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

Sommario

- Obiettivi-chiave nel disegno di studio
- ***Precisione*** e dimensione
- Il concetto di ***effect size***
- **Potenza** di un test
- Esempi di calcolo: RCT
- Bayes factor (cenni)

«Just as a recipe makes more sense if the cook is somewhat familiar with the ingredients, sample size calculations are easier if the investigator is acquainted with the basic concepts"

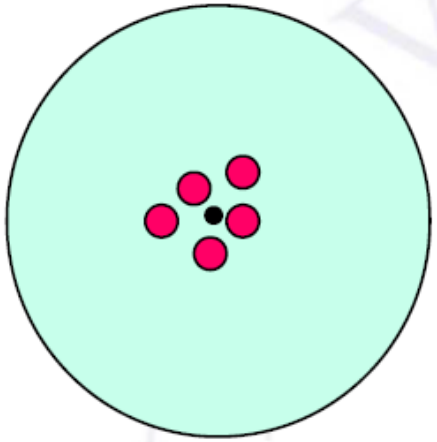
Obiettivi-chiave nel Disegno di Studio

Validità: evitare l'errore «sistematico» nelle stime

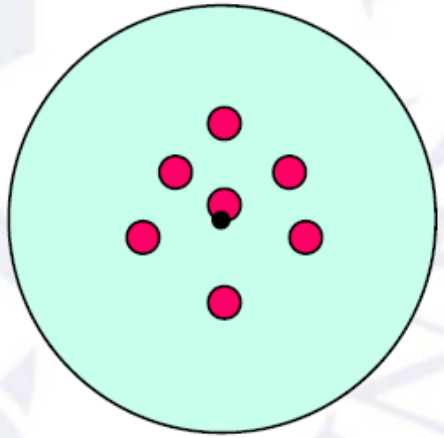
- **Validità interna** = le differenze nell'*outcome* sono «causate» dall'esposizione o da un errore sistematico/presenza di confondenti?
- **Validità esterna** = è possibile generalizzare quanto osservato ad altre popolazioni? [«rappresentatività» del campione]

Precisione/stima di effetti: data la variabilità *random* delle stime -> ampiezza degli intervalli di confidenza/potenza dei test

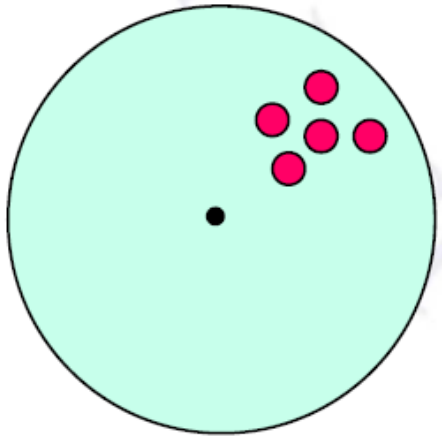
- **Dimensione campionaria**/bilanciamento dei gruppi/covariate …



Valide e precise



Valide
non precise



Non valide ma precise

Validità

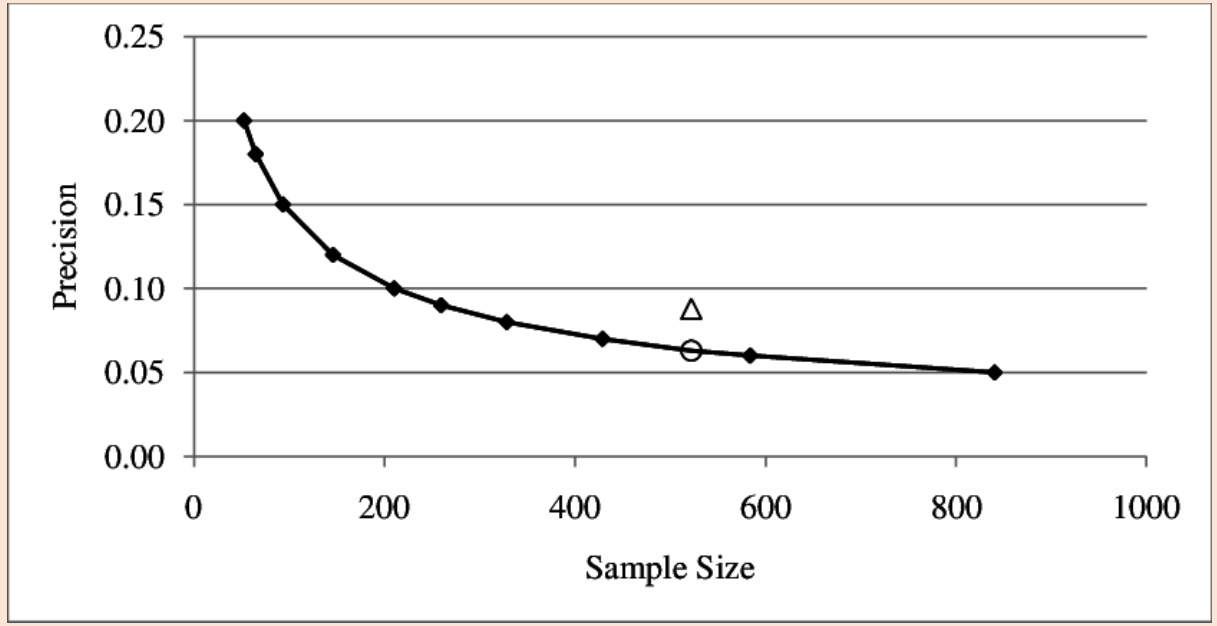
Il grado di accordo tra la stima effettuata e il valore "vero" del fenomeno

Precisione

Il grado di variabilità (dispersione) delle misure effettuate intorno al valore "vero"

Due strategie di base per la dimensione campionaria

Precisione (intervalli di confidenza)



Potenza del test di ipotesi (**effect size**)

		True state of H_0 (Unknown)	
		H_0 true	H_0 false
Decision (sample data)	Reject H_0	Type I error*	ok
	Do Not reject H_0	ok	**Type II error****

Precisione sulla stima della media

Teorema del limite centrale: $\bar{y} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

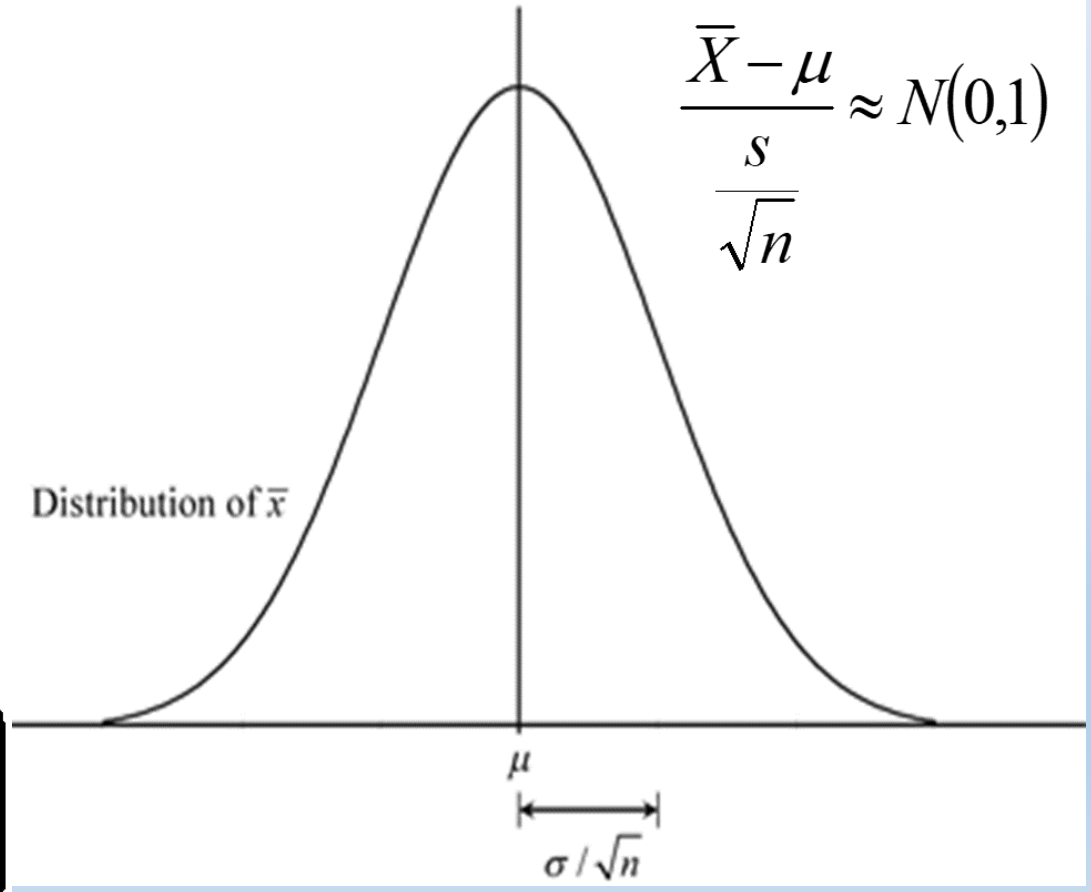
$z_{\frac{\alpha}{2}} = \frac{\alpha}{2}$ *percentile* $N(0,1)$

