Estadística

Alejandro Caceres

2022-09-25

Contents

Chapter 1

Objetivo

- Este es el curso de introducción a la estadística de la EEBE (UPC).
- Las fechas de exámenes y material de estudio adicional se pueden encontrar en ATENEA

1.1 Lectura recomendada

- Douglas C. Montgomery and George C. Runger. "Applied Statistics and Probability for Engineers" 4th Edition. Wiley 2007.

Chapter 2

Descripción de datos

2.1 Objetivo

- Datos: discretos, continuos
- Resumir datos en tablas y figuras.

2.2 Estadísticas

- Resolver problemas de manera sistemática (ciencia, tecnología e ingeniería)
- ¡Los humanos modernos usamos un **método** general históricamente desarrollado durante miles de años! … y aún en desarrollo.
- Tiene tres componentes principales: observación, lógica y generación de nuevo conocimiento.

2.3 Metodo científico

2.4 Resultado

Observación o *Realización*

- Una **observación** es la adquisición de un número o una característica de un experimento

… 1 0 0 1 0 **1** 0 1 1 … (el número en negrita es una observación en una repetición del experimento)

Resultado

- Un **resultado** es una de las posibles observaciones de un experimento.

1 es un resultado, **0** es el otro resultado

2.5 Tipos de resultado

- **Categórico**: Si el resultado de un experimento solo puede tomar valores discretos (número de piezas de automóvil producidas por hora, número de leucocitos en sangre)
- **Continuo**: Si el resultado de un experimento solo puede tomar valores continuos (estado de carga de la batería, temperatura del motor).

2.6 Experimentos aleatorios

Definición:

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

Ejemplos:

- en el mismo objeto (persona): temperatura, niveles de azúcar.
- sobre objetos diferentes pero de la misma medida: el peso de un animal.
- sobre eventos: número de correos electrónicos recibidos en una hora.

2.7 Frecuencias absolutas

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

Resumimos las observaciones **categóricas** contando cuántas veces vimos un resultado en particular.

Frecuencia absoluta:

$$n_i$$

es el número de veces que observamos el resultado i

2.8 Ejemplo

Experimento aleatorio: extraiga un leucocito de **un** donante y anote su tipo. Repita el experimento $N = 119$ veces.

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

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

- Por ejemplo: $n_1 = 34$ es el número total de células T
- $N = \sum_i n_i = 119$

2.9 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 registran $n_1 = 34$ células T, por lo que la frecuencia relativa nos da la proporción de células T de un total de 119.

2.10 Ejemplo

```
##      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
```

Tenemos

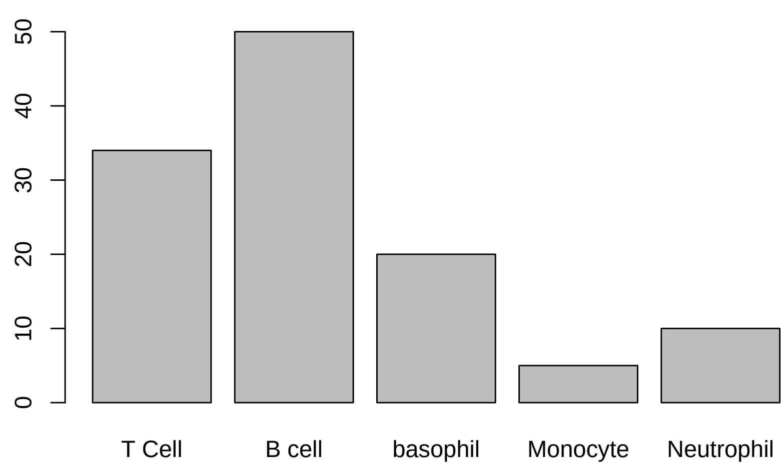
$\sum_{i=1..M} n_i = N$

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

donde M es el número de resultados.

2.11 Diagrama de barras

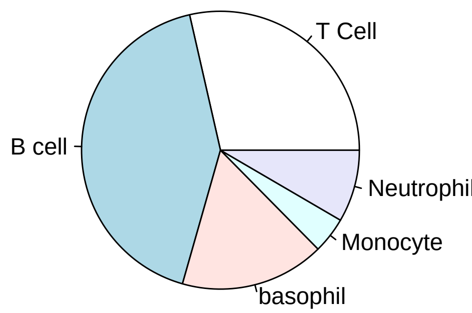
Podemos graficar n_i Vs los resultados, dándonos un gráfico de barras



2.12 Gráfico de sectores

Podemos visualizar las frecuencias relativas con un gráfico de sectores

- Donde el área del círculo representa el 100% de las observaciones (proporción = 1) y las secciones las frecuencias relativas de todos los resultados.



2.13 Variables categóricas y ordenadas

Los tipos de células no están ordenados de manera lógica en relación con los resultados. Sin embargo, a veces las variables **categóricas** se pueden **ordenar**.

Estudio de misofonía:

- 123 pacientes fueron examinados por misofonía: ansiedad/ira producida por ciertos sonidos

- Se clasificaron en 4 grupos diferentes según la gravedad.

2.14 Ejemplo

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
```

y su tabla de frecuencias

```
##   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
```

2.15 Frecuencias acumuladas absolutas y relativas

La gravedad de la misofonía es **categórica** y **ordenada**.

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

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

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

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

2.16 Tabla de frecuencia

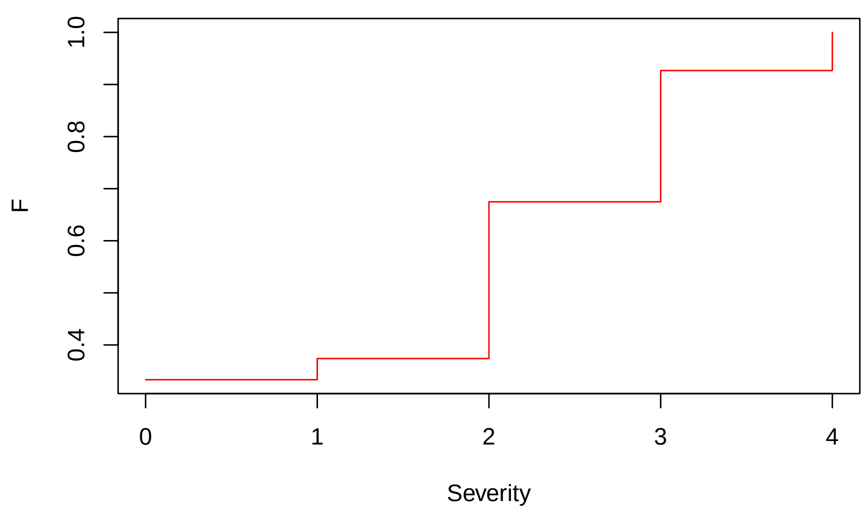
```
##   outcome ni         fi  Ni        Fi
## 0       0 41 0.33333333  41 0.3333333
## 1       1  5 0.04065041  46 0.3739837
```

```
## 2        2 37 0.30081301  83 0.6747967
## 3        3 31 0.25203252 114 0.9268293
## 4        4  9 0.07317073 123 1.0000000
```

- **67%** de los pacientes tenían misofonía hasta la gravedad **2**
- **37%** de los pacientes tienen una gravedad menor o igual a **1**

2.17 Gráfica de frecuencia acumulada

También podemos graficar la frecuencia acumulada Vs los resultados



2.18 Variables continuas

El resultado de un experimento aleatorio también puede dar resultados continuos.

En el estudio de 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 para la convexidad de la mandíbula (grados)

```
##   [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
```

2.19 Contenedores

¡Los resultados continuos no se pueden contar!

Las transformamos en variables categóricas ordenadas

- Cubrimos el rango de las observaciones en intervalos regulares del mismo tamaño (bins)

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

2.20 Crear una variable categórica a partir de una continua

- Asignamos 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]"
```

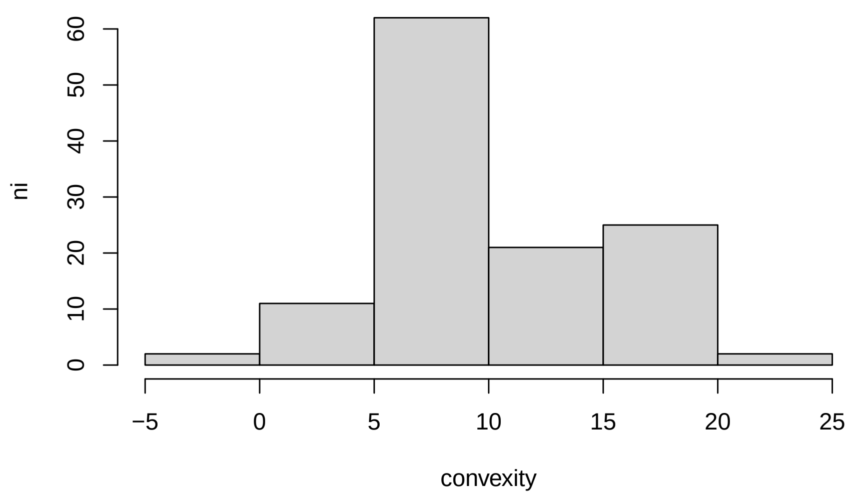
2.21 Tabla de frecuencias para una variable continua

```
##        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
```

2.22 Histograma

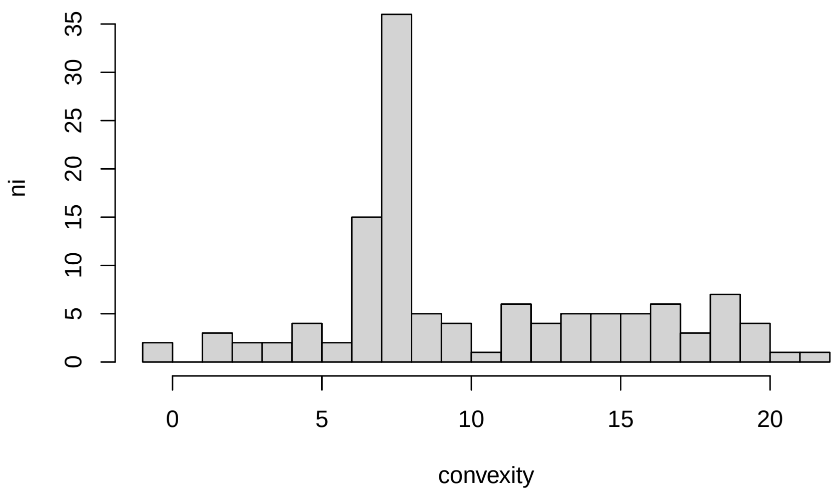
El histograma es la gráfica de n_i o f_i Vs los resultados (bins). El histograma depende del tamaño de los contenedores.

