ESTADÍSTICA Y PROBABILIDAD 523250.

Profesores: Andrea Fernández & Jean Paul Navarrete

[1]Universidad de Concepción. Facultad de Ciencias Matemáticas.
Departamento de Estadística

Marzo, 2024

Contenido

Contenido

Medidas de tendencia central

- **Media aritmética:**
- **Para datos no agrupados**
  Si $\{x_1, \ldots, x_n\}$ son datos, entonces su media aritmética es:

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

Medidas de tendencia central

- **Para datos agrupados**
  - Si tenemos la siguiente tabla

Variable	**Frecuencia absoluta**	**Frecuencia relativa**	**Frec. Abs. Acum.**
x_1	n_1	n_1/n	N_1
x_2	n_2	n_2/n	N_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_k	n_k/n	N_k

La media es dada por

$$\overline{x} = x_1\, n_1 + \cdots + x_k\, n_k = \frac{1}{n} \sum_{i=1}^{k} x_i\, n_i$$

Medidas de tendencia central

- **Para datos agrupados**
  - Si tenemos la siguiente tabla

Clases	Marca de clase	Frecuencia absoluta	Frecuencia relativa	Frec. Abs. Acum.
$[L_0, L_1)$	c_1	n_1	n_1/n	N_1
$[L_1, L_2)$	c_2	n_2	n_2/n	N_2
$[L_2, L_3)$	c_3	n_3	n_3/n	N_3
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[L_{k-1}, L_k]$	c_k	n_k	n_k/n	N_k

La media es dada por:

$$\overline{x} \approx \frac{1}{n} \sum_{i=1}^{k} c_k \cdot n_k$$

donde k es el número de intervalos.

Algunos inconvenientes de la media

Remark 1

En general, la media aritmética (promedio) obtenida a partir de las marcas de clase c_k, diferirá de la media obtenida con los valores reales, x_i.

La perdida de precisión será mayor si las longitudes $b_i = L_i - L_{i-1}$ de los intervalos son grandes.

Remark 2

Notemos que la media es afectada por valores extremos o valores atípicos de la muestra o población, lo que implica que no es recomendable usar la media como medida central en las distribuciones muy asimétricas.

Remark 3

Si consideramos una variable discreta, por ejemplo, el número de hijos en las familias chilenas, el valor de la media puede no pertenecer al conjunto de valores de la variable; por ejemplo $\overline{x} = 1.2$ hijos.

Ejemplo

Example 1.1

Un ingeniero agrega un polímetro de látex a un mortero de cemento, para determinar los efectos del polímetro sobre la resistencia a la tensión (en kgf/cm^2). Los datos obtenidos de este experimento son:

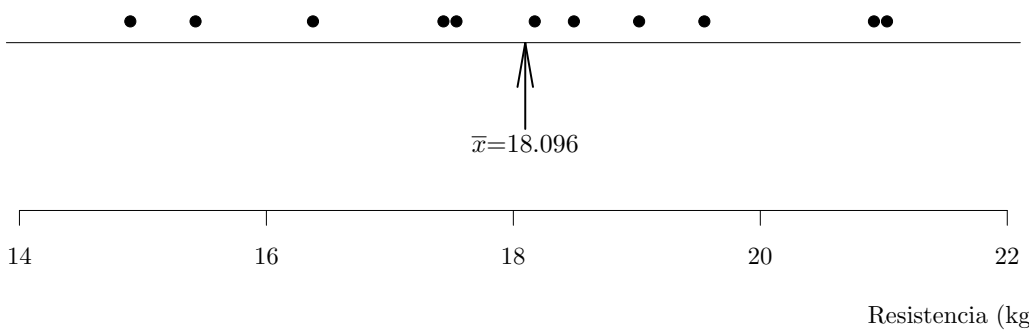
$$14.95, 15.40, 16.36, 17.45, 18.52, 19.04, 20.96, 18.15, 19.59, 17.57 y 21.07$$

Ejemplo

Example 1.2

La media muestral de la resistencia a la tensión de las 11 observaciones es:

$$\overline{x} = \frac{\sum_{i=1}^{11} x_i}{11} = \frac{14.95 + 15.40 + \cdots + 21.07}{11} = 18.09636$$



Ejemplo Remark (2)

Example 1.3

Se supone que una muestra de los ingresos por ventas mensuales en miles de dólares para cinco meses es de 56, 67, 52, 45, y 67. La media o promedio se calcula

