Anàlisi univariant

PID_00268323

Jordi Mas Elias

Temps mínim de dedicació recomanat: 3 hores

Universitat Oberta de Catalunya

Jordi Mas Elias

L'encàrrec i la creació d'aquest recurs d'aprenentatge UOC han estat coordinats pel professor: Jordi Mas Elias (2019)

Primera edició: setembre 2019

Av. Tibidabo, 39-43, 08035 Barcelona
Realització editorial: FUOC

Índex

Introducció

Tal com el seu nom indica, l'anàlisi univariant significa l'anàlisi d'una sola variable. La naturalesa de la variable que volem estudiar determinarà en bona part els instruments que utilitzarem per a l'anàlisi univariant: si és una variable categòrica farem un tipus de tractament, mentre que si és una variable numèrica en farem un altre. Per aquest motiu serà molt important que tinguem identificada la variable en R amb el vector més apropiat.

En moltes ocasions, l'anàlisi univariant requereix utilitzar més d'una variable. Per exemple, podem voler observar una variable tenint en compte els valors que prenen altres variables del marc de dades. Això pot portar confusió amb la feina de l'anàlisi bivariant, que s'ocupa d'estudiar la relació entre dues variables. Per això distingirem dues tasques crucials de l'anàlisi de dades: la descripció i l'explicació (King i altres, 1994). Aquest mòdul s'ocupa de la descripció. Per tant, la utilització de dues variables s'orienta a descriure una variable i a fer comparacions dels seus valors entre els diversos subgrups d'una altra variable (Babbie, 2013). En cap moment, aquest procés, s'orienta a suggerir si les dues variables en qüestió estan o no associades entre elles, que és feina de l'anàlisi bivariant.

En la primera part del mòdul veurem diferents representacions gràfiques que podem utilitzar en l'anàlisi univariant. El principal objectiu d'una bona visualització és representar de manera clara com es distribueixen els valors en una variable. En gran part, la manera com representem aquests valors estarà determinada per si la variable és de tipus categòric o numèric. En la segona part del mòdul passarem de la visualització a la quantificació. És a dir, buscarem com resumir la distribució de les dades en una variable de manera numèrica, mitjançant un o pocs nombres. Mirarem mesures de freqüència, de centralitat, de dispersió i de localització.

1. Representacions gràfiques

En aquesta secció combinarem funcions dels paquets *dplyr* i *ggplot2* per a representar gràficament la distribució d'una variable del marc de dades *gapminder*. Per a poder fer aquestes representacions haurem d'estar suficientment familiaritzats amb la gramàtica de *ggplot2*, ja que en les pàgines següents aplicarem diverses de les geometries que ofereix el paquet. Segons el tipus de variable que vulguem representar i la manera com la vulguem representar, utilitzarem una de les següents formes de representació gràfica:

- l'histograma,
- el diagrama de línies,
- el diagrama de barres,
- el diagrama de caixes i
- el diagrama de dispersió.

1.1. Histograma

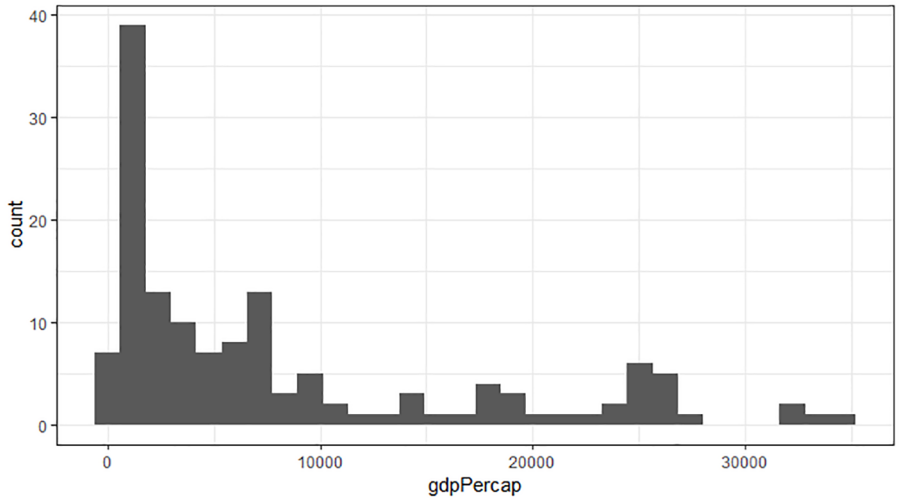
L'histograma ens permet visualitzar amb diverses barres verticals la distribució dels valors d'una variable numèrica. Cada una de les barres representa un interval de valors de la variable i l'alçada de la barra correspon al nombre de casos de cada interval. Cal tenir en compte, doncs, que les barres de l'histograma no representen els valors numèrics originals de la variable, sinó una funció estadística que separa les dades en diferents intervals i les apila per columnes. Amb R, representarem aquesta figura amb la funció `geom_histogram()`.

Pèrdua d'informació amb l'histograma

L'histograma ens permet obtenir una imatge molt nítida de la forma que té una distribució numèrica. La contrapartida, no obstant això, és que perdem informació, ja que els intervals que creem homogeneïtzen el valor de les dades que contenen. Una manera visualment menys nítida però que permet conservar informació és el `geom_dotplot()`, on cada punt representa una observació. Aquesta visualització és útil amb un nombre de casos baix.

En la figura 1 hem creat un histograma per a observar com estan distribuïts els valors de la variable *gdpPercap* l'any 1992. Cada barra horitzontal representa un interval de valors de la variable que indicarem a l'eix de les x a dins dels estètics de *ggplot*. No cal que indiquem l'eix vertical de les y, ja que sempre veurem el recompte de casos de cada columna (de la quantitat de casos en direm freqüències). Per defecte, l'histograma talla la variable en 30 intervals de la mateixa amplada i representa la quantitat de valors que hi ha a cada interval amb l'alçària de les columnes . En el nostre cas, cada barra representa un interval d'uns 1.300 dòlars. Així, en el primer interval, hi trobarem el nombre de països situats entre l'interval de 0 i 1.300 dòlars; en el segon, els situats entre 1.300 i 2.600 dòlars; en el tercer entre 2.600 i 3.900 i així successivament.

Figura 1. Histograma del PIB per càpita




```
gapminder %>%
  filter(year == 1992) %>%
  ggplot(aes(x = gdpPercap)) +
  geom_histogram() +
  theme_bw()
```

En l'histograma que hem creat veiem com l'interval amb més freqüències és el segon, probablement situat entre uns 1.300 i 2.600 dòlars per càpita al mes (de la barra més alta en direm moda, com veurem més endavant). Si mirem en l'eix vertical el recompte de freqüències, veiem com en aquest interval se situen prop de 40 països. A la vora dels 30.000 dòlars, en canvi, no hi ha cap cas.[1]

Quan creeu un histograma la consola us indica que ha creat 30 intervals de dades com a mesura per defecte.[2] Podem canviar l'amplada dels intervals que representa cada columna de dues maneres:

1) Indicant la quantitat d'intervals. Si, en lloc dels 30 per defecte volem observar-ne només 20, en la funció afegirem `geom_histogram(bins = 20)`. Podeu fer diverses proves introduint nombres diferents per veure com canvia la visualització.

(1)Percentatges a l'eix de les *y*: si en lloc de visualitzar el nombre de freqüències a l'eix vertical preferiu visualitzar els percentatges (també anomenat densitat), podeu introduir el següent estètic a dins de la geometria: `geom_histogram(aes(y = ..density..))`.

(2)Seria el mateix que introduir l'argument `bins = 30` a dins de la funció de l'histograma. En la consola us hauria d'especificar el retorn: *`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.*

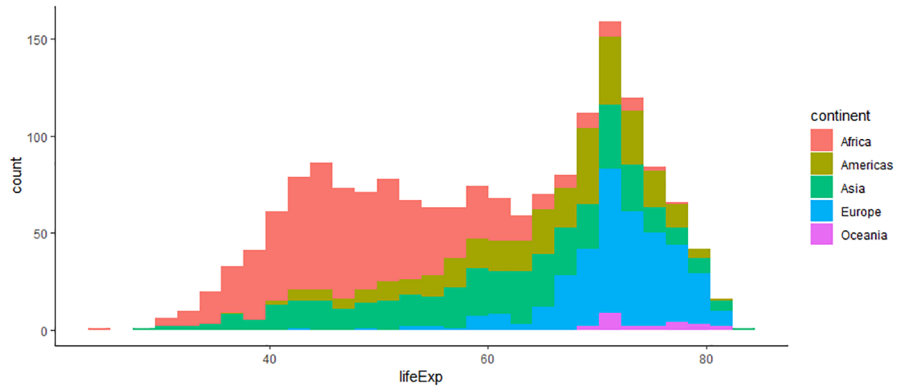
2) Indicant l'amplada dels intervals. Si volem, per exemple, que cada un dels intervals ens representi 3.000 dòlars per càpita, farem servir l'argument *binwidth* de la manera següent: `geom_histogram(binwidth = 3000)`. També podeu fer diferents proves per a veure com canvia la visualització, i també canviar el color de les barres indicant a dins de la geometria el color amb l'atribut *fill*, per exemple, `geom_histogram(fill = "dark blue")`.

Com es calcula l'interval de cada columna

Per a calcular l'amplada de les columnes d'un histograma R mira el rang de les dades. És a dir, el valor màxim (en aquest cas uns 40.000 dòlars) i el valor mínim (que se situa a prop de 0) i divideix el rang en 30 intervals de la mateixa longitud. Per exemple, l'interval de tots els valors de la variable PIB per càpita és: `diff(range(gapminder$gdpPercap)) / 30`. Trobareu més informació sobre aquest procés si mireu la fitxa dels histogrames. Només heu d'entrar `?geom_histogram` a la consola i mirar la descripció dels arguments.

El color també ens pot servir per a introduir una nova variable en el gràfic. Encara que l'histograma s'utilitzi per a veure la distribució d'una variable numèrica, també podem veure a la vegada la distribució d'aquests valors en diverses categories. Observeu la figura 2, on visualitzem la distribució de l'esperança de vida als diversos continents. En el codi, hem omplert les barres de l'histograma utilitzant l'argument `fill = continent` com a estètic per a indicar la variable categòrica.

Figura 2. Histograma amb variable categòrica




```
gapminder %>%
  ggplot(aes(x = lifeExp,
fill = continent)) +
  geom_histogram(position =
"stack") + theme_classic()
```

En el codi, podeu observar com hem afegit un argument nou dins de la geometria: `position = "stack"`. Si traiem aquest argument veureu que no canvia res en la visualització. Això és, perquè, per defecte, la posició de les columnes és *stack* (apilada). Això vol dir que, en cada columna, apilarà les categories una sobre l'altra. Podria, alternativament, posar les categories una al costat de l'altra o superposar-les.

