Hypothesis Testing
Testing The Distribution Shape Of Continuous Data

Sorana D. Bolboacă

OBJECTIVES

- Testing hypothesis: General Approach
- Data follows a normal distribution?
  - Could be applied on any type of distribution (Binomial, Poisson, etc.)
- Two distributions has the same shape?
  - Could not answer to the question: “Which is the shape of the distribution?”
  - Tell us if the shapes of distributions of two samples are the same or are different

TESTING HYPOTHESIS

- 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

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

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 client 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 nott true in the population.

Statistical Test Frequently used in Medicine

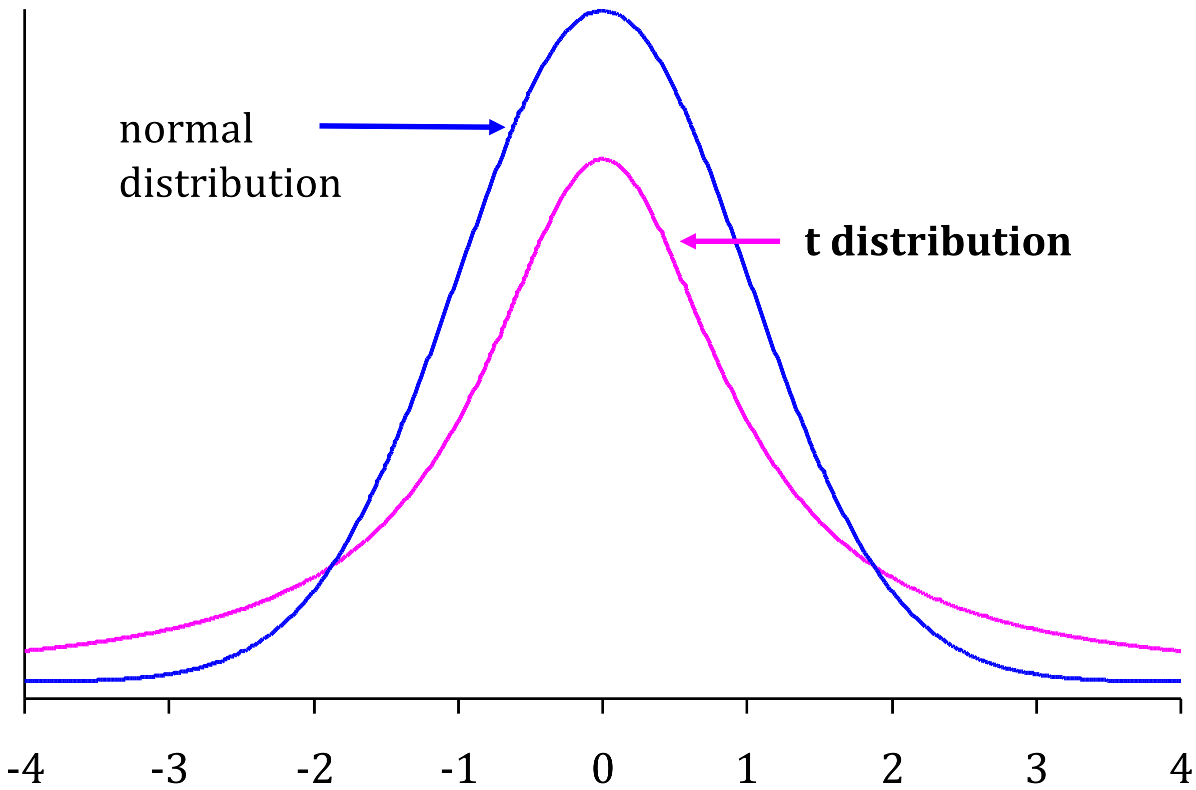
- Parametric tests (quantitative normal distributed data):
  - T-test for dependent or independent samples (2 groups)
  - ANOVA (2 or more groups)

