CDL in MEDICINA & CHIRURGIA

Statistica Medica

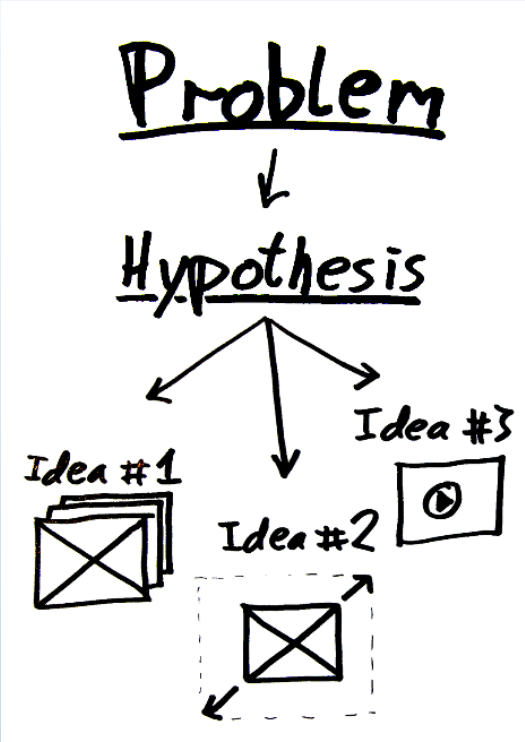
gbarbati@units.it

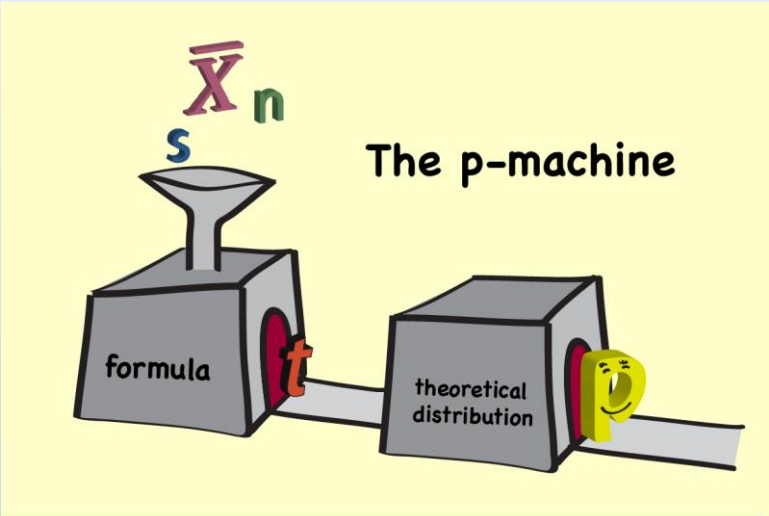
A.A. 2025-26

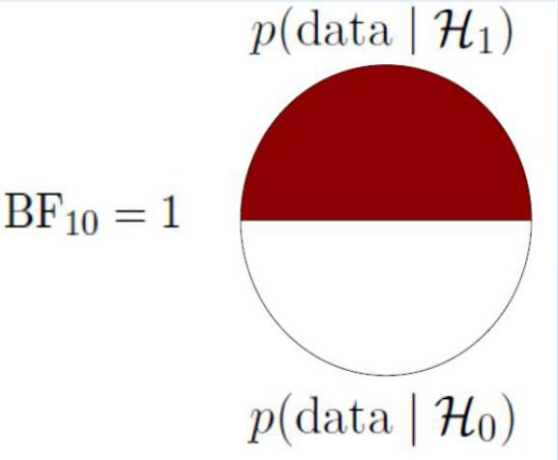
UNITÀ DI BIOSTATISTICA

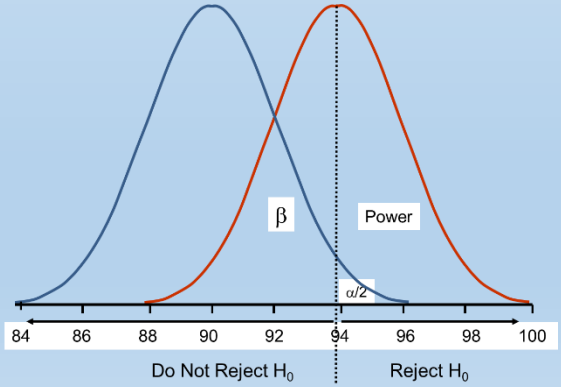
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

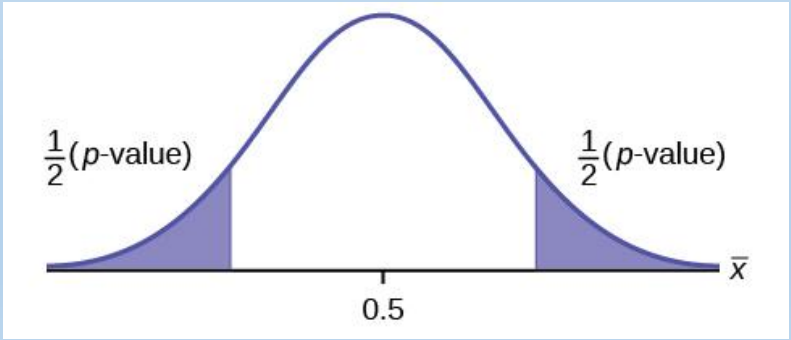
Il meccanismo del *test di ipotesi*





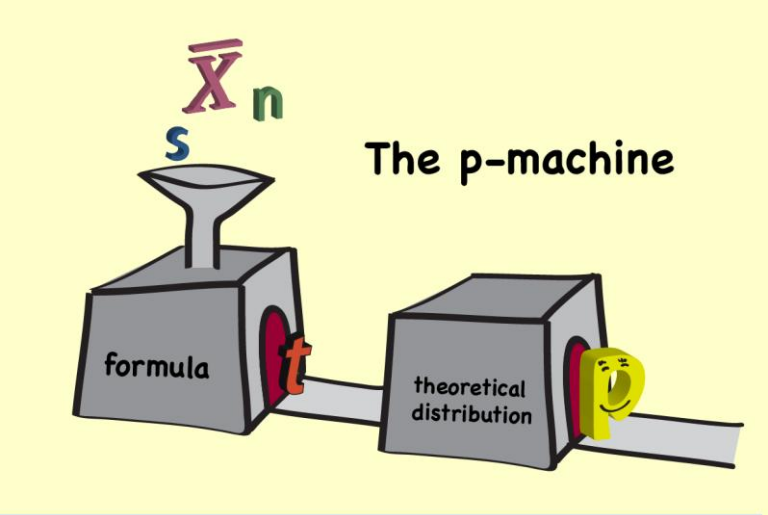


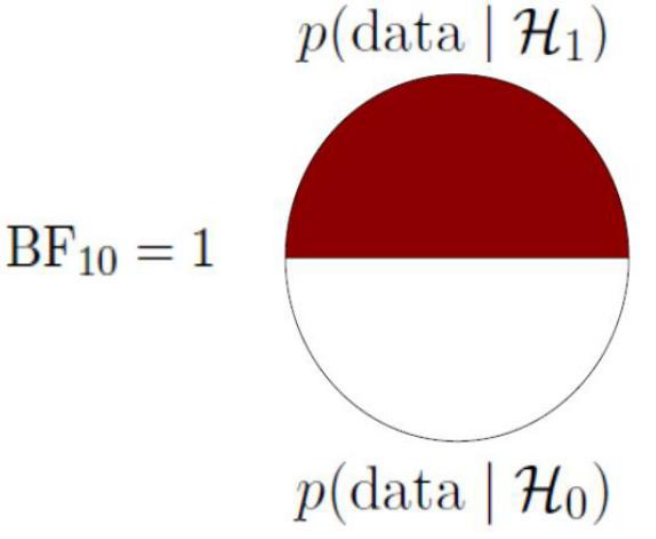




Interpretation of Bayes factors

Bayes factor	Evidence for Hypothesis A
1 to 3.2	"Not worth more than a bare mention"
3.2 to 10	"Positive"
10 to 100	"Strong"
> 100	"Very strong"





Sommario

- Il test di ipotesi (*frequentista*)
  - proporzione / medie
  - Test *non parametrico*
  - Test per dati categoriali

- Il Bayes Factor (cenni)

