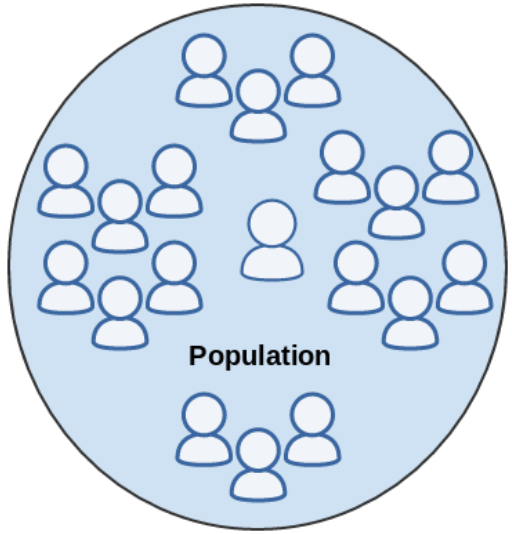
Statistical testing

February 12, 2020

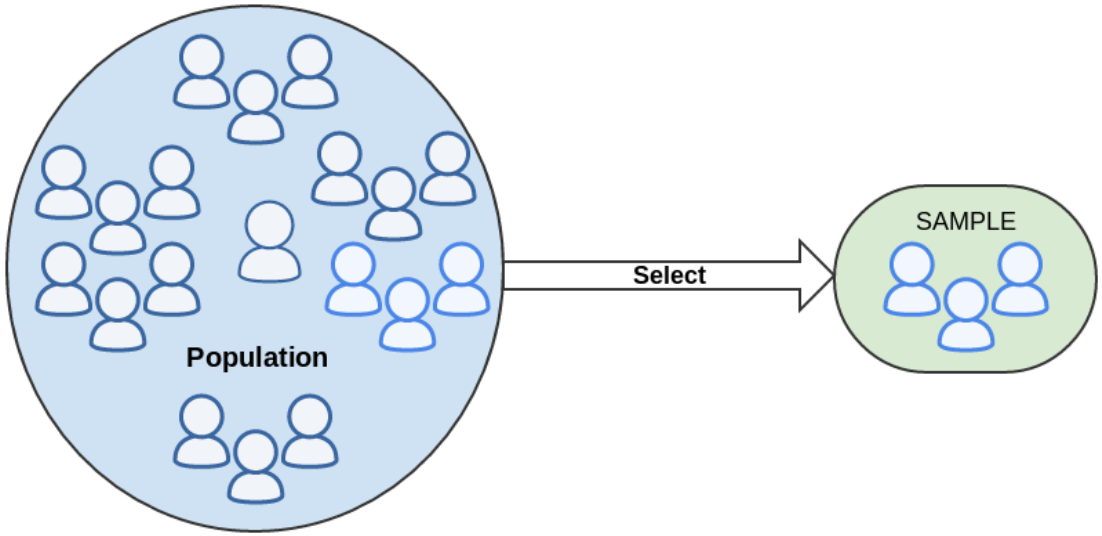
Recap

- **Population** represents the set of individuals having a number of common characteristics.



