UNIVERSITÀ DEGLI STUDI DI TRIESTE

Dipartimento di scienze economiche, aziendali, matematiche e statistiche "Bruno de Finetti"

Statistica

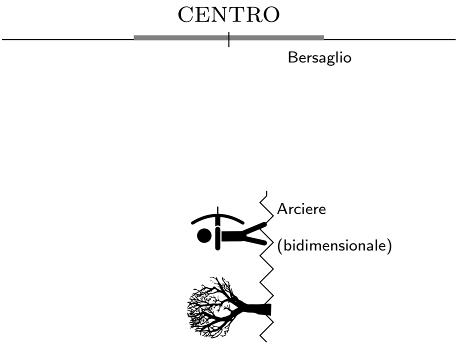
Inferenza: introduzione

Francesco Pauli

A.A. 2016/2017

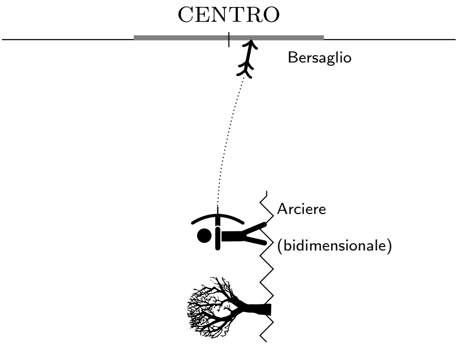
Indice

L'arciere e la freccia



Un arciere lancia una freccia contro il centro di un bersaglio.

L'arciere e la freccia

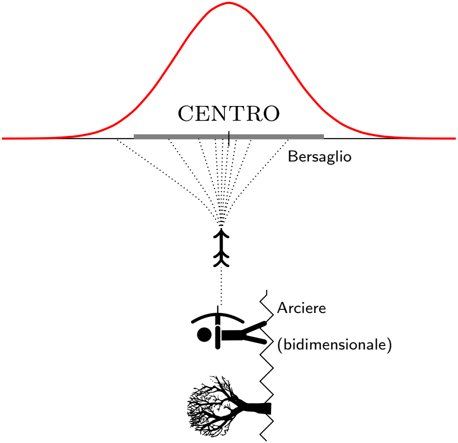


Un arciere lancia una freccia contro il centro di un bersaglio.

—

L'arciere non è perfetto, la freccia non colpirà esattamente il centro: il punto colpito è casuale.

L'arciere e la freccia



Un arciere lancia una freccia contro il centro di un bersaglio.

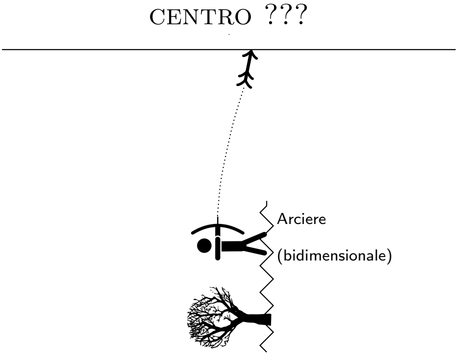
—

L'arciere non è perfetto, la freccia non colpirà esattamente il centro: il punto colpito è casuale.

—

Non è neppure del tutto incapace: il punto colpito si trova più probabilmente vicino al centro che lontano dal centro.

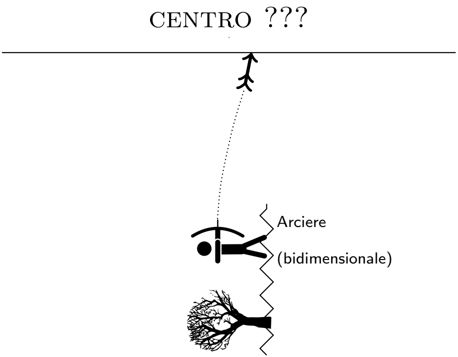
L'arciere e la freccia



Un arciere lancia una freccia contro il centro di un bersaglio.

Sappiamo dove la freccia ha colpito, non dove fosse il centro.

L'arciere e la freccia



Un arciere lancia una freccia contro il centro di un bersaglio.

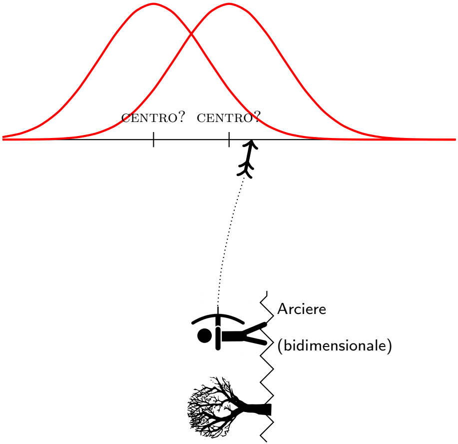
—

Sappiamo dove la freccia ha colpito, non dove fosse il centro.

—

Se l'arciere fosse infallibile → centro sulla freccia.

L'arciere e la freccia



Un arciere lancia una freccia contro il centro di un bersaglio.

—

Sappiamo dove la freccia ha colpito, non dove fosse il centro.

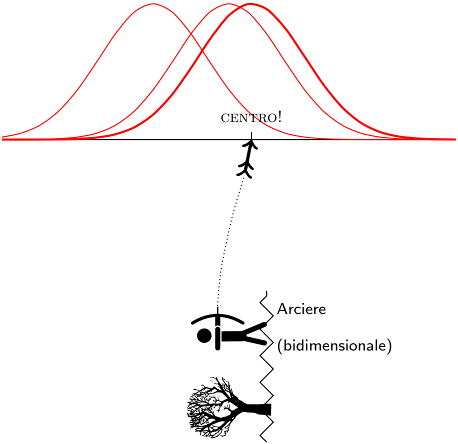
—

Se l'arciere fosse infallibile $\rightarrow$ centro sulla freccia.

—

Non è così, però è più verosimile che il centro fosse piazzato vicino alla freccia che lontano.

L'arciere e la freccia



Un arciere lancia una freccia contro il centro di un bersaglio.

—

Sappiamo dove la freccia ha colpito, non dove fosse il centro.

—

Se l'arciere fosse infallibile $\rightarrow$ centro sulla freccia.

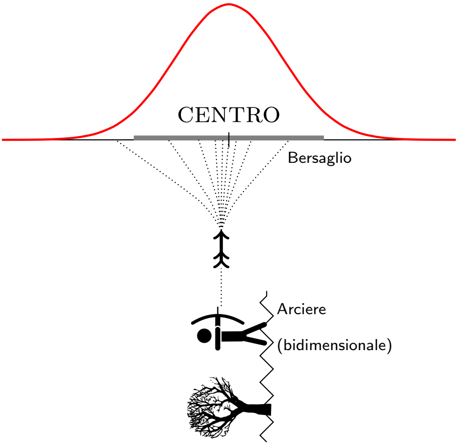
—

Non è così, però è più verosimile che il centro fosse piazzato vicino alla freccia che lontano.

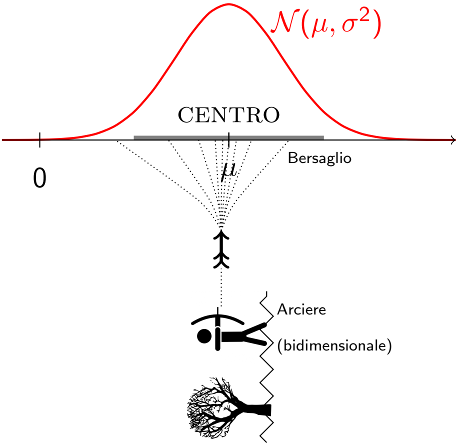
—

Migliore ipotesi: il centro era dove è caduta la freccia.

