Biostatistics
Week IV

Ege Ülgen, MD, PhD

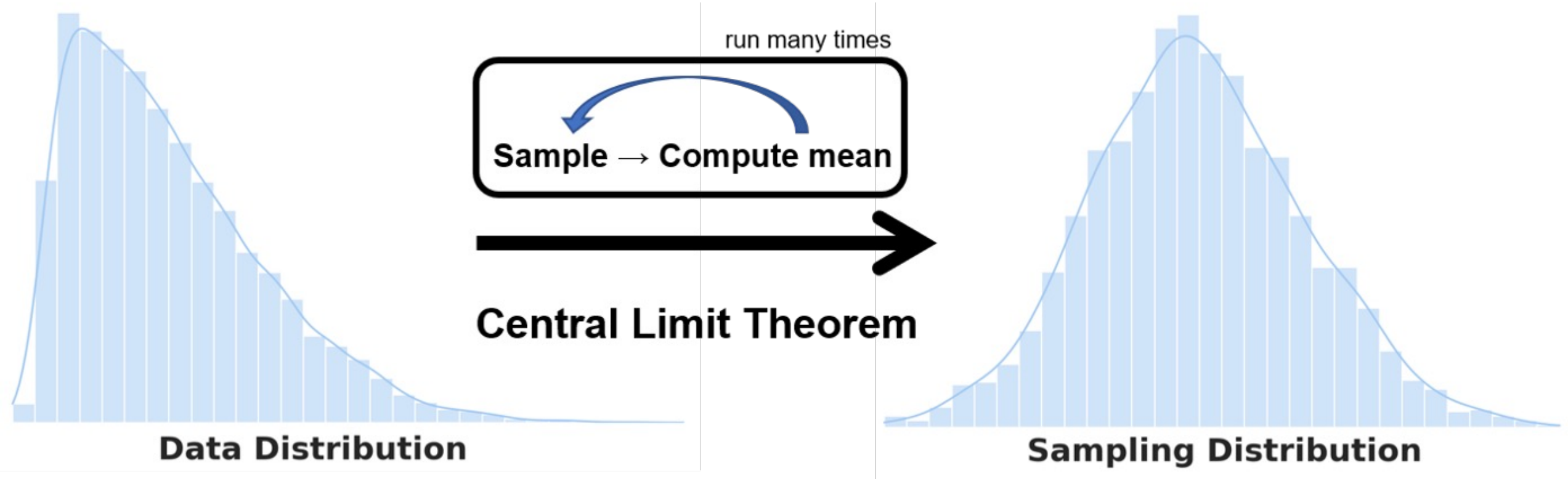
27 October 2022

ACIBADEM
MEHMET ALİ AYDINLAR ÜNİVERSİTESİ

Sampling Distribution

- Population Distribution
- Sample Distribution
- **Sampling Distribution**
  - theoretical probability distribution of a statistic obtained through a specific number of samples drawn from a specific population
  - if samples are randomly selected, the sample means will be somewhat different from the population mean (sampling error)

https://towardsdatascience.com/data-distribution-vs-sampling-distribution-what-you-need-to-know-294819109796



Sampling Distribution - cont.

- If sample size is large enough, the sampling distribution of the sample mean will be approximately normal

- the mean of the sample means will be the same as the population mean
- the standard deviation of the sample means = $\frac{\sigma}{\sqrt{n}}$

https://www.scribbr.com/statistics/confidence-interval/

Confidence Interval

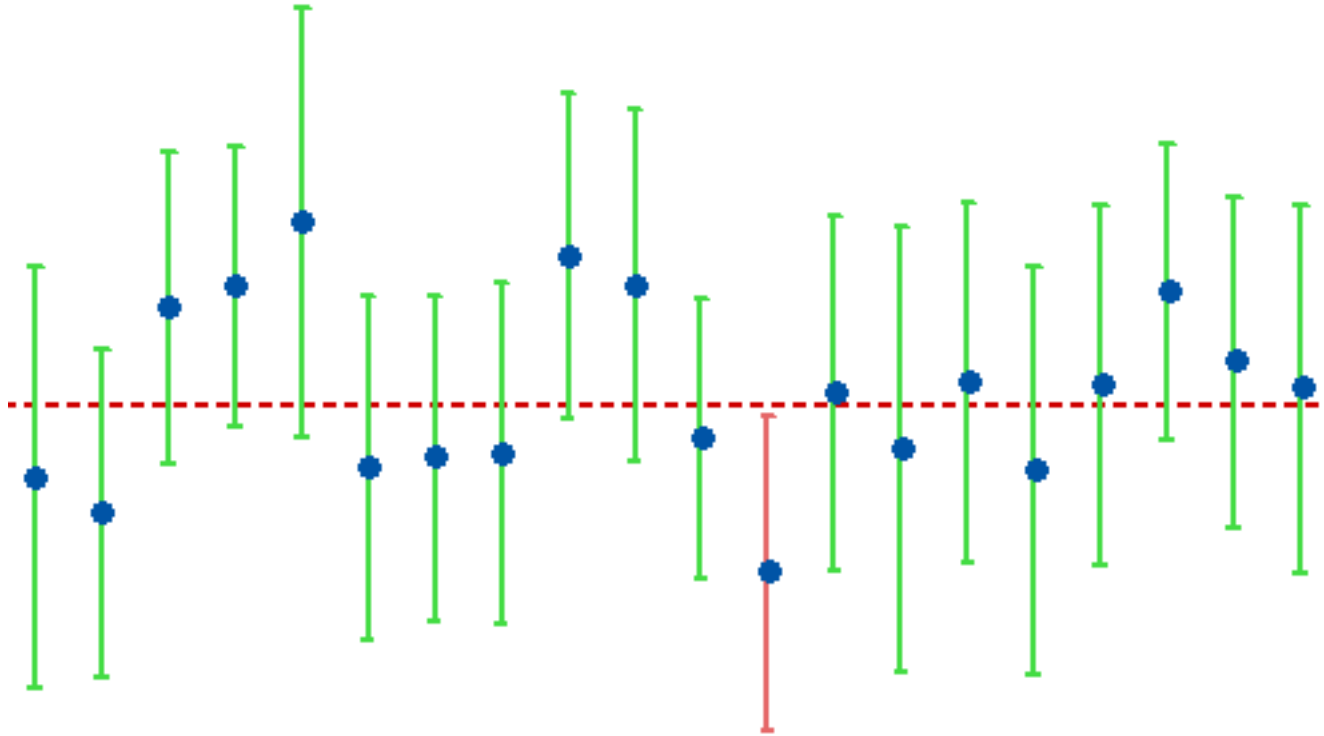
- When you make an estimate in statistics, there is always uncertainty around that estimate because the number is based on a single sample
- The confidence interval is the **range of values that you expect your estimate to fall between a certain percentage of the time** if you run your experiment again (re-sample the population in the same way)

