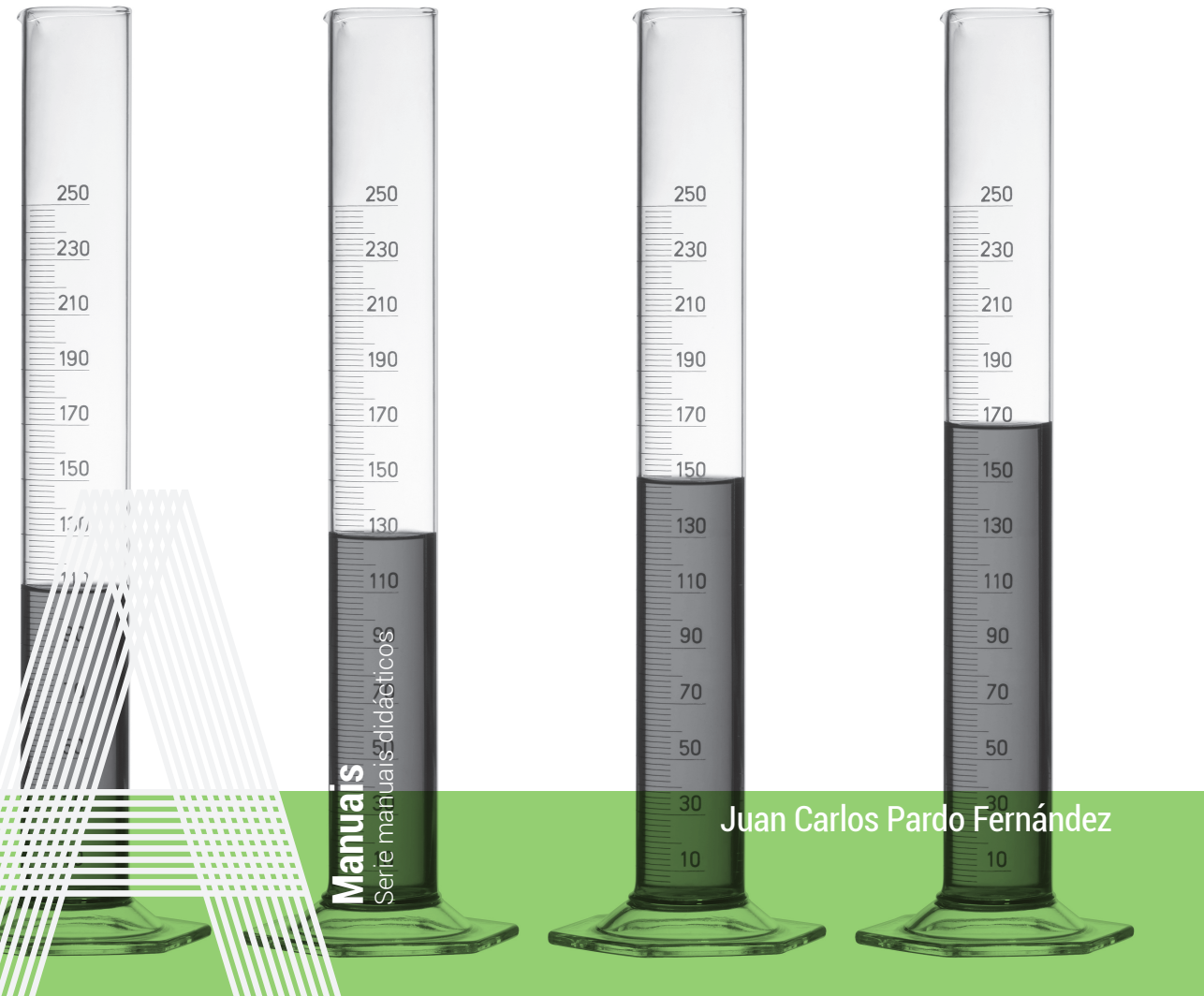
Bioestatística para a Enxeñaría Biomédica



Manuais
Serie manuais didácticos

Juan Carlos Pardo Fernández

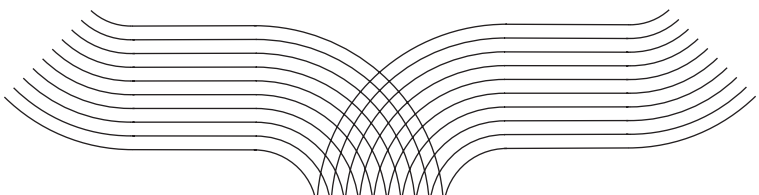
Juan Carlos Pardo Fernández

Juan Carlos Pardo Fernández é licenciado en Matemáticas e doutor en Estatística pola Universidade de Santiago de Compostela. Completou a súa formación académica con estadías de investigación predoutorais e posdoutorais na Universidade Federal de São Carlos (Brasil), na Université catholique de Louvain (Bélxica), na Université Toulouse III-Paul Sabatier (Francia) e na Universidad de Buenos Aires (Arxentina). Desde 2011 é profesor titular do Departamento de Estatística e Investigación Operativa da Universidade de Vigo.

O seu labor investigador enmárcase no ámbito da inferencia estatística non paramétrica e semiparamétrica, concretamente nos campos dos contrastes de bondade de axuste en modelos de regresión e das curvas ROC. Publicou artigos científicos en revistas de referencia na súa área, como *Annals of Statistics*, *Statistical Methods in Medical Research*, *Journal of Business and Economic Statistics* ou *Scandinavian Journal of Statistics*. Ademais participa con frecuencia en congresos e seminarios científicos.

Conta cunha ampla experiencia docente. Desde hai máis de vinte anos imparte materias de estatística en titulacións de grao (especialmente no campo da enxeñaría) e de mestrado (entre outros, nos Mestrados interuniversitarios en Técnicas Estatísticas e en Economía). Ademais foi invitado a impartir cursos de doutoramento e de especialización relacionados cos seus temas de investigación na KULeuven (Bélxica), na Universidad de Sevilla (España), na Universidade de Cabo Verde (Cabo Verde) ou na Universidad de Buenos Aires (Arxentina).

Servizo de Publicacións
Universidade Vigo



Manuais

Serie de manuais didácticos

n.º 082

Edición
Universidade de Vigo
Servizo de Publicacións
Rúa de Leonardo da Vinci, s/n
36310 Vigo

Deseño da portada
Tania Sueiro Graña
Área de Imaxe
Vicerreitoría de Comunicacións e Relacións Institucionais

Maquetación
Tórculo Comunicación Gráfica, S. A.

Fotografía da portada
Adobe Stock

Impresión
Tórculo Comunicación Gráfica, S. A.

ISBN (Libro impreso)
978-84-8158-968-9

Depósito legal
VG 195-2023

Servizo de Publicacións
Universidade Vigo

Bioestatística para a Enxeñaría Biomédica

Juan Carlos Pardo Fernández

Prefacio

Este manual foi concebido como material docente da materia *Bioestatística*, que se imparte no terceiro curso do Grao en Enxeñaría Biomédica da Escola de Enxeñaría Industrial da Universidade de Vigo. Tamén pode ser empregado noutras materias que aborden temas similares.

Os contidos estrutúranse en seis capítulos: técnicas descritivas (táboas de frecuencias, gráficos e medidas resumo), modelos probabilísticos relevantes en bioestatística (breve revisión de variables aletorias e introdución aos problemas de clasificación e á curva ROC), técnicas inferenciais (estimación de parámetros, intervalos de confianza e tests de hipóteses), táboas de continxencia (distribucións conxuntas e condicionadas, test de independencia e riscos relativos), modelos de regresión (modelo lineal, modelos con variables nominais, regresión loxística) e técnicas de análise multivariante (análise de compoñentes principais, métodos cluster e análise discriminante).

A orientación do manual é eminentemente práctica, incluíndo numerosos exemplos e exercicios para ilustrar e comprender os métodos estudados. Ademais, tamén serve para que o alumnado adquira competencias no manexo do software estatístico de distribución libre R, que na actualidade conta cunha ampla difusión no ámbito científico-técnico. Os conxuntos de datos aos que se fai referencia ao longo deste manual están dispoñibles a través da web do autor (juancp.webs.uvigo.gal).

Vigo, marzo de 2023.

Contidos

Capítulo 1

Revisión de técnicas descritivas e software R

Contidos

1.1. Introdución

1.1.1. Obxectivos da Estatística e da Bioestatística

Hoxe en día case todos os estudos técnicos e científicos están acompañados ou baseados en **datos** e polo tanto esixen o seu tratamento mediante técnicas estatísticas adecuadas.

En termos xerais, a **Estatística** ten dous grandes obxectivos:

- Resumir os datos dunha forma axeitada e comprensible. Neste caso, as técnicas empregadas englóbanse dentro da **estatística descritiva**.
- Facer inferencias a partir dos datos para xeralizar as conclusións dos resultados obtidos a grupos ou poboacións máis amplos. Neste caso, as metodoloxías empregadas están dentro da **inferencia estatística**.

A Estatística aplicada a problemas do ámbito médico ou biolóxico recibe o nome de **Bioestatística** ou tamén Biometría.

Por que se precisa a Estatística?

A necesidade de resumir e facer inferencia vén do feito de que en calquera proceso mediante o que se observan datos teremos **variación** ou **variabilidade**, é dicir, os valores observados dunha determinada característica poden variar de individuo a individuo. A análise desta variabilidade inherente aos datos é o obxectivo principal das técnicas estatísticas. Se non houbese variación, entón non habería necesidade da estatística.

Que acadamos co uso da Estatística?

- **Resumir** os datos mediante a obtención tendencias xerais e aspectos comúns.
- Atopar **relacións** entre variables mediante o uso de técnicas gráficas, medidas de asociación, modelos de regresión ou técnicas de análise multivariante.
- Facer **inferencias**, é dicir, xeralizar as conclusións do estudo obtidas a partir dunha mostra representativa a un universo máis amplo mediante o uso de metodoloxías inferenciais, en particular, os contrastes de hipóteses.

1.1.2. Algúns conceptos básicos

Enuméranse a continuación algúns conceptos básicos relacionados cos estudos estatísticos:

- A **poboación** é a maior clase á cal se poden xeralizar os resultados do estudo ou investigación.

- Cada elemento da poboación chámase **individuo**, **caso**, **suxeito** ou simplemente **elemento da poboación**.
- Habitualmente, non podemos observar todos os elementos da poboación. No seu lugar, seleccionamos unha **mostra** segundo un determinado método de mostraxe. O número de elementos da mostra chámase **tamaño mostral**.
- Para cada individuo incluído na mostra observaremos unha ou varias características. Cada unha desas características chámase **variable estatística**, ou simplemente **variable**.
- A información recollida, convenientemente gardada e organizada (habitualmente nun arquivo informático), chámase **conxunto de datos**.

1.1.3. Organización dun conxunto de datos

Os conxuntos de datos normalmente recóllense e gárdanse en arquivos informáticos (por exemplo, en follas de cálculo). Moitas veces é útil traballar con documentos de texto en formatos `.txt` ou `.csv` debido ao seu fácil manexo. Os documentos de texto poden obterse a partir das follas de cálculo.

Cando creamos conxuntos de datos (especialmente cando se emprega unha folla de cálculo), convén seguir as seguintes **recomendacións** para a súa correcta organización:

- As variables codifícanse en columnas.
- A primeira cela de cada columna (no caso dunha folla de cálculo) ou a primeira liña (no caso dun documento de texto) resérvase para o nome da variable. É conveniente que este nome sexa sinxelo pero que, dalgunha maneira, permita identificar a variable da que se trata de xeito fácil.
- As filas (ou liñas) representan aos distintos individuos ou elementos da poboación sobre os que observamos as variables.
- As variables de natureza numérica recóllense obviamente con números, preferentemente coa maior precisión posible. As variables non numéricas (que máis adiante lles chamaremos variables cualitativas ou factores) poden codificarse con códigos ou abreviaturas que nos permitan traballar comodamente.
- Se sobre un mesmo individuo se observan varias variables, entón as observacións correspondentes deben estar sempre na mesma fila (ou liña) do conxunto de datos. Desta maneira non se perde a conexión entre esas observacións e permitirá o estudo de posibles relacións entre as variables. É importante ter isto en conta cando se reorganiza ou reordena o conxunto de datos.

Exemplo. Os documentos `datos-diabetes.xlsx` e `datos-diabetes.txt` conteñen o mesmo conxunto de datos en formato folla de cálculo e en formato texto[1]. A este conxunto de datos chamarémoslle DIABETES. A continuación aparece parte do conxunto de datos:

[1] Os conxuntos de datos aos que se fai referencia ao longo deste manual están dispoñibles a través da web do autor (juancp.webs.uvigo.gal).

		variables						
				↙	↓	↘		
		id	glu	bp	skin	bmi	age	npreg	type
		...	...	...	...	...	...	...	...
		3	77	82	41	35.80	35	5	No
	↗	4	165	76	43	47.90	26	0	No
individuos	→	5	107	60	25	26.40	23	0	No
ou **casos**	↘	6	97	76	27	35.60	52	5	Yes
		7	83	58	31	34.30	25	3	No
		...	...	...	...	...	...	...	...

□

1.1.4. R e RStudio. Importar conxuntos de datos en R

Neste manual empregaremos o software de distribución libre **R** (www.r-project.org) para facer as nosas análises estatísticas. A interface **RStudio** (dispoñible a través da web posit.co) resulta cómoda, así que tamén a empregaremos.[2] Por suposto, hai outros softwares dispoñibles. Non obstante, R presenta varias vantaxes: é de distribución libre e gratuíta, é un proxecto colaborativo, actualmente conta cunha ampla difusión entre a comunidade científico-técnica e permite intercomunicación con outras linguaxes de programación.

Para traballar en R, necesitaremos importar os nosos conxuntos de datos á sesión de traballo. O primeiro que temos que facer ao iniciar R é que indicarlle cal é o noso **directorio de traballo** ("Working Directory" en R). Para isto empregaremos a función `setwd(./path/.)` para especificar o directorio onde están gardados os datos e no que queremos gardar as nosas análises e gráficos. En RStudio pode facerse a través do menú Session > Set Working Directory.

Para cargar os conxuntos de datos en R temos as seguinte funcións:

`read.table()` para documentos `.txt`;

`read.csv()` ou `read.csv2()` para documentos `.csv`.

Os comandos e instrucións empregados nas nosas análises gárdanse nun documento de texto que se chama **script** e ten extensión `.R`.

Exemplo. Para importar a R o documento de texto `datos-diabetes.txt`, primeiro establecemos como directorio de traballo o directorio onde o temos gardado e despois escribimos

[2] Instrucións básicas para instalar R e RStudio:

- Ir á web https://www.r-project.org e seguir as instrucións que aparecen en "download R".
- Ir á web https://posit.co > Download RStudio > RStudio Desktop Free. Clicar en "Download" e seguir as instrucións.

```
> diabetes <- read.table("datos-diabetes.txt",header=TRUE)
```

Para importar o documento `datos-diabetes.csv` en R, escribimos

```
> diabetes <- read.csv("datos-diabetes.csv",header=TRUE,dec=".",sep=",")
```

Agora o obxecto `diabetes` é un **conxunto de datos** en R ("**data frame**" na nomenclatura propia de R). Esta clase de obxectos é a forma máis conveniente de manexar os conxuntos de datos en R. Esencialmente, un "data frame" é unha matriz na cal as columnas teñen nomes. As columnas correspóndense coas variables e as filas, cos individuos ou casos.

Para acceder a unha variable deste conxunto de datos empregamos

```
> diabetes$glu
```

ou

```
> diabetes[,1]
```

ou

```
> diabetes[,"glu"]
```

Nótese que nos comandos anteriores sempre nos referimos primeiro ao conxunto de datos a través do seu nome. Para evitar escribir o nome do conxunto de datos cada vez que precisemos chamar a unha variable podemos empregar a función `attach()` para cargalo á sesión de traballo:

```
> attach(diabetes)
```

Unha vez que rematemos de traballar con este coxunto de datos, debemos "descargalo" coa función `detach()` para evitar confusións cos nomes das variables doutros conxuntos de datos:

```
> detach(diabetes)
```

Exemplo. O conxunto de datos DIABETES contén observacións de varias variables relacionadas cun estudo sobre diabetes nunha comunidade do pobo indíxena Pima, que vive cerca da cidade de Phoenix (Arizona, Estatos Unidos). Neste estudo recolléronse datos de 200 mulleres que foron diagnosticadas ou non de padecer diabetes segundo os criterios da Organización Mundial da Saúde (OMS). O conxunto de datos contén información sobre as seguintes variables:

- *glu*: concentración de glucosa no sangue (mg/dl).
- *bp*: presión sanguínea diastólica (mm Hg).
- *skin*: grosor do pregamento cutáneo do tríceps (mm).

- *bmi*: índice de masa corporal (*body mass index*, en inglés). Calcúlase coa fórmula (masa en kg)/(altura en m)^2.
- *age*: idade (en anos).
- *npreg*: número de embarazos.
- *type*: Si (Yes) / Non (No) indicando se padece ou non diabetes segundo os criterios da OMS.

A función `head()` permite visualizar unha parte do conxunto de datos:

```
> head(diabetes)
```

```
  glu bp skin  bmi age npreg type
1  86 68   28 30.2  24     5   No
2 195 70   33 25.1  55     7  Yes
3  77 82   41 35.8  35     5   No
4 165 76   43 47.9  26     0   No
5 107 60   25 26.4  23     0   No
6  97 76   27 35.6  52     5  Yes
```

Tamén podemos empregar a función `str()`, que amosa a estrutura do conxunto de datos:

```
> str(diabetes)
```

```
'data.frame':        200 obs. of  7 variables:
 $ glu  : int  86 195 77 165 107 97 83 193 142 128 ...
 $ bp   : int  68 70 82 76 60 76 58 50 80 78 ...
 $ skin : int  28 33 41 43 25 27 31 16 15 37 ...
 $ bmi  : num  30.2 25.1 35.8 47.9 26.4 35.6 34.3 25.9 32.4 43.3 ...
 $ age  : int  24 55 35 26 23 52 25 24 63 31 ...
 $ npreg: int  5 7 5 0 0 5 3 1 3 2 ...
 $ type : chr  "No" "Yes" "No" "No" ...
```

□

1.1.5. Tipos de variables

Cando analizamos unha variable estatística é importante coñecer a súa natureza, xa que as técnicas empregadas dependerán do tipo de variable coa que esteamos traballando.

Unha clasificación moi básica das variables vén dada pola súa natureza numérica/non numérica:

- **Variables cualitativas:** as súas observacións non son cantidades numéricas, senón cualidades. Estas variables tamén reciben o nome de **factores** e os seus posibles valores chámanse **niveis**.
- **Variables cuantitativas:** as súas observacións son cantidades numéricas. Entre as variables cuantitativas, distinguimos dous tipos:
  - Variables discretas: son aqueles que toman valores nun conxunto finito ou nun conxunto numerable que se pode identificar cos números enteiros.
  - Variables continuas: son aquelas que toman valores nun intervalo (posiblemente non acotado) de números reais.

  Exemplo. No conxunto de datos DIABETES a variable *type* é un factor con dous niveis (diabética / non diabética). As variables *glu*, *age* e *npreg* son cuantitativas. A variable *npreg* é discreta. A variable *glu* é continua. A variable *age* está rexistrada como discreta (aínda que estritamente falando sería continua). □

Unha clasificación máis exhaustiva vén dada polos catro seguintes tipos de variables:

- **Nominal:** os seus valores son cualidades que non posúen ningún tipo de orde.
- **Ordinal:** os seus valores son cualidades medidas nunha escala non numérica, pero que poden ser ordenadas dalgún xeito e cumpren unha relación de transitividade (se a cualidade a é menos ca b e b é menos ca c, entón a tamén é menos ca c).
- **De intervalo:** este tipo de variables están medidas nunha escala numérica que non ten un cero absoluto. Nesta escala, as magnitudes das diferenzas entre observacións teñen sentido numérico (cousa que non ocorre nas variables de tipo ordinal), pero a escala non se pode interpretar en sentido absoluto porque o cero é arbitrario.

  Nalgunhas ocasións, a clasificación dunha variable como ordinal ou de intervalo pode non ser completamente clara.

  Exemplo. O exemplo típico de variable de intervalo é a temperatura. O feito de asignar o número 0 a unha determinada temperatura é completamente arbitrario (compárense, por exemplo, o significado de 0 graos Celsius con 0 graos Farenheit). Se nas localidades A e B se rexistran temperaturas de 20°C e 10°C, respectivamente, entón podemos dicir que a temperatura é 10 graos máis alta en A ca en B, pero carece de sentido dicir que a temperatura é o dobre en A ca en B. □
- **De razón** ou **de ratio:** este tipo de variables están medidas nunha escala que ten un cero absoluto. Chámanse así porque as razóns ou cocientes entre as observacións teñen significado numérico.

As variables de tipo nominal son sempre cualitativas. As variables de intervalo e de razón son sempre cuantitativas. A variables ordinais poden ser cualitativas ou cuantitativas, dependendo do contexto. Usualmente, as variables nominais e ordinais son consideradas como factores nas análises estatísticas.

Exemplo. Nun laboratorio estúdase o efecto dun determinado tratamento sobre o peso dunhas crías de rata. O documento de texto `datos-ratpups.txt` contén a información correspondente. A este conxunto de datos chamarémoslle RATPUPS. Contén as seguintes variables:

- *weight*: peso (en g).
- *sex*: male (macho) / female (femia).
- *litter*: número identificador da camada.
- *littersize*: número de individuos nacidos na camada.
- *treatment*: tratamento administrado, con tres niveis (control < low < high).

```
> ratpups <- read.table(file="datos-ratpups.txt",header=TRUE)
> str(ratpups)

'data.frame':        322 obs. of  5 variables:
 $ weight    : num  6.6 7.4 7.15 7.24 7.1 6.04 6.98 7.05 6.95 6.29 ...
 $ sex       : chr  "male" "male" "male" "male" ...
 $ litter    : int  1 1 1 1 1 1 1 1 1 1 ...
 $ littersize: int  12 12 12 12 12 12 12 12 12 12 ...
 $ treatment : chr  "control" "control" "control" "control" ...
```

□

Exercicio 1.1. *De que tipo son as variables do conxunto de datos* RATPUPS*?*

1.2. Ferramentas básicas para a descrición de datos

Nesta sección veremos diversas técnicas para resumir a información contida nun conxunto de datos que permiten describir as características fundamentais das variables estudadas. Estas técnicas inclúense dentro da **estatística descritiva**. Simultaneamente veremos como aplicar estas técnicas con R.

1.2.1. Táboas de frecuencias

Unha forma moi común de resumir datos é a través de **táboas de frecuencias**. A construción deste tipo de táboas consiste en listar os valores observados dunha variable xunto co número de veces que aparece cada un deses valores. Primeiro teremos que establecer un conxunto de clases mutuamente excluíntes nas cales se clasificarán as observacións. A continuación organizaranse os valores obtidos nunha táboa. Podemos traballar coas seguintes cantidades:

- **Frecuencia absoluta**: número de observacións en cada clase.
- **Frecuencia relativa**: cociente da frecuencia absoluta entre o número total de observacións. As frecuencias relativas habitualmente danse en porcentaxes (multiplicando por 100) e resultan máis informativas ca as frecuencias absolutas.

Exemplo. Consideremos o conxunto de datos[3] DIABETES. Para obter as frecuencias absolutas da variable *type* en R empregamos

```
> table(type)
```

```
type
 No Yes
132  68
```

e para obter as frecuencias relativas, primeiro obtemos o número total de observacións

```
> n <- length(type)
```

e despois facemos

```
> table(type)/n
```

```
type
  No  Yes
0.66 0.34
```

ou ben

```
> 100*table(type)/n
```

```
type
 No Yes
 66  34
```

Cando incorporamos a táboa a un documento de traballo (informe, artigo etc.) debemos ser coidadosos co seu formato (véxase, por exemplo, a Táboa 1.1).

A táboa de frecuencias obtida directamente da variable *age* resulta

```
> table(age)
```

[3] Ao longo deste documento asumimos que se empregan convenientemente as funcións `attach()` e `detach()` para cargar os conxuntos de datos coas súas variables á sesión de traballo de R.

Táboa 1.1: Conxunto de datos DIABETES. Frecuencias absolutas e relativas da variable *type*.

	frecuencia absoluta	frecuencia relativa (%)
non diabética	132	66.0 %
diabética	68	34.0 %
total	200	100.0 %

```
age
21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43
21 17 13 17  6 10  8 12  6  6  7  4  5  4  2  2  4  3  3  5  8  3  3
44 45 46 48 49 50 51 52 54 55 57 58 59 60 62 63
 1  3  5  1  1  1  2  3  1  2  1  3  2  2  2  1
```

Obviamente, non é moi informativa xa que temos moitos valores distintos da idade. Cando se traballa cunha variable cuantitativa que presenta moitos valores distintos, resulta máis axeitado contruír unha táboa de frecuencias por intervalos. Por exemplo, no caso da variable *age* podemos establecer os grupos de idades da forma $21-25, 26-30, \ldots, 56-60, 61-65$. En R facemos o seguinte:

```
> table(cut(age,breaks=seq(20,65,by=5)))

(20,25] (25,30] (30,35] (35,40] (40,45] (45,50] (50,55] (55,60]
     74      42      22      17      18       8       8       8
(60,65]
      3

> 100*table(cut(age,breaks=seq(20,65,by=5)))/n

(20,25] (25,30] (30,35] (35,40] (40,45] (45,50] (50,55] (55,60]
   37.0    21.0    11.0     8.5     9.0     4.0     4.0     4.0
(60,65]
    1.5
```

No caso das variables continuas, empregar intervalos para construír táboas de frecuencias resulta imprescindible. A menos que haxa un redondeo moi forte, o habitual é que nas variables continuas non aparezan valores repetidos. Polo tanto non ten ningún sentido contar cantas veces aparece cada valor, xa que probablemente aparecerá unha única vez. Por exemplo, a variable *glu* toma valores entre 56 e 199. Podemos empregar intervalos da forma $(50, 75], (75, 100], \ldots, (150, 175], (175, 200]$:

```
> table(cut(glu,breaks=c(50,75,100,125,150,175,200)))

  (50,75]  (75,100] (100,125] (125,150] (150,175] (175,200]
        6        49        62        42        23        18
```

No caso das variables cuantitativas tamén pode resultar interesante incluír na táboa as **frecuencias acumuladas** absolutas e/ou relativas. As frecuencias acumuladas indican o número (no caso das absolutas) ou a proporción/porcentaxe (no caso das relativas) de individuos que presentan un valor da variable que é menor ou igual ca un valor dado. Por exemplo, no caso da variable *glu*, as frecuencias relativas acumuladas correspondentes aos intervalos antes citados son:

```
> 100*cumsum(table(cut(glu,breaks=c(50,75,100,125,150,175,200))))/n

  (50,75]  (75,100] (100,125] (125,150] (150,175] (175,200]
      3.0      27.5      58.5      79.5      91.0     100.0
```

A Táboa 1.2 amosa a táboa completa de frecuencias da variable *glu*. Se miramos na segunda e terceira columnas desta táboa, vemos que 62 mulleres (ou, equivalentemente, o 31.0 %) presentan unha concentración de glucosa entre 100 e 125. As columnas cuarta e quinta indícannos que 117 mulleres (58.5 %) presentan unha concentración de glucosa de como moito 125.

Táboa 1.2: Conxunto de datos DIABETES. Táboa de frecuencias da variable *glu*.

concentración de glucosa	frequencia absoluta	frecuencia relativa (%)	frecuencia absoluta acumulada	frecuencia relativa acumulada (%)
$(50, 75]$	6	3.0	6	3.0
$(75, 100]$	49	24.5	55	27.5
$(100, 125]$	62	31.0	117	58.5
$(125, 150]$	42	21.0	159	79.5
$(150, 175]$	23	11.5	182	91.0
$(175, 200]$	18	9.0	200	100.0
total	200	100.0		

□

Como vemos nos dous exemplos anteriores, ao facer a clasificación por intervalos temos que decidir o número de intervalos co que imos traballar. Convén escoller un número nin moi pequeno nin moi grande, xa que en ambos os dous casos obteriamos unha táboa pouco informativa. Unha regra que nos permite ter unha idea dun número de intervalos razoable é a **regra de Sturges**, que suxire tomar un número de intervalos próximo a $1 + \log_2 n$, onde n é o número total de observacións.

Exercicio 1.2. *Constrúe unha táboa de frecuencias para resumir a variable* bmi *do conxunto de datos* DIABETES.

1.2.2. Gráficos

Outra posibilidade para resumir datos é o uso de ferramentas gráficas. Veremos agora tres exemplos moi básicos: o gráfico de sectores, o gráfico de barras e o histograma. Máis adiante estudaremos o gráfico de caixas ou boxplot.

Gráfico de sectores e gráfico de barras para variables nominais

Os gráficos de sectores e de barras son útiles para describir variables nominais. Os dous conteñen a mesma información. O **gráfico de sectores** representa os valores ou niveis da variable como sectores circulares e o tamaño de cada sector é proporcional á frecuencia do nivel que representa. O **gráfico de barras** consiste en varias barras de altura proporcional aos valores das frecuencias do niveis correspondentes. Os dous tipos de gráficos poden empregarse para variables nominais. Non obstante, cando o número de clases é grande, o gráfico de sectores pode deixar de ser informativo e polo tanto debe evitarse o seu uso. No caso de variables de tipo ordinal resulta máis axeitado o gráfico de barras para indicar claramente a orde entre os niveis.

Exemplo. A Figura 1.1 amosa o gráfico de sectores e o gráfico de barras da variable *type* do conxunto de datos DIABETES. Para obtelos en R escribimos os comandos seguintes:

```
> pie(table(type))
> barplot(table(type))
```

Nótese que as funcións `pie()` e `barplot()` non se aplican directamente sobre unha variable, senón sobre unha táboa de frecuencias.

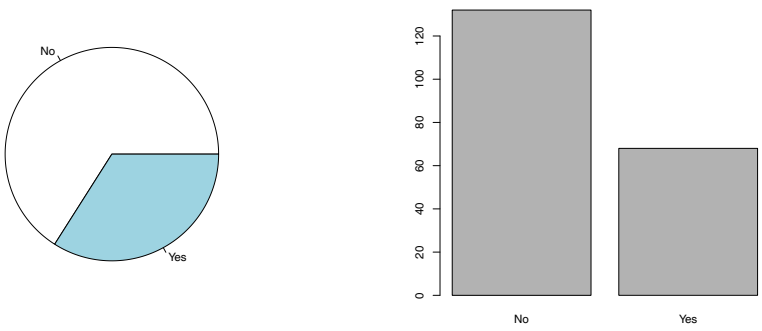


Figura 1.1: Conxunto de datos DIABETES. Gráfico de sectores (esquerda) e gráfico de barras (dereita) da variable *type*.

Os gráficos de barras tamén se poden empregar para comparar dúas ou máis variables. Por exemplo, no caso do conxunto de datos RATPUPS, podemos facer un gráfico que combine as variables *sex* e *treatment*, tal e como se amosa na Figura 1.2.

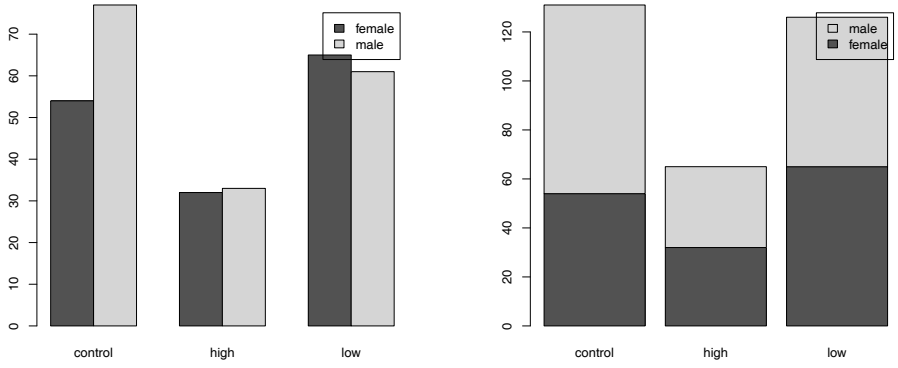


Figura 1.2: Conxunto de datos RATPUPS. Dous exemplos de gráfico de barras combinando as variables *sex* e *treatment*. Nótese que os niveis da variable de tipo ordinal *treatment* non aparecen na orde natural control – low – high.

```
> barplot(table(ratpups$sex,ratpups$treatment),beside=TRUE,legend=TRUE)
```

Nótese que nos gráficos da Figura 1.2 non se usa a orde natural dos niveis da variable *treatment*, que é de tipo ordinal. Isto débese a que R non é capaz de identificar a orde correcta a menos que llo especifiquemos. Para facer isto, escribimos:

```
> ratpups$treatment <- factor(ratpups$treatment,
+                             levels=c("control","low","high"))
```

A Figura 1.3 amosa os mesmos gráficos da Figura 1.2 cos niveis da variable *treatment* ordenados correctamente. □

Histogramas para variables continuas

Para representar unha variable numérica nun gráfico, debemos empregar un gráfico de barras no caso das variables discretas con poucos posibles valores. No caso de ter moitos valores distintos ou cando temos unha variable continua, entón debemos empregar un histograma. O **histograma** é similar ao gráfico de barras no que cada barra se reempraza por un rectángulo que ten por base o intervalo de valores que representa e a altura constrúese de tal maneira que a área do rectángulo sexa proporcional á frecuencia relativa do intervalo correspondente. O máis habitual é construír os histogramas con intervalos da mesma lonxitude. Para elixir o número de intervalos pódese empregar a regra de Sturges que comentamos no apartado das táboas de frecuencias.

Exemplo. A Figura 1.4 amosa histogramas da concentración de glucosa para mulleres non diabéticas e diabéticas. En R escribimos:

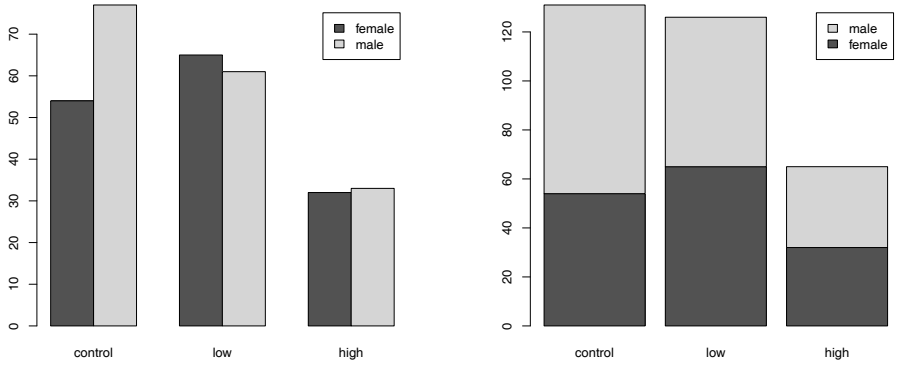


Figura 1.3: Conxunto de datos RATPUPS. Dous exemplos de gráfico de barras combinando as variables *sex* e *treatment*. Nótese que agora os niveis da variable ordinal *treatment* aparecen na orde correcta control – low – high.

```
> hist(glu[type=="No"])
> hist(glu[type=="Yes"])
```

Nótese que os gráficos que obtemos en R non coinciden exactamente cos da Figura 1.4. Isto débese a que se modificaron algúns dos seus parámetros gráficos. En particular, para obter o gráfico da dereita, escribimos

```
> hist(glu[type=="Yes"],
+      freq=FALSE,main=" ",xlab="glu",ylab="densidade",xlim=c(40,200))
```

□

Exercicio 1.3. *Investiga os axustes gráficos que aparecen no código anterior. Podes empregar a axuda de R escribindo* `?hist` *na consola.*

Exercicio 1.4. *Para o conxunto de datos* RATPUPS*, prepara táboas de frecuencias e gráficos para as variables* treatment *e* weight*. Como combinarías estas variables de forma gráfica?*

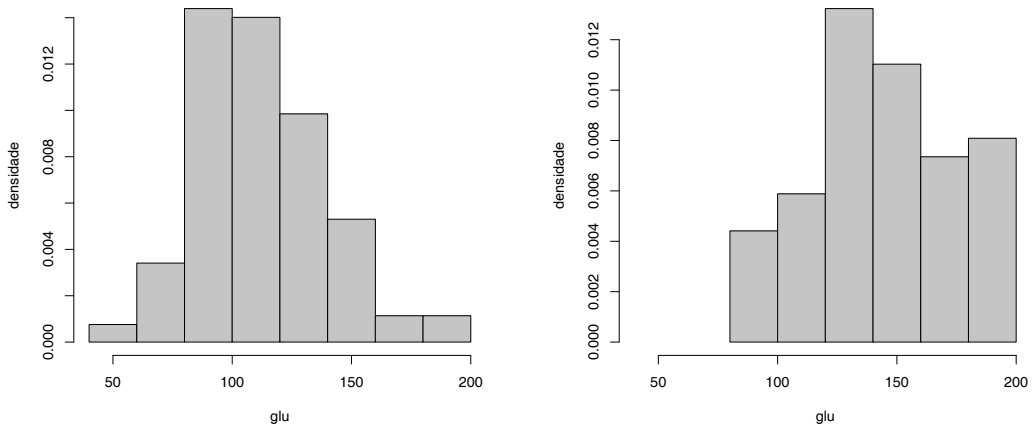


Figura 1.4: Conxunto de datos DIABETES. Histogramas da variable *type* para mulleres non diabéticas (esquerda) e diabéticas (dereita).

1.2.3. Medidas resumo

Ademais de táboas e gráficos, moitas veces tamén resulta conveniente dar cantidades numéricas como resumo dos datos. Estas cantidades proporcionarán información sobre diversas características dos datos: localización, dispersión, simetría etc. Obviamente, os resumos numéricos só se poden empregar con variables numéricas.

Medidas de localización

As medidas de localización, tamén chamadas medidas de centralidade ou de tendencia central, indican onde se sitúan os valores da mostra con respecto á escala na que están rexistrados os seus valores. As principais medidas de localización son a media mostral, a mediana mostral e os cuantís mostrais.

Para poder dar as fórmulas destas cantidades de forma precisa necesitamos algo de notación matemática. Supoñamos que $X_1, X_2, \ldots, X_n$ representa unha mostra de n observacións da variable cuantitativa X.

- A **media mostral** é

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Esta cantidade representa o "centro de gravidade" dos datos, xa que se cumpre que

$$\sum_{i=1}^{n}(X_i - \bar{X}) = 0.$$

- A **mediana mostral** é o valor que divide a mostra ordenada en dúas metades, de xeito que o 50 % dos datos son menores ca a mediana e o 50 % son maiores ca a mediana. Unha vantaxe da mediana mostral sobre a media mostral é que a mediana é máis robusta ante a presenza de datos atípicos na mostra.

  Cando o histograma dos datos é simétrico, entón a media e a mediana mostrais serán case iguais. Diferenzas entre a media e a mediana indican posibles asimetrías na distribución dos datos.

- Os cuantís mostrais son xeralizacións da mediana. Dada unha proporción $0 < p < 1$ (ou, equivalentemente, unha porcentaxe $p100\,\%$), o **cuantil mostral de orde** p (tamén chamado **percentil mostral** $p100\,\%$) é o valor que divide a mostra ordenada en dúas partes, de xeito que a primeira contén o $p100\,\%$ dos datos a segunda, o $(1-p)100\,\%$.

  Os cuantís de orde 0.25, 0.50 e 0.75 chámanse **cuartís mostrais** (primeiro cuartil, segundo cuartil, que coincide coa mediana, e terceiro cuartil, respectivamente).

En R temos funcións que permiten calcular estas medidas resumo: `mean()` para calcular a media mostral, `median()` para calcular a mediana mostral, `quantile()` para calcular cuantís mostrais.

Exemplo. Consideremos por exemplo o conxunto de datos DIABETES:

```
> mean(age)

[1] 32.11

> median(age)

[1] 28

> quantile(age,c(0.10,0.25,0.50,0.75,0.90))

  10%   25%   50%   75%   90%
21.00 23.00 28.00 39.25 49.10
```

A función `summary()` devolve varias destas cantidades á vez:

```
> summary(age)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  21.00   23.00   28.00   32.11   39.25   63.00
```

Cando traballamos con conxuntos de datos que inclúen variables cualitativas (factores) que permiten formar grupos de individuos en función dos seus niveis, entón pode resultar interesante comparar as medidas resumo en cada grupo. Isto pódese facer de varias maneiras. Por exemplo, no conxunto de datos DIABETES, se queremos comparar as medidas resumo da variable *age* en cada un dos dous grupos definidos polos niveis do factor *type*, podemos escribir

```
> summary(age[type=="No"])

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  21.00   23.00   26.00   29.23   31.25   63.00

> summary(age[type=="Yes"])

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  21.00   29.00   36.00   37.69   45.25   62.00
```

ou, equivalentemente, empregar a función `tapply()`

```
> tapply(age,type,summary)

$No
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  21.00   23.00   26.00   29.23   31.25   63.00

$Yes
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  21.00   29.00   36.00   37.69   45.25   62.00
```

□

Medidas de dispersión

As medidas de dispersión proporcionan información sobre a variabilidade dos datos. As máis empregadas son a varianza e a desviación estándar mostrais, o rango mostral, o rango intercuartílico mostral e o coeficiente de variación mostral.

- A **varianza mostral** é

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

  e mide a variabilidade da mostra arredor da media mostral. As unidades da varianza mostral son as unidades da variable ao cadrado (por exemplo, as unidades da varianza mostral da variable *age* son anos^2), o cal non resulta interpretable na maior parte dos casos. Por iso se emprega tamén a **desviación estándar mostral**

$$S = \sqrt{S^2},$$

  que ten as unidades orixinais da variable.

- O **rango mostral** é a distancia entre os valores mínimo e máximo da mostra.

- O **rango intercuartílico** é a distancia entre o primeiro cuartil e o terceiro cuartil. Nótese que o intervalo delimitado entre os cuartís primeiro e terceiro contén o 50 % central dos datos.
- O **coeficiente de variación mostral** é

$$CV = \frac{S}{|\bar{X}|}.$$

  Esta cantidade non ten unidades, e polo tanto pode empregarse para comparar a dispersión de variables de distinta natureza.

En R, as funcións `var()`, `sd()`, `range()` e `IQR()` calculan a varianza mostral, a desviación estándar mostral, o rango mostral (máis concretamente, o mínimo e o máximo) e o rango intercuartílico mostral, respectivamente.

Exemplo. No conxunto de datos DIABETES, as desviacións estándar mostrais das variables *glu* e *skin* son `sd(glu)`=31.667 e `sd(skin)`=11.725, respectivamente. Estas cantidades non teñen as mesmas unidades e polo tanto non podemos comparar os seus valores. Para comparar a dispersión destas dúas variables debemos empregar os seus coeficientes de variación mostrais, que son `sd(glu)/abs(mean(glu))`=0.255 e `sd(skin)/abs(mean(skin))`=0.401, respectivamente. Vemos que, de feito, *skin* presenta maior variabilidade ca *glu*. □

Gráfico de caixas ou boxplot

O **gráfico de caixas** ou **boxplot** é unha ferramenta gráfica que recolle información sobre a localización e a dispersión dos datos. Resulta especialmente útil para comparar os valores dunha variable numérica con respecto aos valores dun factor. Tamén permite detectar asimetría na distribución dos datos.

A construción do boxplot baséase nas seguintes cantidades mostrais:

- O mínimo $m = \text{mín}\{X_1, \dots, X_n\}$ e o máximo $M = \text{máx}\{X_1, \dots, X_n\}$.
- 1^{o} cuartil (Q_1), mediana (med.), 3^{o} cuartil (Q_3).
- As cantidades

$$L = \text{máx}\{m, Q_1 - 1.5(Q_3 - Q_1)\} \quad \text{e} \quad U = \text{mín}\{M, Q_3 + 1.5(Q_3 - Q_1)\}.$$

- As observacións que quedan fóra do intervalo (L, U), é dicir, a unha distancia maior de 1.5 veces o rango intercuartílico respecto do primeiro ou do terceiro cuartil, reciben o nome de **datos atípicos** (ou **outliers**).

O gráfico constrúese como se indica no diagrama da Figura 1.5. Nel represéntanse a mediana, os cuartís e o mínimo e o máximo dos valores non atípicos (que denotamos m_L e M_U, respectivamente). Ademais, os datos atípicos indícanse con pequenos círculos ou cruces. Nótese que no diagrama non hai datos atípicos pola parte da esquerda, polo que neste caso $L = m_L$, que á súa vez coincide co mínimo dos datos.

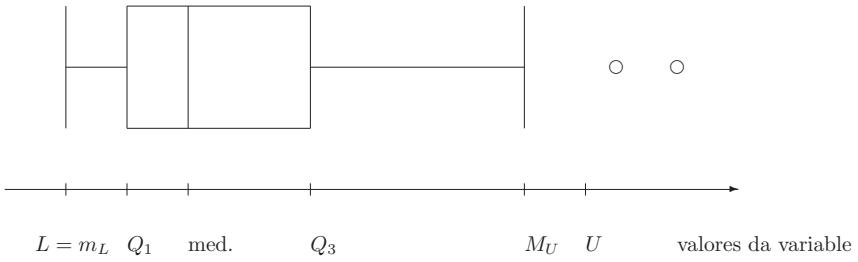


Figura 1.5: Construción do boxplot.

Exemplo. No conxunto de datos DIABETES, podemos dividir os valores da variable *glu* en dous grupos de acordo cos niveis do factor *type* e comparar os correspondentes valores mediante boxplots. En R simplemente temos que escribir

```
> boxplot(glu~type)
```

Na Figura 1.6 podemos ver que a variable *glu* presenta valores máis elevados no grupo de mulleres que sufren diabetes. No grupo de mulleres non diabéticas aparecen 4 datos atípicos.

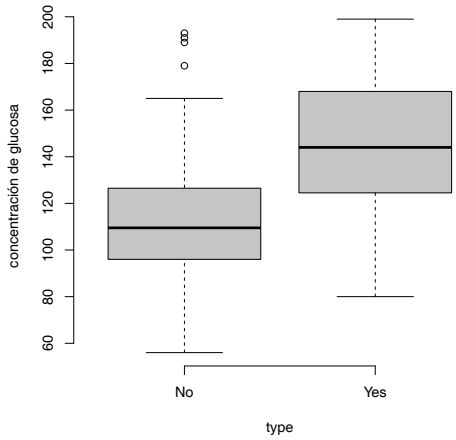
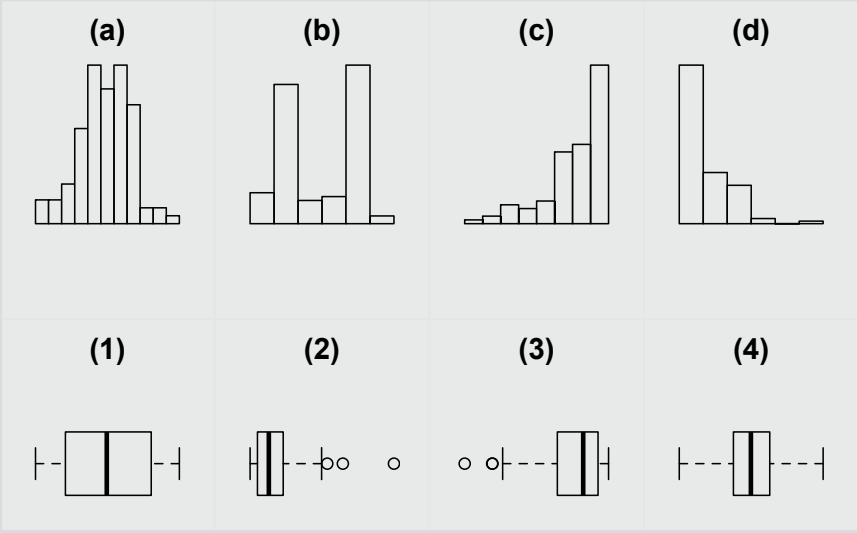


Figura 1.6: Conxunto de datos DIABETES. Boxplots da variable *glu* segundo os niveis da variable *type* ("No" - non diabética, "Yes" - diabética).

□

Exercicio 1.5. *Constrúe boxplots da variable* bp *en función dos niveis do factor* type *do conxunto de datos* DIABETES. *Hai diferenzas claras? E no caso da variable* bmi*?*

Exercicio 1.6. *Relaciona de forma razoada cada histograma co diagrama de caixas que representa o mesmo conxunto de datos.*



En R xa vimos como extraer unha parte dos valores dunha variable para un nivel particular dun factor. Por exemplo `glu[type=="Yes"]` extrae as concentracións de glucosa das 68 mulleres diabéticas. As variables son tratadas como vectores, así que podemos empregar as ferramentas dispoñibles para estes obxectos. A xeito de resumo:

- `glu[1:10]` extrae as concentracións de glucosa das 10 primeiras mulleres que aparecen no coxunto de datos.
- `glu[type=="Yes"]` selecciona as concentracións de glusosa das mulleres diabéticas. Se queremos empregar a negación escribimos `glu[type!="Yes"]`, que extrae as concentracións de glucosa das mulleres non diabéticas (neste caso coincide con `glu[type=="No"]`, pero cando hai máis de dous niveis pode resultar útil empregar a negación).
- `glu[bmi>25]` extrae as concentracións de glucosa das mulleres que teñen valores da variable *bmi* por riba de 25. Pódese facer o mesmo co resto de operadores de comparación numérica `>=` (maior ou igual), `<` (menor) e `<=` (menor ou igual).
- `glu[glu>100]` extrae as concentracións de glucosa das mulleres que á súa vez teñen valores da concentración de glucosa por riba de 100.
- Os operadores lóxicos son `&` (e) e `|` (ou). Por exemplo, `glu[bmi>25 & bmi<=30]` extrae as concentracións de glucosa das mulleres que teñen valores da variable *bmi* maiores ca 25 e menores ou iguais ca 30.

Capítulo 2

Modelos de probabilidade en bioestatística

Contidos

2.1. Probabilidade

Cando observamos unha variable estatística sobre distintos individuos dunha poboación, sabemos que os valores observados cambian dun individuo a outro. Isto débese a que neste proceso de observación intervén o azar. Desde un punto de vista xeral, podemos asumir que o comportamento da poboación da cal observamos unha mostra está gobernado por un mecanismo abstracto que produce os resultados observados. Este procedemento mediante o que obtemos as observacións e no que intervén o azar é un **experimento aleatorio**. A **probabilidade** vincula este mecanismo teórico e o feito de que observemos un valor específico ou grupo de valores da nosa variable. Como sabemos, a probabilidade é un número entre 0 e 1 que mide a "frecuencia teórica" dun determinado suceso (por exemplo, que observemos un posible valor ou un conxunto de valores da variable de interese).

Exemplo. Consideremos un exemplo moi sinxelo. Supoñamos que unha empresa produce compoñentes electrónicas funxibles que se empregan en equipos para a obtención de imaxes por resonancia magnética. Por diversos desaxustes no proceso de fabricación, algunhas das pezas resultan defectuosas. A empresa sabe que, a longo prazo, a porcentaxe de pezas defectuosas é do 3 %. Se seleccionamos unha peza ao azar, temos un experimento aleatorio no que podemos definir dous sucesos: D="a peza seleccionada é defectuosa", que ocorre con probabilidade $P(D) = 0.03$, e $\bar{D}$="a peza seleccionada non é defectuosa", con probabilidade $P(\bar{D}) = 0.97$. Nótese que estes dous sucesos son complementarios entre si (por iso os denotamos por D e $\bar{D}$).

Quere isto dicir que se seleccionamos 100 unidades ao azar, necesariamente 3 serán defectuosas? Non! En realidade, as probabilidades dadas arriba son unha descrición teórica do experimento aleatorio, pero non implican que necesariamente en cada lote de 100 unidades haxa 3 defectuosas. Máis ben o número de pezas defectousas poderá ser quizais 3, ou 2, ou 4, ou...

Dúas preguntas sinxelas a modo de **exercicio**: hipoteticamente, cantas pezas defectuosas podería haber nun lote de 100 pezas? Serán todos eses posibles valores igual de probables? □

2.2. Variables aleatorias

Cando o resultado dun experimento aleatorio é un número, entón chamámoslle **variable aleatoria**. Para coñecer o comportamento probabilístico dunha variable aleatoria, debemos saber os posibles valores numéricos que pode tomar e as correspondentes probabilidades coas que ocorren estes valores. Chamámoslle **soporte** da vararíable aleatoria ao conxunto dos seus posibles valores e chamámoslle **distribución** da variable aleatoria á descrición detallada do seu comportamento probabilístico. Normalmente, denotaremos as variables aleatorias por letras maiúsculas: X, Y etc.

Exemplo. (cont.) No exemplo anterior, X ="número de pezas defectuosas nun lote de 100 unidades escollidas ao azar" é unha variable aleatoria. O seu soporte é o subconxunto de números naturais $\{0, 1, 2, \dots, 100\}$. Para dar a distribución da variable aleatoria X teriamos que detallar as probabilidades $P(X = k)$, para $k = 0, 1, 2, \dots, 100$. □

Distinguimos dous tipos de variables aleatorias: **discretas** e **continuas**.

2.2.1. Variables aleatorias discretas

Unha variable aleatoria é **discreta** cando os seus posibles valores forman un conxunto finito ou se poden identificar cos números naturais. En xeral, empréganse para **contar** o número de veces que ocorre algo.

Exemplos.

- O número de pezas defectuosas nun lote.
- O número de persoas que acuden a un servizo de urxencias nun día.
- O número de persoas do cadro de persoal dun hospital en situación de baixa médica.
- O número de ... □

Sexa X unha variable aleatoria discreta que ten soporte $\{x_1, x_2, \dots, x_n, \dots\}$. A distribución de X vén dada pola súa **masa de probabilidade**, que é a colección de probabilidades

$$p_i = P(X = x_i), \qquad \text{para } i = 1, 2, \dots$$

As probabilidades p_i suman 1: $\sum_i p_i = 1$.

Exemplo. (cont.) Consideremos a variable X ="número de pezas defectuosas nun lote de 100 unidades escollidas ao azar". Esta variable segue un modelo moi coñecido: a distribución **Binomial**, concretamente unha $Binomial(100, 0.03)$. Denotamos isto por $X \sim Binomial(100, 0.03)$. Como dixemos, o soporte de X é o conxunto $\{0, 1, 2, \dots, 100\}$. A masa de probabilidade de X é[1]

$$P(X = k) = \binom{100}{k} 0.03^k (1 - 0.03)^{100-k}, \qquad \text{para } k = 0, 1, 2, \dots, 100.$$

Isto describe completamente o comportamento de X. En particular, podemos calcular, por exemplo, a probabilidade de que nun lote haxa 4 pezas defectuosas:

$$P(X = 4) = \binom{100}{4} 0.03^4 (1 - 0.03)^{100-4} = 0.1706,$$

é dicir, o 17.06 % de lotes de 100 unidades conteñen 4 unidades defectuosas. En R podemos obter esta probabilidade mediante a función `dbinom()`:

```
> dbinom(4,size=100,prob=0.03)

[1] 0.1706056
```

Supoñamos que a empresa se compromete e reembolsar o custo do lote se este contén 8 ou máis pezas defectuosas. A probabilidade de ter que reembolsar un lote é polo tanto $P(X \geq 8) = 1 - P(X \leq 7)$. En R podemos calcular facilmente esta probabilidade coa axuda da función `pbinom()`, que calcula as probabilidades acumuladas:

[1] Recordemos que $\binom{n}{k}$ é o número combinatorio n sobre r, é dicir, $\binom{n}{k} = \frac{n!}{k!(n-k)!}$.

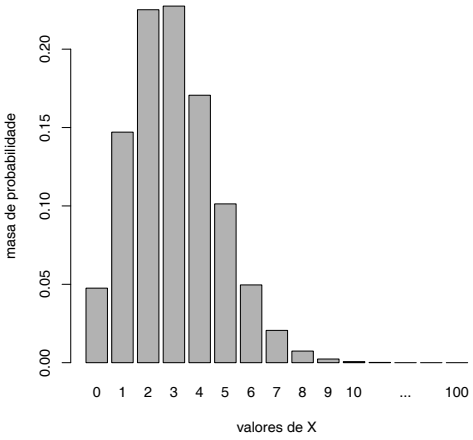


Figura 2.1: Masa de probabilidade dunha variable aleatoria con distribución $Binomial(100, 0.03)$.

```
> 1-pbinom(7,size=100,prob=0.03)

[1] 0.01062381
```

Polo tanto, a probalidade de ter que reembolsar o custo dun lote é do 1.06 %.

Na Figura 2.1 amósase a masa de probabilidade de X. Nótese que, aínda que o soporte da variable son todos os posibles números naturais entre 0 e 100, en realidade as probabilidades concéntranse nuns poucos valores (esencialmente, do 0 ao 9). □

Características dunha variable aleatoria discreta

Como características fundamentais dunha variable aleatoria discreta temos a súa esperanza e a súa varianza. A **media** ou **esperanza matemática** da variable aleatoria discreta X é

$$\mathrm{E}(X) = \sum_i x_i p_i.$$

Algunhas propiedades da media:

- Sexa X unha variable aleatoria discreta, e sexan a e b dúas constantes. Entón

$$\mathrm{E}(aX + b) = a\mathrm{E}(X) + b.$$

- Sexan X e Y dúas variables aleatorias discretas. Entón

$$\mathrm{E}(X + Y) = \mathrm{E}(X) + \mathrm{E}(Y).$$

- Supoñamos que transformamos unha variable aleatoria discreta X por unha función h para obter unha nova variable aleatoria $T = h(X)$. Entón a media da variable transformada pode calcularse como

$$\mathrm{E}(T) = \mathrm{E}[h(X)] = \sum_i h(x_i)p_i.$$

A **varianza** de X é

$$\mathrm{Var}(X) = \mathrm{E}[(X - \mathrm{E}(X))^2] = \sum_i (x_i - \mathrm{E}(X))^2 p_i.$$

A **desviación estándar** (ou desviación típica) de X é

$$\mathrm{sd}(X) = \sqrt{\mathrm{Var}(X)}.$$

Se X é unha variable aleatoria discreta e a e b son constantes, entón a varianza e a desviación estándar cumpren as seguintes propiedades:

- $\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X)$.
- $\mathrm{sd}(aX + b) = |a| \mathrm{sd}(X)$.

Exemplo. (cont.) A media de $X \sim Binomial(100, 0.03)$ é

$$\mathrm{E}(X) = 100 \cdot 0.03 = 3$$

e a desviación estándar é

$$\mathrm{sd}(X) = \sqrt{100 \cdot 0.03 \cdot (1 - 0.03)} = 1.706.$$

En xeral, **se $X \sim Binomial(n, p)$, entón a media e a desviación estándar de X son**

$$\mathrm{E}(X) = np \qquad \textbf{e} \qquad \mathrm{sd}(X) = \sqrt{np(1-p)},$$

respectivamente. □

Algúns modelos importantes de variables aleatorias discretas son os seguintes: **Bernoulli**, **Binomial**, **Hiperxeométrica**, **Xeométrica**, **Poisson** etc.

Exemplo. Supoñamos agora que un hospital adquire un lote que contén 5 unidades defectuosas. Para reparar a máquina de resonancia precísanse 3 unidades. Para isto, cóllense as 3 unidades ao azar do lote (obviamente, sen reemprazamento). Cal é a probabilidade de que haxa algunha defectuosa entre as 3 seleccionadas?

Neste caso, o máis doado é pensar na probabilidade do suceso complementario, é dicir, calcular a probabilidade de que entre as 3 pezas seleccionadas non haxa ningunha unidade defectuosa:

$$\begin{aligned} & P(\text{polo menos 1 unidade defectuosa entre as 3 seleccionadas}) \\ = \; & 1 - P(\text{ningunha unidade defectuosa entre as 3 seleccionadas}) \\ = \; & 1 - \frac{95}{100} \cdot \frac{94}{99} \cdot \frac{93}{98} = 0.144. \end{aligned}$$

Esta probabilidade está relacionada coa masa de probabilidade da variable aleatoria **Hiperxeométrica**. De que xeito? (**exercicio**). □

Exercicio 2.1. *A distribución de* **Poisson** *é un modelo de variable aleatoria discreta amplamente empregado. Pode utilizarse para contar o número de veces que ocorre un evento por unidade de tempo. Depende dun único parámetro, λ, que é o número medio de ocorrencias por unidade de tempo. A variable aleatoria X segue unha distribución de Poisson de media λ, que denotamos por $X \sim Poisson(\lambda)$, se a súa masa de probabilidade é*

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}, \qquad \text{para } k = 0, 1, 2, ...$$

Supoñamos que número de persoas que chegan a un servizo de urxencias nunha hora é unha variable aleatoria de **Poisson** *de media 2 persoas por hora.*

(a) Cal é a probabilidade de que na próxima hora non chegue ningunha persoa?

(b) Cal é a probabilidade de que nunha hora cheguen máis de 4 persoas?

(c) Que distribución ten a variable aleatoria que conta o número de persoas que chegan nun turno de 8 horas?

(d) Cal é a probabilidade de que nun turno de 8 horas cheguen máis de 32 persoas? Compara esta probabilidade coa do apartado (b). Algunha posible explicación para esta diferenza entre as probabilidades? (Pista: compara as desviacións estándar das variables aleatorias correspondentes).

En R, as funcións `dpois()` *e* `ppois()` *permiten calcular probabilidades asociadas á distribución de Poisson. Investiga o seu uso.*

Exercicio 2.2. ***Relación entre a Binomial e a Poisson.*** *Un estudo científico demostra que 1 de cada 200 persoas portan unha modificación xenética que pode causar cancro de colon hereditario. Nunha mostra de 1000 individuos, o número de persoas portadoras deste xen é unha variable aleatoria con distribución Binomial$(1000, 1/200)$. Neste caso, esta variable aleatoria tamén se pode modelizar mediante unha distribución de Poisson de media $1000/200$.*

(a) Calcula en R as masas de probabilidade destas dúas distribucións e compáraas.

(b) Calcula a probabilidade de que na mostra haxa 7 ou máis persoas portadoras do xen. Emprega a Binomial e a Poisson e compara os resultados.

(c) Investiga as condicións que deben cumprir os parámetros da Binomial e da Poisson para que se poida facer esta identificación entre as dúas distribucións.

2.2.2. Variables aleatorias continuas

Unha variable aleatoria é **continua** cando os seus posibles valores forman un intervalo (posiblemente infinito) dos números reais. Empréganse para **medir** magnitudes.

Exemplos.

- A concentración de glucosa no sangue dunha persoa.
- A presión arterial dunha persoa.
- O tempo que tarda unha persoa en facer un test psicotécnico.
- O índice de masa corporal dunha persoa.
- O contido dun vial dunha vacina.
- O tempo que tarda unha médica en atender a un paciente.
- ... □

Función de distribución e función de densidade

Unha variable aleatoria continua queda caracterizada matematicamente pola súa función de distribución ou pola súa función de densidade.

