TEACHING STATISTICAL METHODS FOR THE HEALTH SCIENCES IN THE POST-GENOMIC ERA

Nibia Aires The Cardiovascular Institute, Department of Medicine, Sahlgrenska University Hospital/Östra and School of Information Science, Computer and Electrical Engineering, Halmstad University, Sweden

In many complex diseases researchers have observed that neither genetic factors nor environmental factors alone determine the disease. This observation generates the hypothesis of multifactorial causes of disease, which means that human disease is caused by both genetic and environmental factors that act together. On the other hand, the recent compilation of the draft human genome sequence opens the possibility of detecting candidate genes for complex diseases and to study these in relation to environmental factors. This gene-environmental interaction may not be easy to analyse due to the complex structure. The involving factors have different nature that should be treated at different stages of the study. Particular attention should be paid to the study size and design. Epidemiological studies with particular interest in identifying candidate genes that contribute to complex diseases as well as detection of intergenetic or gene-environment interactions require large sample size because many variables are studied simultaneously. The larger patient populations ensure that individual subgroups retain adequate power to detect significant results with narrow confidence intervals. In the paper we focus on the advantages/disadvantages of classic multifactorial statistical methods applied to the health sciences and the genome scan.

Introduction

Any biological function can be assessed as a consequence of the interaction between external factors and genetic disposition. The latter has been designated genetic susceptibility and is largely a result of genetic polymorphism. Modern biology aims at estimating how large part of the phenotype variation that can be explained by genetic polymorphism, and to relate this to the underlying physiological functions.

Risk of disease, or disease liability may be considered a phenotype variation where the risk is a function of this interaction as well as a direct effect of external and genetic factors.

The completation of the human genome mapping opens new possibilities to be able to study the interactions between genetic and environmental factors in relation to complex diseases, but on the other hand it brings up to discussion the statistical methodology used in these analyses.

In this paper we discuss three issues of the statistical methodology which are extensive used in the area of the health sciences, namely the problem of multiple testing, sample size determination and parametric vs. non-parametric tests. We stress the importance of discussing these issues among medicine and genetics students emphasising that in many situations there is not a straightforward solution.

MULTIPLE TESTING

The gene-environmental interaction may not be easy to analyse due to the large number of variables and measurements involved in the analysis and their complex structure. Assessing gene-environmental interaction may lead to multiple testing. Once the data sets related to these studies are constructed, they often consist of a large number of variables that researchers wish to test from many angles. A major drawback of multiple testing is the increased probability of getting "false-positive" results, that is statistically significant associations, interactions, etc, that are not in reality or associations that either cannot be replicated or corroborated by other similar studies.

From our experience adjustment for multiple tests is common in studies related to the health sciences and even in other areas. In epidemiological studies once the data are collected, researchers may perform multiple tests to detect interesting relationships not assumed a priori or by analysing subpopulations. In clinical trials, the problem may arise when analysing few experimental units and a large number of variables.

In teaching statistics, it is important to emphasise the consequences of using the statistical methodology erroneously, i.e. lacking of hypotheses at start of the study, performing multiple tests without computing adjusted p-values, etc. The conclusions under such circumstances can be very misleading and far away from the true results. Moreover, using inappropriate correction for multiple testing may also lead to erroneous conclusions. In fact, using a weak correction may increase the false-positive levels, but on the contrary, a strong correction may decrease the possibility of detecting (statistical power) associations, differences, interactions or effects.

However, in genetic and genetic-epidemiological studies, where many factors are involved, and the aim of the study is to assess associations and interactions, multiple testing cannot be avoided. Indeed multiple hypotheses may be part of the aim of the study and then multiple testing has to be carried out to provide multiple inferences. It is important to inform the students of medical sciences the existence of alternative adequate methods to prevent erroneous conclusions when dealing with multiple inference.

Such methods are the ones involving permutations, also referred as resampling methods, which involve computer-intensive simulation analysis. Although they can be computationally time-consuming, they can also provide accurate information about extreme events. Some examples of resampling methods are the bootstrap methods, permutation analysis, and parametric simulation methods. The basic concept underlying the resampling methods is to randomly re-assign the observed value of the variable to treatment groups, and to re-compute the test statistics. This procedure is performed many times and the original test statistic is compared to the resampling distribution. If the original test statistic is extreme in comparison to the simulated distribution, then it is considered unusual, otherwise it is considered typical. Using this method one might compute the p-value. Suppose we want to compare the means of two groups, using resampling. If data are randomly assigned to the groups, there should be no differences between groups except for differences by chance. The test statistic is re-computed and compared with the original test statistic as above. The original statistic is at last classified as unusual or a typical according to the simulated distribution. The resampling-based p-value is then the proportion of the resampled data sets giving a test statistic as extreme as the original test statistic, c.f. Westfall and Young (1993).

