Hypothesis Testing

Sorana D. Bolboacă

MEDICAL STATISTICS

"Medical students may not like statistics, but as doctors they will."

Martin Bland, Letter to the Editor, 1998. BMJ; 316:1674.

"Medical students may not like statistics, but as good doctors they will have to understand statistics."

John Chen, 2004, Advice to GCRC & Surgery Fellows and Residents

OBJECTIVES

- Understand the principles of hypothesis-testing
- To be able to correctly interpret P values
- To know the steps needed in application of a statistical test
- To be able to state statistical hypotheses: null and alternative

DEFINITIONS

- **Statistical hypothesis test** = a method of making statistical decisions using experimental data.
- A result is called **statistically significant** if it is unlikely to have occurred by chance.
- Statistical hypothesis = an assumption about a population parameter. This assumption may or may not be true.
- Clinical hypothesis = a single explanatory idea that helps to structure data about a given subject in a way that leads to better understanding, decision-making, and treatment choice.

[Lazare A. The Psychiatric Examination in the Walk-In Clinic: Hypothesis Generation and Hypothesis Testing. Archives of General Psychiatry 1976;33:96-102.]

DEFINITIONS

Clinical hypothesis:

- A proposition, or set of propositions, set forth as an explanation for the occurrence of some specified group of phenomena, either asserted merely as a provisional conjecture to guide investigation (working hypothesis) or accepted as highly probable in the light of established facts
- A tentative explanation for an observation, phenomenon, or scientific problem that can be tested by further investigation.
- Something taken to be true for the purpose of argument or investigation; an assumption.

HYPOTHESIS TESTING

- Hypothesis testing is a form of inferential statistics used to determine the probability (or likelihood) that a conclusion based on analysis of data from a sample is true.
- We use hypothesis testing to make comparisons between:
  - One sample and a population
  - Between 2 or more samples
- A statistical hypothesis test produces a **p-value**, or the probability of obtaining the results (or more extreme results) from tests of samples, if the results really were not true in the population.

Statistical Test Frequently used in Medicine

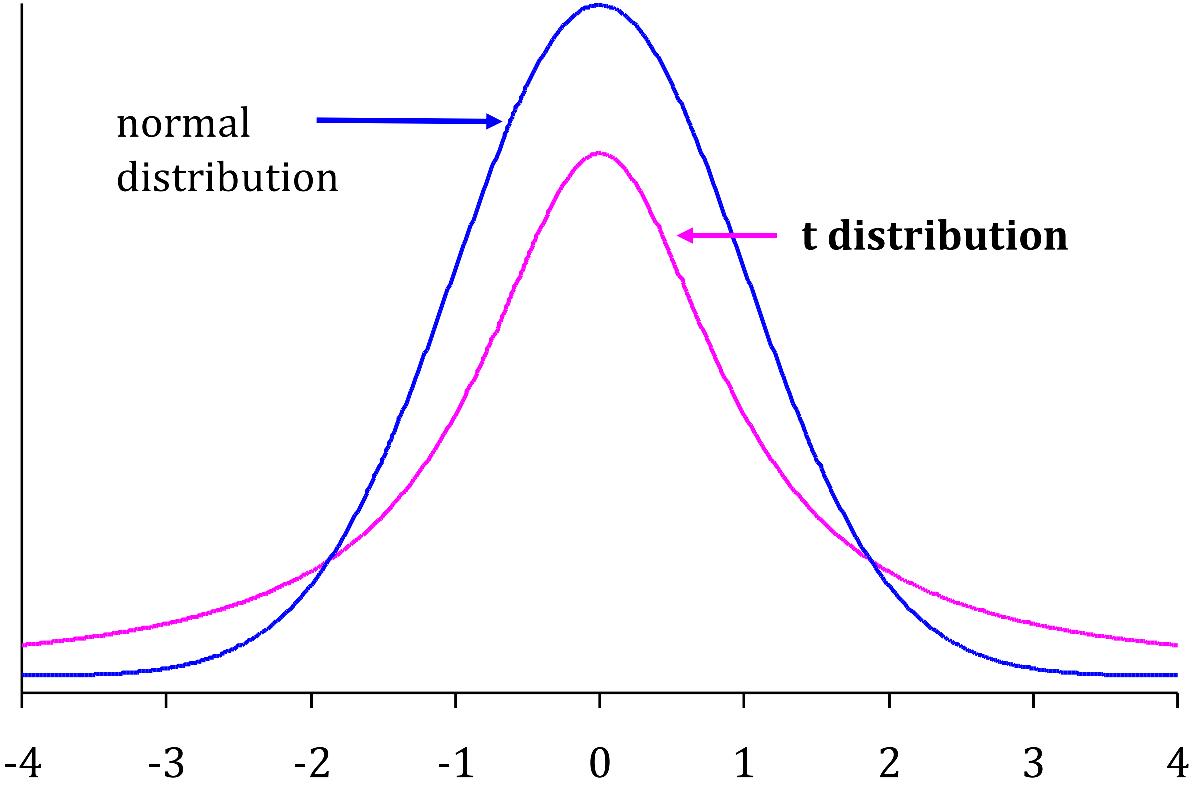
- Parametric tests (quantitative normal distributed data):
  - T-test for dependent or independent samples (2 groups)
  - ANOVA (2 or more groups)

- Non-parametric tests (qualitative data – nominal or ordinal):
  - Chi-Square test
  - Fisher’s exact test

- Test for associations (quantitative & qualitative data):
  - Correlation (Pearson & Spearman) & Regression (Linear & Logistic)

