Statistica

Introduzione alla stima per intervalli: intervalli di confidenza

Andrea Giommi

Dipartimento di Statistica, Informatica, Applicazioni (DiSIA)
Università degli Studi di Firenze

Scuola di Psicologia
Corso di Studio in Scienze e Tecniche Psicologiche

Intervalli di confidenza

- La stima puntuale utilizza le osservazioni di un campione casuale per ottenere una stima del parametro tramite un singolo valore numerico
- Tale approccio possiede un punto di debolezza: La stima ottenuta sul campione osservato potrebbe differire molto dal valore del parametro nella popolazione
- È dunque opportuno che l'inferenza su un certo parametro si basi non solo sulla stima puntuale ma dia informazioni anche su quanto precisa sia la stima, ossia su quanto è probabile che la stima sia vicina al vero valore del parametro
- A tal fine si considera oltre alla stima puntuale, un *intervallo* di stime plausibili al quale sia associato un fissato *livello di fiducia/confidenza*
- Obiettivo: Determinare un intervallo di numeri intorno alla stima puntuale che ci aspettiamo contenga, con un certo livello di fiducia, il valore del parametro

Intervalli di confidenza

- Un intervallo di confidenza è un intervallo di numeri centrato sulla stima puntuale che con un fissato livello di probabilità contiene il vero valore del parametro
- La probabilità che l'intervallo di confidenza contenga il vero valore del parametro è detto livello di confidenza o livello di fiducia
- In genere si scelgono livelli di confidenza prossimi a 1: $0.9, 0.95, 0.99$

Intervallo di confidenza per un parametro

Obiettivo: determinare due statistiche campionarie:

$$L_I = L_I(Y_1, \ldots, Y_n) \qquad \text{e} \qquad L_S = L_S(Y_1, \ldots, Y_n)$$

- $L_I \leq L_S$ per ogni possibile campione; e
- L'intervallo $[L_I, L_S]$ contiene il parametro θ con probabilità $1 - \alpha$

$$P(L_I \leq \theta \leq L_S) = 1 - \alpha$$

- $1 - \alpha$ è detto livello di fiducia o livello di confidenza
- Una volta estratto il campione si ottiene l'intervallo di confidenza stimato: $[\ell_I; \ell_S]$
- Non è possibile sapere se l'intervallo stimato contenga o meno il valore vero del parametro
- La chiave per costruire un intervallo di confidenza è la distribuzione campionaria dello stimatore utilizzato per ottenere la stima puntuale
- La distribuzione campionaria dello stimatore permette di determinare la probabilità che lo stimatore produca una stima che cade entro una certa distanza dal parametro

Intervalli di confidenza

- Se la distribuzione campionaria dello stimatore è Normale, (anche approssimativamente), allora
  - ✓ con probabilità di circa il 95% lo stimatore produrrà una stima del parametro che ricade a 2 errori standard dal parametro
  - ✓ con probabilità di circa il 99.7% lo stimatore produrrà una stima del parametro che ricade a 3 errori standard dal parametro
  - ✓ minore è l'errore standard, maggiore è la precisione dello stimatore
- Un intervallo di confidenza si può dunque costruire aggiungendo e sottraendo dalla stima puntuale un multiplo dell'errore standard dello stimatore
- Il multiplo dell'errore standard dello stimatore è detto margine di errore e dipende dalla variabilità (errore standard) della distribuzione campionaria dello stimatore
- Forma tipica degli intervalli di confidenza:

Stima puntuale ± Margine di Errore

Intervallo di confidenza per una proporzione: Campioni di dimensione elevata

- Si supponga che la variabile di interesse Y sia binaria
- La distribuzione di Y nella popolazione è Bernoulliana con probabilità di successo π: $Y \sim \text{Bernoulli}(\pi)$ e π è il parametro di interesse
- Stimatore puntuale di π:

  Proporzione di successi campionaria = Media campionaria

$$\hat{\pi} = \overline{Y} = \frac{Y_1 + \cdots + Y_n}{n}$$

- Si ricorda che

$$E(\hat{\pi}) = \mu_{\hat{\pi}} = \pi \qquad \text{e} \qquad V(\hat{\pi}) = \sigma^2_{\hat{\pi}} = \frac{\pi \cdot (1 - \pi)}{n}$$

  Quindi l'errore standard della proporzione campionaria è

$$\sigma_{\hat{\pi}} = \sqrt{\frac{\pi \cdot (1 - \pi)}{n}}$$

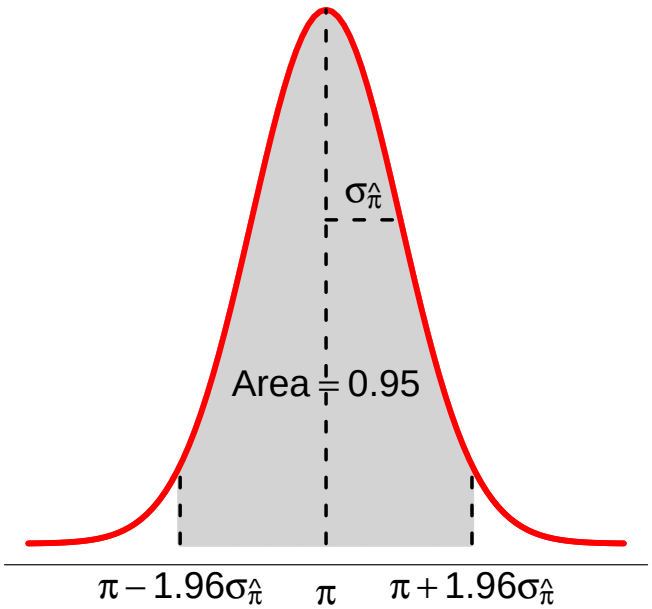
Intervallo di confidenza per una proporzione: Campioni di dimensione elevata

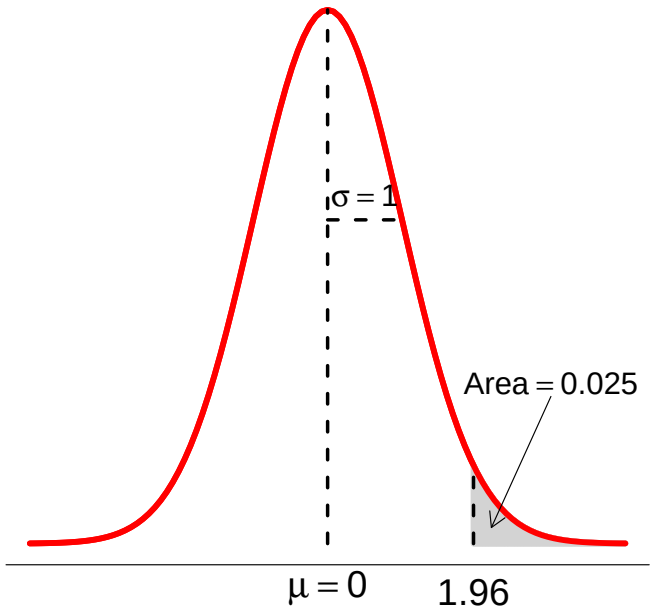
- La proporzione campionaria $\hat{\pi}$ è una media campionaria, quindi per campioni di dimensioni sufficientemente elevate la sua distribuzione campionaria si può approssimare con una distribuzione Normale per il teorema del limite centrale
- Formalmente, per il il teorema del limite centrale, se n è sufficientemente grande

$$\hat{\pi} \approx N\left(\pi, \frac{\pi \cdot (1-\pi)}{n}\right) \quad \text{e quindi} \quad \frac{\hat{\pi} - \pi}{\sqrt{\frac{\pi \cdot (1-\pi)}{n}}} \approx N(0,1)$$

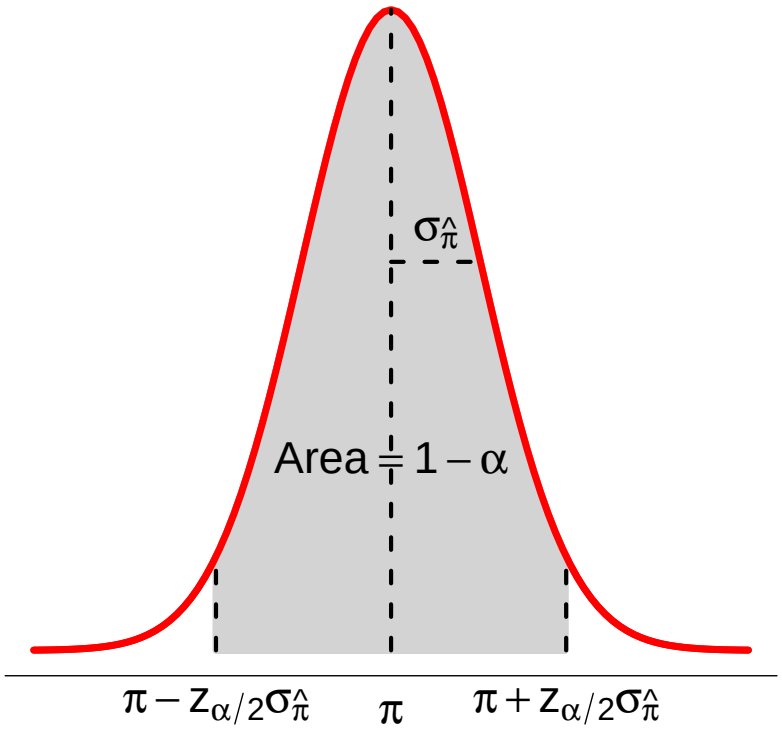
Intervallo di confidenza per una proporzione: Campioni di dimensione elevata

