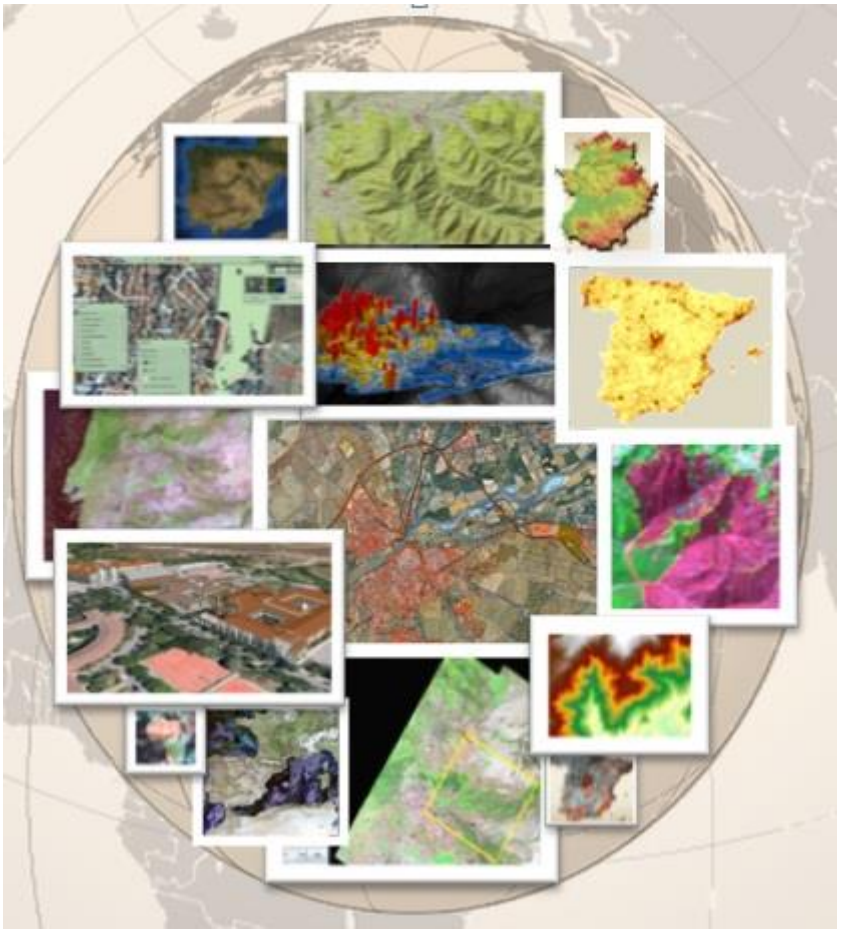
Máster en Tecnologías de la Información Geográfica: SIG y Teledetección



GEOESTADÍSTICA Y CALIDAD DE LA INFORMACIÓN

Herramientas estadísticas para el análisis de datos: datos escalares

María Eugenia Polo García

1. Datos y variables

Definimos un **dato** como un hecho verificable de la realidad. Los datos son la base de la mayoría de la actividad científica y de la práctica totalidad de la actividad tecnológica. Como no podremos manejarlos todos al mismo tiempo necesitaremos resumirlos en determinadas medidas. A estas medidas se les llama **descriptivas** y las iremos analizando a lo largo de este tema. Para eso necesitamos unos conocimientos básicos de estadística y un software apropiado, que en nuestro caso será R y RStudio.

Pero empecemos por el principio: definir una serie de conceptos y diferenciar los tipos de datos que vamos a usar [1] [2].

1.1 Definiciones

Aunque conocidos, recordamos los siguientes conceptos:

- **Población**: conjunto de elementos sobre los que se desea estudiar una característica.
- **Muestra**: subconjunto de la población sobre los que se aplican los estudios estadísticos, por la dificultad o imposibilidad de trabajar con la población completa. La muestra es representativa de la población si recoge los rasgos característicos de la misma.
- **Medidas descriptivas**: son valores numéricos obtenidos de los datos de la población o de la muestra que resumen la información contenida en ellos.
- **Variable**: característica que puede tomar valores distintos en distintos elementos. Por ejemplo, la altura de una persona o la precipitación obtenida en un determinado lugar.
- **Datos**: son números que representan las modalidades de las variables. Para la variable altura, un dato es 1.75 m.
- **Estadística descriptiva:** resume la información que contienen los datos de una muestra indicando las características más importantes y permitiendo obtener conclusiones sólo de la muestra observada.
- **Inferencia estadística:** pretende obtener (inferir) propiedades de una población a partir de los datos de una muestra, siendo necesario un ajuste de un modelo probabilístico adecuado. Obvia decir que la muestra debe ser representativa de la población.
- **Geoestadística:** comprende un conjunto de técnicas y herramientas que sirven para analizar y predecir los valores de una variable que está distribuida en el espacio o en el tiempo de forma continua. Es decir, es una estadística que está relacionada con los datos geográficos.
- **Parámetro**: propiedad descriptiva de la población. Los parámetros se escriben con letras griegas.
- **Estadístico**: propiedad descriptiva de la muestra. Los estadísticos se escriben con letras latinas.
- **Frecuencia:** la cantidad de veces que se repite un evento durante un experimento o muestra.

1.2 Tipos de datos y variables

Ya sabemos que los datos son la base de nuestro trabajo. Podemos obtener datos sobre pluviosidad en el mes de abril en una determinada provincia, el número de estudiantes universitarios en España o las coordenadas obtenidas en una medición topográfica.

En este punto, definimos los tipos de variables como:

- **Variables cualitativas**, representan una cualidad y sus valores no son numéricos, aunque su valor se puede representar por un número. Pueden ser cualitativas nominales (sexo, color…) o cualitativas ordinales. Por ejemplo, si asigno los valores de

1 a 5 a las siguientes categorías: Muy bueno (5), bueno (4), medio (3) malo (2) o muy malo (1), estas siguen siendo variables cualitativas y no puedo aplicar herramientas propias de las variables cuantitativas, como puede ser la media aritmética.

- **Variables cuantitativas**, representan cantidades. Existen variables cuantitativas discretas que sólo toman valores enteros, por ejemplo, el número de personas ingresadas en un hospital. Tenemos también variables cuantitativas continuas, que admiten cualquier valor de un intervalo de números reales, como el volumen de agua de una piscina.

En cuanto a los datos, los clasificamos como:

- **Datos escalares o lineales**. Pueden proceder de la observación de variables cualitativas o cuantitativas. En el caso de variables cuantitativas, una magnitud escalar es aquella que queda completamente determinada con un número y sus correspondientes unidades, por ejemplo, un valor de temperatura o una medida de distancia. Son representables por una escala numérica.
- **Datos vectoriales**. Una magnitud vectorial es aquella que, además de un valor numérico y su unidad (módulo) tiene una dirección y sentido, por ejemplo, la velocidad y dirección del viento. Si trabajamos en el plano, esta dirección y sentido se analiza como dato circular, pero si estamos en el espacio de 3 dimensiones hablamos de dato esférico.
- **Datos espaciales**. Son aquellos datos que, además de las variables que se estén considerando, aparece su localización geográfica.

2. Medidas descriptivas para datos escalares

Empezamos describiendo las medidas descriptivas para un conjunto de datos escalares. Este tipo de datos toma valores en la recta real, donde a partir del valor 0, se sitúan a la derecha los valores positivos y a la izquierda los negativos. Ya sabemos que están definidos por un número y su unidad. Como ejemplo nos sirven, los valores de temperatura de una determinada zona o el volumen de líquido contenido en una botella. En este texto hablaremos de datos escalares o lineales [3] [4].

2.1 Medidas de tendencia central

Estos estadísticos describen la localización de la distribución y suelen situarse hacia el centro de la distribución de los datos. Se aplican a datos cuantitativos, aunque algunos como la moda se aplique a datos cualitativos. Las principales medidas de tendencia central para una serie de datos escalares $x_1, x_2, x_3, \ldots, x_n$, son:

- **Media aritmética o media (** $\bar{x}$ **)**. Es el promedio aritmético de las observaciones. Suele ser muy sensible a valores extremos y viene dada por:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Si las observaciones vienen agrupadas en una tabla de frecuencias (f), la expresión a utilizar será:

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \ldots + x_n f_n}{n}$$

Si llamamos **residuos o desviaciones** a las diferencias entre cada una de las observaciones y la media aritmética, podemos enunciar las propiedades de la media aritmética:

- La suma de los residuos es cero.
- La suma de los cuadrados de los residuos es mínima.
- Si a todos los valores se les suma un mismo número, la media aritmética queda aumentada en dicho número.
- Si a todos los valores se les multiplica por un mismo número, la media aritmética queda multiplicada por dicho número.

