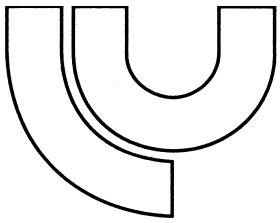

Limburgs Universitair Centrum

Faculteit Wetenschappen

Likelihood based analysis of clustered binary data with applications in developmental toxicity studies

Proefschrift voorgelegd tot het behalen van de graad van

Doctor in de Wetenschappen

aan het Limburgs Universitair Centrum te verdedigen door

LIEVEN DECLERCK

Promotoren: Prof. dr. M. Aerts

Prof. dr. G. Molenberghs

1999

Dankwoord

Bij het afwerken van dit proefschrift wens ik graag een aantal personen van harte te bedanken. Vooreerst gaat mijn dank naar m'n promotoren Marc Aerts en Geert Molenberghs. De afgelopen jaren hebben zij mij met heel veel inzet en enthousiasme geholpen bij het uitvoeren van dit onderzoeksproject over gecorreleerde binaire gegevens! Tijdens de vele vergaderingen gaven zij een heldere kijk op de volgende stapjes in dit boeiend onderzoeksgebied. Deskundige uitleg over dit domein van toegepaste statistiek werd aangevuld met vele, ludieke momenten.

Ook Paul Janssen ben ik veel dank verschuldigd. Onder zijn leiding heb ik tijdens de voorbije zes jaar de kans gekregen om ABOS-studenten te begeleiden tijdens hun biostatistiek-opleiding. Het was tof om op die manier een steentje te kunnen bijdragen tot deze ontwikkelingssamenwerking! De interesse van de ABOS- en andere biostatistiek-studenten voor m'n onderzoeksproject waren een extra aanmoediging.

Graag bedank ik hier trouwens alle professoren van het Centrum voor Statistiek van het LUC. Mede door hen kon ik m'n loopbaan een flinke duw in de goede richting geven!

Verder wens ik alle leden van het Centrum voor Statistiek te danken voor de aangename werksfeer! Het was leuk om als dertiger te werken tussen de vele (nog jongere) assistenten van de groep statistiek.

Tot slot wil ik m'n ouders en andere familieleden bedanken voor hun belangstelling in m'n onderzoek. Hun aanmoedigingen waren een extra motivatie voor het uitvoeren en afwerken van dit proefschrift!

Lieven Declerck

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Reproductive and developmental toxicity

Society has become increasingly concerned about the effects of several *types of exposure* on reproduction and development of humans. In addition to the therapeutic effects of a medicine, a patient can also experience adverse effects due to that drug. Food additives and materials such as phthalic acid esters which are used extensively as plasticizers for packing food and drinks, can also be considered as exposures and hence, concern has been raised about their possible toxic effects. In chemical factories, people are exposed to solvents and other chemicals. Furthermore, due to the environment, man is exposed to radiation and chemicals. This is illustrated by recent discussions about combustion of waste in incinerators. One of the key questions deals with the existence of a link between the concentration of certain chemicals in the exhaust-gases of an incinerator and the occurrence of birth defects in the neighbourhood of these plants.

Questions are raised about the relationship between environmental and other exposures on the one hand and *reproductive and developmental toxicity* on the other hand. More specifically, interest is in the effects of chemicals, radiation,... on infertility of men and women, on pregnancy, on early pregnancy losses, on stillbirths, on birth defects (e.g., types of malformation and low birth weight) and on postnatal developmental complications.

Regulatory agencies, such as the U.S. Environmental Protection Agency (EPA) and the Food and Drug Administration (FDA) stimulate reproductive and developmental toxicity research. One of the objectives is to understand the causes of

these problems. Also, one wants to better protect people from exposures resulting in increased risks. Besides EPA and FDA, there is also support for this type of research from other organizations, such as NATO.

There are several strategies to investigate the implications of these exposures on reproduction and development. For example, epidemiological studies can be used in this respect. In contrast to animal experiments, there is no extrapolation needed from animal data to human risk. In case of epidemiological studies focusing on effects of exposure on developing fetuses, the outcomes of the fetuses in a particular study can be considered as independent in most cases since a pregnant woman has in general only one fetus. As a consequence, the analyses are simpler as in the case of animal experiments, in which dams typically have multiple offspring. However, reliable epidemiological information is often limited or even unavailable. As a consequence, other types of studies are performed in order to have a better understanding of the effects of environmental and other agents on reproduction and development.

1.2 Animal toxicity experiments

Toxicity experiments on animals are alternatives of epidemiological studies on humans. Under some conditions, there is ethical justification to administer a dose of some toxic agent to animals. One might opt for rodents (mice, rats,...) as testing animals because of the large database in control animals regarding reproductive performance and incidence of malformations (Lindström *et al.*, 1990). Such laboratory experiments play an important role in testing and regulating substances with potential danger to humans. The results of the animal studies can be extrapolated to human beings for whom a safe dose is determined.

While the extrapolation from animal data to humans is always difficult, use of animal data on reproduction and development raises some very relevant and interesting questions. Furthermore, animal experiments have the big advantage that there is a much better control of all kinds of factors which might influence the outcomes.

Several experimental protocols are used in reproductive and developmental studies. Three test designs (Segments I, II and III) were established by the U.S. FDA in 1966 to assess specific types of effects (Food and Drug Administration, 1966).

Male and female fertility and general reproductive ability are evaluated in Segment I designs. Segment II designs are suitable when interest lies in the effects of exposure during the period of major organogenesis and structural development of fetuses. In Segment III designs, the focus is on later effects and involves exposures from late gestation through lactation.

This research deals with developmental toxicity studies (Price *et al.*, 1985; George *et al.*, 1987; Price *et al.*, 1987; Tyl *et al.*, 1988; Lindström *et al.*, 1990; Ryan, 1992) in which the emphasis is on assessing potential adverse effects of exposures on developing *fetuses*. Hence, Segment II designs are of most importance here and are used in the experiments considered in this thesis.

Rats, mice and occasionally rabbits are usually chosen as the animal model in Segment II studies. Administration of the exposure is generally by the clinical or environmental route(s) most closely mimicking human exposure. Timed-pregnant animals (dams) are exposed during the critical period of major organogenesis (days 6–15 for mice and rats; 6–19 for rabbits) and sacrificed just prior to normal delivery. At that time, the uterus is removed and the contents are thoroughly examined for defects. A standard segment II study includes a control group and three or four dosed groups, exposed to the test substance. About 20 to 30 pregnant dams are randomized to the dose groups. Typical litter sizes, i.e., the number of live-born offspring, for control animals range from 8 for rabbits, 12 for mice and 14 for rats.

1.3 Data structure of a developmental toxicity study

Several types of data are collected in developmental toxicity studies. For each dam, the number of implants as well as the dose which was administered to that dam, is registered. Furthermore, there are a number of outcomes which are observed in these experiments. Outcomes include the number of fetal deaths and resorptions (very early deaths that are detectable at the time of maternal sacrifice as a small mark on the uterine wall), the number of malformed fetuses according to several types of malformation, weight,...

The outcomes of a developmental toxicity experiment as well as some notation, are represented in Figure 1.1. Consider a developmental toxicity study involving C pregnant dams, each one resulting in a cluster. The number of implants in the

ith dam is denoted by m_i. This number does not depend on the dose given to the dam since dams are randomly assigned to a dose group and exposure to the toxic agent occurs after mating. When sacrificing the dam near the end of the gestation period, an implant is developed to a viable or a non-viable animal. The number of viables in dam i is represented by n_i, while the remaining $m_i - n_i \equiv r_i$ fetuses are non-viable. Deaths during gestation can be classified into several subcategories, including resorptions. However, in analyses of developmental toxicity experiments, one usually does not make a distinction between a resorbed and a dead fetus. Hence, in this research, they are collapsed into a "non-viable category". The viable animals are investigated according to their weight and malformations. There are a number of malformation types, but usually, they are classified into three categories: external, skeletal and visceral (Williams and Ryan, 1997). External malformations are those that the teratologist can observe with the naked eye, i.e., missing limbs, cleft palate,... Skeletal malformations are detected through specialized staining techniques. Finally, visceral malformations are those affecting internal organs, such as the heart, brain, lung and are detectable only after dissection. Typically, the malformation outcomes are binary, i.e., either present or absent, or equivalently, either a malformed or a healthy fetus according to that indicator. Here, the number of malformed fetuses in dam i according to a specific malformation type, is represented by z_i. The malformation indicators can be collapsed into a single, binary outcome, indicating if a fetus has at least one type of malformation. Such a collapsed outcome is used e.g., when specific malformations are rare.

1.4 National Toxicology Program data

The developmental toxicity studies considered here, are conducted at the Research Triangle Institute, which is under contract to the National Toxicology Program of the U.S. (NTP data). These studies investigate the effects in mice of five chemicals: ethylene glycol (Price *et al.*, 1985), triethylene glycol dimethyl ether (George *et al.*, 1987), diethylene glycol dimethyl ether (Price *et al.*, 1987), di(2-ethylhexyl)phthalate (Tyl *et al.*, 1988) and theophylline (Lindström *et al.*, 1990). In all studies, one control and three or four active dose levels are included. These are standardized such that the control dose level is zero and the highest level is one. When sacrificing the dam, the fetal weight is recorded, as well as information about each fetus being dead

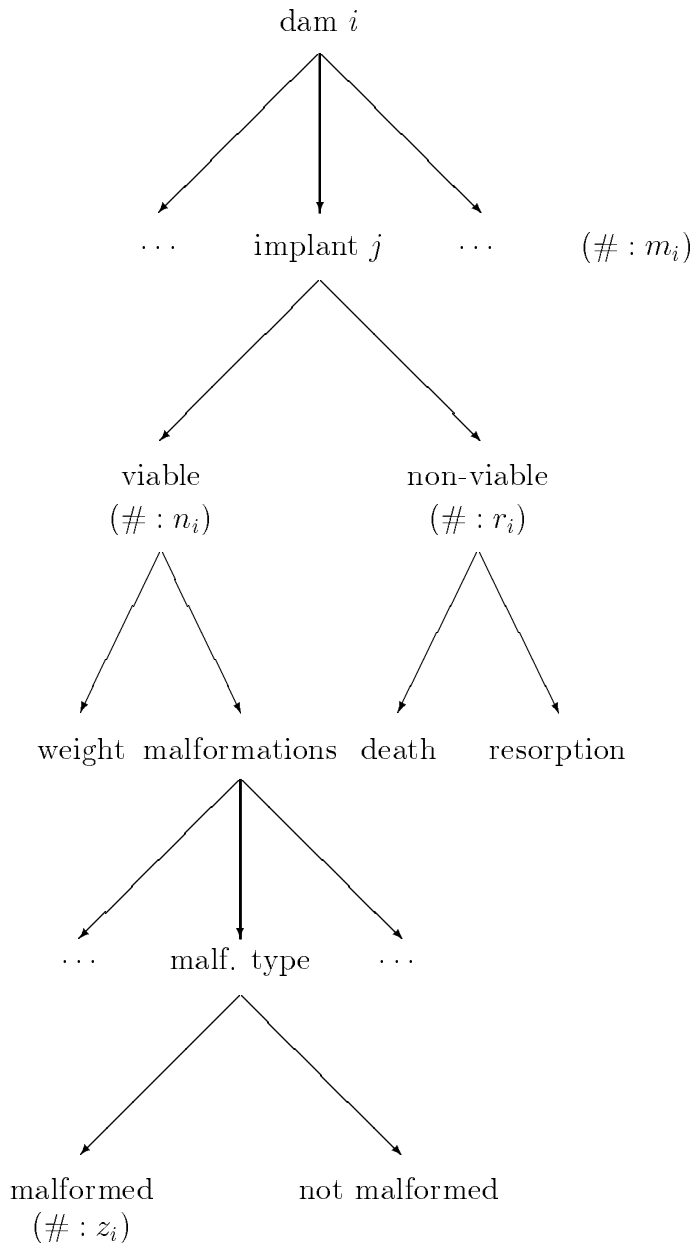

Figure 1.1: *Data structure of a developmental toxicity study.*

or viable. Indicators for external, skeletal and visceral malformation are considered. For each malformation type, one records if a fetus is malformed or healthy. It turns out that several fetuses exhibit two or three of these kinds of malformation. A collapsed, binary outcome indicating for every fetus if at least one type of malformation is present, is also investigated here.

1.4.1 Ethylene glycol

Ethylene glycol (EG) is also called 1,2-ethanediol and can be represented by the chemical formula $HOCH_2CH_2OH$. It is a high-volume industrial chemical with many applications. EG is used as an antifreeze in cooling and heating systems, as one of the components of hydraulic brake fluids, as an ingredient of electrolytic condensers and as a solvent in the paint and plastics industries. Furthermore, EG is employed in the formulation of several types of inks, as a softening agent for cellophane and as a stabilizer for soybean foam used to extinguish oil and gasoline fires. Also, one uses EG in the synthesis of various chemical products, such as plasticizers, synthetic fibers and waxes (Windholz, 1983).

EG may represent little hazard to human health in normal industrial handling, except possibly when used as an aerosol or at elevated temperatures. EG at ambient temperatures has a low vapour pressure and is not very irritating to the eyes or skin. However, accidental or intentional ingestion of antifreeze products, of which approximately 95% is EG, is toxic and may result in death (Rowe, 1963; Price *et al.*, 1985).

In the EG study, Price *et al.* (1985) consider the dose levels 0, 750, 1500 and 3000 mg/kg/day. Table 1.1 represents the number of dams containing at least one implant, as well as the number of dams having at least one viable fetus. These frequencies are listed for each dose of EG and of the other four NTP toxic agents under investigation. The distribution of the number of implants is given in Table 1.2 for each of these five chemicals. It is shown that clusters consisting of 10–15 implants occur frequently.

Figure 1.2 represents some of the data of this study. For each dose group, cumulative relative frequencies of the number of clusters are plotted for the number of implants in a cluster, the number of viable fetuses, the number of dead fetuses, the number of abnormals (i.e., dead or malformed fetuses), the number of external, skeletal and visceral malformations and the number of fetuses with at least one type

Table 1.1: Number of dams with at least one implant and number of dams with at least one viable fetus, by NTP study and by dose.

Exposure	dose (mg/kg/day)	# dams with at least 1 implant	# dams with at least 1 viable fetus
EG	0	25	25
	750	24	24
	1500	23	22
	3000	23	23
	overall	95	94
TGDM	0	27	26
	250	26	26
	500	26	24
	1000	28	26
	overall	107	102
DYME	0	21	21
	62.5	20	20
	125	24	24
	250	23	23
	500	22	22
	overall	110	110
DEHP	0	30	30
	44	26	26
	91	26	26
	191	24	17
	292	25	9
	overall	131	108
THEO	0	26	25
	282	26	25
	372	33	29
	396	23	17
	overall	108	96

Table 1.2: Frequency distribution of the number of implants, by NTP study.

Number of implants	EG	TGDM	DYME	DEHP	THEO
1	0	1	0	1	2
2	0	0	0	1	2
3	1	1	1	0	1
4	0	3	1	2	1
5	1	1	0	0	0
6	0	1	0	2	3
7	2	2	2	0	0
8	1	0	2	4	0
9	8	2	2	5	6
10	4	7	7	7	4
11	8	21	10	18	14
12	19	26	15	21	17
13	16	19	27	26	21
14	11	10	19	21	19
15	16	8	9	10	12
16	6	4	10	8	3
17	1	1	5	2	2
18	0	0	0	2	1
19	1	0	0	1	0
	95	107	110	131	108

of malformation.

1.4.2 Triethylene glycol dimethyl ether

Triethylene glycol dimethyl ether (TGDM) is also referred to as triglyme or tetraoxadodecane. Its chemical formula is $CH_3O(CH_2)_2O(CH_2)_2O(CH_2)_2OCH_3$ (Windholz, 1983). TGDM is a member of the glycol ether class of industrial solvents. These solvents are widely used in the manufacture of protective coatings (NIOSH, 1983).

Although field studies have not adequately evaluated the potential of glycol ethers to produce human reproductive toxicity, some glycol ethers have been identified as reproductive toxicants in several mammalian species (Clapp, Zaebst and Herrick, 1984; George *et al.*, 1987).

The pregnant dams of the TGDM study are exposed to 0, 250, 500 or 1000 mg/kg/day (George *et al.*, 1987). In Figure 1.3, some of the data of the TGDM study are shown.

1.4.3 Diethylene glycol dimethyl ether

Other names for diethylene glycol dimethyl ether (DYME) are diglyme and bis(2-methoxyethyl) ether. DYME has as chemical formula $CH_3O(CH_2)_2O(CH_2)_2OCH_3$ (Windholz, 1983). Like TGDM, this chemical also belongs to the glycol ether class of industrial solvents and is involved in the production of protective coatings (NIOSH, 1983).

Motivation for this developmental toxicity experiment is the same as for the TGDM study.

Price *et al.* (1987) use the doses 0, 62.5, 125, 250 and 500 mg DYME/kg/day. A representation of the data in the DYME study, is given in Figure 1.4.

1.4.4 Di(2-ethylhexyl)phthalate

Di(2-ethylhexyl)phthalate (DEHP) is also called octoil, dioctyl phthalate or 1,2-benzenedicarboxylic acid bis(2-ethylhexyl) ester. It can be represented by $C_{24}H_{38}O_4$. DEHP is used in vacuum pumps (Windholz, 1983). Furthermore, this ester as well as other phthalic acid esters, are used extensively as plasticizers for numerous plastic

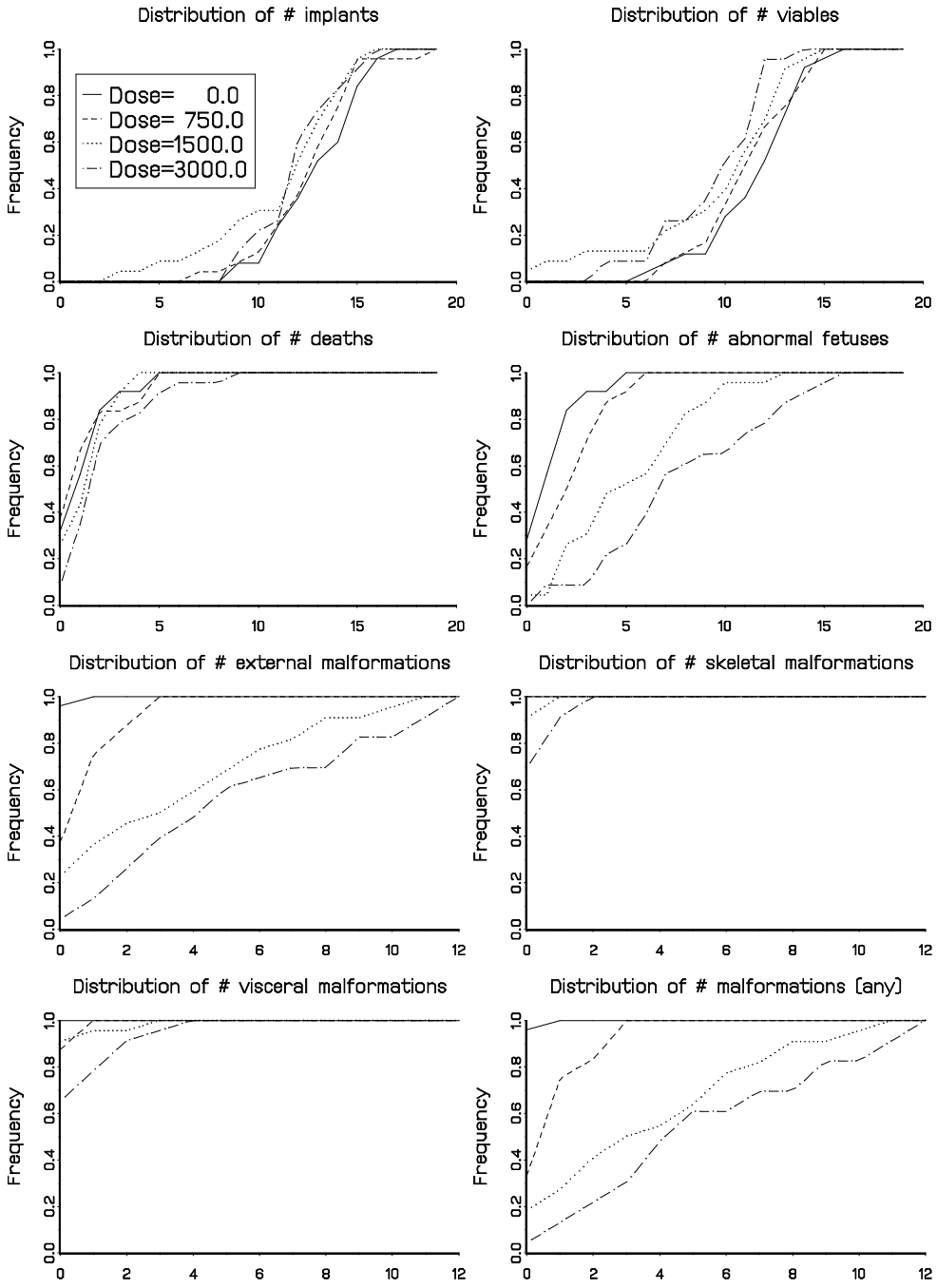

Figure 1.2: *Cumulative relative frequencies of the number of clusters representing some of the data in the EG study.*

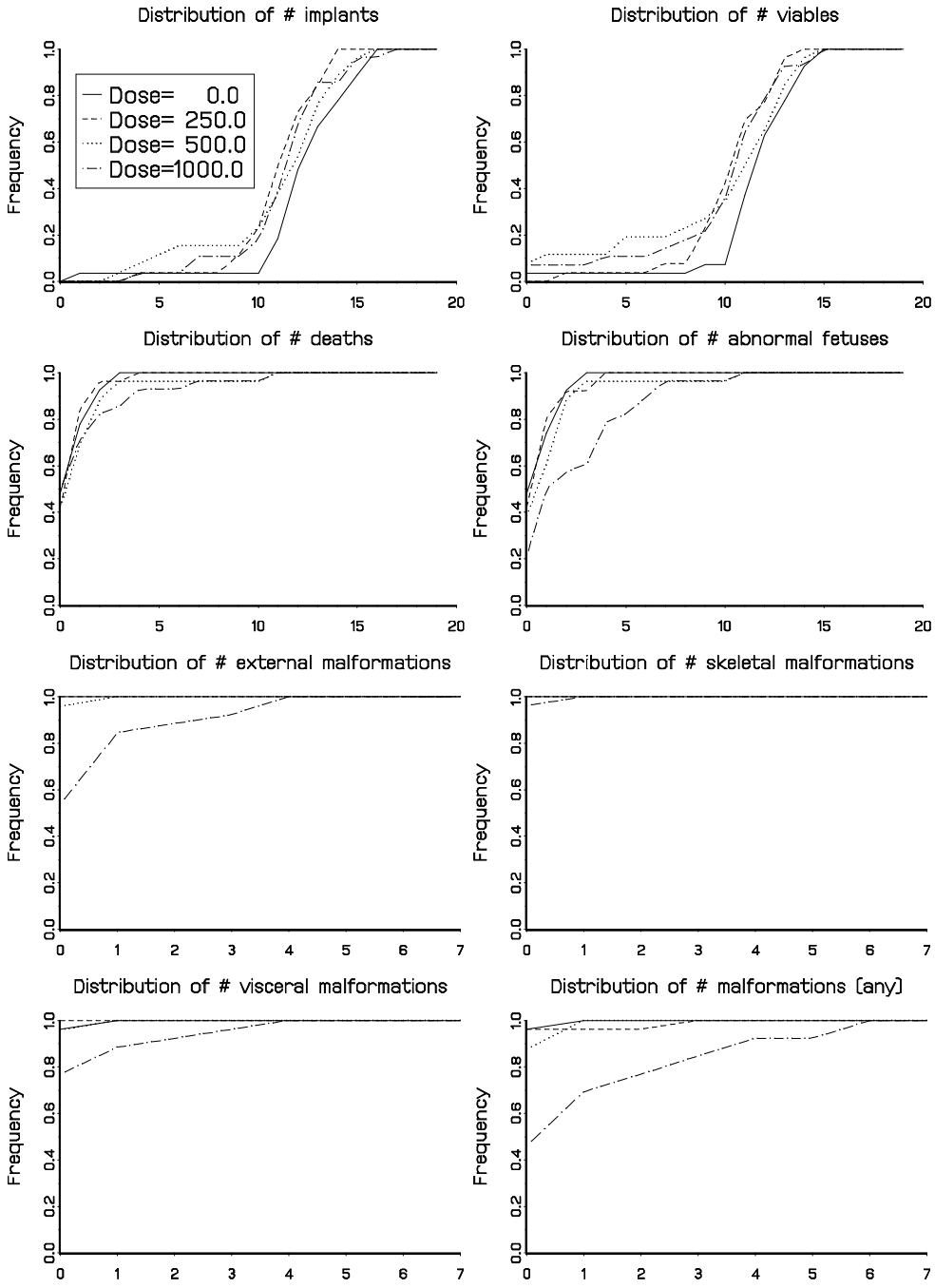

Figure 1.3: *Cumulative relative frequencies of the number of clusters representing some of the data in the TGDM study.*

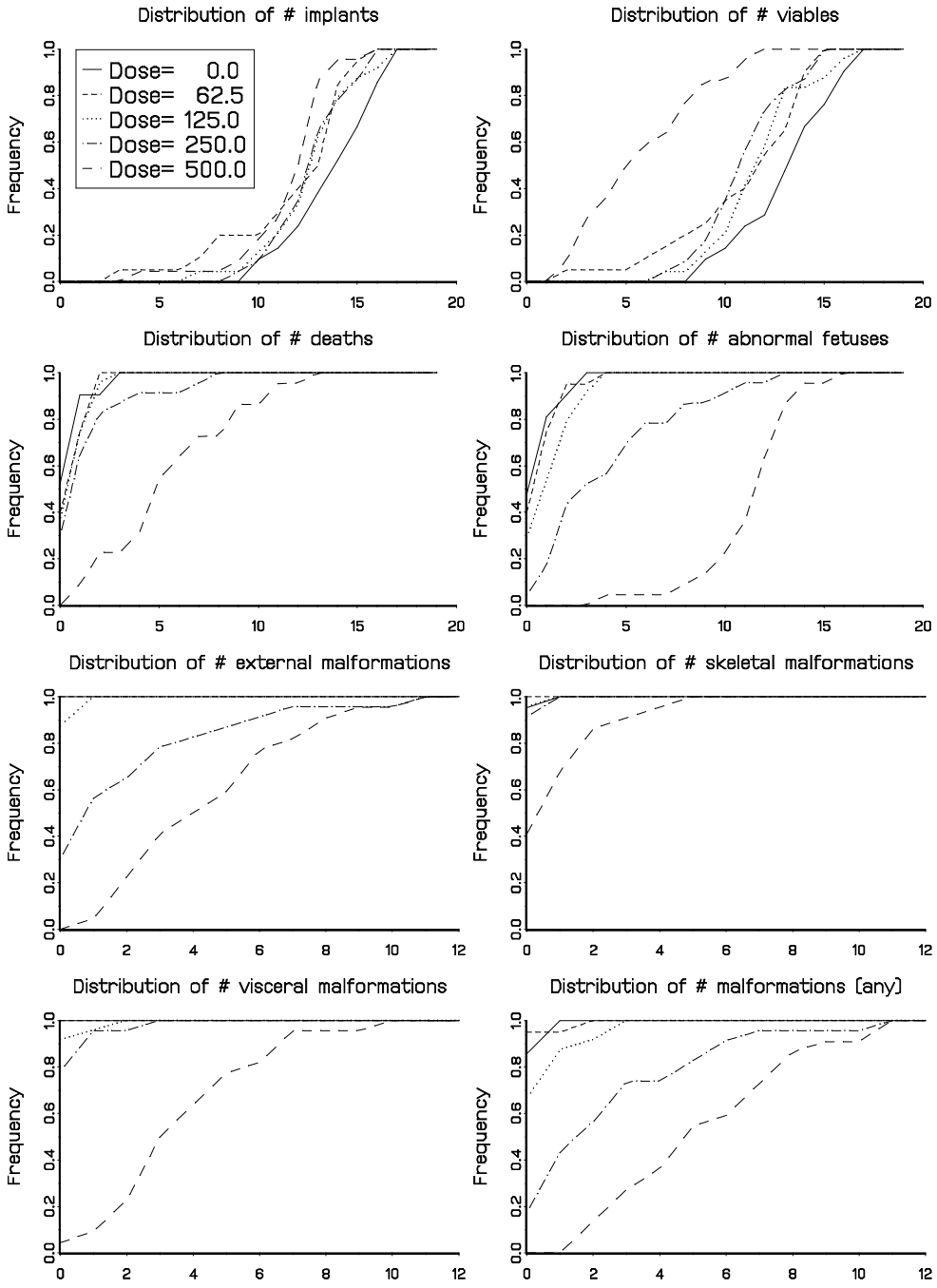

Figure 1.4: *Cumulative relative frequencies of the number of clusters representing some of the data in the DYME study.*

devices made of polyvinyl chloride. DEHP provides the finished plastic products with desirable flexibility and clarity (Shiota, Chou and Nishimura, 1980).

It has been well documented that small quantities of phthalic acid esters may leak out of polyvinyl chloride plastic containers in the presence of food, milk, blood or various solvents. Due to their ubiquitous distribution and presence in human and animal tissues, considerable concern has developed as to the possible toxic effects of the phthalic acid esters (e.g., Autian, 1973).

In the DEHP study, Tyl *et al.* (1988) consider the concentrations 0, 0.025, 0.05, 0.1 and 0.15%, corresponding to a DEHP consumption of 0, 44, 91, 191 and 292 mg/kg/day respectively. Some of the data of this study are shown in Figure 1.5.

1.4.5 Theophylline

Theophylline (THEO) has many other names, among others 1,3-dimethylxanthine, theocin and 3,7-dihydro-1,3-dimethyl-1H-purine-2,6-dione. One can represent THEO by $C_7H_8N_4O_2$ (Windholz, 1983). THEO plays an important role in the management of asthma, both as a prophylactic drug and in the treatment of prolonged attacks. It is one of the drugs of choice in the treatment of asthma during pregnancy (e.g., Kayser and Cupit, 1978).

Evaluation of the potential for THEO to cause developmental toxicity is undertaken in consideration of the widespread use of this chemical for the treatment of asthma in pregnant women and also its presence in beverages such as coffee, tea and chocolate (Lindström *et al.*, 1990).

In the THEO experiment, Lindström *et al.* (1990) expose the dams to the concentrations 0, 0.075, 0.15 and 0.2%, which correspond to a consumption of 0, 282, 372 and 396 mg THEO/kg/day respectively. Figure 1.6 represents some of the data of this experiment.

1.5 Risk assessment

Risk assessment deals with safety issues and regulation of environmental and other exposures with a potential for adverse human health effects. It investigates among others the implications of possibly toxic agents on reproduction and development via animal experiments. A number of topics can be considered in the area of risk assessment.

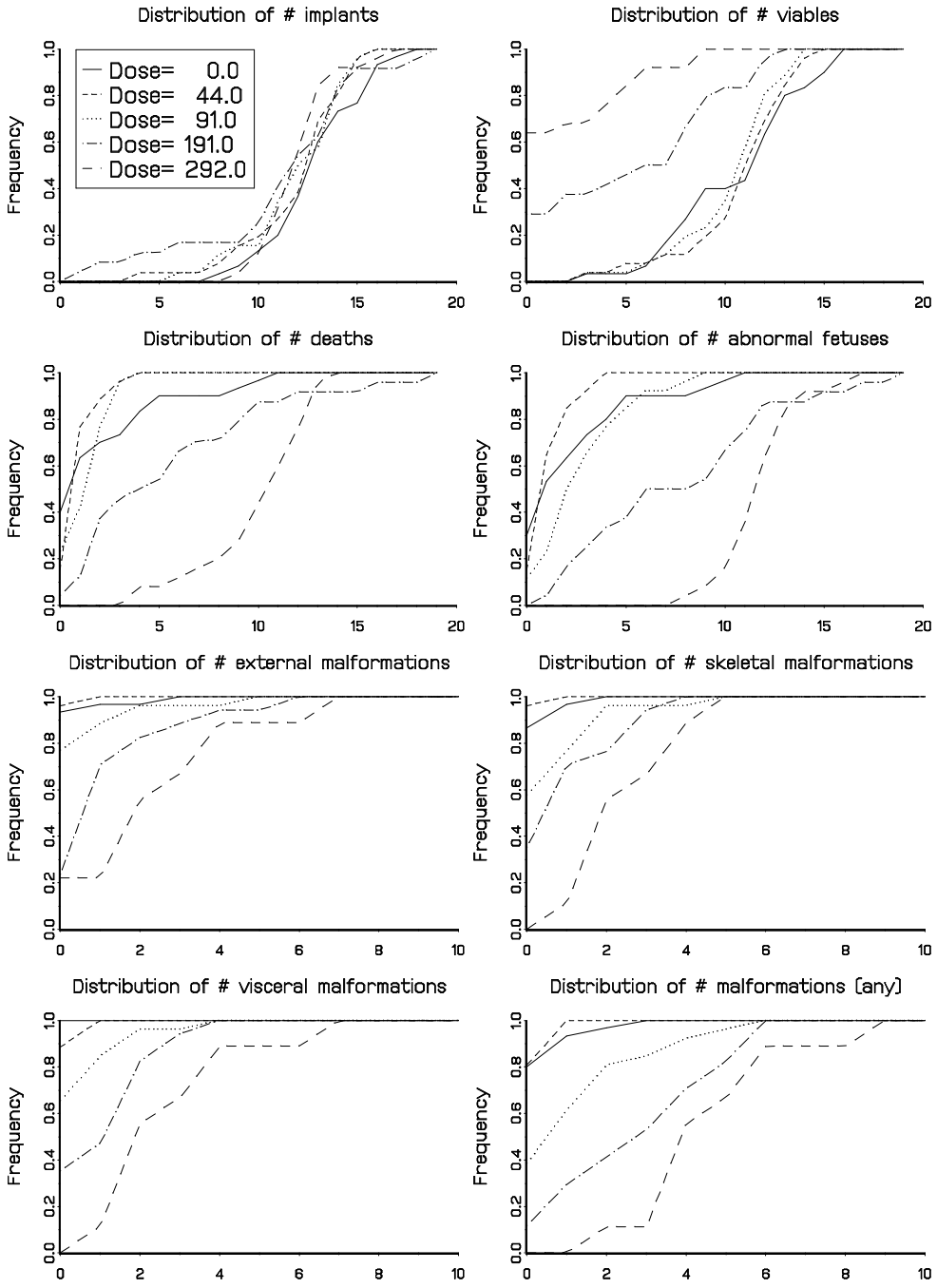

Figure 1.5: *Cumulative relative frequencies of the number of clusters representing some of the data in the DEHP study.*

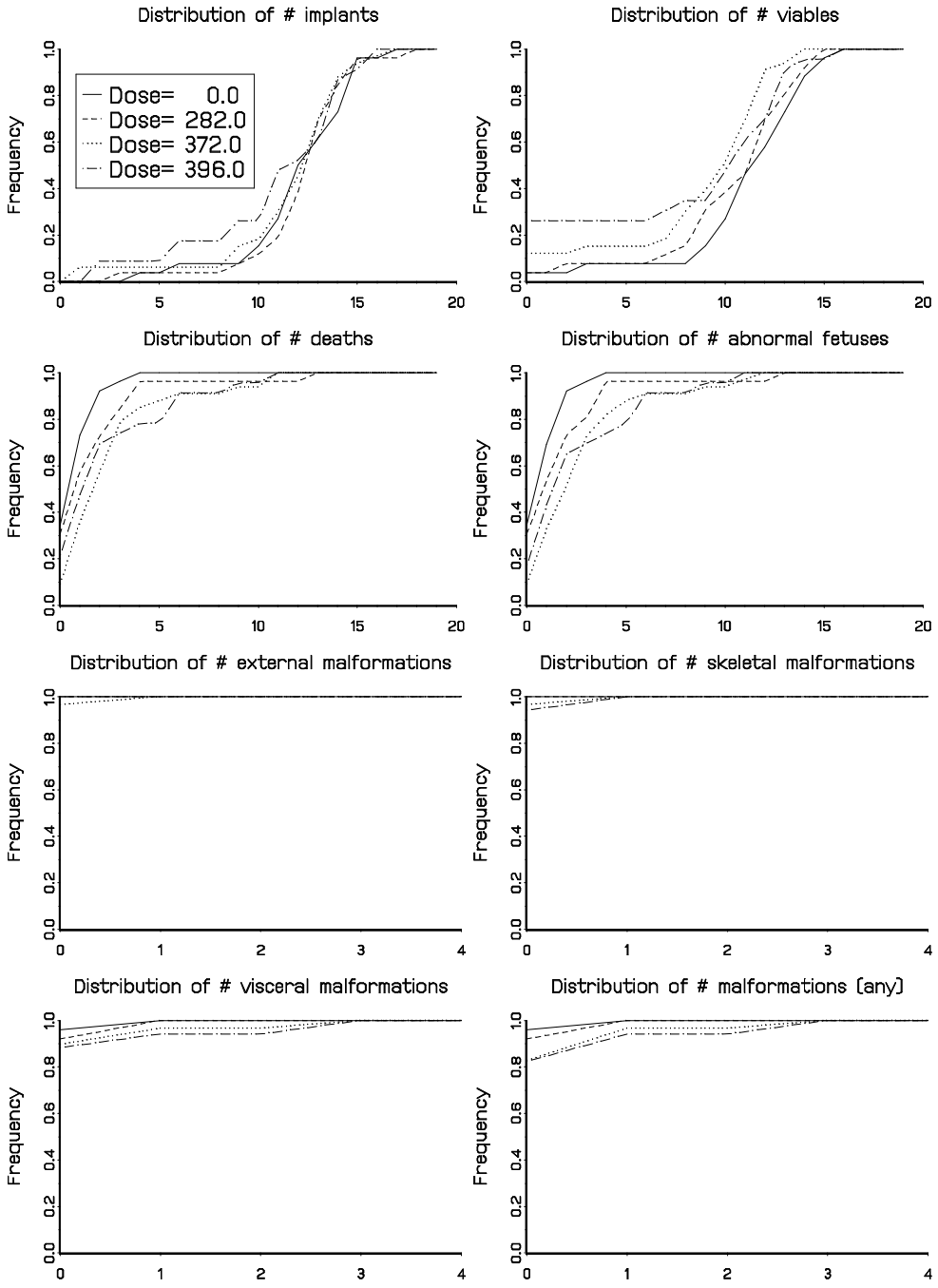

Figure 1.6: *Cumulative relative frequencies of the number of clusters representing some of the data in the THEO study.*

First, one can characterize the dose-response relationship, i.e., the dependence of a particular outcome (the number of deaths, the risk of a malformed fetus,...) on the dose which is administered to the dam. In Chapter 3 and 4, the focus is on dose-response modelling.

Secondly, a critically important topic is *quantitative risk assessment*. One of its objectives is to determine a safe dose of a toxic agent for humans. Quantitative risk assessment can be based on the dose-response curve. A number of different routes can be followed, thus leading to different approaches. This subject is discussed in Chapter 5 and 6. Furthermore, quantitative risk assessment can be performed via the NOAEL-safety factor approach, which is the topic of the following section.

1.6 The NOAEL-safety factor approach

Recently, regulatory agencies such as the Environmental Protection Agency (EPA) and the Food and Drug Administration (FDA) were basing the determination of a safe dose of an exposure on the No Observed Adverse Effect Level (NOAEL). The assumption made here is that if the dose administered to a dam is below some value (the threshold), then there will be no adverse effects on the fetuses of that dam (Williams and Ryan, 1997). The NOAEL is defined as the experimental dose level immediately below the lowest dose that produces a statistically or biologically significant increase in an adverse effect in comparison with the control. An "acceptably safe" daily dose for humans is then calculated by dividing the NOAEL by a safety factor (commonly 100 or 1000). In this way, sensitive subgroups of the population and extrapolation from animal experiments to human risk are taken into account. This safe daily dose is called reference dose by the EPA and allowable daily intake by the FDA.

Basing the determination of a safe dose on the NOAEL-safety factor approach, suffers from a number of serious statistical drawbacks (e.g., Leisenring and Ryan, 1992). First, the NOAEL of a specific toxic agent for a specified adverse event, depends very much on the experimental design. More specifically, results are highly sensitive to the number of doses, their spacing and the sample size. In larger experiments, the power to detect small differences is higher and as a consequence, this safe dose procedure leads to lower NOAELs than in the case of smaller studies. Secondly, this approach does not allow to calculate a measure of statistical variability. Also,

the actual risks at the NOAEL or reference dose may vary considerably from one developmental toxicity experiment to another. As a consequence, it is difficult to classify environmental and other agents based on the NOAELs. Another disadvantage of this approach is that the procedure is typically based on individual outcomes, instead of taking the complete process of fetal development into account. As a consequence, a NOAEL is calculated for every adverse event under investigation and the minimum NOAEL is taken for regulatory purposes.

Due to the disadvantages of using the NOAEL-safety factor approach, there is an increased interest in developing techniques of dose-response modelling when assessing safe dose levels. Rather new regulatory guidelines focus on the use of quantitative methods for risk assessment in analogy with cancer risk assessment (Environmental Protection Agency, 1991). The approach based on dose-response models, is more complicated as compared to the NOAEL procedure, but it benefits from a number of important advantages. This risk assessment procedure allows to add a measure of variability to the point estimation of a safe dose. Also, dose-response modelling is flexible to incorporate special features of the structure of developmental toxicity studies, such as hierarchical structure and multiple outcomes. Furthermore, fitting dose-response models enables to incorporate not only dose, but also duration, timing of exposure,...

1.7 Issues in risk assessment based on dose-response modelling

Because of the above mentioned statistical disadvantages of using the NOAEL-safety factor approach and because of the benefits of basing quantitative risk assessment on dose-response modelling, the latter approach will be considered in the present work. Dose-response modelling raises some highly relevant and interesting questions.

First, most species of laboratory animals are multiparous. The fetuses of a particular pregnant dam are genetically related and are developing under analogous conditions (e.g., the dose administered to the mother is the same). As a consequence, fetal outcomes are in general associated. In the literature, this phenomenon is called *litter effect*. Models that try to approximate the complex data generating mechanism of a developmental toxicity study, have to take the litter effect into account. The resulting data of developmental toxicity studies can be classified as

repeated measures, more specifically as *clustered data*. Analysis of clustered data must account for the extra variation in the responses of fetuses from the same dam, as compared to binomial and multinomial distributions. In this thesis, the clustered data are taken from the NTP studies and will illustrate the topics under investigation.

Secondly, besides the litter effect induced by the clustering of offspring within dams, dose-response modelling (and the NOAEL-safety factor approach) is complicated by the hierarchical structure of developmental toxicity studies. Outcomes include among others the number of fetal deaths and the number of malformed fetuses among the viables according to several types of malformation. Continuous outcomes (e.g., fetal weight) are also considered in developmental toxicity experiments. The ultimate purpose is the modelling of the number of viable fetuses, various malformation indicators, weight and clustering, as a function of exposure variables. Some attempts have been made in the literature to model the number of viable fetuses in a dam (or equivalently, the number of deaths) and the number of malformations jointly. Prominent are models of the Dirichlet-multinomial type (Chen and Kodell, 1989; Chen and Li, 1994; Zhu, Krewski and Ross, 1994). Efforts have also been directed towards modelling of multivariate malformation types (Lefkopoulou, Moore and Ryan, 1989; Geys, Molenberghs and Ryan, 1999) and to model malformation (categorical) and weight (continuous) simultaneously (Catalano and Ryan, 1992; Geys *et al.*, 1999b).

In dose-response modelling, a multitude of subproblems can be considered. Emphasis can be put on describing the dose-response relationship, on estimating a dose effect parameter, on testing the null hypothesis of no dose effect, on investigating the implications of model misspecification on dose effect, on determining a safe dose,... In the next section, an overview of the topics which are studied in this research, is given.

1.8 Topics discussed in this thesis

A fundamental question is which dose-response model should be used. Different types of models (marginal, random effects, conditional models) are available. An overview of existing models is given in Chapter 2. Some simplifications made in this research, as well as the implemented models, are also discussed in that chapter.