https://www.scribbr.com/statistics/confidence-interval/

Confidence Interval

- The **confidence level** is the percentage of times you expect to reproduce an estimate between the upper and lower bounds of the confidence interval
  - if you construct a confidence interval with a 95% confidence level, you are confident that 95 out of 100 times the estimate will fall between the upper and lower values specified by the confidence interval

https://statisticsbyjim.com/hypothesis-testing/confidence-interval/



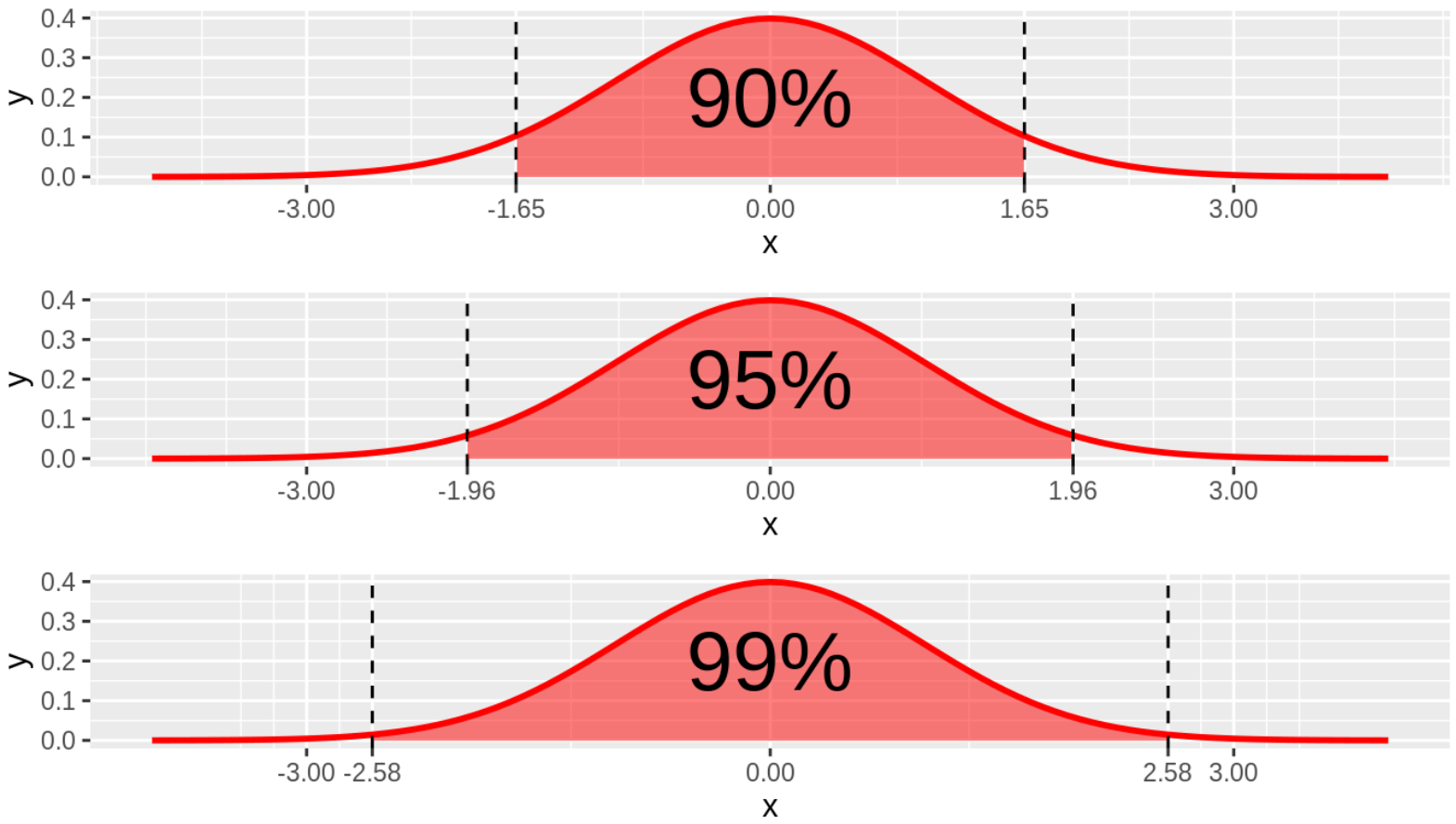
a 95% confidence interval [10 15] indicates that we can be 95% confident that the parameter is within that range

However, it does NOT indicate that 95% of the sample values occur in that range

Confidence Interval

$$CI = \bar{x} \pm Z * \frac{s}{\sqrt{n}}$$

$$CI = \bar{x} \pm t * \frac{s}{\sqrt{n}}$$



Confidence Interval - Example

id	week_1	cd4_1	week_2	cd4_2	perc_benefit
361	0	26	7.43	3	-11.905994
1017	0	13	7.00	10	-3.296703
519	0	3	8.14	5	8.190008
1147	0	65	33.00	97	1.491841
1216	0	36	8.00	31	-1.736111
52	0	16	9.43	31	9.941676
660	0	34	8.43	32	-0.697788
1145	0	41	8.00	71	9.146341
697	0	33	8.00	45	4.545455
560	0	21	8.00	27	3.571429

- Mean percentage benefit is 1.925015
- What is the 95% confidence interval of the mean percentage benefit?

Confidence Interval - Example (cont.)