$$\overline{x} = \frac{56 + 67 + 52 + 45 + 67}{5} = 57.4$$

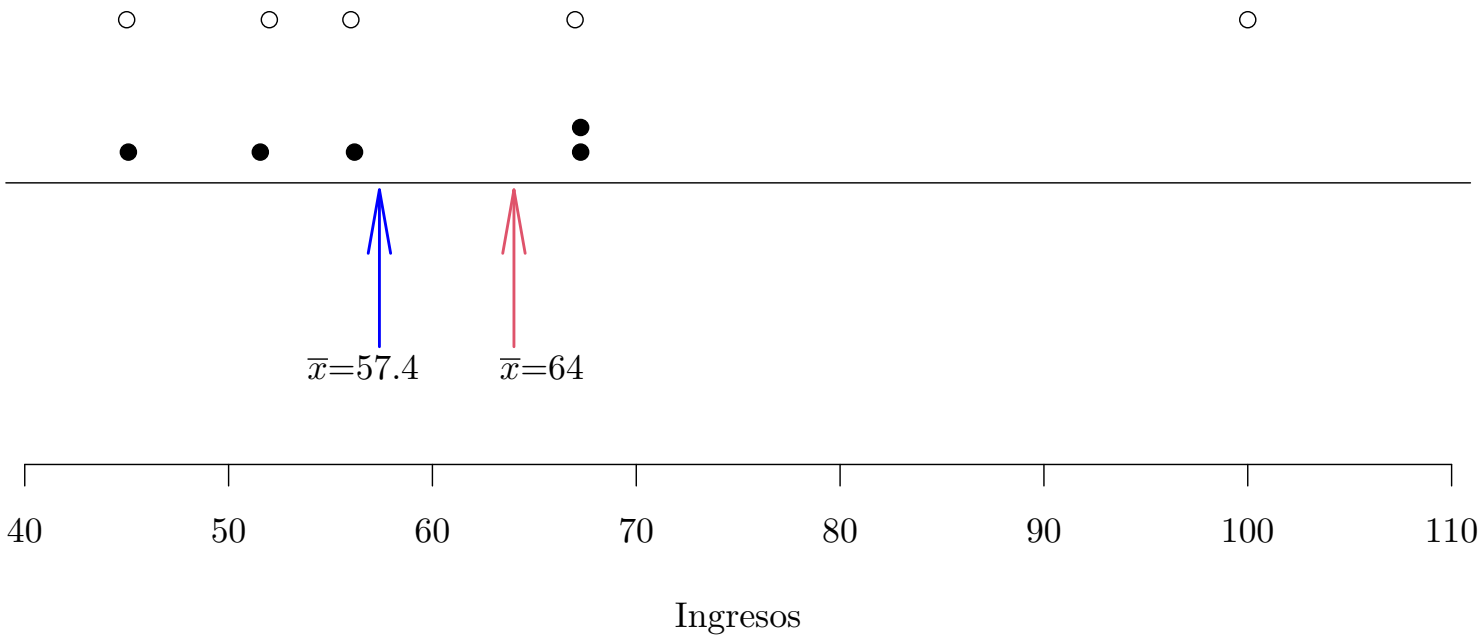
Si la última observación fuera 100 en lugar de 67, la media sería

$$\overline{x} = \frac{56 + 67 + 52 + 45 + 100}{5} = 64$$

El código en `R` para calcular la media es

```
X=c(56,67,52,45,67)
mean(X)
```

Medidas de tendencia central



Algunos inconvenientes de la media

Remark 4

Observe que la media muestral $\overline{x}$ representa el valor promedio de todas las observaciones en la muestra. También es posible pensar en el cálculo del valor promedio poblacional, es decir;

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N},$$

conocida como media poblacional.

Medidas de tendencia central

- **Mediana: Datos no ordenados** Valor numérico que verifica que sus datos ordenados de menor a mayor, el 50 % son menores o igual a él y el otro 50 % restante son mayores o iguales a él.
  Se denota Me.
  El cálculo de Me es como sigue:
  1). Ordene sus datos $\{x_1, \ldots, x_n\}$ de menor a mayor $x_{(1)} < \cdots < x_{(n)}$.
  2). Si n es impar, entonces $Me = x_{((n+1)/2)}$.
  Si n es par, entonces $Me = 1/2 \cdot \left(x_{(n/2)} + x_{((n/2)+1)}\right)$
- Observe que la mediana toma en cuenta el orden de los datos y no su magnitud.
  Por esto, es una medida más robusta que la media frente a datos extremos.

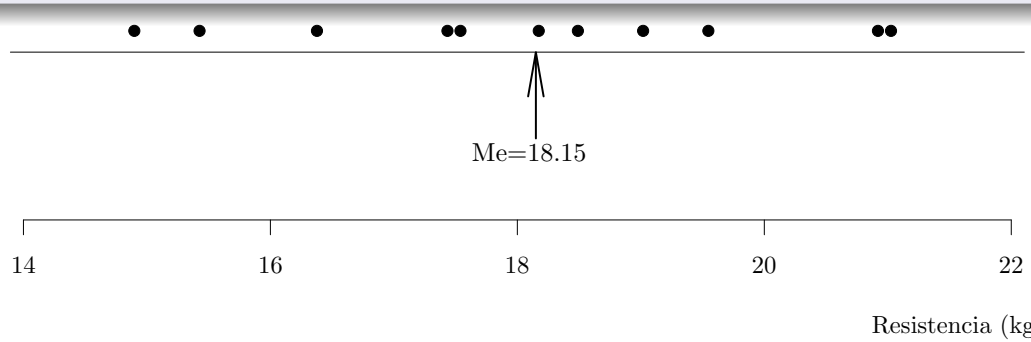
Example 1.4

Calcular la mediana de la resistencia a la tensión del mortero con polímero de látex: Los datos ordenados son:

$$14.95, 15.40, 16.36, 17.45, 17.57, 18.15, 18.52, 19.04, 19.59, 20.96, 21.07,$$

luego la mediana se calcula

$$Me = x_{((n+1)/2)} = x_{(6)} = 18.15$$



Medidas de tendencia central

Example 1.5

Se supone que una muestra de los ingresos por ventas mensuales en miles de dólares para cinco meses es de 56, 67, 52, 45, y 67. Los datos ordenas son:

$$45, 52, 56, 67, 67.$$

Luego la mediana se calcula

$$Me = x_{((n+1)/2)} = x_{(3)} = 56$$

Esto significa que en la mitad de los meses las ventas estuvieron por debajo de US$56.000, y en la mitad de los meses los ingresos excedieron dicha suma. En R `median()`

Remark 5

La media muestral en este caso es $\overline{x} = 57.4$.

