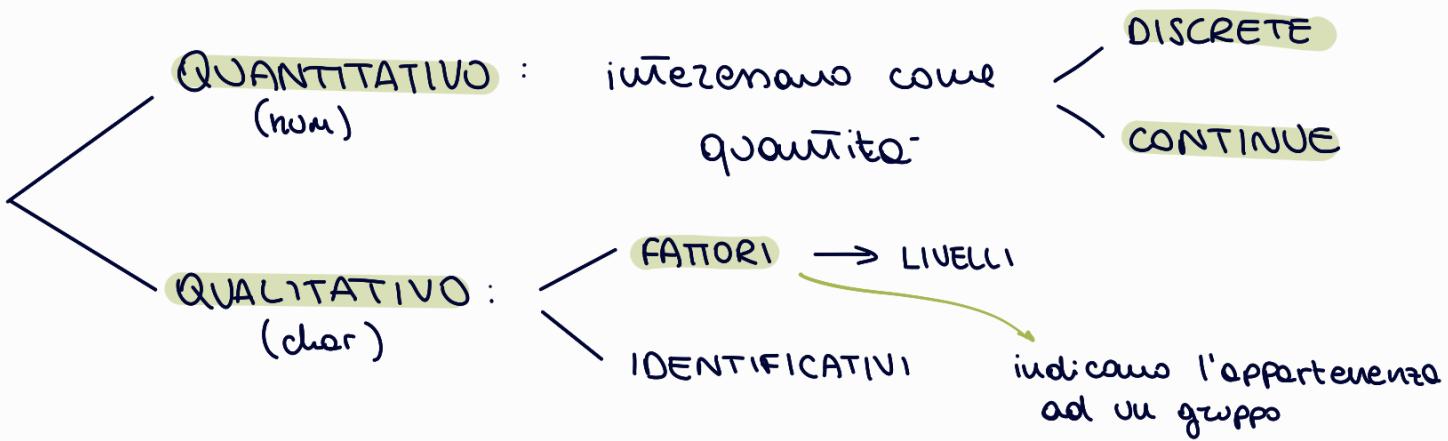


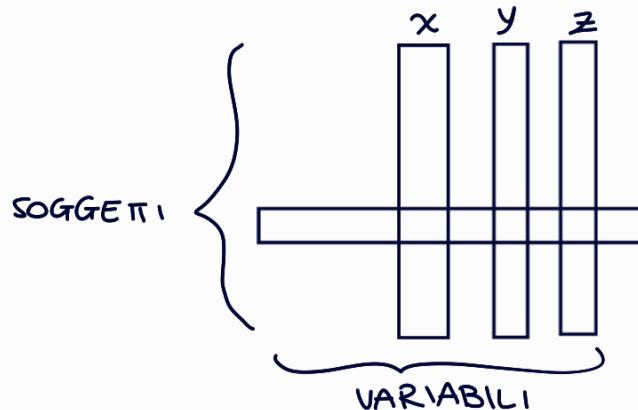
TIPI DI VARIABILE



USIAMO $\text{R} \rightarrow \text{R Studio}$

STATISTICA DESCRITTIVA

Dataset :
CAMPIONE CASUALE (SAMPLE)



x_1, y_1, z_1
 x_2, y_2, z_2
 x_n, y_n, z_n

n: numero di SOGGETTI MISURATI /
n° di osservazioni /
taglia del campione

- DESCRIVERE
- ① NUMERIC SUMMARIES : riassumiamo le osservazioni in singoli numeri → INDICI STATISTICI
 - ② PRODUCIAMO OPPORTUNI GRAFICI (DATA VISUALIZATION)

NUMERIC SUMMARIES

- TIPI :
- 1 INDICI DI POSIZIONE - dove i dati si trovano sull'asse \mathbb{R}
 - 2 INDICI DI VARIABILITÀ / DISPERSIONE
 - 3 INDICI DI FORMA

INDICI DI POSIZIONE

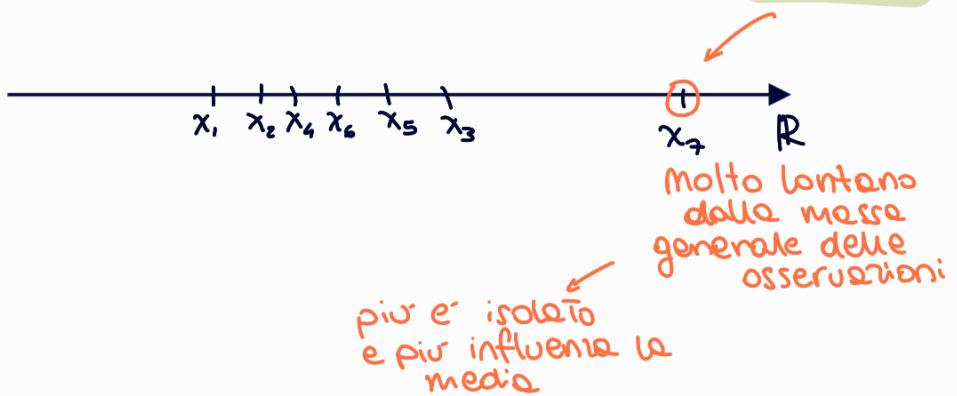
MEDIA CAMPIONARIA

= Media aritmetica delle osservazioni

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

Oss: ogni osservazione viene pesata con lo stesso peso $\frac{1}{n}$,

quindi la media campionaria è molto sensibile a valori estremi



MEDIANA CAMPIONARIA

= OSSERVAZIONE CENTRALE

come si calcola?

ordino le osservazioni;

la mediana lascia metà delle osservazioni a sx e metà a dx

se n dispari è la $\frac{n+1}{2}$ esima

se n pari è la media di $\frac{n}{2}$ esima e $(\frac{n}{2}+1)$ esima
aritmetica

→ 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** ⇒ valori estremali verso dx
ASIMMETRIA
- se **media < mediana** ⇒ valori estremali verso sx

P-ESIMO PERCENTILE

→ generalizza il concetto di mediana

↳ il P-esimo percentile lascia il P% delle osservazioni a sx

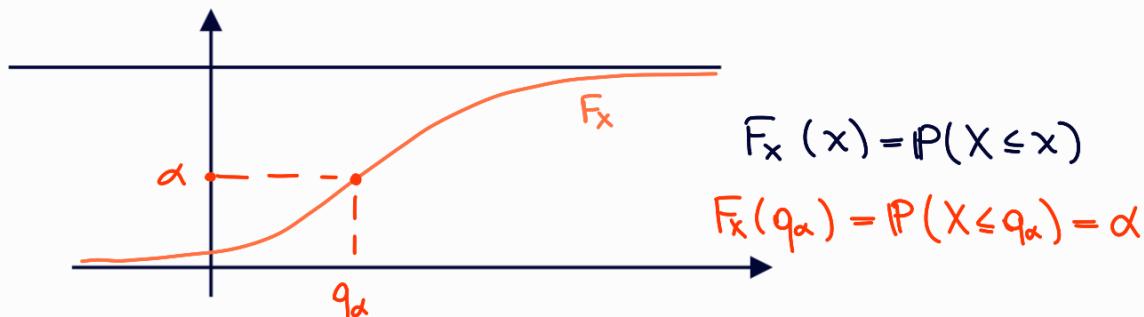
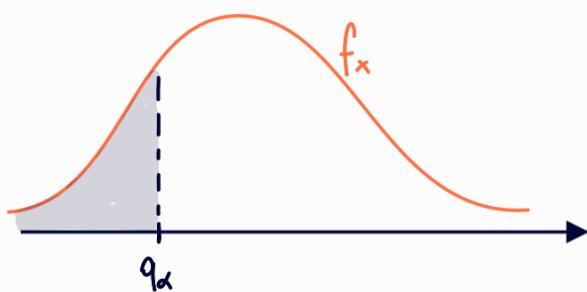
↳ **mediana = 50° percentile**

il P% delle osservazioni sono valori < del P-esimo percentile

hanno una def. corrispondente sulla v.a. e non sul campione

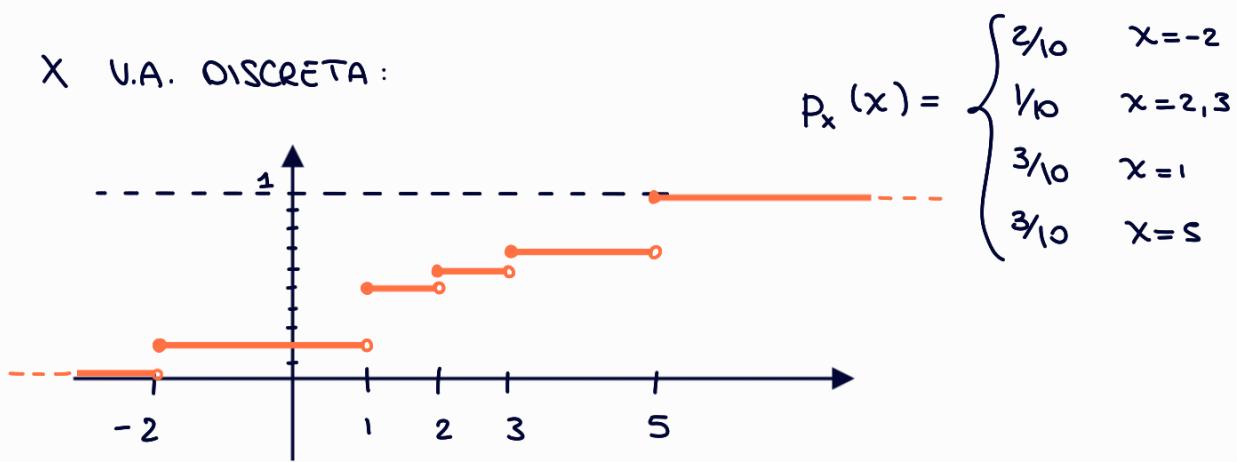
Def. QUANTILE di ordine α (α -QUANTILE) è il valore t.c.

$$P(X \leq q_\alpha) = \alpha$$



COSA SUCEDE NEL DISCRETO?

