Inferenza Statistica

Domenico De Stefano

A.A. 2020/2021

Indice

Popolazione e parametro

- Consideriamo una popolazione,
  - ad es.: corpo elettorale della CdD italiana;

Popolazione e parametro

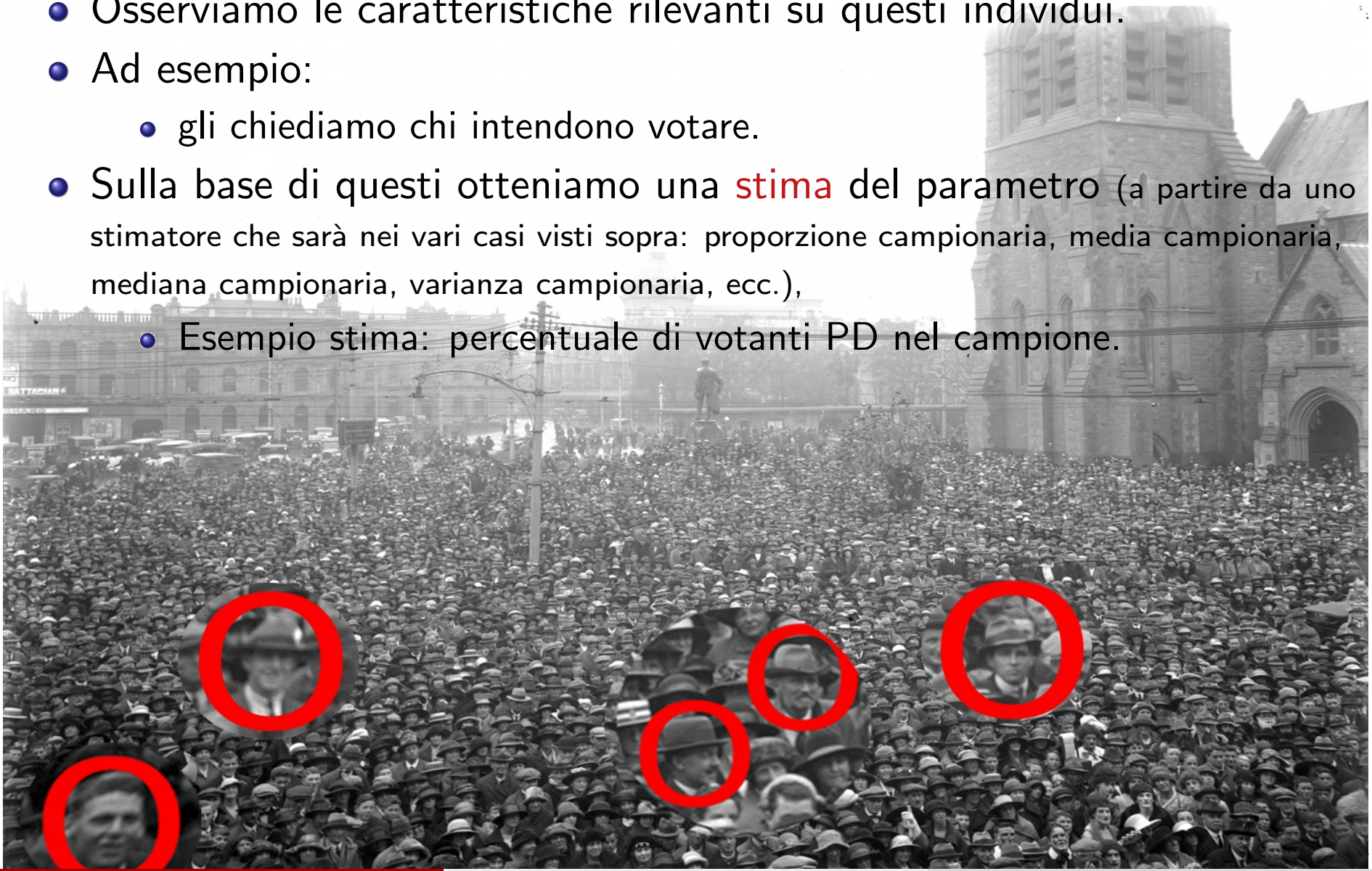
- Consideriamo una popolazione,
  - ad es.: corpo elettorale della CdD italiana;
- Di questa popolazione interessa una certa caratteristica, questa è detta parametro, che per noi sarà un certo indice: ad esempio una percentuale (o proporzione o frequenza relativa), una media, una mediana o una varianza di una data variabile),
  - Esempio parametro: percentuale di elettori del PD.

Campione

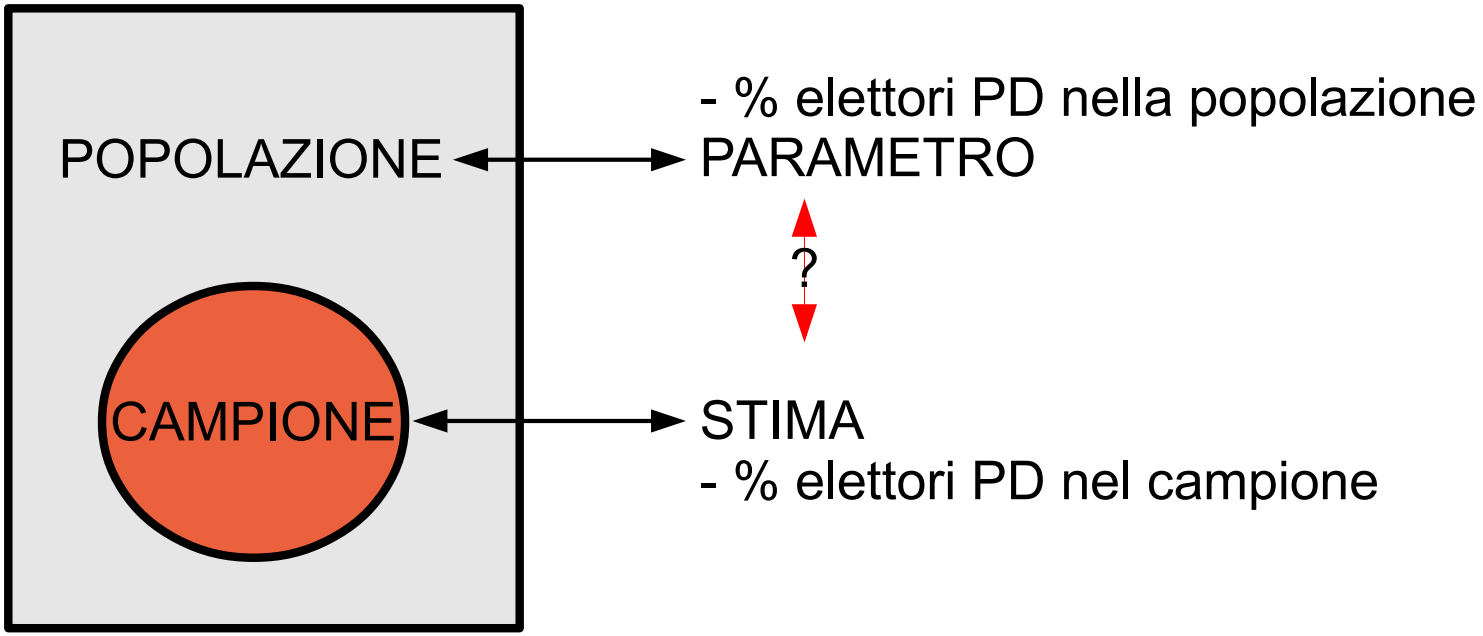
- Non possiamo/vogliamo osservare tutta la popolazione (troppo grande, non c'è tempo),
- Scegliamo **a caso** alcuni individui: il campione.

Campione e stima

- Osserviamo le caratteristiche rilevanti su questi individui.
- Ad esempio:
  - gli chiediamo chi intendono votare.
- Sulla base di questi otteniamo una stima del parametro (a partire da uno stimatore che sarà nei vari casi visti sopra: proporzione campionaria, media campionaria, mediana campionaria, varianza campionaria, ecc.),
  - Esempio stima: percentuale di votanti PD nel campione.