Introduzione al Test di Ipotesi (frequentista): concetti di base

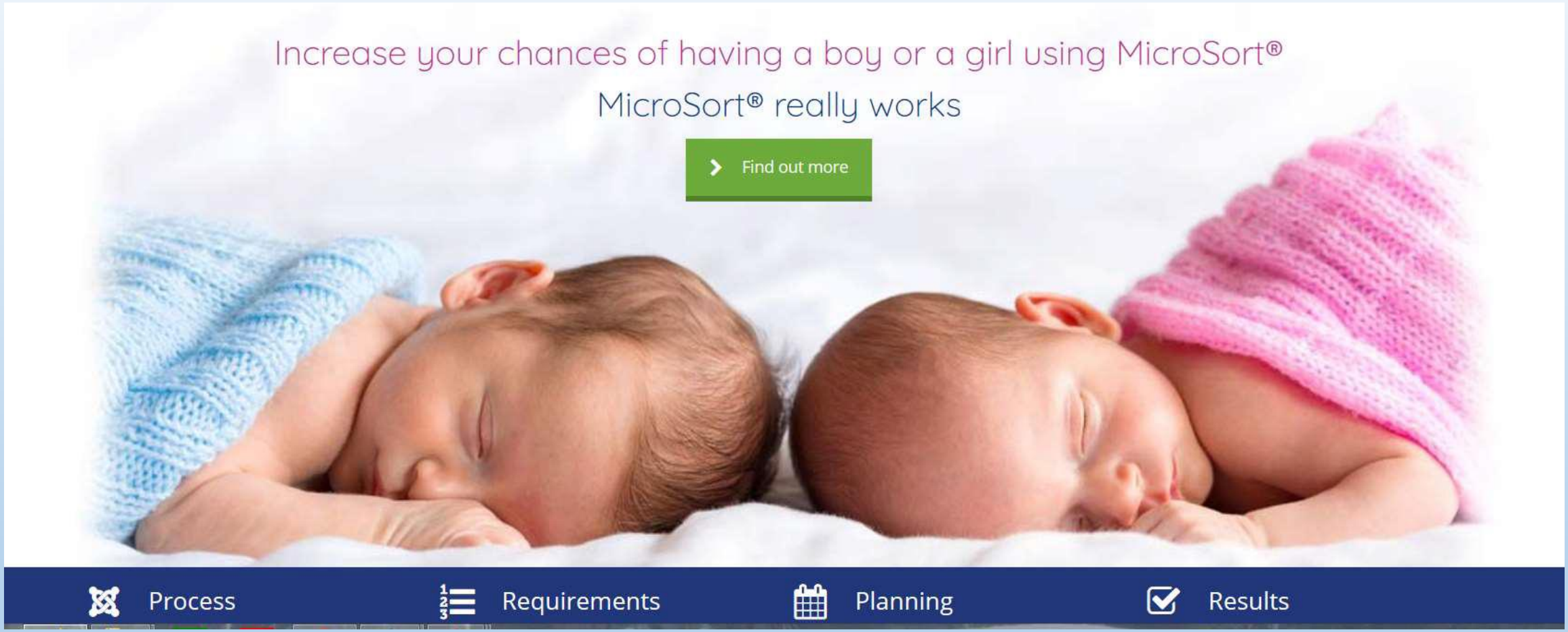
- Si ha una specifica ipotesi su un certo fenomeno nella popolazione che si vuole verificare (ipotesi **nulla** *vs* ipotesi **alternativa**)

- 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

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



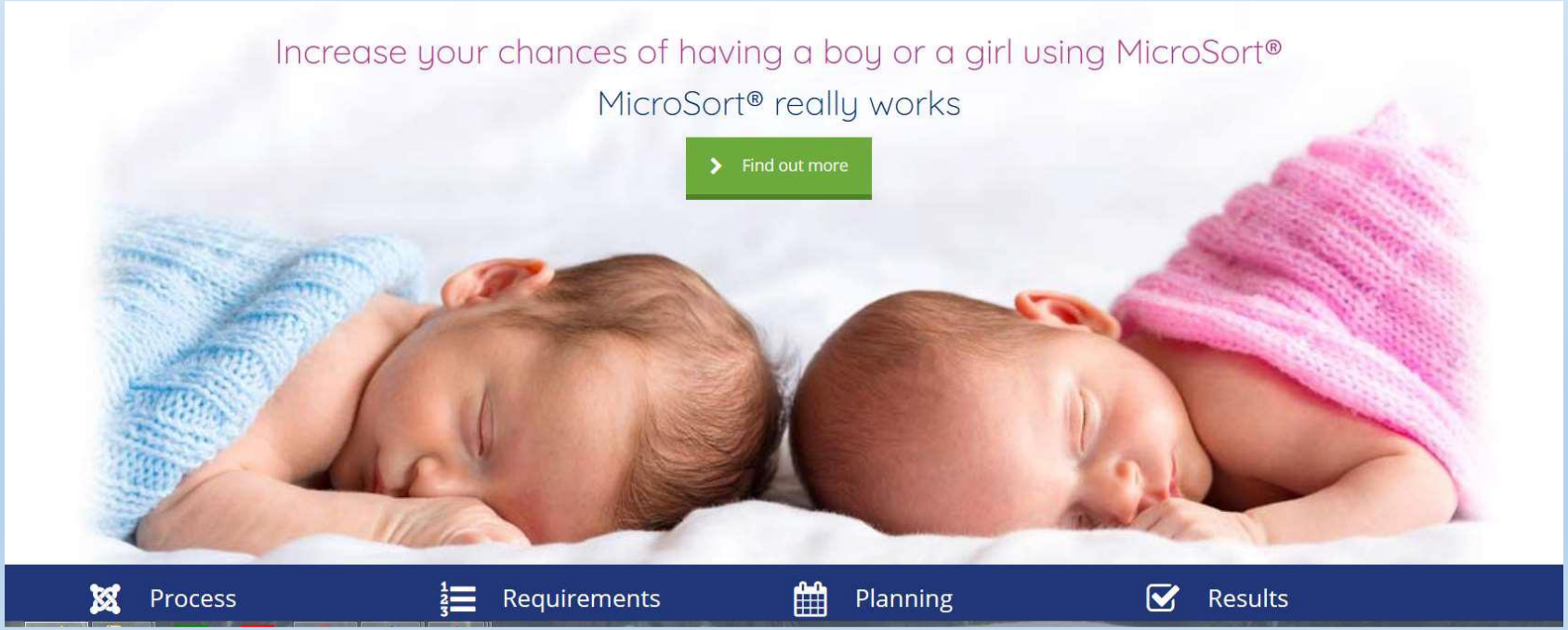
The Genetics & IVF Institute claims that its XSORT method allows couples **to increase** the probability of having a baby girl.

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…

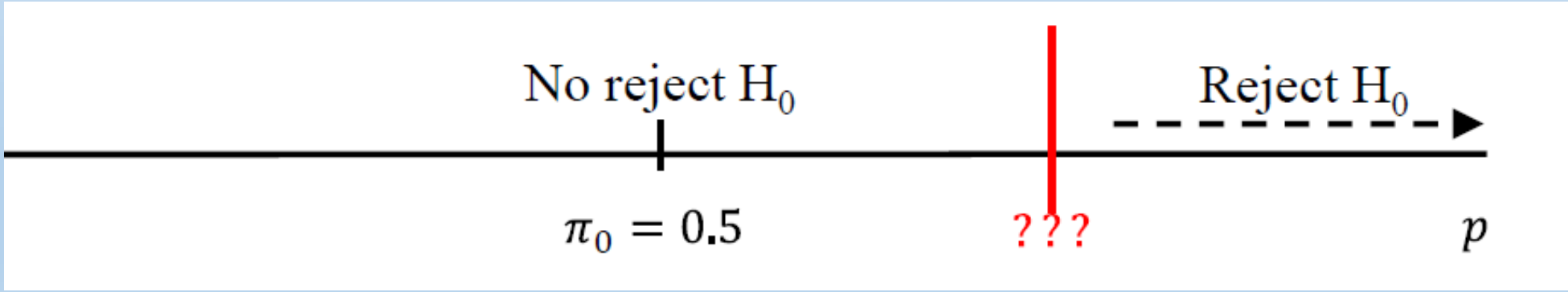
H_0: $\pi = \pi_0 = 0.5$ (null hypothesis: e.g. *the probabilty to get a girl is 50%*)

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)