A **función de distribución** dunha variable aleatoria X é a función que lle asigna a cada número real x a probabilidade de que o valor observado da variable aleatoria sexa como moito x , é dicir,

$$F(x) = P(X \leq x).$$

Algunhas propiedades da función de distribución:

- $F(x)$ é unha probabilidade, polo tanto $0 \leq F(x) \leq 1$ para todo x.
- F é non decrecente, é dicir, se x_1 e x_2 son tales que $x_1 < x_2$, entón $F(x_1) \leq F(x_2)$.
- $\lim_{x\to-\infty} F(x) = 0$ e $\lim_{x\to+\infty} F(x) = 1$.
- En realidade a función de distribución está definida tanto para variables aleatorias discretas como continuas:
  - Se X é unha variable aleatoria discreta, entón a súa función de distribución é unha función constante por pedazos con discontinuidades nos valores que forman o soporte da variable. Véxase a Figura 2.2-(a).
  - Se X é unha variable aleatoria continua, entón a súa función de distribución F é unha función continua. Véxase a Figura 2.2-(b).

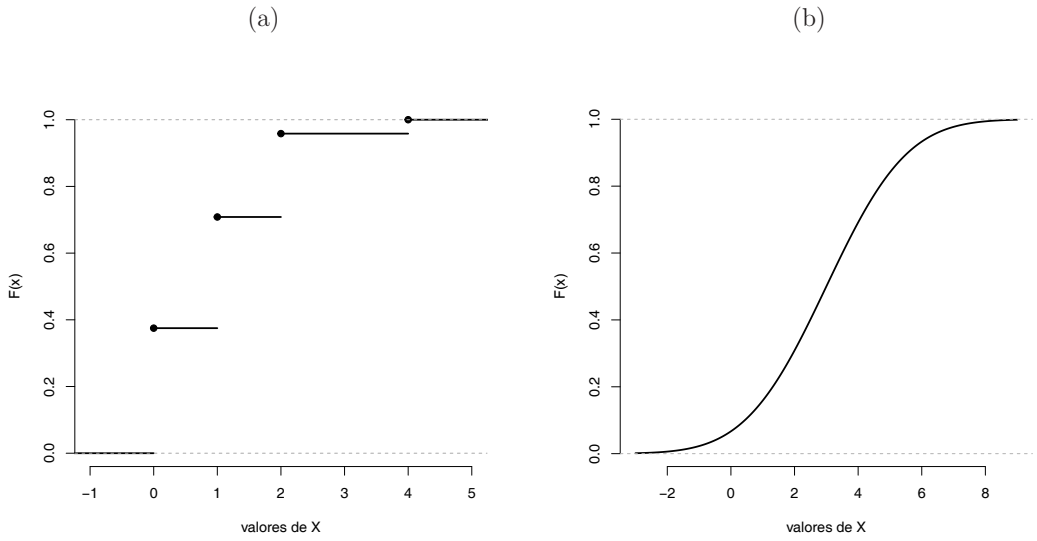


Figura 2.2: Exemplos de funcións de distribución. (a) Función de distribución dunha variable aleatoria discreta con soporte $\{0, 1, 2, 4\}$. (b) Función de distribución dunha variable aleatoria continua.

Resulta que se X é unha variable aleatoria continua, entón a súa función de distribución, ademais de ser continua, tamén é unha función derivable. A **función de densidade** de X é a derivada da función de distribución

$$f(x) = F'(x),$$

ou equivalentemente, a función de distribución é a integral da función de densidade

$$F(x) = \int_{-\infty}^{x} f(t)dt.$$

Algunhas propiedades da función de densidade:

- $f(x) \geq 0$.
- $\int_{-\infty}^{\infty} f(t)dt = 1$.
- $P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a) = \int_a^b f(t)dt$.
- Se X é unha variable aleatoria continua, entón a probabilidade de que a variable tome un valor concreto x é cero: $P(X = x) = 0$.
- Polo tanto, se X é unha variable aleatoria continua entón as seguintes probabilidades coinciden e pódense calcular como unha integral da función de densidade

$$P(a < X < b) = P(a < X \leq b) = P(a \leq X < b) = P(a \leq X \leq b) = F(b) - F(a) = \int_a^b f(t)dt.$$

É dicir, no caso das variables aleatorias continuas só ten sentido calcular probabilidades sobre intervalos. A probabilidade de que a variable aleatoria tome valores nun intervalo pódese calcular como a diferenza das imaxes da función de distribución nos extremos do intervalo ou ben como a correspondente integral da función de densidade. Tal e como se amosa na Figura 2.3, a probabilidade $P(a \leq X \leq b)$ é a área debaixo do gráfico da función de densidade no intervalo $[a, b]$.

A función de densidade tamén se pode interpretar como o límite dos histogramas. Supoñamos que temos un número moi grande de observacións dunha variable aleatoria continua de tal xeito que podemos construír histogramas con intervalos de amplitude moi pequena. Se facemos que a amplitude dos intervalos se faga cada vez máis pequena, entón o histograma parécese cada vez máis á función de densidade da variable aleatoria, tal e como se amosa na Figura 2.4. Polo tanto, podemos interpretar a función de densidade de forma similar a como interpretabamos o histograma: se a densidade é alta quere dicir que é moi probable observar valores de X nesa zona do soporte; en cambio se a densidade é baixa entón hai menos probabilidade de observar valores nesa zona do soporte.

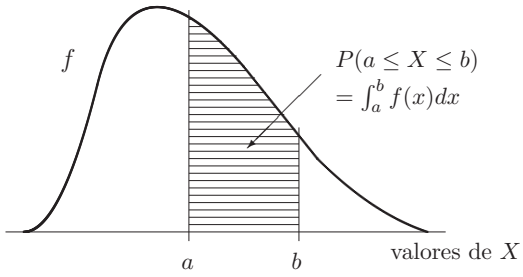


Figura 2.3: As probabilidades asociadas a unha variable aleatoria continua son integrais da función de densidade.

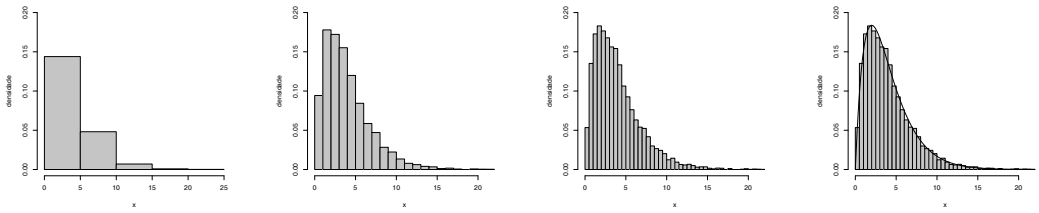


Figura 2.4: Relación entre o histograma e a función de densidade.

Exercicio 2.3. *Na Figura 2.5 aparecen distintos tipos de funcións de densidade e as súas correspondentes funcións de distribución. Para cada unha delas identifica o soporte e as zonas dese soporte onde hai maior probabilidade de observar valores das correspondentes variables aleatorias.*

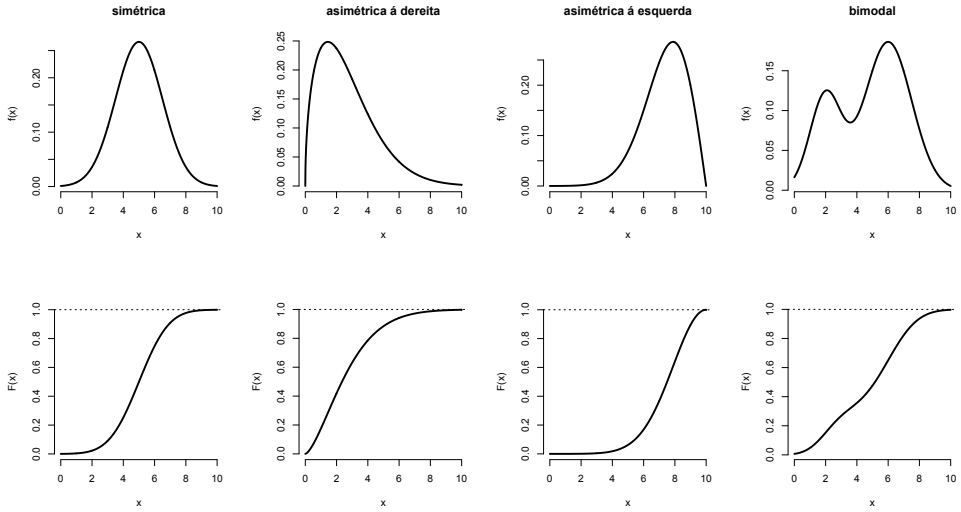


Figura 2.5: Distintos tipos de funcións de densidade (arriba) e as súas correspondentes funcións de distribución (abaixo).

Exemplo. A variable aleatoria **Exponencial** pode empregarse para modelizar tempos de duración de compoñentes electrónicas sinxelas. Supoñamos que as compoñentes do exemplo anterior teñen un tempo de vida útil (en meses) que se comporta como unha variable aleatoria Y que ten función de densidade

$$g(y) = \begin{cases} 0.4\, e^{-0.4y} & \text{se } y \geq 0, \\ 0 & \text{se } y < 0. \end{cases}$$

Neste caso, a variable aleatoria Y recibe o nome de **Exponencial** de parámetro 0.4, ou $Y \sim Exponencial(0.4)$. A función de distribución de Y é

$$G(y) = \int_{-\infty}^{y} g(t)dt = \begin{cases} 0 & \text{se } y < 0, \\ \int_0^y 0.4e^{-0.4t}dt = 1 - e^{-0.4y} & \text{se } y \geq 0. \end{cases}$$

Cal é a probabilidade de que unha peza dure máis de 2.5 meses?

$$P(Y > 2.5) = 1 - P(Y \leq 2.5) = 1 - F(2.5) = 1 - (1 - e^{-0.4 \cdot 2.5}) = e^{-1} = 0.3679.$$

En R, a función `pexp()` calcula a función de distribución dunha variable aleatoria Exponencial. Para obter a probabilidade anterior, escribimos:

```
> 1-pexp(2.5,rate=0.4)

[1] 0.3678794
```

□

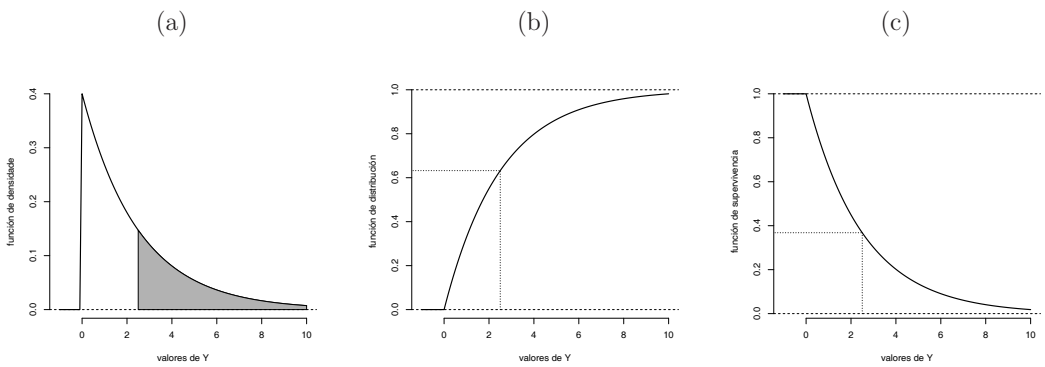


Figura 2.6: Función de densidade (a), de distribución (b) e de superviviencia (c) dunha variable aleatoria con distribución $Exponencial(0.4)$. A área sombreada en (a) é a probabilidade de que a variable supere o valor 2.5.

Cando a variable aleatoria que se está estudando é un tempo (por exemplo, o tempo de duración dunha compoñente electrónica, o tempo que tarde unha persoa en recuperarse dunha determinada enfermidade etc.), entón é moi común traballar coa función de supervivencia. A **función de supervivencia** dá a probabilidade de que a variable aleatoria X supere un determinado valor x, é dicir, é a función

$$S(x) = P(X > x).$$

Obviamente, hai unha relación inmediata entre a función de distribución e a función de supervivencia: se F é a función de distribución de X, entón a función de supervivencia de X é $S(x) = 1 - F(x)$.

A Figura 2.6 amosa as funcións de densidade, de distribución e de supervivencia dunha variable aleatoria con distribución $Exponencial(0.4)$.

Características dunha variable aleatoria continua

Sexa X unha variable aleatoria continua con función de distribución F e función de densidade f. A **media** ou **esperanza matemática** de X é a integral

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f(x) dx.$$

En moitas ocasións, denotaremos a media de X por μ_X, ou simplemente por μ.

Algunhas propiedades da media:

- Sexa X unha variable aleatoria continua e sexan a e b dúas constantes. Entón

$$\mathrm{E}(aX + b) = a\mathrm{E}(X) + b.$$

- Sexan X e Y dúas variables aleatorias continuas. Entón

$$\mathrm{E}(X+Y) = \mathrm{E}(X) + \mathrm{E}(Y).$$

- Sexa X unha variable aleatoria continua con función de densidade f. Supoñamos que transformamos X por unha función h para obter unha nova variable aleatoria $T = h(X)$. Entón a media da variable transformada pode calcularse como

$$\mathrm{E}(T) = \mathrm{E}[h(X)] = \int_{-\infty}^{\infty} h(x) f(x) dx.$$

- Se a función de densidade de X é simétrica respecto dun valor a, é dicir, $f(a+c) = f(a-c)$ para todo $c > 0$, entón $\mathrm{E}(X) = a$ (**exercicio**).

Exemplo. Para obter a media da variable aletoria Y con distribución $Exponencial(0.4)$ temos que calcular a integral

$$\mathrm{E}(Y) = \int_{-\infty}^{\infty} y\, g(y) dy = \int_{0}^{\infty} y\, 0.4\, e^{-0.4y} dy.$$

Nótese que esta é unha integral impropia e polo tanto debe resolverse como un límite:

$$\int_{0}^{\infty} y\, 0.4\, e^{-0.4y} dy = \lim_{c \to +\infty} \int_{0}^{c} y\, 0.4\, e^{-0.4y} dy = \cdots \textbf{(exercicio)} \cdots = \frac{1}{0.4} = 2.5 \text{ meses.}$$

Podemos xeralizar este resultado sobre a media dunha Exponencial substituíndo 0.4 por calquera outro valor positivo: **se Y é unha variable aleatoria con distribución $Exponencial(\lambda)$, entón** $\mathrm{E}(Y) = 1/\lambda$. □

A **varianza** dunha variable aleatoria continua X é

$$\mathrm{Var}(X) = \mathrm{E}[(X - \mathrm{E}(X))^2] = \int_{-\infty}^{\infty} (x - \mathrm{E}(X))^2 f(x) dx.$$

En moitas ocasións, denotaremos a varianza de X por σ_X^2, ou simplemente σ^2.

A **desviación estándar** de X é a raíz cadrada da varianza

$$\mathrm{sd}(X) = +\sqrt{\mathrm{Var}(X)},$$

que denotaremos por σ_X ou simplemente por σ.

Algunhas propiedades da varianza (estas propiedades son válidas tanto para variables aleatorias discretas como continuas):

- Sexa X unha variable aleatoria e sexan a e b dúas constantes. Entón

$$\mathrm{Var}(aX + b) = a^2\, \mathrm{Var}(X) \qquad \text{e} \qquad \mathrm{sd}(aX + b) = |a|\, \mathrm{sd}(X).$$

- Sexan X e Y dúas variables aleatorias. Entón

$$\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) - 2\,\text{Cov}(X,Y),$$

onde

$$\text{Cov}(X,Y) = E\left[(X - E(X))(Y - E(Y))\right]$$

é a **covarianza** entre X e Y.

Dúas variables aleatorias X e Y son **independentes** se, para calquera par de valores x e y, se cumpre que

$$P(X \leq x \text{ e } Y \leq y) = P(X \leq x)P(Y \leq y).$$

Intuitivamente, que dúas variables aleatorias sexan independentes significa que non se pode establecer ningunha relación entre elas, é dicir, unha non inflúe na outra, e viceversa. En cambio, se hai algún tipo de relación entre elas, entón serán variables dependentes.

Exemplo. No exemplo das compoñentes electrónicas, é bastante razoable pensar que a variable aletoria "vida útil dunha compoñente" será independente da variable aleatoria "cantidade de chuvia recollida en Vigo o día que se fabricou a compoñente". En cambio, parece razoable pensar que as variables aleatorias "vida útil dunha compoñente" e "voltaxe eléctrica á que está sometida a compoñente" poidan presentar algún tipo de dependencia. □

Dúas variables aleatorias son **incorreladas** se $\text{Cov}(X,Y) = 0$. Independencia implica incorrelación: **se X e Y son independendes, entón X e Y son incorreladas**. Polo tanto, **se X e Y son independentes** (e polo tanto incorreladas), **entón** $\text{Var}(X+Y) = \text{Var}(X)+\text{Var}(Y)$.

Exemplo. Para calcular a varianza da variable aleatoria Y con distribuión $Exponencial(0.4)$ temos que resolver a integral impropia

$$\begin{aligned}\text{Var}(Y) &= \int_{-\infty}^{\infty} (y - \text{E}(Y))^2\, g(y)dy = \int_0^{\infty} \left(y - \frac{1}{0.4}\right)^2 0.4\, e^{-0.4y} dy \\ &= \cdots \textbf{(exercicio)} \cdots = \frac{1}{0.4^2} = 6.25 \text{ meses}^2.\end{aligned}$$

Polo tanto, a desviación estándar de Y é $\text{sd}(Y) = 1/0.4 = 2.5$ meses. En xeral, **se Y ten distribución $Exponencial(\lambda)$, entón** $\text{Var}(Y) = 1/\lambda^2$ **e** $\text{sd}(Y) = 1/\lambda$. □

A **moda** dunha variable aleatoria continua é o valor (ou valores) que maximiza a súa función de densidade.

A **mediana** dunha variable aletoria continua X é o valor divide o seu soporte en dous intervalos de igual probabilidade. Polo tanto, a mediana de X é o valor $M = \text{Mediana}(X)$ tal que

$$F(M) = \int_{-\infty}^{M} f(x)dx = 0.5.$$

Os cuantís son xeralizacións da mediana. O **cuantil de orde p** (tamén chamado **percentil $100p\,\%$**) de X é o valor x_p tal que

$$F(x_p) = \int_{-\infty}^{x_p} f(x)dx = p.$$

p	cuantil de orde p
0.10	0.2634
0.25	0.7192
0.50	1.7329
0.75	3.4657
0.80	4.0236
0.90	5.7565

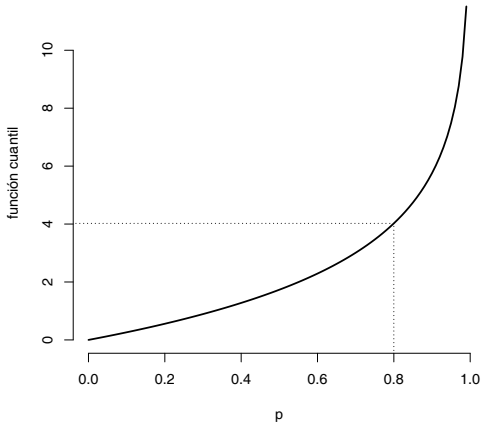


Figura 2.7: Esquerda: algúns cuantís dunha variable aleatoria con distribución *Exponencial*(0.4). Dereita: función cuantil da *Exponencial*(0.4). Aparece sinalado o cuantil de orde 0.80.

Os cuantís de orde 0.25, 0.50 e 0.75 chámanse **primeiro cuartil** (que denotamos por $x_{0.25}$), **segundo cuartil** ($x_{0.50}$, que coincide coa mediana) e **terceiro cuartil** ($x_{0.75}$), respectivamente. O **rango intercuartílico** dunha variable aleatoria é $x_{0.75} - x_{0.25}$.

A función que a cada probabilidade p lle asocia o cuantil de orde p chámase **función cuantil** de X. A función cuantil é a inversa da función de distribución. Formalmente, a función cuantil asociada á función de distribución F defínese como

$$F^{-1}(p) = \text{ínf}\{x \in \mathbb{R} : p \leq F(x)\}, \qquad \text{para } 0 < p < 1.$$

Con esta definición, a mediana de X é $F^{-1}(0.5)$, e os cuartís de X son $F^{-1}(0.25)$, $F^{-1}(0.50)$ e $F^{-1}(0.75)$.

Exemplo. (cont.) A función cuantil da variable aleatoria Y con distribución *Exponencial*(0.4) é

$$G^{-1}(p) = -\frac{\log(1-p)}{0.4} \qquad \text{para } 0 < p < 1. \qquad \textbf{(exercicio)}$$

A mediana de Y é $G^{-1}(0.5) = -\log(1-0.5)/0.4 = 1.7329$ meses. Este valor quere dicir que o 50 % das compoñentes duran menos de 1.7329 meses e o outro 50 % superan esa duración. Compárese este valor co da media, que era 2.5 meses. Esta diferenza entre a media e a mediana débese ao carácter asimétrico da función de densidade.

Na parte esquerda da Figura 2.7 recóllense algúns cuantís de Y. Por exemplo, o feito de que o cuantil de orde 0.90 sexa 5.7565 quere dicir que o 90 % das compoñentes fallan antes dos 5.7565 meses, ou equivalentemente, o 10 % das compoñentes superan esta duración. A parte da dereita da Figura 2.7 amosa o gráfico da función cuantil de Y.

A función `qexp()` de R calcula os cuantís dunha variable aleatoria con distribución Exponencial. Por exemplo, para obter a mediana escribimos

```
> qexp(0.5,rate=0.4)

[1] 1.732868
```

e para obter os cuantís de orde 0.70, 0.80 e 0.90 escribimos

```
> qexp(c(0.70,0.80,0.90),rate=0.4)

[1] 3.009932 4.023595 5.756463
```

□

Exercicio 2.4. *Unha barra metálica de* 30 *cm de lonxitude suxéitase polos dous extremos e aplícaselle unha forza ata que rompe. Consideremos a variable aletoria X que mide a distancia entre o extremo esquerdo e o punto de rotura. A función de densidade de X é*

$$f(x) = \begin{cases} \frac{1}{150}x\left(1-\frac{x}{30}\right) & \text{se } 0 \leq x \leq 30, \\ 0 & \text{noutro caso.} \end{cases}$$

(a) Calcula a función de distribución e grafica as funcións de densidade e de distribución. Identifica o soporte de X e a zona de maior probabilidade.

(b) Calcula as seguintes probabilidades: $P(X \leq 10)$, $P(10 \leq X \leq 20)$ e $P(X > 20)$.

(c) Calcula $E(X)$ e $\text{VAR}(X)$.

(d) Calcula os cuantís de orde 0.10 *e* 0.90 *de X. Que relación existe entre eles?*

2.2.3. Modelos de variables aleatorias

Algunhas variables aleatorias teñen distribucións moi coñecidas e estudadas. Un **modelo probabilístico** é unha representación teórica do comportamento da variable aleatoria. Noutras palabras, un modelo probabilístico é unha descrición matemática completa da distribución da variable. Xa mencionamos algúns:

- Entre as variables aleatorias discretas, xa falamos da distribución Binomial, a distribución de Poisson ou a distribución Hiperxeométrica.
- En canto ás variables aleatorias continuas, xa falamos da distribución Exponencial, que, xunto coa Weibull, forma parte do grupo de variables aleatorias que serven para modelizar tempos. Outras variables continuas importantes son a distribución Normal ou a distribución Uniforme.
- Hai algunhas variables aleatorias que se empregan nas distintas metodoloxías da inferencia estatística. Algunhas delas son a Normal Estándar, a distribución t de Student, a distribución Chi-cadrado de Pearson ou a distribución F de Snedecor.

Exercicio 2.5. ***Mesturas de variables aletorias.*** *Un servizo de atención ao cliente ten dous postos, que identificamos con A e B. O tempo (en minutos) que tarda o posto A en atender a un cliente, que denotamos por T_A, é unha variable aleatoria Exponencial(0.5), é dicir, unha Exponencial de media 2. En cambio, o tempo que tarda o posto B, que denotamos por T_B é unha Exponencial(0.25). Supoñamos que os clientes son asignados aleatoriamente a un dos dous postos. Denotemos por p a probabilidade de que un cliente sexa atendido polo posto A e polo tanto $1-p$ é a probabilidade de que sexa atendido no posto B. Sexa T a variable aleatoria que mide o tempo que tarda en ser atendido un novo cliente. A variable aleatoria T compórtase como T_A con probabilidade p e como T_B con probabilidade $1-p$. Dise entón que T é unha* **mestura** *das variables aleatorias T_A e T_B con probabilidades p e $1-p$.*
Para calcular a función de distribución de T podemos facer o seguinte razoamento con axuda do **Teorema das Probabilidades Totais**. *Definamos os sucesos $A =$"o cliente é atendido no posto A" e $B =$"o cliente é atendido no posto B". Obsérvese que A e B son sucesos complementarios con probabilidades p e $1-p$, respectivamente. Polo tanto*

$$F(t) = P(T \leq t) = P(T \leq t \mid A)P(A) + P(T \leq t \mid B)P(B).$$

Agora abonda ter en conta que a probabilidade condicionada $P(T \leq t \mid A)$ é precisamente a función de distribución de $T_A \sim$ Exponencial(0.5) e, analogamente, $P(T \leq t \mid B)$ é a función de distribución de $T_B \sim$ Exponencial(0.25). Polo tanto, a función de distribución de T será

$$\begin{aligned} F(t) = P(T \leq t) &= P(T_A \leq t)p + P(T_B \leq t)(1-p) \\ &= (1-e^{-0.5t})p + (1-e^{-0.25t})(1-p), \end{aligned}$$

para $t \geq 0$ e $F(t) = 0$ para $t < 0$.

(a) Calcula a función de densidade de T.

(b) Grafica as funcións de distribución e de densidade de T para $p = 0.25, 0.50$ e 0.75.

(c) Calcula a media de T en función de p. Particulariza o valor da media para $p = 0.25, 0.50$ e 0.75.

(d) Deduce unha fórmula que relacione a media de T coas medias de T_A e T_B.

2.3. A distribución Normal

A **distribución Normal** é o modelo máis importante entre as variables aleatorias continuas. Pódese empregar en moitas situacións prácticas, xa que moitas variables cuantitativas teñen distribucións que se axustan moi ben á Normal. Podemos citar como exemplos as medidas antropométricas (altura, peso), erros de medida en experimentos científicos, as puntuacións en tests psicolóxicos etc.

A Normal ten as seguintes características xenéricas:

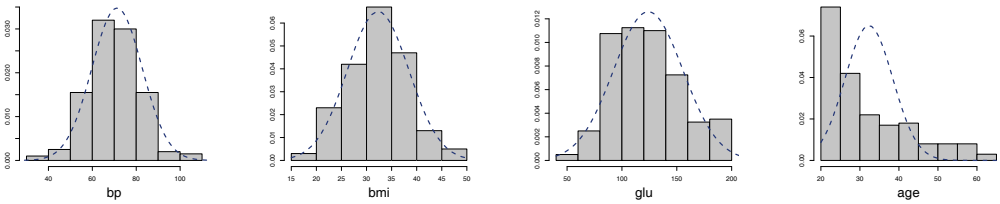


Figura 2.8: Conxunto de datos DIABETES. Histogramas das variables *bp*, *bmi*, *glu* e *age*. A curva sobreimposta é a densidade dunha variable aleatoria Normal con media e desviación estándar estimadas a partir das correspondentes versións mostrais.

- Os posibles valores da variable distribúense de forma simétrica arredor dun valor central.
- É máis probable que se observen valores próximos a ese valor central ca a valores que están alonxados.
- O histograma construído con observacións dunha variable aleatoria Normal ten forma de campá (chamada campá de Gauss).

A Figura 2.8 contén os histogramas das variables *bp*, *bmi*, *glu* e *age* do conxunto de datos DIABETES. Claramente, a variable *age* non ten distribución Normal, xa que o histograma é completamente asimétrico. O histograma da variable *glu* tamén presenta unha certa asimetría, polo tanto a distribución Normal quizais tampouco resulte axeitada neste caso. En cambio, no caso das variables *bmi* e *bp* parece moito máis razoable supoñer que seguen unha distribución Normal.

As características anteriores só son unha descrición intuitiva da distribución Normal. A especificación matemática da Normal vén dada pola súa función de densidade. Unha variable aleatoria X segue unha distribución **Normal con media μ e desviación estándar σ**, que denotamos por $X \sim N(\mu, \sigma)$, se a súa función de densidade é

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

Como vemos na fórmula anterior, para especificar a densidade da Normal necesitamos dar dous parámetros: μ e σ. A Figura 2.9 amosa funcións de densidade e de distribución de variables aleatorias con distribución Normal.

- Como vemos na Figura 2.9, o valor de μ é o punto de simetría da densidade da Normal, polo tanto μ é a media da variable aleatoria. O cambio en media simplemente se traduce nunha translación da densidade, tal e como se pode ver coas densidades $N(50, 5)$ e $N(80, 5)$.
- O parámetro σ é a desviación estándar da variable. Un cambio na desviación estándar supón máis ou menos dispersión dos posibles valores arredor de μ. Compárense os gráficos das densidades $N(50, 5)$, $N(50, 10)$ e $N(50, 15)$.

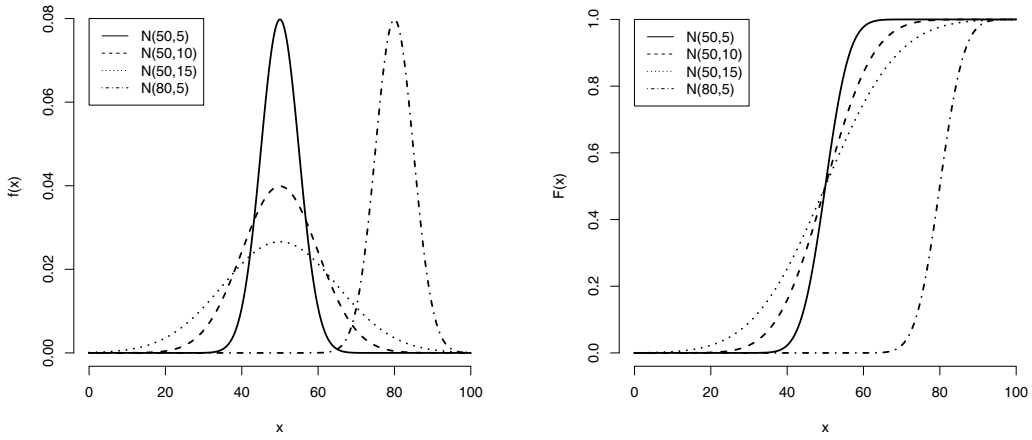


Figura 2.9: Funcións de densidade (esquerda) e funcións de distribución (dereita) de variables aleatorias con distribución Normal.

Para calcular probabilidades asociadas á Normal temos que calcular integrais sobre a función de densidade, tal e como se indica na Figura 2.10. A primitiva da función de densidade da Normal non ten unha forma analítica explícita, e polo tanto tampouco dispoñemos dunha fórmula explícita para a función de distribución. Para calcular as probabilidades (ou integrais) recórrese á integración numérica.

Exemplo. Á vista dos histograma da Figura 2.8 parece razoable supoñer que a presión arterial diastólica se comporta como unha variable aleatoria Normal. Supoñamos que a súa media e desviación estándar son 70 e 11, respectivamente. Cal é a probabilidade de que a presión arterial supere 80 mm Hg? Cal é a probabilidade de que a presión arterial estea entre 60 e 80?

Sexa X a variable aleatoria "presión arterial". A distribución de X é $N(70, 11)$. Temos que calcular $P(X \geq 80)$. A función `pnorm()` de R dá os valores da función de distribución da Normal aproximados numericamente. Entón só temos que escribir

```
> 1-pnorm(80,mean=70,sd=11)
```

```
[1] 0.1816511
```

Polo tanto, o 18.17 % das persoas presentan unha presión maior de 80 mm Hg. Para calcular $P(60 \leq X \leq 80)$ escribimos

```
> pnorm(80,mean=70,sd=11)-pnorm(60,mean=70,sd=11)
```

```
[1] 0.6366979
```

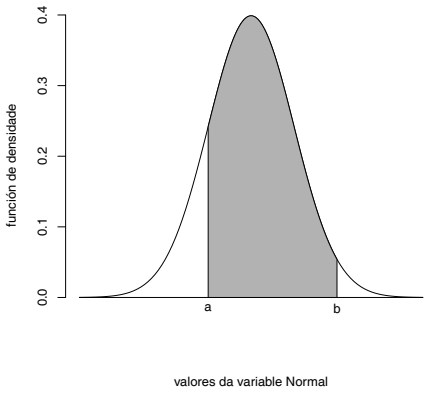


Figura 2.10: Densidade dunha variable aleatoria Normal. A área sombreada é $P(a \leq X \leq b)$. A área total debaixo da curva é 1.

Supoñamos agora que queremos coñecer a partir de que valor se sitúa o 5 % dos individuos que presentan valores da presión máis altos. Para iso teremos que calcular o cuantil de orde 95 % de X. En R empregamos a función `qnorm()`:

```
> qnorm(0.95,mean=70,sd=11)

[1] 88.09339
```

□

A distribución Normal verifica a seguinte propiedade importante que se denomina estandarización da Normal:

Estandarización da Normal

- **Se X é unha variable aleatoria con distribución $N(\mu, \sigma)$, entón a versión estandarizada $Z = \frac{X-\mu}{\sigma}$ é unha variable aleatoria $N(0, 1)$.**

  Isto quere dicir que se $X \sim N(\mu, \sigma)$ e $Z \sim N(0, 1)$, entón

$$P(X \leq x) = P\left(Z \leq \frac{x-\mu}{\sigma}\right).$$

- **Analogamente, se Z é $N(0, 1)$ entón $X = \mu + \sigma Z$ é $N(\mu, \sigma)$.**

  Isto quere dicir que se z_p é o cuantil de orde p da $N(0, 1)$, entón $x_p = \mu + \sigma \cdot z_p$ é o cuantil de orde p da $N(\mu, \sigma)$.

A partir da propiedade de estandarizadión da Normal dedúcese que as probabilidades ou os cuantís de calquera distribución Normal se poden expresar en termos dos da $N(0,1)$. A $N(0,1)$ chámase **Normal Estándar** e denótase por Z.

Exemplo. (cont.) Se X é $N(70,11)$, entón $Z = \frac{X-70}{11}$ é $N(0,1)$. Polo tanto

$$P(X \geq 80) = P\left(\frac{X-70}{11} \geq \frac{80-70}{11}\right) = P(Z \geq 0.909).$$

Comparemos en R:

```
> 1-pnorm(80,mean=70,sd=11)

[1] 0.1816511

> 1-pnorm((80-70)/11,mean=0,sd=1)

[1] 0.1816511
```

Analogamente, se $z_{0.95}$ é o cuantil de orde 0.95 da $N(0,1)$, entón $70 + 11 \cdot z_{0.95}$ é o cuantil de orde 0.95 da $N(70,11)$. De novo podemos verificalo en R:

```
> qnorm(0.95,mean=70,sd=11)

[1] 88.09339

> 70+11*qnorm(0.95,mean=0,sd=1)

[1] 88.09339
```

En R cando traballamos coa $N(0,1)$ non é necesario especificar os valores da media (argumento `mean`) e da desviación estándar (argumento `sd`) nas fucións `pnorm()` e `qnorm()`. Podemos escribir simplemente

```
> 1-pnorm((80-70)/11)

[1] 0.1816511

> 70+11*qnorm(0.95)

[1] 88.09339
```

Vexamos agora como se comportan as probabilidades asociadas a unha variable aleatoria Normal. Sexa X unha $N(\mu, \sigma)$ e sexa $k > 0$ un número positivo. Calcularemos as probabilidades de que X estea a menos de k desviacións estándar da media, é dicir, a probabilidade de que X tome un valor no intervalo $(\mu - k\sigma, \mu + k\sigma)$. Pola propiedade da estandarización da Normal é doado comprobar que

$$P(\mu - k\sigma \leq X \leq \mu + k\sigma) = P(-k \leq Z \leq k).$$

Para $k = 1, 2, 3$ e 4 obtemos as seguintes probabilidades (**exercicio**):

$$\begin{array}{rcccc} P(\mu - \sigma \leq X \leq \mu + \sigma) & = & P(-1 \leq Z \leq 1) & = & 0.683 \\ P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) & = & P(-2 \leq Z \leq 2) & = & 0.954 \\ P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) & = & P(-3 \leq Z \leq 3) & = & 0.997 \\ P(\mu - 4\sigma \leq X \leq \mu + 4\sigma) & = & P(-4 \leq Z \leq 4) & = & \text{aprox. } 1 \end{array}$$

Como vemos, en realidade os valores naturais para unha variable aleatoria Normal podemos dicir que están no intervalo que se estende 3 (ou, se queremos ser máis estritos, 4) desviacións estándar á esquerda e á dereita da media. Por exemplo, se $X \sim N(70, 11)$, entón os seus posibles valores estarán aproximadamente no intervalo $(70 - 3 \cdot 11, 70 + 3 \cdot 11) = (37, 103)$, ou, sendo moi estritos, no intervalo $(70 - 4 \cdot 11, 70 + 4 \cdot 11) = (26, 114)$.

Exercicio 2.6. *Supoñamos que X segue unha distribución $N(\mu, \sigma)$. Sexan $x_{0.25}$ e $x_{0.75}$ o primeiro e o terceiro cuartil de X, é dicir, $P(X \leq x_{0.25}) = 0.25$ e $P(X \leq x_{0.75}) = 0.75$. Atopa a probabilidade $P(x_{0.25} - 1.5(x_{0.75} - x_{0.25}) \leq X \leq x_{0.75} + 1.5(x_{0.75} - x_{0.25}))$. Como se relaciona isto coa construción do boxplot que vimos no Capítulo 1? Cal é a probabilidade de observar un dato atípico se a variable observada é Normal? Depende esta probabilidade dos valores de μ e σ?*

Exercicio 2.7. *Supoñamos que X ten distribución $N(20, 3)$. Atopa dous valores a e b tales que a probabilidade de que X estea entre a e b é 0.95. Xeraliza o resultado para unha Normal calquera.*

2.3.1. Sumas de variables aleatorias

En moitos problemas de estatística aparecen sumas de variables aleatorias. A distribución Normal xoga un papel fundamental neste campo. Se as variables aleatorias que se suman son Normais, entón a suma tamén é Normal; isto chámase Reprodutividade da Normal. Pero para que a suma sexa Normal non é necesario que os sumandos sexan Normais. O Teorema Central do Límite garante que, en xeral, cando o número de sumandos non é moi pequeno, a suma de variables aleatorias é aproximadamente Normal.

Formalicemos estas propiedades:

- **Reprodutividade da Normal.** Supoñamos que temos n variables aleatorias independentes con distribución Normal, é dicir, $X_i \sim N(\mu_i, \sigma_i)$, para $i = 1, \dots, n$, entón a suma $S = \sum_{i=1}^{n} X_i$ tamén ten distribución Normal:

$$S = \sum_{i=1}^{n} X_i \sim N\left(\sum_{i=1}^{n} \mu_i, \sqrt{\sum_{i=1}^{n} \sigma_i^2}\right).$$

  En particular, se as variables X_i son identicamente distribuídas, é dicir, $X_i \sim N(\mu, \sigma)$ para $i = 1, \dots, n$, entón

$$S = \sum_{i=1}^{n} X_i \sim N(n\mu, \sqrt{n}\sigma).$$

- **Teorema Central do Límite.** Supoñamos que temos n variables aleatorias independentes X_i, tales que $\mathrm{E}(X_i) = \mu_i$ e $\mathrm{SD}(X_i) = \sigma_i$ para $i = 1, \dots, n$, entón a suma $S = \sum_{i=1}^{n} X_i$ ten aproximadamente distribución Normal:

$$S = \sum_{i=1}^{n} X_i \quad \text{é aproximadamente} \quad N\left(\sum_{i=1}^{n} \mu_i, \sqrt{\sum_{i=1}^{n} \sigma_i^2}\right)$$

  cando n non é moi pequeno.

  En particular, se as variables X_i teñen a mesma media e desviación estándar, é dicir, $\mathrm{E}(X_i) = \mu$ e $\mathrm{SD}(X_i) = \sigma$ para $i = 1, \dots, n$, entón

$$S = \sum_{i=1}^{n} X_i \quad \text{é aproximadamente} \quad N(n\mu, \sqrt{n}\sigma).$$

  cando n non é moi pequeno.

Obviamente, a expresión "cando n non é moi pequeno" é moi imprecisa. En realidade a aproximación á Normal que dá o Teorema Central do Límite funciona sorprendentemente ben para valores de n relativamente pequenos. Por exemplo acostúmase dicir que en xeral $n \geq 30$ é suficiente para ter unha boa aproximación. Non obstante, a calidade da aproximación depende fundamentalmente do carácter simétrico ou asimétrico das distribucións das variables que se suman. Se as distribucións son simétricas, pódese ter unha moi boa aproximación para valores moi pequenos de n. En cambio, se as distribucións das variables que se suman non son simétricas, entón n ten que ser maior para obter unha boa aproximación.

Exemplo. Vexamos como funciona o Teorema Central do Límite na práctica mediante unha **simulación**. Supoñamos que sumamos variables aleatorias con distribución Uniforme. Unha variable aletoria X ten distribución **Uniforme** no intervalo $[a, b]$, que denotamos por $X \sim Uniforme[a, b]$, se a súa función de densidade é

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{se } a < x < b, \\ 0 & \text{noutro caso.} \end{cases}$$

É doado de comprobar que $\mathrm{E}(X) = (a+b)/2$ e $\mathrm{SD}(X) = \sqrt{(b-a)^2/12}$ (**exercicio**).

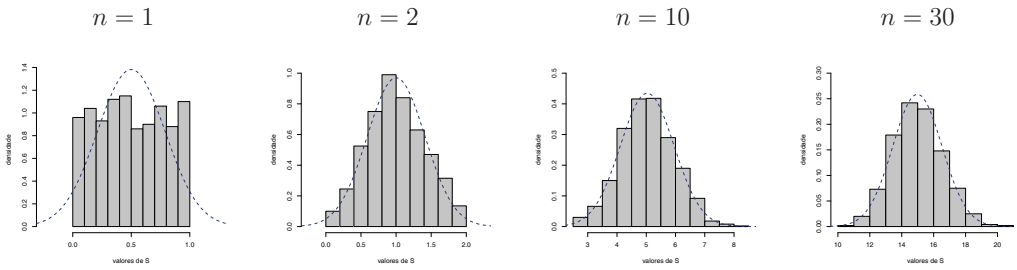


Figura 2.11: Histogramas da suma de n variables aleatorias independentes con distribución $Uniforme[0,1]$. A curva punteada é a función de densidade da $N(0.5n, \sqrt{n/12})$.

Supoñamos que temos n variables aleatorias independentes con distribución uniforme no intevalo $[0,1]$, é dicir, $X_i \sim Uniforme[0,1]$ para $i = 1, \ldots, n$. Neste caso $\mathrm{E}(X_i) = 0.5$ e $\mathrm{SD}(X_i) = \sqrt{1/12}$. Polo tanto, segundo o Teorema Central do Límite, $S = \sum_{i=1}^{n} X_i$ ten aproximadamente distribución $N(0.5n, \sqrt{n/12})$.

En R, podemos simular observacións destas variables mediante a función `runif()`. Por exemplo, para simular 5 observacións dunha $Uniforme[0,1]$ escribimos

```
> runif(5)

[1] 0.84441470 0.86878314 0.07010354 0.75317325 0.13466467
```

Agora podemos simular n valores e sumalos, e á súa vez este proceso podemos repetilo tantas veces como queiramos (digamos, por exemplo, 1000 veces). Con eses 1000 valores faremos histogramas para ver como se comporta a distribución da suma. En R:

```
> n <- 10                                          # número de sumandos
> simulacions <- matrix(runif(1000*n),ncol=n) # matriz 1000 x n
> observacions.de.S <- rowSums(simulacions)   # rowSums suma cada fila
> hist(observacions.de.S)                          # histog. dos 1000 valores de S
```

Cando $n = 1$ só temos un sumando. No primeiro histograma da Figura 2.11 vemos que a distribución Uniforme é moi distinta da distribución Normal. En cambio, conforme se aumenta o número de sumandos, a aproximación da distribución da suma á Normal é cada vez mellor. Para $n = 10$ as distribucións son xa practicamente indistinguibles. □

Exercicio 2.8. *Repite a simulación do exemplo anterior, pero substituíndo a distribución Uniforme pola distribución Exponencial*(1)*. Conséguese a mesma calidade na aproximación á Normal para valores pequenos de n? Como de grande debe ser n para que a aproximación sexa boa?*
Nota: a función `rexp()` *de R permite simular observacións dunha variable aleatoria Exponencial. Por exemplo, para simular 5 observacións dunha Exponencial*(1) *escribimos* `rexp(5,rate=1)`*.*

Exercicio 2.9. *A distribución Normal resulta axeitada para modelizar* **erros de medida**. *Un xarope antitusivo véndese en frascos de* 200 *ml. Por distintas razóns, o proceso de enchido dos frascos non é absolutamente preciso e provoca que a cantidade de líquido (en ml) en cada frasco sexa unha variable aleatoria Normal de media* 200 *ml e desviación estándar* 3 *ml.*

(a) Cal é a probabilidade de que un frasco conteña menos de 195 *ml?*

(b) Cal é a probabilidade de que un frasco conteña máis de 210 *ml?*

(c) Cal é o valor de c tal que o intervalo $(200 - c, 200 + c)$ *inclúe o* 98 *% dos frascos?*

(d) Se un lote está composto por 25 frascos, cal é a distribución da cantidade total de xarope nun lote?

(e) Cal é a probabilidade de que un lote conteña máis de 5.02 *l?*

2.4. Distribucións empregadas en inferencia estatística

En inferencia estatística aparecen unha serie de distribucións que son fundamentais para aplicar os distintos métodos inferenciais. As máis importantes son a Normal Estándar, a distribución t de Student, a distribución Chi-cadrado de Pearson e a distribución F de Snededor.

Distribución Normal Estándar $N(0, 1)$

A distribución **Normal Estándar**, habitualmente denotada por Z, é a distribución Normal con media 0 e desviación estándar 1, é dicir, $Z \sim N(0, 1)$. Esta distribución é amplamente empregada en inferencia estatística.

En R as funcións `pnorm()` e `qnorm()` permiten obter a función de distribución e función cuantil da Normal Estándar:

```
> pnorm(1.43)

[1] 0.9236415

> qnorm(0.70)

[1] 0.5244005
```

Distribución t de Student

A distribución t **de Student** (ou simplemente, distribución t) depende dun número chamado "graos de libertade" (*"degrees of freedom"*, en inglés). Para cada valor dos graos de liberdade, g, temos unha distribución t distinta. Denotamos cada unha destas distribucións por t_g.

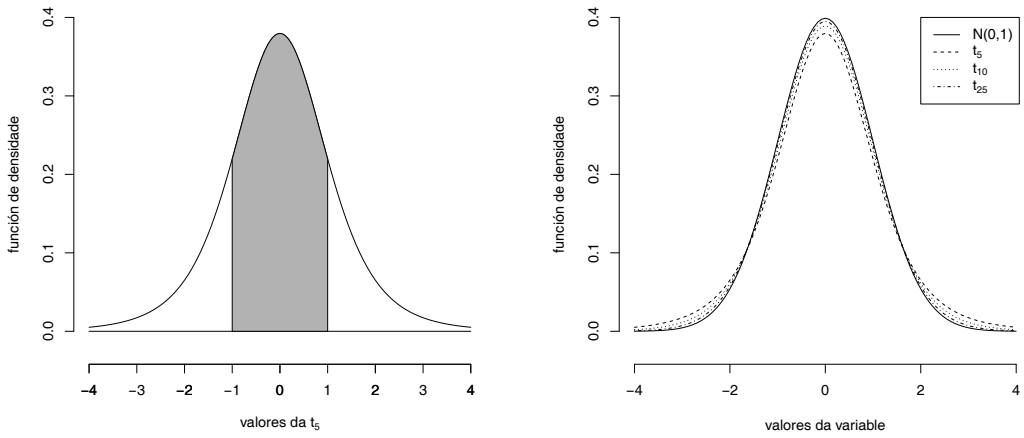


Figura 2.12: Esquerda: densidade dunha variable aleatoria con distribución t_5; a área sombreada é a probabilidade $P(-1 \leq t_5 \leq 1)$. Dereita: comparativa das densidades t_5, t_{10}, t_{25} e $N(0,1)$.

A densidade da t é moi similar á da $N(0,1)$. Está centrada no 0, é simétrica e ten forma de campá. A única diferenza aparece nas colas da distribución, nas que a t presenta unha maior probabilidade. A t_g pode tomar valores máis extremos ca os da $N(0,1)$. Conforme se incrementa o valor de g, a distribución t_n parécese máis á $N(0,1)$. Para $g \geq 100$ a t_g e a $N(0,1)$ son practicamente indistinguibles (véxase a Figura 2.12). En R empregamos as funcións `pt()` e `qt()` para obter probabilidades e cuantís da distribución t.

Exemplo. Calculemos as probabilidades da forma $P(-k \leq t_5 \leq k)$, para $k = 1, 2, 3, 4$, e comparémolas coas correspondentes probabilidades que xa calculamos para a Normal. Para calcular $P(-1 \leq t_5 \leq 1)$ escribimos en R

```
> pt(1,df=5)-pt(-1,df=5)

[1] 0.6367825
```

De forma análoga podemos obter

$$\begin{aligned} P(-1 \leq t_5 \leq 1) &= 0.637 \\ P(-2 \leq t_5 \leq 2) &= 0.898 \\ P(-3 \leq t_5 \leq 3) &= 0.970 \\ P(-4 \leq t_5 \leq 4) &= 0.989 \end{aligned}$$

Nótese que estas probabilidades son lixeiramente menores ca as correspondentes probabilidades da $N(0,1)$ que calculamos na páxina 47. □

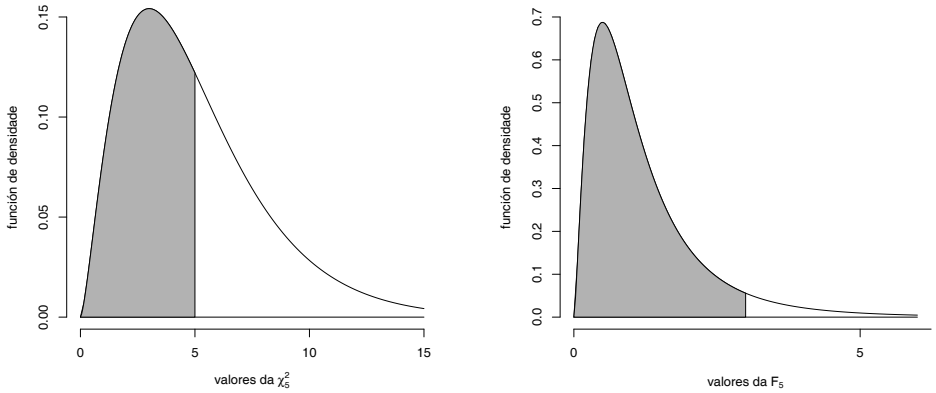


Figura 2.13: Esquerda: densidade dunha χ^2_5; a área sombreada é a probabilidade $P(0 \leq \chi^2_5 \leq 5)$. Dereita: densidade dunha $F_{5,10}$; a área sombreada é a probabilidade $P(0 \leq F_{5,10} \leq 3)$.

Exercicio 2.10. *Emprega as funcións* `qt()` *e* `qnorm()` *para atopar o valor a tal que: (a)* $P(-a \leq t_5 \leq a) = 0.95$*; (b)* $P(-a \leq t_{10} \leq a) = 0.95$*; (c)* $P(-a \leq t_{50} \leq a) = 0.95$*; (d)* $P(-a \leq t_{100} \leq a) = 0.95$*; (e)* $P(-a \leq N(0,1) \leq a) = 0.95$.

Distribución Chi-cadrado

A **Chi-cadrado** de Pearson é unha distribución asimétrica con soporte $[0, +\infty)$. Igual ca no caso da t, depende dos graos de liberdade, g, de xeito que para cada valor de g temos unha distribución distinta. A notación é χ^2_g. A Figura 2.13 amosa a función de densidade dunha variable aleatoria con distribución χ^2_5.

En R as funcións `pchisq()` e `qchisq()` permiten calcular probabilidades e cuantís asociados á distribución chi-cadrado.

Exemplo. Para calcular por exemplo $P(1.50 \leq \chi^2_5 \leq 7.25)$ en R escribimos `pchisq(7.25,df=5)-pchisq(1.50,df=5)`, que resulta 0.7104. Para obter os cuantís da chi-cadrado empregamos a función `qchisq()`: por exemplo, `qchisq(0.95,10)`, que resulta 18.31. □

Distribución F de Snedecor

A **F de Snedecor** (ou simplemente, F) é unha distribución asimétrica con soporte $[0, +\infty)$ que depende de dous valores de graos de liberdade, g_1 e g_2. A notación é F_{g_1,g_2}. A Figura 2.13 amosa a función de densidade dunha variable aleatoria con distribución $F_{5,10}$. En R as funcións `pf()` e `qf()` permiten calcular probabilidades e cuantís asociados á distribución F.

Exemplo. En R para obter o cuantil de orde 0.95 dunha $F_{5,10}$ escribimos `qf(0.95,5,10)`, que resulta 3.33. □

2.5. Conceptos relevantes en biomedicina

2.5.1. Clasificación. Sensibilidade e especificidade. Prevalencia e incidencia

Un gran número de problemas en estatística poden describirse da seguinte maneira: unha poboación está dividida en **dúas clases** disxuntas ás que nos referiremos xenericamente como "**clase Positiva**" (P) e "**clase Negativa**" (N) e queremos clasificar un individuo nunha desas dúas clases. Un **procedemento de asignación** consiste en clasificar un individuo nunha das dúas clases en base á información dispoñible sobre ese individuo. Desafortunadamente, o procedemento de asignación case nunca será perfecto e, ás veces, cometerá erros: algúns individuos serán clasificados incorrectamente porque a información da que se dispón pode ser enganosa. Debido a esta imperfección, necesitamos avaliar a calidade do procedemento de asignación.

Exemplos.

- Diagnose médica: detectar se unha persoa sofre unha determinada enfermidade ou non.
- Diagnose médica: decidir se un paciente sofre a enfermidade A ou a enfermidade B.
- Computación: filtrar os correos electrónicos para decidir se son verdadeiros ou se son spam.
- Teoría do sinal: distinguir entre sinal e ruído.
- ... □

Consideremos unha **proba médica** deseñada para detectar a presenza ou ausencia dunha determidada enfermidade. A partir dun protocolo (cuestionario, análises clínicas, procedementos médicos etc.) a proba proporciona unha resposta binaria:

- **Positivo** (**P**): o individuo é diagnosticado como "enfermo".
- **Negativo** (**N**): o individuo é diagnosticado como "non enfermo/san".

Gustaríanos que a proba clasificase a todos os individuos correctamente, dando un valor Positivo a todos os enfermos e un valor Negativo a todos os non enfermos. Non obstante, na maioría dos casos prácticos isto non ocorrerá e algúns individuos serán clasificados incorrectamente. Aparecen entón os seguintes **erros de clasificación**:

- algúns individuos que realmente padecen a enfermidade, debido ás súas características particulares, non serán diagnosticados como enfermos (**falsos negativos**);
- e, de xeito análogo, algúns individuos non enfermos serán incorrectamente clasificados como enfermos (**falsos positivos**).

Cando se aplica un procedemento de asignación aparecen polo tanto as catro situacións que se recollen no cadro seguinte:

		status	
		enfermo (status P)	**non enfermo** (status N)
diagnose	**enfermo** (diagnose P)	**verdadeiro positivo** VP	**falso positivo** FP
	non enfermo (diagnose N)	**falso negativo** FN	**verdadeiro negativo** VN

A calidade dun procedemento de asignación mídese a través das seguintes probabilidades condicionadas:

- **fracción de verdadeiros positivos** = FVP = $P(\text{diagnose P} \mid \text{status P})$
- **fracción de verdeiros negativos** = FVN = $P(\text{diagnose N} \mid \text{status N})$
- **fracción de falsos positivos** = FFP = $P(\text{diagnose P} \mid \text{status N}) = 1 - \text{FVN}$
- **fracción de falsos negativos** = FFN = $P(\text{diagnose N} \mid \text{status P}) = 1 - \text{FVP}$

A fracción de verdadeiros positivos, FVP, chámase **sensibilidade**[2]. A fracción de verdadeiros negativos, FVN, chámase **especificidade**[3]. Por outra parte, das catro probabilidades anteriores só dúas son relavantes xa que $\text{FFP} = 1 - \text{FVN}$ e $\text{FFN} = 1 - \text{FVP}$. Habitualmente, para describir a precisión dunha proba diagnóstica emprégase o par de probabilidades $(\text{FFP}, \text{FVP}) = (1 - \textbf{especificidade}, \textbf{sensibilidade})$:

- Unha proba perfecta tería $\text{FFP} = 0$ (especificidade = 1) e $\text{FVP} = 1$ (sensibilidade = 1).
- Unha proba completamente inútil tería FFP = FVP (a proba non ten ningunha relación coa enfermidade).
- En xeral, as probas terán $\text{FFP} \geq 0$ e $\text{FVP} \leq 1$.
- As probabilidades FFP e FVP tamén se poden empregar para comparar distintos procedementos de asignación.

Exemplo. En novembro de 2020 a Comisión Europea adoptou unha serie de recomendacións para o uso de tests rápidos de antíxenos para a detección de infección por SARS-CoV-2. En particular, esixiu que os tests presentasen como mínimo unha sensibilidade do 80 % e unha especificidade do 97 %. □

Outros dous conceptos importantes no estudo de probas diagnósticas, especialmente desde o punto de vista epidemiolóxico, son a prevalencia e a incidencia:

[2] En inglés: *"sensitivity"*.
[3] En inglés: *"specificity"*.

- A **prevalencia** é a proporción de individuos que padecen a enfermidade nun determinado momento. Por exemplo, no caso de enfermidades infecciosas, como a COVID-19, fálase de "casos activos", é dicir o número de persoas que están contaxiadas con infección activa nun momento determinado. Noutro tipo de enfermidades de longa duración, como por exemplo a diabetes, a prevalencia é simplemente a proporción de individuos que padecen esta enfermidade.
- A **incidencia** dá información sobre o número de casos novos que contraen a enfermidade ao longo dun período de tempo determinado. Este é un concepto importante no caso das enfermidades infecciosas de curta duración. Por exemplo, no caso da COVID-19 empregouse a "incidencia nos últimos 14 días por cada 100.000 habitantes" para analizar a evolución da epidemia nunha zona xeográfica.

Se a prevalencia da enfermidade na poboación total é $\pi = P(\text{status P})$, entón podemos aplicar o Teorema das Probabilidades Totais para obter a probabilidade **global de clasificación errónea** (M) da proba diagnóstica:

$$\begin{aligned} P(M) &= P(M \mid \text{status P})P(\text{status P}) + P(M \mid \text{status N})P(\text{status N}) \\ &= (1 - \text{FVP})\pi + \text{FFP}(1 - \pi). \end{aligned}$$

Exemplo. (cont.) Supoñamos que, seguindo as recomendacións a Comisión Europea, un test rápido de antíxenos ten unha sensibilidade do 85 % e unha especificidade do 98 %. Se nun momento determinado a prevalencia é do 1.2 %, entón a probabilidade global de clasificación errónea é

$$(1 - 0.85) \cdot 0.012 + (1 - 0.98) \cdot (1 - 0.012) = 0.02156.$$

Polo tanto, o 2.16 % dos diagnósticos serán erróneos (incluíndo falsos positivos e falsos negativos). □

Esta probabilidade global de clasificación errónea pódese empregar nalgunhas situacións para resumir a calidade do procedemento de asignación, pero habitualmente non resulta informativa porque as consecuencias dos dous tipos de erro non son simétricas:

- Un falso negativo significa que o individuo está realmente enfermo, pero non se lle diagnostica a enfermidade. Polo tanto non recibirá o tratamento adecuado ou, no caso dunha enfermidade contaxiosa, poderá contaxiar a outras persoas ao non tomar as medidas preventivas oportunas. En xeral as consecuencias dun falso negativo son bastante graves.
- Aínda que obviamente depende moito da enfermidade concreta, os falsos positivos son en xeral menos graves. Se un individuo san é diagnosticado como enfermo, seguramente terá que seguir algún tratamento ou tomar certas medidas preventivas innecesarias no seu caso. Á parte destes inconvenientes persoais, as consecuencias non son en xeral graves.

Esa é a razón pola que se prefire o par de probabilidades (FFP, FVP) para describir a calidade da proba. Ademais, como vimos antes, a probabilidade global de clasificación errónea depende da prevalencia, e esta cantidade moitas veces é descoñecida ou difícil de estimar.

As probabilidades anteriores (sensibilidade, especificidade, prevalencia) son interesantes desde o punto de vista médico ou epidemiolóxico. Pero desde o punto de vista do individuo resulta

máis importante contestar ás seguintes preguntas: se a proba deu un resultado positivo, cal é a probabilidade de que padeza a enfermidade?, ou se a proba deu un resultado negativo, cal é a probabilidade de non ter a enfermidade? Estas probabilidades condicionadas denomínanse **valores preditivos**:

- **valor preditivo positivo** = VPP = $P(\text{status P} \mid \text{diagnose P})$,
- **valor preditivo negativo** = VPN = $P(\text{status N} \mid \text{diagnose N})$.

Unha proba diagnóstica perfecta tería que diagnosticar perfectamente a enfermidade, é dicir tería que ter VPP = 1 e VPN = 1. Por outra parte, unha proba completamente inútil desde o punto de vista da diagnose non daría ningunha información sobre o status real do paciente, é dicir, tería

$$\text{VPP} = P(\text{status P} \mid \text{diagnose P}) = P(\text{status P}) = \pi$$

e

$$\text{VPN} = P(\text{status N} \mid \text{diagnose N}) = P(\text{status N}) = 1 - \pi,$$

onde π é a prevalencia.