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

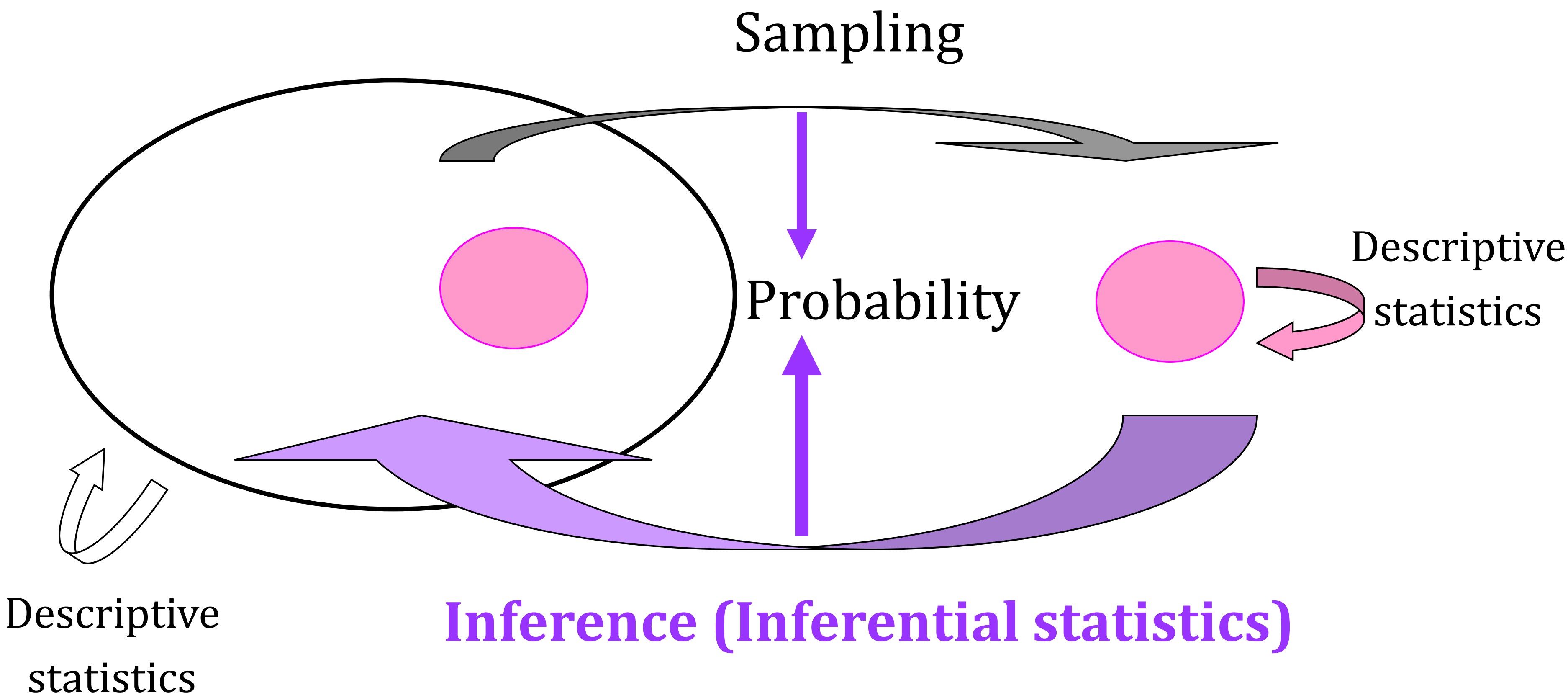
- Test for associations (quantitative & qualitative data):
  - Correlation (Pearson & Spearman) & Regression (Linear & Logistic)

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



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: μ=110 $H_{1/a}$: μ < 110 OR $H_{1/a}$: μ > 110	H_0: μ = 110 $H_{1/a}$: μ ≠ 110

