Elementi di statistica descrittiva 1

Test per le organizzaioni

Ottavia M. Epifania
ottavia.epifania@unipd.it

Margherita Calderan
margherita.calderan@unipd.it

Università di Padova

1. Indici di variabilità
   - Campo di variazione
   - La differenza interquartilica

Indici di variabilità

- Il concetto di variabilità si riferisce a quanto i punteggi di una distribuzione sono sparsi ovvero quanto siano simili o dissimili tra loro
- Il ricorso ad un indice di tendenza centrale comporta una forte semplificazione, e da solo non fornisce informazioni esaurienti sulla distribuzione.
- E' fondamentale capire quanto i dati siano dispersi intorno all'indice di tendenza centrale.
- La variabilità è una caratteristica fondamentale delle distribuzioni "*Variability is the essence of statistics*" (Cobb, 1992)

Consideriamo i risultati dei compiti di Psicometria ottenuti dagli studenti di tre diversi Professori:

```
prof1 = c(18, 22, 24, 16, 19, 22 , 18, 21)
mean(prof1)
```

```
[1] 20
```

```
prof2 = c(10, 10, 12, 10, 30, 28 , 30, 30)
mean(prof2)
```

```
[1] 20
```

```
prof3 = c(20, 20, 20, 20, 20, 20 , 20, 20)
mean(prof3)
```

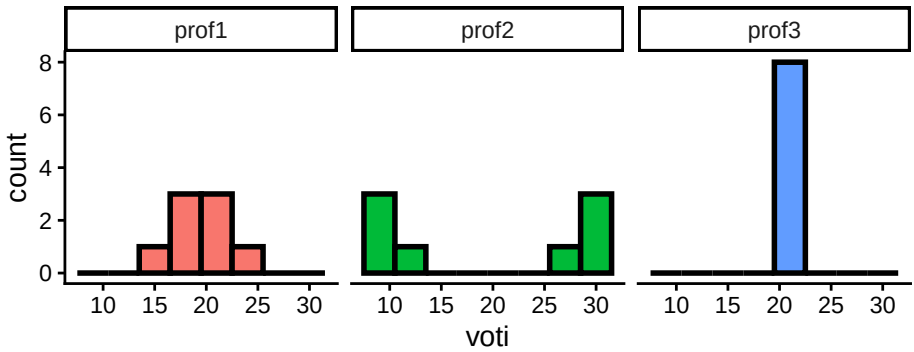
```
[1] 20
```

In ciascun gruppo di studenti la media dei voti è pari a 20, ma è evidente una diversa dispersione intorno a tale valore.

Indici di variabilità o disperisione

- Dobbiamo quindi considerare la la variabilità (o dispersione) di una distribuzione di dati.
- Gli indici di variabilità possono assumere solo valori positivi (non ha senso parlare di dispersione negativa) o nulli (quando i dati osservati hanno tutti lo stesso valore).
- La variabilità minima possibile è 0 e si riferisce a distribuzioni in cui tutti i punteggi sono uguali e dunque non c'è variabilità nei dati.

Quale gruppo è più variabile? In quale gruppo le osservazioni si discostano di più dalla media?



Campo di variazione

Campo di variazione (o gamma) di una distribuzione di dati è la differenza tra il valore massimo e il valore minimo osservato:

$$\text{gamma} = X_{max} - X_{min}$$

```
prof1
```

```
[1] 18 22 24 16 19 22 18 21
```

```
max(prof1) - min(prof1)
```

```
[1] 8
```

Esempio 2:

```
x1 = c(10, 10, 27, 29, 30, 28, 30, 30)
x2 = c(10, 10, 16, 17, 30, 20, 21, 30)

max(x1) - min(x1)
```

```
[1] 20
```

```
max(x2) - min(x2)
```

```
[1] 20
```

Commenti?

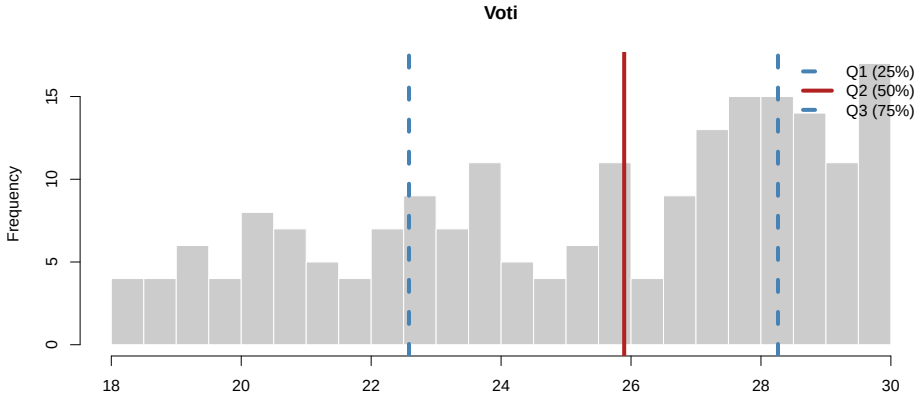
La differenza interquartilica

La differenza interquartilica di una distribuzione è la differenza tra il terzo e il primo quartile (o equivalentemente tra il 75—esimo e il 25—esimo percentile) dei dati:

$$Q = Q_{75} - Q_{25}$$

Cosa vuol dire?

