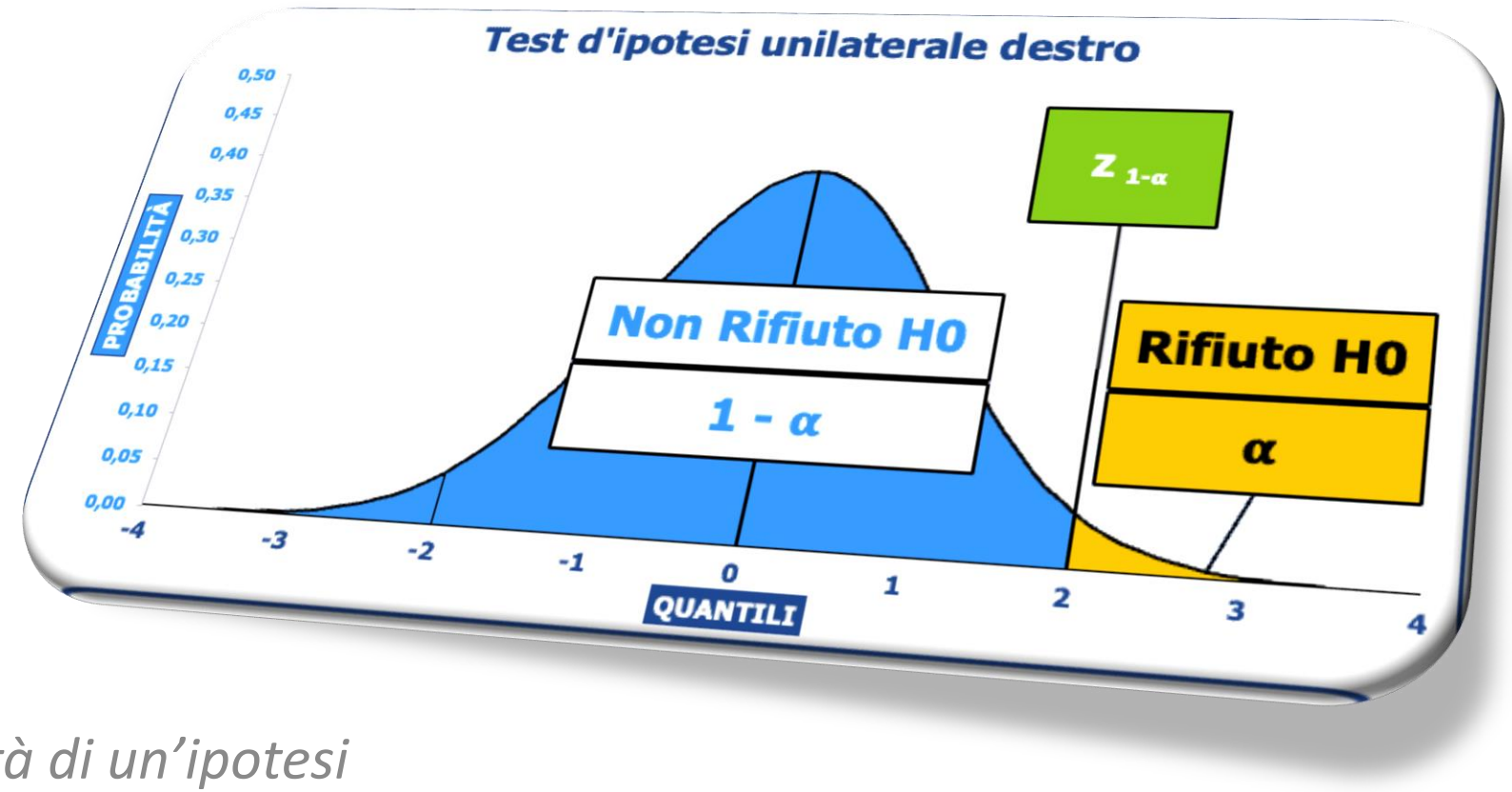
Analisi Dati con Excel

Giovanni Della Lunga

giovanni.dellalunga@gmail.com

La prima regola di ogni tecnologia è che l'automazione applicata ad un'operazione efficiente ne aumenterà l'efficienza. La seconda è che l'automazione applicata ad un'operazione inefficiente ne aumenterà l'inefficienza.

Bill Gates



Test di Ipotesi

Come verificare la veridicità di un'ipotesi

Che cosa sono le ipotesi e i test di ipotesi?

