Sommario

Tipi di variabile:

QUANTITATIVA (numerica)	→ Interessano come quantità:	- Discreta - Continua
QUALITATIVA (caratteri)	→	- Fattori - Identificativi

Statistica descrittiva UNIVARIATA

Per variabili QUANTITATIVE

Abbiamo un dataset (campione casuale / sample) che può essere visto come una matrice in cui le righe sono i soggetti osservati e le colonne rappresentano le diverse variabili.

La taglia del campione corrisponde al numero di osservazioni e quindi al numero delle righe.

Per descrivere un dataset utilizziamo:

1) Numeric summaries (riassumiamo le osservazioni tramite indici statistici)
2) Data visualization (produciamo opportuni grafici)

NUMERIC SUMMARIES

Sono dei numeri e ce ne sono tre tipi:

1) Indici di posizione – indicano dove i dati si trovano sull'asse reale
2) Indici di variabilità / dispersione
3) Indici di forma

1. INDICI DI POSIZIONE

- MEDIA CAMPIONARIA: media aritmetica delle osservazioni $x_1, \dots, x_n$

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Ogni osservazione viene pesata con lo stesso peso $\frac{1}{n}$, quindi la media campionaria molto sensibile ai valori estremali (= valori molto lontani dalla massa generale delle osservazioni)

R: `mean(…)`

- MEDIANA CAMPIONARIA: valore dell'osservazione centrale

  R: `median(…)`

La mediana si legge con la media: le calcolo entrambe e confronto i valori ottenuti.

- Se media ≈ mediana ⟶ campione simmetrico e senza valori estremali
- Se media >> mediana ⟶ asimmetria con valori estremali verso destra
- Se media << mediana ⟶ asimmetria con valori estremali verso sinistra

- P-ESIMO PERCENTILE:

  Lascia il P% delle osservazioni a sinistra:
  il P% delle osservazioni sono i valori minori del P-esimo percentile.

  La mediana corrisponde al 50-esimo percentile.

  Il QUANTILE di ordine α (α-quantile) è quel valore tale che $\mathbb{P}(X \leq q_\alpha) = \alpha$

  R: `quantile(…, alpha)`

2. *INDICI DI DISPERSIONE*

- RANGE: $\max(x_1, \dots, x_n) - \min(x_1, \dots, x_n)$

  R: `range(…)`

- VARIANZA CAMPIONARIA:
  è la media aritmetica degli scarti quadratici dalla media campionaria

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$$

  R: `var(…)`

- DEVIAZIONE STANDARD CAMPIONARIA: $\sqrt{s^2}$

  R: `sd(…)`

- Z-SCORES: trasformo il campione per renderlo confrontabile con altri (ovvero lo standardizzo)

  $x_1, \dots, x_n \rightarrow d_1, \dots, d_n$

  CV – coefficiente di variazione: $\frac{s}{\bar{x}_n} = \begin{cases} > 1 & se\ s > \bar{x}_n \\ \sim 1 & se\ s \approx \bar{x}_n \\ < 1 & se\ s < \bar{x}_n \end{cases}$

3. *INDICI DI FORMA*

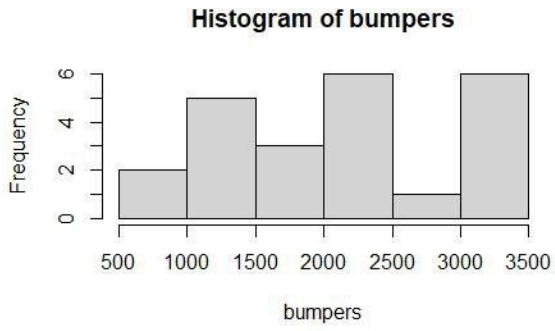
- SKEWNESS:

$$sk = \frac{1}{n}\sum_{i=1}^{n} d_i^3 \quad \begin{cases} > 0 & se\ ho\ più\ osservazioni > \bar{x}_n \\ \approx 0 & se\ ho\ dati\ simmetrici \\ < 0 & se\ ho\ più\ osservazioni < \bar{x}_n \end{cases}$$

