Help to Difficult Statistical Notions

Contents

1 Population, sample realization and random sample

Population

By population we mean a source of data (data generator) with some specific properties which are reflected in the generated data. The properties are expressed in probabilities not values themselves. E.g. "the probabilities of the generated values are equal on a certain interval" (uniform distribution) or "the data are generated from a fixed value and are affected by many small independent random errors" (normal distribution) or "only two values are generated with probabilities p and $1 - p$" (Bernoulli distribution).

- The population properties are fixed and apply to all generated data.
- The characteristics of population are fixed and they are computed by means of the probability (density) function by integration or summation over all possible data.

Example

1. *Speeds of cars measured in s specified point of the communication.*
2. *Severity of traffic accidents in a specified traffic region.*
3. *Queue lengths in the arms of a controlled crossroads.*

Sample realization

Sample realization is the set of data measured in the generator. It is set of values (vectors). Thus, the sample is fixed and also its characteristics (mean, variance etc.) are fixed.

- Sample realization is an ordinary dataset.
- Different data samples have different values, because the sampling is random.

Example

1. *A set of 50 values of the car speeds measured in a specified point.*
2. *A set of 25 records with severity of traffic accidents in a specified region. The severity is: 1 = light accident, 2 = serious accident, 3 = accident with injury or death.*
3. *A set of 100 vectors of the measured queue lengths in four arms of the crossroads.*

Random sample

Random sample is a vector of equally distributed and independent random variables. The sample realizations are values of this random vector.

Explanation

The sample realization can theoretically be repeated (even if in practice only one sample realization is always taken). When repeated, each sample realization differs from• the others (sampling is random). So, if we take e.g. the first position of the sample realization, we obtain different values in each sample realization. But this variability after each measurement is the main characteristics of a random variable. So, we can say, that in the first position of the random sample is a random variable whose realizations are the first numbers in each sample realization. And the same holds also for the remaining positions of the sample.

Condition "equally distributed" means, that all the measurements are taken always from the same population (e.g. speed measurements are measured at the same point).

Condition "independent" means that there are no preferences in measurements (e.g. all cars are measured not only Mercedes and Audi).

Consequence

Characteristics (mean, variance etc.) of random sample are random variables and as such they have also their characteristics (sample average, sample variance etc.)

Example (turning cars in T-junction)

Situation: Monitoring cars in T-junction at given time with results direction of turning.

Population (random variable): two possible values: 1-turning left, 2-turning right. The probabilities of turning are fixed and given by many circumstances, e.g. size of the regions in left / right direction, density and composition of population in the areas, workplaces of people from the regions and many others. These probabilities are not exactly known.

Random sample: measurement of n speeds of randomly chosen cars.

Sample realization: a set of n values taking by measuring the speeds of specific cars. Probabilities can be guessed as fractions of the numbers of cars turning to the left / right divided by the number of measurements n. They are not exactly the probabilities of the population but they are close to them.

2 Data ranks

Ranks of data are their orders in the sorted dataset.

We denote: x - data, s - sorted data, r - ranks.

For the data

$$x = [x_1, x_2, x_3, x_4] = [5.3,\ 2.8,\ 4.5,\ 1.7]$$

where

$$x_1 = 5.3,\ x_2 = 2.8,\ x_3 = 4.5,\ x_4 = 1.7$$

Sorted data are

$$s = [1.7,\ 2.8,\ 4.5,\ 5.3] = [x_4, x_2, x_3, x_1]$$

so the ranks (indexes of the sorted vector) are

$$r = [4,\ 2,\ 3,\ 1]$$

Repeated values

If the values of the data repeat, the rank is the average of the position of repeated values. E.g. for

$$x = [3,\ 5,\ 2,\ 5,\ 2,\ 2]$$

the sorted data and their indexes i are

$$s = \left[\ 2,\ 2,\ 2,\ 3,\ 5,\ 5\ \right]$$

$$i = [1,\ 2,\ 3,\ 4,\ 5,\ 6]$$

So, 2 spread over indexes 1, 2, 3 with the average equal to 2. The value 3 has position 4 and the average of indexes for 5 is 5.5. The ranks are

$$r = [2,\ 2,\ 2, 4, 5.5]\,.$$

Exercise

Determine ranks for

$$x = [3,\ 3,\ 2,\ 4,\ 3,\ 5,\ 1,\ 3,\ 2,\ 3]$$

Result

$$r = [6,\ 6,\ 2.5,\ 9,\ 6,\ 10,\ 1,\ 6,\ 2.5, 6]$$

Consequence of ranking

1. Transformation of data to their ranks naturally suppresses outliers.
2. Ranks instead of data are frequently used in tests for data which do not come from normal distribution (nonparametric tests).

