

**STATISTICAL ANALYSIS OF  
WARRANTY CLAIMS DATA**

**J.D. Kalbfleisch and J.F. Lawless**

**IIQP Research Report  
RR-93-03**

**February 1993**

# STATISTICAL ANALYSIS OF WARRANTY CLAIMS DATA

*J.D. Kalbfleisch and J.F. Lawless*

Department of Statistics and Actuarial Science  
University of Waterloo  
February 1993

## *ABSTRACT*

Manufacturers generally collect fairly detailed data on warranty claims. Besides providing a record of claims and their costs, such data may be used to predict future claims, to compare claims experience for different product lines, models or places of manufacture, and to provide information on field reliability. To do this it is necessary to collect good data and employ sound methods of analysis. This report presents methods for analyzing claims which are simple and widely applicable. Car warranty data are examined as an illustration.

*Key words and phrases: Age rates of claims; Reporting delays; Prediction; Costs; Field reliability; Log linear models.*

## CONTENTS

1. INTRODUCTION
2. SIMPLE AGE-SPECIFIC CLAIMS ANALYSIS
  - 2.1 Notation and assumptions
  - 2.2 Age-specific analysis
  - 2.3 Grouped data
3. ADJUSTMENTS FOR REPORTING DELAYS
4. EXAMPLES
5. ADJUSTMENTS WHEN SALES ARE ESTIMATED
6. PREDICTION
7. COVARIATES AND REGRESSION ANALYSIS
8. CALENDAR TIME EFFECTS
9. ANALYSIS OF COSTS
10. ESTIMATION OF FIELD RELIABILITY
11. CONCLUSION

## 1. INTRODUCTION

Manufacturers generally collect fairly detailed data on warranty claims. Besides providing a record of claims and their costs, such data may be used to predict future claims, to compare claims experience for different components or product lines, and to study variations in claims relative to variables such as time and place of manufacture or usage environment. These data also provide information on the field reliability of products and so may influence design and manufacturing decisions. To take full advantage of such possibilities, it is important to obtain good data and to employ sound analysis. This chapter presents some statistical methods of quite broad applicability.

We suppose that units of a product are sold over time and that information becomes available on the time and type of all warranty claims for each unit. We assume that when a claim is recorded, the date of sale of the unit is noted; this is typically necessary in order to verify that the unit is under warranty. In some sections, we also assume that the manufacturer knows the dates of sale for all the units that have been sold. This is the typical situation for expensive items such as automobiles, but not for many other less expensive products.

Our objectives are to analyze warranty claims and, to a lesser extent, costs. Important aspects include comparisons, the assessment of trends and explanatory variables, and prediction. The emphasis is on aggregate claims for the population of units sold rather than on the detailed examination of claims patterns for individual units. We do, however, discuss individual claims and field reliability estimation briefly in section 10.

Section 2 of this chapter presents methods of age-specific claims analysis when complete data both on claims and on units sold are available. Section 3 discusses adjustments for delays in the reporting of claims, and section 4 presents some examples. Section 5 deals with adjustments when data on sales are incomplete. Section 6 deals with prediction, section 7 with covariate and regression analysis, and section 8 with calendar time effects on claims. The main focus of this chapter is on numbers of claims, but section 9 briefly discusses how

similar methods apply to the analysis of costs. Section 10 discusses the utilization of warranty claims data for reliability estimation, and section 11 concludes with a few additional remarks.

## 2. SIMPLE AGE-SPECIFIC CLAIMS ANALYSIS

### 2.1 Notation and Assumptions

We assume to start that time is measured in discrete units, and that data are available on the numbers of units sold and numbers of claims in each time period. For convenience of exposition we will take the time to be days. We assume that units are sold on calendar days  $0, 1, \dots, \tau$ , with

$$N(d) = \text{number of units sold on day } d, \quad d = 0, 1, \dots, \tau.$$

We consider claims data that have accrued up to day  $T \geq \tau$ , and assume that we can determine

$$n^*(t, a) = \text{number of claims on day } t \text{ for units} \\ \text{sold on day } t - a, \quad 0 \leq a \leq t \leq T.$$

Here the variable  $a$  is the age of the unit at the time  $t$  of a claim.

Sometimes it is useful to classify claims according to the day of sale. For this purpose we define  $n(d, a)$  to be the number of claims at age  $a$ , for units sold on day  $d$ . Clearly

$$n(d, a) = n^*(d + a, a), \quad a \geq 0, \quad d \geq 0, \quad a + d \leq T.$$

In some cases claims may be grouped or aggregated according to one or more of age, time period or date of sale. To deal with this later (see section 2.3) we use  $n^*$  and  $n$  with capital letter arguments:

$$n^*(P, A) = \sum_{t \in P} \sum_{a \in A} n^*(t, a) \\ n(D, A) = \sum_{d \in D} \sum_{a \in A} n(d, a),$$

where  $D, P, A$  represent sets of sales days, time period days and age days, respectively.

In what follows, we suppose that  $N(d)$ , the number of units sold on day  $d$  is known,  $d = 0, 1, \dots, \tau$ . To develop methods of analysis we assume that claims occur according to

some random process such that, given  $N(0), \dots, N(\tau)$ ,  $n^*(t, a)$  and  $n(d, a)$  have (conditional) expected values

$$\mu^*(t, a) = E\{n^*(t, a)\}, \quad \mu(d, a) = E\{n(d, a)\}.$$

We similarly write  $\mu^*(P, A)$  and  $\mu(D, A)$  for grouped data.

The notation above refers to one specific type of claim for a unit. More generally, we may want to consider claims of several types simultaneously. In such cases, we subscript the quantities above to designate the type. For example,  $n_j^*(t, a)$  refers to the number of type  $j$  claims on day  $t$ , for units of age  $a$ . Generally, different claim types can be analyzed separately, so we will proceed as though only one type is being considered.

## 2.2 Age-Specific Analysis

We first consider situations where the average or expected number of claims per unit per day may depend on the age of the unit but is roughly independent of other factors. This leads us to define age-specific expected claim frequencies:

$$\begin{aligned} \lambda(a) &= \text{expected number of claims per unit while at age } a \\ \Lambda(a) &= \sum_{u=0}^a \lambda(u) = \text{expected number of claims per unit up to age } a. \end{aligned}$$

Note that  $\lambda(a)$  is typically not the expected number of failures per unit at age  $a$ , since some units sold may no longer be in use or under warranty at age  $a$ . In addition, we note that with warranties where a failed unit may be replaced, the “age” with respect to the definition above is the number of days since the unit that originated the warranty coverage was sold.