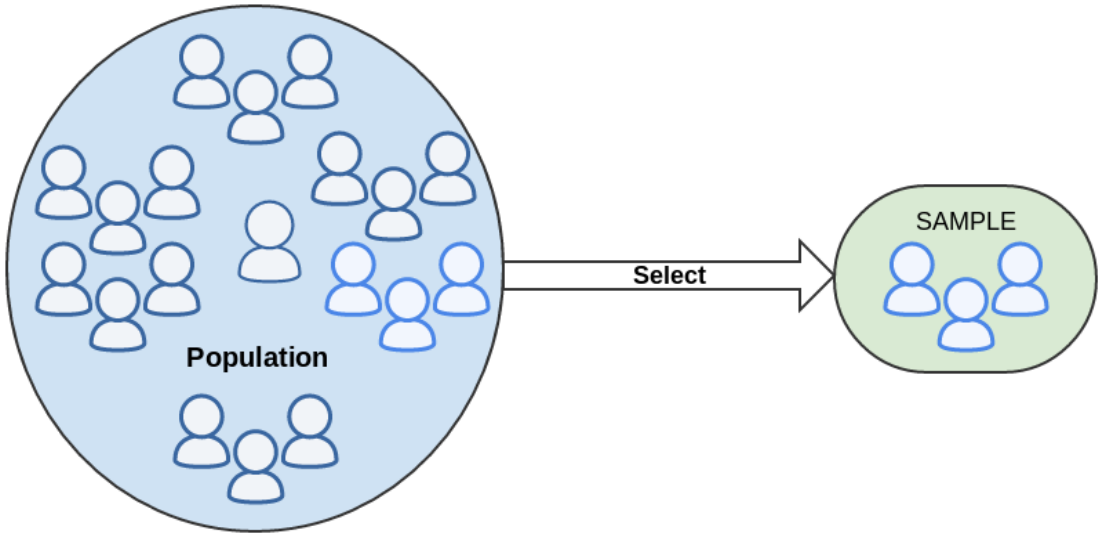
Recap

- **Population** represents the set of individuals having a number of common characteristics.
- **Sample** represents a subgroup of the analyzed population.



Recap

- **Population** represents the set of individuals having a number of common characteristics.
- **Sample** represents a subgroup of the analyzed population.



- Representative sample?

Central tendency and spread

Central tendency:

- Average: represents the arithmetical mean of values; appropriate for numerical values;

Central tendency and spread

Central tendency:

- Average: represents the arithmetical mean of values; appropriate for numerical values;
- Median: is the value that splits an ordered array into two equal parts; suitable for ordinal and numeric values;

Central tendency and spread

Central tendency:

- Average: represents the arithmetical mean of values; appropriate for numerical values;
- Median: is the value that splits an ordered array into two equal parts; suitable for ordinal and numeric values;
- Mode: is the value that appears more frequent in a set of values (sample); appropriate for nominal, ordinal and numeric values;

Central tendency and spread

Central tendency:

- Average: represents the arithmetical mean of values; appropriate for numerical values;
- Median: is the value that splits an ordered array into two equal parts; suitable for ordinal and numeric values;
- Mode: is the value that appears more frequent in a set of values (sample); appropriate for nominal, ordinal and numeric values;

Spread:

Central tendency and spread

Central tendency:

- Average: represents the arithmetical mean of values; appropriate for numerical values;
- Median: is the value that splits an ordered array into two equal parts; suitable for ordinal and numeric values;
- Mode: is the value that appears more frequent in a set of values (sample); appropriate for nominal, ordinal and numeric values;

Spread:

- Standard deviation: represents the distance from the average in a sample

Central tendency and spread

Central tendency:

- Average: represents the arithmetical mean of values; appropriate for numerical values;
- Median: is the value that splits an ordered array into two equal parts; suitable for ordinal and numeric values;
- Mode: is the value that appears more frequent in a set of values (sample); appropriate for nominal, ordinal and numeric values;

Spread:

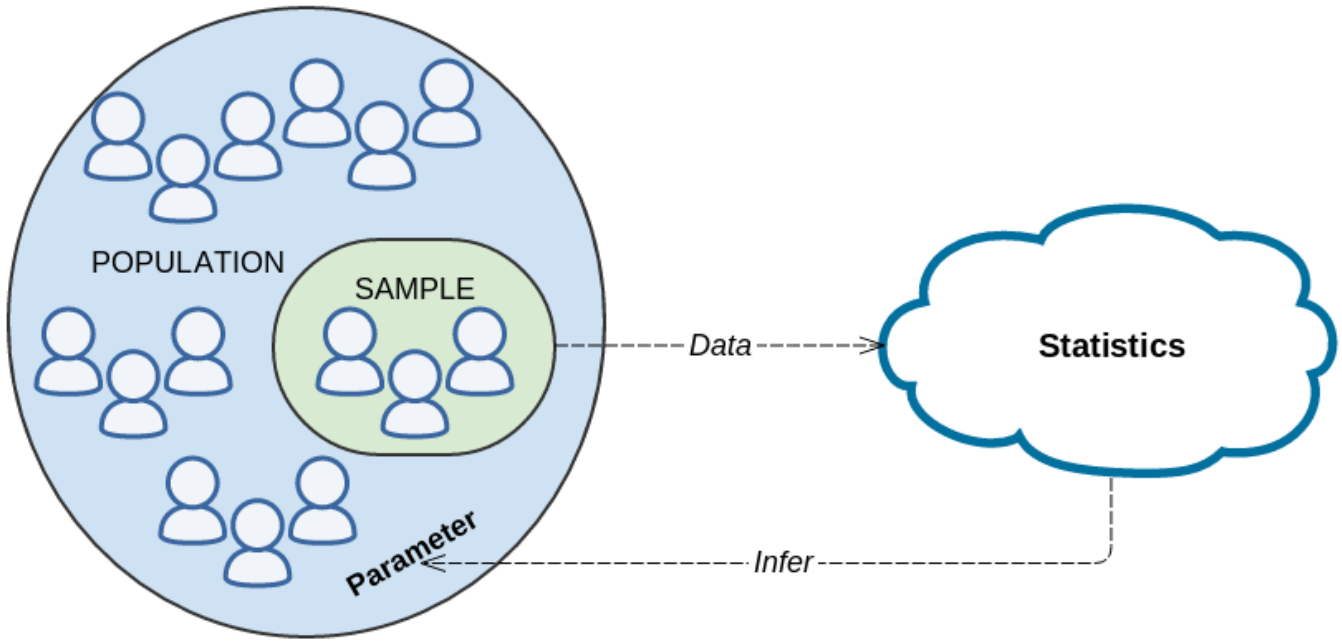
- Standard deviation: represents the distance from the average in a sample
- Standard Error of Mean: represents the spread of mean values of the sample with respect to the mean of the population