  R: `sum(((dataset-mean(…))/sd(…))^3)/length(dataset)`

GRAFICI

ISTOGRAMMA

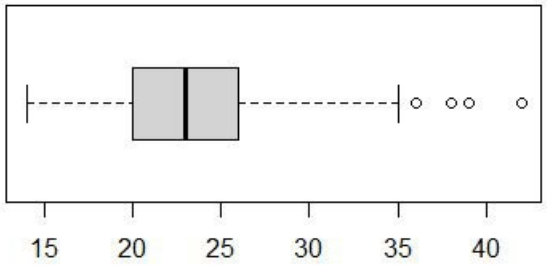



```r
library(UsingR)
data("bumpers")
str(bumpers)

#istogramma
hist(bumpers, breaks=6)

```

BOX PLOT




```r
#boxplot
boxplot(firstchi, horizontal=TRUE)
```

Le linee al lato rappresentano i baffi, mentre i pallini rappresentano i valori estremali (outlier). La linea centrale, più spessa, è la mediana (distanza interquartile), mentre gli estremi della "scatola" sono il 25-esimo e 75-esimo percentile (quartili campionari).

Per variabili QUALITATIVE / CATEGORIALI (fattori)

L'operazione principale è la tabulazione, per contare quante osservazioni cadono in ciascuna categoria.

R: `table(…)`

Statistica descrittiva BIVARIATA

Posso fare un'analisi prendendo in considerazione due variabili:

- Quantitativa/Qualitativa

Vedi l'esempio in *codiciLezioniR -> bivariata.R*

- Quantitativa/Quantitativa

Mi chiedo se c'è una relazione tra le due variabili

Covarianza di X e Y:

$$Cov(X,Y) = \mathbb{E}[(X - \mathbb{E}(X)) * (Y - \mathbb{E}(Y))]$$

R: `cov(x,y)`

Correlazione di X e Y:

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X) * Var(Y)}} \quad \in [-1,1]$$

R: `cor(x,y)`

Vedi gli esempi in *codiciLezioniR -> bivariata.R*

INDICI: covarianza campionaria e correlazione campionaria

$$cov = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n)$$

$$corr = \rho = \frac{1}{n-1}\sum_{i=1}^{n}\frac{x_i - \bar{x}_n}{s_X} * \frac{y_i - \bar{y}_n}{s_Y}$$

Possiamo aggiungere dei parametri su R per fare diverse analisi dei campioni:

- Correlazione di Pearson – mostra se c'è relazione lineare
- Correlazione di Spearman – evidenzia relazioni di monotonia

Vedi gli esempi in *codiciLezioniR -> bivariata.R*

- Qualitativa/Qualitativa

Realizzo una tabella a doppia entrata.

Vedi esempio in *codiciLezioniR -> bivariata.R*

Statistica inferenziale (parametrica)

INFERENZA: processo che ci porta a fare affermazioni riguardanti l'intera popolazione di interesse e non solo sul campione casuale.
Devo tener conto della variabilità del campione casuale tramite intervalli di confidenza e test di ipotesi.

STIMATORI:

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

Ci sono delle famiglie di teoremi che ci permettono di affermare che $\bar{X}_n$ restituisce valori vicini a $\mathbb{E}(X)$ e che S^2 restituisce valori vicini a $Var(X)$.

Teoremi dei grandi numeri

- Sia $(X_i)_{i=1}^{+\infty}$ una collezione di v.a. INDIPENDENTI e IDENTICAMENTE DISTRIBUITE (IID), allora:

$$per\ n \to +\infty: \qquad \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \to \quad \mathbb{E}(X_1)$$

  Se T è stimatore per θ e $\mathbb{E}(T) = \theta$ allora dico che T è STIMATORE CORRETTO per θ.