Medidas de tendencia central

Example 1.6

Si la última observación fuera 100 en lugar de 67. Entonces

$$45, 52, 56, 67, 100.$$

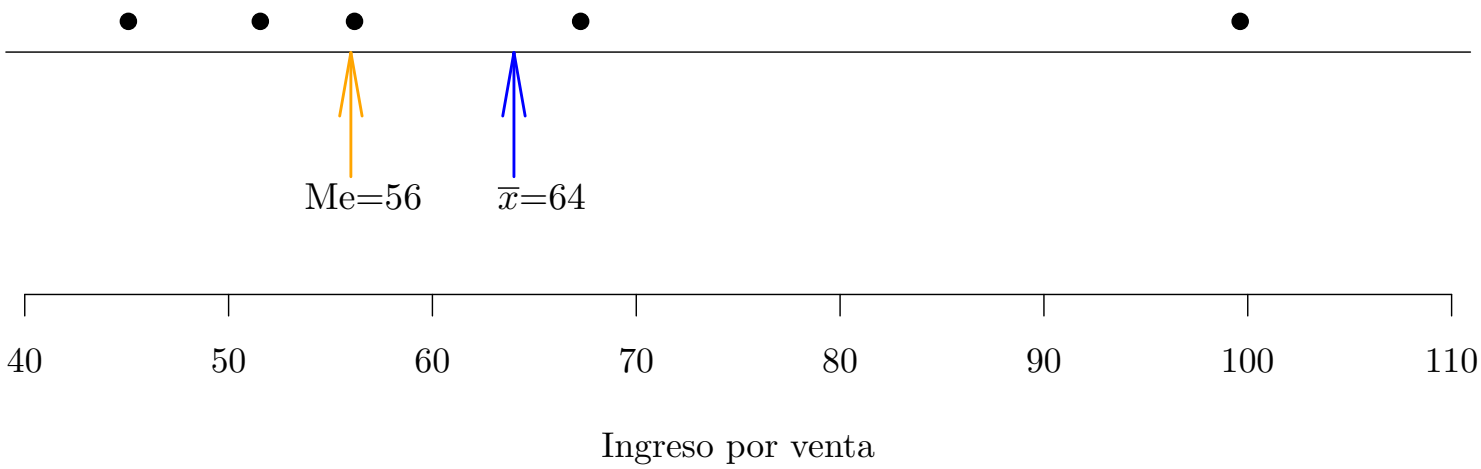
Luego la mediana se calcula

$$Me = x_{((n+1)/2)} = x_{(3)} = 56$$

Remark 6

La media muestral en este caso es $\overline{x} = 64$. Dado que la mediana no se ve afectada por este valor extremo, representa mejor las cinco observaciones

Medidas de tendencia central



Medidas de tendencia central

- **Mediana: Datos ordenados:** Sea $[L^*_{j-1}, L^*_j)$ el intervalo donde hemos encontrado que por debajo están el 50 % de las observaciones. Entonces se obtiene **la mediana** a partir de las frecuencias absolutas acumuladas:

$$\boxed{M_{ed} = L^*_{j-1} + \frac{\frac{n}{2} - N_{j-1}}{n_j} \cdot b^*_j}$$

Esto equivale a decir que la mediana divide al histograma en dos partes de areas iguales a $\frac{1}{2}$.

Propiedades de la Mediana

Remark 7

Como medida descriptiva, tiene la ventaja de no estar afectada por las observaciones extremas, ya que no depende de los valores que toma la variable, sino del orden de las mismas. Por ello es adecuado su uso en distribuciones asimétricas [a].

[a] Este concepto será analizado al final de esta clase

Medidas de tendencia central

- **Moda:** Valor más frecuente entre las observaciones (datos).
  Se denota por Mo.
- *Cómo calcular la moda con datos agrupados en intervalos?*
  - Sea $[L^*_{j-1}, L^*_j)$ el intervalo modal (aquel intervalo donde se encuentra la moda) y sea $b^*_j = L^*_j - L^*_{j-1}$ su amplitud.
  - Denote $n_{(j)}$ a la frecuencia absoluta del intervalo modal.
  - Con esto,

$$\boxed{Mo = L^*_{j-1} + \frac{n_j - n_{(j-1)}}{(n_j - n_{(j-1)}) + (n_j - n_{(j+1)})} \cdot b^*_j}$$

Propiedades de la Moda

Remark 8 (8)

Puede no ser única. Por ejemplo, considera las siguientes mediciones de temperaturas

$$1, 1, 1, 2, 2, 2, 2, 3, 4, 5, 5, 6, 6, 6, 7, 7, 7, 7, 7, 8, 10, 10, 10, 10, 10$$

Example 1.7 (3)

Como estadístico de una aerolínea se le solicita recopilar y agrupar los datos sobre el número de pasajeros que han decidido viajar con L&P (en miles). Los datos correspondientes a los últimos 21 días aparecen en la tabla

58	42	51	54	40	39	49
56	58	57	59	63	58	66
70	72	71	69	70	68	64

$$\begin{aligned} k &= \sqrt{21} = 4.58 \approx 5 \\ IC &= \frac{72-39}{5} = \frac{33}{5} = 6.6 \approx 7 \end{aligned}$$

Medidas de tendencia central

Example 1.8 (Cont.)

Clases	Marca de clase	Frecuencia absoluta	Frecuencia relativa	Frec. Abs. Acum.	Frec. Rel. Acum.
$[38, 45)$	*41,5*	*3*	*0,1428*	*3*	*0,1428*
$[45, 52)$	*48,5*	*2*	*0,0952*	*5*	*0,2381*
$[52, 59)$	*55,5*	*7*	*0,3333*	*12*	*0,5714*
$[59, 66)$	*62,5*	*3*	*0,1428*	*15*	*0,7143*
$[66, 73]$	*69,5*	*6*	*0,2857*	*21*	≈ 1
		21	≈ 1		

Determine las medidas de tendencia central e interprete.

