EEBE Estadística

Alejandro Cáceres (alejandro.caceres.dominguez@upc.edu)

2023-09-11

Contents

Chapter 1

Objetivo

Este es el curso de introducción a la estadística de la EEBE (UPC).

La estadística es un **lenguaje** que permite afrontar problemas nuevos, sobre los que no tenemos solución, y en donde interviene la **aleatoridad**.

En este curso trataremos los **conceptos fundamentales** de estadística.

- 3 horas de **teoría** por semana: Explicaremos los conceptos, haremos ejercicios.
- 6 horas de **estudio individual** por semana: Notas notas de curso y los recursos en ATENEA.
- 2 horas de Solución de problemas con **R**: Sesiones presenciales con ordenador (Prácticas).

Las fechas de exámenes y material de estudio adicional se pueden encontrar en **ATENEA metacurso**:

Activitat	Data	Pes
Q1 (T1 – T2)	**11/10/2023** (00:05) – **13/10/2023** (23:55)	10%
EP1 (T3 – T4)	**19/10/2023**, 15.30 h	25%
Q2 (T5 - T6)	**21/11/2023 – 4/12/2023** (en hora de clase)	20%
EP2 (T7 - T8)	**18/01/2024**, 16.00 h	40%
CG	**18/01/2024**, 16.00 h	5%

EP1: Evaluación presencial escrita
EP2: Evaluación presencial con ordenador o tablet que el estudiantado llevará a la prueba

Q1: Cuestionario asíncrono
Q2: Cuestionario síncrono

CG: Competencia Genérica

Objetivos de evaluación:

Q1 (10%): Prueba en ordenador duración 2h en las fechas indicadas.

a. Dominio de comandos básicos en R (Prácticas)
b. Capacidad de calcular estadísticos descriptivos y gráficos, en situaciones concretas (Teoría/Práctica)
c. Conocimiento sobre la regresión lineal (Prácticas)

EP1 (25%): Prueba escrita (2-3 problemas)

a. Capacidad de interpretación de enunciados en fórmulas de probabilidad (Teoría).
b. Conocimiento de las herramientas básicas para solucionar problemas de probabilidad conjunta y probabilidad condicional (Teoría).
c. Dominio matemático de funciones de probabilidad para calcular sus propiedades básicas (Teoría).

Q2 (20%): Prueba en ordenador duración 2h en horario de clase en las fechas indicadas

a. Capacidad de identificación de modelos de probabilidad en problemas concretos (Teoría/Práctica).
b. Uso de funciones de R para calcular probabilidades de modelos probabilísticos (Práctica/Teoría)
c. Capacidad de identificación de un estadístico de muestreo y sus propiedades (Teoría/Práctica)
d. Conocimiento de cómo calcular la probabilidad de los estadísticos de muestreo (Teoría/Práctica)
e. Uso de comandos en R para calcular probabilidades y hacer simulaciones de muestras aleatorias (Prácticas)

EP2 (40%): Prueba escrita (2-3 problemas)

a. Capacidad matemática para determinar estimadores puntuales de modelos de probabilidad.
b. Conociemiento de las propiedades de los estimadores puntuales.
c. Conocimiento de los intervalos de confianza y sus propiedades (Teoría).
d. Capacidad de identificar el tipo de intervalo de confianza en un problema concreto (Teoría).
e. Capacidad de interpretación del tipo de hipótesis a usar en un problema concreto (Teoría).
f. Uso de comandos en R para resolver problemas de intervalos de confianza y prueabas de hipótesis (Práctica).

CG (5%): Prueba escrita (2 preguntas sobre un texto)

a. Capacidad de expresión escrita sobre un tema relacionado a la estadística.

coordinadores:

- Luis Mujica (luis.eduardo.mujica@upc.edu)
- Pablo Buenestado (pablo.buenestado@upc.edu)

1.1 Lectura recomendada

- Los apuntes de clase se nuestra sección estarán accesibles en ATENEA en pdf y en html.
- Douglas C. Montgomery and George C. Runger. "Applied Statistics and Probability for Engineers" 4th Edition. Wiley 2007.

Chapter 2

Descripción de datos

En este capítulo, presentaremos herramientas para describir datos.

Lo haremos utilizando tablas, figuras y estadísticos descriptivos de tendencia central y dispersión.

También presentaremos conceptos clave en estadística como experimentos aleatorios, observaciones, resultados y frecuencias absolutas y relativas.

2.1 Método científico

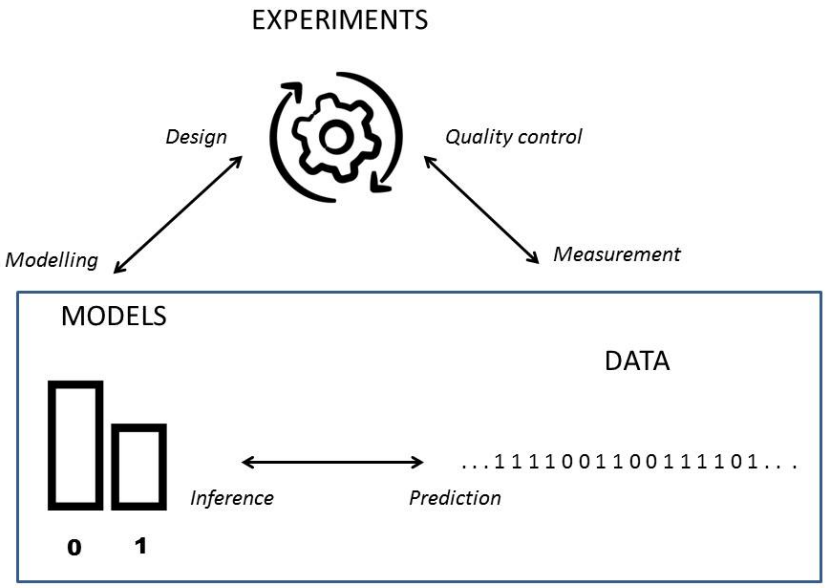
Uno de los objetivos del método científico es proporcionar un marco para resolver los problemas que surgen en el estudio de los fenómenos naturales o en el diseño de nuevas tecnologías.

Los humanos modernos han desarrollado un **método** durante miles de años que todavía está en desarrollo.

El método tiene tres actividades humanas principales:

- *Observación* caracterizada por la adquisición de **datos**
- *Razón* caracterizada por el desarrollo de **modelos** matemáticos
- *Acción* caracterizada por el desarrollo de nuevos **experimentos** (tecnología)

Su compleja interacción y resultados son la base de la *actividad científica*.