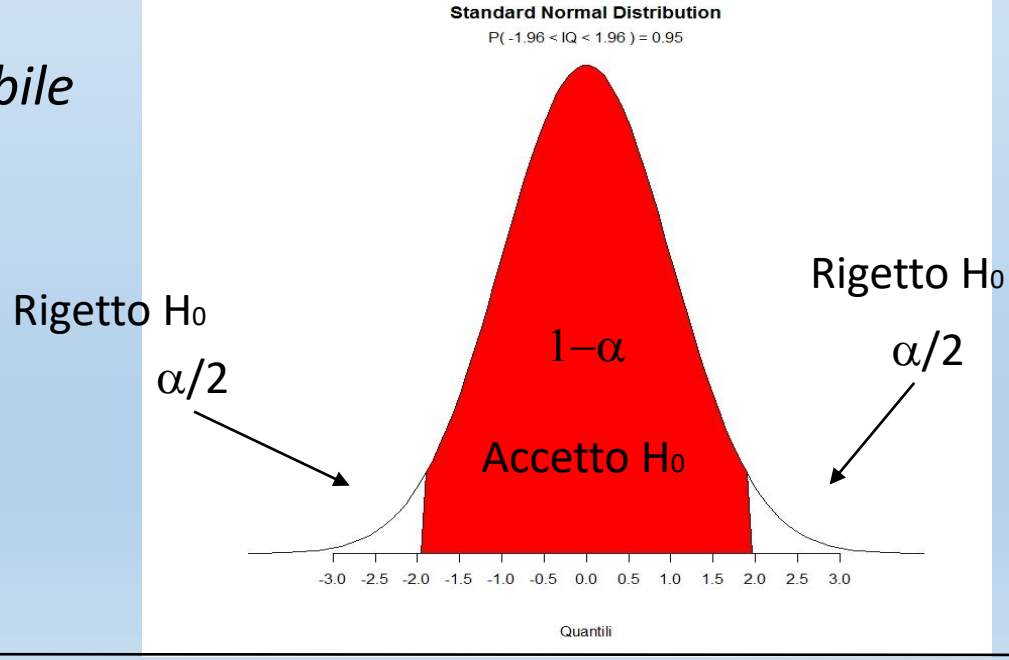
(a) Ipotesi *nulla* **H_0** *versus* Ipotesi *alternativa* **H_1** (mutualmente esclusive)

(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**

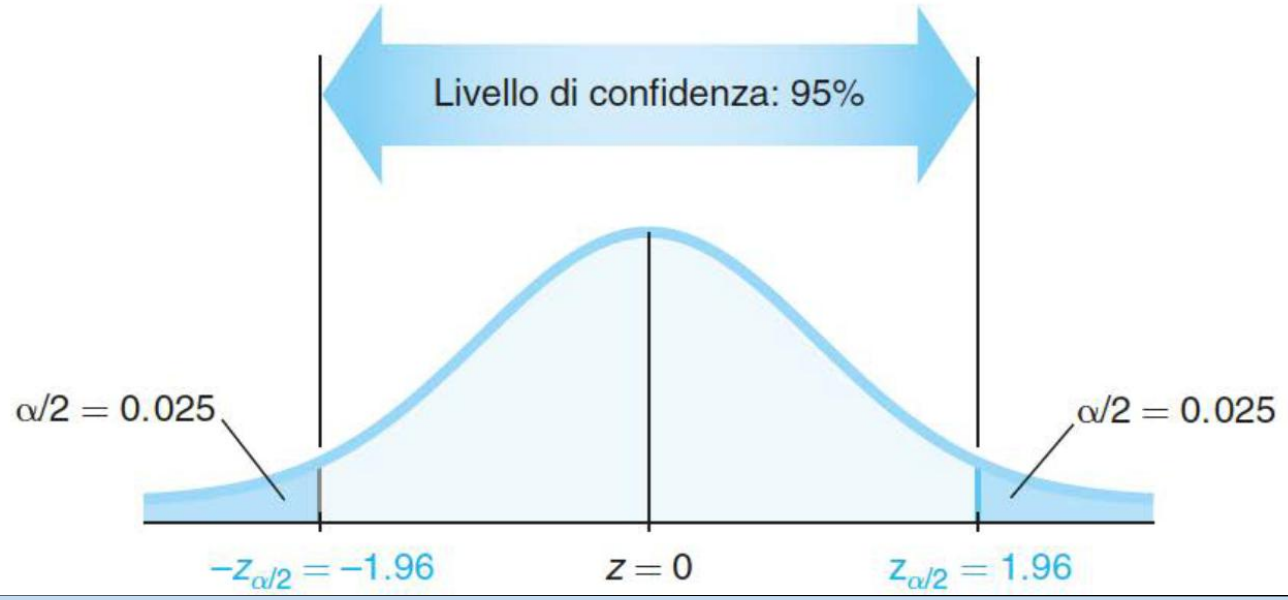
***p-value*= probabilità sotto H_0 che la variabile casuale T abbia il valore t osservato sui dati campionari o un valore più «*estremo*»**

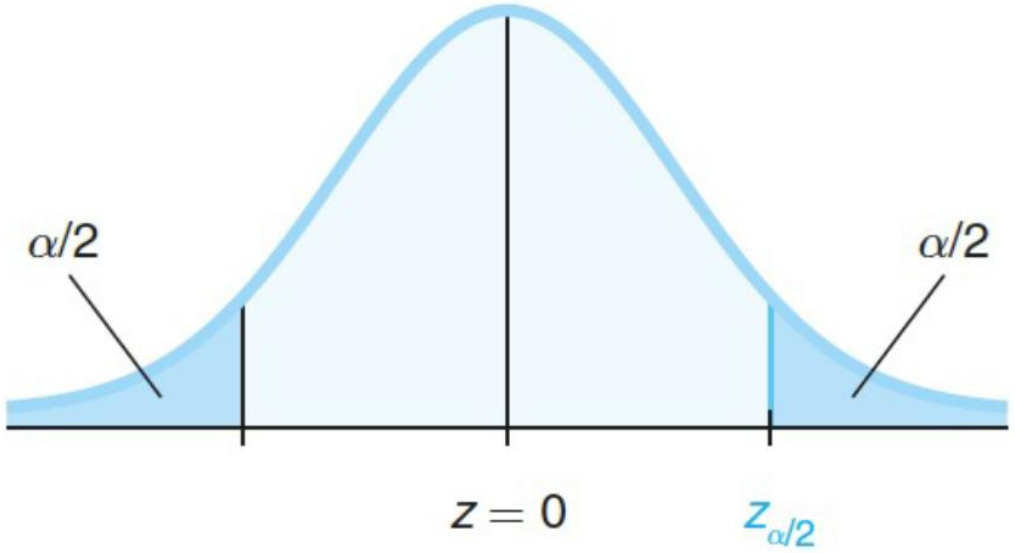


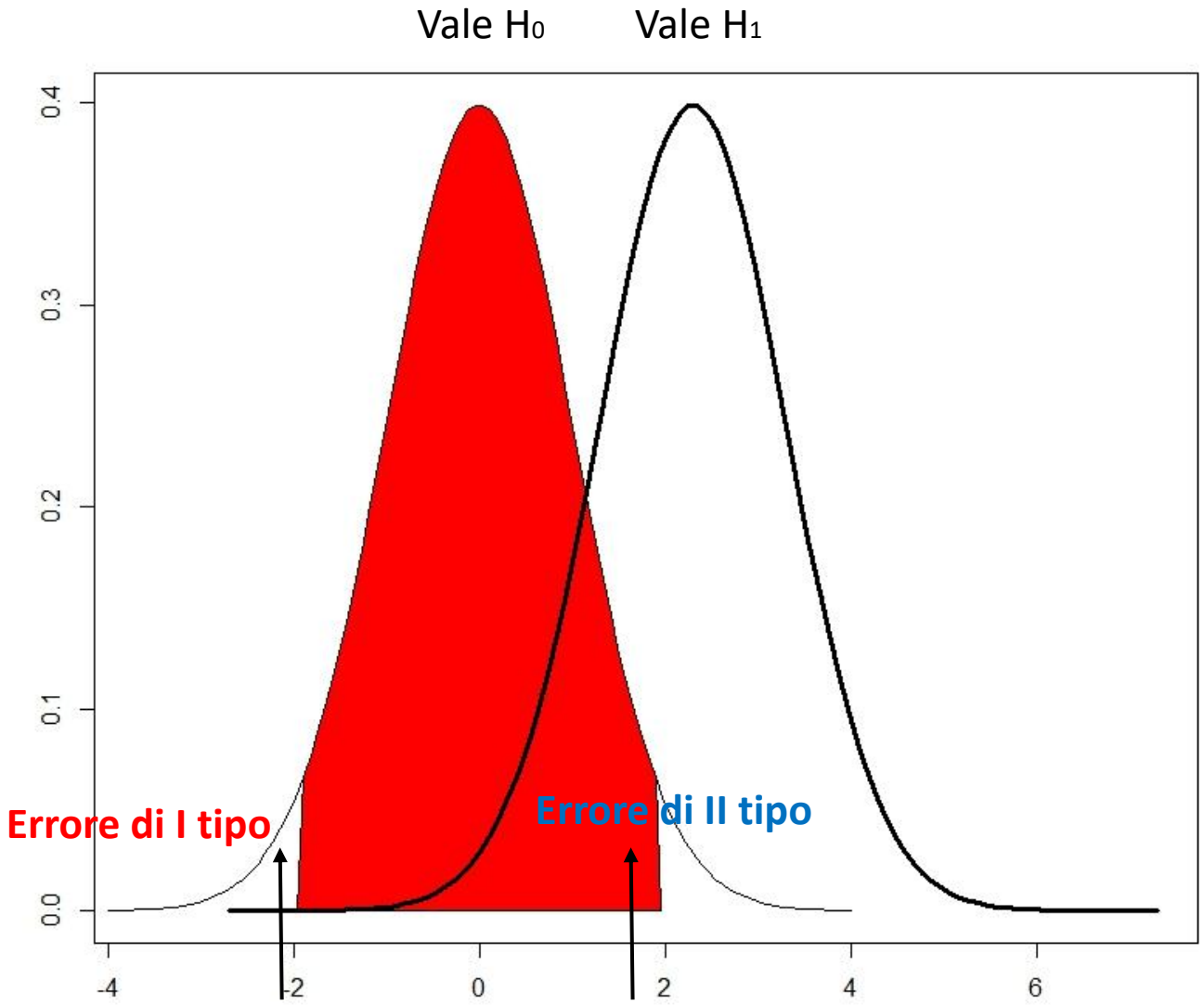
Un **valore critico** è il valore corrispondente al quantile della distribuzione che delimita l'area con maggiore possibilità di verificarsi.

