CDL in MEDICINA & CHIRURGIA

Statistica Medica

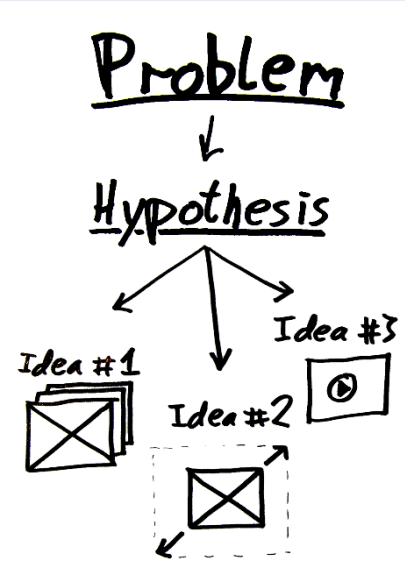
gbarbati@units.it

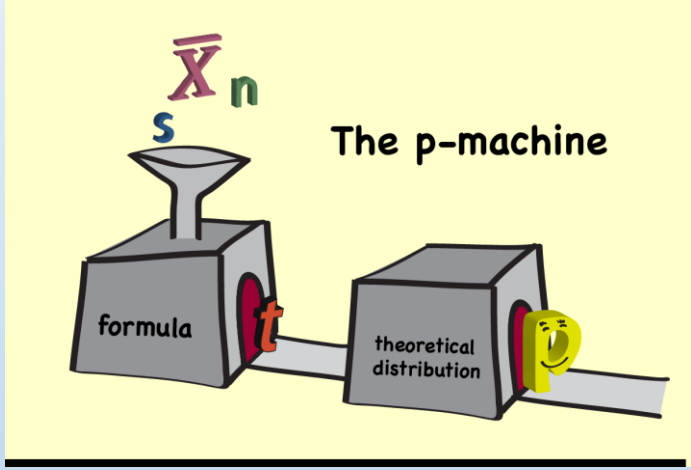
A.A. 2024-25

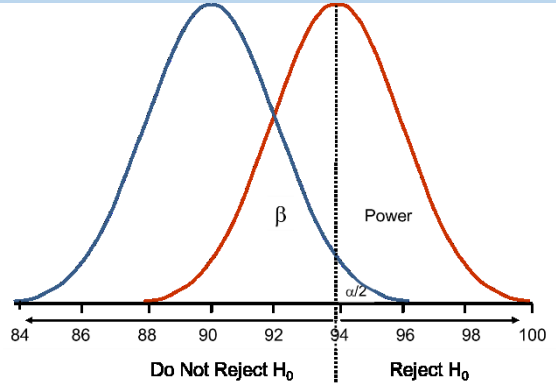
UNITÀ DI BIOSTATISTICA

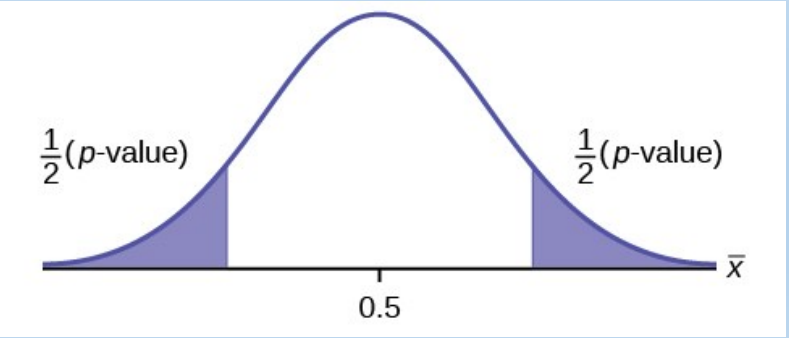
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

Il meccanismo del test di ipotesi









Introduzione al Test di Ipotesi: concetti di base

- Si ha una specifica ipotesi su un certo fenomeno nella popolazione che si vuole verificare (ipotesi **nulla** *vs* ipotesi **alternativa**)

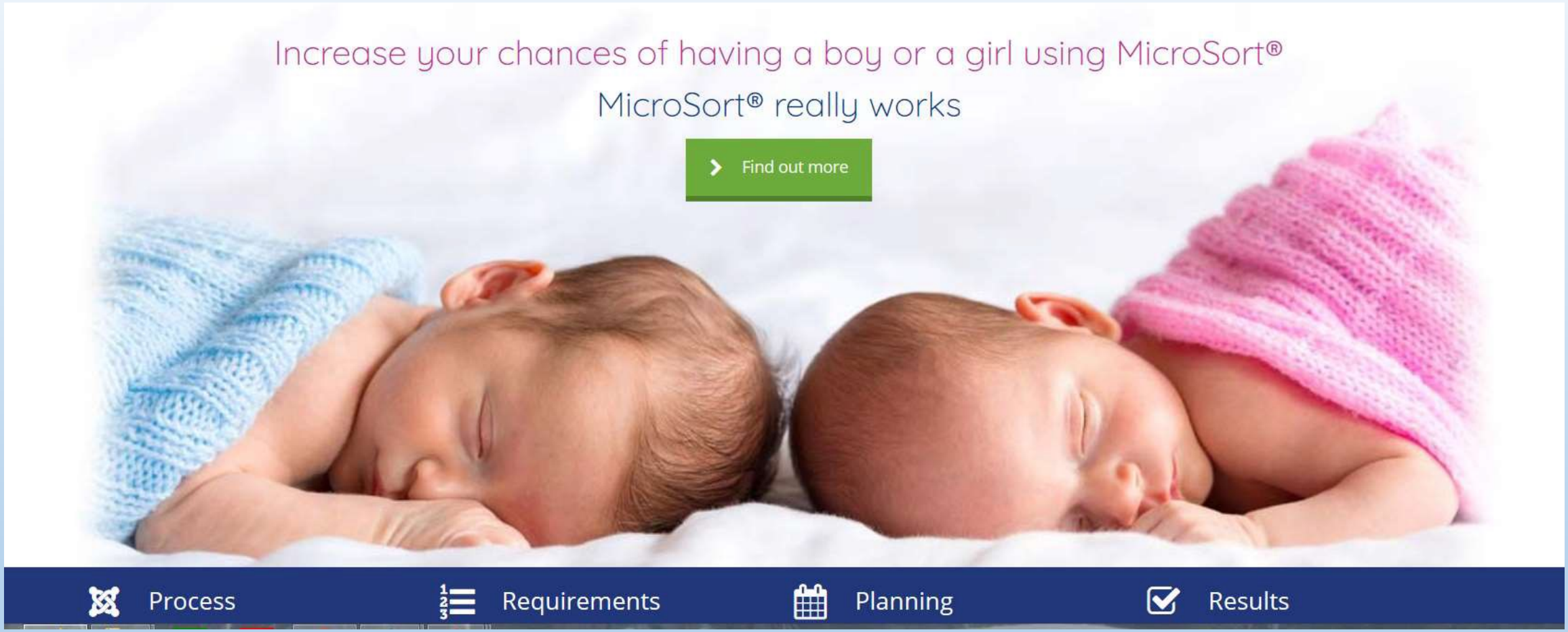
- Si raccolgono dei dati attinenti al problema (dati campionari) [di numerosità ***sufficiente…***]

- Si combinano i pezzi di informazione per ottenere una «**misura di evidenza**» a favore o contro l'ipotesi

- Si decide se c'è abbastanza **evidenza** dai dati per ***rigettare*** l'ipotesi nulla

[approccio **frequentista**]

Example: Does the **MicroSort** Method of Gender Selection Increase the Likelihood That a Baby Will Be a Girl?



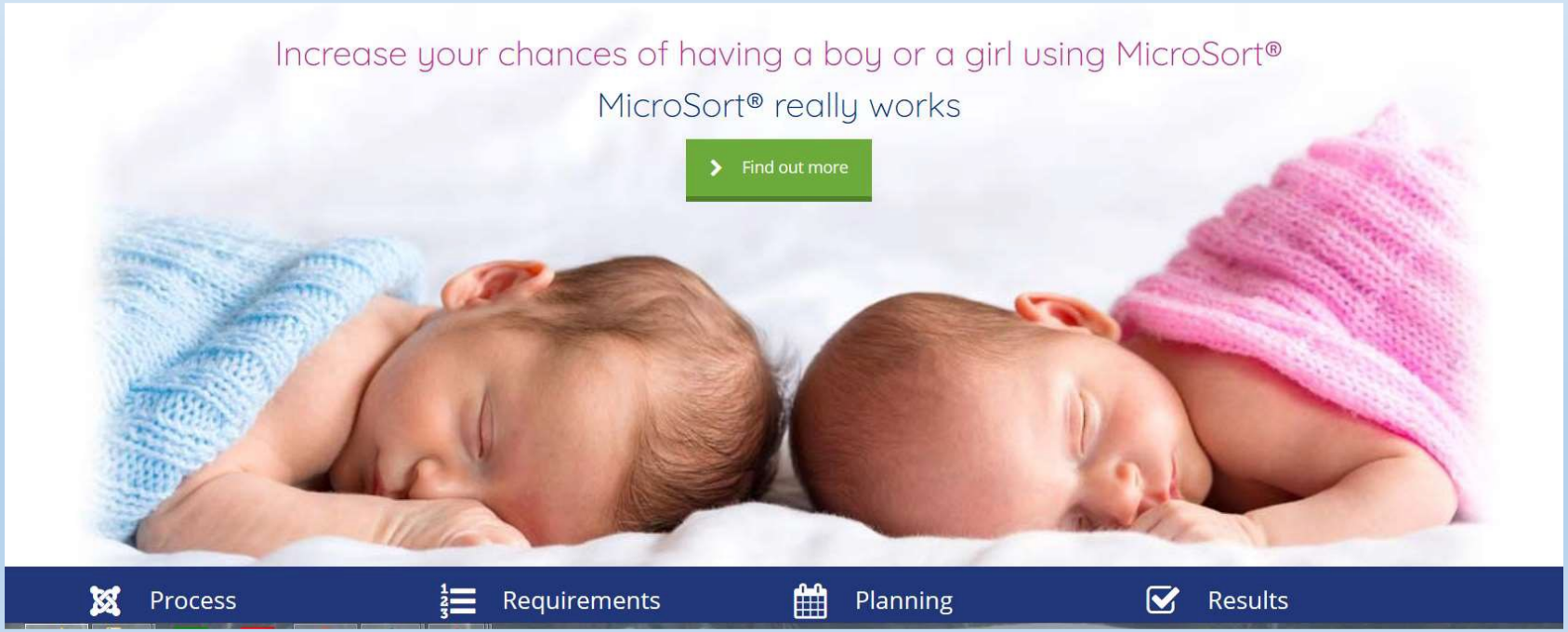
The Genetics & IVF Institute claims that its XSORT method allows couples **to increase** the probability of having a baby girl.

Preliminary results:

- **14** babies born to couples using the XSORT method of gender selection
- **13** of the babies were girls.

Under normal circumstances with no special treatment, girls occur in about 50% of births.
(Actually, the current birth rate of girls is 48.8%, but we will use 50% to keep things simple.)

Can we *actually support* the claim that the XSORT technique is effective in increasing the probability of a girl?



Test di Ipotesi: tentativo di analogia

Una persona viene accusata di un reato: viene arrestata e portata davanti ad un tribunale

(a) Ipotesi nulla (H_0): presunzione di innocenza
(b) Ipotesi alternativa (H_1): colpevolezza dell'indagato
(c) Si raccolgono informazioni (**evidenze=dati**) sulla questione
(d) Il giudice **valuta** gli indizi raccolti
(e) Il giudice decide se incolpare o meno l'indagato

Il principio fondamentale:
Evidenza non sufficiente -> Verdetto di non colpevolezza (in dubio pro reo)

Purtroppo: può succedere che un innocente vada in galera,
così come un colpevole sia lasciato libero…

A result of **8 girls out of 14** (or 57.1%) could easily occur *by chance* under normal circumstances with no treatment, so 8 is not significantly high.

