Notas de Clase:

Introducción a la Estadística con Apoyo Computacional

Felipe Osorio y Ronny Vallejos

Departamento de Matemática, Universidad Técnica Federico Santa María

Índice general

Capítulo 1

Estadística Descriptiva

1.1. Preliminares: Sumas y Productos

En esta sección se introduce notación que tiene por objetivo escribir de forma compacta sumas y productos de secuencias de números $a_1, a_2, \ldots$, donde $a_i \in \mathbb{R}$, para todo i.

DEFINICIÓN 1.1. Considere una secuencia de números $a_1, a_2, \ldots$. Se define la *sumatoria* de esta secuencia, como:

$$\sum_{i=1}^{n} a_i = a_1 + a_2 + \cdots + a_n, \tag{1.1}$$

donde i denota el índice de la sumatoria, mientras que a_i representa un elemento genérico. En este caso, n indica la cantidad de elementos que se están sumando.

Es posible apreciar que la suma en (1.1) puede ser escrita de manera análoga como

$$\sum_{1 \leq i \leq n} a_i = a_1 + a_2 + \cdots + a_n. \tag{1.2}$$

Debemos notar que, si $n = 0$ el valor de la sumatoria se define como cero.

OBSERVACIÓN 1.2. A partir de la Ecuación (1.2) podemos introducir una notación mucho más general. En efecto, sea R un conjunto de índices. Así, basta considerar el conjunto $R = \{1, 2, \ldots, n\}$, para re-escribir la suma en (1.2) como:

$$\sum_{i \in R} a_i = a_1 + \cdots + a_n. \tag{1.3}$$

OBSERVACIÓN 1.3. Aunque frecuentemente la notación dada en la Ecuación (1.3) es utilizada para sumas finitas, esta puede ser adaptada con facilidad para sumas infinitas. Por ejemplo,

$$\sum_{i=1}^{\infty} a_i = \sum_{i \geq 1} a_i = a_1 + a_2 + \cdots .$$

Más formalmente, debemos escribir

$$\sum_{i=1}^{\infty} a_i = \lim_{n \to \infty} \sum_{i=1}^{n} a_i.$$

RESULTADO 1.4 (Regla constante). *Sea a un número real. De este modo,*

$$\sum_{i=1}^{n} a = a + a + \cdots + a = na.$$

En general, para $r < n$ tenemos

$$\sum_{i=r}^{n} a = (n - r + 1)a, \qquad a \in \mathbb{R}.$$

RESULTADO 1.5. *Considere la secuencia $x_1, \ldots, x_n$ y sea a una constante. Entonces,*

$$\sum_{i=1}^{n} a\, x_i = a\, x_1 + \cdots + a\, x_n = a(x_1 + \cdots + x_n) = a \sum_{i=1}^{n} x_i.$$

En general, sean $x_1, \ldots, x_n$ y $y_1, \ldots, y_n$ dos secuencias de números y $a, b \in \mathbb{R}$. Entonces,

$$\sum_{i=1}^{n} (a\, x_i + b\, y_i) = a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} y_i.$$

Note también que las sumatorias pueden ser *descompuestas* en varias sumas. En efecto, para una secuencia de números $a_1, \ldots, a_n$. Tenemos que

$$\sum_{i=1}^{n} a_i = \sum_{i=1}^{k} a_i + \sum_{i=k+1}^{n} a_i, \qquad k < n.$$

En general, sea $R = R_1 \cup R_2$, tal que $R_1 \cap R_2 = \varnothing$. Entonces,

$$\sum_{i \in R} a_i = \sum_{i \in R_1} a_i + \sum_{i \in R_2} a_i.$$

EJEMPLO 1.6 (Propiedad telescópica). Suponga $a_0, a_1, \ldots, a_n$ una secuencia de números reales, y considere

$$\begin{aligned} \sum_{i=1}^{n} (a_i - a_{i-1}) &= (a_1 - a_0) + (a_2 - a_1) + \cdots + (a_{n-1} - a_{n-2}) + (a_n - a_{n-1}) \\ &= -a_0 + (a_1 - a_1) + \cdots + (a_{n-1} - a_{n-1}) + a_n \\ &= a_n - a_0. \end{aligned}$$

EJEMPLO 1.7 (Suma de una progresión geométrica). Asuma que $x \neq 1$ y $n \geq 0$. Entonces,

$$\begin{aligned} a + ax + \cdots + ax^n &= \sum_{j=0}^{n} ax^j \\ = a + \sum_{j=1}^{n} ax^j &= a + (ax + \cdots + ax^n) = a + x(a + \cdots + ax^{n-1}) \qquad (1.4) \\ = a + x \sum_{j=1}^{n} ax^{j-1} &= a + x \sum_{j=0}^{n-1} ax^j \qquad (1.5) \\ = a + x \sum_{j=0}^{n} ax^j - ax^{n+1}. & \qquad (1.6) \end{aligned}$$

Tomando la primera y última relaciones, sigue que

$$\sum_{j=0}^{n} ax^j = a + x \sum_{j=0}^{n} ax^j - ax^{n+1},$$

es decir,

$$(1-x)\sum_{j=0} ax^j = a - ax^{n+1}.$$

De este modo, obtenemos finalmente

$$\sum_{j=0}^{n} ax^j = a\Big(\frac{1-x^{n+1}}{1-x}\Big).$$

En este ejemplo se han utilizado diversos elementos que permiten notar algunas de las propiedades de las sumatorias. En efecto, en Ecuación (1.4) se utilizó una muy particular versión del Resultado 1.5. Mientras que en (1.5) y (1.6) se realizó un cambio en el índice de la suma y reorganizó los términos (es decir el dominio sobre el que opera la suma) para obtener el resultado deseado.

Las siguientes son igualdades que **no satisface** la suma:

- Sean $a_1, \ldots, a_n$ y $b_1, \ldots, b_n$ dos secuencias de números reales. Entonces

$$\sum_{i=1}^{n} a_i b_i \neq \Big(\sum_{i=1}^{n} a_i\Big)\Big(\sum_{i=1}^{n} b_i\Big). \tag{1.7}$$

  En efecto, basta notar que la cantidad de términos involucrados en cada uno de los lados de la ecuación anterior es diferente.

- Un caso particular del anterior es

$$\sum_{i=1}^{n} x_i^2 \neq \Big(\sum_{i=1}^{n} x_i\Big)^2.$$

- En general, si $f : \mathbb{R} \to \mathbb{R}$ es una función no lineal. Entonces

$$\sum_{i=1}^{n} f(x_i) \neq f\Big(\sum_{i=1}^{n} x_i\Big).$$

En ocasiones disponemos de secuencias de números indexados mediante dos (o más) índices, es decir $\{a_{ij}\}$. Por ejemplo, consideremos $\boldsymbol{A} = (a_{ij}) \in \mathbb{R}^{m\times n}$ (es decir $i = 1, \ldots, m$, $j = 1, \ldots, n$) y suponga que deseamos sumar todos los elementos de la matriz $\boldsymbol{A}$. Es decir,

$$\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij} = a_{11} + \cdots + a_{1n} + a_{21} + \cdots + a_{m1} + \cdots + a_{mn}.$$

Notamos fácilmente que podemos intercambiar el orden de las sumas. En efecto,

$$\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij} = \sum_{j=1}^{n}\sum_{i=1}^{m} a_{ij}$$

Observación 1.8. Se debe resaltar que la operación de intercambiar el orden de las sumas **no** siempre es válido para series infinitas.

Contrariamente al resultado de la Ecuación (1.7), es válido considerar

$$\Big(\sum_{i=1}^{m} a_i\Big)\Big(\sum_{j=1}^{n} b_j\Big) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j, \tag{1.8}$$

asimismo

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

Para comprender mejor la Ecuación (1.8) considere un caso especial

$$\begin{aligned}\Big(\sum_{i=1}^{2} a_i\Big)\Big(\sum_{j=1}^{3} b_j\Big) &= (a_1 + a_2)(b_1 + b_2 + b_3)\\ &= (a_1b_1 + a_1b_2 + a_1b_3) + (a_2b_1 + a_2b_2 + a_2b_3)\\ &= \sum_{i=1}^{2}\Big(\sum_{j=1}^{3} a_i b_j\Big).\end{aligned}$$

Otros ejemplos de sumas útiles (que pueden ser probadas usando inducción) son:

- $\displaystyle\sum_{k=1}^{n} k = 1 + 2 + \cdots + n = \frac{n(n+1)}{2}.$
- $\displaystyle\sum_{k=1}^{n} k^2 = 1^2 + 2^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6}.$

Existe una notación análoga para productos. En efecto, considere la siguiente definición

DEFINICIÓN 1.9. Sea $a_1, a_2, \ldots$ una secuencia de números. Se define la *productoria* de esta secuencia, como:

$$\prod_{i=1}^{n} a_i = a_1 a_2 \cdots a_n. \tag{1.9}$$

En general, podemos escribir

$$\prod_{i\in R} a_i,$$

donde R representa un conjunto de índices. Note que si no existe algún entero $i \in R$, el producto se define con el valor uno.

EJEMPLO 1.10 (factorial de un número). Un ejemplo del uso de productorios es:

$$1 \cdot 2 \cdot 3 \cdots (n-1) \cdot n = \prod_{j=1}^{n} j = n!$$

que se denomina n factorial. Recuerde que 0! por definición es 1.

OBSERVACIÓN 1.11. En efecto, el factorial tiene una propiedad interesante. Es muy fácil notar que:

$$n! = (n-1)!\, n,$$

lo que permite definir el factorial de manera *recursiva*.

1.2. Estadísticas de Resumen

1.2.1. Medidas de Posición. Se introduce una serie de medidas que permiten resumir un gran volumen de información y cuantifican el valor central de un conjunto de datos.

DEFINICIÓN 1.12 (Media muestral o promedio). Sea $x_1, \ldots, x_n$ valores muestrales. Se define el *promedio* o *media muestral* como:

$$\overline{x} = \frac{1}{n}(x_1 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{1.10}$$

Suponga que la observación i-ésima, digamos x_i, se repite n_i veces. Entonces tenemos que

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} n_i x_i = \sum_{i=1}^{n} f_i x_i, \tag{1.11}$$

donde $f_i = n_i/n$ es la frecuencia relativa. Considere "pesos" o ponderaciones $\omega_1, \ldots, \omega_n$ asociados a las observaciones $x_1, \ldots, x_n$. En este caso tenemos,

$$\overline{x} = \frac{1}{\sum_{j=1}^{n} \omega_j} \sum_{i=1}^{n} \omega_i x_i, \tag{1.12}$$

si consideramos la proporción

$$p_i = \frac{\omega_i}{\sum_{j=1}^{n} \omega_j}, \qquad i = 1, \ldots, n,$$

entonces podemos re-escribir (1.12) como

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

Note que el promedio en Ecuación (1.11) es un caso particular donde $\sum_i f_i = 1$.

EJEMPLO 1.13. Considere el conjunto de datos $\{1, 2, 2, 2, 3, 3, 8\}$. Tenemos $n = 7$, y

$$\sum_{i=1}^{7} x_i = 1 + 3 \cdot 2 + 2 \cdot 3 + 8 = 21,$$

así $\overline{x} = 21/7 = 3$. Note también que el gráfico de *tallo y hoja*, adopta la forma:

1	*
2	* * *
3	* *
4	
5	
6	
7	
8	*

EJEMPLO 1.14 (Datos de accidentes). Suponga el siguiente conjunto de datos:

Número de accidentes (x_i)	Frecuencia (n_i)	$n_i x_i$
0	55	0
1	14	14
2	5	10
3	2	6
4	0	0
Total	76	30

De este modo, $\overline{x} = 30/76 = 0.395$ es el número promedio de accidentes.

Considere que el conjunto de observaciones $x_1, \ldots, x_n$ es ordenado de menor a mayor como:

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)},$$

tal que $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$, donde $x_{(k)}$ se denomina la k-ésima estadística de orden.

DEFINICIÓN 1.15 (Mediana). Sean $x_{(1)}, \ldots, x_{(n)}$ observaciones ordenadas. Si n es impar, entonces la *mediana* se define como la observación central, es decir

$$\mathsf{me} = x_{((n+1)/2)},$$

en el caso que n sea par, entonces

$$\mathsf{me} = \frac{1}{2}\Big(x_{(n/2)} + x_{(n/2+1)}\Big).$$

OBSERVACIÓN 1.16. En ocasiones escribiremos $\mathsf{me}(\boldsymbol{x})$ para indicar cual es el conjunto de datos sobre el que se calcula la mediana.

DEFINICIÓN 1.17 (Moda). La *moda* o *valor modal* es el valor observado con la más alta ocurrencia.

OBSERVACIÓN 1.18. Respecto de las medidas de tendencia central introducidas anteriormente podemos apreciar que:

- El promedio puede verse fuertemente afectado por *datos atípicos*.
- La mediana "divide" el conjunto de datos en dos, es decir, el 50 % de los datos están por debajo de la mediana, mientras que el 50 % se encuentran por sobre este valor.
- En general, la moda no es única y puede no existir.
- Es interesante notar la diferencia entre la complejidad de cálculo del promedio versus el de la mediana.

OBSERVACIÓN 1.19 (Otras medidas de tendencia central). Sea $f(x)$ cualquier función de números reales. Entonces podemos definir

$$\overline{f} = \frac{1}{n}\sum_{i=1}^{n} f(x_i) = \frac{1}{n}\Big(f(x_1) + \cdots + f(x_n)\Big).$$

Los siguientes casos particulares son de interés:

(a) *Média cuadrática.* Considere $f(x) = x^2$. Entonces se define la media cuadrática, Q como:

$$Q = \frac{1}{n}\sum_{i=1}^{n} x_i^2.$$

(b) *Média armónica.* Sea $f(x) = 1/x$. Entonces, decimos que H es la media armónica si

$$\frac{1}{H} = \frac{1}{n}\Big(\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}\Big) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}.$$

Es decir, H es el inverso de la media aritmética de los inversos de los valores observados. De donde sigue que,

$$H = \frac{n}{\sum_{i=1}^{n} 1/x_i}.$$

(c) *Média geométrica.* Considere $f(x) = \log x$. Entonces la media geométrica G es definida por la fórmula

$$\log G = \frac{1}{n}\Big(\log x_1 + \cdots + \log x_n\Big) = \frac{1}{n}\sum_{i=1}^{n}\log x_i. \tag{1.13}$$

Es decir,

$$G = \Big(\prod_{i=1}^{n} x_i\Big)^{1/n}.$$

OBSERVACIÓN 1.20. Un procedimiento bastante usado para el cálculo de la média geométrica es obtener la média aritmética dada en la Ecuación (1.13), y luego considerar

$$G = \exp\Big(\frac{1}{n}\sum_{i=1}^{n}\log x_i\Big).$$

Sin embargo, este método aunque correcto, puede inducir algunos errores en la precisión de los resultados. Una alternativa que intenta corregir esta situación, es basada en el trabajo de Graillat (2009), quien propuso un esquema compensado para la evaluación de productos utilizando un mecanismo de multiplicación y suma fundidas (FMA). La biblioteca fastmatrix (Osorio y Ogueda, 2022), contiene una implementación de la média geométrica usando este enfoque. El siguiente ejemplo, con comandos en R ilustra el uso de la función `geomean`.

```
# introduciendo datos en la consola de R
> x <- c(2.2, 0.3, 0.5, 0.4, 0.2, 1.9)

# cargando biblioteca 'fastmatrix'
> library(fastmatrix)
> geomean(x)
[1] 0.6072855
> exp(mean(log(x))) # equivalente a 'geomean'
[1] 0.6072855
```

1.2.2. Medidas de Dispersión. Considere los conjuntos de datos:

$$D_1 = \{10, 20, 30\}, \qquad D_2 = \{5, 5, 20, 35, 35\}, \qquad D_3 = \{20, 20, 20\}.$$

Tenemos los gráficos de tallo-y-hoja:

Datos D_1:

5	
10	*
15	
20	*
25	
30	*
35	

Datos D_2:

5	* *
10	
15	
20	*
25	
30	
35	* *

Datos D_3:

5	
10	
15	
20	* * *
25	
30	
35	

Sea $\overline{x}_j$ y me_j el promedio y la mediana asociada al conjunto de datos D_j $(j = 1, 2, 3)$. Entonces,

$$\begin{aligned}
\overline{x}_1 &= \frac{1}{3}(10 + 20 + 30) = \frac{60}{3} = 20, \\
\overline{x}_2 &= \frac{1}{5}(2 \cdot 5 + 20 + 2 \cdot 35) = \frac{100}{5} = 20, \\
\overline{x}_3 &= \frac{3 \cdot 20}{3} = 20.
\end{aligned}$$

Además, $\mathsf{me}_j = 20$ para $j = 1, 2, 3$. Es decir, tenemos tres configuraciones de datos con valores centrales idénticos.

Esto motiva la introducción de medidas que permitan caracterizar la variabilidad o dispersión de un conjunto de observaciones. Algunas medidas simples corresponden a los *cuartiles*. En efecto, sea Q_1 y Q_3 las medianas de la mitad inferior y superior de los datos, respectivamente. Entonces, esto lleva a definir la siguiente medida de dispersión:

$$IQR = Q_3 - Q_1,$$

el que es conocido como *rango intercuartílico*. Es interesante notar que algunos software estadísticos (por ejemplo, R/S-Plus, Stata, entre otros) reportan:

$$\text{mín}\{x_i\}_{i=1}^n, Q_1, \mathsf{me}, Q_3, \text{máx}\{x_i\}_{i=1}^n.$$

Así, también podemos considerar el *rango* de la muestra como

$$R = \text{máx}\{x_i\}_{i=1}^n - \text{mín}\{x_i\}_{i=1}^n = x_{(n)} - x_{(1)}.$$

Por otro lado, se ha sugerido utilizar subdivisiones más finas que los cuartiles. Por ejemplo, considere subdividir los datos ordenados $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ en secciones de 100 %, llamados *percentiles*. Precisamente, si disponemos de los datos ordenados $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, entonces el percentil de orden j $(1 \leq j \leq 100)$ está dado por:

$$P_j = x_{(j(n+1)/100)}. \tag{1.14}$$

Debemos notar además que, el primer cuartil Q_1 corresponde al percentil $25^{\underline{o}}$, mientras que la mediana (o $2^{\underline{o}}$ cuartil, Q_2) representa el percentil $50^{\underline{o}}$ y Q_3 corresponde al percentil $75^{\underline{o}}$.

