MAT-042: Taller 1

Felipe Osorio

http://fosorios.mat.utfsm.cl

Departamento de Matemática, UTFSM

Ejercicio 1.a)

Interpretación:

Las plantas sometidas al tratamiento Cross tienen una altura promedio de 161.5 unid., su distribución presenta asimetría negativa (hacia la izquierda) con curtosis positiva indicando que la distribución es aguda (leptocúrtica). Se puede apreciar además que existe una única observación atípica. Comparativamente con las plantas sometidas al tratamiento Self la distribución tiene un nivel promedio y variabilidad mayores.

EX UMBRA IN SOLEM

Ejercicio 1.b)

Tenemos

$$\overline{z} = \frac{1}{n}\sum_{i=1}^{n} z_i = \frac{1}{n}\sum_{i=1}^{n} x_i - \frac{1}{n}\sum_{i=1}^{n} y_i = \overline{x} - \overline{y},$$

mientras que

$$\begin{aligned} s_z^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \overline{z})^2 = \frac{1}{n-1}\sum_{i=1}^{n}[(x_i - y_i) - (\overline{x} - \overline{y})]^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}[(x_i - \overline{x}) - (y_i - \overline{y})]^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}[(x_i - \overline{x})^2 - 2(x_i - \overline{x})(y_i - \overline{y}) + (y_i - \overline{y})^2] \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 + \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2 = s_x^2 + s_y^2, \end{aligned}$$

pues $\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = 0$.

Ejercicio 1.b)

Para los datos del Ejercicio, obtenemos

$$\overline{z} = 161.5 - 140.6 = 20.9$$

$$s_z^2 = 28.94^2 + 16.41^2 = 1106.8117$$

$$\mathrm{CV}_z = \frac{\sqrt{1106.8117}}{20.9} = 1.5918$$

Ejercicio 1.c)

Sabemos que

$$\begin{aligned}\overline{u} &= -1.76\,\overline{y} + 408.96 \\ &= -1.76 \cdot 140.6 + 408.96 = 161.5040\end{aligned}$$

y

$$s_u = |-1.76|\,s_y = 1.76 \cdot 16.41 = 28.8816$$

Ejercicio 2

a) Considere $x_i = u_i + 5$, $i = 1, \ldots, 40$. De este modo,

$$\overline{x} = \overline{u} + 5 = 45.4 + 5 = 50.4, \qquad s_x = s_u = 12.8.$$

Mientras que

$$\text{CV}_x = \frac{s_x}{\overline{x}} = \frac{12.8}{50.4} = 0.2540.$$

b) En este caso, $y_i = 1.1\, v_i$, $i = 1, \ldots, 13$. Así,

$$\overline{y} = 1.1\, \overline{v} = 1.1 \cdot 41.8 = 45.98, \qquad s_y = 1.1\, s_v = 1.1 \cdot 17.8 = 19.58.$$

Además

$$\text{CV}_y = \frac{s_y}{\overline{y}} = \frac{17.8}{41.8} = 0.4258.$$

c) La mayor variabilidad se obtuvo en el Paralelo P101.

Ejercicio 3

Tenemos el conjunto de datos:

$$\boldsymbol{x} = \{\underbrace{10\,000, 10\,000, \cdots, 10\,000}_{500 \text{ observaciones}}, 10\,001, \underbrace{10\,002, 10\,002, \cdots, 10\,002}_{500 \text{ observaciones}}\}.$$

De este modo, es evidente que

$$\operatorname{me}(\boldsymbol{x}) = 10\,001,$$

mientras que el promedio muestral es dado por:

$$\begin{aligned}
\overline{x} &= \frac{500 \cdot 10\,000 + 10\,001 + 500 \cdot 10\,002}{1001} = \frac{500 \cdot 10\,000 + 10\,000 + 1 + 500(10\,000 + 2)}{1001} \\
&= \frac{500 \cdot 10\,000 + 10\,000 + 500 \cdot 10\,000 + 1 + 500 \cdot 2}{1001} = \frac{10\,000 \cdot 1001 + 1001}{1001} \\
&= \frac{1001(10\,000 + 1)}{1001} = 10\,001.
\end{aligned}$$

Ejercicio 3

Sea $u_i = x_i - \overline{x}$, para $i = 1, \ldots, 1001$. Es decir, tenemos:

$$\boldsymbol{u} = \{\underbrace{-1, -1, \cdots, -1}_{500 \text{ obs}}, 0, \underbrace{1, 1, \cdots, 1}_{500 \text{ obs}}\}.$$