2.23 Tabla de frecuencias para una variable continua



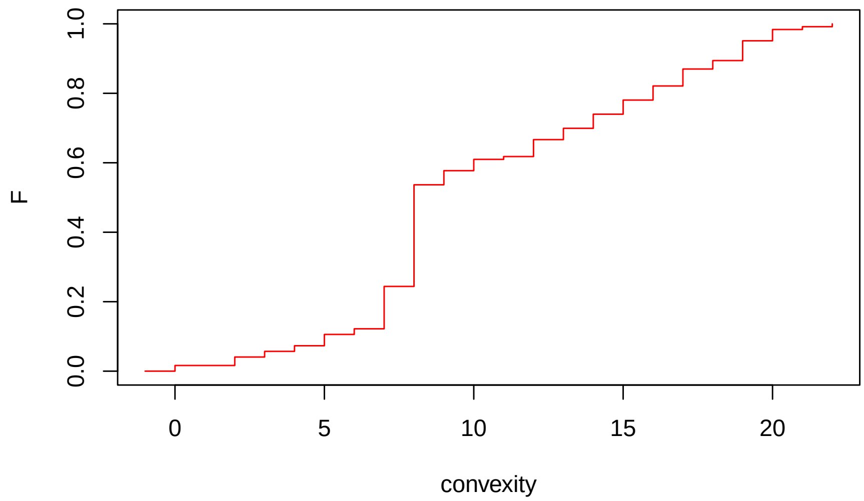
2.24 Histograma

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



2.25 Gráfica de frecuencia acumulada: Variables continuas

También podemos graficar la frecuencia acumulada Vs los resultados



2.26 Resumen estadístico

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

Valores límite:

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

Valor central para los resultados

- El promedio se define como

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

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

2.27 Promedio

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

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

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

2.28 Promedio (ordenado categóricamente)

Para las variables **ordenadas categóricamente**, podemos usar la tabla de frecuencias para calcular el promedio

```
##   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
```

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

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

$= 0 * f_0 + 1 * f_1 + 2 * f_2 + 3 * f_3 + 4 * f_4 = 1,691057$

(note el cambio de N a M en la segunda suma)

2.29 Promedio (ordenado categóricamente)

En términos de los **resultados** de las variables ordenadas categóricas, el **promedio** se puede escribir como

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

de un total de M posibles resultados (número de niveles de gravedad).

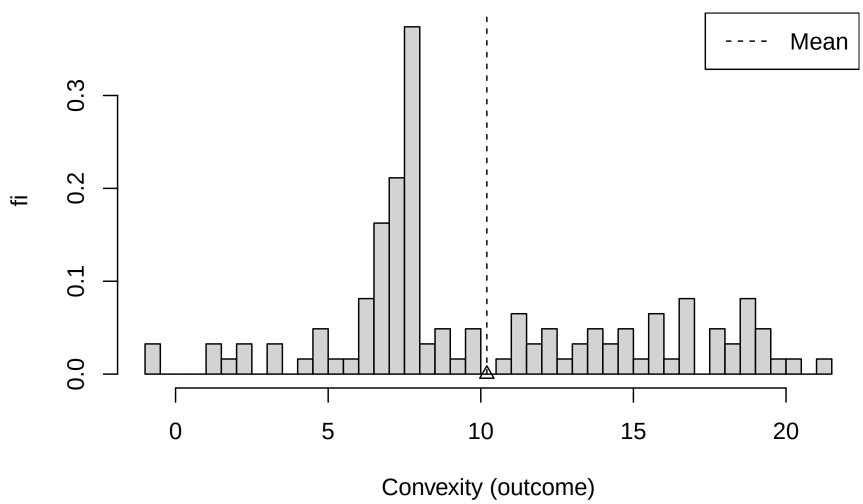
$\bar{x}$ es el **valor central** o centro de gravedad de los resultados. Como si cada resultado tuviera una densidad de masa dada por f_i.

2.30 Promedio

- El promedio no es el resultado de una observación (experimento aleatorio).
- Es el resultado de una serie de observaciones (muestra).
- Describe el número donde se equilibran los valores observados.

Por eso escuchamos, por ejemplo, que un paciente con una infección puede contagiar a una media de 2,5 personas.

2.31 Promedio



2.32 mediana

Otra medida de centralidad es la mediana. La mediana $q_{0.5}$ es el valor x_p

$$mediana(x) = q_{0.5} = x_p$$

debajo del cual encontramos la mitad de las observaciones

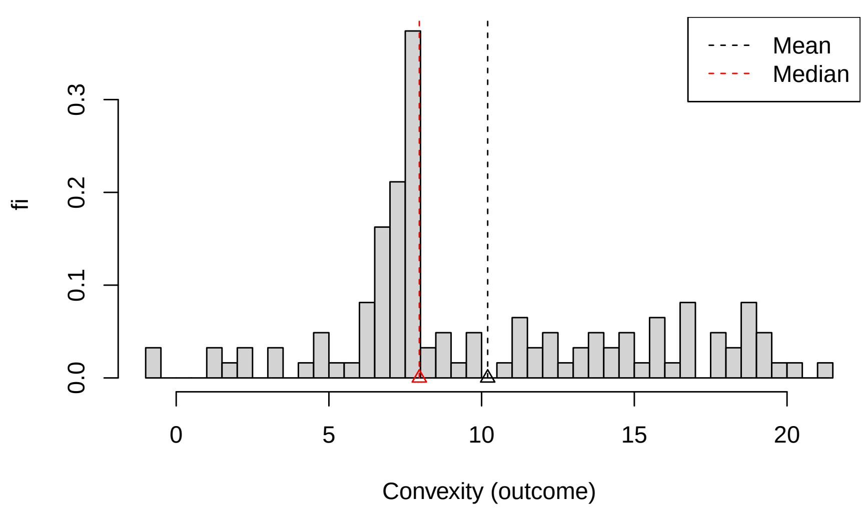
$$\sum_{x \leq x_p} 1 = \frac{N}{2}$$

o en términos de frecuencias, es el valor x_p que hace que la frecuencia acumulada F_p sea igual a 0.5

$$q_{0.5} = \sum_{x \leq x_p} f_x = F_p = 0.5$$

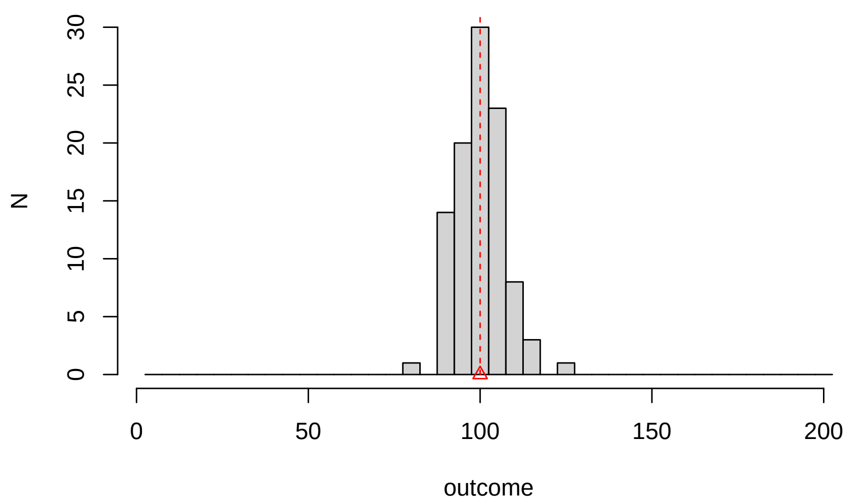
2.33 Mediana Vs Promedio

- Promedio: Centro de masa (compensa valores distantes)
- Mediana: La mitad de la masa

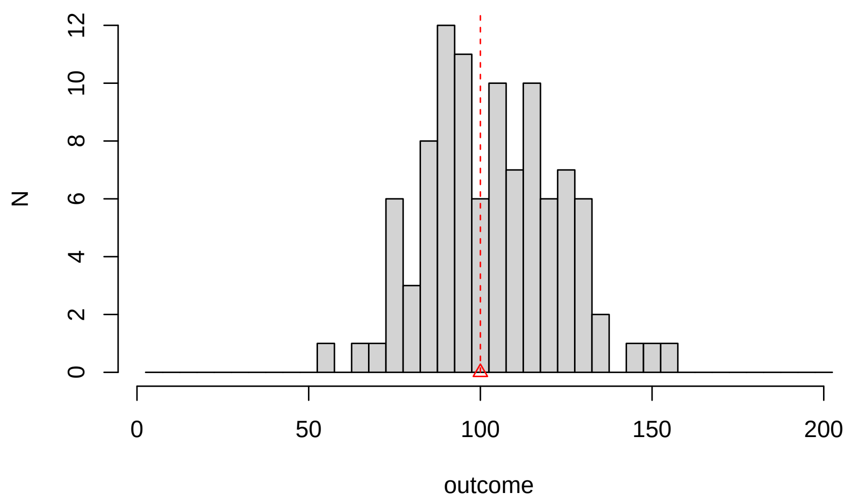


2.34 Dispersión

Una medida importante de los resultados es su **dispersión**. Muchos experimentos pueden compartir su media, pero difieren en la dispersión de los valores.



2.35 Dispersión



2.36 Variación de la muestra

La dispersión con respecto a la media se mide con el

- La varianza muestral:

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

Mide la distancia cuadrada promedio de las **observaciones** al promedio. La razón de $N-1$ se explicará cuando hablemos de inferencia.

2.37 Variación de la muestra

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

$$s^2 = \frac{N}{N-1}\sum_x (x-\bar{x})^2 f_x$$

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

2.38 Desviación Estándar

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

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 $5,086707$.

