USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA FACULTADE DE ENFERMARÍA

Estadística y metodología de la investigación
Curso 2012-2013
Pedro Faraldo, Beatriz Pateiro

Tema 1. Estadística Descriptiva

1 Introducción

La **estadística descriptiva** es un conjunto de técnicas numéricas y gráficas para describir y analizar un grupo de datos, sin extraer conclusiones (inferencias) sobre la población a la que pertenecen. En este tema se introducirán algunas técnicas descriptivas básicas, como la construcción de tablas de frecuencias, la elaboración de gráficas y las principales medidas descriptivas de centralización, dispersión y forma que permitirán realizar la descripción de datos.

Ejemplo 1: Con objeto de hacer un estudio sobre la salud de los habitantes de una ciudad con edades entre 18 y 60 años, se recogen en un centro médico datos sobre análisis realizados a 100 pacientes mayores de 18 años y menores de 60 que aparentemente no presentan problemas de salud graves. De los análisis realizados se recoge el sexo del paciente, el antígeno del grupo sanguíneo (A, B, AB o 0), el pH de la sangre y el ácido úrico, además de la edad. La distribución de los antígenos en la población Española es de 45 % para el 0, 42 % para el A, 10 % para el B y 3 % para el AB.
Además, los valores normales del pH en sangre están entre 7.35 y 7.45 y los del ácido úrico están entre 2.4 y 7 mg/dL.

2 Conceptos generales

En cualquier análisis estadístico el objetivo último es extraer conclusiones sobre un colectivo de interés denominado población. En ocasiones, el tamaño de la población (formada por individuos) puede hacer inabordable el estudio individualizado de las características de cada uno de ellos. Si se quisiera realizar un estudio sobre el nivel de glucemia en los varones adultos en España, sería imposible realizar una toma de glucemia en cada uno de ellos. Para solucionar este problema, dichas mediciones se realizaran sobre una muestra.

- **Población**: colectivo de individuos sobre los que se quiere extraer alguna conclusión.
- **Individuo**: cada uno de los elementos de la población (unidad estadística).
- **Muestra**: subconjunto (representativo) de la población, que se selecciona con el objetivo de extraer información.

En el Ejemplo 1, la población está formada por los habitantes de la ciudad que tienen entre 18 y 60 años. Cada uno de ellos es un individuo de la población. Los 100 pacientes sobre los que se recoge la información forman la muestra.

Las técnicas de **estadística descriptiva** permiten describir y analizar un grupo dado de datos, sin extraer conclusiones (inferencias) sobre la población a la que pertenecen. Se tendrá que recurrir a la **inferencia estadística**, que es la parte de la Estadística que trata las condiciones bajo las cuales las inferencias extraídas a partir de una muestra son válidas, para extraer conclusiones sobre la población de interés. Para aplicar una técnica descriptiva, numérica o gráfica, será necesario analizar previamente el tipo de variable con la que se está trabajando.

- **Variable estadística**: cada una de las características consideradas con el propósito de describir a cada individuo de la muestra.
- **Tipos de variables**: distinguiremos dos tipos de variables. Las variables cualitativas o categóricas (aquellas que no se pueden expresar a través de una cantidad numérica) y las variables cuantitativas (se puede expresar a través de un número). A su vez, estas últimas pueden clasificarse en discretas y continuas, según el tipo de valores que tomen. En el Cuadro 1 se incluyen algunos ejemplos.

Tipo	Clases	Ejemplo
Cualitativa	Nominal	Sexo, raza, color de ojos,...
	Ordinal	Grado de contaminación, calificación,...
Cuantitativa	Discreta	Nº de hermanos, nº de materias, ...
	Continua	Peso, altura, ...

Cuadro 1: Tipos de variables estadísticas.

Volviendo al Ejemplo 1, el sexo y el antígeno del grupo sanguíneo son variables estadísticas cualitativas (nominales). El pH en sangre y el ácido úrico son variables cuantitativas continuas y la edad es cuantitativa discreta. La edad como puede presentar muchos valores (desde 18 a 60, si se mide en años), por lo que para su tratamiento podrían utilizarse técnicas propias de las variables cuantitativas continuas.

---

3 Distribuciones de frecuencias

Las tablas de frecuencias son una de las técnicas básicas para el resumen de información a partir de una muestra de datos. Su construcción es sencilla pero en conjuntos de datos de un tamaño moderado o grande su cálculo puede resultar laborioso, aunque se pueden obtener utilizando cualquier paquete estadístico.

- **Tablas de frecuencias**: las tablas de frecuencias se utilizan para representar la información contenida en una muestra de tamaño n extraída de una población, $(x_1, \ldots, x_n)$.
- **Modalidades**: cada uno de los valores que puede tomar una variable (cualitativa o cuantitativa discreta). Se denotan como: $c_i, i = 1, \ldots, k$.

  El número de individuos de la muestra en cada modalidad c_i se denota por n_i.
- **Frecuencia absoluta**: para cada modalidad c_i, la frecuencia absoluta es $n_i, i = 1, \ldots, k$.
- **Frecuencia relativa**: para cada modalidad c_i, la frecuencia relativa es $f_i = n_i/n, i = 1, \ldots, k$.
- **Frecuencia absoluta acumulada**: la frecuencia absoluta acumulada de una modalidad c_i es $N_i = \sum_{j=1}^{i} n_j = n_1 + \ldots + n_i, i = 1, \ldots, k$.
- **Frecuencia relativa acumulada**: la frecuencia relativa acumulada de una modalidad c_i es $F_i = \sum_{j=1}^{i} f_j = f_1 + \ldots + f_i = \frac{N_i}{n}, i = 1, \ldots, k$.