X V.A. DISCRETA:



$$F_x(x) = P(X \leq x)$$

$$P(X \leq 2) = \underbrace{P(X=2)}_{1/10} + P(X=1) + P(X=-2)$$

$$q_\alpha \mid P(X \leq q_\alpha) = \alpha$$

$$q_\alpha = \sup \{x \mid P(X \leq x) \leq \alpha\}$$

$$\text{se } \alpha = 0.8 \quad q_{0.8} = \sup \{x \mid P(X \leq x) \leq 0.8\}$$

$$= \frac{8}{10}$$

$$\Rightarrow q_{0.8} = 3 \rightarrow \begin{pmatrix} P(X \leq 3) = 0.7 \\ P(X \leq 4) = 1 \end{pmatrix}$$

INDICI DI DISPERSIONE

$$\boxed{\text{RANGE}} = \max(x_1, \dots, x_n) - \min(x_1, \dots, x_n)$$

VARIANZA
CAMPIONARIA

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

↳ media campionaria

e' "media aritmetica"
degli scarti quadratici
dalla media campionaria

$$\sqrt{s^2} = \boxed{\text{DEVIAZIONE STANDARD CAMPIONARIA}}$$

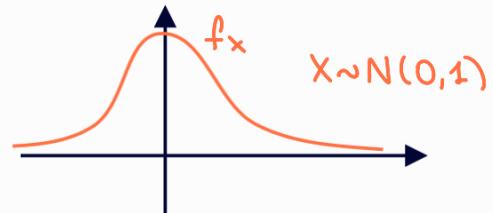
Z - SCORES

→ trasformo il campione per renderlo confrontabile con altri (standardizzazione)

$$\begin{array}{c} x_1 \\ x_2 \\ \vdots \\ x_n \end{array} \rightarrow \begin{array}{l} d_1 = \frac{x_1 - \bar{x}_n}{s} \\ d_2 = \frac{x_2 - \bar{x}_n}{s} \\ \vdots \\ d_n = \frac{x_n - \bar{x}_n}{s} \end{array}$$

z-scores

sto centrandola variabile attorno allo zero e con dispersione ≈ 1



CV - COEFFICIENTE DI VARIAZIONE

$$= \frac{s}{\bar{x}_n} \left\{ \begin{array}{ll} > 1 & \text{se } s > \bar{x}_n \\ \approx 1 & \text{se } s \approx \bar{x}_n \\ < 1 & \text{se } s < \bar{x}_n \end{array} \right.$$

INDICI DI FORMA

SKEWNESS

$$SK = \frac{1}{n} \sum_{i=1}^n d_i^3$$

$> 0 \rightarrow$ ho più osservazioni $> \bar{x}_n$

$\approx 0 \rightarrow$ dati simmetrici

$< 0 \rightarrow$ ho più osservazioni $< \bar{x}_n$

GRAFICI

ISTOGRAMMA

x_1
 x_2
 |
 x_n

→ DATASET
OSSEVAZIONI

1) Costruisco una collezione di intervalli su \mathbb{R}

tutti con stessa ampiezza :
irregolari

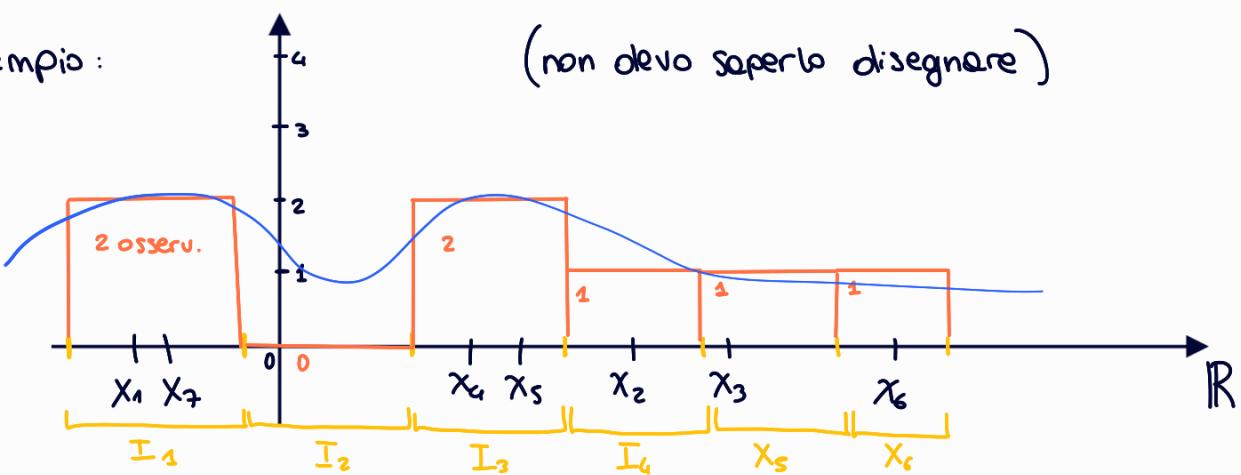
(si chiamano BINS)

2) Su ciascun intervallo costruisco una colonna / rettangolo con area proporzionale al n° di osservazioni che cadono nell'intervallo :

intervalli regolari → guardo le aree = altezze
intervalli irregolari → guardo le aree

esempio:

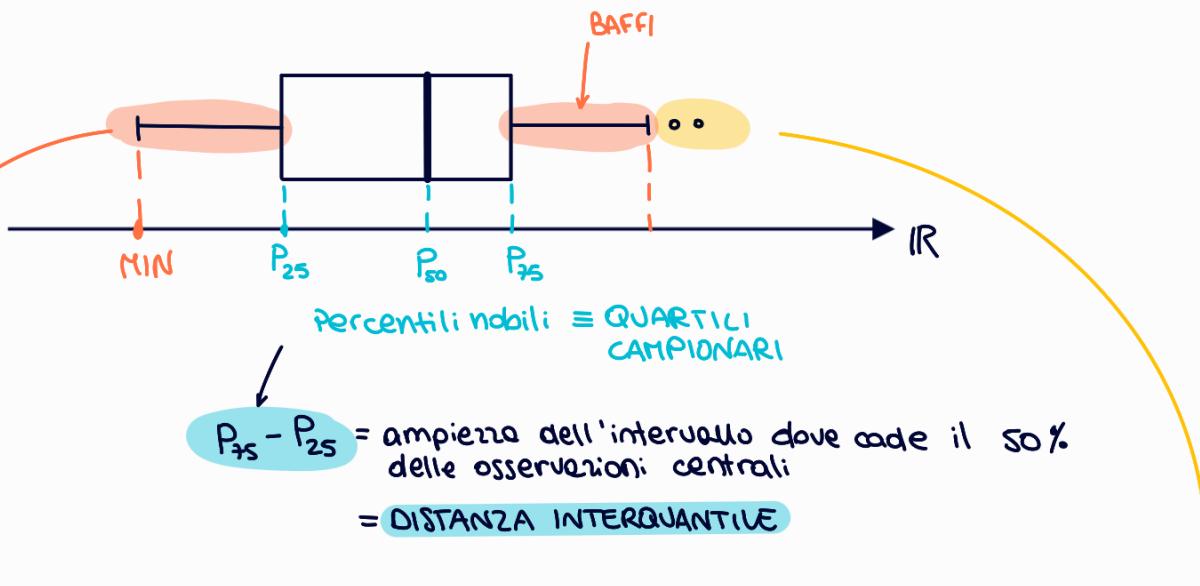
(non devo saperlo disegnare)



L'idea è che le colonne sono più alte negli intervalli con più osservazioni.

- L'andamento è simile alle densità

BOX PLOT



BAFFI arrivano fino a 1.5 volte la distanza interquartile (se ho osservazioni)

ALTRIMENTI arrivano fino a MIN e MAX

Il resto delle osservazioni = PALLINI → **VALORI ESTREMALI**
(OUTLIER)

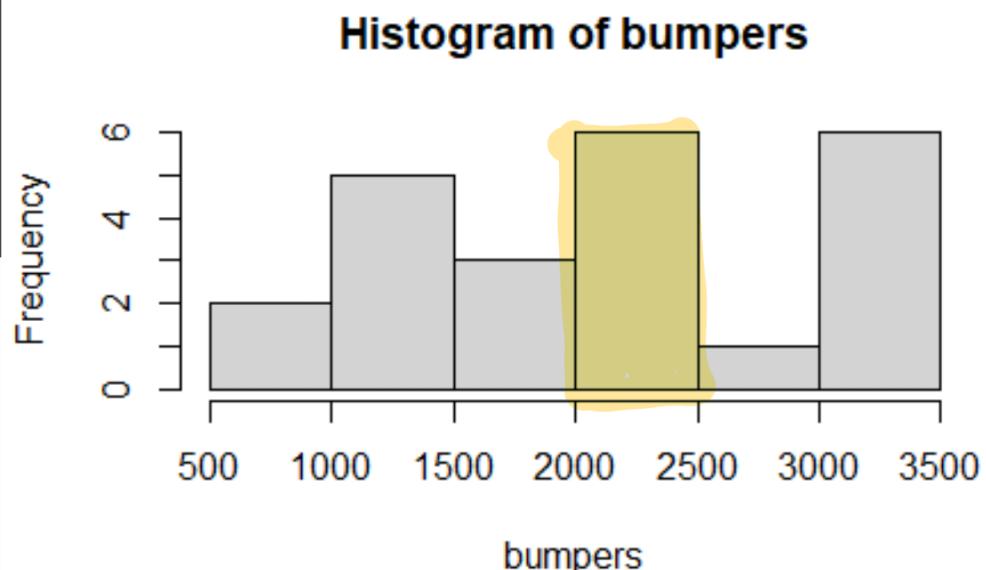
(1) 2.31

2.31 For the data sets **bumpers** ([UsingR](#)), **firstchi** ([UsingR](#)), and **math** ([UsingR](#)), make histograms. Try to predict the mean, median, and standard deviation from the graphic. Check your guesses with the appropriate R commands.

