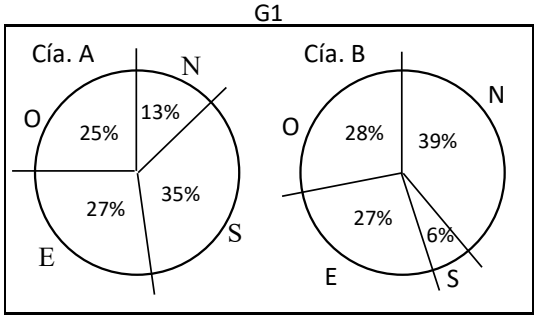
Guía de Ejercitación 1. Estadística Descriptiva

Ejercicio 1. Decirlo con un gráfico. ¿Cuál de las seis gráficas presentadas parece más adecuada a los datos de la tabla? Discutir el énfasis del mensaje que transmiten.

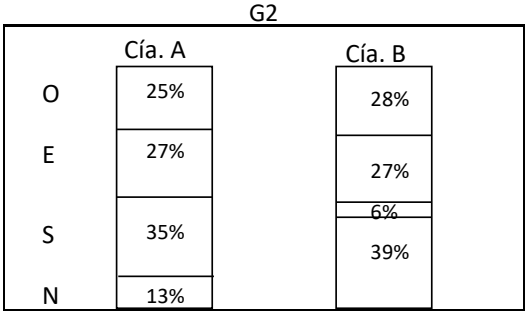
	Compañía A	Compañía B
Norte	13%	39%
Sur	35%	6%
Este	27%	27%
Oeste	25%	28%

Porcentaje de ventas por región durante Enero

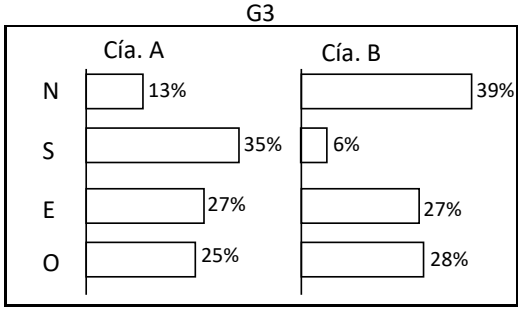
G1



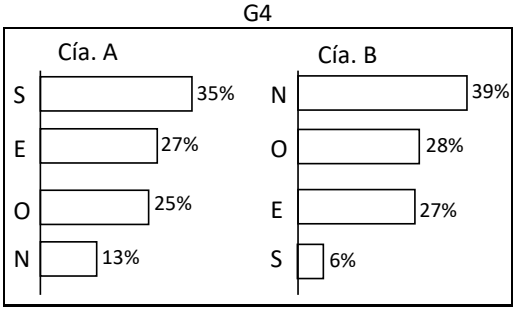
G2



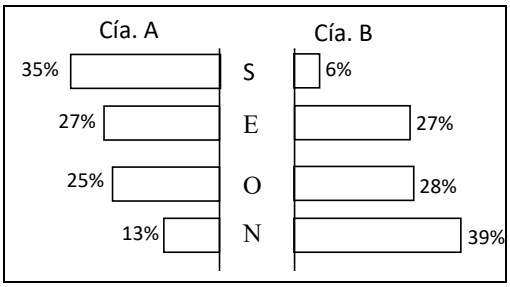
G3



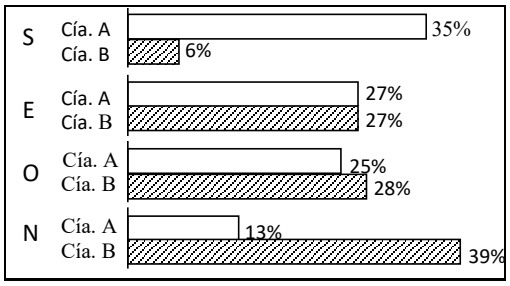
G4



G5



G6



Ejercicio 2. Dados los siguientes tres conjuntos de datos calcular la media, mediana y dispersión para cada conjunto:

CI	1	2	3	4	5	6
CII	1	1	1	6	6	6
CIII	-13	2	3	4	5	20

¿Qué se puede concluir?

Ejercicio 3. Se tiene un conjunto de observaciones de duración de piedras de esmeril, en miles de piezas trabajadas.

104 25 66 30 22 33 35 69 52 63 71

Usando las expresiones correspondientes y una calculadora hallar:

- **a)** La mediana, la media y la desviación estándar.
- **b)** Los 4 cuartiles y el recorrido intercuartílico.
- **c)** Los fractiles (o cuantiles) 0,1 y 0,7 (llamados también percentiles 10 y 70 respectivamente).
- **d)** ¿Qué porcentaje de piedras tendrá una vida superior a las 40 mil piezas?

Ejercicio 4. El estallido del transbordador espacial Challenger con sus astronautas, el 28 de Enero de 1986 a los 73 segundos del despegue, condujo a efectuar varios estudios para investigar los motivos de su fracaso al elevarse. La atención se dirigió rápidamente al comportamiento de los "anillos-O" (*O-rings*) del motor del cohete (juntas tóricas que deben asegurar la perfecta estanqueidad de los cohetes aceleradores). Los datos (x_i) de las observaciones de la temperatura del anillo (en grados Fahrenheit, °*F*) para cada ignición de prueba o de despegue real del motor cohete, como el del transbordador (*Presidential Commission on the Space Shuttle Challenger Accident*, Vol. 1, 1986, pp. 129-131) son los siguientes:

84	49	61	40	83	67	45	66	70	69	80	58
68	60	67	72	73	70	57	63	70	78	52	67
53	67	75	61	70	81	76	79	75	76	58	31

- **a)** Calcular la media $(\bar{x})$, mediana $(\tilde{x})$ y la desviación estándar muestral (s) con una herramienta de preferencia personal. Los datos se encuentran digitalizados en el archivo "DatosGED.xlsx", en la hoja "Temperaturas de O-Ring".
- **b)** Calcular los cuartiles y el recorrido intercuartílico.
- **c)** A partir de los valores calculados previamente confeccionar el diagrama de cajas y bigotes (a mano). ¿Hay datos inusuales y/o extremos?
- **d)** ¿Qué porcentaje de los valores cae en el intervalo $(\bar{x} - s; \bar{x} + s)$? ¿Cuál porcentaje en $(\bar{x} - 2s; \bar{x} + 2s)$?