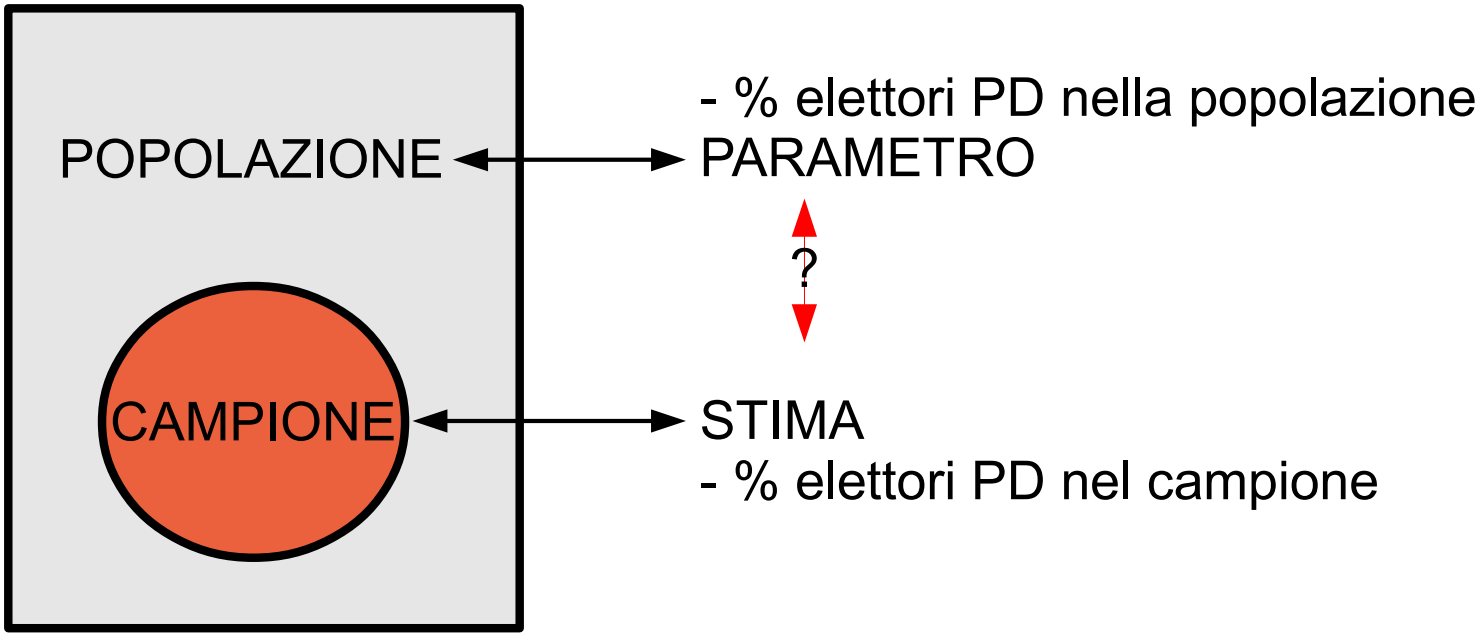


Schematicamente: popolazione-parametro, campione-stima



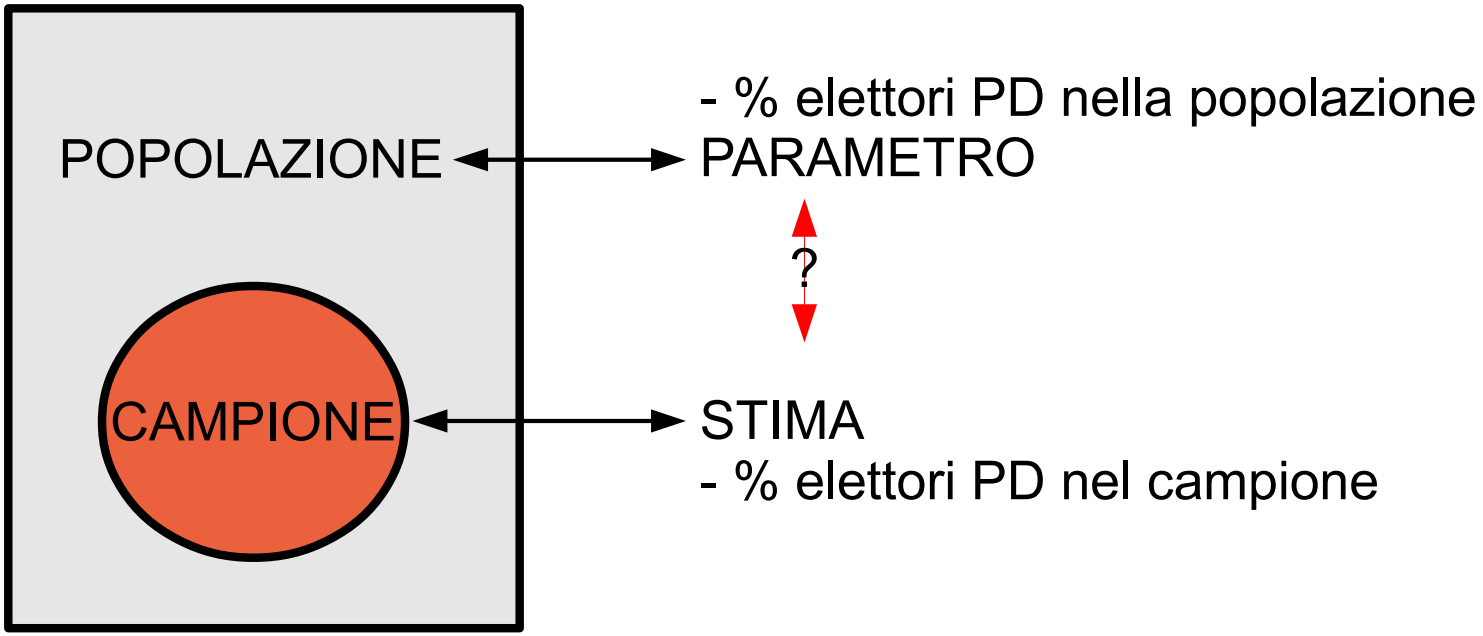
- Non è ovvio, ancora, cosa sia una stima e come sia collegata al parametro.

Schematicamente: popolazione-parametro, campione-stima



- Non è ovvio, ancora, cosa sia una stima e come sia collegata al parametro.
- Per intanto una stima è una caratteristica quantitativa del campione che si ritiene 'simile' al parametro.

Schematicamente: popolazione-parametro, campione-stima



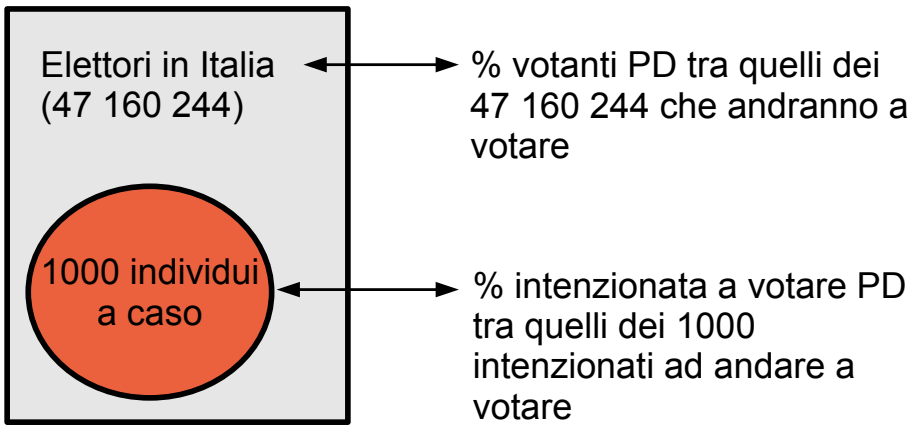
- Vediamo intanto alcuni esempi, poi discuteremo perché questa dovrebbe essere 'simile' al parametro.

Esempio: sondaggio elettorale

Quanti voteranno PD alla Camera dei Deputati nelle prossime elezioni?