Nomi dei DATASET
 ↓
 ↓
 ↓
 NOMI DEI PACCHETTI
 CHE CONTENGONO I DATASET
`library(UsingR)`
`data()`
`str()`

```

1 library(UsingR)
2 data("bumpers")
3 str(bumpers)
4
5 #istogramma
6 hist(bumpers, breaks=6)
7
  
```



Media \approx Mediana \approx 2000

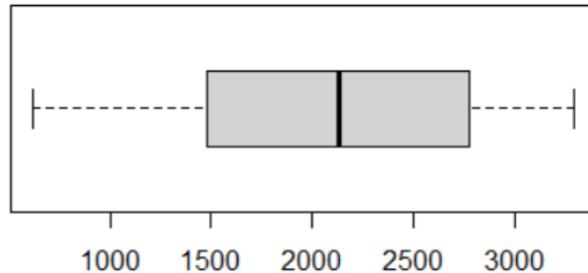
Verifico con R

Deviazione Standard

```

> mean(bumpers)
[1] 2122.478
> median(bumpers)
[1] 2129
> sd(bumpers)
[1] 798.4574
>
  
```

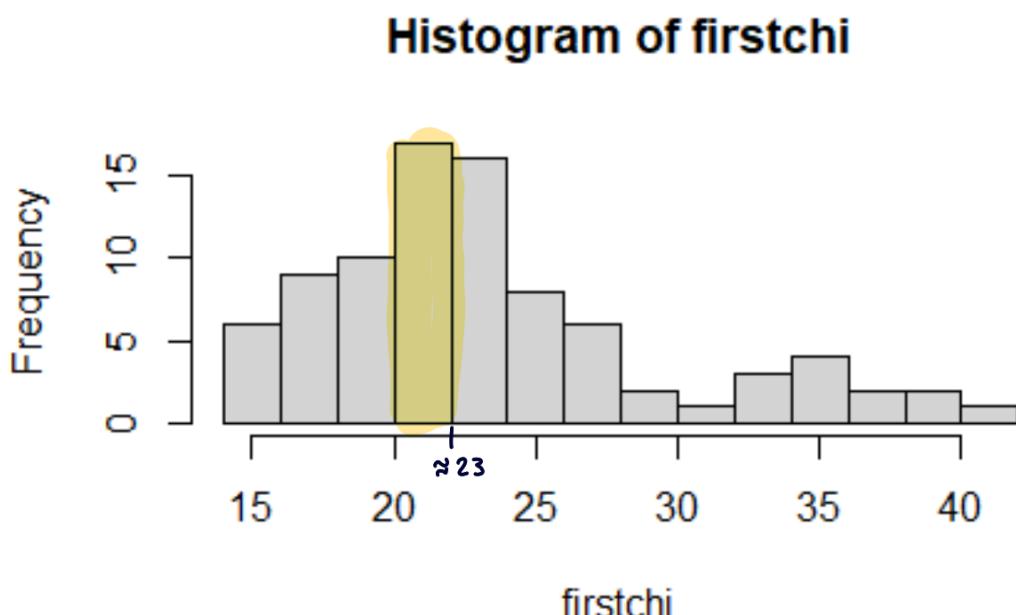
Box PLOT :



```
firstchi :
```

```
data("firstchi")
str(firstchi)

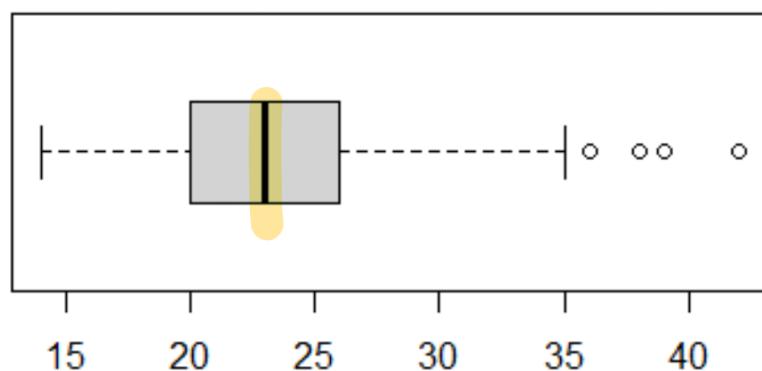
#istogramma
hist(firstchi, breaks=10)
```



```
mean(firstchi)
median(firstchi)
sd(firstchi)
```

```
> mean(firstchi)
[1] 23.97701
> median(firstchi)
[1] 23
> sd(firstchi)
[1] 6.254258
> |
```

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



DESCRITTIVA UNIVARIATA PER VAR. CATEGORIALI (FACTORI)

L'operazione principale e' la tabulazione

⇒ PROONCO / DISEGNO TABULE

→ CONTO QUANTE OSSERVAZIONI CADONO IN
CIASCUA CATEGORIA

TABELLE DI CONTINGENZA
DELE FREQUENZE

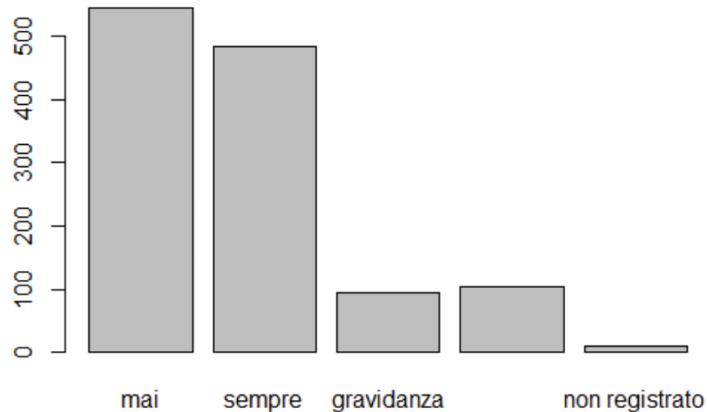
- BAR CHARTS
- DOTS CHARTS
- PIE CHARTS

] OK

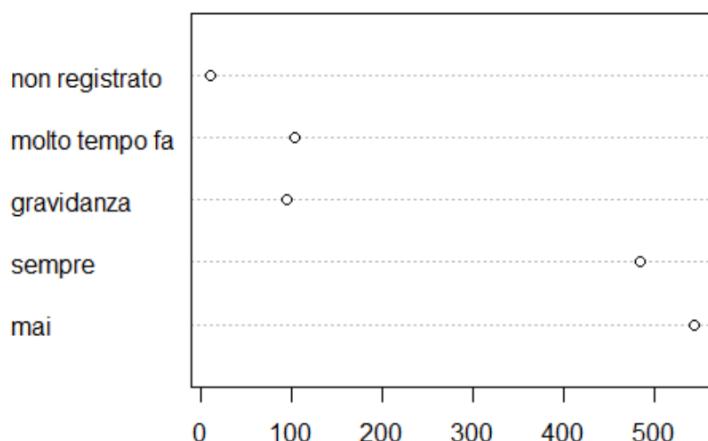
] deprecate

Vedi codice Lezioni R → plots.R

BAR PLOT :

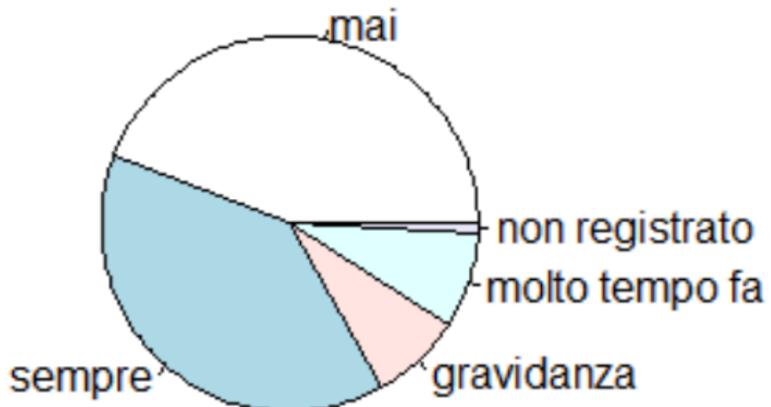


DOT PLOT :



PIE CHART :

↳ poco intuitivo
 ↓
 meglio non usarlo



DESCRITTIVA BIVARIATA

- QUANTITATIVA / QUANTITATIVA → RIPETO LA DESCRITT. UNIVARIATA DELLA V.A. QUANTITATIVA PER CIASCUNA CATEGORIA DELLA VAR. FATTORE

vedi codiceLezioniR → bivariate.R

- QUANTITATIVA / QUANTITATIVA

X	Y
x_1	y_1
x_2	y_2
⋮	⋮
x_n	y_n

es. peso/altezza

DOMANDA STATISTICA: c'è una relazione tra le due variabili?