Ejercicio 5. Duración de baterías. En ocasiones el analista no tiene acceso a los datos individuales, sino que se le da una tabla con las cantidades de observaciones por categorías. Considerar los siguientes datos de duración (en semanas) de una muestra de baterías de un mismo tipo.

x [Semanas]	$[0; 50)$	$[50; 75)$	$[75; 100)$	$[100; 125)$	$[125; 150)$	$[150; 200)$	$[200; 700)$
Baterías	17	25	59	88	134	150	27

Responder a las siguientes cuestiones.

- **a)** Calcular media, mediana, moda y desvío estándar de la muestra con los datos agrupados dados. ¿Cuáles de estas medidas de tendencia central y dispersión son preferibles? Justificar.
- **b)** ¿Cuál es la duración garantizada con 90% de confianza?
- **c)** Evaluar el porcentaje de baterías con duraciones superiores o iguales a 100 semanas.
- **d)** Hallar la duración promedio de las baterías que duran al menos 100 semanas.
- **e)** Trazar un histograma de frecuencias absolutas y el de frecuencias relativas ajustando la ordenada según los anchos de los intervalos (histograma de densidad de frecuencias). Compararlos. ¿Cuál parece más apropiado? Justificar.
- **f)** Comparar la media y la mediana. Relacionarlo con el histograma.

Ejercicio 6. Boxplot con variables categóricas. En una planta existen tres líneas de llenado de un producto. En el archivo "DatosGED.xlsx", en la hoja "Volumen de llenado" se dan 100 valores de los volúmenes (en cm^3) que llenó cada máquina en un turno.

a) Trazar en una misma gráfica el boxplot correspondiente a cada máquina.

b) Visualizar el gráfico e indicar las siguientes cuestiones, ¿Hay diferencias entre las producciones promedio? ¿Qué se puede decir de los desvíos estándar de las producciones de cada máquina?

Ejercicio 7. Histogramas a partir de datos individuales. Realizar este ejercicio a mano y con distintas herramientas. En este problema, por única vez, se deberá efectuar el histograma de frecuencias relativas a mano y luego compararlo con los que se obtienen usando una planilla de cálculo como Excel y el que se obtiene con R. También se puede usar el Infostat u otro software estadístico. Antes de comenzar a resolverlo, organizar los distintos pasos que es conveniente llevar a cabo para ejecutar lo pedido sin el auxilio del uso de una computadora o elemento similar.

A continuación, se presentan 100 datos de los saldos de cajas de ahorro de una sucursal bancaria expresados en miles de pesos (también están digitalizados en el archivo "DatosGED.xlsx", en la hoja "Saldos bancarios"):

272	299	328	305	198	397	286	179	486	890
823	1557	704	497	434	1365	257	574	191	345
125	158	513	563	102	149	513	449	456	781
502	443	200	631	772	1104	254	204	242	220
232	82	298	146	112	434	414	103	288	420
320	381	363	258	779	261	226	130	976	1782
185	151	176	207	561	244	559	562	121	330
290	186	525	234	303	297	112	1222	399	96
229	276	249	275	175	329	449	443	427	359
627	575	244	697	249	333	362	418	545	211

a) Construir un histograma.

b) Calcular la media, la mediana y el desvío estándar a partir de los datos de la muestra.

c) Calcular la media, la mediana y el desvío estándar a partir de los datos agrupados. Compararlos.

d) Usando Excel (u otra herramienta computacional) ordenar los datos y calcular los 3 cuartiles, y el recorrido intercuartílico.

e) A partir de la tabla de frecuencias relativas calcular el porcentaje de cuentas con saldo superior a $500000.

Ejercicio 8. En el archivo "DatosGED.xlsx", en la hoja "Ciclo Combinado" se encuentran datos de la potencia entregada por una central térmica de ciclo combinado. Se relevaron datos diarios de la potencia máxima entregada (PE, en MW) por la planta funcionando en capacidad máxima. Los registros fueron tomados entre los años 2006 y 2011. La variable "HighTemp" vale 1 si la temperatura media diaria fue superior a 20°C en el día en el que se tomó el dato y vale 0 en caso contrario.

a) Realizar un histograma con los datos de PE, ¿Cuántas modas parece haber? ¿Tiene sentido pensar que estamos ante una mezcla de poblaciones? Justificar.

b) Trazar los histogramas individuales para los días en que la temperatura fue mayor a 20°C y menor o igual a 20°C respectivamente.

c) Eliminar la mitad de las observaciones mayores que corresponden a los días en los cuáles la temperatura fue inferior o igual a 20°C y volver a trazar el histograma manteniendo todos los datos para aquellos días en los cuáles la temperatura fue superior a 20°C. ¿Qué se puede observar?

Ejercicio 9. Problema adaptado del libro "Estadística para Ingenieros y científicos". William Navidi (2006). El artículo "Virgin Versus Recycled Wafers for Furnace Qualification: Is the Expense Justified?" (V. Czitrom y J. Reece, en Statistical Case Studies for Industrial Process Improvement, ASA y SIAM, 1997:87-104) describe un proceso para el crecimiento de una capa delgada de dióxido de silicio sobre placas de silicio que se usan en la fabricación de semiconductores. La tabla presenta las mediciones del espesor, en Angstroms $\left(\text{Å}; 1\text{Å} \equiv 1 \times 10^{-8} cm\right)$, de la capa de óxido para 24 placas. Se hicieron nueve mediciones en cada placa. Las placas se fabricaron en dos corridas distintas, con 12 placas por cada corrida. Los datos se encuentran digitalizados en el archivo "DatosGED.xlsx", en la hoja "Espesor de placas"