Os valores preditivos poden escribirse en termos da sensibilidade, especificidade e prevalencia grazas ao **Teorema de Bayes**:

$$\begin{aligned}
\text{VPP} &= P(\text{status P} \mid \text{diagnose P}) \\
&= \frac{P(\text{diagnose P} \mid \text{status P})P(\text{status P})}{P(\text{diagnose P} \mid \text{status P})P(\text{status P}) + P(\text{diagnose P} \mid \text{status N})P(\text{status N})} \\
&= \frac{\text{FVP}\ \pi}{\text{FVP}\ \pi + \text{FFP}\ (1-\pi)} \\
&= \frac{\text{sensibilidade} \cdot \pi}{\text{sensibilidade} \cdot \pi + (1 - \text{especificidade}) \cdot (1-\pi)}, \\
\text{VPN} &= \frac{(1-\text{FFP})\ (1-\pi)}{(1-\text{FFP})\ (1-\pi) + (1-\text{FVP})\ \pi} \\
&= \frac{\text{especificidade} \cdot (1-\pi)}{\text{especificidade} \cdot (1-\pi) + (1 - \text{sensibilidade}) \cdot \pi}. \qquad \textbf{(exercicio)}
\end{aligned}$$

Exemplo. (cont.) Seguindo co mesmo exemplo, supoñamos que un test de antíxenos ten unha sensibilidade do 85 %, unha especificidade do 98 % e que a prevalencia é do 1.2 %. Os valores preditivos son

$$\begin{aligned}
\text{VPP} &= \frac{0.85 \cdot 0.012}{0.85 \cdot 0.012 + (1 - 0.98) \cdot (1 - 0.012)} = 0.34, \\
\text{VPN} &= \frac{0.98 \cdot (1 - 0.012)}{0.98 \cdot (1 - 0.012) + (1 - 0.85) \cdot 0.012} = 0.998.
\end{aligned}$$

Nótese o valor sorprendentemente baixo do VPP. Por que ocorre isto? (**exercicio**) □

Exercicio 2.11. *No exemplo anterior:*

(a) Mantendo unha especifidade do 98 *% e a prevalencia igual ao* 1.2 *%, podemos alcanzar un valor preditivo positivo igual a* 1 *para algún valor da sensibilidade? Cal é o máximo valor posible do valor preditivo positivo?*

(b) Calcula os valores preditivos se a prevalencia é do 5 *% (mantendo a sensibilidade no* 85 *% e a especifidade no* 98 *%). Podería agora alcanzarse un valor preditivo positivo de* 1 *para algún valor da sensibilidade?*

(c) Programa unha función en R que teña como argumentos a sensibilidade, a especificidade e a prevalencia e devolva os valores preditivos positivo e negativo.

2.5.2. A curva ROC

En moitas ocasións, a información sobre os individuos en estudo é de tipo **continuo**. Por exemplo, nunha proba médica, pódense empregar diversos **biomarcadores** continuos para a diagnose dunha enfermidade (peso, índice de masa corporal, concentración de glucosa, presión arterial etc.). De feito, habitualmente a información contén variables de distinto tipo (categóricas, ordinais, continuas). A partir de agora consideraremos que toda a información relativa a un individuo se reduce a unha variable continua Y, que lle chamaremos **variable de diagnose**. Esta variable pode construírse na práctica como combinación doutras variables.

Como a información vén dada a través dunha variable continua, o procedemento de asignación ten que facerse comparando esa variable cun certo **punto de corte**, c. Supoñamos que **os individuos da clase Positiva tenden a presentar valores da variable de diagnose maiores ca os valores que presentan os individuos da clase Negativa**. Entón o procedemento de asignación é o seguinte:

- **Se $Y > c$ entón o individuo é asignado á clase Positiva (P).**
- **Se $Y \leq c$ entón o individuo é asignado á clase Negativa (N).**

Se foramos capaces de atopar un punto de corte c de tal xeito que todos os individuos da clase Positiva presentasen valores da variable de diagnose por riba de c e todos os da clase Negativa presentasen valores por debaixo de c, entón o procedemento de asignación sería perfecto. Pero esta situación ideal non ocorre na práctica porque normalmente o conxunto de posibles valores da variable de diagnose nas dúas clases (é dicir, os soportes) solápanse. Isto lévanos de novo aos dous tipos de erros que xa coñecemos:

- Algúns individuos da clase Positiva poden presentar un valor na variable de diagnose menor que c e polo tanto serán asignados á clase Negativa (**falsos negativos**).
- Analogamente, o valor dalgúns individuos da clase Negativa pode exceder o punto de corte c e polo tanto serán asignados á clase Positiva (**falsos positivos**).

Por suposto, **os erros de clasificación dependen do punto de corte** c. Se cambiamos o punto de corte, entón as probabilidades de cometer cada un dos posibles erros tamén cambirán.

A partir de agora asumiremos que a variable de diagnose se modeliza mediante dúas variables aleatorias: Y_P na clase Positiva e Y_N na clase Negativa. A súas funcións de distribución son

- na clase Positiva: $F_P(c) = P(Y_P \leq c)$,
- na clase Negativa: $F_N(c) = P(Y_N \leq c)$.

Tal e como xa dixemos, estamos supoñendo que os individuos da clase Positiva tenden a ter valores máis altos da variable de diagnose ca os da clase Negativa. En termos probabilísticos, formalmente isto significa que Y_P é *estocasticamente maior* ca Y_N, é dicir, $F_P(c) \leq F_N(c)$ para todo c.

Para un punto de corte c, as probabilidades que empregabamos para analizar a calidade do procedemento de asignación son

$$\begin{aligned} \text{FVP} &= P(Y_P > c) = 1 - F_P(c) = \textbf{sensibilidade}, \\ \text{FFP} &= P(Y_N > c) = 1 - F_N(c) = 1 - \textbf{especificidade}. \end{aligned}$$

Exemplo. Supoñamos que as variables de diagnose na clase Negativa e Positiva teñen distribucións $Y_N \sim N(0, 0.75)$ e $Y_P \sim N(2, 1)$, respectivamente, e tomemos $c = 1.0$. Entón

$$\begin{aligned} \text{FVP} &= P(Y_P > c) = P(N(2,1) > 1.0) = 0.84 \\ \text{FFP} &= P(Y_N > c) = P(N(0, 0.75) > 1.0) = 0.09 \end{aligned}$$

Con este punto de corte alcanzaríase polo tanto unha sensibilidade de 0.84 e unha especificidade de $1 - 0.09 = 0.91$. □

Se o punto de corte varía, as probabilidades FFP e FVP tamén o farán, tal e como podemos ver na Figura 2.14. Entón podemos consideralas como funcións de c:

$$\text{FVP}(c) = P(Y_P > c) = 1 - F_P(c) \qquad \text{e} \qquad \text{FFP}(c) = P(Y_N > c) = 1 - F_N(c).$$

Se movemos o punto de corte c de forma continua ao longo do soporte das variables Y_P e Y_N e representamos os pares de puntos

$$(\text{FFP}(c), \text{FVP}(c)) = (1 - \text{especificidade}(c), \text{sensibilidade}(c))$$

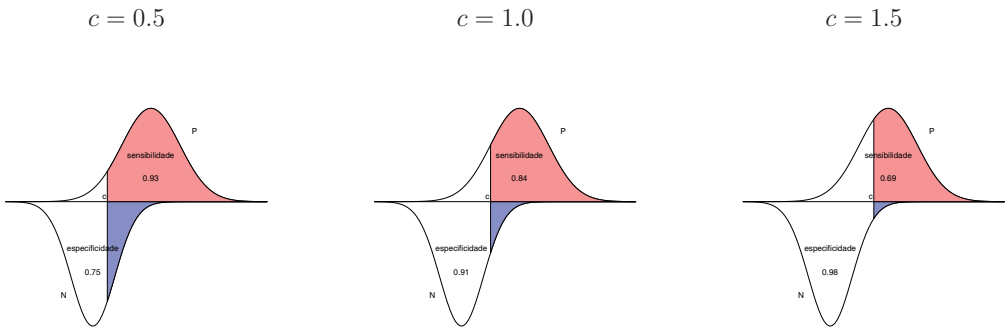
obtemos unha curva que se chama **curva ROC** (do inglés *Receiver Operating Characteristic Curve*), tal e como se ilustra na Figura 2.15.

A curva ROC é polo tanto o lugar xeométrico do plano formado polos puntos

$$\Big\{ \big(\text{FFP}(c), \text{FVP}(c)\big),\ c \in \mathbb{R} \Big\} = \Big\{ \big(1 - F_N(c), 1 - F_P(c)\big),\ c \in \mathbb{R} \Big\}.$$

Se facemos o cambio de variable $p = 1 - F_N(c)$ entón podemos reparametrizar a curva ROC como

$$\Big\{ \big(p, 1 - F_P(F_N^{-1}(1-p))\big), \quad 0 < p < 1 \Big\},$$



FVP $= 0.93$
FFP $= 0.25$

FVP $= 0.84$
FFP $= 0.09$

FVP $= 0.69$
FFP $= 0.02$

Figura 2.14: Ao variar o punto de corte c obtemos distintos valores para as probabilidades FFP (1−especificidade) e FVP (sensibilidade).

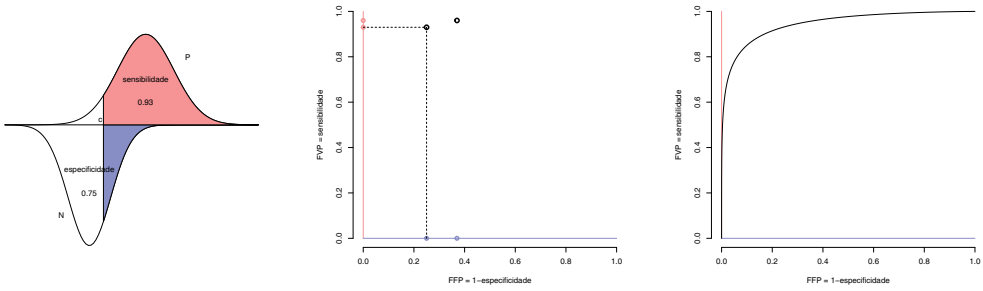


Figura 2.15: Construción da curva ROC. Ao variar o punto de corte c obtemos os pares $(\text{FFP}(c), \text{FVP}(c))$, que, ao representalos no cadrado unidade, permiten obter a curva ROC.

Nótese que cando c varía en $(-\infty, +\infty)$, p varía en $(0, 1)$. Incluso podemos considerar a curva ROC como unha función matemática de p expresándoa como

$$\text{ROC}(p) = 1 - F_P(F_N^{-1}(1-p)), \qquad \text{para } 0 < p < 1,$$

onde F_P é a función de distribución da variable de diagnose na clase Positiva e F_N^{-1} é a función cuantil da variable de diagnose na clase Negativa.

2.5.3. Propiedades da curva ROC

- A curva ROC está contida no cadrado unidade $[0, 1] \times [0, 1]$.
- A curva ROC conecta os puntos $(0, 0)$ e $(1, 1)$. (**exercicio**)
- A curva ROC é unha función monótona crecente, é dicir, se $p_1 < p_2$ entón $\text{ROC}(p_1) \leq \text{ROC}(p_2)$. (**exercicio**)
- Se as distribucións da variable de diagnose na poboación positiva e na poboación negativa son iguais, entón $\text{FFP}(c) = \text{FVP}(c)$ para calquera valor de c. Neste caso a curva ROC é a diagonal do cadrado, tal e como se amosa na Figura 2.16-(a). O procedemento de asignación resulta completamente inútil.
- Por outra parte, un procedemento de asignación perfecto separaría perfectamente as poboacións Positiva e Negativa. Isto é, para algún valor do punto de corte c (ollo! pero non necesariamente para todos) teriamos $\text{FVP}(c) = 1$ e $\text{FFP}(c) = 0$. Neste caso a curva ROC situaríase ao longo do lado esquerdo e da parte superior do cadrado unidade, tal e como se amosa na Figura 2.16-(b).
- A maior parte dos procedementos de asignación presentan curvas ROC que están entre as dos dous casos anteriores, como a que se amosa na Figura 2.16-(c).
- Se a curva ROC está por debaixo da diagonal, entón significa que as distribucións das variables de diagnose están orientadas ao revés, é dicir, os individuos da clase Positiva tenden a tomar valores menores ca os da clase Negativa. Neste caso podemos facer dúas cousas: ou ben cambiar de signo a variable de diagnose ou ben facer a asignación ao revés (asignar á clase Positiva cando a variable de diagnose é menor ca c).
- A curva ROC é invariante ante transformacións estritamente crecentes da variable de diagnose (por exemplo, non se ve afectada por cambios de escala ou incluso por transformacións logarítmicas).
- As curvas ROC poden empregarse para comparar procedementos de asignación. Comparando por exemplo as curvas ROC que aparecen na Figura 2.17 podemos dicir que o procedemento baseado no test A é mellor ca o procedemento baseado no test B porque a curva ROC do test A está sempre por riba da curva ROC do test B. Para calquera valor da FFP, p, a FVP do test A é sempre maior ca a do test B. De xeito similar, se escollemos puntos de corte c_A e c_B para os cales $\text{FVP}_A(c_A) = \text{FVP}_B(c_B)$, entón as FFPs están ordenadas en favor do test A xa que $\text{FFP}_A(c_A) < \text{FFP}_B(c_B)$.

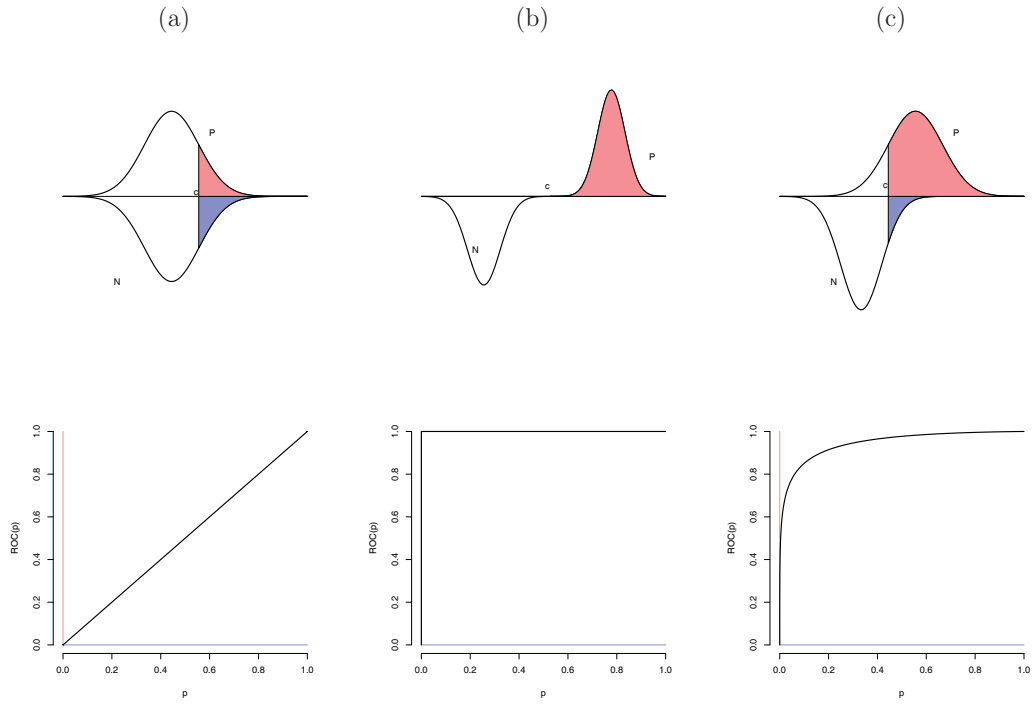


Figura 2.16: Distintos tipos de curvas ROC: (a) o procedemento de clasificación é completamente inútil; (b) o procedemento de clasificación é perfecto; (c) exemplo de curva ROC habitual na práctica.

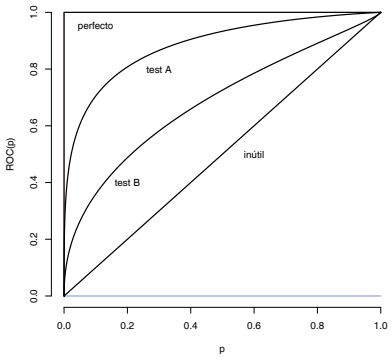


Figura 2.17: Comparación de dous procedementos de asignación a través das súas curvas ROC.

- Desde un punto de vista práctico, a curva ROC dá unha descrición completa dun procedemento de asignación baseado nunha variable de diagnose continua, facilita a comparación e a combinación de información para obter procedementos de asignación mellores e axuda na elección do punto de corte en problemas aplicados.

2.5.4. Valores resumo da curva ROC: a área debaixo da curva e o índice de Youden

Como acabamos de ver, a curva ROC é unha descrición completa sobre o comportamento dun procedemento de asignación baseado nunha variable de diagnose continua. En moitas ocasións a curva ROC acompáñase de medidas resumo que describen algún aspecto particular. Os máis empregados son a área debaixo da curva e o índice de Youden.

A **área debaixo da curva** (**AUC**, do inglés *area under the curve*) é a medida resumo da curva ROC máis empregada. Defínese como

$$\text{AUC} = \int_0^1 \text{ROC}(p)dp.$$

Vista a construción da curva ROC, a AUC tomará valores entre 0.5 e 1:

- AUC $= 0.5$ correspóndese coa curva ROC que coincide coa diagonal, e polo tanto cun procedemento de asignación inútil.
- AUC $= 1$ correspóndese cun procedemento de asignación perfecto.
- A maior parte das curvas ROC na práctica teñen AUCs entre 0.5 e 1.
- Se dous procedementos, A e B, verifican que $\text{ROC}_A(p) \geq \text{ROC}_B(p)$ para todo $0 < p < 1$, entón obviamente $\text{AUC}_A \geq \text{AUC}_B$. Non obstante, o contrario non é necesariamente certo (pensemos por exemplo en curvas ROC que se crucen).
- A AUC ten unha interpretación probabilística importante: se Y_P e Y_N son as variables de diagnose de dous individuos escollidos de forma independente, o primeiro da clase Positiva e o segundo da Negativa, entón $\text{AUC} = P(Y_P > Y_N)$.

O **índice de Youden** (**IY**) é a máxima diferenza entre as probabilidades $\text{FVP}(c)$ e $\text{FFP}(c)$. Pode expresarse de varias formas:

$$\begin{aligned}
\text{IY} &= \sup_c |\text{FVP}(c) - \text{FFP}(c)| \\
&= \sup_c |\text{sensibilidade}(c) + \text{especificidade}(c) - 1| \\
&= \sup_c |(1 - F_P(c)) - (1 - F_N(c))| \\
&= \sup_c |F_N(c) - F_P(c)| \\
&= \max_p |\text{ROC}(p) - p|.
\end{aligned}$$

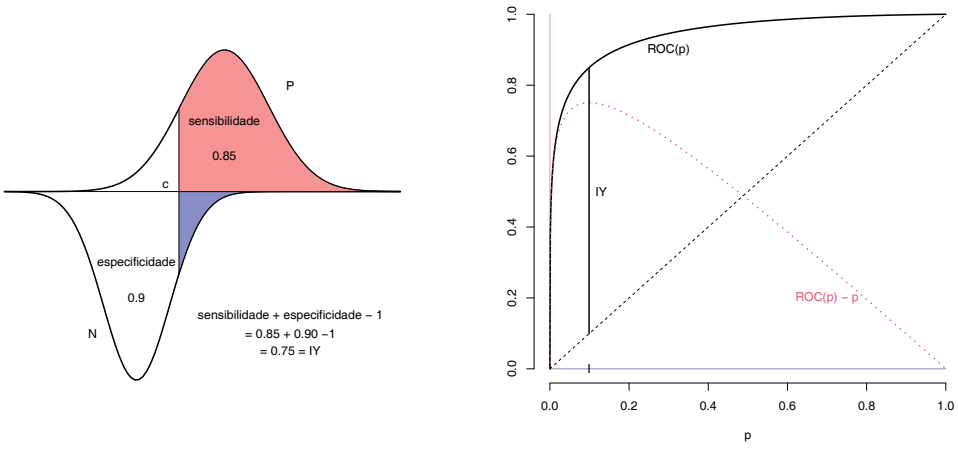


Figura 2.18: Construción do índice de Youden.

Da segunda expresión vemos que o índice de Youden maximiza a suma da sensibilidade e a especificidade. Da penúltima expresión, vemos que o índice de Youden é a máxima separación entre as funcións de distribución da variable de diagnose nas clases Positiva e Negativa. Da última expresión vemos que o índice de Youden tamén é a máxima distancia vertical entre a curva ROC e a diagonal, tal e como se pode comprobar na Figura 2.18. Os valores do índice de Youden están entre 0 (procedemento inútil) e 1 (procedemento perfecto).

O punto de corte asociado ao índice de Youden pode empregarse como **punto de corte óptimo**, no sentido que será o punto que maximice a suma de sensibilidade e especificidade. Sexa p_0 o valor para o cal $\text{IY} = \text{ROC}(p_0) - p_0$, entón o punto de corte óptimo é

> o cuantil de orde $(1 - p_0)$ de Y_N, é dicir, $F_N^{-1}(1 - p_0)$,

ou ben

> o cuantil de orde $(1 - \text{ROC}(p_0))$ de Y_P, é dicir, $F_P^{-1}(1 - \text{ROC}(p_0))$.

Desde o punto de vista teórico, estes dous valores coinciden.

Exemplo. (cont.) Sexan $Y_N \sim N(0, 0.75)$ e $Y_P \sim N(2, 1)$. A curva ROC correspondente aparece na Figura 2.19. Ten as seguintes características:

- $\text{AUC} = 0.95$.
- $\text{IY} = 0.75$, que se obtén para $p = 0.099$. Nese caso a especifidade é $1 - 0.099 = 0.901$ e a sensibilidade é 0.850.

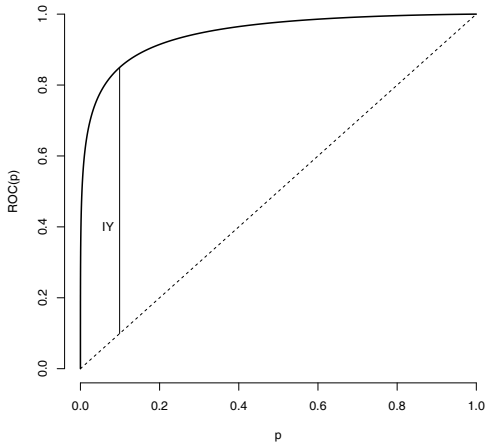


Figura 2.19: Curva ROC correspondente ás variables de diagnose $Y_N \sim N(0, 0.75)$ e $Y_P \sim N(2, 1)$.

- O punto de corte asociado ao índice de Youden é o cuantil de orde 0.901 da $N(0, 0.75)$, ou o cuantil de orde $1 - 0.850 = 0.15$ da $N(2, 1)$. Nos dous casos o valor do punto de corte resulta 0.965.

2.5.5. A curva ROC binormal

Cando as distribucións da variable de diagnose nas poboacións Positiva e Negativa son Normais, entón a curva ROC correspondente chámase **curva ROC Binormal**. Sexa $Y_P \sim N(\mu_P, \sigma_P)$ e $Y_N \sim N(\mu_N, \sigma_N)$. Entón a curva ROC Binormal é

$$\text{ROC}(p) = \Phi(a + b\ \Phi^{-1}(p)),$$

onde Φ e Φ^{-1} son a función de distribución e a función cuantil da Normal Estándar, respectivamente, e

$$a = \frac{\mu_P - \mu_N}{\sigma_P} \quad \text{e} \quad b = \frac{\sigma_N}{\sigma_P}.$$

A AUC asociada á curva ROC Binormal é

$$\text{AUC} = \Phi\left(\frac{\mu_P - \mu_N}{\sqrt{\sigma_P^2 + \sigma_N^2}}\right) = \Phi\left(\frac{a}{\sqrt{1 + b^2}}\right).$$

Exercicio 2.12. *Supoñamos que se está estudando a posibilidade de empregar a concentración de glucosa medida sobre unha gota de sangue obtida mediante unha sinxela punción nun dedo para diagnosticar a diabetes. Sábese que as distribucións da variable de diagnose nas poboacións Positiva (individuos diabéticos) e Negativa (individuos non diabéticos) son* $Y_P \sim N(130, 25)$ *e* $Y_N \sim N(90, 15)$*, respectivamente.*

(a) Coa axuda de R, calcula a curva ROC e represéntaa graficamente.

(b) Calcula o valor da AUC.

(c) Acha o índice de Youden e o correspondente punto de corte. Cales son os valores da sensibilidade e da especificidade para ese punto de corte? Nota: para atopar o índice de Youden pódese facer unha búsqueda numérica nun vector de puntos p da forma $(0.001, 0.002, \ldots, 0.999, 1)$*.*

(d) Representa nun mesmo gráfico as funcións de densidade das variables Y_P *e* Y_N *e identifica o punto de corte asociado ao índice de Youden no soporte. Que característica gráfica cumpre?*

Exercicio 2.13. *Supoñamos que as densidades das variables de diagnose nas poboacións Negativa e Positiva son as seguintes:*

$$f_N(y) = \begin{cases} 0 & \text{se } y \leq 0, \\ e^{-y} & \text{se } y > 0. \end{cases} \qquad f_P(y) = \begin{cases} 0 & \text{se } y \leq 1, \\ e^{-y+1} & \text{se } y > 1. \end{cases}$$

Nótese que $Y_N \sim Exponencial(1)$*. Por outra parte,* Y_P *compórtase como* $E + 1$*, onde* $E \sim Exponencial(1)$*.*

(a) Obtén a fórmula explícita da curva ROC e represéntaa graficamente.

(b) Atopa o valor da AUC.

(c) Atopa o valor do índice de Youden e o correspondente punto de corte. Cales son a sensibilidade e a especificidade asociadas a ese punto de corte?

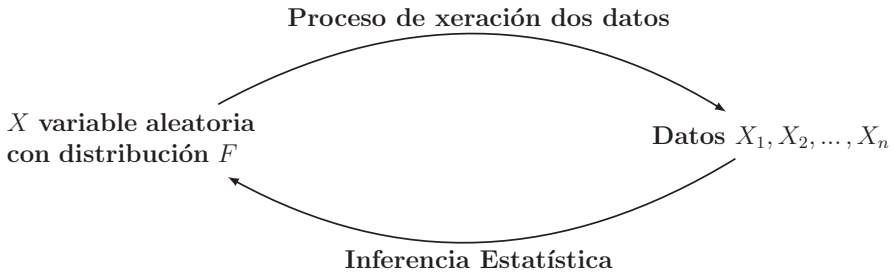
Capítulo 3

Métodos inferenciais

Contidos

3.1. Que é a inferencia estatística?

A **Inferencia Estatística** é unha colección de métodos que emprega os datos observados para inferir algún coñecemento sobre distintas características da distribución da variable aleatoria que xerou eses datos. De xeito máis formal, a pregunta á que responde a inferencia estatística é: **dado un conxunto de observacións dunha variable aleatoria X con distribución descoñecida F, como podemos inferir algo sobre F?**



Tipos de métodos inferenciais:

- **Inferencia non paramétrica**: o obxectivo principal é facer inferencia sobre características xerais da distribución da variable de interese sen asumir ningún modelo particular.

  Exemplo. Dada unha variable aleatoria X con media $\mu = \mathrm{E}(X)$ e función de distribución F, podemos facer inferencia sobre μ, ou incluso sobre F, sen asumir ningún modelo específico para F.

- **Inferencia paramétrica**: o obxectivo é facer inferencia sobre un ou varios parámetros que caracterizan a distribución da variable de interese baixo a suposición de que esta segue algún modelo paramétrico.

  Exemplo. Supoñamos que a variable de interese X ten distribución $N(\mu, \sigma)$. Queremos facer inferencia sobre o cuantil de orde 90 % de X. Poderemos empregar dalgún xeito o feito de que X sexa Normal no proceso inferencial?

Metodoloxías inferenciais:

- **Estimación puntual** de cantidades numéricas (por exemplo os parámetros que caracterizan unha distribución, estimación de probabilidades, estimación de cuantís etc.) e estimación de curvas (estimación da función de distribución, da función de densidade, da curva ROC etc.).
- **Intervalos de confianza** (para parámetros ou outras cantidades numéricas) e bandas de confianza (para curvas).
- **Tests de hipóteses** (comparación de medias, comparación de varianzas, tests de bondade de axuste etc).

3.2. Algúns conceptos básicos da inferencia estatística

A **poboación** é o conxunto de obxectos ou individuos de interese no estudo estatístico. Dada unha poboación de obxectos ou individuos, o estudo centrarase nalgunha característica particular destes. Esta característica será unha variable aleatoria, X, á cal lle chamaremos **variable aleatoria poboacional**, ou simplemente **poboación**.

Exemplo. Estase a realizar un estudo sobre hipertensión en homes de máis de 65 anos sen patoloxías previas. En particular, deséxase estudar o comportamento da presión arterial sistólica. A variable aleatoria poboacional é X =“presión arterial sistólica (dun home de máis de 65 anos sen patoloxías previas)”. Obviamente a presión arterial pode variar dun individuo a outro.

Na maior parte das ocasións, non poderemos observar a todos os individuos da poboación. No seu lugar, traballaremos cunha **mostra**, que é un subconxunto pequeno de individuos da poboación. A mostra debe ser seleccionada adecuadamente por algún **método de mostraxe**. O número de observacións na mostra chámase **tamaño mostral**.

Obviamente, non poderemos medir a presión arterial de *todos* os homes de máis de 65 anos. No seu lugar, selecciónase unha mostra de 250 persoas, convenientemente seleccionadas entre o grupo de homes maiores de 65 anos sen patoloxías previas.

Cada un dos elementos da mostra proporcionará unha observación da variable aleatoria poboacional X. Se o tamaño mostral é n, entón a mostra é a colección de variables aleatorias $X_1, X_2, \dots, X_n$. Os **datos** son as realizacións numéricas destas variables aleatorias. Neste curso asumiremos que as observacións son variables aleatorias independentes. Isto dá lugar a unha **mostra aleatoria simple**. Tamén se di que as variables que forman a mostra son **independentes e identicamente distribuídas** (i.i.d.).

Os valores da presión arterial sistólica recollidos sobre os 250 homes da mostra son:

$$154, 130, 158, 142, 125, \dots, 136, 164$$

Desde un punto de vista estatístico, interesaranos algunha característica particular da variable aleatoria poboacional X (por exemplo, a súa media, a súa varianza, un cuantil, unha certa probabilidade etc.)

Neste estudo sobre hipertensión interésanos facer inferencia sobre o cuantil de orde 0.90 da presión arterial sistólica.

A característica de interese obxecto de estudo chámase **parámetro**. O parámetro está relacionado coa distribución de X.

Neste caso o parámetro é o cuantil 0.90 de X, é dicir, $F^{-1}(0.90)$, onde F^{-1} é a función cuantil de X.

Para facer inferencia sobre o parámetro empregaremos algún cálculo sobre os datos que obtivemos na mostra. Un **estatístico** é calquera función dos datos $T = T(X_1, \dots, X_n)$. Como T é unha función de variables aleatorias, T **tamén será unha variable aleatoria** coa súa propia distribución (denominada **distribución na mostraxe** de T), a súa propia media, a súa varianza, os seus cuantís etc.

Por exemplo, para estimar o cuantil 0.90 poboacional parece bastante razoable empregar o cuantil mostral. Entón o cuantil mostral de orde 0.90 (que podemos calcular a partir da mostra) será o estatístico empregado para estimar o correspondente cuantil poboacional (que é unha cantidade descoñecida). □

Os estatísticos empréganse con dous obxectivos:

(a) Para estimar un parámetro. Neste caso, o estatístico recibe o nome de **estimador**. Algúns exemplos importantes son:

- a media mostral, que permite estimar a media poboacional;
- a varianza mostral e a desviación estándar mostral, que permiten estimar a varianza poboacional e a desviación estándar poboacional, respectivamente;
- os cuantís mostrais, que permiten estimar os cuantís poboacionais;
- as proporcións mostrais ou frecuencias, que permiten estimar probabilidades;
- etc.

(b) Para levar a cabo outros procedementos inferenciais, como por exemplo a construción de **intervalos de confianza** ou **tests de hipóteses**.

3.3. Estimación puntual

Supoñamos que estamos interesados en facer inferencia sobre un certo **parámetro** θ relacionado coa distribución da variable aleatoria poboacional X. A **estimación puntual** consiste en dar un valor numérico para o parámetro a través dun estimador calculado cos datos obtidos a partir da mostra. O estimador é un estatístico (é dicir, unha función dos datos) especialmente deseñado para estimar o parámetro de interese.

En moitas ocasións poden existir varios posibles estimadores para un mesmo parámetro. Nese caso teremos necesidade de comparar os estimadores para decidir cal deles ten un mellor comportamento na práctica. Para iso establecemos unha serie de propiedades teóricas que nos permiten estudar a calidade dos estimadores. En xeral, estas propiedades poden comprobarse sen necesidade de coñecer o verdadeiro valor do parámetro.

Denotamos por θ o parámetro poboacional que queremos estimar. Supoñamos que dispoñemos dunha mostra aleatoria simple $X_1, X_2, \ldots, X_n$ de X. Denotamos o estimador de θ por $\hat{\theta}_n$ (enfatizamos así que o estimador depende do tamaño mostral).

Exemplo. No estudo sobre a hipertensión, o parámetro de interese é o cuantil 0.90 da distribución da variable aleatoria poboacional "presión arterial sistólica". Como xa dixemos, un estimador razoable podería ser o cuantil mostral de orde 0.90. Pero será este un bo estimador para o parámetro de interese? □

Innesgadez e innesgadez asintótica. O **nesgo**[1] de $\hat{\theta}_n$ como estimador do parámetro θ é

$$\textsc{Bias}(\hat{\theta}_n) = \mathrm{E}(\hat{\theta}_n) - \theta.$$

Un estimador é **innesgado** se $\textsc{Bias}(\hat{\theta}_n) = 0$, é dicir, se $\mathrm{E}(\hat{\theta}_n) = \theta$. Isto significa que, como variable aleatoria, o estimador está centrado no verdadeiro valor do parámetro.

[1] En inglés, *bias*.

Nalgunhas ocasións, esixirlle a un estimador que sexa innesgado resulta moi restritivo, por iso se pode relaxar a propiedade a que o estimador sexa **asintoticamente innesgado**, que é aquel que cumpre

$$\lim_{n\to\infty} \text{BIAS}(\hat{\theta}_n) = 0.$$

Obviamente, todo estimador innesgado tamén será asintoticamente innesgado.

Consistencia. O erro cadrático medio de $\hat{\theta}_n$ como estimador de θ é

$$\text{ECM}(\hat{\theta}_n) = \text{E}\left[(\hat{\theta}_n - \theta)^2\right] = \text{VAR}(\hat{\theta}_n) + \left[\text{BIAS}(\hat{\theta}_n)\right]^2.$$

O estimador $\hat{\theta}_n$ é **consistente** se

$$\lim_{n\to\infty} \text{ECM}(\hat{\theta}_n) = 0.$$

Á vista da definición do erro cadrático medio, para que un estimador sexa consistente terá que ser asintóticamente innesgado e tal que a súa varianza se faga pequena ao aumentar o tamaño mostral.

O erro cadrático medio mide a distancia entre o estimador e o parámetro, polo tanto valores pequenos do erro cadrático medio significan que o estimador é moi preciso. Cando dispoñemos de varios estimadores para un mesmo parámetro, podemos comparalos a través dos seus erros cadráticos medios. Máis formalmente, sexan $\hat{\theta}_1$ e $\hat{\theta}_2$ dous estimatores do parámetro θ. O estimador $\hat{\theta}_1$ é **máis eficiente** ca $\hat{\theta}_2$ se $\text{ECM}(\hat{\theta}_1) < \text{ECM}(\hat{\theta}_2)$, ou, equivalentemente, $\text{ECM}(\hat{\theta}_1)/\text{ECM}(\hat{\theta}_2) < 1$. O cociente $\text{ECM}(\hat{\theta}_1)/\text{ECM}(\hat{\theta}_2)$ chámase **eficiencia relativa**.

Erro estándar e normalidade asintótica. O **erro estándar** dun estimador é a súa desviación estándar

$$\text{SE}(\hat{\theta}_n) = +\sqrt{\text{VAR}(\hat{\theta}_n)}.$$

Esta cantidade xoga un papel fundamental na inferencia estatística. Normalmente será descoñecida e polo tanto tamén terá que ser estimada.

Un estimador $\hat{\theta}_n$ é **asintoticamente Normal** se

$$\frac{\hat{\theta}_n - \theta}{\text{SE}(\hat{\theta}_n)} \quad \text{é aproximadamente } N(0,1).$$

3.3.1. Estimación da media: a media mostral

Exemplo. Consideremos o conxunto de datos DIABETES. Queremos estimar a media da concentración de glucosa (variable *glu*) no grupo de mulleres non diabéticas. Sexa X a variable aleatoria poboacional "concentración de glucosa dunha muller non diabética".

```
> diabetes <- read.table(file="datos-diabetes.txt",header=TRUE)
> attach(diabetes)
```

Seleccionemos a nosa mostra:

```
> mostra <- glu[type=="No"]
> n <- length(mostra)
```

O tamaño mostral é $n = 132$. O parámetro que queremos estimar é $\mu = \mathrm{E}(X)$. O estimador natural de μ é a media mostral, que en R se calcula coa función `mean()`:

```
> mean(mostra)

[1] 113.1061
```

Pero terá este estimador boas propiedades teóricas? Vexámolo a continuación. □

Supoñamos que $X_1, X_2, \ldots, X_n$ é unha mostra aleatoria simple de X. Supoñamos que $\mu = \mathrm{E}(X)$ é o parámetro que queremos estimar. Ademais, supoñamos que $\mathrm{Var}(X) = \sigma^2$. A **media mostral** é

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

A media mostral ten as seguintes propiedades:

- $\hat{\mu} = \bar{X}$ **é un estimador innesgado para** μ, xa que $\mathrm{E}(\bar{X}) = \mu$. Podemos demostralo facilmente:

$$\mathrm{E}(\bar{X}) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \overset{(a)}{=} \frac{1}{n}\mathrm{E}\left(\sum_{i=1}^{n} X_i\right) \overset{(b)}{=} \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(X_i) \overset{(c)}{=} \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{n\mu}{n} = \mu.$$

  En (a) empregamos que as constantes multiplicativas poden sacarse da esperanza; en (b) empregamos que a esperanza da suma é a suma das esperanzas; en (c) empregamos que a mostra está formada por variables aleatorias identicamente distribuídas e polo tanto todas teñen esperanza μ.

- $\hat{\mu} = \bar{X}$ **é un estimador consistente para** μ, xa que, como acabamos de ver, é innesgado e ademais a súa varianza é

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \overset{(a)}{=} \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) \overset{(b)}{=} \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) \overset{(c)}{=} \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

  En (a) empregamos que as constantes multiplicativas saen da varianza ao cadrado; en (b) empregamos que a varianza da suma é a suma das varianzas porque as variables son independentes (recordemos que estamos traballando cunha mostra aleatoria simple); en (c) empregamos que a mostra está formada por variables aleatorias identicamente distribuídas e polo tanto todas teñen a mesma varianza, que á súa vez coincide con σ^2. Concluímos así que a varianza de $\bar{X}$ vai a cero cando $n \to \infty$:

$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n} \overset{n\to\infty}{\longrightarrow} 0.$$

  Ademais, de aquí tamén deducimos que o **erro estándar da media mostral** é

$$\mathrm{se}(\bar{X}) = \frac{\sigma}{\sqrt{n}},$$

  onde σ é a desviación estándar poboacional. O erro estándar de $\bar{X}$ depende polo tanto da desviación estándar poboacional, que tamén terá que ser estimada na práctica.

- A **distribución na mostraxe de $\bar{X}$** depende da distribución da variable poboacional X. Non obstante, como a media mostral é esencialmente unha suma de variables aleatorias, o Teorema Central do Límite garante que a distribución de $\bar{X}$ sempre pode ser aproximada por unha distribución Normal. A aproximación vén dada por

$$\bar{X} \text{ é aproximadamente } N\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

  Equivalentemente, podemos reescribir a expresión anterior como

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \frac{\bar{X}-\mu}{\text{SE}(\bar{X})} \quad \text{é aproximadamente } N(0,1),$$

  é dicir, $\bar{X}$ é **asintoticamente Normal**.

- Cando dispoñemos de información adicional sobre a variable poboacional X, entón poderemos dar máis detalles sobre a distribución na mostraxe de $\bar{X}$.

  O caso máis importante é cando a poboación X é **Normal**. Supoñamos $X \sim N(\mu, \sigma)$. Debido á reprodutividade da Normal, inmediatamente obtemos que $\bar{X}$ tamén é unha variable aleatoria Normal(ollo! exactamente Normal, non só aproximadamente coma no apartado anterior):

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \text{ ou, equivalentemente, } \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \frac{\bar{X}-\mu}{\text{SE}(\bar{X})} \sim N(0,1).$$

3.3.2. Estudos de Monte Carlo

Os estudos de Monte Carlo son estudos de simulación que se empregan para analizar o comportamento práctico dos estimadores ou doutros procedementos estatísticos. O algoritmo básico dun estudo de Monte Carlo é o seguinte:

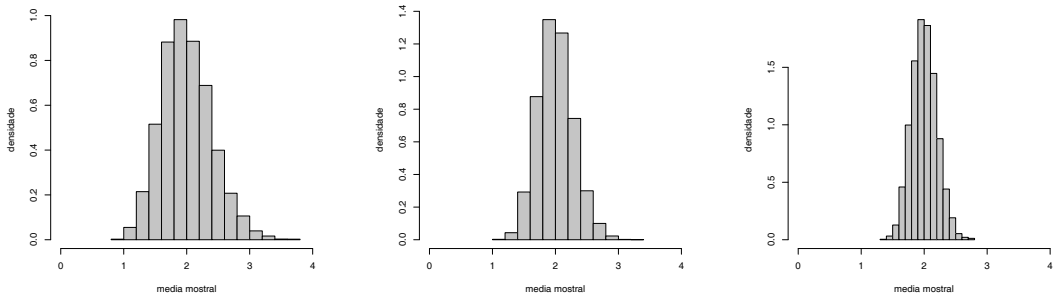
- Repetir os seguintes pasos un número grande de veces (por exemplo 10.000 veces):
  1. Simular unha mostra dunha distribución coñecida.
  2. Calcular o estimador na mostra simulada e gardar o seu valor.
- Analizar os valores gardados no paso 2: aproximar o nesgo, a varianza, o erro cadrático medio, a distribución na mostraxe etc.

Exemplo. A Táboa 3.1 recolle aproximacións do nesgo (Bias), varianza (Var) e erro cadrático medio (ecm) da media mostral como estimador da media poboacional cando $X \sim Exponencial(1/\mu)$ obtidas a partir de 10.000 simulacións de Monte Carlo. O verdadeiro valor do parámetro fixouse en $\mu = 2$. Xa sabemos que o estimador é innesgado, por isto observamos que o nesgo aproximado (columna Bias na táboa) é despreciable. A varianza e o erro cadrático medio diminúen ao aumentar o tamaño mostral n, como xa sabemos que ten que ocorrer porque o estimador é consistente. Podemos comparar os valores aproximados por Monte Carlo (parte esquerda da táboa) cos valores teóricos $\text{Var}(X)/n = \mu^2/n$ (parte dereita da táboa).

Na Figura 3.1 podemos comprobar como é a distribución na mostraxe do estimador: debido á normalidade asintótica do estimador, ao aumentar o tamaño mostral a distribución parécese cada vez máis á dunha Normal. Nótese tamén que o soporte do estimador se vai concentrando cada vez máis ao redor de 2, que é o verdadeiro valor do parámetro na nosa simulación. □

Táboa 3.1: Nesgo (BIAS), varianza (VAR) e erro cadrático medio (ECM) da media mostral aproximados mediante 10.000 simulacións de Monte Carlo cando os datos proceden dunha *Exponencial*(1/2).

	Aproximación Monte Carlo			Valores teóricos		
n	BIAS	VAR	ECM	BIAS	VAR	ECM
25	−0.0037	0.1612	0.1612	0	0.1600	0.1600
50	−0.0006	0.0795	0.0795	0	0.0800	0.0800
100	0.0001	0.0407	0.0407	0	0.0400	0.0400



$n = 25$ $n = 50$ $n = 100$

Figura 3.1: Histogramas obtidos a partir de 10.000 simulacións de Monte Carlo da media mostral de mostras de tamaño n dunha *Exponencial*(1/2).

3.3.3. Estimación da varianza poboacional: a varianza mostral

Supoñamos que temos unha mostra aleatoria simple $X_1, X_2, \ldots, X_n$ dunha variable aleatoria X. Queremos estimar o parámetro $\sigma^2 = \text{VAR}(X)$. O estimador natural de σ^2 é a **varianza mostral**

$$\hat{\sigma}^2 = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

que é innesgado e consistente[2]. O estimador da desviación estándar poboacional, σ, será polo tanto

$$\hat{\sigma} = S = +\sqrt{S^2},$$

que se denomina **desviación estándar mostral**. Este estimador é asintoticamente innesgado e consistente. En R, as funcións `var()` e `sd()` calculan S^2 e S, respectivamente.

Un uso práctico fundamental de S é a **estimación do erro estándar da media mostral**. Recordemos que o erro estándar da media mostral era $\text{SE}(\bar{X}) = \sigma/\sqrt{n}$, que se estima por

$$\widehat{\text{SE}}(\bar{X}) = \frac{S}{\sqrt{n}}.$$

Exemplo. (cont., datos DIABETES) O estimador do erro estándar da media mostral no exemplo anterior é

```
> se <- sd(mostra)/sqrt(n)
> se

[1] 2.318505
```

□

3.3.4. Estimación dunha proporción: a proporción mostral

Supoñamos unha poboación na que unha proporción p de individuos presenta unha determinada característica. Desexamos estimar esta proporción. Para iso tomamos unha mostra de tamaño n e contamos o número de elementos da mostra que presentan a característica de interese. Denotemos por K esta cantidade. Claramente K é unha variable aleatoria $Binomial(n, p)$. Un bo estimador para p é a **proporción mostral** de elementos da mostra que presentan a característica:

$$\hat{p} = \frac{\text{número de elementos da mostra que presentan a característica de interese}}{n} = \frac{K}{n}.$$

Nótese que $n\hat{p} = K \sim Binomial(n, p)$. Recordemos que a media e a varianza dunha $Binomial(n, p)$ son np e $np(1-p)$, respectivamente. Tendo isto en conta é doado demostrar que $\hat{p}$ é innesgado, consistente e asintoticamente Normal:

- A media de $\hat{p}$ é

$$\text{E}(\hat{p}) = \text{E}\left(\frac{n\hat{p}}{n}\right) = \frac{1}{n}\text{E}(K) = \frac{np}{n} = p.$$

  Polo tanto $\hat{p}$ é un estimador innesgado para estimar p.

[2]En moitos libros S^2 denomínase "cuasivarianza mostral" para distinguilo do estatístico $s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_i)^2$, que denominan "varianza mostral". Para tamaños mostrais non demasiado pequenos, S^2 e s^2 son practicamente iguais. Obviamente $(n-1)S^2 = ns^2$. A vantaxe fundamental de S^2 sobre s^2 é que S^2 é innesgado, mentras que s^2 é só asintoticamente innesgado.

Exercicio 3.1. ***Comparación de estimadores mediante un estudo de Monte Carlo.*** *Supoñamos que $X \sim N(\mu, \sigma)$ e que o parámetro de interese é μ. Como sabemos, μ é a media de X, polo tanto un estimador axeitado para μ é a media mostral $\bar{X}$. Pero μ tamén é a mediana de X, xa que a densidade de X é simétrica, logo tamén parece razoable estimar μ mediante a mediana mostral.*

(a) Simula datos dunha distribución $N(\mu, \sigma)$ mediante a función `rnorm(n,mean=mu,sd=sigma)`*, onde* `n` *é o tamaño mostral (digamos, por exemplo, 25),* `mu` *é o verdadeiro valor do parámetro (digamos, por exemplo, 7) e* `sigma` *é a desviación estándar (digamos, por exemplo, 1). Estima μ mediante a media mostral (función* `mean()`*) e mediante a mediana mostral (función* `median()`*) e garda os valores obtidos. Repite 10.000 veces e aproxima o nesgo, a varianza e o erro cadrático medio. Enche a seguinte táboa:*

	Media mostral			Mediana mostral		
n	BIAS	VAR	ECM	BIAS	VAR	ECM
25						
50						
100						
200						

(b) É a mediana mostral un estimador innesgado? E consistente? Por que?

(c) Para cada tamaño mostral, calcula a eficiencia relativa entre os dous estimadores. Que conclusión sacas? Que estimador ten un mellor comportamento práctico?

- A varianza de $\hat{p}$ é

$$\text{VAR}(\hat{p}) = \text{VAR}\left(\frac{n\hat{p}}{n}\right) = \frac{1}{n^2}\text{VAR}(K) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n},$$

que converxe a 0 cando $n \to \infty$ porque p é unha constante. Polo tanto, $\hat{p}$ é consistente para estimar p.

Nótese que, dado que $p \in [0, 1]$, a varianza de $\hat{p}$ pode acotarse por

$$\text{VAR}(\hat{p}) = \frac{p(1-p)}{n} \leq \frac{0.25}{n},$$

que non depende do verdadeiro valor poboacional de p.

- O erro estándar de $\hat{p}$ é

$$\text{SE}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}},$$

que pode estimarse mediante

$$\widehat{\text{SE}}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Tendo en conta a acotación da varianza de $\hat{p}$, o erro estándar tamén se pode acotar por

$$\text{SE}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}} \leq \sqrt{\frac{0.25}{n}} = \frac{0.5}{\sqrt{n}}.$$

- Sabemos que $n\hat{p} \sim Binomial(n, p)$. Debido á aproximación da Binomial pola Normal, temos que $n\hat{p}$ é aproximadamente $N(np, \sqrt{np(1-p)})$, ou equivalentemente

$$\frac{n\hat{p} - np}{\sqrt{np(1-p)}} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{\hat{p} - p}{\text{SE}(\hat{p})} \quad \text{é aproximadamente } N(0,1),$$

  o cal significa que $\hat{p}$ é asintoticamente Normal.

Exemplo. (cont., datos DIABETES) Supoñamos agora que queremos estimar a proporción de mulleres non diabéticas que presentan unha concentración de glucosa maior de 150, é dicir, se X representa a concentración de glucosa dunha muller non diabética, entón queremos estimar $p = P(X > 150)$. O que faremos é comprobar cantas observacións son maiores ca 150 e calcularemos a correpondente proporción mostral:

```
> K <- sum(mostra > 150)
> p.estimado <- K/n
> round(p.estimado,digits=3)  # a función "round" serve para redondear
```

```
[1] 0.098
```

Tamén podemos facer simplemente

```
> p.estimado <- mean(mostra > 150)
> round(p.estimado,digits=3)
```

```
[1] 0.098
```

Estimamos nun 9.8 % a porcentaxe de mulleres non diabéticas que presentan unha concentración de glucosa maior de 150. O estimador do erro estándar de $\hat{p}$ é

```
> se.p.estimado <- sqrt(p.estimado*(1-p.estimado))/sqrt(n)
> round(se.p.estimado,digits=3)
```

```
[1] 0.026
```

Compárese este valor coa cota $0.5/\sqrt{132} = 0.044$. □

Exercicio 3.2. *Calcula a probabilidade de que a proporción mostral $\hat{p}$ estea no intervalo $(p-0.1, p+0.1)$ cando...*

(a) $p = 0.2$ *e* $n = 50$*;* *(b)* $p = 0.2$ *e* $n = 100$*;*

(c) $p = 0.5$ *e* $n = 50$*;* *(d)* $p = 0.5$ *e* $n = 100$*;*

(e) $p = 0.8$ *e* $n = 50$*;* *(f)* $p = 0.8$ *e* $n = 100$*.*

Atopa as probabilidades baseándote na distribución Binomial de $n\hat{p}$ e na súa aproximación pola Normal. Comenta os resultados.

3.3.5. Estimación dun cuantil: o cuantil mostral

Supoñamos agora que queremos estimar o cuantil de orde p da variable X, é dicir, o parámetro de interese é $x_p = F^{-1}(p)$. Un estimador razoable é o correspondente **cuantil mostral de orde** p, que en R se calcula coa función `quantile(mostra,probs=p)`. Pódese demostrar que o cuantil mostral é asintoticamente innesgado e consistente (as propiedades teóricas do cuantil mostral son máis complicadas de obter ca as dos estimadores anteriores, así que non veremos os detalles aquí).

Exemplo. (cont., datos DIABETES) Supoñamos que queremos estimar o cuantil de orde 0.80 da concentración de glucosa dunha muller non diabética. En R escribimos `quantile(mostra, probs=0.80)`, que resulta 136.8. □

3.3.6. Métodos de construción de estimadores

Existen diversos métodos de construción de estimadores. Os máis relevantes son o **Método dos Momentos** e o **Método de Máxima Verosimilitude**. Non entraremos nos detalles de ningún deles. Simplemente veremos un exemplo no que de xeito natural poden xurdir varios estimadores razoables para un mesmo parámetro e faremos un estudo de Monte Carlo para comparalos.

Exemplo. Estimadores do parámetro dunha $Uniforme[0,\theta]$**.** Supoñamos que a variable poboacional X ten distribución $Uniforme[0,\theta]$ e que dispoñemos dunha mostra aleatoria simple $X_1, X_2, \dots, X_n$ de X. O parámetro de interese é θ, extremo superior do soporte de X. Podemos pensar en dous estimadores para θ que se constrúen de forma completamente distinta:

- Tendo en conta que $\mathrm{E}(X) = \theta/2$ e que $\bar{X}$ é un estimador consistente para a media, entón parece bastante razoable considerar o dobre da media mostral como estimador de θ, é dicir, $\hat{\theta}_1 = 2\bar{X}$. Este estimador é o que se obtén polo método dos momentos. Facendo uso das propiedades teóricas da media mostral, é doado ver este estimador é innesgado, consistente e asintoticamente Normal (**exercicio**).

- Tendo en conta que θ é o extremo superior do soporte de X, tamén parece razoable considerar o máximo dos valores observados como estimador de θ, é dicir $\hat{\theta}_2 = \text{máx}\{X_1, \dots, X_n\}$.

Este é o estimador que se obtén mediante o método de máxima verosimilitude. As propiedades teóricas deste estimador son máis complicadas de obter, por iso nos será útil facer un estudo de simulación.

O seguinte esquema ilustra o funcionamento destes dous estimadores. Os datos están representados mediante puntos negros. A media mostral está marcada cunha aspa. O estimador $\hat{\theta}_1$ é o dobre da media mostral. O estimador $\hat{\theta}_2$ é o máximo das observacións.

Compararemos os dous estimadores mediante un pequeno estudo de Monte Carlo. Para iso fixamos por exemplo $\theta = 2$ e simulamos 10.000 mostras dunha variable aleatoria $Uniforme[0, 2]$ (en R: `runif(n,min=0,max=2)`). A Táboa 3.2 recolle os resultados obtidos.

Táboa 3.2: Resultados do estudo de Monte Carlo para comparar os estimadores $\hat{\theta}_1$ e $\hat{\theta}_2$.

	$\hat{\theta}_1$			$\hat{\theta}_2$			
n	Bias	Var	ecm	Bias	Var	ecm	$\text{ecm}(\hat{\theta}_2)/\text{ecm}(\hat{\theta}_1)$
25	−0.0048	0.0545	0.0545	−0.0779	0.0056	0.0117	0.2140
50	0.0012	0.0264	0.0264	−0.0398	0.0015	0.0031	0.1174
100	0.0005	0.0134	0.0134	−0.0198	0.0004	0.0008	0.0590

Xa sabemos que $\hat{\theta}_1$ é innesgado, cousa que se reflicte nos valores case despreciables na aproximación do nesgo. Do estimador $\hat{\theta}_2$ non sabemos se é innesgado ou non. O seu nesgo aproximado proporciona valores negativos (por que?) e maiores en valor absoluto ca os de $\hat{\theta}_1$, se ben fanse cada vez máis pequenos ao aumentar n. Parece polo tanto que $\hat{\theta}_2$ é lixeiramente nesgado pero asintoticamente innesgado (de feito pódese demostrar matematicamente que $\text{E}(\hat{\theta}_2) = \frac{n}{n+1}\theta$).

Por outra parte, os dous estimadores son consistentes, xa que os seus erros cadráticos medios se fan cada vez máis pequenos ao aumentar n. Non obstante hai unha gran diferenza no seu comportamento práctico, tal e como se pode comprobar na última columna, que recolle a eficiencia relativa (é dicir, o cociente entre os erros cadráticos medios). Como podemos comprobar, $\hat{\theta}_2$ ten un erro cadrático medio substancialmente menor ca o de $\hat{\theta}_1$. Isto quere dicir que o estimador $\hat{\theta}_2$ emprega a información da mostra dun xeito moito máis eficiente. Por exemplo, cando $n = 25$, $\hat{\theta}_2$ é case 5 veces máis eficiente ca $\hat{\theta}_1$. □

3.3.7. Estimación da función de distribución e da función de densidade

No capítulo anterior vimos que a función de distribución e a función de densidade son moi importantes na análise dunha variable aleatoria. Estas funcións poden estimarse de forma global.

Exercicio 3.3. ***Comparación de estimadores paramétricos e non paramétricos.*** *Sexa X unha variable aleatoria con función de distribución F. Queremos estimar o cuantil de orde* 0.95 *de X. Denotemos por $\tau = F^{-1}(0.95)$ o parámetro de interese. Podemos pensar en dous estimadores para τ:*

- *O cuantil mostral de orde* 0.95 *(en R* `quantile(mostra,0.95)`*), que denotamos por $\hat{\tau}_1$. Este é un* **estimador non paramétrico**, *xa que non depende de ningún modelo particular para a distribución de X.*
- *Tamén podemos pensar nun estimador paramétrico. Supoñamos que $X \sim N(\mu, \sigma)$. Tendo en conta que o cuantil de orde* 0.95 *de X é $\mu + 1.645\ \sigma$ (recorda que* 1.645 *é o cuantil* 0.95 *da Normal estándar), parece bastante razoable estimar τ mediante $\hat{\tau}_2 = \bar{X} + 1.645\ S$. En R:* `mean(mostra)+qnorm(0.95)*sd(mostra)`.

Un estudo de Monte Carlo baseado en 10.000 mostras proporcionou os resultados da táboa que aparece abaixo. As mostras obtivéronse de dúas distribucións: (a) $N(7, 2)$, polo tanto cúmprese o modelo paramétrico; e (b) Exponencial(0.5)*, que é unha distribución completamente distinta da Normal.*

		$\hat{\tau}_1$ (non paramétrico)			$\hat{\tau}_2$ (paramétrico)		
distribución de X	n	BIAS	VAR	ECM	BIAS	VAR	ECM
$N(7,2)$	50	-0.1786	0.2828	0.3147	-0.0175	0.1886	0.1889
	100	-0.1021	0.1627	0.1731	-0.0119	0.0955	0.0956
	200	-0.0477	0.0847	0.0869	-0.0026	0.0466	0.0466
	400	-0.0263	0.0441	0.0448	-0.0040	0.0239	0.0239
Exponencial(0.5)	50	-0.2929	1.1964	1.2821	-0.7548	0.7450	1.3147
	100	-0.1588	0.6857	0.7108	-0.7255	0.3777	0.9040
	200	-0.0779	0.3652	0.3713	-0.7149	0.1908	0.7019
	400	-0.0405	0.1807	0.1824	-0.7079	0.0945	0.5956

Na medida do que podemos ver no estudo de Monte Carlo:

(a) É o estimador non paramétrico innesgado nos dous escenarios simulados? E asintoticamente innesgado? E o paramétrico? Por que?

(b) No caso $X \sim$ Exponencial(0.5)*, $\bar{X}$ estima $E(X) = 2$ e S estima* SD$(X) = 2$*, polo tanto o estimador paramétrico $\hat{\tau}_2 = \bar{X} + 1.645S$ estimará $2 + 1.645 \cdot 2 = 5.29$. Coincide este valor co cuantil* 0.95 *da Exponencial*(0.5)*? Cal é a diferenza entre os entre os dous valores? Como se relaciona esta diferenza co nesgo?*

(c) Podemos concluír que os dous estimadores son consistentes nos dous escenarios?

(d) No caso da $N(7, 2)$, cal é a eficiencia relativa entre os dous estimadores para cada valor de n? Interpreta os resultados obtidos.

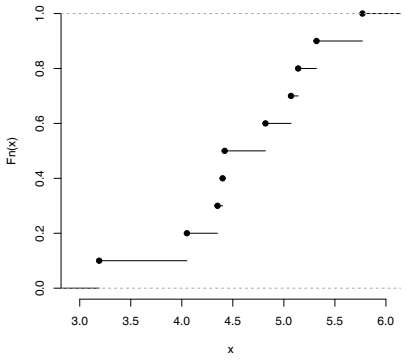


Figura 3.2: Función de distribución empírica da mostra 3.19, 4.05, 4.35, 4.40, 4.42, 4.82, 5.07, 5.14, 5.32, 5.77. O tamaño mostral é $n = 10$.

Tendo en conta que $F(x) = P(X \leq x)$ é unha probabilidade, entón podemos estimar $F(x)$ pola correspondente proporción mostral

$$\hat{F}(x) = \frac{\text{número de observacións} \leq x}{n}.$$

Se consideramos o estimador anterior como función de x, entón obtemos a **función de distribución empírica**, que formalmente se escribe como

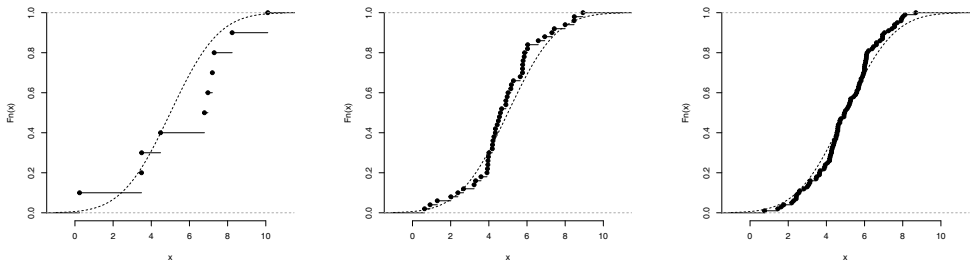
$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{I}(X_i \leq x),$$

onde $\mathrm{I}(\cdot)$ é a función indicadora, é dicir, $\mathrm{I}(X_i \leq x) = 1$ se $X_i \leq x$ e $\mathrm{I}(X_i \leq x) = 0$ se $X_i > x$. Esta función é un estimador non paramétrico global da función de distribución. A súa peculiaridade máis salientable é que se trata dunha función constante por pedazos. Os saltos prodúcense en cada unha das observacións e a magnitude de cada salto é $1/n$. A Figura 3.2 amosa a función de distribución empírica da mostra 3.19, 4.05, 4.35, 4.40, 4.42, 4.82, 5.07, 5.14, 5.32, 5.77. En R obtense coa función `ecfd()` (do inglés *empirical cumulative distribution function*):

```
> x <- c(3.19,4.05,4.35,4.40,4.42,4.82,5.07,5.14,5.32,5.77)
> plot(ecdf(x))
```

Ao aumentar o tamaño mostral, a función de distribución empírica acércase cada vez máis á función de distribución poboacional, tal e como se ilustra na Figura 3.3.