Related to the choice of the model, one needs to decide which estimation method should be taken. Estimation methods range from full likelihood to methods based on quasi-likelihood and generalized estimation equations. In this thesis, parameters are most often estimated using likelihood based methods. Besides point estimation, the focus can also be on performing a test of no dose effect, which is a crucial item in risk assessment. These two main issues are discussed in Chapter 3. In particular, the effect of misspecifying the parametric response model on the assessment of dose effect is investigated. When the model chosen to fit the data is inappropriate, several problems can occur. A few questions that will be addressed in that chapter are: Can a true dose effect be identified by a misspecified model and with which power? Do different test procedures, likelihood ratio and Wald tests in particular, behave similarly? Does this depend on the true values of the parameters of the underlying model? What is the impact of the magnitude of the correlation parameter?

The focus of Chapter 4 is on the behaviour of the likelihood ratio test statistic when a Bahadur model is fitted to the data. Bahadur (1961) proposed a now well-known although not very frequently used model for correlated binary data. In its general form, it combines marginal logits with pairwise and higher order correlations, to describe the joint response distribution. In many applications, the third and higher order correlations are set equal to zero. In that chapter, the implications of this simplification on the likelihood ratio test statistic are investigated.

A main issue in risk assessment deals with the determination of a safe dose of the toxic agent under investigation. This topic is addressed in Chapter 5 and 6. In the literature, different definitions and approaches leading to a safe dose and its lower bound, a so-called *virtually safe dose* (VSD), have been introduced. Moreover, a range of parametric dose-response models have been developed. The aim of Chapter 5 is to compare a number of VSDs and to study the effect of misspecifying the probabilistic model.

An important question in safe dose determination is whether risk assessment should be based on the fetus or the litter level. In Chapter 6, fetus and litter-based risks that properly account for cluster size are defined and compared for two models for clustered binary data. It is also studied how the hierarchical structure of non-viable implants and viable but malformed offspring can be incorporated. Risks based on a joint model for death and malformation are contrasted with risks based on an adverse event defined as either death or malformation. Another item is how

estimation of the cluster size distribution affects variance estimation.

Finally, some aspects of non-linear modelling are considered in Chapter 7. Rather than modelling the parameter of interest (or a link function of that parameter) as a linear function of dose, some power predictor is investigated. Under the null hypothesis of no dose effect, the regression parameters are unidentifiable. It is studied how Bayesian statistics can provide a procedure to test for no dose effect in case of power models.

Most of the results in this thesis have been submitted for publication and were published in statistical journals, as indicated in the reference list. Molenberghs, Declerck and Aerts (1998), Declerck, Aerts and Molenberghs (1998) and Aerts, Declerck and Molenberghs (1997) deal with Chapter 3, 4 and 5 respectively. Declerck, Molenberghs, Aerts and Ryan (1999) is accepted for publication and focuses on the contents of Chapter 6. The issue discussed in Chapter 7 is a topic of current research. Part of the results in this thesis have also been published in proceeding papers (Molenberghs, Declerck and Aerts, 1995; Declerck, Molenberghs and Aerts, 1997) and in a keynote paper (Molenberghs *et al.*, 1998).

Chapter 2

Models for clustered binary data

2.1 Some simplifications

The structure of developmental toxicity studies results in clustering of offspring within dams and as a consequence, in clustered data. The simplifications made in this research are discussed now. Notation of the number of implants, viables, malformations,... was given in Section 1.3.

Although continuous and discrete outcomes are recorded in developmental toxicity experiments, the analysis of fetal weight and possibly other continuous responses is not of primary interest in this thesis. Here, the focus is on binary death and malformation indicators.

In this type of studies, one generally considers multiple malformation indices. For example, external, skeletal and visceral malformations are examined in the NTP experiments. Here, malformation is analysed univariately in the sense that each type of malformation is modelled separately. Furthermore, a collapsed malformation outcome indicating if a fetus has at least one malformation is considered. Suppose Y_{ij} indicates whether the jth fetus in litter i is affected ($Y_{ij} = 1$) or not ($Y_{ij} = 0$) according to a particular type of malformation. Then, the number of such malformations of that dam is

$$Z_i = \sum_{j=1}^{n_i} Y_{ij}.$$

Covariates of interest are the dosing d_i that was administered to dam i, as well as the number of implants m_i and and the number of viable fetuses n_i.

Furthermore, exchangeability is assumed, which is a natural assumption in this kind of studies. This implies on the one hand that each fetus within a cluster has the

same marginal adverse event probability. On the other hand, it implies that within a cluster, the associations of any particular order are constant, i.e., the association between any pair of fetuses is equal, as well as between any triplet, any quartet,... of fetuses of the same dam.

Analysing data of the hierarchical structure expressed in Figure 1.1, one can opt for collapsing the number of dead fetuses r_i and the number of malformed fetuses z_i of a dam. Alternatively, a joint model for the number of deaths and for the number of malformations can be fitted. The effect of dose d_i on cluster i with m_i implants can be assessed by modelling the following joint distribution:

$$f(r_i, z_i|m_i, d_i) = f(r_i|m_i, d_i)f(z_i|r_i, m_i, d_i) = f(r_i|m_i, d_i)f(z_i|n_i, m_i, d_i). \qquad (2.1)$$

Often, it will be reasonable to assume that $f(z_i|n_i, m_i, d_i) = f(z_i|n_i, d_i)$. This assumption is made here. In cases where this is not acceptable, m_i can be included in the modelling strategy, such as in Catalano *et al.* (1993).

When both components of (2.1) are affected by dose, i.e., when the number of deaths (or equivalently the litter size) and the number of malformations show a dose effect, then the entire effect can be assessed by modelling both components. However, under a correctly specified model and assuming that different parameters describe these components, the likelihood factors into two parts that can be maximized separately. Even when the parameters are not disjoint, contemplating only one component does not result in bias, but merely in efficiency loss. Therefore, it is warranted, as a final simplification, to study malformation only, ignoring death. The latter simplification is considered in this thesis, except in Chapter 6.

A concise overview of existing models for clustered binary data is given in the following section. Some details of the models studied in this thesis, together with information about their likelihood implementation, are discussed then.

2.2 Overview of models for clustered binary data

There are several ways to handle clustering. While dose-response modelling is relatively straightforward in uncorrelated settings, it is less so in the clustered context. Of course, one can ignore the clustering altogether by treating the littermates as if they were independent. However, this will in general be a very strong assumption. Also, the litter effect issue can be avoided by modelling the probability of an affected cluster via e.g., a logistic regression model. Such models are generally too simplistic

but there is a multitude of models which do consider clustering. Several likelihood models for clustered binary data can be formulated, e.g., the beta-binomial model (Skellam, 1948; Kleinman, 1973), the Bahadur model (Bahadur, 1961), the correlated binomial models of Altham (1978), the double binomial model of Efron (1986), the multivariate Dale model (Molenberghs and Lesaffre, 1994), the folded logistic model of George and Bowman (1995) and the conditional exponential family model of Molenberghs and Ryan (1999).

In random effects models, the intracluster correlation is assumed to arise from natural heterogeneity in the parameters across litters. There are two routes to introduce randomness into the model parameters. Stiratelli, Laird and Ware (1984) assume the parameter vector to be normally distributed. Alternatively, Skellam (1948) introduced the beta-binomial model, in which the adverse event probability of any fetus of a particular cluster comes from a beta distribution. Hence, this model can also be viewed as a random effects model.

Marginal, random effects and conditional models for multivariate correlated binary data are discussed in Diggle, Liang and Zeger (1994). A thorough review of methods for the analysis of clustered binary data is given in Pendergast *et al.* (1996).

Due to the clustering that remains in the model, it is not obvious which model would be preferable. In the present work, attention is restricted to a selection of models for univariate clustered binary outcomes: the Bahadur model and the George-Bowman model (marginal models), the beta-binomial model (a random effects model) and the conditional model of Molenberghs and Ryan (1999). Details of these models are given in the following sections. As mentioned before, the estimation methods presented are likelihood based. Connections with second order generalized estimating equations (Liang, Zeger and Qaqish, 1992) are mentioned in Chapter 3.

2.3 The Bahadur model

The binary response Y_{ij} indicates if fetus j of cluster i has the adverse event under investigation. In order to be more specific, some type of malformation is considered here. The marginal distribution of Y_{ij} is Bernoulli with $E(Y_{ij}) = P(Y_{ij} = 1) \equiv \pi_{ij}$, i.e., the probability that the fetus is affected according to the specified malformation type.

In order to describe the association between binary outcomes, the pairwise probability $P(Y_{ij} = 1, Y_{ik} = 1) = E(Y_{ij}Y_{ik}) \equiv \pi_{ijk}$ has to be characterized. This "success probability" of two fetuses of the same dam can be modelled in terms of the two marginal probabilities π_{ij} and π_{ik}, as well as an association parameter.

Dealing with binary responses, common choices for the association parameter are the marginal odds ratio, the marginal correlation and the kappa coefficient (Agresti, 1990).

The marginal odds ratio is given by

$$\psi_{ijk} = \frac{\pi_{ijk}(1 - \pi_{ij} - \pi_{ik} + \pi_{ijk})}{(\pi_{ij} - \pi_{ijk})(\pi_{ik} - \pi_{ijk})}.$$

Using the odds ratio, the joint malformation probability π_{ijk} can be expressed in terms of two marginal malformation probabilities and an odds ratio association parameter, obtaining the expression of the bivariate Dale model (Dale, 1986):

$$\pi_{ijk} = \begin{cases} \dfrac{a_{ijk} - [a_{ijk}^2 - 4\psi_{ijk}(\psi_{ijk} - 1)\pi_{ij}\pi_{ik}]^{1/2}}{2(\psi_{ijk} - 1)} & \text{if } \psi_{ijk} \neq 1 \\ \pi_{ij}\pi_{ik} & \text{if } \psi_{ijk} = 1 \end{cases},$$

where $a_{ijk} = 1 - (1 - \psi_{ijk})(\pi_{ij} + \pi_{ik})$.

The marginal correlation coefficient assumes the form

$$\text{Corr}(Y_{ij}, Y_{ik}) \equiv \rho_{ijk} = \frac{\pi_{ijk} - \pi_{ij}\pi_{ik}}{[\pi_{ij}(1 - \pi_{ij})\pi_{ik}(1 - \pi_{ik})]^{1/2}}.$$

In terms of this association parameter, the joint probability π_{ijk} can then be written as

$$\pi_{ijk} = \pi_{ij}\pi_{ik} + \rho_{ijk}[\pi_{ij}(1 - \pi_{ij})\pi_{ik}(1 - \pi_{ik})]^{1/2}.$$

Hence, given the marginal correlation coefficient ρ_{ijk} and the univariate probabilities π_{ij} and π_{ik}, the pairwise probability π_{ijk} can easily be calculated. Other expressions for the associations and the pairwise probabilities can be found in Cox (1972). Bahadur (1961) and Cox (1972) consider the marginal correlation ρ_{ijk} to measure the association.

The first and second moments of the distribution have been specified. However, a likelihood-based approach requires the complete representation of the joint probabilities of the vector of binary responses in each litter. The full joint distribution $f(\boldsymbol{y})$ of $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^t$ is multinomial with a 2^{n_i} probability vector.

In this chapter, it is shown how the Bahadur, George-Bowman, beta-binomial and conditional models put restrictions on the 2^{n_i} joint probabilities of $\boldsymbol{Y}_i$.

A model first suggested by Bahadur (1961) and later by Cox (1972), henceforth called the Bahadur model, is described now. This model has been used by a few authors in the context of toxicological experiments (Altham, 1978; Kupper and Haseman, 1978). As a consequence, it is treated in this thesis as a representative of the marginal family. The Bahadur model gives a closed form expression for the joint distribution $f(\boldsymbol{y})$. The association between binary responses is expressed in terms of marginal malformation probabilities and correlation coefficients of second, third,... order.

Let

$$\varepsilon_{ij} = \frac{Y_{ij} - \pi_{ij}}{\sqrt{\pi_{ij}(1-\pi_{ij})}} \quad \text{and} \quad e_{ij} = \frac{y_{ij} - \pi_{ij}}{\sqrt{\pi_{ij}(1-\pi_{ij})}},$$

where y_{ij} is an actual value of the binary response variable Y_{ij}. Further, let $\rho_{ijk} = E(\varepsilon_{ij}\varepsilon_{ik})$, $\rho_{ijkl} = E(\varepsilon_{ij}\varepsilon_{ik}\varepsilon_{il})$, ..., $\rho_{i12\ldots n_i} = E(\varepsilon_{i1}\varepsilon_{i2}\ldots\varepsilon_{in_i})$.

Then, the general Bahadur model can be represented by $f(\boldsymbol{y}_i) = f_1(\boldsymbol{y}_i)c(\boldsymbol{y}_i)$, where

$$f_1(\boldsymbol{y}_i) = \prod_{j=1}^{n_i} \pi_{ij}^{y_{ij}} (1-\pi_{ij})^{1-y_{ij}}$$

and

$$c(\boldsymbol{y}_i) = 1 + \sum_{j<k} \rho_{ijk} e_{ij} e_{ik} + \sum_{j<k<l} \rho_{ijkl} e_{ij} e_{ik} e_{il} + \ldots + \rho_{i12\ldots n_i} e_{i1} e_{i2} \ldots e_{in_i}.$$

Thus, the probability mass function is the product of the independence model $f_1(\boldsymbol{y}_i)$ and the correction factor $c(\boldsymbol{y}_i)$. The factor $c(\boldsymbol{y}_i)$ can be viewed as a model for overdispersion.

As indicated in Section 2.1, the focus is on the special case of exchangeable littermates. This implies on the one hand that each fetus within a litter has the same malformation probability, i.e., $\pi_{ij} = \pi_i$ for $j = 1, \ldots, n_i$ and for $i = 1, \ldots, C$. On the other hand, it implies that within a litter, the associations of a particular order are constant, i.e., $\rho_{ijk} = \rho_{i(2)}$ for $j < k$, $\rho_{ijkl} = \rho_{i(3)}$ for $j < k < l$,..., $\rho_{i12\ldots n_i} = \rho_{i(n_i)}$, with $i = 1, \ldots, C$. Under exchangeability, the Bahadur model reduces to

$$f_1(\boldsymbol{y}_i) = \pi_i^{z_i}(1-\pi_i)^{n_i - z_i}$$

and

$$c(\boldsymbol{y}_i) = 1 + \sum_{r=2}^{n_i} \rho_{i(r)} \sum_{s=0}^{r} \begin{pmatrix} z_i \\ s \end{pmatrix} \begin{pmatrix} n_i - z_i \\ r - s \end{pmatrix} (-1)^{s+r} \lambda_i^{r-2s}, \tag{2.2}$$

with $\lambda_i = \sqrt{\pi_i/(1-\pi_i)}$. The probability mass function of Z_i, the number of malformations in cluster i, is given by

$$f(z_i) = \binom{n_i}{z_i} f(\boldsymbol{y}_i).$$

In addition setting all three- and higher-way correlations equal to zero, the probability mass function of Z_i simplifies further to:

$$\begin{aligned} f(z_i) \quad \equiv \quad & f(z_i|\pi_i, \rho_{i(2)}, n_i) = \binom{n_i}{z_i} \pi_i^{z_i} (1-\pi_i)^{n_i - z_i} \\ & \times \left[1 + \rho_{i(2)} \left\{ \begin{pmatrix} n_i - z_i \\ 2 \end{pmatrix} \frac{\pi_i}{1-\pi_i} - z_i(n_i - z_i) + \begin{pmatrix} z_i \\ 2 \end{pmatrix} \frac{1-\pi_i}{\pi_i} \right\} \right]. \end{aligned} \tag{2.3}$$

This very tractable expression of the Bahadur probability mass function is advantageous over other representations, such as an odds ratio representation for which no closed form solution for the joint distribution is possible. However, a drawback is the fact that the correlation between two responses is highly constrained when the higher order correlations are removed. Even when higher order parameters are included, the parameter space of marginal parameters and correlations is known to be of a very peculiar shape. Bahadur (1961) discusses restrictions on the correlation parameters. The second order approximation in (2.3) is only useful if it is a probability mass function. Bahadur indicates that the sum of the probabilities of all possible outcomes is one. However, depending on the values of π_i and $\rho_{i(2)}$, expression (2.3) may fail to be nonnegative for some outcomes. The latter results in restrictions on the parameter space which, in case of the second order approximation, are described by Bahadur (1961). From these, it can be deduced that the lower bound for $\rho_{i(2)}$ approaches zero as the cluster size increases. However, it is important to notice that also the upper bound for this correlation parameter is constrained. Indeed, even though it is one for clusters of size two, the upper bound varies between $1/(n_i - 1)$ and $2/(n_i - 1)$ for larger clusters. Taking a (realistic) litter of size 12, the upper bound is in the range $(0.09; 0.18)$. Kupper and Haseman (1978) present numerical

values for the constraints on $\rho_{i(2)}$ for choices of π_i and n_i. Restrictions for a specific version where a third order association parameter is included as well, are studied by Prentice (1988), while a more general situation is discussed in Chapter 4 of this thesis.

The marginal parameters π_i and $\rho_{i(2)}$ can be modelled using a composite link function. Since Y_{ij} is binary, the logistic link function for π_i is a natural choice. In principle, any link function, such as the probit link, the log-log link or the complementary log-log link, could be chosen. A convenient transformation of $\rho_{i(2)}$ is Fisher's z-transform. This leads to the following generalized linear regression relations

$$\begin{pmatrix} \ln(\frac{\pi_i}{1-\pi_i}) \\ \ln(\frac{1+\rho_{i(2)}}{1-\rho_{i(2)}}) \end{pmatrix} \equiv \boldsymbol{\eta}_i = X_i\boldsymbol{\beta}, \tag{2.4}$$

where X_i is a design matrix and $\boldsymbol{\beta}$ is a vector of unknown parameters. For example, a linear marginal logit model and a constant association $\rho_{i(2)} = \rho_{(2)}$ implies:

$$X_i = \begin{pmatrix} 1 & d_i & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_d \\ \beta_2 \end{pmatrix}. \tag{2.5}$$

Obviously, this model can be extended by changing the design matrix and the vector of regression parameters, such that the logit of π_i depends on dose via e.g., a quadratic or a higher order polynomial function. Also, the association parameter $\rho_{i(2)}$ can be modelled as some function of dose.

Denote the log-likelihood contribution of the ith cluster by $\ell_i = \ln f(z_i|\pi_i, \rho_{(2)}, n_i)$. The maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ is defined as the solution to the score equations $\boldsymbol{U}(\boldsymbol{\beta}) = \mathbf{0}$. The score function $\boldsymbol{U}(\boldsymbol{\beta})$ can be written as

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^{C} X_i^t (T_i^t)^{-1} L_i \tag{2.6}$$

where C is the number of clusters in the dataset,

$$T_i = \frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\Theta}_i} = \begin{pmatrix} \frac{\partial \eta_{i1}}{\partial \pi_i} & \frac{\partial \eta_{i2}}{\partial \pi_i} \\ \frac{\partial \eta_{i1}}{\partial \rho_{(2)}} & \frac{\partial \eta_{i2}}{\partial \rho_{(2)}} \end{pmatrix} = \begin{pmatrix} \frac{1}{\pi_i(1-\pi_i)} & 0 \\ 0 & \frac{2}{(1-\rho_{(2)})(1+\rho_{(2)})} \end{pmatrix},$$

$$L_i = \frac{\partial \ell_i}{\partial \boldsymbol{\Theta}_i} = \begin{pmatrix} \frac{\partial \ell_i}{\partial \pi_i} \\ \frac{\partial \ell_i}{\partial \rho_{(2)}} \end{pmatrix} \text{ and}$$

$$\boldsymbol{\Theta}_i = (\pi_i, \rho_{(2)})^t, \text{ the set of natural parameters.}$$

A Newton-Raphson algorithm is used to obtain the maximum likelihood estimates $\hat{\boldsymbol{\beta}}$. An estimate of the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ is obtained from the observed information matrix at maximum.

When including higher order correlations, implementing the score equations and the observed information matrices becomes increasingly cumbersome. While the functional form (2.6) does not change, the components T_i and L_i become fairly complicated. Therefore, analytical expressions are only used up to the three-way Bahadur model. For higher orders, the numerical optimizer OPTMUM of GAUSS is employed. Fisher's z transform is applied to all correlation parameters $\rho_{i(r)}$. The design matrix X_i is extended in a straightforward fashion. Unfortunately, fitting a higher order Bahadur model, whether through numerical or analytical maximization, is not straightforward, due to increasingly complex restrictions on the parameter space.

Observing that in the studies considered, interest is restricted to the marginal mean function and the pairwise association parameter, one can replace a full likelihood approach by estimating equations where only the first two moments are modelled and working assumptions are adopted about third and fourth order moments. A thorough treatment is found in Liang, Zeger and Qaqish (1992). Obviously, an important special form for these working assumptions is given by setting the higher order parameters equal to zero, thereby avoiding the need for moment-based estimation of nuisance parameters. Consistent point estimates are supplemented with *robust* standard errors (following from the sandwich estimator), rather than with purely model-based (or naive) standard errors. Often, point estimates differ only slightly from their likelihood counterparts, while test statistics may change considerably. This point will be illustrated in Chapter 3.

2.4 The George-Bowman model

Besides the Bahadur model, another marginal model is introduced here. George and Bowman (1995) propose a model for the analysis of exchangeable binary data. The probability mass function for the number of malformations Z_i in litter i consisting

of n_i viable fetuses, is presented as:

$$f(z_i|\lambda_{i,z_i},\lambda_{i,z_i+1},\ldots,\lambda_{i,n_i},n_i)=\binom{n_i}{z_i}\sum_{\ell=0}^{n_i-z_i}(-1)^\ell\binom{n_i-z_i}{\ell}\lambda_{i,z_i+\ell}, \tag{2.7}$$

in which

$$\lambda_{i,k}=\begin{cases} P(Y_{i1}=1,Y_{i2}=1,\ldots,Y_{ik}=1) & \text{if } k=1,\ldots,n_i,\\ 1 & \text{if } k=0.\end{cases}$$

As a consequence, the parameter $\lambda_{i,k}$ can be interpreted as the probability that in litter i, all fetuses in a set of k exhibit the adverse event under consideration. The mean of the number of malformed fetuses and the second order correlation between two responses of the same litter can be expressed in terms of $\lambda_{i,k}$ parameters:

$$E(Z_i)=\sum_{j=1}^{n_i}E(Y_{ij})=n_iP(Y_{ij}=1)=n_i\lambda_{i,1}$$

and

$$\begin{aligned}\text{Corr}(Y_{ij},Y_{ik}) &= \frac{E(Y_{ij}Y_{ik})-E(Y_{ij})E(Y_{ik})}{E(Y_{ij}^2)-(E(Y_{ij}))^2}\\ &= \frac{P(Y_{ij}=1,Y_{ik}=1)-P(Y_{ij}=1)P(Y_{ik}=1)}{P(Y_{ij}=1)-P(Y_{ij}=1)^2}\\ &= \frac{\lambda_{i,2}-\lambda_{i,1}^2}{\lambda_{i,1}(1-\lambda_{i,1})}.\end{aligned}$$

George and Bowman (1995) also give expressions for higher order moments of Z_i and for higher order correlations.

Under independence of the n_i responses of litter i,

$$\lambda_{i,k}=P(Y_{i1}=1)\ldots P(Y_{ik}=1)=\lambda_{i,1}^k$$

and (2.7) can be written as:

$$\begin{aligned}f(z_i|\lambda_{i,z_i},\lambda_{i,z_i+1},\ldots,\lambda_{i,n_i},n_i) &= \binom{n_i}{z_i}\sum_{\ell=0}^{n_i-z_i}(-1)^\ell\binom{n_i-z_i}{\ell}\lambda_{i,1}^{z_i}\lambda_{i,1}^{\ell}\\ &= \binom{n_i}{z_i}\lambda_{i,1}^{z_i}\sum_{\ell=0}^{n_i-z_i}\binom{n_i-z_i}{\ell}(-\lambda_{i,1})^\ell\\ &= \binom{n_i}{z_i}\lambda_{i,1}^{z_i}(1-\lambda_{i,1})^{n_i-z_i}.\end{aligned}$$

Hence, under independence, the George-Bowman model with parameters λ_{i,z_i}, λ_{i,z_i+1}, $\ldots$, λ_{i,n_i} and n_i, reduces to a binomial model with parameters n_i and $\lambda_{i,1}$.

George and Bowman focus attention on the so-called *folded logistic* parameterization:

$$\lambda_{i,z_i+\ell}(\boldsymbol{\beta}) = \frac{2}{1+\exp\left[-X_i\boldsymbol{\beta}\ln(z_i+\ell+1)\right]} \tag{2.8}$$

where $X_i = (1, d_i)$ and $\boldsymbol{\beta} = (\beta_0, \beta_d)^t$. Hence, (2.8) can be rewritten as:

$$\lambda_{i,z_i+\ell}(\boldsymbol{\beta}) = \frac{2}{1+(z_i+\ell+1)^{-\beta_0-\beta_d d_i}}.$$

When the responses of a litter are independent, it has been shown that the "general" George-Bowman model (2.7) reduces to the binomial model. However, it turns out that the "specific" George-Bowman model with the folded logistic parameterization does not simplify to the binomial model in this case.

The maximum likelihood estimates of the George-Bowman model with this specific parameterization are found by the Newton-Raphson algorithm. George and Bowman prove that $X_i\boldsymbol{\beta} < 0$ is necessary and sufficient in order to have a valid probability mass function. In this thesis, limited attention is payed to this model.

2.5 The beta-binomial model

Rather than modelling marginal functions directly, a popular approach is to assume a random effects model in which each litter has a random parameter (vector). Skellam (1948), Kleinman (1973) and Williams (1988) assume the malformation probability P_i of any fetus in litter i to come from a beta distribution with mean π_i and conditional on P_i, the number of malformations Z_i in the ith cluster follows a binomial distribution. This leads to the well-known beta-binomial model. In a litter of size n_i, the probability mass function of Z_i is expressed by

$$f(z_i \mid \pi_i, \rho_i, n_i) = \binom{n_i}{z_i} \frac{B(\pi_i(\rho_i^{-1}-1)+z_i, (1-\pi_i)(\rho_i^{-1}-1)+n_i-z_i)}{B(\pi_i(\rho_i^{-1}-1), (1-\pi_i)(\rho_i^{-1}-1))}, \tag{2.9}$$

where $B(.,.)$ denotes the beta function. The only association parameter of this model is ρ_i, which is the correlation between two binary responses of litter i. The higher order correlations of the beta-binomial model can be expressed as a function of the mean malformation probability π_i and ρ_i. The association in both the beta-binomial and the Bahadur model is expressed by means of the intraclass correlation. It turns

out that both models have the same first and second moments. As a consequence, the parameter ρ_i of the beta-binomial model equals $\rho_{i(2)}$ of the Bahadur model. The parameters π_i and ρ_i of the beta-binomial model have a marginal interpretation and therefore, they are the parameters in the derived marginal model as well. This results in similarities between the beta-binomial and marginal models, such as the Bahadur model.

It can be shown (Williams, 1975) that the contribution of the ith cluster to the log-likelihood, $\ln f(z_i|\pi_i, \rho_i, n_i) \equiv \ell_i$, can be written as

$$\ell_i = \sum_{r=0}^{z_i-1} \ln\left(\pi_i + \frac{r\rho_i}{1-\rho_i}\right) + \sum_{r=0}^{n_i-z_i-1} \ln\left(1-\pi_i + \frac{r\rho_i}{1-\rho_i}\right) - \sum_{r=0}^{n_i-1} \ln\left(1 + \frac{r\rho_i}{1-\rho_i}\right), \tag{2.10}$$

with $i = 1, ..., C$. It follows from (2.10) that if the association parameter ρ_i equals zero, then the beta-binomial model reduces to the logistic regression model.

Assuming the same generalized linear regression relations (2.4) and (2.5) for π_i and ρ_i, the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ is the solution to $\boldsymbol{U}(\boldsymbol{\beta}) = \mathbf{0}$ with the score function for $\boldsymbol{\beta}$ defined as in (2.6).

Kupper and Haseman (1978) compare the Bahadur model to the beta-binomial model. They conclude that the models perform similarly in three clustered data experiments, whereas they both outperform the (naive) binomial model.

2.6 A conditional model

Molenberghs and Ryan (1999) propose a likelihood-based conditional model for multiple clustered binary outcome variables. This model is based on the multivariate exponential family model as proposed by Cox (1972). Related work is done by Fitzmaurice, Laird and Tosteson (1999). The conditional model of Molenberghs and Ryan is described here for the special case of a univariate clustered outcome and restricting the association to pairwise effects (analogous to the second order approximation of the Bahadur model). The focus is again on exchangeability.

This conditional model is used in this thesis to describe several adverse events, e.g., the number of fetal deaths R_i of dam i. Molenberghs and Ryan consider Y_{ij} to be equal to 1 if the jth fetus in cluster i exhibits the adverse event and -1 otherwise. This coding is preferred above the 1 and 0 coding, since it provides a parameterization that more naturally leads to desirable properties when the roles of

success and failure are reversed (Molenberghs and Ryan, 1999). They propose the following joint density function of the vector of responses $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{im_i})^t$:

$$f(\boldsymbol{y}_i|\psi_i^*, \phi_i^*, m_i) = \exp\left\{\psi_i^* \sum_{j=1}^{m_i} y_{ij} + \phi_i^* \sum_{j<j'} y_{ij}y_{ij'} - A_i^*\right\}, \tag{2.11}$$

where A_i^* is a constant such that (2.11) is a density. Representing the number of dead fetuses in cluster i by R_i, expression (2.11) can be rewritten as

$$f(\boldsymbol{y}_i|\psi_i^*, \phi_i^*, m_i) = \exp\left\{\psi_i^*(2r_i - m_i) + \phi_i^*\left[\binom{m_i}{2} - 2r_i m_i + 2r_i^2\right] - A_i^*\right\}. \tag{2.12}$$

After absorbing the constant terms into the normalizing constant and after a simple reparameterization ($\psi_i = 2\psi_i^*$ and $\phi_i = 2\phi_i^*$), one obtains from (2.12):

$$f(\boldsymbol{y}_i|\psi_i, \phi_i, m_i) = \exp\left\{\psi_i r_i - \phi_i r_i(m_i - r_i) - A_i\right\}. \tag{2.13}$$

From formula (2.13), the probability mass function of the number of fetal deaths follows:

$$f(r_i|\psi_i, \phi_i, m_i) = \binom{m_i}{r_i} \exp\left\{\psi_i r_i - \phi_i r_i(m_i - r_i) - A_i\right\}. \tag{2.14}$$

Hence, the normalizing constant A_i can be written as:

$$A_i = \ln\left\{\sum_{r_i=0}^{m_i} \binom{m_i}{r_i} \exp\left\{\psi_i r_i - \phi_i r_i(m_i - r_i)\right\}\right\} \equiv A(\psi_i, \phi_i, m_i).$$

Based on (2.14), the conditional logit for a dead fetus given the number of deaths in the group of remaining fetuses can be written as a linear function of ψ_i and ϕ_i:

$$\text{logit}[P(\text{fetus } j \text{ dead} \mid r_i \text{ of the other fetuses also dead})] = \psi_i - \phi_i(m_i - 2r_i - 1),$$

where $j = 1, \ldots, m_i$. This implies that if the number of implants is odd, then the parameter ψ_i equals the logit for a dead fetus given that one half of the remaining fetuses are dead as well. Also, (2.14) results in

$$\psi_i = \frac{1}{m_i} \ln\left(\frac{P(R_i = m_i)}{P(R_i = 0)}\right). \tag{2.15}$$

From (2.14) and (2.15), it follows that the parameter $\psi_i = 0$ if and only if the distribution of R_i is symmetric around $m_i/2$. Furthermore, it can be shown that the parameter ϕ_i is one half of the log odds ratio for a pair of fetuses given the number

of deaths in the remaining group of fetuses. Thus clearly, the parameters in the model of Molenberghs and Ryan have a conditional interpretation.

A special case of the conditional model (2.14) is obtained when the association parameter $\phi_i = 0$:

$$f(r_i|\psi_i, m_i) = \binom{m_i}{r_i} \exp(\psi_i r_i - A_i). \tag{2.16}$$

Let

$$\psi_i = \ln\left\{\frac{\omega_i}{1-\omega_i}\right\}.$$

Then, formula (2.16) can be reexpressed as

$$\begin{aligned} f(r_i|\psi_i, m_i) &= \binom{m_i}{r_i} \left\{\frac{\omega_i}{1-\omega_i}\right\}^{r_i} \exp(-A_i) \\ &= \binom{m_i}{r_i} \omega_i^{r_i}(1-\omega_i)^{m_i - r_i}. \end{aligned}$$

Hence, if in the conditional model with parameters ψ_i, ϕ_i and m_i, the parameter ϕ_i is set equal to zero, then this model reduces to the logistic regression model with parameters m_i and

$$\omega_i = \frac{1}{1+\exp(-\psi_i)}.$$

Furthermore, one notices from (2.14) that positive and negative values of ϕ_i correspond to overdispersion and underdispersion respectively. Also, there are no restrictions on the parameter space of the conditional model, even in case of underdispersion (Molenberghs and Ryan, 1999).

The parameters ψ_i and ϕ_i can be modelled as

$$\begin{pmatrix} \psi_i \\ \phi_i \end{pmatrix} = X_i \boldsymbol{\beta},$$

with X_i and $\boldsymbol{\beta}$ as in (2.5). Estimation of these model parameters can easily be carried out using maximum likelihood techniques. Grouping the summary statistics in

$$\boldsymbol{W}_i = \begin{pmatrix} R_i \\ -R_i(m_i - R_i) \end{pmatrix},$$

the contribution of the ith cluster to the log-likelihood is given by $\ell_i = \boldsymbol{w}_i^t X_i \boldsymbol{\beta} - A_i$, whence the score function becomes

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^{C} X_i^t(\boldsymbol{w}_i - E(\boldsymbol{W}_i)).$$

The expectation of R_i/m_i, the marginal death probability in a cluster of m_i implants, is clearly a (non-linear) function of m_i:

$$\pi_i = E\left(\frac{R_i}{m_i}\right) = \frac{\sum_{r_i=0}^{m_i} r_i \binom{m_i}{r_i} \exp\{(\beta_0 + \beta_d d_i) r_i - \beta_2 r_i (m_i - r_i)\}}{\sum_{r_i=0}^{m_i} m_i \binom{m_i}{r_i} \exp\{(\beta_0 + \beta_d d_i) r_i - \beta_2 r_i (m_i - r_i)\}}. \quad (2.17)$$

Methods similar to those of Cox and Wermuth (1994) could be invoked to develop approximate expressions for the marginal means and odds ratios. Because the model is conditional in nature, the marginal parameter (2.17) does not simplify in general. As a consequence, the conditional model implies a natural dependence of π_i on the number of implants, in contrast to marginal models. Furthermore, only when the clustering parameters are equal to zero, the conditional model, the Bahadur model and the beta-binomial model reduce to logistic regression. In Section 2.4, it has been shown that the general expression of the George-Bowman model reduces to the logistic regression model too if the fetuses of a cluster are independent. The previous discussion implies that the parameters of the conditional model are not directly comparable to their counterparts in the Bahadur, George-Bowman and beta-binomial models. Molenberghs and Ryan (1999) consider simple linear models of the form $\psi_i = \beta_0 + \beta_d d_i$ and show that a score test provides a flexible way of testing a broad class of hypotheses. This linear model will generally be too simple in the context of dose-response modelling. So, one could consider instead quadratic or power models (of which some issues are discussed in Chapter 7), which can still be fitted using the methods described in Molenberghs and Ryan (1999). They devote most attention to constant association models, i.e., $\phi_i = \phi$. However, it is possible to let the association depend on dose as well (Claeskens and Aerts, 1999; Geys, Molenberghs and Ryan, 1999; Ryan and Molenberghs, 1999). More details about model properties and inference of the conditional model can be found in Molenberghs and Ryan (1999) and Ryan and Molenberghs (1999).

2.7 The logistic regression model

As indicated in the previous section, treating the littermates as being independent is in general a very strong assumption. Nevertheless, the logistic regression model is

considered in this thesis, enabling the comparison of the results obtained from the binomial model and from models taking the clustering into account.

Chapter 3

Implications of misspecifying the likelihood on dose effect assessment

In this chapter, the effect of misspecifying the parametric response model on dose effect estimation and hypothesis testing is investigated. When the model chosen to fit the data is incorrect, several problems can occur. It is plausible that the parameter in the incorrect model included to capture the effect of dose will fail to do so or will do it only partly, resulting in an apparent reduced effect. On the other hand, the dose parameter in the incorrect alternative model might capture not only part of the true dose effect, but also other effects that were misspecified or omitted. This implies that no clear prediction of the overall effect of misspecification can be done and quantitative assessment is needed.

Assessment of dose effect is done on the one hand by estimating the effect of dose, mainly via maximum likelihood techniques. On the other hand, the null hypothesis of no dose effect is tested via likelihood ratio and Wald statistics. Other tests, such as score tests, are not considered here.

Classical theoretical tools like bias computation and asymptotic efficiency calculations are less attractive since the models are not nested. As an alternative strategy, the models described in Chapter 2 are compared by asymptotic calculations as suggested by Rotnitzky and Wypij (1994). This technique is adapted here to compute and compare the large sample (asymptotic) values of test statistics. The true model is in turn chosen to be Bahadur, beta-binomial and conditional. In each case, the test statistics are computed from the three models, thus including one correct and two incorrect models. The results are reported in Section 3.1. To investigate whether the

conclusions carry over to samples encountered in real applications, a small sample simulation study is performed (Section 3.2). Emphasis is on rejection probabilities of the null hypothesis of no dose effect for a range of dose parameters. Finally, in Section 3.3, the results of real data analyses of the NTP data are examined for their agreement with the findings of the simulated data. To supplement insight in the restrictions of the parameter space of the Bahadur model, the GEE2 version of the Bahadur model is added to the discussion. In order to investigate the characteristics of the George-Bowman model, this marginal model is also fitted to the NTP data.

3.1 Asymptotic study

If the correct model is fitted to a dataset of finite sample size, it is well known that both Wald and likelihood ratio test statistics have the same asymptotic χ^2 distribution under the null model and under contiguous alternatives (Serfling, 1980). For fixed alternatives, the picture is less clear and when fitting the incorrect model, asymptotic theory is not always available. In order to get asymptotic information on the effect of model misspecification on the assessment of dose effect, the ideas of Rotnitzky and Wypij (1994) are followed here. Indeed, in order to compute the asymptotic bias or the asymptotic covariance matrix of the maximum likelihood estimator, an artificial sample can be constructed, where each possible realization is weighted according to its true probability. In this case, one needs to consider all realizations of the form (d_i, n_i, z_i), i.e., each combination of dose d_i, number of viable fetuses n_i and number of malformations z_i, is considered and is assigned a weight equal to the probability $f(d_i, n_i, z_i)$ that this combination occurs in the underlying model. Thus, one has to specify: (1) $f(d_i)$, the relative frequencies of the dose groups, as prescribed by the design; (2) $f(n_i|d_i)$, the probability with which a litter size can occur, possibly dependent on the dosing (here, it is assumed that $f(n_i|d_i) = f(n_i)$) and (3) $f(z_i|n_i, d_i)$, the actual model probabilities. In this research, the technique introduced by Rotnitzky and Wypij is adapted to compute "asymptotic" values of test statistics. Under each model, Bahadur (Bah), beta-binomial (BB) or conditional (Cond), the value of the Wald (W) and likelihood ratio (LR) test statistic is computed for this artificial sample. These so-called *population test statistics* can be interpreted as follows. Suppose a very large sample of size N is obtained, in which a given cluster occurs exactly with the multiplicity predicted by

the model, then the corresponding test statistics are N times the population values. Therefore, this method is useful to investigate the large sample effect of choosing an incorrect model.

Throughout this and the next section, it is assumed that there are four dose groups, with one control group ($d_i = 0$) and three active groups ($d_i = 0.25, 0.5, 1$). The number n_i of viable fetuses per cluster is assumed to follow a local linear smoothed version of the frequency distribution proposed by Kupper *et al.* (1986), which is considered representative of that encountered in actual experimental situations. Least squares cross-validation has been used to choose the bandwidth (Aerts, Augustyns and Janssen, 1997). The absolute and relative frequency distribution used by Kupper *et al.*, as well as the smoothed relative frequencies, are presented in Table 3.1.

The NTP data are analysed in order to obtain realistic ranges for parameters (Section 3.3). The intercepts for Bahadur and beta-binomial correspond to a low baseline malformation rate: a value of -5.5 (-4.5) corresponds to 0.4% (1%). For the conditional model, the baseline malformation rate is a function of both intercept and association parameter. Only the intercepts closest to zero are used for the asymptotic study. A range of dose effects is considered. Parameter settings are summarized in Table 3.2. A transformed correlation of 0.1 (correlation of about 0.05) in the Bahadur model must be interpreted as considerable, given the restrictions on the association parameter. Although one usually finds higher correlations with the beta-binomial and higher associations in the conditional model, it is opted also for an association parameter of 0.1 in this case since otherwise, the Bahadur model becomes prohibitively difficult to fit.

Figures 3.1, 3.2 and 3.3 show population values for test statistics (likelihood ratio and Wald for Bahadur, beta-binomial and conditional model), arising from (1) choosing the Bahadur, beta-binomial or conditional model as the true one, (2) choosing the true association parameter to be 0.0 or 0.1. The picture obtained when the true model is Bahadur is exactly the one obtained for beta-binomial when the correlation is zero. This is to be expected because here both true models reduce to ordinary logistic regression. Although the same holds for the conditional model, a difference is seen because a different intercept is used. Further, the parameters have a conditional rather than a marginal meaning, which is reflected in the asymptotic covariance matrix. Although W(Bah) and W(BB) are the same, this is not true for

Table 3.1: Absolute and relative frequencies of the number of viable fetuses.

Number of viables	absolute frequency	relative frequency	smoothed relative frequency
1	2	0.0038	0.0046
2	3	0.0057	0.0057
3	4	0.0076	0.0099
4	9	0.0172	0.0139
5	8	0.0153	0.0147
6	6	0.0115	0.0148
7	10	0.0191	0.0225
8	20	0.0382	0.0321
9	19	0.0364	0.0475
10	38	0.0727	0.0766
11	64	0.1224	0.1179
12	82	0.1568	0.1529
13	93	0.1778	0.1605
14	73	0.1396	0.1424
15	58	0.1109	0.0975
16	19	0.0364	0.0542
17	12	0.0229	0.0207
18	1	0.0019	0.0086
19	2	0.0038	0.0030
	523	1	1

Table 3.2: Parameter settings.

Parameter	Bahadur model beta-bin. model	cond. model
intercept β_0	$-5.5; -4.5$	$-3.5; -2.5$
dose effect β_d	0.0(0.5)8.0	0.0(0.5)5.5
association β_2	$0.0; 0.1$	$0.0; 0.1$

the corresponding likelihood ratio test statistic. When the true model is correlated, the three pictures separate, but the true beta-binomial and true Bahadur plots cannot be distinguished by visual inspection only.

All population test statistics are very close when the true dose effect is small, i.e., until about 2 when the true model is Bahadur (or beta-binomial) and until about 1 when the true model is conditional. For higher dose effects, one observes that

$$\text{LR(Bah)} \gg \text{LR(Cond)} > \text{LR(BB)}$$

and

$$\text{W(Bah)} \geq \text{W(BB)} \gg \text{W(Cond)}.$$

Curves clearly standing apart for large dose effects, are LR(Bah) and W(Cond). This holds *regardless of the model used to generate the data.* The LR(Bah) tends to be higher because the higher order correlations in the Bahadur model are set equal to zero. Indeed, this implies the likelihood of the null model to be *much lower* than when higher order correlations are allowed. The precise quantification of this statement is the subject of Chapter 4.

For the conditional model, one should bear in mind that all parameters, including the dose effect parameter, are conditional in nature. A marginal dose effect is likely to depend in a complex way on the model parameters. Since the Wald test is known to depend on the particular parameterization (in contrast to likelihood ratio and score tests), it might be a less relevant measure, in particular for conditional models. It will be illustrated in Section 3.3 that the correlation between $\hat{\beta}_d$ and $\hat{\beta}_2$ is much larger in the conditional model than in the other models.

