TEORIA DEI CAMPIONI

Psicometria 1 - Lezione 10
Lucidi presentati a lezione

AA 2000/2001 dott. Corrado Caudek

Nella teoria statistica per ***popolazione*** si intende la totalità delle unità potenziali d'osservazione.

L'insieme dei valori assegnati a ciascuna unità di osservazione costituisce la distribuzione della popolazione.

Solitamente questa distribuzione è formata da un numero molto grande di casi.

La media della distribuzione della popolazione si indica con la lettera m e la varianza della distribuzione della popolazione si indica con s^2.

Le variabili (come m e s^2) che descrivono le proprietà della popolazione sono chiamate ***parametri***.

Le variabili che descrivono le corrispondenti proprietà di un campione estratto dalla popolazione (come la media e la varianza denotate da $\overline{X}$ e S^2), invece, sono dette ***statistiche***.

Più specificamente, per *statistica* si intende una qualunque funzione delle *n* variabili aleatorie che costituiscono un campione.

Dato che sono anch'esse delle variabili aleatorie, anche le statistiche possiedono una distribuzione di probabilità.

Tale distribuzione è detta ***distribuzione campionaria***.

DISTRIBUZIONE CAMPIONARIA

DISTRIBUZIONE # 1

Consideriamo la popolazione degli studenti universitari ed esaminiamo la variabile costituita dall'età.

L'insieme dei valori dell'età di ciascuno studente universitario costituisce la ***distribuzione della popolazione*** di questa variabile.

Possiamo calcolare la media e la varianza di questa distribuzione. Questi valori saranno i ***parametri della popolazione***.

Solitamente i parametri della popolazione non sono direttamente accessibili. Sarebbe molto difficile, per esempio, trovare i dati anagrafici di tutti gli studenti universitari.

DISTRIBUZIONE # 2

Consideriamo ora un campione di studenti universitari (ad esempio, quelli in questa aula - *notate che questo non è un campione casuale*). Diciamo che ci sono 123 persone.

La ***distribuzione del campione*** sarà l'insieme dei valori dell'età di tutti gli studenti presenti in questa aula.

La media e la varianza di questo insieme di valori sono due ***statistiche*** di questo campione.

La grandezza di questo campione corrisponde al numero di studenti che lo compongono, diciamo $n = 123$.

DISTRIBUZIONE # 3

Consideriamo ora un terzo tipo di distribuzione.

Supponiamo di esaminare tutti i possibili ***campioni casuali*** di grandezza $n = 123$ che si possono estrarre *con reimissione* dalla popolazione degli studenti universitari.

Per ciascuno di questi campioni possiamo calcolare una data statistica.

Il valore di questa statistica varierà da campione a campione.

L'insieme di tutte queste statistiche genera una nuova distribuzione. Questa distribuzione si chiama ***distribuzione campionaria***.

Se la statistica calcolata è la media, possiamo definire la *distribuzione campionaria della media*;

se la statistica calcolata è la varianza, possiamo definire la *distribuzione campionaria della varianza*.

Una volta calcolata la distribuzione campionaria di queste due statistiche possiamo chiederci, ad esempio,

quali sono la media e la varianza della distribuzione campionaria della media,

quali sono la media e la varianza della distribuzione campionaria della varianza.

La distribuzione campionaria è una distribuzione di probabilità teorica che ci fornisce un modello della distribuzione di frequenza che si otterrebbe per tutti i possibili valori di una data statistica basata su un campione di *n* casi ***se il processo di campionamento venisse ripetuto infinite volte***.

La distribuzione campionaria non è un concetto del tutto nuovo, in quanto in precedenza abbiamo già preso in esame una distribuzione campionaria.

La distribuzione binomiale infatti è una distribuzione campionaria che ci mostra la probabilità di osservare ciascuno dei possibili numeri di successi che si possono ottenere in un campione di *n* prove di un processo bernoulliano, per tutte le possibili grandezze del campione.

Abbiamo anche esaminato, ad esempio, la distribuzione campionaria *t* di Student, per la statistica

$$t = \frac{\bar{Y} - m}{s/\sqrt{n}}$$

Così come le distribuzioni della popolazione, anche le distribuzioni campionarie possono essere discrete o continue.

La distribuzione binomiale, ad esempio, è una distribuzione discreta, mentre la distribuzione *t* è una distribuzione continua.

ALCUNE DISTRIBUZIONI CAMPIONARIE ASSOCIATE ALLA DISTRIBUZIONE NORMALE

1. Teorema.

Sia $Y_1, Y_2, \ldots, Y_n$ un campione casuale di grandezza n tratto da una distribuzione normale con media m e varianza σ^2. La distribuzione campionaria della media $\overline{Y}$ di queste osservazioni è normale con media $m_{\overline{Y}} = m$ e varianza $s^2_{\overline{Y}} = s^2/n$

In base al teorema precedente, dunque, la variabile Z

$$Z = \frac{\bar{Y} - m_{\bar{Y}}}{s_{\bar{Y}}} = \frac{\bar{Y} - m}{s / \sqrt{n}}$$

sarà distribuita normalmente con media 0 e varianza unitaria.

2. Teorema.

Sia $Y_1, Y_2, \ldots, Y_n$ un campione casuale di grandezza n tratto da una distribuzione normale con media m e varianza σ^2. Sia $Z_i = (Y - m)/s$. Allora, la variabile

$$\sum_{i=1}^{n} Z_i^2$$

sarà distribuita come c^2 con n gradi di libertà.

3. Teorema.

Sia $Y_1, Y_2, \ldots, Y_n$ un campione casuale di grandezza n tratto da una distribuzione normale con media m e varianza σ^2. Allora, la variabile

$$(n-1)\frac{s^2}{\sigma^2}$$

sarà distribuita come c^2 con $(n - 1)$ gradi di libertà.

4. Teorema.

Sia $Y_1, Y_2, \ldots, Y_n$ un campione casuale di grandezza n tratto da una distribuzione normale con media m e varianza σ^2. La variabile

$$T = \frac{\bar{Y} - m}{s/\sqrt{n}}$$

seguirà la distribuzione *t* di Student con $(n - 1)$ gradi di libertà.

5. Teorema.

Siano W_1 e W_2 due variabili aleatorie indipendenti distribuite come c^2 con ν_1 e ν_2 gradi di libertà, rispettivamente. Allora, la variabile

$$F = \frac{W_1 / n_1}{W_2 / n_2}$$

seguirà la distribuzione F con ν_1 gradi di libertà al numeratore e ν_2 gradi di libertà al denominatore.

IL CAMPIONAMENTO

Campionamento casuale

Campionamento a grappoli

Campionamento stratificato

Campionamento a più stadi

Una volta chiarita la nozione di distribuzione campionaria passiamo ad esaminare le proprietà degli stimatori.

PROPRIETA’ DEGLI STIMATORI

Un problema centrale della statistica inferenziale è la stima dei parametri della popolazione per mezzo delle statistiche del campione.

Il fatto che un campione rappresenti soltanto una piccola parte della popolazione fa sì che sia pressoché impossibile che una data statistica corrisponda esattamente al corrispondente parametro della popolazione.

Inoltre, possiamo dire che statistiche diverse si approssimano in modo diverso ai parametri della popolazione.

