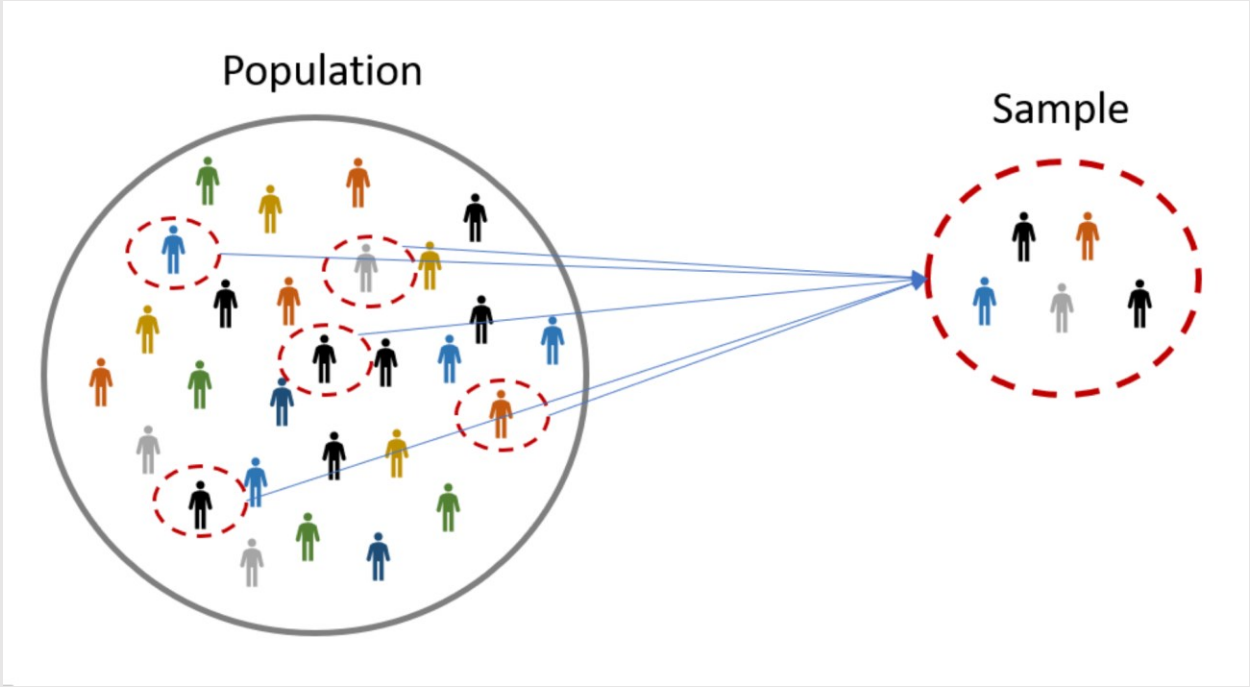
Sample Size [basics] II

- **Effect size** approach for Sample Size
- *Difference* between means
- *Difference* between proportions
- SS for OR/RR
- Concluding remarks

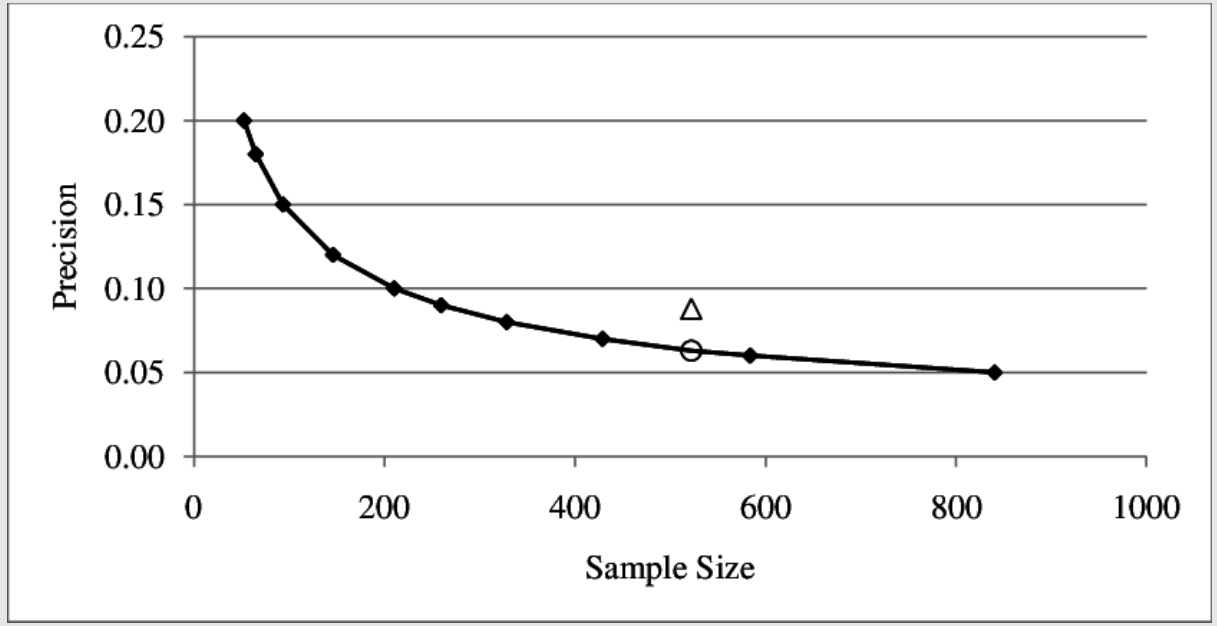


$$n = \frac{\$ \text{ available}}{\$ \text{ per sample}}$$

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

Two strategies for sample size

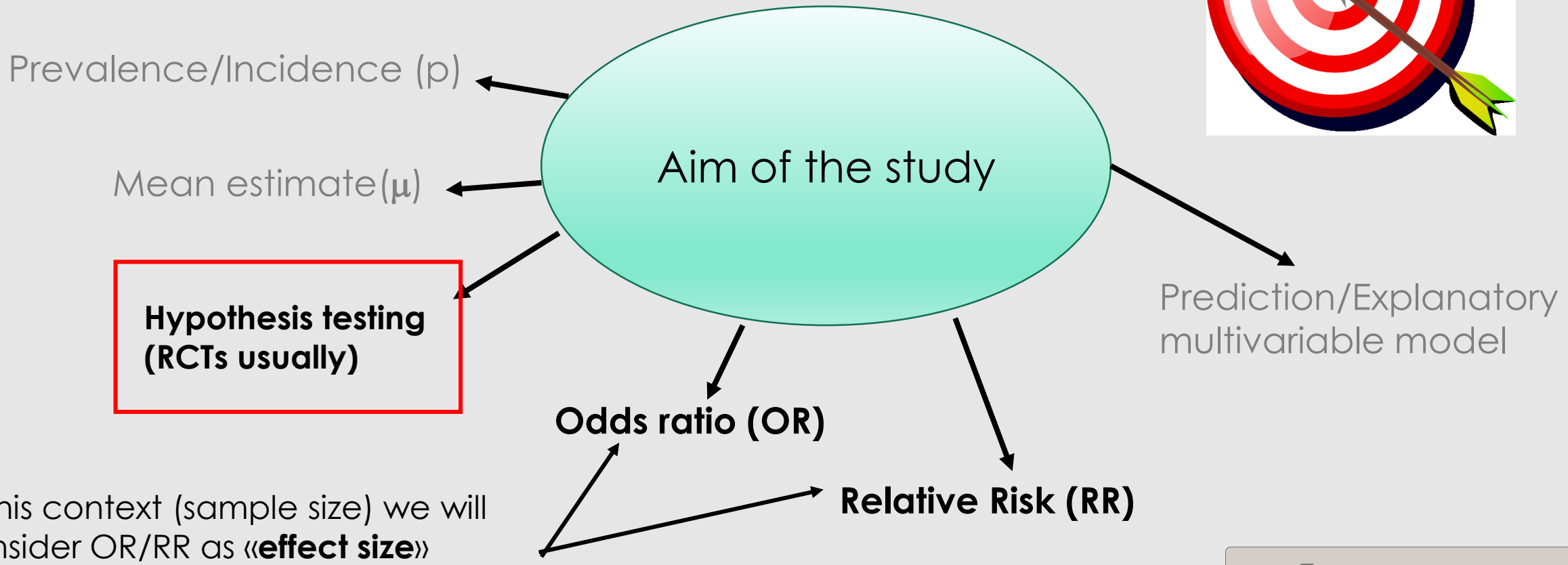
Precision (confidence intervals)



Power of the statistical test (**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**

Sample size calculations **depend on the primary objective** of the study design:



In this context (sample size) we will consider OR/RR as «**effect size**» (ingredient of the formula !!)