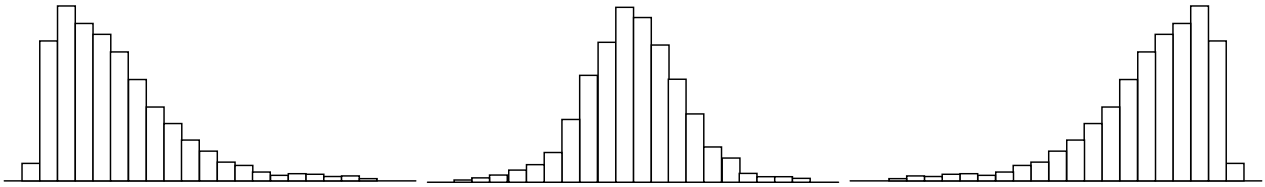
	Placa	**Espesor** $[\text{Å}]$								
Corrida 1	1	90,0	92,2	94,9	92,7	91,6	88,2	82,0	98,2	96,0
	2	91,8	94,5	93,9	77,3	92,0	89,9	87,9	92,8	93,3
	3	90,3	91,1	93,3	93,5	87,2	88,1	90,1	91,9	94,5
	4	92,6	90,3	92,8	91,6	92,7	91,7	89,3	95,5	93,6
	5	91,1	89,8	91,5	91,5	90,6	93,1	88,9	92,5	92,6
	6	76,1	90,2	96,8	84,6	93,3	95,7	90,9	100,3	95,2
	7	92,4	91,7	91,6	91,1	88,0	92,4	88,7	92,9	92,6
	8	91,3	90,1	95,4	89,6	90,7	95,8	91,7	97,9	95,7
	9	96,7	93,7	93,9	87,9	90,4	92,0	90,5	95,2	94,3
	10	92,0	94,6	93,7	94,0	89,3	90,1	91,3	92,7	94,5
	11	94,1	91,5	95,3	92,8	93,4	92,2	89,4	94,5	95,4
	12	91,7	97,4	95,1	96,7	77,5	91,4	90,5	95,2	93,1
Corrida 2	1	93,0	99,9	93,6	89,0	93,6	90,9	89,8	92,4	93,0
	2	91,4	90,6	92,2	91,9	92,4	87,6	88,9	90,9	92,8
	3	91,9	91,8	92,8	96,4	93,8	86,5	92,7	90,9	92,8
	4	90,6	91,3	94,9	88,3	87,9	92,2	90,7	91,3	93,6
	5	93,1	91,8	94,6	88,9	90,0	97,9	92,1	91,6	98,4
	6	90,8	91,5	91,5	91,5	94,0	91,0	92,1	91,8	94,0
	7	88,0	91,8	90,5	90,4	90,3	91,5	89,4	93,2	93,9
	8	88,3	96,0	92,8	93,7	89,6	89,6	90,2	95,3	93,0
	9	94,2	92,2	95,8	92,5	91,0	91,4	92,8	93,6	91,0
	10	101,5	103,1	103,2	103,5	96,1	102,5	102,0	106,7	105,4
	11	92,8	90,8	92,2	91,7	89,0	88,5	87,5	93,8	91,4
	12	92,1	93,4	94	94,7	90,8	92,1	91,2	92,3	91,1

Las 12 placas en cada corrida eran de varios tipos y se procesaron en diferentes posiciones en el horno. El propósito en la recopilación de datos fue determinar si el espesor de la capa de óxido se afectaba ya sea por el tipo de placa o por la posición en el horno. Por tanto, éste fue un experimento factorial, con los factores "tipo de placa" y "posición en el horno" y como resultado "el espesor de la capa de óxido". El experimento se diseñó de tal manera que no se supuso ninguna diferencia sistemática entre las capas de una corrida a otra. El primer paso en el análisis es construir un diagrama de caja para los datos de cada corrida con el propósito de ayudar a determinar si esta condición se satisfacía realmente y analizar si existen valores *outliers*.

a) Construir y comparar histogramas para ambas muestras.
b) Construir y comparar Box-Plots para ambas muestras.
c) ¿Qué conclusiones preliminares se puede sacar de las comparaciones? ¿Qué gráfico resultó más apropiado?

Ejercicios Misceláneos de Estadística Descriptiva

Ejercicio 10. Es habitual que la imagen visual presentada por el histograma se conozca como "forma de la distribución" de los datos. Con frecuencia en estadística se clasifican las distribuciones en base a su forma. Una de las posibles clasificaciones se basa en el sesgo. Una distribución es sesgada si la mayor parte de los valores están agrupados hacia el borde izquierdo –lo que demuestra un sesgo positivo o asimetría positiva– o el derecho del histograma –esto es, un sesgo negativo. Una distribución es simétrica si los valores se agrupan en la parte media sin sesgo positivo o negativo. En la figura se presentan ejemplos de estos tres tipos de formas.



Sesgo Positivo Distribución Simétrica Sesgo Negativo

Indicar cuál es la distribución esperable, entre las tres dadas, para las siguientes variables: consumo de energía eléctrica por usuario en una gran ciudad capitalina; sueldos en una fábrica; medidas antropométricas de individuos de un grupo homogéneo –por ejemplo estatura–; rendimientos de cosecha en kg por parcela en experimentos agropecuarios diseñados; tiempo de vida de los seres de una misma especie con aplicación de fármacos eficientes. Proponer algunos otros ejemplos para casos de sesgo positivo, insesgado y sesgo negativo.