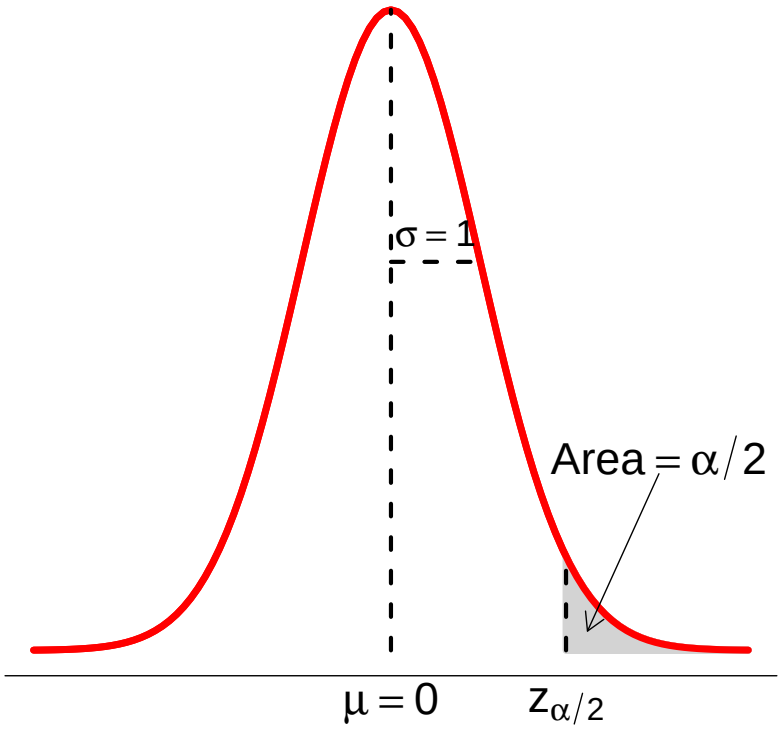
- Nella distribuzione Normale il 95% delle osservazioni è compreso entro 1.96 deviazioni standard dalla media
- Utilizzando l'approssimazione Normale si ha che la proporzione campionaria, come stimatore, assumerà valori che si trovano entro $1.96 \cdot \sigma_{\hat{\pi}}$ unità dal parametro π con probabilità pari a 0.95, dove 1.96 è il valore che nella Normale standard lascia alla sua destra un'area pari a 0.025





- Fissato $1-\alpha$, utilizzando l'approssimazione Normale si ha che la proporzione campionaria, come stimatore, assumerà valori che si trovano entro $z_{\alpha/2} \cdot \sigma_{\hat{\pi}}$ unità dal parametro π con probabilità $1-\alpha$, dove $z_{\alpha/2}$ è il valore che nella Normale standard lascia alla sua destra un area pari a $\alpha/2$





Intervallo di confidenza per una proporzione: Campioni di dimensione elevata

- Una volta osservato il campione, $y_1, \ldots, y_n$, si ha un solo valore dello stimatore

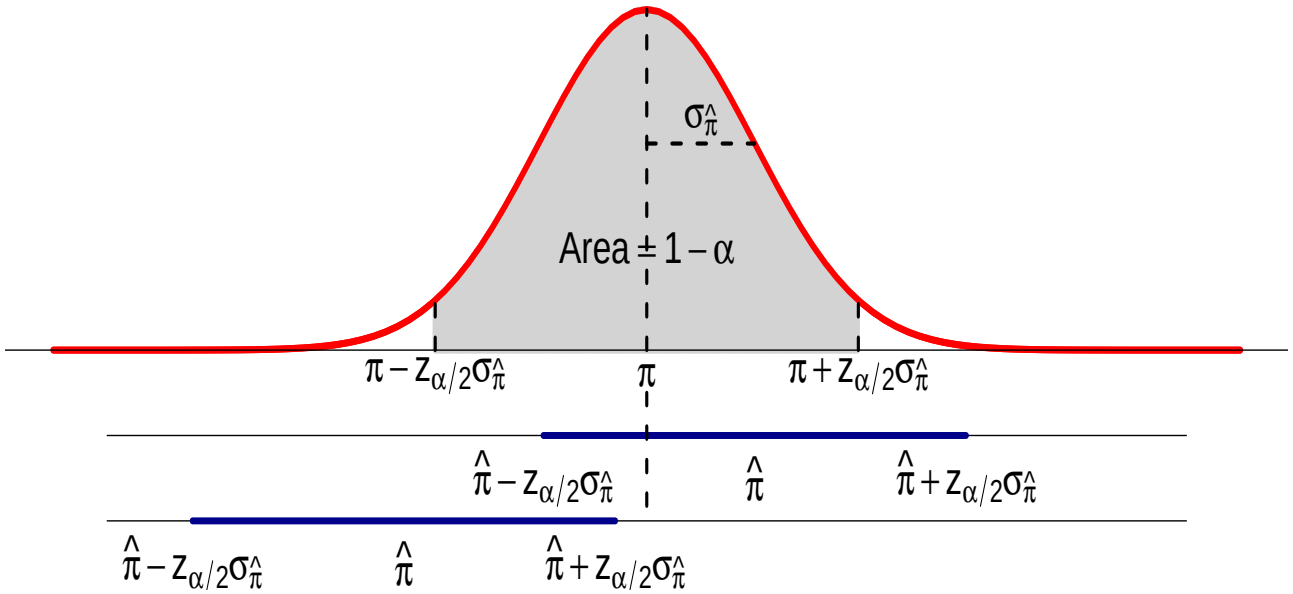
$$\hat{\pi} = \frac{y_1 + \ldots + y_n}{n}$$

e non è noto se tale valore si trova entro $z_{\alpha/2} \cdot \sigma_{\hat{\pi}}$ unità da π

- Se $\hat{\pi}$ si trova entro $z_{\alpha/2} \cdot \sigma_{\hat{\pi}}$ unità da π allora l'intervallo di estremi

$$\hat{\pi} \pm z_{\alpha/2} \cdot \sigma_{\hat{\pi}}$$

contiene π, altrimenti tale intervallo non contiene π



- In pratica, il valore dell'errore standard della proporzione campionaria,

$$\sigma_{\hat{\pi}} = \sqrt{\frac{\pi \cdot (1 - \pi)}{n}}$$

non è noto perché dipende dal parametro π ignoto che interessa stimare

- L'errore standard della proporzione campionaria viene stimato sostituendo a π la proporzione campionaria

$$se_{\hat{\pi}} = \hat{\sigma}_{\hat{\pi}} = \sqrt{\frac{\hat{\pi} \cdot (1 - \hat{\pi})}{n}}$$

- Usando la notazione del libro, il simbolo *se* rappresenta la stima dell'errore standard

Intervallo di confidenza al livello di confidenza $1-\alpha$ per la proporzione

$$IC_{1-\alpha}(\pi) = \left[\hat{\pi} - z_{\alpha/2} \cdot \sqrt{\frac{\hat{\pi} \cdot (1-\hat{\pi})}{n}}; \hat{\pi} + z_{\alpha/2} \cdot \sqrt{\frac{\hat{\pi} \cdot (1-\hat{\pi})}{n}}\right]$$

dove $z_{\alpha/2}$ è il valore che nella Normale standard lascia alla sua destra un'area pari a $\alpha/2$: $P(Z > z_{\alpha/2}) = \alpha/2$

- Con un livello di confidenza pari a $1-\alpha$ si ha una probabilità pari a α che il metodo produca un intervallo di confidenza che *non* contiene il vero valore del parametro
  - ✓ Con un livello di confidenza pari a $1-\alpha = 0.95$ si ha una probabilità pari a $\alpha = 0.05$ che il metodo produca un intervallo di confidenza che *non* contiene il vero valore del parametro

- Ampiezza dell'intervallo di confidenza di livello di confidenza $1-\alpha$

  Estremo superiore - Estremo inferiore $=$

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

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

  ossia:

  $$\text{Ampiezza} = 2 \cdot \text{Margine di Errore} = 2 \cdot z_{\alpha/2} \cdot \sqrt{\frac{\hat{\pi} \cdot (1-\hat{\pi})}{n}}$$

- Il Margine di Errore è uguale al prodotto tra il valore $z_{\alpha/2}$ e l'errore standard

Intervallo di confidenza per una proporzione: Campioni di dimensione elevata

- Maggiore è il livello di confidenza, maggiore sarà la possibilità che l'intervallo di confidenza contenga il vero valore del parametro
- Maggiore è il livello di confidenza, maggiore è l'ampiezza dell'intervallo di confidenza (ossia maggiore è il margine di errore) e quindi minore la precisione (accuratezza) della stima
- Con un livello di confidenza $1 - \alpha = 1$, l'intervallo di confidenza per la proporzione sarebbe $[0, 1]$, che non è di alcun aiuto perché include tutti i possibili valori per π
- La scelta del livello di confidenza é il risultato di un compromesso tra desiderio che l'inferenza sia corretta e precisione della stima: al migliorare di un aspetto l'altro peggiora e viceversa
- Valori tipici del livello di confidenza: $1 - \alpha = 0.90, 0.95, 0.99$