HYPOTHESIS TESTING: STEP 2

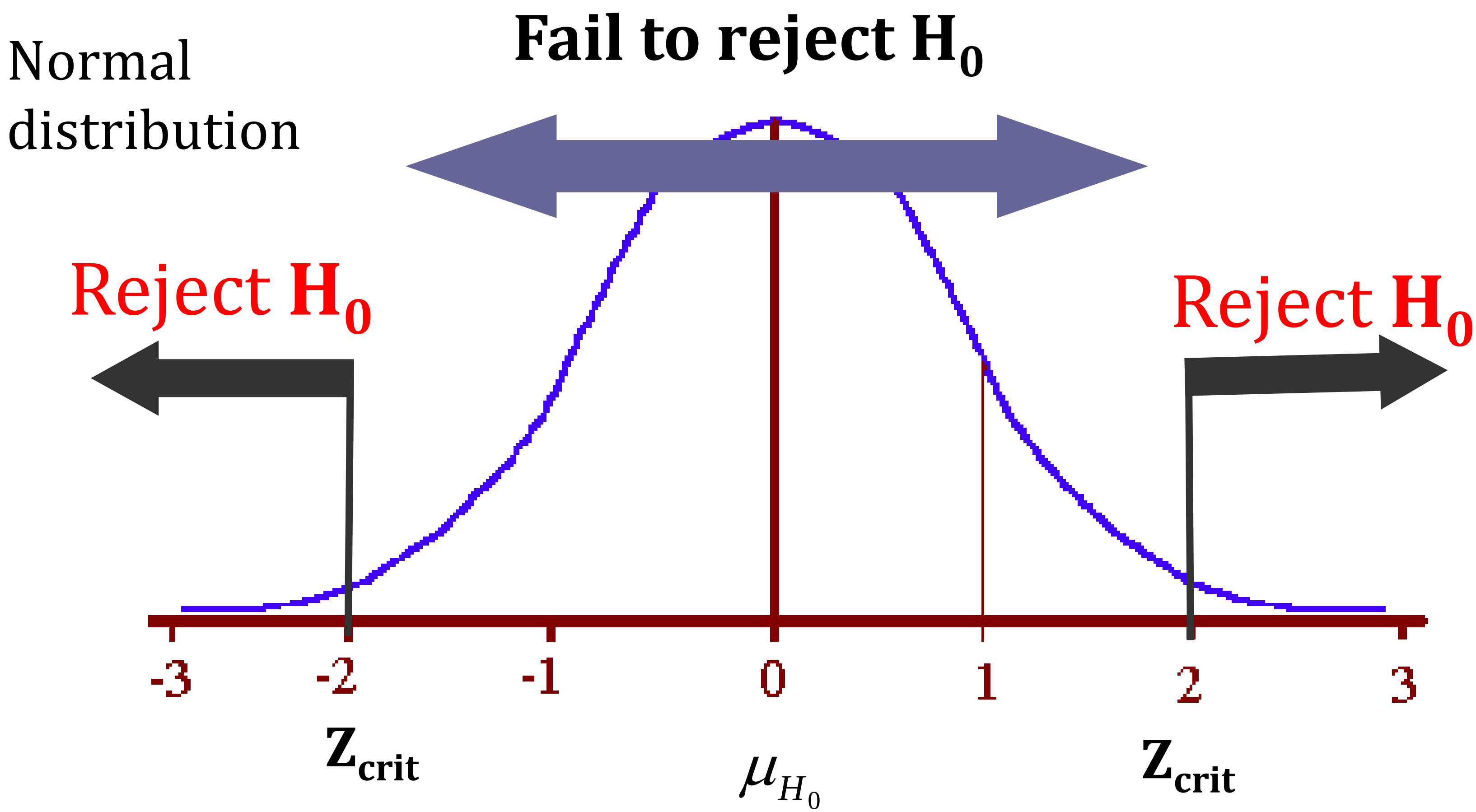
- Set decision criterion:
  - Decide what p-value would be "too unlikely"
  - This threshold is called the alpha level.
  - When a sample statistic surpasses this level, the result is said to be significant.
  - Typical **alpha levels** are **0.05** and **0.01**.
- Alpha levels (level of significance) = 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:
  - 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