2.2 Estadística

La estadística se ocupa de la interacción entre *modelos* y *datos* (la parte inferior de la figura).

Las preguntas de tipo estadístico son:

- ¿Cuál es el mejor modelo para mis datos (inferencia)?
- ¿Cuáles son los datos que produciría un determinado modelo (predicción)?

2.3 Datos

Los datos se presentan en forma de observaciones.

Una **Observación** o *Realización* es la adquisición de un número o una característica de un experimento.

Por ejemplo, tomemos la serie de números que se producen por la repetición de un experimento (1: éxito, 0: fracaso)

… 1 0 0 1 0 **1** 0 1 1 …

El número en negrita es **una observación** en una repetición del experimento

Un **resultado** es una **posible** observación que es el resultado de un experimento.

1 es un resultado, **0** es el otro resultado del experimento.

Recuerda que la observación es **concreta** es el número que obtienes un día en el laboratorio. El resultado **abstracto** es una de las características del tipo de experimento que estás realizando.

2.4 Tipos de resultado

En estadística nos interesan principalmente dos tipos de resultados.

- **Categóricos**: Si el resultado de un experimento es una cualidad. Pueden ser nominales (binario: sí, no; múltiple: colores) u ordinales cuando las cualidades pueden jerarquizarse (gravedad de una enfermedad).
- **Numéricos**: Si el resultado de un experimento es un número. El número puede ser discreto (número de correos electrónicos recibidos en una hora, número de leucocitos en sangre) o continuo (estado de carga de la batería, temperatura del motor).

2.5 Experimentos aleatorios

Se puede decir que el tema de estudio de la estadística son los experimentos aleatorios, el medio por el cual producimos datos.

Definición:

Un **experimento aleatorio** es un experimento que da diferentes resultados cuando se repite de la misma manera.

Los experimentos aleatorios son de diferentes tipos, dependiendo de cómo se realicen:

- en el mismo objeto (persona): temperatura, niveles de azúcar.
- sobre objetos diferentes pero de la misma medida: el peso de un animal.
- sobre eventos: el número de huracanes por año.

2.6 Frecuencias absolutas

Cuando repetimos un experimento aleatorio con resultados **categóricos**, registramos una lista de resultados.

Resumimos las observaciones contando cuántas veces vimos un resultado particular.

Frecuencia absoluta:

$$n_i$$

es el número de veces que observamos el resultado i.

Ejemplo (leucocitos)

Extraigamos un leucocito de **un** donante y anotemos su tipo. Repitamos el experimento $N = 119$ veces.

```
(célula T, célula T, neutrófilo, ..., célula B)
```

La segunda **célula T** en negrita es la segunda observación. La última **célula B** es la observación número 119.

Podemos listar los **resultados** (categorías) en una **tabla de frecuencia**:

```
##      outcome ni
## 1     T Cell 34
## 2     B cell 50
## 3   basophil 20
## 4   Monocyte  5
## 5 Neutrophil 10
```

De la tabla, podemos decir que, por ejemplo, $n_1 = 34$ es el número total de células T observadas en la repetición del experimento. También notamos que el número total de repeticiones $N = \sum_i n_i = 119$.

2.7 Frecuencias relativas

También podemos resumir las observaciones calculando la **proporción** de cuántas veces vimos un resultado en particular.

$$f_i = n_i/N$$

donde N es el número total de observaciones

En nuestro ejemplo se registraron $n_1 = 34$ células T, por lo que nos preguntamos por la proporción de células T del total de 119. Podemos agregar estas proporciones f_i en la tabla las frecuencias.

```
##      outcome ni         fi
## 1     T Cell 34 0.28571429
## 2     B cell 50 0.42016807
## 3   basophil 20 0.16806723
## 4   Monocyte  5 0.04201681
## 5 Neutrophil 10 0.08403361
```

Las frecuencias relativas son **fundamentales** en estadística. Dan la proporción de un resultado en relación con los otros resultados. Más adelante las entenderemos como las observaciones de las probabilidades.

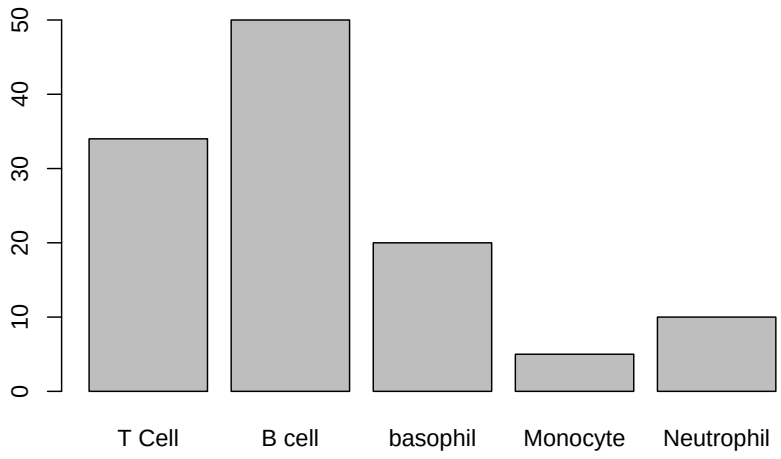
Para las frecuencias absolutas y relativas tenemos las propiedades

- $\sum_{i=1..M} n_i = N$
- $\sum_{i=1..M} f_i = 1$

donde M es el número de resultados.

2.8 Diagrama de barras

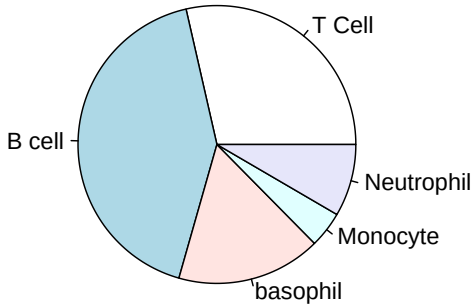
Cuando tenemos muchos resultados y queremos ver cuáles son los más probables, podemos usar un gráfico de barras que es una cifra de n_i Vs los resultados.



2.9 Gráfico de sectores (pie)

También podemos visualizar las frecuencias relativas con un gráfico de sectores.

El área del círculo representa el 100% de las observaciones (proporción = 1) y las secciones las frecuencias relativas de cada resultado.



2.10 Variables categóricas ordinales

El tipo de leucocito de los ejemplos anteriores es una variable nominal **categórica**. Cada observación pertenece a una categoría (cualidad). Las categorías no siempre tienen un orden determindado.