The actual result of **13 girls** (or 92.9%) appears to be *significantly* high…

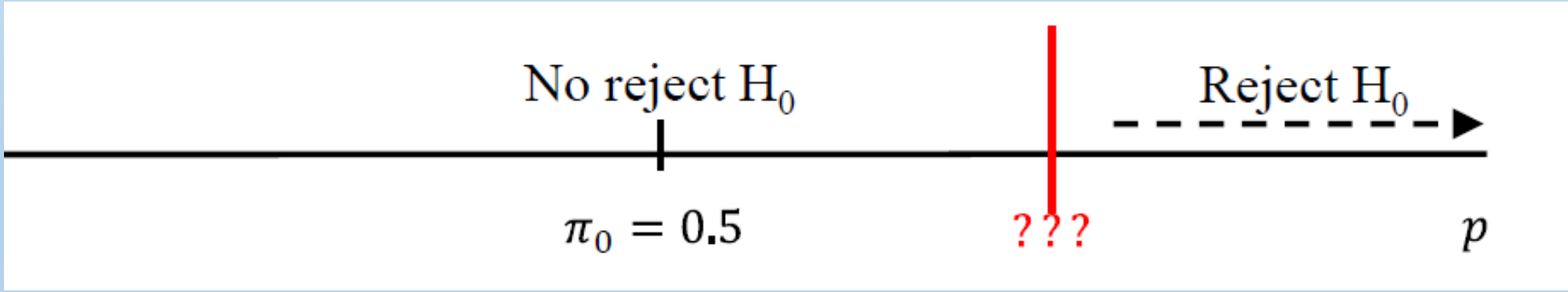
H_0: $\pi = \pi_0 = 0.5$ (null hypothesis: e.g. *the probabilty to get a girl is 50%*)

H_1: $\pi > 0.5$ (alternative hypothesis e.g. *the probabilty to get a girl is higher then 50%*)

Even if the *true probability* of getting a girl is 50% it is possible that *by chance* we observe a **sample probability** which is higher than 50%.

Even if the *true probability* of getting a girl is higher than 50% it is however possible that a **sample probability** is observed that is very close to 50% (or even lower).

In defining the *reject region* we need to control **randomness** or the probability of making mistakes and this can be done by using statistical inference.



Errori di I e di II tipo

Studio campionario → Possibilità di decisioni sbagliate

		VERITA' (Ignota)	
		H_0 è vera	H_0 è falsa
Decisione presa sui dati campionari	Rigetto H_0	Errore di I tipo *	ok
	Non rigetto H_0	ok	Errore di II tipo **

***Errore di I tipo:**

- Rigettare H_0 quando in realtà è vera; (falso positivo = innocente in galera);
- Si associa una probabilità α di commettere questo errore: livello di significatività
- α è ***sotto controllo***, perché il test è disegnato in modo tale che α non sia più grande di una soglia pre-specificata

****Errore di II tipo:**

- Non rigettare H_0 quando in realtà è falsa; (falso negativo = colpevole in libertà);
- Si associa una probabilità β di commettere questo errore: **$1-\beta$=Potenza del test**
- β non è solitamente sotto controllo, perché la distribuzione della statistica di test è nota solo sotto l'ipotesi nulla…

Effettuare un test statistico (strategia generale)

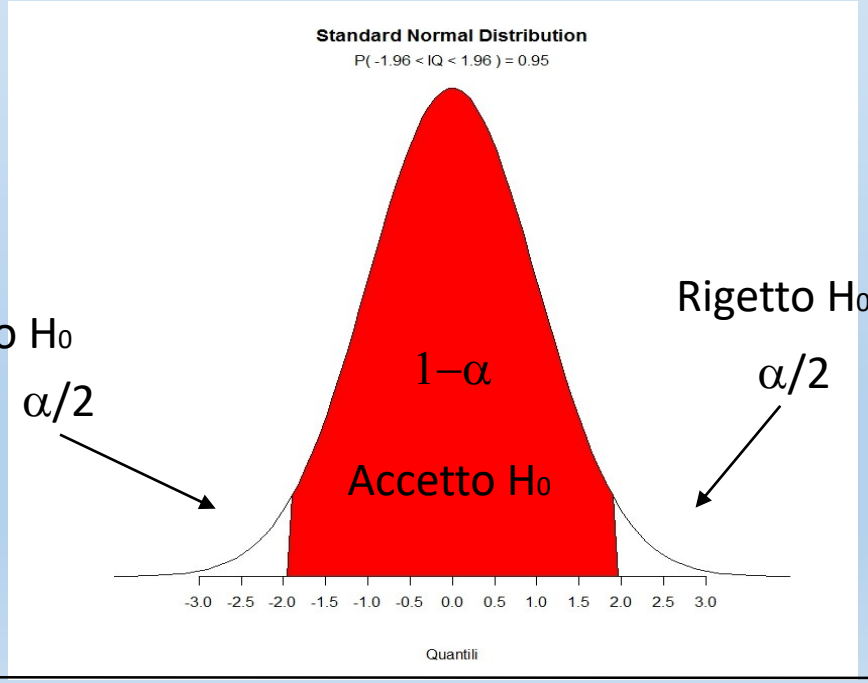
(a) Ipotesi *nulla* **H_0** *versus* Ipotesi *alternativa* **H_1** (mutualmente esclusive)

(b) Si raccoglie un campione di dati **x** (es: glicemia/bimbi nati da coppie in trattamento…)

(c) Si calcola sui dati una *statistica di test* **T(x1, x2,…)=t** la cui distribuzione di probabilità è nota se vale H_0 (e differisce rispetto a quello che avrebbe sotto H_1)

(d) Si rigetta H_0 se il valore osservato di **t** è troppo *poco probabile* (se H_0 fosse vera): ***p* <= α**

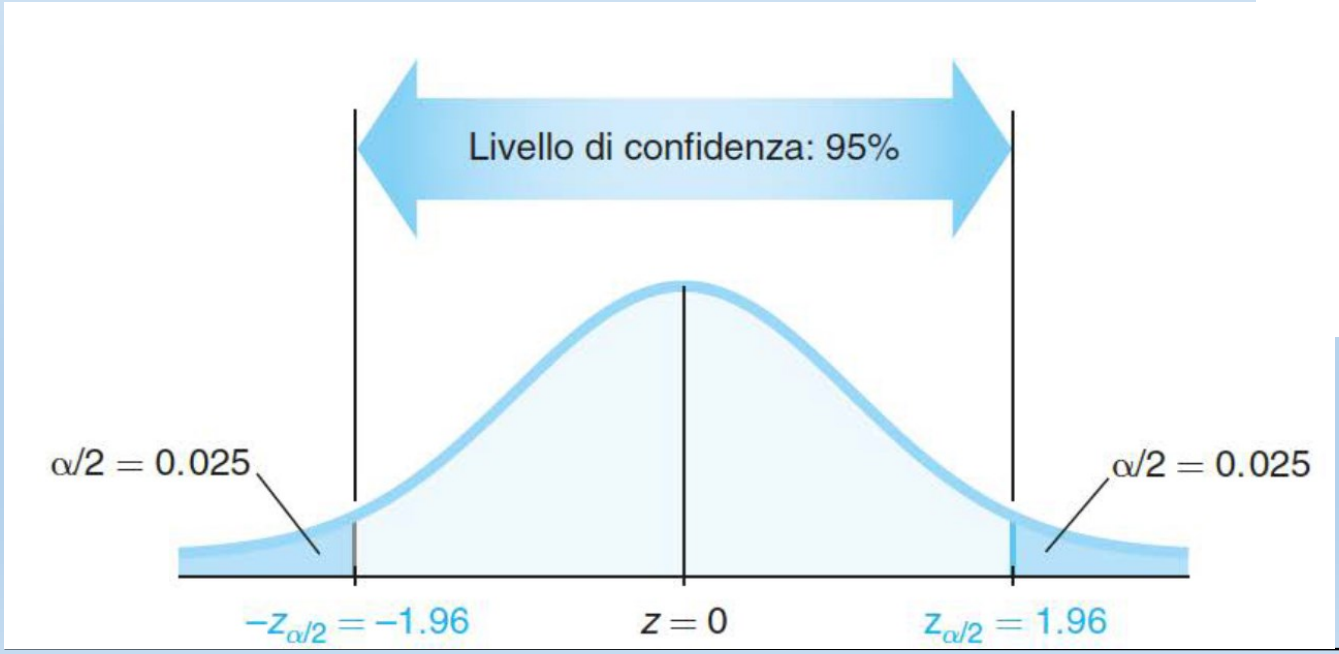
***p-value*= probabilità sotto H_0 che la variabile casuale T abbia il valore t osservato sui dati campionari o un valore più «*estremo*»**