Demo in R

- Mean percentage benefit is 1.925015
- Standard deviation is 6.702202
- Sample size is 10

$$95\%\ CI = [\bar{X} - t^* \frac{s}{\sqrt{n}}, \bar{X} + t^* \frac{s}{\sqrt{n}}] \qquad (t^* \sim t_{n-1} = t_9)$$

Hypothesis Testing

- **Hypothesis**: an assumption that can be tested based on the evidence available
  - A novel drug is efficient in treating a certain disease
  - Regular smoking leads to lung cancer
  - Overweight individuals who (1) consume greasy food and (2) consume a low amount vegetables (1) have high levels of cholesterol and (2) have a higher risk of cardiovascular diseases
- **Hypothesis test**: investigation of the hypothesis using the sample
  - Assessing evidence provided by the data against the null claim (the claim which is to be assumed true unless enough evidence exists to reject it)

Null and Alternative Hypotheses

- H_0 – Null hypothesis
  - The mean of a variable is not different than *c*
  - There is no difference between the two groups' means
  - There is no difference compared to baseline
  - …
- H_a or H_1 – Alternative hypothesis
  - There is a difference between the two groups' means
  - The mean in group A is higher than group B
  - …

One- vs. Two-tailed Tests

- The coin is biased

Two-tailed

H_0: $p = 0.5$

H_a: $p \neq 0.5$

- The probability of heads is larger (or smaller) than 0.5

One-tailed

H_0: $p \leq 0.5$ (or $p \geq 0.5$)

H_a: $p > 0.5$ (or $p < 0.5$)

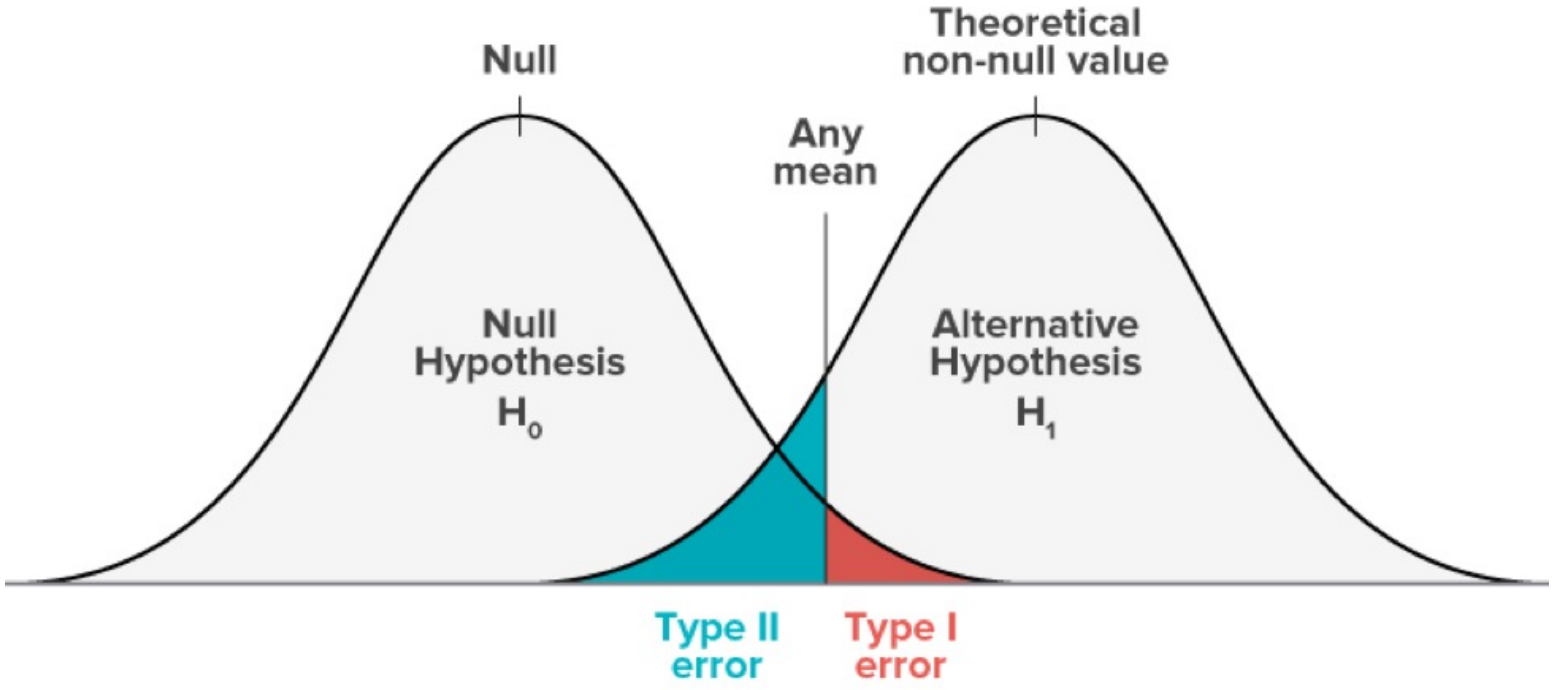
H_0	Decision	
	Fail to reject	Reject
True	Correct decision	**Type I Error** α
False	**Type II Error** **ß**	Correct decision

Barai A. Hypothesis Testing | All a beginner needs to know [Internet]. Medium. 2020 [cited 2021 Oct 4]. Available from: https://medium.com/analytics-vidhya/hypothesis-testing-all-a-beginner-needs-to-know-dd1bdeebdbf1

Hypothesis Testing

- P(Type 1 error) = α = P(reject H_0| H_0 is true)
- P(Type 2 error) = β = P(fail to reject H_0| H_0 is false)

- As α gets larger β gets smaller, vice versa
- As *n* gets large, both α and β get smaller