Ejercicio 11. Se presenta a continuación información referida a diferentes medidas de resumen de una muestra de datos. Completar la información faltante indicando claramente los cálculos realizados. Una vez completados los cuadros, ¿la información es suficiente como para deducir si hay o no algún dato inusual? Justificar

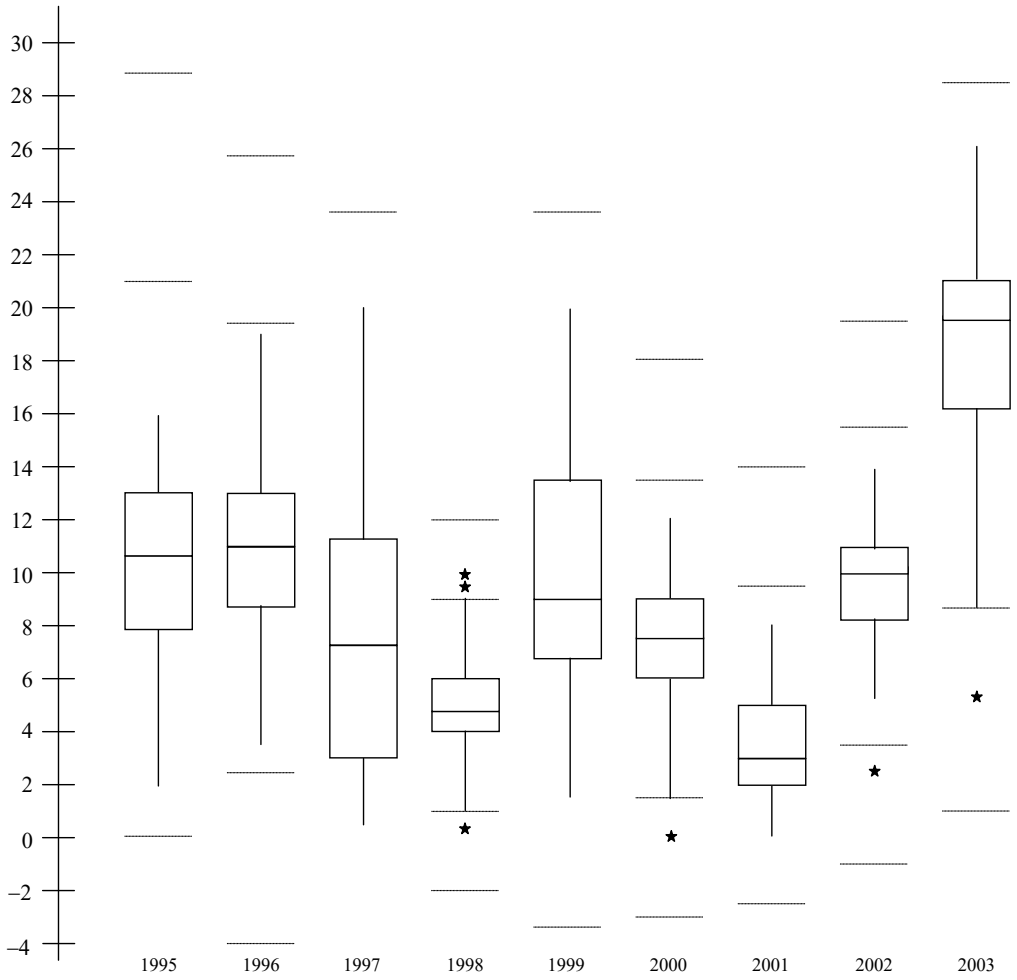
Tamaño muestra	Media	Mediana	Desvío estándar	Varianza	Rango	Mínimo	Máximo	Suma total
108		23,75		8,577	15		31,5	2594,5

Percentiles					
10	20	25	50	60	75
20,5	21,5			24,5	26

Rango intercuartílico	Cuartil 1	Cuartil 2	Cuartil 3
	22		

Ejercicio 12. Comparación con diagramas de caja y bigote. La siguiente información tabular corresponde a las ganancias en cientos de $ (un 5 representa $500, un 10 a $1000, etc.) obtenidas por hectárea en 52 campos durante la cosecha de los años mencionados. La figura que continúa a la tabla representa la misma información, pero en diagramas de caja y bigotes (box-plot) para cada año. Los datos se encuentran digitalizados en el archivo "DatosGED.xlsx", en la hoja "Ganancia campos".

Fila	1995	1996	1997	1998	1999	2000	2001	2002	2003
1	5	5	2	4,5	1,5	1,5	1	5	9,5
2	10	5	5	9	4	4	3	8	17,5
3	11	11	4	9	6	6	3	11	18
4	11,25	10	5	9,5	7	5	3	9	20
5	11	7	4	10	7	6	5	9,5	20,5
6	11	13	3	7	10	5	3,5	10	20
7	9	10,5	2,5	7,5	7	5	3	12	19
8	9	10	2	7	7	6	4,5	10	21
9	9	13	2	8	10	8	5	11,5	19
10	9	12	2	7	7	5	5	10	26
11	6	10	3	7	6	8	6	11	20
12	7	9	2,5	5	7	7	5	10	23
13	5	11	2	5	9	10	4	10,5	22
14	7	7	3	4	9	9	6	10	20
15	10	8	3	3	8	9	5	11	20
16	13	10,5	3	4	12	9	6,5	11	24
17	5,5	10	4	4	13	7	4	11,5	23
18	10,5	11,5	5	4	15	7	6	12	23
19	9	9	6	4,5	17	7	3	12,5	20
20	6	11	8	3,5	20	7	6	13	21
21	8	13	7	6	16	8	5	11	23
22	15	15	11	5	17	9	8	11	23
23	13	14	9	6	15	7	4,5	11	22
24	14	17	10	6	12	8	4	13	22
25	15	16	10	5	13	8	4	11	21
26	14,5	14	11	7	13	8	6,5	13	20
27	14	11	14	6	18	8	4	12	18
28	14	16	15	5	20	8	5	14	21
29	15	12	12	3	16	7	3	10	16
30	16	11,5	10	5	16	7	5,5	13	16
31	13	11	12	4	16	10	3	8	15
32	12	13	13	4,5	16	8	2	7,5	16
33	9	12	9,5	4	15	7	2	8	15
34	11,5	12	8	4	11	7	2	8	15
35	9,5	13	8	4	11	11	2	7	17
36	12	9	9	4	11	11	3	10	17
37	13	15	11	6	8,5	12	2	7	21
38	13,5	12	13	6	10	10	2	10	17
39	15	7	16	4	6	11	2	8	20
40	15,5	7	18	3	10	11	2	10	16
41	14	10	12	3,5	8	11	2	9	20
42	14	5	15	8	8	8,5	1,5	10	15
43	10	3,5	14	6	6	10	1,5	8	17
44	13	5	20	4,5	8	9	1,5	11	17
45	11	14	14	6	6	11	1	10	20
46	8,5	13	7	4,5	6	7	1	7	15
47	7,5	19	7,5	4	7	5	1	5	14
48	6	11,5	4	3	3	3	1	7	12
49	5	10	5	3	3	3	0,5	8	11
50	4	6	3	2	2	1,5	0,5	6	11
51	3	6,5	1,5	1	2	1	0	7,5	8,5
52	2	5	0,5	0,5	2	0	0	2,5	2,5