EJERCICIO 1.21. Considere el conjunto de datos $\boldsymbol{x} = (4, 7, 18, 1, 7, 13, 2)^\top$ y suponga que deseamos calcular P_{75} y el rango intercuartílico IQR. Primeramente es necesario ordenar el conjunto de datos:

$$(x_{(1)}, x_{(2)}, x_{(3)}, x_{(4)}, x_{(5)}, x_{(6)}, x_{(7)})^\top = (1, 2, 4, 7, 7, 13, 18)^\top.$$

Disponemos de $n = 7$ datos, luego para obtener el 1er y 3er cuartiles podemos usar la fórmula del percentil en (1.14). En efecto,

$$Q_1 = P_{25} = x_{(25\cdot(7+1)/100)} = x_{(1\cdot 8/4)} = x_{(2)} = 2,$$
$$Q_3 = P_{75} = x_{(75\cdot(7+1)/100)} = x_{(3\cdot 8/4)} = x_{(6)} = 13.$$

De este modo, $IQR = Q_3 - Q_1 = 13 - 2 = 11$.

DEFINICIÓN 1.22 (Varianza muestral). Considere $x_1, x_2, \ldots, x_n$ valores observados, se define su *varianza muestral* como:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2. \tag{1.15}$$

Note que s^2 corresponde a un "promedio" de los desvios al cuadrado con relación a la media. También se suele anotar s_x^2 o bien $\text{var}(x)$.

Aunque en ocasiones es recomendable dividir la suma de cuadrados en (1.15) por n. En el capítulo sobre inferencia veremos la razón de utilizar $n - 1$.

OBSERVACIÓN 1.23. $s = \sqrt{s^2}$ se denomina desviación estándar.

Basados en Ecuación (1.15) podemos definir otras medidas de dispersión:

- *Desviación absoluta en torno de la media*

$$\frac{1}{n}\sum_{i=1}^{n}|x_i - \overline{x}|. \tag{1.16}$$

- *Desviación absoluta en torno de la mediana*

$$\frac{1}{n}\sum_{i=1}^{n}|x_i - \mathsf{me}\,|. \tag{1.17}$$

OBSERVACIÓN 1.24. En general podemos considerar, por ejemplo:

$$g(T) = \frac{1}{n}\sum_{i=1}^{n} h(x_i - T(\boldsymbol{x})),$$

donde $T(\boldsymbol{x})$ es alguna estadística de la muestra $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$. Note que si $T(\boldsymbol{x}) = \overline{x}$ y $h(z) = z^2$, obtenemos la varianza. Mientras que para $T(\boldsymbol{x}) = \mathsf{me}$ y $h(z) = |z|$ obtenemos (1.17).

De este modo, podemos definir el r-ésimo momento centrado en torno de a, como

$$M_r(a) = \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^r. \tag{1.18}$$

PROPIEDAD 1.25. A continuación se describen una serie de propiedades del promedio y la varianza de un conjunto de datos observados.

(a) $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$. En efecto,

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

(b) (Fórmula de Köning)

$$\sum_{i=1}^{n}(x_i-\overline{x})^2=\sum_{i=1}^{n}(x_i^2-2x_i\overline{x}+\overline{x}^2)=\sum_{i=1}^{n}x_i^2-2\overline{x}\sum_{i=1}^{n}x_i+n\overline{x}^2$$
$$=\sum_{i=1}^{n}x_i^2-n\overline{x}^2=\sum_{i=1}^{n}x_i^2-\frac{1}{n}\Big(\sum_{i=1}^{n}x_i\Big)^2 \tag{1.19}$$

(c) $\overline{x}$ es el valor que minimiza la función $S(a)=\sum_{i=1}^{n}(x_i-a)^2$. En efecto, note que

$$\frac{\mathsf{d}}{\mathsf{d}a}S(a)=\sum_{i=1}^{n}\frac{\mathsf{d}}{\mathsf{d}a}(x_i-a)^2=-2\sum_{i=1}^{n}(x_i-a),$$

resolviendo la condición de primer orden, tenemos

$$\sum_{i=1}^{n}(x_i-\widehat{a})=0,$$

desde donde sigue que $\widehat{a}=\overline{x}$. Además

$$\frac{\mathsf{d}^2}{\mathsf{d}a^2}S(a)=-2\sum_{i=1}^{n}\frac{\mathsf{d}}{\mathsf{d}a}(x_i-a)=2n,$$

y como la segunda derivada es positiva (para cualquier valor de n), obtenemos que $\overline{x}$ es mínimo global.

Una manera alternativa para probar este resultado puede ser obtenida mediante notar que

$$S(a)=\sum_{i=1}^{n}(x_i-a)^2=\sum_{i=1}^{n}\{(x_i-\overline{x})+(\overline{x}-a)\}^2$$
$$=\sum_{i=1}^{n}(x_i-\overline{x})^2+2(\overline{x}-a)\sum_{i=1}^{n}(x_i-\overline{x})+\sum_{i=1}^{n}(\overline{x}-a)^2.$$

Por la propiedad en (a) y notando que el término $(\overline{x}-a)^2$ es constante para la suma. Obtenemos

$$S(a)=\sum_{i=1}^{n}(x_i-\overline{x})^2+n(\overline{x}-a)^2,$$

y de este modo, resulta evidente que $S(a)$ alcanza su mínimo para $\widehat{a}=\overline{x}$.

(d) Sea $x_1,x_2,\ldots,x_n$ y considere la transformación

$$y_i=ax_i+b.$$

Entonces

$$\overline{y}=a\overline{x}+b,\qquad s_y^2=a^2s_x^2.$$

Es fácil notar que,

$$\overline{y}=\frac{1}{n}\sum_{i=1}^{n}y_i=\frac{1}{n}\sum_{i=1}^{n}(ax_i+b)=\frac{1}{n}\Big(a\sum_{i=1}^{n}x_i+b\Big)$$
$$=a\Big(\frac{1}{n}\sum_{i=1}^{n}x_i\Big)+b=a\overline{x}+b.$$

Mientras que

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2,$$

como $y_i - \overline{y} = ax_i + b - (a\overline{x} + b) = a(x_i - \overline{x})$, sigue que

$$\begin{aligned} s_y^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{1}{n-1}\sum_{i=1}^{n}\{a(x_i - \overline{x})\}^2 \\ &= a^2\,\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = a^2 s_x^2. \end{aligned}$$

En particular, si (que corresponde a una *estandarización* del conjunto de datos $x_1, \ldots, x_n$)

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

Entonces $\overline{z} = 0$ y $s_z^2 = 1$. En efecto, en la Propiedad 1.25 (d) basta hacer $a = 1/s$ y $b = \overline{x}/s$.

(e) Sea $x_1, \ldots, x_n$ un conjunto de n observaciones. Considere aplicar la transformación

$$y_i = g(x_i), \qquad i = 1, \ldots, n,$$

con $g(\cdot)$ función dos veces diferenciable y suponga que utilizamos una aproximación de Taylor de primer orden en torno del promedio. Es decir,

$$g(x_i) \approx g(\overline{x}) + g'(\overline{x})(x_i - \overline{x}).$$

De este modo, $\overline{y} = \overline{g}(\boldsymbol{x})$, y

$$\begin{aligned} \overline{g}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} g(x_i) &\approx \frac{1}{n}\sum_{i=1}^{n}\{g(\overline{x}) + g'(\overline{x})(x_i - \overline{x})\} \\ &= g(\overline{x}) + g'(\overline{x})\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}) = g(\overline{x}). \end{aligned}$$

Mientras que,

$$\begin{aligned} s_y^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\big(g(x_i) - \overline{g}(\boldsymbol{x})\big)^2 \approx \frac{1}{n-1}\sum_{i=1}^{n}\big(g(x_i) - g(\overline{x})\big)^2 \\ &\approx \frac{1}{n-1}\sum_{i=1}^{n}\big(g(\overline{x}) + g'(\overline{x})(x_i - \overline{x}) - g(\overline{x})\big)^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}\big(g'(\overline{x})(x_i - \overline{x})\big)^2 = \{g'(\overline{x})\}^2\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \\ &= \{g'(\overline{x})\}^2 s_x^2. \end{aligned}$$

Para $g(x_i) = ax_i + b$ una transformación lineal, tenemos $g'(x_i)$ y luego, se recupera los resultados dados en el item (d). Por otro lado, si consideramos una aproximación de Taylor de segundo orden (en torno de $\overline{x}$),

$$y_i \approx g(\overline{x}) + g'(\overline{x})(x_i - \overline{x}) + \frac{g''(\overline{x})}{2}(x_i - \overline{x})^2.$$

Obtenemos $\overline{y} \approx g(\overline{x}) + g''(\overline{x})s_x^2/2$.

EJEMPLO 1.26. Considere un conjunto de datos $x_1, \ldots, x_n$. Verifique que

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

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,

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

Usando la Propiedad 1.25 (a), 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.

PROPIEDAD 1.27. La mediana es el valor que minimiza la función

$$Q(a) = \sum_{i=1}^{n} |x_i - a|.$$

DEMOSTRACIÓN. Podemos escribir,

$$Q(a) = \sum_{x_i \geq a} (x_i - a) - \sum_{x_i < a} (x_i - a),$$

así

$$\frac{\mathsf{d}}{\mathsf{d}\, a} Q(a) = - \sum_{x_i \geq a} 1 + \sum_{x_i < a} 1,$$

resolviendo la condición de primer orden $\mathsf{d}\, Q(a) / \mathsf{d}\, a = 0$, tenemos

$$\sum_{x_i \geq a} 1 = \sum_{x_i < a} 1 \qquad \Longrightarrow \qquad \widehat{a} = \mathsf{me},$$

por la propia definición de mediana. □

OBSERVACIÓN 1.28. Recuerde que

$$z = \text{signo}(z) \cdot |z|.$$

De este modo, $\text{signo}(z) = z/|z|$ (o bien $|z| = \text{signo}(z) \cdot z$). En efecto, tenemos que

$$\frac{\mathsf{d}}{\mathsf{d}\, z} |z| = \text{signo}(z), \qquad \text{para } z \neq 0.$$

Por tanto, una forma alternativa de escribir $\mathsf{d}\, Q(a)/\,\mathsf{d}\, a$ es

$$\frac{\mathsf{d}}{\mathsf{d}\, a} Q(a) = \sum_{i=1}^{n} \text{signo}(x_i - a),$$

es decir

$$\sum_{i=1}^{n} \text{signo}(x_i - a) = \sum_{i=1}^{n} \frac{1}{|x_i - a|}(x_i - a) = \sum_{i=1}^{n} \omega_i(a)(x_i - a), \tag{1.20}$$

donde $\omega_i(a) = |x_i - a|^{-1}$ (Ecuación (1.20) es un promedio ponderado). Sin embargo, resolver la condición de primer orden

$$\sum_{i=1}^{n} \omega_i(a)(x_i - a) = 0, \tag{1.21}$$

con relación a a es bastante difícil. Una alternativa es considerar el Algoritmo 1 presentado a continuación.

Algoritmo 1: Cálculo de la mediana usando un procedimiento basado en promedios iterativamente ponderados.

Entrada: Conjunto de n datos $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$, una aproximación inicial, a_0 (por ejemplo, $a_0 = \overline{x}$) y un valor de tolerancia $(= 1 \cdot 10^{-6})$

Salida : Aproximación para el valor de la mediana, $\widehat{a} = \mathsf{me}$.

begin

Calcular

$$\omega_i(a_0) \leftarrow \frac{1}{|x_i - a_0|}, \qquad i = 1, \ldots, n.$$

Actualizar:

$$a_1 = \frac{1}{\sum_{j=1}^{n} \omega_j(a_0)} \sum_{i=1}^{n} \omega_i(a_0) x_i. \tag{1.22}$$

if $|a_1 - a_0| <$ tolerancia **then**

  return $\mathsf{me} = a_1$, y detener el algoritmo.

else

  hacer, $a_0 \leftarrow a_1$

  volver a Paso **2**.

end

end

Debemos destacar que la etapa de actualización dada en la Ecuación (1.22), equivale a resolver

$$\sum_{i=1}^{n} \omega_i(a_0)(x_i - a) = 0,$$

con relación a "a", que es muchísimo más simple que resolver (1.21) pues ahora las ponderaciones $\omega_i(a_0)$ están *fijas*.

OBSERVACIÓN 1.29. ¿Qué deficiencias presenta el Algoritmo 1?

- Este algoritmo converge a la mediana (aunque su velocidad de convergencia puede ser bastante lenta).
- En la convergencia, digamos $\widehat{a} = \mathsf{me}$, tendremos exactamente un "peso" $\omega_i(\widehat{a})$ indefinido (¿Ud. podría 'adivinar' cuál?).
- Una alternativa para el punto anterior es usar que[1]

$$\text{signo}(z) \approx \frac{z}{\sqrt{z^2+\epsilon^2}}.$$

Por tanto, "cerca" del óptimo ($\widehat{a} = \mathsf{me}$) podemos considerar los *pesos modificados*:

$$\widetilde{\omega}_i(a) = \frac{1}{\sqrt{(x_i-a)^2+\epsilon^2}}, \qquad i = 1, \ldots, n.$$

EJEMPLO 1.30. Considere el siguiente conjunto de datos:

$$\boldsymbol{x} = (2.40, 2.70, 2.80, 3.03, 3.40, 3.70, 28.95)^\top.$$

Evidentemente el valor 28.95 es una observación que se destaca y puede ser considerada como atípica. Los valores de la media muestral y el desviación estándar están dados por $\overline{x} = 6.711$ y $s_x = 9.816$, respectivamente. En efecto, es posible especular que el valor de la media muestral no representa una buena estimación del centro de los datos. Suponga que $\boldsymbol{z}$ denota el conjunto donde hemos eliminado el valor "sospechoso"de 28.95. En este caso obtenemos,

$$\overline{z} = \frac{18.03}{6} = 3.005, \qquad s_z^2 = \frac{1.14075}{5} = 0.2282,$$

con $s_z = 0.4777$. Mientras que, cuando calculamos el valor de la mediana para ambos conjuntos de datos obtenemos

$$\mathsf{me}(\boldsymbol{x}) = 3.03, \qquad \mathsf{me}(\boldsymbol{z}) = \frac{2.80+3.03}{2} = 2.915.$$

Es decir, la mediana es un procedimiento muy apropiado para cuantificar el valor central de un conjunto de datos en presencia de observaciones aberrantes o atípicas. Se dice entonces que la mediana es una estadística *robusta*.
Por otro lado, una alternativa robusta a la desviación estandar se conoce como la *desviación mediana absoluta (MAD)*, que es definida como:

$$\text{MAD}(\boldsymbol{x}) = \mathsf{me}(|\boldsymbol{x} - \mathsf{me}(\boldsymbol{x})|).$$

Evaluando esta cantidad para los conjuntos de datos $\boldsymbol{x}$ y $\boldsymbol{z}$, obtenemos

$$\text{MAD}(\boldsymbol{x}) = \mathsf{me}\{0.63, 0.33, 0.23, 0.00, 0.37, 0.67, 25.92\} = 0.37$$

$$\text{MAD}(\boldsymbol{z}) = \mathsf{me}\{0.515, 0.215, 0.115, 0.115, 0.485, 0.785\} = \frac{0.215+0.485}{2} = 0.35.$$

Para hacer el MAD comparable con la desviación estándar, se define el *MAD normalizado* como:

$$\text{MADN}(\boldsymbol{x}) = \frac{\text{MAD}(\boldsymbol{x})}{0.6745}.$$

[1] Aproximación que mejora conforme $\epsilon \to 0$.

De este modo, obtenemos MADN($\boldsymbol{x}$) = 0.5486 mientras que MADN($\boldsymbol{z}$) = 0.5189. Comparativamente con $s_z = 0.4777$, claramente MAD no se ve influenciado fuertemente por la presencia de observaciones atípicas. A continuación presentamos un fragmento de comandos en R para el cálculo de algunas estadísticas de resumen:

```
# introduciendo datos en la consola de R
> x <- c(2.40, 2.70, 2.80, 3.03, 3.40, 3.70, 28.95)
> x
[1] 2.40 2.70 2.80 3.03 3.40 3.70 28.95

# removiendo la 7a observación
> z <- x[-7]
> z
[1] 2.40 2.70 2.80 3.03 3.40 3.70

# cálculo de la media muestral
> mean(x)
[1] 6.711429
> mean(z)
[1] 3.005

# cálculo de la mediana
> median(x)
[1] 3.03
> median(z)
[1] 2.915

# cálculo de la desviación estandar
> sd(x)
[1] 9.815978
> sd(z)
[1] 0.4776505
```

En R está disponible la función `mad`, para el cálculo del MAD o de su versión normalizada.

```
# cálculo del MAD
> abs(x - median(x))
[1] 0.63  0.33  0.23  0.00  0.37  0.67 25.92
> sort(abs(x - median(x)))
[1] 0.00  0.23  0.33  0.37  0.63  0.67 25.92

> mad(x, constant = 1) # MAD
[1] 0.37
> mad(x) # MAD normalizado
[1] 0.548562

> mad(z, constant = 1)
[1] 0.35
> mad(z)
[1] 0.51891
```

DEFINICIÓN 1.31 (Coeficiente de variación). Este coeficiente es una medida que compara la desviación estándar con el promedio de una muestra y es definido como

$$\text{CV} = s/\overline{x}, \qquad \overline{x} \neq 0.$$

El coeficiente es particularmente útil para comparar dos o más muestras (o grupos). Un valor pequeño para el CV está asociado a una muestra homogénea.

OBSERVACIÓN 1.32. CV es una medida adimensional. Debemos resaltar que, en Econometría, $1/\,\text{CV}$ es conocido como *razón de Sharpe*.

1.2.3. Cálculo del promedio y varianza muestrales. Es interesante notar que una mala implementación computacional puede hacer que un buen algoritmo sea inútil. Un ejemplo de esto son las pobres implementaciones para el cálculo de estadísticas básicas que son ofrecidas en Microsoft Excel. En efecto, McCullough y Wilson (1999, 2002, 2005) reportan una serie de falencias de los procedimientos estadísticos presentes en Excel. Por otro lado, software estadístico como R (R Core Team, 2019), o bien hojas de cálculo como Gnumeric[2] disponen de algoritmos confiables (ver, por ejemplo, Keeling y Pavur, 2007).

La definición de la varianza muestral dada en la Ecuación (1.15) permite sugerir un algoritmo en 2-pasos para el cálculo de s^2 (ver Algoritmo 2). Aunque se ha demostrado que este es un algoritmo estable (Chan y Lewis, 1979; Chan et al., 1983), puede no ser recomendable para grandes volúmenes de datos debido a que requiere *"pasar a través de los datos dos veces"*, es decir, requiere usar dos ciclos for. Mientras que la fórmula de Köning en (1.19) lleva a un algoritmo de 1-paso. En efecto, basado en la fórmula:

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

tenemos la implementación dada en el Algoritmo 3. Desafortunadamente, aunque este procedimiento es más veloz que el Algoritmo 2 (2-pasos), es bien sabido que puede llevar a cancelamientos 'catastróficos' y por tanto no es un método recomendable (ver Chan et al., 1983; Barlow, 1993). Para evitar este tipo de dificultades se ha propuesto algoritmos que explotan la definición de la media y varianza muestrales y que solo requieren de pasar por los datos una única vez.

