Corso di Statistica
Stima intervallare

Domenico De Stefano

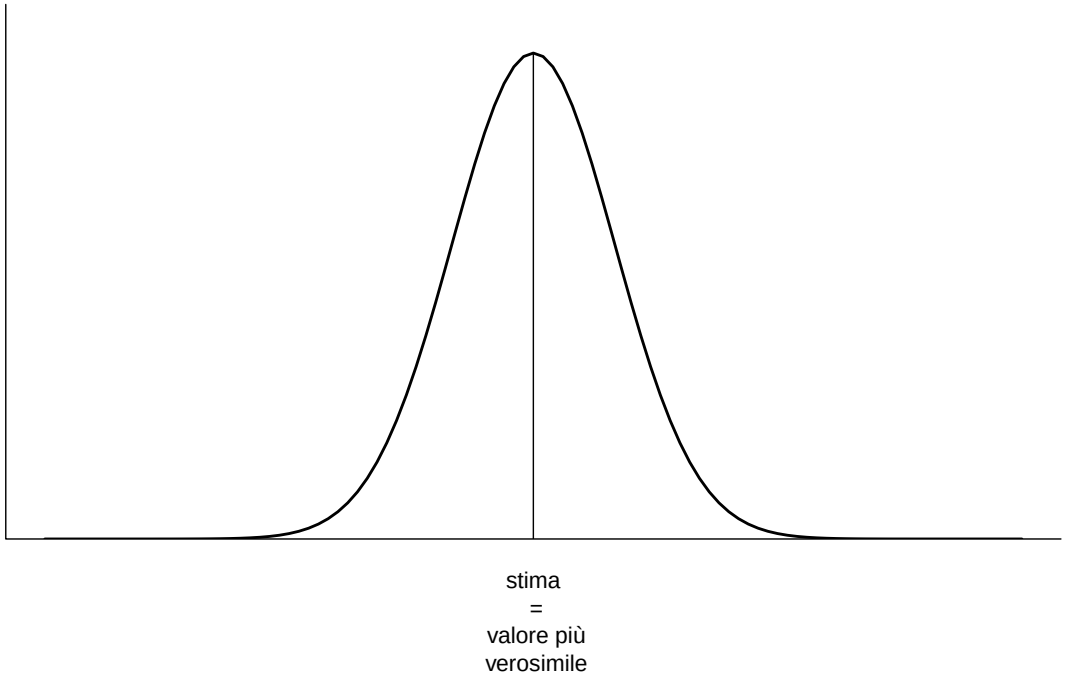
a.a. 2016/2017

Indice

Stima puntuale e intervallare

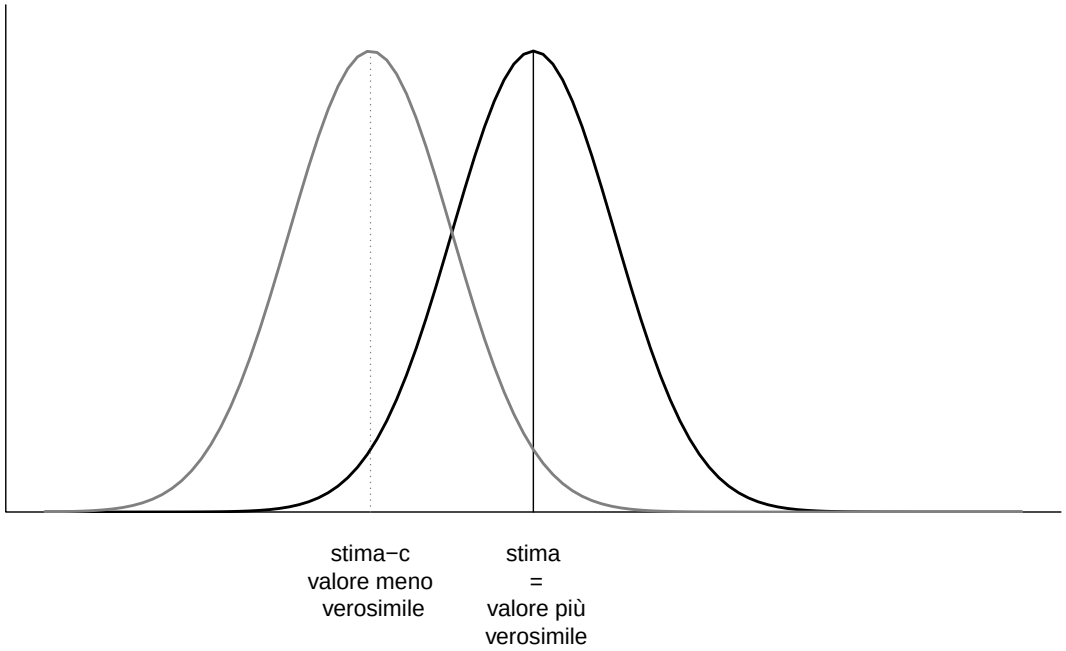
- Nella stima puntuale, calcoliamo dal campione un valore plausibile per il parametro.
- Il c.d.p. ci dice che quel valore è probabilmente vicino al vero valore del parametro.
- Ci dice qualcosa di più, perché ci dice, attraverso la distribuzione campionaria, quanto distante.
- Possiamo usare questa informazione per ottenere una stima intervallare.
- Un intervallo è un modo di esprimere stima e incertezza insieme.

