Recap.

Remember...

- Statistical inference is the process that allows us to generalize the characteristics of a sample to an entire population.

Recap.

Remember...

- Inference
- Estimations starting from given facts (statistics) we proceed in estimating parameters about our population (confidence intervals).

Recap.

Remember...

- Inference
- Estimations
- Research hypothesis certifies that there is a difference between the studied groups or a possible association between factors.

Recap.

Remember...

- Inference
- Estimations
- Research hypothesis
- Statistical Hypothesis: two of them H_0 and H_a stating that
  - H_0 makes the statement that between the elements that we are comparing there are **NO** significant differences; denoted $H_0 : \mu_1 = \mu_2$ or $\mu_1 - \mu_2 = 0$
  - H_a makes the statement that between the elements that we are comparing **there are** significant differences; denoted $H_a : \mu_1 \neq \mu_2$ (two tailed) and $\mu_1 < \mu_2$ (single tailed)

Recap.

Remember...

- Inference
- Estimations
- Research hypothesis
- Statistical Hypothesis
- **p**-values represents the computed result of a statistical test and it is the probability that the observed differences appear due to chance.

Recap.

Remember...

- Inference
- Estimations
- Research hypothesis
- Statistical Hypothesis
- **p**-values
- Errors when testing

	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** $p = 1 - \beta$

- α : probability of Type I error $\Rightarrow 1 - \alpha$ confidence level
- β: probability of Type II error $\Rightarrow 1 - \beta$ power of the test

Statistical tests

February 12, 2020

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** $\frac{\bar{x}-\mu_0}{s/\sqrt{n}}$

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** $\frac{\bar{x}-\mu_0}{s/\sqrt{n}}$
- 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** $\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$
- Check the table for the computed test and retrieve the p-value

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: $H_0 : \mu = \mu_0$

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: $H_a : \mu \neq \mu_0$

Example[1]

Let us suppose that we are interested in comparing the cholesterol levels among recent Asian immigrants to the USA with typical levels found in the general US population. Suppose we assume the cholesterol levels in women ages 21-40 in the US are approximately normally distributed with mean 190 mg/dL. It is unknown whether cholesterol levels among recent Asian immigrants are lower or higher than those in the general US population. Let us assume that levels among recent female Asian immigrants are normally distributed with unknown mean μ. Hence we wish to test the null hypothesis: $H_0 : \mu_0 = \mu = 190$ vs. the alternative hypothesis $H_a : \mu_0 \neq \mu$. Blood tests are performed on 100 female Asian immigrants ages 21-41, and the mean level ($\bar{x}$) is 181.52 mg/dL with standard deviation = 40 mg/dL. **Test** the hypothesis that the mean cholesterol level of recent female Asian immigrants is different from the mean in the general US population.

[1] Taken from B. Rosner - Fundamentals of biostatistics, p.216, ex. 7.20

Example[1]

We compute the test statistic:

$$\begin{aligned} t &= \frac{\bar{x} - \mu_0}{s_x/\sqrt{n}} \\ &= \frac{181.52 - 190}{40/\sqrt{100}} \\ &= \frac{-8.48}{4} = -2.12 \end{aligned} \tag{1}$$

We now need to compute the *p-value* for the test. We use Excel for this and call the TDIST function. As arguments since we have a bilateral H_a we use $\|t\|$, second argument represents the degrees of freedom $d.f = n - 1$ and specify as third argument to TDIST the value 2.

$$p = TDIST(2.12, 99, 2) = .037 \tag{2}$$

[1] Taken from B. Rosner - Fundamentals of biostatistics, p.216, ex. 7.20

Example[1]

We set $\alpha = .05$ and we have $t = -2.12$ and $p = .037$.

We compare $p = .037$ with $\alpha = .05$

[1] Taken from B. Rosner - Fundamentals of biostatistics, p.216, ex. 7.20

Example[1]

We set $\alpha = .05$ and we have $t = -2.12$ and $p = .037$.

We compare $p = .037 < \alpha = .05$

$\Rightarrow$ with confidence .05 we can Reject H_0.

And conclude that mean cholesterol level of recent Asian immigrants is significantly different from that of the general US population.

[1] Taken from B. Rosner - Fundamentals of biostatistics, p.216, ex. 7.20

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)

Examples

Let us assume that we are interested in the relationship between oral contraceptives (OC) and blood pressure (SBP) in women. We collect the following data from women while using OCs and while not using OCs.

Examples

Let us assume that we are interested in the relationship between oral contraceptives (OC) and blood pressure (SBP) in women. We collect the following data from women while using OCs and while not using OCs.

