Clase práctica 15

Nahuel Arca

03 de octubre de 2019

1. Contenidos

- Estimar probabilidades, esperanza y varianza.
- Implementar y graficar la función de distribución empírica.
- `mean(x,trim=0)`, `median`, desvío estándar, `quantile`.
- Hacer histogramas (superponiendo curvas), boxplots y qqplots.
- Relación entre asimetría, media y mediana.

2. Notas

Consideremos una *población* $\mathcal{S}$ y para cada $A \subset \mathcal{S}$ sea $P(A)$ la proporción de *individuos* en $\mathcal{S}$ que también están en A. Es razonable pensar que esto modela correctamente el experimento aleatorio "se extrae al azar un individuo de la población".

Se tiene una *variable* $X : \mathcal{S} \to \mathbb{R}$ y se desea estudiar su distribución. Para ello se extraen con reposición n individuos. Este experimento aleatorio tendrá entonces espacio muestral $\mathcal{S}^n$ y su probabilidad estará dada por la ley del producto:

$$P(A_1 \times \cdots \times A_n) = P(A_1) \cdots P(A_n)$$

El i-ésimo individuo seleccionado en el experimento aleatorio es $\omega_i = \pi_i(\omega)$. El *valor de la variable* (*dato*, *observación* o *medición*) correspondiente a tal individuo es entonces $X(\pi_i(\omega)) = X_i(\omega)(=: x_i)$, siendo $X_i := X \circ \pi_i$. Las variables aleatorias $X_1, \ldots, X_n$ son la *muestra*.

2.1. Medidas de Resumen

Supongamos que tenemos un conjunto de n datos $x_1, \ldots, x_n$. Los datos ordenados serían $x_{(1)} \leq \ldots \leq x_{(n)}$ y

$$x_{(t)} = (t - [t])(x_{([t]+1)} - x_{([t])}) + x_{([t])}$$

para todo $t \in [1, n)$.

Dado $k \in \mathbb{N}_0$ tal que $k < n/2$, se pueden considerar los datos eliminando $x_{(1)}, \ldots, x_{(k)}$ y $x_{(n)}, \ldots, x_{(n-k+1)}$. La media calculada con estos datos es la *media α-podada* de los datos originales con $\alpha = k/(n+1)$, y se denota $\overline{x}_\alpha$. En general

$$\overline{x}_\alpha = (\alpha(n+1) - [\alpha(n+1)])(n+1)\left(\overline{x}_{\frac{[\alpha(n+1)]+1}{n+1}} - \overline{x}_{\frac{[\alpha(n+1)]}{n+1}}\right) + \overline{x}_{\frac{[\alpha(n+1)]}{n+1}}$$

para todo $\alpha \in [1/(n+1), (\lceil n/2 \rceil - 1)/(n+1))$.

Tipo	Medida	Poblacional (continua)	Muestral
De posición o centrado	Media	$E(X)$	$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (promedio)
	Mediana	$x_{1/2}$	$\tilde{x} = x_{\left(\frac{n+1}{2}\right)}$
	$100p$-percentil	x_p	$x_{(p(n+1))}$
	Media α-podada	$E(X \mid x_\alpha < X < x_{1-\alpha})$	$\overline{x}_\alpha$
De dispersión o variabilidad	Varianza	$E((X-\mu_X)^2)$	$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2$
	Desvío estándar	σ_X	S_x
	Distancia Intercuartil	$x_{3/4} - x_{1/4}$	$x_{(3(n+1)/4)} - x_{(n+1)/4}$

Más allá de estas medidas, las distribuciones acampanadas se comparan con la normal mediante otras dos medidas:
Asimetría:

$$E\left(\left(\frac{X-\mu_X}{\sigma_X}\right)^3\right)$$

Si la asimetría es positiva, se dice que hay *asimetría a derecha* (la cola de la derecha es más pesada que la de la izquierda). Vale lo mismo intercambiando "izquierda" y "derecha", y reemplazando "positiva" por "negativa". Otra medida de la asimetría de una curva es $(\mu_X - x_{1/2})/\sigma_X$.

Curtosis:

$$E\left(\left(\frac{X-\mu_X}{\sigma_X}\right)^4\right)$$

Si la curtosis es mayor que 3 (la curtosis de una normal), se dice que las colas son *pesadas*. Si es menor que 3, se dice que son *livianas*.

2.2. Histograma

1. Se divide el rango de los datos en *intervalos* o *clases*, que no se superpongan. Las clases deben ser excluyentes y exhaustivas.

2. Se cuenta la cantidad de datos en cada intervalo o clase, es decir la frecuencia.

3. Se grafica el histograma en un par de ejes coordenados representando en las abscisas los intervalos y sobre cada uno de ellos un rectángulo cuya área sea proporcional a la frecuencia de dicho intervalo.

- No es necesario que todos los intervalos tengan la misma longitud, pero es recomendable que así sea. Esto facilita su interpretación.

- Es recomendable tomar

$$\text{altura del rectángulo} = \frac{\text{frecuencia relativa}}{\text{longitud del intervalo}}$$

  De esta manera el área es 1 y dos histogramas son fácilmente comparables independientemente de la cantidad de observaciones en las que se basa cada uno.

2.3. Box-Plots

1. Representamos una escala vertical.
2. Dibujamos una caja cuyos extremos son los cuartiles y dentro de ella un segmento que corresponde a la mediana.
3. A partir de cada extremo dibujamos un segmento hasta el dato más alejado que está a lo sumo $1{,}5d_I$ del extremo de la caja.
4. Marcamos con $\circ$ a aquellos datos que están a más de $1{,}5d_I$ de cada extremo. Estos datos se llaman outliers.

2.4. QQ-plot

Dada una distribución F continua y creciente, un qq-plot es un gráfico que consiste de los puntos

$$\left(F^{-1}\left(\frac{i}{n+1}\right), x_{(i)}\right)$$

Notar que se trata de pares compuestos por los percentiles $100i/(n+1)\,\%$ de la distribución y de los datos. De esta manera, se espera que si los datos siguen la distribución F, los puntos tiendan a estar en la recta $x = y$.
El caso más habitual es tomar $F = \Phi$. En este caso, si los datos siguen una distribución $N(\mu, \sigma^2)$, tenderán a estar en la recta $y = \sigma x + \mu$.

3. Ejercicios

1. Se pretende estudiar la distribución de Y_n, el 90-percentil de n v.a. independientes $X_i \sim U(0,1)$ ($n = 19, 49, 299, 4999$). Para esto se toma en cada caso una muestra de tamaño 3000. Para cada n:

   a) Estimar $P(0{,}6 < Y_n < 0{,}8)$.

   b) Implementar y graficar la función de distribución empírica. ¿Cómo se estima la probabilidad de que Y_n se encuentre en un determinado intervalo?

   c) Estimar la esperanza y la varianza. Calcular las siguientes medidas muestrales: mediana, media 0,1-podada, desvío estándar y distancia intercuartil. ¿Qué se puede decir de la asimetría?

   d) Hacer histogramas, boxplots y qqplots. Interpretar.

   Finalmente, hacer un boxplot paralelo e interpretarlo.

2. Hacer lo mismo para el 75-percentil y para el máximo.