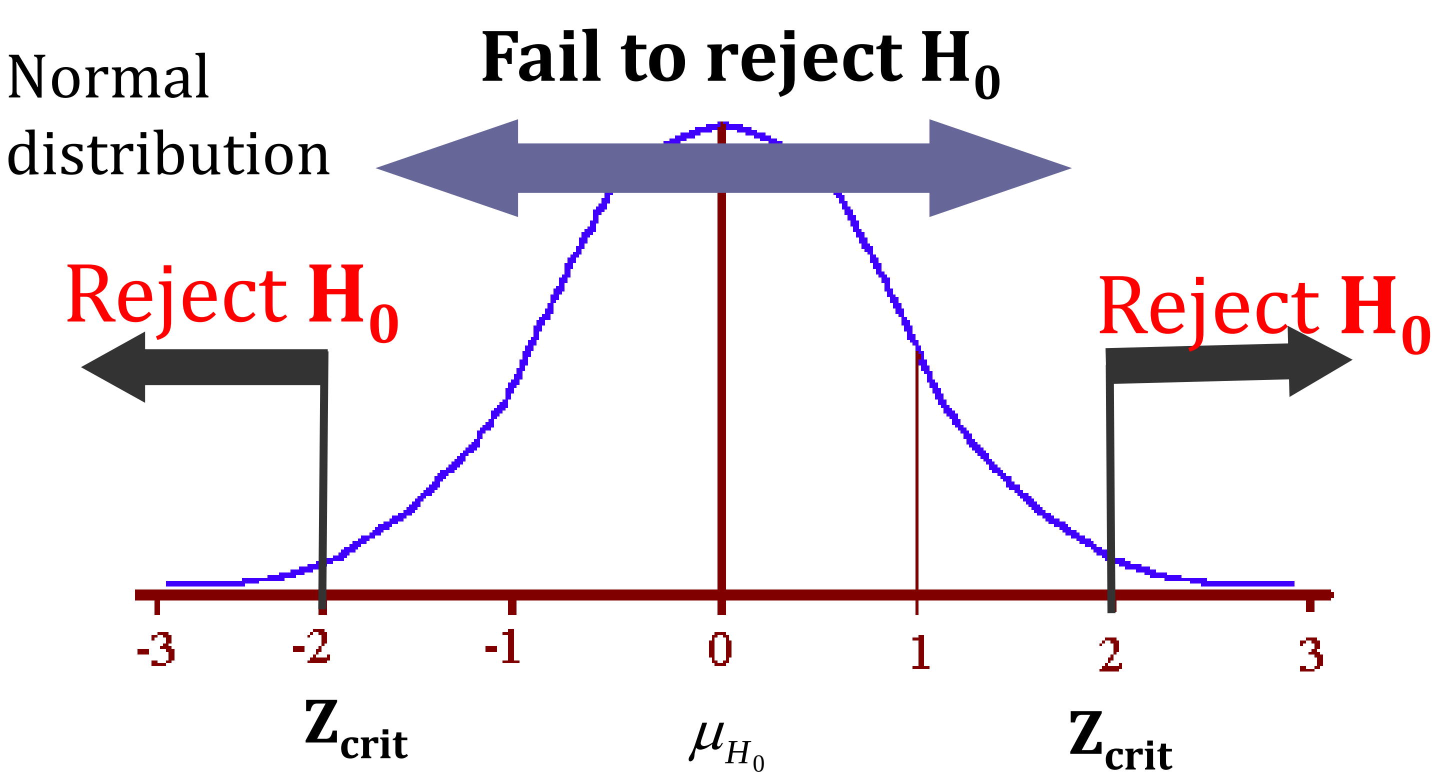


HYPOTHESIS TESTING: STEP 4

- Compute sample statistics
- A test statistic (e.g. Ztest, Ttest, or Ftest) 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 (α = 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

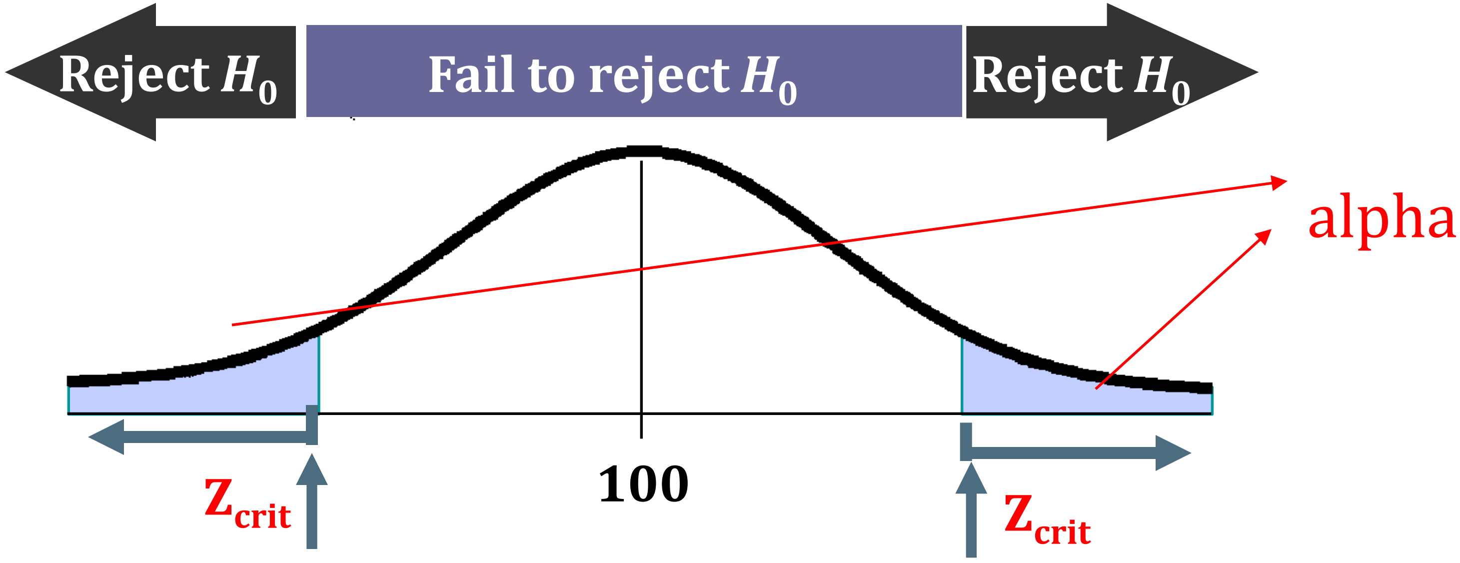
- H_a is that m is *either* greater or less than μ
  - H_a: $m \neq \mu$