The maximum error E around the value of μ that one is willing to accept could be then defined as:

$$E = |\bar{y} - \mu| = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$n \geq \frac{z_{\frac{\alpha}{2}}^2 * \sigma^2}{E^2}$$

$$P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$



Suppose that an estimate is desired of the average price of tablets of a tranquilizer.

A random sample of pharmacies is selected. The estimate is required to be **within 10 cents [under/over]** of the true average price with 95% confidence. *Based on a small pilot study* the standard deviation in price can be estimated as 85 cents.

How many pharmacies should be randomly selected ?

$$n^* = \left(\frac{z_{\frac{\alpha}{2}} * \sigma}{d^*}\right)^2 = \left(\frac{1.96 * 0.85}{0.10}\right)^2 = 277.56$$

a sample of **278** pharmacies should be taken.

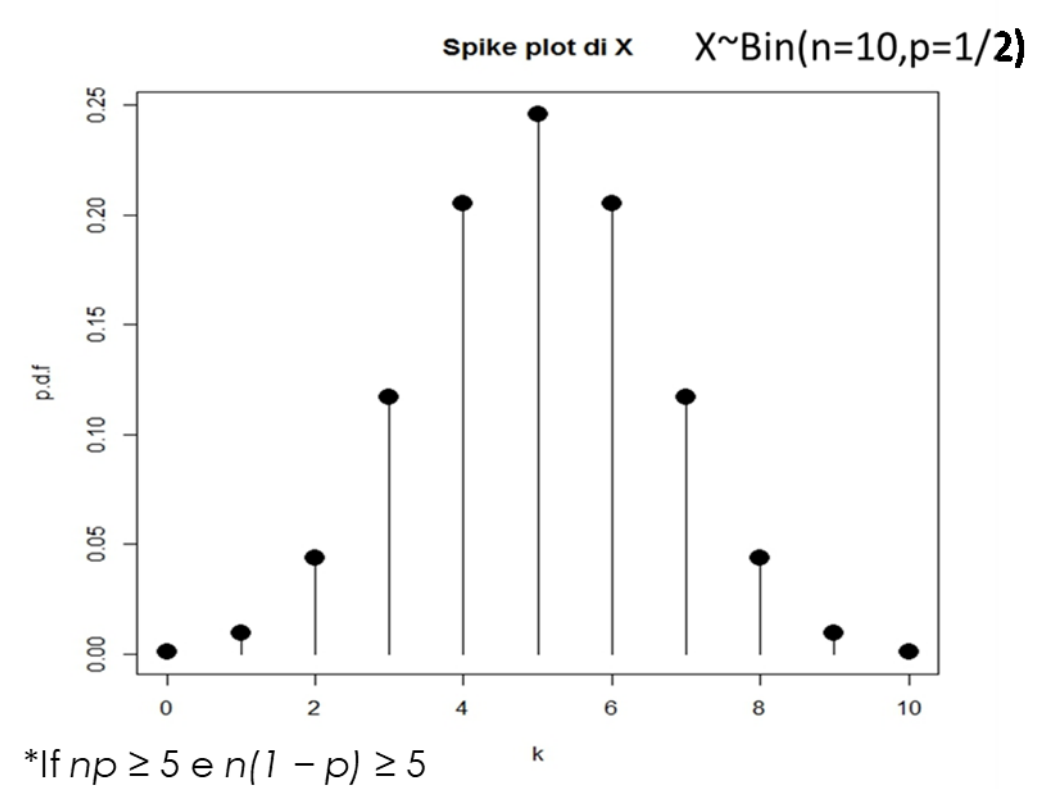
σ^2 is **crucial** for determining the optimal sample size. We can use previous studies or pilot studies.

However, it is advisable to *overestimate* the variance (rather than underestimate it) as it is better to use a sample size that is *too high* rather than one that is *too low*.

Precisione sulla stima di una proporzione

Here too we must set the **precision** that we want for the estimate.

How large the sample has to be to estimate an (unknown) proportion p with a precision of E ?



$$n \geq \frac{z_{\frac{\alpha}{2}}^2 * p * (1-p)}{E^2}$$

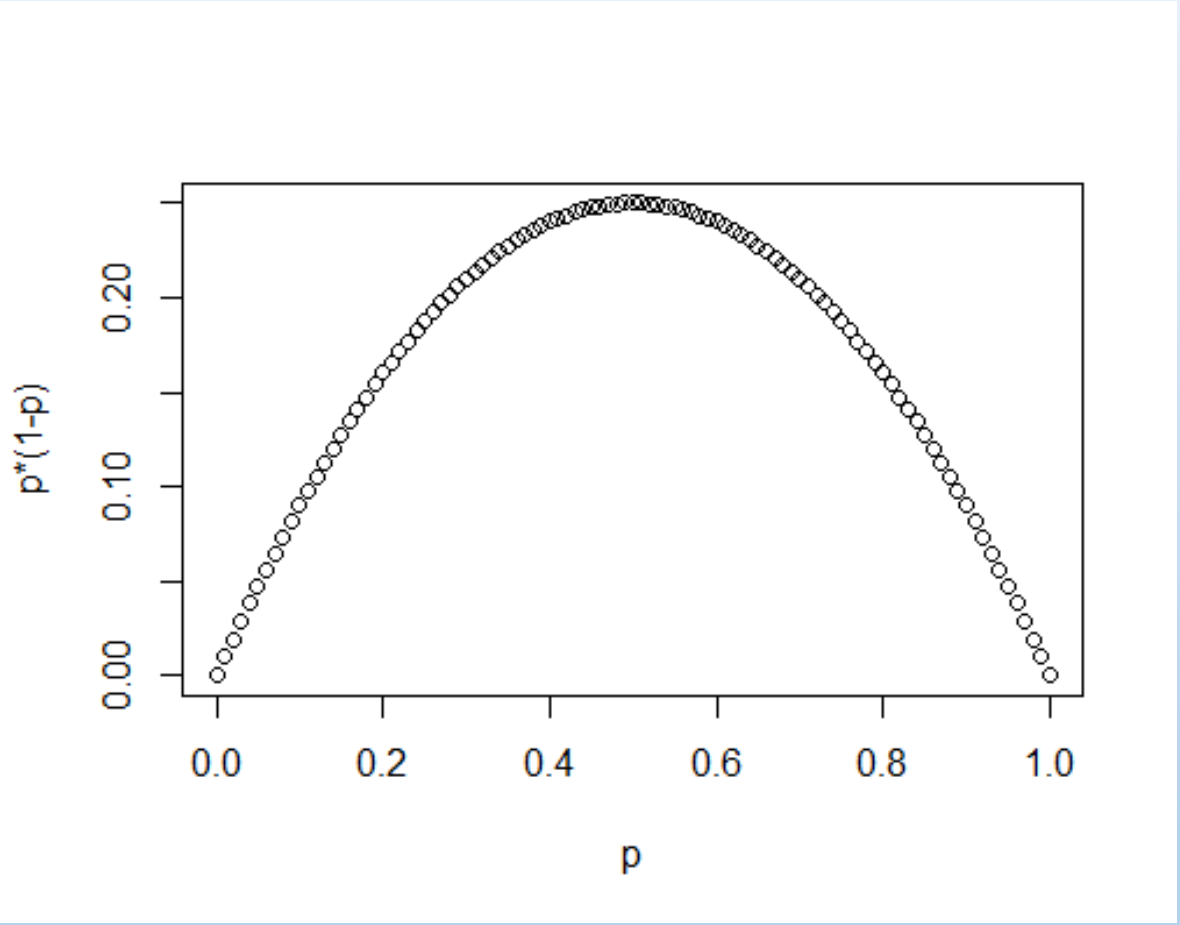
N.B. Assuming p= 0.5 for the unknown proportion we will always have the "largest" sample possible !!