For $\beta_2 = 0$, LR(Bah)>W(Bah) except for small to moderate dose effects, while W(BB)>LR(BB) for all dose effects in the range considered (but a cross-over seems to appear for higher dose effects) and clearly LR(Cond)≫W(Cond). For $\beta_2 = 0.1$, the dominance of the LR(Bah) to the corresponding Wald(Bah) statistic is more pronounced. The cross-over for the beta-binomial model already occurs for moderately high dose effects.

Finally, all Wald tests show a non-monotone trend, an aberrant behaviour in agreement with Hauck and Donner (1977). These authors show that, in the context of testing for a single parameter in a logistic model, Wald's test statistic decreases to zero as the distance between the parameter estimate and the null value increases (for any fixed sample size). Likelihood ratio test statistics all increase with increasing dose.

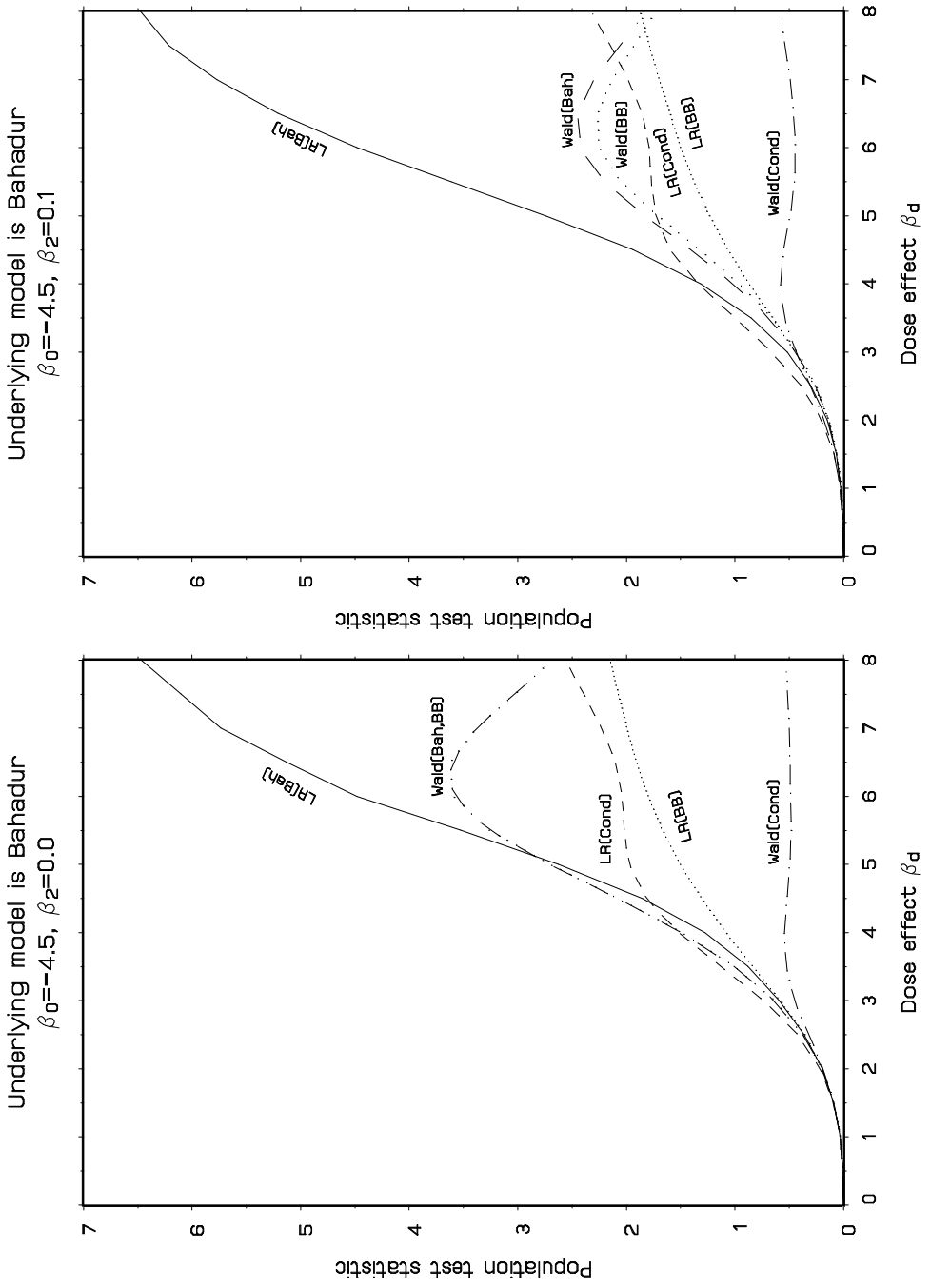

Figure 3.1: *Population values for likelihood ratio and Wald test statistics when the underlying model is Bahadur.*

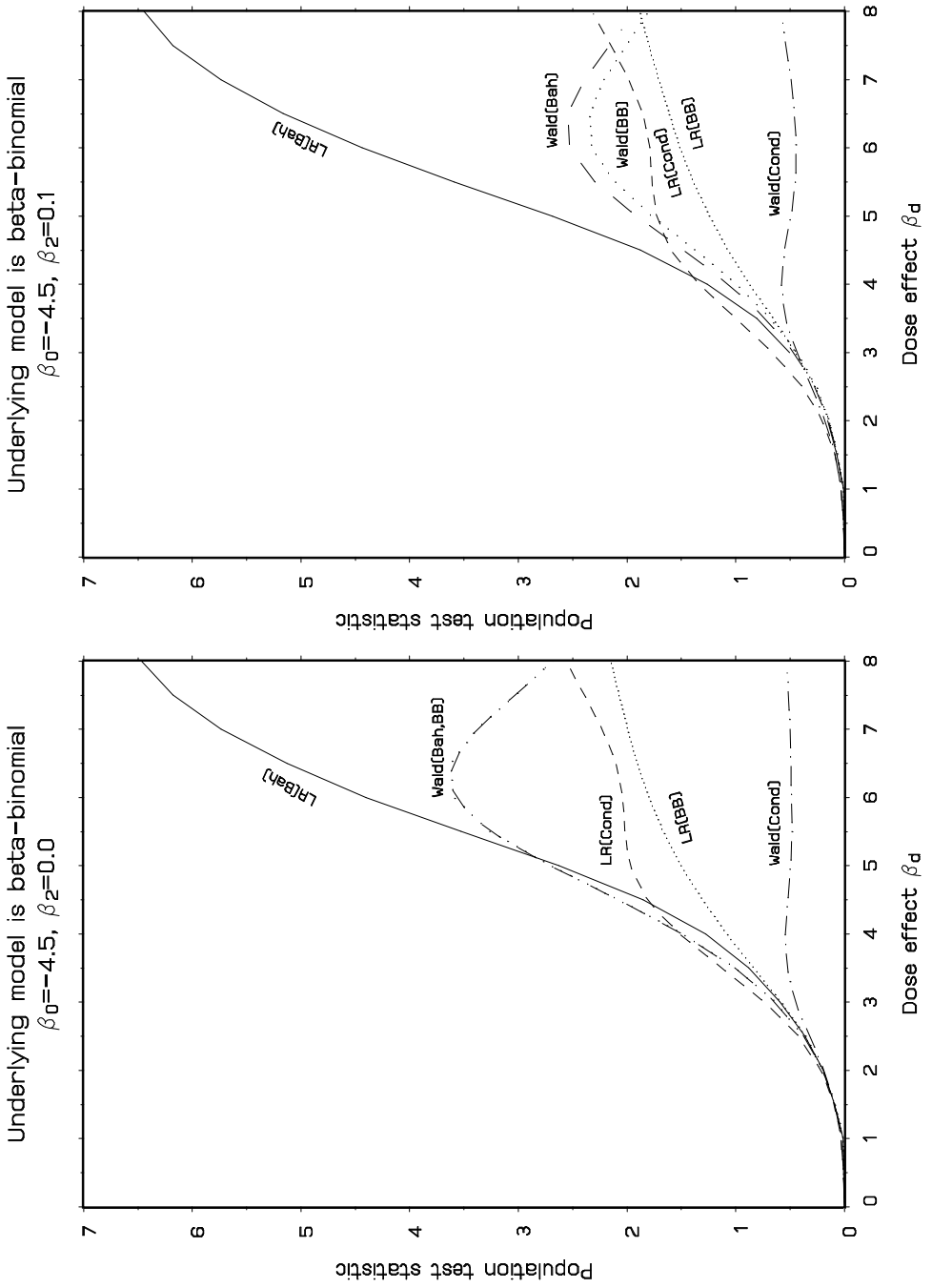

Figure 3.2: *Population values for likelihood ratio and Wald test statistics when the underlying model is beta-binomial.*

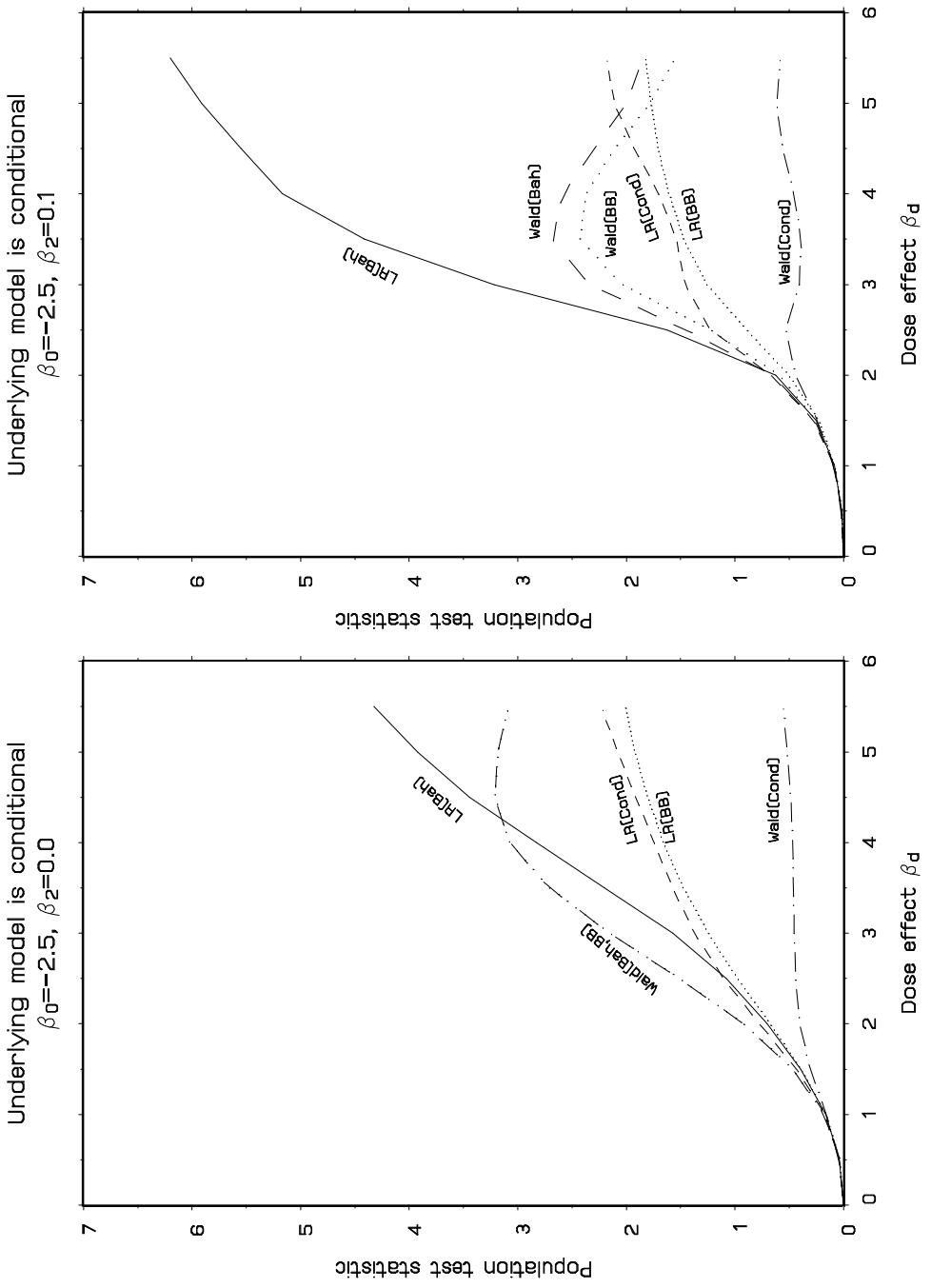

Figure 3.3: *Population values for likelihood ratio and Wald test statistics when the underlying model is conditional.*

3.2 Small sample simulations

A small sample simulation study is performed to supplement the information obtained from the asymptotic study. While in both cases a known data generating mechanism is used, a small sample simulation enables the assessment of whether realistic levels of random noise in the data either moderate or reinforce the effects of model misspecification on the operational characteristics of dose effect tests.

The parameter settings of the asymptotic study are used again (Table 3.2). For each parameter setting, 500 datasets were generated as follows. For a given dose level d, using the asymptotic method, the probability for each possible realization $f_i = f(n_i, z_i|d)$ $(i = 1, \ldots, I)$ is computed. Then, the cumulative probabilities

$$g_i = \sum_{j=1}^{i} f_j$$

are calculated. Since in an ordinary segment II design, between 20 and 30 pregnant dams are randomized to each dose level, a series of 30 random uniform numbers u_k are generated for each dose group, using the built-in GAUSS routine RNDU (multiplicative-congruential method). Hence, an equal number of 30 clusters is assigned to the control group and to the three dosed groups ($d = 0.25, 0.5, 1$). Observation k in dose group d is then (n_i, z_i) if $g_{i-1} \leq u_k < g_i$ (with $g_0 = 0$). Repeating this procedure for each of the four dose groups yields datasets of 120 observations. At the start of each simulation run, the seed is stored in order to enable repetition of the experiment.

To each dataset, the Bahadur, beta-binomial and conditional models were fitted. A Wald and LR test for the null hypothesis of no dose effect $H_0 : \beta_d = 0$ was calculated. Based on the frequency of rejecting H_0, rejection probabilities were estimated. Results are reported in Tables 3.3 (true model is Bahadur), 3.4 (beta-binomial), and 3.5 (conditional). Apart from rejection probabilities, the mean of the estimated dose effect $\hat{\beta}_d$ is given. Since from $\beta_d = 4$ (Bahadur and beta-binomial) or $\beta_d = 2$ (conditional) onwards, rejection probabilities are 100%, these results are not reported.

First, the focus is on the type I error (size) of the test. When the true model is Bahadur, beta-binomial or conditional, the estimated size (expressed as percentages) varies in the range (0.73; 11.80), (2.11; 11.94) and (3.80; 7.10) respectively. In particular, under a correctly specified model, the size is close to the nominal level for the

Table 3.3: Simulation study. Data generated under a Bahadur model. Each dataset consists of 4 dose groups with 30 litters in each. Each simulation result is based on 500 replications. Estimated rejection percentages of $H_0 : \beta_d = 0$ are shown. The significance level is 0.05. Div. indicates the number of divergences.

	Bahadur				beta-binomial				conditional			
β_d	Wald	LR	mean	div.	Wald	LR	mean	div.	Wald	LR	mean	div.
					$\beta_0 = -4.5; \beta_2 = 0.0$							
0.0	8.25	8.93	-0.093	209	7.17	8.19	-0.097	207	6.40	7.60	-0.087	0
0.5	13.80	13.40	0.505	0	13.70	13.50	0.532	137	11.40	11.80	0.474	0
1.0	54.21	53.78	1.064	37	51.47	50.79	1.043	59	47.20	47.20	0.983	0
2.0	99.80	99.80	2.048	9	99.79	99.79	2.036	33	99.60	99.60	2.031	0
					$\beta_0 = -5.5; \beta_2 = 0.0$							
0.0	4.55	10.61	-0.174	434	2.43	5.99	-0.255	216	2.25	5.53	-0.235	12
0.5	6.58	9.98	-0.157	59	6.72	9.89	0.424	136	6.26	9.70	0.419	5
1.0	27.56	29.49	1.264	344	19.09	23.17	1.062	90	18.67	22.49	1.037	2
2.0	88.50	88.63	2.252	78	88.47	88.27	2.234	108	85.80	86.00	2.048	0
					$\beta_0 = -4.5; \beta_2 = 0.1$							
0.0	3.04	4.66	-0.051	6	4.23	5.44	-0.062	4	9.40	11.80	-0.082	0
0.5	11.44	12.95	0.462	29	12.63	14.06	0.499	2	20.80	21.40	0.461	0
1.0	35.07	36.27	1.007	1	33.87	34.07	1.008	1	45.60	46.00	0.952	0
2.0	98.00	98.00	2.007	0	97.80	97.80	2.033	1	98.00	98.00	1.877	0
					$\beta_0 = -5.5; \beta_2 = 0.1$							
0.0	0.82	4.41	-0.276	137	0.73	4.16	-0.357	91	3.94	9.85	-0.367	43
0.5	3.75	8.43	0.437	49	4.34	5.91	0.422	43	9.54	12.66	0.440	18
1.0	10.69	14.47	0.990	23	11.95	15.38	0.924	32	23.42	25.66	0.905	9
2.0	77.42	78.43	2.163	4	75.75	77.56	2.197	1	83.97	84.77	2.127	1

Table 3.4: Simulation study. Data generated under a beta-binomial model. Each dataset consists of 4 dose groups with 30 litters in each. Each simulation result is based on 500 replications. Estimated rejection percentages of $H_0 : \beta_d = 0$ are shown. The significance level is 0.05. Div. indicates the number of divergences.

	Bahadur				beta-binomial				conditional			
β_d	Wald	LR	mean	div.	Wald	LR	mean	div.	Wald	LR	mean	div.
					$\beta_0 = -4.5; \beta_2 = 0.0$							
0.0	5.11	6.39	-0.053	187	4.73	6.31	-0.057	183	2.80	4.20	-0.058	0
0.5	19.70	18.97	0.600	94	18.14	17.73	0.595	94	16.60	16.80	0.574	0
1.0	54.00	54.22	1.088	50	51.64	51.17	1.065	72	47.60	47.60	1.004	0
2.0	99.80	99.80	2.012	7	99.79	99.57	1.998	32	99.80	99.80	2.005	0
					$\beta_0 = -5.5; \beta_2 = 0.0$							
0.0	4.41	11.94	-0.267	433	2.11	7.25	-0.124	169	2.65	6.94	-0.182	10
0.5	6.82	10.23	0.395	412	2.94	7.96	0.429	299	5.48	8.52	0.470	7
1.0	31.21	31.85	1.143	343	30.95	32.93	1.171	336	19.72	21.93	0.979	3
2.0	90.36	90.73	2.225	90	88.10	89.01	2.177	118	86.80	88.00	2.077	0
					$\beta_0 = -4.5; \beta_2 = 0.1$							
0.0	5.83	6.67	-0.017	20	3.29	4.40	-0.023	45	9.60	10.60	-0.009	0
0.5	12.68	14.11	0.488	11	12.39	13.04	0.504	40	18.60	19.20	0.477	0
1.0	43.78	43.78	1.037	2	40.17	40.83	1.045	20	51.40	51.60	1.024	0
2.0	98.00	97.80	1.998	0	98.40	98.40	2.052	0	98.60	98.60	1.921	0
					$\beta_0 = -5.5; \beta_2 = 0.1$							
0.0	4.85	8.44	-0.167	192	2.73	6.97	-0.139	170	6.09	10.92	-0.168	24
0.5	8.31	13.02	0.395	139	5.90	10.14	0.369	135	10.81	14.76	0.408	19
1.0	16.54	18.80	0.945	101	12.78	17.55	0.952	124	20.08	23.53	0.905	7
2.0	75.98	76.39	2.047	13	74.84	76.73	2.052	53	80.00	80.52	2.018	2

Table 3.5: Simulation study. Data generated under a conditional model. Each dataset consists of 4 dose groups with 30 litters in each. Each simulation result is based on 500 replications. Estimated rejection percentages of $H_0 : \beta_d = 0$ are shown. The significance level is 0.05. Div. indicates the number of divergences.

	Bahadur				beta-binomial				conditional			
β_d	Wald	LR	mean	div.	Wald	LR	mean	div.	Wald	LR	mean	div.
					$\beta_0 = -2.5; \beta_2 = 0.0$							
0.0	6.20	6.20	-0.010	0	6.00	5.80	-0.009	0	5.40	5.40	-0.007	0
0.5	59.20	57.80	0.512	0	58.80	57.40	0.510	0	56.20	56.20	0.512	0
1.0	99.80	99.60	1.015	0	99.60	99.60	1.014	0	99.60	99.80	1.029	0
1.5	100	100	1.508	0	100	100	1.506	0	100	100	1.540	0
					$\beta_0 = -3.5; \beta_2 = 0.0$							
0.0	5.63	5.84	-0.065	3	5.45	5.45	-0.064	5	4.40	5.20	-0.052	0
0.5	28.40	27.60	0.486	0	27.22	26.87	0.482	5	27.60	27.40	0.488	0
1.0	87.40	86.80	1.001	0	86.52	86.52	0.999	3	87.40	87.40	1.008	0
1.5	99.80	99.80	1.523	0	99.80	99.80	1.520	0	99.80	99.80	1.532	0
					$\beta_0 = -2.5; \beta_2 = 0.1$							
0.0	6.61	7.10	0.000	35	6.71	6.78	0.006	13	5.20	5.80	0.004	0
0.5	27.23	26.48	0.539	43	28.34	26.18	0.545	11	27.40	27.20	0.501	0
1.0	87.89	87.47	1.114	21	88.69	88.03	1.111	7	87.40	87.20	0.994	0
1.5	99.80	99.80	1.769	12	99.80	99.80	1.758	2	99.80	99.80	1.513	0
					$\beta_0 = -3.5; \beta_2 = 0.1$							
0.0	4.32	4.68	-0.026	222	4.24	5.52	-0.038	29	3.80	5.80	-0.037	0
0.5	16.38	16.38	0.573	146	12.73	12.81	0.503	16	12.00	12.80	0.467	0
1.0	46.75	46.32	1.075	38	44.86	44.74	1.061	44	44.40	45.00	0.978	0
1.5	87.30	87.10	1.615	4	86.57	87.24	1.609	22	89.20	89.20	1.506	0

conditional model and acceptable for the beta-binomial model, whereas considerable departures are seen for the Bahadur model. The size of Wald is always smaller than the one of LR when data are generated from the Bahadur and beta-binomial models. For the conditionally generated data, the sizes are comparable. Generating samples from Bahadur or beta-binomial, LR can result in sizes that are too high, especially when the Bahadur model is fitted (smallest intercept and no association) and when the conditional model is fitted in case of association. Very small sizes are found for the Wald test when the true model is Bahadur and the fitted model is either Bahadur or beta-binomial (smallest intercept and correlated outcomes). Generating data from the conditional model, the estimated sizes are very reasonable. Several possible explanations for the discrepancies between the size and the nominal level can be suggested. At first sight, random variability is able to explain only part of the effect. Indeed, samples of 120 dams are fairly large and simulation runs of length 500 should yield a fairly accurate picture. However, even moderate sample sizes, in combination with low background rates and small dose effects might still lead to very small numbers of malformations. Apart from introducing small sample effects, this phenomenon can lead to divergence of the numerical maximization process. Tables 3.3–3.5 report on the number of divergences. The divergence is considerable when fitting a Bahadur model, while it is almost not an issue with the conditional model, a clear advantage of the latter one. For instance, a pathological case is seen in Table 3.4 ($\beta_0 = -5.5, \beta_2 = 0.0$). The large number of divergences is probably due to sparseness in conjunction with the numerical complexity of the Bahadur model. The number of affected littermates in an entire dataset is on average 6 ($\beta_d = 0.0$), 7 ($\beta_d = 0.5$), 10 ($\beta_d = 1.0$), 18 ($\beta_d = 2.0$) and 90 ($\beta_d = 4.0$). In other words, divergence is most likely to occur for $\beta_d = 0$. In fact, convergence problems suggest great care with both Bahadur and beta-binomial models. It seems that, especially with low background rates and hence with a very small response probability, using these models is not advisable. Finally, fitting a misspecified model distorts the distribution of the test statistic and hence, referring to a χ^2_1 distribution might be misleading. However, also discrepancies under the correct model are noticed.

Next, the attention is turned to the parameter settings with a non-zero dose effect, i.e., to the estimated power of the tests. It was decided not to adjust the power for size. While this option could be debatable, it is argued that it reflects common data analysis practice. Of course, powers can then vary by the size of the

test alone. In all three tables, one observes that the power is lower for the correlated true model than for the uncorrelated version. This is because in a correlated model, the information contributed by a littermate is reduced. Further, since β_0 affects the background rate, the power increases with β_0. When the true model is conditional (Table 3.5), the power is fairly stable across test statistics and models fitted. When the true model is Bahadur or beta-binomial, the picture is less clear. The LR test seems to be somewhat more powerful, especially for dose effects $\beta_d = 0.5$ and 1.0. The number of divergences in Table 3.3 ($\beta_0 = -5.5, \beta_2 = 0.0$) exhibits an anomalous behaviour. Several strategies were tried to reduce the divergence, overall and especially for $\beta_d = 1.0$, including a grid search and logistic regression to determine initial estimates for β_0 and β_d, step halving, ... However, the results could not be improved. One of the main problems is that the parameter spaces of the Bahadur and beta-binomial models are not rectangular, in contrast to the parameter space of the conditional model. Especially for the Bahadur model, there are severe restrictions (Bahadur 1961; Prentice, 1988), as will be discussed in Chapter 4.

A striking feature is that the dramatic differences seen between the test statistics in the asymptotic study, seem to disappear here. It is claimed that there are two main reasons for this apparent discrepancy. First, when the true model is Bahadur or beta-binomial, all population test statistics are very close when the true dose effect is smaller than about 2. Secondly, for larger dose effects, although the curves start to separate, a very modest sample size is likely sufficient to obtain reasonable power. A sample size of 4×30 will inevitably lead to very high powers for all test statistics and will flatten out the observed asymptotic differences. The combination of both effects implies that for the design under consideration, the power is almost independent of the fitted model.

For each simulation study, the mean of $\hat{\beta}_d$ is also reported. They are usually in good agreement with the true effect. Medians are also calculated. However, since the agreement between means and medians is very good to excellent, the latter are not reported.

Finally, there is good agreement between the observed variability of $\hat{\boldsymbol{\beta}}$ (the covariance of all estimates in a simulation study) and the average of the asymptotic covariance matrices, estimated from each dataset. For $\hat{\beta}_d$, the ratio between both precision estimates varied within $(0.95; 1.3)$ with an average that was very close to 1.0.

Table 3.6: Parameter estimates (standard errors) for the DEHP study.

Outcome	Parameter	Bah	Bah(GEE2)	BB	Cond
External	β_0	-4.93(0.39)	-4.98(0.37)	-4.91(0.42)	-2.81(0.58)
	β_d	5.15(0.56)	5.29(0.55)	5.20(0.59)	3.07(0.65)
	β_2	0.11(0.03)	0.15(0.05)	0.21(0.09)	0.18(0.04)
Skeletal	β_0	-4.67(0.39)	-5.23(0.40)	-4.88(0.44)	-2.79(0.58)
	β_d	4.68(0.56)	5.35(0.60)	4.92(0.63)	2.91(0.63)
	β_2	0.13(0.03)	0.18(0.02)	0.27(0.11)	0.17(0.04)
Visceral	β_0	-4.42(0.33)	-4.49(0.36)	-4.38(0.36)	-2.39(0.50)
	β_d	4.38(0.49)	4.52(0.59)	4.42(0.54)	2.45(0.55)
	β_2	0.11(0.02)	0.15(0.06)	0.22(0.09)	0.18(0.04)
Collapsed	β_0	-3.83(0.27)	-5.23(0.40)	-3.83(0.31)	-2.04(0.35)
	β_d	5.38(0.47)	5.35(0.60)	5.59(0.56)	2.98(0.51)
	β_2	0.12(0.03)	0.18(0.02)	0.32(0.10)	0.16(0.03)

3.3 Analysis of NTP data

To amplify the findings from the analyses of simulated data, the models described in Chapter 2 are applied to the NTP data. Apart from the external, skeletal and visceral malformation outcomes, a collapsed malformation outcome was considered, which is one if a fetus exhibits at least one type of malformation and zero otherwise. Tables 3.6 and 3.7 contain maximum likelihood estimates (MLE) and standard errors for the Bahadur, beta-binomial and conditional models. Estimates of the Bahadur model parameters obtained by a GEE2 method, are also shown.

Bahadur (MLE and GEE2) and beta-binomial parameters have the same interpretation, but they are not directly comparable with the parameters of the conditional model. The intercepts β_0 and dose effect parameters β_d have similar numerical values but the situation is slightly different for β_2. In 6 out of 8 cases, $\beta_2(\text{Bah}) < \beta_2(\text{GEE2}) < \beta_2(\text{BB})$. The only exceptions are EG (visceral), where the association is not statistically significant and EG (collapsed), where the three estimates are very close. In the other cases, the beta-binomial MLE for β_2 is typically about double the corresponding Bahadur MLE. This is due to range restrictions on β_2 in the Bahadur model. For instance, the allowable range of β_2 for the external

Table 3.7: Parameter estimates (standard errors) for the EG study.

Outcome	Parameter	Bah	Bah(GEE2)	BB	Cond
External	β_0	-5.25(0.66)	-5.63(0.67)	-5.32(0.71)	-3.01(0.79)
	β_d	2.63(0.76)	3.10(0.81)	2.78(0.81)	2.25(0.68)
	β_2	0.12(0.03)	0.15(0.05)	0.28(0.14)	0.25(0.05)
Skeletal	β_0	-2.49(0.11)	-4.05(0.33)	-2.89(0.27)	-0.84(0.17)
	β_d	2.96(0.18)	4.77(0.43)	3.42(0.40)	0.98(0.20)
	β_2	0.27(0.02)	0.30(0.03)	0.54(0.09)	0.20(0.02)
Visceral	β_0	-7.38(1.30)	-7.50(1.05)	-7.45(1.17)	-5.09(1.55)
	β_d	4.25(1.39)	4.37(1.14)	4.33(1.26)	3.76(1.34)
	β_2	0.05(0.08)	0.02(0.02)	0.04(0.09)	0.23(0.09)
Collapsed	β_0	-2.51(0.09)	-4.07(0.71)	-2.51(0.09)	-0.81(0.16)
	β_d	3.05(0.17)	4.89(0.90)	3.05(0.17)	0.97(0.20)
	β_2	0.28(0.02)	0.26(0.14)	0.28(0.02)	0.20(0.02)

outcome in the DEHP data is $(-0.0164; 0.1610)$ when β_0 and β_d are fixed at their MLE. This range excludes the MLE under a beta-binomial model. It translates to $(-0.0082; 0.0803)$ on the correlation scale. A GEE2 estimate is valid as soon as the second, third and fourth order joint probabilities are nonnegative, whereas the likelihood analysis requires all joint probabilities to be nonnegative. Thus, a correlation valid for GEE2 estimation is allowed to violate the full likelihood range restrictions. The standard errors, obtained by Bahadur and GEE2 are very similar, except for EG(skeletal) and EG(collapsed). It is no coincidence that exactly in these cases, β_2 attains very high values, probably very close to the boundary of the admissible range, implying that boundary effects might distort large sample approximations to the null distribution. The beta-binomial model features all positive correlations. Hence, the dominant ordering of the estimated β_2 parameters reflects the severity of the parameter restrictions.

Since the conditional model has no restrictions on the parameters, it is easier to fit than the others. In all 8 examples, standard starting values (all parameters equal to zero) led to convergence.

Besides the Bahadur model, the beta-binomial model and the conditional model, also the George-Bowman model with the folded logistic parameterization, is fitted to

the NTP data. Maximum likelihood estimates and corresponding standard errors of the regression parameters of this model are listed in Table 3.8. The marginal George-Bowman model differs in several respects from the previous models. First, a fundamental drawback is that reversing the coding for malformed and non-malformed subjects has a non-trivial effect on the estimation. For example, the parameters for the external outcome in the DEHP study change from the ones reported in Table 3.8 to $\hat{\beta}_0 = 0.27(0.05)$ and $\hat{\beta}_d = -1.86(0.26)$. Also the log-likelihood at maximum changes from -172.40 to -166.41. The LR and Wald statistics change from 101.35 to 79.86 and from 101.43 to 55.59 respectively. Secondly, there is no explicit association parameter included. Rather, the folded logistic parameter vector describes the moments of all orders. While parsimony is no doubt desirable, the model fails to render direct quantitative evidence for the within cluster association. Thirdly, even though the folded logistic parameterization guarantees the joint probabilities to be nonnegative, one has to ensure that the vector $\boldsymbol{\beta}$ itself is valid. When dose d satisfies $0 \leq d \leq 1$, this implies the constraint $\beta_0 + \beta_d < 0$. For example, the dose effect in the external outcome for the DEHP study is restricted to the range $]-\infty; 8.05]$. Of course, there are also restrictions on the parameter space of the Bahadur model. Fourthly, in contrast with the Bahadur model, the beta-binomial model and the conditional model, it turns out that when the littermates are independent, then the George-Bowman model using the folded logistic parameterization, does not reduce to the binomial model. These features make that the model should be applied with caution.

After having discussed the parameter estimates of the fitted models, the focus is now on the problem of testing the null hypothesis of no dose effect. Results are summarized in Table 3.9. They are in agreement with previous findings.

For the LR tests, one observes that LR(Bah) dominates the others. LR(BB) is considerably smaller and the smallest values are found with LR(Cond). This picture is seen in 7 out of 8 cases. A slightly different picture is seen for EG (visceral and external outcomes), where all three statistics are in fact very close to each other. However, although there are discrepancies between the magnitudes of the LR statistics, they all clearly reject the null hypothesis.

Comparing LR to Wald tests, the former ones are seen to dominate the latter in most cases: LR(Cond)>W(Cond) in all 8 cases and LR(Bah)>W(Bah) in 6 cases. However, LR(BB)>W(BB) in only two cases and, more importantly, agree-

Table 3.8: Maximum likelihood estimates (standard errors) for the George-Bowman model.

Outcome	Parameter	DEHP	EG	DYME
External	β_0	-8.05(0.52)	-9.25(0.97)	-11.18(0.92)
	β_d	6.77(0.67)	4.42(1.15)	10.18(0.96)
Skeletal	β_0	-8.05(0.53)	-5.49(0.32)	-8.22(0.48)
	β_d	6.56(0.70)	4.42(0.39)	7.79(0.51)
Visceral	β_0	-7.41(0.46)	-11.88(1.89)	-11.23(1.16)
	β_d	6.01(0.65)	6.46(2.04)	8.04(1.26)
Collapsed	β_0	-6.27(0.32)	-5.36(0.31)	-7.55(0.41)
	β_d	6.07(0.37)	4.38(0.38)	7.34(0.43)

ment between both test statistics is very close, providing evidence for approximate equivalence of both tests under a range of alternatives. This feature is in agreement with the asymptotic findings. Recall that both Bahadur test statistics might differ due to a misspecified higher order correlation structure, whereas for the conditional model, the Wald statistic could be low due to sensitivity of the test to the particular parameterization adopted. For example, the correlations between $\hat{\beta}_d$ and the other parameter estimates for the external outcome in the DEHP study are $\text{corr}(\hat{\beta}_0, \hat{\beta}_d) = -0.96$ and $\text{corr}(\hat{\beta}_a, \hat{\beta}_d) = -0.79$, as opposed to -0.91 and 0.27 for Bahadur and -0.90 and 0.23 for beta-binomial.

Among the Wald tests, W(Bah) and W(BB) are reasonably close to each other, apart from two aberrant cases (EG skeletal and EG collapsed). Misspecification might be one of the sources for the observed discrepancies. When GEE2 based tests are believed to correct for (at least part) of aforementioned misspecification, then their values should be smaller than W(Bah) and much closer to W(BB). This effect is indeed seen, but it is not clear whether the naive or robust test is the best statistic to achieve this correction, although only W(robust) is compatible with the philosophy of generalized estimating equations. The most striking phenomenon is that the two aberrant W(Bah) values in the EG data are indeed largely corrected downward by the GEE2 versions.

Liang and Hanfelt (1994) have shown that assuming a constant intraclass correlation in the beta-binomial model might substantially bias mean parameter esti-

Table 3.9: Wald and likelihood ratio test statistics.

Outcome	Model	Statistic	DEHP	EG
External	Bahadur	LR	96.48	15.05
		Wald	85.94	11.89
	BB	LR	71.58	13.18
		Wald	76.78	11.61
	Cond	LR	43.20	14.43
		Wald	22.30	10.78
	GEE2	Wald(naive)	79.40	12.50
	GEE2	Wald(robust)	92.41	14.70
Skeletal	Bahadur	LR	76.40	182.45
		Wald	71.02	261.22
	BB	LR	58.51	63.88
		Wald	61.39	74.89
	Cond	LR	38.95	49.95
		Wald	21.46	23.15
	GEE2	Wald(naive)	70.72	120.99
	GEE2	Wald(robust)	78.87	58.53
Visceral	Bahadur	LR	81.28	16.00
		Wald	78.82	9.40
	BB	LR	59.78	17.37
		Wald	66.80	11.82
	Cond	LR	33.72	13.98
		Wald	19.91	7.81
	GEE2	Wald(naive)	71.45	10.32
	GEE2	Wald(robust)	58.23	14.64
Collapsed	Bahadur	LR	164.75	189.99
		Wald	130.31	314.40
	BB	LR	91.66	65.36
		Wald	98.46	71.58
	Cond	LR	74.48	50.74
		Wald	33.71	23.39
	GEE2	Wald(naive)	113.75	121.87
	GEE2	Wald(robust)	92.27	29.69

mation and testing. Therefore, it is useful to study at least one possible departure from the constant association model. In the Bahadur, beta-binomial and conditional models, the association parameter β_2 was allowed to vary linearly with dose level, $\beta_{2i} = \beta_{20} + \beta_{2d} d_i$, extending the three parameter families $(\beta_0, \beta_d, \beta_2)$ to four parameter versions $(\beta_0, \beta_d, \beta_{20}, \beta_{2d})$. Reconsider now the problem of testing the null hypothesis of no dose effect, neither on the malformation rate nor on the association. The corresponding test statistics are shown in Table 3.10. First, values in bold correspond to those cases where the null hypothesis of a constant association parameter $H_0 : \beta_{2d} = 0$, was rejected on the basis of a one degree of freedom likelihood ratio test. Clearly, a non-constant association in one model (e.g., the conditional model) does not necessarily imply the same for the other models (e.g., the Bahadur and beta-binomial models). Next, the test statistics for dose effect are considered, which in the four parameter model becomes $H_0 : \beta_d = \beta_{2d} = 0$. In most cases, the statistics vary only mildly, although W(Bah) tends to increase somewhat more. The discrepancy is larger when the null hypothesis of a constant association is rejected. Of course, one has to bear in mind that these test statistics should be compared to a null χ^2 distribution with *two* degrees of freedom, diluting power when there is no evidence for non-constant association. Finally, failure to detect a linear trend on the association does not imply that the association is constant, since the association function might have an entirely different shape (e.g., quadratic). In a real data analysis, it is advisable to explore these functions in a bit more detail (Molenberghs and Ryan, 1999).

3.4 Concluding remarks

Marginal, random effects and conditional models were studied to describe dose-response curves based on a binary outcome in clustered experiments. Bahadur (1961) provided a general description of the joint distribution of correlated binary data which can be simplified to exchangeability. This model suffers from severe range restrictions, which complicates the numerical performance of maximization. It was observed at several occasions that setting higher order associations equal to zero might introduce discrepancies. Most likely, this issue accounts at least partly for the differences observed between Bahadur and beta-binomial models, who have the same first and second moments. The effect of including these parameters on

Table 3.10: Wald and likelihood ratio test statistics with linear dose effect on the association parameter. Bold figures refer to cases where the dose effect on the association was significant at the 5% nominal level.

Outcome	Model	Statistic	DEHP	EG
External	Bahadur	LR	99.09	**19.99**
		Wald	93.04	**18.07**
	BB	LR	71.90	13.57
		Wald	73.70	11.20
	Cond	LR	44.99	16.19
		Wald	25.98	12.39
Skeletal	Bahadur	LR	76.59	**192.77**
		Wald	69.60	**207.93**
	BB	LR	58.65	65.74
		Wald	63.07	63.50
	Cond	LR	40.70	**58.90**
		Wald	24.98	**30.72**
Visceral	Bahadur	LR	**86.13**	
		Wald	**90.00**	
	BB	LR	61.56	17.55
		Wald	67.17	9.48
	Cond	LR	33.72	14.89
		Wald	19.99	7.44
Collapsed	Bahadur	LR	**173.07**	**196.06**
		Wald	**157.37**	**211.28**
	BB	LR	**97.98**	67.45
		Wald	**107.17**	63.88
	Cond	LR	75.61	**60.01**
		Wald	33.00	**30.35**

the operational characteristics of the models is studied in the next chapter, where it is established that at least fourth order interactions are needed to relieve the restrictions on the Bahadur model. The beta-binomial model combines simplicity with interpretational ease (first and second order moment parameters have intuitive meaning) and with mild restrictions on the parameter space. The numerical performance (speed and stability) is much better than with the Bahadur model, although still a nonnegligible number of divergences was observed. As a consequence, the beta-binomial model is preferred among the marginal and random effects models. Even though GEE2 estimation yields less severely constrained parameters than the likelihood version of the Bahadur model, it suffers from other problems (e.g., likelihood ratio tests and joint probabilities are unavailable). These conclusions are in agreement with the prominent role played by the beta-binomial model (or its generalizations such as the Dirichlet-multinomial model) in quantitative risk assessment.

The conditional model is very different from the others and hence, the main effect and association parameters cannot be compared directly and the parameters have to be interpreted in conditional rather than in marginal terms. Since there are no parameter space constraints, interpretation is not further complicated and the model is the easiest one to fit. This was seen in both finite sample simulations as well as in the NTP data. In both cases, little problems with divergence were encountered. With the other models, moderate (beta-binomial) to extreme (Bahadur) care had to be taken with the numerical maximization process (choosing starting values and monitoring convergence).

When rejecting the null hypothesis of no dose effect is of primary importance, one could still suggest the use of the beta-binomial model, where both LR and Wald tests might be used, although LR tests have better theoretical properties. In addition, the LR test of the conditional model also seems very appropriate. Indeed, since reasonably strong effects, such as the ones encountered in the NTP datasets, were found with all methods, one might opt for the model which is easiest to fit. From this point of view, the best model in this respect is undoubtedly the conditional model. Because the Wald test in the conditional model performed very poorly, the likelihood ratio test should be recommended. The disadvantage of the GEE2 based model is that it does not entail the same range of inferential tools (such as likelihood ratio tests) as the likelihood-based conditional model.

The model proposed by George and Bowman (1995) using the folded logistic

parameterization, has been considered in the analysis of the NTP data. This model suffers from a number of drawbacks. In contrast with the Bahadur model, the beta-binomial model and the conditional model, it turns out that in case of independence of the littermates, the specific George-Bowman model with the folded logistic parameterization does not simplify to the binomial model. Another disadvantage of the George-Bowman model is the dependence of coding of success and failure. This implies that not only the maximum likelihood estimates change when the coding is reversed, but also the values of the maximized log-likelihood and the derived test statistics. Further, as is the case with the Bahadur model, there are non-trivial restrictions on the parameter space of the George-Bowman model. Finally, there is no explicit association parameter included and therefore, the model fails to render direct quantitative evidence for the within cluster association.

Other models than the ones presented here deserve consideration, e.g., the odds ratio model. Also, other test statistics such as a score test, could have been considered.

In conclusion, procedures that incorporate the effect of dose on both death and malformation are worthwhile to consider. The latter extension enables the study of the relation between dosing and observed cluster size. However, in the simplified setting considered in this chapter, some of the problems, advantages and drawbacks of different estimation and testing procedures have been identified.

