3.11 Legge dei grandi numeri

Hp:

Supponiamo di avere una successione $\{X_i\}$ di variabili aleatorie statisticamente indipendenti con identica funzione di ripartizione. Definiamo la nuova variabile aleatoria:

$$\overline{X}_n = \frac{X_1 + X_2 + \ldots + X_n}{n} \qquad (media \quad campionaria)$$

supponiamo inoltre che

$$\mathbb{E}[X_i] = \mu \quad , \quad V[X_i] = \sigma^2 \qquad i = 1, \ldots, n \tag{3.26}$$

Ts:

$$\overline{X}_n \xrightarrow{d} M$$

dove

a) M è una variabile aleatoria che assume il valore μ con probabilità 1

b) $\xrightarrow{d}$ è un particolare tipo di convergenza

Sia $\begin{cases} X_n & \text{variabile aleatoria con funzione di ripartizione } F_n \\ X & \text{variabile aleatoria con funzione di ripartizione F} \end{cases}$

allora $X_n \xrightarrow{d} X \Leftrightarrow \lim_{n\to\infty} F_n = F$

La legge dice che medie molto grandi di variabili aleatorie tendono alla media vera, ovvero

$$\mathbb{E}[\overline{X}_n] \longrightarrow \mu$$

3.12 Teorema del limite centrale

La legge dei grandi numeri ci dice che $\overline{X}_n \longrightarrow M$, ma non ci dice con quale rapidità (cioè a partire da quale valore di n).

Sotto le stesse ipotesi della legge dei grandi numeri, allora:

$$\overline{X}_n \xrightarrow{d} X \simeq N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \tag{3.27}$$

cioè per n grande $\overline{X}_n$ si distribuisce come una normale $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$, i cui parametri dipendono da quelli delle variabili X_i (cioè da μ e σ).

Dim. Dalla definizione di variabili aleatorie statisticamente indipendenti (2.17), dalle (2.10) e (3.26) si ha

$$V[\bar{X}_n] = V\left[\sum_{i=1}^{N} \frac{X_i}{n}\right] = \frac{1}{n^2} V\left[\sum_{i=1}^{N} X_i\right] = \frac{1}{n^2} \sum_{i=1}^{N} V\left[X_i\right] = \frac{1}{n^2} \sum_{i=1}^{N} \sigma^2 = \frac{\sigma^2}{n}$$

e quindi la deviazione standard della media campionaria e' pari a $\sigma/\sqrt{n}$. Per dimostrare che la media campionaria si distribuisce come una normale di parametri μ e $\sigma/\sqrt{n}$, bisogna sfruttare la disequazione di Chebychev, che esula dai nostri scopi.

Nella pratica il teorema del Limite Centrale ci dice che per n *sufficientemente grande*, la variabile aleatoria media campionaria ha come funzione di distribuzione una normale di parametri μ e $\sigma/\sqrt{n}$, indipendentemente dalla distribuzione della popolazione. Cio' accade per $\mathbf{n \geq 30}$.

Capitolo 4

Stime di parametri

La statistica *inferenziale* consente di dedurre caratteristiche particolari di una popolazione analizzando un numero finito e piccolo (preferibilmente) di suoi individui detto *campione*. Quando le caratteristiche che si vogliono individuare sono esprimibili numericamente, allora prendono il nome di *parametri*.

4.1 Problema del campionamento

Devo determinare le caratteristiche di una popolazione con un numero limitato di individui. Infatti, se la popolazione è vasta, per risparmiare tempo e denaro conviene analizzare un piccolo campione.

4.1.1 Strategie di campionamento

i) Casuale: Associo ad ogni individuo un numero e con un generatore di numeri casuali ne estraggo un certo numero.

Attenzione: se la statistica va fatta sugli abitanti di una città, non si deve fare il campionamento ad esempio tra abbonati al telefono o quelli che si incontrano per strada. Non si prenderebbero in considerazione chi non ha telefono e chi esce poco.

ii) Stratificato: La popolazione è suddivisa in gruppi con stesse caratteristiche (es. età, sesso, etc..)

ii) A grappoli: Si suddivide la popolazione in gruppi *eterogenei*, in modo che ogni singolo gruppo rappresenti l'intera popolazione.

Sia X il carattere della popolazione su cui si è interessati a fare delle inferenze (es. peso, altezza, etc... su una popolazione di persone). Il valore assunto da questo carattere varia a seconda dell'individuo considerato e viene indicato con x. Quindi X è una variabile aleatoria con *distribuzione sconosciuta*, che corrisponde a quella che si otterrebbe facendo ricorso alle tecniche della statistica descrittiva e potendo quindi utilizzare l'intera popolazione.

Definizione 4.1 - *Campione casuale di numerosità n*
É una n-upla $(X_1, X_2, \ldots, X_n)$ di variabili aleatorie indipendenti (estratte da una popolazione) aventi ognuna la stessa distribuzione del carattere X della popolazione. I valori assunti da questa n-upla

$$(x_1, x_2, \ldots, x_n)$$

sono le misure fatte e sono dette *realizzazioni* di $(X_1, X_2, \ldots, X_n)$.

