

Master di II livello in DATA SCIENCE

Attività formativa **Statistical Models**

Richiami di Inferenza Statistica

Prof. Anthony Cossari

Dipartimento di Economia, Statistica e Finanza (DESF)

Cubo 0C – III piano

a.cossari@unical.it

approfondimenti

- **Cicchitelli G., *Probabilità e statistica*, seconda edizione, Maggioli Editore (2001)**
- Mood A.M., Graybill F.A., Boes D.C., *Introduzione alla statistica*, McGraw-Hill Italia (1988)
- Zenga M., *Inferenza Statistica*, Giappichelli (1996)
- Cicchitelli G, D'Urso P., Minozzo M., *Statistica: principi e metodi*, terza edizione, Pearson (2017)
- Piccolo D., *Statistica*, terza edizione, il Mulino (2010)

Inferenza statistica

analisi condotte su un **insieme di dati campionari** per trarre conclusioni sull'intera **popolazione** di riferimento

conclusioni incerte

(basate sul calcolo delle probabilità)

popolazione

- finita (reale)
- infinita (virtuale)

variabile **numerica** (o eventualmente **codificata**) di interesse nella popolazione

rappresentata da una **variabile casuale** X (v.c.) (aleatoria) (stocastica)

grandezza numerica il cui valore è incerto
(determinato da un esperimento casuale)

- estrazione casuale di un individuo tra i residenti di un Comune per verificarne il sex (variabile casuale X *dicotomica* a valori 0/1)

(1) X variabile casuale **Bernoulliana** \Rightarrow popolazione **Bernoulliana**

- selezione casuale di un appartamento tra quelli di una città per misurarne la superficie (variabile casuale X a *valori reali*)

(2) X variabile casuale **Normale** (per ipotesi) \Rightarrow popolazione **Normale**

(1) e (2) **casi tipici** nella pratica dell'inferenza

Definizione 2.5. Una variabile aleatoria ha distribuzione Bernoulliana se la sua funzione di probabilità è espressa da

$$f(x) = p^x (1-p)^{1-x}, x=0, 1. \quad (2.6)$$

x valore di X

variabile casuale **discreta**

$$f(x) \geq 0, \forall x, \sum_x f(x) = 1.$$

supporto di X

$$f(1) = p \text{ probabilità di } \mathbf{successo}$$

notazione: $X \sim B(p)$

$$f(0) = 1 - p \text{ probabilità di } \mathbf{insuccesso}$$

$f(x)$ modello descrittivo (distributivo) della popolazione (funzione di probabilità)
(p **parametro** del modello)

Teorema 2.2. La media e la varianza della distribuzione di Bernoulli sono date rispettivamente da

$$E(X) = p, \text{Var}(X) = p(1-p).$$

Definizione 2.2. Si chiama *media o valore atteso* della variabile aleatoria discreta X , e la si denota con $E(X)$, la quantità

$$E(X) = \sum_x x f(x), \quad (2.2)$$

dove la somma è estesa a tutti i valori della variabile aleatoria.

notazione tipica $\Rightarrow E(X) = \mu$

Definizione 2.3. Sia X una variabile aleatoria e sia $\mu = E(X)$. Si chiama *varianza* di X , e la si denota con σ^2 o con $\text{Var}(X)$, il valore atteso di $(X - \mu)^2$, ossia

$$\text{Var}(X) = E(X - \mu)^2. \quad (2.3)$$

X discreta \Rightarrow $\text{Var}(X) = \sum_x (x - \mu)^2 f(x)$

notazione tipica $\Rightarrow \text{Var}(X) = \sigma^2$

Definizione 2.16. Una variabile aleatoria continua ha distribuzione normale se la sua funzione di densità è espressa da

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, \quad -\infty < x < +\infty, \quad (2.23)$$

dove μ e σ^2 sono la media e la varianza della variabile aleatoria.

variabile casuale continua

notazione: $X \sim N(\mu, \sigma^2)$

supporto di X

$f(x)$ modello descrittivo (distributivo) della popolazione (funzione di densità)
(μ e σ^2 **parametri** del modello)

La distribuzione normale con media 0 e varianza 1 è chiamata *normale standardizzata* ed assume la forma

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

$$\text{Se } X \sim N(\mu, \sigma^2) \Rightarrow \frac{X-\mu}{\sigma} \sim N(0,1)$$

Definizione 2.11. Una variabile aleatoria X definita nell'intervallo (l, L) è detta continua se esiste una funzione $f(x)$, chiamata funzione di densità di probabilità di X , tale che:

1) $f(x) \geq 0$

2) $\int_l^L f(x) dx = 1$

3) $\int_a^b f(x) dx = P(a < X < b)$

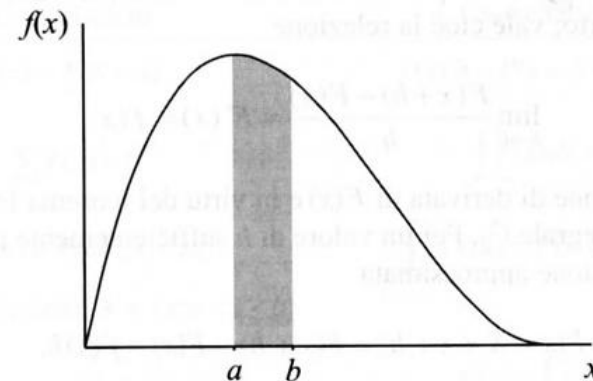


Fig. 2.13. Area sottostante alla curva $f(x)$ nell'intervallo (a, b)

dove a e b sono due valori di X tali che $a < b$.

Definizione 2.13. Si chiama *media o valore atteso* della variabile aleatoria continua X la quantità

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx, \quad (2.17)$$

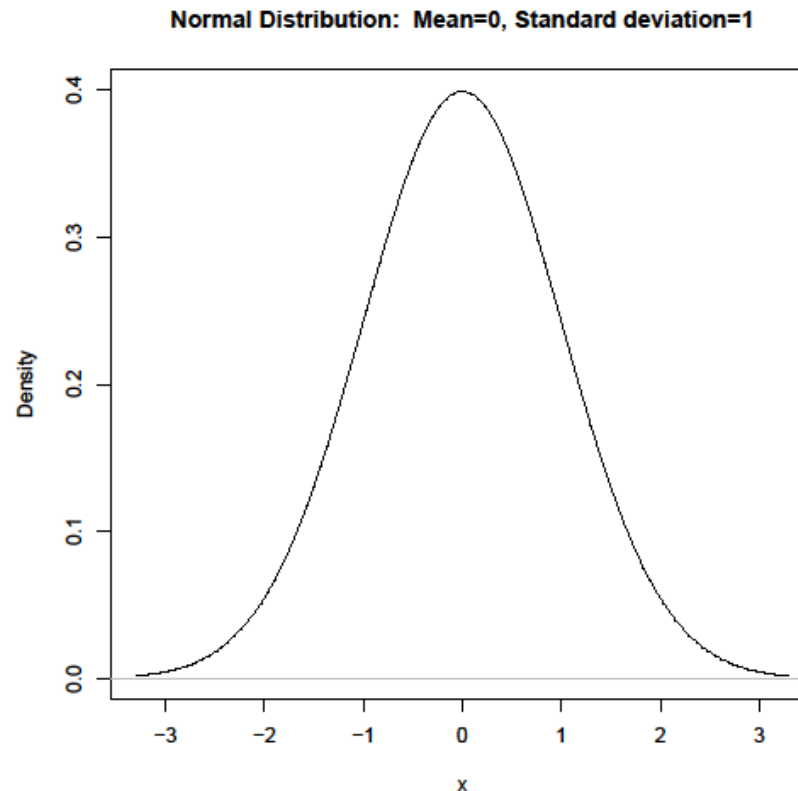
purché l'integrale esista e sia finito ⁽³⁾.

Definizione 2.14. Sia X una variabile aleatoria e sia $\mu = E(X)$ la sua media si chiama *varianza di X* la quantità

$$\text{Var}(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx. \quad (2.18)$$

studio della funzione di densità Normale

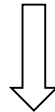
- è simmetrica, avendo come asse di simmetria la retta $x = \mu$;
- è crescente nell'intervallo $(-\infty, \mu)$ e decrescente nell'intervallo $(\mu, +\infty)$;
- ha due punti di flesso in $x = \mu - \sigma$ e $x = \mu + \sigma$;
- è concava (verso il basso) nell'intervallo $(\mu - \sigma, \mu + \sigma)$ e convessa altrove;
- ha come asintoto l'asse delle x .



es. Normale
standardizzata

Campionamento e distribuzioni campionarie

estrazione casuale di **un individuo** tra i residenti di un Comune per verificarne il **sex**



estrazione casuale di **n individui** tra i residenti di un Comune per verificarne il **sex**

Definizione 5.1. *Data una popolazione finita di N unità, un campione casuale di ampiezza n si ottiene estraendo a sorte, con estrazioni successive, n unità dalla popolazione, riponendo, dopo ogni selezione, l'unità estratta nella popolazione.*

estrazioni **con riposizione** \Rightarrow estrazioni **indipendenti**

ad ogni estrazione, si considera la **medesima variabile casuale X**

X_1 è la variabile casuale X alla prima estrazione

X_2 è la variabile casuale X alla seconda estrazione

\vdots

X_n è la variabile casuale X alla n -esima estrazione

campione casuale

variabile casuale **multipla** (X_1, X_2, \dots, X_n)

t.c. X_1, X_2, \dots, X_n sono **indipendenti e identicamente distribuite**

estrazioni **indipendenti** $\Rightarrow X_1, X_2, \dots, X_n$ v.c. **indipendenti**

X_1, X_2, \dots, X_n v.c. **identicamente distribuite** (hanno il **medesimo** modello distributivo)

nell'esempio modello di Bernoulli con il **medesimo** parametro p

popolazione Bernoulliana $\Leftrightarrow X \sim B(p) \Leftrightarrow X_1, X_2, \dots, X_n$ tutte v.c. $B(p)$

campione casuale (X_1, X_2, \dots, X_n) estratto da una popolazione Bernoulliana di parametro p

popolazione Bernoulliana

Esempio 5.2. Si consideri un campione casuale di 100 individui estratto dalla popolazione attiva della Toscana per stabilire il tasso di disoccupazione. Si tratta di una popolazione Bernoulliana, in quanto la variabile aleatoria X_i associata alla singola estrazione può assumere solo i valori 1 e 0 (1 è associato convenzionalmente a “occupato”, 0 a “disoccupato”, o viceversa) con probabilità p e $1 - p$, rispettivamente, essendo p la frequenza relativa della modalità “occupato” nella popolazione.

popolazione Normale

Esempio 5.4. Si supponga che il campione (X_1, X_2, \dots, X_n) sia connesso con l'osservazione della pressione del sangue di 20 persone sane aventi la stessa età. È lecito ammettere che X_i abbia distribuzione normale con media μ e varianza σ^2 . Si può dire, in modo equivalente, che il campione è estratto da una popolazione normale $N(\mu, \sigma^2)$.

popolazione Bernoulliana

Esempio 5.3. Nel controllo di un processo produttivo viene esaminato un elemento ogni mille pezzi prodotti, per accertarne la qualità. Quello che si ottiene a fine giornata è un campione proveniente da una popolazione infinita di tipo Bernoulliano. La variabile associata alla generica osservazione assumerà infatti il valore 1 (pezzo non difettoso) con probabilità p e il valore 0 (pezzo difettoso) con probabilità $1 - p$.

popolazione **virtuale** (potenzialmente **infinita**) costituita da tutti i pezzi (potenzialmente **infiniti**) che possono essere prodotti, **nelle stesse condizioni** (processo produttivo sotto controllo).
Le estrazioni dei pezzi non influiscono sul processo produttivo, sono quindi **estrazioni indipendenti**

campione casuale

variabile casuale **multipla** (X_1, X_2, \dots, X_n)

t.c. X_1, X_2, \dots, X_n sono **indipendenti e identicamente distribuite**

popolazione **finita** o **infinita**

Definizione 5.2. Si chiama *campione casuale di ampiezza n* la variabile aleatoria multipla (X_1, X_2, \dots, X_n) le cui componenti, X_1, X_2, \dots, X_n associate alle varie osservazioni sono indipendenti e identicamente distribuite secondo la funzione di probabilità e di densità $f(x)$, essendo $f(x)$ il modello descrittivo della popolazione.