Immaginiamo di avere i voti di 200 studenti:



Il quantile al $\%p$ è il valore tale che circa $\%p$ dei voti è minore o uguale a quel valore.

- Q1 (25%) indica una soglia sotto cui cade circa un quarto della classe;
- Q2 (50%) indica una soglia sotto cui cade la metà della classe;
- Q3 (75%) lascia sotto di sé circa tre quarti dei voti.

Cos'è Q2?

```r
head(voti, n = 6) # primi 6 voti
```

```
[1] 21.19 22.47 24.87 28.90 20.42 28.78
```

```r
tail(voti, n = 6) # ultimi 6 voti
```

```
[1] 27.85 28.77 27.33 29.52 27.95 29.35
```

Per parlare di percentili/quantili dobbiamo prima ordinare i voti dal più basso al più alto:

```
v_sorted = sort(voti) # riordino

head(v_sorted, n = 6) # primi 6 voti
```

```
[1] 18.16 18.16 18.28 18.43 18.71 18.74
```

```
tail(v_sorted, n = 6) # ultimi 6 voti
```

```
[1] 29.78 29.82 29.83 29.90 29.90 29.91
```

Obiettivo: Trovare il valore sotto cui cade circa il 25% dei voti.

```
n = length(voti) # numero di osservazioni

p = 0.25 # voglio il 25° percentile

k = n * p  # # quanti voti rappresentano il 25% del totale (25% di 200)

Q_25 = v_sorted[k] # predo il voto in quella posizione

sum(voti <= Q_25)  # quanti sono <= soglia
```

```
[1] 50
```

Esempio 2:

```
voti = c(21.4, 20, 30, 27.3, 26.4, 20, 18.7, 18.2, 29.4, 28.3)

# riordino
voti_sorted = sort(voti)

voti_sorted
```

```
 [1] 18.2 18.7 20.0 20.0 21.4 26.4 27.3 28.3 29.4 30.0
```

Qual'è il valore sotto cui cade il 50% dei voti?

Qual'è il valore sotto cui cade il 50% e il 75% dei voti?

```
n = length(voti) # numero di osservazioni
p = .50 # che percentile voglio
p50 = n*p # il 50% di 10

voti_sorted[p50]
```

```
[1] 21.4
```

```
p = .75 # che percentile voglio
p75 = n*p # il 75% di 10

# ?ceiling computa il numero intero più grande vicino al valore (7.5)
ceiling(p75)
```

```
[1] 8
```

```
voti_sorted[ceiling(p75)]
```

```
[1] 28.3
```

Fortunatamente esiste la funzione `quantile()` che permette di calcolare i quantili:

```
# probs specifica che quantili vogliamo

quantile(voti, probs = c(0.25,0.50,0.75))
```

```
  25%   50%   75%
20.00 23.90 28.05
```

La differenza interquartilica è un indice di variabilità robusto, risente cioè poco della presenza di valori anomali (outliers) nei dati.

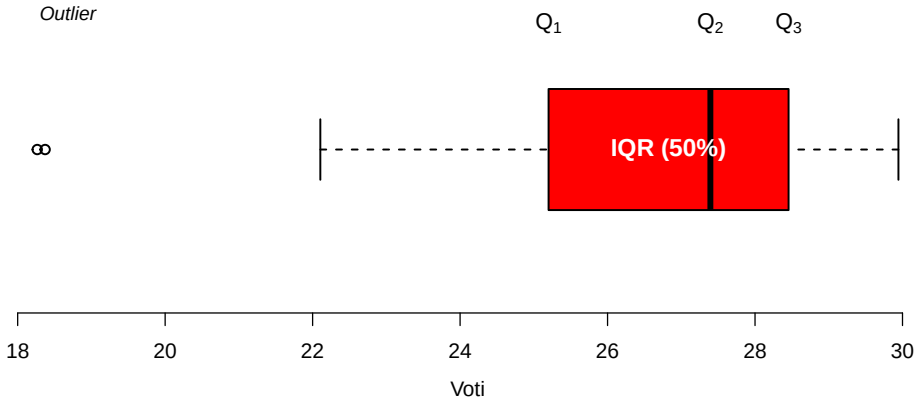
```
q75 = quantile(voti, probs = 0.75)
q25 = quantile(voti, probs = 0.25)

q75-q25
```

```
 75%
8.05
```

```
IQR(voti) # funzione per il calcolo automatico
```

```
[1] 8.05
```

Boxplot



Outilier

- I baffi si estendono fino all'ultimo dato entro 1.5 volte l'IQR dai quartili. Tutto ciò che sta oltre i baffi è un outlier.
- Un outlier è un valore che dista dai quartili (Q_1 o Q_3) più di 1.5 volte la differenza interquartile (IQR).
- È un outlier qualsiasi osservazione che cade fuori dall'intervallo $[Q_1 - 1.5 \times IQR \quad ; \quad Q_3 + 1.5 \times IQR]$.