Estadística I

Guillermo Julián Moreno

13/14 C1

Índice

1. Estadística descriptiva de datos univariantes

La estadística descriptiva es el conjunto de técnicas para resumir la información proporcionada por una gran masa de datos. El primer objetivo natural es resumir la información que proporcionan esos datos.

1.1. Estadísticos de tendencia central

Definición 1.1 **Media**.

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Es la medida de tendencia central más utilizada. Es bastante sensible a los valores atípicos (*outliers*), observaciones anormalmente grandes que aparecen en el conjunto de datos por errores de transcripción o medición.

Definición 1.2 **Mediana**. Es el valor que divide a los datos en dos mitades, de tal forma que la mitad son menores y la otra mitad mayores que la mediana.

La mediana se calcula de la siguiente forma: dado un conjunto de datos $\{x_1, \ldots, x_n\}$, la mediana es $x_{\frac{n+1}{2}}$ si n es impar y el promedio entre $x_{\frac{n}{2}}$ y $x_{\frac{n}{2}+1}$.

1.2. Estadísticos de dispersión

Definición 1.3 **Varianza**.

$$\sigma^2 = \frac{1}{n}\sum_{i=1} n\left(x_i - \overline{x}\right)^2 = \frac{1}{n}\sum_{i=1} n x_i^2 - \overline{x}^2$$

Definición 1.4 **Desviación típica**.

$$\sigma = \sqrt{\sigma^2}$$

La desviación típica es la raíz de la varianza.

Definición 1.5 **Cuantil**. Para $p \in (0,1)$ se llama cuantil p o q_p al valor que deja el $100p\,\%$ de los datos a la izquierda.

Definición 1.6 **Cuartil**. Los cuartiles son los tres datos que dejan a la izquierda el 25, 50 y 75 por ciento de los datos respectivamente. Es decir:

- $Q_1 = q_{0,25}$
- $Q_2 = q_{0,5}$. El cuartil dos es la mediana.

- $Q_3 = q_{0,75}$

Hay varios métodos para el cálculo de cuantiles. Para hacerlo a mano, podemos usar el siguiente método.

Si el dato en la posición $p(n+1)$ no es un número entero, entones se interpola entre las observaciones ordenadas que están en la posición $\lfloor p(n+1) \rfloor$ y $\lfloor p(n+1) \rfloor + 1$ de la siguiente forma: sea j la parte entera de $p(n+1)$ y m la parte decimal. Entonces,

$$q_p = (1-m)x_j + mx_{j+1}$$

Definición 1.7 **Coeficiente de asimetría**. El tercer momento con respecto a la media se define como

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3$$

que, en su versión adimensional dividimos por σ^3.

Al ser una función cúbica, los valores que se alejen mucho de la media tendrán un valor muy alto en valor absoluto (positivo o negativo según se aleje por la derecha o izquierda, respectivamente). Si la distribución de datos es muy asimétrica, los valores más altos no se cancelan con los valores altos del otro lado (porque no hay) y saldrá un valor más alejado de cero.[1]

1.3. Representación gráfica de datos

Definición 1.8 **Box-plot**. El diagrama de caja o *box-plot* (imagen 1) nos permite visualizar las medidas de dispersión respecto a la mediana. Hay que añadir una nueva medida, el **rango intercuartílico**, la diferencia entre el primer y el tercer cuartil:

$$RI = Q_3 - Q_1$$

Definición 1.9 **Histograma**. El histograma se trata de una aproximación discreta a la función de densidad continua $f(t)$ de la variable que estamos midiendo. Es un diagrama de frecuencias que *mantiene la forma* de esa función de densidad.