Chiediamoci, dunque, come si possono rappresentare le relazioni tra le statistiche e i parametri della popolazione.

Una parametro può essere stimato secondo due modalità:

- mediante la ***stima puntuale***

- mediante la ***stima intervallare***

STIMA PUNTUALE

Quattro proprietà sono ritenute auspicabili per le statistiche usate quali stimatori dei parametri della popolazione:

1. ***correttezza,***
2. ***efficienza,***
3. ***consistenza,***
4. ***sufficienza.***

1. CORRETTEZZA

Definiamo come "predittore equilibrato" (*unbiased*) o “centrato sul parametro” della popolazione l’indice statistico il cui valore atteso è uguale al parametro corrispondente dell’universo.

Detto F un generico indice statistico, e detto φ il parametro corrispondente, diremo che F è un predittore equilibrato di φ se

$$E(F) = j$$

2. EFFICIENZA

L'assenza di distorsione non è l'unica proprietà desiderabile di uno stimatore.

Uno stimatore, infatti, potrebbe essere centrato sul parametro perché errori molto grandi di segno positivo vengono bilanciati da errori molto grandi di segno negativo.

Una seconda proprietà desiderabile di uno stimatore è chiamata "efficienza".

Supponiamo di avere due stimatori F_1 e F_2 centrati sul parametro f e calcolati su campioni di eguale grandezza.

Siano $V(F_1)$ e $V(F_2)$ le varianze dei due stimatori.

L'efficienza relativa di F_1 rispetto a F_2 è definita dal rapporto:

$$eff\left(F_1, F_2\right) = \frac{V\left(F_2\right)}{V\left(F_1\right)}$$

Se F_1 e F_2 sono entrambi stimatori senza distorsione del parametro f, l'efficienza di F_1 relativa a F_2 è maggiore di 1 solo se la varianza di F_2 è maggiore della varianza di F_1.

In queste circostanze, F_1 è uno stimatore migliore di F_2.

Supponiamo, ad esempio, di stimare la media della popolazione usando due stimatori, la mediana (F_1) di un campione di grandezza n e la media (F_2) di un campione di eguale grandezza.

Può essere dimostrato che, per campioni di grandi dimensioni, la varianza della mediana è

$$V(F_1) = 1.2533^2 (s^2/n) .$$

Ne segue che l'efficienza della mediana relativamente alla media campionaria è

$$eff\left(F_1, F_2\right) = \frac{V\left(F_2\right)}{V\left(F_1\right)} = \frac{s^2/n}{(1.2533)^2 s^2/n} = 0.6366$$

La variabilità associata alla media campionaria è dunque circa il 64% della variabilità associata alla mediana campionaria, il che ci consente di concludere che è preferibile usare la media campionaria anziché la mediana quale stimatore della media della popolazione.

http://www.ruf.rice.edu/~lane/stat_sim/sampling_dist/index.html

3. CONSISTENZA

Uno stimatore F del parametro f si dice consistente quando la dispersione di F intorno a f diminuisce all'aumentare della grandezza del campione.

Esempio. Supponiamo di effettuare n lanci di una moneta. Se i lanci sono indipendenti, il numero Y degli esiti "testa" segue una distribuzione binomiale con probabilità di osservare l'esito "testa" uguale a p.

Una stima della probabilità p è fornita dalla variabile aleatoria corrispondente al rapporto tra Y e il numero n di lanci della moneta (Y/n).

Dal punto di vista intuitivo, potremmo aspettarci che la probabilità

$$P\left(\left|\frac{Y}{n} - p\right| \leq e\right)$$

tenda a 1 al crescere di n, per qualsiasi arbitrario numero positivo e.

Questo in effetti si verifica con $n \to \infty$, e la variabile aleatoria (Y/n) viene detta uno stimatore consistente di p.

3. SUFFICIENZA

Una statistica F si dice sufficiente per un parametro f se riassume tutta l'informazione rilevante per f che si trova nel campione.

In altre parole, se F è uno stimatore sufficiente per f allora la stima di f non può essere migliorata considerando altri aspetti dei dati che non siano già stati considerati dallo stimatore F.

Consideriamo n lanci di una moneta con probabilità p di osservare l'esito "testa".

Ciascun lancio $X_1, X_2, \ldots, X_n$ è una variabile aleatoria indipendente con la seguente distribuzione di probabilità:

$$X_i = \begin{cases} 1, & \text{con probabilità } p \\ 0, & \text{con probabilità } q = 1 - p \end{cases}$$

In precedenza, per stimare la probabilità p abbiamo usato la variabile aleatoria M_n:

$$M_n = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{Y}{n}$$

Potremmo ora chiederci se esiste una diversa funzione di $X_1, X_2, \ldots, X_n$ in grado di fornire altre informazioni a proposito di p che non siano già contenute in M_n.

Dato che può essere dimostrato che M_n riassume tutta l'informazione concernente p presente nelle $X_1, X_2, \ldots, X_n$, possiamo concludere che M_n è una statistica sufficiente per p.

Analogamente, è stato dimostrato che la media campionaria è uno stimatore sufficiente della media incognita m di una popolazione normale.

VALORI ATTESI E ERRORI STANDARD DEGLI STIMATORI PUNTUALI PIU' COMUNI

Parametro	***Grandezza del(i) Campione(i)***	***Stimatore***	***Valore atteso***	***Errore standard***
m	n	$\bar{Y}$	m	$\frac{s}{\sqrt{n}}$
p	n	$\hat{p} = \frac{}{n}$	p	$\frac{pq}{\sqrt{n}}$
$m_1 - m_2$	n_1 *e* n_2	$\bar{Y}_1 - \bar{Y}_2$	$m_1 - m_2$	$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

STIMA INTERVALLARE

Le statistiche del campione (per esempio, la media) non sono mai esattamente uguali ai parametri della popolazione a causa degli errori introdotti dal processo di campionamento. E' dunque necessario stabilire l'entità di questo errore.

Solitamente questo problema viene affrontato calcolando un ***intervallo di fiducia***, ovvero la gamma dei valori che hanno una probabilità prestabilita di contenere il vero parametro della popolazione (per esempio, la media).

Come si interpreta un intervallo di fiducia?

L'intervallo di confidenza si calcola in base a certe regole sulla base delle informazioni fornite dal campione. Supponiamo di costruire l'intervallo di fiducia del 95%, ovvero quell'intervallo che contiene il parametro della popolazione con probabilità .95.

Estraiamo un campione casuale dalla popolazione e calcoliamo l'intervallo di fiducia del 95%. Otteniamo in questo modo due valori che delimitano l'intervallo. Ripetiamo questo processo con un altro campione casuale. Calcoliamo anche in questo caso l'intervallo di fiducia del 95%. Otterremo così altri due valori che delimitano l'intervallo di confidenza.

Se ripetessimo questo processo infinite volte, ci accorgeremmo che non tutti gli intervalli calcolati in base agli infiniti campioni casuali estratti dalla popolazione contengono il parametro stimato della popolazione.

Se abbiamo calcolato l'intervallo di fiducia del 95%, allora il parametro della popolazione sarà contenuto nell'intervallo calcolato ***soltanto*** nel 95% dei casi.

