CAPITOLO 10

INFERENZA STATISTICA

Fare inferenza significa risalire dal particolare al generale (dal latino in-fero, portare dentro). Si tratta, quindi, di estendere il risultato delle considerazioni statistiche da una piccola parte alla generalità dei casi, cioè a quello che, con terminologia statistica, definiamo *universo*.

Definiamo:

1. *statistica campionaria* = valore caratteristico del campione, cioè una misura statistica ad esso riferita e su di esso calcolata. Generalmente, la statistica campionaria si indica con la lettera S e i valori che può assumere con le lettere latine minuscole;
2. *parametri dell'universo* = valori caratteristici della popolazione, misure statistiche ad essa riferite, non note e che si intende stimare.

$$S\begin{Bmatrix}\overline{x}\\ s\\ s^2\\ b\\ r\\ p\end{Bmatrix} term.\,noti \quad \rightarrow \quad \hat{\Theta} \quad \rightarrow \quad \Theta\begin{Bmatrix}\mu\\ \sigma\\ \sigma^2\\ \beta\\ \rho\\ \pi\end{Bmatrix} term.\,ignoti$$

Lo scopo della Inferenza è quello di valutare attraverso le statistiche campionarie (S) il valore dei parametri incogniti Θ della popolazione.

Utilizzando le statistiche campionarie (valori noti perché calcolabili direttamente sul campione) ed opportune metodologie, potremo, quindi, *stimare* i parametri incogniti della popolazione.

L'inferenza si avvale di due metodologie:

1. stima dei parametri;

2. verifica delle ipotesi (rientra nella c.d. teoria delle decisioni).

Statistica campionaria	S
Parametri	Θ

Tra *S* e Θ vi é la funzione $\hat{\Theta}$, detta **stimatore**.

$\hat{\Theta}$ stimatore di Θ è una funzione delle $X_1, X_2, ..., X_n$ osservazioni campionarie le quali sono variabili casuali indipendenti, identicamente distribuite (in statistica spesso si legge l'acronimo i.i.d.), con stessa distribuzione della popolazione e, quindi, con stessa media e stessa varianza.

$\hat{\Theta}$ è una variabile casuale del tipo $\hat{\Theta}$ = h ($X_1, X_2, ..., X_n$) ed il valore così calcolato (detto stima) è funzione delle determinazioni campionarie.

Attraverso lo stimatore è possibile configurare l'intero universo campionario di una prefissata dimensione *n* nonché l'insieme di tutte le statistiche campionarie potenzialmente calcolabili sull'universo campionario configurato. E', quindi, attraverso l'operazione di stima che è possibile ipotizzare il calcolo delle statistiche campionarie.

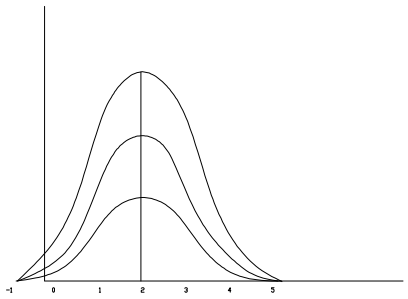
La successione delle *n* statistiche campionarie calcolate, organizzata in distribuzione di frequenza origina la *distribuzione campionaria della statistica S.*

10.1. Proprietà degli stimatori

Per stimare un certo parametro, esistono in Statistica più stimatori. E', quindi, opportuno introdurre dei metodi che consentono di confrontare tra di loro gli stimatori o che consentano di stabilire se un dato stimatore possiede talune proprietà considerate, in genere, desiderabili. Tali metodi prendono il nome di proprietà. Vediamone nel seguito alcune.

1. *Centratura*: uno stimatore si dice centrato (non distorto, non tendenzioso) quando: $E(\hat{\Theta}_n) = \Theta$. La proprietà è verificata per
   a. la media: $E(\bar{x}) = \mu$
   b. la mediana : $E(Me) = Me$
   c. la moda: $E(Mo) = Mo$
   d. non vale per la varianza, infatti abbiamo che $E(s^2) = \sigma^2 \rightarrow$ non vero

2. *Efficienza*: dati due o più stimatori $\Theta_1; \Theta_2; \Theta_3$ tutti egualmente centrati, si preferisce lo stimatore con la varianza minore.



Le ascisse dei punti di flesso ci danno lo scarto quadratico medio(s.q.m. opp σ).

A parità di centratura scelgo il parametro con minore varianza.

3. *Sufficienza*: Data una variabile casuale, cui si associa una famiglia di distribuzioni di probabilità parametrizzate tramite il vettore θ, e una statistica T(.), T(x) è *sufficiente per* θ se la distribuzione di probabilità

della X data T(X) non dipende da θ. L'idea di fondo è che uno stimatore possa dirsi sufficiente quando racchiude ed esaurisce tutte le informazioni riguardanti il parametro incognito e contenute nel campione casuale. Sia X=(X1, X2,…,Xn) un campione casuale generato dalla v.c. X che segue una distribuzione del tipo f(x; θ) dove θ appartiene ad Ω(θ) è il parametro oggetto di stima. Diremo che Tn è uno stimatore sufficiente per θ

$$T_n \text{ è sufficiente per } \theta \Leftrightarrow \varphi_X\left(x_1, x_2, ..., x_n \middle| T_n = t_0\right) \text{ non dipende da } \theta$$

4. *Consistenza*: uno stimatore è consistente se:

$$\lim_{n \to +\infty} \Pr\left\{\left|\hat{\Theta}_n - \Theta\right| < \varepsilon\right\} = 1$$

$\hat{\Theta}_n$ = valore dello stimatore, dipendente da n;

Θ = parametro;

n = dimensione campionaria;

ε = prefissato piccolo.