It follows that the expected values of  $n^*(t, a)$  and  $n(d, a)$  are

$$\mu^*(t, a) = N(t - a)\lambda(a), \quad \mu(d, a) = N(d)\lambda(a). \quad (2.1)$$

When the  $N(d)$ 's are known, we can estimate the  $\lambda(a)$ 's from the warranty claim counts  $n^*(t, a)$  up to day  $T$  (i.e. with  $0 \leq a \leq t \leq T$ ) as

$$\hat{\lambda}(a) = \frac{n_T(a)}{R_T(a)} \quad a = 0, 1, \dots, T \quad (2.2)$$

where

$$n_T(a) = \sum_{t=a}^T n^*(t, a) = \sum_{d=0}^{T-a} n(d, a)$$

is the total number of age  $a$  claims occurring up to day  $T$ , and

$$R_T(a) = \sum_{d=0}^{T-a} N(d) \tag{2.3}$$

is the total number of units that have reached age  $a$  on or before day  $T$ . The estimates (2.2) are maximum likelihood estimates when the  $n(d, a)$ 's are independent Poisson random variables but are valid quite generally as long as (2.1) is true. These estimates and the corresponding estimates of cumulative claims

$$\hat{\Lambda}(a) = \sum_{u=0}^a \hat{\lambda}(u) \tag{2.4}$$

are very useful; examples are given in section 4. In the following we will for convenience suppress the “ $T$ ” in  $n_T(a)$  and  $R_T(a)$ , writing just  $n(a)$  and  $R(a)$ .

Variance estimates for  $\hat{\lambda}(a)$  and  $\hat{\Lambda}(a)$  can be obtained under various assumptions and derivations are outlined in the appendix. If the  $n(d, a)$ 's are independent Poisson variables, then  $\text{Var}\{n(a)\} = R(a)\lambda(a)$  and the variance of  $\hat{\Lambda}(a)$  is estimated by

$$\hat{V}_p(a) = \widehat{\text{Var}}\{\hat{\Lambda}(a)\} = \sum_{u=0}^a \frac{\hat{\lambda}(u)}{R(u)}. \tag{2.5}$$

This estimate is reasonable if units generate claims randomly and in an identical fashion. However, even though (2.1) may be acceptable, there is often extra-Poisson variation evident in the claim frequencies, and sometimes a degree of correlation. Since (2.5) may in those cases substantially underestimate  $\text{Var}\{\hat{\Lambda}(a)\}$ , we will note two approaches that allow for extra-Poisson variation. A third approach is mentioned in the appendix.

A simple approach is to assume that the  $n(d, a)$ 's (or equivalently the  $n^*(t, a)$ 's) are independent random variables with means  $\mu(d, a) = N(d)\lambda(a)$  and variances

$$v(d, a) = \text{Var}\{n(d, a)\} = \sigma^2 \mu(d, a) \tag{2.6}$$

where  $\sigma > 0$  is a dispersion parameter. The case  $\sigma = 1$  gives the Poisson variance function and  $\sigma > 1$  gives extra-Poisson variation. The model (2.6) often provides a reasonable approximation to reality. The parameter  $\sigma^2$  may be estimated in several ways, by equating expressions for observed and expected second moments. For example, the fact that  $E\left\{\frac{[n(d,a)-\mu(d,a)]^2}{\sigma^2\mu(d,a)}\right\} = 1$  suggests the estimators

$$\hat{\sigma}^2 = \frac{1}{df} \sum_d \sum_a \frac{[n(d,a) - \hat{\mu}(d,a)]^2}{\hat{\mu}(d,a)} \quad (2.7)$$

and

$$\hat{\sigma}^2 = \frac{\sum_d \sum_a [n(d,a) - \hat{\mu}(d,a)]^2}{\sum_d \sum_a \hat{\mu}(d,a)}$$

where  $df$  equals the number of terms in the sum less the number of  $\lambda(a)$ 's that are estimated. In the case where we observe all  $n(d,a)$ 's with  $0 \leq d+a \leq T$ , we have  $df = T(T+1)/2$ . The adequacy of (2.1), (2.6) and the independence of the  $n(d,a)$ 's can be assessed by examining the residuals

$$r(d,a) = \frac{n(d,a) - \hat{\mu}(d,a)}{\hat{\sigma}\hat{\mu}(d,a)^{1/2}}. \quad (2.8)$$

If the overall model is adequate the  $r(d,a)$ 's should look roughly like independent standard normal variables, provided the  $\mu(d,a)$ 's are not too small. If the  $n(d,a)$ 's are small (this will depend on the time units being used to record the counts as well as the claim rates and number of units sold), then it is better to use larger age- and time-intervals in defining (2.7) and (2.8), or to consider other types of residuals; McCullagh and Nelder (1989, chs. 2, 12) may be consulted for more information about residual analysis.

Under the model (2.6) and independent counts, the estimated variance of  $\hat{\Lambda}(t)$  is simply  $\hat{\sigma}^2$  times the Poisson estimate (2.5):

$$\hat{V}(a) = \hat{\sigma}^2 \sum_{u=0}^a \frac{\hat{\lambda}(u)}{R(u)}. \quad (2.9)$$

Extra-Poisson variation may arise from several sources, including inherent variation in the robustness of units, variations in usage environment, and non-Poisson claim patterns for individual units. Kalbfleisch, Lawless and Robinson (1991) consider an approach that



is motivated by assuming that there is an unobservable random variable  $\alpha_i$  associated with each unit sold. The  $\alpha_i$ 's are assumed to be i.i.d., and the  $i$ 'th unit is assumed to generate claims according to a Poisson model with age-specific expected claims  $\alpha_i \lambda(a)$ . This leads to both extra-Poisson variation and a degree of correlation in claim counts corresponding to the same day of sale; we have

$$\begin{aligned} \text{Var}\{n(d, a)\} &= \mu(d, a)[1 + \sigma^2 \lambda(a)] \\ \text{cov}\{n(d, a_1), n(d, a_2)\} &= \sigma_\alpha^2 N(d) \lambda(a_1) \lambda(a_2) \quad a_1 \neq a_2 \\ \text{cov}\{n(d_1, a_1), n(d_2, a_2)\} &= 0 \quad d_1 \neq d_2 \end{aligned} \tag{2.10}$$

where  $\sigma_\alpha^2$  is the variance of the  $\alpha_i$ 's. The parameter  $\sigma_\alpha^2$  may be estimated by equations analogous to (2.7), for example,

$$\sum_d \sum_a \frac{[n(d, a) - \hat{\mu}(d, a)]^2}{\hat{\mu}(d, a)[1 + \sigma^2 \hat{\lambda}(a)]} = df, \tag{2.11}$$

where  $\hat{\mu}(d, a)$ ,  $\hat{\lambda}(a)$  and  $df$  are the same as for (2.7). Under this model,  $\text{Var}\{\hat{\Lambda}(a)\}$  is estimated by

$$\hat{V}(a) = \hat{V}_p(a) + \hat{\sigma}_\alpha^2 \sum_{u=0}^a \frac{\hat{\lambda}(u)^2}{R(u)} + 2\hat{\sigma}_\alpha^2 \sum_{u=1}^a \hat{\lambda}(u) \hat{V}_p(u-1), \tag{2.12}$$

where  $\hat{V}_p(a)$  is the Poisson estimate (2.5). A derivation of this is sketched in the appendix.

As with the previous variance specification, the assumptions (2.1) and (2.10) can be assessed by examining residuals, now defined as

$$r(d, a) = \frac{n(d, a) - \hat{\mu}(d, a)}{\{\hat{\mu}(d, a)[1 + \hat{\sigma}^2 \hat{\lambda}(a)]\}^{1/2}}. \tag{2.13}$$

In many situations either of the variance estimation methods will be satisfactory and in that case, (2.6) is slightly simpler. An additional check on (2.9) and (2.12) is mentioned in the appendix. More refined variance function modelling is beyond the scope of this chapter; in any event, with most types of warranty data detailed modelling of the variance function is not warranted since there will generally be departures from the mean specification (2.1).