In altre parole, dire che un intervallo di fiducia del 95% contiene il vero parametro della popolazione con probabilità .95 significa dire che, se ripetessimo il processo di campionamento infinite volte e calcolassimo l'intervallo di fiducia per ciascun campione, allora otterremmo un intervallo che effettivamente contiene il vero parametro della popolazione nel 95% dei casi.

http://www.ruf.rice.edu/~lane/stat_sim/conf_interval/index.html

Come si calcola un ***intervallo di fiducia***? Questo problema si risolve decidendo a priori un determinato livello di probabilità e poi trovando due valori, *a* e *b*, che contengono nel loro intervallo il parametro della popolazione (ad esempio, la media) al livello di probabilità prestabilito.

Troviamo ora l'intervallo che ha probabilità .90 di contenere la media della popolazione.

LA DISEGUAGLIANZA DI CEBICEV

Per costruire un intervallo di fiducia è possibile usare la disuguaglianza di Cebicev.

La diseguaglianza di Cebicev afferma che, per qualunque distribuzione di probabilità, deve essere vero che:

$$P\left(\left|X-m\right|\geq e\right)\leq\frac{s^{2}}{e^{2}}$$

ovvero

$$P\left(\left|X-m\right|\leq e\right)\geq 1-\frac{s^{2}}{e^{2}}$$

Sia $e = ks$. Possiamo allora scrivere:

$$P\left(\left|X - m\right| \geq ks\right) \leq \frac{s^2}{k^2 s^2}$$

$$P\left(\left|X - m\right| \geq ks\right) \leq \frac{1}{k^2}$$

$$P\left(\left|\frac{X - m}{s}\right| \geq k\right) \leq \frac{1}{k^2}$$

$$P\left(\left|z\right| \geq k\right) \leq \frac{1}{k^2} \quad \text{ovvero} \quad P\left(\left|z\right| \leq k\right) \geq 1 - \frac{1}{k^2}$$

In altre parole: la probabilità che il *valore assoluto* di un punteggio standardizzato tratto a caso da una distribuzione qualunque sia maggiore o uguale a *k* è sempre minore o uguale a $1/k^2$.

Ad esempio, data una distribuzione qualunque con un certa media e varianza, la probabilità di estrarre a caso un osservazione standardizzata con un valore maggiore o uguale a 2 (in valore assoluto) deve essere minore o uguale a 1/4.

La probabilità di estrarre a caso un osservazione standardizzata con un valore maggiore o uguale a 10 (in valore assoluto) deve essere minore o uguale a 1/100.

Se possiamo assumere, inoltre, che la distribuzione è simmetrica e unimodale, allora la relazione diventa:

$$P\left(|z| \geq k\right) \leq \frac{4}{9}\left(\frac{1}{k^2}\right)$$

Esempio. Si calcoli la probabilità di estrarre a caso da una distribuzione un'osservazione con un valore di 3 o più deviazioni standard dalla media.

$$P(|z| \geq k) \leq \frac{1}{k^2}$$

$$P(|z| \geq 3) \leq \frac{1}{3^2}$$

Se possiamo assumere che la distribuzione è unimodale e simmetrica, allora la probabilità cercata diventa uguale a:

$$P(|z| \geq 3) \leq \left(\frac{4}{9}\right)\frac{1}{3^2}$$

Allo stesso modo, possiamo dire che vi sono almeno 1 - 4/9 osservazioni contenute tra ±1 deviazione standard dalla media.

$$P\left(|z| \leq 1\right) \geq 1 - \left(\frac{4}{9}\right)\left(\frac{1}{1^2}\right)$$

Vi sono almeno 1 - (4/9) (1/4) osservazioni contenute tra ±2 deviazione standard dalla media.

$$P\left(|z| \leq 2\right) \geq 1 - \left(\frac{4}{9}\right)\left(\frac{1}{2^2}\right)$$

APPLICHIAMO ORA LA DISEGUAGLIANZA DI CEBICEV ALLA DISTRIBUZIONE CAMPIONARIA DELLA MEDIA

Nel caso della ***distribuzione campionaria della media***, sarà vero che:

$$P\left(\left|\frac{\overline{X}-m}{s_{\overline{X}}}\right| \geq k\right) \leq \frac{1}{k^2}$$

| z | > *a* significa z > *a* e z < -*a*. Dunque, possiamo riscrivere la diseguaglianza precedente come:

$$P\left(-k \geq \frac{\overline{X}-m}{s_{\overline{X}}} \geq k\right) \leq \frac{1}{k^2}$$

$$P\left(-k \geq \frac{\overline{X}-m}{s_{\overline{X}}} \geq k\right) \leq \frac{1}{k^2}$$

$$P\left(-ks_{\overline{X}} \geq \overline{X}-m \geq ks_{\overline{X}}\right) \leq \frac{1}{k^2}$$

$$P\left(ks_{\overline{X}} \leq -\overline{X}+m \leq -ks_{\overline{X}}\right) \leq \frac{1}{k^2}$$

$$P\left(\overline{X}+ks_{\overline{X}} \leq m \leq \overline{X}-ks_{\overline{X}}\right) \leq \frac{1}{k^2}$$

In conclusione, la probabilità dell'evento complementare (ovvero, la media della popolazione è compresa *all'interno* dell'intervallo) sarà dunque

$$P\left(\overline{X} + k s_{\overline{X}} \geq m \geq \overline{X} - k s_{\overline{X}}\right) \geq 1 - \frac{1}{k^2}$$

A questo risultato possiamo attribuire la seguente interpretazione:

se consideriamo la distribuzione campionaria della media, allora l'intervallo compreso tra $\overline{X} - ks_{\overline{X}}$ e $\overline{X} + ks_{\overline{X}}$ avrà una probabilità *almeno* uguale a $1 - 1/k^2$ di contenere il valore m della media della popolazione.

Esempio. Quale è la probabilità che la media μ della popolazione si trovi in un intervallo di ± 2 errori standard dalla media del campione?

$$P\left(\overline{X} - k s_{\overline{X}} \leq m \leq \overline{X} + k s_{\overline{X}}\right) \geq 1 - \frac{1}{k^2}$$

In questo caso, $k = 2$, quindi la probabilità che cerchiamo è maggiore o uguale a $1 - 1/2^2 = .75$.

Esempio. Supponiamo di disporre di un campione casuale di $n = 50$ osservazioni indipendenti tratte da una popolazione con media m non conosciuta e deviazione standard conosciuta e uguale a 20. La media del campione è uguale a 124.

Si determini la probabilità che la vera media della popolazione sia contenuta nell'intervallo $\overline{X} \pm 3s_{\overline{X}}$

Si rendano inoltre espliciti i limiti dell'intervallo di fiducia.

$$P\left(\overline{X} - 3s_{\overline{X}} \leq m \leq \overline{X} + 3s_{\overline{X}}\right) \geq 1 - \frac{1}{3^2}$$

La probabilità che cerchiamo è dunque uguale a .89.

$$s_{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{20}{\sqrt{50}} = 2.83$$

124 + 3(2.83) = 132.49

124 - 3(2.83) = 115.51

In conclusione, l'intervallo [115.51, 132.49] contiene la vera media della popolazione con una probabilità maggiore o uguale a .89.

DISTRIBUZIONE CAMPIONARIA DELLA MEDIA

In precedenza abbiamo visto che la distribuzione campionaria della media di campioni di dimensioni n tratti da una ***popolazione normale*** con media m e varianza s^2 è ***normale con media*** m ***e varianza*** s^2/n.