distribuzione di probabilità del campione casuale

\Rightarrow distribuzione di probabilità congiunta della v.c. multipla (X_1, X_2, \dots, X_n)

funzione di **probabilità** (o di **densità**) congiunta

$$f(x_1, x_2, \dots, x_n) = f(x_1)f(x_2)\dots f(x_n)$$

(fattorizzazione dovuta all'**indipendenza** delle X_i)

$$f(x_1) = f(x_2) = \dots = f(x_n) = f(x)$$

(x_1, x_2, \dots, x_n) campione **osservato**
(campione numerico)

\Rightarrow determinazione della v.c. multipla (X_1, X_2, \dots, X_n)

spazio campionario \Rightarrow insieme di tutti i possibili campioni
(discreto o continuo a seconda della natura di X)

(x_1, x_2, \dots, x_n) punto nello spazio euclideo a n dimensioni

popolazione Bernoulliana

(es. abitanti maggiorenni di un Comune)

$$f(x; p) = p^x (1 - p)^{1-x}, x = 0, 1$$

notazione

p è la probabilità di **successo**

(es. **frazione** di coloro che guardano un programma TV)

campione di ampiezza $n \Rightarrow$

$$f(x_1, x_2, \dots, x_n; p) = p^{\sum_{i=1}^n x_i} (1 - p)^{n - \sum_{i=1}^n x_i}$$

$p = 0.4$ noto

studio **potenziale** dei dati campionari

$n = 10$

$$f(1, 0, 0, 0, 1, 0, 0, 1, 0, 0; 0,4) = 0,4^3 \cdot 0,6^7 = 0,0018$$

$$f(0, 0, 1, 1, 0, 0, 1, 0, 1, 1; 0,4) = 0,4^5 \cdot 0,6^5 = 0,0008$$

problema diretto (calcolo delle probabilità)

problema inverso \Rightarrow dal campione (dati) alla popolazione (parametro **incognito**)

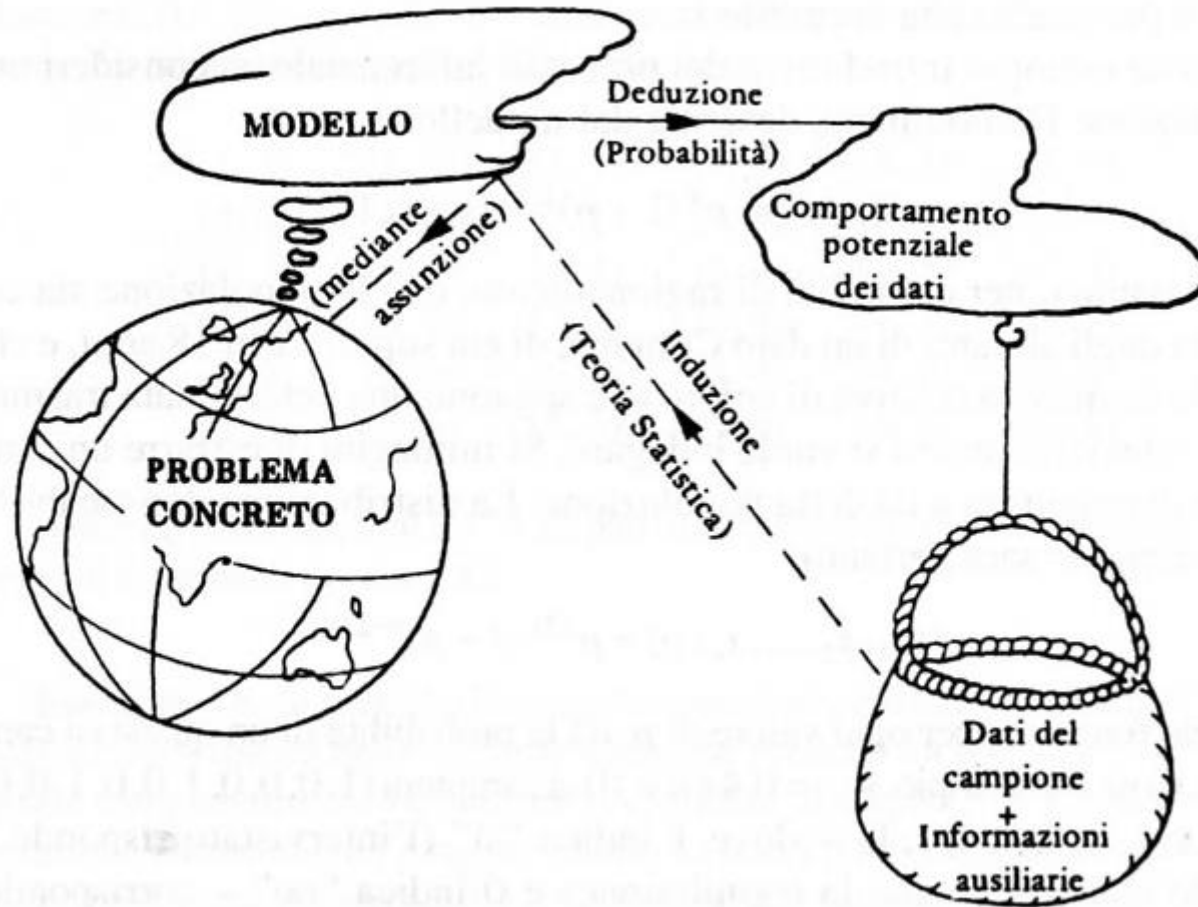


Fig. 5.1. *Problema diretto e problema inverso*
(figura tratta da Barnett, 1975)

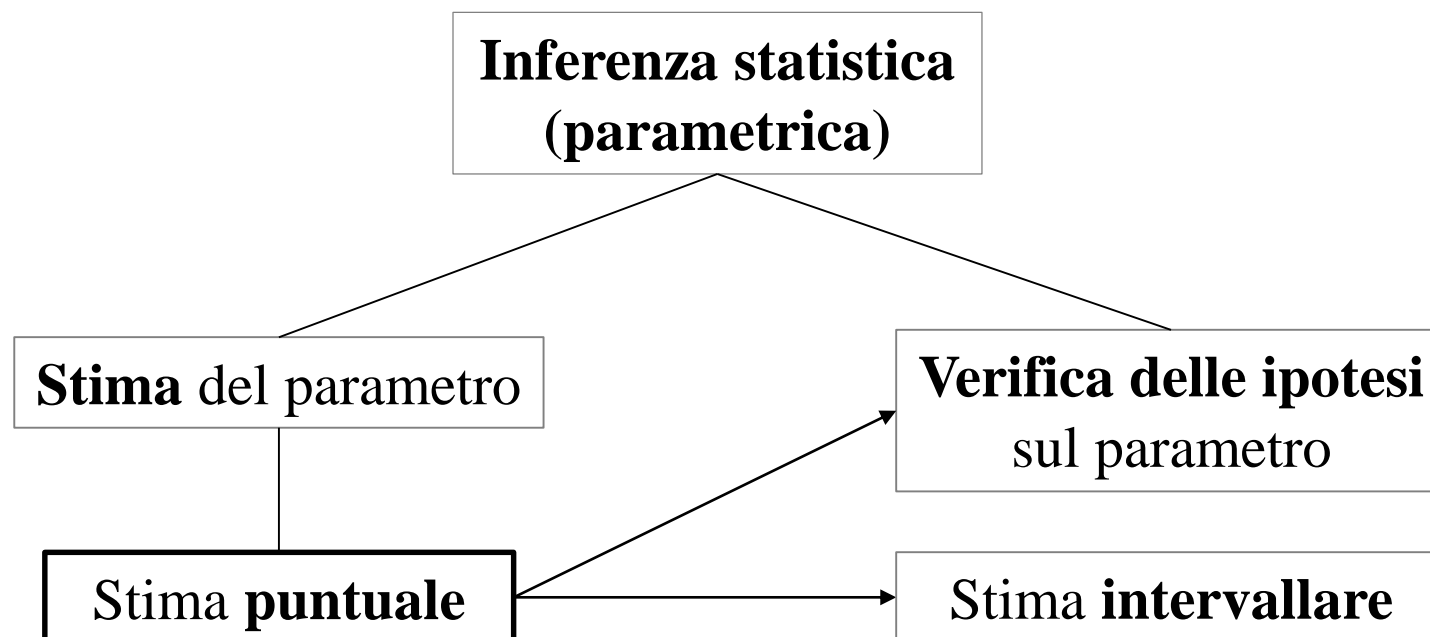
Inferenza statistica \Rightarrow metodi di risoluzione del **problema inverso**

$f(x; \theta)$ modello **parametrico** (funzione di probabilità o di densità)

$\theta \in \Theta$ parametro oggetto di inferenza (Θ spazio parametrico)

(X_1, X_2, \dots, X_n) campione casuale

$(x_1, x_2, \dots, x_n) \in \Omega$ campione osservato (Ω spazio campionario)



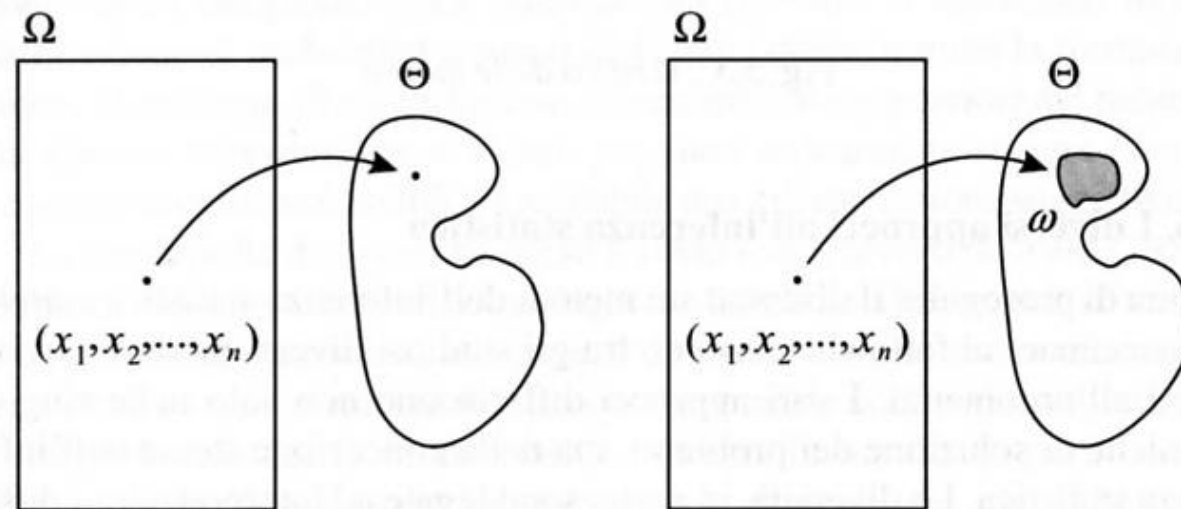


Fig. 5.2. *Stima puntuale e stima per intervallo*

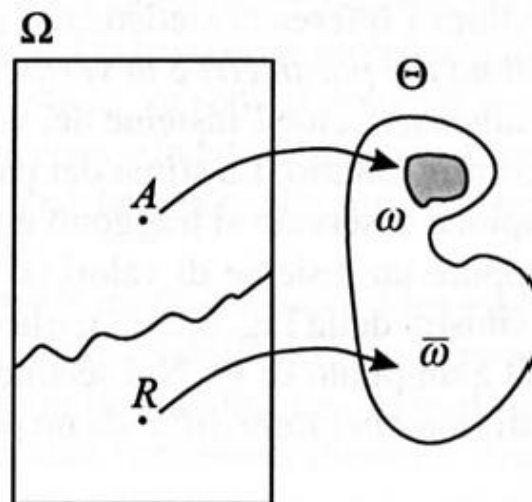


Fig. 5.3. *Verifica delle ipotesi*

Approccio classico all'inferenza statistica

(Fisher, Neyman, Pearson)

metodologie basate sulle **distribuzioni campionarie**
delle **statistiche campionarie** (o di loro funzioni)

(concezione frequentista della probabilità)

Definizione 5.4. Dato un campione casuale (X_1, X_2, \dots, X_n) , si chiama *statistica campionaria* una qualsiasi funzione $g(X_1, X_2, \dots, X_n)$, del campione.