2.39 RIC

- La dispersión de datos también se puede medir con respecto a la mediana por el **rango intercuartílico**
- Definimos el **primer** cuartil como el valor x_p que hace que la frecuencia acumulada F_p sea igual a $0,25$

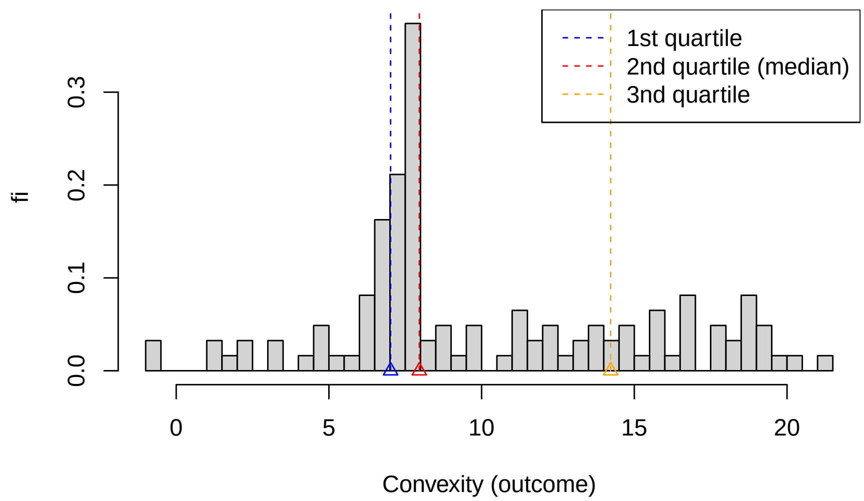
$$q_{0.25} = \sum_{x \leq x_p} f_x = F_p = 0.25$$

- También definimos el **tercer** cuartil como el valor x_p que hace que la frecuencia acumulada F_p sea igual a $0,75$

$$q_{0.75} = \sum_{x \leq x_p} f_x = F_p = 0.75$$

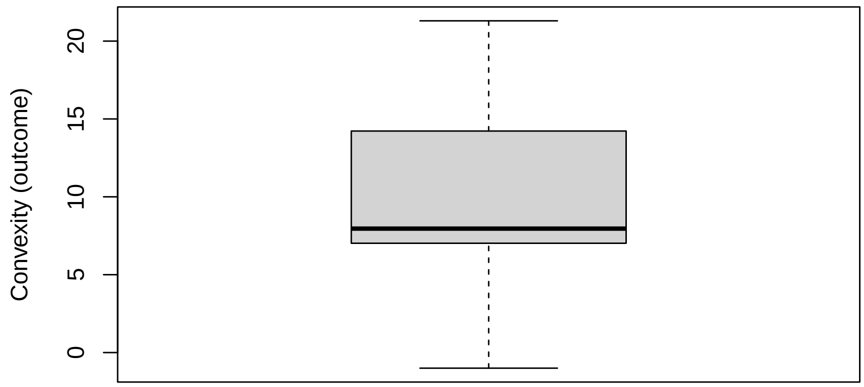
2.40 RIC

La distancia entre el tercer cuartil y el primer cuartil se denomina **rango intercuartílico** (RIC) y captura el 50 % central de las observaciones



2.41 Diagrama de caja

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


Convexity (outcome)
0
5
10
15
20

Chapter 3

Probabilidad

3.1 Objetivo

- Definición de probabilidad
- Álgebra de probabilidad
- Probabilidad conjunta

3.2 Experimentos aleatorios

Observación

- y **observación** es la adquisición de un número o una característica de un experimento

Salir

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

Experimento aleatorio

- Un experimento que da resultados **diferentes** cuando se repite de la misma manera.

3.3 Probabilidad

La **probabilidad** de un resultado es una medida de cuán seguros estamos de observar ese resultado al realizar un experimento aleatorio.

- 0: Estamos seguros de que la observación **no** ocurrirá.
- 1: Estamos seguros de que la observación sucederá.

3.4 Ejemplo

- Considere las siguientes observaciones de un experimento aleatorio:

1 5 1 2 2 1 2 2

- ¿Qué tan seguro estamos de obtener 2 en la siguiente observación?

3.5 Ejemplo

La tabla de frecuencias es

```
##   outcome ni    fi
## 1       1  3 0.375
## 2       2  4 0.500
## 3       5  1 0.125
```

La **frecuencia relativa** f_i

- es un número entre 0 y 1.
- mide la proporción del total de observaciones que observamos un resultado particular.
- parece una medida de probabilidad razonable.

Como $f_2 = 0.5$ entonces estaríamos 50 seguros de obtener 2 en la siguiente repetición del experimento.

3.6 Frecuencia relativa

¿f_i es una buena medida de certeza?

Digamos que repetimos el experimento 12 veces más:

1 5 1 2 2 1 2 2 **3 1 1 3 3 1 6 3 5 6 4 4**

La tabla de frecuencias es ahora

```
##   outcome ni  fi
## 1       1  6 0.3
## 2       2  4 0.2
## 3       3  4 0.2
## 4       4  2 0.1
## 5       5  2 0.1
## 6       6  2 0.1
```

Aparecieron nuevos resultados y f_2 ahora es 0.2, ahora estamos un 20% seguros de obtener 2 en el próximo experimento… la probabilidad no debería depender de N

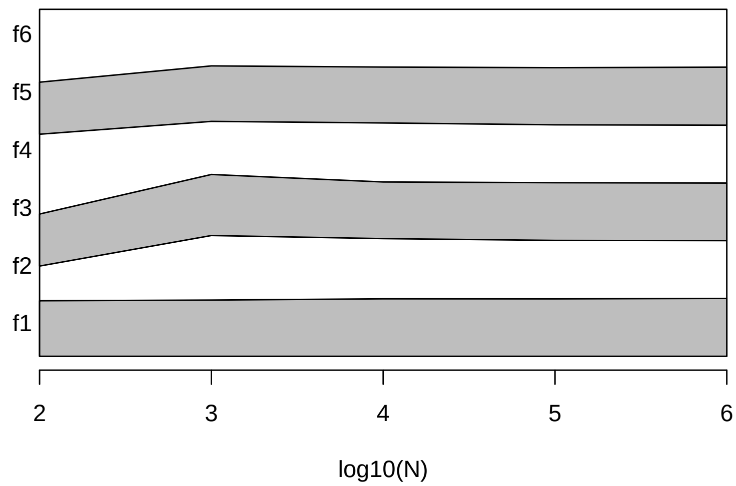
3.7 En el infinito

Digamos que repetimos el experimento 1000 veces:

```
##   outcome  ni    fi
## 1       1 178 0.178
## 2       2 149 0.149
## 3       3 156 0.156
## 4       4 187 0.187
## 5       5 182 0.182
## 6       6 148 0.148
```

Encontramos que f_i está convergiendo a un valor constante

$$lim_{N\to\infty} f_i = P_i$$



3.8 Probabilidad frecuentista

Llamamos **Probabilidad** P_i al límite cuando $N \to \infty$ de la **frecuencia relativa** de observar el resultado i en un experimento aleatorio.

Defendida por Venn (1876)

La interpretación frecuentista de probabilidades se deriva de datos/experiencia (empírica).

- No observamos P_i, observamos f_i
- Cuando **estimamos** P_i con f_i (normalmente cuando N es grande), escribimos:

$$\hat{P}_i = f_i$$

3.9 Probabilidad clásica

Cada vez que un experimento aleatorio tiene M resultados posibles que son todos **igualmente probables**, la probabilidad de cada resultado es $\frac{1}{M}$.

Defendida por Laplace (1814).

Dado que cada resultado es **igualmente probable**, declaramos una completa ignorancia y lo mejor que podemos hacer es distribuir equitativamente la misma probabilidad para cada resultado.

¿Y si te dijera que nuestro experimento fue tirar un dado? entonces

$P_2 = 1/6 = 0.166666$.

$$P_i = lim_{N\to\infty} \frac{n_i}{N} = \frac{1}{M}$$

3.10 Probabilidades clásicas y frecuentistas

3.11 Probabilidad

La probabilidad es un número entre 0 y 1 que se asigna a cada miembro E de una colección de **eventos** de un **espacio muestral** (S) de un experimento aleatorio.

$$P(E) \in (0, 1)$$

donde $E \in S$

3.12 Espacio muestral

Empezamos razonando cuáles son todos los valores posibles (resultados) que podría dar un experimento aleatorio.

Tenga en cuenta que no tenemos que observarlos en un experimento en particular: estamos usando **razón/lógica** y no observación.

Definición:

- El conjunto de todos los resultados posibles de un experimento aleatorio se denomina **espacio muestral** del experimento

- El espacio muestral se denota como S.

3.13 Ejemplos de espacios muestrales

- temperatura 35 y 42 grados centígrados
- niveles de azúcar: 70-80mg/dL
- el tamaño de un tornillo de una línea de producción: 70 mm-72 mm
- número de correos electrónicos recibidos en una hora: 0-100
- un lanzamiento de dados: 1, 2, 3, 4, 5, 6

3.14 Espacios muestrales discretos y continuos

- Un espacio muestral es discreto si consiste en un conjunto de resultados finito o infinito numerable.
- Un espacio muestral es continuo si contiene un intervalo (ya sea de longitud finita o infinita) de numeros reales.

3.15 Evento

Definición:

Un **evento** es un **subconjunto** del espacio muestral de un experimento aleatorio. Es una **colección** de resultados.

Ejemplos de eventos:

- El evento de una temperatura saludable: temperatura 37-38 grados centígrados
- El evento de producir un tornillo con un tamaño: de 71,5 mm
- El evento de recibir más de 4 correos electrónicos en una hora.
- El evento de obtener un número menor de 3 en el lanzamiento de un dado

Un evento se refiere a un posible conjunto de **resultados**.

3.16 Operaciones de eventos

Para dos eventos A y B, podemos construir los siguientes eventos derivados:

- Complemento A': el evento de **no** A
- Unión $A \cup B$: el evento de A **o** B
- Intersección $A \cap B$: el evento de A **y** B

3.17 Ejemplo de operaciones de eventos

Tomar

- Evento $A : \{1, 2, 3\}$ un número menor o igual a tres en el lanzamiento de un dado
- Evento $B : \{2, 4, 6\}$ un número par en el lanzamiento de un dado

Nuevos eventos:

- No menos de tres: $A' : \{4, 5, 6\}$
- Menor o igual a tres **o** par: $A \cup B : \{1, 2, 3, 4, 6\}$
- Menor o igual a tres **y** par $A \cap B : \{2\}$

3.18 Resultados

Los resultados son eventos que son **mutuamente excluyentes**

Definición:

Dos eventos denotados como E_1 y E_2, tales que

$$E_1 \cap E_2 = \emptyset$$

No pueden ocurrir al mismo tiempo.