A veces, las variables **categóricas** se pueden **ordenar** cuando cumplen una clasificación natural. Esto permite introducir **frecuencias acumulativas**.

Ejemplo (misofonía)

Este es un estudio clínico en 123 pacientes que fueron examinados por su grado de misofonía. La misofnía es ansiedad/ira descontrolada producida por ciertos sonidos.

Cada paciente fue evaluado con un cuestionario (AMISO) y se clasificaron en 4 grupos diferentes según la gravedad.

Los resultados del estudio son

```
##   [1] 4 2 0 3 0 0 2 3 0 3 0 2 2 0 2 0 0 3 3 0 3 3 2 0 0 0 4 2 2 0 2 0 0 0 3 0 2
##  [38] 3 2 2 0 2 3 0 0 2 2 3 3 0 0 4 3 3 2 0 2 0 0 0 2 2 0 0 2 3 0 1 3 2 4 3 2 3
##  [75] 0 2 3 2 4 1 2 0 2 0 2 0 2 2 4 3 0 3 0 0 0 2 2 1 3 0 0 3 2 1 3 0 4 4 2 3 3
## [112] 3 0 3 2 1 2 3 3 4 2 3 2
```

Cada observación es el resultado de un experimento aleatorio: medición del nivel de misofonía en un paciente. Esta serie de datos se puede resumir en términos

de los resultados en la tabla de frecuencia

```
##   outcome ni         fi
## 1       0 41 0.33333333
## 2       1  5 0.04065041
## 3       2 37 0.30081301
## 4       3 31 0.25203252
## 5       4  9 0.07317073
```

2.11 Frecuencias acumuladas absolutas y relativas

La gravedad de la misofonía es **categórica ordinal** porque sus resultados pueden ordenarse en relación con su grado.

Cuando los resultados se pueden ordenar, es útil preguntar cuántas observaciones se obtuvieron hasta un resultado dado. Llamamos a este número la **frecuencia acumulada absoluta** hasta el resultado i:

$$N_i = \sum_{k=1..i} n_k$$

También es útil para calcular la **proporción** de las observaciones que se obtuvo hasta un resultado dado

$$F_i = \sum_{k=1..i} f_k$$

Podemos agregar estas frecuencias en la **tabla de frecuencias**

```
##   outcome ni         fi  Ni        Fi
## 0       0 41 0.33333333  41 0.3333333
## 1       1  5 0.04065041  46 0.3739837
## 2       2 37 0.30081301  83 0.6747967
## 3       3 31 0.25203252 114 0.9268293
## 4       4  9 0.07317073 123 1.0000000
```

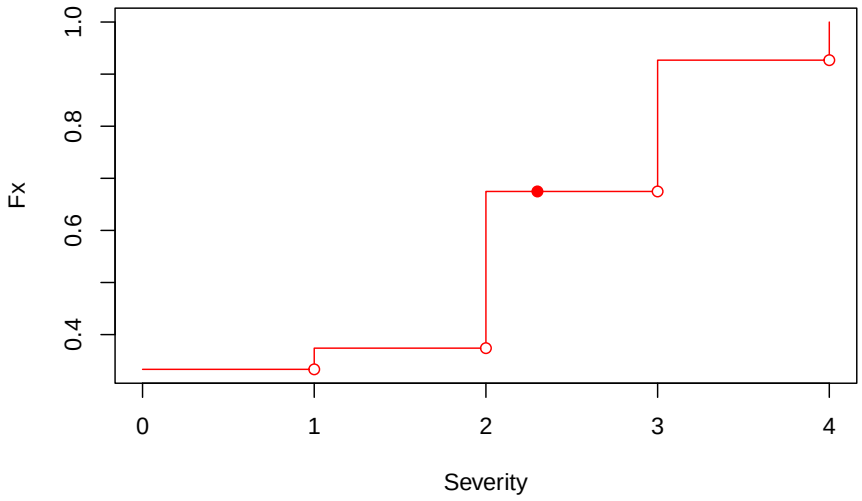
Por lo tanto, el **67 %** de los pacientes tenían misofonía hasta la gravedad **2** y el **37 %** de los pacientes tenían una gravedad inferior o igual a **1**.

2.12 Gráfica de frecuencia acumulada

F_i es una cantidad importante porque nos permite definir la acumulación de probabilidades hasta niveles intermedios.

La probabilidad de un nivel intermedio x ($i \leq x < i+1$) es solo la acumulación hasta el nivel inferior $F_x = F_i$.

F_x es por lo tanto una función de rango **continuo**. Podemos dibujarla con respecto a los resultados.



Por lo tanto, podemos decir que el **67 %** de los pacientes tenían misofonía hasta gravedad 2.3, aunque 2.3 no es un resultado observado.

2.13 Variables numéricas

El resultado de un experimento aleatorio puede producir un número. Si el número es **discreto**, podemos generar una tabla de frecuencias, con frecuencias absolutas, relativas y acumulativas, e ilustrarlas con gráficos de barras, de sectores y acumulativos.

Cuando el número es **continuo** las frecuencias no son útiles, lo más probable es que observemos o no un número contínuo en particular.

Ejemplo (misofonía)

Los investigadores se preguntaron si la convexidad de la mandíbula afectaría la gravedad de la misofonía. La hipótesis científica es que el ángulo de convexidad de la mandíbula puede influir en el oído y su sensibilidad. Estos son los resultados de la convexidad de la mandíbula (grados) para cada paciente:

```
##   [1]  7.97 18.23 12.27  7.81  9.81 13.50 19.30  7.70 12.30  7.90 12.60 19.00
##  [13]  7.27 14.00  5.40  8.00 11.20  7.75  7.94 16.69  7.62  7.02  7.00 19.20
##  [25]  7.96 14.70  7.24  7.80  7.90  4.70  4.40 14.00 14.40 16.00  1.40  9.76
##  [37]  7.90  7.90  7.40  6.30  7.76  7.30  7.00 11.23 16.00  7.90  7.29  6.91
```

```
##  [49]  7.10 13.40 11.60 -1.00  6.00  7.82  4.80 11.00  9.00 11.50 16.00 15.00
##  [61]  1.40 16.80  7.70 16.14  7.12 -1.00 17.00  9.26 18.70  3.40 21.30  7.50
##  [73]  6.03  7.50 19.00 19.01  8.10  7.80  6.10 15.26  7.95 18.00  4.60 15.00
##  [85]  7.50  8.00 16.80  8.54  7.00 18.30  7.80 16.00 14.00 12.30 11.40  8.50
##  [97]  7.00  7.96 17.60 10.00  3.50  6.70 17.00 20.26  6.64  1.80  7.02  2.46
## [109] 19.00 17.86  6.10  6.64 12.00  6.60  8.70 14.05  7.20 19.70  7.70  6.02
## [121]  2.50 19.00  6.80
```

2.14 Transformando datos continuos

Como los resultados continuos no se pueden contar (de manera informativa), los transformamos en variables categóricas ordenadas.