è una v.c.

funzione **solo** del campione

(eventualmente anche di parametri **noti**)

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i,$$

media campionaria

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

varianza campionaria

$$M_r = \frac{1}{n} \sum_{i=1}^n X_i^r$$

momento (campionario) di ordine r

Altri esempi di statistiche: range campionario, mediana campionaria

distribuzione di probabilità della statistica $g(X_1, X_2, \dots, X_n)$
o della funzione di una statistica (o di due o più statistiche)

distribuzione campionaria

funzione di probabilità o funzione di densità

Dipende dal modello descrittivo della popolazione
possibilità

- distribuzione esatta
- distribuzione approssimata (per un grande campione)

Casi tipici: distribuzione Normale, t di Student, Chi-quadrato, F di Fisher

Nella **stima puntuale** \Rightarrow caratteristiche sintetiche come media e varianza

Distribuzione campionaria della media campionaria (media e varianza – popolazione qualsiasi)

Data una qualsiasi popolazione con media μ e varianza σ^2 si ha

$$E(\bar{X}) = \mu, \text{ Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

indipendentemente dal modello descrittivo della popolazione

$\mu = E(X)$ è la media di X (di X_i)

$\sigma^2 = \text{Var}(X)$ è la varianza di X (di X_i)

Distribuzione campionaria della media campionaria (popolazione Normale)

Teorema 5.1. Sia (X_1, X_2, \dots, X_n) un campione casuale proveniente da una popolazione normale $N(\mu, \sigma^2)$. Allora la distribuzione di probabilità della media campionaria è normale. In simboli,

$$\bar{X} \sim N(\mu, \sigma^2 / n). \quad (5.5)$$

$$\Rightarrow Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

$$X \sim N(175, 42)$$

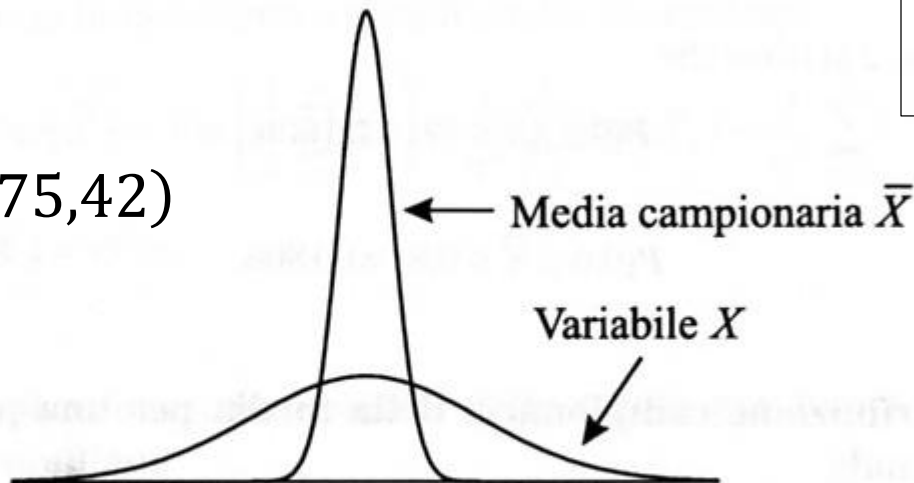


Fig. 5.4. Distribuzioni di X e della media campionaria, \bar{X} per $n = 20$

Distribuzione campionaria di rapporti funzioni di media campionaria e varianza campionaria (popolazione normale – un campione)

scenario in cui la varianza di popolazione non è nota

Teorema 5.5. *Sia data una popolazione normale con media μ e varianza σ^2 . Siano \bar{X} e S^2 la media e la varianza di un campione di dimensione n . Allora il rapporto*

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \quad (5.14)$$

ha distribuzione t di Student con $n - 1$ gradi di libertà.

T funzione di due statistiche

errore standard di \bar{X} (v.c. o suo valore)

T è generata dalla **studentizzazione** di \bar{X}

distribuzione t di Student con r gradi di libertà
simmetrica attorno a 0 (come la Normale standardizzata)

$$E(T) = 0 \ ; \ \text{Var}(T) = \frac{r}{r-2} \quad (r > 2)$$

T ha distribuzione limite (asintotica) $N(0,1)$

$$T \xrightarrow[r \rightarrow \infty]{} N(0,1)$$

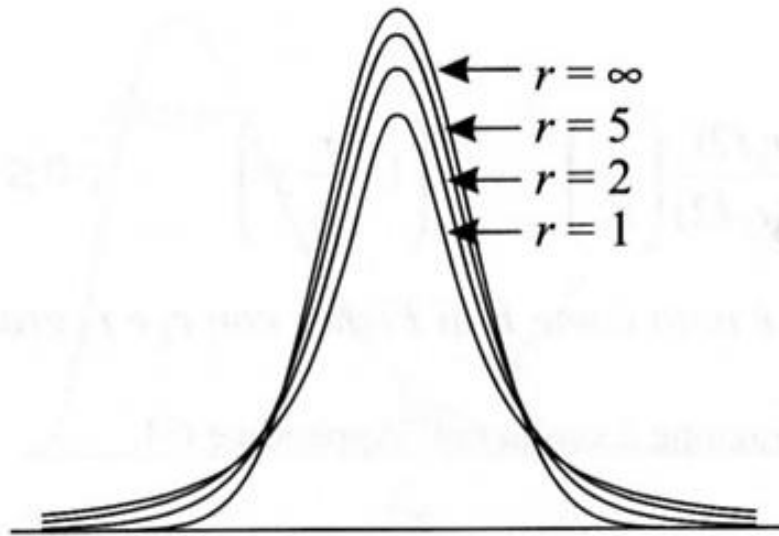


Fig. 4.3. *Distribuzioni t di Student per alcuni valori di r*

Distribuzione campionaria della media campionaria

(popolazione qualsiasi – grande campione)

Teorema del **limite centrale**

Teorema 5.2. Sia (X_1, X_2, \dots, X_n) un campione casuale proveniente da una popolazione qualsiasi. Sia $\bar{X} = \sum_1^n X_i / n$ la media campionaria.

Allora \bar{X} ha distribuzione limite Normale

$$\bar{X} \xrightarrow{n \rightarrow \infty} N\left(\mu, \frac{\sigma^2}{n}\right)$$

**vale anche se la
varianza è ignota**

per n sufficientemente elevato

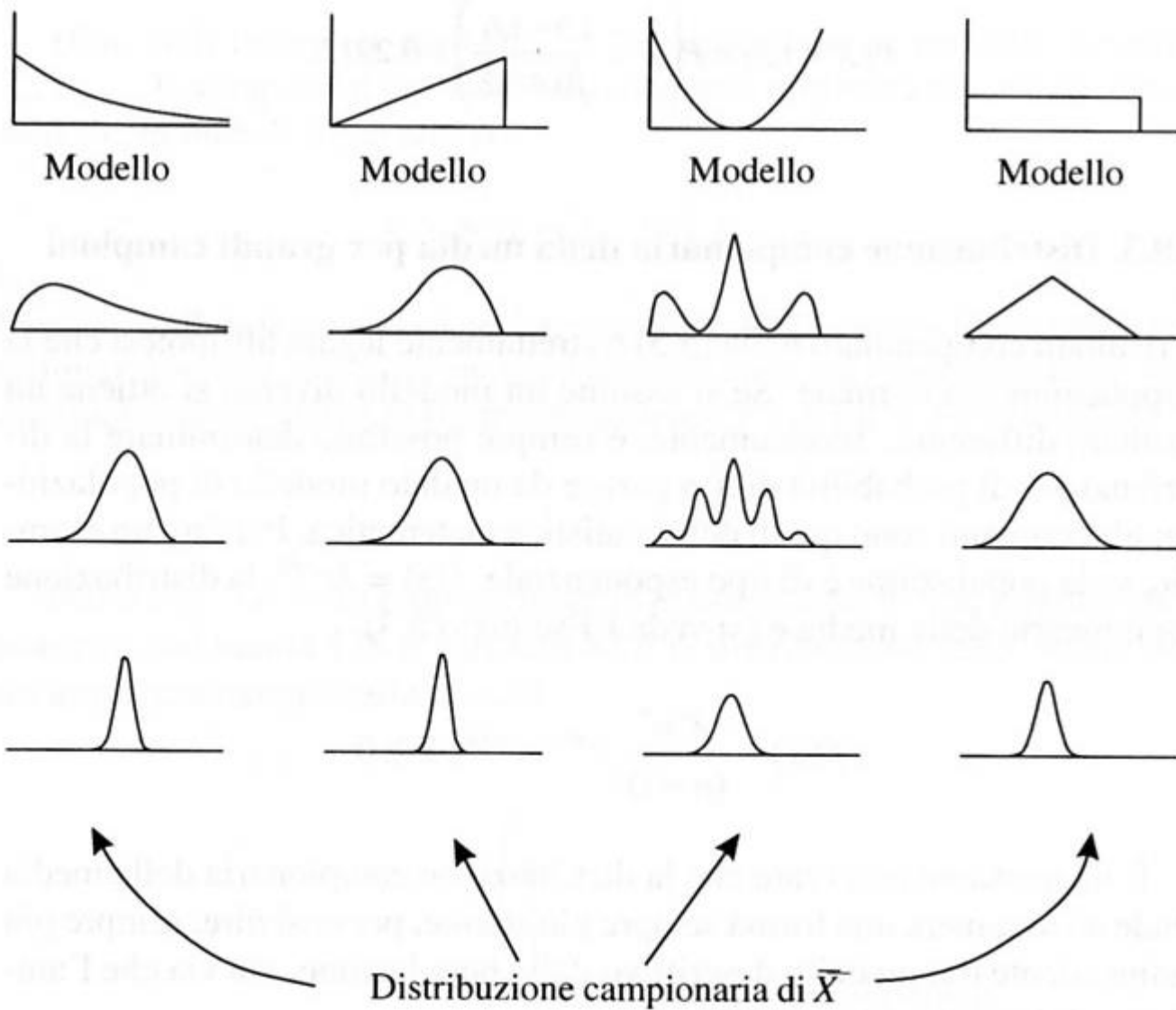
$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

(in modo approssimato)

quanto grande n ?

$$Z = \frac{\bar{X} - \mu}{S / \sqrt{n}} \sim N(0,1)$$

$$Z = \frac{\bar{X} - \mu}{\sqrt{\frac{n-1}{n} S^2 / n}} \sim N(0,1)$$



convergenza meno
veloce per modelli
asimmetrici

$$n \geq 30$$

Fig. 5.5. *Distribuzione campionaria di \bar{X} per alcune popolazioni e per $n = 2, 4, 25$*
(Lapin, 1990)

Distribuzione campionaria della media campionaria (popolazione Bernoulliana)

(X_1, X_2, \dots, X_n) espresso da una **sequenza** di n valori 1 o 0

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

frequenza relativa di successi

$$\left(\frac{0}{n}, \frac{1}{n}, \frac{2}{n}, \dots, \frac{n}{n} \right)$$

$n\bar{X}$ numero di successi $(0, 1, 2, \dots, n)$

$$E(\bar{X}) = p, \quad \text{Var}(\bar{X}) = \frac{p(1-p)}{n}$$

(teorema limite centrale) $\Rightarrow \bar{X}$ ha distribuzione limite Normale

Distribuzione campionaria della differenza tra due medie campionarie (media e varianza – popolazioni qualsiasi)

Si considerino due popolazioni, \mathcal{P}_1 e \mathcal{P}_2 , la prima avente media μ_1 e varianza σ_1^2 , la seconda media μ_2 e varianza σ_2^2 . Sia $(X_1, X_2, \dots, X_{n_1})$ un campione casuale di ampiezza n_1 proveniente dalla popolazione \mathcal{P}_1 e $(Y_1, Y_2, \dots, Y_{n_2})$ un campione casuale di ampiezza n_2 proveniente dalla popolazione \mathcal{P}_2 .