Chapter 4

Behaviour of the likelihood ratio test statistic under a Bahadur model

In the previous chapter, the Bahadur, George-Bowman, beta-binomial and conditional models are compared with respect to their performance in estimating dose effect and testing the null hypothesis of no dose effect using likelihood ratio (LR) and Wald statistics. A simplified Bahadur model is used in which the association structure is confined to pairwise correlations and higher order correlations are set equal to zero. One of the findings is that for strong dose effects, the LR statistic under the Bahadur model is much larger than any other statistic considered. This feature turns out to be more pronounced with larger true dose effects. While desirable at first sight, it is claimed that this property is primarily due to misspecifying the correlation structure. Generating data from a two-way Bahadur model, the alternative model is by construction correctly specified, whence the effect of misspecification is necessarily confined to the associated null model. This is in line with the finding that the Wald statistic does not inflate as the dose effect becomes stronger.

In Section 4.1, an asymptotic study is performed, while the NTP data are analysed in Section 4.2. Both studies support this claim. In addition, it is shown that adding three-way correlations to the Bahadur model induces little change, while including fourth order correlation helps closing the gap between the Bahadur and beta-binomial likelihood ratios. The restrictions are further investigated in Section 4.3.

4.1 Asymptotic study

In the previous chapter, the behaviour of the two-way Bahadur, beta-binomial and conditional models was studied based on, among others, asymptotic calculations. Artificial samples were generated based on some underlying model, following a suggestion of Rotnitzky and Wypij (1994). Details about this procedure and about the selected values of the intercept parameter β_0, the dose effect parameter β_d and the association parameter β_2, can be found in Section 3.1. The key feature of the method of Rotnitzky and Wypij is that each combination of dose (d_i), number of viable fetuses (n_i) and number of malformations (z_i) is weighted according to the probability of occurrence in the underlying model. For each parameter combination, such an artificial sample is generated. Then, the Bahadur, beta-binomial and conditional models are fitted to these samples, based on the maximum likelihood procedure. Finally, testing the null hypothesis of no dose effect is based on the Wald and the likelihood ratio (LR) statistics.

When a Bahadur model is fitted to a sample with a non-zero dose effect, the LR statistic is inflated in comparison to its counterparts from the beta-binomial and conditional models (Section 3.1). This is better seen for larger dose effects and holds irrespective of the underlying model and irrespective of the presence of association (Figures 3.1 – 3.3). In particular, the differences between the LR-trajectories of the Bahadur model and the beta-binomial model (which can also be considered as a marginal model) will turn out to be a useful guide to investigate this anomalous behaviour.

When the underlying model shows no clustering, both Bahadur and beta-binomial models reduce to a logistic regression model and hence yield the same likelihood. Since a difference in LR is observed, the null models (of no dose effect) must be responsible for the difference. This is illustrated in Figure 4.1. Artificial samples are generated from a Bahadur model with intercept $\beta_0 = -4.5$ and no association ($\beta_2 = 0$). The dose effect β_d ranges from 0 to 8 (step size 0.5).

One observes that the negative log-likelihood of the beta-binomial null model increases until a dose effect of about 5. For the Bahadur null model the increase does not stop. As a consequence, both curves start separating.

In addition to the likelihood, also the parameter estimates of the fitted models are investigated. In particular, the ML estimate of pairwise association $\hat{\beta}_2$ when fitting a null model, is explored. The settings are the same as in Figure 4.1. The

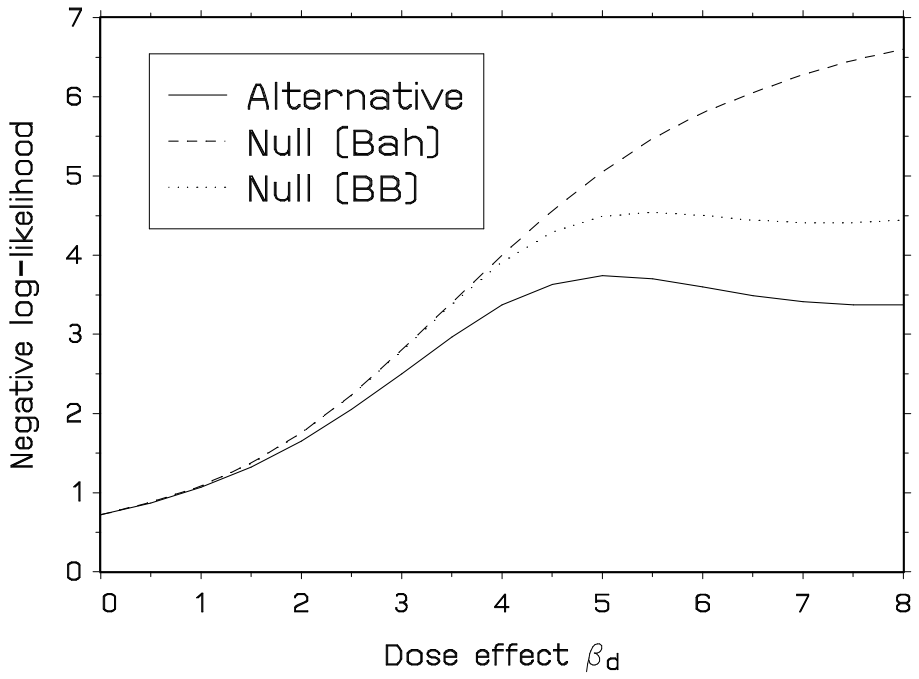

Figure 4.1: *Negative log-likelihood of the two-way Bahadur and beta-binomial alternative and null models fitted to artificial samples.*

relationship between $\hat{\beta}_2$ and dose effect β_d in the underlying model, is illustrated in Figure 4.2 for the beta-binomial and Bahadur null models.

Fitting the null model, an apparent association, clearly stemming from the omitted dose effect, is captured by the association parameter. As dose effect becomes stronger, this quantity clearly increases for the beta-binomial model, while it levels off for the Bahadur model. This bound occurs at $\beta_d \approx 4$, i.e., where the LR statistics of both models start to separate.

Based on Figures 4.1 and 4.2, it will be shown in Section 4.3 that the striking behaviour of the LR statistic under Bahadur, is linked with the presence of parameter restrictions. In the beta-binomial model, all non-negative correlations are allowable. However, in the Bahadur model, the parameter space is constrained, as is discussed in Section 4.3. Support for this statement will be sought by extending the Bahadur model with higher order correlation parameters. In turn, the third and fourth order correlations will be added to the Bahadur model.

First, the model is extended with a three-way association parameter, although it sacrifices the computational ease of this model, even in the case of exchangeable

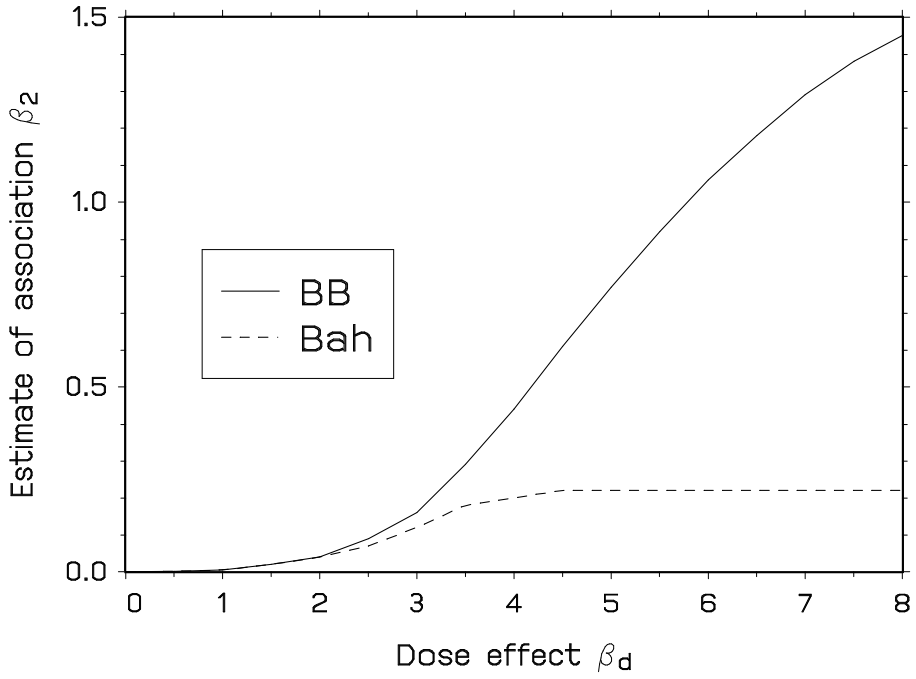

Figure 4.2: *Association of the two-way Bahadur and beta-binomial null models fitted to artificial samples.*

data. A Newton-Raphson algorithm based on analytically calculated derivatives, is used to obtain ML estimates of the four model parameters, i.e., intercept β_0, dose effect β_d, two-way association β_2 and three-way association β_3. This extended model is fitted to the same artificial samples as in Figures 4.1 and 4.2. Adding this parameter does not substantially change the relationship between the LR statistic and true dose effect. Also Figure 4.2 remains virtually identical when Bah(3) is fitted rather than Bah(2). The suffix p in Bah(p) denotes the highest order correlation that is still in the model.

Next, the four-way correlation is added to the model. It is extremely hard to fit this model to the artificial samples of Figures 4.1 and 4.2. Therefore, it was decided to generate artificial samples based only on clusters of size 12, which are very frequently encountered in rodent experiments. The other settings remain the same as in the previous figures. Figure 4.3 depicts the trajectory of the LR statistic versus true dose effect for Bah(2), Bah(3) and Bah(4), as well as for the beta-binomial model. In contrast to the addition of a third order correlation, the LR statistic corresponding to a strong dose effect decreases considerably when including

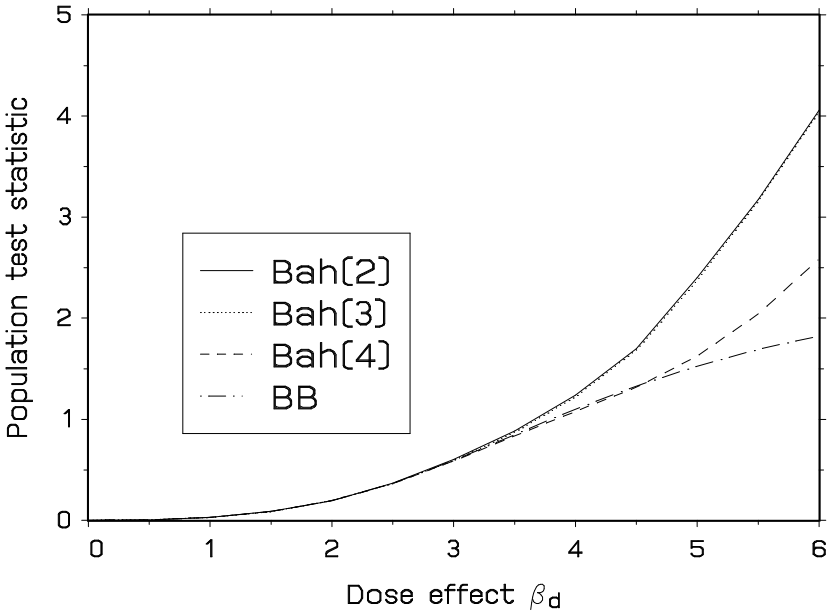

Figure 4.3: *Trajectories of the likelihood ratio statistic of Bahadur and beta-binomial models fitted to artificial samples consisting of clusters of size 12.*

all association parameters up to the fourth order. Analogous to Figure 4.2, the relationship between the ML estimate of the pairwise association $\hat{\beta}_2$ of the null model and the true dose effect was also investigated for the settings of Figure 4.3. In Figure 4.4, it is shown that the results for the Bah(2) and beta-binomial models are similar to Figure 4.2. The trajectory of Bah(3) is again virtually the same as the one of Bah(2). However, the trajectory of Bah(4) is more comparable to the one of the beta-binomial model.

In conclusion, it appears that adding a three-way correlation does not help in changing the behaviour of the LR statistic, while the four-way correlation seems crucial. It is claimed that the net effect of including the fourth order association is a relaxation of the restrictions on the pairwise correlation.

4.2 Analysis of NTP data

To further illustrate the findings of the asymptotic study, the Bahadur and beta-binomial models are fitted to the NTP data. The effects in mice of the exposures DEHP, DYME and EG are investigated and malformations are classified as being external, skeletal and visceral. Also a collapsed outcome is considered, which is one

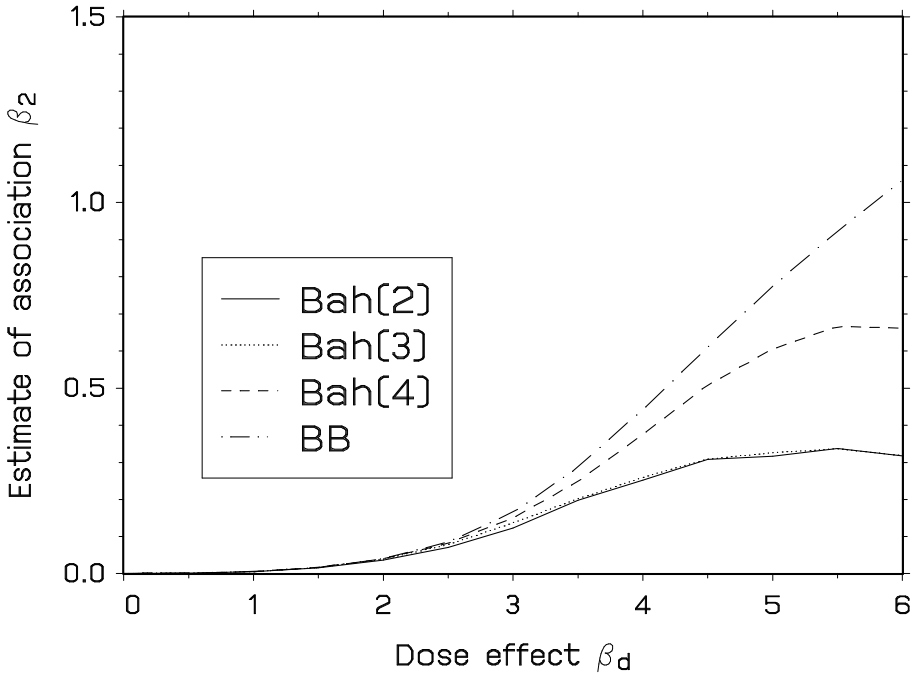

Figure 4.4: *Association of the Bahadur and beta-binomial null models fitted to artificial samples consisting of clusters of size 12.*

if a fetus has at least one malformation and zero otherwise.

The LR statistics to test the null hypothesis of no dose effect are computed. To this end, the Bah(2), Bah(3), Bah(4) and beta-binomial models are fitted (Table 4.1). Missing values in this table (and the following tables) are due to the lack of convergence. In general, the actual values of the LR test statistic of Bah(2) and Bah(3) are strikingly larger than for the beta-binomial model. The LR under Bah(4) is still larger than, but better comparable to the LR under beta-binomial.

Analogous to Figure 4.1, the likelihood of the Bahadur and beta-binomial alternative and null models is examined (Table 4.2). The likelihood of the Bahadur alternative model increases only slightly when including higher order associations in Bah(2). The beta-binomial alternative model is most often more likely than Bah(4). However, the differences in LR when fitting Bahadur and beta-binomial models, are primarily due to differences in the likelihood of the null models (i.e., without dose effect). The likelihood of the Bah(2) and Bah(3) null models are in general markedly smaller than the likelihood of the Bah(4) null model, which is in turn strikingly smaller than the likelihood of the beta-binomial null model. This is in agreement

Table 4.1: Likelihood ratio test statistic for $H_0 : \beta_d = 0$, after fitting Bahadur and beta-binomial models to DEHP, DYME and EG data.

Exposure	Malformation	Bah(2)	Bah(3)	Bah(4)	BB
DEHP	External	96.5	96.2	76.2	71.6
	Skeletal	76.4	78.8	68.1	58.5
	Visceral	81.3	80.5	61.0	59.8
	Collapsed	164.7	166.1	123.7	91.7
DYME	External	241.5	243.6	.	112.0
	Skeletal	310.8	315.3	.	117.4
	Visceral	63.9	48.2	.	43.3
	Collapsed	420.7	421.7	.	151.7
EG	External	15.1	6.0	.	13.2
	Skeletal	182.4	184.6	116.2	63.9
	Visceral	16.0	19.2	.	17.4
	Collapsed	190.0	191.2	116.8	65.4

with the results of the asymptotic study.

In analogy with Figure 4.2, the focus is now on the MLE of the second order association parameter $\hat{\beta}_2$ when fitting Bahadur and beta-binomial null models to the NTP data (Table 4.3). In general, $\hat{\beta}_2$ of Bah(2) is comparable to the one of Bah(3), but both are strikingly smaller than $\hat{\beta}_2$ of Bah(4), which is in turn markedly smaller than $\hat{\beta}_2$ of the beta-binomial model. Again, this is similar to the asymptotic study.

In agreement with the asymptotic study, the results obtained here suggest that the behaviour of the LR statistic under Bahadur is related to restrictions on the association parameters. Adding a fourth order correlation to a three-way Bahadur model seems to result in a relaxation of these bounds and leads to values of the LR statistic being more comparable to the ones of the beta-binomial model. In order to get evidence for this claim, the parameter space of the Bahadur model is explored.

Table 4.2: Negative log-likelihood evaluated at the maximum for Bahadur and beta-binomial alternative and null models fitted to DEHP, DYME and EG data.

H	Exposure	Malformation	Bah(2)	Bah(3)	Bah(4)	BB
H_1	DEHP	External	171.60	171.57	169.94	170.35
		Skeletal	171.24	169.12	168.37	166.57
		Visceral	196.96	196.93	194.66	194.99
		Collapsed	281.39	280.25	278.44	274.27
	DYME	External	135.84	134.80	134.43	135.52
		Skeletal	207.59	205.08	197.68	187.33
		Visceral	95.18	95.10	.	91.58
		Collapsed	206.02	204.85	201.49	196.63
	EG	External	90.29	89.80	.	88.82
		Skeletal	375.89	374.52	367.92	360.40
		Visceral	50.04	46.10	.	50.10
		Collapsed	384.63	383.70	376.91	368.04
H_0	DEHP	External	219.84	219.68	208.04	206.14
		Skeletal	209.44	208.52	202.44	195.82
		Visceral	237.60	237.16	225.17	224.88
		Collapsed	363.76	363.29	340.30	320.10
	DYME	External	256.58	256.58	.	191.52
		Skeletal	362.97	362.72	.	246.03
		Visceral	127.14	119.21	.	113.21
		Collapsed	416.39	415.71	.	272.49
	EG	External	97.82	92.81	.	95.41
		Skeletal	467.11	466.81	426.02	392.34
		Visceral	58.04	55.70	.	58.78
		Collapsed	479.63	479.32	435.31	400.72

Table 4.3: Maximum likelihood estimate of association $\hat{\beta}_2$ of the Bahadur and beta-binomial null models fitted to DEHP, DYME and EG data.

Exposure	Malformation	Bah(2)	Bah(3)	Bah(4)	BB
DEHP	External	0.25	0.26	0.51	0.78
	Skeletal	0.28	0.29	0.41	0.73
	Visceral	0.22	0.22	0.48	0.64
	Collapsed	0.30	0.30	0.57	1.09
DYME	External	0.33	0.33	.	1.65
	Skeletal	0.35	0.35	.	1.80
	Visceral	0.19	0.43	.	0.92
	Collapsed	0.32	0.32	.	1.82
EG	External	0.15	0.33	.	0.38
	Skeletal	0.32	0.32	0.54	0.96
	Visceral	0.11	-0.82	.	0.13
	Collapsed	0.31	0.32	0.53	0.98

4.3 Restrictions on the Bahadur model parameters

In Sections 4.1 and 4.2, it was suggested that the difference between the LR under a Bahadur model and under a beta-binomial model is related to restrictions on the model parameters. The beta-binomial model features all non-negative correlations, implying that there are only very mild constraints on the parameter space of this model. The restrictions on the Bahadur model parameters are much more complicated and stringent. Bahadur (1961) indicates that the sum of the probabilities of all possible outcomes is one, even when higher order correlations are set equal to zero. However, the requirement of having non-negative probabilities for all possible outcomes results in restrictions on the parameters. This holds even in the case of a Bahadur model with all higher order associations involved.

In this section, the subscript referring to the cluster is omitted in order to simplify notation.

Bahadur (1961) discusses the restrictions on the second order correlation when all higher order associations are left out. He shows that the second order approximation

is a probability distribution if and only if

$$-\frac{2}{n(n-1)}\min\left(\frac{\pi}{1-\pi},\frac{1-\pi}{\pi}\right)\leq\rho_{(2)}\leq\frac{2\pi(1-\pi)}{(n-1)\pi(1-\pi)+0.25-\gamma_0}, \qquad (4.1)$$

where

$$\gamma_0=\min_{z=0}^{n}\left\{[z-(n-1)\pi-0.5]^2\right\}.$$

Bounds of the second order correlation $\rho_{(2)}$ are graphically represented in Figure 4.5 for smaller litter sizes ($n=2,3,4,5$) and in Figure 4.6 for larger litters ($n=7,10,12,15$). The lower bound for $\rho_{(2)}$ in a two-way Bahadur model, attains its smallest value $-2/(n(n-1))$ at the malformation probability $\pi=0.5$. This bound quickly approaches zero as the litter size n increases. When $n=2$, the upper bound for $\rho_{(2)}$ is one, independent of π. For larger values of n, the upper bound depends on π and varies between $1/(n-1)$ and $2/(n-1)$. As a consequence, the upper bound is in the range $(0.09;0.18)$ for litters of size 12. As litter size increases, the restrictions on $\rho_{(2)}$ of Bah(2) become more severe.

Kupper and Haseman (1978) also consider the two-way Bahadur model and present numerical values for the constraints on $\rho_{(2)}$ for choices of π and n. Prentice (1988) studies the constraints in Bah(2) for any n. Furthermore, when the size of the clusters equals three, he argues that including the third order correlation removes the upper bound on $\rho_{(2)}$. However, it will be shown here that the requirements he verifies are necessary but not sufficient.

The parameter space of the general Bahadur model seems to be only partially known. The upper and lower bound of the second order correlation in Bah(3) and Bah(4) will be studied here. This leads to a clearer insight in the properties and usefulness of this model in general and in the behaviour of the LR statistic in particular.

First, the focus is on the three-way Bahadur model. An analytical procedure that can handle any cluster size is developed. In Appendix A of this chapter, explicit expressions for the bounds of $\rho_{(2)}$ are derived. These bounds are constructed such that for any value of $\rho_{(2)}$ between the lower and upper bound (both depending on the specified values of n and π), there exists at least one value of $\rho_{(3)}$ leading to a valid probability mass function.

The constraints on $\rho_{(2)}$ in Bah(3) for $n=3$ are depicted in Figure 4.5. Although this model is saturated in the sense that only clusters of size three are considered

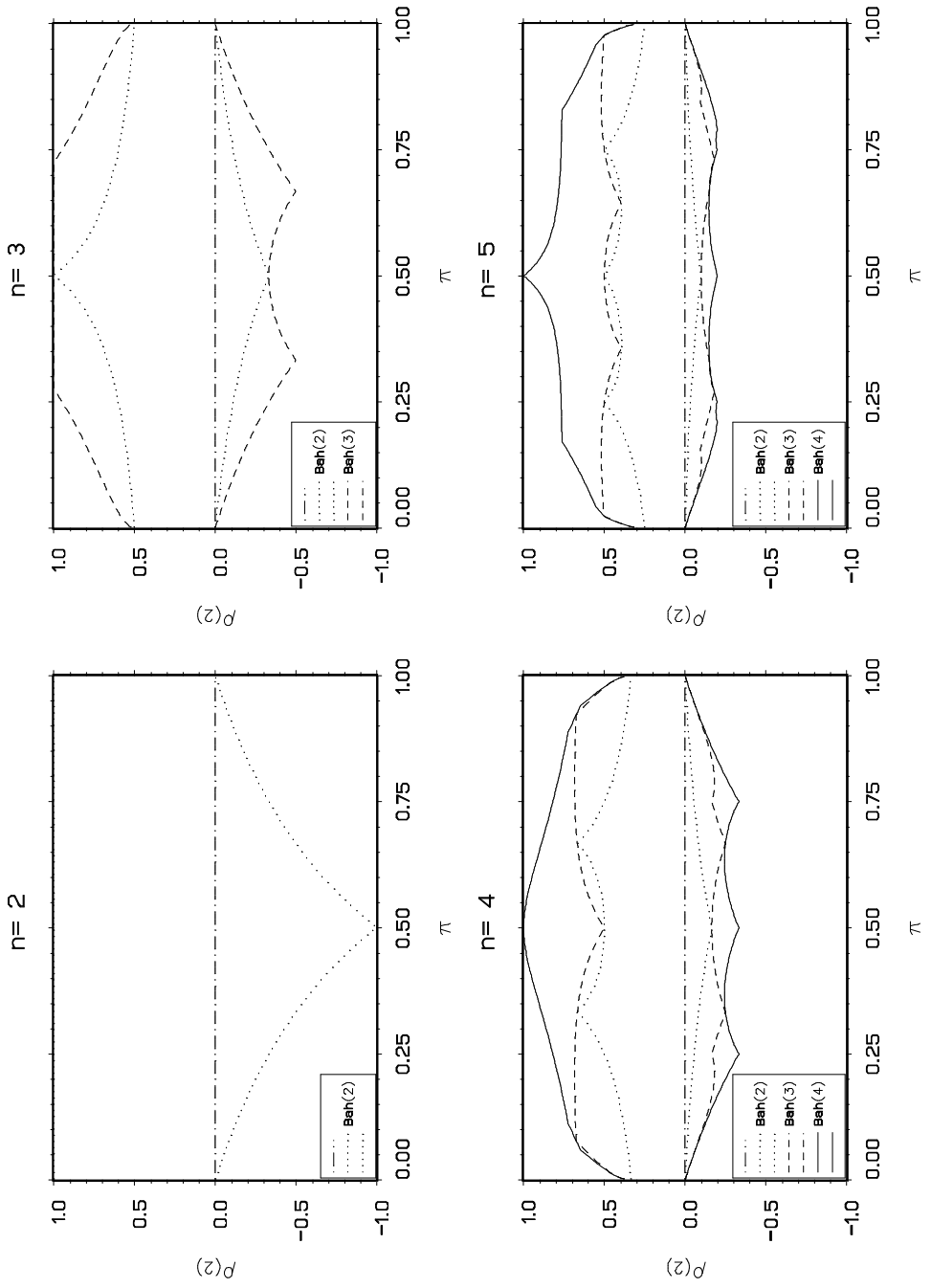

Figure 4.5: *Boundaries for the second order correlation of the two-way, three-way and four-way Bahadur model for some smaller litter sizes.*

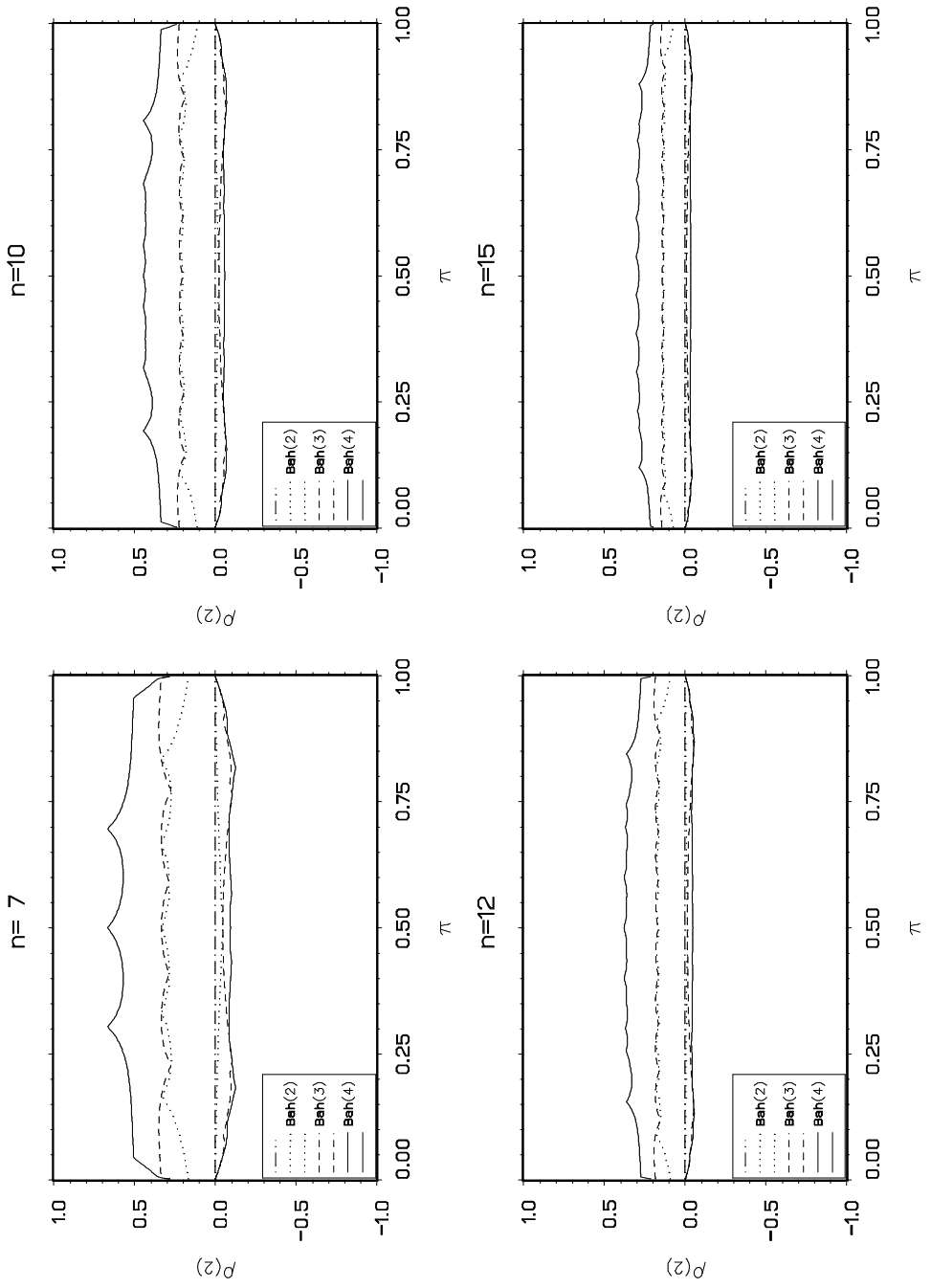

Figure 4.6: *Boundaries for the second order correlation of the two-way, three-way and four-way Bahadur model for some larger litter sizes.*

and two- and three-way correlations are included, still not all positive pairwise correlations are allowed. This is due to the condition that $\rho_{(3)} \in [-1, 1]$.

Constraints on $\rho_{(2)}$ for $n = 4$ are also shown in Figure 4.5. The small values of the upper bound for extreme malformation probabilities are again due to the constraint $-1 \leq \rho_{(3)} \leq 1$.

For larger cluster sizes, boundaries for $\rho_{(2)}$ in a three-way Bahadur model, are represented in Figures 4.5 and 4.6. For these clusters, the upper bound is very similar to the one of Bah(2), except for extreme values of π. Furthermore, it seems that the effect of adding a third order correlation to a two-way Bahadur model, results in an upper bound for $\rho_{(2)}$ being almost independent on π. Compared to Bah(2), the range of negative $\rho_{(2)}$ is enlarged for small and large π.

Next, the focus is on the four-way Bahadur model. The analytical method described in Appendix A for Bah(3), is extended to Bah(4) in Appendix B. Developing first an expression for the constraints on $\rho_{(4)}$, the restrictions on $\rho_{(3)}$ are then derived, which finally result in the bounds for $\rho_{(2)}$. For any specified values of n and π, the lower and upper bound for the pairwise correlation is such that for any $\rho_{(2)}$ between these bounds, there exists at least one pair $(\rho_{(3)}, \rho_{(4)})$ leading to a valid probability mass function. Dealing with large clusters, Figure 4.6 shows that compared to Bah(3), the range of allowable positive pairwise correlations of Bah(4) increases markedly, except for extreme values of π. The range of negative second order correlations remains very narrow.

In principle, constraints on $\rho_{(2)}$ for five- and higher-way Bahadur models, can be calculated by generalizing the analytical procedure given for Bah(4).

Besides this analytical method, also a numerical procedure was developed to compute the bounds for the pairwise correlation in Bah(3) and Bah(4). This second procedure is used to check the calculations of the constraints on $\rho_{(2)}$. The numerical method is described here for Bah(3). First, the upper bound is calculated corresponding to some specified malformation probability. The starting value for the upper bound for $\rho_{(2)}$ is based on expression (4.1). Then, an increment is given to the starting value and by screening the interval [-1,1], a value of $\rho_{(3)}$ leading to a valid probability mass function is searched for. If such a value is found, an increment is given to the improved $\rho_{(2)}$ and the procedure is repeated. Otherwise, step halving is used and it is investigated whether a value of $\rho_{(3)}$ can be found resulting in non-negative probabilities for all outcomes. When improvements of the upper

bound become smaller than some cut-off value, computations corresponding to the specified malformation probability are stopped. Next, an increment is given to π and the values of $\rho_{(2)}$ and $\rho_{(3)}$ corresponding to the previous π, are used as starting values for the current malformation probability. The upper bound for $\rho_{(2)}$ is found for a grid of values of π. An analogous procedure is used to get lower bounds. It turns out that the results of the analytical and numerical procedures are essentially identical.

The findings here are consistent with the results from both the asymptotic study and the analysis of the NTP data. Fitting a Bah(2) null model, the association parameter $\hat{\beta}_2$ captures part of the omitted dose effect. However, due to the (in general severe) restrictions on the second order association, this parameter is tied to a small range. This has some implications when dealing with strong dose effects in the underlying model. On the one hand, this results in values of $\hat{\beta}_2$ being smaller than for the beta-binomial model. On the other hand, the likelihood of the two-way Bahadur null model is smaller than the one of beta-binomial for which the constraints on the association parameter are very mild. In the case of the asymptotic study, the likelihood of the Bah(2) and beta-binomial alternative models are equal when there is no association in the underlying model (as in the settings of Figures 4.1 and 4.2). In the case of the NTP data, the difference between the likelihood of these two alternative models is minor relative to the null models. In conclusion, the likelihood of alternative and null models results in inflated values of the LR statistic when testing the null hypothesis of no dose effect.

In the previous discussion, artificial samples are generated without correlation in the underlying Bahadur model. Now, the focus is on the correlated case. With increasing dose effect, the association parameter of the two-way Bahadur null model will reach more quickly the boundary since there is already an association in the absence of dose effect. Here, an analogous explanation as for the case without association can be given for the behaviour of the LR statistic.

Finally, the inclusion of a third order correlation into a two-way Bahadur model hardly changes the upper bound for $\rho_{(2)}$. This leads to values of the LR statistic being comparable to the ones under Bah(2). When adding a fourth order correlation, the constraints on $\rho_{(2)}$ are relaxed strikingly. As a consequence, the Bah(4) null model is much more likely than the Bah(2) and Bah(3) null models. Hence, the LR statistic results in values closer to the ones under beta-binomial. This finding

clearly needs to be addressed carefully. In order to gain some additional insight, some binomial, beta-binomial and Bahadur distributions are displayed in Figure 4.7. All distributions assume $\pi = 0.5$ and the cluster size is chosen to be $n = 20$. One striking observation is that the probability mass for the Bahadur model with only two-way association is bimodal for $\rho_{(2)}$ sufficiently large. It can be shown that when $\rho_{(2)}$ increases, the trough between the two modi reaches zero when the second order correlation reaches its upper bound, i.e., $\rho_{(2)} = 0.1$. When $\rho_{(3)}$ is added, the mass function is skewed as is obvious from definition (2.2). Considering the curve for $\rho_{(2)} = 0.05$, $\rho_{(3)} = 0$ and $\rho_{(4)} = 0.01$ is very insightful. Indeed, relative to the curve with only two-way association, the bimodal shape has disappeared, the curve is much closer to the binomial model, but the tails are thicker, which is in line with the concept of kurtosis. Thus, it seems that a plausible form of overdispersion is captured, not by merely adding $\rho_{(2)}$, but by adding $\rho_{(2)}$ and $\rho_{(4)}$. Observe however, that the form of this distribution is still fairly different from the beta-binomial one. Since in the analysis of the NTP data, overdispersion seems to be more of an issue than skewness, $\rho_{(3)}$ adds little to the picture in this case. In general, since $\rho_{(3)}$ merely skews the distribution, rather than pulling up the trough, it is not surprising that $\rho_{(3)}$ only marginally relieves the bounds on $\rho_{(2)}$, whereas $\rho_{(4)}$ has a considerably stronger effect. This effect of $\rho_{(4)}$ is seen not only by the disappearance of the bimodal shape; in addition, this unimodal distribution is much closer to the binomial distribution.

4.4 Concluding remarks

Fitting a two-way Bahadur model, an anomalous behaviour of the LR test statistic for the null hypothesis of no dose effect is observed when analysing data from artificial samples and from developmental toxicity studies. Dealing with artificial samples, the LR statistic inflates as the dose effect becomes stronger. Analysing the NTP data, the values of this test statistic are in general strikingly larger than when fitting a beta-binomial model. Adding a third order correlation to the Bahadur model most often results in the same phenomena. However, considering Bah(4), the values of LR are more comparable to the ones under a beta-binomial model.

The behaviour of the LR statistic when fitting a Bahadur model is explained by investigating the parameter space. Requiring a valid probability mass function, the parameters of the Bahadur model are subject to constraints. By means of

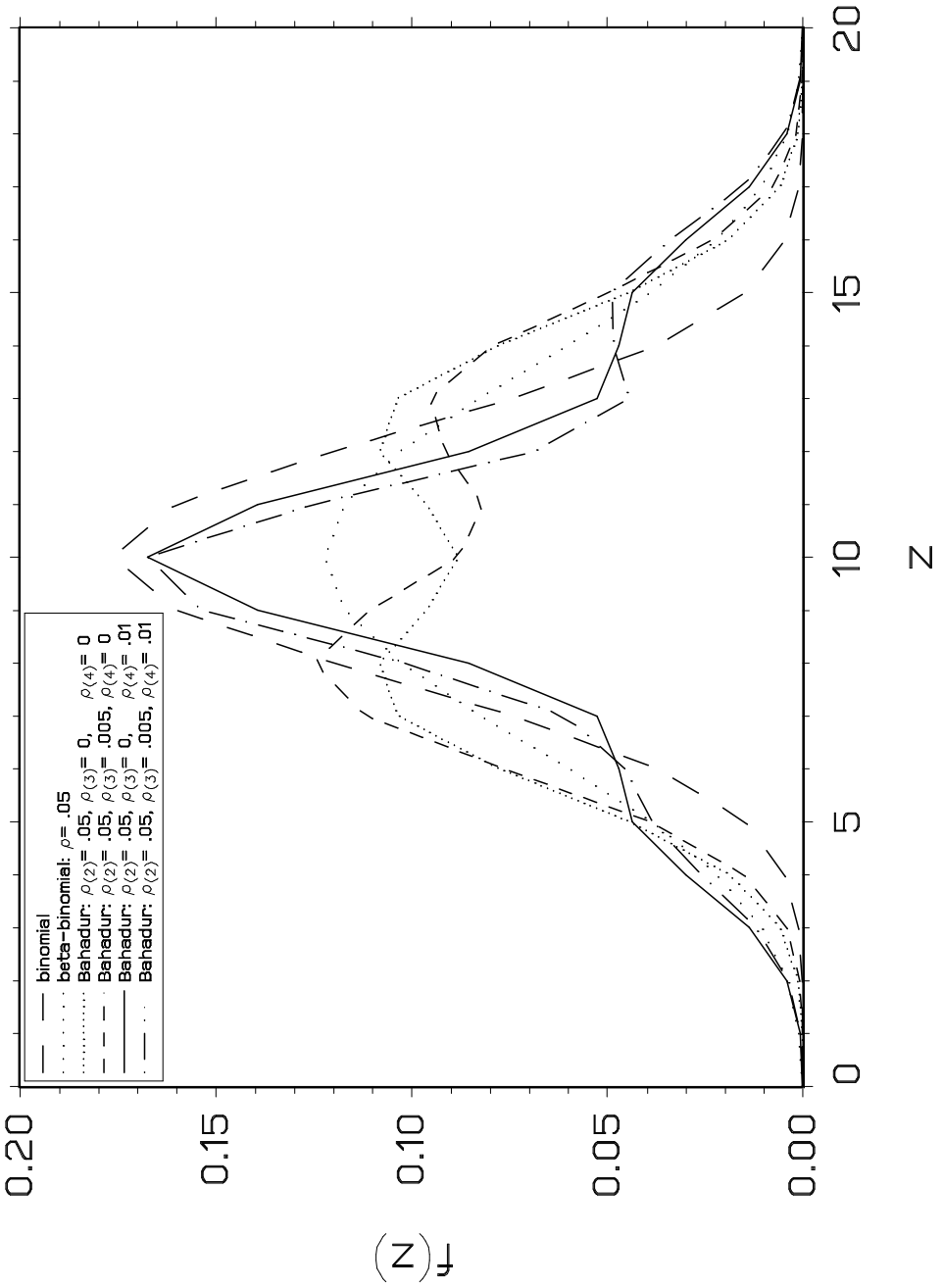

Figure 4.7: *Distribution of binomial, beta-binomial and Bahadur models.*

analytical, numerical and graphical methods, it is shown that the inclusion of a third order association (playing the role of a skewness parameter) does not relax the upper bound of the second order correlation. In Bah(4), the range of positive second order associations is enlarged markedly. The combination of the second and fourth order correlations captures a more standard form of overdispersion (by means of a unimodal distribution) than a model with the second order association parameter only. This form shows a better resemblance with the overdispersion captured by a beta-binomial distribution, although there are still differences. Hence, in comparison with the LR statistic under Bah(2), the LR under Bah(4) is more comparable to the beta-binomial version.

The price to pay for including higher order associations is computational ease. While fitting Bah(2) is already more complex than fitting the beta-binomial model, the conditional model or GEE versions of the Bahadur model (as explained in the previous chapter), even more difficulties are encountered with the Bah(3) and Bah(4) versions. This difficulty is not due to increased computation time, but to the complicated restrictions on the parameter space, which easily leads to divergence. It seems that even careful convergence monitoring is not able to fully relieve this problem.

Appendix A: Restrictions on $\rho_{(2)}$ in a three-way Bahadur model

An analytical method for the bounds of the second order correlation in a three-way Bahadur model is described. Let the coefficient of $\rho_{i(r)}$ in expression (2.2) be denoted by $g_r(\lambda, n, z)$. Hence, the three-way Bahadur model under exchangeability can be written as

$$f(\mathbf{y}) = \pi^z(1-\pi)^{n-z}\left[1 + \sum_{r=2}^{3} \rho_{(r)} g_r(\lambda, n, z)\right].$$

Let the values for n and λ (or equivalently π) be arbitrary but fixed and drop them from notation. Hence, $g_r(\lambda, n, z)$ is abbreviated as $g_r(z)$. Restrictions on the Bahadur model parameters are resulting from the condition that the probability mass function has to be non-negative for all possible outcomes (Bahadur, 1961), which for Bah(3) implies that

$$1 + \rho_{(2)} g_2(z) + \rho_{(3)} g_3(z) \geq 0, \tag{4.2}$$

for $z = 0, 1, \ldots, n$. Let $\boldsymbol{z}_P$, $\boldsymbol{z}_Z$ and $\boldsymbol{z}_N$ be the vectors containing the values of z for which $g_3(z)$ is positive, zero and negative respectively. Denote a general element of $\boldsymbol{z}_P$, $\boldsymbol{z}_Z$ and $\boldsymbol{z}_N$ by z_P, z_Z and z_N respectively. For each of the elements of $\boldsymbol{z}_P$, (4.2) can be expressed as

$$\rho_{(3)} \geq -\frac{1 + \rho_{(2)} g_2(z_P)}{g_3(z_P)}.$$

Analogously for $\boldsymbol{z}_N$, one obtains

$$\rho_{(3)} \leq -\frac{1 + \rho_{(2)} g_2(z_N)}{g_3(z_N)}.$$

Taking into account that $\rho_{(3)} \in [-1, 1]$, the constraints for $\rho_{(3)}$ are:

$$\max\left[\max_{z_P}\left(-\frac{1 + \rho_{(2)} g_2(z_P)}{g_3(z_P)}\right), -1\right] \leq \rho_{(3)} \leq \min\left[\min_{z_N}\left(-\frac{1 + \rho_{(2)} g_2(z_N)}{g_3(z_N)}\right), 1\right]. \tag{4.3}$$

In particular, in the case of clusters of size $n = 3$, expression (4.3) reduces to

$$\begin{aligned}\max(-\lambda^{-1} + (2\lambda^{-1} - \lambda)\rho_{(2)}, -\lambda^3 - 3\lambda\rho_{(2)}, -1) \leq \rho_{(3)} \leq \\ \min(\lambda^{-3} + 3\lambda^{-1}\rho_{(2)}, \lambda + (\lambda^{-1} - 2\lambda)\rho_{(2)}, 1).\end{aligned} \tag{4.4}$$

For clusters of size $n = 4$, the expression analogous to (4.4) depends on π lying in the first, second, third or fourth quarter of the [0,1]-interval. The derivation is straightforward but lengthy and is omitted here.

