Statistical methodology in epidemiology

1 Introduction

Medical students and physicians often believe that broader knowledge of statistics is not necessary. They are sure that to compute basic characteristics of data is simple and statisticians and computers can provide more complicated computing. However, we will show the reasons, why this way of thinking is too week.

Statistics is in same sense the language for data gathering, their manipulating and quantitative interpretation. In fact a physician deals with similar activities Therefore he should to learn about basic statistical approaches and to know when he can use them correctly.

Questions that are asked by physicians have often a statistical nature. For example: Is a new drug better then the one previously applied?, What is the risk of side effects?, How many patients should used the drug to prove its positive effect?, Also two basic questions that are asked by patient, i.e.: What is my diagnosis?, and What is prognosis of my disease?.

Therefore it is necessary to learn what methods should be used to receive a correct answer and what is really revealed by a statistical analysis.

Our procedures for generalization of evidence from data are based on many assumptions under consideration. These assumptions are not often explicitly stated. However, using statistical methods for evaluation of data it is necessary to know and estimate, whether it is possible to consider their validity for given problem or not.

Explosion of computer technology that enter health care sector makes possible to use computers and complicated statistical methods also by not statistically knowledgeable persons. These possibilities have very negative face in increased danger of selection not correct statistical procedure for data evaluation and consequently for wrong conclusions based on this analysis.

In published papers in medical journals we can find application of statistical approaches for data analysis very often. Therefore, nowadays physicians cannot evaluate collected data or to study national and international medical journals without basic statistical knowledge. This is very important assumption for physicians to be able to publish own results and critically judge other papers.

Now we will give an overview on basic statistical concepts and methods and we will concentrate on their application in epidemiology.

2 Basic statistical notions

Statistics is a science that is focused on gathering, evaluating and analyzing of the data obtained by counting or measuring the properties of populations of natural phenomena. Statistics can be divided into two organically connected parts - descriptive statistics and inductive statistics.

Descriptive statistics deals with ordering of samples, their description and summarization. Their historical basis lies in antiquity, when the summarization of citizens, fields and other things important for the strength of the state has started. Nowadays the word statistics also can express a description of sample or its summarization, e.g. statistics of mortality, statistics of morbidity. Descriptive statistics enables us to create conclusions from given sample only for this sample. Methods of the descriptive statistics are based on well-arranged results using tables, graphs or basic statistical characteristics.

Inductive statistics provides methods that enable us from empirical evidence to formulate scientifically judged conclusions. It is based on probability theory results and it has been strongly developed much later, namely in twentieth century.

Human way of thinking is based on two main thinking approaches, i.e. deductive and inductive thinking. *Deductive thinking* is a process, when under generally valid assumptions or laws we create conclusion for individual cases. This way of thinking is mainly developed in mathematics, when we make conclusions in abstract mathematical models. Practical sense has deductive way of thinking only as the part of thinking chain, when also other thinking types are applied.

Let us show an example. We will give to a physician a new drug and we will instruct him on its application. Then (if physician follows instructions) his way of thinking is in fact deductive. „The drug has positive effect on all persons suffering from a given disease. The patient suffers from this disease. Therefore the drug will have a positive effect and I will apply it“. In his thinking physician uses general knowledge about the effect of the drug when it is applied and his treatment in individual cases is based on them.

To reveal effect of a new drug we have to proceed in the opposite way, i.e. from observation of the drug effect in individual cases to generalized conclusion. This part of human thinking, we call inductive thinking. *Inductive thinking* enables us to crate general conclusions based on observation of individual cases. Inductive way of thinking plays very important role for human being behaviour in the real word. The ability to learn from experience makes human being possible to adapt on changing living conditions. However, conclusions of inductive thinking procedures are influenced by subjective attitudes and have limited validity. For example in certifying a positive influence of drug on a large sample of patients suffering from the given disease, e.g. all diseased persons in our country in given calendar year, we never will have all sample of patients we can treat with the drug in future. We have to use inductive way of thinking that in opposite to the deductive one can yield decision error. However, using subjective way of thinking we never can reveal how the error is large.

Methods of inductive statistics (so called *statistical induction*) can under given assumptions to make general conclusions and to objectively enumerate their degree of confidence. The main aim of inductive statistics is to elaborate procedures how to create general conclusions from empirical data that can substitute subjective inductive thinking by objective inductive thinking based on concepts of probability theory. Central role in methods of inductive statistics play two important concepts - population and sample.