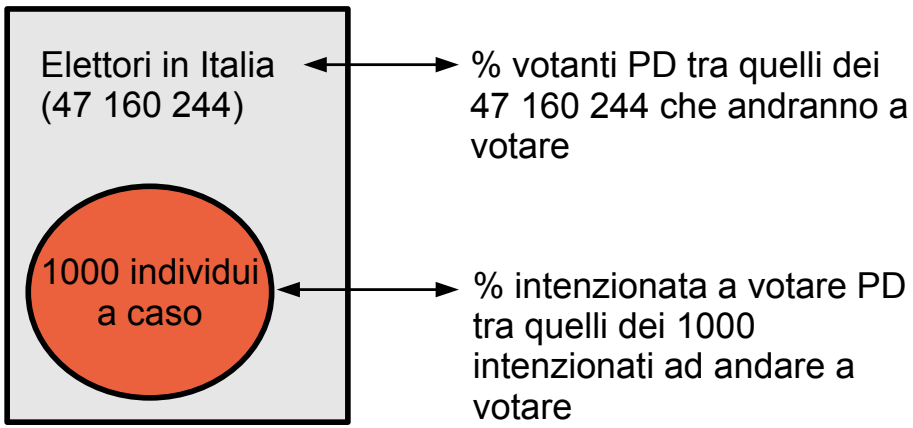
Esempio: sondaggio elettorale

Quanti voteranno PD alla Camera dei Deputati nelle prossime elezioni?



Esempio: sondaggio elettorale

Quanti voteranno PD alla Camera dei Deputati nelle prossime elezioni?



In un sondaggio del 28-29 gennaio 2013 la percentuale è risultata pari a 31,5% .

Esempio: stranieri e lavoro

Quanti studenti di seconda generazione hanno subito almeno un episodio offensivo, non rispettoso e/o violento da parte di altri ragazzi nell'ultimo mese?

Esempio: stranieri e lavoro

Quanti studenti di seconda generazione hanno subito almeno un episodio offensivo, non rispettoso e/o violento da parte di altri ragazzi nell'ultimo mese?

Esempio: stranieri e lavoro

Quanti studenti di seconda generazione hanno subito almeno un episodio offensivo, non rispettoso e/o violento da parte di altri ragazzi nell'ultimo mese?

Secondo l'ISTAT la percentuale è intorno al 49.5% (Indagine ISTAT sulle Seconde generazioni: identità e percorsi di integrazione, `https://www4.istat.it/it/archivio/210595`).

Esempio: consumi

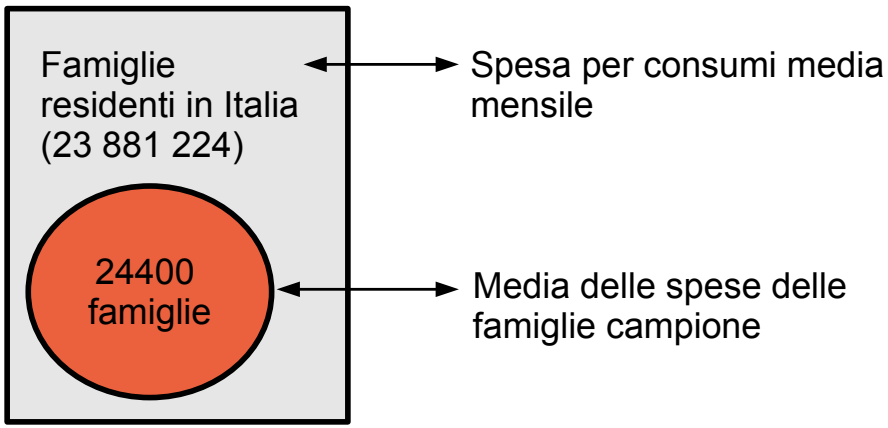
Non sempre la popolazione è composta da individui.

Quanto spende in media una famiglia italiana ogni mese?

Esempio: consumi

Non sempre la popolazione è composta da individui.

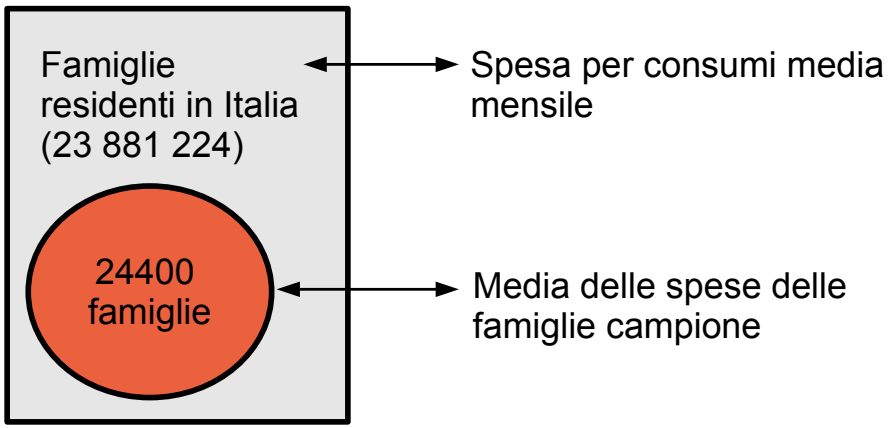
Quanto spende in media una famiglia italiana ogni mese?



Esempio: consumi

Non sempre la popolazione è composta da individui.

Quanto spende in media una famiglia italiana ogni mese?



Secondo l'ISTAT la spesa media è 2480 € e l'intervallo di confidenza è 2450-2510 €.

Come dev'essere il campione

Quando diciamo che il campione sono n individui selezionati nella popolazione, questo non vuol dire che qualunque gruppo di n individui vada bene.

Campione "rappresentativo"

Un campione "rappresentativo" è un sottoinsieme della popolazione che ne riflette le caratteristiche. (Una versione in miniatura della popolazione.)

È il fatto che il campione è rappresentativo che consente di generalizzare i risultati che si ottengono sulla base di calcoli fatti sul campione, alla popolazione.

Come NON dev'essere il campione

NON si ottiene un campione rappresentativo

- prendendo le persone presenti in quest'aula,
- prendendo gli amici/parenti/conoscenti,
- ponendo una domanda in una trasmissione televisiva e invitando il pubblico a rispondere via telefono o sms o internet.

questi gruppi di persone hanno caratteristiche peculiari, non possiamo escludere che queste siano legate alle caratteristiche che stiamo indagando, quindi introdurremmo delle distorsioni.
Per grande che sia, un campione non rappresentativo non consente generalizzazioni.

Come ottengo un campione rappresentativo?

L'idea è di selezionare le unità da includere nella popolazione in modo casuale, poi ci sono diversi metodi

- Campione casuale semplice (ccs): Il modo più semplice è scegliere n individui in modo che ciascun individuo della popolazione abbia la stessa probabilità di essere estratto.
- Altre opzioni sono spesso usate allo scopo di
  - migliorare la rappresentatività;
  - semplificare la procedura (risparmio).

  tra queste
  - campione stratificato;
  - campione a grappoli;
  - campione a più stadi.

  come nel ccs tutti possono essere estratti, però le probabilità possono variare.

Campione casuale $\neq$ stime non distorte

Anche se il campione è scelto in modo casuale, non è detto che si siano eliminati tutti i problemi.

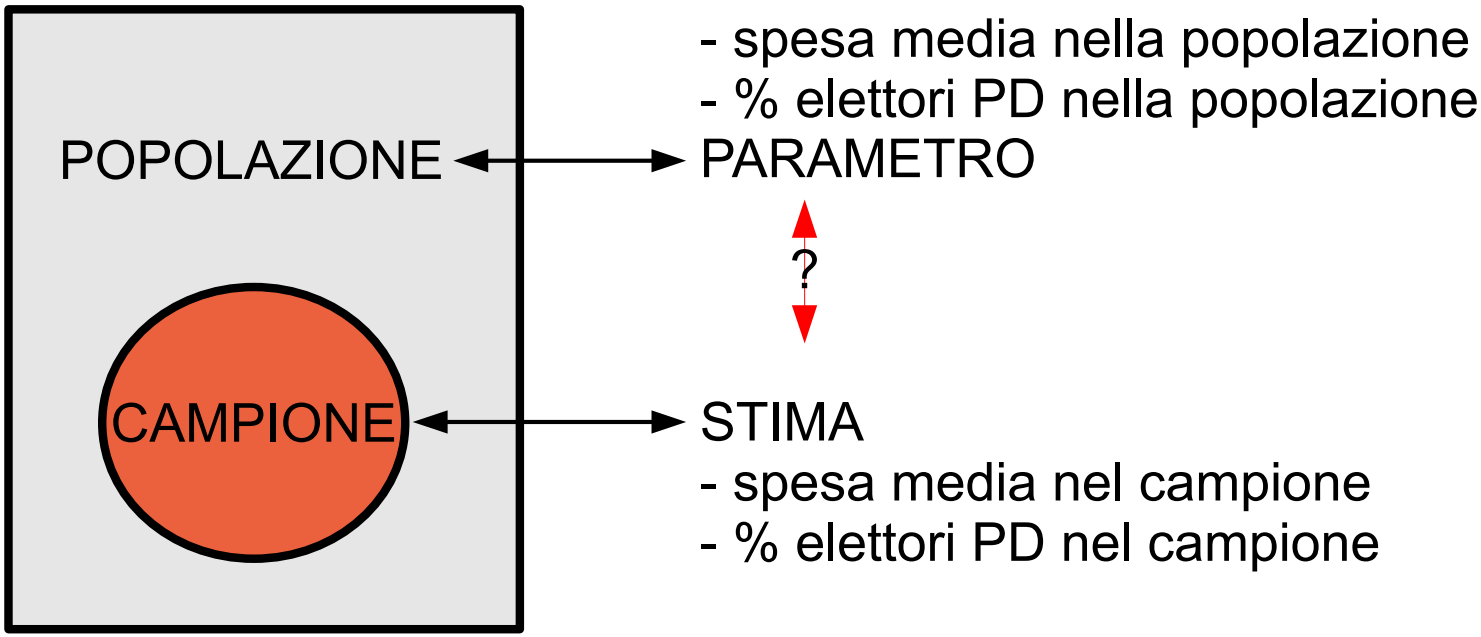
- Sottocopertura: può essere che il campione venga dalla popolazione sbagliata, un indagine telefonica 'manca' tutti quelli che non hanno telefono fisso.