Definizione 4.2 - *Parametro e stima*
Un *parametro* è un valore numerico che descrive una caratteristica di una popolazione, ed è una grandezza associata ad una sua distribuzione (quale il valore atteso e la varianza).
Una *stima* del parametro è una misura fatta sul campione.

Esempio 4.1

X="costo al mq degli appartamenti della città"

Sia $(X_1, X_2, \ldots, X_{80})$ un campione di numerosità 80. Si considera il parametro valore atteso:

μ="costo medio al mq degli appartamenti della città"

Ovviamente non si conosce μ, perchè non si hanno i dati relativi a tutti gli appartamenti della città.

Una stima del parametro μ puo' essere fatta con la media sugli 80 valori

$$\overline{X}_n = \frac{1}{80}\sum_{i}^{n} x_i$$

presumibilmente il valore vero di μ sarà diverso da $\overline{X}_n$.

Allora le stime sono anch'esse delle variabili aleatorie definite in funzione del campione

$$H_n = f(X_1, \ldots, X_n)$$

Esse prendono il nome di *statistiche campionarie* e le loro distribuzioni sono chiamate *distribuzioni campionarie.*

4.2 Principali distribuzioni campionarie

Sia X il carattere di una popolazione con distribuzione cumulativa F, valore atteso μ, varianza σ sconosciuti. Vediamo come stimare questi due parametri incogniti.

Definizione 4.3 - *Media campionaria n-esima di un campione casuale*

$$\overline{X}_n = \frac{X_1 + \ldots + X_n}{n} \tag{4.1}$$

Proposizione 4.1

Se le variabili aleatorie X_i del campione sono tutte indipendenti e

$$\mathbb{E}[X_i] = \mu, \qquad V[X_i] = \sigma^2 \qquad i = 1, n$$

allora

$$\mathbb{E}[\overline{X}_n] = \mu, \qquad V[\overline{X}_n] = \frac{\sigma^2}{n} \qquad (dimostrare) \tag{4.2}$$

In tal caso il valore atteso $\overline{X}_n$ non dipende dalla numerosità del campione, mentre la sua varianza ne dipende. Questo significa che la media campionaria sara' piu' vicina al valore incognito μ quanto piu' e' grande la numerosita' del campione.

Osservazione:

Per la legge dei grandi numeri ed il teorema del limite centrale, per $n \geq 30$, $\overline{X}_n$ si può approssimare con una variabile aleatoria avente distribuzione normale di parametri μ e $\frac{\sigma}{\sqrt{n}}$:

$$\overline{X}_n \simeq N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Definizione 4.4 - *Varianza campionaria n-esima*

Sia $(X_1, \ldots, X_n)$ *un campione estratto da una popolazione avente distribuzione* F*, media* μ *e deviazione standard* σ*. Si definisce* varianza campionaria n-esima*:*

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \quad . \tag{4.3}$$

La distribuzione di S_n^2 si chiama *distribuzione della varianza campionaria n-esima.*

Si prova che

$$\mathbb{E}[S_n^2] = \frac{n-1}{n}\ \sigma^2 \tag{4.4}$$

$$V[S_n^2] = \frac{1}{n}\left(\mathbb{E}[X^4] - \frac{n-3}{n-1}\ \sigma^4\right)$$

Ancora per il teorema del limite centrale si prova che per $n \geq 30$, la sua distribuzione si può approssimare con

$$N(\overline{\mu}, \overline{\sigma})$$

$$\overline{\mu} = \frac{n-1}{n}\ \sigma^2$$

$$\overline{\sigma} = V[S_n^2] \quad .$$

A causa del fattore $\frac{n-1}{n}$ in (4.4), si preferisce considerare una nuova statistica

Definizione 4.5 - *Varianza campionaria n-esima corretta*

$$\widehat{S}_n^2 \ = \ \frac{n}{n-1}\ S_n^2 \ = \ \frac{1}{n-1}\sum_i (X_i - \overline{X}_n)^2 \tag{4.5}$$

da cui ovviamente

$$\mathbb{E}[\widehat{S}_n^2] = \sigma^2 \quad . \tag{4.6}$$

Abbiamo visto che una stima del valore atteso μ di una popolazione e' data dalla media campionaria $\overline{X}_n$, mentre una stima di σ e' data dalla varianza campionaria S_n^2 o quella corretta $\widehat{S}_n^2$.

FUNZIONI EXCEL 4.1

- DEV.ST $= \widehat{S}_n = \sqrt{\dfrac{n}{n-1}}\, S_n$ da usare per un campione
- DEV.ST.POP $= S_n = \sqrt{\dfrac{1}{n}\sum_i (X_i - \overline{X}_n)^2}$ da usare su tutta la popolazione

4.3 Stimatori puntuali

Sia θ un parametro incognito di una popolazione X. Estratto dalla popolazione un campione di numerosita' n, chiamo **estimatore puntuale** $\hat{\theta}$ un numero costruito a partire dalle realizzazioni $(x_1, x_2, ..., x_n)$ del campione casuale. Per esempio la media campionaria e' un estimatore puntuale del parametro valore di aspettazione, come anche la varianza campionaria.