A partir de sus definiciones, se pueden demostrar algunas propiedades de las frecuencias absolutas y relativas que se calculan en las tablas de frecuencias. Así, se tiene que:

- Las frecuencias absolutas: $0 \leq n_i \leq n$, $i = 1, \ldots, k$.
- Las frecuencias relativas: $0 \leq f_i \leq 1$, $i = 1, \ldots, k$.
- Las frecuencias absolutas acumuladas: $N_k = \sum_{j=1}^{k} n_j = n_1 + \ldots + n_k = n$.
- Las frecuencias relativas acumuladas: $F_k = \sum_{j=1}^{k} f_j = f_1 + \ldots + f_k = 1$

A continuación se muestra la disposición de los distintos elementos de una tabla de frecuencias.

Modalidad	**Frecuencia absoluta**	**Frecuencia relativa**	**Fr. abs. acumulada**	**Fr. rel. acumulada**
c_1	n_1	f_1	N_1	F_1
c_2	n_2	f_2	N_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
c_i	n_i	f_i	N_i	F_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
c_k	n_k	f_k	$N_k = n$	$F_k = 1$
Total	n	1		

Cuadro 2: Tabla de frecuencias.

Para un grupo de 21 pacientes de la muestra, se tienen los siguientes datos sobre el antígeno.

Paciente	1	2	3	4	5	6	7	8	9	10	11
Grupo	AB	0	A	B	0	0	B	A	B	0	B
Paciente	12	13	14	15	16	17	18	19	20	21	
Grupo	A	0	0	A	B	B	0	0	0	AB	

Para estos datos, podemos construir una tabla de frecuencias, calculando frecuencias absolutas y relativas, así como las respectivas acumuladas. ¿Cuál es la proporción de individuos con grupo A en la muestra? ¿Y con grupo A o B?

En el caso de variables cualitativas o cuantitativas discretas con pocos valores, es posible determinar las modalidades de la variable. Sin embargo, en el caso de variables cuantitativas continuas (o cuantitativas discretas con muchos valores), se tendrán que construir modalidades artificiales de manera que se agrupen valores por intervalos. Estas nuevas modalidades se denominan intervalos de clase.

- **Intervalos de clase**: para variables cuantitativas continuas, se agrupan los distintos valores obtenidos en la muestra en intervalos. Cada intervalo representará una *modalidad* en el caso de variables cuantitativas continuas. A partir de una muestra, los intervalos de clase se construyen de la siguiente forma:
  - Denotamos por $e_0 < e_1 < \ldots < e_k$ los extremos de los k intervalos de clase. Cada intervalo será de la forma (e_{i-1}, e_i).
  - Amplitud del intervalo: $a_i = e_i - e_{i-1}$.
  - Marca de clase: $c_i = \dfrac{e_{i-1} + e_i}{2}$.
  - Para seleccionar el número de intervalos, consideramos el entero más próximo a $\sqrt{n}$, donde n es el tamaño de la muestra observada. El número de intervalos suele estar entre 5 y 20. Para determinar la amplitud de los intervalos (en principio, todos de la misma amplitud), tenemos que ver antes cuál es el rango de variación de los datos (diferencia entre el máximo y el mínimo), y construir los intervalos de manera que cubran todo el rango.

4 Representaciones gráficas

La clasificación de variables que se ha expuesto en la sección anterior, distinguiendo entre variables cualitativas y cuantitativas (discretas y continuas) es de crucial importancia a la hora de construir representaciones gráficas. De modo esquemático, se introducen las principales técnicas de representación para variables cualitativas, variables cuantitativas discretas y cuantitativas continuas. En el caso de variables cuantitativas discretas, si tienen pocos valores, se puede hacer uso de las representaciones descritas para variables cualitativas (diagramas de barras y sectores). Si por el contrario toman muchos valores, entonces se pueden utilizar las representaciones para variables cuantitativas continuas.

- **Variables cualitativas.** Para la representación de variables cualitativas se suelen utilizar el diagrama de barras o el diagrama de sectores. Para construir un diagrama de barras, en el eje horizontal se representan las categorías o modalidades de la variable que se quiere representar y se levantan barras de altura proporcional a la frecuencia de cada modalidad (absoluta o relativa). En el diagrama de sectores también se representan las distintas modalidades y su frecuencia, de manera que el círculo se reparte de forma proporcional a la frecuencia de cada modalidad. Algunos ejemplos de estas representaciones para datos de participación en redes sociales en un grupo de 180 jóvenes se muestran en la Figura 1.

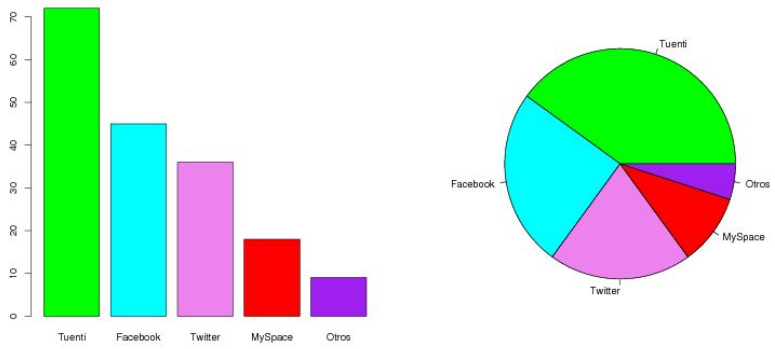


Figura 1: Diagrama de barras y diagrama de sectores para datos de pertenencia a redes sociales.