Consideriamo ora in maggiore dettaglio la distribuzione campionaria della media nel caso in cui non si possa assumere la normalità della popolazione (e dunque il teorema precedente non si applica).

Inzieremo a trovare il valore atteso e la varianza della ***somma*** di n variabili aleatorie con valore atteso m e varianza s^2.

Considereremo poi il valore atteso e la varianza della ***media*** di n variabili aleatorie con valore atteso m e varianza s^2.

Enunceremo infine il ***teorema del limite centrale***.

VALORE ATTESO E VARIANZA DELLA SOMMA

Sia X una variabile aleatoria con valore atteso $E(X) = m$ e varianza $V(X) = s^2$.

Sia S_n la somma di n variabili indipendenti X_i:
$S_n = X_1 + X_2 + \ldots + X_n$.

$$E(S_n) = E(X_1 + X_2 + \ldots + X_n) =$$

$$= E(X_1) + E(X_1) + \ldots + E(X_n) = nm$$

$$V(S_n) = V(X_1 + X_2 + \ldots + X_n) =$$

$$= V(X_1) + V(X_1) + \ldots + V(X_n) = n s^2$$

In conclusione, la variabile aleatoria S_n ha valore atteso

$\boldsymbol{E(S_n) = nm}$ e varianza $\boldsymbol{V(S_n) = n s^2}$.

VALORE ATTESO E VARIANZA DELLA MEDIA

Sia $M_n = S_n / n$.

$$E(M_n) = E\left(\frac{S_n}{n}\right) = \frac{1}{n} E(S_n) = \frac{1}{n} nm = m$$

$$V(M_n) = V\left(\frac{S_n}{n}\right) = \frac{1}{n^2} V(S_n) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

In conclusione, la media M_n ha valore atteso $\boldsymbol{E(M_n) = m}$

e varianza $\boldsymbol{V(M_n) = s^2/n}$.

Si noti che la varianza della distribuzione campionaria della media diminuisce al crescere delle dimensioni del campione: $V(M_n) = s^2/n$.

Questo significa che la media del campione è uno stimatore consistente per la media della popolazione. E' stato inoltre dimostrato che la media del campione è uno stimatore massimamente efficiente, sufficiente e corretto.

Una volta trovata la media e la varianza della distribuzione campionaria della media, chiediamoci ora quale è la forma della distribuzione campionaria della media.

La risposta a questa domanda ci viene data dal ***Teorema del Limite Centrale:***

http://www.stat.sc.edu/~west/javahtml/CLT.html

http://www.ruf.rice.edu/~lane/stat_sim/sampling_dist/index.html

TEOREMA DEL LIMITE CENTRALE

Se $X_1, X_2, X_3, \ldots$ è un insieme di variabili aleatorie con valore atteso uguale a μ e varianza uguale a σ^2 , allora la distribuzione di

$$z = \frac{\overline{X} - m}{s / \sqrt{n}}$$

tende alla distribuzione normale standardizzata con

$n \longrightarrow \infty$

… in altre parole:

se tutti i possibili campioni di grandezza n vengono estratti da una popolazione con media m *e varianza* s^2*, all'aumentare di n le medie di questi campioni approssimeranno una* ***distribuzione normale*** *con media* m *e varianza* s^2/n.

Il teorema del limite centrale è così importante perchè ci consente di specificare completamente la distribuzione campionaria della media di campioni casuali di grandezza *n* senza fare nessuna assunzione a proposito della forma della distribuzione della popolazione.

ESERCIZI

E1 Esercizi 1 - 8 Hays, p. 214-215.

APPROSSIMAZIONE NORMALE ALLA BINOMIALE

Il teorema del limite centrale afferma che i valori di probabilità della distribuzione binomiale tendono a quelli della distribuzione normale standardizzata al crescere di *n*, quando il numero di successi *r* in *n* prove bernoulliane viene trasformato in unità standard *x* in base alla formula:

$$x = \frac{r - np}{\sqrt{npq}}$$

Esempio. Una moneta viene lanciata 100 volte. Si trovi la probabilità che la proporzione di esiti *T* sia compresa nell'intervallo 40 - 60.

Per risolvere questo problema, trasformiamo i limiti dell'intervallo in punteggi standardizzati e usiamo la curva normale per stimare la probabilità

$$E(X) = n\,p = 100\ .5 = 50.$$

$$SD(X) = Sqrt(npq) = Sqrt(100\ .5\ .5) = 5.$$

$$z_1 = (40 - 50)/5 = -2$$

$$z_2 = (60 - 50)/5 = 2$$

L’area sottesa alla curva normale tra -2 e +2 è 0.9545.

Dunque, la probabilità di osservare tra i 40 e i 60 esiti *T* nel caso di 100 lanci di una moneta onesta è .9545.

Si noti che, per risolvere questo problema, non abbiamo sommato le probabilità della distribuzione binomiale per tutti i valori compresi tra 40 e 60 successi. Abbiamo invece usato l'approssimazione normale alla binomiale.

Esempio. Nelle università americane gli studenti fanno domanda di ammissione e l'università decide se ammettere o meno gli studenti in base al punteggio che hanno conseguito nei test di ammissione.

Non tutti gli studenti che vengono accettati però si iscrivono, dato che ciascuno studente fa domanda a più università e si iscrive all'università migliore tra quelle che lo hanno accettato.

Supponete che un college americano non possa ammettere più di 1060 studenti.

Dalle statistiche effettuate negli anni passati si è stabilito che uno studente accettato in questa università ha una probabilità di .6 di iscriversi.

Supponete inoltre che l'università invii una lettera di accettazione a 1700 studenti in un anno.

Quale è la probabilità che a questa università si iscrivano troppi studenti?

Consideriamo l'iscrizione come un processo bernulliano (iscrizione = successo, non iscrizione = insuccesso) e usiamo l'approssimazione normale alla binomiale.

Il valore atteso del numero di iscrizioni, X, è:

$E(X) = n\ p = 1700\ .6 = 1020.$

$ES = Sqrt(npq) = Sqrt(1700\ .6\ .4) = 20.$

Il valore critico che non dobbiamo eccedere è 1060.
Il punteggio standardizzato si ottiene come:

$z = (1060 - 1020) / 20 = 2.$

I punteggi in eccesso di 1060 rappresentano gli eventi che vogliamo evitare. Il problema è di calcolare la probabilità di questi eventi.

Per trovare questa probabilità dobbiamo trovare l'area sottesa alla curva normale tra 2.0 e $+\infty$.

Questa probabilità è uguale a 0.02275.

In conclusione, la probabilità che si verifichi un evento indesiderato per gli amministratori dell'università avendo accettato 1700 studenti, dunque, è molto piccola.

ESERCIZIO

E2 Sia S la somma del numero di esiti T in 100 lanci di una moneta onesta. Si usi il *TLC* per stimare:

(a) $P(S < 45)$

(b) $P(45 < S < 55)$

(c) $P(S > 63)$

DISTRIBUZIONE CAMPIONARIA DELLA VARIANZA

In precedenza abbiamo detto che la *media campionaria* fornisce una stima priva di errore sistematico della media della popolazione. In altre parole, la media della distribuzione campionaria della media campionaria è uguale alla media della popolazione.