Podemos calcular la varianza muestral como:

$$s^2 = \frac{1}{1001 - 1} \sum_{i=1}^{1001} u_i^2 = \frac{1}{1000} \left(500(-1)^2 + 0 + 500(1)^2\right) = \frac{1000}{1000} = 1.$$

Como $s = 1$, sigue que

$$z_i = \frac{x_i - \overline{x}}{s} = u_i, \qquad i = 1, \ldots, 1001.$$

Ejercicio 3

Lo anterior permite calcular

$$b_1 = \frac{1}{1001}\sum_{i=1}^{1001}\left(\frac{x_i - \overline{x}}{s}\right)^3 = \frac{1}{1001}\sum_{i=1}^{1001} z_i^3 = \frac{1}{1001}\left(500(-1)^3 + 0 + 500(1)^3\right) = 0.$$

Finalmente,

$$\begin{aligned} b_2 &= \left\{\frac{1}{1001}\sum_{i=1}^{1001}\left(\frac{x_i - \overline{x}}{s}\right)^4\right\} - 3 = \frac{1}{1001}\sum_{i=1}^{1001} z_i^4 - 3 \\ &= \frac{1}{1001}\left(500(-1)^4 + 0 + 500(1)^4\right) - 3 = \frac{1000}{1001} - 3 \\ &= -2.001 \end{aligned}$$

Ejercicio 4

Es fácil notar que

$$(x_i - x_j)^2 = ((x_i - \overline{x}) - (x_j - \overline{x}))^2 = (x_i - \overline{x})^2 + (x_j - \overline{x})^2 - 2(x_i - \overline{x})(x_j - \overline{x}).$$

De este modo,

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^2 &= \sum_{i=1}^{n}\sum_{j=1}^{n}\{(x_i - \overline{x})^2 + (x_j - \overline{x})^2 - 2(x_i - \overline{x})(x_j - \overline{x})\} \\ &= n\sum_{i=1}^{n}(x_i - \overline{x})^2 + n\sum_{j=1}^{n}(x_j - \overline{x})^2 - 2\sum_{i=1}^{n}(x_i - \overline{x})\sum_{j=1}^{n}(x_j - \overline{x}).\end{aligned}$$

Tenemos $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$ (y análogamente para $\sum_{j=1}^{n}(x_j - \overline{x}) = 0$), luego

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^2 = 2n\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

Lo que lleva a

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{2n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^2.$$

que es el resultado deseado.

Ejercicio 5

En efecto,

$$\begin{aligned}\overline{x}_n &= \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\Big(\sum_{i=1}^{n-1} x_i + x_n\Big) \\ &= \frac{1}{n}\Big((n-1)\overline{x}_{n-1} + x_n\Big) = \frac{1}{n}\big(n\overline{x}_{n-1} - \overline{x}_{n-1} + x_n\big) \\ &= \overline{x}_{n-1} + \frac{1}{n}(x_n - \overline{x}_{n-1})\end{aligned}$$

Ejercicio 5

Por otro lado,

$$\overline{x}_n = \frac{1}{W_n}\sum_{i=1}^{n}\omega_i x_i = \frac{1}{W_n}\Big(\sum_{i=1}^{n-1}\omega_i x_i + \omega_n x_n\Big)$$

Ahora,

$$W_n = \sum_{i=1}^{n}\omega_i = \sum_{i=1}^{n-1}\omega_i + \omega_n = W_{n-1} + \omega_n,$$

es decir $W_{n-1} = W_n - \omega_n$. Así,

$$\begin{aligned}\overline{x}_n &= \frac{1}{W_n}\Big(W_{n-1}\overline{x}_{n-1} + \omega_n x_n\Big) = \frac{1}{W_n}(W_n\overline{x}_{n-1} - \omega_n\overline{x}_{n-1} + \omega_n x_n)\\ &= \overline{x}_{n-1} + \frac{\omega_n}{W_n}(x_n - \overline{x}_{n-1})\end{aligned}$$

Ejercicio 6.b)

La tabla de frecuencias para el total de trabajadores adopta la forma:

Ingreso (UM)	C_i	n_i	f_i	N_i	F_i
65 – 75	70	10	0.048	10	0.048
75 – 85	80	15	0.071	25	0.119
85 – 95	90	65	0.310	90	0.429
95 – 105	100	25	0.119	115	0.548
105 – 115	110	60	0.285	175	0.833
115 – 125	120	25	0.119	200	0.952
125 – 135	130	10	0.048	210	1.000
Total	–	210	1.000	–	–

En este caso, tenemos $n = 210$ y $\sum_{i=1}^{7} n_i C_i = 21\,150$. De este modo, la media aritmética es:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{5} n_i C_i = \frac{21\,150}{210} = 100.714 \text{ (UM)}.$$

Ejercicio 6.b)

Para calcular la mediana, primero debemos ubicar el intervalo mediano. En efecto, debemos ubicar la primera frecuencia relativa acumulada (F_i) que supere 0.5.