Given $\rho_{(2)}$ in expression (4.3), there exists at least one value of $\rho_{(3)}$ leading to a valid probability mass function if and only if the lower bound is not larger than the upper bound. Equivalently, both terms on the left hand side have to be smaller than or equal to both terms on the right hand side. First, this implies for any pair (z_P, z_N) that

$$-\frac{1 + \rho_{(2)} g_2(z_P)}{g_3(z_P)} \leq -\frac{1 + \rho_{(2)} g_2(z_N)}{g_3(z_N)}. \tag{4.5}$$

Let

$$\Delta(z_P, z_N) = g_3(z_N) - g_3(z_P)$$

and

$$\tau(z_P, z_N) = g_2(z_N) g_3(z_P) - g_2(z_P) g_3(z_N).$$

Expression (4.5) can then be rewritten as

$$\rho_{(2)}\tau(z_P, z_N) \geq \Delta(z_P, z_N).$$

On the one hand, for any (z_P, z_N) for which $\tau > 0$, it implies that

$$\rho_{(2)} \geq \frac{\Delta(z_P, z_N)}{\tau(z_P, z_N)}.$$

On the other hand, for any (z_P, z_N) for which $\tau < 0$, it results in

$$\rho_{(2)} \leq \frac{\Delta(z_P, z_N)}{\tau(z_P, z_N)}.$$

Based on these inequalities, boundaries for $\rho_{(2)}$ are derived:

$$\max_{(z_P,z_N):\tau>0} \left(\frac{\Delta(z_P, z_N)}{\tau(z_P, z_N)} \right) \leq \rho_{(2)} \leq \min_{(z_P,z_N):\tau<0} \left(\frac{\Delta(z_P, z_N)}{\tau(z_P, z_N)} \right). \tag{4.6}$$

Expression (4.3) also implies for any element of $\boldsymbol{z}_P$ that

$$-\frac{1 + \rho_{(2)} g_2(z_P)}{g_3(z_P)} \leq 1.$$

On the one hand, for any z_P for which $g_2 > 0$, it leads to

$$\rho_{(2)} \geq -\frac{1 + g_3(z_P)}{g_2(z_P)}.$$

On the other hand, for any z_P for which $g_2 < 0$, it results in

$$\rho_{(2)} \leq -\frac{1 + g_3(z_P)}{g_2(z_P)}.$$

Based on these inequalities, boundaries for $\rho_{(2)}$ are derived:

$$\max_{z_P:g_2>0} \left(-\frac{1 + g_3(z_P)}{g_2(z_P)} \right) \leq \rho_{(2)} \leq \min_{z_P:g_2<0} \left(-\frac{1 + g_3(z_P)}{g_2(z_P)} \right). \tag{4.7}$$

Analogously, the condition that

$$-1 \leq -\frac{1 + \rho_{(2)} g_2(z_N)}{g_3(z_N)}$$

for any element of $\boldsymbol{z}_N$, implies that

$$\max_{z_N:g_2>0} \left(-\frac{1 - g_3(z_N)}{g_2(z_N)} \right) \leq \rho_{(2)} \leq \min_{z_N:g_2<0} \left(-\frac{1 - g_3(z_N)}{g_2(z_N)} \right). \tag{4.8}$$

Also the effects of the elements of $\boldsymbol{z}_Z$ on the constraints on $\rho_{(2)}$ need to be studied. Expression (4.2) then reduces to

$$1 + \rho_{(2)} g_2(z_Z) \geq 0$$

for any z_Z. For z_Z for which $g_2 > 0$, it leads to

$$\rho_{(2)} \geq -\frac{1}{g_2(z_Z)},$$

while for z_Z for which $g_2 < 0$, it results in

$$\rho_{(2)} \leq -\frac{1}{g_2(z_Z)}.$$

Based on these inequalities, boundaries for $\rho_{(2)}$ are derived:

$$\max_{z_Z : g_2 > 0} \left(-\frac{1}{g_2(z_Z)} \right) \leq \rho_{(2)} \leq \min_{z_Z : g_2 < 0} \left(-\frac{1}{g_2(z_Z)} \right). \tag{4.9}$$

From (4.6)–(4.9) and the constraint $-1 \leq \rho_{(2)} \leq 1$, it follows that the upper and lower bound for $\rho_{(2)}$ in a three-way Bahadur model can be written as

$$\max \left[\max_{(z_P, z_N) : \tau > 0} \left(\frac{\Delta}{\tau} \right), \max_{z_P : g_2 > 0} \left(-\frac{1 + g_3}{g_2} \right), \max_{z_N : g_2 > 0} \left(-\frac{1 - g_3}{g_2} \right), \max_{z_Z : g_2 > 0} \left(-\frac{1}{g_2} \right), -1 \right]$$
$$\leq \rho_{(2)} \leq$$
$$\min \left[\min_{(z_P, z_N) : \tau < 0} \left(\frac{\Delta}{\tau} \right), \min_{z_P : g_2 < 0} \left(-\frac{1 + g_3}{g_2} \right), \min_{z_N : g_2 < 0} \left(-\frac{1 - g_3}{g_2} \right), \min_{z_Z : g_2 < 0} \left(-\frac{1}{g_2} \right), 1 \right].$$

Appendix B: Restrictions on $\rho_{(2)}$ in a four-way Bahadur model

In contrast with Appendix A, this appendix deals with an analytical procedure for the derivation of the constraints on the second order correlation in a four-way Bahadur model. Again, represent the coefficient of $\rho_{i(r)}$ in formula (2.2) by $g_r(\lambda, n, z)$. One can then express the four-way Bahadur model under exchangeability as

$$f(\mathbf{y}) = \pi^z (1 - \pi)^{n - z} \left[1 + \sum_{r=2}^{4} \rho_{(r)} g_r(\lambda, n, z) \right].$$

Here, the values for n and λ are arbitrary but fixed and hence, the coefficient $g_r(\lambda, n, z)$ is represented by $g_r(z)$. Constraints on the parameters of the Bahadur model are due to the condition that the density function needs to be non-negative for all outcomes (Bahadur, 1961), which for the four-way Bahadur model results in

$$1 + \rho_{(2)} g_2(z) + \rho_{(3)} g_3(z) + \rho_{(4)} g_4(z) \geq 0, \tag{4.10}$$

for $z = 0, 1, \ldots, n$. The vectors containing the values of z for which $g_4(z)$ is positive, zero and negative are denoted by $\boldsymbol{z}_P$, $\boldsymbol{z}_Z$ and $\boldsymbol{z}_N$ respectively. Let a general element of $\boldsymbol{z}_P$, $\boldsymbol{z}_Z$ and $\boldsymbol{z}_N$ be represented by z_P, z_Z and z_N respectively. Then, expression (4.10) can be written as

$$\rho_{(4)} \geq -\frac{1 + \rho_{(2)} g_2(z_P) + \rho_{(3)} g_3(z_P)}{g_4(z_P)}$$

for each of the elements of $\boldsymbol{z}_P$. For $\boldsymbol{z}_N$, it follows from (4.10) that

$$\rho_{(4)} \leq -\frac{1 + \rho_{(2)} g_2(z_N) + \rho_{(3)} g_3(z_N)}{g_4(z_N)}.$$

Adding the constraint $\rho_{(4)} \in [-1, 1]$, the restrictions on the four-way correlation coefficient are:

$$\begin{aligned} \max\left[\max_{z_P}\left(-\frac{1 + \rho_{(2)} g_2(z_P) + \rho_{(3)} g_3(z_P)}{g_4(z_P)}\right), -1\right] \\ \leq \rho_{(4)} \leq \\ \min\left[\min_{z_N}\left(-\frac{1 + \rho_{(2)} g_2(z_N) + \rho_{(3)} g_3(z_N)}{g_4(z_N)}\right), 1\right]. \end{aligned} \tag{4.11}$$

For a particular value of $\rho_{(2)}$ and $\rho_{(3)}$ in expression (4.11), there exists at least one value of $\rho_{(4)}$ resulting in a valid density function if and only if the lower bound in (4.11) is not larger than the upper bound. Hence, each of the terms on the left hand side needs to be smaller than or equal to each of the terms on the right hand side. This implies among others, that for any pair (z_P, z_N),

$$-\frac{1 + \rho_{(2)} g_2(z_P) + \rho_{(3)} g_3(z_P)}{g_4(z_P)} \leq -\frac{1 + \rho_{(2)} g_2(z_N) + \rho_{(3)} g_3(z_N)}{g_4(z_N)}. \tag{4.12}$$

Let

$$\Delta(z_P, z_N, \rho_{(2)}) = g_4(z_N)\left[1 + \rho_{(2)} g_2(z_P)\right] - g_4(z_P)\left[1 + \rho_{(2)} g_2(z_N)\right] \tag{4.13}$$

and

$$\tau(z_P, z_N) = g_3(z_N)g_4(z_P) - g_3(z_P)g_4(z_N). \tag{4.14}$$

One can then reexpress formula (4.12) as

$$\rho_{(3)}\tau(z_P, z_N) \geq \Delta(z_P, z_N, \rho_{(2)}).$$

For any (z_P, z_N) for which $\tau > 0$, it results in

$$\rho_{(3)} \geq \frac{\Delta(z_P, z_N, \rho_{(2)})}{\tau(z_P, z_N)}.$$

Also, for any (z_P, z_N) for which $\tau < 0$, it implies that

$$\rho_{(3)} \leq \frac{\Delta(z_P, z_N, \rho_{(2)})}{\tau(z_P, z_N)}.$$

From these inequalities, constraints on the third order correlation coefficient are obtained:

$$\max_{(z_P,z_N):\tau>0} \left(\frac{\Delta(z_P, z_N, \rho_{(2)})}{\tau(z_P, z_N)} \right) \leq \rho_{(3)} \leq \min_{(z_P,z_N):\tau<0} \left(\frac{\Delta(z_P, z_N, \rho_{(2)})}{\tau(z_P, z_N)} \right). \tag{4.15}$$

Formula (4.11) also results in

$$-\frac{1 + \rho_{(2)}g_2(z_P) + \rho_{(3)}g_3(z_P)}{g_4(z_P)} \leq 1$$

for any element of $\boldsymbol{z}_P$. On the one hand, for any z_P for which $g_3 > 0$, it implies that

$$\rho_{(3)} \geq -\frac{1 + g_4(z_P) + \rho_{(2)}g_2(z_P)}{g_3(z_P)}.$$

On the other hand, for any z_P for which $g_3 < 0$, it leads to

$$\rho_{(3)} \leq -\frac{1 + g_4(z_P) + \rho_{(2)}g_2(z_P)}{g_3(z_P)}.$$

Based on these inequalities, restrictions on $\rho_{(3)}$ are found:

$$\max_{z_P:g_3>0} \left(-\frac{1 + g_4(z_P) + \rho_{(2)}g_2(z_P)}{g_3(z_P)} \right) \leq \rho_{(3)} \leq \min_{z_P:g_3<0} \left(-\frac{1 + g_4(z_P) + \rho_{(2)}g_2(z_P)}{g_3(z_P)} \right). \tag{4.16}$$

In an analogous way, the condition that

$$-1 \leq -\frac{1 + \rho_{(2)}g_2(z_N) + \rho_{(3)}g_3(z_N)}{g_4(z_N)}$$

for any element of $\boldsymbol{z}_N$, results in

$$\max_{z_N:g_3>0}\left(-\frac{1-g_4(z_N)+\rho_{(2)}g_2(z_N)}{g_3(z_N)}\right)\leq\rho_{(3)}\leq\min_{z_N:g_3<0}\left(-\frac{1-g_4(z_N)+\rho_{(2)}g_2(z_N)}{g_3(z_N)}\right). \tag{4.17}$$

Also the effects of the elements of $\boldsymbol{z}_Z$ on the boundaries for the third order correlation have to be considered. Formula (4.10) then simplifies to

$$1+\rho_{(2)}g_2(z_Z)+\rho_{(3)}g_3(z_Z)\geq 0$$

for any z_Z. For z_Z for which $g_3>0$, it implies that

$$\rho_{(3)}\geq-\frac{1+\rho_{(2)}g_2(z_Z)}{g_3(z_Z)},$$

while for z_Z for which $g_3<0$, it results in

$$\rho_{(3)}\leq-\frac{1+\rho_{(2)}g_2(z_Z)}{g_3(z_Z)}.$$

From these inequalities, restrictions on $\rho_{(3)}$ are derived:

$$\max_{z_Z:g_3>0}\left(-\frac{1+\rho_{(2)}g_2(z_Z)}{g_3(z_Z)}\right)\leq\rho_{(3)}\leq\min_{z_Z:g_3<0}\left(-\frac{1+\rho_{(2)}g_2(z_Z)}{g_3(z_Z)}\right). \tag{4.18}$$

Based on (4.15)–(4.18) and the restriction $-1\leq\rho_{(3)}\leq 1$, the lower and upper bound for the third order correlation coefficient in a four-way Bahadur model can be expressed as

$$\max\left[\max_{(z_P,z_N):\tau>0}\left(\frac{\Delta}{\tau}\right),\max_{z_P:g_3>0}\left(-\frac{1+g_4+\rho_{(2)}g_2}{g_3}\right),\max_{z_N:g_3>0}\left(-\frac{1-g_4+\rho_{(2)}g_2}{g_3}\right),\max_{z_Z:g_3>0}\left(-\frac{1+\rho_{(2)}g_2}{g_3}\right),-1\right]$$

$$\leq\rho_{(3)}\leq$$

$$\min\left[\min_{(z_P,z_N):\tau<0}\left(\frac{\Delta}{\tau}\right),\min_{z_P:g_3<0}\left(-\frac{1+g_4+\rho_{(2)}g_2}{g_3}\right),\min_{z_N:g_3<0}\left(-\frac{1-g_4+\rho_{(2)}g_2}{g_3}\right),\min_{z_Z:g_3<0}\left(-\frac{1+\rho_{(2)}g_2}{g_3}\right),1\right].$$

Based on the previous formula, restrictions on the second order correlation coefficient can be derived. For a particular value of $\rho_{(2)}$, there exists at least one value of $\rho_{(3)}$ resulting in a valid distribution if and only if the lower bound for $\rho_{(3)}$ is smaller than or equal to the upper bound. Hence, each of the five terms in the lower bound

of $\rho_{(3)}$ needs to be smaller than or equal to each of the five terms in the upper bound. More specifically, this implies among others, that

$$\max_{(z_P,z_N):\tau>0}\left(\frac{\Delta(z_P,z_N,\rho_{(2)})}{\tau(z_P,z_N)}\right) \leq \min_{(z_P,z_N):\tau<0}\left(\frac{\Delta(z_P,z_N,\rho_{(2)})}{\tau(z_P,z_N)}\right). \tag{4.19}$$

Represent a pair (z_P, z_N) for which $\tau > 0$ by (z_{P1}, z_{N1}) and a pair (z_P, z_N) for which $\tau < 0$ by (z_{P2}, z_{N2}). Inequality (4.19) can be reexpressed as

$$\frac{\Delta(z_{P1},z_{N1},\rho_{(2)})}{\tau(z_{P1},z_{N1})} \leq \frac{\Delta(z_{P2},z_{N2},\rho_{(2)})}{\tau(z_{P2},z_{N2})} \tag{4.20}$$

for any combination of (z_{P1}, z_{N1}) and (z_{P2}, z_{N2}). Let

$$\begin{aligned}\nu(z_{P1},z_{N1},z_{P2},z_{N2}) &= \tau(z_{P1},z_{N1})\left[g_2(z_{N2})g_4(z_{P2}) - g_2(z_{P2})g_4(z_{N2})\right] \\ &- \tau(z_{P2},z_{N2})\left[g_2(z_{N1})g_4(z_{P1}) - g_2(z_{P1})g_4(z_{N1})\right]\end{aligned}$$

and

$$\begin{aligned}\omega(z_{P1},z_{N1},z_{P2},z_{N2}) &= \tau(z_{P1},z_{N1})\left[g_4(z_{N2}) - g_4(z_{P2})\right] \\ &- \tau(z_{P2},z_{N2})\left[g_4(z_{N1}) - g_4(z_{P1})\right].\end{aligned}$$

Using also expression (4.13) for $\Delta(z_P, z_N, \rho_{(2)})$, it follows from (4.20) that

$$\rho_{(2)}\nu(z_{P1},z_{N1},z_{P2},z_{N2}) \geq \omega(z_{P1},z_{N1},z_{P2},z_{N2}).$$

For any $(z_{P1}, z_{N1}, z_{P2}, z_{N2})$ for which $\nu > 0$, it implies that

$$\rho_{(2)} \geq \frac{\omega(z_{P1},z_{N1},z_{P2},z_{N2})}{\nu(z_{P1},z_{N1},z_{P2},z_{N2})}.$$

Furthermore, for any $(z_{P1}, z_{N1}, z_{P2}, z_{N2})$ for which $\nu < 0$, it follows that

$$\rho_{(2)} \leq \frac{\omega(z_{P1},z_{N1},z_{P2},z_{N2})}{\nu(z_{P1},z_{N1},z_{P2},z_{N2})}.$$

Based on these inequalities, constraints for $\rho_{(2)}$ are derived:

$$\max_{(z_{P1},z_{N1},z_{P2},z_{N2}):\nu>0}\left(\frac{\omega(z_{P1},z_{N1},z_{P2},z_{N2})}{\nu(z_{P1},z_{N1},z_{P2},z_{N2})}\right) \leq \rho_{(2)} \leq$$

$$\min_{(z_{P1},z_{N1},z_{P2},z_{N2}):\nu<0}\left(\frac{\omega(z_{P1},z_{N1},z_{P2},z_{N2})}{\nu(z_{P1},z_{N1},z_{P2},z_{N2})}\right).$$

In an analogous way, the other constraints for the second order correlation coefficient in a four-way Bahadur model are obtained. The derivation is straightforward but tedious and hence, that part is omitted here.

Chapter 5

Implications of misspecifying the likelihood on safe dose determination

In Chapters 3 and 4, the focus was on dose-response modelling of data from developmental toxicity experiments. Characteristics of the Bahadur, beta-binomial and conditional models and the behaviour of the Wald and likelihood ratio test statistics were studied for the NTP data, as for simulated small samples and in the large sample context.

Besides investigating suitable dose-response models, another issue is quantitative risk assessment. This critically important area of risk assessment is based on the relationship between dose and response to derive a safe dose. In quantitative risk assessment, there are a number of choices which have to be made, resulting in a variety of approaches.

First, safe exposure levels can be derived from the NOAEL-safety factor approach, as discussed in Section 1.6. Alternatively, dose-response modelling can be used to determine safe doses. Due to the disadvantages of the NOAEL and the benefits of dose-response models, this chapter and the following one are concerned with statistical procedures to predict safe exposure levels based on the modelling approach.

Secondly, there are several ways to handle clustering. While dose-response modelling is relatively straightforward in uncorrelated settings, it is less so in the clustered context. Of course, one can ignore the clustering altogether by treating the littermates as if they were independent. Also, the litter effect issue can be avoided by modelling the probability of an affected litter. Such models are generally too

simplistic but there is a multitude of models which do consider clustering. As discussed in Chapter 2, different types of models (marginal, random effects, conditional) for clustered binary data can be formulated. In this chapter, the Bahadur, beta-binomial and conditional probability models are considered. The discussion is briefly extended with another marginal model introduced by George and Bowman (1995). Besides the choice of an appropriate dose-response model, model parameters can be estimated via several inferential procedures. Estimation methods range from full likelihood to quasi-likelihood and generalized estimating equations. Dealing with quantitative risk assessment in this thesis, parameters are estimated using maximum likelihood methodology. Furthermore, the implications of fitting some model (rather than basing on the unknown data generating mechanism) on the assessment of safe doses can be investigated. While in Chapter 3, the effect of misspecifying the parametric response model on the assessment of dose effect was investigated, this chapter focuses on the implications of likelihood misspecification on the estimation of a safe dose.

Thirdly, quantitative risk assessment can be based on either fetus or litter-based risks. In order to perform dose-response modelling and assessment of safe doses, most authors take a fetus-based perspective, where the excess risk over background for an affected fetus is determined as a function of dose. In this chapter, safe doses and lower confidence limits are computed for several types of malformation, using a fetus-based approach. For models formulated in terms of a fetus-specific marginal probability, this is particularly straightforward. However, a disadvantage of this approach is that it may raise biological questions. From a biological perspective, modelling litter-based excess risks is a very appealing alternative. Arguably, the entire litter is more representative of a human pregnancy than a single fetus. The following chapter will contrast fetus and litter-based perspectives.

Fourthly, one needs to acknowledge the stochastic nature of the number of implants and the number of viable fetuses (i.e., the litter size) in a dam. Some methods (Ryan, 1992) condition on the observed litter size when modelling the number of malformations. Others (e.g., Catalano *et al.*, 1993) allow response rates to depend on litter size and then calculate a safe dose at an "average" litter size, thereby avoiding the need for direct adjustment. Krewski and Zhu (1995) use a model formulation that causes litter size to drop from the expression for excess risk. Rai and Van Ryzin (1985) compute risks by integrating over the litter size distribution. This approach

will be used in this chapter and the following one.

The relatively complex data structure forces several other decisions: (1) Are linear or non-linear predictors used ? (2) Are the malformation indices studied separately, collapsed into a single indicator or treated as a multivariate outcome vector per fetus within a litter ? (3) Is death ignored, studied separately without taking into account malformations among the viable fetuses, combined with a collapsed malformation indicator into a new indicator for *abnormality* (i.e., death or malformation) or studied jointly with the malformation outcomes ? (4) Are continuous responses, such as birth weight, excluded from the model or not ? Chen *et al.* (1991), Ryan (1992), Catalano *et al.* (1993), Krewski and Zhu (1994) discuss statistical models that allow for exposure effects on death and malformation, formulating the problem as a trinomial model with overdispersion. Catalano and Ryan (1992) and Catalano *et al.* (1993) propose models that incorporate fetal weight in addition to death and malformation. In this chapter and in Chapter 6, linear predictors are used for the parameters of the implemented models. Here, the malformation indicators external, skeletal and visceral are analysed, as well as a collapsed malformation indicator. The developmental toxicity endpoints "death" and "fetal birth weight" are not highlighted here.

In Section 5.1, two approaches to calculate a safe dose are followed. First, as suggested by many authors (e.g., Crump and Howe, 1985), a likelihood ratio based version is discussed. Secondly, an alternative method based on profile likelihood is explored. For both methods, two versions are contrasted: entirely model based and linearly extrapolated. In Section 5.2, the different methods are compared by asymptotic calculations based on a method of Rotnitzky and Wypij (1994). Their technique is adapted to compute "asymptotic values" of several safe dose estimators. These results are then contrasted with analyses of the NTP data in Section 5.3.

5.1 Determination of a safe dose

Suppose one wishes to estimate a safe level of exposure, based on the models discussed in Chapter 2. The standard approach to quantitative risk assessment requires the specification of an adverse event, along with the risk of this event expressed as a function of dose. In this chapter, events of interest are malformed fetuses according to a specific type and malformation according to any type. The risk $r(d)$ represents

the probability that the event of malformation occurs at dose level d. In this chapter, the focus is on the fetus-based risk, which is the probability that a fetus has the considered adverse event of malformation. In the Bahadur model, this risk equals the model parameter π, being the marginal probability of a malformed fetus. Basing on the modelling of the Bahadur parameters as expressed in (2.5),

$$r(d) = \frac{1}{1 + \exp(-\beta_0 - \beta_d d)}. \tag{5.1}$$

As a consequence, the background risk

$$r(0) = \frac{1}{1 + \exp(-\beta_0)}. \tag{5.2}$$

The same formulas are derived in case of the beta-binomial model. When dealing with the George-Bowman model and its folded logistic parameterization as introduced in Section 2.4, the risk is

$$r(d) = \lambda_1(d) = \frac{2}{1 + 2^{-\beta_0 - \beta_d d}}. \tag{5.3}$$

The background risk then equals

$$r(0) = \lambda_1(0) = \frac{2}{1 + 2^{-\beta_0}}. \tag{5.4}$$

For the conditional model, the expression for the risk is more complicated. Using (2.17), the probability of malformation at dose d for this model can be written as

$$r(d) = \sum_n \frac{nP(n)}{\sum_n nP(n)} \left(\frac{\sum_{z=0}^n z \binom{n}{z} \exp\{(\beta_0 + \beta_d d)z - \beta_2 z(n-z)\}}{\sum_{z=0}^n n \binom{n}{z} \exp\{(\beta_0 + \beta_d d)z - \beta_2 z(n-z)\}} \right), \tag{5.5}$$

where $\{P(n)\}_n$ is the probability distribution of the litter sizes. This unknown distribution is estimated here using the local linear smoothed frequencies of the number of viables, as presented in Table 3.1 (Aerts, Augustyns and Janssen, 1997). The background rate in the conditional model can be written as:

$$r(0) = \sum_n \frac{nP(n)}{\sum_n nP(n)} \left(\frac{\sum_{z=0}^n z \binom{n}{z} \exp\{\beta_0 z - \beta_2 z(n-z)\}}{\sum_{z=0}^n n \binom{n}{z} \exp\{\beta_0 z - \beta_2 z(n-z)\}} \right). \tag{5.6}$$

Instead of the risk $r(d)$ itself, one might prefer to use the additive excess risk, which is the excess risk above the background rate, i.e., $r(d) - r(0)$. Assuming that at any non-zero value of dose, the chemical under investigation has more toxic effects than at dose level 0, the additive excess risk function ranges from 0 to $1 - r(0)$. This type of risk does not relate the difference in risk at dose d and at dose 0 to the background rate. This is in contrast with the relative excess risk function $r^*(d)$. It is a "multiplicative" risk function, measuring the relative increase in risk above background and is defined as (Crump, 1984)

$$r^*(d) = \frac{r(d) - r(0)}{1 - r(0)}. \tag{5.7}$$

In this thesis, $r^*(d)$ is called the *excess risk*. Assuming again that the chemical results in more adverse effects at non-zero dose d compared to dose level 0, the excess risk ranges from 0 to 1. An expression for the excess risk in the Bahadur and the beta-binomial models can be derived from formulas (5.1) and (5.2):

$$r^*(d) = \frac{1 - \exp(-\beta_d d)}{1 + \exp(-\beta_0 - \beta_d d)}. \tag{5.8}$$

In the George-Bowman model, the excess risk can be computed using (5.3) and (5.4):

$$r^*(d) = \frac{2(2^{-\beta_d d} - 1)}{(2^{-\beta_d d} + 2^{\beta_0})(1 - 2^{-\beta_0})}. \tag{5.9}$$

From the relationship $r^*(d)$, a safe level of exposure can be determined. The terminology used to describe a "virtually safe dose" or a "benchmark dose" is not standardized and depends on the area of application (carcinogenicity, developmental toxicity) and the regulatory authorities involved (Environmental Protection Agency, Food and Drug Administration,...). A useful overview is given in Williams and Ryan (1997). A benchmark dose is defined as the statistical lower confidence limit on a dose corresponding to a risk in the range of 1 to 10% (Crump, 1984). The *virtually safe dose* (VSD) can be defined in several ways (Crump and Howe, 1985; Gart *et al.*, 1986; Chen and Kodell, 1989). For instance, it can be defined as the lower confidence limit on a dose corresponding to an excess risk of 10^{-4}. The dose itself is referred to here as the *effective dose* (ED). Hence, the ED is the dose at which the excess risk over the background rate is small, say 10^{-4} and the VSD is the lower confidence limit of the effective dose.

Using $r(d) \equiv r(d;\boldsymbol{\beta})$, i.e., the marginal probability of a malformed fetus at dose level d corresponding to the parameter vector $\boldsymbol{\beta}$ in the model considered, the ED can be defined as the value d that solves $r^*(d;\boldsymbol{\beta}) = 10^{-4}$. Equating (5.8) to 10^{-4}, the ED in case of a Bahadur or a beta-binomial model can be calculated:

$$ED_{Bah,BB} = \frac{\ln\left(\frac{1+10^{-4}\exp(-\beta_0)}{1-10^{-4}}\right)}{\beta_d}. \tag{5.10}$$

Analogously, the ED in case of the George-Bowman model is derived from (5.9):

$$ED_{GB} = \frac{\log_2\left(\frac{2+10^{-4}(2^{-\beta_0}-1)}{2+10^{-4}(2^{\beta_0}-1)}\right)}{\beta_d}. \tag{5.11}$$

From (5.5) and (5.6), it is clear that there is no simple expression for the effective dose when dealing with the conditional model.

The ML estimate of the effective dose is the solution to $\hat{r}^*(d) = r^*(d;\hat{\boldsymbol{\beta}}) = 10^{-4}$ where $\hat{\boldsymbol{\beta}}$ is the ML estimate of $\boldsymbol{\beta}$. As a consequence, for the Bahadur and the beta-binomial models, the ML estimate of the ED is found by replacing β_0 and β_d in expression (5.10) by their ML estimates. Analogously, the ML estimate of the ED in case of the George-Bowman model is obtained from (5.11).

For setting confidence limits in low dose extrapolation, i.e., to determine the VSD, two approaches are considered. Crump and Howe (1985) recommend to use the asymptotic distribution of the likelihood ratio. According to this method, an approximate $100(1-\alpha)\%$ lower confidence limit for the ED (denoted by VSD(1)), corresponding to an excess risk of 10^{-4}, is defined as

$$\min\{d(\boldsymbol{\beta}) : r^*(d;\boldsymbol{\beta}) = 10^{-4} \text{ over all } \boldsymbol{\beta} \text{ such that } 2(\ell(\hat{\boldsymbol{\beta}}) - \ell(\boldsymbol{\beta})) \leq \chi^2_p(1-2\alpha)\}, \tag{5.12}$$

with p the number of regression parameters. This might imply that a dose-response model with more regression parameters (and thus more uncertainty) leads to a larger confidence region and thus to a smaller VSD. For the Bahadur and beta-binomial models, the 97.5% lower limit can be written explicitly as

$$\min\left\{\frac{\ln\left(\frac{1+10^{-4}e^{-\beta_0}}{1-10^{-4}}\right)}{\beta_d} \text{ over all } \beta_0, \beta_d \text{ s.t. } 2(\ell(\hat{\beta}_0,\hat{\beta}_d,\hat{\beta}_2) - \ell(\beta_0,\beta_d,\beta_2)) \leq \chi^2_3(0.95)\right\}, \tag{5.13}$$

with $\ell = \sum_{i=1}^{N} \ell_i$ determined by (2.3) or (2.10) and with $\chi^2_3(0.95) = 7.81$. An analogous expression for the George-Bowman model is

$$\min \left\{ \frac{\log_2 \left(\frac{2+10^{-4}(2^{-\beta_0}-1)}{2+10^{-4}(2^{\beta_0}-1)} \right)}{\beta_d} \text{ over all } \beta_0, \beta_d \text{ s.t. } 2(\ell(\hat{\beta}_0, \hat{\beta}_d) - \ell(\beta_0, \beta_d)) \leq \chi^2_2(0.95) \right\}, \tag{5.14}$$

where $\ell = \sum_{i=1}^{N} \ell_i$ is based on (2.7) and where $\chi^2_2(0.95) = 5.99$. The procedure is somewhat more involved for the conditional model. The expressions (5.5) and (5.6) for the risk in the conditional model at dose levels d and zero respectively, are plugged into (5.7), whence (5.12) is solved numerically. Of course, it is crucial that a likelihood ratio test be available, making the method less straightforward to use in non-likelihood settings. For pseudo-likelihood, the proposal of Geys, Molenberghs and Ryan (1999) for pseudo-likelihood ratio tests could be followed.

A second approach, denoted by VSD(2), is based on the profile likelihood method (Morgan, 1992). This procedure is explained first for the Bahadur, beta-binomial and conditional models. For a specified excess risk, three parameters fully specify the set of four $(\beta_0, \beta_d, \beta_2, d)$ parameters, i.e., given the parameters $\boldsymbol{\beta}$, the corresponding dose d is uniquely determined. Similarly, β_0, β_2 and d determine β_d. In other words, given β_0 and β_2, either member of the pair (β_d, d) contains the same information, provided a monotonic relationship exists between β_d and d. For the Bahadur and beta-binomial models, expression (5.10) shows that the relationship between β_d and d is indeed monotone. For the conditional model however, this transformation is most often not monotone. Therefore, VSD(2) will not be applied to this model. In case the relationship between β_d and d is monotonic, the following procedure is suggested. First, transform the likelihood depending on β_0, β_d and β_2 to a likelihood expressed in terms of β_0, d and β_2. Next, calculate the logarithm of the profile likelihood, $l_P(d)$, by maximizing the logarithm of the transformed likelihood over β_0 and β_2:

$$l_P(d) = \max_{\beta_0, \beta_2} l(\beta_0, d, \beta_2).$$

The final step in the construction of a profile likelihood based confidence interval for the effective dose is the computation of its $100(1-\alpha)\%$ lower limit. The lower bound VSD(2) satisfies the conditions

$$VSD(2) < \hat{ED} \quad \text{and} \quad 2(l_P(\hat{ED}) - l_P(VSD(2))) = \chi^2_1(1-2\alpha).$$

Hence, constructing the lower limit of a 97.5% confidence interval for the effective

dose is based on the percentile $\chi^2_1(0.95) = 3.84$. In case of the George-Bowman model, the profile likelihood procedure is analogous to the method described above, except that $\boldsymbol{\beta}$ consists only of β_0 and β_d. Formula (5.11) expresses the relationship between β_d and d for this model.

A variation on this theme, suggested by many authors (Chen and Kodell, 1989; Gaylor, 1989; Ryan, 1992), first determines a lower confidence limit, e.g., corresponding to an excess risk of 1% and then linearly extrapolates it to a VSD. The main advantage quoted for this procedure is that the determination of a VSD is less model dependent. This assertion will be explored in the next sections.

Several other methods have been proposed. Using the delta method, a Wald based version can be obtained. Several authors have indicated that this method suffers from drawbacks, especially with low dose extrapolation (Krewski and Van Ryzin, 1981; Crump, 1984; Crump and Howe, 1985). One of the disadvantages of a Wald based confidence interval for the effective dose is that its lower limit may fail to be positive. The NOAEL provides another alternative. This method is discouraged by several authors (Kimmel and Gaylor, 1988). It is very ad hoc and often leads to a VSD which is considerably larger than with the dose-response based methods (Gaylor, 1989).

In the next section, the ED will be applied in an asymptotic study, while both procedures VSD(1) and VSD(2) will be applied to the NTP data in Section 5.3.

5.2 Asymptotic study

Asymptotic information on the implications of model misspecification on the assessment of safe doses, is based on the ideas of Rotnitzky and Wypij (1994), as explained in Section 3.1. An artificial sample is constructed, in which each possible realization of dose d_i, number of viables n_i and number of malformed fetuses z_i, is weighted according to its probability under a given underlying model. The joint probability $f(d_i, n_i, z_i)$ is factorized in the same way as in Section 3.1. This technique is adapted here to compute "asymptotic" dose values. Notice that this will not be a lower limit and hence not a VSD, since asymptotically it can be determined without error. Therefore, in this section, the focus is on the effective dose (ED). Data are generated from Bahadur (Bah), beta-binomial (BB) and conditional (Cond) models in turn. For each sample, the ED is determined under all three models, based on an

entirely model based estimator, as well as on an extrapolated version of the 1% ED.

Again, four dose groups are considered, with one control group ($d_i = 0$) and three active groups ($d_i = 0.25, 0.5, 1$). Here, the distribution of the number of viable fetuses in a dam, is a local linear smoothed version of the relative frequency distribution given in Table 1 of Kupper *et al.* (1986). The smoothed frequencies are presented in Table 3.1.

For the Bahadur and beta-binomial models, the selected intercept β_0 equals -4.5 and corresponds to a low baseline adverse event rate of 1%. The baseline malformation rate of the conditional model depends on both intercept (here, $\beta_0 = -2.5$) and association parameter. A grid of values for dose effect is considered. Table 3.2 represents the parameter settings used in this simulation study. More details about the choice of the values for the transformed correlations were given in Section 3.1.

Results are graphically summarized in Figures 5.1 (Bahadur), 5.2 (beta-binomial) and 5.3 (conditional). The "true" ED is found by fitting the correct model (i.e., the model under which the data were generated) and by calculating the purely model based ED. This is the Bahadur model in Figure 5.1, the beta-binomial model in Figure 5.2 and the conditional model in Figure 5.3.

First, the results from the three underlying models under investigation are compared to each other. When the underlying model is Bahadur or beta-binomial, the effective doses are the same in case of no association in the true model. This is expected because both models then reduce to logistic regression and since the same intercept parameter is considered ($\beta_0 = -4.5$). When there is association between the littermates in the underlying Bahadur model ($\beta_2 = 0.1$), the results are very close to the ones of the beta-binomial model. One also notices that in case of the underlying conditional model, the effective doses are larger than for the other two models.

Secondly, the curves for a particular underlying model are investigated for both values of the association parameter. When the true model is Bahadur or beta-binomial and there is no association between littermates, there are only two curves: the model based and the extrapolated version. The latter one yields the lowest ED. When association between littermates is introduced, the curves separate slightly. When the true model is conditional and no association is assumed, all procedures yield virtually the same results. A somewhat different conclusion is obtained for as-

sociated outcomes: essentially the model-based and extrapolated curves are grouped.

Thus, typically the extrapolated version is lower than the model based one. While this may seem cautious, one should not forget that in this study the true ED is known (the model-based version under the correct model) and hence, the extrapolated version is found to be too low here. In this respect, Morgan (1992) (p. 175) and the Scientific Committee of the Food Safety Council (1980) point out that blind adherence to a conservative procedure is to be regarded as scientifically indefensible.

5.3 Analysis of NTP data

The knowledge gathered from the asymptotic study is supplemented with analyses of the NTP data. The studies investigate the effects in mice of the toxic agents DEHP, DYME and EG. Details were provided in Section 1.4.

Two estimates of the effective dose corresponding to an excess risk of 10^{-4} are provided. Besides an entirely model based (MB) effective dose, a linear extrapolation (EP) version is computed. Furthermore, four quantities for the lower confidence limit of the ED are determined. In addition to the determination of the confidence region based VSD(1), the profile likelihood version VSD(2) is calculated. For both methods, a model-based and an extrapolated version is implemented. The VSD(2) is not calculated for the conditional model, since a confidence interval for dose effect β_d does not necessarily transform monotonically into an interval for dose d. In addition to the three models used in the previous section, the model proposed by George and Bowman (1995), which was introduced in Section 2.4, is included.