A continuación se describe el *algoritmo online* (1-paso) propuesto por West (1979). Considere una muestra de tamaño n y sea

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

De este modo, evidentemente tenemos que

$$\begin{aligned}\overline{x}_n &= \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\Big(\sum_{i=1}^{n-1} x_i + x_n\Big) = \frac{1}{n}\Big((n-1)\overline{x}_{n-1} + x_n\Big) \\ &= \frac{1}{n}\Big(n\overline{x}_{n-1} - \overline{x}_{n-1} + x_n\Big) = \overline{x}_{n-1} + \frac{x_n - \overline{x}_{n-1}}{n}. \end{aligned} \tag{1.23}$$

[2]URL: www.gnumeric.org

Algoritmo 2: Varianza muestral usando un algoritmo de 2-pasos.

Entrada: Conjunto de n datos $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$.
Salida : Promedio y varianza muestrales, $\overline{x}$ y s^2.

begin
  $M \leftarrow x_1$
  for $i = 2$ **to** n **do**
    $M \leftarrow M + x_i$
  end
  $M \leftarrow M/n$
  $T \leftarrow (x_1 - M)^2$
  for $i = 2$ **to** n **do**
    $T \leftarrow T + (x_i - M)^2$
  end
  $\overline{x} \leftarrow M$
  $s^2 \leftarrow \dfrac{1}{n-1} T$
end

Algoritmo 3: Varianza muestral usando un algoritmo de 1-paso.

Entrada: Conjunto de n datos $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$.
Salida : Promedio y varianza muestrales, $\overline{x}$ y s^2.

begin
  $M \leftarrow x_1$
  $T \leftarrow x_1^2$
  for $i = 2$ **to** n **do**
    $M \leftarrow M + x_i$
    $T \leftarrow T + x_i^2$
  end
  $\overline{x} \leftarrow M/n$
  $s^2 \leftarrow \dfrac{1}{n-1} T - \dfrac{n}{n-1} \overline{x}^2$
end

La base del algoritmo propuesto por West (1979) es la relación recursiva definida en la Ecuación (1.23). Sea $\delta_n = x_n - \overline{x}_{n-1}$, también podemos definir un algoritmo recursivo para el cálculo de la varianza muestral. En efecto, considere

$$\begin{aligned} T_n &= \sum_{i=1}^{n} (x_i - \overline{x}_n)^2 = \sum_{i=1}^{n-1} (x_i - \overline{x}_n)^2 + (x_n - \overline{x}_n)^2 \\ &= \sum_{i=1}^{n-1} (x_i - \overline{x}_{n-1} - \delta_n/n)^2 + (x_n - \overline{x}_{n-1} - \delta_n/n)^2 \\ &= \sum_{i=1}^{n-1} \left[(x_i - \overline{x}_{n-1})^2 - 2\frac{\delta_n}{n}(x_i - \overline{x}_{n-1}) + \frac{\delta_n^2}{n^2} + \left(\delta_n - \frac{\delta_n}{n}\right)^2 \right] \end{aligned}$$

Sabemos por Propiedad 1.25 (a), que

$$\sum_{i=1}^{n-1}(x_i - \overline{x}_{n-1}) = 0,$$

De este modo, podemos escribir la suma de cuadrados T_n, como:

$$\begin{aligned} T_n &= \sum_{i=1}^{n-1}(x_i - \overline{x}_{n-1})^2 + (n-1)\frac{\delta_n^2}{n^2} + \left(1 - \frac{1}{n}\right)^2 \delta_n^2 \\ &= T_{n-1} + \left(1 - \frac{1}{n}\right)\delta_n^2. \end{aligned} \tag{1.24}$$

Ecuaciones (1.23) y (1.24) llevan al Algoritmo 4 (West, 1979), cuya definición se presenta a continuación.

Algoritmo 4: Promedio y varianza muestrales usando un algoritmo online.

Entrada: Conjunto de n datos $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$.
Salida : Promedio y varianza muestrales, $\overline{x}$ y s^2.

begin
  $M \leftarrow x_1$
  $T \leftarrow 0$
  for $i = 2$ **to** n **do**
    $\delta \leftarrow (x_i - M)/i$
    $M \leftarrow M + \delta$
    $T \leftarrow T + i(i-1)\delta^2$
  end
  $\overline{x} \leftarrow M$
  $s^2 \leftarrow \dfrac{1}{n-1}T$
end

Chan y Lewis (1979) introdujeron una medida que permite evaluar la sensibilidad de una muestra $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ cuando debemos desarrollar el cálculo de su varianza muestral.

Sea

$$\|\boldsymbol{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2},$$

la norma Euclidiana para el conjunto de datos $\boldsymbol{x}$. De este modo, se define el *número condición de una muestra*, como:

$$\kappa = \frac{\|\boldsymbol{x}\|_2}{\sqrt{n-1}s}.$$

Esta medida permite cuantificar el efecto de introducir errores en los datos y como éstos son magnificados en el cálculo de la varianza. Es interesante notar que el número condición puede ser utilizado para evaluar la estabilidad de un algoritmo para el cálculo de s^2. Además podemos verificar fácilmente que el número condición está relacionado con el coeficiente de variación. Considere el siguiente ejemplo.

EJEMPLO 1.33. Sea $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ una muestra de datos. Verifique que

$$\kappa^2 = 1 + \frac{n}{n-1}\,\mathrm{CV}^{-2}\,. \tag{1.25}$$

En efecto, sabemos que

$$\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2 + n\overline{x}^2,$$

dividiendo ámbos términos por $(n-1)s^2$, obtenemos

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

Además $\sum_{i=1}^n (x_i - \overline{x})^2 = (n-1)s^2$, entonces podemos escribir

$$\kappa^2 = 1 + \Big(\frac{n}{n-1}\Big)\Big(\frac{\overline{x}}{s}\Big)^2.$$

tenemos $CV = s/\overline{x}$ y el resultado sigue.

Desde (1.25), sigue que

$$\kappa = \sqrt{1 + n\,\mathrm{CV}^{-2}/(n-1)},$$

es decir, a menos que CV sea bastante grande $\kappa \approx \mathrm{CV}^{-1}$ será una buena aproximación.

Finalmente debemos destacar que Chan et al. (1983) propusieron un algoritmo extremadamente estable para el cálculo de la varianza basado en un procedimiento de *suma acumulada por pares* el cual puede ser fácilmente paralelizado.

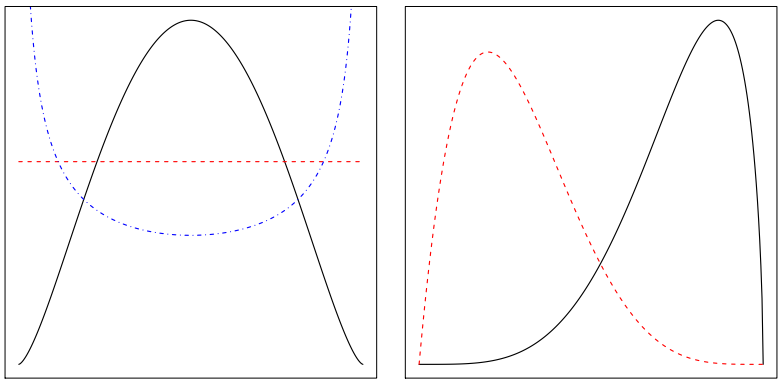
1.2.4. Medidas de forma. Momentos de orden mayor, o estadísticas involucrando potencias de orden mayor de los datos observados permiten caracterizar la forma de la densidad que describe el mecanismo que genera las n realizaciones $x_1, \ldots, x_n$ de nuestra variable de interés. A continuación revisamos la definición de los coeficientes de asimetría y curtosis, los que caracterizan el grado de asimetría de una distribución en torno de su promedio y el grado de agudeza o achatamiento de una distribución al ser comparada contra la distribución normal (gaussiana).

DEFINICIÓN 1.34 (Coeficiente de asimetría). Considere M_3 el tercer momento muestral en torno del promedio. Entonces, se define el *coeficiente de asimetría* (o sesgo) como:

$$b_1 = \frac{M_3}{s^3} = \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{x_i - \overline{x}}{s}\Big)^3.$$

OBSERVACIÓN 1.35. b_1 también es conocido como coeficiente de asimetría de Fisher y se caracteriza por ser una medida adimensional así como por ser invariante bajo traslaciones del origen y transformaciones de escala. Considere los gráficos desplegados en la Figura 1

- Si $b_1 = 0$ la distribución es simétrica con relación a la media.
- Si $b_1 > 0$ la distribución tiene sesgo positivo (o hacia la derecha), en cuyo caso la distribución tiende a concentrarse en valores altos de la variable. En caso contrario ($b_1 < 0$), diremos que su sesgo es negativo (o que es asimétrica hacia la izquierda).



(a) (b)

Figura 1. (a) distribuciones simétricas, (b) distribución con asimetría negativa (—) y asimetría positiva (– –).

OBSERVACIÓN 1.36. Se han definido varios índices de simetría, por ejemplo el *coeficiente de asimetría de Galton*:

$$b_{\mathsf{G}} = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1}.$$

DEFINICIÓN 1.37 (Coeficiente de curtosis). Considere M_4 el cuarto momento muestral en torno del promedio. Entonces, se define el *coeficiente de curtosis* (o achatamiento) como:

$$b_2 = \frac{M_4}{s^4} - 3 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s}\right)^4 - 3.$$

OBSERVACIÓN 1.38. El término -3 hace que $b_2 = 0$ cuando los datos siguen una distribución normal (gaussiana) en cuyo caso decimos que la distribución es *mesocúrtica*. Considere la Figura 2, si $b_2 > 0$ la distribución de los datos es más aguzada que la distribución normal (distribución *leptocúrtica*), mientras que si $b_2 < 0$ la distribución es más achatada que la normal (distribución *platicúrtica*).

OBSERVACIÓN 1.39. Debemos destacar que Spicer (1972) propuso calcular momentos centrales de hasta cuarto orden, M_2, M_3 y M_4 usando un algoritmo online. La función `moments` desde la biblioteca fastmatrix permite el cálculo de los coeficientes b_1 y b_2.

1.2.5. Estadísticas de resumen para datos agrupados. Cuando los datos han sido organizados en una *tabla de frecuencias* se dice que los datos se encuentran agrupados. El objetivo de esta sección es proporcionar fórmulas para las medidas de posición (o tendencia central), de dispersión y de forma sin la necesidad de desagregar los datos. Primeramente vamos a suponer que la variable de interés es discreta. Considere el siguiente ejemplo:

EJEMPLO 1.40. Se consultó las fichas de los empleados de una fábrica, registrándose el *número de cargas familiares* y se obtuvo los siguientes datos:

1	2	4	2	2	2	3	2	1	1	0	2	2
0	2	2	1	2	2	3	1	2	2	1	2	

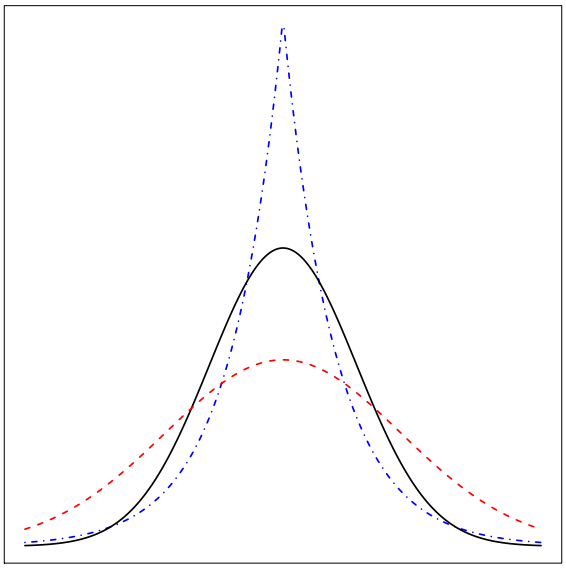


Figura 2. Distintos grados de curtosis: distribución leptocúrtica (– · –), mesocúrtica (—) y platicúrtica (– –).

En cuyo caso, tenemos que la tabla de frecuencias asociada asume la forma:

Cargas familiares	Número de empleados	Porcentaje de empleados	Num. acumulado de empleados	Porcentaje acumulado
0	2	8 %	2	8 %
1	6	23 %	8	32 %
2	14	56 %	22	88 %
3	2	8 %	24	96 %
4	1	4 %	25	100 %
Total	25	100 %	–	–

Podemos notar que los datos de interés corresponden a una variable discreta x (número de cargas familiares) que tiene k categorías. De este modo, podemos construir la siguiente tabla

Variable	Frecuencia Absoluta	Frecuencia Relativa	Frec. Abs. Acumulada	Frec. Rel. Acumulada
x_1	n_1	f_1	N_1	F_1
x_2	n_2	f_2	N_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_k	f_k	N_k	F_k
Total	n	1	–	–

donde

$$f_i = \frac{n_i}{n}, \qquad N_i = \sum_{j=1}^{i} n_j, \qquad F_i = \sum_{j=1}^{i} f_j, \tag{1.26}$$

para $i = 1, \ldots, k$. Evidentemente tenemos que

$$\sum_{i=1}^{k} n_i = n, \qquad N_k = n, \qquad \sum_{i=1}^{k} f_i = 1, \qquad F_k = 1. \tag{1.27}$$

En este contexto tenemos que la media y varianza muestrales están dados por

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \sum_{i=1}^{k} f_i x_i,$$

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

Para los datos del Ejemplo 1.40, es fácil notar que

$$\overline{x} = \frac{2\cdot 0 + 6\cdot 1 + 14\cdot 2 + 2\cdot 3 + 1\cdot 4}{25} = \frac{44}{25} = 1.760$$

$$s^2 = \frac{2(0-1.76)^2 + 6(1-1.76)^2 + 14(2-1.76)^2 + 2(3-1.76)^2 + 1(4-1.76)^2}{25-1}$$

$$= \frac{18.56}{24} = 0.773$$

Para datos continuos $x_1, \ldots, x_n$ tenemos la tabla de frecuencias:

Marca de clase	Frecuencia Absoluta	Frecuencia Relativa	Frec. Abs. Acumulada	Frec. Rel. Acumulada
C_1	n_1	f_1	N_1	F_1
C_2	n_2	f_2	N_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
C_k	n_k	f_k	N_k	F_k
Total	n	1	–	–

donde f_i, N_i y F_i, para cada una de las k categorías ha sido definido en (1.26) y (1.27), mientras que la *marca de clase* es definida como:

$$C_i = \frac{L_i + U_i}{2}, \qquad i = 1, \ldots, k,$$

con L_i y U_i los límites inferior y superior de cada intervalo, respectivamente. Note que la marca de clase es un representante de la clase (intervalo) respectiva. Análogamente a caso de una tabla de frecuencias para datos discretos, tenemos que:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_i C_i = \sum_{i=1}^{k} f_i C_i,$$

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

EJEMPLO 1.41. Considere los *precios de cierre* de acciones una determinada empresa nacional, dados por:

179	173	181	170	158	174	172	166	194	185
162	187	198	177	178	165	154	188	166	171
175	182	167	169	172	186	172	176	168	187

cuya tabla de frecuencias asume la forma:

Precio de cierre	Marca de clase	Número de días	Porcentaje de días	Num. de días acumulado	Porcentaje acumulado
(150,160]	155	2	7 %	2	7 %
(160,170]	165	8	27 %	10	34 %
(170,180]	175	11	36 %	21	70 %
(180,190]	185	7	23 %	28	93 %
(190,200]	195	2	7 %	30	100 %
Total	–	30	100 %	–	–

De este modo, podemos calcular la media y varianza muestrales desde la tabla de frecuencias, como:

$$\overline{x} = \frac{2 \cdot 155 + 8 \cdot 165 + 11 \cdot 175 + 7 \cdot 185 + 2 \cdot 195}{30} = \frac{5240}{30} = 174.667$$

mientras que,

$$\sum_{i=1}^{5} n_i (C_i - \overline{x})^2 = 2(155 - 174.667)^2 + 8(165 - 174.667)^2 + 11(175 - 174.667)^2 + 7(185 - 174.667)^2 + 2(195 - 174.667)^2 = 3096.667,$$

de este modo, $s^2 = 3096.667/(30-1) = 106.782$. Es interesante notar que los valores de la media y varianza muestrales obtenidos a partir de una tabla de frecuencias corresponden a una *aproximación* de los valores obtenidos a partir de los *datos a granel*. En efecto, el siguiente fragmento en R considera los datos crudos:

```
# precios de cierre (datos a granel)
> x <- c(179, 173, 181, 170, 158, 174, 172, 166, 194, 185,
+   162, 187, 198, 177, 178, 165, 154, 188, 166, 171, 175,
+   182, 167, 169, 172, 186, 172, 176, 168, 187)

# cálculo de media y varianza muestrales
> mean(x)
[1] 175.0667
> var(x)
[1] 105.7195
```

Se ha sugerido una serie de procedimientos para la construcción de tablas de frecuencia. Considere los siguientes pasos:

- Determinar el *número de categorías*, k usando por ejemplo:

$$k = \sqrt{n},$$
$$k = 1 + 3.3 \log_{10}(n), \qquad \text{(regla de Sturges)}$$
$$k = \lceil 2n^{1/3} \rceil, \qquad \text{(regla de Rice)}$$

- Calcular el rango,

$$R = \text{máx}\{x_i\}_{i=1}^{n} - \text{mín}\{x_i\}_{i=1}^{n},$$

- Determinar la *longitud de los intervalos* a_i para $i = 1, \ldots, k$. Usualmente, podemos elegir una longitud constante, $a_i = a$ como:

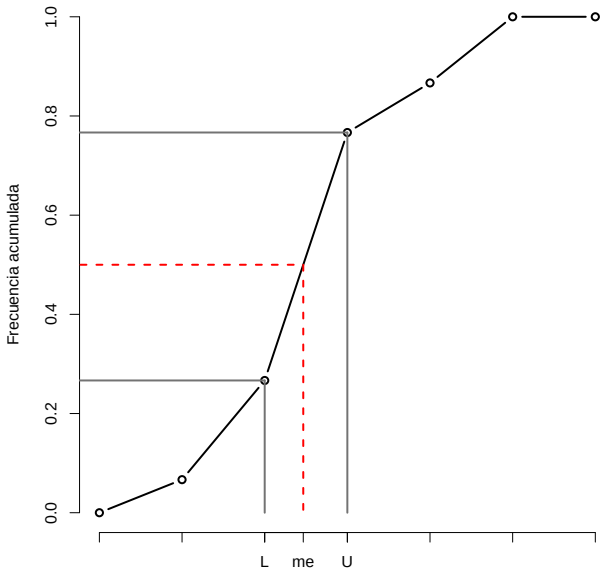
$$a = \frac{R}{k},$$

- Finalmente, construir los *límites de clases*

$$\begin{aligned} L_1 &= \text{mín}\{x_i\}_{i=1}^n - \triangle, & U_1 &= L_1 + a, \\ L_2 &= U_1, & U_2 &= L_2 + a, \\ &\vdots \end{aligned}$$

Debemos resaltar que aunque este tipo de procedimientos suelen ser bastante apropiados en la construción de tablas de frecuencia, deben ser entendidos meramente como *reglas de trabajo*.