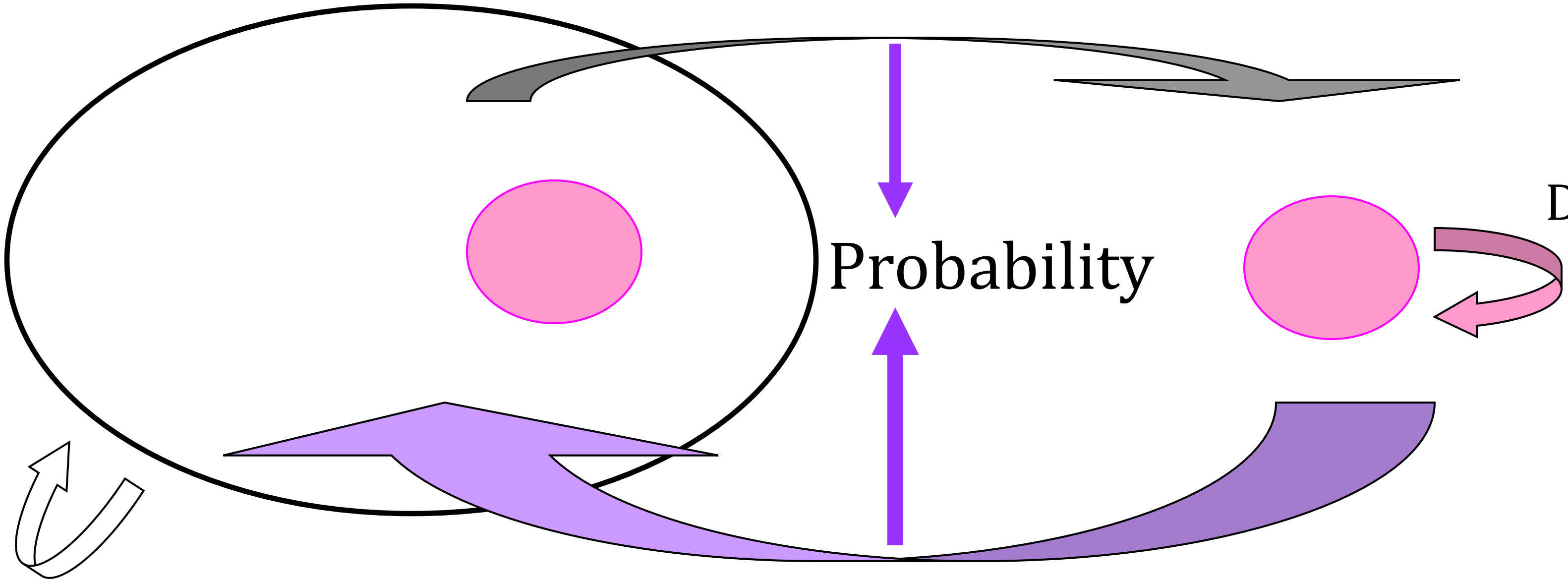
HYPOTHESIS TESTING

- In all hypothesis testing, the numerical result from the statistical test is compared to a probability distribution to determine the probability of obtaining the result if the result is not true in the population.



From Probability to Hypothesis Testing

Sampling



Probability

Descriptive statistics

Descriptive statistics

Inference (Inferential statistics)

Population

Parameters (μ, σ)

Sample

Statistics (m, s)

From Probability to Hypothesis Testing

What we Learned from Probability

1) The mean of a sample can be treated as a random variable.

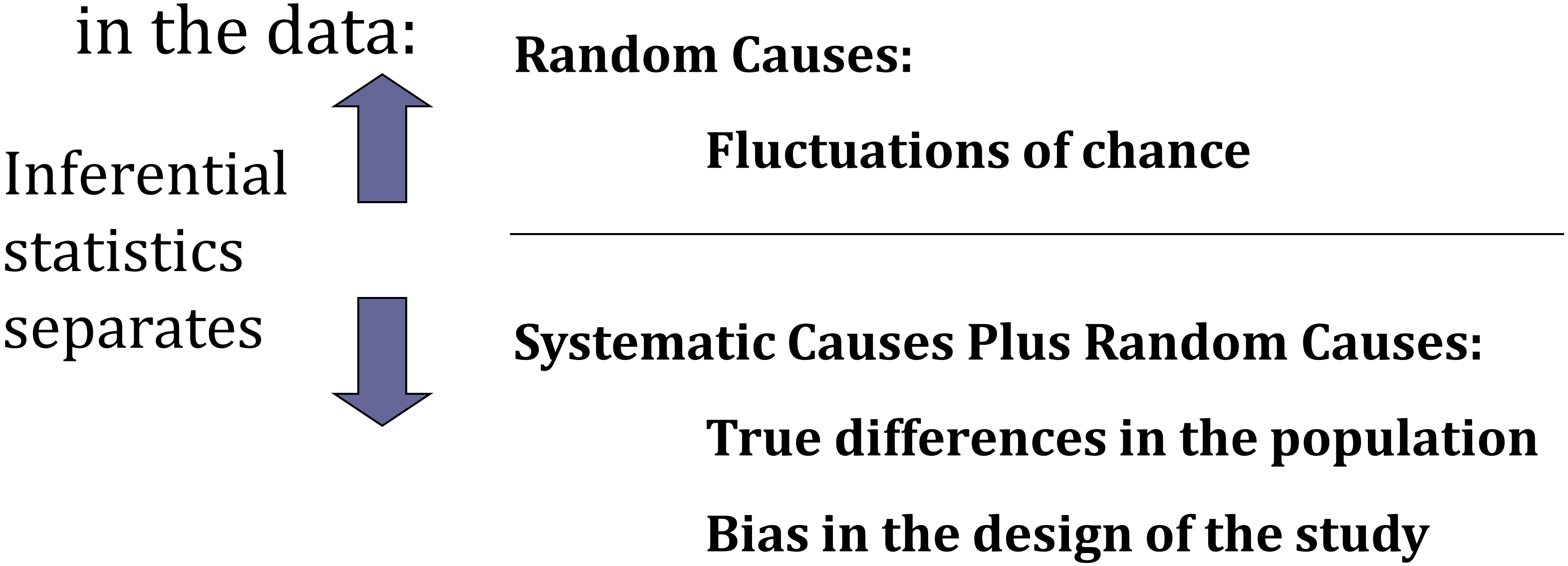
2) By the central limit theorem, sample means will have a normal distribution (for n > 30) with $\mu_{\overline{X}} = \mu$ and $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$