The precision of the estimator depends on the value assumed by p which is unknown.

It could be appropriate to use (conservative estimate) :
p =0.5 ; p *(1- p)=0.25

If we wish to calculate the minimum size required to have an interval for p that does not exceed the half maximum amplitude E^* we must look for the minimum value of n such that:

$$z\frac{\alpha}{2} * \sqrt{\frac{0.25}{n}} \leq E^*$$



$$n^* \geq 0.25 * \left(\frac{Z\frac{\alpha}{2}}{E^*}\right)^2$$

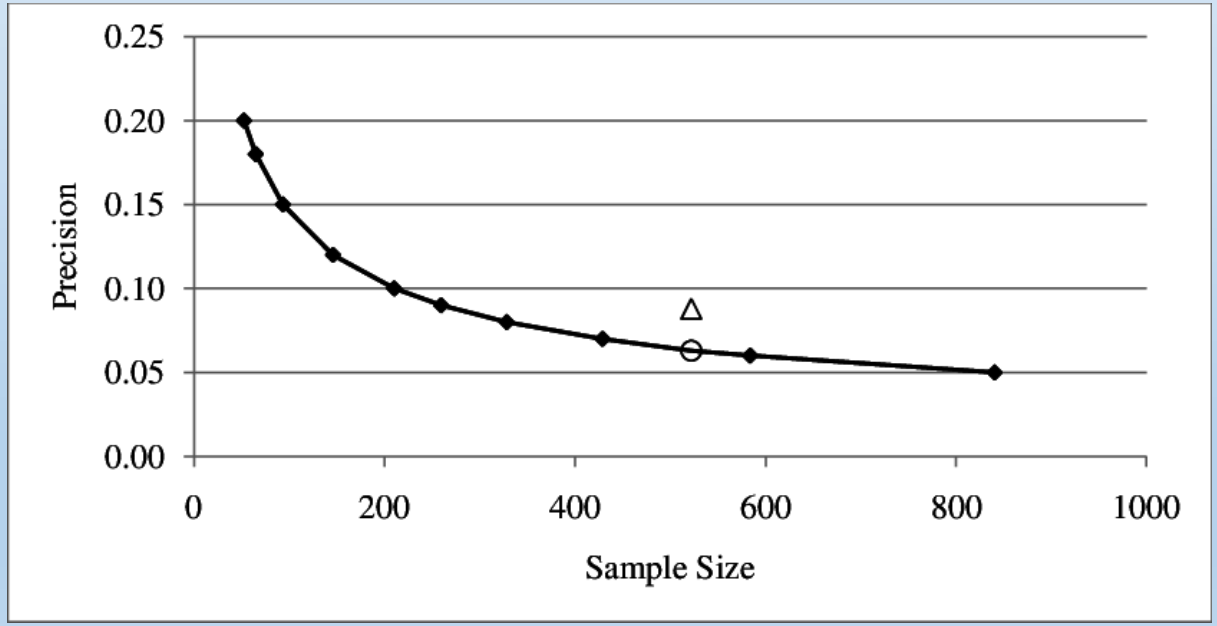
An epidemiological study is planned to estimate the prevalence of subjects with a certain disease and we want the confidence interval at level $1-\alpha = 0.95$ not to exceed the maximum error ($E^* = 0.01$, i.e. a global amplitude of 0.02) we will need a minimum of:

$$n^* \geq 0.25 * \left(\frac{1.96}{0.01}\right)^2 = 9604$$

at least 9604 subjects from the target population...

Due strategie di base per la dimensione campionaria

Precisione (intervalli di confidenza)



Potenza del test di ipotesi (**effect size**)

		True state of H_0 (Unknown)	
		H_0 true	H_0 false
Decision (sample data)	Reject H_0	Type I error*	ok
	Do Not reject H_0	ok	**Type II error****

Ronald Fisher
(1890 – 1962)

Jersey Neyman
(1894 – 1981)

Egon Pearson
(1895 – 1980)

Neyman e Pearson lavorarono sull'aspetto della **potenza** del test e sulla dimensione campionaria minima richiesta

Fisher introdusse il concetto di test di ipotesi focalizzandosi principalmente sull'errore di primo tipo

		True state of H_0 (Unknown)	
		H_0 true	H_0 false
Decision (sample data)	Reject H_0	**Type I error***	ok
	Do Not reject H_0	ok	**Type II error****

RCTs: perché calcolare la dimensione campionaria a priori?

- Uno studio **sotto-dimensionato** non identificherà l'effetto di interesse;
- Uno studio **sovra-dimensionato** è uno spreco di risorse;
  - **Non è etico** arruolare più pazienti/cavie del minimo necessario a stabilire se esista un effetto significativo;
  - **Non è etico** esporre più pazienti/cavie del necessario agli interventi sperimentali o al placebo, diminuendo così la loro probabilità di essere arruolati in altri studi.

Primo passo:

Specificare l'*outcome* primario:

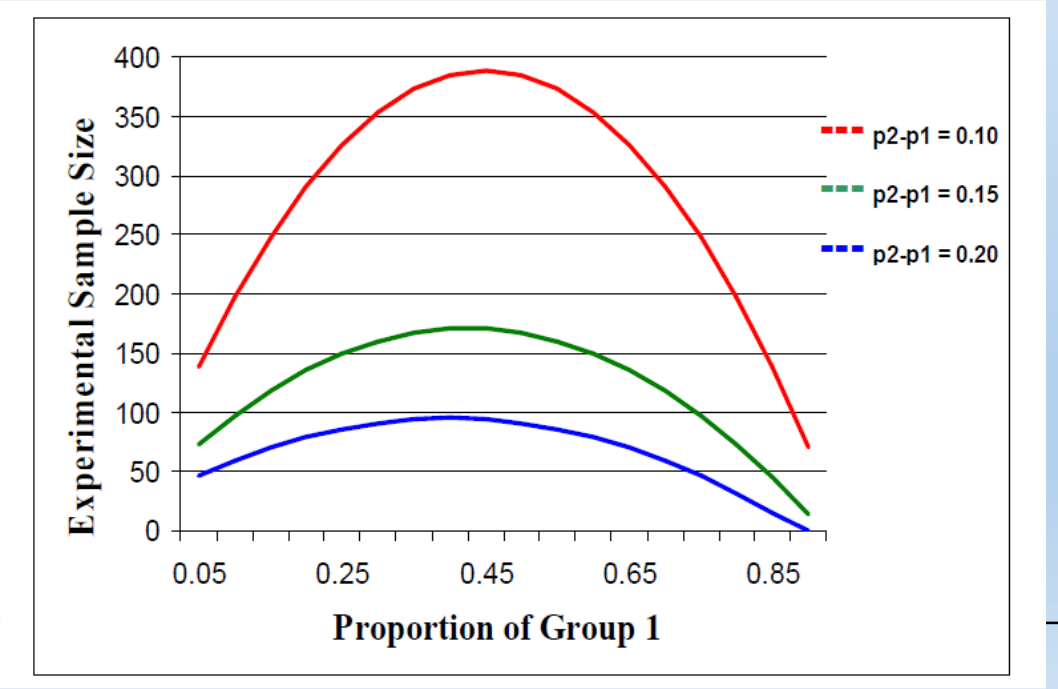
- Che **tipo** di variabile è/quale **scala di misura** ? (continua, conteggio, categorica, binaria, tempo ad evento…)
- Se sono presenti **outcomes secondari**, la dimensione campionaria minima necessaria deve essere calcolata anche per questi, prendendo poi la maggiore.