Approximate confidence limits for  $\Lambda(a)$  may be obtained by using the fact that, if (2.1) and the variance specification (2.6) or (2.10) used are correct, then  $[\hat{\Lambda}(a) - \Lambda(a)]/\hat{V}(a)^{1/2}$  has an approximate standard normal distribution. For example, an approximate .95 confidence interval for  $\Lambda(a)$  is given by

$$\hat{\Lambda}(a) \pm 1.96\hat{V}(a)^{1/2}.$$

### 2.3 Grouped Data

Frequently data are grouped into total claims for units in various age groups, over specific time periods. For example, the number of claims on units aged 0-30 days, 31-60 days etc. might be given on a monthly basis. If the age groups and reporting periods are of the same fixed length then one possibility is to use the procedures in the previous section with the periods taken as the units of time. However, if the age groups and reporting periods are of different lengths or if the periods are fairly long relative to the time units in which the warranty period is specified, some adjustments are called for.

Suppose that claims are grouped into age intervals  $A_i = [a_{i-1}, a_i)$  with  $a_0 = 0 < a_1 < a_2 < \dots$ . In this case we seek to estimate the expected number of claims for that interval,

$$\Lambda(A_i) = \sum_{a=a_{i-1}}^{a_i-1} \lambda(a). \quad (2.14)$$

A natural estimate is

$$\hat{\Lambda}(A_i) = \frac{n(A_i)}{R(A_i)} \quad (2.15)$$

where, extending the notation used for (2.2) and suppressing  $T$ ,

$$\begin{aligned} n(A_i) &= \sum_t n^*(t, A_i) = \text{number of claims on units of age } a \in A_i \\ R(A_i) &= \frac{1}{a_i - a_{i-1}} \sum_{a=a_{i-1}}^{a_i-1} R(a) \end{aligned} \quad (2.16)$$

where  $R(a) = \sum_{d=0}^{T-a} N(d)$ , as before. To motivate (2.15), we note that

$$\begin{aligned} E\{n(A_i)\} &= \sum_{a=a_{i-1}}^{a_i-1} \sum_{d=0}^{T-a} E\{n(d, a)\} \\ &= \sum_{a=a_{i-1}}^{a_i-1} \lambda(a)R(a), \end{aligned}$$

and assume (with little practical consequence if  $a_i - a_{i-1}$  is not too large) that the  $\lambda(a)$ 's are constant for  $a_{i-1} \leq a < a_i$ . Thus  $E\{n(A_i)\} = \Lambda(A_i)R(A_i)$ , which motivates (2.15).

If sales data are only available in aggregate for different time periods, then (2.16) has to be estimated. The simplest way to do this is to estimate the daily sales figures  $N(d)$  from the available period data, thus providing estimates of the  $R(a)$ 's and of (2.16); the  $N(d)$ 's can be estimated using plausible assumptions about sales patterns. Provided that the sales periods are not too long relative to the age periods, errors due to estimation of the  $R(A_i)$ 's should not be substantial. Further discussion of estimated sales data is given in section 4.

The easiest approach to variance estimation is to assume independence of the  $n^*(t, a)$ 's along with (2.6). Then  $\text{Var}\{n(A_i)\} = \sigma^2 \Lambda(A_i)R(A_i)$ , where we continue to assume that the  $\lambda(a)$ 's are constant over  $A_i$ . Thus by (2.15),  $\text{Var}\{\hat{\Lambda}(A_i)\}$  is estimated by

$$\hat{V}(A_i) = \hat{\sigma}^2 \frac{\hat{\Lambda}(A_i)}{R(A_i)}. \quad (2.17)$$

A suitable estimate for  $\sigma^2$  is

$$\hat{\sigma}^2 = \frac{1}{df} \sum_j \sum_i \frac{\{n^*(P_j, A_i) - \hat{\mu}(P_j, A_i)\}^2}{\hat{\mu}(P_j, A_i)} \quad (2.18)$$

where the sum ranges over all of the time periods ( $P_j$ ) and age intervals ( $A_i$ ) in the data,  $df$  equals the number of terms in the sum minus the number of age intervals, and  $\mu(P_j, A_i) = E\{n^*(P_j, A_i)\}$  is estimated by

$$\hat{\mu}(P_j, A_i) = \frac{\hat{\Lambda}(A_i)}{a_i - a_{i-1}} \sum_{a=a_{i-1}}^{a_i-1} \sum_{t \in P_j} N(t - a). \quad (2.19)$$

As before, the  $N(d)$ 's will need to be estimated if sales data are available only in aggregate form.

Confidence limits for  $\Lambda(A_i)$ 's or sums of  $\Lambda(A_i)$ 's may be obtained by treating the  $\hat{\Lambda}(A_i)$ 's as independent and approximately normally distributed with means  $\Lambda(A_i)$  and variances  $\hat{V}(A_i)$ .

### 3. ADJUSTMENTS FOR REPORTING DELAYS

There are frequently delays in the reporting or ratification of warranty claims and hence delays in their being recorded in the data base used for analysis. Recent claim counts in that case are an underestimate of the number of claims actually made. In this case, one approach to analysis is to exclude recently reported claims: for example, if reporting delays are up to 60 days duration we could, at any point in time, use only data on claims that had occurred at least 60 days before. Another approach, which is important when prompt reviews of warranty claims are needed, is to adjust recent claim counts for underreporting. We describe the techniques presented by Kalbfleisch, Lawless and Robinson (1991).

We assume a stationary reporting delay distribution and let

$$f(r) = Pr(\text{a claim is reported exactly } r \text{ days after it occurs}) \quad r = 0, 1, \dots$$

and  $F(r) = f(0) + \dots + f(r)$ . The term “reported” here refers to the point at which the claim is entered into the data base used for analysis. The  $f(r)$ ’s can be estimated from previous claims data, and we will thus consider them known; Kalbfleisch et al. (1991) and Kalbfleisch and Lawless (1992) discuss estimation of the  $f(r)$ ’s. The procedure for estimating the  $\lambda(a)$ ’s is now to consider counts  $n(d, a, r)$  = the number of age  $a$  claims that had a reporting delay of  $r$  days, for units that were sold on day  $d$ . Then, assuming that  $E\{n(d, a, r)\} = N(d)\lambda(a)f(r)$ , we find that  $\lambda(a)$  can be estimated by the same kind of expression as (2.2),

$$\hat{\lambda}(a) = \frac{n_T(a)}{R_T(a)}, \quad (3.1)$$

where now (temporarily reintroducing the  $T$  in  $n(a)$ ,  $R(a)$  for clarity)

$$\begin{aligned} n_T(a) &= \text{total number of age } a \text{ claims reported up to day } T \\ &= \sum_{d=0}^{T-a} \sum_{r=0}^{T-d-a} n(d, a, r) \\ R_T(a) &= \sum_{d=0}^{T-a} N(d)F(T-a-d). \end{aligned} \quad (3.2)$$

Thus, as in (2.2),  $n_T(a)$  is the number of age  $a$  claims reported, but  $R_T(a)$  is now a discounted number of units reaching age  $a$ , the discounting factor being the probability that a claim occurring on day  $t = a + d$  would be reported by day  $T$  and hence be in the data.