Statistical inference

Statistical inference is the process that generalizes the characteristics of a sample to an entire population.

Statistical inference

Statistical inference is the process that generalizes the characteristics of a sample to an entire population.
Starting with a *Sample*, we are interested in finding parameters for the entire *Population* . . . but we can compute statistics only for the sample.



Statistical inference

Statistical inference is the process that generalizes the characteristics of a sample to an entire population.

	Population Parameters Reality Greek letters	Sample Statistics Estimates reality Latin letters
Average	μ	X
Standard Deviation	σ	SD
Number of observations	N	n

Confidence Intervals

- We are unable to calculate exactly the parameters of a population

Confidence Intervals

- We are unable to calculate exactly the parameters of a population
- The value of a statistic is biased by the selection of the sample

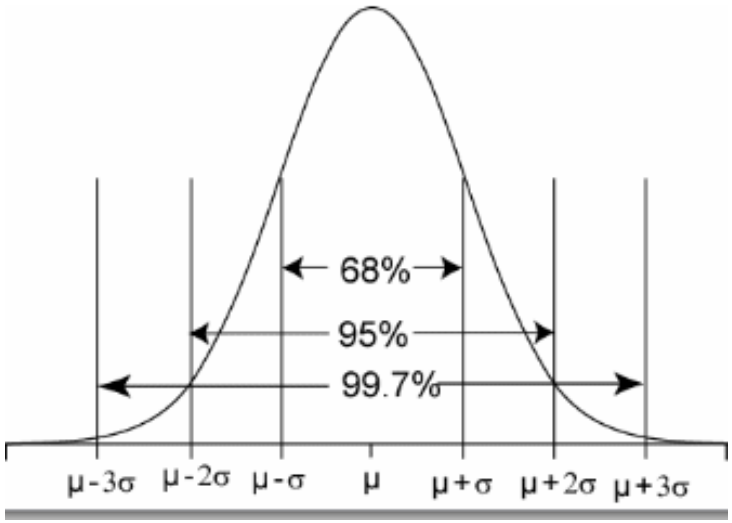
Confidence Intervals

- We are unable to calculate exactly the parameters of a population
- The value of a statistic is biased by the selection of the sample
- By using confidence intervals we are able to estimate a parameter of a population regarding the variations of statistics in samples

Confidence intervals

Definition: Confidence Interval defines an interval in which is probable to have the parameter of the population
Estimating a value interval:

- 68% of individuals can be found at $\pm 1 * \sigma$ around μ
- 95% of individuals can be found at $\pm 2 * \sigma$ around μ
- 99.7% of individuals can be found at $\pm 3 * \sigma$ around μ