- **Variables cuantitativas discretas.** Además del diagrama de barras descrito para las variables cualitativas, que también se puede utilizar para variables cuantitativas discretas, para la representación de este tipo de variables se tiene el diagrama acumulativo de frecuencias. El diagrama acumulativo de frecuencias se construye representando, para cada modalidad de la variable c_i, los puntos (c_i, N_i) (o bien (c_i, F_i)) y uniéndolos con segmentos horizontales y verticales, de forma que se obtiene una función escalonada. Si se utilizan las frecuencias relativas acumuladas, el valor máximo del diagrama acumulativo se alcanza en el 1, mientras que si se construye con las frecuencias absolutas acumuladas, el máximo será el número de datos de la muestra. Se muestran el diagrama de barras y el diagrama acumulativo de frecuencias para la variable *"número de hijos de una familia"* en la Figura 2 .

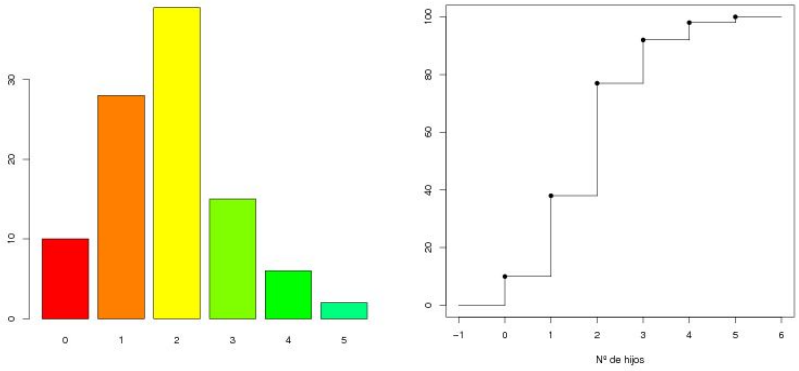
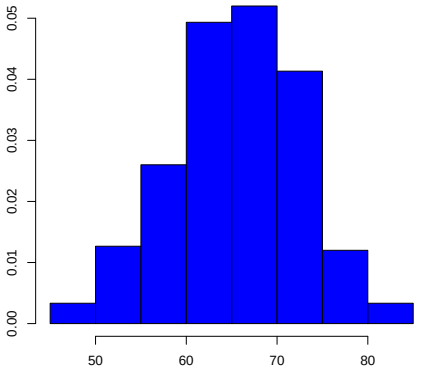


Figura 2: Diagrama de barras y diagrama acumulativo de frecuencias para el número de hijos de una familia.

- **Variables cuantitativas continuas.** En el caso de variables cuantitativas continuas, podemos construir el polígono (acumulativo) de frecuencias, de igual modo que el diagrama acumulativo de frecuencias explicado para variables cuantitativas discretas, pero considerando las marcas de clase de cada intervalo e_i en la representación. Sin embargo, son más usuales otras representaciones como el histograma y el diagrama de tallo y hojas.

  El histograma equivale en cierto modo al diagrama de barras, pero en el caso continuo, de forma que las barras aparecen contiguas. En el eje horizontal se representan los intervalos de clase de la variable, y

sobre ellos se levantan barras de altura $h_i = n_i/a_i$ (o bien $h_i = f_i/a_i$), donde n_i es la frecuencia absoluta de cada intervalo (f_i es la frecuencia relativa) y a_i es la amplitud del mismo. Si el histograma se construye con frecuencias relativas, la suma de las áreas de las barras es igual a 1. El histograma da una idea clara de la *distribución* de los datos, pero es muy sensible a la elección de los intervalos de clase (véase Figura 3, panel izquierdo).




```
46 | 4450
48 | 8
50 | 3556
52 | 1578289
54 | 112356794568
56 | 23779011455678
58 | 000112367912455677789
60 | 00111233445667778999012566778899
62 | 23456799011122234444556788899
64 | 00111256777889000001122334566778899
66 | 00111222233445555667788901244558899999999
68 | 11122346679112356788888
70 | 11123333889111335567889
72 | 00223455568899000112447778
74 | 0233366677799934778
76 | 13455358
78 | 81167
80 | 3
82 | 1
84 | 227
```

Figura 3: Histograma y diagrama de tallo y hojas para datos de peso de personas adultas.

El diagrama de tallo y hojas es una representación que permite observar los datos y que a la vez da una idea de la distribución de los mismos. Primero se seleccionan el número de cifras significativas (tallo) que se colocan a la izquierda, se traza una línea vertical y se incluyen al lado las cifras siguientes de cada dato observado (hojas). Se puede ver un ejemplo de representación para el peso de 300 personas en la Figura 3. Si se gira el diagrama de tallo y hojas 90º en el sentido contrario a las agujas del reloj, se puede observar una forma muy similar a la del histograma.

■ Para representar las observaciones de las variables del ejemplo debemos tener en cuenta si son cualitativas o cuantitativas. El sexo y el antígeno del grupo sanguíneo pueden representarse utilizando un diagrama de barras o un diagrama de sectores. Para el pH en sangre y el ácido úrico se puede utilizar un histograma o un diagrama de tallo y hojas. La edad, cuantitativa discreta, puede representarse con un diagrama de barras si no toma muchos valores distintos. En otro caso, se puede probar con un diagrama acumulativo de frecuencias o con alguna de las representaciones propias de variables cuantitativas continuas (histograma o diagrama de tallo y hojas).

