SOME STATISTICAL ASPECTS
OF
THE GENERALIZABILITY OF OCCUPATIONAL HEALTH STUDIES

Design cover: Pennarts Graphics Design, Holland

Printed by: Casparie Heerhugowaard, Holland

SOME STATISTICAL ASPECTS
OF
THE GENERALIZABILITY OF OCCUPATIONAL HEALTH STUDIES

Statistische aspecten betreffende de generaliseerbaarheid
van studies uitgevoerd binnen de bedrijfsgezondheidszorg

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Erasmus Universiteit Rotterdam
op gezag van de rector magnificus
Prof. Dr C.J. Rijnvos
en volgens besluit van het College van Dekanen.

De openbare verdediging zal plaatsvinden op
woensdag 22 januari 1992 om 13.45 uur

door

Dirk Lugtenburg

geboren te Zuidland

PROMOTIE COMMISSIE

Promotor: Prof. R. van Strik

Overige leden: Prof. Dr Ir. J.D.F. Habbema
Prof. Dr J.J. Kolk
Prof. Dr W. Molenaar

.... incompleteness may be due to the method of investigation itself such as the double sample procedures in large scale sample surveys. If the measurement of the character to be estimated is expensive, a small sample of individuals is chosen to supply measurements of this character together with a number of concomitant characters which are relatively inexpensive to measure. Then a large sample is taken for the concomitant characters only. The former sample provides the regression equation and the latter the estimates of mean values of the concomitant characters which are substituted in the regression equation to obtain the estimates for the main character under study.

C. R. Rao (1956)

Aan mijn ouders

CONTENTS *page*

CHAPTER 1

INTRODUCTION

"... it is hardly an exaggeration to say that all studies published to date contain at least some systematic errors .... "

S. Hernberg (1986)

1.1 SOME STATISTICAL ASPECTS OF THE GENERALIZABILITY OF OCCUPATIONAL HEALTH STUDIES

Occupational health departments screen employees during voluntary periodic health assessments for possible early indications of ill health. These assessments are based, among other things, on reference ranges for liver parameters, kidney parameters and other blood parameters. After a period of, say, ten years, data thus collected can also be used for large-scale epidemiologic studies. In such studies, long-term risk assessments may be made, either for early death or for specific diseases such as ischaemic heart disease. Conclusions drawn from these studies can have a direct influence on the short-term advice to specific employees on the basis of findings from periodic health assessments, as well as on the contents of these assessments. Factors such as exposure to chemicals, noise and physical exertion are being studied in these epidemiologic studies (Monson, 1980). Furthermore, known determinants of disease in general populations may be studied separately in occupational populations, in order to investigate whether these determinants are less important in such healthy cohorts. Apart from epidemiologic studies, well-designed trials are currently being introduced in occupational health environments, in order to assess the effects of interventions in life style habits of employees.

Hartley and Hocking mentioned in 1971 that "the evil of incompleteness is most frequently encountered with data routinely collected .... such as .... based on surveys". Their remark, alas, is still valid. Incomplete records still constitute a major problem in studies based on data routinely collected at periodic health assessments. The quality of the available data is always debatable, too, and should be studied critically. Since in a large occupational health study, both problems are quite often encountered at the same time, they should be addressed by means of special methodology, e.g., through small studies on either incomplete observations or on the quality of data, the results of which can be used to adjust the conclusions from the large study for bias. Such a systematic approach of these problems of incomplete records and questionable quality of data enables generalizations from such large occupational health studies.

With regard to the principles of reliability of a study it is useful to recall two relevant concepts: internal and external validity (Karvonen and Mikheev, 1986). Internal validity is achieved when no systematic errors are present within a study, whereas external validity goes beyond the population actually studied. Internal validity of a study comprises validity of selection, information and comparison. To ensure validity of selection, the probability of selection of individuals should not depend, in any way, on the parameter under study. This refers directly to the problem of missing data. Validity of information means that information gathered from all the subjects is similar, irrespective of the category to which they belong, which refers to the problem of quality of data. The validity of comparison has to do with correctly defining reference populations and is embedded in the previous two items. The present thesis discusses methods for assessing validity of selection and validity of information. To this end, it presents methods that may help to remove bias caused by either incomplete records or poor quality of data. Extrapolation of results or methods towards the area of external validity is beyond the scope of this thesis.

The problems of missing data and misclassification are also conceivable in small health studies. Here, however, they should be avoided by taking adequate measures at the design phase. If this is not done, any statistical significance observed might be due to "possibly erroneously collected data", which, certainly for large-scale studies, would not be acceptable as a way out. An additional problem with small studies is that asymptotic results cannot be used in an adequate description of the behaviour of test statistics. Generalizability can only be achieved if this behaviour is known, at least under the null hypothesis. This will be a separate topic in this thesis.

1.2 THE SIMILARITY BETWEEN INCOMPLETENESS OF OBSERVATIONS AND DUBIOUS QUALITY OF DATA

The most simple way of dealing with incomplete records is to remove them before calculating the statistics on which interest focusses. The conclusions, however, are then to be ristricted to the subsample of subjects of whom all data are (or can be) known. These strong restrictions can only be removed through the use of a strategy to deal with missing data. Three general approaches are found in the literature. Firstly, there is gaining actual information on the missing-data-generating process by resampling in the subsample of individuals with incomplete records, resulting in adjustment of conclusions derived from the sample of complete records. Secondly, postulating various missing-data-generating processes to be operative and studying their effects on important conclusions. Thirdly, imputing 'probable values' for missing data, and performing standard statistical analyses. In this thesis, a review of relevant recent literature will be presented, from which it will become clear that, if no extra information can be gathered by the first approach, a sensitivity analysis by the second approach may help to predict any dependence of relevant findings on assumptions concerning the missing-data-generating mechanism.

In the area of misclassification there are again three general approaches to be distinguished. Firstly, as in the missing data context, extra information can be gathered and used for adjustment of relevant statistics. For example, the concepts of sensitivity and specificity are useful to represent the proportions of possible misclassifications in the special case of dichotomous variables. Begg (1987) discusses problems in the design of studies made to assess these error rates when using a fallible diagnostic instrument, and Marshall (1989) discusses methods to combine results of several fallible instruments. The use of such estimated error rates can give an improvement in the estimates for prevalence (Lew and Levy, 1989), odds ratios (Espeland and Hui, 1987) and risk ratios (Clayton, 1985 and Greenland, 1989). The second approach is to postulate specific structures to have created the non-ideal data, for example in the case of continuous variables with measurement errors. Reference is made to a series of papers from a workshop on errors-in-variables published in *Statistics in Medicine* (Byar and Gail, 1989). The third general approach originates from the area of robust statistical analysis. When dubious data are suspected to be present, a possible way to address the problem is using a robust analysis. Such analyses are fairly insensitive to errors in data-collecting or data-filing. Well-known robust statistics are the median and the median absolute deviation as alternatives to the sample mean and the sample standard deviation. Recent publications discuss research on "robustified" regression analysis (Rousseeuw and Leroy, 1987) as well as on logistic regression (Künsch, Stefanski and Carroll, 1989) and on proportional hazards models (Lin and Wei, 1989).

From this short summary it should be clear that the first two approaches mentioned are applicable to both problems, since they consider generating mechanisms. The first approach is based on resampling, which might give evidence that a specific mechanism is operative and might, consequently, make adjustments feasible. The second approach is to postulate various mechanisms, with consequences calculated in some sort of sensitivity analysis. Therefore, both problems of generalizability can be addressed in a similar way. When data are missing, the problematic observations are easily recognized and isolated. When dubious data occur, the problem might be in any of the observations, which in a sense makes it more difficult to tackle.

1.3 SMALL-SAMPLE BEHAVIOUR OF STATISTICS

Two general approaches can be distinguished in the assessment of the small-sample behaviour of a statistic, if asymptotic results are no longer assumed to be valid. As most asymptotic results are correct up to order $O(1/n)$, the first approach is to develop statistics that are correct up to a higher order. In general, the approach stems from the fact that the expected value of a statistic (under a specified hypothesis) can be calculated up to any order, if required. In practice, it is only approximations up to the second order that can usually be calculated. Most authors in this research area use such second-order approximations to improve on the first-order results. Correction factors like Bartlett's (1937) can be applied for that purpose.

If statistics, correct up to the second order, are still not acceptable, another approach is to generate all possible realizations of the statistic and then to calculate *exactly* the statistic of interest. This method has been used for calculating p-values of non-parametric (rank-)statistics under the null hypothesis of a uniform distribution of all observed ranks (Lehmann, 1975). It is possible to extend this approach towards parametric models postulated to have generated the data.

1.4 EXAMPLES OF OCCUPATIONAL HEALTH STUDIES

The present thesis discusses two examples of occupational health studies conducted at Shell Nederland Raffinaderij B.V./Shell Nederland Chemie B.V. at Pernis (Shell Pernis for short). The first study concerns the effectiveness of a "back school" education programme. One of the most frequent complaints in an industrialized population is low back pain. Up to eighty per cent of such a population may be troubled by it at some time or other (Ingber, 1989). Most of the complaints eventually disappear, even without treatment. However, the percentage recurrences is high: between 56 and 90 per cent (Knibbe, 1989). A "back school" is a course intended to improve the knowledge and skills of patients through education and selective training (Keysers et al, 1989). Up till now, back schools have been used as a therapy for patients with acute or chronic low back pain. A trial was started at the occupational health, hygiene and biomedical department of Shell Pernis in 1990 to study the possible effects of a back school education programme on the percentage recurrences of back pain. This trial is a secondary prevention trial with the underlying hypothesis that intervention with a back school education programme will give an improved flexibility of the spine. The overall hypothesis is that an increased flexibility of the spine reduces the probability of recurrence. This should be reflected in a reduced number of sickness absence days.

Problems concerning missing data that may be encountered in such a trial are threefold. Firstly, some people who are invited may refuse to participate, well before randomization. This will affect generalization of results to the population of all the persons invited. Secondly, people may refuse to participate in the programme after randomization, in which case they are labelled drop-outs. All relevant information from these drop-outs would still be collected in order to enable intention-to-treat analysis. Thirdly, there may be persons on whom no sickness absence data can be retrieved, because they left the population. A worst-case approach would then be taken, in which these persons are assumed to have had sickness leave since their departure. In designing our trial, we decided to study the phenomenon of non-participation before randomization in a separate study; the results of which are presented in this thesis.

The second example concerns a study of liver function tests, such as γ-glutamyltransferase (GGT) determinations, which form part of periodic health assessments. This study is part of a project to explore possibilities of using these data to predict the state of health of an employee within two years after liver function determination. As a measure of health condition, data from absenteeism records were used. Four different potential predictors are derived from the liver function tests, which makes it possible to distinguish between cross-sectional and longitudinal characterizations. In the cross-sectional approach, persons are labelled as having a high liver function test value if it is outside the reference range, based on interindividual variations alone. The longitudinal approach requires at least two measurements of the liver function to be made on a person. If the change in value represents a significant biologic change, this may be seen as a potential predictor. If four (or more) test results per individual are available, the three previous measurements can be used to calculate a person-related interval of reference values; the persons whose fourth value is outside his interval are labelled. A fourth approach is to calculate the variation with time, and to derive a potential predictor from a significant trend.

Problems may arise if persons do not show up at a periodic health assessment to have their liver function tested. The degree of participation for the period 1985 - 1989 was studied on the basis of a total of 5279 male persons. It was found that, during this four-year period, between 88 and 93 per cent of the employees in different departments actually attended the periodic health assessment. For employees leaving

Shell Pernis within this period, attendance rates were somewhat lower, although for each year only about 6 per cent. This indicates that about 84 per cent of these persons, too, would have attended the periodic health assessment, if they had been present for the whole period. This result makes validity of selection plausible. Therefore, the conclusions from this study are considered to be valid for the total population, between 1985 and 1989.

In a longitudinal study it is necessary to know the accuracy of the specific tests, such as GGT determinations, performed during the relevant period. Therefore, the long-term analytical bias of GGT determinations in the biomedical laboratory of Shell Pernis was assessed, in comparison with results found in other laboratories using the same analytical methods, and the results are reported in this thesis.

1.5 STRUCTURE OF THE THESIS

The present thesis discusses both the methodology as developed, and applications in practice. Methods of dealing in the statistical analysis with the occurrence of missing data are presented in Chapter 2 as a review of recent literature on this topic. Complete issues of *Biometrics, Biometrika, Journal of the American Statistical Association, Statistics in Medicine* and the *American Journal of Epidemiology* have been screened from 1986 up to 1991 on this point. Relevant literature from other sources has also been included.

Two important statistics used in occupational health epidemiology are relative risk and prevalence. A study of possible effects on the relative risk in the event of inequality of the fractions of deaths and alives of which vital status is known, is described in Chapter 3. Adjustment procedures are proposed for calculating conservative confidence regions of the "true" relative risk, using only little extra information. With respect to prevalence, adjustment procedures have been proposed by several authors. An attempt to combine the imputation method with postulating a missing-data-generating mechanism is presented in Chapter 4, where missing-data-corrected estimates for the prevalence are presented. These estimates are compared with respect to their likelihood through a likelihood ratio test.

Two general approaches concerning small-sample behaviour of statistics are exemplified in Chapters 5 and 6. Many parameter estimates that are used in the analysis of continuous outcomes, have well-known small-sample behaviour. This is not true for the two statistics that are often used in the analysis of categorical variables. Therefore Chapter 5 discusses an approximate Bartlett adjustment for testing the goodness-of-fit of multinomial regression models. This adjustment is applicable to any multinomial regression model, since it only needs estimated category frequencies and does not need any derivatives to be calculated. To illustrate the second approach, the consequences of calculating exact p-values for the Wilcoxon-Mann-Whitney test statistic in the case of ordinal response variables are presented in Chapter 6 with up to ten observations in each sample. From this it appears that the performance of conditional statistics, based on observed ranks only, is not so good as that of unconditional statistics based on an underlying, plausible, distribution.

The trial protocol of the back school education programme intervention study included an investigation of differences between subgroups of participants and non-participants, in order to gain some understanding of the distinctions between those who want to participate and those who do not. If a determinant of participation is related to the primary outcome variable, adjustments have to be made to allow generalization of conclusions. Results obtained are presented in Chapter 7.

Chapter 8 describes a cross-sectional study of person-related determinants of GGT levels. Model-based person-related tolerance regions are developed to facilitate detection of potential determinants. The main finding is that persons with an observed GGT value under 30 U/l in general do not need further screening, whereas persons with an observed GGT value above 60 U/l do. For persons with intermediate GGT values person-related determinants should be taken into consideration. Thus, each individual is either "labelled" to undergo further follow-up or not. Persons who are so labelled are to be compared with persons who are not labelled in the actual liver-function project. In Chapter 9, a model is postulated for correcting for long-term analytical bias of a specific laboratory compared to a pool of comparable laboratories in an external quality control programme.

1.6 REFERENCES

Bartlett MS (1937). Properties of sufficiency and statistical tests. *Proc. R. Soc. A*, 160, 268-282.

Begg CB (1987). Biases in the assessment of diagnostic tests. *Statistics in Medicine*, 6, 411-423.

Byar DP and Gail MH (1989). Introduction errors-in-variables workshop. *Statistics in Medicine*, 8, 1027-1029.

Clayton D (1985). Using test-retest reliability data to improve estimates of relative risk: an application of latent class analysis. *Statistics in Medicine*, 4, 445-455.

Espeland MA and Hui SL (1987). A general approach to analyzing epidemiologic data that contain misclassification errors. *Biometrics*, 43, 1001-1012.

Greenland S (1989). On correcting for misclassification in twin studies and other matched-pair studies. *Statistics in Medicine*, 8, 825-829.

Hartley HO and Hocking RR (1971). The analysis of incomplete data. *Biometrics*, 27, 783-823.

Ingber RS (1989). Iliopsoas myofascial dysfunction: a treatable cause of "failed" low back syndrome. *Arch. Phys. Med. Rehabil.*, 70, 382-386.

Karvonen M and Mikheev MI (1986). *Epidemiology of occupational health.* WHO Regional Publications, European Series. No 20. Copenhagen.

Keysers JFEM, Bouter LM, Steenbakkers WHL and Meertens RM (1989). Methodological quality and inter-comparability of investigation into the effectiveness of back schools (in Dutch), *Fysiotherapie*, 99, 112-116.

Knibbe JJ (1989). Epidemiology of lower back complaints - a study into the need for secondary prevention (in Dutch), *Fysiotherapie*, 99, 169-174.

Künsch HR, Stefanski LA and Carroll RJ (1989). Conditionally unbiased bounded-influence estimation in general regression models, with applications to generalized linear models. *J. Am. Statist. Assoc.* 84, 460-466.

Lehmann EL (1975). *Nonparametrics, Statistical methods based on ranks.* Holden-Day, San Francisco.

Lew RA and Levy PS (1989). Estimation of prevalence on the basis of screening tests. *Statistics in Medicine*, 8, 1225-1230.

Lin DY and Wei LJ (1989). The robust inference for the Cox Proportional Hazards Model. *J. Am. Statist. Assoc.* 84, 1074-1078.

Marshall RJ (1989). The predictive value of simple rules for combining two diagnostic tests. *Biometrics*, 45, 1213-1222.

Monson RR (1980). *Occupational Epidemiology.* CRC Press. Boca Raton, Florida.

Rousseeuw PJ and Leroy AM (1987). *Robust regression and outlier detection.* John Wiley and Sons, New York.

CHAPTER 2

MISSING-DATA-GENERATING MECHANISMS AND THEIR EFFECTS ON THE SELECTION OF METHODS OF STATISTICAL ANALYSIS: A REVIEW *

SUMMARY

In medical and public health research, conclusions concerning a specific population are usually derived from a sample which is assumed to have been randomly drawn and is therefore considered representative of this population. If missing data occur in the sample, the question arises whether the observed data are still sufficiently representative to base conclusions upon. In many situations, inference is based only upon those individuals for which all variables have been observed, the "complete-cases-only analysis". Generalization of the conclusions to the total population can only be done by assuming a very specific mechanism for generating the missing data. In this review chapter it will be argued that such an assumption is unrealistic. However, if additional information on the mechanism is available, correction for the possible bias due to the occurrence of missing data becomes feasible. In recent literature, several options for postulating such mechanisms have been mentioned. Besides introducing these mechanisms, this chapter discusses their merits with regard to an increase in precision and unbiasedness of results, as well as their drawbacks. The major drawback is the fact that underlying assumptions cannot usually be verified with only the observed data at hand, but have to be based upon additional information. If no additional data can be collected, a sensitivity analysis should be performed to see whether relevant conclusions change with changes in the underlying conjectured missing-data-generating process.

* After: Lugtenburg D and Van Strik R (submitted) *Statistica Neerlandica*

2.1 INTRODUCTION

In empirical investigations on humans there are only few situations where missing data never occur. Therefore, it is important to know whether valid conclusions can be derived from a sample selected randomly from a population of individuals, if some observations are lacking. Methodology concerning the analysis of samples with missing data is developing rapidly. This chapter presents a review of relevant literature, leaving it to the reader to consult the standard text on missing data by Little and Rubin (1987) for additional information.

The approach taken most frequently for dealing with missing data in the analysis of a study is to remove every individual from the sample with at least one missing value. One of the reasons for using this "complete-cases-only analysis" is that most software packages need complete records to be applicable. Let us consider an illustrative example. Suppose that in a specific research project, ten variables are to be measured in a large sample of individuals. Let us further suppose that, for various reasons, 5 per cent of the values turn out to be missing for each variable. If these percentages are mutually independent, only 60 per cent of the individuals (0.95^{10} x 100 %) will have complete records. In the complete-cases-only approach it is these individuals which are used in the analysis. It will be clear that in this situation, confidence intervals for estimated parameters will generally be much wider than they would have been if all the data had been available. If conclusions from this analysis are suggested to be valid for the total population, the underlying assumption is that individuals with full records are still a random sample of that population, implying that individuals with incomplete records are a random sample too, of course. If 60 per cent of a random sample is included in the analysis, this assumption may be far from justifiable in usual practice. Hence, parameter estimates and significance tests may yield biased results and at least will lack precision. There is a clear need for more flexible algorithms that do not remove data automatically from the sample to be analysed.

It is useful to distinguish three general approaches to deal with missing data. The first approach is to use a specific assumption for the underlying missing-data-generating process, which enables us to estimate the relevant parameters, using all the measurements made. If the missing-data-generating process depends only on completely observed variables, and analyses are based on the likelihood principle, conclusions may be considered to be valid for the total population. This approach is exemplified in section 2.2, in which the likelihood principle is clarified as well. A well-known algorithm for calculating estimates following the likelihood principle is the Expectation Maximization

(EM) algorithm of Dempster, Laird and Rubin (1977). Because of the widespread use of this algorithm, it will be discussed briefly in section 2.3, by means of a theoretical example. The second approach is to impute the missing data and then use the complete sample to estimate the parameter of interest using a complete-cases-only analysis. This approach is discussed in section 2.4. The third approach, which is the preferred one, is to gather extra information about the incomplete records by resampling. In this way, data can actually be used to test assumptions on the missing-data-generating process.

Some applications of these three approaches are presented in section 2.5, namely for estimating either the prevalence of a binary feature, or the mean value of a continuous variable, when only the outcome variable is subject to incomplete observation. With respect to repeated outcome measurements, section 2.6 discusses relevant literature. In section 2.7, the situation where only explanatory variables are subject to missing data, is considered. Finally, section 2.8 presents conclusions and recommendations.

2.2 MISSING-DATA-GENERATING PROCESSES AND LIKELIHOOD-BASED ANALYSES

If missing data occur strictly randomly amongst the individuals of the total sample, the data are said to be missing-completely-at-random (MCAR). Complete-cases-only analyses, which use all individuals with complete records, will lead to valid conclusions for the total population. Before considering a less stringent mechanism, we will present a formal introduction to likelihood-based analyses, of which linear regression, logistic regression and log-linear regression are some well-known examples.

In many analyses, the outcome variable is assumed to follow a specific known distribution, subject to some unknown parameters, e.g. a normal distribution with unknown mean and variance. Suppose a specific parameter is written as a function of explanatory variables, known except for some parameters, and that these latter parameters have to be estimated. In order to do so, one can express the probability (or likelihood) that the observations were gathered from such a specific distribution as a function of the unknown parameters. Likelihood-based analyses maximize this function with respect to the unknown parameters. Estimates of the parameters that are calculated in this way are called maximum likelihood estimates.

If missing data occur, the mechanism that is postulated to generate the missing data is included in the likelihood function. Maximization of the function now means that one should also optimize over parameters related to that mechanism. If the occurrence of missing data for a specific variable per individual is related only to one or more other variables that are observed for this individual, these missing data are said to be missing-at-random (MAR). On the basis of this assumption, Rubin (1976) has shown that the likelihood can be split into a missing-data-dependent part and an observed-data-dependent part. The two parts are mutually independent. Consequently, the likelihood of the observed-data-dependent part can be maximized irrespective of the missing-data-dependent part of the likelihood. This is expressed in the statement that the missing-data-generating process is "ignorable", not meaning that maximization is easy, but only that estimates are valid for the total population.

If the occurrence of missing data for a specific variable per individual is related to the (unobserved) value of that variable, the missing data are said to be missing-not-at-random (MNAR). In that case likelihood-based analyses should include this mechanism in the likelihood in order to obtain valid results. The most generally known example of this is survival analysis with right truncation, in which each individual with a truncated survival time, t, is analysed as having a survival time of at least t.

The three mechanisms, MCAR, MAR and MNAR, form a complete and disjunct division of all the missing-data-generating processes. A summary is presented in Table 2-I. With MCAR the underlying conditions to be fulfilled are the most severe, but make analysis easy. MNAR has the least stringent underlying conditions, but often makes analysis difficult. Irrespective of the ease of analysis, it is obviously important to pursue an accurate description of the underlying missing-data-generating process. A paper by Stasny (1986) provides an example of violation of this principle. When estimating the increase in unemployment due to persons losing jobs, he postulated his missing data to be MNAR, whereas in fact they were analysed to be MAR. Therefore, he mistakenly related nonresponse to observed previous employment status instead of to unobserved present employment status.

TABLE 2-I

Summary of missing-data-generating processes and examples where conclusions are valid for the total population

Process	Example
MCAR	complete-cases-only analyses
MAR	likelihood-based analyses
MNAR	survival analysis with right truncation

For discriminating between the several mechanisms, appropriate tests are indispensable. For example, to decide whether MCAR applies, a simple test is to model the probability of becoming a missing observation as a function of relevant variables. If one of these variables is found to be a predictor of that probability, the data are not MCAR. For multivariate data, Little (1988) proposes an omnibus test for testing missing data to be MCAR. Several authors have proposed tests of the hypothesis that missing data are MAR. However, such tests can never be applied without extra information gained either through resampling or through using external information. For example, Diggle (1989) ignores this principle in presenting a method for testing whether missing data are MAR within each of several treatment groups. However, if data are not observed at a specific moment in time, because the value at the preceding moment was high, then the test statistic of Diggle erroneously leads to the conclusion that the missing data are MNAR, whereas in fact this is a classical example of missing data being MAR.

2.3 THE EM ALGORITHM

The most popular algorithm for calculating maximum likelihood estimates when data are missing is the EM algorithm. This algorithm nearly always converges to the correct maximum. As it does not use second-order derivatives, this algorithm does not converge so quickly as a method like the Newton Raphson method, which, however, does not converge always.

The EM algorithm consists of two parts that are repeated iteratively until convergence is reached. Before implementation, a representation of the maximum likelihood estimates is derived as if all data were present. These estimates, often to be described as a function of sufficient statistics, i.e. statistics in which all information of the data is provided under the condition of a specific underlying distribution, are the basis for the EM algorithm. In the expectation step (step I) of iteration $j+1$, the expected values of the sufficient statistics are calculated as functions of observed values and of estimates of the model parameters from iteration j. In the maximization step (step II) of iteration $j+1$, the new estimates for the model parameters are calculated by maximizing a function of these expected values. Dempster, Laird and Rubin (1977) present several applications, including analyses with missing data. They also present a proof that the iterative procedure converges to the maximum likelihood estimate.

Let us consider a simple illustration of the performance of the EM algorithm as discussed by Little and Rubin (1987, pp 130-131). Suppose a random sample of n observations is taken on a normally distributed variable with unknown mean and variance. Let *S1* be the arithmetic mean of the sample values and *S2* the same of the squared sample values. The maximum likelihood estimate for the mean is *S1* and the maximum likelihood estimate for the variance is $(S2)$-$(S1)^2$. Therefore, two functions of the observations y_i are relevant, being (in algebraic notation): $S1=(\Sigma\ y_i)/n$ and $S2=(\Sigma y_i^2)/n$. These functions are sufficient statistics, since they contain all sample information under a normal distribution whose two parameters (the mean and variance), can be estimated from *S1* and *S2*. Now suppose that only *m* observations are known, i.e. $(n$-$m)$ observations were lost. When, at the j^{th} iteration, estimates $\mu(j)$ and $\sigma(j)$ of the mean and variance are known, the expected value for *S1* is a function of the observed values y_i and $\mu(j)$:

$$E(S1)=\frac{(n-m)\mu(j)+\sum_{i=1}^{m} y_i}{n} .$$

With regard to the expected value for *S2*, it is useful to note that for a complete sample the expected value of the squared standard deviation is:

$$E(S2)-E(S1)^2=\sigma^2 .$$

This has to be imputed in the incomplete part, leading to the expected value for S2:

$$E(S2)=\frac{(n-m)(\mu(j)^2+\sigma(j)^2)+\sum_{i=1}^{m} y_i^2}{n} .$$

The new estimates for the mean and the variance are given by:

$$\mu(j+1)=E(S1)$$

and

$$\sigma(\ j+1)^2=E(S2)-E(S1)^2 .$$

A point of interest is that this approach is not an imputation method with estimated values as imputations for missing values. The term including $\sigma(j)$ in the above equation for *E(S2)* can be considered proof for this statement. There are, however, applications where the estimated value is imputed, viz. if sufficient statistics are sums of observations, e.g. models in which the response variate follows a multinomial distribution.

2.4 IMPUTATION METHODS

A commonly used imputation technique is the imputation of means. The mean of the variable for which missing data occur is imputed as surrogate for the missing data. This, naturally, leads to strong underestimation of the variance of parameter estimates if the fact, that such surrogate observations by definition contribute zero to the estimated variance, is ignored. To obviate this problem, a random error term may be generated and then "added" to a simple imputation. Many of the imputation techniques currently known follow from the assumption that missing data are MAR. Chiu and Sedransk (1986) have shown, however, how information from other sources, about how the data became incorrect or incomplete, can be included to permit a general specification of the nonresponse model. Some imputation procedures for estimating the population mean will now be discussed under the condition that missing data are MAR.

Little and Smith (1987) estimate the mean and covariance structure of a set of variables through a stepwise algorithm. First, observations with missing or highly unlikely (extreme) data are located, after which the corresponding variables are isolated and lacking observations are imputed, including an additional random error term. The problem of underestimation of variances, however, is still there because the imputed values are treated as if they were observed, yielding confidence intervals that are too narrow because the variability from not knowing the missing values is ignored (Rubin and Schenker, 1986). The latter authors present a multiple imputation technique as an alternative in which two or more complete data sets are created by imputing the missing data several times. From every complete data set, the relevant population parameter is estimated. Then, the combined estimate for the population parameter is a function (mostly the mean) of the sample estimates. The estimate of the variance includes both the within-imputation variance and the between-imputation variance. For this purpose, simple multiple imputation methods may be as effective as more complicated ones (Heitjan and Rubin, 1990).

Imputation methods are subject to criticism because data are "created" on the basis of observed data. However, on the assumption of missing data to be MAR, there is no objection. Additional research has to be done to compare these procedures with direct likelihood approaches as discussed in the previous section.

2.5 ANALYSES CONDUCTED WHEN MISSING DATA OCCUR IN THE OUTCOME VARIABLE

In this section, we review some methods of analysis to be used if missing data occur only in the response variable, as in follow-up studies where baseline data are collected on all individuals and follow-up is only possible for a subset of this baseline cohort. Here, commonly used population parameters are means for continuous variables and prevalence and incidence for binomial variables. If missing data are not MCAR, it is possible under specific circumstances to correct for the bias in estimates of parameters. The literature on estimating prevalence and means in such situations is growing. A general method for adjusting for nonresponse has also been described (Alho, 1990). We will now consider methods for estimating prevalence, followed by methods for estimating means.

For estimating the prevalence of a feature of interest, imputation and likelihood-based methods are generally used simultaneously. For example, in estimating the incidence of home injury deaths, it has been found that a large proportion of reports on injury deaths do not specify the place of occurrence (Conn, Lui and McGee, 1989). By means of a logistic regression model, the probability of having a home injury was estimated and used in a one-time imputation phase. Relaxing the corresponding assumption that missing data are MAR is possible by specifying "certainty rates" for the proportion of the missing data having this feature, which are then imputed. For example, if all missing data are supposed to have this feature (e.g. a specific disease), an upper limit on the estimate for the prevalence can be calculated. If none of the missing data are supposed to have the feature, a lower limit on the estimate for the prevalence is obtained. Dinse (1986) discusses a nonparametric estimator that uses "certainty rates" as introduced above. By postulating a model that incorporates both extreme "certainty rates" together with more plausible alternatives, Lugtenburg and Mulder (accepted for publication in *Kwantitatieve Methoden*) advocate likelihood ratio tests as performance criteria for the various imputation options.

Besides imputing the missing data, collecting information on the missing-data-generating process is a third approach for dealing with missing data. Sampling persons within the group of individuals with incomplete follow-up will present this additional information. For example, in wildlife animal counting, this approach is used to estimate the total number of missed observations, after a first counting. Zeh et al. (1986) advocate this approach by not using only a first camp, but also a second camp to record the number of bowhead whales, together with a measure of the visibility. They use these data to estimate the size of the stock. If another way of sampling were feasible, without problems like visibility, maybe more expensive and therefore not to be used as a standard method, it would be possible to compare the two methods and end up with a formula to correct for visibility. Steinhorst and Samuel (1989) assume that visibility problems do not come up in counting radio-labelled animals and use a double sampling method to adjust for visibility bias. In a similar category falls a paper by Aebischer (1986), who uses a sight-record resampling method of eider ducks to estimate the proportion of uncatchable animals, which can be found in situations where a specific group is "trap shy" for a specific sampling strategy.