Definizione 4.6 - *Estimatore corretto o non-distorto (unbiased)*
Un estimatore θ di una variabile aleatoria si dice ***corretto o non-distorto (unbiased)****, se il suo valore di aspettazione coincide con il valore vero.*

Dalle proprieta' $(4.2)_1$, (4.6) ne segue che la media campionaria e la varianza campionaria corretta sono estimatori non-distorti, mentre la varianza campionaria e' un estimatore distorto.

Osserviamo che per stimare un parametro θ possiamo definire diversi estimatori corretti. Un criterio per stabilire quale sia preferibile e' il seguente

Definizione 4.7 *Siano $H_{1,n}, H_{2,n}$ due estimatori corretti del parametro θ di una popolazione. Allora diremo che H_1 e' piu' efficiente di H_2 se vale*

$$V[H_{1,n}] \leq V[H_{2,n}]$$

Esempio 4.2 *Si considerino i seguenti estimatori della media μ di una popolazione*

$$H_{1,n} = \frac{X_1 + X_2 + ... + X_n}{n} \quad , \quad H_{2,n} = \frac{X_1}{n} + \frac{X_2 + X_3 + ... + X_n}{2(n-1)}$$

proviamo che il primo e' piu' efficiente del secondo.

Si osservi che il secondo estimatore da' piu' importanza alla prima componente del campione, a cui e' assegnato un peso di valore $\frac{1}{2}$ invece di $\frac{1}{2(n-1)}$. Supposto che tutte le variabili aleatorie del campione siano indipendenti (si veda (4.2)), e' facile vedere che i due estimatori sono corretti, cioe'

$$\mathbb{E}[H_{1,n}] = \mathbb{E}[H_{2,n}] = \mu$$

mentre

$$V[H_{1,n}] = \frac{\sigma^2}{n} \quad \leq \quad V[H_{2,n}] = \left[\frac{1}{4} + \frac{1}{4(n-1)}\right]\sigma^2 \quad \square$$

4.3.1 Altri metodi

Per effettuare delle stime puntali di parametri esistono altri metodi. Quelli principali sono il *metodo dei momenti* e di *massima verosimiglianza*. Nel seguito considereremo solo il secondo metodo con il seguente esempio

Esempio 4.3 *Supponiamo di estrarre da una popolazione X distribuita secondo un'Esponenziale (3.6) di parametro incognita λ, il campione*

$$(3.0, 4.1, 2.8, 5.5, 1.5, 2.2, 6.01.2, 3.2, 0.9) \quad .$$

Vogliamo stimare λ con il metodo di massima verosimiglianza.

Siccome i campioni sono indipendenti, allora

$$\mathcal{L}(x_1, ..., x_{10}, \lambda) = \prod_{i=1}^{10} \lambda e^{-\lambda x_i} = \lambda^{10} e^{-30.4\lambda}$$

La stima del parametro si ottiene determinando quel valore di λ per cui risulta massima la funzione $\mathcal{L}$, ovvero

$$\mathcal{L}' = \lambda^9 e^{-30.4\lambda}(10 - 30.4\lambda) = 0 \quad \rightarrow \quad \lambda = 0.329$$

Osservazione

Nel nostro caso la media campionaria $\bar{x} = 3.04$. Per una variabile aleatoria con distribuzione esponenziale sappiamo che

$$\mathbb{E}[X] = \frac{1}{\lambda}$$

e se poniamo $\mathbb{E}[x] \simeq \bar{x}$ otteniamo ancora $\lambda = 0.329$. Questa osservazione pero' non puo' essere generalizzata a tutte le stime ottenute con il metodo di massima verosimiglianza.

Esempio 4.4 *Supponiamo di effettuare N misure della stessa grandezza fisica x_i, che siano tra loro statisticamente indipendenti ed inoltre affette da errori casuali distribuiti secondo la legge di Gauss. La densità di probabilità corrispondente all'evento casuale costituito dall'osservazione degli N valori (applicando il teorema della probabilità composta)*

$$\prod_{i=1}^{N} \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{x^\star - x_i}{2\sigma_i^2}\right)$$

dove $x^\star$ e' il valore vero (incognito) e σ_i gli errori quadratici medi (noti). Detta funzione di verosimiglianza

$$\mathcal{L}(x_1, ..., x_N, x) = \prod_{i=1}^{N} \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{x - x_i}{2\sigma_i^2}\right)$$

la stima più verosimile di $x^\star$ è quella che rende massima L, rispetto alla variabile incognita x

Si prova che questo massimo esiste ed e' unico e vale

$$\bar{x} = \frac{1}{K} \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} \quad , \quad K = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \tag{4.7}$$

4.4 Campionamento da una distribuzione normale

Consideriamo una popolazione normalmente distribuita $N(\mu, \sigma)$ ed estraiamo da essa un campione di numerosita' n. Ci chiediamo se anche il campione ha una distribuzione normale e con quali parametri. Valgono i seguenti teoremi:

Teorema 4.1 *Sia $(X_1, .., X_n)$ un campione estratto da $N(\mu, \sigma)$, allora la media campionaria $\overline{X}_n$ e' ancora distribuita secondo una normale ma con parametri μ e $\sigma/\sqrt{n}$, ovvero*

$$\overline{X}_n \simeq N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \quad . \tag{4.8}$$

Osservazione

Notiamo che il risultato (4.8) vale per qualsiasi n. Un simile risultato e' ottenuto con il teorema del limite centrale (3.27), dove pero' l'unica ipotesi e' che $n \geq 30$.