Idea generale I



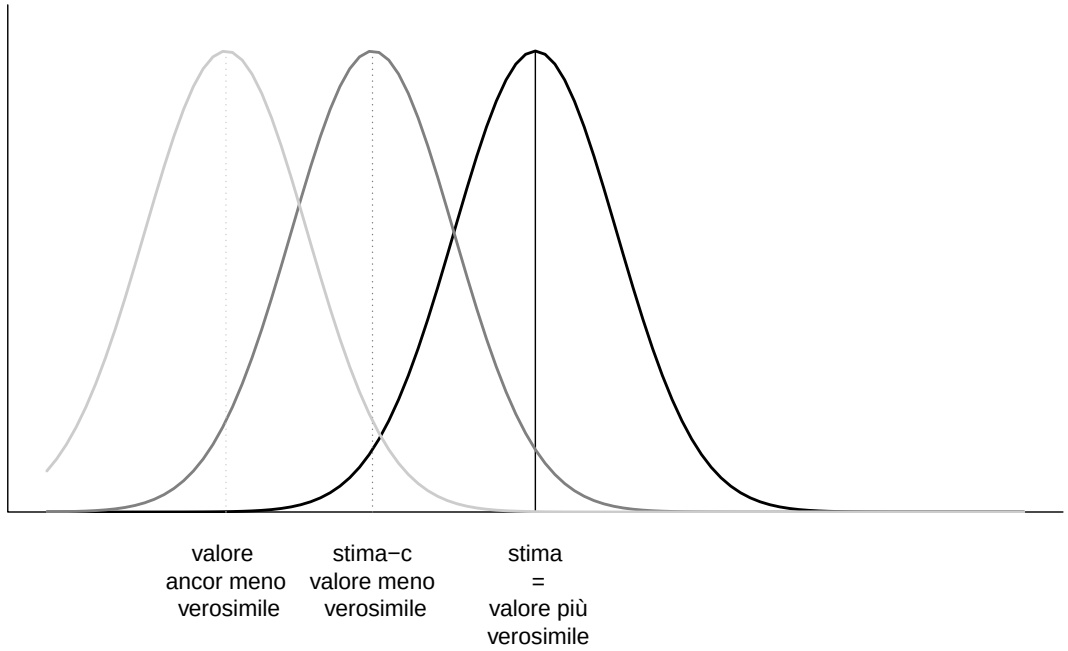
- Dal campione si calcola la stima, che è il valore più verosimile alla luce del campione.
- Sappiamo, un po' vagamente, che il parametro "non è tanto distante" dalla stima, probabilmente.

Idea generale II



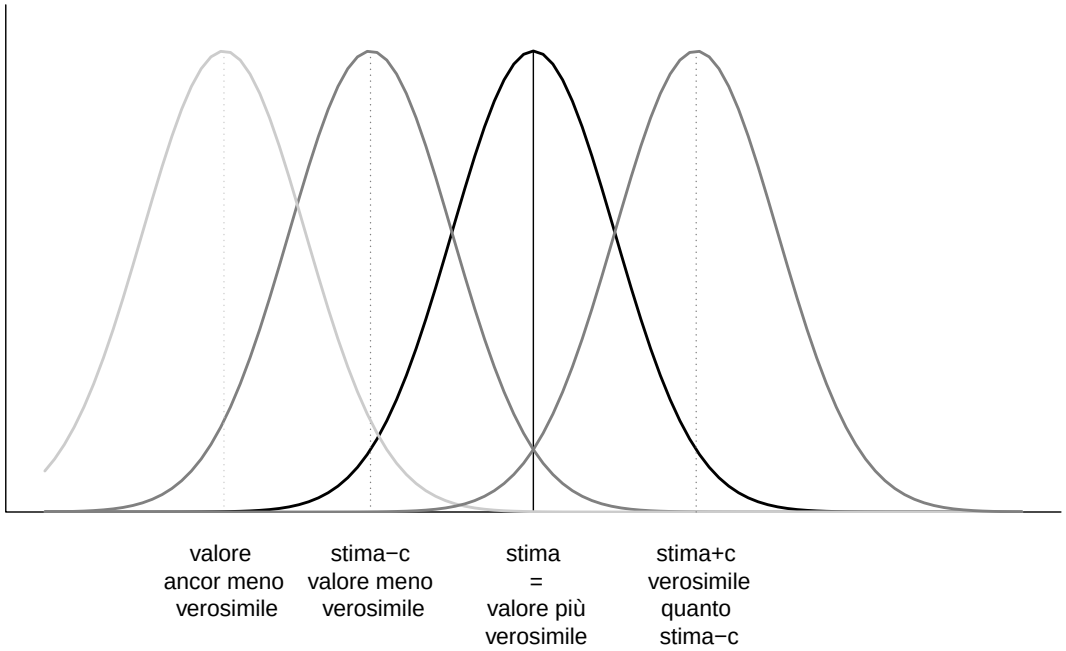
- Se ci discostiamo dalla stima, in stima $- c$ per esempio, il valore è meno verosimile.

Idea generale III



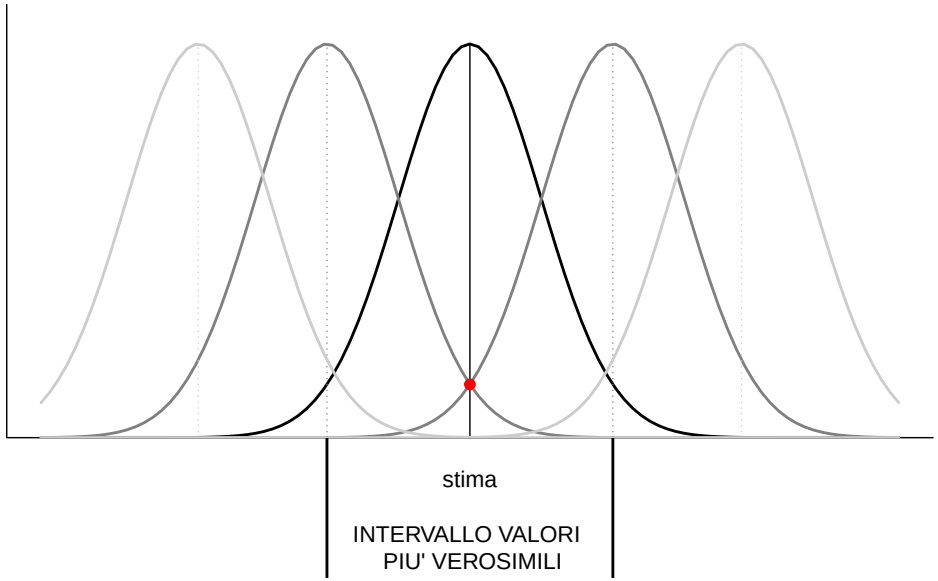
- Se ci discostiamo dalla stima, in stima $-$ c per esempio, il valore è meno verosimile.
- Più ci spostiamo meno verosimile è il valore.