Remark

The effect of ranking lies in this: A particular type of distribution is characterized by the fact that its realizations are denser in some areas than in others. It means, that in denser areas the data points are closer to one another than in others. This means that in more densely populated areas, data points are closer together than in others. The ranking suppresses these specific distances between data points, and therefore the developed methods are valid regardless of the type of distribution.

3 Moments

Moments are important characteristic of data as well as random variables. Moments of data correspond to measured values which form a sample realization. Moments of random variable relate to population.

In estimation theory and hypothesis testing, we use measured data to draw inferences about population parameters. We define statistics, a function of data, whose values point at the estimated parameter. The basic parameters are population characteristics (expectation, variance, proportion) and the statistics are the corresponding sample characteristics. That is why their knowledge is very important.

Raw moments (of order r)

Data	Continuous r. variable	Discrete r. variable
$\frac{1}{n}\sum_{i=1}^{n} x_i^r$	$\int_{-\infty}^{\infty} x^r f(x)\,dx$	$\sum_{x_i \in X} x_i^r f(x_i)$

Especially: first raw moment is

- sample average

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- expectation of continuous variable

$$E[X] = \int_{-\infty}^{\infty} x f(x)\,dx$$

- expectation of discrete variable
  (if $x_i \in \{0, 1\}$ we call it proportion)

$$E[X] = \sum_{x_i \in X} x_i f(x_i)$$

Central moments (of order r)

Data	Continuous r. variable	Discrete r. variable
$\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^r$	$\int_{-\infty}^{\infty} (x - E[X])^r f(x)\,dx$	$\sum_{x_i \in X} (x_i - E[X])^r f(x_i)$

Especially: second central moment is

- second moment for data

$$s^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

sample variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

which is unbiased estimate of variance

- second moment of continuous variable

$$D\left[X\right] = \int_{-\infty}^{\infty}\left(x - E\left[X\right]\right)^2 f\left(x\right)dx$$

- second moment of discrete variable

$$D\left[X\right] = \sum_{x_i \in X}\left(x_i - E\left[X\right]\right)^2 f\left(x_i\right)$$

Second mutual moment (covariance)

- data covariance

$$\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)$$

- covariance of continuous variables

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(x - E\left[X\right]\right)\left(y - E\left[Y\right]\right) f\left(x,y\right)dxdy$$

- covariance of discrete variables

$$\sum_{x_i \in X}\sum_{y_i \in Y}\left(x_i - E\left[X\right]\right)\left(y_i - E\left[Y\right]\right) f\left(x_i, y_i\right)$$

Correlation coefficient

$$\frac{\text{covariance}\left(x,y\right)}{\sqrt{\text{variance}\left(x\right)\text{variance}\left(y\right)}}$$

Use in estimation and testing

parameter		**statistics**
expectation	$\rightarrow$	sample average
variance	$\rightarrow$	sample variance
proportion	$\rightarrow$	sample proportion
independence	$\rightarrow$	sample covariance

4 The statistics for estimation

The first thing we need to realize is this:

We have some population described by density (probability) function $f(x, \theta)$ with an unknown parameter θ. What does it mean?

We have some experiment (mostly monitoring some variable on a real system) from which we get data. The experiment is random and the data are produced according to some inner rule which is expressed by density (probability) function, specifically by the parameter θ. The distribution is known to us up to the parameter θ - e.g. the distribution is normal with know variance but the expectation (which is position) is unknown.

Example: We measure speeds of passing cars at a place with speed restricted to 80 km/h. If we would be able to take into account all cars (in the past, present time and future) we could construct the density function of the speeds and to determine the real expectation of the random variable "speed of the passing cars". But this is only a fiction. We will never be able to do such monitoring.

So, we have to estimate!

As we have said, the form of th distribution is assumed known (it is the distribution of the data), only the parameter is unknown and has to be estimated.

Let X be the random variable (experiment from which we get data) and $f(x, \theta)$ be its distribution, with θ being an unknown parameter (e.g. expectation) which we want to estimate.

