COESIONE ITALIA 21-27 EMILIA-ROMAGNA

Cofinanziato dall'Unione europea

DATA LAB

GUARDA AVANTI

Big Data, nuove competenze per nuove professioni.

FORM INDUSTRIA EMILIA ROMAGNA

ifoa Sapere utile

REISS ROMOLI la passione della conoscenza

Università degli Studi di Ferrara

UNIMORE

UNIVERSITÀ DI PARMA

POLITECNICO MILANO 1863

UNIVERSITÀ CATTOLICA del Sacro Cuore

"Anticipare la crescita con le nuove competenze sui Big Data" Operazione Rif. PA 2023-19167/RER approvata con DGR n° 843 del 29 maggio 2023 e co-finanziata dal Fondo Sociale Europeo Plus 2021-2027 Regione Emilia-Romagna

ARTIFICIAL INTELLIGENCE & MACHINE LEARNING

Operazione Rif. PA 2023-19167/RER/10/1, "ANTICIPARE LA CRESCITA CON LE NUOVE COMPETENZE SUI BIG DATA", approvata dalla Regione Emilia-Romagna con DGR n° 843 del 29/05/2023 e co-finanziata dal Fondo Sociale Europeo Plus 2021-2027

DATA LAB

INTERVALLI DI CONFIDENZA
TEST D'IPOTESI

Intervallo di confidenza: media

Supponiamo di considerare la media campionaria $\bar{X}$ e assumiamo che

- la popolazione X ha una distribuzione normale, oppure consideriamo un campione grande con $n \geq 30$
- la varianza della popolazione σ^2 è nota.

Allora costruiamo un intervallo di confidenza per la media campionaria come:

$$\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

dove $z_{\frac{\alpha}{2}}$ è il valore critico della distribuzione normale standard che lascia nella coda destra una probabilità pari ad $\frac{\alpha}{2}$.

Intervallo di confidenza: media

L'intervallo di confidenza è costruito aggiungendo e sottraendo alla media campionaria un margine d'errore

$$z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}.$$

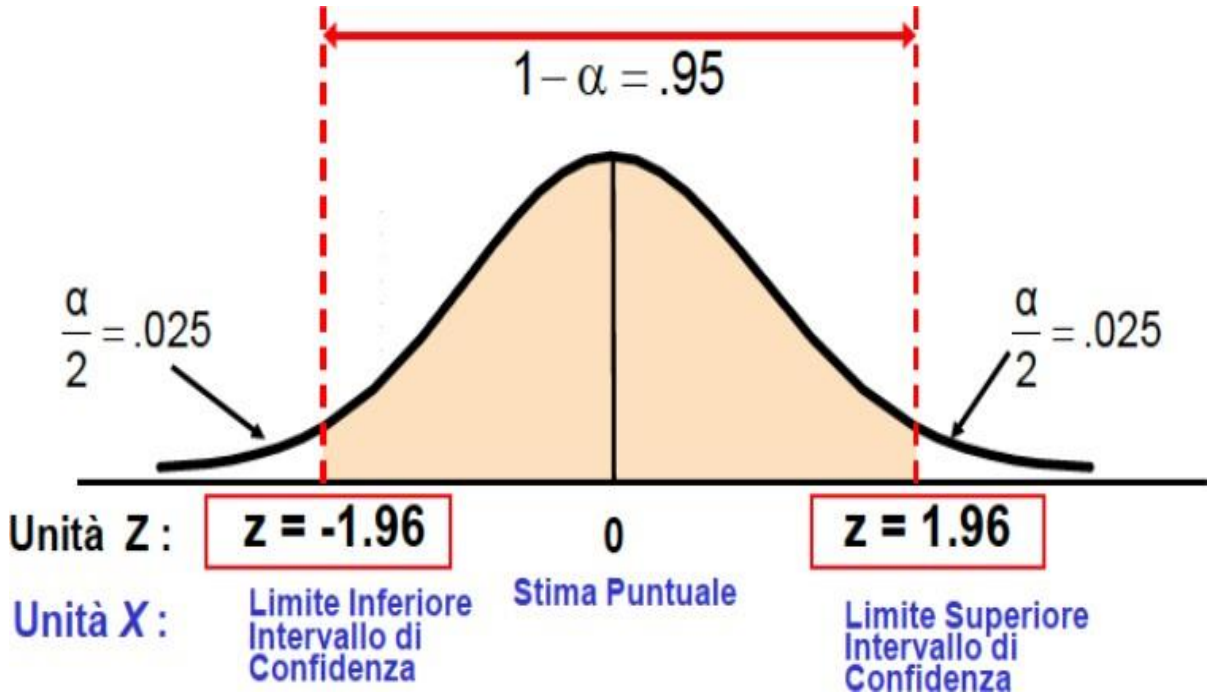
L'intervallo per la media campionaria è simmetrico e la sua ampiezza è il doppio del margine d'errore.