Debemos recordar que:

- La media se puede calcular sólo para variables cuantitativas.
- La media es muy sensible a las observaciones extremas, por lo que se puede recurrir a calcular una **media recortada**, suprimiendo un porcentaje de los datos más extremos.

- **Media ponderada (** $\bar{x}_w$ **)**. Se usará cuando las observaciones tengan un peso (*w*) o importancia diferente. Su expresión es:

$$\bar{x}_w = \frac{x_1 w_1 + x_2 w_2 + \ldots + x_n w_n}{w_1 + w_2 + \ldots + w_n}$$

- **Media armónica (*H*)**. Se suele usar para promediar rendimientos, tiempos, velocidades… y se expresa por:

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \ldots + \frac{1}{x_n}}$$

En ciertos casos es más representativa que la media aritmética, pero no se aconseja su empleo en valores pequeños por ser muy sensible a estos valores.

- **Media geométrica** ($\bar{x}_g$) es la raíz enésima del producto de todos sus números. Su expresión es:

$$\bar{x}_g = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$$

- **Mediana.** Colocadas ordenadas todas las observaciones, la mediana es el valor que las separa por la mitad. Si la serie de observaciones es impar, la mediana es el valor central y si es par es la media aritmética de los dos valores centrales. La mediana no se ve afectada por las puntuaciones extremas como sí ocurre con la media aritmética, es decir, es más robusta.
- **Moda**. Es el valor que más veces se repite en un conjunto de observaciones. No tiene por qué ser único pues podemos tener distribuciones de datos unimodales, bimodales, multimodales…

2.2 Medidas de dispersión

Cuantifican la separación de valores respecto a las medidas de tendencia central complementando la información proporcionada por estas. Cuanto mayor es la dispersión de un conjunto de datos menos representativa de la muestra son las medidas de tendencia central. Definimos las siguientes medidas de dispersión:

- **Rango (*R*).** Es la diferencia entre el valor máximo y mínimo de las observaciones. Cuanto mayor es el rango, más dispersos están los datos.
- **Rango intercuartílico (R_1).** Los cuantiles (que se explicarán a continuación como medida de posición) son valores de la distribución de datos que la dividen en partes iguales. Los cuartiles dividen la muestra en cuatro partes iguales (Q_1, Q_2, Q_3 y Q_4). El rango intercuartílico es la diferencia entre Q_3 y Q_1:

$$R_1 = Q_3 - Q_1$$

- **Desviación típica o desviación estándar.** Describe el grado de homogeneidad de las observaciones. Cuanto más dispersas son las observaciones con respecto a la media, mayor es la desviación. Este estadístico se basa en la normalidad de la distribución y toma valores no negativos. La expresión para la desviación típica muestral es:

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

Mientras que la desviación típica poblacional se expresa por:

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

- **Varianza.** Es el cuadrado de la desviación típica y es un valor no negativo. Al igual que la desviación típica, cuanto más dispersas son las observaciones con respecto a la media, mayor es la varianza.
- **Coeficiente de variación (CV).** Expresa la desviación como un porcentaje de la media, siempre que esta sea distinta de cero. Cuanto mayor de 1 sea el CV, mayor es la variabilidad de nuestros datos.

$$V = \frac{s}{|\bar{x}|}$$

2.3 Medidas de forma

Nos referimos a los coeficientes de asimetría y curtosis que informan sobre la forma de la distribución de una variable.

- **Coeficiente de asimetría (skewness)**. Caracteriza la simetría de la distribución. Una distribución unimodal y simétrica como la distribución normal o gaussiana (se explicará más adelante) tiene asimetría 0. Cuando la distribución tiene una cola izquierda desarrollada, toma valores negativos, si es la derecha la desarrollada, positivos. En una distribución simétrica la mediana, la moda y la media aritmética coinciden.

  Se suele considerar que si el coeficiente de asimetría se encuentra entre -0.5 y 0.5 los datos son casi simétricos. Si el coeficiente se encuentra entre -1 y -0.5, para asimetría

negativa y entre 0.5 y 1, para asimetría positiva, los datos son moderadamente asimétricos. Con valores mayores de 1 y -1 consideramos asimetría alta.