- Un’ipotesi è una formulazione di una congettura, un’assunzione sulla popolazione che deve essere verificata e può essere vera o falsa
- Le conclusioni che traiamo sulla popolazione vengono valutate esaminando un campione estratto da quella popolazione
- Le ipotesi fatte sono relative a parametri di quella popolazione
- Il test di ipotesi è la procedura che ci permette di stabilire se l’ipotesi è vera o falsa

Ipotesi

- Nei test di ipotesi le ipotesi sono due: l'ipotesi nulla e l'ipotesi alternativa.
- L'ipotesi nulla è quella di uguaglianza o mancanza dell'effetto, quella alternativa è quella di differenza/diversità

Es. Se vogliamo valutare se una moneta è truccata (perciò la probabilità che esca testa non è uguale alla probabilità che esca croce) allora formuleremo l'ipotesi come:

Ipotesi nulla: $p_{testa} = p_{croce}$ Ipotesi alternativa: $p_{testa} \neq p_{croce}$

Oppure

Ipotesi nulla: $p_{testa} = 0.5$ Ipotesi alternativa: $p_{testa} \neq 0.5$

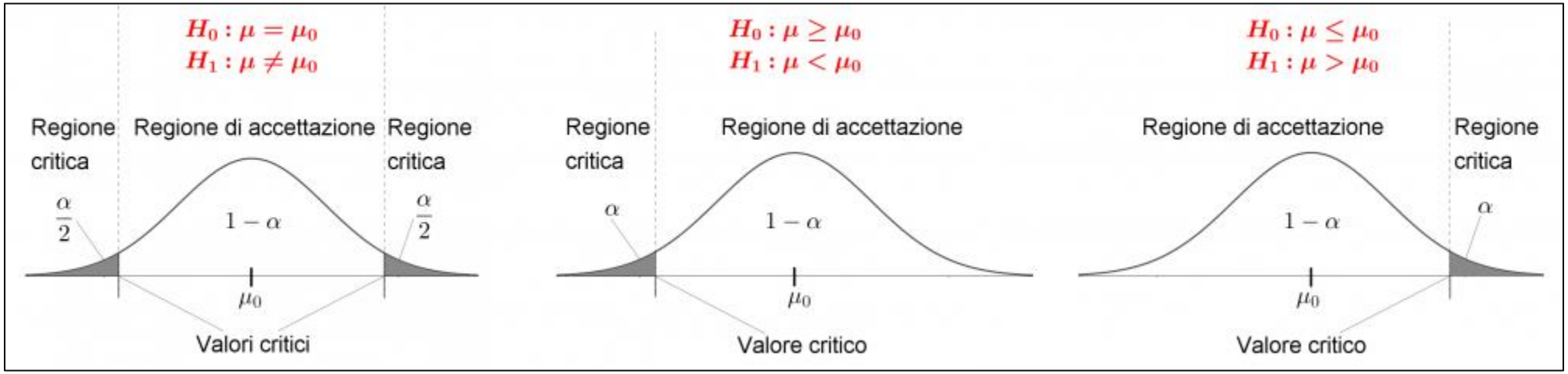
Ipotesi

bilaterale

$$\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu \neq \mu_0 \end{cases}$$

unilaterale

$$\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu < \mu_0 \end{cases} \qquad \begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu > \mu_0 \end{cases}$$



Esiti

H_0 è l'ipotesi nulla
H_1 è l'ipotesi alternativa

→ Preferiamo rifiutare l'ipotesi nulla a favore di quella alternativa

Oppure

H_0 è l'ipotesi nulla
H_1 è l'ipotesi alternativa

→ Non rifiutare H_0 non significa che la condizione descritta dall'ipotesi nulla è vera, ma significa che non abbiamo sufficiente evidenza per rifiutarla!

La Statistica Test

- Per stabilire se rifiutare o non rifiutare l'ipotesi nulla dobbiamo calcolare la statistica test
- In generale la statistica test è calcolata come:

$$T = \frac{stima - valore\ ipotizzato}{errore}$$

- Ogni statistica test ha una distribuzione nota

La statistica test della media

- Consideriamo la seguente ipotesi: «la pressione arteriosa media negli anziani è maggiore di 120»

- **Come è strutturato il test di ipotesi?**

$$\begin{cases} H_0: \mu = 120 \\ H_1: \mu > 120 \end{cases}$$

Cos’è l’errore standard campionario

Definizione

- La formula $\epsilon = \frac{\sigma}{\sqrt{n}}$ rappresenta l'**errore standard della media campionaria**.
- Questa formula è fondamentale in statistica per capire quanto la media di un campione $\bar{X}$ è probabile che differisca dalla media vera della popolazione μ.