---

5 Medidas características: posición, dispersión, forma

Denotando por X la variable estadística de interés y por x_i la observación en el individuo i, se introducirán en este apartado algunas de las principales medidas características para describir la información contenida en una muestra $x_1, \ldots, x_n$ de tamaño n. Dichas medidas se utilizan para resumir la información atendiendo a tres aspectos principales: alrededor de qué valores se encuentran los datos, cuánto se dispersan y si se distribuyen de manera similar a una *campana de Gauss*, que será el modelo que se tome como referencia. Por ello, se distinguirán tres tipos de medidas: medidas de posición, medidas de dispersión y medidas de forma.

5.1 Medidas de posición

Las medidas de posición o localización nos indican el valor o valores alrededor de los cuales se sitúan los datos observados. Distinguiremos medidas de localización de tendencia central (media, mediana y moda) y de tendencia no central (cuartiles, deciles y percentiles).

5.1.1 Medidas de posición de tendencia central.

Como medidas de posición de tendencia central se introducirán la media aritmética o media muestral, la mediana y la moda. Estas medidas nos proporcionan valores alrededor de los cuales se distribuyen los datos observados en la muestra.

- **Media aritmética**. Se define como:

$$\bar{x} = \frac{x_1 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

  La media aritmética (media muestral) presenta las siguientes propiedades, que son fáciles de deducir a partir de la definición.

  - Toma valores entre el mínimo y el máximo:

$$\text{mín}\{x_1, \ldots, x_n\} \leq \bar{x} \leq \text{máx}\{x_1, \ldots, x_n\}.$$

  - La media aritmética es lineal. Si consideramos los datos $y_i = ax_i + b$, la media de los nuevos datos se obtendrá como $\bar{y} = a\bar{x} + b$.
  - La media de las desviaciones con respecto a la media es cero:

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

  - La media de los cuadrados de las desviaciones con respecto a una constante es mínima para la media:

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

  El valor de la media no tiene porqué pertenecer al conjunto de posibles valores de la variable. Por ejemplo, puede resultar que el número medio de hermanos de una muestra no sea un número entero.

  Uno de los problemas que presenta la media es que no es una medida robusta, es decir, su valor se ve influenciada por datos anormalmente altos o bajos. Los datos que difieren numéricamente de las demás observaciones se denominan valores atípicos. Algunas modificaciones para corregir la falta de robustez son la media truncada y media recortada. En la media truncada, un porcentaje de los datos atípicos se elimina del cálculo y para obtener una media recortada, estos valores atípicos se substituyen por el punto de corte, es decir, el dato inmediatamente inferior a los que se eliminan, para datos altos, y el inmediatamente superior para los datos bajos.

  Otra modificación es la media ponderada en la cual se asigna distintos pesos a las observaciones. En la media aritmética cada observación tiene una contribución de peso $1/n$ al valor de $\overline{x}$. En la media ponderada, cada observación tendrá una ponderación ω_i, de tal modo que $\sum_{i=1}^{n} \omega_i = 1$.

En el caso de que se disponga de datos agrupados en una tabla de frecuencias, la media aritmética se calcula como:

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

donde c_i es la marca de clase y k denota el número de intervalos de clase de los que se dispone. Las propiedades anteriormente descritas también se aplican a este caso.

- **Mediana**. Si suponemos que los datos de la muestra están ordenados de menor a mayor, la mediana es el valor hasta el cual se encuentran el 50 % de los casos. Por tanto, la mediana dejará la mitad de las observaciones por debajo de su valor y la otra mitad por encima. Así, si la muestra consta de un número impar de datos (n impar), la mediana será el dato central. Si el tamaño de la muestra n es par, entonces se tomará como mediana la media de los dos datos centrales.

  En el caso de tener la variable representada en una tabla de frecuencias, podemos definir el intervalo mediano, que será aquel cuya frecuencia relativa acumulada en el extremo inferior es menor que 1/2 y en el extremo superior mayor que 1/2.

  La mediana, a diferencia de la media, es una medida robusta ya que su valor se ve poco afectado por la presencia de datos atípicos. Si de una muestra se obtienen la media y la mediana y sus valores difieren sustancialmente, esto será indicativo de la presencia de datos atípicos.

- **Moda**. Para variables discretas o cualitativas, la moda es el valor o valores que más se repiten. Esto implica que la moda no tiene porqué ser única. Para variables cuantitativas continuas, el intervalo modal es aquel con mayor frecuencia. La moda se denotará por Mo.

  Si los datos se encuentran agrupados, se puede obtener el intervalo modal como aquel que tiene una mayor frecuencia.

5.1.2 Medidas de posición de tendencia no central

Como medidas de posición de tendencia no central, introduciremos los cuartiles, deciles y percentiles.

- **Cuartiles**. Los cuartiles Q_1, Q_2 y Q_3 dividen la muestra en cuatro partes iguales, de manera que por debajo de Q_1 tenemos el 25 % de los datos, entre Q_1 y Q_2 se encuentra otro 25 % y por encima de Q_3 otro 25 %. La idea de dividir la muestra en partes iguales se puede generalizar a la construcción de los deciles ($d_1, \ldots, d_9$, dividen la muestra el 10 partes iguales) y los percentiles ($p_1, \ldots, p_{99}$, dividen la muestra el 100 partes iguales).

  En general, se define el cuantil de orden p ($0 < p < 1$) como el valor que deja por debajo (a lo sumo) np observaciones (por tanto, $n(p-1)$ observaciones por encima). El cuantil p se denotará por q_p.