Ganancias anuales por hectárea, en cientos de $, en 52 campos en distintos años

a) Notar la facilidad de la expresión gráfica en cuanto a la información global de las ganancias año tras año y sus variaciones. Comparar la ganancia en los distintos años a partir de la mediana de los mismos.

b) ¿Cuáles fueron los dos años de mayor variabilidad?

c) ¿Hay alguna evidencias de que cuanto más bajas son las ganancias anuales menor es la variabilidad en el conjunto?

d) Mencionar una desventaja del tipo de expresión gráfica presentada con respecto a la información dada en la tabla.

Ejercicios de mirada conceptual/teórica

Ejercicio I. Sea un conjunto de datos $x_1, x_2, \ldots, x_n$ donde $x_i \in \mathbf{R}$ con $i = 1, 2, \ldots, n$.

a) ¿Para qué valor de c la cantidad definida a continuación es mínima?

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

Sugerencia: tomar la derivada primera de la expresión en función de *c*, igualar a cero y resolver.

b) Demostrar que la suma de los desvíos de las observaciones con respecto a la media muestral es nula. Esto es:

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

Esta demostración permite interpretar físicamente la ubicación de la media muestral como el punto de equilibrio en un sistema donde cada observación representa una pesa ubicada en su

posición sobre una barra con respecto a una escala numérica adosada a ella. La media muestral es el punto donde colocar el *fulcro* para equilibrar el sistema.

c) Sean a y b constantes y sea $y_i = ax_i + b$ para $i = 1,2,\cdots,n$ (cambio de escala y traslación). ¿Cuál es la relación entre los valores $\bar{x}$ y de $\bar{y}$? ¿Y entre los valores de $\tilde{x}$ y de $\tilde{y}$? ¿Entre s_x^2 s_x^2 y s_y^2? ¿Entre s_x y s_y?

d) Utilizando las propiedades del cambio de escala y de desplazamiento de los valores de un conjunto de datos analizadas en el ítem anterior sobre cómo afectan a las medidas de tendencia central y variabilidad, indicar los valores en Celsius (°*C*) de la media y el desvío estándar muestral halladas en el ejercicio 4 de las temperaturas del anillo-O en las pruebas para el transbordador Challenger que se dan en grados Fahrenheit. El vínculo entre ambas escalas de temperatura es lineal dado por $C = (5/9)(F-32)$, donde F y C representan las temperaturas medidas en grados Fahrenheit y Celsius, respectivamente.

Ejercicio II. De la vida cotidiana. Describir situaciones reales donde estén involucrados los siguientes conceptos, no necesariamente todos en una misma situación: población estadística, censo, muestra, variable categorial, variable numérica, datos univariados, datos bivariados, datos multivariados, muestreo estratificado, estudio de observación, experimento aleatorizado, grupo de control, grupo de tratamiento, placebo, experimento ciego, experimento doble ciego.

A tener en cuenta. Esta guía de ejercicios se completa con la entrega del primer trabajo práctico grupal con asistencia informática. A cada equipo se le indicará en clase el problema que deba resolver, forma y la fecha de entrega correspondiente. En igual forma, las actividades de autoevaluación del CVG.

Algunas definiciones y observaciones

Medidas numéricas descriptivas.

Sea un conjunto de datos $x_1, x_2, \dots, x_n$ donde $x_i \in \mathbf{R}$ con $i = 1, 2, \dots, n$. Dentro de los números que pueden servir para caracterizarlo se encuentran fundamentalmente dos tipos de medidas de interés: **tendencia central** y **variabilidad**.

Medidas de tendencia central. Indican la disposición de los datos para agruparse ya sea alrededor del centro o de ciertos valores numéricos: media, mediana, moda.

Media muestral:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad (\text{promedio aritmético})$$

Mediana muestral: $\tilde{x}$, ordenando primero las n observaciones de menor a mayor (incluyendo eventuales valores repetidos)

$$\tilde{x} \equiv \begin{cases} x_{k+1} & n = 2k+1 \quad (n \text{ impar}) \\ \dfrac{x_k + x_{k+1}}{2} & n = 2k \quad (n \text{ par}) \end{cases} \quad (\text{valores ordenados}),$$

divide al conjunto de datos en dos partes de igual tamaño.

Moda. Es el valor de las observaciones que más se repite en el conjunto. Un conjunto de datos puede no tener una moda definida, o bien ser bi o tri-modal.

Otras medidas de localización.

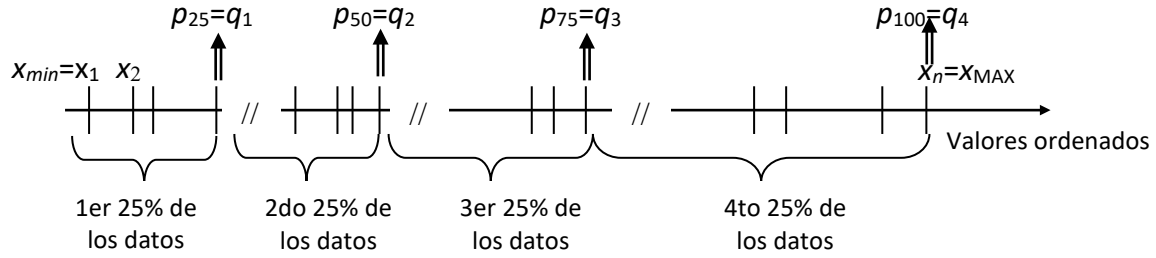
Media recortada (o acotada) al $a\%$ (con $0 < a < 100$): se define como el promedio aritmético del conjunto de datos eliminando el $a/2\%$ de los valores más pequeños y el $a/2\%$ de los valores más grandes.