Table 5.1 shows model-based VSD's, whereas Table 5.2 gives the extrapolated versions. For the quantities "ED - MB", "ED - EP", "VSD(1) - MB" and "VSD(1) - EP", the conditional model results in the largest value in 8, 11, 9 and 10 cases (out of 12) respectively. As a consequence, the conditional model in general yields the highest values for both ED and VSD(1), in both the extrapolation and the model-based methods, but the effect is somewhat clearer in the extrapolation procedure. To some extent, this result differs for the three chemicals under consideration. In the DEHP, DYME and EG experiments, the conditional model results in the largest values of ED and VSD(1) in 16, 8 and 14 cases (out of 16) respectively. As a consequence, this picture is slightly amended in the DYME study where the largest

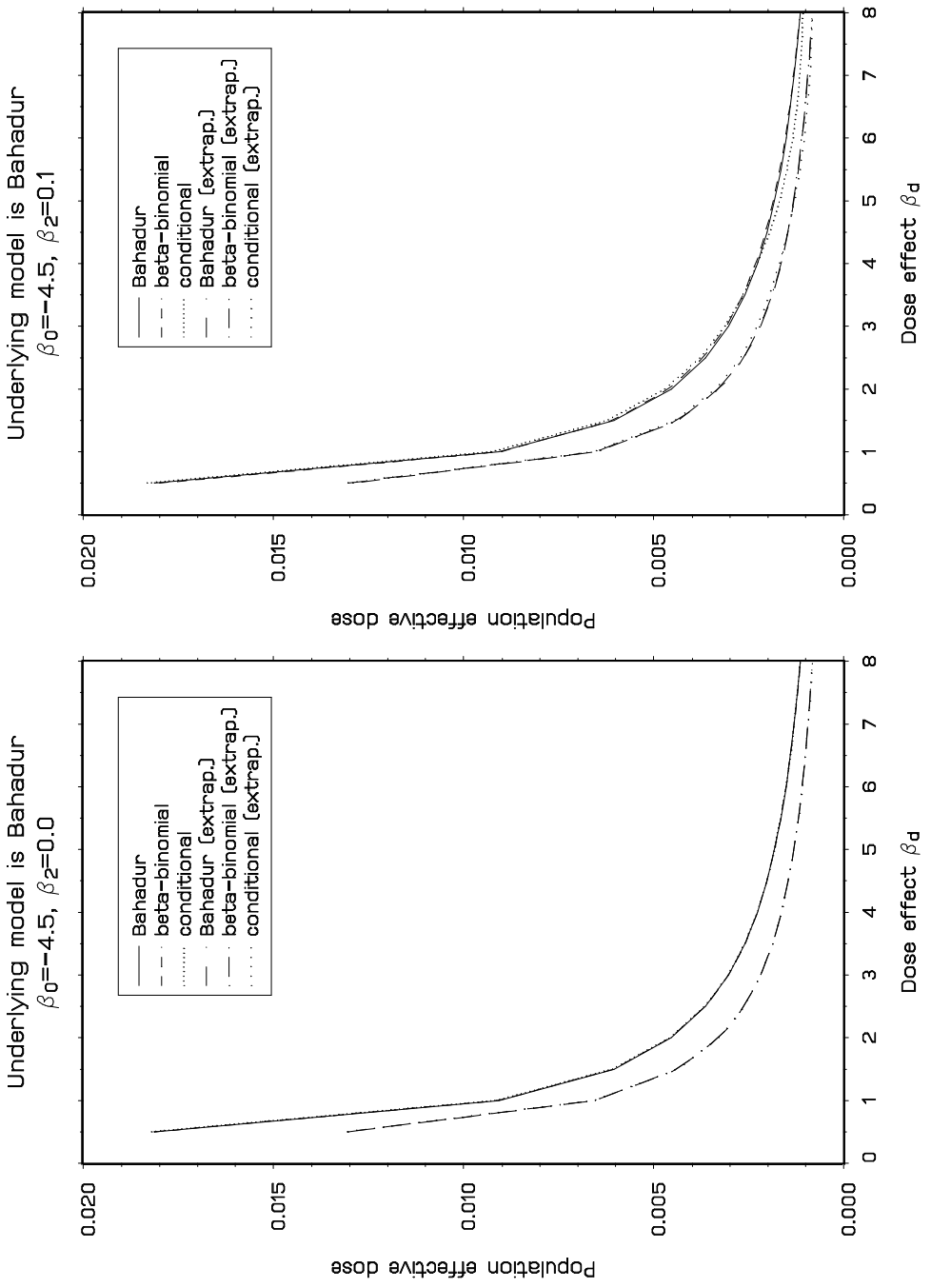

Figure 5.1: *Population values of the effective dose when the underlying model is Bahadur: model-based and extrapolated estimators.*

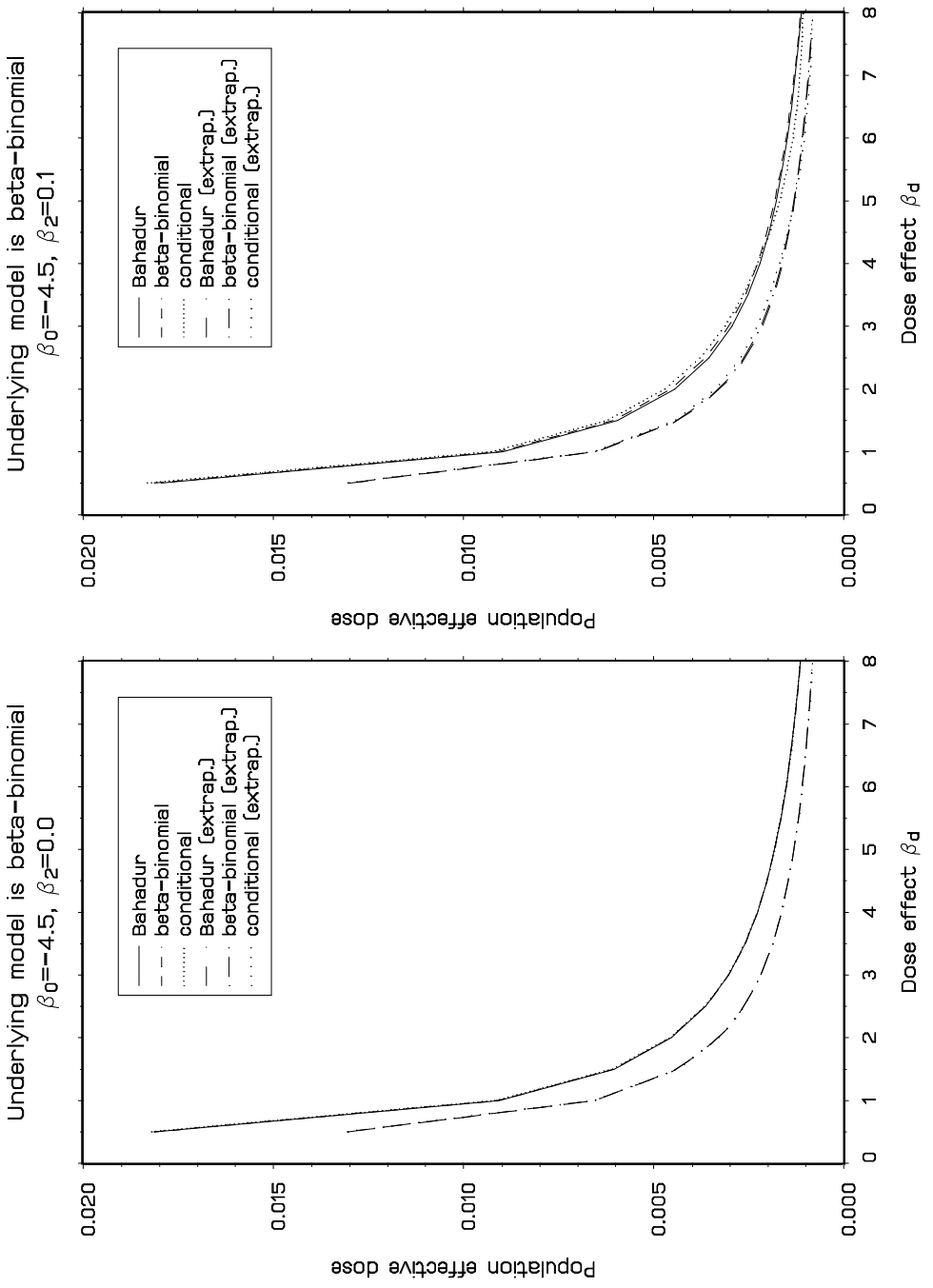

Figure 5.2: *Population values of the effective dose when the underlying model is beta-binomial: model-based and extrapolated estimators.*

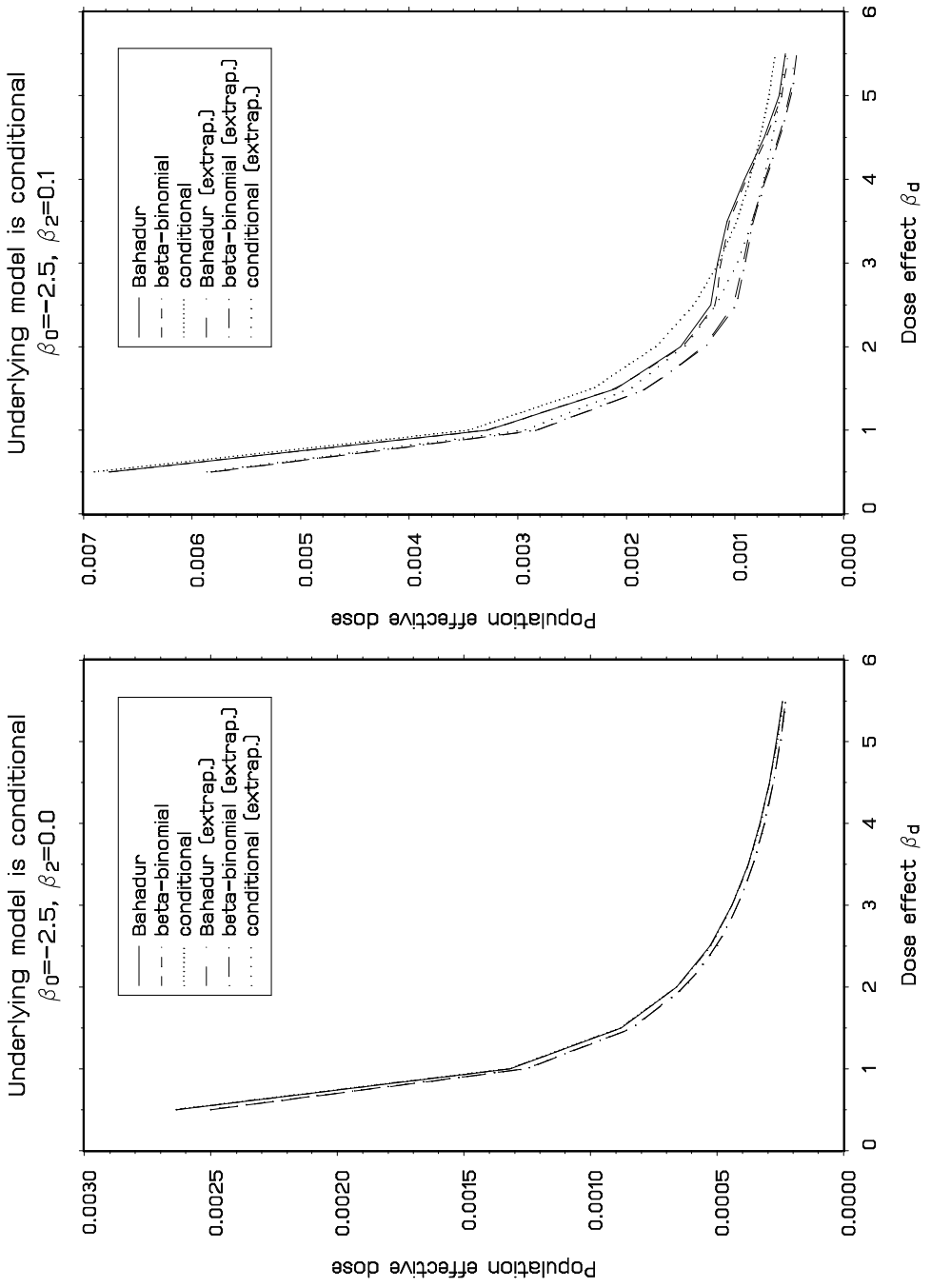

Figure 5.3: *Population values of the effective dose when the underlying model is conditional: model-based and extrapolated estimators.*

results are most often obtained under the conditional model (in 8 out of 16 cases as indicated above) or under the George-Bowman model (in 6 out of 16 cases).

Turning attention to the marginal models, the highest values of ED, VSD(1) and VSD(2) are found with the George-Bowman model. The number of cases in which the Bahadur model results in higher values relative to the beta-binomial model, is comparable to the number of cases with lower values for the Bahadur model.

One also notices that the extrapolation method yields much smaller values in all cases. This is in line with the asymptotic study, which has shown that effective doses computed by means of the extrapolation procedure results in lower values compared to the model-based estimates.

Now, the focus is on the various procedures to calculate VSD's. First, one observes that VSD(2) is virtually always higher than VSD(1). This is to be expected, since it is based on a one degree of freedom procedure, whereas for VSD(1), three degrees of freedom are spent in case of the Bahadur and beta-binomial models and two degrees of freedom are used in the George-Bowman model. Of course, as pointed out in the previous section, a lower VSD is "safer", but one should be careful not to be overly cautious (Morgan, 1992). Secondly, the linear extrapolated versions of Table 5.2 are smaller than their purely model based counterparts of Table 5.1. These two observations yield the following ordering:

$$\text{VSD(1,EP)} \leq \text{VSD(2,EP)} \leq \text{VSD(1,MB)} \leq \text{VSD(2,MB)}.$$

This ordering is found in 30 out of 36 cases. In addition, the discrepancies between the VSD's from different models and between VSD(1) and VSD(2), are smaller with the linear extrapolation method.

Visceral malformation in the EG study is an interesting exception as it is the only one for which no clustering is found. The conclusion here agrees with the left hand panel of Figures 5.1 and 5.2. Indeed, when there is no association, there are essentially only two curves: model based and extrapolated.

5.4 Concluding remarks

In this chapter, a few models describing dose-response curves based on a binary outcome in clustered experiments, have been studied . From a modelling perspective, both marginal and conditional models can be defended. Model properties w.r.t. estimation of a safe dose were investigated. A likelihood confidence region based

Table 5.1: Estimates of effective doses and lower confidence limits. Entirely model based computation. All quantities shown should be divided by 10^4.

Outcome	Model	Statistic	DEHP	EG	DYME
External	Bahadur	ED	27	72	165
		VSD(1)	15	47	48
		VSD(2)	18	55	63
	GB	ED	28	98	156
		VSD(1)	16	50	60
		VSD(2)	20	68	71
	BB	ED	26	73	168
		VSD(1)	14	45	47
		VSD(2)	17	56	62
	Cond	ED	36	124	141
		VSD(1)	22	66	55
Skeletal	Bahadur	ED	23	4	13
		VSD(1)	14	4	9
		VSD(2)	16	4	7
	GB	ED	29	7	27
		VSD(1)	17	6	16
		VSD(2)	21	6	18
	BB	ED	27	6	25
		VSD(1)	14	4	11
		VSD(2)	18	5	14
	Cond	ED	34	11	25
		VSD(1)	20	9	17
Visceral	Bahadur	ED	19	350	171
		VSD(1)	13	189	48
		VSD(2)	15	126	72
	GB	ED	20	385	203
		VSD(1)	13	83	100
		VSD(2)	16	135	87
	BB	ED	18	367	98
		VSD(1)	11	199	63
		VSD(2)	14	131	40
	Cond	ED	28	504	202
		VSD(1)	18	134	95
Collapsed	Bahadur	ED	9	4	25
		VSD(1)	6	4	13
		VSD(2)	7	4	15
	GB	ED	9	7	18
		VSD(1)	6	5	12
		VSD(2)	7	6	13
	BB	ED	8	5	27
		VSD(1)	6	4	13
		VSD(2)	6	4	15
	Cond	ED	14	11	27
		VSD(1)	9	8	18

Table 5.2: Estimates of effective doses and lower confidence limits. Linear extrapolation method. All quantities shown should be divided by 10^4.

Outcome	Model	Statistic	DEHP	EG	DYME
External	Bahadur	ED	17	41	34
		VSD(1)	12	38	23
		VSD(2)	14	39	25
	GB	ED	18	46	36
		VSD(1)	13	37	26
		VSD(2)	15	44	28
	BB	ED	17	41	34
		VSD(1)	12	37	22
		VSD(2)	13	39	25
	Cond	ED	24	57	36
		VSD(1)	17	42	25
Skeletal	Bahadur	ED	16	4	10
		VSD(1)	11	4	8
		VSD(2)	13	4	6
	GB	ED	19	7	17
		VSD(1)	13	5	12
		VSD(2)	16	6	13
	BB	ED	17	5	14
		VSD(1)	12	4	9
		VSD(2)	14	5	10
	Cond	ED	23	10	18
		VSD(1)	16	8	14
Visceral	Bahadur	ED	14	67	44
		VSD(1)	11	62	28
		VSD(2)	12	62	34
	GB	ED	15	67	46
		VSD(1)	11	46	38
		VSD(2)	13	62	37
	BB	ED	14	67	36
		VSD(1)	10	62	31
		VSD(2)	11	62	26
	Cond	ED	21	78	64
		VSD(1)	15	55	47
Collapsed	Bahadur	ED	7	4	14
		VSD(1)	5	4	9
		VSD(2)	6	4	10
	GB	ED	8	6	13
		VSD(1)	5	5	9
		VSD(2)	6	5	10
	BB	ED	7	5	14
		VSD(1)	5	4	9
		VSD(2)	6	4	11
	Cond	ED	11	10	17
		VSD(1)	8	8	13

and a profile likelihood based procedure were implemented. In both cases, a purely model based and a partly linearly extrapolated version were considered.

Based on this study, extrapolation to compute a VSD is recommendable over an entirely model based determination, since in the former procedure, there is a closer agreement between the results calculated from different models. In the settings considered here, the conditional model often, but not always, yields the highest VSD.

Estimating a VSD was built on the strategy developed for a marginal model: the probability of a malformation occurring in an individual animal is used as a tool to define excess risk from which a VSD is derived. However, the probability of malformation occurring in the litter as a whole could be used instead, with a human being represented by a litter, rather than by an individual fetus. Whether this procedure is advantageous, is discussed in the following chapter.

Other models than the ones presented here deserve consideration, e.g., the odds ratio model (Molenberghs and Lesaffre, 1994). Other methods of estimation, e.g., estimating equations (Liang and Zeger, 1986) can be investigated in the context of misspecifying the model and the implications for safe dose assessment. Further, procedures that incorporate the effect of dose on both death and malformation are worthwhile to consider. The latter extension enables the study of the relation between dosing and observed litter size and is incorporated into Chapter 6. However, in the simplified setting considered here, some of the advantages, problems and drawbacks have been identified.

Chapter 6

Litter-based methods in safe dose determination

In the previous chapter, quantitative risk assessment was performed using dose-response models. Risks were based on the probability that a fetus exhibits the adverse event under investigation. This approach is straightforward for marginal models, which are expressed in terms of this marginal adverse event probability (Diggle, Liang and Zeger, 1994; Pendergast *et al.*, 1996). However, one might base risk assessment also on the cluster of fetuses of a dam. Then, the probability that at least one fetus of a dam has the adverse event under consideration, is crucial. In this chapter, fetus and so-called litter-based risks are contrasted in the determination of safe doses basing on dose-response modelling. In analogy with the previous chapter, the stochastic behaviour of the number of implants and the number of viables is taken into account when calculating risks, via integration over the cluster size distribution (Rai and Van Ryzin, 1985).

Here, the emphasis is on the beta-binomial model (Skellam, 1948; Kleinman, 1973), as well as on the conditional model of Molenberghs and Ryan (1999). Both models were introduced in Chapter 2. The Bahadur model (1961) is not considered here because of the previously observed drawbacks of this model, which are primarily due to restrictions on its parameters. The parameters of the selected models in this chapter are modelled by means of linear predictors. It is shown how the beta-binomial and the conditional models can easily handle litter-based rates. Furthermore, it is demonstrated how the conditional model leads to a natural formulation of the fetus-based excess risk on the number of implants in a dam, unlike marginal models such as the beta-binomial model.

In this chapter, four approaches are focused: (1) an indicator for death, (2) a col-

lapsed malformation indicator ignoring dead fetuses, (3) an indicator for abnormality (i.e., death or malformation) and (4) a joint model for death and malformation. A multivariate approach for malformation and the incorporation of weight into the model, are subjects of future research.

In Section 6.1, fetus and litter-based risks are derived for the beta-binomial and conditional models. The collapsed outcome "abnormality" is discussed, as well as the hierarchically structured outcomes death and malformation. Section 6.2 compares the fetus and litter-based approach and contrasts a model for abnormality with a joint model for death and malformation, based on the NTP data. Section 6.3 illustrates these items based on asymptotic samples. In Section 6.4, the focus is on the variability of the excess risk estimator.

6.1 Expressing risks

As indicated in Section 5.1, the specification of an adverse event is required in quantitative risk assessment, together with its risk as a function of administered dose. A safe dose level can then be derived based on this relation. In the literature, several definitions are used to express the concept of a safe dose. For instance, a virtually safe dose can be defined as the lower confidence limit on a dose corresponding to a very small excess risk, e.g. 10^{-4}. Here, the dose itself is referred to as the effective dose (ED). In this section, different risks and corresponding excess risks are presented for the beta-binomial and conditional models. They can be fetus or litter based and they can be defined for a single adverse event like "death" or "malformation" as well as for both events jointly. In the next sections, these different approaches to risk and ED estimation are compared for the NTP data and for so-called asymptotic samples.

6.1.1 Fetus and litter-based risks

The main issue deals with the choice between fetus and litter-based risks. Here, for simplicity, the presentation is restricted to the adverse event "abnormality" in a litter with m implants. A fetus-based approach focuses on the risk of a fetus as a function of the level of exposure d given to the dam. Let $q_F(m;d)$ be the probability that a fetus is abnormal, given that the fetus is selected from a litter with m implants. Consider all values of the number of implants m with non-zero probability $P(m)$. Administering some specified dose d to M dams, the fetus-based

risk is:

$$r_F(d) = \frac{\sum_m MP(m)mq_F(m;d)}{\sum_m MP(m)m} = \frac{\sum_m P(m)mq_F(m;d)}{\sum_m P(m)m}. \tag{6.1}$$

Hence, the fetus-based risk at some specified dose is an average of conditional probabilities $q_F(m;d)$ with weights $MP(m)m$, i.e., the expected number of fetuses in litters with m implants resulting from the M dams.

In marginal models such as the beta-binomial model, the probability $q_F(m;d)$ does not depend on the number of implants m (except when it is explicitly incorporated in the model as a covariate) and hence, $r_F(d) = q_F(d)$. It will be shown that this is in contrast with the conditional model of Molenberghs and Ryan (1999), where q_F is related to the number of implants m in a natural way.

In a litter-based approach, the event of interest is whether at least one fetus in a litter is abnormal. Let $q_L(m;d)$ be the probability that at least one fetus in a litter of size m has the adverse event. The litter-based risk is

$$r_L(d) = \sum_m P(m)q_L(m;d), \tag{6.2}$$

which is an average of conditional probabilities $q_L(m;d)$ with weights $P(m)$.

Since a particular adverse effect in one or more fetuses of a litter is at least as probable as the occurrence of this adverse event in a specific fetus, it follows that $q_F(m;d) \leq q_L(m;d)$. Considering this inequality for any number of implants m with non-zero probability $P(m)$, it follows that

$$\sum_m P(m)q_F(m;d) \leq r_L(d) = \sum_m P(m)q_L(m;d).$$

For a single adverse event in a marginal model, the conditional probability $q_F(m;d) = q_F(d)$ and hence, the first sum equals $r_F(d)$. In this case, the fetus-based risk is smaller than or equal to the litter-based risk. Notice however that in general, the first sum is different from $r_F(d)$. One can easily find examples in which $r_F(d)$ is smaller than, equal to or greater than $r_L(d)$. Indeed, consider the case of two litters, litter 1 with one fetus being abnormal and litter 2 with two fetuses being healthy. Then, $r_F(d)$ for the adverse event "abnormality" is 1/3, while $r_L(d) = 1/2$. If litter 1 would have had two abnormal fetuses, then $r_F(d) = r_L(d) = 1/2$. Finally, if litter 1 consisted of three abnormal fetuses, then $r_F(d) = 3/5 > r_L(d) = 1/2$.

The excess risk $r^*(d)$ has been introduced in Section 5.1. Again, there are cases where the fetus-based excess risk $r_F^*(d)$ is smaller than, equal to or greater than the litter-based excess risk $r_L^*(d)$. This is illustrated in Table 6.1.

Table 6.1: Cases in which the fetus-based excess risk is smaller than, equal to or larger than the litter-based excess risk. A and H indicate an abnormal and a healthy fetus respectively.

	Case 1	Case 2	Case 3
Litter 1, dose=0	A,A	A,A	A,H
Litter 2, dose=0	H,H	H,H	H,H
Litter 3, dose=d>0	A,A	A,A	A,A
Litter 4, dose=d>0	A,H	A,A	H,H
$r_F(0)$	1/2	1/2	1/4
$r_L(0)$	1/2	1/2	1/2
$r_F(d)$	3/4	1	1/2
$r_L(d)$	1	1	1/2
$r_F^*(d)$	1/2	1	1/3
$r_L^*(d)$	1	1	0

For a specific model, these risks can be estimated by replacing all parameters by their maximum likelihood estimates. Also the values $P(m)$, i.e., the distribution of the number of implants in a litter, have to be estimated. This is discussed in more detail in Section 6.4.

In what follows, fetus and litter-based risks will be discussed for the beta-binomial model and the conditional model of Molenberghs and Ryan. For both models, the approach of a single adverse event (focusing on "abnormality") will be given, as well as the approach where the adverse events "death" and "malformation" are studied jointly.

6.1.2 Risks for a beta-binomial model for abnormality

The probability q_F that a fetus is abnormal, given that the fetus is selected from a litter with m implants, is π. Based on (2.4) and (2.5),

$$q_F = \frac{1}{1 + \exp(-\beta_0 - \beta_d d)}.$$

As mentioned before, the probability q_F does not depend on the number of implants m and hence, the fetus-based excess risk equals

$$r_F^* = \frac{q_F(d) - q_F(0)}{1 - q_F(0)} = \frac{1 - \exp(-\beta_d d)}{1 + \exp(-\beta_0 - \beta_d d)}.$$

Since r_F^* does not depend on the correlation parameter ρ, the above expression is also valid for the ordinary logistic regression model.

The probability that at least one fetus of a litter is abnormal is

$$q_L = 1 - \frac{B(\pi(\rho^{-1} - 1), (1 - \pi)(\rho^{-1} - 1) + m)}{B(\pi(\rho^{-1} - 1), (1 - \pi)(\rho^{-1} - 1))}.$$

This expression can be rewritten as

$$q_L = 1 - \prod_{k=0}^{m-1} \left(1 - \pi + \frac{k\pi\rho}{1 + (k-1)\rho}\right).$$

Notice that, in cases of overdispersion, the litter-based probability of an adverse event q_L is smaller than the probability $1 - (1 - \pi)^m$, corresponding to $\rho = 0$ (no clustering). From (5.7) and (6.2), the litter-based excess risk can be computed as

$$r_L^* = 1 - \frac{\sum_m P(m) \prod_{k=0}^{m-1} (1 - \pi(d) + k\pi(d)\rho/(1 + (k-1)\rho))}{\sum_m P(m) \prod_{k=0}^{m-1} (1 - \pi(0) + k\pi(0)\rho/(1 + (k-1)\rho))}.$$

In case of no clustering, this expression reduces to

$$r_L^* = 1 - \frac{G(1 - \pi(d))}{G(1 - \pi(0))}$$

where $G(\cdot)$ is the probability generating function of the number of implants. For $m = 1$, $G(z) = z$ such that $r_L^* = r_F^*$.

6.1.3 Risks for a beta-binomial model for death and malformation jointly

Here, it is proposed to model both components of (2.1) with a model similar to (2.9):

$$f(r \mid m, d) = \binom{m}{r} \frac{B(\pi_{dth}(\rho_{dth}^{-1} - 1) + r, (1 - \pi_{dth})(\rho_{dth}^{-1} - 1) + m - r)}{B(\pi_{dth}(\rho_{dth}^{-1} - 1), (1 - \pi_{dth})(\rho_{dth}^{-1} - 1))}, \tag{6.3}$$

$$f(z \mid n, d) = \binom{n}{z} \frac{B(\pi_{mal}(\rho_{mal}^{-1} - 1) + z, (1 - \pi_{mal})(\rho_{mal}^{-1} - 1) + n - z)}{B(\pi_{mal}(\rho_{mal}^{-1} - 1), (1 - \pi_{mal})(\rho_{mal}^{-1} - 1))}. \tag{6.4}$$

Again, one can distinguish between risk assessment at fetus level and at litter level.

The probability that fetus j is dead or malformed, given that the number of implants equals m, is

$$\begin{aligned} q_F &= P(\text{fetus } j \text{ is dead} \mid m \text{ implants}) + P(\text{fetus } j \text{ is malformed} \mid m \text{ implants}) \\ &= \pi_{dth} + \sum_{r=0}^{m-1} (P(\text{fetus } j \text{ is alive and } R = r \mid m \text{ implants}) \\ &\quad \times P(\text{fetus } j \text{ is malformed} \mid \text{fetus } j \text{ alive \& } r \text{ deaths out of } m \text{ implants})) \end{aligned} \tag{6.5}$$

where R denotes the number of deaths in a litter. This can be reexpressed as

$$\begin{aligned} q_F &= \pi_{dth} + \frac{\pi_{mal}}{B(\pi_{dth}(\rho_{dth}^{-1} - 1), (1 - \pi_{dth})(\rho_{dth}^{-1} - 1))} \\ &\quad \times \sum_{r=0}^{m-1} \binom{m-1}{r} B(\pi_{dth}(\rho_{dth}^{-1} - 1) + r, (1 - \pi_{dth})(\rho_{dth}^{-1} - 1) + m - r). \end{aligned} \tag{6.6}$$

Expressions (5.7), (6.1) and (6.6) enable the calculation of the fetus-based excess risk.

The probability that at least one fetus is dead or malformed, given m, is based upon (2.1) and reduces to

$$q_L = 1 - P(R = 0, Z = 0 \mid m, d) = 1 - P(R = 0 \mid m, d)P(Z = 0 \mid n, d).$$

Explicitly, in terms of (6.3) and (6.4),

$$q_L = 1 - \prod_{k,\ell=0}^{m-1} \left(1 - \pi_{dth} + \frac{k\pi_{dth}\rho_{dth}}{1 + (k-1)\rho_{dth}}\right) \left(1 - \pi_{mal} + \frac{\ell\pi_{mal}\rho_{mal}}{1 + (\ell - 1)\rho_{mal}}\right). \tag{6.7}$$

Formulas (5.7), (6.2) and (6.7) allows one to compute the litter-based excess risk.

6.1.4 Risks for a conditional model for abnormality

For the conditional model, the probability q_F that fetus j is abnormal, given implant size m, can be expressed in several ways. The probability q_F can be written as in (2.17) in terms of S, the number of abnormals. It can also be computed based on (2.14), by summing over the distribution of the outcomes of the other littermates:

$$\begin{aligned} q_F &= \sum_{s=1}^{m} \binom{m-1}{s-1} \exp\{\psi s - \phi s(m-s) - A(\psi, \phi, m)\} \\ &= \exp\{-A(\psi, \phi, m) + A(\psi + \phi, \phi, m-1) + \psi - \phi(m-1)\} \end{aligned}$$

with s the number of abnormal fetuses in a litter. Finally, one can derive an expression for q_F by calculating the probability that fetus j is healthy, given m implants:

$$\begin{aligned} q_F &= 1 - \sum_{s=0}^{m-1} \binom{m-1}{s} \exp\{\psi s - \phi s(m-s) - A(\psi, \phi, m)\} \\ &= 1 - \exp\{-A(\psi, \phi, m) + A(\psi - \phi, \phi, m-1)\}. \end{aligned} \tag{6.8}$$

Using the expression for a healthy fetus is slightly more convenient. These formulas show how, for the conditional model, the probabilities q_F depend on the number of implants. Based on (5.7), (6.1) and (6.8), the fetus-based excess risk follows.

Now, the probability that at least one fetus is abnormal, given m, is $q_L = 1 - P(S = 0)$, which, based on (2.14), is given by

$$q_L = 1 - \exp\{-A(\psi, \phi, m)\}. \tag{6.9}$$

This result is an appealing counterpart to (6.8). It differs from (6.8) by the deletion of one normalizing constant. Expression (5.7), (6.2) and (6.9) can be used to calculate the litter-based excess risk.

6.1.5 Risks for a conditional model for death and malformation jointly

The conditional model for death and malformation jointly is the product of

$$f(r \mid m, d) = \binom{m}{r} \exp\{\psi_{dth} r - \phi_{dth} r(m-r) - A(\psi_{dth}, \phi_{dth}, m)\}, \tag{6.10}$$

$$f(z \mid n, d) = \binom{n}{z} \exp\{\psi_{mal} z - \phi_{mal} z(n-z) - A(\psi_{mal}, \phi_{mal}, n)\}. \tag{6.11}$$

Using (6.5), the conditional probability that fetus j exhibits an adverse event, given that a litter contains m implants, can be rewritten as:

$$\begin{aligned} q_F &= \pi_{dth} + \sum_{r=0}^{m-1} \binom{m-1}{r} \exp\{\psi_{dth} r - \phi_{dth} r(m-r) - A(\psi_{dth}, \phi_{dth}, m)\} \\ &\quad \times \{1 - \exp\{-A(\psi_{mal}, \phi_{mal}, m-r) + A(\psi_{mal} - \phi_{mal}, \phi_{mal}, m-r-1)\}\}. \end{aligned}$$

Based on (6.8), this expression can be simplified to

$$q_F = 1 - \exp\{-A(\psi_{dth}, \phi_{dth}, m)\} \sum_{r=0}^{m-1} \binom{m-1}{r} \exp(B(r)) \tag{6.12}$$

with

$$B(r) = \psi_{dth} r - \phi_{dth} r(m-r) - A(\psi_{mal}, \phi_{mal}, m-r) + A(\psi_{mal} - \phi_{mal}, \phi_{mal}, m-r-1).$$

Again, by means of (5.7), (6.1) and (6.12), the fetus-based excess risk can be computed.

Considering the conditional model for death and malformation jointly, the probability that a litter has at least one adverse event, given m, can be based on (2.1), (6.10) and (6.11):

$$q_L = 1 - \exp\left\{-A(\psi_{dth}, \phi_{dth}, m) - A(\psi_{mal}, \phi_{mal}, m)\right\}. \tag{6.13}$$

The rather complicated sum in (6.12) is replaced by a normalizing constant. Expressions (5.7), (6.2) and (6.13) enable the calculation of the litter-based excess risk.

6.2 Analysis of NTP data

The different risk and corresponding ED estimators of Section 6.1 are compared for the NTP data, introduced in Section 1.4. Excess risk functions are estimated by maximum likelihood for a grid of dose values, based on the beta-binomial model (BB) and the conditional model (Cond). These results are also compared to those of the logistic model for which all information of a litter is collapsed into a single, binary variable indicating whether there is at least one abnormal fetus. For the toxic agents DEHP, DYME, EG, TGDM and THEO, the resulting curves for the adverse events "abnormality" and "death and malformation jointly", are shown in Figures 6.1 – 6.5. In these figures, two groups of curves can be distinguished: fetus-based versus litter-based excess risk curves. As expected intuitively, litter-based excess risks are clearly larger than fetus-based risks at the same dose level. Within the set of litter-based curves, the ordering

$$\text{Cond, joint} < \text{Cond, abnormal} < \text{BB, joint} < \text{BB, abnormal}$$

is frequently observed. This is also true for the fetus-based risks except for large dose values. For small dose levels, the risk for the logistic regression model is most often somewhat higher than the other litter-based risks.

Besides the adverse effects considered in Figures 6.1 – 6.5, also "death" and "malformation among the viable fetuses" are investigated for the five toxic agents

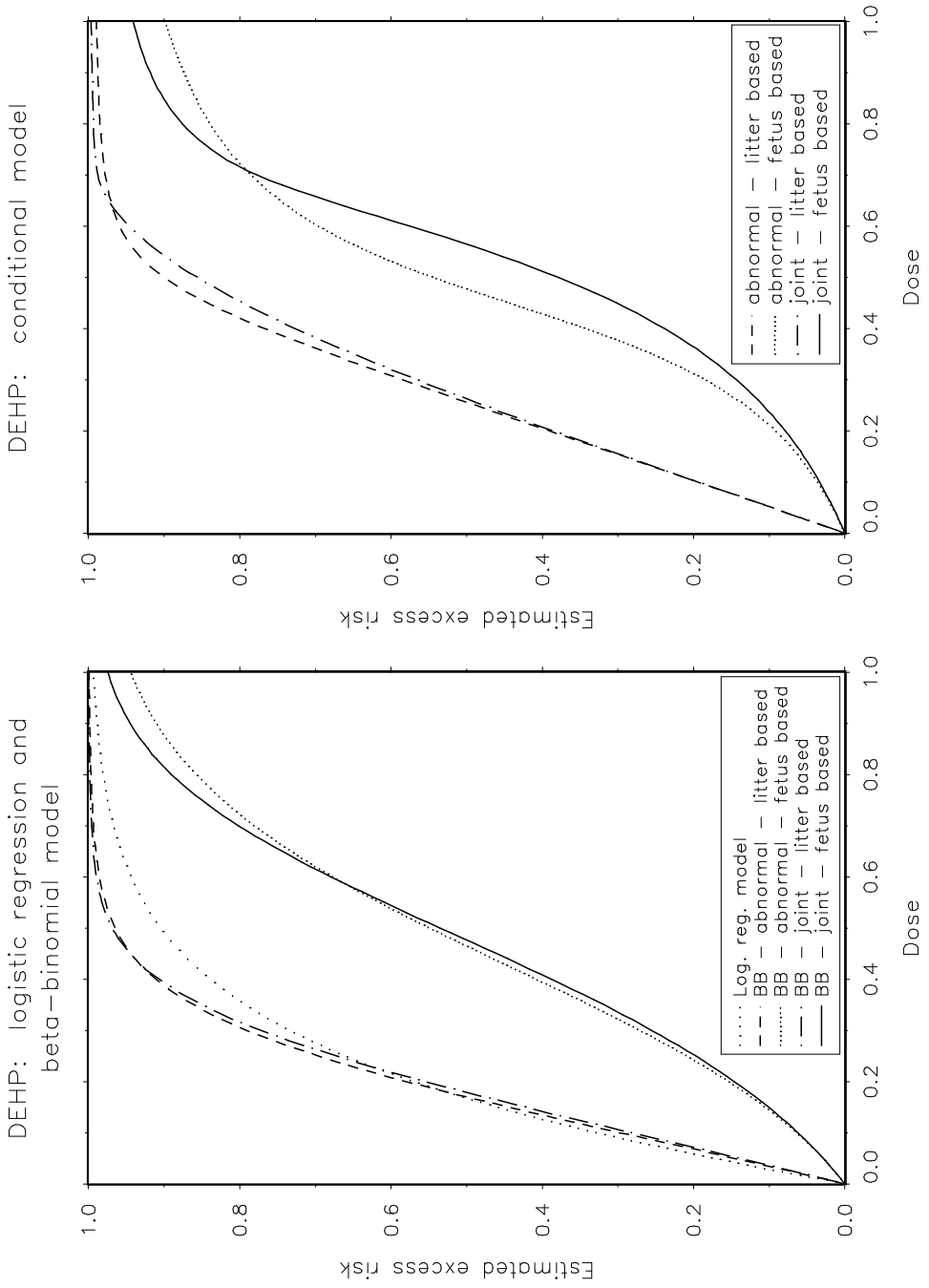

Figure 6.1: *Excess risk curves for DEHP based on the logistic regression, beta-binomial and conditional models.*

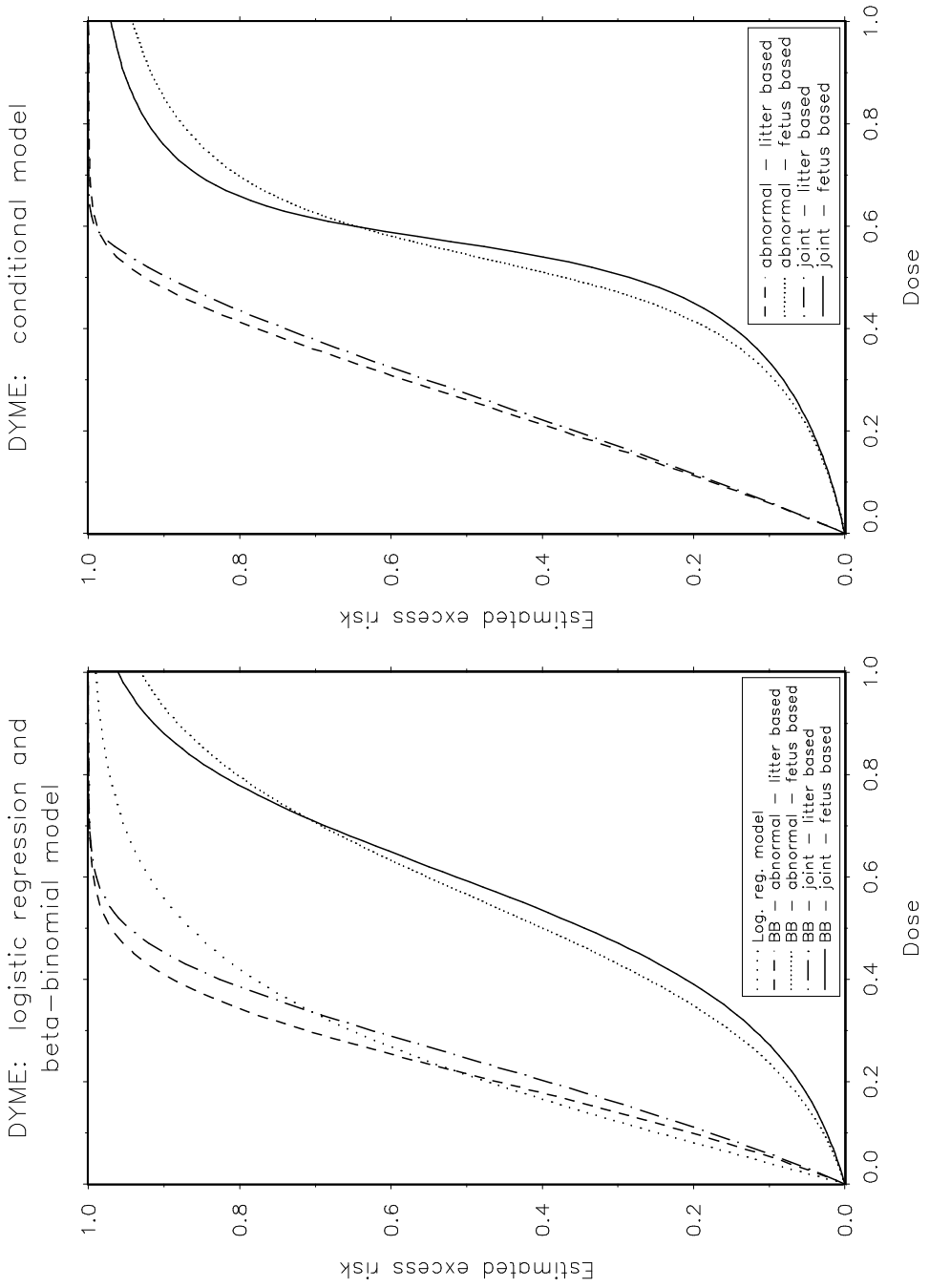

Figure 6.2: *Excess risk curves for DYME based on the logistic regression, beta-binomial and conditional models.*

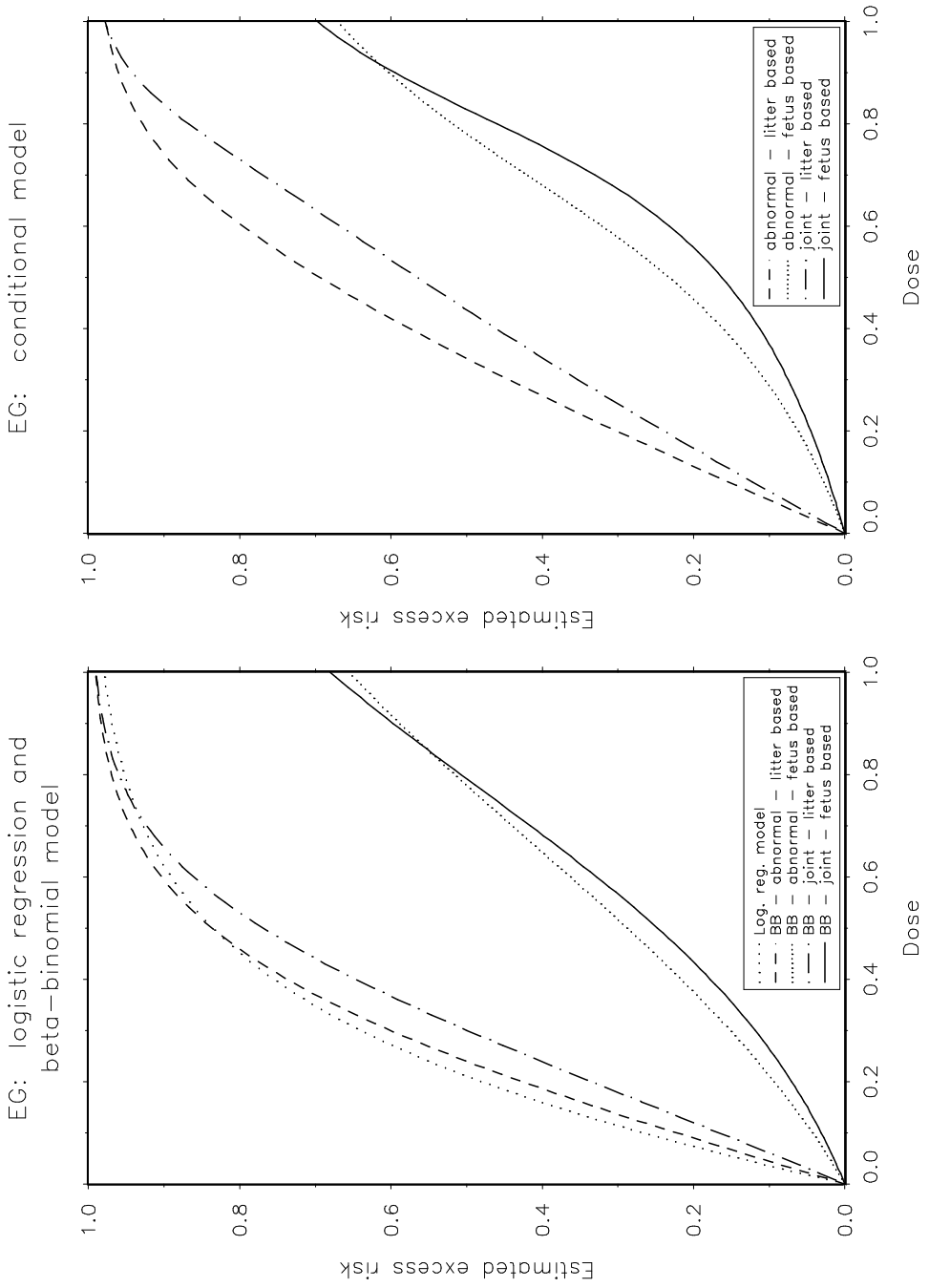

Figure 6.3: *Excess risk curves for EG based on the logistic regression, beta-binomial and conditional models.*