Esprime una legge di <u>*convergenza in probabilità*</u>, non in valore (facciamo inferenza, non lavoriamo, quindi, con fatti certi). Se si fa riferimento alla media campionaria, l'insieme di tutte le medie calcolate su tutti i possibili campioni, di dimensione *n* appartenenti ad un certo universo campionario origina la *distribuzione campionaria della media.*

Il concetto, con i dovuti accorgimenti di cui si dirà in seguito, è ripetibile per tutte le misure statistiche calcolabili sui campioni.

Essa comprende tutti i possibili valori della statistica nell'universo campionario prescelto. E' una distribuzione particolare da non confondere con la distribuzione di un carattere che si studia nella popolazione.

Posto $\bar{x}$ media del campione con $\bar{x}_1, \bar{x}_2 ....., \bar{x}_n$, possiamo pensare di strutturare la successione delle medie in forma di distribuzione di frequenze, ottenendo la *distribuzione campionaria della media.*

Dimostriamo che $E(\bar{x}) = \mu$ il valore atteso della V.C. distribuzione campionaria della media è uguale al valore del parametro μ :

$$E(\bar{x}) = E\left[\frac{1}{n}(X_1 + X_2 + X_i + ...X_n)\right]$$

(il valore atteso di una somma = somma valori attesi)

$$E(\bar{x}) = \frac{1}{n}\left[EX_1 + EX_2 + EX_i + ...EX_n\right]$$

$$E(\bar{x}) = \frac{1}{n}\left[\mu + \mu + \mu + ...\right] \Rightarrow E(\bar{x}) = n\mu \cdot \frac{1}{n} = \mu$$

Dimostriamo ora che:

1. $Var_{\bar{x}} = \frac{\sigma^2}{n}$ La varianza della *distribuzione campionaria della* media è uguale alla vacanza della popolazione divisa per la numerosità campionaria *n*
2. Var $(\bar{x}) =$ var $\left[\frac{1}{n}(X_1 + X_2 + ...X_n)\right]$
3. Var $(\bar{x}) = \frac{1}{n^2}\left[\text{var}\, X_1 + \text{var}\, X_2 + ... \text{var}\, X_n\right] = = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$

La media della popolazione è un valore sul quale è centrata la distribuzione campionaria della media.

Nel caso di campioni di piccole dimensioni e/o con la varianza della popolazione, σ^2, non nota la media campionaria si distribuisce secondo una T di Student con n-1 gradi di libertà.

La distribuzione T di Student è:

- ✓ definita positiva,

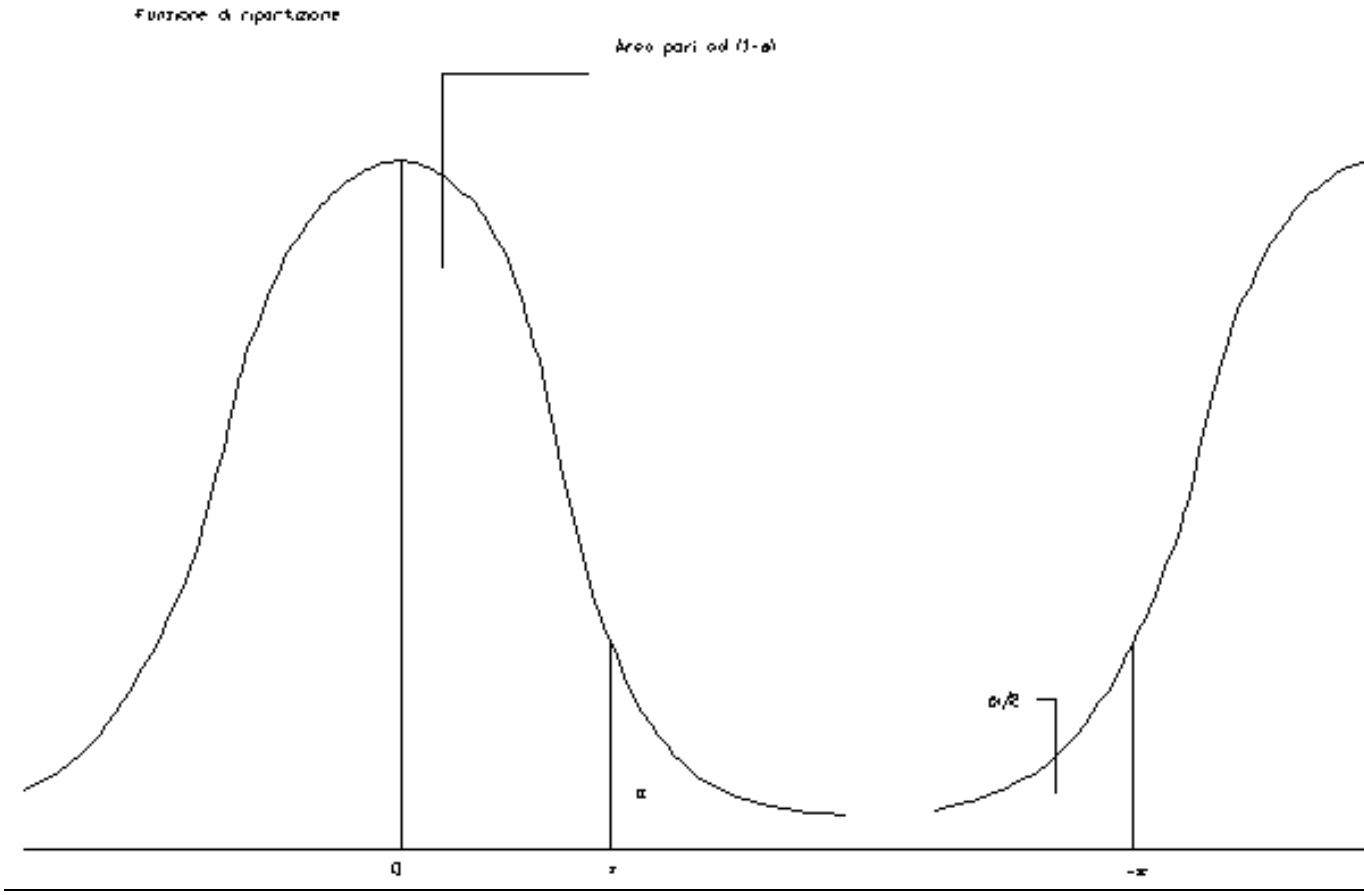
- ✓ dipende da n (numerosità campionaria) per cui esistono infinite curve tutte simmetriche ed asintotiche, più piatte della Normale (sono affette da curtosi – sono platicurtiche o iponormali),
- ✓ per n che tende ad infinito la distribuzione T tende alla normale