De este modo, el intervalo mediano es $(95, 105]$, tenemos que $a_i = 10$ para todo los intervalos. Luego,

$$\mathrm{me} = L_i + \frac{1/2 - F_{i-1}}{f_i}\, a_i,$$

donde $L_i = 95$, $F_{i-1} = 0.429$, $f_i = 0.119$ y $a_i = 10$, luego

$$\mathrm{me} = 95 + \frac{0.500 - 0.429}{0.119} \cdot 10 = 95 + 0.5966 \cdot 10 = 100.966 \text{ (UM)}.$$

Ejercicio 6.b)

Por otro lado,

$$s^2 = \frac{1}{n-1}\Big(\sum_{i=1}^{7} n_i C_i^2 - n\overline{x}^2\Big).$$

En nuestro caso,

$$\sum_{i=1}^{7} n_i C_i^2 = 2\,176\,500, \qquad \overline{x}^2 = 10\,143.367.$$

Así,

$$\begin{aligned} s^2 &= \frac{1}{210-1}(2\,176\,500 - 210 \cdot 10\,143.367) = \frac{1}{209}(2\,176\,500 - 2\,130\,107.143) \\ &= \frac{46\,392.857}{209} = 221.975\ \text{(UM)}^2. \end{aligned}$$

Además, $s = \sqrt{221.975}$ UM $= 14.899$ UM y $\mathrm{CV} = 14.899/100.714 = 0.148$

Ejercicio 6.b)

Podemos evaluar la simetría usando el coeficiente de Galton. Tenemos que Q_1 y Q_3, son dados por

$$
\begin{aligned}
Q_1 &= \ 85 + \frac{0.250 - 0.119}{0.310} \cdot 10 = \ 89.226,\\
Q_3 &= 105 + \frac{0.750 - 0.548}{0.285} \cdot 10 = 112.088,
\end{aligned}
$$

de este modo $IQR = 22.862$, y

$$
\begin{aligned}
b_{\mathsf{G}} &= \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1}\\
&= \frac{(112.088 - 100.966) - (100.966 - 89.226)}{22.862} = \frac{11.122 - 11.740}{22.862} = -0.027
\end{aligned}
$$

Es decir, la distribución de los datos presenta una ligera asimetría.

Ejercicio 7.a)

```
# chequear su espacio de trabajo, usando:
> getwd()

# archivo 'levis.csv' en mi espacio de trabajo
> levis <- read.csv("levis.csv")
> levis
   Fabrica.A Fabrica.B
1       0.12      1.64
2       1.01     -0.60
3      -0.20     -1.16
4       0.15     -0.13
5      -0.30      0.40
6      -0.07      1.70
7       0.32      0.38

...

20      0.30      0.85
21      0.24      0.60
22      0.13      0.29
```

Ejercicio 7.a)

Sea

$$X : \text{mediciones de desgaste por defectos en la fábrica } A$$
$$Y : \text{mediciones de desgaste por defectos en la fábrica } B$$

Tenemos

$$\sum_{i=1}^{22} x_i = 9.95, \qquad \sum_{i=1}^{22} y_i = 19.43,$$

de ahí que

$$\overline{x} = \frac{9.95}{22} = 0.4523, \qquad \overline{y} = \frac{19.43}{22} = 0.8832.$$

Ejercicio 7.a)

```
# extrae variables desde 'levis'
> x <- levis[,1]
> x
 [1] 0.12  1.01 -0.20  0.15 -0.30 -0.07 0.32 0.27 -0.32 -0.17 0.24
[12] 0.03  0.35 -0.08  1.94  0.28  1.30 4.27 0.14  0.30  0.24 0.13
> y <- levis[,2]
> y
 [1] 1.64 -0.60 -1.16 -0.13  0.40  1.70 0.38 0.43  1.04  0.42 0.85
[12] 0.63  0.90  0.71  0.43  1.97  0.30 0.76 7.02  0.85  0.60 0.29

# cálculo de promedios
> mean(x)
[1] 0.4522727
> mean(y)
[1] 0.8831818

# alternativamente
> nobs <- length(x) # nrow(levis)
> sum(x) / nobs
[1] 0.4522727
> sum(y) / nobs
[1] 0.8831818
```

