Statistical tests - II

Dr. Tudor Călinici

2023

Chi square test

- Is used to test the independence (association) between qualitative variables

- The null hypothesis: there is no association between variables
- The alternative hypothesis: there is association between variables

Chi square by example – dichotomial variables

- We are searching for hypothetical association between smoking and lung cancer. For that, from oncology department records, we select a group of 160 people who were diagnosed with lung cancer. We also select a control group of 240 persons which were never diagnosed with lung cancer. The study of this sample leads us to the following contingency table:

	Cancer present	Cancer absent	Total
Smoking yes	80	50	130
Smoking No	80	190	270
Total	160	240	400

The hypotheses, level of significance, critical region

- H0: there is no association between smoking and lung cancer
- H1: there is association between smoking and lung cancer

- We'll use a level of significance of 5%
- Critical region for 5% level of significance is $[3.84, +\infty)$

The parameter of the test

- The test consists by the comparation of the observed contingency table with the theoretical contingency table. The theoretical contingency table contains the theoretical distribution of the data when null hypothesis should be true

	Cancer present	Cancer absent	Total
Smoking yes	a^o	b^o	a^o+b^o
Smoking No	c^o	d^o	c^o+d^o
Total	a^o+c^o	b^o+d^o	$a^o+b^o+c^o+d^o$

Observed contingency table

	Cancer present	Cancer absent	Total
Smoking yes	a^t	b^t	a^t+b^t
Smoking No	c^t	d^t	c^t+d^t
Total	a^t+c^t	b^t+d^t	$a^t+b^t+c^t+d^t$

Theoretical contingency table

Computing the theoretical table

- The null hypotheses change the distribution of the data in the contingency table, but the totals will remain the same

	Cancer present	Cancer absent	Total
Smoking yes	$(a^o+c^o) * (a^o+b^o) / (a^o+b^o+c^o+d^o)$	$(b^o+d^o) * (a^o+b^o) / (a^o+b^o+c^o+d^o)$	a^o+b^o
Smoking No	$(a^o+c^o) * (c^o+d^o) / (a^o+b^o+c^o+d^o)$	$(b^o+d^o) * (c^o+d^o) / (a^o+b^o+c^o+d^o)$	c^o+d^o
Total	a^o+c^o	b^o+d^o	$a^o+b^o+c^o+d^o$

$$a^t=(a^o+c^o) * (a^o+b^o) / (a^o+b^o+c^o+d^o)$$

$$b^t=(b^o+d^o) * (a^o+b^o) / (a^o+b^o+c^o+d^o)$$

$$c^t=(a^o+c^o) * (c^o+d^o) / (a^o+b^o+c^o+d^o)$$

$$d^t=(b^o+d^o) * (c^o+d^o) / (a^o+b^o+c^o+d^o)$$

Compute the theoretical table

	Cancer present	Cancer absent	Total
Smoking yes	$\frac{160 * 130}{400}$	$\frac{240 * 130}{400}$	130
Smoking No	$\frac{160 * 270}{400}$	$\frac{270 * 240}{400}$	270
Total	160	240	400

	Cancer present	Cancer absent	Total
Smoking yes	52	78	130
Smoking No	108	162	270
Total	160	240	400

Chi square value

	Cancer present	Cancer absent	Total
Smoking yes	a^o	b^o	a^o+b^o
Smoking No	c^o	d^o	c^o+d^o
Total	a^o+c^o	b^o+d^o	$a^o+b^o+c^o+d^o$

	Cancer present	Cancer absent	Total
Smoking yes	a^t	b^t	a^t+b^t
Smoking No	c^t	d^t	c^t+d^t
Total	a^t+c^t	b^t+d^t	$a^t+b^t+c^t+d^t$

	Cancer present	Cancer absent	Total
Smoking yes	80	50	130
Smoking No	80	190	270
Total	160	240	400

	Cancer present	Cancer absent	Total
Smoking yes	52	78	130
Smoking No	108	162	270
Total	160	240	400

$$Chi\ square = \frac{(a^o - a^t)^2}{a^t} + \frac{(b^o - b^t)^2}{b^t} + \frac{(c^o - c^t)^2}{c^t} + \frac{(d^o - d^t)^2}{d^t}$$

$$Chi\ square = \frac{(80 - 52)^2}{52} + \frac{(50 - 78)^2}{78} + \frac{(80 - 108)^2}{108} + \frac{(190 - 162)^2}{162} = 37.2$$

Decision

- If Chi square is in the critical region, we can reject H0 and accept h1
- If Chi square is not in the critical region, we cannot reject H0 and cannot demonstrate H1

- Chi square = 37.2 in [3.84, +∞) , we can reject H0 and accept H1
- With 95% of confidence there is association between smoking and lung cancer

Chi square using statistical software

- If you use a statistical software to apply Chi square test, you can make the decision by interpreting p value
- If $p <= 0,05$ you can reject H0 and accept H1 with 95% of confidence
- If $p > 0,05$ you cannot reject H0

- For our example $p = 0,00000000105$

Chi Square Corrections

When we apply Chi square?

