Testing Psicologico

Lezione 2

Filippo Gambarota

@Università di Padova

07/11/2022

Case Study 👩‍🚀

Un (fake) dataset

Il dataset `psych.csv` contiene un dataframe di dati simulati riguardo a **diagnosi psicologiche**, **questionari** e **variabili sociodemografiche**. Le variabili sono:

- `eta`: corrisponde all'età dei pazienti (età minima dei pazienti 10 anni)
- `diagnosi`: rappresenta la diagnosi psicologica
- `ses`: rappresenta lo status socio-economico
- `dep_score`: rappresenta il punteggio ad un questionario di depressione dove 0 è il punteggio minimo (assenza di aspetti depressivi) e 100 è il punteggio massimo di depressione.
- `ans_score`: rappresenta il punteggio ad un questionario di ansia dove 0 è il punteggio minimo (assenza di aspetti ansiosi) e 100 è il punteggio massimo di ansia
- `self_esteem_score`: rappresenta il punteggio ad item di autostima dove 1 è bassa autostima e 6 è alta autostima

Before starting...

- Importiamo il dataset (attenzione all'*estensione*, al *separatore* e agli *argomenti* della funzione che scegliete)
- Controllare struttura e tipo di variabili del dataset
- Controllare **anomalie** nel dataset (valori strani o mancanti) e sistemarli. Gli errori anche in raccolta dati sono sempre dietro l'angolo
  - nel caso di valori mancanti, rimuovere quelle righe (vedi la funzione `complete.cases()`)
  - nel caso di valori anomali, creare un nuovo dataset con le righe anomale e rimuoverle dal dataset principale
- Aggiungere una colonna che indica il numero del paziente da 1 a quanti sono i pazienti
- Aggiungere una colonna `eta_bin` che prende i valori di "maggiorenne" e "minorenne" in base all'età
- Rimuovere dal dataset le righe associate a pazienti con diagnosi `altro`

Before starting...

Before starting...

```
dat <- read.csv("../../data/psych.csv", header = TRUE, sep = ";") # importare
```

Before starting...

```r
dat <- read.csv("../../data/psych.csv", header = TRUE, sep = ";") # importare
```

```r
str(dat) # struttura e tipo di variabili
```

```
## 'data.frame':    185 obs. of  6 variables:
##  $ eta              : int  17 44 21 14 44 40 9 19 20 12 ...
##  $ diagnosi         : chr  "ocd" "depressione" "ocd" "ansiasociale" ...
##  $ ses              : chr  "basso" "basso" "alto" "basso" ...
##  $ dep_score        : int  11 100 10 28 71 80 84 83 8 68 ...
##  $ ans_score        : int  72 37 68 78 40 48 41 61 62 45 ...
##  $ self_esteem_score: int  5 1 3 2 6 1 6 1 2 1 ...
```

Before starting...

```r
dat <- read.csv("../../data/psych.csv", header = TRUE, sep = ";") # importare
```

```r
str(dat) # struttura e tipo di variabili
```

```
## 'data.frame':    185 obs. of  6 variables:
##  $ eta              : int  17 44 21 14 44 40 9 19 20 12 ...
##  $ diagnosi         : chr  "ocd" "depressione" "ocd" "ansiasociale" ...
##  $ ses              : chr  "basso" "basso" "alto" "basso" ...
##  $ dep_score        : int  11 100 10 28 71 80 84 83 8 68 ...
##  $ ans_score        : int  72 37 68 78 40 48 41 61 62 45 ...
##  $ self_esteem_score: int  5 1 3 2 6 1 6 1 2 1 ...
```

```r
# ?complete.cases # a cosa serve?
complete.cases(dat)
```

```
##  [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [19] TRUE TRUE
##  [ reached getOption("max.print") -- omitted 165 entries ]
```

```r
dat <- dat[complete.cases(dat), ] # selezioniamo solo quelli senza valori mancanti
```

Before starting...

Before starting...

Abbiamo detto che l'età dovrebbe essere maggiore/uguale a 10 anni. Vediamo se abbiamo dei casi anomali:

```
dat[dat$eta < 10, ] # errori di codifica o raccolta dati
```

```
##     eta    diagnosi  ses dep_score ans_score self_esteem_score
## 7     9 depressione alto        84        41                 6
## 95    8 depressione alto        74        32                 6
## 127   9 depressione alto        74        34                 3
```

```
dat_errori <- dat[dat$eta < 10, ] # dataset con errori
dat <- dat[dat$eta >= 10, ] # dataset senza errori
```

Before starting...

