PARAMETER ESTIMATION IN A TWO-STAGE MODEL OF CARCINOGENESIS

by

SHIH-CHIA CHANG LIU

Department of Biostatistics
University of North Carolina

Institute of Statistics
Mimeo Series No. 2121T

September 1993

PARAMETER ESTIMATION IN A TWO-STAGE MODEL OF CARCINOGENESIS

by

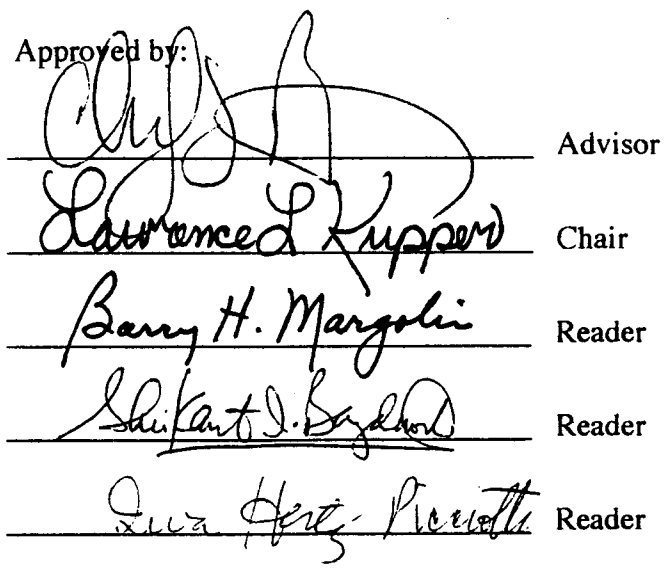
SHIH-CHIA CHANG LIU

A Dissertation submitted to the faculty of The University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics.

Chapel Hill

1993

Approved by:



Advisor

Chair

Reader

Reader

Reader

SHIH-CHIA CHANG LIU. Parameter estimation in a two-stage model of carcinogenesis (Under the direction of Christopher J. Portier)

ABSTRACT

The distribution of the number of pre-malignant lesions (PMLs), is assumed to be Poisson in the two-stage model of carcinogenesis. In the computation of almost all quantities, the Poisson assumption is the most fundamental. Therefore, the validity of this assumption is examined, and corrections are made where necessary.

In analyzing the number of the PMLs from several studies, we observed severe over-dispersion. To modify the Poisson assumption on the distribution of the number of PML's, two major approaches using classical Bayesian and quasi-likelihood methods, are taken. We set prior distributions, namely gamma and uniform, on the mutation rate and then computed the marginal distribution of the number of PML's. The quasi-likelihood method is also used assuming a quadratic relationship between the mean and variance of the number of PML's instead of assuming the structure of the distribution. So we have four models: the Poisson model, the Poisson-uniform model, the Poisson-gamma (or negative binomial) model, and the quasi-likelihood model.

The above four models are applied both to animal data and simulated data. We found that the negative binomial model, i.e. when a gamma prior is set on the mutation rate, performs the best in terms of parameter estimation and model fitting.

Table of Contents

Page

LIST OF TABLES v

LIST OF FIGURES vii

REFERENCES 137

Chapter

I. Introduction and literature review 1
- 1.1 Long-term carcinogenicity studies 1
- 1.2 Initiation-Promotion (IP) Studies 4
  - 1.2.1 History and Background 4
  - 1.2.2 Statistical Analysis of I-P studies 8
- 1.3 Initiation-Promotion-Stop promotion (IPS) Studies 9
- 1.4 Mathematical modelling of carcinogenesis 10
  - 1.4.1 The one-hit, one-stage model of Iversen and Arley 10
  - 1.4.2 The multi-hit, multi-stage theory of Stocks and Nordling 11
  - 1.4.3 The one-hit, multi-stage theory of Armitage and Doll 12
  - 1.4.4 The one-hit, two-stage and one-hit, three-stage theory of Neyman and Scott 13
  - 1.4.5 The two stage stochastic model by Moolgavkar, Venzon, and Knudson 14
  - 1.4.6 The damage-fixation model by Portier and Kopp 19
- 1.5 Goal of research 22

II. Statistical analysis of over-dispersed data 23
- 2.1 Index of dispersion 23
- 2.2 Poisson regression 32
- 2.3 Mean-variance plot 33

III. Classical Bayesian Estimation in two-stage models 35
- 3.1 Prior information 35
- 3.2 A gamma prior on the mutation rate 37
- 3.3 A uniform prior on the mutation rate 39

3.4 Application of the Bayesian methods to animal data..... 45

IV. Empirical methods for over-dispersion..... 50
4.1 Introduction to the quasi-likelihood method..... 50
4.2 Definition of the quasi-likelihood method..... 50
4.2.1 Covariance function..... 50
4.2.2 Quasi-likelihood function..... 51
4.2.3 Parameter estimation..... 54
4.3 Application of the quasi-likelihood methods to animal data..... 55

V. Simulation Study..... 58
5.1 Introduction to simulation..... 58
5.2 Statistical model..... 59
5.2.1 Poisson model..... 60
5.2.2 Negative binomial model..... 62
5.2.3 Poisson-uniform model..... 62
5.3 Description of simulation study..... 63
5.3.1 Random variates generation..... 63
5.3.1.1 Gamma random variates..... 63
5.3.1.2 Poisson random variates..... 64
5.3.2 Experimental design..... 64
5.3.3 Sample size determination..... 66
5.3.4 Choice of maximization method..... 68
5.3.5 Method of Moments..... 68
5.3.5.1 $\mu X \sim Gamma(\gamma,\delta)$..... 68
5.3.5.2 $\mu X \sim Uniform(a,b)$..... 70
5.3.5.3 μX is a constant, P(N) is Poisson..... 71
5.4 Simulation Study Results..... 72
5.4.1 Summary Statistics..... 72
5.4.2 Interpretations..... 85

VI. Directions..... 93
6.1 Recommendations..... 93
6.2 Size of the premalignant lesions (PML's) in an IP protocol..... 96
6.3 Future research direction..... 101

LIST OF TABLES

Table 1.1 Long-term carcinogenicity studies..... 2
1.2 Characters of initiators and promoters..... 5
2.1 Mean and variance of number of NNM rat liver foci: all dose groups 24
2.2 The correlation between the number of PML's and length of exposure for the NNM rat liver foci data 24
2.3 Mean and variance of number of rat liver foci: NNM=0 ppm..... 27
2.4 Mean and variance of number of rat liver foci: NNM=0.1 ppm..... 27
2.5 Mean and variance of number of rat liver foci: NNM=1 ppm..... 28
2.6 Mean and variance of number of rat liver foci: NNM=5 ppm..... 28
2.7 Mean and variance of number of rat liver foci: NNM=10 ppm..... 29
2.8 Mean and variance of number of rat liver foci: NNM=20 ppm..... 29
2.9 Mean and variance of number of rat liver foci: NNM=40 ppm..... 29
2.10 Mouse skin papilloma data (Index of dispersion test)..... 30
2.11 TCDD-rat liver foci-DEN data (Index of dispersion test)..... 31
2.12 TCDD-rat liver foci-saline data (Index of dispersion test..... 31
2.13 Poisson regression analysis of NNM rat liver foci data..... 32
2.14 Poisson regression analysis of TCDD rat liver foci data..... 33
3.1 Mean-variance relationship of different priors on μX assuming $N|\mu X$ is Poisson 43
3.2 NNM rat liver foci data fitting Negative binomial model..... 47
3.3 NNM rat liver foci data fitting Poisson-uniform model..... 47
3.4 NNM rat liver foci data fitting Poisson model..... 48
3.5 Dioxin rat liver foci data fitting Negative binomial model..... 48
3.6 Dioxin rat liver foci data fitting Poisson-uniform model..... 49
3.7 Dioxin rat liver foci data fitting Poisson model..... 49
4.1 Mean-variance relationships in quasi-likelihood methods..... 53
4.2 NNM rat liver foci data fitting the quasi-likelihood method..... 56
4.3 Dioxin rat liver foci data fitting the quasi-likelihood method..... 56
5.1 Assumptions of different models..... 60
5.2 Birth rate / mutation rate combination design..... 64
5.3 Experimental design for the simulation study..... 65
5.4 Sample size needed versus percentage of difference allowed..... 67
5.5 Poisson data fitting Poisson model..... 73
5.6 Negative binomial data fitting Poisson model..... 74

5.7 Poisson-uniform data fitting Poisson model..... 75
5.8 Poisson-uniform data fitting Poisson-uniform model..... 76
5.9 Negative binomial data fitting Poisson-uniform model..... 77
5.10 Poisson data fitting Poisson-uniform model..... 78
5.11 Negative binomial data fitting Negative binomial model 79
5.12 Poisson-uniform data fitting Negative binomial model..... 80
5.13 Poisson data fitting Negative binomial model..... 81
5.14 Poisson data fitting Quasi-likelihood method..... 82
5.15 Poisson-uniform data fitting Quasi-likelihood method..... 83
5.16 Negative binomial data fitting Quasi-likelihood method..... 84
5.17 Percentage bias when the simulated data is Poisson..... 90
5.18 Percentage bias when the simulated data is Poisson-uniform..... 91
5.19 Percentage bias when the simulated data is negative binomial..... 92
6.1 Type I and Type II errors..... 96
6.2 Priors on the birth rate, the mutation rate, or both..... 101
6.3 Correlation between the size and number of PML's of NNM data.. 103

LIST OF FIGURES

Figure 1.1	Solt and Farber's assay procedure	105
1.2	Modern IP protocol for mouse skin	106
1.3	Modern IP protocol for rat liver	107
1.4	The design of Initiation-Promotion Studies	108
1.5	The one-hit, one-stage model of Iversen and Arley	109
1.6	The multi-cell theory of Fisher and Holloman	109
1.7	The multi-stage theory of Stocks and Nordling	110
1.8.1	The multistage theory of carcinogenesis proposed by Armitage and Doll.	111
1.8.2	The two-stage theory of carcinogenesis proposed by Armitage and Doll	111
1.9	The one-hit,two stage model of Neyman and Scott	112
1.10	The one-hit, three-stage model of Neyman and Scott	112
1.11	The two stage stochastic model of carcinogenesis of Moolgavkar, Venzon, and Knudson	113
1.12	The damage-fixation-multistage model of Portier and Kopp	114
2.1	Number of rat liver foci vs. length of exposure, NNM=0 ppm	115
2.2	Number of rat liver foci vs. length of exposure, NNM=0.1 ppm	116
2.3	Number of rat liver foci vs. length of exposure, NNM=1 ppm	117
2.4	Number of rat liver foci vs. length of exposure, NNM=5 ppm	118
2.5	Number of rat liver foci vs. length of exposure, NNM=10 ppm	119
2.6	Number of rat liver foci vs. length of exposure, NNM=20 ppm	120
2.7	Number of rat liver foci vs. length of exposure, NNM=40 ppm	121
2.8	Variance versus mean plot for NNM rat liver foci data	122
2.9	Variance versus mean plot for mouse skin papilloma data	123
2.10	Variance versus mean plot for TCDD-liver foci-DEN data	124
3.1	Gamma distribution with shape parameter=1	125
3.2	Gamma distribution with shape parameter=3	126
3.3	Gamma distribution with shape parameter=9	127
3.4	Gamma distribution with shape parameter=20	128
3.5	NNM rat liver foci data fitting Negative binomial model: residual plot	129
3.6	NNM rat liver foci data fitting Poisson-uniform model: residual plot	130
3.7	NNM rat liver foci data fitting Poisson model: residual plot	131
3.8	Dioxin rat liver foci data fitting Negative binomial model: residual plot	132
3.9	Dioxin rat liver foci data fitting Poisson-uniform model: residual plot	133
3.10	Dioxin rat liver foci data fitting Poisson model: residual plot	134
4.1	NNM rat liver foci data fitting the quasi-likelihood: residual plot	135

4.2 Dioxin rat liver foci data fitting the quasi-likelihood: residual plot.. 136
6.1 Flowchart of the decision process in the analysis of PML data..... 95
6.2 Initiation-Promotion Protocol..... 97

Chapter 1

1.1 Long-term carcinogenicity studies

Thirty years ago, the geneticist Rachel Carson issued a warning about the effects of chemicals on mankind in her famous book Silent Spring :" If you want living things to adapt to these chemicals, it will take...several generations. But laboratories are continually manufacturing new chemicals, and people and animals must continually adapt. This far exceeds the rate of adaptability of mankind." Though it is almost impossible to stop the manufacture of new chemicals, there are ways that laboratories can reduce the toxic effects of chemicals on humans. Animal carcinogenicity studies are one important way.

Table 1.1 shows how long-term carcinogenicity studies have evolved. Initially, most carcinogenicity studies were based upon exposure to one compound. The corresponding designs have changed from using a single dose of a chemical to multiple doses, and the time of observation has changed from the observation of immediate death to two weeks or even longer after the agents were applied. Gradually, the design of animal studies has come to involve continuous application of a given chemical by various methods: inhalation, ingestion, skin painting, intubation and intraperitoneal injection. Of these methods, the first three involve daily application; while the last two are generally applied once every two weeks.

In 1983, the National Toxicology Program (Haseman,1983) redefined the standard animal bioassay guidelines first proposed by the National Cancer Institute (Sontag et al., 1976). Use of both male and female animals of at least two species was specified by this

Design	Outcome	Time of observation
Single chemical:Single dose	Death	Immediate
Single dose	Death	Two weeks
Multiple doses	Death	Longer
Continuous	Other (cancer)	Lifetime (2 years)
Multiple chemicals (simultaneously)	Cancer	Lifetime

Table 1.1 Long-term carcinogenicity studies

new guideline. Rodents such as mice, rats, and Syrian hamsters are most commonly used due to their small size, short life span, and ready availability. Dose levels are determined by subchronic toxicity studies which define for each species and sex the maximum tolerated dose (MTD), i.e. the highest dose which can be applied that does not cause a significant decrease in survival from effects other than carcinogenicity . If the purpose of the study is the measurement of the carcinogenic potential of an agent on a qualitative basis as in a screening assay, only three experimental groups are required: high dose (MTD), low dose (1/2 or 1/4 MTD), and control (untreated). If the purpose is determination of a possible dose-response relationship, then three dose levels plus an untreated control are used. A minimum of 50 animals of each sex, species, and dose level for testing each potential carcinogen is recommended in most bioassay designs. In addition, the route of administration is chosen to mimic the route of exposure in humans. To randomize animals to experimental groups, animals are first stratified by factors which may affect tendency to tumor development, such as body weight, and then randomly allocated to dose level within strata.

A few weeks after weaning, treatment is initiated. It continues for about two years, which constitutes 60-70% of the life span of mice and rats (Portier et al., 1986). Animals are examined periodically for evidence of tumors or any other toxic effects. For any animal found dead, a complete necropsy is performed including 40 tissue sites such as liver, lung, stomach, brain, and skeleton. At the end of the treatment period, all surviving animals are sacrificed and subjected to a complete necropsy and histopathologic examination. Interim sacrifices, or planned kills prior to study termination, are done occasionally to get more information on tumor progression. Since these sacrifices reduce the number of animals surviving to the end of the studies, more animals should be included in an animal bioassay design with interim sacrifices (Feron et al. 1980).

In later experimental designs, multiple chemicals were used simultaneously on the same animal. This is partly due to the elucidation of a synergistic effect among cigarette smokers. It was observed that asbestos workers who smoked had about 92 times the risk of dying of bronchogenic carcinoma as men who neither work with asbestos nor smoke cigarettes. The above risk ratio was much larger than the one compared to people who work with asbestos but do not smoke or people who smoke but do not work with asbestos (Selikoff, Hammond & Churg,1968). It was also noted (Rothman and Keller,1972) that alcohol consumption and smoking produced a multiplicative effect. These two examples demonstrate the phenomenon of synergism, and the discovery of synergistic effects led to the development of animal cancer studies using more than 1 chemical rather than examining the effects of a single chemical. This helped to direct observations on cancer-related data.

However, long-term animal studies have certain inherent disadvantages. First, the time period of such studies is too long. They involve one year to set up, two years to run, and another two years to analyze data, making a total of five years to finish an animal experiment. Second, long term studies are expensive; it is not uncommon to spend US $ 1~2 million to conduct one long-term animal study (Haseman, 1990)!

1.2 Initiation-Promotion Studies

1.2.1 History and background

Carcinogens which cause cancer at the point of application in animals are direct-acting carcinogens. Carcinogens which do not act at the point of application but often affect specific tissues are procarcinogens. The direct-acting carcinogens or procarcinogens are also called initiators. Products that have little or no carcinogenic activity themselves but

Initiator	Promoter
1.Affect initiation:	1.Stimulate and expand the population
increase mutation rate	of initiated cells
2.Bind to cellular DNA	2.Bind to cell surface or
(genetic damage)	intercellular receptors
3.Cumulative	3.Not cumulative when the administration
	interval is long

Table 1.2 Characteristics of initiators and promoters

that considerably augment the effect of carcinogenicity are called promoters. Experimental promoters have been useful in studies of the mechanism of action of all types of carcinogens (Weisburger,1975). Table 1.2 illustrates some of the characteristics of initiators and promoters:

It was shown beginning in the late 1930s that cancer induction in the skin of rabbits or mice with chemicals, especially with small doses, requires two separate external manipulations (Farber,1984 a). From these studies evolved the paradigm of "two-stage carcinogenesis", initiation and promotion. Initiation is the process of a normal cell mutating to an intermediate or initiated cell, while promotion is the process of an initiated cell proliferating to more initiated cells and thus increasing the full rate of transformation from initiated cells to malignant cells.

Solt and Farber (1976) developed an approach which offered the first quantitative assay for populations of initiated cells, as well as the possibility of a sequential analysis of some of the early premalignant hepatocytes. They assumed that an early change induced by a carcinogen might be the induction of an altered metabolic pattern in some liver cells, such that these cells would have a potential growth advantage over the majority of original hepatocytes in a cytotoxic environment. They presumed that it should be possible to detect the altered cells soon after initiation by creating an intense stimulus for proliferation in the presence of a toxic growth inhibitor requiring metabolic activation. Initiated cells with a decreased capacity for activation might proliferate rapidly. Based on this theory, they derived an assay system for such resistant cells (Fig. 1.1).

The system consists of three components: an initiator, diethylnitrosamine (DEN), a selective growth inhibitor, 2-acetylaminofluorene (2-AAF), and a generalized potent growth stimulus, in this case, partial hepatectomy (PH). DEN was given intraperitoneally

to Fischer-344 rats at a dose of 200 mg/kg body wt. Following a 2-week period of recovery from the initial cell damage, the animals were fed a standard diet (24% protein) containing 0.02% 2-AAF for 1 week and were then subjected to 67% partial hepatectomy.

The results indicated that animals pretreated with DEN, fed with 2-AAF, and subjected to PH (group a) produced multiple focal islands of proliferating intensely basophilic hepatocytes within about 30 hours after operation. They grew rapidly, so that within 7-10 days they were visible as small, grayish white nodules about 1 mm in diameter. No such foci were seen in animals treated with saline in place of DEN (group c), given 2-AAF without PH (group d) or subjected to PH without concomitant exposure to 2-AAF (group b).

Modern IP experiments for the rat liver, however, are slightly different from the one stated above. The procedure described by Schere and Emmelot (1975) and carried out in the laboratory of Pitot et al.(1978) is currently used (Fig. 1.2). Animals are partially hepatectomized at time 0 followed within 24 hours by a single gastric instillation of an initiator. The animals are then allowed to remain on a normal basal diet for two months. At the end of this time one set of animals is placed on a diet containing a promoter (e.g. phenobarbitol,PB), whereas the other is fed the same diet but no promoter. At the end of 6 months the animals and their livers are examined grossly and morphologically as well as for the histochemical presence of three different enzymes. It is clear that PH and DEN (or any other appropriate carcinogen) combined serve as initiators, and PB (or any other appropriate cocarcinogen) serves as the promoter, while PH serves as the promoter in Solt and Farbers' (1976) research.

For the mouse skin IP experiment, 2 cm^2 (Kopp-Schneider et al., 1991) of the back skin of mice are shaved around 7 weeks of age and painted with an initiator (e.g. benzopyrene) three days later (Chouroulinkov et al., 1989). Soon after, they are exposed to promoters (e.g. 12-O-tetradecanoylphorbol-13-acetate or otherwise known as TPA).

The mice in the control group are either given the promoter alone without initiator or given the initiator alone without promoter. Each animal is individually examined and recorded for papilloma appearance or regression, as well as the induction of enzyme (e.g. ornithine decarboxylase or otherwise known as ODC) activity. The classic outline for the two-stage mouse skin carcinogenesis experiment is seen schematically in Fig. 1.3 (Pitot et al, 1978).

1.2.2 Statistical Analyses of Initiation-Promotion Studies

1.2.2.1 Design

The main study group for an IP study is a single exposure to an initiator followed by continuous applications of promoters as shown in Fig. 1.4-a. A variety of other experimental groups used as controls are also illustrated in Fig.1.4: such as a reverse order of initiator and promoter, a lifetime application of the initiator with no promoter, a lifetime application of the promoter with no initiator, or a single application of the initiator in the beginning but no promoters, and/or a control group with no initiator or promoter at all.

1.2.2.2 Data collection

The data collected in the studies basically belong to two categories, i.e. tissue level or cellular level. At the tissue level, number, size, and age at onset of pre-malignant lesions, PMLs (e.g. papillomas in mouse skin and altered foci in rat liver), or tumors are recorded. At the cellular level, types of alterations of cells, cell replication, cell death, and/or other biological changes are sometimes observed. In contrast to the usual long-term carcinogenicity study, much of these data are restricted to a single target tissue such as the liver.

1.2.2.3 Hypothesis testing for screening

Different statistical analyses are performed for different types of data. If the data consist of the age at tumor onset, time-to-tumor analyses as described in the literature review of Williams (1990) can be used. If the data consist of the size of a tumor or PML, general linear models (ANOVA) or growth curve methods can be applied. If the data consist of number of tumors or PMLs, Poisson regression will be an adequate tool.

1.3 Initiation-Promotion-Stop promotion (IPS) studies

An important question concerning the genesis and development of premalignant lesions (PMLs) is the stability of these lesions in the presence of and, especially, following the withdrawal of the promoter. This question is important in both prevention of carcinogenesis through careful public health controls on promoters and on the use of chemotherapeutics, many of which have secondary carcinogenic effects which are thought to be the result of promotional mechanisms. In some biological models, PMLs regress following the removal of the proliferative stimulus. For the mouse skin experiment, it is found (Stenback et al., 1974) that if promoters are removed, the kinetics revert to normal, and the papillomas disappear. A decrease in the number of papilloma would thus be a direct measure of loss of promotion. For the rat liver experiment, altered hepatic foci (AHF) and nodules are observed (Farber, 1984 b) to decline after stop-dosing of the promoter. Yet in other models (Goldsworthy et al., 1986; Scherer et al., 1975), the number of AHFs remain stable. The kinetics of the system is obviously changed after the removal of the promoter, so IPS experiments should be performed to help elucidate the mechanism of carcinogenesis.

A protocol has been devised which we will refer to as the Initiation-Promotion-Stop promotion (IPS) experiment . Rats are partially hepatectomized at time 0 followed by a

single intragastric dose of an initiator (DEN) after 20 hours. One week later, rats are fed a diet containing a 0.05% PB for 6 months. At this time, half of the rats are killed; the other half are withdrawn from PB for 10 days before sacrifice (Glauert et al. 1986). Then the liver sections are fixed in formalin and processed for histological analysis. This is a rather novel experiment which has been used only rarely in the past. However, there are several pressing issues concerning the mechanism of carcinogenesis which may be addressed using the IPS experiment.

1.4 Mathematical modeling of carcinogenesis

From the above IP experimental protocols, it is obvious that IP experiments have certain advantages in that they are shorter (30-50 weeks), cheaper, and mechanistically based. One parametric manner in which to analyze these experiments is to use mechanistically-based mathematical models of carcinogenesis. A brief history and description of these models follow.

1.4.1 The one-hit, one stage model of Iversen and Arley

The earliest quantitative theory of carcinogenesis is that of Iversen and Arley (1952). Their model is based on a one-step transition process occurring in a single cell and is illustrated by Figure 1.5.

In this model, a normal cell (state N in Fig. 1.5) is transformed into a malignant cell (state M in Fig. 1.5) at a rate $\mu(d)$ assumed by Iversen and Arley to be a linear function of the level of the carcinogen (i.e., $\mu(d)=\mu_1+\mu_2 d$). The intercept of the function (μ_1) would then represent the background or spontaneous transition probability obtained in the absence of the carcinogen. After transition to a malignant cell, they assumed that the growth of a

clone is a pure birth process with a birth rate, α_1, independent of the concentration of the carcinogen. A clone is assumed to become a detectable tumor when it reaches a certain size. The above model is usually called the "one-hit, one stage" model and suggests that the expected number of malignant cells generated from a single normal cell depends only upon the total dose of carcinogen and not upon the pattern of exposure.

1.4.2 The multistage theory of Stocks and Nordling

Shimkin and Gubareff (1967) and White (1967) suggested that the subdivision of a single dose of urethane into a large number of small subdoses administered at short time intervals decreases the crop of tumors considerably, so the pattern of exposure is actually very important. Neyman and Scott (1967) also pointed out that the relation of tumor to dose is far from linear, but rather parabolic.

Around the time that Iversen and Arley proposed their model, several investigators found that the death rates for many forms of human cancer increased proportionately with the fifth or sixth power of age (Fisher and Holloman,1951; Nordling,1953; Stocks,1953; Armitage and Doll,1954) They assumed that the time from tumor detection until death was negligible compared to the time required for transformation and growth so they assumed that the mortality rate from cancer was the same as the incidence rate. They also assumed that there was exposure to a constant concentration of carcinogen from birth to death. Fisher and Holloman (1951) explained the power law relationship as six or seven different cells which have to be transformed in a single tissue in order to form a tumor, a concept known as the multicell theory. This multicell theory is illustrated in Figure 1.6. Unfortunately, this theory is in serious disagreement with the data (Whittemore and Keller, 1978; Doll, 1971).

But Stocks (1953) and Nordling (1953) proposed another explanation; namely, that a single cell has to undergo a certain number of changes or mutations, say k, before it can generate a tumor. This is called the multistage theory, and is illustrated in Fig. 1.7. Both Nordling and Stocks assumed that the transition probabilities from one state to the next were the same and were proportional to both time and the level of the carcinogen, i.e. $\mu_i(t,d)=\mu td$. Their model predicts that incidence of cancer is proportional to a high power of both dose and time. This did not agree with the results of human and animal data. To avoid the disagreement, Armitage and Doll (1954) modified the theory of Nordling and Stocks.

1.4.3 The multistage and two-stage models of Armitage and Doll

Armitage and Doll (1954) assumed that the transition probabilities between events were not all equal and that only some of the transitional events were dependent upon the carcinogen while the remainder had a spontaneous probability of transition independent of the level of the carcinogen. This model differs from the model of Stocks and Nordling in that it did not assume that the transition probabilities from one state to the next were the same. With this modification, the model then became consistent with both human and animal data that showed tumor incidence as being related to either dose or the square of dose, but not higher powers. The model is shown in Fig. 1.8.1

This multistage model adequately describes most of the existing data on adult tumor incidence. It has been tested by examining the age-specific mortality rates for 17 types of cancer, and generally speaking, the results accord well with the theory (Armitage and Doll, 1954).

The main biological defect of the Armitage and Doll model is the lack of any direct experimental evidence that cancer occurs in more than two stages. So Armitage and Doll

(1957) modified their theory to permit cells in intermediate stages to multiply more rapidly than normal cells, while in the original theory only the cells in the final stage were assumed to multiply.

Figure 1.8.2 depicts the two-stage model of Armitage and Doll. Cells in stage zero are normal, those in stage one generate clones which are assumed to grow exponentially at some rate k, and clones in stage two are tumors. The subject is exposed to two separate agents: the initiating agent which work on the normal cells only, and the promoting agent which work on the 'changed' or initiated cells only. The chance of inducing either change is directly proportional to the doses of the agents, so the resulting incidence of cancer is proportional to $d_2 n_t$. Here d_2 is the dose of the promoting agent, and n_t is the number of initiated cells present at a time t after exposure to the initiating agent. This n_t will be proportional to the dose of the initiating agent, d_1, and will also increase exponentially at a rate k with the passage of time, i.e., n_t will be proportional to $d_1 e^{kt}$. Hence the incidence of cancer will be proportional to $d_1 d_2 e^{kt}$.

The theory that cancer is induced in two stages can be extended to include the postulate that the first stage results in the production of a clone of cells which have a slight selective advantage over the surrounding and unaffected cells. With this added postulate, it is shown that the expected incidence of cancer at all ages between 25 and 74 years is similar to data on mortality from cancer of the stomach, intestines, rectum and pancreas in both sexes (Armitage and Doll, 1957).

1.4.4 The one-hit, two stage and the one-hit, three stage models of Neyman and Scott

Hyperplastic foci, which are clones of modified cells in the lungs of mice caused by injection of a carcinogen, were observed (Shimkin and Polissar, 1955) to increase up to

four weeks with time between injection and sacrifice. The number of tumors was also found to increase monotonically with time. Hence it was interesting to model the behavior of hypercellularity. Neyman and Scott (1967) proposed a one hit, two stage model of carcinogenesis which is illustrated by Fig. 1.9.

In the Neyman-Scott model, it is assumed that a hit on a normal cell turns this cell immediately into a first order mutant (which is 1° in Fig. 1.9) subject to three time independent risks: the risk of division at rate α; the risk of death at rate $\beta>\alpha$; and the risk of a second order mutation at rate μ. The mutation rate μ is assumed to be a linear function of $d \cdot f(t)$ where d is the dose and f(t) is a function of age, t. In addition, all cells and clones are assumed to be mutually independent. Apart from second order mutant (which is 2° in Fig. 1.9), all of the cells descending from the original first order mutants were assumed to have the same properties, forming a subcritical birth and death process "with emigration". The second order mutants were considered cancer cells, each generating a supercritical birth and death process. Unfortunately, this model was not consistent with the data of Shimkin and Polissar (Neyman and Scott, 1967).

To be consistent with this biological data, Neyman and Scott revised their model into a one-hit, three-stage model as shown in Fig. 1.10. The postulated distinction between primary and secondary first order mutants was suggested by the fact that mutation in irradiated flies manifest themselves, not in these same flies, but in their progeny and further descendants. This revised model, qualitatively at least, agreed with the changes in hypercellularity observed by Shimkin and Polissar (1955).

1.4.5 The two-stage stochastic model by Moolgavkar, Venzon, and Knudson

More recently, a two-stage stochastic model popularized by Moolgavkar, Venzon, and Knudson (Moolgavkar & Venzon, 1979; Moolgavkar & Knudson, 1981) allowing for

birth and death of normal and intermediate cells has been used to model human cancer incidence. At first glance, it seems that this model (Fig.1.11) differs from the model of Neyman and Scott (Fig.1.9) in that the latter did not allow for birth and death of normal cells. But in their computations (Moolgavkar & Venzon, 1979; Moolgavkar & Knudson, 1981), the birth and death rate of normal cell have not been used. Therefore, the distinction between these two models becomes vague.

This two-stage model embracing the concepts of "initiation", "promotion" and "progression" allows for the clonal expansion of cells, especially initiated (I) cells, which is confirmed in the work of Slaga, Pitot and others (Slaga et al., 1980; Pitot et al., 1978). In addition, this model seems to be capable of describing spontaneous and induced carcinogenesis in general (Moolgavkar et al, 1979,1981). The MVK model allows the cells to replicate, differentiate, or die through a linear birth and death process. Figure 1.11 illustrates the model.

The main assumptions are: In a small time interval Δt, a normal cell may

(1) divide into two normal cells with probability $\alpha_0\Delta t+o(\Delta t)$;

(2) divide into one normal and one intermediate cell with probability $\mu_1\Delta t+o(\Delta t)$;

(3) die(or differentiate) with probability $\beta_0\Delta t+o(\Delta t)$.

It is assumed that the probability of more than one event occurring in an interval Δt is $o(\Delta t)$ which is defined such that $\lim_{\Delta t\to 0} \frac{o(\Delta t)}{\Delta t} = 0$. The same assumptions are made about the intermediate cells, with parameters α_1, μ_2, and β_1, respectively.

It has been shown (Emmelot and Scherer, 1980; etc.) that, during the process of hepatocarcinogenesis (carcinogenesis of the liver), foci of altered hepatocytes precede the development of hepatocellular cancer. These foci, known as altered hepatic foci (AHF) or enzyme-altered foci, include cells which exhibit qualitatively altered enzyme activity (Pitot

et al., 1978). The appearance of enzyme-altered foci has been correlated with the later development of malignant cells (Kunz et al., 1983). These foci are thought to be clonally expanded colonies of initiated cells. With this data in mind, one would want to estimate the expected number of observable foci, the distribution function for the number of cells in a observable focus, and the distribution function for the radii of the sections of foci in two and in three dimensions. This has been done by Moolgavkar et al. (1990).

In order to utilize this model, it must be assumed that the transformations from both normal cells to intermediate cells and intermediate cells to malignant cells are stochastic. Let X(s) be the number of normal susceptible cells in the liver at age s. The number of initiated cells that arise from the normal cells by age t is assumed to be a Poisson random variable with expectation $\int_0^t \mu(s)X(s)ds$, where $\mu(s)$ is the rate of initiation of a normal cell. The number of non-extinct clones of intermediate cells at age t has a Poisson distribution with mean $\Lambda(t)$ given by

$$\Lambda(t)=\int_0^t \mu(s)X(s)[\ 1-p(t;s)]ds \tag{1.1}$$

where p(t;s) is the probability that a clone of intermediate cells that formed at time s, together with all its daughters, is extinct by time t. Let W(t) represent the number of cells in a randomly chosen (non-extinct) clone at time t. For any q=1,2,....,

$$P[W(t)=q]=\frac{1}{\Lambda(t)}\int_o^t \mu(s)x(s)[\ \frac{\alpha}{\beta}p(t;s)][1-\frac{\alpha}{\beta}p(t;s)]^{q-1}ds \tag{1.2}$$

(I) Translation from three-dimensions into two-dimensions:

To examine the foci, it is impossible for researchers to examine an entire population of rat hepatocytes (liver cells) under the microscope. Instead, a cross-sectional slice of rat liver is taken with a width of one or two hepatocytes. Although the focus is itself three-dimensional, what is observed is only two-dimensional. Thus, it is necessary to translate the model-derived three-dimensional quantities into two-dimensional quantities by using

standard stereological procedures. Moolgavkar et.al. used the formula attributed to Wicksell (1925). Let $G_2(y)$ be the cumulative distribution function for Y, which is the radii of the sections of foci in two dimensions, and let $g_3(r)$ be the density function for the radii in 3-dimensions, then

$$(1.3) \qquad G_2(y)=1 - \int \frac{1}{\tilde{r}_3}(r^2-y^2)^{1/2} g_3(r)dr,$$

where $\tilde{r}_3 = \int_0^\infty r g_3(r)dr$ is the mean radius in three dimensions.