A variance estimate for  $\hat{\Lambda}(a) = \sum_{u=0}^a \hat{\lambda}(u)$  based on the Poisson model is once again given by (2.5), with  $R(a)$  now given by (3.2). It is sensible, however, to allow for extra-Poisson variation. A simple approach is to use the model (2.6) with independent counts, which yields the variance estimate (2.9). In this case  $\sigma^2$  may be estimated by moment methods, as with (2.7). One reasonable estimator, analogous to the second estimator in (2.7), is

$$\hat{\sigma}^2 = \sum_{d=0}^T \sum_{a=0}^{T-d} [n(d, a, \cdot) - \hat{\mu}(d, a, \cdot)]^2 / \sum_{d=0}^T \sum_{a=0}^{T-d} \hat{\mu}(d, a, \cdot), \quad (3.3)$$

where  $\hat{\mu}(d, a, \cdot) = N(d)\hat{\lambda}(a)F(T - d - a)$ . Another way to estimate  $\sigma^2$  is to use the total observed and expected claims for each product unit: this gives

$$\hat{\sigma}^2 = \sum_{i=1}^m [y_i - \hat{\mu}_i]^2 / \sum_{i=1}^m \hat{\mu}_i \quad (3.4)$$

where  $y_i$  is the total number of claims for unit  $i$  ( $i = 1, \dots, m$ ) and  $\hat{\mu}_i = \sum_{a=0}^{T-d_i} F(T - d_i - a)\hat{\lambda}(a)$ , where  $d_i$  is the day unit  $i$  was sold, estimates  $E(y_i)$ .

Kalbfleisch et al. (1991) discuss another model for extra-Poisson variation. They also discuss adjustments when claims or sales data are aggregated, as in section 2.3. In particular, (2.15) and (2.17) still hold, with  $R(a)$  in (2.16) replaced by (3.2).

#### 4. EXAMPLES

We start with an artificial but fairly realistic illustration of age-specific analysis. Consider a product with a one-year warranty period and suppose that  $N(d) = 100$  units are sold on each day  $d = 0, 1, \dots, 364$  of a one-year period. We generated artificial data by assuming that the numbers of claims at age  $a$  for any unit are independent Poisson random variables with constant age-specific claim rates  $\lambda(a) = .002$  for  $a = 0, 1, \dots, 364$ . We assume that reporting delays are distributed over 0 to 59 days, with probabilities  $f(r) = 1/30$  for  $r = 20, \dots, 39$  and  $f(r) = 1/120$  for  $r = 0, \dots, 19$  and  $40, \dots, 59$ . In a real situation we would of course not

have the same number of units sold on each day, and days when businesses are closed would preclude claims being made or reported on certain days, but for simplicity of illustration we ignore such features. Note also that the  $N(d)$ 's are assumed known as each day passes; we discuss how to handle uncertainty about sales in section 5.

Figure 1 shows estimates  $\hat{\Lambda}(a)$ ,  $a = 0, 1, \dots, 364$  based on (3.1) and (3.2) and data on claims reported up to one year after the first units were sold (i.e.  $T = 364$ ). Table 1 shows part of the calculations involved: the first three columns show for selected ages  $a$  the value of  $R_T(a)$  of (3.2) and  $\hat{\Lambda}(a) = \sum_{u=0}^a \hat{\lambda}(u)$ . Figure 1 also shows (lower curve) the biased estimate of  $\Lambda(a)$  that results if we ignore the reporting delays and incorrectly use (2.2) for  $\hat{\lambda}(a)$  with (2.3) for  $R_T(a)$ . The underestimation of  $\Lambda(a)$  is especially severe for  $a$  close to 364 since only a small fraction of claims made close to 364 days (and none past day 345 in this example) have been reported by  $T = 364$  days.

If claims data are grouped into larger age classes, we use (2.15) to estimate the expected claims per unit for each class. The last five columns of Table 1 show results when age at claim is assigned to classes 0-30 days, 31-60 days, and so on, as shown in column 4. Column 5 shows the total claims  $n(A_i)$  (see (2.16)) for each age class  $A_i$ , and column 6 shows  $R(A_i)$  as in (2.16) but with  $R(a)$  defined as in (3.2) to allow for reporting delays. Column 7 shows the estimated average claims per unit  $\hat{\Lambda}(A_i)$  for each age class, and column 8 the cumulative estimates  $\hat{\Lambda}(a_i) = \sum_{j=1}^i \hat{\Lambda}(A_j)$ . It is noted that the  $\hat{\Lambda}(a_i)$ 's are virtually identical to those obtained from the raw data (column 3).

We now consider some real warranty data for a system on a particular car model and year. As of the final data base update 36,683 cars had been sold and 5,760 claims had been reported. For the discussion here day 0 is defined as the day on which the first sales occurred.

Figure 2 shows the pattern of sales by week. We estimated claim reporting delay probabilities from the data; see Kalbfleisch et al. (1991, section 1) and Kalbfleisch and Lawless (1992) for a discussion of this. The estimated cumulative probability function  $F(r)$  is summarized in Table 2. Figure 3 shows the estimated average cumulative claims per vehicle,

$\hat{\Lambda}(a)$ ; for illustration we have shown the curves that result from the sales and claims reported up to each of  $T = 91, 182, 273, 365, 456$  and  $547$  days, respectively. That is, these are the estimates that we would obtain approximately 3, 6, 9, 12, 15 and 18 months, respectively, after the first cars of that model year were sold. We observe that the estimates at 9-18 months agree well but those for  $T = 91$  and  $182$  are somewhat lower, suggesting that cars sold early in the model year had a somewhat lower frequency of early claims than cars sold later. Kalbfleisch et al. (1991) discuss and suggest explanations for this. One point to note is that the estimates  $\hat{\Lambda}(a)$  for  $T = 91$  have rather large standard errors.

Plots like Figure 3 are very useful for tracking warranty claims experience as time progresses. They may also be used to compare claim rates for different time periods or groups of products. In Figure 4 we show estimates of  $\Lambda(a)$  for cars that were manufactured in each of six two-month production periods, going from mid-July to mid-September (period 1) to mid-May to mid-July the following year (period 6). The plots are based on all claims reported up to  $T = 547$  days, of which there were 5701. They reveal the striking fact that average claims per vehicle appear to be very similar for all periods except period 3 (November-January), for which claims are much higher. It would be interesting to determine the source of this difference.

