Hypothesis testing

Prof. Giuseppe Verlato
Unit of Epidemiology & Medical Statistics,
Department of Diagnostics & Public Health,
University of Verona

From '900 hypotheses have only been falsified (Karl Popper)

In'800 hypotheses were demonstrated (Claude Bernard)

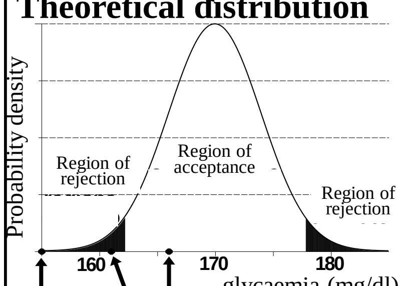
HYPOTHESIS TESTING

Null Hypothesis (H_0)
(All observed differences are due to chance)

Alternative Hypothesis (H_1)

FALSIFYING NULL HYPOTHESIS

Theoretical distribution



~~Theoretical distribution~~

Non-parametric or distribution-free tests

Data are sorted in ascending order:

AAABAABBABBBABBBBB

The values in group A tends to be lower than the values in group B

In the Middle Ages mankind was right in the middle of everything:

. of space (geocentric theory)

. of time (using the Bible a scholar had computed that the Earth had been created 4,000 years before Christ)

. of the biologic world (fixism theory)

Nowadays mankind is NO LONGER in the middle:

. of space (heliocentric theory, the Sun peripherally located in the Milky Way, 100-200 billions galaxies in the Universe)

. of time (the age of the Earth is about 4.5 billion years, the Big Bang happened 13.82 billion years ago)

. of the biologic world (Darwin's theory of species evolution)

«sentì che era un punto al limite di un continente, sentì che era un niente»
(La canzone della bambina portoghese, Radici, 1972, Francesco Guccini)

SCIENTIFIC HYPOTHESIS:

A statement that can be supported or refuted through experimentation or observation. In other words, an essential property of scientific hypothesis is falsifiability.

Is aspirin effective against cold? ~~Does the soul survive after death~~?

A scientific hypothesis is hold true until it is falsified.

STATISTICAL HYPOTHESIS:

A statement about a **population** characteristic which can be supported or refuted according to available information, usually obtained from a **sample**.

HYPOTHESIS TESTING

H_0: NULL HYPOTHESIS
All observed differences are simple random variations

H_1: ALTERNATIVE HYPOTHESIS
Differences observed among sample statistics reflect real differences among parameters of source populations

Example:

Italian diabetic patients have the same glycaemia as American diabetic patients

Glycaemia differs between Italian and American diabetic patients

STATISTICAL TEST:

Objective mathematical rule which allows to discriminate between sample observations which allow to accept or reject the null hypothesis (H_0).

The probability, that the final decision (acceptance/refusal) is correct, is also reported.

Data from samples

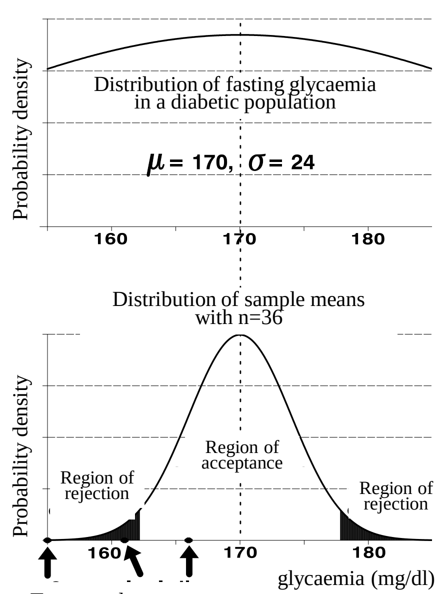
Statistical test

P>0.05 = the probability that observed differences are due to chance is greater than 5%

Null hypothesis (H0) is accepted
All differences observed among samples can be attributed to chance

P<0.05 = the probability that observed differences are due to chance is lower than 5%