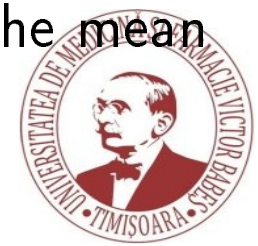
Estimating the mean

$$SEM = \frac{SD}{\sqrt{N}}$$

- $(1 - \alpha) = 0.68 \Rightarrow X = \mu \pm 1 * SEM$
- $(1 - \alpha) = 0.95 \Rightarrow X = \mu \pm 2 * SEM$
- $(1 - \alpha) = 0.997 \Rightarrow X = \mu \pm 3 * SEM$

Work in class

Assume we have a group of 32 smokers from whom we collect information about their blood pressure (mmHg). The sample has mean 95 and standard deviation 4. a) Compute the interval where 95% of the blood pressure values will be located. b) With 95% confidence estimate the mean blood pressure value for the smokers.



Statistical testing?

Statistical tests represent the process through which we can decide, with a specific degree of belief, if the differences between statistical parameters are significant or not.

Statistical testing?

Statistical tests represent the process through which we can decide, with a specific degree of belief, if the differences between statistical parameters are significant or not.

- Non-significant Differences
- Significant differences
- Significance level α

Statistical testing?

Statistical tests represent the process through which we can decide, with a specific degree of belief, if the differences between statistical parameters are significant or not.

- Non-significant Differences are those differences having a **high** probability of appearance due to chance (main reason is variability of the sample).
- Significant differences
- Significance level α

Statistical testing?

Statistical tests represent the process through which we can decide, with a specific degree of belief, if the differences between statistical parameters are significant or not.

- Non-significant Differences
- Significant differences are those differences having a **low** probability of appearance due to chance.
- Significance level α

Statistical testing?

Statistical tests represent the process through which we can decide, with a specific degree of belief, if the differences between statistical parameters are significant or not.

- Non-significant Differences
- Significant differences
- Significance level α it is the conventional value under which we start considering that the *differences are significant*. In practice we normally use $\alpha = 0.05$ or 5%.

Stages of a statistical study

1. Formulating the research (clinical) hypothesis

Stages of a statistical study

1. Formulating the research (clinical) hypothesis
2. Formulating the statistical hypothesis

Stages of a statistical study

1. Formulating the research (clinical) hypothesis
2. Formulating the statistical hypothesis
3. Selecting a sample - collecting data about the sample

Stages of a statistical study

1. Formulating the research (clinical) hypothesis
2. Formulating the statistical hypothesis
3. Selecting a sample - collecting data about the sample
4. Find the statistics of the test

Stages of a statistical study

1. Formulating the research (clinical) hypothesis
2. Formulating the statistical hypothesis
3. Selecting a sample - collecting data about the sample
4. Find the statistics of the test
5. Statistical testing - evaluate the results

Stages of a statistical study

1. Formulating the research (clinical) hypothesis
2. Formulating the statistical hypothesis
3. Selecting a sample - collecting data about the sample
4. Find the statistics of the test
5. Statistical testing - evaluate the results
6. Choose one statistical hypothesis (accepting or rejecting H_0)

Stages of a statistical study

1. Formulating the research (clinical) hypothesis
2. Formulating the statistical hypothesis
3. Selecting a sample - collecting data about the sample
4. Find the statistics of the test
5. Statistical testing - evaluate the results
6. Choose one statistical hypothesis (accepting or rejecting H_0)
7. Define the clinical conclusion

Statistical Hypothesis

Research hypothesis: certifies that there is a difference between the studied groups or a possible association between factors

Statistical Hypothesis

Research hypothesis: certifies that there is a difference between the studied groups or a possible association between factors
Statistical testing of the hypothesis allows us to quantify the risk of error that is involved in the statistical inference mechanism.

Statistical Hypothesis

Statistical hypothesis is a sentence (phrase) that contains a positive or negative affirmation with respect to a parameter of a population.
There are two statistical hypothesis H_0 and H_a.

The null hypothesis (H_0)

Denoted by H_0, is called **the null hypothesis**.

H_0 usually makes the statement that *between the elements that we are comparing there are **NO** significant differences*.

The null hypothesis (H_0)

Denoted by H_0, is called **the null hypothesis**.

H_0 usually makes the statement that *between the elements that we are comparing there are **NO** significant differences*.

Allows us to compare:

- mean values with a given means (theoretical means)
- two means
- a theoretical distribution with an experimental distribution
- two experimental distributions
- more than two averages, etc.