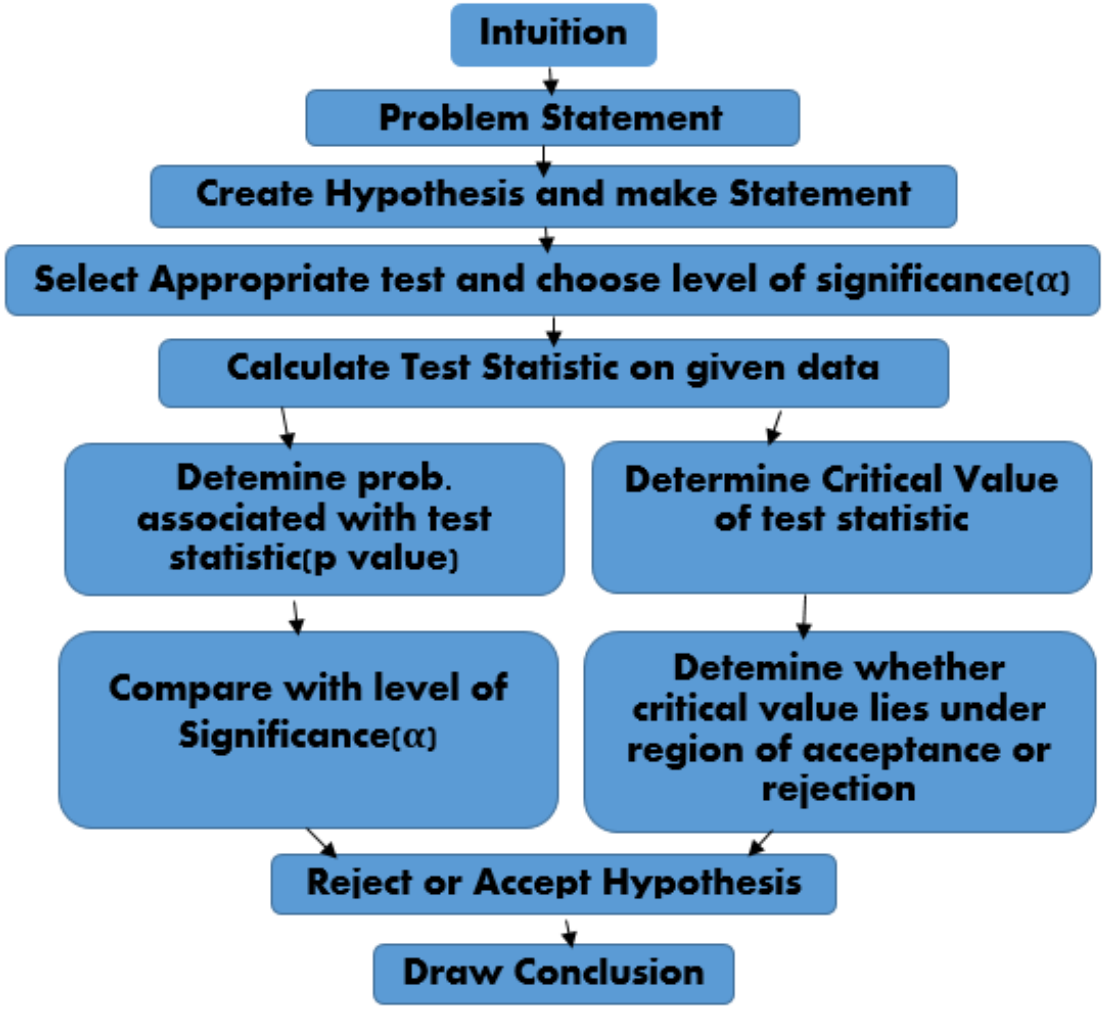
Hypothesis Testing

H_0	Decision	
	Fail to reject	Reject
True	Correct decision	**Type I Error** α
False	**Type II Error** ß	Correct decision

- **Confidence level** = $1 - \alpha$
  - P(fail to reject H_0| H_0 is true)
- **Statistical power** = $1 - \beta$
  - P(reject H_0| H_0 is false)

Barai A. Hypothesis Testing | All a beginner needs to know [Internet]. Medium. 2020 [cited 2021 Oct 4]. Available from: https://medium.com/analytics-vidhya/hypothesis-testing-all-a-beginner-needs-to-know-dd1bdeebdbf1

Hypothesis Testing - Steps



Hypothesis Testing - Steps

1. **Check assumptions, determine H_0 and H_a, choose α**
   - Assumptions differ based on the test
   - The null hypothesis always contains equality (=)
2. **Calculate the appropriate test statistic**
   - z, t, χ^2, …
3. **Calculate critical values/p value**
   - With the aid of precalculated tables/software
4. **Decide whether to reject/fail to reject H_0**
   - Reject if the statistic is within the critical region/$p \leq \alpha$

Test Statistic

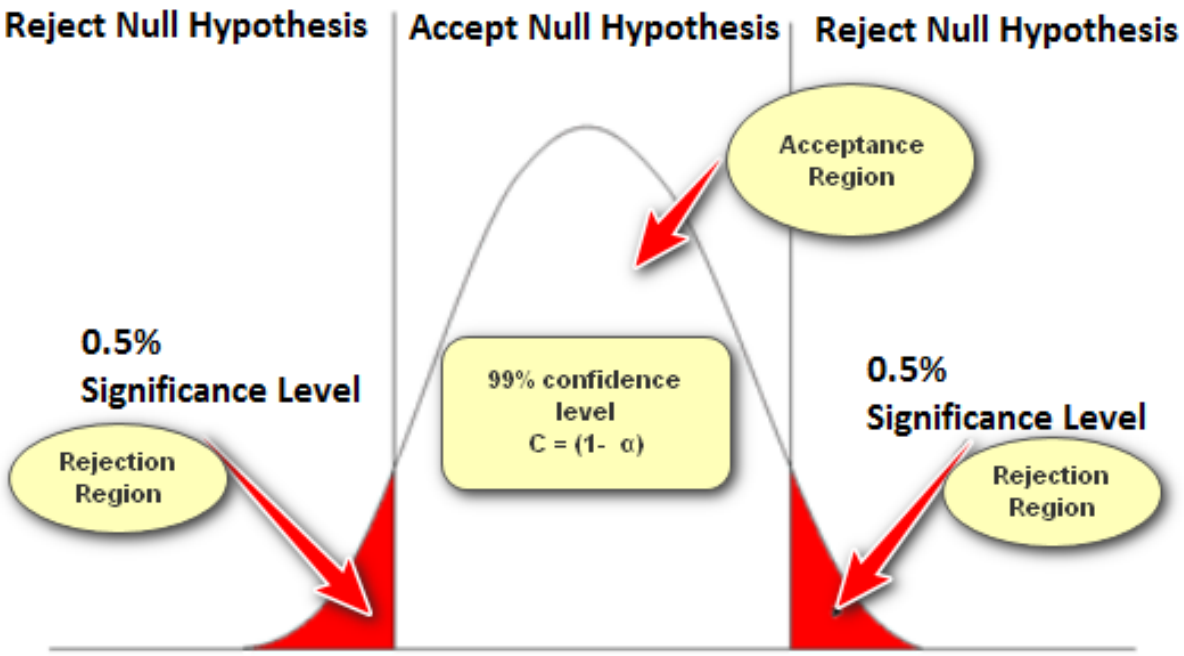
$$\text{test statistic} = \frac{\text{estimator} - \text{null value}}{\text{standard error of estimator}}$$