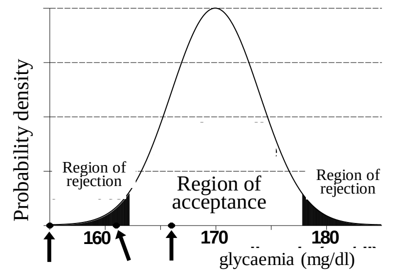
Null hypothesis is rejected and alternative hypothesis is favored
Differences observed among samples reflect real differences among source populations



Two sample means support H0 rejection, the other one H0 acceptance

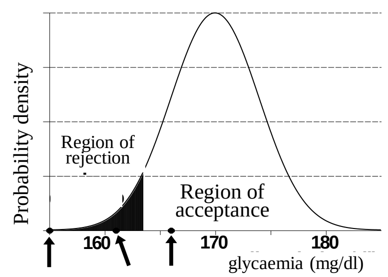
Two-sided (two-tailed) test

H_0: $\mu_{Italians} = \mu_0$ H_1: $\mu_{Italians} \neq \mu_0$



One-sided (one-tailed) test

H_0: $\mu_{Italians} \geq \mu_0$ H_1: $\mu_{Italians} < \mu_0$



Null hypothesis (H_0)

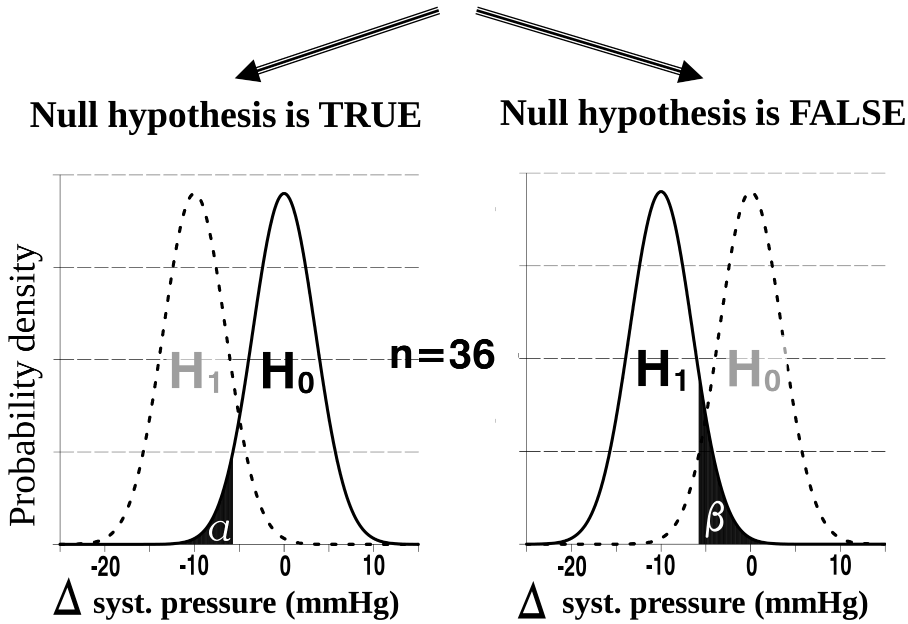
	true	false
H_0 accepted	**O.K.**	**Type II error**
H_0 rejected	**Type I error**	**O.K.**

P(type I error) = α (alpha)

P(type II error) = β (beta)

In hypothesis testing probability of type I error is usually set at 5% (0.05). Hence H_0 will be rejected (and the statistical test will turn out to be significant) at RANDOM in one in twenty tests, even if H_0 is true. In statistical terms significance level is set at 5%.

For instance, if in hypothesis testing *P*<0.01, H_0 can be rejected with a probability of type I error lower than 1%; in other words, the probability that observed differences are due to chance is lower than 1%.



In the current scientific literature the exact probability associated with a given statistical test is usually reported.

For instance a ***P*-value=0.003** is given, rather than reporting ***P*<0.05**. In statistical terms, a **significance test** is performed rather than a **hypothesis testing**.

However, while ***P*<0.05 is the preselected probability that a "significant" difference is due to chance under H_0**,

***P*=0.003 is the probability of the observed result or of a more extreme result under H_0. American statisticians call it PTOME (Probability of This Or More Extreme)**.

