

# Statistical tests

## Intro to Statistical Inference Part 2

Chuvakin Sergey

«School of Advanced Studies»

December 16, 2020

# Outline

- ▶ Compare more than two groups
- ▶ Chi-squared test
- ▶ Correlations

## Compare more than two groups

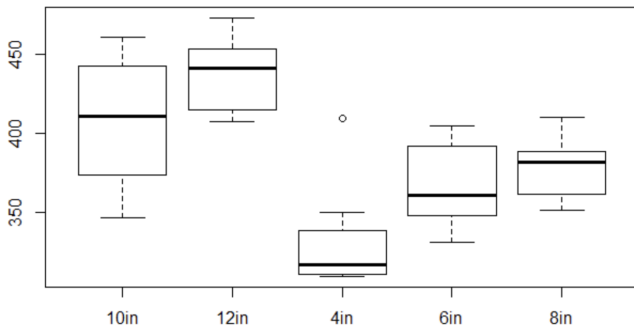
ANOVA - ANalysis Of VAriance - statistics that help to point out the difference between two or more groups. Detailed information [here](#)

⚠ Variances of groups should be homogenous

⚠ The post hoc test is required

## Compare more than two groups

Post hoc test - is a test that usually conducted after analysis («post hoc»), to precise inferences. I.e. it could be box plots!



# Chi-Square test

$\chi^2$  - Popular statistics to test if matrix has non uniform distribution.  
It return Chi square that can be easily transformed to P-value.  
Let's use example for explanation!

H0 - Gender and preference for cats or dogs are independent.

H1 - Gender and preference for cats or dogs are not independent.

# Chi-Square test

	Cat	Dog
Men	207	282
Women	231	242

# Chi-Square test

	Cat	Dog	
Men	207	282	489
Women	231	242	473
	438	524	962

## Chi-Square test

	Cat	Dog	
Men	$\frac{489 \times 438}{962}$	$\frac{489 \times 524}{962}$	489
Women	$\frac{473 \times 438}{962}$	$\frac{473 \times 524}{962}$	473
	438	524	962



# Chi-Square test

	Cat	Dog	
Men	222.64	266.36	489
Women	215.36	257.64	473
	438	524	962

# Chi-Square test

Subtract expected from observed, square it, then divide by expected:

In other words, use formula

$$\frac{(O - E)^2}{E}$$

where:

O = Observed (actual) value

E = Expected value

# Chi-Square test

	Cat	Dog	
Men	$\frac{(207-222.64)^2}{222.64}$	$\frac{(282-266.36)^2}{266.36}$	489
Women	$\frac{(231-215.36)^2}{215.36}$	$\frac{(242-257.64)^2}{257.64}$	473
	438	524	962

Which gets us:

	Cat	Dog	
Men	1.099	0.918	489
Women	1.136	0.949	473
	438	524	962

## Chi-Square test

Now add up those calculated values:

$$1.099 + 0.918 + 1.136 + 0.949 = 4.102$$

*Chi-Square* is 4.102

Than look into table [here](#).

Make inference!

# Correlation

Correlation - is the most popular way to answer the question: what is the relation between two variables. It returns metrics between 0 and 1 that shows how two vector move together. It's standartized version of covariance

- ▶ Reference point:  $\text{cov } x,y=0$  means X and Y have nothing incommon
- ▶ Positive covariance values indicate positive association: the bigger X the bigger Y
- ▶ Negative covariance values indicate negative association: thebigger X the smaller Y

# Correlation

One of the assumptions of correlation is - variables should be measured in one space. For example we can compare number of death per capita and per 100 000.

Standardization is a general method of transforming input variables (e.g.  $X$  or  $Y$ ) or some statistical measures of interest (e.g. covariance coefficient) in a way that eliminates scale effects.

E.g., standardizing a normally distributed variable means converting it to the standard normal distribution (with zero mean and unit standard deviation). After standardization, this variable is no longer measured in meters, years, or dollars, but in standard deviations

# Correlation

Generic standardization formula (Z-score formula):

$$z_i = \frac{x_i - \hat{x}}{\sigma_x}$$

Covariance

$$\text{Cov}(X, Y) = E[(X - \mu_x)(Y - \mu_y)]$$

Correlation

$$r = \frac{\text{cov}_{x,y}}{\sigma_x \times \sigma_y}$$

# Correlation

## Assumptions

- ▶ Continuous variables
- ▶ Normal distribution



# Correlation

## Assumptions

Otherwise use non-parametric alternatives!

- ▶ Spearman's  $\rho$  (also reads as rho)
- ▶ Kendall's  $\tau$  (reads as tau)
- ▶ Interpretation is pretty much the same as with Pearson's  $r$ : same effect size cut-offs, same p-value thresholds
- ▶ Use the method = argument of `cor()` or `cor.test` to choose a relevant correlation coefficient (three options: `pearson`(default), `spearman`, and `kendall`, all should be typed in lower case)