Considere la función de distribución acumulada



de este modo, podemos usar interpolación lineal para determinar la mediana en datos agrupados. En efecto,

$$\frac{1/2 - F_{i-1}}{F_i - F_{i-1}} = \frac{\mathsf{me} - L_i}{U_i - L_i},$$

es decir,

$$(1/2 - F_{i-1})(U_i - L_i) = (\mathsf{me} - L_i)(F_i - F_{i-1}),$$

o bien

$$\mathsf{me} = L_i + \frac{1/2 - F_{i-1}}{F_i - F_{i-1}}(U_i - L_i) = L_i + \frac{1/2 - F_{i-1}}{F_i - F_{i-1}}\, a_i,$$

donde $a_i = U_i - L_i$, representa la amplitud de la clase mediana. Recordando que $F_i = \sum_{j=1}^{i} f_j$, tenemos

$$F_i - F_{i-1} = \sum_{j=1}^{j} f_j - \sum_{j=1}^{i-1} f_j = f_i + \sum_{j=1}^{i-1} f_j - \sum_{j=1}^{i-1} f_j = f_i.$$

Además, como $f_i = n_i/n$ y $F_i = N_i/n$, podemos re-escribir la *mediana para datos agrupados* como:

$$\begin{aligned}\mathsf{me} &= L_i + \frac{1/2 - F_{i-1}}{f_i}\, a_i = L_i + \frac{1/2 - F_{i-1}}{n_i/n}\, a_i \\ &= L_i + \frac{n/2 - N_{i-1}}{n_i}\, a_i\end{aligned}$$

Siguiendo exactamente la misma lógica, tenemos que el percentil k-ésimo, digamos P_k es dado por

$$P_k = L_i + \frac{k/100 - F_{i-1}}{f_i}\, a_i,$$

o bien

$$P_k = L_i + \frac{n(k/100) - N_{i-1}}{n_i}\, a_i.$$

EJEMPLO 1.42. Los trabajadores de una empresa, cuya tarea es clasificar y envasar fruta, obtuvieron los siguientes salarios semanales (clasificados según sexo).

Ingreso (UM)	Mujeres	Hombres
65 – 75	10	0
75 – 85	15	0
85 – 95	60	5
95 – 105	15	10
105 – 115	10	50
115 – 125	0	25
125 – 135	0	10
Total	110	100

Consideraremos solamente el grupo de mujeres y dejaremos el análisis del grupo de hombres y el total de trabajadores como ejercicio. Así, la tabla de frecuencias para el grupo de mujeres adopta la forma:

Ingreso (UM)	C_i	n_i	f_i	N_i	F_i
65 – 75	70	10	0.090	10	0.090
75 – 85	80	15	0.136	25	0.226
85 – 95	90	60	0.548	85	0.774
95 – 105	100	15	0.136	100	0.910
105 – 115	110	10	0.090	110	1.000
Total	–	110	1.000	–	–

En este caso, tenemos $n = 110$ y $\sum_{i=1}^{5} n_i C_i = 9\,900$. De este modo, la media aritmética para el grupo de mujeres es:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{5} n_i C_i = \frac{9900}{110} = 90 \text{ (UM)}.$$

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 (o

bien, frecuencia absoluta acumulada (N_i) que supere $n/2$). De este modo el intervalo mediano es $(85, 95]$. Además, $a_i = 10$ para todos los intervalos. De este modo,

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

donde $L_i = 85$, $F_{i-1} = 0.226$, $f_i = 0.548$ y $a_i = 10$, luego

$$\mathsf{me} = 85 + \frac{0.500 - 0.226}{0.548} \cdot 10 = 85 + 0.5 \cdot 10 = 90 \text{ (UM)}.$$

Por otro lado,

$$s^2 = \frac{1}{n-1}\Big(\sum_{i=1}^{5} n_i C_i^2 - n\overline{x}^2\Big).$$

En nuestro caso,

$$\sum_{i=1}^{5} n_i C_i^2 = 902\,000, \qquad \overline{x}^2 = 8\,100.$$

Así,

$$\begin{aligned} s^2 &= \frac{1}{110-1}(902\,000 - 110 \cdot 8\,100^2) = \frac{1}{109}(902\,000 - 891\,000) \\ &= \frac{11\,000}{109} = 100.9174 \text{ (UM)}^2. \end{aligned}$$

Además, tenemos que $s = \sqrt{11\,000/109} = 10.0458$ (UM). Mientras que

$$\text{CV} = \frac{10.0458}{90} = 0.1116.$$

Podemos evaluar la simetría usando el coeficiente de Galton. Por tanto, debemos calcular Q_1 y Q_3, como:

$$\begin{aligned} Q_1 &= 85 + \frac{0.250 - 0.226}{0.548} \cdot 10 = 85.438 \\ Q_3 &= 85 + \frac{0.750 - 0.226}{0.548} \cdot 10 = 94.562, \end{aligned}$$

de este modo $IQR = 9.1240$, y

$$\begin{aligned} \gamma_{\mathsf{G}} &= \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} \\ &= \frac{(94.562 - 90) - (90 - 85.438)}{9.124} = \frac{4.562 - 4.562}{9.124} = 0.000 \end{aligned}$$

Es decir, la distribución de los datos es simétrica.

Mientras que la moda interpolada, es dada por

$$\mathsf{mo} = L_{\mathsf{m}} + \frac{\Delta_1}{\Delta_1 + \Delta_2}\, a_{\mathsf{m}},$$

donde

- L_{m} es el límite inferior de la clase modal.
- $\Delta_1 = n_{\mathsf{m}} - n_{\mathsf{m}-1}$ con n_{m} la frecuencia absoluta de la clase modal, mientras que $n_{\mathsf{m}-1}$ es la frecuencia absoluta de la clase anterior a la clase modal.
- $\Delta_2 = n_{\mathsf{m}} - n_{\mathsf{m}+1}$ con $n_{\mathsf{m}+1}$ la frecuencia absoluta de la clase posterior a la clase modal.
- a_{m} es la amplitud de la clase modal.

OBSERVACIÓN 1.43. La clase modal es aquella que tiene la mayor frecuencia relativa. Note además, que en una tabla de frecuencia podría existir más de una clase modal.

1.3. Covarianza y correlación

Considere $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ y $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ dos vectores de datos con n observaciones. En esta sección se introducen medidas de asociación entre variables continuas y ordinales. Primeramente se introduce el concepto de covarianza.

DEFINICIÓN 1.44 (Covarianza). Para el conjunto $(x_1, y_1), \ldots, (x_n, y_n)$, se define la covarianza como una medida de variabilidad conjunta de dos variables cuantitativas, como:

$$\mathsf{cov}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}).$$

OBSERVACIÓN 1.45. En efecto, es fácil apreciar que $\mathsf{cov}(\boldsymbol{x}, \boldsymbol{x}) = s_x^2$. Además, evidentemente la covarianza está relacionada con el producto interno entre dos vectores n-dimensionales, $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \sum_{i=1}^{n} u_i v_i$.

EJEMPLO 1.46. Tal como en la Propiedad 1.25 (b), a continuación derivamos una expresión alternativa para la covarianza. En efecto,

$$\begin{aligned} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) &= \sum_{i=1}^{n} x_i (y_i - \overline{y}) - \overline{x} \sum_{i=1}^{n} (y_i - \overline{y}) \\ &= \sum_{i=1}^{n} x_i y_i - \overline{y} \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i - n \overline{x}\, \overline{y}. \end{aligned}$$

De este modo, podemos escribir

$$\mathsf{cov}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{n-1} \Big(\sum_{i=1}^{n} x_i y_i - n \overline{x}\, \overline{y} \Big).$$

PROPIEDAD 1.47. Tenemos que

$$\mathsf{cov}(a\boldsymbol{x} + b, c\boldsymbol{y} + d) = ac\, \mathsf{cov}(\boldsymbol{x}, \boldsymbol{y}).$$

DEMOSTRACIÓN. Sea $z_i = ax_i + b$ y $w_i = cy_i + d$ para $i = 1, \ldots, n$. Entonces $\overline{z} = a\overline{x} + b$ y $\overline{w} = c\overline{y} + d$, luego

$$\mathsf{cov}(\boldsymbol{z}, \boldsymbol{w}) = \frac{1}{n-1} \sum_{i=1}^{n} (z_i - \overline{z})(w_i - \overline{w}),$$

como $z_i - \overline{z} = a(x_i - \overline{x})$ y $w_i - \overline{w} = c(y_i - \overline{y})$. Entonces

$$\mathsf{cov}(\boldsymbol{z}, \boldsymbol{w}) = \frac{1}{n-1} ac \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}),$$

que es el resultado deseado. □

La covarianza permite medir la asociación, pero depende de la unidad de medida. Una alternativa es usar una medida conocida como *correlación*.

DEFINICIÓN 1.48 (Correlación). La correlación entre $\boldsymbol{x}$ e $\boldsymbol{y}$ es la covarianza de sus versiones estandarizadas. Es decir,

$$\begin{aligned}\mathsf{cor}(\boldsymbol{x},\boldsymbol{y}) &= \frac{1}{n-1}\sum_{i=1}^{n}\Big(\frac{x_i-\overline{x}}{s_x}\Big)\Big(\frac{y_i-\overline{y}}{s_y}\Big)\\ &= \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2\sum_{i=1}^{n}(y_i-\overline{y})^2}}\\ &= \frac{\mathsf{cov}(\boldsymbol{x},\boldsymbol{y})}{s_x s_y}.\end{aligned}$$

OBSERVACIÓN 1.49. $\mathsf{cor}(\boldsymbol{x},\boldsymbol{y})$ es una medida adimensional.

EJERCICIO 1.50. Muestre que

$$\mathsf{cor}(a\boldsymbol{x}+b, c\boldsymbol{y}+d) = \pm\,\mathsf{cor}(\boldsymbol{x},\boldsymbol{y}).$$

EJEMPLO 1.51. Verifique que

$$s^2_{x+y} = s^2_x + s^2_y + 2\,\mathsf{cov}(\boldsymbol{x},\boldsymbol{y}) = s^2_x + s^2_y + 2\,\mathsf{cor}(\boldsymbol{x},\boldsymbol{y})s_x s_y.$$

En efecto, podemos notar que

$$\begin{aligned}s^2_{x+y} &= \frac{1}{n-1}\sum_{i=1}^{n}\big((x_i-\overline{x})+(y_i-\overline{y})\big)^2\\ &= \frac{1}{n-1}\Big(\sum_{i=1}^{n}(x_i-\overline{x})^2+\sum_{i=1}^{n}(y_i-\overline{y})^2+2\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})\Big)\\ &= s^2_x + s^2_y + 2\,\mathsf{cov}(\boldsymbol{x},\boldsymbol{y})\end{aligned}$$

OBSERVACIÓN 1.52. Cuando $\mathsf{cor}(\boldsymbol{x},\boldsymbol{y}) = 0$ diremos que $\boldsymbol{x}$ e $\boldsymbol{y}$ son *no correlacionados*.

OBSERVACIÓN 1.53. Si $\boldsymbol{x}$ e $\boldsymbol{y}$ son no correlacionados, entonces

$$s^2_{x+y} = s^2_x + s^2_y.$$

PROPIEDAD 1.54. Tenemos que $\{\mathsf{cor}(\boldsymbol{x},\boldsymbol{y})\}^2 \leq 1$.

DEMOSTRACIÓN. Usando la desigualdad de Cauchy-Schwarz[3], tenemos que

$$\Big(\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})\Big)^2 \leq \Big(\sum_{i=1}^{n}(x_i-\overline{x})^2\Big)\Big(\sum_{i=1}^{n}(y_i-\overline{y})^2\Big),$$

es decir,

$$\begin{aligned}\Big(\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})\Big)^2 &\leq \Big(\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2\Big)\Big(\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\overline{y})^2\Big)\\ \{\mathsf{cov}(\boldsymbol{x},\boldsymbol{y})\}^2 &\leq s^2_x s^2_y.\end{aligned}$$

De este modo,

$$\frac{\mathsf{cov}^2(\boldsymbol{x},\boldsymbol{y})}{s^2_x s^2_y} \leq 1.$$

□

[3]Recuerde que $\langle \boldsymbol{u},\boldsymbol{v}\rangle^2 \leq \|\boldsymbol{u}\|^2\|\boldsymbol{v}\|^2$

Evidentemente, desde la propiedad anterior tenemos $-1 \leq \mathsf{cor}(\boldsymbol{x}, \boldsymbol{y}) \leq 1$.

Para datos que no son de naturaleza continua, una alternativa es reemplazar el valor de x_i (y_i) por su rango, R_i (S_i) que corresponde a un valor en el conjunto $\{1, 2, \ldots, n\}$. Considere la siguiente definición

DEFINICIÓN 1.55 (Coeficiente de correlación de Spearman). Suponga el conjunto de datos pareados $(x_1, y_1), \ldots, (x_n, y_n)$ y sea R_i el rango de x_i y S_i el rango de y_i, para $i = 1, \ldots, n$. Entonces el coeficiente de correlación de Spearman es dado por

$$r_{\mathsf{S}} = \frac{\sum_{i=1}^{n}(R_i - \overline{R})(S_i - \overline{S})}{\sqrt{\sum_{i=1}^{n}(R_i - \overline{R})^2 \sum_{i=1}^{n}(S_i - \overline{S})^2}}.$$

Considere

$$D = \sum_{i=1}^{n}(R_i - S_i)^2,$$

y suponga que no existen empates entre los x's e y's, entonces podemos escribir

$$r_{\mathsf{S}} = 1 - \frac{6D}{n(n^2 - 1)}.$$

OBSERVACIÓN 1.56. Análogamente a los resultados expuestos en la Sección 1.2.3, podemos llevar a cabo el cálculo de la covarianza entre $\boldsymbol{x}$ e $\boldsymbol{y}$ de manera eficiente considerando,

$$C_n = \sum_{i=1}^{n}(x_i - \overline{x}_n)(y_i - \overline{y}_n) = \sum_{i=1}^{n-1}(x_i - \overline{x}_n)(y_i - \overline{y}_n) + (x_n - \overline{x}_n)(y_n - \overline{y}_n).$$

Sea $\delta_n = x_n - \overline{x}_{n-1}$, de este modo podemos escribir:

$$x_i - \overline{x}_n = x_i - \overline{x}_{n-1} - \frac{\delta_n}{n}, \qquad x_n - \overline{x}_n = \Big(1 - \frac{1}{n}\Big)\delta_n,$$

y análogamente,

$$y_i - \overline{y}_n = y_i - \overline{y}_{n-1} - \frac{\eta_n}{n}, \qquad y_n - \overline{y}_n = \Big(1 - \frac{1}{n}\Big)\eta_n,$$

con $\eta_n = y_n - \overline{y}_{n-1}$. Notando que

$$\begin{aligned}(x_i - \overline{x}_n)(y_i - \overline{y}_n) &= \Big\{(x_i - \overline{x}_{n-1}) - \frac{\delta_n}{n}\Big\}\Big\{(y_i - \overline{y}_{n-1}) - \frac{\eta_n}{n}\Big\} \\ &= (x_i - \overline{x}_{n-1})(y_i - \overline{y}_{n-1}) - (x_i - \overline{x}_{n-1})\frac{\eta_n}{n} \\ &\quad - \frac{\delta_n}{n}(y_i - \overline{y}_{n-1}) + \frac{\delta_n \eta_n}{n^2}.\end{aligned}$$

Sumando sobre $\{1, \ldots, n-1\}$ y recordando que

$$\sum_{i=1}^{n-1}(x_i - \overline{x}_{n-1}) = 0, \qquad \sum_{i=1}^{n-1}(y_i - \overline{y}_{n-1}) = 0,$$

sigue que

$$\begin{aligned}C_n &= \sum_{i=1}^{n-1}(x_i - \overline{x}_{n-1})(y_i - \overline{y}_{n-1}) + \Big(\frac{n-1}{n^2}\Big)\delta_n\eta_n + \Big(\frac{n-1}{n}\Big)^2\delta_n\eta_n \\ &= C_{n-1} + \Big(\frac{n-1}{n}\Big)\delta_n\eta_n.\end{aligned}$$

Así, haciendo $\mathsf{cov}_n = \mathsf{cov}(\boldsymbol{x}, \boldsymbol{y})$, tenemos

$$\mathsf{cov}_n = \frac{1}{n-1} C_n = \frac{1}{n-1}\Big\{C_{n-1} + \Big(\frac{n-1}{n}\Big)\delta_n\eta_n\Big\},$$

pero $C_{n-1} = \big((n-1)-1\big)\,\mathsf{cov}_{n-1} = (n-1)\,\mathsf{cov}_{n-1} - \mathsf{cov}_{n-1}$. De este modo,

$$\begin{aligned}\mathsf{cov}_n &= \frac{1}{n-1}\Big\{(n-1)\,\mathsf{cov}_{n-1} - \mathsf{cov}_{n-1} + \Big(\frac{n-1}{n}\Big)\delta_n\eta_n\Big\} \\ &= \mathsf{cov}_{n-1} - \frac{\mathsf{cov}_{n-1}}{n-1} + \frac{(x_n - \overline{x}_{n-1})(y_n - \overline{y}_{n-1})}{n}.\end{aligned}$$

Este desarrollo lleva al siguiente algoritmo:

Algoritmo 5: Covarianza muestral usando un algoritmo online.

Entrada: Conjuntos de n datos $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ e $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$.
Salida : Covarianza muestral, $\mathsf{cov}(\boldsymbol{x}, \boldsymbol{y})$.

begin
  $M \leftarrow x_1$
  $N \leftarrow y_1$
  $C \leftarrow 0$
  for $i = 2$ **to** n **do**
    $\delta \leftarrow (x_i - M)/i$
    $\eta \leftarrow (y_i - N)/i$
    $M \leftarrow M + \delta$
    $N \leftarrow N + \eta$
    $C \leftarrow C - \dfrac{C}{i-1} + i\,\delta\,\eta$
  end
  $\mathsf{cov}(\boldsymbol{x}, \boldsymbol{y}) \leftarrow C$.
end

Además, es fácil notar que modificando ligeramente el Algoritmo 5 podemos obtener también $\overline{x}_n$, $\overline{y}_n$, s_x^2 y s_y^2.