Dimensione campionaria in caso di test di ipotesi

- **Tipo** di outcome primario;
- **Dimensione** dell'effetto di interesse;
- «*Guess-estimate*» della **variabilità** dell'outcome
- Numero di pazienti/cavie **massimo** disponibile (se esiste) «eligibile» o «compliante»
- **Tempo** necessario a completare lo studio.

Il calcolo della dimensione campionaria è un **SET** di calcoli…possibilmente presentato sotto forma di tabella o grafico:

Esempio di calcolo di sample size per confrontare due proporzioni →



Quanti ***dropouts*** sono attesi?
(aggiustare nel calcolo per questo dato)

Parametri necessari per il calcolo della dimensione campionaria

- **Livello di significatività alpha** (α):
  probabilità di concludere che c'è un effetto significativo quando non c'è (5%);
- **Potenza** ($1-\beta$):
  probabilità di non mancare un effetto significativo, quando c'è (80%);
- **Differenza/effetto clinico/effect size: la differenza/effetto che si ritiene sia rilevante…**

- Stima della **dimensione dell'effetto atteso**:

[se non è disponibile da studi precedenti/letteratura si può effettuare uno **studio pilota** per determinarla]

$$ES = \frac{\mu_1 - \mu_2}{\sigma}$$

Sullivan GM, Feinn R, "Using Effect Size—or Why the *P* Value Is Not Enough", J Grad Med Educ. 2012 Sep; 4(3): 279–282.

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]

Che cosa è l' «effect size»?

Ampiezza della «differenza» tra i gruppi

(a) Differenza assoluta:

$$ES = \mu_1 - \mu_2$$

(b) Differenza «standardizzata»:

$$ES = \frac{\mu_1 - \mu_2}{\sigma}$$

(a) i parametri hanno un significato numerico chiaro: pressione sistolica «media», numero di eventi di ospedalizzazione…

(b) i parametri non hanno una interpretazione numerica diretta: punteggi su una scala; misure su scale differenti; rilevante variabilità …

Perché riportare una misura di effect size ?

La significatività statistica (=il *p value*) afferma che un effetto è verosimile ma non dice nulla circa la sua **dimensione**; la significatività «sostanziale/clinica» (=*effect size*) deve essere riportata come risultato principale dello studio.

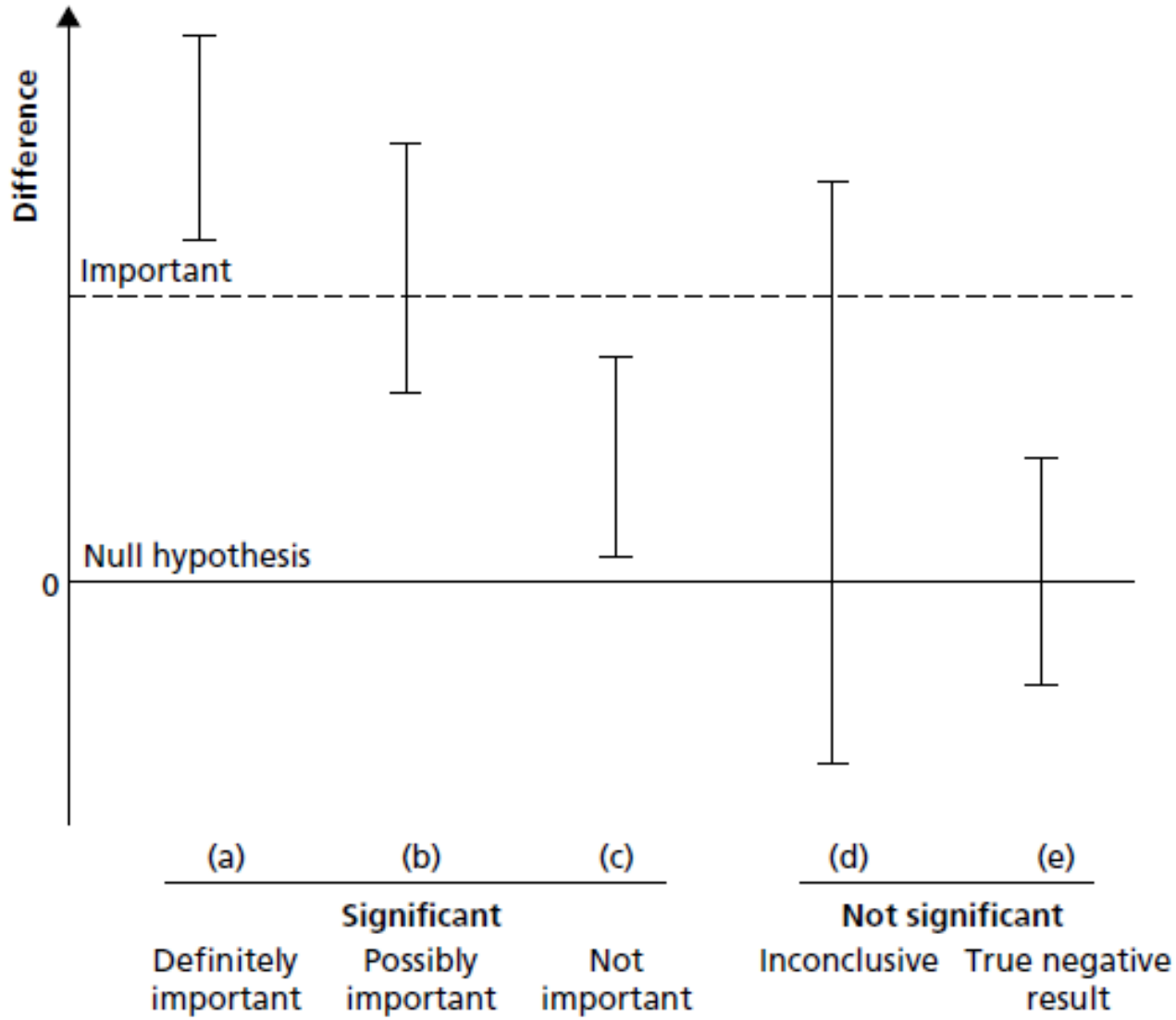
Una stima dell' «*effect size*» deve essere fatta **prima dell'inizio dello studio** per determinare la dimensione campionaria, assumendo come «costanti» le probabilità di commettere errori nei test di ipotesi…

Perché il p value **non è** sufficiente?

La significatività statistica (=p value) corrisponde alla probabilità che la differenza che abbiamo ottenuto tra i gruppi sia dovuta al caso.

Se il p value è > 5% per convenzione si decide che la differenza osservata è spiegata dalla variabilità casuale dello studio campionario, ma non riflette una reale differenza «biologica».