Population is given by exact definition of its objects. Objects of a population are given by enumeration or by explicit rule (for example given common property) that enables us to decide for any object whether belongs to given population or not.

Examples of populations are: population of citizens of the given town during specified time period, population of patients suffering from a given disease or population of samples of animal tissues. Therefore objects of populations are of different types, e.g. persons, experimental animals, families, blood samples or EEG records.

Population can be finite (e.g. demographic populations) or infinite, that is an abstract population existing only as our idea. Infinite are mostly populations defined by properties of their objects, e.g. population of results of experiments provided under given experimental conditions or population of all patient suffering form a given disease (time period is not specified).

Objects of populations are described by variables. A *variable* is a quantity with values that may vary from object to object. Variables are of two main types.

Qualitative (categorical) variables have values that are intrinsically nonnumeric (categorical). Therefore qualitative variables always are always *categorical variables*.

Quantitative variables have values that are intrinsically numerical. If quantitative variable can reach any value from a given range, it is called *continuous* variable (e.g. weight, height). If quantitative variable can reach only separated numerical values it is called *discrete* (e.g. number of children in a family, number of deaths). Continuous variables are sometimes transformed in categorical variables by grouping their values in categories (classes). However this transformation can lead to a loss of information.

Discrete variables are categorical variables, when we consider their values as categories. Quantitative variables are measured on the interval scale. *Interval scale* can order values of objects according to the degree of measured property and also determine their exact distance. We can distinguish between two types of interval scale. The first one is the interval scale with fixed origin the second one with changeable origin. The interval scale with *changeable origin* is the null position given by a choice (e.g. measuring of temperature in degrees of Celsius or any time scale). Null position on the interval scale with a *fixed origin* is given and it expresses absence of a given property (e.g. scaling of the weight in kg or height in cm).

Qualitative variables can be measured on nominal or ordinal scale. *Nominal* scale enables us only to name values of the nominal variable, (e.g. nominal variable *family status* for a man can have values single, divorced, married, widower). *Ordinal* scale (in addition) can have values ordered according to a specified criterion, (e.g. ordinal variable *education* can have ordered values basic education, high school education, university education). With values of qualitative variables we often attach numbers. However, these numbers can serve only as the codes that in case of ordinal variables reflect also the ordering.

Methods of statistical induction allow us to create conclusions on population from given elements in so-called *sample*. Because we know that the inductive conclusions are joined with uncertainty, the scientific nature of statistical inductive conclusions is based on the fact, that the degree of the uncertainty can be objectively expressed. This can be done only in case that samples were created by *random (probability) sampling method.*

3 Sampling

The purpose of taking a sample is to make statements about a whole by just examining a part. The theory of sampling is concerned with how to take the sample and the types of statements you can make about the whole based on looking at a part. Any sampling procedure should be completely objective. The main aim is to get the best value for money where value is defined in a statistical sense. Before we select any sample we should get to know our population. A population is any group of objects (people, things etc.) that includes all possible members of that group (e.g. a population of hospitals, a population of readings on a sphygmomanometer, a population of lung cancer deaths).

Any individual member of a given population is known as a *unit*. A *sample* is a part or subset of the population, which is chosen in such a way that it is *representative* of the population, and it is used to gain information about the population. Firstly, we need to define who are our population. A population is the list of all elements. It must be precisely defined in terms of who is to be included. Sometimes it is easier to define who is to be excluded. It is very important prior to selecting the sample to define the population precisely. Some headings to consider are as follows:

- Who is to be included?
- Who is to be excluded?
- Are there any special characteristics?
- Geographical considerations?
- Time?

A further step is to examine if any lists of the population units exist. A unit may be a household or a person, depending on the survey. A *sampling frame* is a list of all elements in the population. We usually try to use a frame that already exists. This makes the task of selecting a sample much easier. It is very time consuming to construct a frame. One approach is to examine how other studies have approached the problem. One of the most common sampling frames in use us the *Electoral Register* lists. This is not a perfect list of all objects in the population of eligible voters but we often have to comprise and use what ever is available.

If we have the time and money available we may want to check the lists for duplicate, missing or foreign elements. We often redefine our population to coincide with some existing sampling frame.

Finally, in some cases it is important to investigate how these lists or population elements are physically arranged. Hopefully the lists are computerized as this makes the tasks of sampling much easier. Sometimes the elements are files and it is important to see how they are physically stored.

