Taller Iniciación Estadística Aplicada a la Investigación

Hugo J. Bello

2024/04

Estadística y la investigación en ciencias de la salud

- El primer médico que utilizó métodos matemáticos para cuantificar variables de pacientes y sus enfermedades fue el francés **Pierre Charles-Alexandre Louis** (1787-1872).
- **Pierre Simon Laplace** publicó en 1812 Théorie analytique des probabilités, sugiriendo que tal análisis podría ser una herramienta valiosa para resolver problemas médicos.
- Los primeros trabajos bioestadísticos en enfermería los realizó, a mediados del siglo xix la enfermera inglesa **Florence Nightingale**.
- los problemas conceptuales ligados a la comprensión de la relación entre la genética y el darwinismo condujeron a un acalorado debate entre biométricos (Weldon, **Pearson**) y mendelianos (Davenport, Bateson).
- En los años 30, **Ronald Fisher** desarrolló varios métodos básicos de la estadística en su libro The Genetical Theory of Natural Selection.

Estadística y la investigación en ciencias de la salud

- Durante las últimas décadas, las ciencias de la salud han experimentado un importante proceso de **cuantificación**: además del uso tradicional de información cualitativa, como puede ser el aspecto de una herida, o el estado general del enfermo, se ha aprovechado el desarrollode la tecnología para la determinación cantidades numéricas que pudieran tener alguna relación con la salud del paciente, como pueden ser la presión sanguínea, el nivel de glucosa en suero, etcétera
- Se generan por lo tanto grandes cantidades de información que requieren análisis exhaustivo: **un análisis estadístico**
- Una situación típica en la que resultan útiles los métodos estadísticos es en la comparación de dos o más tratamientos, o la comparación de dos o más grupos o poblaciones con relación a una característica de interés, relación entre dos factores o características de los pacientes...

Primeros pasos: Cómo trabajamos con datos

> *La* **estadística descriptiva** *estudia diversas técnicas útiles en la presentación y el resumen de un conjunto de datos.*

Cuando recogemos datos para una investigación o estudio

- Los datos se colocan filas, una fila por paciente.
- Las columnas representan **valiables estadísticas**
- Debemos incorporar los datos de manera ordenada y consistente.
  - decimales todos de la misma manera
  - si los datos faltan, dejar esa celda de la columna en blanco
  - ser consistentes con las variables no numéricas

Ejemplo: a evitar

paciente	peso	sexo	edad	obesidad	tiempo de atención
1	79.05	H	55.0	No	9.45
2	132.93	H	77.0	Si	266.05
3	79.6	Hombre	40.0	Si	9.13
4	48.99	M	64.0	No	3.05
5	72.82	M	69.0	No	13.38
6	89.57	H	22.0	Si	14.45
7	80.78	M	73.0	Si	5.02
8	61.37	H	35,0	n	12.77
9	80.96	H	47.0	Si	15.25
10	91.55	H	61.0	Si	ocho minutos

Ejemplo

paciente	peso	sexo	edad	obesidad	tiempo de atención
1	79.05	H	55.0	No	19.45
2	132.93	H	77.0	Si	266.05
3	79.6	H	40.0	Si	9.13
4	48.99	M	64.0	No	50.05
5	72.82	M	69.0	No	113.38
6	89.57	H	22.0	Si	14.45
7	80.78	M	73.0	Si	5.02
8	61.37	H	35.0	No	112.77
9	80.96	H	47.0	Si	15.25
10	91.55	H	61.0	Si	8.14

Síntesis de datos: principales estadísticos descriptivos

La estadística descriptiva diseña se ocupa que puedan resultar útiles en la presentación y el resumen de un conjunto de datos.
Una **variable estadística** es un aspecto del fenómeno estudiado que puede ser medido. Las hay de dos tipos:

- **Cuantitativas.** Peso, altura...
- **Cualitativas.** Sexo, enfermedad...