L'arciere e la freccia, qualcosa di più preciso



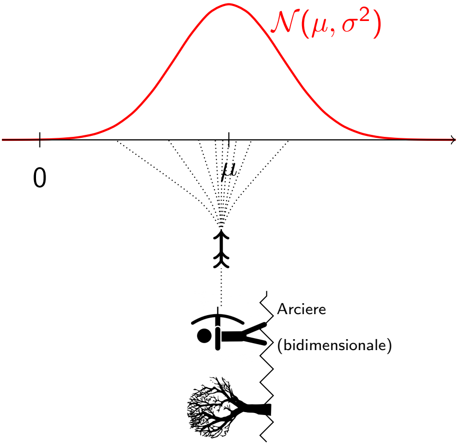
L'arciere e la freccia, qualcosa di più preciso



Aggiungiamo un asse di riferimento

- μ è il centro
- il punto che la freccia colpirà ha coordinata $X \sim \mathcal{N}(\mu, \sigma^2)$

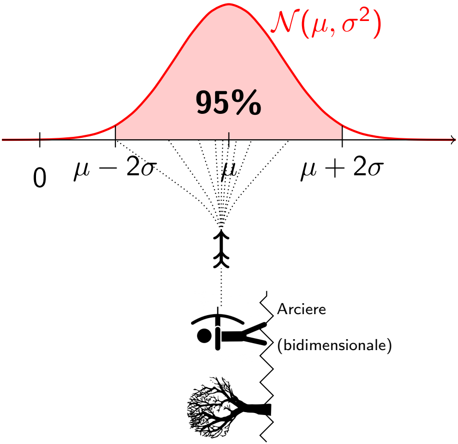
L'arciere e la freccia, qualcosa di più preciso



Aggiungiamo un asse di riferimento

- μ è il centro
- il punto che la freccia colpirà ha coordinata $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$

L'arciere e la freccia, qualcosa di più preciso



Aggiungiamo un asse di riferimento

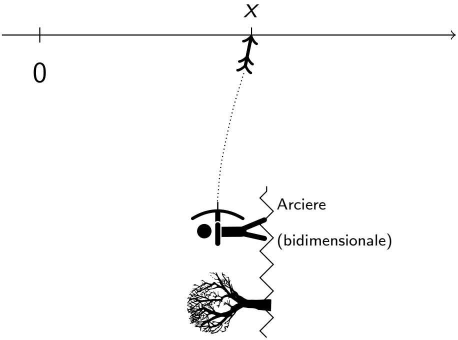
- μ è il centro
- il punto che la freccia colpirà ha coordinata $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$

Allora

$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 0.95$$

cioè sono abbastanza sicuro (al 95%) che la freccia (x) cada entro 2σ dal centro (μ).

L'arciere e la freccia, qualcosa di più preciso



Aggiungiamo un asse di riferimento

- μ è il centro
- il punto che la freccia colpirà ha coordinata $X \sim \mathcal{N}(\mu, \sigma^2)$

Allora

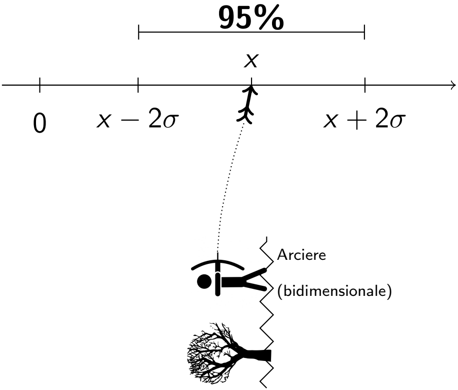
$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 0.95$$

cioè sono abbastanza sicuro (al 95%) che la freccia (x) cada entro 2σ dal centro (μ).

—

Ribaltando l'affermazione, una volta visto x, sono abbastanza sicuro (al 95%) che μ si trovi entro 2σ da x.

L'arciere e la freccia, qualcosa di più preciso



Aggiungiamo un asse di riferimento

- μ è il centro
- il punto che la freccia colpirà ha coordinata $X \sim \mathcal{N}(\mu, \sigma^2)$

Allora

$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 0.95$$

cioè sono abbastanza sicuro (al 95%) che la freccia (x) cada entro 2σ dal centro (μ).

Ribaltando l'affermazione, una volta visto x, sono abbastanza sicuro (al 95%) che μ si trovi entro 2σ da x.

Inferenza statistica

L'operazione per cui, a partire da dove la freccia ha colpito, cerchiamo di inferire dov' il centro (il punto a cui l'arciere hamirato) la chiamiamo **inferenza statistica**.

Essa non dà la certezza su dove fosse il centro, ma al più una ragionevole certezza.

In altre parole, sappiamo che sbagliamo, ma possiamo quantificare quanto probabilmente stiamo sbagliando.

Nel seguito, caliamo il ragionamento appena fatto in alcune situazioni più interessanti e generali.

Inferenza statistica, un po' di terminologia

Il centro del bersaglio, quello che vorrei conoscere ma non osservo, è detto parametro, esso è una caratteristica di una popolazione.

Il punto colpito, X, è detto stimatore.

Un aspetto chiave è capire "quanto è preciso l'arciere", cioè come è fatta la distribuzione del punto colpito X.

La distribuzione di X è detta distribuzione campionaria, nella maggior parte dei casi che vedremo è una normale, si tratterà di trovare la varianza.

Dobbiamo capire cosa sia l'arciere (o la freccia), in pratica è il campione, cioè il risultato dell'osservazione di una parte casuale della popolazione.

Indice

Esempio: sondaggi elettorali

- Della popolazione degli elettori francesi vogliamo conoscere la proporzione che voterà Macron
  senza (prima di) osservare tutta la popolazione.
- Scegliamo **a caso** alcuni individui: il campione.
- Sulla base di questi otteniamo una stima del parametro percentuale di elettori di Macron,

Campioni e stime

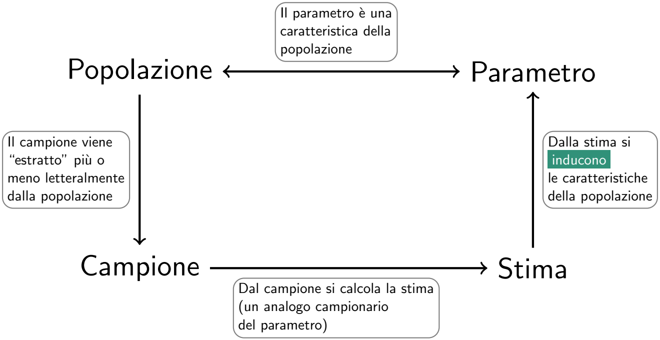
La prima cosa da osservare è che il risultato di un campionamento è casuale, nel contesto dei sondaggi elettorali possiamo osservarlo.

Il problema infatti se lo pongono in molti, e molti fanno dei sondaggi, ad esempio qui ne sono due del 24/4.

sondaggista	Macron	Le Pen
Ipsos	64	36
Harris	62	38

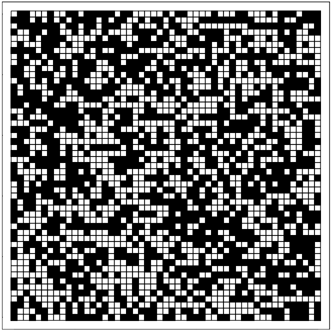
- Nei due sondaggi s'ottengono stime diverse, eppure sono contemporanei e popolazione e parametro sono gli stessi.
- Sono però intervistate persone diverse, e quindi il risultato è casuale.
- Esso è comunque utile in quanto è più probabile che sia vicino alla caratteristica di interesse della popolazione che non che sia lontano.

Popolazione-campione-parametro-stima



Si noti che questo schema è utile per capire il principio alla base dell'inferenza, è però restrittivo e non è adatto a descriverne tutte le applicazioni.

Esempio dei bianchi e dei neri: deduzione probabilistica

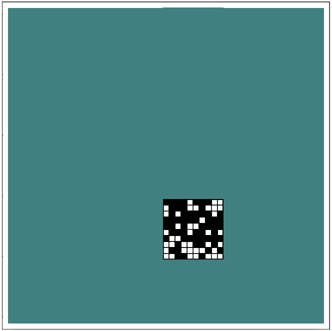


Popolazione dei
bianchi e dei neri.

La popolazione è nota: il $\pi = 50\%$ dei quadratini è nero.
(Nota: di solito il parametro si indica con una lettera greca, qui π.)

Sapendo che $\pi = 50\%$ è nero.

Esempio dei bianchi e dei neri: deduzione probabilistica



Popolazione dei bianchi e dei neri.

La popolazione è nota: il $\pi = 50\%$ dei quadratini è nero.
(Nota: di solito il parametro si indica con una lettera greca, qui π.)

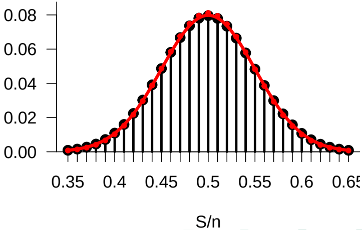
Sapendo che $\pi = 50\%$ è nero.

$\downarrow$

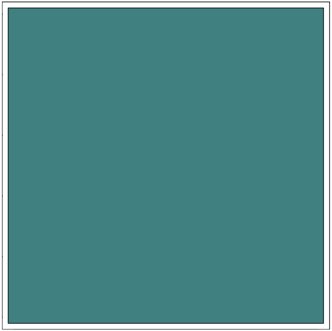
Se estraggo a caso una porzione della popolazione, diciamo di $n = 100$ unità, osservo un numero di neri, S, che è casuale

Le nozioni di CdP consentono di calcolare la probabilità di osservare una proporzione S/n di neri, approssimativamente

$$\bar{S} = S/n \sim \mathcal{N}\left(\pi, \frac{\pi(1-\pi)}{n}\right)$$



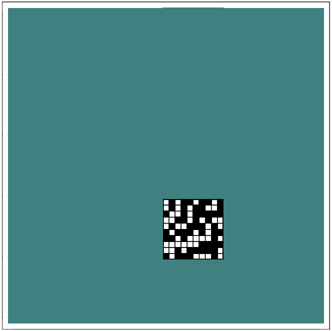
Esempio dei bianchi e dei neri: inferenza statistica



Popolazione dei bianchi e dei neri.