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

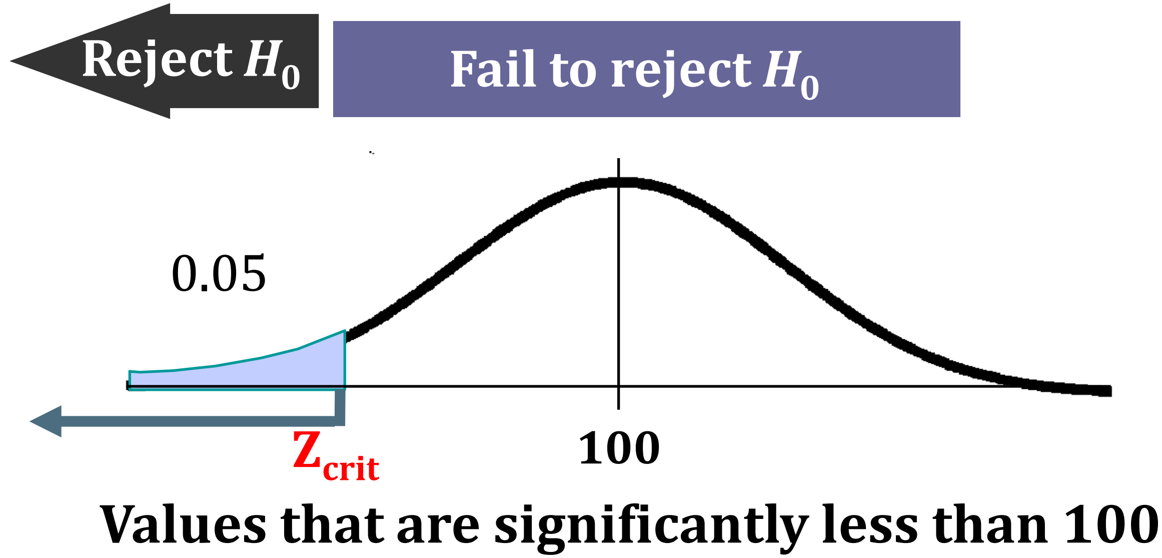
Two-Tailed Hypothesis Testing

H_0: μ = 100

H_1: μ ≠ 100 **Values that differ significantly from 100**

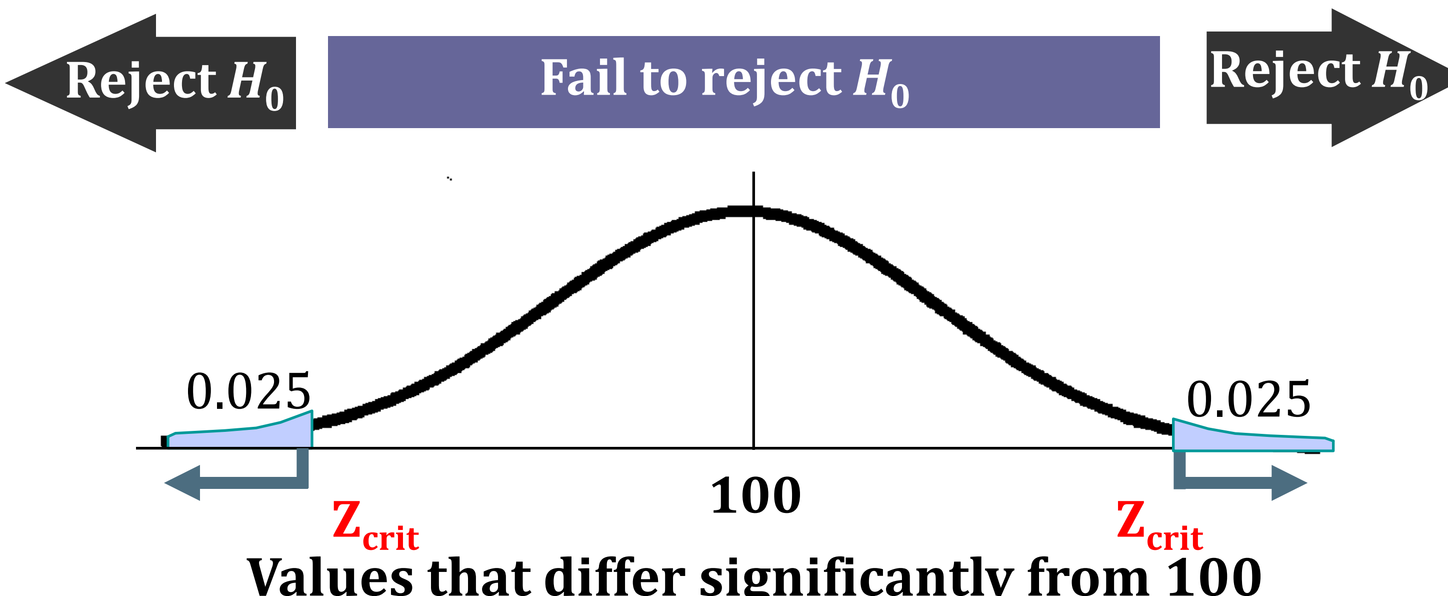


One tailed



Values that differ "significantly" from 100

Two tailed



Testing The Distribution Shape Of Continuous Data

- Test of Normality of a Distribution
- Test of Equality of Two Distributions

TEST OF NORMALITY OF A DISTRIBUTION

- Why to test the normality?
  - The robustness of most tests allows the test to be used if the underlying data are approximately normal
  - Test that required a normal distribution assumption:
    - Student t test
    - Z test
    - Tests of analysis of variances

NORMALITY TESTS

- Chi-Square goodness-of-fit: 1900
  - Conservative
- Kolmogorov-Smirnov (KS): 1933
  - Conservative
- Shapiro-Wilk: 1965

NORMALITY TESTS

	Prefer less conservative test	Prefer more conservative test
Small sample (5-50)	Shapiro-Wilk	Kolmogorov-Smirnov (one-sample form)
Medium to large sample (> 50)	Shapiro-Wilk	Chi-Square Goodness-of-Fit

SHAPIRO-WILK

- Seek a statistical software to use the Shapiro-Wilk test
- Limitation:
  - Some software packages allow only the sample parameters (means and standard deviation) to be used for the theoretical normal against which the data are being tested

KOLMOGOROV-SMIRNOV: ONE SAMPLE FORM

- Is the sample normal in shape without any a priori mean and standard deviation specified?