4 Types of sampling techniques - basic definitions

Sampling is divided into two types - probability and non-probability sampling.
In *probability sampling (random sampling)* every object of the population has a known non-zero chance of being selected. In *non-probability sampling* choice of selection of sampling units depends entirely on the decision of a sampler. Inductive statistics is based on probability (random) sampling methods only. We will give same examples of random sampling methods.

Simple Random Sampling (SRS)

Each object has an equal chance of being selected. This is the type of sampling assumed by most of the statistical packages and it is standard to which all methods are compared.

How do I select a simple random sample?

Define the population. Construct a frame if one does not exist already. In other words produce a list of all elements in the population. You may be lucky and one may exist on a computer. Give each element a unique ID starting from 1 to the number of elements in the population N. Use only the minimum number of digits. Select people at random using tables of random number or computer generated random numbers.

Multi-Stage Sampling

As the name suggests the sample is selected in stages. The population is divided up into hierarchical units. Clusters of the elements that occur naturally in the population usually form these units, e.g. they live near each other. For example we are interested in selecting a sample of people for interview face-to-face. We first select a cluster and then from the selected cluster we select the people. This type of sampling is used in some way in the vast majority of all social surveys, which involve interviews.

Advantages of multi-stage sampling

The first major advantage of multi-stage sampling is the reduction in cost relative to simple random sampling. The second major advantage arises when there are no suitable sampling frames for the entire population. We first construct a list of all the areas in the population but we need only construct a sampling frame - list of all the elements for the selected areas. This substantially reduces the cost of drawing up a sampling frame. An area is often called a Primary Sampling Unit (PSU).

Disadvantages of multi-stage sampling

For the same sample size, the value (measured in terms of precision) for multi-stage is usually less than simple random sampling.

Stratified sampling

The population is divided up into groups prior to selecting the sample, e.g. areas of the country and separate samples are selected from each group.

Advantages of stratified sampling

Stratification may be desired for administrative convenience. The agency conducting the fieldwork can establish field offices in each of the strata thereby leading to better organization and supervision of fieldwork.

Different sampling techniques may be used in each of the stratum. This may be desirable if the strata correspond to natural characteristics, e.g. cities versus country, as there may be different types of sampling problems, which may need different approaches. Also it may be desirable to use different methods of collecting the data from each of the strata.
Stratification ensures adequate representation of various groups of the population, which may be of interest.
Stratification also ensures a better cross section of the population.
In comparison to SRC, stratified sampling may result in smaller standard errors.

Sys*tematic sampling*

This is a widely used sampling technique. It consists of taking every k-th sampling unit after a random start.

Advantages of systematic sampling

The most important advantage is that it is easy, almost foolproof and flexible to implement. It is especially easy to give instructions to field-workers. If we order our list prior to taking the sample, the sample will reflect this ordering and as such can easily give an implicitly stratified sample.

Disadvantages of systematic sampling

The main disadvantage is if there is an ordering in the list, which is unknown to the researcher, this may bias the resulting estimates.

5 Confidence intervals

It is rarely the case that we make measurements on a set of objects and we are interested in these measured objects only. As usual, the measured objects are regarded as a sample from a much larger population and the properties of the population are what are of interest. For example, if we administer a drug to a group of patients with high blood pressure and measure the reduction in blood pressure then, typically, what we want from the study is to estimate the average reduction in blood pressure that would be seen if the drug were administered to the whole population of such patients. The natural estimate will be the average reduction in the sample but we know that this will be subject to random variation depending on which people are selected for measurement, when they are measured and what measurement errors arise in the process. Accordingly, rather than simply report our sample average as a single point estimate of the population value, usually we prefer to report an interval around the sample average within which the population average is likely to lie.

Let us assume that a variable has population mean μ and standard deviation σ (or variance σ^2) the sample mean. These characteristics can be estimated from sample of the size n as a sample mean $\bar{x}$ and standard deviation s (or variance s^2).

If a sample of size n is selected and the sample mean $\bar{x}$ is calculated then the probability is 0.95 that the value obtained will lie within a distance of $1.96\frac{\sigma}{\sqrt{n}}$ of the population mean μ.

This property can be expressed differently: if we select a sample of size n and calculate $\bar{x}$

then the population mean μ will lie within a distance of $1.96\frac{\sigma}{\sqrt{n}}$ of $\bar{x}$ with probability 0.95,

i.e. the interval $\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$ covers the population men μ with the probability 95%.

The higher the confidence levels the wider (and perhaps less useful) is the confidence interval. Accordingly, there is a trade-off between the level of confidence and the width of the interval.