Abbiamo detto che l'età dovrebbe essere maggiore/uguale a 10 anni. Vediamo se abbiamo dei casi anomali:

```
dat[dat$eta < 10, ] # errori di codifica o raccolta dati
```

```
##     eta    diagnosi  ses dep_score ans_score self_esteem_score
## 7     9 depressione alto        84        41                 6
## 95    8 depressione alto        74        32                 6
## 127   9 depressione alto        74        34                 3
```

```
dat_errori <- dat[dat$eta < 10, ] # dataset con errori
dat <- dat[dat$eta >= 10, ] # dataset senza errori
```

Continuiamo con il pre-processing...

```
dat$id <- 1:nrow(dat) # colonna che identifica il numero del paziente
```

Before starting...

Abbiamo detto che l'età dovrebbe essere maggiore/uguale a 10 anni. Vediamo se abbiamo dei casi anomali:

```
dat[dat$eta < 10, ] # errori di codifica o raccolta dati
```

```
##     eta    diagnosi  ses dep_score ans_score self_esteem_score
## 7     9 depressione alto        84        41                 6
## 95    8 depressione alto        74        32                 6
## 127   9 depressione alto        74        34                 3
```

```
dat_errori <- dat[dat$eta < 10, ] # dataset con errori
dat <- dat[dat$eta >= 10, ] # dataset senza errori
```

Continuiamo con il pre-processing...

```
dat$id <- 1:nrow(dat) # colonna che identifica il numero del paziente
```

```
dat$eta_bin <- ifelse(dat$eta >= 18, "maggiorenne", "minorenne")
```

Before starting...

Abbiamo detto che l'età dovrebbe essere maggiore/uguale a 10 anni. Vediamo se abbiamo dei casi anomali:

```
dat[dat$eta < 10, ] # errori di codifica o raccolta dati
```

```
##     eta    diagnosi  ses dep_score ans_score self_esteem_score
## 7     9 depressione alto        84        41                 6
## 95    8 depressione alto        74        32                 6
## 127   9 depressione alto        74        34                 3
```

```
dat_errori <- dat[dat$eta < 10, ] # dataset con errori
dat <- dat[dat$eta >= 10, ] # dataset senza errori
```

Continuiamo con il pre-processing...

```
dat$id <- 1:nrow(dat) # colonna che identifica il numero del paziente
```

```
dat$eta_bin <- ifelse(dat$eta >= 18, "maggiorenne", "minorenne")
```

```
dat <- dat[dat$diagnosi != "altro", ] # togliamo le diagnosi altro
```

Exploratory Data Analysis 👨‍🚀

EDA

- Calcoliamo delle statistiche descrittive appropriate per ogni tipologia di variabile
- Calcoliamo le frequenze relative delle `diagnosi`
- Calcoliamo le frequenze relative delle `diagnosi` condizionate al `ses`
- Facciamo un istogramma dei punteggi di `ansia` e `depressione`
- Facciamo un barplot dei punteggi di `autostima`
- Facciamo un barplot dei punteggi di `autostima` condizionati al tipo di `diagnosi`
- Calcoliamo `media`, `mediana` e `deviazione standard` e `coefficiente di variazione` dei punteggi di `depressione`, `ansia` e `autostima` per ogni `diagnosi`. Cosa notiamo?

EDA

EDA

```
summary(dat) # funzione utile ma molto generica
```

```
##       eta          diagnosi             ses              dep_score        ans_score
##  Min.   :11.00   Length:149         Length:149         Min.   :  0.00   Min.   : 0.00
##  1st Qu.:20.00   Class :character   Class :character   1st Qu.: 33.00   1st Qu.:35.00
##  self_esteem_score       id           eta_bin
##  Min.   :1.000     Min.   :  1.00   Length:149
##  1st Qu.:1.000     1st Qu.: 47.00   Class :character
##  [ reached getOption("max.print") -- omitted 4 rows ]
```

```
table(dat$ses)
```

```
##
##  alto basso medio
##    68    44    37
```

```
table(dat$diagnosi)
```

```
##
## ansiasociale     bipolare  depressione          ocd
##           27           10           89           23
```

EDA

```
f_diagnosi <- table(dat$diagnosi)/length(dat$diagnosi)
f_diagnosi
```

```
##
## ansiasociale     bipolare  depressione          ocd
##   0.18120805   0.06711409   0.59731544   0.15436242
```

```
sum(f_diagnosi)
```

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

```
# provate ad usare ?prop.table()
```

EDA

```
f_diagnosi <- table(dat$diagnosi)/length(dat$diagnosi)
f_diagnosi
```

```
##
## ansiasociale     bipolare  depressione          ocd
##   0.18120805   0.06711409   0.59731544   0.15436242
```

```
sum(f_diagnosi)
```

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

```
# provate ad usare ?prop.table()
```

```
ds_f <- table(dat$diagnosi, dat$ses) # assolute, come le relativizziamo? vedi prop.table()
ds_f
```