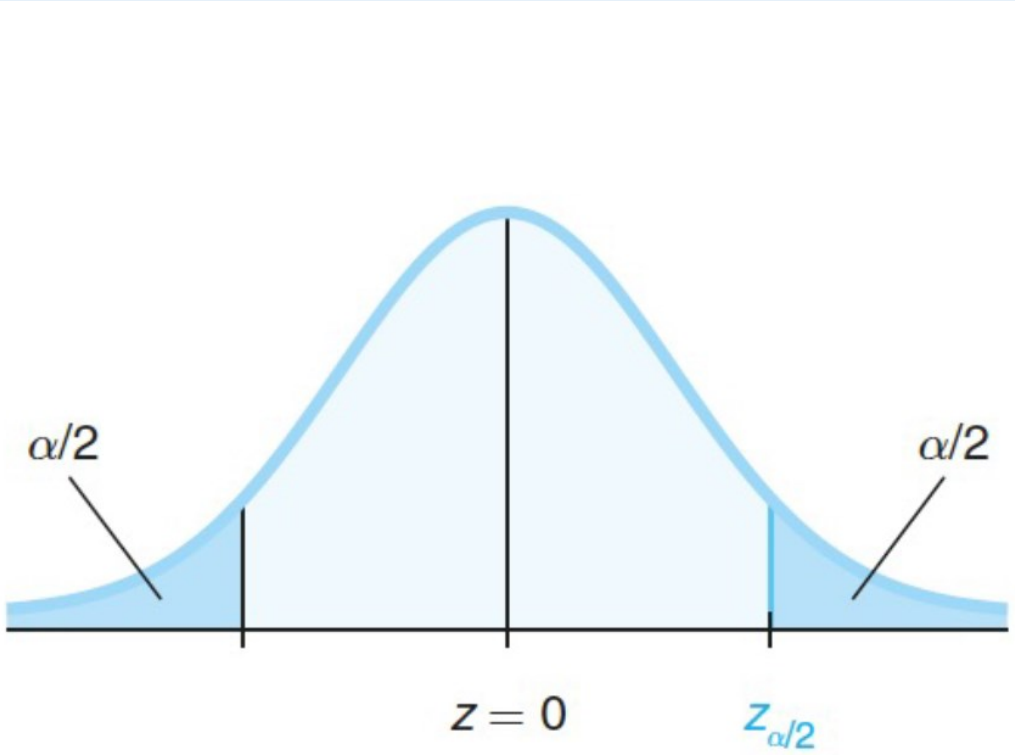


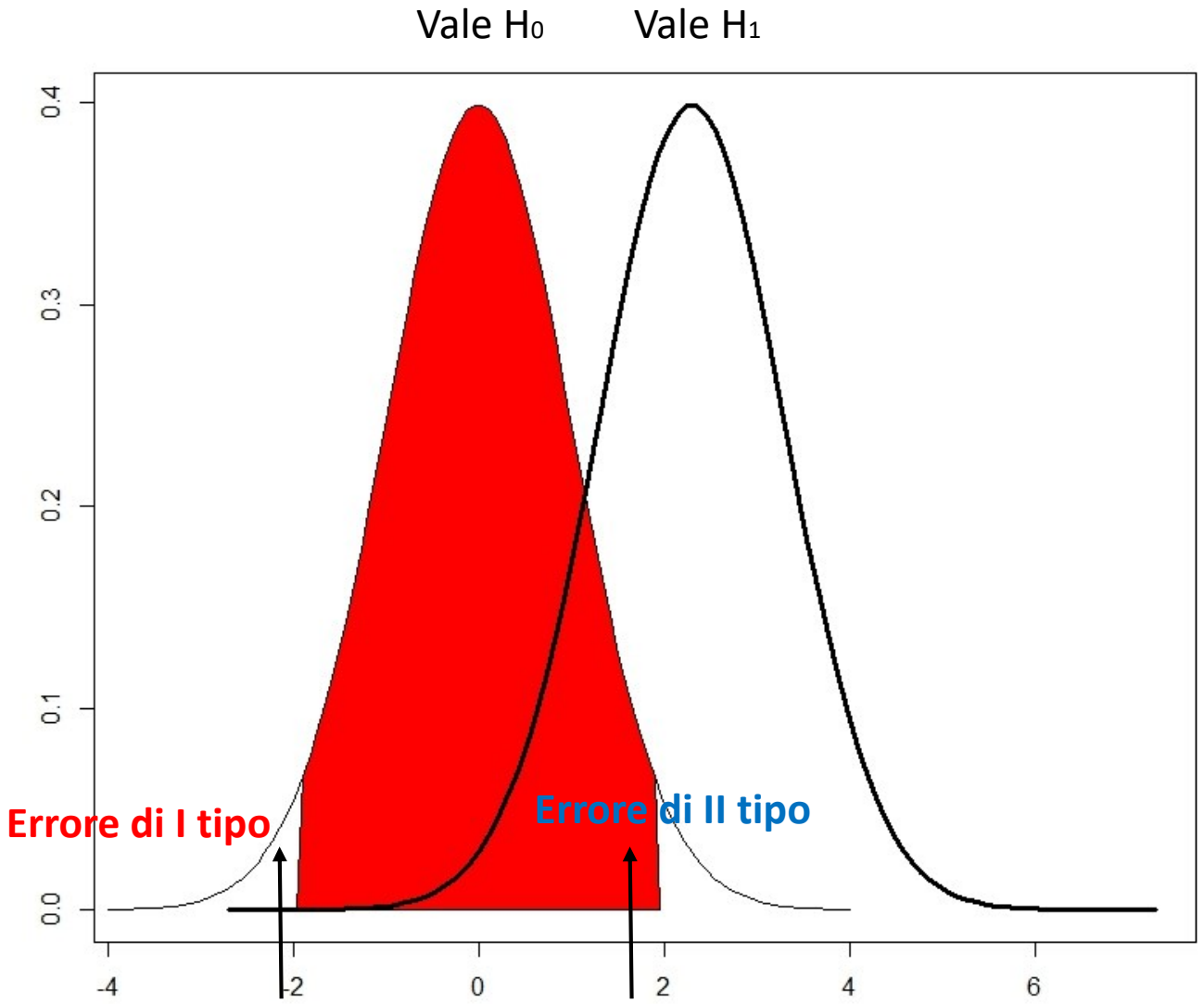
Un **valore critico** è il valore corrispondente al quantile della distribuzione che delimita l'area con maggiore possibilità di verificarsi.

Il quantile $z_{\alpha/2}$ è un valore critico, cioè un valore z che ha la proprietà di separare un'area di probabilità $\alpha/2$ nella coda destra della distribuzione normale standard:

Spesso nei test di ipotesi il valore critico delimita una probabilità complessiva nelle code del **5%**.







		VERITA' (Ignota)	
		H_0 è vera	H_0 è falsa
Decisione presa sui dati campionari	Rigetto H_0	Errore di I tipo *	ok
	Non rigetto H_0	ok	Errore di II tipo **

Conclusioni & Conseguenze

Supponiamo che α sia piccolo, tipicamente $\alpha <= 0.05$

- Se H_0 viene «rigettata» questa decisione **è affidabile**:

la probabilità di errore sempre nota a priori α è piccola («*il risultato del test è statisticamente significativo*»);

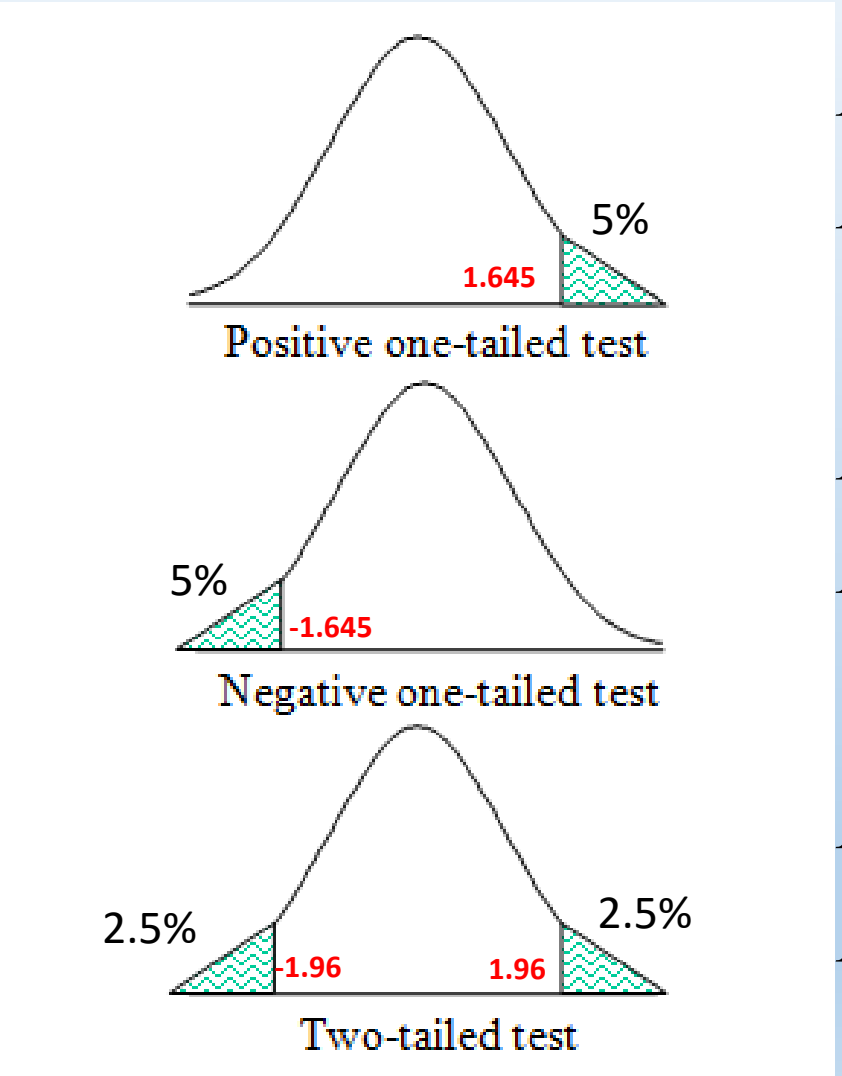
- Se H_0 non è rigettata, si conclude che i dati non offrono sufficiente evidenza per *rigettare* H_0;

questa decisione **può non essere così affidabile**:
la probabilità di errore β non è (generalmente) fissata a priori e
potrebbe essere grande («*il risultato del test non è statisticamente significativo*»);

"absence of evidence is not evidence of absence"

*** una ipotesi H_0 può non essere rigettata (quando invece dovrebbe) perché la dimensione campionaria è troppo piccola!!!**

Test di ipotesi ad una coda oppure a due code?*



$H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 > \mu_2$

Supponiamo di dover confrontare due farmaci «A» e «B».

$H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 < \mu_2$

Se si crede che il farmaco «A» **sia meglio** del farmaco «B», si farà il test ad una coda. In questo caso però si rischia di accettare l'ipotesi nulla di *uguaglianza* anche nel caso che «A» sia *inferiore* a «B»

$H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 \neq \mu_2$

Solo nel caso si consideri questo problema *trascurabile*, si potrà usare il test ad una coda...

*(se la statistica di test ha una distribuzione simmetrica)

We know the **theoretical distribution** of the sample probability:

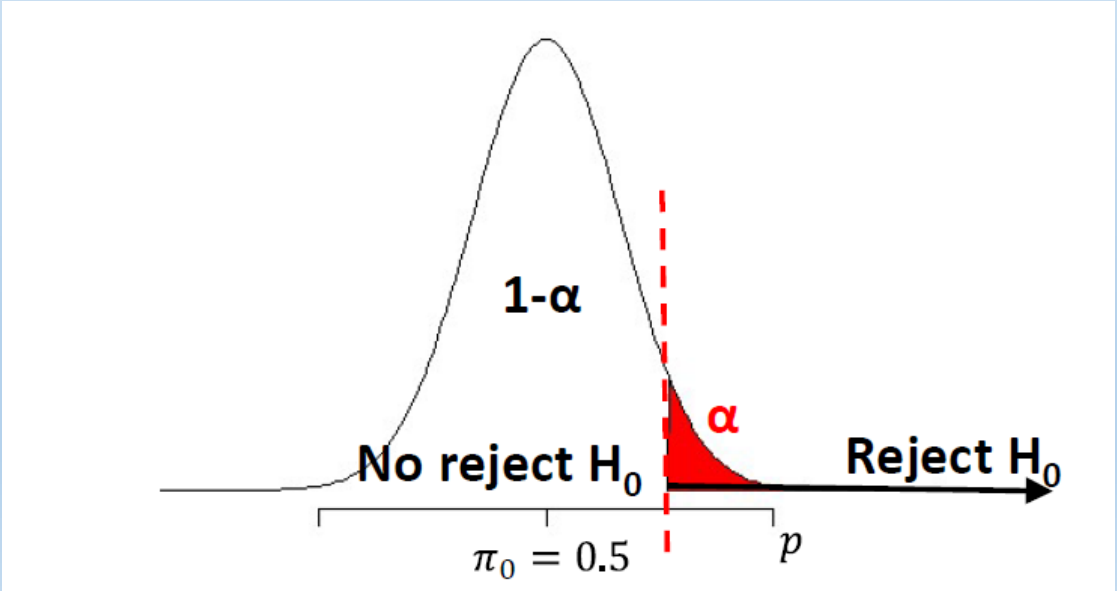
$$\bar{p} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{stimatore di p}$$

$$\frac{\bar{p} - p}{SE(\bar{p})} \approx N(0,1)$$

$$SE(\bar{p}) = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \quad \text{errore standard di } \bar{p}$$

Under the null (true probability of getting a girl is 50%,): $\boldsymbol{H_0: \pi = \pi_0 = 0.5}$

the theoretical distribution of the sample probability: $\boldsymbol{p \sim N(0.5, \sqrt{0.5 * 0.5/14})}$



We can then define the **critical rejection region** in order to establish a priori the probability of making mistakes when we reject H_0.

This probability is called significance level α.