1.4. Estadísticas descriptivas multivariadas

Considere que nuestro interés es estudiar $p \geq 2$ variables (características) de interés asociadas a una muestra aleatoria $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ donde cada $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ es un vector p-dimensional. Note que, podemos disponer la información en una *matriz de datos*

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1^\top \\ \boldsymbol{x}_2^\top \\ \vdots \\ \boldsymbol{x}_n^\top \end{pmatrix}.$$

Análogamente a la media y varianza muestrales $\overline{x}$ y s^2 unidimensionales, respectivamente, podemos definir sus contrapartes multivariadas como:

$$\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i,$$
$$\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^\top.$$

Además, la matriz de correlación entre las p variables de interés, es dada por:

$$\boldsymbol{R} = (r_{ij}),$$

donde

$$r_{ij} = \frac{\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)^2 \sum_{k=1}^{n}(x_{kj} - \overline{x}_j)^2}} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}},$$

con $\boldsymbol{S} = (s_{ij})$ y $\overline{x}_i = \big(\sum_{k=1}^{n} x_{ki}\big)/n$. Sea, $\boldsymbol{D} = \operatorname{diag}(s_{11}, s_{22}, \ldots, s_{pp})$, así podemos escribir

$$\boldsymbol{R} = \boldsymbol{D}^{-1/2}\boldsymbol{S}\boldsymbol{D}^{-1/2}.$$

Observación 1.57. Algunas propiedades del vector de medias y la matriz de covarianza, surgen de escribir formas compactas que dependen de la matriz de datos $\boldsymbol{X} = (x_{ij}) \in \mathbb{R}^{n\times p}$. En efecto,

$$\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i = \frac{1}{n}\boldsymbol{X}^\top \mathbf{1}.$$

Además,

$$\begin{aligned}\boldsymbol{Q} &= \sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^\top = \sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})\boldsymbol{x}_i^\top - \sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})\overline{\boldsymbol{x}}^\top \\ &= \sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^\top - \overline{\boldsymbol{x}}\sum_{i=1}^{n}\boldsymbol{x}_i^\top = \sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^\top - n\overline{\boldsymbol{x}}\,\overline{\boldsymbol{x}}^\top\end{aligned}$$

Es fácil notar que $\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^\top = \boldsymbol{X}^\top\boldsymbol{X}$, de este modo,

$$\begin{aligned}\boldsymbol{S} &= \frac{1}{n-1}\boldsymbol{Q} = \frac{1}{n-1}\Big\{\boldsymbol{X}^\top\boldsymbol{X} - n\Big(\frac{1}{n}\boldsymbol{X}^\top\mathbf{1}\Big)\Big(\frac{1}{n}\boldsymbol{X}^\top\mathbf{1}\Big)^\top\Big\} \\ &= \frac{1}{n-1}\Big(\boldsymbol{X}^\top\boldsymbol{X} - \frac{1}{n}\boldsymbol{X}^\top\mathbf{1}\mathbf{1}^\top\boldsymbol{X}\Big) = \frac{1}{n-1}\boldsymbol{X}^\top\boldsymbol{C}\boldsymbol{X}\end{aligned}$$

con $\boldsymbol{C} = \boldsymbol{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ la matriz de centrado. Es sencillo mostrar que

$$\boldsymbol{C}^\top = \boldsymbol{C}, \qquad \boldsymbol{C}^2 = \boldsymbol{C},$$

es decir $\boldsymbol{C}$ es matriz de proyección. Esto lleva al siguiente resultado.

Propiedad 1.58. La matriz de covarianza $\boldsymbol{S}$, es semidefinida positiva.

Demostración. Sea $\boldsymbol{a} \in \mathbb{R}^p$, vector no nulo. Tenemos que,

$$\begin{aligned}\boldsymbol{a}^\top\boldsymbol{S}\boldsymbol{a} &= \frac{1}{n-1}\boldsymbol{a}^\top\boldsymbol{X}^\top\boldsymbol{C}\boldsymbol{X}\boldsymbol{a} = \frac{1}{n-1}\boldsymbol{a}^\top\boldsymbol{X}^\top\boldsymbol{C}^2\boldsymbol{X}\boldsymbol{a} \\ &= \frac{1}{n-1}\boldsymbol{u}^\top\boldsymbol{u} \geq 0, \qquad \boldsymbol{u} = \boldsymbol{C}\boldsymbol{X}\boldsymbol{a},\end{aligned}$$

es decir, $\boldsymbol{S}$ es matriz semidefinida positiva.[4] □

Considere la siguiente transformación:

$$\boldsymbol{y}_i = \boldsymbol{A}\boldsymbol{x}_i + \boldsymbol{b}, \qquad i = 1, \ldots, n.$$

Entonces, $\overline{y} = \boldsymbol{A}\overline{\boldsymbol{x}} + \boldsymbol{b}$, mientras que

$$\boldsymbol{y}_i - \overline{\boldsymbol{y}} = \boldsymbol{A}\boldsymbol{x}_i + \boldsymbol{b} - \boldsymbol{A}\overline{\boldsymbol{x}} - \boldsymbol{b} = \boldsymbol{A}(\boldsymbol{x}_i - \overline{\boldsymbol{x}}).$$

De este modo,

$$\boldsymbol{S}_y = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{y}_i - \overline{\boldsymbol{y}})(\boldsymbol{y}_i - \overline{\boldsymbol{y}})^\top = \frac{1}{n-1}\sum_{i=1}^{n}\boldsymbol{A}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^\top\boldsymbol{A}^\top$$
$$= \frac{1}{n-1}\boldsymbol{A}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^\top\boldsymbol{A}^\top = \boldsymbol{A}\boldsymbol{S}_x\boldsymbol{A}^\top.$$

En particular, para la transformación (de Mahalanobis),

$$\boldsymbol{z}_i = \boldsymbol{S}^{-1/2}(\boldsymbol{x}_i - \overline{\boldsymbol{x}}), \qquad i = 1, \ldots, n,$$

donde $\boldsymbol{S} = \boldsymbol{S}^{1/2}\boldsymbol{S}^{1/2}$ con $\boldsymbol{S}^{1/2}$ un factor raíz cuadrada de $\boldsymbol{S}$. Entonces, sigue que

$$\overline{\boldsymbol{z}} = \boldsymbol{0}, \quad \text{y} \quad \boldsymbol{S}_z = \boldsymbol{I}_p.$$

DEFINICIÓN 1.59 (Distancia de Mahalanobis). Considere una muestra de n observaciones $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ de este modo, la distancia de Mahalanobis es dada por

$$D_i = \{(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^\top \boldsymbol{S}^{-1}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})\}^{1/2}, \qquad i = 1, \ldots, n,$$

como la distancia de la observación i-ésima hacia el "centro"de los datos, $\overline{\boldsymbol{x}}$ ponderada por la matriz de covarianza.

Sea

$$g_{ij} = (\boldsymbol{x}_i - \overline{\boldsymbol{x}})^\top \boldsymbol{S}^{-1}(\boldsymbol{x}_j - \overline{\boldsymbol{x}}), \qquad i, j = 1, \ldots, n.$$

lo que permite definir medidas de sesgo y curtosis multivariadas (Mardia, 1970), dadas por

$$b_{1p} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} g_{ij}^3, \qquad b_{2p} = \frac{1}{n}\sum_{i=1}^{n} g_{ii}^2,$$

respectivamente. Las cantidades anteriores llevan a un test para evaluar la normalidad multivariada (ver Mardia, 1974).

EJEMPLO 1.60. El conjunto de datos de flores Iris (Fisher, 1936), corresponden a 150 mediciones (en centímetros) del largo y ancho de los sépalos y largo y ancho de los pétalos para flores iris de las especies Setosa, Versicolor y Virginica. Los siguientes comandos en R permiten obtener el diagrama de dispersión así como algunas estadísticas de resúmen para los datos de Iris

```
  # datos Iris (50 observaciones por cada especie)
  > iris
    Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
1            5.1         3.5          1.4         0.2    setosa
2            4.9         3.0          1.4         0.2    setosa
...
```

[4] $\boldsymbol{S}$ será definida positiva si $n \geq p + 1$.

```
# extraémos solamente variables numéricas
> x <- iris[,1:4]
> pairs(x, col = iris$Species) # Fig.3
```

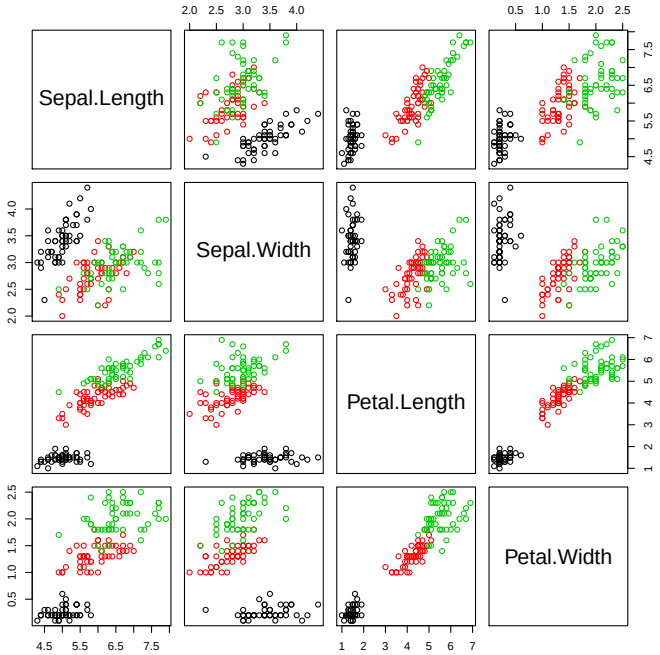


Figura 3. Gráficos de dispersión para los datos de Iris, según variedades Setosa (negro), Versicolor (rojo) y Virginica (verde).

El gráfico de dispersión desplegado en la Figura 3 revela una separación evidente en dos grupos (Setosa versus Versicolor y Virginica). Los siguientes comandos invocan funciones desde la biblioteca fastmatrix

```
# Carga biblioteca 'fastmatrix'
> library(fastmatrix)
# Cálculo de estadisticas de resumen multivariadas
> z <- cov.weighted(x) # por defecto, los 'pesos' son 1
> xbar <- z$mean
> S <- z$cov
> R <- cov2cor(z$cov)
> b1 <- skewness(x)
> b2 <- kurtosis(x)

# análogamente podemos calcular S y R usando:
> S <- cov(x)
> R <- cor(x)
```

En la línea de comandos de R podemos escribir `xbar`, `R`, `b1` y `b2` para obtener, el vector de medias, la matriz de correlación y el coeficiente de sesgo y curtosis muestrales, dados por:

$$\overline{\boldsymbol{x}} = \begin{pmatrix} 5.8433 \\ 3.0573 \\ 3.7580 \\ 1.1993 \end{pmatrix}, \qquad \boldsymbol{R} = \begin{pmatrix} 1.0000 & -0.1176 & 0.8718 & 0.8179 \\ -0.1176 & 1.0000 & -0.4284 & -0.3661 \\ 0.8718 & -0.4284 & 1.0000 & 0.9629 \\ 0.8179 & -0.3661 & 0.9629 & 1.0000 \end{pmatrix},$$

mientras que $b_{1p} = 2.6972$ y $b_{2p} = 23.7397$, respectivamente.

1.5. Regresión lineal simple

Ahora estamos enfocados en situaciones tales que la variable x permite predecir o explicar la respuesta y. En regresión lineal tenemos el modelo,

$$y_i = \alpha + \beta x_i + \epsilon_i, \qquad i = 1, \ldots, n. \tag{1.28}$$

El objetivo es, basado en el conjunto de datos pareados, $(x_1, y_1), \ldots, (x_n, y_n)$ determinar α y β tal que produzcan el mejor ajuste. Utilizaremos el *método de mínimos cuadrados ordinarios* (OLS) definido como:

$$\min_{\theta} S(\boldsymbol{\theta}),$$

con $\boldsymbol{\theta} = (\alpha, \beta)^\top$ y

$$S(\boldsymbol{\theta}) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 = \sum_{i=1}^{n} \epsilon_i^2.$$

Observación 1.61. OLS minimiza las distancias *verticales* a la recta de regresión. Es decir, esto refleja que los x_i's están *fijados* (o bien que se asumen conocidos).[5]

La función $S(\boldsymbol{\theta})$ también es conocida como suma de cuadrados de los errores. De este modo, debemos determinar $\boldsymbol{\theta} = (\alpha, \beta)^\top$, mediante resolver las condiciones de primer orden,

$$\frac{\partial S(\boldsymbol{\theta})}{\partial \alpha} = -2 \sum_{i=1}^{n} (y_i - \alpha - \beta x_i) = 0,$$
$$\frac{\partial S(\boldsymbol{\theta})}{\partial \beta} = -2 \sum_{i=1}^{n} x_i (y_i - \alpha - \beta x_i) = 0,$$

esto lleva a las *ecuaciones normales*:

$$\sum_{i=1}^{n} (y_i - \widehat{\alpha} - \widehat{\beta} x_i) = \sum_{i=1}^{n} y_i - n\widehat{\alpha} - \widehat{\beta} \sum_{i=1}^{n} x_i = 0,$$
$$\sum_{i=1}^{n} x_i (y_i - \widehat{\alpha} - \widehat{\beta} x_i) = \sum_{i=1}^{n} x_i y_i - \widehat{\alpha} \sum_{i=1}^{n} x_i - \widehat{\beta} \sum_{i=1}^{n} x_i^2 = 0.$$

Multiplicando la primera ecuación por n^{-1} y resolviendo para $\widehat{\alpha}$, obtenemos

$$\widehat{\alpha} = \overline{y} - \widehat{\beta} \overline{x}.$$

[5] Una alternativa es usar mínimos cuadrados totales o modelos con errores en las variables.

Substituyendo $\widehat{\alpha}$ en la segunda ecuación, tenemos

$$\begin{aligned}
\sum_{i=1}^{n} x_i y_i - (\overline{y} - \widehat{\beta}\overline{x}) \sum_{i=1}^{n} x_i - \widehat{\beta} \sum_{i=1}^{n} x_i^2 &= 0 \\
\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y} + \widehat{\beta} n \overline{x}^2 - \widehat{\beta} \sum_{i=1}^{n} x_i^2 &= 0 \\
\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y} - \widehat{\beta}\Big(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2 \Big) &= 0 \\
\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) - \widehat{\beta} \sum_{i=1}^{n} (x_i - \overline{x})^2 &= 0.
\end{aligned}$$

Lo que lleva a

$$\widehat{\beta} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} = \frac{\mathsf{cov}(\boldsymbol{x}, \boldsymbol{y})}{\mathrm{var}(\boldsymbol{x})}.$$

De este modo, la *recta de regresión* asume la forma:

$$\widehat{y}_i = \widehat{\alpha} + \widehat{\beta}\, x_i, \qquad i = 1, \ldots, n,$$

y llamamos a $\widehat{y}_i$ el valor *"predicho"* o *"ajuste"* para x_i. Se define

$$e_i = y_i - \widehat{y}_i = y_i - \widehat{\alpha} - \widehat{\beta} x_i, \qquad i = 1, \ldots, n,$$

como el i-ésimo *residuo*, mientras que una medida de variabilidad es dada por:

$$s^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \widehat{\alpha} - \widehat{\beta} x_i)^2 = \frac{1}{n-2}\, S(\widehat{\boldsymbol{\theta}}),$$

con $\widehat{\boldsymbol{\theta}} = (\widehat{\alpha}, \widehat{\beta})^\top$. Debemos resaltar que, el promedio de los valores predichos $\widehat{y}_1, \ldots, \widehat{y}_n$, asume la forma

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^{n} \widehat{y}_i &= \frac{1}{n} \sum_{i=1}^{n} (\widehat{\alpha} + \widehat{\beta} x_i) = \widehat{\alpha} + \widehat{\beta}\, \overline{x} \\
&= (\overline{y} - \widehat{\beta}\, \overline{x}) + \widehat{\beta}\, \overline{x} = \overline{y}.
\end{aligned}$$

Haciendo $R = \mathsf{cor}(\boldsymbol{x}, \boldsymbol{y})$, sigue que

$$\widehat{y}_i - \overline{y} = \widehat{\alpha} + \widehat{\beta} x_i - \overline{y} = \widehat{\beta}(x_i - \overline{x}) = R\, \frac{s_y}{s_x}\, (x_i - \overline{x}),$$

es decir,

$$(\widehat{y}_i - \overline{y})^2 = R^2 \frac{s_y^2}{s_x^2}\, (x_i - \overline{x})^2.$$

Lo que nos lleva a escribir,

$$\sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2 = R^2 \frac{s_y^2}{s_x^2} \sum_{i=1}^{n} (x_i - \overline{x})^2 = R^2 \frac{s_y^2}{s_x^2}\, s_x^2 = R^2 s_y^2 = R^2 \sum_{i=1}^{n} (y_i - \overline{y})^2.$$

A partir de esta ecuación, sigue que

$$R^2 = \frac{\sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} = \frac{s^2_{\mathsf{AJUSTE}}}{s^2_{\mathsf{DATOS}}},$$

que se denomina *coeficiente de determinación*. Usando además que

$$y_i - \overline{y} = (y_i - \widehat{y}_i) + (\widehat{y}_i - \overline{y}),$$

podemos mostrar que

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 + \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2,$$

es decir,

$$s^2_{\mathsf{DATOS}} = s^2_{\mathsf{RESIDUOS}} + s^2_{\mathsf{AJUSTE}}.$$

Como $s^2_{\mathsf{RESIDUOS}} = S(\widehat{\boldsymbol{\theta}})$, tenemos que $s^2_{\mathsf{AJUSTE}} = R^2 s^2_{\mathsf{DATOS}}$. Luego, sigue que:

$$s^2_{\mathsf{RESIDUOS}} = (1 - R^2)s^2_{\mathsf{DATOS}}.$$

Esto permite interpretar R^2 como la proporción de varianza de los datos que puede ser explicada por el ajuste. En efecto, $0 \leq R^2 \leq 1$ permite medir la calidad (o bondad) del ajuste.