(II) Distribution of number of observable sections of foci:

Not all the foci on the slice are observable, but only those above a certain critical size. So let $N_2(t)$ represent the number of observable sections in two dimensions. Then $N_2(t)$ is Poisson with parameter $\lambda_2(t)$ given by

$$(1.4) \qquad \lambda_2(t)= [1-G_2(\varepsilon)]2\tilde{r}_3 \Lambda(t)\}$$

The distribution function of the radii in two dimensions conditional on their being larger than ε is given by (Nychka et. al., 1984) as $G_2^{\varepsilon}(y)=1- \frac{1}{\mu_\varepsilon}\int_y^\infty (r^2-y^2)^{1/2} g_3(r)dr$, where $\mu_\varepsilon=\int_\varepsilon^\infty (r^2- \varepsilon^2)^{1/2} g_3(r)dr$. Thus the density function of the radii in two dimensions conditional on their being larger than ε is

$$(1.5) \qquad g_2^{\varepsilon}(y)=\frac{y}{\mu_\varepsilon} \int_y^\infty (r^2-y^2)^{-1/2} g_3(r)dr$$

(III) Assumptions, expected number of observable foci, and distribution of cells in an observable focus:

To fit the above model to a data set, Moolgavkar et al (1990) assume that exposure to a single agent begins at a fixed age and continues at the same level until death. So based on the above assumption, the birth rate, death rate, and mutation rate may be functions of dose but are independent of time. They also assume that the data will be presented in terms of

numbers of sections per unit area (cm^2). Besides, X(s), which is the number of liver cells per cubic centimeters (instead of total number of hepatocytes), is assumed to be a constant.

With the assumptions above, it follows from equation (5.2) in Dewanji et al. (1989) that expression (1.4) for $\Lambda(t)$ becomes

$$(1.6) \qquad \Lambda(t) = \frac{\mu X}{\alpha}\left[\ln\left(\frac{\beta}{\beta-\alpha p(t)}\right)\right] ,$$

where $$p(t) = \frac{\beta-\beta e^{-(\alpha-\beta)t}}{\alpha-\beta e^{-(\alpha-\beta)t}} .$$

Since r_c is the radius of a cell, for a focus of radius r, $q=(r/r_c)^3$ and

$$(1.7) \qquad g_3(r)dr = \frac{3[(\alpha/\beta)p(t)]^q}{r\ln\left(\frac{\beta}{\beta-\alpha p(t)}\right)}dr$$

The cell division and cell death rates are assumed to be independent of the size of a PML. This assumption is probably reasonable as a first approximation, especially if none of the foci are very large. This is the reason Moolgavkar et. al. (1990) chose to base their analysis on transections no more than R (=500μm in their paper) in radius, i.e. to condition the analysis with radii between ε and R. So R is the upper bound at which we wish to establish a condition, i.e. all sections of foci with radii larger than R are to be dropped from the analysis. Expression (1.4) is then replaced by

$$(1.8) \qquad \tilde{\lambda}_2(t)=\{[G_2(R)-G_2(\varepsilon)]2\tilde{r}_3\Lambda(t)\}$$

and expression (1.5) is replaced by

$$(1.9) \qquad \tilde{g}_2^{\varepsilon}(y)=g_2^{\varepsilon}(y)/\, G_2^{\varepsilon}(R)$$

After fitting the two stage stochastic model, Moolgavkar et al. (1990) found that the liver foci data used in this analysis have a distribution which is significantly different from a Poisson distribution. Future research dealing with this problem is essential.

1.4.6 The Damage-Fixation Model by Portier and Kopp-Schneider

Both the Armitage-Doll model and the two-event stochastic model ignore the ability of damaged cells to recover. Anderson (1987) pointed out that a mutation includes two steps: first, damage to the DNA must occur; second, this damage must then be fixed by replication of the damaged cell. Prior to fixation, this damage may be repaired through many mechanisms in the cell (Hanawalt,1987). Portier and Kopp-Schneider (1991) proposed a 2-stage model of carcinogenesis with clonal expansion of all cell types where the mutation step has been expanded in the way described by Anderson (1987).

This model (see Figure 1.12) is based upon 5 cell types: normal cells (M_0), 2 types of damaged cells (D_i) which are subject to DNA repair and 2 types of mutated cells (M_i) where the DNA damage has been fixed by cell replication, i=1,2. Basically, the model can be described as follow. Normal cells divide, die (differentiate), or mutate with rates of α_0, β_0, and μ_0 respectively. The mutation in these damaged cells are assumed to pertain to a single strand and can be repaired at the rate ρ_1. Damaged cells also replicate or migrate out via a birth and death process with birth rate α_1 and death rate β_1. When cell division occurs in these damaged cells, the DNA damage is fixed in one of the daughter cells resulting in the creation of a single mutated cell, M_1. The other daughter cell was derived from the strand of DNA without damage and is thus a normal cell. The parameters are defined as follows.

In a small time of interval $[t,t+\Delta t)$:

(1) a normal cell (M_0) may:

(a) divide into two normal cells with probability $\alpha_0(t)\Delta t+o(\Delta t)$,

(b) be removed with probability $\beta_0(t)\Delta t+o(\Delta t)$,

(c) or be damaged with probability $\mu_0(t)\Delta t+o(\Delta t)$;

(2) a damaged normal cell (D_1) may:

(a) divide into a normal and an intermediate cell with probability $\alpha_1(t)\Delta t+o(\Delta t)$,

(b) be removed with probability $\beta_1(t)\Delta t+o(\Delta t)$,

(c) or be repaired and become a normal cell with probability $\rho_1\Delta t+o(\Delta t)$;

(3) an intermediate cell (M_1) may:

(a) divide into two intermediate cells with probability $\alpha_2(t)\Delta t+o(\Delta t)$,

(b) be removed with probability $\beta_2(t)\Delta t+o(\Delta t)$,

(c) or be damaged with probability $\mu_2(t)\Delta t+o(\Delta t)$;

(4) a damaged intermediate cell (D_2) may:

(a) divide into an intermediate cell and a malignant cell (M_2) with probability $\alpha_3(t)\Delta t+o(\Delta t)$,

(b) die with probability $\beta_3(t)\Delta t+o(\Delta t)$,

(c) or be repaired and become an intermediate cell with probability $\rho_3(t)\Delta t+o(\Delta t)$;

(5) no change occurs in the system with probability

$$\Delta t(1 - \mu_0(t) - \rho_1(t) - \mu_2(t) - \rho_3(t) - \sum_{i=0}^{3}(\alpha_i(t)+\beta_i(t)))+ o(\Delta t).$$

It is assumed that the probability of more than one event occurring in an interval Δt is $o(\Delta t)$, which is defined in section 1.4.5.

Based on this Damage-Fixation Multistage (DFM) model, Portier and Kopp-Schneider (1991) compute the expected number of cells of each type, and solved the

distribution of tumor onset times using the Kolmogorov backward equations. Kopp-Schneider, Portier and Rippmann (1991) then developed the mathematical formulae needed to apply the DFM model to two-stage skin-painting studies. For example, they approximate the rate of formation of initiated cells using a pseudo-steady state approximation to the number of normal and (specifically) damaged normal cells, and they also computed the number of detectable clones of intermediate cells in an Initiation-Promotion experiment.

The mathematical formulae of the number of detectable clones of initiated cells can be used to model whether the initiated cells are produced according to a Poisson process and whether they are subject to a linear birth and death process.

After fitting the DFM model to a mouse-skin painting experiment, Kopp-Schneider et al. (1991) found that the papilloma data used in this analysis have a distribution which is significantly different from a Poisson distribution. They suggested that future research should incorporate the possibility of overdispersed Poisson or non-Poisson data by possibly putting a prior distribution on the mean number of papillomas per animal.

1.5 Goal of research

As was discussed earlier, the mathematical development related to the multistage modelling of carcinogenesis is based upon an assumption that the number of PMLs follow a Poisson distribution. Since a Poisson assumption on the number of PMLs is the cornerstone of almost all computations, the reliability of this assumption will be examined, and corrections to this assumption will be made where necessary.

After reviewing the literature, we find that the distribution of the number of PMLs does not seem to follow a Poisson distribution (Chapter 2). To examine and modify the Poisson assumption on the distribution of the number of PMLs, two major approaches, i.e., classical Mixture model (Chapter 3) and Quasi-likelihood methods (Chapter 4) are taken. Different priors are put on the mutation rate in Chapter 3, and a relationship between mean and variance is assumed in the Quasi-likelihood method. In chapter 5, a simulation study looking at the operating characteristics of the above two approaches are done .

In Chapter 6, appropriate mathematical formulae for the size distribution of PMLs of the IP (Initiation-Promotion) design are developed based on the two-stage model of Moolgavkar and others. A brief discussion of the results is given and future work suggested.

Chapter 2

In analyzing several datasets from rat liver foci and mouse skin papillomas (detailed descriptions are in section 2.1), it is observed that the variance of the number of PML's is much larger than the mean. This potentially violates the definition of a Poisson distribution, where the variance equals the mean. Several statistical tools are used to check the Poisson assumption of different datasets in Chapter 2.

Moolgavkar et al (1991) tried to explain this extra-Poissoness as due to errors in counting foci, variation in intake of the promoter within a single dose group, and variations in individual susceptibility to the agent. Any of these may be a factor in the extra-Poissoness, but they also stated that "statistical methods for addressing this extra-Poisson variation need to be explored" (Moolgavkar et al, 1990). For unadjusted extra-Poisson variability, possible consequences may include inefficient estimation and under-estimation of variances of parameter estimates (Margolin et al, 1981). Hence it is without doubt that this extra-Poisson variation must be addressed.

2.1 Index of Dispersion

In the Lewis rats experiment (Moolgavkar et al; 1990), the rats were subjected to *N*-nitro-morpholine (NNM), a promoter, at various concentrations (0, 0.1, 1, 5, 10, 20, 40, and 80 ppm.) in the drinking water starting at 14 weeks of age. Those in the highest exposure group drank very little water possibly due to the high toxicity of the NNM. This makes it difficult to analyze data from this group so they were dropped from the analysis. The dependent variable of interest is the number of two-dimensional transections of liver foci.

The remaining data consist of 165 rats in 7 dose groups. Table 2.1 has the basic information (number of rats, mean, and standard deviation in each dose group). The plots in Figure 2.1 to 2.7 show the relationship between the number of PML's and time of exposure for every dose group. It also shows that, except for one outlier (ID = 41, NNM=40, # of PML's= 147), the number of PML's in the dataset are within the range of 0 to 80. Table 2.2 provides the correlation between the number of PML's and the length of exposure, by group.

NNM(ppm)	# of rats	λ (Mean)	s. d.	V (Variance)	ρ
0	37	2.3611	6.7449	45.4936	7.7371
0.1	31	3.6774	5.0026	25.0258	1.5786
1	23	5.3478	7.1960	51.7824	1.6236
5	17	12.7647	18.0435	325.5679	1.9198
10	16	13.2500	17.6956	313.1343	1.7081
20	19	14.8019	17.4749	305.3721	1.3262
40	19	26.6667	33.8248	1144.1171	1.5714

Table 2.1 The mean and variance of the number of liver foci, and the corresponding value of ρ according to the assumed relationship: $V = \lambda + \rho \lambda^2$ from the study of Moolgavkar et al (1990)

NNM (ppm)	Pearson's correlation coefficient	p-value
0	0.3849	0.0187
0.1	0.4755	0.0069
1	0.6761	0.0004
5	0.7176	0.0012
10	0.7684	0.0002
20	0.7936	0.0001
40	0.2640	0.2748

Table 2.2 The correlation between the number of PML's and length of exposure for each dose group in the NNM rat liver foci data of Moolgavkar et al (1990).

All groups show significant correlation between the number of PML's and length of exposure except the last dose group. The groups with medium exposures (NNM=1, 5, 10, 20) have high positive correlation which implies the longer the exposure time, the larger the number of PML's. The low dose groups (including NNM=0.1 and the control group NNM=0) have moderate correlation. The doses are so low that the rats, though constantly exposed to NNM, may still survive hundreds of days with a handful of PML's. The last group has a dose so high (NNM=40) that most of the rats died between day 160 and 220 days (Figure 2.7). This confounds our ability to observe a significant correlation since the rats died so young.

It is observed (Table 2.1) that the higher the dose, the greater the number of PML's. The variances are much larger than the mean across all of the dose groups. To take exposure time into consideration, the data were broken down into different time intervals in Tables 2.3 to 2.9. It is shown that the longer the rats were exposed to NNM in the drinking water, the more liver foci they developed with only a few exceptions. Again, the variance is observed to be larger and, in many cases, much larger than the mean.

To further examine whether the Poisson assumption of number of liver foci (PMLs) is violated, the index of dispersion test is done. Let N be the number of PML's. Assuming N is distributed as a Poisson random variable, E(N) = V(N) where E(N) represents the expected value and V(N) represents the variance. It is possible to test for variance greater than that of a Poisson random variable by testing:

$$H_0: \quad V(N) = E(N)$$

$$\text{vs.} \quad H_A: \quad V(N) > E(N).$$

The test statistic is

$$\frac{(m-1)S^2}{\bar{N}} \sim \chi^2_{m-1}$$

where χ^2_{m-1} is the Chi-squared distribution with m-1 degrees of freedom and m is the number of animals in that group. In 21 of the 26 cases we studied (Table 2.3 to 2.9) the p-value for the test of equal mean and variance is smaller than 0.05. In other words, V(N) was significantly larger than E(N) in most of the cases. So it can be concluded that the Poisson assumption was indeed violated.

We drew the same results from the mouse skin papilloma data of Kopp-Schneider et.al.(1991). These data derive from a mouse skin painting experiment in which 7,12-dimethylbenz[a]anthracene (DMBA) was applied once onto the shaved back of each mouse and, after 1 week lag time, 12-*O*-tetradecanoylphorbol-13-acetate (TPA) was applied twice weekly onto the same area of the skin. Here DMBA acts as the initiator, and TPA as the promoter. Promotion was continued for a different period of time: 5,10,20, and 40 weeks. It is shown in Table 2.10 that in 5 of the 6 groups studied, the p-value for the test of index of dispersion is much smaller than 0.05, again indicating a severe violation of the Poisson assumption.

In another experiment in which DEN is the initiator and 2,3,7,8-tetrachlorobenzo-p-dioxin (TCDD) is the promoter, the rats were exposed to various concentrations (0, 3.5, 10.7, 35.7, and 125 ppm) of TCDD (Trischer et al, 1991). The results of an analysis for overdispersion in the five treatment groups are shown in Table 2.11. As before, all of the variances are much bigger than the means for these groups ($p<0.05$). Similar results hold for the group with saline instead of DEN where the p-values are much smaller than 0.05 in 4 out of 5 cases (Table 2.12).

Time (days)	m	Mean	Variance	χ^2_{m-1}	P - value
$100 \leq T < 200$	5	0	0	.	.
$200 \leq T < 300$	6	0.1667	0.1666	5.000	0.4156
$300 \leq T < 400$	5	0.6000	1.7999	12.000	0.0174
$400 \leq T < 500$	8	0.5000	0.2857	4.000	0.7798
$500 \leq T < 600$	3	4.3333	33.3333	15.385	0.0005
$600 \leq T$	10	9.2000	254.6227	249.087	0.0000

Table 2.3 Mean and variance of the number of liver foci and the results of the index of dispersion test for different time points in the dose group of NNM = 0 ppm from the data of Moolgavkar et al (1990).

Time (days)	m	Mean	Variance	χ^2_{m-1}	P - value
$200 \leq T < 300$	3	1.3333	1.3333	2.000	0.3679
$300 \leq T < 400$	4	0.7500	2.2500	9.000	0.0293
$400 \leq T < 500$	5	1.4000	6.8001	19.429	0.0007
$500 \leq T < 600$	6	4.0000	27.2004	34.000	0.0000
$600 \leq T$	13	7.3077	60.8977	100.000	0.0000

Table 2.4 Mean and variance of the number of liver foci and the results of the index of dispersion test for different time points in the dose group of NNM = 0.1 ppm from the data of Moolgavkar et al (1990).

Time (days)	m	Mean	Variance	χ^2_{m-1}	P - value
$200 \leq T < 300$	6	0.8333	1.7668	10.600	0.0599
$300 \leq T < 400$	5	3.2000	16.7003	20.875	0.0003
$400 \leq T < 500$	4	8.5000	75.0008	26.471	0.0000
$500 \leq T < 600$	4	9.2500	28.9164	9.378	0.0247
$600 \leq T$	4	16.75	150.9163	27.030	0.0000

Table 2.5 Mean and variance of the number of liver foci and the results of the index of dispersion test for different time points in the dose group of NNM = 1 ppm from the data of Moolgavkar et al (1990).

Time (days)	m	Mean	Variance	χ^2_{m-1}	P - value
$200 \leq T < 300$	9	5.2222	51.1940	78.426	0.0000
$300 \leq T < 400$	5	27.8000	763.6988	109.885	0.0000
$400 \leq T < 500$	3	28.6667	86.3338	6.023	0.0492

Table 2.6 Mean and variance of the number of liver foci and the results of the index of dispersion test for different time points in the dose group of NNM = 5.0 ppm from the data of Moolgavkar et al (1990).

Time (days)	m	Mean	Variance	χ^2_{m-1}	P - value
$200 \leq T < 300$	13	6.6154	6.0895	11.047	0.5249
$300 \leq T < 400$	4	51.0000	589.3339	34.667	0.0000

Table 2.7 Mean and variance of the number of liver foci and the results of the index of dispersion test for different time points in the dose group of NNM = 10.0 ppm from the data of Moolgavkar et al (1990).

Time (days)	m	Mean	Variance	χ^2_{m-1}	P - value
$100 \leq T < 200$	3	15.0000	156.0001	20.8	0.0000
$200 \leq T < 300$	13	16.0000	170.3338	127.75	0.0000
$300 \leq T < 400$	3	62.0000	1651.0001	53.258	0.0000

Table 2.8 Mean and variance of the number of liver foci and the results of the index of dispersion test for different time points in the dose group of NNM = 20.0 ppm from the data of Moolgavkar et al (1990).

Time (days)	m	Mean	Variance	χ^2_{m-1}	P - value
$100 \leq T < 200$	10	49.1000	2154.1016	394.845	0.0000
$200 \leq T < 300$	9	19.5714	111.2856	34.117	0.0000

Table 2.9 Mean and variance of the number of liver foci and the results of the index of dispersion test for different time points in the dose group of NNM = 40.0 ppm from the data of Moolgavkar et al (1990).

TPA exposure time (week)	m	Mean	Variance	χ^2_{m-1}	p-value
5	20	0.15	0.02	2.85	0.99999
10	20	4.75	10.98	43.93	0.00096
20	20	32.05	137.30	81.40	0.00000
40	20	65.25	177.38	51.65	0.00007

Table 2.10 Mean and variance of the number of liver foci and the results of the index of dispersion test for four groups from the mouse skin papilloma data of Kopp-Schneider et.al.(1991)

TCDD (ppm)	m	Mean	Variance	χ^2_{m-1}	p-value
0	10	30	416	124.8	0
3.5	9	57.77778	1293.944	179.161	0
10.7	9	63.22222	4025.944	509.434	0
35.7	8	48.125	469.5536	68.299	1.071E-11
125	9	77.8889	1807.111	185.609	0

Table 2.11 Mean and variance of the number of liver foci and the results of the index of dispersion test for four groups from the TCDD rat liver foci data (Trischer et al, 1991).

TCDD (ppm)	m	Mean	Variance	χ^2_{m-1}	p-value
0	10	1.2	2.1778	16.333	0.090480
3.5	9	2	22.0000	88.0	0.000000
10.7	9	1.77778	6.6944	30.125	0.000418
35.7	8	6	78.75	105.0	0.000000
125	9	4	27.4286	48	0.000000

Table 2.12 Mean and variance of the number of liver foci and the results of the index of dispersion test for four groups from the TCDD rat liver foci data (Trischer et al, 1991).

2.2 Poisson regression

Since length of exposure confounds the analysis of overdispersion for our NNM rat liver foci data, a straight linear Poisson regression analysis is done in addition to the index of dispersion test to check if the Poisson assumption is correct. We obtain poor fits when we attempt to fit the Poisson regression model to the NNM rat liver foci data, with scaled deviance ranging from 5 times as large as the degrees of freedom (when NNM = 1 ppm) to 30 times (when NNM=40 ppm) in Table 2.13. The same analysis applied to the TCDD data yields similar results (Table 2.14).

This overdispersion may be due to an incorrect assumption concerning the independence of the response variables, or because some important explanatory variables are not measured, or cannot be measured, and are consequently excluded from the regression relationship. However, it is common that even after allowing for important explanatory variables using the Poisson regression model, the fits obtained are still bad. This is often reflected in high scaled deviance which indicates that, conditional upon the explanatory variables included in the model, the variance of an observation is greater than its mean, implying that the Poisson assumption is incorrect.

NNM (ppm)	α	μX	Scaled deviance	d.f.
0	0.2573	14.9526	235.1892	36
0.1	0.3432	12.3459	161.9021	30
1	0.4727	20.9823	108.6324	22
5	0.8585	21.4007	131.1878	16
10	1.26	31.309	96.5003	15
20	1.149	39.763	189.7139	18
40	1.7109	46.658	519.4161	18

Table 2.13 Analysis of PMLs following exposure to NNM ignoring dispersion and using Poisson regression. α is the birth rate, μ is the mutation rate and X is the number of normal cells per cubic centimeter assumed to be a constant.

TCDD (ppm)	α	μX	Scaled deviance	d.f.
0	0.3581	5.952	135.92	9
3.5	0.4432	8.346	261.21	8
10.7	0.5991	11.477	608.24	8
35.7	0.6123	12.433	831.78	7
125	0.6008	24.312	919.46	8

Table 2.14 Analysis of PMLs following exposure to TCDD (dioxin) ignoring dispersion and using Poisson regression. α is the birth rate, μ is the mutation rate and X is the number of normal cells per cubic centimeter assumed to be a constant.

From all of the above test results, it is quite clear that the variance of N, the number of PML's, is much larger than its mean. Thus, instead of V(N)=E(N), some other relationship may exist between V(N) and E(N). In what follows, we will explore this relationship.

2.3 Mean-variance plot

Figures 2.8 (NNM rat liver foci data), 2.9 (TCDD rat liver foci data) and 2.10 (mouse skin papilloma data) illustrate the mean-variance plots for different datasets. If the Poisson assumption of N is correct, then the mean should be equal to the variance and there should be a straight line. Yet it seems from the figures that the variance is a quadratic function of the mean:

$$(2.1) \qquad V = \lambda + \rho\lambda^2,$$

as shown in Table 2.1.

In Chapter 3, we will impose various prior distributions on the mutation rate. It turns out that the marginal mean-variance relation of the number of PML's (derived in Chapter 3) is similar to what we observed in equation (2.1). This apparent mean-variance relationship in our dataset will also lead us into the quasi-likelihood method in Chapter 4 which is based, not on the assumed distribution, but on the mean-variance relationship of the response variables.

Chapter 3

3.1 Prior information

Data that exhibit an extra-Poisson variation may have a distribution which is dependent upon some covariates. In attempting to identify these covariates, it is best to go back to the biological phenomenon. When a normal cell becomes an initiated cell with a mutation rate of μ_1, then the initiated cell itself may generate two initiated cells with a birth rate α_1 or it may die with a rate β_1 as shown in Figure 1.11. In what follows, we will simplify our problem by setting $\alpha_0=\beta_0=0$ and dropping μ_2, dealing only with the premalignant state. From now on, we will drop the subscripts and use μ, α, and β for mutation rate from normal to initiated, birth rate of initiated cells, and death rate of initiated cells respectively to simplify notation.

Of the above rates, the mutation rate is the most attractive as a source of variation in model response. Mutations play a part in the micro-evolution of all the creatures. It is a rare event in nature with rates that differ from individual to individual. So, even under exactly the same situation, creatures react very differently. In rodent experiments, the mutation rate may not be a constant among all of the rats in the same dose group, but rather a variable which has its own distribution.

In dealing with the problem of extra-Poisson variation, one approach in the thesis will be to set a prior on the distribution of the mutation rate. Introducing a prior distribution may bias the results of the estimation process so as to favor parameter values for which the prior density function is relatively large. However, when the amount of data available is

sufficiently large, the effect of the prior distribution on the parameter estimates is negligible.

Instead of assuming that the distribution of the number of PML's is Poisson, it is assumed that the distribution of the number of PML's conditional on the mutation rate is Poisson. As for the reasonable prior distributions of the mutation rate across individuals, both the gamma and uniform distributions are considered.

The first prior we want to put on the mutation rate is a gamma distribution. A gamma distribution is very flexible: when the shape parameter $\gamma=\nu/2$ and scale parameter $\delta=2$, the distribution is chi-squared with ν degrees of freedom; when the shape parameter $\gamma=1$, the distribution is exponential with mean δ; it is able to approximate the normal distribution as the shape parameter becomes large (Figures 3.1 to 3.4). One of the main reasons for the use of the gamma as a compounding distribution is that it leads to an analytically tractable compound distribution, and it also has the appealing property that the means are preserved. The gamma distribution is an important tool in the evaluation of statistical methods in carcinogenesis.

The mutation rates for subjects within a chemical dose group will have a range, such as, (a,b). Let us assume that it is as equally likely to have a subject i with mutation rate μ_i in the interval (a, a+Δ) as it is to have a subject j with mutation rate μ_j in the interval (b-Δ, b). This would imply that the mutation rate follows a uniform distribution. With a uniform prior, the marginal distribution of PML's is obtained from the conditional distribution of the mutation rate.

From the work of Dewanji et al (1989), we know that the distribution of N conditional on the mutation rate μ (assumed to be a constant in their paper) is Poisson with an expected value $\Lambda(t)$.Assume the mutation rate μ is a random variable with probability

density function $f(\mu)$, and N be the number of PML's. In other words, combining Dewanji's work and our new assumption here, we have

$$N|\mu \sim \text{Poisson}(\Lambda(t))$$

$$\mu \sim f(\mu),$$

where $\Lambda(t)= \frac{\mu X}{\alpha}\left[\ln\left(\frac{\beta}{\beta-\alpha p(t)}\right)\right]$, X is the number of normal cells per cubic centimeter, and $p(t) = \frac{\beta-\beta e^{-(\alpha-\beta)t}}{\alpha-\beta e^{-(\alpha-\beta)t}}$ is the probability of extinction as defined in Chapter 1 from the work of Dewanji et al (1989). To simplify notation in the future, we define

$\Lambda=\frac{1}{\alpha}\left[\ln\left(\frac{\beta}{\beta-\alpha p(t)}\right)\right]$ so $\Lambda(t)=\mu X\Lambda$.

In this section, the quantities of interest are the marginal distribution, the expected value, the variance, and the likelihood function of the number of PML's, based upon various assumed distributions of μ.

3.2 A gamma prior on the mutation rate.

We first assume that the mutation rate has a gamma distribution:

$$(3.1) \qquad f(\mu)=\frac{\mu^{\gamma-1}e^{-\mu/\delta}}{\delta^{\gamma}\Gamma(\gamma)},\ \gamma>0,\ \delta>0,\ \mu\geq 0$$

Traditionally, the symbols for parameters in a gamma distribution are α and β. However, α and β in this dissertation are designated to denote birth rate and death rate. To avoid the confusion of using the same Greek letters, we will instead use γ and δ for the shape and scale parameters.

The range for a gamma variable is from zero to infinity, but the range of the mutation rate cannot be infinity. Since a mutation is a rare event, the mutation rate is in fact very small even under high doses of a chemical agent. Therefore, when we assume that the mutation rate is a gamma random variable, we have to deal with this range problem.

The range problem can be overcome by using the product of the mutation rate μ and X, which is the number of normal cells per cubic centimeter. X is large and assumed to be a constant (Moolgavkar et al, 1990). Here μ is the mutation rate of a normal cell per day, so the combined term μX is the mutation rate of all the cells per cubic centimeter per day. Generally speaking, the range of X is between 10^6 and 10^8. Therefore, μX rather than μ is treated as a random variable which is assumed to have a gamma distribution to avoid the range problem.

The normal cells in this two stage model of carcinogenesis, just like the initiated cells, also follows a stochastic birth-and death process. Since X is so large, its fluctuation is comparably small. However, this stochastic property of the normal cells makes the computation of the rates of the initiated cells extremely complicated. To avoid this, the number of normal cells per unit volume is a assumed to be a constant

The marginal distribution of the number of PML's then becomes a simple problem which has already been worked out in most statistics textbooks. Given $N|\mu X \sim$ Poisson($\Lambda(t)$) and $\mu X \sim$ Gamma (γ, δ), the posterior distribution of N is the following:

$$P(N=n) = \int_0^{\infty} P(N=n\,|\mu X)\, P(\mu X)\, d\mu X$$

$$= \binom{\gamma + n - 1}{\gamma - 1} \left(\frac{\delta\Lambda}{1+\delta\Lambda}\right)^n \left(\frac{1}{1+\delta\Lambda}\right)^{\gamma}$$

Define the new random variable $H = N+\gamma$, then

$$P[H=h] = \binom{h-1}{\gamma-1} p^{\gamma} (1-p)^{h-\gamma} \quad , h=\gamma,\ \gamma+1,\ \ldots\ldots\infty$$

where $p=\frac{1}{1+\delta\Lambda}$. Therefore, H has a negative binomial distribution with parameter γ and p. Its expected value and variance are

$$E(H) = \frac{\gamma}{p} = \frac{\gamma}{\frac{1}{1+\delta\Lambda}} = \gamma(1+\delta\Lambda)$$

$$V(H) = \frac{\gamma(1-p)}{p^2} = \frac{\gamma(1-\frac{1}{1+\delta\Lambda})}{(\frac{1}{1+\delta\Lambda})^2} = \gamma\delta\Lambda(1+\delta\Lambda)$$

The expected value and variance of N are

$$(3.2) \qquad \begin{cases} E(N) = \gamma\delta\Lambda \\ V(N) = \gamma\delta\Lambda(1+\delta\Lambda) \end{cases}$$

In other words,

$$(3.3) \qquad V(N)=E(N) + \frac{[E(N)]^2}{\gamma}.$$

So the smaller the shape parameter, the larger the V(N). This quadratic mean-variance relationship echoes what was observed in Chapter 2 (equation (2.1)).

The log-likelihood for subject i is

$$\text{Log}(P(N=n_i)) = \log(n_i+\gamma-1)! - \log(\gamma-1)! - \log(n_i)! + \gamma \log(p) + n_i \log(1-p)$$

And the log-likelihood for the entire data is the summation of the above over all individuals.

3.3 A uniform prior on the mutation rate

Assume μX has a uniform distribution, i.e.

$$P(\mu X)= \frac{1}{b-a} \text{ if } a < \mu X < b$$

$$=0 \text{ elsewhere.}$$

The marginal distribution of N will be

$$P(N=n)=\int_a^b P(N=n\,|\mu X)\,P(\mu X)\,d\mu X$$

$$=\int_a^b \frac{1}{b-a}\,\frac{(\mu X\Lambda)^n e^{-\Lambda\mu X}}{n!}\,d\mu X$$

$$=\frac{\Lambda^n}{(b-a)n!}\int_a^b (\mu X)^n e^{-\Lambda\mu X}\,d\mu X$$

Let us solve $\int_a^b (\mu X)^n e^{-\Lambda\mu X}\,d\mu X$ first. For simplicity, substitute x for μX in this calculation. After integration by parts,

$$\int_a^b x^n e^{-\Lambda x}\,dx$$

$$=\left[-\frac{e^{-\Lambda x}x^n}{\Lambda}-\frac{e^{-\Lambda x}nx^{n-1}}{\Lambda^2}-\cdots-\frac{e^{-\Lambda x}n(n-1)\cdots4\cdot3\cdot2x}{\Lambda^n}-\frac{e^{-\Lambda x}n!}{\Lambda^{n+1}}\right]_a^b$$

$$=\left[-\frac{e^{-\Lambda x}n!}{\Lambda^{n+1}}\sum_{k=0}^{n}\frac{(\Lambda x)^k}{k!}\right]_a^b=\left[-\frac{n!}{\Lambda^{n+1}}\sum_{k=0}^{n}\frac{(\Lambda x^k)e^{-\Lambda x}}{k!}\right]_a^b.$$

Therefore, $P(N=n)=\frac{\Lambda^n}{(b-a)n!}\left[-\frac{n!}{\Lambda^{n+1}}\sum_{k=0}^{n}\frac{\Lambda x^k e^{-\Lambda x}}{k!}\right]_a^b$

$$(3.4)\quad =\frac{1}{(b-a)\Lambda}\left[\sum_{k=0}^{n}\frac{(\Lambda a)^k e^{-\Lambda a}}{k!}-\sum_{k=0}^{n}\frac{(\Lambda b)^k e^{-\Lambda b}}{k!}\right]$$

This, (3.4), is the marginal distribution of the number of PML's. We can also express (3.4) in a more concrete way. Let N_a, N_b be two Poisson variables with mean Λa and Λb, respectively. Thus we have

$N_a \sim$ Poisson(Λa)

$N_b \sim$ Poisson(Λb),

so that

$$P(N_a \le n) = \sum_{k=0}^{n}\frac{(\Lambda a)^k e^{-\Lambda a}}{k!}$$

$$P(N_b \le n) = \sum_{k=0}^{n}\frac{(\Lambda b)^k e^{-\Lambda b}}{k!}$$

Therefore (3.4) becomes

$$(3.5) \qquad P(N=n) = \frac{1}{(b-a)\Lambda}\left[P(N_a \le n) - P(N_b \le n)\right].$$

Namely, the marginal probability function of N is a fraction of the difference between the cumulative distribution functions (CDF's) of two Poisson variables.

