Continuous Probability Distributions
Point Estimators & Confidence Intervals
Hypothesis Testing

Sorana D. Bolboacă

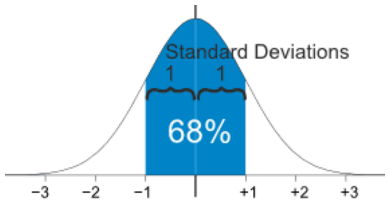
OBJECTIVES

- Continuous distributions: Normal & Student
- Point estimators & Confidence intervals for point estimators
- Testing hypothesis: General Approach

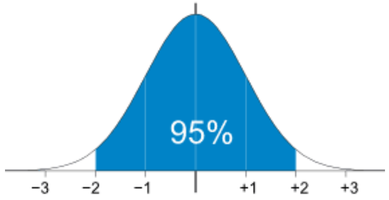
Continuous Probability Distributions

- Normal distribution and its standard form
  - Does your data follow a “bell shaped” pattern? (mean ~ median ~ mode)

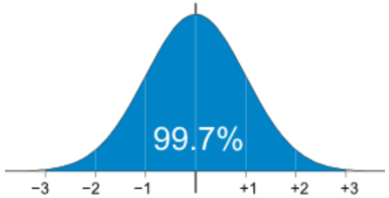
- Also known Gaussian distribution
- Characteristics of normal distribution:
  - ~ 68% of values fall between mean and one standard deviation (in either direction)



  - ~ 95% of values fall between mean and two standard deviations (in either direction)

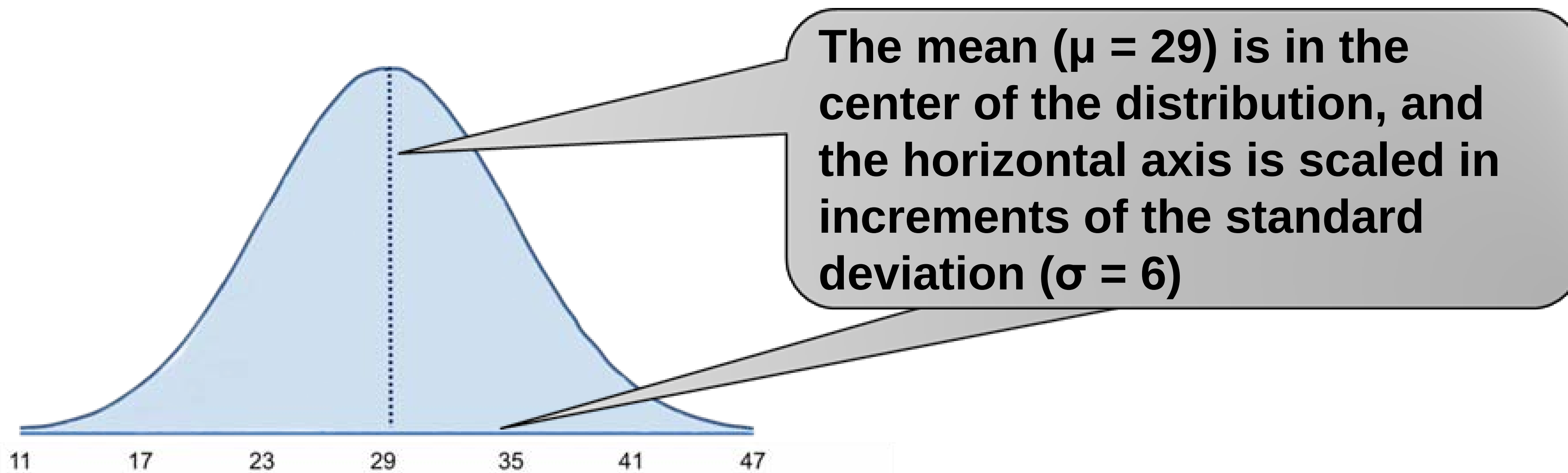


  - ~ 99.7% of values fall between mean and three standard deviations (in either direction)



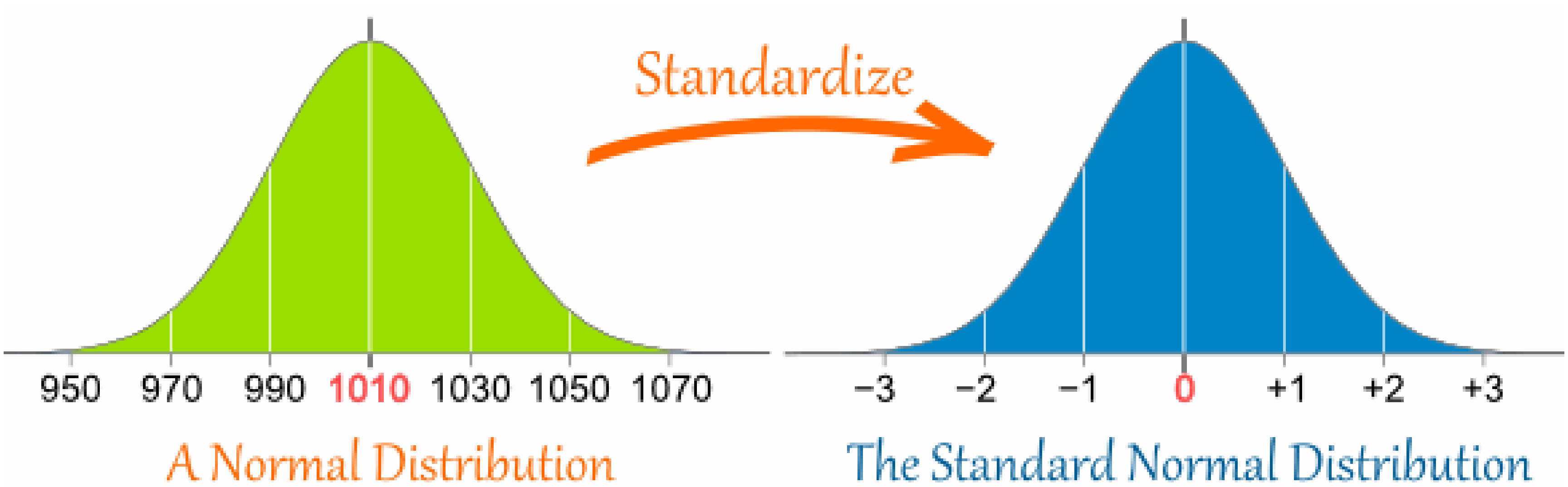
NORMAL DISTRIBUTION

- When we have a normal distributed variable and we know the **population mean (μ)** and **population standard deviation (σ)**, we can compute the probability of particular values using the following formula:
- $Pr(X) = 1/\sigma\sqrt{2\pi}\cdot e^{(-(X-\mu)^2/(2\sigma^2))}$