Para estimar a función de densidade podemos empregar o **histograma**, que tamén é un estimador non paramétrico. En R, a función `hist()` permite graficar o histograma (para obter o estimador da densidade debe empregarse o argumento `freq=FALSE`). A construción do histograma require escoller o número de intervalos ou a lonxitude de cada un deles. R fai esta selección de xeito automático.



$n = 10$ $n = 50$ $n = 100$

Figura 3.3: Función de distribución empírica de mostras de tamaños $n = 10, 50$ e 100 obtidas dunha $N(5, 2)$. A liña punteada representa a verdadeira función de distribución poboacional.

Unha alternativa ao histograma son os estimadores da densidade baseados en **técnicas de suavización**. Un estimador desta clase amplamente empregado na práctica é o **estimador tipo núcleo**, que é

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

onde

- K é a función núcleo (en inglés, *kernel*), que é unha función de densidade simétrica respecto do cero coñecida. Unha función moi empregada é o núcleo de Epanechnikov $K(u) = 0.75(1 - u^2)\mathrm{I}(-1 < u < 1)$.
- h é un número positivo chamado parámetro de suavizado (en inglés, *bandwidth*). O valor de h depende do tamaño mostral. En xeral, para ter un bo comportamento do estimador, h será pequeno para tamaños mostrais grandes e viceversa.

En estimación tipo núcleo, a selección do parámetro de suavizado é un problema moi importante e moi estudado. Nótese que, ao contrario do estimador tipo núcleo da función de densidade, a función de distribución empírica non requería a selección de ningún parámetro para a súa construción. En R, a función `density()` constrúe o estimador núcleo da densidade. O parámetro de suavizado pode ser escollido automaticamente. A Figura 3.4 amosa o histograma e o estimador tipo núcleo obtidos a partir dunha mostra simulada de tamaño 200.

Exemplo. (cont., datos DIABETES) A Figura 3.5 amosa a función de distribución empírica, o histograma e o estimador tipo núcleo obtidos a partir das observacións da concentración de glucosa do grupo de mulleres non diabéticas. A Figura 3.6 contén as estimacións tipo núcleo das funcións de densidade da concentración de glucosa nos grupos de mulleres non diabéticas e diabéticas. Podemos observar unha clara diferenza entre as dúas funcións. □

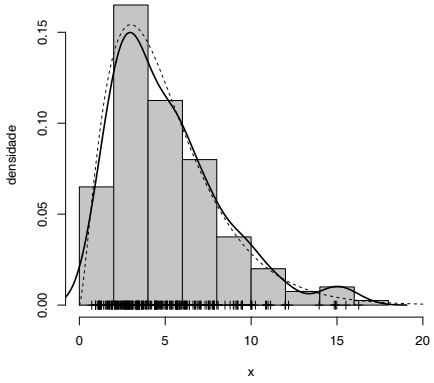
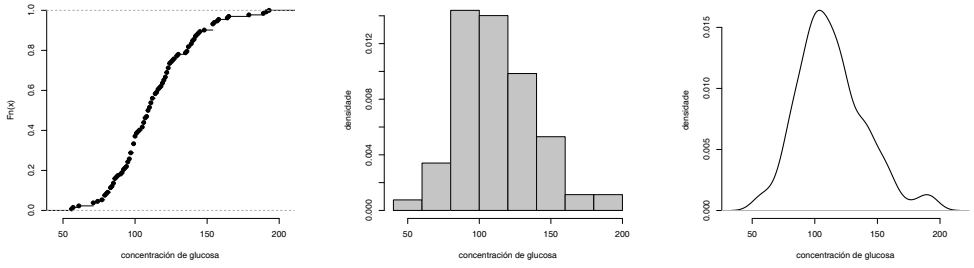


Figura 3.4: Histograma e estimador tipo núcleo (liña negra) baseados nunha mostra simulada de tamaño 200. A liña punteada é a verdadeira función de densidade.



(a) (b) (c)

Figura 3.5: Datos DIABETES. (a) Función de distribución empírica, (b) histograma e (c) estimador tipo núcleo da función de densidade obtidos a partir da mostra da concentración de glucosa das mulleres non diabéticas.

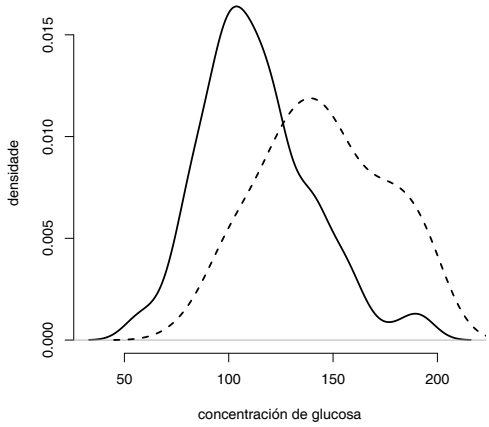


Figura 3.6: Datos DIABETES. Estimacións tipo núcleo das funcións de densidade da concentración de glucosa nas mulleres non diabéticas (en trazo continuo) e diabéticas (en trazo discontinuo).

3.3.8. Estimación da curva ROC

Empregando a función de distribución empírica tamén podemos estimar a curva ROC. Supoñamos que queremos construír a curva ROC baseada nas variables de diagnose Y_N na poboación Negativa e Y_P na poboación Positiva, que teñen funcións de distribución F_N e F_P, respectivamente. As correspondentes mostras son $Y_{N,1}, Y_{N,2}, \ldots, Y_{N,n_N}$ e $Y_{P,1}, Y_{P,2}, \ldots, Y_{P,n_P}$, que teñen tamaños mostrais n_N e n_P, respectivamente. Dado un punto de corte c, estimamos a fracción de falsos positivos como

$$\widehat{\text{FFP}}(c) = 1 - \hat{F}_N(c) = 1 - \frac{1}{n_N}\sum_{i=1}^{n_N} \text{I}(Y_{N,i} \leq c) = \frac{1}{n_N}\sum_{i=1}^{n_N} \text{I}(Y_{N,i} > c)$$

e a fracción de verdadeiros positivos como

$$\widehat{\text{FVP}}(c) = 1 - \hat{F}_P(c) = 1 - \frac{1}{n_P}\sum_{i=1}^{n_P} \text{I}(Y_{P,i} \leq c) = \frac{1}{n_P}\sum_{i=1}^{n_P} \text{I}(Y_{P,i} > c).$$

Tomando valores de c que percorran todo o soporte das variables de diagnose e representando os pares de puntos $(\widehat{\text{FFP}}(c), \widehat{\text{FVP}}(c))$ obtemos a **curva ROC empírica**. Aínda que estritamente falando a curva ROC empírica está formada por puntos aillados, acostuma representarse como unha función con forma de escaleira.

Tendo en conta que a curva ROC tamén se podía expresar como a función matemática $\text{ROC}(p) = 1 - F_P(F_N^{-1}(1-p))$, para $0 \leq p \leq 1$, tamén podemos construír a curva ROC empírica como

$$\widehat{\text{ROC}}(p) = 1 - \hat{F}_P(\hat{F}_N^{-1}(1-p)),$$

onde $\hat{F}_P(\cdot)$ é a función de distribución empírica construída coa mostra de Y_P e $\hat{F}_N^{-1}(1-p)$ é o cuantil mostral de orde $1-p$ da mostra de Y_N. Tomando valores de p entre 0 e 1 podemos estimar toda a función.

A curva ROC empírica pódese empregar para construír estimadores da área debaixo da curva (AUC), do índice de Youden e do punto de corte asociado.

Exemplo. Datos DIABETES. Consideremos que a concentración de glucosa é a variable de diagnose. Construiremos a curva ROC empírica para estudar a capacidade que ten esta variable para distinguir as poboacións de mulleres diabéticas e non diabéticas. Primeiro seleccionamos as mostras:

```
> mostra.Neg <- glu[type=="No"]
> mostra.Pos <- glu[type=="Yes"]
```

Para construír a curva ROC empírica co primeiro método construímos un vector de posibles puntos de corte entre o mínimo e máximo dos valores observados na variable *glu*

```
> c <- seq(min(glu),max(glu),by=0.01)
```

e estimamos as fraccións de falsos positivos e de verdadeiros positivos para cada valor do punto de corte

```
> FFP <- numeric(length(c))
> FVP <- numeric(length(c))
> for (i in 1:length(c)){
+         FFP[i] <- mean(mostra.Neg > c[i] )
+         FVP[i] <- mean(mostra.Pos > c[i] )
+         }
```

Finalmente facemos o gráfico

```
> plot(FFP,FVP,type="s",frame=FALSE)
> abline(a=0,b=1,lty=2)   # diagonal do cadrado
```

A opción `type="s"` na función `plot()` permite facer o gráfico da función con forma de escaleira.

Para construír a curva ROC empírica polo segundo método facemos o seguinte:

```
> p <- seq(0,1,by=0.001)
> ROC.empirica <- numeric(length(p))
> for (i in 1:length(p)){
+         ROC.empirica[i] <- 1 - mean(mostra.Pos <= quantile(mostra.Neg,1-p[i]))
+   }
> plot(p,ROC.empirica,type="s",ylab="ROC(p)",xlim=c(0,1),ylim=c(0,1))
> abline(a=0,b=1,lty=2)   # diagonal do cadrado
```

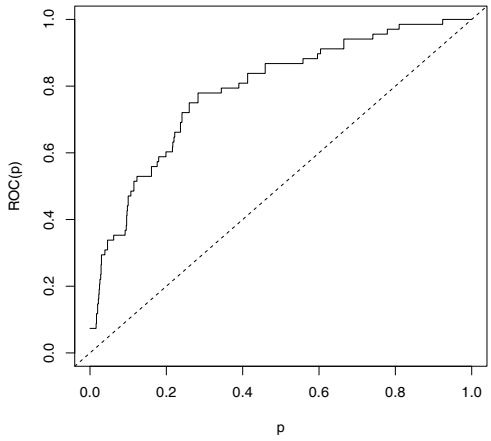


Figura 3.7: Conxunto de datos DIABETES. Curva ROC empírica baseada na concentración de glucosa para distinguir mulleres diabéticas e non diabéticas.

A Figura 3.7 contén a curva ROC empírica obtida co segundo método.

Para estimar a AUC podemos facer unha integración numérica, que resulta inmediata a partir dos valores gardados no vector `ROC.empirica`. Simplemente temos que sumalos e multiplicar pola separación que empregamos na construción do vector que contén os valores de p (0.001 no noso exemplo):

```
> AUC.estimada <- sum(ROC.empirica)*0.001
> AUC.estimada

[1] 0.7892794
```

Resulta unha AUC estimada de 0.789, o cal indica que a concentración de glucosa ten unha capacidade de discriminación bastante alta. □

Exercicio 3.4. *Como estimarías o índice de Youden e o punto de corte asociado a partir da curva ROC empírica? Obtén as estimacións no exemplo anterior. Cal é a sensibilidade e a especificidade para ese punto de corte?*

3.4. Intervalos de confianza

Habitualmente, a estimación puntual dun parámetro acompáñase dun intervalo confianza. Sexa X a variable aleatoria poboacional e sexa θ un parámetro relacionado coa distribución de X. Sexa $X_1, \dots, X_n$ unha mostra aleatoria simple de X. Denomínase **intervalo de confianza para θ con nivel de confianza $1-\alpha$** a un intervalo aleatorio (T_1, T_2) de forma que

$$P(T_1 \leq \theta \leq T_2) = 1 - \alpha,$$

onde $T_1 = T_1(X_1, \dots, X_n)$ e $T_2 = T_2(X_1, \dots, X_n)$ son dous estatísticos que se calculan a partir da mostra.

En moitas aplicacións prácticas non é posible construír intervalos de confianza exactos. No seu lugar, podemos considerar intervalos de confianza asintóticos (ou aproximados), que son aqueles que satisfán a condición anterior asintoticamente, é dicir,

$$\lim_{n\to\infty} P(T_1 \leq \theta \leq T_2) = 1 - \alpha.$$

Os valores do nivel de confianza $1-\alpha$ máis habitualmente usados na práctica son 0.95, 0.99 ou 0.90.

Exemplo. Supoñamos que $X \sim N(\mu, \sigma)$ e que o parámetro de interese é μ. Sexa $X_1, X_2, \dots, X_n$ unha mostra aleatoria simple de X. É ben coñecido que o estatístico

$$\sqrt{n}\frac{\bar{X} - \mu}{S}$$

segue unha distribución t de Student con $n-1$ graos de liberdade, t_{n-1}. Nótese que este estatístico involucra os datos (a través de n, $\bar{X}$ e S) e ao parámetro μ, e polo tanto non se pode observar o seu valor. Sorprendentemente, a súa distribución é completamente coñecida: unha t_{n-1}. Os estatísticos deste tipo chámanse **estatísticos pivotais** e resultan moi útiles para construír intervalos de confianza.

Supoñamos por exemplo que $n = 20$ e que queremos traballar cun nivel de confianza $1-\alpha = 0.95$. Sabemos que 2.093 é o cuantil 0.975 dunha t_{19} (abonda con facer `qt(0.975,df=19)` en R), e polo tanto $P(-2.093 \leq t_{19} \leq 2.093) = 0.95$, ou, para o noso estatístico

$$P\left(-2.093 \leq \sqrt{20}\frac{\bar{X} - \mu}{S} \leq 2.093\right) = 0.95.$$

Agora das desigualdades

$$-2.093 \leq \sqrt{20}\frac{\bar{X} - \mu}{S} \leq 2.093$$

podemos despexar μ:

$$-2.093\frac{S}{\sqrt{20}} \leq \bar{X} - \mu \leq 2.093\frac{S}{\sqrt{20}};$$

$$-\bar{X} - 2.093\frac{S}{\sqrt{20}} \leq -\mu \leq -\bar{X} + 2.093\frac{S}{\sqrt{20}};$$

e multiplicando por -1 (ollo coas desigualdades!) obtemos

$$\bar{X} + 2.093\frac{S}{\sqrt{20}} \geq \mu \geq \bar{X} - 2.093\frac{S}{\sqrt{20}}.$$

Entón, volvendo á probabilidade, temos que

$$P\left(\bar{X} - 2.093\frac{S}{\sqrt{20}} \leq \mu \leq \bar{X} + 2.093\frac{S}{\sqrt{20}}\right) = 0.95,$$

é dicir,

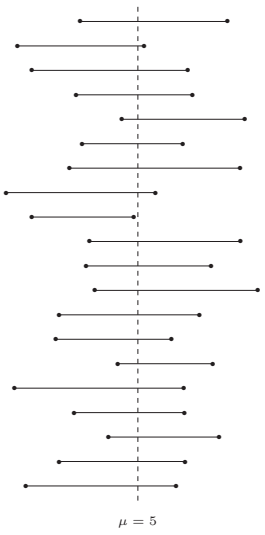
$$\left(\bar{X} - 2.093\frac{S}{\sqrt{20}}, \bar{X} + 2.093\frac{S}{\sqrt{20}}\right)$$

é un intervalo de confianza para μ con nivel de confianza 0.95.

Este intervalo constrúese de forma simétrica respecto da media mostral, abrindo á esquerda e á dereita unha amplitude de $2.093\frac{S}{\sqrt{n}}$. Na construción interveñen a media mostral, a desviación estándar mostral, o nivel de confianza (a través do cuantil da t) e o tamaño mostral.

Vexamos como funciona este intervalo de confianza na práctica. Para isto simulamos mostras de tamaño $n = 20$ dunha $N(5, 2)$. Neste caso, por tratarse dun exemplo simulado, sabemos que o verdadeiro valor de μ é 5 e podemos comprobar cantos destes intervalos "se equivocan" na súa estimación. No esquema da dereita aparecen 20 intervalos. Dos 20, 19 (o 95 %) acertaron e 1 intervalo (é dicir, o 5 %) fallou. □

	mostra	$\bar{X}$	S	intervalo
	1	5.22	2.14	(4.21,6.22)
	2	4.22	1.85	(3.36,5.09)
	3	4.62	2.27	(3.56,5.68)
	4	4.95	1.69	(4.16,5.75)
	5	5.62	1.79	(4.78,6.46)
	6	4.93	1.46	(4.24,5.61)
	7	5.23	2.48	(4.07,6.39)
	8	4.22	2.17	(3.21,5.24)
*	9	4.25	1.49	(3.55,4.95)
	10	5.37	2.20	(4.34,6.40)
	11	5.15	1.83	(4.29,6.00)
	12	5.52	2.38	(4.41,6.64)
	13	4.88	2.04	(3.93,5.84)
	14	4.67	1.69	(3.88,5.46)
	15	5.37	1.39	(4.72,6.02)
	16	4.47	2.47	(3.32,5.63)
	17	4.88	1.61	(4.13,5.64)
	18	5.35	1.61	(4.60,6.10)
	19	4.79	1.84	(3.93,5.65)
	20	4.50	2.19	(3.48,5.52)



Exercicio 3.5. *Deseña en R un estudo de Monte Carlo para reproducir a simulación do exemplo anterior con 1000 réplicas.*

Os intervalos de confianza serán máis informativos canto menor sexa a súa lonxitude. No exemplo anterior vimos que a **lonxitude** é $2 \cdot 2.093\frac{S}{\sqrt{n}}$, que depende do nivel de confianza (a través do cuantil da t), da desviación estándar poboacional (a través do seu estimador S) e do tamaño mostral (a través de $\sqrt{n}$). En xeral temos o seguinte comportamento:

- O aumento do nivel de confianza conleva un aumento na lonxitude do intervalo. Por exemplo, no caso anterior, o feito de pasar dun nivel 0.95 a un nivel 0.99 suporía cambiar o cuantil 2.093 da t por 2.861.

- S é un estimador consistente da desviación estándar poboacional, σ, así que S tomará valores cercanos a σ, que é fixa. Canto maior sexa σ, maior tenderá a ser S e polo tanto maior será a lonxitude do intervalo.

- Tendo en conta que a desviación estándar poboacional σ é fixa, dado un nivel de confianza, o único que está na nosa man para diminuír a lonxitude do intervalo será mediante o incremento do tamaño mostral n. Desafortunadamente, a lonxitude depende do tamaño mostral a través de $\sqrt{n}$. Isto implica que se, por exemplo, quixeramos reducir á metade a lonxitude do intervalo anterior, teriamos que multiplicar o tamaño mostral por 4 (é dicir, pasar de $n = 20$ a $n = 80$). Se quixeramos reducir a lonxitude á décima parte, teriamos que multiplicar o tamaño mostral por 100 (pasar de $n = 20$ a $n = 2000$, cousa que posiblemente non sexa factible na práctica).

3.4.1. O método pivotal

Sexa $X_1, \dots, X_n$ unha mostra aleatoria simple da variable aleatoria X. Denomínase **estatístico pivotal** a calquera estatístico $T(X_1, \dots, X_n; \theta)$ construído coa mostra e co parámetro θ, pero tal que a súa distribución na mostraxe non depende de θ e é completamente coñecida.

Os estatísticos pivotais úsanse para construír intervalos de confianza. O procedemento xeral para construír un intervalo de confianza para θ con nivel de confianza $1 - \alpha$ é o seguinte:

Paso 1. Elixir un estatístico pivotal para θ.

Paso 2. Buscar os valores a e b tales que

$$P(a \leq T(X_1, \dots, X_n; \theta) \leq b) = 1 - \alpha.$$

Paso 3. Despexar θ da expresión $a \leq (X_1, \dots, X_n; \theta) \leq b$. Dese xeito obtense directamente o intervalo de confianza para θ.

No Paso 2, os valores a e b poden obterse explicitamente porque a distribución de T é coñecida.

En principio habería infinitas posibles eleccións para a e b. Fixado o nivel de confianza $1-\alpha$, escóllense α_1 e α_2 tales que $\alpha_1 + \alpha_2 = \alpha$. Entón a é o cuantil de orde α_1 de T e b é o cuantil de orde $1 - \alpha_2$ de T, tal e como se ilustra na Figura 3.8.

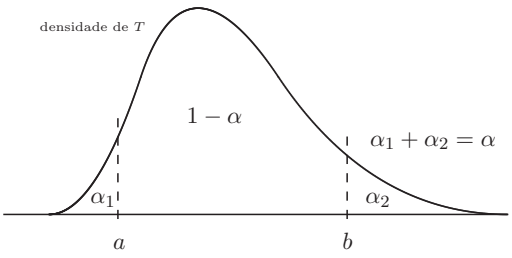


Figura 3.8: Ilustración do método pivotal.

Na práctica, interesa escoller a e b de forma que o intervalo de confianza teña lonxitude mínima. En xeral conseguir isto non é sinxelo, pero no caso en que a distribución do estatístico pivotal T sexa simétrica (por exemplo unha $N(0,1)$ ou unha t de Student), isto conséguese facilmente tomando $\alpha_1 = \alpha_2 = \alpha/2$ (deixando unha probabilidade $\alpha/2$ en cada cola da densidade). Neste caso a é o cuantil de orde $\alpha/2$ de T e b é o cuantil de orde $1-\alpha/2$ de T. Por comodidade, esta selección de a e b tamén se acostuma facer aínda cando a distribución de T non é simétrica.

3.4.2. Estatísticos pivotais en poboacións Normais

Cando $X \sim N(\mu, \sigma)$ temos estatísticos pivotais para os parámetros μ e σ.

Estatístico pivotal para μ:

$$\sqrt{n}\frac{\bar{X}-\mu}{S} \sim t_{n-1}.$$

Por exemplo, o intervalo de confianza para μ de nivel $1-\alpha = 0.95$ cando $n = 20$ é

$$\left(\bar{X} - 2.093\frac{S}{\sqrt{20}}, \bar{X} + 2.093\frac{S}{\sqrt{20}}\right),$$

porque 2.093 é o cuantil de orde 0.975 da t_{19}. Para outros tamaños mostrais hai que cambiar este valor adecuadamente.

Como sabemos, cando n é grande a distribución t_{n-1} parécese moito á $N(0,1)$ e o cuantil da t converxerá ao correspondente cuantil da $N(0,1)$. Polo tanto, para tamaños mostrais grandes (digamos por exemplo, $n \geq 100$), o intervalo de confianza de nivel 0.95 é

$$\left(\bar{X} - 1.96\frac{S}{\sqrt{n}}, \bar{X} + 1.96\frac{S}{\sqrt{n}}\right).$$

Estatístico pivotal para σ:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Por exemplo, o intervalo de confianza para σ de nivel $1-\alpha = 0.95$ cando $n = 20$ é

$$\left(\sqrt{\frac{19S^2}{32.85}}, \sqrt{\frac{19S^2}{8.91}}\right).$$

Nótese que neste caso a distribución do estatístico pivotal é unha chi-cadrado, que non é simétrica. Aínda así, empregamos os cuantís 0.025 e 0.975 igual que antes. En R `qchisq(0.025,df=19)` e `qchisq(0.975,df=19)`.

Exemplo. (cont., datos DIABETES) Á vista da Figura 3.6 parece razoable asumir que a concentración de glucosa nas mulleres non diabéticas segue unha distribución Normal. Para construír o intervalo de confianza para a media de nivel 0.95 facemos o seguinte:

```
> mostra <- glu[type=="No"]
> n <- length(mostra)
> LI <- mean(mostra)-qt(0.975,df=n-1)*sd(mostra)/sqrt(n)
> LS <- mean(mostra)+qt(0.975,df=n-1)*sd(mostra)/sqrt(n)
> c(LI,LS)

[1] 108.5195 117.6926
```

O estimador da media era $\bar{X} = 113.11$ e o correspondente intervalo de confianza de nivel 0.95 é $(108.52, 117.69)$. Se en vez do cuantil da t empregamos o cuantil da $N(0,1)$, obtemos o intervalo $(108.56, 117.65)$, que practicamente coincide co anterior. Isto débese a que o tamaño mostral é 132.

O intervalo de confianza para a desviación estándar de nivel 0.95 pode obterse da seguinte forma:

```
> LI <- sqrt((n-1)*var(mostra)/qchisq(0.975,df=n-1))
> LS <- sqrt((n-1)*var(mostra)/qchisq(0.025,df=n-1))
> c(LI,LS)

[1] 23.76562 30.30535
```

O estimador da desviación estándar era $S = 26.64$ e o intervalo de confianza de nivel 0.95 é $(23.77, 30.31)$. Obsérvese que o punto medio do intervalo, 27.04, non coincide co valor do estimador S. □

Exercicio 3.6. *Deduce a fórmula do intervalo de confianza para σ.*

3.4.3. Estatísticos pivotais asintóticos

Recordemos que un estatimador é asintoticamente Normal cando satisfai

$$\frac{\hat{\theta} - \theta}{\mathrm{SE}(\hat{\theta})} \quad \text{é aproximadamente } N(0,1).$$

Esta expresión pode empregarse como estatístico pivotal sempre que se dispoña dun estimador do erro estándar do estimador, digamos $\widehat{\mathrm{SE}}(\hat{\theta})$. Nese caso, o intervalo de confianza aproximado para θ de nivel 0.95 é

$$\left(\hat{\theta} - 1.96\ \widehat{\mathrm{SE}}(\hat{\theta}), \hat{\theta} + 1.96\ \widehat{\mathrm{SE}}(\hat{\theta})\right).$$

Os intervalos obtidos mediante este método funcionarán ben cando o tamaño mostral non sexa moi pequeno.

Intervalo aproximado para a media

Sexa X unha variable poboacional calquera (non necesariamente Normal) e sexa $\mu = \mathrm{E}(X)$ o parámetro de interese. Sabemos que a media mostral, $\bar{X}$, é un estimador asintoticamente innesgado para estimar μ e que o seu erro estándar pode estimarse por $S/\sqrt{n}$. O intervalo de confianza aproximado para μ de nivel 0.95 é

$$\left(\bar{X} - 1.96\frac{S}{\sqrt{n}}, \bar{X} + 1.96\frac{S}{\sqrt{n}}\right).$$

Intervalo aproximado para unha proporción

Supoñamos que o parámetro de interese é unha proporción p. Sabemos que a proporción mostral, $\hat{p}$, é un estimador asintoticamente innesgado para estimar p e que o seu erro estándar pode estimarse por $\sqrt{\hat{p}(1-\hat{p})}/\sqrt{n}$. O intervalo de confianza aproximado para p de nivel 0.95 é

$$\left(\hat{p} - 1.96\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}, \hat{p} + 1.96\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\right).$$

Exemplo. (cont., datos DIABETES) A estimación da proporción de mulleres cun valor da concentración de glucosa de máis de 150 era $\hat{p} = 0.098$. Para obter o correspondente intervalo de confianza de nivel 0.95 facemos

```
> p.estimado <- mean(mostra>150)
> LI <- p.estimado-qnorm(0.975)*sqrt(p.estimado*(1-p.estimado))/sqrt(n)
> LS <- p.estimado+qnorm(0.975)*sqrt(p.estimado*(1-p.estimado))/sqrt(n)
> c(LI,LS)

[1] 0.04765342 0.14931628
```

O intervalo de confianza é $(0.048, 0.149)$, ou, en porcentaxes, $(4.8\,\%, 14.9\,\%)$. Este intervalo obviamente non é moi informativo por ter moita lonxitude. □

Exemplo. No barómetro de xaneiro de 2021, o Centro de Investigación Sociolóxicas (CIS; web: www.cis.es) inclúe a pregunta *"Está vostede disposto/a a vacinarse da COVID-19 inmediatamente?"*. Dunha mostra de 3862 persoas, o 72.5 % contestou "Si". Neste caso, o intervalo de confianza para a proporción de persoas dispostas a vacinarse é

$$\left(0.725 - 1.96\frac{\sqrt{0.725(1-0.725)}}{\sqrt{3862}}, 0.725 + 1.96\frac{\sqrt{0.725(1-0.725)}}{\sqrt{3862}}\right) = (0.711, 0.739),$$

é dicir $(71.1\,\%, 73.9\,\%)$. Este intervalo ten unha lonxitude pequena e polo tanto é moi informativo. □

En xeral, para ter intervalos de confianza para proporcións que teñan lonxitudes pequenas necesítanse tamaños mostrais moi grandes. Tendo en conta a acotación $\sqrt{\hat{p}(1-\hat{p})} \leq 0.5$, podemos seguir unha estratexia conservadora para construír un intervalo de confianza para p. O

intervalo

$$\left(\hat{p} - 1.96\frac{0.5}{\sqrt{n}}, \hat{p} + 1.96\frac{0.5}{\sqrt{n}}\right).$$

é un intervalo de confianza para p de nivel de confianza maior ou igual que 0.95. A lonxitude deste intervalo é

$$2 \cdot 1.96\frac{0.5}{\sqrt{n}} = \frac{1.96}{\sqrt{n}},$$

que só depende do tamaño mostral. Podemos entón empregar esta fórmula para deducir o tamaño mostral necesario para obter un intervalo de confianza para p cunha lonxitude máxima determinada. Por exemplo, se queremos que a lontitude sexa 0.06 (é dicir, o intervalo é da forma porcentaxe estimada $\mp 3\,\%$), o tamaño mostral terá que ser

$$\frac{1.96}{\sqrt{n}} < 0.06 \;\Leftrightarrow\; \sqrt{n} > \frac{1.96}{0.06} = 32.67 \;\Leftrightarrow\; n > 32.67^2 = 1067.111,$$

é dicir, n debe ser como mínimo 1068.

Exemplo. En moitas enquisas sociolóxicas e de opinión emprégase tamaños mostrais próximos a $n = 1000$ e nivel de confianza 0.955 (este nivel de confianza permite cambiar o cuantil 1.96 por 2 e simplificar os cálculos). Por exemplo, a ficha técnica dunha enquisa publicada pola Voz de Galicia en febreiro de 2021 sobre a vacinación contra a COVID especifica que $n = 1223$ e polo tanto o intervalo de nivel 0.955 será da forma $\mp 100 \cdot 1/\sqrt{1223}\,\% = \mp 2.86\,\%$. □

Exercicio 3.7. ***Intervalo de confianza para a media dunha Exponencial.*** *A xerencia dun hospital quere analizar o tempo que pasan os pacientes no servizo de urxencias. Con este propósito, recóllense os datos dos tempos de permanencia (en horas) de 50 pacientes:*

0.05	*0.07*	*0.08*	*0.15*	*0.18*	*0.21*	*0.43*	*0.48*	*0.52*	*0.62*
0.65	*0.67*	*0.70*	*0.81*	*0.86*	*0.89*	*0.91*	*0.96*	*1.02*	*1.05*
1.19	*1.23*	*1.44*	*1.58*	*1.65*	*1.79*	*1.82*	*1.84*	*1.93*	*1.97*
2.26	*2.27*	*2.34*	*2.59*	*2.76*	*3.09*	*3.39*	*3.54*	*3.62*	*3.70*
4.28	*4.33*	*4.60*	*4.69*	*4.91*	*5.19*	*5.35*	*6.02*	*6.38*	*6.76*

O histograma seméllase ao dunha variable aleatoria Exponencial.

Supoñamos polo tanto que $X \sim$ Exponencial(λ) e sexa $X_1, \ldots, X_n$ unha mostra aleatoria simple de X. Grazas ás propiedades da distribución Exponencial non é difícil de demostrar que

$$2\lambda \sum_{i=1}^{n} X_i \sim \chi^2_{2n}.$$

(a) Emprega o estatístico pivotal anterior para deducir un intervalo de confianza de nivel 0.95 para $E(X) = 1/\lambda$. Particularízao para os datos dos tempos de permanencia en urxencias.

(b) Empregando a normalidade asintótica da media mostral, obtén o intervalo de confianza aproximado para $E(X)$ de nivel 0.95*. Compárao co resultado de (a).*

3.5. Tests de hipóteses

3.5.1. Introdución

En moitas situacións prácticas podemos ter algún coñecemento ou algunha suposición sobre a característica poboacional sometida a estudo. Os **tests de hipóteses** son procedementos inferenciais que nos permiten decidir sobre a validez dunha determinada afirmación sobre esa característica.

Exemplos.

- É ben coñecido que a diabetes pode producir valores altos da concentración de glucosa no sangue. No conxunto de datos DIABETES, vimos que as medias mostrais da variable *glu* no grupo de mulleres non diabéticas e diabéticas eran 113.11 e 145.06, respectivamente. Aparentemente, o valor medio do grupo diabético é maior ca o valor medio do grupo non diabético. Pero, será esta unha diferenza real ou deberase unicamente á variabilidade no proceso de obtención dos datos?

- A Organización Mundial da Saúde define a obesidade como un índice de masa corporal maior que 30. Supera este valor a media da variable *bmi* no grupo de mulleres diabéticas?

- Para comprobar o posible efecto dun certo tratamento no peso duns ratos realizouse un experimento e obtívose a información que se recolle no conxunto de datos RATPUPS. Teñen os distintos niveis da variable *treatment* un efecto no peso dos ratos?

- No estudo correspondente ao conxunto de datos DIABETES, podemos asumir que a variable *glu* ten distribución Normal?

- ...

3.5.2. Elementos dun test de hipóteses

Unha **hipótese** é unha afirmación sobre a variable (ou variables) en estudo. Pode ser unha afirmación sobre algunha característica particular, é dicir, sobre algún parámetro (por exemplo, "a media da variable X é maior ca 30", "as medias das variables X e Y son iguais", "a probabilidade p é menor ca 0.20" etc.) ou sobre a variable completa (por exemplo, "a variable X é Normal").

Para realizar un test temos que establecer dúas hipóteses, que denominaremos **hipótese nula** e **hipótese alternativa**:

- A **hipótese nula** (que denotaremos por H_0) é a afirmación que inicialmente se asume como certa. Habitualmente a hipótese nula indica que non hai ningún efecto presente (por exemplo, "as medias das variables X e Y son iguais" ou "as variables X e Y teñen a mesma distribución").

- A **hipótese alternativa** (que denotaremos por H_1) é a hipótese que pretendemos contrastar fronte á hipótese nula. A hipótese alternativa indica que hai algún efecto presente (por exemplo, "a media da variable X é maior ca a media da variable Y" ou "as variables X e Y teñen distribucións distintas").

Un test de hipóteses é un procedemento estatístico que nos permite decidir se a hipótese nula debería ser rexeitada en base á información proporcionada por unha mostra. **Rexeitarase a hipótese nula H_0 se na mostra se observa suficiente evidencia en contra dela**; nese caso aceptarase a validez de H_1. En cambio, se na mostra non se observa suficiente evidencia en contra de H_0, entón a hipótese nula non poderá ser rexeitada.

Para realizar o test necesitamos un criterio estatístico que permita decidir ata que punto os datos corroboran ou non a hipótese nula. O **estatístico de test** é calquera criterio que permite medir a discrepancia existente entre a hipótese nula H_0 e os datos. O estatístico de test dependerá dos datos e polo tanto será unha variable aleatoria. Unha vez escollido o estatístico de test, divídense os seus posibles valores (é dicir, o seu soporte) en dúas rexións:

- A **rexión de aceptación** son aqueles valores do estatístico de test que non indican unha gran discrepancia entre H_0 e os datos. Se o valor do estatístico de test está na rexión de aceptación entón a hipótese nula H_0 non pode ser rexeitada.

- A **rexión de rexeitamento** ou **rexión crítica** son os valores do estatístico de test que indican unha gran discrepancia entre os datos e a hipótese nula H_0. Se o valor do estatístico de test está na rexión crítica, a hipótese nula H_0 rexéitase en favor da hipótese alternativa H_1.

O valor (ou valores) que separa a rexión de aceptación e a rexión crítica chámase **valor crítico**.

A decisión que se tome a favor dunha ou doutra hipótese está fundamentada na discrepancia observada entre a hipótese nula e a información subministrada por unha única mostra. Resulta obvio que tal decisión pode ser correcta ou errónea. As catro posibles situacións ás que pode dar lugar un test de hipóteses esquematízanse no seguinte cadro:

		Realidade	
		H_0 certa	H_0 falsa
Decisión	Non rexeitar H_0	decisión correcta	**erro de tipo II**
	Rexeitar H_0	**erro de tipo I**	decisión correcta

Damos a continuación unha serie de definicións importantes para establecer a metodoloxía dos tests de hipóteses:

Erro de tipo I e nivel de significación. Nun test de hipóteses, a decisión de rexeitar a hipótese nula H_0 cando é certa denomínase **erro de tipo I**. A probabilidade de cometer dito erro chámase **nivel de significación do test** e denótase por α:

$$\alpha = P(\text{erro tipo I}) = P(\text{rexeitar } H_0 \mid H_0 \text{ certa}).$$

Erro de tipo II. Nun test de hipóteses, a decisión de non rexeitar a hipótese nula H_0 cando é falsa denomínase **erro de tipo II**. A súa probabilidade denótase por β:

$$\beta = P(\text{erro tipo II}) = P(\text{non rexeitar } H_0 \mid H_0 \text{ falsa}).$$

Potencia. A **potencia** é a probabilidade de rexeitar a hipótese nula H_0. Denótase por π:

$$\pi = P(\text{rexeitar } H_0).$$

O ideal sería que as probabilidades dos dous tipos de erro, α e β, fosen pequenas. Desafortunadamente isto non é posible, xa que se intentamos diminuír a probabilidade dun tipo de erro entón aumenta a probabilidade de cometer o erro do outro tipo. Na maior parte dos casos o erro de tipo I é o máis importante, así que na práctica unicamente se controla este. Decídese de antemán a probabilidade máxima de erro tipo I, é dicir, o nivel de significación, e só se rexeita H_0 se a evidencia na súa contra é moi forte. É habitual traballar con $\alpha = 0.05$.

Para realizar o test na práctica, en primeiro lugar temos que escoller un estatístico de test, que, como xa dixemos, é o criterio que permite medir a discrepancia entre a hipótese nula e os datos. O **estatístico de test**, D, é unha función que involucra aos datos e á información proporcionada en H_0, de tal xeito que a súa distribución é coñecida ou pode ser aproximada cando se supón que H_0 é certa. Ao fixar o nivel de significación, α, obtense directamente unha división dos posibles valores do estatístico de test D en dúas rexións:

- A **rexión de aceptación**, de probabilidade $1 - \alpha$ baixo H_0. Se a hipótese nula é certa, entón o estatístico de test D ten unha probabilidade $1 - \alpha$ (digamos por exemplo 0.95) de pertencer á rexión de aceptación. Se ocorre isto, parece polo tanto razoable asumir que non hai razóns para pensar que H_0 é falsa ou, dito doutra forma, se o valor do estatístico de test D calculado na mostra pertence á rexión de aceptación, entón **non** existen razóns suficientes para rexeitar a hipótese nula cun nivel de significación α. Nese caso o test dise que é estatisticamente **non significativo**.
- A **rexión de rexeitamento** ou **rexión crítica**, de probabilidade α baixo H_0. En caso de que H_0 fose certa, o estatístico de test calculado na mostra pertencería a rexión crítica só con probabilidade α (digamos 0.05). Como α é unha probabilidade pequena, parace máis razoable pensar que H_0 non é certa e polo tanto a verdadeira distribución de D non é a que se estableceu baixo a hipótese nula. En resumo, se o valor do estatístico de test D calculado na mostra pertence á rexión crítica quere dicir que os datos contradin á hipótese nula H_0, e polo tanto rexéitase a hipótese nula H_0 en favor da hipótese alternativa H_1. Neste caso dise que o test é estatisticamente **significativo**.

Con esta metodoloxía, a potencia do test comportarase do seguinte xeito:

- Se H_0 é certa, entón a potencia coincide co nivel de significación ($\pi = \alpha$).
- Se H_0 é falsa, entón a potencia é a probabilidade complementaria da probabilidade de erro de tipo II ($\pi = 1 - \beta$).

Un procedemento de test é **consistente** se

(a) no caso de que a hipótese nula sexa certa a potencia efectivamente coincide co nivel de significación establecido de antemán, independentemente do tamaño mostral; e

(b) no caso de que a hipótese nula sexa falsa, a potencia converxe a 1 segundo aumenta o tamaño mostral.

Ás veces non é posible que a condición (a) se cumpra estritamente porque non se dispón dun coñecemento exacto da distribución do estatístico de test baixo a hipótese nula. Isto ocorre por exemplo cando se emprega unha distribución aproximada. Neses casos, o que se lle esixe ao test para ser consistente é que, baixo a hipótese nula, a potencia se aproxime ao nivel de significación ao aumentar o tamaño mostral.

3.5.3. Resumo da metodoloxía dos tests de hipóteses

A resolución dun test de hipóteses consiste nos seguintes pasos:

1. Especificar a hipótese nula, H_0, e a hipótese alternativa, H_1.
2. Fixar o nivel de significación (habitualmente, $\alpha = 0.05$).
3. Escoller o estatístico de test para medir a discrepancia entre a hipótese nula e os datos.
4. Clasificar en dúas rexións os posibles valores da distribución do estatístico de test baixo H_0 (ou, equivalentemente, buscar os valores críticos):
   - A rexión de aceptación, con probabilidade $1 - \alpha$.
   - A rexión crítica, con probabilidade α.
5. Calcular o valor numérico do estatístico de test cos datos e tomar unha decisión:
   - Se o valor do estatístico de test obtido da mostra pertence á rexión de aceptación, entón **non** hai suficiente evidencia en contra da hipótese nula. Neste caso non se rexeita a hipótese nula H_0 e dise que o test é estatisticamente **non significativo**.
   - Se o valor do estatístico de test obtido da mostra pertence á rexión crítica, entón a evidencia en contra da hipótese nula é forte. Neste caso rexéitase a hipótese nula H_0 en favor da hipótese alternativa H_1 e dise que o test é estatisticamente **significativo**.
6. Calcular e interpretar o p-valor (ver seguinte sección).

3.5.4. O p-valor

O **p-valor** é a probabilidade de obter un valor do estatístico de test que sexa tan ou máis contraditorio con H_0 como o que se observou a partir da mostra, supoñendo que H_0 é certa. Intuitivamente, o p-valor é unha medida da evidencia en contra de H_0: canto menor sexa o p-valor, maior é a evidencia en contra de H_0. Así:

- Se p-valor $< \alpha$ entón rexéitase a hipótese nula H_0 (**test significativo**).
- Se p-valor $\geq \alpha$ entón non se rexeita a hipótese nula H_0 (**test non significativo**).

Normalmente trabállase coa seguinte escala de p-valores:

p-valor	evidencia en contra de H_0
< 0.01	evidencia moi forte en contra de H_0
$0.01 - 0.05$	evidencia forte en contra de H_0
$0.05 - 0.10$	evidencia débil en contra de H_0
> 0.1	ningunha evidencia en contra de H_0

Algúns **comentarios/advertencias** sobre o p-valor[3]:

- Cando H_0 é certa, o p-valor compórtase como unha observación dunha variable aleatoria con distribución $Uniforme[0, 1]$. Polo tanto, un p-valor "grande" (digamos por exemplo 0.80) non significa que haxa moita evidencia *a favor* de H_0. p-valores como 0.14, 0.35 ou 0.87 son indistinguibles desde o punto de vista práctico.
- Por outra banda, se H_0 non é certa, a distribución do p-valor tenderá a concentrarse cerca do 0. Por iso a observación dun valor pequeno (digamos por exemplo 0.02 ou 0.007) é evidencia de que a hipótese nula pode ser falsa. Se o procedemento de test é consistente, esta concentración cerca do 0 cando H_0 é falsa tenderá ser maior canto maior sexa o tamaño mostral.
- Un p-valor grande pode ocorrer por dúas razóns:
  - H_0 é certa; ou
  - H_0 é falsa, pero o procedemento de test ten pouca potencia.
- **Importante!**: O p-valor non é a probabilidade de que H_0 sexa certa.

[3] Máis información en

Wasserstein, R.L. & Lazar, N. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, 70, 129–133.

3.5.5. Tests sobre a media nunha poboación Normal

Sexa $X_1, X_2, \ldots, X_n$ unha mostra aleatoria simple dunha variable poboacional X, da cal sabemos que ten distribución $N(\mu, \sigma)$. Pensemos como facer un test sobre a hipótese nula

$$H_0 : \mu = \mu_0,$$

onde μ_0 é un valor pre-especificado. O estatístico de test debe servir para medir a discrepancia entre os datos e H_0. Tendo en conta que a media mostral $\bar{X}$ é un estimador innesgado da verdadeira media poboacional μ, parece bastante razoable comparar $\bar{X}$ con μ_0, por exemplo a través da diferenza $\bar{X} - \mu_0$. Agora teriamos que decidir cales son os valores que nos permiten rexeitar H_0, é dicir, temos que decidir que valores deste estatístico de test son "grandes" e cales son "pequenos". Para facer isto enfrentámonos a dous problemas.

En primeiro lugar, non sabemos cal é a distribución $\bar{X} - \mu_0$. Sabemos que $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$ e supoñendo que H_0 é certa, entón podemos engadir a información $\mu = \mu_0$, pero aínda nos falta σ, da cal non sabemos nada. Aquí podemos botar man a teoría de estatísticos pivotais que xa coñecemos dos intervalos de confianza e empregar que como $X \sim N(\mu, \sigma)$ entón o estatístico pivotal $\sqrt{n}(\bar{X} - \mu)/S$ ten distribución t_{n-1}. Supoñendo que H_0 é certa entón $\mu = \mu_0$ e polo tanto podemos afirmar que

$$D = \sqrt{n}\frac{\bar{X} - \mu_0}{S} \sim t_{n-1} \quad \text{baixo } H_0. \tag{3.1}$$

Agora temos un estatístico que conserva a información sobre a diferenza entre $\bar{X}$ e μ_0, pero coa vantaxe de que ten unha distribución coñecida (sobre a cal poderemos calcular probabilidades, buscar cuantís etc.). Ademais o estatístico D pode ser calculado coa mostra porque n é coñecido, $\bar{X}$ e S calcúlanse a partir dos datos e μ_0 está especificado na hipótese nula. D será polo tanto o noso estatístico de test.

O segundo problema é como decidir cales son os valores grandes ou pequenos do estatístico de test. Parece bastante claro que valores próximos a 0 non suporán evidencia en contra de H_0. Pero cales serán os valores que proporcionarán evidencia en contra de H_0: valores positivos grandes?, valores negativos grandes en valor absoluto?, ambos? Isto depende da hipótese alternativa. Por exemplo, se a hipótese alternativa é $H_1 : \mu > \mu_0$, parece razoable que unicamente haberá evidencia en contra de H_0 cando D tome un valor positivo grande. Esencialmente temos tres posibilidades, tal e como se ilustra na Figura 3.9:

(a) Test bilateral $H_1 : \mu \neq \mu_0$. A rexión crítica está formada polas dúas colas da distribución do estatístico de test baixo H_0, ambas as dúas con probabilidade $\alpha/2$.

(b) Test unilateral $H_1 : \mu > \mu_0$. A rexión crítica está formada pola cola dereita de probabilidade α da distribución do estatístico de test baixo H_0.

(c) Test unilateral $H_1 : \mu < \mu_0$. A rexión crítica está formada pola cola esquerda de probabilidade α da distribución do estatístico de test baixo H_0.

No caso dos tests unilaterais (b) e (c), a construción das rexións críticas serve para os tests análogos con hipótese nula $H_0 : \mu \leq \mu_0$ ou $H_0 : \mu \geq \mu_0$, segundo corresponda.

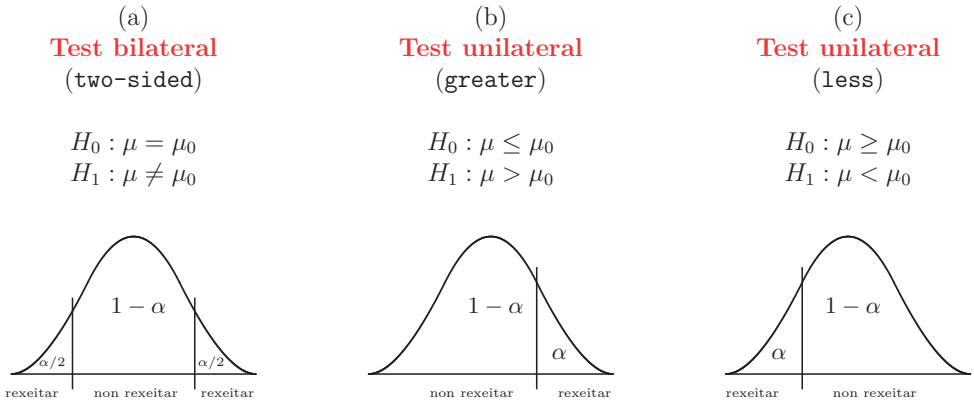


Figura 3.9: Distintos tipos de test en función da hipótese alternativa.

En R disporemos de moitas funcións que permiten facer tests. En xeral, o argumento `alternative` permite establecer se o test é bilateral (valor "`two-sided`") ou unilateral (valores "`greater`" ou "`less`", segundo corresponda).

A Táboa 3.3 recolle as rexións de aceptación e críticas dos tests para a media dunha poboación Normal. Na táboa, $t_{f,p}$ denota o cuantil de orde p dunha distribución t_f (por exemplo, $t_{10,0.40} = -0.26$ e $t_{20,0.95} = 1.725$).

Exemplo. Datos DIABETES. Sábese que a diabetes tipo 2 está dalgunha maneira relacionada co sobrepeso e a obesidade. A OMS define a obesidade como un índice de masa corporal (IMC) maior ou igual que 30 (máis información en https://www.who.int/health-topics/obesity).

Sexa X a variable "índice de masa corporal dunha muller diabética" e sexa $\mu = \mathrm{E}(X)$. Consideremos o test

$$H_0 : \mu = 30 \quad \text{fronte a} \quad H_1 : \mu > 30.$$

Neste caso escollemos un test unilateral porque só estamos interesados en detectar desviacións cara á dereita de 30. Fixamos o nivel de significación en $\alpha = 0.05$.

Para levar á práctica o test imos empregar as $n = 68$ observacións da variable *bmi* dispoñibles no conxunto de datos DIABETES. O histograma destas observacións ten forma de campá, polo que parece bastante razoable supoñer que a variable X é Normal, digamos $N(\mu, \sigma)$. A media mostral é 34.71, un valor aparentemente bastante maior ca 30. Pero este valor tan alto deberase simplemente á variabilidade intrínseca da media mostral ou será que a media poboacional é realmente maior ca 30?

O estatístico de test calculado na mostra é

$$\sqrt{n}\frac{\bar{X} - 30}{S} = \sqrt{68}\frac{34.71 - 30}{4.81} = 8.07.$$

Agora temos que decidir se este valor é suficientemente grande para rexeitar a hipótese nula. Para iso temos que atopar o valor crítico do test. Se o nivel de significación é $\alpha = 0.05$, entón o valor crítico para este test unilateral é o cuantil 0.95 da t_{67}, é dicir, $t_{67,0.95} = 1.668$. Claramente, o valor observado do estatístico é moito maior ca o valor crítico, así que claramente pertence

Táboa 3.3: Tests para a media dunha poboación Normal. O estatístico de test está dado na ecuación (3.1). O nivel de significación é α e d denota o valor do estatístico de test calculado na mostra.

Test	Rexión de aceptación	Rexión crítica	p-valor
$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	$(-t_{n-1,1-\alpha/2},$ $t_{n-1,1-\alpha/2})$	$(-\infty, -t_{n-1,1-\alpha/2})$ $\cup(t_{n-1,1-\alpha/2}, +\infty)$	 $2P(t_{n-1} \geq \lvert d \rvert)$
$H_0 : \mu \leq \mu_0$ $H_1 : \mu > \mu_0$	$(-\infty, t_{n-1,1-\alpha})$	$(t_{n-1,1-\alpha}, +\infty)$	 $P(t_{n-1} \geq d)$
$H_0 : \mu \geq \mu_0$ $H_1 : \mu < \mu_0$	$(-t_{n-1,1-\alpha}, +\infty)$	$(-\infty, -t_{n-1,1-\alpha})$	 $P(t_{n-1} \leq d)$

á rexión crítica. Podemos polo tanto rexeitar a hipótese nula en favor da alternativa, é dicir, á vista dos datos, é claro que a media do IMC da poboación de mulleres diabéticas supera o valor 30. Neste caso o p-valor é $P(t_{67} > 8.07)$, que resulta practicamente 0.

En R podemos empregar a función `t.test()`:

```
> t.test(bmi[type=="Yes"],mu=30,alternative="greater")

        One Sample t-test

data:  bmi[type == "Yes"]
t = 8.0712, df = 67, p-value = 8.955e-12
alternative hypothesis: true mean is greater than 30
95 percent confidence interval:
 33.73574      Inf
sample estimates:
mean of x
 34.70882
```

Nótese que o p-valor é extremadamente pequeno, o cal tamén indica que hai unha evidencia moi forte en contra da hipótese nula. Habitualmente, cando o p-valor é moi pequeno non se acostuma dar o seu valor exacto e simplemente se di que é menor ca 0.001. De feito, moitos softwares estatísticos devolven simplemente "p-valor$<$ 0.001". □

Se na función `t.test()` escollemos un test bilateral poñendo `alternative="two.sided"`, a saída tamén proporciona o intervalo de confianza para a media que estudamos na Sección 3.4.2.

Exercicio 3.8. *Deséxase saber se unha balanza está ben calibrada, é dicir, se o seu erro de medida ten media* 0*. Denotemos por* X *o erro nunha pesada. Queremos polo tanto realizar un test con hipótese nula* $H_0 : \mu = 0$*, onde* $\mu = E(X)$*. Para isto realizáronse* 50 *pesadas dunha pesa de* 1 *kg e obtivéronse os seguintes datos (rexistrados en g):*

1005	*1017*	*1011*	*990*	*1003*	*1007*	*986*	*1013*	*994*	*1002*
1015	*1019*	*1000*	*999*	*998*	*1009*	*989*	*1005*	*1002*	*1004*
996	*1009*	*994*	*1011*	*998*	*1002*	*1008*	*1006*	*991*	*1009*
1005	*1000*	*998*	*1008*	*997*	*994*	*983*	*999*	*1000*	*1001*
1000	*1005*	*1000*	*1020*	*1001*	*1003*	*993*	*1005*	*1023*	*1012*

(a) *Parece razoable supoñer que os datos proceden dunha distribución Normal? Por que?*

(b) *Realiza o test de hipóteses adecuado. Escollerías unha hipótese alternativa bilateral ou unilateral? Analiza e interpreta os resultados.*

Exercicio 3.9. *O conxunto de datos* DIGITRATIO *contén información sobre ratios dixitais, en particular sobre a chamada ratio 2D:4D, que é o cociente entre a lonxitude dos dedos índice (2º dedo) e anular (4º dedo). Esta relación parece que está relacionada coa exposición prenatal aos andróxenos. As variables que contén o conxunto de datos son:*

- age*: idade (en anos).*
- gender*: male (home)/female (muller).*
- D2D4right*: ratio 2D:4D na man dereita.*
- D2D4left*: ratio 2D:4D na man esquerda.*

Realiza un test de hipóteses para comprobar se o valor medio da ratio 2D:4D en cada man é distinta de 1*. Comenta os resultados.*

3.5.6. Que é a potencia dun test?

Supoñamos que a variable de interese é $X \sim N(\mu, 2)$ e queremos facer o seguinte test

$$H_0 : \mu \leq 5 \quad \text{fronte a } H_1 : \mu > 5$$

cun nivel de significación 0.05.

Nótese que agora estamos asumindo non só que a variable é Normal, senón tamén que coñecemos a súa desviación estándar (neste caso, 2), cousa que non é xustificable desde o punto de vista práctico. Aínda así, neste exemplo suporémola coñecida para ilustrar o funcionamento do test en termos de potencia. Con esta suposición, en vez de traballar co estatístico de test do exemplo anterior (o que tiña distribución t_{n-1}) tomaremos o estatístico

$$\sqrt{n}\frac{\bar{X} - 5}{2},$$

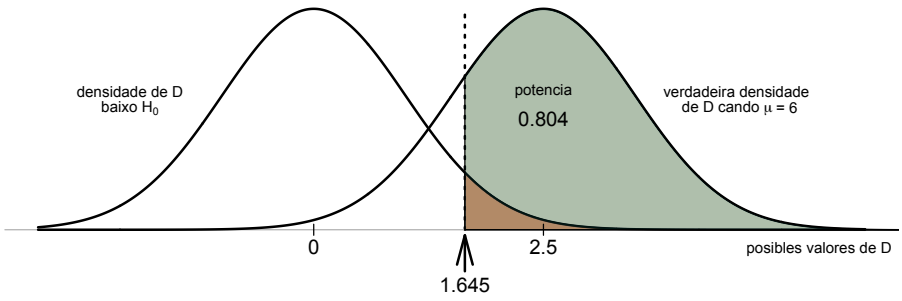


Figura 3.10: Potencia dun test. A curva da esquerda é a densidade da $N(0,1)$ e a da dereita é a densidade da $N(2.5,1)$.

no que esencialmente reemprazamos o estimador S do denominador polo valor poboacional $\sigma = 2$. Isto permítenos traballar máis comodamente, xa que debido ás propiedades da Normal[4]

$$\sqrt{n}\frac{\bar{X}-5}{2} \sim N(0,1) \quad \text{baixo } H_0.$$

Traballaremos con tamaño mostral $n = 25$. Para o nivel de significación $\alpha = 0.05$, a rexión crítica é $(1.645, \infty)$, tal e como se ilustra na Figura 3.10.

Supoñamos agora que a hipótese nula H_0 non é certa porque o verdadeiro valor da media é $\mu = 6$. Entón, a verdadeira distribución do estatístico de test xa non é unha $N(0,1)$, senón que pasa a ser $N(2.5,1)$, xa que

$$\sqrt{25}\frac{\bar{X}-5}{2} = \sqrt{25}\frac{\bar{X}-6+6-5}{2} = \sqrt{25}\frac{\bar{X}-6}{2} + \sqrt{25}\frac{6-5}{2} = \underbrace{\sqrt{25}\frac{\bar{X}-6}{2}}_{N(0,1)} + 2.5 \sim N(2.5,1).$$

Este cambio na distribución do estatístico D é o que lle dá potencia ao test. Recordemos que a potencia é a probabilidade de rexeitar H_0. Cando $\mu = 6$ e $n = 25$ a potencia é

$$P(\text{rexeitar } H_0 \text{ cando } \mu = 6) = P(D > 1.645) = P(N(2.5,1) > 1.645) = 0.804,$$

tal e como se pode ver na Figura 3.10.

A probabilidade que acabamos de calcular depende do verdadeiro valor do parámetro μ e do tamaño mostral n. Para diferentes valores de μ e n obteremos distintos valores da potencia:

- Cando a hipótese nula é certa (é dicir, $\mu = \mu_0$), entón a potencia é sempre α, independentemente do tamaño mostral.
- Cando a hipótese nula é falsa (no noso exemplo $\mu > \mu_0$), entón a potencia aumenta ao aumentar o tamaño mostral. A potencia tamén será maior canto maior sexa a diferenza entre μ (verdadeiro valor do parámetro) e 5 (valor establecido en H_0).

[4]Recordemos que se $X_1, X_2, \ldots, X_n$ é unha mostra aleatoria simple de $X \sim N(\mu, \sigma)$, entón $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$.

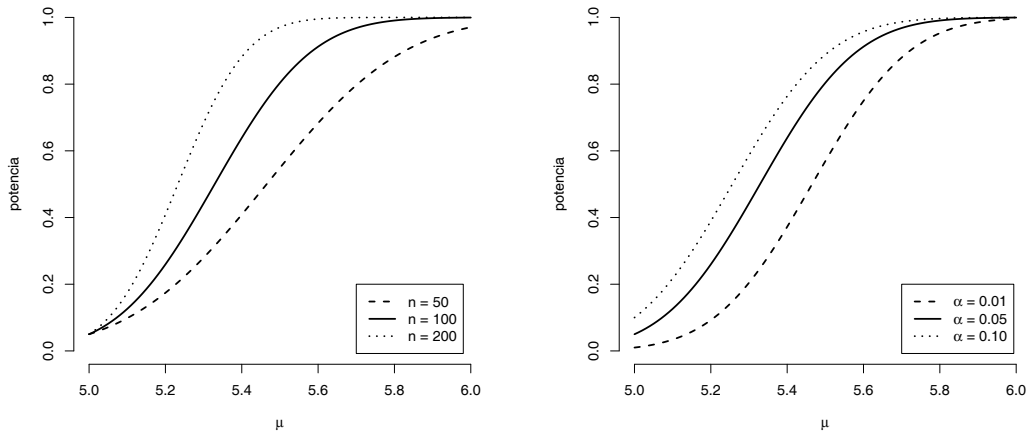


Figura 3.11: Esquerda: efecto de n na potencia (para $\alpha = 0.05$). Dereita: efecto de α na potencia (para $n = 100$).

- Por outra parte, a potencia tamén depende do nivel de significación: canto menor sexa α menor será a potencia.

- Finalmente, a potencia tamén depende da desviación estándar poboacional (no noso exemplo, 2). Se a desviación estándar fose menor, entón a potencia sería maior no caso de que a hipótese nula fose falsa.

Os gráficos da Figura 3.11 amosan a potencia como función de μ para distintos valores de n e α.

No exemplo anterior asumimos que a desviación estándar poboacional era coñecida. Isto permitiunos facer os cálculos de forma sinxela coas densidades Normais. O comportamento é similar cando o estatístico de test é $\sqrt{n}(\bar{X} - 5)/S$ en lugar de $\sqrt{n}(\bar{X} - 5)/2$, pero os cálculos son moito máis complexos.

Exercicio 3.10. *No exemplo anterior, con $\alpha = 0.05$, cal debería ser o tamaño mostral para obter unha potencia de polo menos 0.98 cando o verdadeiro valor de μ é (a) 6, (b) 5.5, (c) 5.25, (d) 5.1. Deduce unha fórmula xeral para este problema.*

3.6. Tests para comparar dúas medias

A comparación das medias de dúas variables é un dos problemas máis comúns na estatística. Existen varios tests para este problema.

Exemplos.

- Cabe esperar que a media da concentración de glucosa en sangue de persoas diabéticas sexa maior ca a das persoas non diabéticas.

  No conxunto de datos DIABETES as medias mostrais da variable *glu* no grupo de mulleres non diabéticas e diabéticas son 113.11 e 145.06, respectivamente. Estas dúas cantidades son moi diferentes, pero significa isto que existe unha verdadeira diferenza entre as medias poboacionais ou pode deberse simplemente á aleatoriedade das mostras?

- Terán os niveis do tratamento aplicado no conxunto de datos RATPUPS un efecto no peso? As medias mostrais da variable *weight* para os niveis de tratamento baixo e alto son 5.93 e 5.89, respectivamente. A diferenza entre estes dous valores parece pequena. Resultará razoable supoñer entón que as medias poboacionais correspondentes son iguais?

- Existe unha diferenza entre as medias das ratios dixitais 2D:4D entre homes e mulleres?

