STATISTICAL METHODOLOGY IN BIOMETRY

G. Molenberghs
Limburgs Universitair Centrum, Hasselt University, Belgium

Keywords: Statistical models, linear regression, categorical data, hierarchical data, survival analysis

Contents

Summary

When choosing a statistical approach in biometry and, in fact, elsewhere, two equally important considerations ought to be made. First, is the method of analysis technically sound and second, is it relevant from a substantive view-point. These considerations apply, irrespective of the complexity of the design and the type of outcome considered. In this chapter, we sketch some of the most commonly used statistical analysis methods and extensions thereof, for normally distributed, non-normally distributed (e.g., categorical), and, in particular, time-to-event outcomes. Both the univariate as well as the correlated data settings are given attention.

1. Introduction

Choosing a statistical approach is a very common task in everyday statistical practice. When choosing a method for analysis, it is important to reflect on whether the methodology is sound from a theoretical point of view and whether it is adequate in terms of the scientific research question of interest. A method chosen should therefore reflect the design, type of outcome, type of covariates, etc. A useful distinction is made between methods for univariate (single) outcomes and methodology for correlated sets of response variables.

The simplest statistical analysis is concerned with a single outcome variable, recorded for a sample of a homogeneous population. Standard procedures include the computation of means or medians (location parameters) and standard errors or interquartile ranges (dispersion parameters). For example, the height of a number of human subjects might be recorded. A first level of complexity arises when a variable is recorded for a sample out of two subgroups (subpopulations) of a larger population (treated and untreated patients, two species, boys and girls): the two-sample problems. A question of interest is whether the means are different in the two populations. The outcome variable might still be height, but we would have an explanatory variable: treatment allocation, or sex. For example, the height of boys can be compared to the height of girls. The outcome variable is often called dependent variable. The predictor is often called covariate or independent variable. The statistical tools for this data setting include analysis of variance (ANOVA), *t* test and Wilcoxon test.

In the previous situation, the dependent variable had only two levels: a binary or dichotomous variable. This is the simplest case. Alternatively, the predictor itself could be a variable with several levels (e.g., dose administered in a clinical trial; one of several species of a plant; race, etc.). In addition, it could potentially have an infinite number of levels, just as is the case with the height response variable. For example, a baseline height at 7 years of age can be compared to the height at 10 years. This leads to a family of models frequently referred to as regression models. When the dependent variable is continuous (height) one often uses linear regression. The independent variable can be continuous, binary, categorical, or discrete. The choice of the statistical analysis method is driven by the outcome or dependent variables, rather than by the predictor variables.

Should the dependent variable be binary (diseased/non diseased; death/alive, etc.), then one would choose logistic regression rather than linear regression. Several alternatives to logistic regression exist, such as probit regression, where an underlying latent normal variable is considered to give rise to the observed binary outcome, after dichotomization, rather than a continuous logistic variable.

Of course, one does not need to be restricted to a single predictor variable. For instance, both treatment allocation and sex of the human subject might be of interest. In such cases, most of the well-known methods easily extend. One-way ANOVA extends to two-way or even multi-way ANOVA. Simple, or single, linear regression extends to so-called multiple regression. Most other techniques, such as logistic regression, are easily extended to encompass multiple covariates. It has to be noted that, while simple in theory, methods for multiple covariates require great care since particular issues are raised that do not occur otherwise. Indeed, such issues as collinearity arise only for multiple covariate models. Often, not all predictors are on equal footing. For example, the relation between an exposure and a disease is of interest, while another variable is merely a confounder. This issue needs careful consideration in all non-randomized settings, such as epidemiological or otherwise observational studies. Thus, model building and interpretation of (regression) coefficients require both expertise as well as subject matter knowledge.

Particular care is needed in cases where the outcome variable is a time to a certain event. In a life sciences context, this is often the time from the beginning of a study, birth, or start of randomization, until a certain medical event occurs, such as death, relapse of onset of disease, complete cure, pregnancy, etc. This methodological area is often referred to as survival analysis or lifetime data analysis. There are two main reasons why standard (linear) regression is seldom appropriate. First, survival times tend to show skew rather than symmetric distributions, unlike in the normal distribution. Second, and more important, is the potential occurrence of censoring, i.e., the follow up time for a subject is not sufficiently long in order to observe the actual survival time. In such a case, it is clear that the actual survival time exceeds the end of follow-up, i.e., the survival time is larger than the censoring time. This means that partial, or coarse, information is present. Nevertheless, such information needs to be included into the analysis. A lot of research has been devoted to develop parametric and non-parametric methods for the analysis of survival times in the presence of censoring.