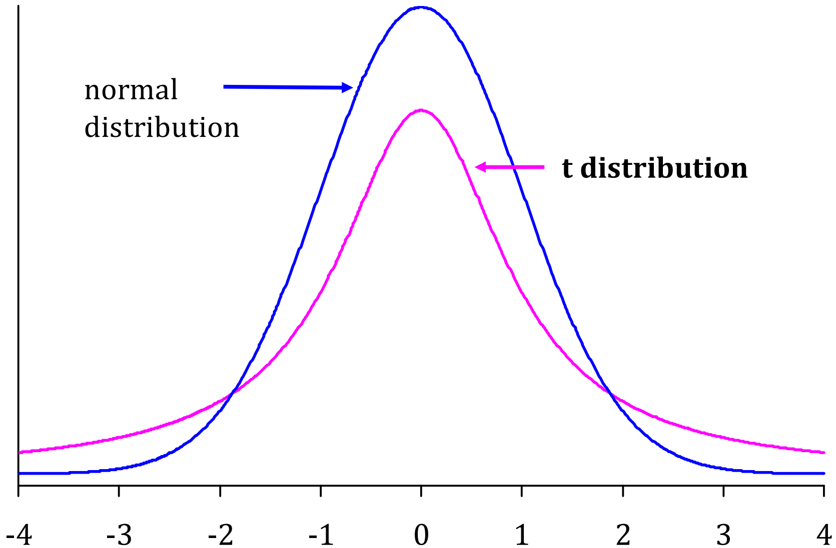
STANDARD NORMAL DISTRIBUTION

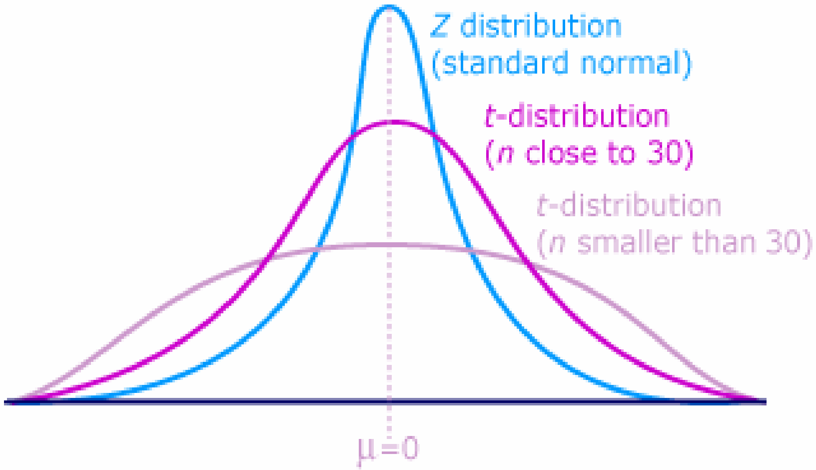
- The standard normal distribution is a normal distribution with a mean of zero and standard deviation of 1.



STUDENT T-DISTRIBUTION

- Student's t-distribution (or simply the t-distribution):
  - A member of a family of continuous probability distributions that arises when estimating the mean of a normally distributed population in situations where the sample size is small and population standard deviation is unknown.
  - Used by Student t-test, to construct confidence intervals, in linear regression analysis





Normal vs. binomial distribution

- What is the minimum required *n* for a binomial distribution with probability of success of 0.25 to closely follow a normal distribution?
- $n \times 0.25 \geq 10 \rightarrow n \geq 10/0.25 \rightarrow n \geq 40$
- $n \times 0.75 \geq 10 \rightarrow n \geq 10/0.75 \rightarrow n \geq 13.33$

Normal distribution

- A family doctor with ~ 3,000 subjects on the list measure over one year the heart rates (expected to be normal distributed). Three statistics were reported: mean = 75, minimum = 45, and maximum = 105. Which of the following is most likely to be the standard deviation of the distribution?

A. 2 $\rightarrow$ 75 ± 3×2 = (69; 81)

B. 5 $\rightarrow$ 75 ± 3×5 = (60; 90)

C. 10 $\rightarrow$ 75 ± 3×10 = (45; 105)

D. 12 $\rightarrow$ 75 ± 3×12 = (39; 111)

E. 15 $\rightarrow$ 75 ± 3×15 = (30; 120)

RECALL!

- Normal distribution
  - can be used to describe a variety of variables
  - Is bell-shaped, unimodal, symmetric, and continuous; its mean, median, and mode are equal
  - Its standard form has a mean of 0 and a standard deviation of 1
  - Can be used to approximate other distributions to simplify the analysis of data

POINT ESTIMATORS & CONFIDENCE INTERVALS

Inferential Statistics

- Inferential statistics = the process of making guesses about the truth on the population by examining a sample extracted from the population
- Sample statistics = summary measures calculated from data belonging to a sample (e.g. mean, proportion, ratio, correlation coefficient, etc.)
- Population parameter = true value in the population of interest
- Point estimation involves the use of sample data to calculate a single value (known as a statistic) which is to serve as a "best guess" or "best estimate" of an unknown (fixed or random) population parameter.

POINT ESTIMATOR

- Point estimation provide **one value as an estimate of the population parameter** (e.g. the sample mean is a point estimator for population mean)

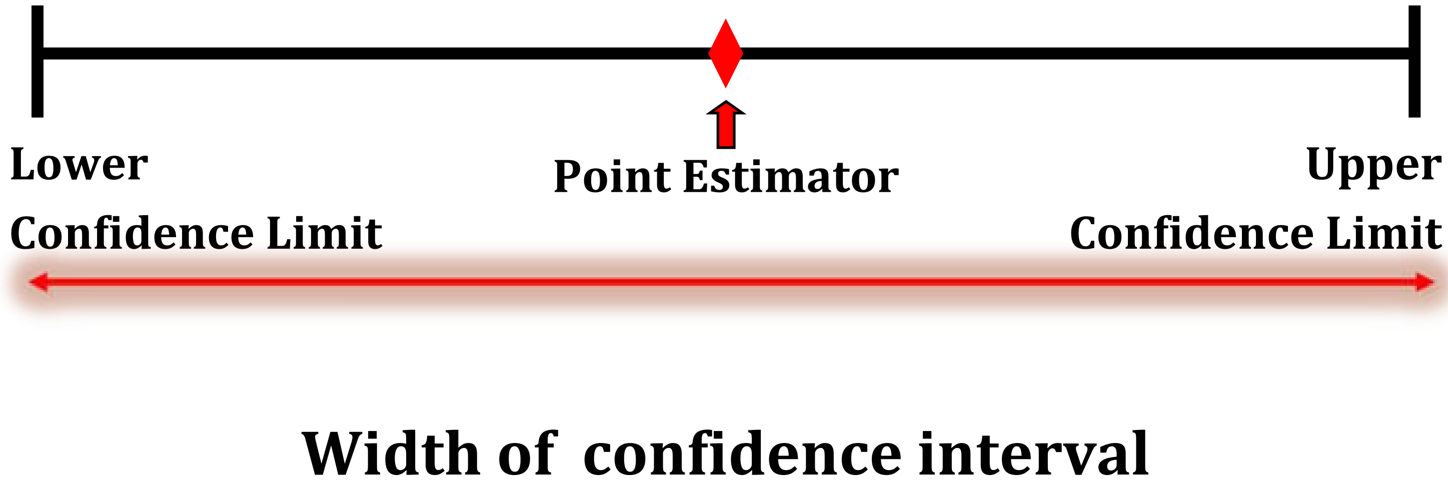
  - We are interested in the mean of height of 10-years-old boys and girls in the Romania. It would be impossible to measure the height of all 10-years-old boys and girls height so we will investigate a random sample of 30 boys and a random sample of 30 girls of 10-years-old. The sample mean for boys is 140 cm and for girls is 132 cm.
    - The sample mean of 140 cm is a point estimator of boys population mean
    - The sample mean of 132 cm is a point estimator of girls population mean

POINT ESTIMATOR VS. INTERVAL ESTIMATION

- Interval estimation: provide a range of values (an interval) that contain with a high probability the unknown parameter
- Confidence interval: the interval that contain an unknown parameter (such as the population mean) with certain degree of confidence
- It is recommended to estimate a theoretical parameter by using a range of value not a single value
  - It is called confidence interval
  - The estimated parameter belong to the confidence intervals with a high probability.