Sometimes the mean of a specific variable has to be estimated and a sample at hand is not representative of the population due to imbalance for another variable related to the variate of interest. Adjustment for this nonresponse bias is possible by calculation of subgroup means, stratified on that variable, and then reweighting the means with respect to the population distribution of that variable. This is referred to as a "direct adjustment". However, if many variables are related both to percentage missing and to the relevant variate, it is preferable that a model-based direct-adjustment approach be taken. By this means the probability of being missed is modelled and the stratification now is on the expected probability to become missing (Rosenbaum, 1987). This strategy can be applied if missing data are MAR. As was discussed in the previous section, imputation techniques can also be used to estimate the mean.

2.6 MISSING DATA IN REPEATED MEASUREMENTS

Situations often occur, where repeated measurements are made on several subjects and inference has to be based on all individuals included in a study, e.g. in trials on the effectiveness of a drug. If the repeated measurements can be condensed into one figure per experimental unit, a "contrast", these contrasts can be used in a between-subjects analysis (Berk, 1987). For example, for the purpose of analysing individual trends, a regression slope may be calculated on each individual as such a contrast. If high values of the variate are more likely to cause missing data, the estimated slope for persons with a steep regression will have a larger variance, because of less data collected, than for persons with a less steep regression (Little, 1988, commentary). Of course, if the slopes are combined, the overall regression slope shrinks towards zero. Several options are compared by Wu and Bailey (1988) to correct for this shrinkage, including a method to model the censoring process (Wu and Carroll, 1988).

Sometimes, however, there are no prior indications to define such a contrast. Other outside information may then be available for use in specific analyses. For example, in an analysis of the total number of relapses, Davis and Wei (1988) use the fact that repeated measurements can only increase. Under a specific nested structure, namely that data missing *at* time t also means that for this experimental unit data are missing *after* time t, Kenward (1987) models the response variate as a function of preceding observations assuming the data to be MAR. The "added significance due to a new measurement" can be assessed in this way and it is possible to determine the moment from which two groups are significantly different. Murray and Findlay (1988) show the use of an equivalent approach (which was incorporated in the protocol, *a priori*) to the correction for the bias in a hypertension trial, caused by drop-outs due to the high value at a previous observation. In a reaction to this procedure, Lewis (1988) presents some alternative methods to handle missing data in that situation. The method of Murray and Findlay seems preferable, however, in answering the question of effectiveness for a subpopulation with a degree of hypertension that does not rise too high while under medication. For categorical data, Lachin and Wei (1988) present estimates for the log odds ratio and log relative risk when two independent groups are measured repeatedly.

Also, models considering separate time points, by defining a specific relationship at each moment in time, have been advocated. The information from all these moments may be combined to search for trends over time. This approach is used both for continuous variables (Wei and Stram, 1988) and for ordinal variables (Stram, Wei and Ware, 1988).

In some cases it may be necessary to model the total multivariate space by estimating a multivariate mean and covariance matrix. For this situation, the BMDP program 5V is available, following an EM algorithm. Jennrich and Schluchter (1986) present several well-known covariance structures and Rochon and Helms (1989) discuss an ARMA covariance structure following Box and Jenkins (1976). These approaches assume the missing data to be MAR. More recently, Brown (1990) used several generalized censoring mechanisms which (in some undefined way) depend on the outcome, to present estimates for the multivariate mean and covariance structure that are robust to certain departures from normality. Such estimates often end up with some added uncertainty about variance and covariance estimates beyond the normal MAR solution, because of the less stringent assumptions being made.

2.7 MISSING DATA IN EXPLANATORY VARIABLES

If a linear regression analysis is required, Azen, Van Guilder and Hill (1989) show that the EM algorithm can often ideally be used to estimate regression parameters and to impute values under general censoring mechanisms. If the occurrence of missing data of covariate values is related to treatment only, imputing values using a probability imputation technique will remove bias in estimating treatment effects, as Schemper and Smith (1990) have shown. In an interesting paper by Simon and Simonoff (1986), regression parameters are described as functions of a statistic F. If the statistic F takes the value 1, the missing data are MAR, otherwise they are not. In this way it is possible to show the dependence between regression parameters and assumptions concerning the mechanism that generates the missing data.

For survival models, omitting a totally balanced covariate can seriously bias the estimated treatment effect, if the covariate is strongly prognostic; this implies the need for thorough comparisons of adjusted and unadjusted analyses, as Chastang, Byar and Piantadosi (1988) have shown. If survival data are censored and only partially observed covariates are present, the EM-algorithm approach of Schluchter and Jackson (1989) can be applied, assuming noninformative censoring mechanisms.

2.8 DISCUSSION AND RECOMMENDATIONS

In this chapter, the three common missing-data-generating mechanisms: MCAR, MAR and MNAR, are discussed. Because of limited space, other ignorable missing-data-generating mechanisms are not considered. For example, Goffinet (1987) has presented alternative conditions for ignoring the missing-data-generating process. Furthermore, the relevant area of longitudinal analyses was not included explicitly because appropriate analyses are closely related to analyses mentioned in section 2.6. In addition, a special issue of *Statistics in Medicine* has been devoted to papers presented at a workshop on analysing longitudinal studies, with an introduction by Laird (1988).

In the design stage of a study it is often wise to presume that in the analysis phase of the project there will be missing data. Therefore the first, trivial, recommendation is: take steps to avoid missing data. In any follow-up study, it is worthwhile to consider possibilities to decrease the rate of loss-to-follow-up. Alternatively, extra information at baseline about nonresponders can be gathered, or it may be feasible to perform a double sampling strategy and obtain extra information about the behaviour of the response variate in nonresponders. In designing a randomized clinical trial it is customary to increase sample size because of the fact that, before the endpoint is reached, a number of losses-to-follow-up will have occured. In such situations sample size adjustment is usually made for specific expected proportions of losses-to-follow-up. In surivival analysis using the logrank test, a sample size correction formula has been advocated (Lachin and Foulkes, 1986), where loss-to-follow-up is assumed to be MCAR within each group and the censoring distribution to be independent of the survival distribution.

If, *a priori*, it is expected that some responders will not tell the truth about having a specific sensitive or hazardous feature, e.g. a drinking problem, another sampling procedure can be considered. For example, a randomized response technique, in which the interviewer does not know to which of two questions the responder reacts, can be used (Warner, 1965). A more refined, computerized alternative to the same problem is presented by Kuk (1990).

Methods to deal with missing data differ according to actual application. Generally speaking, the statistical analysis should take into account that the occurrence of missing data may be related to specific variables. If this is indeed true, and if these specific variables are measured for all individuals, bias due to incomplete response can be adjusted for, by using a likelihood-based approach. This will not be simple, but it is feasible and worthwhile persuing. If a complete-cases-only analysis is performed, conclusions are only valid for this complete case subcohort. Conclusions may be valid for the total population provided missing data are MCAR. It seems only logical to perform additional analyses. In general, if possible, all data observed are to be used in a likelihood-based analysis, under the less stringent assumption that the missing data are MAR. If conclusions are different, it is clear that the complete-cases-only analysis alone is insufficient. If additional information is available (for example through resampling) about the specific nonrandom character of the missing-data-generating process, this extra information can be incorporated into the final analysis.

In many studies, extra information on missing-data-generating processes is not available. Sensitivity analysis may then help to show how relevant conclusions may vary under different assumptions on the underlying missing-data-generating process. This gives some insight into the robustness of these conclusions.

It is likely that in the future, robust statistical analyses will be shown to be insensitive to the occurrence of (relatively) small percentages of missing data. For example, with the "least median of squares" regression method of Rousseeuw (1984), applied to a sample that includes one individual where data are missing completely, imputing extreme values will not change conclusions so much as with an ordinary least squares regression method. This is due to the high breakdown point of the former analysis (50 per cent) compared to the low breakdown point of the latter (0 per cent), the breakdown point being defined as that percentage contamination where estimates can take on uncontrollably large aberrant values.

Although quite some research remains to be done on this subject, it is clearly time to implement procedures that can deal with missing data, both in the design phase and in the analysis phase of a project.

2.9 REFERENCES

Aebischer NJ (1986). Estimating the proportion of uncatchable animals in a population by double-sampling. *Biometrics* 42, 973-979.

Alho JM (1990). Adjusting for nonresponse bias using logistic regression. *Biometrika* 77, 617-624.

Azen SP, Van Guilder M and Hill MA (1989). Estimation of parameters and missing values under a regression model with non-normally distributed and non-randomly incomplete data. *Statistics in Medicine*, 8, 217-228.

Berk K (1987). Computing for incomplete repeated measures. *Biometrics* 43, 385-398.

Box GEP and Jenkins GM (1976). *Time series analysis, forecasting and control*, 2nd edition. San Francisco: Holden-Day.

Brown CH (1990). Protecting against nonrandomly missing data in longitudinal studies. *Biometrics* 46, 143-155.

Chastang C, Byar D and Piantadosi S (1988). A quantitative study of the bias in estimating the treatment effect caused by omitting a balanced covariate in survival model. *Statistics in Medicine*, 7, 1243-1255.

Chiu HY and Sedransk J (1986). A Bayesian procedure for imputing missing values in sample surveys. *J. Am. Statist. Assoc.*, 81, 667-676.

Conn JM, Lui K and McGee DL (1989). A model-based approach to the imputation of missing data: home injury incidences. *Statistics in Medicine*, 8, 263-266.

Davis CS and Wei LJ (1988). Nonparametric methods for analyzing incomplete nondecreasing repeated measurements. *Biometrics* 44, 1005-1018.

Dempster AP, Laird NM and Rubin DB (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B* 39, 1-38.

Diggle PJ (1989). Testing for random dropouts in repeated measurement data. *Biometrics* 45, 1255-1258.

Dinse GE (1986). Nonparametric prevalence and mortality estimators for animal experiments with incomplete cause-of-death data. *J. Am. Statist. Assoc.* 81, 328-336.

Goffinet B (1987). Alternative conditions for ignoring the process that causes missing data. *Biometrika* 74, 437-439.

Heitjan DF and Rubin DB (1990). Inference from coarse data via multiple imputation with application to age heaping. *J. Am. Statist. Assoc.* 85, 304-314.

Jennrich RI and Schluchter MD (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics* 42, 805-820.

Kenward MG (1987). A method for comparing profiles of repeated measurements. *Applied Statistics* 36, 296-308.

Kuk AYC (1990). Asking sensitive questions indirectly. *Biometrika* 77, 436-438.

Lachin JM and Foulkes MA (1986). Evaluation of sample size and power for analyses of survival with allowance for nonuniform patient entry, losses to follow-up, noncompliance and stratification. *Biometrics* 42, 507-519.

Lachin JM and Wei LJ (1988). Estimators and tests in the analysis of multiple nonindependent 2x2 tables with partially missing observations. *Biometrics* 44, 513-528.

Laird NM (1988). Missing data in longitudinal studies. *Statistics in Medicine*, 7, 305-315.

Lewis JA (1988). Correcting for the bias caused by dropouts in hypertension trials. *Statistics in Medicine* 7, 1302-1303.

Little RJ (1988). A test of missing completely at random for multivariate data with missing values. *J. Am. Statist. Assoc.* 83, 1198-1202.

Little RJ (1988). Commentary. *Statistics in Medicine*, 7, 347-355.

Little RJ and Rubin DB (1987). *Statistical analysis with missing data.* New York: John Wiley.

Little RJ and Smith PJ (1987). Editing and imputation for quantitative survey data. *J. Am. Statist. Assoc.* 82, 58-68.

Murray GD and Findlay JG (1988). Correcting for the bias caused by drop-outs in hypertension trials. *Statistics in Medicine* 7, 941-946.

Rochon J and Helms RW (1989). Maximum likelihood estimation for incomplete repeated-measures experiments under an ARMA covariance structure. *Biometrics* 45, 207-218.

Rosenbaum PR (1987). Model-based direct adjustment. *J. Am. Statist. Assoc.* 82, 387-394.

Rousseeuw PJ (1984). Least median of squares regression. *J. Am. Statist. Assoc.* 79, 871-880.

Rubin DB (1976). Inference and missing data. *Biometrika* 63, 581-592.

Rubin DB and Schenker N (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *J. Am. Statist. Assoc.* 81, 366-374.

Schemper M and Smith TL (1990). Efficient evaluation of treatment effects in the presence of missing covariate values. *Statistics in Medicine* 9, 777-784.

Schluchter MD and Jackson KL (1989). Log-linear analysis of censored survival data with partially observed covariates. *J. Am. Statist. Assoc.* 84, 42-52.

Simon GA and Simonoff JS (1986). Diagnostic plots for missing data in least squares regression. *J. Am. Statist. Assoc.* 81, 501-509.

Stasny EA (1986). Estimating gross flows using panel data with nonresponse: an example from the Canadian labour force survey. *J. Am. Statist. Assoc.* 81, 42-47.

Steinhorst RK and Samuel MD (1989). Sightability adjustment methods for aerial surveys of wildlife populations. *Biometrics* 45, 415-425.

Stram DO, Wei LJ and Ware JH (1988). Analysis of repeated ordered categorical outcomes with possibly missing observations and time-dependent covariates. *J. Am. Statist. Assoc.* 83, 631-637.

Warner SL (1965). Randomized response: a survey technique for eliminating evasive answer bias. *J. Am. Statist. Assoc.* 60, 63-69.

Wei LJ and Stram DO (1988). Analysing repeated measurements with possibly missing observations by modelling marginal distributions. *Statistics in Medicine* 7, 139-148.

Wu MC and Bailey K (1988). Analysing changes in the presence of informative right censoring caused by death and withdrawal. *Statistics in Medicine*, 7, 337-346.

Wu MC and Carroll RJ (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics* 44, 175-188.

Zeh JE, Ko D, Krogman BD and Sonntag R (1986). A multinomial model for estimating the size of a whale population from incomplete census data. *Biometrics* 42, 1-14.

-

CHAPTER 3

THE EFFECTS OF LOSS-TO-FOLLOW-UP ON ESTIMATES OF RELATIVE RISK *

SUMMARY

Loss-to-follow-up is a serious problem in longitudinal studies. For estimating a certain relative risk, two subpopulations, one of which is exposed to one or more potential risk factors, are followed in the course of time. After a predetermined period, the endpoint status of all persons is assessed. In many situations, the fraction of loss-to-follow-up will differ amongst endpoint categories. The resulting bias in relative risk estimates is considered and evaluated quantitatively, using survival status as the endpoint. If the fraction of deaths with vital status known (= f_d) differs from the similar fraction of alives (= f_a), there is a bias towards 1 if $f_d>f_a$ and away from 1 if $f_a>f_d$. For low risks of both subpopulations, this bias is small, even for fractions of loss-to-follow-up as large as 20 per cent. Two methods of calculating conservative confidence intervals for the relative risk are presented: a direct approach using simple adjustment factors, and another one using the relation between relative risk and odds ratio. The latter yields a less conservative confidence interval. Data from an actual study are used to illustrate the implications of both methods for estimating relative risks. Authors of papers about relative risk estimates in the case of different fractions of complete follow-up, are encouraged to explicitly present information on such fractions.

* After: Lugtenburg D and Van Strik R (submitted) *Am. J. Epidemiol.*

3.1 INTRODUCTION

Two major problems may present themselves in a follow-up study of relative risk. Firstly, there may be misclassification of exposure (at baseline) or of outcome, which will give biased estimates. Secondly, incomplete follow-up of individuals may present similar problems. One simple reason for incomplete follow-up in mortality studies can be that persons cannot be traced, for instance, due to emigration.

The literature on misclassification is immense, see for example Flegal, Brownie and Haas (1986); Dosemeci, Wacholder and Lubin (1990) and Espeland and Hui (1987). The problem of loss-to-follow-up, however, has not been addressed all that often, although more recently several papers have appeared either on determinants for loss-to-follow-up, for example Vernon et al. (1990); Sonne-Holm et al. (1989) and Jooste et al. (1990), or on adjustment procedures for loss-to-follow-up as presented by Conn, Lui and McGee (1989) and Steinhorst and Samuel (1989). By way of adjustment procedure, a model is generally postulated for the process that generates the incomplete follow-up. Sometimes, data can be collected to support such a model. This model then is used for adjustment of simple statistics and it has led to some interesting results for estimating prevalence (Alho, 1990).

To our knowledge, no adjustment procedure has been published, however, for the relative risk estimate. Only a few papers discuss the fact that biases might occur (Criqui, 1979; Greenland and Criqui, 1981 and Johnson, 1990). This chapter describes effects of incomplete follow-up, if fractions of loss-to-follow-up, the complement of response rates, differ amongst the possible endpoint realizations dead and alive. This results in a correction formula for non-response. If lower and upper bounds for follow-up rates can be derived, maybe from other sources, this formula can be used to adjust the confidence interval for the relative risk. An example from the literature is used to illustrate this adjustment procedure.

3.2 METHODS

3.2.1 The bias in estimating relative risk

Let N be the total number of subjects in the study at baseline. The exposed group has size N_e and the unexposed group has size N_o, so $N=N_e+N_o$. The probability or risk that a person of the exposed group will die during the period of follow-up is denoted by p_e. The risk that a person of the unexposed group will die during follow-up is p_o. The relative risk, or risk ratio, is denoted by $RR=p_e/p_o$. Now suppose that incomplete follow-up data are available. The fraction of deaths of which vital status is known is denoted by f_d. The fraction of alives of which vital status is known is denoted by f_a. Fraction f_a is assumed to be equal in the two subpopulations of exposed and unexposed individuals. Fraction f_d, too, is assumed to be equal in both subpopulations. This causes the estimate of the relative risk to become biased when this estimate is the normal estimate, rr^*, which is (number of certified deaths in the exposed)/(number of exposed where follow-up is complete) divided by (number of certified deaths in the unexposed)/(number of unexposed where follow-up is complete). This estimate is an estimate of a biased relative risk, RR^*. If loss-to-follow-up actually occurs, then the biased relative risk, RR^*, as shown in Appendix 3-A, is the product of the "true" relative risk, RR, and a bias factor, K:

$$RR^* = K \, . \, RR$$

with

$$K=\frac{f_a(1 - p_o)+p_o f_d}{f_a(1 - p_e)+p_e f_d} \tag{1}$$

In the (rare) case where response rates are independent of vital status, i.e. $f_a=f_d$, $K=1$ and the estimate of the relative risk is unbiased. If $p_o=p_e$ ($RR=1$), too, no bias occurs. However, in all other situations where $f_a \neq f_d$ a certain bias will result. In the extreme situation where all subjects that have died at the end of follow-up are known, but all subjects that are still alive at the end of follow-up are unknown ($f_d=1$ and $f_a=0$), both risks are estimated to be 1 and therefore the estimated relative risk equals 1. Consequently, if $f_d > f_a$, the biased relative risk, RR^*, is biased towards unity. To investigate this in more detail, the percentage bias as a function of f_a, f_d, p_e and p_o will now be studied.

The percentage bias is defined as 100 (K-1). Thus, a percentage bias of 10 means that RR^* overestimates the "true" relative risk by 10 per cent. From the foregoing it follows that the percentage bias can be calculated to be:

$$100(K-1)=100\frac{(f_a-f_d)(p_e-p_o)}{f_a \ (1-p_e)+p_e f_d} \tag{2}$$

Since follow-up percentages are positive and $0 \leq p_e \leq 1$, the denominator must also be positive. For $RR>1$, p_e-p_o is positive. In that situation, when f_a-f_d is positive, the percentage bias is positive, too. So, if $f_a > f_d$, the relative risk is overestimated. If $f_a < f_d$, the relative risk is underestimated.

The conclusion above may be derived in another way. If $f_d > f_a$, the estimates of the risks p_e and p_o are biased towards one. Hence, the relative risk will be biased towards unity, too, and is therefore underestimated.

3.2.2 A direct approach to the calculation of the confidence interval of the relative risk

In an actual study it is often possible to estimate from observations the values of f_d, f_a, p_e and p_o, at least within reasonable boundaries. For example, if it is known that out of 1000 persons, 200 persons have died and that vital status is unknown for 100 persons, at most 300 persons have died and at least 200 persons. Therefore f_d lies between 200/300=0.66 and 200/200=1.

Equation (1) is used to obtain adjustment factors for RR^* based on the inequalities $f_{a,min} \leq f_a \leq f_{a,max}$, $f_{d,min} \leq f_d \leq f_{d,max}$, $p_{e,min} \leq p_e \leq p_{e,max}$ and $p_{o,min} \leq p_o \leq p_{o,max}$. For an upper bound to RR^* we then obtain $RR\ (f_{a,max}(1\text{-}p_{o,min}) + p_{o,max}\, f_{d,max})/(f_{a,min}(1\text{-}p_{e,max})+p_{e,min}f_{d,min})$. A lower bound for RR^* is $RR\ (f_{a,min}(1\text{-}p_{o,max}) + p_{o,min}\ f_{d,min})/(f_{a,max}(1\text{-}p_{e,min})+p_{e,max}f_{d,max})$. Consequently, the boundaries for RR are: $1/K_{max}\ RR^* \leq RR \leq 1/K_{min}\ RR^*$ with:

$$\frac{1}{K_{\max}} = \frac{f_{a,\min}(\ 1-p_{e,\max}\ )+p_{e,\min}\ f_{d,\min}}{f_{a,\max}(\ 1-p_{o,\min}\ )+p_{o,\max}\ f_{d,\max}}$$

and

$$\frac{1}{K_{\min}} = \frac{f_{a,\max}(\ 1-p_{e,\min}\ )+p_{e,\max}\ f_{d,\max}}{f_{a,\min}(\ 1-p_{o,\max}\ )+p_{o,\min}\ f_{d,\min}} \ .$$

The boundary values for $1/K_{max}$ and $1/K_{min}$ can be used to adjust confidence intervals for RR^* to obtain conservative confidence limits for RR. Suppose a confidence interval for RR^*, calculated using the log-transform approach (Rothman, 1986), is from cr_1 to cr_2, then the limits of a conservative confidence interval for RR, cr_{min} and cr_{max}, are cr_1/K_{max} and cr_2/K_{min}. A formal proof of this statement is presented in Appendix 3-B.

3.2.3 An indirect approach to the calculation of the confidence interval of the relative risk

The indirect approach uses the relation between odds ratio and risk ratio. The odds ratio, OR, is defined as $OR=(p_e/(1\text{-}p_e))\ /\ (p_o/(1\text{-}p_o))$. The relative risk is easily shown to be directly related to the odds ratio and p_e according to:

$$RR = (\ 1-p_e\ )\ OR + p_e \tag{3}$$

As the odds ratio estimate is unbiased with respect to all realizations of f_a and f_d, see Appendix 3-A, conservative lower and upper 95 per cent confidence boundaries for *RR* are:

$$cr_{\min}=(\ 1-p_{e,\max}\ )\ or_{\min}\ +\ p_{e,\min}$$

$$cr_{\max}=(\ 1-p_{e,\min}\ )\ or_{\max}\ +\ p_{e,\max}$$

where or_{min} and or_{max} denote the 95% confidence limits of *OR*, e.g. using a log-transform approach (Rothman, 1986).

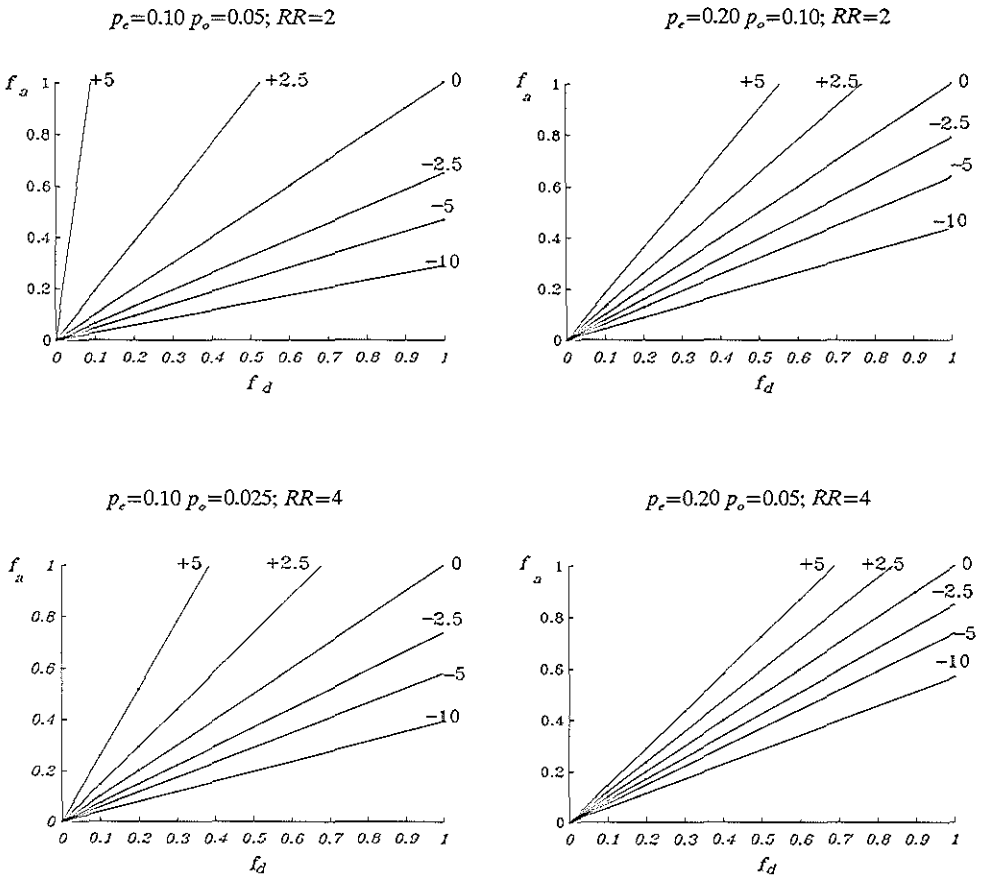
From the above equations it is easy to understand that, in the situation where p_e is small, i.e. if the odds ratio *OR* nearly equals the relative risk *RR*, the use of the above formulae does not change the confidence interval. This is not equally evident for the direct approach discussed in the previous section. An estimate of $p_{e,max}$ is obtained if it is supposed that all persons lost-to-follow-up came from the exposed subpopulation and died during the follow-up period. To estimate $p_{e,min}$ it is supposed that all persons lost-to-follow-up came from the exposed subpopulation and were still alive at the end of the follow-up period.

3.3 RESULTS

The practical implications of the methodology discussed in the previous sections will now be summarized. To begin with, the amount of bias will be considered as a function of the fractions f_a and f_d in the case of incomplete follow-up. Next, a real-life example is used to illustrate how confidence intervals for RR^* can be adjusted to represent a conservative confidence interval for *RR*. In addition, the advantage of using the indirect approach will be illustrated.

3.3.1 Percentage bias as a function of f_a and f_d in the case of incomplete follow-up

Unit contours of the percentage bias according to equation (2) are shown in Figure 3-1 for specific realizations of p_e and p_o and for various combinations of f_a and f_d. A unit contour is defined as a contour connecting possible combinations of f_a and f_d with identical percentages bias. Four different realizations of p_e and p_o are used to show the effect on the percentage bias.



p_e=0.10 p_o=0.05; RR=2

p_e=0.20 p_o=0.10; RR=2

p_e=0.10 p_o=0.025; RR=4

p_e=0.20 p_o=0.05; RR=4

Figure 3-1 Unit contours of percentage bias for four combinations of p_e and p_o as a function of f_a and f_d. A percentage bias of +5 means that the biased relative risk, RR^*, overestimates the true relative risk RR by 5 per cent.

It follows from the definition of K that, with K, p_e and p_o given, f_a is proportional to f_d, so that the contours are straight lines through the origin. For RR=2 and RR=4, it is seen that the relative risk is overestimated as long as f_a exceeds f_d. If f_a equals f_d, the relative risk estimator is unbiased. If f_a is less than f_d, the relative risk is underestimated.

It is also apparent from these graphs that the amount of bias is low in the situations where the risks of death are low in both subpopulations. In the four situations illustrated in Figure 3-1, the absolute percentage bias is less than 2.5 if the fractions of people followed-up exceed 0.90 both for alives and deaths.

3.3.2 A practical example of adjusting for incomplete follow-up

By way of example, use is made of some data from a recent follow-up study in the Netherlands (Doornbos and Kromhout, 1990). A cohort of men, all of them born in 1932, was examined for enlistment in military service in 1950/1951. After 32 years, 78.505 persons had complete follow-up; 3456 persons had died and the remaining 75049 persons were still alive. 150 persons were excluded from the study because their death certificates could not be traced. In addition, 1152 persons were excluded because data were not available on one of the variables under investigation. The original cohort size had therefore been 79807, all male persons who had passed medical examination. The fraction of deaths of which vital status is known, f_d, is between 3456/(3456+150+1152) and 3456/(3456+150), i.e. $0.726 \leq f_d \leq 0.958$, depending on how many of the 1152 persons excluded had died during the 32-year follow-up. The fraction of alives of which vital status is known after 32 years is between 75049/(75049+1152) and 75049/75049, i.e. $0.985 \leq f_a \leq 1$. Obviously, $f_d < f_a$. Therefore, a first conclusion is that unadjusted relative risks will be biased away from one.

As an example of using the direct approach, let us now calculate an adjusted 95 per cent confidence interval for the relative risk of dying within 32 years, having a low education (with a risk p_e of dying) against having a high education (with a risk p_o of dying). Although no absolute frequencies were given, Table 3-I was derived from the percentages as tabulated. From the data in the table rr^* is calculated to be 0.05/0.034=1.47 with a 95 per cent confidence interval for RR^* (see Appendix 3-B) of 1.314 to 1.646. To adjust this confidence interval all available data will be used, including estimates of boundaries for p_e and for p_o. Concerning the upper boundary for p_e it seems logical that $p_e \leq (1892+150+1152)/(37836+150+1152) = 0.082$. Similarly, $p_o \leq (345+150+1152)/(10145+150+1152)= 0.144$. When this reasoning is also applied to lower boundaries, we have $0.0485 \leq p_e \leq 0.082$ and $0.0305 \leq p_o \leq 0.144$. Imputing these boundaries gives $0.85\ RR^* \leq RR \leq 1.19\ RR^*$, leading to an "adjusted" conservative confidence interval for RR given by $1.12 \leq RR \leq 1.96$.

Using the odds-ratio approach, the 95 per cent confidence limits for the odds ratio are found to be: $1.33 \leq OR \leq 1.68$, which can be used for calculating the confidence limits for RR^* by imputing p_e=0.05 in formula (3). Using the boundaries on p_e leads to an adjusted confidence interval $1.27 \leq RR \leq 1.68$, which is somewhat narrower than the interval obtained with the direct approach, which may be due to the fact that only boundaries on p_e are implemented.

TABLE 3-I

Outcomes from a 32-year follow-up of two subpopulations, defined by educational level at baseline

	Vital status		
	Dead	Alive	Unknown
Low education	1892	35944	0
High education	345	9800	0
Unknown education	150	0	1152

3.4 DISCUSSION

This chapter presents a framework to assess the (percentage) bias for estimates of the relative risk, assuming that f_a and f_d, the fractions of alives and deaths, respectively, with known vital status, can be estimated within narrow limits and can be assumed to be nondifferential with respect to exposure. The relative risk of dying is biased towards 1 if $f_d > f_a$ (and away from 1 if $f_a > f_d$). This would indicate that it is feasible to optimize follow-up procedures as long as it is fairly certain that the fraction of deaths with vital status known is at least as high as the fraction of alives with vital status known. However, most follow-up studies will probably show opposite signs, i.e. $f_a > f_d$, as in the example discussed before. So, in general, published relative risks calculated from longitudinal studies, will be biased away from 1 when follow-up is incomplete. In order to ascertain this, extra information is needed, of course. This information might be collected by resampling in the subgroup of incomplete follow-up. Complete, although more expensive, follow-up for a representative subgroup is sometimes attainable. In the example discussed, necessary additional data had been presented in the paper, but very often these are not available. It is clearly important that papers should give data on fractions of follow-up, so that, if relevant, an adjusted analysis can be performed.