- Sia $(X_i)_{i=1}^{+\infty}$ una collezione di v.a. IID, allora:

$$per\ n \to +\infty: \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n) \quad \to \quad Var(X_1)$$

  Vedi esempi in *codiciLezioniR -> leggi_grandi_numeri.R*

  $\bar{X}_n \to \mathbb{E}(X_1)$ vuol dire che $\bar{X}_n$ è STIMATORE CONSISTENTE per $\mathbb{E}(X_1)$

  $S^2 \to Var(X_1)$ vuol dire che S^2 è STIMATORE CONSISTENTE per $Var(X_1)$

Teoremi del limite centrale

Mi permettono di stimare quanto sono vicina a $\mathbb{E}(X_1)$ e $Var(X_1)$.

- Sia $(X_i)_{i=1}^{+\infty}$ una collezione di v.a. IID, allora

  $Per\ n \to +\infty:$

  $$\mathbb{P}\left(\frac{\bar{X}_n - \mathbb{E}(X_1)}{\frac{Stdev(X_1)}{\sqrt{n}}} \leq x\right) \to \mathbb{P}(Z \leq x)$$

  $funzione\ di\ \mathbf{distribuzione}\ (CDF)$

  $dove\ Z = \frac{\bar{X}_n - \mathbb{E}(X_1)}{\frac{Stdev(X_1)}{\sqrt{n}}} \sim N(0,1)\ \ normale\ standard$

  Per *n* grande:

  $$\frac{\bar{X}_n - \mathbb{E}(X_1)}{\frac{Stdev(X_1)}{\sqrt{n}}} \approx Z \sim N(0,1) \qquad \bar{X}_n \approx N\left(\mathbb{E}(X_1), \frac{Var(X_1)}{n}\right)$$

  Vedi esempi in *codiciLezioniR -> teoremi_limite_centrale.R*

CASI PARTICOLARI in cui campioni da v.a. Normali:

- Sia $(X_i)_{i=1}^{+\infty}$ una collezione di v.a. IID con legge $N(\mu, \sigma^2)$

  $$\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \sim Z \sim N(0,1) \quad \forall n$$

- Sia $(X_i)_{i=1}^{+\infty}$ una collezione di v.a. IID con legge $N(\mu, \sigma^2)$

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

  Dove $\chi^2(n-1)$ si legge "chi quadro con n-1 gradi di libertà ed è una v.a. continua la cui PDF, al crescere di n, si avvicina sempre più a quella di una Normale.

Intervalli di confidenza

Si calcolano su R:

- **`binom.test`** per trovare l'IC per proporzioni, ovvero partendo da una v.a. di Bernoulli di parametro p (p = proporzione).
- **`t.test`** per trovare l'IC per differenze di medie, in cui bisogna specificare se i campioni scelti sono indiendenti (trattano soggetti diversi) oppure appaiati (due variabili appartenenti allo stesso soggetto).

Vedi esempi in *codiciLezioniR -> intervalli_confidenza.R*

Test di ipotesi

Sono delle procedure che ci permettono di testare ipotesi (affermazioni sui parametri)

Si considerano sempre due ipotesi:

- IPOTESI NULLA H_0 ovvero l'ipotesi alla quale credi
- IPOTESI ALTERNATIVA H_1

L'idea è: credo nell'ipotesi nulla, raccolgo i dati e faccio un test di ipotesi per capire se posso continuare a credere nella nulla oppure se sono costretta a rifiutarla in favore dell'alternativa (che avevo in mente fin dall'inizio e che deve essere diversa dalla nulla).

Solitamente si fissa un livello di significabilità α (molto piccolo), altrimenti è 0 di default.

Utilizzo ancora **`binom.test`** e **`t.test`** su R e controllo il P-value:

> se P-value è troppo piccolo ($< \alpha$), rifiuto l'ipotesi nulla a favore dell'alternativa.

Vedi esempi in *codiciLezioniR -> intervalli_confidenza.R*