COVARIANZA DI X e Y

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

CORRELAZIONE di X e Y

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \cdot \text{Var}(Y)}}$$

↓
 sta in $[-1, 1]$

→ INDICI : COVARIANZA CAMPIONARIA
e CORRRELAZIONE CAMPIONARIA

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

$$\rho = \text{cov} = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}_n}{s_x} \right) \cdot \left(\frac{y_i - \bar{y}_n}{s_y} \right)$$

CORRRELAZIONE
DI PEARSON

vedi codiciLezioniR → bivariate.R

↳ mostra se c'è relazione lineare

CORRRELAZIONE
DI SPEARMAN

→ evidenzia relazioni di monotonia

CORRRELAZIONE \neq CAUSALITÀ !

QUALITATIVA/QUALITATIVA

→ le tabule

→ tabella e doppie entrate

vedi codiciLezioniR → bivariate.R

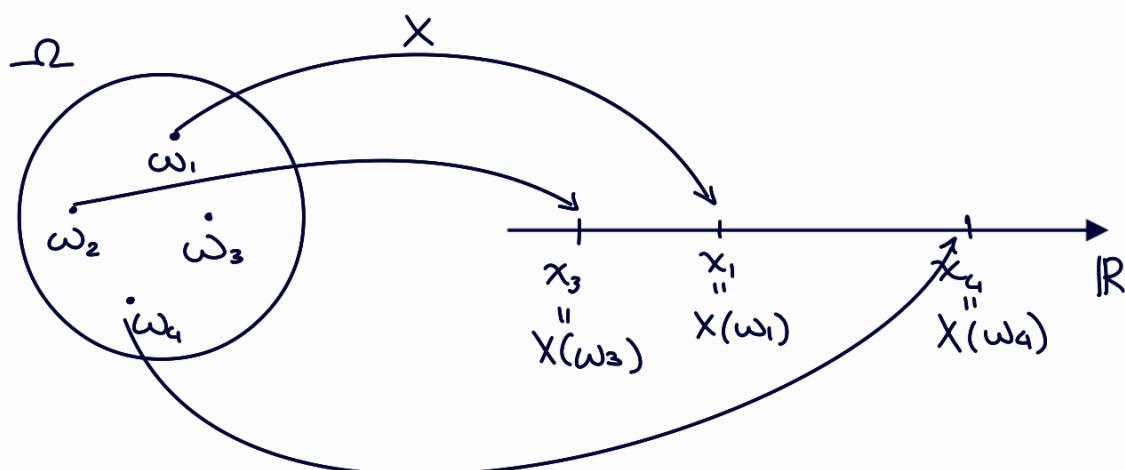
INFERENZA = processo che ci porta a fare affermazioni riguardanti l'intera popolazione di interesse e non solo il campione casuale

ESTRARE INFORMAZIONI
DAI DATI

! VARIABILITÀ DEL CAMPIONE CASUALE !

↓
Lo voglio controllare → **COME ?**

NOTAZIONE



n = TAGLIA DEL CAMPIONE

$$\begin{aligned} x_1 &= X(\omega_1) \\ x_2 &= X(\omega_2) \\ &\vdots \\ x_n &= X(\omega_n) \end{aligned}$$

} come avere n dati identici e indipendenti che lancio una volta

ho n variabili aleatorie x_1, x_2, \dots, x_n
tutte distribuite come X e **INDIPENDENTI**

Per misurare x_1 uso X_1

" " " " x_2 " " X_2

" " " " x_n " " X_n

IL CAMPIONE CASUALE e' $(X_1, X_2, X_3, \dots, X_n)$

→ ho n numeri x_1, \dots, x_n che sono la realizzazione di n V.A. INDEPENDENTI e IDENTICAMENTE DISTRIBUITE X_1, \dots, X_n

PARAMETRICA ?

Ipotizziamo di conoscere le distribuzioni comuni delle v.a. che stiamo campionando X_1, \dots, X_n .

(ad es. so che sono $X_i \sim \text{Exp}$)
oppure $X_i \sim \text{Bernoulli}$



rimangono incogniti solo i parametri

⇒ L'oggetto di interesse sono i PARAMETRI

Per esempio: $X_i \sim N(\mu, \sigma^2)$

voglio conoscere μ e σ^2

"l'altezza media della popolazione e' 175 cm"

$$\mu = E(X)$$

COME FARE ?

• Stimo il parametro = inferisco un valore per il parametro

PASSO 1: COSA DICO?

PASSO 2: CONTROLLO LA VARIABILITA'

STIMA PUNTUALE

① INTERVALLI DI CONFIDENZA

② TEST DI IPOTESI

STIMA PUNTUALE: Siamo oggetti che si chiamano STIMATORI

Definizione • **STIMATORE (STATISTICA)** = una qualunque funzione calcolabile del campione casuale

devo poter ottenere un numero
⇒ NO INCognITE
⇒ NO PARAMETRI !

ESEMPI

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

MEDIA CAMPIONARIA

e' \downarrow uno stimatore !

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

VARIANZA CAMPIONARIA

● COME SUCCIDE CHE GLI STIMATORI SONO LEGATI AI PARAMETRI?

$X \sim \text{Binomiale}(n, p)$

$$E(X) = np$$

$$Var(X) = np(1-p)$$

$X \sim Exp(\lambda)$

$$E(X) = \frac{1}{\lambda}$$

$$Var(X) = \frac{1}{\lambda^2}$$

In realtà ci troveremo a fare la stima NON dei parametri, bensì delle funzioni dei parametri (es. media e varianza)

Se ad esempio prendo $X \sim \text{Bernoulli}(p)$ PROPORTIONE
cercherò di stimare p

► Perche' \bar{X}_n e S^2 mi restituirebbero valori vicini a $E(X)$ e $Var(X)$?



HO DEI TEOREMI CHE LO GARANTISCONO

- ① LEGGI DEI GRANDI NUMERI
- ② TEOREMI DEL LIMITE CENTRALE

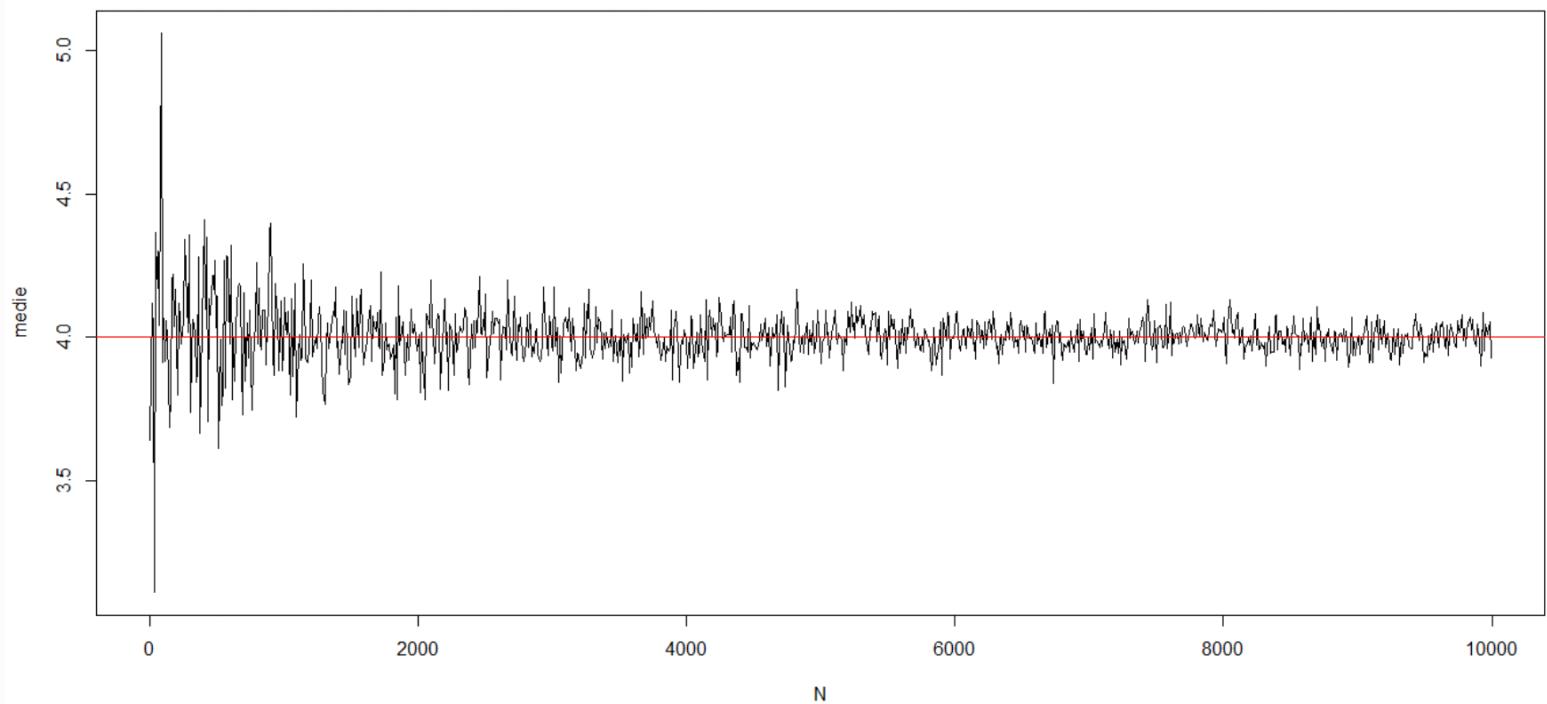
LEGGI DEI GRANDI NUMERI

- Se $(X_i)_{i=1}^{+\infty}$ è collezione di V.A. INDEPENDENTI e IDENTICAMENTE DISTRIBUITE (IID) ALLORA :

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow[n \rightarrow +\infty]{} E(X_i)$$

$$E(X_i) \quad \forall i = 1, \dots, n$$

Vedi codiceLezioniR → leggi-grand-numeri.R



Osservazione:

$$\mathbb{E}(\bar{X}_n) = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}(X_i) = \mathbb{E}(X_i) \quad \forall i \quad (\text{cost})$$