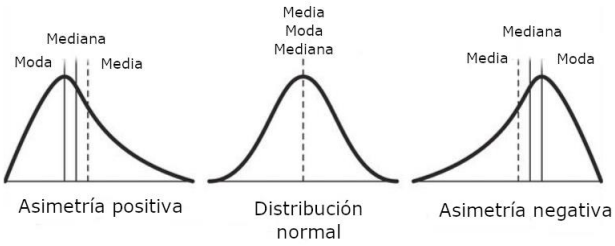


Figura 1. Asimetría de la distribución normal, asimetría positiva y negativa.

- **Coeficiente de apuntamiento o curtosis**. La curtosis caracteriza la forma de la distribución e indica lo picuda o plana que es una distribución. Una distribución unimodal con un pico que ajuste bien a una distribución normal tiene curtosis 3, aunque se trabaja con expresiones en las que se tiene en cuenta el exceso de curtosis para generar un coeficiente que valga 0 para la normal. En este caso, si el pico es mayor la curtosis es positiva y si es menor negativa. La curtosis se relaciona con las colas de la distribución y sirve para medir observaciones discordantes o outliers.

2.4 Medidas de posición

Los **cuantiles** son valores de la distribución de datos que la dividen en partes iguales. Los más usados son:

- **Cuartiles**. Dividen al conjunto de datos en cuatro partes iguales.
- **Quintiles**. Dividen al conjunto de datos en cinco partes iguales.
- **Deciles**. Dividen al conjunto de datos en diez partes iguales.
- **Percentiles**. Dividen al conjunto de datos en cien partes iguales.

Como ejemplo, supongamos que en una clase de 40 alumnos conocemos las notas obtenidas en una determinada asignatura y las ordenamos en orden creciente. Nos dicen que un determinado alumno ha sacado un 6.3 y se corresponde con el percentil 70. Quiere decir que el 70% de las notas están por debajo de ese valor de 6.3.

3. Distribución normal o gaussiana

Una de las distribuciones más habituales para variables continuas es la llamada distribución normal de probabilidades, distribución gaussiana o Campana de Gauss. Esta distribución describe bastante bien los errores de las observaciones y aparece con frecuencia en estadística y en teoría de probabilidades pues modela numerosos fenómenos naturales y sociales. Esta distribución se observó en el siglo XIX y se pensó que se adaptaba a la mayoría de los fenómenos de la naturaleza, por lo que se la denominó ley de probabilidad normal. Depende de dos parámetros, su media poblacional (μ) y su desviación típica (σ). Cuando estos valores son 0 y 1, la curva se denomina normal estándar, *N(0,1)*. Cualquier distribución normal que no siga el esquema *N(0,1),* puede tipificarse o estandarizarse restando su media y dividiéndolo por su desviación típica:

$$Z = \frac{\bar{x} - \mu}{\sigma}$$

La gráfica de esta función tiene forma de campana según se ve en la Figura 2 y su función de densidad es:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

Donde μ es la media, σ es la desviación típica y σ^2 es la varianza.

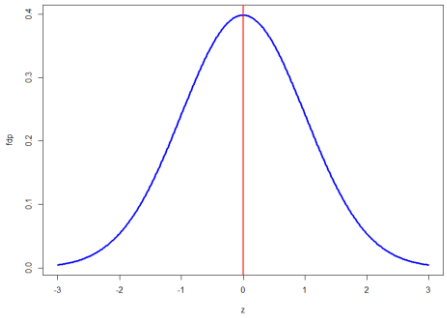


Figura 2. Curva de Gauss estándar, N(0,1).

Las características principales de esta distribución son:

- La media, la moda y la mediana de la distribución son iguales y se localizan en el centro de la distribución.
- La distribución es simétrica con respecto a su media, lo que indica que existen tantos valores positivos como negativos.
- Es asintótica respecto al eje *X*, lo que implica que los valores más extremos tienen una frecuencia menor.
- El área de la campana de Gauss comprendida entre los valores -σ y +σ es el 68.27% del total, es decir, que la probabilidad de que un error esté dentro de ese intervalo es del 68.27%.

Una variable que siga una distribución normal está determinada por dos parámetros, su media aritmética y la desviación típica.