La popolazione non è nota: $\pi =?$ (prop "neri").

Esempio dei bianchi e dei neri: inferenza statistica



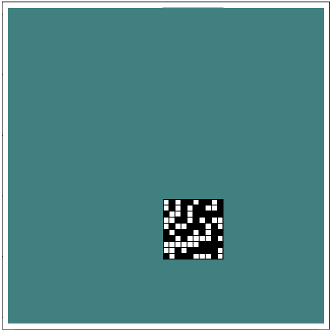
Popolazione dei bianchi e dei neri.

—

La popolazione non è nota: $\pi =?$ (prop "neri").

Osservo solo il campione $S = 53$ su $n = 100$.

Esempio dei bianchi e dei neri: inferenza statistica



Popolazione dei bianchi e dei neri.

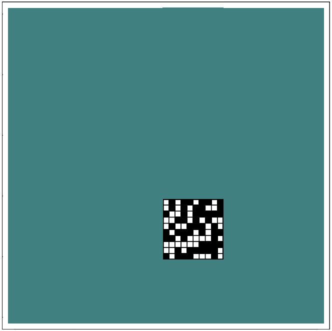
La popolazione non è nota: $\pi =?$ (prop "neri").

Distribuzione campionaria

$$\bar{S} = S/n \sim \mathcal{N}\left(\pi, \frac{\pi(1-\pi)}{n}\right)$$

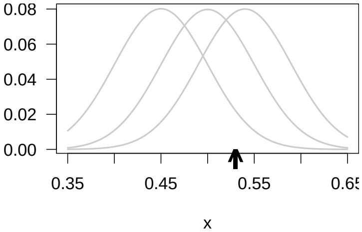
Osservo solo il campione $S = 53$ su $n = 100$.

Esempio dei bianchi e dei neri: inferenza statistica



Popolazione dei bianchi e dei neri.

La popolazione non è nota: $\pi =$? (prop "neri").



$\bar{S}$ è "probabilmente" vicino a π (LGN), esso sarà la mia stima di π: $\hat{\pi} = \bar{S}_n$.

Distribuzione campionaria

$$\bar{S} = S/n \sim \mathcal{N}\left(\pi, \frac{\pi(1-\pi)}{n}\right)$$

Osservo solo il campione $S = 53$ su $n = 100$.

Precisione, al 95%

Dalla distribuzione campionaria di $\hat{\pi}$

$$\hat{\pi} = \bar{S} = S/n \sim \mathcal{N}\left(\pi, \frac{\pi(1-\pi)}{n}\right)$$

possiamo scrivere

$$P\left(\pi - 1.96\sqrt{\frac{\pi(1-\pi)}{n}} \leq \hat{\pi} \leq \pi - 1.96\sqrt{\frac{\pi(1-\pi)}{n}}\right) = 0.95$$

cioè siamo "quasi sicuri" che la proporzione osservata cada a non più di $1.96\sqrt{\frac{\pi(1-\pi)}{n}}$ dalla proporzione nella popolazione, cioè siamo "quasi sicuri" che la proporzione non osservabile π sia nell'intervallo ()

$$\left[\hat{\pi} - 1.96\sqrt{\frac{\pi(1-\pi)}{n}}, \hat{\pi} + 1.96\sqrt{\frac{\pi(1-\pi)}{n}}\right]$$

Precisione, al 95%

Dalla distribuzione campionaria di $\hat{\pi}$

$$\hat{\pi} = \bar{S} = S/n \sim \mathcal{N}\left(\pi, \frac{\pi(1-\pi)}{n}\right)$$

possiamo scrivere

$$P\left(\pi - 1.96\sqrt{\frac{\pi(1-\pi)}{n}} \leq \hat{\pi} \leq \pi - 1.96\sqrt{\frac{\pi(1-\pi)}{n}}\right) = 0.95$$

cioè siamo "quasi sicuri" che la proporzione osservata cada a non più di $1.96\sqrt{\frac{\pi(1-\pi)}{n}}$ dalla proporzione nella popolazione, cioè siamo "quasi sicuri" che la proporzione non osservabile π sia nell'intervallo (sostituendo $\hat{\pi}$ per π nella varianza)

$$\left[\hat{\pi} - 1.96\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}, \hat{\pi} + 1.96\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right]$$

Precisione, al 99%

Se vogliamo una maggiore certezza scriviamo che

$$P\left(\pi - 2.57\sqrt{\frac{\pi(1-\pi)}{n}} \leq \hat{\pi} \leq \pi - 2.57\sqrt{\frac{\pi(1-\pi)}{n}}\right) = 0.99$$

cioè siamo sicuri al 99% che la proporzione osservata cada a non più di $2.57\sqrt{\frac{\pi(1-\pi)}{n}}$ dalla proporzione nella popolazione, cioè che

$$\left[\hat{\pi} - 2.57\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}, \hat{\pi} + 2.57\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right]$$

Naturalmente paghiamo la maggiore certezza, rispetto a prima, con un intervallo più ampio.

Errore e sondaggi

Nel sondaggio Harris, Macron era dato al 62%, cosa significa questo?

- Dovremmo conoscere la numerosità del campione, facciamo finta che sia $n = 1000$ (numero titpico per i sondaggi elettorali).
- Facciamo finta che il risultato provenga da un campione di 1000 individui scelti a caso e che tutti rispondano sinceramente (nei sondaggi elettorali tipicamente non è così, ne parleremo).

Con queste precisazioni, l'aver osservato il 62% significa che, al 95% l'errore non supera

$$1.96\sqrt{\frac{0.62(1-0.62)}{1000}} = 0.03$$

e quindi si ha una "sicurezza al 95%" che la ver apercentuale sia nell'intervallo

$$[0.62 - 0.03, 0.62 + 0.03] \quad \rightarrow \quad [0.59, 0.65]$$

Errore e sondaggi

Nel sondaggio Harris, Macron era dato al 62%, cosa significa questo?

- Dovremmo conoscere la numerosità del campione, facciamo finta che sia $n = 1000$ (numero titpico per i sondaggi elettorali).
- Facciamo finta che il risultato provenga da un campione di 1000 individui scelti a caso e che tutti rispondano sinceramente (nei sondaggi elettorali tipicamente non è così, ne parleremo).

Se vogliamo una "sicurezza al 99%" calcoleremo l'errore

$$2.57\sqrt{\frac{0.62(1-0.62)}{1000}} = 0.04$$

e quindi avremo l'intervallo, più ampio

$$[0.62 - 0.04, 0.62 + 0.04] \quad \rightarrow \quad [0.58, 0.66]$$

Come scelgo n

Un punto, nella scelta di n è capire qual è il mio errore accettabile, calcoliamo allora l'errore per diversi n a un dato livello di "sicurezza", diciamo il 95%

$$n = 1000 \quad \rightarrow \quad 1.96\sqrt{\frac{\pi(1-\pi)}{1000}} \leq 1.96\sqrt{\frac{0.5(1-0.5)}{1000}} = 0.031$$

Come scelgo n

Un punto, nella scelta di n è capire qual è il mio errore accettabile, calcoliamo allora l'errore per diversi n a un dato livello di "sicurezza", diciamo il 95%

$$n = 500 \quad \rightarrow \quad 1.96\sqrt{\frac{\pi(1-\pi)}{500}} \leq 1.96\sqrt{\frac{0.5(1-0.5)}{500}} = 0.043$$

$$n = 1500 \quad \rightarrow \quad 1.96\sqrt{\frac{\pi(1-\pi)}{1500}} \leq 1.96\sqrt{\frac{0.5(1-0.5)}{1500}} = 0.025$$

$$n = 2000 \quad \rightarrow \quad 1.96\sqrt{\frac{\pi(1-\pi)}{2000}} \leq 1.96\sqrt{\frac{0.5(1-0.5)}{2000}} = 0.021$$

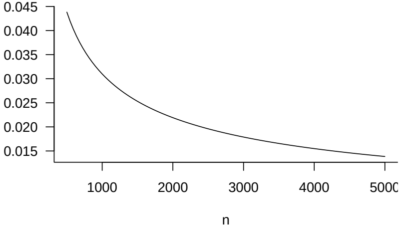
Si noti che diminuisce più lentamente di quanto non aumenti.

Come scelgo n

Un punto, nella scelta di n è capire qual è il mio errore accettabile, calcoliamo allora l'errore per diversi n a un dato livello di "sicurezza", diciamo il 95%

Guardiamo al grafico dell'errore in funzione di n, cioè alla funzione

$$1.96\sqrt{\frac{0.5(1-0.5)}{n}}$$



Per dimezzare l'errore devo quadruplicare il campione!