POINT ESTIMATOR VS. INTERVAL ESTIMATION

- Point estimator = one value obtained on a sample
  - How much uncertainty is associated with a point estimator of parameter?
- An interval provides more information about a population characteristics that does a point estimator confidence interval



INTERVAL ESTIMATION

$$\text{Point Estimator} \pm \underbrace{(\text{Critical Value}) \times (\text{Standard Error})}_{\textbf{Margin of error}}$$

- The margin of error, and hence the width of the interval, gets smaller the as the sample size increases.
- The margin of error, and hence the width of the interval, increases and decreases with the confidence.

INTERVAL ESTIMATION

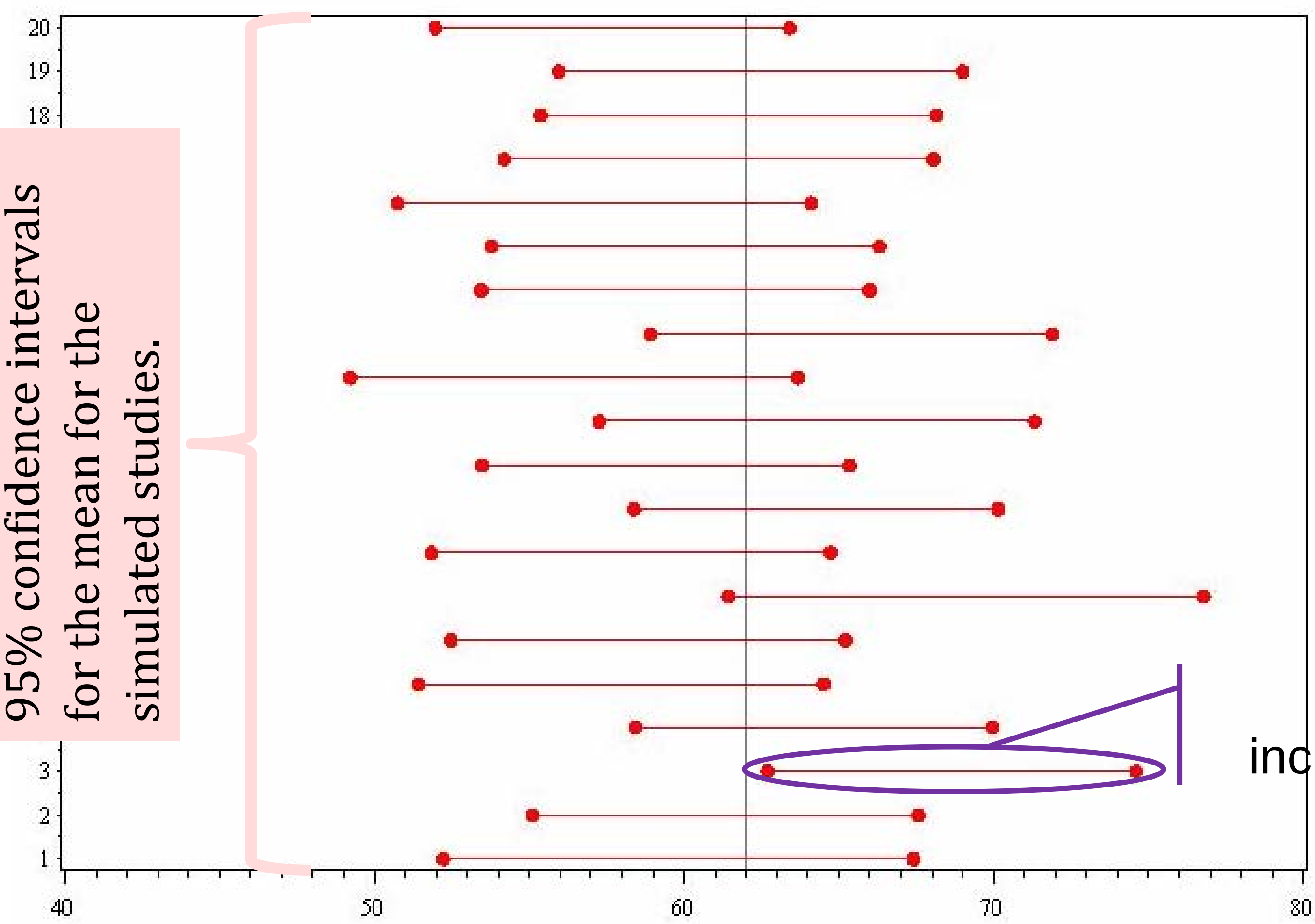
- Significance level α = 5% → 95% confidence interval (CI)
- CI = (1 - α) = 0.95

- Interpretation:
  - If all possible samples of size *n* are extracted from the population and their means and intervals are estimated, 95% of all the intervals will include the **true value of the unknown parameter**
  - A specific interval either will contain or will not contain the true parameter (due to the 5% risk)

INTERVAL ESTIMATION

True mean (62)

95% confidence intervals for the mean for the simulated studies.



This CI did not include the true value

CONFIDENCE INTERVALS

- Provides:
  - A plausible range of values for a population parameter.
  - The precision of an point estimator.
    - When sampling variability is high, the confidence interval will be wide to reflect the uncertainty of the observation.
  - Statistical significance.
    - If the 95% CI does not cross the null value, it is significant at 0.05.