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

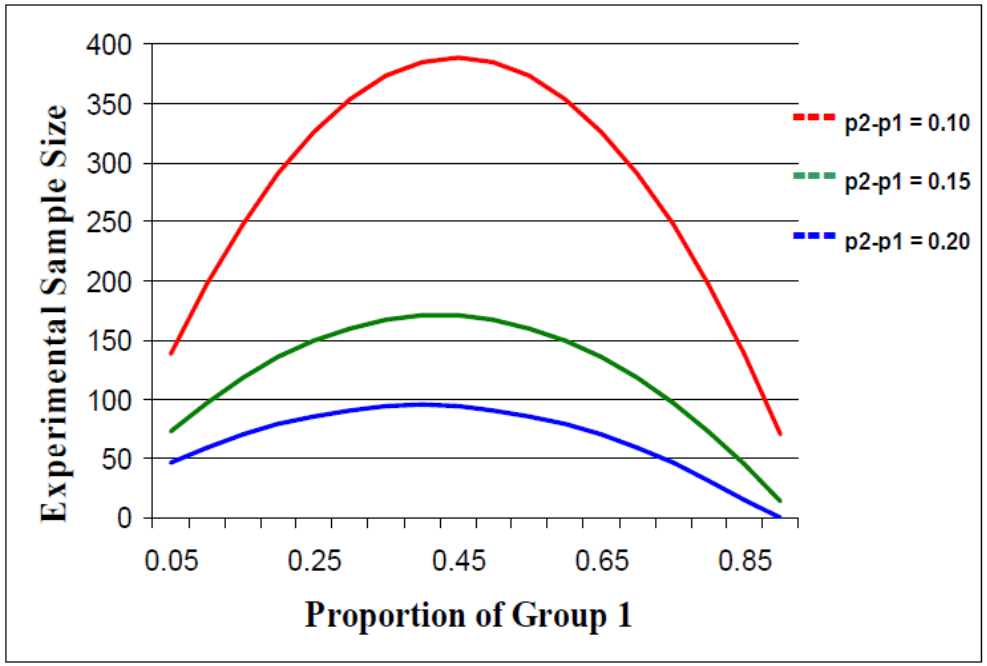
Checklist

- **Type** (scale of measure) of primary outcome
- **Size** of the effect of interest
- *Guess-estimate* of the **variability** of the outcome

- *Maximum number of patients available (if any) "eligible" or "compliant"*
- *Time needed to complete the study*

Calculation of the sample size based on the effect size is a **SET** of calculations ... possibly presented in the form of a table or graph:

Example of sample size calculation to compare two proportions →



How many ***dropouts*** are expected?
(adjust in the calculation for this issue)

Parameters required in input

- **Alpha (α)** significance level:

probability of concluding that there is a significant effect when there is not (5%)

- **Power (1-β)** :

probability of not missing a significant effect, when there is (80%)

- Difference / clinical effect / **effect size**: the difference / effect *believed* to be relevant ...

Effect size:
[if not available from previous studies / literature a pilot study can be carried out to determine it]

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

Sullivan GM, Feinn R, "Using Effect Size - or Why the *P* Value Is Not Enough", J Grad Med Educ. 2012

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 *effect size*?

Amplitude of the **difference** between groups

(a) *Absolute* difference:

$$ES = \mu_1 - \mu_2$$

(b) *Standardized* difference:

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

(a) parameters have a clear numerical meaning: average systolic pressure, number of hospitalization events ...

(b) parameters do not have a direct numerical interpretation:
scores on a scale; measurements on different scales [or they show *significant* variability]

Why report an effect size measurement?

Statistical significance (**p value**) states that an effect *probably exists* but says nothing about its **size**; the *substantial* significance (effect size) must be reported as the main result of the study.

An estimate of the effect size must be made **before** the start of the study to determine the **minimum sample size** required, assuming as *constant* the probabilities of making mistakes in the hypothesis tests ...

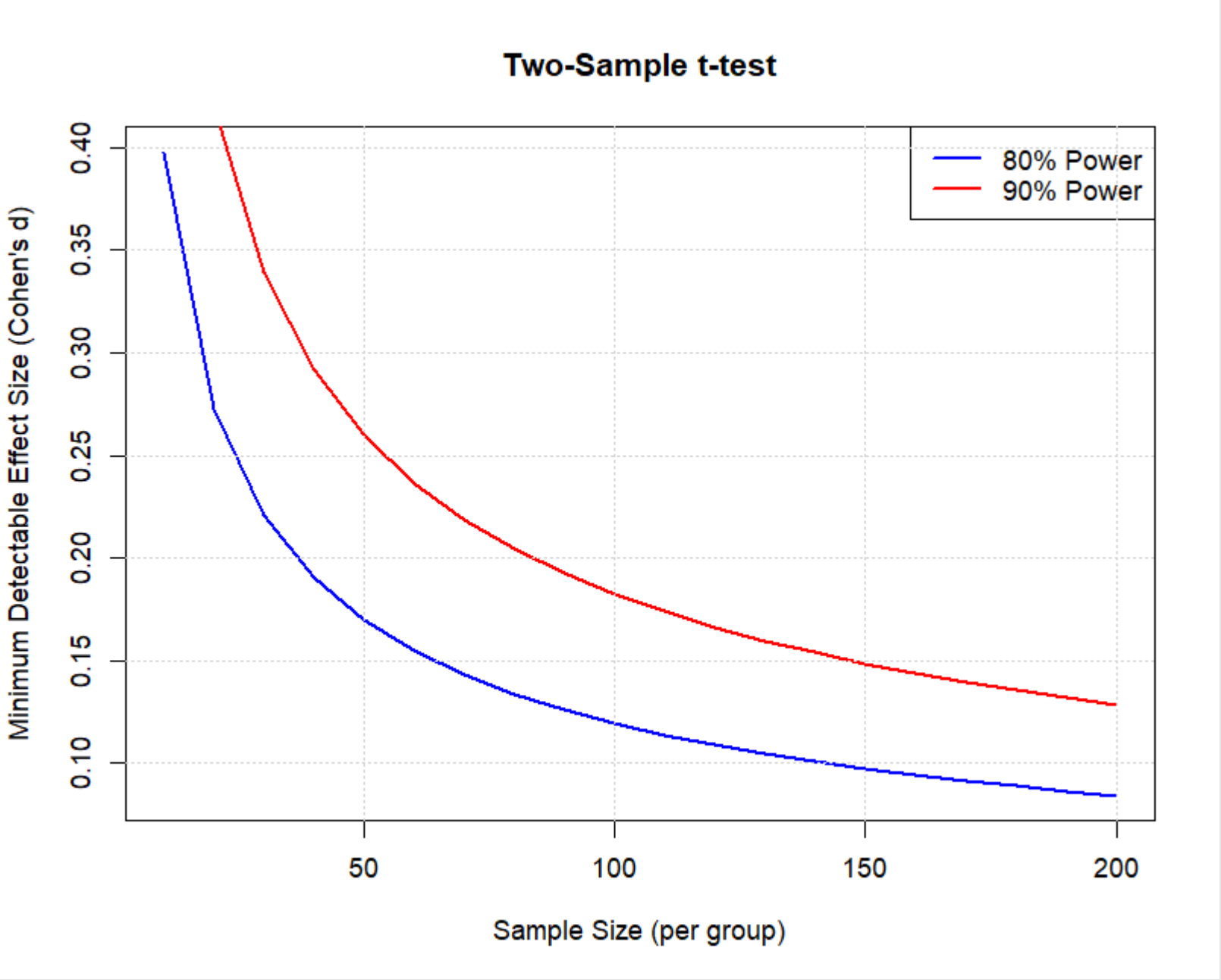
Again: why isn't p value enough??