Obiettivo \Rightarrow inferenza su $\mu_1 - \mu_2$

$$\bar{X} = \frac{1}{n_1} \sum_{i=1}^{n_1} X_i \quad \text{e} \quad \bar{Y} = \frac{1}{n_2} \sum_{i=1}^{n_2} Y_i \quad E(\bar{X} - \bar{Y}) = \mu_1 - \mu_2, \quad \text{Var}(\bar{X} - \bar{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

popolazione Bernoulliana

$$E(\bar{X} - \bar{Y}) = p_1 - p_2 \quad \text{Var}(\bar{X} - \bar{Y}) = \frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}$$

Distribuzione campionaria di rapporti funzioni di media campionaria e varianza campionaria (popolazioni normali – due campioni)

scenario in cui le due popolazioni hanno varianza comune ma ignota

$$\sigma_1^2 = \sigma_2^2 = \sigma^2 \quad \text{ipotesi di omoschedasticità}$$

Teorema 5.6. Siano $(X_1, X_2, \dots, X_{n_1})$ e $(Y_1, Y_2, \dots, Y_{n_2})$ due campioni indipendenti provenienti, rispettivamente, dalle popolazioni normali $N(\mu_1, \sigma^2)$ e $N(\mu_2, \sigma^2)$. Siano $S_1^2 = \sum_{i=1}^{n_1} (X_i - \bar{X})^2 / (n_1 - 1)$ e $S_2^2 = \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2 / (n_2 - 1)$ le varianze dei due campioni. Allora il rapporto

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \quad (5.16)$$

ha distribuzione *t* di Student con $n_1 + n_2 - 2$ gradi di libertà.

T funzione di quattro statistiche

Stima puntuale

$f(x; \theta)$ modello **parametrico** (funzione di probabilità o di densità)

$\theta \in \Theta$ parametro oggetto di inferenza (Θ spazio parametrico)

(X_1, X_2, \dots, X_n) campione casuale

$(x_1, x_2, \dots, x_n) \in \Omega$ campione osservato (Ω spazio campionario)

stima puntuale di θ

(attribuzione di un singolo valore a θ)

Definizione 6.1. Si chiama *stimatore del parametro θ* la statistica $T = t(X_1, X_2, \dots, X_n)$ utilizzata per stimare θ .

$t = t(x_1, x_2, \dots, x_n) \Rightarrow$ **stima campionaria**
(**valore osservato** dello stimatore T)

esempi di stimatori

media campionaria $\bar{X} \Rightarrow$ stimatore di $\theta = \mu = E(X)$

varianza campionaria $S^2 \Rightarrow$ stimatore di $\theta = \sigma^2 = \text{Var}(X)$

(\bar{x} e s^2 sono le **stime campionarie** corrispondenti)

stimatori per analogia

(alternativa \Rightarrow stimatori di massima verosimiglianza)

La **qualità** degli stimatori influenza **stima intervallare** e **verifica di ipotesi**
(basate sulla **stima puntuale**)

dato un problema di stima \Rightarrow **pluralità** di stimatori possibili

esempio

- media campionaria
- mediana campionaria \Rightarrow **stimatori possibili di $\mu=E(X)$**
- valore centrale del campione

quale stimatore scegliere?

Le **proprietà** dello stimatore T sono basate sullo studio della sua **distribuzione campionaria**
(modello distributivo, media, varianza, ecc.)

I **valori potenziali** di T (uno di essi sarà la stima campionaria) tendono a **discostarsi** dal valore incognito del **parametro**
(errori di stima)

necessità \Rightarrow T abbia errore di stima **medio** il più **piccolo** possibile

Approccio tipico \Rightarrow restringere l'attenzione agli **stimatori non distorti**

Definizione 6.3. *Uno stimatore $T = t(X_1, X_2, \dots, X_n)$ di θ si dice non distorto se e solo se*

$$E(T) = \theta, \forall \theta.$$

distorsione (bias)

$$B(T) = E(T) - \theta$$

\bar{X} stimatore non distorto di $\mu = E(X)$ $[E(\bar{X}) = \mu]$

S^2 stimatore non distorto di $\sigma^2 = \text{Var}(X)$ $[E(S^2) = \sigma^2]$

indipendentemente dal modello descrittivo della popolazione

Esempio 6.3. Dovendo stimare la varianza σ^2 di una qualsiasi popolazione, si può assumere come stimatore la quantità $(n-1)S^2/n = \sum_1^n (X_i - \bar{X})^2/n$, che rappresenta l'ordinaria formula della varianza che si trova nei testi di statistica descrittiva (si veda, ad esempio, Leti, 1983, p. 379). Lo stimatore è distorto; la sua distorsione è data da

$$B[(n-1)S^2/n] = E[(n-1)S^2/n] - \sigma^2 = (n-1)\sigma^2/n - \sigma^2 = -\sigma^2/n.$$

L'errore di stima **medio**, per uno stimatore T **non distorto**, è la sua **varianza**

$$\text{Var}(T) = E(T - \theta)^2 \quad \text{con } \theta = E(T)$$

esempio: **media campionaria** \bar{X} stimatore di $\theta = \mu = E(X)$ (popolazione qualsiasi)

$$\text{Var}(\bar{X}) = E(\bar{X} - \mu)^2 = \frac{\sigma^2}{n} \quad [\mu = E(\bar{X})]$$

confronto di due stimatori non distorti

dati T_1 e T_2 stimatori non distorti di θ ,

T_1 è **più efficiente** di T_2 se $\text{Var}(T_1) \leq \text{Var}(T_2) \forall \theta$

T_1 stimatore **non distorto** con **varianza uniformemente minima** ($\forall \theta$) rispetto a T_2

Possibile, in generale, determinarlo **per l'intera classe** degli stimatori **non distorti**

La sua determinazione dipende dal modello descrittivo della popolazione

Stima di massima verosimiglianza

$f(x; \theta)$ modello **parametrico** (funzione di probabilità o di densità)

$\theta \in \Theta$ parametro oggetto di inferenza (Θ spazio parametrico)

(X_1, X_2, \dots, X_n) campione casuale

$(x_1, x_2, \dots, x_n) \in \Omega$ campione osservato (Ω spazio campionario)

$$f(x_1, x_2, \dots, x_n; \theta) = f(x_1; \theta) f(x_2; \theta) \dots f(x_n; \theta)$$

distribuzione di probabilità del campione

popolazione **Bernoulliana** di parametro **noto** $p = 0,3$

campione di
ampiezza $n = 8$

$$f(x_1, x_2, \dots, x_8; 0,3) = 0,3^{\sum_{i=1}^8 x_i} (1 - 0,3)^{8 - \sum_{i=1}^8 x_i}$$

probabilità del campione
(problema diretto)

p ignoto

campione osservato

$$(x_1, x_2, \dots, x_8) = (1, 0, 1, 0, 0, 1, 0, 0)$$

$$f(1, 0, 1, 0, 0, 1, 0, 0; p) = p^3 (1 - p)^5$$

Valori di p	Valori della funzione $f(1, 0, 1, 0, 0, 1, 0, 0; p)$
0,1	0,00059
0,2	0,00262
0,3	0,00454
0,4	0,00498
0,5	0,00391
0,6	0,00221
0,7	0,00083
0,8	0,00016
0,9	0,00001

probabilità del
campione osservato
in **funzione** dei
valori di p

verosimiglianza
dei valori di p

Definizione 6.7. Sia X una variabile discreta e sia (x_1, x_2, \dots, x_n) un campione osservato proveniente dalla popolazione descritta dal modello $f(x; \theta)$. Si chiama funzione di verosimiglianza la probabilità congiunta del campione (x_1, x_2, \dots, x_n) interpretata come funzione del parametro θ . In simboli, si può scrivere

$$L(\theta) = f(x_1; \theta) f(x_2; \theta) \dots f(x_n; \theta). \quad (6.6)$$

caso continuo \Rightarrow *mutatis mutandis*, analoga definizione

Definizione 6.8. Sia (x_1, x_2, \dots, x_n) un campione osservato proveniente da una popolazione descritta dal modello $f(x; \theta)$. Si chiama stima di massima verosimiglianza di θ un valore $\hat{\theta} \in \Theta$ che massimizza la funzione di verosimiglianza $L(\theta)$. In simboli, si può scrivere

$$L(\hat{\theta}) = \max_{\theta \in \Theta} L(\theta). \quad (6.7)$$

$\hat{\theta} = t(x_1, x_2, \dots, x_n)$ **stima** di massima verosimiglianza
 $\hat{\theta} = t(X_1, X_2, \dots, X_n)$ **stimatore** di massima verosimiglianza

massimizzazione di $L(\theta) \Rightarrow$ *procedura matematica*

Esempio 6.10. Sia dato un campione (x_1, x_2, \dots, x_n) estratto da una popolazione Bernoulliana con parametro p . Qual è la stima di massima verosimiglianza di p ? La funzione di verosimiglianza del campione è

$$L(p) = p^{\sum_1^n x_i} (1-p)^{n-\sum_1^n x_i}.$$

si può concludere che $\hat{p} = \sum_1^n x_i / n$ è la stima di massima verosimiglianza di p .

nell'esempio precedente $\Rightarrow \hat{p} = \bar{x} = 3/8 = 0,375$

Stima intervallare

Intervallo di confidenza per la media (popolazione Normale – varianza nota)

Sia (X_1, X_2, \dots, X_n) un campione proveniente da una popolazione $N(\mu, \sigma^2)$ con σ^2 nota; sia $1 - \alpha$ il coefficiente fiduciario. Si ha allora

$$\bar{X} \sim N(\mu, \sigma^2 / n)$$

stimatore puntuale

e quindi

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0, 1).$$

non è una statistica!

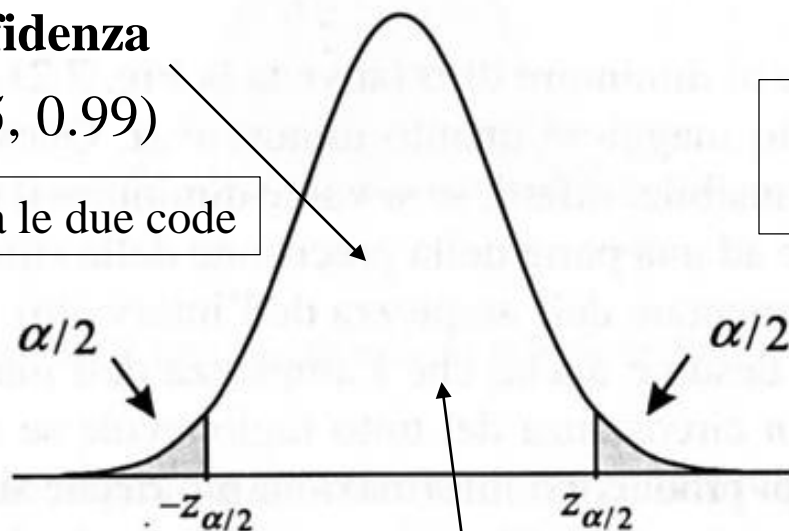
μ ignoto

funzione pivot (quantità pivotale)

funzione della statistica \bar{X} (lo stimatore) e funzione **monotona** di μ
(con distribuzione **indipendente** da μ)

$1-\alpha$ livello di confidenza
(valori tipici 0.95, 0.99)

α ripartito equamente tra le due code



$z_{\alpha/2}$ centile **superiore**
di ordine $\alpha/2$

$-z_{\alpha/2}$ centile **inferiore**
di ordine $\alpha/2$

$-z_{\alpha/2}, z_{\alpha/2}$ tali che $P(Z < -z_{\alpha/2}) = P(Z > z_{\alpha/2}) = \frac{\alpha}{2}$

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

grazie alle proprietà
della funzione pivot

la diseuguaglianza viene «pivotata»

$$P(\bar{X} - z_{\alpha/2} \sigma/\sqrt{n} < \mu < \bar{X} + z_{\alpha/2} \sigma/\sqrt{n}) = 1 - \alpha.$$

probabilità che l'**intervallo casuale** $(\bar{X} - z_{\alpha/2} \sigma/\sqrt{n}, \bar{X} + z_{\alpha/2} \sigma/\sqrt{n})$ contenga μ pari a $1 - \alpha$

deviazione standard di \bar{X}

$$(\bar{X} - z_{\alpha/2} \sigma / \sqrt{n}, \bar{X} + z_{\alpha/2} \sigma / \sqrt{n})$$

intervallo di confidenza (casuale) per μ al $(1-\alpha)100\%$ (95%, 99%, ecc.)

intervallo **casuale** che contiene μ con **probabilità $1-\alpha$**

α è la probabilità che l'intervallo **casuale** non contenga μ

misura del rischio di errore

(0.05, 0.01)

$$(\bar{x} - z_{\alpha/2} \sigma / \sqrt{n}, \bar{x} + z_{\alpha/2} \sigma / \sqrt{n})$$

intervallo di confidenza **osservato**
(intervallo **numerico**)