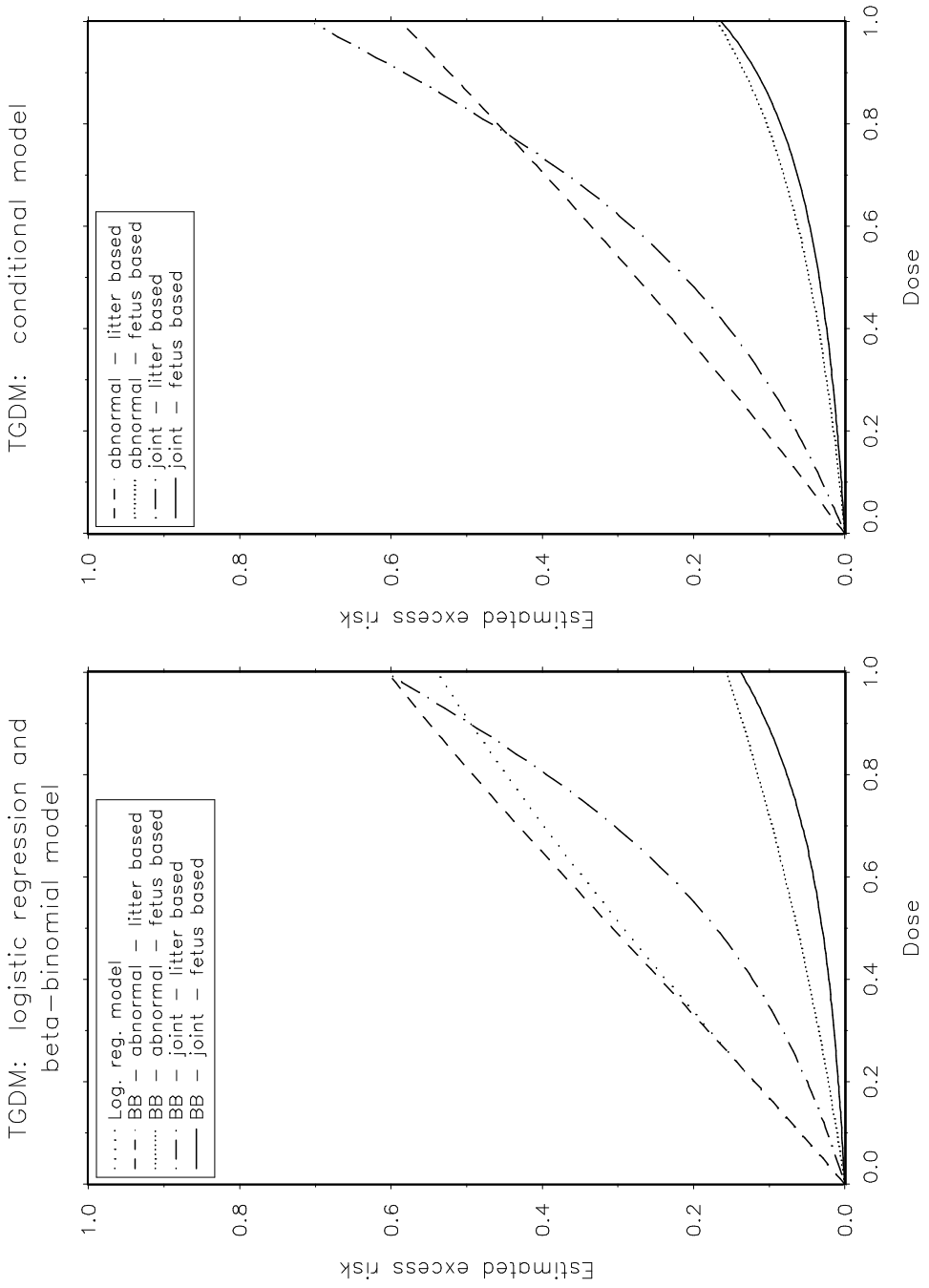

Figure 6.4: *Excess risk curves for TGDM based on the logistic regression, beta-binomial and conditional models.*

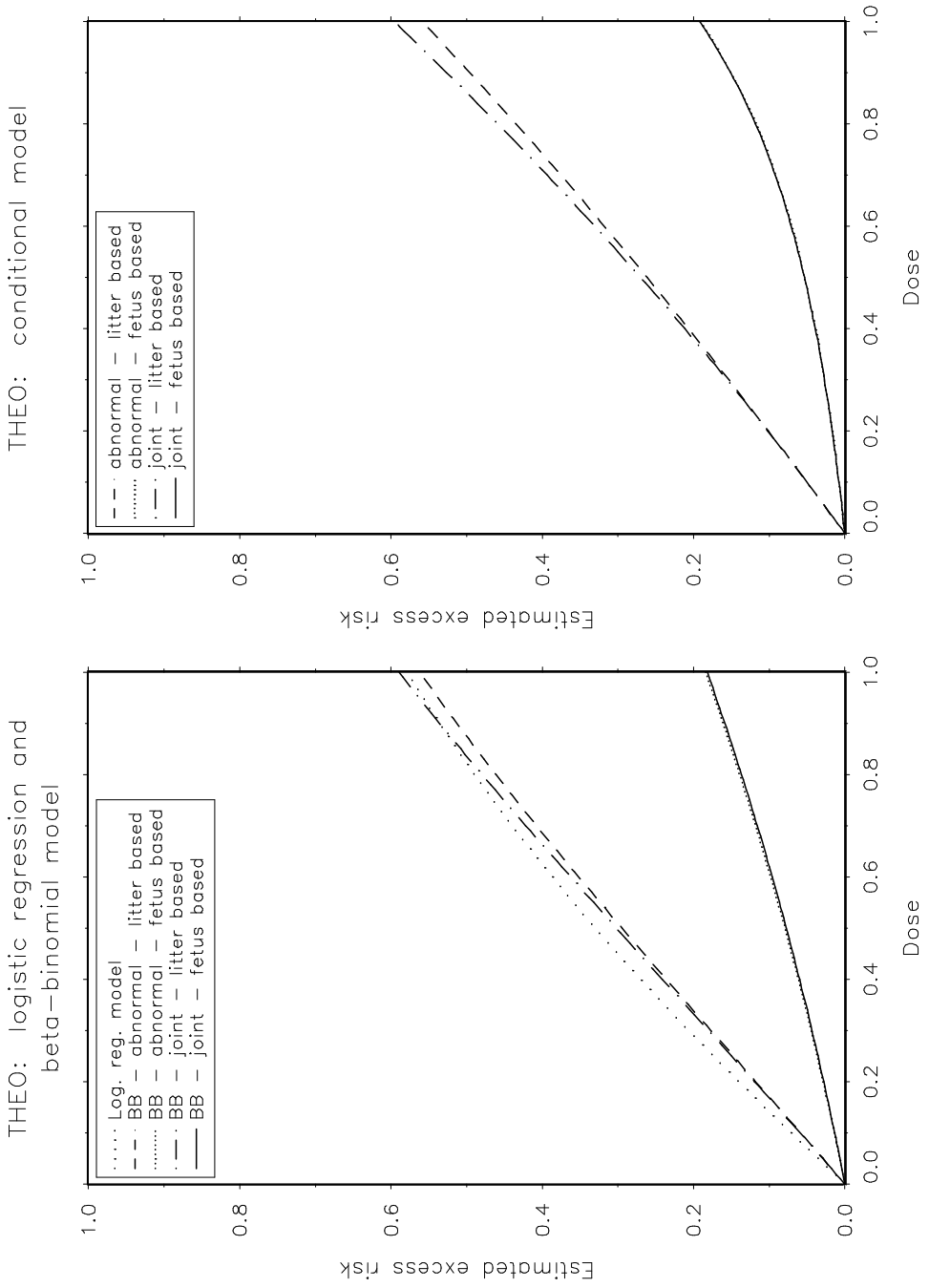

Figure 6.5: *Excess risk curves for THEO based on the logistic regression, beta-binomial and conditional models.*

under consideration. Effective doses are calculated for several adverse events. These are shown in Table 6.2. The ED of a fetus-based risk curve is in general about 5 to 10 times larger than the corresponding litter-based ED. This is in line with the excess risk curves of Figures 6.1 – 6.5. The effective doses of "abnormal" and of "joint" are well comparable, except for the chemical TGDM. Comparing the three models under investigation, the ED of the conditional model is most often larger than the ED of the beta-binomial model. The logistic model results in the smallest ED. Since the models considered come from fundamentally different modelling families (conditional and marginal), a somewhat different behaviour in key aspects is not unexpected. Indeed, in Chapter 5 which addressed ED determination in the fetus-based setting, it was concluded that EDs tend to be somewhat higher in the conditional model, as opposed to the beta-binomial, the Bahadur and the George-Bowman models.

The next section examines whether these findings are confirmed by a large sample simulation study.

6.3 Asymptotic study

In order to compare the asymptotic effect of a fetus-based versus a litter-based approach on the effective dose, the ideas of Rotnitzky and Wypij (1994) are followed here. An artificial (asymptotic or "large") sample is constructed where each possible realization of dose d, number of implants m, number of deaths r and number of malformations z is weighted according to the probability in the underlying model. Precisely, all realizations of the form (d, m, r, z) are included and are assigned the weight $f(d, m, r, z)$ where f denotes a probability mass function. Hence, one has to specify: (1) $f(d)$, the relative frequency of the dose group as prescribed by the design; (2) $f(m|d)$, which equals $f(m)$ since a dam is randomly assigned to a dose group and exposure occurs after mating; (3) $f(r|m, d)$, the actual model probability for the occurrence of r deaths and (4) $f(z|r, m, d) = f(z|n, m, d)$, which is assumed here to be $f(z|n, d)$, the actual model probability for z malformations. Again, the doses 0, 0.25, 0.5 and 1 are considered when generating asymptotic samples and each dose is assigned a relative frequency of 1/4. The distribution of the number of implants, $f(m)$, is based on the NTP data. The relative frequencies of m for all NTP datasets under investigation are smoothed via a local linear smoothing technique. Least squares cross-validation has been used to choose the bandwidth.

Table 6.2: Effective doses of DEHP, DYME, EG, TGDM and THEO corresponding to an excess risk of 10^{-4}. All quantities shown should be divided by 10^4.

Model	Unit	Adverse event	DEHP	DYME	EG	TGDM	THEO
Logistic regression	Litter	Abnormal	0.3	0.4	0.3	1.7	1.4
Beta-binomial	Fetus	Dead	2.5	7.8	12.3	40.4	8.1
		Malformed	7.7	15.6	5.1	81.3	150.6
		Abnormal	1.9	4.6	2.5	10.0	7.5
		Joint	1.9	5.2	3.6	27.1	7.7
	Litter	Dead	0.5	0.9	1.5	7.7	1.9
		Malformed	1.4	2.6	1.1	12.1	17.6
		Abnormal	0.3	0.6	0.4	1.7	1.7
		Joint	0.4	0.6	0.6	4.7	1.7
Conditional	Fetus	Dead	5.2	10.7	15.9	31.7	14.9
		Malformed	9.8	15.4	8.2	93.6	182.4
		Abnormal	3.4	6.6	4.0	15.2	13.7
		Joint	3.5	6.7	5.3	23.7	13.8
	Litter	Dead	0.8	1.0	1.8	5.9	2.3
		Malformed	1.1	1.5	1.2	8.6	16.8
		Abnormal	0.5	0.6	0.6	2.0	2.0
		Joint	0.5	0.6	0.8	3.6	2.0

The absolute and relative frequency distribution resulting from the NTP data, as well as the smoothed relative frequencies, are presented in Table 6.3. The conditional model is used for generating the number of deaths and the number of malformations as in (6.10) and (6.11). The parameters are modelled as

$$\psi_{dth} = \beta_{0,dth} + \beta_{d,dth} d, \qquad \phi_{dth} = \beta_{2,dth},$$
$$\psi_{mal} = \beta_{0,mal} + \beta_{d,mal} d, \qquad \phi_{mal} = \beta_{2,mal}.$$

Based on the parameter estimates from the conditional model for each NTP dataset, 60 parameter combinations are selected (Table 6.4). Next, for each parameter vector, an asymptotic sample is generated based on a conditional model for death and malformation jointly. Fetus and litter-based excess risk curves are computed for death and malformation jointly as well as for abnormality.

Figure 6.6 shows a selection of curves for six parameter combinations. Again, fetus-based excess risks are markedly smaller than litter-based excess risks. For $\beta_{0,dth} = \beta_{0,mal} = 0$, the difference is less pronounced. In general, fetus and litter-based curves are relatively close to each other for large background rates for death and malformation. The plots of Figure 6.6 also show that the curve for "abnormality" and the corresponding curve for "death and malformation jointly", are relatively close to each other. This is true for the fetus-based as well as for the litter-based approach. The 54 other parameter combinations considered here, result in excess risk functions for "abnormal" and for "joint" which are often comparable. However, there are a number of parameter combinations for which the curve for "abnormal" is strikingly larger than for "joint", i.e., the overly simplistic model leads to higher excess risks than the correctly specified joint model.

In general, for an increasing value of the ratio $f = \beta_{0,dth}/\beta_{0,mal}$, the risk curves seem to get closer to each other. The same holds for increasing values of $\beta_{0,mal}$ and of $\beta_{d,mal} = \beta_{d,dth}$. Furthermore, in case of association ($\beta_{2,mal} = \beta_{2,dth} = 0.2$), the excess risk at a particular dose is in general smaller than in the case of independence.

6.4 Variability of the excess risk estimator

Previous sections have shown how the different risk approaches and their corresponding estimators compare. A further question is how large the sample variability is of these risk estimators and how to properly estimate their variances. A small simu-

Table 6.3: Absolute and relative frequencies of the number of implants.

Number of implants	absolute frequency	relative frequency	smoothed relative frequency
1	4	0.0073	0.0073
2	3	0.0054	0.0063
3	4	0.0073	0.0081
4	7	0.0127	0.0094
5	2	0.0036	0.0074
6	6	0.0109	0.0092
7	6	0.0109	0.0113
8	7	0.0127	0.0189
9	23	0.0417	0.0376
10	29	0.0526	0.0676
11	71	0.1289	0.1226
12	98	0.1779	0.1712
13	109	0.1978	0.1812
14	80	0.1452	0.1469
15	55	0.0998	0.1002
16	31	0.0563	0.0579
17	11	0.0200	0.0249
18	3	0.0054	0.0084
19	2	0.0036	0.0036
	551	1	1

Table 6.4: Parameter settings.

Parameter	values
$\beta_{0,mal}$	-6;-4;-2;0
$\beta_{0,dth} = f\beta_{0,mal}$ with	
f	0.25;0.5;1
$\beta_{d,mal} = \beta_{d,dth}$	2;4;6
$\beta_{2,mal} = \beta_{2,dth}$	0.0;0.2

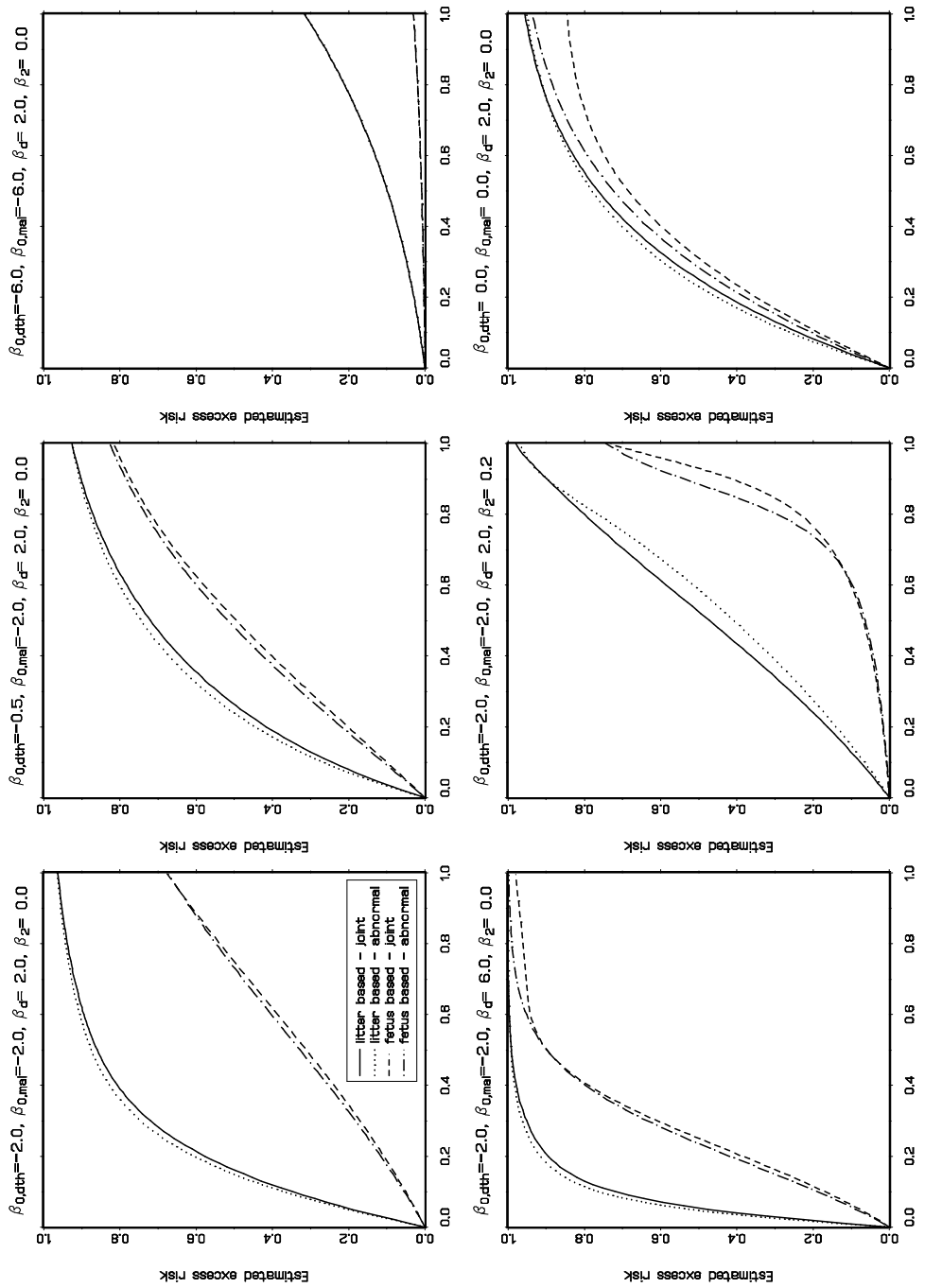

Figure 6.6: *Excess risk versus dose for asymptotic samples based on the conditional model.*

lation study showed that there are no substantial differences in variability between the different strategies to determine risk. Therefore, attention is focused on the estimation of the standard error $se(\hat{r}^*(d))$ for one particular setting. Using simulated data, a conditional model is fitted to the number of abnormal fetuses in a dam, after which the litter-based excess risk estimator $\hat{r}^*(d)$ is computed.

The standard error of $\hat{r}^*(d)$ is estimated based on the delta method. In general, the excess risk does not only depend on the dose administered to a dam, but also on unknown parameters $(\theta_1, \ldots, \theta_k) \equiv \boldsymbol{\theta}$. As a consequence, the excess risk is represented here by $r^*(d; \boldsymbol{\theta})$. Under some regularity conditions,

$$C^{1/2}(\hat{r}^*(d; \boldsymbol{\theta}) - r^*(d; \boldsymbol{\theta})) \xrightarrow{D} N(0; \boldsymbol{\Delta}\Sigma\boldsymbol{\Delta}'),$$

where C is the number of clusters in the dataset, $\hat{r}^*(d; \boldsymbol{\theta}) \equiv r^*(d; \hat{\boldsymbol{\theta}})$, $\boldsymbol{\Delta} = \frac{\partial r^*(d; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \equiv \boldsymbol{\Delta}(\boldsymbol{\theta})$ and Σ is the asymptotic variance-covariance matrix of $C^{1/2}\hat{\boldsymbol{\theta}}$. Here, the parameter $\boldsymbol{\theta}$ is estimated via maximum likelihood methodology. Hence, for large samples, the variance of $r^*(d; \hat{\boldsymbol{\theta}})$ can be approximated by $\boldsymbol{\Delta}(\hat{\boldsymbol{\theta}})I(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{\Delta}(\hat{\boldsymbol{\theta}})'$, where $I(\boldsymbol{\theta})$ is the information matrix of $\boldsymbol{\theta}$.

A key issue when studying excess risks is the estimation of the distribution $P(m)$. In addition to the estimators for the regression parameters, also the estimator for the distribution of the number of implants, $P(m)$, contributes to the variability of $\hat{r}^*(d)$. Clearly, in simulated data, the distribution $P(m)$ is known. Assuming that the distribution of $P(m)$ is known, is referred to as approach 1. Approach 2 stands for replacing $P(m)$ by an estimator $\hat{P}(m)$ but ignoring the variability associated with this estimation. Of course, when analysing real data, approach 1 is impossible and approach 2 is incorrect. They are included to assess the difference with approach 3: the delta method which correctly accounts for all sources of variability.

A small sample simulation study is performed with the conditional model for death and malformation jointly as the underlying generating probability model, with parameters $\beta_{0,dth} = \beta_{0,mal} = \beta_{2,dth} = \beta_{2,mal} = 0$ and $\beta_{d,dth} = \beta_{d,mal} = 2$. The distribution $P(m)$ of the number of implants is taken as in Table 6.3 (referred to as NTP), uniform on $\{1, \ldots, 19\}$ or a truncated Poisson model. A total number of 120 and 1920 clusters is considered, equally distributed over dose levels 0, 0.25, 0.5 and 1. Excess risks are calculated at doses 0.1 and 0.9, based on 500 simulation runs. For each approach, the standard error of the excess risk estimates is compared with the mean of the standard errors obtained by applying the delta method on the data for each run. Three estimators for the distribution $P(m)$ are chosen:

the multinomial model on $\{1, \ldots, 19\}$, a truncated Poisson model and a smoothed multinomial model.

The simulation results are shown in Table 6.5. For each setting, both measures of variability, i.e., the simulated error $se(r^*)$ and the mean estimated error $\overline{se}$, are very comparable for approach 1. This is to be expected since $P(m)$ is known and not estimated. In the first setting with $P(m)$ as NTP and $\hat{P}(m)$ multinomial, both measures are quite different for approach 2. For the smaller sample size, the extra variability induced by estimating $P(m)$ is not taken into account correctly by approach 3. Since several probabilities $P(m)$ in Table 6.3 are small ($P(m) < 0.01$ for 8 values of m), observed relative frequencies (the ML estimators for the multinomial probabilities) for these sizes can be zero, which result in zero rows and zero columns in the estimated large sample variance-covariance matrix used in the delta method. Indeed, in the simulation study with 120 clusters, only in 1% of runs are all observed frequencies non-zero. In the other runs, the number of zero elements varies between 1 and 7. In the simulation study with 1920 clusters in each experiment, all 500 runs result in non-zero observed frequencies. Thus, the discrepancy observed with approach 3 for experiments of size 120 is clearly related to the occurrence of zero frequencies. In order to get further evidence, the previous simulations are repeated with the distribution of the number of implants uniform on $\{1, \ldots, 19\}$. The results are shown in the second part of Table 6.5. Now, both error measures for approach 3 result in similar values. It turns out that for the smaller sample size, all observed frequencies are non-zero in 492 out of 500 runs, which supports the stated claim.

Returning to $P(m)$ based on the NTP data, the use of a Poisson model for $\hat{P}(m)$ is investigated as an alternative for the multinomial model. Assuming that there is at least one implant in a litter (as in Table 6.3), a modified Poisson distribution is considered: $\hat{P}(m) = \exp^{-\hat{\gamma}} \hat{\gamma}^{m-1}/(m-1)!$ if $m = 1, 2, \ldots$ and 0 if otherwise. The ML estimator $\hat{\gamma}$ is $\overline{m} - 1$. In order to deal with the infinite number of terms in the expression of the excess risk, the Poisson distribution is truncated at $m = 19$ and rescaled to unit sum. The third part of Table 6.5 shows that the results of approaches 2 and 3 are close to each other, but markedly different from the results of approach 1. This is probably caused by the rather poor fit of the modified Poisson model to the cluster sizes of Table 6.3. Omitting the variability induced by $\hat{\gamma}$ in the truncated Poisson distribution, turns out to have no considerable effect on the variability of the excess risk estimator. Rather than using a smoothed version

Table 6.5: Standard errors for the excess risk estimator based on three approaches (appr.) and two error measures, for some combinations of $P(m)$, $\hat{P}(m)$, dose and number of dams.

$P(m)$	$\hat{P}(m)$	dose	# dams	measure	appr. 1	appr. 2	appr. 3
NTP	multinomial	0.1	120	$se(r^*)$	0.030471	0.149240	0.149240
				$\overline{se}$	0.030005	0.037478	0.047863
			1920	$se(r^*)$	0.007241	0.015885	0.015885
				$\overline{se}$	0.007364	0.007425	0.015201
		0.9	120	$se(r^*)$	0.024368	0.040869	0.040869
				$\overline{se}$	0.024324	0.017122	0.019023
			1920	$se(r^*)$	0.005957	0.007926	0.007926
				$\overline{se}$	0.006068	0.005958	0.007942
uniform	multinomial	0.1	120	$se(r^*)$	0.031070	0.038265	0.038265
				$\overline{se}$	0.030853	0.031213	0.037202
			1920	$se(r^*)$	0.007543	0.008616	0.008616
				$\overline{se}$	0.007586	0.007588	0.008875
		0.9	120	$se(r^*)$	0.023693	0.025229	0.025229
				$\overline{se}$	0.024345	0.023902	0.025443
			1920	$se(r^*)$	0.006104	0.006309	0.006309
				$\overline{se}$	0.006119	0.006120	0.006405
NTP	Poisson	0.1	120	$se(r^*)$	0.030471	0.049841	0.049841
				$\overline{se}$	0.030005	0.048727	0.049203
			1920	$se(r^*)$	0.007241	0.012224	0.012224
				$\overline{se}$	0.007364	0.012223	0.012347
		0.9	120	$se(r^*)$	0.024368	0.004211	0.004211
				$\overline{se}$	0.024324	0.003924	0.003953
			1920	$se(r^*)$	0.005957	0.000849	0.000849
				$\overline{se}$	0.006068	0.000844	0.000852
Poisson	Poisson	0.1	120	$se(r^*)$	0.049558	0.050028	0.050028
				$\overline{se}$	0.048639	0.048352	0.048853
			1920	$se(r^*)$	0.011853	0.011738	0.011738
				$\overline{se}$	0.012238	0.012169	0.012294
		0.9	120	$se(r^*)$	0.004142	0.004367	0.004367
				$\overline{se}$	0.003554	0.003677	0.003706
			1920	$se(r^*)$	0.000845	0.000871	0.000871
				$\overline{se}$	0.000846	0.000875	0.000882
NTP	multinomial + smoothing	0.1	120	$se(r^*)$	0.030471	0.044408	0.044408
				$\overline{se}$	0.030005	0.031062	0.056985
			1920	$se(r^*)$	0.007241	0.015306	0.015306
				$\overline{se}$	0.007364	0.007424	0.014920
		0.9	120	$se(r^*)$	0.024368	0.026427	0.026427
				$\overline{se}$	0.024324	0.023111	0.030569
			1920	$se(r^*)$	0.005957	0.007766	0.007766
				$\overline{se}$	0.006068	0.005963	0.007863

of the relative frequencies of the NTP data for the distribution $P(m)$, a truncated Poisson distribution as described above, is considered now. This allows one to get an idea of the effect of the type of distribution $P(m)$ on the simulation results when a truncated Poisson model is chosen for $\hat{P}(m)$. A value of γ is computed based on the NTP data. Results are given in the fourth part of Table 6.5. All measures of variability of $\hat{r}^*$ are similar now, which was expected since the three strategies consider a truncated Poisson distribution for the computation of the excess risk estimate and its standard error. Again, the influence of including the variability related to the estimation of γ in the calculation of the standard error is negligible.

In order to propose an appropriate distribution $\hat{P}(m)$ in the case $P(m)$ is a local linear smoothed cluster frequency based on the NTP data, a multinomial model combined with a smoothing technique is studied. Rather than considering the observed relative frequencies, a simple form of smoothing is applied (Santner and Duffy, 1989, p. 53)

$$\left(\text{observed number of clusters of size } m + \frac{1}{2}\right) / \left(\text{total number of clusters} + \frac{19}{2}\right).$$

The last part of Table 6.5 indicates that for small experiments (size=120), approach 3 now seems to overestimate the variance of $\hat{r}^*$ to some extent. This is still preferable to approach 2 which use would result in inappropriate confidence limits. Comparing these last results with a multinomial model without smoothed relative frequencies, it turns out that this simple smoothing technique leads to satisfying results for approach 3.

6.5 Concluding remarks

Developmental toxicity studies are complicated by the hierarchical (death, malformation, healthy fetus), clustered (fetuses within litters) and multivariate (several malformation indicators and low birth weight) nature of the data. As a consequence, a multitude of modelling strategies, with varying degrees of simplification, have been proposed in the literature. Such choices are often subjective and can affect the quantitative risk assessment based on the fitted models.

While ignoring others for conciseness, the emphasis was on the choice between (1) the beta-binomial model versus the conditional model proposed by Molenberghs and Ryan (1999), (2) modelling death only, modelling malformation only, modelling a collapsed outcome indicating death or malformation (termed "abnormal") or a

joint model for death and malformation. The main emphasis has been put on (3) the distinction between fetus-based and litter-based risk assessment.

It has been argued that effective doses calculated from the litter-based approach are between 5 and 10 times smaller than those obtained from the fetus-based approach. Thus, while the latter seems to be the standard in practice, it is deduced that a litter-based approach should be considered far more often. Furthermore, from a biological perspective, one could argue that litter-based inference makes sense since a litter represents the typical pregnancy outcome in a rodent, compared with a single birth in humans. However, in general, litter-based risk assessment has not been widely studied, nor compared with fetus-based risk assessment in a systematic way. While this chapter does not resolve the issue of whether to use fetus or litter-based risk assessment procedures, it raises the question in a new way and provides a convincing argument that further work, statistical and biological, is needed on this topic.

In most cases, the beta-binomial model yields somewhat smaller ED's than the conditional approach, but the differences are less pronounced. A joint model for death and malformation yields in some cases approximately the same risk as a collapsed indicator for abnormality, but there are regions in the parameter space where the former yields considerably larger ED's. As a result, a joint modelling strategy is recommended.

Whenever risk assessment is based on the conditional model, as well as for litter-based risks under the beta-binomial model, the distribution $P(m)$ of the number of implants needs to be estimated and sampling variability needs to be incorporated in the estimator. Whenever some frequencies are near zero, a careful reflection on the estimator for $P(m)$ is necessary. A multinomial model may be inadequate due to sampling zeros, whereas more parsimonious models such as a Poisson model can provide an inadequate fit. As a compromise, it was found that a smoothed multinomial model performs relatively accurately.

Chapter 7

Frequentist versus Bayesian inference in power models

In the previous chapters, simple forms for the linear predictors describing main effects and associations have been considered. The predictors chosen were linear functions of the dose administered to a pregnant dam. Due to the simplicity of these predictors, there are no specific inferential issues related to this type of predictors. However, these models can be too restrictive to adequately describe the dose-response relationship in real applications. Also, simple expressions of linear predictors may fail to estimate effective doses and virtually safe doses accurately. Since quantitative risk assessment is based on extrapolation to very low excess risks (e.g., 10^{-4}), appropriate models for the parameters should be aimed at.

Rather than modelling the parameters by means of a linear function of dose, quadratic effects could be added to the predictor:

$$g(\xi_i) = \beta_0 + \beta_1 d_i + \beta_2 d_i^2,$$

where g is some link function, ξ_i is a model parameter and d_i is the dose administered to dam i. Clearly, quadratic models can be generalized to polynomial models:

$$g(\xi_i) = \beta_0 + \beta_1 d_i + \beta_2 d_i^2 + \ldots + \beta_p d_i^p,$$

in which the order of the polynomial p is a positive integer. Alternatives of polynomial predictors are fractional polynomials (Royston and Altman, 1994), which are used e.g., in Geys *et al.* (1999a).

In contrast with the previous predictor functions, which are linear in the parameters, the class of predictors can be enlarged by including non-linear predictors. In this chapter, the focus is on a subgroup of non-linear models, called power models.

More specifically, power predictors of the following type will be studied:

$$g(\xi_i) = \alpha + \beta d_i^{\gamma}. \tag{7.1}$$

One notices that by setting $\gamma \equiv 1$, expression (7.1) simplifies to a linear model. Furthermore, if the covariate is positive, model (7.1) can be expressed by means of an exponential function of dose:

$$g(\xi_i) = \alpha + \beta e^{\gamma d_i^*},$$

with $d_i^* = \ln(d_i)$. This alternative way of expressing a power model is used e.g., by Cox and Hinkley (1978, p. 92). It is worthwhile to investigate whether these power models can allow a better dose-response modelling and improve the quantitative risk assessment procedure.

Throughout this chapter, attention is confined to independent binary responses. In the context of developmental toxicity experiments, the adverse event of having a litter with at least one malformed fetus, can be taken as an illustration. Other examples are the adverse event of a cluster with at least one dead fetus or a cluster with at least one abnormal fetus. Focusing on the logit link function, the power model can be represented by:

$$\text{logit}(\pi_i) = \alpha + \beta d_i^{\gamma}. \tag{7.2}$$

The implementation of power predictors in models which take the litter effect into account, such as the beta-binomial model and the conditional model, is a topic of future research.

In dose-response modelling, power models are used as alternatives of linear models, since γ in (7.1) is a shape parameter and allows more flexibility of the dose-response curve. It seems that power models are commonly implemented primarily in order to get a better fit to the data, rather than for testing purposes. However, in this chapter, the emphasis will be on the latter aspect.

The use of power models invokes some interesting statistical issues. The key item is linked with the effect of the covariate (e.g., dose given to a dam) on the non-linear predictor. The case of no effect of dose d_i on the model parameter π_i can be rephrased as $\beta = 0$ *or* $\gamma = 0$. One notices that this case corresponds to the union of two planes in the parameter space of α, β and γ, i.e., the planes with equation $\beta = 0$ and $\gamma = 0$. Furthermore, the condition that $\beta = 0$ or $\gamma = 0$ is equivalent

with $\beta\gamma = 0$. As a consequence, the restriction put on the model by implying no dose effect, is no longer linear in the parameters. One notices that the parameter γ is not identifiable if $\beta = 0$, since the model then reduces to $\text{logit}(\pi_i) = \alpha$. If $\gamma = 0$, the model simplifies to $\text{logit}(\pi_i) = \alpha + \beta$. In that case, one cannot identify α and β separately, although their sum is identifiable. Hence, when the covariate has no influence on the power predictor, the parameters α, β and γ are not identifiable anymore.

The non-identifiability of regression parameters results in some interesting statistical problems. First, fitting models with power predictors might be complicated if the dose effect is weak, since convergence problems can be expected in that case. If the dose effect is absent, then the regression parameters of the power model under consideration are non-identifiable. Secondly, the effect of dose d_i on π_i can be investigated via testing the null hypothesis $H_0 : \beta\gamma = 0$. Performing a test of no dose effect can be approached e.g., from a frequentist point of view. Section 7.1 shows that this approach leads to complications which are also due to non-identifiable parameters. In Section 7.2, it is shown how Bayes factors can provide a way out here.

7.1 A frequentist approach

In this section, the effect of dose d_i on the non-linear predictor $\text{logit}(\pi_i)$ is investigated via testing the null hypothesis $H_0 : \beta\gamma = 0$, from a frequentist point of view.

Using the likelihood ratio (LR) methodology, the following generalized LR test statistic can be considered:

$$\Lambda_n \equiv \lambda(Y_1, \ldots, Y_n) = \frac{\sup_{\boldsymbol{\theta} \in \Theta_0} L(\boldsymbol{\theta}; Y_1, \ldots, Y_n)}{\sup_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}; Y_1, \ldots, Y_n)},$$

where $Y_1, \ldots, Y_n$ are the binary response variables and where $\boldsymbol{\theta} = (\alpha, \beta, \gamma)^t$. The parameter space Θ consists of all parameter vectors $\boldsymbol{\theta} \in I\!R^3$, while the parameter subspace Θ_0 contains all $\boldsymbol{\theta}$ for which $\beta\gamma = 0$. The likelihood in the numerator of Λ_n reaches its maximum for some value of π, say π^*, or equivalently, some value of the logit of π equal to $\text{logit}(\pi^*) \equiv \delta$. The denominator of Λ_n equals $L(\hat{\boldsymbol{\theta}}; Y_1, \ldots, Y_n)$, with $\hat{\boldsymbol{\theta}}$ the ML estimator of $\boldsymbol{\theta}$. For regular cases, Λ_n has an asymptotic χ^2 null distribution. However, in the settings considered here, there is a problem. Indeed,

transforming the value π^* to the regression parameters, the maximum of the likelihood under the null model is obtained for

$$\text{all } \boldsymbol{\theta} \text{ for which } \alpha = \text{logit}(\pi^*), \beta = 0$$

and

$$\text{all } \boldsymbol{\theta} \text{ for which } \alpha + \beta = \text{logit}(\pi^*), \gamma = 0.$$

As a consequence, under the hypothesis of no dose effect, the likelihood is maximized at any parameter combination on these two intersecting lines. Even if the generalized likelihood ratio test statistic is asymptotically χ^2 distributed, a remaining question would be the number of degrees of freedom (one, two,...).

7.2 A Bayesian approach

This section considers a Bayesian framework for testing the null hypothesis of no dose effect on the predictor $\text{logit}(\pi_i)$. Several Bayesian testing procedures have been proposed in the literature (Kass and Raftery, 1995). Bayesian hypothesis testing can be performed by means of *Bayes factors*, which are introduced now.

In this context, the null model corresponds to the hypothesis H_0 of no effect of dose d_i on the predictor:

$$H_0 : \text{logit}(\pi_i) = \delta,$$

with δ some constant. Under the alternative model, the dose administered to a dam has an effect on the predictor via a power function:

$$H_1 : \text{logit}(\pi_i) = \alpha + \beta d_i^{\gamma}.$$

In this chapter, the data considered are assumed to be independent binary responses and are represented by $(y_1, \ldots, y_n) \equiv \boldsymbol{y}$. The Bayes factor can be defined as

$$B_{10} = \frac{P(\boldsymbol{y} \mid H_1)}{P(\boldsymbol{y} \mid H_0)}. \tag{7.3}$$

From (7.3), it follows that the Bayes factor can be viewed as measuring the relative success of H_0 and H_1 at predicting the data (Kass and Raftery, 1995). Representing the a priori probabilities of the null and alternative hypotheses by $P(H_0)$ and $P(H_1)$ respectively, one notices that

$$B_{10}\frac{P(H_1)}{P(H_0)} = \frac{P(H_1, \boldsymbol{y})}{P(H_0, \boldsymbol{y})} = \frac{P(H_1 \mid \boldsymbol{y})}{P(H_0 \mid \boldsymbol{y})},$$

i.e., the Bayes factor multiplied by the prior odds of H_1 results in the posterior odds of H_1. If both hypotheses are equally probable a priori, then $P(H_0) = P(H_1) = 0.5$ and the Bayes factor B_{10} equals the posterior odds in favour of H_1.

The two components of the Bayes factor, i.e., $P(\boldsymbol{y} \mid H_0)$ and $P(\boldsymbol{y} \mid H_1)$, are computed by integrating the joint density of the data and the regression parameters of the corresponding model, over its parameters. Hence, the probability $P(\boldsymbol{y} \mid H_0)$ is calculated by means of the expression

$$P(\boldsymbol{y} \mid H_0) = \int_{\Theta_0} P(\boldsymbol{y} \mid \delta, H_0)\omega(\delta \mid H_0)d\delta, \tag{7.4}$$

where Θ_0 is the parameter space of δ and where $\omega(\delta \mid H_0)$ is the prior density of δ in the null model. The probability $P(\boldsymbol{y} \mid H_1)$ is found in an analogous way:

$$P(\boldsymbol{y} \mid H_1) = \int_{\Theta_1} P(\boldsymbol{y} \mid \boldsymbol{\theta}_1, H_1)\omega(\boldsymbol{\theta}_1 \mid H_1)d\boldsymbol{\theta}_1, \tag{7.5}$$

with $\boldsymbol{\theta}_1$ the parameter vector under the alternative model, i.e., $\boldsymbol{\theta}_1 = (\alpha, \beta, \gamma)^t$, with Θ_1 the parameter space of $\boldsymbol{\theta}_1$ and with $\omega(\boldsymbol{\theta}_1 \mid H_1)$ the prior density of $\boldsymbol{\theta}_1$ in the alternative model. The two components of the Bayes factor are also called marginal likelihoods or integrated likelihoods (Kass and Raftery, 1995). From (7.4) and (7.5), it follows that the marginal likelihood is a weighted average of the likelihood, using the prior distribution as a weight function.

Kass and Raftery (1995) provide categories for the Bayes factor, expressing the evidence against the null hypothesis. Table 7.1 lists classes for B_{10}, as well as for $2 \ln B_{10}$, which is on the same scale as e.g., the likelihood ratio test statistic. These categories are a rough descriptive statement about standards of evidence in scientific investigation (Kass and Raftery, 1995). By comparing the computed Bayes factor of a data analysis with the classes of Table 7.1, one can make a conclusion about the effect of dose d_i on the predictor $\text{logit}(\pi_i)$.

A remaining issue is the computation of the marginal likelihoods $P(\boldsymbol{y} \mid H_0)$ and $P(\boldsymbol{y} \mid H_1)$. In the literature, an extended number of methods have been proposed (Kass and Raftery, 1995). Some of these procedures are briefly introduced here.

In a limited number of cases, the marginal likelihood can be evaluated analytically. Due to the type of method, it results in an exact value of $P(\boldsymbol{y} \mid H_k)$ with $k = 0, 1$. However, in most cases, this procedure is intractable and thus, numerical methods are needed to approximate the marginal likelihood.

A conceptually simple numerical method is the partitioning of the parameter space Θ_k into small rectangular parallelepipeds. One then evaluates the likelihood

Table 7.1: Categories for the Bayes factor expressing evidence against the null hypothesis.

B_{10}	$2 \ln B_{10}$	Evidence against H_0
1 to 3	0 to 2	not worth more than a bare mention
3 to 20	2 to 6	positive
20 to 150	6 to 10	strong
>150	>10	very strong

function $P(\boldsymbol{y} \mid \boldsymbol{\theta}_k, H_k)$ and the prior density function $\omega(\boldsymbol{\theta}_k \mid H_k)$ in a central point of that object. The parameter $\boldsymbol{\theta}_k$ equals δ if k=0 and as indicated before, $\boldsymbol{\theta}_k = (\alpha, \beta, \gamma)^t$ if k=1. Next, one multiplies this likelihood value and the value of the prior density with the volume of the parallelepiped. Finally, by summing all contributions over Θ_k, an approximation of the marginal likelihood is obtained. On the one hand, this method requires neither an ML estimate of the parameters $\boldsymbol{\theta}_k$, nor an estimate of the posterior mode, i.e., the mode of posterior density function $P(\boldsymbol{\theta}_k \mid \boldsymbol{y}, H_k)$. On the other hand, this procedure can be very inefficient in the sense that a lot of computer time might be needed for calculating an approximation of the marginal likelihood with a specified level of precision.

Rather than considering all parallelepipeds in the constructed grid of the parameter space Θ_k, one can opt to select a smaller number of points in Θ_k by means of some random mechanism and to evaluate the likelihood function in these points. This idea is used in the simple Monte Carlo estimate of the marginal likelihood:

$$\hat{P}(\boldsymbol{y} \mid H_k) = \frac{1}{m} \sum_{i=1}^{m} P(\boldsymbol{y} \mid \boldsymbol{\theta}_k^{(i)}, H_k),$$

where $\boldsymbol{\theta}_k^{(1)}, \ldots, \boldsymbol{\theta}_k^{(m)}$ is a sample from the prior density $\omega(\boldsymbol{\theta}_k \mid H_k)$. Hence, $\hat{P}(\boldsymbol{y} \mid H_k)$ is an unweighted average of the likelihoods of the sampled parameter values (Hammersley and Handscomb, 1964). Again, this method does not need an ML estimate of the parameters or an estimate of the posterior mode. However, a disadvantage of this procedure is that the simulation process can be quite inefficient (McCulloch and Rossi, 1991).

The precision of simple Monte Carlo integration can be increased by the technique of importance sampling. Rather than generating a sample of parameter values

from the prior distribution, one generates parameter values $\boldsymbol{\theta}_k^{(1)}, \ldots, \boldsymbol{\theta}_k^{(m)}$ from another distribution, say $\omega^*(\boldsymbol{\theta}_k \mid H_k)$. Then, a weighted average of the likelihood evaluated at these sampled parameter values is computed:

$$\hat{P}(\boldsymbol{y} \mid H_k) = \frac{\sum_{i=1}^m w_i P(\boldsymbol{y} \mid \boldsymbol{\theta}_k^{(i)}, H_k)}{\sum_{i=1}^m w_i}, \tag{7.6}$$

with $w_i = \omega(\boldsymbol{\theta}_k^{(i)} \mid H_k)/\omega^*(\boldsymbol{\theta}_k^{(i)} \mid H_k)$. The density $\omega^*(\boldsymbol{\theta}_k \mid H_k)$ is called the importance sampling function (Geweke, 1989). If this function is chosen to be the posterior distribution $P(\boldsymbol{\theta}_k \mid \boldsymbol{y}, H_k) = P(\boldsymbol{y} \mid \boldsymbol{\theta}_k, H_k)\omega(\boldsymbol{\theta}_k \mid H_k)/P(\boldsymbol{y} \mid H_k)$, then expression (7.6) yields

$$\hat{P}(\boldsymbol{y} \mid H_k) = \left\{ \frac{1}{m} \sum_{i=1}^m P(\boldsymbol{y} \mid \boldsymbol{\theta}_k^{(i)}, H_k)^{-1} \right\}^{-1}.$$

Hence, when generating parameter values from the posterior distribution, the estimator of the marginal likelihood is the harmonic mean of the likelihood values (Newton and Raftery, 1994). No ML estimate or posterior mode of $\boldsymbol{\theta}_k$ are required here. Furthermore, it is easy to calculate and experience suggests that it often gives results that are accurate enough for interpretation on the logarithmic scale of Table 7.1. However, this approximation is unstable (Rosenkranz, 1992).