Under the null (H_0): XSORT does not work

Better standardise: $z = \frac{p-\pi_0}{se(p|H_0)}$ $\quad Z \sim N(0,1)$

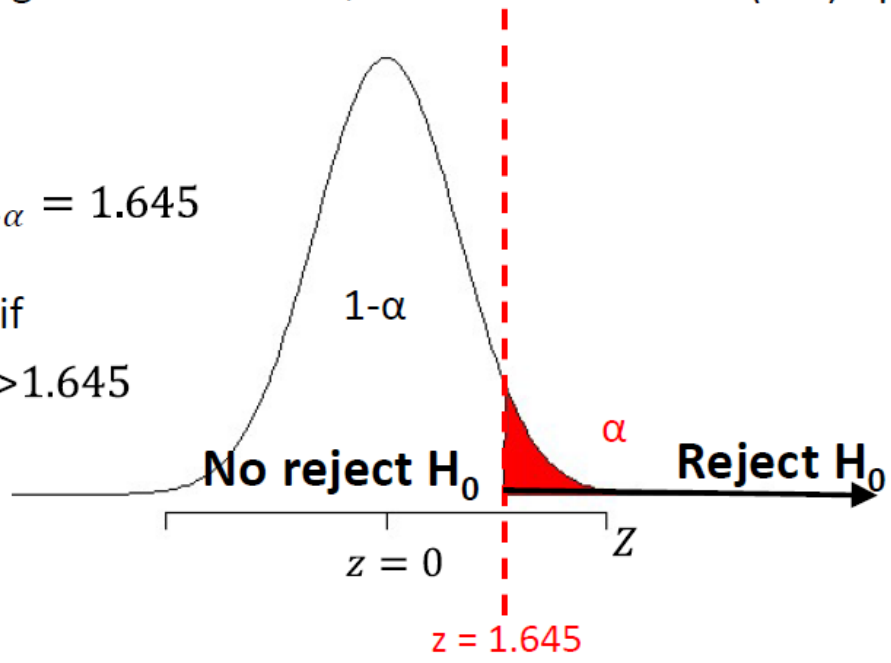
$$z = \frac{p - 0.5}{\sqrt{0.5 * 0.5/14}}$$

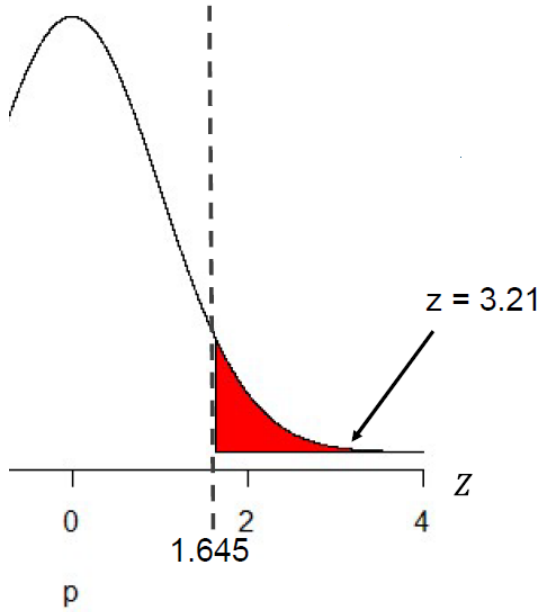
When the level of significance α is set, the threshold is the (1-α)th percentile $z_{1-\alpha}$

Ex. $\alpha = 0.05$ -> $z_{1-\alpha} = 1.645$

Thus we will reject if

$$z = \frac{p-0.5}{\sqrt{(0.5*0.5)/14}} > 1.645$$





n=14

p=13/14=0.929

$\alpha = 0.05 \quad z_{0.95} = 1.645$

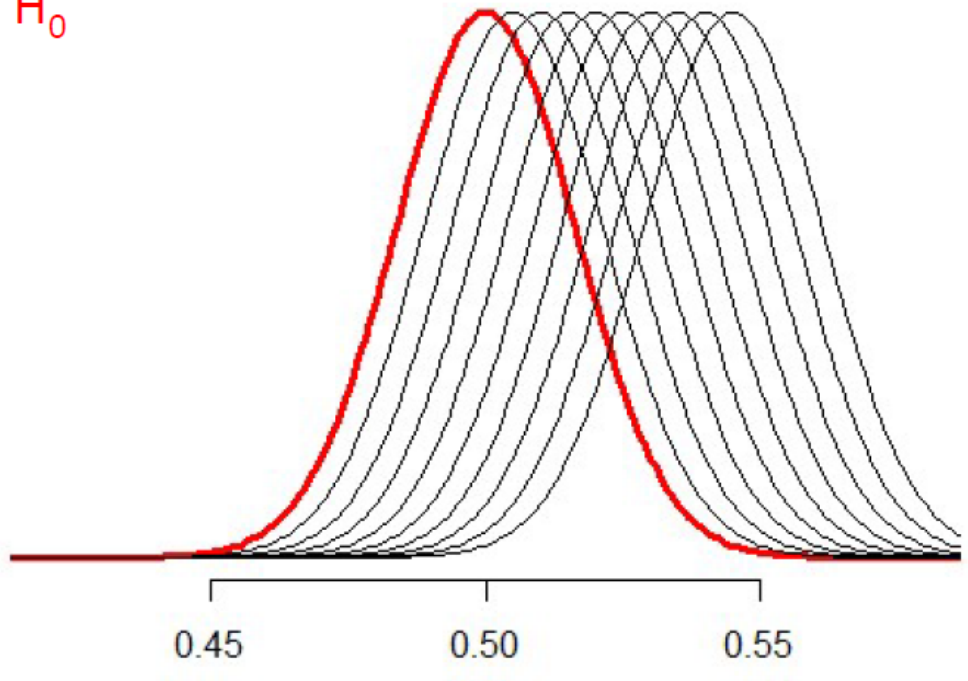
$$z = \frac{p-\pi_0}{se(p)}$$

$$z = \frac{0.929 - 0.5}{\sqrt{0.5 * 0.5/14}} = 3.21$$

Reject the null hypothesis! There is sufficient sample evidence to support the claim that for couples using the «XSORT gender selection method», most (more than half) of their babies are girls.

The logic of the hypothesis test:

The statistical hypthesis test is based on the disproof of a specific hypothesis H_0



p

H_0: π=0.5

Under a specific single hypothesis is possible to find the sampling distribution

H_1: π>0.5

The alternative hypothesis includes an infinity of values and their related sampling distributions

Torneremo su questo concetto più nello specifico quando parleremo della dimensione campionaria, dell' **effect size** e della potenza del test.

Esempio di test di ipotesi per la differenza tra due medie

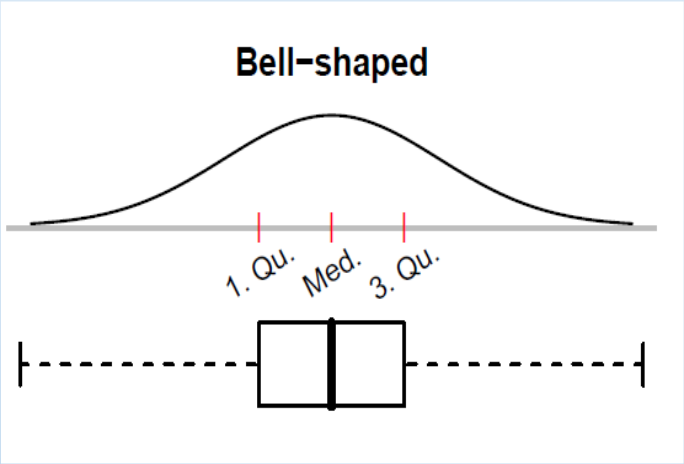
Dati due campioni del carattere x* misurato in due gruppi:

$$H_0: \varepsilon = 0$$

$$H_1: \varepsilon \neq 0$$

$$\varepsilon = \bar{x}_1 - \bar{x}_2$$

* per i quali abbia senso utilizzare la media come indice di posizione…



- Se le VA sono **indipendenti** → t-test
  Es: due gruppi diversi di pazienti
- Se le VA *non sono indipendenti* → t-test per dati **accoppiati**
  Es: stesso gruppo pre-post

NB: nell'utilizzo del t-test occorre specificare se le varianze dei due campioni sono simili oppure no (F-test)

Test di ipotesi per la differenza tra due medie: esempio

C'è una differenza nella risposta dell'antigene p24 nei malati di HIV utilizzando 300 mg giornalieri di AZT vs 600mg ? A 10 pz vengono somministrati 300 mg e ad altri 10 pz 600 mg.

$$H_0 : \mu_{300} = \mu_{600}$$

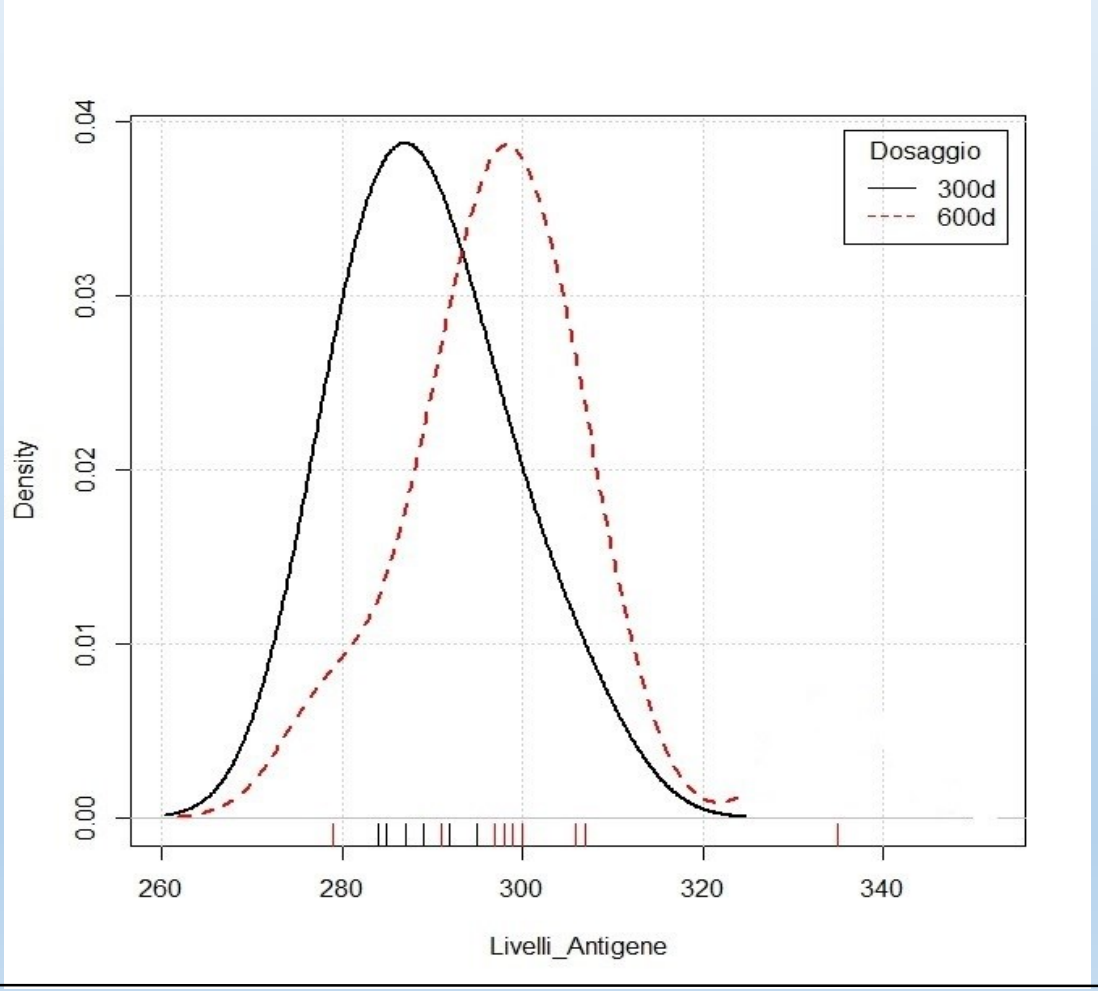
$$H_1 : \mu_{300} \neq \mu_{600}$$

1. Verifichiamo la *gaussianità* delle distribuzioni:

```
with(antigene_300d, shapiro.test(Livelli_Antigene))
```

```
    Shapiro-Wilk normality test

data:  Livelli_Antigene
W = 0.95218, p-value = 0.6943
```

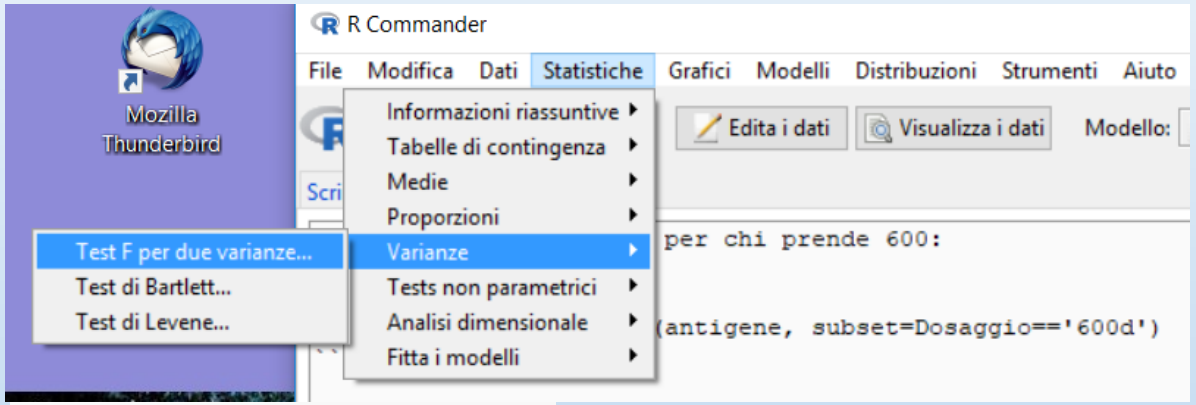
```
> with(antigene_600d, shapiro.test(Livelli_Antigene))

    Shapiro-Wilk normality test

data:  Livelli_Antigene
W = 0.8734, p-value = 0.1095
```