5.1.3 Medidas de dispersión absolutas

Las medidas de posición o localización indican en torno a qué valores se sitúan los datos, pero para obtener una descripción más precisa de los mismos, es necesario conocer cuál es la dispersión que presentan. Las medidas de dispersión absolutas dependen de las unidades en las que se miden las observaciones, siendo las más conocidas la varianza muestral y la desviación típica muestral, que no es más que la raíz cuadrada de la varianza muestral.

- **Varianza (s^2) y desviación típica (s)**. La varianza, s^2, se calcula como:

$$s^2 = \frac{(x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

La varianza está medida en las unidades de los datos al cuadrado, por lo que no se puede comparar directamente con las medidas de posición, por ejemplo, con la media. Para obtener una medida en las unidades de los datos, se considera la desviación típica:

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

Dada una muestra $x_1, \ldots, x_n$, si consideramos la media $\bar{x}$ como medida de posición de tendencia central, se podría pensar en medir la dispersión a través de las diferencias de los valores a la media: $(x_i - \bar{x})$, para todo $i = 1, \ldots, n$. Una forma de contabilizar todas estas diferencias sería a través de la suma: $\sum_{i=1}^{n}(x_i - \bar{x})$. Sin embargo, en este caso es previsible que muestras grandes nos den valores altos de esta suma de diferencias, por la intervención de un mayor número de datos. Para corregir el efecto del número de datos, se podría pasar a un promedio, de manera que la dispersión se mediría a través de: $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})$. Por las propiedades de la media muestral, vimos que la media de las diferencias con respecto a la media es nula, así que esta expresión siempre resultará cero. En este caso, las diferencias positivas y negativas a la media se compensan, por lo que para hacerlas positivas podríamos pensar en medir estas diferencias al cuadrado: $(x_i - \bar{x})^2$. De este modo se obtiene la varianza.
La varianza tiene las siguientes propiedades, fáciles de deducir a partir de la definición.

- Toma valores no negativos, puesto que se trata de un promedio de valores no negativos (diferencias al cuadrado).
- La varianza no es lineal. Si consideramos los datos $y_i = ax_i + b$, la varianza de los nuevos datos será $s_y^2 = a^2 s_x^2$. Es decir, la varianza no se ve afectada por traslaciones (sumar o restar una constante), pero sí por los cambios de escala al multiplicar los valores por un factor.
- Una expresión alternativa para el cálculo de la varianza es:

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

Aunque la varianza es la medida más común, en capítulos posteriores se introducirá una nueva medida de dispersión, denominada cuasi-varianza:

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

La diferencia entre varianza y cuasi-varianza radica en el denominador. En la varianza, se hace un promedio, dividiendo por el número de datos. En la cuasi-varianza, se divide por el número de datos de los que obtenemos información, sabiendo la media.

Consideremos el siguiente ejemplo: supongamos que tenemos una muestra de tamaño $n = 4$, $\{2, x, 6, 8\}$ cuya media es $\bar{x} = 5'25$. Con esta información es fácil deducir que $x = 5$. En general, si se conoce el valor de la media y $(n-1)$ valores de la muestra, podemos determinar el que falta. Esta corrección es importante en muestras pequeñas o de tamaño moderado. Al igual que para la varianza, también se puede definir la cuasi-desviación típica, S.

Otras medidas de dispersión absoluta (es decir, que también dependen de las unidades de los datos) son el rango muestral (R) y el rango intercuartílico (RIC):

$$R = \text{máx}\{x_i\} - \text{mín}\{x_i\}, \quad RIC = Q_3 - Q_1.$$

Para el cálculo del rango se utilizan sólo dos observaciones, la más grande y la más pequeña, por lo que se ve afectado por la presencia de datos atípicos.
Aunque las aquí expuestas son las medidas de dispersión absolutas más usuales, también existen otras medidas de dispersión que en lugar de incluir un cuadrado para evaluar las diferencias entre los datos y las medidas de centralización (en el caso de la varianza, las diferencias entre los datos y la media) utilizan un valor absoluto. Así, se tienen la desviación absoluta con respecto a la media y la desviación absoluta con respecto a la mediana:

$$D_{\overline{x}} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \overline{x}|, \quad D_{Me} = \frac{1}{n}\sum_{i=1}^{n}|x_i - Me|.$$

Una medida de dispersión robusta (poco influenciada por la presencia de datos atípicos) es la *MEDA* que se calcula como:

$$MEDA = Me\{|x_i - Me|;\ i = 1, \ldots, n\}.$$

5.1.4 Medidas de dispersión relativa

Las medidas de dispersión absolutas dependen de las unidades de los datos, por lo que no son adecuadas para comparar variables. Una de las medidas de dispersión relativa (no depende de las unidades de los datos) mas usual es el coeficiente de variación:

$$CV = \frac{s}{\bar{x}}.$$

El coeficiente de variación permiten comparar variables aunque estas estén registradas en distintas unidades de medida. También es de utilidad para comparar variables que, aunque de la misma magnitud, están en escalas distintas. Por ejemplo, para comparar las longitudes del diámetro del tímpano (normalmente, entre 8 y 10 milímetros) y de la columna vertebral (en centímetros), podríamos transformar todas las observaciones a la misma escala pero seguramente la dispersión (medida en desviación típica) que encontraríamos en las longitudes del diámetro del tímpano sería prácticamente nula.

5.1.5 Medidas de forma

Consideraremos dos medidas que proporcionan una idea de la forma de cómo se distribuyen los datos. Su cálculo no es tan sencillo como el de las medidas de posición y dispersión estudiadas y lo que nos interesa es su interpretación.