Medidas estadísticas

Estudiaremos *medidas estadísticas* también llamados *estadísticos*. Se trata de cálculos que tratan de capturar la información de los datos. Veremos:

- **Medidas de concentración**
- **Medidas de orden**
- **Medidas de dispersión**

Media

La **media** de unos datos $x_1, \ldots x_n$ se calcula usando

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \ldots + x_n)$$

Algunas propiedades

- Fácil de calcular y muy efectiva para ver el **lugar central de los datos**
- se ve afectada por valores extremos.

en el caso anterior si tomamos $x = \textit{tiempo de atención}$ vemos que

$$\overline{x} = 61.36$$

Esto nos da una idea del **tiempo de atención promedio**, y nos permite hacernos una idea de qué pasa con los datos.

Incluso podemos plantearnos separar los datos en dos grupos: **los que tienen obesidad y los que no**

Ejemplo

En los datos anteriores

paciente	peso	sexo	edad	obesidad	tiempo de atención
1	79.05	H	55.0	No	19.45
2	132.93	H	77.0	Si	266.05
3	79.6	H	40.0	Si	9.13
4	48.99	M	64.0	No	50.05
5	72.82	M	69.0	No	113.38
6	89.57	H	22.0	Si	14.45
7	80.78	M	73.0	Si	5.02
8	61.37	H	35.0	No	112.77
9	80.96	H	47.0	Si	15.25
10	91.55	H	61.0	Si	8.14

Ejemplo

En este caso:

- la media del tiempo de atención de los que presentan obesidad es de 53.00
- la de los que no tienen obesidad es de 73.91.

En vista a esto podríamos llegar a la conclusión de que **en presencia de obesidad el tiempo de atención es menor.**

Sin embargo fijándonos bien vemos un dato **sospechoso**

Ejemplo

[redacted]	peso	sexo	edad	enfermedad	tiempo de atención
1	79.05	H	55.0	No	9.45
2	132.93	H	77.0	Si	**266.05**
3	79.6	H	40.0	Si	9.13
4	48.99	M	64.0	No	50.05
5	72.82	M	69.0	No	113.38
6	89.57	H	22.0	Si	14.45
7	80.78	M	73.0	Si	5.02
8	61.37	H	35.0	No	112.77
9	80.96	H	47.0	Si	15.25
10	91.55	H	61.0	Si	8.14

El paciente 2 tiene un tiempo de atención de 266.05. Un valor inusualmente mayor que está haciendo que la media sea extremadamente elevada. Se trata de un dato **anómalo**

Para evitar este problema (la sensibilidad a datos extremos o anómalos) utilizamos la **mediana**.

Mediana

La **mediana** de unos datos $x_1, x_2, \ldots, x_n$ representa el valor de la variable de **posición central en un conjunto de datos ordenados**. Es el dato que que separa la mitad inferior y la mitad superior de los datos.

Para calcularla

- ordenamos los datos
- buscamos el que ocupa la posición central. Si no hay ninguno que ocupe exactamente esa posición (porque no hay una posición central como tal) tomamos la media de los dos que la ocupen

Se suele denotar por $Med(x)$. Y a diferencia de la media **No se ve afectada por valores extremos**

Mediana:Ejemplo sencillo

- Para la muestra $9, 3, 7, 6, 3, 8, 1$. Si los ordenamos tenemos

$$1, 3, 3, \mathbf{6}, 7, 8, 9$$

En este caso la mediana es 6

- Para la muestra $4, 9, 3, 7, 6, 3, 8, 1$. Si los ordenamos tenemos

$$1, 3, 3, \mathbf{4}, \mathbf{6}, 7, 8, 9$$

En este caso la mediana es $(4 + 6)/2 = 5$

Ejemplo

En los datos anteriores

paciente	peso	sexo	edad	obesidad	tiempo de atención
1	79.05	H	55.0	No	19.45
2	132.93	H	77.0	Si	266.05
3	79.6	H	40.0	Si	9.13
4	48.99	M	64.0	No	50.05
5	72.82	M	69.0	No	113.38
6	89.57	H	22.0	Si	14.45
7	80.78	M	73.0	Si	5.02
8	61.37	H	35.0	No	112.77
9	80.96	H	47.0	Si	15.25
10	91.55	H	61.0	Si	8.14