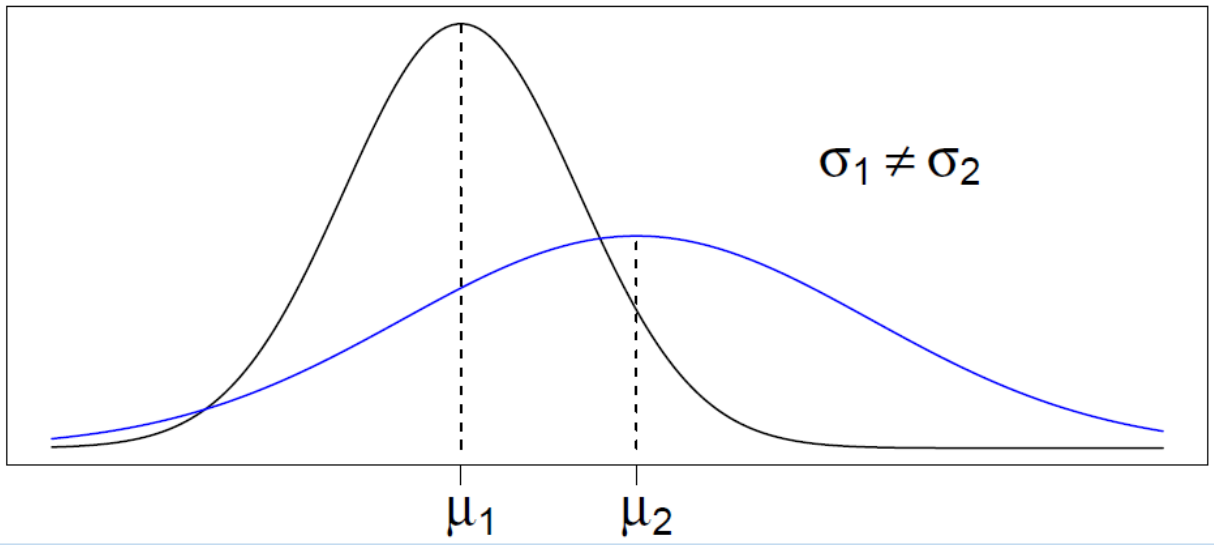


Test di ipotesi per la differenza tra due medie: esempio

2. Prima di effettuare il test sulla media, dobbiamo verificare se le varianze* delle distribuzioni siano omogenee (si utilizza la statistica di test F):

$$H_0 : \sigma_X = \sigma_Y$$
$$H_1 : \sigma_X \neq \sigma_Y$$





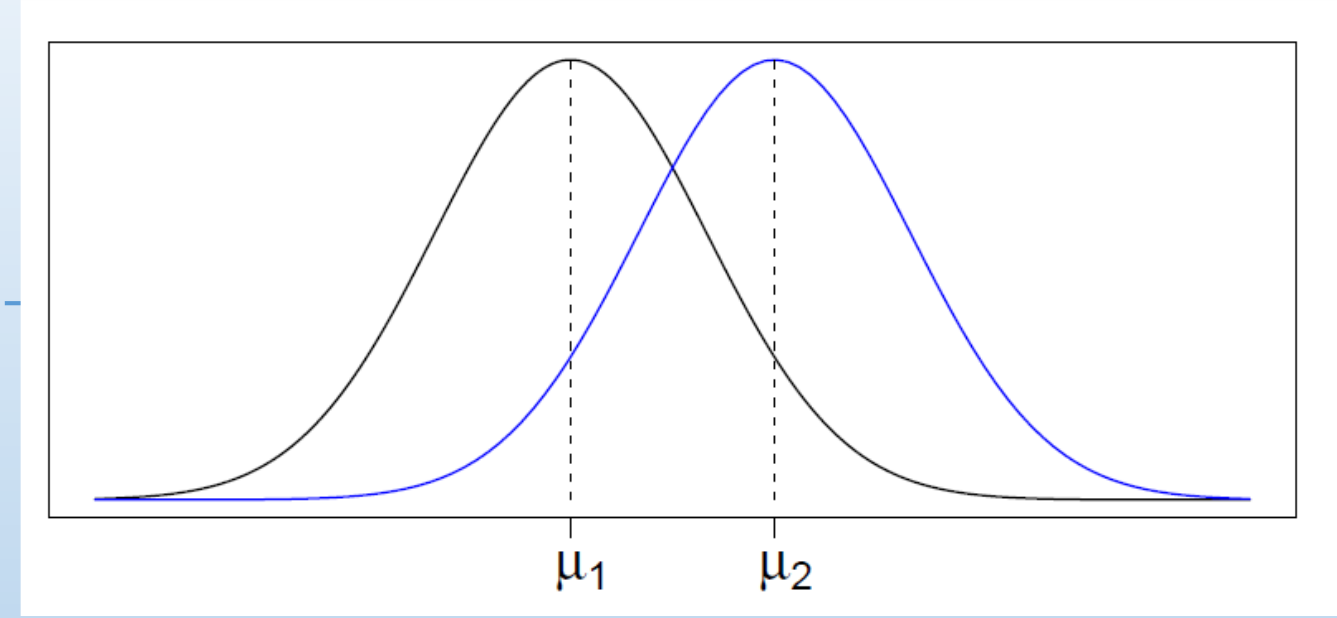

```
F test to compare two variances

data:  Livelli_Antigene by Dosaggio
F = 0.34183, num df = 9, denom df = 9, p-value = 0.1256
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.0849059 1.3762082
sample estimates:
ratio of variances
         0.3418306
```

Accettiamo l'ipotesi nulla di omogeneità delle varianze

*Il rapporto fra due varianze segue la distribuzione F

Test di ipotesi per la differenza tra due medie: esempio

3. Effettuiamo adesso il t-test:




```
        Two Sample t-test

data:  m300 and m600
t = -2.034, df = 18, p-value = 0.05696

95 percent confidence interval:
 -22.158       0.3584

sample estimates: mean of x 289.4 mean of y 300.3
```

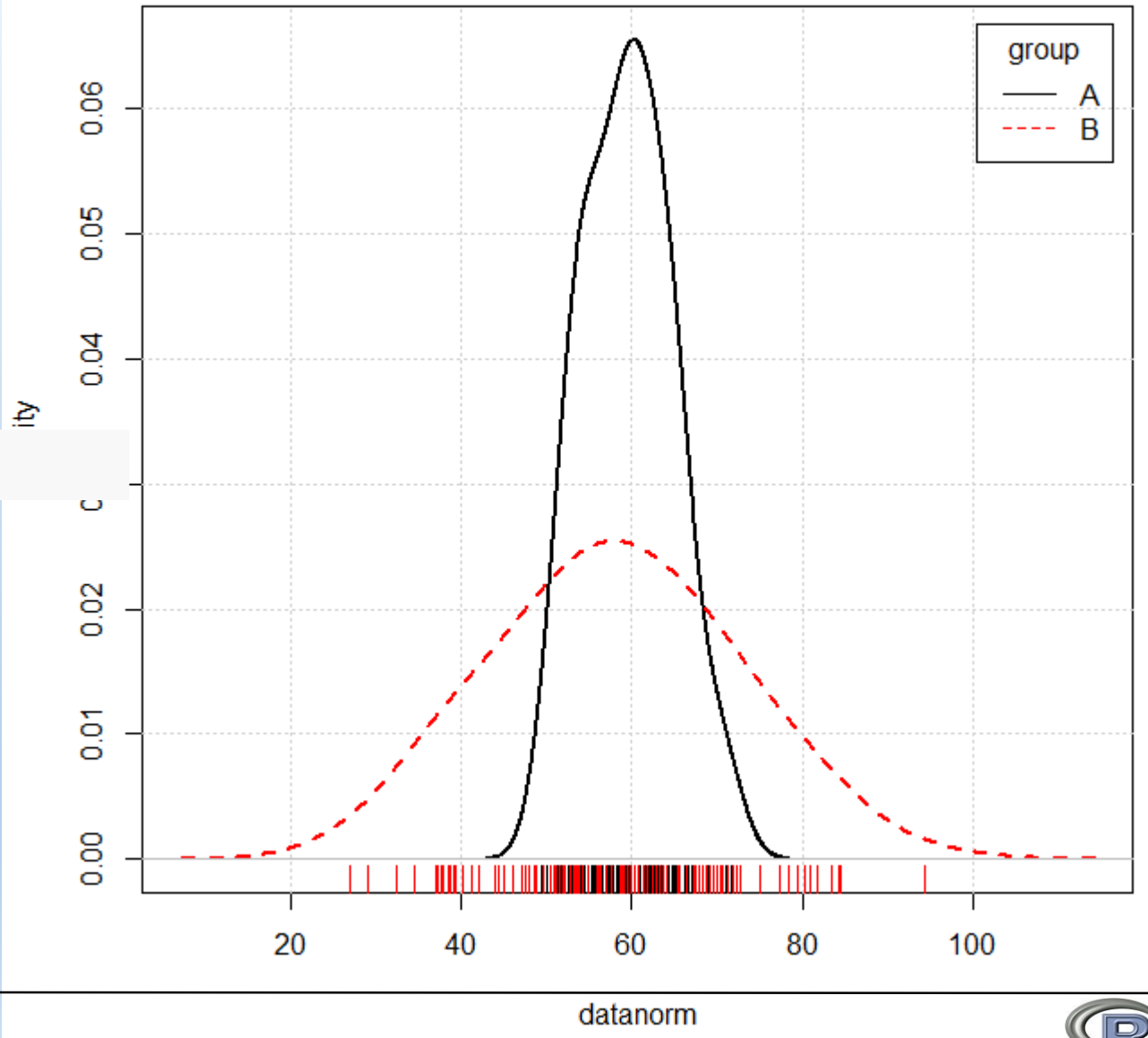
L'ampiezza dell'IC per la differenza delle medie ci fa concludere che i dati non supportano una differenza significativa (contiene zero)

Suggerisce una differenza fra le medie ma non significativa (alla soglia standard di 0.05)