Statistical significance (p value) corresponds to the probability that the difference between the groups *is due to chance*

If p value > 5%, it is decided that the observed difference is explained by *random variability* of the sample study, but does not reflect a *true* difference

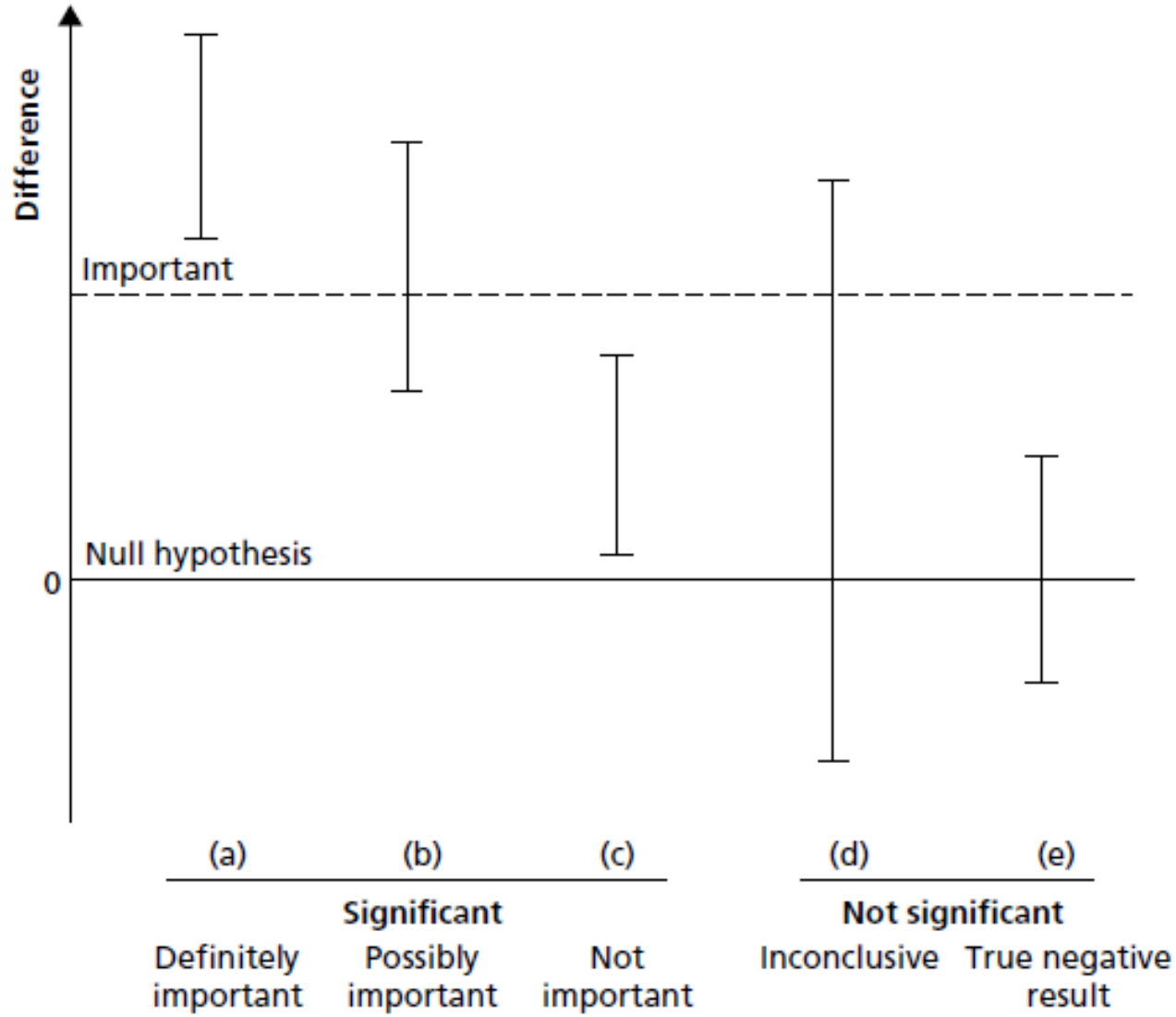
Problem: in ***sufficiently large*** samples the statistical test **will always** produce a p value < 5% ... even for **irrelevant** observed differences (= negligible effect size)..

Similarly, in ***small*** sample studies, the p value > 5% also for relevant differences….



In ***sufficiently large*** samples the statistical test **will always** produce a p value < 5% ... even for **irrelevant** observed differences (= negligible effect size)..

Similarly, in ***small*** **sample** studies, the p value > 5% also for **relevant** differences….



(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).

How do we define and calculate the effect size?

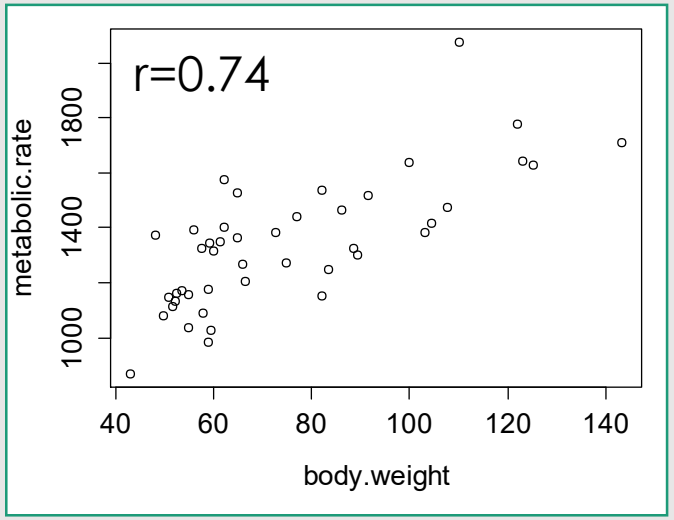
1. **Differences/Ratios** between groups: based on the outcome measurement scale (numeric / binary)

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)	$\frac{\text{Group 1 odds of outcome}}{\text{Group 2 odds of outcome}}$ If OR = 1, the odds of outcome are equally likely in both groups	Small 1.5 Medium 2 Large 3	For binary outcome variables Compares odds of outcome occurring from one intervention vs another
Relative risk or risk ratio (RR)	Ratio of probability of outcome in group 1 vs group 2; If RR = 1, the outcome is equally probable in both groups	Small 2 Medium 3 Large 4	Compares probabilities of outcome occurring from one intervention to another

How do we define and calculate the effect size?

2. **Associations** [continuous variables]: correlation/linear regression*

Index	Description[b]	Effect Size	Comments
Measures of association			
Pearson's *r* correlation (linear 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% variability of the metabolic rate explained by body weight

$$0 \leq r \leq 0.25 \; \boldsymbol{low} \; correlation$$

$$0.25 < r \leq 0.50 \; \boldsymbol{medium} \; level \; of \; correlation$$

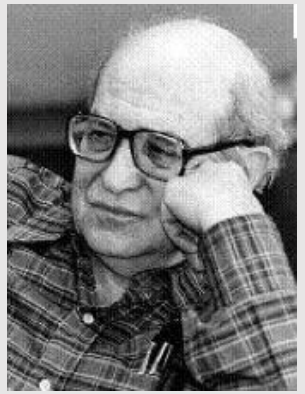
$$0.50 < r \leq 0.75 \; \boldsymbol{good} \; correlation$$

$$r > 0.75 \; \boldsymbol{very \; good} \; correlation$$

Again (!) basic **ingredients** to determine the sample size:

- **Alpha** (α) significance level, probability of concluding that there is a significant effect when there is not (5%)
- **Power** (1-β), probability of not "missing" a significant effect, when there is (80%)
- Difference/clinical effect/**effect size**: effect believed to be relevant

Threshold for β (= 20%) was proposed by Cohen, who stated that since a first type error (false positive) ***was more relevant**** than a second type one (false negative) it could be tolerated that it happened with a 4 times greater probability



1923-1998

*It depends on the context !!

And now a quick reminder…

Hypothesis test*: basic concepts

- There is a hypothesis on a certain phenomenon in the population to be tested (**null** hypothesis vs **alternative** hypothesis)
- We collect data relevant to the problem (**sample** data)
- The pieces of information are combined to obtain a measure of **evidence** in favor of against the null hypothesis
- It is decided whether there is **enough evidence** from the data to accept or reject the null hypothesis

*__frequentist__ point of view

Hypothesis test: an analogy attempt

A person is accused of a crime: he/she is arrested and brought in a court

Null hypothesis: Presumption of innocence
Alternative hypothesis: the suspect is guilty
Information (evidence = **data**) is collected on the matter

The judge evaluates the evidence collected

The judge decides whether to blame the suspect or not

The basic principle:
Not enough evidence -> Not guilty verdict (*in dubio pro reo*)

Unfortunately: it can happen that an innocent goes
to jail,
just as a guilty is left free ...

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di
Scienze Mediche Chirurgiche e della Salute

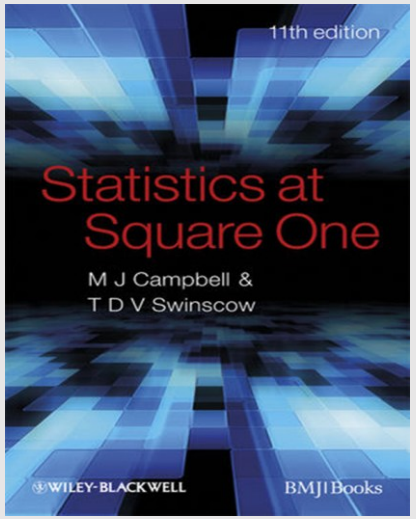
Type I and II errors

Sample study → Probability of *errors*

		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****

***Type I error:**
Reject H_0 when it is actually true (innocent in jail)

A probability is associated with this error: **level of significance α** is ***under control***, because *the test is designed in such a way that α is not larger* than a pre-specified threshold.