Anche se il campione è costruito perfettamente, le stime possono essere distorte da altri fenomeni

- Non risposte: i democratici tendono a rispondere più frequentemente dei conservatori.
- False risposte: si pensi a un'indagine sul consumo di stupefacenti.

Stima e parametro

Abbiamo individuato i personaggi e la relazione tra alcuni di essi.



- Il punto è: perché sono legati stima e parametro?
- Ci aspettiamo che la stima sia vicina al valore del parametro, ma perché?
- Cerchiamo di capire meglio cos'è la stima.

Campioni e stime

- Consideriamo di nuovo l'esempio relativo alle elezioni, alla ricerca della percentuale di elettori del PD.
- Il problema se lo pongono in molti, e molti fanno dei sondaggi, ad esempio il 14 gennaio 2013 ne sono stati fatti 3.
- tutti dati disponibili al sito: `http://www.sondaggipoliticoelettorali.it`, che siete invitati a visitare!

committente	sondaggista	num.camp	PD
RAI BallarÃ	Ipsos	800	33.40
Sky	TecnÃ©	600	31.80
Reuters AffariItaliani.it	Piepoli	1003	32.50

Campioni e stime

- Consideriamo di nuovo l'esempio relativo alle elezioni, alla ricerca della percentuale di elettori del PD.
- Il problema se lo pongono in molti, e molti fanno dei sondaggi, ad esempio il 14 gennaio 2013 ne sono stati fatti 3.
- tutti dati disponibili al sito: `http://www.sondaggipoliticoelettorali.it`, che siete invitati a visitare!

committente	sondaggista	num.camp	PD
RAI BallarÃ	Ipsos	800	33.40
Sky	TecnÃ©	600	31.80
Reuters AffariItaliani.it	Piepoli	1003	32.50

- Nei 3 sondaggi si ottengono stime diverse, eppure sono contemporanei e popolazione e parametro sono gli stessi.

Campioni e stime

- Consideriamo di nuovo l'esempio relativo alle elezioni, alla ricerca della percentuale di elettori del PD.
- Il problema se lo pongono in molti, e molti fanno dei sondaggi, ad esempio il 14 gennaio 2013 ne sono stati fatti 3.
- tutti dati disponibili al sito: `http://www.sondaggipoliticoelettorali.it`, che siete invitati a visitare!

committente	sondaggista	num.camp	PD
RAI BallarÃ	Ipsos	800	33.40
Sky	TecnÃ©	600	31.80
Reuters AffariItaliani.it	Piepoli	1003	32.50

- Nei 3 sondaggi si ottengono stime diverse, eppure sono contemporanei e popolazione e parametro sono gli stessi.
- Qualcuno si sbaglia?

Campioni e stime

- Consideriamo di nuovo l'esempio relativo alle elezioni, alla ricerca della percentuale di elettori del PD.
- Il problema se lo pongono in molti, e molti fanno dei sondaggi, ad esempio il 14 gennaio 2013 ne sono stati fatti 3.
- tutti dati disponibili al sito: `http://www.sondaggipoliticoelettorali.it`, che siete invitati a visitare!

committente	sondaggista	num.camp	PD
RAI BallarÃ	Ipsos	800	33.40
Sky	TecnÃ©	600	31.80
Reuters AffariItaliani.it	Piepoli	1003	32.50

- Nei 3 sondaggi si ottengono stime diverse, eppure sono contemporanei e popolazione e parametro sono gli stessi.
- Qualcuno si sbaglia?
- In un certo senso tutti sbagliano.

Campioni e stime

- Consideriamo di nuovo l'esempio relativo alle elezioni, alla ricerca della percentuale di elettori del PD.
- Il problema se lo pongono in molti, e molti fanno dei sondaggi, ad esempio il 14 gennaio 2013 ne sono stati fatti 3.
- tutti dati disponibili al sito: `http://www.sondaggipoliticoelettorali.it`, che siete invitati a visitare!

committente	sondaggista	num.camp	PD
RAI BallarÃ	Ipsos	800	33.40
Sky	TecnÃ©	600	31.80
Reuters AffariItaliani.it	Piepoli	1003	32.50

- Nei 3 sondaggi si ottengono stime diverse, eppure sono contemporanei e popolazione e parametro sono gli stessi.
- Qualcuno si sbaglia?
- In un certo senso tutti sbagliano.
- Il fatto è che sono intervistate persone diverse e quindi il risultato è

Stimatore, stima, distribuzione campionaria

- Una stima è il risultato di un esperimento casuale:
  - estraggo a caso il campione,
  - applico lo stimatore (ad es. la proporzione campionaria),
  - ottengo la stima (valore assunto dallo stimatore),
- se estraessi un altro campione otterrei una diversa stima!

- Quindi otteniamo una risposta a caso?

Stimatore, stima, distribuzione campionaria

- Quindi otteniamo una risposta a caso?
- La risposta è sì, la risposta che ottengo è casuale.
- Questo però non significa che non sia informativa.
- La stima è sì casuale, ma è probabile che sia vicina al parametro. (vero valore della proporzione di votanti il PD, nel nostro esempio)
- (Sempre se sto facendo tutto bene.)
- Per capire cosa significa questo, immagineremo di estrarre tanti campioni, cioè di ripetere più volte la procedura.

Indice

Dati: tempi alla Bavisela

- Disponiamo dei tempi registrati, in ore, da 20 dei 695 concorrenti maschi alla Bavisela 2012,
- il parametro d'interesse è il tempo medio registrato tra i 695 concorrenti (che, nota la popolazione, potremmo facilmente calcolare).
- Non conosciamo tutti i tempi, usiamo un campione di 20 unità:

$$2.933, 3.081, 3.32, 3.519, 3.559, 3.631, 3.676, 3.836, 3.859, 3.887$$

$$3.941, 3.951, 4.091, 4.105, 4.241, 4.274, 4.289, 4.447, 4.999, 5.203$$

- La media del campione è 3.9421, è una buona stima? Perché?

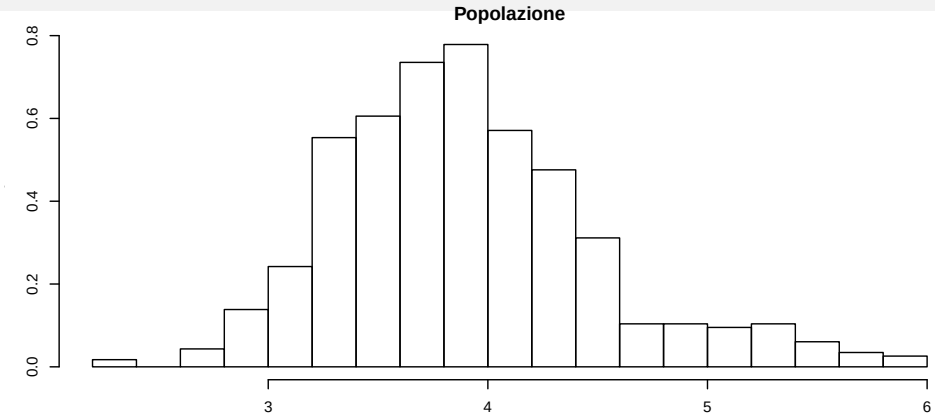
Principio del campionamento ripetuto

Per valutare la bontà di uno stimatore (qui la media campionaria) per un parametro (qui la media della popolazione) si ragiona così

- immagino di ripetere il campionamento (ottenere cioè altri campioni);
- a ogni campione corrisponde una stima (media campionaria) diversa;
- come sono messe queste stime rispetto al parametro?