Confidence Intervals

- Are calculated taking into consideration:
  - The sample or population size
  - The type of investigated variable (qualitative OR quantitative)

- Formula of calculus comprised two parts:
  - One estimator of the quality of sample based on which the population estimator was computed (standard error)
    - Standard error: is a measure of how good our best guess is.
      - Standard error: the bigger the sample, the smaller the standard error.
      - Standard error: is always smaller than the standard deviation
  - Degree of confidence (standard values)

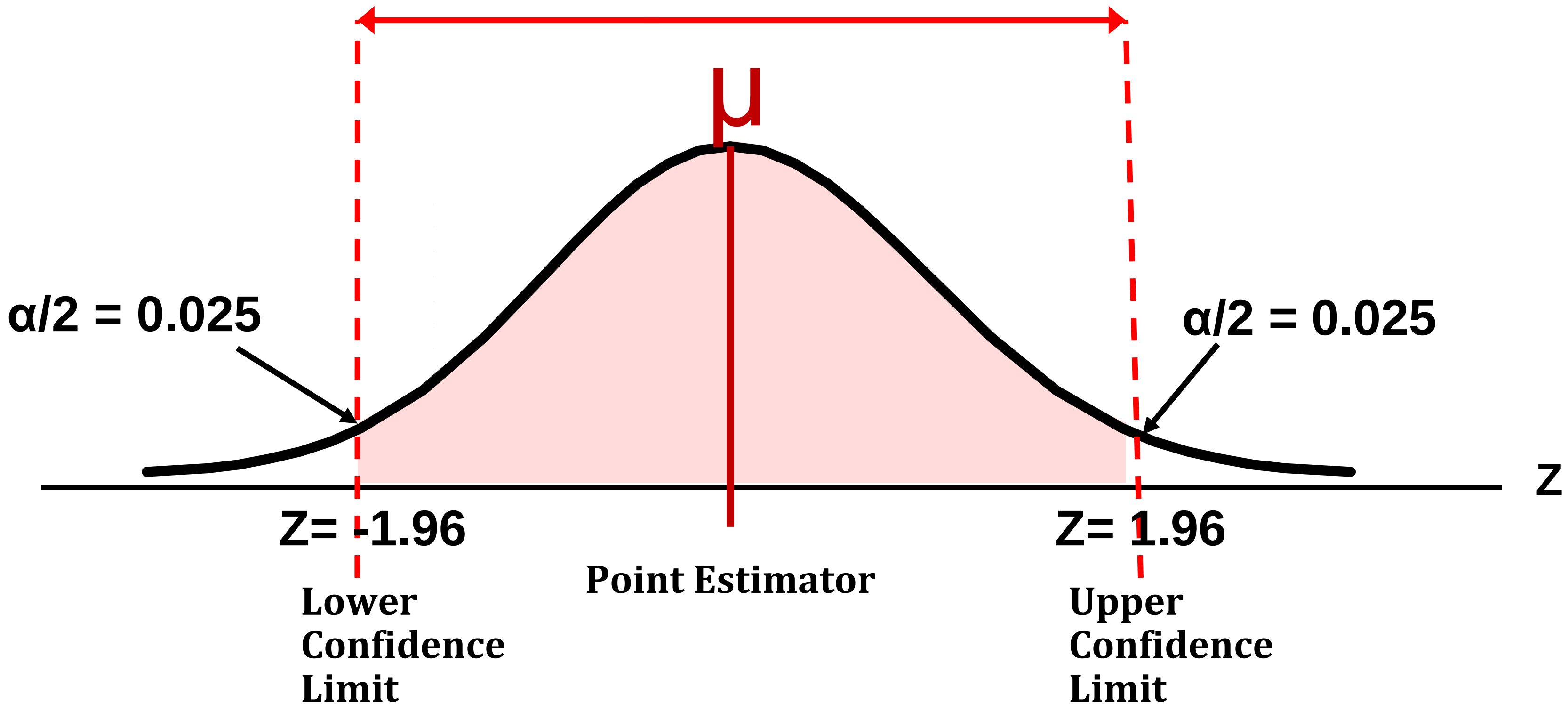
CONFIDENCE INTERVALS FOR MEANS

- Assumptions:
  - Population standard deviation (σ) is known
  - Population is normally distributed

$$\left[\overline{X} - Z_{\alpha}\frac{\sigma}{\sqrt{n}};\overline{X} + Z_{\alpha}\frac{\sigma}{\sqrt{n}}\right]$$

where Z is the normal distribution's critical value for a probability of α/2 in each tail

- Consider a 95% confidence interval:
- $1-\alpha = 0.95$ & $\alpha = 0.05$ & $\alpha/2 = 0.025$



CONFIDENCE INTERVALS FOR MEANS

- Consider the distribution of serum cholesterol levels for all female Romanian who are hypertensive and overweight. This population has an unknown mean (μ) and a standard deviation (σ) of 30 mg/dl. We extracted from this population a random sample of 20 subjects and we found a mean of serum cholesterol level ($\bar{X}$) equal with 220 mg/dl.
  - $\bar{X}$ = 220 mg/dl is a point estimator of the unknown mean serum cholesterol level (μ) in the population
  - Because of the sampling variability, it is important to construct the interval able to take into account the sampling variability:

$$95\%CI = \left(220 - 1.96\frac{30}{\sqrt{20}}, 220 + 1.96\frac{30}{\sqrt{20}}\right) = (207, 233)$$

Length = 233-207 = 26

$$99\%CI = \left(220 - 2.58\frac{30}{\sqrt{20}}, 220 + 2.58\frac{30}{\sqrt{20}}\right) = (203, 237)$$

Length = 237-203 = 34

CONFIDENCE INTERVALS FOR MEANS

- Unknown population mean (μ) & unknown population standard deviation (σ): student t-distribution with n-1 degree of freedom will be used

$$P\left(\overline{X} - t_{\alpha/2}\frac{s}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right)=0.95$$

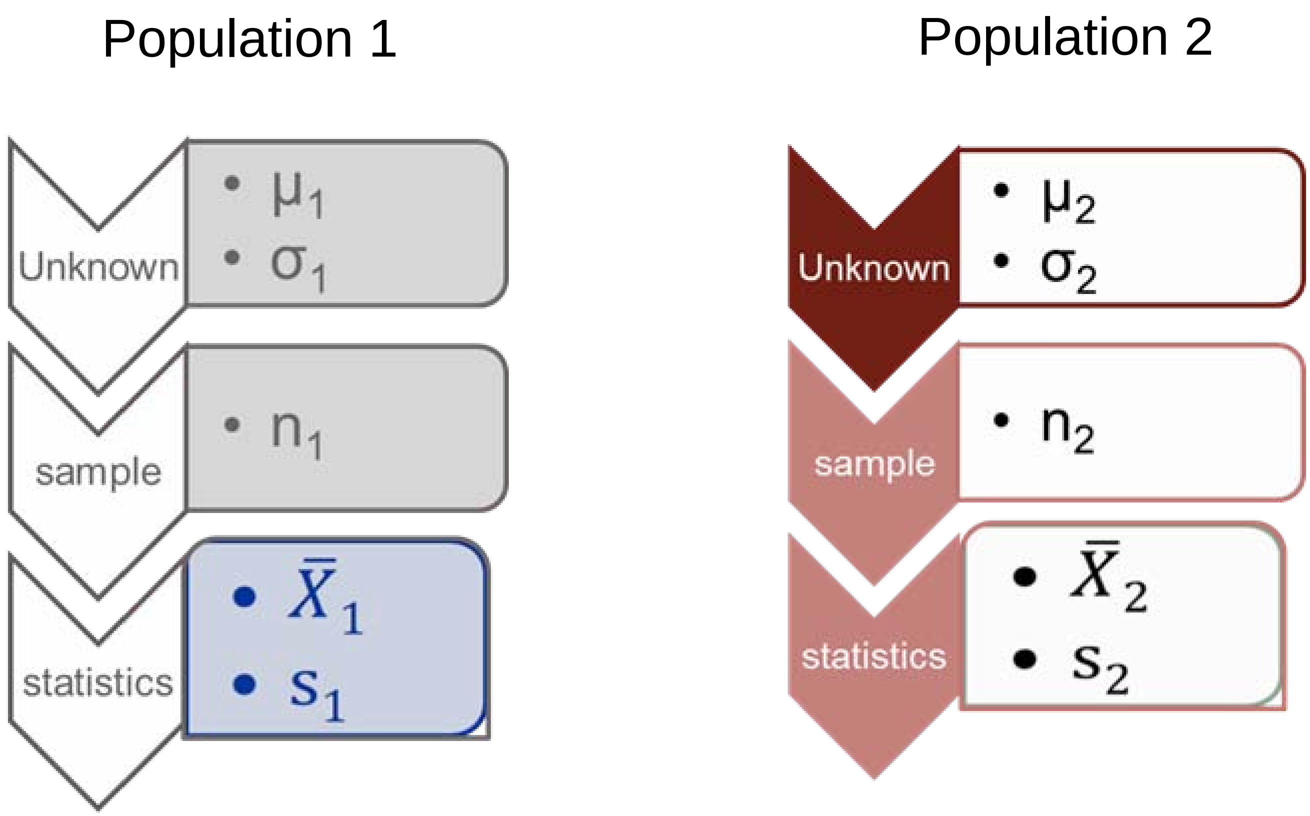
- A sample of 20 female students gave a mean weight of 60kg and a standard deviation of 8 kg. Assuming normality, find the 90, 95, and 99 percent confidence intervals for the population mean weight.

