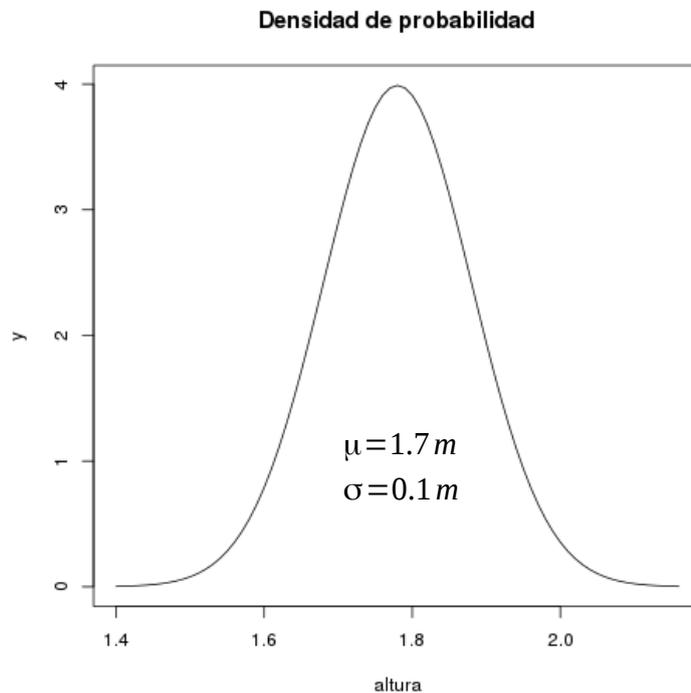
- **Población:** Conjunto finito o infinito que incluye la totalidad de los elementos de estudio. **Ejemplo:** las personas con edades mayores de 20 años.
- **Variable aleatoria:** Una función que asigna un valor numérico al resultado de un experimento aleatorio sobre una población. **Ejemplo:** La altura 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 reales

- **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 $\sim 1.81m$.

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.**



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

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

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

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

- 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

$$\text{Var}[T_n] > \text{Var}[S_n] \rightarrow S_n \text{ mas eficiente}$$

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

$$\lim_{n \rightarrow \infty} E[T_n] = \theta$$

$$\lim_{n \rightarrow \infty} \text{Var}[T_n] = 0$$

- 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

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

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

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

→ Puede verse también que la media es consistente ya que su valor esperado siempre es el parámetro y...

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

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...

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

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

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

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

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

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

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 $\mu = 1.78\text{m}$ y $\sigma = 0.1\text{ m}$. 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 la media muestral para una muestra de tamaño N (T_N). Generar un número alto $M = 10000$ de muestras y estudiar la distribución $(\mu - T_N)$, 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 . Repetir el mismo procedimiento usando la varianza.
- 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 $(\sigma - S_N)$, 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?