La **media** aritmetica es:

$$\overline{x} = \frac{41.5 \cdot 3 + ... + 69.5 \cdot 6}{21} = 57.83$$

(Interprete)

La **mediana** es el valor de la variable que deja por debajo de si a la mitad de las n observaciones. La primera frecuencias absoluta acumulada que supera el valor $n/2 = 11.5$ es $N_i = 12$ vemos que eso ocurre en la tercera clase, es decir:

$$\begin{aligned} j &= 3 \quad \text{Observación} \\ [L_2^*, L_3^*) &= [52, 59) \quad \text{Intervalo donde se encuentra la mediana} \\ M_{ed} &= 52 + \frac{\frac{21}{2} - 5}{7} \cdot 7 = 57.5 \in [L_{j-1}, L_j) \end{aligned}$$

(Interprete)

Para el cálculo de la **moda**, lo primero es encontrar los intervalos modales, buscando los máximos relativos en la columna de las frecuencias absolutas, n_j. Vemos que corresponde a la clase $j = 3$. Asi el intervalo modal es: $[L_2^*, L_3^*) = [52, 59)$, la moda se calcula como:

$$\begin{aligned} Mo &= L_{j-1}^* + \frac{n_j - n_{(j-1)}}{(n_j - n_{(j-1)}) + (n_j - n_{(j+1)})} \cdot b_j^* \\ &= 52 + \frac{5}{(4+5)} \cdot 7 = 55.89 \end{aligned}$$

(Interprete)

Estadísticos de posición

Definition 1.1

Los estadísticos de posición van a ser valores de la variable caracterizados por superar a cierto porcentaje de observaciones en la población o muestra. Tenemos fundamentalmente a los **percentiles** *como medidas de posición, y asociados a ellos veremos también los* **cuartiles** *y* **deciles**.

Otras medidas de tendencia central

Definition 1.2 (**Percentil)**

El 100 k-ésimo percentil P_k es un valor tal, que al menos el 100k % de las observaciones están en el valor o por debajo de él, y al menos el 100(1-k) % están en el valor o por encima de él

Notemos que:

$$M_{ed} = P_{50}$$

- *Cálculo de percentiles:*

$$P_k = \begin{cases} x_{([(n\,k)/100]+1)} & \text{si } (n\,k\,\%) \text{ no es un entero} \\ \dfrac{x_{((n\,k)/100)} + x_{((n\,k)/100+1)}}{2} & \text{si } (n\,k\,\%) \text{ es un entero} \end{cases}$$

- $P_0 = \text{mín}\{x_1, \ldots, x_n\}$.
- $P_{100} = \text{máx}\{x_1, \ldots, x_n\}$.

Otras medidas de tendencia central

- **Cuartiles:** Percentiles que dividen a la distribución en cuatro partes iguales.
  - $P_{25} = Q_1$ es el primer cuartil.
  - $P_{50} = Q_2 = M_{ed}$ es el segundo cuartil, o sea la mediana.
  - $P_{75} = Q_3$ es el tercer cuartil.

Otras medidas de tendencia central

Example 1.9

A continuación se presentan 20 observaciones en orden de falla, en horas, de un material aislante eléctrico

204	228	252	300	324	444	624	720	816	912
1176	1296	1392	1488	1512	2520	2856			
3192	3528	3710							

Encuentre Q_1, Q_2 *y* Q_3.

Otras medidas de tendencia central

- **Quintiles:** Percentiles que dividen a la distribución en 5 partes iguales.
  - P_{20} es el primer quintil.
  - $\vdots \quad \vdots$
  - P_{80} es el cuarto quintil.
- **Deciles:** Percentiles que dividen a la distribución en 10 partes iguales.
  - P_{10} es el primer decil.
  - $\vdots \quad \vdots$
  - P_{90} es el noveno decil.

Otras medidas de tendencia central

- *Cómo calcular P_p si los datos están agrupados en intervalos?*
  - Sea $[L^*_{i-1}, L^*_i)$ el intervalo en el cual se encuentra el se encuentra P_p. Sea $b^*_i = L^*_i - L^*_{i-1}$ su amplitud.
  - Sea $n_{(i)}/n$ la frecuencia relativa del intervalo $[L^*_{i-1}, L^*_i)$.
  - Sea $N_{(i-1)}/n$ la frecuencia relativa acumulada del intervalo inmediatamente anterior a $[L^*_{i-1}, L^*_i)$.
  - Con esto,

$$\boxed{P_k = L^*_{i-1} + \frac{\frac{nk}{100} - N_{(i-1)}}{n_{(i)}} \cdot b^*_i}$$

Example 1.10

Calcular los cuartiles en la siguiente distribución de una variable continua resumida en la siguiente tabla estadística:

$[l_{i-1} \quad l_i)$	n_i	N_i
0 - 1	10	10
1 - 2	12	22
2 - 3	12	34
3 - 4	10	44
4 - 5	7	51
	$n = 51$	

Example 1.11 (Cont.)

Solución: Sea X la variable aleatoria continua en estudio.

- *Primer cuartil ($Q_1 = x_{(\frac{n}{4})}$): $\frac{n}{4} = 12.75$ buscar la Primera f.a.a $N_i > n/4$, es decir, $N_2 = 22$. la clase $i = 2$ es la del intervalo $[1, 2)$.*

$$Q_1 = l^*_{i-1} + \frac{\frac{n}{4} - N_{(i-1)}}{n_{(i)}} \times b^*_i = 1 + \frac{12.75 - 10}{12} \times 1 = 1.23$$

Otras medidas de tendencia central

Example 1.12 (Cont.)