CHAPTER 6

P-values, power, type I and type II errors

****Type II error:**
Do not reject H_0 when it is actually false (free guilty)

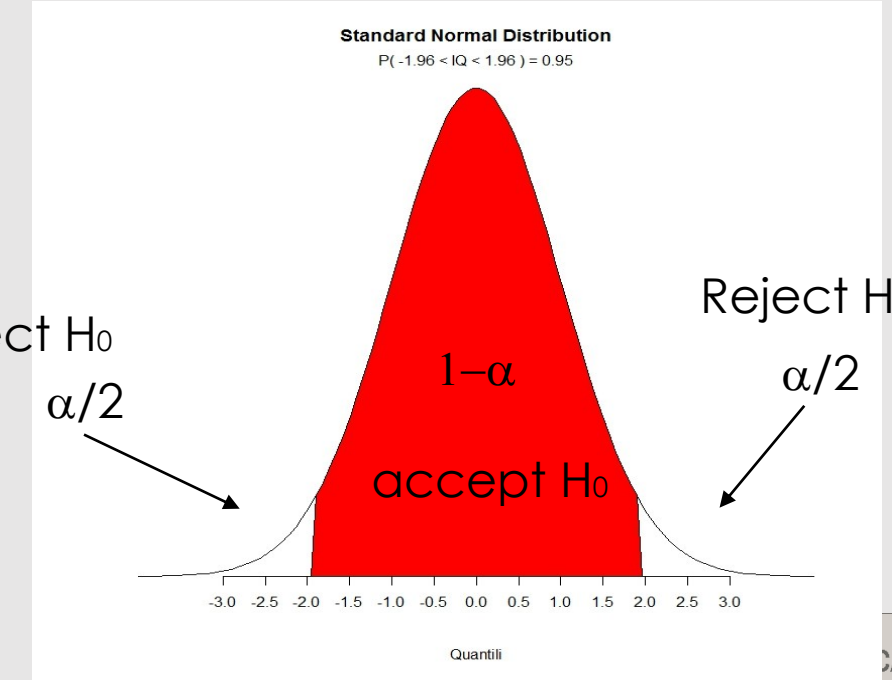
A probability β is associated with this error:
$1-\beta$ = Test **power**
β is not ***usually under control***, because the distribution of the test statistics is known *only* under the null hypothesis ...

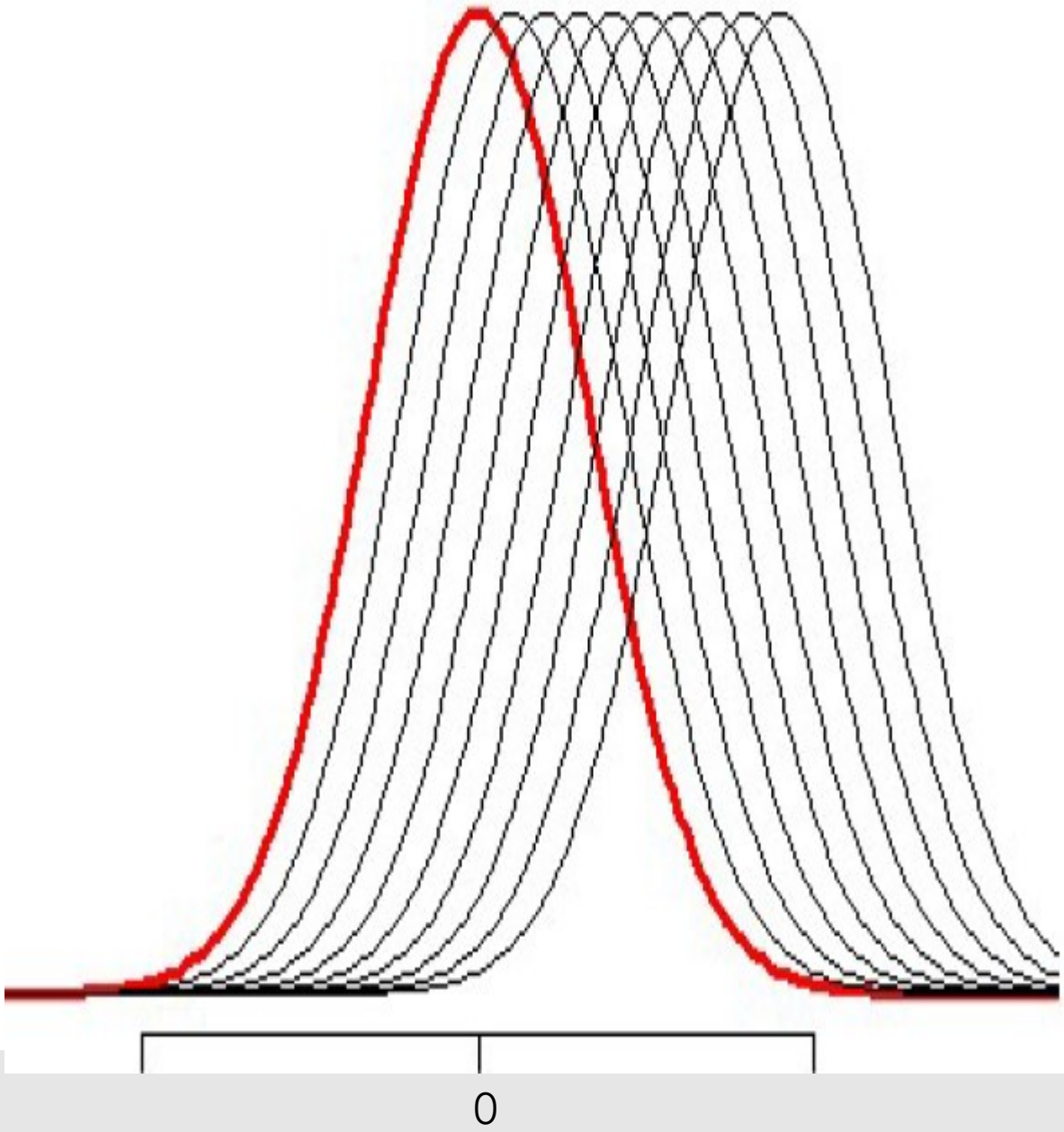
UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

Perform a statistical test (general strategy)

- **Null** Hypothesis H_0 versus **Alternative** hypothesis H_1 (mutually exclusive)
- The study is **designed** with the RV (= Random Variables) relevant to the problem: X1, X2, ...
- [A plausible ***model*** (data generating mechanism) is/*could be assumed* for RV]
- A test statistic **T (x1, x2, ...) = t** is calculated **on the random sample**; the probability distribution of T is known **if H_0 holds** (and differs from what it would have under H_1): the p-value is obtained
- H_0 is **rejected** if p value *is too unlikely* (if H_0 were true): if and only if $p <= \alpha$

***p-value*:** probability under H_0 that the RV T has the value *t* observed on the sample data or a more "extreme" value observed on the sample data

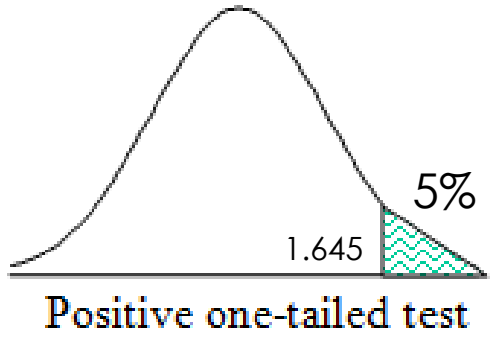




		True state of H_0 (Unknown)	
		H_0 true	H_0 false
Decision (sample data)	**Reject H_0**	**Type I error***	ok
	Do Not reject H_0	ok	**Type II error****

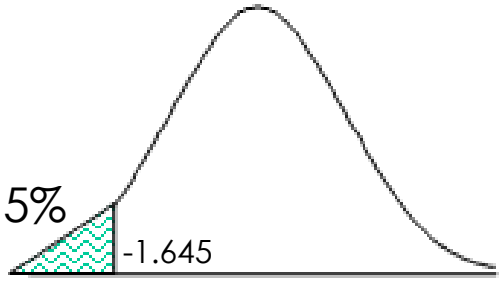
Defining the ***minimal*** *clinically relevant difference* (effect size) helps in determining the **sample size** required **to control** the **type II error**

One tail or two tails ??*



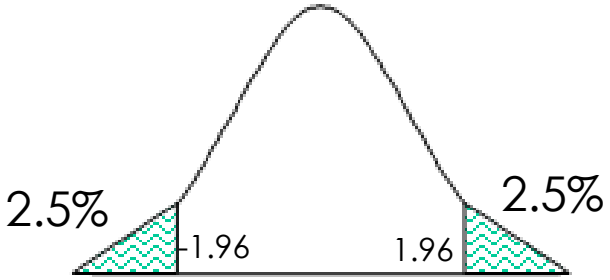
Positive one-tailed test

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



Negative one-tailed test

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



Two-tailed test

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

Suppose we compare two drugs A and B.