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

Barai A. Hypothesis Testing | All a beginner needs to know [Internet]. Medium. 2020 [cited 2021 Oct 4]. Available from: https://medium.com/analytics-vidhya/hypothesis-testing-all-a-beginner-needs-to-know-dd1bdeebdbf1

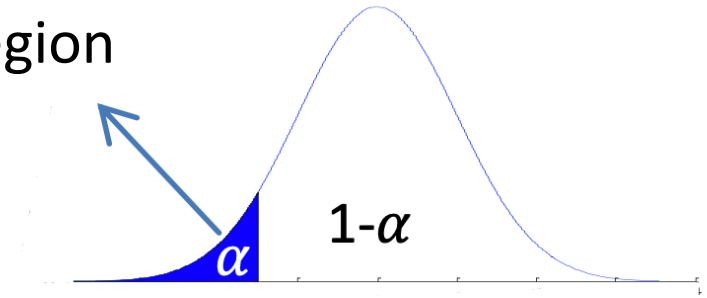
Critical Value/Rejection Region

- We select α (**significance level**) prior to performing a hypothesis test
  - Some common values for α are 0.01, **0.05** an 0.10
- Based on the selected α, the critical values are calculated, and the rejection region is determined
  - the region where the null hypothesis is rejected



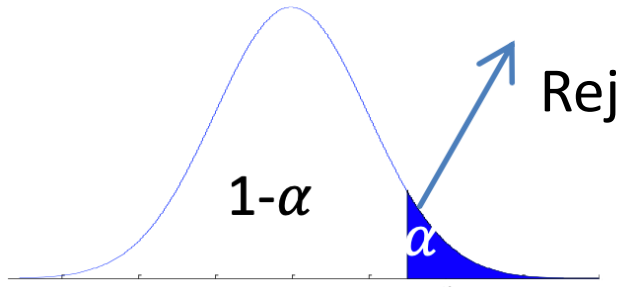
$$H_0: \mu = \mu_0$$
$$H_1: \mu < \mu_0$$