Intervallo di confidenza per una proporzione: Campioni di dimensione elevata

- Fissato $1-\alpha = 0.90$, $z_{0.05} = 1.645$, quindi, si ha

$$IC_{0.90}(\pi) = \left[\hat{\pi} - 1.645 \cdot \sqrt{\frac{\hat{\pi} \cdot (1-\hat{\pi})}{n}}; \hat{\pi} + 1.645 \cdot \sqrt{\frac{\hat{\pi} \cdot (1-\hat{\pi})}{n}}\right]$$

- Fissato $1-\alpha = 0.95$, $z_{0.025} = 1.96$, si ha

$$IC_{0.95}(\pi) = \left[\hat{\pi} - 1.96 \cdot \sqrt{\frac{\hat{\pi} \cdot (1-\hat{\pi})}{n}}; \hat{\pi} + 1.96 \cdot \sqrt{\frac{\hat{\pi} \cdot (1-\hat{\pi})}{n}}\right]$$

- Fissato $1-\alpha = 0.99$, $z_{0.005} = 2.58$, si ha

$$IC_{0.99}(\pi) = \left[\hat{\pi} - 2.58 \cdot \sqrt{\frac{\hat{\pi} \cdot (1-\hat{\pi})}{n}}; \hat{\pi} + 2.58 \cdot \sqrt{\frac{\hat{\pi} \cdot (1-\hat{\pi})}{n}}\right]$$

Esempio: Il problema del fumo negli adolescenti

- Per analizzare il fenomeno del fumo tra gli adolescenti, si estrae un campione casuale semplice di $n = 900$ adolescenti.
- In tale campione il numero di adolescenti che fumano è pari a 180
- Nella popolazione la variabile di interesse $Y \sim \text{Bernoulli}(\pi)$
- Obiettivo: Stimare π, la proporzione di adolescenti che fumano nell'intera popolazione degli adolescenti
- Stima puntuale e stima dell'errore standard:

$$\hat{\pi} = \overline{y} = 180/900 = 0.2$$

$$\text{se}_{\hat{\pi}} = \sqrt{\frac{\hat{\pi} \cdot (1 - \hat{\pi})}{n}} = \sqrt{\frac{0.2 \cdot 0.8}{900}} = \sqrt{0.00018} = 0.0133$$

- Intervallo di confidenza al livello di confidenza del 95%

$$IC_{1-0.05}(\pi) = [0.2 - 1.96 \cdot 0.0133; 0.2 + 1.96 \cdot 0.0133] = [0.174; 0.226]$$

- La percentuale dei adolescenti che fumano è (al livello di confidenza del 95%) non meno di 17.4% e non più del 22.6%

Dimensione del campione e accuretezza della stima

Campioni di dimensione maggiore permettono di ottenere intervalli di confidenza più stretti

- Si ricorda che

$$\text{Ampiezza} = 2 \cdot \text{ Margine di errore} = 2 \cdot z_{\alpha/2} \cdot \sqrt{\frac{\hat{\pi} \cdot (1 - \hat{\pi})}{n}}$$

- Dato $(1 - \alpha)$ al crescere di n si riduce *se* e con esso il margine di errore e l'ampiezza dell'IC
- Esempio: Il problema del fumo negli adolescenti
  - ✓ Con $n = 900$ e $\hat{\pi} = 0.2$, si ha:

$$\text{Ampiezza} = 2 \cdot 1.96 \cdot \sqrt{\frac{0.2 \cdot 0.8}{900}} = 0.052$$

  - ✓ Con $n = 3600 = 4 \cdot 900$ e nell'ipotesi che $\hat{\pi} = 0.2$ (720 successi):

$$\text{Ampiezza} = 2 \cdot 1.96 \cdot \sqrt{\frac{0.2 \cdot 0.8}{3600}} = 0.026$$

  - ✓ L'intervallo di confidenza ottenuto con $n = 3600$ è ampio la metà dell'intervallo di confidenza ottenuto con $n = 900$

Dimensione del campione e accuretezza della stima

- Il margine di errore è inversamente proporzionale alla radice quadrata di n
- L'errore standard è massimo per $\pi = 0.5$
- Massimo valore del margine di errore per $1 - \alpha$ fissato

$$z_{\alpha/2} \cdot \sqrt{\frac{0.5 \cdot 0.5}{n}} = z_{\alpha/2} \cdot \frac{1}{\sqrt{4 \cdot n}} = z_{\alpha/2} \cdot \frac{1}{2 \cdot \sqrt{n}}$$

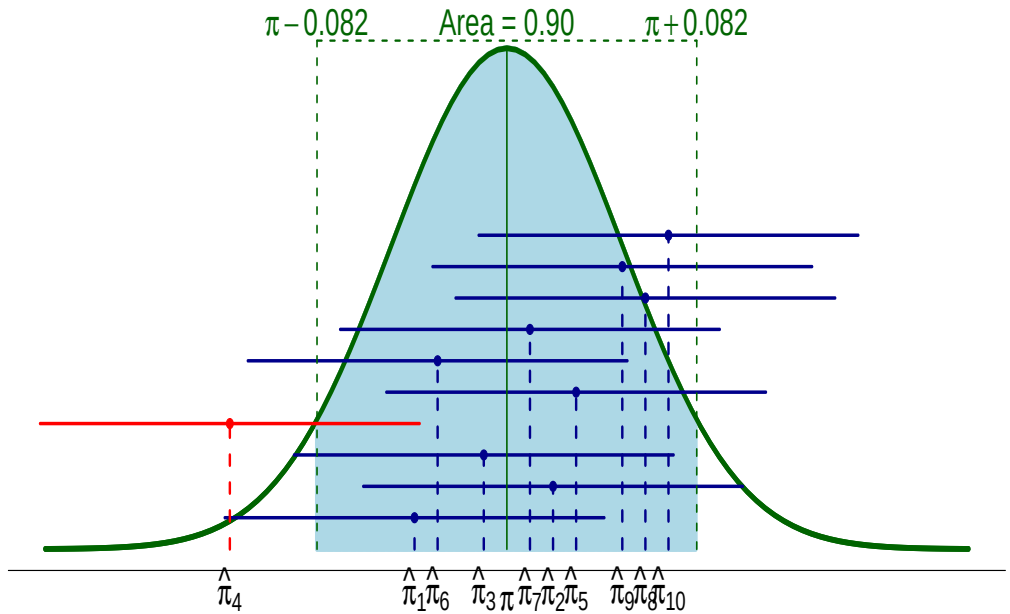
- La dimensione del campione deve essere quadruplicata per ottenere un intervallo di confidenza di ampiezza dimezzata (ossia per ottenere una doppia precisione)
- In sintesi
  - ✓ L'ampiezza dell'intervallo di confidenza per la proporzione cresce al crescere del livello di confidenza
  - ✓ L'ampiezza dell'intervallo di confidenza per la proporzione decresce al crescere della dimensione del campione
  - ✓ Queste proprietà valgono per tutti gli intervalli di confidenza

Interpretazione del livello di confidenza

- Il livello di confidenza ci dice come si comporta il metodo, utilizzato per la costruzione dell'intervallo di confidenza, quando venga applicato ripetutamente a differenti campioni casuali
- Se venissero selezionati più campioni casuali di una certa dimensione e ogni volta venisse costruito un intervallo di confidenza al livello di confidenza $1 - \alpha$ allora circa il $(1 - \alpha)\%$ degli intervalli conterrebbe π
  - ✓ Se venissero selezionati più campioni casuali di una certa dimensione e ogni volta venisse costruito un intervallo di confidenza al livello di confidenza $1 - \alpha = 0.90$ allora circa il $(1 - \alpha)\% = 10\%$ degli intervalli conterrebbe π

Esempio: $Y \sim Ber(\pi = 0.5)$ - Campioni di dimensione $n = 100$
Intervalli di confidenza con livello di confidenza del 90%

Campione	$\hat{\pi}$	ℓ_I	ℓ_S
1	0.46	0.38	0.54
2	0.52	0.44	0.60
3	0.49	0.41	0.57
4	**0.38**	**0.30**	**0.46**
5	0.53	0.45	0.61
6	0.47	0.39	0.55
7	0.51	0.43	0.59
8	0.56	0.48	0.64
9	0.55	0.47	0.63
10	0.57	0.49	0.65



Il 10% degli intervalli di confidenza a livello di confidenza del 90% non includono il vero valore del parametro

Intervallo di confidenza per la proporzione: Osservazioni conclusive

- In pratica viene selezionato un solo campione di dimensione prestabilita e si costruisce un unico intervallo di confidenza utilizzando le osservazioni dell'unico campione selezionato
- Non si può sapere se l'intervallo di confidenza contenga o meno il vero valore del parametro, π
- Il livello di confidenza è una quantità che è relativa alle proprietà del metodo utilizzato per costruire l'intervallo di confidenza