3) Because of this, we can find the probability that a given population might randomly produce a particular range of sample means.

$$P(\overline{X} > something) = P(Z > something) = \text{Use standard table}$$

INFERENTIAL STATISTICS

- Once we have got our sample
- The key question in statistical inference:
  - Could random chance alone have produced a sample like ours?
- Distinguishing between 2 interpretations of patterns in the data:



Inferential statistics separates

Random Causes:

Fluctuations of chance

Systematic Causes Plus Random Causes:

True differences in the population

Bias in the design of the study

REASONING OF HYPOTHESIS TESTING

1. Make a statement (the null hypothesis) about some unknown population parameter.

2. Collect some data.

3. Assuming the null hypothesis is TRUE, what is the probability of obtaining data such as ours? (this is the “p-value”).

4. If this probability is small, then reject the null hypothesis.

HYPOTHESIS TESTING: STEP 1

- State the research question in terms of a statistical hypothesis
  - Null hypothesis (the hypothesis that is to be tested): abbreviated as H_0
    - Straw man: “Nothing interesting is happening”
  - Alternative hypothesis (the hypothesis that in some sense contradicts the null hypothesis): abbreviated as H_a or H_1
    - What a researcher thinks is happening
    - May be one- or two-sided

HYPOTHESIS TESTING: STEP 1

- Hypotheses are in terms of population parameters

One-sided	Two-sided
H_0: $\mu=110$ $H_{1/a}$: $\mu < 110$ OR $H_{1/a}$: $\mu > 110$	H_0: $\mu = 110$ $H_{1/a}$: $\mu \neq 110$

HYPOTHESIS TESTING: STEP 2

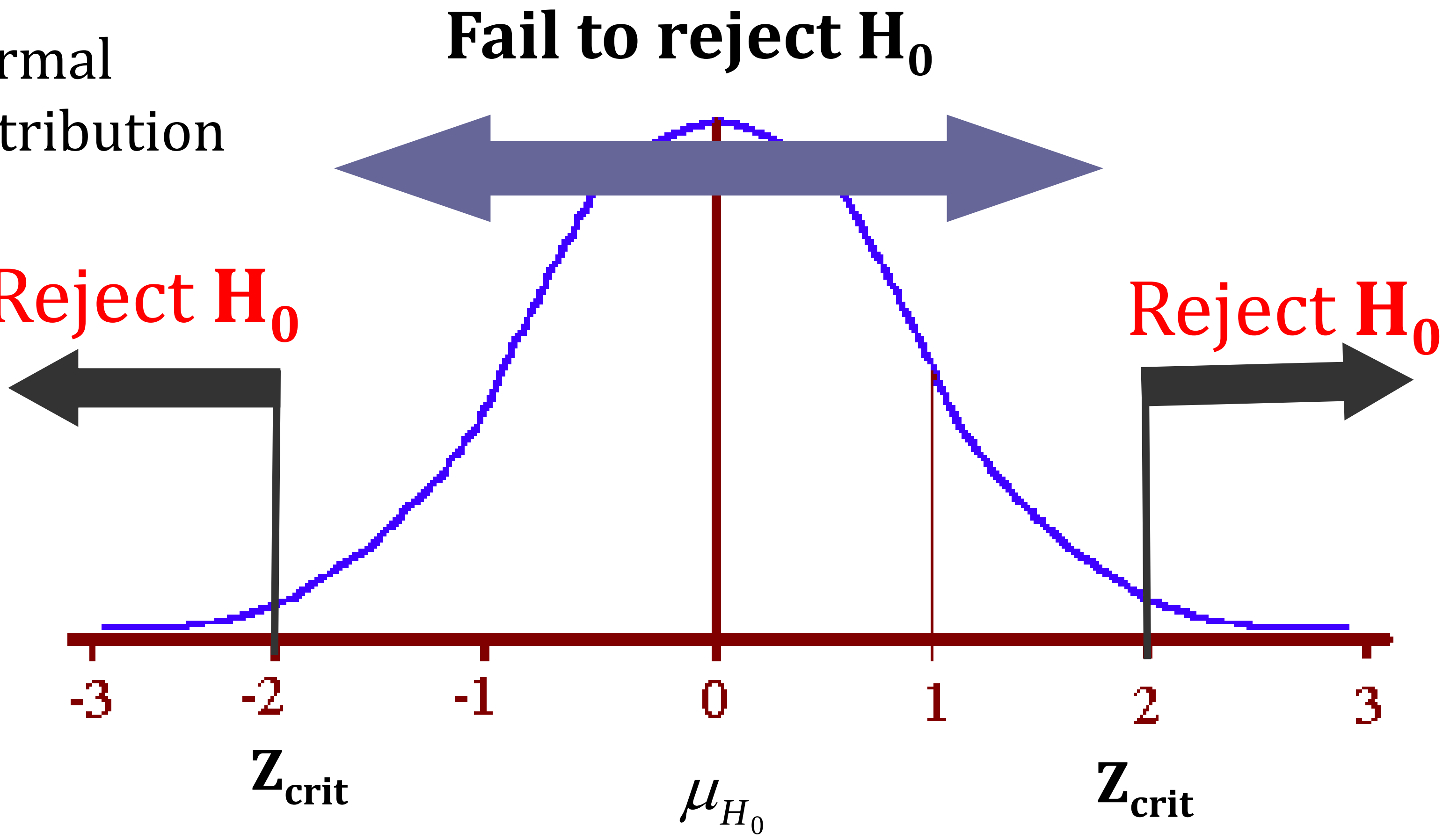
- Set decision criterion:
  - Decide what p-value would be “too unlikely”
  - This threshold is called the **alpha level / significance level**.
  - When a sample statistic surpasses this level, the result is said to be significant.
  - Typical **alpha levels** are **0.05** and **0.01**.
- Alpha levels (level of significance) = probability of a type I error (the probability of rejecting the null hypothesis even that H0 is true)
- The probability of a type II error is the probability of accepting the null hypothesis given that H_1 is true. The probability of a Type II error is usually denoted by β.