En general, `"stack"` serà la millor posició que tindrem per a visualitzar una variable categòrica dins d'un histograma. Alternativament, podem provar de visualitzar les columnes en altres posicions. Proveu, per exemple, de canviar la posició de l'anterior histograma a `"fill"`. Aquesta opció omple la barra fins a la part superior del gràfic i és útil per veure els percentatge de cada categoria a cada interval. Amb la posició `"dodge"` veurem de costat les barres separades

[3]En els histogrames, la posició `"dodge"` és només recomanable quan tenim pocs intervals. Proveu, per exemple, combinar *dodge* amb un nombre reduït d'intervals amb la funció `geom_histogram(position = "dodge", bins = 10)`.

per categories[3], mentre que amb "identity" les columnes no s'apilen sinó que conserven la seva identitat real. Aquesta darrera, però, es pot observar millor amb un diagrama de densitat.

Diagrama de densitat: l'alternativa a l'histograma

Una altra opció d'anàlisi univariant per a variables numèriques és el diagrama de densitat. Aquest diagrama resumeix les dades dibuixant una àrea que ens permet tenir una visió general de la forma que té la distribució. El diagrama de densitat és útil quan volem afegir informació addicional al gràfic (perquè ens el deixa molt net) i també per a superposar diverses àrees. Proveu els codis següents:

```
gapminder %>% ggplot(aes(x = lifeExp)) + geom_density(fill = "orange", bw = 1)
gapminder %>% ggplot(aes(x = lifeExp, fill = continent)) + geom_density(alpha = 0.4)
gapminder %>% filter(year %in% c(1962, 1977, 1992, 2007)) %>% ggplot(aes(x = lifeExp,
  fill = factor(year))) + geom_density(alpha = 0.4, bw = 5)
gapminder %>% filter(year %in% c(1962, 1977, 1992, 2007)) %>% ggplot(aes(x = lifeExp,
  fill = factor(year))) + geom_density(bw = 5) + facet_wrap(~ year)
```

1) En el primer gràfic hem indicat l'amplada (*bw*, que és una abreviació de *binwidth*).

2) En el segon gràfic hem representat diversos continents i hem donat una transparència de 0,4 als diagrames de densitat. Amb les mateixes dades, podeu provar de fer un histograma especificant dins de la geometria `position = identity`.

3) En el tercer gràfic hem visualitzat el PIB per càpita en quatre anys diferents. Per això hem hagut d'indicar a R que tracti com a categòrica la variable *any* i la converteixi en factor.

4) El darrer gràfic és semblant al tercer, però en aquest cas hem reduït l'amplada dels intervals i hem separat les dades per *facets*. Proveu de fer aquest mateix gràfic amb un histograma.

Fixeu-vos que l'eix vertical sempre veiem la densitat (percentatge) i que l'àrea que cobreix cada distribució està per defecte normalitzada, de manera que totes les àrees cobreixen exactament la mateixa superfície.

Hem vist com hi ha diverses maneres de representar un histograma, per exemple, modificant la quantitat d'intervals de l'eix horitzontal o afegint una variable categòrica i canviant la posició de les barres. No hi ha una regla concreta per a decidir quina és la millor visualització, sinó que la decisió estarà determinada pel que vulguem comunicar, la quantitat de dades disponibles i la distribució d'aquestes. El millor que podem fer és provar diverses variants i utilitzar el sentit comú per a decidir quina és la millor visualització.

1.2. Diagrama de línies

El diagrama de línies s'utilitza normalment per a visualitzar la tendència d'una variable numèrica en el temps. A l'eix horitzontal de les *x*, hi situaríem la variable temporal (en el cas de *gapminder* és l'any, però podrien ser mesos, dies...), mentre que a l'eix vertical de les *y*, hi situaríem la variable numèrica que volem explicar. Amb R, representarem aquesta figura geomètrica amb la funció `geom_line()`. En el codi següent hem fet un sumari de la mitjana de l'esperança de vida mundial agrupada per any. Tot seguit hem demanat un diagrama de línies i hem especificat diversos atributs a la geometria:

- el color,
- la mida,
- el tipus de línia.

Si imprimim el codi observarem que hi ha hagut un augment important de la mitjana de l'esperança de vida per país al llarg del temps, ja que ha passat d'estar per sota dels 50 anys el 1952 a superar els 65 anys poc abans de l'any 2000.

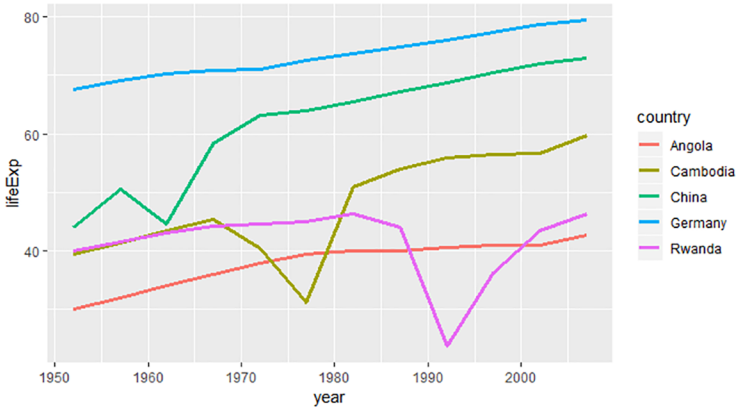
Els tipus de vectors *date*

Els anys s'emmagatzemen com a vectors enters, però si necessitem operar amb franges més curtes de temps ens pot interessar convertir el vector en *date*, un tipus de vector especial que emmagatzema valors numèrics però els visualitza com si fos una data. El paquet *lubridate* (https://lubridate.tidyverse.org/) us ajudarà a fer la conversió. Per a fer una prova, demaneu-li a R l'hora que és amb `Sys.time()` i a continuació proveu de fer dues operacions. Primer, mireu la classe del vector resultant. Segon, convertiu el vector resultant en numèric amb `as.numeric(Sys.time())`. Executeu aquesta operació vàries vegades i descobrireu com R emmagatzema realment aquest vector.

```
gapminder %>%
  group_by(year) %>%
  summarize(mean_lifeExp = mean(lifeExp)) %>%
  ggplot(aes(x = year, y = mean_lifeExp)) +
  geom_line(col = "dark green", size = 1.2, lty = 5)
```

Hi ha centenars de combinacions que podem fer amb els atributs del diagrama línia. Proveu de canviar-los modificant el color, la mida o el tipus de línia a dins de la funció de la geometria. També podem fer que cada línia representi una variable categòrica, mitjançant l'estètic de color. En la figura 3 hem filtrat cinc països de *gapminder* i hem demanat un diagrama de línies per a veure l'evolució temporal de la seva esperança de vida. Hem situat l'any a l'eix horitzontal, l'esperança de vida a l'eix vertical i hem inclòs un estètic addicional, el color, que representa la variable *country*. Cada color representarà l'evolució de l'esperança de vida en un país diferent.

Figura 3. Diagrama de línia amb variable categòrica




```
gapminder %>%
  filter(country %in% c("Cambodia", "Germany", "China", "Angola", "Rwanda")) %>%
  group_by(year, country) %>%
  summarize(mean_lifeExp = mean(lifeExp)) %>%
  ggplot(aes(x = year, y = mean_lifeExp, col = country)) +
  geom_line(size = 1.2)
```

Aquesta figura podem observar més detalls que no amb la que hem generat amb el codi anterior. Fixem-nos que en el casos de la Xina, Cambodja i Rwanda observarem reduccions sobtades de l'esperança de vida, que marquen períodes històrics rellevants en aquests països. El millor estètic per a diferenciar diferents línies és el color, encara que també podem utilitzar el tipus de línia (*lty*) o una combinació de color i tipus de línia.

L'estètic *group* per a agrupar sense estètics

També podem utilitzar *group* en els estètics per a construir línies que representin diferents valors d'una variable categòrica. En aquest cas, però, no ens mostrarà cap diferència visual entre les línies.

1.3. Diagrama de barres

El diagrama de barres no és un histograma. Per a diferenciar-los, direm que el diagrama de barres pren com a punt de partida una variable categòrica, mentre que l'histograma sempre ens mostrarà una variable numèrica. Cada una de les columnes del diagrama de barres representa els diferents valors que pren la variable categòrica i l'alçària de les barres representa uns determinats valors que volem comparar entre categories. Hi ha dues maneres de representar un diagrama de barres.

Diferència entre diagrama de barres i histograma

Visualment, fixeu-vos que hi ha dues diferències molt importants entre l'histograma i el diagrama de barres. En primer lloc, les etiquetes de l'eix horitzontal estan situades a dins de les barres del diagrama de barres perquè representen la categoria; en canvi, en l'histograma estan situades entre les barres perquè representen la divisió entre intervals. En segon lloc, en l'histograma no hi ha separació entre barres perquè representen la continuïtat de la variable. En canvi, com que el diagrama de barres representa variables categòriques (no contínues), hi ha separació entre les barres.

1) La primera és que l'alçària de les barres representi el nombre de freqüències de cada categoria. En aquest cas, utilitzarem la funció geomètrica `geom_bar()` per a representar el diagrama. Fixeu-vos en el codi següent. En aquest cas, només necessitem indicar la variable categòrica que representarem en l'eix de les *x*. En l'eix vertical ens generarà automàticament un recompte de freqüències en cada categoria.

