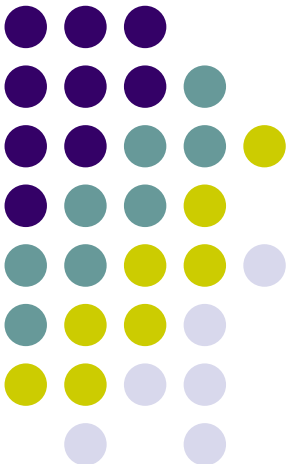
Statistica

Intervalli di confidenza



Insegnamento: Statistica

Corso di Laurea Triennale in Economia

Dipartimento di Economia e Management, Università di Ferrara

Docenti: Prof. S. Bonnini, Dott.ssa V. Mini, Dott.ssa M.Brunori

Argomenti

- Intervallo di confidenza per la media (σ noto)
- Intervallo di confidenza per la media (σ non noto)
- La distribuzione *t* di Student
- Intervallo di confidenza per la proporzione
- Determinazione dell'ampiezza campionaria

L'intervallo di confidenza

- L'**inferenza statistica** è il processo attraverso il quale i risultati campionari vengono utilizzati per trarre conclusioni sulle caratteristiche di una popolazione
- Tale processo consente di **stimare** caratteristiche non note della popolazione come i parametri (ad es. la media per le var. numeriche o la proporzione per le var. categoriali) che caratterizzano la distribuzione della variabile di interesse
- Ci sono due approcci fondamentali di stima: le stime puntuali e le stime per intervalli
- Uno **stimatore puntuale** è una statistica (cioè una funzione dei dati campionari) che viene definita allo scopo di fornire una sintesi su un parametro di interesse

L'intervallo di confidenza

- La **stima puntuale** è lo specifico valore assunto da una statistica, calcolata in corrispondenza dei dati campionari e che viene utilizzata per stimare il vero valore non noto di un parametro di una popolazione
- Uno **stimatore per intervallo** è un intervallo costruito attorno allo stimatore puntuale, in modo tale che sia nota e fissata la probabilità che il parametro appartenga all'intervallo stesso
- Tale probabilità è detta **livello di confidenza** ed è in generale indicato con $(1-\alpha)\%$ dove α è la probabilità che il parametro si trovi al di fuori dell'intervallo di confidenza
- Quindi la confidenza è il grado di fiducia che l'intervallo possa contenere effettivamente il parametro di interesse

Intervallo di confidenza per la media (σ noto)

Esempio: si consideri un processo industriale di riempimento di scatole di cereali e sia assuma che il peso X delle scatole sia $X \sim N(\mu; 15^2)$. Dato un campione casuale di $n=25$ scatole con peso medio 362.3 grammi si vuole costruire un intervallo di confidenza al 95% per μ.

Per la proprietà della distribuzione normale e della media campionaria risulta che