- **Two groups**
- **A qualitative variable of interest**
- **The association between two qualitative variables**
- **Comparing the observed distribution with theoretical distribution**

Cochran rule

	G1	G2	…	Total
E1	a	b	…	a +c+…
E2	c	d	…	c+d+…
…	…	…	…	…
Total	a +c+…	b +d+…	…	a+b+c+d+…

$$Chi^2 = \sum \frac{(O-T)^2}{T}$$

- **Chi square test result is valid only when 80% of theoretical frequencies are more than 5 and all theoretical frequencies are more than 1**

Theoretical contingency table – COCHRANE rule

	Group 1	Group 2	Total
Var 1	a^t	b^t	a^t+b^t
Var 2	c^t	d^t	c^t+d^t
Total	a^t+c^t	b^t+d^t	$a^t+b^t+c^t+d^t$

Theoretical table

All Values >5 – Chi square test

More values between 0 and 5– Fisher Exact test

One value between 2 and 5– Yates Corrected Chi Square

Yates corrected chi-square

- **Yates correction, involves reducing by 0.5 units the difference between the observed and the probable frequency within the Chi square before squaring.**

$$\boldsymbol{Chi\ square\ = \sum \frac{(O - T - 0,5)^2}{T}}$$

Example 1

- We are studying the hypothetical association between alcohol consumption and the occurrence of hepatic cirrhosis. For that, in a sample of 32 persons we have the following distribution. What statistical test should we use?

	Cirrhosis yes	Cirrhosis no	Total
Alcohol Yes	8	4	12
Alcohol No	6	14	20
Total	14	18	32

	Cirrhosis yes	Cirrhosis no	Total
Alcohol Yes	5,25	6,75	12
Alcohol No	8,75	11,25	20
Total	14	18	32

All theoretical frequencies > 5
Chi Square test

P=0,04

Example 2

- We are studying the hypothetical association between alcohol consumption and the occurrence of hemochromatosis. For that, in a sample of 25 persons we have the following distribution. What statistical test should we use?

	Hemochromatosis yes	Hemochromatosis no	Total
Alcohol Yes	4	3	7
Alcohol No	3	15	18
Total	7	18	25

P (uncorrected) = 0,04

	Hemochromatosis yes	Hemochromatosis no	Total
Alcohol Yes	1,96	5,04	7
Alcohol No	5,04	12,96	18
Total	7	18	25

P (Yates Corrected) = 0,05

One theoretical frequency <5
Yates corrected Chi Square test

Example 3

- We are studying the hypothetical association between alcohol consumption and the occurrence of Wilson disease. For that, in a sample of 11 persons we have the following distribution. What statistical test should we use?

P (Fisher Exact test) = 0,55

	Wilson Disease Yes	Wilson Disease No	Total
Alcohol Yes	2	3	5
Alcohol No	1	5	6
Total	3	8	11

	Wilson Disease Yes	Wilson Disease No	Total
Alcohol Yes	1,36	3,64	5
Alcohol No	1,64	4,36	6
Total	3	8	11

More than one theoretical frequency <5 – Fisher exact test

Quantitative analysis
ANOVA TEST.