HYPOTHESIS TESTING: STEP 3

- Setting the rejection region (comparison of means):
  - The range of sample mean values that are “likely” if H_0 is true.
    - If your sample mean is in this region, retain the null hypothesis.
  - The range of sample mean values that are “unlikely” if H_0 is true.
    - If your sample mean is in this region, reject the null hypothesis.

Hypothesis Testing: Step 3

Normal distribution



Fail to reject H_0

Reject **H_0**

Reject **H_0**

Z_{crit}

μ_{H_0}

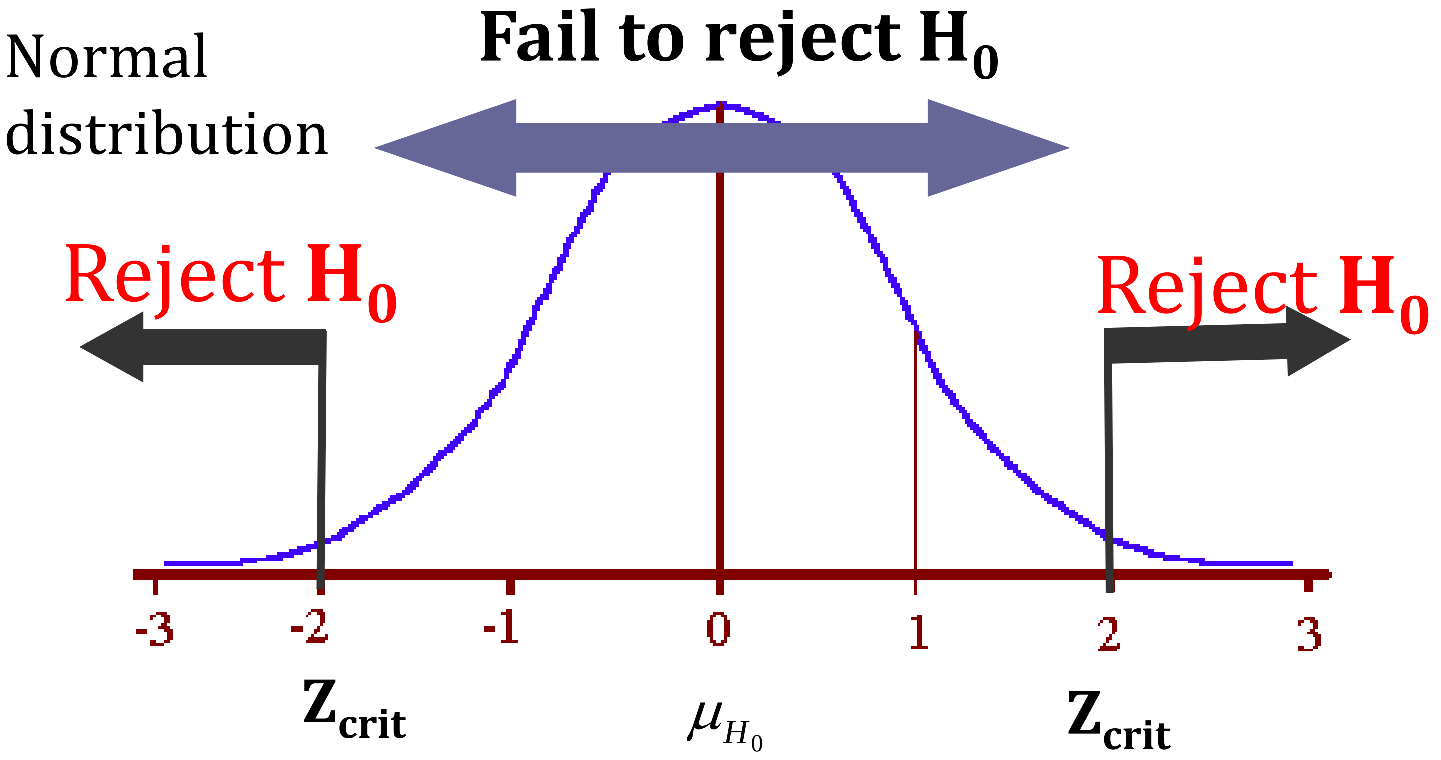
Z_{crit}

HYPOTHESIS TESTING: STEP 4

- Compute sample statistics
- A test statistic (e.g. Z-test, T-test, or F-test) is information we get from the sample that we use to make the decision to reject or keep the null hypothesis.