1) Primero cubrimos el rango de las observaciones en intervalos regulares del mismo tamaño (contenedores)

```
## [1] "[-1.02,3.46]" "(3.46,7.92]"  "(7.92,12.4]"  "(12.4,16.8]"  "(16.8,21.3]"
```

2) Luego mapeamos cada observación a su intervalo: creando una variable categórica **ordenada**; en este caso con 5 resultados posibles

```
##   [1] "(7.92,12.4]"  "(16.8,21.3]"  "(7.92,12.4]"  "(3.46,7.92]"  "(7.92,12.4]"
##   [6] "(12.4,16.8]"  "(16.8,21.3]"  "(3.46,7.92]"  "(7.92,12.4]"  "(3.46,7.92]"
##  [11] "(12.4,16.8]"  "(16.8,21.3]"  "(3.46,7.92]"  "(12.4,16.8]"  "(3.46,7.92]"
##  [16] "(7.92,12.4]"  "(7.92,12.4]"  "(3.46,7.92]"  "(7.92,12.4]"  "(12.4,16.8]"
##  [21] "(3.46,7.92]"  "(3.46,7.92]"  "(3.46,7.92]"  "(16.8,21.3]"  "(7.92,12.4]"
##  [26] "(12.4,16.8]"  "(3.46,7.92]"  "(3.46,7.92]"  "(3.46,7.92]"  "(3.46,7.92]"
##  [31] "(3.46,7.92]"  "(12.4,16.8]"  "(12.4,16.8]"  "(12.4,16.8]"  "[-1.02,3.46]"
##  [36] "(7.92,12.4]"  "(3.46,7.92]"  "(3.46,7.92]"  "(3.46,7.92]"  "(3.46,7.92]"
##  [41] "(3.46,7.92]"  "(3.46,7.92]"  "(3.46,7.92]"  "(7.92,12.4]"  "(12.4,16.8]"
##  [46] "(3.46,7.92]"  "(3.46,7.92]"  "(3.46,7.92]"  "(3.46,7.92]"  "(12.4,16.8]"
##  [51] "(7.92,12.4]"  "[-1.02,3.46]" "(3.46,7.92]"  "(3.46,7.92]"  "(3.46,7.92]"
##  [56] "(7.92,12.4]"  "(7.92,12.4]"  "(7.92,12.4]"  "(12.4,16.8]"  "(12.4,16.8]"
##  [61] "[-1.02,3.46]" "(12.4,16.8]"  "(3.46,7.92]"  "(12.4,16.8]"  "(3.46,7.92]"
##  [66] "[-1.02,3.46]" "(16.8,21.3]"  "(7.92,12.4]"  "(16.8,21.3]"  "[-1.02,3.46]"
##  [71] "(16.8,21.3]"  "(3.46,7.92]"  "(3.46,7.92]"  "(3.46,7.92]"  "(16.8,21.3]"
##  [76] "(16.8,21.3]"  "(7.92,12.4]"  "(3.46,7.92]"  "(3.46,7.92]"  "(12.4,16.8]"
##  [81] "(7.92,12.4]"  "(16.8,21.3]"  "(3.46,7.92]"  "(12.4,16.8]"  "(3.46,7.92]"
##  [86] "(7.92,12.4]"  "(12.4,16.8]"  "(7.92,12.4]"  "(3.46,7.92]"  "(16.8,21.3]"
##  [91] "(3.46,7.92]"  "(12.4,16.8]"  "(12.4,16.8]"  "(7.92,12.4]"  "(7.92,12.4]"
##  [96] "(7.92,12.4]"  "(3.46,7.92]"  "(7.92,12.4]"  "(16.8,21.3]"  "(7.92,12.4]"
## [101] "(3.46,7.92]"  "(3.46,7.92]"  "(16.8,21.3]"  "(16.8,21.3]"  "(3.46,7.92]"
## [106] "[-1.02,3.46]" "(3.46,7.92]"  "[-1.02,3.46]" "(16.8,21.3]"  "(16.8,21.3]"
## [111] "(3.46,7.92]"  "(3.46,7.92]"  "(7.92,12.4]"  "(3.46,7.92]"  "(7.92,12.4]"
## [116] "(12.4,16.8]"  "(3.46,7.92]"  "(16.8,21.3]"  "(3.46,7.92]"  "(3.46,7.92]"
## [121] "[-1.02,3.46]" "(16.8,21.3]"  "(3.46,7.92]"
```

Por tanto, en lugar de decir que el primer paciente tenía un ángulo de convexidad de 7.97, decimos que su ángulo estaba entre el intervalo (o **bin**) $(7.92, 12.4]$.

Ningún otro paciente tenía un ángulo de 7.97, pero muchos tenían ángulos entre $(7.92, 12.4]$.

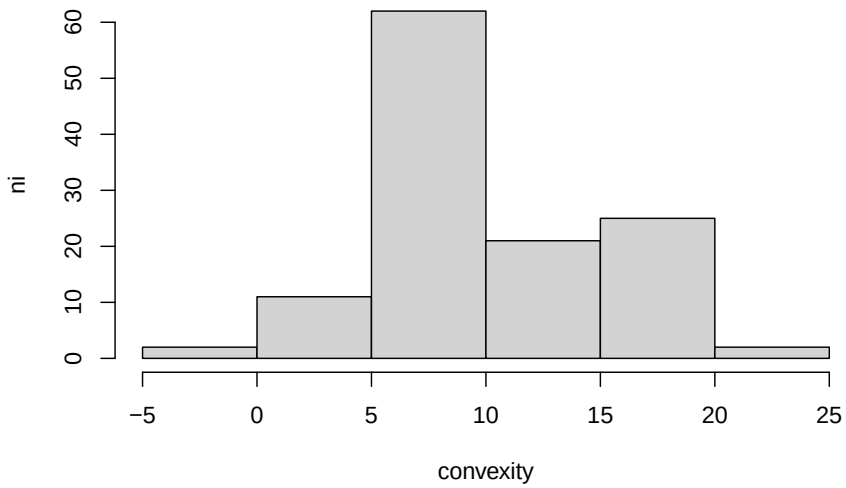
2.15 Tabla de frecuencias para una variable continua