Another conclusion from this chapter is that, in many practical situations, the resulting bias due to incomplete follow-up can be expected to be relatively small. With a follow-up rate of at least 90 per cent, the percentage bias will usually not exceed 10.

If $f_{a,min} > f_{d,max}$ (or $f_{a,max} < f_{d,min}$), statistical hypotheses testing is not affected and p-values calculated for RR^* are still appropriate for RR, as may be shown as follows. If $f_{a,min}>f_{d,max}$ then $f_a>f_d$. Therefore, it is known from equation (1) that either $1<RR<RR^*$ or

$1=RR=RR^*$ or $1>RR>RR^*$. Suppose that a test of the hypothesis $RR^*=1$ at $\alpha=0.05$ leads to rejection and results in the statement $RR^* > 1$. The conclusion would consequently be that also $RR>1$ (at $\alpha=0.05$). This is confirmed in the example discussed before, where both approaches of calculating confidence regions led to the conclusion that $RR>1$.

Given that the influence of incomplete follow-up will usually be relatively small, the direct approach of adjusting a relative risk seems rather pessimistic as it leads to rather large increases of the confidence interval. The increases are less if the indirect approach of adjusting an odds ratio to represent a relative risk is used. The odds ratio approach is conjectured to give less conservative estimates and confidence limits and therefore might be preferable in practical situations.

In the introduction it was stated that two problems occur in longitudinal studies, misclassification and incomplete follow-up. The simultaneous occurrence might result in a smaller bias of the relative risk estimate, since misclassification generally leads to a bias towards 1, whereas incomplete follow-up generally leads to a bias away from 1. More research needs to be done to establish to what degree this occurs in actual practice.

3.5 REFERENCES

Alho JM (1990). Adjusting for nonresponse bias using logistic regression. *Biometrika* 77: 617-624.

Conn JM, Lui K and McGee DL (1989). A model-based approach to the imputation of missing data: home injury incidences. *Statistics in Medicine* 8: 263-266.

Criqui MH (1979). Response bias and risk ratios in epidemiologic studies. *Am. J. Epidemiol.* 109: 394-399.

Doornbos G and Kromhout D (1990). Educational level and mortality in a 32-year follow-up study of 18-year old men in the Netherlands. *Int. J. Epidemiol.* 19: 374-379.

Dosemeci M, Wacholder S and Lubin JH (1990). Does nondifferential misclassification of exposure always bias a true effect toward the null value? *Am. J. Epidemiol.* 132: 746-748.

Espeland MA and Hui SL (1987). A general approach to analyzing epidemiologic data that contain misclassification errors. *Biometrics* 43: 1001-1012.

Flegal KM, Brownie C and Haas JD (1986). The effects of exposure misclassification on estimates of relative risk. *Am. J. Epidemiol.* 123: 736-751.

Greenland S and Criqui MH (1981). Are case-control studies more vulnerable to response bias? *Am. J. Epidemiol.* 114: 175-177.

Johnson ES (1990). Bias on withdrawing lost subjects from the analysis at the time of loss, in cohort mortality studies, and in follow-up methods. *J. Occup. Med.* 32: 250-254.

Jooste PL, Yach D, Steenkamp HJ, Botha JL and Rossouw JE (1990). Drop-out and newcomer bias in a community cardiovascular follow-up study. *Int. J. Epidemiol.* 19: 284-289.

Rothman KJ (1986). *Modern Epidemiology.* Little, Brown and Company Boston p. 173.

Sonne-Holm S, Sorenson TIA, Jensen G and Schnohr P (1989). Influence of fatness, intelligence, education and sociodemographic factors on response rate in a health survey. *J. Epidem. Comm. Health* 43: 369-374.

Steinhorst RK and Samuel MD (1989). Sightability adjustment methods for aerial surveys of wildlife populations. *Biometrics* 45: 415-425.

Vernon SW, Acquavella JF, Yarborough CM, Hughes JI and Thar WE (1990). Reasons for participation and nonparticipation in a colorectal cancer screening program for a cohort of high risk polypropylene workers. *J. Occup. Med.* 32: 46-51.

APPENDIX 3-A

Derivation of equation (1)

On the basis of the notation introduced in the methodology section, it is intuitively clear that the expected value of (number of deaths in the exposed group)/(number of exposed where follow-up is complete) is $p_{e,observed} = \mathrm{p}_e f_d \,/\, ((1\text{-}p_e)f_a + p_e f_d)$ and that the expected value of (number of deaths in the unexposed group)/(number of unexposed where follow-up is complete) is $p_{o,observed} = p_o f_d \,/\, ((1\text{-}p_o)f_a + p_o f_d)$. Therefore, rr^* estimates the population entity: $RR^* = p_{e,observed}/p_{o,observed} = RR\ .\ K$.

Independence of odds ratio of f_a or f_d

The simple estimate of the odds of the exposed group is an estimate of $p_{e,observed}/(1\text{-}p_{e,observed})$:

$$odds(\ exposed\ ) = \frac{p_e\, f_d}{(1-p_e)\, f_a}\ .$$

Similarly, the simple estimate of the odds of the unexposed group is an estimate of $p_{o,observed}/(1-p_{o,observed})$:

$$odds(\ unexposed\ )=\frac{p_o\, f_d}{(1-p_o)\, f_a}\ .$$

Hence, the simple estimate of the odds ratio of both observed groups estimates the quantity: $OR^* = odds(exposed)/odds(unexposed) = OR.$

APPENDIX 3-B

Interval cr_1/K_{max} to cr_2/K_{min} provides a conservative confidence interval of *RR*

Suppose A_1 deaths are observed in the exposed subpopulation of which N_1 persons had complete follow-up, and suppose A_2 deaths observed in the unexposed subpopulation of which N_2 persons had complete follow-up. For RR^* the 95 per cent confidence interval is obtained following Rothman (1986) as:

$$\exp\left[\ \ln\, rr^* \pm 1.96\sqrt{Var(\ln\, rr^*)}\ \right] =$$

$$\exp\left[\ \ln\frac{A_1\ /N_1}{A_2\ /N_2} \pm 1.96\sqrt{\frac{N_1-A_1}{A_1N_1}+\frac{N_2-A_2}{A_2N_2}}\ \right]\ .$$

From $RR= RR^*/K$ it follows that an estimate, *rr*, of *RR* is rr^*/K.
$\mathrm{Var}(\ln(rr))=\mathrm{Var}(\ln(rr^*/K\ ))=\mathrm{Var}(\ \ln rr^*\ -\ln(K)\ )=\mathrm{Var}(\ln rr^*)$ since K is constant.

Hence, conservative 95 per cent confidence limits for *RR* are

$$\exp\left[\ln\, rr^* - \ln\, K_{\max} - 1.96\sqrt{Var(\ \ln\, rr^*)}\ \right]$$

and

$$\exp\left[\ln\, rr^* - \ln\, K_{\min} + 1.96\sqrt{Var(\ \ln\, rr^*)}\ \right]\ .$$

A conservative 95 per cent confidence interval is thus given by cr_1/K_{max} to cr_2/K_{min} .

CHAPTER 4

MODELLING MANIPULATED DISCRETE PROCESSES: ESTIMATING PREVALENCE IN THE CASE OF NON-RESPONSE *

SUMMARY

Missing data may cause a problem in the estimation of the prevalence of a certain disease in a population by means of a survey. This chapter presents a model-based approach for dealing with missing data. The model involves several strategies, each resulting in a different estimate of this prevalence. The strategies will be compared by means of likelihood ratio testing procedures within the parametric setting. Of course, properly testing goodness of fit is not possible solely through a likelihood ratio test as long as missing data are not replaced with resampled information. The parametric model is shown to be applicable also for situations in which there is observational error.

* After: Lugtenburg D and Mulder PGH (submitted) *Applied Statistics*

4.1 INTRODUCTION

In some situations observations collected are not correct or complete. One reason may be that, before outcomes of the relevant process can be observed, a second process has eliminated part of these outcomes or changed part of them. This second process "manipulates" the outcomes of the first process. Probably, the manipulating processes encountered in practice most often are "vanishing" and "recategorizing". The latter is a process where only part of the outcomes of the first process are observed correctly. If the former process is active, part of the outcomes of the first process become missing.

The nature of "vanishing" processes as they may occur in practice varies widely. For instance, in surveys where questionnaires are sent to persons of specific populations, it is only rarely that all questionnaires are returned or are completely useful for analysis. There is a certain percentage of non-response. The mechanisms that underly the existence of such missing data may affect the appropriate analysis, see e.g. Rubin (1976). An excellent introduction to this subject has been written by Little and Rubin (1987).

Three general approaches are suitable to deal with the "vanishing" problem. First, extra information about the "missing group" can be gathered by resampling. Second, the missing data can be "filled-in" or "imputed" and corresponding analyses performed. Finally, a specific structure of missingness can be assumed and conclusions can be derived from that assumption. For estimating prevalence all three general approaches have been proposed in the literature. In estimating the size of the western Arctic stock of bow head whales, Zeh et al. (1986) use the assumption that missingness is only related to visibility and obtain estimates by resampling. If several sampling procedures are available (the one that is the most expensive being the most precise) a correction formula can be derived. An example is the procedure of counting radio-labelled animals, which is not interfered with by visibility problems (Steinhorst and Samuel, 1989). The second and third of these general approaches are sometimes used simultaneously, as in the example of this chapter. By supposing that all missing data show the outcome of interest (e.g. have the disease or have high blood pressure) an upper limit of the prevalence is obtained. The model postulated in this chapter makes it possible to evaluate such imputation procedures. Dinse (1986) discusses similar but nonparametric estimators.

The following example will be discussed. A chemical industry needed an estimate of the percentage employees with hypertension. In a survey all employees were invited to have a medical examination. Some employees, however, decided not to attend. These refusers presumably did not form a random subgroup of the total population. A non-random process may have caused some employees to attend and others not to attend the examination. Assumptions concerning the non-randomness of this underlying missing-data-generating process are needed in order to properly estimate the prevalence of hypertension. Several simple sets of assumptions will be discussed, as well as likelihood ratio tests for discriminating between these sets.

Besides the problem of missing data, there is also the problem of misclassification. An observer may make mistakes in allocating an observed outcome to the correct category. Many authors have written about this phenomenon. The first and the last of the three general approaches referred to earlier are proposed in the literature to deal with it. Obviously, resampling by means of repeated measurements can improve the estimates considerably, as discussed by Clayton (1985). Also, models have been proposed which *a priori* allow for misclassification. Copas (1988), for example, discusses binary regression models.

In this chapter, a framework is presented that may facilitate the modelling of both the recategorizing and the vanishing process. As has been remarked before, we will assume that the outcomes of the first process are manipulated by a second process. Therefore, only certain combinations of outcomes of both processes can be observed. First, the underlying principles will be discussed, illustrated step by step by showing the implications of modelling a vanishing process. This generally applicable model is presented in section 4.2 for binomial successive processes. The distribution of the actually observed outcome variate will be demonstrated to be multinomial. The model will be specified and the procedure to obtain maximum likelihood estimators (MLE) for the model parameters will be described. For the algorithm we use the composite link approach as introduced by Thompson and Baker (1981). The several different imputation procedures to estimate prevalence will be discussed within the general framework. In section 4.3, the above-mentioned example (prevalence of hypertension) is dealt with in more detail. The analysis is elaborated as far as necessary to serve illustrative purposes. Section 4.4 describes the recategorizing process within the general framework, and a discussion is presented in section 4.5.

4.2 THE MODEL

4.2.1 Introduction of the model

Consider two successive binomial processes with the second process manipulating the outcomes of the first process. The first process has two possible outcomes, z_1 and z_2. The probability that the first process actually has outcome z_1 is denoted q_1. The second process has two possible outcomes, z_{21} and z_{22}, where the probability that outcome z_{21} actually occurs may depend on the outcome of the first process. As a consequence, another two probabilities are introduced: q_2 is the probability that outcome z_{21} occurs when z_1 is the outcome of the first process, q_3 being the corresponding probability when z_2 is the outcome of the first process. The resulting process is illustrated in Figure 4-1.

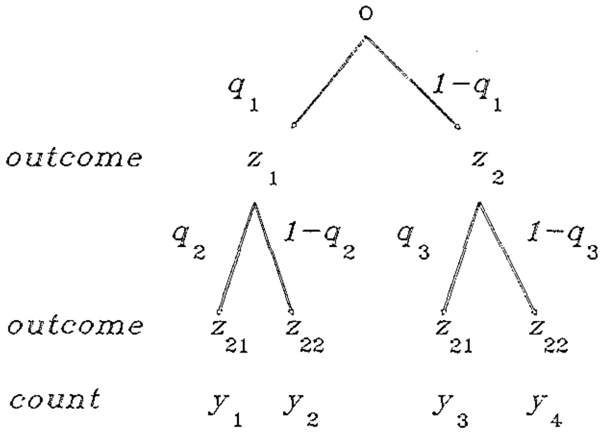


Figure 4-1 Schematic description of two successive processes.

There are four response categories r: (z_1, z_{21}), (z_1, z_{22}), (z_2, z_{21}) and (z_2, z_{22}) with r being 1 to 4, respectively. The response category r $(r=1,..,4)$ is observed y_r times. The four response categories are complete (there are no other responses possible) and mutually exclusive. For each response category r the corresponding category probability, p_r, is assumed constant with $p_1 = q_1.q_2$, $p_2 = q_1.(1-q_2)$, $p_3 = (1-q_1).q_3$ and $p_4 = (1-q_1).(1-q_3)$. Consequently, the response variate, which describes the frequency with which this category will actually be observed, has a multinomial distribution, see e.g. Johnson and Kotz (1969).

If the second process is a missing-data-generating process in which z_{21} means "not missing" and z_{22} means "missing", it is easy to see that in fact one cannot observe all category counts separately. The counts that can be observed are (y_1), (y_3) and the sum of y_2 and y_4: (y_2+y_4). In matrix notation: $y_{observed} = C\,y$, with $y=(y_1, y_2, y_3, y_4)^T$ and

$$C = \begin{bmatrix} 1000 \\ 0010 \\ 0101 \end{bmatrix} .$$

In short, only combined category counts are observed. Somehow, one has to deal with this limited information when estimating the dependence of the probabilities q_j on a vector of explanatory variables. This will be discussed in subsection 4.2.4. In the next subsection, 4.2.2, the dependence model for the probabilities q_j is specified.

4.2.2 Postulation of the model

For design point i (i=1,..., s) the process probabilities q_j (j=1,2,3) are assumed to depend on p explanatory variables assembled in a vector x_i:

$$q_{ij}=q_j(x_i) \qquad j=1,2,3; \;\; i=1,...,s.$$

For every q_j, a logistic model is postulated as follows:

$$logit(q_j)=\beta_j^T x_i \qquad j=1,2,3; \;\; i=1,..., s$$

with β_j a vector of p coefficients specific for each probability q_j. The first element of the vector x_i equals 1 in order to provide intercepts β_{j1}. Some of the other coefficients in β_j may be zero, meaning that the corresponding explanatory variable in x_i is not incorporated in the j^{th} process.

4.2.3 Several strategies for estimating prevalence

Suppose that persons having the disease of interest, e.g. hypertension, fall in category z_1. Again, the second process is a missing-data-generating process where z_{21} means "not-missing" and z_{22} means "missing". Then an estimator for the prevalence (percentage observations falling in category z_1) is defined as follows. When the estimated category counts for the category counts y_{ir} are denoted $\hat{y}_{ir}$ then the prevalence estimator is:

$$prevalence=\frac{100\sum_{i=1}^{s}(\hat{y}_{i1}+\hat{y}_{i2})}{\sum_{i=1}^{s}n_i}.$$

The conditions that define independence of both processes are straightforward. Both processes are independent if and only if $\Pr(z_{21}|z_1)=\Pr(z_{21}|z_2)$ for all design points i, which is tantamount to $q_{i2}=q_{i3}$, or $\beta_2^T x_i=\beta_3^T x_i$, for all i (this means $\beta_2=\beta_3$). This independence condition results in a "simple-guess-estimator" for the prevalence.

A generalization is the following:

$$logit(q_{i2})-logit(q_{i3})=\psi$$

for all i. If $\psi=0$ this simplifies to the above case $q_{i2}=q_{i3}$. If $\psi \to \infty$, then $q_{i2} \to 1$, meaning in the example introduced above that all missings are counted as normotensives. This leads to a "lower-boundary estimator" of the prevalence.

If $\psi \to -\infty$, then $q_{i3} \to 1$, meaning that all missings are counted as hypertensives. This results in an "upper-boundary estimator" of the prevalence.

It will be obvious that these last two situations are extreme situations and unlikely to be real. However, "how" unlikely are they? Interval estimation for the prevalence can be evaluated by means of a likelihood ratio testing procedure as follows. First, consider the following three situations: $\psi = -\infty$ (upper bound), $\psi = 0$ (simple guess) and $\psi = \infty$ (lower bound). For every situation the corresponding deviance is calculated. Secondly, the deviance corresponding to the MLE of ψ is calculated. This leads to an "optimal-guess estimator" for the prevalence. For each fixed ψ the likelihood ratio test statistic is calculated as the difference: deviance(ψ_{fixed}) minus deviance(ψ_{optimal}). When this statistic is supposed to have a chi-squared distribution with 1 degree of freedom (DF for short) on the assumption that ψ_{fixed} is the true ψ, then all values of ψ with a deviance less than 3.84 (chi-squared value, 1 DF and $\alpha=0.05$) away from the deviance corresponding to ψ_{optimal}, correspond to values that are in the 95 per cent confidence interval of the prevalence, because of the monotonic relationship between ψ and the prevalence.

4.2.4 Fitting the model

If the second process is a missing-data-generating process and $q_{i2}=q_{i3}$ or equivalently $\beta_2=\beta_3$, MLEs for the model parameters β_1 and β_2 can be calculated by applying standard binomial algorithms. Two separate binomial models can be fitted. The first for calculating MLEs for the model parameters of the first process: y_{i1} positive responders out of a total of $y_{i1}+y_{i3}$ observations at design point *i*. For the second process this is: $y_{i1}+y_{i3}$ as positive responders out of a total of n_i observations at design point *i*. By adding the corresponding binomial log-likelihoods it can be shown that the total log-likelihood is the appropriate multinomial log-likelihood. The more extreme models with either $q_{i2}=1$ or $q_{i3}=1$, too, can be fitted with standard binomial algorithms.

In all other situations, no standard binomial algorithms can be used. As a general method to maximize the multinomial log-likelihood, a Poisson reparametrization will be introduced.

A sufficient condition for the outcomes y_{ir} $(r=1,..,4)$ at design point i to behave like a multinomial distribution for given p_{ir} and $n_i=y_{i1}+y_{i2}+y_{i3}+y_{i4}$ is a Poisson distribution for the y_{ir}, stratified on design point i. Birch (1963) showed that both likelihoods are proportional to each other and so lead to identical MLEs in a log-linear model specification. Palmgren (1981) showed that the inverses of the Fisher information matrices are identical so that the asymptotic covariance matrices of the estimates also coincide.

We now specify the following dependency model for the expected responses E(Y_{ir}) with a model parameter μ_i representing the stratification on design point i, so that there is a one-to-one relationship between the linear logit(q_{ij}) predictors (with coefficient vectors ß$_j$) and the linear log(p_{ir}) predictors:

$$E(Y_{i\,1})= n_i\, p_{i\,1}= \exp(\mu_i+\gamma_1^T x_i+\beta_2^T x_i)$$

$$E(Y_{i\,2})= n_i\, p_{i\,2}= \exp(\mu_i+\gamma_1^T x_i)$$

$$E(Y_{i\,3})= n_i\, p_{i\,3}= \exp(\mu_i+\beta_3^T x_i)$$

$$E(Y_{i\,4})= n_i\, p_{i\,4}= \exp(\mu_i)$$

with:

$$\beta_1^T x_i=\gamma_1^T x_i+\ln(\frac{\exp(\beta_2^T x_i\ )+1}{\exp(\beta_3^T x_i\ )+1})$$

resulting for ψ in:

$$\psi=\beta_2^T x_i-\beta_3^T x_i\ .$$

As ψ has to be constant across i, ψ is the difference of the intercepts of process 2 and process 3: ψ = ß$_{21}$ - ß$_{31}$, the other coefficients in ß$_2$ being equal to those in ß$_3$.

The above reparametrization to a Poisson regression model provides a way of handling the problem in question, namely that the y_{ir} at design point i are not observed directly, but that only certain combinations over r of the y_{ir} are observed: $y_i(obs) = C\, y_i$ with $y_i(obs)$ being an observed column vector of dimension $k<4$ and y_i a column vector of elements y_{ir} $(r=1,...,4)$; the $(k$*$4)$-matrix C being the composite link matrix which is the same for all design points i. A typical C-matrix has zeroes and ones as elements.

Composite Poisson models are described by Thompson and Baker (1981), who introduced the composite link model as a generalization of the generalized linear model (McCullagh and Nelder, 1983). The algorithm described by Thompson and Baker has also been applied to the example of this chapter. For the vanishing process the *C*-matrix already was presented in subsection 4.2.1. The Poisson reparametrization provides a generally applicable algorithm without any further conditions.

4.3 AN APPLICATION: THE PREVALENCE OF HYPERTENSION

A survey was conducted on 6287 employees of Shell Pernis, ranging in age from 20 to 59 years, to find the total percentage of employees with hypertension. This was defined as a diastolic blood pressure of at least 95 mm Hg or a systolic blood pressure of at least 160 mm Hg. The target population was the total working population on January 1, 1982 at the site. The employees were invited to undergo a physical examination that year. Of about 25 per cent of the employees, however, no blood pressure values were recorded. This is partly due to the fact that a number of employees refused to attend, but also because of some computer storage problems. The latter is known because the observed blood pressure values of some employees, who were known to have attended the examination, could not be retrieved at the time of analysis in 1989.

In the general context of this chapter the first process is considered to be the process that determines whether a person has high blood pressure. The second process determines whether the blood pressure values could be retrieved for the analysis in 1989. A very relevant explanatory variable for the first process is age, which was grouped into four classes: 20-29, 30-39, 40-49 and 50-59 years. One explanatory variable for the second process is the number of times a person reported sick during 1982. The underlying reasoning is simple. If an employee was sick, he did not get an invitation to attend the examination. Two classes were used, namely "reporting sick less than four times" and "reporting sick at least four times". No other explanatory variables were used. This is also an obvious determinant for process 1, especially for correctly estimating the prevalence. A summary of frequencies in the resulting 8 age-sickness classes is presented in Table 4-I.

In this example the first binomial process was defined as the process that determines whether a person has high blood pressure with the probability parameter only varying across age groups. The outcomes z_1 and z_2 mean "hypertensive" and "normotensive", respectively. The second binomial process is a missing-data-generating process with z_{21} and z_{22} meaning "not missing" and "missing", respectively.

TABLE 4-I

Observed percentages of definite hypertension in a survey of an industrial population for different age and sickness-absence categories

Age (years)	Reporting sick at least four times	Number of employees	Percentages data that are unknown	Within the data that are known, the percentage persons with definite hypertension
20 - 29	no	1197	24.6	8.4
20 - 29	yes	245	33.1	12.2
30 - 39	no	1318	27.5	9.6
30 - 39	yes	203	36.0	13.8
40 - 49	no	1267	21.7	14.3
40 - 49	yes	171	35.7	20.0
50 - 59	no	1642	24.5	20.7
50 - 59	yes	244	32.0	23.5

As the second process may be considered to randomly delete blood pressure records within a given age-sickness category, the percentage unknowns should be the same for employees that have outcome z_1 and for employees that have outcome z_2, implying $q_{i2}=q_{i3}$ for all classes i. This also implies that ψ, introduced in section 4.2.3, equals zero. This "simple-guess approach" is evaluated first. In this situation, it turns out for the first process that, besides a linear trend of the logit of the response probability on age category, the number of sickness absences (<4 or >3) in 1982, too, has explanatory power. The second process is assumed to be related only to the number of sickness absences. After fitting this model for the total process, a deviance of 16.22 with 11 DF was found, which indeed suggests a reasonable fit. Since independence applies, this deviance can easily be divided into two. For the first process this leads to a deviance of 3.52 with 5 DF and for the second process to a deviance of 12.70 with 6 DF, which latter result is not satisfactory. Various other alternatives for ψ instead of the simple-guess approach ($\psi = 0$) are to be evaluated. The results are presented in Table 4-II.

TABLE 4-II

Summary of several strategies towards the randomness of the underlying missing-data-generating process

Strategy name	Second process	Estimated prevalence of hypertension	Deviance	DF
simple guess	$q_{i2}=q_{i3}$	14.3	16.22	11
lower bound	$q_{i2}=1$	10.6	15.73	11
upper bound	$q_{i3}=1$	36.5	81.44	11
optimal guess	ψ estimated (0.91)	12.2	14.97	10

It seems that in this example the simple guess, the lower bound and the optimal guess are nearly equivalent strategies. However, the upper bound strategy with $\psi = -\infty$ is highly unlikely ($p << 0.001$). The optimal guess estimate for the prevalence is 12.2 per cent. In Figure 4-2 the deviance (as a function of the estimated prevalence) of the different strategies is interpolated; it can be seen that the 95 per cent tolerance region for the prevalence is from 10.6 to about 16 per cent, including both the lower bound estimate and the simple guess estimate.

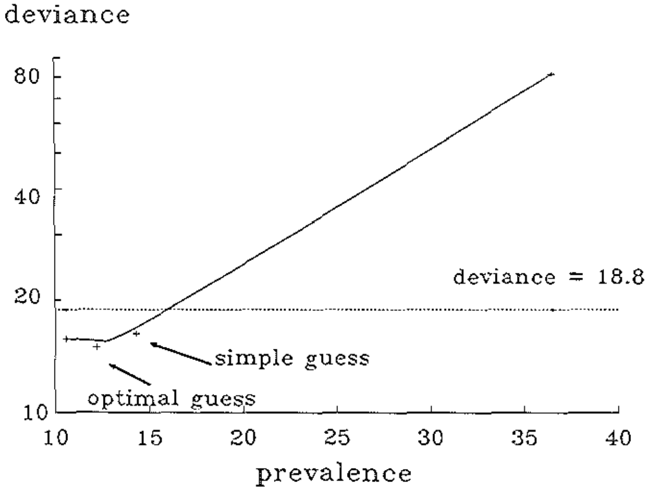


Figure 4-2 Deviance as a function of estimated prevalence of hypertension.

4.4 MODELLING A RECATEGORIZATION PROCESS

If the second process is a recategorizing process, then this process, too, can be described in terms of the general framework presented in this chapter. This will be shown in this section.

If the second process is a recategorizing mechanism, in which z_{21} means "correctly reproduced" and in which z_{22} means "falsely reproduced", it is easy to see that in fact one can only observe the following combined counts: (y_1+y_4) and (y_2+y_3).

When the Poisson reparametrization is used, the composite link matrix C is:

$$C = \begin{bmatrix} 1001 \\ 0110 \end{bmatrix}$$

If the second process recategorizes the outcomes of the first process, direct estimation by binomial algorithms is possible if $q_2=1-q_3$ (or $\beta_2=-\beta_3$). In this case only the parameters β_2 can be estimated. However, for a recategorizing process this condition is very unrealistic. Moreover, one is interested in β_1 rather than β_2.

4.5 DISCUSSION

This chapter presents a comprehensive approach to tackling the situation where two processes are observed and interest lies in only one, because the other process is a "manipulating" process. Additional information in the form of explanatory variables may be incorporated into the approach, provided that this information is available for all units. An algorithm is presented, applying the composite link approach. It is shown that sometimes it is also possible to apply standard binomial algorithms to calculate the MLEs. Of course, the EM algorithm (Dempster, Laird and Rubin, 1977) can be used as well.

By means of an example the problem of testing various strategies of "correcting" for missing data is presented. The choice is based on a likelihood ratio testing procedure by incorporating in a generalized model several strategies, of which the "simple-guess" strategy relates to the model-based direct-adjustment procedure of Rosenbaum (1987). Closely related, too, is the method of Conn, Lui and McGee (1989). These authors deal with the problem of estimating the incidence of home injury deaths in a situation where the place of occurrence is often unspecified. They use a logistic regression to estimate the probability of having a home injury and use this function in a one-time imputation phase for estimating the incidence of home injury deaths.

One should be aware that the appropriateness of the model can never be proved by goodness-of-fit testing. It can only be proved by resampling within the missing-data subgroup. However, the assumption, stated in the example, that the missing-data-generating process does not depend on age seems to be supported by the data from Table 4-II. For this specific generalization of the model, a likelihood ratio test was performed. The observed small age effects were not statistically significant. If the model specification had given a bad fit, also a super-binomial model could have been postulated and fitted with the computer program GLIM. For the present example, this did not seem necessary. In the example, the fitted value for ψ was 0.91. This means that there is a slight tendency towards an overrepresentation of normotensives within the missing-data group, so that the optimal-guess estimate for the prevalence goes toward the lower boundary. Additional data were available for only 51 per cent of the missing cohort. In this group, attending the periodic health examination in one of the following three years, the prevalence of hypertension turned out to be 11.1 per cent. Although this subsample, too, is no random subsample of the missing data cohort, this finding supports the parametric findings.

If the first process has more than two possible realizations, e.g., in misclassification problems of ordinal responses, the algorithm may be tedious to build, but the approach is essentially not different from the approach described in this chapter. For the McCullagh (1980) models the composite link matrix C may become a band matrix because misclassification for ordinal data will most probably consist of a shift of one category away from the true category.

4.6 REFERENCES

Birch MB (1963). Maximum likelihood in three way contingency tables. *J. R. Statist. Soc. B*, 25, 220-33.

Clayton D (1985). Using test-retest reliability data to improve estimates of relative risk: an application of latent class analysis. *Statistics in Medicine*, 4, 445-455.

Conn JM, Lui KJ and McGee DL (1989). A model-based approach to the imputation of missing data: home injury incidences. *Statistics in Medicine* 8, 263-266.

Copas JB (1988). Binary regression models for contaminated data. *J. R. Statist. Soc. B*, 50, 225-265.

Dempster AP, Laird NW and Rubin DB (1977). Maximum likelihood for incomplete data via the EM algorithm. *J. R. Statist. Soc. B*, 39, 1-38.

Dinse GE (1986). Nonparametric prevalence and mortality estimators for animal experiments with incomplete cause-of-death data. *J. Am. Statist. Assoc.* 81, 328-336.

Johnson NL and Kotz S (1969). *Distributions in statistics.* Discrete distributions. Houghton Mifflin Company, Boston.

Little RJA and Rubin DB (1987). *Statistical analysis with missing data.* John Wiley & Sons, New York.

McCullagh P (1980). Regression models for ordinal data. *J. R. Statist. Soc. B*, 42, 109-142.

McCullagh P and Nelder JA (1983). *Generalized linear models.* Chapman and Hall, London.

Palmgren J (1981). The Fisher information matrix for log-linear models arguing conditionally on observed explanatory variables. *Biometrika*, 68, 563-566.

Rosenbaum PR (1987). Model-based direct adjustment. *J. Am. Statist. Ass.* 398, 387-394.

Rubin DB (1976). Inference and missing data. *Biometrika*, 63, 581-592.

Steinhorst RK and Samuel MD (1989). Sightability adjustment methods for aerial surveys of wildlife populations. *Biometrics* 45, 415-425.

Thompson R and Baker RJ (1981). Composite link functions in generalized linear models. *Applied Statistics*, 30, 125-131.

Zeh JE, Ko D, Krogman BD and Sonntag R (1986). A multinomial model for estimating the size of a whale population from incomplete census data. *Biometrics* 42, 1-14.