LINEARITÀ

$$= \frac{1}{n} \cdot n \mathbb{E}(X_1) = \mathbb{E}(X_1)$$

\bar{X}_n e' STIMATORE CORRETTO (UNBIASED) PER LA
MEDIA DELLA POPOLAZIONE

IN GENERALE: se T e' STIMATORE per θ e $\mathbb{E}(T)=\theta$
allora dico che T e' STIMATORE CORRETTO
per θ

Osservazione:

$$\text{Var}(\bar{X}_n) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \text{Var e' quadratica}$$

\times le cost. moltiplicative

$$= \frac{1}{n^2} \cdot \text{Var}\left(\sum_{i=1}^n X_i\right)$$

\downarrow
 X_i indipendenti

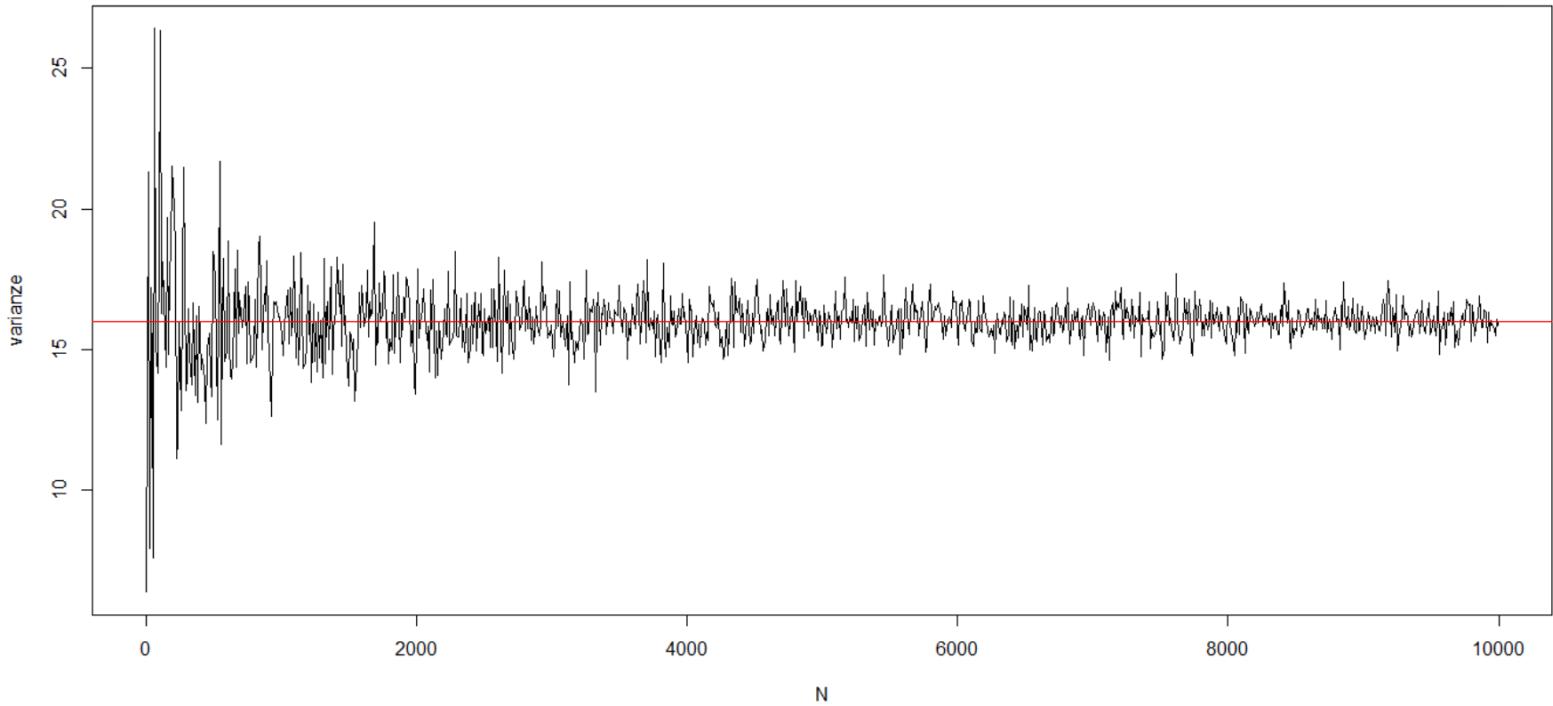
$$\Rightarrow \text{Var}(\Sigma) = \sum \text{Var}$$

$$= \frac{1}{n^2} \cdot \sum_{i=1}^n \text{Var}(X_i) = \frac{1}{n^2} \cdot n \text{Var}(X_1)$$

$$= \frac{1}{n} \text{Var}(X_1)$$

\nearrow

$$\Rightarrow \lim_{n \rightarrow +\infty} \text{Var}(\bar{X}_n) = \lim_{n \rightarrow +\infty} \frac{1}{n} \text{Var}(X_1) = 0$$



► $\mathbb{E} \bar{X}_n$
 Le media delle media campionaria e' costante \bar{h}_n ed e'-
 uguale alla media della popolazione.
 $\mathbb{E}(X_1)$

La varianza delle medie campionarie va a zero
 per $n \rightarrow +\infty$

• Sia $(X_i)_{i=1}^{+\infty}$ una collezione di v.a. iid

allora :

$$S^2 = \frac{1}{n-1} \cdot \sum_{i=1}^n (X_i - \bar{X}_n) \xrightarrow{n \rightarrow +\infty} \text{Var}(X_1)$$

$\bar{X}_n \rightarrow E(X_1)$ vuol dire che \bar{X}_n e' STIMATORE
CONSISTENTE per $E(X_1)$

$S^2 \rightarrow \text{Var}(X_1)$ vuol dire che S^2 e' STIMATORE
CONSISTENTE per $\text{Var}(X_1)$

- RIESCO A QUANTIFICARE QUANTO SONO VICINO
A $E(X_1)$ o $\text{Var}(X_1)$?



TEOREMI DEL LIMITE CENTRALE

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

$$P\left(\frac{\bar{X}_n - E(X_1)}{\frac{\text{Stdev}(X_1)}{\sqrt{n}}} \leq x\right) \xrightarrow{n \rightarrow +\infty} \underline{P(z \leq x)}$$

dove $Z \sim N(0,1)$

funzione di
distribuzione
(CDF)

e' una v.a.

- Per n grande posso usare la CDF della $N(0,1)$ per calcolare probabilita' che riguardano

$$\frac{\bar{X}_n - E(X_1)}{\frac{\text{Stdev}(X_1)}{\sqrt{n}}}$$

se n grande
e' circa distribuito come $Z \sim N(0,1)$

COSA ABBIAMO FATTO A \bar{X}_n ?

1) $\bar{X}_n - \mathbb{E}(\bar{X}_n)$

se ho Y e considero $Y - \mathbb{E}(Y)$

ho media 0 . Infatti: $\mathbb{E}[Y - \mathbb{E}(Y)] = \mathbb{E}(Y) - \mathbb{E}(Y) = 0$

2) $\frac{\text{StDev}(\bar{X}_n)}{\sqrt{n}}$

ha il significato di essere $\text{Var}(\bar{X}_n) = \frac{1}{n} \text{Var}(X_1)$

$\text{StDev}(\bar{X}_n) = \sqrt{\text{Var}(\bar{X}_n)} = \sqrt{\frac{\text{Var}(X_1)}{n}} = \frac{\text{StDev}(X_1)}{\sqrt{n}}$

Oss: $\text{Var}\left(\frac{Y}{\text{StDev}(Y)}\right) = \frac{1}{\text{Var}(Y)} \cdot \text{Var}(Y) = 1$

1) e 2) hanno lo scopo di STANDARDIZZARE X_n

- $\mathbb{E}(\cdot) = 0$
- $\text{Var}(\cdot) = 1$

Posso leggere il TLC in due modi diversi:

• $\frac{\bar{X}_n - \mathbb{E}(X_1)}{\frac{\text{StDev}(X_1)}{\sqrt{n}}} \underset{\text{per } n \text{ grande}}{\approx} Z \sim N(0,1)$

• $\bar{X}_n \underset{\text{per } n \text{ grande}}{\approx} N\left(\mathbb{E}(X_1), \frac{\text{Var}(X_1)}{n}\right)$