- A test statistic converts the original measurement (e.g. a sample mean) into units of the null distribution (e.g. a Z-score), so that we can look up probabilities in a table.

HYPOTHESIS TESTING: STEP 4



- If we want to know where our sample mean lies in the null distribution, we convert X-bar to our test statistic Z_{test}
- If an observed sample mean were lower than Z=-1.65 then it would be in a critical region where it was more extreme than 95% of all sample means that might be drawn from that population

HYPOTHESIS TESTING: STEP 5

- State the test conclusion:
  - If our sample mean turns out to be extremely unlikely under the null distribution, maybe we should revise our notion of μ_{H0}
  - We never really "accept" the null hypothesis. We either **reject it**, or **fail to reject it**.

STEPS IN HYPOTHESIS TESTING

Step 1: State hypothesis (H_0 and H_1/H_a)

Step 2: Choose the level of significance ($\alpha = 5\%$)

Step 3: Setting the rejection region

Step 4: Compute test statistic (Z_{test}) and get a p-value

Step 5: Make a decision

ONE- VS. TWO-TAILED TESTS

- In theory, should use one-tailed when
  1. Change in opposite direction would be meaningless
  2. Change in opposite direction would be uninteresting
  3. No rival theory predicts change in opposite direction

- By convention/default in the social sciences, two-tailed is standard

- Why? Because it is a more stringent criterion (as we will see). A more conservative test.

One- vs. Two-Tailed Tests

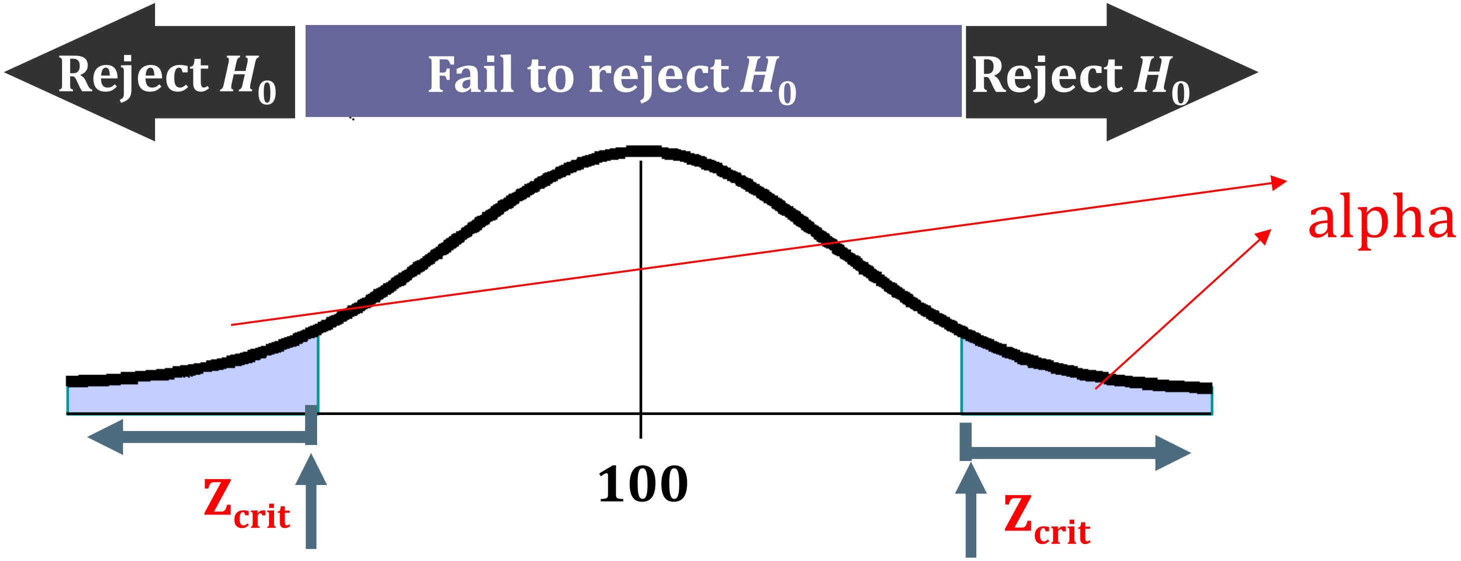
- H_a is that $\bar{X}$ is *either* greater or less than μ
  - H_a: $\bar{X} \neq \mu$