Media ponderada:

$$x_p = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}{w_1 + w_2 + \cdots + w_n} = \frac{\sum_{i=1}^{n} w_1 x_1}{\sum_{i=1}^{n} w_i}, w_i \in \mathbf{R}_{\geq \mathbf{0}}, i = 1,2,\cdots,n; \sum_{i=1}^{n} w_i \neq 0.$$

Cuartiles $q_1, q_2, q_3,\ q_4$, dividen en forma aproximada al conjunto de datos en cuatro partes iguales.

Percentiles $p_1 = x_{min}, p_2, \dots, p_{25}, \dots,\ p_{50} = \tilde{x}, \dots,\ p_{75}, \dots,\ p_{99}, p_{100} = x_{MAX}$, con ellos se puede dividir el conjunto de datos en 100 partes iguales.



Medidas de variabilidad. Indican la dispersión de los datos en el conjunto.

Recorrido (R) o intervalo. Diferencia entre los valores máximo y mínimo de la muestra.

$$R = x_{MAX} - x_{min}$$

Desviación media ($D.M.$). Es el promedio de los valores absolutos de las diferencias entre cada observación y la media de las observaciones.

$$D.M. = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|$$

Desviación mediana ($D.Md.$). Es el promedio de los valores absolutos de las diferencias entre cada observación y la mediana de las observaciones.

$$D.Md. = \frac{1}{n}\sum_{i=1}^{n}|x_i - \tilde{x}|$$

Recorrido intercuartílico o cuarta dispersión. La diferencia entre los cuartiles tercero y primero.

$$f_s = q_3 - q_1 = p_{75} - p_{50}$$

Varianza muestral s^2 se define por:

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

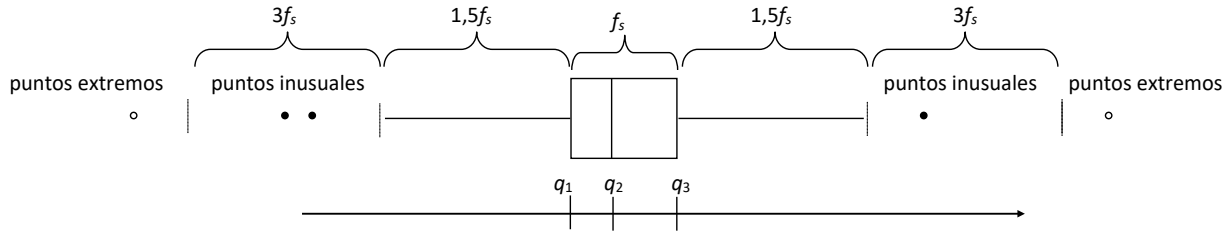
Desviación, dispersión o desvío estándar muestral $s = +\sqrt{s^2}$.

Diagrama de caja y bigote (*boxplot*).

En años recientes se ha empleado con éxito un resumen gráfico llamado *diagrama de caja y bigote* (*boxplot*), para describir varias de las características más destacadas de un conjunto de datos. Entre estas características están: centro, dispersión, naturaleza y magnitud de cualquier desviación de la simetría, y la identificación de puntos inusuales (atípicos), o sea, observaciones que están muy lejos del cuerpo principal de los datos. Debido a que incluso un solo punto inusual puede afectar de manera drástica el valor de algunas medidas numéricas, como el promedio o la dispersión muestral, un diagrama de caja está basado en medidas que son resistentes a la presencia de unos cuantos puntos atípicos: la mediana ($p_{50} = q_2 = \tilde{x}$) y una medida de la dispersión llamada cuarta dispersión o recorrido intercuartílico, $f_s = q_3 - q_1 = p_{75} - p_{50}$.

Primero se traza una escala horizontal de medición. A continuación, se pone un rectángulo sobre ese eje; su borde izquierdo está en el cuarto inferior q_1 y el derecho en el cuarto superior q_3, por lo que el ancho de la caja es f_s. Se traza un segmento de recta vertical o algún otro símbolo dentro del rectángulo, en el lugar de la mediana; la posición del símbolo de la mediana en relación con las dos orillas plasma la información de la asimetría en el 50% intermedio de los datos. Por último, se trazan "bigotes" o brazos que salen de ambos lados del rectángulo. En la primera versión de estos diagramas el largo de estos bigotes se correspondía con ser una vez y media f_s. Si existe alguna observación más alejada de $1{,}5f_s$., del cuarto más cercano es *inusual* o *outlier*. Un valor inusual es *extremo* si está a más de $3f_s$., y es *moderado* en cualquier otro caso. Cada valor inusual moderado se representa con un tipo de símbolo (por ejemplo, círculo lleno), y cada valor inusual extremo con otro (por ejemplo, círculo vacío). Esta descripción del procedimiento fue adaptada del texto Probabilidad y Estadística para ingeniería y ciencias, de Jay L. Devore, 5ta. edición, 2001, Ed. Thomson Learning.

También se puede trazar un diagrama de caja con orientación vertical, haciendo las modificaciones obvias al proceso anteriormente descrito.



En las versiones más modernas, el bigote que parte de cada extremo de la caja se recorta, si fuera el caso, antes de $1{,}5f_s$., llevándose hasta que terminen en las observaciones mínima y máxima que no sean puntos atípicos (mínimo y máximo de los puntos moderados).

Manejo de un gran número de datos.