Numerosità campionaria per un dato errore

È utile anche invertire la formula sopra, a partire dalla generica

$$P\left(|\hat{\pi}-\pi| \leq z_{1-\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}}\right) = 1-\alpha$$

cioè con probabilità $1-\alpha$ l'errore è inferiore a

$$z_{1-\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}} \leq z_{1-\alpha/2}\frac{0.5}{\sqrt{n}}$$

Se dunque vogliamo che l'errore sia inferiore a un certo livello c con probabilità $1-\alpha$ occorre che

$$c \geq z_{1-\alpha/2}\frac{0.5}{\sqrt{n}} \Rightarrow \sqrt{n} \geq 2\frac{z_{1-\alpha/2}}{c} \Rightarrow n \geq 4\left(\frac{z_{1-\alpha/2}}{c}\right)^2$$

Indice

Numerosità campionaria e precisione dello stimatore

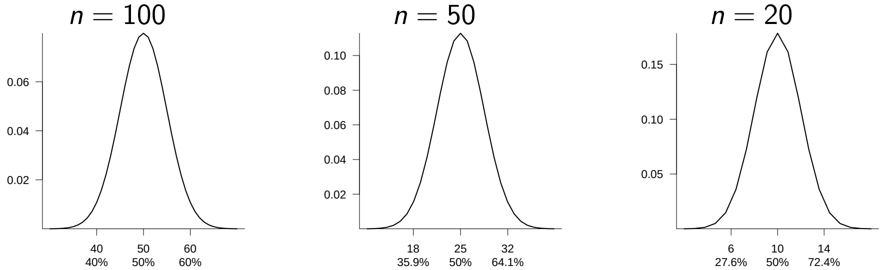
- Nella cittadina di Dunwich, che conta 100 000 abitanti, si terranno a breve le elezioni comunali.
- Ci sono due soli candidati, Arthur Dupin (attuale sindaco) e Auguste Pym.
- L'attuale sindaco Arthur Dupin commissiona un sondaggio: 500 persone vengono intervistate (e tutti rispondono, è una storia 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?

Bianchi e neri: dimensione campionaria

Confrontiamo le distribuzioni del numero e della percentuale di neri in campioni di diversa numerosità



Si noti che in tutti i casi la distribuzione è centrata sul 50%, ma **quanto più è grande il campione tanto più è concentrata intorno al 50%.**

Indice

Ruolo della dimensione della popolazione

- Auguste Pym che, ricordiamolo, è dato favorito come candidato sindaco di Dunwich (100 000 abitanti) secondo un sondaggio che ha coinvolto 1000 persone, si vanta con un amico, William Wilson, anche lui politico, del fatto appunto di essere il favorito nella sua competizione elettorale.
- L'amico William è anch'egli candidato sindaco, ma nella città di Innsmouth (1 000 000 di 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 100 000 (l'1 per cento), il tuo 1000 su 1 000 000 (l'1 per mille), io sono molto più sicuro di te.

Auguste ha ragione?

Campionare è come assaggiare una zuppa durante la cottura

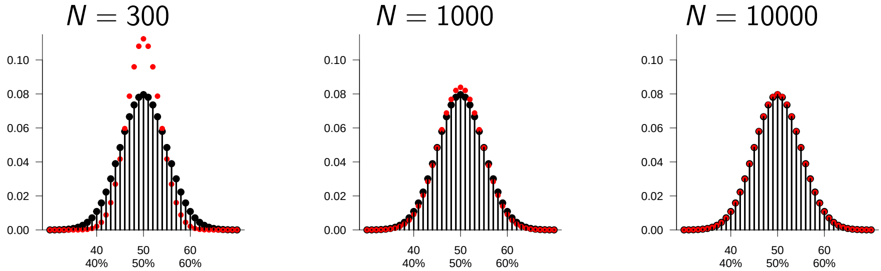
- Quando osserviamo un campione osserviamo una parte della popolazione e ne inferiamo le caratteristiche dell'intera popolazione.
- Quando assaggiamo la zuppa per vedere se è giusta di sale ne mangiamo un cucchiaio, verifichiamo che quel cucchiaio sia correttamente salato.
- Se la zuppa è stata adeguatamente 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).

Bianchi e neri: dimensione della popolazione

Distribuzioni campionarie del numero di neri in un campione di dimensione 100 da popolazioni di diverse dimensioni, in nero l'approssimazione binomiale, in rosso la distribuzione esatta.



L'approssimazione binomiale non cambia, evidentemente, al variare della dimensione della popolazione.

Viceversa se la popolazione è piccola rispetto al campione, la distribuzione ipergeometrica è più concentrata intorno al vero valore.

Indice

Come dev'essere il campione

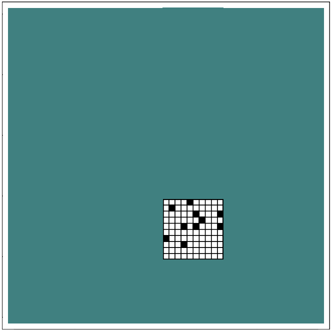
Quando diciamo che il campione sono n individui selezionati nella popolazione, questo non vuol dire che qualunque gruppo di n individui vada bene.

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.

Bianchi e neri: campione distorto



Popolazione dei bianchi e dei neri.

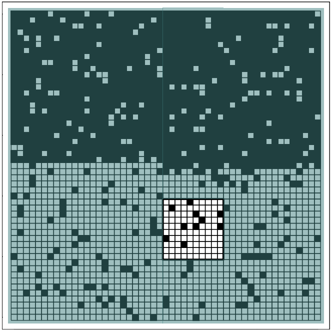
La popolazione non è nota: $\pi =?$ (prop "neri").

La popolazione ha la stessa composizione della precedente (50% di neri).

—

Oserviamo un campione aprendo una "finestra" come fatto precedentemente, si osservano 10 neri su 100.

Bianchi e neri: campione distorto



Popolazione dei bianchi e dei neri.

La popolazione non è nota: $\pi =?$ (prop "neri").

La popolazione ha la stessa composizione della precedente (50% di neri).

—

Oserviamo un campione aprendo una "finestra" come fatto precedentemente, si osservano 10 neri su 100.

—

Si ha però che bianchi e neri non sono omogenei nel quadrato, per cui con la finestra abbiamo ottenuto un campione **non rappresentativo.**

Come NON devo costruire 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 (a meno di ipotesi aggiuntive).

Come ottengo un campione rappresentativo?

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

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

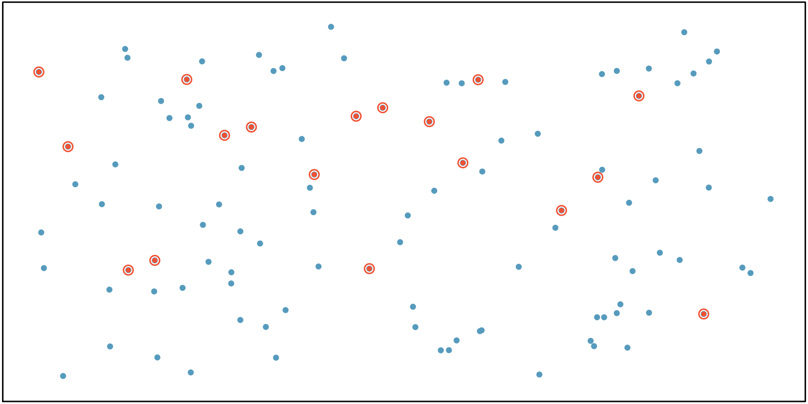
  tra queste
  - campione stratificato;
  - campione a grappoli;
  - campione a più stadi.

  Vale sempre il principio che tutti possono essere estratti, le probabilità possono variare.

Indice

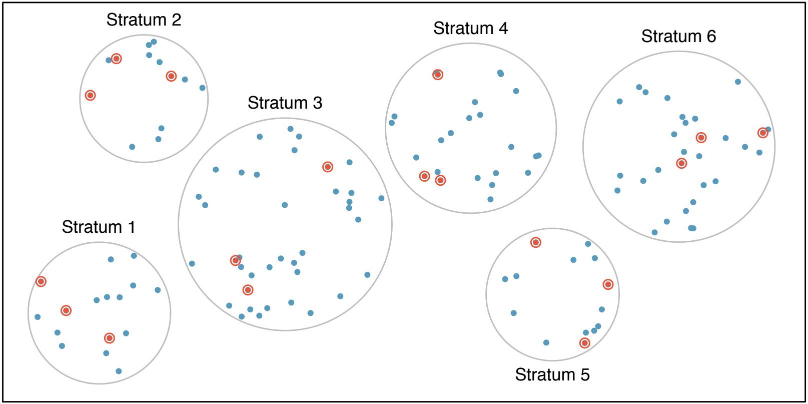
Campionamento casuale semplice (ccs)

Si selezionano le unità statistiche completamente a caso.