The null hypothesis (H_0)

Denoted by H_0, is called **the null hypothesis**.

H_0 usually makes the statement that *between the elements that we are comparing there are **NO** significant differences*.

Let us assume that we are interested in comparing average height of men vs women. Then the null hypothesis is written symbolic (mathematical) as:

$$H_0 : \mu_B = \mu_F$$

And it reads to: *Between the average height of men and the average height of women there are no significant differences.*

Alternative hypothesis (H_a)

Alternative hypothesis is true when H_0 is said to be not true (false).

H_a can classified as follows:

- **two tailed alternative hypothesis**: $\mu_M \neq \mu_F$, *average height of men is different from average height of women.*
- **single tailed alternative hypothesis**: $\mu_M > \mu_F$, *average height of men is greater than average height of women* **OR** $\mu_M < \mu_F$, *average height of men is lower than average height of women*

Significance level (α)

Significance level (α) represents the value from which we start considering that the differences that appeared between the compared entities are significant.

- Usually, in practice we use $\alpha = 0.05$ that is $\alpha = 5\%$.

Significance level (α)

Significance level (α) represents the value from which we start considering that the differences that appeared between the compared entities are significant.

- Usually, in practice we use $\alpha = 0.05$ that is $\alpha = 5\%$.

Besides the significance level we define the confidence level as being $1 - \alpha$. Usually, in practice we use $1 - \alpha = 0.95 = 95\%$

Significance level (α)

Significance level (α) represents the value from which we start considering that the differences that appeared between the compared entities are significant.

- Usually, in practice we use $\alpha = 0.05$ that is $\alpha = 5\%$.

Besides the significance level we define the confidence level as being $1 - \alpha$. Usually, in practice we use $1 - \alpha = 0.95 = 95\%$

Coefficient (result of a statistical test) **p** represents the probability that the observed differences appear due to chance.

Statistical decision

Our statistical decision is based on the computed value of **p**, as follows:

- if $p \geq \alpha$ we accept $H_0 \Rightarrow$ differences are not statistically significant

Statistical decision

Our statistical decision is based on the computed value of **p**, as follows:

- if $p \geq \alpha$ we accept $H_0 \Rightarrow$ differences are not statistically significant
- if $p < \alpha$ we reject $H_0 \Rightarrow$ differences are statistically significant

Statistical decision

Our statistical decision is based on the computed value of **p**

Assume we have $\alpha = 0.05$ and we reject H_0 doctors conclude:

- if $p < 0.05 \Rightarrow$ the differences are statistically significant

Statistical decision

Our statistical decision is based on the computed value of **p**

Assume we have $\alpha = 0.05$ and we reject H_0 doctors conclude:

- if $p < 0.05 \Rightarrow$ the differences are statistically significant
- if $p < 0.005 \Rightarrow$ the differences are **very** significant

Statistical decision

Our statistical decision is based on the computed value of **p**

Assume we have $\alpha = 0.05$ and we reject H_0 doctors conclude:

- if $p < 0.05 \Rightarrow$ the differences are statistically significant
- if $p < 0.005 \Rightarrow$ the differences are **very** significant
- if $p < 0.0005 \Rightarrow$ the difference are **extremely** significant

Statistical Errors

There are two main categories of statistical errors:

1. **Type I Error**: it appears when we reject H_0 although it is true

Statistical Errors

There are two main categories of statistical errors:

1. **Type I Error**: it appears when we reject H_0 although it is true
2. **Type II Error**: this type of error appears when we accept H_0 although it is false

Statistical Errors

There are two main categories of statistical errors

	Reality	
Decision	H_0 is True	H_0 is False
Accept H_0	**Correct decision** $p = 1 - \alpha$	**Type II Error** $p = \beta$
Reject H_0	**Type I Error** $p = \alpha$	**Correct** $p = 1 - \beta$

Statistical tests classification

- **Parametric tests** the distribution is given (usually the normal distribution is used)

Statistical tests classification

- **Parametric tests**
- **Non-parametric tests** the distribution is unknown (more general tests; in case we find out the actual distribution to be normal the results obtained using these tests are similar to the parametric ones)

Statistical tests classification

- **Parametric tests**
- **Non-parametric tests**
- **Significance tests** check whether a given parameter is equal with an estimated parameter (average, percent, etc.). t-test (Student) and z-test.