CHAPTER 5

AN APPROXIMATE BARTLETT ADJUSTMENT FOR TESTING THE GOODNESS OF FIT OF MULTINOMIAL REGRESSION MODELS *

SUMMARY

For up to four categories and given sufficient replications per design point (eight or more seem to be enough), an adjusted likelihood ratio statistic for testing the goodness of fit of multinomial regression models is proposed. This adjustment does not depend on the multinomial model used and therefore will have a wide range of applications. It is certainly useful for the proportional odds model, introduced by McCullagh. This proportional odds model is used in some simulation studies to investigate the performance of the new statistic.

* After: Lugtenburg D (1991) *J. R. Statist. Soc. B*, 53, 393-398

5.1 INTRODUCTION

Ordinal response variables may often be appropriately analysed by application of ordinal regression models, as introduced by McCullagh (1980). The likelihood ratio (LR) statistic, which is often used as the goodness-of-fit test statistic, is a special member of a more general family of goodness-of-fit statistics; see Cressie and Read (1984). For small sample sizes, however, this statistic is not distributed as the corresponding asymptotic chi-squared distribution; see Koehler (1986) and Agresti and Yang (1987) for the case of log-linear models.

One way of improving the statistic is to use Bartlett's (1937) correction factor; see also Barndorff-Nielsen and Cox (1984), in which a simple proof is presented of the correctness of this statistic up to order $O(N^{-3/2})$ where N is the number of observations. Recently, Barndorff-Nielsen and Hall (1987) proved that, in regular cases, this adjustment is even correct up to order $O(N^{-2})$. The correction factor has been applied in many contexts, for example in multivariate normal distributions (Møller, 1986; Porteous, 1985). Results obtained by Lawley (1956) can be applied in the computation of the adjusted statistic in more complicated situations. Cordeiro (1983) applies this approach to univariate generalized linear models. In another paper Cordeiro (1987) presents Bartlett adjustments when the dispersion parameter is unknown. A parameter invariant alternative to Lawley's results has been derived by Ross (1987) for curved exponential families.

In the present chapter, an approximate Bartlett adjustment is proposed for testing the goodness of fit of multinomial regression models. This adjustment does not correct the distribution of the test statistic up to any order, but it will be shown to be adequate in some situations. In section 5.2, the LR statistic will be separated into two parts: a model-dependent part and a model-independent part. The approximate Bartlett correction factor, derived in section 5.3, is based primarily on the model-independent part of the LR statistic. In section 5.4, some simulation results are presented, which illustrate the performance of the improved statistic.

5.2 SEPARATION OF LIKELIHOOD RATIO STATISTICS

The number of times that response category j is observed for the categorical response variate at design point i is denoted by y_{ij} ($i=1,2,..., s$; $j=1,2,...,k$), where s is the number of design points and k is the number of response categories. For a model under investigation with p unknown parameters, the fitted value f_{ij} is the maximum likelihood estimate (MLE) of the expected value of the category count y_{ij}. The resulting maximum multinomial log-likelihood is:

$$L_p=\sum_{i=1}^{s}\sum_{j=1}^{k}y_{ij}\ \log(\frac{f_{ij}}{n_i}),$$

with

$$n_i=\sum_{j=1}^{k}y_{ij}.$$

For the saturated model this simplifies to

$$L_n=\sum_{i=1}^{s}\sum_{j=1}^{k}y_{ij}\ \log(\frac{y_{ij}}{n_i}).$$

The value of the LR test statistic is $2(L_n\text{-}L_p)$. This statistic asymptotically, as $\min_i n_i \rightarrow \infty$, has a chi-squared null distribution with *s(k-1)-p* degrees of freedom (DF); see, for example, McCullagh and Nelder (1983).

If p_{ij} denotes the true but unknown category probabilities, the exact log-likelihood *L* is

$$L=\sum_{i=1}^{s}\sum_{j=1}^{k}y_{ij}\ \log(p_{ij}).$$

The LR statistic now can be separated into a model-dependent and a model-independent part: $LR=LR_{mi}\text{-}LR_{md}$ with $LR_{mi}=2(L_n\text{-}L)$ and $LR_{md}=2(L_p\text{-}L)$.

If the model is correct, LR_{mi} has, asymptotically, the chi-squared distribution with *s(k-1)* DF, and LR_{md} is asymptotically chi-squared distributed with *p* DF; see, for example, Dobson (1983).

A simulation study was performed, to check whether the larger part of the bias, caused by considering the null distribution of the test statistic to follow a chi-squared distribution with *s(k-1)-p* DF, is associated with LR_{mi} or with LR_{md}. Some results of this study are presented in Table 5-I. Observations were generated for six different parameterizations of the proportional odds model of McCullagh (1980). All six models relate to an ordinal scale of response with four response categories (scores j=1,2,3 or 4).

TABLE 5-I

Simulation results*

Model	ß$_1$	ß$_2$	Model-independent mean (expected: 18)	Model-dependent mean (expected: 5)
I	0.2	0.0	21.48	4.80
II	0.3	0.0	21.76	5.57
III	0.4	0.0	21.19	4.78
IV	0.1	0.2	22.36	5.72
V	0.2	0.4	21.17	5.08
VI	0.3	0.6	20.09	5.04

* For 100 simulations per study the observed means of the model-dependent part and the model-indepent part of the LR statistic are listed.

Cumulative probabilities are denoted by

$$\gamma_{ij}=\sum_{t=1}^{j} p_{it}$$

The model used in the simulations can be described by

$$logit(\gamma_j)=\theta_j e_j-\beta_1 x_1-\beta_2 x_2$$

for $j=1,2,3$, with

$$\gamma_j=(\gamma_{1j},\ldots,\gamma_{sj})^T;$$

e_j is an s-vector with elements unity. The vectors x_1 and x_2, too, are of dimension s.

The three cut-off points θ_j are treated as fixed (θ_1=-1,θ_2=0 and θ_3=1). Six design points are considered for each model (s=6), with x_1=(0 0 0 1 1 1)^T and x_2=(0 1 2 0 1 2)^T. The six models differ only with respect to the values of the parameters $\beta 1$ and $\beta 2$.

The expected value of LR$_{mi}$ is asymptotically equal to 18 and, if the model is correct, the expected value of LR$_{md}$ is asymptotically equal to 5. Table 5-I shows that LR$_{mi}$ is subject to the larger bias. Similar conclusions with respect to this statistic are reported by Koehler and Larntz (1980).

5.3 APPROXIMATE BARTLETT CORRECTION FACTOR

To apply Bartlett's correction factor the expected value of the second-order Taylor approximation of LR$_{mi}$, denoted by E*(LR$_{mi}$), and of LR$_{md}$ need to be derived. Since LR$_{mi}$

shows the larger bias, only $E^*(\mathrm{LR}_{mi})$ is calculated. The formulae for k=2 up to k=4 are presented in Appendix 5-A. The derivation for k=2 can be found in Cordeiro (1983). In Appendix 5-A helpful remarks are given for k=4. For k>4, direct calculation is not trivial, but calculations can be simplified with the help of software like REDUCE or MAPLE.

The following correction factor $c(p_{11},..,p_{sk})$ will be used for the LR statistic:

$$c(p_{11},..., p_{sk}) = \frac{s\ (k-1)-p}{E^*(LR_{mi})-p}\ ,$$

in which $E^*(\mathrm{LR}_{mi})$ is a function of the unknown exact category probabilities p_{ij}. Since the p_{ij} are not known in practice, the proposal is to use correction factor $c(p^*_{11},..,p^*_{sk})$ with $p^*_{ij}=f_{ij}/n_i$, to calculate the corrected test statistic $\mathrm{LR}_c=c(p^*_{11},..,p^*_{sk})\ \{2(Ln\text{-}Lp)\}$.

5.4 SIMULATION RESULTS

To compare the two statistics LR and LR_c, simulations were performed for each of the six models described in section 5.2. The results, which can be found in Table 5-II, are strongly in favour of LR_c. In these simulations, the null hypothesis is correct. A study of the non-null behaviour is beyond the scope of this chapter and will not be illustrated further.

TABLE 5-II

Number of exceedances of LR_c for three values of α *

Model	α=0.20	α=0.10	α=0.05
I	28 (48)	16 (30)	6 (18)
II	19 (44)	12 (22)	4 (13)
III	27 (41)	17 (27)	7 (20)
IV	32 (46)	17 (35)	9 (20)
V	21 (42)	11 (23)	5 (11)
VI	17 (32)	6 (19)	4 (12)

* For the six different models, the table lists the number of times (out of 100) that the corrected test statistic LR_c exceeds the α percentage point of the chi-squared distribution with 13 DF. The corresponding numbers for the uncorrected statistic LR are given in parantheses.

5.5 REFERENCES

Agresti A and Yang MC (1987). An empirical investigation of some effects of sparseness in contingency tables. *Comput. Statist. Data Anal.*, 5, 9-21.

Barndorff-Nielsen OE and Cox DR (1984). Bartlett adjustments to the likelihood ratio statistics and the distribution of the maximum likelihood estimator. *J.R.Statist.Soc. B*, 46, 483-495.

Barndorff-Nielsen OE and Hall P (1987). On the level-error of the Bartlett adjustment of the likelihood ratio statistic. *Biometrika*, 75, 374-378.

Bartlett MS (1937). Properties of sufficiency and statistical tests. *Proc. R. Soc. A*, 160, 268-282.

Cordeiro GM (1983). Improved likelihood ratio statistics for generalized linear models. *J.R.Statist.Soc. B*, 45, 404-413.

Cordeiro GM (1987). On the corrections to the likelihood ratio statistics. *Biometrika*, 74, 265-274.

Cressie N and Read TRC (1984). Multinomial goodness-of-fit tests. *J.R.Statist.Soc. B*, 46, 440-464.

Dobson AJ (1983). *Introduction to statistical modelling.* Chapman and Hall, London.

Johnson NL and Kotz S (1969). *Distributions in statistics. Discrete distributions.* Houghton Boston: Houghton Mifflin.

Koehler KJ (1986). Goodness-of-fit tests for log-linear models in sparse contingency tables. *J. Am. Statist.Ass.*, 81, 483-493.

Koehler KJ and Larntz K (1980). An empirical investigation of goodness-of-fit statistics for sparse multinomials. *J. Am. Statist. Ass.*, 75, 336-344.

Lawley DN (1956). A general method for approximating to the distribution of the likelihood ratio criteria. *Biometrika*, 43, 295-303.

McCullagh P (1980). Regression models for ordinal data (with discussion). *J.R.Statist.Soc. B*, 42, 109-142.

McCullagh P and Nelder JA (1983). *Generalized Linear Models*. London: Chapman and Hall.

Møller J (1986). Bartlett adjustments for structured covariances. *Scand. J. Statist.*, 13, 1-15.

Porteous BT (1985). Improved likelihood ratio statistics for covariance selection models. *Biometrika*, 72, 97-101.

Ross WH (1987). The expectation of the likelihood ratio criterion. *Int. Statist. Rev.*, 55, 315-330.

APPENDIX 5-A

Suppose that there is one design point, $y_{ij}=y_j$, $y_{ij}/n_i=q_j$, $n_i=n$ and $p_{ij}=p_j$. For up to four categories the second-order Taylor approximations of $E(\mathrm{LR}^*_{\mathrm{m}l}/2)$ are then:

$$E\ (LR^*_{ml}/2) = A_0(k)+\sum_{t=1}^{k-1} A_1(t),$$

$$E\ (LR^*_{ml}/2) = A_0(k)+\sum_{t=1}^{k-1} A_1(t)+\sum_{t=1}^{k-1} A_2(t,k)+\sum_{t=1}^{k-1}\ \sum_{j=t+1}^{k-1} A_3(t,\ j,k),$$

$$E\ (LR^*_{ml}/2) = A_0(k)+\sum_{t=1}^{k-1} A_1(t)+\sum_{t=1}^{k-1} A_2(t,k)+\sum_{t=1}^{k-1}\ \sum_{j=t+1}^{k-1} A_3(t,\ j,k)\ +$$

$$+\sum_{t=1}^{k-1}\ \sum_{j=t+1}^{k-1}\ \sum_{h=j+1}^{k-1} A_4(t,\ j,h,k)$$

for $k=2$, $k=3$ and $k=4$, respectively, with

$$A_0(k)=(k-1)/2,$$

$$A_1(t)=(-p_t+1/p_t)/12n,$$

$$A_2(t,k)=\{p_t(1-2p_t)p_k^2/6-p_k\ p_t(1-2p_t)(1-p_k-p_t)/3\ +$$

$$+p_k\ p_t^2(1-p_t)/4-3p_t^2(1-p_t)(1-p_k-p_t)/4\}/np_k^3,$$

$$A_3(t,\ j,k)=\{2(p_t\ p_j)^2+p_t\ p_j(1-p_t)(1-p_j)\}/2np_k^3,$$

$$A_4(t,\ j,h,k)\ =\ -p_t\ p_j\ p_h\ /np_k^2.$$

With s design points these formulae are appropriate for every design point i separately as functions of the corresponding category probabilities p_{ij}, with $j=1,\ldots,k$ The relevant formulae in these cases are simply the corresponding sums.

The derivation of the formula for k=4 will be presented in five steps.

(a) The Taylor series around $L(p_1,..,p_{k-1})$ of $L(q_1,..,q_{k-1})$ is:

$$L(q_1,...,q_{k-1}) = L(p_1,...,p_{k-1})+\sum_{j=1}^{k-1}(q_j-p_j)(\frac{\partial L}{\partial q_j})_{q=r} +$$

$$+\frac{1}{2!}\sum_{t=1}^{k-1}\sum_{j=1}^{k-1}(q_t-p_t)(q_j-p_j)(\frac{\partial^2 L}{\partial q_t \partial q_j})_{q=r} +$$

$$+\frac{1}{3!}\sum_{j=1}^{k-1}\sum_{t=1}^{k-1}\sum_{h=1}^{k-1}(q_j-p_j)(q_t-p_t)(q_h-p_h)(\frac{\partial^3 L}{\partial q_j \partial q_t \partial q_h})_{q=r} +...$$

with $q=(q_1 ,.., q_{k-1})$, $r=(p_1 ,.., p_{k-1})$ and with the constraint $\Sigma^k_{j=1}q_j$=1.

(b) The log-likelihood $L(q_1 ,.., q_{k-1})$ is a function of the observed relative frequencies q_j: $L(q_1 ,.., q_{k-1})=\Sigma^k_{j=1}nq_j\log q_j$ with the constraint $\Sigma^k_{j=1} q_j$=1. This results in the following formulae for $j,t \in \{1,..,k\text{-}1\}$.

$$\frac{\partial L}{\partial q_t}=n(\log q_t-\log q_k)$$

$$\frac{\partial^2 L}{\partial q_t^2}=\frac{n}{q_t}+\frac{n}{q_k}$$

and

$$\frac{\partial^2 L}{\partial q_t q_j}=\frac{n}{q_k}$$

for $t \neq j$, and so on.

(c) After the derivatives have been substituted into the Taylor formula, we can calculate the expected value of the difference of both log-likelihoods $L(q)$ and $L(r)$, correct up to second order. For this, formulae for the product moments are needed. The multinomial cumulants and the transformations needed to give the product moments are both listed in Johnson and Kotz (1969). Formulae like $E(q_1\text{-}p_1)(q_2\text{-}p_2)$ = - $np_1\,p_2\,(1/n^2)$ are thereby derived.

(d) The terms of the expected value involving a first-order derivative are

$$\sum_{t=1}^{k-1} E(q_t-p_t)(n \log p_t-n \log p_k)=0.$$

The terms of the expected value involving a second-order derivative are

$$\frac{1}{2}\sum_{t=1}^{k-1} E(q_t-p_t)^2(\frac{n}{p_t}+\frac{n}{p_k})+E\{(q_1-p_1)(q_2-p_2)+(q_1-p_1)(q_3-p_3)\ +$$

$$+(q_2-p_2)(q_3-p_3)\}\frac{n}{p_k}=\frac{1}{2}\left\{\sum_{t=1}^{k-1}\frac{1}{n^2}np_t(1-p_t)(\frac{n}{p_t}+\frac{n}{p_k})\right\}\ +$$

$$-(p_1\ p_2-p_1\ p_3-p_2\ p_3)\frac{1}{p_k}=\frac{1}{2}(k-1)=A_0(4)\ .$$

The terms of the expected value involving third-order and fourth-order derivatives are derived as above. $\Sigma\, A_1\ (t)$ and $\Sigma\, A_2\ (t,k)$ result from combining the terms involving a third-order derivative for $t=j$ and involving a fourth-order derivative for $t=j=h$, for terms respectively independent of and dependent on p_k. $\Sigma\, A_3\ (t,j,k)$ results from combining the terms involving a third-order derivative for $t \neq j$ and involving a fourth-order derivative for one pair of the indices (t,j,h) being identical. $\Sigma\, A_4(t,j,h,k)$ results from combining all remaining terms.

(e) The total of the terms involving up to the fourth-order derivative (higher order derivatives do not add terms of second-order precision) provides the formula for k=4, as presented at the beginning of this appendix.

CHAPTER 6

SMALL-SAMPLE RESULTS OBTAINED WITH THE UNCONDITIONAL WILCOXON-MANN-WHITNEY TEST STATISTIC FOR ORDINAL DATA

SUMMARY

For up to 10 observations per sample, and for up to five categories, tables with unconditional critical values are presented for the Wilcoxon-Mann-Whitney test statistic when the response variate is ordinal. As the underlying process is considered to be a multinomial process, these critical values are not dependent on ties but, instead, on the category probabilities of this multinomial process. If these category probabilities equal 1/*k*, where *k* is the number of categories, the critical values appear to be conservative. Extensive comparisons with respect to the difference between nominal and actual p-values suggest that this unconditional statistic is superior to the conditional test statistic.

6.1 INTRODUCTION

The Wilcoxon-Mann-Whitney statistic tests the null hypothesis that the distributions of two populations are identical, given independent random samples. More specifically, it tests the hypothesis that $\Gamma=\Pr(T_1>T_2)=1/2$, with T_1 being the observed outcome on a subject from sample 1 and T_2 the observed outcome on a subject from sample 2. This statistic is suitable in particular in the testing for a shift between the distributions. If the response variate has an ordinal scale with up to k possible response categories, the problem of ties occurs. Lehmann (1975) has shown how for this situation it is possible to calculate exact p-values. His approach conditions the possible outcomes on sample sizes and on category frequencies. Calculation is not difficult but tedious. Mehta et al. (1984) and Verbeek and Kroonenberg (1985) present algorithms that reduce some of the calculations. Emerson and Moses (1985) suggest that this conditional exact test should always be used for sample sizes of 10 or less and asymptotic results only for larger sample sizes.

For large samples an unconditional approach has been presented for calculating confidence limits for Γ, see Halperin, Hamdy and Thall (1989), who postulate a multinomial underlying process. They follow the approach discussed in Halperin, Gilbert and Lachin (1987). Postulating such a multinomial underlying process makes it possible to calculate critical values of the Wilcoxon-Mann-Whitney statistic, as functions of the multinomial parameters. Whenever these multinomial category probabilities are supposed to equal $1/k$, this appears to lead to a conservative test.

Comparisons between conditional and unconditional tests have been made in the 2x2 situation by Suissa and Shuster (1985), using a Z-statistic for sample sizes larger than 10. They explicitly stated to be the first authors recommending unconditional testing to obtain an increase in power. They calculated sample sizes accordingly, with the nuisance parameter being eliminated following Basu (1977), resulting in a conservative test. The availability of several levels of prior information on the nuisance parameter has led to a probability model with an unconditional alternative to Fisher's exact test, see Rice (1988).

For the more general kx2 situation, we will make a comparison in this chapter between the two approaches for the Wilcoxon statistic with respect to the difference between the actual and nominal p-values. For a wide range of multinomial category parameter settings, for k=3 and k=4 and for sample sizes that range from 4 to 8, exact actual p-values will be calculated for both approaches. In all these situations, the unconditional approach, using conservative critical values, outperforms the conditional approach. For the most heterogeneous situation (category probabilities equalling $1/k$), in which the unconditional approach has actual p-values that are mostly at least 90% of the nominal p-values, the unconditional approach sometimes has a coverage of only 50%.

In section 6.2, Lehmann's approach is discussed. The unconditional approach is presented in section 6.3. Also, critical values are listed there for the case where the category probabilities equal $1/k$, for sample sizes up to ten, with k being three to five and α equal to 0.01, 0.05 and 0.10. With the aid of several parameter settings of the multinomial process, the magnitude of conservatism for both approaches is shown in section 6.4. Power comparisons between the two approaches are presented in section 6.5. Finally, section 6.6 presents conclusions and recommendations.

6.2 THE CONDITIONAL WILCOXON STATISTIC

If we denote the m observations from the first sample X_u and the n observations from the second sample Y_v, the outcome of the Wilcoxon-Mann-Whitney statistic for testing the null hypothesis of no difference against the one-sided alternative of a shift to the right ($X<Y$) corrected for ties, is:

$$W=\sum_{u=1}^{m}\sum_{v=1}^{n}\Phi(X_u,Y_v) \tag{1}$$

where $\phi(x,y)=\delta_1$ if $x<y$, δ_2 if $x=y$ and δ_3 if $x>y$. Here, $\delta_1=1$, $\delta_2=0.5$ and $\delta_3=0$. This, naturally, is equivalent (apart from a constant) to the sum of ranks for the Y_v's evolving from ranking all $m + n$ observations.

Another expression, suitable in particular for ordinal data with up to k possible outcomes in order of magnitude 1,2,..k, is the following. The number of observations that fall into category j for sample i is denoted f_{ij} with i equalling 1 or 2 and j ranging from 1 to k. The outcome W of the test statistic, w, is:

$$W=\sum_{j=1}^{k} f_{1j}\ (\delta_2\, f_{2j}+\delta_1\sum_{t=j+1}^{k} f_{2t}\ ) \tag{2}$$

For deriving the conditional distribution of this statistic, the $n + m$ (not necessarily different) ranks are used as a starting point to generate all possible permutations of m ranks. For each permutation the outcome of the test statistic with its probability are calculated, assuming a uniform distribution of all ranks. The proportion of generated outcomes of the test statistic that at least equal the observed outcome *W*, represent the one-sided p-value. It is easy to see that this underlying distribution is conditional both on the sample sizes (*m* out of *m* + *n* ranks) and on the category frequencies. For example, if category 2 was actually observed four times, the corresponding mean rank can be selected from the *n* + *m* possible ranks not more than four times. This distribution has to be generated for every single observed outcome. In the next section a different method is presented, which generates critical values applicable in every practical situation, independent of ties that may have occurred.

6.3 THE UNCONDITIONAL WILCOXON STATISTIC

Suppose observations from population *i* (*i*=1,2) follow a multinomial distribution with category probabilities p_{ij} (*j*=1,..,*k*). One specific realization (f_{i1} ,.., f_{ik}) for sample *i* then has log-probability

$$\log(p_i) = f_{i1}\log(p_{i1}) + \dots + f_{ik}\log(p_{ik}) + \log n_i! - \log f_{i1}! - \dots - \log f_{ik}!$$

with

$$n_i = \sum_{j=1}^{k} f_{ij}$$

If for each possible realization (f_{11} ,..., f_{1k} , f_{21} ,..., f_{2k}) the probability to occur, $p_1 p_2$, is calculated, then for each possible realization *W* of the Wilcoxon-Mann-Whitney statistic, its one-sided p-value, p, is calculated as:

$$p = \sum_{f_{11},\dots,f_{2k} \text{ with } w \geq W} p_1\, p_2$$

in which *W* is calculated as in (2).

Obviously, critical values for corresponding α values are easy to calculate if the multinomial parameters $p_{1j}=p_{2j}$ are known. In practice this situation rarely occurs. For sufficiently large samples, maximum likelihood estimates can be used to estimate these parameters. For small sample sizes one might search for a specific parameter set that results in conservative critical values. In the next section it will be shown that the parameter set $p_{ij}=1/k$ appears to be such a parameter set. Using this parameter setting, for *k*=3 to *k*=5, the lower critical values W_L have been calculated for sample sizes up to 10 and α=0.01, 0.05 and 0.10 (so: $\Pr(w \leq W_L) \leq \alpha$). Tables 6-Ia to 6-Ic show the results.

TABLE 6-Ia

Lower critical values W_L for which $\Pr(w \le W_L) \le \alpha$ with $\alpha=0.01$, $\alpha=0.05$ and $\alpha=0.1$ on the first, second and third rows, respectively, for $k=3$

	n							
m	3	4	5	6	7	8	9	10
3	*							
	0.5							
	1.5							
4	*	0.5						
	1.0	2.5						
	2.5	3.5						
5	0.5	1.0	2.0					
	2.0	3.5	5.0					
	3.0	4.5	6.5					
6	1.0	1.5	3.5	4.5				
	2.5	4.5	6.5	8.0				
	4.0	6.0	8.0	10.0				
7	1.0	2.5	4.0	5.5	7.5			
	3.5	5.5	7.5	10.0	12.0			
	5.0	7.5	9.5	12.5	15.0			
8	1.5	3.5	5.0	7.5	9.5	11.5		
	4.0	6.5	9.5	11.5	14.5	17.0		
	5.5	8.5	11.5	14.5	17.5	20.0		
9	2.5	4.0	6.5	8.5	11.0	13.5	16.0	
	4.5	7.5	10.5	13.5	17.0	19.5	23.0	
	6.5	9.5	13.0	16.5	19.5	23.0	26.5	
10	2.5	5.0	7.5	10.0	13.0	15.5	18.5	21.5
	5.5	8.5	12.0	15.5	19.0	22.5	25.5	29.5
	7.5	11.0	14.5	18.5	22.5	26.0	30.0	33.5

TABLE 6-Ib

Lower critical values W_L for which $\Pr(w \leq W_L) \leq \alpha$ with $\alpha=0.01$, $\alpha=0.05$ and $\alpha=0.1$ on the first, second and third rows, respectively, for $k=4$

				n				
m	3	4	5	6	7	8	9	10
3	*							
	0.5							
	1.0							
4	*	0.5						
	1.0	2.5						
	2.0	3.5						
5	0.0	1.0	2.0					
	2.0	3.0	4.5					
	3.0	4.5	6.0					
6	0.5	1.5	3.0	4.0				
	2.5	4.5	6.0	7.5				
	4.0	5.5	8.0	10.0				
7	1.0	2.5	4.0	5.5	7.0			
	3.0	5.5	7.5	9.5	11.5			
	4.5	7.0	9.5	12.0	14.5			
8	1.0	3.0	5.0	7.0	9.0	11.0		
	4.0	6.5	9.0	11.5	14.0	16.5		
	5.5	8.0	11.0	14.0	17.0	19.5		
9	2.0	3.5	6.0	8.5	10.5	13.0	15.5	
	4.5	7.5	10.5	13.0	16.0	19.0	22.0	
	6.5	9.5	12.5	16.0	19.5	22.5	26.0	
10	2.5	4.5	7.0	9.5	12.5	15.0	18.0	20.5
	5.0	8.5	11.5	15.0	18.5	22.0	25.5	28.5
	7.0	10.5	14.5	18.0	22.0	25.5	29.5	33.0

TABLE 6-Ic

Lower critical values W_L for which $\Pr(w \leq W_L) \leq \alpha$ with $\alpha=0.01$, $\alpha=0.05$ and $\alpha=0.1$ on the first, second and third rows, respectively, for $k=5$

	n							
m	3	4	5	6	7	8	9	10
3	*							
	0.5							
	1.0							
4	*	0.0						
	1.0	2.0						
	2.0	3.0						
5	0.0	0.5	1.5					
	1.5	3.0	4.5					
	3.0	4.5	6.0					
6	0.5	1.5	2.5	4.0				
	2.5	4.0	6.0	7.5				
	3.5	5.5	7.5	9.5				
7	0.5	2.0	3.5	5.5	7.0			
	3.0	5.0	7.0	9.5	11.5			
	4.5	7.0	9.5	11.5	14.0			
8	1.0	3.0	4.5	6.5	8.5	10.5		
	3.5	6.0	8.5	11.0	13.5	16.5		
	5.5	8.0	11.0	14.0	16.5	19.5		
9	1.5	3.5	5.5	8.0	10.5	12.5	15.0	
	4.5	7.5	10.0	13.0	16.0	19.0	22.0	
	6.0	9.5	12.5	16.0	19.0	22.5	26.0	
10	2.0	4.0	6.5	9.5	12.0	14.5	17.5	20.5
	5.0	8.5	11.5	15.0	18.0	21.5	25.0	28.5
	7.0	10.5	14.5	18.0	21.5	25.5	29.0	33.0

6.4 CONSERVATISM OF CRITICAL VALUES FROM BOTH APPROACHES

The upper critical values, W_U, of w for the case where $p_{ij}=0.25$ ($k=4$) and $n=m=8$ are calculated as described in the previous section for the unconditional statistic. These are 44.5, 47.5 and 53, respectively, for $\alpha=0.1$, 0.05 and 0.01. These values can be calculated from Table 5-Ib as well, using the relation $W_U=nm-W_L$ from the lemma:

$$if\ Pr(w\leq W_L)\leq\alpha\ then\ Pr(w\geq n\ m-W_L)\leq\ \alpha$$

which is obvious for the unconditional approach. So, if $\alpha=0.1$, $W_U=64\text{-}19.5=44.5$. Considering different parametrizations of the multinomial underlying process, we can calculate the exact (or actual) probability of a more extreme value than this specific critical value. In several figures, results of such calculations are summarized for 30 different underlying multinomial distributions, covering a range of possibilities for the difference between the maximum category probability, p_{max} and the minimum category probability, p_{min}. The exact probabilities calculated are shown as functions of the range $R=p_{max}\text{-}p_{min}$ in Figure 6-1 for $k=3$: ($n=4$, $m=4$), ($n=4$, $m=8$), and ($n=8$, $m=8$) and in Figure 6-2 for $k=4$: ($n=4$, $m=4$), ($n=4$, $m=8$) and ($n=8$, $m=8$).

For the conditional test statistic, the following approach has been applied. For all possible combinations of r_1, r_2, r_3, r_4 (in which $r_j=f_{1j}+f_{2j}$) and observed value W of the test static w, the p-value corresponding to this W is calculated as described in Section 6.2. If p≤0.05 this is denoted by

$$S\ (\ r_1,r_2,r_3,r_4,W\ )=1\ ,$$

otherwise

$$S\ (\ r_1,r_2,r_3,r_4,W\ )=0\ .$$

Suppose that for the multinomial model the probability that combination r_1, r_2, r_3, r_4, W actually occurs is denoted by $p(r1,r2,r3,r4,W)$. The p-value, which is the probability that the null-hypothesis will be rejected if the null hypothesis (including the multinomial assumption) is true, then is:

$$p\text{-}value=\sum_{r_1,r_2,r_3,r_4,W} p\ (\ r_1,r_2,r_3,r_4,W\ )\ S\ (\ r_1,r_2,r_3,r_4,W\ )\ .$$

Results from the conditional approach for the situations where the unconditional approach was used are shown in Figures 6-1 and 6-2 as well.