Ejemplo:

- El resultado de obtener 1 **y** el resultado de obtener 5 en el lanzamiento de un dado son mutuamente excluyentes:

- El evento de obtener 1 y 5 está vacío:

$$\{1\} \cap \{5\} = \emptyset$$

3.19 Definición de probabilidad

Una probabilidad es un número que se asigna a cada evento posible (E) de un espacio muestral (S) de un experimento aleatorio que cumple las siguientes propiedades:

- $P(S) = 1$
- $0 \leq P(E) \leq 1$
- cuando $E_1 \cap E_2 = \emptyset$

$$P(E_1 \cup E_2) = P(E_1) + P(E_2)$$

Propuesto por Kolmogorov (1933)

3.20 Propiedades de probabilidad

Kolmogorov dice que podemos construir una tabla de probabilidad (al igual que la tabla de frecuencia relativa)

resultado	Probabilidad
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6
$P(1 \cup 2 \cup ... \cup 6)$	1

Como $\{1, 2, 3, 4, 5, 6\}$ son mutuamente excluyentes, entonces

$$P(S) = P(1 \cup 2 \cup ... \cup 6) = P(1) + P(2) + ... + P(n) = 1$$

3.21 Regla de adición

Cuando A y B no son mutuamente excluyentes, entonces:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Donde $P(A)$ y $P(B)$ se denominan **probabilidades marginales**

3.22 Ejemplo de regla de adición

Tomar

- Evento $A : \{1, 2, 3\}$ un número menor o igual a tres en el lanzamiento de un dado
- Evento $B : \{2, 4, 6\}$ un número par en el lanzamiento de un dado

después:

- $P(A) : P(1) + P(2) + P(3) = 3/6$
- $P(G) : P(2) + P(4) + P(6) = 3/6$
- $P(A \cap B) : P(2) = 1/6$

$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 3/6 + 3/6 - 1/6 = 5/6$

Nota: $P(2)$ aparece en $P(A)$ y $P(B)$ por eso lo restamos con la intersección

---

3.23 Diagrama de Venn

Tenga en cuenta que siempre se puede descomponer el espacio muestral en conjuntos **mutuamente excluyentes** que involucran las intersecciones:

$S = \{A \cap B, A \cap B', A' \cap B, A' \cap B'\}$

Marginales:

- $P(A) = P(A \cap B') + P(A \cap B) = 2/6 + 1/6 = 3/6$
- $P(B) = P(A' \cap B) + P(A \cap B) = 2/6 + 1/6 = 3/6$

3.24 Tabla de probabilidades

Veamos la tabla de probabilidades.

resultado	Probabilidad
$A \cap B$	$P(A \cap B)$
$A \cap B'$	$P(A \cap B')$
$A' \cap B$	$P(A' \cap B)$
$A' \cap B'$	$P(A' \cap B')$
suma	1

3.25 Ejemplo de tabla de probabilidades

También escribimos $A \cap B$ como (A, B) y lo llamamos la **probabilidad conjunta** de A y B

En nuestro ejemplo:

resultado	Probabilidad
(A, B)	$P(A, B) = 1/6$
(A, B')	$P(A, B') = 2/6$
(A', B)	$P(A', B) = 2/6$
(A', B')	$P(A', B') = 1/6$
suma	1

Nota: cada resultado tiene *dos* valores (uno para la característica del tipo A y otro para el tipo B)

3.26 Tabla de contingencia

Podemos organizar la probabilidad de **resultados conjuntos** en una **tabla de contingencia**

	B	B'	suma
A	$P(A, B)$	$P(A, B')$	$P(A)$

	B	B'	suma
A'	$P(A', B)$	$P(A', B')$	$P(A')$
suma	$P(B)$	$P(B')$	1

marginales:

- $P(A) = P(A, B') + P(A, B)$
- $P(B) = P(A', B) + P(A, B)$

3.27 Ejemplo de tabla de contingencia

- Evento $A : \{1, 2, 3\}$ un número menor o igual a tres en el lanzamiento de un dado
- Evento $B : \{2, 4, 6\}$ un número par en el lanzamiento de un dado

	B	B'	suma
A	1/6	2/6	3/6
A'	2/6	1/6	3/6
suma	3/6	3/6	1

Tres formas de la **regla de la suma**:

$P(A \cup B)$

$$= P(A) + P(B) - P(A \cap B)$$

$$= P(A \cap B) + P(A \cap B') + P(A' \cap B)$$

$$= 1 - P(A' \cap B')$$

3.28 Estudio de misofonía

En el estudio de misofonía, se evaluó a los pacientes según la gravedad de su misofonía **y** si estaban deprimidos.

El resultado de un experimento aleatorio es medir la gravedad de la misofonía **y** el estado de depresión de un paciente. La repetición del experimento aleatorio consistía en realizar las mismas dos mediciones en otro paciente.

```
##    Misofonia.dic depresion.dic
## 1              4             1
## 2              2             0
## 3              0             0
## 4              3             0
## 5              0             0
## 6              0             0
## 7              2             0
## 8              3             0
## 9              0             1
## 10             3             0
## 11             0             0
## 12             2             0
## 13             2             1
## 14             0             0
## 15             2             0
## 16             0             0
## 17             0             0
## 18             3             0
## 19             3             0
## 20             0             0
## 21             3             0
## 22             3             0
## 23             2             0
## 24             0             0
## 25             0             0
## 26             0             0
## 27             4             1
## 28             2             0
## 29             2             0
## 30             0             0
## 31             2             0
## 32             0             0
## 33             0             0
## 34             0             0
## 35             3             0
## 36             0             0
## 37             2             0
## 38             3             1
## 39             2             0
## 40             2             0
## 41             0             0
## 42             2             0
## 43             3             0
## 44             0             0
## 45             0             0
```

```
## 46                2              0
## 47                2              0
## 48                3              0
## 49                3              0
## 50                0              0
## 51                0              0
## 52                4              1
## 53                3              0
## 54                3              1
## 55                2              1
## 56                0              1
## 57                2              0
## 58                0              0
## 59                0              0
## 60                0              0
## 61                2              0
## 62                2              0
## 63                0              0
## 64                0              0
## 65                2              0
## 66                3              1
## 67                0              0
## 68                1              0
## 69                3              0
## 70                2              0
## 71                4              1
## 72                3              0
## 73                2              1
## 74                3              0
## 75                0              1
## 76                2              0
## 77                3              0
## 78                2              0
## 79                4              1
## 80                1              0
## 81                2              0
## 82                0              0
## 83                2              0
## 84                0              0
## 85                2              0
## 86                0              1
## 87                2              0
## 88                2              0
## 89                4              1
## 90                3              0
## 91                0              1
```

```
## 92             3              0
## 93             0              0
## 94             0              0
## 95             0              0
## 96             2              0
## 97             2              0
## 98             1              0
## 99             3              0
## 100            0              0
## 101            0              0
## 102            3              1
## 103            2              0
## 104            1              0
## 105            3              0
## 106            0              0
## 107            4              1
## 108            4              1
## 109            2              0
## 110            3              0
## 111            3              0
## 112            3              1
## 113            0              0
## 114            3              0
## 115            2              0
## 116            1              0
## 117            2              0
## 118            3              1
## 119            3              0
## 120            4              1
## 121            2              0
## 122            3              0
## 123            2              0
```

3.29 Tabla de contingencia para frecuencias

- Para el número de observaciones $n_{i,j}$ de cada resultado (x_i, y_i), misofonía: $x \in \{0, 1, 2, 3, 4\}$ y depresión $y\ en\{0, 1\}$ (no:0, sí:1)

```
##
##                Depression:0 Depression:1
##   Misophonia:4            0            9
##   Misophonia:3           25            6
```

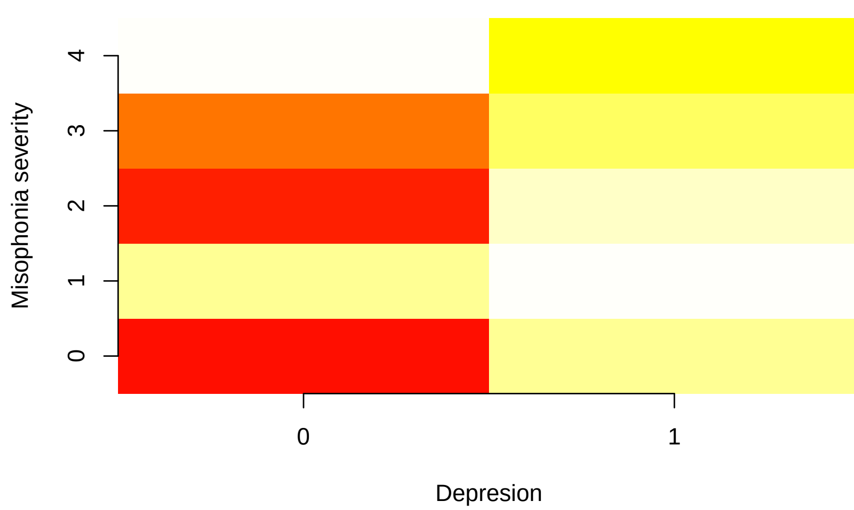
```
##    Misophonia:2           34             3
##    Misophonia:1            5             0
##    Misophonia:0           36             5
```

- para las frecuencias relativas $f_{i,j}$

```
##
##                  Depression:0 Depression:1
##    Misophonia:4   0.00000000   0.07317073
##    Misophonia:3   0.20325203   0.04878049
##    Misophonia:2   0.27642276   0.02439024
##    Misophonia:1   0.04065041   0.00000000
##    Misophonia:0   0.29268293   0.04065041
```

3.30 Mapa de calor

La tabla de contingencia se puede trazar como un **mapa de calor**



3.31 Variables continuas

En el estudio de misofonía también se midió la protrusión mandibular como posible factor cefalométrico de la enfermedad.