When the significance level is set at 5%, one test in 20 turns out to be significant simply by chance. If 100 statistical tests are performed, 5 will be significant by chance.

Hence the scientific literature is at risk to be flooded with new discoveries that later on will appear to be false. This would inflate alpha, the probability of type I error. This alpha inflation is called “multiple testing bias” (*distorsione da test ripetuti*).

For instance multiple testing bias occurs when:

1) considering several outcomes,
2) statistical tests are repeated in different subgroups
3) statistical tests are repeated at different times.

SUBGROUP ANALYSIS

In the international study ISIS2 (1998) aspirin proved to be superior to placebo in the treatment of myocardial infarction and in the prevention of further ischemic episodes.

However, in the Twins Zodiac sign placebo was more effective than aspirin.

References

ISIS-2 (Second International Study of Infarct Survival) Collaborative Group (1998) Randomized trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction: ISIS-2. Lancet ii: 349-360.

How can we prevent **multiple testing bias**?

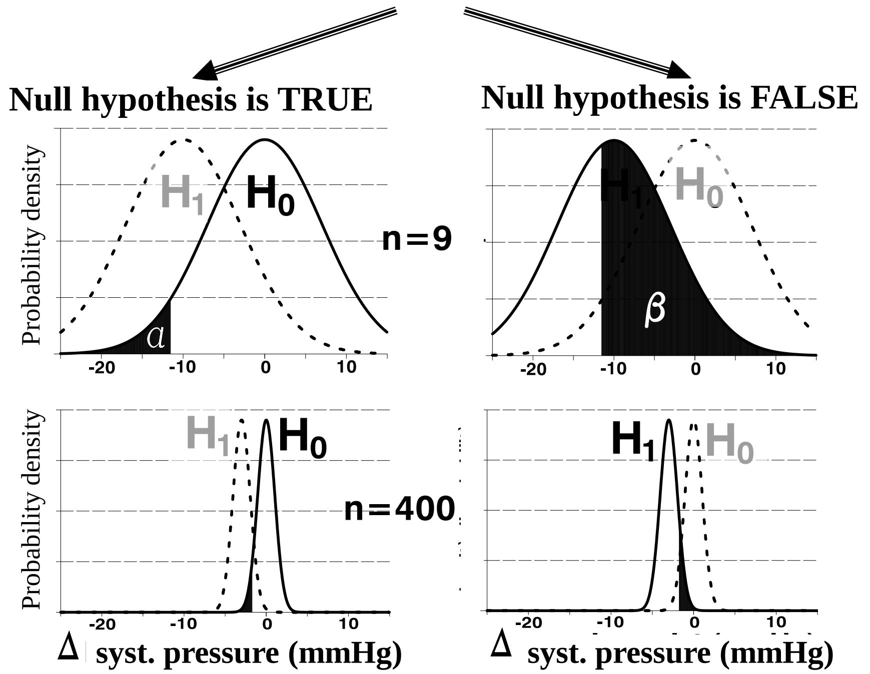
The family physician dealing with laboratory tests	The biostatistician dealing with several statistical tests
- finds out the most important tests	- distinguishes between primary end-points (1 or 2) and secondary end-points (several)
- takes into particular account biochemical values largely out of the normal range	- adopts more conservative significance values ($p<0.05 \rightarrow p<0.01$) or corrects observed *P*-values (Bonferroni's correction)
- check the agreement among different tests (biomarkers of liver disease, of viral infection)	- check the agreement among different end-points

POWER of a statistical test=1-beta=1-P(type II error)

It is the probability to falsify the null hypothesis when null hypothesis is indeed false.

In other words, the power of a statistical test reflects its ability to detect differences, when these differences do exist.

The statistical test has been developed in order to keep constant the level of significance, irrespective of sample size. But this result has been achieved at the expenses of statistical power, which is largely affected by sample size.



STATISTICAL SIGNIFICANCE ***versus***
CLINICAL IMPORTANCE

An epidemiological survey, performed on a large number of people, highlighted that smokers sleep less than non-smokers.