Ejercicio 7.a)

Además

$$\sum_{i=1}^{22}(x_i - \overline{x})^2 = 21.1348, \qquad \sum_{i=1}^{22}(y_i - \overline{y})^2 = 49.5031,$$

esto permite obtener

$$\operatorname{var}(\boldsymbol{x}) = \frac{21.1348}{22-1} = 1.0064, \qquad \operatorname{var}(\boldsymbol{y}) = \frac{49.5031}{22-1} = 2.3573.$$

Por tanto,

$$\mathrm{CV}_x = \frac{s_x}{\overline{x}} = \frac{\sqrt{1.0064}}{0.4523} = \frac{1.0032}{0.4523} = 2.2181,$$

$$\mathrm{CV}_y = \frac{s_y}{\overline{y}} = \frac{\sqrt{2.3573}}{0.8832} = \frac{1.5353}{0.8832} = 1.7384.$$

Ejercicio 7.a)

```
# suma de cuadrados de los desvios
> sum((x - mean(x))^2)
[1] 21.13479
> sum((y - mean(y))^2)
[1] 49.50308

# cálculo de varianzas
> var(x)
[1] 1.006418
> var(y)
[1] 2.357289

# cálculo de desvios estándar
> sd(x)
[1] 1.003204
> sd(y)
[1] 1.535347
```

Ejercicio 7.a)

Por otro lado, para los datos de la fábrica A:

$$Q_1(\boldsymbol{x}) = -0.0450, \qquad Q_2(\boldsymbol{x}) = 0.1950, \qquad Q_3(\boldsymbol{x}) = 0.3150.$$

De ahí que,

$$\begin{aligned} b_{\mathsf{G}}(\boldsymbol{x}) &= \frac{(Q_3(\boldsymbol{x}) - Q_2(\boldsymbol{x})) - (Q_2(\boldsymbol{x}) - Q_1(\boldsymbol{x}))}{Q_3(\boldsymbol{x}) - Q_1(\boldsymbol{x})} \\ &= \frac{(0.3150 - 0.1950) - (0.1950 + 0.0450)}{0.3150 + 0.0450} = \frac{0.1200 - 0.2400}{0.3600} \\ &= -\frac{0.1200}{0.3600} = -0.3333. \end{aligned}$$

Ejercicio 7.a)

```
# datos 'x' ordenados:
> sort(x)
 [1] -0.32 -0.30 -0.20 -0.17 -0.08 -0.07  0.03  0.12  0.13  0.14  0.15
[12]  0.24  0.24  0.27  0.28  0.30  0.32  0.35  1.01  1.30  1.94  4.27

# cálculo de la mediana
> median(x)
[1] 0.195

# cálculo de quantiles
> quantile(x) # por defecto
    0%    25%    50%    75%   100%
-0.320 -0.045  0.195  0.315  4.270

# algunos 'cuartiles' selectos
> Q <- quantile(x, probs = c(.25, .50, .75))
> Q <- as.vector(Q)

# cálculo del coef. asimetría de Galton
> IQR <- Q[3] - Q[1]
> bG <- ((Q[3] - Q[2]) - (Q[2] - Q[1])) / IQR
> bG
[1] -0.3333333
```

EX UMBRA IN SOLEM

Análogamente, para los datos de la fábrica B,

$$Q_1(\boldsymbol{y}) = 0.3850, \qquad Q_2(\boldsymbol{y}) = 0.6150, \qquad Q_3(\boldsymbol{y}) = 0.8875.$$

Esto lleva a

$$\begin{aligned} b_{\mathsf{G}}(\boldsymbol{y}) &= \frac{(0.8875 - 0.6150) - (0.6150 - 0.3850)}{0.8875 - 0.3850} = \frac{0.2725 - 0.2300}{0.5025} \\ &= \frac{0.0425}{0.5025} = 0.0846. \end{aligned}$$

Ejercicio 7.b)

Los datos transformados, $z_i = x_i - y_i$, para $i = 1, \dots, 22$, son:

-1.52	1.61	0.96	0.28	-0.70	-1.77	-0.06	-0.16	-1.36	-0.59
-0.61	-0.60	-0.55	-0.79	1.51	-1.69	1.00	3.51	-6.88	-0.55
-0.36	-0.16								

Es fácil notar que

$$\begin{aligned}\overline{z} &= \frac{1}{n}\sum_{i=1}^{n} z_i = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i) = \frac{1}{n}\sum_{i=1}^{n} x_i - \frac{1}{n}\sum_{i=1}^{n} y_i = \overline{x} - \overline{y} \\ &= 0.4523 - 0.8832 = -0.4309.\end{aligned}$$