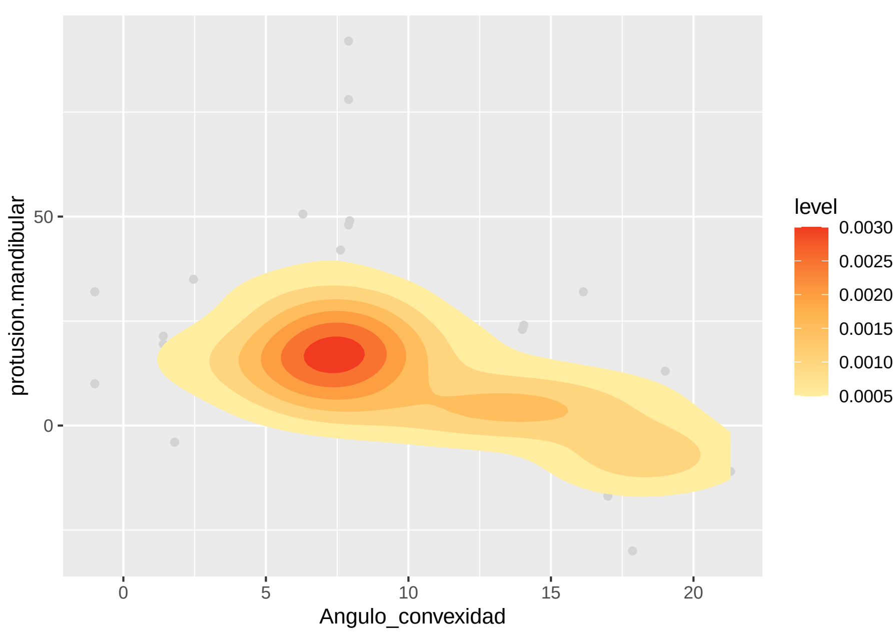
```
##    Angulo_convexidad protusion.mandibular
## 1               7.97                13.00
## 2              18.23                -5.00
## 3              12.27                11.50
## 4               7.81                16.80
## 5               9.81                33.00
## 6              13.50                 2.00
## 7              19.30                -3.90
## 8               7.70                16.80
## 9              12.30                 8.00
## 10              7.90                28.80
## 11             12.60                 3.00
## 12             19.00                -7.90
## 13              7.27                28.30
## 14             14.00                 4.00
## 15              5.40                22.20
## 16              8.00                 0.00
## 17             11.20                15.00
## 18              7.75                17.00
## 19              7.94                49.00
## 20             16.69                 5.00
## 21              7.62                42.00
## 22              7.02                28.00
## 23              7.00                 9.40
## 24             19.20               -13.20
## 25              7.96                23.00
## 26             14.70                 2.30
## 27              7.24                25.00
## 28              7.80                 4.90
## 29              7.90                92.00
## 30              4.70                 6.00
## 31              4.40                17.00
## 32             14.00                 3.30
## 33             14.40                10.30
## 34             16.00                 6.30
## 35              1.40                19.50
## 36              9.76                22.00
## 37              7.90                 5.00
## 38              7.90                78.00
## 39              7.40                 9.30
## 40              6.30                50.60
```

```
## 41             7.76           18.00
## 42             7.30           18.00
## 43             7.00           10.00
## 44            11.23            4.00
## 45            16.00           13.30
## 46             7.90           48.00
## 47             7.29           23.50
## 48             6.91           37.60
## 49             7.10           15.00
## 50            13.40            5.10
## 51            11.60           -2.20
## 52            -1.00           32.00
## 53             6.00           25.00
## 54             7.82           24.00
## 55             4.80           33.60
## 56            11.00            3.30
## 57             9.00           31.50
## 58            11.50           12.80
## 59            16.00            3.00
## 60            15.00            6.00
## 61             1.40           21.40
## 62            16.80          -10.00
## 63             7.70           19.00
## 64            16.14           32.00
## 65             7.12           15.00
## 66            -1.00           10.00
## 67            17.00          -16.90
## 68             9.26            2.00
## 69            18.70          -10.10
## 70             3.40           12.20
## 71            21.30          -11.00
## 72             7.50            5.20
## 73             6.03           16.00
## 74             7.50            5.80
## 75            19.00            5.20
## 76            19.01           13.00
## 77             8.10           13.60
## 78             7.80           16.10
## 79             6.10           33.20
## 80            15.26            4.00
## 81             7.95           12.00
## 82            18.00           -1.50
## 83             4.60           18.30
## 84            15.00            3.00
## 85             7.50           15.80
## 86             8.00           27.10
```

```
## 87              16.80          -10.00
## 88               8.54           25.00
## 89               7.00           27.10
## 90              18.30           -8.00
## 91               7.80           12.00
## 92              16.00           -8.00
## 93              14.00           23.00
## 94              12.30            5.00
## 95              11.40            1.00
## 96               8.50           18.90
## 97               7.00           15.00
## 98               7.96           22.00
## 99              17.60           -3.50
## 100             10.00           20.00
## 101              3.50           12.20
## 102              6.70           14.70
## 103             17.00           -5.00
## 104             20.26           -4.15
## 105              6.64           11.00
## 106              1.80           -4.00
## 107              7.02           25.00
## 108              2.46           35.00
## 109             19.00           -5.00
## 110             17.86          -30.00
## 111              6.10           12.20
## 112              6.64           19.00
## 113             12.00            1.60
## 114              6.60           20.00
## 115              8.70           17.10
## 116             14.05           24.00
## 117              7.20            7.10
## 118             19.70          -11.00
## 119              7.70           21.30
## 120              6.02            5.00
## 121              2.50           12.90
## 122             19.00            5.90
## 123              6.80            5.80
```

3.32 Variables continuas

En el estudio de misofonía también se midió la protrusión mandibular como posible factor cefalométrico de la enfermedad.

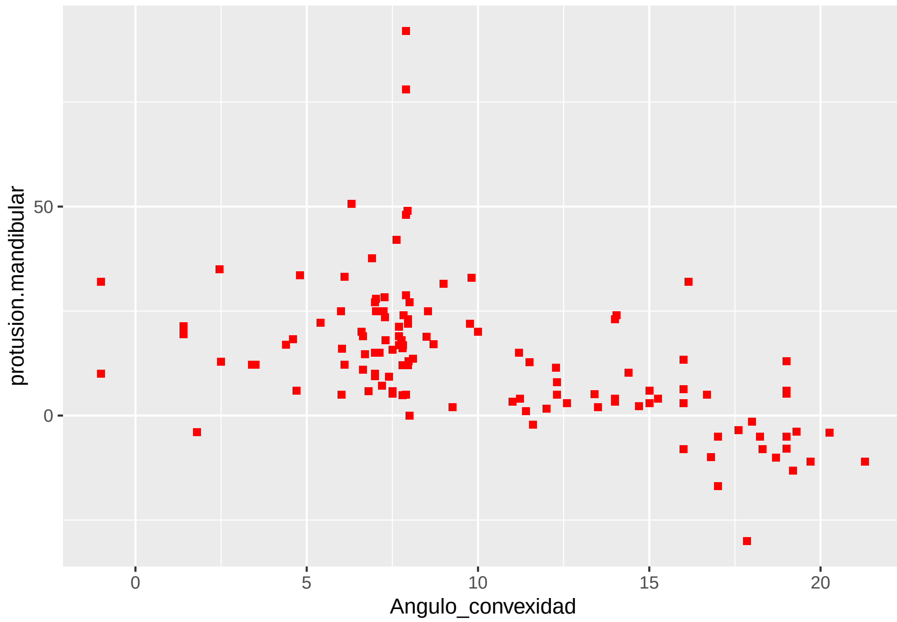


3.33 Gráfico de dispersión

- El **histograma** depende del tamaño del contenedor (píxel).

- Si el píxel es lo suficientemente pequeño como para contener una sola observación, el mapa de calor da como resultado un **diagrama de dispersión**

El diagrama de dispersión es la ilustración de una "tabla de contingencia" para variables continuas cuando el contenedor (píxel) es lo suficientemente pequeño como para contener una sola observación (que consta de un par de valores).



protusion.mandibular
50
0
0
5
10
15
20
Angulo_convexidad

Chapter 4

Probabilidad condicional

4.1 Objetivo

- Probabilidad condicional
- Independencia
- Teorema de Bayes

4.2 Probabilidad conjunta

La probabilidad conjunta de dos eventos A y B es

$$P(A,B) = P(A \cap B)$$

Imaginemos un experimento aleatorio que mide dos tipos diferentes de resultados.

- altura y peso de un individuo: (h, w)
- hora y lugar de una carga eléctrica: (p, t)
- una tirada de dos dados: (n_1,n_2)
- cruzar dos semáforos en verde: $(\bar{R}_1, \bar{R}_2)$

En muchos casos, nos interesa saber si los valores de un resultado **condicionan** los valores del otro.

4.3 Diagnósticos

Consideremos una **herramienta de diagnóstico**

Queremos encontrar el estado de un sistema (s):

- inadecuado (sí)
- adecuado (no)

con una prueba (t):

- positivo
- negativo

Probamos una batería para saber cuánto tiempo puede vivir. Tensamos un cable para saber si resiste llevar cierta carga. Realizamos una PCR para ver si alguien está infectado.

4.4 Prueba de diagnóstico

Consideremos diagnosticar una infección con una nueva prueba.

Estado de infección:

- si (infectado)
- no (no infectado)

Prueba:

- positivo
- negativo

4.5 Observaciones

Cada individuo es un experimento aleatorio con dos medidas: (Infección, Prueba)

Asunto	Infección	Prueba
s_1	si	positivo
s_2	no	negativo
s_3	si	positivo
...	...	...

Asunto	Infección	Prueba
s_i	no	positivo*
...	...	...
...	...	...
s_n	si	negativo*

4.6 Tablas de contingencia

- Por el número de observaciones de cada resultado

	Infección: sí	Infección: no	suma
Test: positivo	18	12	30
Test: negativo	30	300	330
suma	48	312	360

- Para las frecuencias relativas, si $N >> 0$ tomaremos $f_{i,j} = \hat{P}(x_i, y_j)$

	Infección: sí	Infección: no	suma
Test: positivo	0.05	0.0333	0.0833
Test: negativo	0.0833	0.833	0.9166
suma	0.133	0.866	1

4.7 La probabilidad condicional

Pensemos primero en términos de aquellos que están **infectados**

Dentro de los que están infectados (**sí**), ¿cuál es la probabilidad de dar positivo?

- Sensibilidad (tasa de verdaderos positivos)

$$\hat{P}(positivo|s) = \frac{n_{positivo,s}}{n_s}$$

$$= \frac{\frac{n_{positivo,s}}{N}}{\frac{n_s}{N}} = \frac{f_{positivo,s}}{f_s}$$

Por lo tanto, en el límite, esperamos tener una probabilidad del tipo

$$P(positivo|s) = \frac{P(positivo, s)}{P(s)} = \frac{P(positivo \cap s)}{P(s)}$$

4.8 La probabilidad condicional

Definición: La probabilidad condicional de un evento B dado un evento A, denotado como $P(A|B)$, es

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- se puede probar que la probabilidad condicional satisface los axiomas de probabilidad.
- la probabilidad condicional es la probabilidad bajo el espacio muestral dado por B: S_B.

4.9 Tabla de contingencia condicional

	Infección: Sí	Infección: No
Test: positivo	P(positivo | sí)	P(positivo | no)
Test: negativo	P(negativo | sí)	P(negativo | no)
suma	1	1

- Tasa de verdaderos positivos (Sensibilidad): La probabilidad de dar positivo **si** se tiene la enfermedad $P(positivo|s)$
- Tasa de verdaderos negativos (Especificidad): La probabilidad de dar negativo **si** no se tiene la enfermedad $P(negativo|no)$
- Tasa de falsos positivos: La probabilidad de dar positivo **si** no se tiene la enfermedad $P(positivo|no)$
- Tasa de falsos negativos: la probabilidad de dar negativo **si** se tiene la enfermedad $P(negativo|s)$

4.10 Ejemplo de tabla de contingencia condicional

Tomando las frecuencias como estimaciones de las probabilidades, entonces

	Infección: Sí	Infección: No
Test: positivo	18/48 = 0.375	12/312 = 0.038
Test: negativo	30/48 = 0.625	300/312 =0.962
suma	1	1

Nuestra herramienta de diagnóstico tiene baja sensibilidad (0.375) pero alta especificidad (0.962).