In questo caso la statistica di riferimento sarà

$$t = \frac{\bar{x} - \mu}{\hat{\sigma}_{\bar{x}}}$$

La differenza con la statistica z consiste nel denominatore: nella z al denominatore troviamo l'errore standard della media calcolato attraverso l'utilizzo dello scostamento quadratico medio della popolazione (*la varianza della popolazione è nota)* nella statistica t al denominatore l'errore standard della media è calcolato utilizzando la stima dello scostamento quadratico medio della popolazione (*la varianza della popolazione non è nota)*.

Lo scostamento quadratico medio della popolazione (e quindi anche la varianza) è stimato utilizzando la statistica campionaria *scostamento quadratico medio campionario corretto. Graficamente:*


Funzione di ripartizione
Area pari ad (1-α)
α/2
α
0
z
-z

11.2. Teorema del Limite Centrale.

Con questo nome vengono indicati un gruppo di teoremi che risultano indispensabili per la teoria delle distribuzioni, necessaria allo sviluppo della statistica inferenziale.

Questi teoremi costituiscono in pratica un modo per "quantificare la legge dei grandi numeri".

Ricordiamo brevemente le due formulazioni fondamentali di tale legge

1. per prove ripetute indipendenti, dove il risultato di ciascuna prova può essere classificato come successo o insuccesso, si può affermare che *(teorema di Bernulli),* ***al crescere del numero delle prove, la frequenza relativa dei successi converge alla probabilità di successo di una prova.***
2. per prove ripetute indipendenti in cui il risultato di ciascuna prova è il valore x di una variabile aleatoria X (ad esempio, una misura di lunghezza, peso, durata) si può asserire che (*teorema di Cebicev),* ***per un numero sufficientemente grande di prove indipendenti, la media aritmetica dei valori osservati di una variabile aleatoria converge in probabilità alla sua speranza matematica.***

Tutte le formulazioni della legge dei grandi numeri stabiliscono che i risultati delle singole prove influiscono poco sul risultato medio di un numero elevato di prove: le deviazioni dalla media, inevitabili in una singola prova, si livellano reciprocamente quando il numero delle prove è elevato.

Questo significa che quando il numero di prove, E, è elevato, il risultato medio diventa stabile e, quindi, può essere previsto.

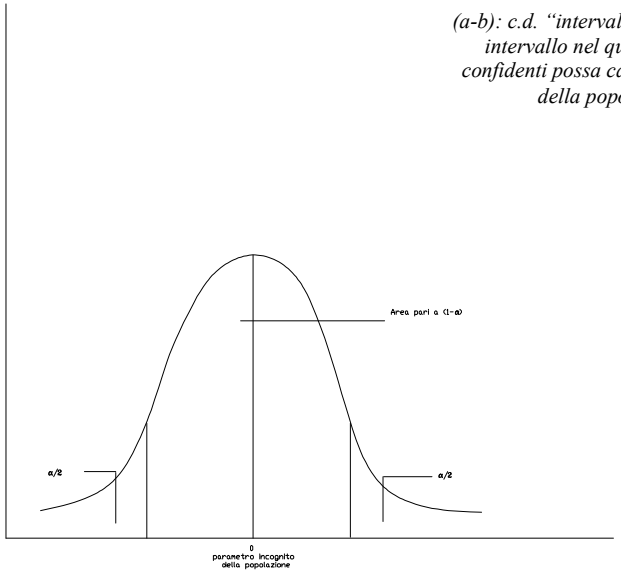
La possibilità di effettuare tali previsioni sono rese maggiori dal *teorema del limite centrale* che stabilisce quale distribuzione segue la somma di un numero sufficientemente grande di variabili aleatorie.

Tale teorema, detto "centrale" proprio per la sua importanza, permette di definire delle ipotesi e di stimare la loro probabilità di verificarsi.