Idea generale IV



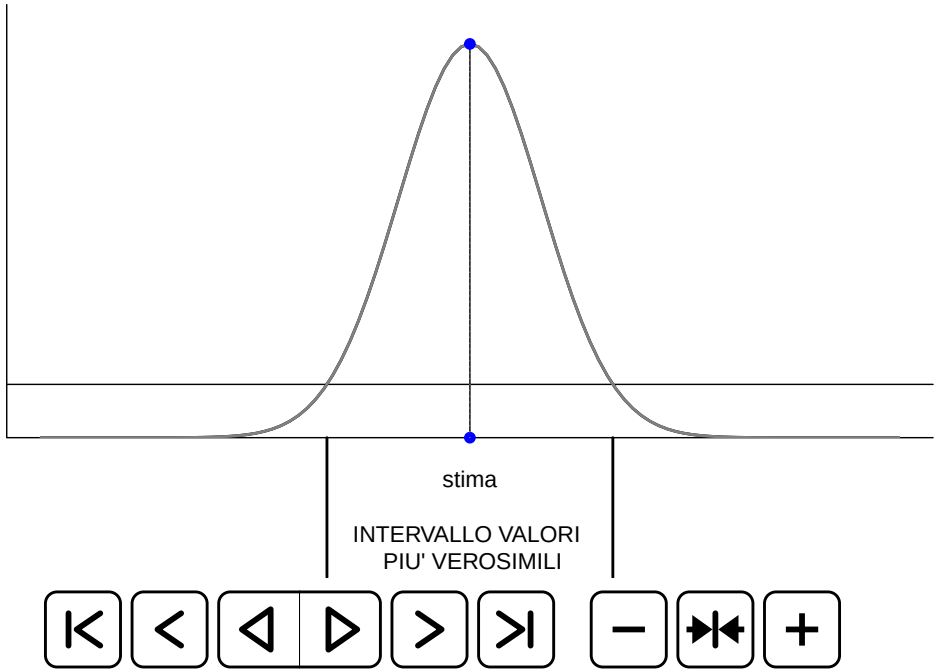
- Se ci discostiamo dalla stima, in stima $- c$ per esempio, il valore è meno verosimile.
- Più ci spostiamo meno verosimile è il valore.
- Se ci spostiamo dall'altra parte, accade lo stesso, simmetricamente.

Idea generale V



Ha senso quindi determinare un intervallo di valori, più verosimili degli altri.

Idea generale VI



Stima puntuale

Media μ di una popolazione

- Campione $X_1, \ldots, X_n$.
- Stimatore (media campionaria): $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$.
- Definiamo anche la varianza campionaria

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

Proporzione π di una popolazione

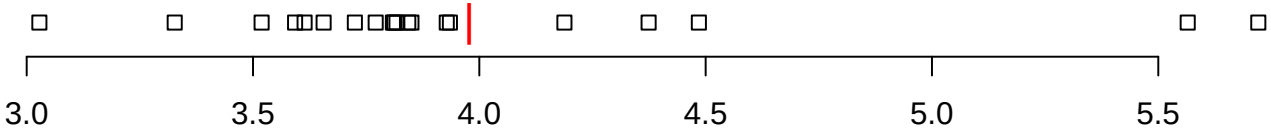
- Campione $X_1, \ldots, X_n$.
- Stimatore (proporzione campionaria): $\hat{\pi} = \frac{1}{n}\{\#X_i = 1\}$.

Indice

Indice

Dati: tempi alla Bavisela

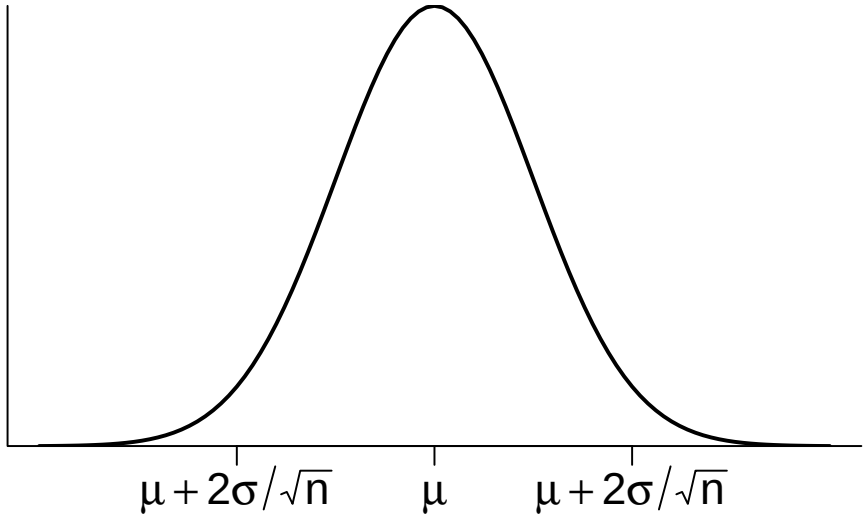
Disponiamo dei tempi registrati, in ore, da 20 dei concorrenti maschi alla Bavisela 2012, il parametro d'interesse è il tempo medio registrato tra i 695 concorrenti.



La media campionaria 3.978 (3:58:40.8) è la stima della media nella popolazione (=parametro).
Assumiamo che i tempi, nella popolazione, siano distribuiti secondo una $N(\mu, \sigma^2)$ (il parametro è μ).
Assumiamo che il campione sia costituito da individui presi a caso e indipendenti (non un gruppo di amici o gli iscritti a una data società sportiva).

Distribuzione campionaria del tempo medio (campionario)

La distribuzione della media campionaria è una $N(\mu, \sigma^2/n)$. Non nota perché dipende da due parametri della popolazione: la media μ e la varianza σ^2.