- *Segundo cuartil* ($Q_2 = x_{(\frac{n}{2})}$)

$$Q_2 = l^*_{i-1} + \frac{\frac{n}{2} - N_{(i-1)}}{n_{(i)}} b^*_i = 2 + \frac{25.5 - 22}{12} \times 1 = 2.29$$

- *tercer cuartil* ($Q_3 = x_{(\frac{3n}{4})}$)

$$Q_3 = l^*_{i-1} + \frac{\frac{3\,n}{4} - N_{(i-1)}}{n_{(i)}} b^*_i = 3 + \frac{38.25 - 34}{10} \times 1 = 3.425$$

Ejemplo 6, Clase 1

Example 1.13

Calcular los cuartiles de los datos tabulados en el ejemplo 6, clase 1 (Interprete los resultados) y compare los valores con el boxplot.

	Clases	Marca de clase	Frecuencia absoluta	Frecuencia relativa	Frec. Abs. Acum.	Frec. Rel. Acum.
1	[50-51.1)	50.55	4	0.08	4	0.08
2	[51.1-52.2)	51.65	5	0.10	9	0.18
3	[52.2-53.3)	52.75	9	0.18	18	0.36
4	[53.3-54.4)	53.85	12	0.24	30	0.60
5	[54.4-55.5)	54.95	10	0.20	40	0.80
6	[55.5-56.6)	56.05	5	0.10	45	0.90
7	[56.6-57.7)	57.15	3	0.06	48	0.96
8	[57.7-58.8)	58.25	2	0.04	50	1.00

Cuadro: *Tabla de frecuencias agrupada con intervalos de clase.*

Medidas de posición

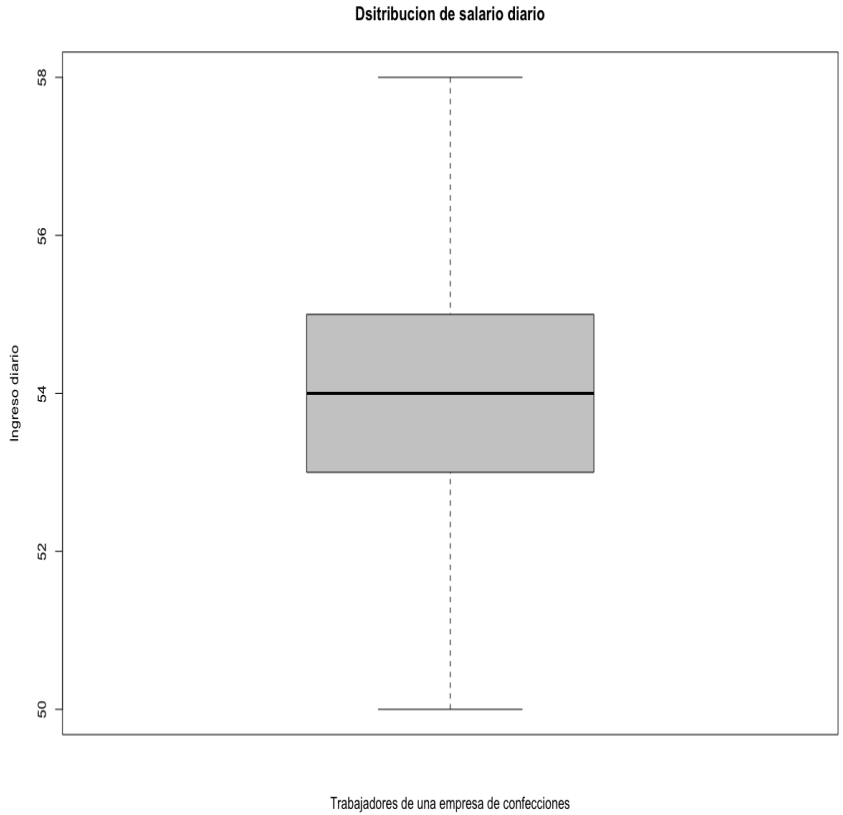
Example 1.14 (6 Cont.)

- *Primer cuartil* $Q_1 = P_{25} = 52.63$
- *Segundo cuartil* $Q_2 = P_{50} = 53.94$
- *Tercer cuartil* $Q_3 = P_{75} = 55.23$

En R

```
quantile(salario, p=c(0.25,0.50,0.75))
boxplot(salario, col="gray80",
    main="Dsitribucion de salario diario",
    xlab="Trabajadores de una empresa de confecciones",
    ylab="Ingreso diario")
```

Boxplot

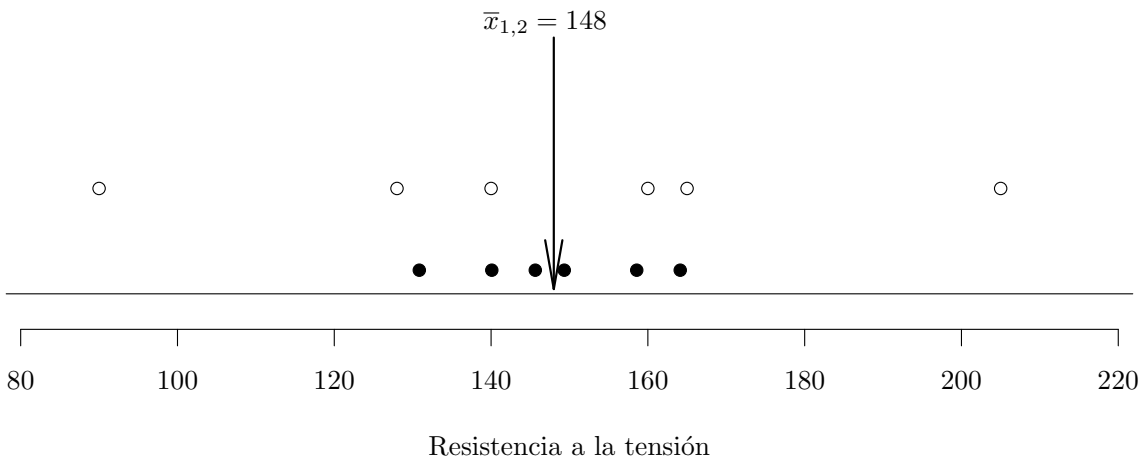

Dsitribucion de salario diario
Ingreso diario
50
52
54
56
58
Trabajadores de una empresa de confecciones