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

ampiezza dell'intervallo

Interpretazione dell'intervallo di confidenza osservato (e del corrispondente livello di confidenza)

specificazione numerica del problema di stima

popolazione $X \sim N(\mu, 44)$

$n = 27$

$1 - \alpha = 0.95$

$$(\bar{X} - z_{\alpha/2} \sigma / \sqrt{n}, \bar{X} + z_{\alpha/2} \sigma / \sqrt{n})$$

$$(\bar{X} - 2.5, \bar{X} + 2.5)$$

intervallo di confidenza (casuale) al 95%

intervallo **casuale** che contiene μ con **probabilità 0,95**

$\bar{x} = 174.5 \Rightarrow$ intervallo di confidenza osservato (172,177)

$A=5$

intervallo di confidenza al 95%

(intervallo **numerico** che contiene μ con una **confidenza del 95%**)

interpretazione frequentista della probabilità $1 - \alpha$

$$\bar{X} \sim N(\mu, 44/27)$$

popolazione $X \sim N(\mu, 44)$

Campioni

1

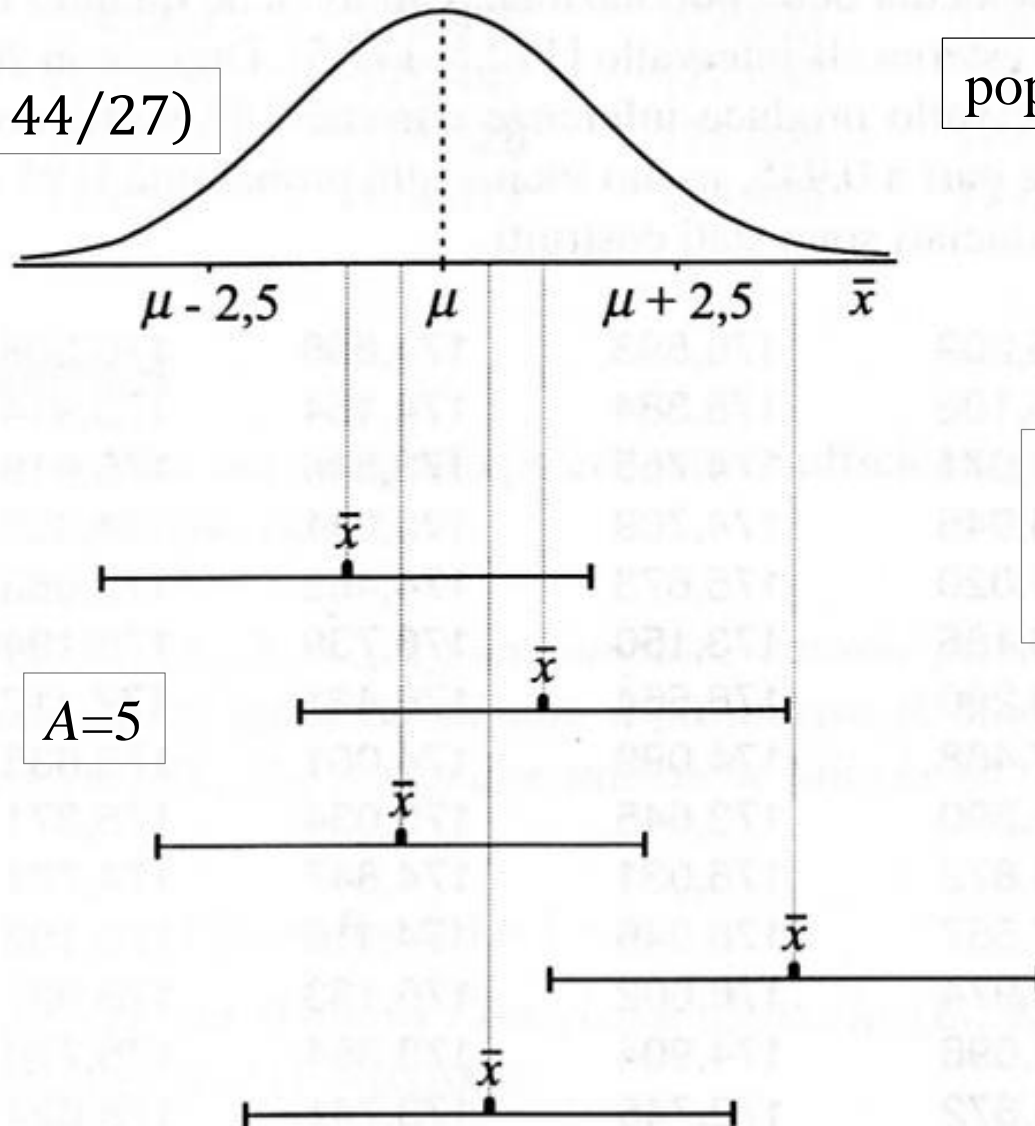
2

3

4

5

$A=5$



L'intervallo contiene μ

\Leftrightarrow

$$\mu - 2,5 \leq \bar{x} \leq \mu + 2,5$$

μ resta incognito

**Qual è la frequenza
di intervalli validi?
(contenenti μ)**

Fig. 7.1. Intervalli $(\bar{x} - 2,5, \bar{x} + 2,5)$ nel campionamento ripetuto dalla popolazione $N(\mu, 44/27)$

μ supposto pari a 175

$X \sim N(175, 44) \Rightarrow 200$ campioni ($n = 27$)

medie campionarie \Rightarrow intervallo $(\bar{x} - 2,5, \bar{x} + 2,5)$

175,209	176,593	174,866	176,308	175,081
175,106	175,384	174,154	175,914	176,739
176,071	174,765	173,544	175,618	173,821
175,945	174,709	175,041	174,620	175,205
176,020	175,673	174,462	*172,063	173,368
173,185	173,150	176,739	175,194	176,463
173,290	176,564	176,431	177,117	172,824
177,488	174,099	174,001	175,635	175,097
175,390	173,645	175,034	175,371	177,157
174,872	176,031	174,847	174,781	172,510
*177,557	*178,046	174,116	175,102	176,441

intervallo valido $\Leftrightarrow \bar{x} \in [\mu - 2,5, \mu + 2,5] = [172.5, 177.5]$

frequenza relativa di intervalli validi $= \frac{189}{200} = 0.945$

interpretazione frequentista di $1-\alpha$

qualità della stima intervallare

- **livello α** (ovvero $1 - \alpha$)

stima **meno incerta** se α più piccolo

- **ampiezza** dell'intervallo

stima **più precisa** se ampiezza più piccola

idealmente, α e ampiezza **entrambi piccoli!**

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

$\alpha \downarrow \quad (1 - \alpha) \uparrow \quad z_{\alpha/2} \uparrow \quad A \uparrow$

soluzione \Rightarrow **fissato α** (ovvero $1 - \alpha$), si ricavano intervalli di **ampiezza minima**

(gli intervalli di uso comune sono di ampiezza minima)

$n \uparrow \quad A \downarrow$

stima più precisa se **n più grande**

$\sigma \downarrow \quad A \downarrow$

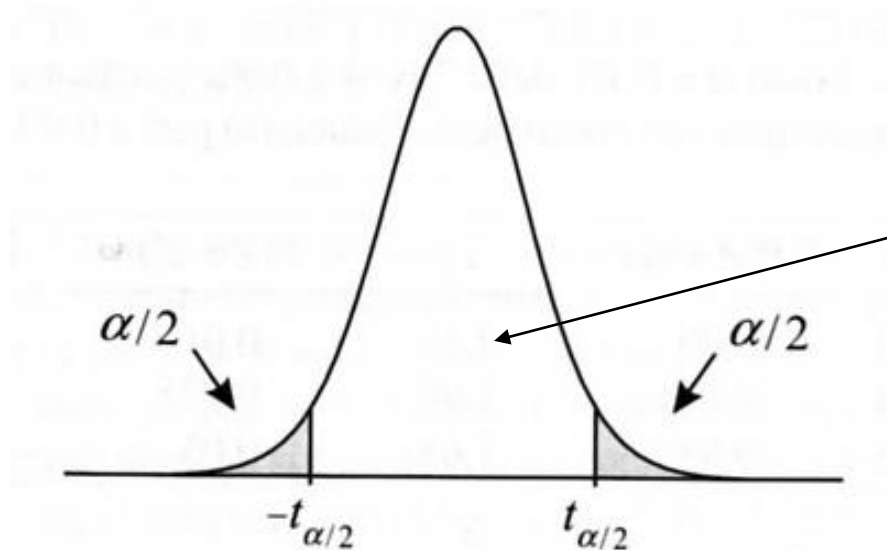
stima più precisa se variabilità
intrinseca dei dati più piccola
(**non controllabile**)

Intervallo di confidenza per la media (popolazione Normale – varianza ignota)

(X_1, X_2, \dots, X_n) campione casuale da $N(\mu, \sigma^2)$ (σ^2 ignota)

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1) \quad \text{funzione pivot}$$

$-t_{\alpha/2}, t_{\alpha/2}$ tali che $P(T < -t_{\alpha/2}) = P(T > t_{\alpha/2}) = \alpha/2$



$$P\left(-t_{\alpha/2} < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{\alpha/2}\right) = 1 - \alpha$$

$$P(\bar{X} - t_{\alpha/2} S/\sqrt{n} < \mu < \bar{X} + t_{\alpha/2} S/\sqrt{n}) = 1 - \alpha$$

errore standard di \bar{X}

$$(\bar{X} - t_{\alpha/2} S/\sqrt{n}, \bar{X} + t_{\alpha/2} S/\sqrt{n})$$

intervallo di confidenza (casuale) per μ al $(1-\alpha)100\%$ (varianza **ignota**)

errore standard di \bar{X}

$$(\bar{x} - t_{\alpha/2} s/\sqrt{n}, \bar{x} + t_{\alpha/2} s/\sqrt{n})$$

intervallo di confidenza osservato

ipotesi

X : quantitativo di catrame $\sim N(\mu, \sigma^2)$

Esempio 7.3. Il produttore di una certa marca di sigarette desidera controllare il quantitativo medio di catrame in esse contenuto. A questo scopo egli osserva un campione di 30 sigarette in cui trova che $\bar{x} = 10,92$ mg e $s = 0,51$ mg. Sulla base di questi dati, si determini l'intervallo fiduciario per μ al 99%.

Poiché $t_{0,005} = 2,756$ (i gradi di libertà sono 29), gli estremi dell'intervallo fiduciario per μ sono $10,92 \pm 2,756 (0,51/\sqrt{30})$. Si può dunque affermare che μ è compreso verosimilmente nell'intervallo (10,66, 11,18).

ordine di grandezza
del quantitativo (medio) di catrame

comando di R `t.test()` non utilizzabile con dati di sintesi

```
n <- 30  
est <- 10.92  
var <- 0.51^2  
conf <- 0.99  
se <- sqrt(var)/sqrt(n)  
q <- qt((1-conf)/2, df=n-1, lower.tail=FALSE)  
ic <- c(est-q*se, est+q*se)
```

centile

```
qt(0.005, df=29, lower.tail=FALSE)
```

ampiezza

```
(est+q*se)-(est-q*se)
```

Intervallo di confidenza per la media

(popolazione Bernoulliana – grandi campioni)

(in generale – teorema limite centrale)
$$Z = \frac{\bar{X} - \mu}{\sqrt{\frac{n-1}{n} S^2 / n}} \xrightarrow{n \rightarrow \infty} N(0,1)$$

$$\frac{n-1}{n} S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

popolazione Bernoulliana

$$= \frac{1}{n} \left(\sum_{i=1}^n X_i^2 - n \bar{X}^2 \right) = \frac{1}{n} \left(\sum_{i=1}^n X_i - n \bar{X}^2 \right) = \bar{X}(1 - \bar{X})$$

per n sufficientemente elevato

funzione pivot
$$Z = \frac{\bar{X} - p}{\sqrt{\bar{X}(1 - \bar{X})/n}} \sim N(0,1) \quad (\text{in modo approssimato})$$

$$P\left(-z_{\alpha/2} < \frac{\bar{X} - p}{\sqrt{\bar{X}(1-\bar{X})/n}} < z_{\alpha/2}\right) \approx 1 - \alpha$$

errore standard di \bar{X}

$$(\bar{X} - z_{\alpha/2}\sqrt{\bar{X}(1-\bar{X})/n}, \bar{X} + z_{\alpha/2}\sqrt{\bar{X}(1-\bar{X})/n})$$

intervallo di confidenza (casuale) per p al $(1-\alpha)100\%$
(approssimato)

