1 Overview of hypotheses tests

1.1 Tests with one sample

Parametric tests (normality required)

- expectation (known × unknown variance) - test of true average
  Ex: *A company declares that its production is more than 150 products per day. Somebody opposes and says that it is less.*
  Prg: z_test, t_test
- proportion - test of a part from the whole
  Ex: *City manager says that only 5% of drivers exceed the permitted speed at certain street. Police is convinced that the ratio is higher.*
  Prg: prop_test
- variance - test of variability of a variable
  Ex: *Quality of production is given by the dispersion of weight of products is. If it is higher then a given level, the machines must be adjusted. Test, if the machines are OK or it is necessary to tune them.*
  Prg: var_test

Nonparametric tests (normality is not required)

- Wilcoxon test: tests median of rv from one sample
  – H0: median is equal to the assumed value
  – test is all sided
  Ex: *Compare caloric intake measured at 11 selected women with the recommended value 7725 kJ.*
  Prg: wilcoxon_test

Tests of distribution type

- w/s test of normality (statistics = range / std)
  – H0: rv is normal
  Prg: norm_test
- Kolmogorov-Smirnov test: tests given distribution. It is based on comparison of assumed and empirical DF.
  – H0: rv has assumed distribution
  – right sided test with special crit. vals
  Prg: ks_test_cont, ks_test_disc, ks_test_2 (comp.of two distr.)
- Chi-square test of homogeneity: test of distribution type. It compares observed and expected frequencies.
  – H0: rv has the assumed distribution
  – right sided test
  Ex: *We have measured number of accidents for week days and weekends. Test if they are uniformly distributed.*
  Prg: chisquare_test

1.2 Tests with two samples

Parametric

- two expectations (independent × paired samples)
  Ex (indep): *Company A claims that its production is greater than that of B. Assistant of company B denies it. Test.* ... how to determine the side.
  Prg: t_test_2s, t_test_2n
  Ex (paired): *Uniformity of tire removal at the front wheels of cars of a specific mark has been investigated. The producer of the cars proclaims uniformity. Test it.*
  Prg: t_test_2p
- two proportions
  Ex: *Ratio of drivers violating rules in town is greater than outside. Test it.*
  Prg: prop_test_2
- two variances
  Ex: *Variability of weights of products from company A is greater than those, from company B. Test it.*
  Prg: var_test_2

Nonparametric

- Mann-Whitney test: tests equality of two medians (independent samples)
  – H0: the medians are equal
  – both sided test
  Ex: *Marks form math were checked at two classes of secondary school. 5 marks from the first and 8 marks from the second class were recorded. Compare the classes.*
  Prg: mannwhit_test
- Wilcoxon: tests two medians (paired samples)
  – H0: medians are equal
  – all sided test
  Ex: *At a secondary school an improvement of students in math was checked. In the 1st class eight students were selected and their marks recorded. In the 2nd class the marks of the same students were recorded again. Test, if the results of individual students are improved.*
  Prg: wilcoxon_test
- McNemar: tests improvement after some action. Data are yes/no - two by two table of frequencies.
  – H0: no improvement
  – right sided
  Ex: *22 selected people were tested for cold (yes/no). Then, they received some drug and after a week they were tested again. Test the effectiveness of the drug.*
  Prg: mcnemar_test

1.3 Tests with more samples

Parametric

- Analysis of variance: tests equality of several expectations
  – H0: expectations are equal
  – right sided test
  Ex: *Test if the power of engine of vehicles of five marks is the same.*
  Prg: anova_1

- Anova with two factors: tests equality in columns and rows.
  Ex: *Five cars are tested by three drivers. Test the cars and the drivers.*
  Prg: anova_2

Auxiliary tests to anova

- Bartlett - test of equality of more variances
  Prg: bartlett_test

- Scheffé - detects different samples
  Prg: scheffe_test

Nonparametric

- Kruskal-Wallis: nonparametric anova.
  – H0: medians are equal
  – right sided test
  Ex: as for anova1
  Prg: kruskal_test

- Friedman - block test of equality of medians
  – H0: medians are equal
  – test is right sided
  Ex. *5 shops are rated by 3 inspectors (each shop is rated by each inspector; inspectors are factors of no interest = block). Evaluate quality of the shops.*
  Prg: friedman_test