- **Coeficiente de asimetría**. El coeficiente de asimetría de Fisher toma valor 0 cuando la distribución de los datos es simétrica con respecto a la media. Valores positivos de este coeficiente indicarán la presencia de asimetría positiva (más datos con valores superiores a la media), mientras que valores negativos son indicativos de una asimetría negativa (más datos con valores inferiores a la media). Se calcula como:

$$\gamma_F = \frac{1}{s^3}\frac{(x_1 - \bar{x})^3 + \ldots + (x_n - \bar{x})^3}{n} = \frac{1}{s^3}\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3.$$

  Para cuantificar la asimetría de unos datos, podemos utilizar los cuartiles. Si la distribución es simétrica, la distancia entre Q_3 y Q_2 (que contiene un 25 % de la muestra) y entre Q_2 y Q_1 (otro 25 %), debería ser la misma (es decir, $Q_3 - Q_2 = Q_2 - Q_1$). Así, si $Q_3 - Q_2 > Q_2 - Q_1$, es indicativo de asimetría positiva. Por otro lado, si $Q_3 - Q_2 < Q_2 - Q_1$, tendríamos indicios de asimetría negativa. Para que el resultado no dependa de la dimensión de los datos, podemos utilizar el siguiente índice de asimetría que toma valores en $[-1, 1]$, basado en los cuartiles:

$$\gamma_Q = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)}.$$

Otro coeficiente de asimetría, que resulta útil en el caso de que los datos presenten una única moda. El coeficiente de asimetría de Pearson viene dado por:

$$\gamma_{Mo} = \frac{\overline{x} - Mo}{s}.$$

Basado en la mediana, tenemos el siguiente índice:

$$\gamma_{Me} = \frac{3(\overline{x} - Me)}{s}.$$

- **Coeficiente de curtosis**. El coeficiente de curtosis mide el grado de apuntamiento de la distribución. Su fórmula es:

$$\gamma_C = \frac{1}{s^4}\frac{(x_1 - \bar{x})^4 + \ldots + (x_n - \bar{x})^4}{n} = \frac{1}{s^4}\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4.$$

Si $\gamma_C > 3$, se dice que la distribución de frecuencias es leptocúrtica. Si $\gamma_C < 3$, la distribución de frecuencias es platicúrtica. También se puede modificar la expresión anterior y considerar $\gamma_C^* = \gamma_C - 3$, ya que 3 es el valor del coeficiente cuando los datos vienen de una distribución Normal (que es la de referencia). De este modo, tendremos distribuciones leptocúrticas si $\gamma_C^* > 0$ y platicúrticas si $\gamma_C^* < 0$.

5.2 Representación de medidas: el diagrama de caja

Las representaciones gráficas que se han descrito en la sección anterior utilizan los datos observados para su construcción o la información que se obtiene en las tablas de frecuencias. A partir de las medidas características que se han descrito, se puede construir una nueva representación, el diagrama de caja.
El diagrama de caja se construye a partir de las siguientes medidas:

- El primer y el tercer cuartil, Q_1 y Q_3, que delimitan la caja central (véase Figura 4). La longitud de la caja viene dada por el RIC, que es una medida de dispersión absoluta.
- Los límites inferior y superior (en la Figura 4, son los segmentos horizontales superior e inferior) se calculan como:

$$LI = \text{máx}\{\text{mín}\{x_i\}, Q_1 - 1.5(Q_3 - Q_1)\},$$

$$LS = \text{mín}\{\text{máx}\{x_i\}, Q_3 + 1.5(Q_3 - Q_1)\}.$$

  En el cálculo de los límites inferior y superior se utiliza el $RIC = Q_3 - Q_1$.
- La mediana (Q_2) se representa con una línea horizontal en la caja central.

El diagrama de caja se utiliza para determinar los valores atípicos de la muestra, que son datos que difieren numéricamente de los demás. Formalmente, los datos atípicos son aquellos datos que quedan fuera del intervalo (LI, LS). Si en lugar de considerar los límites inferior y superior construimos el intervalo (LI_e, LS_e) donde $LI_e = Q_1 - 3RIC$ y $LS_e = Q_3 + 3RIC$, los datos que caen fuera de este intervalo se denominan extremos. Algunos paquetes estadísticos hacen la distinción entre atípicos y extremos, representándolos de distintas formas en las salidas gráficas.
En la Figura 4 se puede observar la presencia de datos atípicos altos, representados con puntos. Sin embargo, un problema del diagrama de caja es que no permiten observar la presencia de multimodalidad.

5.3 Tipificación de datos

El coeficiente de variación, como ya hemos visto, se utiliza para comparar la dispersión de variables. Si lo que queremos es comparar individuos de distintos grupos, debemos utilizar la tipificación de datos. A partir de una

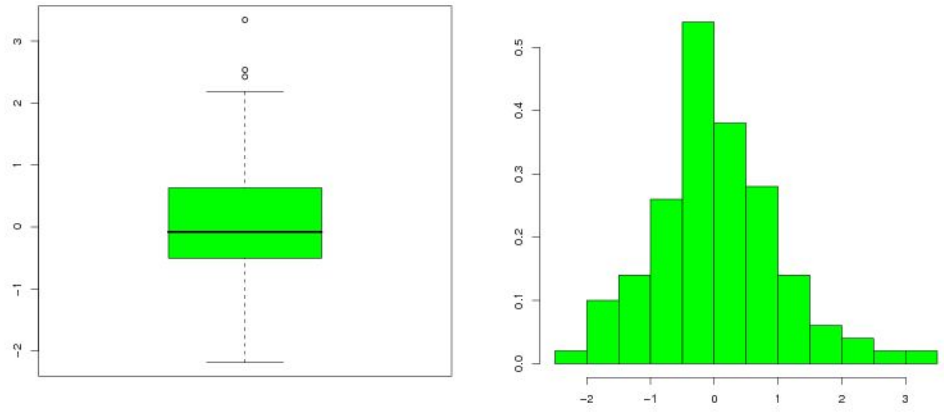


Figura 4: Diagrama de caja e histograma correspondiente.