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*»);

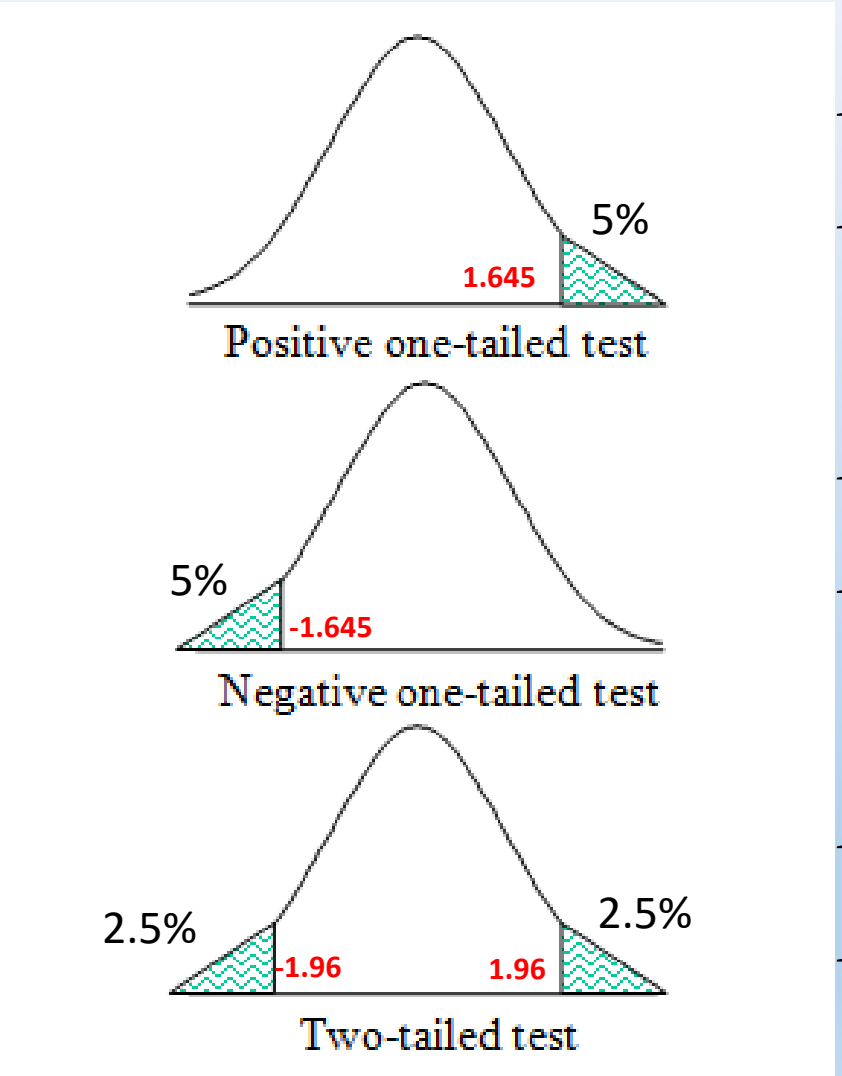
- Se H_0 non è rigettata, si conclude che i dati non offrono sufficiente evidenza per *rigettare* H_0;

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)

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%*)

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

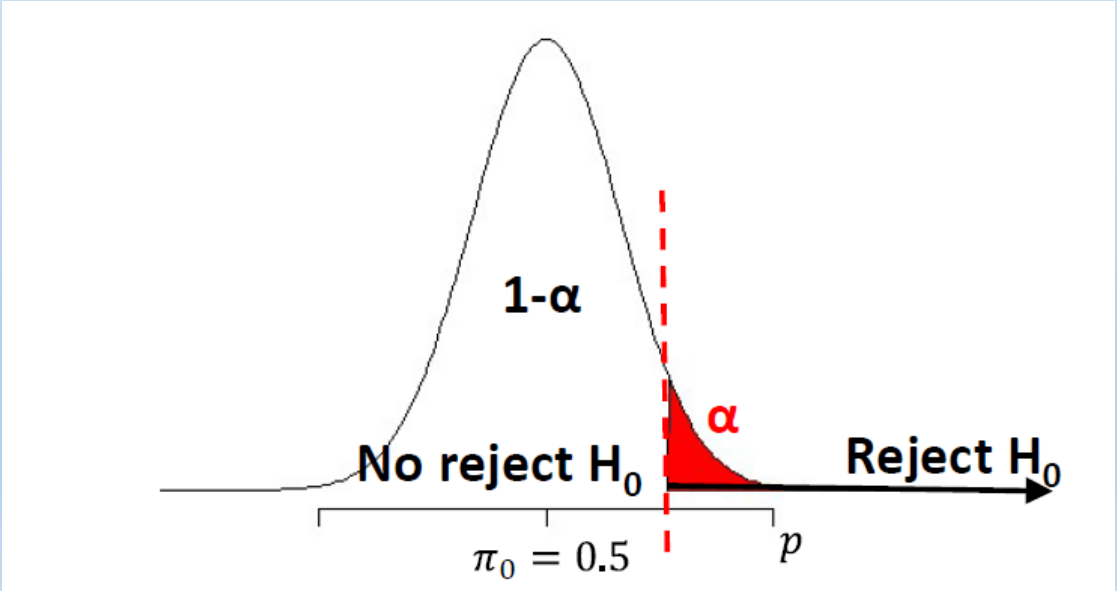
$$\bar{p} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{stimatore di p}$$

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

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

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

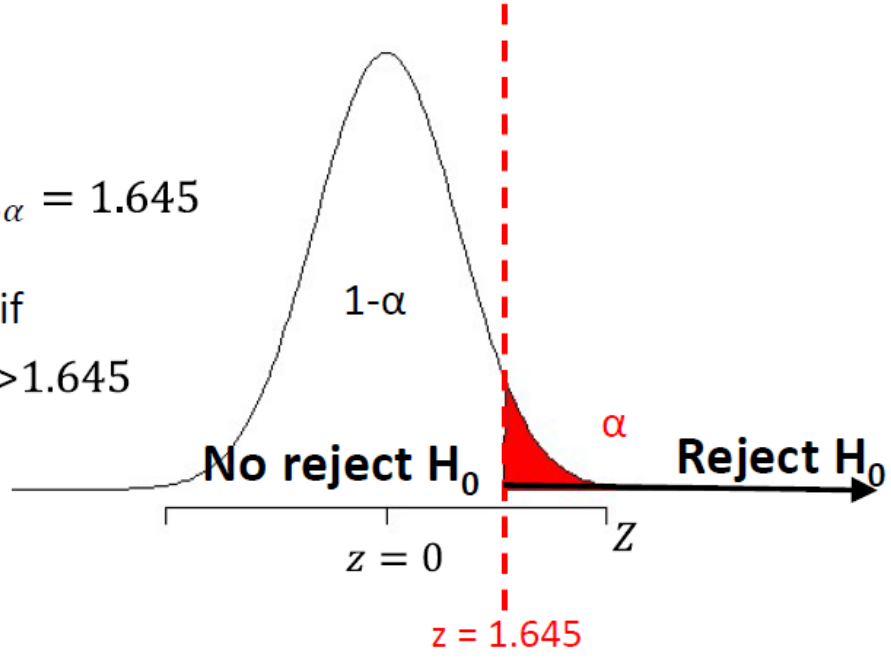
$$z = \frac{p - 0.5}{\sqrt{0.5 * 0.5/14}}$$