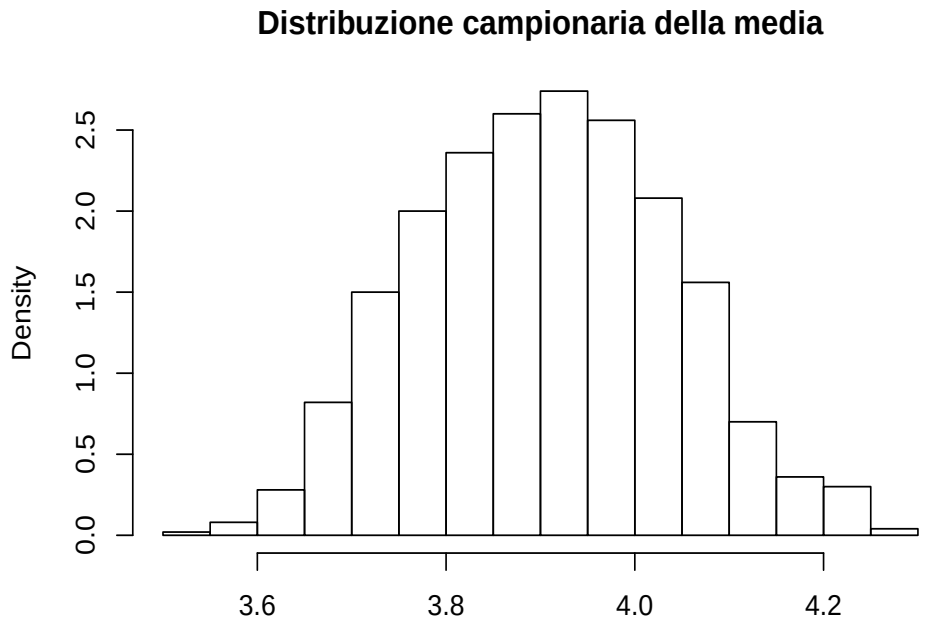
Per questo esempio possiamo davvero farlo, vediamo che succede.

Distribuzione campionaria



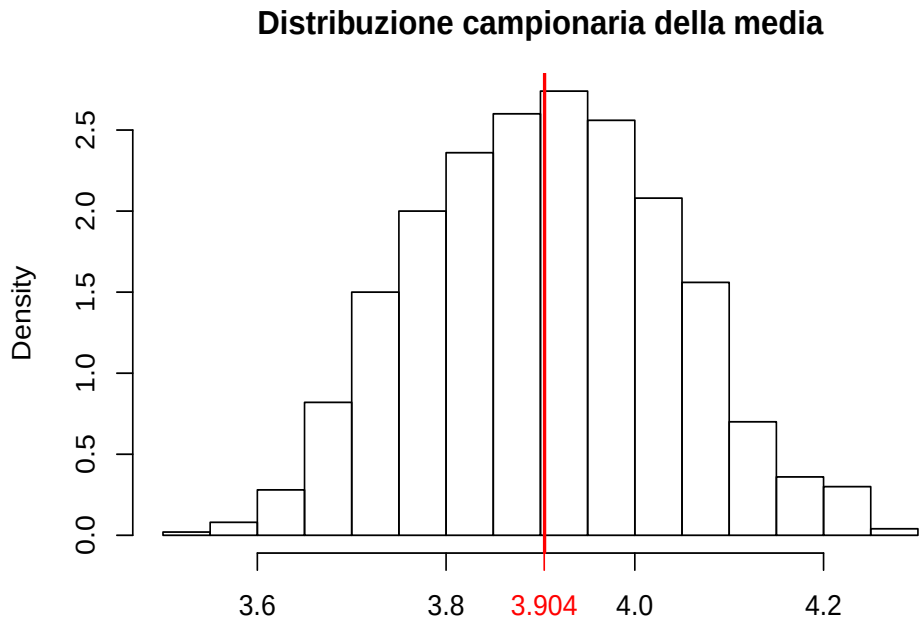
Distribuzione campionaria

Le 1000 medie derivate dai nostri 1000 campioni si distribuiscono così



Distribuzione campionaria

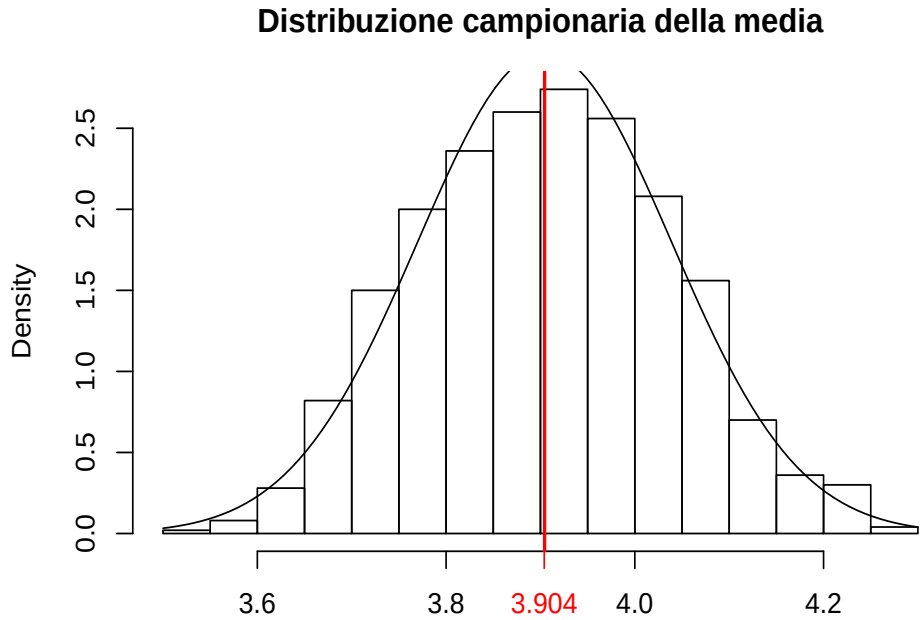
Le 1000 medie derivate dai nostri 1000 campioni si distribuiscono così



- Intorno al vero valore del parametro, nel senso che sono più probabili valori vicini alla media della popolazione, 3.9, che valori lontani.

Distribuzione campionaria

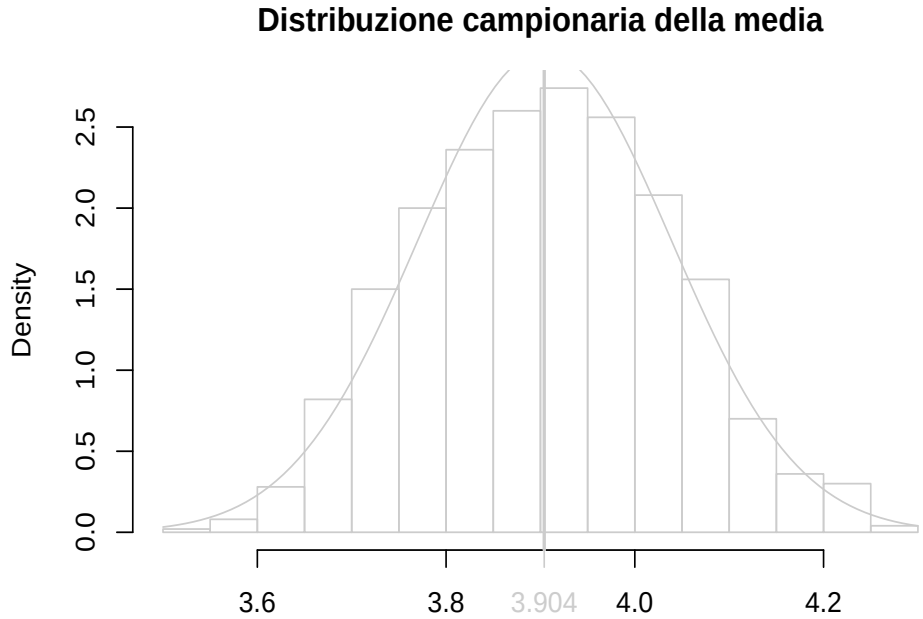
Le 1000 medie derivate dai nostri 1000 campioni si distribuiscono così



- Intorno al vero valore del parametro, nel senso che sono più probabili valori vicini alla media della popolazione, 3.9, che valori lontani.
- Semplifichiamo la rappresentazione, immaginiamo di avere un'infinità di campioni e passiamo alla funzione di densità.