Il margine d'errore può essere ridotto se:

- la dimensione del campione aumenta;
- il livello di confidenza $(1-\alpha)$ diminuisce.

Distribuzione campionaria: media

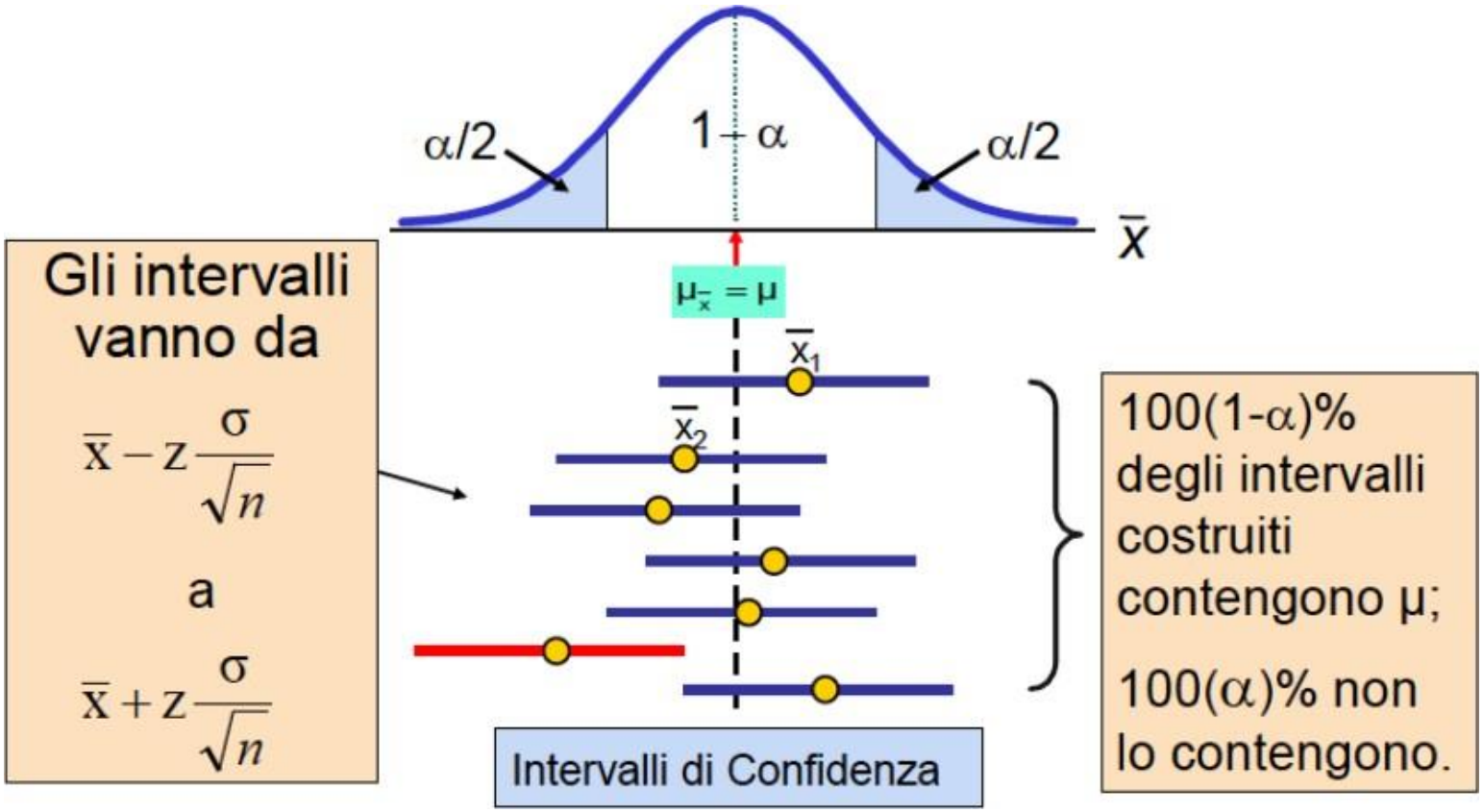
Ad esempio se $(1-\alpha) = 0.95$ allora:



Dove 1.96 viene ricavato dalla tavola normale standard. Si ha che se $(1-\alpha) = 0.95$ allora $\frac{\alpha}{2} = 0.025$ e $z_{0.025} = 1.96$.
Come per la proporzione, i livelli di confidenza comunemente usati sono: 90%, 95% e 99%.

Intervallo di confidenza: media

Costruiamo l'intervallo di confienza per la media, allora, considerata la distribuzione della media campionaria si ha:



Intervallo di confidenza: esempio

In una piantagione si sa che le misure del diametro (dbh) delle piante sono distribuite normalmente con $\sigma = 1.6$. Viene selezionato all'interno della stessa piantagione un campione di 16 alberi. La misura media del diametro per il campione era $\bar{x} = 12.6$ cm. Calcolare l'intervallo di confidenza al 95% e al 99% per la media sconosciuta della popolazione.
Calcoliamo la deviazione standard:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{1.6}{\sqrt{16}} = 0.4$$

Se $(1-\alpha) = 0.95$ allora $z_{\frac{\alpha}{2}} = 1.96$ e

$$12.6 - 1.96 \cdot 0.4 < \mu < 12.6 + 1.96 \cdot 0.4$$

$$11.82 < \mu < 13.38$$

Intervallo di confidenza: esempio

Se $(1-\alpha) = 0.99$ allora $z_{\frac{\alpha}{2}} = 2.58$ e

$$12.6 - 2.58 \cdot 0.4 < \mu < 12.6 + 2.58 \cdot 0.4$$

$$11.57 < \mu < 13.63.$$

Concludiamo che:

- Siamo confidenti al 95% che il vero diametro medio sia compreso tra 11.82 e 13.38 cm. Sebbene la vera media possa o meno essere inclusa in questo intervallo, il 95% degli intervalli costruiti con questo metodo conterrà il vero valor medio.
- Siamo confidenti al 99% che il vero diametro medio sia compreso tra 11.57 e 13.63 cm.

Intervallo di confidenza: media

Come ci si comporta se la varianza σ^2 non è nota ma va stimata dai dati?

In quel caso costruiamo l'errore standard a partire dalla varianza campionaria, ossia:

$$se = \frac{s}{\sqrt{n}}$$

Anche il margine d'errore sarà allora costruito a partire da questa quantità. Bisogna stabilire come vanno calcolati i valori critici da moltiplicare allo standard error in questo caso.
L'aver introdotto un'ulteriore stima nello standard error aumenta la possibilità d'errore e l'approssimazione con la normale non è più sufficiente.

Intervallo di confidenza: media

Consideriamo un campione aleatorio di n osservazione

- con media $\bar{X}$e deviazione standard S
- estratto da una distribuzione normale con media μ.