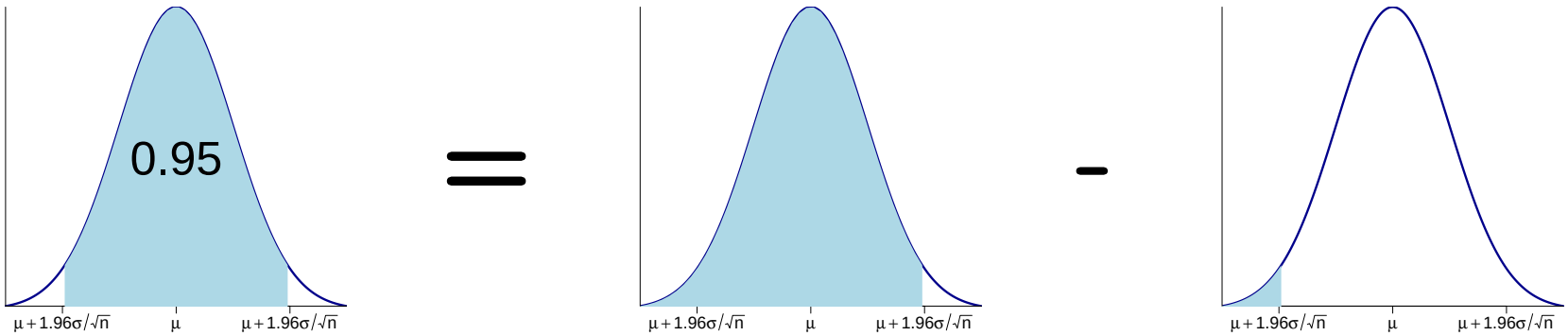


A significare che valori vicini a μ sono più probabili che valori lontani. (N.B.: questa affermazione richiede che ci sia indipendenza e identica distirbuzione.)

Distribuzione campionaria e probabilità

Quanto vicina è la media campionaria al parametro (media della popolazione)?

$$P(\mu - 1.96\sigma/\sqrt{n} < \bar{X} < \mu + 1.96\sigma/\sqrt{n}) = 0.95$$

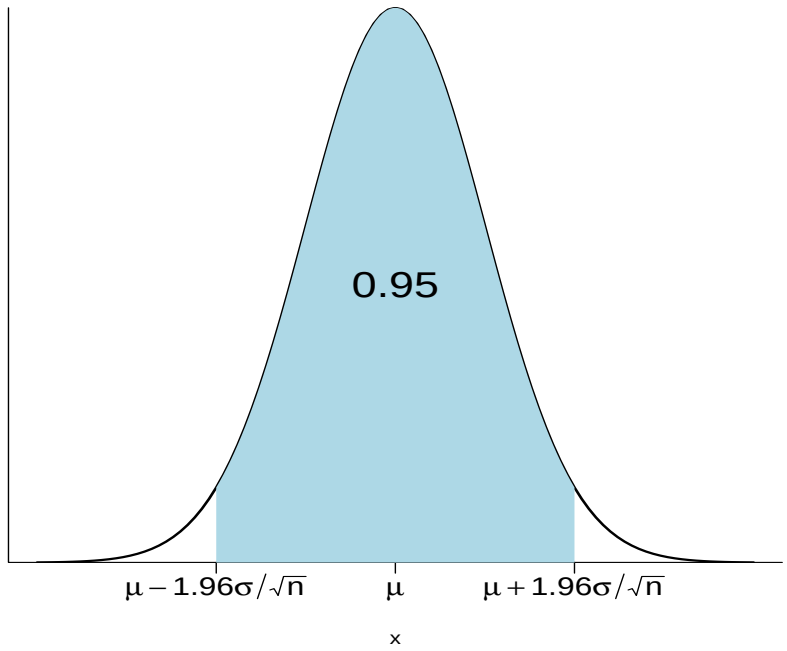


$$0.95 = \Phi(1.96) - \Phi(-1.96)$$

Distribuzione campionaria e probabilità

Quanto vicina è la media campionaria al parametro (media della popolazione)?

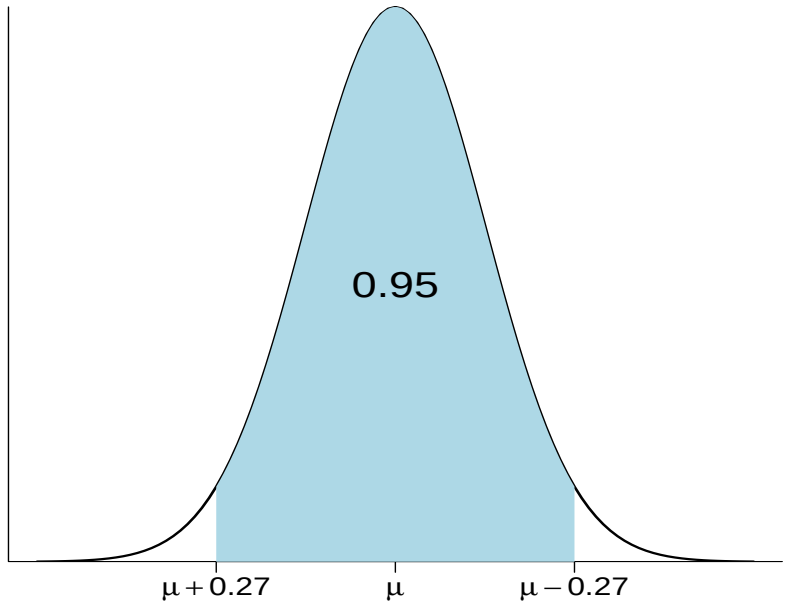
$$P(\mu - 1.96\sigma/\sqrt{n} < \bar{X} < \mu + 1.96\sigma/\sqrt{n}) = 0.95$$



Distribuzione campionaria e probabilità

Supponiamo di conoscere $\sigma = 0.602$, sostituiamo

$$P(\mu - 0.27 < \bar{X} < \mu + 0.27) = 0.95$$



C'è il 95% di probabilità che la media campionaria cada entro 0.27 dal parametro (media della popolazione).