muestra $x_1, \ldots, x_n$ con media $\overline{x}$ y varianza s^2, los datos tipificados se construyen como:

$$z_i = \frac{x_i - \bar{x}}{s}$$

de manera que la muestra resultante $z_1, \ldots, z_n$ tendrá media 0 y varianza 1. La tipificación de datos permite comparar distintos grupos, así como la posición relativa de las observaciones dentro de cada uno.

5.4 Desigualdad de Tchebychev

La desigualdad de Tchebychev permite construir intervalos centrados en la media y con amplitudes proporcionales a la desviación típica que contienen (al menos) un determinado porcentaje de las observaciones.

En una muestra $x_1, \ldots, x_n$ con media $\overline{x}$ y varianza s^2, en el intervalo $(\overline{x} - ks, \overline{x} + ks)$ tendremos al menos el $100(1 - 1/k^2) \times 100\,\%$ de los datos. Si tomamos $k = 2$, tendremos al menos el 75 % de las observaciones; si $k = 3$, tendremos en el intervalo al menos el 88'89 % de los datos y así sucesivamente.

6 Recta de regresión

Existen muchas situaciones que requieren el análisis combinado de dos ó más variables, debido a las posibles relaciones entre ellas. Para variables cuantitativas (continuas), una forma de representar la dependencia entre ellas es a través de la recta de regresión. En esta sección introduciremos las medidas características usuales en este contexto (vector de medias y matriz de varianzas-covarianzas) y veremos cómo se construye una recta de regresión.

6.1 Vector de medias. Covarianza y correlación

Supongamos que tenemos una variable bidimensional (X, Y) y que disponemos de las observaciones en una muestra de tamaño n, $\{(x_i, y_i)\}_{i=1}^n$. Se denomina **vector de medias** al vector cuyas componentes son las medias muestrales de las variables: $(\bar{x}, \bar{y})$.
Para representar la dispersión podemos considerar los valores de las varianzas de cada variable por separado, es decir, s_x^2 y s_y^2, pero quedaría sin resumir la variabilidad conjunta de ambas. Por eso debemos introducir la

covarianza. La covarianza entre dos variables X e Y, que es una medida que indica la variabilidad conjunta de X e Y. Se calcula como:

$$S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{xy}.$$

A partir de las varianzas y la covarianza se obtiene la **matriz de varianzas-covarianzas**:

$$S = \begin{pmatrix} s_x^2 & S_{xy} \\ S_{xy} & s_y^2 \end{pmatrix}$$

Covarianza y correlación

El signo de la covarianza proporciona información sobre el tipo de relación que puede existir entre las variables. De este modo:

a) Si la relación entre las variables es directa, entonces $S_{xy} > 0$.
b) Si la relación entre las variables es inversa, entonces $S_{xy} < 0$.
c) Si no existe relación lineal entre las variables, entonces $S_{xy} = 0$.

Las parejas de datos datos (x_i, y_i) con $i = 1, \ldots, n$, de las dos variables (X, Y) (también llamada variable bidimensional), se pueden representar a partir de una **nube de puntos** o **diagrama de dispersión**. Esta representación gráfica se construye representando sobre un plano los valores de los puntos observados. En la Figura 5 podemos ver dos ejemplos de relaciones entre variables. La covarianza de los datos de la izquierda es positiva, mientras que la covarianza de los datos de la derecha es negativa. Así, diremos que la relación entre X e Y es directa cuando valores altos de X se corresponden con valores altos de Y. La relación se dice que es inversa si valores altos de X se corresponden con valores bajos de Y, o viceversa.

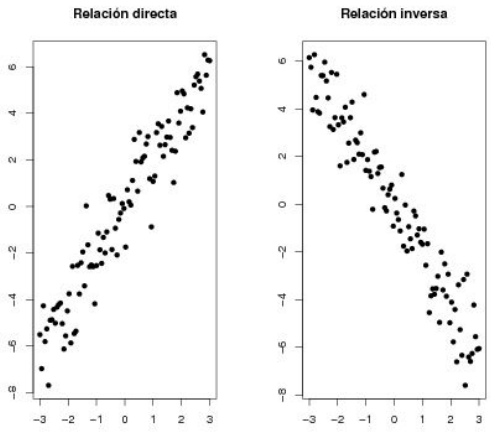


Figura 5: Ejemplo de diagramas de dispersión. Relaciones directa e inversa.

La covarianza está afectada por las unidades de medida de las variables, por lo que definiremos una medida característica para explicar la relación lineal entre variables que sea adimensional: el **coeficiente de correlación lineal**. A partir de una muestra de datos $\{(x_i, y_i)\}_{i=1}^{n}$, el coeficiente de correlación lineal se calcula como:

$$r = \frac{S_{xy}}{s_x s_y},$$

donde S_{xy} es la covarianza muestral y s_x, s_y son las respectivas desviaciones típicas muestrales.