The difference was highly significant ($p<0.001$), i.e. it was unlikely to be due to chance.

The difference consisted in 3 minute decrease in sleep duration in smokers as compared to non-smokers.

A test POWER depends on:

1) sample size

2) variability of the characteristic under study

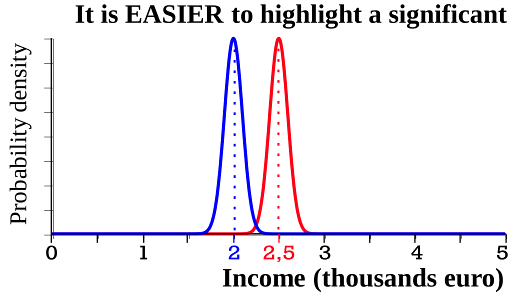
3) minimal difference to be highlighted

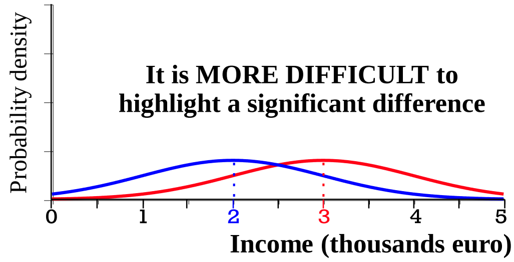
4) significance level adopted.

The most important way to achieve an adequate power is to plan an adequate sample size in the study protocol.

$$\text{Statistical significance} \approx \frac{\text{Observed difference}}{\text{Random variability}}$$

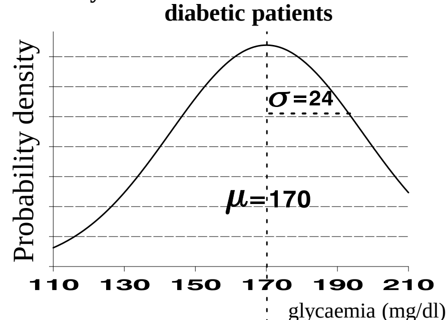
It is EASIER to highlight a significant difference





Confidence interval as hypothesis testing

Glycaemia distribution in American diabetic patients



Confidence interval is an interval estimate, but it can be considered also as an hypothesis test.

NULL HYPOTHESIS:
Glycaemia distribution in Italian diabetic patients is the same as in American diabetic patients.
Hence sample means with n=36 is normally distributed with:

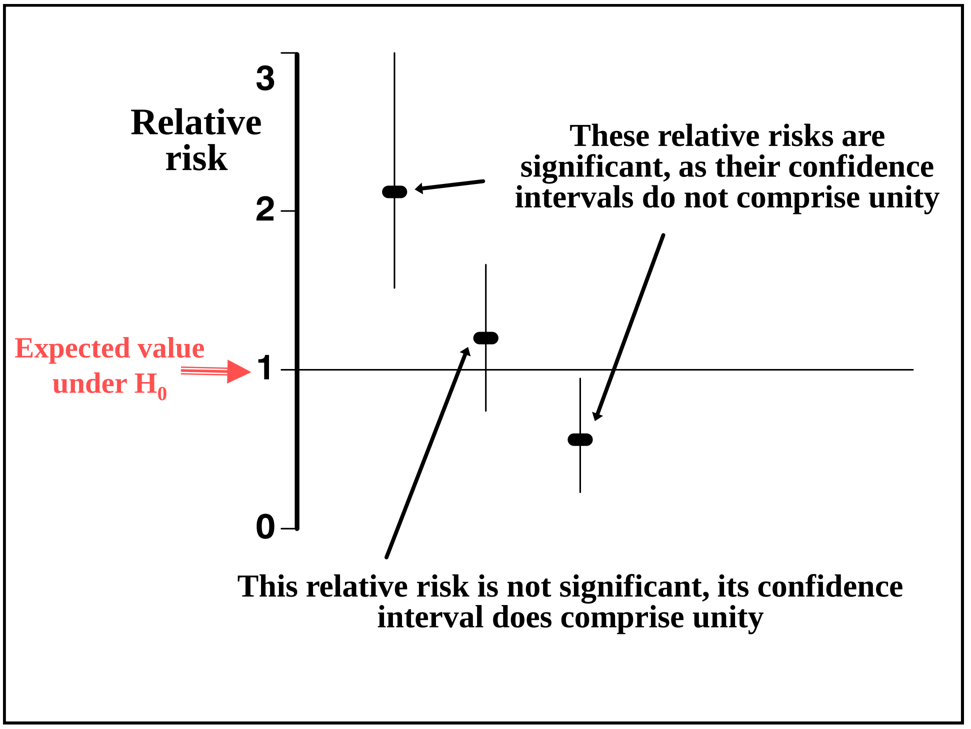
$\mu = 170$

$\sigma/\sqrt{n} = 24/\sqrt{36} = 4$

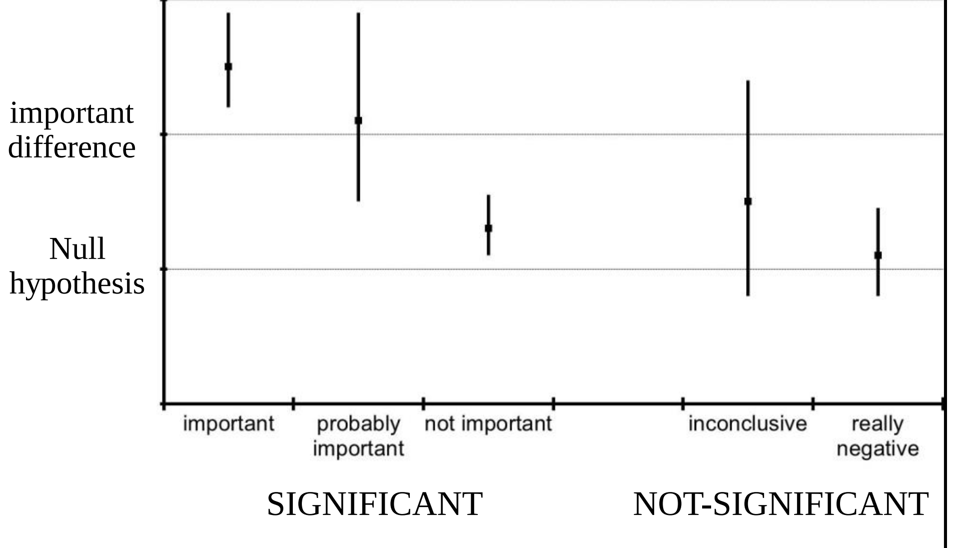
The first two confidence intervals are significant as they do not comprise 170 mg/dl (expected value under H0)

Interval estimates of μ

$155 \pm 1{,}96*4$	147,2 —— 162,8
$161 \pm 1{,}96*4$	153,2 —— 168,8
$166 \pm 1{,}96*4$	158,2 —— 173,8



Evaluating a confidence interval from a clinical/statistical point of view



"Overemphasis on hypothesis testing - and the use of P values to dichotomise significant or non-significant results - has detracted from more useful approaches to interpreting study results, such as estimation and confidence intervals.

In medical studies investigators are usually interested in determining the size of difference of a measured outcome between groups, rather than a simple indication of whether or not it is statistically significant ...

Confidence intervals, if appropriate to the type of study, should be used for major findings in both the main text of a paper and its abstract."

Gardner MJ, Altman DG (1986) Confidence intervals rather than P values: estimation rather than hypothesis testing. British Medical Journal, 292: 746-750

International Committee of Medical Journal Editors

"When possible, quantify findings and present them with appropriate indicators of measurement error or uncertainty (such as confidence intervals). Avoid sole reliance on statistical hypothesis testing, such as the use of P values, which fails to convey important quantitative information."

International Committee of Medical Journal Editors (1992) Uniform requirements for manuscripts submitted to biomedical journals [Special Report] N Engl J Med, 324: 424-428.

Choice of the statistical test

When starting a new statistical analysis, the following question must be answered first: "*Which type of variable is the outcome of the study* ?"