- en el caso anterior si tomamos $x = \textit{tiempo de atención}$. Los datos de la columna ordenados son

5.02, 8.14, 9.13, 14.45, **15.25**, **19.45**, , 50.05, 112.77, 113.38, 266.05

vemos que

$$Med(x) = \frac{15.25 + 19.45}{2} = 17.35$$

- Si ahora nos planteamos las diferencias entre los que presentan una enfermedad vs los que no en términos de la mediana, repetiríamos esto para los dos grupos (tomando solo los datos de cada grupo y calculando la mediana igual que antes). El resultado es que:
  - El grupo con obesidad (amarillo) tiene mediana: **11.79**
  - El grupo sin obesidad grave tiene mediana **81.41**

Los valores de estas medianas contrastan con los valores de las medias anteriores. En el caso de la obesidad vemos como baja drásticamente al reducrse el efecto del valor extremo, además vemos que los valores de las medianas son más cercanos entre sí.

Las **medidas de orden o posición** se basan en ordenar los datos y buscar qué datos ocupan ciertas posiciones identificativas.

- El **percentil** k de una muestra, es el valor que, una vez ordenados los datos de menor a mayor, queda por encima del k por ciento de las observaciones. Por ejemplo, el percentil 20 es el valor bajo el cual se encuentran el 20 % de las observaciones, y el 80 % restante son mayores. Se denota por P_k
- Se calculan de forma similar a la mediana. Se ordenan los datos y se busca cual es que está en la posición $k\%$.
- Los **cuartiles** Q_1, Q_2, Q_3 son los percentiles $25, 50$ y 75 respectivamente. El segundo cuartil coincide con la mediana.

Usos y ejemplos

- Los percentiles se usan para medir la inteligencia respecto a tests estandarizados como alternativa del coeficiente de inteligencia.
- La desviación respecto a los cuartiles se utiliza para medir la variabilidad de presión arterial.
- Los cuartiles se utilizan para medir la dispersión a través de el rango intercuartílico y los diagramas de cajas.

Medidas de dispersión

Las **medidas de dispersión** miden cuánto se separa los datos de su lugar central.

Completan la información aportada por las medidas de concentración puesto que nos aportan como de irregulares o dispersas (disimilares) son las observaciones.

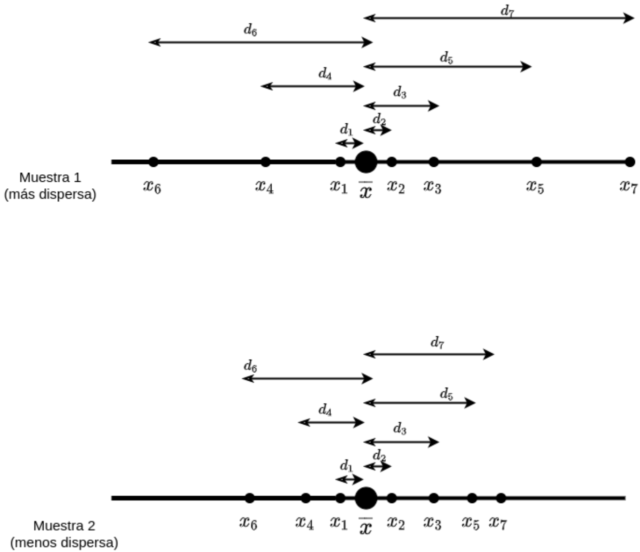
Varianza

La **varianza** de unos datos $x_1, \ldots, x_n$ se calcula mediante

$$S_x^2 = \frac{1}{n} \cdot \left((x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2 \right)$$
$$= \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