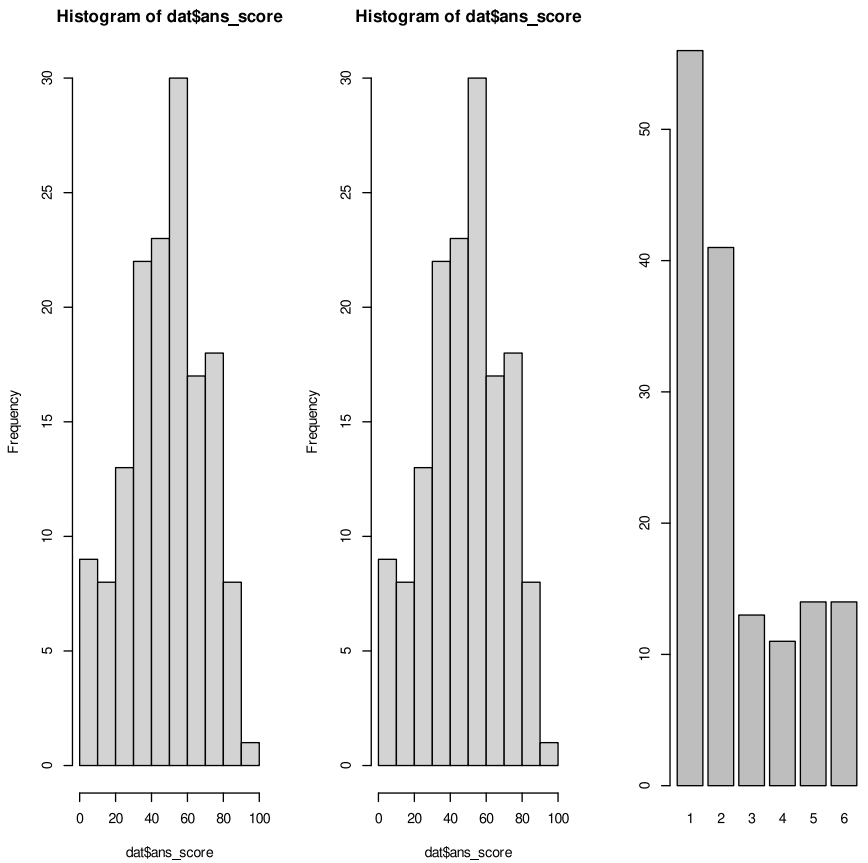
```
##
##                alto basso medio
##   ansiasociale   11     9     7
##   bipolare        4     5     1
##   depressione    45    26    18
##   ocd             8     4    11
```

```
prop.table(ds_f, margin = 1) # a cosa serve margin?
```

```
##
##                     alto     basso     medio
```

EDA

```
par(mfrow = c(1,3)) # per fare 2 grafici
hist(dat$ans_score)
hist(dat$ans_score)
barplot(table(dat$self_esteem_score))
```



EDA

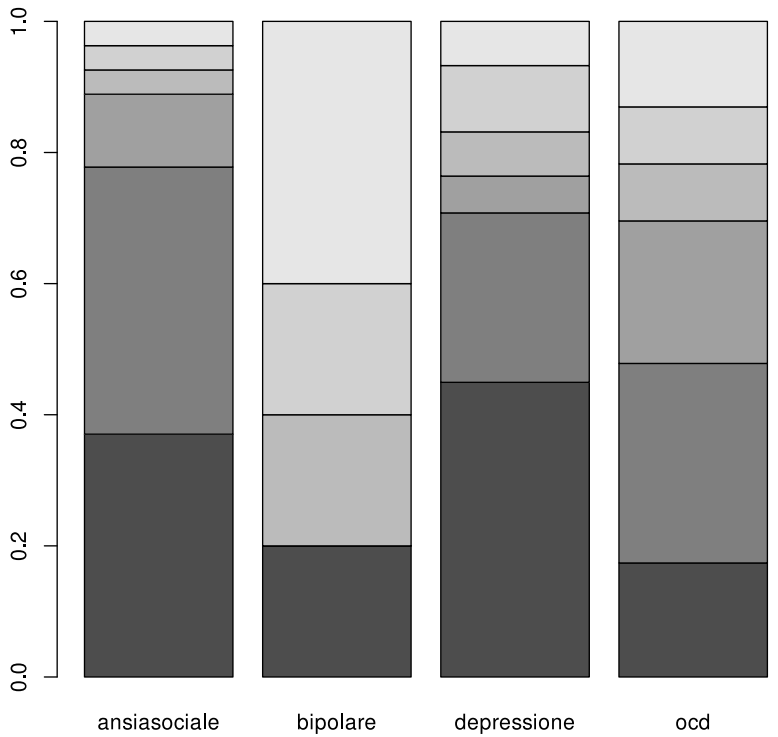
Il barplot condizionato è molto utile per esplorare visivamente la relazione tra due variabili categoriali:

```
barplot(table(dat$self_esteem_score, dat$diagnosi)) # quale è il problema qui?
```

EDA

Dobbiamo fare le frequenze relative per fare in modo che sia più efficacie:

```
freq_self_diag <- prop.table(table(dat$self_esteem_score, dat$diagnosi), 2)
barplot(freq_self_diag)
```



EDA

EDA

Calcoliamo `media`, `mediana` e `deviazione standard` e `coefficiente di variazione` dei punteggi di `depressione`, `ansia` e `autostima` per ogni `diagnosi`. Cosa notiamo?