Definimos una serie, las marcas de intervalos $a_1^n, \ldots, a_n^n$, donde n es el número de intervalos y la longitud de cada intervalo es $h_n = a_{j+1}^n - a_j^n$. Sea el conjunto $\{x_i\}_{i=0,\ldots,m}$ los datos de nuestra muestra. Entonces, el estimador, la función $\hat{f}_n$, se define de la siguiente forma:

$$\hat{f}^n(t) = \frac{\#\left(i \nearrow x_i \in \left(a_j^n, a_{j+1}^n\right]\right)}{nh_n} = \frac{\sum_{i=1}^{m} \mathbb{1}_{(a_j^n, a_{j+1}^n]}(x_i)}{nh_n}$$

Recordemos que

$$\mathbb{1}_A(n) = \begin{cases} 1 & n \in A \\ 0 & n \notin A \end{cases}$$

[1] Está explicado como el p. culo, ya.

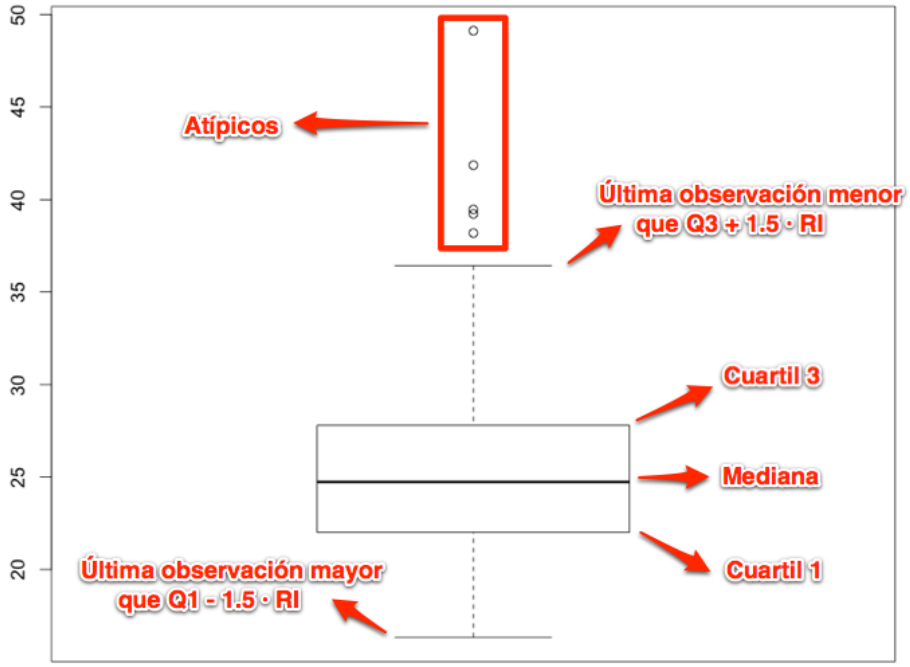

Figura 1: Diagrama de caja

A grandes rasgos, lo que hace en una función es definir un número de intervalos fijos de ancho h_n. Al evaluar $\hat{f}^n(t)$ buscamos en qué intervalo cae t y contamos cuántas de nuestras mediciones caen también en ese intervalo.

1.3.1. Estimadores núcleo o kernel

Definición 1.10 **Método de ventana móvil**. El método de ventana móvil nos da una estimación de la función de densidad en un punto t midiendo los x_i que están en el intervalo de radio h_n centrado en t. Matemáticamente:

$$\hat{f}_n(t) = \frac{1}{n2h_n} \sum_{i=1}^{n} \mathbb{1}_{[t-h_n, t+h_n]}(x_i) = \frac{1}{n2h_n} \sum_{i=1}^{n} \mathbb{1}_{[-1,1]} \left(\frac{t - x_i}{h_n} \right)$$

Podemos reemplazar la función $\frac{1}{2}\mathbb{1}_{[-1,1]}$ por otra, llamada la función de densidad K, kernel o núcleo:

Definición 1.11 **Estimador núcleo**. Dada una función de densidad K simétrica, no necesariamente positiva, definimos el estimador kernel como:

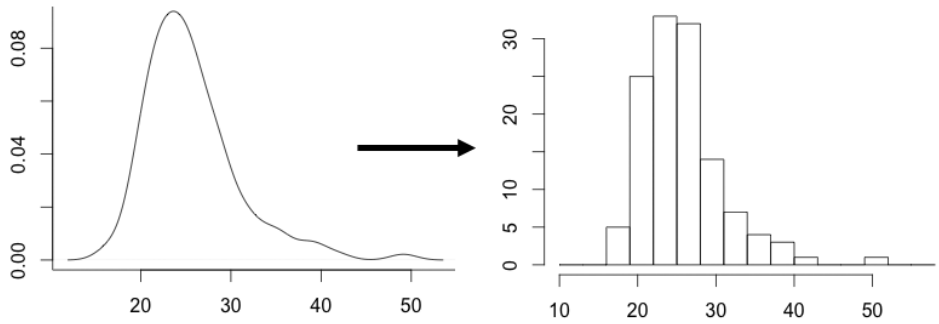

Figura 2: El histograma es una aproximación de la función de densidad real en base a la muestra que hemos obtenido.

$$\hat{f}_n(t) = \frac{1}{n}\sum_{i=1}^{n} K_h(t - x_i) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{t - x_i}{h_n}\right)$$

con $K_h(x) = \frac{1}{h}K(\frac{x}{h})$.

La elección del núcleo K no afecta especialmente a lo bien aproximada que esté la función de densidad. Sin embargo, sí que influye la selección de la ventana h_n (figura 3), también llamada *bandwith* en inglés. Si escogemos una ventana muy pequeña, damos demasiado peso a los datos de nuestra muestra. Si elegimos una ventana muy grande, nuestra muestra pierde importancia y podemos perder información importante.

La elección del h_n más habitual es el que minimiza la distancia L^2 entre $\hat{f}$ y f, es decir, el parámetro que minimice $\int \left(\hat{f}_h - f\right)^2$. Sin embargo, hay un problema: no sabemos qué es f. Hay trucos que imagino que veremos más tarde.

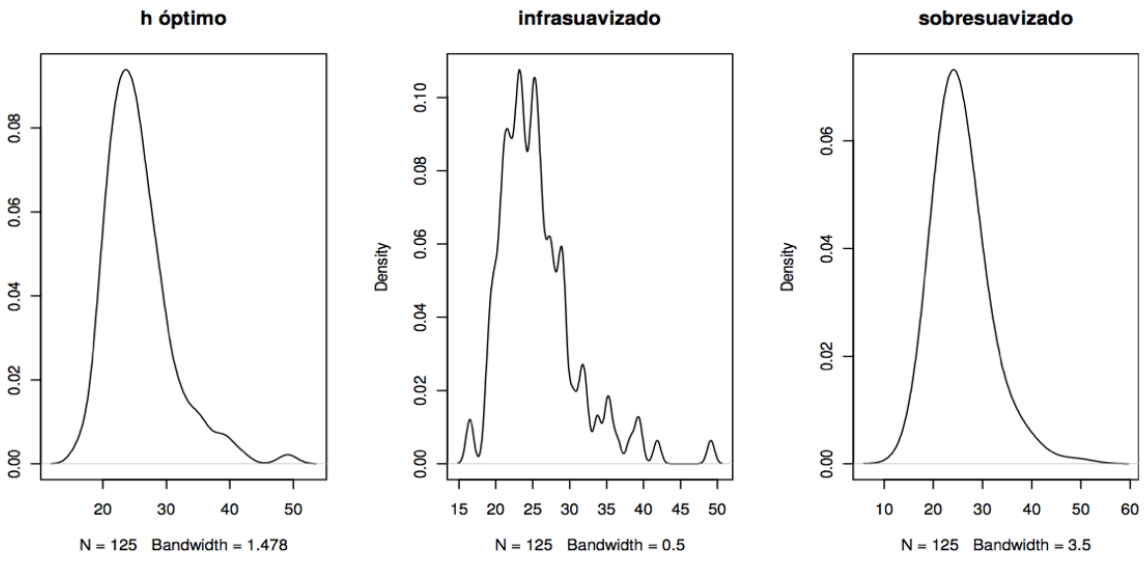

Figura 3: Los efectos que causa elegir una ventana más grande o más pequeña en el estimador

Las funciones kernel más usadas son la uniforme, $\frac{1}{2}\mathbb{1}_{[-1,1]}$, la gaussiana $\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}$ y la de Epanechnikov, que matemáticamente es la que mejor aproxima f.