- α is divided equally between the two tails of the critical region

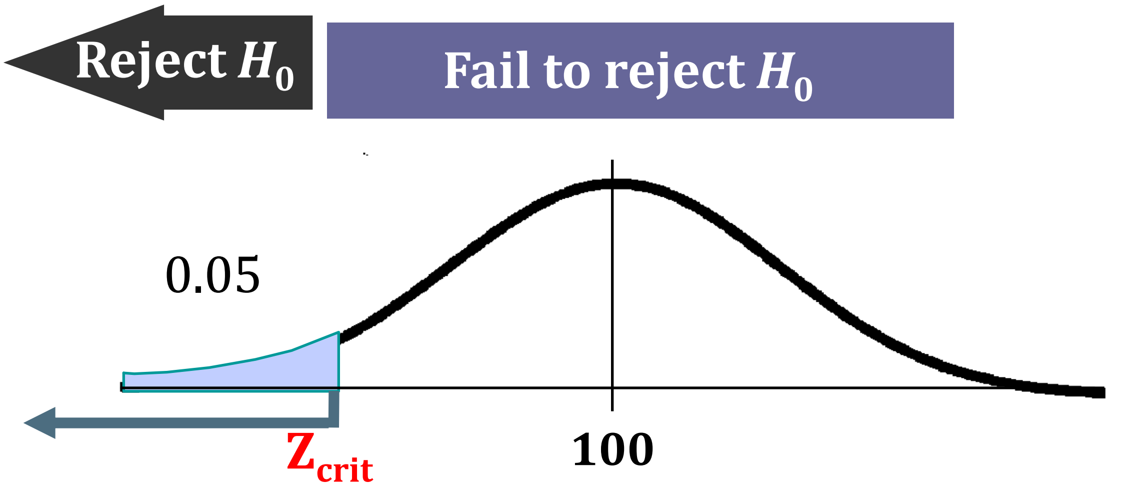
Two-Tailed Hypothesis Testing

H_0: μ = 100

H_1: μ ≠ 100   Values that differ significantly from 100



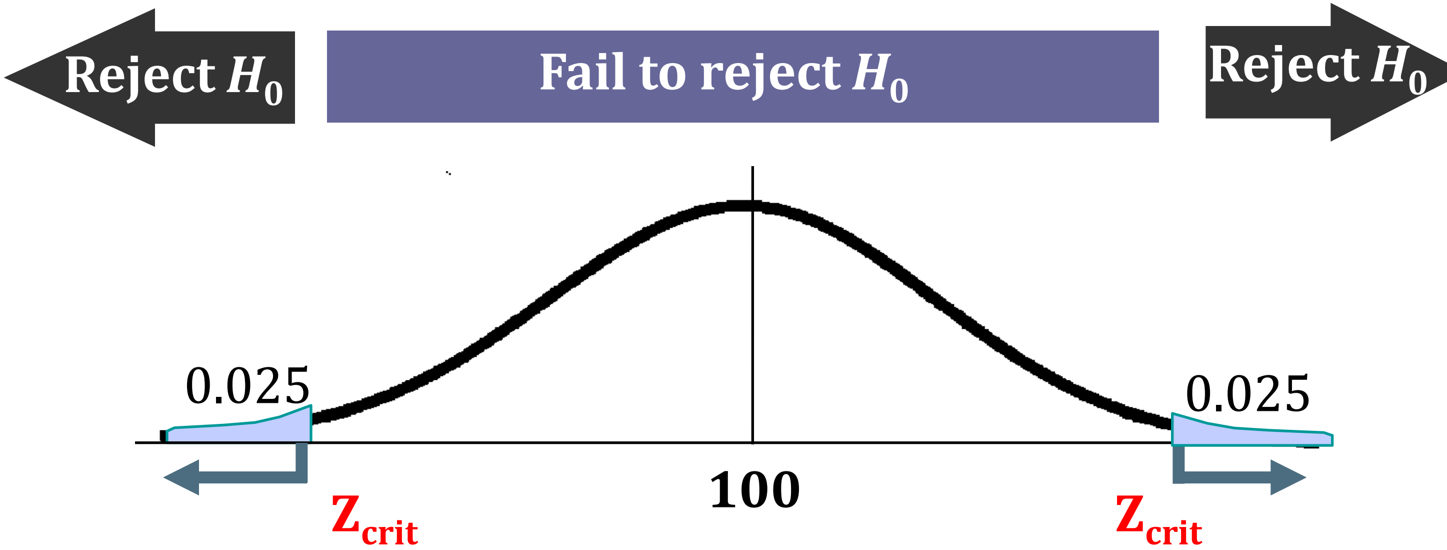
One tailed



Values that
differ “significantly”
from 100

Values that are significantly less than 100

Two tailed



Values that differ significantly from 100

P VALUES vs. CONFIDENCE INTERVALS

- A **P** value measures the strength of evidence against the null hypothesis.
- A **P** value is the probability of getting a result as, or more, extreme if the null hypothesis was true.
- It is easy to compare results across studies using **P** values
- **P** values are measures of **statistical significance**
- Confidence intervals give a plausible range of values in clinically interpretable units
- Confidence intervals enable easy assessment of **clinical significance**

P VALUES vs. CONFIDENCE INTERVALS

Statistical significance can be obtained from a confidence interval as well as a hypothesis test

&

Confidence intervals convey more information than p values

- For this reason, most medical journals now prefer that results be presented with confidence intervals rather than p-values.
- If the NULL VALUE for a statistical hypothesis test using alpha = 0.05 is contained within the 95% confidence interval, we can conclude that there is <u>NO</u> statistical significance at a significance level of 5% <u>without doing the hypothesis test</u>

P VALUES vs. CONFIDENCE INTERVALS

- For a 95% confidence interval for a difference between two values [-7.7 to 2.1]
    - The 95% CI includes 0, so there is no statistically significant difference between my values. In addition, we have information about the precision of our estimate of the difference, which cannot be obtained from p values alone.

!!! This is a relatively wide confidence interval maybe because the sample size is small

Relation of Confidence Intervals with Hypothesis Testing

- A general purpose approach to constructing confidence intervals is to define a 100*(1−α)% confidence interval to consist of all those values θ_0 for which a test of the hypothesis $\theta=\theta_0$ is not rejected at a significance level of 100α%.
  - ❑ Such an approach may not always be available since it presupposes the practical availability of an appropriate significance test.
  - ❑ Naturally, any assumptions required for the significance test would carry over to the confidence intervals.