Allora la variabile

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

ha una distribuzione t di Student con $n - 1$ gradi di libertà (df).

Distribuzione t

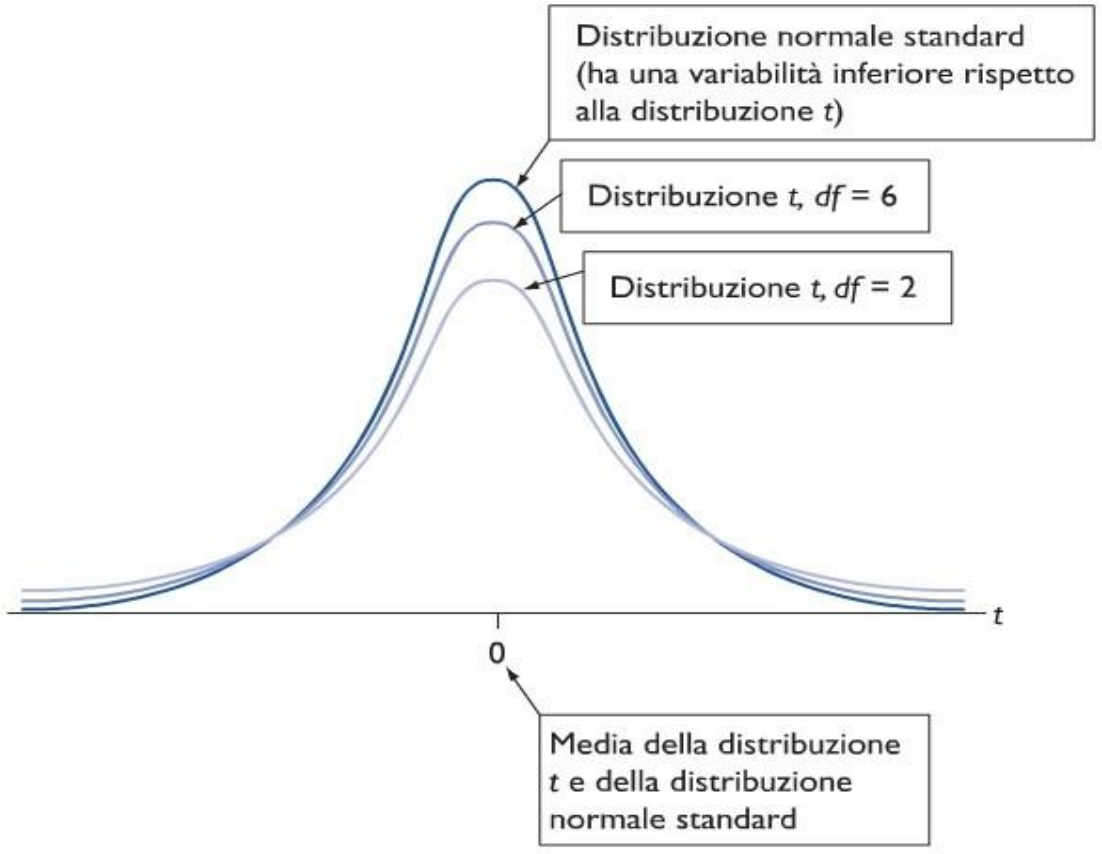
La distribuzione t assomiglia alla distribuzione normale standard. Ha una forma a campana ed è centrata sullo 0.

La sua deviazione standard è un po' più grande di 1, e il suo valore preciso dipende da una quantità detta gradi di libertà, indicata con df.

Quando vogliamo eseguire un'inferenza sulla media di una popolazione, i gradi di libertà sono $df = n-1$, vale a dire un valore pari alla dimensione del campione meno uno.

La distribuzione t ha le code più pesanti e una maggiore variabilità rispetto alla normale standard. Tuttavia al crescere dei gradi di libertà df , la sua forma si avvicina sempre di più a quella della normale standard. Quando df è 30 o più, le due distribuzioni sono quasi identiche.