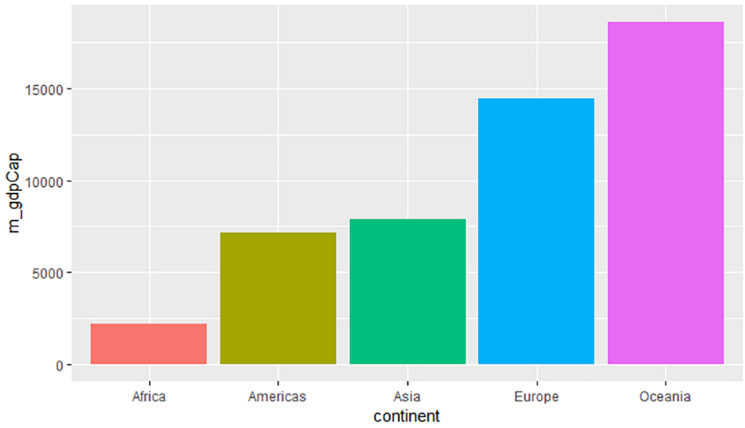
```
ggplot(gapminder, aes(x = continent)) + geom_bar()
```

2) La segona manera de representar un diagrama de barres és situant una variable categòrica a l'eix de les *x* però indicant que l'alçària de les barres ens mostri el sumari d'una altra variable, normalment numèrica, que indicarem en l'eix de les *y*. En aquest cas, utilitzarem la funció geomètrica `geom_col()` per a representar el diagrama . Per a poder aconseguir aquesta visualització les funcions de *dplyr* `group_by()` i `summarize()` ens seran de gran ajuda, com veiem en el codi associat a la figura 4. En primer lloc, hem agrupat les dades per continent i hem demanat un sumari de la mitjana del PIB per càpita a cada continent. D'aquesta manera, en l'eix horitzontal situarem la variable categòrica continent i en l'eix vertical el sumari de la variable numèrica agrupada per continent. Per a visualitzar amb més claredat les diferències entre columnes també hem introduït l'estètic *fill* perquè ens mostri un color per cada continent. Per a evitar redundàncies, hem eliminat la llegenda.

Atributs que diferencien *geom_bar* i *geom_col*

L'única diferència entre aquestes dues maneres de representar diagrames de barres és la funció estadística que tenen associada per defecte. La funció `geom_bar()` fa un recompte de freqüències de cada categoria de *x*, ja que té per defecte l'atribut `stat = "count"`. La funció `geom_col()` ens mostra la identitat de *y*, ja que té per defecte l'atribut `stat = "identity"`.

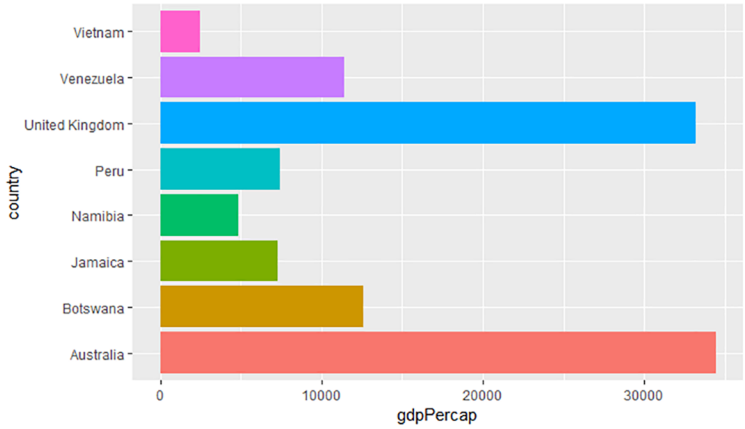
Figura 4. Diagrama de barres




```
gapminder %>%
  group_by(continent) %>%
  summarize(m_gdpCap = mean(gdpPercap)) %>%
  ggplot(aes(x = continent, y = m_gdpCap, fill = continent)) +
  geom_col(show.legend = FALSE)
```

Els diagrames de barres no han de ser necessàriament verticals. Ens podem trobar que vulguem representar un nombre relativament elevat de categories i no puguem distingir bé els seus noms en l'eix horitzontal. També ens podem trobar que tenim poques categories però els noms de cada categoria són molt llargs. En aquests casos, podem rotar els eixos del gràfic amb l'opció `coord_flip()`, que intercanvia de posició *x* i *y*. En la figura 5 hem rotat els eixos del diagrama de barres perquè teníem vuit categories per a mostrar.

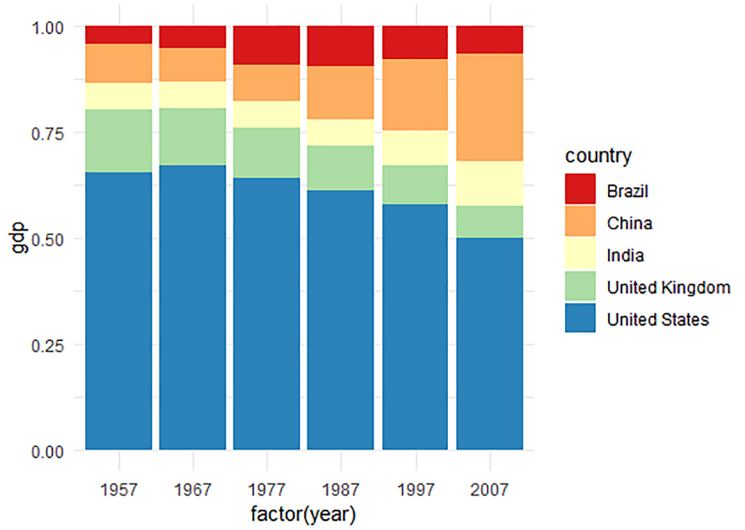
Figura 5. Diagrama de barres horitzontals




```
gapminder %>%
  filter(year == 2007, country %in% c("Australia", "Botswana", "Namibia", "Jamaica",
  "Peru", "United Kingdom", "Venezuela", "Vietnam")) %>%
  arrange(desc(country)) %>%
  ggplot(aes(x = country, y = gdpPercap, fill = country)) +
  geom_col(show.legend = FALSE) +
  coord_flip()
```

En els exemples que hem vist fins ara, l'alçària de les barres del diagrama han representat nombres absoluts. És a dir, les barres ens indicaven un determinat valor d'una variable. Una altra eina que tenim és la possibilitat de fer que les columnes ens representin proporcions en lloc de valors. Així, podrem veure la dimensió relativa de cada valor dins d'una categoria concreta. En la figura 6 hem volgut observar l'evolució, en termes relatius, del PIB entre grans potències mundials: Brasil, Xina, Índia, Regne Unit i Estats Units. En aquest cas, no ens interessa tant com ha crescut el PIB en nombres absoluts sinó com ha variat la relació entre ells. Aquesta visualització ens la permet l'argument `position = "fill"` dins de la geometria.

Figura 6. Diagrama de barres amb proporcions



```
gapminder %>%
  filter(country %in% c("Brazil", "China", "India", "United Kingdom", "United States"),
  year %in% c(1957, 1967, 1977, 1987, 1997, 2007)) %>%
  group_by(year, country) %>%
  summarize(gdp = gdpPercap * pop) %>%
  ggplot(aes(x = factor(year), y = gdp, fill = country)) +
  geom_col(position = "fill") +
  scale_fill_brewer(palette = 9, type = "div") + theme_minimal()
```

Fixeu-vos que aquesta opció ens permet representar tres variables. En l'eix horitzontal hem passat la variable any a factor perquè ens la consideri categòrica, en l'eix vertical ens mostra el PIB per càpita en termes relatius en comptes de termes absoluts, mentre que les barres estan fraccionades per països. Podem observar que la mida relativa del PIB ha disminuït clarament en el cas dels Estats Units i el Regne Unit, i ha augmentat en el cas de la Xina i l'Índia.

Evolució dels diagrames de barres

A l'hora de representar el resum d'una variable numèrica seguint els valors d'una variable categòrica, els diagrames de barres se substitueixen progressivament per altres representacions més sofisticades que poden representar millor aquesta relació. Pensem-ho bé. La barra representa una sola dada, el sumari d'una sèrie de valors de *y*, però amb la barra no podem visualitzar altra informació útil com la distribució dels valors, no sabem el mínim ni el màxim, ni tampoc la quantitat de valors. Per a fer-nos una idea de l'evolució dels diagrames de barres, imprimiu el codi següent, que mostra un gràfic que representa exactament els mateixos valors (i molta més informació) que la figura 4.

```
gapminder %>%
  ggplot(aes(x = continent, y = gdpPercap, col = continent)) +
  geom_point(alpha = 0.2, position = position_jitter(width = 0.2),
             show.legend = FALSE) +
  stat_summary(fun.y = mean, geom = "point", col = "black") +
  theme_light()
```

Quan utilitzem l'estètic *fill* per a representar diversos colors a dins de les barres, podem triar la posició de les columnes tal com també hem pogut fer amb l'histograma. Tant `geom_bar()` com `geom_col()` tenen per defecte l'argument `position = "stack"`, que apila les barres una a sobre l'altra, però podrem escollir tres maneres més de visualitzar la posició de les columnes.

Alternatives per a visualitzar les columnes

Amb `"fill"` ens omple cada columna fins a dalt de tot i ens ensenya les proporcions de cada una. La posició `"dodge"` ubica les barres una al costat de l'altra (va bé quan tenim les variables *x* o *fill* amb poques categories, per exemple una variable binària), mentre que `"identity"` ens mostra l'alçària real de cada barra sense apilar-les. Aquesta darrera opció es visualitza millor amb transparències (atribut `alpha`).

1.4. Diagrama de caixes

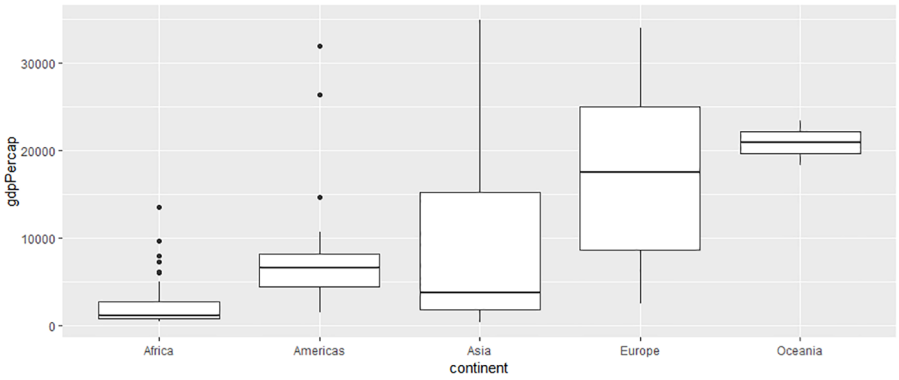
El diagrama de caixes ens ajuda a observar i comparar com una variable numèrica està distribuïda segons els valors d'una variable categòrica. A diferència del diagrama de barres, que només ens permet visualitzar un sol valor per mitjà de l'alçària de cada barra, la utilitat del diagrama de caixes és principalment que ens permet visualitzar alguns estadístics descriptius, com la mediana de cada distribució o els casos extrems. Amb R, representarem aquesta figura geomètrica amb la funció `geom_boxplot()`. En el codi següent hem demanat un diagrama de caixes per a veure com estaven distribuïts l'any 1992 els valors de la variable *gdpPercap* en diversos continents.

```
gapminder %>% filter(year == 1992) %>%
```

```
ggplot(aes(x = continent, y = gdpPercap)) + geom_boxplot()
```

Aquest codi imprimeix un gràfic que dona diversos estadístics descriptius de la manera com la variable numèrica està distribuïda segons els valors de la variable categòrica que hem representat a l'eix de les *x*. És per això que el diagrama de caixes és un diagrama bastant sofisticat, que conté molts elements, i requereix pràctica per a poder ser interpretat adequadament. En la figura 7 representem el diagrama de caixes que genera el codi anterior.