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

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



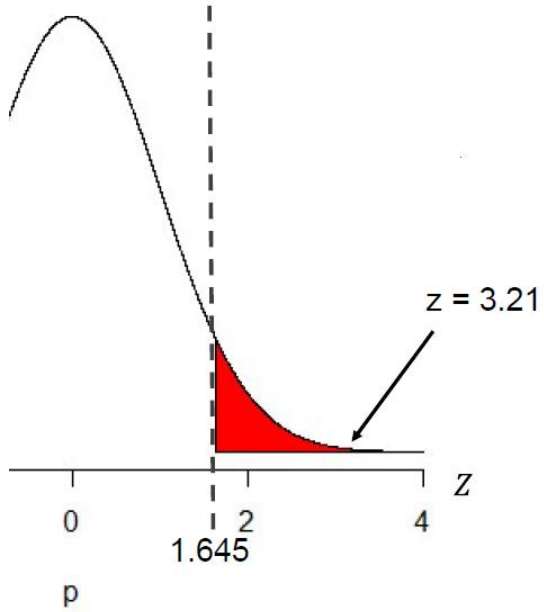
n=14

p=13/14=0.929

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

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

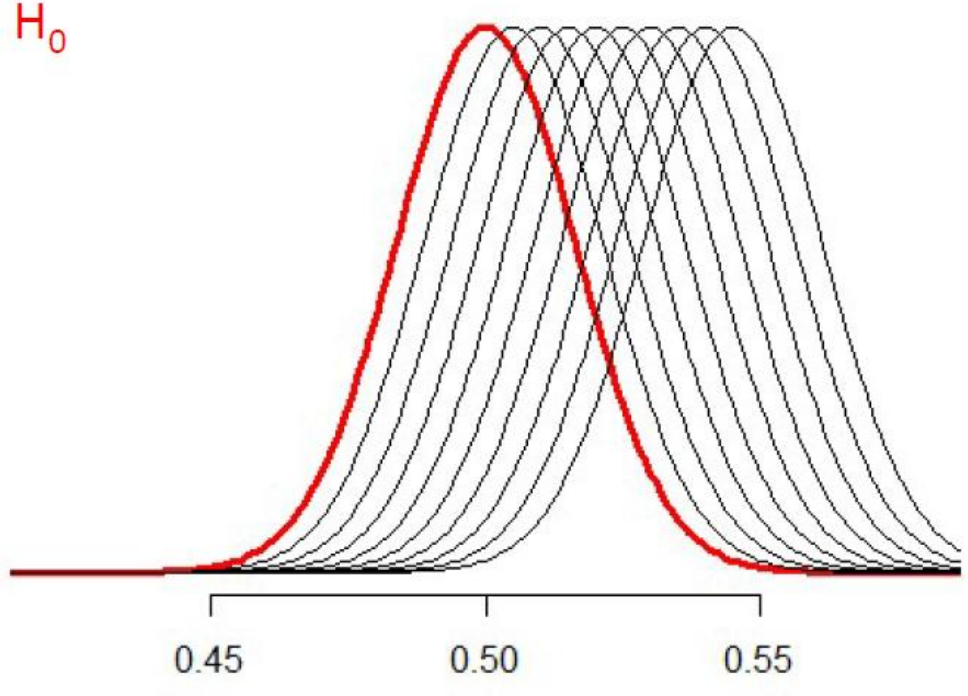
$$z = \frac{0.929 - 0.5}{\sqrt{0.5 * 0.5/14}} = 3.21$$



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.

Test di ipotesi per la differenza tra due medie: il test «t»

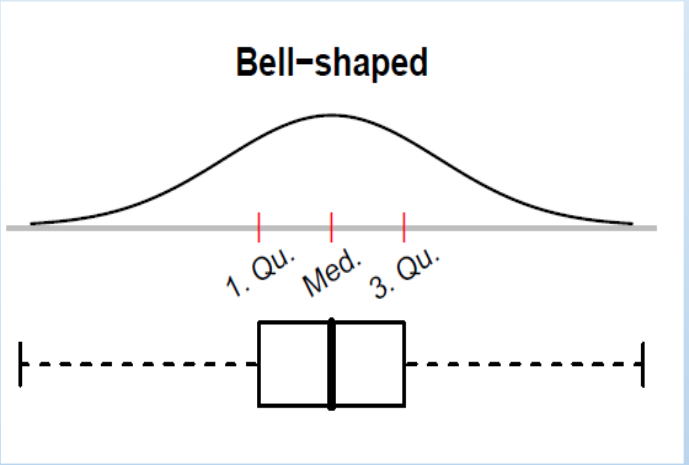
Dati due campioni del carattere numerico misurato in due gruppi:

$$H_0: \varepsilon = 0$$

$$H_1: \varepsilon \neq 0$$

$$\varepsilon = \mu_1 - \mu_2$$

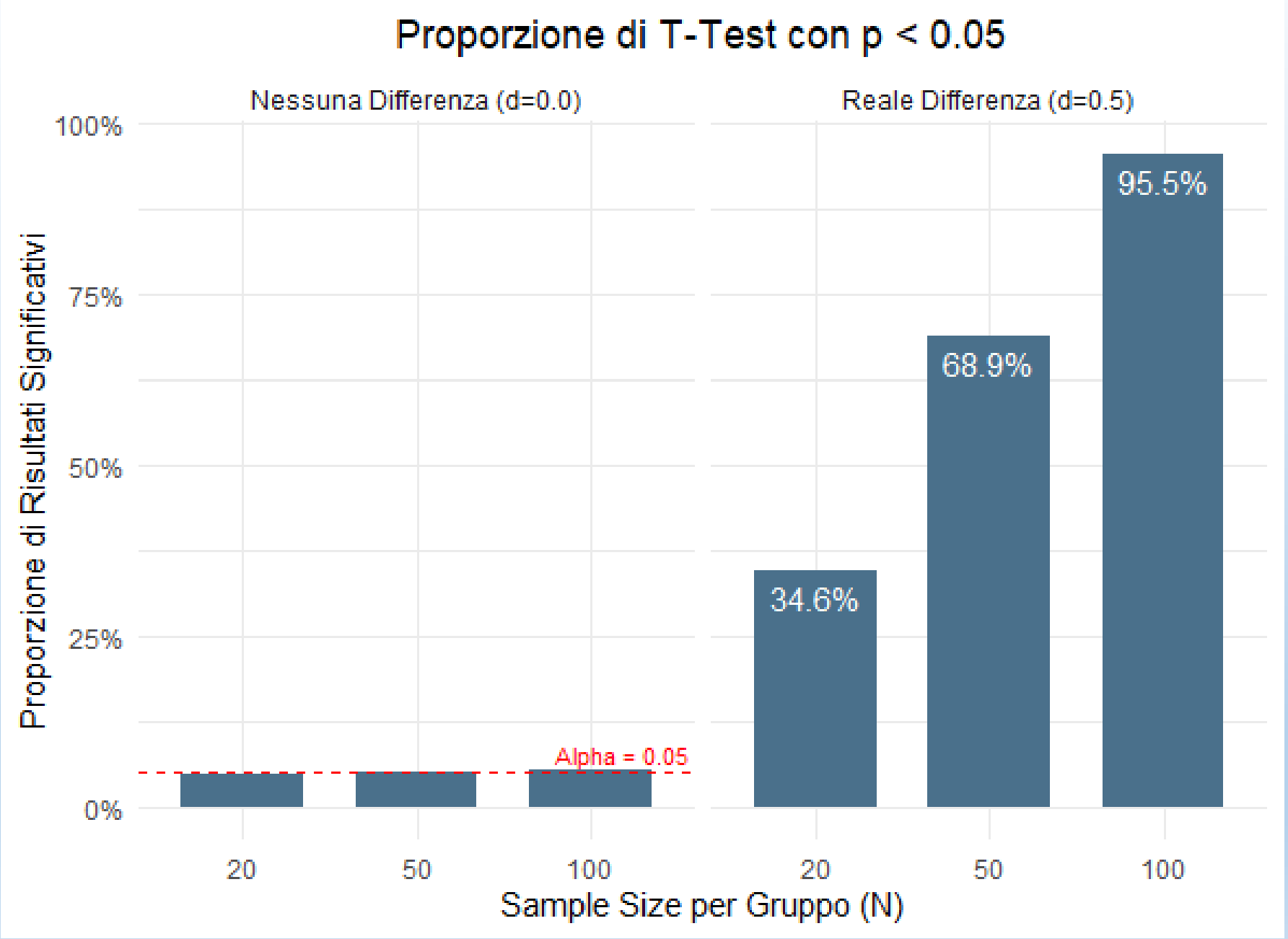
* per i quali abbia senso utilizzare la media come indice di posizione…