If it is believed that drug A **could be only better** than drug B, the **one-tailed** test will be performed.

Note that there is a risk of accepting the null hypothesis of equality even if A is ***worse*** than B.

Only if this probability is considered **negligible**, the one-tailed test could be used.

* If the sampling distribution of the test statistic is symmetrical

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

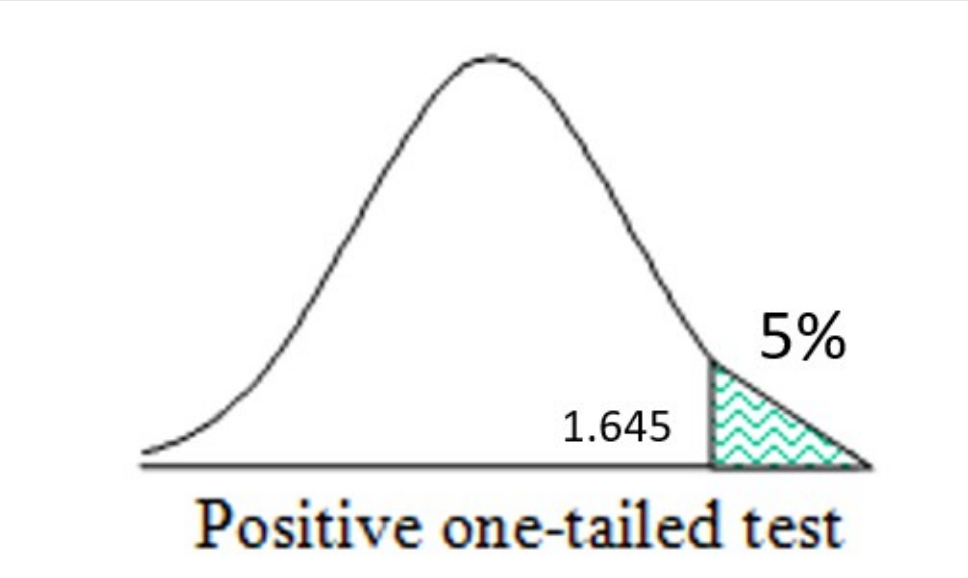
One-tail test [with a specific effect size]