Esempio: vedi codiciLezioniR → Teoremi_Limite_Centrale.R

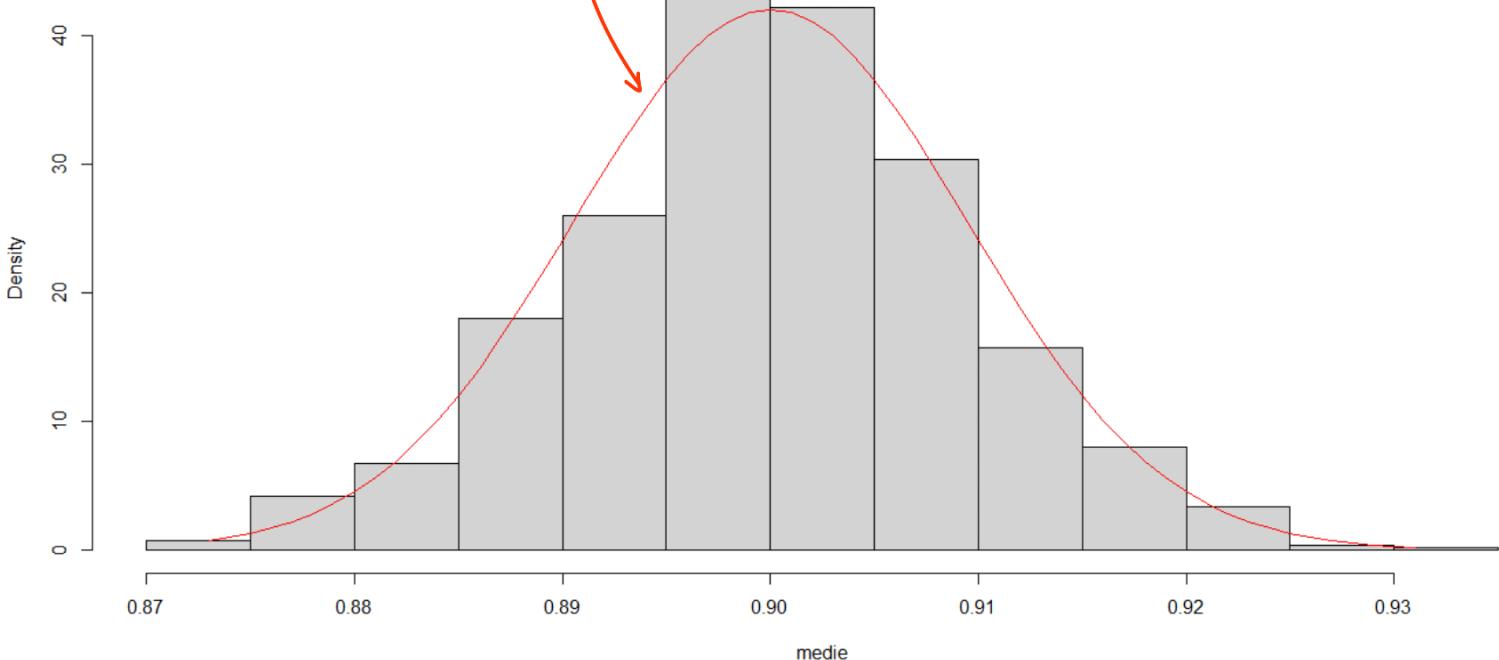
$$X_i \sim \text{Bernoulli}(p)$$

$$\mathbb{E}(X_i) = p$$

$$\text{Var}(X_i) = p(1-p)$$

$$\bar{X}_{1000} \approx N(p, \frac{p(1-p)}{n})$$

Histogram of medie

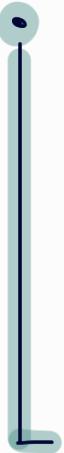


CASI PARTICOLARI → campioni da normali

TEOREMA • $(X_i)_{i=1}^{+\infty}$ collezione di v.e. IID con legge $N(\mu, \sigma^2)$

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

$$\rightarrow \bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

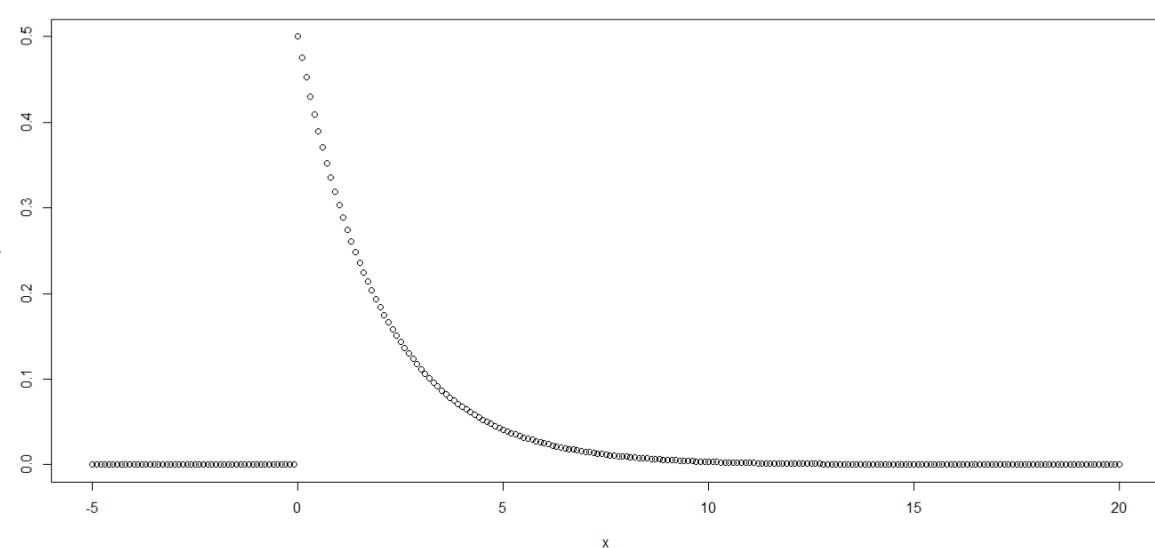
TEOREMA  $(X_i)_{i=1}^{+\infty}$ collezione di v.o. IID
con legge $N(\mu, \sigma^2)$

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

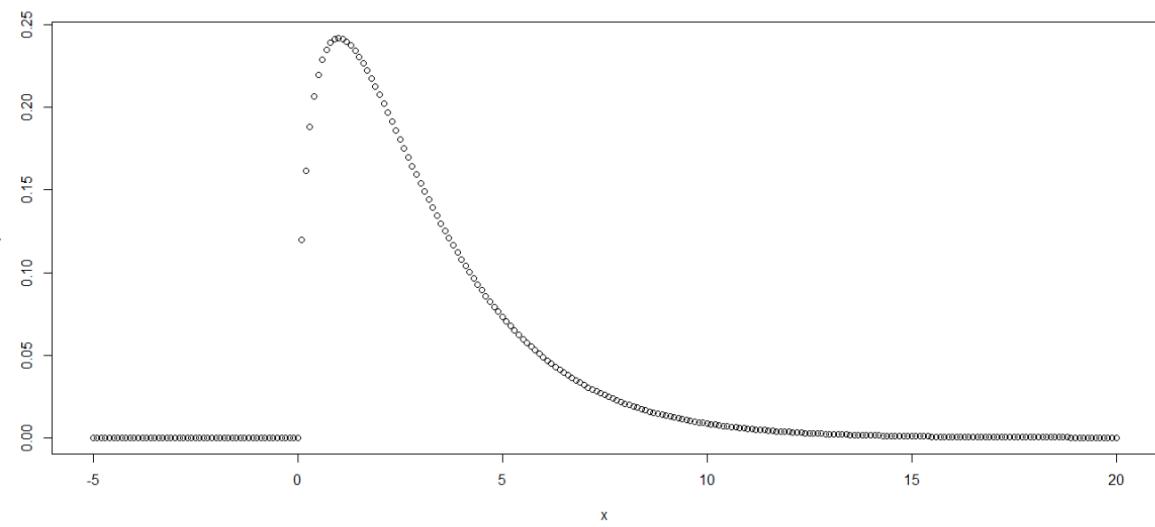
"CHI quadro"
con $n-1$ gradi
di libertà

COME E' FATTA LA $\gamma \sim \chi^2(n-1)$?

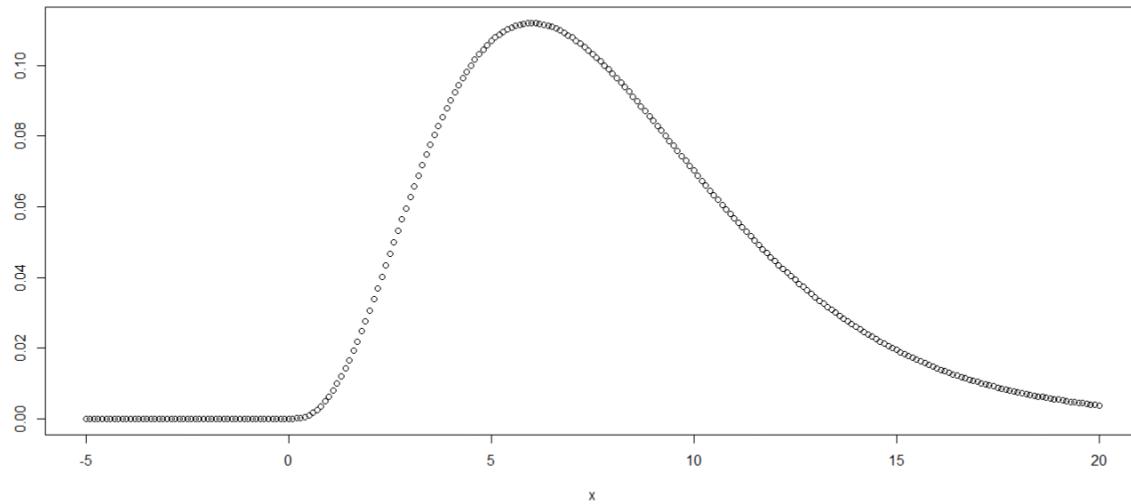
- E' UNA V.A. CONTINUA
- VADO A DISEGNARE LA PDF