  - $\boldsymbol{\sigma}$: È la deviazione standard della popolazione, che misura quanto i dati della popolazione si discostano in media dalla media della popolazione μ.
  - **n** : È il numero di osservazioni o la dimensione del campione.

Cos’è l’errore standard campionario

Interpretazione della Formula

- L'errore standard $\epsilon = \frac{\sigma}{\sqrt{n}}$ è una misura di dispersione che indica quanto la media campionaria è probabile che si discosti dalla media della popolazione:
  - **σ**: Se la deviazione standard è alta, significa che i dati sono molto dispersi rispetto alla media. Questo porta ad un errore standard più alto.
  - $\sqrt{n}$: Il denominatore della formula è la radice quadrata del numero di dati nel campione. Aumentare la dimensione del campione riduce l'errore standard, rendendo la media campionaria una stima più precisa della media della popolazione.

N. B. l'errore standard della media si riduce con l'aumentare della dimensione del campione ma solo come la radice quadrata del campione. In particolare, quando raddoppiamo la dimensione del campione, l'errore standard non si dimezza, ma si riduce di un fattore pari alla radice quadrata di 2.

La statistica test della media

Che cosa facciamo per condurre la verifica di questa ipotesi??

1. Estraggo un campione dalla popolazione
2. Su tutte le unità valuto la pressione arteriosa
3. Calcolo la statistica test della media che è data da:

$$T = \frac{stima\ -valore\ ipotizzato}{errore} = \frac{\bar{x} - 120}{\sigma/\sqrt{n}}$$

La statistica test per la media ha una **distribuzione Gaussiana** di media 0 e deviazione standard 1

La regione di accettazione e rifiuto

» Per capire se rifiutare o non rifiutare l'ipotesi nulla devo vedere se il valore assunto dalla statistica test appartiene alla regione di accettazione o alla regione di rifiuto

» La regione di accettazione è l'insieme di valori per cui non rifiuto H_0

» La regione di rifiuto è l'insieme di valori per cui rifiuto H_0

Livello di significatività

» L’insieme dei valori che appartengono alla regione di rifiuto/di accettazione è definito dal livello di significatività (α)

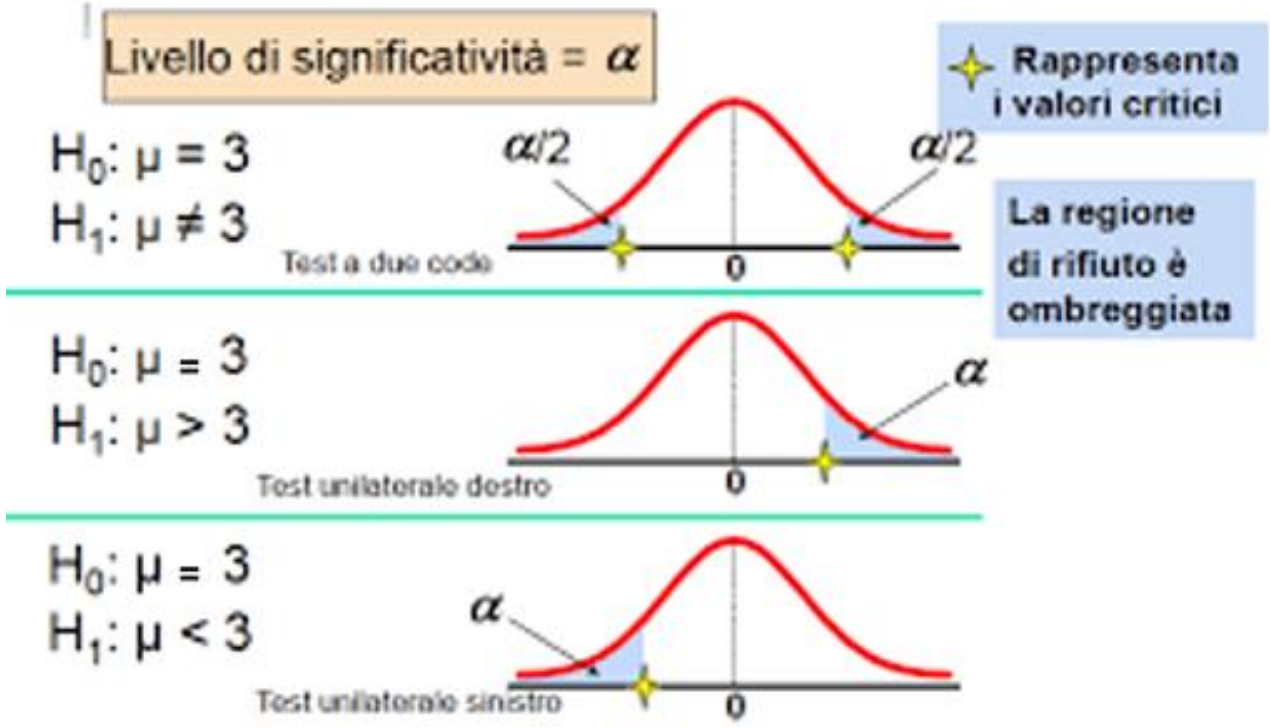
» Il livello di significatività è una probabilità, cioè la probabilità di rifiutare l’ipotesi nulla quando questa è vera