- Se mide en unidades al cuadrado (respecto de las de los datos). Aunque no se suelen poner.
- $(x_i - \overline{x})^2$ representa cuánto se desvía x_i respecto a la media.
- En algunos textos se define dividiendo entre $n - 1$ en vez de n

Medidas de dispersión



Desviación típica

La **desviación típica** de unos datos $x_1, \ldots, x_n$ se define simplemente como la raíz cuadrada de la varianza

$$\begin{aligned} S_x &= \sqrt{S_x^2} \\ &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} \end{aligned}$$

- Se mide en las mismas unidades que los datos.
- Se utiliza en el mismo contexto que la varianza, pero el uso de la raíz cuadrada la hace más interpretable.
- También se denota por σ.

Rango intercuartílico

El **rango intercuartílico** de unos datos $x_1, \ldots, x_n$ se se define como la diferencia entre el tercer y primer cuartil

$$IQR_x = Q_3 - Q_1$$

- Se mide en las mismas unidades que los datos.
- Se utiliza en el mismo contexto que las anteriores para determinar la dispersión. Al igual que ocurre con la mediana **No se ve tan afectada por valores extremos**

En el ejemplo que hemos venido utilizando para la variable $x =$ *tiempo de atención* vimos que la media resultaba $\overline{x} = 61.36$
Para calcular la **varianza** haríamos:

$$S_x^2 = \frac{1}{10}\left((9.45 - 61.36)^2 + (266.05 - 61.36)^2 + \ldots + (8.14 - 61.36)^2\right)$$
$$= 6219.26$$

Por otra parte la **desviación típica** resulta

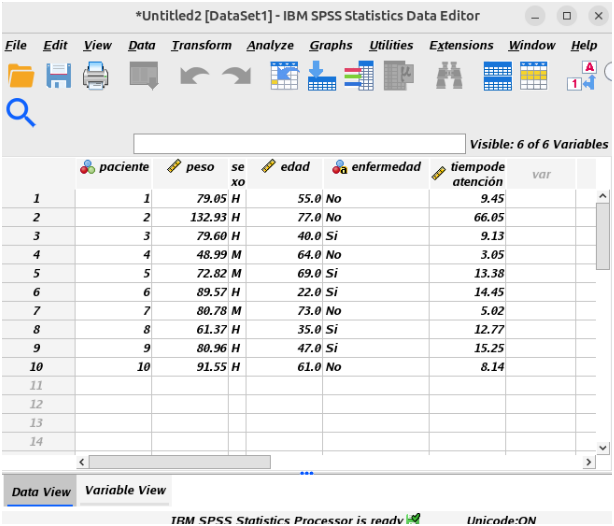
$$S_x = \sqrt{6219.26} = 78.86$$

En este caso los cuartiles son: $Q_1 = 10.46$, $Q_3 = 97.09$, por lo tanto $IQR_x = 97.09 - 10.46 = 86.63$

Al igual que hicimos antes, ahora nos planteamos las diferencias entre los que presentan una enfermedad vs los que no en términos de la **dispersion**, resultado es que:

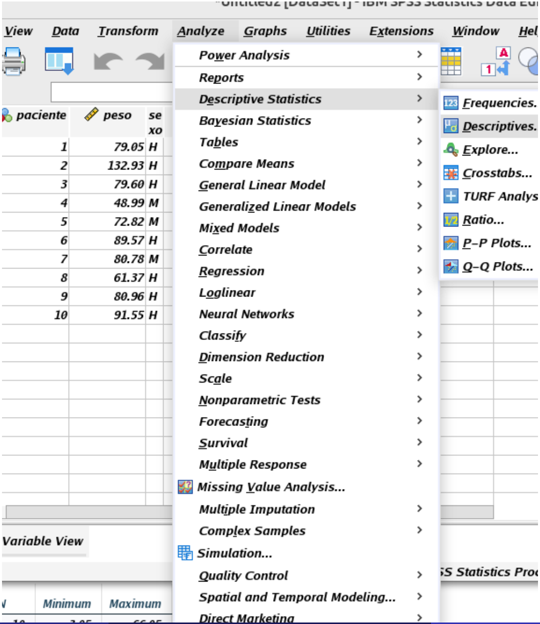
- El grupo con obesidad (amarillo) tiene:
  - **Varianza** $S_x^2 = 9090.09$
  - **Desviación típica** $S_x = 95.34$
  - **Rango intercuartílico** $IQR_x = 6.66$
- El grupo sin obesidad tiene
  - **Varianza** $S_x^2 = 1650.79$
  - **Desviación típica** $S_x = 40.62$
  - **Rango intercuartílico** $IQR_x = 70.52$