Distribuzione t



La distribuzione *t* e la distribuzione normale standard. Al crescere dei gradi di libertà (*df*), la distribuzione *t* si avvicina sempre più alla distribuzione normale standard. Le due distribuzioni sono sostanzialmente identiche quando $df \geq 30$.

Intervallo di confidenza: media (σ^2 non nota

Assunzioni:

- Deviazione standard della popolazione non nota
- Popolazione distribuita normalmente
- Se la popolazione non è distribuita normalmente si utilizzano grandi campioni

In queste situazioni si fa ricorso alla distribuzione t di Student, e calcoliamo l'intervallo di confidenza come:

$$\bar{x} - t_{(\frac{\alpha}{2},n-1)} \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{(\frac{\alpha}{2},n-1)} \frac{s}{\sqrt{n}}$$

Dove $t_{(\frac{\alpha}{2},n-1)}$ è il valore critico della distribuzione t con $n-1$ gradi di libertà che lascia nella coda destra una probabilità pari a $\frac{\alpha}{2}$.

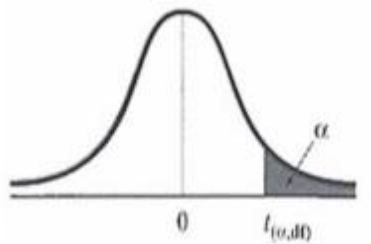
Disribuzione t

$t_{(\frac{\alpha}{2},n-1)}$ è tale che

$$P\left(t_{(n-1)} > t_{(\frac{\alpha}{2},n-1)}\right) = \frac{\alpha}{2}.$$

Ad esempio, se $\frac{\alpha}{2} = 0.025$ e abbiamo un campione con $n = 10$ allora $t_{(\frac{\alpha}{2},n-1)} = t_{(0.025,9)} = 2.262$

Tavola della distribuzione T di Student



Gradi di libertà	Area nella coda di destra								
	0.1	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	3.078	6.314	12.706	15.894	31.821	63.656	127.321	318.289	636.578
2	1.886	2.920	4.303	4.849	6.965	9.925	14.089	22.328	31.600
3	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.214	12.924
4	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.894	6.869
6	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587

La tavola contiene all'interno i valori t e non le probabilità.

Intervallo di confidenza: esempio

Un campione casuale di $n = 25$ ha $\bar{x} = 50$ e $s = 8$. Determinare un intervallo di confidenza al 95% per μ

- **df** = $n - 1 = 24$, allora $t_{n-1,\alpha/2} = t_{24,.025} = 2.064$

L'intervallo di confidenza è

$$\bar{x} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$$

$$50 - (2.064)\frac{8}{\sqrt{25}} < \mu < 50 + (2.064)\frac{8}{\sqrt{25}}$$

$$46.698 < \mu < 53.302$$

Intervallo confidenza: esempio 2

Supponiamo che nell'esempio sulla piantagione visto sopra, la deviazione standard per dbh non sia nota ma stimata dal campione come $s = 2.09$ cm. Ricordiamo che la misura media del diametro per il campione era $\bar{x} = 12.6$ cm. Calcolare l'intervallo di confidenza al 95% e al 99% per la media sconosciuta della popolazione.

$$se_{\bar{X}} = \frac{S}{\sqrt{n}} = \frac{2.09}{\sqrt{16}} = 0.5225$$

Se $(1-\alpha) = 0.95$ allora $t_{(\frac{\alpha}{2},n-1)} = t_{(0.025,15)} = 2.13$ e

$$12.6 - 2.13 \cdot 0.5225 < \mu < 12.6 + 2.13 \cdot 0.5225$$

$$11.49 < \mu < 13.71$$

Intervallo confidenza: esempio 2

Se $(1-\alpha) = 0.99$ allora $t_{(\frac{\alpha}{2},n-1)} = t_{(0.005,15)} = 2.95$ e

$$12.6 - 2.95 \cdot 0.5225 < \mu < 12.6 + 2.95 \cdot 0.5225$$

$$11.06 < \mu < 14.14$$

Test d'ipotesi

Il test d'ipotesi o test di significatività è il secondo grande metodo per eseguire inferenze statistiche relative a una popolazione.