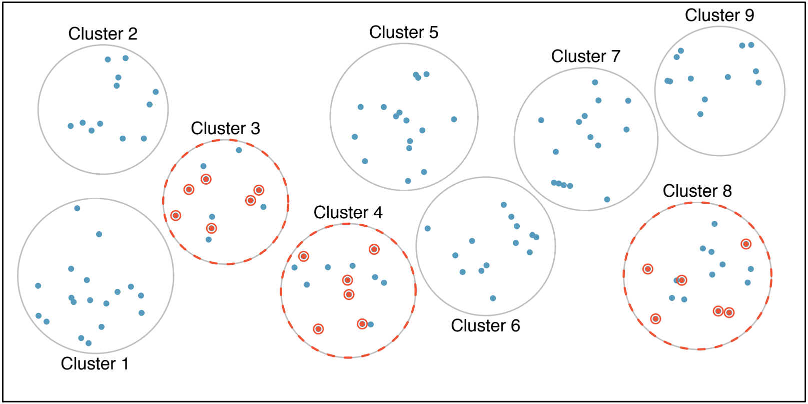
Campionamento stratificato

La popolazione è divisa in strati di unità statistiche simili. Si prende un ccs da ogni strato.



Campionamento a grappolo

La popolazione è divisa in grappoli, che usualmente sono formati da osservazioni non omogenee. Si prende un ccs di grappoli. Nel campionamento a stadi, prima si prende un ccs di grappoli, poi un ccs da ciascun grappolo.



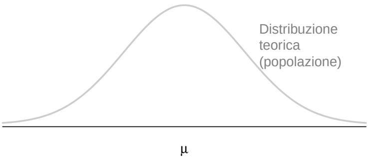
Campione nei sondaggi elettorali

I sondaggi elettorali sono fatti con una logica diversa, e i risultati che vengono forniti sono basati anche su considerazioni politologiche, e pone alcni problemi

- non copre tutta la popolazione ma solo le persone reperibili per telefono (spesso solo telefono fisso);
- la non risposta potrebbe essere collegata con il voto (si afferma spesso che chi vota a destra risponde meno frequentemente);
- momento del sondaggio temporalmente distante dall'elezione, le persone possono cambiare idea.
- Spesso, s'impiegano campioni per quote:
  - La popolazione è divisa in strati (e.g. per età e sesso).
  - All'interno degli strati il campione è scelto per convenienza.
  - Il campione somiglierà alla popolazione quanto alla frequenza delle caratteristiche oggetto di stratificazione.
  - Può esserci distorsione da selezione.
  - La generalizzabilità non è garantita.

Indice

Esempio gaussiana: controllo di processo



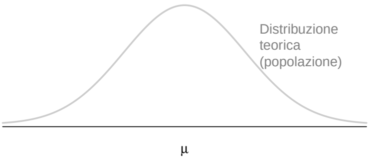
In grigio è rappresentata la distribuzione del peso di confezioni di un prodotto alimentare, che si suppone

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

In un processo industriale di confezionamento di un prodotto, si assume che il peso X di una confezione si distribuisca come una $\mathcal{N}\left(\mu, \sigma^2\right)$.

Il produttore è interessato a valutare μ (ad esempio per verificare che non si discosti troppo dal peso dichiarato sulle confezioni).

Esempio gaussiana: controllo di processo



In grigio è rappresentata la distribuzione del peso di confezioni di un prodotto alimentare, che si suppone

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

In un processo industriale di confezionamento di un prodotto, si assume che il peso X di una confezione si distribuisca come una $\mathcal{N}\left(\mu, \sigma^2\right)$.

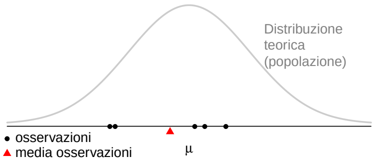
Il produttore è interessato a valutare μ (ad esempio per verificare che non si discosti troppo dal peso dichiarato sulle confezioni).

Se peso una confezione, il suo peso è $X_1 \sim \mathcal{N}\left(\mu, \sigma^2\right)$, col solito ragionamento ho la "sicurezza al 95%" che μ sia nell'intervallo

$$[X_1 - 1.96\sigma, X_1 + 1.96\sigma]$$

Posso però fare qualcosa di meglio.

Esempio gaussiana: controllo di processo



In grigio è rappresentata la distribuzione del peso di confezioni di un prodotto alimentare, che si suppone

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

In un processo industriale di confezionamento di un prodotto, si assume che il peso X di una confezione si distribuisca come una $\mathcal{N}\left(\mu, \sigma^2\right)$.

—

Il produttore è interessato a valutare μ (ad esempio per verificare che non si discosti troppo dal peso dichiarato sulle confezioni).

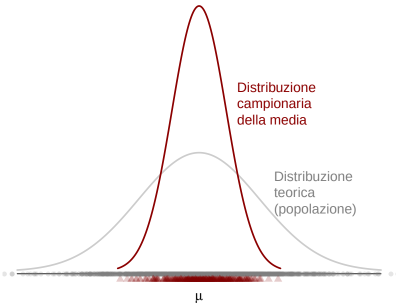
—

Pesiamo n confezioni a caso, il campione, e calcoliamo il peso medio

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

In termini di variabili aleatorie, $X_1, \ldots, X_n$ sono **indipendenti e identicamente distribuite (IID)**

Esempio gaussiana: controllo di processo



In grigio è rappresentata la distribuzione del peso di confezioni di un prodotto alimentare, che si suppone

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

In un processo industriale di confezionamento di un prodotto, si assume che il peso X di una confezione si distribuisca come una $\mathcal{N}(\mu, \sigma^2)$.

Il produttore è interessato a valutare μ (ad esempio per verificare che non si discosti troppo dal peso dichiarato sulle confezioni).

Pesiamo n confezioni a caso, il campione, e calcoliamo il peso medio

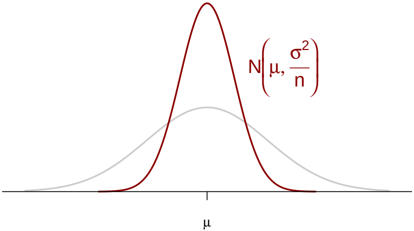
$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

In termini di variabili aleatorie, $X_1, \ldots, X_n$ sono **indipendenti e identicamente distribuite (IID)**, la distribuzione campionaria di $\bar{X}$ è

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

Precisione dello stimatore media campionaria

Consideriamo lo stimatore media campionaria, $\bar{X}$ per la media di una normale



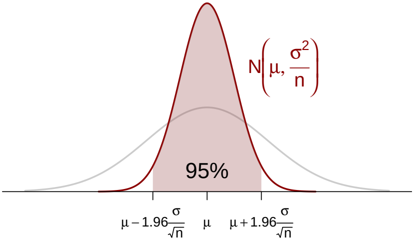
$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

La distribuzione è centrata su μ, sono più probabili valori vicini a μ che valori lontani.

Precisione dello stimatore media campionaria

Consideriamo lo stimatore media campionaria, $\bar{X}$ per la media di una normale



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

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

La distribuzione è centrata su μ, sono più probabili valori vicini a μ che valori lontani.

Possiamo però essere più precisi, sappiamo dalle proprietà della normale che

$$P\left(\mu - 1.96\frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu - 1.96\frac{\sigma}{\sqrt{n}}\right) \approx 0.95$$

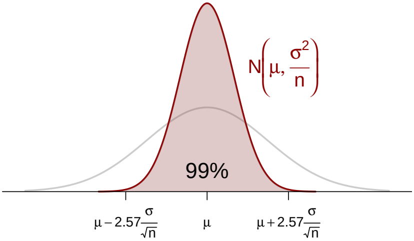
quindi possiamo scrivere

$$P\left(|\bar{X} - \mu| \leq 1.96\frac{\sigma}{\sqrt{n}}\right) \approx 0.95$$

cioè che l'errore è, al 95%, minore di $1.96\frac{\sigma}{\sqrt{n}}$.

Precisione dello stimatore media campionaria

Consideriamo lo stimatore media campionaria, $\bar{X}$ per la media di una normale



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

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

La distribuzione è centrata su μ, sono più probabili valori vicini a μ che valori lontani.

Possiamo avere maggiori certezze, ma il prezzo è ammettere un errore maggiore

$$P\left(|\bar{X} - \mu| \leq 2.57 \frac{\sigma}{\sqrt{n}}\right) \approx 0.99$$

cioè l'errore è, al 99%, minore di $2.57 \frac{\sigma}{\sqrt{n}}$.

Esempio

Si considera un processo industriale di confezionamento di un prodotto alimentare.

Si assume che il peso X di una confezione si distribuisca normalmente con s.q.m. $\sigma = 5$ intorno a una media μ.