Sexan X e Y as variables que queremos comparar e sexan μ_X e μ_Y as súas medias. O test de hipóteses pode establecerse nos seguintes termos. A hipótese nula afirma a igualdade das medias

$$H_0 : \mu_X = \mu_Y \qquad \text{(as medias poboacionais de } X \text{ e } Y \text{ son iguais).}$$

En canto á hipótese alternativa, temos dúas opcións. A primeira é unha hipótese alternativa **bilateral**:

$$H_{1,\text{bilateral}} : \mu_X \neq \mu_Y \qquad \text{(as medias poboacionais de } X \text{ son } Y \text{ diferentes).}$$

A segunda posibilidade é a hipótese alternativa **unilateral**:

$$H_{1,\text{unilateral}} : \mu_X > \mu_Y \text{ (a media poboacional de } X \text{ é maior ca a media poboacional de } Y\text{).}$$

3.6.1. t-test para mostras independentes

Supoñamos que temos dúas mostras independentes, $X_1, X_2, \ldots, X_n$ de X e $Y_1, Y_2, \ldots, Y_m$ de Y. Nótese que neste caso os tamaños mostrais n e m poden ser diferentes. O estatístico de test é

$$D = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}, \tag{3.2}$$

onde $\bar{X}$ e $\bar{Y}$ son as medias mostrais e S_X^2 e S_Y^2 son as varianzas mostrais. Baixo a hipótese nula H_0, a distribución de D é:

- Se as poboacións X e Y son Normais, entón D ten unha distribución t_f, onde os graos de liberdade f dependen dos tamaños mostrais (non daremos os detalles aquí).

Táboa 3.4: t-test para comparar dúas medias. O estatístico de test está dado na ecuación (3.2). As rexións de aceptación e crítica están calculadas sobre a distribución $N(0,1)$. O nivel de significación é α e d denota o valor do estatístico de test calculado coas mostras.

Test	Rexión de aceptación	Rexión crítica	p-valor
$H_0 : \mu_X = \mu_Y$ $H_1 : \mu_X \neq \mu_Y$	$(-z_{1-\alpha/2}, z_{1-\alpha/2})$	$(-\infty, -z_{1-\alpha/2})$ $\cup(z_{1-\alpha/2}, +\infty)$	 $2P(N(0,1) \geq \lvert d \rvert)$
$H_0 : \mu_X = \mu_Y$ $H_1 : \mu_X > \mu_Y$	$(-\infty, z_{1-\alpha})$	$(z_{1-\alpha}, +\infty)$	 $P(N(0,1) \geq d)$

- Cando os tamaños mostrais son moderados ou grandes (digamos, $n > 50$ e $m > 50$), entón D é aproximadamente $N(0,1)$.

A Táboa 3.4 contén as rexións de aceptación e críticas dos tests para comparar dúas medias.

Esta metodoloxía para a comparación de dúas medias baseada na distribución t foi desenvolta por William Gosset a principios do século XX. Os seus métodos foron mellorados por Bernard L. Welch na década de 1940. Os traballos de Gosset, que asinaba os seus artigos científicos co pseudónimo Student, son claves no desenvolvemento da estatística moderna.

Exemplo. Consideremos o conxunto de datos RATPUPS. Realizaremos un test de hipóteses para comprobar se as medias do peso (variable *weight*) son distintas para os niveis de tratamento baixo ("low") e alto ("high") da variable *treatment*. Máis concretamente, comprobaremos se a media do peso é menor no grupo cun nivel alto de tratamento. Sexa X o "peso dun rato que recibiu o nivel baixo do tratamento", con media μ_X, e sexa Y o "peso dun rato que recibiu o nivel alto do tratamento", con media μ_Y. Dispoñemos de mostras de X e Y con tamaños 126 e 65, medias mostrais 5.93 g e 5.89 g, e varianzas mostrais 0.18 e 0.41, respectivamente. Neste caso, parece razoable facer un test unilateral

$$H_0 : \mu_X = \mu_Y \quad \text{fronte a} \quad H_1 : \mu_X > \mu_Y.$$

En R escribimos

```
> ratpups <- read.table(file="datos-ratpups.txt",header=TRUE)
> attach(ratpups)

> t.test(weight[treatment=="low"],weight[treatment=="high"],
+        alternative="greater")

        Welch Two Sample t-test
```

```
data:  weight[treatment == "low"] and weight[treatment == "high"]
t = 0.48394, df = 93.86, p-value = 0.3148
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.1041092        Inf
sample estimates:
mean of x mean of y
 5.928333  5.885538
```

O argumento `alternative="greater"` é a forma de indicarlle a R que queremos facer un test unilateral. O valor observado do estatístico de test é 0.48, que para un nivel de significación $\alpha = 0.05$ pertence á rexión de aceptación na $N(0,1)$ ou en calquera distribución t. O p-valor é 0.3148. Este é un valor "grande", o cal significa que non hai evidencia en contra da hipótese nula. Non podemos rexeitar que o peso medio para os dous niveis de tratamento sexa igual. O test é non significativo. □

Exercicio 3.11. *Como acabamos de ver no exemplo anterior relativo ao conxunto de datos* RATPUPS*, non parece haber diferenza entre os pesos medios para os niveis de tratamento baixo e alto. Considera agora os niveis baixo ("low") e alto ("high") como un único grupo e realiza un test de hipóteses para comprobar se hai diferenzas significativas entre a media dese grupo e a do grupo que non recibiu ningún tratamento ("control"). Comenta os resultados.*

Exercicio 3.12. *Considera o conxunto de datos* DIABETES*.*

(a) Fai un test de hipóteses para comprobar se as medias da concentración de glucosa (variable glu*) no grupo de mulleres diabéticas e non diabéticas son distintas. Comenta os resultados obtidos.*

(b) Fai un test de hipóteses para comprobar se as medias da presión arterial (variable bp*) no grupo de mulleres diabéticas e non diabéticas son distintas. Comenta os resultados obtidos.*

(c) Fai un test de hipóteses para comprobar se as medias do pregamento cutáneo do triceps (variable skin*) no grupo de mulleres diabéticas e non diabéticas son distintas. Comenta os resultados obtidos.*

3.6.2. t-test para mostras dependentes ou apareadas

No apartado anterior supuxemos que as mostras son independentes, é dicir, que proceden de poboacións independentes. Isto aplícase por exemplo cando os grupos que queremos comparar se forman por algunha variable categórica con dous valores (por exemplo home/muller,

enfermo/non enfermo, tratamento/control). En cambio, nalgúns casos prácticos, a comparación establécese entre dúas variables medidas sobre os mesmos individuos. Nese caso falamos de mostras **dependentes** ou **apareadas**.

Exemplo. Estamos interesados en comparar a media da ratio dixital 2D:4D medida na man dereita e na man esquerda dunha mesma persoa. Neste caso trátase dunha situación con variables dependentes, xa que para un mesmo individuo medimos dúas variables. Isto é o que ocorre coas variables *D2D4left* e *D2D4right* do conxunto de datos DIGITRATIO. □

Neste caso, a mostra consiste nun conxunto de pares de observacións da forma (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n). Nótese que agora hai un único tamaño mostral, n. Sexan μ_X e μ_Y as medias poboacionais do par de variables (X, Y). Para facer o test de comparación de medias de X e Y empregaremos unha nova variable artificial $T = X - Y$. A hipótese nula $H_0 : \mu_X = \mu_Y$ é equivalente a $H_0 : \mu_T = 0$, onde μ_T é a media poboacional da variable $T = X - Y$, é dicir, $\mu_T = \mathrm{E}(T) = \mathrm{E}(X - Y) = \mathrm{E}(X) - \mathrm{E}(Y) = \mu_X - \mu_Y$.

O test está baseado no estatístico

$$D = \sqrt{n}\frac{\bar{T}}{S_T},$$

onde $\bar{T}$ e S_T son a media mostral e a desviación estándar mostral da mostra formada polos valores $T_i = X_i - Y_i$, para $i = 1, \ldots, n$. A hipótese nula será rexeitada cando o valor observado do estatístico de test sexa moi distinto de 0 (lembremos que a decisión depende da hipótese alternativa: bilateral ou unilateral). Cando a hipótese nula é certa, entón a distribución do estatístico de test é:

- t_{n-1} cando as variables X e Y son Normais.
- Aproximadamente $N(0, 1)$ cando o tamaño mostral non é demasiado pequeno (habitualmente, $n \geq 40$ é suficiente).

As rexións de aceptación e crítica son as mesmas ca as dadas na Táboa 3.3.

Exemplo. No conxunto de datos DIGITRATIO compararemos as medias das ratios 2D:4D na man dereita e na man esquerda (variables *D2D4right* e *D2D4left*). Como non temos ningunha información previa sobre unha alternativa específica, faremos un test bilateral.

Primeiro faremos o test para as mostras completas e despois distinguiremos por sexo. En R empregamos a función `t.test()` e especificamos o argumento `paired=TRUE`:

```
> t.test(D2D4right,D2D4left,paired=TRUE)

        Paired t-test

data:  D2D4right and D2D4left
t = -2.0938, df = 40, p-value = 0.04266
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.0217139097 -0.0003836513
```

```
sample estimates:
mean of the differences
            -0.01104878
```

Cal é a conclusión do test? É fiable esta conclusión? (comproba se as variables son Normais ou se o tamaño mostral é suficientemente grande). (**exercicio**) □

> **Exercicio 3.13.** *Realiza un test para comparar as medias das ratios dixitais 2D:4D na mans dereita e esquerda, separando o grupo de homes e o de mulleres. Son fiables os resultados obtidos? Comproba se os tamaños mostrais son suficientemente grandes ou se é razoable asumir que as variables son Normais.*

3.6.3. Tests non paramétricos: o test de Wilcoxon-Mann-Whitney

Existen situacións nas que os tests descritos anteriormente non resultan axeitados porque as poboacións non son Normais e os tamaños mostrais son moi pequenos. Para solucionar este problema, adoitan empregarse tests non paramétricos en lugar dos t-tests para dúas mostras. O principal obxectivo destes procedementos é comparar as posicións de dúas distribucións nun marco xeral, sen supoñer ningún modelo de distribución particular nas poboacións. En xeral, só son consistentes baixo alternativas unilaterais.

Sexan X e Y as poboacións que queremos comparar. A hipótese nula agora é

$$H_0: \text{ os centros das distribucións de } X \text{ e } Y \text{ son iguais.}$$

Esta afirmación é bastante vaga e pode interpretarse en termos das medias ou das medianas.

Sexan $X_1, X_2, \ldots, X_n$ e $Y_1, Y_2, \ldots, Y_m$ dúas mostras independentes de X e Y, respectivamente. O **test de Wilxocon-Mann-Whitney** traballa cos pares (X_i, Y_j) construídos artificialmente combinando cada observación de X con cada observación de Y e conta en cantos destes pares a observación de Y é maior ca a observación de X mediante o estatístico

$$D_{WMW} = \sum_{i=1}^{n} \sum_{j=1}^{m} \mathrm{I}(Y_j > X_i).$$

Nótese que hai nm pares artificiais da forma (X_i, Y_j). Cando as distribucións de X e Y son similares, entón o estatístico D_{WMW} tomará un valor moderado. Por outra banda, se Y tende a ser maior ca X, entón D_{WMW} tomará un valor grande, mentres que se X tende a ser maior ca Y entón D_{WMW} tomará un valor pequeno. Cando as distribucións de X e Y son iguais, e polo tanto cando se cumpre a hipótese nula, a distribución de D_{WMW} pódese calcular exactamente (cousa que fixeron de forma independente F. Wilcoxon en 1945 e H.B. Mann e D.R. Whitney en 1947). En R a función `wilcox.test()` realiza este test.

Exemplo. Nun estudo sobre os efectos do tabaco na calidade do sono unha variable importante é o tempo que se tarda en conciliar o sono. Sospéitase que as persoas fumadoras tardan máis tempo en conciliar o sono. Para corroborar esta hipótese, realizouse un experimento no que, en

condicións controladas, se rexistraron os tempos de conciliación do sono de 27 individuos, dos cales 12 son fumadores e 15 son non fumadores. Os tempos observados (en minutos) son:

Non fumadores (X):	28.6	25.1	26.4	34.9	29.8	28.4	38.5	30.2
	30.6	31.8	41.6	21.1	36.0	37.9	13.9	
Fumadores (Y):	69.3	56.0	22.1	47.6	53.2	48.1	23.2	13.8
	52.7	34.4	60.2	43.8				

Os tamaños mostrais son pequenos e os histogramas amosan que asumir que as distribucións de X e Y son Normais non é razoable. Polo tanto, para comprobar se o hábito de fumar ten un efecto nocivo no tempo de conciliación do sono resultará máis adecuado aplicar o test de Wilcoxon-Mann-Whitney para saber se os tempos no grupo de fumadores efectivamente tenden a ser maiores ca os do grupo de non fumadores. En R facemos o seguinte:

```
> tempo.nonfumador <- c(28.6, 25.1, 26.4, 34.9, 29.8, 28.4, 38.5, 30.2,
+                 30.6, 31.8, 41.6, 21.1, 36.0, 37.9, 13.9)
> tempo.fumador <- c(69.3, 56.0, 22.1, 47.6, 53.2, 48.1, 23.2, 13.8,
+              52.7, 34.4, 60.2, 43.8)

> wilcox.test(tempo.fumador,tempo.nonfumador,alternative="greater")

        Wilcoxon rank sum exact test

data:  tempo.fumador and tempo.nonfumador
W = 134, p-value = 0.01606
alternative hypothesis: true location shift is greater than 0
```

O p-valor é pequeno (0.016), o cal indica que hai unha diferenza significativa entre os tempos dos fumadores e os non fumadores, tendendo a ser maiores os dos fumadores. □

Convén facer unha aclaración sobre o uso do argumento `alternative` na función `wilcox.test()`. No exemplo anterior puxemos `alternative="greater"` para comprobar que os tempos dos non fumadores son menores ca os dos fumadores. Isto é así porque en realidade este argumento refírese á posición das respectivas funcións de distribución: se os tempos dos non fumadores son menores, entón a correspondente función de distribución tenderá a estar por riba ("greater") da función de distribución dos tempos dos fumadores.

Exemplo. O test de Wilcoxon-Mann-Whitney tamén se pode adaptar para mostras apareadas. Por exemplo, no conxunto de datos DIGITRATIO, o histograma das variables *D2D4right* e *D2D4left* no grupo de homes non parecen axustarse a unha distribución Normal. Ademais o tamaño mostral é 19, que non é suficiente para empregar a aproximación á $N(0,1)$ do t-test. Polo tanto, parece máis recomendable empregar o test de Wilcoxon-Mann-Whitney, neste caso para mostras apareadas. En R

```
> wilcox.test(D2D4right[gender=="male"],D2D4left[gender=="male"],paired=TRUE)
```

```
        Wilcoxon signed rank test with continuity correction

data:  D2D4right[gender == "male"] and D2D4left[gender == "male"]
V = 54.5, p-value = 0.1074
alternative hypothesis: true location shift is not equal to 0
```

Neste caso o p-valor (0.1074) non amosa diferenza significativa. □

3.7. Tests de bondade de axuste

Moitos procedementos inferenciais baséanse na suposición de que a distribución poboacional segue un determinado modelo paramétrico, como por exemplo a Normal. O uso deses procedementos pode resultar inapropiado se a verdadeira distribución da poboación difire moito do modelo paramétrico suposto. O termo **bondade de axuste** refírese a procedementos que permiten comprobar a calidade dun modelo de distribucións á hora de describir a verdadeira distribución da variable de interese. O caso máis relevante na práctica é o da distribución Normal.

Exemplo. No conxunto de datos DIABETES, o histograma da variable *bmi* permite supoñer que as observacións proceden dunha variable Normal. Como podemos asegurarnos de que isto é realmente así? □

Para comprobar a bondade de axuste dunha familia de distribucións podemos empregar técnicas gráficas (como por exemplo o xa coñecido histograma ou o qq-plot, que estudaremos agora) ou tests de hipóteses.

3.7.1. qq-plots

Dispoñemos dunha mostra $X_1, X_2, \dots, X_n$ dunha variable aleatoria X e desexamos saber se resulta verosímil que X teña un tipo particular de distribución, como por exemplo a Normal. Un método efectivo, aínda que subxectivo, para comprobar esta hipótese é o **qq-plot** (tamén chamado gráfico cuantil-cuantil). A idea fundamental deste gráfico é que en caso de que a distribución postulada sexa correcta, entón os puntos do gráfico aparacerán cerca dunha liña recta. Por outra parte, se a verdadeira distribución de X é moi diferente da distribución empregada para construír o gráfico, entón os puntos desviaranse substancialmente do patrón lineal.

Centrémonos no caso da distribución Normal. O **qq-plot de Normalidade** consiste no gráfico dos pares de puntos

$$\left\{\left(F_0^{-1}(\hat{F}(X_i)), X_i\right),\ i = 1, \dots, n\right\},$$

onde F_0^{-1} é a función cuantil da $N(0,1)$ e $\hat{F}$ é un estimador non paramétrico da función de distribución da variable X (por exemplo, a función de distribución empírica). Se a verdadeira distribución dos datos é unha Normal, entón os puntos aparecerán sobre unha liña recta debido á relación de linealidade existente entre os cuantís dunha Normal calquera cos cuantís da $N(0,1)$.

En cambio, se os datos non proceden dunha distribución Normal, entón os puntos deben amosar algunha desviación clara con respecto á liña recta. En R a función `qqnorm()` realiza o qq-plot de Normalidade.

A Figura 3.12 amosa os histogramas, boxplots e qq-plots de Normalidade obtidos a partir de 500 observacións simuladas de distribucións $N(60, 10)$, $Exponencial(2)$ e $Uniforme[0, 1]$. Como podemos comprobar, cando os datos proceden da distribución Normal, o qq-plot presenta un patrón rectilíneo. En cambio, se os datos non proceden dunha distribución Normal, entón hai unha clara desviación con respecto á recta.

(a) $X \sim N(60, 10)$ (b) $X \sim Exponencial(2)$ (c) $X \sim Uniforme[0, 1]$

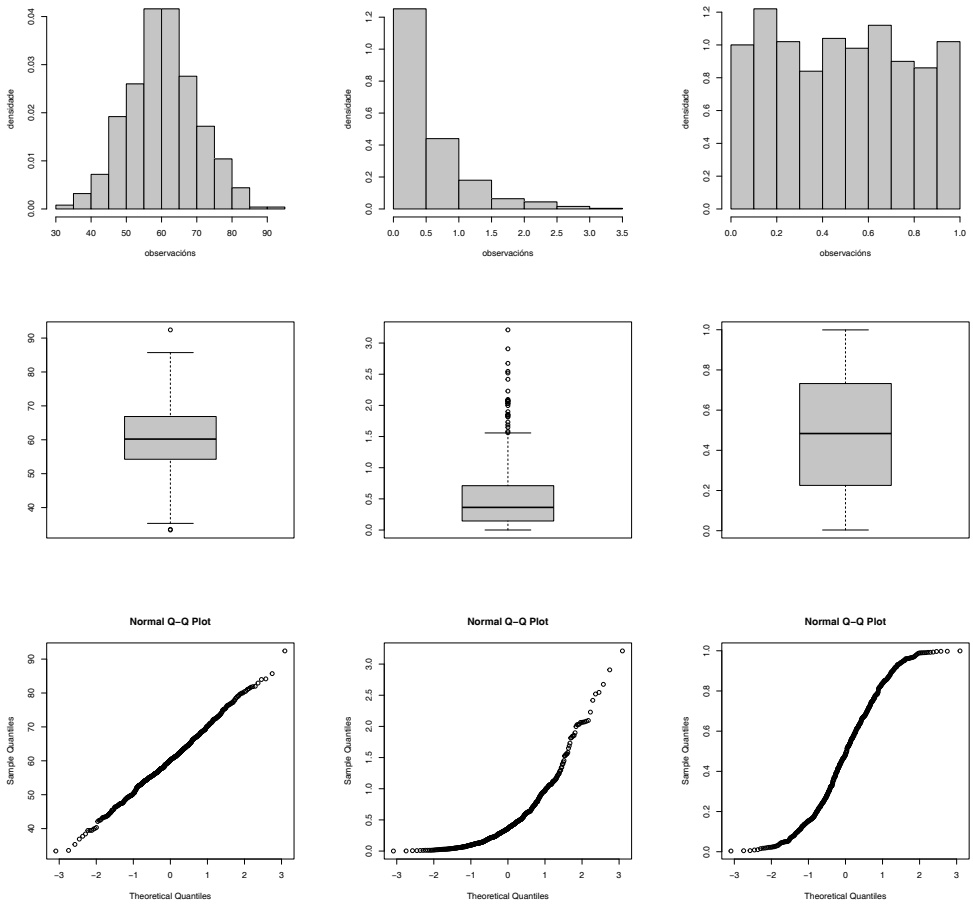


Figura 3.12: Histograma, boxplot e qq-plot de Normalidade obtidos a partir de 500 observacións simuladas con distribución (a) $N(60, 10)$; (b) $Exponencial(2)$; e (c) $Uniforme[0, 1]$.

Exemplo. A Figura 3.13 amosa os qq-plots das variables *bmi* e *glu* do conxunto de datos DIABETES. Para obter estes gráficos en R abonda facer

```
> qqnorm(bmi)
> qqnorm(glu)
```

A variable *glu* non parece seguir unha distribución Normal. □

Exercicio 3.14. *Realiza os qq-plots de Normalidade da variable* glu *separando os grupos de mulleres diabéticas e non diabéticas. Podemos supoñer que as variables correspondentes son Normais?*

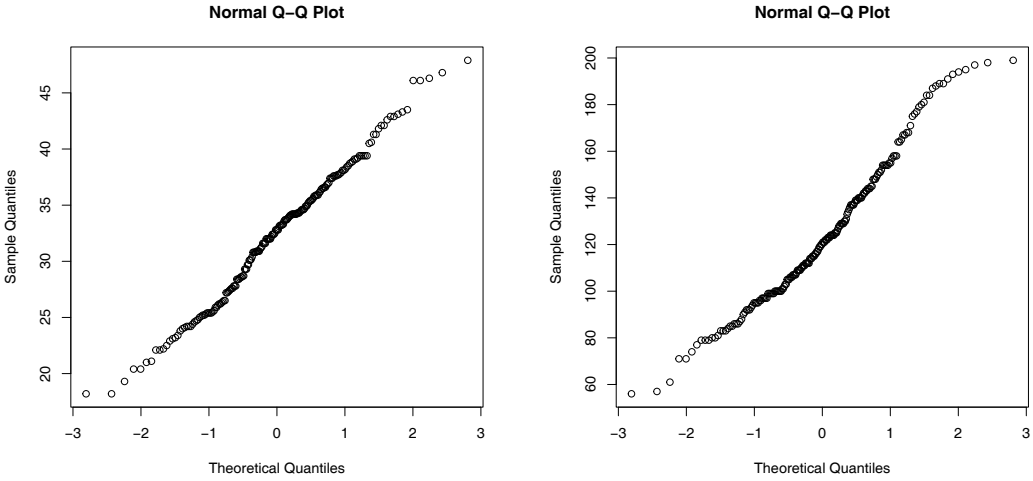


Figura 3.13: Datos DIABETES. qq-plots das variables *bmi* (esquerda) e *glu* (dereita).

3.7.2. Tests de bondade de axuste de Normalidade

Á parte das ferramentas gráficas xa estudadas (histograma e qq-plot), resulta imprescindible dispoñer tamén dalgún test de hipóteses para verificar se un modelo de distribución resulta axeitado para explicar o comportamento dunha variable aleatoria poboacional.

Sen dúbida o caso máis importante é o da distribución Normal. O problema é polo tanto o seguinte. Dada unha variable aleatoria poboacional X da cal se obtivo unha mostra $X_1, X_2, \ldots, X_n$, desexamos facer un test para a hipótese nula

$$H_0 : X \text{ ten distribución Normal}$$

fronte a unha alternativa xeral

$$H_1 : H_0 \text{ non é certa.}$$

Nótese que na hipótese nula se afirma que X é unha Normal calquera, sen especificar os seus parámetros μ e σ.

Existen moitos procedementos para resolver este problema. Describiremos brevemente dous deles: o test de Lilliefors e o test de Shapiro-Wilk.

- **O test de Lilliefors** baséase no estatístico

$$D = \sup_x |\hat{F}(x) - F_{(\hat{\mu},\hat{\sigma})}(x)|,$$

  onde $\hat{F}(x)$ é a función de distribución empírica da mostra (ver sección 3.3.7) e $F_{(\hat{\mu},\hat{\sigma})}(x)$ é a función de distribución dunha Normal con media estimada por $\hat{\mu} = \bar{X}$ e desviación estándar estimada por $\hat{\sigma} = S$. A hipótese nula de Normalidade rexéitase para valores grandes do estatístico de test. Os valores críticos foron tabulados por H.W. Lilliefors en 1967. En R, a función `lillie.test()`[5] realiza este test.

- **O test de Shapiro-Wilk** está baseado no qq-plot de Normalidade. Lembremos que o feito de que no qq-plot observemos unha disposición dos puntos sobre unha liña recta é un indicativo da Normalidade dos datos. Polo tanto, o coeficiente de correlación dos puntos que aparecen no qq-plot pode empregarse como un indicador do axuste do modelo Normal. Non daremos a fórmula explícita do estatístico de test porque é moi complicada. Como o estatístico de test está baseado no coeficiente de correlación, baixo H_0 o estatístico debe estar próximo a 1; por outra banda, cando a hipótese nula non é certa, o estatístico tomará valores menores ca 1. A hipótese nula rexéitase para valores pequenos do estatístico de test. Este test foi proposto e estudado por S.S. Shapiro e M. Wilk en 1965. Os valores críticos están tabulados. En R emprégase a funcion `shapiro.test()` para realizar o test.

Exemplo. No conxunto de datos DIABETES, fagamos os test de Normalidade das variables *bmi* e *glu*:

```
> lillie.test(bmi)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  bmi
D = 0.053375, p-value = 0.1779

> shapiro.test(bmi)
```

[5] A función `lillie.test()` non pertence á instalación básica de R. Requísere a instalación do paquete `nortest`. Para instalar e cargar o paquete, escribe na consola

```
>install.packages("nortest")
>library(nortest)
```

```
        Shapiro-Wilk normality test

data:  bmi
W = 0.99104, p-value = 0.2523
```

Os p-valores obtidos por ambos os dous tests son grandes, así que non existe evidencia para rexeitar que a variable *bmi* segue unha distribución Normal.

```
> lillie.test(glu)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  glu
D = 0.072282, p-value = 0.01288

> shapiro.test(glu)

        Shapiro-Wilk normality test

data:  glu
W = 0.97294, p-value = 0.0006624
```

Neste caso os dous tests devolven p-valores pequenos, polo tanto deberiamos rexeitar que a variable *glu* segue unha distribución Normal.

Que ocorre coa variable *glu* cando distinguimos os grupos de mulleres diabéticas e non diabéticas? (**exercicio**) □

Exercicio 3.15. *Considera o conxunto de datos* RATPUPS. *Proporciona a distribución Normal un bo axuste para a variable* weight*? Considera subgrupos separando por tratamento e sexo. Interpreta os resultados.*

3.8. Análise da varianza (ANOVA)

O termo Análise da Varianza (ANOVA, do inglés *ANalysis Of VAriance*) refírese a un conxunto de técnicas inferenciais deseñadas para comparar as medias dunha variable en k subgrupos, sendo $k > 2$. Habitualmente, os subgrupos fórmanse de acordo aos valores dun ou varios factores.

Recordemos que para $k = 2$ podemos empregar os t-tests xa estudados. Cando $k \geq 3$ podería resultar tentador facer as comparacións entre todos os posibles pares, pero este procedemento resulta inválido na práctica porque produce un incremento enorme do erro de tipo I, tal e como podemos ver no seguinte exemplo.

Táboa 3.5: Probabilidade de falso descubrimento cando se fan comparacións de grupos por pares.

		$P(R>0)$	
k	$k(k-1)/2$	$\alpha=0.05$	$\alpha=0.01$
2	1	0.050	0.010
3	3	0.143	0.030
4	6	0.265	0.059
5	10	0.401	0.096
6	15	0.537	0.140
7	21	0.659	0.190
8	28	0.762	0.245
9	36	0.842	0.304
10	45	0.901	0.364
15	105	0.995	0.652
20	190	1.000	0.852

Exemplo. O problema dos tests múltiples. Supoñamos que temos k grupos de individuos sobre os que medimos unha certa variable e que queremos comparar as medias correspondentes para poder descubrir diferenzas entre elas. Denotemos por μ_j a media poboacional no grupo j, para $j=1,\dots,k$. Poderiamos pensar en facer os tests

$$H_0^{j,\ell}:\mu_j=\mu_\ell \qquad \text{fronte a} \qquad H_1^{j,\ell}:\mu_j\neq\mu_\ell$$

para todo $j,\ell\in\{1,\dots,k\}$, con $j\neq\ell$. Nótese que existen $\binom{k}{2}=\frac{k(k-1)}{2}$ comparacións posibles. Chamémoslle "descubrimento" ao feito de que rexeitemos algunha destas hipóteses nulas. Se todos os tests se realizan a un nivel de significación α, cal é a probabilidade de que fagamos un descubrimento cando en realidade todas as medias son iguais (e polo tanto non hai nada que descubrir)? Esta probabilidade medra rapidamente ao aumentar k. Definamos a variable aleatoria $R=$"número de rexeitamentos entre as $\frac{k(k-1)}{2}$ comparacións". Se todas as hipóteses nulas son certas (é dicir, todas as medias son iguais), entón

$$R\sim Binomial\left(\frac{k(k-1)}{2},\alpha\right)$$

e a probabilidade de facer un (falso) descubrimento é $P(R>0)$. A Táboa 3.5 contén os valores desta probabilidade para distintos valores de k e para niveis de significación $\alpha=0.05,0.01$. Claramente, a probabilidade de facer falsos descubrimentos aumenta moito ao incrementar o número de grupos (e polo tanto o de comparacións). Obviamente, isto non son boas noticias. Está claro que esta non será a forma correcta de comparar máis de 2 grupos. □

Os métodos de tipo ANOVA evitan este problema facendo a comparación entre os grupos de forma simultánea. Aquí só veremos o caso máis sinxelo, que é aquel no que os grupos se forman a partir dos valores dun único factor e son independentes. Estes métodos foron ideados polo estatístico británico Ronald Fisher a partir dos anos 20 do século XX e son unha ferramenta fundamental no **deseño e análise de experimentos**.

3.8.1. ANOVA dun factor

O **ANOVA dun factor** (tamén chamado ANOVA dunha vía) emprégase para comparar k medias a partir de mostras independentes. Estas mostras obtéñense de poboacións ou grupos definidos polos valores dunha variable cualitativa ou, dito doutra forma, polos niveis dun factor. Ás veces a estes niveis tamén se lles chama tratamentos. A situación práctica é a seguinte:

/grupo/nivel/ /tratamento	media poboacional	mostra	tamaño mostral	media mostral
grupo 1	μ_1	$X_{11}, X_{12}, \dots, X_{1n_1}$	n_1	$\bar{X}_1$
grupo 2	μ_2	$X_{21}, X_{22}, \dots, X_{2n_2}$	n_2	$\bar{X}_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
grupo k	μ_k	$X_{k1}, X_{k2}, \dots, X_{kn_k}$	n_k	$\bar{X}_k$
total			$n = \sum_{i=1}^{k} n_i$	$\bar{X}$

No cadro anterior temos os seguintes elementos:

n_i é o tamaño mostral no grupo i.

$n = \sum_{i=1}^{k} n_i$ é o tamaño mostral total.

X_{ij} denota a j-ésima observación da i-ésima mostra.

$\bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}$ é a media mostral no grupo i.

$\bar{X} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij} = \frac{1}{n} \sum_{i=1}^{k} n_i \bar{X}_i$ é a media mostral global (tamén chamada "gran media").

A hipótese nula é

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k,$$

e a hipótese alternativa é

$$H_1 : \text{existe algunha diferenza entre as medias.}$$

Nótese que H_1 é a negación de H_0 e polo tanto non necesariamente implica que todas as medias sexan distintas entre si. O único que afirma é que existe algunha diferenza (podería ser por exemplo $\mu_1 = \mu_2$ e $\mu_2 \neq \mu_3$).

A idea do test ANOVA é estudar as fontes de variación nos datos. Cada observación pode descompoñerse como

$$X_{ij} = \bar{X} + \underbrace{(\bar{X}_i - \bar{X})}_{\text{desviación entre grupos}} + \underbrace{(X_{ij} - \bar{X}_i)}_{\text{desviación dentro dos grupos}}$$

e as fontes de variación nos datos poden resumirse nos seguintes dous termos:

- Variación entre grupos (representada por $\bar{X}_i - \bar{X}$):

$$\text{SSG} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{X}_i - \bar{X})^2 = \sum_{i=1}^{k} n_i (\bar{X}_i - \bar{X})^2.$$

- Variación dentro dos grupos (representada polos *residuos* $X_{ij} - \bar{X}_i$):

$$\text{SSR} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2.$$

Pódese demostrar facilmente que

$$\text{SSG} + \text{SSR} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2 = \text{TSS},$$

que é o numerador da varianza mostral da mostra conxunta. A esta cantidade, que denotamos por TSS, chámaselle variación total.

Habitualmente, toda esta información organízase na **táboa ANOVA**:

fonte de variación	graos de liberdade	suma de cadrados (SS)	media de cadrados (MS)	F-ratio
grupos	$k-1$	$\text{SSG} = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2$	$\text{MSG} = \frac{\text{SSG}}{k-1}$	$\frac{\text{MSG}}{\text{MSR}}$
residuos	$n-k$	$\text{SSR} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$	$\text{MSR} = \frac{\text{SSR}}{n-k}$	
total	$n-1$	$\text{TSS} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2$		

Se todas as medias son iguais, é dicir, non hai efecto do grupo sobre as medias, entón as medias mostrais en cada grupo deberían ser próximas entre si e á súa vez tamén próximas á media mostral global. Isto quere dicir que MSG debería tomar un valor pequeno e o mesmo debería ocorrer co cociente MSG/MSR. Por outra parte, se realmente existe algunha diferenza entre as medias, entón o cociente MSG/MSR debería tomar un valor grande. Polo tanto, a hipótese nula será rexeitada para valores grandes de MSG/MSR. Baixo determinadas condicións (ver sección 3.8.3), se a hipótese nula é certa, entón

$$\frac{\text{MSG}}{\text{MSR}} \sim F_{k-1,n-k},$$

onde $F_{k-1,n-k}$ denota a distribución F de Snedecor con $k-1$ e $n-k$ graos de liberdade. Os valores críticos e os p-valores poden obterse inmediatamente desta distribución. En R empregamos a función `aov()`.

Exemplo. (adaptado do libro de Vidakovic, 2017). Deséxase realizar un estudo para saber se distintos tipos de implantes cocleares teñen distinto efecto sobre a capacidade de comprensión

e de audición en persoas que sofren xordeira profunda. En particular, disponse de tres tipos de implantes: A, B e C. Os implantes A e B son similares en canto á tecnoloxía que empregan, mentras que o implante de tipo C é máis moderno. Cada individuo realiza unha proba despois da cal recibe unha puntuación. Neste exemplo, centrarémonos no recoñecemento das consoantes mediante son. O obxectivo é saber se a media da puntuacións dos individuos empregando os tres tipos de implantes son iguais (hipótese nula) ou existe algunha diferenza entre elas (hipótese alternativa). Aplicaremos un test ANOVA.

Para levar á práctica o test, empregaremos os datos recollidos no conxunto de datos IMPLANTECOCLEAR, que contén as seguintes variables:

- *implante*: tipo de implante (A, B ou C).
- *CS*: puntuación no recoñecemento de consoantes a través do son.
- *CV*: puntuación no recoñecemento de consoantes a través da visión.
- *CSV*: puntuación no recoñecemento de consoantes a través do son e da visión.

O conxunto de datos contén información de 78 individuos, dos cales 24 empregaron o implante de tipo A, 24 empregaron o implante de tipo B e 30 empregaron o implante de tipo C. En primeiro lugar facemos un boxplot para comprobar visualmente se hai diferenzas entre as puntuacións (ver Figura 3.14).

```
> implantecoclear <- read.table("datos-implantecoclear.txt",header=TRUE)
> attach(implantecoclear)

> boxplot(CS~implante,frame=FALSE)
```

Comprobemos agora se as medias da variable *CS* difiren para cada tipo de implante. En R a función `aov()` crea unha obxecto que contén a información da análise ANOVA. O primeiro argumento da función é unha fórmula do tipo `x` $\sim$ `factor`, onde `x` indica a variable da cal queremos comparar as medias segundo os grupos formados polos niveis do `factor`[6].

No noso exemplo:

```
> anova.implantes <- aov(CS~implante)
> summary(anova.implantes)

            Df Sum Sq Mean Sq F value   Pr(>F)
implante     2  12312    6156   15.05 3.21e-06 ***
Residuals   75  30688     409
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

[6]Nalgunhas ocasións o niveis dos factores poden estar codificados mediante números. Nese caso, é posible que R non recoñeza esa variable como factor, cousa que daría resultados erróneos no ANOVA. Para facelo correctamente, débese empregar a función `as.factor()` empregando a fórmula `x` $\sim$ `as.factor(factor)`.

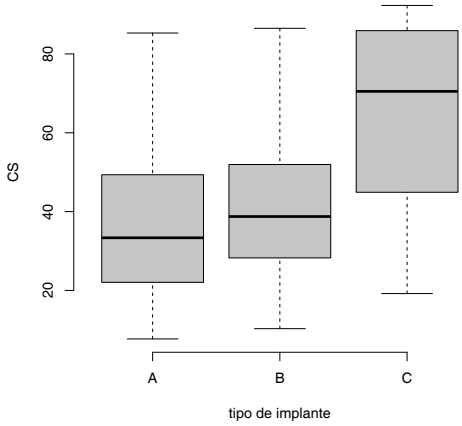


Figura 3.14: Conxunto de datos IMPLANTECOCLEAR. Boxplot da variable *CS* en función do tipo de implante.

Na táboa ANOVA podemos ver que o valor do estatístico de test é 15.05. O valor crítico para o nivel de significación 0.05 é `qf(0.95,2,75) =` 2.658, así que o valor observado do estatístico de test claramente pertence á rexión crítica. De feito, o p-valor é moi pequeno ($3.21 \cdot 10^{-6}$). Isto indica unha forte evidencia en contra da hipótese nula de igualdade de medias. Parece claro que existen diferenzas significativas nas puntuacións medias para os distintos tipos de implante. □

3.8.2. Tests *post-hoc* para comparacións múltiples

Na práctica, en caso de rexeitar a hipótese nula, gustaríanos saber onde reside a diferenza entre os grupos. Isto debe ser feito a través dalgún procedemento que teña en conta que se realizarán comparacións múltiples e que ao mesmo tempo controle o erro de tipo I. Os procedementos deseñados con este fin chámanse tests *post-hoc*. Posiblemente o máis empregado é o **test HSD de Tukey**[7].

Exemplo. (cont.) Apliquemos o test post-hoc de Tukey ao noso exemplo:

```
> TukeyHSD(anova.implantes)

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = CS ~ implante)

$implante
```

[7] As siglas HSD refírense a *honestly significant difference*.

```
         diff       lwr      upr     p adj
B-A  4.095833 -9.866669 18.05834 0.7633873
C-A 27.660833 14.414841 40.90683 0.0000111
C-B 23.565000 10.319007 36.81099 0.0001757
```

Á vista dos p-valores axustados (columna `p adj` da saída de R) atopamos diferenzas significativas entre os implantes C–A e C–B, pero non entre os implantes B–A. □

3.8.3. Suposicións do test ANOVA

O test ANOVA depende das seguintes **suposicións**:

- Normalidade en cada grupo (debemos usar técnicas gráficas ou tests de bondade de axuste para comprobalo).
- Varianzas iguais en todos os grupos. Isto pode comprobarse formalmente co **test de Bartlett**. En R emprégase a función `bartlett.test()`.

De todos os xeitos, a análise ANOVA é robusta respecto a desviacións das suposicións anteriores, especialmente cando os tamaños mostrais non son demasiados pequenos.

En caso de que estas suposicións sexan claramente violadas e os tamaños mostrais sexan pequenos pode substituírse a análise ANOVA por polo **test de Kruskal-Wallis**, que é un test non paramétrico. En R emprégase a función `kruskal.test()`.

Exemplo. (cont.) O test de Bartlett aplicado ao noso exemplo amosa que a hipótese de varianzas iguais non pode ser rexeitada, xa que o p-valor é grande:

```
> bartlett.test(CS~implante)


        Bartlett test of homogeneity of variances

data:  CS by implante
Bartlett's K-squared = 1.1619, df = 2, p-value = 0.5594
```

Os test de bondade de axuste tamén permiten supoñer que os datos non teñen unha gran desviación da Normalidade (**exercicio**). □

Exercicio 3.16. *Considera o conxunto de datos* IMPLANTECOCLEAR. *Realiza análises ANOVA para comparar as medias das variables* CV *e* CSV *segundo o tipo de implante. Comproba se se cumpren as suposicións de traballo do ANOVA.*

Exercicio 3.17. *Considera o conxunto de datos* RATPUPS. *Anteriormente comparamos as medias da variable* weight *para os niveis 'high' e 'low' do factor* treatment. *Non obstante, en realidade hai 3 niveis ('high', 'low' e 'control'), polo que se queremos facer un test global de comparación de medias debemos facer un test de tipo ANOVA.*

(a) Realiza o test adecuado para comparar as medias da variable weight *segundo os niveis do factor* treatment.

(b) Comproba as suposicións da análise ANOVA. Cúmprese a Normalidade? Son as varianzas homoxéneas?

(c) Compara os resultados do test ANOVA cos do test de Kruskal-Wallis.

(d) En caso de que o test ANOVA rexeite a hipótese nula de igualdade de medias, aplica o procedemento HSD de Tukey. Cal é a conclusión? Entre que pares de niveis se atopan diferenzas?

Exercicio 3.18. *Realizouse un experimento para comprobar se o diámetro dunha prótese metálica pode ter algún efecto sobre o desgaste na zona de unión. Realizáronse probas de desgaste con próteses de diamétros* 16, 28, 36 *e* 42 *mm e medíronse as rugosidades correspondentes, obténdose os seguintes resultados (en nm):*

diámetro	*rugosidade*									
16	*2.42*	*1.91*	*1.86*	*2.45*	*2.52*	*2.88*	*2.84*	*1.89*	*2.75*	*1.52*
	2.34	*1.16*	*2.47*	*1.66*	*2.21*					
28	*2.43*	*1.37*	*2.81*	*2.93*	*2.44*	*1.53*	*1.38*	*1.37*	*2.00*	*2.21*
	2.48	*2.75*								
36	*2.43*	*3.43*	*2.52*	*1.82*	*3.19*	*3.10*	*2.82*	*3.05*	*3.64*	*2.55*
42	*3.48*	*2.51*	*3.71*	*2.75*	*2.71*	*3.23*	*3.23*	*3.74*	*3.14*	*3.47*

Realiza un test ANOVA para comprobar se existen diferenzas nas medias da rugosidade en función do diámetro empregado. En caso de atopar diferenzas, explica cales son.

3.9. Test de Kolmogorov-Smirnov para comparar dúas distribucións

Supoñamos que temos dúas variables aleatorias continuas, X e Y, con funcións de distribución F e G, respectivamente. O **problema das dúas mostras** consiste en facer un test con hipótese nula

$$H_0 : F(x) = G(x) \text{ para todo } x,$$

fronte á alternativa xeral

$$H_1 : F(x) \neq G(x) \text{ para algún } x.$$

Nótese que neste caso a hipótese nula é máis restritiva ca a do t-test, xa que agora estamos poñendo unha condición sobre as distribucións completas, mentras que no t-test unicamente se poñía unha condición sobre as medias.

Para facer o test, disporemos de dúas mostras: $X_1, X_2, \dots, X_n$ de X e $Y_1, Y_2, \dots, Y_m$ de Y. Sexan

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}(X_i \leq x) \qquad \text{e} \qquad \hat{G}(y) = \frac{1}{m}\sum_{j=1}^{m} \mathrm{I}(Y_j \leq y)$$

as funcións de distribución empíricas obtidas a partir das mostras de X e Y, respectivamente (ver a sección 3.3.7 para máis detalles). O **estatístico de Kolmogorov-Smirnov** mide a distancia entre estas dúas funcións a través da distancia do supremo:

$$D_{KS} = \sup_x |\hat{F}(x) - \hat{G}(x)|.$$

Aínda que o estatístico de Kolmogorov-Smirnov se define a partir dun supremo, na práctica o seu cálculo resulta moi doado. Tendo en conta que as funcións de distribución empíricas son funcións constantes por pedazos, para calcular D_{KS} abonda con obter o seguinte máximo:

$$D_{KS} = \max_{1 \leq i \leq n+m} |\hat{F}(Z_i) - \hat{G}(Z_i)|,$$

onde $Z_1, Z_2, \dots, Z_{n+m}$ é a mostra *conxunta* $X_1, \dots, X_n, Y_1, \dots, Y_m$.

Cando a hipótese nula é certa, as funcións de distribución empíricas estarán cerca unha da outra porque as dúas estiman a mesma función. Nese caso esperamos que o estatístico tome valores pequenos. En cambio, cando a hipótese nula é falsa, as funcións de distribución empíricas estimarán funcións distintas, e polo tanto esperamos que a distancia entre elas sexa grande. Polo tanto, rexeitaremos a hipótese nula para valores grandes do estatístico de test.

A distribución do estatístico D_{KS} baixo a hipótese nula depende dos tamaños mostrais n e m, pero sorprendentemente non depende da distribución común F ou G, de xeito que pode estudarse exactamente e consecuentemente pódense obter os correspondentes cuantís e p-valores asociados. Para valores pequenos de n e m incluso existen táboas da distribución de D_{KS}. Este procedemento para a comparación de dúas distribucións foi proposto e estudado polos matemáticos rusos Andrei Kolmogorov e Nikolai Smirnov na década de 1930.

En R, a función `ks.test()` realiza o test de Kolmogorov-Smirnov.

Exemplo. Consideremos o conxunto de datos RATPUPS. Son diferentes as distribucións da variable *weight* nos grupos de tratamento alto e baixo? Recordemos que o t-test para comparar as medias non atopou diferenzas significativas.

Sexan F e G as funcións de distribución da variable *weight* no grupos de tratamento alto (*high*) e baixo (*low*), respectivamente. Queremos polo tanto realizar un test con hipótese nula $H_0 : F = G$ fronte a $H_1 : F \neq G$. Os tamaños mostrais son 65 e 126. A Figura 3.15 amosa as correspondentes funcións de distribución empíricas. Para realizar os test en R, escribimos

```
> attach(ratpups)
> ks.test(weight[treatment=="low"],weight[treatment=="high"])
```

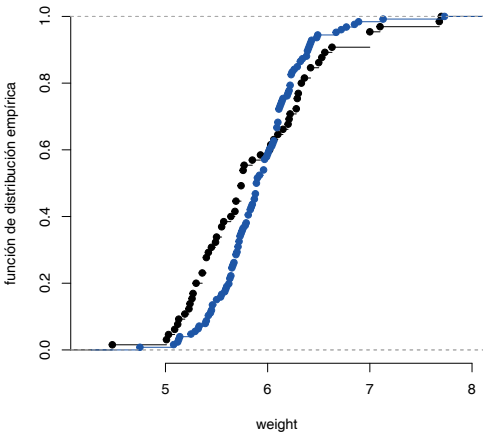


Figura 3.15: Conxunto de datos RATPUPS. Funcións de distribución empíricas da variable *weight* segundo os valores *high* (en negro) e *low* (en azul) da variable *treatment*.

```
        Two-sample Kolmogorov-Smirnov test

data:  weight[treatment == "low"] and weight[treatment == "high"]
D = 0.21795, p-value = 0.03403
alternative hypothesis: two-sided
```

A un nivel de significación do 5 %, rexéitase a hipótese nula. Parace polo tanto que hai diferenzas entre as funcións de distribución F e G. Quizas estas diferenzas estean na variabilidade das variables correspondentes. □

Exercicio 3.19. *Supoñamos que temos dúas variables aleatorias das cales sospeitamos que poden ter distribucións diferentes. Que procedemento será mellor para detectar esa posible diferenza: o t-test ou o test de Kolmogorov-Smirnov? Deseña un estudo de Monte Carlo para analizar a potencia destes dous tests ante distintas alternativas. Considera, por exemplo, situacións onde...*

- *a diferenza estea nas medias, como unha* $N(10, 3)$ *e unha* $N(12, 3)$*;*
- *ou nas varianzas, como unha* $N(10, 3)$ *e unha* $N(10, 5)$*.*

Realiza a simulación para varios tamaños mostrais para comprobar o seu efecto na potencia dos tests.

3.10. As etapas do método científico

A Estatística, e en particular as técnicas inferenciais, xoga un papel fundamental no desenvolvemento da ciencia. Todos os ámbitos científicos empregan métodos inferenciais para describir e confirmar os seus avances. Convén recordar que, en liñas xerais, o **método científico** consiste nos seguintes pasos:

1. Formular unha **hipótese de investigación**/teoría/problema.
2. Deseñar un experimento e recoller **datos**.
3. Analizar os datos e aplicar algún **método inferencial** axeitado para o problema en cuestión (por exemplo, un test de hipóteses). É importante ter en conta que calquera método estará baseado nunha serie de suposicións que teñen que ser verificadas para garantir o seu correcto funcionamento.
4. Interpretar os **resultados** e tomar as decisións que correspondan.

É importante ter presentes estes pasos cando se apliquen técnicas inferenciais.

Capítulo 4

Táboas de continxencia

Contidos

4.1. Introdución

As **táboas de continxencia** empréganse para organizar a información relativa a dúas (ou máis) variables cualitativas. Resultan útiles tanto desde o punto descritivo como para estudar posibles relacións entre as variables.

Exemplo. O conxunto de datos SAUDEGALICIA2017 recolle información extraída da Enquisa Nacional de Saúde que realizou o INE en 2017. En particular, este conxunto de datos recolle información sobre 400 persoas residentes en Galicia. As variables son as seguintes:

- *sexo*: muller (M) / home (H)
- *idade* (en anos)
- *nivel.estudos*. Máximo nivel de estudos alcanzado, coas seguintes categorías:
  1 primaria incompleta
  2 primaria completa
  3 secundaria
  4 bacharelato e ensinanzas profesionais de grao medio
  5 ensinanzas profesionais de grao superior
  6 estudos universitarios
- *estado.saude*. Estado xeral de saúde, cos seguintes niveis:
  1 moi bo
  2 bo
  3 regular
  4 malo
  5 moi malo
- *hipertension*: pacede hipertensión? Non (0) / Si (1)
- *artrose*: pacede artrose? Non (0) / Si (1)
- *diabetes*: pacede diabetes? Non (0) / Si (1)
- *colesterol*: ten colesterol alto? Non (0) / Si (1)
- *osteoporose*: pacede osteoporose? Non (0) / Si (1)
- *gafas*: usa gafas ou lentes de contacto? Non (0) / Si (1)
- *audifono*: usa audífonos? Non (0) / Si (1)
- *medicamentos*: consumiu medicamentos nas últimas dúas semanas? Non (0) / Si (1)
- *altura* (en cm)

- *peso* (en kg)
- *act.principal*. Tipo de actividade física que realiza na súa actividade diaria principal, cos seguintes niveis:

  1 sentado/a a maior parte da xornada

  2 de pé a maior parte da xornada sen efectuar grandes desprazamentos ou esforzos

  3 camiñando, levando algún peso, efectuando desprazamentos frecuentes

  4 realizando tarefas que requiren gran esforzo físico

- *act.fisica*. Frecuencia coa que realiza algunha actividade física no seu tempo libre, cos seguintes niveis:

  1 non fai exercicio, o tempo libre ocúpao de forma case totalmente sedentaria

  2 fai algunha actividade física ou deportiva ocasional

  3 fai actividade física varias veces ao mes

  4 fai entrenamento deportivo ou físico varias veces á semana

De que tipo é cada unha destas variables? (**exercicio**).

A Táboa 4.1 é a táboa de continxencia das variables sexo e estado xeral de saúde (*estado.saude*). Cada entrada da táboa representa a frecuencia absoluta de individuos que cumpren as características correspondentes. Por exemplo, nos nosos datos hai

- 186 homes;
- 24 persoas (homes e mulleres) que declaran un estado de saúde bo (nivel 1);
- e 14 homes que declaran un estado de saúde moi bo (nivel 1).

Táboa 4.1: Datos SAUDEGALICIA2017. Táboa de continxencia das variables *sexo* e *estado.saude*.

	estado de saúde					
sexo	1	2	3	4	5	total
home	14	96	62	10	4	186
muller	10	104	74	21	5	214
total	24	200	136	31	9	400

En R facemos

```
> saudegalicia2017 <- read.csv(file="datos-saudegalicia2017.csv")
> attach(saudegalicia2017)
> table(sexo,estado.saude)
```

e para obter os totais por filas e columnas podemos facer simplemente

```
> table(sexo)
> table(estado.saude)
```

□

De xeito máis formal, sexan dúas variables cualitativas X (con niveis $A_1, A_2, \dots, A_k$) e Y (con niveis $B_1, B_2, \dots, B_m$) das cales se observa unha mostra $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$. A **táboa de continxencia** de (X, Y) é

	niveis de Y				
niveis de X	B_1	B_2	$\cdots$	B_m	total
A_1	N_{11}	N_{12}	$\cdots$	N_{1m}	$N_{1\bullet}$
A_2	N_{21}	N_{22}	$\cdots$	N_{2m}	$N_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_k	N_{k1}	N_{k2}	$\cdots$	N_{km}	$N_{k\bullet}$
total	$N_{\bullet 1}$	$N_{\bullet 2}$	$\cdots$	$N_{\bullet m}$	n

onde, para $j = 1, \dots, k$ e $l = 1, \dots, m$,

$$\begin{aligned} N_{jl} &= \text{“número de observacións } (X_i, Y_i) \text{ tales que } X_i = A_j \text{ e } Y_i = B_l\text{”}, \\ N_{j\bullet} &= \text{“número de observacións } X_i \text{ tales que } X_i = A_j\text{”}, \\ N_{\bullet l} &= \text{“número de observacións } Y_i \text{ tales que } Y_i = B_l\text{”}. \end{aligned}$$

4.2. Distribución conxunta, marxinal e condicionada

Se dividimos as frecuencias absolutas N_{jl} polo tamaño mostral n obtemos as frecuencias relativas, que forman a **distribución conxunta**. Por outra parte, as frecuencias relativas das variables X e Y por separado (é dicir, as proporcións da forma $N_{j\bullet}/n$ e $N_{\bullet l}/n$, respectivamente) forman as **distribucións marxinais** correspondentes. As marxinais tamén poden obterse sumando as filas ou columnas da distribución conxunta.

Exemplo. (cont.) Para obter a distribución conxunta en R podemos facer

```
> taboa.continxencia <- table(sexo,estado.saude)
> n <- sum(taboa.continxencia)
> distrib.conxunta <- taboa.continxencia/n
```

e para as marxinais

```
> distrib.marxinal.sexo <- table(sexo)/n
```

A distribución conxunta e as marxinais deste exemplo están recollidas na Táboa 4.2. □

Se desexamos facer unha comparativa entre as frecuencias recollidas na táboa de continxencia para comprobar se o seu comportamento é similar ao longo dos distintos niveis dunha

Táboa 4.2: Datos SAUDEGALICIA2017. Distribución conxunta e distribucións marxinais das variables *sexo* e *estado.saude*.

	estado de saúde					
sexo	1	2	3	4	5	total
home	0.035	0.240	0.155	0.025	0.010	0.465
muller	0.025	0.260	0.185	0.052	0.013	0.535
total	0.060	0.500	0.340	0.077	0.023	1

das variables, temos que ter en conta que os tamaños mostrais totais poden ser distintos e polo tanto a comparación non se pode facer directamente. O mesmo ocorre coas proporcións da distribución conxunta.

Exemplo. (cont.) No noso exemplo, poderiamos estar interesados en saber como se distribúen os distintos niveis do estado de saúde no grupo de homes e de mulleres para intentar ver se hai diferenzas entre eles. Como o número de homes e mulleres é distinto (186 e 214, respectivamente), entón non podemos facer a comparación directamente. □

Nesta situación, é máis conveniente traballar coas **distribucións condicionadas**, que nos indican como se distribúe unha variable coa suposición de que a outra toma un determinado valor. A **distribución de X condicionada polo valor B_l de Y** está formada polas frecuencias relativas

$$\frac{N_{jl}}{N_{\bullet l}}, \quad \text{para } j = 1, \dots, k.$$

Analogamente, **a distribución de Y condicionada polo valor A_j de X** está formada polas frecuencias relativas

$$\frac{N_{jl}}{N_{j\bullet}}, \quad \text{para } l = 1, \dots, m.$$

Exemplo. (cont.) A distribución da variable *estado.saude* condicionada polo valor "home" da variable *sexo* pode empregarse para saber como se comportan as frecuencias dos distintos niveis do estado de saúde exclusivamente dentro do grupo de homes. En R podemos facer

```
> table(sexo,estado.saude)[1,]/sum(sexo=="H")
```

e obteremos a distribución recollida na Táboa 4.3.

Táboa 4.3: Datos SAUDEGALICIA2017. Distribución condicionada da variable *estado.saude* no grupo de homes.

estado de saúde	1	2	3	4	5	total
frecuencia relativa	0.075	0.516	0.333	0.054	0.022	1

□

Exercicio 4.1. *Atopa a distribución condicionada do estado xeral de saúde no grupo de mulleres. Paréceche que hai diferenzas entre o grupo de homes e o grupo de mulleres?*

Exercicio 4.2. *Calcula a táboa de continxencia das variables* sexo *e* act.fisica*. Calcula a distribución conxunta e as distribucións marxinais. Calcula a distribución condicionada da variable* actividade.fisica *no grupo de homes e no grupo de mulleres. Aprécianse diferenzas entre as dúas distribucións?*

Exercicio 4.3. *No conxunto de datos* SAUDEGALICIA2017 *hai algunhas variables cuantitativas (*idade, peso *e* altura*). Tamén podemos facer táboas de continxencia destas variables para relacionalas con outras variables nominais. Para iso teremos que discretizalas en intervalos facendo uso da función* `cut()` *que vimos na Sección 1.2.1.*
A partir das variables peso *e* altura *calcula o índice de masa corporal e despois clasifica os seus valores segundo os intervalos que establece a OMS*[a] *(infrapeso:* < 18.5*; peso normal:* $18.5 - 24.9$*; pre-obesidade:* $25.0 - 29.9$*; obesidade clase I ou obesidade leve:* $30.0 - 34.9$*; obesidade clase II ou obesidade media:* $35.0 - 39.9$*; obesidade clase III ou obesidade mórbida:* ≥ 40*). Contrúe a continuación unha táboa de continxencia da variable* diabetes *e a versión discretizada do índice de masa corporal. Hai diferenzas nas distribucións condicionadas para o grupo de persoas diabéticas e non diabéticas?*

[a]Para máis información, véxase a web https://www.euro.who.int/en/health-topics/disease-prevention/nutrition/a-healthy-lifestyle/body-mass-index-bmi

Exercicio 4.4. *Fai unha táboa de continxencia das variables* artrose *e* idade*. Compara as distribucións condicionadas.*

4.3. O gráfico de mosaico

Unha posible visualización gráfica das táboas de continxencia é o **gráfico de mosaico**. Consiste en rectánculos de área proporcional aos valores da distribución conxunta. En R obtense a partir da función `mosaic()`[1]:

```
> library(vcd)
> mosaic(table(sexo,estado.saude))
```

A Figura 4.1 amosa dous gráficos de mosaico. O da esquerda é o correspondente á táboa de continxencia das variables *sexo* e *estado.saude*. O da dereita é o que se obtén coas variables *artrose* e *estado.saude*. Nótese que neste segundo exemplo na parte inferior esquerda non aparece

[1]A función `mosaic()` pertence ao paquete `vcd`. Antes de empregar esta función hai polo tanto que instalar o paquete mediante `install.packages("vcd")` e despois cargalo á sesión de traballo mediante `library(vcd)`.

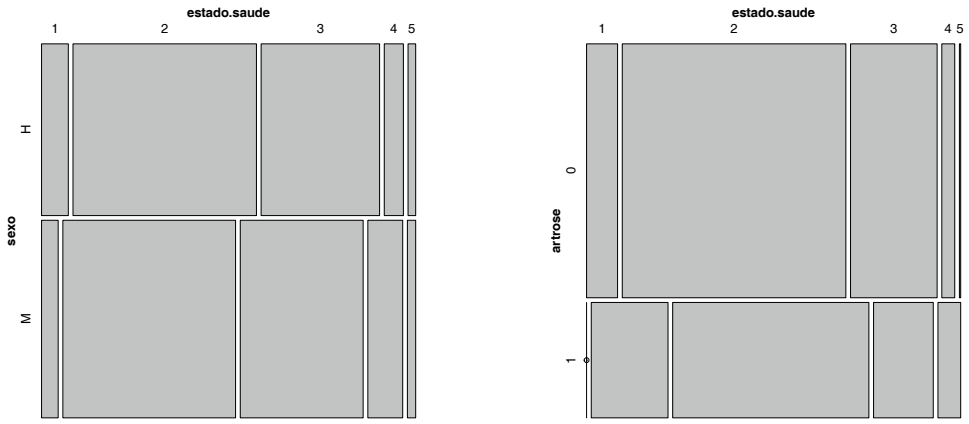


Figura 4.1: Datos SAUDEGALICIA2017. Gráficos de mosaico. Esquerda: variables *sexo* e *estado.saude*. Dereita: variables *artrose* e *estado.saude*.

un rectángulo, senón nunha liña marcada cun círculo. Isto emprégase para indicar que nos datos non hai ningún individuo que padeza artrose e declare estado de saúde "moi bo" (nivel 1).

Que conclusións podemos sacar destes gráficos? (**exercicio**).

4.4. O test Chi-cadrado de independencia

As táboas de continxencia tamén se poden empregar para comprobar a posible dependencia entre variables. A dependencia entre variables pode entenderse de moitas formas. No noso exemplo coas variables sexo e estado de saúde podemos pensar as seguintes posibilidades para hipóteses nulas e alternativas:

(a) H_0 : o estado de saúde dunha persoa é independente do seu sexo (é dicir, o estado de saúde non está asociado co sexo),

fronte a

H_1 : o estado de saúde dunha persoa depende do sexo (é dicir, o estado de saúde está asociado co sexo).

(b) H_0 : a ratio entre homes e mulleres é a mesma para cada nivel do estado de saúde,

fronte a

H_1 : a ratio entre homes e mulleres non é a mesma en cada nivel do estado de saúde.

(c) H_0 : as proporcións de persoas en cada un dos niveis do estado de saúde son as mesmas para os dous sexos,

fronte a

H_1 : as proporcións de persoas en cada un dos niveis do estado de saúde non son as mesmas para os dous sexos.

Calquera destas formulacións é válida para verificar a posible relación entre as variables. No caso (a) temos unha hipótese nula de *independencia*. Nos casos (b) e (c) temos hipóteses nulas de *homoxeneidade*, que basicamente din que as distribucións condicionadas dunha das variables son iguais para cada valor da outra variable. As tres formulacións son equivalentes entre si.

Para formalizar o test, pensemos entón que temos unha mostra $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ do par de variables (X, Y). Queremos facer un test con hipótese nula

$$H_0 : X \text{ e } Y \text{ son } \textbf{independentes}$$

fronte á alternativa

$$H_1 : X \text{ e } Y \text{ non son independentes.}$$

Rexeitar a hipótese nula significaría polo tanto que hai evidencia a favor de que existe algún tipo de relación entre as variables.

O procedemento máis empregado é o **test da Chi-cadrado**, ideado por Karl Pearson a principios do século XX. Esencialmente o que fará este test será comparar as frecuencias N_{jl} da táboa de continxencia (chamadas "frecuencias observadas") coas frecuencias que se esperarían se as variables fosen independentes (chamadas "frecuencias esperadas"). O estatístico de test é

$$\chi^2 = \sum_{\text{todas as clases}} \frac{(\text{frecuencias observadas} - \text{frecuencias esperadas})^2}{\text{frecuencias esperadas}}.$$

As frecuencias esperadas constrúense as partir das distribucións marxinais da seguinte forma:

$$\frac{N_{j\bullet} N_{\bullet l}}{n}, \quad \text{para } j = 1, \dots, k \quad \text{e} \quad l = 1, \dots, m.$$

Así, o estatístico de test resulta

$$\chi^2 = \sum_{j=1}^{k} \sum_{l=1}^{m} \frac{(N_{jl} - N_{j\bullet} N_{\bullet l}/n)^2}{N_{j\bullet} N_{\bullet l}/n}.$$

A hipótese nula será rexeitada para valores grandes do estatístico de test. Baixo a hipótese nula de independencia, a distribución do estatístico χ^2 pódese aproximar por unha chi-cadrado con $(k-1)(m-1)$ graos de liberdade, $\chi^2_{(k-1)(m-1)}$. A obtención dos correspondentes valores críticos e p-valores é polo tanto inmediata.

Exemplo. (cont.) Fagamos un test para comprobar a posible relación entre o estado de saúde e o sexo. A táboa de frecuencias observadas era

	estado de saúde					
sexo	1	2	3	4	5	total
home	14	96	62	10	4	186
muller	10	104	74	21	5	214
total	24	200	136	31	9	400

Para calcular a primeira frecuencia esperada facemos $186 \cdot 24/400 = 11.16$. A táboa completa de frecuencias esperadas é:

	estado de saúde					
sexo	1	2	3	4	5	total
home	11.16	93.00	63.24	14.41	4.18	186
muller	12.84	107.00	72.76	16.59	4.82	214
total	24	200	136	31	9	400

O estatístico de test é

$$\chi^2 = \frac{(14 - 11.16)^2}{11.16} + \frac{(96 - 93.00)^2}{93.00} + \cdots + \frac{(5 - 4.82)^2}{4.82} = 4.12.$$

Os graos de liberdade son $(2-1)\cdot(5-1) = 4$ e o p-valor resulta $P(\chi^2_4 > 4.12) = 0.39$. Polo tanto non hai evidencia suficiente para rexeitar a hipótese nula. Non parece que haxa ningún tipo de relación entre estas dúas variables. Dito doutra forma, o estado de saúde non depende do sexo.

En R, a función `chisq.test()` calcula o estatístico e o correspondente p-valor do test da Chi-cadrado. O argumento é unha táboa de continxencia:

```
> chisq.test(table(sexo,estado.saude))

        Pearson's Chi-squared test

data:  table(sexo, estado.saude)
X-squared = 4.12, df = 4, p-value = 0.39
```

□

Nas aplicacións prácticas deben terse en conta as seguintes **recomendacións** en relación ao uso do test da Chi-cadrado:

- Para que a aproximación á distribución chi-cadrado sexa correcta, as frecuencias esperadas estimadas non deben ser moi pequenas. En xeral, acostúmase pedir que as frecuencias esperadas sexan ≥ 5. Non obstante, esta condición pode resultar bastante restritiva (especialmente se traballamos cun número grande de niveis) e en moitos casos é innecesaria para que a aproximación sexa boa. Se a cantidade de frecuencias observadas que a incumpren é pequena e ademais esas frecuencias non son excesivamente pequenas (digamos, son maiores de 2), entón non debemos preocuparnos.

- Se hai moitas frecuencias que incumpren a condición ou algunhas son moi pequenas, entón pódense combinar dúas ou máis filas ou columnas para obter frecuencias esperadas de magnitude suficiente. Obviamente, isto conleva unha perda de información.

- O test da Chi-cadrado tamén se pode empregar para variables continuas sempre que se lles aplique unha discretización. Non obstante, hai que ter en conta que distintas discretizacións poderían dar lugar a conclusións distintas.

Exemplo. Consideremos agora o caso das variables *artrose* e *estado.saude*:

```
> chisq.test(table(artrose,estado.saude))

        Pearson's Chi-squared test

data:  table(artrose, estado.saude)
X-squared = 97.405, df = 4, p-value < 2.2e-16
```

Á vista do p-valor parece bastante claro que entre estas dúas variables existe unha relación.

Neste caso hai unha frecuencia esperada moito menor que 5, en concreto a correspondente á combinación *artrose* $= 1$ e *estado.saude* $= 5$:

```
> chisq.test(table(artrose,estado.saude))$expected

       estado.saude
artrose    1     2    3       4      5
      0 16.5 137.5 93.5 21.3125 6.1875
      1  7.5  62.5 42.5  9.6875 2.8125
```

Poderiamos pensar entón en colapsar as columnas cuarta e quinta. Para facer isto creamos unha nova variable a partir de *estado.saude* na que os niveis 4 e 5 aparecen colapsados no nivel 4:

```
> estado.saude2 <-
+          estado.saude*(estado.saude==1 | estado.saude==2 | estado.saude==3)
>          + 4*(estado.saude== 4 | estado.saude==5)
```

Fagamos o test con esta nova variable:

```
> chisq.test(table(artrose,estado.saude2))

        Pearson's Chi-squared test

data:  table(artrose, estado.saude2)
X-squared = 95.953, df = 3, p-value < 2.2e-16
```

O resultado é moi similar ao que obtivemos antes de combinar as columnas. □

Exercicio 4.5. *Empregando a táboa de continxencia obtida no exercicio 4.3, realiza un test de hipóteses para saber se o feito de padecer diabetes garda algún tipo de relación co índice de masa corporal. Comproba as recomendacións do test da Chi-cadrado, e, en caso de que non se cumpran, considera unha discretización distinta do índice de masa corporal.*

Unha vez que rexeitamos a hipótese de independencia pode interesarnos cuantificar o grao de asociación entre as variables. Existen moitas posibilidades para facer isto, pero quizais a medida máis coñecida é o **coeficiente V de Cramér**, que se calcula como

$$V = \sqrt{\frac{\chi^2}{n \min\{k-1, m-1\}}},$$

onde χ^2 é o valor do estatístico do test da Chi-cadrado. O coeficiente V sempre está entre 0 e 1. Se as variables son independentes, V será próximo a 0. Por outra parte, canto máis próximo a 1, maior será o grao de asociación entre as variables.

Exemplo. Calculemos a V de Cramér para os exemplos que estabamos analizando:

- variables *sexo* e *estado.saude*: $V = \sqrt{4.12/(400 \cdot 1)} = 0.101$.
- variables *artrose* e *estado.saude*: $V = \sqrt{97.405/(400 \cdot 1)} = 0.493$. □

4.5. Táboas 2×2: proporcións, riscos relativos e odd-ratios

O caso das táboas 2×2 resulta útil para identificar o efecto dunha das variables sobre a outra a través de proporcións.

Supoñamos que temos unha táboa de continxencia de dúas variables nominais **binarias** (si/non, home/muller, enfermo/san, tratamento/control, tratamento A/tratamento B etc.). Para simplificar a notación, identificamos os dous posibles valores con 0 e 1. A táboa de continxencia será da forma seguinte:

	Y		
X	0	1	total
0	a	b	$a+b$
1	c	d	$c+d$
total	$a+c$	$b+d$	$a+b+c+d=n$

Sobre esta táboa podemos calcular proporcións interesantes e tamén candidades relacionadas con elas. Vexámolas a partir dun exemplo.

Exemplo. Estamos interesados en saber se existe algunha relación entre o sexo e o feito de padecer artrose. A táboa de continxencia que se obtén a partir dos datos SAUDEGALICIA2017 é a seguinte:

	artrose		
sexo	non	si	total
home	143	43	186
muller	132	82	214
total	275	125	400

Agora traballaremos con algunhas proporcións/probabilidades que poden resultar interesantes na práctica.

Dentro do grupo de homes, cal é a proporción de persoas que padecen artrose? Podémola estimar por

$$\frac{43}{186} = 23.1\,\%$$

Formalmente, esta proporción corresponderíase cunha probabilidade condicionada da forma $P(Y = 1 \mid X = 0)$, que, a partir da táboa xenérica será estimada mediante a proporción mostral

$$\frac{b}{a+b}.$$

No grupo de mulleres, a proporción correspondente é

$$\frac{82}{214} = 38.3\,\%.$$

De xeito análogo, a probabilidade correspondente será $P(Y = 1 \mid X = 1)$, que se estima a través da proporción mostral

$$\frac{d}{c+d}.$$

Parece polo tanto que a artrose é máis prevalente no grupo de mulleres. O cociente entre dúas probabilidades chámase **risco relativo** e permítenos comparalas. Por exemplo, un risco relativo interesante é

$$\text{risco relativo} = \frac{P(Y = 1 \mid X = 1)}{P(Y = 1 \mid X = 0)},$$

que nos permite cuantificar o efecto que teñen as características de X sobre a variable Y. A partir da táboa 2×2 esta cantidade estímase por

$$\frac{d/(c+d)}{b/(a+b)}.$$

No noso exemplo, o risco relativo de padecer artrose no grupo de mulleres con respecto ao grupo de homes é

$$\frac{82/214}{43/186} = 1.66,$$

é dicir, o feito de ser muller parece que incrementa o risco de padecer artrose.

Outra candidade interesante é a chamada odds. Dada unha probabilidade p, a **odds** correspondente é $o = p/(1-p)$. Obviamente, $p = o/(1+o)$. A odds pode interpretarse como a relación entre as probabilidades de que ocorra e non ocorra o correspondente suceso. Por exemplo, se $p = 0.75$, entón $o = 3$, o cal quere dicir que hai unha vantaxe 3 a 1 de que ocorra o evento. O cociente de odds tamén se pode empregar para comparar probabilidades e chámase **odds ratio**. Por exemplo, a odds ratio correspondente ao risco relativo anterior será

$$\text{odds ratio} = \frac{\dfrac{P(Y = 1 \mid X = 1)}{1 - P(Y = 1 \mid X = 1)}}{\dfrac{P(Y = 1 \mid X = 0)}{1 - P(Y = 1 \mid X = 0)}} = \frac{\dfrac{P(Y = 1 \mid X = 1)}{P(Y = 0 \mid X = 1)}}{\dfrac{P(Y = 1 \mid X = 0)}{P(Y = 0 \mid X = 0)}},$$

que se pode estimar por

$$\frac{\dfrac{d/(c+d)}{c/(c+d)}}{\dfrac{b/(a+b)}{a/(a+b)}} = \frac{d/c}{b/a} = \frac{a\,d}{b\,c}.$$

No exemplo da artrose/sexo obteriamos

$$\frac{143 \cdot 82}{43 \cdot 132} = 2.07.$$

En xeral é máis doado de interpretar o risco relativo ca a odds ratio. Cando a probalidade do evento de interese é pequena entón $o \approx p$, e polo tanto o risco relativo e a odds ratio son moi parecidos. □

Exercicio 4.6. *Considera as variables* sexo *e* osteoporose *do conxunto de datos* SAUDEGALICIA2017. *Calcula os risco relativo e a odds ratio de padecer osteoporose no grupo de mulleres con respecto ao grupo de homes. Realiza o test chi-cadrado para comprobar a posible relación entre estas dúas variables.*

Exemplo. Ás veces o cálculo do risco relativo pode facerse incluso tendo unicamente coñecemento parcial das probabilidades involucradas. A Dirección Xeral de Tráfico (DGT) lanzou no ano 2021 unha campaña para sensibilizar sobre o uso do cinto de seguridade. Nun dos vídeos que formaban parte desta campaña informábase que o 26 % das persoas falecidas en accidente de tráfico durante o ano 2020 non levaba o cinto de seguridade posto. Que conclusión podemos sacar desta información? Se o 26 % dos falecidos non levaba posto o cinto, entón quere dicir que o 74 % restante si que o levaba. A conclusión inxenua podería ser que como falece máis xente con cinto que sen cinto, entón é mellor non levalo posto. Obviamente, isto non é certo. Vexamos por que.

Claramente o 26 % ao que fai referencia da DGT é unha estimación da probabilidade de "non levar o cinto" sabendo que esa persoa "faleceu nun accidente de tráfico". Máis formalmente, a probabilidade á que fai referencia a DGT é a probabilidade condicionada

$$P(\bar{C} \mid F),$$

onde F ="falecer nun accidente de tráfico" e C ="levar o cinto de seguridade". A estimación desta probabilidade pode facerse perfectamente a partir dos rexistros da DGT, xa que é bastante razoable pensar que todos os accidentes con vítimas mortais quedarán perfectamente documentados. De aí sae o valor 0.26.

Realmente o que resulta máis interesante é estudar o efecto de levar o cinto de seguridade sobre a posibilidade de falecer nun accidente, é dicir, gustaríanos ter información sobre a probabilidade $P(F \mid \bar{C})$. Esta probabilidade é difícil de estimar, pero podemos obter información sobre ela traballando con riscos relativos. Consideremos o cociente

$$\frac{P(F \mid \bar{C})}{P(F \mid C)},$$

é dicir, estudaremos como o feito de levar ou non o cinto de seguridade modifica a probabilidade de falecer nun accidente de tráfico. Isto non deixa de ser un risco relativo. Aplicando o Teorema

de Bayes podemos reescribir o cociente anterior como

$$\text{risco relativo} = \frac{P(F \mid \bar{C})}{P(F \mid C)} = \frac{\dfrac{P(\bar{C} \mid F)\cancel{P(F)}}{P(\bar{C})}}{\dfrac{P(C \mid F)\cancel{P(F)}}{P(C)}} = \frac{P(\bar{C} \mid F)}{P(C \mid F)} \frac{P(C)}{P(\bar{C})}.$$

Afortunadamente, na expresión anterior puidemos desfacernos da probabilidade $P(F)$, que realmente sería moi difícil de estimar. En cambio as probabilidades $P(\bar{C} \mid F)$ e $P(C)$ son fáciles de estimar. A primeira delas precisamente podemos estimala polo 0.26 que menciona a DGT na súa campaña. Por outra parte, $P(C)$ pódese estimar mediante observación ou mediante enquisas. De feito, a propia DGT refire un estudo de 2017 que afirma que o uso do cinto de seguridade en España ronda o 80 %. Combinando estas informacións, obtemos que

$$\text{risco relativo} = \frac{0.26}{(1-0.26)} \frac{0.80}{(1-0.80)} = 1.41,$$

é dicir, o feito de non levar o cinto de seguridade aumenta nun 41 % a probabilidade de falecer nun accidente de tráfico.

A estimación do 0.80 para $P(C)$ semella bastante baixa. Un estudo do Comisariado Europeo do Automóbil de 2019 afirma que o uso do cinto de seguridade supera o 95 %, cousa que parece máis axustada á realidade. Con esta estimación, o erro relativo queda

$$\text{risco relativo} = \frac{0.26}{(1-0.26)} \frac{0.95}{(1-0.95)} = 6.68,$$

é dicir, o feito de non levar o cinto de seguridade multiplica case por 7 o risco de falecer nun accidente de tráfico. □

Capítulo 5

Regresión

Contidos

5.1. Gráfico de dispersión e coeficiente de correlación

En moitas ocasións recóllense varias variables dun mesmo individuo, tal e como ocorre nos nosos conxuntos de datos. Ademais de estudar as características particulares de cada variable, tamén é interesante a análise das posibles relacións entre elas.

En termos estatísticos, dicimos que dúas (ou máis) variables son **dependentes** se os valores dunha delas inflúen nos valores da outra, é dicir, se o feito de observar un determinado valor nunha delas nos dá algunha información sobre a(s) outra(s). En caso contrario, as variables son **independentes**.

As técnicas para analizar as posibles relacións entre variables dependen da súa natureza:

- Variables cualitativas: a análise pode facerse a través de táboas de continxencia (ver Capítulo 4).
- Variables cuantitativas: empregamos gráficos de dispersión, coeficiente de correlación, técnicas de regresión ou técnicas de análise multivariante.

Primeiro estudaremos técnicas para variables bidimensionais. Supoñamos que dispoñemos dunha mostra de n observacións (X_1, Y_1), (X_2, Y_2), ... , (X_n, Y_n) dunha variable bidimensional (X, Y). O **gráfico de dispersión** pode usarse para comprobar visualmente a existencia de posibles relacións entre X e Y. Constrúese simplemente debuxando os pares (X_i, Y_i) no plano.

Exemplo. Conxunto de datos SAUDEGALICIA2017. A Figura 5.1 amosa o gráfico de dispersión do par de variables (*altura*, *peso*). En R facemos

```
> saudegalicia2017 <- read.csv(file="datos-saudegalicia2017.csv",header=TRUE)
> attach(saudegalicia2017)
> plot(altura,peso)
```

Á vista do gráfico de dispersión, hai algún tipo de dependencia entre estas dúas variables? Por que? (Comproba, por exemplo, cales son os valores observados do peso cando o valor da altura é 150 e cando é 180.) De que tipo é esa relación? □

Para cuantificar numericamente a dependencia entre variables existen varias medidas. O máis empregado é o **coeficiente de correlación mostral**, que se define como

$$r_{XY} = \frac{S_{XY}}{S_X S_Y},$$

onde

$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$ é a **covarianza mostral** entre X e Y,

S_X é a desviación estándar mostral de X,

S_Y é a desviación estándar mostral de Y.

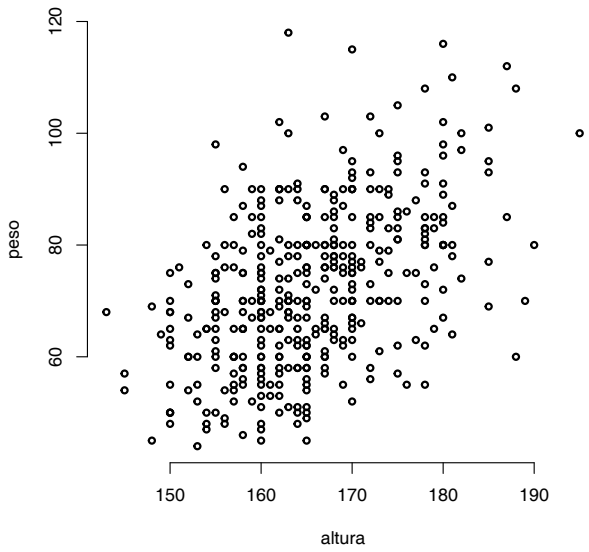


Figura 5.1: Datos SAUDEGALICIA2017. Gráfico de dispersión do par de variables (*altura*, *peso*).

Algunhas propiedades do coeficiente de correlación mostral (ver Figura 5.2):

- Non ten unidades.
- Emprégase para cuantificar o grao de relación lineal entre X e Y.
- Toma valores entre -1 e 1:
  - Valores de r_{XY} próximos a 1 indican unha relación lineal con pendente positiva entre X e Y. Valores de r_{XY} próximos a -1 indican unha relación lineal con pendente negativa entre X e Y.
  - Canto máis próximo sexa o valor de r_{XY} a 1 ou -1, máis forte é a dependencia entre X e Y.
  - Se $r_{XY} = 1$ (respectivamente, $r_{XY} = -1$) entón a variable Y pódese expresar exactamente a través dunha recta de X de pendente positiva (respectivamente, negativa).
  - Se $r_{XY} = 0$ entón non existe unha relación lineal entre X e Y, pero pode existir outro tipo de relación.
- Se X e Y son independentes, entón r_{XY} é aproximadamente cero (exactamente cero a nivel poboacional). O recíproco non é certo, é dicir, poder haber pares de variables que teñan coeficiente de correlación nulo, pero que non sexan independentes.

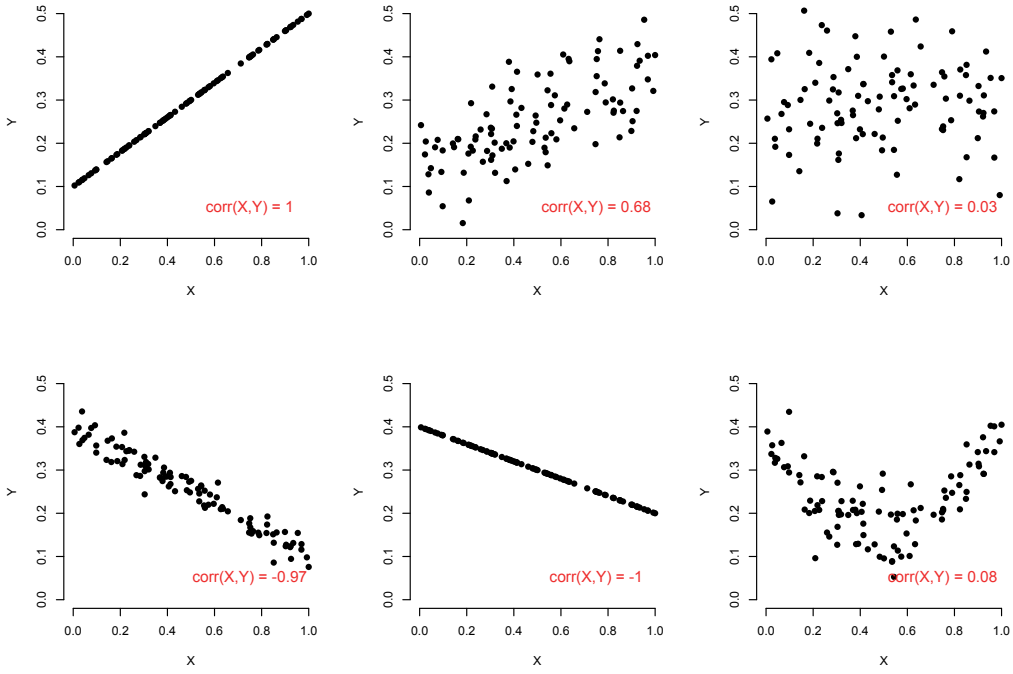


Figura 5.2: Exemplos de gráficos de dispersión e os seus correspondentes coeficientes de correlación.

En R as funcións `cov()` e `cor()` calculan a covarianza mostral e o coeficiente de correlación mostral, respectivamente.

Exemplo. (cont.) O coeficiente de correlación mostral entre *altura* e *peso* é 0.475. En R escribimos `cor(altura,peso)`. □

5.2. Regresión lineal simple: a recta de regresión

5.2.1. Estimación

O **modelo de regresión lineal simple** ou **recta de regresión** describe a relación entre unha **variable resposta** (tamén chamada **variable dependente**), Y, e unha **covariable** (tamén chamada **variable preditora**), X, a través dunha liña recta. Dada unha mostra de n observacións (X_i, Y_i), $i = 1, \dots, n$, dunha variable bidimensional (X, Y), o modelo de regresión lineal simple é

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

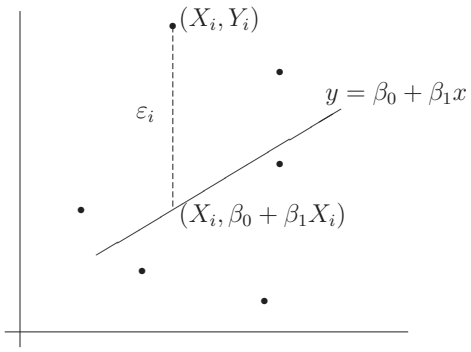


Figura 5.3: Representación gráfica dos erros de regresión: dada a recta $y = \beta_0 + \beta_1 x$, o erro de regresión correspondente á observación (X_i, Y_i) é $\varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$.

onde

- β_0 é o **intercepto**,
- β_1 é a **pendente**,
- ε_i, $i = 1, \ldots, n$, son os **erros de regresión**.

Na práctica, os parámetros da recta de regresión (intercepto, β_0, e pendente , β_1) son descoñecidos e terán que ser estimados a partir das observacións. Os estimadores obtéñense mediante **método dos mínimos cadrados**[1], que consiste en buscar os valores de β_0 e β_1 que minimizan a suma de erros ao cadrado (ver Figura 5.3). Para iso consideramos a seguinte función de (β_0, β_1):

$$m(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2.$$

Os estimadores dos coeficientes da recta de regresión polo método dos mínimos cadrados son as solucións do problema de optimización

$$\arg \min_{(\beta_0, \beta_1)} m(\beta_0, \beta_1).$$

A función m é derivable, así que para atopar o seu mínimo simplemente tomamos derivadas parciais con respecto a β_0 e β_1 e igualámolas a cero (nótese que as observacións X_i e Y_i son

[1] O método dos mínimos cadrados foi ideado a principios do século XIX polo matemático alemán Karl Gauss (1777–1855) para resolver un problema de astronomía. Simultaneamente tamén foi estudado polo matemático francés Adrien-Marie Legendre (1752–1833)

tratadas como constantes á hora de derivar)

$$\begin{aligned}\frac{\partial m(\beta_0,\beta_1)}{\partial \beta_0} &= -2\sum_{i=1}^{n}(Y_i-(\beta_0+\beta_1 X_i))=0,\\ \frac{\partial m(\beta_0,\beta_1)}{\partial \beta_1} &= -2\sum_{i=1}^{n}(Y_i-(\beta_0+\beta_1 X_i))X_i=0,\end{aligned}$$

e resolvemos o sistema de ecuacións lineais correspondente, que resulta

$$\begin{aligned} n\beta_0+\left(\sum_{i=1}^{n}X_i\right)\beta_1 &= \sum_{i=1}^{n}Y_i,\\ \left(\sum_{i=1}^{n}X_i\right)\beta_0+\left(\sum_{i=1}^{n}X_i^2\right)\beta_1 &= \sum_{i=1}^{n}X_iY_i,\end{aligned}$$

ou, escrito en forma matricial

$$\mathbf{X}^t\mathbf{X}\mathbf{B}=\mathbf{X}^t\mathbf{Y},$$

onde

$$\mathbf{X}=\begin{pmatrix}1 & X_1\\ 1 & X_2\\ \vdots & \vdots\\ 1 & X_n\end{pmatrix},\qquad \mathbf{B}=\begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix},\qquad \text{e}\qquad \mathbf{Y}=\begin{pmatrix}Y_1\\ Y_2\\ \vdots\\ Y_n\end{pmatrix}.$$

A solución é

$$\hat{\mathbf{B}}=(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y},$$

que admite as seguintes expresións explícitas dos **estimadores de mínimos cadrados** dos coeficientes da recta de regresión

$$\hat{\beta}_0=\frac{\left(\sum_{i=1}^{n}Y_i\right)\left(\sum_{i=1}^{n}X_i^2\right)-\left(\sum_{i=1}^{n}X_i\right)\left(\sum_{i=1}^{n}X_iY_i\right)}{n\sum_{i=1}^{n}X_i^2-\left(\sum_{i=1}^{n}X_i\right)^2},$$

$$\hat{\beta}_1=\frac{n\sum_{i=1}^{n}X_iY_i-\left(\sum_{i=1}^{n}Y_i\right)\left(\sum_{i=1}^{n}X_i\right)}{n\sum_{i=1}^{n}X_i^2-\left(\sum_{i=1}^{n}X_i\right)^2}.$$

Estes estimadores tamén se poden reescribir como

$$\hat{\beta}_0=\bar{Y}-\frac{S_{XY}}{S_X^2}\bar{X}\qquad \text{e}\qquad \hat{\beta}_1=\frac{S_{XY}}{S_X^2},$$

onde $\bar{X}$ e $\bar{Y}$ son as medias mostrais de X e Y, respectivamente, S_{XY} é a covarianza mostral entre X e Y e S_X^2 é a varianza mostral de X. A recta de regresión estimada é

$$y=\hat{\beta}_0+\hat{\beta}_1 x=\bar{Y}+\frac{S_{XY}}{S_X^2}(x-\bar{X}).$$

A **pendente** da recta de regresión, $\hat{\beta}_1$, pódese interpretar como a variación media da variable resposta Y por cada unidade de incremento na covariable X.

En R, a función `lm()` axusta modelos de regresión lineais, incluída a recta de regresión. A función crea un obxecto que contén moitos elementos relacionados coa análise de regresión.

Exemplo. (cont.) Datos SAUDEGALICIA2017. Calculemos os estimadores dos coeficientes da recta de regresión da variable resposta *peso* con respecto á covariable *altura*. O estimador do intercepto, $\hat{\beta}_0$, é

```
> mean(peso) - cov(altura,peso)*mean(altura)/var(altura)

[1] -52.43113
```

e o estimador da pendente, $\hat{\beta}_1$, é

```
> cov(altura,peso)/var(altura)

[1] 0.7552768
```

Resulta moito máis cómodo empregar a función `lm()`:

```
> rr.pesoaltura <- lm(peso~altura)
```

A pendente é 0.755. Isto significa que, en promedio, o peso aumenta en 0.755 kg por cada cm de incremento da altura.

A recta de regresión pódese incluír no gráfico de dispersión facendo

```
> abline(rr.pesoaltura)
```

En realidade o obxecto creado pola función `lm()` contén moita máis información que empregaremos para analizar a calidade do modelo de regresión e para facer inferencia:

```
> summary(rr.pesoaltura)

Call:
lm(formula = peso ~ altura)

Residuals:
    Min      1Q  Median      3Q     Max
-29.561  -9.251   0.055   7.810  47.321

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -52.43113   11.62094  -4.512 8.47e-06 ***
altura        0.75528    0.07009  10.776  < 2e-16 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.5 on 398 degrees of freedom
Multiple R-squared:  0.2259,        Adjusted R-squared:  0.2239
F-statistic: 116.1 on 1 and 398 DF,  p-value: < 2.2e-16
```

□

5.2.2. Variabilidade explicada. Coeficiente R^2

Para analizar a calidade do axuste do modelo de regresión aos datos empregamos a seguinte descomposición da variabilidade total da variable resposta

$$\underbrace{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}_{\text{variabilidade total}} = \underbrace{\sum_{i=1}^{n}\hat{\varepsilon}_i^2}_{\text{variabilidade residual}} + \underbrace{\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2}_{\text{variabilidade explicada}},$$

onde

$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ son os **valores axustados** e

$\hat{\varepsilon}_i = Y_i - \hat{Y}_i$ son os **residuos**.

O **coeficiente** R^2 é

$$\begin{aligned} R^2 &= \frac{\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2}{\sum_{i=1}^{n}(Y_i-\bar{Y})^2} = \frac{\text{variabilidade explicada}}{\text{variabilidade total}} \\ &= 1 - \frac{\sum_{i=1}^{n}\hat{\varepsilon}_i^2}{\sum_{i=1}^{n}(Y_i-\bar{Y})^2} = 1 - \frac{\text{variabilidade residual}}{\text{variabilidade total}} \end{aligned}$$

Esta fórmula do coeficiente R^2 é xenérica e pódese empregar en moitos modelos de regresión. No caso da recta de regresión é doado demostrar que o coeficiente R^2 coincide co cadrado do coeficiente de correlación,

$$R^2 = \left(\frac{S_{XY}}{S_X S_Y}\right)^2 = r_{XY}^2.$$

O coeficiente R^2 ten as seguintes propiedades:

- $0 \leq R^2 \leq 1$.
- R^2 é a proporción de variabilidade explicada polo modelo de regresión. Noutras palabras, representa a proporción de información da variable resposta explicada directamente pola covariable a través do modelo de regresión.
- Se R^2 é próximo a 1, entón o modelo axústase moi ben aos datos e explica moi ben a relación entre X e Y.

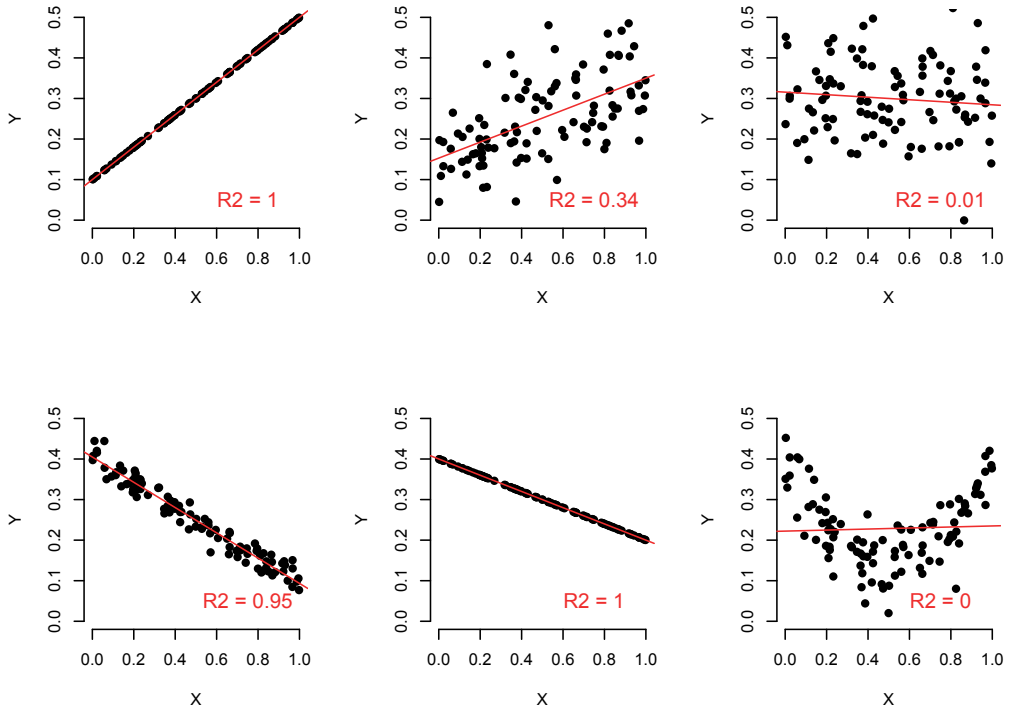


Figura 5.4: Exemplos de gráficos de dispersión coas rectas de regresión estimadas e os correspondentes coeficientes R^2.

- Se R^2 é próximo a 0, entón o axuste do modelo non é bo. Isto pódese deber a dúas razóns (ver Figura 5.4):

  - A variabilidade da resposta é moi alta.
  - O modelo de regresión escollido non é apropiado para explicar os datos.

Exemplo. (cont.) Datos SAUDEGALICIA2017. O coeficiente R^2 da recta de regresión entre as variables *peso* e *altura* é `cor(altura,peso)^ 2 =` 0.2259. Esta información tamén aparece como "`Multiple R-squared`" cando se aplica a función `lm()`. Neste exemplo, o valor do coeficiente R^2 é relativamente baixo. Aínda así, parece que a recta de regresión é un modelo axeitado para describir a relación entre estas dúas variables. □

5.2.3. Análise de residuos

A análise dos residuos do modelo axustado tamén resulta de axuda para comprobar a validez do modelo de regresión. Lembremos que os residuos son

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i), \qquad i = 1, \dots, n.$$

Se o modelo de regresión é correcto, entón os residuos e os valores axustados non deberían amosar ningunha correlación, é dicir, non se debería observar ningunha tendencia particular no diagrama de dispersión dos residuos fronte aos valores axustados.

Por exemplo, na Figura 5.5 podemos observar que no primeiro caso a recta de regresión se axusta ben aos datos. Por outra banda, os residuos dos exemplos segundo e terceiro presentan unha clara tendencia con respecto aos valores axustados, polo que nestes casos a recta de regresión non é un modelo axeitado. Nótese, así e todo, que o coeficiente R^2 do terceiro exemplo é moi alto.

Ademais, para realizar inferencias (tests de hipóteses e intervalos de confianza) sobre os coeficientes de regresión tamén debemos comprobar que os erros seguen unha distribución Normal. Isto pode facerse mediante o histograma ou o qq-plot dos residuos.

Exemplo. (cont.) Datos SAUDEGALICIA2017. Os gráficos correspondentes á análise dos residuos (Figura 5.6) pode obterse en R mediante

```
> plot(rr.pesoaltura)
```

□

Exercicio 5.1. *No conxunto de datos* SAUDEGALICIA2017 *a análise dos residuos da recta de regresión do* peso *sobre a* altura *indica que hai 3 datos que poderían considerarse como atípicos. Aparecen identificados no gráfico de dispersión dos residuos fronte aos valores axustados e no qq-plot dos residuos. Identifica estes individuos na mostra e axusta a recta de regresión despois de eliminar as observacións correspondentes. Obtense un axuste mellor?*

5.2.4. Tests de hipóteses en regresión

Para facer inferencia sobre os coeficientes da recta de regresión necesitamos engadir algúns supostos adicionais ao modelo. En particular, supoñemos que

os erros de regresión ε_i **son** $N(0, \sigma)$.

Ademais da Normalidade, esta suposición implica que a variabilidade dos erros é constante ao longo dos distintos valores da covariable. Esta condición denomínase **homocedasticidade**.

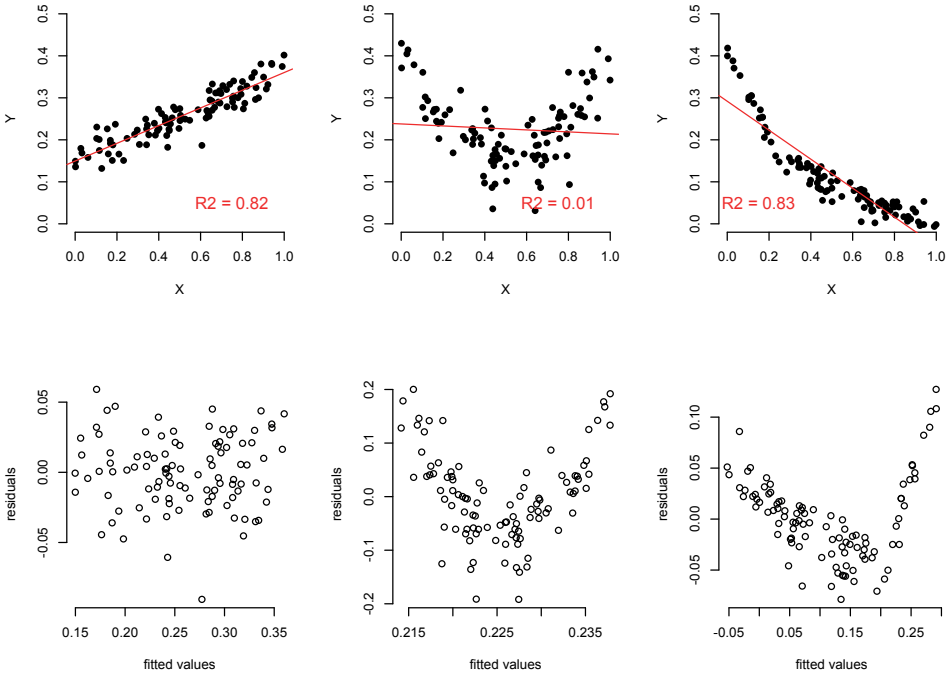


Figura 5.5: Primeira fila: gráficos de dispersión. Segunda fila: gráficos de dispersión dos residuos fronte aos valores axustados.

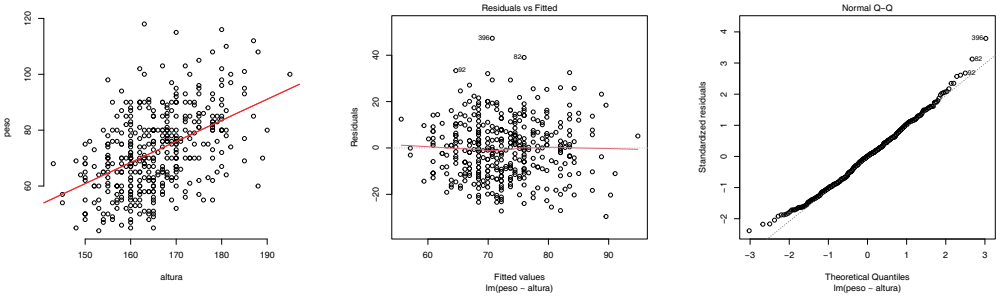


Figura 5.6: Conxunto de datos SAUDEGALICIA2017. Esquerda: gráfico de dispersión de *peso* fronte a *altura* e recta de regresión axustada. Centro: residuos fronte a valores axustados. Dereita: qq-plot de Normalidade dos residuos.

Test de utilidade global do modelo

En primeiro lugar podemos pensar nun test xeral que nos permita saber se o modelo de regresión contén algunha información sobre a relación entre a reposta e a covariable. A hipótese nula é

H_0 : o modelo de regresión non contén ningunha información sobre a relación entre Y e X,

e a hipótese alternativa é

$$H_1 : H_0 \text{ é falsa.}$$

Este test denomínase tamén test de utilidade global do modelo. O estatístico de test é

$$F = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\frac{1}{n-2}\sum_{i=1}^n \hat{\varepsilon}_i^2},$$

que baixo a hipótese nula se distribúe como unha F de Snedecor con 1 e $n-2$ graos de liberdade, $F_{1,n-2}$. A información sobre este test tamén se pode organizar nunha táboa de tipo ANOVA:

fonte de variabilidade	graos de liberdade	suma de cadrados (SS)	media de cadrados (MS)	estatístico F
explicada	1	$\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$	$\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$	$F \sim F_{1,n-2}$
residual	$n-2$	$\sum_{i=1}^n \hat{\varepsilon}_i^2$	$\frac{1}{n-2}\sum_{i=1}^n \hat{\varepsilon}_i^2$	
total	$n-1$	$\sum_{i=1}^n (Y_i - \bar{Y})^2$		

En R, a información sobre este test está contida no obxecto creado pola función `lm()`. Tamén pode obterse coa función `anova()`.

Exemplo. (cont.) Datos SAUDEGALICIA2017. A función `lm()` indica que

```
F-statistic: 116.1 on 1 and 398 DF,  p-value: < 2.2e-16
```

así que o modelo é claramente significativo. Tamén podemos facer

```
> anova(rr.pesoaltura)

Analysis of Variance Table

Response: peso
           Df Sum Sq Mean Sq F value    Pr(>F)
altura      1  18148 18148.4  116.12 < 2.2e-16 ***
Residuals 398  62205   156.3
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

□

Tests sobre os coeficientes

Baixo as condicións habituais do modelo de regresión lineal (homocedasticidade e erros con distribución Normal), pódese demostrar que

$$\frac{\hat{\beta}_1 - \beta_1}{\widehat{\mathrm{SE}}(\hat{\beta}_1)} \sim t_{n-2},$$

onde

$$\widehat{\mathrm{SE}}(\hat{\beta}_1) = \sqrt{\frac{\frac{1}{n-2}\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

é o estimador do erro estándar do estimador $\hat{\beta}_1$. Este resultado pódese empregar para obter intervalos de confianza e facer tests sobre β_1.

> **Exercicio 5.2.** *Considera o conxunto de datos* SAUDEGALICIA2017. *Constrúe un intervalo de confianza de nivel 0.95 para a pendente da recta de regresión da variable* peso *sobre a variable* altura.

Nun modelo de regresión, é interesante comprobar se algún dos seus coeficientes é distinto de 0. No caso da recta de regresión, o habitual é facer unicamente o test sobre a pendente:

$$H_0 : \beta_1 = 0 \qquad \text{fronte a} \qquad H_1 : \beta_1 \neq 0.$$

A hipótese nula afirma que a pendente da recta de regresión é 0, é dicir, a covariable non ten efecto sobre a resposta[2]. O estatístico de test é

$$t = \frac{\hat{\beta}_1}{\widehat{\mathrm{SE}}(\hat{\beta}_1)}.$$

A distribución do estatístico de test baixo a hipótese nula é t_{n-2}. A función `lm()` proporciona o estimador do erro estándar. Os valores críticos e os p-valores correspondentes poden obterse de xeito inmediato.

Exemplo. (cont.) Datos SAUDEGALICIA2017. O obxecto creado pola función `lm()` contén os tests sobre os coeficientes da recta de regresión:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -52.43113   11.62094  -4.512 8.47e-06 ***
altura        0.75528    0.07009  10.776  < 2e-16 ***
```

O estimador da pendente da recta de regresión é 0.75528, cun erro estándar estimado de 0.07009. Polo tanto, o estatístico de test é de $0.75528/0.07009 = 10.776$. O p-valor é practicamente 0 (de feito, $< 2 \cdot 10^{-16}$), polo que a pendente é significativamente distinta de cero. A altura ten un efecto significativo e positivo sobre o peso. □

[2]No caso da recta de regresión, o test $H_0 : \beta_1 = 0$ baseado no estatístico t é equivalente ao test de utilidade global do modelo baseado no estatístico F. É doado demostrar que $t^2 = F$.

5.2.5. Predición: intervalos de confianza para a media da resposta e intervalos de predición

Unha vez que temos o modelo de regresión estimado, podemos facer predicións. Para un valor particular da covariable, x, o **valor predito** da resposta é $\hat{Y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$. Esta predición pode acompañarse de intervalos de confianza de nivel $1-\alpha$:

- **Intervalo de confianza para o valor medio da resposta** cando o valor da covariable é x:

$$\left(\hat{Y}(x) \mp t_{n-2,1-\alpha/2}\, \hat{\sigma}_R \sqrt{\frac{1}{n} + \frac{(x-\bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}\right),$$

  onde $t_{n-2,1-\alpha/2}$ é o cuantil de orde $(1-\alpha/2)$ da distribución t_{n-2} e

$$\hat{\sigma}_R = \sqrt{\frac{1}{n-2}\sum_{i=1}^n \hat{\varepsilon}_i{}^2}$$

  é o estimador da desviación estándar dos erros de regresión.

- **Intervalo de predición para o valor da resposta** cando o valor da covariable é x:

$$\left(\hat{Y}(x) \mp t_{n-2,1-\alpha/2}\, \hat{\sigma}_R \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}\right).$$

Exemplo. Datos SAUDEGALICIA2017. Consideremos o valor $x = 170$ cm da variable *altura*.

- O valor predito do *peso* é $-52.43113 + 0.75528 \cdot 170 = 75.97$ kg.
- O intervalo de confianza de nivel 0.95 para o valor medio do *peso* é $(74.59, 77.34)$.
- O intervalo de predición de nivel 0.95 para o valor do *peso* é $(51.35, 100.58)$.

A función `predict()` calcula os valores preditos, os intervalos de confianza e os intervalos de predición. A información sobre os valores da covarible nos que se desexa facer a predición e calcular os intervalos de confianza deben ser introducidos como un `data.frame`.

```
> puntos.predicion <- data.frame(altura=c(170))
> predict(rr.pesoaltura,newdata=puntos.predicion)

       1
75.96593

> predict(rr.pesoaltura,interval="confidence",newdata=puntos.predicion)

       fit      lwr      upr
1 75.96593 74.59318 77.33868
```

```
> predict(rr.pesoaltura,interval="prediction",newdata=puntos.predicion)
```

```
       fit      lwr      upr
1 75.96593 51.34998 100.5819
```

O gráfico da Figura 5.7 amosa os intervalos de confianza para os valores medios e os intervalos de predición da variable *peso* segundo os valores da variable *altura*. Nótese que os intervalos de predición son moito máis amplos ca os intervalos de confianza para a media da resposta. Obsérvese tamén que se obteñen intervalos máis estreitos para valores próximos á media da covariable. □

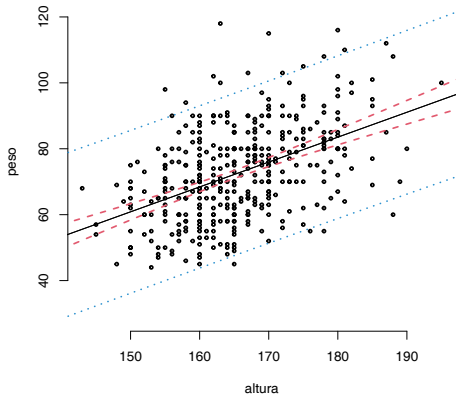


Figura 5.7: Conxunto de datos SAUDEGALICIA2017. Gráfico de dispersión da variable *peso* fronte á variable *altura*, recta de regresión, intervalos de confianza de nivel 0.95 para o valor medio (liñas discontinuas) e intervalos de predición de nivel 0.95 (liñas punteadas).

Exercicio 5.3. *Considera o conxunto de datos* SAUDEGALICIA2017.

(a) Fai un gráfico de dispersión da variable peso *con respecto á variable* altura *distinguindo os homes e as mulleres con dúas cores.*

(b) Axusta as rectas de regresión por separado para cada sexo e incorpóraas ao gráfico.

(c) As rectas resultantes son case paralelas. *Que conclusión podes sacar deste feito? Como interpretarías os coeficientes?*

Exercicio 5.4. *A partir do conxunto de datos* SAUDEGALICIA2017*, crea unha nova variable que recolla o índice de masa corporal de cada individuo.*

(a) Calcula a recta de regresión do índice de masa corporal en función da altura. Ten algunha utilidade este modelo? É a pendente da recta distinta de cero? Como interpretas este feito?

(b) Calcula a recta de regresión do índice de masa corporal en función da idade. Analiza o axuste do modelo. Como interpretas os coeficientes?

5.3. Clasificación dos modelos de regresión

En xeral, dada unha variable resposta Y e unha covariable (posiblemente multidimensional) X, un modelo de regresión é unha relación da forma

$$Y = m(X) + \varepsilon,$$

onde m é a **función de regresión** e ε é o erro de regresión. Asúmese que o valor medio do erro ε é cero, polo que o valor da función $m(x)$ representa a media condicional da resposta para un valor particular da covariable.

Os modelos de regresión pódense clasificar en varios tipos:

- Segundo a **dimensión** da covariable:
  - Modelo de regresión **simple**: a covariable é unidimensional (como, por exemplo, a recta de regresión).
  - Modelo de regresión **múltiple**: a covariable é multidimensional (como, por exemplo, os modelos que se estudarán na sección 5.4).
- Segundo a **forma** da función m:
  - Modelo de regresión **lineal**: a función m é lineal, é dicir $m(x_1, \dots, x_d) = \beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d$. A recta de regresión é un exemplo de modelo lineal.
  - Modelo de regresión **non lineal**: a función m non é lineal, é dicir, a función de regresión involucra polinomios, funcións exponenciais etc. (véxase a seccion 5.6.3).

5.4. Regresión lineal múltiple

5.4.1. O modelo de regresión lineal múltiple. Estimación

A regresión lineal múltiple é unha extensión natural da regresión lineal simple. Supoñamos que dispoñemos de observacións dun conxunto de d covariables $(X_1, X_2, \dots, X_d)$ e da resposta

Y. Os datos son da forma

$$(X_{i1}, X_{i2}, \dots, X_{id}, Y_i), \qquad i = 1, \dots, n.$$

O **modelo de regresión lineal múltiple** é

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_d X_{id} + \varepsilon_i, \qquad i = 1, \cdots, n,$$

ou, en forma matricial

$$\mathbf{Y} = \mathbf{XB} + \boldsymbol{\varepsilon},$$

onde

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1d} \\ 1 & X_{21} & \cdots & X_{2d} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & \cdots & X_{nd} \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_d \end{pmatrix} \qquad \text{e} \qquad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

A matriz $\mathbf{X}$ chámase **matriz de deseño**. O vector $\mathbf{B} = (\beta_0, \beta_1, \dots, \beta_d)^t$ é o **vector de coeficientes**.

O método de mínimos cadrados consiste en atopar o vector de coeficientes $\mathbf{B}$ que é solución do problema de minimización

$$\arg\min_{\mathbf{B}} \sum_{i=1}^{n} \varepsilon_i^2 = \arg\min_{\mathbf{B}} \boldsymbol{\varepsilon}^t \boldsymbol{\varepsilon}.$$

Tendo en conta que

$$\boldsymbol{\varepsilon}^t \boldsymbol{\varepsilon} = (\mathbf{Y} - \mathbf{XB})^t (\mathbf{Y} - \mathbf{XB}) = \mathbf{Y}^t\mathbf{Y} - \mathbf{Y}^t\mathbf{XB} - \mathbf{B}^t\mathbf{X}^t\mathbf{Y} + \mathbf{B}^t\mathbf{X}^t\mathbf{XB}$$

e que a derivada de $\boldsymbol{\varepsilon}^t\boldsymbol{\varepsilon}$ respecto do vector $\mathbf{B}$ é

$$\frac{\partial \boldsymbol{\varepsilon}^t \boldsymbol{\varepsilon}}{\partial \mathbf{B}} = -\mathbf{Y}^t\mathbf{X} - (\mathbf{X}^t\mathbf{Y})^t + 2\mathbf{B}^t\mathbf{X}^t\mathbf{X} = -2\mathbf{Y}^t\mathbf{X} + 2\mathbf{B}^t\mathbf{X}^t\mathbf{X},$$

entón simplemente temos que atopar a solución do sistema de ecuacións (nótese que $\mathbf{X}^t\mathbf{X}$ é unha matriz simétrica)

$$-2\mathbf{Y}^t\mathbf{X} + 2\mathbf{B}^t\mathbf{X}^t\mathbf{X} = 0 \iff \mathbf{B}^t\mathbf{X}^t\mathbf{X} = \mathbf{Y}^t\mathbf{X} \iff \mathbf{X}^t\mathbf{XB} = \mathbf{X}^t\mathbf{Y}.$$

A solución é

$$\hat{\mathbf{B}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}.$$

As mesmas ideas que explicamos no apartado de regresión simple pódense aplicar á regresión múltiple:

- O vector $\hat{\mathbf{B}} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_d)^t$ contén os estimadores dos $d+1$ coeficientes de regresión.
- $\hat{\beta}_0$ é o estimador do intercepto (pouco interesante na maior de casos prácticos).

- O estimador $\hat{\beta}_j$ que acompaña á covariable X_j interprétase como o cambio esperado na resposta asociado a un incremento de 1 unidade en X_j *supoñendo que as outras covariables permanecen fixas.*
- Na análise de residuos, teremos que comprobar que os residuos non presentan ningunha correlación con valores axustados e que teñen distribución Normal.
- Para realizar un test sobre a utilidade global do modelo de regresión usamos o F-test.
- Para realizar tests particulares sobre os coeficientes de regresión usamos os t-tests.

5.4.2. Tests en modelos de regresión múltiples

Para facer inferencia sobre todo o modelo, os coeficientes ou as predicións, precisamos asumir as condicións habituais en regresión lineal: homocedasticidade e erros con distribución Normal. Estas condicións deben comprobarse na análise dos residuos.

Test de utilidade global do modelo

O test de utilidade global do modelo completo ten hipótese nula

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_d = 0$$

e a hipótese alternativa é

$$H_1 : \text{ hai polo menos un } \beta_j \neq 0, \quad j = 1, \ldots, d.$$

Nótese que o intercepto β_0 non se inclúe na formulación deste test. O estatístico de test é

$$F = \frac{\frac{1}{d}\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\frac{1}{n-d-1}\sum_{i=1}^{n}\hat{\varepsilon}_i^2},$$

que, baixo a hipótese nula, ten unha distribución $F_{d,n-d-1}$. En R, a información deste test está contida no obxecto xerado pola función `lm()`.

Tests sobre os coeficientes do modelo de regresión

Nun modelo de regresión múltiple podemos facer tests sobre a significación de cada coeficiente β_j:

$$H_0 : \beta_j = 0 \qquad \text{fronte a} \qquad H_1 : \beta_j \neq 0.$$

A hipótese nula establece que a covariable X_j non ten efecto sobre a resposta. O estatístico de test é

$$t = \frac{\hat{\beta}_j}{\widehat{\text{SE}}(\hat{\beta}_j)},$$

onde $\widehat{\text{SE}}(\hat{\beta}_j)$ é un estimador do erro estándar de $\hat{\beta}_j$ (non damos aquí a fórmula detallada). En R, a función `lm()` calcula este erro estándar e o estatístico de test e devolve o correspondente p-valor.

Exemplo. O conxunto de datos CALIDADEAIRENY recolle medidas diarias da calidade do aire de New York (este conxunto de datos forma parte doutro máis amplo contido en R). Contén as seguintes variables:

- *Ozone*: media da cantidade de ozono (en partes por 10^9).
- *Solar.R*: radiación solar (en Langleys).
- *Wind*: promedio da velocidade do vento (en millas por hora).
- *Temp*: temperatura máxima diaria (en graos Fahrenheit).

Consideraremos a variable *Ozone* como resposta e as variables *Solar.R*, *Wind* e *Temp* como covariables.

```
> calidadeaireNY <- read.table(file="datos-calidadeaireNY.txt",header=TRUE)
> attach(calidadeaireNY)
```

Primeiro podemos comprobar o grao de dependencia entre as variables facendo un gráfico de dispersión múltiple empregando a función `pairs()` (ver Figura 5.8)

```
> pairs(calidadeaireNY)
```

e calculando os coeficientes de correlación mostrais

```
> cor(calidadeaireNY)
```

```
         Ozone Solar.R   Wind   Temp
Ozone    1.000   0.348 -0.612  0.699
Solar.R  0.348   1.000 -0.127  0.294
Wind    -0.612  -0.127  1.000 -0.497
Temp     0.699   0.294 -0.497  1.000
```

A maior correlación aparece entre *Ozone* e *Temp*. A análise do modelo de regresión lineal simple entre estas dúas variables proporciona un coeficiente R^2 de 0.488. Intentemos mellorar o modelo incorporando as outras dúas covariables. Para facer isto en R simplemente empregamos a función `lm()` do seguinte xeito:

```
> modelo.Ozone <- lm(Ozone~Temp+Wind+Solar.R)
> summary(modelo.Ozone)

Call:
lm(formula = Ozone ~ Temp + Wind + Solar.R)

Residuals:
```

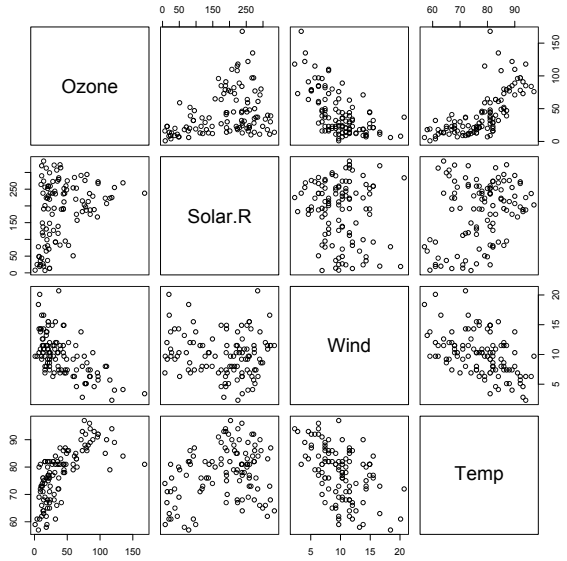


Figura 5.8: Conxunto de datos CALIDADEAIRENY. Gráficos de dispersión das variables do estudo.