La masa de probabilidad que hay bajo la curva es 1. Así, el área acumulada hasta un determinado punto, representa un porcentaje de los casos. Por ejemplo, en la Figura 3 , el área marcada en color magenta entre la curva y el eje de abscisa hasta x=-0.5, se corresponde con la probabilidad del 30.85% de que la variable de distribución N(0,1) tome un valor menor que -0.5 (ver en una tabla de distribución normal).

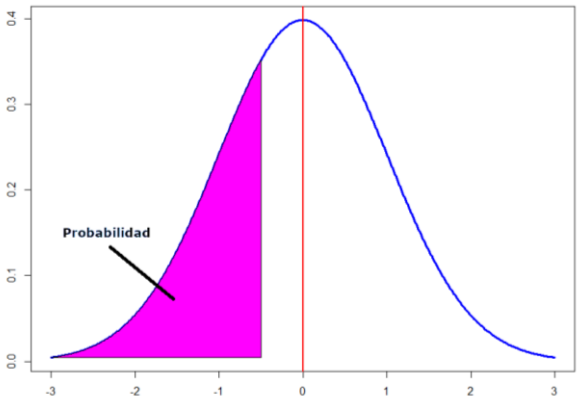


Figura 3. Área acumulada hasta x=-0.5.

Disponemos de diversas pruebas o test para detectar la normalidad de los datos como el test de *Kolmogorov-Smirnov* o el test de *Shapiro-Wilk*. Este último test es recomendado para muestras menores de 50 valores. Aplicaremos estos test una vez recordado el concepto de contraste de hipótesis y p-valor.

4. Contraste de hipótesis

Se denomina hipótesis estadística (*H*) a una afirmación respecto a una característica de la población. El **contraste de hipótesis** es uno de los métodos estadísticos más habituales que nos permite aceptar o rechazar una determinada afirmación en función de los valores obtenidos en la realidad. El primer paso es definir las hipótesis que se quieren contrastar o comparar. Por ejemplo, si el tratamiento con un medicamento modifica el nivel estándar de colesterol, si dos variables en un estudio son independientes o si una variable de estudio sigue una determinada distribución.

Se establece una hipótesis de partida que se desea contrastar y se denomina **hipótesis nula**, que se representa por H_0. La otra hipótesis recibe el nombre de **hipótesis alternativa** (H_1 o Ha) y será la que aceptaremos si la H_0 es rechazada**.** Se trata de decidir por técnicas estadística qué hipótesis creemos correcta**,** aceptándola o rechazándola.

Habrá que tomar una muestra de valores para poder contrastar las hipótesis y tomar un modelo probabilístico para la variable que se observa, que suele ser la distribución normal. Si es posible usar un modelo, se habla de contrastes paramétricos, en caso contrario tenemos contrastes no paramétricos. Los primeros son más potentes que los segundos, por lo que se preferirá su uso siempre que sea posible.

Lo que se afirma es que, si se rechaza H_0, la probabilidad de que H_0 sea errónea va a ser muy alta. Si no rechazo H_0, no puedo decir que la probabilidad de que sea falsa es muy alta, sino que no tengo datos suficientes para poder rechazarla. La explicación más habitual que se hace en este caso es como si fuera un juicio. La hipótesis inicial (nula) es que se es inocente y la hipótesis alternativa que se es culpable. Si rechazo H_0 en un juicio de que alguien es inocente debo de tener una seguridad muy alta, con muchas pruebas. Si no tengo suficientes pruebas, no estoy asegurando que es inocente, estoy diciendo que no tengo evidencias de que sea culpable.

La idea que subyace en este tipo de métodos estadísticos es la siguiente: si la hipótesis nula planteada es cierta, la muestra observada debería tener una determinada distribución de probabilidad. Si extraída una muestra al azar ocurre un suceso que tenía poca probabilidad de ocurrir de ser cierta H_0, se puede deber a dos motivos: o hemos tenido mala suerte de elegir una muestra extraña, o lo más probable, que la hipótesis nula fuera falsa.

En este proceso de decisión podemos cometer dos tipos de errores:

Rechazar la hipótesis nula cuando es cierta→ error de tipo I.

Aceptar la hipótesis nula cuando es falsa → error de tipo II.