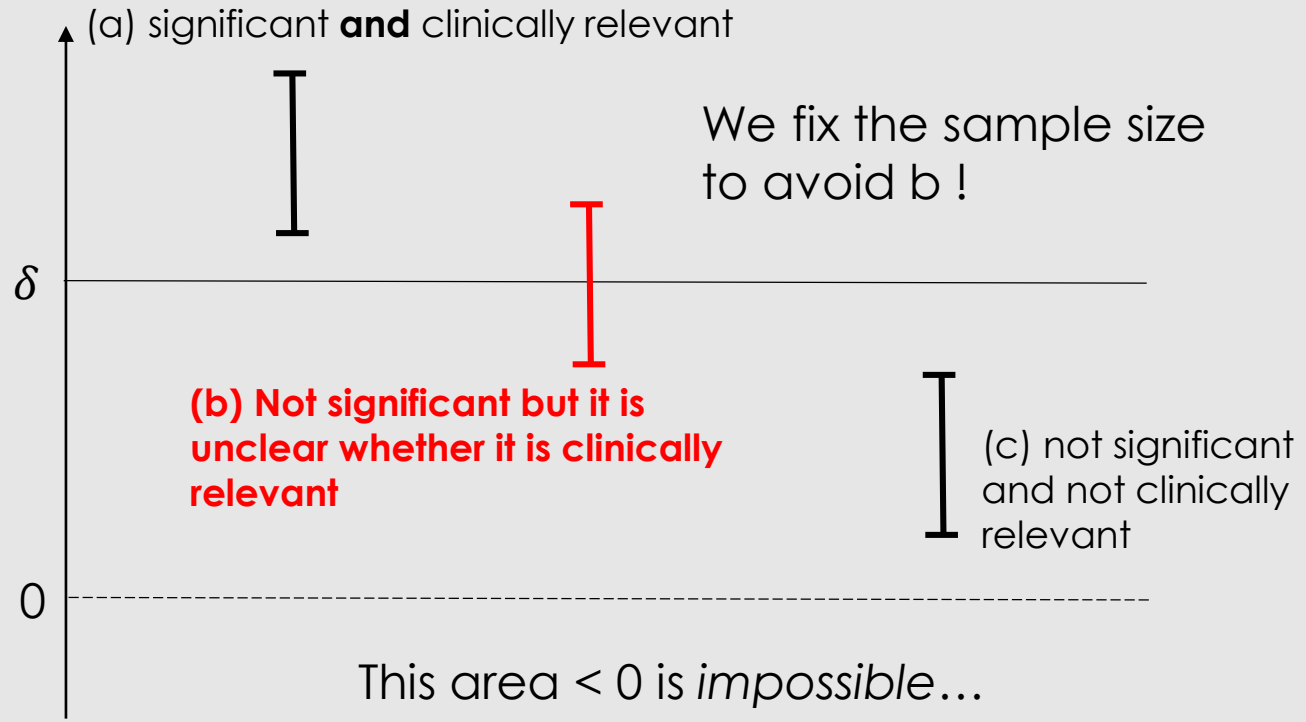
$\varepsilon = \mu_2 - \mu_1$

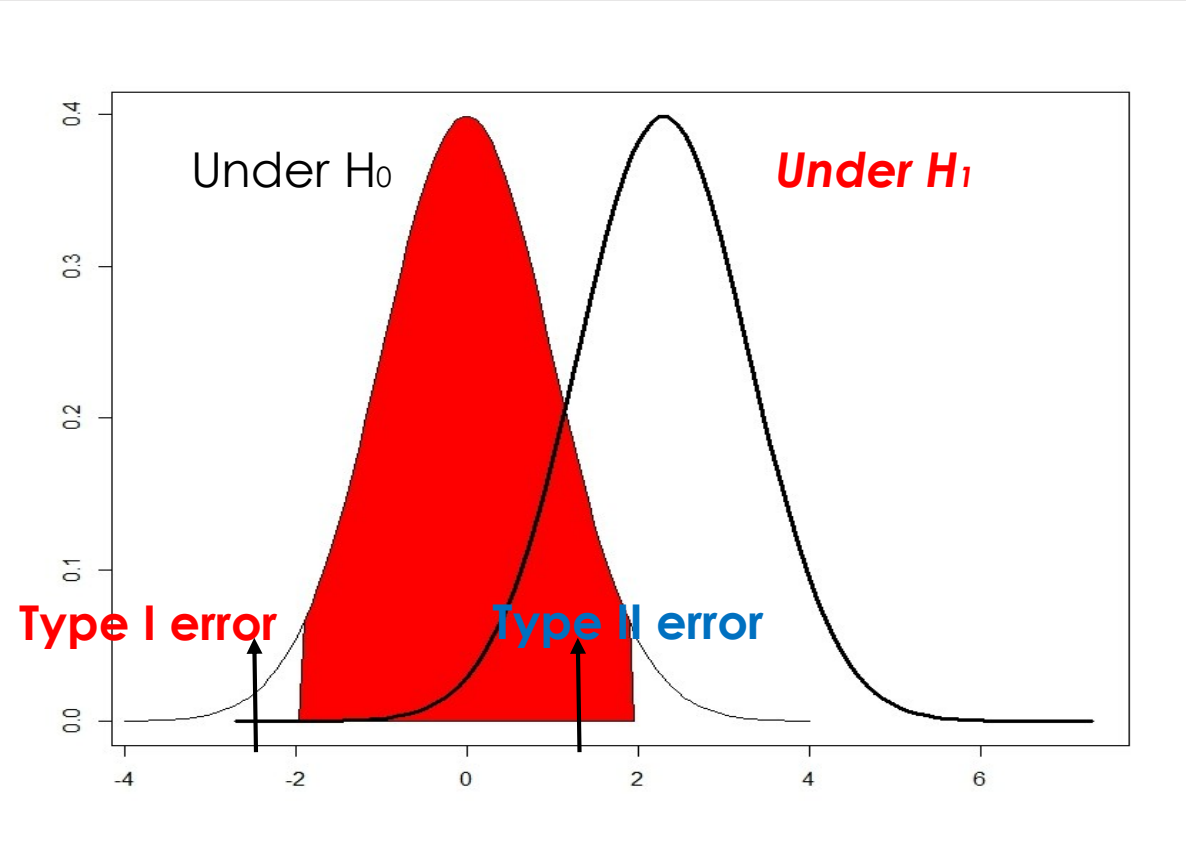
H_0: $\varepsilon \leq \delta$

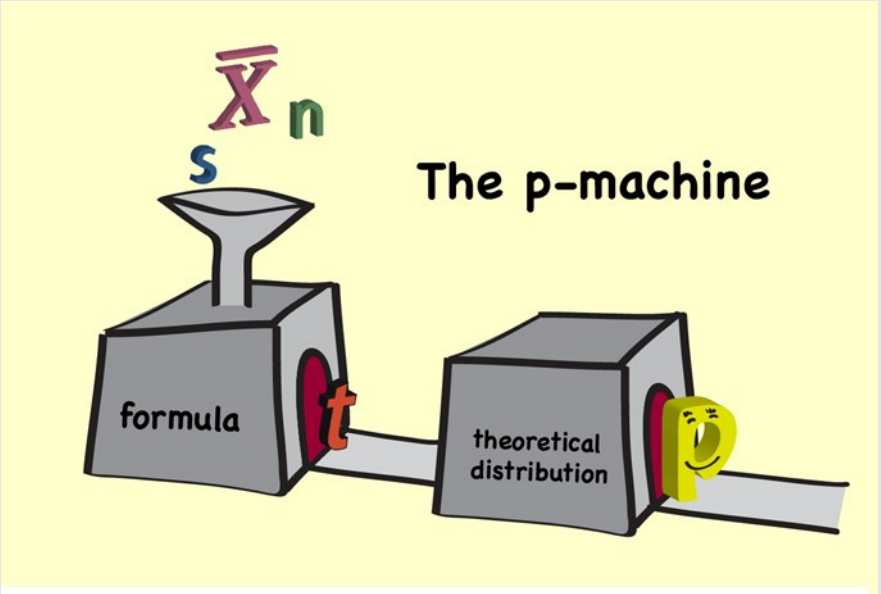
H_1: $\varepsilon > \delta$

What change in the **one-tail test** is only the **constant** of the gaussian distribution that fix the rejection threshold: for a 5% type I error instead of using 1.96, we will use 1.64.









- The calculation of test statistic and p-value is usually done by the **statistical software**
- The value of *p* quantifies the ***plausibility*** of the null hypothesis: the smaller it is, the less plausible (likely) H0 appears…
- Unless an **effect size** is fixed it is not possible to define the probability distribution of the test statistic under H1

Conclusions & Consequences

Suppose that α is small, typically $\alpha \leq 0.05$ **and that you have not fixed a priori** an effect size

- If H_0 is rejected : this decision is considered **reliable**. Type I probability of error α **is always fixed a priori** ("test result is statistically significant")

- If H_0 **is not** rejected, it is concluded that the data ***do not offer sufficient evidence*** to reject the null; but because β has been not fixed a priori could be ***large…***

"absence of evidence is not evidence of absence"

H_0 may not be rejected because the sample size *is too small* …

Test Power: the *hidden* ingredient

For a fixed α, and for a (typically not known and not modifiable) σ (variability of the outcome) the power of the test answers these questions:

- Given a sample size *N* and a "difference" (ES) between treatments Δ, what is the power $(1-\beta)$ to identify this difference, i.e. to conclude in favor of H_1?
- Given a certain power $(1-\beta)$ and a "difference" (ES) Δ, what sample size *N* is needed to identify the difference (i.e. support H_1)?
- Given a certain sample size *N* and a power $(1-\beta)$, what is the minimum difference Δ (ES) that can be identified with a $1-\beta$ probability?

Power = P(reject H_0 | *true* effect ≥ ***Effect Size***)

To calculate power it is necessary to have an estimate of σ and of the difference Δ (ES) under study

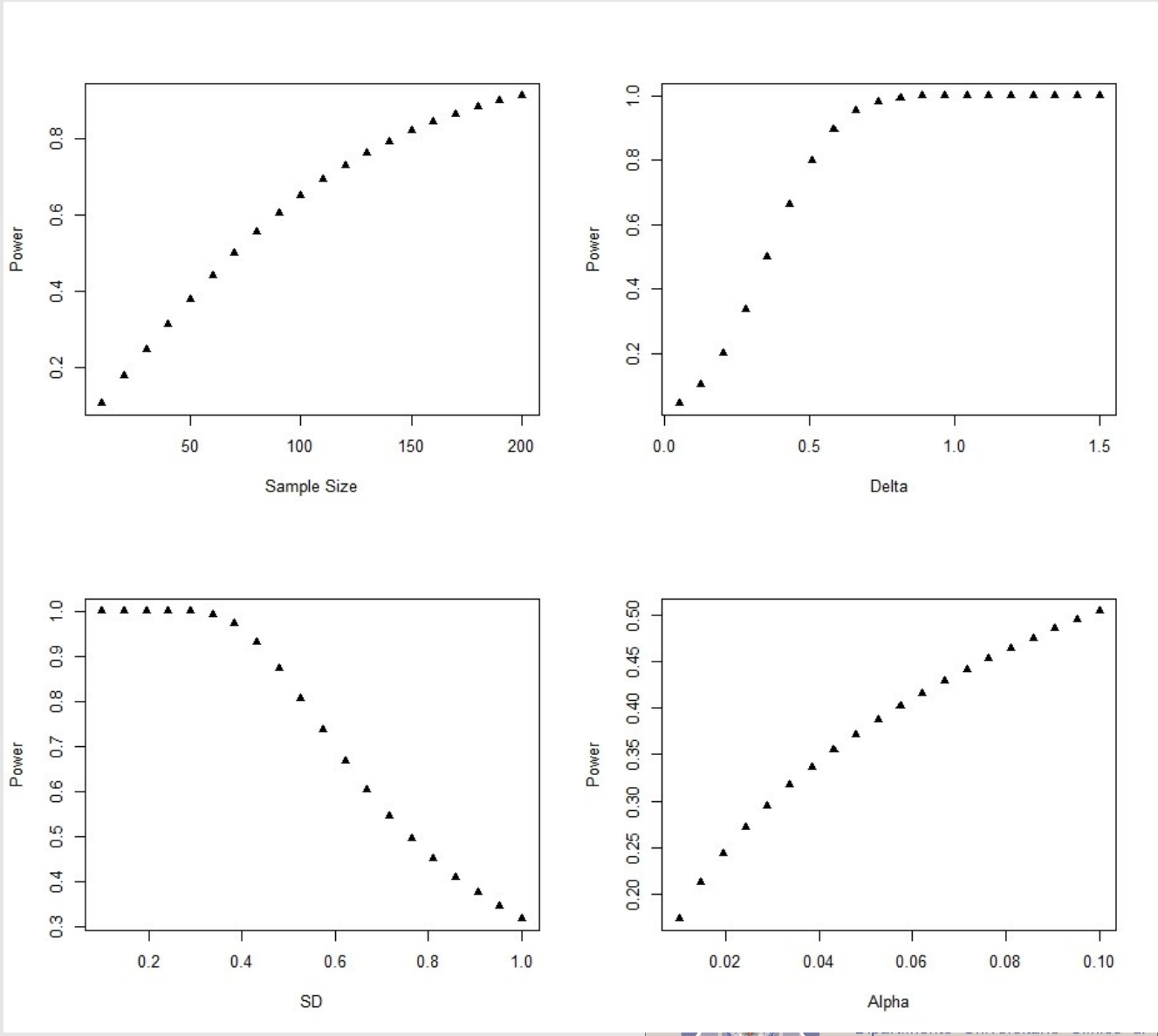
UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

Power of a test

t-test for two samples:
$H_0: \mu_1=\mu_2$ vs $H_1: \mu_1\neq\mu_2$
with $\sigma_1=\sigma_2=\sigma$

Power of the test is function of:
$\Delta=\mu_1-\mu_2(/\sigma)$ (ES), σ,n, α

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



Power of a test

Example (given the sample size):

Gold Standard vs *New* treatment: outcome is numerical (continous) with means μ_1 e μ_2 and standard deviations $\sigma_1=\sigma_2=\sigma$

New treatment vs Gold standard are considered clinically different if the mean difference is at least:

$\Delta=|\mu_1-\mu_2|$ >=**0.3** units

We have $\mathbf{n_1=n_2=50}$ patients eligible for each study arm

Significance level α=5% ; previous studies give an estimate of $\sigma \leq 0.9$ units

what is the **power** of the test?

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

Example

Given n=50 and $|\Delta|=0.3$ (with $\sigma=0.9$ and $\alpha=5\%$) what is the power?