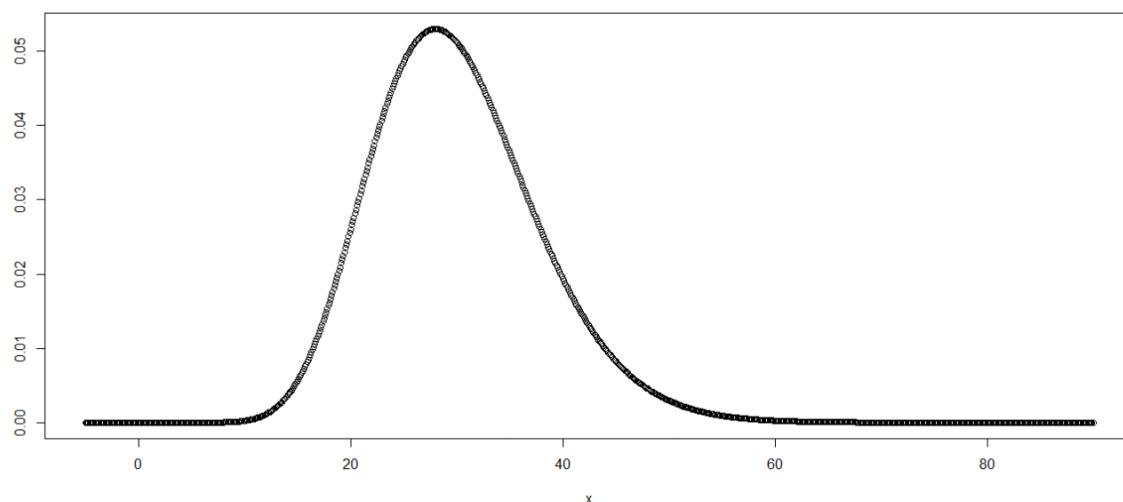
per 2 gradi
di libertà



per 3 gradi
di libertà



8 gradi di libertà



30 gradi di libertà

```
grad <- 30
x <- seq(-5,100,0.001)
y <- dchisq(x,grad)

plot(x,y)
```

All' aumentare di $n-1$ la curva diventa una campana avvicinandosi sempre più alla Normale

Gli stimatori ci restituiscono stime che ci aspettiamo vicine al parametro da stimare e che sappiamo dire quanto vicine

$$P(\bar{X}_n \in [3,5]) \underset{\text{TLC}}{\approx} P(X \in [3,5])$$

$$\text{con } X \sim N\left(\mu(X_1), \frac{\text{Var}(X_1)}{n}\right)$$

⇒ Controllo della variabilità :

- ① INTERVALLI DI CONFIDENZA
 - ② TEST DI IPOTESI

INTERVALLI DI CONFIDENZA

vedi CodiceLezioni.R → intervalli -
confidenza.R

IDEA: invece che dare solo la stima puntuale del
parametro, restituisco due cose:

$$[a, b] \quad + \quad (1 - \alpha)$$

UN INTERVALLO CONFIANOZA $\in [0, 1]$

LI LEGGO COSÌ: le IP che l'intervallo $[a,b]$ contiene il valore vero del parametro e- almeno pari a $(1-\alpha)$

$$\mathbb{E}(X_1) = \mu \quad [a, b] \quad (1-\alpha)$$

$$P(a \leq \mu \leq b) \geq 1-d$$

le vorrei grande vicine a 1

SI CALCOLA SU 

↓

sulle slides delle prof. Sirovich c'e' un esempio di solvola 2022 B

(lezione 1 dic. 22)
corso A

①

METODO DELLA QUANTITA' PIVOTALE

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

Q.TA' PIVOTALE
"

funzione del campione casuale
e dei parametri incogniti.

So come e' distribuita

$$s = \sqrt{s^2} \rightarrow \text{VARIANZA CAMPIONARIA}$$

↓
dev. Standard CAMPIONARIA

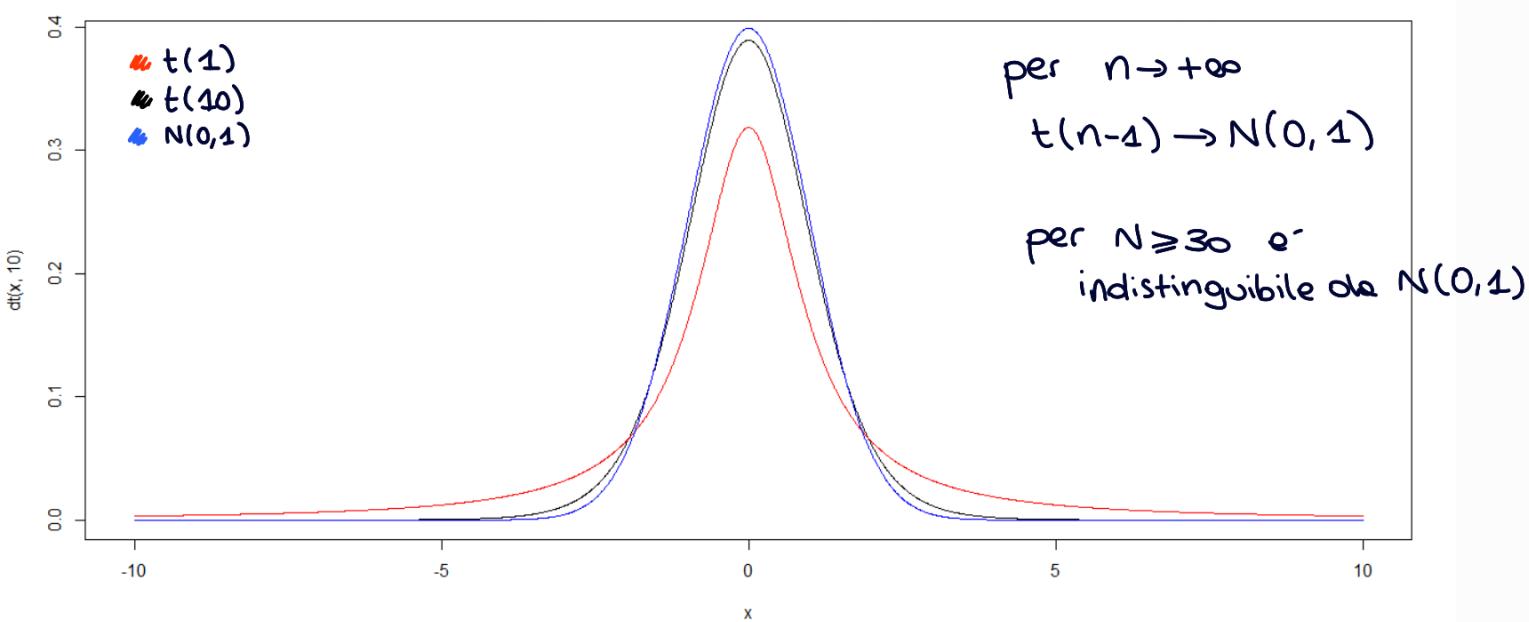
NON E' PIU' UNA $N(0,1)$ E HO BISOGNO DI CAPIRE
COM'E' DISTRIBUITA

TEOREMA 1 • Se $(x_1, \dots, x_n) \sim N(\mu, \sigma^2)$

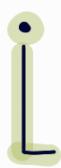
$$\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \sim t(n-1)$$

e' una t di Student
con $n-1$ gradi di liberta'

DISEGNO LA PDT SU \mathbb{R}



TEOREMA 2 Per n grande : (≥ 30)



$$\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \approx t(n-1) / N(0,1)$$

circa distribuita come

- [Se n e' grande $\rightarrow \approx t$ di Student
- [Se n piccolo \rightarrow Devo poter ipotizzare che $X_i \sim N(\mu, \sigma^2)$

② Cerco q_1 e q_2 t.c.

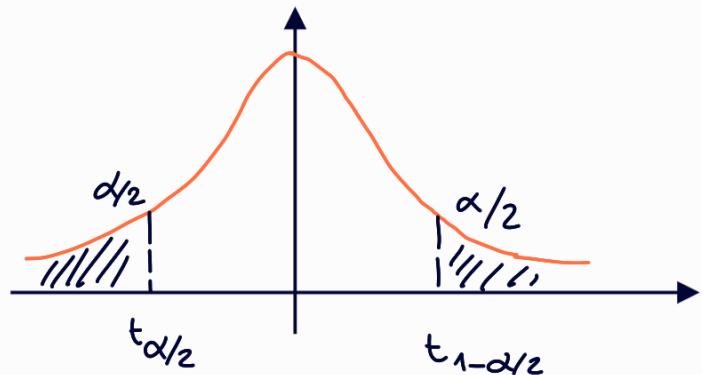
$$P\left(q_1 < \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} < q_2\right) = 1 - \alpha$$

↓ Quantili

li prendo dalla t.Student

es. fisso $1 - \alpha = 0.95$
 $\alpha = 0.05$

li prendo simmetrici:



Supponiamo di avere $n=10$ osservazioni :

```
> #t_alpha/2
> qt(alpha/2, 9)
[1] -2.26157163
> #t_(1-alpha/2)
> qt(1-alpha/2, 9)
[1] 2.26157163
```

$$t_{\alpha/2} = -2.26 \quad t_{1-\alpha/2} = 2.26 \quad) \text{ sono simmetrici}$$

③ ESPLICATO μ : $t_{\alpha/2} \leq \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \leq t_{1-\alpha/2} \Rightarrow a \leq \mu \leq b$

$$\rightarrow a = \bar{X} - t_{1-\alpha/2} \frac{s}{\sqrt{n}} \quad b = \bar{X} - t_{\alpha/2} \frac{s}{\sqrt{n}} \quad) \text{ R fa questo}$$

t. test

Vedi esempio R

Vedi eserciziR → intervalli-confidenza.R (fino a 8.18)

IC PER ALTRI PARAMETRI

1) IC PER PROPORZIONI

$$X_i \sim \text{Bernoulli}(p)$$

PROB. DI SUCCESSO

$$\begin{bmatrix} 1 & p \\ 0 & 1-p \end{bmatrix}$$

Sono IC per il parametro p quando campioni dalle Bernoulli

① STIMA PUNTUALE PER p $\rightarrow \frac{\sum_{i=1}^n X_i}{n} = \frac{1}{n} \sum_{i=1}^n X_i$ medio campionario