Medidas de dispersión

La localización o tendencia central no necesariamente proporciona información suficiente para describir los datos de manera adecuada. Por ejemplo, considere los datos de resistencia a la tensión (en libras por pulgada cuadrada (psi)) de dos muestras de aleación de aluminio-litio:

Muestra 1: $130, 150, 145, 158, 165, 140$

Muestra 2: $90, 128, 205, 140, 165, 160$



Medidas de dispersión

Las medidas de dispersión miden el grado de dispersión (variabilidad) de los valores de la variable.

- **Rango ó Recorrido** El rango o amplitud o recorrido se define como:

$$R = x_{(n)} - x_{(1)}$$

Example 1.15

Para el par de muestras en las que se media la resistencia a la tensión, los recorridos son

$$\begin{aligned} R_1 &= 165 - 130 = 35 \\ R_2 &= 205 - 90 = 115 \end{aligned}$$

Esta medida es sensible si en los datos existen valores extremos, esto se puede solucionar calculando:

- **Longitud intervalo intercuantilico** se define tambien como recorrido intercuantilico es dado por:

$$IQR = Q_3 - Q_1$$

contiene el 50 % central de la población o muestra.

Example 1.16

Calcule los IQR para las dos muestras de resistencia a la tensión de aleación de aluminio-litio

$$\begin{aligned} IQR_1 &= 158 - 140 = 18 \\ IQR_2 &= 165 - 128 = 37 \end{aligned}$$

El IQR es menos sensible a los valores extremos de la muestra, que el rango muestral.

Medidas de dispersión

- **Rango interdecilico** tambien puede utilizarse el rango interdecilico

$$D_9 - D_1$$

que sera la longitud del intervalo que contiene el 80 % central de la población o muestra.

Medidas de dispersión

Las principales medidas de variabilidad tienen que ver con las **desviaciones a partir de la media**, es decir $x_1 - \overline{x}, x_2 - \overline{x}, \ldots, x_n - \overline{x}$.

Example 1.17

Muestre que

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

Medidas de dispersión

- **Desviación media** Se define la desviación media como la media de las diferencias en valor absoluto de los valores de la variable a la media, es decir, si tenemos un conjunto de n observaciones, $x_1, ..., x_n$, entonces

$$\boxed{D_M = \frac{1}{n}\sum_{i=1}^{n} |x_i - \overline{x}|}$$

Si los datos estan agrupados

$$\boxed{D_M = \frac{1}{n}\sum_{j=1}^{m} |x_j - \overline{x}|\, n_j}$$

Medidas de dispersión

- **Varianza Poblacional y desviación típica**
  La varianza poblacional, σ^2, se define como la media de las diferencias cuadráticas de n datos con respecto a su media aritmética, es decir

$$\boxed{\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

  Para datos agrupados en tablas, usando las notaciones establecidas anteriormente, la varianza se puede escribir como

$$\boxed{\sigma^2 = \frac{1}{N}\sum_{j=1}^{m}(x_j - \mu)^2 n_j}$$

Medidas de dispersión

- **Varianza muestral y desviación estándar muestral**
  La varianza muestral, $\mathcal{S}^2$, se define como la media de las diferencias cuadráticas de n datos con respecto a su media aritmética, es decir

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

  Para datos agrupados en tablas, usando las notaciones establecidas anteriormente, la varianza se puede escribir como

$$\boxed{S^2 = \frac{1}{n-1}\sum_{j=1}^{m}(x_j - \overline{x})^2\, n_j}$$

Medidas de dispersión

- **Varianza muestral y desviación estándar muestral**
  Una fórmula equivalente para el cálculo de la varianza es:

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

  Si los datos están agrupados en tablas la formula equivalente es:

$$\mathcal{S}^2 = \frac{1}{n-1}\left(\sum_{j=1}^{m} x_j^2\, n_j - n\overline{x}^2\right)$$

Medidas de dispersión

La varianza no tiene la misma magnitud que las observaciones (ej. si las observaciones se miden en metros, la varianza lo hace en metros^2). Si queremos que la medida de dispersión sea de la misma dimensionalidad que las observaciones bastará con tomar su raíz cuadrada. Por ello se define la desviación estándar muestral, $\mathcal{S}$, como:

$$\boxed{\mathcal{S} = \sqrt{\mathcal{S}^2}}$$

La desviación estándar muestral muestral se representa por S ó SD.