$$90\%CI = \left[60 - 1.73\frac{8}{\sqrt{20}}, 60 + 1.73\frac{8}{\sqrt{20}}\right] = [56.91, 63.09]$$

$$95\%CI = \left[60 - 2.09\frac{8}{\sqrt{20}}, 60 + 2.09\frac{8}{\sqrt{20}}\right] = [56.26, 63.74]$$

$$99\%CI = \left[60 - 2.86\frac{8}{\sqrt{20}}, 60 + 2.86\frac{8}{\sqrt{20}}\right] = [54.88, 65.12]$$

Confidence Intervals for Means Difference



Estimate $(\mu_1 - \mu_2)$ with $\bar{X}_1 - \bar{X}_2$

CONFIDENCE INTERVALS FOR MEANS DIFFERENCE

$$(\bar{X}_1 - \bar{X}_1) \pm t_{df}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_{12}}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

Group 1	7	7	8	8	8	6	9	6	5
Group 2	8	10	9	6	10	8	9	7	8

	Group 1	Group 2
Mean	7.11	8.33
s	1.27	1.32
s^2	1.61	1.75

df=15.97

for α = 0.05 $t_{15.97} = 2.13$

$$(7.11 - 8.33) \pm 2.13\sqrt{0.18 + 0.19}$$

-1.22 ± 2.13*0.61

-1.22 ± 1.30 [-2.52, 0.08]

CONFIDENCE INTERVALS

- Interpretation of CI for difference between two means
  - If 0 is contains by the confidence intervals, there is no significant difference between means.
  - If 0 is NOT contains by the confidence intervals, there is a significant difference between means.

COMPARING MEANS USING CONFIDENCE INTERVALS

http://www.biomedcentral.com/content/pdf/1471-2458-12-1013.pdf

Table 1 Living conditions of the MS-MV and the immigrant population (CASEN survey 2006)

	IMMIGRANT POPULATION 1% total sample, n = 154 431 weighted population (1877 real observations)		**MS-MV GROUP 0.67% total sample, n = 108 599 weighted population (1477 real observations)**	
	% or mean	**95% CI**	**% or mean**	**95% CI**
DEMOGRAPHICS				
Mean age**	X = 33.41	31.81–35.00	X = 26.13	23.41–28.26
Age categories:				
<16 years old**	13.60	11.29–16.28	45.25	39.53–51.10
16-65 years old**	79.08	75.92–81.93	47.26	41.64–52.94
>65 years old	7.32	5.33–9.97	7.49	5.31–10.46
Sex (female = 1)	45.21	41.74–48.72	51.27	47.99–55.41
Marital status:				
Single**	45.81	42.06–49.62	64.30	59.36–68.95
Married**	45.49	41.66–49.36	29.39	25.09–34.10

CONFIDENCE INTERVAL FOR FREQUENCY

$$\left[f - Z_{\alpha}\sqrt{\frac{f(1-f)}{n}};\ f + Z_{\alpha}\sqrt{\frac{f(1-f)}{n}} \right]$$

- Could be computed if:
  - n×f > 10, where n = sample size, f = frequency

- We are interested in estimating the frequency of breast cancer in women between 50 and 54 years with positive family history. In a randomized trial involving 10,000 women with positive history of breast cancer were found 400 women diagnosed with breast cancer.
- What is the 95% confidence interval associated frequently observed?

- f = 400/10000 = 0.04

$$\left[0.04 - 1.96\sqrt{\frac{0.04 \cdot 0.96}{10000}};\ 0.04 + 1.96\sqrt{\frac{0.04 \cdot 0.96}{10000}} \right]$$

- [0.04-0.004; 0.04+0.004]
- [0.036; 0.044]

CONFIDENCE INTERVALS FOR OTHER ESTIMATORS

http://www.biomedcentral.com/content/pdf/1471-2458-12-1013.pdf

Table 3 Odds Ratio (OR) of presenting **any disability and any chronic condition or cancer**, adjusted by different sets of factors separately (CASEN survey 2006)

	ANY DISABILITY				**ANY CHRONIC CONDITION OR CANCER**			
	International immigrants		**MS-MV**		**International immigrants**		**MS-MV**	
	OR	95% CI	OR	95% CI	OR	95% CI	OR	95% CI
DEMOGRAPHICS								
Age	**1.04***	**1.02-1.06**	**1.04***	**1.02-1.06**	**1.05***	**1.02-1.08**	**1.02***	**1.01-1.04**
Sex (female = 1)	0.56	0.25-1.25	**0.39***	**0.20-0.75**	**2.78****	**1.26-6.71**	1.05	0.46-2.36

FORMULAS!

- One mean: population standard deviation is known

$$\left[\overline{X} - Z_{\alpha}\frac{\sigma}{\sqrt{n}}; \overline{X} + Z_{\alpha}\frac{\sigma}{\sqrt{n}}\right]$$

- One mean: population standard deviation is not known

$$\left[m - t_{n-1, 1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}; m + t_{n-1, 1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right]$$

- Difference between two means

$$\left((m_1 - m_2) - t_{critical} \times SE; (m_1 - m_2) + t_{critical} \times SE\right)$$

$$SE(m_1 - m_2) = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1){S_2}^2}{n_1 + n_2 - 2}}$$

FORMULAS!

- One frequency

$$\left[f - Z_{\alpha}\sqrt{\frac{f\left(1-f\right)}{n}}; f + Z_{\alpha}\sqrt{\frac{f\left(1-f\right)}{n}} \right]$$

- Difference between two frequencies

$(f_1-f_2) \pm Z_{critical} \times SE$

$SE = sqrt((f_1*(1-f_1)/n_1)+(f_2*(1-f_2)/n_2))$

RECALL!