Figura 7. Diagrama de caixes



Observem el diagrama pas a pas. En primer lloc, la línia més gruixuda que hi ha a dins les caixes és la mediana de la distribució (el nombre que separa la distribució ordenada en dos parts iguals). Veiem que l'únic continent amb una mediana superior als 20.000 dòlars és Oceania. Això significa que si ordenem per PIB per càpita tots els països d'Oceania, de més gran a més petit, el que està a la meitat de la distribució supera lleugerament els 20.000 dòlars.[4] Europa se situa aproximadament en els 17.000 mentre que la resta de continents no superen els 10.000 dòlars de mediana. Les parts superior i inferior de cada caixa representen el 75 i el 25 per cent de la distribució. És a dir, si tinguéssim 100 països a cada continent, la part superior de la caixa ens indica el 25è país més ric i la part inferior el 75è. Les caixes d'Àfrica, Amèrica i Oceania són molt petites, la qual cosa significa que hi ha poca diferència de riquesa entre el país que fa el 25 i el país que fa el 75 per cent de la distribució ordenada: a l'Àfrica gairebé tots són molt pobres, a Amèrica gairebé tots són més aviat pobres, mentre que a Oceania tots són bastant rics. En canvi, les caixes d'Àsia i Europa són molt més llargues, la qual cosa significa que tant a Àsia com a Europa la distribució és més dispersa i hi ha tant països rics com països pobres.

[4] Val a dir que Oceania només té dos països a la mostra, Austràlia i Nova Zelanda. Per tant, la xifra que veiem és la que està entremig del PIB per càpita de cada país.

Fora de les caixes, hi trobem dues figures geomètriques: línies i punts. Les línies que surten de les caixes cap amunt i cap avall ens indiquen l'interval on es troben els països per sota del 25 per cent i per sobre del 75 per cent. Quan hi ha casos extrems, que són en aquest cas països molt allunyats de la resta de la distribució, es representen amb punts.

Línies i casos extrems

En el cas de *ggplot2*, les línies fan com a molt 1,5 de la mida de la caixa. A partir d'aquesta distància, la llargada de la línia s'estableix buscant el darrer valor per sota d'1,5. Els punts extrems són tots aquells valors que superen la llargada de la línia.

Fixeu-vos que en aquest diagrama de caixes hem representat exactament les mateixes dades que en l'histograma de la figura 1 i semblant a les dades de la figura 4. Tots tenen els seus avantatges i inconvenients, i per a seleccionar el diagrama més adequat hem de pensar quin és el nostre propòsit i com es visualitzaran millor les dades. L'histograma i el diagrama de densitat ens permeten fer una ullada general a la forma que té la distribució d'una variable. Si la distribució és bimodal, cosa que significa que tenim dos intervals amb concentracions altes de valors, aquests gràfics ens permetran distingir amb claredat aquestes dues puntes. En canvi, aquesta observació no la podrem fer amb un diagrama de barres ni amb el diagrama de caixes.

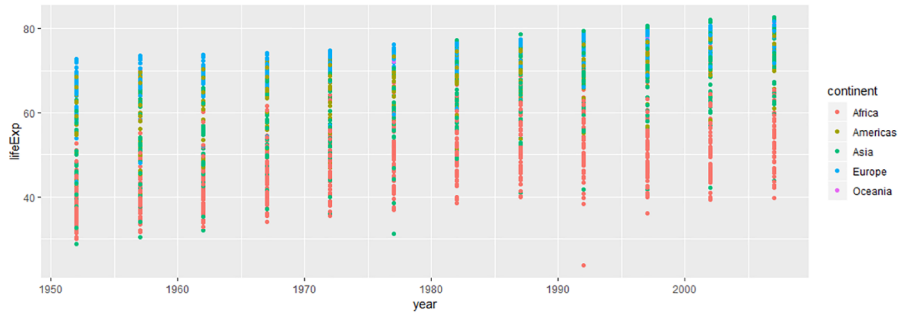
L'avantatge dels diagrames de caixes i de barres és que podem veure la distribució en orientació horitzontal o vertical afegint la capa de coordenades `coord_flip()`. El diagrama de caixes permet fer més comparacions, ja que descriu diverses propietats de les dades, com els valors que es troben en el 25, el 50 i el 75 per cent de la distribució, l'amplitud de la distribució i també els casos extrems.

1.5. Diagrama de dispersió

El diagrama de dispersió és una figura de dues dimensions que representa amb punts la relació entre dues variables numèriques. Amb R, representarem aquesta figura geomètrica amb la funció `geom_point()`. Més enllà del que ja hàgim pogut veure fins ara sobre el diagrama de dispersió, en aquest apartat destacarem la manera com podem solucionar gràficament un dels problemes habituals d'aquest tipus de gràfic: la sobrerepresentació. Aquest fenomen és habitual en variables ordinals i en variables numèriques discretes, quan els punts representen una quantitat de valors limitada i es mouen en un rang més aviat petit. Això fa que els punts es puguin encavalcar entre ells. La sobrerepresentació també és habitual en el cas de variables nominals quan tenim una gran quantitat de punts per a representar. Visualment, els punts s'encavalquen i no podem distingir bé els uns dels altres.

Per posar un exemple, pensem en les variables *any* i *esperança de vida*. Són dues variables numèriques que haurien de poder ser representades amb un diagrama de dispersió. No obstant això, observeu què passa a la figura 8 quan intentem representar aquestes dues variables del marc de dades *gapminder*. És evident que tenim un problema important de sobrerepresentació. La variable *any* admet pocs valors i està tipificada com a vector enter, mentre que la variable *esperança de vida* pot tenir poca variació entre els seus valors. Amb tot, els punts s'encavalquen.

Figura 8. Diagrama de dispersió amb sobrerepresentació



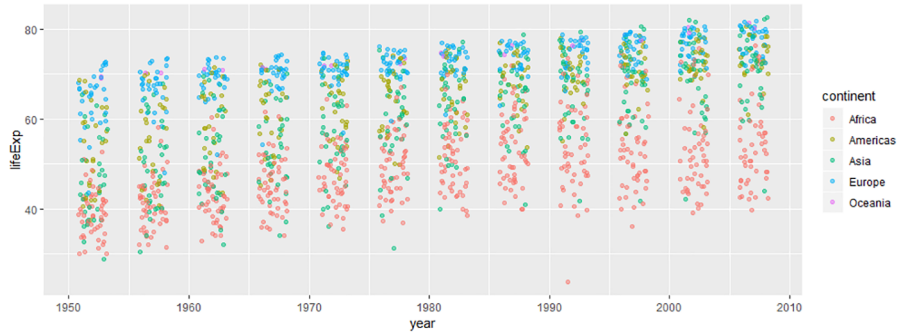
Si ens fixem en els colors que representen els continents, sospitem que els països africans tenen menys esperança de vida i els europeus la tenen més elevada, però ens agradaria veure les dades de manera més nítida per a arribar a conclusions més concretes. Una bona solució és substituir `geom_point()` per `geom_jitter()`[5], com veiem en el codi següent, que afegirà un «soroll» o moviment aleatori a la posició dels punts.

[5]Aquesta funció fa el mateix que `geom_point(position = "jitter")`. Per defecte, la posició de `geom_point()` és `"identity"`.

```
gapminder %>% ggplot(aes(x = year, y = lifeExp, col = continent)) + geom_jitter(width = 1.2,
height = 0.1, alpha = 0.5)
```

El *jittering* separa els punts entre si mitjançant la introducció d'un moviment aleatori horitzontal i vertical que ens permet observar millor el diagrama de dispersió. Si no afegim res a dins el parèntesi de `geom_jitter()` tindrem la mateixa aleatorietat en vertical que en horitzontal. No obstant això, hem volgut especificar amb l'argument `width` que el moviment aleatori en horitzontal sigui elevat, i en canvi amb l'argument `height` hem ordenat que el moviment aleatori en vertical sigui mínim. A la figura 9 veiem el resultat.

Figura 9. Diagrama de dispersió amb *jittering*



Un altre mètode que hem fet servir per a reduir la sobrerepresentació és l'argument `alpha`, que introdueix transparència als punts.

1.6. Retocs finals

Hi ha milers de retocs que es poden fer en un gràfic mitjançant *ggplot2* per a millorar la visualització. Aquí us explicarem dos retocs:

- posar títols al gràfic i als eixos,

- eliminar casos extrems.

Per a canviar els títols ho podem fer de tres maneres:

1) La primera consisteix a especificar-ho en les escales de *x* i de *y*, com veiem en la primera línia del codi següent. En l'exemple, si treballem amb una variable numèrica en l'eix de les *x*, especificarem l'argument *name* a la funció `scale_x_continuous()`.[6]

[6]Sempre posarem *scale*, el nom de l'estètic i si la variable és *continuous* o *discrete*. Això també implica estètics com el color. Per exemple, en el cas d'un estètic categòric de color introduirem `scale_color_discrete()`.

2) Si no necessitem modificar cap escala ens serà més fàcil i ràpid introduir arguments a dins de la funció `labs()` indicant el títol general amb `title`, l'eix horitzontal amb *x* i l'eix vertical amb *y*. Si tenim una llegenda que mostra els valors d'un estètic addicional només caldrà introduir el nom de l'estètic (*size*, *color*, *fill*...) seguit del títol.