EJEMPLO 1.62. Debemos ser precavidos al llevar un análisis de regresión y confiar de medidas tales como el R^2 para evaluar la calidad del ajuste. Para notar este tipo de situaciones, considere los datos introducidos por Anscombe (1973). Este corresponde a cuatro conjunto de datos con 11 observaciones que aunque son bastante diferentes tienen estadísticas de resumen idénticas (promedio, varianza, correlación, coeficientes de regresión, etc.). En efecto, considere la siguiente figura

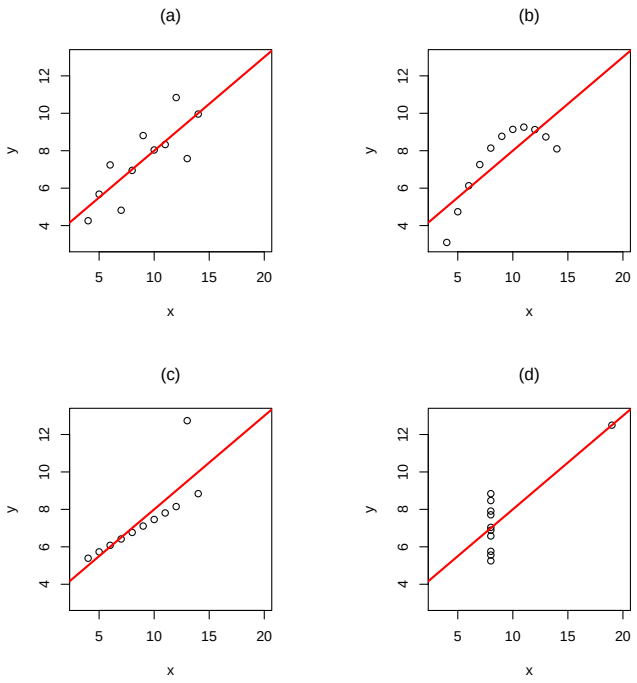


Figura 4. Cuarteto de regresiones idénticas de Anscombe (1973).

Por ejemplo, en la Figura 4 (b), los datos presentan una tendencia cuadrática, mientras que en los Paneles (c) y (d) se aprecia el efecto de observaciones atípicas sobre la recta de regresión. Más aún, eliminando la observación extrema en el conjunto de datos en la Figura 4 (d), la regresión deja de tener sentido. Para obtener los resultados del ajuste mediante regresión por mínimos cuadrados podemos considerar los siguientes comandos de R

```
  # datos de Anscombe
  > anscombe
   x1 x2 x3 x4    y1   y2    y3    y4
1  10 10 10  8  8.04 9.14  7.46  6.58
2   8  8  8  8  6.95 8.14  6.77  5.76
3  13 13 13  8  7.58 8.74 12.74  7.71
...
10  7  7  7  8  4.82 7.26  6.42  7.91
11  5  5  5  8  5.68 4.74  5.73  6.89

  # Ajuste mediante OLS (para cada uno de los modelos)
  > fm1 <- lm(y1 ~ x1, data = anscombe)
  > fm2 <- lm(y2 ~ x2, data = anscombe)
  > fm3 <- lm(y3 ~ x3, data = anscombe)
  > fm4 <- lm(y4 ~ x4, data = anscombe)

  # Figura 4.a (otros paneles son análogos)
  > par(pty = "s")
  > plot(y1 ~ x1, data = anscombe, xlim = c(4,20),
  +      ylim = c(3,13), xlab = "x", ylab = "y")
  > abline(coef(fm1), col = "red", lwd = 2) # línea en rojo

  # Salida de resultados:
  > summary(fm1)
  Call:
  lm(formula = y1 ~ x1, data = anscombe)

  Residuals:
       Min       1Q   Median       3Q      Max
  -1.92127 -0.45577 -0.04136  0.70941  1.83882

  Coefficients:
              Estimate Std. Error t value Pr(>|t|)
  (Intercept)   3.0001     1.1247   2.667  0.02573 *
  x1            0.5001     0.1179   4.241  0.00217 **

  Residual standard error: 1.237 on 9 degrees of freedom
  Multiple R-squared: 0.6665, Adjusted R-squared: 0.6295
  F-statistic: 17.99 on 1 and 9 DF, p-value: 0.00217

  # Almacena coeficientes, residuos y valores predichos
  > cf <- coef(fm1)
  > res <- resid(fm1)
  > fit <- fitted(lm1)
```

```
# Figura 5.a (otros paneles son análogos)
> par(pty = "s")
> plot(fit, res, xlab="Valores predichos", ylab="Residuos")
> abline(h = 0, lty = 2, col = "gray", lwd = 2) # línea gris
```

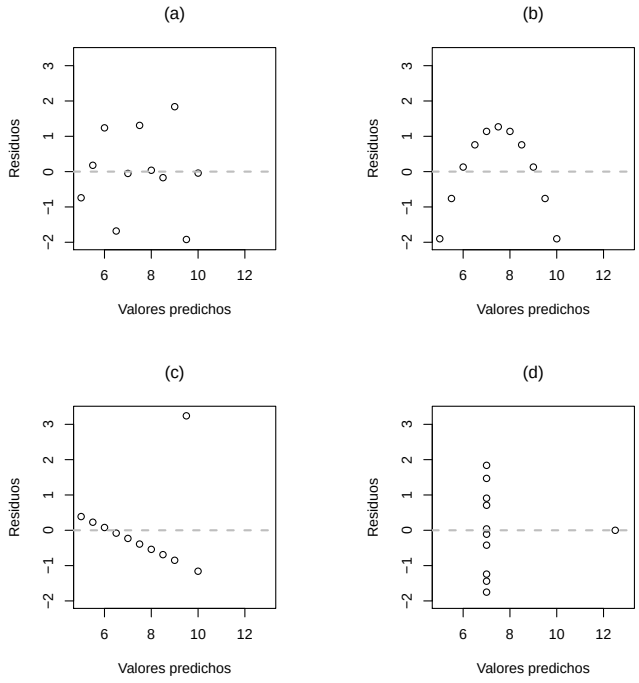


Figura 5. Gráfico de residuos para el cuarteto de regresiones idénticas de Anscombe (1973).

Para el modelo en Ecuación (1.28), es posible mostrar que:

$$\sum_{i=1}^{n} e_i = 0, \qquad \text{y} \qquad \sum_{i=1}^{n} e_i \widehat{y}_i = 0.$$

De este modo, el grafico de dispersión de residuos contra valores predichos dado en la Figura 5 no debería presentar algún tipo de comportamiento sistemático. Es decir, el único conjunto de datos donde el modelo de regresión lineal parece razonable es el desplegado en el Panel (a) de la Figura 4. En efecto, el gráfico de residuos en la Figura 5 (a) es el único que no presenta algún tipo de tendencia. En cambio, se debe incluir un término cuadrático al modelo de la Figura 4 (b). Mientras que el rol de las observaciones atípicas visibles en los Paneles (c) y (d) de la Figura 4 se aprecia claramente en los gráficos de residuos. Es decir, corresponden a un outlier (observación alejada en la 'respuesta') y un leverage (observación alejada en los 'regresores'), respectivamente.

Debemos destacar que para este ejemplo sencillo es bastante fácil determinar que el modelo lineal no es apropiado para las configuraciones de datos en los Paneles

(b), (c) y (d) presentados en la Figura 4. En general, no resulta fácil identificar situaciones donde se aprecie que el modelo propuesto está mal ajustado. Este tipo de situaciones es más difícil de determinar conforme se tiene de un mayor volumen de datos, o bien la cantidad de regresores aumenta.[6]

1.6. Resúmenes gráficos

Para introducir ideas consideremos una imagen digital de tamaño $r \times c$. Una imagen digital es un arreglo de tamaño $r \times c$ que puede ser conceptualizada como una matriz de datos $\boldsymbol{X} = (X_{ij})$, para $i = 1, \ldots, r$, $j = 1, \ldots, c$, donde el valor de la variable X en la posición (i, j) corresponde a un valor en una escala de grises en el intervalo $[0, 1]$ tal que 0 representa el color negro y 1 representa el blanco. De este modo, $\{X_{ij} : 1 \leq i \leq r, 1 \leq j \leq c\}$ representa el conjunto de intensidades de gris en cada una de las $r \times c$ posiciones en el espacio bidimensional. Cada posición (i, j) es denominada un *pixel* (picture element). La Figura 6 presenta una imagen muy popular en ingeniería, específicamente en el área de tratamiento de imágenes, llamada frecuentemente *Lenna* y fue popularizada hace varias décadas atrás.[7]

Figura 6. Imagen *Lenna* de tamaño 512×512.

Podemos considerar una imagen como un vector $\boldsymbol{x} = \text{vec}(\boldsymbol{X})$ en el que hemos concatenado las columnas de $\boldsymbol{X}$. Para la imagen Lenna tenemos $n = 512 \times 512 = 262\,144$ observaciones. Un resumen para el conjunto de datos *Lenna* es obtenido usando los siguientes comandos en R:

```
# carga datos de 'Lenna' desde URL
> lena <- "http://fosorios.mat.utfsm.cl/files/data/lena.png"
> download.file(lena, z, mode = "wb")
> library(png) # biblioteca para cargar imágenes PNG
> lena <- readPNG(z)[,,1] # sólo el 1er canal es necesario
```

[6] En la asignatura *Análisis de Regresión* se revisará el aspecto de la crítica del modelo con mayor profundidad.

[7] Para más detalles sobre esta imagen, consultar la página web: http://www.lenna.org/

```
# carga biblioteca 'SpatialPack'
> library(SpatialPack)
> lena <- normalize(lena)
> plot(as.raster(lena)) # Figura 6

# convierte en un vector de datos
> x <- as.vector(lena)
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.2712  0.4443  0.4316  0.5809  1.0000

# cuantiles más utilizados
> qx <- quantile(x)
> qx
      0%       25%       50%       75%      100%
0.0000000 0.2712455 0.4442928 0.5808888 1.0000000

> qx <- as.vector(qx)[-c(1,5)] # remueve extremos
> IQR <- qx[3] - qx[1]
> IQR # rango intercuartílico
[1] 0.3096433
```

Podemos notar que, debido a la normalización de los datos, el valor mínimo y máximo corresponden a 0 y 1, respectivamente, mientras que el rango intercuartílico, esto es, la distancia en la que se concentra el 50 % central de la muestra resulta $IQR = Q_3 - Q_1 = 0.5808 - 0.2712 = 0.3096$.

A continuación se presenta dos herramientas gráficas para una muestra $x_1, x_2, \ldots, x_n$ que permiten visualizar de forma simple la distribución de una variable de interés X, conocidas como *histograma* y *boxplot* (o diagrama de cajón-con-bigotes).

1.6.1. Histograma. Un histograma es una aproximación a la densidad desconocida construído a partir de los datos observados. Considere la siguiente definición.

DEFINICIÓN 1.63. Sea $x_1, x_2, \ldots, x_n$ una muestra de n observaciones y suponga que los datos son subdivididos en k intervalos, digamos $I_1, I_2, \ldots, I_k$. Entonces, la frecuencia absoluta de la clase I_j es la cantidad de observaciones n_j de la muestra, que pertenecen al intervalo I_j. Evidentemente tenemos que:

$$n_j \geq 0, \qquad \text{y} \qquad \sum_{j=1}^{k} n_j = n.$$

La frecuencia relativa asociada a la clase I_j se define como:

$$f_j = \frac{n_j}{n}, \qquad j = 1, \ldots, k.$$

En este caso es fácil ver que $f_j \geq 0$ y $\sum_{j=1}^{k} f_j = 1$. De este modo, un *histograma* es un gráfico de f_j (o n_j) versus I_j.

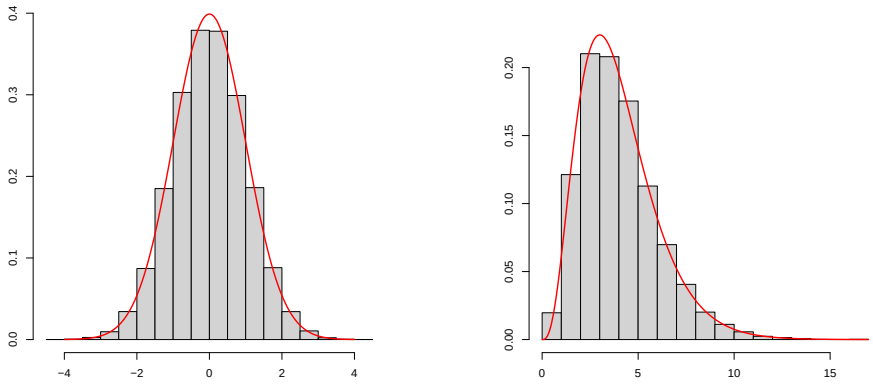
Un histograma es un diagrama de frecuencias y resume la cantidad de observaciones por unidad de longitud. Luego, este diagrama permite visualizar la distribución de la variable de interés, tal como se presenta en la Figura 7. En este caso, ambos

histogramas tienen 14 clases. Este corresponde a un parámetro gráfico que puede modificarse para adaptarse a la estructura de la muestra que se tiene disponible. En algunos casos se grafica f_j/h versus I_j, donde h es la amplitud de los intervalos I_j. Este gráfico tiene la particularidad de que el área bajo la curva es dada por

$$A = \sum_{j=1}^{k} A_j = \sum_{j=1}^{k} (f_j/h) \cdot h = \sum_{j=1}^{k} f_j = 1.$$

En efecto, resultará claro más adelante que este propiedad está asociada al concepto de probabilidad.

Debemos enfatizar que un histograma es una herramienta descriptiva. Más adelante estudiaremos algunas curvas llamadas *funciones de densidad de probabilidad* asociada a ciertas variables de interés. Sin embargo, histogramas han sido criticados por tener algunas inconvenientes notables. Además de la necesidad de escoger el número de intervalos k (o equivalentemente la amplitud de los intervalos h), frecuentemente pueden resultar engañosos si se interpretan en exceso.



(a) (b)

Figura 7. Histograma y densidad teórica para una muestra de 10 000 observaciones provenientes desde una distribución (a) normal y (b) gama.

1.6.2. Boxplot (Diagrama de cajón con bigotes). Este tipo de diagrama corresponde a un tipo diferente de visualización que nos permite explorar la ubicación, la escala, la asimetría y las colas de una densidad así como la presencia de observaciones atípicas (outliers). En contraste con el histograma el boxplot ofrece una descripción mucho más gruesa de la estructura de la muestra y no requiere la especificación de parámetros para la calibración del gráfico. Esta representación usualmente se realiza en forma de una caja, lo que explica el nombre del gráfico. A continuación se presenta el procedimiento para su construcción.

DEFINICIÓN 1.64. Considere $x_1, x_2, \ldots, x_n$ una muestra de n observaciones y sea:

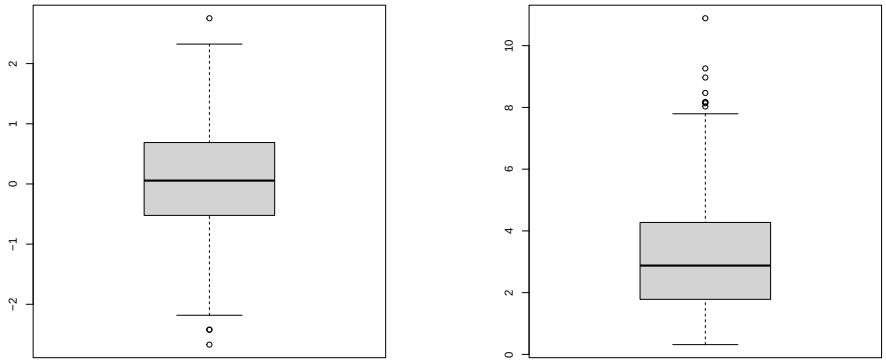
- Q_1, me y Q_3 respectivamente, el primer cuartil, la mediana y el tercer cuartil de $\{x_1, \ldots, x_n\}$.
- Cálculo de los "bigotes" W_1 y W_2, como:

$$W_1 = \min_{1 \leq j \leq n} \{x_j : x_j \geq Q_1 - 1.5IQR\}$$
$$W_2 = \max_{1 \leq j \leq n} \{x_j : x_j \leq Q_3 + 1.5IQR\}$$

- $O = \{i \in \{1, 2, \ldots, n\} : x_i \notin [W_1, W_2]\}$.

De este modo, el *boxplot* de $x_1, \ldots, x_n$ es el registro de $W_1, Q_1, \mathsf{me}, Q_3, W_2$ y $\{x_j : j \in O\}$ sobre la recta real.

Aunque la definición puede ser un tanto difícil de visualizar, los gráficos en la Figura 8 permiten comprender la idea. La línea central corresponde a la mediana, mientras que los extremos del cajón representan Q_1 y Q_3. Note que W_1 y W_2 representan los extremos del bigote y las observaciones atípicas son indicados por puntos. En el Panel (a) tenemos una distribución simétrica en torno de cero, mientras que la distribución en el Panel (b) presenta asimetría positiva.



(a) (b)

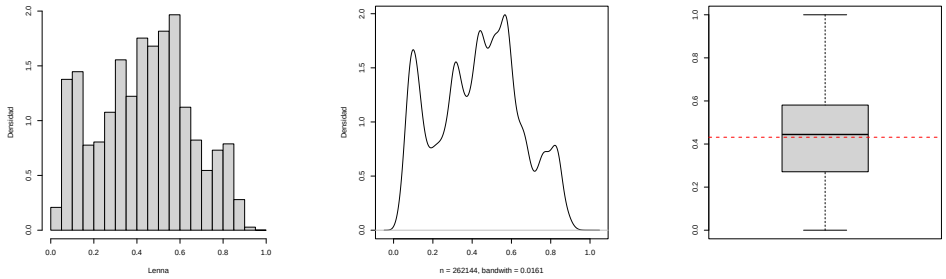
Figura 8. Boxplot para una muestra de 300 observaciones provenientes desde una distribución (a) normal y (b) gama.

Un aspecto destacable del boxplot, es que permite la comparación de varias muestras permitiendo la visualización simultánea de medidas de tendencia central y dispersión mediante construir una secuencia de gráficos de cajón-con-bigotes.

Retomando el conjunto de datos de la imagen Lenna, el siguiente código en R, permite obtener el histograma, la curva de densidad estimada y el boxplot:

```
# histograma y densidad estimada (Fig. 9)
> hist(x, freq = FALSE, main = "", xlab = "Lenna",
+      ylab = "Densidad")
```

```
> plot(density(x), main = "", xlab = "bandwith = 0.0161",
+      ylab = "Densidad")
> boxplot(x)
> xbar <- mean(x) # promedio
> abline(h = xbar, lty =2, col = "red") # línea segmentada
```



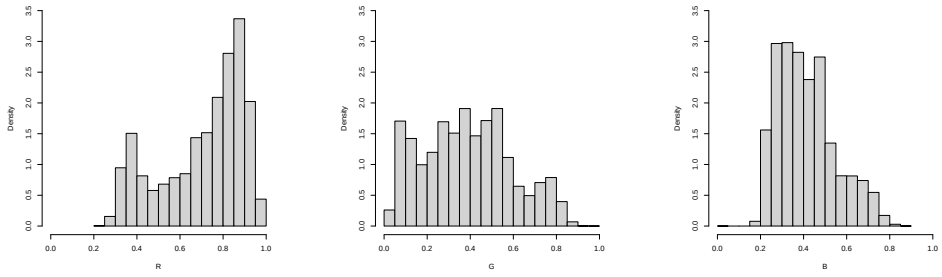
(a) (b) (c)

Figura 9. Histograma, densidad estimada y boxplot para los datos de la imagen Lenna.