Sample Size

It is clear from the formula that the size of the sample controls the width of the confidence interval for a given standard deviation and a given confidence level. Thus suppose we wanted a 95% confidence interval for the mean height in population of boys aged 9.5-10 years and we required an interval accuracy ± 0,5 cm. The calculated interval would be

$$\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

and we require $1.96\frac{\sigma}{\sqrt{n}} = 0.5$.

Therefore for σ = 08 we will calculate n = $(1.96\ (08/0.5))^2 \cong 568$.

In order to get a confidence interval, which was ± 0.5 cm, we would require a sample size of 568 boys.

What happens when we do not know the population standard deviation? Let us take a random sample of boys (n = 3231) and calculate the average height of our sample ($\bar{x}$ = 138.52 cm, s = 08). The confidence interval now becomes $\bar{x} \pm t\frac{s}{\sqrt{n}}$. What is the t value?

Basically what happens here is that when we do not know the population standard deviation σ we replace it by the sample standard deviation s. This, however, implies somewhat more uncertainty in our calculation. To take account of this, instead of using the value derived form the standardized normal distribution, (i.e.1.96 for 95% confidence interval), we use the corresponding value from a slightly different (but related) distribution so-called. Student's t-distribution. This is another theoretical distribution like the normal. So instead of looking up the normal tables for a value, we look the t-tables. To look up the t-tables we need to know what is called the degrees of freedom - this is just the size of the sample minus one, i.e. (*n- 1*). The t-distribution is very like the normal distribution for large samples (*n* > *100*). Therefore the numbers in the t-table corresponding to the last row (∞) are the values corresponding to the normal tables.

In case of our example we have n =3231 therefore the number of degrees of freedom is n –1 = 3230. Then we can find that the t value is equal to 1.96, and the estimate of the required sample size is 568.

6 Hypothesis testing: Introduction

The best way to analyze and report the results of comparative studies is through the use of confidence intervals since the results are thereby not only established as real, i.e. as not being due to chance variation, but the uncertainty in the measured change is made explicit. It is, however, still the case that statistical hypothesis tests are widely used in scientific work and almost mandatory in some journals. Accordingly, it is appropriate that we should consider the issues that arise in using such tests.

The majority of statistical analyses involve comparisons, most obviously between treatments and procedures or between groups of subjects. We state a hypothesis called the null hypothesis and alternative hypothesis. The following is summarizing possible outcomes of a statistical test.

Table 1 Possible outcomes of a statistical test

Sample	Population	
	H_0 True	H_0 False
Reject H_0	Type I Error	Correct decision
Failed to reject H_0	Correct decision	Type II Error

The medical hypothesis is the basis for formulation of both statistical hypotheses, i.e. null hypothesis H_0 and the alternative hypothesis H_1 as we can see on the examples given below.

7 Hypotheses testing in fourfold and contingency tables

A chi-squared test is used to determine whether there is an association between two variables, which may be:

- qualitative
- discrete quantitative
- continuous quantitative, whose values have been grouped.

When there are two such variables the data are arranged in a c*ontingency table*. The categories for one variable define the rows and the categories for the other variable define the columns. Individuals are assigned to the appropriate cell of the contingency table according to their values for the two variables. A chi-squared (χ^2) test is used to test whether there is evidence for an association between the row variable and column variable. When the table has only two rows or two columns this is equivalent to the comparison of proportions. In this case it is called *four-fold table*.

Example: The medical hypothesis is that progressive polyarthritis (PAP) is associated with the HLA-DR4 antigen. Statistical testing is based on reformulating the medical hypothesis in two statistical hypotheses, i.e. null hypothesis H_0 and the alternative hypothesis $H_{1.}$ For our medical hypothesis the statistical hypotheses are as follows:

H_0: There is no association of PAP with HLA-DR4

H_1: PAP is associated with HLA-DR4.

We intend to verify the null hypothesis on 5% level using data given in Tab.4

Solution: We will perform a statistical test of above stated statistical hypotheses at 5% level of significance using sample data (*observed frequencies*) given in Tab.4

Table 4 Observed frequencies in the sample of 308 patients divided according to presence of PAP and HLA-DR4

PAP	HLA-DR4		Total
	Present	Absent	
Present	46	28	74
Absent	50	184	234
Total	96	212	308

Expected frequencies for each cell in the table are numbers that *we would expect to find* if the null hypothesis was true. Generally the expected frequency in the cell of the i-th row and j-th column can be calculated as the sum of the i-th row multiplied by the sum of the j-th column and divided by the total number of patients *n*. Tab.5 summarizes calculated expected frequencies in four-fold table.