SAMPLE SIZE DETERMINATION

The size determination of the sample is very important in the design of clinical trials and epidemiological studies. The power of a study is influenced by many factors, such as the size of the effect and the sample size.

The general procedure for sample size determination can be based upon two approaches: controlling Type I and Type II errors under a hypothesis structure or controlling the width of the confidence interval under point estimation. Sample size determination might also be a compromise between the available resources and the objectives of the study. In Lachin (1981), a presentation of methods for sample size determination for clinical trials is given. The paper treats the problem of sample size determination and power analysis for simple and even more complicated study designs. The determination of sample size based upon the confidence estimation approach is given by McHugh and Le (1984).

However, when dealing with a complex data structure, i.e. studies in complex disease, where many variables are involved, the calculation of the sample size is not straightforward. Assumptions about the disease information that is not known with certainty, or assumptions about the size of effects and variation, have to be made, when these parameters are unknown. In such situations, the standard approach for sample size determination is not feasible and simulation procedures are recommended to assess the sample size of the study. Although computer programs for assessing power and sample size determination are available, the investigator often has to provide information about the underlying model and some value of variation to initiate the simulation process. Sample sizes should be chosen to assume non-optimal conditions, such us weak effects, large variances, etc. Many of these computer programs are often based on Monte Carlo methods, Haines and Perick-Vance (1998).

It is important to point out, when teaching statistical methods for the health sciences, that sample size determination for complex designs involving gene-environmental interactions assessments is an intricate problem that requires the co-operation of a multidisciplinary team, i.e. geneticists, epidemiologists, statisticians, programmers, etc. Even though the discussion about power is limited to approximate estimates of sample size required to detect an association based on assumptions about some underlying model.

PARAMETRIC VS. NON-PARAMETRIC TESTS

Another important issue that we want to bring up for discussion is the choice of hypothesis tests. This choice depends upon the assumptions that the observations come from a known probability distribution or not. Some of the well-known distributions are the normal or Gaussian, the binomial, the uniform, Poisson, exponential, multinomial, gamma, beta and Weibull distributions. Parameter estimations using this approach are examples of parametric inferences. In practice, some researchers carry out parametric tests even if the normality assumption is incorrect, based on the central limit theorem. The central limit theorem states that the shape of the sampling distribution of the sample mean is approximately normal for large sample sizes, even if the parent distribution is not normal (and the larger the sample size, the more closely is this distribution). The central limit theorem applies for large samples; thus these assumptions are often inappropriate among small samples. Moreover, parametrical methods are inappropriate for analysing qualitative data, such as data from rating scales, dichotomous variables, etc. The method of inference for qualitative data is non-parametric, since calculations based on adding or subtracting ordinal data are not appropriate, Svensson (2001). Test of normality are available in most relevant statistical programs and in many cases it is easy to see that the sample is not from a normal distribution, since normality implies certain properties of symmetry and spread.

It should be also emphasised that there exist a list of non-parametric tests to deal with different problems and in many situations non-parametric methods are the only ones available for data that simply specify ranks or counts of individuals in various categories, Sprent and Smeeton (2001).

REFERENCES

Haines J.L., Pericak-Vance M.A. (1998) Approaches to gene mapping in complex human diseases, John Wiley & Sons.

Lachin J.M. (1981) Introduction to Sample Size Determination and Power Analysis for Clinical Trials, *Controlled Clinical Trials*, 2, 93-113.

McHugh R.B., Le C.T. (1984) Confidence Estimation and the Size of a Clinical Trial, *Controlled Clinical Trials*, 5, 157-163.

Sprent P., Smeeton N.C. (2001), Applied nonparametric statistical methods, 3rd. edition, Chapman & Hall.

Svensson E. (2001), Guidelines to statistical evaluation of data from rating scales and questionnaires, *J Rehab Med*, 33: 47-48.

Westfall P.H., Young S.S. (1993), Resampling-based multiple testing: Examples and methods for p-value adjustment. Wiley.