La formulazione più generale del *teorema del limite centrale* è la seguente: *sia* S_n *una variabile aleatoria somma di n variabili aleatorie indipendenti* X_i *aventi ciascuna la stessa distribuzione di probabilità, speranza matematica* μ *e varianza al crescere di n* S_n *tende ad assumere una distribuzione normale con media* $n\mu$ *e*

$$S_n = X_1 + X_2 + \cdots\cdots + X_n$$

La variabile $Z = \dfrac{S_n - n\mu}{\sqrt{n\sigma}}$ *è la corrispondente variabile normale standardizzata.* Graficamente,



Il teorema del limite centrale risulta valido per n sufficientemente grande, qualunque sia la distribuzione della variabile X_i; inoltre, è generalizzabile al caso di variabile aleatoria ***con distribuzione di probabilità qualsiasi***, alla sola condizione che ciascuna di esse abbia media e varianza finite e non risulti predominante rispetto alle altre.

Se le variabili X_i hanno distribuzione normale con media μ e varianza σ allora la variabile S_n ha sempre distribuzione normale, qualunque sia il valore di n.

Nel caso si debba stimare σ^2 non noto, in presenza di un campionamento senza ripetizione si userà, partendo dai dati campionari il solito criterio della varianza campionaria corretta (stimatore corretto di σ^2) tenendo presente il fattore di correzione per la varianza stimata su dati campionari per campioni estratti senza ripetizione:

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

Va considerato che poi quando si trasporterà tale stima nella espressione:

$$\hat{\sigma}^2_{\bar{x}} = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$$

si avrà:

$$\hat{\sigma}^2_{\bar{x}} = s^2 \cdot \frac{n}{n-1} \cdot \frac{N-1}{N} \cdot \frac{N-n}{N-1} \cdot \frac{1}{n}$$

che per effetto delle semplificazioni diventa:

$$\hat{\sigma}^2_{\bar{x}} = \frac{s^2}{n-1} \cdot \frac{N-n}{N}$$

consentendo direttamente la determinazione dell'errore standard della D.C. della media per campionamento senza ripetizione.

Stima:

- ✓ <u>puntuale</u>: si stima mediante la statistica di un campione $\hat{\Theta} = \Theta$
- ✓ <u>intervallo di fiducia</u>: intervallo entro il quale cade Θ

La <u>stima puntuale</u> non ci da la possibilità di conoscere la probabilità di errore.

La stima puntuale di $\hat{\Theta} = \Theta$ presenta un errore dovuto al fatto che si rileva un solo campione e non tutta la popolazione. Tale errore non è valutabile. Si preferisce determinare in base alle osservazioni campionarie un intervallo (a-b), appartenente allo spazio di Θ, entro cui cade $\hat{\Theta}$ con una certa prefissata probabilità.

Nell'intervallo di confidenza cade il parametro incognito della popolazione

- (a, b) = limiti di confidenza o di fiducia;
- (a – b)= intervallo di confidenza = d.

L'errore è l'area esterna ad a e b

- α = probabilità che Θ sia esterno ad (a, b)
- $1-\alpha =$ probabilità che Θ sia interno al (a, b) - Probabilità $a\langle\Theta\langle b$
- $1-\alpha =$ livello di confidenza

Poiché Θ può essere esterno sia per eccesso che per difetto, la probabilità di errore α va ripartita sulle due code della distribuzione dello stimatore.

Se α è la probabilità che Θ sia esterno ad (a-b) possiamo affermare che α/2 è la probabilità che Θ sia o minore di a o maggiore di b.

$$\Pr(\Theta \le \alpha) = \alpha/2 \qquad \Pr(\Theta \ge b) = \alpha/2$$

$(1-\alpha)$= livello di confidenza indica la probabilità che il parametro incognito sia interno ad (a – b).

$$\Pr(1-\alpha) = \Pr \alpha < \Theta < b$$

$P = (1-\alpha)$ deve essere alta ma non può tendere ad 1 perché altrimenti l'intervallo (a, b) si dilata e non ha senso fare inferenza, parità di α, più è piccolo l'intervallo (a, b) più è precisa la stima.

$\Pr\{a \le \vartheta \le b\} = 1 - \alpha$

un tale intervallo si dice centrato.

Un intervallo di confidenza può essere costruito per qualsiasi parametro.

Volendo stimare μ la media incognita della popolazione, ricaviamo l'intervallo che contiene μ con $\Pr = 1 - \alpha$.

Ricordiamo che la distribuzione campionaria della media, ricorrendo le condizioni dettate dal *teorema del limite centrale,* segue una legge Normale e come tale standardizzabile, quindi la statistica di riferimento sarà:

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

avendo fissato il rischio di errore α si tratterà di considerare tutti i valori compresi tra:

$$-z_{\frac{\alpha}{2}} \le \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \le z_{\frac{\alpha}{2}}$$

Fissando α = 0.05 e ricordando che la stima può essere errata sia per difetto che per eccesso, bisogna individuare quei valori di $z_{\alpha/2}$ che delimitano l'intervallo di confidenza. Essendo l'area sottesa alla curva che descrive l'andamento dello stimatore, media campionaria, uguale ad 1, l'area interna all'intervallo di confidenza (1 – α) sarà uguale a 1-0.05= 0.95. Tenuto conto che la probabilità di errore (α) deve essere ripartita sulle due code e che la distribuzione è simmetrica, il valore 0.95 va diviso per 2 . 0.95 : 2= 0.475.

Ipotizzando di utilizzare una distribuzione Normale alla Probabilità (area) pari a 0.475 si associa il valore ± 1,96.