Le cose sono diverse, invece, per quel che riguarda la varianza di un campione. La varianza campionaria, infatti, non fornisce una stima priva di errore sistematico della varianza della popolazione. In altre parole, la media della distribuzione campionaria della varianza campionaria è diversa dalla varianza della popolazione.

Si può dimostrare, infatti, che la media della distribuzione campionaria della varianza campionaria, $E(S^2)$, è uguale alla differenza tra la varianza della popolazione e la varianza della distribuzione campionaria della media:

$$E\left(S^2\right) = \sigma^2 - \sigma^2_{\bar{X}}$$

In generale, questa differenza non sarà uguale alla varianza della popolazione dato che la varianza della distribuzione campionaria della media non è uguale a zero. Quindi, la varianza del campione tende ad essere più piccola della varianza della popolazione.

Per correggere questo errore sistematico notiamo che:

$$E\left(S^2\right) = s^2 - \frac{s^2}{n} = \frac{ns^2 - s^2}{n} = \left(\frac{n-1}{n}\right)s^2$$

La varianza campionaria media, dunque, è più piccola della varianza della popolazione di un fattore uguale a *(n- 1)/n*.

Per ottenere una stima priva di errore sistematico della varianza della popolazione modifichiamo dunque la varianza del campione nel modo seguente:

$$s^2 = \frac{n}{n-1} S^2$$

E' infatti chiaro che il valore atteso della distribuzione campionaria della statistica s^2 sarà uguale alla varianza della popolazione.

ESERCIZIO

Si dimostri l’affermazione precedente.

Una stima priva di errore sistematico della varianza della popolazione può essere calcolata direttamente dai dati del campione:

$$s^2 = \frac{n}{n-1} S^2 = \frac{n}{n-1} \frac{\sum_i \left(X_i - \overline{X}\right)^2}{n} = \frac{\sum_i \left(X_i - \overline{X}\right)^2}{n-1}$$

In conclusione, dunque, usiamo la statistica S^2 per indicare la varianza del campione quale indice descrittivo, e la statistica s^2 quale stimatore privo di errore sistematico della varianza della popolazione s^2.

Un ulteriore complicazione nasce dal fatto che la deviazione standard è uguale alla radice quadrata della varianza.

Dato che la radice quadrata non è una funzione lineare, questo significa che la deviazione standard non fornisce una stima priva di errore sistematico della deviazione standard della popolazione.

Ci sono dei metodi che ci consentono di correggere questo errore, ma non li esamineremo dato che, quando la grandezza del campione è ragionevolmente grande, l'entità di questo errore è trascurabile.

La distribuzione campionaria della varianza di campioni estratti da una popolazione normale non è normale, ma è collegata alla distribuzione c^2.

Abbiamo visto in precedenza che il rapporto tra la varianza del campione s^2 e la varianza della popolazione, moltiplicato per $(n - 1)$, si distribuisce secondo la legge c^2 con $\nu = n - 1$ gradi di libertà.

In base a questo principio, è possibile costruire gli intervalli di fiducia per la varianza della popolazione sulla base delle informazioni fornite dal campione.

ERRORE STANDARD STIMATO DELLA MEDIA CAMPIONARIA

In base al teorema del limite centrale, nel caso di $n > 30$, possiamo dire che la distribuzione campionaria della media è normale con media uguale alla media della popolazione e varianza uguale alla varianza della popolazione divisa per n.

In generale, però, questa conclusione ci è di poco aiuto dato che la varianza della popolazione non è conosciuta.

E' dunque necessario stimare la varianza della popolazione per calcolare l'errore standard della media.

Chiediamoci ora come sia possibile stimare l'errore standard della distribuzione campionaria della media a partire dai dati di un campione.

Questa stima può essere calcolata in due modi:

$$\hat{s}_{\bar{X}} = \sqrt{\frac{s^2}{n}} = \sqrt{\frac{s^2}{n}} = \frac{s}{\sqrt{n}}$$

$$\hat{s}_{\bar{X}} = \sqrt{\frac{s^2}{n}} = \sqrt{\left(\frac{n}{n-1}\right)\frac{S^2}{n}} = \frac{S}{\sqrt{n-1}}$$

In conclusione, le caratteristiche della distribuzione campionaria della media sono le seguenti.

Se possiamo assumere che la popolazione sia normale, con media μ e varianza σ^2, allora la distribuzione campionaria della media sarà

- ***normale***
- ***con media uguale a*** $E(\bar{X}) = \mu$
- ***con errore standard uguale a*** $\sigma_{\bar{X}} = \sigma / \sqrt{n}$

Se la popolazione ha media μ e varianza σ^2, ma non è distribuita normalmente, allora in base al teorema del limite centrale, con $n > 30$, la distribuzione campionaria della media sarà approssimativamente

- ***normale***

- ***con media uguale a*** $E(\overline{X}) = \mu$

- ***con errore standard uguale a*** $\sigma_{\overline{X}} = \sigma / \sqrt{n}$

In entrambi i casi (popolazione normale oppure popolazione non normale con campione > 30 osservazioni), la varianza della popolazione (solitamente non conosciuta) può essere stimata con s^2.

Se dunque possiamo ritenere che le medie dei campioni siano distribuite normalmente con media m e varianza s^2, allora la quantità

$$z = \frac{\overline{X} - m}{\hat{s}_{\overline{X}}} = \frac{\overline{X} - m}{s/\sqrt{n}}$$

sarà distribuita come una variabile normale standardizzata.

STIMA DEI PARAMETRI DELLA POPOLAZIONE NEL CASO DI PIU' CAMPIONI

Supponiamo di avere diversi campioni indipendenti e di volere stimare la media e la varianza della popolazione.

Iniziamo considerando il caso di due campioni indipendenti. Per ciascun campione calcoliamo la media. La stima congiunta della media della popolazione sarà:

$$\hat{m} = \frac{n_1 \overline{X}_1 + n_2 \overline{X}_2}{n_1 + n_2}$$

ovvero, la media ponderata delle medie dei due campioni.

Nel caso di 3 campioni avremo:

$$\hat{m} = \frac{n_1 \overline{X}_1 + n_2 \overline{X}_2 + n_3 \overline{X}_3}{n_1 + n_2 + n_3}$$

Il fatto che la stima congiunta della media della popolazione sia da preferire alla stima separata fornita da ciascun campione è dimostrato dall'errore standard della stima congiunta della media.

L'errore standard di una distribuzione campionaria, infatti, ci fornisce un'indicazione del grado di errore che compiamo usando la *statistica* del campione per stimare il *parametro* della popolazione.

Per due campioni indipendenti, ciascuno composto da *n* osservazioni tratte dalla medesima popolazione, l'errore standard è:

$$s_{\bar{X}} = \frac{s}{\sqrt{n_1 + n_2}}$$

che è necessariamente più piccolo dell'errore standard calcolato a partire da $\bar{X}_1$ e $\bar{X}_2$ considerate isolatamente.

Per 3 campioni avremo:

$$s_{\bar{X}} = \frac{s}{\sqrt{n_1 + n_2 + n_3}}$$

Questi risultati, però, sono espressi nei termini della deviazione standard della popolazione (σ) che, solitamente, non è nota. Possiamo però stimare questo parametro usando la varianza del campione.