It is a good idea to compute variance estimates or standard errors for  $\hat{\Lambda}(a)$ 's so that we may be aware of the uncertainty about  $\Lambda(a)$ . For example, if we use the variance function (2.6) and estimate  $\sigma^2$  separately for the six production period groups using (3.4), we obtain  $\hat{\sigma}^2 = 1.86, 1.67, 1.87, 1.59, 1.65$  and  $1.30$  respectively. These may be used in (2.9) to estimate  $\text{Var}\{\hat{\Lambda}(a)\}$  and to obtain approximate confidence limits for  $\Lambda(a)$  as described at the end of section 2.2. As an illustration, let us consider groups 1 and 3 and the average claims per vehicle up to age one year (i.e.  $\Lambda(364)$ ). Using (2.9) and the estimates of  $\sigma^2$  we find that the standard errors  $\hat{V}(364)^{1/2}$  for  $\hat{\Lambda}(364)$  are approximately .006 and .0085 for groups 1 and 3, respectively. Approximate .95 confidence limits (i.e.  $\hat{\Lambda}(364) \pm 1.96\hat{V}(364)^{1/2}$ ) for the two periods are roughly .15 - .17 and .21 - .25. It seems clear that the two average claim values

are different. In section 7 we test the equality of the  $\Lambda(a)$  functions for all six groups and find that there is very strong evidence against equality.

We conclude this example with a few additional remarks. We note first that the rather large jumps at the right hand ends of some of the plots in Figures 3 and 4 is due to the fact that for ages  $a$  close to  $T$ ,  $R_T(a)$  is small and the variability in  $\hat{\lambda}(a)$  is consequently large. For point estimation of  $\Lambda(a)$  one may want to smooth the last portion of the plots; standard errors for  $\hat{\Lambda}(a)$  in such cases make it clear that there is considerable uncertainty about  $\Lambda(a)$ , based solely on the data. We remark also that the warranty in this example is one-year, and also has a 12,000 mile limit. Consequently one should not see any claims at ages above one year: Figures 3 and 4 show that there are indeed a few claims allowed, presumably on a courtesy basis or due to data recording errors. Note too that the fact that some vehicles reach the 12,000 mile limit in less than a year and cannot have claims after that does not affect the analysis presented here, since  $\lambda(a)$  is defined as the average number of age  $a$  claims per vehicle sold, taking into account that some vehicles drop out of the warranty coverage before one year. A third point is that we have assumed reporting delay probabilities to be constant over the entire period of the data. This may be checked by examining the delays for claims reported in different time periods. There is no evidence here that delay probabilities are changing but if there were it would be important to adjust the analysis to deal with it; this is readily done. Finally, there are some days on which claims cannot be made or reported, in particular weekend days and holidays. It is usually not worth incorporating this additional complexity into the models since there is little effect on estimates of  $\lambda(a)$  or  $\Lambda(a)$ , and we have not done it here.

## 5. ADJUSTMENTS WHEN SALES ARE ESTIMATED

Frequently the exact number of product units sold in different time periods is unknown to the manufacturer. If sales up to time  $T$  are known to a fairly close approximation, then close approximations to  $R_T(a)$  in (2.2) are available and we may proceed as though these values were known. If only imprecise estimates of sales are available, however, it may be



desirable to reflect this by including an extra term in variance estimates for  $\hat{\lambda}(a)$  or  $\hat{\Lambda}(a)$ , or by a sensitivity analysis to determine the effect of uncertainty about sales.

It may be shown that if  $\hat{R}(a)$ 's are approximately unbiased estimates of  $R(a)$ 's, then the variance of  $\hat{\Lambda}(a)$  is estimated by

$$\hat{V}_1(a) = \hat{V}(a) + \widehat{\text{Var}}\left\{\sum_{u=0}^a \lambda(u) \frac{\hat{R}(u)}{R(u)}\right\}, \quad (5.1)$$

where  $\hat{V}(a)$  is the estimated variance for  $\hat{\Lambda}(a)$  without the adjustment; this is given by (2.5) for the Poisson model or, more generally, (2.9) for the model (2.6) that incorporates extra-Poisson variation. The second term in the right hand side of (5.1) accounts for the fact that the  $R(a)$ 's are estimated. In order to evaluate this term we need to have an estimated covariance matrix for the  $\hat{R}(a)/R(a)$  values, from which an algebraic expression can be obtained. However, the simplest approach is to approximate the distribution of the estimated sales data  $\hat{N}(d)$  by some distribution, and to estimate the right hand part of (5.1) by simulation.

An alternative procedure that is adequate for most practical situations is to ignore the additional variance term in (5.1), but to carry out a check on the sensitivity of estimates  $\hat{\Lambda}(a)$  and  $\hat{V}(a)$  to the estimated  $R(a)$  values. By varying the  $R(a)$ 's in ways that are considered plausible, we can see to what extent  $\hat{\Lambda}(a)$  and  $\hat{V}(a)$  vary, and use this to modify confidence limits for  $\Lambda(a)$  in an informal way.

## 6. PREDICTION

We have so far focussed on the estimation of expected claim counts for a hypothetical infinite population of units, of which those sold are considered a random sample. Such estimates are very useful for summarizing claims experience and for comparing different groups of units. However, we should remember that the population of units that is eventually sold is finite. Thus, if the total number of units sold over the time period  $(0, \tau)$  is  $N. = \sum_{d=0}^{\tau} N(d)$ , then the actual (finite population) average number of claims per unit at age  $a$  is

$$m(a) = \frac{\sum_{d=0}^{\tau} n(d, a)}{N.} \quad a = 0, 1, 2, \dots$$

and

$$M(a) = \sum_{u=0}^a m(u)$$

is the average number of claims per unit up to age  $a$ . If data on claims up to time  $T$  are available, then the  $m(a)$ 's are partially known, and the problem of estimating them reduces to one of predicting the  $n(d, a)$ 's that are still unobserved. This problem is discussed in detail by Kalbfleisch et al. (1991), and we summarize the main points.

Sensible estimates of  $m(a)$  and  $M(a)$  based on data to time  $T$  are clearly

$$\hat{m}(a) = \hat{\lambda}(a), \quad \hat{M}(a) = \hat{\Lambda}(a) \quad (6.1)$$

where  $\hat{\lambda}(a)$  and  $\hat{\Lambda}(a)$  are given by (2.2) and (2.4). The difference that consideration of the finite population makes is in the variance estimates: under the extra-Poisson dispersion model with variance function (2.6), we have (assuming  $\tau = T$ , so that  $N. = N$ )

$$\text{Var}\{\hat{m}(a) - m(a)\} = \sigma^2 \left\{ \frac{N - R(a)}{NR(a)} \right\} \lambda(a). \quad (6.2)$$

In addition, the  $\hat{m}(a)$ 's are independent, and

$$V_{Pred}(a) = \text{Var}\{\hat{M}(a) - M(a)\} = \sigma^2 \sum_{u=0}^a \left\{ \frac{N - R(u)}{NR(u)} \right\} \lambda(u). \quad (6.3)$$

The overdispersion parameters  $\sigma^2$  and the  $\lambda(a)$ 's are estimated as in section 2.2; setting  $\sigma^2 = 1$  gives the Poisson distribution results. Note that if  $R(u) = N$  for  $0 \leq u \leq a$  (which implies that all units sold over  $0, 1, \dots, T$  have already reached age  $a$  by time  $T$ ) then  $V_{Pred}(a) = 0$ . This is as it should be, since in that case  $M(a) = \hat{M}(a)$  is fully known.