Problema: in studi campionari di dimensioni «sufficientemente grandi» il test statistico produrrà «sempre» un p value < 5%... anche per differenze osservate (effect size) **irrilevanti**..
Analogamente, in studi su piccoli campioni, il p value > 5% anche per differenze **rilevanti**....



(a) La differenza è significativa e clinicamente rilevante

(b) La differenza è significativa ma non è chiaro se sia clinicamente rilevante

(c) La differenza è significativa ma non clinicamente rilevante

(d) La differenza non è significativa ma può essere clinicamente rilevante

(e) La differenza non è significativa e non è clinicamente rilevante

L'obiettivo quando si pianifica uno studio dovrebbe essere di «garantirci» che, **se una differenza clinicamente rilevante esiste**, allora riusciremo ad identificarla tramite il test statistico (-> dimensione campionaria).

Come definiamo e calcoliamo l'effect size ?

1. **<u>Differenze tra gruppi</u>**: in base alla scala di misura dell'outcome (numerica/binaria)

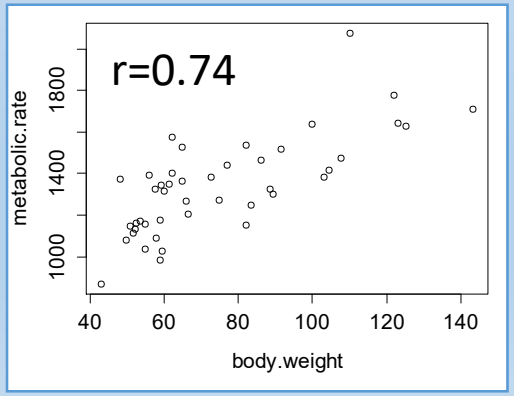
Index	Description[b]	Effect Size	Comments
Between groups			
Cohen's *d*[a]	$d = M_1 - M_2 / s$ $M_1 - M_2$ is the difference between the group means (*M*); *s* is the standard deviation of either group	Small 0.2 Medium 0.5 Large 0.8 Very large 1.3	Can be used at planning stage to find the sample size required for sufficient power for your study
Odds ratio (OR) **OR=(a/c):(b/d)**			For binary outcome variables Compares odds of outcome occurring from one intervention vs another
Relative risk or risk ratio (RR) **RR=(a/a+b):(c/c+d)**			Compares probabilities of outcome occurring from one intervention to another

	Malati	Sani
Esposti	**a**	**b**
Non Esposti	**c**	**d**

Come definiamo e calcoliamo l'effect size ?

2. **Associazioni fra variabili**: in base alla scala di misura dell'outcome (correlazione/regressione)

Index	Description[b]	Effect Size	Comments
Measures of association			
Pearson's *r* correlation	Range, −1 to 1	Small ±0.2 Medium ±0.5 Large ± 0.8	Measures the degree of linear relationship between two quantitative variables
r^2 coefficient of determination	Range, 0 to 1; Usually expressed as percent	Small 0.04 Medium 0.25 Large 0.64	Proportion of variance in one variable explained by the other



r^2=0.54
54% variabilità spiegata del tasso metabolico dal peso corporeo

- $0 \leq r \leq 0.25$: poca o nessuna associazione lineare;
- $0.25 < r \leq 0.50$: discreto grado di associazione lineare;
- $0.50 < r \leq 0.75$: grado di associazione lineare tra moderato e buono;
- $r > 0.75$: grado di associazione lineare tra molto buono ed eccellente.

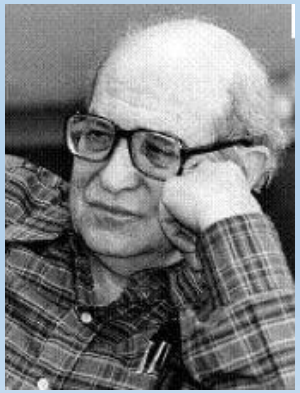
Torniamo quindi alla lista:

- **Livello di significatività alpha** (α):
probabilità di concludere che c'è un effetto significativo quando non c'è (5%);

- **Potenza** (1-β):
probabilità di non «mancare» un effetto significativo quando c'è (80%);

- ***Differenza/effetto clinico/effect size: la differenza/effetto che si ritiene sia rilevante.***

*La soglia convenzionale per β (=20%) fu proposta da Cohen: affermò che poiché un errore di primo tipo (falso positivo) era *più rilevante* di un errore di secondo tipo (falso negativo) si poteva tollerare che accadesse con una probabilità 4 volte maggiore.



1923-1998

Potenza di un test: l'ingrediente nascosto…

Per una soglia α prefissata, e per una (tipicamente non nota e non modificabile) σ (variabilità del fenomeno) la potenza del test risponde a queste domande:

1. Data una dimensione campionaria ***n*** ed una «differenza» **(ES)** tra trattamenti $\Delta \neq 0$, quale è la potenza **(1-β)** di identificare questa differenza, i.e. di concludere *in favore* di H_1?

2. Data una certa potenza **(1-β)** ed una «differenza» **(ES)** $\Delta \neq 0$, quale dimensione campionaria ***n*** è necessaria per identificare la differenza (ossia supportare H_1)?

3. Data una certa dimensione campionaria ***n*** e una potenza **(1-β),** quale è la ***minima differenza*** Δ **(ES)** che può essere identificata con una probabilità 1-β?

POTENZA= P(rigettare H_0 | *Vero* effetto>= *Effect Size*)

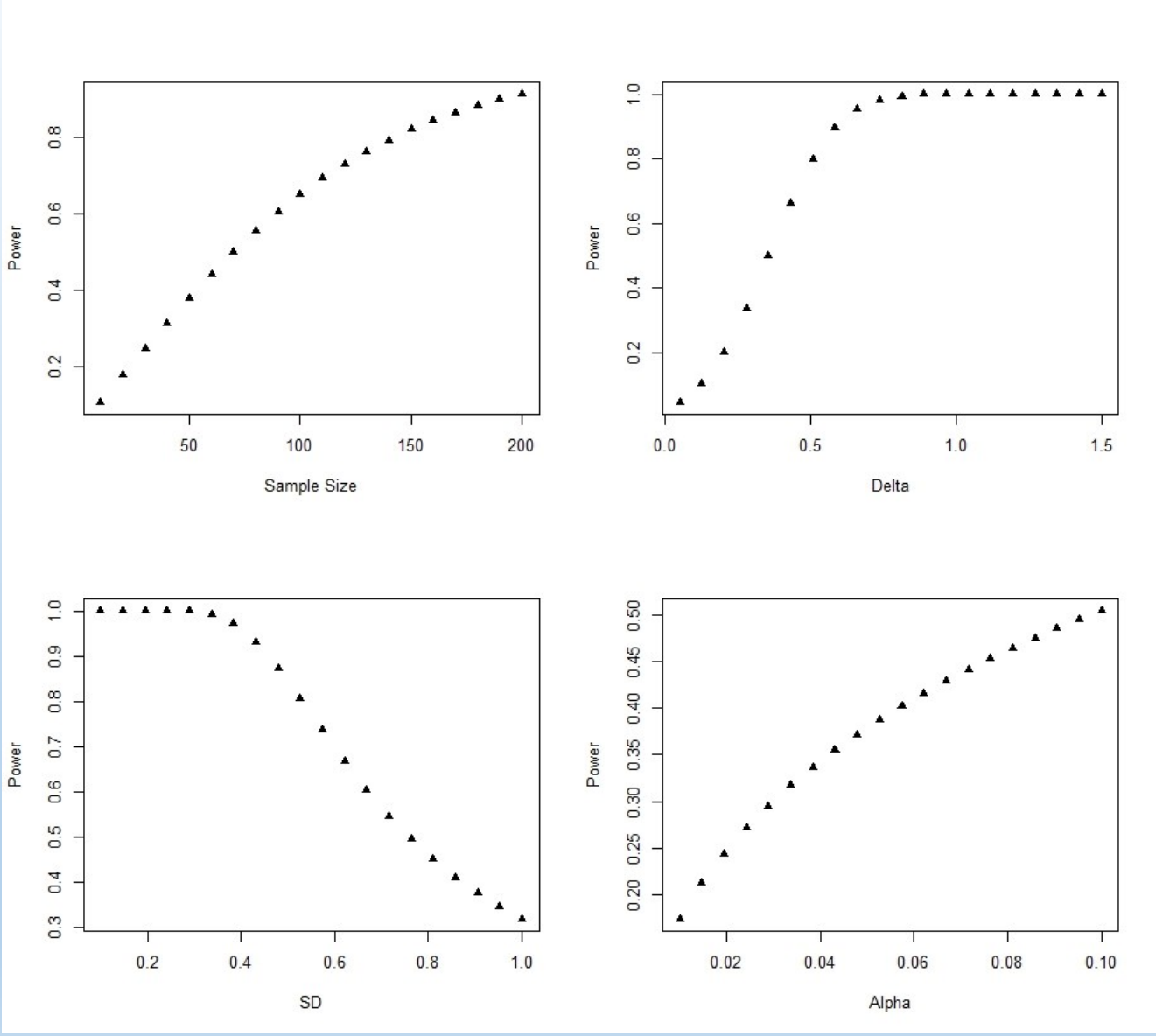
Per calcolare la potenza occorre però avere una «stima» di σ e della differenza Δ (ES) oggetto di studio…