Nel caso di due campioni indipendenti, la varianza stimata della popolazione ($\hat{s}^2$) si può calcolare come:

$$\hat{s}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

e l'errore standard stimato della distribuzione campionaria della media sarà uguale a:

$$\hat{s}_{\bar{X}} = \frac{\hat{s}}{\sqrt{n_1 + n_2}}$$

INTERVALLO DI FIDUCIA PER LA MEDIA (CAMPIONI DI GRANDI DIMENSIONI)

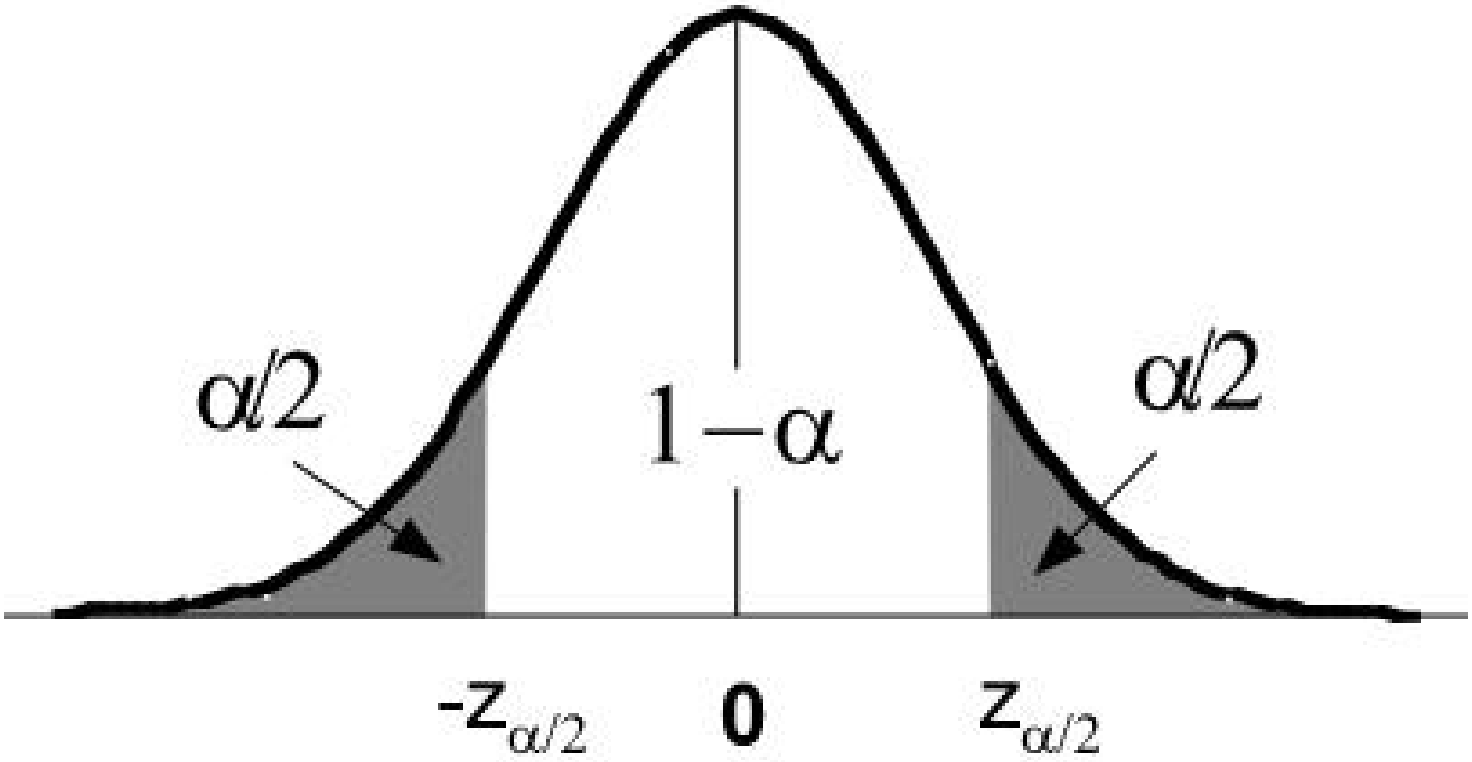
In base al ***teorema del limite centrale*** abbiamo stabilito che, per campioni di grandi dimensioni, la quantità

$$Z = \frac{\bar{Y} - m}{s/\sqrt{n}}$$

ha una distribuzione normale standardizzata.

Troviamo ora due valori tali per cui

$$P\left(-z_{a/2} \le Z \le z_{a/2}\right) = 1 - a$$


α/2
1−α
α/2
-z α/2
0
z α/2

$$P\left(-z_{a/2} \le \frac{\overline{X}-m}{\hat{s}_{\overline{X}}} \le z_{a/2}\right)=1-a$$

$$P\left(-z_{a/2}\hat{s}_{\overline{X}} \le \overline{X}-m \le z_{a/2}\hat{s}_{\overline{X}}\right)=1-a$$

$$P\left(-\overline{X}-z_{a/2}\hat{s}_{\overline{X}} \le -m \le -\overline{X}+z_{a/2}\hat{s}_{\overline{X}}\right)=1-a$$

$$P\left(\overline{X}+z_{a/2}\hat{s}_{\overline{X}} \ge m \ge \overline{X}-z_{a/2}\hat{s}_{\overline{X}}\right)=1-a$$

I due limiti dell'intervallo di fiducia del *100(1 - a)%* sono dunque uguali a:

$$\overline{X} + z_{a/2} \frac{s}{\sqrt{n}}$$

$$\overline{X} - z_{a/2} \frac{s}{\sqrt{n}}$$

Esempio. Il tempo necessario a 64 individui per completare il test XYZ è stato estratto a caso da un database che contiene i dati di tutti gli individui che si sono sottoposti al test. La media e la varianza di questo campione sono, rispettivamente, 33 minuti e 256.

Si trovi l'intervallo che contiene la vera media della popolazione con una probabilità di .90.

$$\hat{s}_{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{\sqrt{256}}{\sqrt{64}} = 2$$

$$z_{a/2} = z_{.05} = 1.645$$

$$\bar{Y} - z_{a/2}\hat{s}_{\bar{X}} = 33 - 1.645 \times 2 = 29.71$$

$$\bar{Y} + z_{a/2}\hat{s}_{\bar{X}} = 33 + 1.645 \times 2 = 36.29$$

Anche se non possiamo essere sicuri che l’intervallo calcolato (29.71, 36.29) contenga effettivamente la media della popolazione, possiamo però affermare che, se ripetessimo il processo di campionamento e calcolassimo l’intervallo di fiducia, gli intervalli così calcolati conterrebbero la vera media della popolazione nel 90% dei casi.

ESERCIZI

E3 Un gerontologo studio le abitudini alimentari delle donne con un'età superiore ai 70 anni. Il gerontologo ipotizza che, in questa fascia d'età, le abitudini alimentari delle donne siano mutate nel corso degli ultimi 50 anni.

I dati di uno studio di 50 anni fa indicano che la quantità media dei calorie assunte giornalmente dalle donne in questa fascia d'età era uguale a 2032 calorie.