Rejection region



$$H_0: \mu = \mu_0$$
$$H_1: \mu > \mu_0$$

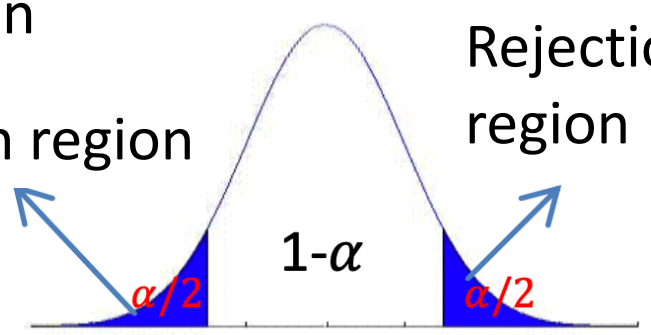
Rejection region



$$H_0: \mu = \mu_0$$
$$H_1: \mu \neq \mu_0$$

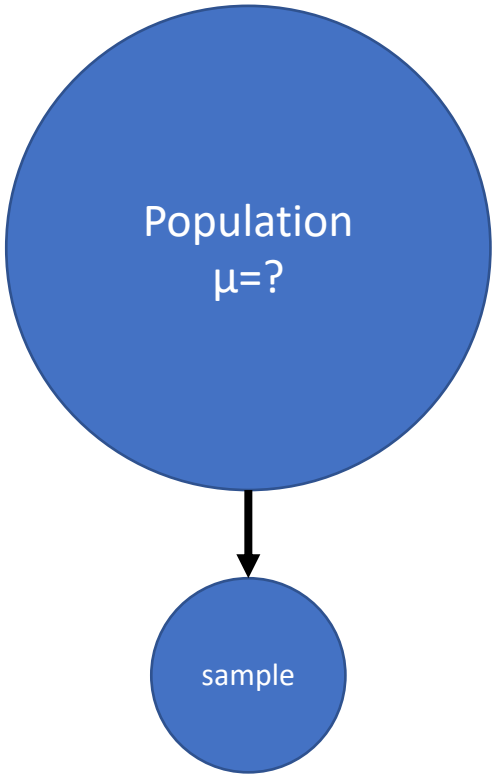
Rejection region

Rejection region

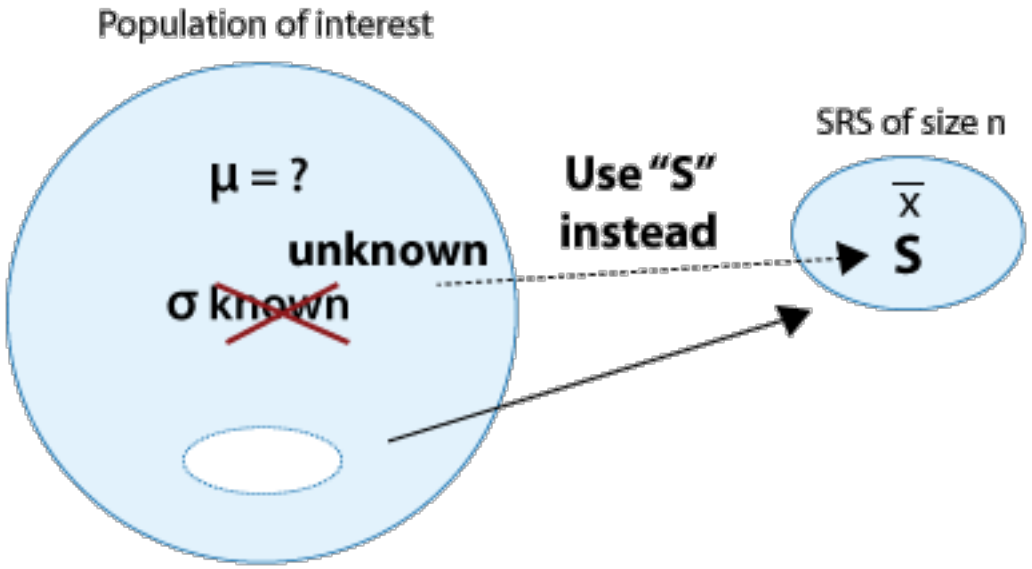


One-Sample t-Test

- a statistical hypothesis test used to determine whether an unknown population mean is different from a specific value



One-Sample t-Test



Means(All steps) [Internet]. [cited 2021 Oct 6]. Available from: https://bolt.mph.ufl.edu/6050-6052/unit-4/module-12/means-all-steps/

One-Sample t-Test – Example I

id	week_1	cd4_1	week_2	cd4_2	perc_benefit
361	0	26	7.43	3	-11.905994
1017	0	13	7.00	10	-3.296703
519	0	3	8.14	5	8.190008
1147	0	65	33.00	97	1.491841
1216	0	36	8.00	31	-1.736111
52	0	16	9.43	31	9.941676
660	0	34	8.43	32	-0.697788
1145	0	41	8.00	71	9.146341
697	0	33	8.00	45	4.545455
560	0	21	8.00	27	3.571429

- Mean percentage benefit is 1.925015
- Is it due to chance? Or does it indicate positive impact of the novel treatment?
  - What would be the value of mean percentage benefit what if you selected another set of 10 patients?

One-Sample t-Test – Example I (cont.)

1. Check assumptions, determine H_0 and H_a, choose α
   - Normality of the variable is checked (Quantile-quantile plot)
   - $H_0: \mu = 0$ $H_a: \mu \neq 0$
   - $\alpha = 0.05$

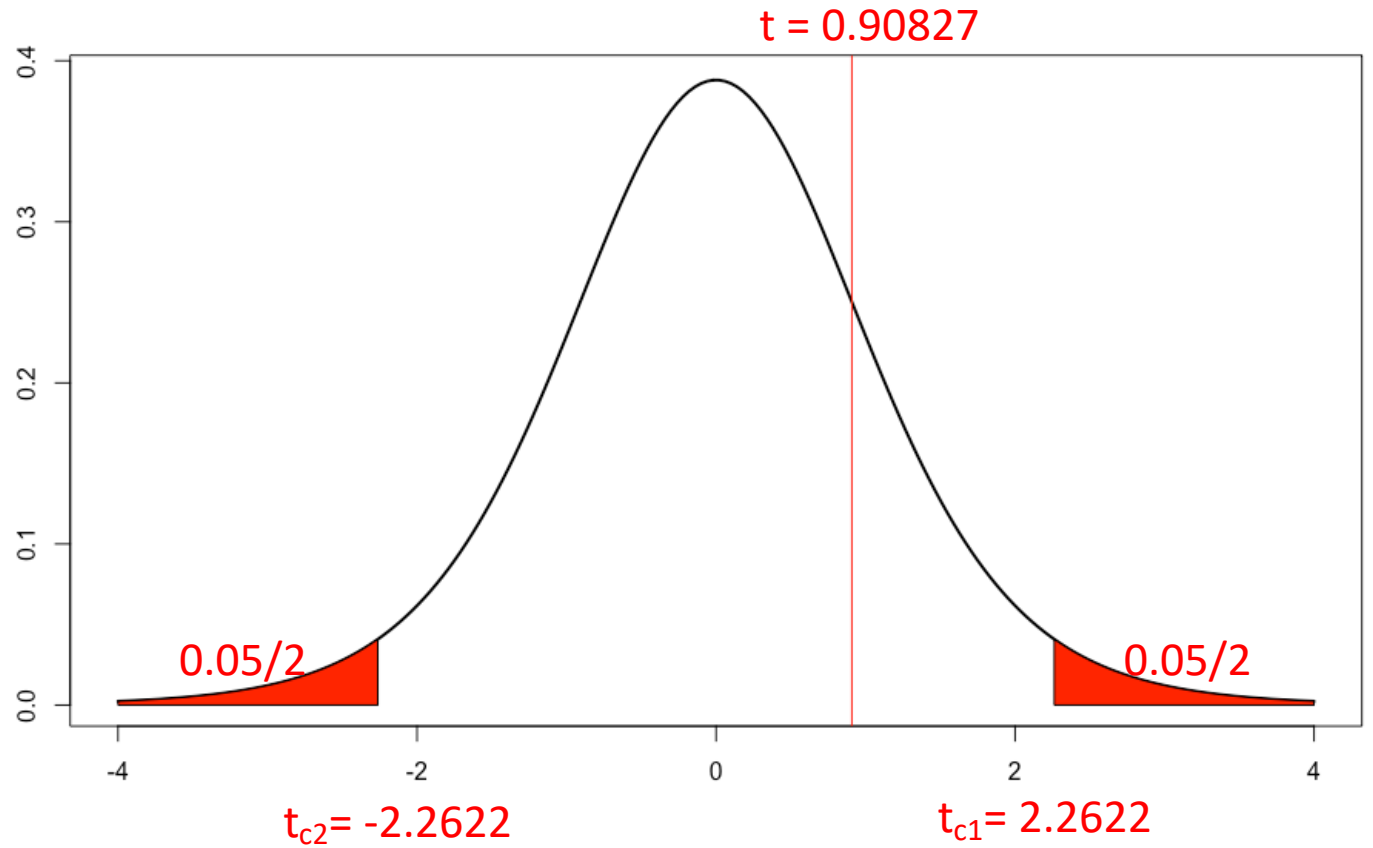
One-Sample t-Test – Example I (cont.)