  - H_0: The form of the distribution is normal
  - Used the sample's mean and standard deviation
- Did the sample arise from a particular normal distribution with a postulated mean and standard deviation?
  - H_0:The distribution form which the sample was drawn is not different from a specified normal

KOLMOGOROV-SMIRNOV: ONE SAMPLE FORM

- The critical values for different significance levels:
  - $\alpha = 0.01$: $(1.63/\sqrt{n})-(1/3.5\cdot n)$
  - $\alpha = 0.05$: $(1.36/\sqrt{n})-(1/4.5\cdot n)$
  - $\alpha = 0.10$: $(1.22/\sqrt{n})-(1/5.5\cdot n)$

1. Arrange the n sample values in ascending order
2. Let *x* denote the sample value each time it changes:
   - If sample values are 1, 1, 2, 2, 3
     - First $x = 2$
     - Second $x = 3$

KOLMOGOROV-SMIRNOV: ONE SAMPLE FORM

3. Let k denote the number of sample members less than *x*
   - If sample values are 1, 1, 2, 2, 3, x = 2 and k = 2 (we have 2 values less than x)
4. Let $F_n(x)$ denote k/n for each x
   - = the sample cumulative frequency sum
5. The expected cumulative sum against which to test the sample
   - Calculated for each *x*
   - $z=(x-m)/s$

KOLMOGOROV-SMIRNOV: ONE SAMPLE FORM

6. For each *z* value, find the expected *Fe(x)* as the area under the normal distribution to the left to *z*
7. Calculate the absolute values of the difference between the F's | $F_n(x)$ - $F_e(n)$ |
8. Test statistic is the largest of these differences, say *L*.
9. If L exceeds the critical values, reject H_0; otherwise, do not reject H_0.