Si osserva un campione di dimensione 10 e risulta la stima $\bar{X} = 98$, cosa possiamo dire dell'errore che commettiamo?

Con una probabilità del 95% l'errore è inferiore a

$$2\frac{5}{\sqrt{10}} = 2 \times 1.58 = 3.16$$

mentre con una probabilità del 99% l'errore è inferiore a

$$3\frac{5}{\sqrt{10}} = 3 \times 1.58 = 4.74$$

Notiamo che il margine di errore non dipende dal valore di $\bar{X}$.

Esempio

Si considera un processo industriale di confezionamento di un prodotto alimentare.

Si assume che il peso X di una confezione si distribuisca normalmente con s.q.m. $\sigma = 5$ intorno a una media μ.

Si osserva un campione di dimensione 10 e risulta la stima $\bar{X} = 98$, cosa possiamo dire dell'errore che commettiamo?
Se aumentiamo la numerosità campionaria il margine si riduce, con un campione di dimensione 20 l'errore è al 95% inferiore a

$$2\frac{5}{\sqrt{20}} = 2 \times 1.12 = 2.24$$

Potremmo chiederci allora quanto dovrebbe essere grande il campione per avere un margine di ± 1 al 95%, impostiamo l'equazione

$$\frac{\sigma}{\sqrt{n}} = 0.5 \Leftrightarrow \sqrt{n} = \frac{1}{0.5}\sigma = 10 \Rightarrow n = 100$$

Numerosità campionaria per un dato errore

Campione normale

Nel caso della normale si ha,

$$P\left(|\bar{X}-\mu| \leq z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

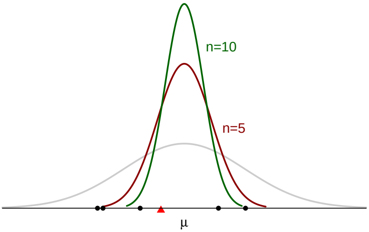
cioè con probabilità $1-\alpha$ l'errore è inferiore a

$$z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Se dunque vogliamo che l'errore sia inferiore a un certo livello c con probabilità $1-\alpha$ occorre che

$$c \geq z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \Rightarrow \sqrt{n} \geq \frac{\sigma z_{1-\alpha/2}}{c} \Rightarrow n \geq \left(\frac{\sigma z_{1-\alpha/2}}{c}\right)^2$$

Esempio gaussiana: dimensione campionaria

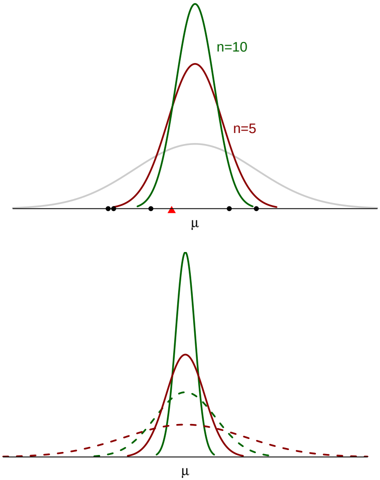


La distribuzione campionaria è

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

la varianza decresce quindi con n.

Esempio gaussiana: dimensione campionaria



La distribuzione campionaria è

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

la varianza decresce quindi con n.

Osserviamo che la varianza della media campionaria è più bassa allorché è più bassa la varianza della popolazione.

Indice

Stimatore per la varianza

Consideriamo nuovamente il caso $X \sim \mathcal{N}(\mu, \sigma^2)$, abbiamo definito uno stimatore per μ ($\bar{X}$, la media campionaria) ma non per σ^2, quando questa non è nota, però, andrà anch'essa stimata.

Un'idea intuitiva sarebbe la varianza campionaria

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

si mostra tuttavia che questo non è lo stimatore migliore, va preferito

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

detto **varianza campionaria corretta**.

Curiosità: distribuzione campionaria di S^2

Si mostra che, se la popolazione è normale, $X_i \sim \mathcal{N}(\mu, \sigma^2)$, e

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

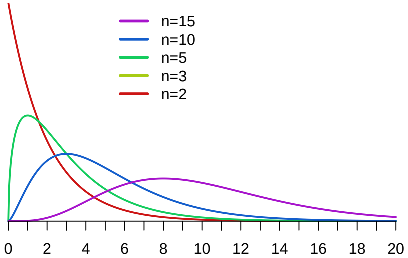
allora la variabile

$$V = \frac{(n-1)S^2}{\sigma^2}$$

è distribuita come un χ^2 con $n-1$ gradi di libertà.

(Non servirà, solo per dire che non è che tutte le distribuzioni campionarie sono normali.)

Curiosità: distribuzione campionaria di S^2: χ^2_n (chi-quadrato)



n	0.9	0.95	0.975	0.990	0.995	0.999
	Probabilità					
1	3.08	6.31	12.71	31.82	63.66	318.30
2	1.89	2.92	4.30	6.96	9.93	22.33
3	1.64	2.35	3.18	4.54	5.84	10.21
4	1.53	2.13	2.78	3.75	4.60	7.17
5	1.48	2.02	2.57	3.37	4.03	5.89
6	1.44	1.94	2.45	3.14	3.71	5.21
7	1.42	1.90	2.37	3.00	3.50	4.79
8	1.40	1.86	2.31	2.90	3.35	4.50
9	1.38	1.83	2.26	2.82	3.25	4.30
10	1.37	1.81	2.23	2.76	3.17	4.14
⋮	⋮	⋮	⋮	⋮	⋮	⋮

La distribuzione χ^2 con $n \in \mathbb{N}$ gradi di libertà, χ^2_n, ha f.d.d.

$$f_n(x) = \frac{1}{2^{n/2}\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-x/2}$$

per $x > 0$. Ha media n e varianza $2n$.

La tavola del χ^2 funziona analogamente a quella della t di Student.

Il quantile 0.99 del χ^2_7 è

$$\chi^2_{7,0.99} = 3.00$$

Precisione di uno stimatore: normale con varianza non nota

Nel caso di una normale con varianza nota abbiamo usato

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

che abbiamo anche scritto come

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

da cui

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

e quindi riguardo all'errore

$$P\left(|\bar{X} - \mu| \leq z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Precisione di uno stimatore: normale con varianza non nota

Nel caso di una normale con varianza nota abbiamo usato

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

che abbiamo anche scritto come

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

da cui

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

e quindi riguardo all'errore

$$P\left(|\bar{X} - \mu| \leq z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Questo però non funziona sostituendo σ con la stima $S = \sqrt{S^2}$.

Distribuzione campionaria della media della normale con varianza non nota

Si mostra che se $X_1, \ldots, X_n$ è un campione IID da una $\mathcal{N}(\mu, \sigma^2)$ la statistica

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

è distribuita come una t di Student con $n-1$ gradi di libertà.
Valgono quindi relazioni del tipo

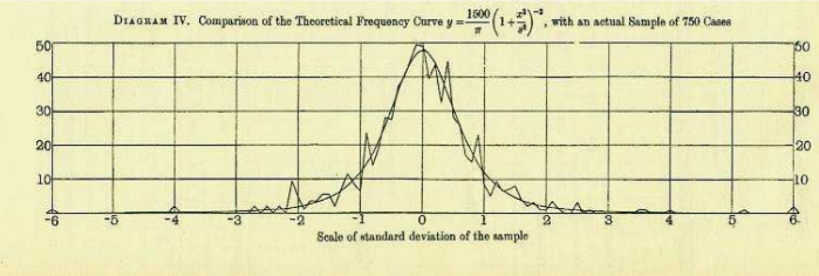
$$P\left(-t_{n-1,1-\alpha/2} \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq t_{n-1,1-\alpha/2}\right) = 1 - \alpha$$

$$P\left(|\bar{X} - \mu| \leq t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

Cioè, in sostanza, devo usare i quantili della distribuzione t e non quelli della normale.

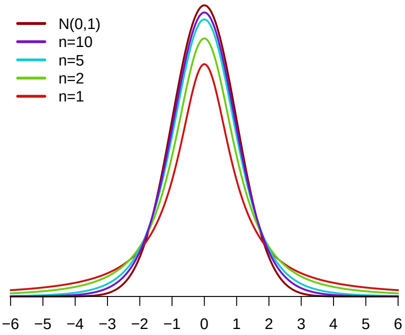
t di Student

La t di Student viene introdotta da William Gossett (alias Student, 1876-1937), responsabile della birrifcazione per la Guinness, per confrontare mediante campioni la qualità di diverse produzioni di birra.



t di Student

La t di Student viene introdotta da William Gossett (alias Student, 1876-1937), responsabile della birrifcazione per la Guinness, per confrontare mediante campioni la qualità di diverse produzioni di birra.