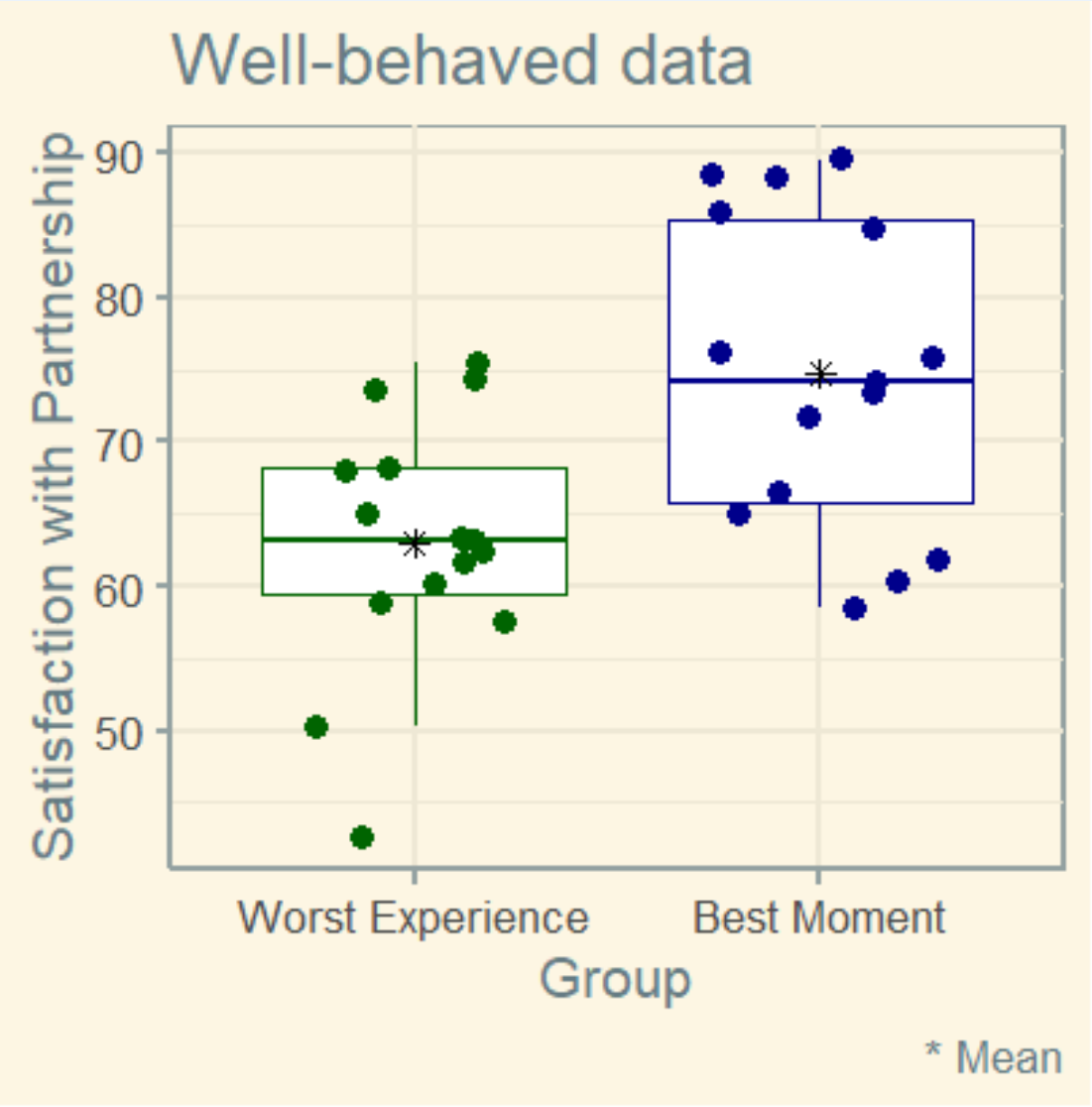
Test di ipotesi per la differenza tra due medie: esempio per varianze non omogenee

```
F test to compare two variances

data:  datanorm by group
F = 0.13816, num df = 99, denom df = 99, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.09296229 0.20534354
sample estimates:
ratio of variances
         0.1381637
```

```
> t.test(datanorm~group, alternative='two.sided', conf.level=.95,
+   var.equal=FALSE, data=dati.diff.var)

    Welch Two Sample t-test

data:  datanorm by group
t = 1.027, df = 125.84, p-value = 0.3064
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.410123  4.452409
sample estimates:
mean in group A mean in group B
       59.56966        58.04851
```



Impatto della **asimmetria/outliers** sugli esiti del test

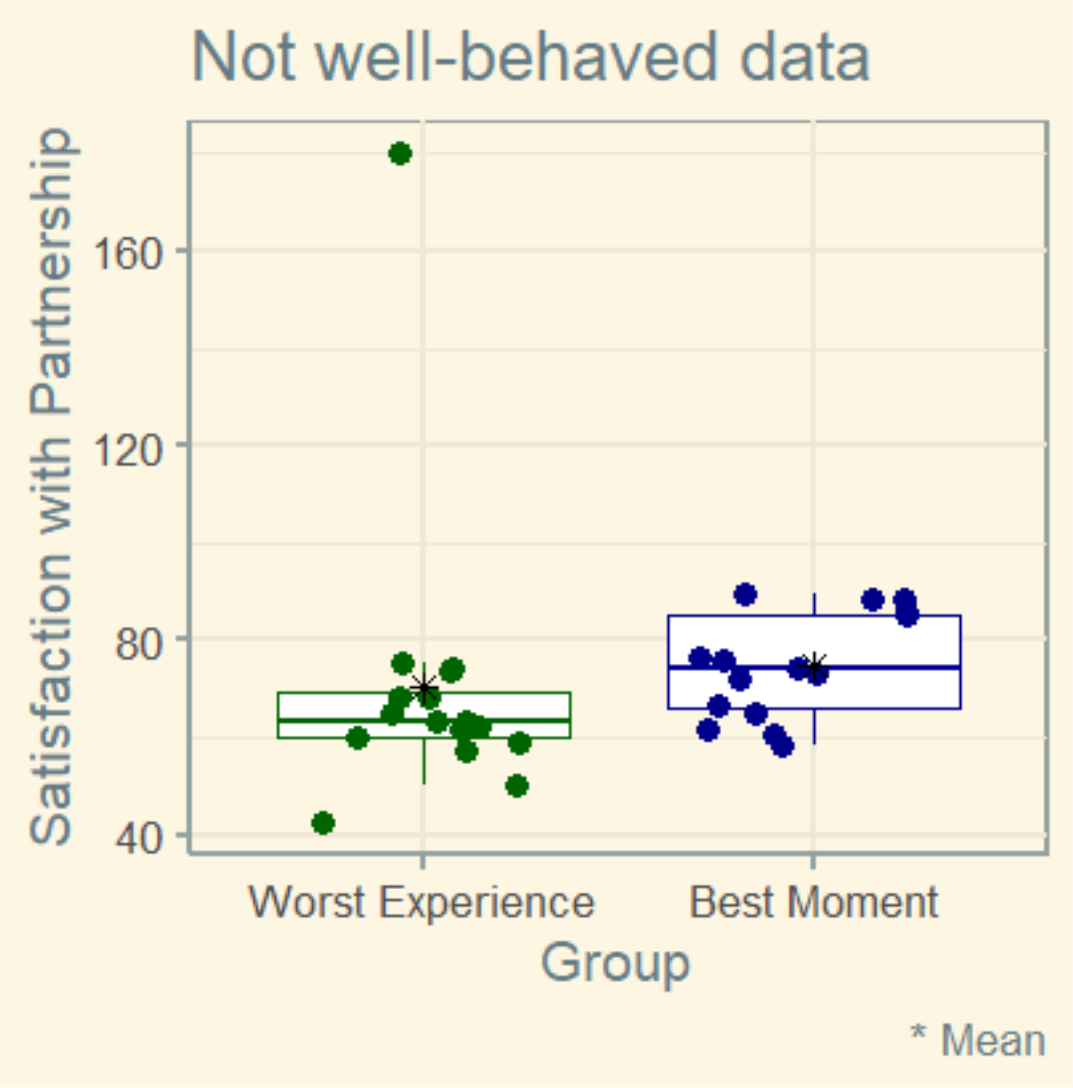


Characteristic	Worst Experience, N = 15[1]	Best Moment, N = 15[1]	p-value[2]
satisfaction	63 (9)	75 (11)	0.003

[1] Mean (SD)

[2] Welch Two Sample t-test

Il t-test (opzione varianze non omogenee) trova una differenza significativa tra le medie di questi due gruppi.



Characteristic	Worst Experience, N = 16[1]	Best Moment, N = 15[1]	p-value[2]
satisfaction	70 (30)	75 (11)	0.6

[1] Mean (SD)

[2] Welch Two Sample t-test

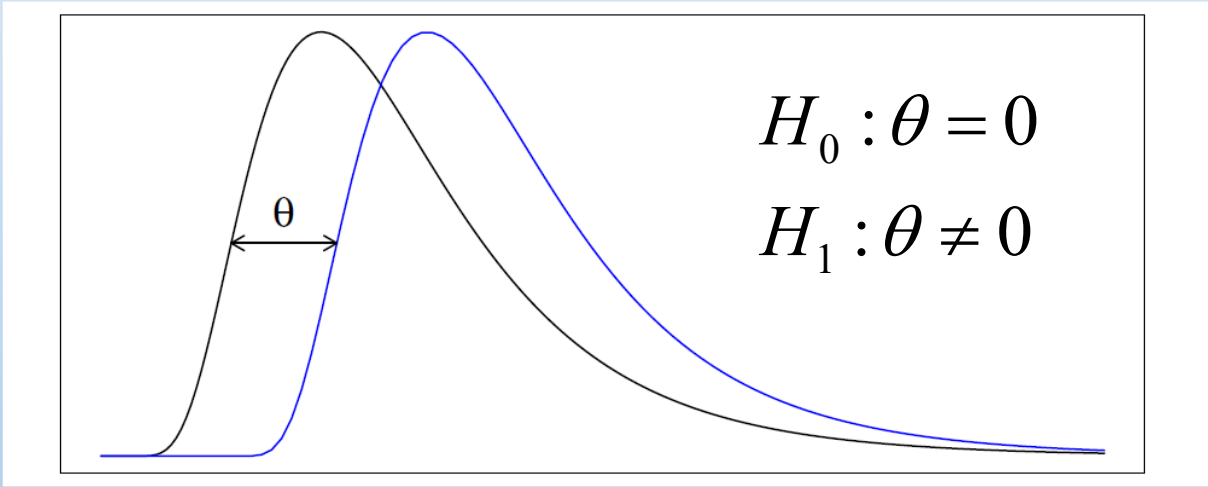
Il t-test (opzione varianze non omogenee) **non trova** una differenza significativa tra le medie di questi due gruppi ….

Come risolvere [se non è possibile trasformare]?

Test di ipotesi per la differenza tra due distribuzioni *non gaussiane*

Quando le distribuzioni non sono «gaussiane» la media può non essere più rappresentativa…

Il test di significatività può essere fatto sui «**ranghi**»* della distribuzione:



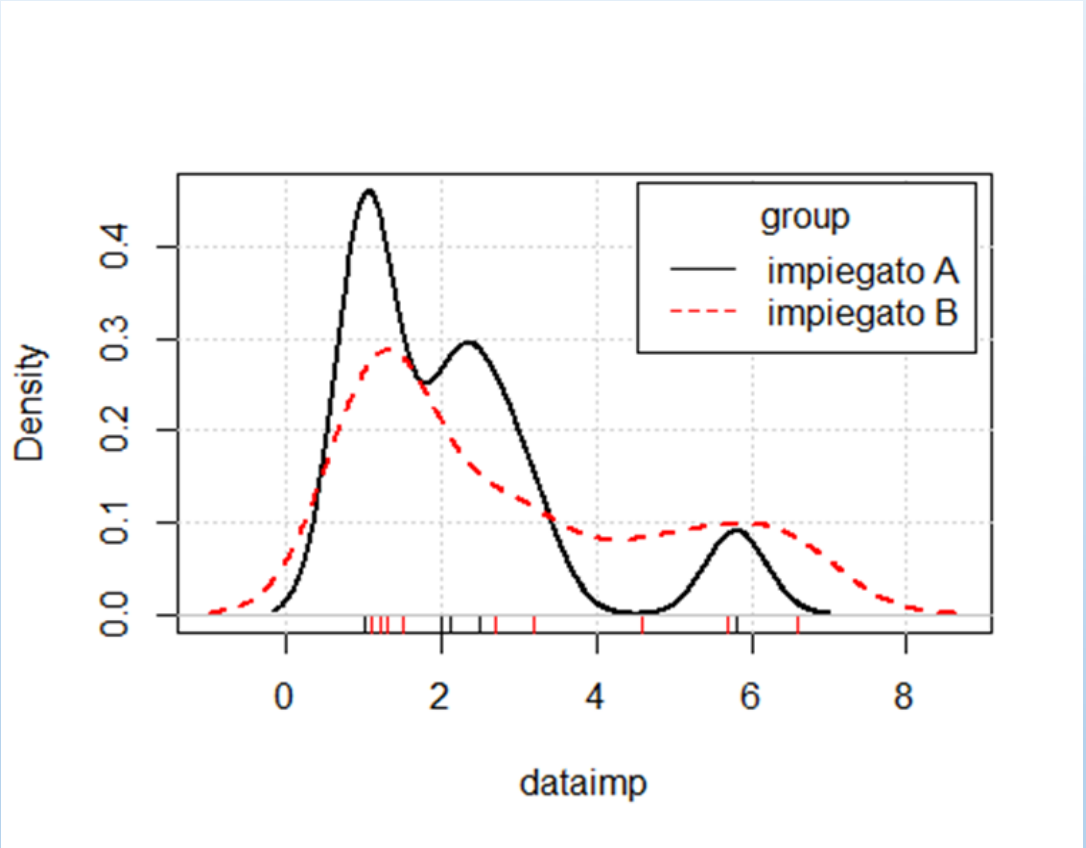
- Se le VA sono indipendenti → Wilcoxon-Mann-Whitney test (Wilcoxon rank sum test)
- Se le VA non sono indipendenti → Wilcoxon test per dati accoppiati (Wilcoxon signed rank test)