② Q.TA' PIVOTALE $E(X_i) = p$
 $\text{Var}(X_i) = p(1-p)$) non efficiente

$$\Rightarrow \frac{\bar{X}_1 - p}{\sqrt{\frac{p(1-p)}{n}}} \rightarrow \text{IC} \text{ tramite R con binom.test}$$

Vedi eserciziR → intervalli-confidenza.R (8.6 e 8.9)

2) IC PER DIFFERENZE DI MEDIE

↓
e' il parametro

Voglio confrontare medie di una variabile in diverse situazioni



CAMPIONI INDEPENDENTI

→ confronto una variabile misurata su soggetti diversi

es. temperatura corporea in maschi e femmine

CAMPIONI APPAIATI

→ misura una stessa q.ta sullo stesso soggetto ma in tempi / condizioni diverse

es. peso di un soggetto prima e dopo una dieta

X Y

μ_x μ_y

$(\mu_x - \mu_y)$ diff. di medie

Devo fare attenzione nel dire a R se considerare i campioni appaiati o indipendenti

Uso t.test con parametri differenti

Vedi eserciziR → intervalli-confidenza.R (ultimi due)

TEST DI IPOTESI

Sono delle procedure che ci permettono di testare IPOTESI

↓
Sono affermazioni/proposizioni
sui PARAMETRI
 $\left(\begin{array}{l} \text{- media} \\ \text{- proporzioni} \\ \text{- differenza di medie} \end{array} \right)$

es. affermo che
• $\mu = 5$
• $\mu > 5$
• $p < 0.1$
• $\mu_1 - \mu_2 = 0$

► Si considerano SEMPRE DUE IPOTESI:

- IPOTESI NULLA (H_0) → è l'hp alla quale credi
- IPOTESI ALTERNATIVA (H_1 o H_a)

CREDO IN H_0 , RACCOLGO I DATI E FACCIO UN TEST DI HP
PER CAPIRE SE POSSO CONTINUARE A CREDERE IN H_0
(ALLA LUCE DEI DATI RACCOLTI) OPPURE SE SONO COSTRETTA
AD ABBANDONARE/RIFIUTARE H_0

IN FAVORE DI H_1 , OVVERO
L'IPOTESI ALTERNATIVA

che avevo a mente
fin dall'inizio

$$H_0 \neq H_1$$

↓
"I DATI PRODUCONO
SUFFICIENTE EVIDENZA
STATISTICA"

- esempio:
- $H_0: \mu = 5$
 - $H_1: \mu > 5$
- ONE-SIDED
(TEST AD UNA CODA)

- $H_0: p = 0.1$
 - $H_1: p \neq 0.1$
- TWO-SIDED
(TEST A DUE CODE)

TEST SULLA MEDIA DELLA POPOLAZIONE : μ

$$H_0: \mu = \mu_0 \quad \text{dove } \mu_0 \text{ e' un valore (es. } \mu_0 = 5)$$

$$H_1: \mu \neq \mu_0$$

$\mu > \mu_0 \quad \textcircled{B}$ → vediamo questo

$\mu < \mu_0 \quad \textcircled{C}$

Considero una STATISTICA DEL TEST
e le calcolo sui miei dati e posso dire se assume un valore probabile o improbabile

funzione del campione casuale

Q.TA' CHE SO CORRE E' DISTRIBUITA SOTTO H_0
(CONSIDERO H_0 VERO)

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \approx N(\mu_0, \frac{\sigma^2}{n})$$

con σ^2 noto

Io calcolo: ad es. $\bar{X}_n = 7.32$

e' un valore probabile o improbabile sotto H_0 ?

Facciamo un disegno

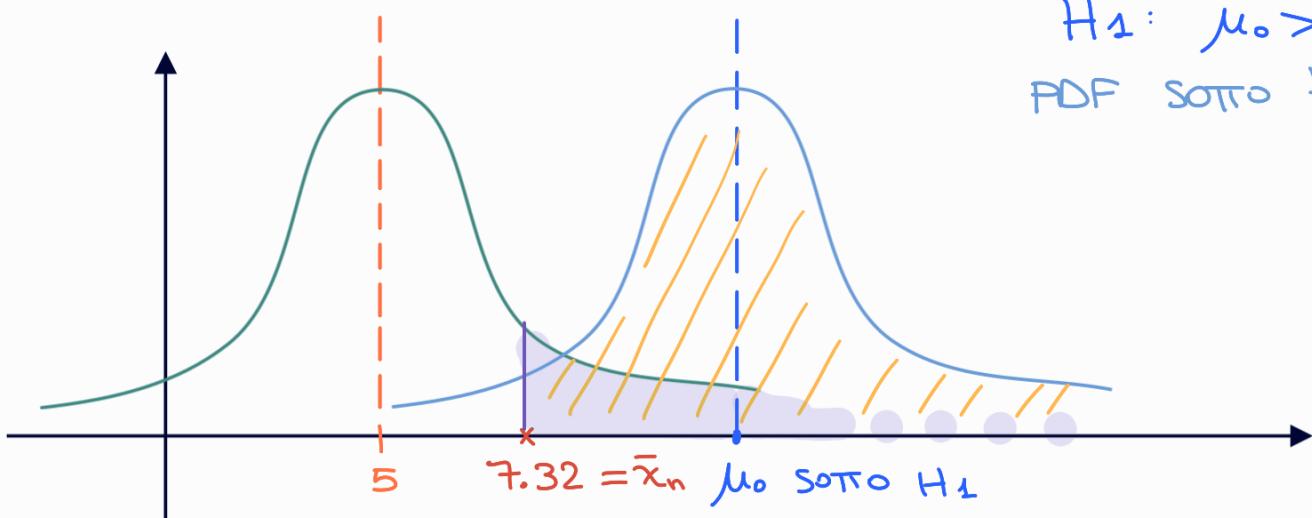


per $\mu_0 = 5$:

mi aspetto un oggetto con asse di simmetria 5

PDF di \bar{X}_n SOTTO H_0

$H_1: \mu > 5$
PDF SOTTO H_1



QUANTO È PROBABILE?

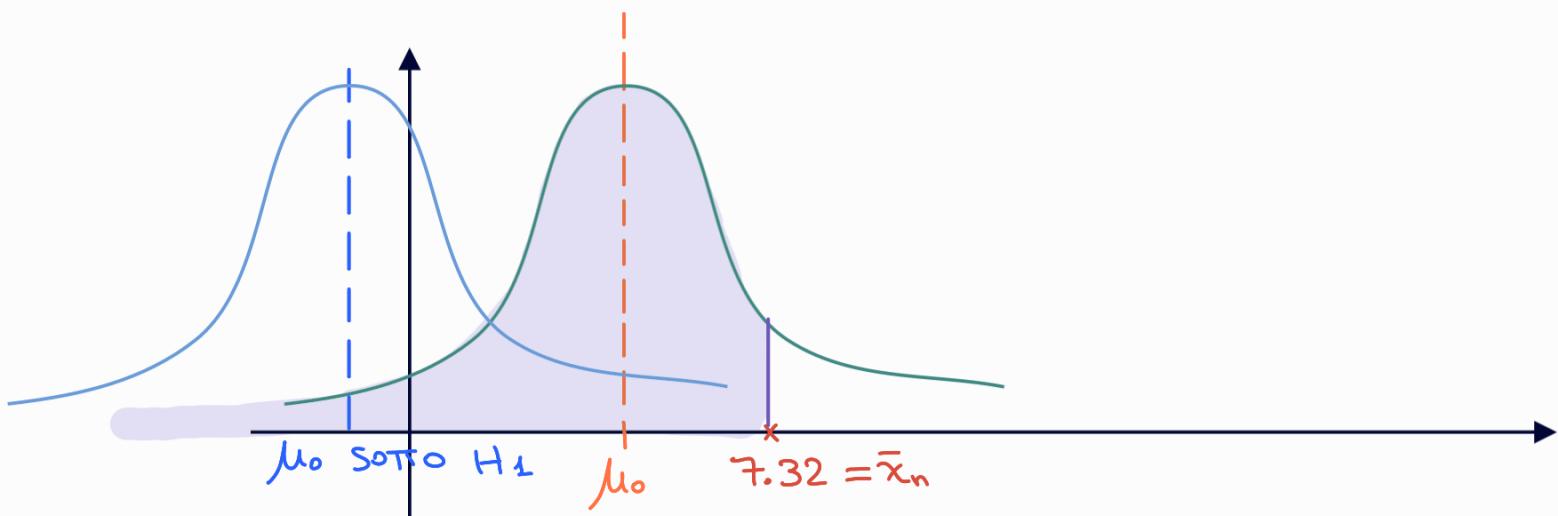
$P(\bar{x}_n > 7.32) = p$ → se l'area p è troppo piccolo
e' meglio H_1 //

più estremo di
quello campionario
//
PIÙ PROBABILE

SE CAMBIO L'ALTERNATIVA:

$$H_0: \mu = \mu_0$$

$$H_1: \mu < \mu_0 \rightarrow \text{il disegno e' al contrario}$$



$P(\bar{x}_n < 7.32) = p$ è grande ⇒ non abbandono H_0

INVECE se faccio TEST A DUE CODE :

$$H_0: \mu = \mu_0$$

$$H_1: \mu \neq \mu_0$$



calcolo le IP che il valore
sia troppo estremo a dx
o sx

→ abbandono H_0 per
valori che sono troppo
a dx o troppo a sx

t.test e binom.test
l'oggetto che calcola **R** e' **P** (**P-VALUE**)

- Se P-VALUE e' troppo piccolo RIFIUTO H_0

$< \alpha$

(VALORE FISSATO PICCOLO)
= LIVELLO DI SIGNIFICABILITÀ

Vedi eserciziR → test_ipotesi.R