	H0 verdadera	**H0 Falsa**
Aceptar H0	No hay error	Error tipo II
Rechazar H0	Error tipo I	No hay error

Nos preocupa más el error de tipo I. Volviendo al ejemplo del juicio, si la hipótesis nula es que se es inocente, y rechazamos esa hipótesis siendo cierta, aceptamos la culpabilidad. Es decir, continuando con el ejemplo, sería preferible soltar a un culpable, que encarcelar a un inocente. De ahí que se busque minimizar el error de tipo I.

La probabilidad de cometer error de tipo I se llama **nivel de significación** y se representa por α. La probabilidad de cometer un error de tipo II se suele representar por β.

Además, se establece una región crítica y una región de aceptación siempre en función de un estadístico adecuado. La región crítica comprende el conjunto de valores del estadístico donde se rechaza la hipótesis nula. El conjunto de datos complementario se llama región de aceptación y se corresponde con los valores del estadístico para los que se acepta la hipótesis nula. Las hipótesis, además, pueden ser simples o compuestas si están formadas por un solo parámetro o varios, respectivamente.

Para realizar el contraste hay que definir una **medida de discrepancia** entre los datos muestrales y la hipótesis nula y determinar cuál es la máxima discrepancia admisible. Para la región de rechazo de la hipótesis debemos fijar el nivel de significación α (por ejemplo 0.05 o 0.01). Esto implica que no se consideran aceptables discrepancias que tengan una probabilidad menor de 0.05 de ocurrir.

Se llama p-valor a la probabilidad de obtener una discrepancia mayor que la observada. Se rechazará H_0 cuando el p-valor sea pequeño (menor de 0.05 o 0.01). El cálculo del p-valor permite valorar la decisión de aceptar o rechazar la hipótesis nula.

De entrada, establecemos nuestro estadístico Z sobre una población normal tipificada y vemos:

Se acepta H_0 si $Z = \frac{|\bar{x}-\mu|}{s/\sqrt{n}} =\leq t_{n-1;\propto/2}$

Se rechaza H_0 si $Z = \frac{|\bar{x}-\mu|}{s/\sqrt{n}} =\geq t_{n-1;\propto/2}$

Por ejemplo, supongamos que tenemos una muestra de datos y aplicamos el test de normalidad de *Shapiro-Wilk*. En este test se establecen una hipótesis nula (H0) y una hipótesis alternativa (H1)

H0: Los datos siguen una distribución normal.

H1: Los datos no siguen una distribución normal.

Se rechazará la hipótesis nula si el p-valor es menor que el nivel de significación.

```
        Shapiro-Wilk normality test

data:  x1
W = 0.99072, p-value = 0.7223
```

Figura 4. Resultado de p-valor en prueba de Shapiro-Wilk, donde se acepta la hipótesis nula de normalidad de distribución.

5. Representación gráfica

Una de las formas más comunes de representar gráficamente la frecuencia de los datos es a través de **histogramas**. Los datos se dividen en intervalos de clases y se representan en barras con una superficie proporcional a la frecuencia de valores de cada clase. La elección del número de intervalos de clase corresponde al usuario según el tamaño de la muestra y el rango de la variable. Si las clases se toman de la misma longitud, las frecuencias son proporcionales a la altura. Uno de los criterios a utilizar para elegir el número de intervalos es hacerlo igual a la raíz del tamaño de la muestra. En la figura siguiente se muestra el histograma de frecuencia para un conjunto de datos dado.

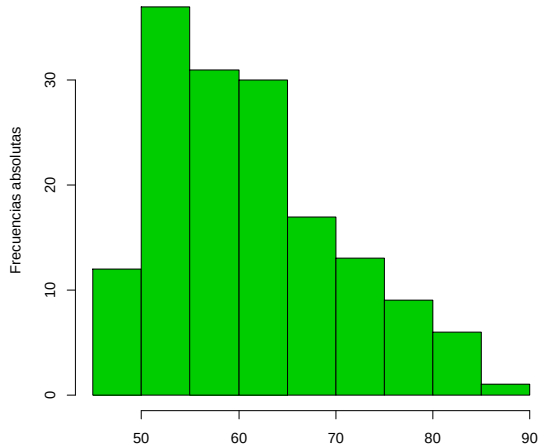


Figura 5. Histograma de distribución de datos.

La forma del histograma ofrece ya una indicación de la simetría y la modalidad de la distribución.