Teorema 4.2 *Sia $(X_1, .., X_n)$ un campione estratto da $N(\mu, \sigma)$ di cui supponiamo di non conoscere μ. Allora la varianza campionaria S_n^2 e' tale che*

$$S_n^2 \simeq \frac{\sigma^2}{n-1}\chi^2_{n-1} \quad . \tag{4.9}$$

Teorema 4.3 *Sia $(X_1, .., X_n)$ un campione estratto da $N(\mu, \sigma)$. Allora la variabile aleatoria*

$$T_n = \frac{\overline{X}_n - \mu}{\frac{\widehat{S}_n}{\sqrt{n}}} \simeq t_{n-1} \tag{4.10}$$

Osservazione

Per calcolare T_n occorre conoscere μ (media su tutta la popolazione) ed anche la media e la varianza campionaria, ma non la σ. Inoltre T_n e' distribuita secondo una t di Student che ha come solo parametro n (cioè il grado di libertà),e che quindi non dipende più dai parametri μ e σ.

4.5 Stime intervallari

La stima di un parametro di una popolazione data da un solo numero è detta *stima puntuale* (es. valore medio), mentre se è data da 2 numeri si dice *stima intervallare*.
Ad esempio: "La misura di una distanza è 5,28 mt (stima puntuale)"; "La misura di una distanza è compresa tra gli estremi 5,28±0,03 mt (stima intervallare)".
Sia θ un parametro incognito di una popolazione e $\widehat{\theta}$ un suo estimatore puntuale costruito a partire da un campione casuale estratto dalla popolazione. Un intervallo del tipo

$$I = [\widehat{\theta} - e_1, \widehat{\theta} + e_2] \subseteq \mathbb{R}$$

conterrà il valore θ con maggiore o minore probabilità a seconda dell'ampiezza.

se e_1, e_2 sono grandi, $\mathbb{P}(\theta \in I) \simeq 1$

se $e_1 \simeq e_2 \simeq 0$, $\mathbb{P}(\theta \in I) \simeq 0$

Definizione 4.8 - *Intervallo di confidenza per il parametro θ*
Fissato $\alpha \in [0, 1]$, si chiama *intervallo di confidenza con livello di fiducia* α, quell'intervallo:

$$[\widehat{\theta} - e_1, \widehat{\theta} + e_2]$$

tale che

$$\mathbb{P}(\theta \in [\widehat{\theta} - e_1, \widehat{\theta} + e_2]) = 1 - \alpha$$

solitamente

$$\alpha = 0,1 \ (90\%) \qquad \alpha = 0,05 \ (95\%) \qquad \alpha = 0,01 \ (99\%)$$

E' possibile, in generale, costruire intervalli di confidenza per i parametri *media* e *varianza* di una popolazione.

Esempio 4.5 *Si vuole stimare il parametro θ = altezza media della popolazione composta da tutti gli italiani.*

Un modo (poco pratico) e' quello di misurare l'altezza di tutti gli italiani e quindi calcolarne la media. Invece, possiamo considerare un suo estimatore puntuale $\widehat{\theta}$ = *altezza media di un campione di numerosita' n* che, per la legge dei grandi numeri, sappiamo convergere al valore vero θ per n molto grande.
Fissato il livello di fiducia $\alpha = 0.01$, supponiamo di ottenere l'intervallo $I = [1.62, 1.85]$. Allora concluderemo che il valore vero del parametro sconosciuto θ = *altezza media degli italiani*, cade nell'intervallo I con probabilita' 1 - α. Vedremo nel seguito che questo intervallo I dipende sia dal livello di fiducia α che dalla numerosita' del campione n.

4.5.1 Intervallo di confidenza per la media

a) *Popolazione non normalmente distribuita e varianza σ^2 nota*

Dal teorema del limite centrale sappiamo che la media campionaria $\overline{X}_n$ è approssimabile (per $n \geq 30$) ad una variabile aleatoria con distribuzione normale con media μ (incognita) e deviazione standard $\frac{\sigma}{\sqrt{n}}$ (nota)

$$\overline{X}_n \simeq N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

che si può normalizzare definendo

$$Z = \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \simeq N(0,1)$$

I valori assunti da Z dipendono dal campione $(X_1, \ldots, X_n)$.