Es interesante destacar que imágenes en color se suelen representar como un arreglo con tres canales, usualmente rojo (R), verde (G) y azul (B), lo que son combinados de manera apropiada para producir una representación a color. Considere la imagen Lenna en formato a color,

Figura 10. Imagen Lenna en versión a color.

De este modo, podemos considerar que las intensidades de rojo, verde y azul, representan 3 conjuntos de datos. Con fines comparativos, Figuras 11 y 12 presentan histogramas y boxplots para cada uno de los canales RGB de la imagen Lenna.



(a) canal rojo (R)

(b) canal verde (G)

(c) canal azul (B)

Figura 11. Histogramas para cada uno de los canales RGB de los datos de la imagen Lenna.

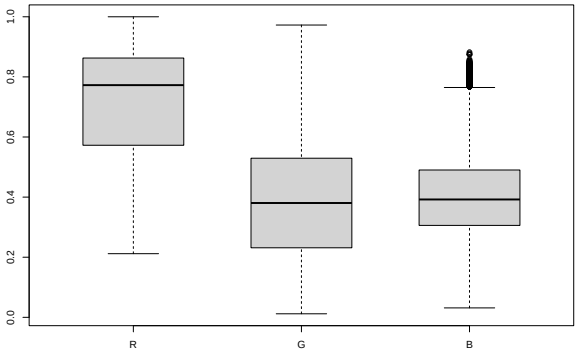


Figura 12. Boxplots para cada uno de los canales RGB de los datos de la imagen Lenna.

Es notable como los boxplots en la Figura 12 nos permite hacer una comparación rápida de la configuración de cada uno de los canales RGB.

El campo de visualización es tópico amplio con desafíos muy relevantes dado la demanda por resumir grandes volúmenes de información usando un despliegue gráfico. Para mayores detalles sobre visualización así como por técnicas de análisis exploratorio consultar Tukey (1977), Cleveland (1993) y Wilkinson (1999).

Capítulo 2

Nociones de Probabilidad

2.1. Preliminares

DEFINICIÓN 2.1. El conjunto Ω, de todos los resultados posibles de un experimento aleatorio es llamado *espacio muestral*.

EJEMPLO 2.2. Suponga que lanzamos de una moneda. De este modo, tenemos sólo dos resultados posibles

$$\Omega = \{C, S\}.$$

EJEMPLO 2.3. Para el caso de las notas de la asignatura MAT-021 para un grupo de estudiantes escogidos al azar, sigue que

$$\Omega = \{0, 1, \ldots, 99, 100\}.$$

EJEMPLO 2.4. El espacio muestral para el tiempo de duración de un artículo es dado por:

$$\Omega = [0, \infty).$$

DEFINICIÓN 2.5. Un evento (o suceso) es cualquier colección de resultados posibles de un experimento aleatorio, esto es cualquier subconjunto de Ω (incluyendo al propio Ω).

OBSERVACIÓN 2.6. Sea $A \subset \Omega$, diremos que A ocurre si $\omega \in A$ con $\omega \in \Omega$ es un resultado asociado a un experimento aleatorio. Evidentemente, $\omega \notin A$ si y sólo si A no ocurre.

A continuación describimos algunas operaciones sobre conjuntos que usaremos frecuentemente:

$$A \subset B \iff x \in A \Longrightarrow x \in B.$$
$$A = B \iff A \subset B \text{ y } B \subset A.$$

Unión: La unión entre A y B denotada por $A \cup B$ es definida como:

$$A \cup B = \{x : x \in A, \text{ o } x \in B\}.$$

Intersección: La intersección entre A y B escrita como $A \cap B$ se define como:

$$A \cap B = \{x : x \in A, \text{ y } x \in B\}.$$

Complemento: El complemento de A, denotado como A^c es el conjunto de todos los elementos que *no* están en A:

$$A^c = \{x : x \notin A\}.$$

PROPIEDAD 2.7. Sean A, B y C tres eventos definidos en Ω. Tenemos:

(a) $A \cup B = B \cup A$, $A \cap B = B \cap A$.

(b) $A \cup (B \cup C) = (A \cup B) \cup C$,
$A \cap (B \cap C) = (A \cap B) \cap C$.
(c) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$,
$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.
(d) $(A \cup B)^c = A^c \cap B^c$, y $(A \cap B)^c = A^c \cup B^c$.

La Propiedad 2.7 (d) es conocida frecuentemente como *Leyes de De Morgan*. La demostración de cada una de estas propiedades se dejan como ejercicio para el lector.

Note que las operaciones de unión e intersección se pueden extender a colecciones infinitas de conjuntos,

$$\bigcup_{i=1}^{\infty} A_i = \{x \in \Omega : x \in A_i, \text{para algún } i\}.$$
$$\bigcap_{i=1}^{\infty} A_i = \{x \in \Omega : x \in A_i, \text{para todo } i\}.$$

DEFINICIÓN 2.8. Dos eventos A y B son disjuntos (o excluyentes) si

$$A \cap B = \varnothing.$$

Los eventos $A_1, A_2, \ldots$, son disjuntos por pares (o mutuamente excluyentes) si

$$A_i \cap A_j = \varnothing, \quad i \neq j.$$

DEFINICIÓN 2.9. Si $A_1, A_2, \ldots$ son disjuntos por pares y $\cup_{i=1}^{\infty} = \Omega$, entonces la colección se denomina una *partición* de Ω.

EJEMPLO 2.10. Los conjuntos $A_i = [i, i+1)$, para $i = 0, 1, 2, \ldots$ forman una partición de $[0, \infty)$.

EJEMPLO 2.11. $\Omega = A \cap A^c$ es una partición.

DEFINICIÓN 2.12. Sea A, B dos sucesos, diremos que A implica B si y sólo si $A \subseteq B$.

2.2. Conceptos básicos

Para todo evento $A \subset \Omega$ deseamos asociar un número entre cero y uno llamado probabilidad de A.

DEFINICIÓN 2.13. Una colección de subconjuntos de Ω es llamado σ-álgebra y es denotada por $\mathcal{F}$ si satisface las propiedades:

(a) $\varnothing \in \mathcal{F}$.
(b) Si $A \in \mathcal{F} \Longrightarrow A^c \in \mathcal{F}$.
(c) Si $A_1, A_2, \cdots \in \mathcal{F} \Longrightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$

Note que $\varnothing \subset \Omega$ y $\Omega = \varnothing^c$, así por la Propiedad (a) y (b) sigue que $\Omega \in \mathcal{F}$.

Además, si $A_1, A_2, \cdots \in \mathcal{F}$ entonces $A_1^c, A_2^c, \cdots \in \mathcal{F}$ y de este modo, $\bigcup_{i=1}^{\infty} A_i^c \in \mathcal{F}$. Por las leyes de De Morgan, tenemos

$$\Big(\bigcup_{i=1}^{\infty} A_i^c\Big)^c = \bigcap_{i=1}^{\infty} A_i.$$

Es decir, tenemos que $\bigcap_{i=1}^{\infty} A_i \in \mathcal{F}$.

Observación 2.14. Asociado a un espacio muestral Ω puede haber muchas σ-álgebras. Por ejemplo, la colección $\{\varnothing, \Omega\}$ es σ-álgebra (minimal).

Definición 2.15. Sea Ω un espacio muestral y sea $\mathcal{F}$ un σ-álgebra, decimos que el par $(\Omega, \mathcal{F})$ es un espacio medible. Si $A \in \mathcal{F}$ decimos que A es medible.

Definición 2.16 (Probabilidad). Dado un espacio muestral Ω y un σ-álgebra asociada $\mathcal{F}$, una función de probabilidad P, $\mathrm{P} : \mathcal{F} \to \mathbb{R}$, satisface:

(a) $\mathrm{P}(A) \geq 0$, para todo $A \in \mathcal{F}$.
(b) $\mathrm{P}(\Omega) = 1$.
(c) Si $A_1, A_2, \ldots$ son disjuntos por pares (mutuamente excluyentes), entonces

$$\mathrm{P}\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} \mathrm{P}(A_i).$$

Sea $(\Omega, \mathcal{F})$ un espacio de medida y P una medida de probabilidad definida en $\mathcal{F}$. Entonces $(\Omega, \mathcal{F}, \mathrm{P})$ se denomina *espacio de probabilidad*.

Resultado 2.17. *Si* P *es una función de probabilidad y A es cualquier conjunto en $\mathcal{F}$, entonces*

(a) $\mathrm{P}(\varnothing) = 0$.
(b) $\mathrm{P}(A) \leq 1$.
(c) $\mathrm{P}(A^c) = 1 - \mathrm{P}(A)$.

Demostración. Primero considere (c). Como A y A^c son una partición de Ω, sigue que

$$\mathrm{P}(A \cup A^c) = \mathrm{P}(\Omega) = 1,$$

además $A \cap A^c = \varnothing$ son disjuntos, luego

$$\mathrm{P}(A \cup A^c) = \mathrm{P}(A) + \mathrm{P}(A^c) = 1,$$

lo que permite mostrar el resultado. Como $\mathrm{P}(A^c) \geq 0$, (b) sigue desde (c). Finalmente, para probar (a) note que $\Omega = \Omega \cup \varnothing$ y como $\Omega \cap \varnothing = \varnothing$, tenemos

$$1 = \mathrm{P}(\Omega) = \mathrm{P}(\Omega \cup \varnothing) = \mathrm{P}(\Omega) + \mathrm{P}(\varnothing),$$

de este modo $\mathrm{P}(\varnothing) = 0$. □

Resultado 2.18. *Si* P *es una función de probabilidad y $A, B \in \mathcal{F}$. Entonces,*

(a) $\mathrm{P}(B \cap A^c) = \mathrm{P}(B) - \mathrm{P}(A \cap B)$.
(b) $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$.
(c) *Si* $A \subseteq B \Longrightarrow \mathrm{P}(A) \leq \mathrm{P}(B)$.

Demostración. Para notar (a) considere que para A y B conjuntos cualquiera

$$B = (B \cap A) \cup (B \cap A^c)$$

(pues $B = B \cup (A \cap A^c) = B \cup \varnothing$). Luego,

$$\mathrm{P}(B) = \mathrm{P}(\{B \cap A\} \cup \{B \cap A^c\}) = \mathrm{P}(B \cap A) + \mathrm{P}(B \cap A^c).$$

En efecto, $(B \cap A) \cap (B \cap A^c) = B \cap (A \cap A^c) \cap B = \varnothing$. Para probar (b), note que

$$(A \cup B) = A \cup (B \cap A^c).$$

Además,

$$A \cap (B \cap A^c) = (A \cap A^c) \cap B = \varnothing \cap B = \varnothing.$$

Por tanto, sigue que

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B \cap A^c) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B),$$

por parte (a).

Si $A \subseteq B$ entonces $A \cap B = A$. De este modo, usando (a) tenemos

$$0 \leq \mathrm{P}(B \cap A^c) = \mathrm{P}(B) - \mathrm{P}(A),$$

y (c) es verificado. □

Como $\mathrm{P}(A \cup B) \leq 1$, reagrupando términos tenemos

$$\mathrm{P}(A \cap B) \geq \mathrm{P}(A) + \mathrm{P}(B) - 1,$$

la desigualdad anterior es un caso particular de la desigualdad de Bonferroni.

RESULTADO 2.19. *Si* P *es una función de probabilidad. Entonces,*

(a) $\mathrm{P}(A) = \sum_{i=1}^{\infty} \mathrm{P}(A \cap C_i)$ *para cualquier partición* $C_1, C_2, \ldots$.
(b) $\mathrm{P}(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} \mathrm{P}(A_i)$ *para conjuntos* $A_1, A_2, \ldots$ *cualquiera.*[1]

DEMOSTRACIÓN. Dado que $C_1, C_2, \ldots$ forman una partición, tenemos $C_i \cap C_j = \varnothing$ para todo $i \neq j$ y $\Omega = \bigcup_{i=1}^{\infty} C_i$. De ahí que

$$A = A \cap \Omega = A \cap \Big(\bigcup_{i=1}^{\infty} C_i \Big) = \bigcup_{i=1}^{\infty} (A \cap C_i).$$

De este modo,

$$\mathrm{P}(A) = \mathrm{P}\Big(\bigcup_{i=1}^{\infty} (A \cap C_i) \Big) = \sum_{i=1}^{\infty} \mathrm{P}(A \cap C_i),$$

pues, dado que los C_i's son disjuntos, tambien lo es la secuencia $\{A \cap C_i\}_{i=1}^{\infty}$.

Para establecer (b) se construye una colección disjunta $A_1^*, A_2^*, \ldots$ tal que

$$\bigcup_{i=1}^{\infty} A_i^* = \bigcup_{i=1}^{\infty} A_i.$$

Defina A_i^*, como

$$A_1^* = A_1,$$
$$A_i^* = A_i - \Big(\bigcup_{j=1}^{i-1} A_j \Big), \quad i = 2, 3, \ldots,$$

donde $A - B = A \cap B^c$ denota la diferencia entre conjuntos. Podemos notar que $\bigcup_{i=1}^{\infty} A_i^* = \bigcup_{i=1}^{\infty} A_i$, luego

$$\mathrm{P}\Big(\bigcup_{i=1}^{\infty} A_i \Big) = \mathrm{P}\Big(\bigcup_{i=1}^{\infty} A_i^* \Big) = \sum_{i=1}^{\infty} \mathrm{P}(A_i^*),$$

[1] Esta es conocida como la *Desigualdad de Boole.*

pues los A_i^*'s son disjuntos. En efecto,

$$\begin{aligned} A_i^* \cap A_k^* &= \Big\{A_i - \Big(\bigcup_{j=1}^{i-1} A_j\Big)\Big\} \cap \Big\{A_k - \Big(\bigcup_{j=1}^{k-1} A_j\Big)\Big\} \\ &= \Big\{A_i \cap \Big(\bigcup_{j=1}^{i-1} A_j\Big)^c\Big\} \cap \Big\{A_k \cap \Big(\bigcup_{j=1}^{k-1} A_j\Big)^c\Big\} \\ &= \Big\{A_i \cap \Big(\bigcap_{j=1}^{i-1} A_j^c\Big)\Big\} \cap \Big\{A_k \cap \Big(\bigcap_{j=1}^{k-1} A_j^c\Big)\Big\}, \end{aligned}$$

si $i > k$ entonces la primera intersección está contenida en A_k^c, luego esa intersección será vacia. Si $i < k$ el argumento es similar. Además, por construcción $A_i^* \subset A_i$ de modo que $\mathrm{P}(A_i^*) \leq \mathrm{P}(A_i)$. Por tanto,

$$\sum_{i=1}^{\infty} \mathrm{P}(A_i^*) \leq \sum_{i=1}^{\infty} \mathrm{P}(A_i),$$

estableciendo el resultado. □

Observación 2.20. Usando la desigualdad de Boole, tenemos

$$\mathrm{P}\Big(\bigcup_{i=1}^{n} A_i^c\Big) \leq \sum_{i=1}^{n} \mathrm{P}(A_i^c),$$

y como $\cup A_i^c = (\cap A_i)^c$ y $\mathrm{P}(A_i^c) = 1 - \mathrm{P}(A_i)$, tenemos

$$1 - \mathrm{P}\Big(\bigcap_{i=1}^{n} A_i\Big) \leq n - \sum_{i=1}^{n} \mathrm{P}(A_i),$$

es decir

$$\mathrm{P}\Big(\bigcap_{i=1}^{n} A_i\Big) \geq \sum_{i=1}^{n} \mathrm{P}(A_i) - (n-1),$$

que es conocida como la desigualdad de Bonferroni.

2.3. Espacios muestrales finitos

En esta sección consideraremos que

$$\Omega = \{\omega_1, \omega_2, \ldots, \omega_k\},$$

para caracterizar $\mathrm{P}(A)$ supondremos eventos elementales, es decir $A = \{\omega_i\}$ y definimos $p_i = \mathrm{P}(\{\omega_i\})$ la probabilidad de $\{\omega_i\}$ tal que,

(a) $p_i \geq 0$, $i = 1, \ldots, k$.
(b) $p_1 + p_2 + \cdots + p_k = 1$.

Suponga ahora que $A = \{\omega_{j_1}, \ldots, \omega_{j_r}\}$ está formado por r elementos de Ω, luego

$$\mathrm{P}(A) = p_{j_1} + \cdots + p_{j_r}.$$

Adicionalmente, supondremos que cada $\{\omega_i\}$ es igualmente probable. Entonces,

$$p_i = \mathrm{P}(\{\omega_i\}) = \frac{1}{k}.$$

Luego, para un evento $A = \{\omega_{j_1}, \ldots, \omega_{j_r}\}$ sigue que

$$\mathrm{P}(A) = \frac{r}{k},$$

o bien,

$$\mathrm{P}(A) = \frac{\#(A)}{\#(\Omega)},$$

donde $\#(A)$ denota la cardinalidad del conjunto A. Debemos resaltar que esta *no* es una definición general, sino que apropiada *sólo* bajo el supuesto de espacios muestrales finitos y equiprobables.

Ejemplo 2.21. Considere un lote de 100 artículos con 20 defectuosos y 80 no defectuosos. Se eligen 10 artículos al azar, sin substituir un artículo antes de elegir el próximo. ¿Cuál es la probabilidad de que exactamente la mitad de los artículos escogidos sean defectuosos? Note que el espacio muestral consta de todos los vectores $(i_1, i_2, \ldots, i_{10})$ extraídos desde el lote de 100 artículos ¿Cómo contar cuántos son? ¿Cuáles satisfacen la condición que define $\mathrm{P}(A)$?

2.4. Técnicas de conteo

El objetivo es determinar el número de subconjuntos distintos que se pueden formar con un conjunto de elementos dados.

2.4.1. Principio multiplicativo. Suponga que se desarrolla un experimento y digamos que podemos subdividir este experimento en 2 etapas. Si la primera etapa puede desarrollarse de n_1 maneras y la segunda etapa en n_2 manera, entonces el experimento puede llevarse a cabo en

$$n_1 \cdot n_2,$$

maneras.

Observación 2.22. Esto puede extenderse a un experimento que se desarrolla en k etapas. Así, el experimento puede ser desarrollado en

$$n_1 \cdot n_2 \cdots n_k,$$

maneras.

2.4.2. Principio aditivo. Suponga que un experimento consta de dos etapas, tal que la primera se puede realizar de n_1 maneras y la segunda de n_2 maneras. Además considere que ambas etapas *no* se pueden realizar juntas. Entonces el experimento se puede llevar a cabo en

$$n_1 + n_2,$$

maneras.

Observación 2.23. Evidentemente podemos extender lo anterior para un experimento que se desarrolla en k etapas. En cuyo caso tenemos que hay

$$n_1 + n_2 + \cdots + n_k,$$

formas de desarrollar el experimento.