La t di student con $n \in \mathbb{N}$ gradi di libertà, t_n, ha f.d.d.

$$f_n(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}\left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

Ha una forma simile alla normale ma con code tanto più pesanti quanto più piccolo è n.

Tavola della t di Student

La tavola della t, diversamente da quella della normale, non riporta la FdR ma i quantili corrispondenti a particolari livelli di probabilità.

Le righe corispondono ai gradi di libertà.

Le colonne alle probabilità.

n	...	...	p	...
...	...	...	...	...
k	...	...	$t_{k,p}$	...
...	...	...	...	...

$t_{k,p}$ è il quantile di ordine p della t di Student con k gradi di libertà.

	Probabilità					
n	0.9	0.95	0.975	0.990	0.995	0.999
1	3.08	6.31	12.71	31.82	63.66	318.30
2	1.89	2.92	4.30	6.96	9.93	22.33
3	1.64	2.35	3.18	4.54	5.84	10.21
4	1.53	2.13	2.78	3.75	4.60	7.17
5	1.48	2.02	2.57	3.37	4.03	5.89
6	1.44	1.94	2.45	3.14	3.71	5.21
7	1.42	1.90	2.37	3.00	3.50	4.79
8	1.40	1.86	2.31	2.90	3.35	4.50
9	1.38	1.83	2.26	2.82	3.25	4.30
10	1.37	1.81	2.23	2.76	3.17	4.14
11	1.36	1.80	2.20	2.72	3.11	4.03
12	1.36	1.78	2.18	2.68	3.06	3.93
13	1.35	1.77	2.16	2.65	3.01	3.85
14	1.34	1.76	2.14	2.62	2.98	3.79
15	1.34	1.75	2.13	2.60	2.95	3.73
16	1.34	1.75	2.12	2.58	2.92	3.69
17	1.33	1.74	2.11	2.57	2.90	3.65
18	1.33	1.73	2.10	2.55	2.88	3.61
19	1.33	1.73	2.09	2.54	2.86	3.58
20	1.32	1.73	2.09	2.53	2.85	3.55
21	1.32	1.72	2.08	2.52	2.83	3.53
22	1.32	1.72	2.07	2.51	2.82	3.50
23	1.32	1.71	2.07	2.50	2.81	3.48
24	1.32	1.71	2.06	2.49	2.80	3.47
25	1.32	1.71	2.06	2.48	2.79	3.45
26	1.31	1.71	2.06	2.48	2.78	3.44
27	1.31	1.70	2.05	2.47	2.77	3.42
28	1.31	1.70	2.05	2.47	2.76	3.41
29	1.31	1.70	2.04	2.46	2.76	3.40
30	1.31	1.70	2.04	2.46	2.75	3.38

Esempio

Da una popolazione $\mathcal{N}\left(\mu, \sigma^2\right)$ estraggo il campione IID di $n = 8$ osservazioni

$$(3.5, 4.3, 5.5, 4.6, 5.3, 5.5, 5.1, 5.1)$$

Da esso ricaviamo media e varianza campionaria corretta

$$\bar{X} = 4.9, \quad S^2 = 0.5$$

Per ottenere un intervallo per cui siamo "al 95% sicuri" che la media μ vi caschi all'interno consideriamo il quantile 0.975 della t di Student con $n - 1 = 7$ gradi di libertà:

$$t_{7,0.975} = 2$$

e quindi otteniamo l'intervallo di estremi

$$4.9 \pm 2 \times \sqrt{\frac{0.5}{8}} \quad \rightarrow \quad [4.4, 5.4]$$

Indice

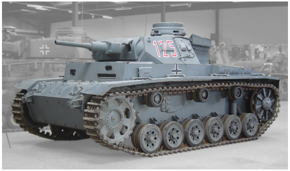
Le forze del nemico

- Siamo nel 1942, in piena IIGM.
- Uno degli elementi più importanti della strategia militare nella IIGM è costituito dai carri armati.
- Ovviamente, sapere di quanti carri armati disponesse il nemico costituirebbe un vantaggio.

- Mettiamoci nei panni degli alleati.
- Vorremmo stabilire quanti carri armati hanno a disposizione i tedeschi e quanti ne producono.
- Come?
  - ricorrendo alle spie (ad esempio, qualcuno con accesso agli archivi del ministero degli armamenti tedesco);
  - ricorrendo a informazioni prebelliche sulle capacità dell'industria tedesca;
  - C'è però anche un'opzione 'statistica'.

I numeri di serie!

- I carri armati sono numerati.
- Si sa che i carri sono numerati in sequenza.
- Si conoscono i numeri di serie di alcuni carri catturati.



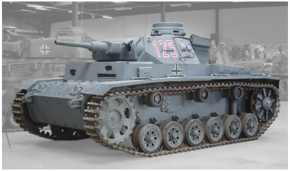
I numeri di serie!

- I carri armati sono numerati.
- Si sa che i carri sono numerati in sequenza.
- Si conoscono i numeri di serie di alcuni carri catturati.

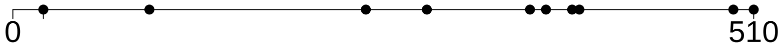
- **Popolazione**: θ carri, numerati da 1 a θ.
- **Parametro**: θ.
- **Campione**: n carri catturati, e i loro numeri, siano

$$243, 94, 510, 285, 385, 390, 367, 496, 21, 356$$

Come stimiamo θ?



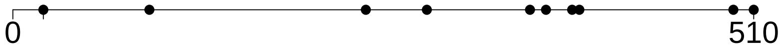
I numeri di serie!



Alcune statistiche:

- Media: $m = 314.7$;
- Mediana: $\text{Me} = 361.5$;
- Deviazione Standard: $\sigma = 160$;
- Minimo: $x_{(1)} = 21$;
- Massimo: $x_{(n)} = 510$

I numeri di serie!



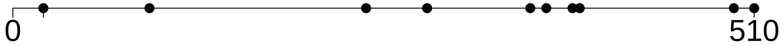
Alcune statistiche:

- Media: $m = 314.7$;
- Mediana: $\mathrm{Me} = 361.5$;
- Deviazione Standard: $\sigma = 160$;
- Minimo: $x_{(1)} = 21$;
- Massimo: $x_{(n)} = 510$

Stimatori proponibili

- $2m_1 = 629.4$
- $2\mathrm{Me} = 723$
- $m_1 + 2\sigma = 632.4$
- $x_{(n)} = 510$
- $x_{(n)} + x_{(1)} = 531$

I numeri di serie!



Alcune statistiche:

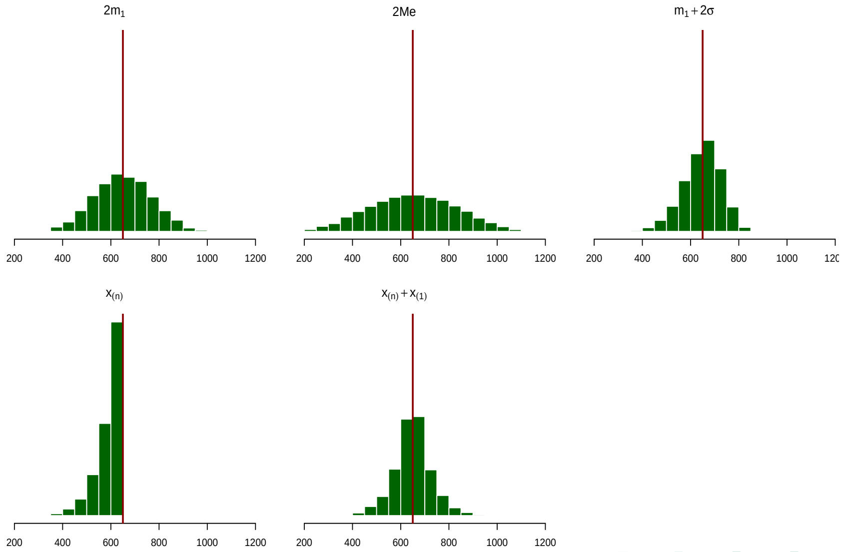
- Media: $m = 314.7$;
- Mediana: $\mathrm{Me} = 361.5$;
- Deviazione Standard: $\sigma = 160$;
- Minimo: $x_{(1)} = 21$;
- Massimo: $x_{(n)} = 510$

Stimatori proponibili

- $2m_1 = 629.4$
- $2\mathrm{Me} = 723$
- $m_1 + 2\sigma = 632.4$
- $x_{(n)} = 510$
- $x_{(n)} + x_{(1)} = 531$

Come facciamo a decidere qual è meglio?
Se ci riferiamo a questo solo campione è impossibile, se ci riferiamo alla distribuzione campionaria invece è possibile.

Distribuzione campionaria, $C = 650$



La realtà

La situazione reale era più complicata

- i numeri di serie sono più complessi (contengono di solito anno e mese di produzione, produttore, impianto);
- si possono usare i numeri di serie di varie componenti (ad esempio per i carri armati risultava molto utile quello della scatola del cambio).