Si fissa un livello di fiducia α e si cerca quell'intervallo

$$[-Z_{1-\frac{\alpha}{2}}, Z_{1-\frac{\alpha}{2}}]$$

tale che

$$\mathbb{P}(Z \in [-Z_{1-\frac{\alpha}{2}}, Z_{1-\frac{\alpha}{2}}]) = 1 - \alpha$$

e per la gaussiana normalizzata avrò:

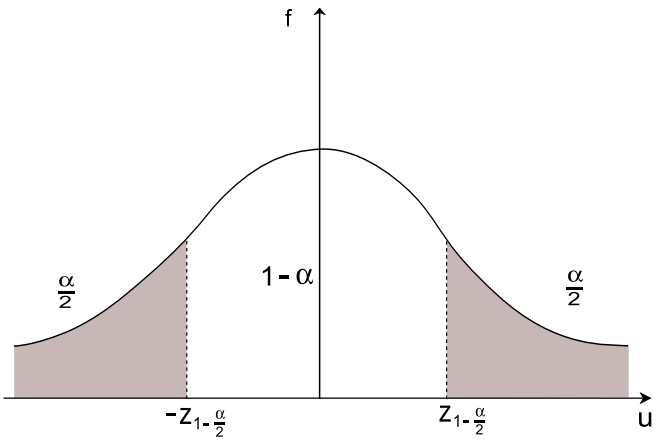

Figura 4.1: Intervallo di confidenza

$(1-\alpha)$ è l'area staccata dal segmento $[-Z_{1-\frac{\alpha}{2}}, Z_{1-\frac{\alpha}{2}}]$ sulla curva e quindi al di fuori di esso ho due regioni simmetriche aventi area $\frac{\alpha}{2}$, infatti $1-\alpha+\frac{\alpha}{2}+\frac{\alpha}{2}=1$.
$Z_{1-\frac{\alpha}{2}}$ è detto *quantile* della distribuzione normale standardizzata ed è tale che alla sua destra lascia un'area pari a $\frac{\alpha}{2}$ e alla sua sinistra un'area pari a $1-\alpha+\frac{\alpha}{2}=1-\frac{\alpha}{2}$.
Essendo

$$\mathbb{P}(X \leq z) = \int_{-\infty}^{z} f(u)du$$

allora avremo

$$\mathbb{P}(X \leq Z_{1-\frac{\alpha}{2}}) = \int_{-\infty}^{Z_{1-\frac{\alpha}{2}}} f(u)du = 1-\frac{\alpha}{2} \quad .$$

Una volta conosciuto il quantile

$$\mathbb{P}(Z \in [-Z_{1-\frac{\alpha}{2}}, Z_{1-\frac{\alpha}{2}}]) = 1-\alpha$$

$$Z = \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$\Downarrow$$

$$\mathbb{P}\left(\frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \in [-Z_{1-\frac{\alpha}{2}}, Z_{1-\frac{\alpha}{2}}]\right) = 1-\alpha$$

che risolta rispetto al parametro μ incognito:

$$-Z_{1-\frac{\alpha}{2}} \quad \leq \quad \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \quad \leq \quad Z_{1-\frac{\alpha}{2}}$$

$$-Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \quad \leq \quad \overline{X}_n - \mu \quad \leq \quad Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

$$\overline{X}_n - Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \quad \leq \quad \mu \quad \leq \quad \overline{X}_n + Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

$$\mathbb{P}\left(\overline{X}_n - Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X}_n + Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

Quindi assegnata la media campionaria $\overline{X}_n$ allora il valore medio (vero) μ della popolazione, con probabilità $1-\alpha$, sta nell'intervallo

$$\left[\overline{X}_n - Z_{1-\frac{\alpha}{2}}\ \frac{\sigma}{\sqrt{n}}\ ,\ \overline{X}_n + Z_{1-\frac{\alpha}{2}}\ \frac{\sigma}{\sqrt{n}}\right] \quad . \tag{4.11}$$

FUNZIONI EXCEL 4.2

Con EXCEL 2010 è possibile calcolare facilmente l'intervallo di confidenza in questo caso. Basta chiamare la funzione CONFIDENZA.NORM che richiede come argomenti α, σ e la dimensione del campione n, il cui risultato è $Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$.

FUNZIONI EXCEL 4.3

Se vogliamo calcolare il solo quantile $Z_{1-\frac{\alpha}{2}}$, allora si possono utilizzare altre funzioni

- DISTRIB.NORM.ST $= \mathbb{P}(X \leq z) = \displaystyle\int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}}\ \exp\left(-\frac{u^2}{2}\right)\ du$
  Fissato z questa funzione mi restituisce l'area della Normale nell'intervallo]-∞, z], ovvero
  $$z \quad \rightarrow \quad \mathbb{P}(X \leq z)$$
- INV.NORM.ST Fissata l'area della Normale, questa funzione mi restituisce la z che corrisponde a quest'area, ovvero
  $$P(X \leq z) \quad \rightarrow \quad z$$
  quindi INV.NORM.ST è l'inversa di DISTRIB.NORM.ST .