KOLMOGOROV-SMIRNOV: ONE SAMPLE FORM

Example:

- Do the ages of the 10 patients form a normal distribution?
- n = 10:
- $\alpha = 0.05$
  - Critical value: $(1.36/\sqrt{10})-(1/4.5\cdot 10) = 0.408$

Kolmogorov-Smirnov: one sample form

Age	x	k	$F_n(x)$	z	$F_e(x)$	$\lvert F_n(x)-F_e(x)\rvert$
54						
61	61	1	0.1	(61-65.1)/7=-0.586	0.279	0.179
61						
61						
62	62	4	0.4	(62-65.1)/7=-0.443	0.329	0.071
62						
68	68	6	0.6	(68-65.1)/7=0.414	0.661	0.061
73	73	7	0.7	(73-65.1)/7=1.129	0.871	0.171
74	74	8	0.8	(74-65.1)/7=1.271	0.898	0.098
75	75	9	0.9	(75-65.1)/7=1.414	0.921	0.021

m = (54+61+61+61+62+62+68+73+74+75)/10 = 651/10 = 65.1; s = 7

L = 0.171

$1.171 \leq 0.408 \rightarrow$ Do not reject H_0: The distribution from which the sample was drawn has not been shown to be different from normal.

CHI-SQUARED GOODNESS-OF-FIT

- H_0: the distribution from which the sample was drawn is normal (alternatively is normal with parameters μ and σ)
- H_1: the distribution is different.
- If calculated χ^2 is greater than critical χ^2, reject H_0; otherwise, accept H_o

$$\chi^2 = \sum \frac{(n_i - e_i)^2}{e_i} = \sum \frac{n_i^2}{e_i}$$

α (one tail)	0.10	0.05	0.025	0.01
df = 6	10.64	12.59	14.45	16.81
df = 8	13.36	15.51	17.53	20.09
df = 9	14.68	16.92	19.02	21.67

Chi-Squared goodness-of-fit

- $\chi^2 = 7.895 < 16.92$: Do not reject H_0

Interval	Standard normal z to end of interval	p	Expected frequencies (e_i)	Observed frequencies (n_i)
<50	-2.0691	0.0195	5.87	3
50-<55	-1.4591	0.0542	16.31	17
55-<60	-0.8346	0.1286	38.71	32
60-<65	-0.2173	0.2120	63.81	74
65-<70	0.4000	0.2407	72.45	62
70-<75	1.0173	0.1900	57.19	57
75-<80	1.6346	0.1035	31.15	37
80-<85	2.2519	0.0359	10.81	14
85-<90	2.8691	0.0145	4.36	4
≥90	∞	0.0019	0.57	1

Interpretation of Normality Tests using p-value

- If the p-value is less than the chosen alpha level, then the null hypothesis is rejected
  - One concludes the data are not from a normally distributed population
- If the p-value is greater than the chosen alpha level, then one does not reject the null hypothesis that the data came from a normally distributed population.

NORMALITY ANALYSIS BY EXAMPLE

Variable	Tests of Normality (BD) N	max D	K-S p	Lilliefors p	W	p
CHOLESTEROL	655	0.058620	p < .05	p < .01	0.985239	0.000004

One-Sample Kolmogorov-Smirnov Test

		GLYCEMIA
N		655
Normal Parameters[a,b]	Mean	121.62
	Std. Deviation	45.590
Most Extreme Differences	Absolute	.219
	Positive	.219
	Negative	-.166
Kolmogorov-Smirnov Z		5.618
Asymp. Sig. (2-tailed)		0.00E+00