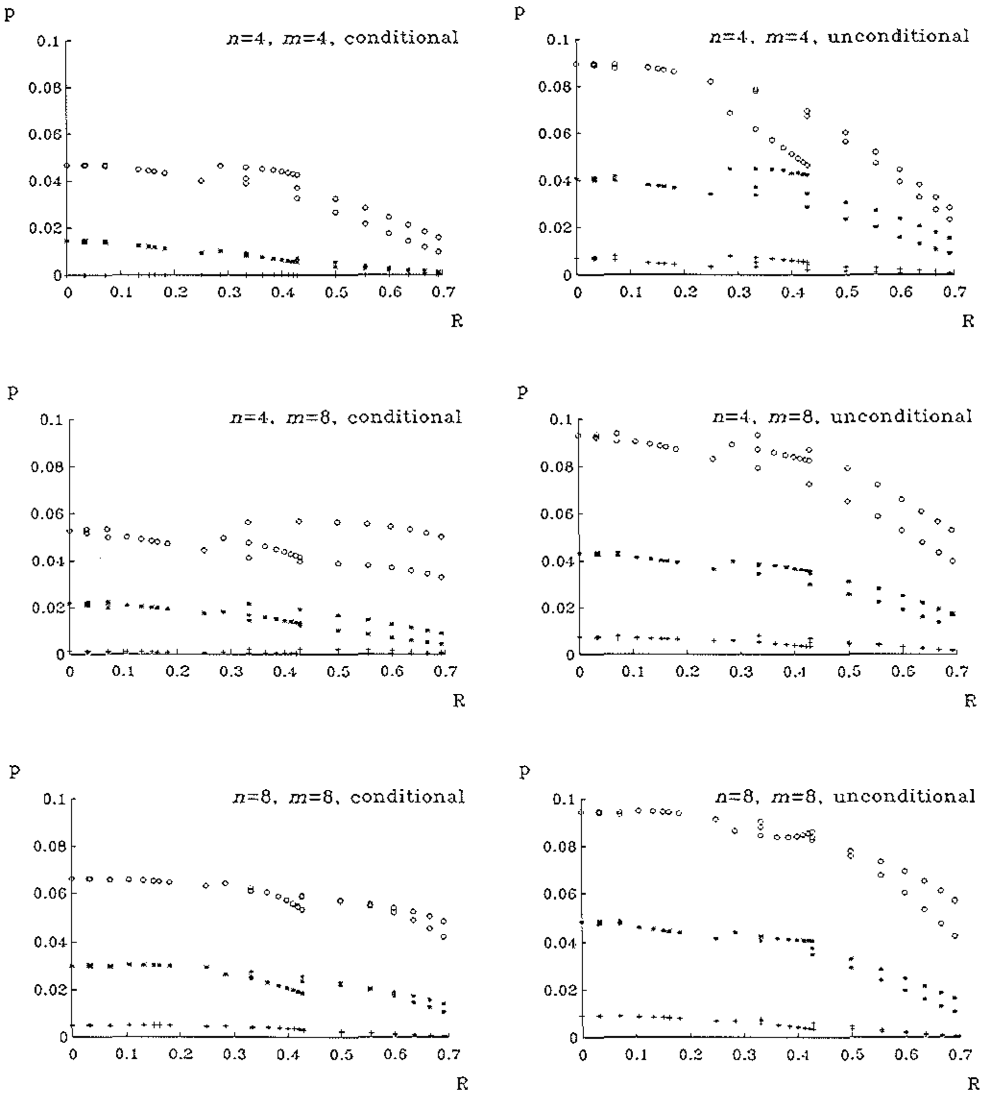


Figure 6-1 Exact (or actual) p-values for specific α-levels (o=0.10, $*$=0.05, +=0.01) for k=3 as a function of the range R for n=4 with m=4, n=4 with m=8 and n=8 with m=8 (conditional and unconditional).

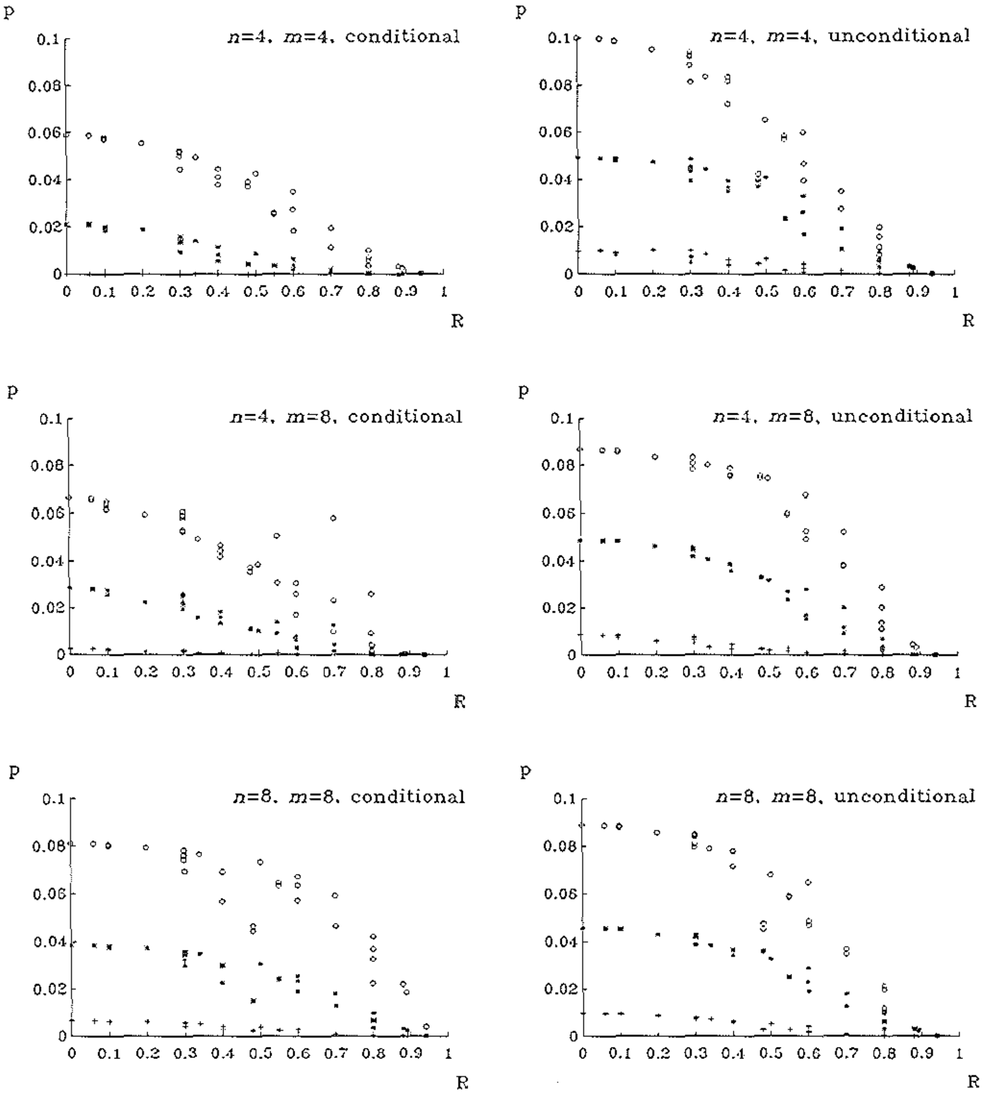


Figure 6-2 Exact (or actual) p-values for specific α-levels (o=0.10, *=0.05, +=0.01) for k=4 as a function of the range R for n=4 with m=4, n=4 with m=8 and n=8 with m=8 (conditional and unconditional).

The graphs show that the $p_{ij}=1/k$ substitution generally leads to conservative critical values. For the unconditional approach in which $\alpha=0.05$, this conservatism is less than 30% in most cases, viz. with $R(p_{.j})<0.3$, and $p_{min}>0.05$. Only for cases with extreme variations in p_{i}, like $R(p_{.j})>0.8$, the exact p-value is considerably overestimated. For $\alpha=0.01$ more cases show considerable overestimation.

For the conditional approach, this conservatism is even much higher. Here, one can see, that even in the situation where $R(p_{.j})=0$, the exact coverage of critical values might be even less than 60% of the nominal value. This is true, in particular if there are few observations, so if the probability of zero-frequencies is considerable. When there is a zero-frequency, the conditional approach ignores that category completely, whereas the probability that this frequency is zero by chance is not 0, especially not if the multinomial underlying probabilities are specified.

6.5 COMPARISON OF THE POWER OF THE TWO STATISTICS

In considering the power of a test statistic, an alternative hypothesis has to be specified to show the probability of rejecting the null hypothesis if this alternative is true. Because of the nonparametric approach, this may cause difficulties for the Wilcoxon statistic. In testing for a shift in distribution, Collings and Hamilton (1988) used a general concept of the alternative, approximating the power with a bootstrap method. In this section, however, we follow the approach of Whitehead (1983), who discusses a general way to look at the power of the Wilcoxon statistic. If we denote γ_{ij} the probability that a response in category j or worse is observed on a subject from population i with $i=1$ or 2, population differences are defined as:

$$\Delta_j = \mathit{logit}\ (\gamma_{1j}) - \mathit{logit}\ (\gamma_{2j})\ ,\ \text{where}\ j = 1,\ldots,k\text{-}1.$$

Whenever the two population distributions are equal $\Delta_j=0$ for all j. If $\Delta_j < 0$ for all j, then there is a shift to the left for the distribution of the second population. Although many realizations of Δ_j may occur, the following is chosen to investigate the power of the Wilcoxon statistic:

$$\Delta_j = \Delta \text{ for all } j.$$

This choice of a parameter to study the power of a statistic with regard to a given realization of Δ, also relates this concept to the cumulative log odds model, discussed by McCullagh (1980). He shows that in a comparison of two populations, with underlying logistic distributions, the comparison may be done with a model as described above. In this case a pure shift-model emerges with Δ being the difference, or shift, between the two population distributions.

Both strategies for deriving one-sided p-values of the Wilcoxon-statistic will now be considered.

For the conditional test statistic the following approach is used. For all possible combinations of r_1 ,r_2 ,r_3 ,r_4 (in which $r_j=f_{1j}+f_{2j}$) and observed value W of the test statistic w, the p-value corresponding to this W is calculated as described in section 6-2. If p≤0.05 this is denoted by

$$S(\, r_1,r_2,r_3,r_4,W\,)=1,$$

otherwise

$$S\ (\, r_1,r_2,r_3,r_4,W\,)=0\ .$$

For the multinomial model, the probability that combination r_1 ,r_2 ,r_3 ,r_4 ,W actually occurs is denoted $p^*(r_1,r_2,r_3,r_4,W)$. If the null hypothesis is not met, this is not equivalent to $p(r_1,r_2,r_3,r_4,W)$ from the previous section. The power, which is the probability that the null hypothesis will be rejected if the alternative hypothesis is true, then is:

$$POWER=\sum_{r_1,\,r_2,\,r_3,\,r_4,\,W} p^*\ (\, r_1,r_2,r_3,r_4,W\,)\ S\ (\, r_1,r_2,r_3,r_4,W\,)\ .$$

Exact power calculations are presented in Table 6-II for various values of Δ when k=4 and n=m=8, using several null distributions, including p_{ij}=$1/k$.

For the unconditional test statistic the power calculations are made by calculating Pr(w≤W_L) with W_L=16.5.

From tables like this it has been concluded that the exact conditional test in general has less power to detect a difference Δ than the conservative unconditional test has.

TABLE 6-II

Exact power for one-sided tests with $\alpha=0.05$ and as alternative hypothesis a shift to the left, under specific null hypotheses for two statistics where $k=4$ and $n=m=8$

Alternative hypothesis	Unconditional Wilcoxon	Conditional Wilcoxon
(1) $\Delta = 0.0$	0.0458	0.0384
$\Delta = -0.5$	0.1247	0.1081
$\Delta = -1.0$	0.2613	0.2348
$\Delta = -2.0$	0.6036	0.5785
(2) $\Delta = 0.0$	0.1247	0.1081
$\Delta = -0.5$	0.2689	0.2408
$\Delta = -1.0$	0.4590	0.4240
$\Delta = -2.0$	0.7901	0.7681

(1) If $p_{1j}=0.25, j=1,...,4$
(2) If $p_{11}=0.17, p_{12}=0.21, p_{13}=0.27, p_{14}=0.35$

6.6 CONCLUSIONS AND RECOMMENDATIONS

In this chapter the conditional and unconditional approach of calculating p-values for the Wilcoxon-Mann-Whitney statistic have been compared with respect to nominal and actual p-values. It appears that the conditional approach is comparable to the conservative unconditional approach. One drawback of the unconditional approach, viz. its dependence on the underlying multinomial distribution, is outweighed by its advantages: it is independent of the number of ties and permits the use of conservative critical value tables.

The overall recommendation should be that if the underlying multinomial assumption is considered met, the unconditional approach should always be used. If the multinomial null-parameters are known *a priori*, then the conservative critical values can be replaced with exact critical values. This might be related to the use of different weights for specific categories in the conditional Wilcoxon statistic.

6.7 REFERENCES

Basu D (1977). On the elimination of nuisance parameters. *J. Am. Statist. Ass.* 72, 355.

Collings BC and Hamilton MA (1988). Estimating the power of the two-sample Wilcoxon test for location shift. *Biometrics* 44, 847-860.

Emerson JD and Moses LE (1985). A note on the Wilcoxon-Mann-Whitney test for 2xk ordered tables. *Biometrics* 41, 303-309.

Halperin M, Gilbert PR and Lachin JM (1987). Distribution-free confidence intervals for Pr(X1<X2). *Biometrics* 43, 71-80.

Halperin M, Hamdy MI and Thall PF (1989). Distribution-free confidence intervals for a parameter of Wilcoxon-Mann-Whitney type for ordered categories and progressive censoring. *Biometrics* 45, 509-521.

Lehmann EL (1975). *Nonparametrics: statistical methods based on ranks.* San Francisco: Holden-Day.

McCullagh P (1980). Regression models for ordinal data. *J.R.Statist.Soc.B* 42, 109-142.

Mehta CR, Patel NR and Tsiatis AA (1984). Exact significance testing to establish treatment equivalence with ordered categorical data. *Biometrics* 40, 819-825.

Rice WR (1988). A new probability model for determining exact p-values for 2x2 contingency tables when comparing binomial proportions. *Biometrics* 44, 1-22.

Suissa S and Shuster JJ (1985). Exact unconditional sample sizes for the 2x2 binomial trial. *J. R. Statist. Soc. A* 148, 317-327.

Verbeek A and Kroonenberg PM (1985). A survey of algorithms for exact distributions of test statistics in r x c contingency tables with fixed margins. *Comp. Stat. and Data Anal.* 3, 159-185.

Whitehead J (1983). *The design and analysis of sequential clinical trials.* Chicester: Ellis Horwood.

CHAPTER 7

PARTICIPANTS AND NON-PARTICIPANTS IN A TRIAL OF A BACK SCHOOL EDUCATION PROGRAMME *

SUMMARY

Two groups of persons with a recent history of back complaints have been invited to participate in a trial, in order to evaluate the effectiveness of a back school education programme, held outside working hours and intended as a secondary prevention. The first group (A) had been on sickness leave in connection with back complaints. The second group (B) had stated to have frequent back complaints in a questionnaire at their periodic health examination. The present chapter assesses the degree of association of several variables with the willingness to participate in this trial.

The overall attendance figure for the trial was 43 per cent. Different determinants of willingness to participate were found to be active at the four different stages of the intake process. Relevant determinants for non-participation were high body mass index, low educational level, smoking and short duration of the latest absenteeism. Personal reasons mentioned by employees for non-participation indicate that non-participants are not troubled by back pain so seriously or have developed a method to deal with it. On the other hand, participants have more complaints or no method to deal with them; they are hence more willing to invest time in the programme. Because categories of participants and non-participants appear not to be quite similar, persons from the non-participant categories, too, have to be followed over time, in order to be able to generalize the outcome of the trial.

* After: Lugtenburg D, Wiemer GR, Drooger W and Rensink HJT (submitted) *Spine*

7.1 INTRODUCTION

Between October 1, 1990 and April 1, 1991, 439 male employees of Shell Pernis were asked to participate in a trial on a back school education programme. Back school education programmes had previously been used with either acute[1,2] or chronic[3-5] low back pain patients. The Pernis back school education programme, however, was not intended as therapy, but as a secondary prevention for normally working persons with a (recent) history of back complaints. In the overall evaluation of this trial, non-participation has to be taken into account, especially as a secondary prevention trial is bound to have a participation figure far below 100 per cent.

Persons who run a high risk of developing specific diseases are of greatest benefit to health-promotion programmes. Therefore, studies have been conducted to see whether these persons participate.[6,7] Personal reasons for not participating[8] and also more objective "baseline" data for participants and non-participants have been studied, the latter collected after[9] or before[10] the moment of invitation. The determinants of non-participation in population-based studies, such as higher age[10] and lower socioeconomic status/education[10], are not found consistently in work-site studies, except for the higher prevalence of smoking, which is a determinant in both kinds of studies.[9,11] These findings might explain results from follow-up studies of participants and non-participants, both in population-based surveys[11] and in work-site medical surveillance programs[12], from which it is known that non-participants, in general, are less healthy than participants.

A health-promotion programme and a secondary prevention trial differ in that for the latter, the proposed programme has as yet to be evaluated for effectiveness. Hence, determinants of participation will differ and participation in either of the two settings has to be studied separately, preferably in a prospective setting to reduce recall-bias. A trial also makes it possible to identify different levels of non-participation. Determinants of non-participation must be assessed separately for each category of non-participants.[13]

Our study was designed to gather information prospectively with respect to variables concerning sickness absence and periodic health assessments from employees invited to participate in a trial of a back school education programme, which was held after working hours. Our study included participants and non-participants of this trial, so that an unbiased assessment of several determinants in the different phases of participation was feasible. Facts about health habits, health status and demographics are relevant for describing the subpopulation studied and also make possible a discussion of the external validity or generalizability of the trial.

7.2 METHODS

7.2.1 Population

At Shell Pernis, about 4500 persons, predominantly male, work in a petrochemical complex consisting of a refinery and a number of chemical plants. Absenteeism due to sickness is reported to the department of occupational health. At the end of a sickness period both its duration and the disease code (following the WHO ICD-codes of 1975), are stored in a computer. The department of occupational health also conducts on a regular basis periodic health assessments. Non-operational employees are screened every four years, operators and maintenance people every two years. The periodic examination comprises the measurement of the characteristics body weight, body length and diastolic and systolic blood pressure. Further, a questionnaire is completed concerning smoking habits, alcohol consumption and sporting activities. Also, questions about general health are answered, one of them being whether the person has frequent back complaints.

7.2.2 Study population

Between October 1, 1990 and April 1, 1991, a number of male employees were invited to participate in the trial; they were selected on either of two characteristics. One group (group A) consisted of persons reporting back in the week before invitation, after a sickness absence period due to back complaints. The other group (group B) consisted of persons stating that they had frequent back complaints at a periodic health examination one week before invitation. The number of persons invited varied between 2 and 15 per week.

7.2.3 Intake procedure

Persons from the study population were invited by means of a letter which provided extensive information. They were asked to make an intake appointment with a physician at the department of occupational health. If, after two weeks, no appointment had been made, they received a second invitation with identical information. The persons who, after this second letter, still did not respond, formed response category *I*. The persons who responded, but stated that they had no interest, formed response category *II*. The remaining persons made an intake appointment. During this intake appointment a physical examination was conducted by the back school physiotherapist and inclusion and exclusion criteria were applied by the physician. Persons whose complaints could not be defined as caused by *low* back pain were excluded. Medical contra-indications were Bechterew, herniated disk, steroid therapy, anticoagulant therapy, (beginning) hiparthrosis, malignancies and severe psychological problems. Also, persons who were unable to read or understand Dutch and persons who knew *a priori* that they would be unable to attend all lessons, were excluded. So, response category *III* consisted of persons who were excluded from the trial because of one of the exclusion criteria. The remaining persons were asked to participate in the trial. Persons who, for various (personal) reasons, did not participate after all, formed response category *IV*, and persons who agreed and participated formed response category *V*. Figure 7-1 presents a schematic diagram explaining this intake procedure and the results obtained.

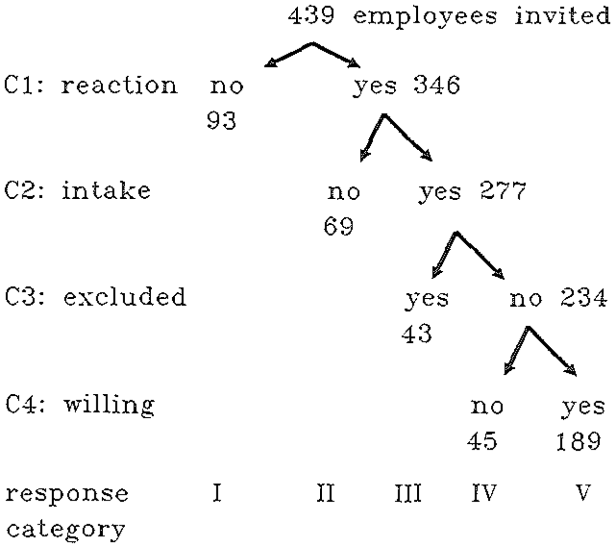


Figure 7-1 Schematic representation of the intake procedure followed in the Pernis back school education trial.

7.2.4 The trial

In the trial, participants were randomized into two groups: a waiting-list group and a group that attended the back school education programme in subgroups of eight persons. The programme consisted of eight sessions, in which information was given on physical and psychological factors related to back complaints. Furthermore, a number of exercises for the back were instructed according to the principles of McKenzie.[14] After a one-year follow-up period, the possible effects of the back school education programme on, among other things, absenteeism will be assessed.

7.2.5 Baseline variables considered to be potential determinants

At each of the four consecutive stages of intake (cf. Figure 7-1), the categories of persons from the study population willing to participate were compared with the category of persons who were not.

For every individual, age, work-to-residence distance, registered car pooling (yes/no), shift work (yes/no), kind of department (chemical, refinery, technical/maintenance and remaining) and educational level (low, medium, high) were known. With respect to educational level, 'low' means that a person only had primary education, 'high' means a university degree or a higher vocational education level and 'medium' means all intermediate educational levels. Since different determinants of participation may operate within groups A and B comparisons between categories were made separately for A and B. In group A the influence of the number of days of the latest period of sickness leave was studied. In group B the influence of the health characteristics diastolic blood pressure, systolic blood pressure, body mass index (weight/height^2), smoking habits (smoker: yes/no), alcohol intake (less than or at least 3 drinks a week), sports (yes/no) and medication (yes/no) were studied.

7.2.6 Statistical methods

In the sequence of comparisons, the intake procedure was followed closely, see also Figure 7-1. In the first comparison, *C1*, persons who did not react (category *I*) were compared with those who did (categories *II*, *III*, *IV* and *V*). In the second comparison, *C2*, persons who did not want to make an intake appointment (category *II*) were compared with those who did (categories *III*, *IV* and *V*). In *C3*, persons who were excluded from the trial (category *III*) were compared with those who were not (categories *IV* and *V*). In *C4*, persons who did not want to participate (category *IV*) were compared with those who did (category *V*).

For each comparison (*C1* to *C4*), univariate analyses were done to assess the influences of baseline variables on the willingness to participate. The Wilcoxon-Mann-Whitney test[15] was used to compare categories with respect to the mean value of continuous baseline variables, viz. age and distance to residence (for both groups A and B), period of absenteeism (group A only) and diastolic blood pressure, systolic blood pressure and body mass index (group B only). Chi-square tests[15] were used to compare categories with respect to discrete baseline variables, viz. kind of department and educational level. Fisher's exact test[15] was used to compare categories with respect to dichotomous variables, viz. registered car pooling and shift work, and, for group B only, smoking, drinking, sports and medication.

On the basis of the results of the univariate analyses, logistic regression analysis[16] was applied to evaluate simultaneously all potential determinants of participation for every individual comparison, *C1* to *C4*. Since the response categories are mutually exclusive and complete, the response variable can be considered to follow a multinomial distribution. Using logistic regression analysis for each comparison, however, implies that the outcome variable for each comparison follows a binomial distribution. But, given the comparison sequence as discussed above, the likelihood, when optimized, relates to the multinomial distribution.[17] By this approach the probability of a person not further participating in the intake procedure was modelled at every stage. The resulting probabilities can straightforwardly be compared between two subgroups using the odds ratio. If we denote the probability that a person from a particular subgroup *S1* ends participation at a certain stage p_1 and the probability that another person from a subgroup *S2* ends participation p_2, the odds ratio of ending participation for subgroup *S1* compared to subgroup *S2* is defined as $p_1/(1-p_1)$ divided by $p_2/(1-p_2)$. Odds ratios were estimated with corresponding confidence intervals (CI) for all relevant determinants using a logistic regression model.

7.3 RESULTS

7.3.1 Study population

A total of 439 persons were invited to participate in the trial; group A, which consisted of 207 persons reporting back in the week before invitation after a period of sickness absence due to back complaints, and group B, which consisted of 232 persons stating to have frequent back complaints at a periodic health examination one week before invitation. Baseline characteristics are summarized in Table 7-I. No substantial differences between A and B are observed, except for the distribution of departments (which is fully explained by the agenda according to which departments were invited for their periodic health examination) and the distribution of educational level (persons with higher education have less sickness absence due to back complaints).

7.3.2 Univariate comparisons

Baseline variables with a statistically significant effect at the 5 per cent level (or of borderline significance) on willingness to participate in either group (A, B) at some stage in the univariate analysis, are listed in Table 7-II. Variables with no significant effect were not included in this table.

Table 7-II shows that in group A, 19 % did not react at all at stage 1 and in group B 23 % (comparison *C1*). A comparison between workers from the various departments showed that for group A 32% of the employees in chemical departments did not react, compared to 14% of the employees in other departments and for group B 37% compared to 18%.

Of those reacting, 20 % were not interested in the trial in both groups (comparison *C2*). Workers in a chemical plant (those who gave a reaction) had a lower percentage of non-interest: 9% compared to 24% in the remaining departments in group A and 8% compared to 23% in group B. Duration of latest absenteeism was another determinant for non-interest in group A. In group B, low alcohol consumption and low body mass index were also indicative of unwillingness to participate.

At the intake appointment, 16 % were excluded in group A and 15 % in group B (comparison *C3*), low education being a determinant for exclusion in both groups, 55% compared to 12% in group A and 45% compared to 13% in group B. Some variables indicative of exclusion in the univariate analysis only were age and duration of latest absenteeism in group A and medication in group B.

At the final stage 18 % in group A refused to participate and 20 % in group B (comparison *C4*). Only in group B was a difference observed between workers with low educational level (67 %) and persons with higher education (18 %). In group B, a high body mass index was found to be indicative of unwillingness to cooperate at the final stage.

TABLE 7-I

Baseline characteristics for subjects in both groups (A: sickness absence, B: frequent complaints) who received written invitations to participate in a back school education trial

	Group A (n=207)	Group B (n=232)
Age in years: mean (sd)	42.4 (9.5)	42.5 (9.2)
Distance to residence in km: mean (range)	18.0 (2-97)	20.1 (2-101)
Registered car poolers (%)	12.6	12.1
Shift workers (%)	61.4	55.4
Department (%)		
chemical	30.4	25.4
refinery	21.7	11.2
technical/maintenance	31.9	47.4
remaining	16.0	15.9
Education (%)		
low	9.2	8.6
medium	83.6	76.7
high	7.2	14.7
Period of latest absenteeism in days: mean (range)	12.0 (1-88)	*
Diastolic blood pressure in mm Hg: mean (sd)	*	81.4 (10.2)
Systolic blood pressure in mm Hg: mean (sd)	*	133.4 (14.8)
Body mass index in kg/m^2: mean (sd)	*	25.2 (3.1)
Smokers (%)	*	36.2
Alcohol intake < 3 units/week (%)	*	35.3
Sports (%)	*	40.5
Any medication (%)	*	15.5

*: not determined for that specific group

TABLE 7-II

Variables considered to be potential determinants for unwillingness to participate. Percentages indicative of unwillingness are shown by variable and by intake stage. Significance levels are based on the ungrouped baseline data

Comparison at stage	*C1*		*C2*		*C3*		*C4*	
Group	A	B	A	B	A	B	A	B
Overall percentages	19	23	20	20	16	15	18	20
Age:								
below 40 years	21	26	25	21	6†	11	16	21
at least 40 years	18	21	17	19	22	18	19	20
Education:								
low	32	40	15	8	55‡	45‡	20	67‡
higher	18	21	21	20	12	13	18	18
Chemical department:								
yes	32‡	37‡	9‡	8‡	13	15	21	14
no	14	18	24	23	17	15	17	23
Duration of absenteeism:								
less than 1 week	20	*	28‡	*	11†	*	21	*
at least 1 week	19	*	14	*	19	*	16	*
Body mass index								
below 25 kg/m^2	*	19	*	24†	*	17	*	10‡
at least 25 kg/m^2	*	27	*	14	*	14	*	30
Alcohol intake								
low	*	27	*	28†	*	19	*	17
not low	*	21	*	15	*	14	*	22
Medication								
yes	*	31	*	24	*	32†	*	31
no	*	21	*	19	*	13	*	19
Smoking								
yes	*	30¶	*	20	*	21	*	22
no	*	19	*	19	*	12	*	20

* :	not determined for that specific group
C1:	no reaction vs reaction
C2:	no interest vs interest
C3:	excluded vs not excluded
C4:	not willing to cooperate vs willing to cooperate
† :	statistically significant difference between percentages concerned ($p \leq 0.05$), only by univariate analysis, uncorrected for confounding variables
‡ :	statistically significant difference between percentages concerned, also by multivariate analysis, corrected for confounding variables
¶ :	statistically significant difference between percentages concerned, by multivariate analysis, borderline significance by univariate analysis

7.3.3 Simultaneous assessments of potential determinants

Generally speaking, a univariate analysis can give only a preliminary result as to which variable (out of a number of more or less correlated variables) may be a potential determinant. Since various comparisons showed more than one determinant in the univariate approach, it was obvious that a thorough analysis was needed to see whether some of these effects could be ascribed to a "confounding" variable. Therefore, logistic regression analyses were performed to take account of such influences.

With respect to initial reaction (*C1*), working in a chemical department was found to be a determinant. The odds ratio estimate for not reacting in group A: 2.88 with a 95% CI (1.42, 5.87) and in group B, 2.83 with a 95% CI (1.46, 5.52). In group B, smoking was another independent determinant for not reacting. The odds ratio estimate for smokers: 1.92 with a 95% CI (1.01, 3.65).

With regard to the intake procedure (*C2*) for those reacting at the first stage, not working in a chemical plant was found to be a determinant for non-interest. The odds ratio for non-interest in group A is 0.32 with a 95% CI (0.10, 0.96) and in group B 0.30 with a 95% CI (0.10, 0.92). In group A duration of the latest absenteeism is another independent determinant for non-interest, with the odds ratio estimate for persons with a period of latest absenteeism less than 1 week being 2.40 with a 95 % CI (1.10, 5.26).

With respect to exclusions (*C3*), only one independent determinant remained as indicative of exlusion from the trial, viz. low education. The odds ratio estimate for exclusion of persons with a low education in group A: 8.6 with a 95% CI (2.32, 31.5) and in group B: 5.3 with a 95% CI (1.64, 17.4).

With respect to the final stage (*C4*) in group B the two independent determinants for unwillingness were high body mass index, with odds ratio estimate: 3.4; CI (1.23, 9.09), and low educational level, with an estimated odds ratio of 6.93 and 95% CI (1.13, 42.5).

7.3.4 Reasons of invited individual persons for not participating

At each stage of the intake process, data were collected about reasons mentioned for not participating any further. As these data concern personal opinions, collected only from non-participants, they cannot be used in assessing differences between participants and non-participants. However, they may help to find out whether the reasons given correlate with the actual comparisons of baseline data in Table 7-II.

The reasons given for non-interest in the trial (*C2*) were as follows: not having *low* back complaints (n=7), back complaints due to a cold (n=9), in therapy with a specialist (n=14), back complaints not serious or already doing exercise (n=26). The remaining 13 persons stated other reasons.

Reasons for exclusions according to criteria written down in the study protocol (*C3*) were as follows: not understanding or not reading Dutch correctly (n=6), not having *low* back complaints (n=9), other specific objective back complaints like Bechterew and herniated disk (n=11), serious psychological problems (n=2), not working at own department due to incapability (n=8), excluded for other reasons (n=7).

Personal reasons given by the 45 persons who did not cooperate in the final stage (*C4*) were as follows: would have participated if problems concerning transportation could have been solved (n=20), not liking questionnaires (n=5), low expectations (n=6), other personal problems (n=14).

7.4 DISCUSSION

In this study, non-participation in a trial on a back school education programme is evaluated, following the intake process step by step. Altogether 439 persons were invited, of which 43 per cent ultimately participated. The distinction between two groups, A and B, pertains to the fact that sickness absence characteristics were known for group A and periodic health characteristics were known for group B. However, the two groups have in common that they consist of persons with low back pain with quite similar baseline characteristics (cf. Table 7-I). Therefore in this discussion, the distinction between the two groups will not be mentioned. Univariate analyses may produce results that are due to "confounding" by other variables; in multivariate analysis, based on a logistic model, the problem of confounding is circumvented. Therefore, results from logistic regression analyses are given more weight.