In definitiva, per calcolare il quantile $Z_{1-\frac{\alpha}{2}}$, poichè per definizione esso lascia alla sua sinistra un'area pari a $1-\dfrac{\alpha}{2}$:
Fisso $\alpha \longrightarrow$ calcolo $1-\dfrac{\alpha}{2} \longrightarrow$ INV.NORM.ST$\left(1-\dfrac{\alpha}{2}\right) \longrightarrow Z_{1-\frac{\alpha}{2}}$

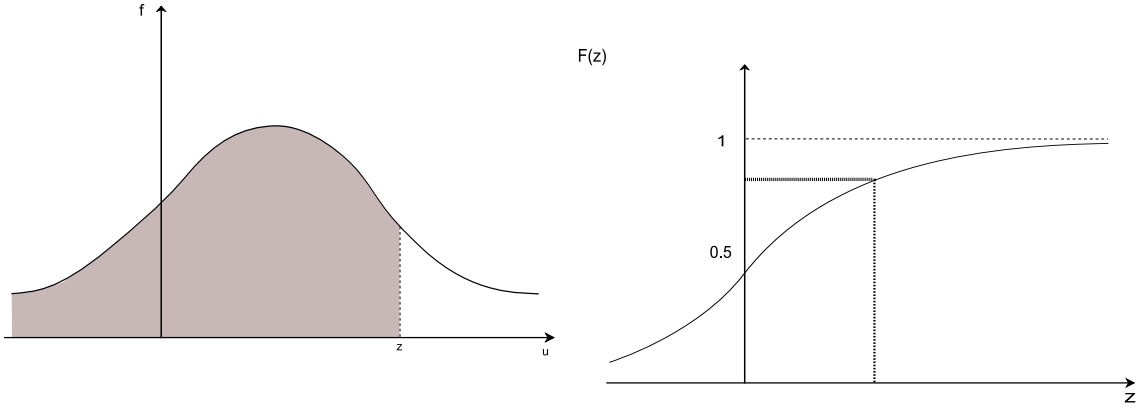

Figura 4.2: DISTRIB.NORM.ST e INV.NORM.ST

b) *Popolazione non normalmente distribuita e varianza σ^2 sconosciuta*

Si ragione come nel caso *a*) sostituendo a σ una sua stima:

$$\sigma \sim \widehat{S}_n = \sqrt{\frac{n}{n-1}}\, S_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}$$

Si prova che per n grande anche la variabile aleatoria:

$$\widehat{Z} = \frac{\overline{X}_n - \mu}{\frac{\widehat{S}_n}{\sqrt{n}}}$$

è distribuita come una normale standardizzata ed il valore medio μ della popolazione sta in

$$\left[\overline{X}_n - Z_{1-\frac{\alpha}{2}}\ \frac{\widehat{S}_n}{\sqrt{n}}\ ,\ \overline{X}_n + Z_{1-\frac{\alpha}{2}}\ \frac{\widehat{S}_n}{\sqrt{n}}\right] \tag{4.12}$$

con probabilità $1 - \alpha$

c) *Popolazione normalmente distribuita e varianza σ^2 nota*

In questo caso prova che la media campionaria $\overline{X}_n$ è una variabile aleatoria con $N(\mu, \frac{\sigma}{\sqrt{n}})$. Si ragiona come nel caso a) con la differenza che n può essere qualunque (nel caso a) invece era $n \geq 30$).

d) *Popolazione normalmente distribuita e varianza σ^2 sconosciuta*

La variabile aleatoria che si usa in questo caso e'

$$T_n = \frac{\overline{X}_n - \mu}{\frac{\widehat{S}_n}{\sqrt{n}}} \simeq t_{n-1} \tag{4.13}$$

che per la proprieta' (4.10) e' distribuita secondo una t di Student con $n-1$ **gradi di liberta'**. Fissato $\alpha \in [0,1]$ determino quel valore $t_{1-\frac{\alpha}{2}}$ (con n-1 gradi di libertà)

$$\mathbb{P}\left(-t_{1-\frac{\alpha}{2}} \ \leq \ T_n \ \leq \ t_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

Ovvero l'area della densità di probabilità di Student $f_X(t)$ compresa in $\left[-t_{1-\frac{\alpha}{2}} \ , \ t_{1-\frac{\alpha}{2}}\right]$ è $1-\alpha$

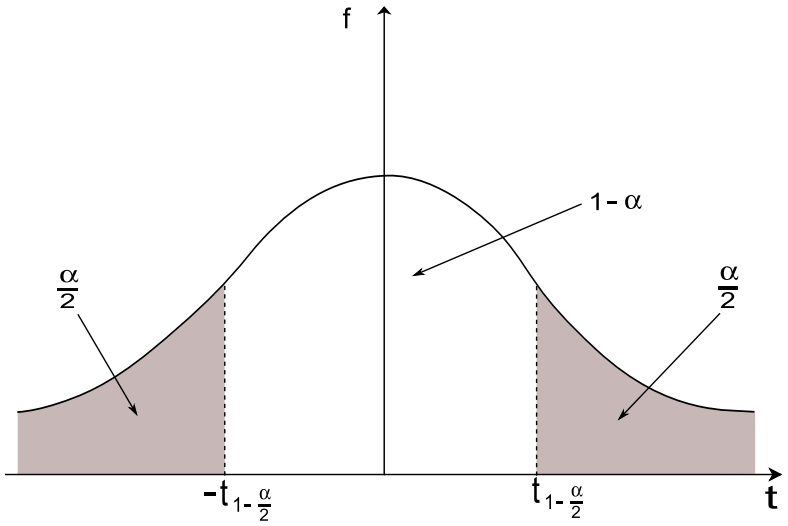

Figura 4.3: Densità di probabilità di Student

$t_{1-\frac{\alpha}{2}}$ è il *quantile* della t di Student ad n -1 gradi di libertà ed è tale che lascia alla sua sinistra un'area di $1-\dfrac{\alpha}{2}$.