Example: Let the true distribution of the speeds has normal form with expectation $\mu = 79$ and the variance $\sigma^2 = 8$. Let us suppose that the expectation can be considered known (it is given by the restriction) but the variance (which speaks about general keeping the restricted speed) is unknown and has to be estimated. So the population distribution is $f(x, \theta) = N_x\left(79, \sigma^2\right)$.

How to estimate?

Example: The random variable is just the generator of the data - here $N_x(79, 8)$. *Notice: It is constant - does not depend on sampling.* From it we can take a sample realization: say $x = 80$, 79, 78, 77, 82. *Notice: if we repeat sampling, we surely obtain different sample realization*[1].

Now, important: The data themselves do not point at the estimated variance. To be able to get information about the variance from the data, we need to transform them. This transformation of the data sample gives a function whose values point at the estimated parameter is called

[1] However, in practice we take only one sample. If we want more data, we add them to the original one. The repetitive sampling is only theoretical and it shows that the sample realization is only random and instead this one another sample could have been chosen. That is why the information brought by a sample realization is not precise.

statistics. In our case the function will be the sample variance whose general form is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

So, the values of the statistics with the above sample will be

$$T(x) = \frac{1}{5-1}\left[(82-79)^2 + (77-79)^2 + \cdots + (82-79)^2\right] = 9.5,$$

where 79 is the average. We can see that the value of the statistics is near to the true values of the population variance (that is 8), even if the values of x themselves are quite different.

Now comes a very important piece of information !!!

As we have already said, if we measure new samples and calculate the sample variance from each of them, we would get different values of the statistics. And we also said that a variable that gives different values after each measurement is the random variable. So, in general, the statistic is a random variable. Its values are its realizations. And because statistics is a random variable, we can talk about its distribution.

Thus, we have a random variable X (population) from which we measure data x_i that form the sample realization $x = [x_1, x_2, \cdots .x_n]$. The sample is transformed to the statistics T whose values point at the estimated parameter θ. The estimate of this parameter is simply given by the statistics with the sample realization substituted $\hat{\theta} = T(x)$. Both, the data generator X with values x and the statistics T with values t are random variables and have their distribution $f(x)$ and $f(t)$. Thus we can determine

- probability that data x are from an interval, say (a, b) is

  $$\int_a^b f(x)dx$$

  where distribution of data $f(x)$ is used

- probability that a parameter θ is from an interval, say (c, d) is

  $$\int_c^d f(t)\,dt$$

  where the statistics T with the distribution $f(t)$ is used.

Consequence

The confidence intervals and tests of hypotheses deal with parameters. they are based on intervals where the parameters occur with given probabilities. So, their derivations, concerning critical region or p-value always deal with the **distribution $f(t)$ of the statistics T**.

5 Properties of the statistics

Definition: Statistics is the function of random sample.

It means - statistics T for estimation of parameter θ is a formula which after inserting the sample realization produces value which is near to the parameter θ.

It should have the following properties

1. Is **unbiased**
   It holds

   $$E(T) = \theta$$

   e.g. for sample average $T = \bar{x}$ and expectation $\theta = \mu$: the average from all possible sample averages (made from all possible sample realizations) is exactly equal to μ.

2. Is **consistent**
   For unbiased estimate T_n where n is the sample length it holds

   $$\lim_{n\to\infty} D[T_n] = 0$$

   i.e. for sample length n going to infinity the estimate is precise.
   For sample average:

   $$\lim_{x\to\infty} D(\bar{x}) = \lim_{x\to\infty} \frac{\sigma^2}{n} = 0$$

3. Is **efficient**
   For comparison of two unbiased statistics it holds: The statistics with smaller variance is better (more efficient).
   For sample variance

   $$D[X_1] < D[X_2] \rightarrow \frac{\sigma^2}{n_1} < \frac{\sigma^2}{n_2}$$

   if $n_1 > n_2$.

6 What is p-value

There are two basic ways how express results of testing. They are (i) critical region and realized statistics and (ii) p-value. The letter is preferred as it expresses also the strength of rejection (or not rejection).

Basic notions:
The distribution of data is $f(x,\theta)$ and point estimate of θ is T which is a function of the sample x, i.e. $x \to T$ (e.g. $\bar{x} = \frac{1}{n}\sum_i x_i$). When we apply this transformation on the data distribution we get the distribution of the statistics T (which of course also depends on θ)

$$f(x,\theta) \to f(T,\theta)$$

Similarly as for data it holds

$$\int_a^b f(x,\theta)\,dx = P(x \in (a,b))$$

it also holds for the statistics

$$\int_c^d f(T,\theta)\,dT = P(T \in (c,d))$$

and as $\hat{\theta} = T$ the last probability holds also for the estimate $\hat{\theta}$. I.e. the parameter estimates are located by the distribution of the statistics.