4.11 Regla de multiplicación

Ahora imaginemos la situación real, donde queremos obtener la probabilidad **conjunta** de la probabilidad **condicional**

- Se (realizaron) PCR para coronavirus [https://www.nejm.org/doi/full/10.1056/NEJMp2015897] en personas en el hospital que estamos seguros de estar infectadas. Este test tiene una sensibilidad del 70%. También se ha probado en el laboratorio en condiciones sin infección con una especificidad del 96 %.
- Un estudio de prevalencia en España mostró que $P(s) = 0.05$, $P(no) = 0.95$ antes del verano.

Con estos datos, ¿cuál era la probabilidad de que una persona seleccionada al azar de la población diera positivo **y** estuviera infectada: $P(s \cap positivo) = P(s, positivo)$?

4.12 Rendimiento de diagnóstico

Para estudiar el rendimiento de una nueva prueba diagnóstica:

- selecciona muestras que son inadecuadas (enfermedad: **sí**) y aplica la prueba, tratando de encontrar su sensibilidad: $P(positivo|s)$ (0.70 para PCR)

- selecciona muestras que son adecuadas (enfermedad: **no**) y aplica la prueba, tratando de encontrar su especificidad: $P(negativo|no)$ (0.96 para PCR)

	Infección: Sí	Infección: No
Test: positivo	P(positivo|sí)=0.7	P(positivo|no)=0.06
Test: negativo	P(negativo|sí)=0.3	P(negativo|no)=0.94
suma	1	1

De esta matriz, ¿podemos obtener $P(s, positivo)$?

4.13 Regla de multiplicación

¿Cómo se recupera la probabilidad conjunta de la probabilidad condicional?

Para dos eventos A y B tenemos la regla de la multiplicación

$$P(A, B) = P(A|B)P(B)$$

que se sigue de la definición de probabilidad condicional.

4.14 Tabla de contingencia en términos de probabilidades condicionales

	Infección: Sí	Infección: No	suma
Test: positivo	P(positivo | sí)P(sí)	P(positivo | no)P(no)	P(positivo)
Test: negativo	P(negativo | sí)P(sí)	P(negativo | no) P(no)	P(negativo)
suma	P(sí)	P(no)	1

Por ejemplo, la probabilidad de dar *positivo* y estar infectado *s*:

- $P(positivo, s) = P(positivo \cap s) = P(positivo|s)P(s)$

4.15 Árbol condicional

4.16 Tabla de contingencia en términos de probabilidades condicionales

	Infección: sí	Infección: no	suma
Test: positivo	0.035	0.057	0.092
Test: negativo	0.015	0.893	0.908
suma	0.05	0.95	1

- $P(positivo, si) = 0.035$

Pero también encontramos la probabilidad marginal de ser **positivo**:

- $P(positivo) = 0.092$

4.17 Regla de probabilidad total

	Infección: Sí	Infección: No	suma
Test: positivo	P(positivo | sí)P(sí)	P(positivo | no)P(no)	P(positivo)
Test: negativo	P(negativo | sí)P(sí)	P(negativo | no) P(no)	P(negativo)
suma	P(sí)	P(no)	1

Cuando escribimos las marginales desconocidas en términos de sus probabilidades condicionales, lo llamamos **regla de probabilidad total**

- $P(positivo) = P(positivo|s)P(s) + P(positivo|no)P(no)$
- $P(negativo) = P(negativo|s)P(s) + P(negativo|no)P(no)$

4.18 Árbol condicional

Regla de probabilidad total para la marginal de B: ¿De cuántas maneras puedo obtener el resultado B?

$P(B) = P(B|A)P(A) + P(B|A')P(A')$

4.19 Encontrar probabilidades inversas

De la tabla de contingencia condicional

	Infección: Sí	Infección: No
Test: positivo	P(positivo | sí)	P(positivo | no)
Test: negativo	P(negativo | sí)	P(negativo | no)
suma	1	1

¿Cómo podemos calcular la probabilidad de estar infectado si la prueba da positivo: $P(s|positivo)$?

4.20 Recuperar probabilidades conjuntas

1. Recuperamos la tabla de contingencia para probabilidades conjuntas

	Infección: Sí	Infección: No	suma
Test: positivo	P(positivo | sí)P(sí)	P(positivo | no)P(no)	P(positivo)
Test: negativo	P(negativo | sí)P(sí)	P(negativo | no) P(no)	P(negativo)
suma	P(sí)	P(no)	1

4.21 Condicionales inversas

2. Calculamos las probabilidades condicionales para la prueba:

$$P(infeccin|prueba) = \frac{P(prueba|infeccin)P(infeccin)}{P(prueba)}$$

	Infección: Sí	Infección: No	suma
Test: positivo	P(sí|positivo)	P(sin|positivo)	1
Test: negativo	P(sí|negativo)	P(sin|negativo)	1

Por ejemplo:

$$P(s|positivo) = \frac{P(positivo|s)P(s)}{P(positivo)}$$

como normalmente no tenemos $P(positivo)$, usamos la regla de **probabilidad total** en el denominador

$$P(s|positivo) = \frac{P(positivo|s)P(s)}{P(positivo|s)P(s) + P(positivo|no)P(no)}$$

4.22 Teorema de Bayes

La expresion:

$$P(s|positivo) = \frac{P(positivo|s)P(s)}{P(positivo|s)P(s) + P(positivo|no)P(no)}$$

se llama **teorema de Bayes**

Teorema

Si $E1, E2, ..., Ek$ son k eventos mutuamente excluyentes y exhaustivos y B es cualquier evento,

$$P(Ei|B) = \frac{P(B|Ei)P(Ei)}{P(B|E1)P(E1) + ... + P(B|Ek)P(Ek)}$$

Permite invertir los condicionales:

$$P(B|A) \rightarrow P(A|B)$$

O **diseñe** una prueba B en condición controlada A y luego utilícela para **inferir** la probabilidad de la condición cuando la prueba es positiva.

4.23 Ejemplo: teorema de Bayes

Teorema de Bayes:

$$P(s|positivo) = \frac{P(positivo|s)P(s)}{P(positivo|s)P(s) + P(positivo|no)P(no)}$$

sabemos:

- $P(positivo|s) = 0.70$
- $P(positivo|no) = 1 - P(negativo|no) = 0.06$
- la probabilidad de infección y no infección en la población: $P(s) = 0.05$ y $P(no) = 1 - P(s) = 0.95$.

Por lo tanto:

$$P(s|positivo) = 0.47$$

Las pruebas no son tan buenas para **confirmar** infecciones.

4.24 Ejemplo: teorema de Bayes

Apliquémoslo ahora a la probabilidad de no estar infectado si la prueba es negativa.

$$P(no|negativo) = \frac{P(negativo|no)P(no)}{P(negativo|no)P(no) + P(negativo|s)P(s)}$$

La sustitución de todos los valores da

$$P(no|negativo) = 0.98$$

Las pruebas son buenas para **descartar** infecciones.

4.25 Independencia estadística

En muchas aplicaciones, queremos saber si el conocimiento de un evento condiciona el resultado de otro evento.

- hay casos en los que queremos saber si los eventos no están condicionados

4.26 Independencia estadística

Considere los conductores para los cuales medimos sus fallas superficiales y si su capacidad de conducción es defectuosa.

Las **probabilidades conjuntas** estimadas son

	defectos (F)	sin defectos (F')	suma
defectuoso (D)	0.005	0.045	0.05
sin defectos (D')	0.095	0.855	0.95
suma	0.1	0.9	1

donde, por ejemplo, la probabilidad conjunta de F y D es

- $P(D, F) = 0.005$

Las probabilidades marginales son

- $P(D) = P(D, F) + P(D, F') = 0.05$
- $P(F) = P(D, F) + P(D', F) = 0.1.$

4.27 Independencia estadística

¿Cuál es la **probabilidad condicional** de observar un conductor defectuoso si tiene un defecto?

	F	F'
D	P(D|F) = 0.05	P(D|F')=0.05
D'	P(D'|F)=0.95	P(D'|F')=0.95
suma	1	1

¡Las probabilidades marginales y condicionales son las mismas!

- $P(D|F) = P(D|F') = P(D)$
- $P(D'|F) = P(D'|F') = P(D')$

La probabilidad de observar un conductor defectuoso **no** depende de haber observado o no un defecto.

$$P(D) = P(D|F)$$

4.28 Independencia estadística

Dos eventos A y B son estadísticamente independientes si

- $P(A|B) = P(A)$; A es independiente de B
- $P(B|A) = P(B)$; B es independiente de A

y por la regla de la multiplicación, su probabilidad conjunta es

- $P(A \cap B) = P(A|B)P(B) = P(A)P(B)$

la multiplicación de sus probabilidades marginales.

4.29 Productos de productos marginales

	F	F'	suma
D	0.005	0.045	0.05
D'	0.095	0.855	0.95
suma	0.1	0.9	1

Confirme que todas las entradas de la matriz son el producto de los marginales.

Por ejemplo:

- $P(F)P(D) = P(D \cap F)$
- $P(D')P(F') = P(D' \cap F')$

4.30 Ejemplo

Resultados de lanzar dos monedas: $S = (H,H),(H,T),(T,H),(T,T)$

	H	T	suma
H	1/4	1/4	1/2
camiseta	1/4	1/4	1/2
suma	1/2	1/2	1

- Obtener cara en la primera moneda no condiciona obtener cruz en el resultado de la segunda moneda $P(T|H) = P(T) = 1/2$
- la probabilidad de obtener cara y después cruz es el producto de cada resultado independiente $P(H,T) = P(H) * P(T) = 1/4$

Chapter 5

Ejercicios

5.1 Descripción de datos

5.1.0.1 Ejercicio 1

Hemos realizado un experimento 12 veces con los siguientes resultados

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

Responde las siguientes preguntas:

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

5.1.0.2 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
```

Considere 10 contenedores de tamaño 1: [0,1], (1,2]…(9,10).

Responde las siguientes preguntas:

- Calcula las frecuencias relativas de cada resultado y dibuje el histograma

- Calcula las frecuencias acumulativas de cada resultado y dibuje la gráfica acumulativa.

- Dibuja un diagrama de caja.

5.2 Probabilidad

5.2.0.1 Ejercicio 1

El resultado de un experimento aleatorio es medir la gravedad de la misofonía **y** el estado de depresión de un paciente.

- Gravedad de la misofonía: $x \in \{0, 1, 2, 3, 4\}$
- Depresión: $y \in \{0, 1\}$ (no:0, si:1)