Relation of Confidence Intervals with Hypothesis Testing

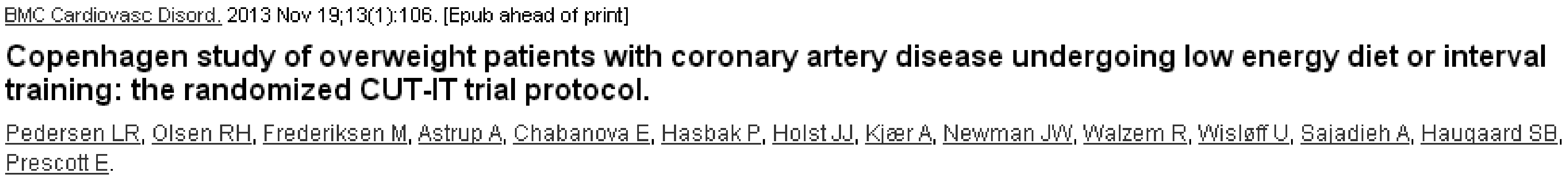
- It may be convenient to make the general correspondence that parameter values within a confidence interval are equivalent to those values that would not be rejected by an hypothesis test, but this would be dangerous.

- In many instances the confidence intervals that are quoted are only approximately valid, perhaps derived from "plus or minus twice the standard error", and the implications of this for the supposedly corresponding hypothesis tests are usually unknown.

HYPOTHESIS TESTING BY EXAMPLE

- **Null Hypothesis (H_0):**
  - The means of cholesterol level are not significantly different in case and control groups
  - There is no relationship between the weight and cholesterol level

- **Alternative hypothesis:**
  - The means of cholesterol level are significantly different in case and control groups
  - There is no relationship between the weight and cholesterol level



BMC Cardiovasc Disord. 2013 Nov 19;13(1):106. [Epub ahead of print]

Copenhagen study of overweight patients with coronary artery disease undergoing low energy diet or interval training: the randomized CUT-IT trial protocol.

Pedersen LR, Olsen RH, Frederiksen M, Astrup A, Chabanova E, Hasbak P, Holst JJ, Kjær A, Newman JW, Walzem R, Wisløff U, Sajadieh A, Haugaard SB, Prescott E.

The primary endpoint of the study is change in coronary flow reserve after the first 12 weeks' intervention.

The participants were consecutively enrolled during the inclusion period and randomized (1:1) into two groups:

1. 12 weeks of AIT (aerobic interval training) three times a week, followed by 40 weeks’AIT twice weekly.
2. 8–10 weeks’ LED (low energy diet) followed by 2–4 weeks of transition to a high protein/low glycemic index diet and 40 weeks of weight loss maintenance and AIT twice weekly.

Null Hypothesis: There is no significant difference in regards of coronary flow reserve between the investigated groups.

Alternative hypothesis: There is a significant difference in regards of coronary flow reserve between the investigated groups.

Hypothesis Testing by Example

Null and alternative hypotheses are either non-directional (two-tailed) or directional (one-tailed):

Non-directional (two-tailed):

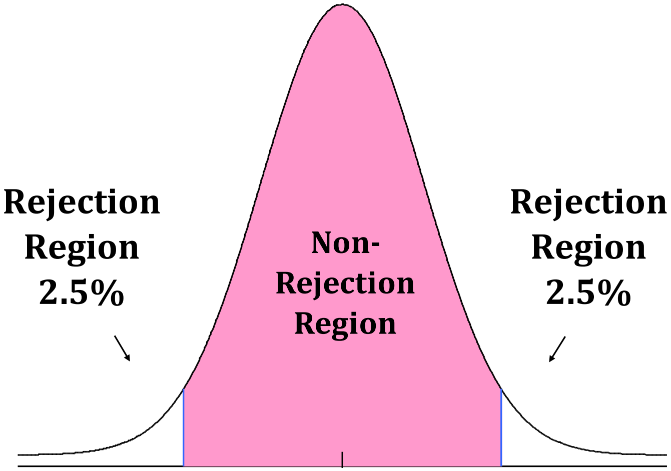
H_0: Coronary flow reserved AIT = Coronary flow reserved LED

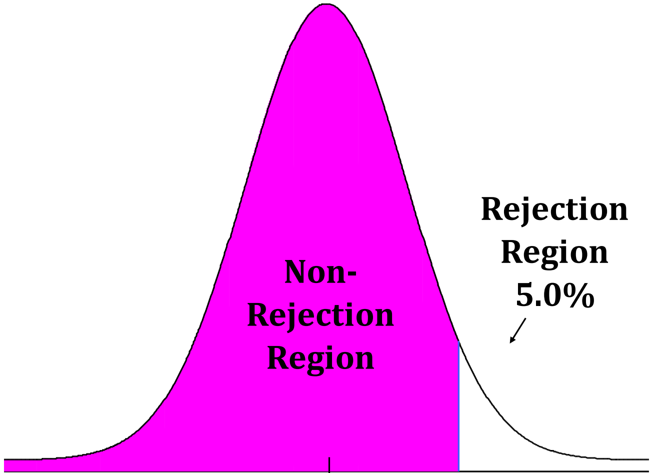
$H_{1/a}$: Coronary flow reserved AIT ≠ Coronary flow reserved LED

Directional (one-tailed):

H_0: Coronary flow reserved AIT ≤ Coronary flow reserved LED or H_0: Coronary flow reserved AIT ≥ Coronary flow reserved LED