2. Calculate the appropriate test statistic
   - Mean percentage benefit is 1.925015
   - Standard deviation is 6.702202
   - Sample size is 10

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{1.925015 - 0}{6.702202/\sqrt{10}} = 0.9082736 \quad (\sim t_{n-1} = t_9)$$

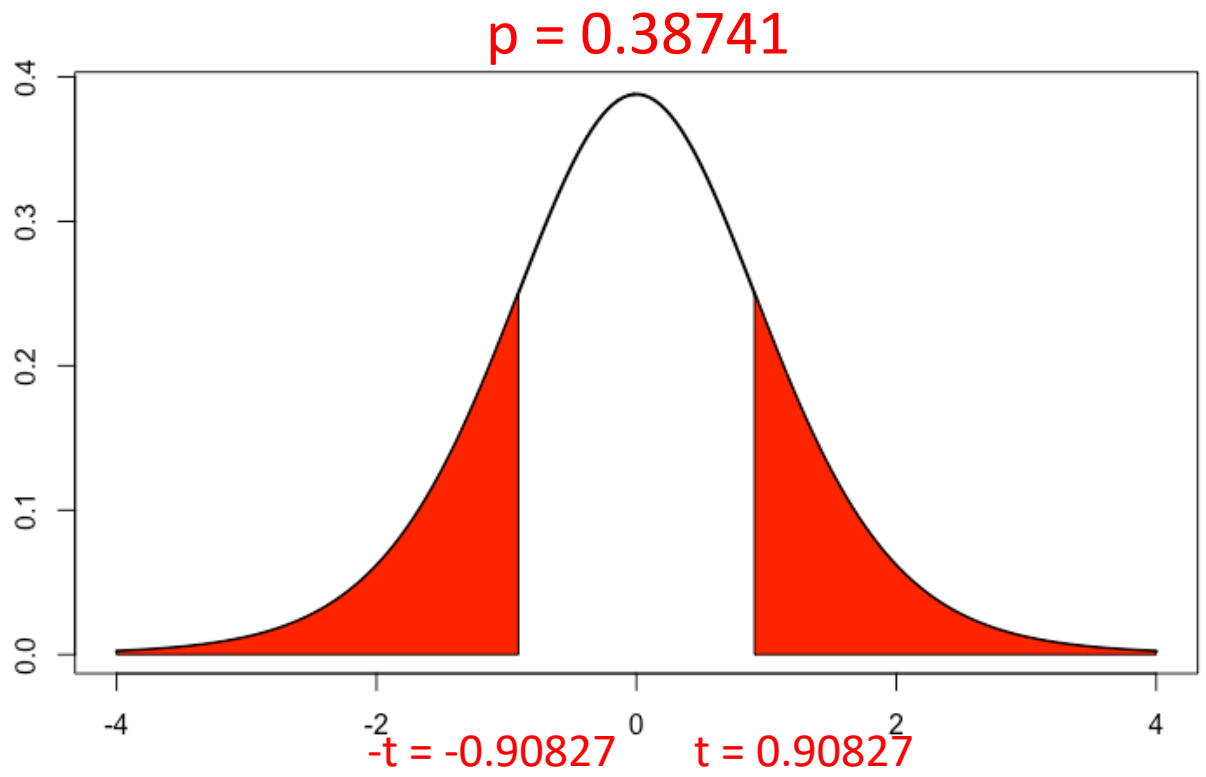
One-Sample t-Test – Example I (cont.)

3. Calculate **critical values**/p value
4. Decide whether to reject/fail to reject H_0



One-Sample t-Test – Example I (cont.)

3. Calculate critical values/**p value**
4. Decide whether to reject/fail to reject H_0



One-Sample t-Test – Example II

- It is claimed that the post-treatment tumor volume of glioblastoma patients subject to a novel treatment is different than 5 cm^3
- The mean tumor volume of 41 randomly-selected patients is 5.9 cm^3
- Sample standard deviation is 1.74

One-Sample t-Test – Example II (cont.)

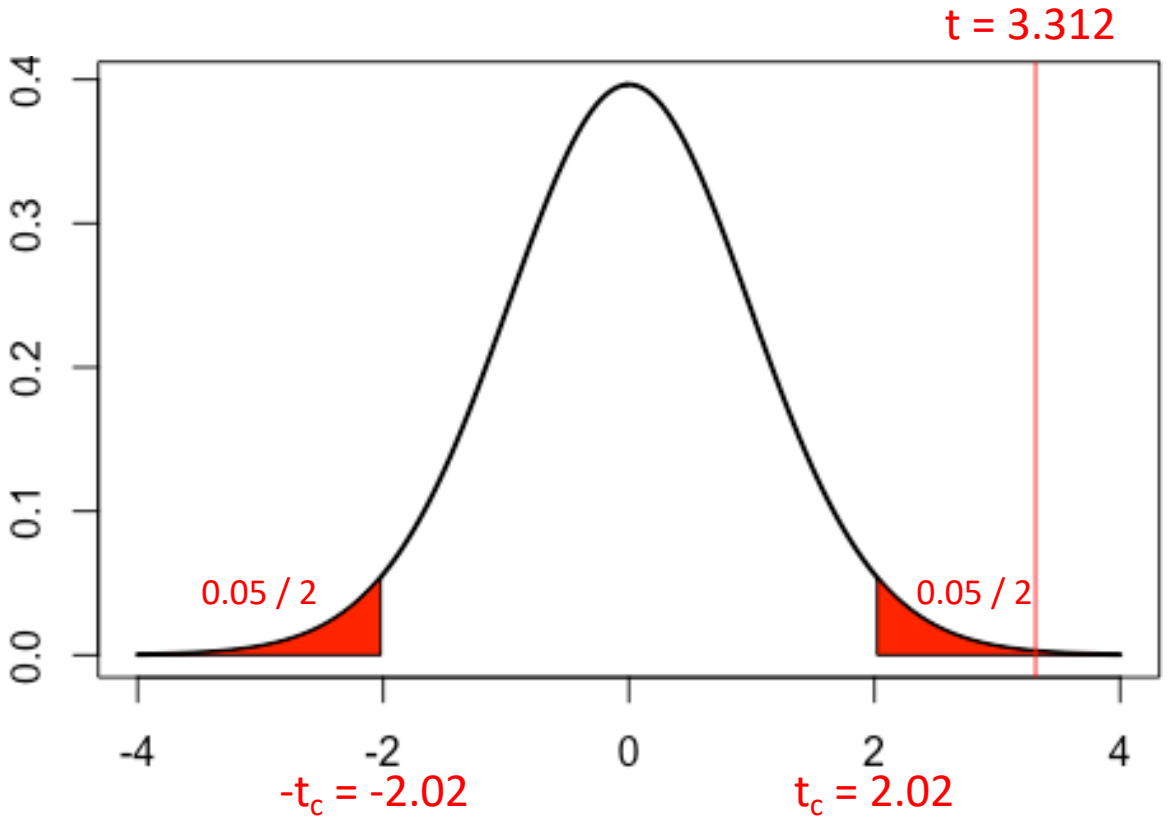
1. Check assumptions, determine H_0 and H_a, choose α
   - Normality of the variable is checked
   - H_0: $\mu = 5$ H_a: $\mu \neq 5$
   - $\alpha = 0.05$