Next we generate the expected value and variance through the probability generating function for the number of initiated cells, $\phi(s)$.

$$\phi(s) = E(s^n) = \sum_{n=0}^{\infty} s^n \frac{1}{(b-a)\Lambda} \sum_{k=0}^{n} \left[\frac{(\Lambda a)^k e^{-\Lambda a}}{k!} - \frac{(\Lambda b)^k e^{-\Lambda b}}{k!}\right], \; 0 \le k \le n < \infty$$

$$= \frac{1}{(b-a)\Lambda} \sum_{k=0}^{\infty} \left[\frac{(\Lambda a)^k e^{-\Lambda a}}{k!} - \frac{(\Lambda b)^k e^{-\Lambda b}}{k!}\right] \sum_{n=k}^{\infty} s^n$$

$$= \frac{1}{(b-a)\Lambda} \sum_{k=0}^{\infty} \left[\frac{(\Lambda a)^k e^{-\Lambda a}}{k!} - \frac{(\Lambda b)^k e^{-\Lambda b}}{k!}\right] \frac{s^k}{1-s}$$

$$= \frac{1}{(s-1)(a-b)\Lambda} \sum_{k=0}^{\infty} \left[\frac{s^k(\Lambda a)^k e^{-\Lambda a}}{k!} - \frac{s^k(\Lambda b)^k e^{-\Lambda b}}{k!}\right]$$

$$= \frac{1}{(s-1)(a-b)\Lambda} \left\{ e^{-\Lambda a} \sum_{k=0}^{\infty} \left[\frac{s^k(\Lambda a)^k}{k!}\right] - e^{-\Lambda b} \sum_{k=0}^{\infty} \left[\frac{s^k(\Lambda a)^k}{k!}\right] \right\}$$

$$= \frac{1}{(s-1)(a-b)\Lambda} \left[e^{-\Lambda a} e^{\Lambda a s} - e^{-\Lambda b} e^{\Lambda b s} \right]$$

$$= \frac{1}{(s-1)(a-b)\Lambda} \left[e^{\Lambda a(s-1)} - e^{\Lambda b(s-1)}\right]$$

$$= \frac{1}{(s-1)(a-b)\Lambda} \sum_{k=0}^{\infty} \left[\frac{(\Lambda a)^k (s-1)^k}{k!} - \frac{(\Lambda b)^k (s-1)^k}{k!}\right]$$

$$= \frac{1}{(s-1)(a-b)\Lambda} \sum_{k=0}^{\infty} \frac{1}{k!}\left[(\Lambda a)^k - (\Lambda b)^k\right](s-1)^k$$

So,

$$(3.6)\qquad \phi(s)= \frac{1}{(a-b)\Lambda}\sum_{k=0}^{\infty}\left[\frac{(\Lambda a)^{k+1}-(\Lambda b)^{k+1}}{(k+1)!}(s-1)^{k}\right],\ |s|<1.$$

To get the expected value and variance of N, the first derivative and second derivative of the above probability generating function have to be found first:

$$\phi'(s)= \frac{1}{(a-b)\Lambda}\sum_{k=1}^{\infty}\left[\frac{(\Lambda a)^{k+1}-(\Lambda b)^{k+1}}{(k+1)!}k(s-1)^{k-1}\right]$$

$$\phi''(s)= \frac{1}{(a-b)\Lambda}\sum_{k=2}^{\infty}\left[\frac{(\Lambda a)^{k+1}-(\Lambda b)^{k+1}}{(k+1)!}k(k-1)(s-1)^{k-2}\right]$$

As s approaches 1, we can get the expected value and variance of N:

$$(3.7)\qquad E(N) = \phi'(s)\big|_{s=1} = \frac{(a+b)\Lambda}{2}$$

$$(3.8)\qquad V(N) = E(N^2) - (E(N))^2 = E(N(N-1)) + E(N) - (E(N))^2$$

$$= \phi''(s)\big|_{s=1} + \phi'(s)\big|_{s=1} - (\phi'(s)\big|_{s=1})^2$$

$$= \frac{\Lambda^2(a-b)^2 + 6(a+b)\Lambda}{12}$$

$$= E(N) + [E(N)]^2\frac{(a-b)^2}{3(a+b)^2}$$

So when the lower bound, a, for the mutation rate is near zero, the mean-variance relationship is

$$V(N) = E(N) + \frac{[E(N)]^2}{3}$$

From the form of the expected value and variance, the marginal distribution of the number of PML's bears a resemblance to the uniform distribution.

The likelihood contribution for subject i with n liver foci will be :

$$L_i = \frac{1}{(b-a)\Lambda}\left[P(N_a \le n_i) - P(N_b \le n_i)\right]$$

And the log-likelihood for subject i is

$$(3.9)\qquad \text{Log}(P(N=n)) = -\log(b-a) - \log\Lambda + \log\sum_{k=0}^{n}\left[\frac{(\Lambda a)^k e^{-\Lambda a}}{k!} - \frac{(\Lambda b)^k e^{-\Lambda b}}{k!}\right]$$

The log-likelihood function for the entire group is simply a summation of the above formula.

As a side issue, we were interested in trying to find classes of prior distribution which might have a mean-variance relationship which was not quadratic. Table 3.1 extends the results from Chapter 2 to additional members of the exponential family.

Distribution of μX	Probability function $f(\mu X)$	Mean-variance relationship
uniform	$\frac{1}{b-a}$, $a \le \mu X \le b$	$V(N)=E(N)+[E(N)]^2\frac{(a-b)^2}{3(a+b)^2}$
gamma	$\frac{\mu X^{\gamma-1} e^{-\mu X/\delta}}{\delta^{\gamma}\Gamma(\gamma)}$, $\gamma>0$, $\delta>0$, $\mu X \ge 0$	$V(N)=E(N)+\frac{[E(N)]^2}{\gamma}$
exponential	$\frac{e^{-\mu X/\delta}}{\delta}$, $\delta>0$, $\mu X \ge 0$	$V(N)=E(N)+[E(N)]^2$
chi-square	$f(\chi^2)=\frac{(\chi^2)^{(\nu/2)-1} e^{-\chi^2/2}}{2^{\nu/2}\Gamma(\nu/2)}$, $\chi^2>0$	$V(N)=E(N)+\frac{[E(N)]^2}{(\nu/2)}$
normal	$\frac{1}{\sigma\sqrt{2\pi}}\exp[-(\frac{1}{2\sigma^2})(\mu X-\varpi)^2]$	$V(N)=E(N)+\frac{[E(N)]^2}{\varpi/\sigma}$

Table 3.1:Mean-variance relationship of different priors on mX assuming N|mX is Poisson

The mean-variance relationships for uniform and gamma priors are worked out and shown in equation (3.8) and (3.3), respectively. We know that the exponential distribution is a special case of the gamma distribution when the shape parameter γ is 1, and the chi-

square distribution is another when ($\gamma = \nu/2, \delta=2$) where ν is the degrees of freedom. From this, it is simple to derive the mean-variance relationship for the exponential prior:

$$V(N)=E(N)+ [E(N)]^2$$

and for the chi-square prior:

$$V(N)=E(N)+ \frac{[E(N)]^2}{(\nu/2)}.$$

The normal distribution can be approximated by the gamma distribution as the shape parameter γ becomes large (Figures 3.1 to 3.4). We know that the mean , ϖ , of this approximated normal distribution is also the mean of the gamma distribution $\gamma\delta$, and the standard deviation σ is the scale parameter δ. Hence the mean-variance relationship for the normal prior is approximately

$$V(N)=E(N)+ \frac{[E(N)]^2}{\varpi / \sigma}$$

The mean-variance relationships of the five priors of μX, namely uniform, gamma, exponential, chi-square, and normal, are (approximately, at least) of the same quadratic form

$$V(N)=E(N)+ \frac{[E(N)]^2}{\kappa}$$

with κ equals $\frac{(a-b)^2}{3(a+b)^2}$, γ, 1, $\nu / 2$, and ϖ / σ for the five priors respectively. Therefore, these five priors all fit into the category of "negative binomial" in the quasi-likelihood method table. Which means their quasi-likelihoods will have the form:

$$n \log(\frac{E(N)}{\kappa+E(N)}) + \kappa \log(\frac{\kappa}{\kappa+E(N)})$$

with κ equals the above 5 coefficients.

Based upon the results of in the above table, we may conclude that the mean-variance relationship is quadratic if the prior is any of these continuous distributions. In the following section, we will use the uniform prior and the gamma prior to analyze animal data.

3.4 Application of the mixture model methods

Of the three datasets we have, namely the rat (exposed to NNM) liver foci, rat (exposed to dioxin) liver foci, and mouse (exposed to TPA) skin papilloma data, we will only use the NNM and the dioxin data. The reason is that for the mouse skin data, mice in different groups were exposed to the promoter (TPA) for various lengths of time, then the promoter was dropped. It was a study with changing doses of the test compound and any analysis of these data would require piecewise constant rates. This is left as a future extension of our current methods.

The parameter estimates along with their standard errors for these two models and for the Poisson model, i.e. the "wrong" model, are listed in Tables 3.2 to 3.4 for the NNM data and in Tables 3.5 to 3.7 for the dioxin data. Those tables alone do not tell us which model performs best, for we have no idea what the true value might be.

On the other hand, residual plots can reveal the goodness of fit by showing the residuals visually. Standardized residuals are plotted against the fitted mean of the number of PML's for the three models. The (standardized) residual plots (Figure 3.5 to 3.7) for the NNM rat liver foci data fitting the Poisson, PU and NB models show that the residuals of all three models are within the range (-2,4), similar to the range (-1,4) in the quasi-likelihood method (Figure 4.1). The residuals of the Poisson model (Fig.3.7) scatter widely between -2 and +4, which is reasonable since the data are over-dispersed. The residuals of the PU model (Fig.3.6), except 3 dots, scatter around the zero line between -2

and +2. The residuals of the NB model (Fig.3.5) are, in fact, have a similar pattern as those in the NB model but are not as widely scattered.

The residuals of the quasi-likelihood (QL) method have the smallest range: (-1,4). All but two dots scatter around the zero line between -0.5 and +2, which may indicate the best fit among the four models.

As for the dioxin data, most of the residuals are positive in the Poisson model, which indicates that the observed number of PML's are (much) larger than the expected values. This is not surprising since the data is obviously overdispersed. Generally speaking, the residuals of all of the three models (the Poisson, PU and NB models, Figures 3.8 to 3.10) are within the range (-1,3), same as the range (-1,3) in the quasi-likelihood method (Figure 4.2).

NNM(ppm)	α	s.e.(α)	γ	s.e.(γ)	δ	s.e.(δ)	μX
0	1.13	0.03	0.15	0.03	33.54	36.28	4.94
0.1	0.82	0.25	0.26	0.06	13.93	38.86	3.50
1	0.97	0.05	0.45	0.11	22.87	44.13	10.30
5	1.37	0.09	0.60	0.14	62.98	154.3	37.61
10	1.29	0.09	0.76	0.41	127.6	276.9	96.63
20	1.47	0.19	0.80	0.44	187.9	778.1	150.94
40	2.09	0.13	0.92	0.52	196.5	461.3	181.41

Table 3.2 NNM rat liver foci data fitting Negative binomial model. α is the birth rate, μ is the mutation rate and X is the number of normal cells per cubic centimeter assumed to be a constant. In this model, μX is assumed to be a gamma random variable, with shape parameter γ, and scale parameter δ.

NNM(ppm)	α	s.e.(α)	a	s.e.(a)	b	s.e.(b)	μX
0	1.29	0.89	0.10	0.03	41.26	19.65	20.69
0.1	4.59	0.31	0.10	0.05	31.28	1.51	15.69
1	5.15	0.86	0.10	0.13	99.09	6.64	49.59
5	4.27	19.79	82.2	0.01	82.2	0.01	82.20
10	8.74	4.35	116.5	0.01	116.5	0.01	116.5
20	1.63	0.67	152.4	0.01	152.4	0.01	152.4
40	1.99	0.78	164.5	0.33	209.9	0.48	187.2

Table 3.3 NNM rat liver foci data fitting Poisson-uniform model α is the birth rate, μ is the mutation rate and X is the number of normal cells per cubic centimeter assumed to be a constant. In this model, μX is assumed to be a uniform random variable, with parameters (a,b), $a < b$.

NNM (ppm)	α	s.e.(α)	μX	s.e.(μX)
0	0.5560	0.0003	3.6306	0.0772
0.1	0.7813	0.0001	4.1520	0.0012
1	0.8225	0.0002	11.7421	0.0988
5	1.1147	0.0001	57.7541	0.0030
10	1.3854	0.0002	61.5678	0.3605
20	1.5049	0.0001	137.8113	0.3307
40	2.1460	0.0002	147.8675	0.5251

Table 3.4 NNM rat liver foci data fitting Poisson model α is the birth rate, μ is the mutation rate and X is the number of normal cells per cubic centimeter assumed to be a constant. In this model, μX is assumed to be a constant.

TCDD (ppm)	α	s.e.(α)	γ	s.e.(γ)	δ	s.e.(δ)	μX
0	0.45	0.14	35.5	5.11	0.65	0.29	23.1
3.5	0.42	0.25	38.7	6.22	0.68	0.37	26.3
10.7	0.39	0.03	49.6	5.14	0.75	0.31	37.2
35.7	0.44	0.11	58.7	7.12	0.71	0.14	41.7
125	0.41	0.09	70.4	6.41	0.81	0.09	57.0

Table 3.5 TCDD(dioxin) rat liver foci data of Trischer et al (1991) fitting Negative binomial model. α is the birth rate, μ is the mutation rate and X is the number of normal cells per cubic centimeter assumed to be a constant. In this model, μX is assumed to be a gamma random variable, with shape parameter γ, and scale parameter δ.

TCDD (ppm)	α	s.e.(α)	a	s.e.(a)	b	s.e.(b)	μX
0	0.43	0.09	11.9	2.06	38.3	5.65	25.1
3.5	0.44	0.29	15.8	3.04	42.8	3.51	29.3
10.7	0.46	0.56	19.2	3.99	56.8	5.64	38.0
35.7	0.47	0.39	22.6	5.01	59.0	4.31	40.8
125	0.46	0.78	26.3	7.33	80.7	9.47	53.5

Table 3.6 TCDD(dioxin)-DEN rat liver foci data of Trischer et al (1991) fitting Poisson-uniform model α is the birth rate, μ is the mutation rate and X is the number of normal cells per cubic centimeter assumed to be a constant. In this model, μX is assumed to be a uniform random variable, with parameters (a,b), a < b.

TCDD (ppm)	α	s.e.(α)	μX	s.e.(μX)
0	0.4822	0.12	18.12	1.08
3.5	0.4769	0.09	22.34	2.01
10.7	0.4758	0.21	31.00	3.09
35.7	0.4811	0.11	34.75	4.77
125	0.4775	0.33	48.52	3.52

Table 3.7 TCDD (dioxin) rat liver foci data of Trischer et al (1991) fitting Poisson model α is the birth rate, μ is the mutation rate and X is the number of normal cells per cubic centimeter assumed to be a constant. In this model, μX is assumed to be a constant.

Chapter 4

4.1 Introduction to quasi-likelihood method

In the previous sections, we were trying to solve the extra-Poissonness problem via the classical mixture model approach, i.e., to find a closed form solution for the marginal distribution of the number of PMLs. However, the quasi-likelihood method introduced by Wedderburn (1974) provides another route to deal with the mixed Poisson problem. Wedderburn showed that often it is not necessary to make specific detailed assumptions regarding the random variation. Instead, it is necessary to know how the variance of each observation changes with its mean. Wedderburn (1974) states that "the form of the mean-variance relationship is often much easier to postulate (than the distribution); this is what makes quasi-likelihoods useful."

In other words, if we know the distribution of a variable, we can build a likelihood function and find the maximum likelihood estimates; if we don't know the distribution but we do know that there is a relationship between the variance and mean, we then can build a quasi-likelihood function and find the maximum quasi-likelihood estimates. From Table 2.1, it seems that the mean and the variance may follow a quadratic relationship:

$$V = \lambda + \rho\lambda^2.$$

In the plots in Figure 2.8 to 2.10, this relationship becomes more evident. This mean-variance relationship will be our foundation for the quasi-likelihood method.

4.2 Definition of the Quasi-likelihood function

4.2.1 Covariance functions

In this section, we use the problem faced in this dissertation to illustrate the definition of quasi-likelihoods. Let m be the number of subjects, N_i be the number of PMLs for subject i (i=1,2,....,m), $\lambda_i = \Lambda(t)$ be the expected value of N_i, and **N** be the response vector. Suppose that the components of **N** are independent with mean vector $\boldsymbol{\lambda}$ and covariance matrix $V(\mathbf{N})=\sigma^2\mathbf{V}(\boldsymbol{\lambda})$, where σ^2 may be unknown and $\mathbf{V}(\boldsymbol{\lambda})$ is a matrix of known functions.

It is assumed that the vector of parameters of interest $\boldsymbol{\delta}$, consisting of p parameters, are dependent upon a vector of covariates **x**, which is exposure time in our case. So we write $\boldsymbol{\lambda}(\boldsymbol{\delta})$ to absorb the covariates into the regression function. An important point is that σ^2 does not depend on $\boldsymbol{\delta}$. Since the components of Y are independent, the matrix $\mathbf{V}(\boldsymbol{\lambda})$ must be diagonal. One further assumption is required concerning the function $V_{ii}(\boldsymbol{\lambda})$, namely $V_{ii}(\boldsymbol{\lambda})$ must depend only on the i-th component of $\boldsymbol{\lambda}$. Thus we write

$$\mathbf{V}(\boldsymbol{\lambda}) = \text{diag}\{ V_{11}(\lambda_1), ..., V_{mm}(\lambda_m)\}$$

From now on we will use $V(\lambda)$ to denote the variance for a single response variable, $V_{ii}(\lambda_i)$, to simplify notation.

4.2.2 Construction of the quasi-likelihood function

Consider first a single component of the response vector **N**, which we write as n without subscripts. Under the conditions listed above, the function

$$U=\frac{n-\lambda}{\sigma^2 V(\lambda)}$$

has the following properties in common with a log-likelihood derivative:

$$E(U) = 0$$

$$V(U) = \frac{1}{V(\lambda)}$$

$$-E\left(\frac{\partial U}{\partial \lambda}\right) = \frac{1}{\sigma^2 V(\lambda)}$$

Since most first-order asymptotic theory connected with the likelihood function is founded on these three properties (McCullagh and Nelder, 1989), it is not surprising that the quasi-likelihood function for a single observation

$$Q(\lambda;n) = \int_n^{\lambda} \frac{n-t}{\sigma^2 V(t)} dt$$

should behave like a log-likelihood function for λ under the mild assumptions stated previously. Some examples of quasi-likelihoods for a number of common variance functions are given in Table 4.1 Many, but not all, of these quasi-likelihoods correspond to real log likelihoods for known distributions.

The mean-variance relationships in quasi-likelihood methods are diverse . It may be linear, quadratic, cubic, quadruple, or any other forms which are appropriate. For example, if the marginal distribution of N is inverse Gaussian (or Wald) distribution, then the mean-variance relationship is cubic. In fact, the marginal distribution does not have to be specified at all.

Often the mean-variance relationship in quasi-likelihood methods are determined by the characteristic of the data. For example, Wedderburn (1974) tried to model the incidence of leaf blotch. The relationship $V(N)=(E(N))^2(1-E(N))^2$ was adopted according to the property of the data, instead of the result of any known distribution.

The PML data in this dissertation are observed to have a quadratic mean-variance relationship. That is why we assume a quadratic relationship ($V= \lambda + \rho\lambda^2$), instead of linear or cubic, in the quasi-likelihood method. For future research, according to the nature of different data, we surely will assume mean-variance relationship other than quadratic where appropriate.

In this section, we use the problem faced in this dissertation to illustrate the definition of quasi-likelihoods. Let m be the number of subjects, N_i be the number of PMLs for subject i (i=1,2,....,m), $\lambda_i = \Lambda(t)$ be the expected value of N_i, and **N** be the response vector. Suppose that the components of **N** are independent with mean vector $\boldsymbol{\lambda}$ and covariance matrix $V(\mathbf{N})=\sigma^2\mathbf{V}(\boldsymbol{\lambda})$, where σ^2 may be unknown and $\mathbf{V}(\boldsymbol{\lambda})$ is a matrix of known functions.

It is assumed that the vector of parameters of interest $\boldsymbol{\delta}$, consisting of p parameters, is dependent upon a vector of covariates **x**, which is exposure time in our case. So we write $\boldsymbol{\lambda}(\boldsymbol{\delta})$ to absorb the covariates into the regression function. An important point is that σ^2 does not depend on $\boldsymbol{\delta}$. Since the components of Y are independent, the matrix $\mathbf{V}(\boldsymbol{\lambda})$ must be diagonal. One further assumption is required concerning the function $V_{ii}(\boldsymbol{\lambda})$, namely $V_{ii}(\boldsymbol{\lambda})$ must depend only on the i-th component of $\boldsymbol{\lambda}$. Thus we write

$$\mathbf{V}(\boldsymbol{\lambda}) = \text{diag}\{ V_{11}(\lambda_1), ..., V_{mm}(\lambda_m)\}$$

From now on we will use $V(\lambda)$ to denote the variance for a single response variable, $V_{ii}(\lambda_i)$, to simplify notation.

4.2.2 Construction of the quasi-likelihood function

Consider first a single component of the response vector **N**, which we write as n without subscripts. Under the conditions listed above, the function

$$U=\frac{n-\lambda}{\sigma^2 V(\lambda)}$$

has the following properties in common with a log-likelihood derivative:

$$E(U) = 0$$

$$V(U) = \frac{1}{V(\lambda)}$$

$$-E(\frac{\partial U}{\partial \lambda}) = \frac{1}{\sigma^2 V(\lambda)}$$

Since most first-order asymptotic theory connected with the likelihood function is founded on these three properties (McCullagh and Nelder, 1989), it is not surprising that the quasi-likelihood function for a single observation

$$Q(\lambda;n) = \int_n^{\lambda} \frac{n-t}{\sigma^2 V(t)} dt$$

should behave like a log-likelihood function for λ under the mild assumptions stated previously. Some examples of quasi-likelihoods for a number of common variance functions are given in Table 4.1 Many, but not all, of these quasi-likelihoods correspond to real log likelihoods for known distributions.

The mean-variance relationships in quasi-likelihood methods are diverse . It may be linear, quadratic, cubic, quadruple, or any other forms which are appropriate. For example, if the marginal distribution of N is inverse Gaussian (or Wald) distribution, then the mean-variance relationship is cubic. In fact, the marginal distribution does not have to be specified at all.

Often the mean-variance relationship in quasi-likelihood methods are determined by the characteristic of the data. For example, Wedderburn (1974) tried to model the incidence of leaf blotch. The relationship $V(N)=(E(N))^2(1-E(N))^2$ was adopted according to the property of the data, instead of the result of any known distribution.

The PML data in this dissertation are observed to have a quadratic mean-variance relationship. That is why we assume a quadratic relationship ($V= \lambda + \rho\lambda^2$), instead of linear or cubic, in the quasi-likelihood method. For future research, according to the nature of different data, we surely will assume mean-variance relationships other than quadratic where appropriate.

Distribution name	Variance function	Quasi-likelihood
Poisson	$V(N)=E(N)$	$n \log(E(N)) - E(N)$
Inverse Gaussian	$V(N)=(E(N))^3$	$-n/(2(E(N))^2) + 1/E(N)$
Negative binomial	$V(N)=E(N)+\frac{[E(N)]^2}{\kappa}$	$n \log(\frac{E(N)}{\kappa+E(N)})+\kappa \log(\frac{\kappa}{\kappa+E(N)})$
Binomial	$V(N)=E(N)(1-E(N))$	$n \log(\frac{E(N)}{1-E(N)}) + \log(1-E(N))$
Wedderburn*	$V(N)=(E(N))^2(1-E(N))^2$	$(2n-1)\log(\frac{E(N)}{1-E(N)})$ $-\frac{n}{E(N)} - \frac{1-n}{1-E(N)}$

Table 4.1: Mean-variance relationships in quasi-likelihood methods. This table is edited from Table 9.1 of McCullagh and Nelder (1984).* This special mean-variance relationship was used by Wedderburn (1974) to model the incidence of leaf blotch in 1965.

Since the components of **N** are independent by assumption, the quasi-likelihood for the complete data is the sum of the individual contributions. Let $Q_i(\lambda_i;n_i)$ be the quasi-likelihood function for the i-th subject, then the quasi-likelihood for the complete dataset is $Q(\boldsymbol{\lambda};\mathbf{n})$, where

$$Q(\boldsymbol{\lambda};\mathbf{n})=\sum_{i=1}^{n} Q_i(\lambda_i;n_i)$$

By analogy, the quasi-deviance function corresponding to a single observation is

$$D(n;\lambda)=-2\sigma^2 Q(\lambda;n)= 2\int_{\lambda}^{n} \frac{n-t}{V(t)}dt$$

which is strictly positive except at $n=\lambda$. The total deviance $D(\mathbf{n};\boldsymbol{\lambda})$, obtained by adding over the components, is a computable function depending on **n** and $\boldsymbol{\lambda}$ alone; it does not depend on σ^2.

4.2.3 Parameter estimation

The quasi-likelihood estimating equations for the regression parameters $\boldsymbol{\delta}$, obtained by differentiating $Q(\boldsymbol{\lambda};\mathbf{n})$ may be written in the form $U(\hat{\boldsymbol{\delta}})=0$, where $\hat{\boldsymbol{\delta}}$ are the estimates of $\boldsymbol{\delta}$ and

$$\mathbf{U}(\boldsymbol{\delta})=\mathbf{D}^T\mathbf{V}^{-1}(\mathbf{N}-\boldsymbol{\lambda})/\sigma^2$$

This is called the quasi-score function. In this expression the $(i,r)^{th}$ component of $\mathbf{D}$, of order p*p, where p is number of parameters, is $D_{ir}=\frac{\partial\lambda_i}{\partial\delta_r}$ (r=1,..p), the derivative of $\lambda(\delta)$ with respect to the parameters.

The covariance matrix of $\mathbf{U}(\boldsymbol{\delta})$, which is also the negative expected value of $\frac{\partial\mathbf{U}(\boldsymbol{\delta})}{\partial\boldsymbol{\delta}}$ is

$$\text{cov}(\mathbf{U}(\boldsymbol{\delta}))=\frac{\mathbf{D}^T\mathbf{V}^{-1}\mathbf{D}}{\sigma^2}$$

For quasi-likelihood functions, this matrix plays the same role as the Fisher information matrix for the ordinary likelihood function. Beginning with a reasonable initial value $\hat{\boldsymbol{\delta}}_0$ sufficiently close to $\hat{\boldsymbol{\delta}}$, the sequence of parameter estimates generated by the Newton-Raphson method with Fisher scoring is

$$\hat{\boldsymbol{\delta}}_1 = \hat{\boldsymbol{\delta}}_0 + (\hat{\mathbf{D}}_0^T\hat{\mathbf{V}}_0^{-1}\hat{\mathbf{D}}_0^T)^{-1}\hat{\mathbf{D}}_0^T\hat{\mathbf{V}}_0^{-1}(\mathbf{n}-\hat{\boldsymbol{\lambda}}_0)$$

$$\hat{\boldsymbol{\delta}}_i = \hat{\boldsymbol{\delta}}_{i-1} + (\hat{\mathbf{D}}_{i-1}^T\hat{\mathbf{V}}_{i-1}^{-1}\hat{\mathbf{D}}_{i-1}^T)^{-1}\hat{\mathbf{D}}_{i-1}^T\hat{\mathbf{V}}_{i-1}^{-1}(\mathbf{n}-\hat{\boldsymbol{\lambda}}_{i-1})$$

where $\hat{\mathbf{D}}_0$, $\hat{\mathbf{V}}_0$, and $\hat{\boldsymbol{\lambda}}_0$ are calculated from the initial guess $\hat{\boldsymbol{\delta}}_0$. The quasi-likelihood estimate $\hat{\boldsymbol{\delta}}$ may be obtained by iterating until convergence occurs. An important property of the sequence so generated is that it does not depend on the value of σ^2.

In all of the above aspects, the quasi-likelihood behaves just like an ordinary log likelihood. However, for the estimation of σ^2, the quasi-likelihood function does not behave like a log likelihood. The estimate of σ^2 is a moment estimator based on the residual vector $\mathbf{N}-\hat{\boldsymbol{\lambda}}$, namely

$$\hat{\sigma}^2 = \frac{1}{m-p}\sum_{i=1}^{m}(N_i - \hat{\lambda}_i)^2/V_i(\hat{\lambda}_i) = \frac{\chi^2}{m-p}$$

where χ^2 is the generalized Pearson statistic. In other words, we calculate the parameter estimates first, plug those in $\hat{\lambda}$, then use the above formula to compute $\hat{\sigma}^2$.

In our case, from the relationship $V = \lambda + \rho\lambda^2 = \lambda(1+\rho\lambda)$ shown previously in Figure 2.8 to 2.10, we can treat the number of PMLs as a Poisson random variable with a prior weight $\frac{1}{1+\rho\lambda}$, since the variance for each subject is equal to the specified variance, which is λ, divided by its prior weight. Once we get the estimates of α and μX, we subsequently obtain $\hat{\lambda}_i$, which is the expected mean for subject i. It is then easy to calculate $\hat{\rho}$. Namely, since $\rho = \frac{V-\lambda}{\lambda^2}$, so

$$\hat{\rho} = \frac{1}{m-p}\sum_{i=1}^{m}\frac{1}{\hat{\lambda}_i}\left(\frac{(N_i - \hat{\lambda}_i)^2}{\hat{\lambda}_i} - 1\right)$$

where m is number of subjects in that group, and p is the number of parameters.

The advantage of the above Quasi-likelihood is that the estimating equations for mean and variance parameters are strictly unbiased (Davidian & Carrol, 1988). This guarantees that the variance parameters are consistently estimated, provided that the model has been correctly specified.

4.3 Application of the quasi-likelihood method

Applying the quasi-likelihood method to the NNM and TCDD rat liver foci data, we obtain a much better fit than using the Poisson regression in Chapter 2. Comparing Table 4.2 of the quasi-likelihood method to Table 2.13 of the Poisson regression for the NNM data (and Table 4.3 to Table 2.14 for the dioxin data), it is clear that the scaled deviance of the former is much smaller than the latter, indicating a much better fit (McCullagh and Nelder, 1989).

NNM (ppm)	α	μX	ρ	Scaled deviance	d.f.
0	0.2573	14.9527	6.960	8.845	35
0.1	0.3423	12.3458	1.344	11.435	29
1	0.4727	20.9823	1.705	19.538	21
5	0.8585	21.4008	1.101	13.0089	15
10	1.1260	31.310	1.084	14.6740	14
20	1.1490	39.763	1.3733	16.8925	17
40	1.7110	46.658	1.2370	16.8250	17

Table 4.2: Analysis of PML's following exposure to NNM data of Moolgavkar et al. (1990) using the quasi-likelihood method, where α is the birth rate, μ is the mutation rate and X is the number of normal cells per cubic centimeter.

NNM (ppm)	α	μX	ρ	Scaled deviance	d.f.
0	0.441	27.33	73.54	8.445	8
3.5	0.452	30.45	69.77	7.102	7
10.7	0.494	39.34	71.33	8.411	7
35.7	0.611	42.11	82.7	5.198	6
125	0.602	57.44	99.8	5.144	7

Table 4.3: Analysis of PMLs following exposure to TCDD data of Tritscher et al (1992) using the quasi-likelihood method, where α is the birth rate, μ is the mutation rate and X is the number of normal cells per cubic centimeter.

To understand more of the use of this quasi-likelihood method, a simulation study is performed as described in the next chapter. Three kinds of data, i.e. Poisson, negative binomial, and Poisson-uniform data, are generated to simulate the behavior of the number of PML's.

Chapter 5 Simulation

5.1 Introduction to Simulation

From Chapter 1, it is clear that the two dimensional (2-D) mathematical equations derived in Moolgavkar's paper (1990) involve several numerical integrations. However, the maximum likelihood estimates generated varied with the chosen algorithms of numerical integration. To avoid the instability inherent with 2-D data, a simulation study using three dimensional (3-D) data is conducted to examine the operating characteristics of our solutions to the extra-Poisson problem.

Simulation, according to Naylor et al. (1966), is a numerical technique for conducting experiments on a digital computer. This involves certain types of mathematical and logical models that describe the behavior of business, economic, or biological systems over extended periods of real time. Although simulation is frequently viewed as a last resort to be adopted, recent advances in simulation methodologies, availability of software, and technical developments have made simulation a widely used and accepted tool in many research areas (Rubinstein,1981).

Simulation analysis can be justified for several reasons (Naylor et al, 1966). The experience of designing a computer simulation model itself may have even greater value than the actual simulation. And the knowledge gained in designing a simulation study often suggests changes in the system being simulated. The effects of these changes can then be tested via simulation before implementing them on the actual system. In addition, simulation can be used to experiment with new situations about which we have little or no information to enable us to prepare for what may happen.

The results obtained from simulation are much the same as observations or measurements that might have been made in the experiment. Many programming systems have been developed which incorporate simulation languages. Some of them are general-purpose in nature, while others are designed for specific types of systems.

However, simulation is not perfect. Though it is indeed a valuable and very versatile tool in those problems where analytic techniques are inadequate, it is, in fact, an imprecise technique. It only provides statistical estimates rather than exact results, and it only compares alternatives rather than identifying the optimal result. Simulation is also a slow way to study a problem. It usually requires a large amount of time for analysis and programming. Finally, simulation yields only numerical data about the performance of the system, and sensitivity analysis of the model parameters is very expensive. The only approach available is to conduct a series of simulation runs with different parameter values.

5.2 Statistical model

Four models are employed for this simulation: the Poisson model, the negative binomial (NB) model, the Poisson-Uniform (PU) model, and the quasi-likelihood (QL) model. Of the four models, the Poisson model assumes that the marginal distribution of N is really Poisson, the NB and PU models assume the mutation rate has its own distribution (so NB and PU are mixture models), and the QL model assumes a mean-variance relationship of the number of PML's. Each model can be described with its characteristic assumptions given in Table 5.1.

	Assumption	
Model	on the marginal distribution of # of PML's	on mutation rate
Poisson	Poisson	constant
Poisson Uniform	Poisson Uniform	Uniform
Negative binomial	Negative binomial	Gamma
Quasi-likelihood	None (mean-variance relationship)	None

Table 5.1 Assumptions of different models

5.2.1 Poisson model

In this section, the probability of the truncated number of foci in 3-D, the likelihood function for each subject, and finally the likelihood function and log-likelihood function for the entire group are computed.

(i) The number of observable foci (or PMLs) in 3-D

From section 1.4.5, for the Moolgavkar-Venzon-Knudson model, $g_3(r)$ is the density function for the radius r in 3-dimensions. According to formula (1.7) of that section,

$$g_3(r)dr = \frac{3[(\alpha/\beta)p(t)]^q}{r \ln(\frac{\beta}{\beta-\alpha p(t)})}dr ,$$

where α is the birth rate, β is the death rate, $q = (r/r_c)^3$ is the number of cells in one PML, r is the radius of one PML, and r_c is the radius of one liver cell assumed to be 12 μm.

Let r be the radius in 3-dimensions, and $G_3(r)$ be the distribution function for r. So

$$(5.1) \qquad G_3(r) = \int_0^r g_3(s)ds = \int_0^r \frac{3[(\alpha/\beta)p(t)]^q}{s\ \ln(\frac{\beta}{\beta-\alpha p(t)})}ds\ .$$

Now, let N(t) represent the number of observable foci in 3-D at time t. Suppose that a focus is observable when its radius is larger than ε and R is the upper bound on which we want to condition (see Chapter 1). Then N(t) is Poisson (Dewanji, 1989) with parameter $\Lambda_3(t)$

$$\Lambda_3(t) = [G_3(R)-G_3(\varepsilon)]\Lambda(t)$$

$$= \Lambda(t)[\int_\varepsilon^R g_3(r)dr]$$

Let $\int_\varepsilon^R g_3(r)dr = k$, the 3-D Poisson mean of the truncated N is

$$(5.2) \qquad \Lambda_3(t) = \mu X \Lambda k,$$

(ii). <u>The likelihood function for the entire group of subjects in 3-D</u>

Let n_i be the number of observable PML's for subject i at time t, i.e. $N(t) = n_i$. From (i), the likelihood function for subject i is

$$L_i = \frac{[\Lambda_3(t_i)]^{n_i}\exp(-\Lambda_3(t_i))}{n_i!}$$

So the likelihood function for the entire group with m subjects will be $L = \prod_{i=1}^{m} L_i$

The log-likelihood function for the entire group is given as

$$\text{LogL} = \sum_{i=1}^{m} n_i \log[\Lambda_3(t_i)] - \Lambda_3(t_i) - \log[n_i!]$$

5.2.2 Negative binomial model

According to the results in Chapter 3, the marginal PDF (probability density function) of the number of PML's under the negative binomial model for subject i is

$$P(N=n_i) = \binom{\gamma + n_i - 1}{\gamma - 1}(1-p)^{n_i} p^{\gamma}$$

where $\Lambda = \frac{1}{\alpha}\left[\ln\left(\frac{\beta}{\beta - \alpha p(t)}\right)\right]$, $\gamma + n = h$, $p = \frac{1}{1+\delta\Lambda}$. Therefore, the likelihood function for the entire group is

$$L = \prod_{i=1}^{m} P(N = n_i) = \prod_{i=1}^{m} \binom{\gamma + n_i - 1}{\gamma - 1}(1-p)^{n_i} p^{\gamma}$$

And the log-likelihood function is

$$\log \prod_{i=1}^{m} P(N = n_i) = \sum_{i=1}^{m} \log(P(N = n_i))$$

$$= \sum_{i=1}^{m} \log(n_i+\gamma-1)! - \log(\gamma-1)! + \gamma(\log(p)) + n(\log(1-p)),$$

which is the function to be maximized.