Intervallo di confidenza

Questa affermazione ci dice che media campionaria aspettarci conoscendo il parametro (media della popolazione)

$$P(\mu - 0.27 < \bar{X} < \mu + 0.27) = 0.95$$

possiamo pero rovesciarla e scrivere

$$P(\bar{X} - 0.27 < \mu < \bar{X} + 0.27) = 0.95$$

e reinterpretarla scrivendo che

C'è una probabilità del 95% che l'intervallo

$$[\bar{X} - 0.27, \bar{X} + 0.27]$$

includa il vero valore della media, μ.

Intervallo di confidenza: campionamento ripetuto

C'è una probabilità del 95% che l'intervallo

$$[\bar{X} - 0.27, \bar{X} + 0.27]$$

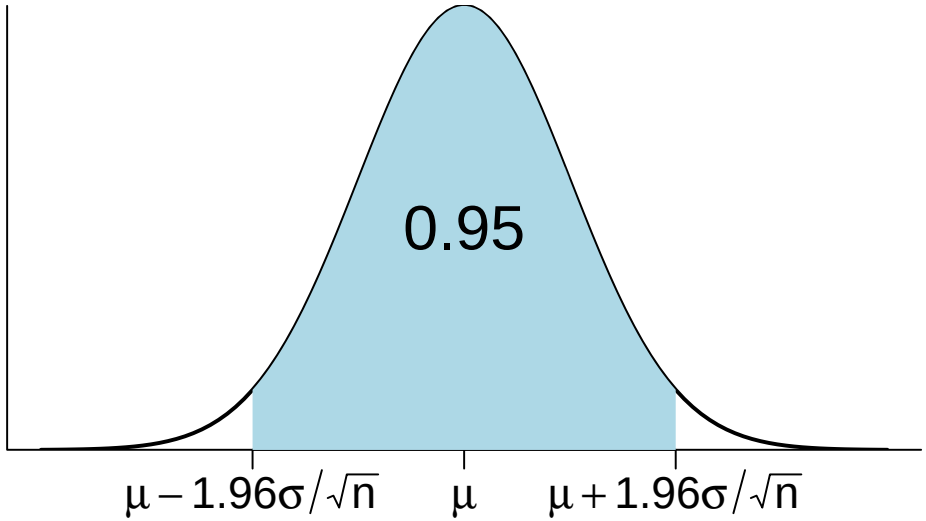
includa il vero valore della media, μ.

Possiamo interpretare questa affermazione in questi termini

- ripetiamo 1000 volte il campionamento
- a ogni campione corrisponde una diversa media $\bar{X}$ e quindi un diverso intervallo $[\bar{X} - 0.27, \bar{X} + 0.27]$
- dei 1000 intervalli, mediamente il 95% includono il vero valore del parametro (media della popolazione).

Intervallo di confidenza: campionamento ripetuto

Cambiamo il livello



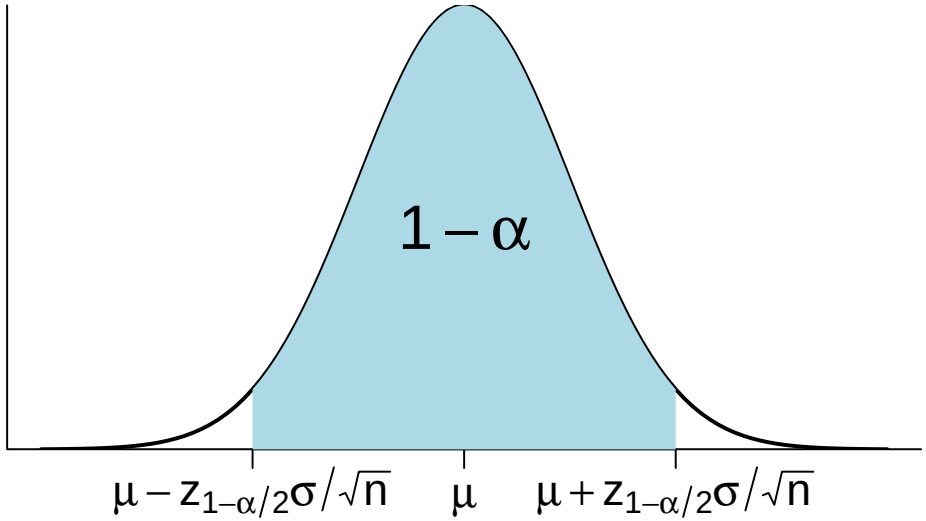
L'intervallo

$$\bar{X} \pm 1.96\sigma/\sqrt{n}$$

contiene μ con probabilità 95% in quanto

$$P(\mu - 1.96\sigma/\sqrt{n} \leq \bar{X} \leq \mu + 1.96\sigma/\sqrt{n})$$

Cambiamo il livello



L'intervallo

$$\bar{X} \pm z_{1-\alpha/2}\sigma/\sqrt{n}$$

contiene μ con probabilità $1-\alpha$ in quanto

$$P(\mu - z_{1-\alpha/2}\sigma/\sqrt{n} \leq \bar{X} \leq \mu + z_{1-\alpha/2}\sigma/\sqrt{n})$$

Esempio: tempi della Bavisela

Nel caso dei tempi della Bavisela

- Dal campione, di 20 osservazioni, otteniamo la media 3.978;
- La varianza è nota (calcolata sull'intera popolazione), $\sigma = 0.602$;
- indicando con *se* l'errore standard dello stimatore (la deviazione standard dello stimatore $\bar{X}$, ovvero radice quadrata della varianza di $\bar{X}$) si ottiene quindi

$$\text{se}(\bar{X}) = \frac{\sigma}{\sqrt{n}} = \frac{0.602}{\sqrt{20}} = 0.1346113$$

- abbiamo $z_{1-\alpha/2} = 1.96$ e otteniamo l'intervallo al 95% (quantile 0.975 della normale);
- si ottiene

$$3.978 \pm 0.2646 \rightarrow [3.7134, 4.2426]$$

- tradotto in ore:minuti:secondi si ha, per la media della popolazione, l'intervallo

$$[3:42:48.24, 4:14:33.36]$$

Indice

Intervallo di confidenza per la media I

Qui abbiamo visto un caso particolare, in generale la formula per un intervallo di confidenza per la media di una popolazione è

Intervallo di confidenza per la media

L'intervallo di confidenza di livello $1-\alpha$ per la media μ di una popolazione si ottiene come

$$\bar{X} \pm c_\alpha \text{se}(\bar{X})$$

dove

- $\bar{X}$ è la media campionaria;
- $\text{se}(\bar{X})$ è la deviazione standard di $\bar{X}$;
- c_α un coefficiente che dipende dalla distribuzione campionaria di $\bar{X}$ e dal livello scelto.