Another important set of situations, different from the univariate settings considered thus far, occurs when several dependent variables are recorded simultaneously. This concept harbors a large and ever growing portion of statistical methodology of use in health sciences and elsewhere. The most classical setting is multivariate analysis, where different outcomes are measured on the same subject. Alternatively, the same measurement can be taken on the same unit or on correlated units. Examples include longitudinal studies, where the same measurement is repeatedly made over time, spatial statistics, where the connection of a response to its geographical location is of interest, clustered data (e.g., in animal litters or in family studies), hierarchical survey data, such as arising from multistage or cluster sampling, etc. A particular area of such dependent, or hierarchical, data is given by meta-analysis in clinical trials, where information is pooled from several clinical trials. Apart from general methodological considerations, one is then confronted with pragmatic issues such as which study to include, etc.

Section 2 is devoted to classical univariate modeling, including linear regression, generalized linear models, and logistic regression. Section 3 offers a perspective on hierarchical data, encompassing multivariate as well as repeated measures and multilevel modeling. Finally, some comments about survival analysis are made, and its connection to the other modeling frameworks is highlighted.

2. Linear Regression, Generalized Linear Models, Exponential Family, and Logistic Regression

2.1 Gaussian Outcomes

The analysis of continuously distributed responses, especially when they are normally distributed, has received a lot of attention. Next to the *t* test, analysis of variance and linear regression have received a lot of attention. The general linear regression model is customarily written as

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i ,$$

where Y_i is a response variable for subject $i = 1, \cdots, N$ in a study, X_{ij} is the value for the j th predictor variable and ε_i is an error term. There are some important special cases. For example, when $p = 1$ then there are no covariates and the one-sample problems results. When $p = 2$, so-called simple or single regression is obtained, where the outcome variable is regressed on a single covariate. When all of the covariates are dummy variables (0 or 1 depending on whether a certain characteristic is absent or present within a subject), possibly resulting from a multi-categorical covariate, then analysis of variance is obtained. Analysis of variance and regression are often treated as different entities in introductory texts. This makes sense because on the one hand linear regression generalizes ANOVA, while on the other hand a larger number of results and tools is available for the ANOVA setting than for the more general regression setting. In a sense, ANOVA refers to categorical covariates, whereas regression focuses on continuous covariates, or a combination of continuous and categorical covariates.

Regarding the error term, two views can be taken. First, one can restrict attention to specification of its moments only. Most commonly, one assumes a zero mean and a constant variance, σ^2 say. This results in the so-called ordinary least squares (OLS) approach to linear regression. Alternatively, the error term, and subsequently the response variable itself, can be considered normally distributed. In the first case, sampling-based or frequentist inference results, in the second case, full maximum likelihood follows. Both approaches yield the same parameter estimates and almost the same estimates of precision, given that they are asymptotically equivalent. The OLS approach is valid under weaker assumptions, i.e. that the errors are not necessarily normally distributed, but if one is comfortable with the normal distribution, use can be made of fully parametric inference. This is one of many instances in statistical modelling: if one is prepared to make assumptions, more results become available, but the risk of incorrect assumptions is always present. This is why careful assessment of assumptions is important.

Throughout statistical analysis, and not only in linear regression, the normal distribution is omnipresent. Let us reflect on this phenomenon. For a simple random sample with just one outcome variable, the mean and the standard deviation, and/or the standard error of the mean, are easily computed. This is independent of the true distribution of the data. However, for some distributions, a mean and standard deviation will be less meaningful. This includes, for example, bimodal distributions. Even though they may have a mean, primary scientific interest may lie in identification of the two modes and other characteristics thereof. Another example is provided the Cauchy distribution, which does not have finite mean and variance.

The normal distribution has easy interpretations for the mean and standard deviation of samples drawn from it. The usual definitions of mean and standard deviation are the least squares estimators (and maximum likelihood estimators) of the population quantities. Very importantly, under regularity conditions, the sample mean converges to the location parameter of the true distribution, even if it is not normal. This is based on the so-called law of large numbers (central limit theory), which means, roughly speaking, that distributions of estimators from large samples show a normal spread, even when the samples themselves are drawn from non-normal distributions. In

addition, the researcher disposes of an alternative set of tools consisting of transformation methods. This allows to transform responses or residuals that are non-normal, to (more) normally distributed ones. For all of these reasons, the normal distribution is a convenient working paradigm in a number of statistical areas.