5.2.3 Poisson uniform model

The marginal PDF of the number of PML's for subject i is

$$P(N=n_i) = \frac{1}{(b-a)\Lambda}\left[\sum_{k=0}^{n_i} \frac{(\Lambda a)^k e^{-\Lambda a}}{k!} - \sum_{k=0}^{n_i} \frac{(\Lambda b)^k e^{-\Lambda b}}{k!} \right]$$

And the log-likelihood function is

$$\sum_{i=1}^{m} \log\left[\sum_{k=0}^{n_i} \frac{(\Lambda a)^k e^{-\Lambda a}}{k!} - \sum_{k=0}^{n_i} \frac{(\Lambda b)^k e^{-\Lambda b}}{k!} \right] - \log(b-a) - \log\Lambda - \log(k!)$$

5.3 Description of Simulation Study

The simulation study in this chapter provides a way of comparing the different marginal distributions described in Chapter 3 and the quasi-likelihood method described in Chapter 4 in terms of their ability to properly estimate parameters and variances. A 3-D dataset is generated, and the simulated data is used to establish a likelihood function. Our interest rests mainly with the ability of these models to estimate parameters with sample sizes similar to established experiments.

Several preliminary issues must be addressed before the simulation study is performed. The first requirement is that there must be an established method of generating data consistent with a specified distribution.

5.3.1 Random Variate Generation

The algorithms used to generate data under the above models were adapted from the call routines of random variate generators in SAS (SAS Language Guide, version 6.03). All of the random number generations originated with a linear congruential generator of deviates from the uniform distribution on the unit interval (Fishman and Moore, 1982).

5.3.1.1 Gamma random variates

One assumption is that the parameter μX has a gamma distribution with (γ, δ) as the shape and scale parameters. In the congruential routine which is used to generate gamma random variates, only the shape parameter is allowed to appear; in other words, it has to be a standard form with $(\gamma, 1)$ as the parameters. The generated gamma random variate is then multiplied by δ to transform back to the original gamma density function with parameters (γ, δ). If the shape parameter $\gamma > 1$, an acceptance-rejection method developed by Cheng (1977) is used. If $\gamma \leq 1$, an acceptance-rejection method developed by Fishman (1978) is then used.

5.3.1.2 Poisson random variates

We also assume that conditional on μX being a gamma r.v., N has a Poisson distribution. Since we assume mutation rate varies from animal to animal, for each μ generated, we employ the Poisson random variate generator routine to generate one Poisson r.v. with the mean shown above. If the Poisson mean $\Lambda_3(t) < 100$, an inverse transformation method applied to a uniform variable (generated from the SAS routine called RANUNI) is used. If $\Lambda_3(t) \geq 100$, the normal approximation of a Poisson random variable is used. In that case, the Box-Muller transformation of a uniform variate is used.

5.3.2 Experimental Design

Three kinds of data, Poisson , NB, and PU, are generated and each one of the four statistical models mentioned above are fit to these data. In order to cover a broader range of the birth rates and mutation rates, a design with a good combination of high and low birth rates and mutation rates is desired. Based upon biological considerations, the following combinations were chosen as reasonably representative: $(\alpha,\mu x)=(0.5, 10), (0.5, 200), (1.0, 10), (1.0, 200)$. Hence we have this low-low, low-high, high-low, and high-high design shown in Table 5.2. In all cases, the ratio of the death rate to the birth rate was chosen as $\beta/\alpha=0.99$.

α(birth rate)	μX (mutation rate times # of normal liver cells/cm^3)
0.5 (low)	10 (low)
0.5 (low)	200 (high)
1.0 (high)	10 (low)
1.0 (high)	200 (high)

Table 5.2 Birth rate/ mutation rate combination design

Model being fit	Model From Which Data is Generated	(α, μx) combinations
Poisson	Poisson	(0.5, 10), (0.5, 200), (1.0, 10), (1.0, 200)
	Negative binomial	(0.5, 10), (0.5, 200), (1.0, 10), (1.0, 200)
	Poisson uniform	(0.5, 10), (0.5, 200), (1.0, 10), (1.0, 200)
Negative binomial (NB)	Poisson	(0.5, 10), (0.5, 200), (1.0, 10), (1.0, 200)
	Negative binomial	(0.5, 10), (0.5, 200), (1.0, 10), (1.0, 200)
	Poisson uniform	(0.5, 10), (0.5, 200), (1.0, 10), (1.0, 200)
Poisson Uniform (PU)	Poisson	(0.5, 10), (0.5, 200), (1.0, 10), (1.0, 200)
	Negative binomial	(0.5, 10), (0.5, 200), (1.0, 10), (1.0, 200)
	Poisson uniform	(0.5, 10), (0.5, 200), (1.0, 10), (1.0, 200)
Quasi-likelihood (QL)	Poisson	(0.5, 10), (0.5, 200), (1.0, 10), (1.0, 200)
	Negative binomial	(0.5, 10), (0.5, 200), (1.0, 10), (1.0, 200)
	Poisson uniform	(0.5, 10), (0.5, 200), (1.0, 10), (1.0, 200)

Table 5.3 Experimental design for the simulation study

For any given data set, regardless of the model from which the data were generated, the MLE's for α and μX are obtained under the four birth rate/mutation rate structure in Table 5.2. Table 5.3 describes the layout of this simulation study.

For each of the 4x3x4=36 sets, 20 experimental animals are used since this number is routinely used in carcinogenesis studies of PML's. For each model-data set combination, the corresponding distributions are randomly sampled to generate the observed number of PML's. The resulting 48 sets constitute the observed data from which the MLE and related statistics are obtained. Fitting the models under these different assumptions makes it possible to investigate the impact of incorrect modelling on the statistical properties, e.g. biases and variances, of the MLE's for the parameters.

5.3.3 Sample size determination

The number of samples to simulate depends upon the desired power for testing bias. Each simulation can be viewed as an independent Bernoulli trial. The parameters of interest are μX and α. Because μX is assumed to have its own distribution instead of being a constant, we will focus on μX in sample size determination.

We can specify the minimum bias between the true μX and the estimated $\mu_0 X$ that we would be interested in finding. Suppose this difference is 5% of the true value. By specifying the minimum difference, we are saying that any difference less than 5% is too small to be of practical interest. The following formula is used to determine the minimum runs, n_s, of the simulation required:

$$n_s = \frac{(Z_\alpha - Z_\beta)^2}{(\mu X - \mu_0 X)^2/\sigma^2},$$

where α is the type I error, β is the type II error, Z_α is the value of Z that cuts off the α region of the sampling distribution of $\mu_0 X$ and Z_β is the value of Z that cuts off the β region of the sampling distribution of μX.

Based on the above formula, it is possible to calculate the number of replicates we need in this study. In one of the pilot simulation studies, the true value for μX is 10.414, and the standard deviation is 1.5374. If the percentage difference allowed is 5%, then α=0.05. Usually β is 4 fold of α, so β is 0.20. In this case, Z_α=1.645 and $-Z_\beta$ =-0.84

$$n_s = \frac{(1.645-(-0.84))^2}{(10.414*5\%)^2/(1.5374)^2} = 53.83.$$

The following is a summary of the percentage of difference allowed versus the sample size needed for α=0.05, β=0.20, $\sigma^2 = (1.5374)^2$, μX = 10.414.

% of difference allowed	sample size needed
5%	54
4%	85
3%	150
2%	337
1%	1346

Table 5.4 Sample size needed versus % of difference allowed based on μX

To assure that the difference between the estimates and the true values are under 5% if the model is correct, 50 samples will be used for each of the 48 groups in the experimental design. This number of sample is also tractable in the finite amount of time needed to finish this thesis.

5.3.4 Choice of maximization method

After building up the likelihood function, we need to maximize the function and get the MLE's of the parameters. Several optimization methods are available to find the MLE's.

The method of steepest descent is not recommended for general use because of its poor convergence properties (Walsh, 1975). The Newton-Raphson method has much better convergence properties, it works especially well if a good initial estimate of the optimal point can be found. However, it may fail to converge from a bad initial guess if the optimal point is far away. Also, the evaluation of the elements of the Hessian matrix and the inversion of this matrix, may pose formidable computational problems.

If the function to be optimized can be expressed as the sum of squares of functions, then special methods are available, e.g. the methods of Marquardt (Marquardt, 1963) and of Powell (Powell, 1965). These methods are expected to converge faster than the general purpose methods. They are derived from the "generalized least squares method". Marquardt use the first derivatives of the objective function, while Powell's method requires function evaluations only.

In this thesis, the method of Marquardt is chosen since the second derivative of the nonlinear function with respect to α is extremely messy.

5.3.5 Method of Moments

5.3.5.1 $\mu X \sim Gamma(\gamma,\delta)$

We will use the method of moments to decide on initial values for the parameters of interest. According to section 5.2, the 3-D expected value of truncated N is

$\Lambda_3(t)=\mu X\Lambda[\int_{60}^{500} g_3(r)dr]$. In Chapter 3, one assumption is that the distribution of N, conditional on the mutation rate being a gamma random variable, is a Poisson distribution. The marginal distribution of N is a negative binomial with mean and variance given by formula 3.2. These formulae are for the 3-D case neglecting truncation. If we take truncation into consideration, the formula 3.2 of Chapter 3 has to be modified. In Chapter 3, the 3-D Poisson mean of N is $\Lambda(t)$, and the 3-D Poisson mean of the truncated N is $\Lambda_3(t)$. Therefore, the marginal distribution of truncated N is still a negative binomial with mean and variance different from formula 3.2:

$$E(N)= \gamma\delta(\Lambda k)$$

$$V(N)= \gamma\delta(\Lambda k)(1+\delta(\Lambda k))$$

$$= \gamma\delta(\Lambda k) + \gamma(\delta\Lambda k)^2.$$

Now information from the real data can be used to find a reasonable estimate of γ and δ. Let $\bar{X}$ be the sample mean and S^2 be the sample variance of a specific dose group at a specific time interval of exposure. Then we have

$$\bar{X} = \gamma\delta\Lambda k$$

$$S^2 = \gamma\delta\Lambda k + \gamma(\delta\Lambda k)^2$$

By inverting these formulae, we can obtain the shape and scale parameters of the gamma density function from the data:

$$\frac{\bar{X}^2}{S^2-\bar{X}} = \hat{\gamma}$$

$$\frac{S^2-\bar{X}}{\bar{X}\Lambda k} = \hat{\delta}$$

where $k = [\int_{60}^{500} g_3(r)dr]$ can be obtained through numerical integration.

5.3.5.2 μX~Uniform(a,b)

Taking the truncation into consideration and following the results of Chapter 3, we determine the marginal mean and variance:

$$E(N) = \frac{(a+b)\Lambda k}{2}$$

$$V(N) = \frac{(\Lambda k)^2(a-b)^2 + 6(a+b)\Lambda k}{12}$$

$$= E(N) + [E(N)]^2 \frac{(a-b)^2}{3(a+b)^2}$$

From the sample we have

$$\bar{X} = \frac{(a+b)\Lambda k}{2}$$

$$S^2 = \frac{(\Lambda k)^2(a-b)^2 + 6(a+b)\Lambda k}{12}$$

So

$$(5.3)\quad \frac{S^2-\bar{X}}{\bar{X}^2} = \frac{(b-a)^2}{3(b+a)^2}$$

$$(5.4)\quad S^2-\bar{X} = \frac{(\Lambda k)^2(b-a)^2}{12}$$

Let $b-a=\Delta$, $b+a=a+2\Delta$. Divide (5.4) by (5.3):

$$\bar{X}^2 = \frac{\frac{(\Lambda k)^2(b-a)^2}{12}}{\frac{(b-a)^2}{3(b+a)^2}} = \frac{(\Lambda k)^2(a+2\Delta)^2}{4}$$

So we get $(a+2\Delta)^2 = \frac{4\,\bar{X}^2}{(\Lambda k)^2}$, which means

$$(5.5) \qquad a+2\Delta = \frac{2\,\bar{X}}{\Lambda k}$$

From (5.3), we get

$$\Delta^2 = (b-a)^2 = \frac{S^2-\bar{X}}{\bar{X}^2}\ 3\left(\frac{4\,\bar{X}^2}{(\Lambda k)^2}\right)=12\,\frac{S^2-\bar{X}}{(\Lambda k)^2}$$

which leads to

$$(5.6) \qquad \Delta = \frac{\sqrt{12(S^2-\bar{X})}}{\Lambda k}$$

So using (5.5) and (5.6), the moment estimates of a and b are

$$\hat{a} = \frac{2\,\bar{X} - \sqrt{12(S^2-\bar{X})}}{2\Lambda k}$$

$$\hat{b} = \frac{2\,\bar{X} + \sqrt{12(S^2-\bar{X})}}{2\Lambda k}$$

5.3.5.3 <u>μX is a constant, P(N) is Poisson</u>

If the marginal distribution of the number of PML's is really Poisson, then

$E(N) = V(N) = \mu X \Lambda k$

From the sample we have

$\bar{X} = \mu X \Lambda k$

So the moment estimate is

$$\frac{\bar{X}}{\Lambda k} = \hat{\mu} X$$

5.4 Simulation Study Results

5.4.1 Summary Statistics

For each of the 48 parameter sets based on the four (α, μX) combinations (0.5, 10), (0.5, 200), (1, 10), (1, 200), four models (Poisson, NB, PU, QL method) and three kinds of data (Poisson, NB, PU), we generated $n_s = 50$ samples, each sample consisting of 20 subjects. Thus, the data for each simulation consisted of information on n_i, i=1, ..., 20, where n_i is the number of observed PML's for subject i.

Let $\hat{\alpha}_L$ and $\hat{\mu X}_L$ denote the ML estimators of α and μX obtained from the L-th sample, L = 1, ..., n_s. Then, with mean $\hat{\alpha}_\cdot = \sum_{L=1}^{n_s} \frac{\hat{\alpha}_L}{n_s}$, we define

$$\text{Bias}(\hat{\alpha}) = (\hat{\alpha}_\cdot - \alpha)$$

and $$V_E(\hat{\alpha}) = \frac{1}{n_s - 1} \sum_{I=1}^{n_s} (\hat{\alpha}_\cdot - \alpha)^2$$

to be, respectively, the bias and variance of the ML estimator α. Completely analogous definitions for Bias ($\hat{\mu}X$) and $V_E(\hat{\mu}X)$ can be made. Note that $V_E(\hat{\alpha})$ given above is an estimate of the unknown exact variance of $\hat{\alpha}$. In practice, it is appropriate to use the diagonal element of the inverse of the observed sample information matrix to estimate the variability of $\hat{\alpha}$.

The validity of the maximum likelihood-based variance estimate depends on asymptotic theory. The asymptotic variance of $\hat{\alpha}$, $V_A(\hat{\alpha})$, is the appropriate diagonal element of the inverse of the average of the L sample information matrices for the specific likelihood under study. Hence, $V_A(\hat{\alpha})$ represents the asymptotic variance of $\hat{\alpha}$ under standard maximum likelihood theory, so $V_A(\hat{\alpha})$ is expected to be less than $V_E(\hat{\alpha})$ when the fitted and true models agree.

However, the asymptotic theory-based estimate of the variability of $\hat{\alpha}$ we use, i.e. $V_I(\hat{\alpha})$, is different from $V_A(\hat{\alpha})$. Rather than use the inverse of the average, we use the

average of the inverse, $V_I(\hat{\alpha})$. Namely, $V_I(\hat{\alpha})$ is the appropriate diagonal element of the average of the inverses of the L sample information matrices for the specific likelihood under study, so the inequality $V_I(\hat{\alpha}) > V_A(\hat{\alpha})$ should hold. Note that $V_I(\hat{\alpha})$ is a measure of the expected value of the variance estimator of $\hat{\alpha}$ employed in actual practice, and hence a comparison between $V_I(\hat{\alpha})$ and $V_E(\hat{\alpha})$ is important. Table 5.5 through 5.16 summarize the numerical results of this simulation study. Section 5.4.2 contains the interpretations of the patterns suggested by the entries in these tables.

$(\alpha, \mu X)$			Mean	Bias	V_E	V_I
(0.5, 10)	α	0.5	0.49867	-0.00133	0..0021123	0.0012479
	μX	10	10.067	0.067	3.4318	1.04784
(1.0, 10)	α	1.0	1.04067	0.04067	0.0024291	0.0010909
	μX	10	9.887	-0.113	0.1549	0.1195
(0.5, 200)	α	0.5	0.49967	-0.00033	0.0000167	0.0000036
	μX	200	200.580	0.580	18.8548	1.6425
(1.0, 200)	α	1.0	1.00812	0.00812	0.0040032	0.000001
	μX	200	198.000	-2.000	59.7905	0.1659

Table 5.5 Mean, bias and variance estimates from a simulation study in which we generate a Poisson number of pre-malignant lesions and estimate parameters in the two stage model using a Poisson-based likelihood.

NB parameters (γ, δ)	(α, μX)			Mean	Bias	V_E	V_I
(1,10)	(0.5, 10)	α	0.5	0.655333	0.155333	0.021683	0.000983
		μX	10	8.913509	-1.1086491	1.59396	0.190404
	(1.0, 10)	α	1.0	1.208329	0.208329	0.035789	0.000215
		μX	10	9.293415	-0.706585	3.063158	0.007032
(20,10)	(0.5, 200)	α	0.5	0.498551	-0.001449	0.000074	9.92E-8
		μX	200	202.1226	2.1226	170.603	4.380745
	(1.0, 200)	α	1.0	1.482667	0.482667	0.003614	0.001714
		μX	200	187.4489	-12.5511	0.732217	0.547566

Table 5.6 Mean, bias and variance estimates from a simulation study in which we generate a negative binomial number of pre-malignant lesions and estimate parameters in the two stage model using a Poisson-based likelihood.

PU parameters (a,b)	(α, μX)			Mean	Bias	V_E	V_I
(1,19)	(0.5, 10)	α	0.5	0.701498	0.201498	0.010521	3.283E-6
		μX	10	16.68733	6.687330	0.876107	2.015980
	(1.0, 10)	α	1.0	1.081333	0.081333	0.027755	0.001785
		μX	10	10.24000	0.240000	2.590270	0.269891
(1,399)	(0.5, 200)	α	0.5	0.492716	-0.00728	0.000755	6.506E-6
		μX	200	207.5432	7.5432	151.5268	4.765489
	(1.0, 200)	α	1.0	1.291998	0.291998	0.002638	0.000022
		μX	200	194.0800	-5.9200	0.263784	0.001340

Table 5.7 Mean, bias and variance estimates from a simulation study in which we generate a Poisson-uniform number of pre-malignant lesions and estimate parameters in the two stage model using a Poisson-based likelihood.

$(\alpha, \mu X)$			Mean	Bias	V_E	V_I
(0.5, 10)	α	0.5	0.504999	0.004999	0.000071	1.0E-8
	μX	10	10.199445	0.199445	0.0435715	2.5E-4
	a	1	1.049888	0.049888	0.004737	0.00101
	b	19	19.35001	0.35001	0.169549	1.0E-8
(1.0, 10)	α	1.0	0.999594	-0.000406	2.339E-7	4.356E-9
	μX	10	10.17421	0.17421	0.087105	0.0145585
	a	1	1.190531	0.190531	0.009308	0.000684
	b	19	19.15789	0.15789	0.289849	0.057550
(0.5, 200)	α	0.5	0.505001	0.005001	0.000072	0.000009
	μX	200	203.92594	3.92594	12.29132	0.07101
	a	1	1.111997	0.111997	0.0163813	0.00745
	b	399	406.7399	7.7399	49.1489	0.2766
(1.0, 200)	α	1.0	1.010004	0.010004	0.000286	0.000007
	μX	200	199.02745	-0.97255	0.641703	0.0097415
	a	1	1.310987	0.310987	0.012301	0.000236
	b	399	396.7449	-2.25510	2.55451	0.038730

Table 5.8 Mean, bias and variance estimates from a simulation study in which we generate a Poisson-uniform number of pre-malignant lesions and estimate parameters in the two stage model using a Poisson-uniform based likelihood.

NB parameters (γ,δ)	(α, μX)			Mean	Bias	V_E	V_I
(1,10)	(0.5, 10)	α	0.5	0.49195	-0.00805	0.00039	6.959E-6
		μX	10	9.84389	-0.15611	5.36139	4.855
		a	1	3.48374		3.0396	19.4
		b	19	16.2040		14.1789	0.02
	(1.0, 10)	α	1.0	0.96742	-0.03258	0.00688	0.00004
		μX	10	9.77607	-0.22393	3.94105	15.0003
		a	1	3.34767		2.26758	57.7236
		b	19	16.20446		11.1406	10.2780
(20,10)	(0.5, 200)	α	0.5	0.49765	-0.00235	0.00001	0.00051
		μX	200	211.3257	11.3257	198.529	1.93772
		a	1	141.2501		5589.36	7.75091
		b	399	281.4012		304.266	2.27E-23
	(1.0, 200)	α	1.0	0.99746	-0.00254	0.00001	0.00002
		μX	200	209.678	9.678	163.542	0.07711
		a	1	137.176		390.366	0.30844
		b	399	282.179		279.594	2.98E-17

Table 5.9 Mean, bias and variance estimates from a simulation study in which we generate a negative binomial number of pre-malignant lesions and estimate parameters in the two stage model using a Poisson-uniform based likelihood.

$(\alpha, \mu X)$			Mean	Bias	V_E	V_I
(0.5, 10)	α	0.5	0.76417	0.26417	0.00674	1.0076
	μX	10	10.1467	0.1467	2.7549	0.6751
	a	1	6.1569		1.4683	1.6864
	b	19	14.1364		9.6120	1.0180
(1.0, 10)	α	1	0.887575	0.112425	0.002882	1.007032
	μX	10	11.97465	1.97465	1.568152	0.55765
	a	1	11.71275		3.22144	1.21485
	b	19	12.23654		1.27964	1.01576
(0.5, 200)	α	0.5	0.35466	-0.14534	0.00025	1.0101
	μX	200	213.3062	13.3058	33.4775	0.50001
	a	1	213.3058		33.4321	1.00004
	b	399	213.3066		33.4229	1.00002
(1.0, 200)	α	1.0	0.637	-0.363	0.0086	1.01
	μX	200	209.48	9.48	23.59	0.60
	a	1	208.83		39.94	1.003
	b	399	210.13		27.52	1.000

Table 5.10 Mean, bias and variance estimates from a simulation study in which we generate a Poisson number of pre-malignant lesions and estimate parameters in the two stage model using a Poisson-uniform based likelihood.

$(\alpha, \mu X)$			Mean	Bias	V_E	V_I
(0.5, 10)	α	0.5	0.477333	-0.022667	0.005926	0.000127
	μX	10	9.900347	-0.099653	96.042176	0.1729189
	γ	1	1.011333	0.011333	0.022622	0.071460
	δ	10	9.800111	-0.199889	1.771812	2.419801
(1.0, 10)	α	1.0	0.962667	-0.037333	0.021147	0.000627
	μX	10	10.31587	0.31587	107.4906	0.7308224
	γ	1	0.998667	-0.001333	0.021072	0.055912
	δ	10	10.42004	0.42004	2.124430	1.307094
(0.5, 200)	α	0.5	0.482002	-0.017998	0.004573	2.485E-7
	μX	200	195.4065	-4.5935	1898.8818	12.581869
	γ	20	20.14667	0.14667	2.260224	2.636007
	δ	10	9.71333	-0.28667	2.47432	1.89600
(1.0, 200)	α	1.0	0.956000	-0.044	0.019259	0.000001
	μX	200	207.7813	7.7813	2150.4443	194.69358
	γ	20	20.27333	0.27333	2.897942	2.786961
	δ	10	10.30111	0.30111	1.097315	1.985874

Table 5.11 Mean, bias and variance estimates from a simulation study in which we generate a negative binomial number of pre-malignant lesions and estimate parameters in the two stage model using a negative binomial based likelihood.

PU parameters (a,b)	(α, μX)			Mean	Bias	V_E	V_I
(1,19)	(0.5, 10)	α	0.5	0.500992	0.000992	8.804E-6	0.003236
		μX	10	10.13659	0.13659	1.050531	4606.4394
		γ	1	3.299912		2.519087	2.754573
		δ	10	3.557992		1.388024	1672.288
	(1.0, 10)	α	1.0	0.95854	-0.04146	0.001515	0.000049
		μX	10	9.849287	-0.150713	1.791094	9744.4314
		γ	1	3.450974		2.789104	3.840942
		δ	10	3.278185		1.302284	2536.99
(1,399)	(0.5, 200)	α	0.5	0.500151	0.000151	0.000025	0.004822
		μX	200	199.0157	-0.9843	513.6135	2.4E16
		γ	20	2.151279		0.47079	1.852672
		δ	10	100.1951		821.73	129E14
	(1.0, 200)	α	1.0	1.010615	0.010615	0.001599	0.027356
		μX	200	196.8969	-3.11031	828.3489	5.8E267
		γ	20	2.222738		0.616399	1.135327
		δ	10	97.81957		1041.976	5.07E267

Table 5.12 Mean, bias and variance estimates from a simulation study in which we generate a Poisson-uniform number of pre-malignant lesions and estimate parameters in the two stage model using a negative binomial based likelihood.

$(\alpha, \mu X)$			Mean	Bias	V_E	V_I
(0.5, 10)	α	0.5	0.494878	0.005122	0.000389	1.000418
	μX	10	9.915591	0.084408	0.645444	5.55E571
	γ	1	7.854E12		8.252E26	4.59E285
	δ	10	0.059672		0.006802	1.21E286
(1.0, 10)	α	1.0	0.99212	0.00788	0.000645	1.002017
	μX	10	10.3856	0.3856	2.118092	2.6E303
	γ	1	8.884E12		1.499E27	3.26E304
	δ	10	0.045714		0.056798	0.07899
(0.5, 200)	α	0.5	0.49633	-0.00367	0.000207	0.000141
	μX	200	200.012732	0.012732	55031.97	183238.11
	γ	20	20425.34878		1.6769E8	3143883.6
	δ	10	0.103729		0.010782	0.058284
(1.0, 200)	α	1.0	1.000671	0.000671	0.097694	0.00137
	μX	200	204.5214	4.5214	657836.7	5.4561.77
	γ	20	2.184E14		3.544E9	6387666.4
	δ	10	0.524444		0.056798	0.07899

Table 5.13 Mean, bias and variance estimates from a simulation study in which we generate a Poisson number of pre-malignant lesions and estimate parameters in the two stage model using a negative binomial based likelihood.

$(\alpha, \mu X)$			Mean	Bias	V_E
(0.5, 10)	α	0.5	0.489833	-0.010167	4.275E-7
	μX	10	10.38937	0.38937	1.297382
	ρ	0	-0.01437	-0.01437	0.000112
(1.0, 10)	α	1.0	0.9229	-0.0771	1.825E-7
	μX	10	10.34314	0.34314	0.582157
	ρ	0	0.012844	0.012844	0.000426
(0.5, 200)	α	0.5	0.497782	-0.002218	6.875E-6
	μX	200	199.2648	-0.7322	106.5742
	ρ	0	0.016737	0.016737	0.017971
(1.0, 200)	α	1.0	0.989774	-0.010225	6.009E-7
	μX	200	196.9195	-3.0805	8.42416
	ρ	0	0.0152	0.0152	0.00124

Table 5.14 Mean, bias and variance estimates from a simulation study in which we generate a Poisson number of pre-malignant lesions and estimate parameters in the two stage model using a quasi-likelihood method.

PU parameters (a,b)	(α, μX)			Mean	Bias	V_E
(1,19)	(0.5, 10)	α	0.5	0.489713	0.010287	1.158E-7
		μX	10	10.64186	0.64186	0.311969
		ρ	0.27	0.284956	0.014956	0.00821
	(1.0, 10)	α	1.0	0.992174	0.007826	3.72E-6
		μX	10	10.178789	0.178789	3.228951
		ρ	0.27	0.299929		0.005879
(1,399)	(0.5, 200)	α	0.5	0.50025	0.00025	3.27E-6
		μX	200	198.968	-1.032	874.333
		ρ	0.33	0.350081		0.258487
	(1.0, 200)	α	1.0	1.041518	0.041518	3.146E-6
		μX	200	201.9577	1.9577	996.6004
		ρ	0.33	0.301191	-0.028809	0.293329

Table 5.15 Mean, bias and variance estimates from a simulation study in which we generate a Poisson-uniform number of pre-malignant lesions and estimate parameters in the two stage model using a quasi-likelihood method.

NB parameters (γ,δ)	$(\alpha, \mu X)$			Mean	Bias	V_E
(1,10)	(0.5, 10)	α	0.5	0.490353	0.009747	8.801E-7
		μX	10	9.556224	0.043776	2.605203
		ρ	1	1.076794	0.076794	0.607333
	(1.0, 10)	α	1.0	0.952548	0.047452	2.205E-6
		μX	10	10.49765	0.49765	7.420093
		ρ	1	0.939143	-0.060857	0.400584
(20,10)	(0.5, 200)	α	0.5	0.48573	-0.01427	7.216E-6
		μX	200	205.329	5.329	326.2709
		ρ	0.05	0.056176	0.006176	0.070244
	(1.0, 200)	α	1.0	0.9458976	-0.051024	2.454E-6
		μX	200	208.1029	8.1029	73.81437
		ρ	0.05	0.054735	0.0054735	0.048723

Table 5.16 Mean, bias and variance estimates from a simulation study in which we generate a negative binomial number of pre-malignant lesions and estimate parameters in the two stage model using a quasi-likelihood method.

5.4.2 Interpretations

We will first discuss the entries in Tables 5.5, 5.6, and 5.7, the cases in which the Poisson was fit to varying types of data. It can be seen that, in all four experimental design groups of Table 5.5, the use of Poisson data to fit the Poisson model produces negligible bias in the estimation of α and μX. The bias for μX in the $(\alpha, \mu X) = (1.0, 200)$ group is -2.00, which seems large at first glance. But since the true value for μX is 200, the difference is only 1%. So percentage-wise, this bias is still negligible.

When negative binomial data (Table 5.6) or Poisson-Uniform data (Table 5.7) are used to fit the Poisson model, noticeable bias is observed in all four groups especially for μX. The bias in Table 5.6 and 5.7 is much larger than the one observed in Table 5.5 as expected, because we use the extra-Poisson data to fit a Poisson model.

The shape parameter γ of the gamma distribution for the upper half of Table 5.6 is 1, and that for the lower half of Table 5.6 is 20 (note that overdispersion is inversely proportional to Y). As shown previously (Chapter 3 and Figure 3.1 to 3.4) as the γ increases (up to 9), the shape of the curve becomes similar to the normal probability density curve. Thus the lower half of Table 5.6 is better described as Poisson-normal data fitting Poisson model.

Recall that in Chapter 3, we showed that the marginal mean and variance of N has the following relationship if the mutation rate has a gamma distribution:

$$(3.3)\quad V(N) = E(N) + \frac{[E(N)]^2}{\gamma}$$

where γ is the shape parameter. The smaller the γ, the larger the variance, which means more dispersion. It is obvious that the bias of the cases with γ=1 is larger than that with $\gamma = 20$, except for one group. In other words, the more dispersed the data, the larger the bias since variance should equal the mean in a Poisson model.

Next, we will examine the cases in which the Poisson-uniform (PU) model is fit to the simulated data (Tables 5.8 to 5.10). In the Poisson-Uniform model, we assume that $\mu X \sim$ Uniform(a,b). So a and b, instead of μX ($\mu X = (a+b)/2$), enter the likelihood function. The bias section for a and b are empty in Tables 5.9 and 5.10 for the simulated data are generated based on the value of α and μX, not a and b.

When we use PU data (Table 5.8) to fit the PU model, almost every bias is within a 5% difference of the true value. When the underlying data is NB, the PU model (Table 5.9) results are also good: all of the bias percentages are within 5% (Table 5.19). Both NB and PU models have a quadratic mean-variance relationship; this may be the reason that the PU model is also good for the NB data.

On the other hand, when we use the Poisson data to fit the Poisson-uniform (PU) model (Table 5.10), it generates the biggest bias among the four approaches (P, PU, NB, QL, Table 5.17). Most of the bias percentages for the above situation are well over 5%, which is much more than their peers. Recall that in Chapter 3, the log likelihood we want to maximize in the PU model is

$$(3.9) \qquad \text{Log}(P(N=n)) = -\log(b-a) - \log\Lambda + \log\left(\sum_{k=0}^{n} \left[\frac{(\Lambda a)^k e^{-\Lambda a}}{k!} - \frac{(\Lambda b)^k e^{-\Lambda b}}{k!} \right] \right)$$

The PU model assumes that $\mu X \sim$ uniform(a,b), so when the data is really Poisson, the values of (a,b) will be very close or even identical. In this case, (b-a) and $\sum_{k=0}^{n} \left[\frac{(\Lambda a)^k e^{-\Lambda a}}{k!} - \frac{(\Lambda b)^k e^{-\Lambda b}}{k!} \right]$ are very small or zero. Therefore, when we want to maximize the log likelihood to obtain the MLE, the only thing we maximize is $-\log\Lambda$. That is to say, the response variable n (the number of PML's) is no longer in the log likelihood, so we get very poor information from the data. This explains the big bias percentage observed in Table 5.17.

Table 5.11 to 5.13 contain the analysis of the fits using the negative-binomial model. For a negative binomial model, μX should have the following properties:

$$E(\mu X) = \gamma\delta = \bar{X}$$

$$V(\mu X) = \gamma\delta^2 = S^2$$

So the moment estimates for γ and δ will be

$$\hat{\gamma} = \frac{\bar{X}^2}{S^2}$$

$$\hat{\delta} = \frac{S^2}{\bar{X}}$$

When the data are Poisson-Uniform, and $\mu X \sim$ Uniform (a,b), then

$$\hat{\gamma} = \frac{(\frac{a+b}{2})^2}{\frac{(b-a)^2}{12}} = \frac{3(a+b)^2}{(b-a)^2}$$

$$\hat{\delta} = \frac{\frac{(b-a)^2}{12}}{\frac{a+b}{2}} = \frac{(b-a)^2}{6(a+b)}$$

So when (a, b) = (1, 19), $(\hat{\gamma}, \hat{\delta})$ = (3.7,2.7), and when (a, b) = (1, 399), $(\hat{\gamma}, \hat{\delta})$ = (3.03,66.0).

Negative binomial data are used to fit the negative binomial model in Table 5.11. As expected, the bias for α and μX are small. When PU data are used to fit the model in Table 5.12, the bias for α and μX is also small, each one within 5%; the reason may be due to the fact that both the NB and the PU models have a quadratic mean-variance relationship.