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

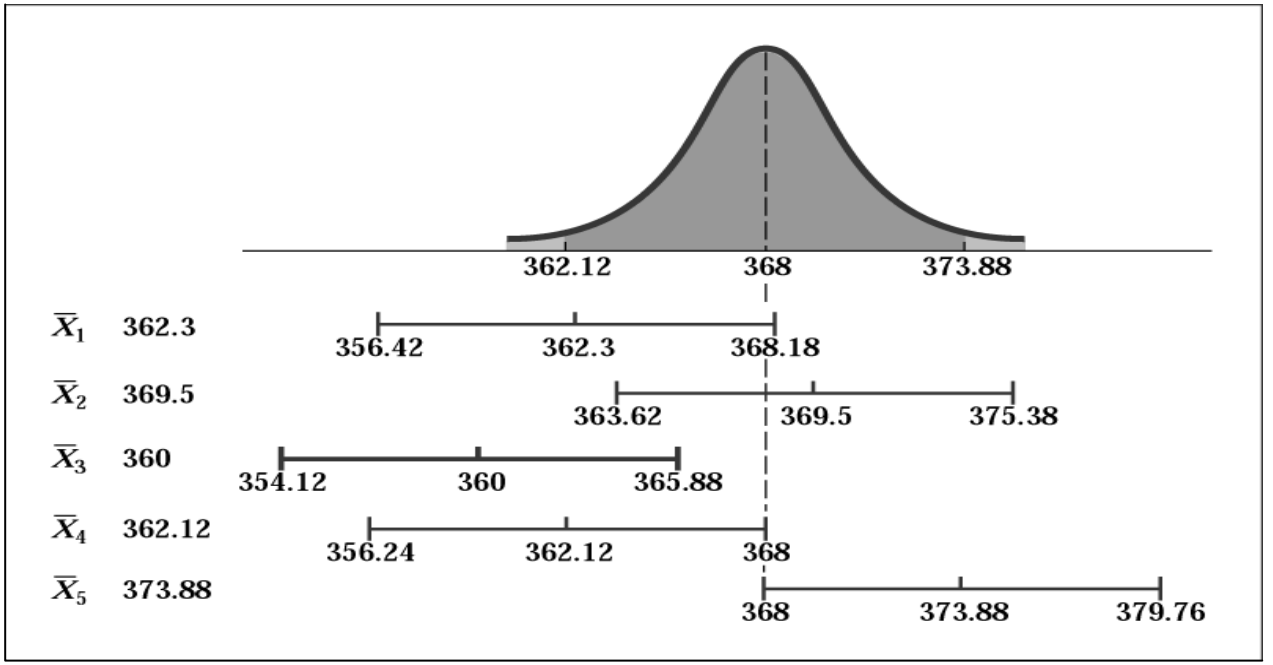
quindi un intervallo di confidenza all'(1−α)% per μ è dato da

$$\bar{X} - Z_{\alpha/2} \cdot \sigma/\sqrt{n} \leq \mu \leq \bar{X} + Z_{\alpha/2} \cdot \sigma/\sqrt{n}$$

Nel caso specifico si ottiene 356.42 ≤ μ ≤ 368.18.

Intervallo di confidenza per la media (σ noto)

Ipotizziamo che μ sia uguale a 368. Per comprendere a fondo il significato della stima per intervallo e le sue proprietà è utile fare riferimento all'ipotetico insieme di tutti i possibili campioni di ampiezza *n* che è possibile ottenere.



Osserviamo che per alcuni campioni la stima per intervalli di μ è corretta, mentre per altri non lo è.

Intervallo di confidenza per la media (σ noto)

Nella pratica estraiamo un solo campione e siccome non conosciamo la media della popolazione non possiamo stabilire se le conclusioni a cui perveniamo sono corrette o meno.

Tuttavia possiamo affermare di avere una fiducia all'(1−α)% che la media appartenga all'intervallo stimato.

Quindi, l'intervallo di confidenza all'(1−α)% della media con σ noto si ottiene utilizzando l'equazione (8.1).

Intervallo di confidenza per la media con σ noto

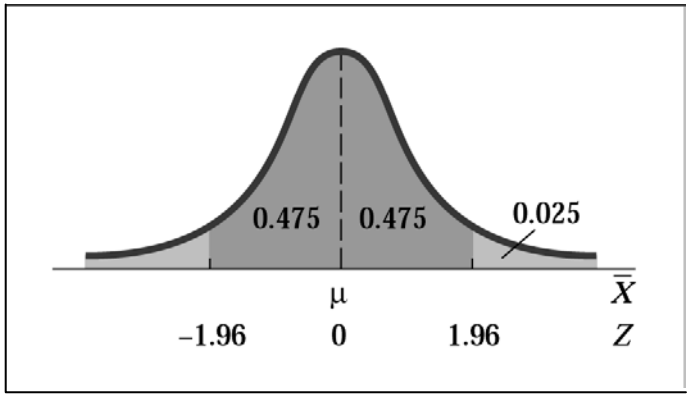
$$\bar{X} - Z_{\alpha/2} \cdot \sigma / \sqrt{n} \leq \mu \leq \bar{X} + Z_{\alpha/2} \cdot \sigma / \sqrt{n} \qquad \textbf{(8.1)}$$