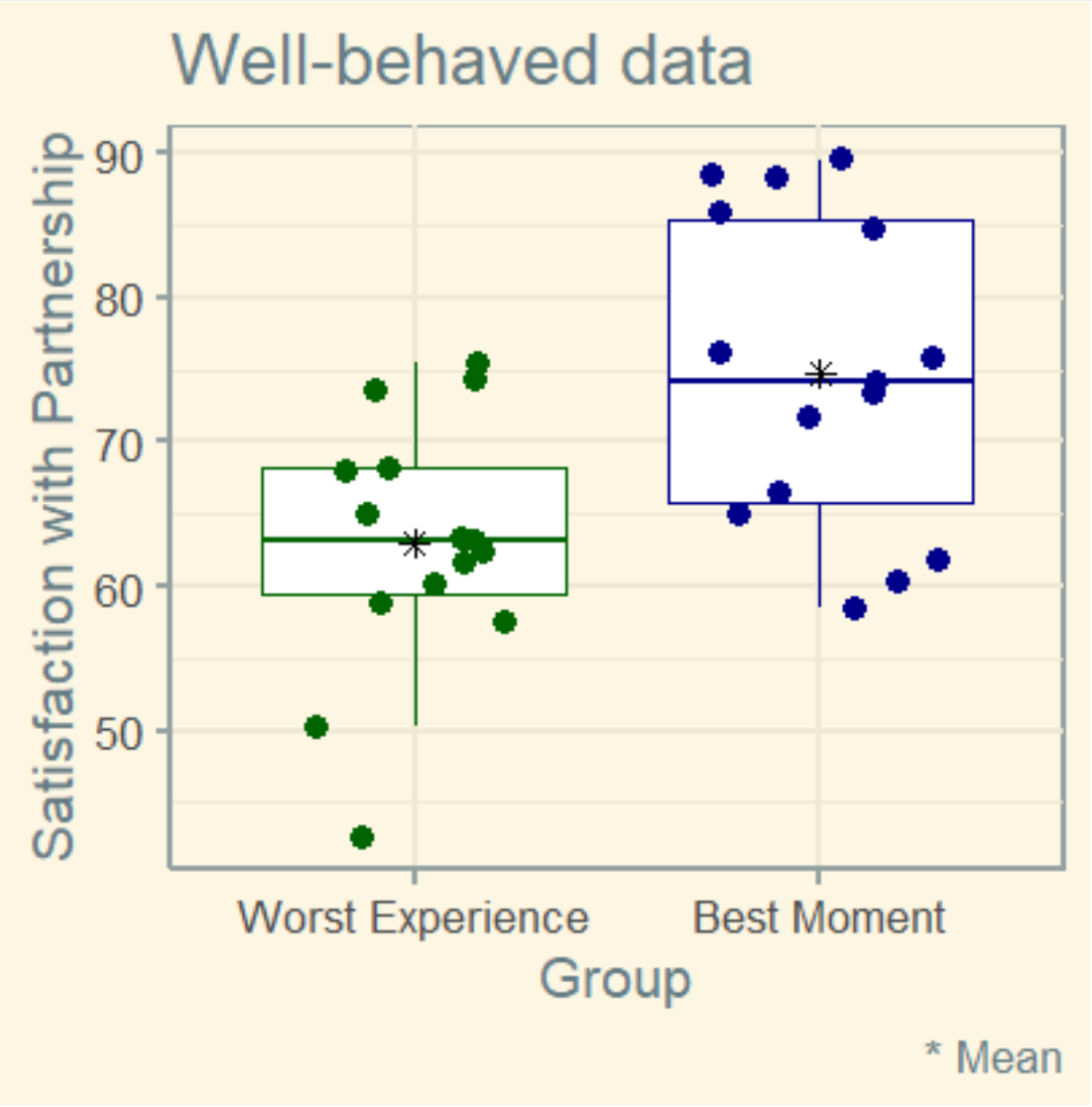
- Se le VA sono **indipendenti** → **t-test***
  Es: due gruppi diversi di pazienti
- Se le VA *non sono indipendenti* → **t-test*** per dati **accoppiati**
  Es: stesso gruppo pre-post

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

* Nella formula del t-test si divide la differenza fra le medie per lo standard error (~**effect size**)


Proporzione di T-Test con p < 0.05
Nessuna Differenza (d=0.0)
Reale Differenza (d=0.5)
Proporzione di Risultati Significativi
100%
75%
50%
25%
0%
Alpha = 0.05
34.6%
68.9%
95.5%
20
50
100
20
50
100
Sample Size per Gruppo (N)

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

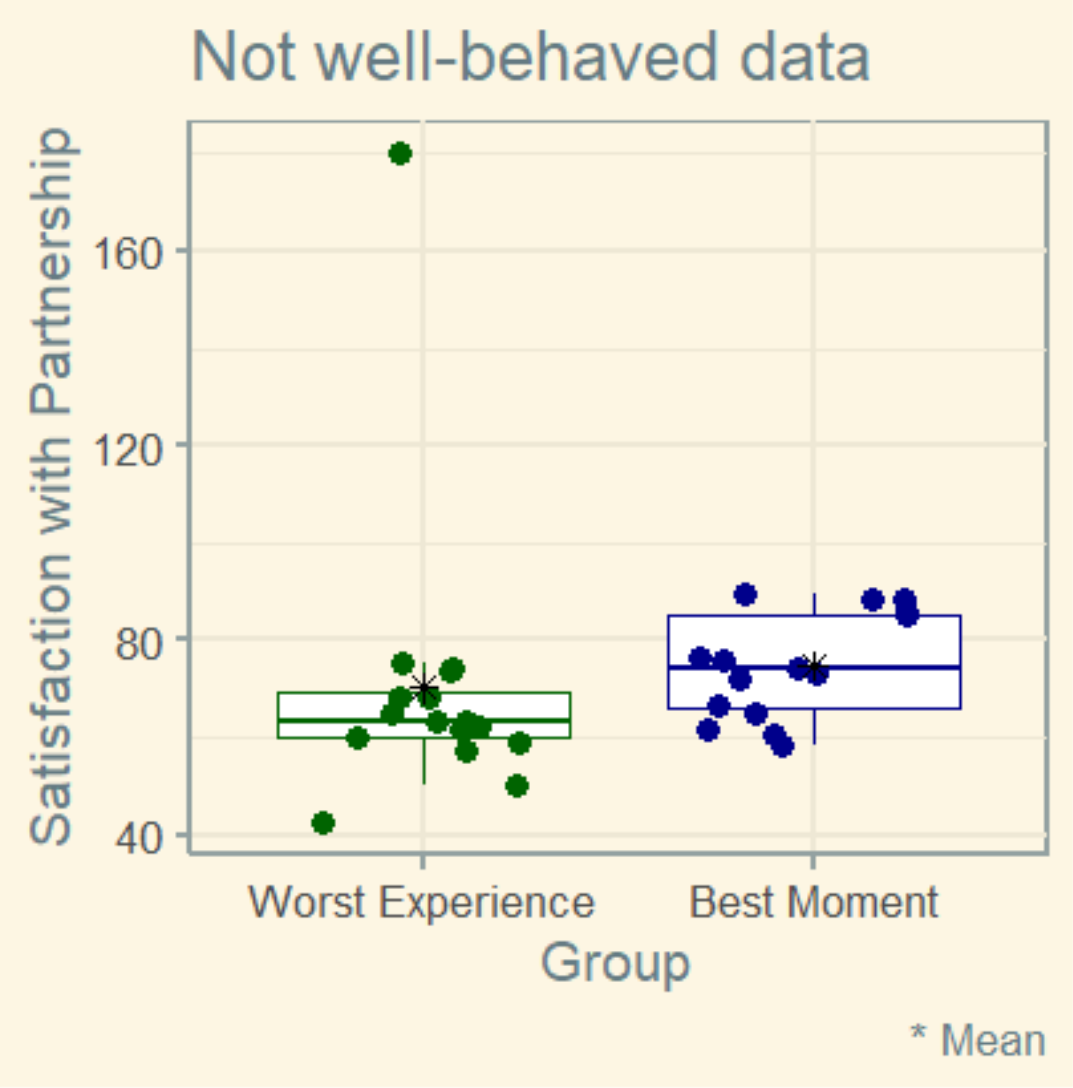


Characteristic	Worst Experience, N = 15[1]	Best Moment, N = 15[1]	p-value[2]
satisfaction	63 (9)	75 (11)	0.003

[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

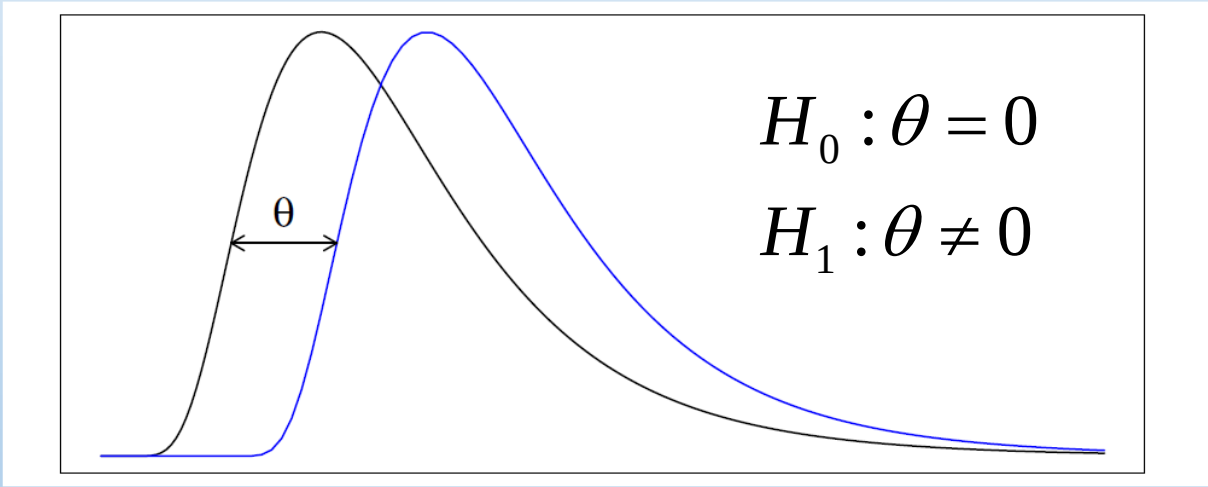
Il t-test (opzione varianze non omogenee) **non trova** una differenza significativa tra le medie di questi due gruppi ….