For the upper half of Table 5.12, where (a, b) = (1, 19), the MLE for (γ, δ) are (3.30, 3.56) and (3.45, 3.28), close to the moment estimates (3.70, 2.72). As for the

lower half where (a, b) = (1, 399), the MLE for (γ, δ) are (2.15, 100.2) and (2.22, 97.82), which differ, especially in b, from the moment estimates (3.03, 66.00). When Poisson data are used to fit the model in Table 5.13, again, the bias for α and μX is very small, each within 5%. Since μX is a constant, S^2 should be zero, so

$$\hat{\gamma} = \frac{\bar{X}^2}{S^2} \doteq \infty$$

$$\hat{\delta} = \frac{S^2}{\bar{X}} = 0$$

Therefore we observe a very big $\hat{\gamma}$ and $\hat{\delta}$ close to zero for all four (α, μX) pairs in Table 5.13.

The variance estimates in Table 5.5 through 5.13 based upon asymptotic constructs (V_I) are smaller than the corresponding observed simulation variances (V_E) when the correct model is employed with only a few exceptions. The few exceptions may be due to the fact that $V_I > V_A$ (the asymptotic variance), so V_I is not the lower bound of V_E. If the model is incorrect, V_I can become very large due to the fluctuation of the values of the inverse of the L sample information matrices.

The results of fitting the QL method are given in Tables 5.14 to 5.16. The mean-variance relationship, instead of a specific assumption of the distribution, is the foundation of the QL method. So the results of the QL method using Poisson data (Table 5.14) should be comparable to fitting the Poisson model to Poisson data (Table 5.5). Using the QL method with PU data (table 5.15) should be comparable to fitting the PU model to PU data (Table 5.8), and applying the QL method to NB data (Table 5.16) is comparable to fitting the NB model to NB data (Table 5.11).

After comparing the above pairs, the estimates for α and μX are pretty close. As we take a closer look, it is clear that the bias for the QL method is slightly larger than that of the

previous three models. Numerical integration involved in the expected value of N (formula 5.2) may be the reason for this larger bias since the QL analysis is carried out in GLIM, a statistical package with only single precision, while the analysis of the other three models are conducted in SAS.

In terms of parameter estimation, we have the following conclusion. The Poisson model is only good if the data being analyzed is really Poisson. If the data is over-dispersed, then the adoption of the Poisson model will result in MLE's with significant bias. The Poisson-uniform model is good for NB data or PU data, but not good for the Poisson data due to the nature of the model. The negative binomial model and the quasi-likelihood methods appear to be the best in parameter estimation. These two models always give estimates with negligible bias to any of the three kinds of data used here.

Percentage Bias($\hat{\alpha}$)

(α, μX)	P	PU	NB	QL
(0.5, 10)	-0.266	-52.834	1.024	-2.033
(1.0, 10)	4.067	11.24	0.788	-0.710
(0.5, 200)	-0.066	-29.068	-0.734	-0.444
(1.0, 200)	0.812	-36.31	0.067	-1.023

Percentage Bias($\hat{\mu X}$)

(α, μX)	P	PU	NB	QL
(0.5, 10)	0.67	1.467	0.844	3.894
(1.0, 10)	-1.13	19.75	3.856	3.431
(0.5, 200)	0.29	6.6529	0.006	-0.366
(1.0, 200)	-1.00	4.74	2.261	-1.540

Table 5.17 Percentage bias estimate from a simulation study in which we generate a Poisson number of pre-malignant lesions and estimate parameters in the two stage model using 4 statistical approaches.

Percentage Bias($\hat{\alpha}$)

(α, μX)	P	PU	NB	QL
(0.5, 10)	40.299	1.000	0.198	2.057
(1.0, 10)	8.133	-0.041	-4.146	0.783
(0.5, 200)	-1.456	1.000	0.0002	0.0005
(1.0, 200)	29.199	1.000	0.030	0.093

Percentage Bias($\hat{\mu}$X)

(α, μX)	P	PU	NB	QL
(0.5, 10)	66.873	1.994	1.366	6.419
(1.0, 10)	2.400	1.742	-1.507	1.788
(0.5, 200)	3.772	1.963	-0.492	-0.516
(1.0, 200)	-2.962	-0.486	-1.555	-0.014

Table 5.18 Percentage bias estimate from a simulation study in which we generate a Poisson-uniform number of pre-malignant lesions and estimate parameters in the two stage model using 4 statistical approaches. (a,b)=(1,19) if μX=10, (a,b)=(1,399) if μX=200.

Percentage Bias($\hat{\alpha}$)				
(α, μX)	P	PU	NB	QL
(0.5, 10)	31.067	-1.61	-4.533	1.949
(1.0, 10)	-0.145	-3.26	-3.733	4.745
(0.5, 200)	41.665	-0.47	-3.599	-2.854
(1.0, 200)	48.267	-0.25	-4.400	-5.102

Percentage Bias($\hat{\mu X}$)				
(α, μX)	P	PU	NB	QL
(0.5, 10)	-11.087	-1.56	-0.997	0.438
(1.0, 10)	21.226	-2.24	3.159	4.977
(0.5, 200)	-0.353	5.66	-2.297	2.665
(1.0, 200)	-6.276	4.84	3.891	4.051

Table 5.19 Percentage bias estimate from a simulation study in which we generate a negative binomial number of pre-malignant lesions and estimate parameters in the two stage model using 4 statistical approaches. $(\gamma,\delta)=(1,10)$ if $\mu X=10$, $(\gamma,\delta)=(20,10)$ if $\mu X=200$.

Chapter 6 Directions

6.1 Recommendations

Based on the two-stage carcinogenesis model of Moolgavkar et al (1981), examination of the Poisson assumption of the number of PML's is taken and corrections are made and carried out using both animal data and simulated data.

The QL method and the NB-based likelihood almost always generate MLE's with small bias, whether the data is Poisson or over-dispersed Poisson, according to the results of our simulation study. The advantage of the QL method is that, instead of assuming any form of distribution, it only assumes a mean-variance relationship and consequently avoids mathematically complicated problems. This mean-variance relationship may be linear, quadratic, cubic or any form depending on the nature of the data (as described in Chapter 4). Statistically speaking, this is appealing because of its simplicity and flexibility. However, it does not hold any assumption about the birth rate or the mutation rate; it really makes no attempt to offer an explanation concerning the source of this over-dispersion.

On the other hand, the NB model, in addition to its capacity for parameter estimation, does try to point out the source of over-dispersion by putting a gamma prior on the mutation rate. The gamma distribution is so flexible that it can approximate the normal, the beta, the chi-square and the exponential distributions depending on the values of its shape and scale parameters. In other words, the gamma distribution covers a broad range of distributions. If the mutation rate is a continuous random variable, then the gamma distribution would be a likely choice to correct for over-dispersion and is also supported by the data analysis in Chapter 3.

The mixture models approach puts an assumption on the mutation rate (in this dissertation), the birth rate, or both (future research). This assumption may be correct or wrong, we have no idea. However, if the mean-variance relationship assumption of the QL method is made correctly (according to the data), the QL method can serve as a reference point. When one of the mixture models turns out to have the best fit (e.g. with the smallest residuals) or the closest result following parameter estimation via the quasi-likelihood method, it shows that its assumption may be approximately correct.

Above all, if the purpose of the study (of the over-dispersed PML's data with an underlying assumption of two-stage model of carcinogenesis) is merely parameter estimation, the QL method with an appropriate mean-variance relationship (of the data) is recommended. If, in addition to parameter estimation, the purpose is to try to explain the sources of over-dispersion, then the mixture model with the best fit is recommended. To summarize the above conclusion and recommendations, a "decision process" flow chart is given in Figure 6.1.

Having a PML dataset in front of us, the first question we want to ask is "Is it Poisson?" because the Poisson assumption concerning the number of PML's is fundamental in the two stage stochastic model of carcinogenesis. The index of dispersion test is used to determine whether the Poisson assumption is violated. We performed the test using the animal data but we did not do it using the simulated data since the data were generated to be over-dispersed Poisson or Poisson according to need. Thus, our simulation results are not directly applicable to this flowchart; they only suggest support.

It was found, in the animal data we studied, that almost all of the p-values are significantly smaller than 0.05. Thus, we rejected the null hypothesis, H_0, which states the mean and the variance are the same. There stands a chance that H_0 is true but we reject it and thus commit a type II error, α, or the alternative hypothesis H_A (mean and variance are

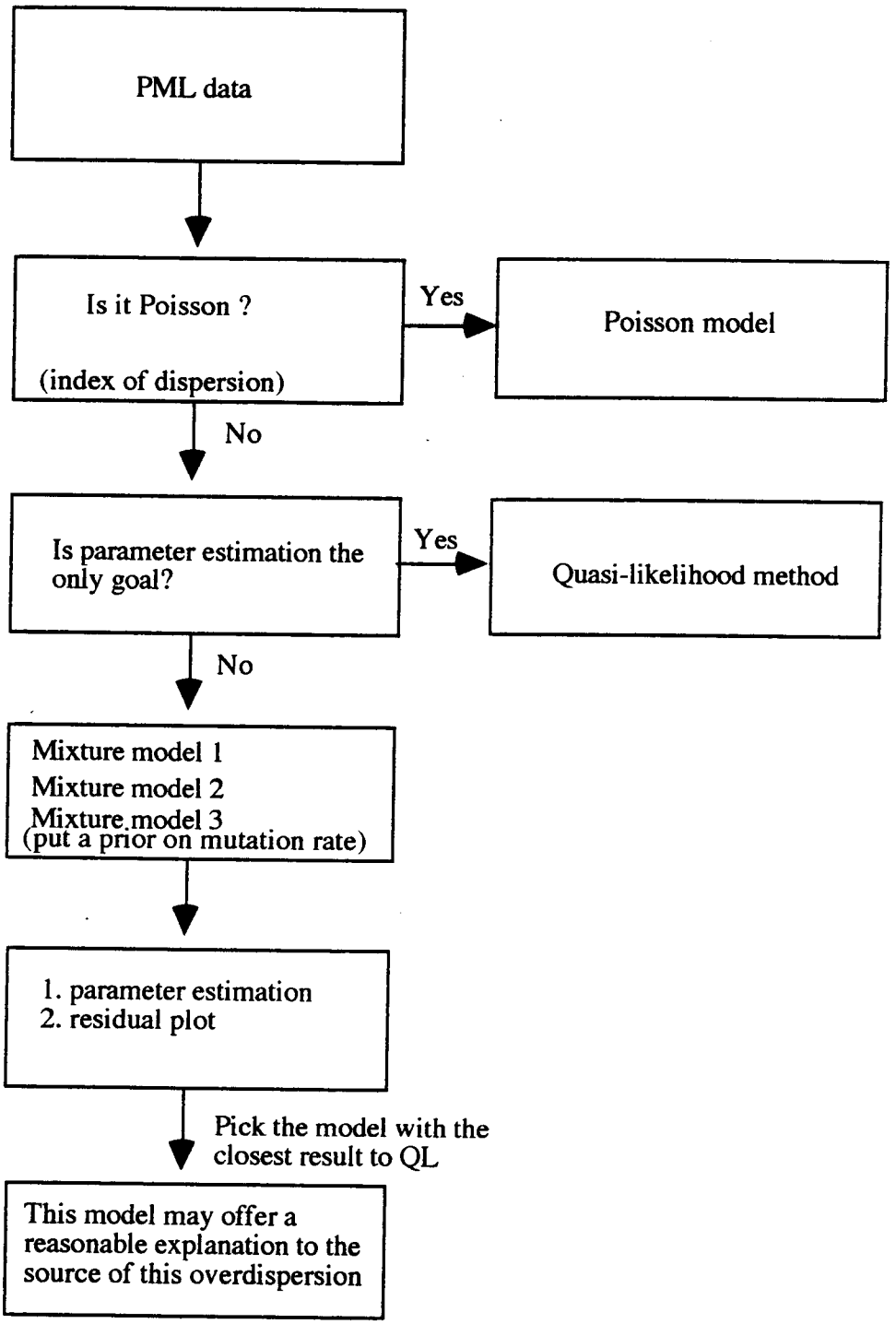


Figure 6.1 Flowchart of the decision process proposed for deciding on which method to use in the analysis of PML data.

not the same) is true but we accept H_0 and commit a type I error, β. Table 6.1 illustrates these situations:

	H_0 (true)	H_A (true)
H_0 (accept)	$1-\alpha$	β
H_0 (reject)	α	$1-\beta$

Table 6.1 Type I and Type II errors

If H_A is true but we accept H_0 (type I error, β), we treat truly overdispersed data as Poisson. This results in inefficient estimation and under-estimation of the variance of the model parameters will be likely (Margolin et al, 1981). If H_0 is true but we reject it (Type I error, α), then we are not likely to have a problem. The simulation study shows that when the data is Poisson, the QL method and the NB method generate parameters with small bias (within 5% difference). So even if we made a mistake from the index of dispersion test, we do not lose much except time and energy.

6.2 Size of the premalignant lesions (PML's)

In the previous chapters, our focus has been on the distribution of number of the premalignant lesions (PML's). Yet to fully understand the mechanistic model of carcinogenesis, the size of the PML's is also important. Therefore in Chapter 6.2, our focal point is on the size of the PML's for Initiation-Promotion (IP) studies.

In evaluating the size problem, the design of the standard IP experiment is considered, and the key treatment group for the usual IP protocol is illustrated in Figure 6.2.

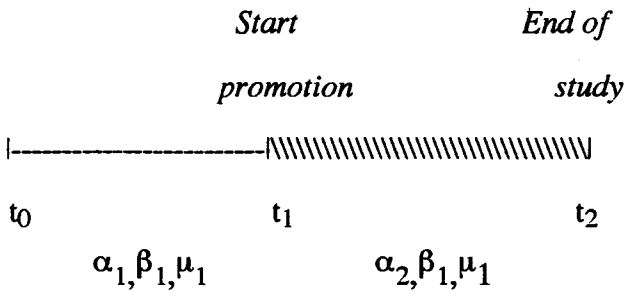


Fig. 6.2 Initiation-Promotion Protocol

In this protocol, promotion is started at time t_1 and continued until study termination at t_2. When the external environment changes, the birth rate will adjust accordingly. In the I phase, normal cells are mutated to initiated cells. After the promoter is administered, the growth of initiated cells get stimulated and proliferate more rapidly, which indicates that the birth rate of the initiated cells increases in the P phase. So the birth rate is assumed to be piece-wise constant: before the promoter is applied, the birth rate is α_1, but after the application, the birth rate increases to α_2. There may be a proportion, Π, of the animals sacrificed at t_1, but in general, all data will be collected at t_2.

First, assumptions must be made relating the biology to the mathematical model. It is assumed that only the birth rate of initiated cells α_1 will change during the course of this experiment. The death rate, β_1, and mutation rate, μ_1, will remain the same. The change in the birth rate occurs during the promotion phase, changing from the background rate of α_1 to a promoted rate of α_2. Two intervals are of interest; the I phase prior to promotion which we denote as $I_1 = [t_0, t_1]$ and the P phase during promotion denoted as $I_2 = [t_1, t_2]$. Let s be the occurrence time of the first mutation of a clone, and let t be the observation time. It is reasonable to assume that the first mutation takes place in the I phase and is observed in P phase, i.e. $s \in I_1, t \in I_2$. From Dewanji et al (1989), we know that for $s \in I_1, t \in I_2$, the probability of extinction is

(6.1) $$p(t,s) = \frac{\beta_1-\beta_1E_1-(\beta_1-\alpha_1E_1)p_2(t-t_1)}{\alpha_1-\beta_1E_1-(\alpha_1-\alpha_1E_1)p_2(t-t_1)} = \frac{A}{C}, \quad \text{where}$$

(6.2) $$E_1 = e^{-(\alpha_1-\beta_1)(t_1-s)},$$

(6.3) $$p_2(t-t_1) = \frac{\beta_1-\beta_1E_2}{\alpha_1-\beta_1E_2},$$

(6.4) $$E_2 = e^{-(\alpha_2-\beta_1)(t-t_1)}.$$

Let W(t,s) denote the number of cells in one focus at time t given the focus was formed at time s. The probability generating function, $\Phi_{w(t,s)}(u) = E[u^{W(t,s)}]$ is given by

(6.5) $$\Phi_{w(t,s)}(u) = \frac{A+Bu}{C+Du}, \text{ where}$$

(6.6) $$A = (\beta_1-\beta_1E_1)(\alpha_2-\beta_1E_2)-(\beta_1-\alpha_1E_1)(\beta_1-\beta_1E_2),$$

(6.7) $$B = -(\beta_1-\beta_1E_1)(\alpha_2-\alpha_2E_2)+(\beta_1-\alpha_1E_1)(\beta_1-\alpha_2E_2),$$

(6.8) $$C = (\alpha_1-\beta_1E_1)(\alpha_2-\beta_1E_2)-(\alpha_1-\alpha_1E_1)(\beta_1-\beta_1E_2),$$

(6.9) $$D = -(\alpha_1-\beta_1E_1)(\alpha_2-\alpha_2E_2)+(\alpha_1-\alpha_1E_1)(\beta_1-\alpha_2E_2)$$

The above result will be used to find the conditional probability distribution of W(t,s) given that the clone is detected: P[W(t,s) = q |W(t,s)>0] in the IP model.

By definition of a probability generating function, it is well known that:

(6.10) $$P[W(t,s) = q \,|W(t,s)>0] = \frac{1}{c!}\Phi^{(m)}_{W(t,s)>0}(u)\,|_{u=0}.$$

where $\Phi^{(m)}_{W(t,s)>0}(u) = \frac{\partial^q E[u^{W(t,s)}|W(t,s)>0]}{\partial u^q}$ is the c-th derivative with respect to u of $\Phi_{W(t,s)>0}(u) = E[u^{W(t,s)}|W(t,s)>0]$, which is the probability generating function of W(t,s) conditional on the clone not being extinct by time t.

Theorem 6.1: In the IP protocol, if the PML is formed in the initiation phase (time s) and observed in the promotion phase (time t), then the conditional density function of W(t,s) given this PML is not extinct at time t is :

$$(6.11)\quad P[W(t,s)=x \mid W(t,s)>0] = C[\frac{K}{A}p(t,s)][1-\frac{K}{A}p(t,s)]^{x-1},$$

where A,C, and p(t,s) are defined in (6.6), (6.8), (6.1) and we define $K = (\alpha_1-\beta_1)(\alpha_2-\beta_1)E_1E_2$.

Proof: This proof is by induction.

(I) Consider the case of x = 1. By definition, p(t,s) = P[W(t,s) = 0]. Using (6.5), it follows that

$$(6.12)\quad \Phi_{w(t,s)>0}(u) = \frac{\phi_{W(t,s)}(u)-p(t,s)}{1-p(t,s)} = \frac{\frac{A+Bu}{C+Du} - p(t,s)}{1-p(t,s)}.$$

The derivative with respect to u of $\Phi_{w(t,s)>0}(u)$ evaluated at u=0 is

$$\psi_{t,s}^{(1)}(u)\Big|_{u=0} = \frac{\frac{\partial}{\partial u}(A+Bu)/(C+Du)\Big|_{u=0}}{1-p(t,s)}$$

$$= \frac{1}{1-p(t,s)}\frac{B(C+Du)-D(A+Bu)}{(C+Du)^2}\Big|_{u=0}$$

$$= \frac{1}{1-p(t,s)}\frac{BC-AD}{(C+Du)^2}\Big|_{u=0}$$

$$(6.13)\quad = \frac{1}{1-p(t,s)}\frac{BC-AD}{C^2}\Big|_{u=0}$$

Using (6.1), it follows that

$$1\text{-}p(t,s) = 1 - \frac{A}{C} = \frac{C\text{-}A}{C}$$

$$= \frac{(\alpha_2-\beta_1E_2)(\alpha_1-\beta_1E_1-\beta_1+\beta_1E_1)+(\beta_1-\beta_1E_2)(\beta_1-\alpha_1E_1-\alpha_1+\alpha_1E_1)}{C}$$

$$(6.14) \quad = \frac{(\alpha_1-\beta_1)(\alpha_2-\beta_1)}{C}$$

Hence from (6.13), (6.14), and (6.1) it follows that

$$\psi_{t,s}^{(1)}(u)\Big|_{u=0} = \frac{C}{(\alpha_1-\beta_1)(\alpha_2-\beta_1)}\frac{BC\text{-}AD}{C^2}$$

$$= \frac{1}{(\alpha_1-\beta_1)(\alpha_2-\beta_1)}(B - \frac{A}{C}D)$$

$$(6.15) \quad = \frac{1}{(\alpha_1-\beta_1)(\alpha_2-\beta_1)}(B - p(t,s)D)$$

Using (6.1), (6.7), and (6.9), after some algebra, it is found that

$$(6.16) \quad B\text{-}p(t,s)D = (\alpha_1-\beta_1)^2(\alpha_2-\beta_1)^2E_1E_2$$

Therefore $\psi_{t,s}^{(1)}(u)\Big|_{u=0} = (\alpha_1-\beta_1)(\alpha_2-\beta_1)E_1E_2$. This matches the Thm.6.1 for x = 1.

Now suppose Thm 6.1 holds for $j \geq 1$. Then

$$\psi_{t,s}^{(j)}(u)\Big|_{u=0} = (j!)C[\frac{K}{A}p(t,s)][1-\frac{K}{A}p(t,s)]^{j-1}$$

Using this, we shall deduce the corresponding result for x = j + 1.

$$\psi_{t,s}^{(j+1)}(u)\Big|_{u=0} = \frac{\partial(BC\text{-}AD)/(1\text{-}p(t,s))\ (j!)(-D)^{j-1}(C+Du)^{-(j+1)}\Big|}{\partial u}_{u=0}$$

$$= \frac{BC\text{-}AD}{1\text{-}p(t,s)}(j)!(-D)^{j-1}(D)(-(j+1))(C+Du)^{-(j+2)}\Big|_{u=0} = (j+1)\frac{-D}{C}\psi_{t,s}^{(j)}(u)\Big|_{u=0}$$

Using (6.8), (6.9) and the definition of K in Thm.6.1, we found that $-D = C - K$ after some algebra. So

$$\psi_{t,s}^{(j+1)}(u)\Big|_{u=0} = (j+1)\,\frac{C-K}{C}\,\psi_{t,s}^{(j)}(u)\Big|_{u=0} = (j+1)\left(1-\frac{K}{C}\right)\psi_{t,s}^{(j)}(u)\Big|_{u=0}$$

$$= (j+1)\left[1-\frac{K}{A}p(t,s)\right]\psi_{t,s}^{(j)}(u)\Big|_{u=0}$$

$$= (j+1)!\,C\left[\frac{K}{A}p(t,s)\right]\left[1-\frac{K}{A}p(t,s)\right]^{j}$$

Thus, by induction, Thm.6.1 is proved to be true. Q.E.D.

6.3 Future research direction

One topic of interest in the future is to assume that the birth rate is a random variable. In this dissertation, only the mutation rate is assumed to be a random variable, while the birth rate is assumed to be a constant. In fact, the birth rate may also be a random variable which differs from individual to individual. We may consider three situations in the future:

	Assumption		
Category	Birth rate	Mutation rate	Conditional dist'n of N
1	constant	random	$N\|\mu \sim$ Poisson
2	random	constant	$N\|\alpha \sim$ Poisson
3	random	random	$N\|\mu,\alpha \sim$ Poisson

Table 6.2 Priors on the birth rate, the mutation rate, or both

There are three major assumptions in Table 6.2. In each assumption, a set of prior distributions (e.g. gamma, uniform, or log-normal) are set on the birth rate, the mutation rate, or both. This dissertation focused on comparing various priors and the quasi-likelihood method within category 1, i.e. assuming the birth rate is a constant and the mutation rate is a random variable. It was also intended that we put a prior on the birth rate. However, to obtain the marginal distribution of N, α has to be integrated out. This integration turned out to be messy and complicated because the form of the expected value of N conditional on α had the form

$$\Lambda(t)=\frac{\mu X}{\alpha}\left[\ln\left(\frac{\beta}{\beta-\alpha p(t)}\right)\right],$$

where $p(t)=(\beta-\beta e^{-(\alpha-\beta)t}) / (\alpha-\beta e^{-(\alpha-\beta)t})$. It is easier to integrate out randomness in μ because it is independent of the rest of $\Lambda(t)$. It is much more difficult to integrate out α since it is so entangled in $\Lambda(t)$. Nevertheless, the quasi-likelihood method will still serve as a good reference point to compare the parameter estimates since no assumption is made with respect to the birth rate or the mutation rate, only a mean-variance relationship is assumed. For future research, once we obtain the marginal distribution of N with different priors for each category, we will compare various statistical approaches within and between these categories.

Another topic of interest is to include size of PML's in the estimation of two-stage model parameters. In this dissertation, we focused on N, the number of PML's. Thus, only the probability function of N is in the likelihood function to be maximized. In addition, the size and number of PML's may be correlated. Using the rat (exposed to NNM) liver foci data (Moolgavkar, 1990), we observed the correlations in Table 6.3. Though the size and the number are not highly correlated, they are correlated to some extent (+0.5933). This positive correlation is understandable because the more PML's one has, the more probable it is to have large PML's.

NNM	Pearson's correlation	P-value
0	-0.1824	0.4989
0.1	0.0180	0.9416
1	0.4932	0.0443*
5	0.4988	0.0827
10	0.3230	0.1911
20	0.5933	0.0074*
40	0.2521	0.3129

Table 6.3: Correlation between the size and the number of PML's of rat liver exposed to NNM. * P-value < 0.05

Once we include both the size and the number of PML's, one problem that needs to be worked out is the distribution function of the size of PML's when N does not have a Poisson distribution. If N is truly Poisson, then the distribution of the size of PML's is derived in Dewanji et al (1989). However, when N is not Poisson, the distribution of the size of PML's may not be the same. It needs to be corrected accordingly and incorporated into the likelihood function.

Finally, these methods should be extended to deal with piecewise constant rates. This will allow us to apply them to data from studies with changing doses of the test compound. Methods exist for the cases in which the rates are assumed to change in intervals of time (not continuously) with the application or withdrawal of the promoter. For example, the probability distribution and extinction probabilities for a birth-death process with piecewise

constant rates have been derived, used to model various biological phenomena, and applied to multistage models of carcinogenesis incorporating clonal expansion (Kopp-Schneider, 1992). Incorporating the piecewise constant rates in the design of Table 6.1 may be laborious to do because of the difficulties mentioned in the previous paragraphs. However, it is well worth doing for we will get a fuller picture of the model of carcinogenesis.

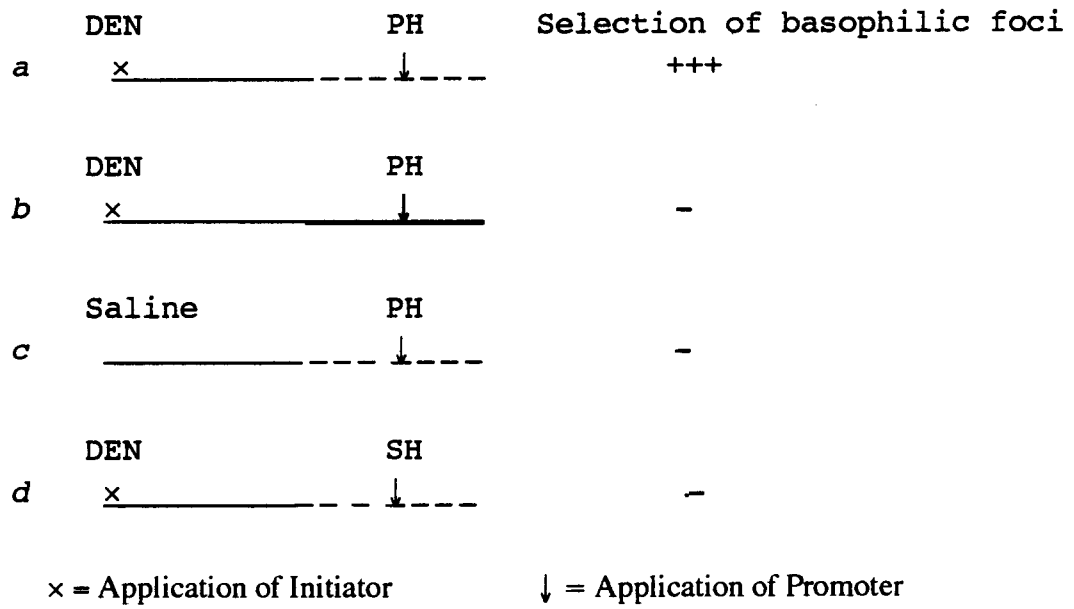


Fig. 1.1 Solt and Farber's assay procedure: An initiator (DEN) or control compound (Saline) is injected to animal soon after birth. Following this exposure, the animal is exposed to a promoter (2-AAF) and followed a growth stimulus (PH or partial hepatectomy). The main experimental group is a and b,c,d are controls (SH= Sham hepatectomy) _ _ _basal diet; ____basal diet plus 2-AAF.

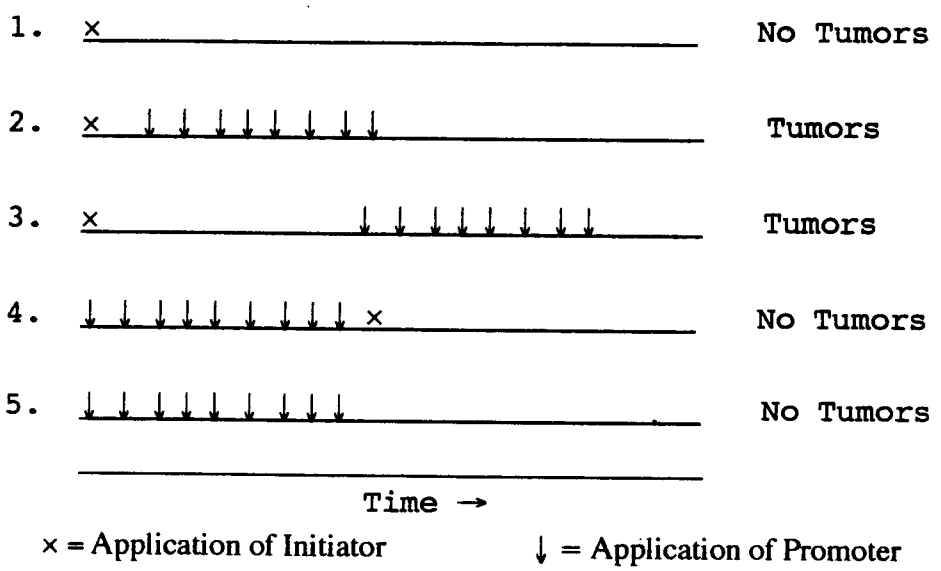


Fig. 1.2 Modern IP protocol for mouse skin: classic outline experiments demonstrating the two-stage natural history of carcinogenesis as exemplified primarily in skin.

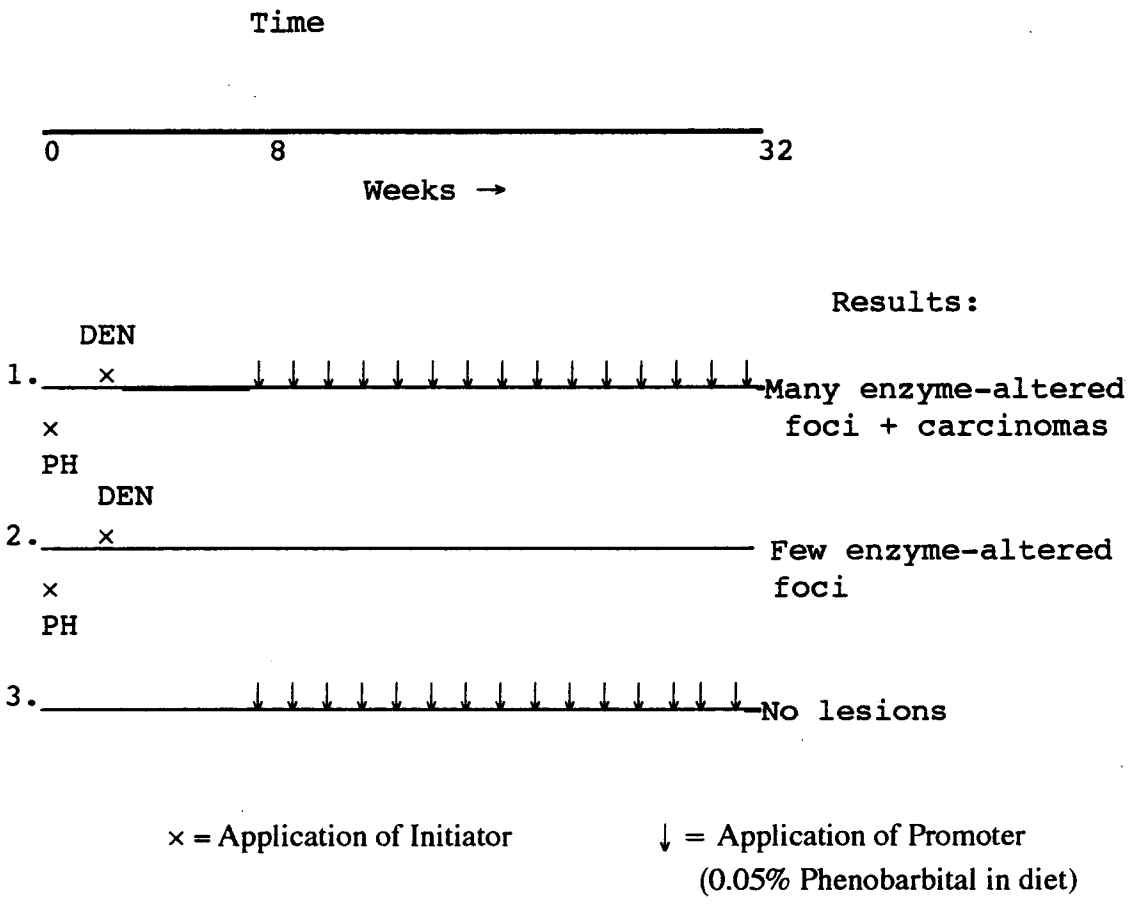


Fig. 1.3 Modern IP protocol for rat liver: Outline of experiments demonstrating a two-stage process in the natural history of hepatocarcinogenesis. Animals were partially hepatectomized at 0 time followed within 24 hours by the administration of DEN at the dose shown. Eight weeks later two groups of animals were placed on a diet containing 0.05% Phenobarbital, whereas the remaining group (group 2) was maintained on the same diet in the absence of the drug.