3) Els eixos horitzontal i vertical, com també el títol general del gràfic, es poden introduir per mitjà de funcions separades: `ggtitle()` permet introduir el títol, `xlab()` permet introduir el títol de l'eix horitzontal i `ylab()` el títol de l'eix vertical.[7]

[7]Si demaneu ajuda de la funció `theme()` obtindreu molta informació sobre tot el que podem arribar a fer amb els títols. També podeu consultar més informació en aquest enllaç (http://www.sthda.com/english/wiki/ggplot2-title-main-axis-and-legend-titles).

```
scale_x_continuous(name = "Títol eix x")
labs(title = "Títol", x = "Títol eix x", y = "Títol eix y")
ggtitle("Títol") + xlab("Títol eix x") + ylab("Títol eix y")
```

El segon retoc final per a generar una millor visualització del gràfic és l'eliminació de casos extrems (*outliers*). Aquests casos deformen les dades i impedeixen observar amb claredat el centre de la distribució. Per exemple, si volem mirar la relació entre el PIB per càpita i l'esperança de vida el 1952, hi ha un cas extrem que ens deforma completament el gràfic.

```
gapminder %>% filter(year == 1952) %>% ggplot(aes(x = gdpPercap, y = lifeExp)) + geom_point()
```

Veiem que només tenim un únic cas superior als 20.000 dòlars per càpita i que s'allunya molt de la resta de la distribució, ja que supera un PIB per càpita de 100.000 dòlars. Per a eliminar aquest cas extrem podem reduir els límits de l'eix de les *x* simplement introduint la funció `xlim(c(0, 20000))`. Una altra opció és introduir un nou argument dins de la funció `filter()` per a indicar que filtri els casos on `gdpPercap < 25000`. Aquestes dues opcions eliminaran el cas extrem i reproduiran la resta de dades.

Podem voler, no obstant això, conservar aquest cas extrem i generar un nou marc de dades en el qual el tinguem assenyalat. En aquest cas, podem utilitzar una combinació de les funcions `filter()` i `mutate()`, com veiem a continuació. Amb la primera fórmula creem el nou marc de dades `gap52` filtrat per

1952 que conté la variable lògica extrem, que és TRUE per als casos amb PIB superior per càpita als 25.000 dòlars i FALSE per a la resta. Tot seguit, fem dues operacions:

1) En la primera, filtrem pel cas extrem, que ens retornarà tots els valors que siguin TRUE en la columna extrem. Aquí podem comprovar que Kuwait és el cas extrem de les nostres dades.

2) En la segona operació filtrem excloent el cas extrem (amb el símbol `!`) i demanem un *ggplot2* de les dades filtrades.

```
gap52 <- gapminder %>% filter(year == 1952) %>% mutate(extrem = gdpPercap > 25000)
> gap52 %>% filter(extrem)
# A tibble: 1 x 6
  country continent  year lifeExp    pop gdpPercap
  <fct>   <fct>     <int>   <dbl>  <int>     <dbl>
1 Kuwait  Asia       1952    55.6 160000   108382.
gap52 %>% filter(!extrem) %>% ggplot(aes(x = gdpPercap, y = lifeExp)) + geom_point()
```

Hi ha altres maneres d'eliminar o filtrar casos extrems. La majoria requeriran la funció `filter()` o bé una mica d'imaginació amb les eines dels paquets d'R que ja coneixem.

2. Mesures d'anàlisi univariant

Fins ara hem vist diverses maneres de representar per mitjà de visualitzacions gràfiques la informació que conté una variable. En aquest apartat buscarem, en lloc d'una visualització, un o pocs nombres que ens puguin sintetitzar la mateixa informació. En altres paraules, el que volem és quantificar la manera com estan distribuïts els valors a la variable. L'exercici de trobar mesures per a sintetitzar distribucions per mitjà de xifres concretes ens permetrà obtenir informació comparable, de manera que podrem contrastar diferents variables entre elles. La manera com sintetitzarem la informació, però, variarà segons el tipus de variable. En el cas de les variables categòriques no podem fer massa més que observar les freqüències de cada valor. En el cas, però, de les variables numèriques, podem resumir la informació mitjançant les diverses mesures que indiquin la centralitat i la dispersió de la distribució i la localització de determinats valors. Bona part d'aquesta feina la podem fer directament amb la funció `summary()`, que retornarà les freqüències d'una variable categòrica i algunes mesures de centralitat i localització d'una variable numèrica.

2.1. Taula de freqüències

La taula de freqüències és una de les modalitats més típiques per a representar variables categòriques, com és el cas de la variable *continent* del marc de dades *gapminder*. Normalment a la taula mostrem el nombre de vegades que es repeteix cada valor categòric en la variable i el seu percentatge sobre el total. Les freqüències les podem obtenir amb les funcions `table()` o `summary()`[8] i el percentatge de freqüències amb la funció `prop.table()`.

[8] La distinció principal entre `table()` i `summary()` és que la primera genera una taula mentre que la segona genera un vector enter.

```
table(gapminder$continent)
prop.table(table(gapminder$continent))
```

Si treballem amb variables categòriques nominals en tindrem prou amb les dues funcions anteriors per a visualitzar les dades, però si es tracta de variables ordinals també ens pot interessar veure l'acumulat per a cada categoria. És a dir, voldrem observar com s'acumulen els valors i els percentatges si els afegim als valors o percentatges de les files anteriors. En el codi següent veiem com crear una taula amb freqüències i percentatges acumulats a partir de la variable `income` del marc de dades `wb`.

Nota

Podeu trobar el codi del marc de dades `wb` a l'annex d'aquest mòdul.

```
> data.frame(categories = rev(unique(wb$income)),
             freq = rev(summary(wb$income))) %>%
     mutate(freq_cum = cumsum(freq),
            percent = round(prop.table(freq), 2),
            percent_cum = cumsum(percent))
    categories freq freq_cum percent percent_cum
```

```
1          low    3        3    0.20        0.20
2 lower-middle    8       11    0.53        0.73
3 upper-middle    3       14    0.20        0.93
4         high    1       15    0.07        1.00
```

La segona columna de la taula que hem generat mostra les freqüències que es repeteixen per a cada categoria i la tercera columna mostra l'acumulat de freqüències, de manera que, per exemple, a la fila «lower-middle» veiem l'acumulat de «lower-middle» i «low», a la fila «upper-middle» veiem l'acumulat per «lower-middle», «low» i «upper-middle», i així successivament. En les darreres dues columnes fem el mateix procediment: primer ensenyem el percentatge de freqüències sobre el total en cada categoria i a continuació el percentatge acumulat.

Si volem fer una taula de freqüències amb una altra variable, només caldrà substituir les dues vegades que apareix `wb$income` per la variable categòrica que vulguem reproduir.

2.2. Mesures de centralitat

Per a sintetitzar una distribució en un sol valor, una manera és buscar on es troba ubicat el centre de la distribució. Fixeu-vos en el vector següent. L'objecte `ev` representa una distribució que ens indica l'esperança de vida, en nombres absoluts, recollida en onze països d'una regió del món. Si volem transformar aquestes dades en una sola dada numèrica que ens resumeixi on es troba el centre d'aquesta distribució, com ho faríem? Quin número creieu que pot resumir millor aquesta distribució?

```
ev <- c(65, 67, 68, 73, 74, 77, 80, 80, 80, 82, 83)
```

La resposta és que dependrà de la pregunta concreta que fem, ja que no hi ha una sola manera de sintetitzar aquesta distribució per a trobar-ne el centre. Principalment, hi ha tres maneres de resumir aquestes dades:

1) La primera és buscar el valor que es repeteix més vegades. D'això se'n diu la moda i, en el nostre cas, seria el número 80, que es repeteix tres cops.

2) La segona és buscar el valor que es troba en el centre de la distribució ordenada. D'això se'n diu la mediana, que en el vector `ev` és el número 77, ja que és el valor que queda just al mig de la distribució si l'ordenem de més petit a més gran.

3) La tercera manera és amb la mitjana, que s'obté sumant tots els valors de la distribució i dividint-los pel nombre de casos.

2.2.1. Mitjana

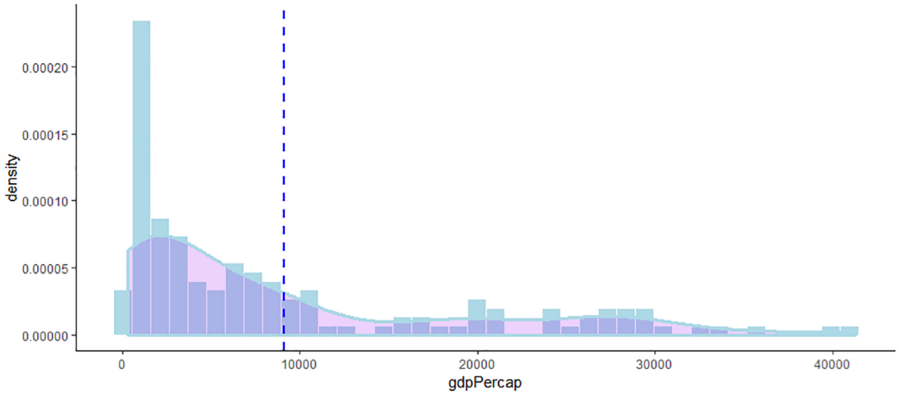
La mitjana és una de les mesures de tendència central més utilitzades per a resumir distribucions. Com veiem a continuació, per a obtenir la mitjana hem de sumar tots els valors d'un vector i dividir el resultat entre el nombre total de valors, que són 11.

```
mitjana <- sum(c(65, 67, 68, 73, 74, 77, 80, 80, 80, 82, 83)) / 11
```

El que fa la mitjana és dir-nos quin seria el valor que obtindríem si anivellés-sim el pes de cada valor de la distribució entre els seus casos. Una manera molt il·lustrativa per a entendre què representa la mitjana és mitjançant la distribució de la riquesa d'un país. En qualsevol país, la riquesa acostuma a estar distribuïda de manera desigual. Hi ha individus que tenen més diners i n'hi ha d'altres que en tenen menys. El que ens diu la mitjana és quants diners li tocaria a cada habitant si agaféssim els diners de tota la població i els repartís-sim a parts iguals entre ells. En totes les distribucions la mitjana fa exactament això: suma tots els valors i els divideix entre el nombre de freqüències.

Fixem-nos en la figura 10, on hem superposat un diagrama de densitat i un histograma que il·lustren com estava distribuït el PIB per càpita en els diferents països del món el 1997. Amb línia blava hem indicat on es troba la mitjana.

Figura 10. Mitjana del PIB per càpita mundial per països



Si agafem els diners dels països més rics i els repartim entre tots els països de manera que els diners quedin distribuïts equitativament veurem que tots els països tindrien 9.090 dòlars per habitant. Podem pensar en la mitjana com el punt que equilibra les dades. Aquest valor divideix el diagrama de densitat en dues àrees aproximadament iguals, de manera que l'àrea de densitat de l'esquerra és igual a l'àrea de densitat de la dreta. La mitjana es calcula amb la funció `mean()`[9]. En el codi següent hem creat l'objecte `gap97_mean` on hem especificat la mitjana amb una línia vertical `geom_vline()`. Imprimiu l'objecte després de crear-lo.