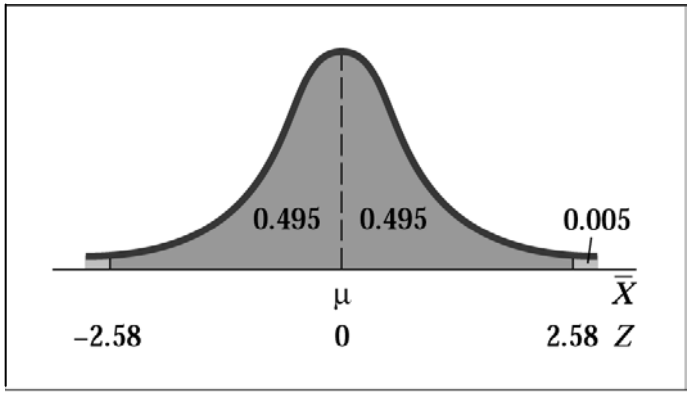
dove $Z_{\alpha/2}$ è il valore a cui corrisponde un'area cumulata pari a (1−α/2) della distribuzione normale standard.

Intervallo di confidenza per la media (σ noto)

In alcuni casi risulta desiderabile un grado di certezza maggiore, ad es. del 99%, ed in altri casi possiamo accettare un grado minore di sicurezza, ad es. del 90%.

Il valore $Z_{\alpha/2}$ di Z che viene scelto per costruire un intervallo di confidenza è chiamato **valore critico**. A ciascun livello di confidenza $(1-\alpha)$ corrisponde un diverso valore critico.





Un livello di confidenza maggiore si ottiene quindi a prezzo di un ampliamento dell'intervallo di confidenza ottenuto: esiste un trade-off tra utilità pratica dell'intervallo e livello di confidenza.

Intervallo di confid. per la media (σ non noto)

In genere lo scarto quadratico medio della popolazione σ, al pari della media μ, non è noto. Pertanto, per ottenere un intervallo di confidenza per la media della popolazione possiamo basarci sulle sole statistiche campionarie $\overline{X}$ e S.

Se la variabile casuale X ha una distribuzione normale allora la statistica

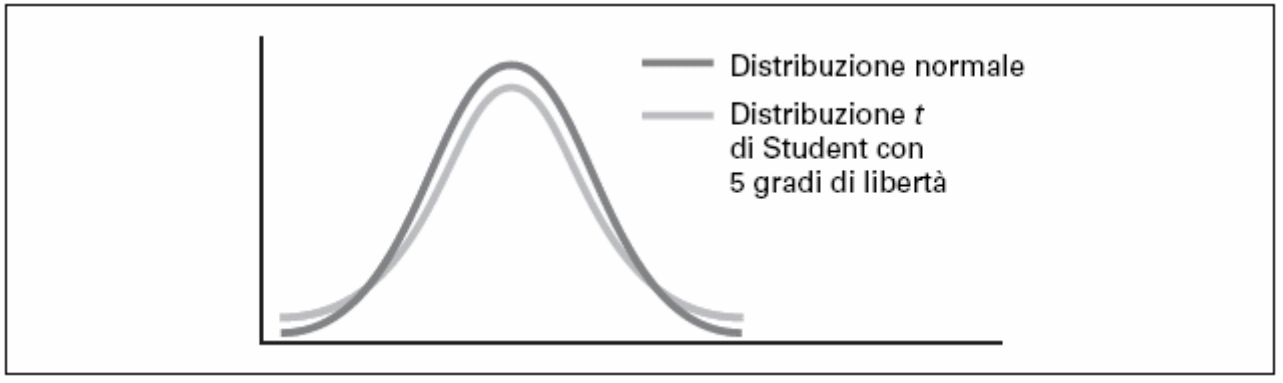
$$t = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

ha una distribuzione *t* di Student con (n−1) **gradi di libertà**.

Se variabile casuale X non ha una distribuzione normale la statistica *t* ha comunque approssimativamente una distribuzione *t* di Student in virtù del Teorema del Limite Centrale.