1.4 Tests of independence

- Gamma coefficient: test of association of two discrete rvs. It compares prediction from marginal and conditional pf.
  – result: how many times the prediction from cond. pf is better than from marginal.
  Ex: *We measure speed and consumption on driven cars. Is there a relation between these two variables?*
  Prg: gamma

- Pearson test: tests independence of two rvs. It tests correlation coefficient. (parametric test)
  – H0: rvs are independent
  – test is both sided
  Ex: *Test the data x and y if they are suitable for linear regression.*
  Prg: pearson_test

- Spearman test: nonparametric Pearson. Works with ranks.
  – H0: rvs are independent
  – test is both sided
  Prg: spearman_test

- Chi-square test of independence: test if independence of two rvs. Compares observed and expected frequencies. Based on the definition of independence $f(x,y) = f(x) f(y)$.
  – H0: rvs are independent
  – test is right sided.
  Ex: *We asked 200 people from three different areas about they pay (low, normal, high). Test if the pay depends on the area.*
  Prg: chisqare_test_i

2 Validation in regression analysis

Regression can be viewed as approximation of dependence of y on x from data sample by some curve - linear, exponential, polynomial etc. However, not each data sample must be convenient for such approximation. Here we will discuss this question.

1. Draw xy-graph: ideal, good, possible and no good regression.

2. Pearson t-test of correlation coefficient

   For approximation of a relation between x and y there mus be any relation. This is expressed in **regression coefficient**

   $$\rho = \frac{C[X,Y]}{\sqrt{D[X]\,D[Y]}} \quad \longleftrightarrow \quad r = \frac{S_{xy}}{\sqrt{S_x S_y}}$$

   where C is covariance, $D$are variances, S are sums

   $$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}), \;\; S_x = \sum (x_i - \bar{x})^2, \;\; S_y = \sum (y_i - \bar{y})^2$$

   The true property of random variables is expressed in population regression coefficient ρ. Its true value is estimated from sample by the statistics r (sample regression coefficient).

   Pearson t-test has H0: $\rho = 0$, HA: $\rho \neq 0$; both sided test with Student distribution.

   H0: x and y are uncorrelated - regression does not have sense. To be able to use regression, H0 has to be rejected.
   Prg: pearson_test

3. Fisher F-test of explained and unexplained variance

   In regression, we have data and predictions of data which lie on the regression line. If we want to characterize data $\{y_i\}_{i=1}^N$ without regression, we can compute the average value $\bar{y}$. Then, for a selected x_i we have the value y_i and its prediction $\hat{y}_i$. Now, the deviation of y_i from $\bar{y}$ can be decomposed as

   $$y_i - \bar{y} = \underbrace{(\hat{y}_i - \bar{y})}_{\text{expl.}} + \underbrace{(y_i - \hat{y}_i)}_{\text{unexpl.}}$$

   where

- $y_i - \bar{y}$ is the error in measurement without taking into account the regression (overall error),
- $\hat{y}_i - \bar{y}$ is a deviation from the average explained by regression (explained error),
- $y_i - \hat{y}_i$ is a deviation of the measured point from the regression line - if regression is precise, all points should lie no the line (unexplained error).

Taking variances, we obtain explained S_r (regression) and unexplained S_e (residual) variances. The statistics is defined as $F = \frac{S_r}{S_e}$ with F distribution. For H0: $F = 0$ is nothing explained and the regression does not have sense. The test is right-sided. Regression has sense, it H0 is rejected.

4. Test of independence of residuals

   Residuals are deviations of the data from regression line. For correct regression the residuals must bu independent. If not, the relations between them could be used to construct better regression line.

   The test has the statistics

$$z = \frac{2b - (n-2)}{\sqrt{n-1}} \sim N(0,1)$$

   where b is number of sequences (deviations from median with the same sign). H0: is independence (for $z = 0$).

5. Test for auto-correlation of residuals

   It is a similar test to the previous one. We test if a current residuum e_i can be estimated from the previous one e_{i-1}. We estimate the dynamical regression

$$e_i = ae_{i-1} + b + \epsilon_i$$

   If $|a| < 0.3$ and $k \to 0$, the regression is OK.

6. Standard error of residuals SE

   Residuals $e_i = y_i - \hat{y}_i$ are errors of approximation of data with regression curve. The smaller the errors are, the better approximation. The standard error is defined as

$$SE = \frac{var(e)}{var(y)}$$

   which is variance of prediction error e_i relative to variance of dependent variable y_i.