X : efficacia del farmaco (SI/NO) $\sim B(p)$

Esempio 7.5. Una ditta farmaceutica è interessata a stabilire l'efficacia di un nuovo farmaco per curare una data malattia. Da un esperimento condotto su 900 pazienti affetti da questa malattia si trova che il farmaco è efficace in 740 casi. Sulla base di questi dati si vuole determinare l'intervallo fiduciario al 95% per la frequenza, p , dei casi in cui il farmaco è efficace (nell'intera popolazione). Poiché $\bar{X} = 0,82$, i limiti fiduciari sono

$$0,82 \pm 1,96\sqrt{(0,82 \times 0,18)/900},$$

ossia 0,80 e 0,84.

efficacia del farmaco tra l'80 e l'84% (con una confidenza del 95%)

prop.test(740,900)

i.c. ottenuto in R con un correttivo
(intervallo *un pò* più ampio ma *più preciso*)

```
n <- 900  
n.succ <- 740  
conf <- 0.95  
est <- n.succ/n  
se <- sqrt(est*(1-est)/n)  
q <- qnorm((1-conf)/2, mean=0, sd=1, lower.tail=FALSE)  
ic <- c(est-q*se, est+q*se)
```

```
qnorm(0.025, mean=0, sd=1, lower.tail=FALSE)
```

confronto ampiezze

```
prop.test(740,900)$conf.int[2]-prop.test(740,900)$conf.int[1]  
(est+q*se)-(est-q*se)
```


Intervallo di confidenza per la differenza tra due medie (popolazioni normali – omoschedasticità)

$(X_1, X_2, \dots, X_{n_1})$ campione casuale da $N(\mu_1, \sigma^2)$
 $(Y_1, Y_2, \dots, Y_{n_2})$ campione casuale da $N(\mu_2, \sigma^2)$ $(\sigma^2 \text{ ignota})$

campioni indipendenti (ampiezze anche diverse)

parametro d'interesse $\mu_D = \mu_1 - \mu_2$

stimatore per $\mu_D \Rightarrow \bar{X} - \bar{Y}$

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2 (1/n_1 + 1/n_2)}}$$

non è una $N(0, 1)$

$$S_c^2 = \frac{S_1^2(n_1 - 1) + S_2^2(n_2 - 1)}{n_1 + n_2 - 2}$$

stimatore di σ^2
 S_c^2 più efficiente di S_1^2 e S_2^2

funzione pivot $T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{S_c^2(1/n_1 + 1/n_2)}} \sim t(n_1 + n_2 - 2)$

$$P\left[-t_{\alpha/2} < \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_c \sqrt{1/n_1 + 1/n_2}} < t_{\alpha/2}\right] = 1 - \alpha,$$

errore standard di \bar{X}

$$[(\bar{X} - \bar{Y}) - t_{\alpha/2} S_c \sqrt{1/n_1 + 1/n_2}, (\bar{X} - \bar{Y}) + t_{\alpha/2} S_c \sqrt{1/n_1 + 1/n_2}].$$

intervallo di confidenza (casuale) per μ_D al $(1-\alpha)100\%$

Esempio 7.6. In un'azienda addetta all'imballaggio di una certa merce sono in uso due macchine diverse. Ci si chiede se i tempi di esecuzione dell'imballaggio sono differenti per le due macchine. A questo scopo vengono osservati i tempi di esecuzione (in secondi) di 10 operazioni di imballaggio, ottenendo, per la prima macchina, un tempo medio $\bar{x} = 55$ con una deviazione standard $s_1 = 1,4$, per la seconda, un tempo medio $\bar{y} = 53$ con una deviazione standard $s_2 = 1,5$.

Si vuole costruire con i dati esposti l'intervallo fiduciario al 95% per $\mu_D = \mu_1 - \mu_2$ (differenza tra i tempi medi di esecuzione dell'imballaggio nelle due macchine). Poiché

$$s_c = \sqrt{\frac{9 \times 1,4^2 + 9 \times 1,5^2}{18}} = 1,45,$$

i limiti fiduciari cercati sono $2 \pm 2,101 \times 1,45 \sqrt{2/10}$, cioè 0,64 e 3,36, essendo $t_{0,025} = 2,101$.

`qt(0.025, df=18, lower.tail=FALSE)`

l'intervallo di confidenza **non contiene lo zero**
 \Rightarrow differenza $\mu_1 - \mu_2$ verosimilmente **positiva**

(tempo medio di imballaggio maggiore per la prima macchina)

```
n1 <- 10  
n2 <- 10  
est1 <- 55  
est2 <- 53  
var1 <- 1.4^2  
var2 <- 1.5^2  
conf <- 0.95  
gl <- (n1+n2-2)  
est <- est1-est2  
var <- ((n1-1)*var1+(n2-1)*var2)/gl  
se <- sqrt(var)*sqrt((1/n1)+(1/n2))  
q <- qt((1-conf)/2, df=gl, lower.tail=FALSE)  
ic <- c(est-q*se, est+q*se)
```

(est+q*se)-(est-q*se)

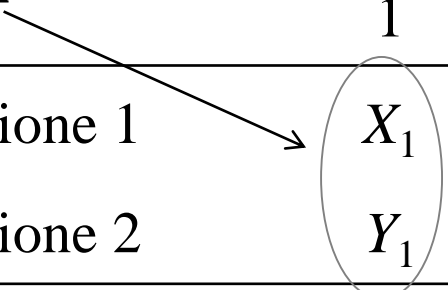
Intervallo di confidenza per la differenza tra due medie (popolazioni normali – dati appaiati)

(X_1, X_2, \dots, X_n) campione casuale da $N(\mu_1, \sigma_1^2)$
 (Y_1, Y_2, \dots, Y_n) campione casuale da $N(\mu_2, \sigma_2^2)$ (σ_1^2, σ_2^2 ignoti)

campioni non indipendenti (medesima ampiezza)

(es.: misura della pressione arteriosa sugli stessi soggetti con due strumenti diversi)

v.a. non indipendenti



	1	2	...	n
campione 1	X_1	X_2	...	X_n
campione 2	Y_1	Y_2	...	Y_n

campione di n coppie di dati

$$D_i = X_i - Y_i \quad (i = 1, \dots, n) \quad (\text{v.a. differenze}) \quad \Rightarrow (D_1, D_2, \dots, D_n)$$

$$E(D_i) = \mu_1 - \mu_2$$

$$\bar{D} = \frac{1}{n} \sum_{i=1}^n D_i \quad E(\bar{D}) = \mu_1 - \mu_2 \quad S_D^2 = \frac{1}{n-1} \sum_{i=1}^n (D_i - \bar{D})^2$$

funzione pivot $T = \frac{\bar{D} - (\mu_1 - \mu_2)}{\sqrt{S_D^2/n}} \sim t(n-1)$

$$P \left[-t_{\alpha/2} \leq \frac{\bar{D} - (\mu_1 - \mu_2)}{\sqrt{S_D^2/n}} \leq t_{\alpha/2} \right] = 1 - \alpha$$

$$\left(\bar{D} - t_{\alpha/2} S_D / \sqrt{n}, \bar{D} + t_{\alpha/2} S_D / \sqrt{n} \right)$$

intervallo di confidenza (casuale) per μ_D al $(1-\alpha)100\%$

Esempio. Si vogliono confrontare 2 *materiali sintetici* (1 e 2) usati per produrre le suole di un certo tipo di scarpe per bambini (**intervallo di confidenza**).

10 bambini calzano uno “speciale” paio di scarpe, una con suola di materiale 1, l’altra con suola di materiale 2 (*scelta casuale*).

I seguenti dati riportano l’*usura* delle suole registrata negli esperimenti.

materiali	bambini									
	1	2	3	4	5	6	7	8	9	10
1	13.2	8.2	10.9	14.3	10.7	6.6	9.5	10.8	8.8	13.3
2	14	8.8	11.2	14.2	11.8	6.4	9.8	11.3	9.3	13.6
d_i	-0.8	-0.6	-0.3	0.1	-1.1	0.2	-0.3	-0.5	-0.5	-0.3

$$\bar{d} = -0.41 \quad s_D^2 = \frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})^2 = 0.149$$

$$1-\alpha = 0.95$$

(intervallo di confidenza per μ_D al 95%)

$$t_{0,025}(9) = 2,262$$

$$-0.41 \mp 2.262\sqrt{0,149/10} \Rightarrow (-0.69, -0.13)$$

l'intervallo di confidenza **non contiene lo zero**
 \Rightarrow differenza $\mu_1 - \mu_2$ verosimilmente **negativa**

(usura media maggiore per il *materiale 2* \Rightarrow migliore il *materiale 1*)

```
t.test(materiale1, materiale2, alternative='two.sided', conf.level=.95, paired=TRUE)
```

dati da imputare

Intervallo di confidenza per la differenza tra due medie (popolazioni Bernoulliane – grandi campioni)

(teorema limite centrale)
$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{n_1 - 1}{n_1} S_1^2 + \frac{n_2 - 1}{n_2} S_2^2}} \xrightarrow{n_1, n_2 \rightarrow \infty} N(0,1)$$

popolazioni Bernoulliane

$$\frac{n_1 - 1}{n_1} S_1^2 = \bar{X} (1 - \bar{X}) \quad ; \quad \frac{n_2 - 1}{n_2} S_2^2 = \bar{Y} (1 - \bar{Y})$$

funzione pivot

per n_1 e n_2 sufficientemente elevati

$$Z = \frac{\bar{X} - \bar{Y} - (p_1 - p_2)}{\sqrt{\bar{X}(1 - \bar{X})/n_1 + \bar{Y}(1 - \bar{Y})/n_2}} \sim N(0,1)$$

(in modo approssimato)

$$P\left[-z_{\alpha/2} < \frac{(\bar{X} - \bar{Y}) - (p_1 - p_2)}{\sqrt{\bar{X}(1 - \bar{X})/n_1 + \bar{Y}(1 - \bar{Y})/n_2}} < z_{\alpha/2}\right] \approx 1 - \alpha$$

$$\hat{\sigma}_{\bar{X} - \bar{Y}}^2 = \bar{X}(1 - \bar{X})/n_1 + \bar{Y}(1 - \bar{Y})/n_2$$

$$[(\bar{X} - \bar{Y}) - z_{\alpha/2} \hat{\sigma}_{\bar{X} - \bar{Y}}, (\bar{X} - \bar{Y}) + z_{\alpha/2} \hat{\sigma}_{\bar{X} - \bar{Y}}]$$

intervallo di confidenza (casuale) per $(p_1 - p_2)$ al $(1 - \alpha)100\%$ (approssimato)

Esempio 7.7. In due sondaggi preelettorali effettuati in due città si osservano i seguenti risultati. Città A: su un campione di 1.000 unità il 29% ha dichiarato di votare per il partito X. Città B: su un campione di 800 unità il 33% ha dichiarato di votare per il partito X. Si vuole determinare l'intervallo fiduciario al 99% per la differenza delle frequenze relative dei votanti per il partito X nelle due popolazioni.