2. Calculate the appropriate test statistic

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{5.9 - 5}{1.74/\sqrt{41}} = 3.312 \qquad (\sim t_{n-1} = t_{40})$$

One-Sample t-Test – Example II (cont.)

3. Calculate **critical values**/p value
4. Decide whether to reject/fail to reject H_0



One-Sample t-Test – Example II (cont.)

5. State a conclusion:

With 95% confidence, we can conclude that there is enough evidence to say that post-treatment tumor volume of glioblastoma patients subject to a novel treatment is different than 5 cm^3.

One-Sample t-Test – Example III

- It is claimed that:
- A novel drug reduces the recovery time of patients to less than 10 days
- Recovery time for 7 randomly-selected patients:

  2, 4, 11, 3, 4, 6, 8 ($\bar{X}$ = 5.43, s = 3.15)

- Test the hypothesis using α = 0.01

One-Sample t-Test – Example III((cont.)

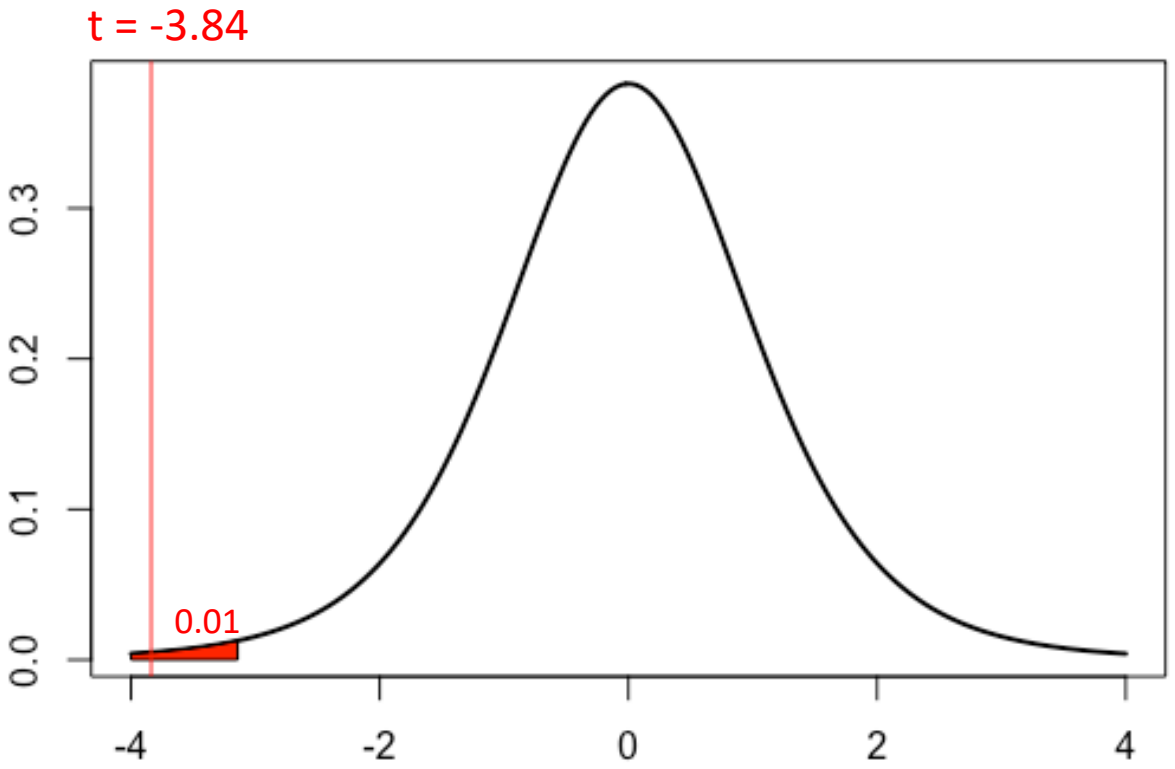
1. Check assumptions, determine H_0 and H_a, choose α
   - Normality of the variable is checked
   - H_0: $\mu \geq 10$ H_a: $\mu < 10$
   - $\alpha = 0.01$

2. Calculate the appropriate test statistic

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{5.43 - 10}{3.15/\sqrt{7}} = -3.84 \quad (\sim t_{n-1} = t_6)$$

One-Sample t-Test – Example III (cont.)

3. Calculate **critical values**/p value
4. Decide whether to reject/fail to reject H_0



Barai A. Hypothesis Testing | All a beginner needs to know [Internet]. Medium. 2020 [cited 2021 Oct 4]. Available from: https://medium.com/analytics-vidhya/hypothesis-testing-all-a-beginner-needs-to-know-dd1bdeebdbf1

Brief Summary