Table 5 Expected frequencies in the sample of 308 patients divided according to presence of PAP and HLA-DR4

PAP	HLA-DR4		Total
	Present	Absent	
Present	23	51	74
Absent	73	161	234
Total	96	212	308

The chi-squared test for association in a fourfold table works like this. The Null Hypothesis is that there is *no association* between the two variables, in our example describing presence or absence of PAP and HLA-DR4. We can now compare the observed and expected frequencies. If the two variables are not associated, the observed and expected frequencies should be close together, any discrepancy being due to random variation. The best way of looking at the *differences between observed and expected frequencies* is to calculate the chi-squared (χ^2) statistic as follows: $\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$,

where the summation is over all cells in the table. For the above example the test statistics is $\chi^2 = 43.61$.

In order to interpret this chi-squared statistic, we need to know the number of degrees of freedom (df) involved. For a contingency table this is given in general by the formula

df = (number of rows - 1) x (number of columns - 1)

In the above example there are 2 rows and 2 columns so we have df = (2-1)(2-1) =1 Turning to the table which shows the percentage points of the χ^2 distribution, we can see the value of 43.61 is greater than 3.84 the critical value for 5% level of significance. Thus the probability is less than 5% that such a large observed difference could have arisen by chance, if there were no real differences.

The use of the chi-squared test is not confined to nominal and ordinal data but can also be used for continuous variables that have been categorized. The procedure described for four-fold table can be easily applied for any contingency table.

There are many other important statistical tests. Very often have been used t-tests for testing different hypotheses on population means. More detail explanation on t-tests and other test used in epidemiology can be found e.g. in the books [1],[2],[3].

8 Simple linear regression

Frequently it is of interest to investigate the relationship between two variables where one variable, the predictor variable (X), is thought of as driving the second variable, the response (Y). Both of these variables are assumed to be quantitative. Other names are dependent variable (Y) and independent variable (X). The first step in such an investigation should be to plot the data. In many cases this will tell much of what is of interest: does there appear to be a relationship?; if yes, do the variables increase or decrease together?, does one decrease when the other increases?, is a straight line a suitable model to describe the relationship between the two variables, and so on. If we want to go beyond this qualitative level of analysis then simple linear regression is often a useful tool. This involves fitting a straight line through our data and investigating the properties of the fitted line.

First, as suggested above, we should plot the data. It is conventional to plot the Y- response variable on the vertical axis and the independent variable X on the horizontal axis. This plot suggests a linear relationship so we proceed to quantify the relationship between and by fitting a regression line through the data points. We could then write the model as follows:

$$Y = \alpha + \beta X + \text{Residual},$$

where α is the intercept.- where the line cuts the Y - axis, β is the slope of the line and the residual is the part that cannot be accounted for by the model. It is clear that all the points do not lie exactly on a straight line. The line fitted by the least squares criterion (this the criterion which is almost invariably used). The above analysis and discussion have dealt with the case of a simple linear model i.e. straight line and 1 independent variable - X. Using regression we can fit many other types of models including those where we have more than one independent variable.

9 Correlation analysis

Sometimes we do not have a clear predictor and a clear response variable. We may be interested in quantifying the relationship between a pair of variables. The regression of X on Y does not give the same regression line as the regression of Y on X. This is because regression analysis presupposes a directional relationship, i.e. X is thought of as influencing Y and not vice versa. Despite this the r^2 value obtained from both regressions will be the same. As such it is a measure of the strength of the linear relationship between X and Y, irrespective of which is considered to influence the other. The square root of r^2 turns out to be exactly the same as a measure called the *correlation coefficient* (variously called Pearson's correlation coefficient or the Product Moment correlation coefficient) which was proposed to measure the strength of *linear relationships* between normally distributed random variables. The correlation coefficient is just the square root of r^2 but has a sign attached: it will be positive if X and Y increase and decrease together and negative if one increases while the other decreases. The correlation coefficient varies from -1 to +1: it is -1 or +1 if all the points lie in a straight line and zero if there is completely random scatter. To sum up when we look at a correlation value we should be interested in the sign and the size (absolute value). The sign tells us the direction while the size tells us how close the points are clustered around a line. It is also crucial to remember that correlation does not imply causation.