Distribuzione campionaria

Le 1000 medie derivate dai nostri 1000 campioni si distribuiscono così



Nella realtà però io

- non conosco la media della popolazione,
- non osservo tanti campioni ma uno.

Senza simulazioni

- In questo esempio abbiamo potuto simulare tanti campioni perchè avvamo la popolazione.
- In generale non è così, useremo risultati teorici.

Distribuzione approssimata della media

Un risultato fondamentale è il seguente teorema che ci dice che, approssimativamente, la media campionaria è sempre normale, con media pari alla media della popolazione e varianza che dipende dalla varianza della popolazione e dalla dimensione del campione.

Teorema del limite centrale

Siano $X_1, \ldots, X_n$ con media μ e varianza σ^2 indipendenti, allora, approssimativamente

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Numerosità campionaria e precisione dello stimatore

- Nella cittadina di Dunwich, che conta 100000 abitanti, si terranno a breve le elezioni comunali.
- Ci sono due soli candidati, Arthur Dupin (attuale sindaco) e Auguste Pym.
- L'attuale sindaco Arthur Dupin commissiona un sondaggio: 500 persone vengono intervistate (e tutti rispondono, è una storia di fantasia), il 50% dichiara che voterà per lui.
- Lo sfidante d'altro canto commissiona un suo sondaggio, sulla base di 1000 rispondenti (anche qui rispondono tutti), si stima al 45% la proporzione di votanti per l'attuale sindaco.

che due campioni portino a stime diverse ormai non ci stupisce, ma la domanda è

> Quale delle due stime è da ritenersi maggiormente affidabile?

Numerosità campionaria e precisione dello stimatore

- Auguste Pym che, ricordiamolo, è dato favorito come candidato sindaco di Dunwich (100000 abitanti) secondo un sondaggio che ha coinvolto 1000 persone, si vanta con un amico, William Wilson, anche lui politico, del fatto appunto si essere il favorito nella sua competizione elettorale.
- L'amico William è anch'egli candidato sindaco nella città di Innsmouth (1000000 abitanti) e anche lui è dato al 55% in un sondaggio effettuato intervistando 1000 persone (tutte rispondenti) e dice all'amico 'Ehi, anch'io sono favorito come te a Innsmouth.'
- Auguste replica però: 'Sì, però il mio sondaggio ha intervistato 1000 persone su 100000 (l'1 per cento), il tuo 1000 su 1000000 (l'1 per mille), io sono molto più sicuro di te.

Auguste ha ragione?

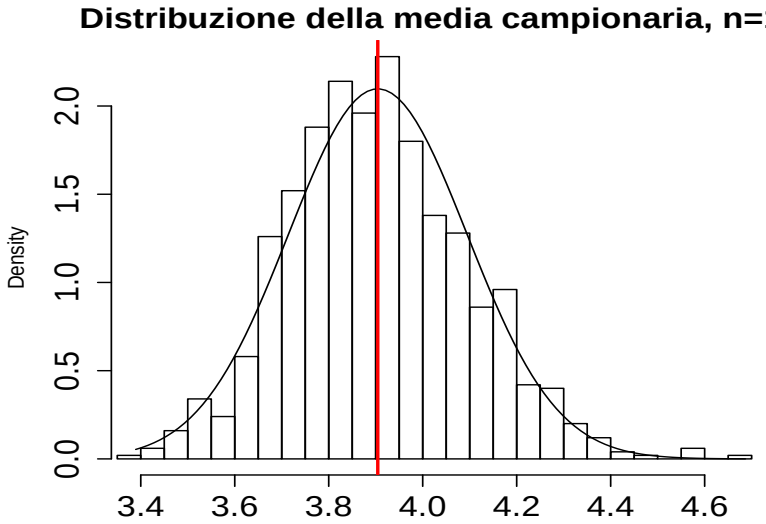
Campionare è come assaggiare una zuppa durante la cottura

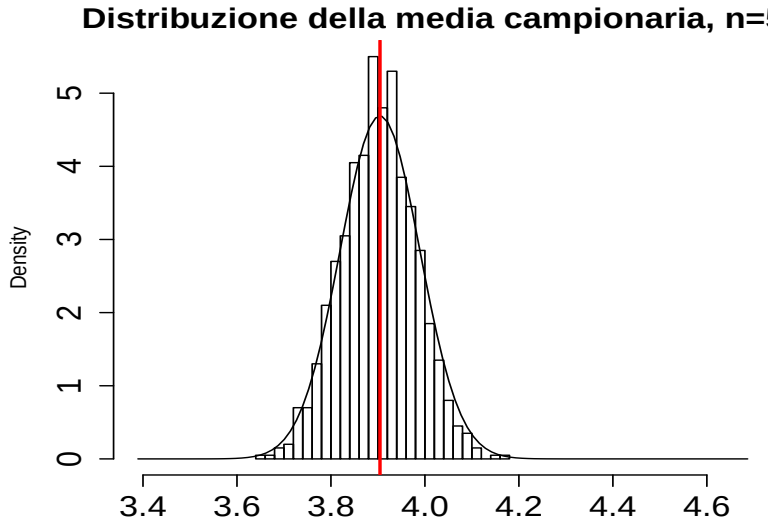
- Quando osserviamo un campione osserviamo una parte della popolazione e ne inferiamo le caratteristiche dell'intera popolazione.
- Quando assaggiamo la zuppa per vedere se è giusta di sale ne mangiamo un cucchiaio, verifichiamo che quel cucchiaio sia correttamente salato.
- Se la zuppa è stata adeguatamente mescolata, riteniamo che l'intera pentola abbia lo stesso gusto del (piccolo) cucchiaio assaggiato.
- Cambia qualcosa che la pentola sia piccola o grande?

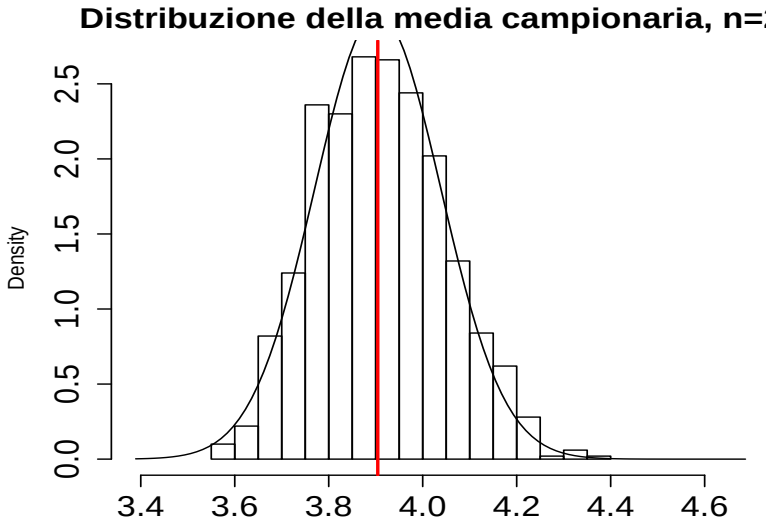
Cosa significa

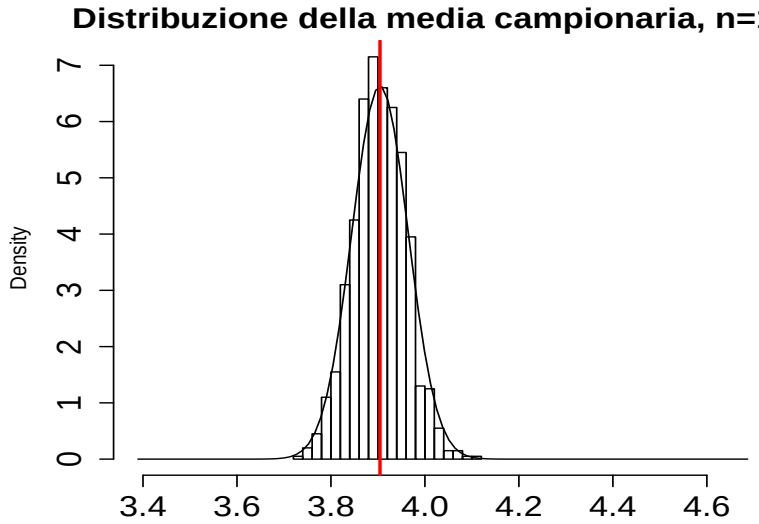
- Che se anche ho una popolazione molto grande, non mi occorre un campione molto grande (un campione di 1000 va egualmente bene per la popolazione della Lombardia, dell'Italia o dell'UE)
- D'altra parte, anche se la popolazione è piccola non è che mi basta un campione più piccolo (un campione di 1000 ha la stessa affidabilità che la popolazione di riferimento sia quella UE, quella italiana o quella di Trieste.)
- Attenzione che tutto questo è vero se comunque la popolazione è grande rispetto al campione (un campione di 1000 abitanti di Opicina ha un valore diverso, ovviamente).