Intervallo di confidenza per la proporzione: Osservazioni conclusive

- Importanza di avere campioni di dimensione elevata
  - ✓ La probabilità che l'intervallo di confidenza contenga il vero valore del parametro, π, è approssimativamente uguale al livello di confidenza: l'approssimazione migliora con campioni di grandi dimensioni
  - ✓ Per il teorema del limite centrale, per n sufficientemente grande, la distribuzione campionaria della proporzione campionaria è approssimativamente Normale
  - ✓ L'approssimazione Normale è in generale adeguata se si hanno almeno 15 osservazioni per categoria
  - ✓ Al crescere della dimensione del campione, l'errore standard stimato della proporzione campionaria tende a assumere valori prossimi al vero errore standard

Intervallo di confidenza per la media

- Supponiamo che il carattere di interesse Y sia quantitativo con media μ nella popolazione
- L'intervallo di di confidenza per la media μ ha la forma

  Stima puntuale $\pm$ Margine di errore

  con il margine di errore multiplo dell'errore standard dello stimatore
- Si distinguono tre casi
  - ✓ Intervallo di confidenza per la media di una *popolazione Normale con varianza nota*
  - ✓ Intervallo di confidenza per la media di una *popolazione Normale con varianza non nota*
  - ✓ Intervallo di confidenza per la media di una *popolazione non Normale per campioni di dimensione elevata*

Intervallo di confidenza per la media di una popolazione Normale con varianza nota

- Distribuzione del carattere nella popolazione: $Y \sim N(\mu, \sigma^2)$ con σ^2 nota
- Campione casuale di dimensione n (qualsiasi): $Y_1, \ldots, Y_n$
- Stimatore della media = Media campionaria

$$\overline{Y} = \frac{Y_1 + \cdots + Y_n}{n}$$

- È noto che se $Y \sim N(\mu, \sigma^2)$ allora $\overline{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ quindi

$$\frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

- Fissato $(1-\alpha)$ con probabilità $(1-\alpha)\%$ lo stimatore media campionaria produce valori della media campionaria entro $z_{\alpha/2}$ errori standard dalla media (non nota) della popolazione

- In tal caso il margine di errore può essere ottenuto moltiplicando l'errore standard della media campionaria, $\sigma_{\overline{Y}} = \sigma/\sqrt{n}$, per un opportuno valore z della distribuzione Normale

$$\text{Margine di errore } = z \cdot \frac{\sigma}{\sqrt{n}}$$

Il valore z dipende dal livello di confidenza

Intervallo di confidenza al livello di confidenza $1-\alpha$ per la media μ di una popolazione Normale con varianza σ^2 nota

$$IC_{1-\alpha}(\mu) = \left[\overline{Y} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}; \overline{Y} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

dove $z_{\alpha/2}$ è il valore che nella Normale standard lascia alla sua destra un'area pari a $\alpha/2$: $P(Z > z_{\alpha/2}) = \alpha/2$

- Ampiezza dell'intervallo di confidenza di livello di confidenza $1-\alpha$

$$\text{Estremo superiore - Estremo inferiore} = \left(\overline{Y} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) - \left(\overline{Y} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 2 \cdot z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

ossia

$$\text{Ampiezza} = 2 \cdot \text{ Margine di errore} = 2 \cdot z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

- La lunghezza dell'intervallo dipende da
  - ✓ la dimensione del campione
  - ✓ il livello di confidenza
  - ✓ la varianza della popolazione
- Intervenendo sulla dimensione del campione o sul livello di confidenza si può aumentare o diminuire la lunghezza dell'intervallo.
- Fissato il livello di confidenza $1-\alpha$ e la dimensione del campione, al variare del campione l'ampiezza dell'intervallo di confidenza rimane costante

Intervallo di confidenza per la media di una popolazione Normale con varianza non nota

- Distribuzione del carattere nella popolazione: $Y \sim N(\mu, \sigma^2)$ con μ e σ^2 non note
- Campione casuale di dimensione n (qualsiasi): $Y_1, \ldots, Y_n$
- Stimatore della media = Media campionaria

$$\overline{Y} = \frac{Y_1 + \cdots + Y_n}{n}$$

- La varianza non è nota, quindi deve essere stimata
- Per stimare la varianza della popolazione si utilizza lo stimatore varianza campionaria corretta:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2 = \frac{\sum_{i=1}^{n} Y_i^2 - n \cdot \overline{Y}^2}{n-1}$$

Intervallo di confidenza per la media di una popolazione Normale con varianza non nota

- Se $Y \sim N(\mu, \sigma^2)$, allora $\overline{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ e quindi $\frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$
- Stimatore della varianza e dell'errore standard della media campionaria

$$\hat{\sigma}^2_{\overline{Y}} = \frac{S^2}{n} \qquad e \qquad se_{\overline{Y}} = \hat{\sigma}_{\overline{Y}} = \frac{S}{\sqrt{n}}$$

  dove $S = \sqrt{S^2}$
- Se si sostituisce la deviazione standard di Y, σ con la deviazione standard campionaria s per ottenere l'errore standard stimato $s/\sqrt{n}$, si introduce un ulteriore fonte di errore
- Sostituendo la varianza incognita con un suo stimatore si ottiene

$$\frac{\overline{Y} - \mu}{\sqrt{S^2/n}} \sim t_{n-1}$$

  dove t_{n-1} è la distribuzione t di Student con $n-1$ gradi di libertà (gdl)

La distribuzione t–Student

William Sealy Gosset

- William Sealy Gosset (Canterbury, 13 giugno 1876 – Beaconsfield, 16 ottobre 1937) è stato uno statistico inglese, meglio noto in statistica come il signor Student della distribuzione t di Student
- Nel 1908 pubblica con lo pseudonimo Student l'articolo nel quale introduce la distribuzione t di Student
- Gosset dovette usare uno pseudonimo poiché la fabbrica Guinness presso la quale lavorava vietava la pubblicazione di articoli per evitare la divulgazione dei segreti di produzione della birra

La distribuzione t–Student

- La distribuzione t di Student dipende da un parametro a valori interi positivi detto *gradi di libertà* (gdl)
- La distribuzione t di Student ha una forma campanulare, simmetrica intorno alla media uguale a zero
- La deviazione standard della distribuzione t di Student è leggermente più grande di 1 (il valore esatto dipende da quelli che vengono chiamati gradi di libertà indicati con gdl)
- La distribuzione t di Student presenta un'ampiezza leggermente diversa per ciascun differente valore dei gdl
- La distribuzione t di Student presenta aree sulle code più grandi (più pesanti) ed è più dispersa rispetto alla distribuzione normale standard
- Quanto più elevato è il valore dei gdl tanto più la distribuzione tenderà a assomigliare a una distribuzione normale standard