Definición 2.24. Sea $A \neq \varnothing$ un conjunto finito, tal que $\#(A) = n$. Se denomina *variación simple* de orden k ($k \leq n$) a todo subconjunto de n elementos distinguiendo los elementos que lo componen y el orden en que son ordenados

EJEMPLO 2.25. Sea $A = \{\omega_1, \omega_2, \ldots, \omega_n\}$. Para formar las variaciones de orden 2, digamos $(n)_2$ considere

$$\begin{pmatrix} \omega_1\omega_2 & \omega_1\omega_3 & \ldots & \omega_1\omega_n \\ \omega_2\omega_1 & \omega_2\omega_3 & \ldots & \omega_2\omega_n \\ \vdots & & & \\ \omega_n\omega_1 & \omega_n\omega_2 & \ldots & \omega_n\omega_{n-1} \end{pmatrix},$$

note que esta matriz tiene n filas y $n-1$ columnas. Así, tenemos un total de $n(n-1)$ elementos, es decir

$$(n)_2 = n(n-1).$$

En general, el numero total de variaciones de orden k es dada por:

$$(n)_k = n(n-1)\cdots(n-k+1) = \frac{n!}{(n-k)!}.$$

OBSERVACIÓN 2.26. Recuerde que, para un entero positivo n, tenemos

$$n! = n(n-1)\cdots 2\cdot 1 = n(n-1)!$$

Además $0! = 1$ por convención.

DEFINICIÓN 2.27. Sea $A \neq \varnothing$ un conjunto finito con $\#(A) = n$ se llama *variación con repetición* a toda variación en la que un mismo elemento puede aparecer n veces repetido.

Sea $A = \{\omega_1, \omega_2, \ldots, \omega_n\}$, cada variación simple de orden $k-1$ da origen a n variaciones de orden k, luego se debe resolver el sistema de ecuaciones

$$\begin{aligned} (n)_1^* &= n \\ (n)_2^* &= n(n)_1^* \\ &\vdots \\ (n)_k^* &= n(n)_{k-1}^*. \end{aligned}$$

Es decir,

$$(n)_k^* = \underbrace{n\cdot n\cdots n}_{k \text{ veces}} = n^k.$$

EJEMPLO 2.28. Considere un grupo de k personas. ¿Cuántas listas posibles se pueden realizar con los días de sus cumpleaños? En efecto, tenemos

$$365\cdot 365\cdots 365 = 365^k.$$

EJEMPLO 2.29 (Problema del cumpleaños). Suponga un grupo de k personas. Se desea calcular la probabilidad de que 2 personas cumplan años el mismo día. Entonces, tenemos

$$\frac{(365)_k}{365^k} = \frac{365\cdot 364\cdots(365-k+1)}{365^k},$$

y por la regla del complemento,

$$\mathrm{P}(\{\text{al menos un par de personas cumplan años el mismo día}\}) = 1 - \frac{(365)_k}{365^k}.$$

Es decir,

k	10	15	20	25	30	35	40	45	50
probabilidad	0.117	0.253	0.411	0.569	0.706	0.814	0.891	0.941	0.970

DEFINICIÓN 2.30 (Permutaciones). Sea $A \neq \varnothing$ un conjunto finito y $\#(A) = n$ se llama permutación a todo subconjunto de n elementos tales que las permutaciones se diferencian por el orden de sus elementos. Así,

$$P_{n,n} = (n)_n = n(n-1)\cdots 2 \cdot 1 = n!$$

OBSERVACIÓN 2.31. $P_{n,n}$ también se denominan permutaciones sin repetición

EJEMPLO 2.32. ¿De cuantas formas ubicar 5 personas en 5 sillas?

$$(5)_5 = 5! = 120$$

Cuando existen elementos iguales en A, las permutaciones se dicen *con* repetición. Considere,

$$A = \{\underbrace{\omega_1, \ldots, \omega_1}_{n_1 \text{ veces}}, \underbrace{\omega_2, \ldots, \omega_2}_{n_2 \text{ veces}}, \ldots, \underbrace{\omega_k, \ldots, \omega_k}_{n_k \text{ veces}}\},$$

tal que $\sum_{j=1}^{k} n_j = n$. Es decir, el número total de combinaciones es $n!$ Sea M el número de permutaciones distintas, entonces

$$M \prod_{j=1}^{k} n_j! = n! \qquad \Longrightarrow \qquad M = \frac{n!}{\prod_{j=1}^{k} n_j!}.$$

EJEMPLO 2.33. Calcular el número de permutaciones distintas con las cifras $\{4, 7, 3, 4, 7, 7, 3\}$. De este modo,

$$M = \frac{7!}{2!2!3!} = \frac{5040}{2 \cdot 2 \cdot 6} = 210.$$

DEFINICIÓN 2.34 (Combinaciones). Sea $A \neq \varnothing$ y $\#(A) = n$, se denomina *combinación* de k elementos a todo conjunto diferente de k elementos tomado desde n.

OBSERVACIÓN 2.35. Dos combinaciones se distinguen entre sí por la naturaleza de los elementos que la componen sin interesar el *orden*.

EJEMPLO 2.36. Considere los objetos $\{a, b, c, d\}$ y $k = 2$. Entonces tenemos,

$$ab, ac, ad, bc, bd, cd,$$

es decir 6 elementos diferentes (note que ab y ba sólo difieren en el orden). Por tanto,

$$k!\binom{n}{k} = (n)_k \qquad \Longrightarrow \qquad \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Algunas propiedades del combinatorio son:

$$\binom{n}{k} = \binom{n}{n-k}$$
$$\binom{n}{k} + \binom{n}{k+1} = \binom{n+1}{k+1}$$
$$\binom{n}{k} = \sum_{r=1}^{n-k-1} \binom{n-k}{k-1}.$$

Además, estos coeficientes aparecen en múltiples ocasiones. Por ejemplo,

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}.$$

2.5. Probabilidad condicional

EJEMPLO 2.37. Considere un lote con 80 artículos sin defectos y 20 defectuosos y suponga que se selecciona 2 artículos (a) con substitución, y (b) sin substitución. Defina los eventos:

$$A = \{\text{el 1er artículo es defectuoso}\},$$
$$B = \{\text{el 2do artículo es defectuoso}\}.$$

Cuando escogemos *con* substitución, tenemos:

$$\mathrm{P}(A) = \mathrm{P}(B) = \frac{20}{100} = \frac{1}{5}.$$

Cuando escogemos *sin* substitución, tenemos que:

$$\mathrm{P}(A) = \frac{20}{100} = \frac{1}{5},$$

pero, ¿Cambia $\mathrm{P}(B)$?

DEFINICIÓN 2.38. Si A y B son dos eventos en Ω y $\mathrm{P}(B) > 0$, entonces la *probabilidad condicional* de A dado B, escrito $\mathrm{P}(A|B)$ es

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}.$$

Note que $\mathrm{P}(B|B) = 1$, es decir, B *actua* como Ω. En efecto, como $A = A \cap \Omega$, tenemos

$$\mathrm{P}(A) = \mathrm{P}(A|\Omega) = \frac{\mathrm{P}(A \cap \Omega)}{\mathrm{P}(\Omega)}.$$

La ocurrencia de A es calibrada con relación a B. En particular, si $A \cap B = \varnothing$, entonces

$$\mathrm{P}(A|B) = \mathrm{P}(B|A) = 0.$$

En el ejemplo anterior, se desea calcular $\mathrm{P}(B|A) = 19/99$, pues si A ya ha ocurrido sólo quedan 19 defectuosos entre los 99 artículos.

Reexpresandola probabilidad condicional, tenemos

$$\mathrm{P}(A \cap B) = \mathrm{P}(A|B)\,\mathrm{P}(B),$$

o bien

$$\mathrm{P}(A \cap B) = \mathrm{P}(B|A)\,\mathrm{P}(A).$$

Las expresiones anteriores permiten "contornar" cálculos complicados, usando[2]

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(B|A)\,\mathrm{P}(A)}{\mathrm{P}(B)}.$$

OBSERVACIÓN 2.39. El espacio de probabilidad definido por $\mathcal{F} \cap B$ permite notar que $\mathrm{P}(A|B)$ es una función de probabilidad, es decir satisface:

(a) $\mathrm{P}(A|B) \geq 0$.
(b) $\mathrm{P}(\Omega|B) = 1$.
(c) Para $\{A_n\}_{n\geq 1}$ sucesión disjunta

$$\mathrm{P}\Big(\bigcup_{n=1}^{\infty} A_n \big| B\Big) = \sum_{n=1}^{\infty} \mathrm{P}(A_n|B).$$

[2]Esto es un caso particular del Teorema de Bayes.

RESULTADO 2.40 (Probabilidad total). *Sea* $(\Omega, \mathcal{F}, \mathrm{P})$ *un espacio de probabilidad y sea* $C_1, C_2, \ldots,$ *una partición medible y contable tal que* $\mathrm{P}(C_i) \geq 0$, $\forall i$. *Entonces, para todo* $A \in \mathcal{F}$,

$$\mathrm{P}(A) = \sum_{i=1}^{\infty} \mathrm{P}(A_i|C_i)\,\mathrm{P}(C_i).$$

DEMOSTRACIÓN. Como los C_i's forman una partición tenemos que

$$A = A \cap \Omega = A \cap \Big(\bigcup_{i=1}^{\infty} C_i\Big) = \bigcup_{i=1}^{\infty} (A \cap C_i).$$

Además,

$$\mathrm{P}(A) = \sum_{i=1}^{\infty} \mathrm{P}(A \cap C_i) = \sum_{i=1}^{\infty} \mathrm{P}(A|C_i)\,\mathrm{P}(C_i).$$

□

RESULTADO 2.41 (Teorema de Bayes). *Sea* $(\Omega, \mathcal{F}, \mathrm{P})$ *espacio de probabilidad y sea* $\{C_i\}$ *partición contable de* Ω *con* $\mathrm{P}(C_i) \geq 0$, $\forall i$. *Entonces, para todo* $A \in \mathcal{F}$, *tenemos que*

$$\mathrm{P}(C_i|A) = \frac{\mathrm{P}(A|C_i)\,\mathrm{P}(C_i)}{\sum_{k=1}^{\infty} \mathrm{P}(A|C_k)\,\mathrm{P}(C_k)},$$

siempre que $\mathrm{P}(A) > 0$.

DEMOSTRACIÓN. Tenemos que

$$\mathrm{P}(C_i|A)\,\mathrm{P}(A) = \mathrm{P}(A|C_i)\,\mathrm{P}(C_i),$$

así

$$\mathrm{P}(C_i|A) = \frac{\mathrm{P}(A|C_i)\,\mathrm{P}(C_i)}{\mathrm{P}(A)}, \qquad \mathrm{P}(A) > 0,$$

desde el Teorema de probabilidad total, sigue que

$$\mathrm{P}(C_i|A) = \frac{\mathrm{P}(A|C_i)\,\mathrm{P}(C_i)}{\sum_{k=1}^{\infty} \mathrm{P}(A|C_k)\,\mathrm{P}(C_k)}.$$

□

EJEMPLO 2.42. Considere un lote de 20 artículos defectuosos y 80 sin defectos, desde los que se escoge 2 artículos sin reemplazo. Sea

$$A = \{\text{el 1er artículo es defectuoso}\},$$
$$B = \{\text{el 2do artículo es defectuoso}\}.$$

Para calcular $\mathrm{P}(B)$ podemos hacer

$$\begin{aligned}\mathrm{P}(B) &= \mathrm{P}(B|A)\,\mathrm{P}(A) + \mathrm{P}(B|A^c)\,\mathrm{P}(A^c) = \frac{19}{99}\frac{1}{5} + \frac{20}{99}\frac{4}{5} \\ &= \frac{1}{5}\frac{1}{99}(19 + 20 \cdot 4) = \frac{1}{5}.\end{aligned}$$

2.6. Independencia estadística

DEFINICIÓN 2.43. Sea $(\Omega, \mathcal{F}, \mathrm{P})$ un espacio de probabilid y sean $A, B \in \mathcal{F}$. Se dice que A y B son independientes si y sólo si

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B).$$

Naturalmente podemos entender la independencia del siguiente modo: "la ocurrencia de un evento B no tiene efecto en la probabilidad de otro evento A". Es decir,

$$\mathrm{P}(A|B) = \mathrm{P}(A).$$

Note tambien que,

$$\mathrm{P}(B|A) = \frac{\mathrm{P}(A|B)\,\mathrm{P}(B)}{\mathrm{P}(A)} = \frac{\mathrm{P}(A)\,\mathrm{P}(B)}{\mathrm{P}(B)} = \mathrm{P}(B),$$

es decir, la ocurrencia de A *no* tiene efecto en B.

RESULTADO 2.44. *Si A y B son independientes, entonces los siguientes pares también son independientes*

(a) *A y B^c.*
(b) *A^c y B.*
(c) *A^c y B^c.*

DEMOSTRACIÓN. Para probar (a), note que

$$\begin{aligned}\mathrm{P}(A \cap B^c) &= \mathrm{P}(A) - \mathrm{P}(A \cap B) = \mathrm{P}(A) - \mathrm{P}(A)\,\mathrm{P}(B) = \mathrm{P}(A)(1 - \mathrm{P}(B)) \\ &= \mathrm{P}(A)\,\mathrm{P}(B^c).\end{aligned}$$

Partes (b) y (c) son análogas y se dejan de ejercicio para el lector. □

DEFINICIÓN 2.45. Una colección de eventos $A_1, A_2, \ldots, A_n$ es mutuamente independiente si para cualquier subcolección $A_{i_1}, \ldots, A_{i_k}$, tenemos

$$\mathrm{P}\Big(\prod_{j=1}^{k} A_{i_j}\Big) = \prod_{j=1}^{k} \mathrm{P}(A_{i_j}).$$

EJEMPLO 2.46. Se lanzan 3 dados de distinto color: blanco, rojo y negro ¿Cuál es la probabilidad de que el dado blanco salga 3 y los otros dos no? Considere A, B y C los eventos

A :resultado del dado blanco es 3,
B :resultado del dado rojo es 3,
C :resultado del dado negro es 3,

tenemos $\mathrm{P}(A) = \mathrm{P}(B) = \mathrm{P}(C) = 1/6$ y se pide calcular

$$\mathrm{P}(A \cap B^c \cap C^c) = \mathrm{P}(A)\,\mathrm{P}(B^c)\,\mathrm{P}(C^c) = \frac{1}{6}\frac{5}{6}\frac{5}{6} = \frac{25}{216}.$$

Bibliografía

Anscombe, F.J. (1973) Graphs in statistical analysis. *The American Statistician* **27**, 17-21.

Barlow, J.L. (1993). Numerical aspects of solving linear least squares problems. En: *Handbook of Statistics, Vol. 9*, C.R. Rao (Ed.). Elsevier, pp. 303-373.

Bolfarine, H., Sandoval, M.C. (2001). *Introdução à Inferência Estatística*. Sociedade Brasileira de Matemática, Rio de Janeiro.

Casella, G., Berger, R.L. (2002). *Statistical Inference (2nd Ed.)*. Duxbury, Pacific Grove.

Chan, T.F., Lewis, J.G. (1979). Computing standard deviations: Accuracy. *Communications of the Association for Computing Machinery* **22**, 526-531.

Chan, T.F., Golub, G.H., LeVeque, R.J. (1983). Algorithms for computing the sample variance: Analysis and recommendations. *The American Statistician* **37**, 242-247.

Clarke, M.R.B. (1971). Algorithm AS 41: Updating the sample mean and dispersion matrix. *Applied Statistics* **20**, 206-209.

Cleveland, W.S. (1993). *Visualizing Data*. Hobart Press, New Jersey.

Frery, A.C., Cribari-Neto, F. (2005). *Elementos de Estatística Computacional Usando Plataformas de Software Livre/Gratuito*. IMPA, Rio de Janeiro.

Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7**, 179-188.

Galbiati, J. (2012). *Tablas de Probabilidad (8a Ed.)*. Instituto de Estadística, Pontificia Universidad Católica de Valparaíso, Chile.

Graillat, S. (2009). Accurate floating-point product and exponentiation. *IEEE Transactions on Computers* **58**, 994-1000.

Grossman, S.I., Turner, J.E. (1974). *Mathematics for the Biological Sciences*. MacMillan Publishing, New York.

Jambu, M. (1991). *Exploratory and Multivariate Data Analysis*. Academic Press, Boston.

Keeling, K.B., Pavur, R.J. (2007). A comparative study of the reliability of nine statistical software packages. *Computational Statistics & Data Analysis* **51**, 3811-3831.

Knuth, D.E. (1997). *The Art of Computer Programming: Vol. 1, Fundamental Algorithms*. Addison-Wesley, Massachusetts.

Mardia, K.V. (1970). Measures of multivariate skweness and kurtosis with applications. *Biometrika* **57**, 519-530.

Mardia, K.V. (1974). Application of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. *Sankhyā, Series B* **36**, 115-128.

McCullough, B.D., Wilson, B. (1999). On the accuracy of statistical procedures in Microsoft Excel 97. *Computational Statistics & Data Analysis* **31**, 27-37.

McCullough, B.D., Wilson, B. (2002). On the accuracy of statistical procedures in Microsoft Excel 2000 and Excel XP *Computational Statistics & Data Analysis* **40**, 713-721.

McCullough, B.D., Wilson, B. (2005). On the accuracy of statistical procedures in Microsoft Excel 2003 *Computational Statistics & Data Analysis* **49**, 1244-1252.

Mood, A.M., Graybill, F.A., Boes, D.C. (1974). *Introduction to the Theory of Statistics (3rd Ed.)*. McGraw-Hill, New York.

Osorio, F., Ogueda, A. (2022). fastmatrix: Fast computation of some matrices useful in statistics. R package version 0.4. URL: faosorios.github.io/fastmatrix/

Panaretos, V.M. (2016). *Statistics for Mathematicians: A Rigorous First Course*. Birkhäuser, Laussane.

Pébay, P., Terriberry, T.B., Kolla, H., Bennett, J. (2016). Numerically stable, scalable formulas for parallel and online computation of higher order multivariate central moments with arbitrary weights. *Computational Statistics* **31**, 1305-1325.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P. (1992). *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, Cambridge.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: www.R-project.org

Spicer, C.C. (1972). Algorithm AS 52: Calculation of power sums of deviations about the mean. *Applied Statistics* **21**, 226-227.

Thisted, R.A. (1988). *Elements of Statistical Computing*. Champan & Hall, Boca Raton.

Tukey, J.W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Massachusetts.

Venables, W.N., Ripley, B.D. (1994). *Modern Applied Statistics with S-Plus*. Springer, New York.

Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference*. Springer, New York.

Welford, B.P. (1962). Note on a method for calculating corrected sums of squares and products. *Technometrics* **4**, 419-420.

West, D.H.D. (1979). Updating mean and variances estimates: An improved method. *Communications of the Association for Computing Machinery* **22**, 532-535.

Wilkinson, L. (1999). *The Grammar of Graphics*. Springer, New York.