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 probability of the generated value is equal within a fixed interval (uniform distribution) or the data are generated from a fixed value and are influenced 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 of data sample is a set of data measured in the generator. It is set of values (vectors) with a specified length (number of measured data). Thus, the sample is fixed and also its characteristics (mean, variance etc.) are fixed, too.

- Sample realization of data sample 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 length in four arms in 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 random sample realization can theoretically be repeated (even if in practice only one sample realization is 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 a different value in each sample realization. And this is he 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 is true 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 on 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 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

$$x = [5.3,\ 2.8,\ 4.5,\ 1.7]$$

we have

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

Sorted data are

$$s = [1,\ 2,\ 4,\ 5]$$

which gives

$$s_1 = 1,\ s_2 = 2,\ s_3 = 4,\ s_4 = 5$$

Now, ranks of x are orders (indexes) of data in the sorted set, i.e. for $x_1 = 5.3$ we look for 5.3 in s. It is $s_4 = 5.3$ and it is at the fourth position (has index 4). That is the rank r_1 of x_1 is $r_1 = 4$.

Similarly, for $x_2 = 2.8$ we find $s_2 = 2.8$ and the rank is $r_2 = 2$.

The rest of ranks is $r_3 = 3$ and $r_4 = 1$.

All the ranks 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 and 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

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

Use of ranks

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 that 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 variable. 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

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

which is unbiased estimate of variance

- second moment of continuous variable

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

- second moment of discrete variable

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

Second mutual moment (covariance)

- data covariance

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

- covariance of continuous variables

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

- covariance of discrete variables

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

Correlation coefficient

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

Use in estimation and testing

parameter		**statistics**
expectation	→	sample average
variance	→	sample variance
proportion	→	sample proportion
independence	→	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) 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. However, this distribution is not known to us.

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!

Basic assumption: The unknown distribution has some form - mostly normal. However, some unknown parameter is in this distribution (expectation, variance or proportion). And our estimation reduces to estimation of this parameter.

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 data sample whose values point at the estimated parameters is called **statistics**. In our case the transformation function will be sample variance whose general form is

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

[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.

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

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

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

Now, very important information comes!!!

As we said, if we try to do new samples and compute sample variance from each of them, we would obtain different values of the statistics. And we have said, that a variable which after each measurement gives different values is random variable. So, generally, the statistics is **random variable.** Values of the measured sample realizations are its realizations. And as the statistics is random variable, we can speak of its distribution.

Thus, we have random variable X (population) from which we measure data x_i that form a sample realization $x=\left[x_1,x_2,\cdots.x_n\right]$. 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 inserted $\hat{\theta}=T\left(x\right)$. Both, the data x and the statistics T are random variables and have their distribution $f\left(x\right)$ and $f\left(T\right)$. 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 $f\left(x\right)$ is used

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

  $$\int_c^d f\left(T\right)dT$$

  where the distribution of the statistics $f\left(T\right)$ is used.

The confidence intervals and tests of hypotheses deal with parameters and form intervals of parameters with given probabilities. So, their derivations, concerning critical region or p-value always deal with the **distribution** $f\left(T\right)$ **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 for inserting the sample realization produces value which is near to the parameter θ.

It should have the following properties

1. Is **unbiased**
   It holds
   $$E\left(T\right)=\theta$$
   i.e. the average from all possible sample averages (made from all possible sample realizations) is exactly equal to θ.

2. Is **consistent**
   For unbiased estimate T_n where n is the sample length it holds
   $$\lim_{n\to\infty} D\left[T_n\right]=0$$
   i.e. for sample length n going to infinity the estimate is precise.

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

6 What is p-value

It is very important to select a correct and solve it in a suitable software. However, perhaps still more important is to make correct conclusions from testing.

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 now preferred.

Critical region and realized statistics This is the basic way how to present a result of testing.

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.

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

p-value This way of result interpretation has the advantage that it is apparent, how strongly we reject or do not reject.

The best way how to introduce p-value is for right-sided test and graphically.

That is: we have population described by distribution $f(x,\theta)$ with unknown parameter θ. This distribution point at data in the sense, that

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

gives th probability that $x \in (a,b)$.

Then we have a statistics T that corresponds to the parameter θ. This statistics is also a random variable with distribution $f\left(\hat{\theta},\theta\right)$[2] - the parameter θ is inherited from the population distribution. This distribution points at the parameter θ (its realizations are point estimates $\hat{\theta}$ of θ). The meaning is the same. The integral

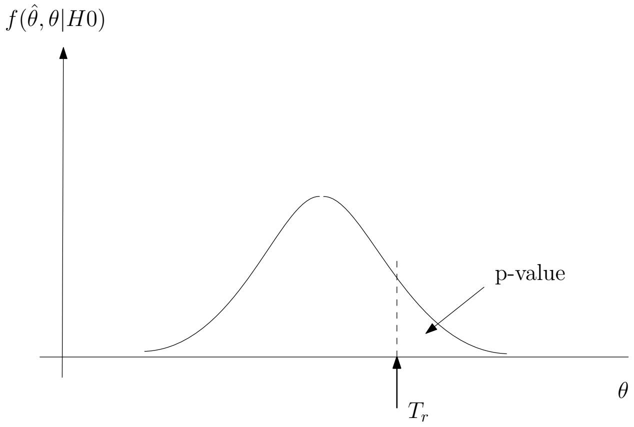
$$\int_c^d f\left(\hat{\theta},\theta\right) dT$$

determines the probability that $\hat{\theta} \in (c,d)$. Here, θ is only a parameter.

From what has been said it is clear, that we shall work with the distribution of the statistics $f\left(\hat{\theta},\theta\right)$

The following picture shows the situation

[2] Here, usually is written $f(T,\theta)$, however, T is random variable with realizations $\hat{\theta}$ and realizations should appear in the argument pf a probability function. Perhaps, the full notation should be used here which is $f_T\left(\hat{\theta}\right)$, which, of course, depends on θ.



As we have shown, the statistics distribution $f\left(\hat{\theta}, \theta\right)$ 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 the situation according to H0. The HA, here in our case, 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. 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 lie (which is determined by data), the smaller is the area under the density and to the right from the realized statistic - which is the **p-value**. Its definition is

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

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

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

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

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

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

If we want to use the confidence level α we can 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 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 is first and which second. Let A is first and B second. What we test is the difference between them A - B.

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

HA then is A - B $<$ 0 and less means left-sided test (see above).

8 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

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

   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.

9 Chi-square test - variants

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)

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

Their product forms joint probability for independent variables

$$f_n = \begin{bmatrix} 0.312 \\ 0.688 \end{bmatrix} [0.456,\, 0.361,\, 0.183] = \begin{bmatrix} 0.142, & 0.113, & 0.057 \\ 0.314, & 0.248, & 0.126 \end{bmatrix}$$

$$E = f_n N = \begin{bmatrix} 124.00, & 98.14, & 49.85 \\ 273.99, & 216.86, & 110.15 \end{bmatrix}$$

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