```
    Min      1Q  Median      3Q     Max
-40.485 -14.219  -3.551  10.097  95.619

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -64.34208   23.05472  -2.791  0.00623 **
Temp          1.65209    0.25353   6.516 2.42e-09 ***
Wind         -3.33359    0.65441  -5.094 1.52e-06 ***
Solar.R       0.05982    0.02319   2.580  0.01124 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.18 on 107 degrees of freedom
Multiple R-squared:  0.6059,        Adjusted R-squared:  0.5948
F-statistic: 54.83 on 3 and 107 DF,  p-value: < 2.2e-16
```

Segundo os resultados obtidos, podemos comprobar que:

- O F-test conclúe que o modelo global é altamente significativo.
- O coeficiente correspondente a *Temp* é 1.65. Segundo o t-test, este coeficiente é distinto de cero (p-valor < 0.001). A variable *Temp* ten un efecto positivo sobre a variable *Ozone*.
- O coeficiente correspondente a *Wind* é -3.33 e o t-test tamén conclúe con claridade que é

distinto de cero (p-valor < 0.001). A variable *Wind* ten un efecto negativo sobre a variable *Ozone*.

- O coeficiente correspondente a *Solar.R* é 0.06. O t-test conclúe que este coeficiente tamén é distinto de cero cun p-valor de 0.011. A variable *Solar.R* ten un efecto positivo sobre a variable *Ozone*.

- A proporción de variabilidade da resposta *Ozone* explicada polo modelo é 0.6059, como podemos comprobar co coeficiente R^2. Non obstante, neste caso é máis adecuado mirar o coeficiente R^2 axustado, que é 0.5948 (ver seguinte sección). □

5.4.3. Análise de residuos. Coeficiente R^2 axustado

Para cuantificar a bondade do axuste dun modelo de regresión múltiple, en vez de empregar o coeficiente R^2 (habitualmente chamado *coeficiente R^2 múltiple*), é máis recomendable empregar o *coeficiente R^2 axustado*. Lembremos que o **coeficiente R^2 múltiple** é

$$R^2 = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = 1 - \frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2},$$

onde

$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_d X_{id}$, $i = 1, \ldots, n$, son os **valores axustados**, e

$\hat{\varepsilon}_i = Y_i - \hat{Y}_i$, $i = 1, \ldots, n$, son os **residuos** do modelo.

O **coeficiente R^2 axustado** defínese como

$$1 - R^2 \text{ axustado} = \frac{(n-1)}{n-d-1}(1 - R^2) \iff R^2 \text{ axustado} = \frac{(n-1)R^2 - d}{n-d-1}.$$

O coeficiente R^2 axustado ten en conta a complexidade do modelo, de tal forma que penaliza os valores grandes de d. Resulta polo tanto axeitado para comparar modelos de distintas complexidades.

Igual que no caso da recta de regresión, na análise dos residuos deberiamos comprobar que a súa distribución non difire da Normal, que os residuos e os valores axustados non están correlados e que ademais os residuos teñen varianza constante.

Exemplo. (cont.) Conxunto de datos CALIDADEAIRENY. A Figura 5.9 amosa a análise de residuos. Teñen os residuos algún tipo de dependencia con respecto aos valores axustados? Pode aceptarse que os residuos teñen distribución Normal? (**exercicio**). □

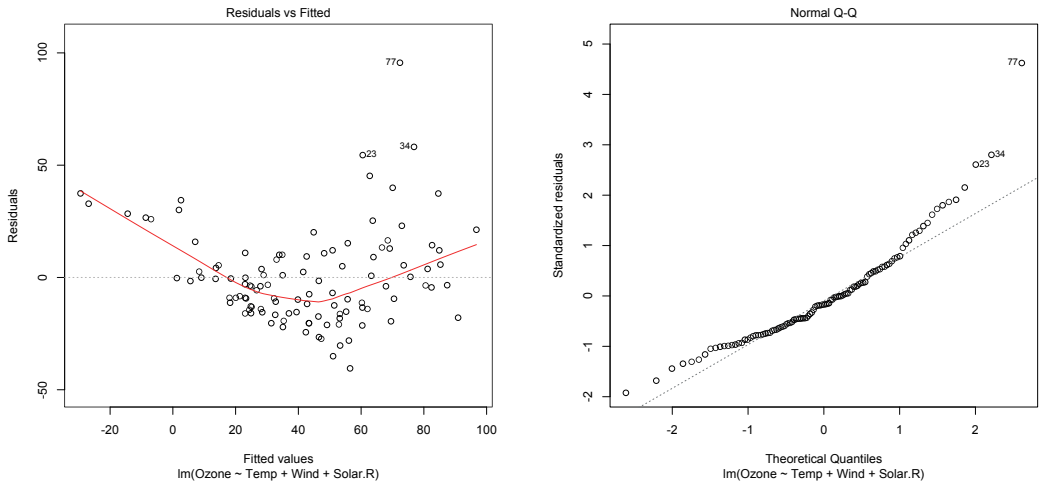


Figura 5.9: Conxunto de datos CALIDADEAIRENY. Análise de residuos.

5.4.4. Comparación de modelos xerárquicos

Cando temos varias covariables dispoñibles, un problema importante na regresión é como elixir o conxunto de covariables que mellor explican o comportamento da resposta. Para conseguilo hai varios métodos, algúns deles moi complicados.

Unha primeira aproximación a este problema consiste na comparación de modelos xerárquicos ou aniñados da forma

$$\begin{aligned} \text{modelo restrinxido:} \quad & Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_d X_{id} + \varepsilon_{r,i}. \\ \text{modelo completo:} \quad & Y_i = \beta'_0 + \beta'_1 X_{i1} + \cdots + \beta'_d X_{id} + \beta'_{d+1} X_{i\,d+1} + \cdots + \beta'_{d+k} X_{i\,d+k} + \varepsilon_{c,i}. \end{aligned}$$

Nótese que os modelos teñen que respectar unha estrutura xerárquica, ou, dito doutra forma, teñen que estar aniñados. Todas as covariables incluídas no modelo restrinxido deben formar parte tamén do modelo completo. Neste caso, é interesante comprobar se as novas covariables incluídas no modelo completo aportan algunha información significativa respecto ao modelo restrinxido. A hipótese nula é

$$H_0 : \beta'_{d+1} = \beta'_{d+2} = \cdots = \beta'_{d+k} = 0,$$

e a alternativa é

$$H_1 : \text{ hai polo menos un coeficiente } \beta'_j \neq 0, \text{ para } d+1 \leq j \leq d+k.$$

O estatístico de test é

$$F = \frac{(\mathrm{SSR}_r - \mathrm{SSR}_c)/k}{\mathrm{SSR}_c/(n-d-k-1)} \underset{\text{baixo } H_0}{\sim} F_{k,n-d-k-1},$$

onde

SSR_r é a suma de residuos ao cadrado do modelo restrinxido:

$$\text{SSR}_r = \sum_{i=1}^{n} \hat{\varepsilon}_{r,i}^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \cdots - \hat{\beta}_d X_{id})^2.$$

SSR_c é a suma de residuos ao cadrado do modelo completo:

$$\text{SSR}_c = \sum_{i=1}^{n} \hat{\varepsilon}_{c,i}^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}'_0 - \hat{\beta}'_1 X_{i1} - \cdots - \hat{\beta}'_{d+k} X_{i\,d+k})^2.$$

En R, a función `anova()` realiza a comparación de modelos xerárquicos.

Exemplo. Conxunto de datos CALIDADEAIRENY. As variables *Ozone* e *Temp* presentan a correlación máis alta. A recta de regresión da variable *Ozone* en termos de *Temp* explica o 48 % da variabilidade da resposta. Será relevante a inclusión das outras dúas variables, *Wind* e *Solar.R*, no modelo?

Consideramos os modelos

$$\begin{aligned} \text{modelo restrinxido:} &\quad Ozone_i = \beta_0 + \beta_1 Temp_i + \varepsilon_{r,i} \\ \text{modelo completo:} &\quad Ozone_i = \beta'_0 + \beta'_1 Temp_i + \beta'_2 Wind_i + \beta'_3 Solar.R_i + \varepsilon_{c,i} \end{aligned}$$

Primeiro creamos os correspondentes modelos en R

```
> modelo.restrinxido <- lm(Ozone~Temp)
> modelo.completo <- lm(Ozone~Temp+Wind+Solar.R)
```

Neste caso $n = 111$, $d = 1$, $k = 3 - 1 = 2$. A sumas de residuos ao cadrado son

```
> ( ssr.r <- sum(resid(modelo.restrinxido)^2) )

[1] 62367.44

> ( ssr.c <- sum(resid(modelo.completo)^2) )

[1] 48002.79
```

O estatístico do F-test e o p-valor son

```
> ( Ftest=((ssr.r-ssr.c)/2)/(ssr.c/(111-1-2-1)) )

[1] 16.00967

> 1-pf(Ftest,2,111-1-2-1)
```

```
[1] 8.270069e-07
```

Á vista do p-valor, parece que as variables *Wind* e *Solar.R* aportan información relevante ao modelo. Toda esta información pódese obter empregando a función `anova()`:

```
> anova(modelo.restrinxido,modelo.completo)

Analysis of Variance Table

Model 1: Ozone ~ Temp
Model 2: Ozone ~ Temp + Wind + Solar.R
  Res.Df   RSS Df Sum of Sq     F   Pr(>F)
1    109 62367
2    107 48003  2     14365 16.01 8.27e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

□

Exercicio 5.5. *Consideremos o conxunto de datos* CALIDADEAIRENY. *Emprega un F-test para decidir se a inclusión da variable* Solar.R *é relevante cando se compara co modelo* $\text{Ozone}_i = \beta_0 + \beta_1 \text{Temp}_i + \beta_2 \text{Wind}_i + \varepsilon_i$. *Nótese que o p-valor do test F neste caso particular coincide co correspondente t-test para o parámetro correspondente a* Solar.R *no modelo completo. De feito, as dúas probas son equivalentes cando o modelo completo só contén unha variable adicional respecto do modelo restrinxido.*

5.5. Problemas en regresión lineal e posibles solucións

No diagnóstico dun modelo de regresión poden xurdir algúns problemas:

- Unha función de tipo lineal non resulta axeitada para describir a relación entre a(s) covariable(s) e a resposta.

  Solución 1: Empregar **regresión non lineal** (por exemplo, regresión polinómica) ou **regresión non paramétrica**.

  Solución 2: Buscar unha **transformación** dos datos de forma que a relación sexa lineal. Por exemplo, transformacións logarítmicas poden ser de utilidade en moitos casos.

- A varianza do erro de regresión non é constante, senón que depende dos valores da covariable, é dicir, hai **heteroscedasticidade**.

  Solución: Empregar **mínimos cadrados ponderados** para obter estimadores dos coeficientes de regresión e facer as inferencias correspondentes.

- O modelo seleccionado axústase ben aos datos, agás por algúns poucos **datos atípicos**. Os datos atípicos poden poden influír moito na estimación dos parámetros.

  Solución: Analizar os datos atípicos en detalle e aplicar **técnicas robustas**.

- A distribución do erro de regresión **non é Normal**.

  Solución: Buscar unha transformación da resposta para que se poida aceptar a normalidade dos erros de regresión. Por exemplo, as transformacións **Box-Cox** son útiles para levar a cabo esta tarefa.

- Os erros presentan estrutura de **dependencia temporal**.

  Solución: Incorporar o tempo como unha covariable relevante no modelo e aplicar técnicas apropiadas para **datos dependentes** (series de tempo etc.).

- As covariables teñen unha forte **colinialidade**, isto é, algunhas covariables están fortemente correladas.

  Solución: Aplicar técnicas de selección de variables e manter no modelo as covariables máis relevantes e con pouca correlación entre elas.

Exercicio 5.6. ***Importante!*** *Considera o conxunto de datos* CALIDADEAIRENY*. A análise dos residuos do modelo lineal* $\text{Ozone}_i = \beta_0 + \beta_1 \text{Temp}_i + \beta_2 \text{Wind}_i + \beta_3 \text{Solar.R} + \varepsilon_i$ *non proporcionou un gran axuste (ver Figura 5.9: o diagrama de dispersión dos residuos fronte aos valores axustados amosa unha marcada tendencia e a normalidade dos residuos tampouco é clara). Considera o modelo lineal múltiple coas mesmas covariables pero substituíndo a variable* Ozone *por* log(Ozone) *como variable resposta. É o axuste mellor agora? Interpreta os resultados e os coeficientes.*

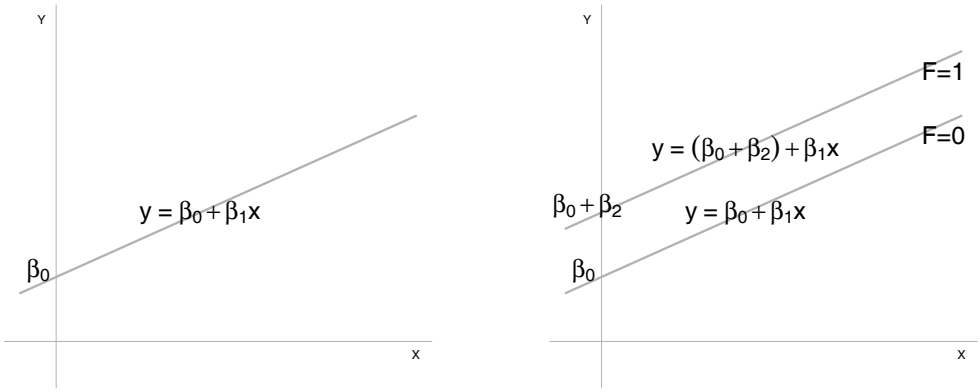
5.6. Modelos de regresión avanzados

5.6.1. Covariables nominais: codificación con variables *dummy*

Ata agora só consideramos o caso en que tanto a variable resposta como a(s) covariable(s) son numéricas. Nos modelos de regresión tamén podemos incluír variables nominais ou factores. A incorporación destas variables realízase mediante **variables dummy**, que son variables indicadoras con valores 0/1 creadas de forma artificial.

Consideremos que desexamos incluír como covariable no modelo un factor F con dous niveis (L_0 e L_1). Para iso creamos unha variable dummy da seguinte maneira:

$$D_i = \begin{cases} 1 & \text{se } F_i = L_1 \\ 0 & \text{se } F_i = L_0 \end{cases}, \qquad \text{para } i = 1, \ldots, n.$$



(a) (b)

Figura 5.10: (a) Modelo cunha covariable numérica. (b) Modelo cunha covariable numérica e un factor con dous niveis (0 e 1).

Neste caso, o nivel L_0 funciona como o "nivel de referencia". Esta asignación é arbitraria cando non se pode establecer ningunha orde entre os niveis.

A variable dummy pódese incorporar ao modelo de regresión lineal. Por exemplo, consideremos un modelo lineal cunha covariable numérica e un factor con dous niveis. Entón escribimos o modelo da forma

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \varepsilon_i, \qquad i = 1, \dots, n.$$

Neste caso, o coeficiente β_1 representa o incremento da resposta cando a covariable X se incrementa nunha unidade. A interpretación do novo coeficiente β_2 tamén é moi sinxela: de feito, representa o cambio na resposta cando pasamos do nivel L_0 (nivel de referencia) a L_1. Nótese que neste modelo pasar do nivel L_0 ao nivel L_1 só ten un efecto no intercepto do modelo lineal correspondente. O modelo contén implicitamente dúas rectas de regresión con interceptos diferentes pero pendentes iguais (ver Figura 5.10).

Cando o factor F ten $k+1$ niveis (digamos $L_0, L_1, \dots, L_k$), entón necesitamos k variables dummy. Para $\ell = 1, \dots, k$, definimos

$$D_{i\ell} = \begin{cases} 1 & \text{se } F_i = L_\ell \\ 0 & \text{noutro caso} \end{cases}, \qquad \text{para } i = 1, \dots, n,$$

e incluímolas no modelo.

Exercicio 5.7. *Describe un modelo de regresión lineal que conteña como covariables unha variable numérica e un factor de tres niveis (L_0, L_1 e L_2). Cantas variables dummy son necesarias? Cantos parámetros ten o modelo? Como se interpretan estes parámetros?*

Exemplo. Datos DIABETES. Diversas razóns médicas explican o feito de que o nivel de glucosa, incluso no caso de persoas que non padecen diabetes, aumenta coa idade. A recta de regresión da variable *glu* con respecto á variable *age* é $glu = 92.15 + 0.99age$, con $R^2 = 0.1179$. Como podemos ver, o axuste é deficiente.

Se no gráfico de dispersión distinguimos os individuos en termos do factor *type*, parece razoable incluír este factor no modelo.

```
> diabetes <- read.table(file="datos-diabetes.txt",header=TRUE)
> attach(diabetes)
> plot(age,glu,type="n")
> points(age[type=="Yes"],glu[type=="Yes"],col=2)
> points(age[type=="No"],glu[type=="No"],col=4)
```

En R cando incorporamos un factor ao modelo a función `lm()` crea as variables dummy necesarias:

```
> modelo.factor <- lm(glu ~ age + type)
```

Podemos comprobar que a matriz de deseño do modelo contén a variable dummy:

```
> model.matrix(modelo.factor)
```

Agora analizamos os resultados da estimación. Nótese que o efecto do factor *type* é significativo.

```
> summary(modelo.factor)

Call:
lm(formula = glu ~ age + type)

Residuals:
    Min      1Q  Median      3Q     Max
-62.693 -16.755  -2.135  15.552  82.825

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  96.7372     6.0355  16.028  < 2e-16 ***
age           0.5599     0.1897   2.951  0.00355 **
typeYes      27.2180     4.3848   6.207 3.13e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27.34 on 197 degrees of freedom
Multiple R-squared:  0.2622,        Adjusted R-squared:  0.2547
F-statistic: 35.01 on 2 and 197 DF,  p-value: 9.778e-14
```

□

Exercicio 5.8. *No exemplo anterior, especifica os dous modelos resultantes para describir a relación entre* glu *e* age *dependendo dos niveis do factor* type *e interpreta os coeficientes.*

Exercicio 5.9. *Considera o conxunto de datos* DIABETES. *Constrúe un modelo lineal para describir a variable* bp *en termos da variable* age *e do factor* type *. É relevante o factor no modelo? Que tipo de modelo se debería empregar no seu lugar? O modelo de regresión correspondente axústase ben aos datos? Por que?*

Exercicio 5.10. *Considera o conxunto de datos* SAUDEGALICIA2017. *No exercicio 4 vimos que as rectas de regresión do* peso *sobre a* altura *para mulleres e homes parecían distintas. Constrúe un modelo de regresión para explicar o* peso *en función da* altura *incorporando o* sexo *como factor.*

Nota importante: Nos conxuntos de datos é habitual que algúns factores estean codificados con números. Cando se desexa incluír un factor no modelo de regresión hai que asegurarse de que R o entende verdadeiramente como factor e non como variable numérica. Por exemplo, no conxunto de datos SAUDEGALICIA2017 hai varias variables codificadas como 0/1 (*nivel.estudos*, *hipertension*, *gafas* etc.). Se queremos facer o modelo para o *peso* en función da *altura* incluíndo o factor *gafas* temos que escribir

```
> lm(peso ~ altura + factor(gafas))
```

Exercicio 5.11. *Considera o conxunto de datos* SAUDEGALICIA2017. *A variable* actividade.fisica *é un factor codificado en catro niveis (1, 2, 3 e 4). Dado que no nivel 4 só hai unha observación, convén non empregalo no modelo de regresión.*

(a) Crea un novo factor con 3 niveis xuntando os niveis 3 e 4 orixinais.

(b) Constrúe un modelo de regresión para explicar o peso *en función da* altura *incluíndo este novo factor con tres niveis. Analiza a calidade do axuste e interpreta os coeficientes.*

5.6.2. Modelos con interaccións

Interacción entre unha variable numérica e un factor

Nalgúns casos, a representación do modelo como suma da covariable e o factor pode ser restritiva. Por exemplo, no modelo

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \varepsilon_i, \qquad i = 1, \dots, n,$$

onde D_i é a variable dummy asociada a un factor con dous niveis, vimos que o modelo correspondente só permite dúas rectas de regresión paralelas (ver Figura 5.10). En moitos casos, forzar que a pendente sexa a mesma para cada nivel do factor non resulta realista.

Un modelo da forma

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 \underbrace{D_i X_i}_{\text{interacción}} + \varepsilon_i, \qquad i = 1, \dots, n,$$

permite interceptos diferentes e pendentes diferentes para cada nivel do factor, tal e como se ilustra na Figura 5.11. O termo $D_i X_i$ chámase **interacción**.

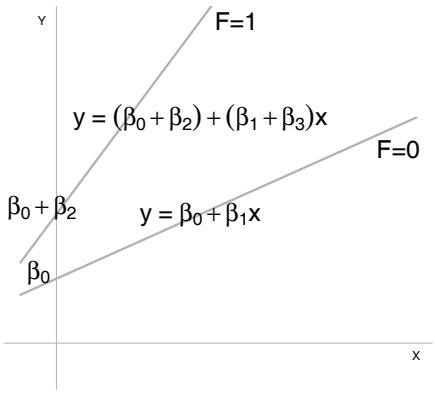


Figura 5.11: Modelo con interacción entre un factor con dous niveis (0 e 1) e unha variable numérica.

Exemplo. Datos DIABETES. Para explicar a variable *bp* en termos da variable *age*, o modelo cun factor non é recomendable, xa que as rectas de regresión correspondentes aos niveis do factor *type* claramente non son paralelas. Para incluír a interacción na función `lm()` simplemente hai que escribir `covariable*factor` ou `covariable:factor`[3]:

[3] A diferenza entre estas dúas formulacións da interacción é a seguinte: `covariable:factor` inclúe unicamente a interacción, `covariable*factor` inclúe a interacción e todos os outros efectos relacionados coas covariables.

```
> modelo.interaccion <- lm(bp~age+type+age*type)
> summary(modelo.interaccion)

Call:
lm(formula = bp ~ age + type + age * type)

Residuals:
    Min      1Q  Median      3Q     Max
-33.652  -7.268   0.171   6.004  36.466

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 54.52808    2.95331  18.463  < 2e-16 ***
age          0.51368    0.09607   5.347 2.47e-07 ***
typeYes     12.53792    5.29687   2.367   0.0189 *
age:typeYes -0.31411    0.14731  -2.132   0.0342 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.49 on 196 degrees of freedom
Multiple R-squared:  0.177,        Adjusted R-squared:  0.1644
F-statistic: 14.05 on 3 and 196 DF,  p-value: 2.474e-08
```

Os resultados amosan que a interacción entre *age* e *type* é significativa. A interpretación dos coeficientes é a seguinte:

- Para o nivel `"No"` de *type*, o modelo de regresión é $bp = 54.53 + 0.51\,age$.
- Para o nivel `"Yes"` de *type*, o modelo de regresión é $bp = (54.53+12.54)+(0.51-0.31)\,age$.

□

Exercicio 5.12.

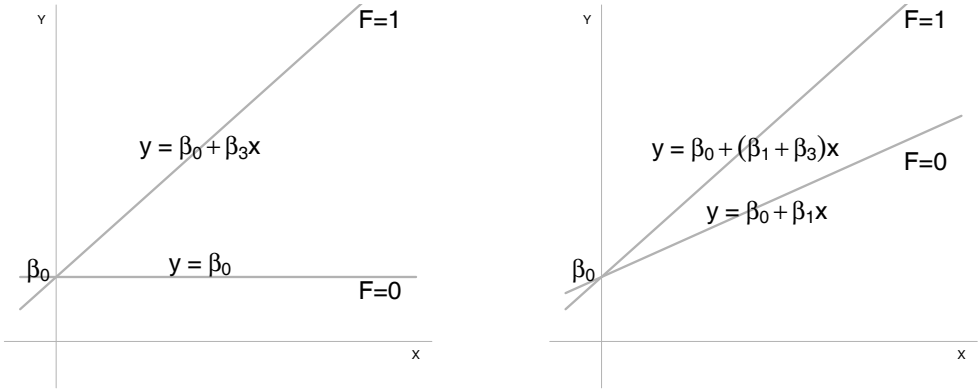
(a) Describe os modelos de regresión correspondentes aos gráficos da Figura 5.12.

(b) Describe un modelo de regresión cunha covariable numérica e un factor con tres niveis (recorda que se necesitan dúas variables dummy). Inclúe a interacción entre eles. Cantos parámetros aparecen no modelo? Interprétaos.

Interacción entre dúas covariables numéricas

Tamén son posibles as interaccións entre covariables numéricas. Por exemplo, consideremos o seguinte modelo con dúas covariables numéricas e unha interacción entre elas:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 \underbrace{X_{i1}X_{i2}}_{\text{interacción}} + \varepsilon_i, \qquad i = 1, \dots, n.$$



(a) (b)

Figura 5.12: Modelos de regresión cunha covariable numérica e un factor con dous niveis. Ver exercicio 12.

Neste caso, a interpretación dos parámetros é máis complicada. Por exemplo, cando a covariable X_1 aumenta en 1 unidade e X_2 permanece fixa, o cambio correspondente na resposta depende do valor de X_2. Sexa x_2 o valor fixo de X_2, entón a repercusión do aumento de 1 unidade en X_1 na resposta é

$$(\beta_0 + \beta_1(x_1 + 1) + \beta_2 x_2 + \beta_3(x_1 + 1)x_2) - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2) = \beta_1 + \beta_3 x_2.$$

De xeito análogo, podemos describir o comportamento da resposta en termos de X_2.

Exemplo. Datos DIABETES. Consideremos a variable *bmi* como resposta e as covariables *age* e *skin*. Probemos un modelo lineal con interacción:

```
> summary(lm(bmi~age+skin+age*skin))

Call:
lm(formula = bmi ~ age + skin + age * skin)

Residuals:
   Min     1Q Median     3Q    Max
-9.940 -2.792  0.197  2.546 11.363

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.446359   2.160150   5.762 3.18e-08 ***
age          0.275393   0.060664   4.540 9.82e-06 ***
```

```
skin         0.706223   0.069839  10.112  < 2e-16 ***
age:skin    -0.009905   0.001788  -5.539 9.71e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.315 on 196 degrees of freedom
Multiple R-squared:  0.512,        Adjusted R-squared:  0.5045
F-statistic: 68.54 on 3 and 196 DF,  p-value: < 2.2e-16
```

O modelo contén catro parámetros, todos eles significativos. Como se interpreta o coeficiente correspondente á interacción (-0.0099)? (**exercicio**) □

> **Exercicio 5.13.** *No exemplo anterior, mellora o axuste a inclusión do factor* type*? O factor pódese incorporar ao modelo de varias maneiras. Considéraas todas. Compara os modelos obtidos e comenta os resultados.*

5.6.3. Modelos non lineais: modelos polinómicos

Nalgúns casos, as relacións lineais non resultan axeitadas para algunhas aplicacións prácticas. No seu lugar pode resultar máis conveniente empregar relacións non lineais.

A familia máis importante de funcións non lineais son os polinomios de grao ≥ 2. Dada unha mostra da forma (X_i, Y_i), $i = 1, \dots, n$, o **modelo polinómico** de grao k é

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \cdots + \beta_k X_i^k + \varepsilon_i.$$

Por suposto, cando $k = 1$ o modelo coincide coa recta de regresión. Para $k \geq 2$, o modelo polinómico non é máis ca un modelo de regresión lineal múltiple con k covariables da forma $X_{ij} = X_i^j$, $j = 1, \dots, k$. A matriz de deseño do modelo polinómico é

$$\begin{pmatrix} 1 & X_1 & X_1^2 & \cdots & X_1^k \\ 1 & X_2 & X_2^2 & \cdots & X_2^k \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_n & X_n^2 & \cdots & X_n^k \end{pmatrix}.$$

Non é recomendable empregar polinomios de graos altos, xa que se k é grande algunhas columnas da matriz de deseño son case colineais e o proceso de estimación pode ser inestable.

Exemplo. O conxunto de datos PEIXES contén información sobre a *Lonxitude* (en cm) e o *Peso* (en g) de 56 percas (peixe de auga doce). O diagrama de dispersión destas dúas variables amosa unha forte relación entre ambas variables, pero non parece que esta relación sexa a través dunha liña recta (ver Figura 5.13). É preferible considerar un modelo polinómico.

Para incorporar os termos do modelo polinómico na función `lm()` empregamos `+I(x^j)`. No noso exemplo, o modelo polinómico de grao 2 resulta

```
> peixes <- read.table(file="datos-peixes.txt",header=TRUE)
> summary(lm(Peso~Lonxitude+I(Lonxitude^2),data=peixes))

Call:
lm(formula = Peso ~ Lonxitude + I(Lonxitude^2), data = peixes)

Residuals:
     Min       1Q   Median       3Q      Max
-174.072  -22.751    0.532   12.276  242.806

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    135.9136    79.4866    1.71  0.09313 .
Lonxitude      -23.6726     6.3135   -3.75  0.00044 ***
I(Lonxitude^2)   1.1651     0.1162   10.02 7.75e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 59.42 on 53 degrees of freedom
Multiple R-squared:  0.9718,        Adjusted R-squared:  0.9708
F-statistic: 914.5 on 2 and 53 DF,  p-value: < 2.2e-16
```

Os resultados amosan que o termo cadrático é significativo e o axuste do modelo e moi bo (ver Figura 5.13).

Tamén podemos comparar modelos polinómicos con distintos graos:

```
> modelo1 <- lm(Peso~Lonxitude,data=peixes)
> modelo2 <- lm(Peso~Lonxitude+I(Lonxitude^2),data=peixes)
> modelo3 <- lm(Peso~Lonxitude+I(Lonxitude^2)+I(Lonxitude^3),data=peixes)
> anova(modelo1,modelo2,modelo3)

Analysis of Variance Table

Model 1: Peso ~ Lonxitude
Model 2: Peso ~ Lonxitude + I(Lonxitude^2)
Model 3: Peso ~ Lonxitude + I(Lonxitude^2) + I(Lonxitude^3)
  Res.Df    RSS Df Sum of Sq       F    Pr(>F)
1     54 541948
2     53 187158  1    354789 99.8814 1.061e-13 ***
3     52 184710  1      2449  0.6894    0.4102
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Os F-tests indican que pasar da recta de regresión a un modelo cadrático representa unha mellora significativa no modelo. Non obstante, o modelo cúbico non aporta ningunha mellora respecto ao cuadrático. Así, deberiamos quedar co modelo cuadrático. □

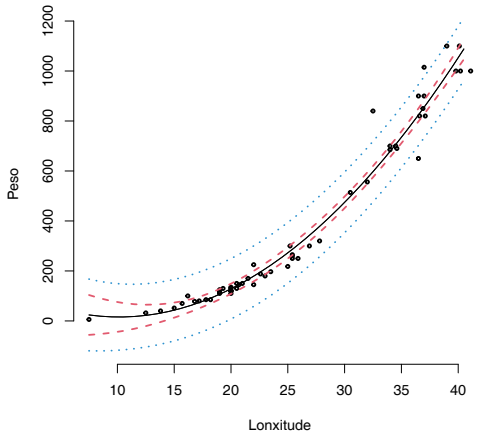


Figura 5.13: Datos PEIXES. Gráfico de dispersión de *Peso* fronte a *Lonxitude*, modelo polinómico de grao 2, intervalos de confianza para o valor medio da resposta (liña discontinua) e intervalos de predición (liña punteada).

Exercicio 5.14. ***Modelos linealizables.*** *Existen algúns modelos non lineais que poden ser analizados como modelos lineais unha vez que se lles aplican certas transformacións ás variables involucradas no modelo. Algúns exemplos coas súas correspondentes transformacións son:*

Modelo	*Función de regresión*	*Transformación*	*Forma linealizada*
Exponencial	$y = \beta_0\ e^{\beta_1 x}$	$T = \log Y$	$t = \log \beta_0 + \beta_1 x$
Logarítmico	$y = \beta_0 + \beta_1 \log x$	$Z = \log X$	$t = \beta_0 + \beta_1 z$
Alométrico ou Potencial	$y = \beta_0\ x^{\beta_1}$	$T = \log Y,\ Z = \log X$	$t = \log \beta_0 + \beta_1 z$

(a) Analiza a forma da función de regresión en cada modelo en función dos valores dos parámetros β_0 e β_1.

(b) Considera o conxunto de datos PEIXES. *Aplica o modelo alométrico para explicar a variable* Peso *en termos da variable* Lonxitude. *Analiza os resultados obtidos e compáraos co modelo polinómico visto no exemplo anterior: cantos parámetros ten cada modelo?, que modelo resulta máis sinxelo de interpretar?*

Exercicio 5.15. ***Exemplo de construción de modelos.*** *O conxunto de datos* TREES *está incluído en R. Contén información sobre a altura (variable* Height*), o diámetro (variable* Girth*) e sobre a cantidade de madeira aproveitada (variable* Volume*) de 31 árbores. Escribe* `?trees` *ou* `str(trees)` *para obter máis detalles.*

(a) Fai unha análise descritiva do conxunto de datos e intenta atopar un modelo de regresión para explicar o volume de madeira en termos do diámetro e da altura.

(b) Tendo en conta a fórmula do volume dun cilindro, parece razoable pensar nun modelo do tipo volume $= \beta_1$ diámetro^2 altura*. Atopa a forma correcta de escribir este modelo en R e comenta os resultados.*

5.7. Regresión con resposta cualitativa: regresión loxística

Os modelos de regresión que estudamos ata o momento son apropiados para relacionar unha resposta numérica cun conxunto de covariables numéricas e factores, pero non son adecuados cando a variable resposta é cualitativa.

Supoñamos que queremos estudar a relación entre un conxunto de covariables $(X_1, X_2, \dots, X_d)$ e unha variable resposta binaria Y que toma valores 0 e 1. Esta variable binaria pódese usar para distinguir entre dous grupos ou dúas poboacións. Neste contexto o modelo lineal

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_d X_{id} + \varepsilon_i, \qquad i = 1, \dots, n,$$

ten varios inconvenientes:

- A resposta é binaria (0/1), mentres que as covariables poden ser continuas, polo que é complicado facer que a relación matemática se cumpra.
- Para que a expresión anterior realmente se cumpra, os erros deberían depender das covariables e a súa distribución non podería ser normal. Isto complicaría moito o tratamento estatístico do modelo en termos de estimación e inferencia.

Unha posible solución consiste en pensar en probabilidades relacionadas coa variable resposta en vez de no seu valor. Para simplificar, consideraremos un modelo cunha única covariable, X. O obxectivo agora é modelizar a probabilidade condicionada

$$p(x) = P(Y = 1|X = x),$$

é dicir, a probabilidade de que a resposta Y tome o valor 1 cando a covariable X toma o valor x. Non obstante, tampouco podemos considerar un modelo lineal para $p(\cdot)$ porque a probabilidade ten que estar contida $[0, 1]$, mentres que unha expresión lineal da covariable non o estará necesariamente. A solución é transformar a expresión lineal anterior a través dunha **función link**, ℓ, da forma

$$p(x) = \ell(\beta_0 + \beta_1 x),$$

onde ℓ é unha función crecente que toma valores en $(0,1)$. A **regresión loxística** obtense cando a función link é a función loxística

$$\ell(x) = \frac{1}{1+e^{-x}}.$$

O **modelo loxístico de regresión** pódese escribir de varias formas equivalentes:

$$p(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}} \iff \frac{p(x)}{1-p(x)} = e^{\beta_0+\beta_1 x} \iff \log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0+\beta_1 x.$$

O cociente $o(x) = \frac{p(x)}{1-p(x)}$ chámase **odds**. Nótese que

$$\log o(x) = \log\left(\frac{p(x)}{1-p(x)}\right) = \log p(x) - \log(1-p(x)) = \beta_0+\beta_1 x,$$

é dicir, o modelo loxístico expresa a diferenza das probabilidades asociadas ás dúas categorías da variable resposta en escala logarítmica en termos da información proporcionada pola covariable. O parámetro β_1 ten unha interpretación en termos do aumento de 1 unidade na covariable, xa que o cociente entre as odds, que se denomina **odds ratio**, é

$$\frac{o(x+1)}{o(x)} = \frac{\frac{p(x+1)}{1-p(x+1)}}{\frac{p(x)}{1-p(x)}} = \frac{e^{\beta_0+\beta_1(x+1)}}{e^{\beta_0+\beta_1 x}} = \frac{\cancel{e^{\beta_0}}\,\cancel{e^{\beta_1 x}}\,e^{\beta_1}}{\cancel{e^{\beta_0}}\,\cancel{e^{\beta_1 x}}} = e^{\beta_1}.$$

A estimación dos parámetros da regresión loxística realízase polo método de **máxima verosimilude**, que é similar ao método dos mínimos cadrados con ponderacións. A calidade do axuste do modelo loxístico mídese en termos da **deviance**, que tamén se pode usar para comparar modelos.

O modelo loxístico está incluído nunha clase máis ampla de modelos de regresión chamados **modelos lineais xeralizados**. En R, a función `glm()` proporciona os estimadores e a análise do modelo loxístico.

Exemplo. Datos DIABETES. Un problema interesante é predicir o efecto dalgunhas das variables numéricas (*glu*, *bmi* etc.) sobre a resposta binaria *type* (diabética/non diabética). Primeiro probamos o modelo incluíndo como covariable unicamente *glu*:

```
> modelo.loxistico <- glm(factor(type)~glu,family=binomial,data=diabetes)
> summary(modelo.loxistico)


Call:
glm(formula = factor(type) ~ glu, family = binomial, data = diabetes)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9714  -0.7795  -0.5292   0.8491   2.2633

Coefficients:
```

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.503636   0.836077  -6.583 4.62e-11 ***
glu          0.037784   0.006278   6.019 1.76e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 256.41  on 199  degrees of freedom
Residual deviance: 207.37  on 198  degrees of freedom
AIC: 211.37

Number of Fisher Scoring iterations: 4
```

Segundo os resultados, a variable *glu* ten un efecto significativo. Como se interpreta o coeficiente (0.038)? (**exercicio**).

A partir do modelo loxístico tamén podemos facer predicións. Neste caso, o modelo predice a probabilidade de padecer diabetes en función do valor da covariable *glu*. Por exemplo, segundo o modelo loxístico, se unha persoa presenta unha concentración de glucosa de 180, a probabilidade de que padeza diabetes será

$$\hat{p}(180) = \frac{1}{1 + e^{-(-5.503636 + 0.037784 \cdot 180)}} = 0.7854.$$

En R facemos:

```
> valor.glu <- data.frame(glu=c(180))
> predict(modelo.loxistico,valor.glu,type="response")

        1
0.7854027
```

A Figura 5.14 amosa a estimación da función $p(x)$ como función da covariable *glu*. Como se pode observar no gráfico, a función outórgalle probabilidades baixas de padecer diabetes a valores baixos da concentración de glucosa e probabilidades altas de padecer a enfermidade a valores altos da concentración de glucosa. □

Exercicio 5.16. *Considera o conxunto de datos* DIABETES.

(a) Estima un modelo loxístico para predicir a probabilidade de padecer diabetes en termos das covariables glu, bmi *e* bp. *Son as tres variables relevantes no modelo?*

(b) Se unha persoa presenta os valores glu = 170, bmi = 30 *e* bp = 100, *cal é a probabilidade de que padeza diabetes segundo o modelo loxístico?*

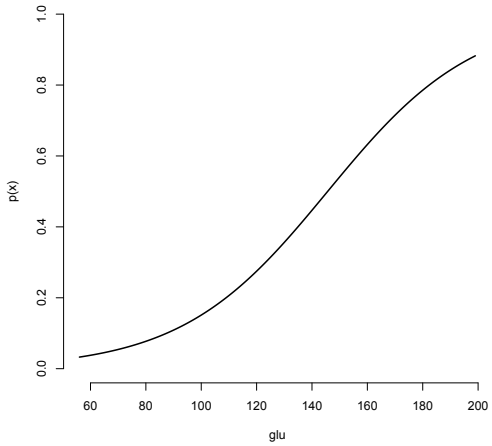


Figura 5.14: Datos DIABETES. Estimación da función $p(x)$ no modelo loxístico.

Clasificación loxística

Unha vez estimado o modelo loxístico, este pode empregarse para **clasificar** novos individuos segundo os valores observados das covariables. Unha regra de clasificación moi sinxela podería ser

- Se $\hat{p}(x) \leq 0.5$, entón clasificar ao individuo no grupo $Y = 0$.
- Se $\hat{p}(x) > 0.5$, entón clasificar ao individuo no grupo $Y = 1$.

Obviamente, este tipo de regras de clasificación deben ser analizados en termos de sensibilidade, especificidade, valores preditivos etc. O valor 0.5 escollido como punto de corte para a probabilidade pode ser modificado en función das necesidades específicas do problema.

Capítulo 6

Técnicas bioestatísticas multivariantes

Contidos

6.1. Introdución á análise multivariante

En moitas ocasións recóllense varias variables sobre un mesmo individuo, tal e como ocorre nos conxuntos de datos cos que traballamos. Isto lévanos aos **datos multivariantes**.

Para estudar datos multivariantes podemos seguir dous enfoques:

- Se unha das variables se identifica como "resposta" e o obxectivo é analizar a **dependencia** entre esta variable e as demais, ou dito doutra forma, se queremos analizar o efecto que as outras variables teñen sobre a resposta, entón empregamos **técnicas de regresión**.

- Se o obxectivo é analizar a **interdependencia** entre as variables, sen darlle máis importancia a ningunha delas sobre as demais, entón empregamos **técnicas multivariantes**.

Estudaremos tres tipos de técnicas para datos multivariantes:

- Técnicas descritivas.

- Técnicas para a redución da dimensión: análise de compoñentes principais.

- Técnicas para crear grupos: métodos cluster e análise discriminante.

6.2. Técnicas descritivas multivariantes

Comparación de ferramentas descritivas para datos univariantes/multivariantes:

- Datos univariantes (variable unidimensional):
  - Medidas de localización: media, mediana, cuantís etc.
  - Medidas de dispersión: varianza, desviación estándar etc.
  - Gráficos: gráfico de sectores, gráfico de barras, histograma etc.

- Datos multivariantes (variable multidimensional):
  - Medidas de localización: vector de medias.
  - Medidas de dispersión e medidas de dependencia: matriz de varianzas-covarianzas, matriz de correlacións, distancia de Mahalanobis.
  - Gráficos: boxplots, gráficos de dispersión por pares.

6.2.1. Representacións gráficas con datos multivariantes

Algunhas representacións gráficas que coñecemos poden adaptarse ao caso multivariante. Dependendo do tipo de datos dos que dispoñamos, teremos que buscar a representación gráfica máis axeitada.

- O boxplot pódese usar para representar unha variable numérica en función das categorías dunha variable nominal ou factor. En R emprégase a función `boxplot(variable.numérica ~ factor)`.
- O diagrama de dispersión por pares consiste en facer gráficos de dispersión para cada par de variables que conforman o conxunto de datos. É útil para buscar relacións entre as variables, aínda que non deixa de ser unha simplificación da posible estrutura de interdependencia multidimensional. En R emprégase a función `pairs(data.frame)`. Nótese que este tipo de gráfico só ten sentido para variables numéricas.

Exemplo. O conxunto de datos MUNDO2016 contén datos demográficos dos países do mundo obtidos da base de datos do ano 2016 do Instituto de Estatística da UNESCO[1]. Contén as seguintes variables:

- País, código do país (tres letras) e zona xeográfica segundo a clasificación da UNESCO (East Asia & Pacific, Europe & Central Asia, Latin America & Caribbean, Middle East & North Africa, North America, South Asia, Sub-Saharan Africa).
- *UE*: membro (si/non) da Unión Europea.
- *OCDE*: membro (si/non) da Organización para a Cooperación e o Desenvolvemento Económico (OCDE).
- *nivel.ingresos*: nivel de ingresos do país de acordo coa clasificación do Atlas do Banco Mundial[2]. Países de ingresos baixos (código B) son aqueles que teñen un produto interior bruto (PIB) per capita de 1005 US$ ou menos; países de ingresos medio-baixos (código MB) son aqueles cun PIB per capita entre 1006 US$ e 3955 US$; países con ingresos medio-altos (código MA) son aqueles cun PIB per capita entre 3956 US$ e 12235 US$; finalmente, países con ingresos altos (código A) son aqueles cun PIB per capita de máis de 12235 US$.
- *esperanza.vida*: esperanza de vida ao nacer (anos).
- *taxa.fertilidade*: taxa de fertilidade (número fillos por muller).
- *taxa.mortalidadeinfantil*: taxa de mortalidade infantil (por cada 1000 nacementos).
- *poboacion.rural*: porcentaxe de poboación rural (%).

[1] http://data.uis.unesco.org
[2] http://www.worldbank.org

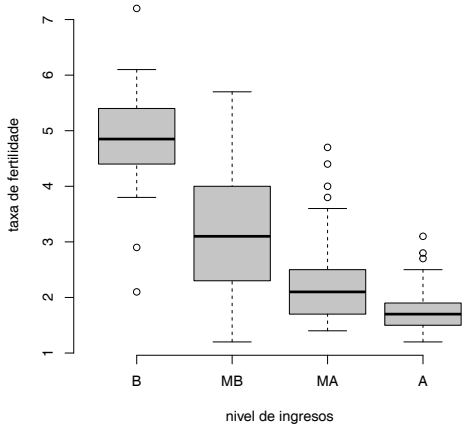


Figura 6.1: Conxunto de datos MUNDO2016. Boxplots da taxa de fertilidade segundo o nivel de ingresos do país.

A Figura 6.1 contén os boxplots da taxa de fertilidade en función do nivel de ingresos (atención á orde dos niveis da variable ordinal *nivel.ingresos*). A Figura 6.2 amosa o gráfico de dispersión por pares das catro variables cuantitativas do conxunto de datos. A curva vermella que aparece en cada gráfico é unha estimación non paramétrica da función de regresión entre as correspondentes variables. □

Exercicio 6.1.

(a) Que conclusións podemos sacar dos boxplots da Figura 6.1.

(b) Fai boxplots do resto de variables numéricas do conxunto de datos MUNDO2016 *en función do nivel de ingresos de cada país. Intenta atopar patróns a partir dos gráficos.*

(c) Que tipo de relacións se observan entre as variables numéricas do conxunto de datos MUNDO2016 *a partir dos gráficos de dipersión da Figura 6.2?*

6.2.2. Vector de medias, matriz de varianzas-covarianzas e matriz de correlacións

Cando as variables son numéricas podemos calcular medidas de localización e dispersión multidimensionais. Supoñamos que se dispoñemos de n observacións dunha variable d-dimensional

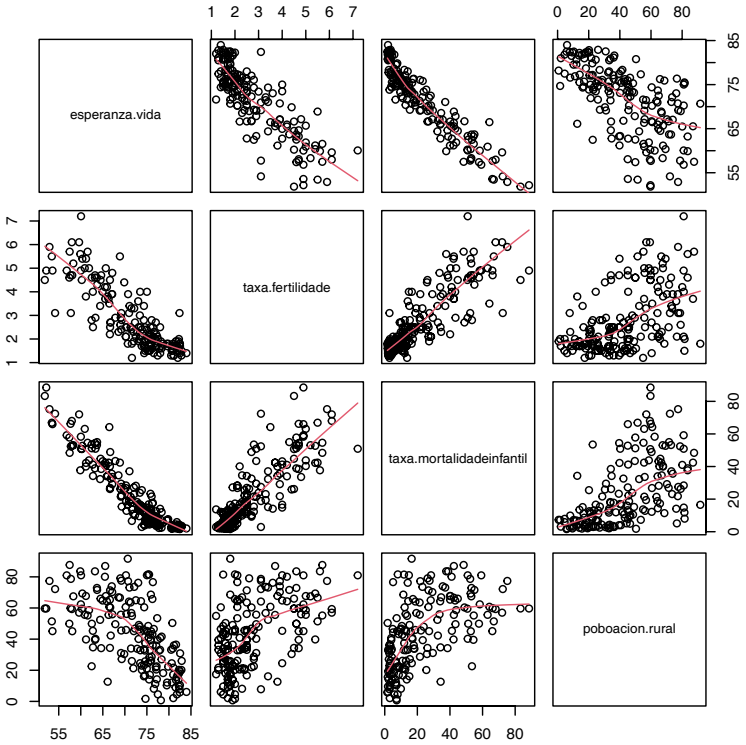


Figura 6.2: Conxunto de datos MUNDO2016. Gráficos de dispersión por pares das variables numéricas.

$(X_1, X_2, \ldots, X_d)$ da forma

$$\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{id}), \qquad i = 1, \ldots, n.$$

Na práctica, os datos organízanse nunha matriz $n \times d$ ou "data frame": cada fila representa un individuo e cada columna representa unha variable. O conxunto de datos está disposto da forma

$$\begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1d} \\ X_{21} & X_{22} & \cdots & X_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{nd} \end{pmatrix}.$$

O **vector de medias** é o vector d-dimensional formado polas medias mostrais de cada variable

$$\bar{\mathbf{X}} = (\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_d),$$

onde

$$\bar{X}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij} \qquad \text{para } j = 1, \dots, d.$$

Pódese empregar como "centro" dos datos.

A **matriz de varianzas-covarianzas** é a matriz $d \times d$ formada polas covarianzas mostrais entre cada par de variables, é dicir

$$\mathbf{S} = \begin{pmatrix} S^2_{X_1} & S_{X_1X_2} & \cdots & S_{X_1X_d} \\ S_{X_2X_1} & S^2_{X_1} & \cdots & S_{X_2X_d} \\ \vdots & \vdots & \ddots & \vdots \\ S_{X_dX_1} & S_{X_dX_2} & \cdots & S^2_{X_d} \end{pmatrix},$$

onde os elementos da diagonal de $\mathbf{S}$ son as varianzas mostrais

$$S^2_{X_j} = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2, \qquad \text{para } j = 1, \dots, d,$$

e os elementos fóra da diagonal son as covarianzas mostrais

$$S_{X_jX_k} = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)(X_{ik} - \bar{X}_k) \qquad \text{para } j, k = 1, \dots, d \text{ e } j \neq k.$$

Nótese que a matriz $\mathbf{S}$ é simétrica, xa que $S_{X_jX_k} = S_{X_kX_j}$.

A **matriz de correlacións** é a matriz $d \times d$ formada polos coeficientes de correlación mostrais entre as variables correspondentes

$$\mathbf{R} = \begin{pmatrix} 1 & r_{X_1X_2} & \cdots & r_{X_1X_d} \\ r_{X_2X_1} & 1 & \cdots & r_{X_2X_d} \\ \vdots & \vdots & \ddots & \vdots \\ r_{X_dX_1} & r_{X_dX_2} & \cdots & 1 \end{pmatrix},$$

onde

$$r_{X_jX_k} = \frac{S_{X_jX_k}}{S_{X_j}S_{X_k}} \qquad \text{para } j, k = 1, \dots, d \text{ e } j \neq k.$$

Lembremos que o coeficiente de correlación mostral toma valores entre -1 e 1 e mide o grao de dependencia lineal entre as variables: canto máis próximo sexa a -1 ou 1, máis forte será o grao de dependencia. A dependencia é directa cando o coeficiente de correlación é positivo e inversa cando o coeficiente de correlación é negativo. Cando o coeficiente de correlación é 0 dicimos que as variables son **incorreladas**.

Exemplo. Conxunto de datos MUNDO2016. En R a función `colMeans()` calcula o vector de medias e as funcións `var()` e `cor()` calculan as matrices de varianzas-covarianzas e de correlacións, respectivamente. Nótese que estes cálculos só teñen sentido cando as variables son numéricas (columnas 7 a 10 do conxunto de datos).

```
> colMeans(mundo2016[,7:10])
```

```
          esperanza.vida         taxa.fertilidade
                  71.579                    2.816
taxa.mortalidadeinfantil          poboacion.rural
                  23.354                   43.530
```

```
> cor(mundo2016[,7:10])
```

```
                        esperanza.vida taxa.fertilidade
esperanza.vida                   1.000           -0.833
taxa.fertilidade                -0.833            1.000
taxa.mortalidadeinfantil        -0.940            0.837
poboacion.rural                 -0.621            0.524
                        taxa.mortalidadeinfantil poboacion.rural
esperanza.vida                            -0.940          -0.621
taxa.fertilidade                           0.837           0.524
taxa.mortalidadeinfantil                   1.000           0.570
poboacion.rural                            0.570           1.000
```

A matriz $\mathbf{R}$ amosa correlacións moi fortes entre as variables esperanza de vida, taxa de fertilidade e taxa de mortalidade infantil. Cal é a explicación intuitiva dos signos das correlacións? (**exercicio**). □

6.2.3. Distancia de Mahalanobis

En moitas ocasións é interesante calcular a distancia entre unha observación e o centro dos datos, que normalmente se identifica co vector de medias. En vez de usar a distancia euclidiana clásica, na análise multivariante prefírese a distancia de Mahalanobis, xa que ten en conta a estrutura de dependencia dos datos e a dispersión de cada variable.

A **distancia Mahalanobis** entre unha observación $\mathbf{X}_i$ e o vector de medias $\bar{\mathbf{X}}$ é

$$D_i = \left[(\mathbf{X}_i - \bar{\mathbf{X}})^t \mathbf{S}^{-1} (\mathbf{X}_i - \bar{\mathbf{X}}) \right]^{1/2},$$

onde $\mathbf{S}^{-1}$ é a inversa da matriz de varianzas-covarianzas. A distancia de Mahalanobis pódese usar para identificar **datos atípicos** nun contexto multivariante.

En R a función `mahalanobis()` calcula o cadrado da distancia de Mahalanobis, D_i^2.

Exemplo. Conxunto de datos MUNDO2016. Para simplificar a explicación, traballaremos só con dúas variables: *taxa.fertilidade* e *esperanza.vida*.

A distancia de Mahalanobis de España con respecto ao vector de medias é

```
> datos.para.mahalanobis <- mundo2016[,c("taxa.fertilidade","esperanza.vida")]
> M <- colMeans(datos.para.mahalanobis)
> S <- cov(datos.para.mahalanobis)
> sqrt(mahalanobis(datos.para.mahalanobis[which(pais=="Spain"),],center=M,cov=S))
```

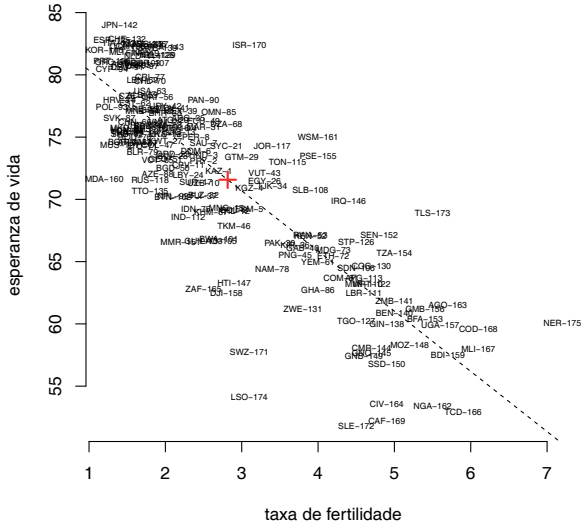


Figura 6.3: Datos MUNDO2016. Gráfico de dispersión das variables taxa de fertilidade e esperanza de vida. Indícase o código do país e a orde que ocupa en termos de distancia de Mahalanobis con respecto ao vector de medias, que está indicado pola cruz vermella. A liña discontinua é a recta de regresión.

```
     148
1.454964
```

Agora calculamos as distancias de Mahalanobis de todos os países facendo uso da función `apply()` de R:

```
> distancias.mahalanobis <-
+   sqrt(apply(datos.para.mahalanobis,1,mahalanobis,center=M,cov=S))
> cbind.data.frame(pais,distancias.mahalanobis)
```

```
                pais distancias.mahalanobis
1        Afghanistan              1.3439229
2            Albania              0.8934653
3            Algeria              1.0296971
4             Angola              2.3405168
5 Antigua and Barbuda              0.6169915
6          Argentina              0.6992731
```

A Figura 6.3 mostra o diagrama de dispersión e a posición de cada país con respecto ao vector de medias. Os números baixos significan distancias de Mahalanobis pequenas, mentras

que os números altos significan distancias de Mahalanobis grandes. Nótese que as posicións non coinciden coas correspondentes distancias euclidianas. As observacións máis próximas á liña de regresión adoitan ter distancias de Mahalanobis máis pequenas. □

Exercicio 6.2. *No exemplo anterior:*

(a) Atopa os 10 países coas maiores e coas menores distancias de Mahalanobis no exemplo anterior e localízaos no gráfico de dispersión da Figura 6.3.

(b) Constrúe boxplots das distancias de Mahalanobis en función do nivel de ingresos de cada país e identifica os valores atípicos en cada grupo.

Exercicio 6.3. *Considera o conxunto de datos* MUNDO2016. *Calcula as distancias de Mahalanobis de cada país con respecto ao conxunto de datos multivariante formado polas catro variables numéricas. Analiza os resultados de forma similar ao exercicio anterior.*

6.3. A distribución Normal multivariante

A distribución Normal multivariante xeraliza a distribución Normal a d dimensións. Dado o vector de medias $\boldsymbol{\mu}$ e a matriz de varianzas-covarianzas Σ, a función de densidade da Normal d-variante é

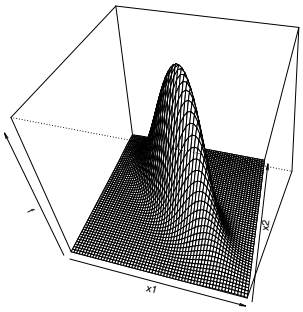
$$f(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}, \qquad \text{para } \mathbf{x} \in \mathbb{R}^d.$$

A distribución Normal multivariante cumpre, entre outras, as seguintes propiedades:

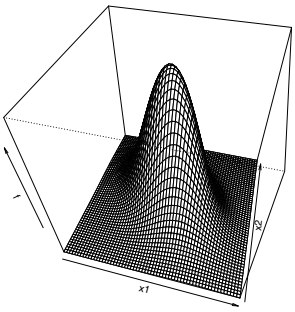
- As distribucións marxinais, é dicir, de cada unha das compoñentes que forman o vector aleatorio multivariante, son Normais.
- Calquera combinación lineal das compoñentes da Normal multivariante tamén é Normal.
- Se dúas compoñentes da Normal multivariante son correladas, entón a relación entre elas ten que ser lineal.

A Figura 6.4 amosa as densidades de distribucións Normais bivariantes con vector de medias $(0,0)$ e matriz de varianzas-covarianzas $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ para varios valores de ρ. Nótese que neste caso ρ representa a coeficiente de correlación entre as dúas compoñentes.

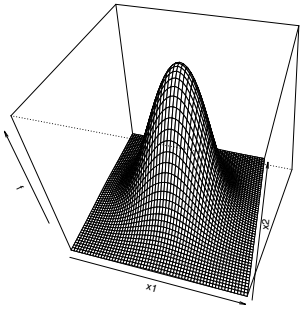
$\rho = -0.80$



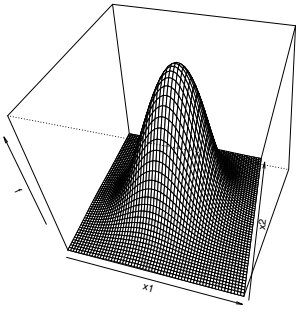
$\rho = -0.40$



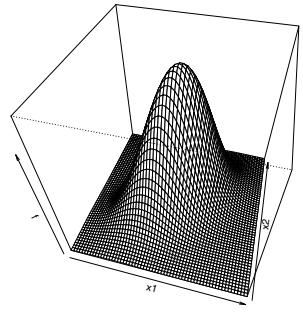
$\rho = 0$



$\rho = 0.25$



$\rho = 0.50$



$\rho = 0.75$

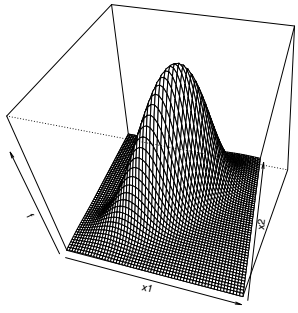


Figura 6.4: Exemplos de densidades de Normais bivariantes.

6.4. Análise de compoñentes principais

6.4.1. Definición, cálculo e propiedades das compoñentes principais

A **Análise de Compoñentes Principais** (**PCA**, do inglés *Principal Component Analysis*) é unha metodoloxía estatística deseñada especificamente para datos multivariantes. O seu obxectivo é realizar unha redución da dimensión e ao mesmo tempo reter a maior cantidade de información posible.

Dado un conxunto de observacións de d variables, o obxectivo da PCA é reexpresar os datos como r variables novas (idealmente, r moito menor ca d) construídas como combinacións lineais das variables orixinais de tal xeito que sexan incorreladas entre si e conteñen a maior cantidade de información (variabilidade) posible. Estas novas variables chámanse **compoñentes principais**.

Os aspectos matemáticos relacionados co cálculo das compoñentes principais consisten esencialmente en atopar autovalores e autovectores da matriz de varianzas-covarianzas ou da matriz de correlacións.

Para ilustrar o funcionamento da PCA, consideremos unha variable bidimensional $\mathbf{X} = (X_1, X_2)$. A idea fundamental da PCA ilústrase na Figura 6.5. Para construír a **primeira compoñente principal** necesitamos atopar unha combinación lineal $Z_1 = X_1u_{11} + X_2u_{12}$ de tal maneira que a varianza de Z_1 sexa máxima. Dado que a varianza pode incrementarse simplemente tomando valores grandes de u_{11} e u_{12}, restrinxímonos a atopar un vector unitario $\mathbf{u}_1$ con compoñentes u_{11} e u_{12}, é dicir $\mathbf{u}_1 = (u_{11}, u_{12})^t$ de forma que $\mathbf{u}_1^t\mathbf{u}_1 = 1$, e tal que a varianza de Z_1 sexa máxima. O problema matemático é polo tanto

$$\textbf{atopar } \mathbf{u}_1 \textbf{ tal que } \mathbf{u}_1^t\mathbf{u}_1 = 1 \textbf{ e } Var(Z_1) = Var(\mathbf{X}\mathbf{u}_1) \textbf{ é máxima.}$$

Nótese que[3] $Var(\mathbf{X}\mathbf{u}_1) = \mathbf{u}_1^t\mathbf{\Sigma}\mathbf{u}_1$, onde $\mathbf{\Sigma}$ é a matriz de varianzas-covarianzas de $\mathbf{X}$. Para resolver este problema, empregamos o método dos multiplicadores de Lagrange. Sexa

$$L(\mathbf{u}_1) = Var(\mathbf{X}\mathbf{u}_1) - \lambda_1(\mathbf{u}_1^t\mathbf{u}_1 - 1) = \mathbf{u}_1^t\mathbf{\Sigma}\mathbf{u}_1 - \lambda_1(\mathbf{u}_1^t\mathbf{u}_1 - 1).$$

Agora temos que derivar L con respecto a u_{11} e u_{12}, igualar a cero e resolver o sistema correspondente. Para isto resulta conveniente escribir as derivadas en forma matricial:

$$\frac{dL(\mathbf{u}_1)}{d\mathbf{u}_1} = \begin{pmatrix} \frac{dL(u_{11},u_{12})}{du_{11}} \\ \\ \frac{dL(u_{11},u_{12})}{du_{12}} \end{pmatrix} \overset{\textbf{(exercicio)}}{=} 2\mathbf{\Sigma}\mathbf{u}_1 - 2\lambda_1\mathbf{u}_1 = 0$$

para concluír que $\mathbf{u}_1$ debe cumprir

$$\mathbf{\Sigma}\mathbf{u}_1 = \lambda_1\mathbf{u}_1.$$

Isto significa que λ_1 é un autovalor de $\mathbf{\Sigma}$ e $\mathbf{u}_1$ é o correspondente autovector. Ademais

$$Var(Z_1) = Var(\mathbf{X}\mathbf{u}_1) = \mathbf{u}_1^t\mathbf{\Sigma}\mathbf{u}_1 = \lambda_1\mathbf{u}_1^t\mathbf{u}_1 = \lambda_1.$$

[3] **Propiedades das distribucións multivariantes.** Sexa $\mathbf{X} = (X_1, X_2, \ldots, X_d)$ un vector aleatorio d-dimensional con vector de medias μ (vector de dimensión d) e matrix de varianzas-covarianzas $\mathbf{\Sigma}$ (matriz de dimensións $d \times d$). Sexan A e b unha matriz $d \times k$ e sexa b un vector k-dimensional de constantes, respectivamente. Entón o vector de medias e a matriz de varianzas-covarianzas do vector aleatorio k-dimensional $\mathbf{X}A + b$ son $\mu A + b$ e $A^t\mathbf{\Sigma}A$, respectivamente.

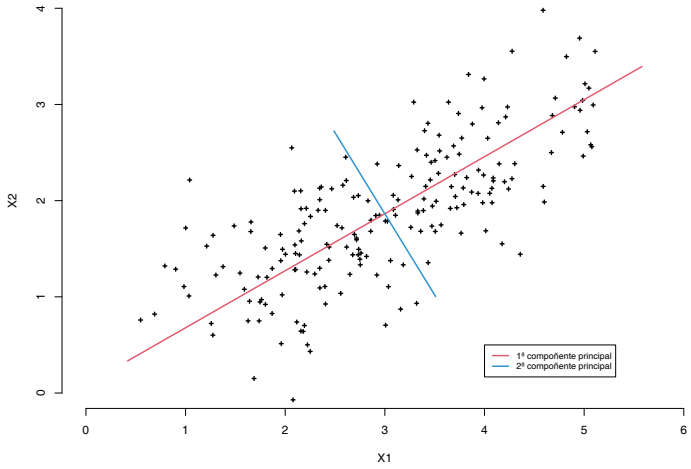


Figura 6.5: Ilustración da PCA con dúas variables.

Polo tanto, se queremos maximizar a varianza de Z_1, simplemente debemos escoller λ_1 como o maior autovalor de $\mathbf{\Sigma}$ e $\mathbf{u}_1$ como o autovector asociado.

Para atopar a **segunda compoñente principal** temos que atopar un vector unitario $\mathbf{u}_2 = (u_{21}, u_{22})^t$ de forma que a combinación lineal $Z_2 = X_1 u_{21} + X_2 u_{22}$ sexa incorrelada coa primeira compoñente principal Z_1 e a varianza de $Var(Z_2)$ sexa máxima. O problema matemático redúcese a atopar un vector unitario $\mathbf{u}_2$ (é dicir, $\mathbf{u}_2^t \mathbf{u}_2 = 1$) ortogonal a $\mathbf{u}_1$ (é dicir, $\mathbf{u}_2^t \mathbf{u}_1 = 0$) e tal que $Var(Z_2) = Var(\mathbf{X}\mathbf{u}_2)$ é máxima. A solución $\mathbf{u}_2$ pódese obter coma antes e resulta ser o autovector de $\mathbf{\Sigma}$ asociado co segundo maior autovector, λ_2. Nese caso, $Var(Z_2) = \lambda_2$.