Moltiplicando per $\sigma(\bar{x})$ tutti i membri della disuguaglianza

$$-z_{\frac{\alpha}{2}} \cdot \sigma(\bar{x}) \leq \mu - \bar{x} \leq z_{\frac{\alpha}{2}} \cdot \sigma(\bar{x})$$

e aggiungendo a destra e a sinistra la media campionaria

$$\Rightarrow \bar{x} - z_{\frac{\alpha}{2}} \cdot \sigma(\bar{x}) \leq \mu \leq \bar{x} + z_{\frac{\alpha}{2}} \cdot \sigma(\bar{x})$$

Riferendosi ad un generico parametro Θ le quantità $-z_{\frac{\alpha}{2}} \cdot \sigma(\bar{x})$ e $z_{\frac{\alpha}{2}} \cdot \sigma(\bar{x})$ vengono indicate con δ e δ' . Se δ = δ' la distribuzione dello stimatore è simmetrica.

Notazione: $\hat{\Theta} - \delta \leq \Theta \leq \hat{\Theta} + \delta'$.

La distribuzione campionaria della media è normale per campioni grandi e si presenta con le seguenti caratteristiche

$$E(\bar{x}) = \mu \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \qquad \sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$

- μ = media della popolazione. Valore su cui è centrata la distribuzione.
- σ_x = errore standard della stima (= scostamento quadratico medio della distribuzione campionaria della media).

Dalla normalità della distribuzione deriva che essa è standardizzabile per cui

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

Nel caso si verifichi che:

1. la dimensione campionaria sia sufficientemente grande *n > 30;*
2. sia nota la varianza della popolazione;
3. il campionamento sia avvenuto con ripetizione,

La costruzione dell'intervallo di confidenza è di immediata e facile soluzione.

Se permangono le condizioni 1 e 2 ed il campionamento avviene senza ripetizione o in blocco, nella costruzione dell'intervallo di confidenza bisogna tenere conto di tale circostanza.

La standardizzata sarà sempre:

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

però, nel calcolo dell'errore standard (il denominatore) si dovrà considerare il fatto che il campionamento sia stato effettuato senza ripetizione o in blocco. Pertanto la distribuzione avrà i seguenti valori:

$$E(\bar{x}) = \mu \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \qquad \sqrt{\frac{N-n}{N-1}} \qquad \sigma^2_{\bar{x}} = \frac{\sigma^2}{n} \quad \left(\frac{N-n}{N-1}\right)$$

In presenza di piccoli campioni e/o di mancata conoscenza della varianza della popolazione, la distribuzione campionaria della media segue una legge T di Student con n – 1 gradi di libertà.

I valori caratteristici sono

$$E(\bar{x}) = \mu \quad \hat{\sigma}^2_{\bar{x}} = \frac{\bar{s}}{\sqrt{n}} \Rightarrow in\ cui\ \bar{s} = s\sqrt{\frac{n}{n-1}}\ da\ cui\ \frac{1}{\sqrt{n}} \cdot s\sqrt{\frac{n}{n-1}} = \hat{\sigma}_{\bar{x}} = \frac{s}{\sqrt{n-1}}$$

Si ricorda che *s è lo scostamento quadratico medio campionario non corretto*

$$\sqrt{\frac{\sum (x-\mu)^2}{n}}$$

che per essere uno stimatore centrato di σ deve essere corretto

$$s\sqrt{\frac{n}{n-1}}$$

Se il campionamento avviene senza ripetizione o in blocco i valori caratteristici della distribuzione campionaria della media sono:

$$E(\bar{x}) = \mu \qquad \hat{\sigma}_{\bar{x}} = \frac{\bar{s}}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \quad \Rightarrow \qquad \text{con } \bar{s} = s\sqrt{\frac{n}{n-1}\times\frac{N-1}{N}}$$

dopo alcune semplificazioni avremo:

$$\frac{\bar{s}}{\sqrt{n}} = \sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{1}{n}\cdot\frac{n}{n-1}\cdot\frac{N-1}{N}\cdot\frac{N-n}{N-1}\cdot S^2} \qquad \hat{\sigma}_{\bar{x}} = s\sqrt{\frac{1}{n-1}\frac{N-n}{N}}$$

Se non si hanno notizie sulla popolazione, non è possibile provarne la normalità e si è in presenza di piccoli campioni, con varianza della popolazione incognita, l'intervallo di confidenza si può costruire ricorrendo alla disuguaglianza di Chebicheff:

$$\text{disuguaglianza di Chebicheff} \rightarrow x < \begin{matrix}\text{popolazione non normale} \\ n < 30\end{matrix} \qquad k = \frac{1}{\sqrt{\alpha}}$$

$$\bar{x} - k\hat{\sigma}_{\bar{x}} \le \mu \le \bar{x} + k\hat{\sigma}_{\bar{x}}$$

11.3. Alcuni esempi

Esempio. Supponiamo di avere a disposizione i dati di un campionamento di tipo bernoulliano con reinserimento

$$E(\bar{x}) = \mu$$

$$\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}$$

Popolazione:

20	21	22	23	24

Dimensione della popolazione (N) :5.

La dimensione dell'universo campionario è, quindi, pari a $2 = 5^2 = 25$. determiniamo, quindi, la distribuzione dei voti medi calcolati su 25 campioni di dimensione 2.

Voti medi	Frequenze
20	1
20,5	2
21	3
21,5	4
22	5
22,5	4
23	3
23,5	2
24	1

Determiniamo, ora, la media e la varianza della nostra distribuzione.

- ✓ $\mu_x = \frac{(20*1)+(20{,}5*3)+\ldots+(24*1)}{25} = 1;$

- ✓ $\sigma^2(x) = \frac{\left[(20-22)^2 * 1\right] + \left[(20{,}5-22)^2 * 2\right] + \ldots + \left[(24-22)^2 * 1\right]}{25} = 2$

Esempio 2: Campionamento in blocco

I dati del problema a nostra disposizione.