EDA

Calcoliamo `media`, `mediana` e `deviazione standard` e `coefficiente di variazione` dei punteggi di `depressione`, `ansia` e `autostima` per ogni `diagnosi`. Cosa notiamo?

```
mean(dat$dep_score[dat$diagnosi == "ansiasociale"])
```

```
## [1] 39.85185
```

```
mean(dat$dep_score[dat$diagnosi == "bipolare"])
```

```
## [1] 8.1
```

```
# ...
```

EDA

Calcoliamo `media`, `mediana` e `deviazione standard` e `coefficiente di variazione` dei punteggi di `depressione`, `ansia` e `autostima` per ogni `diagnosi`. Cosa notiamo?

```
mean(dat$dep_score[dat$diagnosi == "ansiasociale"])
```

```
## [1] 39.85185
```

```
mean(dat$dep_score[dat$diagnosi == "bipolare"])
```

```
## [1] 8.1
```

```
# ...
```

Questo non è il modo più efficace e compatto... 😿

EDA

Possiamo usare la funzione `?tapply()` per applicare una `funzione` a `y` splittando per `x`:

```
cv <- function(x){
  # https://it.wikipedia.org/wiki/Coefficiente_di_variazione
  sd(x, na.rm = TRUE) / abs(mean(x, na.rm = TRUE))
}

# depressione
tapply(X = dat$dep_score, INDEX = dat$diagnosi, FUN = mean, na.rm = TRUE)
```

```
## ansiasociale     bipolare  depressione          ocd
##     39.85185      8.10000     81.65169     14.39130
```

```
tapply(X = dat$dep_score, INDEX = dat$diagnosi, FUN = median, na.rm = TRUE)
```

```
## ansiasociale     bipolare  depressione          ocd
##         40.0          1.5         82.0         11.0
```

```
tapply(X = dat$dep_score, INDEX = dat$diagnosi, FUN = sd, na.rm = TRUE)
```

```
## ansiasociale     bipolare  depressione          ocd
##    10.200693    10.681760    13.026286     9.403767
```

```
tapply(X = dat$dep_score, INDEX = dat$diagnosi, FUN = cv)
```

```
## ansiasociale     bipolare  depressione          ocd
```

EDA - Extra

Facciamo un modo ancora più compatto! 😄

```
my_summary <- function(x){
  c(
    mean = mean(x, na.rm = TRUE),
    median = median(x, na.rm = TRUE),
    sd = sd(x, na.rm = TRUE),
    cv = sd(x, na.rm = TRUE) / abs(mean(x, na.rm = TRUE))
  )
}

tapply(dat$ans_score, dat$diagnosi, my_summary)
```

```
## $ansiasociale
##       mean     median         sd         cv
## 71.9629630 71.0000000  9.4968881  0.1319691
##
## $bipolare
##     mean   median       sd       cv
## 7.600000 4.500000 9.582391 1.260841
##
## $depressione
##       mean     median         sd         cv
## 40.8539326 43.0000000 14.4965126  0.3548376
##
## $ocd
##       mean     median         sd         cv
## 68.0434783 72.0000000 13.7460019  0.2020179
```

```
# ... così per tutte le variabili
```

EDA - Extra Extra Extra

Facciamo un modo ancooooooora più compatto! 😱

```
ys <- list(depressione = dat$dep_score,
           ansia = dat$ans_score,
           autostima = dat$self_esteem_score)

lapply(ys, function(x) tapply(x, dat$diagnosi, my_summary))
```

```
## $depressione
## $depressione$ansiasociale
##       mean     median         sd         cv
## 39.8518519 40.0000000 10.2006927  0.2559653
##
## $depressione$bipolare
##      mean    median        sd        cv
##  8.100000  1.500000 10.681760  1.318736
##
## $depressione$depressione
##       mean     median         sd         cv
## 81.6516854 82.0000000 13.0262856  0.1595348
##
## $depressione$ocd
##       mean     median         sd         cv
## 14.3913043 11.0000000  9.4037668  0.6534339
##
##
## $ansia
## $ansia$ansiasociale
##       mean     median         sd         cv
## 71.9629630 71.0000000  9.4968881  0.1319691
##
## $ansia$bipolare
##     mean   median       sd       cv
## 7.600000 4.500000 9.582391 1.260841
```