Come un intervallo di confidenza per stimare un parametro, il test d'ipotesi impiega la probabilità per trovare un modo di quantificare quanto sia plausibile il valore di un parametro, controllando al tempo stesso la probabilità di eseguire una inferenza non corretta.

Ipotesi statistica

Un'ipotesi è una affermazione (assunzione) circa un parametro della popolazione. In genere nell'affermazione si dichiara che un parametro assume un particolare valore numerico, oppure che è compreso in un certo intervallo di valori:

- media della popolazione

  Esempio: Il tempo di vita medio di una tartaruga marina Caretta Caretta è di 40 anni, oppure il tempo di vita medio di una Caretta Caretta è compreso fra i 30 e i 60 anni..

- proporzione della popolazione

  Esempio: Nel Mediterraneo la proporzione di tartarughe marine tartarughe che muoiono dopo essere state catturate da attrezzi di pesca risulta $p = 0.26$.

L'ipotesi si riferisce sempre al parametro della popolazione e mai alla statistica campionaria.

Test d'ipotesi: esempio

Si vuole verificare se le lattine di caffè confezionate automaticamente da una ditta contengono in media il peso dichiarato $\mu = 250$ g. A tale scopo si estrae un campione di 30 lattine, se ne pesa il contenuto e si calcola il peso medio, per stabilire se il peso medio differisce da 250g

Test d'ipotesi

La verifica delle ipotesi statistiche inizia con la definizione del problema in termini di ipotesi sul parametro oggetto di studio.

Per prima cosa si stabilisce l'ipotesi da sottoporre a test, detta ipotesi nulla, indicata con H_0. Con l'ipotesi nulla si afferma che il parametro assume un particolare valore.

Oltre all'ipotesi nulla occorre specificare anche un'adeguata ipotesi alternativa, indicata con H_a, ossia un'affermazione che contraddice l'ipotesi nulla. Si afferma che il valore del parametro è uno tra quelli presenti in un certo intervallo di valori alternativi.

Test d'ipotesi: esempio

Si vuole verificare se le lattine di caffè confezionate automaticamente da una ditta contengono in media il peso dichiarato $\mu = 250$ g. A tale scopo si estrae un campione di 30 lattine, se ne pesa il contenuto e si calcola il peso medio, per stabilire se il peso medio differisce da 250g.

Il tutto viene tradotto come:

$$H_0 : \mu = 250$$
$$H_a : \mu /= 250$$

Test d'ipotesi

- In un test d'ipotesi, si presume che l'ipotesi nulla sia vera a meno che i dati non producano forti evidenze contro di essa. L'ipotesi nulla è posta con lo scopo di essere screditata, quindi ciò che si oppone alla conclusione che il ricercatore cerca di raggiungere rappresenta l'ipotesi nulla.
- Il ricercatore afferma che a essere vera è l'ipotesi alterna- tiva. Nell'ipotesi alternativa viene messo ciò che si spera o ci si aspetta di poter concludere come risultato del test.
- Nell'ipotesi nulla deve sempre comparire un segno di uguaglianza ($=, \geq, \leq$).
- Le due ipotesi sono complementari, ossia considerate in- sieme esauriscono tutte le possibilità riguardanti il valore che può assumere il parametro in esame.

Si può fare un'analogia con quanto avviene in un'auladi tribunale, quando una giuria deve decidere sulla colpevolezza o sull'innocenza di un imputato.

L'ipotesi nulla, che corrisponde all'assenza di un effetto, è che l'imputato sia innocente.