SBP while using OCs	SBP while not using OCs
128	115
115	112
106	107
128	119
122	115
145	138
132	126
109	105
102	104
117	115

Examples

1. We define the statistical hypothesis:

Examples

1. We define the statistical hypothesis:
   - $H_0 : \mu_{withoutOC} = \mu_{withOC}$

Examples

1. We define the statistical hypothesis:
   - $H_0 : \mu_{withoutOC} = \mu_{withOC}$
   - $H_a : \mu_{withoutOC} \neq \mu_{withOC}$

Examples

1. We define the statistical hypothesis:
   - $H_0 : \mu_{withoutOC} = \mu_{withOC}$
   - $H_a : \mu_{withoutOC} \neq \mu_{withOC}$
2. We compute the p-value of the test using Excel:
   $p = TTEST(range\_with\_OC; range\_without\_OC; tails; type)$
   In our case we have a two tailed test hence $tails = 2$ and it is paired, hence $type = 1$.

Examples

1. We define the statistical hypothesis:
   - $H_0 : \mu_{withoutOC} = \mu_{withOC}$
   - $H_a : \mu_{withoutOC} \neq \mu_{withOC}$
2. We compute the p-value of the test using Excel:
   $p = TTEST(range\_with\_OC; range\_without\_OC; tails; type)$
   In our case we have a two tailed test hence $tails = 2$ and it is paired, hence $type = 1$.
3. Having computed $p = .0087$ and $\alpha = .05$ we can safely reject H_0
   $\Rightarrow$ we can conclude that there are significant differences in woman's blood pressure wile using oral contraceptives vs while not using oral contraceptives.

A U.S.[2] magazine, Consumer Reports, carried out a survey of the calorie and sodium content of a number of different brands of hotdog. There were three types of hotdog: beef, meat (mainly pork and beef but can contain up to 15% poultry) and poultry. The results below are the calorie content of the different brands of beef and poultry hotdogs.
Beef hotdogs: {186, 181, 176, 149, 184, 190, 158, 139, 175, 148, 152, 111, 141, 153, 190, 157, 131, 149, 135, 132}
Poultry hotdogs: {129, 132, 102, 106, 94, 102, 87, 99, 170, 113, 135, 142, 86, 143, 152, 146, 144}
We are interested if there are significant differences in the calorie count of beef and poultry hotdogs.

[2] Data taken from Moore and McCabe Introduction to the Practice of Statistics

1. We define the statistical hypothesis:

1. We define the statistical hypothesis:
   - $H_0 : \mu_{beef} = \mu_{poultry}$

1. We define the statistical hypothesis:
   - $H_0 : \mu_{beef} = \mu_{poultry}$
   - $H_a : \mu_{beef} \neq \mu_{poultry}$

1. We define the statistical hypothesis:
   - $H_0 : \mu_{beef} = \mu_{poultry}$
   - $H_a : \mu_{beef} \neq \mu_{poultry}$
2. We compute the p-value of the test using Excel:
   $p = TTEST(range\_beef; range\_poultry; tails; type)$
   In our case we have a two tailed test hence $tails = 2$ and it is un-paired, hence $type = 3$.

1. We define the statistical hypothesis:
   - $H_0 : \mu_{beef} = \mu_{poultry}$
   - $H_a : \mu_{beef} \neq \mu_{poultry}$
2. We compute the p-value of the test using Excel:
   $p = TTEST(range\_beef; range\_poultry; tails; type)$
   In our case we have a two tailed test hence $tails = 2$ and it is un-paired, hence $type = 3$.
3. Having computed $p = .00014$ and $\alpha = .05$ we can safely reject H_0 $\Rightarrow$ we can conclude that there are extremely significant difference in calorie count for beef hotdogs vs poultry hotdogs.

Analysis of Variance (ANOVA)

- We are interested in comparing “n” series of independent data (Single Factor Anova)

Analysis of Variance (ANOVA)

- We are interested in comparing "n" series of independent data (Single Factor Anova)
- Formulate the statistical hypothesis:
  - $H_0 : \mu_1 = \mu_2 = \cdots = \mu_n$
  - $H_a : \mu_1 \neq \mu_2 \neq \cdots \neq \mu_n$

Example

Below you find the salaries of people who have a degree in economics, medicine or history.

economics	medicine	history
42	69	35
53	54	40
49	58	53
53	64	42
43	64	50
44	55	39
45	56	55
52		39
54		40

Example

Below you find the salaries of people who have a degree in economics, medicine or history.

1. We define the statistical hypothesis:

Example

Below you find the salaries of people who have a degree in economics, medicine or history.