An intriguing finding was that persons working in chemical departments were less willing to participate initially than persons working in other departments, in the sense that they were less inclined to react to the invitation (*C1*); the odds of not responding are about three times higher than in other categories. However, this effect was clearly compensated for in the next stage (*C2*), where persons working in a chemical plant were more willing to make an intake appointment than other employees. Persons from chemical plants were not found to be different from persons from other departments with respect to mean age and mean distance from work to residence. As might be expected, the percentage shift workers was higher in chemical plants, but a more refined analysis of participation within both groups (chemical plants and other departments) showed that shift work was no determinant of non-participation within these groups. Therefore, the observed difference in the percentage shift workers between the two groups provides no explanation for the difference of non-participation between them. A similar analysis of the observed differences in the distribution of educational levels between the two groups did not provide any explanation either. Overall, there is apparently no difference between departments in willingness to make an appointment for the intake procedure. However, people from different departments may show their non-interest in different ways: either by not reacting at all, like the workers in a chemical plant in this study, or by reacting through declaring explicitly to have no interest in such a trial, like other workers in our investigation.

Short duration of recent absence because of low back complaints turned out to be a determinant for non-participation. The participants with a relatively long sickness absence considered the trial more attractive due to their more serious back complaints. Some confirmation was found from individual reasons mentioned by 69 non-participants, where "few complaints" was the main motivator not to participate.

On the basis of these results we would assume that there is an expected-cost-benefit ratio for willingness to participate in such a trial. The expected cost is the personal investment in time outside working hours, the expectation of benefit being based on the individual health status and the seriousness of the recent complaint. On this assumption we may expect a fairly high percentage of participation in back school education programmes from low back patients without any current therapy, against a low percentage in primary prevention programmes. The percentage participation in our secondary prevention programme should be somewhere in between. This is confirmed by a comparison of the attendance figure after the first two stages in our study (63 %) with the attendance at a general fitness programme, where voluntary participation outside scheduled work time was found to be low (28 %)[16] and with the attendance at a back school education programme with a directly available consultant at the department of occupational health for persons with acute back problems, where it was high (100 %)[1]. It seems important to note that (secondary) prevention should follow actual back complaint periods very closely to get a high attendance and that severity of back complaints seems a relevant trigger for participation. It could be argued that the seriousness of a complaint is felt less after an elapsed period of time. It may be interesting to investigate whether participation increases with introducing the back school trial at the time when a person reports back after sickness absence or when completing the questionnaire.

In the third stage, *C3*, 16 per cent of persons were excluded, mainly related to low education, partly because these people were unable to read and understand Dutch. If we exclude the six persons involved from our analysis, low educational level remains a determinant, although statistically significant only for group A. Our conclusion is that the remaining exclusion criteria do not favour the inclusion of persons with low education.

It is interesting to note that body mass index, being a known determinant of willingness to enter a study[10] turned out to be a determinant in this study too, although only in the last stage after admission (*C4*). At this stage, however, low educational level was a determinant for non-participation only in group B and not at all in group A. It may be that different mechanisms are operating in the two groups.

The reasons mentioned at the final stage for non-participation, for example car pooling and distance from work to residence, are not found as determinants in the analysis of the baseline data. It might therefore be speculated that these reasons are some sort of excuse for not wishing to participate in persons of low motivation.

In none of the four stages does shift work appear to have any effect on non-response. At the time of design of the trial, there was some scepticism about the "degree of inconvenience" involved for both shift- and day-workers. Consequently, the programme was planned after working hours, so that inconvenience would be the same for both groups. Since no effect of shift work on participation was found, this problem seems to have been addressed adequately.

Concerning the personal reasons mentioned for non-participation, they seem to indicate that non-participants do not have serious complaints or have developed a method to deal with them. Presumably, participants are persons who either have more serious complaints or have no method to deal with them, and are therefore more willing to invest the time involved. Although until now the back school education programme is offered outside working hours, it might be considered recommendable to have this changed to within working hours in order to increase motivation to participate, at all stages. Therefore, the various groups of non-participants will be followed with respect to the main outcome used in the trial and be compared with the control group in the trial in order to advise management on the desirability to present the programme during working hours.

7.5 REFERENCES

1 Bergquist-Ullman M and Larsson U. Acute low back pain in industry. *Acta Orthop. Scand. Suppl. 170.* 1977.

2 Lindequist S, Lundberg B, Wikmar R et al. Information and regime at low back pain *Scand. J. Rehab. Med.* 1984; 16: 113-116.

3 Hurri H. The Swedish back school in chronic low back pain. Part I: Benefits. *Scand. J. Rehab. Med.* 1989; 21: 33-40.

4 Hurri H. The Swedish back school in chronic low back pain. Part II: Factors predicting the outcome. *Scand. J. Rehab. Med.* 1989; 21: 41-44.

5 Klaber Moffett JA, Chase SM, Portek I and Ennis JR. A controlled, prospective study to evaluate the effectiveness of a back school in the relief of chronic low back pain. *Spine* 1986; 2: 120-122.

6 Stange KC, Strogatz D, Schoenbach VJ, Shy C, Dalton B and Cross AW. Demographic and health characteristics of participants and nonparticipants in a work site health-promotion program. *J. Occup. Med.* 1991; 33: 474-478.

7 Stange KC, Strecher VJ, Schoenbach VJ, Strogatz D, Dalton B and Cross AW. Psychosocial predictors of participation in a work site health-promotion program. *J. Occup. Med.* 1991; 33: 479-485.

8 Vernon SW, Acquavella JF, Yarborough CM, Hughes JI and Thar WE. Reasons for participation and nonparticipation in a colorectal cancer screening programme for a cohort of high risk polypropylene workers. *J. Occup. Med.* 1990; 32: 46-51.

9 Settergren SK, Wilbur CS, Tyler DH and Rassweiler JH. Comparison of respondents and nonrespondents to a worksite health screen. *J. Occup. Med.* 1983; 25: 475-480.

10 Sonne-Holm S, Sorenson TIA, Jensen G and Schnahr P. Influence of fatness, intelligence, education and sociodemographic factors on response rate in a health survey. *J. Epidem. Comm. Health* 1989; 43: 369-374.

11 Criqui MH, Barrett-Connor E and Austin M. Differences between respondents and non-respondents in a population-based cardiovascular disease study. *Am. J. Epidem.* 1978; 108: 367-372.

12 Bond GB, Lipps TE, Stafford BA and Cook RR. A comparison of cause-specific mortality among participants and nonparticipants in a work-site medical surveillance program. *J. Occup. Med.* 1991; 33: 677-680.

13 Ohlson CG and Ydreborg B. Participants and non-participants of different categories in a health survey: a cross-sectional register study. *Scand. J. Soc. Med.* 1985; 13: 67-74.

14 McKenzie R. *Treat your own back.* Spinal publications 1980. Wright and Carman Limited, New Zealand.

15 Siegel S and Castellan NJ. *Nonparametric statistics for the behavioral sciences.* Scnd. edition 1988. McGraw-Hill, New York.

16 Walker SH, Duncan DB. Estimation of the probability of an event as a function of several independent variables. *Biometrika* 1967; 54: 167-179.

17 McCullagh P and Nelder JA. *Generalized linear models.* Chapman and Hall, London, 1983.

18 Lynch WD, Golaszewski TJ, Clearie AF et al. Impact of a facility-based corporate fitness programme on the number of absences from work due to illness. *J. Occup. Med.* 1990; 32: 9-12.

CHAPTER 8

A STATISTICAL METHOD FOR EVALUATING GAMMA-GLUTAMYLTRANSFERASE IN OCCUPATIONAL HEALTH SCREENING *

SUMMARY

Data on 5222 male employees in a refinery and petrochemical complex have been used in a cross-sectional study to identify determinants influencing the levels of serum γ-glutamyltransferase (GGT). Multiple regression analysis showed that GGT levels were associated, in order of importance, with body mass index, age, alcohol consumption, cigarette smoking, medication and physical activity. On the basis of the person-related determinants a model was developed, which was used to study the influence of work-related variables (e.g. chemical exposure) on GGT. None of the work-related variables was found to be of influence. In addition, the model was used to predict employees' individual GGT values and to compare these with the laboratory-determined values. These comparisons were used in defining reference intervals which may be applied in the identification of possible hepatocellular effects. We concluded that persons with GGT levels below 30 U/l are unlikely to have liver impairment and do not require additional testing. For persons with GGT levels above 60 U/l additional testing of other liver enzymes is indicated. For persons who have observed GGT levels between 30 and 60 U/l the need for additional testing depends on the personal characteristics mentioned before.

* After: De Koningh AGJ, Lugtenburg D and Van Sittert NJ (submitted) *Clinical Chemistry*

8.1 INTRODUCTION

Gamma-glutamyltransferase (GGT; EC 2.3.2.2) is a sensitive indicator of various types of hepatobiliary diseases.[1-3] Several determinants are known to influence the concentration of GGT in serum of human populations. The most important ones are age and sex[4-7], body mass index[5,7], alcohol consumption[5,8-10] and medication[1,5,11,12]. The measurement of serum GGT levels in industrial workers potentially exposed to hepatotoxic chemicals is generally considered to be a sensitive test for monitoring hepatocellular effects.[2,13,14] However, the inter-person variation of GGT levels will be influenced by the above-mentioned person-related determinants. In a study of possible association between exposure to chemicals and serum GGT activities, it is essential to adjust for these determinants.[2] This is also necessary for the detection of liver impairment, particularly if GGT is marginally increased.

The main aim of the cross-sectional study reported here was to find out whether we could confirm the association between the above-mentioned determinants and GGT levels in a large industrial population. On the basis of these determinants a statistical model was developed, which allowed the prediction of an individual's theoretical GGT level and its tolerance interval. This model was then used to study the influence of work-related factors, e.g. chemical exposure, on GGT. In addition, the model was used to compare predicted values and laboratory-determined values, after which observed differences were studied to find out whether they were statistically significant as indicators of possible hepatocellular effects.

8.2 MATERIALS AND METHODS

8.2.1 Description of the study population

A total number of 6310 men are registered as having been employed at Shell Pernis for some time between 1 April 1985 and 1 October 1989. Of this group 5222 persons attended a voluntary Periodic Health Assessment (PHA) or an obligatory Pre-Employment Health Assessment (PEHA). These men formed our study population. The frequency of the PHA depends on job-specific hazards (e.g. employment in chemical operations, in refinery operations or in administrative jobs) and varies from yearly to once every 4 years. This was the rationale for choosing an observation period of 4.5 years. Employees who started employment during this period had at least had an obligatory PEHA.

During the PHA and PEHA every person had been requested to complete a questionnaire with regard to health experience, working conditions and lifestyle habits; biochemical and hematological substances in blood had been measured, as well as several physical parameters (body weight, height). The personal and health assessment data had been stored in an HP 3000 computer. Health assessment data of the first PHA were used for the present study, except when these were not available, in which case data of the PEHA were used.

8.2.2 Variables studied

The following person-related variables were considered in the statistical analysis: age (years); body mass index (kg/m^2); body weight (kg); score alcohol consumption (SAC) graded 0-4, i.e. 0, 1-2, 3-10, 11-25, >25 drinks a week; score smoking habits (SSH) graded 0-4, i.e. 0, 1-5, 6-10, 11-20, >20 cigarettes a day; medication (no/yes); physical activity during leisure time (no/yes); educational level, i.e., low, medium, medium/high and high education.

The work-related variables considered in the analysis were: duration of employment (years); intake procedure (PHA or PEHA); exposure category, i.e. employment in chemical operations, in refinery operations or in an administrative job; shift work (yes/no); individual perception of working conditions, i.e. complaints of exposure to dust, smoke or vapour (yes/no).

8.2.3 Specimen collection and GGT analysis

Blood specimens had been collected between 08.00 and 10.00 hours from subjects who had not been asked to fast. Blood had been taken from the antecubital vein with the subject seated. Tourniquet pressure had been released immediately on venepuncture using the vacutainer system. About 15 ml of blood had been collected in tubes for preparation of serum for the measurements of GGT and other blood substances. The serum samples had been kept at 4°C and analysed within 24 hours. No losses of activity of GGT had been observed when serum had been stored under these conditions. GGT had been determined with a Hitachi 705 autoanalyser (Boehringer, Mannheim) at 30°C, using L-γ-glutamyl-3-carboxy-4-nitroanilide (2.9 mmol/l) as a substrate.

The precision and accuracy of the GGT measurements over the whole study period was estimated by participating in quality assessment programmes. The day-to-day precision over the whole observation period, expressed as coefficient of variation, was on average 2.4%. It was shown that with our method GGT values were on average 13% lower than with the method recommended by the International Federation of Clinical Chemistry[15]. In another study it was shown that neither precision nor accuracy are dependent on GGT level.[16]

8.2.4 Statistical methods

The influence of work-related variables (x_j) on serum GGT levels, after correction for person-related variables (z_j), was analysed through fitting a linear model. In this model, the logarithm of GGT is tentatively considered to be a linear combination of m person-related variables, (p-m) work-related variables, and a random error term:

$$\ln(GGT) = b_1 z_1 + \ldots\ldots b_m z_m + b_{m+1} x_1 + \ldots. b_p x_{p-m} + \text{error } \epsilon.$$

The error term is assumed to be distributed as $N(0,\sigma^2)$. In a sequence of steps the most relevant main effects and interactions of person-related variables were identified as determinants of GGT, using multiple regression analysis. When the variables were clearly categorical, they were entered into the regression equation by means of dummy variables, which have only two possible values: 1 or 0.

The following method was applied for selecting relevant person-related variables. Starting with the most simple model, ignoring any explanatory variable, we successively considered variables for inclusion in the model, in decreasing order of relevance, using the information from the existing literature[4-12]. Any term was included provided the null hypothesis that the parameter b_j of this extra term z_j equals zero was rejected (F-test). This F-test is based on the difference between two Error Sums of Squares (ESS), i.e. the sum of the squared residuals, calculated for a model including this extra term and for a model excluding it. The corresponding p-value is the probability that the observed reduction in ESS, which results when this extra term has been entered into the model, is due solely to chance. In view of the large number of observations, only variables with p-values less than 0.01 were included in the model. The appropriateness of a model (i.e. its goodness of fit) is characterized by the coefficient of determination, R^2, computed from (SuS-ESS)/(SuS), where SuS is the sum of squares of observed values of ln(GGT). In other words, R^2 equals the proportion of variance of ln (GGT) explained by the model.

In a first step the variables age, body mass index, body weight and alcohol consumption were considered for inclusion in a basic model. Body mass index was included because it had previously been demonstrated that GGT increases with body mass index.[5,7] Age was fitted by a linear as well as a quadratic component because several authors have shown that GGT levels increase up to 50 years and decrease thereafter.[5,18,20]

In subsequent steps the remaining person-related variables were included, one at a time, entering their main effect and their interactions with age, body mass index and alcohol consumption in the model. Various models were fitted this way and their goodness of fit was evaluated. On the basis of this procedure a final model resulted, which incorporated all the relevant person-related variables (reduced PRV model).

The plausibility of model assumptions was evaluated by residuals analysis. In addition, the presence of "influential points", i.e. observations that influence a regression coefficient, was investigated. Whether the F-tests applied in evaluating various models lacked power because of limited diversity in the sample space, was judged through performing collinearity calculations.[17]

The model obtained was then used to calculate a tolerance interval around the predicted value for every person, given his person-related variables only, as:

$$[\text{predicted ln GGT-2.sd, predicted ln GGT+2.sd}] \tag{1}$$

in which sd is an estimator for the residual standard deviation, calculated as the square root of the Mean Square Error, defined as the mean of the squared residuals. In practice the tolerance interval is transformed back to original units, providing the reference interval of a population with identical characteristics.

The person-related-variables model may then be used as a starting point to assess the additional influence of work-related variables. This was done through multiple regression analysis, to detect any further significant reduction of error sums of squares, by taking each variable in turn.

The person-related-variables model can also be used to calculate an individual's theoretical or predicted GGT level and its tolerance interval. If a person's laboratory-determined value (GGT) falls outside the tolerance interval, say, exceeds the upper limit (UL), this may be taken to be an indication of an effect of an unidentified variable such as liver impairment. By plotting for each individual of the population the difference d (d = UL-GGT) against GGT, persons with this indication are easily identified.

8.3 RESULTS

The study population comprised 5222 men participating, out of a total of 6310 men. The age distributions of the participants and non-participants are shown for comparison in Table 8-I. The participation figures of age groups below 50 years were similar, but the age group above 50 years had a lower participation figure.

TABLE 8-I

Age distribution of participants and non-participants

Age group	16-29	30-39	40-49	50-62	Total
Participants	1200 (87.8%)	1345 (88.1%)	1356 (88.9%)	1321 (69.8%)	5222 (82.7%)
Non-participants	166 (12.2%)	182 (11.9%)	169 (11.1%)	571 (30.2%)	1088 (17.3%)
Total population	1366	1527	1525	1892	6310

The distribution of the GGT values for the 5222 men was skewed to the right (Figure 8-1). The median GGT level was 15 U/l and the 2.5 and 97.5 percentiles were 7 and 58 U/l, respectively.

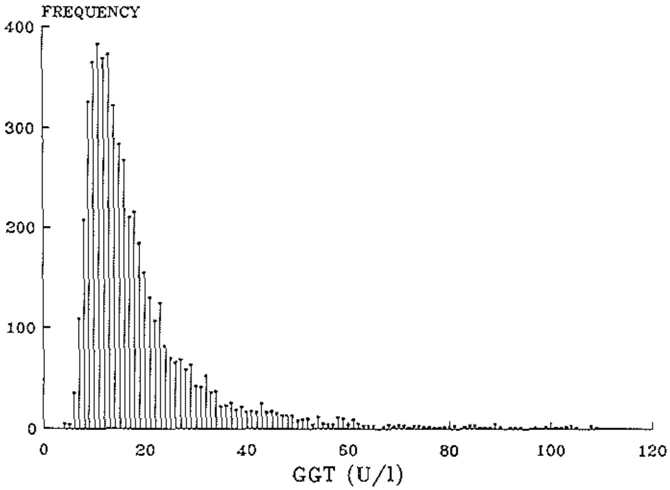


Figure 8-1 The frequency distribution of the GGT values of the study population.

The first part of the present study was directed at identifying person-related variables that influence the level of GGT. The median GGT values at the various levels of variables are presented in Tables 8-IIa and 8-IIb. Because no adjustment for determinants was applied, these tables only give preliminary results. A more refined approach, such as multiple regression analysis, was therefore indicated.

TABLE 8-IIa

Distribution properties of study population (N=5222); subpopulations grouped according to the levels of the person-related variables studied with median GGT values

Variable studied	Levels of variables	Frequency (%)	Median GGT value (U/l)
Age groups (years)	16-29	23	12
	30-39	26	14
	40-49	26	17
	50-62	25	17
Body mass index (kg/m^2)	<20	5	11
	20-22.5	20	12
	22.5-25	34	14
	25-27.5	28	17
	>27.5	14	21
Alcohol Intake	Score Alcohol Consumption (SAC)		
	0="no alcohol"	21	14
	1="1-2 drinks a week"	22	14
	2="3-10 drinks a week"	40	15
	3="11-25 drinks a week"	15	18
	4="more than 25 drinks a week"	2	22
Cigarette smoking	Score Smoking Habits (SSH)		
	0="no cigarettes"	66	15
	1="1-5 cigarettes a day"	5	15
	2="6-10 cigarettes a day"	11	15
	3="11-20 cigarettes a day"	15	16
	4="more than 20 cigarettes a day"	3	18
Medication	yes	12	18
	no	88	15
Physical activity during leisure time	yes	44	14
	no	56	16
Educational level	low	12	18
	medium	33	15
	intermediate	38	14
	high	17	14

TABLE 8-IIb

Distribution properties of study population (N=5222); subpopulations grouped according to the levels of the work-related variables studied with median GGT values

Variable studied	Levels of variables	Frequency (%)	Median GGT value (U/l)
Intake procedure	pre-employment health assessment	3	11
	periodic health assessment	97	15
Duration of employment (years)	0-10	44	13
	11-20	26	16
	21-30	18	17
	>30	13	17
Working schedule	shift work	47	15
	day work	53	15
Exposure category	chemical operations	38	15
	refinery operations	33	15
	administratice staff	29	15
Negative perception of working conditions	yes	2	14
	no	98	15

At this stage, the reduced person-related-variables (PRV) model was derived, following the steps described in section 8.2.4. This model included the variables: age, age^2, body mass index, body weight, alcohol consumption, medication, physical activity, cigarette smoking and the interactions of alcohol intake and body mass index (both linear and quadratic). Analysis of residuals from the fitted model did not reveal signs of inadequacy of the model, except that the distribution showed a skewness to the right. Influential points were not detected. Collinearity was not observed. For all GGT levels below 60 U/l, the variance of the residuals was found to be constant; for levels above 60 U/l this variance increased with GGT level.

For the purpose of estimating a person's ln GGT on the basis of only person-related explanatory variables, the following formula was derived for the PRV model:

$$\text{Predicted ln(GGT)} = 1.05946 + b_1 \text{ body mass index} + b_2 \text{ body weight} + b_3 \text{ age} + \\ + b_4 \text{ age}^2 + b_5 \text{ SAC} + b_6(\text{SAC})^2 + b_7 \text{ SSH} + b_8 \text{ medication} + b_9 \text{ physical activity} + \\ + b_{10} \text{ body mass index.SAC}$$

The regression coefficients (b_1....b_{10}) with their standard errors are shown in Table 8-III. The 'explained variance' was 21.7%. The residual standard deviation equals 0.468.

TABLE 8-III

Regression coefficients, b, with standard errors, s.e., for the reduced PRV-model of ln (GGT)[a)]

	Units	b	s.e.
Intercept		1.06	0.121
Body mass index	kg/m^2	6.0 x 10^{-2}	0.52 x 10^{-2}
Weight	kg	-4.9 x 10^{-3}	1.08 x 10^{-3}
Age	yrs	2.6 x 10^{-2}	0.45 x 10^{-2}
Age^2	yrs^2	-2.6 x 10^{-4}	0.55 x 10^{-4}
SAC [b)]		-2.3 x 10^{-1}	0.55 x 10^{-1}
SAC^2		3.5 x 10^{-2}	0.56 x 10^{-2}
SSH [c)]		2.1 x 10^{-2}	0.54 x 10^{-2}
Medication	1=yes, 0=no	0.11	0.020
Physical activity	1=yes, 0=no	-0.083	0.014
Body mass index times SAC	kg/m^2	8.0 x 10^{-3}	2.16 x 10^{-3}

[a)] $R^2 = 0.217$
MSE (Mean Square Error of residuals) = 0.219

[b)] SAC: Score Alcohol Consumption, graded 0-4, i.e. 0, 1-2, 3-10, 11-25, >25 drinks a week

[c)] SSH: Score Smoking Habits, graded 0-4, i.e. 0, 1-5, 6-10, 11-20, >20 cigarettes a day.

To identify the influence of work-related variables (on a group level) the PRV model was used as a starting point, again using regression analysis. The work-related variables duration of employment (years), intake procedure, exposure category, shift work and perception of working conditions completed the model description by adding to the PRV model each variable as main effect and interactions with age (linear and quadratic), alcohol intake, body mass index (linear and quadratic), medication, cigarette smoking and physical activity, respectively. It should be noted that all the work-related data were not available for every individual person. Only for one variable, "intake (PHA or PEHA)", could all 5222 individual results be used in the analysis. For the remaining work-related variables the number of individual results varied between 5039 and 5052 (170 persons who had a PEHA were unable to provide information on work-related variables, because they were not yet employed at the moment of examination). The result of this approach to the data of our study population was that none of the work-related variables had a significant influence on GGT level.

On an individual basis the use of formula (1) to calculate the tolerance interval around the predicted GGT value makes it possible to judge whether the laboratory-determined GGT value is an indicator for possible hepatocellular effects. For each individual in our study the difference (d) between the GGT level and the calculated UL (upper limit of the tolerance interval) is plotted against GGT in Figure 8-2.

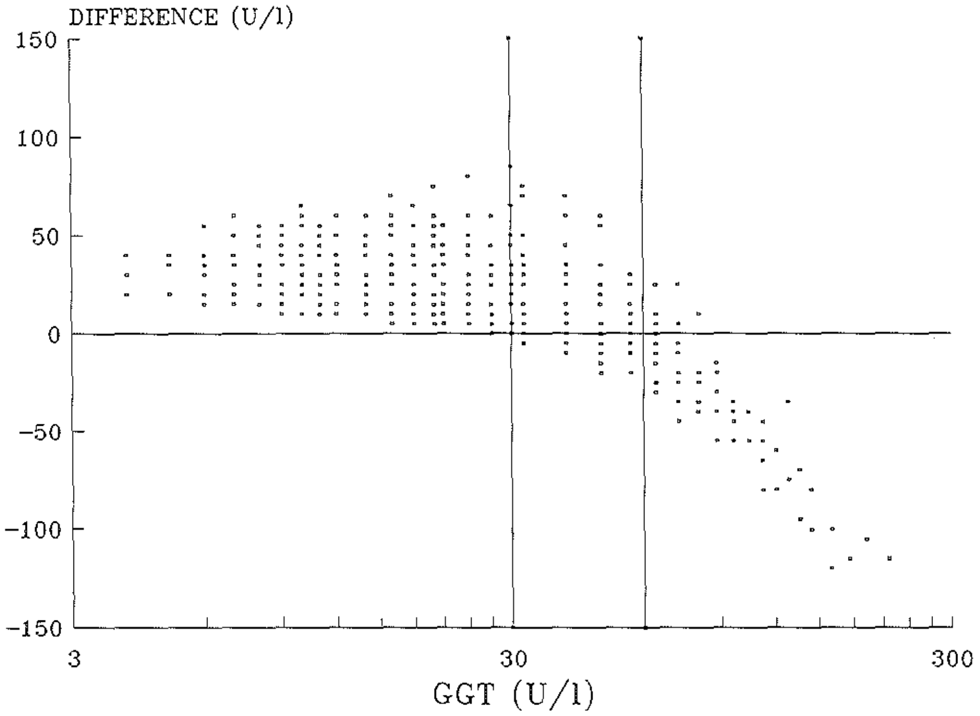


Figure 8-2 The difference d between the laboratory-determined (observed) GGT and the predicted upper limit of the tolerance interval (d=UL-GGT) as a function of the observed GGT.

Figure 8-2 shows that laboratory-determined levels below 30 U/l are nearly always below the corresponding UL, levels above 60 U/l nearly always exceed UL, and the individual UL has to be calculated only at levels between 30 and 60 U/l.

8.4 DISCUSSION

In this cross-sectional study we used data on 5222 men who participated in a voluntary PHA or obligatory PEHA during an observation period of 4.5 years. In a generalization of the conclusions to the total population account should be taken of a possible participation bias. It turned out that fewer persons above 50 years attended the voluntary PHA than persons in the other age groups (70% versus 88%). Therefore, the conclusions are at least valid for the participating 70% and 88% of these age groups. A statistical method has been developed to correct for such a participation bias.[18]

Serum GGT levels in our population were significantly associated with the person-related variables body mass index (and body weight), age, alcohol consumption, cigarette smoking, medication and leisure time physical activity. Altogether, these variables explained approximately 22% of the inter-individual variation of GGT levels, which is similar to the percentage reported for a Norwegian population.[7]

The regression model made it possible to estimate the effect of a variable on the individual GGT, corrected for other variables. The strong positive association between overweight and GGT levels in the present study, adjusted for the other variables in the model, is in agreement with reports of other authors.[5,7,22-24] On the basis of the corresponding estimated model parameters (Table 8-III) it can be concluded that for subjects who do not take alcohol, with similar values of other determinants, an increase of 2 kg/m^2 in body mass index results in about 13% higher predicted GGT levels. With regard to age, predicted GGT levels showed a clear increase to an estimated top at the age of 49 years, and a slow decrease thereafter. This is in agreement with the literature.[5,6,19-21] Persons aged 50 years had 25% higher predicted GGT levels than persons aged 20 years with similar values of other determinants. The effect of alcohol consumption on GGT levels can be illustrated by comparing subjects with SAC=0 and SAC=4, with body mass index 25 kg/m^2 and similar values of other determinants. Heavy drinkers had 58% higher predicted GGT levels than non-drinkers. The association between cigarette smoking and GGT values is not very clearly documented in other studies. In the present study, the predicted GGT value in heavy smokers (SSH=4) was found to be 8.5% higher than in non-smoking men. The positive association between GGT and usage of certain types of medicine, such as the hepatic microsomal enzyme inducers phenytoin, phenobarbital, carbamazepine, and of oral contraceptives is well known.[5,11,12,25-27] In the present study, subjects on medication (without specification of the type of medicine) had 12% higher predicted GGT levels than non-medication persons. The negative association between leisure time physical activity and GGT levels reported previously by Arnesen et al.[7] was confirmed in our population, where persons with leisure time physical activity had 8% lower predicted GGT levels. In summary, the person-related determinants of influence on GGT levels in this industrial population are in agreement with those found in the general population.

The PRV model was then used to assess the influence of the work-related variables on GGT. It was demonstrated that the variables duration of employment, employment in chemical operations, in refinery operations, and shift work had no statistically significant influence on the GGT level. In Table 8-IIb a relation is suggested between the variable "intake (PEHA, PHA)" and median GGT levels (11, 15 U/l, respectively). This relation was not confirmed by multiple regression analysis, since it is merely due to a difference in the distributions of person-related variables (age, body mass index). Another interesting finding suggested by Table 8-IIb was a relationship between the variable "duration of employment" and median GGT values. In the multiple regression analysis (corrected for age) this relation was not confirmed, because of a high correlation between age and duration of employment. Hence, the model enabled us to reject the former relationship suggested by Table 8-IIb, but not the latter.

The PRV model was also used to estimate persons' theoretical GGT values and their tolerance intervals, i.e. reference intervals corrected for personal characteristics. On the assumption that there is no liver impairment if the laboratory-determined GGT is within the tolerance limits of the predicted GGT, we conclude that, if the laboratory-determined GGT level is below 30 U/l, no further medical testing is indicated on the basis of the GGT value alone. Laboratory values above 60 U/l nearly always exceed the upper tolerance limit, and additional liver function tests are indicated for a medical judgment on liver impairment. Additional tests are also recommended for persons with GGT values between 30 and 60 U/l whose laboratory-determined GGT value exceeds the upper tolerance limit calculated on the basis of personal characteristics.

In summary, the results of this study show the possibilities of the PRV model for studying hepatocellular effects of work-related factors and for identifying persons whose laboratory-determined GGT levels indicate further medical testing. A similar approach may also be applicable for other clinical chemical tests.

8.5 REFERENCES

1 Penn R and Worthington DJ. Is serum γ-glutamyltransferase a misleading test? *Brit. Med. J.* 1983; 286: 531-535.

2 Tamburro CH and Liss GM. Tests for hepatotoxicity: usefulness in screening workers. *J. Occup. Med.* 1986; 28: 1034-1044.

3 Cravetto C, Molino G, Biondi AM, Cavanna A, Avagnina P and Frediani S. Evaluation of the diagnostic value of serum bile acid in the detection and functional assessment of liver diseases. *Ann. Clin. Biochem.* 1985; 22: 596-605.

4 Williams GZ, Widdowson GM and Penton J. Individual character of variation in time series studies of healthy people II. Differences in values for clinical chemical analytes in serum among demographic groups, by age and sex. *Clin. Chem.* 1978; 24: 313-320.

5 Schiele F, Guilmin A, Detienne H and Siest G. Gamma-glutamyltransferase activity in plasma: statistical distributions, individual variations, and reference intervals. *Clin. Chem.* 1977; 23: 1023-1028.

6 Mijovic V, Contreras M and Barbara JAJ. Serum alanine aminotrans-ferase (ALT) and γ-glutamyltransferase (γ-GT) activities in north London blood donors. *J. Clin. Pathol.* 1987; 40: 1340-1344.

7 Arnesen E, Huseby NE, Brenn T and Try K. The Tromso Heart Study: distribution of, and determinants for, gamma-glutamyltransferase in a free-living population. *Scand. J. Clin. Lab. Invest.* 1986; 46: 63-70.