El estimador kernel $\hat{f}_n(t)$ es la función de densidad de una medida de probabilidad que es la convolución [2] de dos medidas de probabilidad: una, $K_h(x)$ (el kernel reescalado) y otra que da probabilidad $\frac{1}{n}$ a cada punto de la muestra $\{x_i\}$ (distribución o medida empírica).

Generación de datos del estimador kernel Supongamos que K es el núcelo gaussiano. Podemos generar datos artificiales de la densidad así:

$$x_i^0 = x_i^* + h_n Z_i,\ i = 1, \ldots, k$$

donde x_i^* es una observación elegida al azar entre los datos originales y Z_i una observación aleatoria con probabilidad $N(0,1)$. Es decir, lo que hacemos es añadir un dato aleatorio de la muestra y sumamos una pequeña perturbación aleatoria.

2. Estadística descriptiva de datos bivariantes

En esta sección estudiaremos dos variables (X, Y) para explorar la relación entre ambas y tratar de inferir si existe una relación funcional para predecir los valores de una variable en función de los de la otra.

2.1. Representación gráfica

Definición 2.1 **Diagrama de dispersión**. El diagrama de dispersión representa cada variable en función de la otra para que podamos ver la posible relación entre ambas. Ver figura 4.

2.2. Regresión

Definición 2.2 **Recta de regresión**.

La recta de regresión de y sobre x es la recta de forma $\hat{y} = \hat{a} + \hat{b}x$ que más se aproxima a los datos, minimizando los cuadrados de la distancia:

$$(\hat{a}, \hat{b}) = \arg\min_{a,b} \sum_{i=1}^{n} (y_i - a - bx_i))^2$$

La recta de regresión se calcula obteniendo primero $\hat{b}$:

$$\hat{b} = \frac{\sigma_{x,y}}{\sigma_x^2}$$

donde

$$\sigma_{x,y} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i\right) - \overline{xy}$$

[2] Ya aprenderemos en al algún momento de nuestra vida qué narices es una convolución