En este caso la varianza y desviación típica parecen indicar que el grupo sin obesidad tiene una dispersión mayor si tomamos en cuenta la varianza y desviación típica y una menor si tenemos en cuenta el rango intercuartílico. Esto se debe a que el rango intercuartílico no se ve afectado por el dato extremo que hemos encontrado y esto podría indicar que en realidad los **datos de los obesos están menos dispersos**



	paciente	peso	sexo	edad	enfermedad	tiempode atención	var
1	1	79.05	H	55.0	No	9.45	
2	2	132.93	H	77.0	No	66.05	
3	3	79.60	H	40.0	Si	9.13	
4	4	48.99	M	64.0	No	3.05	
5	5	72.82	M	69.0	Si	13.38	
6	6	89.57	H	22.0	Si	14.45	
7	7	80.78	M	73.0	No	5.02	
8	8	61.37	H	35.0	Si	12.77	
9	9	80.96	H	47.0	Si	15.25	
10	10	91.55	H	61.0	No	8.14	
11							
12							
13							
14							

Figure: Medias y desviaciones SPSS


View
Data
Transform
Analyze
Graphs
Utilities
Extensions
Window
Power Analysis
Reports
Descriptive Statistics
Bayesian Statistics
Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Neural Networks
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
Missing Value Analysis...
Multiple Imputation
Complex Samples
Simulation...
Quality Control
Spatial and Temporal Modeling...
Direct Marketing
Explore...
Crosstabs...
Ratio...
P-P Plots...
Q-Q Plots...
paciente
peso
1 79.05 H
2 132.93 H
3 79.60 H
4 48.99 M
5 72.82 M
6 89.57 H
7 80.78 M
8 61.37 H
9 80.96 H
10 91.55 H
Variable View
Minimum
Maximum

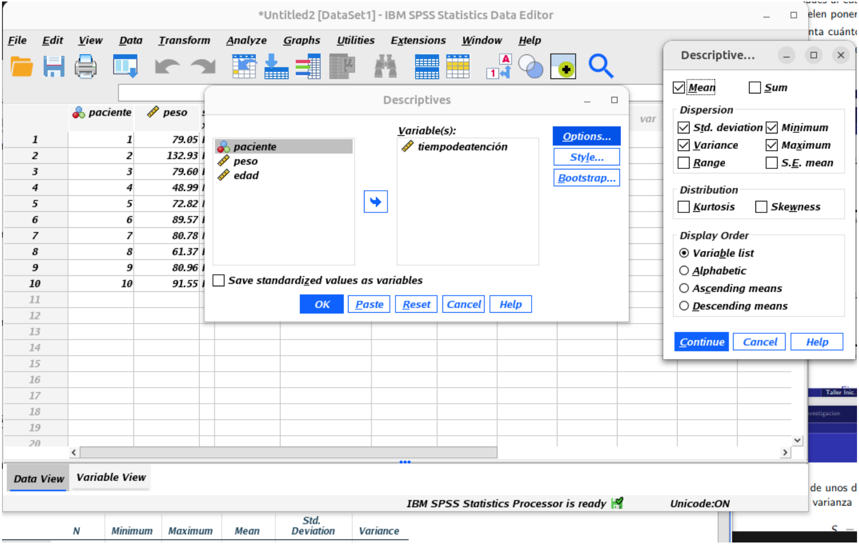
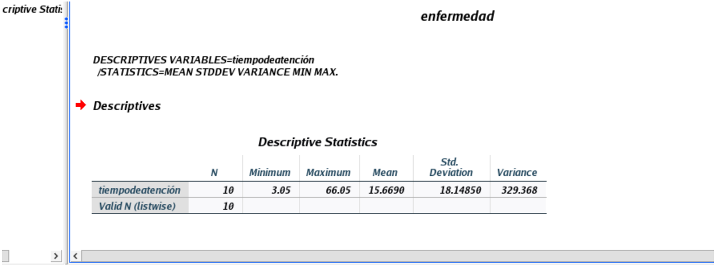