L'ipotesi alternativa è che l'imputato sia colpevole.

La giuria considera l'imputato innocente a meno che chi l'accusa non produca una forte evidenza a favore della sua colpevolezza. L'accusatore deve convincere la giuria che l'imputato sia colpevole.

Il contenuto dichiarato dal produttore delle bottiglie di acqua minerale di una certa marca è 920ml. Un'associazione di consumatori sostiene che in realtà le bottiglie contengono in media una quantità inferiore di acqua.

In questo caso le ipotesi sono:

$$H_0 : \mu \geq 920$$
$$H_a : \mu < 920$$

Le possibili conclusioni per un test di ipotesi sono:

1) se l'ipotesi nulla H_0 è rifiutata, si conclude che l'ipotesi alternativa H_a è probabilmente vera;

2) se l'ipotesi nulla non è rifiutata si conclude che i dati non forniscono una sufficiente evidenza per sostenere l'ipotesi alternativa.

E' importante sottolineare che con la verifica delle ipotesi, e in generale con l'inferenza statistica, non si arriva alla dimostrazione di un'ipotesi; si ha solo un'indicazione del fatto che l'ipotesi sia o meno avvalorata dai dati disponibili.

Statistica test

Dopo aver formulato le ipotesi, occorre specificare quale risultato del campione porterà al rifiuto dell'ipotesi nulla.

Per poterlo fare si definisce una statistica test.

Il parametro al quale le ipotesi fanno riferimento, viene stimato puntualmente a partire dai dati. Una statistica test descrive quanto la stima puntuale si colloca lontano dal valore del parametro specificato nell'ipotesi nulla. Può assumere tanti valori quanti sono i possibili campioni estraibili dalla popolazione.

In generale la distanza descritta dalla statistica test è misurata come numero di errori standard intercorrenti tra la stima puntuale e il parametro.

Statistica test

La distribuzione di campionamento della statistica test ˙, di solito, una distribuzione nota, come la distribuzione normale o la distribuzione t, e ricorriamo a queste distribuzioni per sottoporre a verifica un'ipotesi nulla.

Utilizzando le proprietà della distribuzione di campionamento della statistica soggetta a test, si può identificare un intervallo di valori di quella statistica che verosimilmente non si presentano se l'ipotesi nulla è vera.

Statistica test

La distribuzione di campionamento della statistica test è divisa in due regioni:

1) una **regione di rifiuto** che corrisponde all'insieme dei valori di una statistica test che conducono al rifiuto dell'ipotesi nulla;

2) una **regione di accettazione** che corrisponde all'insieme dei valori di una statistica testche portano invece all'accettazione dell'ipotesi nulla.

Le due regioni sono, delimitate da uno o più valori, detti valori critici.
Se la statistica test, in base ai dati del campione, assume un valore che cade nella regione di rifiuto, l'ipotesi nulla deve essere rifiutata; se al contrario il valore cade nella regione di accettazione, l'ipotesi nulla non può essere rifiutata.