Grouped data are handled similarly, according to the procedures in section 2.3. In particular, quantities  $M(A_i) = \sum_{a=A_i-1}^{A_i-1} m(a)$  (see (2.14)) are estimated as  $\hat{M}(A_i) = \hat{\Lambda}(A_i)$ , from (2.15), and  $\text{Var}\{\hat{M}(A_i) - M(A_i)\}$  is estimated by (compare (2.17))

$$\hat{V}_{Pred}(A_i) = \hat{\sigma}^2 \left\{ \frac{N - R(A_i)}{NR(A_i)} \right\} \hat{\Lambda}(A_i) \quad (6.4)$$

where  $R(A_i)$  is given by (2.16). Calculations are outlined in the appendix.

Reporting delays as discussed in section 3 create no additional problems: we merely replace  $R(a) = \sum_{d=0}^{T-a} N(d)$  with (3.2) in any formulas.

One may wish to predict the total numbers of claims, as opposed to the average claims per unit; this is also easily done. For example, suppose that we wish to predict  $n^*(t, a)$ , the number of age  $a$  claims in a future time period  $t$ . To do this we consider  $n^*(t, a) - \hat{\mu}^*(t, a)$  where  $\hat{\mu}^*(t, a) = N(t - a)\hat{\lambda}(a)$  is a point estimate of  $n^*(t, a)$  or of  $\mu^*(t, a)$ . It is clear that

$$\begin{aligned} \text{Var}\{n^*(t, a) - \hat{\mu}^*(t, a)\} &= \sigma^2 \mu^*(t, a) + N(t - a)^2 \frac{\sigma^2 \lambda(a)}{R(a)}, \\ &= \sigma^2 N(t - a) \lambda(a) \left\{ 1 + \frac{N(t - a)}{R(a)} \right\} \end{aligned} \quad (6.5)$$

where we continue to use the variance model (2.6).

Prediction intervals for quantities such as  $m(a)$ ,  $M(a)$  or  $n^*(t, a)$  are obtained by treating  $m(a) - \hat{m}(a)$ ,  $M(a) - \hat{M}(a)$  or  $n^*(t, a) - \hat{\mu}^*(t, a)$ , respectively, as approximately normal with mean 0 and variance estimated from (6.2), (6.3) or (6.5), respectively. As earlier, such approximate intervals work best when the quantity being predicted is not too small.

## 7. COVARIATES AND REGRESSION ANALYSIS

The methods described so far are appropriate when the expected number of age  $a$  claims per unit does not vary substantially with respect to calendar time or other factors. Sometimes expected claims may depend on factors such as manufacturing conditions or the environment in which the unit is used, however. The car data at the end of section 4 provide a graphic example.

In situations where a few separate conditions can be identified as of possible importance, the simplest procedure is to analyze the data separately under the different conditions; this was the approach taken at the end of section 4 when we examined the claims experience of cars manufactured in six distinct production periods. Another approach that is especially useful when conditions thought to affect claims cannot be categorized with a few groups is regression analysis. In this case covariates are associated with units or groups of units, and a model for the way that the covariates affect the expected claims is postulated. Regression

models for count data such as claims have been widely studied, and standard methods and software are available for analysis (e.g. McCullagh and Nelder, 1989, especially ch. 6). We will merely outline a few points, since a thorough treatment of regression is outside the scope of this chapter.

We suppose that there is a vector of covariates  $z_i$  associated with each individual unit,  $i$ , that enters service. The expected number of claims at age  $a$  for unit  $i$ , entering service on day  $d$ , can be conveniently modelled in the log linear form

$$E\{n_i(d, a)\} = \mu_i(d, a) = \lambda(a)e^{z_i'\beta}, \quad (7.1)$$

where  $\beta$  is a vector of regression parameters. The covariate  $z_i$  can specify, for example, manufacturing characteristics, model lines, and, if they are available, even time-varying factors such as environmental variables. In this latter case,  $z_i$  could vary according to the current calendar time  $t = d + a$ .

The models (7.1), with variance function  $\text{Var}\{n_i(d, a)\} = \sigma^2\mu_i(d, a)$  to allow for extra-Poisson variation, may be fitted by Poisson maximum likelihood or quasi-likelihood methods (McCullagh and Nelder 1989, ch. 6). Most general statistical analysis systems (e.g. SAS, BMDP, GLIM, SPSS, S) perform the necessary calculations. Checks on the form of (7.1) may be made by examination of residuals, as described in McCullagh and Nelder (1989) and other texts.

As a simple example, we consider the comparison of different production periods, described in the second example of section 4, via regression analysis.

We define a covariate vector  $z_i = (z_{i1}, \dots, z_{i5})'$  where  $z_{ij} = 1$  if vehicle  $i$  was produced in period  $j$  and 0 if not ( $j = 1, \dots, 5$ ), and use the regression model (7.1). We can test whether the  $\Lambda(a)$  functions for the six periods are equal by considering the hypothesis  $H : \beta = 0$ . A test of  $H$  may be carried out by obtaining  $\hat{\beta}$  and its estimated asymptotic variance (e.g. see McCullagh and Nelder 1989, ch. 6 or Breslow 1990)  $V(\hat{\beta})$ . If  $H$  is true,  $\hat{\beta}'V(\hat{\beta})^{-1}\hat{\beta}$  is approximately distributed as  $\chi^2_{(5)}$ . Lawless and Nadeau (1993) find  $\hat{\beta}'V(\hat{\beta})^{-1}\hat{\beta} = 85.6$  with one particular choice of variance estimate. The 99.9 percentile of  $\chi^2_{(5)}$  is only 20.52, so there

is clearly very strong evidence against  $H$ , as seemed clear from Figure 4 and the standard errors referred to in section 4.

## 8. CALENDAR TIME EFFECTS

Sometimes there are systematic calendar time effects on rates of claims. For example, a monotonic trend over time might occur because of an improvement or deterioration in the reliability of the units being produced over time, whereas products that experience a pronounced seasonal usage pattern typically show a similar pattern in warranty claims.

Often calendar time effects can be examined simply by estimating age-specific expected claims via the models (2.1), separately for claims occurring in different time periods. When calendar time is an important (perhaps predominant) factor in the occurrence of claims, however, it is best to focus on the counts  $n^*(t, a)$  and their expected values  $\mu^*(t, a)$  described in section 2.1. If  $\mu^*(t, a) = E\{n^*(t, a)\}$  is the expected number of claims in time period  $t$  due to units of age  $a$  (i.e. units sold in period  $t - a$ ) then

$$\mu^*(t, a) = N(t - a)g(t, a), \quad (8.1)$$

where  $g(t, a)$  is the expected number of claims per unit for a unit of age  $a$  at calendar time  $t$ . Models for  $g(t, a)$  may be fitted and examined using the log linear regression software mentioned in section 7. The obvious place to start is with a “main effects” model with

$$g(t, a) = h(t)\lambda(a), \quad (8.2)$$

where the  $h(t)$ 's and  $\lambda(a)$ 's are nonnegative parameters. As (8.2) stands, the parameters are identifiable only up to a multiplicative constant. Identifiability may be achieved by placing a single constraint on the parameters, for example  $h(0) = 1$  or  $\lambda(0) = 1$ . An examination of residuals  $r^*(t, a) = n^*(t, a) - \hat{\mu}^*(t, a)$  will usually indicate whether (8.2) is suitable. Tests may also be carried out for hypotheses of interest; for example, if the  $\lambda(a)$ 's in (8.2) are equal there is no age effect (i.e. the expected number of claims per unit does not vary with age, though it may vary with calendar time), so we may wish to test that  $\lambda(0) = \lambda(1) = \dots = \lambda(T)$ . Such tests are easily carried out with the log linear framework; see McCullagh and Nelder

(1989, ch. 6). Note that our age-specific models of sections 2-5 are given by (8.2) with  $h(t) =$  a constant.