The last estimator of the marginal likelihood described here, is based on the Schwarz criterion

$$S = \ln \frac{P(\boldsymbol{y} \mid \hat{\boldsymbol{\theta}}_1, H_1)}{P(\boldsymbol{y} \mid \hat{\boldsymbol{\theta}}_0, H_0)} - \frac{1}{2}(\dim(1) - \dim(0)) \ln(n),$$

where $\hat{\boldsymbol{\theta}}_k$ is an ML estimate of $\boldsymbol{\theta}_k$ under H_k (with $k = 0, 1$), where $\dim(k)$ is the dimension of $\boldsymbol{\theta}_k$ and where n is the sample size. The Schwarz criterion can be viewed as a rough approximation to the natural logarithm of the Bayes factor B_{10} (Kass and Raftery, 1995). Hence, the Bayes factor B_{10} can be estimated by $\exp(S)$. In the settings considered here,

$$S = \ln \frac{P(\boldsymbol{y} \mid \hat{\boldsymbol{\theta}}_1, H_1)}{P(\boldsymbol{y} \mid \hat{\delta}, H_0)} - \ln(n), \tag{7.7}$$

with $\hat{\boldsymbol{\theta}}_1 = (\hat{\alpha}, \hat{\beta}, \hat{\gamma})^t$. From (7.7), it follows that

$$B_{10} \approx \frac{P(\boldsymbol{y} \mid \hat{\boldsymbol{\theta}}_1, H_1)}{nP(\boldsymbol{y} \mid \hat{\delta}, H_0)}$$

Table 7.2: Some of the data of DEHP, as well as the estimated probabilities of observing a dam with at least one abnormal fetus, based on a power model.

Transformed dose	# dams	# dams with at least one abnormal fetus	relative frequency	estimated frequency based on a power model
1.000	30	21	0.700	0.719
1.151	26	22	0.846	0.804
1.312	26	23	0.885	0.907
1.654	24	24	1.000	0.998
2.000	25	25	1.000	1.000

and that twice the Schwarz criterion equals the usual likelihood ratio test statistic minus twice the natural logarithm of the sample size. Keeping in mind the rough interpretation of the Bayes factor on the logarithmic scale of Table 7.1, it can be shown that in large samples, the Schwarz criterion should provide a reasonable indication of the evidence (Kass and Raftery, 1995). Also, this procedure requires only the value of the likelihood ratio statistic and the number of parameters in both models. Furthermore, no prior distributions are needed.

7.2.1 Analysis of NTP data

In this section, data of the toxic agents DEHP, DYME, EG, TGDM and THEO are analysed by computing an approximation of the Bayes factor in order to test for no dose effect in the power model specified in expression (7.2). Here, the focus is on the adverse event "dam with at least one abnormal (i.e., dead or malformed) fetus". Doses are rescaled first to the [0,1] interval and then shifted to [1,2]. The latter recoding is done in order to avoid numerical problems when fitting the power model, arising from the evaluation of the non-linear predictor when the control group is considered and $\gamma \leq 0$. As an illustration, the data of the chemical DEHP which are relevant in this context, are given in Table 7.2.

Three methods are selected for the estimation of the Bayes factor. Besides the Schwarz criterion, the Bayes factor is approximated by computing the integrals (7.4) and (7.5) numerically. This is performed by partitioning the parameter space into small parallelepipeds. Two types of prior distributions are considered here: a uni-

form and a normal prior. The selected uniform prior distribution for the parameter δ of the null model has probability mass between -1 and 5. For the alternative model, the uniform prior is non-zero for $-2.5 < \alpha, \beta < 3.5$ and $2 < \gamma < 8$. The mean of the normal prior for δ in the model of no dose effect is 2, while the variance of δ equals 3. Finally, the mean vector and the variance-covariance matrix of (α, β, γ) in the power model with normal prior, is $(0.5, 0.5, 5)$ and 3 times the identity matrix respectively. Hence, the first and second moments of the selected priors are equal.

Table 7.3 shows among others, the parameter estimates of the power model under investigation (and the corresponding standard errors), as well as the parameter estimate of the null model for each of the five NTP studies. Based on the parameter estimates of the alternative model, the probability that a dam has at least one abnormal fetus can be estimated for each dose level. In case of DEHP, these probabilities are represented in the last column of Table 7.2. The components of the Schwarz criterion, i.e., the log-likelihood of the power and null models and the number of dams, are also listed in Table 7.3. Furthermore, the Schwarz criterion and the resulting estimate of the Bayes factor are given. Finally, the approximations of the Bayes factors obtained by computing the integrals (7.4) and (7.5) numerically using a uniform or a normal prior distribution, are added to this table. The values of the Bayes factor in case of a uniform prior are comparable to the ones in case of a normal prior, but they are larger than when basing on the Schwarz criterion. These results are interpreted using Table 7.1. For TGDM and THEO, there is no evidence against the null hypothesis of no dose effect in each of these three methods. However, depending on the method, there is positive to strong, strong to very strong and very strong evidence against this null hypothesis in case of EG, DEHP and DYME respectively. The descriptive statistics expressed by means of the distribution of the number of abnormal fetuses in Figures 1.2 – 1.6, are in agreement with these conclusions. Also, the findings of this section are similar to the ones of Section 3.3 in which the likelihood ratio and the Wald statistics are used to test for no dose effect.

7.2.2 Small sample simulations

Besides the NTP data analysis of the previous section, a limited small sample simulation study is performed. Analogous to the small sample simulations of Chapter 3, the doses 0, 0.25, 0.5 and 1 are selected, but due to numerical problems in the

Table 7.3: Analysis results of DEHP, DYME, EG, TGDM and THEO in which three methods are used to approximate the Bayes factor (BF).

	DEHP	DYME	EG	TGDM	THEO
$\hat{\alpha}$	0.484(1.47)	-0.168(0.96)	0.632(1.34)	-0.0761(1.12)	0.575(0.40)
$\hat{\beta}$	0.457(1.21)	0.270(0.64)	0.314(1.06)	0.172(0.82)	0.00716(0.03)
$\hat{\gamma}$	5.04(6.22)	6.14(5.40)	5.08(7.40)	2.99(5.85)	7.65(6.83)
$\hat{\delta}$	1.97	1.17	1.93	0.516	1.25
log-likel.(H_1)	-39.1	-46.6	-29.8	-68.3	-54.5
log-likel.(H_0)	-48.6	-60.2	-36.0	-70.7	-57.2
# dams	131	110	95	107	108
Schwarz	4.66	8.86	1.73	-2.24	-1.96
BF(Schwarz)	106	7038	5.64	0.106	0.141
BF(uniform)	1647	60385	49.8	0.187	0.199
BF(normal)	1754	63359	54.2	0.269	0.209

computation of the power predictor, these doses are shifted to 1, 1.25, 1.5 and 2. The adverse event under study is again a dam with at least one abnormal fetus. For each dose level, 30 binary data are generated, representing the health status of this group of dams. A value of zero indicates a dam without abnormal fetuses, while a value of one refers to the presence of at least one dead or malformed fetus. The number of simulation runs is 100. Several values of the parameter vector (α, β, γ) of the underlying power model are considered. In this study, a chosen combination of these parameters is $\alpha = 0.5$, $\beta = 0.5$ and $\gamma = 5$, which is comparable to the estimates of the DEHP study. Other selected values of (α, β, γ) are obtained by changing β from 0.5 to 0 using a step size of 0.1 (Table 7.4), by changing γ from 5 to 0 using a step size of 1 (Table 7.5) or by changing β and γ simultaneously (Table 7.6). Hence, starting from the parameter vector $(\alpha, \beta, \gamma) = (0.5, 0.5, 5)$, one approaches the null hypothesis of no dose effect via three paths.

The methods used to approximate the Bayes factor, are the same as in the previous section. Tables 7.4 – 7.6 indicate the number of simulation runs which are taken into account. For the approximation of the Bayes factor based on the uniform and the normal priors, all runs are considered in these tables. However, for the estimation of the Bayes factor using the Schwarz criterion, the procedure

for the estimation of the parameters in the power model did not always lead to convergence of the results. For each of the classes of the Bayes factor as indicated in Table 7.1, the percentages of the number of simulation runs taken into account here, are listed in Tables 7.4 – 7.6. When generating data from the null hypothesis, i.e., $(\alpha, \beta, \gamma) = (0.5, 0, 5)$, $(0.5, 0.5, 0)$ or $(0.5, 0, 0)$, the distribution of the number of simulation runs is comparable for the three methods under investigation. Virtually all runs lead to values of the Bayes factor smaller than one, implying that there is no evidence against the null hypothesis. When the underlying model is a power model, the distribution of the number of runs over these classes is similar for the uniform and the normal prior density. For the Schwarz criterion, the percentages of the number of runs in the classes with smaller values of the Bayes factor are larger than in case of the methods using prior distributions.

7.3 Concluding remarks

In this chapter, the focus is on testing the effect of dose on a power predictor. The null hypothesis of no dose effect is equivalent with setting the product of two regression parameters equal to zero. This non-linear restriction of the parameters in the null model, results in parameter unidentifiability in case the effect of dose is absent. In order to avoid the computation of the distribution of the likelihood ratio test statistic in this setting, a Bayesian approach using Bayes factors is considered here.

One of the methods which are applied to approximate the Bayes factor in this chapter, is based on the Schwarz criterion. In the other two methods applied here, the Bayes factor is estimated by integrating the marginal likelihoods numerically making use of a uniform and a normal prior distribution. It would be interesting to investigate the influence of the parameters of these prior densities on the Bayes factor. Also, one could consider other types of priors for the parameters of the null and alternative models and hence, assess the sensitivity of conclusions to the type of prior distributions used. Obviously, many other techniques for the calculation of the Bayes factor can be applied. As indicated in Section 7.2, one can also approximate the marginal likelihood by the simple Monte Carlo estimate or apply the technique of importance sampling. Furthermore, Kass and Raftery (1995) propose other methods for the calculation of the Bayes factor, e.g., Laplace's method.

Table 7.4: Distribution of the Bayes factor obtained from small sample simulations in which the Schwarz criterion is used to approximate the Bayes factor, in addition to integrations over a grid using a uniform or a normal prior density function. Several values of the parameter β of the underlying power model are considered.

α	β	γ	method	# runs	≤ 1	$1 < . \leq 3$	$3 < . \leq 20$	$20 < . \leq 150$	> 150
0.5	0.5	5	Schwarz	62	8.1	3.2	14.5	17.7	56.5
			uniform	100	0	2	6	11	81
			normal	100	0	2	5	7	86
0.5	0.4	5	Schwarz	59	20.3	6.8	20.3	16.9	35.6
			uniform	100	0	2	10	21	67
			normal	100	0	2	10	18	70
0.5	0.3	5	Schwarz	84	15.5	6.0	15.5	25.0	38.1
			uniform	100	0	0	9	20	71
			normal	100	0	0	9	18	73
0.5	0.2	5	Schwarz	83	18.1	2.4	15.7	22.9	41.0
			uniform	100	0	1	7	25	67
			normal	100	0	1	7	24	68
0.5	0.1	5	Schwarz	94	31.9	8.5	18.1	16.0	25.5
			uniform	100	5	8	18	29	40
			normal	100	5	7	20	28	40
0.5	0	5	Schwarz	84	100	0	0	0	0
			uniform	100	100	0	0	0	0
			normal	100	100	0	0	0	0

Table 7.5: Distribution of the Bayes factor obtained from small sample simulations in which the Schwarz criterion is used to approximate the Bayes factor, in addition to integrations over a grid using a uniform or a normal prior density function. Several values of the parameter γ of the underlying power model are considered.

α	β	γ	method	# runs	≤ 1	$1 < . \leq 3$	$3 < . \leq 20$	$20 < . \leq 150$	> 150
0.5	0.5	5	Schwarz	62	8.1	3.2	14.5	17.7	56.5
			uniform	100	0	2	6	11	81
			normal	100	0	2	5	7	86
0.5	0.5	4	Schwarz	81	23.5	9.9	30.9	21.0	14.8
			uniform	100	1	2	20	29	48
			normal	100	1	2	18	30	49
0.5	0.5	3	Schwarz	85	42.4	14.1	15.3	16.5	11.8
			uniform	100	16	7	33	20	24
			normal	100	16	8	32	19	25
0.5	0.5	2	Schwarz	79	89.9	6.3	3.8	0	0
			uniform	100	75	12	11	2	0
			normal	100	73	14	11	2	0
0.5	0.5	1	Schwarz	86	100	0	0	0	0
			uniform	100	96	3	0	1	0
			normal	100	96	3	1	0	0
0.5	0.5	0	Schwarz	89	100	0	0	0	0
			uniform	100	99	1	0	0	0
			normal	100	99	1	0	0	0

Table 7.6: Distribution of the Bayes factor obtained from small sample simulations in which the Schwarz criterion is used to approximate the Bayes factor, in addition to integrations over a grid using a uniform or a normal prior density function. Several values of the parameters β and γ of the underlying power model are considered.

α	β	γ	method	# runs	≤ 1	$1 < . \leq 3$	$3 < . \leq 20$	$20 < . \leq 150$	> 150
0.5	0.5	5	Schwarz	62	8.1	3.2	14.5	17.7	56.5
			uniform	100	0	2	6	11	81
			normal	100	0	2	5	7	86
0.5	0.4	4	Schwarz	86	29.1	5.8	25.6	20.9	18.6
			uniform	100	2	7	16	26	49
			normal	100	2	7	15	25	51
0.5	0.3	3	Schwarz	84	65.5	14.3	13.1	4.8	2.4
			uniform	100	37	20	26	13	4
			normal	100	38	19	27	12	4
0.5	0.2	2	Schwarz	86	96.5	1.2	1.2	0	1.2
			uniform	100	97	0	2	0	1
			normal	100	97	0	2	0	1
0.5	0.1	1	Schwarz	82	98.8	1.2	0	0	0
			uniform	100	99	1	0	0	0
			normal	100	99	1	0	0	0
0.5	0	0	Schwarz	83	100	0	0	0	0
			uniform	100	100	0	0	0	0
			normal	100	100	0	0	0	0

In this chapter, attention is restricted to independent binary data. Considering binary data of developmental toxicity studies at the fetus level by fitting models for clustered data such as the beta-binomial and the conditional models, is a topic of future research. In this respect, a comparison with the results of Aerts and Claeskens (1999) will then be possible. In that paper, these authors analyse the several malformation types of the THEO study, by means of the conditional model. The P-value of the likelihood ratio statistic for the null hypothesis of no dose effect, is computed by means of a parametric bootstrap procedure. A question of interest is about the similarity between their results and the ones obtained by means of the Bayes factor.

References

Aerts, M., Augustyns, I. and Janssen, P., Smoothing sparse multinomial data using local polynomial fitting, *Nonparametric Statistics*, **8** (1997) 127–147.

Aerts, M. and Claeskens, G., Bootstrapping pseudolikelihood models for clustered binary data, *Annals of the Institute of Statistical Mathematics* (1999) In Press.

Aerts, M., Declerck, L. and Molenberghs, G., Likelihood misspecification and safe dose determination for clustered binary data, *Environmetrics*, **8** (1997) 613–627.

Agresti, A., *Categorical Data Analysis,* (Wiley, New York, 1990).

Altham, P.M.E., Two generalizations of the binomial distribution, *Applied Statistics*, **27** (1978) 162–167.

Autian, J., Toxicity and health threats of phthalate esters: Review of the literature, *Environmental Health Perspectives*, **4** (1973) 3–26.

Bahadur, R.R., A representation of the joint distribution of responses to n dichotomous items, in H. Solomon (Ed.) *Studies in Item Analysis and Prediction,* (Stanford Mathematical Studies in the Social Sciences VI. Stanford University Press, Stanford, California, 1961, pp 158–168).

Catalano, P.J. and Ryan, L.M., Bivariate latent variable models for clustered discrete and continuous outcomes, *Journal of the American Statistical Association*, **87** (1992) 651–658.

Catalano, P.J., Scharfstein, D.O., Ryan, L.M., Kimmel, C.A. and Kimmel, G.L., Statistical model for fetal death, fetal weight and malformation in developmental toxicity studies, *Teratology*, **47** (1993) 281–290.

Chen, J.J. and Kodell, R.L., Quantitative risk assessment for teratological effects, *Journal of the American Statistical Association*, **84** (1989) 966–971.

Chen, J.J., Kodell, R.L., Howe, R.B. and Gaylor, D.W., Analysis of trinomial responses from reproductive and developmental toxicity experiments, *Biometrics*, **47** (1991) 1049–1058.

Chen, J.J. and Li, L.-A., Dose-response modeling of trinomial responses from developmental experiments, *Statistica Sinica*, **4** (1994) 265–274.

Claeskens, G., Aerts, M., Bootstrapping local polynomial estimators in likelihood-based models, (1999) Submitted for publication.

Clapp, D.E., Zaebst, D.D. and Herrick, R.F., Measuring exposures to glycol ethers, *Environmental Health Perspectives*, **57** (1984) 91–95.

Cox, D.R., The analysis of multivariate binary data, *Applied Statistics*, **21** (1972) 113–120.

Cox, D.R. and Hinkley, D.V., *Problems and Solutions in Theoretical Statistics,* (Chapman and Hall, London, 1978).

Cox, D.R. and Wermuth, N., A note on the quadratic exponential binary distribution, *Biometrika*, **81** (1994) 403–408.

Crump, K.S., A new method for determining allowable daily intakes, *Fundamental and Applied Toxicology*, **4** (1984) 854–871.

Crump, K.S. and Howe, R.B., A review of methods for calculating statistical confidence limits in low dose extrapolation, in D.B. Clayson, D. Krewski and I. Munro (Eds.) *Toxicological Risk Assessment* Vol. I: Biological and Statistical Criteria, (CRC Press, Boca Raton, 1985, pp 187–203).

Dale, J.R., Global cross-ratio models for bivariate, discrete, ordered responses, *Biometrics*, **42** (1986) 909–917.

Declerck, L., Aerts, M. and Molenberghs, G., Behaviour of the likelihood ratio test statistic under a Bahadur model for exchangeable binary data, *Journal of Statistical Computation and Simulation*, **61** (1998) 15–38.

Declerck, L., Molenberghs, G. and Aerts, M., The influence of simplifying the Bahadur model on the likelihood ratio statistic, in C.E. Minder and H. Friedl (Eds.) *Good Statistical Practice: Proceedings of the 12th International Workshop on Statistical Modelling, Biel/Bienne, Switzerland, 7-11 July, 1997,* (Schriftenreihe der Osterreichischen Statistischen Gesellschaft, Wien, pp 130–134).

Declerck, L., Molenberghs, G., Aerts, M. and Ryan, L., Litter-based methods in developmental toxicity risk assessment, *Environmental and Ecological Statistics* (1999) In Press.

Diggle, P.J., Liang, K.-Y. and Zeger, S.L., *Analysis of Longitudinal Data,* (Clarendon Press, Oxford, 1994).

Efron, B., Double exponential families and their use in generalized linear regression, *Journal of the American Statistical Association*, **81** (1986) 709–721.

Environmental Protection Agency, *Guidelines for Developmental Toxicity and Risk Assessment*, Fed. Regist., 56, 63798 (1991).

Fitzmaurice, G.M., Laird, N.M. and Tosteson, T.D., Polynomial exponential models for clustered binary outcomes, unpublished manuscript (1999).

Food and Drug Administration, *Guidelines for Reproduction and Studies for Safety Evaluation of Drugs for Human Use*, Bureau of Drugs, Rockville, MD (1966).

Gart, J.J., Krewski, D., Lee, P.N., Tarone, R.E. and Wahrendorf, J., *Statistical Methods in Cancer Research, Volume III: The Design and Analysis of Long-Term Animal Experiments,* (International Agency for Research on Cancer, Lyon, 1986).

Gaylor, D.W., Quantitative risk analysis for quantal reproductive and developmental effects, *Environmental Health Perspectives*, **79** (1989) 243–246.

George, E.O. and Bowman, D., A full likelihood procedure for analysing exchangeable binary data, *Biometrics*, **51** (1995) 512–523.

George, J.D., Price, C.J., Kimmel, C.A. and Marr, M.C., The developmental toxicity of triethylene glycol dimethyl ether in mice, *Fundamental and Applied Toxicology*, **9** (1987) 173–181.

Geweke, J., Bayesian inference in econometric models using Monte Carlo integration, *Econometrica*, **57** (1989) 1317–1340.

Geys, H., Molenberghs, G., Declerck, L. and Ryan, L.M., Flexible quantitative risk assessment for developmental toxicity based on fractional polynomial predictors, (1999a) Submitted for publication.

Geys, H., Molenberghs, G. and Ryan, L.M., Pseudo-likelihood modelling of multivariate outcomes in developmental toxicology, *Journal of the American Statistical Association* (1999) In Press.

Geys, H., Regan, M., Catalano, P. and Molenberghs, G., Two latent variable risk assessment approaches for combined continuous and discrete outcomes from developmental toxicity data, (1999b) Submitted for publication.

Hammersley, J.M. and Handscomb, D.C., *Monte Carlo Methods,* (Chapman and Hall, London, 1964).

Hauck, W.W. and Donner, A., Wald's test as applied to hypotheses in logit analysis, *Journal of the American Statistical Association*, **72** (1977) 851–853.

Kass, R.E. and Raftery, A.E., Bayes factors, *Journal of the American Statistical Association*, **90** (1995) 773–795.

Kayser, S.R. and Cupit, G.C., Teratogenicity of selected asthmatic medications, *Drug Intelligence and Clinical Pharmacy*, **12** (1978) 173.

Kimmel, C.A. and Gaylor, D.W., Issues in qualitative and quantitative risk analysis for developmental toxicology, *Risk Analysis*, **8** (1988) 15–19.

Kleinman, J.C., Proportions with extraneous variance: single and independent samples, *Journal of the American Statistical Association*, **68** (1973) 46–54.

Krewski, D. and Van Ryzin, J., Dose-response models for quantal response toxicity data, in M. Csorgo, D. Dawson, J.N.K. Rao and E. Saleh (Eds.), *Statistics and Related Topics,* (North-Holland, New York, 1981, pp 201–231).

Krewski, D. and Zhu, Y., Applications of multinomial dose-response models in developmental toxicity risk assessment, *Risk Analysis*, **14** (1994) 613–628.

Krewski, D. and Zhu, Y., A simple data transformation for estimating benchmark doses in developmental toxicity experiments, *Risk Analysis*, **15** (1995) 29–40.

Kupper, L.L. and Haseman, J.K., The use of a correlated binomial model for the analysis of certain toxicological experiments, *Biometrics*, **34** (1978) 69–76.

Kupper, L.L., Portier, C., Hogan, M.D. and Yamamoto, E., The impact of litter effects on dose-response modeling in teratology, *Biometrics*, **42** (1986) 85–98.

Lefkopoulou, M., Moore, D. and Ryan, L., The analysis of multiple correlated binary outcomes: application to rodent teratology experiments, *Journal of the American Statistical Association*, **84** (1989) 810–815.

Leisenring, W. and Ryan, L., Statistical properties of the NOAEL, *Regulatory Toxicology and Pharmacology*, **15** (1992) 161.

Liang, K.-Y. and Hanfelt, J., On the use of the quasi-likelihood method in teratological experiments, *Biometrics*, **50** (1994) 872–880.

Liang, K.-Y. and Zeger, S.L., Longitudinal data analysis using generalized linear models, *Biometrika*, **73** (1986) 13–22.

Liang, K.-Y., Zeger, S.L. and Qaqish, B., Multivariate regression analyses for categorical data, *Journal of the Royal Statistical Society, Series B*, **54** (1992) 3–40.

Lindström, P., Morrissey, R.E., George, J.D., Price, C.J., Marr, M.C., Kimmel, C.A. and Schwetz, B.A., The developmental toxicity of orally administered theophylline in rats and mice, *Fundamental and Applied Toxicology*, **14** (1990) 167–178.

McCulloch, R.E. and Rossi, P.E., A Bayesian approach to testing the arbitrage pricing theory, *Journal of Econometrics*, **49** (1991) 141–168.

Molenberghs, G., Declerck, L. and Aerts, M., Quantitative risk assessment for clustered binary data, in G.U.H. Seeber, B.J. Francis, R. Hatzinger and G. Steckel-Berger (Eds.) *Statistical Modelling: Proceedings of the 10th International Workshop on Statistical Modelling, Innsbruck, Austria, 10-14 July, 1995,* (Springer-Verlag, pp 193–200).

Molenberghs, G., Declerck, L. and Aerts, M., Misspecifying the likelihood for clustered binary data, *Computational Statistics & Data Analysis*, **26** (1998) 327–349.

Molenberghs, G., Geys, H., Declerck, L., Claeskens, G. and Aerts, M., Analysis of clustered multivariate data from developmental toxicity studies, in R. Payne and P. Green (Eds.) *Proceedings in Computational Statistics, 13th Symposium held in Bristol, Great Britain, 24-28 August, 1998,* (Physica-Verlag, pp 3–14).

Molenberghs, G. and Lesaffre, E., Marginal modelling of correlated ordinal data using a multivariate Plackett distribution, *Journal of the American Statistical Association*, **89** (1994) 633–644.

Molenberghs, G. and Ryan, L.M., An exponential family model for clustered multivariate binary data, *Environmetrics* (1999) In Press.

Morgan, B.J.T., *Analysis of Quantal Response Data,* (Chapman and Hall, London, 1992).

Newton, M.A. and Raftery, A.E., Approximate Bayesian inference by the weighted likelihood bootstrap (with discussion), *Journal of the Royal Statistical Society, Series B*, **56** (1994) 3–48.

NIOSH, U.S. Department of Health and Human Services, Public Health Service, Center for Disease Control, National Institute for Occupational Safety and Health (1983). *Current Intelligence Bulletin 39: Glycol Ethers 2-Methoxyethanol and 2-Ethoxyethanol.*

Pendergast, J.F., Gange, S.J., Newton, M.A., Lindstrom, M.J., Palta, M. and Fisher, M.R., A survey of methods for analyzing clustered binary response data, *International Statistical Review*, **64** (1996) 89–118.

Prentice, R.L., Correlated binary regression with covariates specific to each binary observation, *Biometrics*, **44** (1988) 1033–1048.

Price, C.J., Kimmel, C.A., George, J.D. and Marr, M.C., The developmental toxicity of diethylene glycol dimethyl ether in mice, *Fundamental and Applied Toxicology*, **8** (1987) 115–126.

Price, C.J., Kimmel, C.A., Tyl, R.W. and Marr, M.C., The developmental toxicity of ethylene glycol in rats and mice, *Toxicology and Applied Pharmacology*, **81** (1985) 113–127.

Rai, K. and Van Ryzin, J., A dose-response model for teratological experiments involving quantal responses, *Biometrics*, **47** (1985) 825–839.

Rosenkranz, S., The Bayes factor for model evaluation in a hierarchical Poisson model for area counts, Ph.D. dissertation, University of Washington, Dept. of Biostatistics (1992).

Rotnitzky, A. and Wypij, D., A note on the bias of estimators with missing data, *Biometrics*, **50** (1994) 1163–1170.

Rowe, V.K., Glycols, in F.A. Patty (Ed.) *Industrial Hygiene and Toxicology,* (Interscience, New York, 1963, Vol. 99, pp 1497–1536).

Royston, P. and Altman, D.G., Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling, *Applied Statistics*, **43** (1994) 429–467.

Ryan, L.M., Quantitative risk assessment for developmental toxicity, *Biometrics*, **48** (1992) 163–174.

Ryan, L.M. and Molenberghs, G., Statistical methods for developmental toxicity: analysis of clustered multivariate binary data, *Annals of the New York Academy of Sciences* (1999) In Press.

Santner, T.J. and Duffy, D.E., *The Statistical Analysis of Discrete Data,* (Springer-Verlag, New York, 1989).

Scientific Committee of the Food Safety Council, Proposed system for food safety assessment, *Food and Cosmetic Toxicology*, **16**, Supplement 2, 1–136. Revised report published June 1980 by the Food Safety Council, Washington, DC (1978, 1980).

Serfling, R.J., *Approximation Theorems of Mathematical Statistics*, (Wiley, New York, 1980).

Shiota, K., Chou, M.J. and Nishimura, H., Embryotoxic effects of di-2-ethylhexyl phthalate (DEHP) and di-*n*-butyl phthalate (DBP) in mice, *Environmental Research*, **22** (1980) 245–253.

Skellam, J.G. A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials, *Journal of the Royal Statistical Society, Series B*, **10** (1948) 257–261.

Stiratelli, R., Laird, N. and Ware, J.H., Random effects models for serial observations with binary response, *Biometrics*, **40**, (1984) 961–971.

Tyl, R.W., Price, C.J., Marr, M.C. and Kimmel, C.A., Developmental toxicity evaluation of dietary di(2-ethylhexyl)phthalate in Fischer 344 rats and CD-1 mice, *Fundamental and Applied Toxicology*, **10** (1988) 395–412.

Williams, D.A., The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity, *Biometrics*, **31** (1975) 949–952.

Williams, D.A., Reader Reaction: Estimation bias using the beta-binomial distribution in teratology, *Biometrics*, **44** (1988) 305–307.

Williams, P.L. and Ryan, L.M., Dose-response models for developmental toxicology, in R.D. Hood (Ed.) *Handbook of Developmental Toxicology,* (CRC Press, Boca Raton, 1997, pp 635–666).

Windholz, M., *The Merck Index: An Encyclopedia of Chemicals, Drugs, and Biologicals* (M. Windholz, Ed.), 10th ed., (Merck and Co., Rahway, NJ, 1983).

Zhu, Y., Krewski, D. and Ross, W.H., Dose-response models for correlated multinomial data from developmental toxicity studies, *Applied Statistics*, **43** (1994) 583–598.

Samenvatting

Er is in de maatschappij een grote bezorgdheid ontstaan over de effecten van verschillende soorten toxische blootstellingen op de voortplanting en ontwikkeling van de mens. Geneesmiddelen kunnen naast hun therapeutisch effect ook bijwerkingen vertonen. Additieven gebruikt in de voedingssector kunnen ook beschouwd worden als blootstellingen, net als materialen zoals ftalaten die gebruikt worden voor het verpakken van drank- en voedingswaren. In de scheikundige industrieën worden arbeiders blootgesteld aan oplosmiddelen en andere chemicaliën. Door de vervuiling van het leefmilieu kan de mens eveneens gevolgen ondervinden van scheikundige stoffen en stralingen.

Men stelt zich vragen over de relatie tussen deze blootstellingen enerzijds en reproductie- en ontwikkelingstoxicologie anderzijds. Meer in het bijzonder wenst men de effecten van chemicaliën en stralingen op de vruchtbaarheid van man en vrouw, de zwangerschap, het voorkomen van miskramen en geboortes van dode kinderen, de aanwezigheid van afwijkingen bij pasgeborenen en mogelijke, postnatale complicaties in de ontwikkeling te onderzoeken.

Er zijn verschillende strategieën om de implicaties van zo'n blootstellingen op reproductie en ontwikkeling van de mens te bestuderen. Epidemiologische gegevens kunnen in principe aangewend worden. Gezien deze data worden verzameld bij de mens, is er geen extrapolatie van de resultaten nodig. Echter, in deze context zijn betrouwbare epidemiologische gegevens nauwelijks of niet beschikbaar. Bijgevolg wordt er vaak geopteerd voor toxicologische experimenten op knaagdieren. Eén van de belangrijkste doelstellingen van deze dierenproeven is het bepalen van een "veilige dosis" van de onderzochte toxische verbindingen voor de mens. Alhoewel de extrapolatie van het proefdier naar de mens niet eenvoudig is, leiden deze studies tot vele interessante onderzoeksonderwerpen. Bovendien hebben experimenten op knaagdieren het voordeel dat er een goede controle mogelijk is over allerlei factoren die de resultaten kunnen beïnvloeden.

In dit proefschrift worden statistische technieken toegepast op gegevens die betrekking hebben op de gevolgen van de aanwezigheid van scheikundige stoffen op de ontwikkeling van foetussen. Het Amerikaanse "Research Triangle Institute" heeft dergelijke toxicologische studies uitgevoerd bij muizen en ratten. Zo'n toxicologisch experiment bevat over het algemeen een controle-groep en drie of vier groepen waarbij de zwangere dieren van een gegeven groep worden blootgesteld aan een bepaalde dosis van een chemische verbinding. Meestal worden 20 tot 30 moederdieren bij toeval toegewezen aan een dosis-groep. Deze dieren worden blootgesteld aan die toxische stof gedurende de kritische periode van de dracht. Net voor het baren worden de dieren gedissecteerd en wordt de baarmoeder grondig onderzocht. Er wordt nagegaan of de foetussen levensvatbaar zijn. De levensvatbare foetussen worden gewogen en de aanwezigheid van verschillende types van afwijkingen wordt geregistreerd. Meer specifiek beschouwt de toxicoloog mogelijke afwijkingen betreffende het geraamte en betreffende de ingewanden, naast uitwendige afwijkingen.

Verschillende soorten van gegevens worden dus verzameld in deze toxicologische studies. Van ieder moederdier wordt het aantal inplantingen en de toegediende dosis geregistreerd. Verder wordt de levensvatbaarheid van de foetus genoteerd, net als het gewicht en het al dan niet voorkomen van meerdere types van afwijkingen. Meestal zijn de metingen over deze afwijkingen binair. Behalve de drie reeds vermelde types afwijkingen wordt in dit proefschrift ook een binaire variabele geanalyseerd die aangeeft of een foetus geen enkele soort afwijking vertoont.

De proefdieren in de hier beschouwde toxicologische experimenten zijn muizen. De gegevens van de volgende toxische verbindingen worden geanalyseerd: ethyleen glycol, triethyleen glycol dimethyl ether, diethyleen glycol dimethyl ether, di(2-ethylhexyl)ftalaat en theophylline. De eerste drie chemicaliën worden o.a. gebruikt als oplosmiddel. Di(2-ethylhexyl)ftalaat wordt toegepast bij de productie van voorwerpen bestaande uit polyvinylchloride, waarbij deze toxische verbinding bijdraagt tot de gewenste flexibiliteit van deze voorwerpen. Theophylline ten slotte is een geneesmiddel voor de behandeling van astma tijdens de zwangerschap.

In dit proefschrift staan statistische procedures in het gebied van risico-analyse centraal. Enerzijds wordt de relatie tussen dosis en respons (het aantal dode foetussen in de baarmoeder, het risico op een afwijkende foetus,...) bestudeerd. Anderzijds wordt in het gedeelte over kwantitatieve risico-analyse onderzocht hoe men een veilig niveau van blootstelling aan een bepaalde toxische stof kan schatten. Deze

analyse kan men baseren op de "No Observed Adverse Effect Level" (NOAEL) benadering. Gezien de vele nadelen van deze aanpak, wordt hier verkozen om de kwantitatieve risico-analyse te baseren op de dosis-respons-modellering. In tegenstelling tot de NOAEL benadering laat deze aanpak toe om een maat voor de variabiliteit van de geschatte veilige dosis te bepalen, alsook om de hiërarchische structuur van een toxicologisch experiment in de analyse op te nemen.

Kwantitatieve risico-analyse gesteund op dosis-respons-modellering leidt tot een aantal algemene, relevante onderzoeksonderwerpen. Vooreerst dient men bij de statistische analyse van gegevens uit toxicologische studies, rekening te houden met de genetische verwantschap van foetussen uit eenzelfde nest en de gelijkaardige omstandigheden voor die foetussen in de baarmoeder. Daardoor zijn de data van foetussen uit eenzelfde nest over het algemeen gecorreleerd. Modellen die het complexe mechanisme waaruit de data worden gegenereerd, trachten te benaderen, dienen rekening te houden met dit zogenaamde nest-effect. Verder dient men na te gaan hoe de hiërarchische structuur met dode foetussen enerzijds en levensvatbare maar afwijkende foetussen anderzijds, kan geanalyseerd worden.

In dit onderzoeksdomein kunnen er meerdere, specifieke deelaspecten worden onderscheiden: het beschrijven van de dosis-respons-relatie, het schatten van de dosis-effect-parameter(s), het toetsen van de nulhypothese dat er geen dosis-effect is, het bestuderen van de gevolgen van het foutief specificeren van het model op het dosis-effect en op de veilige dosis,...

Drie types van gegevens worden in dit proefschrift beschouwd. Vooreerst worden de data geanalyseerd van experimenten uitgevoerd door het Research Triangle Institute. Verder wordt een simulatiestudie opgezet met steekproeven van dezelfde grootte als in typische, toxicologische studies. Ten slotte wordt een asymptotische studie uitgevoerd om de effecten van het foutief specificeren van het model te onderzoeken in geval van zeer grote steekproeven. In beide types van simulatiestudies worden data gegenereerd uit een bepaald model, waarna hetzelfde model en andere (foutief gespecificeerde) modellen aangepast worden aan deze gegevens. Op die manier kan men de implicaties van het verkeerd specificeren van het model op de resultaten nagaan.

In Hoofdstuk 2 wordt o.a. vermeld welke vereenvoudigingen in dit proefschrift worden doorgevoerd. Niettegenstaande ook continue responsen zoals het gewicht van een foetus meestal worden beschouwd in dit soort van experimenten, ligt het accent

hier op statistische technieken voor de analyse van binaire metingen. Meer specifiek legt dit proefschrift de klemtoon op het verwerken van binaire data aan de hand van likelihood procedures. Verder wordt verondersteld dat de gegevens van foetussen uit eenzelfde nest uitwisselbaar zijn, d.w.z. dat de marginale kans op een afwijkende foetus dezelfde is voor ieder diertje uit dat nest en dat de associatie tussen elke twee foetussen van eenzelfde moeder gelijk is. Uitwisselbaarheid is in deze context een natuurlijke veronderstelling. Een fundamentele vraag in dosis-respons-modellering is welk model er dient gebruikt te worden bij de analyse van deze gecorreleerde gegevens. Verschillende klassen van modellen komen in aanmerking: marginale en conditionele modellen, naast modellen met toevallige effecten. Een overzicht van deze categorieën van modellen wordt gegeven. Ten slotte worden het Bahadur model (marginaal), het George-Bowman model (eveneens marginaal), het beta-binomiaal model (een model met toevallige effecten) en een conditioneel model geïntroduceerd. Deze modellen worden gebruikt bij de studie van de verschillende deelaspecten.

In Hoofdstuk 3 worden de vermelde modellen vergeleken betreffende de gevolgen van het foutief specificeren van het model op het dosis-effect en op de toets van de nulhypothese dat er geen dosis-effect is. Het gedrag van de likelihood ratio en de Wald toetsstatistieken wordt hier bestudeerd. Eén van de conclusies is dat zowel het beta-binomiale model als het conditionele model een aanvaardbaar gedrag vertonen betreffende het toetsen van de beschouwde nulhypothese. Het conditionele model heeft duidelijke, numerieke voordelen, terwijl de parameters van het beta-binomiale model een eenvoudige, marginale interpretatie hebben.

In Hoofdstuk 4 wordt bijzondere aandacht gegeven aan het gedrag van de likelihood ratio toetsstatistiek wanneer een Bahadur model wordt aangepast aan de data van deze experimenten. In het algemene Bahadur model worden in de gezamenlijke verdeling naast marginale kansen en tweede orde correlaties, ook hogere orde associatie-parameters opgenomen. Meestal worden de derde en hogere orde correlaties weggelaten. Men merkt op dat bij sterke dosis-effecten, deze vereenvoudiging leidt tot opvallend toegenomen waarden van de likelihood ratio toetsstatistiek in vergelijking met dezelfde statistiek bij het beta-binomiale model. Het opnemen van een derde orde associatie-parameter in het Bahadur model verandert het gedrag van deze statistiek nauwelijks. Echter, in het vier-weg Bahadur model (met tweede, derde en vierde orde correlaties), observeert men aanzienlijk lagere waarden van de likelihood ratio toetsstatistiek die beter vergelijkbaar zijn met de waarden van

diezelfde statistiek bij het beta-binomiale model. Dit fenomeen is gerelateerd met de beperkingen op de parameters van dit model. Terwijl het toevoegen van een derde orde correlatie-coëfficiënt deze restricties nauwelijks opheft, wordt getoond hoe een vier-weg Bahadur model leidt tot een duidelijke toename van de parameterruimte.

In kwantitatieve risico-analyse ligt het accent op het schatten van een veilige dosis, welke kan worden gedefinieerd als de dosis waarbij het extra risico op een bijwerking bovenop het achtergrond-risico, gelijk is aan een bepaalde kans, b.v. 10^{-4}. Behalve de puntschatting van deze dosis worden twee methoden toegepast voor de bepaling van een benedengrens voor de veilige dosis. De ene methode is gebaseerd op de limiet-verdeling van de likelihood ratio toetsstatistiek, terwijl de andere verband houdt met de "profile likelihood". Verder worden zowel voor de punt- als de intervalschatting van de veilige dosis, twee procedures beschouwd. Een eerste procedure is volledig gebaseerd op het model. In de tweede procedure wordt eerst gedeeltelijk gebruik gemaakt van de dosis-respons-curve van het model en worden de bekomen resultaten lineair geëxtrapoleerd. In Hoofdstuk 5 wordt de risico-analyse beschouwd op het niveau van een foetus. De vermelde modellen worden vergeleken betreffende de gevolgen van het foutief specificeren van het model op de bepaling van de punt- en intervalschattingen van het veilig niveau van blootstelling.

Een belangrijke vraag in kwantitatieve risico-analyse is of de veilige doses moeten bepaald worden op het niveau van een foetus of op het niveau van een nest. In tegenstelling met Hoofdstuk 5 wordt in Hoofdstuk 6 een vergelijking gemaakt van kwantitatieve risico-analyse op beide niveaus. Er wordt getoond hoe de hiërarchische structuur van dode foetussen en levensvatbare maar afwijkende foetussen kan worden opgenomen in de risico-analyse. In dit hoofdstuk worden uitdrukkingen voor foetus- en nest-gebaseerde risico's opgesteld, zowel voor het beta-binomiale als voor het conditionele model. Telkens wordt een onderscheid gemaakt tussen het samenvoegen van de responsen "levensvatbaarheid van een foetus" en "aanwezigheid van afwijkingen bij een foetus" enerzijds en het analyseren van deze responsen via een hiërarchische structuur anderzijds. Veilige niveaus van blootstelling worden geschat voor elk soort risico. Er wordt ten slotte ook getoond hoe de schatting van de verdeling van nestgroottes een invloed heeft op de variantie van de risico-schatter.

In de vorige hoofdstukken werden lineaire predictoren voor de natuurlijke parameters (of een link-functie van die parameters) van het model gekozen. Men kan zich de vraag stellen of meer complexe predictoren in deze context dienen beschouwd te

worden. Het laatste hoofdstuk bestudeert een specifiek type van niet-lineaire predictoren. In plaats van een lineaire functie van de toegediende dosis, wordt hier een machtsfunctie van die dosis beschouwd. De nulhypothese dat er geen dosis-effect is, is equivalent met het nul stellen van het product van twee regressie-parameters. Onder die nulhypothese zijn de regressie-parameters dus niet identificeerbaar. Dit resulteert in problemen indien voor een frequentistische aanpak wordt gekozen. Echter, in dit hoofdstuk wordt geïllustreerd hoe een Bayesiaanse benadering toelaat om aan de hand van Bayes-factoren deze nulhypothese te toetsen.