1. We define the statistical hypothesis:
   - $H_0 : \mu_{economics} = \mu_{medicine} = \mu_{history}$

Example

Below you find the salaries of people who have a degree in economics, medicine or history.

1. We define the statistical hypothesis:
   - $H_0 : \mu_{economics} = \mu_{medicine} = \mu_{history}$
   - $H_a : \mu_{economics} \neq \mu_{medicine} \neq \mu_{history}$

Example

Below you find the salaries of people who have a degree in economics, medicine or history.

1. We define the statistical hypothesis:
   - $H_0 : \mu_{economics} = \mu_{medicine} = \mu_{history}$
   - $H_a : \mu_{economics} \neq \mu_{medicine} \neq \mu_{history}$
2. We compute the $F - value$ and $F - crit$ of the test using Excel:
   We navigate to Data -> Data Analysis and choose Single Factor ANOVA
   Select as input range the table containing our values and choose $\alpha = 0.05$ and press OK.

Example

Below you find the salaries of people who have a degree in economics, medicine or history.

1. We define the statistical hypothesis:
   - $H_0 : \mu_{economics} = \mu_{medicine} = \mu_{history}$
   - $H_a : \mu_{economics} \neq \mu_{medicine} \neq \mu_{history}$
2. We compute the $F - value$ and $F - crit$ of the test using Excel:
   We navigate to Data -> Data Analysis and choose Single Factor ANOVA
   Select as input range the table containing our values and choose $\alpha = 0.05$ and press OK.
3. Having computed $F = 15.196$ and $F - crit = 3.4433$ we compare them and conclude that $F > F - crit \Rightarrow$ we can reject H_0
   $\Rightarrow$ we can conclude that at least one of the means is different.
   However, ANOVA does not tell us where the difference lies. For that you need to compute pairwise *t-tests*.

Testing categorical variables

Chi-Square (χ^2) test.

- Used in order to show whether or not there is a relationship between two categorical variables
- Can be used to test if the number of outcomes are occurring in equal frequencies or not

The χ^2 equation:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{1}$$

Here, χ^2 = the chi-square statistic, O_i = observed frequency, E_i = expected frequency

Example

Let us suppose that we have a dice and we roll it 600 times obtaining the following results:

Outcome	Frequency (O_i)
1	95
2	72
3	103
4	105
5	97
6	128

Example

Let us suppose that we have a dice and we roll it 600 times obtaining the following results:

Outcome	Frequency (O_i)
1	95
2	72
3	103
4	105
5	97
6	128

In this case the expected frequency (E_i) for each of the numbers would be 100.

Example

Let us suppose that we have a dice and we roll it 600 times obtaining the following results: In this case the expected frequency (E_i) for each of the numbers would be 100. Our table becomes:

Outcome	Frequency (O_i)	Expected freq. (E_i)
1	95	100
2	72	100
3	103	100
4	105	100
5	97	100
6	128	100

Example

Let us suppose that we have a dice and we roll it 600 times obtaining the following results: In this case the expected frequency (E_i) for each of the numbers would be 100.
Putting the test together we get :

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} =$$
$$= \frac{(95-100)^2}{100} + \frac{(72-100)^2}{100} + \cdots + \frac{(128-100)^2}{100}$$
$$= 16.36$$

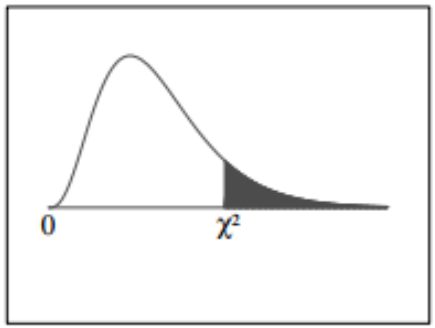
Example

Let us suppose that we have a dice and we roll it 600 times obtaining the following results:

- Having computed $\chi^2 = 16.36$ we still do not know if the value is significant or not
- We compute degree of freedom: throughout this course we compute degree of freedom as being $n - 1$ where n is the sample size
  $\Rightarrow deg.free = 6 - 1 = 5$
- Knowing all the information we check the chi-square table to find the critical value

Example

Chi-Square Distribution Table



The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718

Example

Let us suppose that we have a dice and we roll it 600 times obtaining the following results:

- We computed $\chi^2 = 16.36$
- We retrieved from table for 5 deg. fr. and $\alpha = 0.05$ the critical value 11.07
- Since our computed value is greater than the critical value we conclude that the dice seems to be rolled not fairly $\Rightarrow$ we reject the null hypothesis.