Intervallo di confid. per la media (σ non noto)

La distribuzione *t* di Student ha una forma molto simile a quella della normale standardizzata. Tuttavia il grafico risulta più appiattito e l'area sottesa sulle code è maggiore di quella della normale a causa del fatto che σ non è noto e viene stimato da S. L'incertezza su σ causa la maggior variabilità di *t*.



All'aumentare dei gradi di libertà, la distribuzione *t* si avvicina progressivamente alla distribuzione normale fino a che le due distribuzioni risultano virtualmente identiche.

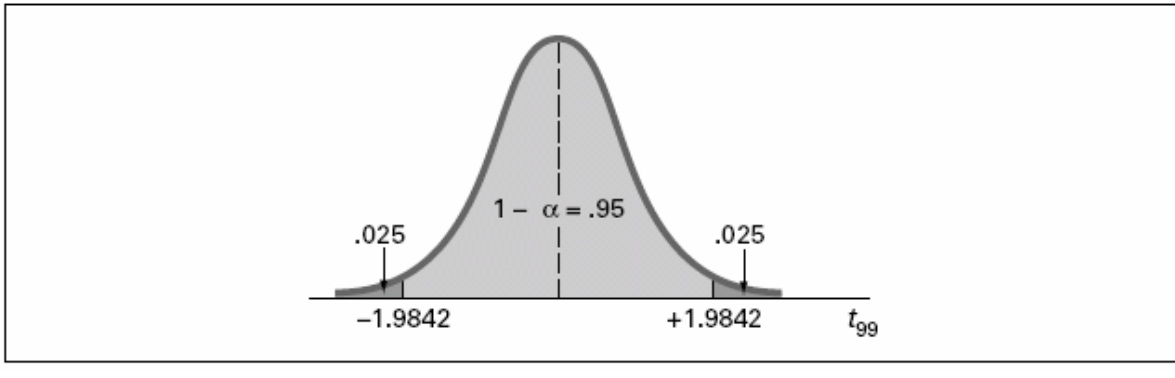
Intervallo di confid. per la media (σ non noto)

I valori critici della distribuzione *t* di Student corrispondenti agli appropriati gradi di libertà si ottengono dalla tavola della distribuzione *t* (Tavola E.3).

Ogni colonna è relativa ad un' area a destra della distribuzione *t*.

	Area nella coda di destra					
Gradi di libertà	0.25	0.10	0.05	0.025	0.01	0.005
1	1.0000	3.0777	6.3138	12.7062	31.8207	63.6574
2	0.8165	1.8856	2.9200	4.3027	6.9646	9.9248
3	0.7649	1.6377	2.3534	3.1824	4.5407	5.8409
4	0.7407	1.5332	2.1318	2.7764	3.7469	4.6041
5	0.7267	1.4759	2.0150	2.5706	3.3649	4.0322
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
96	0.6771	1.2904	1.6609	1.9850	2.3658	2.6280
97	0.6770	1.2903	1.6607	1.9847	2.3654	2.6275
98	0.6770	1.2902	1.6606	1.9845	2.3650	2.6269
99	0.6770	1.2902	1.6604	1.9842	2.3646	2.6264
100	0.6770	1.2901	1.6602	1.9840	2.3642	2.6259

Fonte: Tavola E.3.



Intervallo di confid. per la media (σ non noto)

Il significato dei **gradi di libertà** è legato al fatto che per calcolare S^2 è necessario calcolare preventivamente $\overline{X}$. Quindi, dato il valore di $\overline{X}$, solo $n-1$ osservazioni campionarie sono libere di variare: ci sono quindi $n-1$ gradi di libertà.

L'intervallo di confidenza all'(1−α)% della media quando σ non è noto è definito nell'equazione (8.2).

Intervallo di confidenza per la media (σ non noto)

$$\overline{X} - t_{n-1;\alpha/2} \cdot S/\sqrt{n} \le \mu \le \overline{X} + t_{n-1;\alpha/2} \cdot S/\sqrt{n} \qquad \textbf{(8.2)}$$

dove $t_{n-1;\alpha/2}$ è il valore critico a cui corrisponde un'area cumulata pari a (1−α/2) della distribuzione t di Student con (n−1) gradi di libertà.