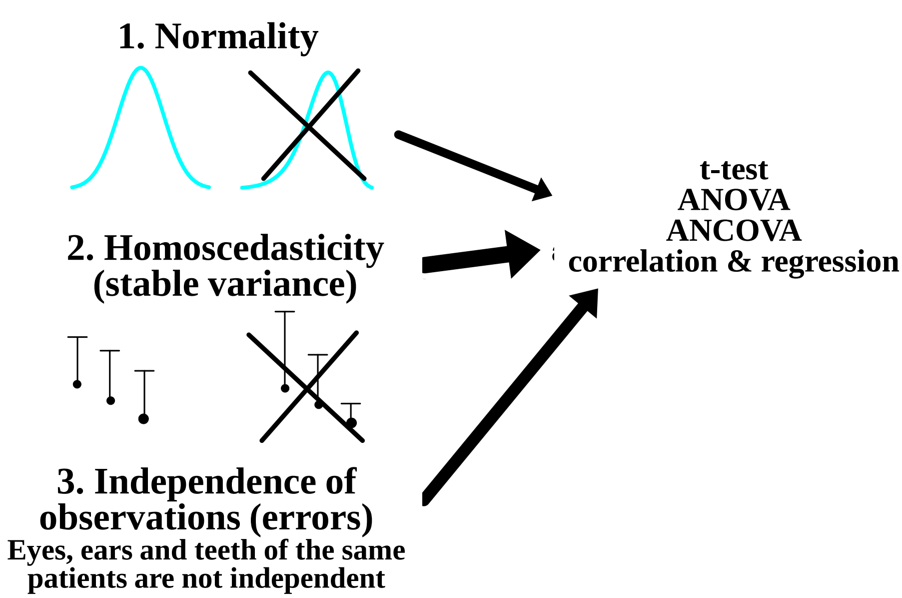
	NOMINAL	ORDINAL	QUANTITATIVE
Examples	Life status (alive/dead) Sex (M/F) Country of origin	Pain intensity Depth of coma	Weight (Kg) Age (years) Glycaemia (mmol)
Suited test	Chi-squared (χ^2) Fisher's exact test McNemar's test	Non-parametric tests	t test for unpaired or paired data Analysis of variance Correlation and regression

ASSUMPTIONS of PARAMETRIC TESTS

1. The outcome should be normally distributed
2. Variability should be approximately the same in different groups

3.The observations should be independent

ASSUMPTIONS OF PARAMETRIC TESTS



Which test should be used with QUANTITATIVE variables ?

Comparison between different subjects		Repeated measures on the same subjects		Relation between different variables
↓	↓	↓	↓	↓
Two groups	Three or more groups	Two measurements	Three or more measurements	
↓	↓	↓	↓	
T test	One-way ANOVA	T test for paired data	ANOVA for repeated measures	Correlation and regression

ANOVA = ANalysis Of VAriance

1) An anthropometric study is performed on university students. The Body Mass Index (weight/height^2) of 1st class students is compared to the Body Mass Index of 3rd class students. Which test should be used for this purpose ?

2) In the same study the Body Mass Index (weight/height^2) is measured twice, at the start of the university career and upon graduation. Which test should be used for this purpose ?

3) In the same survey the relation between height and weight is addressed. Which test should be used for this purpose?

4) In the same survey the relation between eye colour and hair colour is addressed. Which test is suited for this purpose?

A) T test

B) T test for paired data

C) Chi-squared test

D) Correlation and regression

E) other ___________________

NON-PARAMETRIC TESTS

PARAMETRIC TESTS		**NON-PARAMETRIC TESTS**
Heart rate, arterial blood pressure	**Variables**	**pain, Glasgow coma score**
T test	Two independent samples	Mann-Whitney U test, also known as Wilcoxon rank-sum test
One-way ANOVA	K independent samples	Kruskal-Wallis test
T test for paired data	Two dependent samples	Wilcoxon signed-rank test
ANOVA for repeated measures	K dependent samples	Friedman’s test
Correlation and regression	Relation between two different variables	Spearman's correlation coefficient