Para una partición regular dada del intervalo de resultados en intervalos, podemos producir una tabla de frecuencias como antes

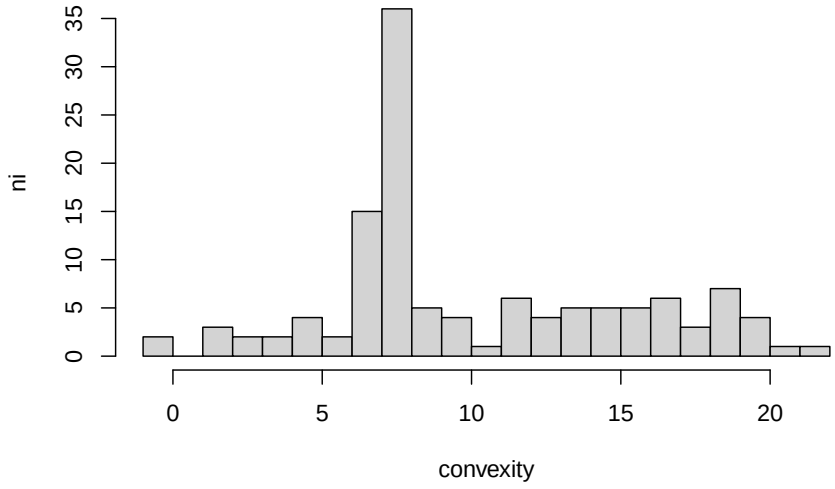
```
##        outcome ni         fi  Ni         Fi
## 1 [-1.02,3.46]  8 0.06504065   8 0.06504065
## 2  (3.46,7.92] 51 0.41463415  59 0.47967480
## 3  (7.92,12.4] 26 0.21138211  85 0.69105691
## 4  (12.4,16.8] 20 0.16260163 105 0.85365854
## 5  (16.8,21.3] 18 0.14634146 123 1.00000000
```

2.16 Histograma

El histograma es la gráfica de n_i o f_i Vs los resultados en intervalos (bins). El histograma depende del tamaño de los bins.



Este es un histograma con 20 bins.



Vemos que la mayoría de las personas tienen ángulos dentro de $(7, 8]$

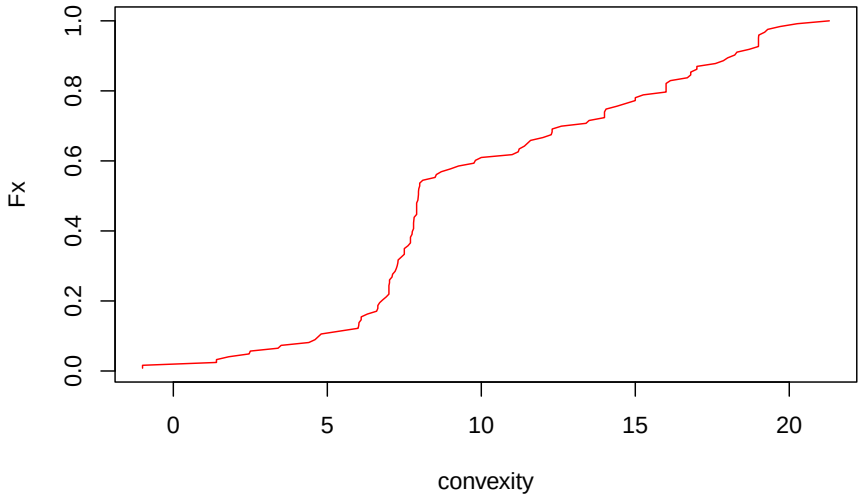
2.17 Gráfica de frecuencia acumulada

También podemos graficar F_x contra los resultados. Como F_x es de rango continuo, podemos ordenar las observaciones $(x_1 < ...x_j < x_{j+1} < x_n)$ y por lo tanto

$$F_x = \frac{k}{n}$$

para $x_k \leq x < x_{k+1}$.

F_x se conoce como la **distribución** de los datos. F_x no depende del tamaño del bin. Sin embargo, su **resolución** depende de la cantidad de datos.



2.18 Estadísticas de resumen

Las estadísticas de resumen son números calculados a partir de los datos que nos dicen características importantes de las variables numéricas (discretas o continuas).

Por ejemplo, tenemos estadísticas que describen los valores extremos:

- **mínimo**: el resultado mínimo observado
- **máximo**: el resultado máximo observado

2.19 Promedio (media muestral)

Una estadística importante que describe el valor central de los resultados (dónde esperar la mayoría de las observaciones) es el **promedio**

$$\bar{x} = \frac{1}{N} \sum_{j=1..N} x_j$$

donde x_j es la **observación** j de un total de N.

Ejemplo (Misofonía)

La convexidad promedio se puede calcular directamente a partir de las **observaciones**

$\bar{x} = \frac{1}{N} \sum_j x_j$

$= \frac{1}{N}(7.97 + 18.23 + 12.27... + 6.80) = 10.19894$

Para variables **categóricamente ordenadas**, podemos usar las frecuencias relativas para calcular el promedio

$\bar{x} = \frac{1}{N} \sum_{i=1...N} x_j = \frac{1}{N} \sum_{i=1...M} x_i * n_i$

$$= \sum_{i=1...M} x_i * f_i$$

donde pasamos de sumar N **observaciones** a sumar M **resultados**.

La forma $\bar{x} = \sum_{i=1...M} x_i f_i$ muestra que el promedio es el **centro de gravedad** de los resultados. Como si cada resultado tuviera una densidad de masa dada por f_i.