Intervallo di confid. per la media (σ non noto)

Esempio: una azienda manifatturiera è interessata a stimare la forza necessaria a rompere un isolatore termico di propria produzione. A questo scopo viene condotto un esperimento dove viene misurato il peso di rottura per un campione di 30 isolatori:

1870	1728	1656	1610	1634	1784	1522	1696	1592	1662
1866	1764	1734	1662	1734	1774	1550	1756	1762	1866
1820	1744	1788	1688	1810	1752	1680	1810	1652	1736

Dai dati campionari si ricava che $\overline{X}$=1723.4 e S=89.55. Dalla tavola E.3 si ottiene il valore critico $t_{29;0.025}$=2.0452, quindi un intervallo di confidenza al 95% per μ è dato da

$$\overline{X} \pm t_{n-1;\alpha/2} \cdot S / \sqrt{n} =$$

$$= 1723.4 \pm (2.0452) \cdot 89.55/\sqrt{30} = 1723.4 \pm 33.44$$

perciò si ottiene $1689.96 \leq \mu \leq 1756.84$.

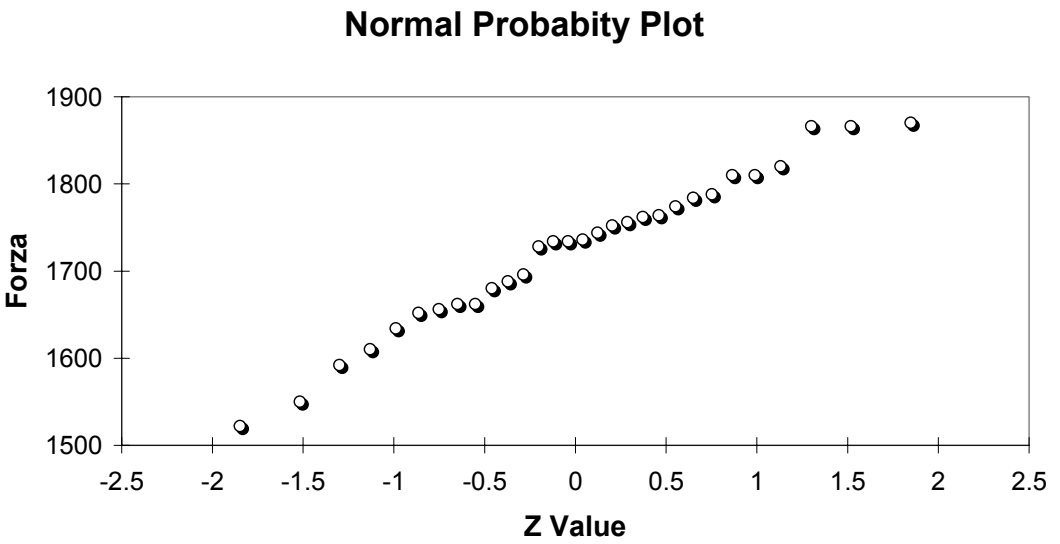
Intervallo di confid. per la media (σ non noto)

Confidence Interval Estimate for the Mean

Data	
Sample Standard Deviation	89.55
Sample Mean	1723.4
Sample Size	30
Confidence Level	95%

Intermediate Calculations	
Standard Error of the Mean	16.35
Degrees of Freedom	29
t Value	2.0452
Interval Half Width	33.44

Confidence Interval	
Interval Lower Limit	1689.96
Interval Upper Limit	1756.84



Possiamo quindi concludere con un livello di confidenza del 95% che la forza media necessaria per rompere un isolatore è compresa tra 1689.96 e 1756.84. La validità dell'intervallo dipende dall'assunzione di normalità per la forza, anche se per campioni di numerosità elevata, questa ipotesi non è così stringente.