Statistical tests classification

- **Parametric tests**
- **Non-parametric tests**
- **Significance tests**
- **Homogeneity tests** compare two parameters (averages, percent, spread, etc). F-test (Fisher).

Statistical tests classification

- **Parametric tests**
- **Non-parametric tests**
- **Significance tests**
- **Homogeneity tests**
- **Concordance tests** compare an experimental distribution with a theoretical distribution; compares two experimental distributions

Statistical tests classification

- Parametric tests
- Non-parametric tests
- Significance tests
- Homogeneity tests
- Concordance tests
- Independence tests check whether some experimental values are independent (contingency tables). χ^2 test

Statistical tests classification

- **Parametric tests**
- **Non-parametric tests**
- **Significance tests**
- **Homogeneity tests**
- **Concordance tests**
- **Independence tests**
- **Correlation tests** evaluate significance of the estimated parameters in risk analysis

Comparing means with a given value

If our distribution is normal, then we can apply the following tests:

- In case the sample volume is less than 30 individuals ($n < 30$) and the spread is **unknown** we apply the t-test
  - Computed as: **t computed** $p = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$

Comparing means with a given value

If our distribution is normal, then we can apply the following tests:

- In case the sample volume is less than 30 individuals ($n < 30$) and the spread is **unknown** we apply the t-test
  - Computed as: **t computed** $p = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$
- In case the sample volume is greater than 30 individuals ($n \geq 30$) and the population spread is **known** then we apply the z-test.
  - Computed as: **z computed** $p = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$
  - acceptance region is $[-1.96; 1.96]$ for $\alpha = 0.05$

Comparing means with a given value

Statistical hypothesis are the following:

- H_0:There are **NO** significant differences between the means of the population and the given value.
  - Symbolically written as: $H_0 : \mu = \mu_0$
- H_a: There **ARE** significant differences between the means of the population and the given value.
  - Symbolically written as: $H_a : \mu \neq \mu_0$

Comparing means with a given value

Statistical hypothesis are the following:

- H_0:There are **NO** significant differences between the means of the population and the given value.
  - Symbolically written as: $H_0 : \mu = \mu_0$
- H_a: There **ARE** significant differences between the means of the population and the given value.
  - Symbolically written as: $H_a : \mu \neq \mu_0$

Testing numerical variables

- **Unpaired t-test**

Testing numerical variables

- **Unpaired t-test**
  - values come from different samples
  - compares two means obtained on independent and normally distributed variables

Testing numerical variables

- **Unpaired t-test**

  Statistical hypothesis are formulated as:

  - H_0: Between the two means there are **NO** significant differences.
    - Symbolically written as: $H_0 : \mu_M = \mu_F$

Testing numerical variables

- **Unpaired t-test**

  Statistical hypothesis are formulated as:

  - H_0: Between the two means there are **NO** significant differences.
    - Symbolically written as: $H_0 : \mu_M = \mu_F$
  - H_a: Between the two means there **are** significant differences.
    - Symbolically written as: $H_a : \mu_M \neq \mu_F$ (**two tailed** H_a)
      $H_a : \mu_M > \mu_F$ or $H_a : \mu_M < \mu_F$ (**single tailed** H_a)

Testing numerical variables

- **Unpaired t-test**
- **Paired t-test**
  - measurements are from the **same** sample in **different** conditions (different time of measurement), variables are dependent
  - compares the means obtained on paired series and on normally distributed variables

Testing numerical variables

- **Unpaired t-test**
- **Paired t-test**

  Statistical hypothesis are formulated as:
  - H_0: Between the two means there are **NO** significant differences.
    - Symbolically written as: $H_0 : \mu_{t1} = \mu_{t2}$

Testing numerical variables

- **Unpaired t-test**
- **Paired t-test**

  Statistical hypothesis are formulated as:
  - H_0: Between the two means there are **NO** significant differences.
    - Symbolically written as: $H_0 : \mu_{t1} = \mu_{t2}$
  - H_a: Between the two means there **are** significant differences.
    - Symbolically written as: $H_a : \mu_{t1} \neq \mu_{t2}$ (**two tailed** H_a)
      $H_a : \mu_{t1} > \mu_{t2}$ or $H_a : \mu_{t1} < \mu_{t2}$ (**single tailed** H_a)