Fisher's approach	**Neyman-Pearson's approach**
Outcome: significant / non-significant	Outcome: accept / reject
p is a measure of evidence against H0	*p* is NOT a measure of evidence and should not be interpreted
An alternative hypothesis **cannot** be specified	An alternative hypothesis **must** be specified
Does not have a concept of "power"	Power has to be specified prior to the experiment
A single rejection of H0 is the start, not the end, of an investigation. Replication needed and meta-analyses are useful	A single rejection is meaningless – the framework only guarantees long-term type-1 and type-2 error rates but does not allow to make inference about a single case.

Esempio:
Consideriamo il t-test per due campioni
H0: $\mu_1=\mu_2$ vs H1: $\mu_1\neq\mu_2$
con $\sigma_1=\sigma_2=\sigma$

La **potenza** del test **è una funzione** di:
$\Delta=\mu_1-\mu_2(/\sigma)$ (ES), di σ, di **n** e di α

Potenza= 1- $\beta(\Delta,\sigma,n,\alpha)$



Potenza di un test

Esempio (data la dimensione campionaria):

«*Gold Standard*» vs «*Nuovo*» trattamento: l'outcome si assume distribuito come una gaussiana in entrambi i gruppi con medie μ_1 e μ_2 e deviazioni standard $\sigma_1=\sigma_2=\sigma$

Il nuovo trattamento ed il Gold standard vengono considerati **clinicamente diversi** se la differenza $\Delta=|\mu_1-\mu_2|$ è almeno di 0.3 unità

Ci sono $n_1=n_2=n=50$ pazienti/cavie eleggibili per ogni braccio dello studio.

Il livello di significatività è $\alpha=5\%$ e precedenti studi giustificano l'assunzione di $\sigma\leq 0.9$ unità.

Quale è la potenza ?

Potenza di un test

Se n=50 e $|\Delta|=0.3$ (con $\sigma=0.9$ e $\alpha=5\%$) quale è la potenza ?

```
power.t.test(n = 50, delta = 0.3, sd = 0.9,sig.level = 0.05, type = "two.sample",alternative =
"two.sided")

     Two-sample t test power calculation

              n = 50
          delta = 0.3
             sd = 0.9
      sig.level = 0.05
!!!!!!!!!!!!!!!!!! power = 0.3784221 !!!!!!!!!!!!!!!!!!
    alternative = two.sided

NOTE: n is number in *each* group
```

Abbiamo una potenza del 38% : probabilità di non rigettare H_0 pur essendo falsa … **62% !!!**

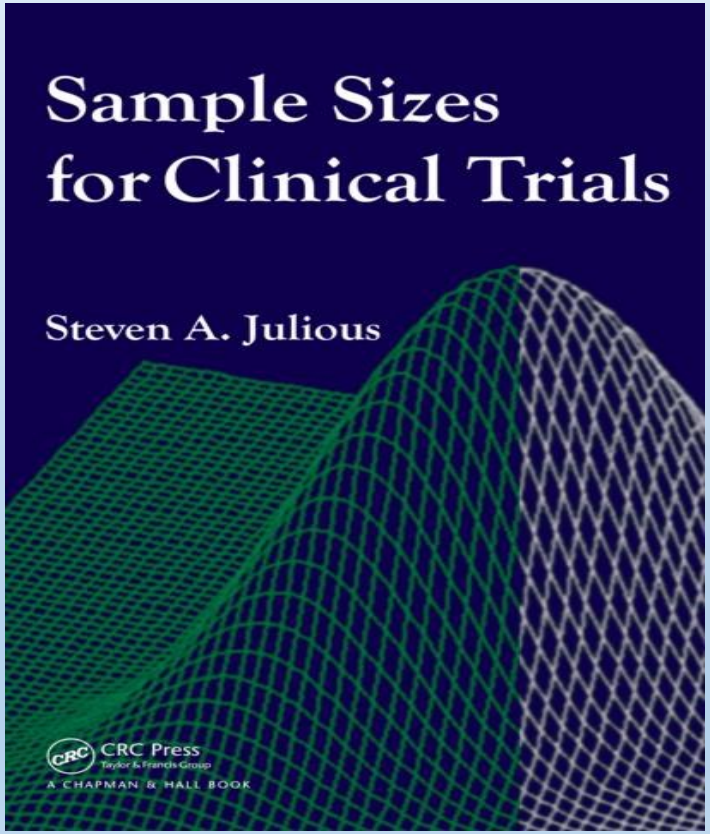
Potenza di un test

Se 1-β=80% e n=50 (con σ=0.9 e α=5%) quale è la più 'piccola' differenza Δ identificabile?

```
power.t.test(n = 50, power=0.8, sd = 0.9,sig.level = 0.05, type = "two.sample",alternative = "two.sided")

     Two-sample t test power calculation

              n = 50
          delta = 0.51
             sd = 0.9
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

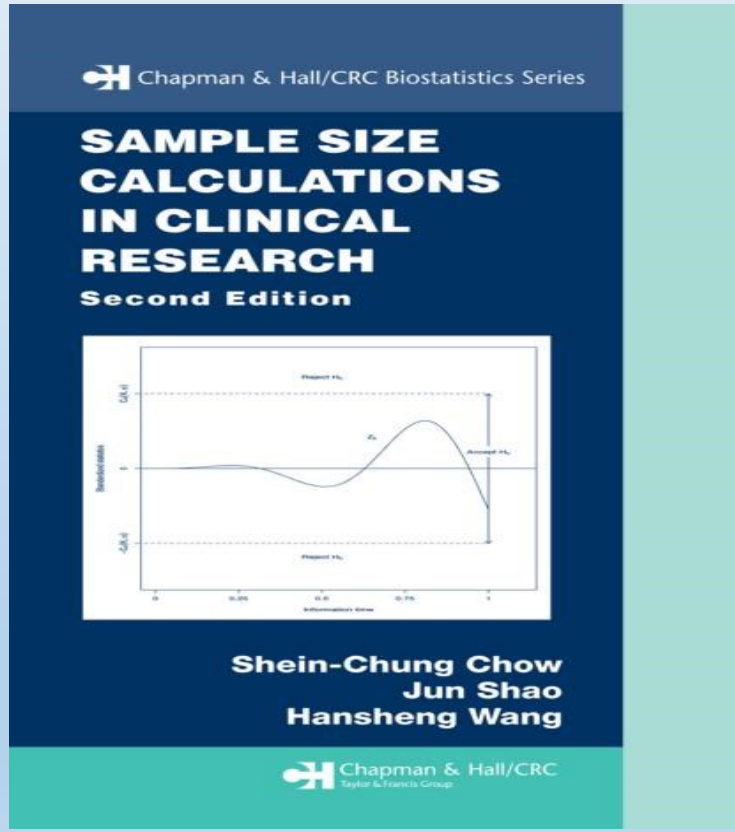
La differenza minima identificabile (stat. significativa) tra i due trattamenti è di **0.51** unità data la dimensione campionaria disponibile e la misura di variabilità dell'outcome.

Esempi (semplici) di calcolo della dimensione campionaria



RCT:

- Differenza tra due medie
- Differenza tra due proporzioni



DIFFERENZA TRA DUE MEDIE

RCT per confrontare due programmi dietetici:

Outcome primario= variazione di peso dopo 6 mesi di trattamento (campioni indipendenti);

Effetto clinico (effect size)= 5 kg = media peso gruppo 1 – media peso gruppo 2

Variabilità attesa del peso= 11 kg

α=5%; **1-**β=80%

77 pazienti per gruppo