Dalla formula precedente, poichè

$$T_n = \frac{\overline{X}_n - \mu}{\frac{\widehat{S}_n}{\sqrt{n}}}$$

si avrà

$$\mathbb{P}\left(-t_{1-\frac{\alpha}{2}}\frac{\widehat{S}_n}{\sqrt{n}} \ \leq \ \overline{X}_n - \mu \ \leq \ t_{1-\frac{\alpha}{2}}\frac{\widehat{S}_n}{\sqrt{n}}\right) = 1-\alpha$$

$$\Downarrow$$

$$\mathbb{P}\left(\overline{X}_n - t_{1-\frac{\alpha}{2}}\frac{\widehat{S}_n}{\sqrt{n}} \ \leq \ \mu \ \leq \overline{X}_n + t_{1-\frac{\alpha}{2}}\frac{\widehat{S}_n}{\sqrt{n}}\right) = 1-\alpha$$

ovvero, conoscendo $\overline{X}_n$ e $\widehat{S}_n$, il parametro μ è compreso nell'intervallo

$$\left[\overline{X}_n - t_{1-\frac{\alpha}{2}}\frac{\widehat{S}_n}{\sqrt{n}} \ , \ \overline{X}_n + t_{1-\frac{\alpha}{2}}\frac{\widehat{S}_n}{\sqrt{n}}\right] \tag{4.14}$$

con probabilità $1-\alpha$.

Poichè la variabile aleatoria T_n è distribuita secondo una t di Student per qualunque valore di n, si parlerà di **statistica per piccoli campioni**. Se n è grande

$$t_n \to N(0,1).$$

FUNZIONI EXCEL 4.4

Con EXCEL 2010 è possibile calcolare facilmente l'intervallo di confidenza in questo caso. Basta chiamare la funzione CONFIDENZA.T che richiede come argomenti $\alpha, \widehat{S}_n$ e la dimensione del campione n, il cui risultato è $t_{1-\frac{\alpha}{2}}\frac{\widehat{S}_n}{\sqrt{n}}$.

FUNZIONI EXCEL 4.5

Se vogliamo calcolare il solo quantile $t_{1-\frac{\alpha}{2}}$ con $n-1$ gradi di libertà, allora dall'eq.(3.23) basta chiamare la funzione INV.T che richiede come argomenti $\alpha, n-1$

$$t_{1-\frac{\alpha}{2}}(n-1) = \text{INV.T}(\alpha, n-1)$$

Possiamo riassumere tutti i casi precedenti con il seguente diagramma di flusso

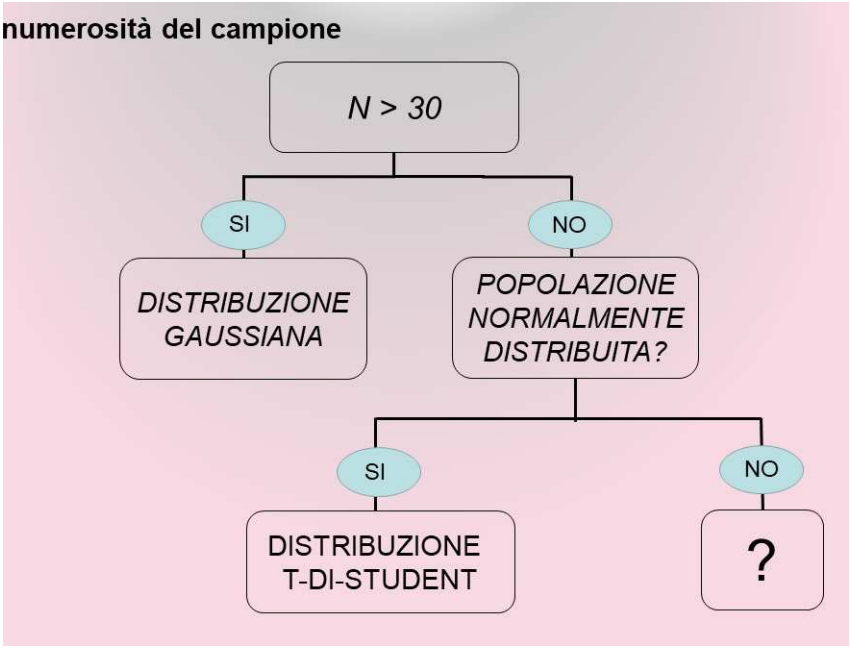

Figura 4.4: Diagramma flusso

Una nota storica

William Gosset (1876-1937, il cui pseudonimo era Student) era un chimico inglese assunto dalla famosa birreria Guinness di Dublino ed eseguiva analisi statistiche su campioni dei prodotti per la mansione che oggi verrebbe chiamata controllo di qualita'. In generale, rilevare un campione costa sempre tempo e denaro. Per questo motivo, spesso Gosset era costretto ad usare per le sue indagini statistiche un numero ridotto di campioni. Gosset si accorse che, avendo una popolazione distribuita secondo una normale di parametri μ e σ incogniti, se prendo un piccolo campione esso non e' ancora distribuito secondo una normale. Gosset scopri' che la variabile casuale T_n (4.13) nella sola incognita μ, e' distribuita secondo una t_{n-1}.