Adicionalmente,

$$\sum_{i=1}^{22}(z_i - \overline{z})^2 = 75.5744, \qquad \operatorname{var}(\boldsymbol{z}) = 3.5988,$$

y $\mathrm{CV}_z = \sqrt{\operatorname{var}(\boldsymbol{z})}/\overline{z} = -1.8970/0.4309 = -4.4024.$

Ejercicio 7.b)

```
# datos transformados
> z <- x - y
> z
 [1] -1.52  1.61  0.96  0.28 -0.70 -1.77 -0.06 -0.16 -1.36 -0.59 -0.61
[12] -0.60 -0.55 -0.79  1.51 -1.69  1.00  3.51 -6.88 -0.55 -0.36 -0.16

# estadísticas de resumen
> mean(z)
[1] -0.4309091
> sum((z - mean(z))^2)
[1] 75.57438
> var(z)
[1] 3.59878
> sd(z)
[1] 1.897045
```

EX UMBRA IN SOLEM

Ejercicio 7.b)

Finalmente, los datos ordenados $z_{(1)}, z_{(2)}, \dots, z_{(22)}$ son:

-6.88	-1.77	-1.69	-1.52	-1.36	-0.79	-0.70	-0.61	-0.60	-0.59
-0.55	-0.55	-0.36	-0.16	-0.16	-0.06	0.28	0.96	1.00	1.51
1.61	3.51								

De ahí que,

$$\mathrm{me}(\boldsymbol{z}) = \frac{1}{2}\big(-0.55 + (-0.55)\big) = -0.55.$$

Ejercicio 7.b)

```
# datos ordenados
> oz <- sort(z)
> oz
 [1] -6.88 -1.77 -1.69 -1.52 -1.36 -0.79 -0.70 -0.61 -0.60 -0.59 -0.55
[12] -0.55 -0.36 -0.16 -0.16 -0.06  0.28  0.96  1.00  1.51  1.61  3.51

# distintas formas de obtener la mediana
> (oz[nobs/2] + oz[nobs/2 + 1]) / 2
[1] -0.55

> mean(oz[c(nobs/2, nobs/2 + 1)])
[1] -0.55

> median(z) # mucho más eficiente
[1] -0.55
```

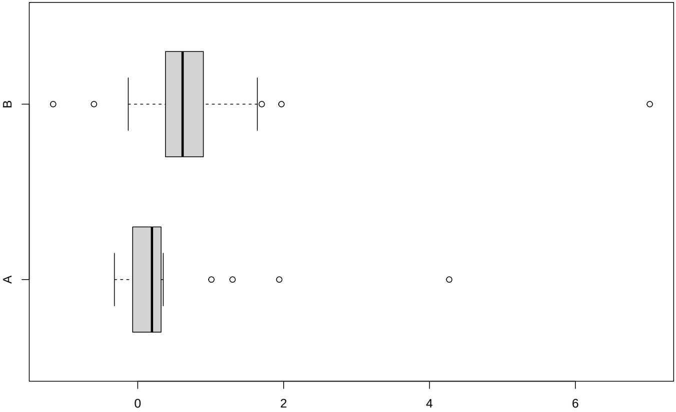
Ejercicio 7.b)

```
# una simpática biblioteca para R
> library(fastmatrix)

# cálculo de momentos de 2do, 3er y 4to orden
# así como coef. de sesgo y curtosis
> moments(z)
$second
[1] 3.435199

$third
[1] -8.714017

$fourth
[1] 91.75744

$skewness
[1] -1.276396

$kurtosis
[1] 4.08485
```

Ejercicio 7.b)

```
# aplicando 'moments' en la base de datos 'levis'
> apply(levis, 2, moments)
$Fabrica.A
$Fabrica.A$second
[1] 0.9606721
$Fabrica.A$third
[1] 2.626377
$Fabrica.A$fourth
[1] 9.962959
$Fabrica.A$skewness
[1] 2.601292
$Fabrica.A$kurtosis
[1] 6.836287

$Fabrica.B
$Fabrica.B$second
[1] 2.25014
$Fabrica.B$third
[1] 9.980491
$Fabrica.B$fourth
[1] 65.65025
$Fabrica.B$skewness
[1] 2.757608
$Fabrica.B$kurtosis
[1] 8.81437
```

Ejercicio 7.b)[1]



[1] Usando la función `boxplot` de R