Power: **38%** [Type II error = **62%**]

```
n <- 50
m <- 50
sigma <- 0.9
delta <- 0.3
```

$$\text{Power} = \Pr\left(\frac{|\overline{X}-\overline{Y}|}{\hat{\sigma}_p\sqrt{\frac{1}{n}+\frac{1}{m}}} > C \right)$$

$$\frac{\Delta}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}}$$

C= quantile of a ***noncentral*** t distribution

With this sample size the study has a **very low power** to detect the *hypothesized* difference

```
C  <- qt(0.975, n + m - 2)
se <- sigma * sqrt( 1/n + 1/m )

power <- 1 - pt(C, n+m-2, ncp=delta/se)
```

Assuming: 1-β=80% and n=50 (σ=0.9 and α=5%) what is the **smallest** difference Δ identifiable?

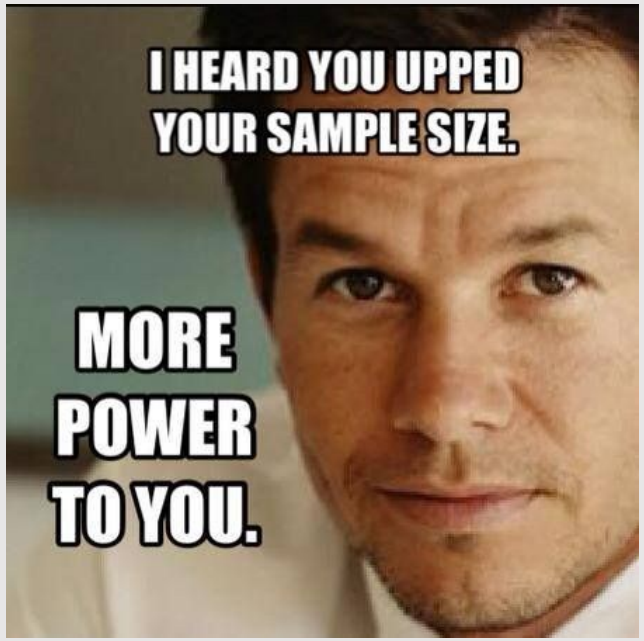
The minimum identifiable difference is **0.51** units given the sample size available and the measure of variability of the outcome

```
n <- 50
m <- 50
sigma <- 0.9
zb <- qt(0.80, n + m - 2)
za <- qt(0.975, n + m - 2)

Delta <- (zb+za)*sigma*sqrt(2/n)
```

Assuming: 1-β=80% and |Δ|=0.3 (σ=0.9 and α=5%) what is the ***'smallest' sample size*** required ?

The **minimum** sample size required is **141** patients in each group…



Basic formula* assuming the normal approximation:

```
zb <- qnorm(0.80)
za <- qnorm(0.975)
Delta <- 0.3
Sigma <- 0.9
```

```
n <- ((2*Sigma**2)*(za+zb)**2)/(Delta)**2
```

$$n = \frac{2 * \left(z_{\frac{\alpha}{2}} + z_{\beta}\right)^2 * \sigma^2}{(\mu_1 - \mu_2)^2}$$

* two-tailed test

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

For a well planned and conducted **RCT**, Type I and Type II errors rank higher as possible explanations for a finding of "*no statistically significant difference*" because randomization has ***overcome*** the potential confounding, the protocol has reduced measurement error, etc…

The idea of statistical power (especially for RCT) is quite simple.

1. We are going to do a study where we will evaluate the evidence using a **significance test**.

2. We decide what the **outcome** variable is going to be and what the comparison is going to be. outcome=systolic blood pressure ; comparison would be between **mean** blood pressure in 2 groups.

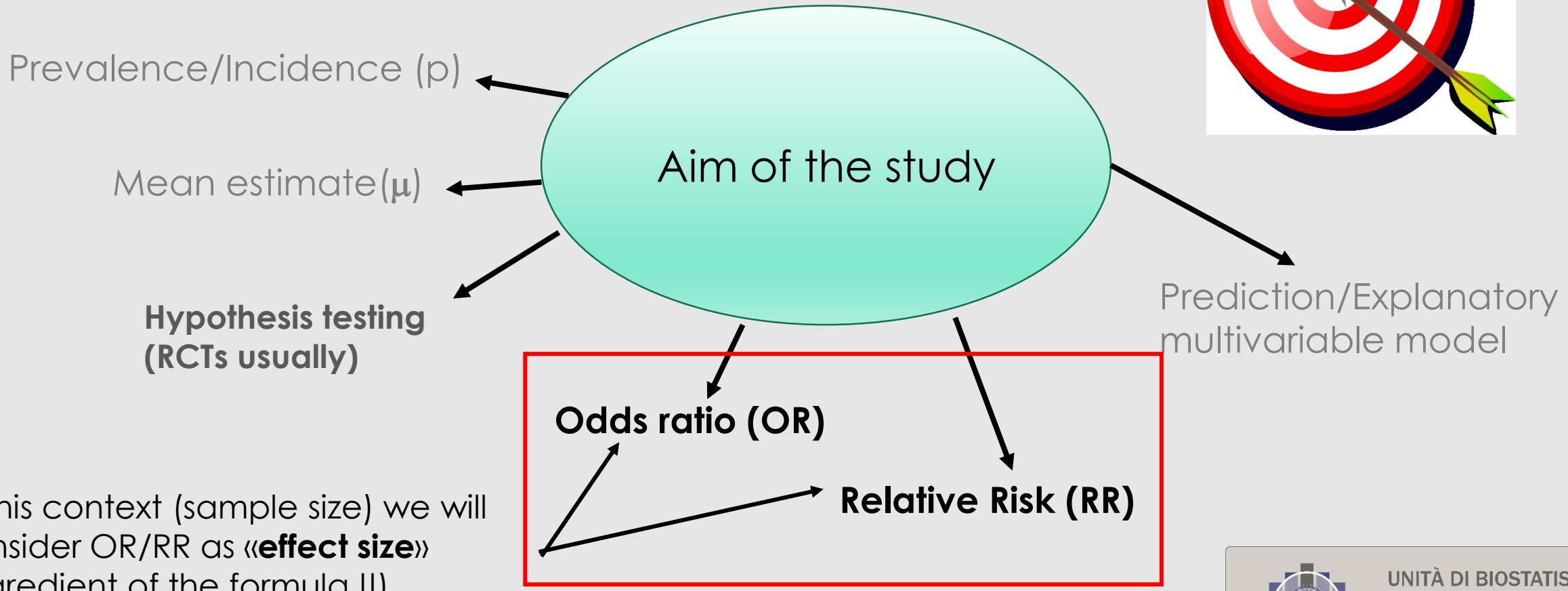
3. We then decide what the test of significance would be (ex: **two sample t test** comparing mean systolic pressure).

4. We decide **how big a difference we want the study to detect** - that is, how big a difference would be worth knowing about. For a two sample t test of mean systolic pressure, this could be the difference in mean pressure that would lead us to adopt the new treatment.

5. We then identify a sample size so that **if this difference were the actual difference in the population**, a large proportion of possible samples would produce a statistically significant difference.

This proportion is the power.

Sample size calculations **depend on the primary objective** of the study design:



In this context (sample size) we will consider OR/RR as «**effect size**» (ingredient of the formula !!)

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

Sample size for the estimate of the *strength* of an association measure

	D	Not D	
E	a	b	a+b
Not E	c	d	c+d
	a+c	b+d	n=a+b+c+d

Additional ingredients:

RR:

- incidence of event in the group of the unexposed
- not exposed / exposed ratio

OR:

- prevalence of exposure in the control group
- case-control ratio

Effect size approach corresponds here to the **strength** of the association that we expect (odds ratio or relative risk).

Sample size for hypothesis testing of the odds ratio

When testing an hypothesis about the OR, the most common H_0 is that of no effect is due to the exposure variable.

Under H_0 : OR=1 and the **% of exposed among cases** is equal to the **% of exposed among the controls**

Thus, H_0 is equivalent to that of equality of the two proportions

For a **specified H_1**, (OR is some number ≠ 1, ***effect size***):

$P_1 = P(E|D)$

$P_2 = P(E|\bar{D})$

Where: $P_1 = \dfrac{OR * P_2}{OR * P_2 + (1 - P_2)}$

P_2 is known: exposure rate among the controls

Ingredients:
- OR
- P_2
- Ratio Cases/Controls
- Power
- Alpha

[Remind: the *outcome* of the analysis is exposure rather than disease, but **symmetry** of the roles!!]

Sample size for hypothesis testing of the relative risk

$$RR = \frac{P(D|E)}{P(D|\bar{E})}$$

	D	Not D	
E	a	b	a+b
Not E	c	d	c+d
	a+c	b+d	n=a+b+c+d

Under H_0 : RR=1 and the **% of disease among exposed** is equal to the **% of disease among the unexposed**.