La distribuzione t–Student

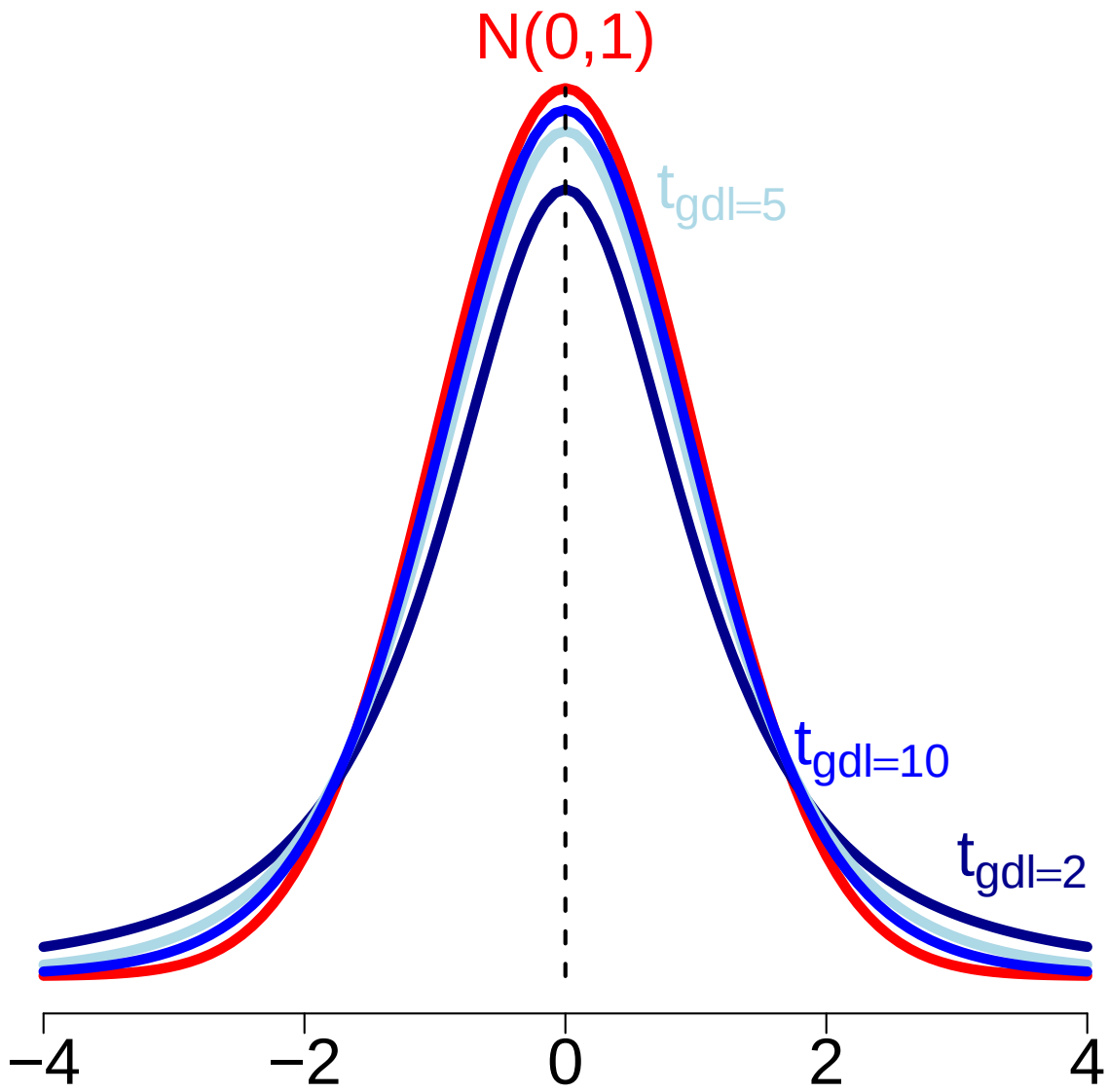
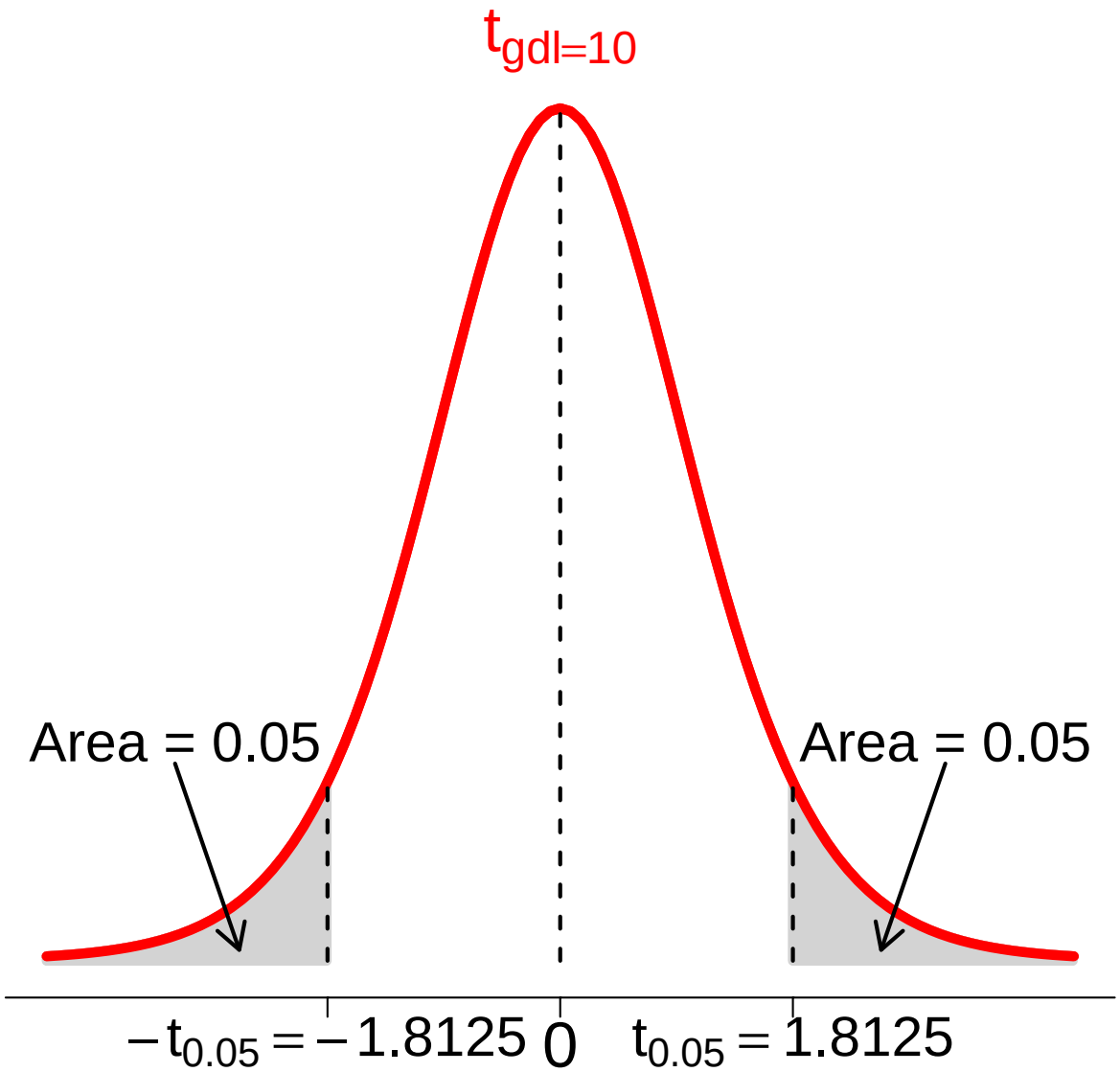


Tavola della distribuzione $t-$ Student

Gradi di Libertà	Area della coda destra della distribuzione *t di Student*						
	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
1	3.0777	6.3138	12.7062	31.8205	63.6567	318.3088	636.6192
2	1.8856	2.9200	4.3027	6.9646	9.9248	22.3271	31.5991
3	1.6377	2.3534	3.1824	4.5407	5.8409	10.2145	12.9240
4	1.5332	2.1318	2.7764	3.7469	4.6041	7.1732	8.6103
5	1.4759	2.0150	2.5706	3.3649	4.0321	5.8934	6.8688
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
10	1.3722	1.8125	2.2281	2.7638	3.1693	4.1437	4.5869
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
30	1.3104	1.6973	2.0423	2.4573	2.7500	3.3852	3.6460
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
50	1.2987	1.6759	2.0086	2.4033	2.6778	3.2614	3.4960
60	1.2958	1.6706	2.0003	2.3901	2.6603	3.2317	3.4602
70	1.2938	1.6669	1.9944	2.3808	2.6479	3.2108	3.4350
80	1.2922	1.6641	1.9901	2.3739	2.6387	3.1953	3.4163
90	1.2910	1.6620	1.9867	2.3685	2.6316	3.1833	3.4019
100	1.2901	1.6602	1.9840	2.3642	2.6259	3.1737	3.3905
Infinito	1.2816	1.6449	1.9600	2.3263	2.5758	3.0902	3.2905

Leggere la tavola della distribuzione t–Student



Intervallo di confidenza per la media di una popolazione Normale con varianza non nota

Dato un campione casuale di dimensione n estratto da una popolazione Normale con media e varianza entrambe ignote, l'intervallo di confidenza per la media a livello di confidenza $1-\alpha$ è dato da:

$$\left[\overline{Y} - t_{(n-1),\alpha/2} \cdot \frac{S}{\sqrt{n}}; \overline{Y} + t_{(n-1),\alpha/2} \cdot \frac{S}{\sqrt{n}}\right]$$

dove $t_{(n-1),\alpha/2}$ è il valore che nella distribuzione t–Student con $n-1$ gdl lascia alla sua destra un'area pari a $\alpha/2$:
$P(T_{n-1} > t_{(n-1),\alpha/2}) = \alpha/2$

Intervallo di confidenza per la media di una popolazione Normale con varianza non nota

- Margine di errore = $t\cdot$ Errore standard = $t \cdot \dfrac{s}{\sqrt{n}}$
- Ampiezza dell'intervallo di confidenza al livello di confidenza $1-\alpha$

  Estremo superiore - Estremo inferiore =

  $$\left(\overline{Y} + t_{(n-1),\alpha/2} \cdot \frac{S}{\sqrt{n}}\right) - \left(\overline{Y} - t_{(n-1),\alpha/2} \cdot \frac{S}{\sqrt{n}}\right) = 2 \cdot t_{(n-1),\alpha/2} \cdot \frac{S}{\sqrt{n}}$$

  ossia

  $$\text{Ampiezza} = 2 \cdot \text{Margine di errore} = 2 \cdot t_{(n-1),\alpha/2} \cdot \frac{S}{\sqrt{n}}$$

- Fissata la dimensione del campione e fissato il livello di confidenza, al variare dei campioni estratti, la lunghezza degli intervalli corrispondenti non rimane costante poiché varia il valore di S

Esempio – La perdita di calcio

- Il latte materno contiene una certa quantità di calcio, una parte del quale deriva direttamente dal calcio contenuto nella struttura ossea
- Alcune donne, quindi, durante l'allattamento possono andare incontro a demineralizzazione ossea
- I ricercatori hanno misurato la variazione percentuale di calcio nelle vertebre di 16 mamme nel corso di tre mesi d'allattamento
- Osservazioni

Mamma$_i$	y_i	Mamma$_i$	y_i
1	−4.7	9	0.4
2	−1.0	10	−0.8
3	−6.5	11	−6.5
4	−4.0	12	0.2
5	0.3	13	−5.2
6	−3.1	14	−2.0
7	−3.0	15	−0.3
8	−4.7	16	1.7

Esempio – La perdita di calcio

- Obiettivo: Costruire un intervallo di confidenza al livello di confidenza $1 - \alpha = 0.95$ per la media della variazione percentuale di calcio nelle vertebre nella popolazione delle mamme
- Condizioni
  - ✓ Le osservazioni sono realizzazioni di un campione casuale di dimensione $n = 16$, $Y_1, \ldots, Y_{16}$, estratto dalla popolazione delle mamme
  - ✓ Nella popolazione delle mamme la variazione percentuale di calcio nelle vertebre, Y, ha distribuzione Normale: $Y \sim N(\mu, \sigma^2)$