- ✓ $E(\bar{x}) = \mu$;
- ✓ $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} * \frac{N-n}{N-1}$;
- ✓ popolazione: 20 – 21 – 22 – 23 – 24
- ✓ dimensione della popolazione: 5.

Svolgendo i calcoli, avremo che:

- ✓ la media è pari a 22
- ✓ la varianza è pari a 2

Ne consegue che l'universo campionario di dimensione 2 = N * (N *- 1) = 20

Esempio 3: determinare la distribuzione dei voti medi calcolati su 20 campioni di dimensione 2.

Voti medi (X)	Frequenze	Prodotti	$(X)^2$	$Fr*(X)^2$
20,5	2	41	420,25	840,5
21	2	42	441	882
21,5	4	86	462,25	1849
22	4	88	484	1936
22,5	4	90	506,25	2025
23	2	46	529	1058
23,5	2	47	552,25	1104,5
Totale	**20**		**3395**	**9695**

media	22
media ^ 2	484
media dei quadrati	484,75
varianza	0,75

<u>Esempio 3</u>: determinare la distribuzione dei voti medi calcolati su 20 campioni di dimensione 2 (*<u>con varianza incognita</u>*).

Le condizioni sulla media non cambiano perché la media campionaria è uno stimatore corretto della media della popolazione → $E(\bar{x}) = \mu$**.**

Ci interessa, quindi, capire se anche la varianza campionaria sia uno stimatore corretto della varianza della popolazione e, cioè, se vale la condizione $E(s^2) = \sigma^2$.

Camp	x	s^2	**σ**	Camp
20-20	20	0	0	23-20
20-21	21	0,25	1	23-21
20-22	21	1	2	23-22
20-23	22	2,25	5	23-23
20-24	22	4	8	23-24
21-20	21	0,25	1	24-20
21-21	21	0	0	24-21
21-22	22	0,25	1	24-22
21-23	22	1	2	24-23
21-24	23	2,25	5	24-24
22-20	21	1	2	
22-21	22	0,25	1	
22-22	22	0	0	
22-23	23	0,25	1	<u>13,75+11,25</u> = 1
22-24	23	1	2	25
		13,8		

Calcoliamo il Valore Atteso della D.C. della varianza campionaria. Costruiamo, quindi, la distribuzione di s^2.

s^2	Fr.
0	5
0,25	8
1	6
2,25	4
4	2
	25

$$s^2 = \frac{0*5+0{,}25*8+1*6+2{,}25*4+4*2}{25} = 1$$

Costruiamo la distribuzione di σ^2:

σ^2	**Fr**	**σ^2 * Fr**
0	5	0
0,5	8	4
2	6	12
4,5	4	18
8	2	16
	25	**50**

$$s^2 = \frac{0*5+0{,}5*8+2*6+4{,}5*4+8*2}{25} = \frac{50}{25} = 2$$

Campionamento in blocco

Camp	X	s^2	**σ**	Camp	x	s^2	**σ**
20-21	20,5	0,25	0,4	23-20	21,5	2,25	3,6
20-22	21	1	1,6	23-21	22	1	1,6
20-23	21,5	2,25	3,6	23-22	22,5	0,25	0,4
20-24	22	4	6,4	23-24	23,5	0,25	0,4

21-20	20,5	0,25	0,4	24-20	22	4	6,4
21-22	21,5	0,25	0,4	24-21	22,5	2,25	3,6
21-23	22	1	1,6	24-22	23	1	1,6
21-24	22,5	2,25	3,6	24-23	23,5	0,25	0,4
22-20	21	1	1,6			**11,25**	**18**
22-21	21,5	0,25	0,4				
22-23	22,5	0,25	0,4		**E(s²) =**	**1,25**	
22-24	23	1	1,6				
		13,75					

Costruiamo la distribuzione di σ^2

0,4	8	3,2
1,6	6	9,6
3,6	4	14,4
6,4	2	12,8
	20	40

$$E(\sigma^2) = 40/20 = 2$$

11.4. Distribuzione campionaria delle frequenze relative.

Si consideri una popolazione con distribuzione binomiale e si dividano i suoi elementi in funzione del fatto che posseggano e non posseggano una determinata caratteristica (modalità). Si individui con il termine *successo* il possesso o la presenza della modalità indicata e con *insuccesso* l'assenza allora avremo:

$x = 1$ presenza della modalità (successo) probabilità: p

$x = 0$ assenza della modalità (insuccesso) probabilità: q= 1-p

In un campione di n elementi (*prove*) avremo media = *np* e s.q.m. = $\sqrt{npq}$

Si consideri la distribuzione campionaria delle frequenze dei successi (*distribuzione campionaria delle frequenze).*

In un campione di n elementi *f* rappresenta la frequenza dei successi: in generale *f* è la variabile aleatoria campionaria.

Allora, in base al teorema del limite centrale per popolazioni normali o per campioni $n \geq 30$ $n \to \infty$ la *distribuzione campionaria delle frequenze* tende ad una legge normale per $n \to \infty$

$$z = \frac{f - np}{\sqrt{npq}} \quad \text{(A)}$$

Dividendo il numeratore ed il denominatore per *n* di z si ottiene

$$z = \frac{fr - p}{\sqrt{\frac{pq}{n}}} \quad \text{(B)}$$

che rappresenta la *distribuzione campionaria delle frequenze relative*.