Un campione di 100 donne con un'età maggiore o uguale a 70 anni viene scelto in maniera casuale e la quantità media di calorie assunte giornalmente da ciascuna di queste donne viene misurata. La media del campione risulta essere di 1847 calorie giornaliere.

Si trovi l'intervallo di fiducia del 95%.

E3 In un esperimento, 200 campioni casuali e indipendenti vengono estratti dalla medesima popolazione. Lo sperimentatore ipotizza che la media della popolazione sia uguale a 67.9. L'intervallo di fiducia del 90% viene calcolato per ciascuno dei 200 campioni. Esaminando i risultati, lo sperimentatore si rende conto che alcuni di questi intervalli di fiducia non coprono il valore di 67.9. Se la media della popolazione fosse effettivamente uguale a 67.9, quanti di questi intervalli "spuri" (ovvero, che non coprono la vera media della popolazione) ci si dovrebbe attendere di trovare?

E4 Un campione casuale di 3000 dichiarazioni dei redditi viene controllata. Viene contato il numero di esenzioni per ciascuna dichiarazione. La media per questo campione risulta essere di 3.78 con una deviazione standard di .97.

Si calcoli l'intervallo di fiducia del 99% relativo al numero di esenzioni per dichiarazione nella popolazione.

INTERVALLO DI FIDUCIA PER LA MEDIA (CAMPIONI DI PICCOLE DIMENSIONI)

Supponiamo di disporre di un campione casuale di piccole dimensioni ($n < 30$) con media $\bar{Y}$ e varianza s^2, tratto da una popolazione normale con media m e varianza s^2.

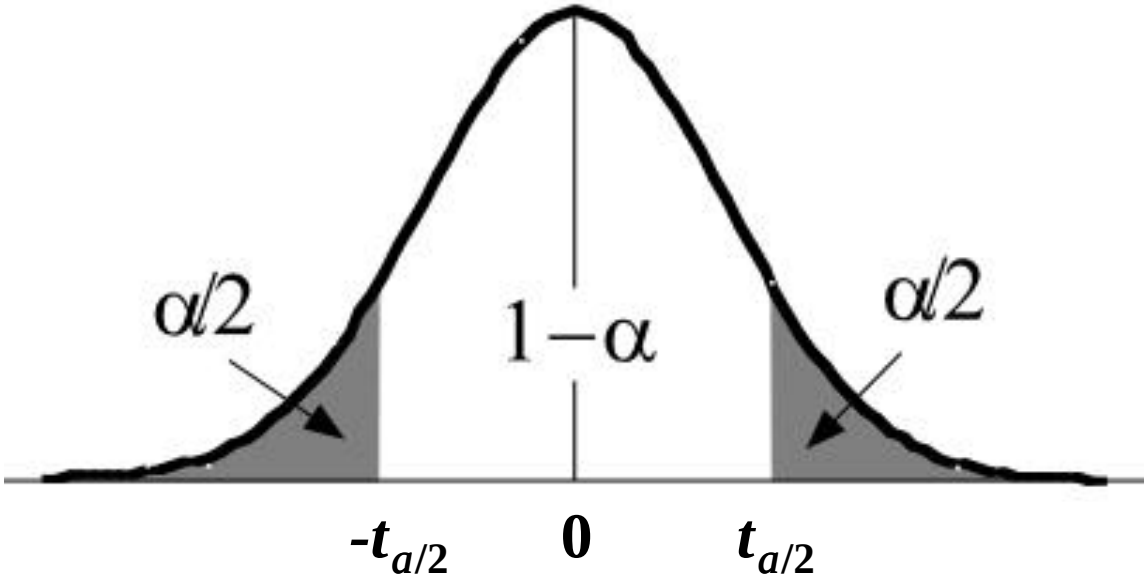
In precedenza abbiamo visto che la quantità

$$T = \frac{\bar{Y} - m}{s/\sqrt{n}}$$

segue la distrbuzione t di Student con (n - 1) gradi di libertà.

Dalle tabelle possiamo trovare i valori $-t_{a/2}$ e $t_{a/2}$ tali per cui

$$P\left(-t_{a/2} \le T \le t_{a/2}\right) = 1 - a$$



In maniera equivalente a ciò che abbiamo fatto in precedenza, i due limiti dell'intervallo di fiducia saranno:

$$\overline{X} + t_{a/2} \frac{s}{\sqrt{n}}$$

$$\overline{X} - t_{a/2} \frac{s}{\sqrt{n}}$$

Esercizio. Da una popolazione normale viene estratto un campione di $n = 8$ osservazioni, con media 2959 e deviazione standard (senza errore sistematico) uguale a 39.1.

Si calcoli l'intervallo di fiducia per la media della popolazione con un coefficiente di confidenza di .95.

Dalle tavole ricaviamo $t_{\alpha/2} = t_{.025} = 2.365$.

L'intervallo di fiducia del 95% sarà:

$$2959 \pm 2.365\ (39.1/\mathrm{Sqrt}(8)) = 2959 \pm 32.7$$

INTERVALLO DI FIDUCIA PER LA VARIANZA

Supponiamo di disporre di un campione di n osservazioni tratto da una popolazione normale con media μ e varianza σ^2.

In precedenza abbiamo notato che la quantità

$$(n-1)\frac{s^2}{\sigma^2}$$

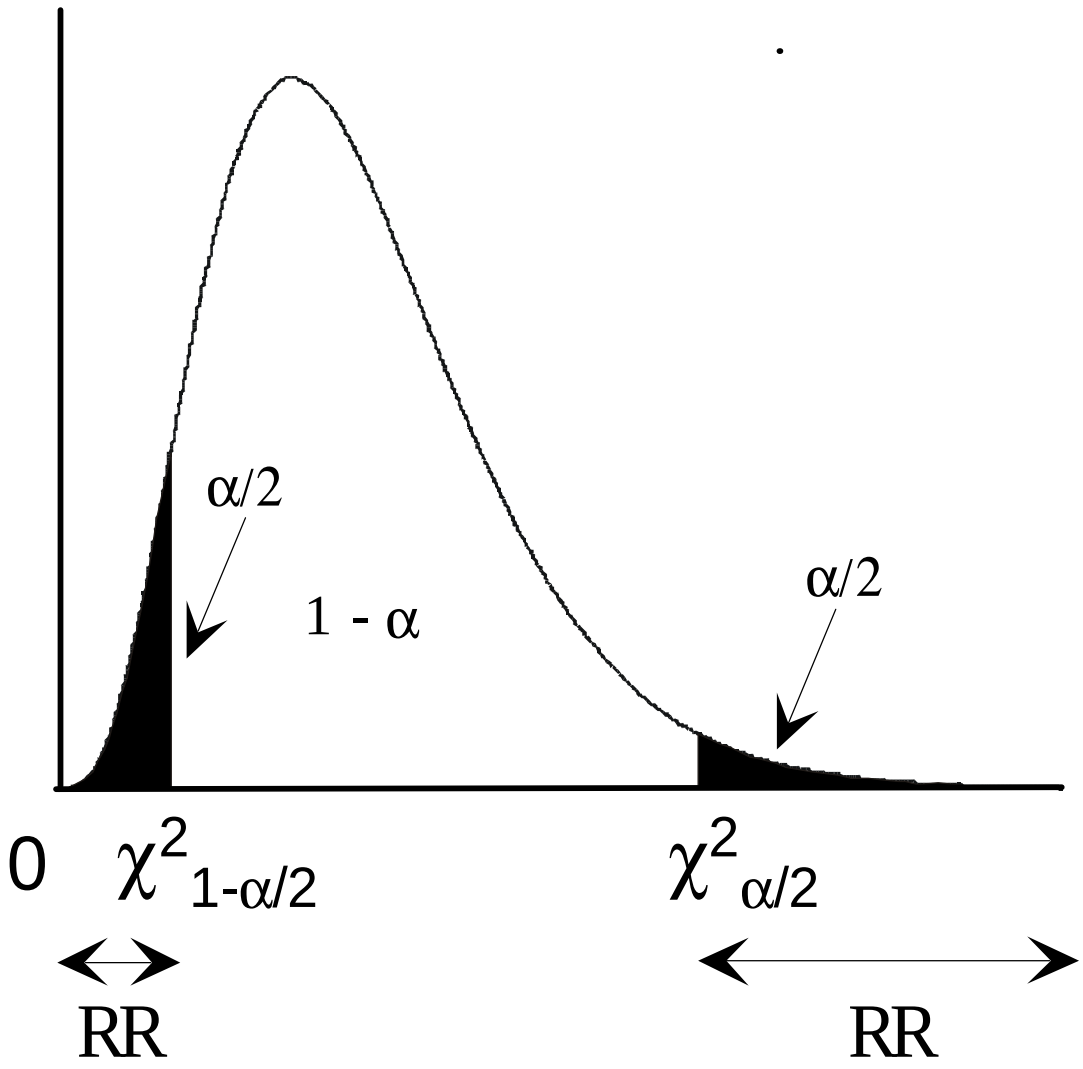
è distribuita come χ^2 con $(n - 1)$ gradi di libertà