En xeral, en vez de 2 variables, teremos unha variable d-dimensional $\mathbf{X} = (X_1, X_2, \dots, X_d)$. Sexa $\mathbf{\Sigma}$ a matriz de varianzas-covarianzas (ou a matriz de correlacións, en caso de que as variables estean estandarizadas). Sexan

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$$

os d autovalores de $\mathbf{\Sigma}$ e sexan $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_d$ os correspondentes autovalores[4]. Hai polo tanto d **compoñentes principais**, que son:

- A primeira compoñente principal é a combinación lineal dada por $Z_1 = \mathbf{X}\mathbf{u}_1$, onde $\mathbf{u}_1$ é o autovector asociado ao maior autovalor de $\mathbf{\Sigma}$, λ_1. Nese caso $Var(Z_1) = \lambda_1$.

- A segunda compoñente principal é a combinación lineal dada por $Z_2 = \mathbf{X}\mathbf{u}_2$, onde $\mathbf{u}_2$ é o autovector asociado ao segundo maior autovalor de $\mathbf{\Sigma}$, λ_2. Nese caso $Var(Z_2) = \lambda_2$.

[4] As matrices de varianzas-covarianzas e de correlacións son simétricas e semidefinidas positivas. Isto implica que os seus autovalores son sempre reais e positivos.

- A terceira compoñente principal ...

  etc.

Na práctica, as compoñentes principais calcúlanse a partir dunha mostra, polo que teremos que substituír $\mathbf{\Sigma}$ por unha estimación. A PCA pode levarse a cabo coa matriz de varianzas-covarianzas mostral, $\mathbf{S}$, ou coa matriz de correlacións, $\mathbf{R}$:

- Debe empregarse a matriz de varianzas-covarianzas, $\mathbf{S}$, cando as variables sexan comparables e/ou estean medidas nas mesmas unidades.
- Debe empregarse a matriz de correlacións, $\mathbf{R}$, cando as variables teñen unidades distintas ou están medidas en distintas escalas, que é a situación máis habitual na práctica. Traballar coa matriz de correlacións é equivalente a traballar coas variables estandarizadas. A PCA baseada na matriz de correlacións chámase **PCA estandarizada**.

Aínda que a **interpretación** das compoñentes principais non sempre é doada, en moitos casos pódense ver como índices creados a partir das variables orixinais (ver exemplo máis abaixo). Tamén é interesante calcular a correlación entre as variables orixinais e as compoñentes principais. Pódese comprobar que **as correlacións entre a j-ésima compoñente principal e o conxunto de variables orixinais** vén dada polo vector $\sqrt{\lambda_j}\mathbf{u}_j$.

En R a función `princomp()` realiza a PCA. Esta función crea un obxecto que contén moitos elementos relacionados coa PCA. Para realizar a PCA estandarizada emprégase o argumento `cor=TRUE`.

Exemplo. O conxunto de datos DECATLON contén a información da final da proba de Décatlon dos Xogos Olímpicos de Rio de Janeiro de 2016. Ademais do nome do atleta e da súa puntuación global na proba (variable *puntos*), o conxunto de datos recolle as marcas obtidas en cada unha das dez probas das que consta o décatlon:

- *carreira.100m*: carreira de 100 metros lisos (en s).
- *salto.lonxitude*: salto de lonxitude (en m).
- *lanzamento.peso*: lanzamento de peso (en m).
- *salto.altura*: salto de altura (en m).
- *carreira.400m*: carreira de 400 metros lisos (en s).
- *carreira.110m*: carreira de 110 metros obstáculos (en s).
- *lanzamento.disco*: lanzamento de disco (en m).
- *salto.pertega*: salto con pértega (en m).
- *lanzamento.xavelina*: lanzamento de xavelina (en m).
- *carreira.1500m*: carreira de 1500 metros (en s).

Aplicaremos a PCA para tratar de reducir a información das dez probas. Como as variables teñen unidades e escalas distintas empregaremos PCA estandarizada.

```
> decatlon <- read.csv(file="datos-decatlon.csv",head=TRUE,sep=",")
> datos.para.PCA=decatlon[,3:12]
> PCA.decatlon <- princomp(datos.para.PCA,cor=TRUE)
```

Os vectores unitarios que se usan para construír as compoñentes principais son as columnas da seguinte matriz, que se chama matriz de cargas (ou matriz de *loadings*):

```
> PCA.decatlon$loadings
```

```
Loadings:
                  Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
carreira.100m      0.284         0.551  0.350  0.346  0.162  0.113
salto.lonxitude   -0.452         0.153 -0.386         0.117  0.184
lanzamento.peso           0.486 -0.300 -0.119  0.606  0.206 -0.473
salto.altura      -0.262  0.242  0.426  0.220 -0.463  0.153 -0.616
carreira.400m      0.470         0.259
carreira.110m      0.419 -0.138  0.122 -0.356         0.515
lanzamento.disco   0.175  0.530        -0.433 -0.353  0.127  0.141
salto.pertega     -0.216  0.141  0.561 -0.391  0.354 -0.335  0.120
lanzamento.xavelina       0.581         0.439                0.533
carreira.1500m     0.410  0.205                -0.167 -0.702 -0.140
                  Comp.8 Comp.9 Comp.10
carreira.100m             0.186  0.544
salto.lonxitude    0.389 -0.488  0.426
lanzamento.peso    0.106 -0.112
salto.altura                    -0.114
carreira.400m     -0.326 -0.752 -0.169
carreira.110m      0.495  0.203 -0.311
lanzamento.disco  -0.441  0.245  0.289
salto.pertega     -0.127  0.160 -0.415
lanzamento.xavelina  0.210 -0.136 -0.328
carreira.1500m     0.470         0.146

               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
SS loadings       1.0    1.0    1.0    1.0    1.0    1.0    1.0
Proportion Var    0.1    0.1    0.1    0.1    0.1    0.1    0.1
Cumulative Var    0.1    0.2    0.3    0.4    0.5    0.6    0.7
               Comp.8 Comp.9 Comp.10
SS loadings       1.0    1.0     1.0
Proportion Var    0.1    0.1     0.1
Cumulative Var    0.8    0.9     1.0
```

Os espazos en branco da matriz anterior significan que o valor correspondente está moi preto de cero. O signo de todo o vector unitario é arbitrario, pero os signos de cada coeficiente

son interesantes: as variables ás que lles corresponden coeficientes co mesmo signo teñen unha relación directa, mentras que as variables con coeficientes de signos opostos teñen unha relación inversa. Neste exemplo, podemos ver que:

- A primeira compoñente principal dálles máis importancia ás variables *salto.lonxitude* (-0.452), *carreira.400m* (0.470), *carreira.110m* (0.419) e *carreira.1500m* (0.410). Nótese que neste caso os signos opostos entre o coeficiente de *salto.lonxitude* e o do resto de variables débese a que nas carreiras valores altos significan peores resultados, mentras que no salto de lonxitude os mellores resultados son aqueles que teñen valores altos.
- A segunda compoñente principal dálles importancia ás tres probas de lanzamento: peso (0.486), disco (0.530) e xavelina (0.581).
- A terceira compoñente principal dálles importancia ás probas de salto de altura e con pértega (0.426 e 0.561, respectivamente) e á proba de 100 metros lisos (0.551).

Os valores das compoñentes principais están gardadas nos `scores` do obxecto creado pola función `princomp()`:

```
> PCA.decatlon$scores
```

```
     Comp.1 Comp.2 Comp.3 Comp.4
[1,] -3.712  0.416 -1.035 -1.597
[2,] -1.854  1.733  0.290 -0.942
[3,] -2.871 -0.013 -1.795 -0.258
[4,] -2.081  0.476  0.063 -0.258
[5,] -2.136 -0.129 -0.507  1.079
[6,] -0.168  1.562  0.607  1.302
```

Os valores de cada compoñente pódense interpretar como un índice do desempeño do atleta nas probas ás que a compoñente correspondente lles dá máis importancia.

As correlacións entre as compoñentes principais e as variables orixinais confirman a interpretación das compoñentes dada anteriormente:

```
> cor(datos.para.PCA,PCA.decatlon$scores)
```

```
                    Comp.1 Comp.2 Comp.3 Comp.4
carreira.100m        0.531  0.084  0.684  0.337
salto.lonxitude     -0.845  0.075  0.190 -0.371
lanzamento.peso      0.079  0.684 -0.373 -0.114
salto.altura        -0.491  0.341  0.529  0.212
carreira.400m        0.879 -0.016  0.322 -0.026
carreira.110m        0.785 -0.194  0.152 -0.343
lanzamento.disco     0.327  0.745 -0.068 -0.417
salto.pertega       -0.404  0.198  0.697 -0.376
lanzamento.xavelina -0.106  0.817 -0.053  0.423
carreira.1500m       0.767  0.288 -0.023 -0.076
```

Por exemplo, a primeira compoñente principal está fortemente correlada coas variables *salto.lonxitude*, *carreira.400m*, *carreira.110m* e *carreira.1500m*. Debemos ter en conta o signo das correlacións para interpretar os índices. Claramente, os mellores desempeños nesta variable serán os correpondentes con índices negativos (valores altos do salto de lonxitude e valores pequenos nos tempos das carreiras). □

6.4.2. Selección do número de compoñentes

Debido ás propiedades dos autovalores, nótese que

$$\text{variabilidade total} = \sum_{j=1}^{d} S_{X_j}^2 = \text{traza}(\mathbf{S}) = \sum_{j=1}^{d} \lambda_j.$$

Polo tanto, as r primeiras compoñentes principais explican a seguinte proporción de variabilidade:

$$\frac{\sum_{j=1}^{r} \lambda_j}{\sum_{j=1}^{d} \lambda_j}.$$

No caso de empregar PCA estandarizada, nótese que a variabilidade total é $\text{traza}(\mathbf{R}) = d$, así que a proporción de variabilidade explicada polas r primeiras compoñentes principais é

$$\frac{\sum_{j=1}^{r} \lambda_j}{d}.$$

O principal obxectivo da PCA é reducir a dimensión, polo que nos quedaremos con algunhas delas (digamos r) sempre que expliquen unha gran proporción de variabilidade. Na práctica, pódense aplicar diversos criterios:

- Manter as r primeiras compoñentes que explican unha gran porcentaxe da variabilidade total (por exemplo, 80 %).
- Manter as compoñentes principais que contribúen cunha cantidade de variabilidade por riba da media, é dicir, cando $\lambda_j >$ (variabilidade total)$/d$. No caso da PCA estandarizada, isto significa manter as compoñentes principais que cumpren $\lambda_j > 1$.
- Usar o *scree-plot* (gráfico de barras de λ_j fronte a j) e buscar un cóbado, é dicir, cando a contribución dunha compoñente principal sexa similar á anterior.

Exemplo. (cont.) A proporción de variabilidade explicada polas compoñentes principais pódese obter facendo

```
> summary(PCA.decatlon)

Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4
Standard deviation     1.8711539 1.4066685 1.2428048 0.96281343
Proportion of Variance 0.3501217 0.1978716 0.1544564 0.09270097
```

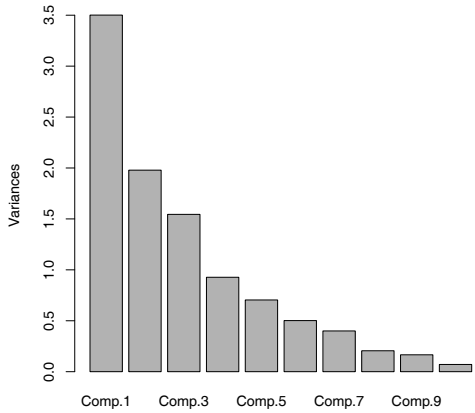


Figura 6.6: Datos DECATLON. Scree-plot correspondente á PCA estandarizada.

```
Cumulative Proportion  0.3501217 0.5479933 0.7024497 0.79515068
                          Comp.5     Comp.6     Comp.7     Comp.8
Standard deviation     0.8392115 0.70830748 0.63259007 0.45335222
Proportion of Variance 0.0704276 0.05016995 0.04001702 0.02055282
Cumulative Proportion  0.8655783 0.91574823 0.95576524 0.97631807
                           Comp.9    Comp.10
Standard deviation     0.40725166 0.26639332
Proportion of Variance 0.01658539 0.00709654
Cumulative Proportion  0.99290346 1.00000000
```

Para obter o scree-plot que aparece na Figura 6.6 en R escribimos

```
> plot(PCA.decatlon)
```

As tres primeiras compoñentes principais explican o 35.0 %, 19.8 % e 15.4 % da variabilidade total, respectivamente. Entre as tres explican o 70.2 %. Estas tres compoñentes son as que teñen variabilidades maiores de 1. Se queremos alcanzar o 80 % teremos que incorporar a cuarta compoñente principal, que aporta outro 9.2 %.

Pódense facer representacións gráficas adicionais. Ás veces é interesante facer gráficos de dispersión dos índices, especialmente os das primeiras compoñentes principais, para identificar a posición de cada individuo segundo os valores das compoñentes. A Figura 6.7 contén os gráficos de dispersión por pares das tres primeiras compoñentes principais. Os gráficos dan unha idea do desempeño de cada atleta en cada unha das compoñentes principais. □

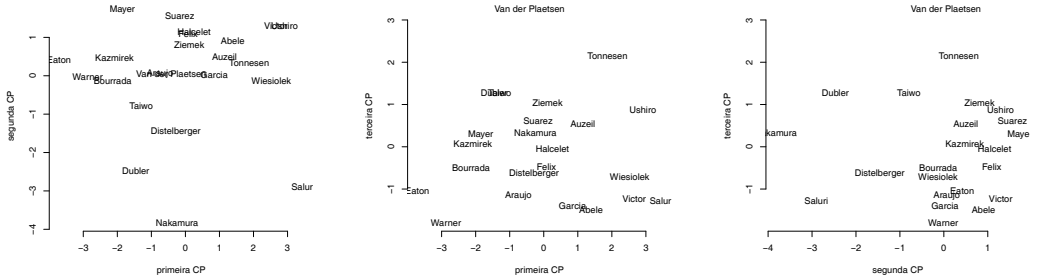


Figura 6.7: Datos DECATLON. Gráficos de dispersión dos scores obtidos da PCA estandarizada.

Exercicio 6.4. *Os medallistas no Décatlon nos xogos Olímpicos de 2016 foron Eaton (ouro), Mayer (prata) e Warner (bronce).*

(a) Localiza estes atletas nos gráficos da Figura 6.7 e interpreta a súa posición.

(b) Calcula as correlacións entre as compoñentes principais e a puntuación global de cada atleta (variable puntos*). Interpreta o resultado.*

Exercicio 6.5. *Realiza unha análise de compoñentes principais das catro variables numéricas do conxunto de datos* MUNDO2016*. Interpreta os resultados.*

Exercicio 6.6. *A matriz de correlacións dunha variable bidimensional (X_1, X_2) é da forma*

$$\mathbf{R} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

onde $\rho \in [-1, 1]$ é o coeficiente de correlación entre X_1 e X_2.

(a) Atopa os autovalores e autovectores de $\mathbf{R}$.

(b) Interpreta os resultados de (a) en termos da PCA estandarizada.

(c) Atopa os valores de ρ para os cales a primeira compoñente principal explica máis do 80 % da variabilidade total.

6.5. Creación de grupos: métodos cluster

O obxectivo principal da **análise cluster** consiste en dividir os individuos da mostra en subgrupos (chamados **clusters**) de acordo a algunha medida de homoxeneidade interna. Os métodos cluster son técnicas de **clasificación non supervisada**.

As cuestións que intenta abordar a análise cluster son:

- Cantos grupos homoxéneos hai na mostra?
- Que individuos pertencen a cada grupo?

O resultado da análise clúster é unha partición da mostra en grupos de tal xeito que

(a) cada individuo pertence a un e só a un dos grupos,

(b) cada individuo queda clasificado nun grupo, e

(c) cada grupo é homoxéneo internamente.

Tipos de métodos cluster:

- **Métodos xerárquicos**. A partición obtense mediante clasificación xerárquica, onde os casos están ordenados en varios niveis de forma que os niveis máis altos conteñen aos niveis máis baixos. A clasificación pódese representar cun **dendograma**.

  Exemplo. Métodos aglomerativos.

- **Métodos non xerárquicos**. Os grupos baséanse en distancias entre os individuos e a clasificación obtida non garante unha estrutura xerárquica.

  Exemplo. Método das K-medias.

6.5.1. Método das K-medias

O método das K-medias asume que existen K grupos homoxéneos (o número K ten que ser especificado de antemán) e procede do seguinte xeito:

1. Toma uns valores iniciais para os K **centroides** ou medias. Para facer isto pódese proceder de varias formas:
   - Formar K grupos aleatoriamente e calcular os seus centroides.
   - Tomar como centroides os K puntos máis distantes.
   - Empregar información a priori sobre os posibles grupos se se dispón dela.
2. Calcula a distancia entre cada individuo a cada centroide e asígnao ao grupo de cuxo centroide diste menos. Despois de cada asignación, os centroides recalcúlanse secuencialmente.

3. O procedemento iterativo remata cando se alcanza un determinado criterio de optimalidade (por exemplo, a homoxeneidade interna dos grupos non mellora cunha nova iteración).

O criterio de optimalidade habitual é minimizar a suma de distancias euclidianas ao cadrado entre as observacións e os centroides dos grupos. Sexa $\mathbf{X}_i = (X_{i1}, X_{i2}, \dots, X_{id}), i = 1, \dots, n$, unha mostra de observacións multivariantes onde cada observación é d-dimensional. O criterio de optimalidade funciona do seguinte xeito. Supoñamos que as n observacións están clasificadas de forma exhaustiva en K grupos, é dicir, cada observación está incluída nun só grupo. Denotemos $\mathbf{X}_{i,k}$ para indicar que a observación $\mathbf{X}_i$ está incluída no grupo k e denotemos por n_k o tamaño mostral do grupo k (obviamente, $\sum_{k=1}^{K} n_k = n$). Calculamos a **suma de cadrados dentro dos grupos**

$$\text{SSWG} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} d(\mathbf{X}_{i,k}, \bar{\mathbf{X}}_k),$$

onde $\bar{\mathbf{X}}_k$ é o vector de medias do grupo k e $d(\mathbf{X}_{i,k}, \bar{\mathbf{X}}_k)$ é o cadrado da distancia euclidiana entre a observación $\mathbf{X}_{i,k}$ e o seu vector de medias.

A solución obtida ao minimizar SSWG proporcionaría a clasificación óptima. Non obstante, acadar isto pode ser computacionalmente moi esixente e normalmente a solución final obtense por métodos heurísticos. O algoritmo pode ser sensible á elección inicial dos centroides, polo que se recomenda usar varios posibles puntos iniciais, especialmente cando K é grande (digamos, por exemplo, para $K > 4$).

En R a función `kmeans()` realiza o método das K-medias.

Exemplo. Consideremos o conxunto de datos CLIMACIDADES, que contén información meteorolóxica sobre 53 cidades de españolas. Consta das seguintes variables con información recollida durante o ano 2015[5]:

- *cidade* e *codigo*: nome da cidade e código de tres letras.
- *temp*: temperatura media anual (en graos Celsius).
- *hsol*: horas de sol durante todo o ano.
- *prec*: cantidade de precipitación anual (mm).
- *lonxitude*, *latitude* e *altitude* da cidade.

```
> climacidades <- read.csv("datos-climacidades.csv",sep=",",dec=".",header=TRUE)
> attach(climacidades)
```

Para ilustrar como funciona o método empregaremos só dúas variables, *hsol* e *prec*. O gráfico de dispersión destas dúas variables aparece na Figura 6.8. A primeira vista, poderiamos dicir que hai dous grupos de cidades claramente separados, aínda que algunhas cidades non resultan fáciles de clasificar.

Primeiro aplicamos o método das K-medias con $K = 2$ grupos:

[5]Fonte: Instituto Nacional de Estadística (INE).

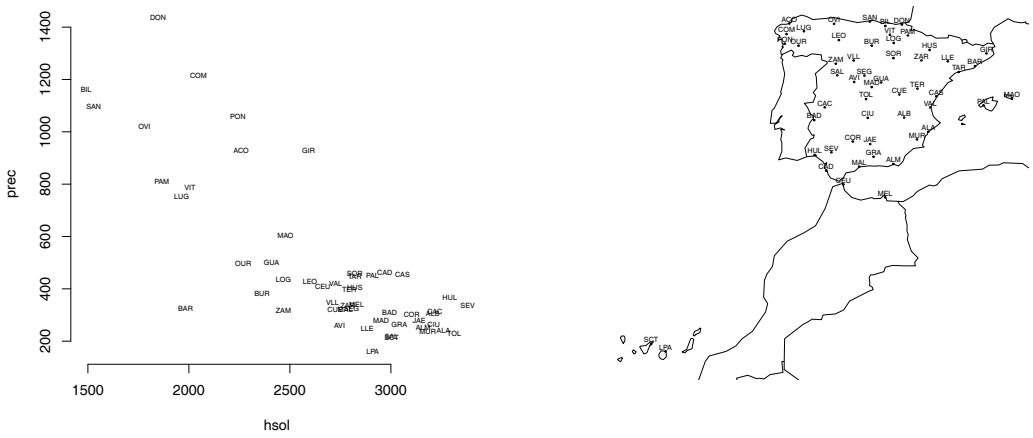


Figura 6.8: Conxunto de datos CLIMACIDADES. Esquerda: gráfico de dispersión das variables *hsol* and *prec*. Dereita: localización xeográfica das cidades.

```
> datos.para.cluster <- climacidades[,c("hsol","prec")]
> Kmedias2 <- kmeans(datos.para.cluster,centers=2)
```

Os dous clusters resultantes son:

```
> as.character(cidade)[Kmedias2$cluster==1]
```

```
 [1] "Oviedo"                  "Santander"
 [3] "Barcelona"               "Girona"
 [5] "A Coruna"                "Lugo"
 [7] "Ourense"                 "Santiago de Compostela"
 [9] "Pontevedra"              "Pamplona"
[11] "Vitoria-Gasteiz"         "Bilbao"
[13] "Donostia-San Sebastian"
```

```
> as.character(cidade)[Kmedias2$cluster==2]
```

```
 [1] "Almeria"                    "Cadiz"
 [3] "Cordoba"                    "Granada"
 [5] "Jaen"                       "Huelva"
 [7] "Malaga"                     "Sevilla"
 [9] "Huesca"                     "Teruel"
[11] "Zaragoza"                   "Mao"
[13] "Palma"                      "Las Palmas de Gran Canaria"
[15] "Santa Cruz de Tenerife"     "Avila"
```

```
[17] "Burgos"                    "Leon"
[19] "Salamanca"                 "Segovia"
[21] "Soria"                     "Valladolid"
[23] "Zamora"                    "Albacete"
[25] "Ciudad Real"               "Cuenca"
[27] "Guadalajara"               "Toledo"
[29] "Lleida"                    "Tarragona"
[31] "Alacant"                   "Castello"
[33] "Valencia"                  "Badajoz"
[35] "Caceres"                   "Madrid"
[37] "Murcia"                    "Logrono"
[39] "Ceuta"                     "Melilla"
```

Os dous grupos claramente distinguen entre cidades moi chuviosas e pouco soleadas (fundamentalmente no norte da Península) e cidades pouco chuviosas e soleadas (fundamentalmente no centro e sur da Península e nos dous arquipélagos), tal e como se pode ver na Figura 6.9.

Para $K = 3$ grupos obtemos os seguintes clusters:

```
> Kmedias3 <- kmeans(datos.para.cluster,centers=3)
> as.character(cidade)[Kmedias3$cluster==1]

 [1] "Oviedo"                 "Santander"
 [3] "A Coruna"               "Lugo"
 [5] "Santiago de Compostela" "Pontevedra"
 [7] "Pamplona"               "Vitoria-Gasteiz"
 [9] "Bilbao"                 "Donostia-San Sebastian"

> as.character(cidade)[Kmedias3$cluster==2]

 [1] "Malaga"      "Huesca"      "Teruel"      "Zaragoza"
 [5] "Mao"         "Avila"       "Burgos"      "Leon"
 [9] "Segovia"     "Soria"       "Valladolid"  "Zamora"
[13] "Cuenca"      "Guadalajara" "Barcelona"   "Girona"
[17] "Tarragona"   "Valencia"    "Ourense"     "Logrono"
[21] "Ceuta"       "Melilla"

> as.character(cidade)[Kmedias3$cluster==3]

 [1] "Almeria"                    "Cadiz"
 [3] "Cordoba"                    "Granada"
 [5] "Jaen"                       "Huelva"
 [7] "Sevilla"                    "Palma"
 [9] "Las Palmas de Gran Canaria" "Santa Cruz de Tenerife"
[11] "Salamanca"                  "Albacete"
```

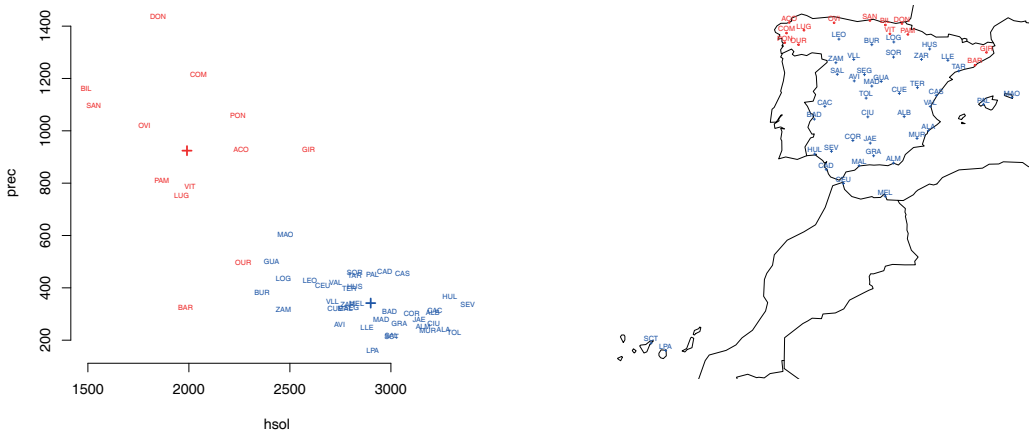


Figura 6.9: Conxunto de datos CLIMACIDADES. Clusters obtidos cando se aplica o método das K-medias con $K = 2$. As cruces indican o centroide de cada grupo.

```
[13] "Ciudad Real"          "Toledo"
[15] "Lleida"               "Alacant"
[17] "Castello"             "Badajoz"
[19] "Caceres"              "Madrid"
[21] "Murcia"
```

A grandes rasgos, os tres grupos formados distinguen entre cidades do norte da Península (moi chuviosas e pouco soleadas), cidades do centro norte (con valores intermedios) e cidades do sur (pouco chuviosas e moi soleadas).

As Figuras 6.9 e 6.10 amosan os clusters obtidos no diagrama de dispersión e no mapa con $K = 2$ e $K = 3$ grupos, respectivamente. Nótese que non se obtén unha clasificación xerárquica, xa que o terceiro cluster cando $K = 3$ está formado por cidades que pertencen a cada un dos dous clusters cando $K = 2$. Hai que ter en conta tamén que a clasificación pode cambiar se aplicamos o método varias veces, especialmente se o número de clusters é grande, xa que R usa unha selección aleatoria de centroides como puntos de partida. □

Cantos clusters deberían empregarse na práctica? Depende da información que teñamos sobre o problema, do conxunto de datos e do interese práctico da aplicación en particular. Unha posible maneira de decidir o número de grupos consiste en facer un gráfico de SSWG fronte a K e buscar unha forma de cóbado, é dicir, buscar o valor de K para o cal non hai unha mellora substancial no SSWG ao pasar de K a $K + 1$.

Exemplo. (cont.) Conxunto de datos CLIMACIDADES. O valor de SSWG pode obterse a partir do obxecto xerado pola función `kmeans()`. Á vista da Figura 6.11 parece que $K = 2$ (ou como moito $K = 3$) é o número de grupos recomendado. □

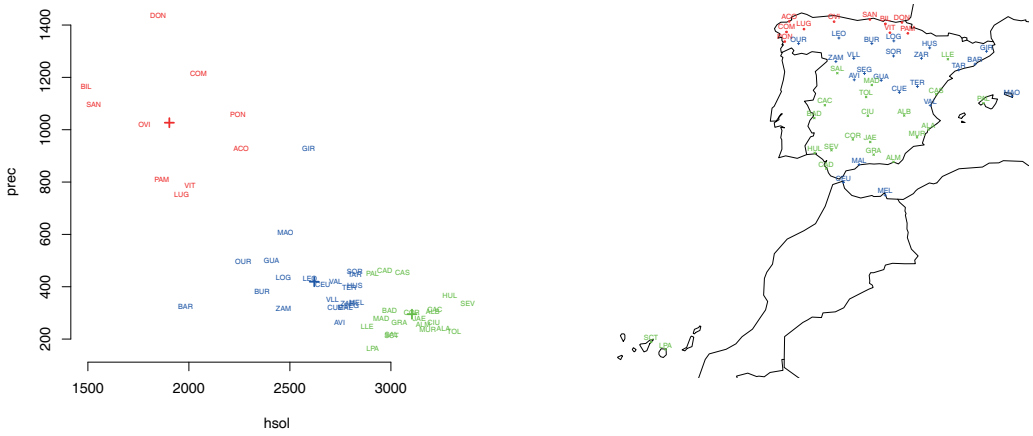


Figura 6.10: Data set CLIMACIDADES. Clusters obtidos cando se aplica o método das K-medias con $K = 3$. As cruces indican o centroide de cada grupo.

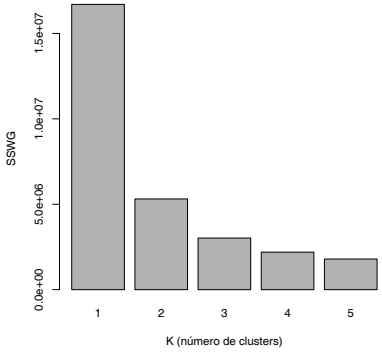


Figura 6.11: Conxunto de datos CLIMACIDADES. Valores de SSWG en función do número de clusters.

Exercicio 6.7.

(a) Repite o exemplo anterior varias veces para comprobar a robustez do método das K-medias. Depende do número de clusters?

(b) Aplica o método das K-medias incluíndo tamén a variable temp. *Compara os resultados cos obtidos no exemplo anterior.*

Exercicio 6.8. *Considera o conxunto de datos* MUNDO2016. *Aplica o método das K-medias para crear grupos baseados nas variables* esperanza.vida, taxa.fertilidade, taxa.mortalidadeinfantil *e* poboacion.rural. *Comproba se os países da OCDE pertencen ao mesmo grupo cando se considera un número pequeno de clusters (2 ou 3).*

6.5.2. Métodos xerárquicos

Os métodos cluster **xerárquicos** traballan de xeito iterativo. Inicialmente cada individuo forma un cluster e a continuación procédese agrupando os individous ou grupos máis cercanos de acordo a uns certos criterios ata conseguir un único cluster. O proceso é aglomerativo.

Estes métodos funcionan con similitudes. Unha **similitude** entre dous individuos con observacións X_i e X_j é unha medida $s(X_i, X_j)$ que satisfai

(a) $0 \leq s(X_i, X_j) \leq 1$;

(b) $s(X_i, X_i) = 1$; e

(c) $s(X_i, X_j) = s(X_j, X_i)$.

Se $s(X_i, X_j)$ é unha similitude, entón

$$d(X_i, X_j) = \sqrt{2(1 - s(X_i, X_j))}$$

define unha distancia. Reciprocamente, se $d(X_i, X_j)$ é unha distancia, entón

$$s(X_i, X_j) = \frac{1}{1 + d(X_i, X_j)}$$

define unha similitude. Pódense usar distintas distancias (por exemplo, a distancia euclidiana ou a distancia de Mahalanobis) para construír similitudes.

Existen moitos métodos cluster de tipo **aglomerativo**. Veremos moi brevemente tres deles:

- **Enlace simple** (tamén chamado "veciño máis próximo"): a similitude entre dous clusters está dada pola máxima similitude (mínima distancia) entre os seus membros.

- **Enlace completo** (tamén chamado "veciño máis alonxado"): a similitude entre dous clusters está dada pola mínima similitude (máxima distancia) entre os seus membros.
- **Método centroide**: a similitude entre dous clusters defínese a partir da distancia entre os seus centroides.

O resultado do cluster aglomerativo pódese representar nun **dendograma** (véxase a Figura 6.12).

En R, a función `hclust()` realiza o cluster xerárquico. Aplícase a unha matriz de distancias obtidas mediante a función `dist()`, que por defecto calcula as distancias euclidianas. O dendograma obtense aplicando a función xenérica `plot()` ao obxecto xerado pola función `hclust()`.

Exemplo. Conxunto de datos CLIMACIDADES. Aplicaremos o método do enlace completo coas variables *hsol* e *prec*.En R facemos

```
> cx.completo <- hclust(dist(datos.para.cluster),method="complete")
```

O dendograma obtense mediante

```
> plot(cx.completo,labels=cidade)
```

e aparece na Figura 6.12. Este método primeiro crea dous grupos que distinguen as cidades pouco chuviosas das cidades chuviosas. □

O dendograma pódese "cortar" ao nivel desexado empregando a función `cutree()` de R.

Exercicio 6.9. *Aplica o método do enlace completo engadindo a variable* temp *ás xa consideradas* hsol *e* prec *do conxunto de datos* CLIMACIDADES. *Son os resultados distintos dos obtidos no exemplo anterior?*

6.6. Análise discriminante

A análise discriminante está relacionada cos problemas de clasificación (ver os conceptos básicos relacionados con estes problemas na sección 2.5 do Capítulo 2). En particular, esta metodoloxía pertence aos métodos de **clasificación supervisada**. Exemplos de análise discriminante son: clasificar a un paciente como san/enfermo, clasificar a un paciente como enfermo de A/enfermo de B, clasificar os clientes bancarios segundo a súa posibilidade de incumprimento ao pagar un crédito, clasificar o correo electrónico entrante como spam/non spam etc.

O problema abordado pola análise discriminante é o seguinte. A poboación divídese en 2 grupos (**grupo 1**/**Negativo** e **grupo 2**/**Positivo**). Cada individuo pertence a un e só a un destes dous grupos. O obxectivo é crear unha **regra de decisión** para clasificar individuos novos en base a d variables cuantitativas recollidas nun vector $\mathbf{X}$.

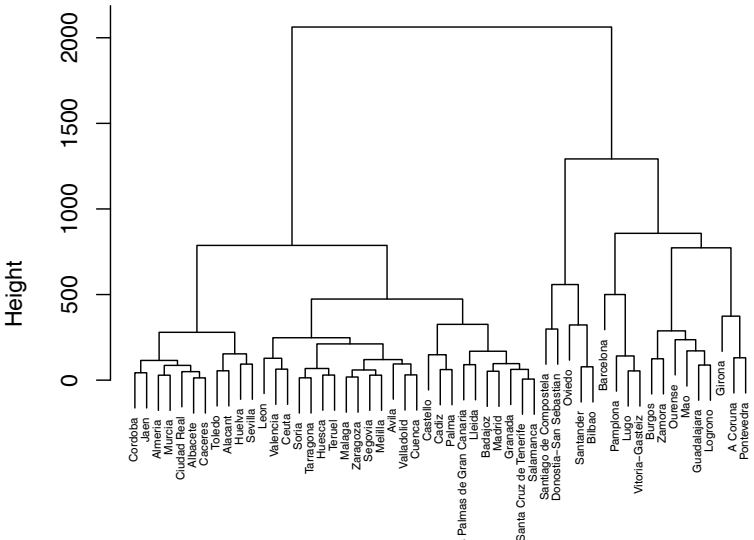


Figura 6.12: Conxunto de datos CLIMACIDADES. Dendograma obtido polo método do enlace completo.

Os elementos que interveñen na construción da regra de decisión son os seguintes:

- Probabilidades *a priori* de pertencer a cada un dos grupos: p_1 e $p_2 = 1-p_1$. Por exemplo, se o problema de clasificación é o diagnóstico dunha enfermidade, entón p_2 será a prevalencia.
- Funcións de densidade multidimensionais en cada un dos grupos:

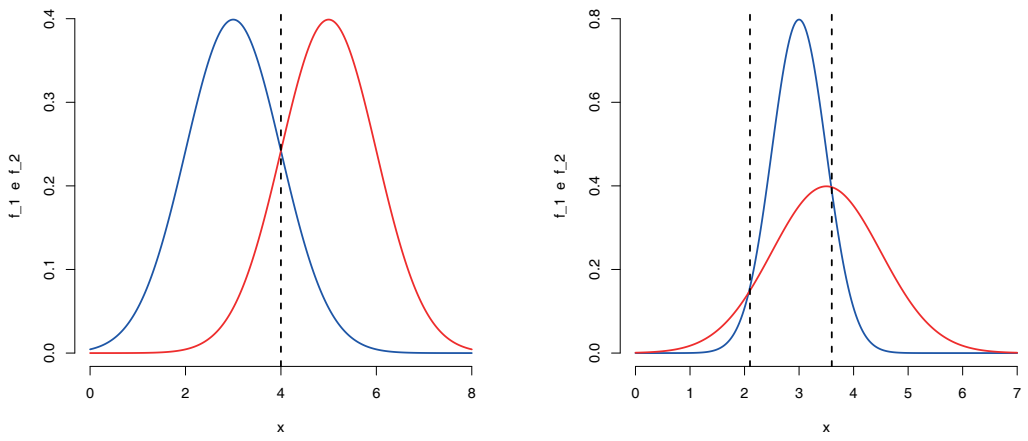
$$f_1(\mathbf{x}) = P(\mathbf{x} \mid \text{grupo 1}) \qquad \text{e} \qquad f_2(\mathbf{x}) = P(\mathbf{x} \mid \text{grupo 2}).$$

- Grazas ao Teorema de Bayes, os dous elementos anteriores pódense combinar para obter as probabilidades *a posteriori* de cada grupo:

$$P(\text{grupo 1} \mid \mathbf{x}) = \frac{f_1(\mathbf{x})p_1}{f_1(\mathbf{x})p_1 + f_2(\mathbf{x})p_2} \qquad \text{e} \qquad P(\text{grupo 2} \mid \mathbf{x}) = \frac{f_2(\mathbf{x})p_2}{f_1(\mathbf{x})p_1 + f_2(\mathbf{x})p_2}.$$

Unha **regra de decisión** natural é a seguinte: un individuo con características $\mathbf{x}$ é asignado ao...

- grupo 1 se $P(\text{grupo 1} \mid \mathbf{x}) > P(\text{grupo 2} \mid \mathbf{x}) \iff f_1(\mathbf{x})p_1 > f_2(\mathbf{x})p_2$.
- grupo 2 se $P(\text{grupo 1} \mid \mathbf{x}) \leq P(\text{grupo 2} \mid \mathbf{x}) \iff f_1(\mathbf{x})p_1 \leq f_2(\mathbf{x})p_2$.



(a) (b)

Figura 6.13: Exemplos de regras de decisión na análise discriminante cando $\mathbf{X}$ é unidimensional e $p_1 = p_2$. (a) Hai un único punto de corte. Os individuos con $x < 4$ son asignados ao grupo 1. (b) Son necesarios dous puntos de corte.

No caso particular no que $p_1 = p_2$, entón a regra redúcese a clasificar no grupo 1 se $f_1(\mathbf{x}) > f_2(\mathbf{x})$ (ver Figura 6.13).

Na práctica, as probabilidades *a priori* e as densidades non son coñecidas. Para estimalas necesitamos unha mostra de individuos para os cales sexa coñecido o grupo ao que pertencen.

6.6.1. Regra discriminante lineal de Fisher

Consideraremos agora un caso sinxelo para construír unha regra de clasificación. Supoñamos que a distribución da variable d-dimensional $\mathbf{X} = (X_1, \ldots, X_d)$ en cada grupo é **Normal Multivariante** con vectores de medias $\boldsymbol{\mu}_1$ e $\boldsymbol{\mu}_2$, respectivamente, e **matrices de varianzas-covarianzas iguais** Σ (a mesma nos dous grupos):

- Densidade no grupo 1:

$$f_1(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^t\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\right\}.$$

- Densidade no grupo 2:

$$f_2(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^t\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right\}.$$

Segundo a regra de decisión explicada na sección anterior, un individuo con características $\mathbf{x}$ será asignado ao grupo 1 se

$$f_1(\mathbf{x})p_1 > f_2(\mathbf{x})p_2,$$

é dicir, se

$$\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^t\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\right\} p_1 > \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^t\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right\} p_2.$$

Eliminando os elementos comúns a ambos os dous lados da igualdade e tomando logaritmos, obtemos

$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^t\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) + \log p_1 > -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^t\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) + \log p_2,$$

ou equivalentemente

$$\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^t\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^t\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) < \log p_1 - \log p_2.$$

Nótese que $(\mathbf{x}-\boldsymbol{\mu}_1)^t\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)$ é o cadrado da distancia de Mahalanobis entre $\mathbf{x}$ e $\boldsymbol{\mu}_1$ no grupo 1 e $(\mathbf{x}-\boldsymbol{\mu}_2)^t\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)$ é o cadrado da distancia de Mahalanobis entre $\mathbf{x}$ e $\boldsymbol{\mu}_2$ no grupo 2. En caso de que $p_1 = p_2$, entón $\log p_1 - \log p_2 = 0$ e a regra de decisión consiste simplemente en asignar $\mathbf{x}$ ao grupo 1 se a súa distancia de Mahalanobis con respecto a $\boldsymbol{\mu}_1$ é menor ca a $\boldsymbol{\mu}_2$.

Despois dalgúns cálculos, tamén podemos escribir

$$\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^t\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^t\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) = (\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^t\Sigma^{-1}\left(\mathbf{x} - \frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}\right).$$

A **regra discriminante lineal de Fisher** é

$$\text{- asignar ao grupo 1 se} \quad \mathbf{w}^t\left(\mathbf{x} - \frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}\right) < c,$$

$$\text{- asignar ao grupo 2 se} \quad \mathbf{w}^t\left(\mathbf{x} - \frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}\right) > c,$$

onde

$c = \log p_1 - \log p_2$, que só depende das probabilidades *a priori*, e

$\mathbf{w}^t = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^t\Sigma^{-1}$, que depende dos vectores de medias e da matriz de varianzas-covarianzas común.

Esta regra ten unha **interpretación xeométrica** en termos dun hiperplano que separa os grupos:

- Se $d = 1$, entón a regra de decisión está baseada nun único punto. Ver Figura 6.13-(a).

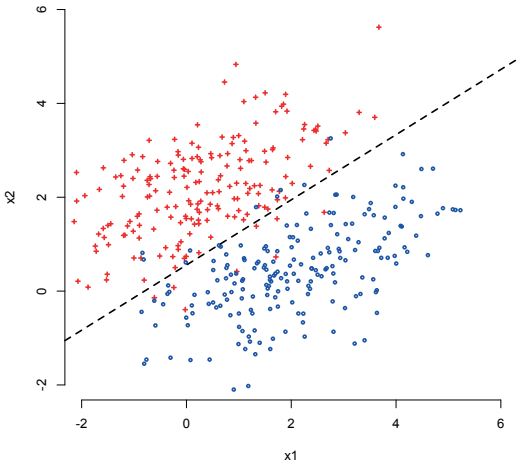


Figura 6.14: Regra discriminante lineal de Fisher cando $d = 2$. A clasificación realízase en base á liña recta que mellor separa os grupos. Algúns individuos quedan mal clasificados.

- Se $d = 2$, entón a regra de decisión está baseada nunha recta. Ver Figura 6.14.
- Se $d = 3$, entón a regra de decisión está baseada nun plano. Etc.

A regra de clasificación en realidade baséase nun **score de clasificación** unidimensional

$$\mathbf{w}^t \left(\mathbf{x} - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2} \right).$$

Na práctica teremos que estimar p_1, p_2, $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ e Σ:

- Se p_1 e p_2 son descoñecidos e os grupos están ben representados na mostra, entón pódense estimar pola correspondentes proporcións mostrais, $\hat{p}_1$ e $\hat{p}_2$ e despois considerar $\hat{c} = \log \hat{p}_1 - \log \hat{p}_2$. Nalgunhas aplicacións, as proporcións p_1 e p_2 poden ser coñecidas ou estimarse a partir doutras mostras ou estudos (por exemplo, cando representan a prevalencia dunha enfermidade ou patoloxía).
- $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ estímanse mediante os vectores de medias mostrais $\bar{\mathbf{X}}_1$ e $\bar{\mathbf{X}}_2$.
- Σ estímase mediante a matriz de varianzas-covarianzas mostral $\mathbf{S}$.

Para clasificar un individuo do cal descoñecemos o grupo e ten características $\mathbf{x}$, calcularemos o score de clasificación e compararémolo co punto de corte. Se o score é menor ca o punto de corte c entón asignámolo ao grupo 1, mentras que se supera o punto de corte será asignado ao

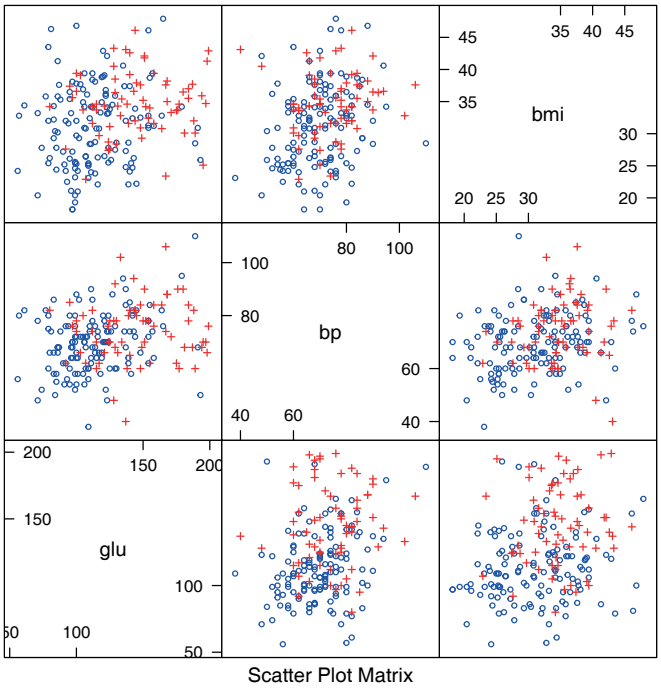


Figura 6.15: Conxunto de datos DIABETES. Gráficos de dispersion das variables *glu*, *bp* e *bmi* distinguindo mulleres non diabéticas (azul) e diabéticas (vermello).

grupo 2. Tamén se pode empregar a probabilidade *a posteriori* de cada grupo e clasificalo no grupo que lle outorga unha maior probabilidade.

En R, a función `lda()`[6] realiza a análise discriminante lineal de Fisher.

Exemplo. Consideremos o conxunto de datos DIABETES.

```
> diabetes <- read.table(file="datos-diabetes.txt",header=TRUE)
> attach(diabetes)
```

Aplicaremos a análise discriminante lineal de Fisher para clasificar ás mulleres como diabéticas/non diabéticas segundo a información proporcionada polas variables cuantitativas *glu*, *bp* e *bmi*. A Figura 6.15 mostra os diagramas de dispersión das variables segundo os grupos diabético/non diabético. Para obter a regra de clasificación en R, facemos

```
> library(MASS)
> lda.diabetes <- lda(type~glu+bp+bmi)
> lda.diabetes
```

[6]A función `lda()` non pertence á instalación básica de R. Requírese o paquete `MASS`.

```
Call:
lda(type ~ glu + bp + bmi)

Prior probabilities of groups:
  No  Yes
0.66 0.34

Group means:
         glu       bp      bmi
No  113.1061 69.54545 31.07424
Yes 145.0588 74.58824 34.70882

Coefficients of linear discriminants:
           LD1
glu 0.03047436
bp  0.01011390
bmi 0.06514745
```

Por defecto as probabilidades *a priori* p_1 e p_2 estímanse a partir das correspondentes proporcións mostrais 0.66 (132 de 200) e 0.34 (68 de 200). Isto sería correcto se a mostra fose seleccionada aleatoriamenta na poboación e estas proporcións representasen as verdadeiras proporcións reais de mulleres diabéticas/non diabéticas na poboación total. Segundo a información da OMS[7], estas proporcións non parecen razoables. Máis ben, parece que a prevalencia da diabetes está arredor do 9 % na poboación xeral (8.3 % no grupo de mulleres). Estas probabilidades deben incorporarse ao modelo a través do argumento `prior` da función `lda()` á hora de clasificar novos individuos. Para simplificar, imos traballar cunha prevalencia do 10 %:

```
> p2 <- 0.10
> p1 <- 1-p2
```

Con estas probabilidades *a priori*, o punto de corte para os scores é $c = \log(p_1) - \log(p_2) = 2.197$. Así, se o score é menor que 2.197, entón o individuo será asignado ao grupo 1 e se o score é maior, entón será asignado ao grupo 2.

Para clasificar novos individuos debemos crear un "data frame" que conteña a información das variables de clasificación e aplicar a regra discriminante. Por exemplo, supoñamos que unha muller presenta as seguintes características: *glu*= 110, *bp*= 50 e *bmi*= 21. Para aplicar a regra de clasificación facemos os seguinte:

```
> muller.1 <- data.frame(110,50,21)
> names(muller.1) <- c("glu","bp","bmi")
> predict(lda.diabetes,muller.1,prior=c(p1,p2))

$class
[1] No
```

[7] Ver a web https://www.who.int/diabetes/country-profiles/es/

```
Levels: No Yes

$posterior
         No        Yes
1 0.9850752 0.01492484

$x
      LD1
1 -1.0748
```

A regra devólvenos o score e a probabilidade *a posteriori* de pertencer a cada un dos grupos. Neste caso, a probabilidade de que esta muller non sexa diabética é 0.985. O score -1.075 non supera o punto de corte $c = 2.197$.

Outro exemplo:

```
> muller.2 <- data.frame(200,85,32)
> names(muller.2) <- c("glu","bp","bmi")
> predict(lda.diabetes,muller.2,prior=c(p1,p2))


$class
[1] Yes
Levels: No Yes

$posterior
         No       Yes
1 0.3495799 0.6504201

$x
       LD1
1 2.738501
```

Neste caso, esta muller é clasificada como diabética cunha probabilidade 0.65. O score 2.739 supera o punto de corte $c = 2.197$. □

6.6.2. Erros de clasificación

Gustaríanos que a regra discriminante clasificase a todos os individuos correctamente, pero como sabemos isto non vai ocorrer. De forma natural na práctica aparecerán **falsos positivos** (individuos do grupo 1 que son incorrectamente asignados ao grupo 2) e **falsos negativos** (individuos do grupo 2 que son incorrectamente asignados ao grupo 1). Recordemos as catro situacións que aparecen nos problemas de clasificación (ver máis detalles na sección 2.5):

		grupo verdadeiro	
		1 (Negativo)	**2 (Positivo)**
asignación	**1 (Negativo)**	verdadeiro negativo	falso negativo
	2 (Positivo)	falso positivo	verdadeiro positivo

Como sabemos, para avaliar a calidade dunha regra de clasificación podemos empregar distintas cantidades: sensibilidade, especificidade e probabilidade global de clasificación errónea. A **táboa de confusión** é unha táboa de continxencia 2×2 que recolle a información sobre o número de individuos clasificados correctamente/incorrectamente. Esta táboa pode empregarse para calcular as cantidades antes referidas.

Exemplo. (cont.) Datos DIABETES. Unha vez obtida a regra de clasificación, podemos aplicala aos individuos da mostra para comprobar a súa calidade en termos de erros de clasificación:

```
> clase.predita <-
+   predict(lda.diabetes,diabetes[,c("glu","bp","bmi")],prior=c(p1,p2))$class
> table(clase.predita,type)
```

```
             type
clase.predita  No Yes
          No  129  58
          Yes   3  10
```

Desafortunadamente, a táboa anterior infraestima as probabilidades de erro porque estamos empregando os mesmos datos para construír a regra e para comprobar a súa precisión. En vez de facer iso, deberiamos construír a táboa de confusión mediante **validación cruzada**. Para conseguir isto construímos a regra discriminante de Fisher eliminando o individuo i-ésimo e despois comprobamos se a clasificación é correcta para ese individuo. Isto repítese para todos os individuos da mostra.

En R, para construír a táboa de confusión mediante validación cruzada no noso exemplo facemos o seguinte:

```
> n <- length(type)
> clase.predita <- factor(rep(NA,n),levels=c("No","Yes"))
> for (i in 1:n){
+ lda.vc <- lda(type~glu+bp+bmi,data=diabetes[-i,])
+ clase.predita[i] <-
+   predict(lda.vc,diabetes[i,c("glu","bp","bmi")],prior=c(p1,p2))$class
+ }
> table(clase.predita,type)
```

```
                type
clase.predita  No Yes
          No  127  58
          Yes   5  10
```

Á vista da táboa de confusión obtida mediante validación cruzada, a especificidade, a sensibilidade e probabilidade global de clasificación errónea poden estimarse mediante as seguintes proporcións:

$$\text{especificidade} = \frac{\#\text{ mulleres clasificadas como non diabéticas}}{\#\text{ mulleres non diabéticas}} = \frac{127}{132} = 96.2\,\%$$

$$\text{sensibilidade} = \frac{\#\text{ mulleres clasificadas como diabéticas}}{\#\text{ mulleres diabéticas}} = \frac{10}{68} = 14.7\,\%$$

$$\text{prob. global clas. errónea} = \frac{\#\text{ mulleres mal clasificadas}}{200} = \frac{5+58}{200} = \frac{63}{200} = 31.5\,\%$$

A regra discriminante ten un funcionamento moi bo en termos de especificidade, pero moi malo en termos de sensibilidade. □

Exercicio 6.10. *Como vimos, a regra de clasificación obtida no exemplo anterior ten un bo comportamento en termos de especificidade, pero é moi pobre en termos de sensibilidade. Isto podería corrixirse cambiando o punto de corte, que basicamente está relacionado coas probabilidades* a priori p_1 *e* p_2.

(a) Empregando o código co que se construíu a táboa de confusión por validación cruzada, extrae os scores mediante `predict(...)$x` *e gárdaos nun vector. Con estes scores constrúe a curva ROC empírica para distinguir o grupo de mulleres diabéticas e non diabéticas. Estima a AUC correspondente.*

(b) Estima o índice de Youden e o punto de corte asociado a partir da curva ROC empírica. Canto valen a sensibilidade e a especifidade nese caso?

(c) Con que valores das probalidades p_1 *e* p_2 *se correspondería o punto de corte asociado ao índice de Youden?*

6.6.3. Regra discriminante cadrática

A análise discriminante lineal baséase na hipótese de que as matrices de varianzas-covarianzas son iguais en ambos os grupos. Se esta hipótese se incumpre de forma clara, entón debería empregarse a **regra discriminante cadrática**. A regra cuadrática require a estimación dunha gran cantidade de parámetros. En moitas situacións prácticas, os resultados obtidos non difiren substancialmente dos obtidos mediante a regra discriminante lineal.

En R, a función `qda()` constrúe a regra de discriminación cuadrática.

Exercicio 6.11. *Investiga a función* `qda()`*. Compara o funcionamento práctico da regra cadrática e da regra lineal no conxunto de datos* DIABETES.

Bibliografía

Contidos básicos sobre técnicas descritivas, probabilidade e variables aleatorias:

- Devore, J.L. (2016). *Probabilidad y Estadística para Ingeniría y Ciencias*. Cengage Learning Editores.

- Navidi, W. (2006). *Estadística para Ingenieros y Científicos*. McGraw-Hill Interamericana.

Software R:

- Dalgaard, P. (2008). *Introductory Statistics with R*. Springer.

- R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing (`https://www.R-project.org/`).

Referencias xerais sobre bioestatística:

- King, A., Eckersley, R. (2019). *Statistics for Biomedical Engineers and Scientists. How to Visualize and Analyze Data*. Academic Press.

- Mirás Calvo, M.A., Sánchez Rodríguez, E. (2018). *Técnicas Estadísticas con Hoja de Cálculo y R. Azar y Variabilidad en las Ciencias Naturales*. Servizo de Publicacións da Universidade de Vigo.

- Vidakovic, B. (2017). *Engineering Biostatistics*. Wiley

- Zar, J.H. (2014). *Biostatistical Analysis*. 5th edition. Pearson.

Referencias específicas sobre distintos contidos da materia:

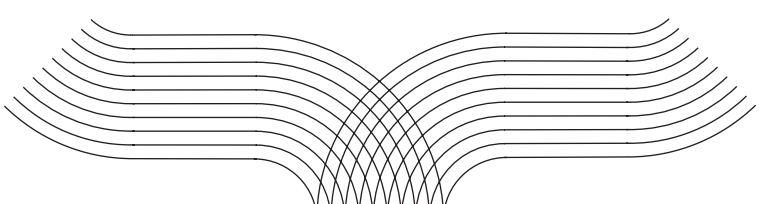
- Everitt, B.S., Hothorn, T. (2013). *Introduction to Applied Multivariate Analysis with R*. Springer. [Técnicas de análise multivariante]

- Lattin, J., Carroll, J.D., Green, P.E. (2003). *Analyzing Multivariate Data*. Thomson. [Técnicas de análise multivariante]

- Pepe, M.S. (2004). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press. [Curvas ROC]

- Sheather, S.J. (2009). *A Modern Approach to Regression with R*. Springer. [Regresión]

- Wasserman, L. (2004). *All of Statistics. A Concise Course in Statistical Inference*. Springer. [Inferencia estatística]



Manuais

Serie de manuais didácticos

Últimas publicacions na colección

Fundamentos de Dereito de Defensa da Competencia (2022)
Julio Costas Comesaña

Sistemas Fluidomecánicos no transporte: Prácticas de simulacións numéricas (2022)
María Concepción Paz Penín, Eduardo Suárez Porto, Jesús Vence Fernández e Adrián Cabarcos Rey.

Álxebra linear: Historia, teoría e práctica (2021)
Ramón González Rodríguez

Manual de programación en Ensamblador: Unha achega teórico-práctica (2021)
Manuel José Fernández Iglesias, Martín Llamas Nistal, Luis Eulogio Anido Rifón, Juan Manuel Santos Gago e Fernando Ariel Mikic Fonte

Elaboración de TFG, TFM e Teses: Claves para o éxito (2021)
Laura Novelle López

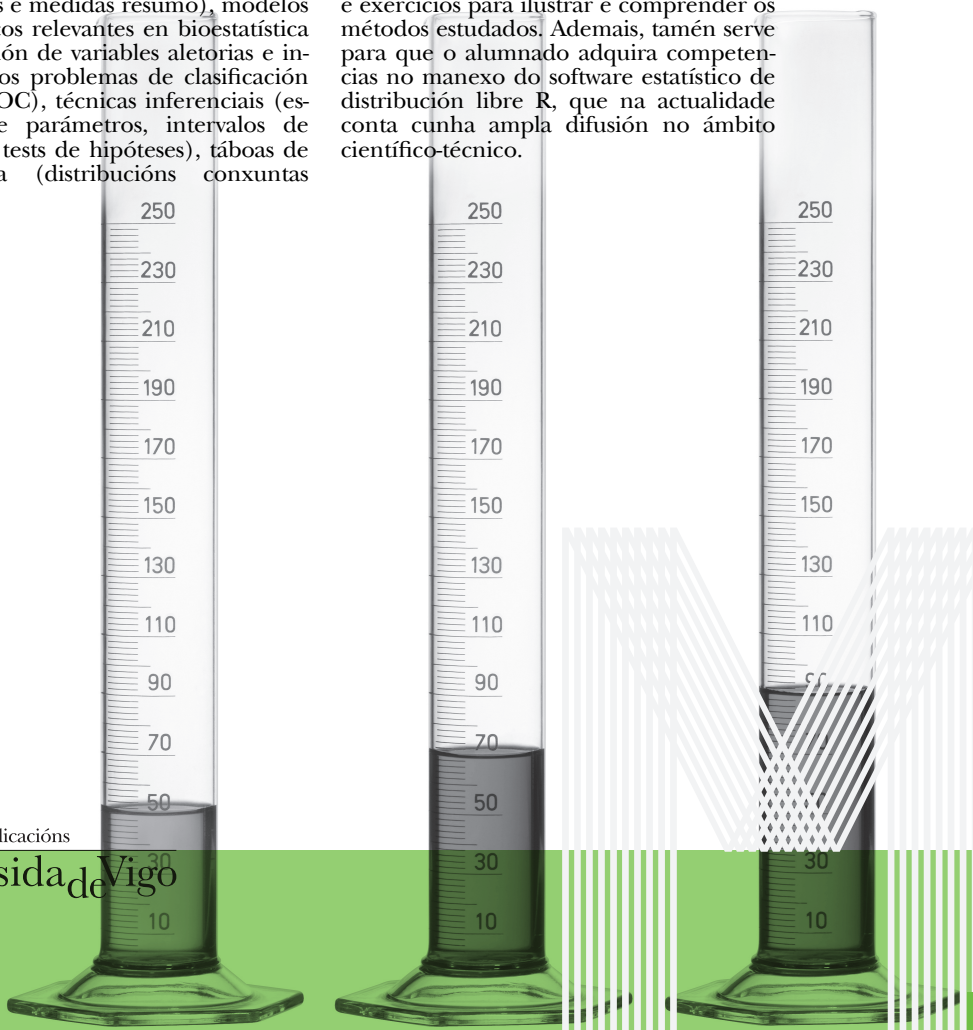
9 7 8 8 4 8 1 5 8 9 6 8 9

Bioestatística para a Enxeñaría Biomédica

Este manual foi concebido como material docente da materia Bioestatística, que se imparte no terceiro curso do Grao en Enxeñaría Biomédica da Escola de Enxeñaría Industrial da Universidade de Vigo. Tamén pode ser empregado noutras materias que aborden temas similares.

Os contidos estrutúranse en seis capítulos: técnicas descritivas (táboas de frecuencias, gráficos e medidas resumo), modelos probabilísticos relevantes en bioestatística (breve revisión de variables aletorias e introdución aos problemas de clasificación e á curva ROC), técnicas inferenciais (estimación de parámetros, intervalos de confianza e tests de hipóteses), táboas de continxencia (distribucións conxuntas e condicionadas, test de independencia e riscos relativos), modelos de regresión (modelo lineal, modelos con variables nominais, regresión loxística) e técnicas de análise multivariante (análise de compoñentes principais, métodos cluster e análise discriminante).

A orientación do manual é eminentemente práctica, incluíndo numerosos exemplos e exercicios para ilustrar e comprender os métodos estudados. Ademais, tamén serve para que o alumnado adquira competencias no manexo do software estatístico de distribución libre R, que na actualidade conta cunha ampla difusión no ámbito científico-técnico.



Servizo de Publicacións
Universidade de Vigo