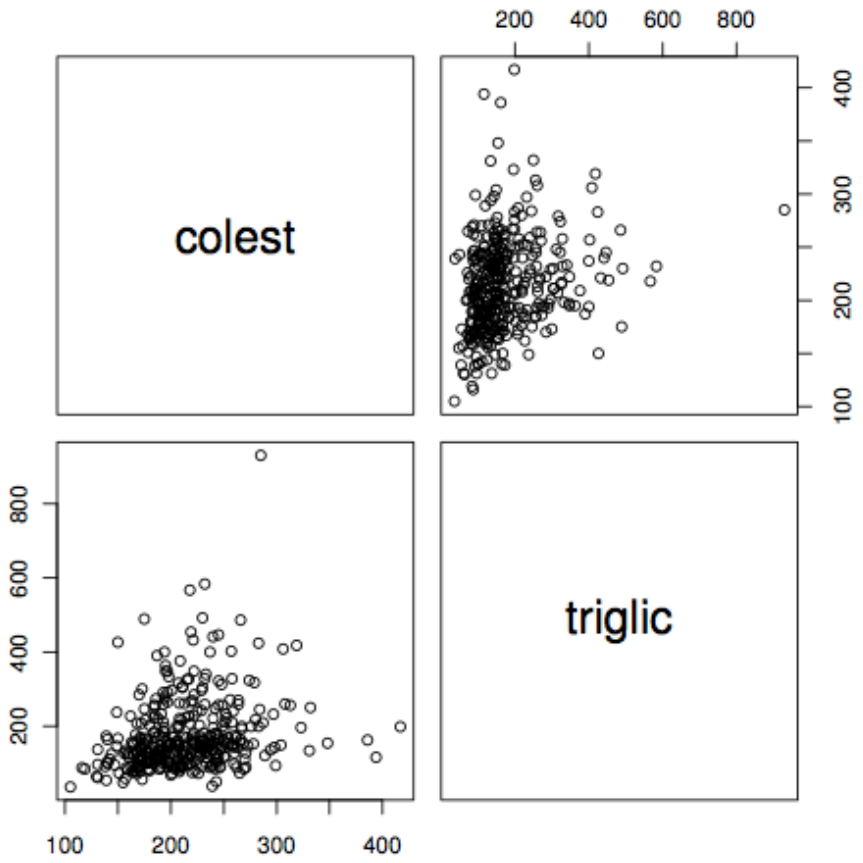

Figura 4: Diagrama de dispersión

y después, sabiendo que la recta pasa por el punto $(\overline{x}, \overline{y})$, obtenemos $\hat{a}$

$$\hat{a} = \overline{y} - \hat{b}\overline{x}$$

El valor b se denomina **coeficiente de regresión lineal** o parámetro de la regresión. Cada valor $e_i = y_i - \hat{y}_i$ se denomina **residuo**. Hay que notar que

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} \left(y_i - \hat{a} - \hat{b}x_i \right) = \sum_{i=1}^{n} \left(y_i - (\overline{y} - \hat{b}\overline{x}) - \hat{b}x_i \right) =$$
$$= \sum_{i=1}^{n} \left(y_i - \hat{b}x_i \right) - n\overline{y} + n\hat{b}\overline{x} = n\overline{y} - n\hat{b}\overline{x} - n\overline{y} + n\hat{b}\overline{x} = 0$$

Esta ecuación ($\sum_{i=1}^{n} e_i = 0$) junto con

$$\sum_{i=1}^{n} x_i e_1 = 0$$

son las dos restricciones entre los residuos que nos dan la recta.

Definición 2.3 **Varianza residual**. La varianza residual s_R^2 o $\hat{\sigma}_e^2$ mide, aproximadamente el *error cuadrático* cometido en la aproximación dada por la recta de regresión:

$$s_R^2 = \hat{\sigma}_e^2 = \frac{1}{n} \sum_{i=1}^{n} e_i^2$$

Definición 2.4 **Coeficiente de correlación lineal**. El coeficiente de correlación lineal o coeficiente de Pearson

$$r = \frac{\hat{\sigma}_{x,y}}{\hat{\sigma}_x \hat{\sigma}_y}$$

que cumple las siguientes condiciones:

$$0 \leq r^2 \leq 1$$
$$\hat{\sigma}_e^2 = \hat{\sigma}_y^2 (1 - r^2)$$
$$r = \hat{b} \frac{\hat{\sigma}_x}{\hat{\sigma}_y}$$

nos indica el grado de ajuste lineal entre las dos variables. Un valor absoluto más cercano a 1 indica una correlación más fuerte. Un valor absoluto cercano a cero indica una correlación débil. El signo, positivo o negativo, indica si la correlación es creciente o decreciente.

3. Muestreo aleatorio

La muestra aleatoria de una cierta v.a. X se denomina como la **muestra aleatoria** o simplemente **muestra**.

Durante este tema, usaremos conceptos de Probabilidad, que repasaré aquí brevemente porque no me apetece escribir demasiado.

3.1. Conceptos de probabilidad

Definición 3.1 **Distribución de una v.a.**.

$$P_X(B) = P(X \in B)$$

Definición 3.2 **Función de distribución**.

$$F(t) = P(X \leq t)$$

Definición 3.3 **Media!de una distribución**. También llamada esperanza de X:

$$E(X) = \int_{-\infty}^{\infty} F(t)\, dt$$

Teorema 3.4 (Teorema de cambio de espacio de integración). *Sea g una función real medible tal que $E(g(X))$ es finita, entonces*

$$E(g(X)) = \int_{\mathbb{R}} g(x)\, dF(x) = \int_{\mathbb{R}} g(x)\, dP(x)$$

.