Le differenze tra le due posizioni (A e B) consistono in:

1. al numeratore di B compaiono le frequenze relative;
2. compare *p* al posto del numero medio di successi *np.*

I valori caratteristici della distribuzione (B) saranno:

$$Mp = p \qquad \sigma_p = \sqrt{\frac{pq}{n}} \qquad z = \frac{fr - p}{\sqrt{\frac{pq}{n}}}$$

Sarà possibile determinare:

Pr (- $z_{\alpha/2}$ $\leq z \geq$ $z_{\alpha/2}$) = *(1-α)*

da cui Pr (- $z_{\alpha/2} \leq \frac{fr - p}{\sqrt{\frac{pq}{n}}} \geq$ *zα/2) = (1-α)* da cui

per l'invertibilità di *z,* abbiamo*:*

$$fr - z_{\alpha/2} \bullet \sqrt{\frac{pq}{n}} \leq p \geq fr + z_{\alpha/2} \bullet \sqrt{\frac{pq}{n}}$$

Operando in tal modo, però, l'intervallo di confidenza contiene *p (parametro da stimare – simbolicamente indicato anche con π)*. Si può ovviare a tale inconveniente introducendo una stima puntuale di *p* ponendo

$$p = \frac{fr}{n} \qquad \text{e} \qquad q = 1 - \frac{fr}{n}$$

operando in tal modo avremo

$$fr - z_{\alpha/2} \cdot \sqrt{\frac{fr(1-fr)}{n}} \leq p \leq fr + z_{\alpha/2} \cdot \sqrt{\frac{fr(1-fr)}{n}}$$

Esempio. Un programma televisivo è stato seguito dal 23% di un campione di 300 spettatori selezionati a caso.

Costruire un intervallo di confidenza al 98% per la proporzione della popolazione del pubblico televisivo in quella fascia oraria.

Dati : n = 300 - proporzione campionaria = 0.23 - $\alpha = 0.02$

Determinazione di $z_{\alpha/2}$

. Si ammette un errore del 2% che può essere commesso per eccesso e per difetto. Il rischio d'errore, α, va ripartito sulle due code della distribuzione di riferimento. Avremo, quindi : 1 – 0.02 = 0.98 ; 0.98 : 2 = 0.49

Dalle tavole della distribuzione normale standardizzata rileviamo che al valore dell'area = 0.49 è associato il valore di z = 2.33

Avremo allora p = 0.23 q = (1 – p) = 0.77.

L'intervallo formale sarà:

$$p - z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}} \le p \le p + z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}}$$

$$0{,}23 - \left(2{,}33 \cdot \sqrt{\frac{0{,}23 \cdot (1-0{,}23)}{300}}\right) \le p \le 0{,}23 + \left(2{,}33 \cdot \sqrt{\frac{0{,}23 \cdot (1-0{,}23)}{300}}\right)$$

$$0{,}23 - (2{,}33 \cdot 0{,}024) \le p \le 0{,}23 + (2{,}33 \cdot 0{,}024)$$

$$0{,}23 - (2{,}33 \cdot 0{,}056) \le p \le 0{,}23 + (2{,}33 \cdot 0{,}056) = [0{,}174;\ 0{,}286]$$

Discutendo della stima della varianza di una popolazione e della statistica T di Student abbiamo introdotto il concetto di

11.5. I gradi di libertà

Si tratta di una nozione che occupa un posto importante nei problemi di inferenza statistica ed è, quindi, opportuno cercare di afferrarne il significato. Assumiamo, per esempio che *l'analisi di un materiale abbia portato i seguenti risultati relativi alla % in peso del componente M*

dati	**scarti**
70.5	*0.057*
7.00	*0.007*
7.05	*0.057*
7.00	*0.007*
6.85	*-0.143*
6.96	*-0.33*
6.94	*-0.053*
7.13	*0.137*
7.00	*0.007*
6.95	*-0.043*
Media: 6.993	*Σ scarti: 0*

I valori che compaiono nella prima colonna della tabella sono stati ottenuti analizzando porzioni di materiali rilevate secondo regole ben precise. Si tratta di valori estratti dalla popolazione con un campionamento casuale e che sono indipendenti tra di loro. Questo significa che non è possibile, conoscendo il primo valore, predire il secondo, o il terzo e così via. In generale, la conoscenza di un certo numero di dati non ci consente di avanzare alcuna ipotesi su quelli che seguono.

Il discorso è diverso se consideriamo gli scarti dalla media: la loro somma è zero (*prima proprietà della media aritmetica*).

Non disponiamo, tra gli scarti, di dieci valori indipendenti fra di loro, ma solo di nove; di conseguenza, nove sono i gradi di libertà della serie di scarti.

Come mai, passando dai singoli dati agli scarti dalla media si perde un grado di libertà?

In pratica è come, se tra i dieci dati a nostra disposizione, uno corrispondesse al valore vero del contenuto percentuale di M e gli altri nove riflettessero l'effetto di fattori aleatori di variazione sulle misure.

E' opportuno sottolineare che il numero dei gradi di libertà viene usato, in questo caso specifico, per stimare la varianza della popolazione. Perciò nei problemi di stima, quando si parla di "*numero dei gradi di libertà della serie di misure"* si deve correttamente intendere *"il numero dei gradi di libertà della serie di misure disponibili per la stima del parametro".* Abbiamo, infatti, visto che i gradi di libertà sono 10 se consideriamo le osservazioni e 9 se ci riferiamo agli scarti dalla media.