```
library(pwr)
d <- 5/11

pwr.t.test(d=5/11,power=0.8,sig.level=0.05,type="two.sample",alternative="two.sided")
```

```
     Two-sample t test power calculation

              n = 76.94921
              d = 0.4545455
      sig.level = 0.05
          power = 0.8
    NOTE: n is number in *each* group
```

Confronto tra proporzioni in due gruppi

RCT per valutare l'efficacia comparativa di due trattamenti per HIV;

Outcome primario= % di pazienti con carica virale (VL) inferiore al limite di soglia a 48 settimane;

Soglia standard= ci si aspetta che il 60% (p1) dei pazienti nel trattamento standard ottenga la soppressione di VL;

Effetto atteso clinico= (p2-p1)= 20%

-> **81** pts per braccio

```
p1=0.6
p2=0.8

h.calc = 2*asin(sqrt(p1))-2*asin(sqrt(p2))

pwr.2p.test(h=h.calc , n =NULL, sig.level =0.05,
power =0.80)
```

```
Difference of proportion power calculation for
binomial distribution (arcsine transformation)

              h = 0.44
              n = 80.29911
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: same sample sizes
```

	Outcome Variable	
Predictor Variable	**Dichotomous**	**Continuous**
Dichotomous	Chi-squared test†	*t* test
Continuous	*t* test	Correlation coefficient

† The chi-squared test is always two-sided; a one-sided equivalent is the Z statistic.

Per **ogni test** c'è una formula di calcolo della sample size …

1. ***t* test:** differenza tra due medie (tra variabili distribuite in modo gaussiano…)…

2. **chi-squared test** (χ_2): associazioni tra variabili qualitative…

3. **z test:** confronto tra proporzioni…

4. **Correlation coefficient:** intensità della associazione lineare tra due variabili su scala continua…

Usi e Abusi dei Test di Ipotesi

Abuso: utilizzare il test di ipotesi come **strumento descrittivo** in studi osservazionali

NON E' la ragione per cui è stato sviluppato questo strumento di statistica inferenziale!!!

	Insomnia (n = 20)	Non-insomnia (n = 20)	$\chi 2$	p value
	n	*n*		
Male (%)	7 (35.0%)	10 (50.0%)	0.672	.180
Feale (%)	13 (65.0%)	10 (50.0%)	0.391	.532
Depression (%)	18 (90.0%)	14 (70.0%)	0.985	**< .001**
	mean (SD)	*mean (SD)*	*t*	*p* value
Age (year)	56.7 (17.1)	53.9 (21.5)	0.467	.643
BMI (Kg/m^2)	22.7 (3.8)	21.6 (4.7)	0.879	.385
PSQI	13.3 (3.5)	6.3 (1.8)	8.193	**< .001**
HDRS	19.1 (6.8)	10.2 (6.2)	1.460	.153
HAMA	18.5 (6.7)	10.4 (7.4)	0.871	.389

Significant differences are presented in bold.

Per confrontare dal punto di vista descrittivo dei gruppi si può utilizzare l'approccio della SMD: Standardized Mean/proportion Difference