Cosa succede se aumenta la numerosità del campione *n*

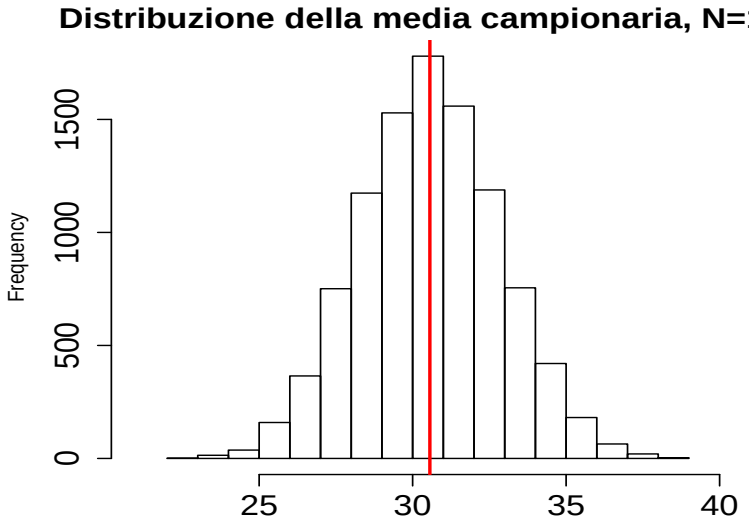


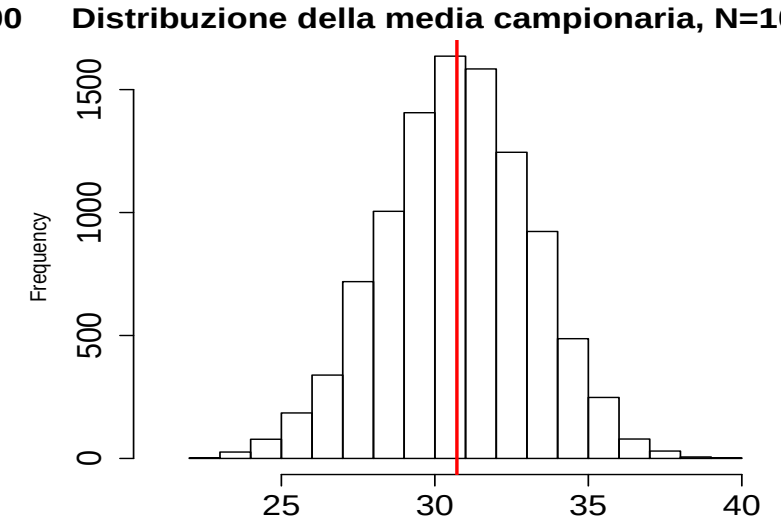


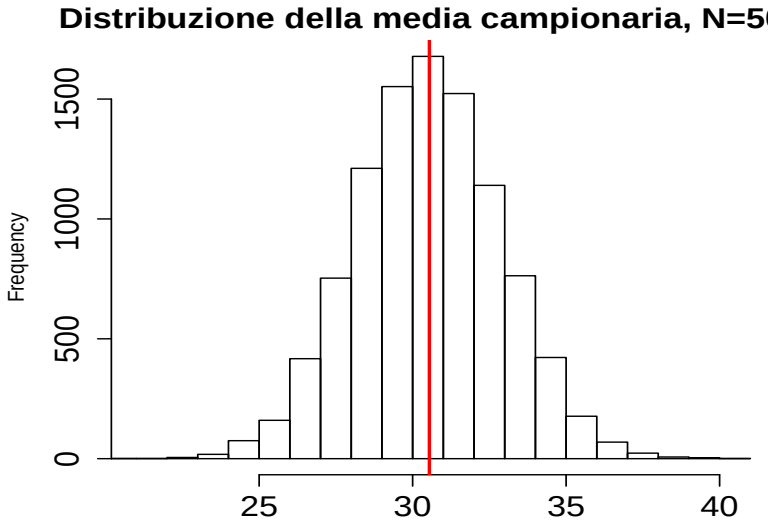


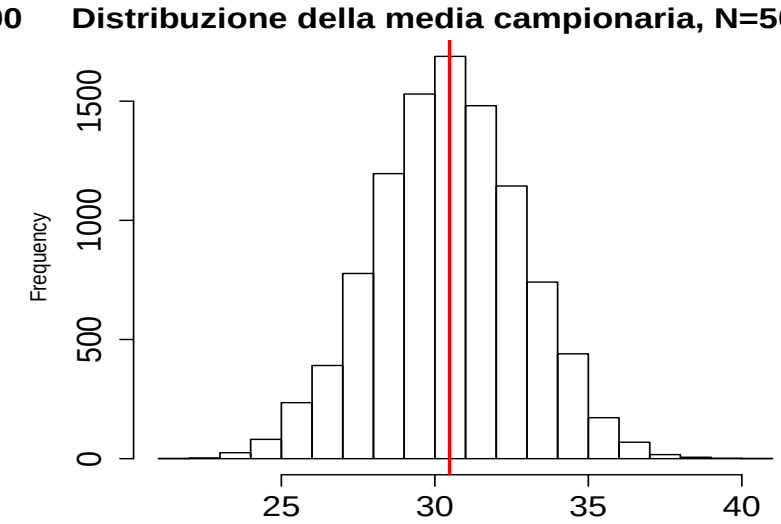


Cosa succede se aumenta la dimensione della popolazione N









Stimatore e distribuzione campionaria

- Lo stimatore è una quantità che calcolo a partire dal campione.
- Esso è buono se la sua distribuzione campionaria si concentra intorno al vero valore del parametro.
- Quanto più è concentrata tanto meglio.
- La distribuzione campionaria servirà a stabilire se non quanto vale il parametro quanto, probabilmente, siamo vicini.

Indice

Elezioni: elettori M5S e approssimazione binomiale I

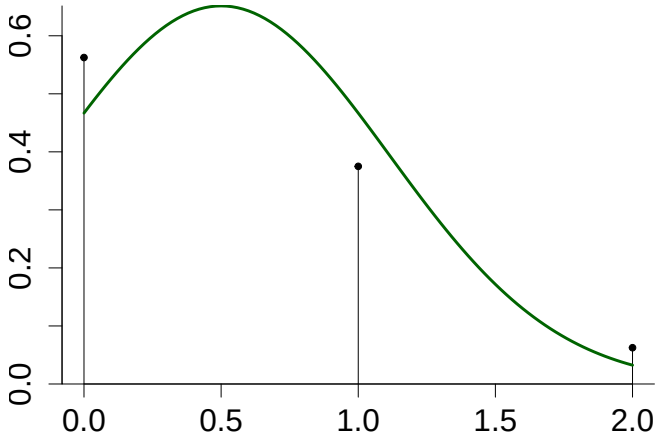
- Abbiamo visto che, preso un votante, la probabilità che abbia votato M5S il 24/25 febbraio è 0.2464.
- Supponiamo estrarre un campione casuale di 1000 elettori, il numero di persone che hanno votato M5S è una binomiale con media 246.4 e varianza 185.68704.
- (In realtà questa è un'approssimazione che presuppone di estrarre con rimpiazzo, cioè di poter estrarre due volte la stessa persona, l'effetto di questa approssimazione è però trascurabile.)
- L'esempio imita un sondaggio elettorale
- Si potrebbe lavorare con la binomiale, in realtà anche qui (dato che n è grande, ossia maggiore di 30 come regola, e la probabilità di successo, p, non è 0 o 1) possiamo usare l'**approssimazione normale**.