[9] Alternativament, també es pot calcular amb `sum()` / `length()`, on primer sumem els valors d'un vector i després dividim el resultat pel recompte d'observacions.

```
gap97_mean <- gapminder %>% filter(year == 1997) %>% ggplot(aes(x = gdpPercap)) +
geom_histogram(aes(y=..density..), colour= "white", fill= "lightblue", bins = 40) +
```

```
geom_density(col = "lightblue", size = 1.2, fill = "purple", alpha = 0.2) +
geom_vline(aes(xintercept = mean(gdpPercap)), color = "blue", linetype = "dashed", size = 1) +
theme_classic()
gap97_mean
```

Les dues maneres més habituals de calcular la mitjana és a dins de la funció `summarize()` o bé mitjançant un vector, com s'indica en el codi següent. En primer lloc, hem creat un vector on filtrem les observacions per l'any 1992 i a continuació hem demanat la mitjana del vector.

```
gdp_cap1992 <- filter(gapminder, year == 1992)
mean(gdp_cap1992$gdpPercap, trim = 0, na.rm = FALSE)
```

En aquest codi hem d'especificar dos arguments addicionals que té per defecte la funció de la mitjana. Amb l'argument *trim* podem eliminar casos extrems de la distribució. Haurem d'especificar un valor de 0 a 0,5, que indicarà el percentatge d'observacions que eliminarà de cada extrem de la distribució. L'argument *na.rm* és extremadament important, ja que elimina els NA de la distribució. El marc de dades *gapminder* no té dades perdudes, però serà freqüent trobar-nos-en en la majoria de marcs de dades. És per això que haurem de canviar l'argument a TRUE perquè ens mostri la mitjana.

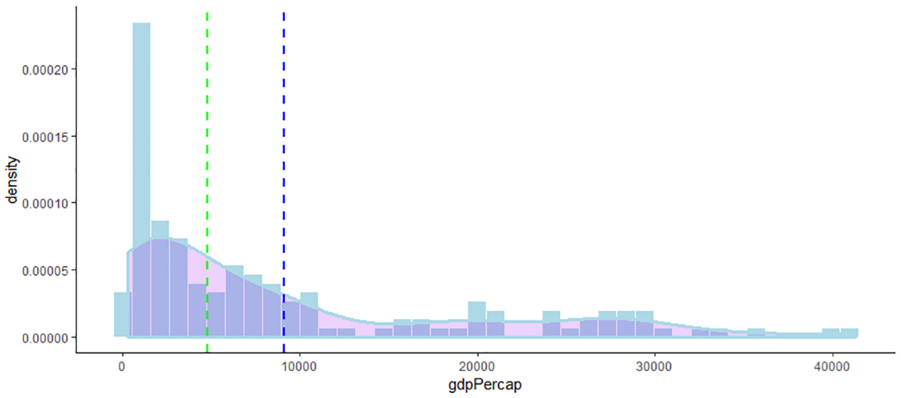
2.2.2. Mediana

La mediana ens indica quin valor té l'observació que es troba exactament al mig de la distribució. Dit d'una altra manera: si tenim una distribució amb 144 països, la mediana ens ordena els valors de més gran a més petit, i ens indica el valor del país que fa 72, que és el que es troba al mig de la distribució (l'obtenim dividint 144 entre dos). Com veiem en el codi següent, hem utilitzat l'objecte `gap97_mean`, ja creat, que conté tota la informació sobre la geometria anterior per afegir-li una línia nova amb informació sobre la mediana.

```
gap97_mean + geom_vline(aes(xintercept = median(gdpPercap)), color="green",
linetype = "dashed", size = 1)
```

En la figura 11 hem il·lustrat el resultat del codi on veiem una nova línia vertical, de color verd, que representa la mediana.

Figura 11. Mediana i mitjana del PIB per càpita mundial per països



És interessant veure com tots dos centres divergeixen clarament. Si calculem la mediana veurem que es troba en els 4.782 dòlars per habitant, mentre que la mitjana es trobava en els 9.090 dòlars. Com que la gran majoria d'observacions ocorren en els valors més baixos de l'histograma, és lògic que la mediana se situï més esbiaixada a l'esquerra de la distribució.

Altres mesures de localització

La mediana és una mesura tant de centralitat com de localització, ja que ens indica el valor per sota del qual tenim ubicat el 50 per cent d'observacions d'una distribució ordenada. Per a referir-nos a altres localitzacions de la distribució diferents a la mediana utilitzarem termes com:

- els percentils (si comptem per unitats de l'1 al 100),
- els decils (si comptem de 10 en 10),
- els quintils (de 20 en 20) i
- els quarts (de 25 en 25).

Localitzar el valor és fàcil amb la funció `quantile()`, on només hem d'indicar el nom de l'objecte i la posició del valor que volem localitzar en l'escala de 0 a 1. Per exemple, si volem trobar el percentil 40 (que també seria el quart decil o el segon quintil) marcarem `quantile(vector, 0.4)`. Podem indicar tantes localitzacions com vulguem si utilitzem el concatenat. Si volem saber tots els quintils introduirem `quantile(vector, c(0, 0.2, 0.4, 0.6, 0.8, 1)`. El zero ens retornarà el valor més baix i l'1 ens retornarà el valor més alt de la distribució.

2.2.3. Moda

L'observació més repetida en una distribució és la moda. Poden ser exemples de moda la religió o el partit polític dominant en un país. La moda és més fàcil de trobar en una variable categòrica, ja que serà senzillament la categoria amb més freqüències. Si imprimim el sumari o demanem una taula de la variable continent de *gapminder* trobarem fàcilment la moda. Veiem clarament que la moda és Àfrica, la distribució amb més freqüències: 624.

```
> summary(gapminder$continent)
> table(gapminder$continent)
  Africa Americas     Asia   Europe  Oceania
     624      300      396      360       24
```

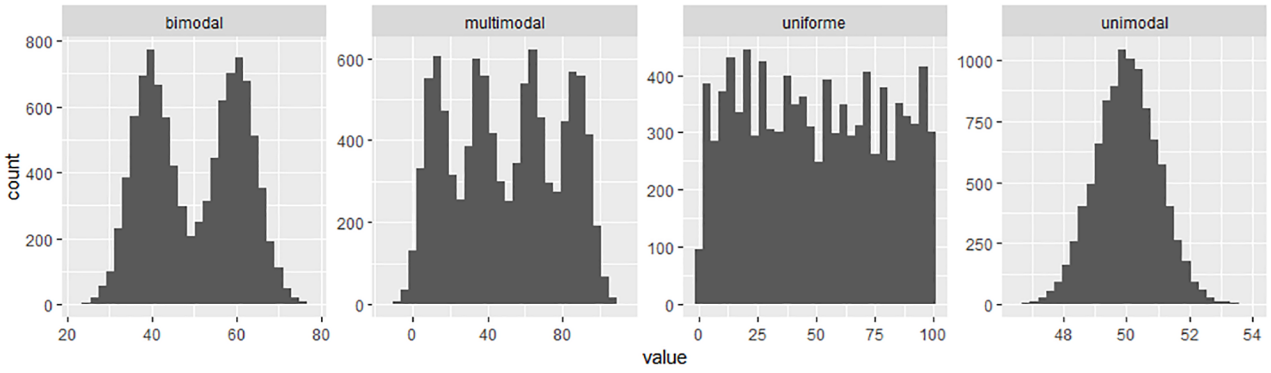
Trobar la moda en una variable numèrica, en canvi, no és un exercici tan evident. Hem de pensar que aquestes variables poden adoptar un nombre infinit de valors i per tant és poc probable trobar algun nombre repetit i, si es dona el cas, no ens aportarà informació més enllà de la casualitat que es repeteixin. La moda només pot ser fàcil de trobar en el cas que la variable en qüestió sigui discreta, com l'edat d'una població. En aquest cas, si tenim un nombre suficientment gran d'observacions en relació amb els valors que pot adoptar la variable podrem obtenir la moda sense problemes.

La discrecionalitat de la moda

Aquest procediment té un punt de discrecionalitat, ja que podem canviar els intervals de l'histograma o la sensibilitat del diagrama de densitat. Això pot fer que la moda que establim visualment varïi en funció dels paràmetres que establim. Feu la prova: agafeu el codi que hem utilitzat per a crear la figura 10 i canvieu els intervals (*binwidth*) a 20. Veureu com ara la moda està més repartida entre dos intervals.

En variables numèriques contínues, no obstant això, ho tenim més complicat. Per a obtenir informació sobre la nostra anàlisi serà més útil, en lloc de trobar un valor concret, saber on estan concentrats els valors de la variable. Aquest exercici el podem fer fàcilment amb un histograma o un diagrama de densitat, en el qual podem visualitzar quin és el punt més alt de la distribució . En la figura 12 podem observar els diferents tipus de distribucions que poden donar lloc a diferents tipus de modes.

Figura 12. Tipus de moda



En direm distribució unimodal quan només hi ha una punta, bimodal quan n'hi ha dues, multimodal quan hi ha més de dues puntes i uniforme quan no s'aprecia cap punta concreta.

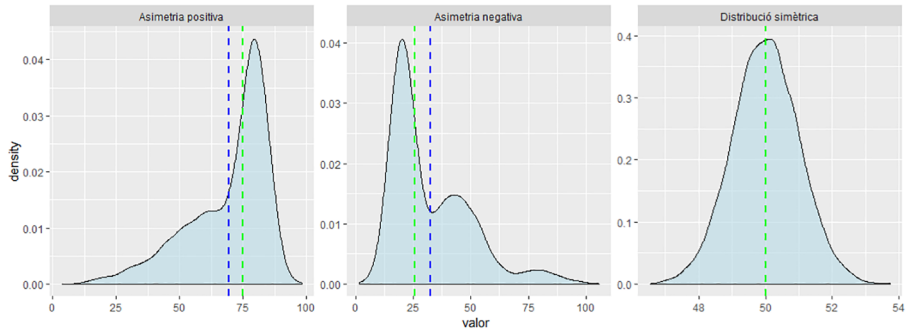
2.2.4. Tipus de transformacions

A mesura que ens anem familiaritzant amb diferents tipus de distribucions, veurem que podem intuir la forma que tenen sense visualitzar-les. En tindrem prou a conèixer la moda, la mitjana i la mediana. Si la mitjana és inferior a la mediana, segurament ens trobarem amb un gràfic semblant al diagrama de densitat de l'esquerra de la figura 13 i direm que és una distribució asimètrica positiva.[10] Com veiem, la cua del gràfic és a l'esquerra i els valors estan esbiaixats a la dreta de la distribució. Si la mitjana, en canvi, és superior a la mediana possiblement ens trobarem amb el gràfic del mig i direm que és una distribució asimètrica negativa. La cua del gràfic, en aquest cas, es troba a la dreta de la distribució i els valors estan esbiaixats a l'esquerra de la distribució. En darrer

[10]Amb el diagrama de densitat es perd informació en relació amb l'histograma, ja que no tenim la informació separada per intervals. Tot i així, és útil quan volem afegir informació a sobre del gràfic, ja que visualment no quedarà sobrecarregat.

lloc, si la moda, la mitjana i la mediana coincideixen més o menys voldrà dir que tenim una distribució simètrica. En aquest cas, la cua a un cantó i a l'altre és semblant i no hi ha biaix.