Esempio: tempi della Bavisela, varianza non nota I

- Se non conosciamo la varianza della popolazione la stimiamo a partire dal campione $X_1, \ldots, X_n$

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 = 0.4296749$$

- Si ha allora la deviazione standard stimata di $\bar{X}$

$$\text{se}(\bar{X}) = \frac{S}{\sqrt{n}} = \frac{\sqrt{0.4296749}}{\sqrt{20}} = 0.1465734$$

- il c_α è questa volta ricavato da una t di Student con $n-1$ gradi di libertà

$$t_{n-1,0.975} = 2.09$$

Esempio: tempi della Bavisela, varianza non nota II

- Si ha quindi l'intervallo

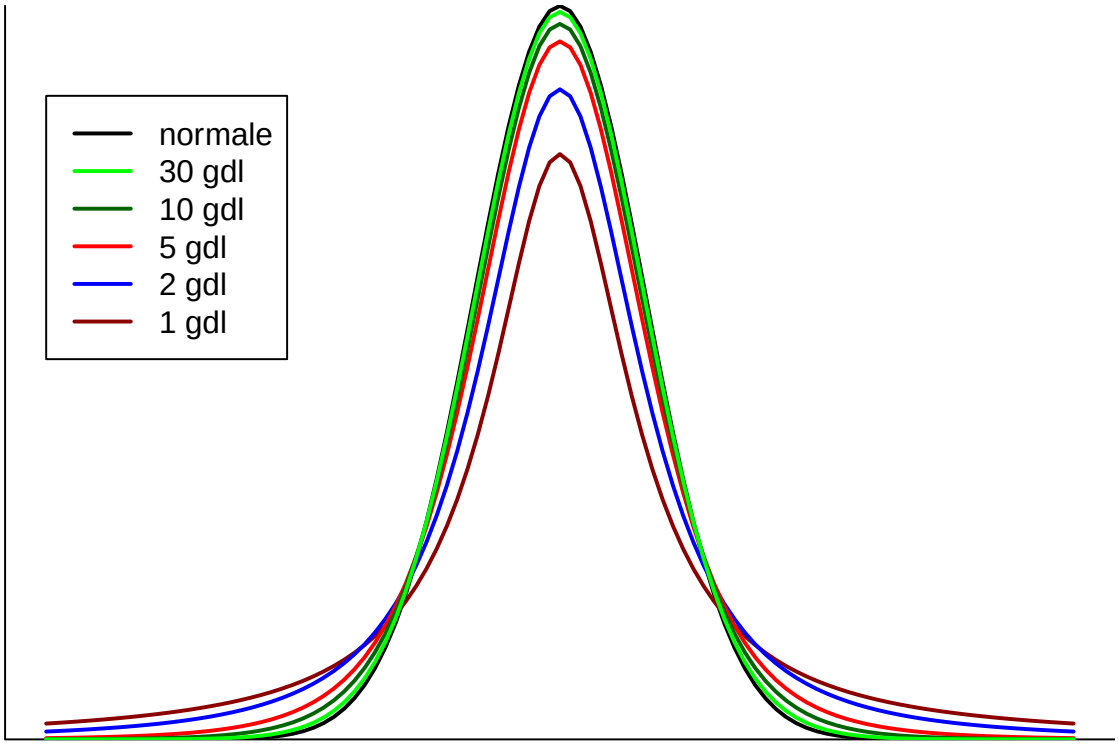
$$3.978 \pm 0.3063383 \rightarrow [3.6716617, 4.2843383]$$

- tradotto in ore:minuti:secondi si ha l'intervallo

$$[3 : 40 : 17.98, 4 : 17 : 3.62]$$

t di student I

Per affrontare il caso in cui la varianza è non nota, non useremo la normale per determinare c_α ma la t di Student.



La t è una distribuzione **simile alla normale standard**, con un parametro in più, i g.d.l.

t di student II

Anche per la t ci sono delle tavole (più semplici)