-
-
-

Bibliography

Abrams, K., Jones, D.R., Sheldon, T.A., Song, F., Sutton, A.J. (2000). *Methods for Meta-analysis in Medical Research*. New York: John Wiley [Excellent overview of meta-analysis in medicine].

Aerts, M., Geys, H., Molenberghs, G., and Ryan, L.M. (2002). *Topics in Modelling of Clustered Binary Data*. London: Chapman & Hall [This book discusses models and inference for repeated binary data, with a focus on clustered data].

Agresti, A. (1990). *Categorical Data Analysis*. New York: John Wiley [A standard text on categorical data, including generalized linear models].

Breslow, N.E. and Day, N.E. (1990). *Statistical Methods in Cancer Research (Vol. 1): The Analysis of Case-Control Studies*. Lyon: World Health Organization [This is a classic textbook on the analysis of data from case-control studies, including logistic regression and methods for analyzing contingency table data].

Cox, D.R. and Hinkley, D.V. (1990). *Theoretical Statistics*. London: Chapman & Hall [Background on the inferential and theoretical aspects of the approaches discussed].

Diggle, P.J., Heagerty, P., Liang, K-Y., and Zeger, S.L. (2002). *Analysis of Longitudinal Data*. New York: Oxford University Press [General text on repeated measures and longitudinal data].

Dunn, G. and Everitt, B. (1995). *Clinical Biostatistics*. London: Arnold [Basic introduction to biostatistical methods. The text is broad and qualitative in nature].

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Heidelberg: Springer-Verlag [This book provides an extensive overview of models for non-normal data].

Johnson, R.A. and Wichern, D.W. (1992). *Applied Multivariate Statistical Analysis*. London: Prentice-Hall [Presents a multitude of concepts and methods of multivariate analysis].

Kalbfleisch, J.D. and Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. New York: John Wiley and Sons [A standard text on survival analysis].

Kleinbaum D.G. (1996). *Survival Analysis, A Self-Learning Text*. New York: Springer-Verlag [This book presents basic information about survival analysis, including the Kaplan-Meier method, the log-rank test, mathematical models for survival analysis, computation of the hazard ratio from these models, and maximum likelihood estimation methods].

Krzanowski, W.J. (1988). *Principles of Multivariate Analysis - A User's Perspective*. Oxford Science Publications [A user-friendly introduction to modern multivariate statistical analysis].

Le, C.T. and Boen, J.R. (1995). *Health and Numbers*. New York: Wiley-Liss [Introductory text on statistical methods in the health sciences. The text is quantitative but easy to follow. It focuses on principles and simple methods rather than on modelling].

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*.London: Chapman & Hall [Classic reference text on generalized linear models].

Neter, J., Kutner, M., and Nachtsheim, C. (1996). *Applied Linear Statistical Models*. (4th ed.) Irwin: Chicago. [Classical introductory text on linear models (regression and ANOVA)].

Pagano, M. and Gauvreau, K. (1992) *Principles of Biostatistics*. Belmont, CA: Duxbury Press [General introductory text on biostatistical methodology].

Rothman K.J. and Greenberg, S. (1998). *Modern Epidemiology*. Philadelphia: Lippincott-Raven Publishers [Basic text on the application of statistical methodology in the context of epidemiology].

Rosner, B. (1994). *Fundamentals in Biostatistics* (4th ed). Belmont, CA: Duxbury Press [General introductory text on biostatistical methodology].

Shoukri, M.M. and Pause, C.A. (1999). *Statistical Methods for Health Sciences* (2nd ed). Boca Raton: CRC Press [General introductory text on statistical methodology in the health and life sciences].

Verbeke, G. and Molenberghs, G. (2000). *Linear mixed models for longitudinal data*, Springer Series in Statistics, Springer-Verlag, New-York [Overview of linear mixed models for longitudinal data, with a lot of examples].

Biographical Sketch

Geert Molenberghs is Professor of Biostatistics at the Limburgs Universitair Centrum in Belgium. He received the B.S. degree in mathematics (1988) and a Ph.D. in biostatistics (1993) from the Universiteit Antwerpen. Geert Molenberghs published methodological work on the analysis of non-response in clinical and epidemiological studies. He serves as an associate editor for Biostatistics and is Joint Editor of Applied Statistics. He is an officer of the Belgian Statistical Society and the Belgian Region of the International Biometric Society. He serves on the Executive Committee of the International Biometric Society. He has held visiting positions at the Harvard School of Public Health (Boston, MA). With Geert Verbeke, he is a co-author of books on longitudinal data.