Figure: Medias y desviaciones SPSS



criptive Stati:

enfermedad

DESCRIPTIVES VARIABLES=tiempodeatención
/STATISTICS=MEAN STDDEV VARIANCE MIN MAX.

Descriptives

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Variance
tiempodeatención	10	3.05	66.05	15.6690	18.14850	329.368
Valid N (listwise)	10					

Figure: Medias y desviaciones SPSS

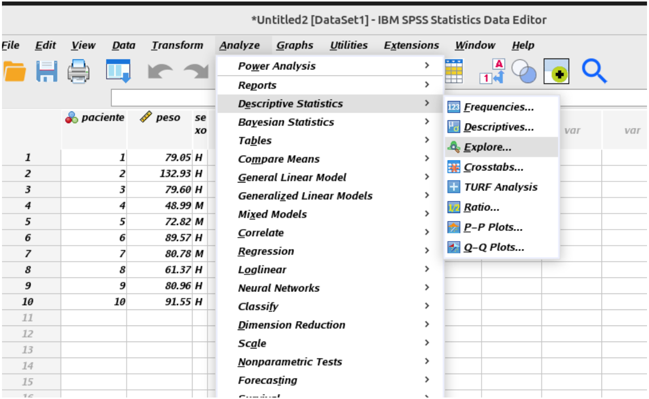


Figure: Percentiles / cuartiles SPSS

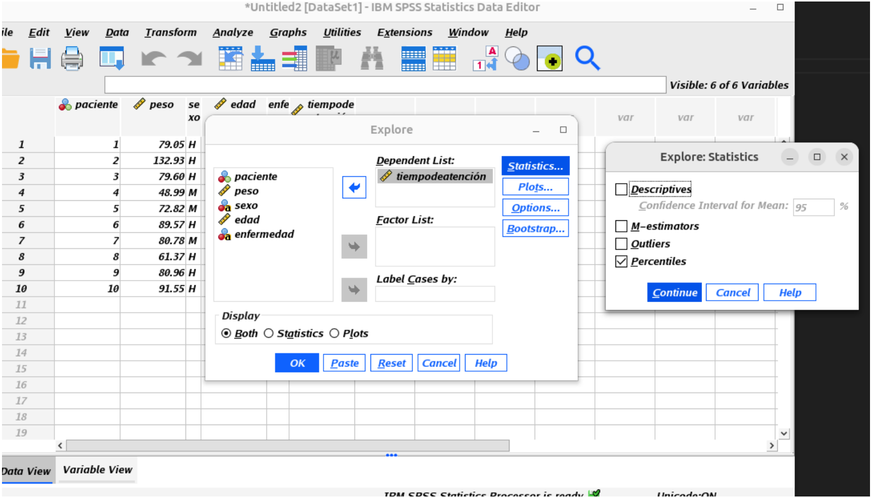


Figure: Percentiles / cuartiles SPSS

tiempodeatención	10	100.0%	0	0.0%	10	100.0%

Percentiles

		Percentiles						
		5	10	25	50	75	90	95
Weighted Average (Definition 1)	tiempodeatención	3.0500	3.2470	7.3600	11.1100	14.6500	60.9700	.
Tukey's Hinges	tiempodeatención			8.1400	11.1100	14.4500		

Figure: Percentiles / cuartiles SPSS