Otra forma de representar nuestros datos es usando un **diagrama de caja** también conocido como diagrama de caja y bigotes o **boxplots**. Este tipo de gráficos consiste en una caja rectangular cuyo lado mayor se corresponde con el rango intercuartílico (RIC), es decir, que los límites de la caja son el Q1 y el Q3 y dentro de la caja están el 50% de los datos. El segmento horizontal de la caja indica la posición de la mediana. Las líneas que sobresalen de la caja se llaman bigotes y se calculan según una determinada cantidad por el rango intercuartílico, generalmente 1.5*RIC. Los valores que quedan fuera de estos bigotes son considerados outliers o valores atípicos. En la Figura 6 se muestra un diagrama de caja.

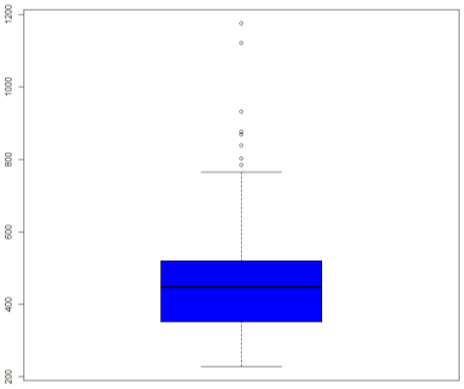


Figura 6. Diagrama de caja.

Un gráfico habitual cuando trabajamos con dos variables cuantitativas es el **gráfico de dispersión** o de **nube de puntos** donde representamos el valor de una variable en el eje de abscisa y la otra en la ordenada. Se suele representar la variable dependiente en el eje de ordenadas y la variable independiente en el eje de abscisas.

En la Figura 7 se representa una nube de puntos donde en el eje de abscisas están los valores de las alturas obtenidas de una serie de estaciones meteorológicas y en la ordenada las precipitaciones medias obtenidas en esas estaciones en un periodo determinado.

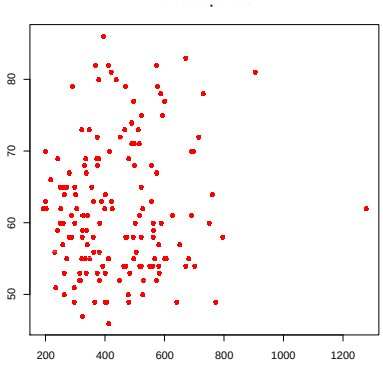


Figura 7. Diagrama de nube de puntos.

6. Relaciones entre dos variables cuantitativas

Los ejemplos anteriores, excepto el último gráfico de nube de puntos, se referían al caso de trabajar con una sola variable, pero lo normal es estar en bases de datos multivariante donde las diferentes variables pueden o no, estar relacionadas entre sí.

En el caso de trabajar con dos variables al mismo tiempo, podemos calcular los estadísticos siguientes:

- **Covarianza.** Es una medida de dispersión conjunta de dos variables estadísticas *X* e *Y*. Dos variables están relacionadas si varían conjuntamente. Si la covarianza es positiva la relación entre las variables es directa, en caso contrario, inversa. Una covarianza próxima a cero indicaría ausencia de relación entre dos variables.

$$COV(X,Y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n}$$

La covarianza presenta dos problemas para su uso, depende de la magnitud de los datos, es por tanto dimensional, y no tiene límite superior ni inferior. Para solucionar estos problemas se prefiere usar el coeficiente de correlación al ser independiente de la magnitud de los datos y tener una acotación entre +1 y -1.

- **Coeficiente de correlación (*r*).** Mide la relación lineal entre dos variables aleatorias cuantitativas *X* e *Y*. Se expresa en función de la covarianza y las desviaciones típicas de cada variable.

$$Coef.Corre(X,Y) = r = \frac{COV(X,Y)}{s_X \cdot s_Y}$$

El coeficiente de correlación, que no tiene unidades, adquiere el signo de la covarianza y oscila entre +1 y -1, esto es, un coeficiente de correlación positivo indica una relación lineal directa. Un valor negativo, una relación lineal inversa. Además, valores próximos a 1 o -1 indican una correlación positiva y negativa fuerte, respectivamente. En el caso de ser cero, puede indicar que son variables independientes o también que existe una relación que es no lineal. Si la relación entre las dos variables no es lineal, es erróneo emplear solo este estadístico.