Quindi in realtà

- Si sono usati metodi più articolati per tenere conto di tutte le informazioni;
- Non si otteneva solo un numero totale, ma maggiore dettaglio, ad esempio la produzione
  - per i vari impianti (utile per scegliere dove bombardare);
  - per diversi periodi (utile per vedere, ad es. se dopo un bombardamento la produzione diminuiva e quindi misurarne il successo).

La metodologia alla base è comunque quella illustrata.

Com'è andata?

Dopo la fine della guerra, si sono potute confrontare le varie previsioni con i dati sulla produzione

Date	Estimated Monthly Production: Serial Number Estimate	Estimated Monthly Production: Munitions Record 10 Aug. 42	Monthly Production Speer Ministry
June, 1940	169	1000	122
June, 1941	244	1550	271
August, 1942	327	1550	342

Indice

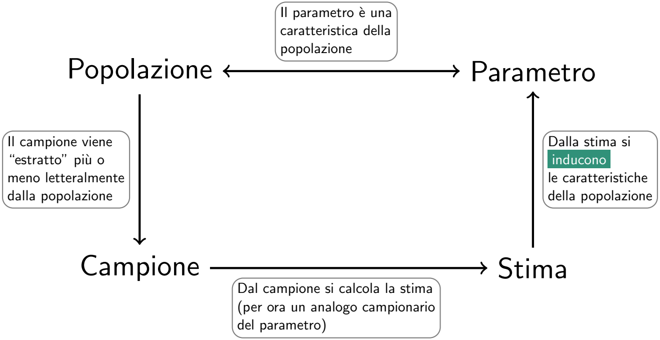
Campionamento e inferenza /1

Nell'accezione concettualmente più semplice, abbiamo un aggregato di interesse (**popolazione**), di cui ci interessa una qualche caratteristica (**parametro**).
Non possiamo osservare l'intero aggregato ma possiamo osservarne una parte (**campione**).
L'inferenza statistica consiste nel trarre dalle caratteristiche del campione conclusioni sulle caratteristiche della popolazione.

Inferenza statistica

L'inferenza statistica consiste nell'indurre le caratteristiche di un aggregato popolazione a partire dall'osservazione di una parte di esso, il campione.

Popolazione-campione-parametro-stima



Si noti che questo schema è utile per capire il principio alla base dell'inferenza, è però restrittivo e non è adatto a descriverne tutte le applicazioni.

Per essere effettivamente generali occorre maggiore astrazione.

Esempio: consumi

Quanto spende in media una famiglia italiana ogni mese?

Secondo l'ISTAT la spesa media è 2480 € (Indagine corrente sui consumi delle famiglie) e l'intervallo di confidenza è 2450-2510 €.

Non sempre la popolazione è composta da individui.

Assicurazioni

Per determinare il premio di una polizza assicurativa, una compagnia deve valutare l'esborso medio per assicurato.

$$
\begin{tikzcd}
\begin{array}{c}\textit{Assicurati}\\ \textit{(futuri assicurati)}\end{array} \arrow[r, leftrightarrow] \arrow[d] & \begin{array}{c}\textit{Esborso medio}\\ \textit{(futuro esborso)}\end{array} \\
\begin{array}{c}\textit{Assicurati attuali}\\ \textit{(e passati)}\end{array} \arrow[r] & \textit{Premio} \arrow[u]
\end{tikzcd}
$$

La compagnia si baserà sull'andamento dei sinistri nelle polizze degli anni passati (cioè, in sostanza, su quanto pagato di indennizzi in media negli anni passati, eventualmente distinguendo per caratteristiche della polizza, motivo per cui nell'RC auto un neopatentato paga più di un 50enne).

- La popolazione di interesse non è un vero aggregato.
- Il campione non è una parte della popolazione e non è scelto casualmente.

Concessione di crediti

Una banca vuole valutare la probabilità che un potenziale cliente onori un debito sulla base di sue caratteristiche.

$$
\begin{tikzcd}
\textit{Generico debitore (futuri debitori)} \arrow[r, leftrightarrow] \arrow[d] & \textit{\% insolventi} \\
\textit{Passati debitori} \arrow[r] & \textit{\# insolventi} \arrow[u]
\end{tikzcd}
$$

Ci si basa sulle insolvenze osservate in passato (tra i passati clienti debitori).

- Il campione non è un sottoinsieme della popolazione: assumiamo che i futuri clienti siano analoghi ai passati (condizionatamente all'avere analoghe caratteristiche) e quindi restituiscano il dovuto con la stessa probabilit'a.
- Correzioni per fattori 'ambientali', ad es. il periodo di crisi.

Mercati finanziari

Consideriamo un fondo di investimento: qual è la probabilità che aumenti di valore il prossimo anno?

Qui i metodi per la previsione sono tanti e diversi, in breve

- si osserva l'andamento dei prezzi dei valori mobiliari nel passato.
- si costruisce un modello matematico/statistico che si pensa possa descrivere l'andamento in futuro.
- (Il futuro ripeterà il passato, o no?)

Efficacia di un farmaco

La verifica dell'efficacia di un farmaco è una delle applicazioni tipiche della statistica oggi.
Una casa farmaceutica vuole verificare l'efficacia di un farmaco contro l'insonnia

- Consideriamo n pazienti che soffrono di insonnia,
- li dividiamo casualmente in due gruppi,
- a uno somministriamo il farmaco, all'altro il placebo.

- Misuriamo in qualche modo la qualità del sonno nei due gruppi, ad esempio con la latenza (tempo necessario ad addormentarsi dopo un risveglio notturno).
- Confrontiamo statisticamente i due gruppi.

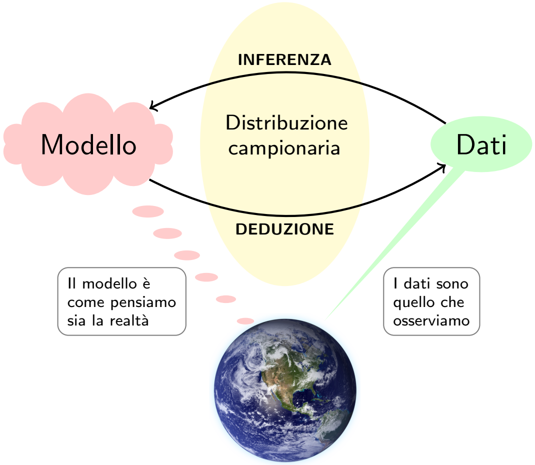
Epidemiologia

Tecniche statistiche possono essere impiegate per rispondere a domande del tipo: **la ferriera di Servola fa male?**.

Più precisamente, possono servire a confermare o smentire l'ipotesi per cui gli inquinanti emessi dalla ferriera portino a un aumento di casi di determinate patologie in chi vi è esposto (lavoratori o abitanti della zona) e anche a quantificare l'aumento di rischio per il singolo.

Questioni tipo i danni da esposizione da amianto o da inquinanti atmosferici sono analoghe.

Schema generale dell'inferenza



Rappresentatività in generale

Cosa succede quando il campione non è un sottoinsieme casuale della popolazione, come in molti esempi sopra?

- Efficacia di un farmaco:
  - il campione non sempre è casuale,
  - le caratteristiche del campione ne determinano la generalizzabilità (ad es. un esperimento fatto su pazienti americani non è immediatamente generalizzabile a europei)
- Concessione di crediti/premi assicurativi:
  - il campione non è casuale e non fa parte della popolazione,
  - assumo che i nuovi clienti si comportino analogamente ai vecchi,
  - (più ragionevole se uso le caratterisitche)
- Finanza:
  - si assume che il futuro "ripeta" il passato,
  - (nel 2008 non è andata tanto bene...)

Infine, in alcuni casi, tipo quello l'esempio citato della ferriera, non c'è un problema di generalizzazione, lo scopo dell'inferenza essendo di analizzare i dati per capire cosa è successo, cioè di interpretare quanto osservato.

Nel seguito

Nel seguito non consideriamo campioni complessi ma il caso del campione casuale semplice.

In pratica, assumiamo di avere un campione di osservazioni indipendenti e aventi tutte la distribuzione della popolazione (di interesse).
In simboli, abbiamo un campione $X_1, \dots, X_n$ indipendente e identicamente distribuito secondo $f(x; \theta)$ dove f è la distribuzione nella popolazione e θ è il parametro (o i parametri).

- Nell'esempio dei bianchi/neri osserviamo $X_1, \dots, X_n$ indipendenti dove X_i vale 1 se l'i-esima osservazione è nera e 0 altrimenti e $P(X_i = 1) = \pi$.
- Nell'esempio del controllo di processo osserviamo $X_1, \dots, X_n$ indipendenti dove $X_i \sim \mathcal{N}(\mu, \sigma^2)$.

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.