*posizioni delle osservazioni nell'ordinamento

Wilcoxon-Mann-Whitney test: esempio sui dati

In un supermercato si vogliono confrontare i tempi alla cassa di due impiegati A e B:

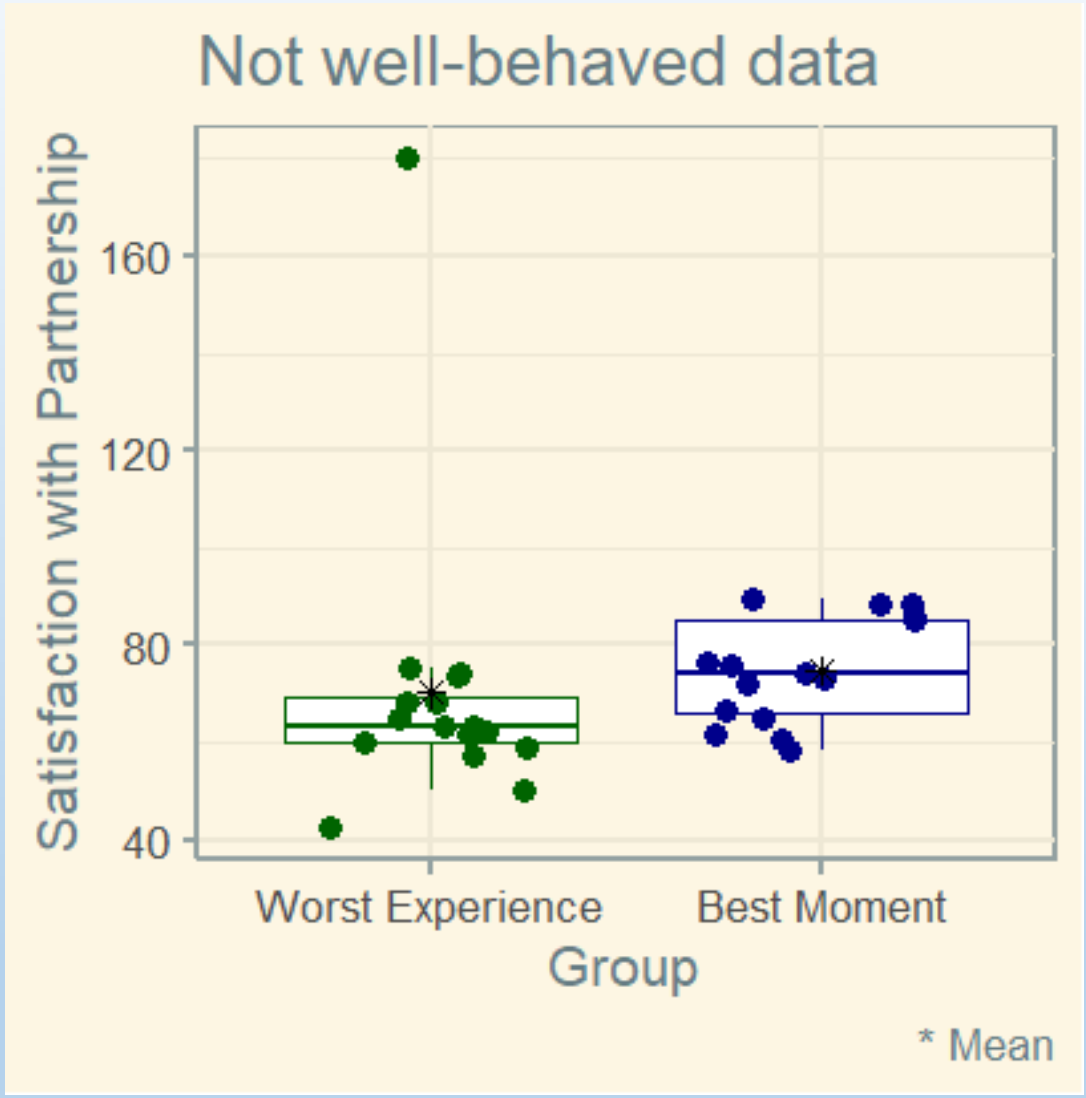
	mean	sd	IQR	0%	25%	50%	75%	100%	data:n
impiegato A	2.136364	1.450705	1.550	1.0	1.050	2.0	2.60	5.8	11
impiegato B	2.910000	2.057210	3.025	1.1	1.225	2.1	4.25	6.6	10




```
    Wilcoxon rank sum test with continuity correction

data:  dataimp by group
W = 37, p-value = 0.2162
alternative hypothesis: true location shift is not equal to 0
```

Non c'è evidenza che indichi una differenza significativa tra le distribuzioni.

*Il test è basato sui «ranghi» cioè sull'ordinamento dei dati dal più piccolo al più grande



Characteristic	Worst Experience, N = 16[1]	Best Moment, N = 15[1]	p-value[2]
satisfaction	63 (60, 70)	74 (66, 85)	0.027

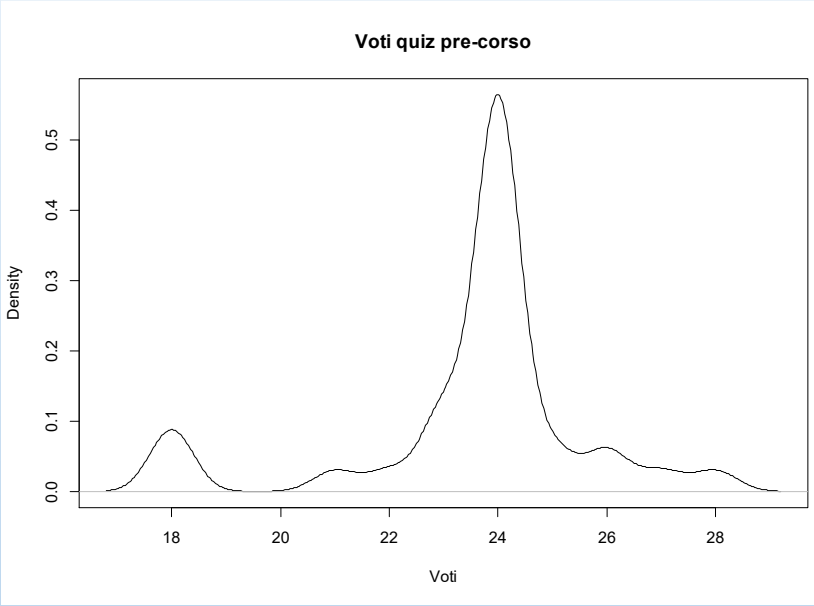
[1] Median (IQR)

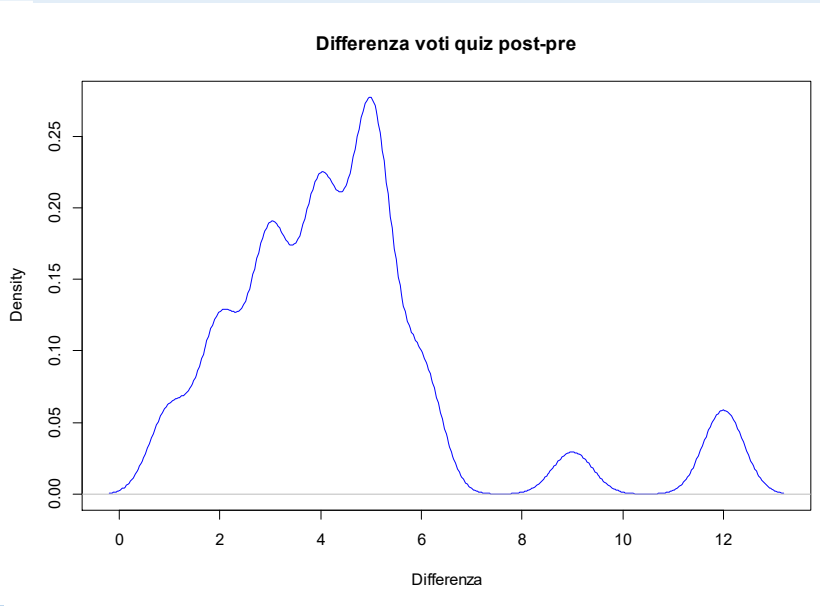
[2] Wilcoxon rank sum exact test

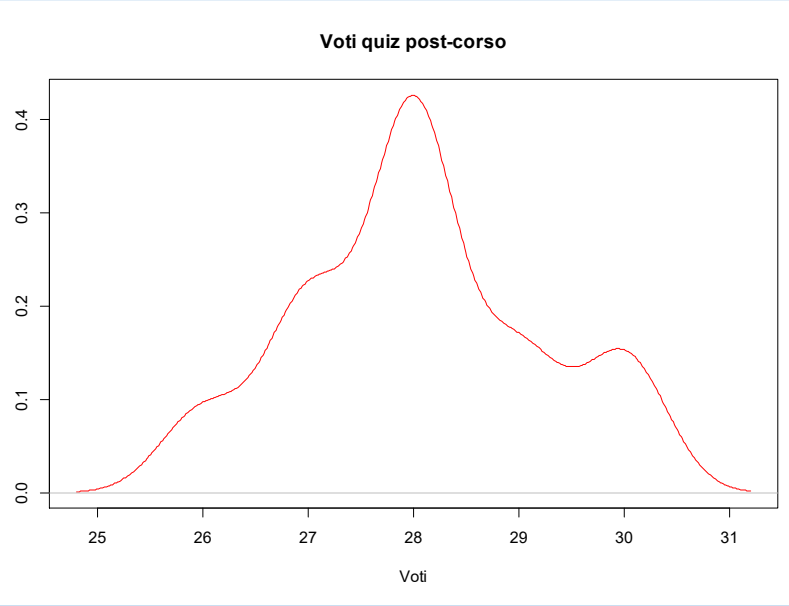
Il test non parametrico **trova correttamente** una differenza significativa nelle distribuzioni di questi due gruppi.

Wilcoxon-test su dati accoppiati: esempio sui dati

Voti ad un quiz di statistica su un gruppo di persone, prima di frequentare questo corso e subito dopo:







La distribuzione delle differenze non è simmetrica intorno alla media. Utilizzeremo il test sui ranghi:

```
wilcoxon signed rank test with continuity correction data:
pre.corso and post.corso

p-value = 3.417e-07

alternative hypothesis: true location shift is not equal to 0
```

Aver frequentato questo corso fa aumentare
In modo significativo i voti al quiz !

Test di ipotesi per una proporzione

Si vuole investigare:

- se l'approvazione verso un politico sia in salita o in discesa…
- se il tasso di disoccupazione stia salendo o scendendo…
- se la frequenza di una certa patologia in una zona sia più o meno alta della media nazionale…

Supponiamo che p_0 sia la *proporzione «attesa»:*

$$H_0 : p = p_0$$

$$H_1 : p \neq p_0 \text{ oppure } p > p_0 \text{ oppure } p < p_0$$

Secondo la letteratura la prevalenza di diabete negli USA è intorno al 11.7%.

Nel 2010 è stata stimata al 15% tramite una indagine campionaria su 1600 persone.