Tanto la covarianza como el coeficiente de correlación son sensibles a valores erráticos, por lo que hay que tener especial cuidado en eliminar observaciones discordantes.

- **Regresión lineal.** Cuando existe una relación entre dos variables con coeficiente de correlación alto, es posible que el conocimiento de una variable proporcione información sobre la otra. Podemos ajustar una línea recta a la nube de puntos, la **recta de regresión lineal**, para describir como varía la variable dependiente (*y*) en función de la independiente (*x*), de la forma:

$$y = \beta + \alpha x$$

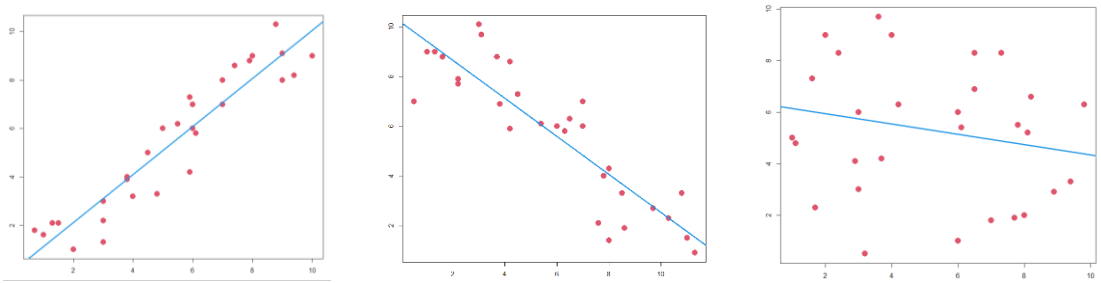
El parámetro α es la pendiente de la recta e indica cuanto aumenta la media de la variable dependiente, cuando la variable independiente aumenta una unidad. El parámetro β es la ordenada en el origen y se corresponde con el valor de la media de la variable dependiente cuando la variable independiente es cero. La recta debe cumplir que las diferencias de distancia entre los puntos observados y la recta sean mínimas y se calcula por un **ajuste de mínimos cuadrados**.

No siempre será posible ajustar la nube de puntos mediante una recta. Además del ajuste lineal, puede existir un ajuste polinómico, exponencial, logarítmico…

Lo que sí es importante es cuantificar la bondad del ajuste de la recta con el llamado **coeficiente de determinación R^2** que es la proporción de varianza total de la variable explicada por la regresión. Este coeficiente oscila entre 0 y 1 de tal forma que, cuanto más se acerque a 1, mejor será el ajuste. A la inversa, un valor próximo a 0 indica un mal ajuste. Este coeficiente lo calculamos al elevar al cuadrado el coeficiente de correlación.

Este tipo de gráficos muestra las posibles relaciones entre ambas variables. En un diagrama de dispersión nos encontramos con 3 situaciones tipo: a) cuando los valores pequeños de la variable tienden a asociarse con los valores pequeños de la otra y lo mismo

para valores grandes, decimos que las variables están correlacionadas positivamente, b) cuando a valores pequeños de una variable se asocian a valores grandes de otra, sería correlación negativa y c) cuando no existe relación entre las variables. En la figura siguiente aparecen ejemplos de estas relaciones que, en el caso *a* y *b* al tender los puntos a colocarse en una recta se dice que la relación es **lineal**. El caso *c*, con una nube de puntos dispersa, indica que no hay relación entre las variables.



a) b) c)

Figura 8. Diagrama de nube de puntos y recta de regresión lineal en a) correlación directa, b) correlación inversa, c) sin correlación.

Los valores del **coeficiente de determinación R^2** para los casos a, b, y c anteriores son 0.882, 0.781 y 0.042, respectivamente.

7. Bibliografía

[1] F. J. Moral García, *La representación gráfica de las variables regionalizadas. Geoestadística lineal*. Cáceres: Universidad de Extremadura. Servicio de publicaciones, 2003.

[2] A. García Pérez, *La interpretación de los datos. Una introducción a la estadística aplicada*. Madrid: UNED.

[3] R. Martínez Quintana, *Estadística básica para Topografía* vol. 66. Cáceres (Spain): Universidad de Extremadura. Servicio de Publicaciones, 2009.

[4] D. Peña and J. Romo, *Introducción a la estadística para las ciencias sociales*. Madrid: McGraw-Hill, 1997.