Distribuciones de frecuencias. Cuando se dispone de un gran número de datos es útil agruparlos en *clases* o *categorías* y determinar el número de individuos pertenecientes a cada clase, que es la *frecuencia de clase*. Una ordenación en forma de tabla de los datos en clases, reunidas las clases y

con las frecuencias correspondientes, se conoce como una *distribución de frecuencias* o *tabla de frecuencias*. Aunque con este proceso de agrupamiento generalmente se pierde parte del detalle original de los datos, tiene la ventaja de presentar una información general en un cuadro que facilita el hallazgo de las relaciones que pueda haber entre ellos. Cada clase está definida por los *límites de clase*, inferior y superior, cuya diferencia determina la *longitud* o *ancho de la clase*. El punto medio del intervalo de clase se conoce como *marca de clase*. Se suele trabajar también con la *frecuencia relativa* de una clase, definida como frecuencia de la clase dividida por el total de frecuencias de todas las clases, confeccionando tablas correspondientes a *distribuciones de frecuencias relativas*.

Histogramas y polígonos de frecuencias. Son dos representaciones gráficas de las distribuciones de frecuencia. Un *histograma* o *histograma de frecuencias* consiste en una serie de rectángulos que tienen, *i*) sus bases sobre un eje horizontal con centros en las marcas de clase y longitud igual al tamaño de los intervalos de clase, *ii*) superficies proporcionales a las frecuencias de clase. Un *polígono de frecuencias* es un gráfico de línea trazado sobre las marcas de clase. Puede obtenerse uniendo los puntos medios de los techos de los rectángulos en el histograma. Se acostumbra a prolongar el polígono hasta las marcas de clase inferior y superior inmediatas, que corresponderían a clases de frecuencia cero, como se ve en la figura a. El área total de los rectángulos en un histograma es igual al área total limitada por el correspondiente polígono de frecuencias y el eje horizontal.

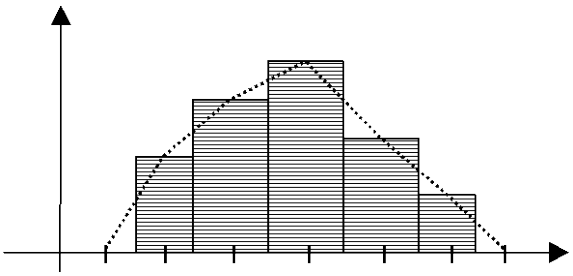


Fig. a. Histograma y polígono de frecuencias

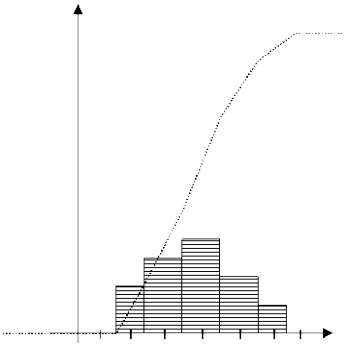


Fig. b. Histograma y polígono de frecuencias acumuladas

Otro gráfico de línea de utilidad lo constituye el llamado *polígono de frecuencias acumuladas* u *ojiva*, ver figura b, o bien de frecuencias relativas acumuladas. El principal uso de la distribución acumulada es lo que comúnmente se conoce como *cuantiles*. Con respecto a una distribución de frecuencia relativa acumulada, se define un *cuantil* como el valor bajo el cual se encuentra una determinada proporción de los valores de la distribución. El cuantil más común es el percentil *p*.

Medidas numéricas descriptivas de un conjunto de datos agrupados. Si el conjunto de *n* datos se ha agrupado en un número k de clases, sea x_i^* es la marca de la i −ésima clase y f_i la frecuencia de la i −ésima clase, de modo que $n = \sum_{i=1}^{k} f_i$.

El valor aproximado de la *media* muestral de los datos agrupados es:

$$\bar{x}_g = \frac{1}{n}\sum_{i=1}^{k} f_i x_i^*.$$

Para datos agrupados, en primer lugar es necesario determinar la clase que contiene el valor de la mediana, para después determinar el valor de la mediana dentro de la clase mediante interpolación. La clase que contiene a la mediana es la primera cuya frecuencia acumulada iguala o excede la mitad del total de observaciones. Una vez que se identifica esta clase, el número *m*, se determina el valor específico mediante el cálculo:

$$\tilde{x}_g = med = L_{i(m)} + \frac{1}{f_m}\left(\frac{n}{2} - \sum_{i=1}^{m-1} f_i\right)\cdot l_m,$$

donde $L_{i(m)}$ es el límite inferior de la clase que contiene a la mediana, n es el número total de observaciones en la distribución de frecuencias, $\sum_{i=1}^{m-1} f_i$ es la suma de las frecuencias de todas las clases por debajo de la clase de la mediana, f_m es la frecuencia de la clase de la mediana y l_m es el tamaño del intervalo de la clase mediana.

Para determinar los cuantiles, deciles y percentiles de datos agrupados, en primer lugar se determina la clase que contiene el punto de interés, de acuerdo con las frecuencias acumuladas, y después se lleva a cabo una interpolación como se hizo para el caso de la mediana. Por ejemplo, el primer cuartil (deja a la izquierda la cuarta parte de los datos), se calcula como

$$q_1 = L_{i(q)} + \frac{1}{f_q}\left(\frac{n}{4} - \sum_{i=1}^{q-1} f_i\right) \cdot l_q,$$

donde $L_{i(q)}$ es el límite inferior de la clase que contiene al primer cuantil, n es el número total de observaciones en la distribución de frecuencias, $\sum_{i=1}^{q-1} f_i$es la suma de las frecuencias de todas las clases por debajo de la clase del primer cuantil, f_q es la frecuencia de la clase del cuantil y l_q es el tamaño del intervalo de dicha clase.

La fórmula para el cálculo de la *varianza* en los datos agrupados es

$$s_g^2 = \frac{1}{n-1}\sum_{i=1}^{k} f_i\left(x_i - \bar{x}_g\right)^2 = \frac{1}{n-1}\left[\sum_{i=1}^{k} f_i (x_i^*)^2 - \frac{1}{n}\left(\sum_{i=1}^{k} f_i x_i^2\right)^2\right].$$

Lectura y comprensión

Histogramas como herramienta en la decisión de autorías

Lectura adaptada del *Capítulo 13*, *Decisión de Autorías*, escrito por Frederick Mosteller y David L. Wallace, en el libro *La Estadística. Una guía de lo desconocido.* Editorial Alianza, 1992.