Choosing the appropriate test

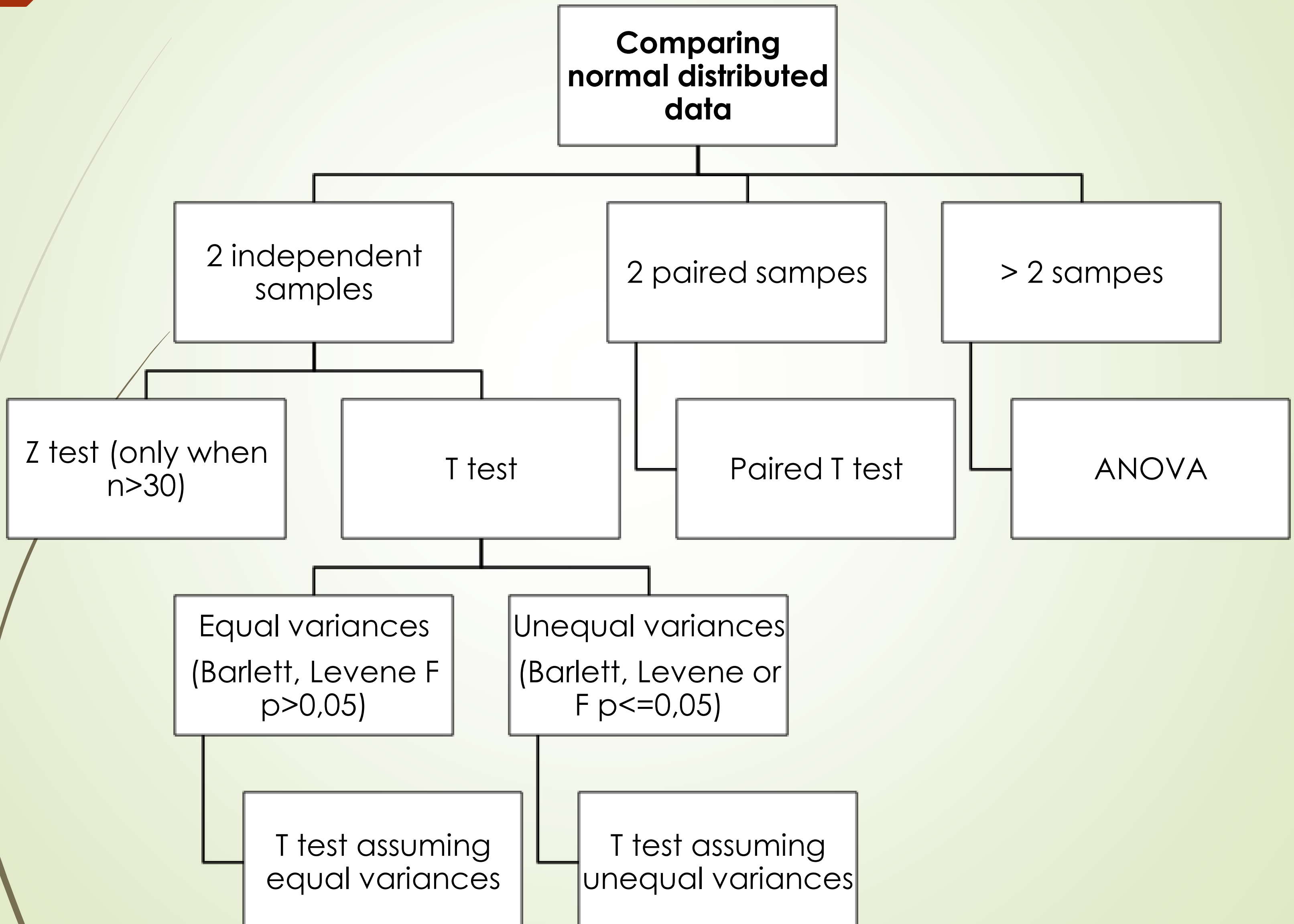
ANOVA TEST

- It is usually used to analyze data obtained from three or more than three samples
- Parametrical test
- Conditions:
  - Independent samples
  - Normal distributed data
  - Equality of the variances between groups
  - No extreme values

Example

- We want to check the effect of different kinds of diets in order to lose weight. For that, we use three samples – people which drink black tee, people which drink green tee and a control group –no tee. The weight of the persons was measured before and after a two months
- H0- there is no difference according to the lose weight between groups(the tee diets have no effect)
- H1 - there is difference according to the lose weight between groups (at least one tee diet has effect)

Choosing the test– quantitative normal distributed data



Non-parametrical tests

- Quantitative variables, data are not normal distributed
- For each parametrical test we have the non-parametrical equivalent

Parametrical and non-parametrical tests

Parametrical test	Non-parametrical equivalent test
Z, ANOVA	Kruskal-Wallis
Student (independent samples) First apply test for variances F, Levine, Barlet, etc.	Mann–Whitney U Mann–Whitney–Wilcoxon Wilcoxon–Mann–Whitney Wilcoxon rank-sum
Student (paired samples)	Wilcoxon signed-rank

Choosing the test – quantitative variables