Figura 13. Distribucions simètriques i asimètriques



Les distribucions dels ingressos en un país acostumen a tenir una asimetria negativa. Això vol dir que la majoria de casos estan situats a l'esquerra de la distribució i la mediana és inferior a la mitjana. En altres paraules, l'individu situat a la meitat de la distribució (la mediana) no té tants diners com el que li tocaria a cada individu si repartíssim tots els diners a parts iguals entre el conjunt de població (la mitjana).

Sensibilitat de la mitjana

La mitjana és sensible a distribucions esbiaixades i valors extrems, de manera que, en aquests casos, pot ser un mètode més adequat per a mesurar la centralitat. Una manera d'evitar que els valors extrems influeixin excessivament en el càlcul de la mitjana és amb l'argument *trim*, que elimina la proporció que indiquem de cada extrem. Per exemple, `mean(variable1, trim = 0.01)` eliminarà el darrer 1 per cent de cada extrem.

Sovint, per a visualitzar millor aquestes distribucions i fer proves de significació estadística, s'acostuma a construir una versió transformada de la variable. En els casos de distribucions amb asimetria negativa, s'aplica l'escala logarítmica amb les funcions `log()`, `log2()` o `log10()`, mentre que en distribucions amb asimetria positiva es pot aplicar l'exponencial amb la funció `exp()`.

2.3. Mesures de dispersió

Les mesures de dispersió ens indiquen la separació o dispersió dels valors en una distribució numèrica. En general, si els valors estan molt concentrats al voltant del centre indicaran un valor baix, mentre que si hi estan molt separats indicaran un valor alt. Hi ha tres maneres principals de mirar la dispersió:

- el rang,
- el rang interquarlític,
- la desviació típica.

2.3.1. Rang

El rang mostra la diferència entre el valor màxim i el valor mínim d'una distribució. És una mesura de dispersió que no s'utilitza gaire perquè és molt sensible als valors extrems, de manera que un sol valor pot tenir un gran efecte en el valor que mostri el rang. En el codi següent indiquem dues maneres de calcular el rang:

1) La primera és utilitzar la funció `range()` per a trobar els valors màxim i mínim d'una distribució i aplicar-hi `diff()` per a veure la diferència entre aquests dos valors.

2) La segona és restar-li el valor mínim `min()` de la distribució al valor màxim `max()`.

```
> diff(range(gapminder$gdpPercap))
> max(gapminder$gdpPercap) - min(gapminder$gdpPercap)
113282
```

Els valors mínim i màxim de la distribució també es poden obtenir mitjançant la funció `summary()`.

2.3.2. Rang Interquartílic

A diferència del rang, el rang interquartílic (*InterQuartile Range*, IQR) és menys sensible als valors extrems, ja que mesura la diferència entre el primer i el tercer quartil de la distribució. En altres paraules, mesura la diferència entre els valors que ocupen els llocs 25 i 75 per cent del rang en una distribució ordenada. Aquesta mesura de dispersió es pot obtenir amb la funció `IQR()`. Alternativament, com veiem en el codi següent, també podem calcular-lo restant el quantil 25 al quantil 75 fent servir la funció `quantile()` i indicant, en escala de 0 a 1, la posició corresponent del primer i el tercer quartil.

L'IQR és la mida de la caixa

Fixeu-vos en la figura 7 d'aquest mòdul. El que fa l'IQR és mesurar de forma quantitativa la mida de les caixes d'un diagrama de caixes.

```
> IQR(gapminder$gdpPercap)
> quantile(gapminder$gdpPercap, 0.75) - quantile(gapminder$gdpPercap, 0.25)
8123.402
```

Els primer i tercer quartil de la distribució també es poden obtenir mitjançant la funció `summary()`. Els trobarem amb el nom de *1st Qu.* i *3rd Qu.*

2.3.3. Desviació típica

La desviació típica és una mesura de dispersió força més complexa en comparació amb les que hem vist anteriorment, i és per això que hi haurem de dedicar més espai i atenció. Aquesta mesura indica la dispersió dels valors respecte de la mitjana. Per a entendre el seu significat numèric haurem de tenir presents tres qüestions importants:

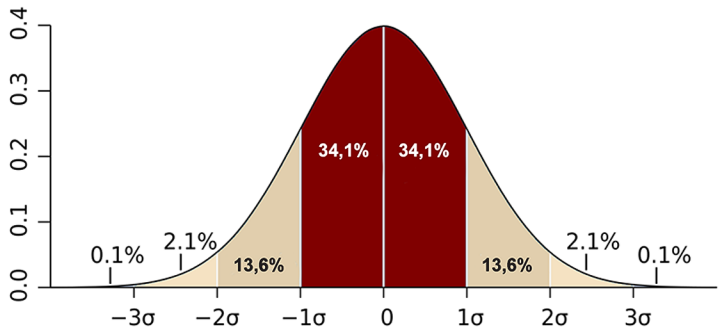
1) En primer lloc, un valor alt de desviació típica voldrà dir que, de mitjana, els valors estan situats lluny de la mitjana, mentre que un valor baix de desviació típica voldrà dir que els valors estan concentrats a prop de la mitjana.

2) En segon lloc, la desviació típica està mesurada amb les mateixes unitats que els valors originals. Això vol dir que si mesurem l'esperança de vida en anys, on els valors oscil·len entre 50 i 80 anys, tindrem desviacions típiques

més baixes que si mesurem esperances de vida en mesos, on tindrem valors més elevats. Si mesurem el PIB per càpita obtindrem segurament desviacions típiques més altes, ja que els valors d'aquesta variable es mouen entre diversos milers. Per tant, la desviació típica ens serà útil per a comparar distribucions, sempre que estiguin mesurades en les mateixes unitats.

3) En tercer i últim lloc, la desviació típica ens indica amb força precisió quin percentatge de valors estan situats al voltant de la mitjana. En una distribució simètrica normal, com en la figura 14, una distribució típica ens indicarà que el 68,2 per cent dels valors estan situats a l'esquerra i a la dreta de la mitjana. Dues desviacions típiques ens indicaran que un 95,4 per cent dels valors estan situats a esquerra i dreta de la mitjana i tres desviacions típiques ens indicaran que el 99,6 dels valors estan situats a esquerra i dreta de la mitjana.

Figura 14. Distribució normal amb desviacions típiques



Font: M. W. Toews, CC BY 2.5. (https://en.wikipedia.org/wiki/Standard_deviation#/media/File:Standard_deviation_diagram.svg)

Per veure el concepte de la desviació típica d'una manera més pràctica, en el codi següent hem creat els vectors `dev_ex1` i `dev_ex2`, formats per nou valors cada un. Amb la funció `mean()` veurem que la mitjana dels dos vectors és la mateixa: 5,11. Això significa que si anem traient pes als valors numèrics més alts i donem aquest pes als valors numèrics més baixos, de manera que al final ens quedi aquest pes repartit de manera equitativa entre tots els valors, la xifra que obtindrem serà 5,11.

```
dev_ex1 <- c(4, 6, 5, 8, 5, 3, 5, 6, 4)
dev_ex2 <- c(3, 9, 4, 8, 6, 1, 5, 8, 2)
```

Sabem que tant en un vector com en l'altre la mitjana és la mateixa, però fins a quin punt és adequat el nombre 5,11 per a representar-nos els valors de cada vector? Serà més adequat si tots els valors són propers a 5. En canvi, 5,11 serà una mesura menys fidel de representar els valors del vector si aquests són valors allunyats de 5, com 1, 2 o 9.

La desviació típica ens ajuda a fer-nos una idea de si els valors d'un vector estan a prop o lluny de la mitjana. Per a calcular-la hem reproduït els passos per al vector `dev_ex1` en el codi següent. En primer lloc calculem la diferència de cada un dels vectors respecte de la mitjana. Per als valors que siguin inferiors

[(11)]Hem posat les fórmules del quadre següent entre parèntesi per poder veure imprès a la consola cada un dels passos per a obtenir la desviació típica.

a la mitjana, aquesta diferència ens sortirà negativa, mentre que per als valors superiors a la mitjana la diferència serà positiva. Per fer que tots els nombres siguin positius, elevarem tots els valors al quadrat i seguidament els sumarem. A continuació dividim aquesta suma pel nombre d'observacions menys ú per evitar que el valor obtingut sigui més gran a mesura que tinguem més observacions. El valor que obtenim fins aquí s'anomena *variància*. La desviació típica s'obté de l'arrel quadrada de la variància.[11]

```
(diff_mitjana <- dev_ex1 - mean(dev_ex1))
(diff_quadrat <- (diff_mitjana)^2)
(suma_diff_quadrat <- sum(diff_quadrat))
(variancia <- suma_diff_quadrat/(length(dev_ex1) - 1))
(desviacio_tipica <- sqrt(variancia))
```

La desviació típica del vector `dev_ex1` és d'1,45, mentre que si canviem els codis i busquem la desviació típica del vector `dev_ex2` veurem que és de 2,84. Per sort, per saber la desviació típica no ens caldrà fer aquestes operacions complicades, sinó que la funció `sd()` ens la calcularà directament.

I ara que tenim la desviació típica d'aquests vectors, com podem interpretar-los? Principalment, el que podem dir és que en el primer vector trobarem aproximadament un 68,2 per cent dels valors entre 3,66 i 6,56, mentre que en el segon vector trobarem el 68,2 per cent dels valors entre 2,27 i 7,95.[12] Així, doncs, si mirem les desviacions típiques d'un i altre vector podem interpretar que els valors del segon vector estan més dispersos en relació amb la mitjana que no pas en el primer. Si afegim o restem dues desviacions típiques a la mitjana podrem saber també on estaran situats el 95,4 i el 99,6 per cent dels valors.

[12] Aquests nombres els obtenim restant una desviació típica a la mitjana (5,11 – 1,45 = 3,66, i 5,11 – 2,84 = 2,27) i sumant una desviació típica a la mitjana (5,11 +1,45 = 6,56, i 5,11 + 2,84 = 7,95). L'estimació assumeix que es tracta d'una distribució simètrica normal amb un nombre elevat de casos.