Finally, we could also fit parametric models for  $h(t)$  or  $\lambda(a)$  of (8.2), or more generally for  $g(t, a)$  in (8.1). If, for example, we suspected a monotonic trend in calendar time, we might consider

$$\mu^*(t, a) = N(t - a)\lambda(a)e^{\beta t}, \quad (8.3)$$

where  $\beta$  and the  $\lambda(a)$ 's are unknown parameters to be estimated.

## 9. ANALYSIS OF COSTS

Warranty costs may also be analyzed using the methods described in this chapter. To do this we assume that claim costs are indexed by  $c = 1, 2, \dots, r$  and  $k(c)$  is the cost of a claim in group  $c$ . The amount of grouping used will depend on the application, but there is no difficulty with  $r$  being large, as for example if dollar amounts were used. We will consider just the case of age-specific analysis, and let  $\lambda_c(a)$  be the expected number of claims of cost  $k(c)$  for a unit at age  $a$ . Under the assumptions of section 2,  $\lambda_c(a)$  is estimated by

$$\hat{\lambda}_c(a) = \frac{n_c(a)}{R(a)} \quad (9.1)$$

where, in an obvious notation,  $n_c(a) = \sum_{d=0}^{T-a} n_c(d, a)$  is the number of claims of age  $a$  and cost  $k(c)$  up to time  $T$ . Examination of the estimated cumulative expected frequencies

$$\hat{\Lambda}_c(a) = \sum_{u=0}^a \hat{\lambda}_c(u) \quad (9.2)$$

is often useful. Cumulative total expected cost per unit up to age  $a$  is estimated by

$$\hat{K}(a) = \sum_{c=1}^r k(c)\hat{\Lambda}_c(a). \quad (9.3)$$

Predictions and associated variance estimates may also be readily obtained. Let

$$M_c(a) = \frac{1}{N} \sum_{u=0}^a \sum_{d=0}^T n_c(d, u)$$

be the actual average number of cost  $k(c)$  repairs per unit, up to age  $a$ , for units sold over  $0, 1, \dots, T$  and

$$K(a) = \sum_{c=1}^r k(c)M_c(a)$$

the average total cost per unit. Clearly  $M_c(a)$  and  $K(a)$  are estimated by (9.2) and (9.3), respectively. In addition,

$$\text{Var}\{\hat{K}(a) - K(a)\} = \sum_{c=1}^r k^2(c)\text{Var}\{M_c(a) - \hat{M}_c(a)\},$$

where  $\text{Var}\{M_c(a) - \hat{M}_c(a)\}$  is given by (6.3), with  $\lambda_c(u)$  in place of  $\lambda(u)$ . Prediction limits for the average total cost  $K(a)$  per unit up to age  $a$  are obtained as in section 6. In particular,  $\hat{K}(a) \pm 1.96\widehat{\text{Var}}\{\hat{K}(a) - K(a)\}^{1/2}$  is an approximate .95 prediction interval for  $K(a)$ .

## 10. ESTIMATION OF FIELD RELIABILITY

Warranty data are a valuable source of information about the reliability of manufactured products. However, for reliability assessments it is desirable to study failure or repair processes for individual units. Therefore, the type of warranty must be taken into consideration, since it affects how long units stay under observation (i.e. under warranty coverage). In addition, the amount of data kept on each unit will affect how well certain quantities may be estimated. It is beyond the scope of this chapter to provide a comprehensive treatment in this area, but we will summarize a few important points.

1. Suppose that a particular type of failure is of interest and that when such failures occur under warranty they result in a claim. If we wish to estimate the expected number of failures  $\lambda^F(a)$  per unit at age  $a$ , and the cumulative expected number  $\Lambda^F(a) = \sum_{u=0}^a \lambda^F(u)$ , then we must adjust the  $\hat{\lambda}(a)$  values discussed earlier in this chapter for the probability  $p(a)$  that a unit failing at age  $a$  is under warranty (and that the claim is reported, if this is not always done). Since  $\lambda(a) = \lambda^F(a)p(a)$ , we can estimate  $\lambda^F(a)$  as

$$\hat{\lambda}^F(a) = \hat{\lambda}(a)/\hat{p}(a), \tag{10.1}$$

provided that we are able to obtain an estimate  $\hat{p}(a)$  of  $p(a)$ . With motor vehicles, for example, warranties typically have both age and mileage limits. In that case we could base  $\hat{p}(a)$  on the warranty limits (e.g. two years or 20,000 miles) and data on the rates at which vehicles in the population accumulate mileage. Vehicle manufacturers usually collect some such data by means of surveys. Note, however, that (10.1) is in this case appropriate only for failures that depend primarily on age, and not on mileage. See points 4 and 5 for additional remarks.

2. Sometimes the distribution of time to the first failure of some type is of interest. If the dates of sale (or of replacement for units replaced under warranty) are available for each unit and if we have only age- or calendar time-based warranty limits, then estimation of the time to failure distribution can be based on well known methods for censored lifetime data (e.g. Kalbfleisch and Prentice 1980, Lawless 1982, Nelson 1982). If, however, dates of sale are not available until a warranty claim is made then estimation either has to be based on data conditional on failure having occurred by a certain age, or on assumptions regarding sales.

To be specific, suppose that  $t_i$  represents the age at failure of the  $i$ 'th unit, and that  $t_i$  has a distribution with density function  $f(t; \theta)$  and cumulative distribution function  $F(t; \theta) = \int_0^t f(u; \theta) du$ , where  $\theta$  is an unknown parameter. Suppose that we consider warranty data for units sold over the calendar time period  $(0, T)$ , and that the warranty age limit is  $A$ . If the  $i$ 'th unit is sold at  $d_i$ , then we will observe its failure time  $t_i$  only if  $t_i \leq \tau_i = \min(T - d_i, A)$ . If the sales dates  $d_i$  are available for all of the  $N$  units sold over  $(0, T)$ , then the  $\tau_i$ 's may be calculated for all  $i = 1, \dots, N$ . Then, if the  $n$  units  $i = 1, \dots, n$  experience failures and the remainder do not, we know that  $t_i > \tau_i$  for items  $i = n + 1, \dots, N$ , and the probability of the data gives the so-called censored data likelihood

$$L_1 = \prod_{i=1}^n f(t_i; \theta) \prod_{i=n+1}^N \{1 - F(\tau_i; \theta)\}. \quad (10.2)$$

If, however, the  $d_i$ 's only become available when a claim is made, then the  $\tau_i$ 's cannot



be calculated for  $i = n + 1, \dots, N$ . In this case, we still have a likelihood based on  $t_1, \dots, t_n$ , given that  $t_i \leq \tau_i$  in each case; this is

$$L_2 = \prod_{i=1}^n \left\{ \frac{f(t_i; \theta)}{F(\tau_i; \theta)} \right\}. \quad (10.3)$$

Kalbfleisch and Lawless (1992) discuss truncated data in some detail.