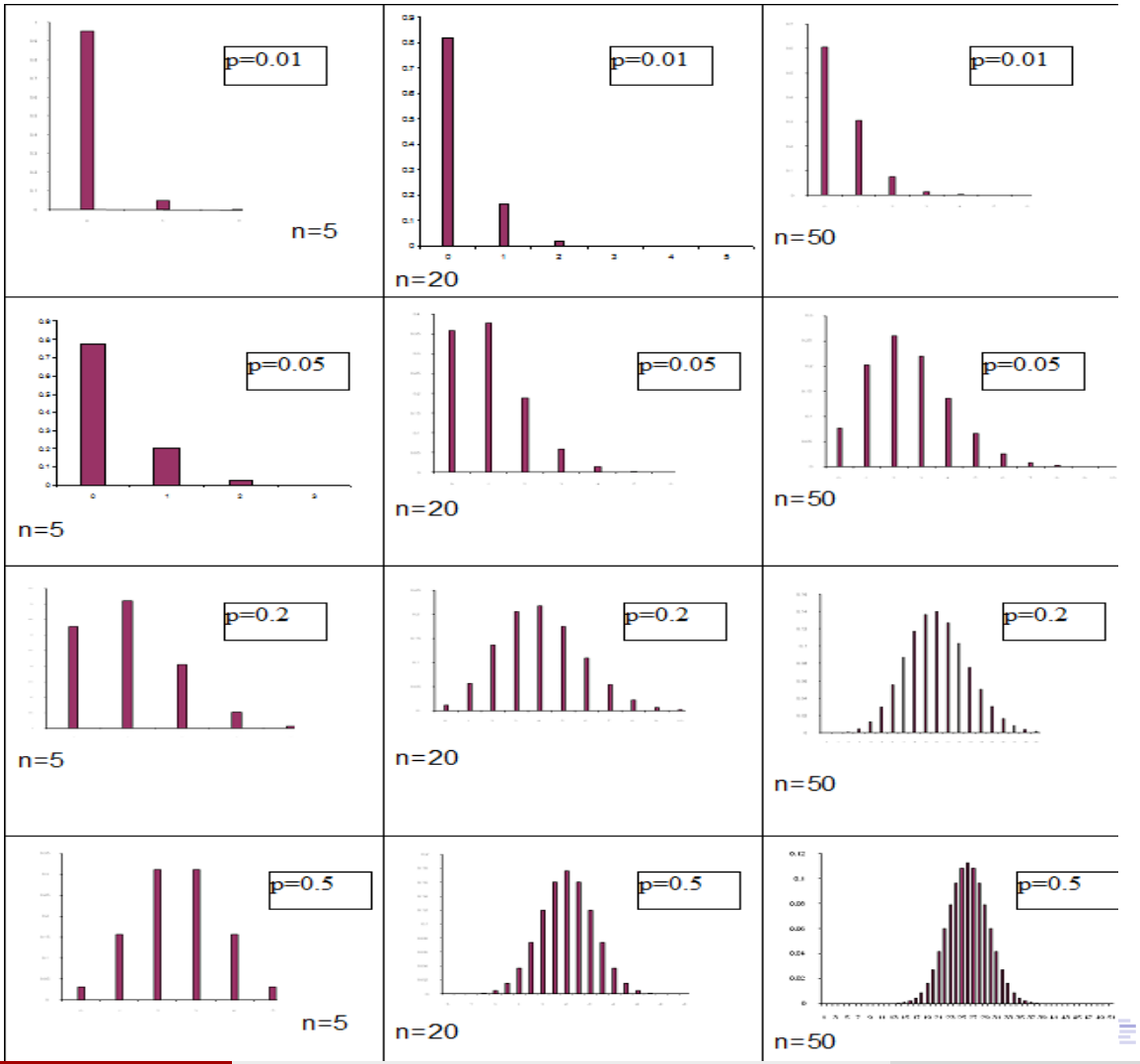
Livello di confidenza	80%	90%	95%	99%
	$t_{0.10}$	$t_{0.05}$	$t_{0.025}$	$t_{0.005}$
1	3.08	6.31	12.71	63.66
2	1.89	2.92	4.30	9.93
3	1.64	2.35	3.18	5.84
4	1.53	2.13	2.78	4.60
5	1.48	2.02	2.57	4.03
10	1.37	1.81	2.23	3.17
15	1.34	1.75	2.13	2.95
20	1.32	1.73	2.09	2.85
25	1.32	1.71	2.06	2.79
30	1.31	1.70	2.04	2.75
inf	1.28	1.65	1.96	2.58

Queste permettono di ricavare alcuni quantili (quelli rilevanti).

Indice

Risultato fondamentale: approssimazione bin-norm

Ricordiamo che per un numero di prove (cioè dimensione del campione) sufficientemente alto ($n \geq 30$) e per probabilità p dell'evento successo ($X = 1$) non molto piccole nè molto grandi la binomiale è ben approssimata dalla distribuzione normale



Elezioni: stima degli elettori del M5S I

- Vogliamo determinare la percentuale di elettori del M5S nella popolazione (parametro, π);
- a tal fine, selezioniamo 1000 persone a caso (c.c.s.) e gli chiediamo se voteranno M5S alle prossime elezioni, 266 risultano elettori M5S;
- la stima del numero di elettori del M5S è quindi $\hat{\pi} = 266/1000 = 0.266$.
- Mentre prima dovevamo stimare una media, ora dobbiamo stimare una proporzione.
- Vale però un discorso analogo, ricordiamo che (approssimazione normale alla binomiale), approssimativamente,

$$\hat{\pi} \sim N\left(\pi, \frac{\pi(1-\pi)}{n}\right)$$

Elezioni: stima degli elettori del M5S II

- Vale però un discorso analogo, ricordiamo che (approssimazione normale alla binomiale), approssimativamente,

$$\hat{\pi} \sim N\left(\pi, \frac{\pi(1-\pi)}{n}\right)$$

- si ha quindi (similmente a quanto visto nel caso della media)

$$P\left(\hat{\pi} - 1.96\text{se}(\hat{\pi}) \leq \pi \leq \hat{\pi} + 1.96\text{se}(\hat{\pi})\right) = 0.95$$

- dove questa volta

$$\text{se}(\hat{\pi}) = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} = \sqrt{\frac{0.266(1-0.266)}{1000}} = 0.014$$

Elezioni: stima degli elettori del M5S III

- si ha quindi, per la proporzione π nella popolazione, l'intervallo di estremi

$$\hat{\pi} \pm c_\alpha \text{se}(\hat{\pi}) = 0.266 \pm 1.96 \times 0.014$$

vale a dire

$$[0.23856, 0.29344]$$

Intervallo di confidenza per la proporzione I

Intervallo di confidenza per la proporzione

L'intervallo di confidenza di livello $1-\alpha$ per la proporzione π di una popolazione si ottiene come