Example 1.18

La siguiente tabla muestra las cantidades necesarias para calcular $\mathcal{S}^2$ y $\mathcal{S}$ de la muestra 2-1 de resistencia de la aleación de aluminio-litio.

i	x_i	x_i^2	x_i	x_i^2
1	90	8.100	130	16.900
2	128	16.384	150	22.500
3	205	42.025	145	21.025
4	140	19.600	158	24.964
5	165	27.225	165	27.225
6	160	25.600	160	19.600
	$\sum_{i=1}^{6} x_i = 888$	$\sum_{i=1}^{6} x_i^2 = 138.934$	$\sum_{i=1}^{6} x_i = 888$	$\sum_{i=1}^{6} x_i^2 = 132.214$

$$
\begin{aligned}
S^2 &= \frac{\sum_{i=1}^{6} x_i^2 - \frac{\left(\sum_{i=1}^{6} x_i\right)^2}{6}}{5} = \frac{138.934 - \frac{888^2}{6}}{5} = 1.502(psi)^2 \\
S_2 &= \sqrt{(1.502)} = 38.8psi. \\
S_1 &= \sqrt{(158)} = 12.57psi
\end{aligned}
$$

Tipificación

- **Tipificación**
  Se conoce por tipificación al proceso de restar la media y dividir por su desviación típica a una variable X. De este modo se obtiene una nueva variable.

$$\boxed{Z = \frac{X - \overline{x}}{\mathcal{S}}}$$

de media $\overline{z} = 0$ y desviación típica $\mathcal{S}_Z = 1$, que denominamos variable tipificada. Esta nueva variable carece de unidades y permite hacer comparables dos medidas que en un principio no lo son, por aludir a conceptos diferentes.

Coeficiente de variación

- **Coeficiente de variación**
  Para comparar el grado de dispersión entre dos o más distribuciones expresadas en distintas unidades de medida, no podemos comparar simplemente las varianzas ó las desviaciones estándar respectivas. El coeficiente de variación tiene en cuenta la proporción existente entre medias y desviación típica. Se define del siguiente modo:

$$\boxed{\mathcal{CV} = \frac{\mathcal{S}_X}{|\overline{x}|} 100\,\%}$$

Otras características

- **Asimetría.** Uno dice que una distribución es simétrica si al graficar distribución de frecuencias y trazar una perpendicular al eje de abcisas por $\overline{x}$ hay el mismo número de valores a ambos lados de la perpendicular, equidistantes de $\overline{x}$ dos a dos y tales que cada par de valores equidistantes a $\overline{x}$ tienen la misma frecuencia.
- **Coeficiente de asimetría (g1)**. Mide si la muestra está igualmente distribuída alrededor de la media.

$$g1 = \frac{1}{(n-1)s^3} \sum_{i=1}^{n} (x_i - \overline{x})^3$$

  - $g1 < 0$ indica asimetría hacia la izquierda de la media.
  - $g1 > 0$ indica asimetría hacia la derecha de la media.
  - $g1 = 0$ indica simetría entorno a la media.

Otras características

- **Coeficiente de curtosis (g2)**. Mide la concentración de los datos alrededor de la media. Uno aplica esta medida a distribuciones simétricas, tomando como referencia a la distribución $N(\mu, \sigma^2)$.

$$g2 = \frac{1}{(n-1)s^4} \sum_{i=1}^{n} (x_i - \overline{x})^4 - 3$$

  - $g2 < 0$ indica histograma más aplastado que el de la distribución normal (distribución platicúrtica).
  - $g1 > 0$ indica histograma menos aplastado que el de la distribución normal (distribución leptocúrtica).
  - $g1 = 0$ indica histograma igual de aplastado que el de la distribución normal (distribución mesocúrtica).

Ejemplo 7

Example 1.19 (7)

El director de vuelo de P&P requiere información respecto a la dispersión del número de pasajeros. Las decisiones que se tomen al respecto a la programación y al tamaño más eficiente de los aviones, dependerá de la fluctuación en el transporte de pasajeros.

	Clases	c_k	n_k	$c_k \cdot n_k$	c_k^2	$c_k^2 \cdot n_k$
1	[50-60)	55	3	165	3025	9075
2	[60-70)	65	7	455	4225	29575
3	[70-80)	75	18	1350	5625	101250
4	[80-90)	85	12	1020	7225	86700
5	[90-100)	95	8	760	9025	72200
6	[100-110]	105	2	210	11025	22050
		n=50		$\sum_{k=1}^{6} c_k n_k = 3.960$		$\sum_{k=1}^{6} c_k^2 n_k = 320.850$

Cuadro: *Tabla de frecuencias agrupada con intervalos de clase.*

Ejemplo 7

Example 1.20 (Cont.7)

La media $\overline{X}$ *es dada por*

$$\overline{X} = \frac{\sum_{k=1}^{6} c_k n_k}{n} = \frac{3.960}{50} = 79.2$$

Podemos ver que P&P transportó un promedio diario de 79.2 pasajeros. Por toro lado, la varianza muestral es dada por

$$\begin{aligned} S^2 &= \frac{\sum_{k=1}^{6} c_k^2 n_k - n\overline{X}^2}{n-1} \\ &= \frac{320.850 - 50(79.2)^2}{49} \\ &= 147.3061 \textit{pasajeros elevado al cuadrado} \end{aligned}$$

Ejemplo 7

Example 1.21 (Cont.7)

Dado lo anterior la desviación estandard es dada por

$$\begin{aligned} S &= \sqrt{S^2} \\ &= \sqrt{147.3061} \\ &= 12.13697 \textit{pasajeros} \end{aligned}$$

El director de vuelo ahora puede decidir si los aviones que se están usando actualmente pueden acomodar fluctuaciones en los niveles de pasajeros tal como lo mide la desviación estándar de 12.14.

Ejemplo 7

Example 1.22 (Cont.7)

El director desea abrir una nueva ruta, para ello desea saber con qué frecuencia los pasajeros están dentro de dos SD de la media, y cuál es dicho intervalo.

SOLUCIÓN

El 75 % de los datos se encuentra dentro del intervalo $[\overline{X} - 2 \cdot SD; \overline{X} + 2 \cdot SD]$, *por lo menos el 75 % de los dias (37 dias), el número de pasajeros estuve entre*

$$[79.2 - 2(12.13697); 79.2 + 2(12.13697)] = [54.92606, 103.4739]$$

Así, el director puede estar seguro que por lo menos en 37 dias, el número de pasajeros estuve entre 55 y 103.