- Correct estimation of a population parameter is done with confidence intervals.
- Confidence intervals depend by the sample size and standard error.
- The confidence intervals is larger for:
  - High value of standard error
  - Small sample sizes

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

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.]

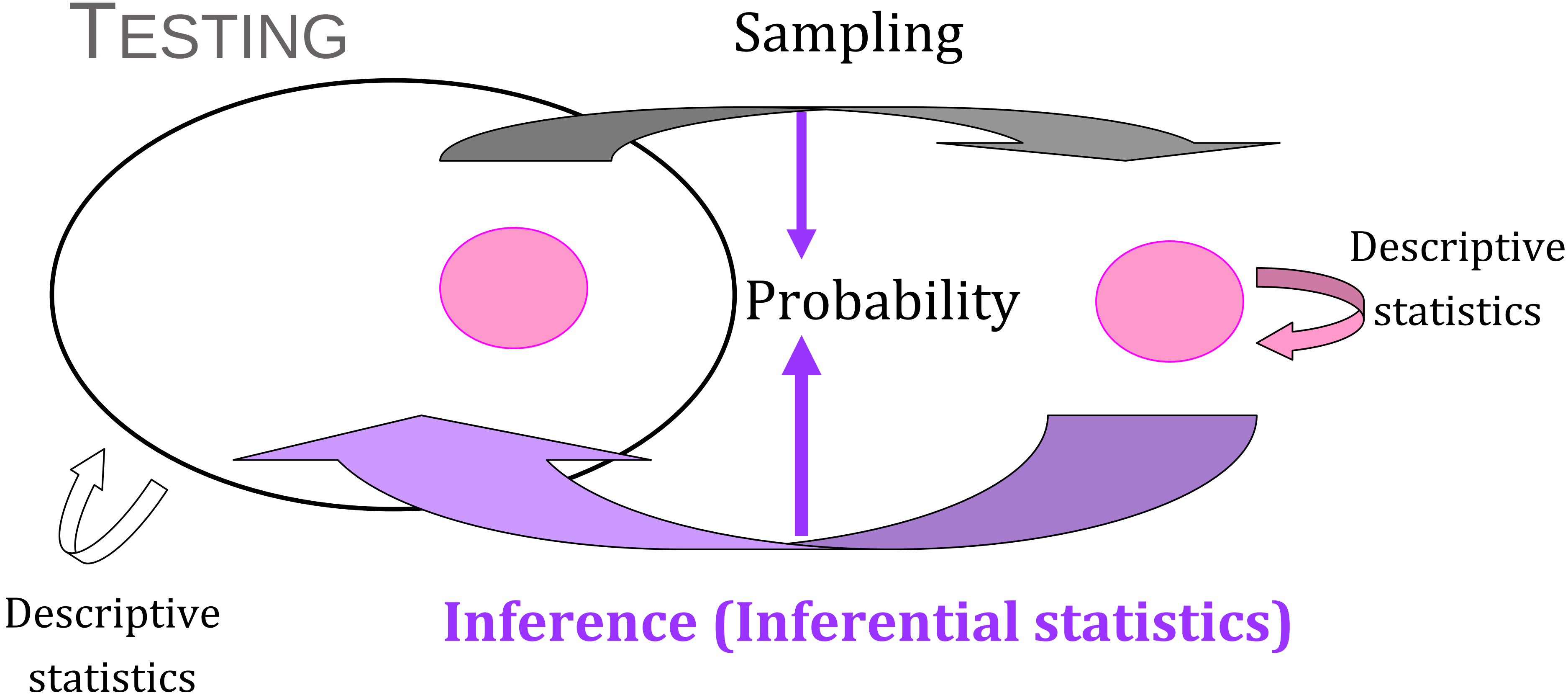
Statistical Test Frequently used in Medicine

- 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)

FROM PROBABILITY TO HYPOTHESIS TESTING



Population

Parameters (μ, σ)

Sample

Statistics (m, s)

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 (e.g. Z_{test}) and get a p-value

Step 5: Make a decision

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

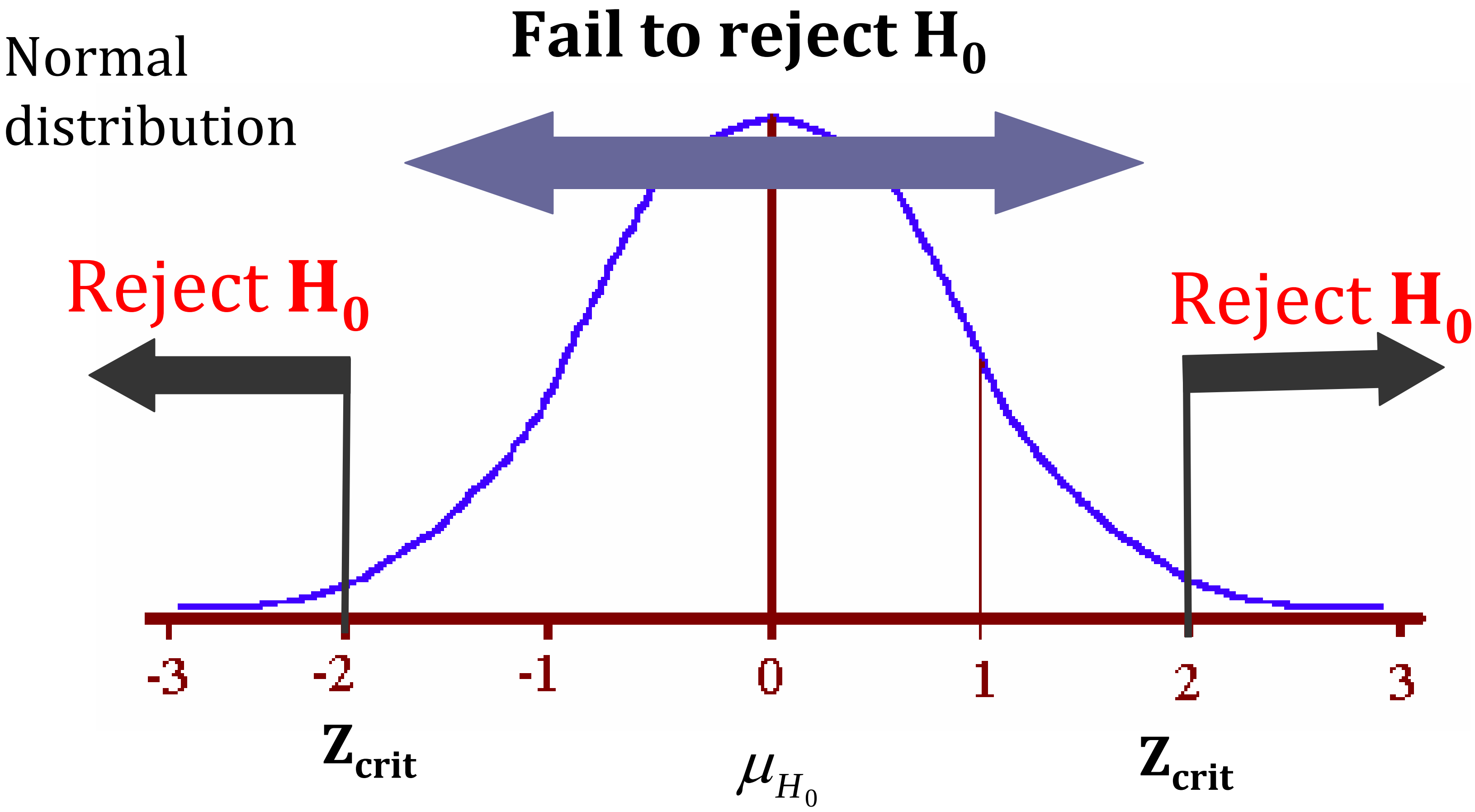
- **Hypotheses are in terms of population parameters!!!**