En particular

$$\mu = \int_{\mathbb{R}} x\, dF(x)$$

y

$$\sigma^2 = \int_{\mathbb{R}} (x-\mu)^2\ dF(x)$$

Definición 3.5 **Momentos**. El momento μ_k es la esperanza de X elevado a una potencia de orden k. Es el valor esperado de la distancia de orden k con respecto a la media

$$\mu_k = E\left((X-\mu)^k\right)$$

3.1.1. Distribuciones aleatorias

Definición 3.6 **Distribución normal**.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, x \in \mathbb{R};\ \mu \in \mathbb{R};\ \sigma > 0$$

Definición 3.7 **Distribución exponencial**.

$$f(x) = \theta e^{-\theta x} \mathbb{1}_{[0,\infty)}(x) \quad \theta > 0$$

con $E(X) = \frac{1}{\theta}$ y $V(X) = \frac{1}{\theta^2}$

La propiedad más interesante es la falta de memoria:

$$P(X > x + a \,|\, X > x) = e^{-\theta a}$$

, es decir, no depende de X. Suponiendo que esta distribución representa, por ejemplo

Definición 3.8 **Distribución gamma**.

$$f(x) = \frac{a^p}{\Gamma(p)} e^{-ax} x^{p-1} \mathbb{1}_{[0,\infty)}(x)\ a > 0;\ p > 0$$

donde $\Gamma(p) = \int_0^\infty x^{p-1} e^{-x}\, dx$, con $E(X) = \frac{p}{a}$ y $V(X) = \frac{p}{a^2}$

Y sudo de copiar más mierda. `http://www.uam.es/personal_pdi/ciencias/abaillo/MatEstI/Algunas-distribuciones-notables.pdf`

A. Ejercicios

A.1. Tema 1 - Estadística descriptiva

Ejercicio 2: Demostrar que

$$\sum_{i=1}^{n} (x_i - \overline{x})^2 = \min_{a \in \mathbb{R}} \sum_{i=1}^{n} (x_i - a)^2$$

Definimos una función

$$g(a) = \sum_{i=1}^{n}(x_i - a)^2$$

, buscamos su derivada

$$g'(a) = -2\sum_{i=1}^{n}(x_i - a)$$

e igualamos a cero:

$$-2\sum_{i=1}^{n}(x_i - a) = 0$$
$$\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} a = 0$$
$$n\overline{x} = na$$
$$\overline{x} = a$$

Esto quiere decir que la media muestral es el valor que minimiza la distancia con cada uno de los datos de la muestra.

Ejercicio 5: Determina si es verdadero o falso:
a) Si añadimos 7 a todos los datos de un conjunto, el primer cuartil aumenta en 7 unidades y el rango intercuartílico no cambia.
b) Si todos los datos de un conjunto se multiplican por -2, la desviación típica se dobla.

Apartado a) *Añadir siete a todos los datos es una traslación, así que la distribución de los datos no cambia.*

Apartado b) *Teniendo en cuenta que si multiplicamos todos los datos del conjunto por -2 la media también se multiplica por -2, y sustituyendo en la fórmula de la varianza:*

$$\sigma' = \sqrt{\frac{1}{n}\sum_{i=1} n(-2x_i)^2 - (-2\overline{x})^2} = \sqrt{\frac{1}{n}\sum_{i=1} 4\left(nx_i^2 - \overline{x}^2\right)} = \sqrt{4\sigma^2} = 2\sigma$$

Por lo tanto, la desviación típica sí se dobla.

Apartado c) *Usando los cálculos del apartado anterior vemos que la varianza se multiplica por cuatro.*

Apartado d) *Efectivamente: cambiar el signo haría una reflexión de los datos sobre el eje Y y la asimetría estaría orientada hacia el lado contrario.*

Índice alfabético