Considerato che $z_{0,005} = 2,576$, e che $\hat{\sigma}_{\bar{X}-\bar{Y}}^2 = (0,29 \times 0,71)/1.000 + (0,33 \times 0,67)/800 = 0,00048$, l'intervallo fiduciario cercato è

$$[(0,29 - 0,33) - 2,576 \times 0,022, (0,29 - 0,33) + 2,576 \times 0,022],$$

ossia $(-0,0967, 0,0167)$.

l'intervallo di confidenza **contiene lo zero**
 \Rightarrow differenza $p_1 - p_2$ verosimilmente **pari a 0**)

(propensione al voto uguale nelle due città)

`prop.test(c(290,264),c(1000,800),conf.level = 0.99)`

```
n1 <- 1000  
n2 <- 800  
# n1.succ <-  
# n2.succ <-  
conf <- 0.99  
est1 <- 0.29    # n1.succ/n1  
est2 <- 0.33    # n2.succ/n2  
est <- est1-est2  
se <- sqrt(est1*(1-est1)/n1+est2*(1-est2)/n2)  
q <- qnorm((1-conf)/2, mean=0, sd=1, lower.tail=FALSE)  
ic <- c(est-q*se, est+q*se)
```

confronto ampiezze

```
prop.test(c(290,264),c(1000,800),conf.level = 0.99) $conf.int[2]-  
prop.test(c(290,264),c(1000,800),conf.level = 0.99) $conf.int[1]  
(est+q*se)-(est-q*se)
```

Verifica di ipotesi

Definizione 8.1. Sia $f(x; \theta)$ il modello descrittivo della popolazione. L'ipotesi statistica è una affermazione o una congettura che riguarda il parametro θ .

es.: $\theta = \theta_0$, $\theta \geq \theta_0$, $\theta < \theta_0$ (θ_0 valore fissato)

H_0 ipotesi nulla (sottoposta a verifica) vs **H_1 ipotesi alternativa**

Definizione 8.2. Sia (X_1, X_2, \dots, X_n) un campione proveniente da una popolazione descritta dal modello $f(x; \theta)$, e sia H_0 l'ipotesi nulla su θ . Si chiama test di significatività il procedimento con cui si decide, alla luce dei dati del campione, se accettare o rifiutare H_0 .

decisioni corrette \Rightarrow

- rifiuto di H_0 quando H_0 è falsa
- accettazione di H_0 quando H_0 è vera

possibili errori \Rightarrow

- rifiuto di H_0 quando H_0 è vera (falso negativo)
- accettazione di H_0 quando H_0 è falsa (falso positivo)

Definizione 8.3. *Si chiama errore di primo tipo la decisione di rifiutare l'ipotesi H_0 quando è vera; si chiama errore di secondo tipo la decisione di accettare l'ipotesi H_0 quando è falsa.*

errore di I tipo più grave dell'errore di II tipo
(implica una modifica della realtà)

Verifica di ipotesi sulla media

(popolazione normale – varianza nota)

(X_1, X_2, \dots, X_n) campione casuale da $N(\mu, \sigma^2)$ (σ^2 nota)

$H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0$ **test unilaterale** (a coda destra)

test basato sul confronto (distanza) tra \bar{x} e μ_0

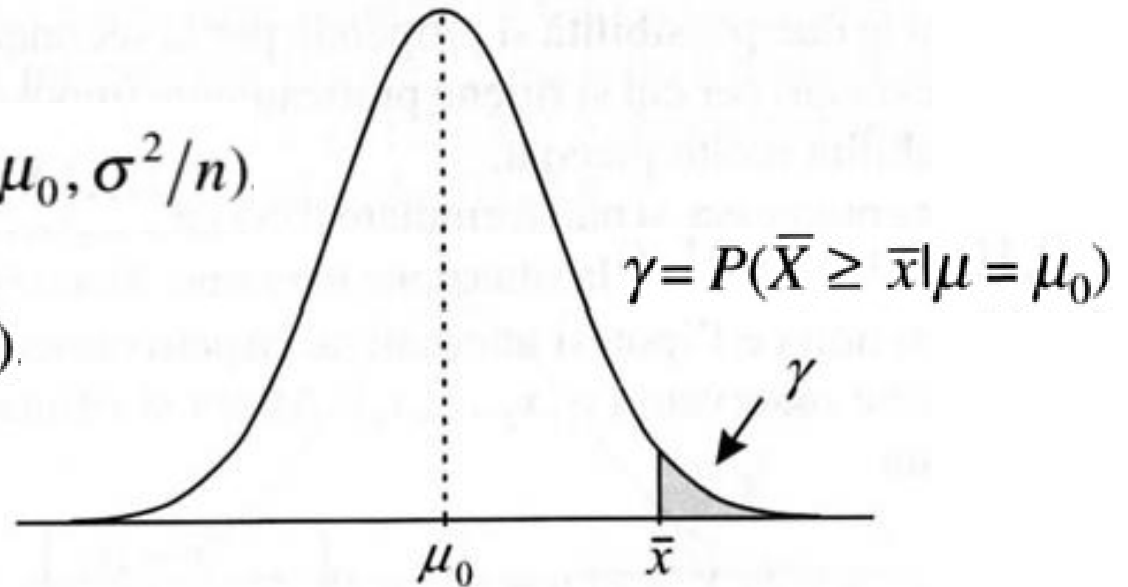
Esempio 8.1. L'ufficio qualità di uno stabilimento che produce pasta alimentare intende controllare se il peso dichiarato nella confezione di 500 gr risponda al vero, oppure se il processo di confezionamento dà luogo ad un peso medio superiore. Poiché sul processo influisce una pluralità di fattori, è ragionevole assumere che il peso di una confezione sia una variabile aleatoria normale. Un campione casuale di n confezioni può quindi essere assunto come proveniente da una popolazione normale. L'ipotesi nulla da verificare è $H_0 : \mu = 500$, l'ipotesi alternativa è $H_1 : \mu > 500$.

In un campione di $n = 25$ unità, l'ufficio qualità trova $\bar{x} = 503,7$. Su questa base deve stabilire se accettare o rifiutare l'ipotesi $H_0 : \mu = 500$. La decisione sarà basata sul confronto tra il valore osservato $\bar{x} = 503,7$ e la quantità $\mu = 500$, valore atteso della media campionaria sotto l'ipotesi nulla.

sotto H_0

$$\bar{X} \sim N(\mu_0, \sigma^2/n)$$

$$\mu_0 = E(\bar{X} | \mu = \mu_0)$$



Lo scarto di \bar{x} da μ_0 è tollerabile (casuale)?

misura di distanza \Rightarrow probabilità γ

\bar{x} è tanto più distante da μ_0 (maggiore di μ_0) quanto minore è γ

γ piccolo \Rightarrow si rifiuta H_0

(valore \bar{x} poco plausibile - \bar{x} **significativamente** maggiore di μ_0)

test significativo

calcolo di γ

$$\gamma = P(\bar{X} \geq \bar{x} | \mu = \mu_0) = P\left(Z \geq z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right)$$

α livello di significatività (valori tipici 0.05, 0.01)

Se $\gamma \leq \alpha \Rightarrow$ si rifiuta H_0

(in caso contrario si accetta H_0)

\bar{X} statistica test (anche Z)

γ *p-value* (livello di significatività osservato)

il rifiuto di H_0 è tanto più plausibile (evidente) quanto minore è γ

α = probabilità errore di I tipo

Esempio 8.2. Proseguendo nell'esempio 8.1, si assuma che la varianza della popolazione sia $\sigma^2 = 42,5$. Allora la probabilità γ è pari a

$$\gamma = P(\bar{X} \geq 503,7 \mid \mu = 500) = P\left(Z \geq \frac{503,7 - 500}{\sqrt{42,5/25}}\right) = 0,0023.$$

test a coda destra

γ piccolo \Rightarrow si rifiuta H_0

$\bar{x} = 503,7$ significativamente maggiore di $\mu_0 = E(\bar{X} \mid H_0) = 500$

$z = \frac{503,7 - 500}{\sqrt{42,5/25}} = 2,84$ significativamente maggiore $E(Z) = 0$

evidenza forte verso il rifiuto di H_0

$H_0 : \mu = \mu_0$ vs $H_1 : \mu < \mu_0$ **test unilaterale** (a coda sinistra)

si rifiuta H_0 se

$$\gamma = P(\bar{X} \leq \bar{x} | \mu = \mu_0) = P\left(Z \leq z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right) \leq \alpha$$

$H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$ **test bilaterale** (a due code)

si rifiuta H_0 se

$$\gamma = P(\bar{X} \geq \bar{x} | \mu = \mu_0) = P\left(Z \geq z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right) \leq \frac{\alpha}{2} \quad (\bar{x} > \mu_0)$$

$$\gamma = P(\bar{X} \leq \bar{x} | \mu = \mu_0) = P\left(Z \leq z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right) \leq \frac{\alpha}{2} \quad (\bar{x} < \mu_0)$$

in alternativa si rifiuta H_0 se **$2\gamma \leq \alpha$** (approccio in R)

Verifica di ipotesi sulla media

(popolazione normale – varianza ignota)

(X_1, X_2, \dots, X_n) campione casuale da $N(\mu, \sigma^2)$ (σ^2 ignota)

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad \begin{cases} H_1 : \mu > \mu_0 \\ H_1 : \mu < \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

sotto $H_0 \Rightarrow T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t(n-1)$

T statistica test

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \quad \text{valore osservato di } T \quad E(T|H_0) = 0$$

si rifiuta H_0 se t è **significativamente** maggiore (minore)(diverso) di/a 0

(\bar{x} è **significativamente** maggiore (minore)(diversa) di/a μ_0)

es.: $H_1 : \mu > \mu_0 \Rightarrow$ si rifiuta H_0 se $\gamma = P\left(T \geq t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}\right) \leq \alpha$

ipotesi X : pulsazioni cardiache $\sim N(\mu, \sigma^2)$

Esempio 8.5. Sono state osservate le pulsazioni cardiache (in battiti per minuto) di 10 studenti maschi al primo anno della facoltà di medicina. La media \bar{x} e la varianza s^2 sono risultate rispettivamente 68,70 e 75,12. È noto che il valore clinico normale della frequenza media di pulsazioni per i maschi giovani è di 72 battiti per minuto. Il quesito cui si deve dare risposta è se i dati del campione sono compatibili con il valore della frequenza media di pulsazioni assunta come normale. L'ipotesi da verificare è dunque $H_0 : \mu = 72$, a fronte dell'alternativa $\mu \neq 72$.

$$t = \frac{68,70 - 72}{\sqrt{75,12/10}} = -1,20.$$

test bilaterale

$$\gamma = P(T \leq t = -1,20) = 0,13 \quad (\text{da confrontare con } \alpha/2)$$

\Rightarrow si accetta H_0

$$\text{ovvero } 2 * 0,13 = 0,26 \quad (\text{da confrontare con } \alpha)$$

($t = -1,20$ non è **significativamente** diverso da 0)

($\bar{x} = 68,70$ non è **significativamente** diverso da $\mu_0 = 72$)

comando di R `t.test()` non utilizzabile con dati di sintesi

calcolo di t

```
n <- 10  
est <- 68.70  
var <- 75.12  
mu.0 <- 72  
se <- sqrt(var)/sqrt(n)  
stat <- (est-mu.0)/se
```

p-value

```
pt(stat, df=n-1, lower.tail=TRUE)
```

se stat ha valore positivo

```
pt(stat, df=n-1, lower.tail=FALSE)
```

Relazione tra verifica di ipotesi (test bilaterali) e stima intervallare

dato un test bilaterale

si accetta $H_0 \Leftrightarrow$

l'intervallo di confidenza corrispondente **include**
il valore del parametro specificato da H_0

si rifiuta $H_0 \Leftrightarrow$

l'intervallo di confidenza corrispondente **non include**
il valore del parametro specificato da H_0

vale per tutti i test bilaterali ed i corrispondenti intervalli di confidenza
(stesso valore di α)

(rivisitazione di esempi)

Verifica di ipotesi sulla media
(popolazione Bernoulliana – grande campione)

$$H_0 : p = p_0 \quad \text{vs} \quad \begin{cases} H_1 : p > p_0 \\ H_1 : p < p_0 \\ H_1 : p \neq p_0 \end{cases} \quad \begin{matrix} \text{sotto } H_0 \\ \\ \end{matrix} \quad Z = \frac{\bar{X} - p_0}{\sqrt{p_0(1-p_0)/n}} \xrightarrow{n \rightarrow \infty} N(0,1)$$

per n sufficientemente elevato

$$Z = \frac{\bar{X} - p_0}{\sqrt{p_0(1-p_0)/n}} \sim N(0,1) \quad (\text{in modo approssimato})$$

Z statistica test (test *normale* approssimato)

Esempio 8.7. Un partito politico ha ricevuto nelle ultime elezioni il 35% dei voti. Quattro anni dopo, da un sondaggio d'opinione basato su 300 interviste si è trovato che il 32% degli intervistati ha dichiarato di essere disposto a votare per quel partito. Ci si chiede se, rispetto al risultato elettorale, la situazione del partito sia peggiorata. L'ipotesi da verificare è $H_0 = p = 0,35$, a fronte dell'alternativa $H_1 : p < 0,35$. Allora si ha

$$z = \frac{0,32 - 0,35}{\sqrt{(0,35 \times 0,65)/300}} = -1,09.$$

test a coda sinistra

$$\gamma = P(Z \leq z = -1,09) = 0,14 \Rightarrow \text{si accetta } H_0$$

($z = -1,09$ non è **significativamente** minore di 0)