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

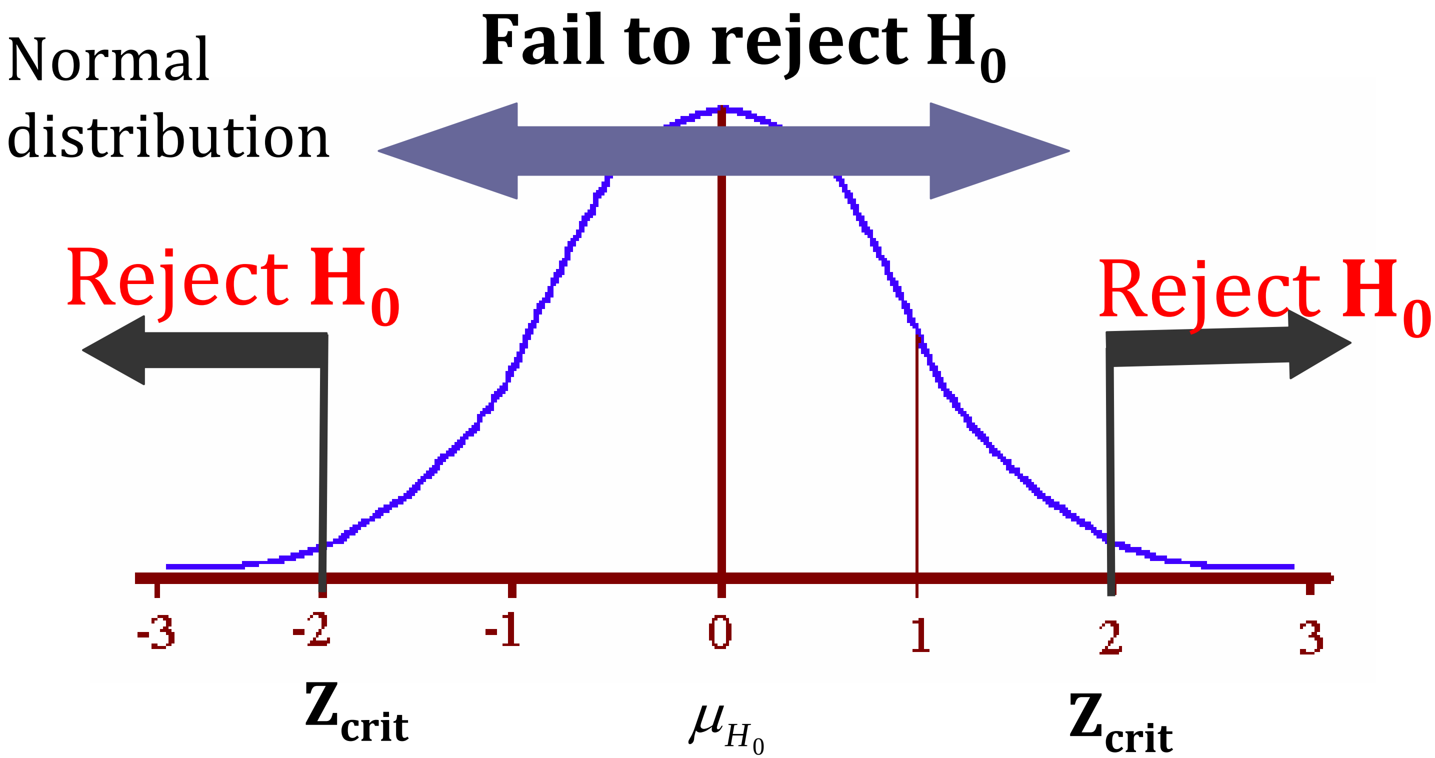
HYPOTHESIS TESTING: STEP 2

- Set decision criterion:
  - Decide what p-value would be “too unlikely”
  - This threshold is called the alpha level.
  - 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



HYPOTHESIS TESTING: STEP 4

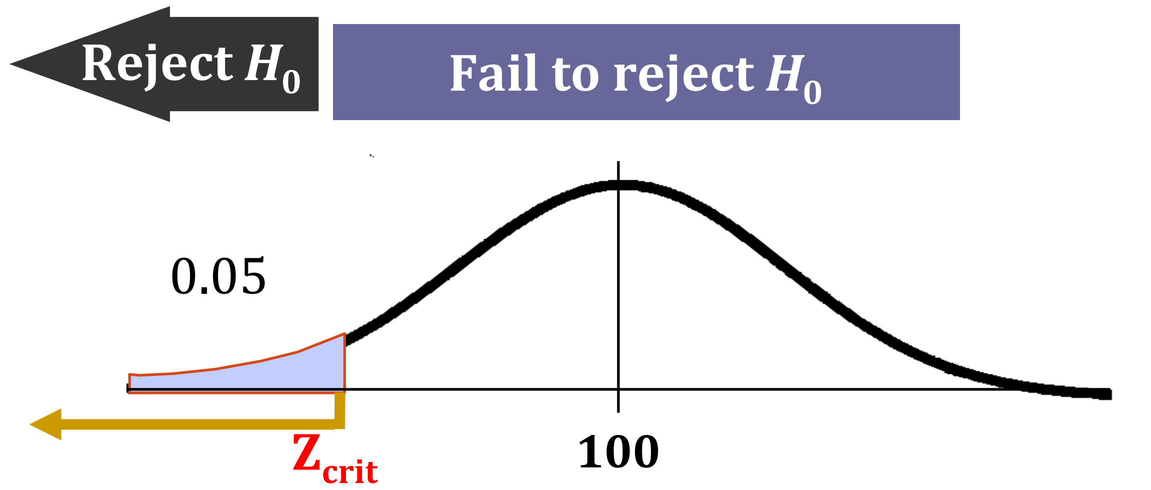


- If we want to know where our sample mean lies in the null distribution, we convert X-bar to our test statistic Z_{test}
- If an observed sample mean were lower than z=-1.65 then it would be in a critical region where it was more extreme than 95% of all sample means that might be drawn from that population

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**.