Esempio – La perdita di calcio

Mamma$_i$	y_i	$(y_i - \bar{y})^2$	y_i^2
1	−4.7	5.0625	22.09
2	−1.0	2.1025	1.00
3	−6.5	16.4025	42.25
4	−4.0	2.4025	16.00
5	0.3	7.5625	0.09
6	−3.1	0.4225	9.61
7	−3.0	0.3025	9.00
8	−4.7	5.0625	22.09
9	0.4	8.1225	0.16
10	−0.8	2.7225	0.64
11	−6.5	16.4025	42.25
12	0.2	7.0225	0.04
13	−5.2	7.5625	27.04
14	−2.0	0.2025	4.00
15	−0.3	4.6225	0.09
16	1.7	17.2225	2.89
Totale	−39.2	103.2000	199.24

- Stima della media di Y

$$\bar{y} = \frac{-39.2}{16} = -2.45$$

- Stima della varianza di Y

$$s^2 = \frac{103.2000}{16-1} = \frac{199.24 - 16 \cdot (-2.45)^2}{16-1} =$$

$$=6.88$$

- Stima della deviazione standard di Y

$$s = \sqrt{6.88} = 2.623$$

Esempio – La perdita di calcio

- Stima dell'errore standard della media campionaria

$$se_{\overline{Y}} = \frac{s}{\sqrt{n}} = \frac{2.623}{\sqrt{16}} = 0.656$$

- Valore $t_{n-1,\alpha/2}$ per $1 - \alpha = 0.95$

$$1 - \alpha = 0.95 \Longrightarrow \alpha = 0.05 \Longrightarrow \alpha/2 = 0.025$$

  Quindi $t_{(n-1),\alpha/2} = t_{15,0.025} = 2.131$

- Intervallo di confidenza per la media μ della variazione percentuale di calcio nelle vertebre delle mamme al livello di confidenza $1 - \alpha = 0.95$

$$IC_{0.95}(\mu) = -2.45 \pm 2.131 \cdot \frac{2.623}{\sqrt{16}} = [-3.85; -1.05]$$

- Ampiezza dell'intervallo

$$\left(-2.45 + 2.131 \cdot \frac{2.623}{\sqrt{16}}\right) - \left(-2.45 - 2.131 \cdot \frac{2.623}{\sqrt{16}}\right) = 2 \cdot 2.131 \cdot \frac{2.623}{\sqrt{16}} = 2.80$$

Esempio – La perdita di calcio

- Si supponga di voler un intervallo di confidenza al livello di confidenza $1 - \alpha = 0.99$
- Valore $t_{n-1,\alpha/2}$ per $1 - \alpha = 0.99$: $t_{(n-1),\alpha/2} = t_{15,0.005} = 2.947$
- Intervallo di confidenza per la media μ della variazione percentuale di calcio nelle vertebre delle mamme al livello di confidenza $1 - \alpha = 0.99$

$$IC_{0.99}(\mu) = -2.45 \pm 2.947 \cdot \frac{2.623}{\sqrt{16}} = [-4.38; -0.52]$$

- Ampiezza dell'intervallo: $2 \cdot 2.947 \cdot \frac{2.623}{\sqrt{16}} = 3.86$
- L'intervallo di confidenza al livello di confidenza $1 - \alpha = 0.99$, $IC_{0.99}(\mu) = [-4.38; -0.52]$, è più ampio di quello al livello di confidenza $1 - \alpha = 0.95$, $IC_{0.95}(\mu) = [-3.85; -1.05]$
- Questo è il costo da pagare per un intervallo con livello di confidenza maggiore

La distribuzione t–Student e la distribuzione Normale

- Al crescere dei gradi di libertà la distribuzione t–Student tende (assomiglia sempre di più) alla distribuzione Normale standard: Se i *gdl* sono sufficientemente grandi, $t_{gdl} \approx N(0,1)$
- Al crescere della dimensione del campione, la distribuzione t–Student diventa sempre meno dispersa e assomiglia sempre di più alla distribuzione Normale
- Se i gradi di libertà sono sufficientemente grandi, $t_{gdl,\alpha/2} \approx z_{\alpha/2}$: Si confrontino i valori sulle tavole

La distribuzione t–Student e la distribuzione Normale

	Area della coda destra della distribuzione *t di Student*						
gdl	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
1	3.0777	6.3138	12.7062	31.8205	63.6567	318.3088	636.6192
2	1.8856	2.9200	4.3027	6.9646	9.9248	22.3271	31.5991
3	1.6377	2.3534	3.1824	4.5407	5.8409	10.2145	12.9240
4	1.5332	2.1318	2.7764	3.7469	4.6041	7.1732	8.6103
5	1.4759	2.0150	2.5706	3.3649	4.0321	5.8934	6.8688
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
30	1.3104	1.6973	2.0423	2.4573	2.7500	3.3852	3.6460
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
90	1.2910	1.6620	1.9867	2.3685	2.6316	3.1833	3.4019
100	1.2901	1.6602	1.9840	2.3642	2.6259	3.1737	3.3905
Infinito	1.2816	1.6449	1.9600	2.3263	2.5758	3.0902	3.2905

Valore	Area della coda destra della distribuzione *Normale*						
z	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
z	1.2816	1.6449	1.9600	2.3263	2.5758	3.0902	3.2905

- Per $gdl > 100$ le differenze tra distribuzione Normale e distribuzione t–Student con gdl gradi di libertà sono molto piccole, quindi non si riportano nella tavola i valori della t

La distribuzione t–Student e la distribuzione Normale

- Si ricordi che se $Y \sim N(\mu, \sigma^2)$ allora $\overline{Y} \sim N(\mu, \sigma^2/n)$
- La distribuzione t–Student viene introdotta per tener conto dell'incertezza sulla varianza σ^2, che, se non nota, deve essere stimata con S^2
- La t–Student è più dispersa della distribuzione Normale standard
- Se la dimensione del campione n è sufficientemente grande,

$$IC_{1-\alpha}(\mu) = \overline{y} \pm t_{n-1,\alpha/2} \cdot \frac{s}{\sqrt{n}} \approx \overline{y} \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

- Lo stimatore S^2 è uno stimatore consistente. Al crescere della dimensione del campione S^2 si avvicina sempre più al vero valore della varianza σ^2 e, di conseguenza, l'errore standard stimato della media campionaria, $\frac{s}{\sqrt{n}}$, approssima sempre meglio il vero errore standard, $\frac{\sigma}{\sqrt{n}}$

Esempio – Qualità delle acque

- Un'associazione ambientalista raccoglie un litro d'acqua in 101 punti scelti a caso lungo un fiume e misura, in ogni punto, la quantità di ossigeno presente
- La media campionaria è $\bar{y} = 4.62$ milligrammi (mg) e la deviazione standard (campionaria) è $s = 0.92$ mg
- Si ipotizza che la quantità di ossigeno presente nel fiume abbia una distribuzione Normale di media μ e varianza σ^2, non note
- Intervallo di confidenza (esatto) al livello di confidenza del 95%: Il valore $t_{101-1,0.025} = 1.984$, quindi

$$IC_{0.95}(\mu) = \left[4.62 - 1.984 \cdot \frac{0.92}{\sqrt{101}}; 4.62 + 1.984 \cdot \frac{0.92}{\sqrt{101}}\right] = [4.438; 4.802]$$

- Intervallo di confidenza (approssimato) al livello di confidenza del 95%: Il valore $z_{0.025} = 1.96$, quindi

$$IC_{0.95}(\mu) = \left[4.62 - 1.96 \cdot \frac{0.92}{\sqrt{101}}; 4.62 + 1.96 \cdot \frac{0.92}{\sqrt{101}}\right] = [4.441; 4.799]$$

Intervallo di confidenza per la media di una popolazione non Normale

- Sia Y una variabile continua che rappresenta il fenomeno di interesse nella popolazione
- La distribuzione di Y nella popolazione non è nota (non Normale)
- Obiettivo: Trovare un intervallo di confidenza per la media di Y nella popolazione
  - ✓ Campioni di dimensione elevata
  - ✓ Campioni di dimensione piccola

Intervallo di confidenza per la media di una popolazione non Normale: Campioni di dimensione elevata

- Si supponga di essere interessati a un carattere Y (continuo) con distribuzione nella popolazione non nota
- Obiettivo: Costruire un intervallo di confidenza per la media, μ, di Y nella popolazione utilizzando un campione casuale $Y_1, \ldots, Y_n$ di dimensione n
- Si ricorda che per il teorema del limite centrale, se la dimensione del campione, n è sufficiente grande, la distribuzione campionaria della media campionaria può essere approssimata dalla distribuzione Normale:

$$\overline{Y} \approx N\left(\mu, \frac{\sigma^2}{n}\right) \text{ per } n \text{ sufficientemente grande}$$

Intervallo di confidenza per la media di una popolazione non Normale: Campioni di dimensione elevata

Quindi, per n sufficientemente grande l'intervallo di confidenza per la media μ di Y al livello di confidenza $1-\alpha$ può essere approssimato come segue