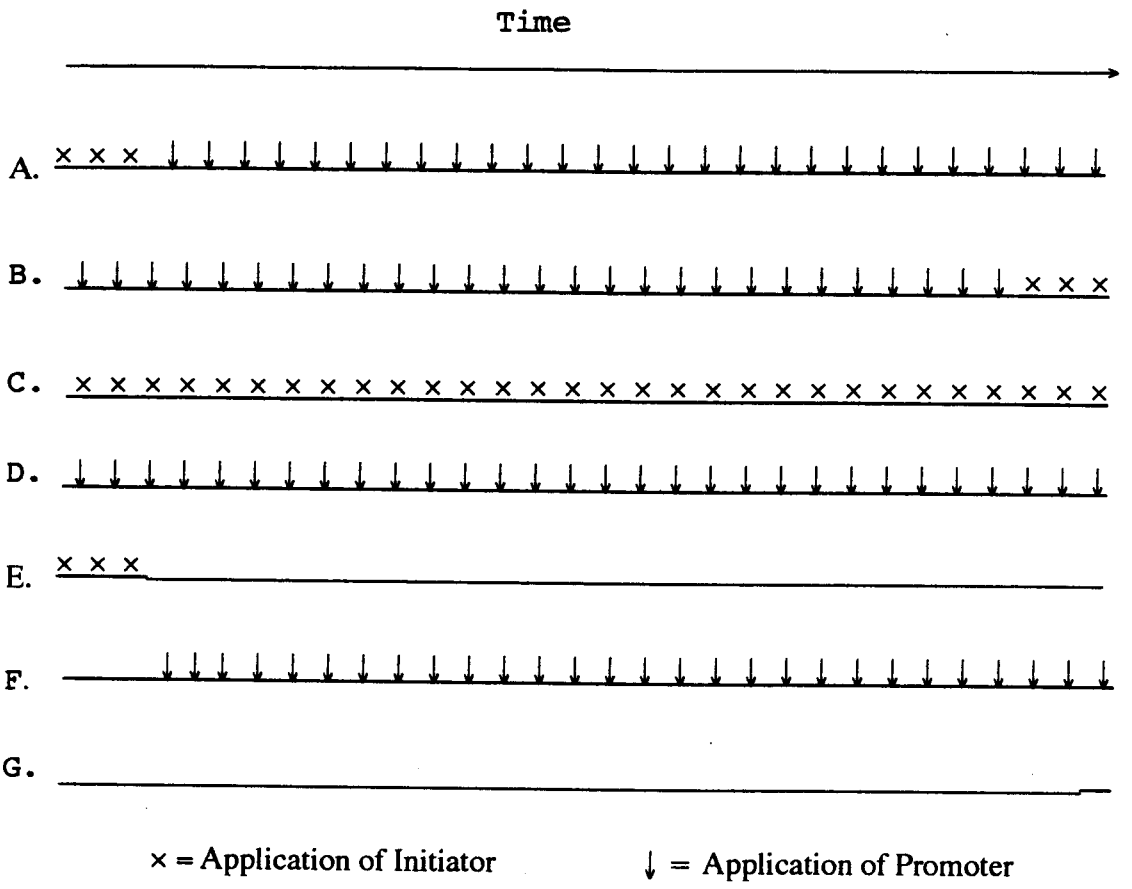


Fig.1.4 The design of Initiation-Promotion Studies

A. Initiation/Promotion

B. Reverse order

C. Lifetime Initiation

D. Lifetime Promotion

E. Initiation/ No promotion

F. No Initiation/ Promotion

G. Unexposed

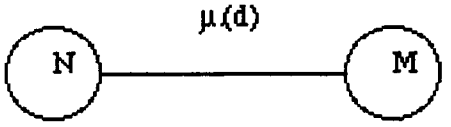


Fig. 1.5 The one-hit, one-stage model of Iversen and Arley (1952)

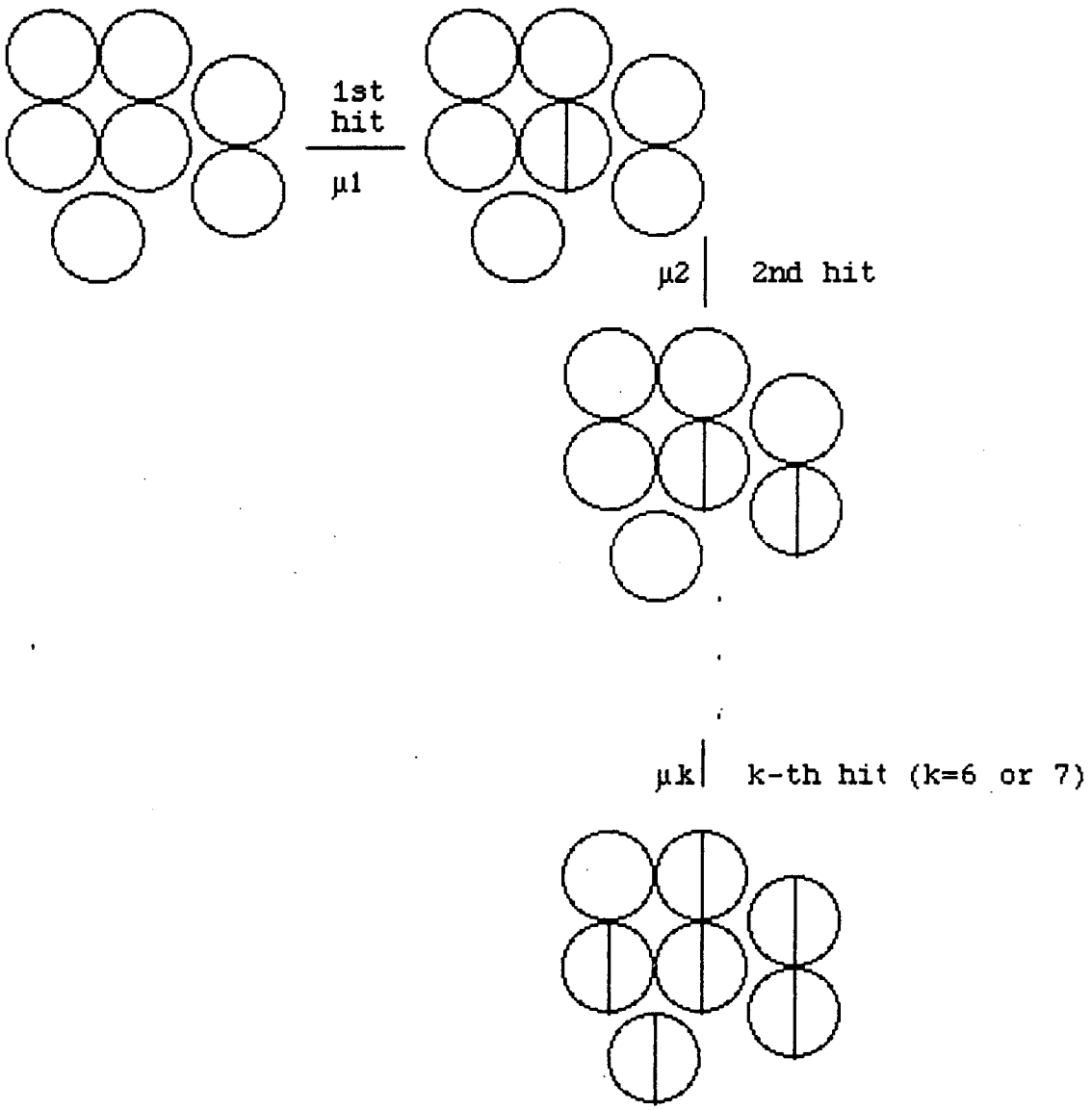


Fig.1.6 The multicell theory of carcinogenesis proposed by Fisher and Holloman (1954)

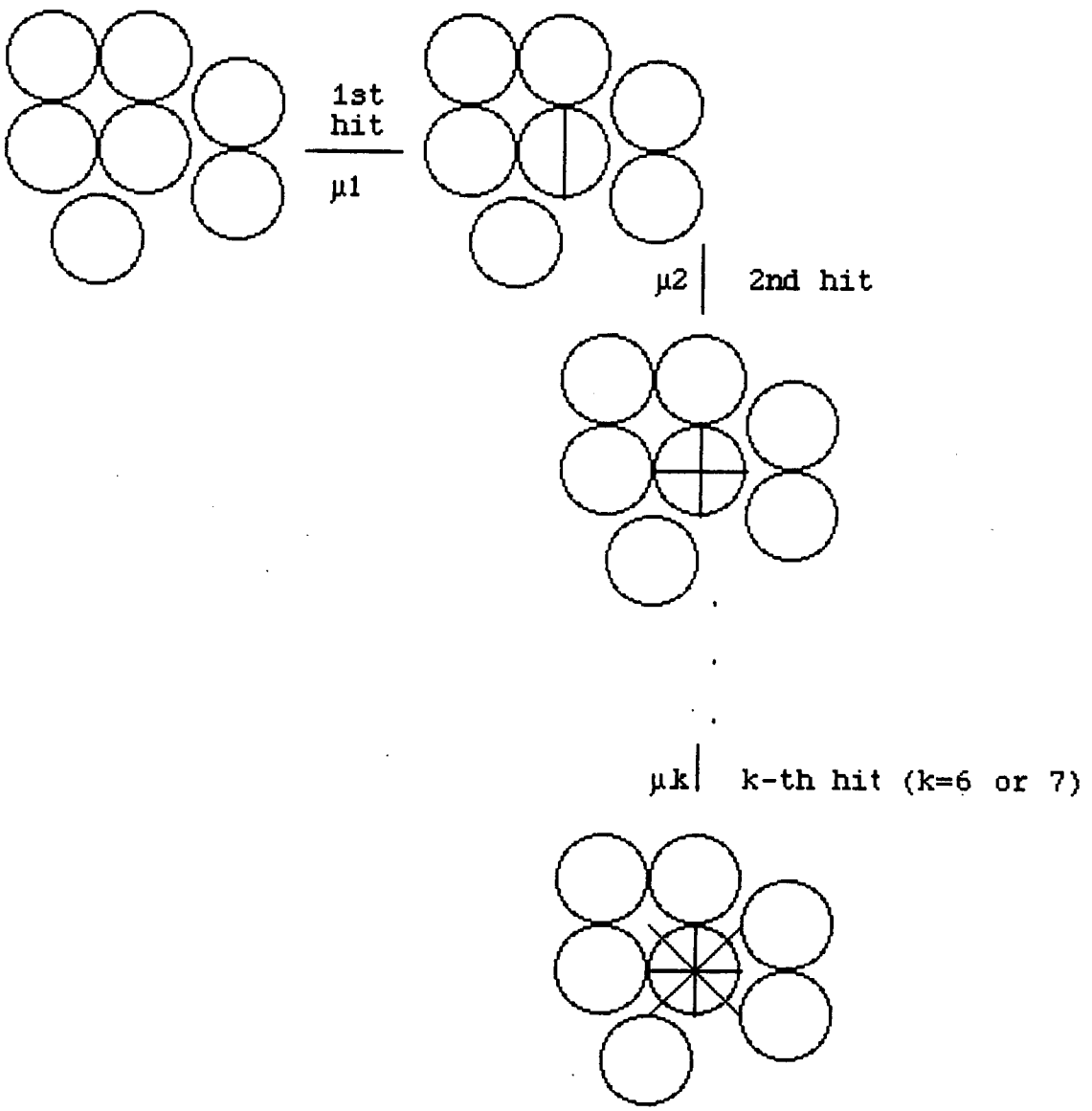


Fig 1.7 The multistage theory of carcinogenesis proposed by Stocks and Nordling (1954)

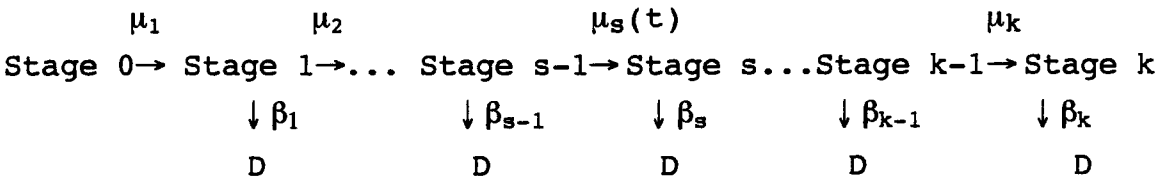


Fig.1.8.1 The multistage theory of carcinogenesis proposed by Armitage and Doll(1954)

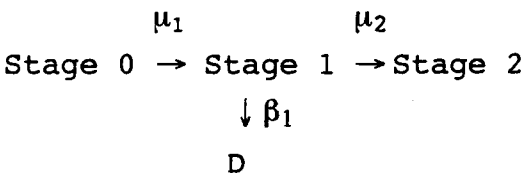


Fig.1.8.2 The two-stage theory of carcinogenesis proposed by Armitage and Doll(1957)

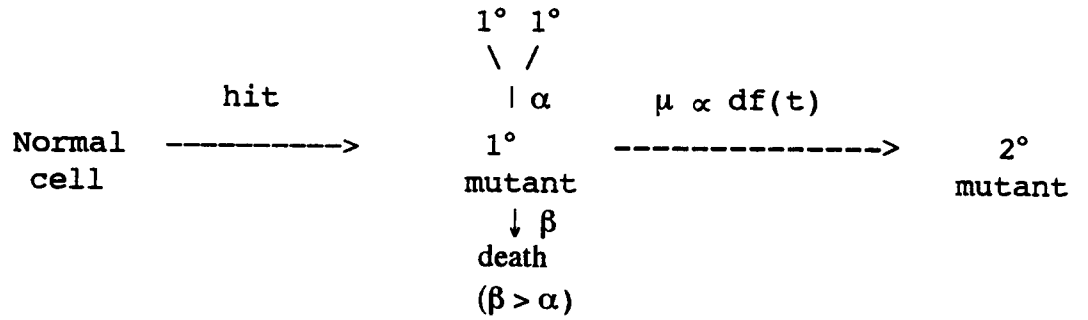


Fig.1.9 The One-hit, two stage model of Neyman and Scott (1954)

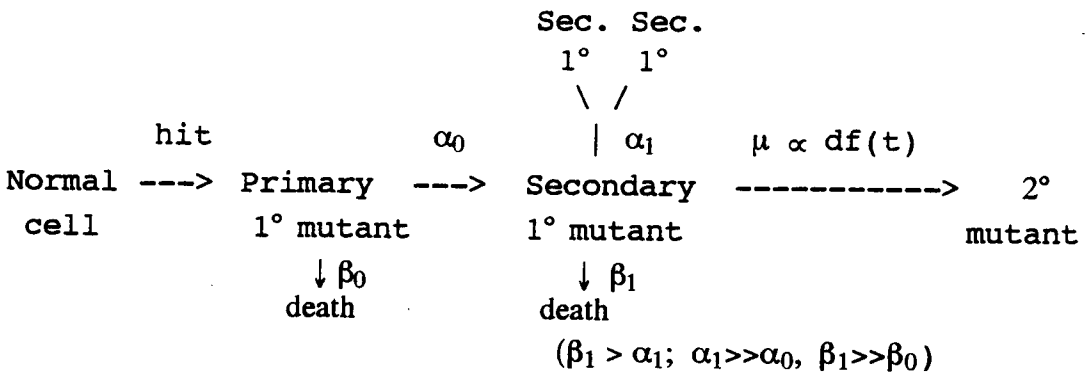


Fig.1.10 The one-hit, three-stage model of Neyman and Scott (1954)
Knudson(M & V,1979; M & K,1981)

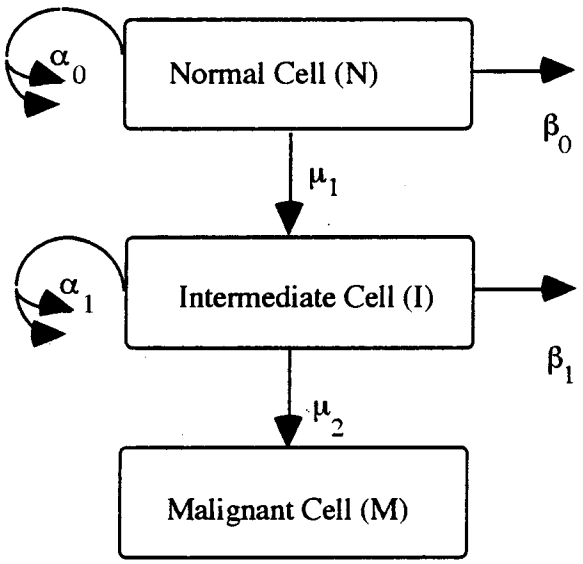


Fig. 1.11 The two stage stochastic model of carcinogenesis of Moolgavkar, Venzon, and Knudson(M & V,1979; M & K,1981)

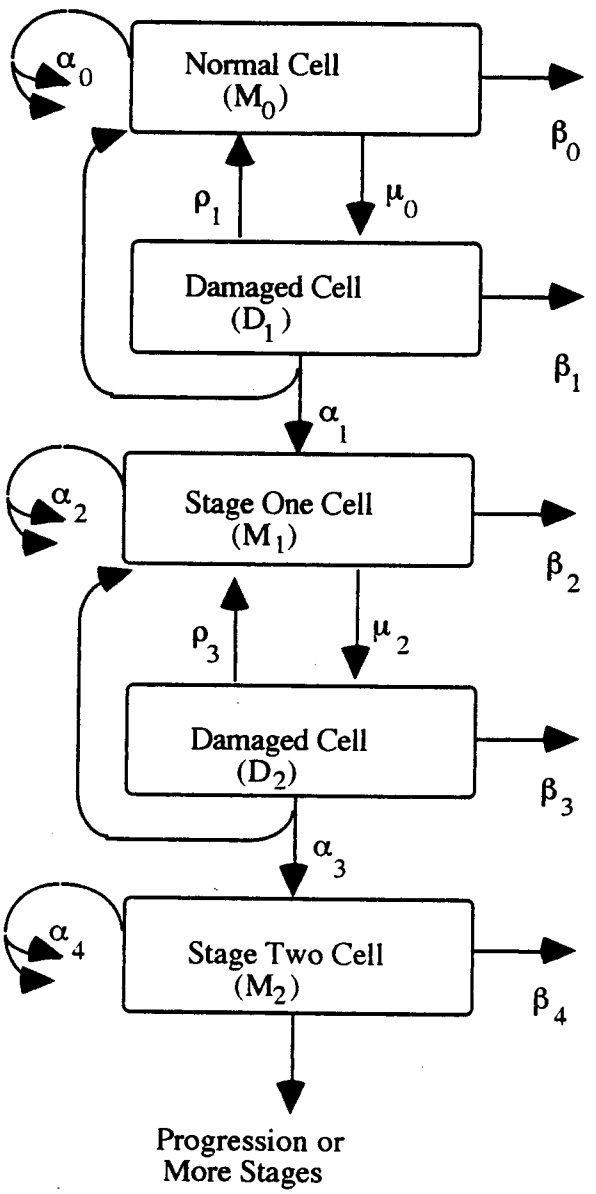


Figure 1.12 The Damage-Fixation-Model of Portier and Kopp (1990)

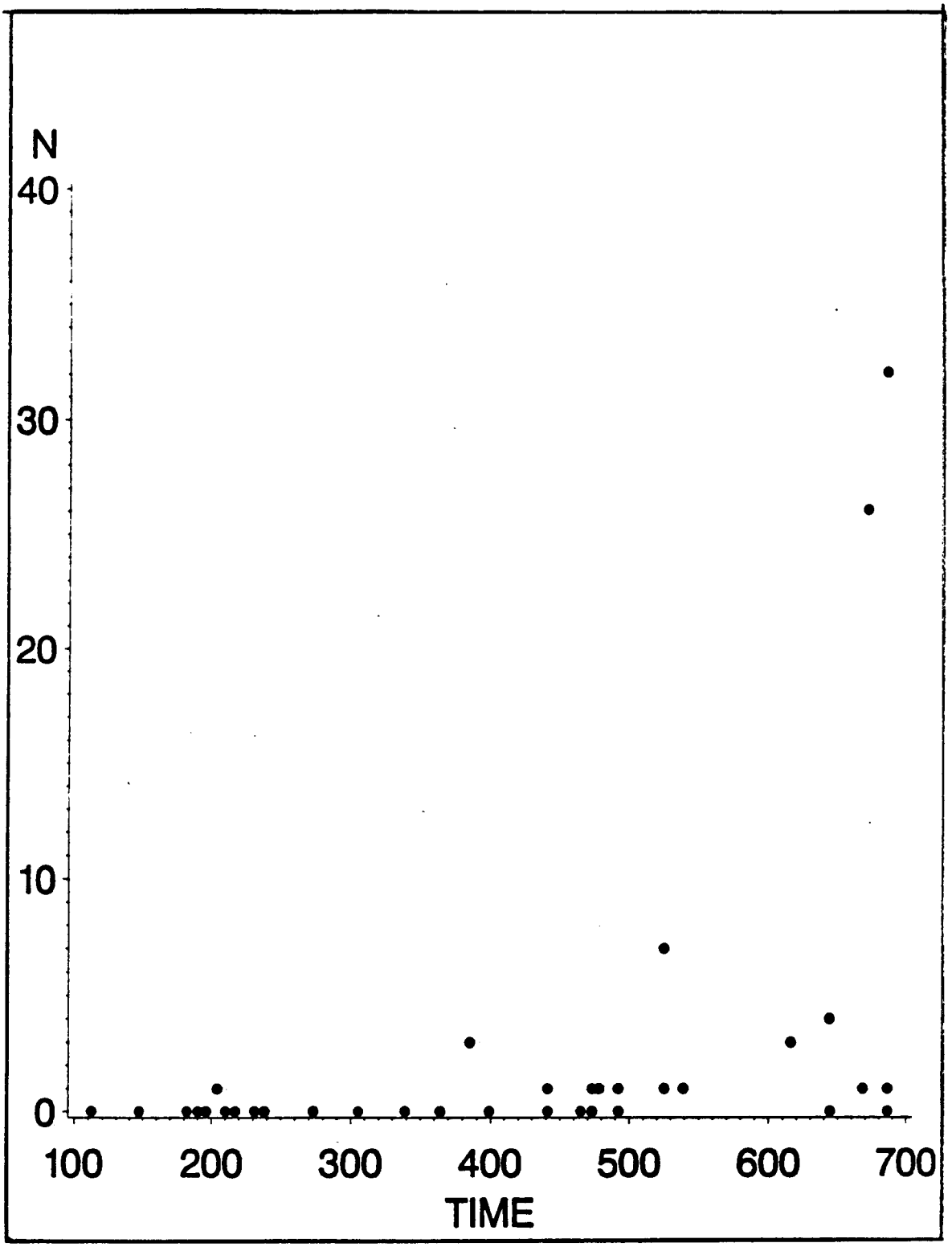


Fig. 2.1 NNM rat liver foci data (Moolgavkar et al, 1990) plot: number of foci vs. length of exposure, NNM=0 ppm.

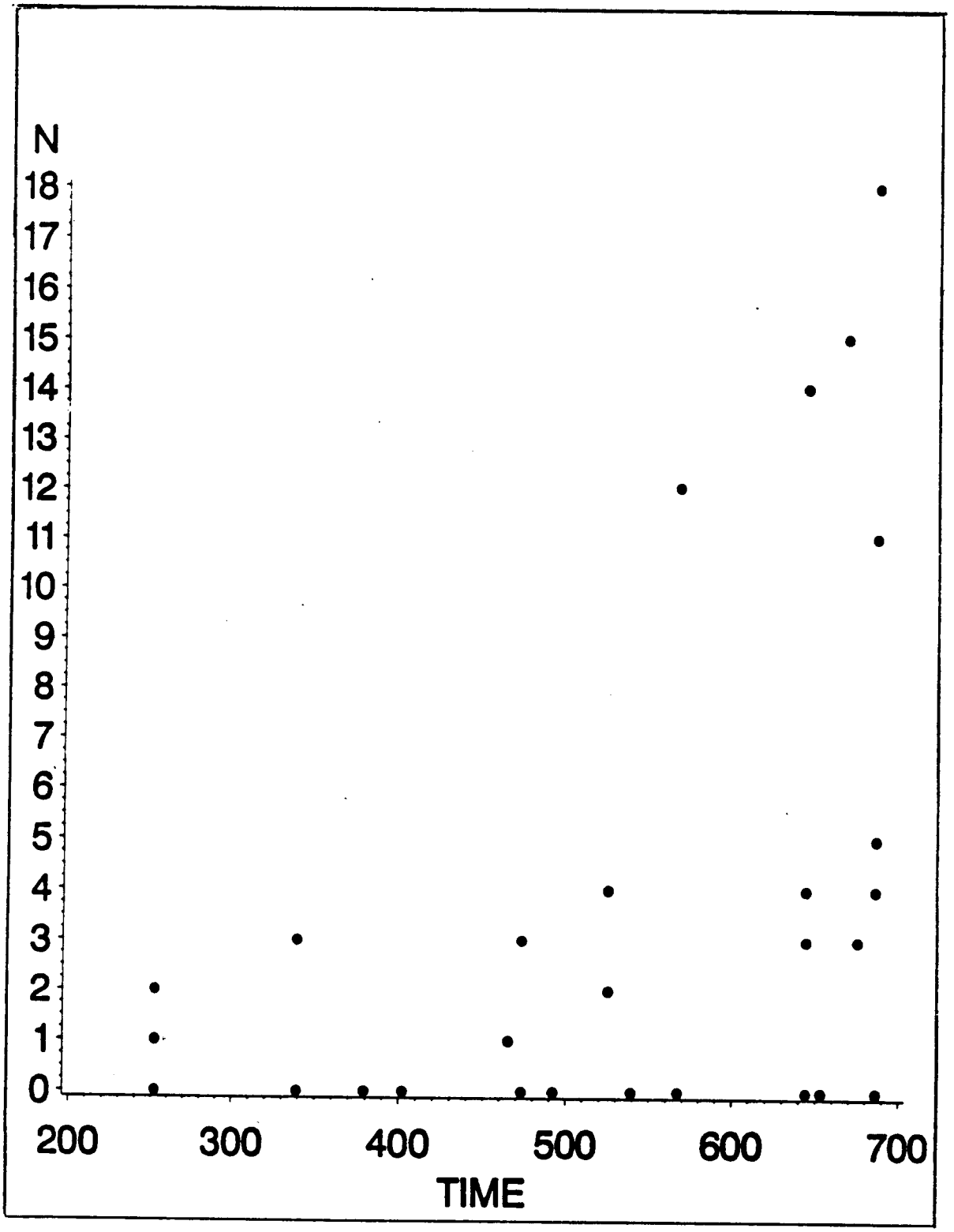


Fig. 2.2 NNM rat liver foci data (Moolgavkar et al, 1990) plot: number of foci vs. length of exposure, NNM=0.1 ppm

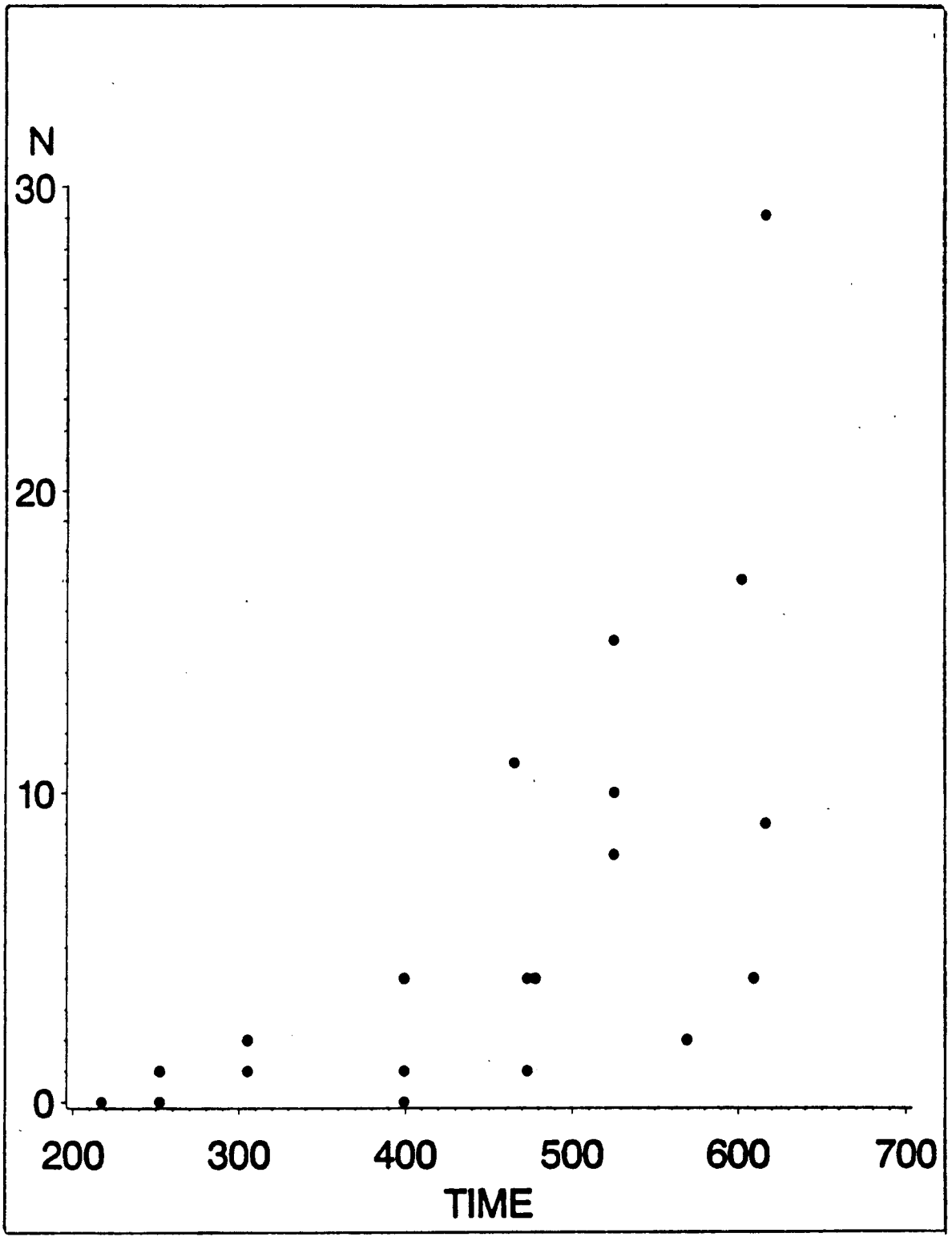


Fig. 2.3 NNM rat liver foci data (Moolgavkar et al, 1990) plot: number of foci vs. length of exposure, NNM=1 ppm.

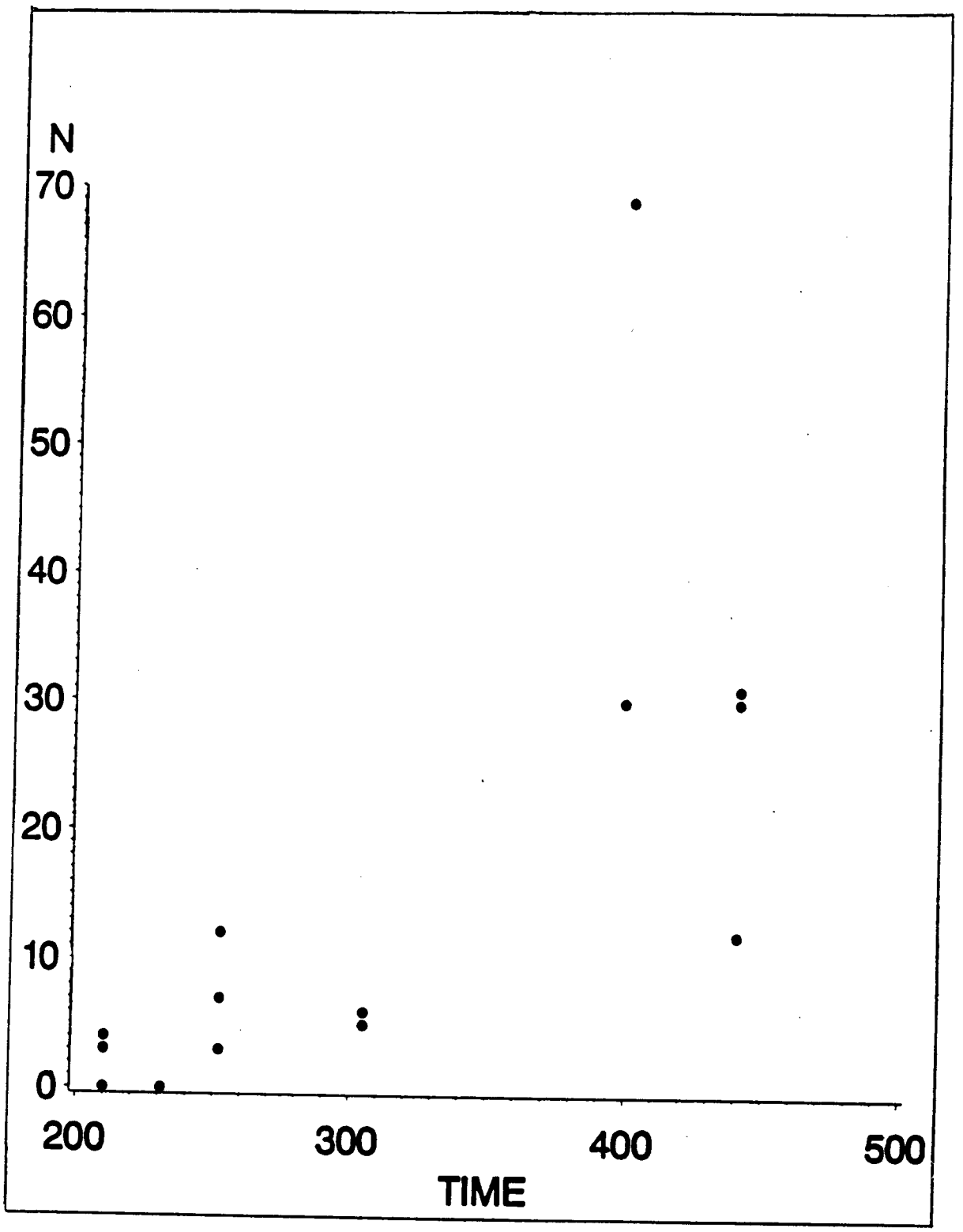


Fig. 2.4 NNM rat liver foci data (Moolgavkar et al, 1990) plot: number of foci vs. length of exposure, NNM=5 ppm

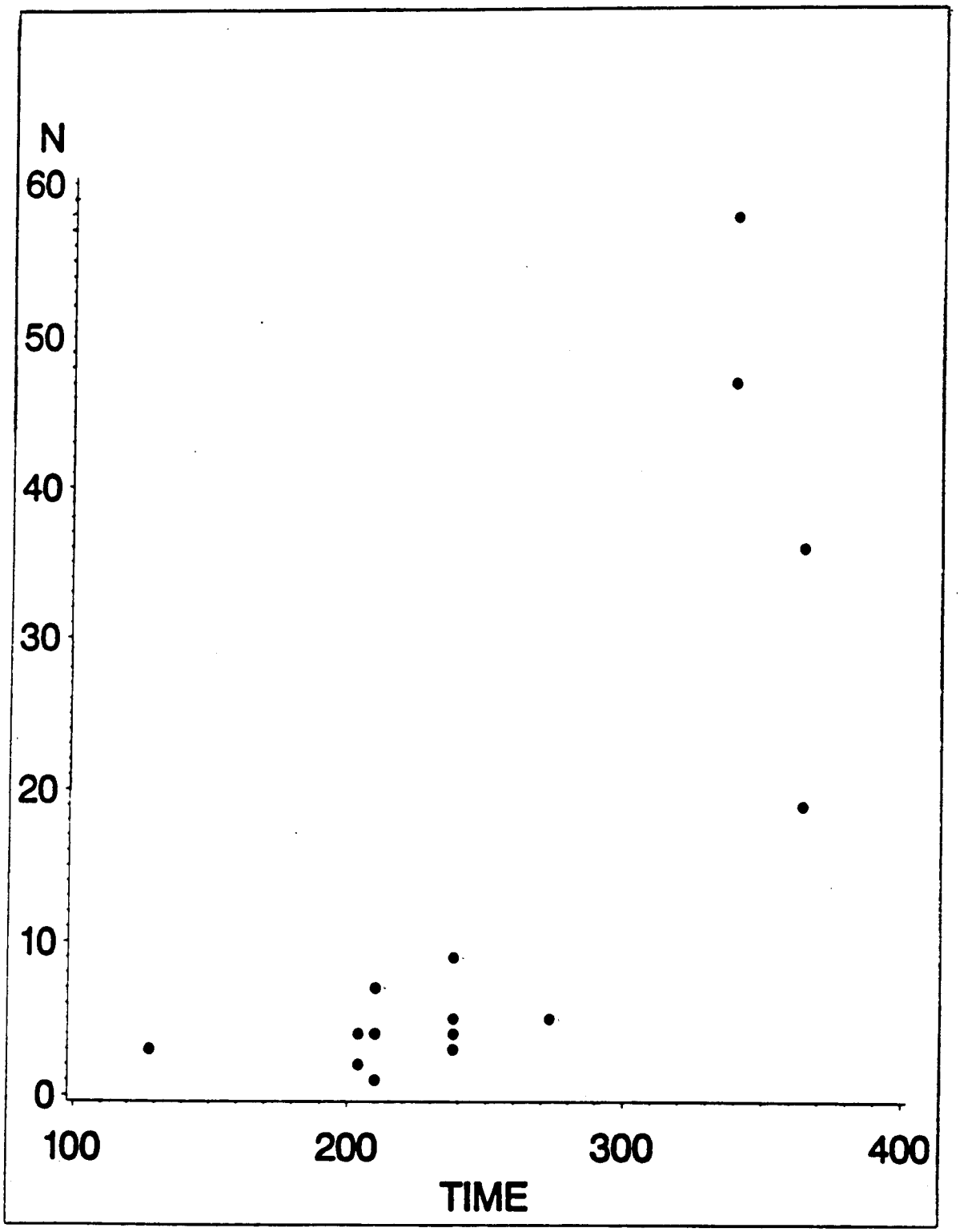


Fig. 2.5 NNM rat liver foci data (Moolgavkar et al, 1990) plot: number of foci vs. length of exposure, NNM=10 ppm.