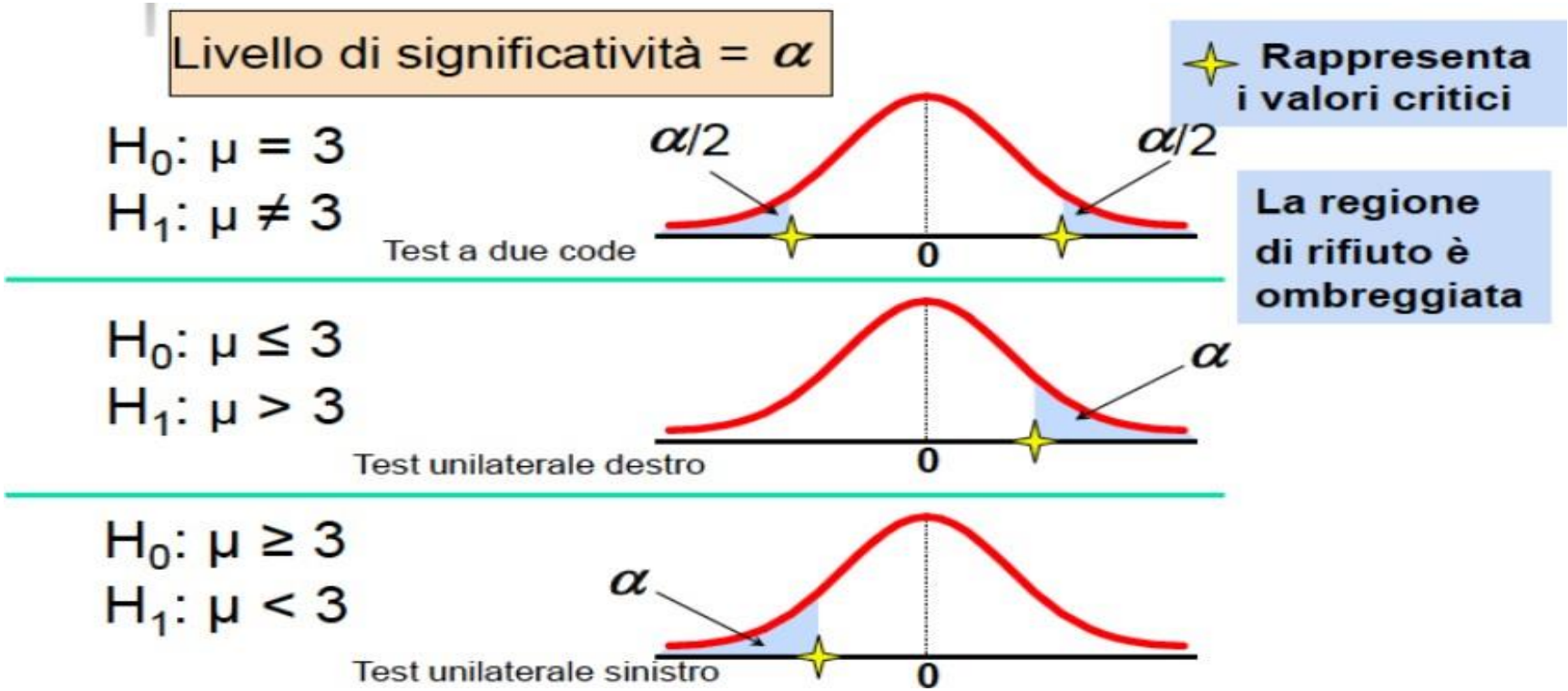
Statistica test

I test di ipotesi possono essere classificati in due gruppi:

1) **test a una coda** (o test unilaterale): la regione di rifiuto è costituita da un intervallo. per un test a una coda nell'ipotesi alternativa compare uno dei segni $>$ oppure $<$.
2) **test a due code** (o test bilaterale): la regione di rifiuto è costituita da due intervalli, ossia da due code della distribuzione. Per un test a due code nell'ipotesi alternativa compare il segno $/=$.

Significatività

I valori della statistica test che definiscono la regione di rifiuto sono definiti a partire dal livello di significatività α del test.
I livelli di significatività usualmente considerati sono $\alpha = 0.1$, $\alpha = 0.05$ o $\alpha = 0.01$.



COESIONE ITALIA 21-27
EMILIA-ROMAGNA

Cofinanziato dall'Unione europea

Regione Emilia-Romagna

DATA LAB

GUARDA AVANTI

Big Data, nuove competenze per nuove professioni.

www.bigdata-lab.it

FORM INDUSTRIA EMILIA ROMAGNA

ifoa Sapere utile

REISS ROMOLI la passione della conoscenza

Università degli Studi di Ferrara

UNIMORE

UNIVERSITÀ DI PARMA

POLITECNICO MILANO 1863

UNIVERSITÀ CATTOLICA del Sacro Cuore