- Se la varianza di Y nella popolazione, σ^2, è nota:

$$IC_{1-\alpha}(\mu) = \left[\overline{Y} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}; \overline{Y} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

- Se la varianza di Y nella popolazione, σ^2 è non nota

$$IC_{1-\alpha}(\mu) = \left[\overline{Y} - z_{\alpha/2} \cdot \frac{S}{\sqrt{n}}; \overline{Y} + z_{\alpha/2} \cdot \frac{S}{\sqrt{n}}\right]$$

dove $S = \sqrt{S^2}$

Esempio – Ricerche di mercato

- Un'azienda che produce piccoli elettrodomestici, incarica una società di sondaggi di stimare le vendite al dettaglio dei suoi prodotti ricavando informazioni da un campione di singoli negozi
- Nell'ultimo mese, in un campione casuale di $n = 175$ negozi, il numero medio di frullatori venduti in ogni negozio è stato pari a $\bar{y} = 24$ con una deviazione standard (campionaria) pari $s = 11$
- Obiettivo: Costruire un intervallo di confidenza al 95% per il numero medio di frullatori venduti in tutti i negozi

Esempio – Ricerche di mercato

- La dimensione del campione è sufficientemente grande da ipotizzare che $\overline{Y} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$
- La varianza del numero di frullatori venduti in tutti i negozi non è nota, quindi la si stima con la varianza campionaria corretta: $s^2 = 11^2$
- L'intervallo di confidenza al 95% per il numero medio di frullatori venduti in tutti i negozi può essere approssimato come segue

$$
\begin{aligned}
IC_{0.95}(\mu) &= \left[\overline{y} - z_{0.025} \cdot \frac{s}{\sqrt{n}}; \overline{y} + z_{0.025} \cdot \frac{s}{\sqrt{n}}\right] = \\
&= \left[24 - 1.96 \cdot \frac{11}{\sqrt{175}}; 24 + 1.96 \cdot \frac{11}{\sqrt{175}}\right] \\
&= [22.37; 25.63]
\end{aligned}
$$

Scelta della dimensione del campione

- La dimensione del campione è un elemento importante che incide sulla precisione dei risultati inferenziali
- Prima di iniziare la raccolta dei dati, in molti studi si cerca di determinare la dimensione campionaria che permetterà di ottenere un determinato grado di precisione
- Un criterio per determinare la dimensione n del campione si basa sull'ampiezza dell'intervallo di confidenza per il parametro di interesse
- Formalmente, si cerca il valore di n per il quale un intervallo di confidenza per il parametro di interesse ha un margine di errore corrispondente a un certo valore
- Gli elementi chiave che influiscono sulla determinazione dell'ampiezza campionaria sono:
  - ✓ Il *margine di errore*, il quale dipende direttamente dall'errore standard della distribuzione campionaria dello stimatore puntuale
  - ✓ L'*errore standard* dello stimatore, il quale dipende dalla

Scelta della dimensione del campione per stimare una proporzione

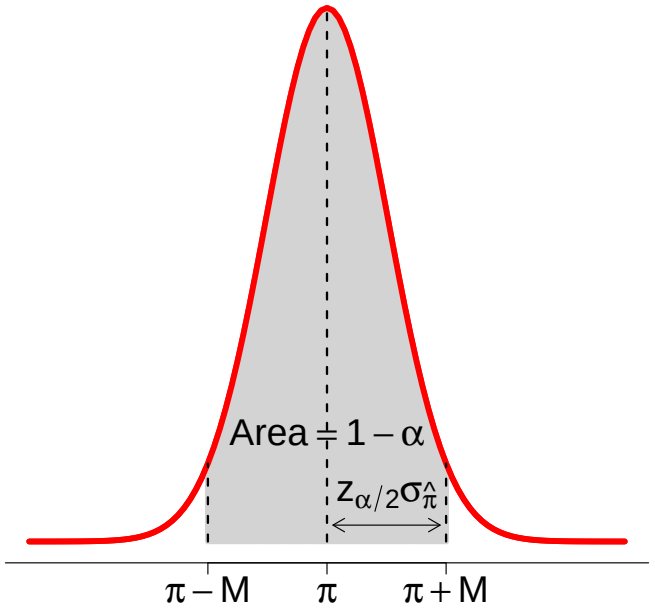
- Carattere di interesse: $Y \sim \text{Bernoulli}(\pi)$
- Obiettivo: Stimare la probabilità di successo, π
- Al fine di determinare la dimensione del campione che garantisca una precisione della stima desiderata si deve
  1. specificare il *margine di errore*
  2. la probabilità con la quale si ottiene quel margine di errore
- Esempio – Sondaggio di opinione in un certo paese: Favorevoli versus contrari all'eutanasia
  - ✓ Obiettivo: determinare n tale che la proporzione dei favorevoli nella popolazione si trovi entro 0.04 punti dal vero valori con probabilità 0.95
  - ✓ Si tratta di determinare la numerosità campionaria necessaria per garantire un margine di errore del 4% e un livello di confidenza del 95%
  - ✓ In altri termini, dobbiamo determinare n in modo tale che l'intervallo di confidenza al livello di confidenza del 95% sia pari a $\hat{\pi} \pm 0.04$

Scelta della dimensione del campione per stimare una proporzione

- Si assuma che la distribuzione campionaria della proporzione campionaria sia ben approssimata da una distribuzione Normale
- Allora la proporzione campionaria assumerà un valore che si trova entro $z_{\alpha/2}$ errori standard da π con probabilità $1 - \alpha$
- Margine di errore = Semi-lunghezza dell'intervallo di confidenza:

$$\text{Margine di errore } = M = z_{\alpha/2} \cdot \sigma_{\hat{\pi}} = z_{\alpha/2} \cdot \sqrt{\frac{\pi \cdot (1 - \pi)}{n}}$$

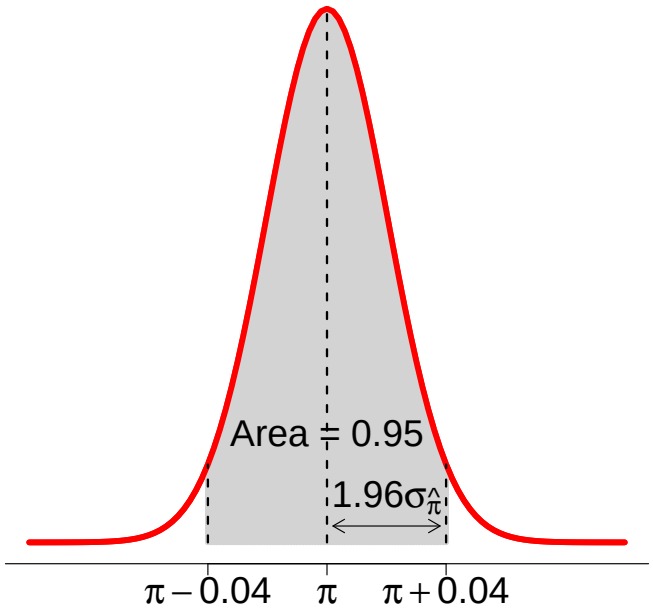
quindi la proporzione campionaria assumerà un valore che si trova entro il margine di errore (ossia entro $z_{\alpha/2} \cdot \sigma_{\hat{\pi}}$) da π con probabilità $1 - \alpha$



Scelta della dimensione del campione per stimare una proporzione

Esempio – Sondaggio di opinione: Favorevoli versus contrari all'eutanasia

- Si ipotizzi che la distribuzione campionaria della proporzione campionaria sia ben approssimata da una distribuzione Normale
- Fissato il livello di confidenza $1 - \alpha = 0.95$ dell'intervallo di confidenza con probabilità $1 - \alpha = 0.95$ la proporzione campionaria assumerà un valore che si trova entro $z_{0.025} = 1.96$ errori standard da π
- Si fissi il margine di errore $M = 0.04$. Allora $M = 0.04 = 1.96 \cdot \sigma_{\hat{\pi}}$ e la proporzione campionaria assumerà un valore che si trova entro 0.04 unità da π



Scelta della dimensione del campione per stimare una proporzione

- Obiettivo: Fissato il margine di errore M e il livello di confidenza dell'intervallo di confidenza $1-\alpha$ trovare n che da un valore dell'errore standard della proporzione campionaria $\sigma_{\hat{\pi}}$ tale che

$$M = z_{\alpha/2} \cdot \sigma_{\hat{\pi}} = z_{\alpha/2} \cdot \sqrt{\frac{\pi(1-\pi)}{n}}$$

- Risolvendo l'equazione si ha

$$\sqrt{n} = z_{\alpha/2} \cdot \frac{\sqrt{\pi(1-\pi)}}{M}$$

ossia

$$n = \left(z_{\alpha/2} \cdot \frac{\sqrt{\pi(1-\pi)}}{M} \right)^2 = z^2_{\alpha/2} \cdot \frac{\pi(1-\pi)}{M^2}$$

- Esempio– Sondaggio di opinione in un certo paese: Favorevoli versus contrari all'eutanasia
  - ✓ Fissato il margine di errore $M = 0.04$ e il livello di confidenza dell'intervallo di confidenza $1-\alpha = 0.95$, n deve essere almeno uguale a

$$n = \left(1.96 \cdot \frac{\sqrt{\pi(1-\pi)}}{0.04} \right)^2 = (1.96)^2 \cdot \frac{\pi(1-\pi)}{0.04^2}$$