Thus, H_0 is equivalent to that of equality of two proportions

For a **specified H_1**, (RR is some number ≠ 1, **effect size**):

$$P_1 = P(D|E) \qquad P_1 = RR * P_2$$

$$P_2 = P(D|\bar{E})$$

P_2 is known: disease rate among the unexposed

Ingredients:
- RR
- P_2
- Ratio Exposed/Unexposed
- Power
- Alpha

In most situations, a case-control study requires a much **smaller** sample size than does a cohort study or exposure-based study for the same problem.

Consider, for example, a case-control study for the smoking and CHD problem:

A sample of men with newly diagnosed CHD will be compared for smoking status (smoker/non smoker) with a sample of controls. Assuming an equal number of cases and controls, how many are needed to detect an ***odds ratio*** of 2 with 90% power using a two-sided 5% test? Government surveys have estimated that 30% of the male population are smokers. A total of **376** men need to be sampled: **188** cases and **188** controls.

To be able to calculate an equivalent value for *SS* in a cohort study, we need an estimate of the chance of a coronary event (morbid or mortal) amongst non smokers: $P_2 = P(D|\bar{E})$

Let us suppose that the cohort study is to last 10 years, and for this period P_2 is estimated from a previous study to be 0.09. Note that here for simplicity we are treating incidence as cumulative (i.e. a **proportion**).

602 subjects need to be enrolled in a cohort study.

In Table are compared *SS* for the two study designs over a range of values for the relative risk (or at least its **approximate value** that could be derived from a case–control study).

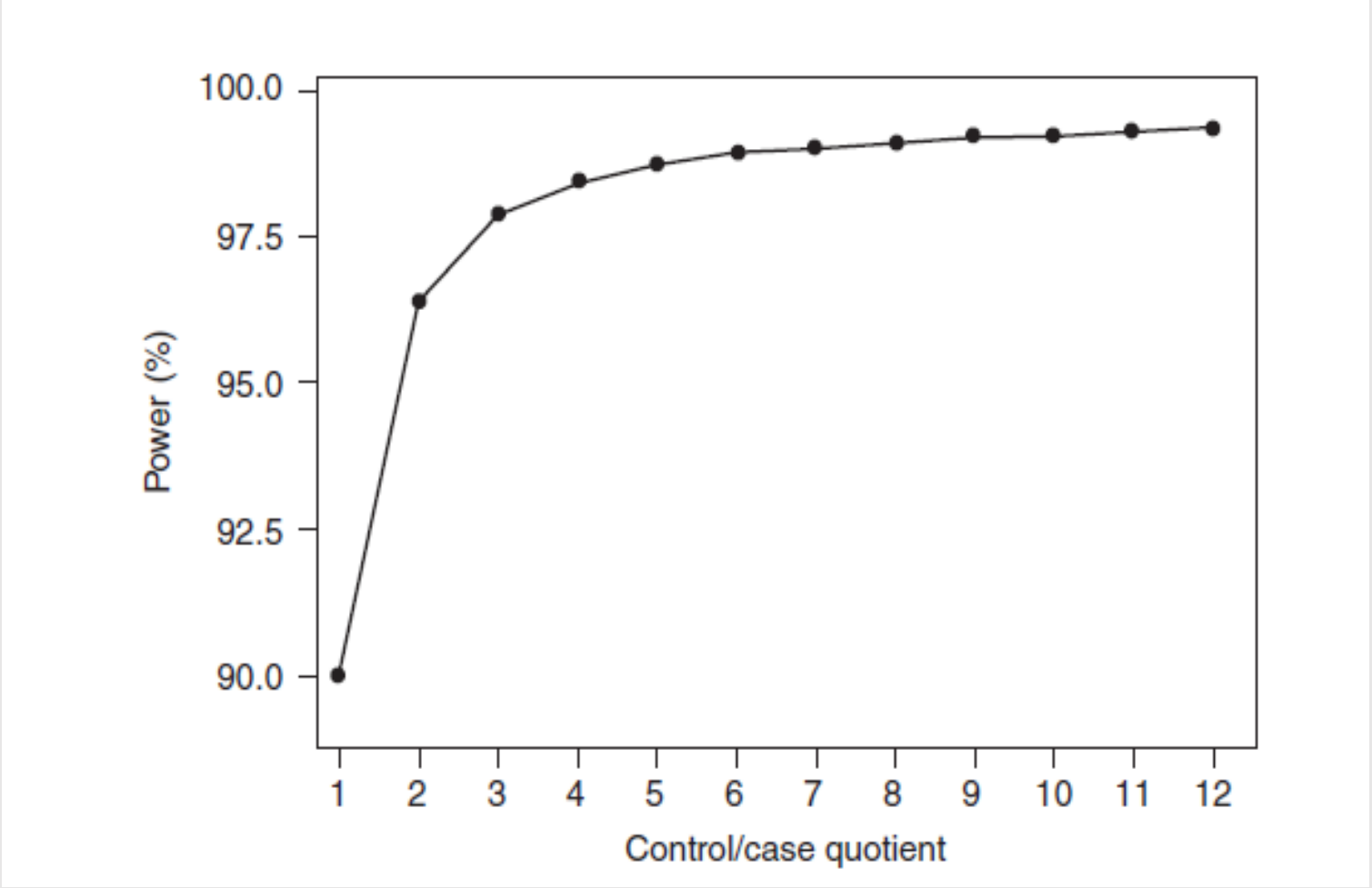
This show the advantage, in terms of **sample size requirements**, of a case-control study when the relative risk to be detected is **small**. Of note: coronary disease is not as rare as many diseases that are the subject of case–control studies (there are even greater savings with very rare diseases).

Sample size requirements to detect a given relative risk with 90% power using two-sided 5% significance tests for cohort and case–control studies.

Relative risk	Cohort study[a]	Case–control study[b]
1.1	44 398	21 632
1.2	11 568	5 820
1.3	5 346	2 774
1.4	3 122	1 668
1.5	2 070	1 138
2.0	602	376
3.0	188	146

[a] assuming an incidence of 0.09 for the nonfactor group.
[b] assuming a prevalence of 0.3 for the risk factor.

Case-Control ratio (or Exposed/Not exposed): it is rarely necessary to include more than 3 or 4 controls (or not exposed) compared to the cases (or exposed).



We reported various examples that give approximate sample size in the most straightforward situations that arise in clinical/epidemiological research.

One of the common requirement is to specify the **effect size** that we want to be able to detect with some high probability.

This requires careful thought. Often the researcher will begin by being *overoptimistic*, specifying an effect size so small that it requires an *enormous* sample to have a good chance of detecting it.

Usually the value ultimately decided upon is some **compromise** between various objectives, including conserving resources.

The ultimate decision may only be obtained after **a few trial calculations**. In this context, the 'inverse' formulae for power and minimum detectable effect size may well be useful.

It is quite possible that the value for *SS* needed to be able to detect the effect size that we would really like to find with high probability is *beyond* our resources. This problem has no easy solution: we must find more resources or accept reduced power.

Throughout, we have assumed that sample size may be determined by considering only **one** variable of interest (*exposure/risk factor*).

Frequently, the study will include **several variables**; for instance, we might be planning a lifestyle survey that will measure height, weight, blood pressure, cholesterol, daily cigarette consumption and several other things.

We might well find that the optimal value of *SS* for analysing height, say, is considerably different from that for analysing cholesterol.

Similarly, there could be **several end-points of interest**.

If there are multiple outcomes, ideally the value for *SS* might be calculated for each criterion.

The *maximum* of all these gives the value for SS(tot) that satisfies all requirements.

This will often be far more than is needed for some of the criteria and hence may be considered too wasteful.

An alternative approach is to pick the **most important criterion (primary outcome)** and use this alone.

A **limitation** with the examples provided is that they make no allowance for **confounding** variables.

That is, they consider only **unadjusted/univariable** comparisons.

In general, the issue of allowing for confounding in sample size estimation is very complex.

We will discuss in Block 3 more specific approaches for SS estimation when the objective is the estimation of a multivariable regression model.

We should remember that the equations for sample size are based upon *probabilities* (through the power) and *assumptions*.

There can be no *absolute* guarantee that the important difference will be detected, when it does exist, even with a very high power specification.

All we can say is that we will run *very little risk of missing it* when we specify a high value for the power. The higher the power is, the lower is the risk (if the assumption hold).

Sample size evaluation is *not an exact science, even when some so-called exact methods are* used, *because assumptions* made may be violated by the data collected.

We can regard the sample size computed only **as a reasonable guide**.

https://www.youtube.com/watch?v=PbODigCZqL8

To end with a little fun:

Biostatistics vs. Lab Research

210.131 visualizzazioni • 6 ago 2010

968 15 CONDIVIDI SALVA ...

J JavaMama926
173 iscritti

ISCRIVITI

How not to collaborate with a biostatistician. This is what happens when two people are speaking different research languages! My current workplace is nothing like this, but I think most biostatisticians have had some kind of similar experiences like this in the past!