L'esigenza di contare unicamente i valori indipendenti fra di loro si presenta in molti problemi di inferenza statistica. E' vero, infatti, che la quantità di informazioni cresce al crescere del numero delle osservazioni, ma è altrettanto vero che se un osservazione non è indipendente dalle altre, l'informazione che essa fornisce è già contenuta nelle altre; è, quindi, logico non contarla tra gli elementi a disposizione per effettuare i calcoli.

Possiamo affermare, quindi, che *il numero dei gradi di libertà di un parametro statistico corrisponde al numero dei valori, indipendenti tra loro, usati per calcolare il parametro in questione.*

Naturalmente non sempre il numero dei gradi di libertà di una serie di osservazioni è dato dal numero delle osservazioni diminuito di uno.

A seconda del parametro che si deve stimare, il numero dei gradi di libertà può essere n-1 ; n – 2 ; n – 3 e così via.

In effetti, per potere fare inferenza sui parametri si deve avere a disposizione le osservazioni che costituiscono il campione , sia altri parametri. Se questi ultimi non sono noti (p.e. la varianza della popolazione) si ricorre alle loro stime, che si ricavano dai dati campionari.

Possiamo allora dire che, per un dato parametro, il numero dei gradi di libertà (g.l. oppure d.f. dall'inglese) è dato dal numero delle osservazioni

(*n=dimensione campionaria*) diminuito del numeri (k) delle stime dei parametri della popolazione che contribuiscono al calcolo del parametro considerato.

In generale, g.l. = n – k. Infatti nel caso della varianza da stimare si ricorre la stima della media della popolazione e quindi k=1.

11.6. Determinazione della numerosità campionaria per inferenza sulla media.

Con la varianza della popolazione (σ^2) nota o stimata con studi pilota , definiti

α = livello di confidenza

e = errore di campionamento $\left(\overline{x} - \mu\right)$ definito dall'equazione $\frac{z \times \sigma}{\sqrt{n}}$ in cui σ è la deviazione standard (s.q.m.) della popolazione, avremo:

$$n = \frac{z^2 \sigma^2}{e^2}$$

Esempio: Il direttore di marketing di un supermercato vuole stimare la cifra media spesa mensilmente per l'abbigliamento dalle clienti donne.

Dovendo fissare l'accuratezza con la quale si intende procedere alla stima, decide che la precisione deve oscillare di ± 10 €. Viene fissata una " confidenza" pari al 95%. Da esperienze pregresse, egli assume $\sigma = 40$ €

I dati sono: $e = 10$; $\sigma = 40$; $z = 1.96$

Quindi: $n = \frac{1.96^2 \times 40^2}{10^2}$ = 61,47 (La dimensione ottimale del campione sarà di **62 unità**).

ESERCIZI DI RICAPITOLAZIONE
SULL'INFERENZA STATISTICA

Esercizio 1

In un processo automatico di confezionamento di biscotti si registra uno s.q.m. pari a 1,5. Determinare la numerosità di un campione affinché l'intervallo di confidenza al 95% ammetta un errore massimo pari a 0,8.

Per determinare la numerosità campionaria si applica la formula

$$n = \frac{z^2_{\alpha/2} * \sigma^2}{\varepsilon^2}$$

in cui

$$z^2_{\alpha/2} == 1{,}96 \ , \sigma^2 = 1{,}5 \, , \varepsilon = 0{,}8$$

Quindi,

$$n = \frac{(1{,}96)^2 * (1{,}5)^2}{(0{,}8)^2} = 13{,}51 \approx 14$$

Esercizio 2

Si vuole conoscere la proporzione di bambini frequentanti la V classe elementare In grado di memorizzare un certo tipo di poesia in un dato tempo. Determinare la numerosità campionaria necessaria affinché la vera proporzione cada i un intervallo al 90% con un errore non superiore al 5%.

La numerosità campionaria si determina con

$$n = \frac{z^2_{\alpha/2} * \pi * (1-\pi)}{\varepsilon^2}$$

Non essendo noto $\boldsymbol{\pi}$ si assume l'ipotesi peggiore in cui $\boldsymbol{\pi}$ sia uguale a 0,5 e quindi $(1-\pi)=0,5$.

Si consideri che per un intervallo di confidenza al 90% il valore di $\alpha/2$ è 1,65.

$$n = \frac{(1,65)^2 * 0.5 * (1-0,5)}{(0,05)^2} = 272,25$$

Esercizio 3

Supponiamo che la perita di peso di n=16 pezzi di metallo, dopo un certo intervallo di tempo sia di 3,42 gr. con una varianza campionaria corretta pari a 0,4624 gr. Costruire u intervallo di confidenza al 99% per la media della perdita di peso.

Si tratta di un piccolo campione n= 16 con varianza della popolazione non nota.

$$\left[\overline{x} - t_{\alpha/2;15} * \frac{\hat{s}}{\sqrt{n}} \leq \mu \leq \overline{x} + t_{\alpha/2;15} * \frac{\hat{s}}{\sqrt{n}} \right] = 1-\alpha$$

in cui $\hat{s}$ è la stima corretta dello scarto quadratico medio. $t_{\alpha/2;15}$ è, invece, il valore teorico (desunto quindi dalle tavole) ed è pari a 2,947.

$$P\left[3,42 - 2,947 * \frac{0,68}{\sqrt{16}} \leq \mu \leq 3,42 - 2,947 * \frac{0,68}{\sqrt{16}} \right] = 0,99$$

Per cui l'intervallo di confidenza è dato da (2,92 – 3,92).