Scelta della dimensione del campione per stimare una proporzione

- Problema: La formula che permette di determinare la dimensione del campione n per la stima di una proporzione π dipende dal parametro che interessa stimare
- È necessario ipotizzare un valore per π
- Se non si hanno informazioni sul possibile valore di π si può prendere un approccio "prudenziale"
- Il massimo valore dell'errore standard della proporzione campionaria si ha per $\pi = 0.5$
- Si determina la dimensione campionaria necessaria per ottenere a un livello di confidenza fissato, $1 - \alpha$ il margine di errore desiderato ponendo $\pi = 0.5$:

$$n = z^2_{\alpha/2} \cdot \frac{0.5(1 - 0.5)}{M^2}$$

- Questo approccio garantisce che al livello di confidenza $1 - \alpha$ il margine di errore non sarà superiore al valore M fissato, qualunque sia il vero valore di π

- Margine di errore $M = 0.04$
- Livello di confidenza dell'intervallo di confidenza $1 - \alpha = 0.95$
- Utilizzando un approccio prudenziale n deve essere almeno uguale a

$$n = (1.96)^2 \cdot \frac{0.5(1 - 0.5)}{0.04^2} = 600.23$$

- Quindi sono necessarie $n = 601$ osservazione per avere un margine di errore del 4% al livello di confidenza $1 - \alpha = 0.95$

Scelta della dimensione del campione per stimare una proporzione

In sintesi

- Fissato il margine di errore M e fissato il livello di confidenza dell'intervallo di confidenza $1-\alpha$ la dimensione del campione necessaria per stimare la proporzione π con la desiderata precisione è

$$n = \left(z_{\alpha/2} \cdot \frac{\sqrt{\pi(1-\pi)}}{M} \right)^2 = z^2_{\alpha/2} \cdot \frac{\pi(1-\pi)}{M^2}$$

  dove $z_{\alpha/2}$ è il valore che nella Normale standard lascia alla sua destra un'area pari al $\alpha/2$

- Tale formula dipende dal parametro π non noto
- È necessario ipotizzare un valore per π
- Se non si hanno informazioni su π si può adottare un approccio prudenziale determinando n nell'ipotesi che Y (e quindi la proporzione campionaria) abbia varianza massima, ossia ponendo $\pi = 0.5$

$$n = z^2_{\alpha/2} \cdot \frac{0.5(1-0.5)}{M^2}$$

Scelta della dimensione del campione per stimare una media

- Si consideri un carattere Y con distribuzione nella popolazione Normale: $Y \sim N(\mu, \sigma^2)$ con σ^2 nota
- Obiettivo: Determinare la dimensione campionaria necessaria per ottenere un intervallo di confidenza per μ di livello di confidenza $1 - \alpha$ per cui il margine di errore non sia superiore a un certo valore fissato M
- Margine di errore = Semi-lunghezza dell'intervallo di confidenza

$$\text{Margine di errore } = M = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

- Quindi

$$\sqrt{n} = z_{\alpha/2} \cdot \frac{\sigma}{M}$$

da cui segue che

$$n = z^2_{\alpha/2} \cdot \frac{\sigma^2}{M^2}$$

- Maggiore è la deviazione standard di Y nella popolazione, maggiore è la dimensione campionaria necessaria per ottenere un prefissato margine di errore per un intervallo di confidenza per la media di Y a un certo livello di confidenza, $1 - \alpha$

Scelta della dimensione del campione per stimare una media

- Si supponga di essere interessati a stimare i minuti che in media studenti del primo anno di università dedicano allo studio la sera
- Si supponga che il tempo di studio nella popolazione degli studenti del primo anno segua una distribuzione Normale con deviazione standard $\sigma = 25$ minuti
- Quanti studenti devono essere intervistati per ottenere un margine di errore pari a $M = 5$ per un intervallo di confidenza per il numero medio di minuti dedicati allo studio con un livello di confidenza 95%?
- Per $1 - \alpha = 0.95$, $z_{\alpha/2} = 1.96$ quindi

$$n = z^2_{\alpha/2} \cdot \frac{\sigma^2}{M^2} = (1.96^2) \cdot \frac{25^2}{5^2} = 96.04$$

Sono necessarie $n = 97$ osservazioni per ottenere un intervallo di confidenza al livello di confidenza 95% non più ampio di 10 minuti

Scelta della dimensione del campione per stimare una media

- Si consideri un carattere Y con distribuzione nella popolazione Normale: $Y \sim N(\mu, \sigma^2)$ con σ^2 non nota
- Obiettivo: Determinare la dimensione campionaria necessaria per ottenere un intervallo di confidenza per μ di livello di confidenza $1 - \alpha$ per cui il margine di errore non sia superiore a un certo valore fissato M
- Margine di errore = Semi-lunghezza dell'intervallo di confidenza

$$\text{Margine di errore } = M = t_{n-1,\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

Quindi, dovrebbe essere $n = t^2_{n-1,\alpha/2} \cdot \frac{s}{M^2}$

- Problema: La deviazione standard campionaria, s, e i gradi di libertà della distribuzione t di Student dipendono da n
- Per risolvere il problema si applica la formula precedente

$$n = z^2_{\alpha/2} \cdot \frac{\sigma^2}{M^2}$$

utilizzando un valore 'prudenziale' per σ^2 e approssimando la distribuzione t di Student con una Normale

Scelta della dimensione del campione per stimare una media

- Si consideri un carattere Y con media μ e σ^2 nella popolazione
- Obiettivo: Determinare la dimensione campionaria necessaria per ottenere un intervallo di confidenza per μ di livello di confidenza $1 - \alpha$ per cui il margine di errore non sia superiore a un certo valore fissato M
- Utilizzando l'approssimazione Normale per la distribuzione campionaria della media campionaria e considerando un valore 'prudenziale' per σ^2 (nel caso in cui σ^2 sia non nota) si ha

$$n = z^2_{\alpha/2} \cdot \frac{\sigma^2}{M^2}$$

Scelta della dimensione del campione per stimare una media: Esempio 1

- Si supponga di voler pianificare uno studio sul livello di istruzione tra gli adulti in Italia
- Tra le variabili rilevate si ha il numero di anni di istruzione
- Obiettivo: Determinare la dimensione campionaria necessaria per ottenere un intervallo di confidenza per il numero di anni di istruzione nella popolazione degli adulti in Italia al livello di confidenza $1 - \alpha = 0.99$ per cui il margine di errore non sia superiore a $M = 1$ anno
- Nessuna informazione sul valore delle deviazione standard è disponibile
- Un valore per la varianza si può ipotizzare pensando che i valori del numero di anni di istruzione si trovano quasi tutti in un intervallo di estremi 0 e 21
- Se la distribuzione del numero di anni di istruzione fosse approssimativamente Normale circa il 99.7% delle osservazioni dovrebbe trovarsi entro tre deviazioni standard dalla media, ossia tra $\mu - 3 \cdot \sigma$ e $\mu + 3 \cdot \sigma$, quindi l'intervallo di lunghezza 21 anni dovrebbe essere uguale a $6 \cdot \sigma$. Quindi $\sigma = 21/6 = 3.5$
- Per $1 - \alpha = 0.99$, $z_{\alpha/2} = 2.58$ quindi

$$n = z_{\alpha/2}^2 \cdot \frac{\sigma^2}{M^2} = (2.58^2) \cdot \frac{3.5^2}{1^2} = 81.54$$

Scelta della dimensione del campione per stimare una media: Esempio 2

- Si supponga di voler pianificare uno studio sul livello di benessere delle famiglie italiane
- Tra le variabili rilevate si ha il reddito annuale della famiglia
- Obiettivo: Determinare la dimensione campionaria necessaria per ottenere un intervallo di confidenza per il reddito annuale medio nella popolazione delle famiglie italiane al livello di confidenza $1 - \alpha = 0.95$ per cui il margine di errore non sia superiore a $M = 1\,000$ euro
- Si supponga che precedenti studi suggeriscano che una deviazione standard di 12 000 euro sia un valore ragionevole per il reddito annuale delle famiglie italiane
- Per $1 - \alpha = 0.95$, $z_{\alpha/2} = 1.96$ quindi

$$n = z_{\alpha/2}^2 \cdot \frac{\sigma^2}{M^2} = (1.96^2) \cdot \frac{12\,000^2}{1\,000^2} = 553.19$$

Sono necessarie $n = 554$ famiglie per ottenere un intervallo di confidenza al livello di confidenza 95% non più ampio di 2 000 euro nell'ipotesi che 12 000 euro sia un valore adeguato per la deviazione standard del reddito annuale delle famiglie italiane

Riepilogo

Parametro	Stima puntuale	Errore Standard	Intervallo di confidenza
Media μ (Varianza σ^2 nota)	$\bar{y}$	$\frac{\sigma}{\sqrt{n}}$	$\bar{y} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
Media μ (Varianza σ^2non nota)	$\bar{y}$	$\frac{s}{\sqrt{n}} = \frac{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2}}{\sqrt{n}}$	$\bar{y} \pm t_{(n-1),\alpha/2}\frac{s}{\sqrt{n}}$
Proporzione π	$\hat{\pi} = \bar{y}$	$\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$	$\hat{\pi} \pm z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$