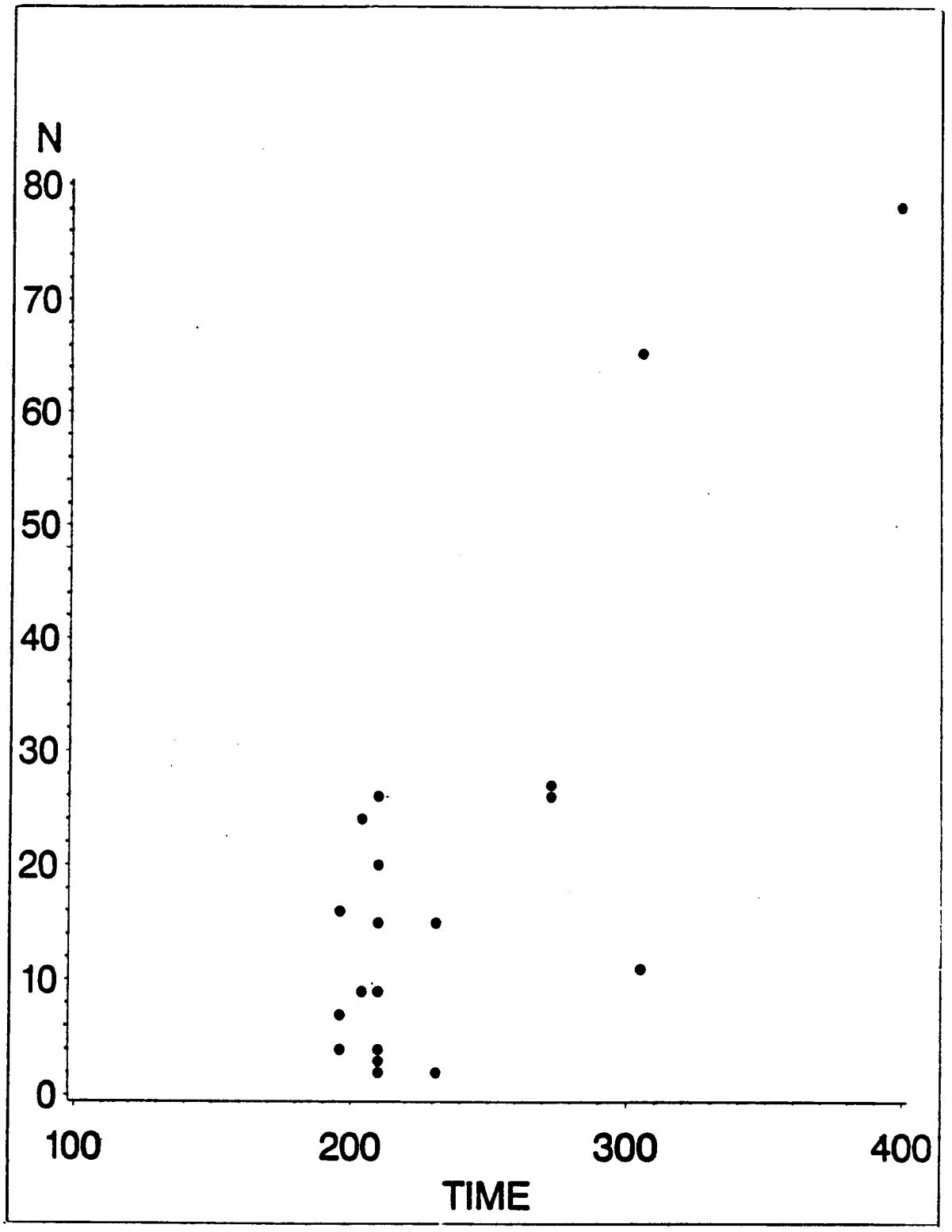


Fig. 2.6 NNM rat liver foci data (Moolgavkar et al, 1990) plot: number of foci vs. length of exposure, NNM=20 ppm

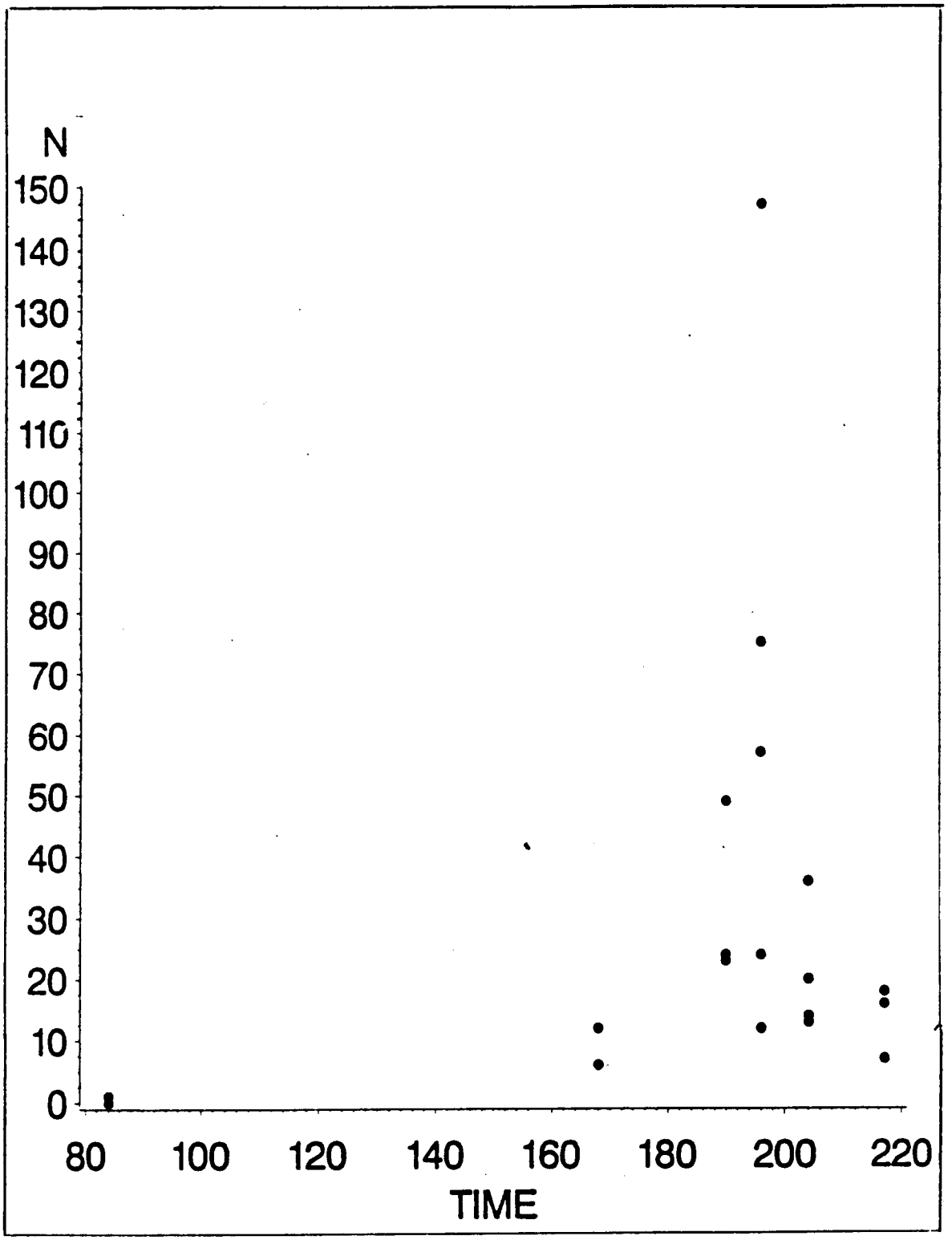


Fig. 2.7 NNM rat liver foci data (Moolgavkar et al, 1990) plot: number of foci vs. length of exposure, NNM=40 ppm.

Fig. 2.8 Variance vs. mean of the number of PML's for NNM rat liver foci data of Moolgavkar et al (1990).

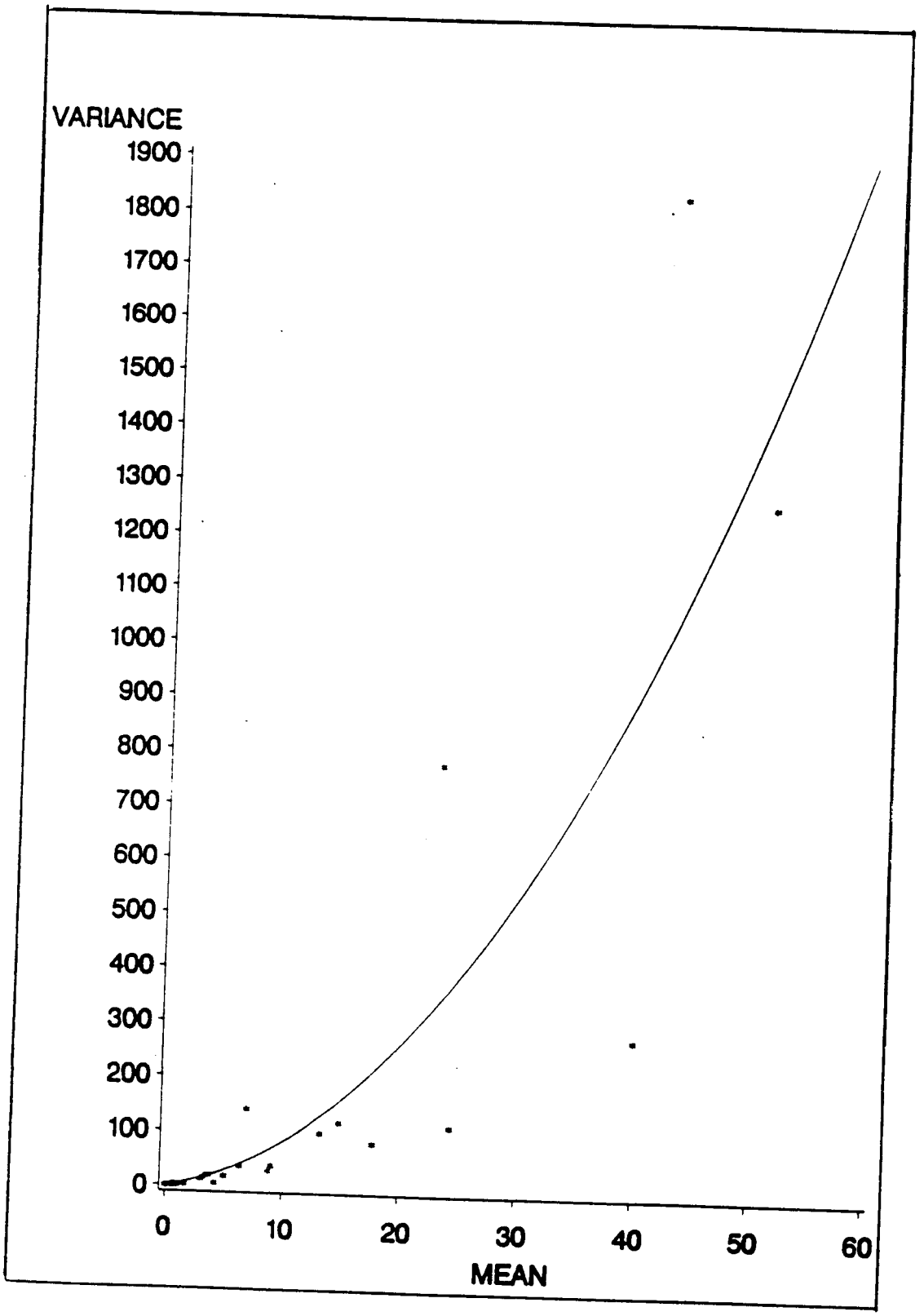


Fig. 2.9 Variance vs. mean of the number of PML's for mouse skin papilloma data of Kopp-Schneider et al (1991).

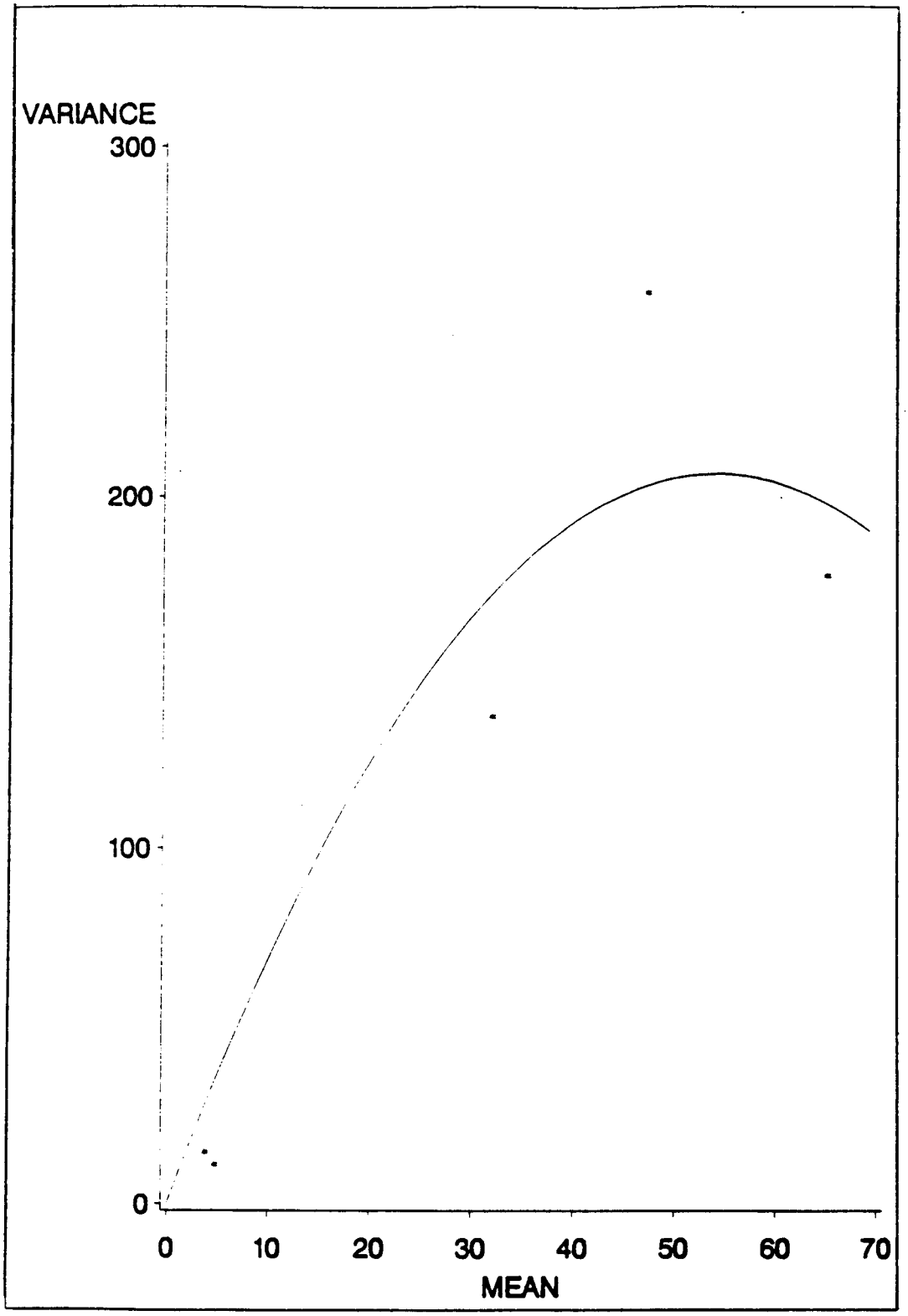


Fig. 2.10 Variance vs. mean of the number of PML's for dioxin rat liver foci data of Trischer et al (1991).

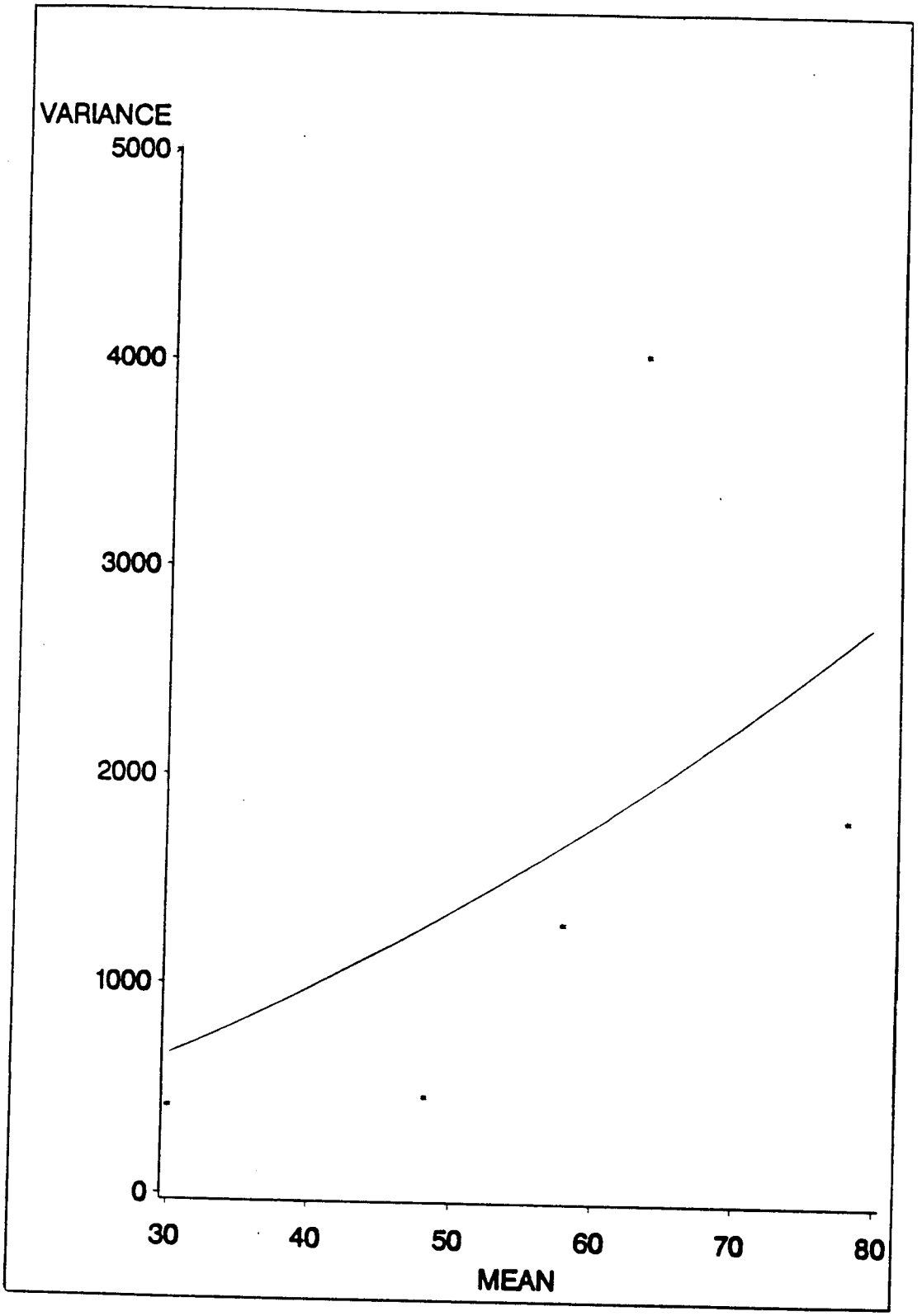


Fig. 3.1 Gamma distribution with shape parameter = 1

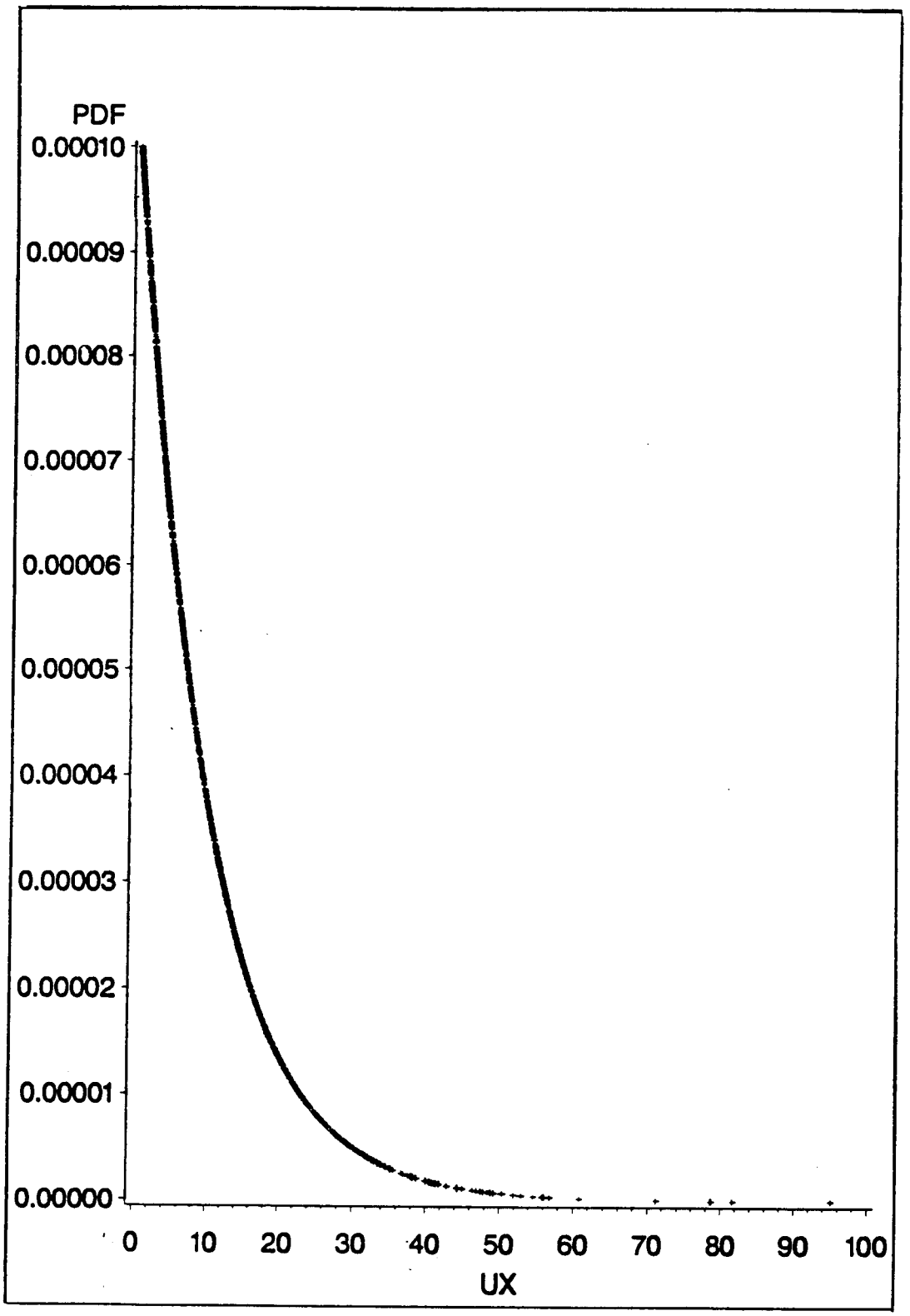


Fig. 3.2 Gamma distribution with shape parameter = 3

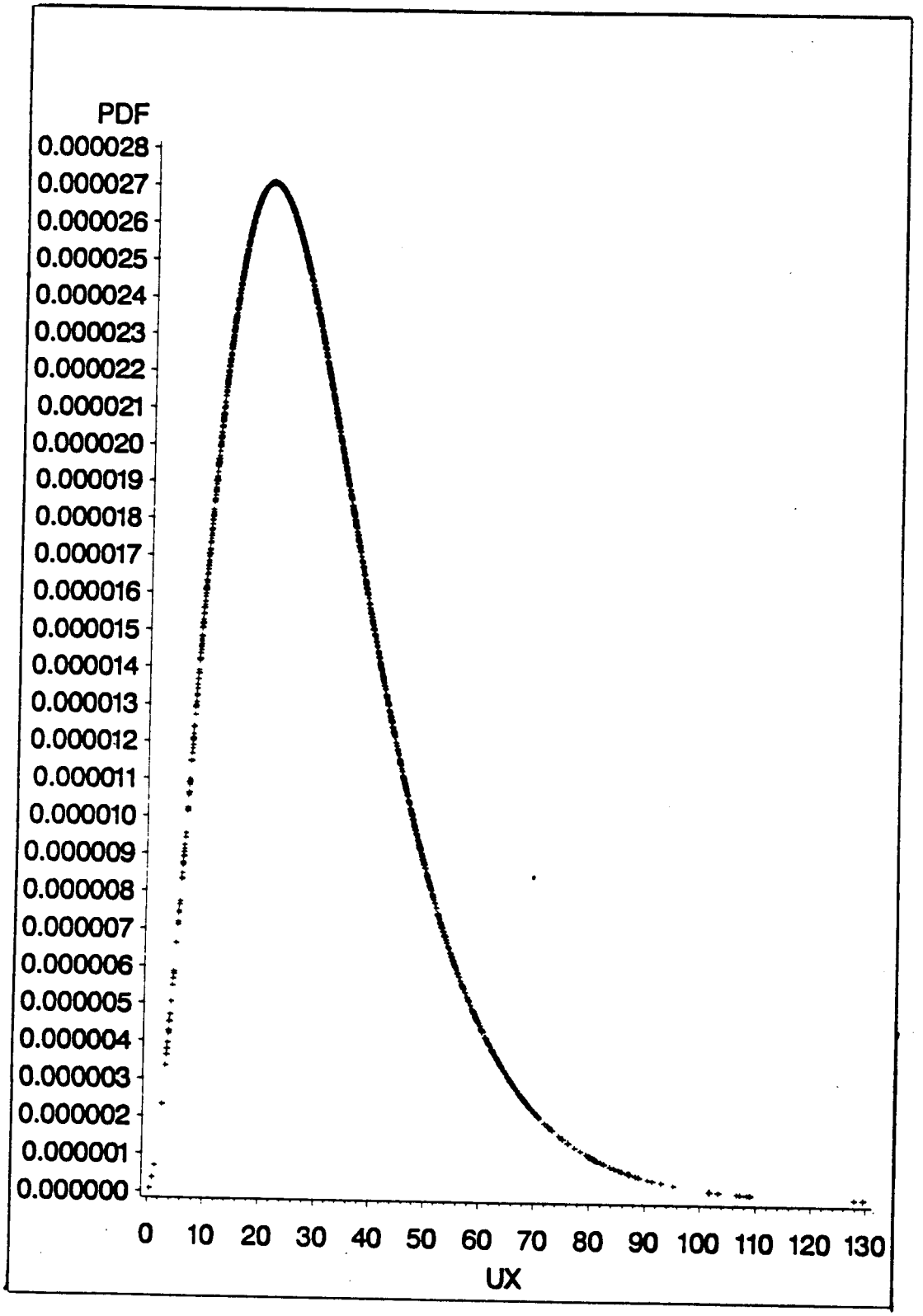


Fig. 3.3 Gamma distribution with shape parameter = 9

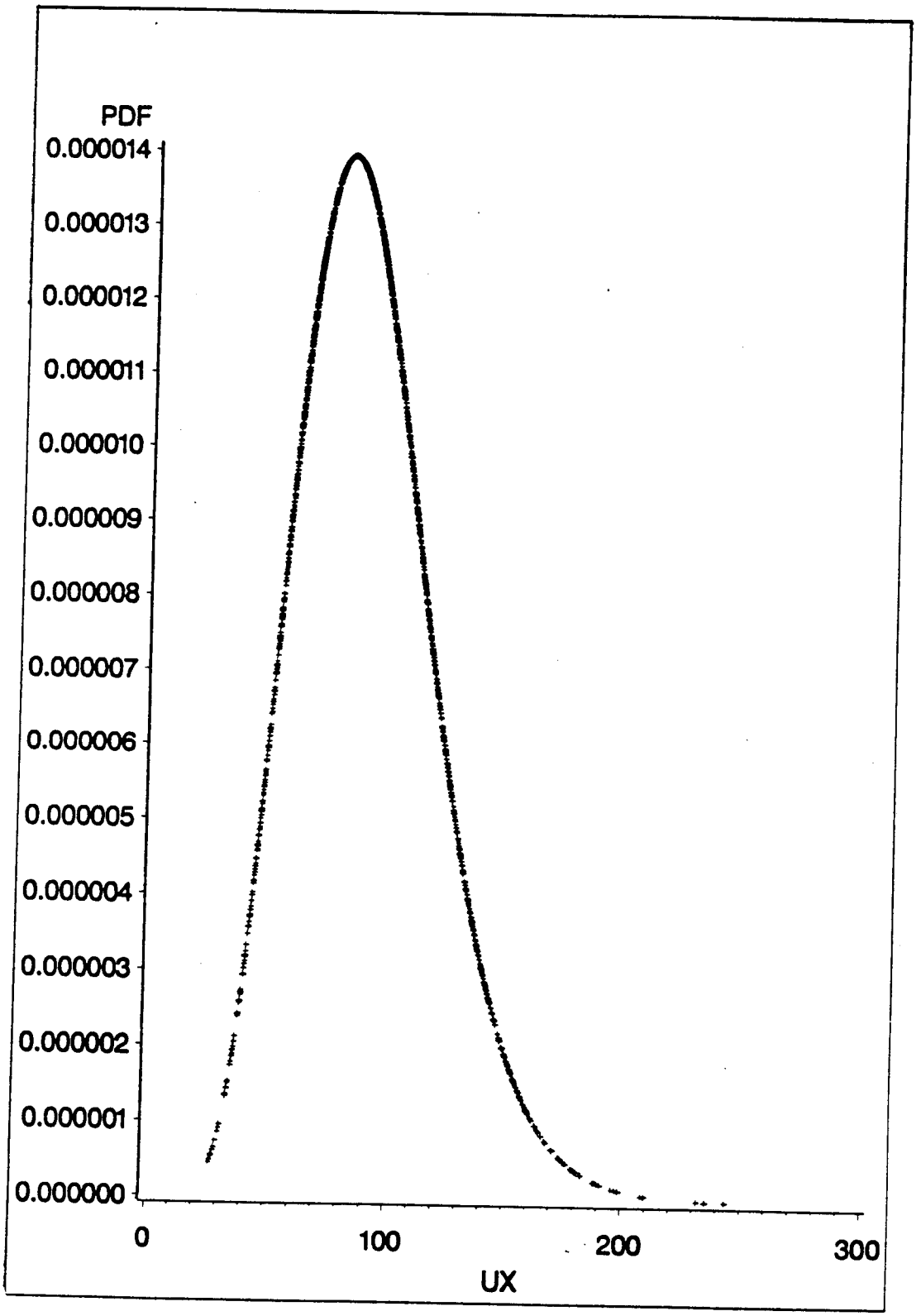
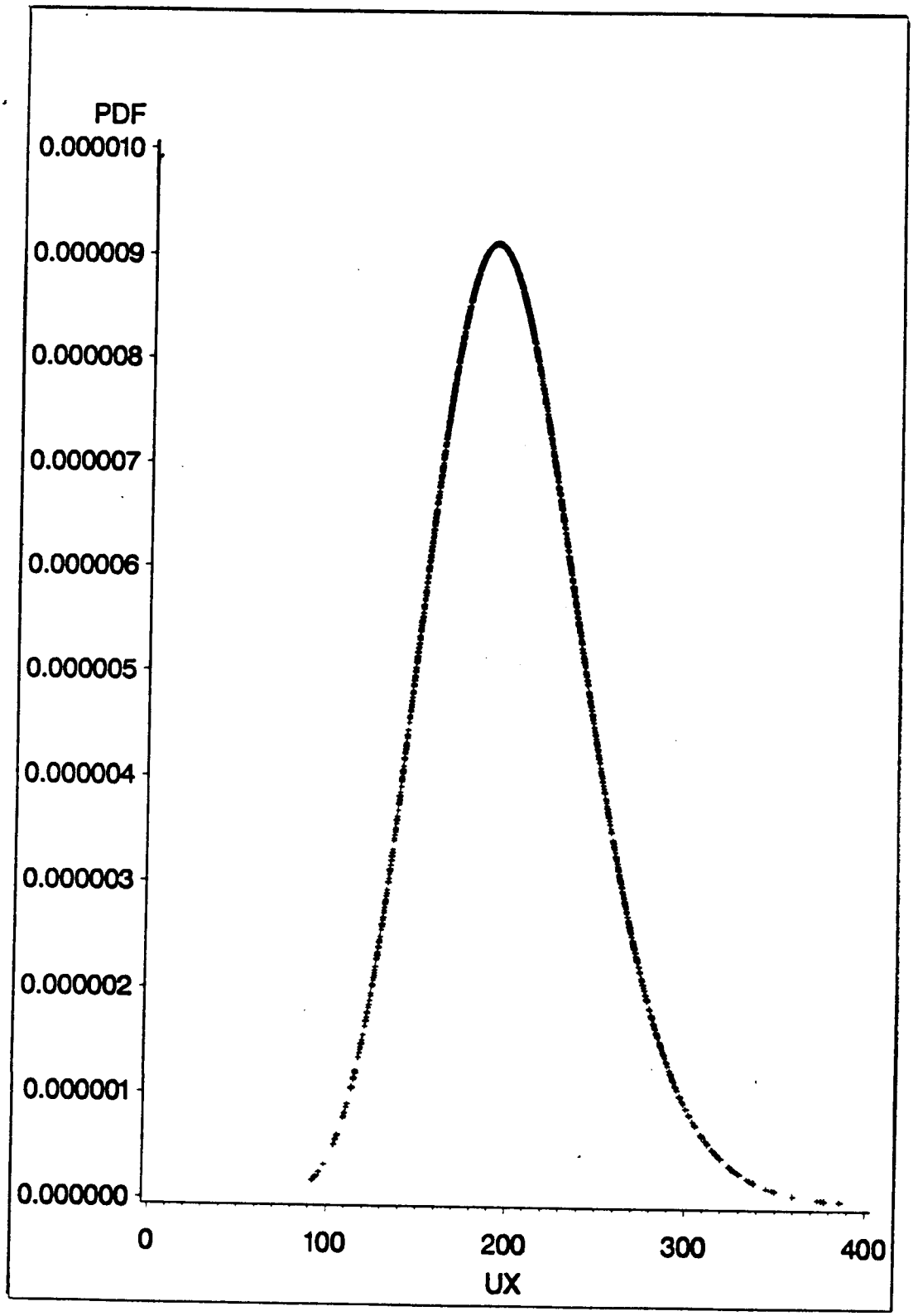


Fig. 3.4 Gamma distribution with shape parameter = 20



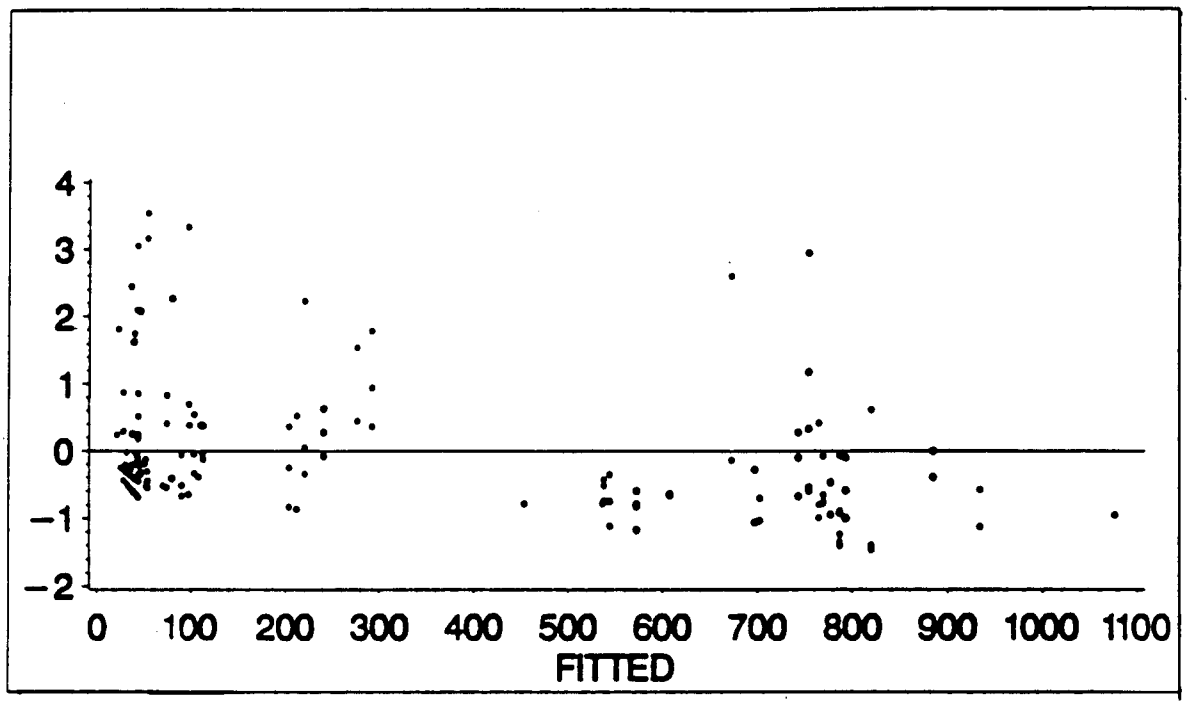


Fig. 3.5 Residual plot of NNM rat liver foci data (Moolgavkar et al, 1990) fitting the negative binomial model: standardized residuals vs. the fitted number of PML's.