Table 1. Baseline Characteristics of Patients Initiating Oral Anticoagulants, Overall and Stratified by Initial Treatment Group

	SMD >0.1	Overall (100%) 7545	DOAC (45.75%) 3452	VKA (54.25%) 4093	SMD
Gender (%)					0.07
Male		3905 (51.76%)	1721 (49.86%)	2184 (53.36%)	
Female		3640 (48.24%)	1731 (50.14%)	1909 (46.64%)	
Year of enrolment (%)	*				1.162
2013-2015		2299 (30.47%)	351 (10.17%)	1948 (47.59%)	
2016-2018		2794 (37.03%)	1190 (34.47%)	1604 (39.19%)	
2019-2021		2452 (32.5%)	1911 (55.36%)	541 (13.22%)	
Province of residence (%)	*				0.269
Gorizia		2553 (33.84%)	1405 (40.7%)	1148 (28.05%)	
Trieste		4992 (66.16%)	2047 (59.3%)	2945 (71.95%)	
Age, y (median [Q1-Q3])		78 [72-84]	79 [72-84]	77 [72-83]	0.075
Averaged GFR before baseline (median [Q1-Q3])	*	68.5 [54.3-81.2]	69.5 [56.0-81.7]	67.5 [52.9-81.0]	0.121
GFR class (%)	*				0.2
15-29		220 (2.91%)	44 (1.27%)	176 (4.30%)	
30-45		818 (10.84%)	350 (10.14%)	468 (11.43%)	
46-60		1577 (20.9%)	717 (20.77%)	860 (21.01%)	
61-90		4232 (56.09%)	2014 (58.34%)	2218 (54.19%)	
>90		698 (9.25%)	327 (9.47%)	371 (9.06%)	

$$d = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2 + s_2^2}{2}}}$$

$$d = \frac{(\bar{p}_1 - \bar{p}_2)}{\sqrt{\frac{\bar{p}_1(1-\bar{p}_1) + \bar{p}_2(1-\bar{p}_2)}{2}}}$$

(per variabili continue non simmetriche, confrontare i quartili… cercare una trasformazione della scala…)

Cenni all'approccio bayesiano nel test di ipotesi:

Is there an effect of group on performance?

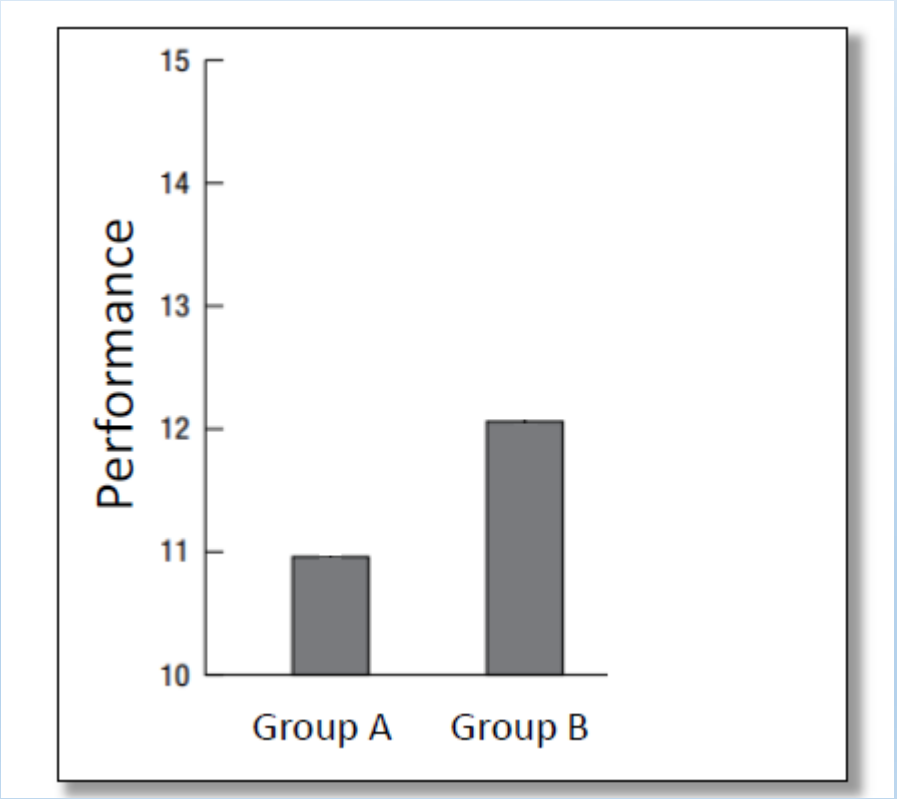
H_0: There is no effect of group on performance
H_1: There is an effect of group on performance

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 (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 "Sudden Infant Death Syndrome" (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.

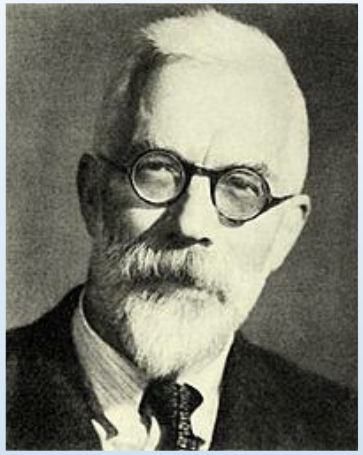
How did it end for Clark?

- 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

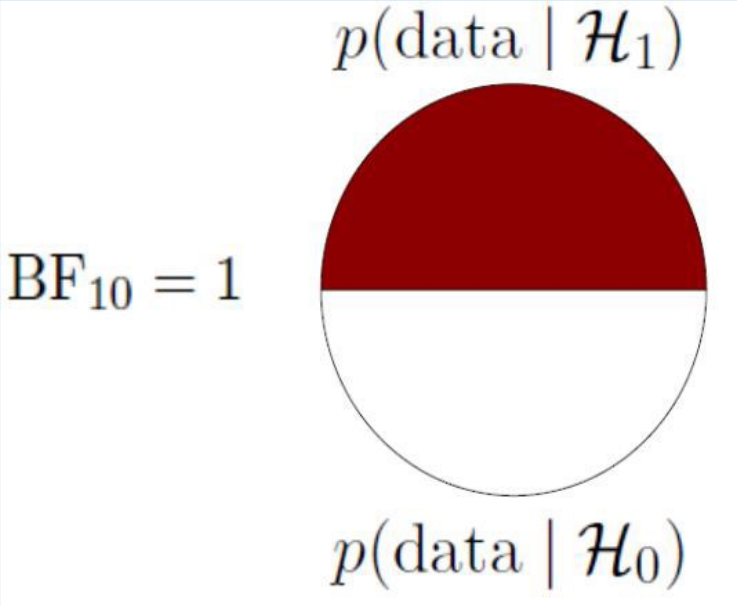
Probability of Hypothesis 1, given the data

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

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

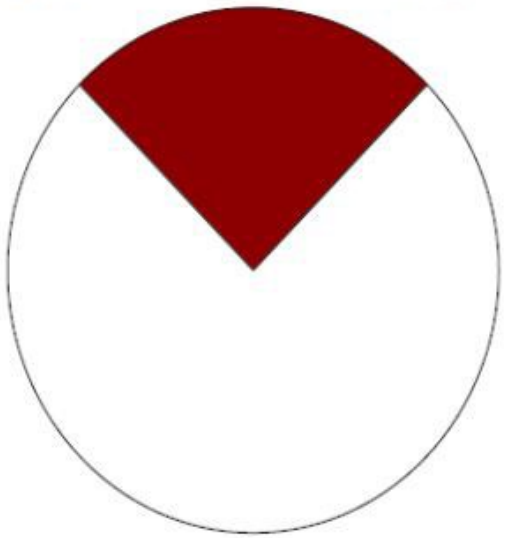
Prior Odds

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

Posterior Odds

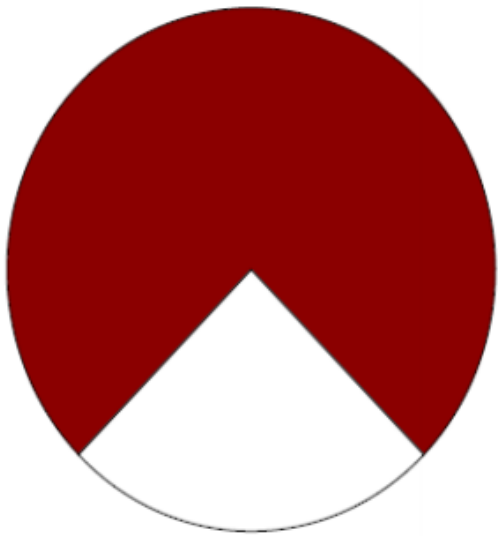
$$\mathrm{BF}_{10} = \frac{1}{3}$$

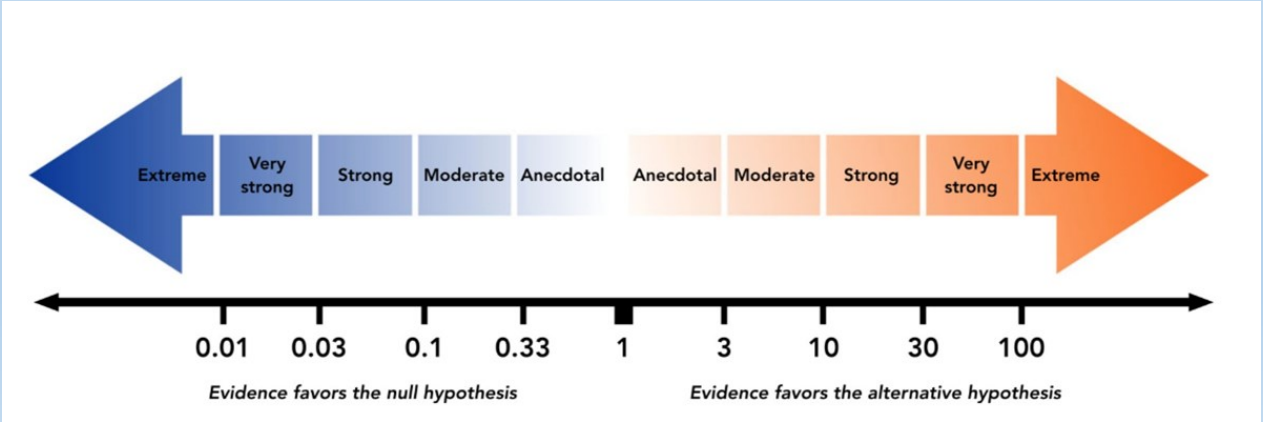
$$\mathrm{BF}_{01} = 3$$



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

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





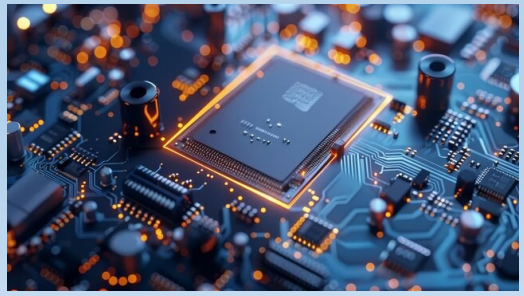
Bayes factor

- 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