» α viene detto anche errore del 1°tipo, e in genere assume valore 0.05 ma può assumere anche valore 0.10 o 0.01 in base all’entità dell’errore che siamo disposti a commettere

Livello di significatività

» La collocazione di α segue l'ipotesi alternativa, perciò se nell'ipotesi alternativa troviamo il ≠ posizioneremo metà errore nella coda sinistra e metà nella coda destra, se è presente il < allora α è tutto a sinistra

» Se il valore della statistica test ricade nell'area bianca allora non rifiuto l'ipotesi nulla, se ricade nella zona celeste rifiuto l'ipotesi nulla



p-value

» Si calcola il p-value come l'area sotto la curva della distribuzione della statistica di test che è pari o più estrema rispetto al valore osservato della statistica di test.

» Per un test a due code, il p-value è la somma delle aree nelle code della distribuzione.

Importante:

Pr (osservazione | ipotesi) ≠ Pr (ipotesi | osservazione)

La probabilità di osservare un risultato data per vera una certa ipotesi non è *equivalente* alla probabilità che l'ipotesi sia vera dato un risultato osservato.

Usando il valore-p come "punteggio" si commette un grave errore logico **la fallacia del condizionale trasposto.**

Il **valore-p** (area verde) è la probabilità di un risultato osservato (o più estremo) supponendo vera l'ipotesi nulla.

p-value

P-Value Approach

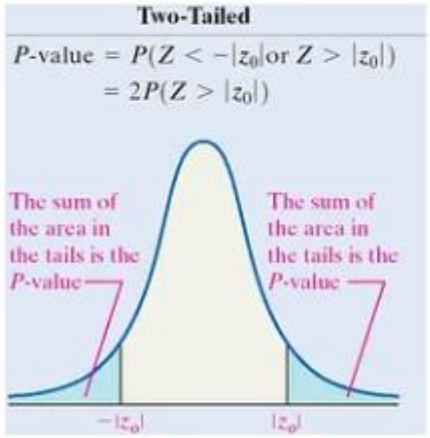
Assume that the null hypothesis is true.

The P-Value is the probability of observing a sample mean that is as or more extreme than the observed.

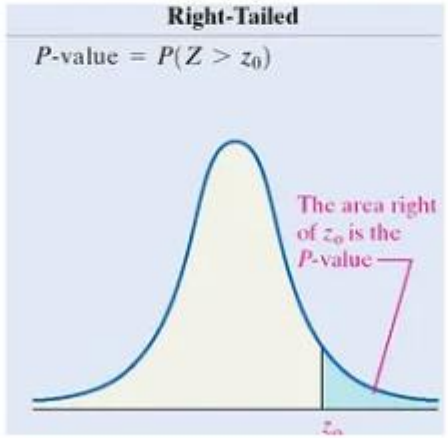
How to compute the P-Value for each type of test:

Step 1: Compute the test statistic $z_0 = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$