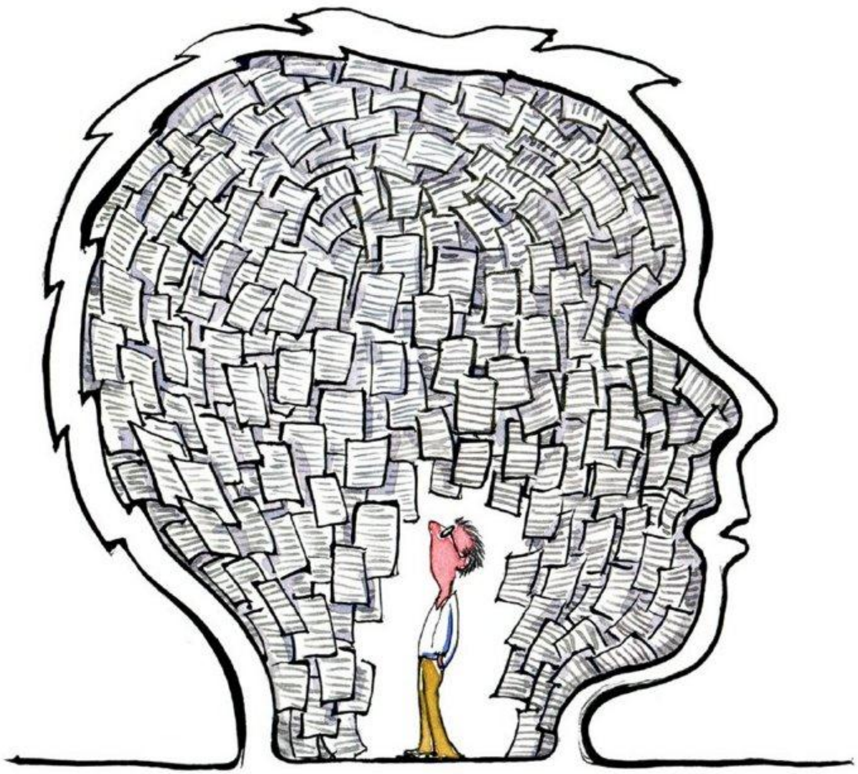
a. Test distribution is Normal.

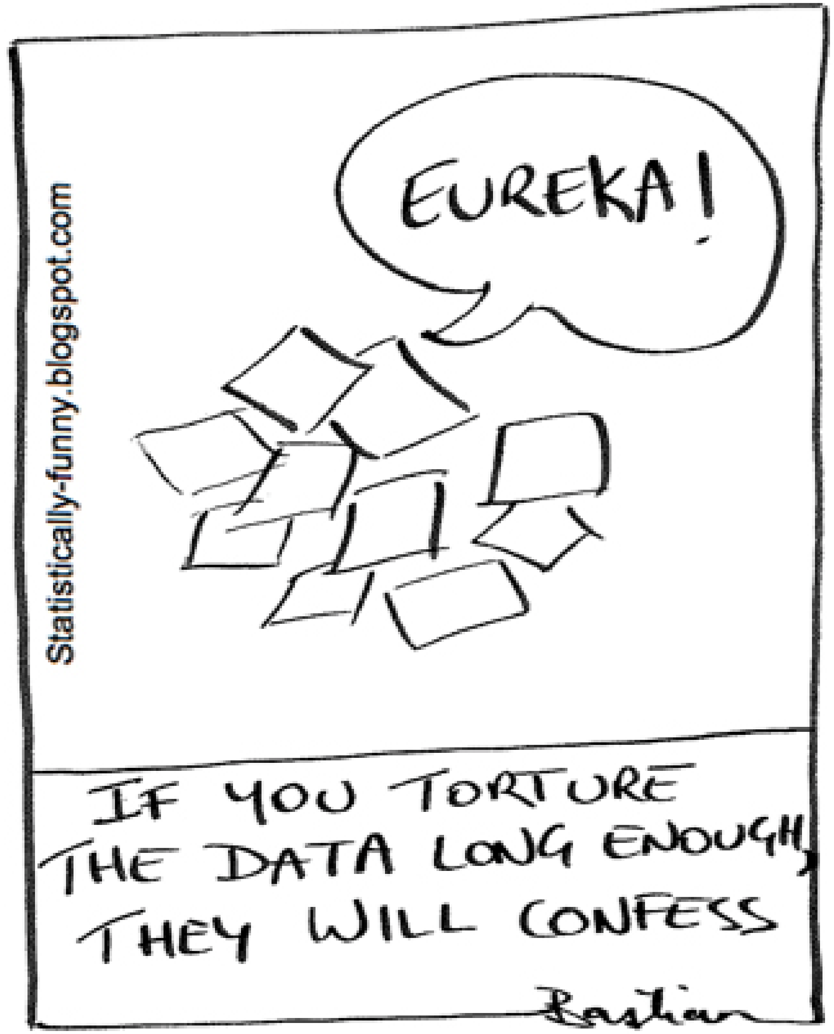
b. Calculated from data.

TESTING NORMALITY!

- The normal shape must be tested in order to be able to applied a series of statistical test
- Normality test could be performed by using statistical software
- It is necessary to know how to interpret a test (critical values vs p-value)

Thank you!




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