Ejemplo 7

Example 1.23 (Cont.7)

Theorem 1.1 (Teorem de Chebyshev)

Establece que para todo conjunto de datos, por lo menos $(1 - \frac{1}{K^2})\,\%$ *de las observaciones están dentro de* K *SD de la media, en donde* K *es cualquier número mayor que 1, es decir, el* $(1 - \frac{1}{K^2})\,\%$ *de los datos se encuentra en el siguiente intervalo*

$$[\overline{X} - K \cdot SD; \overline{X} + K \cdot SD]$$

Ejemplo 7

Example 1.24 (Cont.7)

Calcule la mediana.

	Clases	c_k	n_k	N_k
1	[50-60)	55	3	3
2	[60-70)	65	7	10
3	[70-80)	75	18	28
4	[80-90)	85	12	40
5	[90-100)	95	8	48
6	[100-110]	105	2	50

Cuadro: *Tabla de frecuencias agrupada con intervalos de clase.*

Se debe encontrar el intervalo que contenga el Q_2. *Para ello* $\frac{n}{2} = \frac{50}{2} = 25$ *buscar la Primera f.a.a* $N_j > n/2$, *es decir,*$N_3 = 28$. *la clase* $j = 3$ *es la del intervalo* $[70, 80)$.

Otras medidas de tendencia central

Example 1.25 (Cont. 7)

$$
\begin{aligned}
P_{50} &= Q_2 = M_{ed} = l_{j-1} + \frac{\frac{n}{2} - N_{j-1}}{n_j} b_j \\
&= 70 + \frac{25 - 10}{18} \cdot 10 = 78.3333
\end{aligned}
$$

Se puede concluir que en 25 dias, menos de 78.33 pasajeros volaron en P&P, y en los otros 25 dias, más de 78.33 pasajeros volaron con P&P Airlines.

Example 1.26 (Cont. 7)

Calcule el coeficiente de asímetria.

$$\begin{aligned} g1 &= \frac{3(\overline{X} - M_{ed})}{SD} \\ &= \frac{3(79.2 - 78.33)}{12.13697} = 0.03 \end{aligned}$$

Dado que $g1 > 0$, *los datos sesgado a la derecha.*

Otras medidas de tendencia central

Example 1.27 (Cont. 7)

Suponga que la aerolínea recolecta en el mismo período el número de kilómetros que la aerolínea voló y dicha media y desviación estándar es $12,650$ *y* $1,530$*, respectivamente. Compare ambos registros de datos.*

Contenido

Diagramas de caja o box-plot

- Representación semi-gráfica de una distribución.
- Permite observar principales características de la distribución y también detecta valores atípicos.
- Si Ud. tiene datos $\{x_1, \ldots, x_n\}$, *cómo se construye un box-plot?*
  1). Ordene datos de menor a mayor $x_{(1)} < \cdots < x_{(n)}$ y obtenga los cuartiles Q_1, Q_2 y Q_3.
  2). Así, uno obtiene dos valores más:

$$\text{LI} = Q_1 - 1.5 \cdot (Q_3 - Q_1) \quad \text{(límite inferior)}$$
$$\text{LS} = Q_3 + 1.5 \cdot (Q_3 - Q_1) \quad \text{(límite superior)}$$

Diagramas de caja o box-plot

- Si Ud. tiene datos $\{x_1, \ldots, x_n\}$, *cómo se construye un box-plot? (cont.)*

  (3). Sitúe en un eje estos 5 valores; tomando como base el segmento $[Q_1, Q_3]$, dibuje un rectángulo de altura arbitraria. En él indique la posición de la mediana mediante una línea vertical que divida al rectángulo.

  (4). Desde el centro de los lados verticales del rectángulo, dibuje dos líneas horizontales. Una de ellas debe llegar hasta el mayor dato menor o igual a LI y la otra línea hasta el mayor dato menor o igual a LS.

  (5). Todo dato que se encuentre fuera del intervalo [LI, LS] márquelos con un asterisco $*$ a la altura de las líneas dibujadas en el ítem anterior.

Diagramas de caja o box-plot

Example 2.1 (Cont. 7)

Graficar el box-plot para el número de pasajeros

SOLUCIÓN

Primer cuartil ($Q_1 = x_{(\frac{n}{4})}$): $\frac{n}{4} = 12,5$ *buscar la Primera f.a.a* $N_j > n/4$, *es decir*, $N_3 = 28$. *la clase* $j = 3$ *es la del intervalo* $[70, 80)$.

$$Q_1 = l_{j-1} + \frac{\frac{n}{4} - N_{j-1}}{n_j} \times b_j = 70 + \frac{12,5 - 10}{18} \times 10 = 71.38889$$

Diagramas de caja o box-plot

Example 2.2 (Cont. 7)

Tercer cuartil $Q_3 = x_{(\frac{3n}{4})} = x_{(37.5)}$

$$\mathcal{Q}_3 = L_{j-1} + \frac{\frac{3\,n}{4} - N_{j-1}}{n_j}\, b_j = 80 + \frac{37,5 - 28}{12} \times 10 = 87.91667$$

$$\begin{aligned} LI = Q_1 - 1.5 \cdot (Q_3 - Q_1) = & \quad 71.38889 - 1.5(16.52778) = 46.59722 \\ LS = Q_3 + 1.5 \cdot (Q_3 - Q_1) = & \quad 87.91667 + 1.5(16.52778) = 112.7083 \end{aligned}$$