Ejemplo (Misofonía)

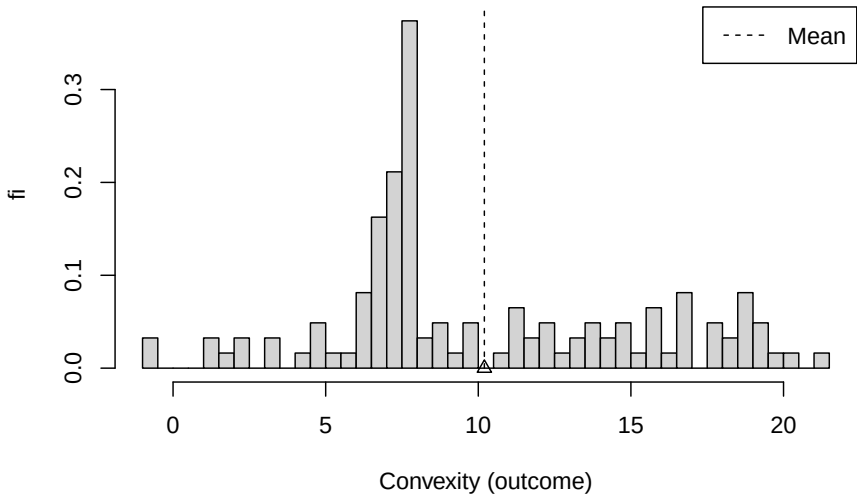
La **severidad** promedio de la misofonía en el estudio se puede calcular a partir de las frecuencias relativas de los **resultados**

```
##   outcome ni         fi
## 1       0 41 0.33333333
## 2       1  5 0.04065041
## 3       2 37 0.30081301
## 4       3 31 0.25203252
## 5       4  9 0.07317073
```

$\bar{x} = 0 * f_0 + 1 * f_1 + 2 * f_2 + 3 * f_3 + 4 * f_4 = 1.691057$

2.20 Promedio

El promedio es también el centro de gravedad de las variables continuas. Ese es el punto donde las frecuencias reativas se equilibran.



2.21 mediana

Otra medida de centralidad es la mediana. La mediana x_m, o $q_{0.5}$, es el valor por debajo del cual encontramos la mitad de las observaciones. Cuando ordenamos las observaciones $x_1 < ...x_j < x_{j+1} < x_N$, las contamos hasta encontrar la mitad de ellas. x_m es tal que

$$\sum_{i \leq m} 1 = \frac{N}{2}$$

Ejemplo (Misofonía)

Si ordenamos los ángulos de convexidad, vemos que 62 observaciones (individuos) ($N/2 \sim 123/2$) están por debajo de 7.96. La **convexidad mediana** es por lo tanto $q_{0.5} = x_{62} = 7.96$

```
##  [1] -1.00 -1.00  1.40  1.40  1.80  2.46  2.50  3.40  3.50  4.40  4.60  4.70
## [13]  4.80  5.40  6.00  6.02  6.03  6.10  6.10  6.30  6.60  6.64  6.64  6.70
## [25]  6.80  6.91  7.00  7.00  7.00  7.00  7.02  7.02  7.10  7.12  7.20  7.24
## [37]  7.27  7.29  7.30  7.40  7.50  7.50  7.50  7.62  7.70  7.70  7.70  7.75
## [49]  7.76  7.80  7.80  7.80  7.81  7.82  7.90  7.90  7.90  7.90  7.90  7.94
## [61]  7.95  7.96
```

```
##  [1]  7.96  7.97  8.00  8.00  8.10  8.50  8.54  8.70  9.00  9.26  9.76  9.81
## [13] 10.00 11.00 11.20 11.23 11.40 11.50 11.60 12.00 12.27 12.30 12.30 12.60
## [25] 13.40 13.50 14.00 14.00 14.00 14.05 14.40 14.70 15.00 15.00 15.26 16.00
```

```
## [37] 16.00 16.00 16.00 16.14 16.69 16.80 16.80 17.00 17.00 17.60 17.86 18.00
## [49] 18.23 18.30 18.70 19.00 19.00 19.00 19.00 19.01 19.20 19.30 19.70 20.26
## [61] 21.30
```

```
## [1] 7.96
```

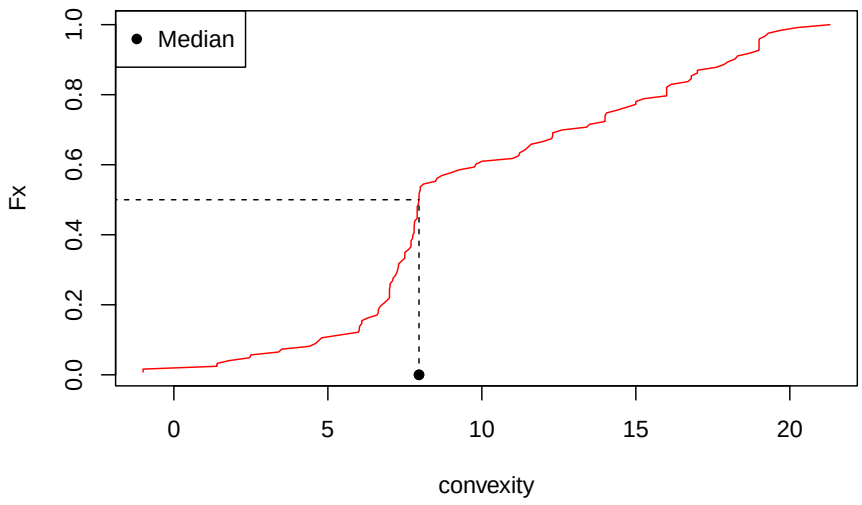
En términos de frecuencias, $q_{0.5}$ hace que la frecuencia acumulada F_x sea igual a 0.5

$$\sum_{i=0,...m} f_i = F_{q_{0.5}} = 0.5$$

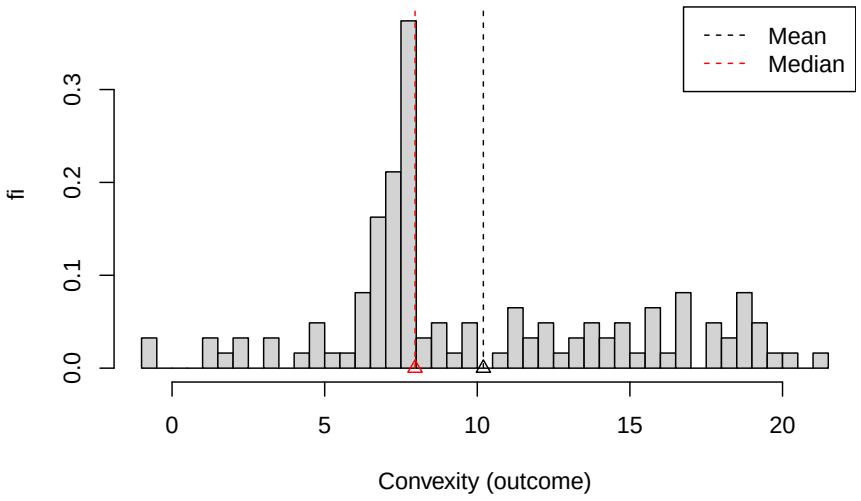
o

$$q_{0.5} = F^{-1}(0.5)$$

En el gráfico de distribución, la mediana es el valor de x en el que se encuentra la mitad del máximo de F.