Come in precedenza:

$$P\left[c_I^2 \le \frac{(n-1)s^2}{s^2} \le c_S^2 \right]$$

Da cui deriva l'intervallo di confidenza per σ^2 di $100(1 - \alpha)$:

$$\left(\frac{(n-1)s^2}{c_{a/2}^2}, \frac{(n-1)s^2}{c_{-a/2}^2} \right)$$


α/2
1 - α
α/2
0
$\chi^2_{1-\alpha/2}$
$\chi^2_{\alpha/2}$
RR
RR

Esercizio. Uno sperimentatore vuole stabilire la varianza delle misure ottenute con uno strumento per la rilevazione del volume sonoro di una data fonte. Tre misure vengono ottenute: 4.1, 5.2, 10.2.

Si stimi la varianza della popolazione σ^2 con un coefficiente di confidenza di .90.

Per i dati presenti, $s^2 = 10.57$. Se possiamo assumere la normalità della popolazione, allora

$$\left(\frac{(n-1)s^2}{c^2_{.05}}, \frac{(n-1)s^2}{c^2_{.95}} \right)$$

$$\left(\frac{(2)10.57}{5.991}, \frac{(2)10.57}{.103} \right)$$

$$(3.53, \quad 205.24)$$

Si noti come l’intervallo di fiducia sia molto grande, il che è dovuto, in primo luogo, al fatto che *n* è piccolo.

INTERVALLO DI FIDUCIA PER UNA PROPORZIONE

Una proporzione non è altro che il rapporto tra il numero di "successi" in *n* prove bernoulliane e il numero delle prove.

In precedenza, discutendo della distribuzione binomiale abbiamo definito il numero di successi in *n* prove bernoulliane come la somma dei valori assunti da *n* variabili che possono assumere soltanto i valori 0 oppure 1.

Abbiamo trovato il valore atteso e la varianza di S_n, (numero di successi in *n* prove bernoulliane).

$$E(S_n) = np$$

$$V(S_n) = npq$$

Il problema che ci poniamo ora è di trovare il valore atteso e la varianza di una proporzione, ovvero della variabile S_n (come è stata definita in precedenza) divisa per il numero n di prove.

$$E(\hat{p}) = E\left(\frac{S_n}{n}\right) = \frac{1}{n}E(S_n) = \frac{1}{n}np = p$$

$$V(\hat{p}) = V\left(\frac{S_n}{n}\right) = \left(\frac{1}{n}\right)^2 V(S_n) = \frac{1}{n^2}npq = \frac{pq}{n}$$

Per ciò che riguarda la forma della distribuzione campionaria di una proporzione, ci limiteremo al caso di campioni di grandi dimensioni.

Nel caso di $n > 100$, in base al teorema del limite centrale possiamo dire che la variabile aleatoria $\hat{p}$ si distribuisce in maniera approssimativamente normale.

La variabile z seguirà dunque la distribuzione normale standardizzata:

$$z = \frac{\hat{p} - p}{\sqrt{pq/n}}$$

In maniera equivalente a ciò che abbiamo fatto in precedenza,
i due limiti dell'intervallo di confidenza saranno:

$$\hat{p} \pm z_{a/2}\sqrt{\frac{pq}{n}}$$

Si noti che nella formula precedente l'intervallo di fiducia
è espresso nei termini dei parametri sconosciuti della popolazione
p, q.

Qualora questi parametri vengano stimati a partire dai dati del campione, l'intervallo di fiducia con un coefficiente di confidenza di 1 - *a* diventa:

$$\hat{p} \pm z_{a/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Esercizio. Un campione casuale di 200 persone viene intervistato. A ciascuno individuo viene posta una domanda a cui si può rispondere affermativamente o negativamente. In questo campione, il 58% degli intervistati risponde affermativamente alla domanda considerata.

Si costruisca l'intervallo di fiducia con un coefficiente di confidenza di .95 per la proporzione di individui che risponderebbero affermativamente alla domanda nella popolazione.

$$\hat{p} \pm z_{a/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

$$.58 \pm 1.96\sqrt{\frac{.58 \times .42}{200}}$$

$$(0.5116, \quad 0.6484)$$

E5 Su 2500 famiglie campionate casualmente, 499 hanno fornito una risposta positiva al quesito proposto dall'intervistatore.

Si costruisca l'intervallo di fiducia del 95%.

E6 In passato un referendum è stato bocciato con il 54% di voti contrari. I proponenti del referendum raccolgono un campione casuale di 1000 potenziali votanti e trovano che il 51% di questi è favorevole alla proposta referendaria.

Si trovi l'intervallo di fiducia del 99% della proporzione di votanti nella popolazione che sono favorevoli al referendum. Che cosa suggerisce questo risultato?

E7 Supponiamo di disporre di un campione con grandezza $n = 600$. La popolazione da cui il campione è estratto ha scarto quadratico medio uguale a 20. La media del campione è 124.

(a) Quali sono i limiti di fiducia corrispondenti a 3 errori standard?

(b) Quale è la probabilità che la vera media della popolazione abbia un valore compreso in questo intervallo in base alla diseguaglianza di Cebicev?

(c) Se assumiamo che la distribuzione campionaria della media sia normale, quale è la probabilità che la media sia compresa all'interno dell'intervallo di confidenza di 3 errori standard?

E8 Se non è possibile fare alcuna assunzione a proposito della distribuzione della popolazione, quale è la probabilità massima di osservare un caso che si scosti più di 1.7 deviazioni standard dalla media?