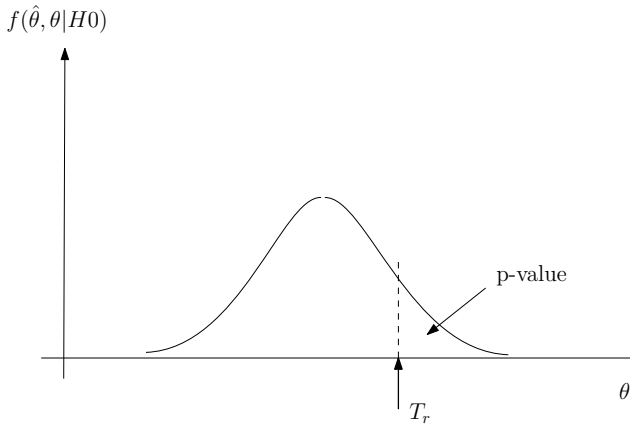
Critical region and realized statistics We construct a confidence interval corresponding to the task (parameter, side etc.). The **critical region** W is a complement of the confidence interval. The **realized statistics** T_r is the value of the statistics with the sample realization inserted.

Then it holds, if $T_r \in W$ we reject H0. Otherwise, we do not reject.

Remark

In practice we formally operate with normalized statistics. I.e. for sample average $\bar{x}$ and right sided test and known variance we have: Critical region is $W = (z_\alpha, \infty)$ and the normalized statistics is $T_r = \frac{\bar{x}-\mu_0}{\sigma}\sqrt{n}$

p-value We will show the definition and meaning of the $p$value for right-sided test in the following picture



As we have shown, the statistics distribution $f(T, \theta)$ depends on the value of the unknown parameter θ. If we substitute for it the value θ_0 according to H0, then the distribution describes location of the parameter according to H0. The HA (for right-sided test) says that the parameter is greater than H0 claims. It means, according to H0 the realizations of T, which are point estimates $\hat{\theta}$ should mostly lie under the peak of the density or to the left. On the other hand, if HA is valid, they should be somewhere more to the right. From the picture we can see that the more to the right the statistics lies the smaller is the area under the density and to the right from the realized statistic T_r.

The **p-value** is the area under the $f(T, \theta)$ from the realized statistics T_r to the right. Its definition is

$$p\text{-value} = P(T > T_r|H0)$$

And it is clear, that the smaller the p-value is the more strictly we reject the H0.

The above definition holds for right-sided test. In the case of left-sided test, the definition is symmetrical

$$p\text{-value} = P(T < T_r|H0)$$

The case of both-sided test is a bit more complicated. The HA says "is not equal" which mean is greater or smaller. But our general assumption is that it can be both - sometimes greater and sometimes smaller. What is reflected in our sample is accidentally one of these cases. So we have to compute both the p-values for $\alpha/2$: right-sided p-value$_R$ and left-sided p-value$_L$. We take the smaller one and multiply by 2. So it is

$$p\text{-value} = (p\text{-value}_R + p\text{-value}_L)/2$$

If we want to use the confidence level α we can do it as follows

If p-value $< \alpha$, reject H0
Otherwise, do not reject.

7 Side of interval or test

H0 always says $\theta = \theta_0$ where θ_0 is the value of the parameter according to H0.

Remark: sometimes we can say e.g. H0: the variance is less than θ_0. However, we mean $\theta = \theta_0$ and we want to stress, that the opposite should be HA: $\theta > \theta_0$.

HA opposes H0.

The direction is always given by the HA

HA: $\theta \neq \theta_0$ - both-sided

HA: $\theta > \theta_0$ - right-sided

HA: $\theta < \theta_0$ - left-sided.

The only difficulty can be if we test two expectations. Then we must say which sample is first and which is second.

Example

Let A is first and B second. What we test is the difference between them A - B.

Now if H0 says A $\geq$ B, then A - B > 0.

HA then is (the opposite), i.e. A - B < 0 and less means left-sided test (see above).