$$\hat{\pi} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

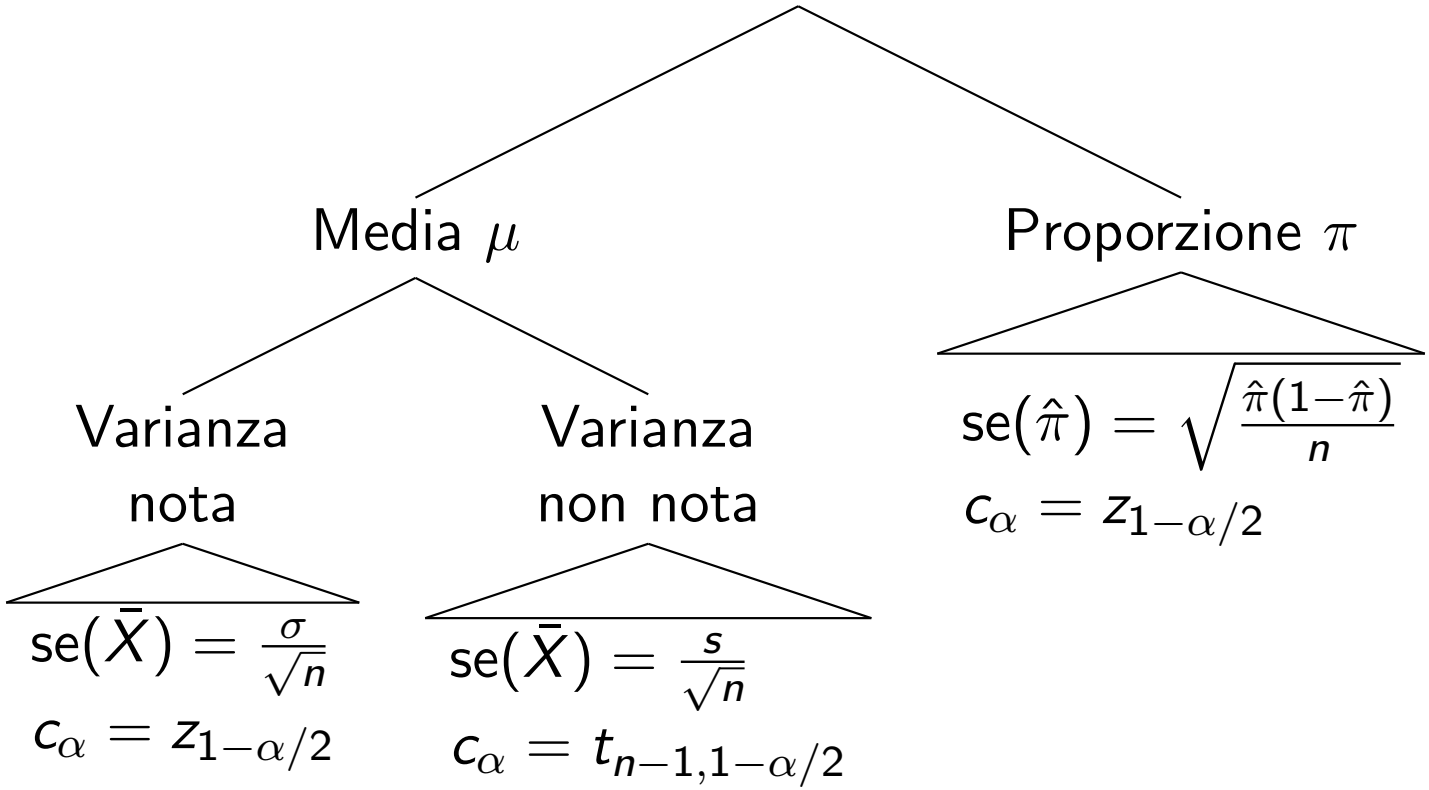
dove

- $\hat{\pi}$ è la proporzione campionaria;
- $z_{1-\alpha/2}$ è il quantile $1-\alpha/2$ della normale standard.

Indice

Casistica

Abbiamo visto tre diverse situazioni



Cosa succede se il campione è più grande?

- Supponiamo di selezionare 5000 persone (invece che 1000);
- ovviamente cambia la stima $\hat{\pi}_2 = 0.2466$;
- soprattutto, però, cambia la deviazione standard, che diventa più piccola

$$\text{se}(\hat{\pi}_2) = \sqrt{\frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n}} = \sqrt{\frac{0.2466(1-0.2466)}{5000}} = 0.0061$$

- l'intervallo diventa più stretto

$$\hat{\pi}_2 \pm c_\alpha \text{se}(\hat{\pi}_2) = 0.2466 \pm 1.96 \times 0.0061$$

vale a dire

$$[0.234644, 0.258556]$$

I.c. di prefissata lunghezza I

- Supponiamo di volere un i.c. di lunghezza non superiore a 2%,
- quanto dovrebbe essere grande il campione?
- la lunghezza dell'intervallo è

$$2 \times 1.96 \sqrt{\frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n}}$$

- vogliamo che

$$2 \times 1.96 \sqrt{\frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n}} < 0.02$$

- cioè

$$1.96 \sqrt{\frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n}} < 0.01$$

I.c. di prefissata lunghezza II

- non conosciamo $\hat{\pi}$, ovviamente, si può però mostrare che

$$\sqrt{\frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n}} < \sqrt{\frac{0.25}{n}}$$

- pertanto, se n è tale che

$$1.96\sqrt{\frac{0.25}{n}} < 0.01$$

  allora n soddisfa il requisito sopra;
- allora dev'essere

$$n > 1.96^2 0.25/0.01^2$$

$$n > 9604$$

Lunghezza i.c. formula generale

Formula per la dimensione campionaria

Si ottiene un intervallo di lunghezza al più $2a$ se

$$n > \frac{1}{4}\left(\frac{c_\alpha}{a}\right)^2$$

L'ampiezza è infatti

$$2c_\alpha\sqrt{\hat{\pi}(1-\hat{\pi})/n} < 2c_\alpha\sqrt{0.25/n}$$

se n soddisfa alla condizione sopra si ha

$$2c_\alpha\sqrt{0.25/n} < 2c_\alpha\sqrt{0.25} \times \sqrt{4}\frac{a}{c_\alpha} = 2a$$

Più o meno confidenza

Abbiamo ottenuto l'intevallo al livello del 95% per i votanti del M5S

$$0.251 \pm z_{0.975} 0.0137 \rightarrow [0.224, 0.278]$$

Se cambiamo il livelo di confidenza otteniamo intervalli più o meno ampi

- al 90% $\rightarrow z_{0.95} = 1.64 \rightarrow [0.228, 0.273]$
  abbiamo meno confidenza su un intervallo meno ampio.
- al 98% $\rightarrow z_{0.99} = 2.33 \rightarrow [0.219, 0.283]$
  abbiamo più confidenza su un intervallo più ampio.