El arte, la música, la literatura y las ciencias sociales, biológicas y físicas, entre otras, comparten una necesidad común de clasificar cosas: ¿Qué artista pintó el cuadro? ¿Quién compuso la obra? ¿Quién escribió el documento? ¿Cometerá otro crimen el prisionero que se encuentra en libertad condicional? ¿Qué señal química está deteriorando el proceso? En el campo de la Estadística estas **cuestiones** reciben el nombre **de clasificación o de discriminación**.

Las cuestiones de autorías son frecuentes y, a veces, importantes. Se ha oído hablar de la controversia Shakespeare-Bacon-Marlowe sobre quién escribió ciertas grandes obras generalmente atribuidas a Shakespeare. Son también ampliamente estudiadas cuestiones referentes a autorías de documentos religiosos antiguos llamados paulinos. Muchos temas de autoría se resuelven fácilmente realizando cálculos sistemáticos; sin embargo, en algunas ocasiones no son suficientes. Un famoso caso, problema especialmente difícil de la historia americana, es la controversia sobre la autoría de doce artículos federalistas reclamada por Alexander Hamilton y James Madison. Allí el análisis estadístico pudo contribuir a la resolución de esta cuestión histórica.

Los artículos de *El Federalista* fueron publicados anónimamente en 1787-88 por A. Hamilton, John Hay y J. Madison para inducir a los ciudadanos del estado de Nueva York a ratificar la Constitución. Aparecieron 77 artículos en forma de carta en periódicos de Nueva York bajo el pseudónimo de Publius. Junto con 8 ensayos más, en 1788 se publicaron en forma de libro, que se ha reeditado continuamente y sigue siendo una importante obra de filosofía política. En términos generales, se sabía quién había escrito El Federalista, pero la asignación pública de artículos específicos a autores no se produjo sino hasta mucho tiempo después y generó varias polémicas. Existe un acuerdo general sobre la autoría de 70 artículos -5 de Jay, 14 de Madison y 51 de Hamilton. De los 15 restantes, 12 se disputan entre Hamilton y Madison, y 3 son trabajos conjuntos. La razón principal de esta disputa es que Madison y Hamilton no registraron sus títulos. Pocos años después se escribir los ensayos, se convirtieron en adversarios políticos –por aquel entonces ambos alegaban a favor de lo mismo. Los contenidos políticos de los escritos en cuestión no sirven para discriminar la autoría. Sus estilos literarios, tampoco. **La longitud media de las frases** en los escritos, una medida usada con éxito para distinguir otros autores, no dio resultado pues Madison y Hamilton tenían medias semejantes de 34,5 y 34,6 palabras por frase en sus artículos conocidos.

Una variable usada en diversos estudios de autoría dudosa son **las razones de ocurrencia de palabras individuales específicas**. Así, en el caso en estudio, los dos autores difieren constantemente al elegir entre las palabras alternativas (sinónimos) *while* y *whilst*. En los 14 ensayos federalistas atribuidos a Madison, *while* no aparece nunca, mientras que *whilst* aparece en ocho de ellos; *while* aparece en 15 de 48 ensayos de Hamilton, pero nunca aparece *whilst*. Tenemos aquí un ejemplo de lo que se denomina **registros**–detalles cuya presencia proporciona una fuerte indicación de la autoría de una persona. Así, la presencia de *whilst* en cinco de los artículos disputados indica que Madison fue el autor de esos cinco artículos.

Cuando se detectan, los **registros** contribuyen mucho a la discriminación, pero también presentan dificultades. En primer lugar, *while* o *whilst* aparecen en menos de la mitad de los artículos; faltan en la otra mitad y, por tanto, no proporcionan evidencia en ninguno de los dos sentidos. Cabría superar esto encontrando bastantes palabras o construcciones **registro** diferentes. Una segunda dificultad, es que, partiendo de la evidencia de los 14 ensayos de Madison, no podemos asegurar que nunca empleara *while*. La presencia de esta palabra es, entonces, una indicación buena, pero no es segura.

Una parte fundamental de la Estadística consiste en hacer inferencias en presencia de incertidumbres. Renunciando a la exactitud, los **registros** nos conducen a un problema estadístico. Debemos encontrar evidencia, valorar su fuerza y combinarlo en una conclusión compuesta.

Una herramienta usada en el caso en estudio, fue la **razón o frecuencia relativa del uso de cada palabra** como una medida para distinguir entre un autor y otro. Por supuesto, muchas palabras no sirven ya que ambos autores las usaron aproximadamente con la misma frecuencia; sin embargo, ya que tenemos miles de palabras válidas, alguna puede ser de utilidad. Las palabras constituyen una enorme fuente de posibles elementos discriminatorios. De una exploración sistemática de esta fuente de palabras, no se encontraron más pares como *while-whilst*, pero sí otras palabras simples usadas regularmente por un autor y raramente por el otro. En este caso se procedió con ellas, pero a veces, el uso de las palabras está determinado por el tema del artículo.

Así, tal como en los procesos de autentificaciones de cuadros, o investigaciones policiales, también se pueden tener en cuenta pequeños detalles. En los escritos, esto es el **uso de palabras no contextuales**. ¿Qué palabras sin significado por sí mismas –que no estén relacionadas directamente con el tema del que trata el artículo sino, más bien, con las preferencias del autor– son buenas candidatas para distinción de autores? Las más atractivas son las palabras de "relleno" del lenguaje, como preposiciones, conjunciones o artículos. A modo de ejemplo presentamos aquí los histogramas de frecuencias relativas de ocurrencia de la palabra "by" en 48 artículos de Hamilton, 50 de Madison y 12 artículos diputados.

Se utilizaron más palabras discriminatorias dando una evidencia contundente de la autoría de Madison de los artículos disputados. Sólo en un artículo la evidencia es más modesta, y a pesar de ello el estudio más minucioso da una ventaja de 80 a 1 a favor de Madison (ventaja fuerte pero no contundente). El siguiente artículo en cuanto a claridad, da una ventaja de 800 a 1 para Madison. Para el resto, los datos son contundentes.