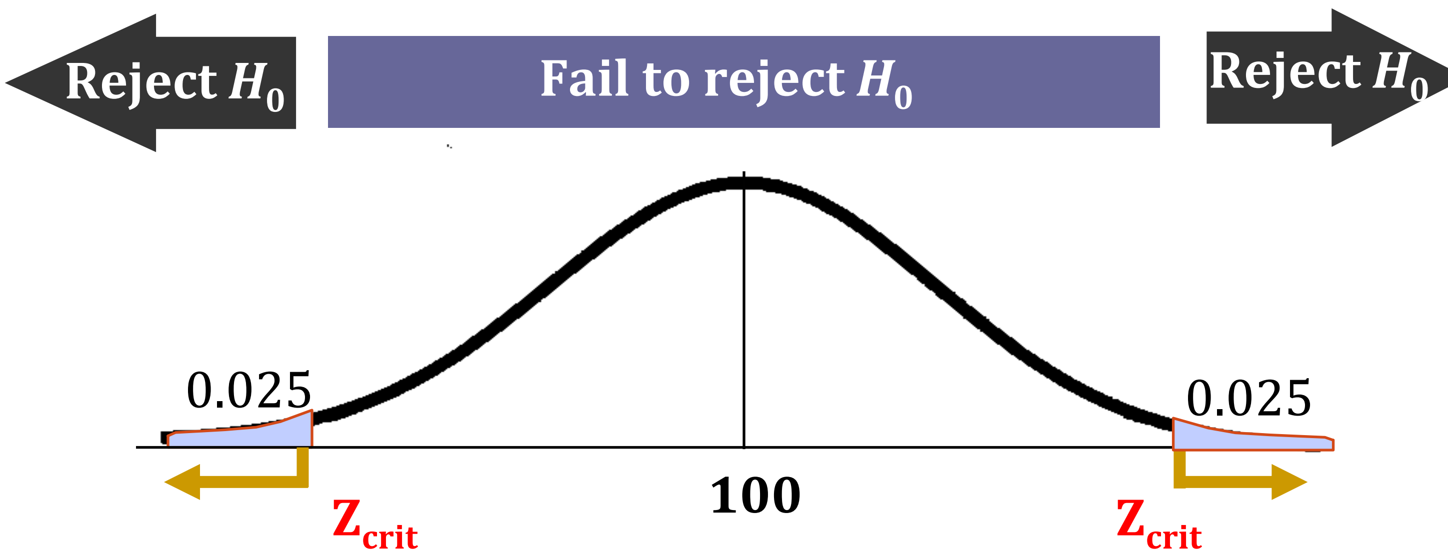
One tailed



Values that differ "significantly" from 100

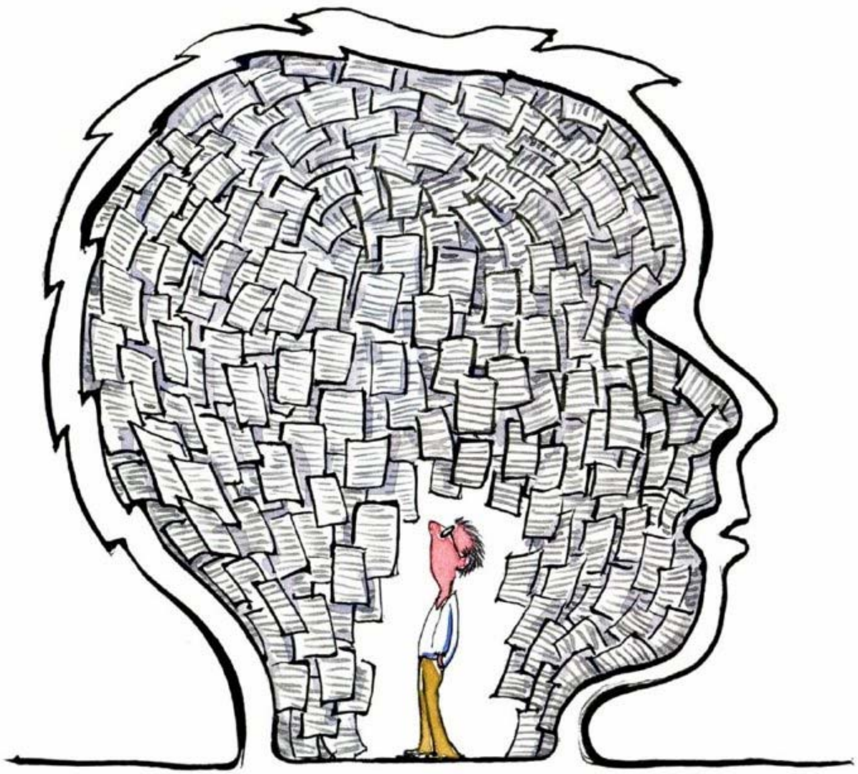
Values that are significantly less than 100

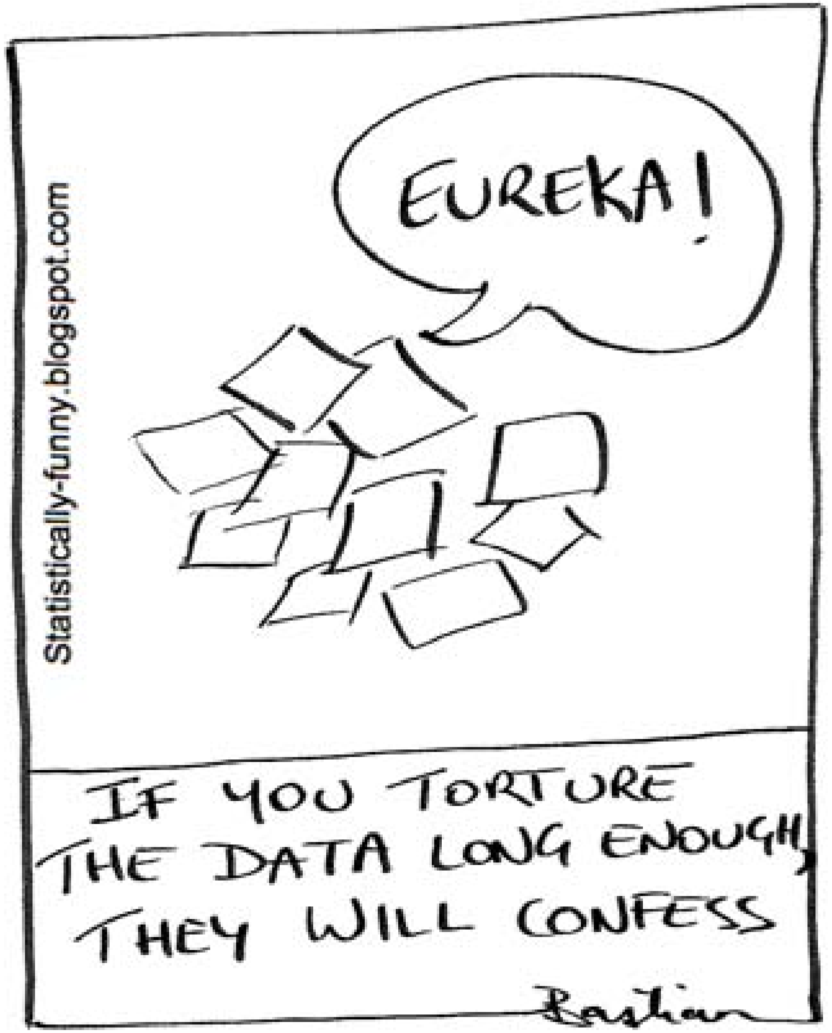
Two tailed



Values that differ significantly from 100

Thank you!




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