El coeficiente de correlación lineal no tiene dimensiones y toma valores en $[-1, 1]$. Valores cercanos a 1 nos indicarían una relación lineal directa, mientras que valores cercanos a -1 darían una relación lineal inversa. En la práctica, si el coeficiente de correlación $r = 0$, esto indica que no existe relación lineal entre las variables, pero podría ocurrir que entre ellas hubiese otro tipo de relación no lineal. Observa que r sólo cuantifica relaciones lineales.
Cuando existe una relación lineal entre dos variables, podemos tratar de buscar un modelo que describa una en función de otra. La regresión lineal simple consiste en aproximar los valores de una variable a partir de los de otra utilizando una relación de tipo lineal. La recta de regresión de Y sobre X tendrá la siguiente expresión:

$$y = a + bx,$$

donde a representa la ordenada en el origen o intercepto y b es la pendiente (indica la razón de cambio en Y cuando X varía en una unidad). Esta expresión nos dice que, cuando $x = 0$, entonces $y = a$. La variable X se denomina variable explicativa o independiente, mientras que la variable Y será la variable respuesta, o variable dependiente.

6.2 Método de Mínimos Cuadrados

En la práctica, a partir de los datos $\{(x_i, y_i)\}_{i=1}^n$ podremos calcular los valores de a y b. El objetivo será obtener los valores a y b que nos proporcionen los residuos más pequeños. Los residuos son las diferencias entre los valores observados de la variable respuesta y_i y los valores que proporciona el ajuste $\hat{y}_i = a + bx_i$ y vienen dados por:

$$e_i = y_i - \hat{y}_i = y_i - a - bx_i, \quad i = 1, \ldots, n.$$

En la Figura 6, los segmentos verticales son los residuos, que representan la diferencia entre el valor observado y el valor que daría la recta ajustada.

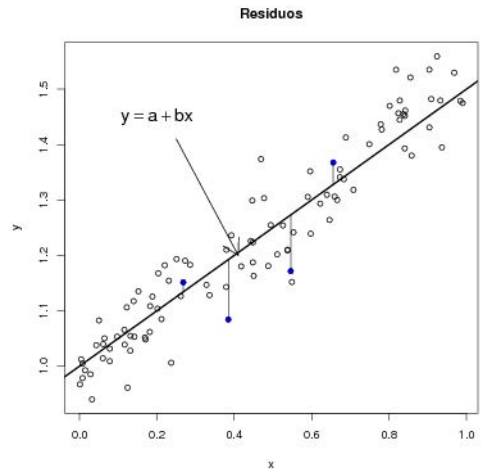


Figura 6: Residuos a minimizar en el Método de Mínimos Cuadrados. Los segmentos verticales representan los residuos e_i.

El **Método de Mínimos Cuadrados** consiste en minimizar la suma de los cuadrados de los residuos, por lo que se buscan los valores a y b que minimizan:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2.$$

A partir del Método de Mínimos Cuadrados, se obtienen los valores para a y b:

$$b = \frac{S_{xy}}{s_x^2}, \quad a = \overline{y} - b\overline{x},$$

donde $\overline{y}$ y $\overline{x}$ denotan las medias muestrales de $y_1, \ldots, y_n$ y $x_1, \ldots, x_n$, respectivamente; s_x^2 es la varianza muestral de X :

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

y S_{xy} es la **covarianza** muestral entre X e Y. En la Figura 6, representamos la recta ajustada, con a y b obtenidos por el método de Mínimos Cuadrados. Se puede comprobar que la recta de regresión ajustada por Mínimos Cuadrados pasa por el vector de medias $(\overline{x}, \overline{y})$.

La recta de regresión de Y sobre X se puede utilizar para predecir valores de Y conocidos los valores de X, pero no al revés. En su construcción, el parámetro pendiente tiene en cuenta la varianza de la variable explicativa s_x^2. Además, las predicciones con la recta de Y sobre X sólo son razonables cuando el valor de X para el que queremos hacer la predicción se encuentra entre el mínimo y el máximo de los valores observados para la variable. Si quisiéramos hacer predicciones sobre el valor de X dado un valor de Y, tendríamos que utilizar la recta de regresión:

$$x = c + dy, \quad \text{con} \quad d = \frac{S_{xy}}{s_y^2},\ c = \overline{x} - d\overline{y}.$$

6.3 Coeficiente de regresión. Coeficiente de determinación

Coeficiente de regresión.

Se denomina **coeficiente de regresión** a la pendiente (parámetro b) de la recta de regresión de Y sobre X. Este coeficiente proporciona información sobre el comportamiento de la variable respuesta Y en función de la variable explicativa X y tiene el mismo signo que la covarianza.

a) Si $b > 0$, al aumentar los valores de X también aumentan los valores de Y.

b) Si $b < 0$, al aumentar los valores de X, los valores de Y disminuyen.

Coeficiente de determinación

Una medida para determinar cómo de bueno es el ajuste del modelo es el **coeficiente de determinación** (r^2) que mide la proporción de variabilidad de Y que explica X a través de la recta de regresión.
El coeficiente de determinación es el cuadrado del coeficiente de correlación lineal, y toma valores entre 0 y 1. Si r^2 toma valores próximos a 1, esto será indicativo de un buen ajuste. El coeficiente de determinación del modelo de regresión lineal simple viene dado por:

$$r^2 = \frac{S_{xy}^2}{s_x^2 s_y^2}.$$

El coeficiente de determinación, y por tanto, la variabilidad explicada por la recta de regresión de Y sobre X y la de X sobre Y es el mismo.