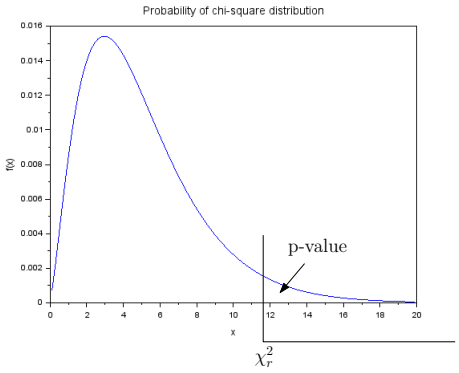
8 Chi-square test (and its variants)

For discrete or discretized variables.

Statistics

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \sim \chi^2(\nu)$$

where O_i are observed and E_i are expected absolute frequencies.



If $\chi^2 = 0$, p-value= 1 - $V_E = 0$ nothing is explained.
If $\chi^2 \to \infty$, p-value$\to$ 0 - $V_U \to 0$ all is explained.

Use:

– test of variance
– test of independence

This test can be used in several situations:

Goodness of fit test

Here we have one sample from a population and we test if the distribution of the population is that we assume.

Example for uniform distribution.

We have measured accident in weekdays, Saturday and Sunday. We obtained the following data

day	weekdays	Saturday	Sunday
number of accidents	53	8	12

Test the assertion (H0) that the accidents occur uniformly (each day).

Solution

The lengths of intervals are: 5, 1, 1. The total number of accidents is 53+8+12 = 63. Number of days is 7. The number of accidents per 1 day is $\frac{63}{7} = 9$. So the expected (uniform) accidents should be 5·9 = 45, 9 and 9.

$$\chi^2 = \frac{(53-45)^2}{45} + \frac{(8-9)^2}{9} + \frac{(12-9)^2}{9} = 2.53$$

$$pv = P\left(\chi^2 > 2.53\right) = 0.28$$

We do not reject uniformity.

Test of homogeneity

We have two samples taken from two subgroups of the population. One sample yields O and the second one E. H0 claims homogeneity of the whole population.

The test follows the previous case.

Test of independence

This test is based of the definition of independence

$$f(x,y) = f(x) \cdot f(y)$$

Example

We asked people from the North (N) and South (S) about their monthly pay grouped into three groups (I, II and III). We obtained data

residence/pay	I	II	III
N	53	128	91
S	345	187	69

Test the independence of pays and place of living.

The table is O observed frequency table. The total number of observations is $N = 873$. The table of relative frequencies (joint probability function) is

0.061	0.147	0.104
0.395	0.214	0.079

Marginals (sums over rows and columns)

$$f\left(res.\right) = \left[\begin{array}{c} 0.312 \\ 0.688 \end{array}\right] \text{ and } f\left(pay\right) = [0.456,\, 0.361,\, 0.183]$$

Their product forms joint probability for independent variables

$$f_n = \left[\begin{array}{c} 0.312 \\ 0.688 \end{array}\right] [0.456,\, 0.361,\, 0.183] = \left[\begin{array}{ccc} 0.142, & 0.113, & 0.057 \\ 0.314, & 0.248, & 0.126 \end{array}\right]$$

$$E = f_n N = \left[\begin{array}{ccc} 124.00, & 98.14, & 49.85 \\ 273.99, & 216.86, & 110.15 \end{array}\right]$$

Now, O and E (rearranged to a vector) can be inserted into the criterion and the statistics and p-value computed.

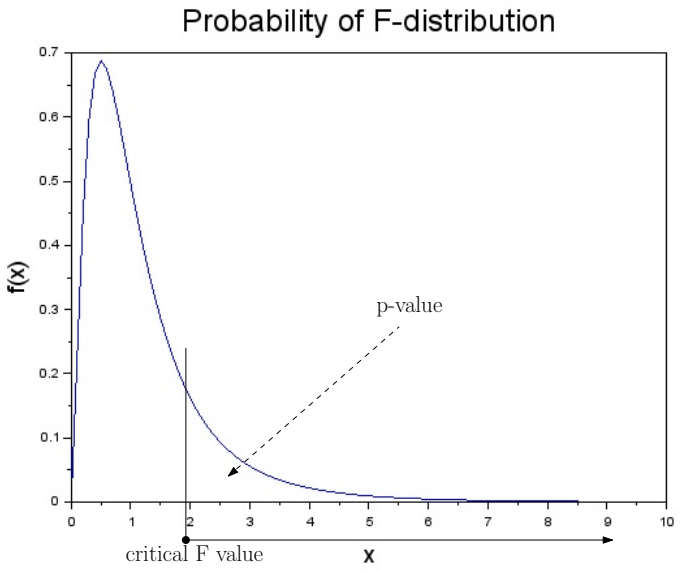
9 F test (and its use)

F test is used to compare ratio of two variances

$$\frac{D_E}{D_U} \sim F$$

where D_E is the explained variance and D_U is the unexplained variance (the specific meaning of these variance will be explained in the examples bellow).