osservazione:

Per intervalli di fiducia che hanno più di un parametro, $\exists\ t_1 \neq t_2$ tali che:

$$\mathbb{P}(X \in [-t_1, t_2]) = 1 - \alpha$$

Nei casi precedenti si sono considerati intervalli simmetrici. La natura del problema determina la scelta dell'intervallo:

i) se si vuole commettere il minimo errore nella scelta di μ si sceglierà l'intervallo simmetrico

ii) se si vuole controllare che μ non raggiunga valori troppo grandi si preferirà un intervallo unilatero:

$$X \in]-\infty, t]$$

Esercizio 4.1 Consideriamo la tabella (1.5) dei carichi di rottura delle travi , relativa ad un campione di n=15.

Supponiamo che detto campione sia estratto da una popolazione con distribuzione normale di parametri incogniti μ e σ. Abbiamo visto in (4.10), che T_n e' distribuita come una t di Student con **n-1** gradi di liberta'.

a) Calcolare l'intervallo di confidenza per μ con livello di fiducia $\alpha = 0.01$

Questo intervallo sarà:

$$\left[\overline{X}_{15} - t_{1-\frac{\alpha}{2}} \frac{\widehat{S}_{15}}{\sqrt{15}} \quad , \quad \overline{X}_{15} + t_{1-\frac{\alpha}{2}} \frac{\widehat{S}_{15}}{\sqrt{15}}\right]$$

con

$\overline{X}_{15} = \ldots$

$\widehat{S}_{15} = \ldots$

$\alpha = 0.01$

$\mathbf{n} = 15$

Per il calcolo del quantile $t_{1-\frac{\alpha}{2}}(n-1)$, come detto in (3.22), si utilizzi INV.T(α, n-1), oppure si utilizzi la funzione CONFIDENZA.T che dà come risultato $t_{1-\frac{\alpha}{2}} \frac{\widehat{S}_{15}}{\sqrt{15}}$. Si ottiene [4770 , 5120]

b) Supponiamo di essere interessati ad una determinazione di μ che non superi un certo valore con $\alpha = 0.01$

Allora devo determinare $t_{1-\alpha}(n-1)$ tale che

$$\mathbb{P}(T_{15} \in \,]-\infty, t_{1-\alpha}]\,) = 1-\alpha$$

$$\Downarrow$$

$$\mathbb{P}\left(-\infty \leq \mu \leq \overline{X}_{15} + t_{1-\alpha} \frac{\widehat{S}_{15}}{\sqrt{15}}\right) = 1-\alpha$$

Si trova $]-\infty, 5099]$, cioè con probabilità 0.99 il parametro μ non supera 5099.

Esercizio 4.2 Riprendere la tabella (1.2) costo al mq di 80 appartamenti e calcolare:

1. la media campionaria $\overline{X}_{80}$

2. la varianza campionaria $\widehat{S}_{80}$

3. per $\alpha = 0.05$ calcolare $Z_{1-\frac{\alpha}{2}}$ per la normale standardizzata

4. calcolare l'intervallo di fiducia per la media

$$\left[\overline{X}_{80} - Z_{1-\frac{\alpha}{2}} \frac{\widehat{S}_{80}}{\sqrt{80}} \quad , \quad \overline{X}_{80} + Z_{1-\frac{\alpha}{2}} \frac{\widehat{S}_{80}}{\sqrt{80}}\right]$$

5. con livello di fiducia $\alpha = 0.05$ stimare il limite superiore per la media μ cioè

$$\mathbb{P}\left(\widehat{Z} \in]-\infty, Z_{1-\alpha}]\right) = 1 - \alpha$$

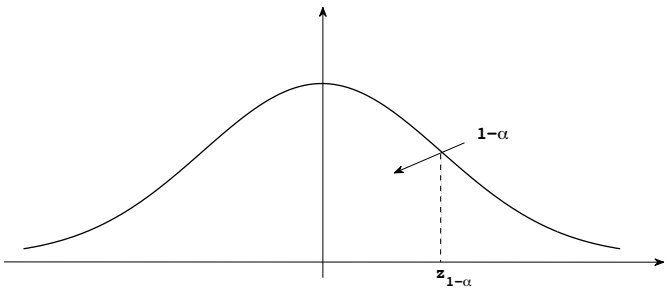

Figura 4.5: Limite superiore per la media

6. É evidente dalla formula (4.12) che l'intervallo di confidenza (a parità di α) dipende da n Determinare n in modo che (con $\alpha = 0.05$) l'intervallo di confidenza simmetrico abbia ampiezza non superiore a 0.03

$$\overbrace{\overline{X}_n - t_{1-\frac{\alpha}{2}}\frac{\widehat{S}_n}{\sqrt{n}} \qquad\qquad \overline{X}_n + t_{1-\frac{\alpha}{2}}\frac{\widehat{S}_n}{\sqrt{n}}}^{0.03}$$

Si possono determinare anche intervalli di confidenza per la varianza σ^2 della popolazione, ma è più complicato. Si hanno risultati significativi soltanto quando la popolazione è *normalmente distribuita*.