Come risolvere [se non è possibile trasformare]?

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

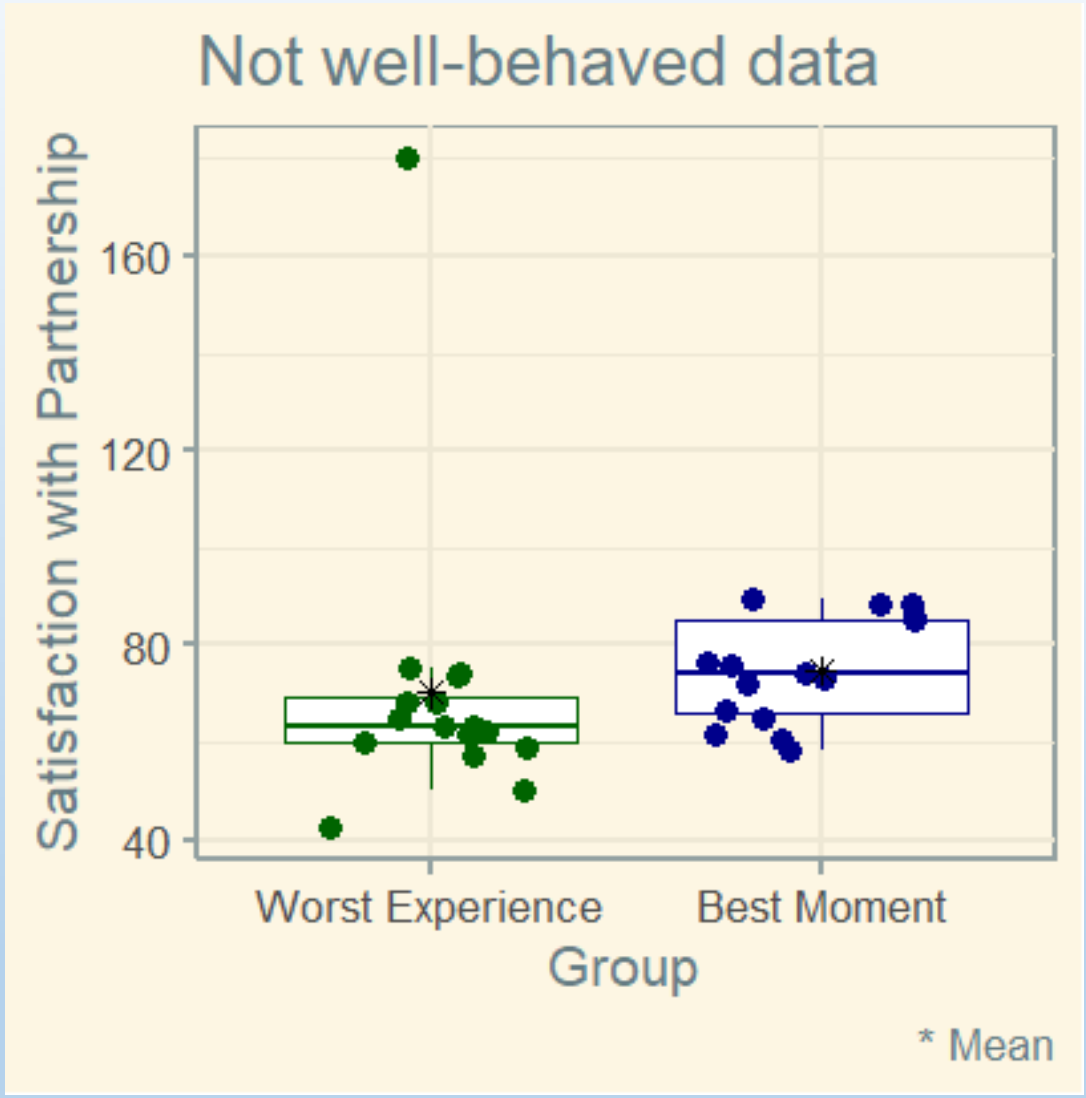
Quando la media può non essere più ***rappresentativa*** della forma della distribuzione…

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



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** una differenza significativa nelle distribuzioni di questi due gruppi.

Test di ipotesi per dati di tipo categorico/classi

- I dati di tipo *categorico* possono essere descritti tramite una **tabella di contingenza**
- Il **test chi-quadrato** confronta i valori osservati in ogni cella della tabella rispetto a quelli che ci saremmo attesi **se non ci fosse associazione** tra la variabile sulle righe e la variabile sulle colonne
- L'ipotesi nulla è quindi: **non c'è associazione** tra i due fenomeni

Criterion 2	Criterion 1					
	1	2	3	. . .	C	Total
1	n_{11}	n_{12}	n_{13}	. . .	n_{1c}	r_1
2	n_{21}	n_{22}	n_{23}	. . .	n_{2c}	r_2
3	n_{31}				⋮	
⋮	⋮				⋮	
r	n_{r1}	. . .	. . .	. . .	n_{rc}	r_r
Total	c_1	c_2			c_c	n

La relazione tra una malattia e un fattore di esposizione può essere descritta tramite una tabella di contingenza (valori osservati=O_i):

Disease			
Exposure	Yes	No	Total
Yes	37	13	50
No	17	53	70
Total	54	66	120

37/54 = 68 % di **individui malati** era esposta
13/66 = 20 % di **individui non malati** era esposta

Questi dati suggeriscono una associazione tra malattia ed esposizione?

Sotto l'ipotesi nulla di **assenza di associazione***, i valori attesi (E_i) nelle celle sarebbero:

Disease			
Exposure	Yes	No	Total
Yes	50/120 x 54 = 22.5	50/120 x 66 = 27.5	50
No	70/120 x 54 = 31.5	70/120 x 66 = 38.5	70
Total	54	66	120

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

Differenze valori osservati vs attesi

Chi quadro= 29.1

La probabilità di ottenere questo valore sotto l'ipotesi nulla è < 0.001

- Il chi-quadro è un test di associazione tra due variabili di tipo categorico
  [o continue ma raggruppate in classi];

- Si può applicare per studiare la associazione tra una malattia ed un fattore di esposizione in uno studio di coorte o in uno studio caso-controllo non appaiato o in uno studio trasversale (cross-sectional study);

- **Il test di McNemar è una variante del Chi-quadro per dati appaiati** e può essere utilizzato negli studi caso-controllo **appaiati**

Adverse effect to a drug	Post-drug condition: present	Post-drug condition: absent
Pre-drug condition: present	…	…
Pre-drug condition: absent	…	…

Il test del Chi-quadro rigetta (o non rigetta) l'ipotesi nulla di INDIPENDENZA tra le due variabili.

Il test del Chi-quadro non ci dice nulla sulla eventuale «FORZA» della associazione tra le due variabili.

*Come facciamo a quantificare la FORZA della associazione tra un **esito** (variabile **dicotomica**) e un fattore di **esposizione** (variabile che può essere su qualsiasi scala di misura) ?...*

Statistical significance is the least interesting thing about the results. You should describe the results in terms of measures of magnitude –not just, does a treatment affect people, but how much does it affect them.

-Gene V. Glass[1]

The primary product of a research inquiry is one or more measures of effect size, not P values.

-Jacob Cohen[2]

What is the ***weight of evidence*** for Hypothesis 1 compared to Hypothesis 0 ??

Frequentist (Fisher) approach:

Compute: p(extremeness of the data | H_0 is true)

Bayesian approach:

Compute: p(data | H_0 is true) / p(data | H_1 is true)

Applying Fisher's approach to the case of Sally Clark

- 1996: Clark's **1st son died** a few weeks after birth ("Sudden Infant Death Syndrome" (SIDS)?)
- 1998: Clark's **2nd son died** a few weeks after birth (SIDS again????)
- 1999: Clark was **found guilty of murder** and given two life sentences

- H_0 : babies died from SIDS aka "crib death"
- SIDS occurence rate is 1 in 8,500
- The chance of this happening twice is 1 in 73 million, i.e., p = 0.0000000137
- Therefore, H_0 is rejected
- Therefore, she *must be* guilty (double murder)

What is wrong with this line of reasoning?