($\bar{x} = 0,32$ non è **significativamente** minore di $p_0 = 0,35$)

(la situazione del partito *non è peggiorata*)

```
n <- 300  
# n.succ <-  
est <- 0.32    # n.succ/n  
p.0 <- 0.35  
se <- sqrt(p.0*(1-p.0)/n)  
stat <- (est-p.0)/se
```

p-value (test a coda sinistra)

```
pnorm(stat, mean=0, sd=1, lower.tail=TRUE)
```

```
prop.test(96,300,p = 0.35,alternative="less")
```

R esegue un test chi-quadrato
(invece del test Normale)

Verifica di ipotesi sulla differenza tra due medie

(popolazioni normali – omoschedasticità)

$(X_1, X_2, \dots, X_{n_1})$ campione casuale da $N(\mu_1, \sigma^2)$
 $(Y_1, Y_2, \dots, Y_{n_2})$ campione casuale da $N(\mu_2, \sigma^2)$
(σ^2 ignota)

campioni indipendenti (ampiezze anche diverse)

Ipotesi H_0	Ipotesi H_1
	$\mu_D = \mu_1 - \mu_2 > 0$
$\mu_1 - \mu_2 = 0$	$\mu_D = \mu_1 - \mu_2 < 0$
	$\mu_D = \mu_1 - \mu_2 \neq 0$

sotto H_0

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim t(n_1 + n_2 - 2)$$

T statistica test

si rifiuta H_0 se t è **significativamente** maggiore (minore)(diverso) di/a 0

se l'ipotesi di omoschedasticità ($\sigma_1^2 = \sigma_2^2 = \sigma^2$) non è valida

problema di Behrens-Fisher
(in R correzione di Welch)

Esempio 8.8. Un ricercatore che lavora alle dipendenze di una industria produttrice di lampadine elettriche afferma di aver trovato un nuovo tipo di filamento che prolunga la durata delle lampadine. Dato che il nuovo filamento è considerevolmente più costoso di quello attualmente in uso, l'industria intende, prima di adottarlo, avere il conforto di una verifica sperimentale. Viene allora formulata l'ipotesi nulla che la durata media, μ_1 , delle lampadine dotate del nuovo filamento sia uguale alla durata media, μ_2 , delle lampadine del vecchio tipo; come ipotesi alternativa si assume $H_1 : \mu_D = \mu_1 - \mu_2 > 0$. Per verificare l'ipotesi, vengono osservati due campioni dei due tipi di lampadine, entrambi di ampiezza 31. Le medie e le varianze dei due campioni risultano essere: $\bar{x} = 1.195,16$, $s_1^2 = 118,13$, $\bar{y} = 1.180,05$ e $s_2^2 = 124,34$. Ammettendo che le varianze delle popolazioni di origine dei due campioni siano uguali ($\sigma_1^2 = \sigma_2^2 = \sigma^2$), viene calcolata la quantità

$$s_c^2 = \frac{30 \times 118,13 + 30 \times 124,34}{60} = 121,23,$$

e quindi il rapporto

$$t = \frac{1.195,16 - 1.180,05}{\sqrt{121,23(1/31 + 1/31)}} = 5,40.$$

test a coda destra

evidenza fortissima

$$\gamma = P(T \geq t = 5,40) = 6 \times 10^{-7} \Rightarrow \text{si rifiuta } H_0$$

(durata maggiore delle nuove lampadine)

```
n1 <- 31  
n2 <- 31  
est1 <- 1195.16  
est2 <- 1180.05  
var1 <- 118.13  
var2 <- 124.34  
gl <- (n1+n2-2)  
est <- est1-est2  
var <- ((n1-1)*var1+(n2-1)*var2)/gl  
se <- sqrt(var)*sqrt((1/n1)+(1/n2))  
  
stat <- este
```

stesso script
usato per gli i.c

p-value (test a coda destra)

```
pt(stat, df= gl, lower.tail=FALSE)
```

Verifica di ipotesi sulla differenza tra due medie

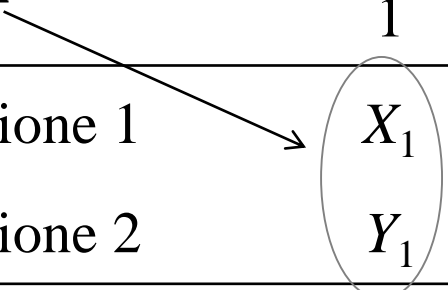
(popolazioni normali – dati appaiati)

(X_1, X_2, \dots, X_n) campione casuale da $N(\mu_1, \sigma_1^2)$
 (Y_1, Y_2, \dots, Y_n) campione casuale da $N(\mu_2, \sigma_2^2)$ (σ_1^2, σ_2^2 ignoti)

campioni non indipendenti (medesima ampiezza)

(es.: misura della pressione arteriosa sugli stessi soggetti con due strumenti diversi)

v.a. non indipendenti



	1	2	...	n
campione 1	X_1	X_2	...	X_n
campione 2	Y_1	Y_2	...	Y_n

campione di n coppie di dati

$$D_i = X_i - Y_i \quad (i = 1, \dots, n) \quad (\text{v.a. differenze}) \quad \Rightarrow (D_1, D_2, \dots, D_n)$$

$$E(D_i) = \mu_1 - \mu_2$$

$$\bar{D} = \frac{1}{n} \sum_{i=1}^n D_i \quad E(\bar{D}) = \mu_1 - \mu_2 \quad S_D^2 = \frac{1}{n-1} \sum_{i=1}^n (D_i - \bar{D})^2$$

sotto H_0

$$T = \frac{\bar{D}}{\sqrt{S_D^2/n}} \sim t(n-1) \quad T \text{ statistica test}$$

$$t = \frac{\bar{d}}{\sqrt{s_D^2/n}} \quad \text{valore osservato di } T$$

Esempio. Si vogliono confrontare 2 *materiali sintetici* (1 e 2) usati per produrre le suole di un certo tipo di scarpe per bambini (**test bilaterale**).

10 bambini calzano uno “speciale” paio di scarpe, una con suola di materiale 1, l'altra con suola di materiale 2 (*scelta casuale*).

I seguenti dati riportano l'*usura* delle suole registrata negli esperimenti.

materiali	bambini									
	1	2	3	4	5	6	7	8	9	10
1	13.2	8.2	10.9	14.3	10.7	6.6	9.5	10.8	8.8	13.3
2	14	8.8	11.2	14.2	11.8	6.4	9.8	11.3	9.3	13.6
d_i	-0.8	-0.6	-0.3	0.1	-1.1	0.2	-0.3	-0.5	-0.5	-0.3

$$\bar{d} = -0.41 \quad s_D^2 = \frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})^2 = 0.149$$

$$t = \frac{-0.41}{\sqrt{0,149/10}} = -3.35$$

test bilaterale

$$\gamma = P(T \leq t = -3,35) = 0,004$$

\Rightarrow si rifiuta H_0 $\alpha = 0.05$

$$\text{ovvero } 2 * 0,004 = 0,008$$

differenza $\mu_1 - \mu_2$ verosimilmente diversa da 0 (verosimilmente **negativa**)

(usura media maggiore per il *materiale 2* \Rightarrow migliore il *materiale 1*)

`t.test(materiale1, materiale2, alternative='two.sided', conf.level=.95, paired=TRUE)`

dati da imputare

Verifica di ipotesi sulla differenza tra due medie (popolazioni Bernoulliane – grandi campioni)

$$\mu_D = p_1 - p_2 \quad H_0 : \mu_D = 0 \quad \text{vs} \quad \begin{cases} H_1 : \mu_D > 0 \\ H_1 : \mu_D < 0 \\ H_1 : \mu_D \neq 0 \end{cases}$$

$$E(\bar{X} - \bar{Y}) = p_1 - p_2 \quad \text{Var}(\bar{X} - \bar{Y}) = \sigma_{\bar{X} - \bar{Y}}^2 = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

$$\boxed{\text{sotto } H_0} \quad \sigma_{\bar{X} - \bar{Y}}^2 = p(1-p) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

$$\boxed{\hat{p} = \frac{n_1 \bar{X} + n_2 \bar{Y}}{n_1 + n_2}} \quad \text{stimatore di } p \quad \Rightarrow \quad \hat{\sigma}_{\bar{X} - \bar{Y}}^2 = \hat{p}(1-\hat{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

sotto H_0

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\hat{p}(1 - \hat{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \xrightarrow{n_1, n_2 \rightarrow \infty} N(0,1)$$

per n_1 e n_2 sufficientemente elevati

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\hat{p}(1 - \hat{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim N(0,1) \quad (\text{in modo approssimato})$$

Z statistica test (test *normale* approssimato)

Esempio 8.9. Un partito politico ha commissionato ad un ente di ricerca demoscopica un'indagine sull'orientamento della popolazione in merito ad un prossimo referendum. Al partito interessa particolarmente sapere se l'opinione dei votanti è la stessa nelle Regioni chiave *A* e *B*. I dati raccolti sono i seguenti: nella Regione *A*, su 500 intervistati, 300 hanno dichiarato che voteranno sì; nella Regione *B*, su 600 intervistati, 340 hanno dichiarato che voteranno sì. L'ipotesi che interessa verificare è $H_0 : \mu_D = p_1 - p_2 = 0$ contro l'alternativa $H_1 : p_1 - p_2 \neq 0$.

$$\hat{p} = \frac{300 + 340}{500 + 600} = 0,58,$$

test bilaterale

$$z = (300/500 - 340/600) / \sqrt{(0,58)(0,42)(1/500 + 1/600)} = 1,11.$$

$$\gamma = P(Z \geq z = 1,11) = 0,13$$

\Rightarrow si accetta H_0

$$\text{ovvero } 2 * 0,13 = 0,26$$

differenza $p_1 - p_2$ verosimilmente pari a 0 (verosimilmente **positiva**)

(propensione al SI maggiore nella regione A)

prop.test(c(300,340),c(500,600))

```
n1 <- 500
n2 <- 600
n1.succ <- 300
n2.succ <- 340
est1 <- n1.succ/n1
est2 <- n2.succ/n2
est <- est1-est2
p.est <- (n1.succ+n2.succ)/(n1+n2)    #(est1*n1+est2*n2)/(n1+n2)
se <- sqrt(p.est*(1-p.est)*(1/n1+1/n2))
stat <- est/se
```

p-value

```
pnorm(stat, mean=0, sd=1, lower.tail=FALSE)
```

se stat ha valore negativo

```
pnorm(stat, mean=0, sd=1, lower.tail=TRUE)
```

test t per un campione – uso del comando t.test()

Il seguente prospetto riporta il contenuto, in litri, di 16 confezioni di latte prelevate a caso dalla linea di produzione:

1.03	1.01	0.94	0.99	1.10	1.03	0.89	1.07
1.12	1.06	1.02	0.94	0.99	1.09	1.01	1.12

Supponendo che i dati provengano da una **popolazione Normale**, verificare l'ipotesi nulla che il contenuto nominale di una confezione sia $\mu = 1$ (attraverso **tests di significatività** e/o **intervalli di confidenza**)

```
t.test(contenuto.latte, alternative='two.sided', mu=1, conf.level=0.95))
```

dati da imputare

testare valori *diversi* degli argomenti del comando

t.test() anche per un **grande campione**, senza l'ipotesi di **normalità**

test t per due campioni – uso del comando t.test()

Due appezzamenti, A e B, di uno stesso frutteto sono stati trattati con due diversi fertilizzanti. In ciascun appezzamento è stato scelto a caso un campione di piante misurando il peso della produzione, in *kg*. I dati sono i seguenti:

campione A 25.3 32.6 18.7 29.4

campione B 31.5 23.4 29.2 34.6 27.5

Supponendo che i dati provengano da due **popolazioni Normali e omoschedastiche**, verificare l'ipotesi nulla che il peso della produzione sia **uguale** nei due appezzamenti (attraverso **tests di significatività e/o intervalli di confidenza**)

```
t.test(peso.produzione~campione, alternative='two.sided', conf.level=0.95, var.equal=TRUE, data=Dataset)
```

dati da imputare

testare valori *diversi* degli argomenti del comando

se FALSE, correzione di Welch (default)

t.test() anche per **grandi campioni**, senza le ipotesi di **normalità e omoschedasticità**