```
##   Misofonia.dic depresion.dic
## 1             4             1
## 2             2             0
## 3             0             0
## 4             3             0
## 5             0             0
## 6             0             0
```

Un estudio en 123 pacientes mostró las frecuencias $n_{x,y}$ dadas en la tabla de contingencia:

```
##
##                Depression:0 Depression:1
##   Misophonia:4            0            9
##   Misophonia:3           25            6
##   Misophonia:2           34            3
##   Misophonia:1            5            0
##   Misophonia:0           36            5
```

Supongamos que $N >> 0$ y que las frecuencias **estiman** las probabilidades $f_{x,y} = \hat{P}(X, Y)$

```
##
##                Depression:0 Depression:1
##   Misophonia:4   0.00000000   0.07317073
##   Misophonia:3   0.20325203   0.04878049
##   Misophonia:2   0.27642276   0.02439024
##   Misophonia:1   0.04065041   0.00000000
##   Misophonia:0   0.29268293   0.04065041
```

- ¿Cuál es la probabilidad marginal de misofonía de gravedad 3?
- ¿Cuál es la probabilidad de no ser misofónico **y** no estar deprimido?
- ¿Cuál es la probabilidad de ser misofónico **o** deprimido?
- ¿Cuál es la probabilidad de ser misofónico **y** deprimido?
- Describir en palabras los resultados con probabilidad 0.

5.2.0.2 Ejercicio 2

Hemos realizado un experimento 10 veces con los siguientes resultados