8 Peterson BO, Trell E, Kristensson H, Fex G, Yettra M and Hood B. Comparison of gamma-glutamyltransferase and other health screening tests in average middle-aged males, heavy drinkers and alcohol non-users. *Scand. J. Clin. Lab. Invest.* 1983; 43: 141-149.

9 Persson J and Magnusson PH. Causes of elevated serum gamma-glutamyltransferase in patients attending outpatient somatic clinics and district health centres. *Scand. J. Prim. Health Care* 1987; 5: 13-23.

10 Rollason JG, Pincherle G and Robinson D. Serum gamma glutamyl transpeptidase in relation to alcohol consumption. *Clin. Chim. Acta* 1972; 39: 75-80.

11 Walter E, Staiger Ch, Vries J de, Weber E, Bitzer, Degott M and Jüngling K. Enhanced drug metabolism after sulfinpyrazone treatment in patients aged 50 and 60 Years. *Klin. Wochenschr.* 1982; 60: 1409-1413.

12 Keeffe EB, Sunderland MC, Gabourel JD. Serum gamma-glutamyl transpeptidase activity in patients receiving chronic phenytoin therapy. *Digestive Disease and Sciences* 1986; 31: 1056-1061.

13 Dundee JW, Fee JPH, McIlroy PDA, Black GW. Prospective study of liver function following repeat halothane and enflurane. *J. R. S. Med.* 1981; 74: 286-291.

14 Anon. Industrial agents and the liver. *The Lancet* 1982; 8307: 1081-1082.

15 Shaw LM, Strömme JH, London JL and Theodorson L. IFCC methods for the measurement of catalytical concentration of enzymes Part 4. IFCC method for γ-glutamyltransferase. *J. Clin. Chem. Clin. Biochem.* 1983; 21: 633-646.

16 Lugtenburg D, Sittert NJ van, Koningh AGJ de. Correcting for variation in analytical performance: the longitudinal effect of age on gamma-GT. Submitted to *Clin. Chem.*

17 Belsley DA, Kuh E and Welsch RE. *Regression diagnostics*, New York: John Wiley and Sons 1980.

18 Lugtenburg D, Mulder PGH. Modelling manipulated discrete processes: estimation, prevalence in case of non-response. *Kwantitatieve Methoden* (in the press).

19 Kawano S, Nakagawa H, Toga H, Okumura Y, Yamagami T, Yamamoto S and Nogawa K. Investigations on physiological values of blood in industrial workers. *Jpn. J. Ind. Health* 1982; 24: 275-282.

20 Gidlow DA, Church JF and Clayton BE. Haematological and biochemical parameters in an industrial workforce. *Ann. Clin. Biochem.* 1983; 20: 341-348.

21 Baadenhuijsen H and Smit JC. Indirect estimation of clinical chemical reference intervals from total hospital patients data: application of a modified Bhattacharya procedure. *J. Clin. Chem. Clin. Biochem.* 1985; 23: 829-839.

22 Nakamura S, Takezawa Y, Nakajima Y and Maeda T. Elevation of glutamic pyruvic transaminase and γ-glutamyl transpeptidase in obesity. *Tohoku J. Exp. Med.* 1980; 132: 473-478.

23 Nomura F, Ohnishi K, Satomura Y, Ohtsuki T, Fukunaga K, Honda M, Ema M, Tohyama T, Sugita S, Saito M, Iida S and Okuda K. Liver function in moderate obesity - study in 534 moderately obese subjects among 4613 male company employees. *Int. J. of Obesity* 1986; 10: 349-354.

24 Kornhuber J, Kornhuber HH, Backhaus B, Kornhuber A, Kaiserauer Ch und Wanner W. GGT-Normbereich bisher falsch definiert: Zur Diagnostik von Bluthochdruck, Adipositas und Diabetes infolge "normalen" Alkoholkonsums. *Versicherungsmedizin* 1989; 3: 78-81.

25 Herbeth B, Bagrel A, Dalo B, Siest G, Leclerc J and Rauber G. Influence of oral contraceptives of differing dosages on α-1-antitrypsin, γ-glutamyltransferase and alkaline phosphatase. *Clin. Chim. Acta* 1981; 112: 293-299.

26 Sano J, Kawada H, Yamaguchi N, Kawakita M and Kobayashi K. Effects of phenytoin on serum γ-glutamyl transpeptidase activity. *Epilepsia* 1981; 22: 331-338.

27 Deisenhammer E, Schwarzbach H und Sommer R. Erhohung der Gamma-GT bei antikonvulsiver Therapie. *Wiener Klin. Wochenschr.* 1982; 21: 584-585.

CHAPTER 9

A GENERAL APPROACH TO THE CORRECTION FOR BIAS IN ANALYTICAL PERFORMANCE IN LONGITUDINAL STUDIES: ESTIMATING THE EFFECT OF AGE ON GAMMA-GLUTAMYLTRANSFERASE *

SUMMARY

In longitudinal studies of variations in biochemical blood parameters with time, results may be influenced by changes in analytical methods, instruments, etc. An observed trend may well represent a drift in analytical performance instead of a truly biological finding. A model has now been developed, which allows retrospective correction of analytical changes with time. This model is based on the concept of adjustment of longitudinal data using the long-term performance of the laboratory compared with other laboratories in an external quality control programme. In practical situations, the model assumptions have to be verified.

Results from the model were applied in a study on the effect of age on γ-glutamyltransferase (GGT) in a cohort of employees between two periodic health assessments in 1984 and in 1989. Cross-sectional results apparently show an increase of GGT up to 50 years of age. However, longitudinal findings, after correction for analytical changes, do not support these results. As a possible explanation a "cohort effect" in GGT is put forward.

* After: Lugtenburg D, Van Sittert NJ, De Koningh AGJ and Van Strik R (submitted) *Clinical Chemistry*

9.1 INTRODUCTION

Every one to four years, employees of Shell Pernis are invited to attend a periodic health assessment. During this assessment, blood specimens are collected. For studying changes with time in biochemical blood parameters it is essential to correct for long-term analytical deviation in each individual's data record. Short-term (and long-term) analytical variation, or analytical imprecision, is usually assessed in internal (intra-) laboratory quality control programmes using quality control specimens, either in the form of commercial freeze-dried survey serum or of a frozen serum pool.[1-5] In this way, the analytical process is monitored. Long-term participation in external (inter-) laboratory quality control programmes permits the comparison of analyte concentrations with those measured in other laboratories using the same method, allowing the assessment of the inaccuracy, or analytical bias, of the analyte under consideration.[6,7] However, no method has been described in the literature on the use of these quality control programmes for the retrospective adjustment of longitudinal data.

Such an approach is investigated in the present chapter, using results collected in an external laboratory quality control programme[6], in which the biomedical laboratory of Shell Pernis participated over the period 1983-1990, and which concerned determinations of γ-glutamyltransferase (GGT) in quality control survey samples. A model was developed which allows the assessment of the amount of analytical bias. Results obtained with this model were applied in a separate study to investigate individual changes with time of serum GGT measurements between 1984 and 1989.

9.2 ANALYTICAL BIAS MODEL

9.2.1 Quality control programme

From 1983 to 1990 the biomedical laboratory of Shell Pernis participated in an external quality control programme.[6] In this external program (EP), analytes were determined in two different samples of quality control survey serum, from human or animal origin, which were sent in freeze-dried form by the organizers to about 200 participating laboratories on specific days once every two months. The findings of each laboratory were reported to the programme committee and, in return, this committee reported to each participating laboratory the overall mean of all the laboratory results and the mean of all laboratories using the same analytical method. For serum enzyme assays only the latter is relevant, because of the large systematic differences between results if different methods are used to measure serum enzyme activity. Consider this mean value, y, as a realization of a stochastic variable, Y (the *EP stochast*). The median of Y is postulated to be the "true" value of the enzyme activity in a quality control serum using a specific method. The amount of bias of a specific laboratory, for that measurement, can then be estimated. The EP, conditional on periods where the process was under control, will serve as a source of information from which analytical bias due to analytical changes with time is assessed. The EP is used because, in comparison with other quality control programmes, it simulates more realistically the actual every-day practice of the routine analysis of samples in a clinical laboratory.

There are three interesting parameters of Y to be estimated, viz. the 2.5th, the 50th and the 97.5th percentile. The median is the "true" value and the other two percentiles form a (analytical-change-corrected) tolerance region for an individual measurement in a quality control serum. These percentiles can be used to see whether the observed change between two consecutive measurements in serum collected from an individual, can be explained either by analytical changes or as an indication of a biological change.

9.2.2 A model for the assessment of analytical bias over time using quality control serum

Using all periods from the EP where the analytical process was under control (resulting in our study in 78 values), a statistical analysis will now be conducted to assess the analytical bias over time. The purpose of the analysis is to estimate three characteristics of the distribution of the stochastic variable Y, viz., the median or 50th percentile, the 2.5th percentile and the 97.5th percentile, when determinants of Y are taken into account. For this, the value reported by the EP committee, y, is modelled as a linear combination of k explanatory variables x_j and a normally distributed error term ϵ with mean zero:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \epsilon$$

The column vector of explanatory variables is denoted generally by $z=(1, x_1,..., x_k)^T$. Under the stated assumptions, the analysis is an ordinary least squares linear regression analysis, leading to estimates b_j for β_j. The covariance matrix of (the estimates of) the parameters β is denoted $\Sigma\ (\beta)$ and the residual mean square is denoted *MSE*. The influence of several possible determinants on Y is assessed on the basis of a linear regression analysis using appropriate F-tests to evaluate the statistical significance of reductions in *MSE*.

The most important underlying assumption is that Y follows a normal distribution with a mean value of $\beta_0 + \beta_1 x_1 + ... + \beta_k x_k$ and constant variance σ^2_ε. This assumption obviously has to be verified empirically. Residual analyses should be performed routinely for that purpose. If the assumption seems not to be met, a transformation of y might perform better, e.g. the ln-transformation. Also, first-order and second-order assumptions of the model can be checked using a test for heteroskedasticity.[8]

In a practical situation, this model is suitable for estimating the median, the 2.5- and the 97.5- percentile of Y if z is known. An estimate $p50(Y|z)$ for the median of Y, given z, is denoted $\hat{y} = b_0 + b_1 x_1 + ... + b_k x_k$, which also estimates the mean of Y under the stated assumptions. In order to estimate the 2.5th and 97.5th percentiles of Y (given z) corresponding to a newly observed column vector z, one should take into account both the residual mean square and the inaccuracy of the estimated model parameters. The estimated variance of a new observation on Y (given z) is $Var(Y|z) = MSE + z^T \Sigma (\beta) z$. The 95 per cent tolerance region of a new observation on Y, given z, is therefore : $\{\hat{y} - 1.96 \surd Var(Y|z), \hat{y} + 1.96 \surd Var(Y|z)\}$. This means that an estimate $p2.5(Y|z)$ for the 2.5th percentile of Y (given z) is $p2.5(Y|z) = \hat{y} - 1.96 \surd Var(Y|z)$ and that an estimate $p97.5(Y|z)$ for the 97.5th percentile of Y (given z) is: $p97.5(Y|z) = \hat{y} + 1.96 \surd Var(Y|z)$. These percentiles will further be denoted, respectively, by the lower limit of Y: $LL(Y)$ and the upper limit of Y: $UL(Y)$.

9.2.3 Determinants of analytical bias studied with quality control serum

In the assessment of the analytical bias in the measurement of GGT, five potential determinants of Y have been considered, viz. (x_1), the measured value of the GGT analyte in quality control serum; (x_2), a dichotomous variable, taking account of a change in the analytical instrument on 1 March 1985, given a value 1 for samples analysed before this date and 0 for samples analysed after it; (x_3 and x_4), source of quality control serum coded by means of two dichotomous variables, x_3 referring to whether the source of the serum is known (yes/no) and x_4 referring to whether it is human or animal provided it is known; both variables are required since the organizers of the quality control programme provided the origin of the quality control serum only from 9 February 1987 onwards; (x_5 to x_{11}), year of measurement using seven dichotomous variables for an eight-year period (1983 yes/no - 1989 yes/no) and (x_{12} to x_{22}): month of measurement using eleven dichotomous variables for 12 months (January yes/no - November yes/no). Since it is unlikely that analytical bias would be smoothly related (e.g. linearly) to either the year of examination or to months if ordered according to a temperature-of-month scale, we did not consider such relationships. Although dichotomous variables will be less powerful in the detection of, for example, a linear relationship, they are more powerful in the detection of "general" departures from the no-effect hypothesis. On the basis of the final (reduced) model the tolerance regions for *Y*, given *z*, are calculated.

9.2.4 Measurements in routine human serum with time

On the basis of the analytical-bias assessment described above using quality control serum, both an estimated median value, $\hat{y}$, and a 95% tolerance region for *Y*, given a newly observed *z*, can be calculated from a single measurement of an analyte in an individual routine serum. We then implicitly postulate that the analytical bias in analysing quality-control serum is representative of the analytical bias in analysing routine human serum. For two consecutive measurements on the same subject the corresponding fitted values and tolerance regions are now calculated as follows. Denote the explanatory column vector *z* at the first assessment z_1. The 95 per cent tolerance region for *Y*, given z_1, is calculated as $\{LL(Y|z_1)\ ,\ UL(Y|z_1)\}$. For a second occasion this vector is denoted z_2. The tolerance region for *Y*, given z_2, is $\{LL(Y|z_2),\ UL(Y|z_2)\}$. Each assessment yields an estimate $p50(Y|z)$ for the median of *Y*, given *z*, denoted by $\hat{y}_1$ and $\hat{y}_2$, respectively.

An individual change with time between two assessments made on the same subject, corrected for analytical bias is defined as $d = \hat{y}_1 - \hat{y}_2$.

Of course, a relevant question is whether the observed change, *d*, is statistically significant. This can be answered as follows. If an individual shows $LL(Y|z_1) > UL(Y|z_2)$, then the decrease is statistically significant and cannot be ascribed to analytical changes with time only (see Figure 9-1). If an individual shows $UL(Y|z_1) < LL(Y|z_2)$, then the increase is statistically significant and cannot be ascribed to analytical changes with time only (see Figure 9-1). This simple approach will result in less than 5 per cent individuals with significant differences under the (null-) assumption that these changes are only due to analytical variation with time.

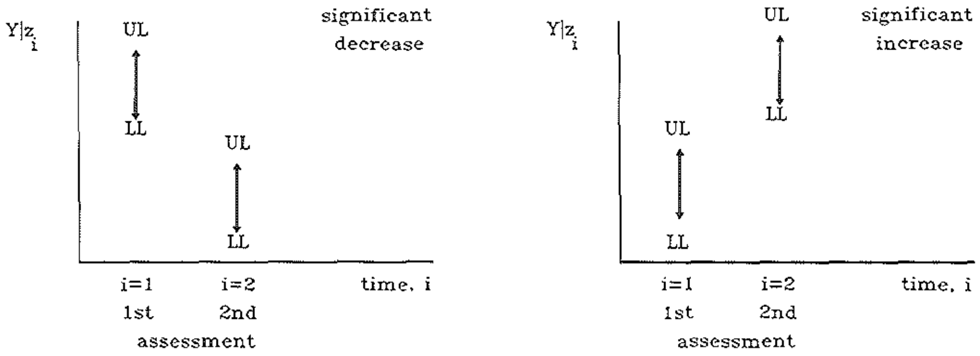


Figure 9-1 Statistically significant decrease and increase between two assessments on one and the same individual, following the simple approach.

Because of the fact that less than 5 per cent individuals will be labelled as having a significant decrease or increase, this simple approach is conservative. It might be replaced with a more refined one, as follows. Under the normality assumptions stated before, the 95% tolerance region of the difference is

$$\{d\text{-}1.96\surd(Var(Y|z_1)+Var(Y|z_2)),\ d+1.96\surd(Var(Y|z_1)+Var(Y|z_2))\}.$$

If this interval does not contain the value zero, the difference cannot be ascribed to analytical changes with time ($\alpha=0.05$). This approach will result in exactly 5 per cent individuals showing significant differences under the (null) assumption that these are only due to analytical variation with time and therefore this approach is not conservative.

If in a population, more than 5 per cent of the individuals show significant changes, then these changes can clearly not be ascribed to analytical variation with time only. This percentage may be determined using a simple binomial test of the null probability of success set at 5 per cent, which furnishes a conservative test in the simple approach.

Both the difference and the percentages significant increases (or decreases) can be related to explanatory variables, for instance age. Hence, on a population level, determinants of (significant) individual changes can be identified.

9.3 THE ANALYTICAL-BIAS MODEL APPLIED

9.3.1 Introduction

We will now describe an application of the results of an analytical-bias model. For each individual employee two GGT measurements were known, one made in 1984 and the other in 1989. For every employee the change in GGT between 1984 and 1989 was assessed quantitatively, corrected for analytical bias. Then it was assessed whether this change was statistically significant. The individual findings were combined to investigate, on a population level, whether age was related to changes in GGT.

9.3.2 Description of study population

The study population comprised all 1019 (out of a total of about 5000) male employees of Shell Pernis, who had attended two voluntary periodic health assessments, one in 1984 and the other in 1989. Many of these persons had also attended health assessments within this interval. However, a time span of 5 years was chosen, to reveal changes in GGT on an individual basis. Of each person, the age was known, the serum GGT had been determined and the body mass index had been measured. Alcohol consumption had been provided by the employees in a questionnaire. Although more variables had been assessed, only those mentioned were included in the present study, because age, body mass index and alcohol consumption are well known determinants of GGT.[9-14]

9.3.3 Specimen collection and GGT analysis

Blood samples had been collected between 8 and 10 a.m. from subjects who had not been asked to fast. Blood had been taken from the antecubital vein with the subject seated. Tourniquet pressure had been released immediately on venepuncture using the vacutainer system. The serum samples had been kept at 4° C and the GGT activity had been determined within 24 hours. No losses of activity of GGT take place if serum is stored under these conditions. In 1983, 1984 and in January/February 1985 GGT had been determined using a programmable discrete analyser PA800 (Vitatron Scientific, Dieren, the Netherlands). Since 1 March 1985, GGT analyses had been carried out with a Hitachi 705 autoanalyser (Boehringer, Mannheim). On both instruments GGT had been determined at 30° C using L-γ-glutamyl-3-carboxy-4-nitroanilide (2.9 mmol/l) as a substrate.

9.3.4 Procedure

Results from the analytical-bias model were applied to study changes in serum GGT values from individual subjects over the period 1984 to 1989. To assess the association between serum GGT and age, two cross-sectional studies were performed using data from the same 1019 subjects separately for 1984 and 1989. Regression analysis was used to estimate the influence of the various determinants on GGT. An additional longitudinal study, using the three definitions described for (significant) individual changes, was carried out for studying the changes in GGT values with time for individuals (intra-individual variation).

9.4 RESULTS

9.4.1 A model for analytical variation with time based on quality control serum

During the period October 1983 to June 1990, the Pernis laboratory participated in the EP, and determined analyte concentrations in 80 quality control samples. Figure 9-2a shows the mean values *y* of the survey samples as reported by the EP committee. Figure 9-2b shows the performance of our laboratory as measured by the ratio y/x_1 during this period. This graph shows that after the change to a different analytical instrument on 1 March, 1985, the relative performance was fairly stable.

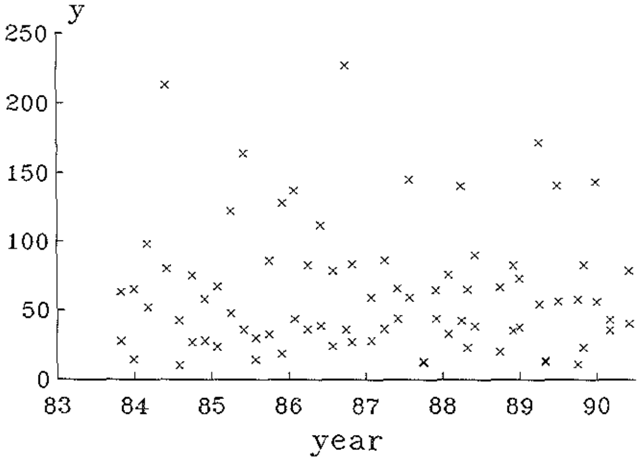


Figure 9-2a Reported mean values of GGT (y) in quality control serum of about 50 laboratories using the same analytical method during the period October 1983 to June 1990.

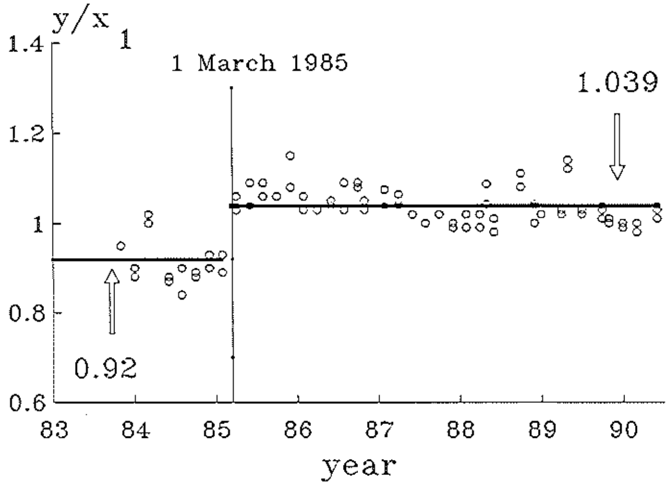


Figure 9-2b The ratio (y/x_1) of the mean GGT values in quality control serum of laboratories using the same analytical method to the GGT measurements in our laboratory during the period October 1983 to June 1990.

Two GGT measurements made in the EP during the period 21 May 1984 to 13 July 1984 were removed, since during this period, an internal quality control programme[15] revealed that the analytical coefficient of variation was above 5 per cent. This left a total of 78 values to be used in the regression analysis.

A residual analysis following the initial regression analysis showed that (1) the residuals did not obey a normal distribution and (2) that variability increased with the magnitude of fitted values. These two findings constitute an example of violating assumptions of the initial regression analysis, which may be remedied by a logarithmic transformation of both the reported value and the laboratory measurement. Therefore, from now on y denotes the (natural) logarithm of the reported value of the EP committee and x_1 denotes the (natural) logarithm of our laboratory measurement. So, Y stands for the (natural) logarithm of the *EP stochast*, i.e. the stochastic variable pertaining to the reported value.

In the regression analysis, it turned out that after the variables x_1 and x_2 (change in analytical instrument) had been included, the year of analysis and an interaction of x_1 and x_2 were relevant determinants of analytical bias (reducing the *MSE* significantly at 5 per cent level, F-test). The interaction x_1, x_2 indicates that the relation between y and x_1 was different during the two periods represented by x_2. These variables taken together determined 74 per cent of the variation in y. The origin of the serum and the month of analysis, however, were not found to have any additional significant influence on y in our study.

After x_1 had been taken into account by simply setting $\beta_1=1$, the variable x_2 appeared to be the most important additional determinant because it explained 60 per cent of the remaining variation in y, which is to be expected when we look at Figure 9-2b. We will now discuss this reduced model, since conclusions derived with it were the same as those found with the unreduced model. For this reduced model, residuals were found to be approximately normally distributed, and first-order and second-order assumptions were not rejected, when tested for appropriateness[8]. The resulting estimates for the model parameters are collected in Table 9-I.

Using a back-transformation from the logarithmic scale to the original scale, we may derive from the estimates of Table 9-I that for a GGT measurement x_1 made in 1984 or made in 1989, an estimate for the median of the *EP stochast*, is 0.92 x_1 or 1.039 x_1, respectively. This means that when the PA800 analyser was used for the GGT determinations, the values reported by our laboratory were on average 8 per cent higher than the overall mean of all the laboratories using the same analytical method. With the Hitachi 705, the values reported were on average 4 per cent lower than the mean. The 95 per cent tolerance interval for the *EP stochast*, for 1984, is from 0.85 x_1 to 0.99 x_1 and for 1989 from 0.96 x_1 to 1.12 x_1 (see Appendix 9-A for details).

TABLE 9-I

Estimates of model parameters for the reduced model, in which only x_2 is assessed as additional determinant of y, taking account of x_1 by $\beta_1=1$

$$\text{Model: } y = 0.03788 + x_1 - 0.1218\,x_2$$

estimated covariance matrix of (estimated) parameters β:

$$\Sigma(\beta)= \begin{vmatrix} 0.00002346 & -0.00002346 \\ -0.00002346 & 0.00013068 \end{vmatrix}$$

MSE=0.001501

On the assumption that the above model developed for survey serum can also be applied when fresh human serum is used, a 95 per cent tolerance interval for the *EP stochast* of a measured GGT level in 1984 is from 0.85 x_l to 0.99 x_l. For a measurement, made in 1989, this interval is from 0.96 x_l to 1.12 x_l. In the following section these results are used to investigate whether individual and population changes have occurred in serum GGT between assessments in 1984 and 1989 and to investigate the influence of age on these changes.

9.4.2 The effect of age on GGT: an application of results from the analytical-bias model

All 1019 persons who had attended a periodic health assessment both in 1984 and in 1989, were included in this analysis. Cross-sectional analyses demonstrated that within each year average GGT values increased with increasing body mass index and with increasing age for the subpopulation of individuals between 20 and 40 years. After the age of 50, average GGT levels showed a slow decrease (Figure 9-3). In addition, a positive association was found between GGT levels and level of alcohol consumption (F-test, $p<0.05$). These findings confirm results of other authors.[9-14]

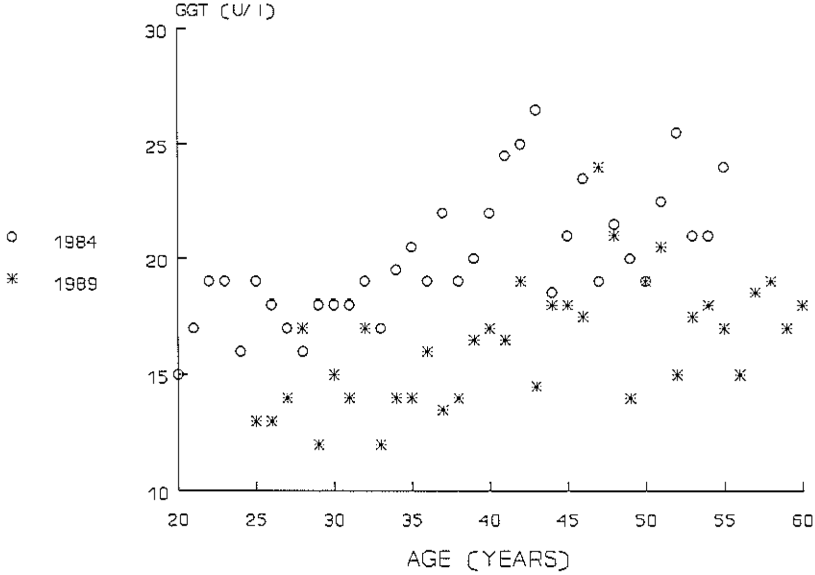


Figure 9-3 The relationship between the median of GGT levels by one-year age-groups and the age at examination, in 1984 and 1989, uncorrected for analytical bias.

Following the approach presented, a longitudinal analysis was performed on changes in GGT levels (corrected for analytical bias) in persons investigated both in 1984 and 1989. In the methodology section we discussed three methods to define individual (significant) changes. As a consequence of the logarithmic transformation, the change in GGT values for an individual is estimated as $p50(EP\ stochast|z_2)\ /\ p50(EP\ stochast|z_1)$.

We found that, compared to 1984, the GGT values in 1989 were on average 2.4 per cent lower. Regression analysis demonstrated that this change was associated with a decrease in body mass index (p=0.0001), but not with age in 1984 (p=0.49). Hence, a relation between age and change in GGT was not found in this way. Additionally, only 6.4 per cent of the intra-individual variation between these two assessments could be explained in this regression analysis. However, data on alcohol consumption were not available in 1984, so a possible effect of changes in alcohol intake on GGT could not be determined. Tentatively assuming that the data on alcohol consumption in 1989 were the same as in 1984, we introduced these in the regression analysis. The above conclusions remained the same.

The percentage subjects in the population with a significant GGT change was also calculated, with the simple method. Results stratified according to age, are shown in Table 9-II. After correction for analytical variation with time, nearly 44 per cent of the subjects showed a significant decrease in GGT after 5 years, and 18 per cent an increase. Hence, the (overall) 62 percent changes cannot be explained from analytical variation alone (binomial test; p=0.001). For each age group the percentage significantly increased GGT values is lower than the percentage decreased values and is fairly constant over the age groups. The conclusion from this second analysis supports the above longitudinal finding that age is not related to an individual change in GGT.

In addition to the easy method to define "increase" or "decrease", the more refined analysis was applied, too, using the formula for the confidence region of an individual change. Here, 22.5 per cent of the individuals showed a significant increase, whereas 52 per cent showed a significant decrease. This is somewhat higher than in the former approach, as was to be expected. Here again, the percentages significant increased and decreased GGT-values were fairly constant over the age groups.

TABLE 9-II

Percentages persons (of a total of N persons) stratified according to age that showed, using the simple method, a significant increase or a significant decrease in GGT over a period of 5 years, after correction for variation in analytical bias

Age in 1984 (years)	Increase (%)	Decrease (%)	N
15-24	11.1	66.7	18
25-29	14.4	47.0	83
30-34	16.6	44.2	181
35-39	20.7	40.0	135
40-44	19.5	35.6	174
45-49	18.4	44.9	147
50-54	20.0	43.9	155
55-60	15.1	50.0	126
Total	18.0	43.6	1019

9.5 DISCUSSION

In this chapter a method has been developed for the retrospective correction of longitudinal data for changes in analytical bias with time. The method developed is based on a statistical model using quality control survey serum measurements in an external quality control programme. It is important that the quality control survey serum used should fully simulate the fresh human serum specimens that are routinely analysed.[16] In the study presented here, no association was found between the origin of quality control serum (animal/human) and the performance of GGT measurement methods. The assumption seems justified that, with respect to GGT, fresh human serum is similar to the survey serum used in the quality control programme. So, a correction for analytical bias using quality control serum is justified. For our specific example, model assumptions concerning the normal distribution of the residuals were met after an ln-transformation.

Results from the analytical-bias model were applied in a study of individual changes in serum GGT of Shell employees between two periodic health assessments carried out in 1984 and 1989. The influence of age on these changes was also investigated. A simple conservative analysis revealed that after correction for analytical variation with time, nearly 44 per cent of all employees showed a significant decrease and 18 per cent a significant increase in GGT after 5 years. These significant changes, therefore, could not be explained from analytical variation alone. The remaining 38 per cent individual changes were possibly explained by analytical variation. Using the more refined method the significant-change figures were higher: 52 and 22 per cent, respectively, as would be expected, because this method is not conservative.

Additional results from analysing the estimated changes *d* show that the average GGT values decreased only 2.4 per cent. In this respect it is clear that the approach using statistically significant changes gives more detailed information, showing there is a clear within-person effect.

In contrast, cross-sectional analyses showed, for both years, a consistent relation between average GGT and age: it consistently increased up to 40 years. So, the conclusion is that age and GGT are related between persons but not within persons, which suggests a "cohort effect". A possible reason for this cohort effect (often found in cross-sectional studies) may be that older people underreport alcohol consumption more seriously than younger people. If this is true, the relationship between age and GGT found in cross-sectional studies[9,10], may be explained by the influence of alcohol intake on GGT. This hypothesis is supported by the lack of a relation between age and GGT in some populations with low alcohol intake.[17] In future cross-sectional studies, age should be considered an alcohol-questionnaire-adjustment variable.

9.6 REFERENCES

1 Koch DD, Oryall JJ, Quam EF, Feldbruegge DH, Dowd DE, Barry PL and Westgard JO. Selection of medically useful quality-control procedures for individual tests done in a multitest analytical system. *Clin. Chem.* 1990; 36: 230-233.

2 Stamm D. A new concept for quality control of clinical laboratory investigations in the light of clinical requirements and based on reference method values. *J. Clin. Chem. Clin. Biochem.* 1982; 20: 817-824.