El promedio y la mediana no siempre son iguales.

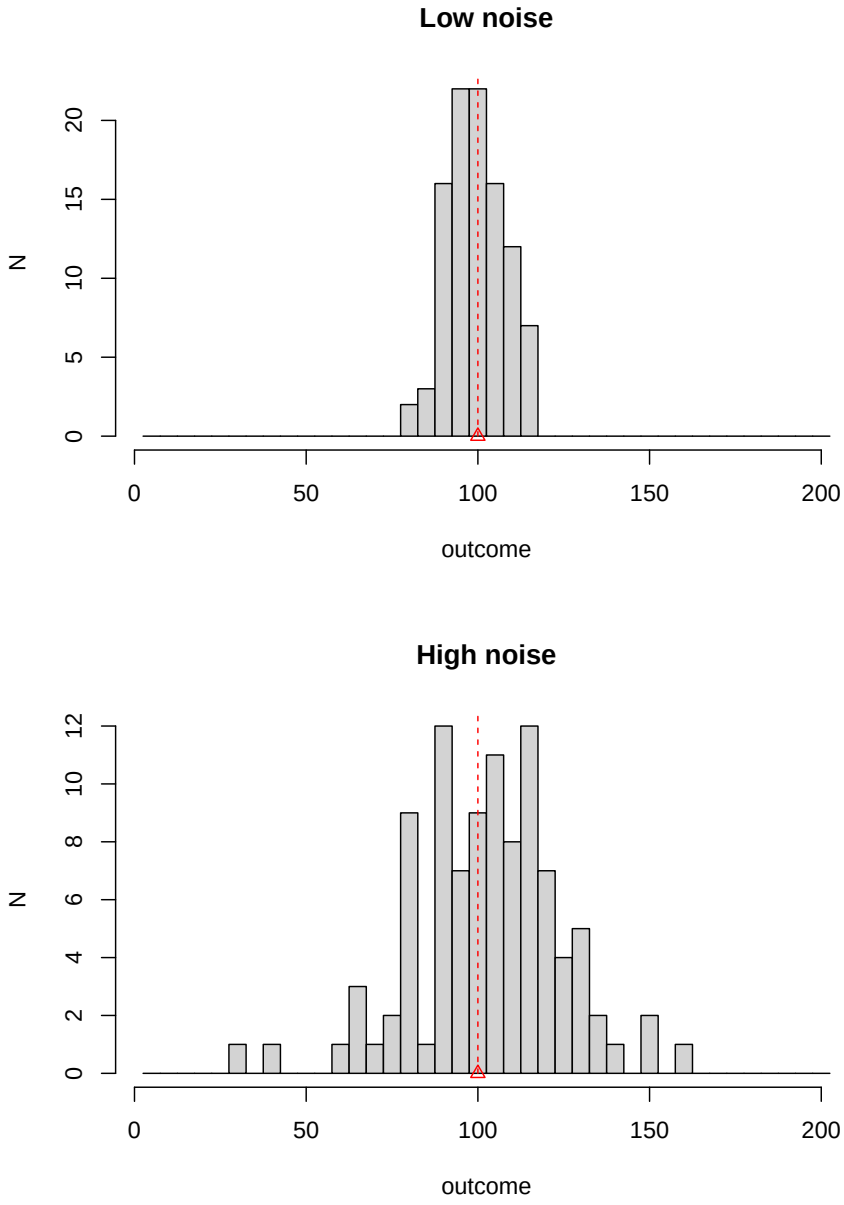


2.22 Dispersión

Otras estadísticas de resumen importantes de las observaciones son las de **dispersión**.

Muchos experimentos pueden compartir su media, pero difieren en cuán **dispersos** son los valores.

La dispersión de las observaciones es una medida del **ruido**.



2.23 Variación de la muestra

La dispersión sobre la media se mide con la varianza muestral

$$s^2 = \frac{1}{N-1} \sum_{j=1..N} (x_j - \bar{x})^2$$

Este número, mide la distancia cuadrada promedio de las **observaciones** al promedio. La razón de $N-1$ se explicará cuando hablemos de inferencia, cuando estudiemos la dispersión de $\bar{x}$, además de la dispersión de las observaciones.

En términos de las frecuencias de las variables **categóricas y ordenadas**

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

s^2 se puede considerar como el **momento de inercia** de las observaciones.

La raíz cuadrada de la varianza de la muestra se denomina **desviación estándar** s.

Ejemplo (Misofonía)

La desviación estándar del ángulo de convexidad es

$s = [\frac{1}{123-1}((7.97 - 10.19894)^2 + (18.23 - 10.19894)^2 + (12.27 - 10.19894)^2 + ...)]^{1/2} = 5.086707$

La convexidad de la mandíbula se desvía de su media en 5.086707.

2.24 Rango intercuartílico (IQR)

La dispersión de los datos también se puede medir con respecto a la mediana usando el **rango intercuartílico**:

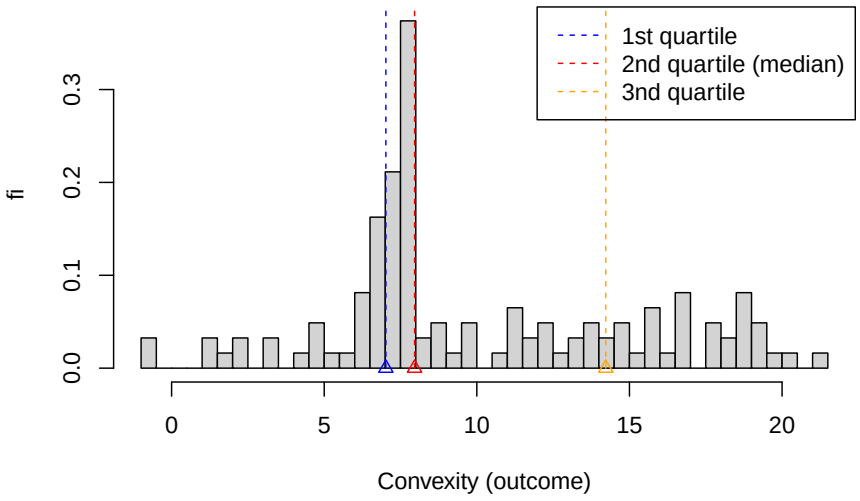
1) Definimos el **primer** cuartil como el valor x_m que hace que la frecuencia acumulada $F_{q_{0.25}}$ sea igual a 0.25 (x donde hemos acumulado una cuarta parte de las observaciones)