Normale come approssimazione della binomiale

- Consideriamo una binomiale di dimensione n e probabilità p

$$P(S = s) = \binom{n}{s} p^s (1-p)^{n-s}.$$

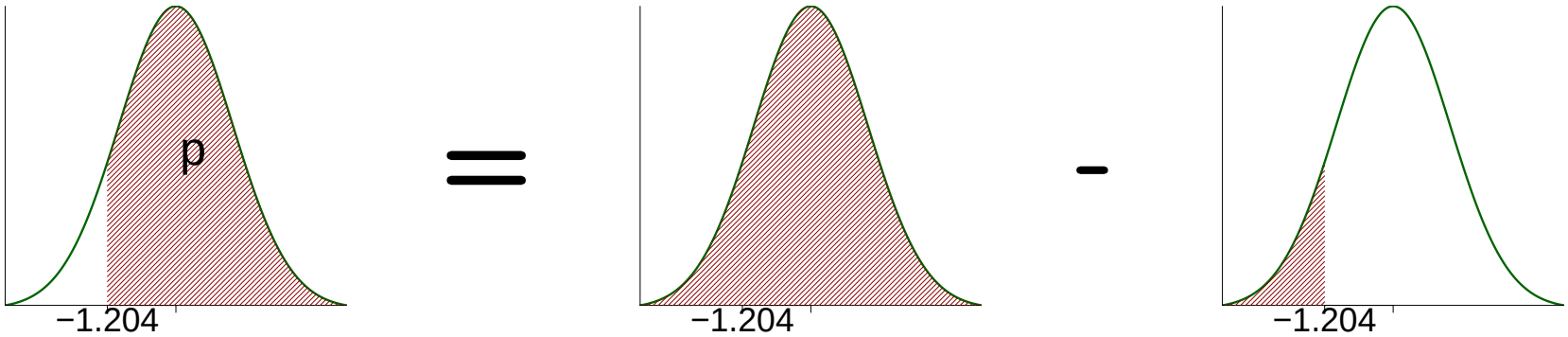
- Questa è approssimabile con una normale di media np e varianza $np(1-p)$.



Distribuzione dei votanti M5S I

- Qual è la probabilità che, tra i 1000 più di 230 abbiano votato M5S
- Usando l'approssimazione normale, è

$$p = 1 - \Phi\left(\frac{230 - 246.4}{\sqrt{185.68704}}\right) = 1 - \Phi\left(-1.2035193\right) = 0.8856123$$



$$\begin{aligned} p &= 1 - \Phi\left(\frac{230-246.4}{\sqrt{185.68704}}\right) \\ &= 1 - \Phi\left(\right) \\ &= 1 - 0.1143877 \end{aligned}$$

Trasformazioni e normalità

Trasformato di una normale

Se X è distribuito secondo una $N(\mu, \sigma^2)$, e a, b sono due numeri reali, allora

$$Y = aX + b$$

è distribuito secondo una $N(a\mu + b, a^2\sigma^2)$.

- Se il numero X di votanti il M5S su n individui è

$$N(np, np(1-p))$$

- La proporzione, $Y = X/n$ è

$$N(p, p(1-p)/n)$$

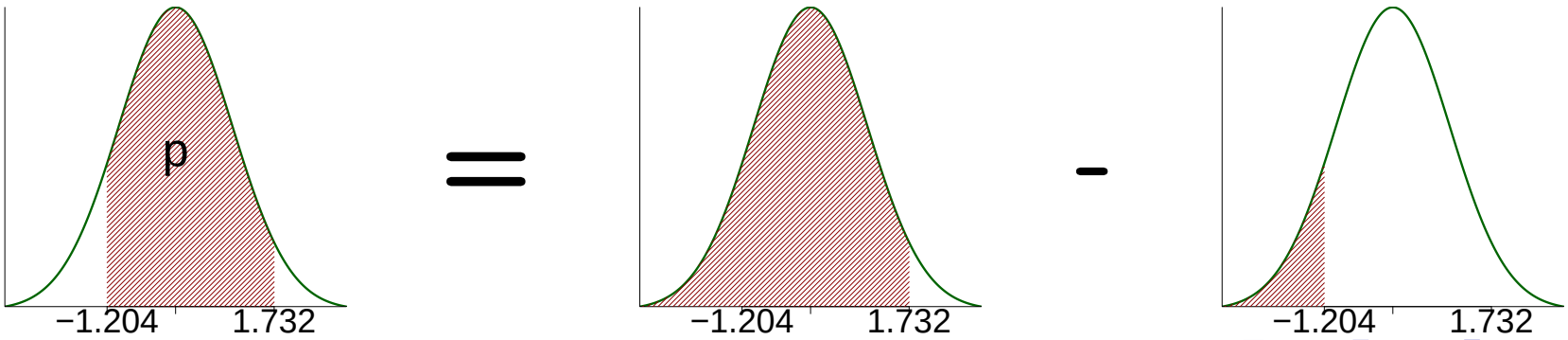
Proporzione di votanti

- Qual è la probabilità che la percentuale di votanti il M5S sia compresa tra 23 e 27 %?
- Sapendo che la proporzione di votanti il M5S su n individui è

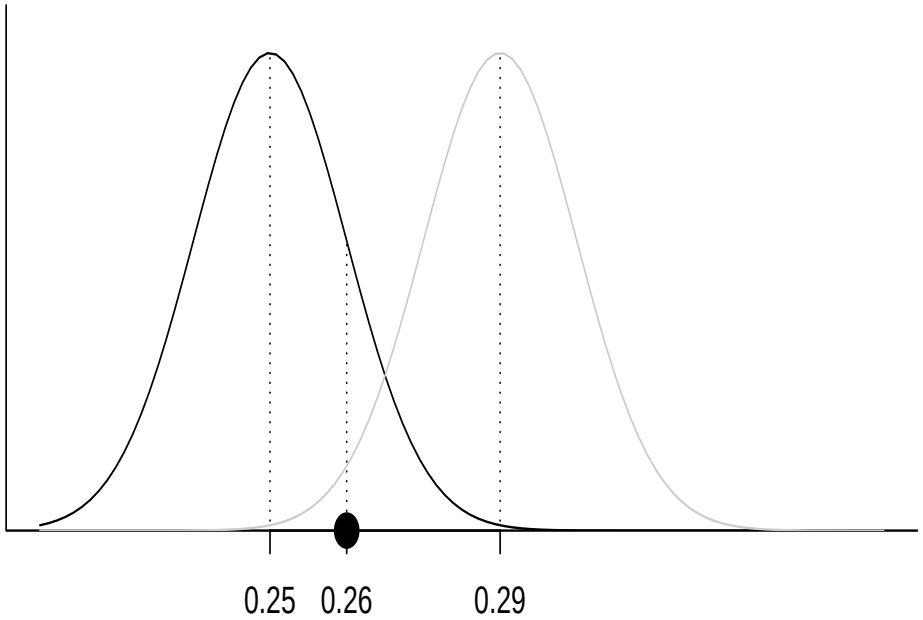
$$N(p, p(1-p)/n)$$

- la probabilità richiesta è allora

$$\begin{aligned} p &= \Phi\left(\frac{0.27 - 0.2464}{\sqrt{0.0001857}}\right) - \Phi\left(\frac{0.23 - 0.2464}{\sqrt{0.0001857}}\right) = \\ &= \Phi(1.7318936) - \Phi(-1.2035193) = \\ &= 0.9583537 - 0.1143877 = 0.843966 \end{aligned}$$



Inferenza per la proporzione



- La proporzione campionaria è 0.26.
- È più compatibile con una proporzione nella popolazione pari a 0.25 che una pari a 0.29.

Indice

Riepilogo

- Popolazione: collettivo su cui vogliamo ricavare informazioni;
- Parametro: l'informazione di interesse, è una caratteristica della popolazione;
- Campione: parte della popolazione che osservo
  - dev'essere rappresentativo;
  - selezione casuale;
- Stimatore: corrispondente del parametro
  - calcolato dal campione;
  - ha una distribuzione di probabilità (campionaria)
    - concentrata intorno al valore del parametro;
    - tanto più concentrata quanto più numeroso è il campione.