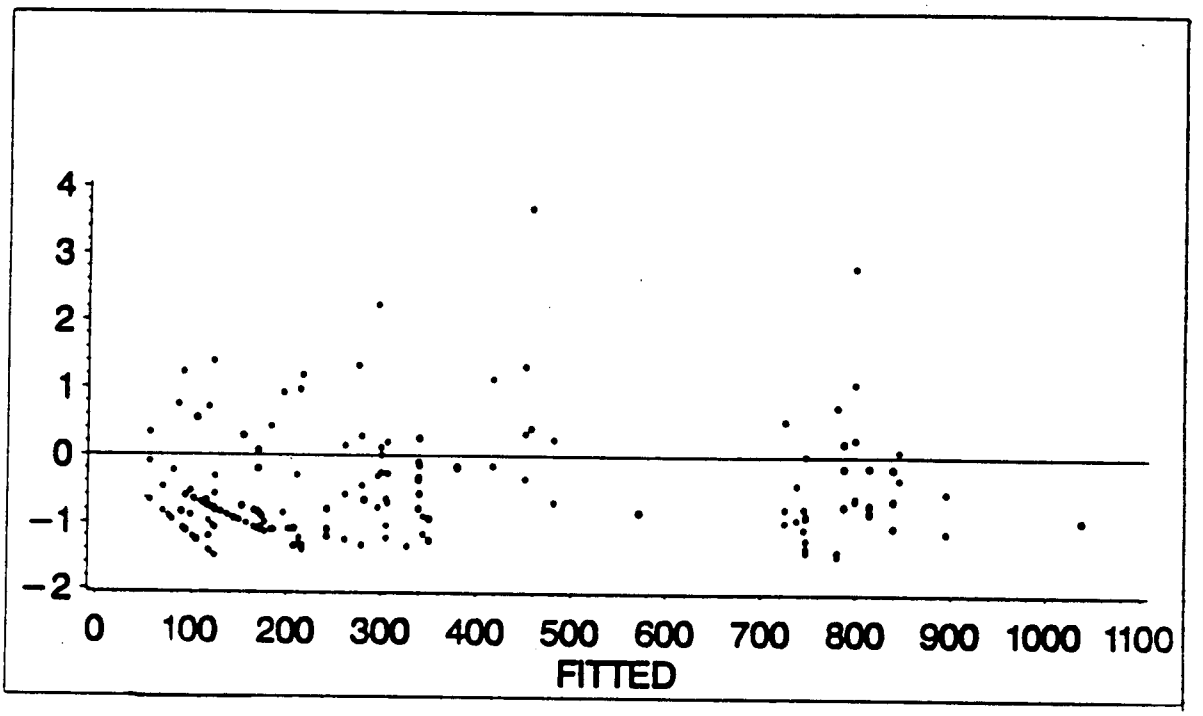


Fig. 3.6 Residual plot of NNM rat liver foci data (Moolgavkar et al, 1990) fitting the Poisson-uniform model: standardized residuals vs. the fitted number of PML's.

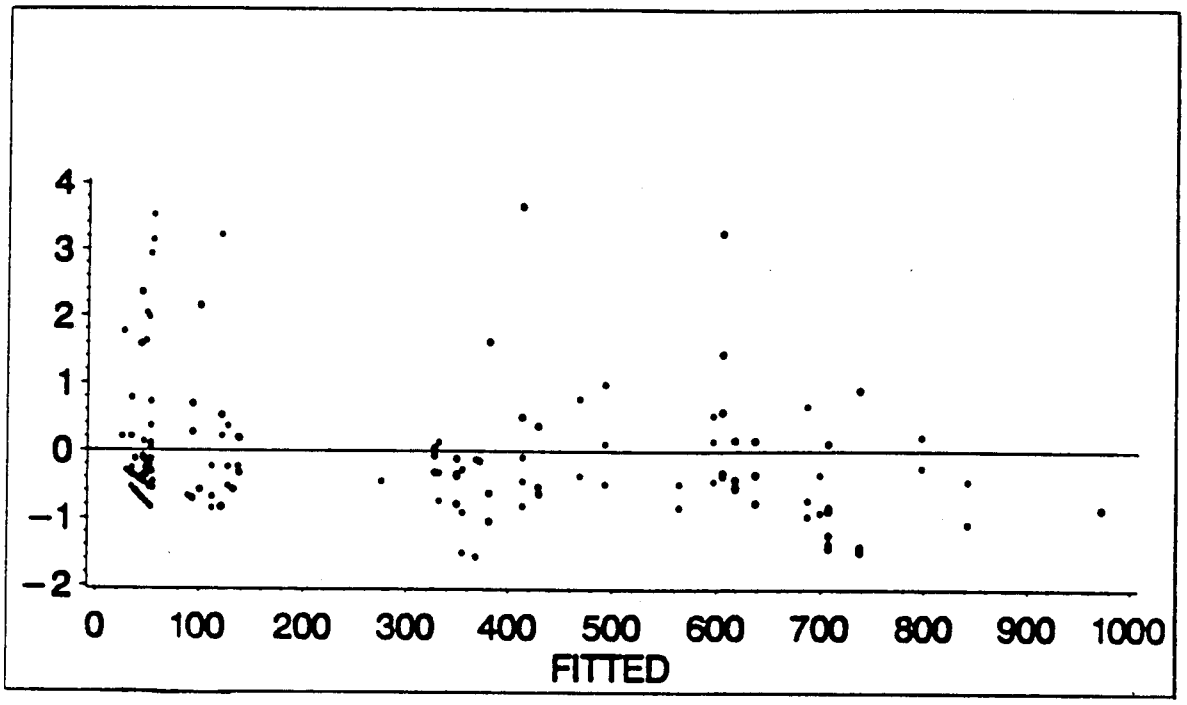


Fig. 3.7 Residual plot of NNM rat liver foci data (Moolgavkar et al, 1990) fitting the Poisson model: standardized residuals vs. the fitted number of PML's.

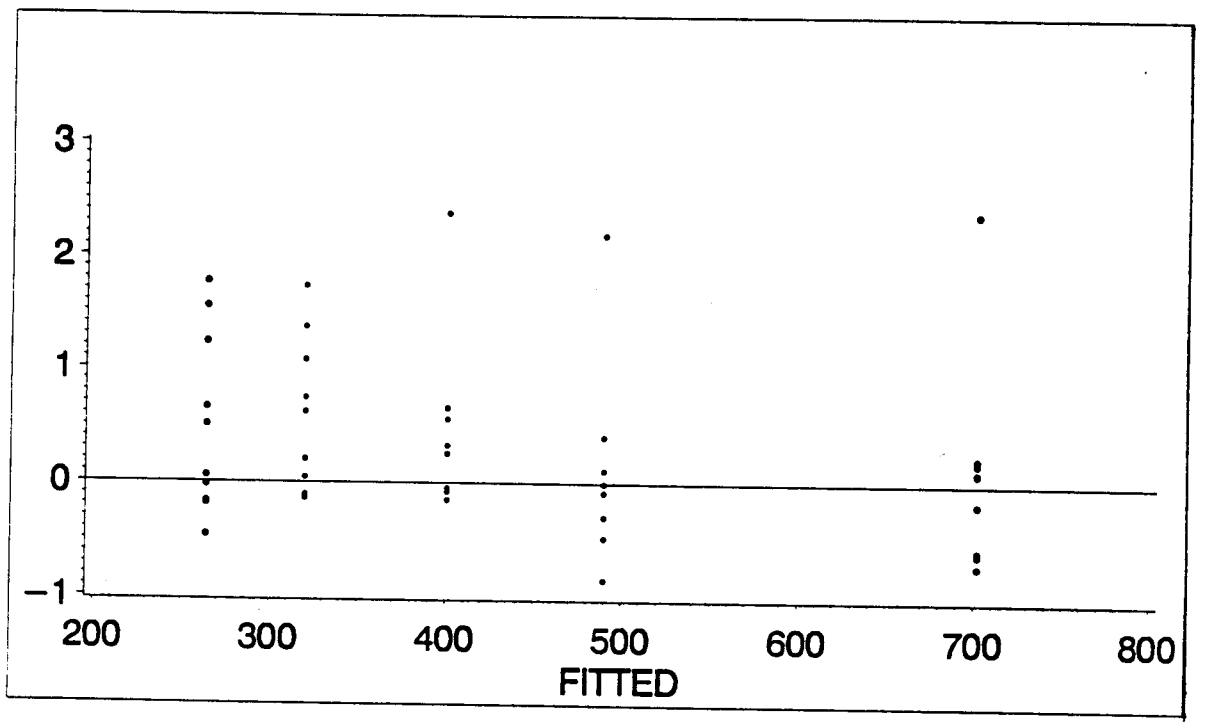


Fig. 3.8 Residual plot of dioxin rat liver foci data (Trischer et al, 1991) fitting the negative binomial model: standardized residuals vs. the fitted number of PML's.

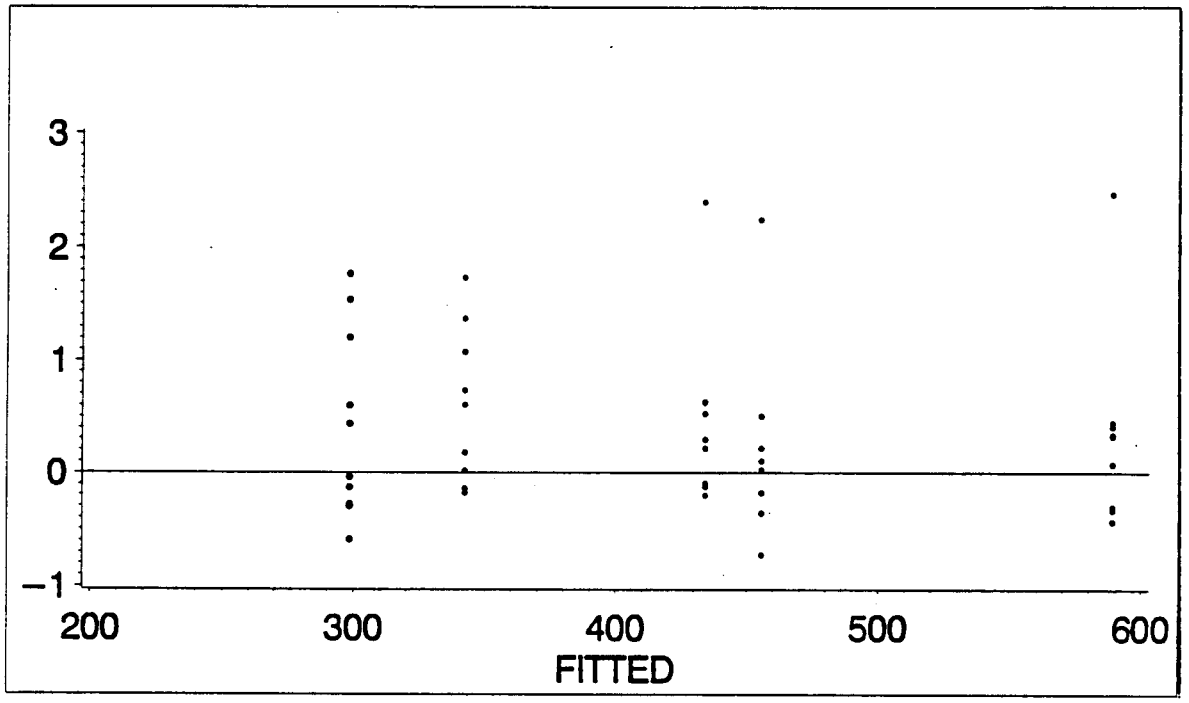


Fig. 3.9 Residual plot of dioxin rat liver foci data (Trischer et al, 1991) fitting the Poisson-uniform model: standardized residuals vs. the fitted number of PML's.

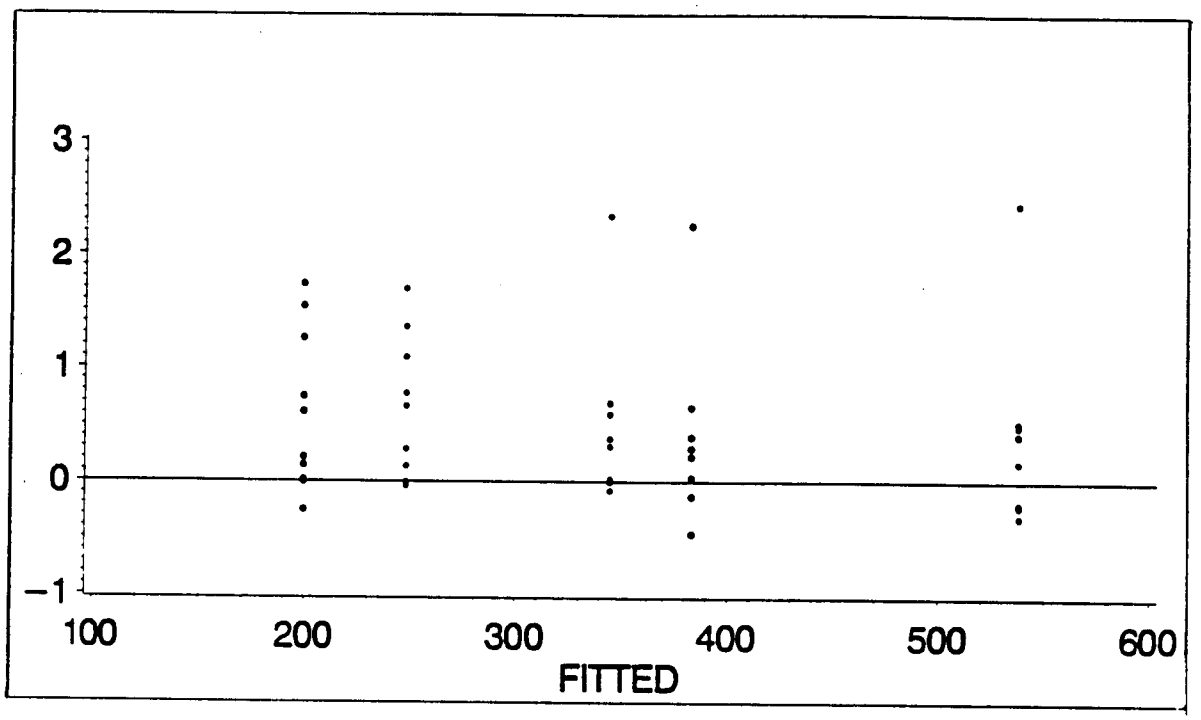


Fig. 3.10 Residual plot of dioxin rat liver foci data (Trischer et al, 1991) fitting the Poisson model: standardized residuals vs. the fitted number of PML's.

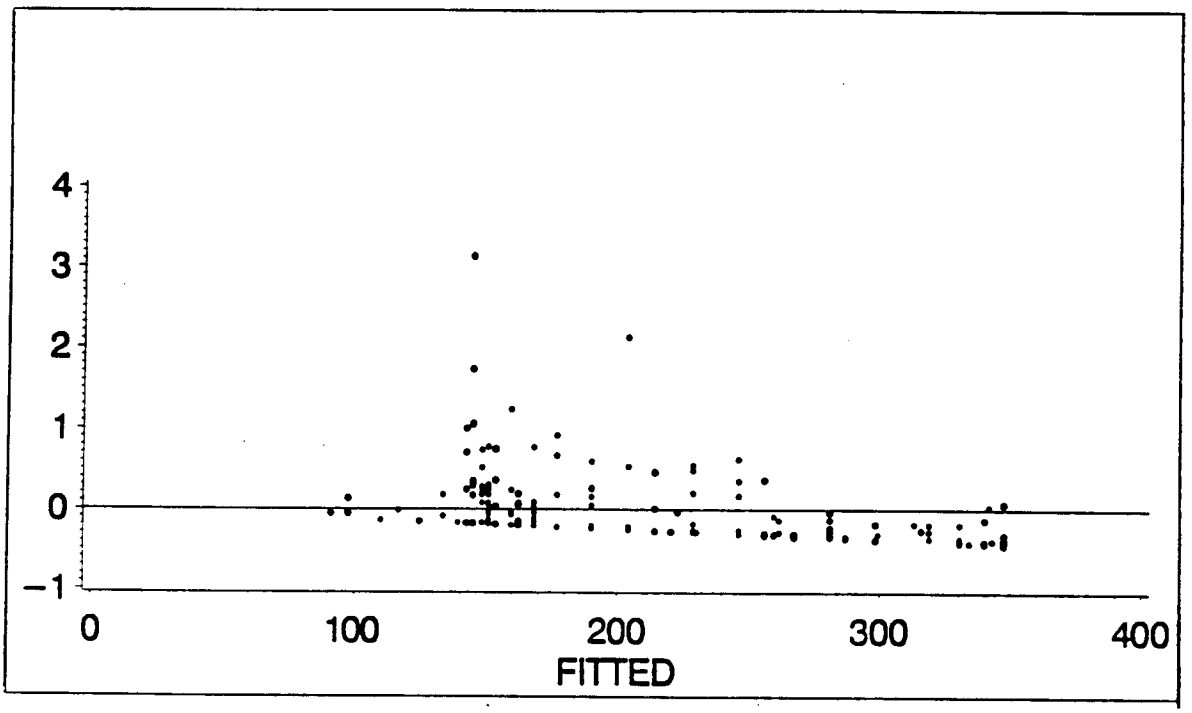


Fig. 4.1 Residual plot of NNM rat liver foci data (Moolgavkar et al, 1990) fitting the quasi-likelihood method: standardized residuals vs. the fitted number of PML's.

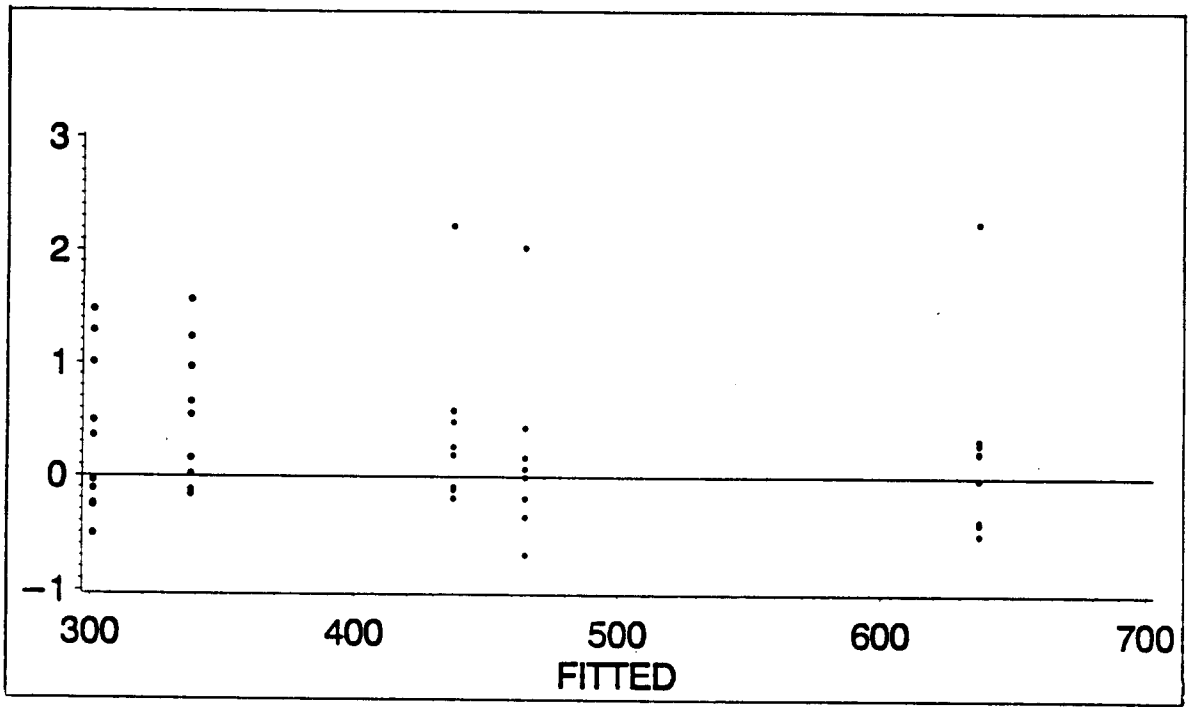


Fig. 4.2 Residual plot of dioxin rat liver foci data (Trischer et al, 1991) fitting the quasi-likelihood method: standardized residuals vs. the fitted number of PML's.

REFERENCES

Anderson, M. (1987) Issues in biochemical applications to risk assessment: How do we evaluate components of multistage models? *Envi. Health Perspective* 76:175-180.

Armitage, P. and Doll, R. (1954) The age distribution of cancer and a multi-stage theory of carcinogenesis, *British J. Cancer* 8: 1-12.

Armitage, P. and Doll, R. (1957) A two-stage theory of carcinogenesis in relation to the age distribution of human cancer.*British J. of Cancer* 11:161-179.

Cheng, R.C.H. (1977) The generation of gamma variables *Applied Statistics* 26: 71-75.

Chouroulinkov,I., Lasne,C., Lowy,R., Wahrendorf,J.,Becher,H., Day,N.E., and Yamasaki,H. (1989) Dose and frequency effect in mouse skin tumor promotion. *Cancer Research* 49:1964-1969.

Davidian, M. , and Carroll, R.J. (1988) A note on extended quasi-likelihood, Journal of the Royal Statistical Society, Ser.B, 50:74–82,.

Dewanji,A., Venzon,D. and Moolgavkar,S. (1989) A stochastic two-stage model of carcinogenesis.II. The number and size of premalignant clones. *Risk Analysis* 9:179-187.

Doll, R. (1971) The age distribution of cancer: implications for models of carcinogenesis. *J.Royal Statist. Soc.*134:133-166.

Emmelot,P. and Scherer,E. (1980) The first relevant cell stage in rat liver carcinogenesis: a quantitative approach. *Biochim. Biophys. Acta* 605:247-304.

Farber, E.(1984 a) Cellular biochemistry of the stepwise development of cancer with chemicals. *Cancer Res.*44:5463-5474.

Farber, E.(1984 b) The multistep nature of cancer development *Cancer Res.*44:4217-4223.

Feron,V.J., Grice,H.C., Griesemer,R., and Peto,R. (1980). Report 1. Basic requirements for long-term assays for carcinogenicity. In: *IARC Monographs on the Long-term and Short-term Screening Assays for Carcinogenesis: A Critical Appraisal.* Lyon: International Agency for Research on Cancer.

Fisher, J.C. and Holloman, J.H. (1951) A new hypothesis for the origin of cancer foci. *Cancer* , 4,916-918.

Fishman, G.S. (1978) Principles of Discrete Event Simulation. New York, John Wiley & Sons, Inc.

Fishman, G.S. and Moore, L.R. (1982) "A statistical evaluation of multiplicative congruential generators with modules (2^{31}-1)," *Journal of the American Statistical Association*, 77, 129-136.

Glauert, H.P., Schwarz, M., and Pitot, H.C (1986). The phenotypic stability of altered hepatic foci: effect of the short-term withdrawal of Phenobarbital and of the long-term feeding of purified diets after the withdrawal of Phenobarbital. *Carcinogenesis* , .7 no.1 pp.117-121.

Goldsworthy,T.L., Hanigan,M.H. and Pitot,H.C. (1986) Models of hepatocarcinogenesis in the rat - contrasts and comparisons. *CRC Crit. Rev. Toxicol.* 17,61-89.

Hanawalt,P.C. Preferential DNA repair in expressed genes.*Environ. Health Perspect.* 76:9-14 (1987).

Harris,T.E.(1963) *The theory of branching processes*. Springer-Verlag, Berlin.

Haseman, J.K.(1983) Statistical support of the proposed National Toxicology Program Protocol. *Toxicology Pathology* 1, pp.77-81.

Haseman, J.K., Hajian, G., Crump, K.S., Selwyn, M.R., Peace, K.E (1990). Dual control groups in rodent carcinogenicity studies In: "*Statistical issues in drug research and development*" edited by Peace, K.E., p.352.Marcel Dekker, Inc.

Iverson,S. and Arley,N.(1952) On the mechanism of experimental carcinogenesis, *Acta. Path. Microbiol. Scand.*,27 pp.773-803.

Kopp-Schneider, A., Portier, C.J.. and Rippmann (1991). The application of a multistage model which incorporates DNA damage and repair to the analysis of initiation/promotion experiments. *Mathematical Biosciences* .

Kopp-Schneider, A (1992). Birth-death processes with piecewise constant rates. *Mathematical Biosciences* .

Kunz,H.W., Tennekes,H.A., Port,R.E., Schwarz,M., Lorke,D. and Schaude,G.(1983) Quantitative aspects of chemical carcinogenesis and tumor promotion in liver. Environ. Health Perspect., 50, 113-122.

Margolin, B.H., Kaplan, N., and Zeiger, E. (1981). "Statistical analysis of the Ames *Salmonella* microsome test," *Proceedings of the National Academy of Science*, 76, 3779-3783.

Marquardt, D.W.(1963), An algorithm for least-squares estimation of nonlinear parameters, *J. Soc. Indus. Appl. Maths.*, 11, 431-441.

McCullagh, P., and Nelder, J.A. (1983), *Generalized Linear Models*, London: Chapman & Hall.

Moolgavkar S. and Venzon D.(1979).Two-event models for carcinogenesis: Incidence curves for childhood and adult tumors. *Mathematical Biosciences* 47:55-77.

Moolgavkar, S. and Knudson, A.(1981). Mutation and cancer: A model for human carcinogenesis. *J.of the National Cancer Institute* 66, 1037-1052.

Moolgavkar,S., Luebeck,E.G.,Gunst,M.D.,Port,R.E. and Schwarz, M (1990). Quantitative analysis of enzyme-altered foci in rat hepatocarcinogenesis experiments-I. Single agent regimen. *Carcinogenesis* Vol.11, No.8, pp.1271-1278.

Naylor, T.J., Balintfy, J. L., Burdick, D, and Chu, K (1966). *Computer Simulation Techniques*, Wiley, New York.

Neyman, J. & Scott, E (1967). Statistical aspect of the problem of carcinogenesis. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, UC press, Berkeley and LA, Vol.4,pp.745-776.

Nychka,D., Wahba,G., Goldfarb,S. and Pugh,T.(1984) Cross validated spline methods for the estimation of 3-dimensional tumor size distributions from observations on 2-dimensional cross-sections J. Am. Stat. Assoc., 79,832-846.

Nordling, C.O. (1953) A new theory on the cancer-inducing mechanism. *Brit. J. Cancer*, Vol.7, pp.68-72.

Pitot, H.C., Barness, L., and Kitagawa, T. (1978) Stages in the Process of Hepatocarcinogenesis in rat liver. *Carcinogenesis*, Vol. 2. *Mechanisms of tumor promotion and cocarcinogenesis*. edited by T.J. Slaga, A.Sivak, and R.K. Boutwell. Raven Press, New York.

Portier, C.J., Hedges, J.C., and Hoel, D.G.(1986). Age specific models of mortality and tumor onset for historical control animals in the National Toxicology Program's carcinogenicity experiments. *Cancer Research* 46, 4372-4378.

Portier, C.J. and Kopp-Schneider(1991), A. A multistage model of carcinogenesis incorporating DNA damage and repair. *Risk Analysis* (in press)

Powell, M.J.D. (1965) A method for minimizing a sum of squares of nonlinear functions without calculating derivatives, The Computer Journal, 7, 303-307.

Rothman, K.J. and Keller, A.(1972). The effect of joint exposure to alcohol and tobacco on risk of cancer of the mouth and pharynx. *J. of Chronic Disease* 25, 711-716.

Rubinstein, R.Y.(1981) Simulation and the Monte Carlo method. Wiley and sons, Inc.

SAS language: Reference, version 6, first edition, SAS Institute Inc., 1990.

Schere, E., and Emmelot, P. (1975): Kinetics of induction and growth of precancerous liver-cell foci, and liver tumor formation by diethylnitrosamine in the rat. *Eur. J. Cancer* , 11:689-701.

Selikoff I.J., Hammond E.C. and Churg,J.(1968) Asbestos exposure, smoking and neoplasis. *J.of the American Medical Association* 204, 106-112.

Shimkin, M.B. and Polissar, M.J. (1955) Some quantitative observations on the induction and growth of primary pulmonary tumors in strain A mice receiving urethane. *J. of Nat'l Cancer Inst.*, Vol.16, pp.75-97.

Shimkin, M.B., Wieder, R., Marzi, D., Gubareff, N. and Suntzeff, V. (1967) Lung tumors in mice receiving different schedules of urethane. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, UC Press, Berkeley and L.A., Vol.4, pp.707-719.

Slaga, T.J.,Fischer, S.M.,Nelson, K. and Gleason, G.L.(1980) Studies on the mechanism of skin tumor promotion: Evidence for several stages in promotion. Proceedings of the Nat'l Academy of Sciences 77, 3659-3663.

Solt, D. & Farber, E (1976). New principle for the analysis of chemical carcinogenesis. Nature Vol.263, pp.701-703.

Sontag, J.A., Page, N.P., and Saffiotti, U.(1976). Guidelines for Carcinogen Bioassay in Small Rodents.,DHEW Publication No. (NIH) 76-801, Department of Health, Education and Welfare, Washington, D.C.

Stenback, F., Garcia, H., Shubik, P.(1974) Present status of the concept of promoting action and co-carcinogenesis in skin. In:Shubik P., ed. *Physiopathology of cancer.* Vol 1, *Biology and biochemistry.* Basel: Karger: 155-225.

Stocks, P.(1953) A study of the age curve for cancer of the stomach in connection with a theory of the cancer producing mechanism. *Brit. J. Cancer*, Vol.7, pp.407-417.

Tritscher, A.M., Goldstein, J.A., Portier, C.J., McCoy, Z., Clark, G.C., and Lucier, G.W. (1992). Dose-response relationships for chronic exposure to 2,3,7,8 tetrachlorobenzo-p-dioxin in a rat tumor promotion model: Quantification and immunolocalization of CYP1A1 and CYP1A2 in the liver. *Carcinogenesis*, 52, 3436-3442.

Walsh, G.R.(1975) Methods of Optimization, John Wiley & Sons, Ltd.

Wedderburn, R. W. M. (1974). Quasi-likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method. *Biometrika*, 61, 439-447.

Weisburger,J.H.(1975) *Toxicology*, Casarett &Doull.

White, M., Grendon, A. and Jones, H.B.(1967) Effects of urethane dose and time patterns on tumor formation. Proceedings *of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, UC Press, Berkeley and L.A., Vol.4, pp.721-743.

Whittemore, A. and Keller, J.B.(1978) Quantitative theories of carcinogenesis. *SIAM Review*, Vol.20, No.1.

Wicksell,D.S.(1925) The corpuscle problem, part I. *Biometrika*,17,87-97.

Williams, P.L. and Portier, C.J.(1989). Analytic expressions for maximum likelihood estimators in a nonparametric model of tumor incidence and death. Institute of Statistics Mimeo Series # 1868T.