3 Van Steirteghem AC, Robertson EA and Young DS. Variance components of serum constituents in healthy individuals. *Clin. Chem.* 1978; 24: 212-222.

4 Flynn FV, Piper KAJ, Garcia-Webb P, McPherson K and Healy MJR. Biological and analytical variation of commonly determined blood constituents in healthy blood donors. *Clin. Chim. Acta* 1976; 70: 179-189.

5 Fraser CG. Analytical goals are applicable to all. JIFCC 1990; 2: 84-86

6 Jansen AP, Van Kampen EJ, Leynse B, Meyers CAM and Van Munster PJJ. Experience in the Netherlands with an external quality control and scoring system for clinical chemistry laboratories. *Clin. Chim. Acta* 1977; 74: 191-201.

7 Steigstra H, Jansen RT and Baadenhuijsen H. Combi scheme: new combined internal/external quality-assessment scheme in the Netherlands. *Clin. Chem.* 1991; 37: 1196-1204.

8 White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrics*, 1980; 48: 817-838.

9 Schiele F, Guilmin AM, Detienne H and Siest G. Gamma-glutamyltransferase activity in plasma: statistical distributions, individual variations, and reference intervals. *Clin. Chem.* 1977; 23: 1023-1028.

10 Baadenhuijsen H and Smit JC. Indirect estimation of clinical chemical reference intervals from total hospital patient data: application of a modified Bhattacharya procedure. *J. Clin. Chem. Clin. Biochem.* 1985; 23: 829-839.

11 Kristenson H, Trell E, Fex G and Hood B. Serum gamma-glutamyltransferase: statistical distribution in a middle-aged male population and evaluation of alcohol habits in individuals with elevated levels. *Preventive Medicine* 1980; 9: 108-119.

12 Nakamura S, Takezawa Y, Nakajima Y and Maeda T. Elevation of glutamic pyruvic transaminase and gamma-glutamyl transpeptidase in obesity. *Tohoku J. Exp. Med.* 1980; 132: 473-478.

13 Nomura F, Ohnishi K, Satomura Y, Ohtsuki T, Fukunaga K, Honda M, Ema M, Tohyama T, Sugita S, Saito M, Iida S and Okuda K. Liver function in moderate obesity-study in 534 moderately obese subjects among 4613 male company employees. *Int. J. of Obesity* 1986; 10: 349-354.

14 Persson J and Magnusson PH. Causes of elevated serum gamma-glutamyl transferase in patients attending outpatient somatic clinics and district health centres. *Scand. J. Prim. Health Care* 1987; 5: 13-23.

15 Jansen RTP and Jansen AP. A coupled external/internal quality control program for clinical laboratories in the Netherlands. *Clin. Chim. Acta*, 1980; 107: 185-201.

16 Lott JA, O'Donnell NJ and Grannis GF. Interlaboratory survey of enzyme analyses III Does college of American pathologists' survey serum mimic clinical specimens? *Am. J. Clin. Pathol.* 1981; 76: 554-566.

17 Arnesen E, Huseby NE, Brenn T and Try K. The Tromsø Heart Study: distribution of, and determinants for, gamma-glutamyltransferase in a free-living population. *Scand. J. Clin. Lab. Invest.* 1986; 46: 63-70

APPENDIX 9-A

Suppose the log of the observed GGT value is denoted x_1 and that the log of the GGT value reported by the EP committee, denoted y, is considered to be a realization of a stochast with estimated mean value $\hat{y}=b_0 + b_1 x_1 +...+ b_k x_k$, then an estimate for the (median) value of the *EP stochast* is x_1.exp{$\hat{y}$}.

From $\hat{y}$ = 0.03788 - 0.1218 x_2 (cf. Table 9-I) it follows for samples analysed before 1 March 1985 (i.e. $x_2=1$) that $\hat{y}$ = - 0.08392, leading to an estimate for the "true" value in 1984 as 0.92 x_1. Similarly, for samples analysed after 1 March 1985 (i.e. $x_2=0$) $\hat{y}=0.03788$, which results in an estimate of 1.039 x_1.

For a normally distributed variable Y, $Var(Y|z)$ is the sum of the residual mean square, *MSE*, and a linear combination of (co-)variances of estimates of parameters β_i, in matrix-notation: $x^T\ \Sigma(\beta)\ x$ (cf. statistical method section). In case of only one explanatory variable x_2, this matrix formula can be rewritten as: $V_1 = x_2^2\ Var(\beta_2) + 2x_2\ Cov(\beta_0,\beta_2) + Var(\beta_0)$, where β_0 and β_2 are the regression parameters to be estimated.

In the example V_1 is the variance of a new observation of Y, being the ln-transformation of the *EP stochast*. An antilog transformation is needed to obtain a tolerance region for the *EP stochast*. Multiplying the estimate for the "true value" by exp{-1.96 $\sqrt{V_1}$} gives a lower limit and multiplying by exp{+1.96 $\sqrt{V_1}$} gives an upper limit for the 95 per cent tolerance region of the *EP stochast*, given z, where V_1 is estimated as described above.

From the results in Table 9-I, it can be seen that MSE=0.001501, $Var(\beta_0)$ is estimated to be 0.00002346, $Var(\beta_2)$ to be 0.00013068 and $Cov(\beta_0,\beta_2)$ to be -0.00002346. For all observed values before 1 March 1985 ($x_2=1$) the estimated value V_1 therefore equals 0.001608. The lower and upper limit of the 95 per cent tolerance region of the *EP stochast* in 1984 can therefore be calculated by multiplying the estimate of the "true" value by 0.92 and 1.08, resulting in 0.85 x_1 and 0.99 x_1, respectively. For all observed values of GGT after 1 March 1985 ($x_2=0$) the estimated value V_1 equals 0.0015244. This means that the lower and upper limit can now be calculated by multiplying the estimate of the "true" value by 0.926 and 1.079, resulting in 0.96 x_1 and 1.12 x_1, respectively.

A reparametrization of the model in which the ln-transformed value of the ratio (reported value) / (own measurement), given specific determinants, would be assumed to follow a normal distribution, leads to identical results.

CHAPTER 10

GENERAL DISCUSSION

10.1 INTRODUCTION

The main subject of interest in this thesis is the generalizability of conclusions derived from statistical analyses of studies with incomplete records and/or dubious data. To investigate the question of generalizability, we have considered several sources of bias. The first source on the part of the investigator would be the omission to take account of the fact that using a model or a test statistic implies that certain assumptions have to be met. These assumptions must therefore be checked (usually on the data) for their plausibility. This aspect of generalizability evidently needs to be considered in every study because, if left unsolved, it is not worthwhile for purposes of generalizability to put any effort into reviewing additional information. In this context, by way of example, consideration has been given to the behaviour of a test statistic under the null hypothesis when samples are small. It is well known that asymptotic results do not apply then. Two well-known alternative approaches have been discussed. A second-order Taylor approximation, discussed in Chapter 5, has the advantage that fairly simple correction formulae might result, but has the disadvantage that it cannot be used if sample sizes are very small. However, for such cases an exact method is available, as presented in Chapter 6. If the null behaviour is misspecified, conclusions derived from values of test statistics may be invalid and generalizability is not warranted.

Although the two examples discussed in Chapters 3 and 4 illustrate relevant issues of generalizability, in practice investigators encounter such generalizability problems only rarely. Whether generalizability is an issue that needs to be addressed separately depends on the research questions to be answered, as discussed in section 10.2. The main results of our study of problems of generalizability are considered in section 10.3. Related statistical topics are presented in section 10.4. Authors of papers reporting on studies involving incomplete records are invited in the final section to present information that may enable the reader to assess generalizability of conclusions.

10.2 GENERALIZABILITY DEPENDS ON RESEARCH QUESTIONS

Let us consider the occurrence of dubious data, due to misclassification for categorical data and due to measurement errors for continuous data. Without adjustment for this source of bias, any conclusions can be generalized only to populations that are identically sampled with the same degree of information loss due to dubious data. In some situations this is inevitable and may have no serious consequences. For example, in screening for a specific disease, a diagnostic instrument is used having sensitivity and specificity that are usually both less than 100 per cent. Whether this is a serious drawback of the instrument depends on whether it still detects correctly the potentially sick persons. Suppose screening on alcohol intake is done by means of a questionnaire, which is bound to have sensitivity and specificity for isolating persons with a drinking problem far below 100 per cent. However, if this questionnaire is nevertheless able to isolate the group of persons with a higher probability for developing liver cirrhosis, assumed to be related to heavy drinking, it can still be a very valuable screening device. The research question in this screening study, therefore, is answered satisfactorily. On the other hand, this questionnaire cannot be used if the research aims at finding out to what degree heavy drinking is related to liver cirrhosis.

In Chapter 1 a clear conceptual similarity was shown between the bias due to such dubious data and bias resulting from incomplete records. For both, resampling provides information on the mechanisms that may have generated the problematic data. This information can be used to adjust for bias. Here, another similarity is pointed out, namely that in some situations it is not necessary to have additional information on incomplete observations. For example, assume that the data collected from persons attending a periodic health assessment (PHA) are used in long-term follow-up studies for detecting possible risk determinants for developing ischaemic heart disease. The determinants found (say age) may be used directly in assessing new findings at periodic health examinations and conclusions will be valid. However, if such determinants would be cited, for instance, in a site-dependent information campaign, it should be kept in mind that participants and non-participants of PHAs differ. Therefore determinants found with participants may differ from those that can be found with non-participants. So, if generalizability towards a wider population than sampled is aimed for, the bias originating from such a sampling procedure has to be removed.

In summary, problems of generalizability are concerned in the first place with using correct statistical models and correct test statistics. If the underlying assumptions are met, the degree of generalizability aimed at will determine whether additional studies about dubious data or incomplete records are needed to reduce bias.

10.3 STATISTICAL METHODS DEALING WITH GENERALIZABILITY

If generalizability of results is at stake, for example because some data are missing or are of dubious quality, feasibility of any bias reduction depends on the postulation of specific underlying mechanisms, for example missing-data-generating mechanisms. These mechanisms may help to predict the effect of bias on relevant conclusions. In Chapter

2, three well-known mechanisms for the occurrence of missing data were discussed, with appropriate methods for statistical analysis.

The use of specific mechanisms helps to show how conclusions might change if circumstances vary. In Chapter 3, such an approach is used to show the effect of incomplete records in the estimation of the relative risk in a follow-up study. In addition, a method to correct for bias is presented. It is furthermore shown that the odds ratio is insensitive to such incomplete follow-up, which might therefore be preferable in situations like these.

One method to assess bias due to incomplete sampling in a specific study is to analyse additional data, either collected at baseline on all individuals, including non-participants, or on a random sample of all individuals near the end of the study. In Chapter 7, participants and non-participants in a trial on the effect of a back school education programme are compared with respect to variables measured *before* invitations were sent. It turned out that participants were fairly comparable with non-participants, except for a few of the observed variables. If such predictors of participation are also determinants of the relevant outcome variable in the trial (which was days of absenteeism in this instance) generalization needs adjustment for these determinants. Therefore, this extra information on non-participants would be needed, if the effect of the intervention is to be generalized to the whole target population. In Chapter 4, data on participants and nonparticipants in a PHA are used to adjust the estimated prevalence of hypertension at a specific site. Obviously, only variables that have been measured for both, participants and non-participants of PHAs, can be used for this purpose. Therefore, it is necessary to consider the relevant variables very carefully at the design of a prevalence study. In Chapter 9, a method is presented to assess long-term analytical bias in the determination of GGT. Without adequate insight into the way these values fluctuate with time due to analytical bias and imprecision alone, the study of relationships between analyte levels and person-related variables would not be feasible in a longitudinal study and not even in a cross-sectional survey, as discussed in Chapter 8. Relevant data collected before the actual study are needed to make such a generalization possible. On the basis of this approach within-person changes in GGT are studied (Chapter 9) which leads to the conclusion that the relation of GGT with age is probably a cohort effect.

As noticed before, implications for small-scale studies are that they are to be planned with sufficient provisions to avoid missing data or similar problems. If this is not done, the sample size is usually too small to enable adequate adjustment for missing-data bias or dubious-data bias. Sensitivity analysis by simulation and use of robust statistics may then be the only alternatives. Of course, in both types of study, all data available should be used in analysing the results.

In conclusion, we would add that sensitivity analyses and simulations may improve insight in the generalizability of the findings from a study. Such sensitivity analyses show possible effects of mechanisms that may have created bias on the relevant conclusions of a study. Relevant data must be collected to support the existence of a specific mechanism for each actual application separately.

10.4 THE METHODS IN A WIDER PERSPECTIVE

Three fields of application in statistics not yet mentioned, are closely related to the subjects discussed in this thesis. They are briefly reviewed here with reference to some relevant literature.

In some studies, missing data or misclassification are a consequence of the design. An example is a study in which measurements on the desired variable are too expensive (Rao, 1956). Surrogate endpoints are then used according to specific designs in different situations (Tosteson and Ware, 1990). Evidently, actual data must be studied to warrant the quality of these surrogate measurements, before they are used. The area of surrogate measurements is closely related to latent class analysis. In such an analysis, it is assumed that the measurement of interest is not observed directly, but that mechanisms can be postulated which result in surrogate measurements, after which the data are analysed accordingly (Kaldor and Clayton, 1985).

With respect to regression analysis, methods and diagnostics have been developed that can help to identify data that have a very strong influence on regression parameters or which will make these parameters highly dependent on each other (Belsley, Kuh and Welsch 1980). A method of dealing with such highly influential points may be to delete them from the analysis. However, these regression diagnostics require that the number of incorrect data points is known *a priori*, which in general is not the case. The method of robust statistical analysis to isolate such points (Atkinson, 1986) therefore was a large step forwards.

In cancer research, the occurrence of "wrong" data has led to an interesting application, which uses the concept of relative survival. In studies where data on cause-specific death, D, are unreliable and where the only reliable outcome is death, the relative survival rate is defined as the ratio of two survival rates: the observed survival rate for the actual group of patients divided by the survival rate to be expected for a similar group taken from the general population. The requirement of similarity to the patients at the beginning of their follow-up period with respect to all possible factors affecting survival, except for D (Hakulinen and Tenkanen, 1987) is easy to understand, but the main problem in actual applications is to find such a similar group.

10.5 PRESENTING OF DATA RELEVANT TO GENERALIZABILITY

Although criteria for internal validity are well established and generally accepted (Chapter 1), in actual applications it is not always made clear whether these criteria have been applied. In particular with regard to validity of selection and validity of information, authors should present information. If inclusion and exclusion criteria have been applied, at least the numbers of subjects that were screened, invited, excluded, participating or withdrawn should be explicitly mentioned. To take a trivial example, it is of interest to any reader to know whether a significant correlation reported in a specific study refers to only 1 per cent of the invited subjects, or to 99 per cent. In both cases, the computations may be right, but generalizability will be quite a different matter. In the field of meta-analyses, problems do arise when reported data are below standards. Studies should be given lower weight, for example, when the quality is doubtful (Spector and Thompson, 1991) since the general purpose is to obtain an objective summary of available evidence. Methods discussed in this thesis can be used to translate quality in quantitative terms. I would therefore conclude by encouraging authors of empirical studies in the life sciences, to present more data relating to generalizability and to include a discussion of generalizability in their future work.

10.6 REFERENCES

Atkinson AC (1986). Masking unmasked. *Biometrika*; 73: 533-541.

Belsley DA, Kuh E and Welsch RE (1980). *Regression diagnostics: identifying influential data and sources of collinearity*. John Wiley and Sons. New York.

Hakulinen T and Tenkanen L (1987). Regression analysis of relative survival rates. *Applied Statistics*; 36: 309-317.

Kaldor J and Clayton D (1985). Latent class analysis in chronic disease epidemiology. *Statistics in Medicine*; 4: 327-335.

Rao CR (1956). Analysis of dispersion with incomplete observations on one of the characters. *J.R.Statist.Soc. B*; 18: 259-264.

Spector TD and Thompson SG (1991). The potential and limitations of meta-analysis. *J. Epidem. Community Health*; 45: 89-92.

Tosteson TD and Ware JH (1990). Designing a logistic regression study using surrogate measures for exposure and outcome. *Biometrika*; 77: 11-21.

SUMMARY

Large-scale studies in occupational health research are often hindered by the problem that the data collected are far from ideal. It may be that part of the observations are either incomplete or of doubtful quality or both. Conclusions based on these observations are subject to criticism, especially if generalizations are made to hypothetical situations where all data are known and of sufficient quality. In this thesis problems are studied that are concerned with generalizability due to incomplete observations or dubious data or both. These problems are shown to be fairly similar with regard to the way in which they may be dealt with.

If all provisions to prevent incomplete records have failed, the statistical analysis needs to assume a specific underlying missing-data-generating mechanism in order to lead to valid, unbiased parameter estimates. Likelihood-based methods generate unbiased estimates if the missing data are missing-at-random. All available data can be used then, even from incomplete records. Other mechanisms, too, may be considered to have generated the missing data. Resampling within the proportion of persons with incomplete records is the only feasible method of checking whether such a mechanism is operative. If resampled information is available, this information can be used to adjust for the bias due to incomplete observations (Chapter 2).

In large-scale occupational health projects, attention often focusses on either estimating a relative risk or on estimating a prevalence. The former may be relevant in mortality studies for assessing possible effects of a specific exposure. The latter is relevant for assessing the feasibility of an intervention project at a specific site. If, in mortality studies, the fraction of persons alive of which the follow-up status is known is smaller than the similar fraction of persons that died, then the simple estimate for relative risk is biased towards one, although in many situations only slightly so. In an actual study, correction for this bias is feasible if estimates for those fractions are present (Chapter 3). Concerning prevalence estimates, some bias will result if determinants of a record becoming incomplete are related to determinants of the feature of interest. However, if data on these determinants are known for all persons at the site, a correction formula may be derived. The merits of extreme imputation procedures, like "all persons of which the outcome is unknown have the feature of interest", as just one realization of a parametric model, can be judged using a sensitivity-based analysis with a likelihood ratio statistic as the measure of sensitivity (Chapter 4).

Obviously, if use is not made in actual applications of the proper statistics, generalization is simply not justified. In small-scale studies, for example, two general approaches are available to derive the (null-) behaviour of statistics. Either a second-order approximation may be used to improve on asymptotic findings (Chapter 5) or, having postulated an underlying model for the data, this (null-) behaviour can be calculated exactly. If an underlying parametric model is assumed, model parameters need to be estimated or postulated. For the Wilcoxon-Mann-Whitney test statistic it is possible to postulate a parameter setting that results in overall conservative p-values, although less conservative than if a conditional approach based on ranks is used (Chapter 6).

As an example of missing data in large-scale studies, a trial on the effect of a back school education programme has been considered where about 57 per cent of persons invited refused to participate. Analysing the data that were available at baseline on both groups, participants and non-participants, revealed inter-group differences regarding educational level, body mass index, smoking habits and duration of latest absenteeism. Therefore, it will be needed to check in the ultimate analysis of the trial, whether these variables are also related to the outcome variable. Our conclusion was that any effects of the intervention, to be demonstrated in the participants group, are likely to be generalizable to the non-participants group possibly after some adjustment (Chapter 7).

A second example concerned a study of methods for the use of longitudinal liver function data, collected in periodic health assessments, for the prediction of imminent sickness absence from work. An important question was whether long-term analytical bias occurred during the period of follow-up. For one of the parameters involved, GGT, it became apparent that such a bias indeed occurred with a change in analytical device in March 1985 (Chapter 9). For data collected after that date no correction for analytical bias was deemed necessary. Therefore, conclusions drawn from a cross-sectional study (Chapter 8) may be considered valid as regards the accuracy of GGT measurements.

In a discussion of results from this thesis (Chapter 10), the relation is stressed between the research question to be answered and the necessity to address generalizability, in practical situations. If generalizability is an issue and incomplete records or dubious data are present, an underlying mechanism has to be postulated for those problematic data. If it is not feasible to collect data which may or may not support this specific mechanism, sensitivity analyses may be performed to show the dependence of relevant conclusions on various underlying mechanisms proposed. A final remark stresses the general desirability of the presentation of more informative data concerning the generalizability in published studies and the use of these specific data to adjust general findings.

SAMENVATTING

In grootschalige onderzoekingen in de bedrijfsgezondheidszorg zijn de verzamelde gegevens soms verre van ideaal, hetzij omdat ze van dubieuze kwaliteit zijn of omdat delen van de gegevens ontbreken. Conclusies gebaseerd op de aldus verkregen gegevens zijn onderhevig aan kritiek, in het bijzonder als er gegeneraliseerd wordt naar situaties met alle gegevens bekend en van een ondubbelzinnige kwaliteit. In dit proefschrift worden problemen die tijdens dit generalisatie proces optreden bestudeerd, waarbij de aanpak van beide problemen overeen blijkt te komen.

Als elke voorzorg om onvolledige "records" te vermijden heeft gefaald moet, om zuivere parameterschattingen te verkrijgen, bij de statistische analyse een onderliggend mechanisme worden gepostuleerd dat de ontbrekende waarden heeft gegenereerd. Om het bestaan van een gespecificeerd mechanisme aannemelijk te maken wordt extra informatie geëist. Indien beschikbaar, kan deze meegenomen worden in de analyse om te adjusteren voor de systematische afwijking, veroorzaakt door onvolledige "records". Indien de ontbrekende waarden "missing-at-random" zijn, zullen "likelihood" georiënteerde statistische methoden zuivere schattingen opleveren. Tevens kunnen dan alle verzamelde gegevens in de analyse betrokken worden (Hoofdstuk 2).

Grootschalig onderzoek binnen de bedrijfsgezondheidszorg richt zich vaak op het schatten van een relatief risico of een prevalentie; het eerste om het mogelijke effect van een bepaalde blootstelling te kwantificeren; het tweede om het mogelijke effect van een grootscheeps interventie-project te bepalen. Als in mortaliteits studies de fractie levenden, waarvan bekend is dat ze levend zijn, kleiner is dan de fractie doden, waarvan bekend is dat ze overleden zijn, dan krijgt de schatter van het relatieve risico een systematische afwijking naar 1; deze afwijking is in veel gevallen echter gering. Wanneer extra informatie aanwezig is, kan er voor deze afwijking geadjusteerd worden (Hoofdstuk 3). De prevalentie schatter krijgt een systematische afwijking indien een determinant van deze prevalentie tevens gerelateerd is met het ontbreken van relevante gegevens. Als voor alle personen uit de populatie informatie betreffende deze determinant aanwezig is, kan hiervoor geadjusteerd worden. Met behulp van een gevoeligheids analyse kunnen verschillende "imputatie"-technieken met elkaar worden vergeleken door ze in een allesomvattend model te plaatsen (Hoofdstuk 4).

Als bij de analyse een verkeerde toetsingsgrootheid wordt gebruikt, is generaliseren vanzelfsprekend niet gerechtvaardigd. Bij de analyse van kleine steekproeven stelt in dit kader bijvoorbeeld de (nul) verdeling van de toetsingsgrootheid ons soms voor problemen. Er zijn twee algemene methoden om deze verdeling te bepalen: een tweede-orde benadering (Hoofdstuk 5) of, als er een onderliggend model wordt verondersteld, een exacte berekening. Het schatten van model parameters is bij de laatste methode evenwel moeilijk. Voor de Wilcoxon-Mann-Whitney toetsingsgrootheid echter lijkt het mogelijk om een parameter set te postuleren, die resulteert in minder conservatieve p-waarden dan wanneer de traditionele conditionele aanpak gevolgd wordt (Hoofdstuk 6).

Als voorbeeld van een grootschalig project met ontbrekende waarden, wordt een onderzoek gebruikt naar het effect van een "rugschool". Slechts 43 procent van de uitgenodigden was bereid in dit onderzoek te participeren. Deze participanten verschilden van niet-participanten met betrekking tot opleidingsnivo, quetelet index, rookgedrag en duur van het laatste verzuim. In het eigenlijke onderzoek zal nagegaan worden of deze variabelen ook determinanten zijn van de primaire uitkomstmaat. Effecten, aangetoond bij participanten, zullen waarschijnlijk (na een adjustering) generaliseerbaar zijn naar de niet-participanten (Hoofdstuk 7).

Een tweede voorbeeld betreft een studie om leverfunctie gegevens, verzameld tijdens periodiek geneeskundig onderzoek, te gebruiken bij het voorspellen van ziekteverzuim. Hierbij werd de vraag gesteld of veranderingen in analytische afwijkingen over een lange termijn optraden. Voor GGT bleek dit inderdaad het geval, hetgeen toe te schrijven was aan het vervangen van het analytische instrument (Hoofdstuk 9). Conclusies getrokken uit een cross-sectioneel onderzoek (Hoofdstuk 8) kunnen wat dit aspect betreft als valide worden beschouwd.

In de algemene discussie (Hoofdstuk 10) wordt de relatie benadrukt tussen de research-vraagstelling en de noodzaak om generaliseerbaarheid te onderzoeken. Als dit laatste geldt terwijl records incompleet zijn of van een dubieuze kwaliteit, dan zal er een onderliggend mechanisme dienen te worden gepostuleerd. Als het niet mogelijk is om dit mechanisme te onderbouwen met informatie, zullen sensitiviteits-analyses nodig zijn om de invloed te laten zien van bepaalde mechanismen op belangrijke conclusies. Een laatste opmerking benadrukt de noodzaak om meer informatieve gegevens betreffende generaliseerbaarheid in gepubliceerde studies te presenteren en deze gegevens te gebruiken bij het adjusteren van belangrijke conclusies.

DANKWOORD

Dit proefschrift had niet tot stand kunnen komen zonder de hulp van een groot aantal mensen. In dit dankwoord wil ik me dan ook graag tot hen richten. Allereerst mijn ouders. Bedankt voor de steun en voor het vertrouwen welke jullie me gegeven hebben op die momenten dat ik ze ook nodig had.

Professor Van Strik, uw vermogen om uit mijn hersenspinsels de relevante zaken naar voren te halen heeft me altijd enorm verbaasd. U kunt uit een document, wat warrig en onsamenhangend is, de relevante zaken isoleren en deze perfect accentueren. Hierdoor hebben alle hoofdstukken aan scherpte gewonnen. Uw snelheid hierbij was bewonderenswaardig. Bedankt voor uw tijdsinvestering.

Paul Mulder, jij hebt me fantastisch opgevangen toen ik drie jaar geleden bij de Erasmus Universiteit begon te werken. Bij het vertalen van praktische problemen naar statistische oplossingen was jouw hulp essentieel. Je invloed is in veel van de hoofdstukken terug te vinden, met name hoofdstuk 4 heeft veel aan waarde gewonnen doordat we samen hieraan hebben gewerkt, zelfs tot aan de koffietafel in Brussel. Bedankt.

Dries de Koningh, je staat aan het hoofd van een efficiënt werkende afdeling, waar ik gedurende drie jaar met veel plezier heb gewerkt. Bedankt dat ik dit onderzoek bij jullie heb mogen verrichten. Juist dat heeft me in de gelegenheid gesteld om met veel mensen van de BGD gedurende enige tijd samen te werken, wat resulteerde in de hoofdstukken 3, 4, 7, 8 en 9. Jijzelf bent, samen met Nico van Sittert, een belangrijke gesprekspartner geweest, met name voor wat betreft de hoofdstukken 8 en 9. Ik wil jullie beiden hiervoor bedanken.

Gerard Braat, bedankt dat jij me als vriend en paranimf vooral organisatorisch hebt geholpen.

Erik Buskens, ex-kamergenoot en paranimf, ook jou bedank ik voor het feit dat je veel van de organisatie van het promotiefeest op je genomen hebt.

Ronald Stolk, bedankt voor de adviezen met betrekking tot tekstverwerking en grafische lay-out, zonder welke dit proefschrift er zeker minder fraai had uitgezien.

De rugschoolgroep, Harry Rensink, Rob Wiemer en Wim Drooger, bedankt voor de vele leuke discussies die we hebben gehad in dit kader. Hoofdstuk 7 is duidelijker geworden door onze besprekingen.

Het gaat te ver om iedereen bij name te bedanken. Het zijn er te veel. Ik ben daardoor genoodzaakt ook hier te generaliseren: ik wil iedereen van het Instituut Epidemiologie en Biostatistiek en iedereen van de Bedrijfs Gezondheids Dienst van Shell Pernis bedanken voor de vele inspirerende gesprekken die ik tijdens deze drie jaren heb mogen voeren.

Tenslotte bedank ik de heer Henk Hazelaar voor het kritisch doornemen en corrigeren van de Engelse tekst.

Lieve Sandra, bedankt voor je aanwezigheid tijdens het laatste gedeelte van dit werk.

CURRICULUM VITAE

Dick Lugtenburg werd op 12 maart 1966 te Zuidland geboren. Hij bezocht de Christelijke Scholengemeenschap "Blaise Pascal" te Spijkenisse, waar hij in 1984 het VWO diploma behaalde. Aansluitend begon hij de studie wiskunde, aan de Technische Universiteit Delft (destijds nog Technische Hogeschool). Na twee jaar specialiseerde hij zich in de richting statistiek, stochastiek en operations research. De begeleiding van zijn afstudeeropdracht, uitgevoerd bij Duphar B.V. te Weesp, was in handen van Peter van Ewijk en Bert van Zomeren. In september 1988 behaalde hij zijn ingenieursdiploma. Van september 1988 tot januari 1989 werkte hij als statisticus bij Duphar. Van januari 1989 tot januari 1992 werkte hij als wetenschappelijk onderzoeker aan het Instituut Epidemiologie en Biostatistiek van de Erasmus Universiteit aan zijn proefschrift, in een gemeenschappelijk project met de Bedrijfs Gezondheids Dienst van Shell Nederland Raffinaderij B.V./Shell Nederland Chemie B.V. Het onderzoek werd uitgevoerd onder supervisie van Professor R. van Strik en Drs. P.G.H. Mulder. Sinds januari 1992 is hij werkzaam bij Shell Nederland B.V. als bedrijfsinformaticus. Tevens is hij sinds die datum regionaal correspondent van de Biometric Society.

STELLINGEN

behorende bij het proefschrift

Some statistical aspects of the generalizability of occupational health studies

1. Het beantwoorden van de vraag betreffende generaliseerbaarheid vereist additionele gegevens.
Dit proefschrift

2. Vergeleken met misclassificatie leidt incompleetheid tot interpretatie problemen van een lagere orde.
Dit proefschrift

3. 10100 11 2, 10101 10 1, 10102 14 M, 10103 10 1, 10104 13 1, 10105 16 M, 10106..
Zonder kennis van de analyse van complete gegevens resulteert de analyse van incomplete gegevens in loze uitspraken.
Dit proefschrift

4. Wetenschappelijke aanpak van de bedrijfsgezondheidszorg vereist het beschouwen van onderzoeksuitkomsten zonder "prior belief".

5. De veronderstelde leeftijds-afhankelijkheid van γ-GT lijkt een cohort effect te zijn.
Dit proefschrift

6. Bij analyse van een ordinale uitkomst variabele in kleine steekproeven, is de conditionele Wilcoxon-Mann-Whitney toetsingsgrootheid overdreven conservatief.
Dit proefschrift

7. Simulatie is een degelijk instrument om de gevoeligheid van conclusies te bepalen.

8. There is no such thing as a routine statistical question, there are only questionable statistical routines.
Sir David R. Cox

9. Bij modellering moet model-ering vermeden worden.

10. Het AIO-systeem leidt tot een positieve discriminatie van twee-verdieners.

11. De eerste van de drie robotica wetten van Asimov is niet te programmeren; daarmee zijn de overige twee ontkracht.

12. Internationale veiligheid is alleen te bereiken wanneer de staten onvoorwaardelijk afzien van een deel van hun vrijheid van handelen.
naar: Einstein en Freud, briefwisseling 1932

13. De wetenschap dat er meer is dan wetenschap geeft onterecht het gevoel dat gevoelens niet wetenschappelijk zijn te onderbouwen.

14. Gezien de tendens tot verdergaande internationalisering verdient het aanbeveling de term "Proefschrift" te vervangen door "Promotie-Research-Document".

D. Lugtenburg

Rotterdam, 22 januari 1992.