The F distribution has the following form



H0: nothing is explained. The test is right-sided. With growing explained variance the statistics grows, too. If the p-value falls below α, H0 is rejected.

ANOVA I

We have data from several sources (populations). We test, if the expectations of the populations are equal.

The data are x in the following table as absolute frequencies

X_1	X_2	X_3
x	x	x
x	x	x
x	x	x
$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_3$
s_1^2	s_2^2	s_3^2

We compute averages $\bar{x}$ and variances D. Then:

- Average of variances s_i^2 correspond to **unexplained variance** D_U - it describes the overall variance in the data.
- Variance of the averages $\bar{x}_i$ corresponds to **explained variance** D_E - it expresses the variance between classes.

If the explained variance D_E is sufficiently larger with respect to the unexplained one D_U then we conclude that the classes are not equal.

Statistics: $F = \frac{D_E}{D_U} \sim F$ distribution (right-sided test)

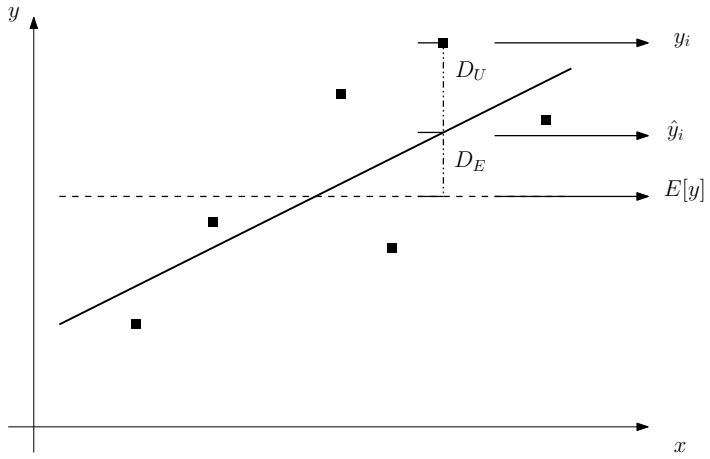
H0: are equal.
HA: are not equal.

ANOVA II

	X_1	X_2	X_3		
Y_1	x	x	x	$\bar{y}_1$	s_{y1}^2
Y_2	x	x	x	$\bar{y}_2$	s_{y2}^2
Y_3	x	x	x	$\bar{y}_3$	s_{y3}^2
Y_4	x	x	x	$\bar{y}_4$	s_{y4}^2
	$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_3$		
	s_{x1}^2	s_{x2}^2	s_{x3}^2		

Two tests - first for columns and second for rows.

Regression



Without the regression assumption, the data are distributed around their mean and their variance is calculated from that mean. Once the regression assumption is in place, the data should lie on the regression line and the variation of the predictions on the line and the average value is explained by the regression assumption. The deviation of the line from the mean constitutes the explained variance. However, the data do not lie on the line. Their deviations from the line are not explained and form the unexplained variance.

Statistics

$$F = \frac{D_E}{D_U}$$

If the unexplained variance is great with respect to the explained one, the value of the statistics is small and the regression is bad.

If all the variance is explained, i.e. the data lie right on the line, the statistics is small and the regression is ideal.

Thus H0 says: the regression is bad. If the p-value is small, H0 is rejected and the regression is good.

10 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 can be convenient for such approximation. Here we will discuss this question.

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

2. Pearson t-test of correlation coefficient

   For approximation of a relation between x and y there mus be any relation. Pearson test can be used.

   Pearson t-test has H0: $\rho = 0$ (no relation), HA: $\rho \neq 0$ (there is a relation); both sided test with Student distribution.

   For good regression, H0 must be rejected.

3. Fisher F-test of explained and unexplained variance

   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 should be independent. If not, the relations between them could be used to construct better regression curve.

   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 standard error is defined as

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

which is variance of prediction error e_i relative to variance of dependent variable y_i. The smaller the errors are, the better approximation. At least it should be smaller than 1.