```
##          A     B
## 1     male  dead
## 2     male  dead
## 3     male  dead
## 4   female alive
## 5     male  dead
## 6   female alive
## 7   female  dead
## 8   female alive
## 9     male alive
## 10    male alive
```

- Crear la tabla de contingencia para el número ($n_{i,j}$) de observaciones de cada resultado (A, B)
- Crear la tabla de contingencia para la frecuencia relativa ($f_{i,j}$) de los resultados
- ¿Cuál es la frecuencia marginal de ser hombre?
- ¿Cuál es la frecuencia marginal de estar vivo?
- ¿Cuál es la frecuencia de estar vivo **o** mujer?

5.3 La probabilidad condicional

5.3.0.1 Ejercicio 1

Se prueba el rendimiento de una máquina para producir varillas de torneado de alta calidad. Estos son los resultados de las pruebas

	Redondeado: Sí	Redondeado: No
superficie lisa: sí	200	1
superficie lisa: no	4	2

- ¿Cuál es la probabilidad estimada de que la máquina produzca una varilla que no satisfaga ningún control de calidad?
- ¿Cuál es la probabilidad estimada de que la máquina produzca una varilla que no satisfaga al menos un control de calidad?
- ¿Cuál es la probabilidad estimada de que la máquina produzca varillas de superficie redondeada y alisada?
- ¿Cuál es la probabilidad estimada de que la barra sea redondeada si la barra es lisa?
- ¿Cuál es la probabilidad estimada de que la varilla sea lisa si es redondeada?
- ¿Cuál es la probabilidad estimada de que la varilla no sea ni lisa ni redondeada si no cumple al menos un control de calidad?

- ¿Son eventos independientes la suavidad y la redondez?

5.3.0.2 Ejercicio 2

Desarrollamos un test para detectar la presencia de bacterias en un lago. Encontramos que si el lago contiene la bacteria, la prueba es positiva el 70% de las veces. Si no hay bacterias, la prueba es negativa el 60% de las veces. Implementamos la prueba en una región donde sabemos que el 20% de los lagos tienen bacterias.

- ¿Cuál es la probabilidad de que un lago que dé positivo esté contaminado con bacterias?

5.3.0.3 Ejercicio 3

Se prueba el rendimiento de dos máquinas para producir varillas de torneado de alta calidad. Estos son los resultados de las pruebas

Máquina 1

	Redondeado: Sí	Redondeado: No
superficie lisa: sí	200	1
superficie lisa: no	4	2

Máquina 2

	Redondeado: Sí	Redondeado: No
superficie lisa: sí	145	4
superficie lisa: no	8	6

- ¿Cuál es la probabilidad de que la barra sea redondeada?
- ¿Cuál es la probabilidad de que la varilla haya sido producida por la máquina 1?
- ¿Cuál es la probabilidad de que la varilla no sea lisa?
- ¿Cuál es la probabilidad de que la varilla sea lisa o redondeada o producida por la máquina 1?
- ¿Cuál es la probabilidad de que la varilla quede redondeada si es alisada y de la máquina 1?
- ¿Cuál es la probabilidad de que la varilla no esté redondeada si no está alisada y es de la máquina 2?
- ¿Cuál es la probabilidad de que la varilla haya salido de la máquina 1 si está alisada y redondeada?
- ¿Cuál es la probabilidad de que la varilla haya venido de la máquina 2 si no pasa al menos uno de los controles de calidad?

5.3.0.4 Ejercicio 4

Queremos cruzar una avenida con dos semáforos. La probabilidad de encontrar el primer semáforo en rojo es 0,6. Si paramos en el primer semáforo, la probabilidad de parar en el segundo es 0,15. Mientras que la probabilidad de detenernos en el segundo si no nos detenemos en el primero es 0,25.

Cuando intentamos cruzar ambos semáforos:

- ¿Cuál es la probabilidad de tener que detenerse en cada semáforo?
- ¿Cuál es la probabilidad de tener que parar en al menos un semáforo?
- ¿Cuál es la probabilidad de tener que detenerse en un solo semáforo?
- Si paré en el segundo semáforo, ¿cuál es la probabilidad de que hubiera tenido que parar en el primero?
- Si tuviera que parar en cualquier semáforo, ¿cuál es la probabilidad de que tuviera que hacerlo dos veces?
- ¿Parar en el primer semáforo es un evento independiente de detenerse en el segundo semáforo?

Ahora, queremos cruzar una avenida con tres semáforos. La probabilidad de encontrar un semáforo en rojo solo depende de la anterior. En concreto, la probabilidad de encontrar un semáforo en rojo dado que el anterior estaba en rojo es de 0,15. Mientras que la probabilidad de encontrar un tráfico justo en rojo dado que el anterior estaba en verde es de 0,25. Además, la probabilidad de encontrar el primer semáforo en rojo es de 0,6.

- ¿Cuál es la probabilidad de tener que parar en cada semáforo?
- ¿Cuál es la probabilidad de tener que parar en al menos un semáforo?
- ¿Cuál es la probabilidad de tener que detenerse en un solo semáforo?

consejos:

- Si la probabilidad de que un semáforo esté en rojo depende únicamente del anterior, entonces $P(R_3|R_2, R_1) = P(R_3|R_2, \bar{R}_1) = P(R_3|R_2)$ y $P(R_3|\bar{R}_2, R_1) = P(R_3|\bar{R}_2, \bar{R}_1) = P(R_3|\bar{R}_2)$
- La probabilidad conjunta de encontrar tres semáforos en rojo se puede escribir como: $P(R_1, R_2, R_3) = P(R_3|R_2)P(R_2|R_1)P(R_1)$

5.3.0.5 Ejercicio 5

Una prueba de calidad en un ladrillo aleatorio se define por los eventos:

- Pasar la prueba de calidad: E, no pasar la prueba de calidad: $\bar{E}$
- Defectuoso: D, no defectuoso: $\bar{D}$

Si la prueba diagnóstica tiene sensibilidad $P(E|\bar{D}) = 0.99$ y especificidad $P(\bar{E}|D) = 0.98$, y la probabilidad de pasar la prueba es $P(E) = 0.893$ entonces

- ¿Cuál es la probabilidad de que un ladrillo elegido al azar sea defectuoso $P(D)$?

- ¿Cuál es la probabilidad de que un ladrillo que ha pasado la prueba sea realmente defectuoso?
- La probabilidad de que un ladrillo no sea defectuoso **y** que no pase la prueba
- ¿Son D y $\bar{E}$ estadísticamente independientes?

5.4 Variables aleatorias

5.4.0.1 Ejercicio 1

Dada la distribución de probabilidad para una variable discreta X

$$F(x) = \begin{cases} 0, & x \leq -1 \\ 0.2, & x \in [-1, 0) \\ 0.35, & x \in [0, 1) \\ 0.45, & x \in [1, 2) \\ 1, & x \geq 2 \end{cases}$$

a) encontrar $f(X)$
b) encontrar $E(X)$ y $V(X)$
c) cuál es el valor esperado y la varianza de $Y = 2X + 3$
d) ¿cuál es la mediana de X?

5.4.0.2 Ejercicio 2

Tenemos un sistema de transmisión de píxeles que es totalmente ruidoso. Estamos probando el sistema y hemos diseñado un experimento para transmitir 3 píxeles.

- ¿Cuál es la probabilidad de recibir 0, 1, 2 o 3 errores en la transmisión de 3 píxeles?
- Dibujar la función de masa de probabilidad
- ¿Cuál es el valor esperado del error?
- ¿Cuál es su varianza?
- Dibujar la distribución de probabilidad
- ¿Cuál es la probabilidad de transmitir al menos 1 error?

consejos:

- Espacio muestral: $\{(0,0,0),(1,0,0),(0,1,0),(0,0,1),(0,1,1),(1,0,1),(1,1,0),(1,1,1)\}$
- donde, por ejemplo, el evento $(0,1,1)$ es el evento de recibir el primer píxel sin errores y el segundo y tercer píxeles con errores.

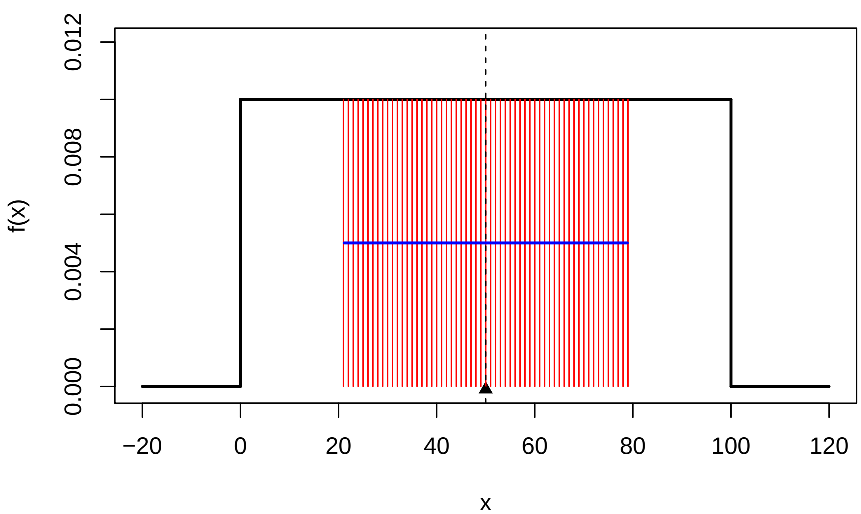
- Todos los eventos son igualmente probables.

5.4.0.3 Ejercicio 3

- para la densidad de probabilidad

$$f(x) = \begin{cases} \frac{1}{100}, & \text{si } x \in (0, 100) \\ 0, & si\,no \end{cases}$$

- calcular la media
- calcule la varianza usando $E(X^2) = V(X) + E(X)^2$
- calcular $P(\mu - \sigma \leq X \leq \mu + \sigma)$
- ¿Cuáles son el primer y tercer cuartiles?



5.4.0.4 Ejercicio 4

Para la densidad de probabilidad

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{si } 0 \leq x \\ 0, & si\,no \end{cases}$$

- Confirmar que se trata de una densidad de probabilidad
- Hallar la distribución de probabilidad $F(a)$
- Calcular la media
- Calcule la varianza usando $E(X^2) = V(X) + E(X)^2$

5.4.0.5 Ejercicio 5

Dada la distribución acumulativa de una variable aleatoria X

$$F(x) = \begin{cases} 0, & x < -1 \\ \frac{1}{80}(17 + 16x - x^2), & x \in [-1, 7) \\ 1, & x \geq 7 \end{cases}$$

calcular:

- $P(X > 0)$
- $E(X)$
- $P(X > 0 | X < 2)$

5.5 Modelos de probabilidad

5.5.0.1 Ejercicio 1

Un motor de búsqueda falla al recuperar información con una probabilidad de 0.1

- Si nuestro sistema recibe solicitudes de búsqueda de 50, ¿cuál es la probabilidad de que el sistema no responda a tres de ellas?
- ¿Cuál es la probabilidad de que el motor complete con éxito búsquedas de 15 antes de la primera falla?
- Consideramos que un buscador funciona suficientemente bien cuando es capaz de encontrar información de 10 solicitudes por cada 2 fallos. ¿Cuál es la probabilidad de que en un ensayo de fiabilidad nuestro motor de búsqueda sea satisfactorio?

5.5.0.2 Ejercicio 2

En una población, la probabilidad de que nazca un niño es $p = 0,51$. Considere una familia de 4 hijos.

- ¿Cuál es la probabilidad de que una familia tenga un solo niño?
- ¿Cuál es la probabilidad de que una familia tenga una sola niña?
- ¿Cuál es la probabilidad de que una familia tenga solo un niño o solo una niña?
- ¿Cuál es la probabilidad de que la familia tenga al menos dos niños?
- ¿Cuál es el número de hijos que debe tener una familia para que la probabilidad de tener al menos una niña sea superior a 0,75?

5.5.0.3 Ejercicio 3

La cantidad promedio de partículas radiactivas que golpean un contador Geiger es de 2.3 segundos.

- ¿Cuál es la probabilidad de contar exactamente 2 partículas en un segundo?
- ¿Cuál es la probabilidad de detectar exactamente 10 partículas en 5 segundos?
- ¿Cuál es la probabilidad de al menos un conteo en dos segundos?
- ¿Cuál es la probabilidad de tener que esperar $2,5$ segundos después de que encendemos el detector?

5.5.0.4 Ejercicio 4

- ¿Cuál es la probabilidad de que la altura de un hombre sea al menos 165cm si la media poblacional es 175cm y la desviación estándar es 10cm?
- ¿Cuál es la probabilidad de que la altura de un hombre esté entre 165cm y 180cm.
- ¿Cuál es la altura que define el 5% de los hombres más pequeños?

5.6 Estimadores puntuales

5.6.0.1 Ejercicio 1

Considere el modelo de probabilidad

$$f(x) = \begin{cases} 1/2 - a, & \text{si } x = -1 \\ 1/2, & \text{si } x = 0 \\ a, & 1\text{si } x = 1 \end{cases}$$

donde a es un parámetro.

Calcule la media y la varianza de la estadística:

$$T = \frac{\bar{X}}{2} + \frac{1}{4}$$

donde $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$

- ¿T es un estimador sesgado de a?
- ¿Es T consistente? es decir, $V(T) \to 0$ cuando $N \to \infty$

5.6.0.2 Ejercicio 2

- ¿Es $\bar{X}^2 = (\frac{1}{N}\sum_{i=1}^{N} X_i)^2$ un estimador imparcial de $E(X)^2$?

5.7 Muestreo y teorema del límite central

5.7.0.1 Ejercicio 1

Un modelo de batería carga hasta 75% de su capacidad en una hora con una desviación estándar de 15%.

- Si cobramos 25, ¿cuál es la probabilidad de que el promedio de la muestra esté a una distancia de 5% del cargo de la media?
- Si cobramos 100, ¿cuál es esa probabilidad?
- Si, en cambio, solo cargamos baterías de 9, ¿cuál es la carga que es superada por el promedio de la muestra con solo 0.015 de probabilidad?

5.7.0.2 Ejercicio 2

Se necesita un componente electrónico para el correcto funcionamiento de un telescopio. Necesita ser reemplazado inmediatamente cuando se desgasta.

La vida media del componente (μ) es de 100 horas y su desviación estándar σ es de 30 horas.

- ¿Cuál es la probabilidad de que el promedio de la vida media de 50 componentes esté dentro de 1 hora de la vida media de un solo componente?
- ¿Cuántos componentes necesitamos para que el telescopio esté operativo 2750 horas consecutivas con una probabilidad de $0,95$?

5.7.0.3 Ejercicio 3

Una máquina automática llena tubos de ensayo con muestras biológicas con una media de $\mu = 130$mg y una desviación estándar de $\sigma = 5$mg.

- para una muestra aleatoria de tamaño 50. ¿Cuál es la probabilidad de que la media muestral (promedio) está entre 128 y 132gr?
- ¿Cuál debe ser el tamaño de la muestra (n) para que la media muestral $\bar{X}$ sea mayor a 131gr con una probabilidad menor o igual a 0.025?

5.7.0.4 Ejercicio 4

En el Caribe, parece haber un promedio de huracanes de 6 por año. Teniendo en cuenta que la formación de huracanes es un proceso de Poisson, los meteorólogos planean estimar el tiempo medio entre la formación de dos huracanes. Planean recolectar una muestra de tamaño 36 para los tiempos entre dos huracanes.

- ¿Cuál es la probabilidad de que su promedio muestral esté entre 45 y 60 días?
- ¿Cuál debe ser el tamaño de la muestra para que tengan una probabilidad de 0.025 de que la media muestral sea mayor a 70 días?

5.7.0.5 Ejercicio 5

La probabilidad de que se encuentre una mutación particular en la población es de 0.4. Si probamos 2000 personas para la mutación:

- ¿Cuál es la probabilidad de que el número total de personas con la mutación esté entre 791 y 809?

sugerencia: use el CLT con una muestra de ensayos de Bernoulli de 2000. Esto se conoce como la aproximación normal de la distribución binomial.

5.8 Máxima verosimilitud

5.8.0.1 Ejercicio 1

Para una variable aleatoria con una función de probabilidad binomial

$$f(x;p) = \binom{n}{x} p^x (1-p)^{n-x}$$

- ¿Cuál es el estimador de máxima verosimilitud de p para una muestra de tamaño 1 de esta variable aleatoria?
- En **un** examen de 100 estudiantes observamos $x_1 = 68$ estudiantes que aprobaron el examen. ¿Cuál es la estimación de p?

5.8.0.2 Ejercicio 2

Tome una variable aleatoria con la siguiente función de densidad de probabilidad

$$f(x) = \begin{cases} (1+\theta)x^\theta, & \text{si } x \in (0,1) \\ 0, & x \notin (0,1) \end{cases}$$

- ¿Cuál es la estimación de máxima verosimilitud para θ?
- Si tomamos una muestra de 5 con observaciones $x_1 = 0,92; \quad x_2 = 0,79; \quad x_3 = 0,90; \quad x_4 = 0,65; \quad x_5 = 0,86$

¿Cuál es el valor estimado del parámetro θ?

5.8.0.3 Ejercicio 3

Tome una variable aleatoria con la siguiente función de densidad de probabilidad

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{si } 0 \le x \\ 0, & si\,no \end{cases}$$

- ¿Cuál es la estimación de máxima verosimilitud para λ?

- Si tomamos una muestra de 5 con observaciones $x_1 = 0.223 \quad x_2 = 0.681; \quad x_3 = 0,117; \quad x_4 = 0,150; \quad x_5 = 0.520$

¿Cuál es el valor estimado del parámetro λ?

5.9 Método de los momentos

5.9.0.1 Ejercicio 1

¿Cuáles son los estimadores de los siguientes modelos paramétricos dados por el método de los momentos?

modelo	f(x)	E(X)
Bernoulli	$p^x(1-p)^{1-x}$	p
binomial	$\binom{n}{x}p^x(1-p)^{n-x}$	np
Geométrico desplazado	$p(1-p)^{x-1}$	$\frac{1}{p}$
Binomial negativo	$\binom{x+r-1}{x}p^r(1-p)^x$	$r\frac{1-p}{p}$
Veneno	$\frac{e^{-\lambda}\lambda^x}{x!}$	λ
Exponencial	$\lambda e^{-\lambda x}$	$\frac{1}{\lambda}$
normales	$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ

5.9.0.2 Ejercicio 2

Tome una variable aleatoria con la siguiente función de densidad de probabilidad

$$f(x) = \begin{cases} (1+\theta)x^\theta, & \text{si } x \in (0,1) \\ 0, & x \notin (0,1) \end{cases}$$

- Calcule $E(X)$ como una función de θ
- ¿Cuál es la estimación de θ utilizando el método de los momentos?
- Si tomamos una muestra de 5 con observaciones $x_1 = 0,92; \quad x_2 = 0,79; \quad x_3 = 0,90; \quad x_4 = 0,65; \quad x_5 = 0,86$

¿Cuál es el valor estimado del parámetro θ?

5.9.0.3 Ejercicio 3

Considere una variable aleatoria discreta X que sigue una distribución binomial negativa con función de masa de probabilidad:

$$f(x) = \binom{x+r-1}{x}p^r(1-p)^x$$

Dado que

- $E(X) = \dfrac{r(1-p)}{p}$
- $V(X) = \dfrac{r(1-p)}{p^2}$

calcular:

- Una estimación del parámetro r y una estimación del parámetro p obtenidas a partir de una muestra aleatoria de tamaño n por el método de los momentos.
- Los valores de las estimaciones de r y p para la siguiente muestra aleatoria:

$$x_1 = 27; \qquad x_2 = 8; \qquad x_3 = 22; \qquad x_4 = 29; \qquad x_5 = 19; \qquad x_5 = 32$$