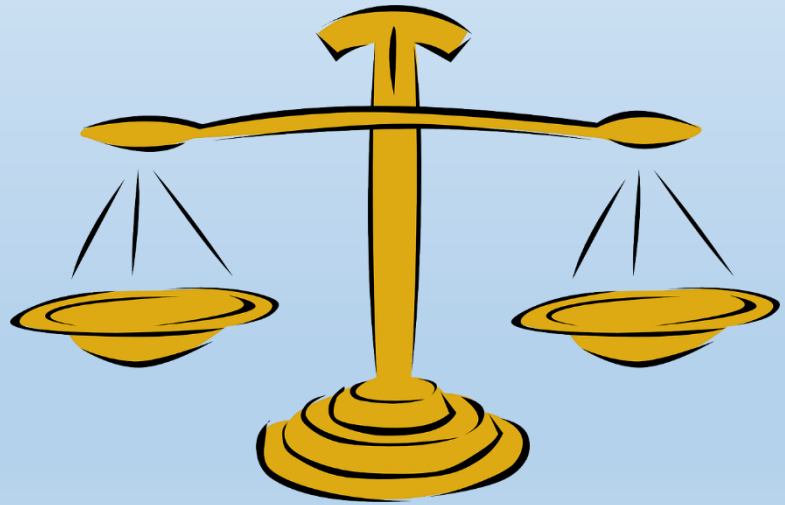
Even though H_0 is unlikely, other hypotheses may be *even more* unlikely!!

Evidence is best treated as a relative concept:

How improbable is H_0 ?

How (im)probable is H_0 *relative to* H_1 ?

- H1: double murder
- Infant murder rate in UK: approximately 1 in 33,000(*)
- The chance of this happening twice is 1 in 1.1 billion, i.e., p = 0.000000000918
- SIDS *is 15 times more likely* than murder!



(*) Marks, M. N., & Kumar, R. (1993). Infanticide in England and Wales. *Medicine, Science and the Law*,*33*(4), 329-339.

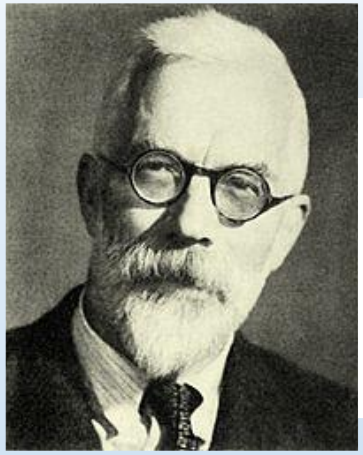
How did it end for Clark?

- 1996: Clark's **first son died** suddenly within a few weeks of his birth
- 1998: Clark's **second son died** suddenly within a few weeks of his birth
- 1999: Clark was **found guilty of murder** and given two life sentences
- 2003: Clark is set free, yet highly traumatized
- 2007: Clark dies from alcohol poisoning

[The deeper problem here:

Some events are unlikely under *any* hypothesis…]

https://unherd.com/2024/05/the-danger-of-trial-by-statistics/



Solution: lower the α value for "rare" events?

. . . no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas.

Sir Ronald A. Fisher (1956)

Some events are unlikely under *any* hypothesis

Should we then reject them all and consider the event unexplainable?

…However: how to do this without knowing the *cause* of the event??

The Bayes factor

$$\frac{p(H_0|D)}{p(H_1|D)}$$

Probability of Hypothesis 0, given the data

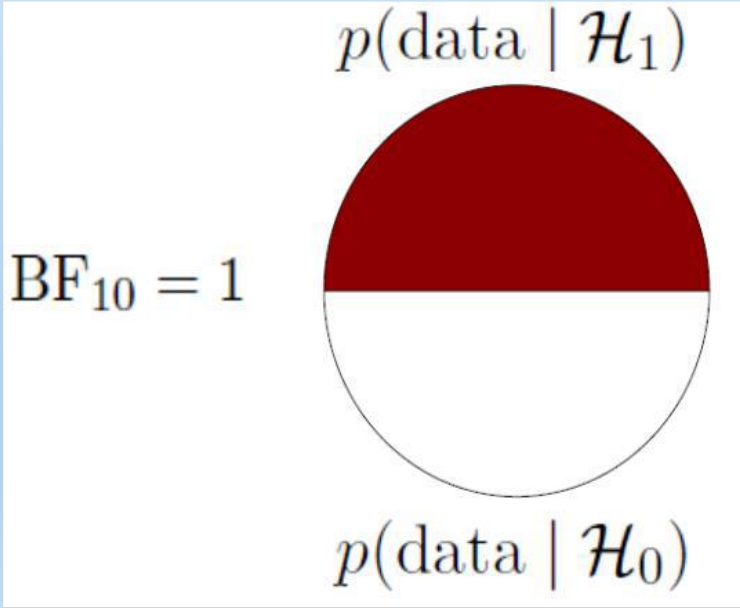
Probability of Hypothesis 1, given the data

$$\frac{p(D|H_0)}{p(D|H_1)}$$

BAYES FACTOR: indicates how many times more likely the data are under H0 compared to H1

BF indicates the change from *prior odds* to *posterior odds* brought about by the data

Visual interpretation of the Bayes factor:



The Bayes factor & the Bayes theorem

$$BF = \frac{p(D|H_0)}{p(D|H_1)} = \frac{\frac{p(H_0|D)p(D)}{p(H_0)}}{\frac{p(H_1|D)p(D)}{p(H_1)}}$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$BF = \frac{p(H_0|D)p(H_1)}{p(H_1|D)p(H_0)}$$

Bayes factor

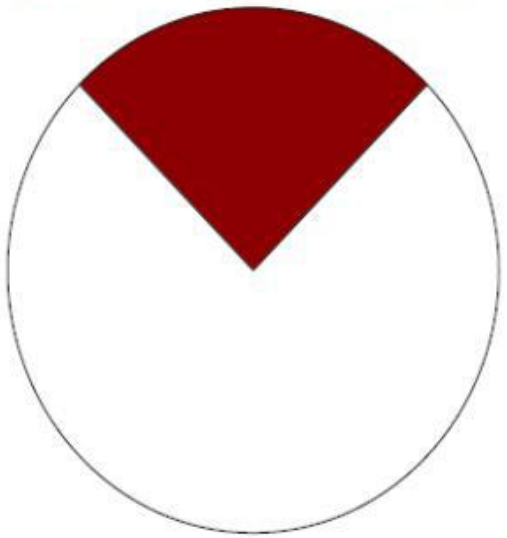
Prior Odds

$$\frac{p(H_0|D)}{p(H_1|D)} = \frac{p(D|H_0)}{p(D|H_1)}\frac{p(H_0)}{p(H_1)}$$

Posterior Odds

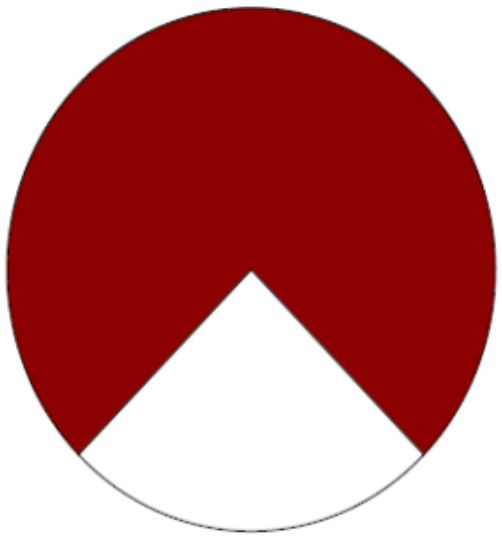
$$\mathrm{BF}_{10} = \frac{1}{3}$$

$$\mathrm{BF}_{01} = 3$$



$$\mathrm{BF}_{10} = 3$$

$$\mathrm{BF}_{01} = \frac{1}{3}$$



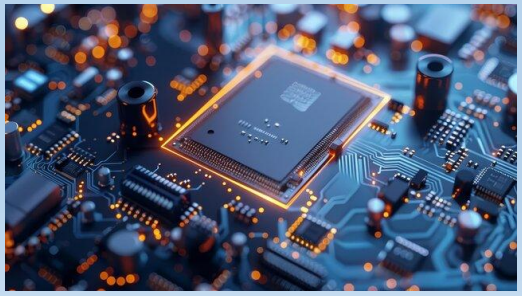
Bayes factor

- Evidence is always **relative**(w.r.t. alternative hypotheses)
- Can **reject** *and* **support** hypotheses
- Less confusing?
- Computationally **expensive**
- Requires specification of **priors**

"Subjective"

"Objective"

Today we have computers much more powerful ….



***p* value**

- Evidence is **absolute** (about single hypothesis)
- Can only **reject** hypotheses
- Confusing for non-statisticians
- Computationally simple