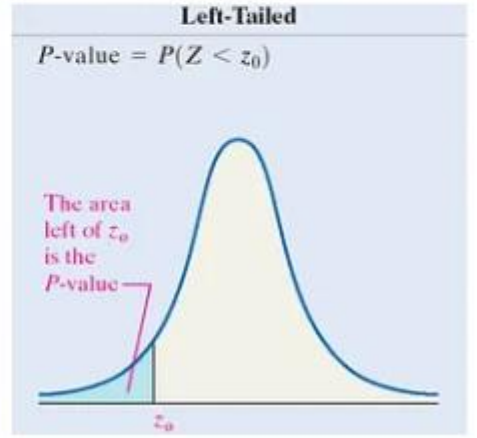
Two-tail



Right Tail



Left Tail



p- value is commonly used for hypothesis testing to evaluate the similarity of difference in a data set. It can be defined in the following ways-It is the degree of confidence with which we can reject the null hypothesis (H0),It is the measure of evidence that we have against the null hypothesis (H0)

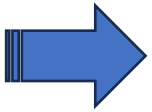
Esempio 1 – Test Z per una Media

» Scenario:

- Supponiamo di voler testare se la media di una popolazione è 50.
- Campione: n = 36, <x> = 52, σ = 10.

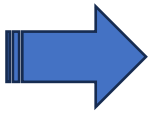
» Passi in Excel:

- Calcolare la statistica Z



```
= (MEDIA - 50) / (DEV.STANDARD / RADQ(n))
= (52 - 50) / (10 / RADQ(36))
```

- Calcolare il p-value (test a due code)



```
=2 * (1 - NORM.DIST(2, 0, 1, VERO))
```

Esempio 1 – Test Z per una Media

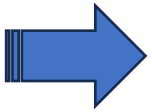
» Scenario:

- Supponiamo di voler testare se la media d
- Campione: n = 36, <x> = 52, σ = 10.

IMPORTANTE! SI NOTI CHE IN QUESTO ESEMPIO SI IPOTIZZA DI CONOSCERE LA STANDARD DEVIATION DELLA POPOLAZIONE

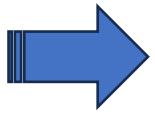
» Passi in Excel:

- Calcolare la statistica Z



```
= (MEDIA - 50) / (DEV.STANDARD / RADQ(n))
= (52 - 50) / (10 / RADQ(36))
```

- Calcolare il p-value (test a due code)



```
=2 * (1 - NORM.DIST(2, 0, 1, VERO))
```

Esempio 2 – Test T per una Media

» Scenario:

- Supponiamo di voler testare se la media di una popolazione è 50.
- Campione: n = 25, <x> = 52, s = 10.

ATTENZIONE! Sembra lo stesso esempio di prima, ma non lo è! Riuscite a vedere la differenza?

Esempio 2 – Test T per una Media

» Scenario:

- Supponiamo di voler testare se la media d
- Campione: n = 25, <x> = 52, s = 10.

IN QUESTO CASO NON CONOSCIAMO LA VARIANZA DELLA POPOLAZIONE MA ABBIAMO SOLO UNA STIMA BASATA SUL CAMPIONE!!!

ATTENZIONE! Sembra lo stesso esempio di prima, ma non lo è! Riuscite a vedere la differenza?

Esempio 2 – Test T per una Media

In questo caso quindi la statistica contiene due variabili aleatorie <x> ed s

$$\frac{\langle x \rangle - 50}{s/\sqrt{25}}$$

In generale si può dimostrare che il rapporto

$$\frac{\langle x_n \rangle - \mu}{S_n/\sqrt{n}}$$

È distribuito come una t di Student con n gradi di libertà.

Esempio 2 – Test T per una Media

» Scenario:

- Supponiamo di voler testare se la media di una popolazione è 50.
- Campione: n = 25, <x> = 52, s = 10.

» Passi in Excel:

- Calcolare la statistica t:

```
= (MEDIA - 50) / (DEV.STANDARD / RADQ(n))
= (52 - 50) / (10 / RADQ(25))
```

- Calcolare il valore P (per test a due code):

```
=2 * (1 - T.DIST(1,96, n-1, VERO))
```

- Decisione: Se il valore di P è minore di 0.05 rifiutare H_0

Riassumendo…

Passi per Eseguire un Test di Ipotesi

1. Formulare H0 e H1.
2. Scegliere il livello di significatività (α).
3. Calcolare la statistica del test.
4. Determinare la regione di rifiuto.
5. Calcolare il valore P.
6. Prendere la decisione:
    1. Rifiutare H0 se il valore P è minore di α.
    2. Non rifiutare H0 se il valore P è maggiore di α.