La desviació típica és molt més sensible als valors extrems que l'IQR, però ho és menys que el rang.

Exercici: exemple de sensibilitat als valors extrems

Per provar la sensibilitat de les mesures sintètiques de dispersió davant de casos extrems, fixeu-vos en aquest vector: `c(1,2,3,4,5,6,7,8,9,10,20)`. Tenim un cas extrem, el número 20. Proveu de calcular el rang `diff(range())`, el rang interquartílic `IQR()` i la desviació típica `sd()`. Després substituïu el número 20 per l'11 i torneu a fer els mateixos càlculs. Quina és la mesura de dispersió més sensible als casos extrems? I quina és la menys sensible?

Resum

En aquest mòdul hem après diferents tècniques de visualització i quantificació en l'anàlisi d'una sola variable. En la primera secció, no solament hem vist com analitzar visualment una variable, sinó també com creuar-la amb una segona variable per a poder observar amb més matisos els seus valors. Això ens serveix també per a repassar alguns procediments del paquet *dplyr* que ens ajuden a transformar variables. Hem de recordar, principalment, que visualitzarem les variables numèriques amb un histograma i les numèriques al llarg del temps amb un diagrama de línies. Amb el diagrama de caixes també podem visualitzar una variable numèrica, i ens és especialment útil quan la creuem amb una variable categòrica, ja que podem comparar més fàcilment alguns dels seus paràmetres estadístics. Utilitzarem el diagrama de dispersió per a analitzar dues variables numèriques, mentre que reservarem el diagrama de caixes per a les variables categòriques.

La segona part d'aquest mòdul, l'hem dedicat a les tècniques per a quantificar variables. Els estadístics descriptius que en podem treure variaran segons si la variable que volem analitzar és numèrica o categòrica. La funció `summary()` il·lustra d'una manera molt clara les possibilitats principals que ens dona cada tipus de variable. És important retenir que l'anàlisi visual i l'anàlisi quantitativa són en part substitutives i en part complementàries. Són substitutives perquè amb la visualització d'una variable ens podem fer la idea de les mesures quantitatives associades a aquesta variable. En altres paraules, visualitzar una distribució ja ens permet intuir com estaran situades la mitjana, la mediana, la moda o la desviació típica. I viceversa, si sabem les mesures de centralitat i dispersió d'una distribució podem fer-nos una idea de quina forma visual tindrà. Són complementàries perquè si ens mirem una variable pels dos cantons, la part visual i la part quantitativa, podrem veure més matisos i ens ajudarà a fer una anàlisi univariant més detallada.

Exercicis d'autoavaluació

Per a un millor l'aprenentatge, intenteu fer mentalment el màxim d'exercicis possible, sense utilitzar R.

1. Canvia la funció perquè ens mostri les proporcions en cada interval.

```
geom_histogram()
```

2. Tenim una distribució amb valors de 10 a 100. Quin argument indicarem per a tenir un histograma format per 15 intervals?

```
geom_histogram()
```

3. Visualitza el diagrama de densitat següent amb una transparència del 50 per cent.

```
geom_density()
```

4. Genera un diagrama de línia amb totes les variables del marc de dades següent.

```
md$exports, md$any, md$paisos
```

5. Canvia els estètics de la geometria de manera que el color de la línia i el tipus de línia mostri la variable `paisos`.

```
geom_line()
```

6. Indica la geometria correcta per a visualitzar el recompte d'observacions.

```
geom_bar(), geom_col()
```

7. Elimina la llegenda del gràfic amb un argument dins la geometria.

```
geom_col()
```

8. Demana un diagrama de caixes amb les variables següents.

```
md$categorica, md$numerica
```

9. Introdueix soroll horitzontal en el diagrama de dispersió següent.

```
ggplot(md, aes(x, y)) + geom_point()
```

10. Sense utilitzar les funcions de *dplyr*, crea el vector lògic `extrem` a partir de la variable següent per detectar casos superiors al valor 5.

```
militarydata$despesa_pib
```

11. Demana una taula de freqüències amb percentatges de la variable següent.

```
tradedata$countries
```

12. Genera una taula de freqüències de classe vector per la variable següent.

```
tradedata$countries
```

13. Quina és la moda del vector següent?

```
c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 10)
```

14. Quina és la mediana del vector següent?

```
c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 10)
```

15. Simplifica les funcions següents per a calcular la mitjana d'aquest objecte.

```
sum(distribution) / length(distribution)
```

16. Calcula el quantil 66 del vector següent.

```
militarydata$despesa_pib
```

17. Demana el valor mínim del vector següent.

```
militarydata$despesa_pib
```

18. Pregunta a R si l'IQR del vector següent és inferior al rang.

```
md$vector
```

19. Quin és el resultat obvi de l'exercici 18?

```
TRUE o FALSE
```

20. Calcula la desviació típica del vector següent.

```
militarydata$vector
```

Solucionari

1. `geom_histogram(position = "fill")`

2. `geom_histogram(bins = 15)` o `geom_histogram(binwidth = 6)`

3. `geom_density(alpha = 0.5)`

4. `ggplot(md, aes(x = any, y = exports, col = paisos)) + geom_line()`

5. `geom_line(aes(col = paisos, lty = paisos)`

6. `geom_bar()`

7. `geom_col(show.legend = FALSE)`

8. `ggplot(md, aes(x = categorica, y = numerica)) + geom_boxplot()`

9. `ggplot(md, aes(x, y)) + geom_jitter(width = 1.2, height = 0)`

10. `extrem <- militarydata$despesa_pib > 5`

11. `prop.table(table(tradedata$countries))`

12. `summary(tradedata$countries )`

13. `10`

14. `6`

15. `mean(distribution)`

16. `quantile(militarydata$despesa_pib, 0.66)`

17. `min(militarydata$despesa_pib)`

18. `IQR(md$vector) < diff(range(md$vector))`

19. `TRUE`

20. `sd(militarydata$vector)`

Glossari

`geom_bar()` Introdueix la geometria d'un diagrama de barres.

`geom_boxplot()` Introdueix la geometria d'un diagrama de caixes.

`geom_col()` Introdueix la geometria d'un diagrama de barres.

`geom_density()` Introdueix la geometria d'un diagrama de densitat.

`geom_histogram()` Introdueix la geometria d'histograma.

`geom_hline()` Introdueix la geometria d'una línia horitzontal.

`geom_jitter()` Introdueix soroll a un diagrama de dispersió.

`geom_line()` Introdueix la geometria d'un diagrama de línia.

`geom_point()` Introdueix la geometria d'un diagrama de dispersió.

`geom_vline()` Introdueix la geometria d'una línia vertical.

`ggtitle()` Permet posar un títol al gràfic.

`IRQ()` Mostra el rang interquartílic d'una distribució.

`labs()` Permet posar etiquetes al gràfic.

`max()` Mostra el valor més alt d'una distribució numèrica.

`mean()` Retorna la suma de tots els valors d'una distribució dividit pel nombre de valors.

`median()` Retorna el valor central d'una distribució ordenada.

`min()` Mostra el valor més baix d'una distribució numèrica.

`prop.table()` Permet obtenir els percentatges d'una taula de freqüències.

`quantile()` Mostra el valor del quantil que especifiquem d'una distribució.

`range()` Retorna el valor mínim i màxim d'una distribució.

`sd()` Mostra la desviació típica d'una distribució.

`table()` Mostra la taula de freqüències d'una variable categòrica.

`xlab()` Permet introduir el títol d'eix de les x.

`ylab()` Permet introduir el títol d'eix de les y.

Bibliografia

Babbie, E. R. (2013). *The practice of social research*. Wadsworth: Cengage Learning.

Brancati, D. (2018). *Social Scientific Research*. Londres: Sage Publications Ltd.

Halperin, S.; Heath, O. (2016). *Political Research: Methods and Practical Skills*. Oxford: Oxford University Press.

King, G.; Keohane, R. O.; Verba, S. (1994). *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton: Princeton University Press.

Annex del mòdul

Codi Apartat 2.1. Taula de freqüències

```
wb <- data.frame(country = c("Antigua and Barbuda", "Belize", "Costa Rica", "Dominica",
"Dominican Republic", "El Salvador", "Guyana", "Guatemala", "Haiti", "Honduras", "Jamaica",
"Nicaragua", "Panama", "Surinam", "Trinidad and Tobago"), income = factor(c("high", "upper-middle",
"upper-middle", "upper-middle", "upper-middle", "lower-middle", "upper-middle", "upper-middle",
"low", "lower-middle", "upper-middle", "lower-middle", "high", "upper-middle", "high")),
stringsAsFactors = FALSE)
```

Codi Figura 12

```
moda <- data.frame(
  unimodal = rnorm(10000, 50), #unimodal
  bimodal = c(rnorm(5000, 60, 5), rnorm(5000, 40, 5)), #bimodal
  multimodal = c(rnorm(2500, 12, 7), rnorm(2500, 37, 7),
                 rnorm(2500, 63, 7), rnorm(2500, 88, 7)),#multimodal
  uniforme = rep(sample(1:100, 5000, replace = T), 2))
moda_ty <- gather(moda, tipus)
ggplot(moda_ty, aes(x = value)) +
  geom_histogram() +
  facet_grid(. ~ tipus, scale = "free")
```

Codi Figura 13

```
simetria <- data.frame(simetrica = rnorm(10000, 50),
              dre = c(rnorm(5500, 80, 5), rnorm(2500, 65, 9),
                      rnorm(1500, 50, 9), rnorm(500, 30, 9)),
              esq = c(rnorm(5500, 20, 5), rnorm(2500, 40, 9),
                      rnorm(1500, 50, 9), rnorm(500, 80, 9)))
distr_ty <- gather(simetria, tipus, valor)
mm <- distr_ty %>%
  group_by(tipus) %>%
  summarize(grp.mean = mean(valor),
            grp.median = median(valor))
ggplot(distr_ty, aes(x = valor)) +
  geom_density(fill = "lightblue", alpha = 0.5) +
  facet_wrap(. ~ tipus, scale = "free",
             labeller = as_labeller(c(dre = "Asimetria positiva",
                                      esq = "Asimetria negativa",
                                      simetrica = "Distribució simètrica"))) +
  geom_vline(data = mm, aes(xintercept=grp.mean),
           color="blue", linetype="dashed",
```

```
            size=1) +
  geom_vline(data = mm, aes(xintercept=grp.median),
             color="green", linetype="dashed",
             size=1)
```