$$F_{q_{0.25}} = 0.25$$

1) Definimos el **tercer** cuartil como el valor x_m que hace que la frecuencia acumulada $F_{q_{0.75}}$ sea igual a 0.75 (x donde hemos acumulado tres cuartos de observaciones)

$$F_{q_{0.75}} = 0.75$$

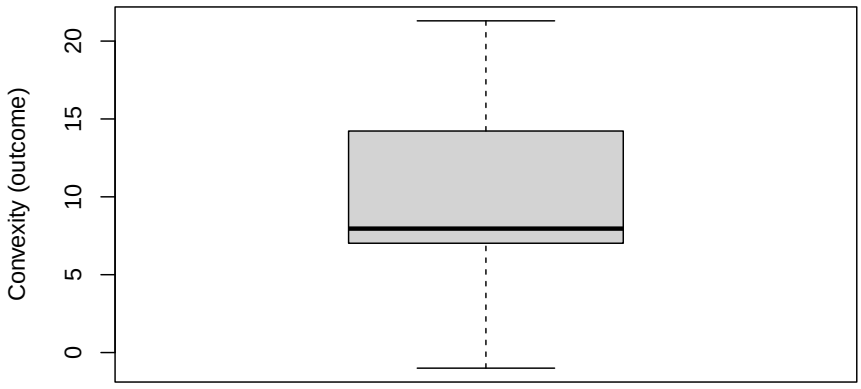
3) El **rango intercuartílico** (IQR) es $IQR = q_{0.75} - q_{0.25}$. Esa es la distancia entre el tercer y el primer cuartil y captura el 50% central de las observaciones



2.25 Diagrama de caja

El rango intercuartílico, la mediana y los 5% y 95% de los datos se pueden visualizar en un **diagrama de caja**.

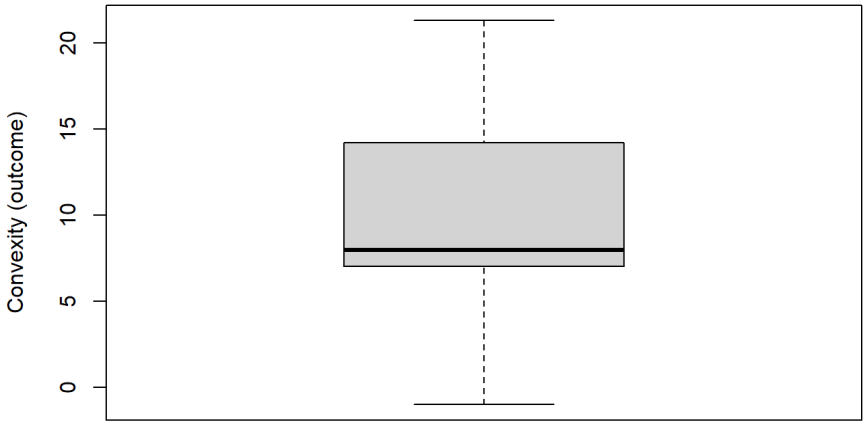
En el diagrama de caja, los valores de los resultados están en el eje y. El IQR es la caja, la mediana es la línea del medio y los bigotes marcan los 5% y 95% de los datos.



2.26 Preguntas

1) En el siguiente diagrama de caja, el primer cuartil y el segundo cuartil de los datos son:

a: $(-1.00, 21.30)$; **b:** $(-1.00, 7.02)$; **c:** $(7.02, 7.96)$; **d:** $(7.02, 14.22)$



2) La principal desventaja de un histograma es que:

a: Depende del tamaño del bin; **b:** No se puede utilizar para variables categóricas; **c:** No se puede usar cuando el tamaño del bin es pequeño; **d:** Se usa solo para frecuencias relativas;

3) Si las frecuencias acumuladas relativas de un experimento aleatorio con resultados $\{1, 2, 3, 4\}$ son: $F(1) = 0.15$, $F(2) = 0.60$, $F(3) = 0.85$, $F(4) = 1$.

Entonces la frecuencia relativa para el resultado 3 es

a: 0.15; **b:** 0.85; **c:** 0.45; **d:** 0.25

4) En una muestra de tamaño 10 de un experimento aleatorio obtuvimos los siguientes datos:

8, 3, 3, 7, 3, 6, 5, 10, 3, 8.

El primer cuartil de los datos es:

a: 3.5; **b:** 4; **c:** 5; **d:** 3

5) Imaginemos que recopilamos datos para dos cantidades que no son mutuamente excluyentes, por ejemplo, el sexo y la nacionalidad de los pasajeros de un vuelo. Si queremos hacer un solo gráfico circular para los datos, ¿cuál de estas afirmaciones es verdadera?

a: Solo podemos hacer un gráfico circular de nacionalidad porque tiene

más de dos resultados posibles; **b:** Podemos hacer un gráfico circular para una variable nueva que marca el sexo **y** la nacionalidad; **c:** Podemos hacer un gráfico circular para la variale sexo o la variable nacionalidad; **d:** Solo podemos elegir si hacemos un gráfico circular para el sexo o un gráfico circular para la nacionalidad.

2.27 Ejercicios

2.27.0.1 Ejercicio 1

Hemos realizado un experimento 8 veces con los siguientes resultados

```
## [1]  3  3 10  2  6 11  5  4
```

Responde las siguientes cuestiones:

- Calcula las frecuencias relativas de cada resultado.
- Calcula las frecuencias acumuladas de cada resultado.
- ¿Cuál es el promedio de las observaciones?
- ¿Cuál es la mediana?
- ¿Cuál es el tercer cuartil?
- ¿Cuál es el primer cuartil?

2.27.0.2 Ejercicio 2

Hemos realizado un experimento 10 veces con los siguientes resultados

```
##  [1] 2.875775 7.883051 4.089769 8.830174 9.404673 0.455565 5.281055 8.924190
##  [9] 5.514350 4.566147
```

Considera 10 bins de tamaño 1: [0,1], (1,2]…(9,10).

Responde las siguientes cuestiones:

- Calcula las frecuencias relativas de cada resultado y dibuja el histograma
- Calcula las frecuencias acumulativas de cada resultado y dibuja la gráfica acumulativa.
- Dibuja un diagrama de caja .