$H_{1/a}$: Coronary flow reserved AIT > Coronary flow reserved LED or $H_{1/a}$: Coronary flow reserved AIT < Coronary flow reserved LED





HYPOTHESIS TESTING BY EXAMPLE

- Alpha (α) is the level of significance in hypothesis testing
- Alpha is a probability specified before the test is performed.
- **Alpha is the probability of rejecting the null hypothesis when it is true.**
- By convention, typical values of alpha specified in medical research are **0.05** and **0.01**.
- Alphas have corresponding **critical values**, the same ones used to calculate confidence intervals – 0.05/1.96, 0.01/2.575

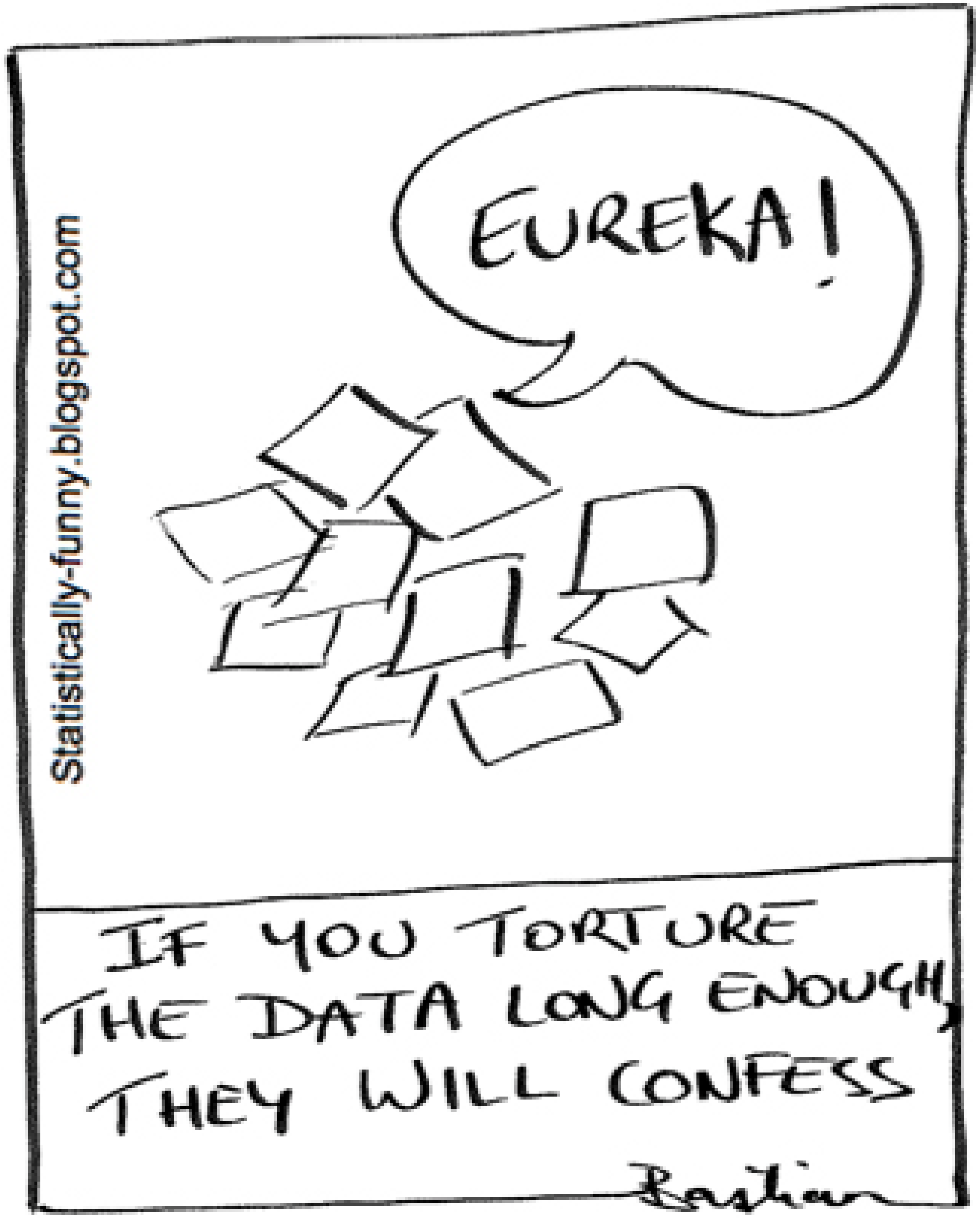
HYPOTHESIS TESTING BY EXAMPLE

- **Beta (β) is the probability of accepting the null hypothesis when it is false.**
- Typical values for beta are 0.10 to 0.20
- Beta is directly related to the power of a statistical test:
- Power is the probability of correctly rejecting the null hypothesis when it is false. Power = 1 - Beta
- A type II error occurs when a false null hypothesis is accepted.

P-VALUES

- Are the actual probabilities calculated from a statistical test, and are compared against alpha to determine whether to reject the null hypothesis or not.
- Example:
  - alpha = 0.05; calculated p-value = 0.008; reject null hypothesis
  - alpha = 0.05; calculated p-value = 0.110; do not reject null hypothesis / failed to reject the null hypothesis
- **A type I error occurs when a true null hypothesis is rejected.**

		True State of Nature	
		H_0 True	H_0 False
Findings	H_0 True	Correct	**Type II Error (β)**
	H_0 False	**Type I Error (α)**	Correct


EUREKA!
Statistically-funny.blogspot.com
IF YOU TORTURE THE DATA LONG ENOUGH, THEY WILL CONFESS
Bastian