Il normal porbablity plot suggerisce che la distribuzione dei dati campionari non è eccessivamente asimmetrica

Intervallo di confidenza per la proporzione

Data una popolazione i cui elementi possiedono una certa caratteristica secondo una data proporzione, indicata dal parametro incognito π, è possibile costruire un intervallo di confidenza per π a partire dal corrispondente stimatore puntuale, dato dalla frequenza campionaria $p=X/n$, dove n è l'ampiezza campionaria e X è il numero di elementi del campione che hanno la caratteristica di interesse.

L'equazione (8.3) definisce l'intervallo di confidenza all'$(1-\alpha)\%$ per la proporzione nella popolazione.

Intervallo di confidenza per la proporzione

$$p - Z_{\alpha/2} \cdot \sqrt{p(1-p)/n} \leq \pi \leq p + Z_{\alpha/2} \cdot \sqrt{p(1-p)/n} \qquad \textbf{(8.3)}$$

dove $Z_{\alpha/2}$ è il valore critico della distribuzione normale standard e si assume che X e $(n-X)$ siano entrambi >5.

Intervallo di confidenza per la proporzione

Esempio: supponiamo che in un campione casuale di 100 fatture, 10 contengano errori e quindi si ha $p=X/n=10/100=0.1$. Fissato un livello di confidenza del 95% si ottiene dalla tavola E.2 il valore critico $Z_{0.025}=1.96$. Quindi, utilizzando l'equazione (8.3) si ottiene

$$p \pm Z_{\alpha/2} \cdot \sqrt{p(1-p)/n} =$$

$$= 0.1 \pm 1.96 \cdot / \sqrt{0.1(1-0.1)/100} = 0.1 \pm 0.0588$$

perciò si ottiene $0.0412 \leq \pi \leq 0.1588$.

Nell'equazione (8.3) si utilizza il valore critico $Z_{\alpha/2}$ della distribuzione normale standard, poichè in virtù del Teorema del Limite Centrale la distribuzione della proporzione campionaria può essere approssimata alla normale, posta l'ampiezza campionaria sufficientemente elevata (e a condizione che $n \cdot p$ e $n \cdot (1-p)$ siano entrambi superiori a 5).

Determinazione dell'ampiezza campionaria

Per determinare l'ampiezza campionaria necessaria per stimare la media dobbiamo considerare l'imprecisione nella stima dovuta alla variabilità campionaria che siamo disposti a tollerare e il livello di confidenza desiderato:

$$\bar{X} \pm Z_{\alpha/2} \cdot \sigma / \sqrt{n} = \bar{X} \pm e$$

La differenza tra la media campionaria e la media della popolazione, indicata con *e*, prende il nome di **errore di campionamento**. Risolvendo per *n* si ottiene l'ampiezza campionaria necessaria per determinare un intervallo di confidenza per la media con errore campionario inferiore ad *e*:

Determinazione dell'ampiezza campionaria - stima della media

$$n = Z_{\alpha/2}^2 \sigma^2 \Big/ e^2 \qquad \textbf{(8.4)}$$

Determinazione dell'ampiezza campionaria

Per determinare l'ampiezza del campione dobbiamo quindi disporre di tre elementi:

1. il livello di confidenza desiderato, che determina il valore di *Z*, il valore critico dalla distribuzione normale standardizzata;
2. l'errore campionario e accettabile;
3. lo scarto quadratico medio σ.

È importante sottolineare che di tali informazioni avremo bisogno prima di estrarre il campione. Nella pratica, può non essere sempre facile determinare queste tre quantità.

Determinazione dell'ampiezza campionaria

Per determinare l'ampiezza campionaria necessaria per stimare la proporzione π dobbiamo conoscere il livello di confidenza desiderato, l'errore campionario accettabile *e* il valore di π.

Determinazione dell'ampiezza campionaria per la stima della proporzione

$$n = Z_{\alpha/2}^2 \pi(1-\pi) \Big/ e^2 \qquad \textbf{(8.5)}$$

Non conoscendo il vero valore di π si potrà inserire nella formula un valore basato su indagini passate o dettato dall'esperienza. Al limite si può inserire $\pi = 0.5$ che è il valore di π che, a parità di *Z* ed *e*, massimizza *n*.