E' in linea alla letteratura questo dato?

```
Exact binomial test data:  number of successes = 240
number of trials = 1600

p-value = 7.104e-05

alternative hypothesis:
true probability of success is not equal to 0.117

95 percent confidence interval: 0.13 0.17

sample estimates: probability of success 0.15
```

Test di ipotesi per la differenza tra due proporzioni

Si confrontano due proporzioni campionarie per decidere se c'è oppure no una differenza significativa:

$$H_0 : p_1 = p_2$$

$$H_1 : p_1 \neq p_2 \text{ oppure } p_1 > p_2 \text{ oppure } p_1 < p_2$$

Nel 2010 la prevalenza di diabete negli USA era stata stimata al 15% tramite una indagine campionaria su 1600 persone.

Nel 2011 è stata fatta un'altra indagine ed è risultato 15.13% (su 1500 persone):

La prevalenza di diabete è significativamente diversa nelle due indagini?

La differenza non è statisticamente significativa.

```
2-sample test for equality of proportions with
continuity correction data:

X-squared = 0.0028599, df = 1, p-value = 0.9574
alternative hypothesis: two.sided

95 percent confidence interval:
-0.02451222 0.02717889

sample estimates:
prop 1 prop 2
0.1513 0.1500000
```

Number Needed to Treat (NNT)

In uno studio clinico con una risposta binaria, come vivo o morto, ci sono molti modi per quantificare la differenza tra due trattamenti.

Ad esempio, è possibile utilizzare la differenza tra due proporzioni: $p_{new} - p_{old}$

p_{new} = proporzione di successi con il nuovo trattamento ; p_{old}= proporzione di successi con il vecchio trattamento

$$NNT = \frac{1}{p_{new} - p_{old}}$$

NNT : # di pazienti da sottoporre al nuovo trattamento per ottenere **un ulteriore successo** rispetto al vecchio trattamento.

$1 \leq NNT < \infty$

- $\frac{1}{1-0}$ Il nuovo trattamento ***è sempre efficace*** mentre il vecchio non lo è mai
- $\frac{1}{0}$ Il nuovo trattamento ***non è in alcun modo efficace*** (la differenza è pari a zero)

Potremmo avere la situazione opposta, il nuovo trattamento è dannoso, e quindi ottenere un valore negativo:

NNT < 0 → NNH= ***number needed to harm*** (numero necessario per avere un danno)

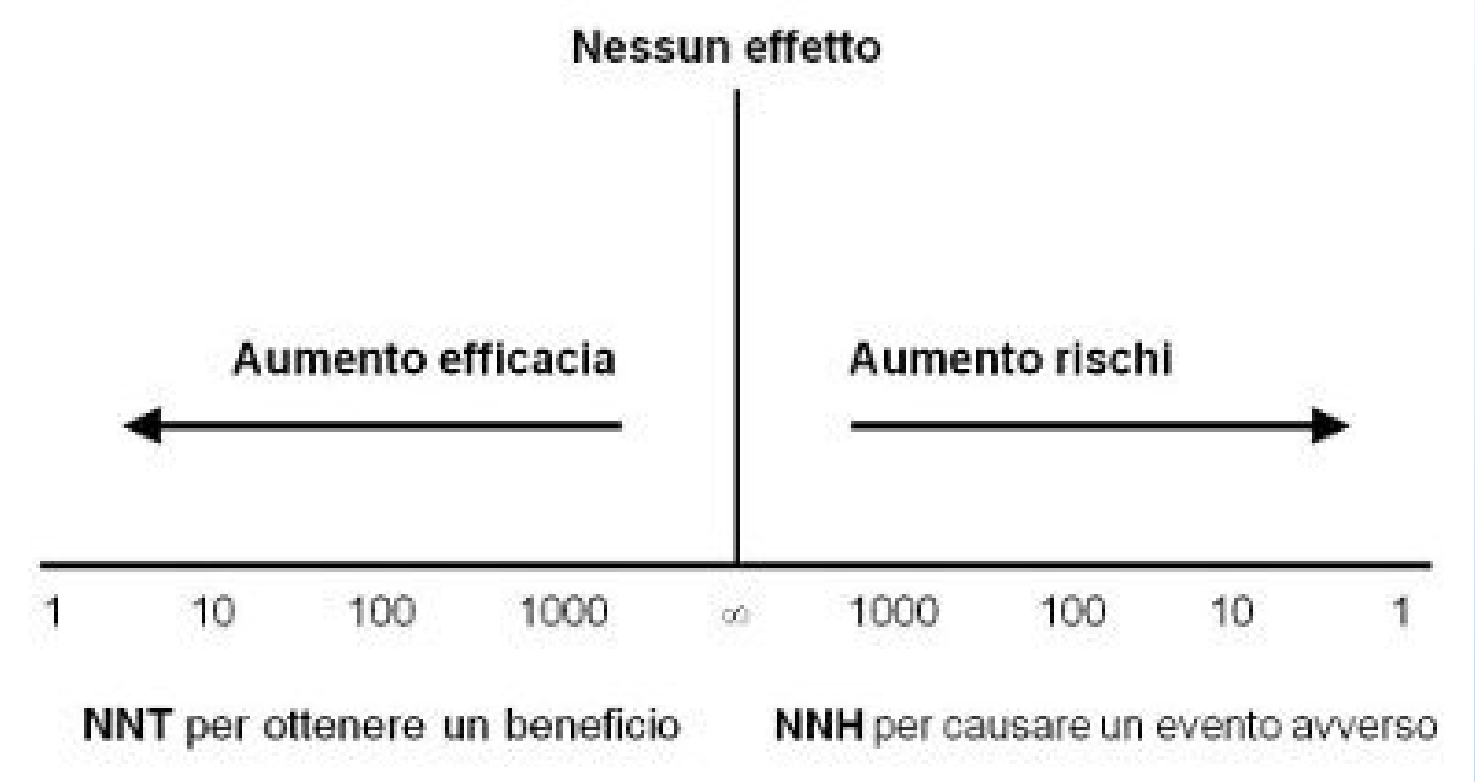
Esempio:
In uno studio sulla streptomicina, nel quale i soggetti erano selezionati tra i pazienti affetti da tubercolosi polmonare, la proporzione di sopravviventi a 6 mesi era 93% nel gruppo trattato vs 73% nei controlli.

$$NNT = \frac{1}{p_{new} - p_{old}} = \frac{1}{0.93 - 0.73} = \frac{1}{0.20} = 5$$

Quindi il numero necessario di pazienti da trattare per prevenire un decesso a 6 mesi era pari a 5.

E' possibile calcolare anche gli intervalli di confidenza attorno a NNT, ma non entriamo in questo dettaglio*.

*Sorgono dei problemi qualora la differenza tra le proporzioni non sia statisticamente significativa, perché l'intervallo di confidenza in quel caso contiene lo zero, in tal caso infinito è un possibile valore per NNT…così come sono ammissibili valori negativi.



Al **diminuire** del NNT aumenta l'efficacia del trattamento, per cui 1 è il NNT ideale: un successo terapeutico per ciascun paziente trattato (misura di **efficacia**).

All'**aumentare** del NNH si riduce la probabilità di eventi avversi e aumenta la sicurezza del trattamento, per cui il NNH ideale tende all'infinito, documentando l'assenza di eventi avversi (misura di **sicurezza**).

- Per (quasi) ogni tipo di ipotesi da sottoporre a verifica si può trovare il test opportuno…(o lo si può *creare*)

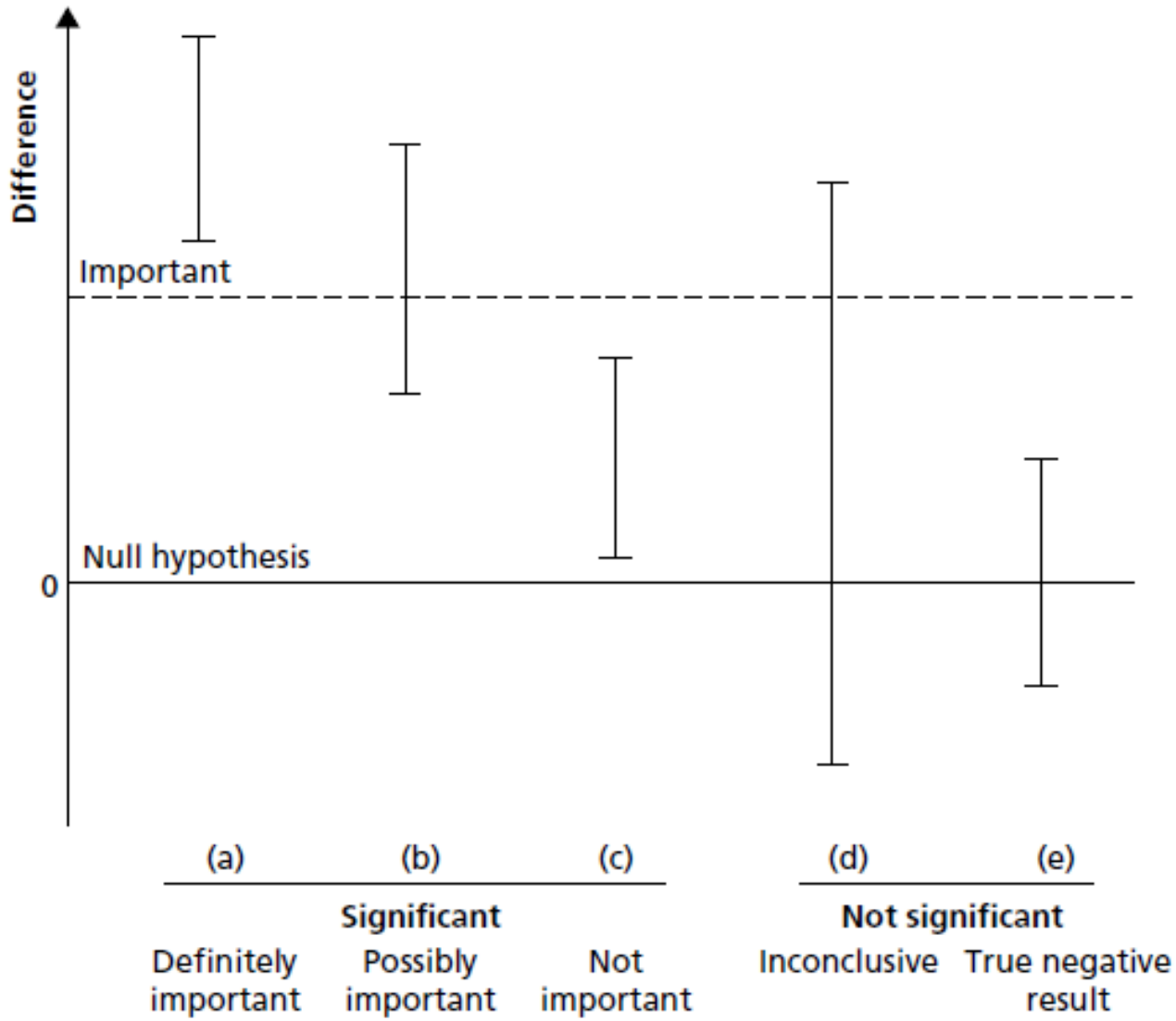
- Il test statistico è soggetto ad **errore**, essendo generato da un meccanismo basato sulla probabilità

- L' errore può essere «*minimizzato*» **ma non eliminato**

- La significatività statistica **non sempre coincide** con la rilevanza clinica, a meno di non definire esplicitamente un ***effect size…***

- E' opportuno consultare lo statistico **nella fase di disegno dello studio** per pianificare le analisi da condurre…ex-post è sempre tardi!

La significatività statistica **non sempre coincide** con la rilevanza clinica:



(a) significant and clinically relevant

(b) significant but it is unclear whether it is clinically relevant

(c) significant but not clinically relevant

(d) not significant but can be clinically relevant

(e) not significant and is not clinically relevant

The **goal** when **planning a study** should be to "guarantee" that **if a clinically relevant difference exists**, then we will be able to identify it through the statistical test (-> sample size/effect size).