Usually  $L_2$  is a lot less informative about  $\theta$  than  $L_1$  (e.g. Kalbfleisch and Lawless 1988). It is therefore desirable to have dates of sale for all items, but if these are not available, it is worth trying to estimate them from sales information. In many cases it is possible to get sufficiently accurate estimates of  $N$  and  $\tau_{n+1}, \dots, \tau_N$  (e.g. from estimates of the numbers of items sold in different days over  $(0, T)$ ) to allow us to use (10.2).

3. A distinction should be made between average failure rates (or other characteristics) across the population of units in service and the form of the failure processes for individual units. Aggregate warranty data as discussed in earlier sections allow us to estimate population average characteristics such as  $\lambda^F(a)$  in point 1, but not to ascertain the nature of individual processes; for this we need data on individual units. For repeated events we would in particular need to keep track of the events occurring for each unit. For example, for engineering purposes it is of much interest whether a certain average failure rate arises from each unit having roughly the same rate, or from a situation where most units experience no failures but a few experience many.
4. Sometimes there is an easily measurable "usage" time and it is wished to estimate failure rates or distributions in terms of that. For example, with motor vehicles it is common to evaluate certain reliability characteristics in terms of mileage. Assuming that mileage is recorded for units experiencing failures under warranty, the main difficulty is similar to that in point 2: the usage (mileage) at which units are censored are usually unknown and consequently even the likelihood (10.3) may not be available. To overcome this it is necessary to postulate a model for mileage accumulation; Lawless and Kalbfleisch (1992, section 3.2) indicate how this can be done.

5. When failures or other events depend upon both age and other factors (including usage), we must have ways of observing or modelling these other factors. In particular, if the factors are observed only for units that experience warranty claims, then a model has to be postulated for how these factors are distributed in the population of units in service. One can learn about this by supplementing the warranty data with random samples from the population; see Suzuki (1985) and Kalbfleisch and Lawless (1988) for methods based on this approach.

## 11. CONCLUDING REMARKS

The methods presented in this chapter are designed to allow us to portray warranty claims data clearly, and to extract as much information as possible from them. The main focus was on analyzing aggregate claims data; this is important for understanding total claims behaviour and factors affecting it either for specific types of claims or for all types combined. The methods also allow claim numbers or costs to be predicted from past data. Section 10 dealt briefly with another issue, namely the use of warranty data for reliability estimation. For thorough analysis it is usually necessary to supplement the warranty data with additional information. Robinson and McDonald (1991) provide a nice discussion.

Finally, we should stress that the foundation of good statistical analysis is good data. If warranty data are to be unambiguous and useful, it is important that claims and sales information be obtained in a timely and accurate way and that it be accessible for analysis.

## APPENDIX

### 1. Variance Estimates for Age-Specific Expected Claims Estimates

Under the Poisson model or the model with variance function (2.6), the  $n(d, a)$ 's are independent, with  $\text{Var}\{n(d, a)\}$  given by (2.5). Thus the  $n_T(a)$ 's in (2.2) are independent for  $a = 0, 1, \dots, T$ , with variance  $\sigma^2 \sum_{d=0}^{T-a} \mu(d, a) = \sigma^2 R_T(a) \lambda(a)$ , so the  $\hat{\lambda}(a)$ 's are independent, with  $\text{Var}\{\hat{\lambda}(a)\} = \sigma^2 \lambda(a) / R_T(a)$ . This leads to (2.5) and (2.9), the former being given by the case where  $\sigma = 1$ .

For the model (2.10) the terms in  $n_T(a) = \sum_{d=0}^{T-a} n(d, a)$  are independent, so  $\text{Var}\{n_T(a)\} = \sum_{d=0}^{T-a} \mu(d, a)[1 + \sigma^2 \lambda(a)] = R_T(a)\lambda(a)[1 + \sigma^2 \lambda(a)]$ , and so

$$\text{Var}\{\hat{\lambda}(a)\} = \frac{\lambda(a)[1 + \sigma^2 \lambda(a)]}{R(a)}.$$

Because of correlation between  $n(d, a_1)$  and  $n(d, a_2)$ ,  $\hat{\lambda}(a_1)$  and  $\hat{\lambda}(a_2)$  are correlated. We have, for  $a_1 \neq a_2$ ,

$$\begin{aligned} \text{cov}\{n_T(a_1), n_T(a_2)\} &= \sum_{d=0}^{\min(T-a_1, T-a_2)} \text{cov}\{n(d, a_1), n(d, a_2)\} \\ &= \sigma^2 \lambda(a_1) \lambda(a_2) R_T\{\max(a_1, a_2)\}. \\ \therefore \text{cov}\{\hat{\lambda}(a_1), \hat{\lambda}(a_2)\} &= \frac{\sigma^2 \lambda(a_1) \lambda(a_2)}{R_T\{\min(a_1, a_2)\}}. \end{aligned}$$

Then,

$$\begin{aligned} \text{Var}\{\hat{\Lambda}(a)\} &= \sum_{u=0}^a \text{Var}\{\hat{\lambda}(u)\} + 2 \sum_{u < v} \text{cov}\{\hat{\lambda}(u), \hat{\lambda}(v)\} \\ &= V_p(a) + \sigma^2 \sum_{u=0}^a \frac{\lambda(u)^2}{R(u)} + 2\sigma^2 \sum_{u=1}^a \lambda(u) V_p(u-1), \text{ giving (2.11)}. \end{aligned}$$

A robust variance estimate for  $\hat{\Lambda}(a)$  can also be obtained (Lawless and Nadeau 1993), provided that data on individual units is available. In this case, let  $m$  be the number of products sold by time  $T$ , and for unit  $i$  let  $d_i$  be the day of sale and  $n_i(u)$  the number of age  $u$  claims. Then the robust estimate for  $\text{Var}\{\hat{\Lambda}(a)\}$  is

$$\hat{V}_R(a) = \sum_{i=1}^m \left\{ \sum_{u=0}^{\min(a, T-d_i)} \frac{1}{R(u)} [n_i(u) - \hat{\lambda}(u)] \right\}^2. \quad (\text{A.1})$$

Variance estimates (2.9) and (2.12), which are based on specific assumptions about variance, may be checked against (A.1), which is valid under quite general conditions.

## 2. Variance Estimates (6.4) for Prediction with Grouped Data

The calculations follow the same lines as in Kalbfleisch et al. (1991, Appendix B). We note that (ignoring reporting delays for simplicity)

$$M(A_i) - \hat{M}(A_i) = \frac{1}{N} \sum_{a=A_i-1}^{A_i-1} \left\{ \sum_{d=0}^{T-a} n(d, a) + \sum_{d=T-a+1}^T n(d, a) \right\} - \sum_{a=A_i-1}^{A_i-1} \sum_{d=0}^{T-a} \frac{n(d, a)}{R(A_i)},$$

where

$$R(A_i) = \frac{1}{a_i - a_{i-1}} \sum_{a=a_{i-1}}^{a_i-1} R(a),$$

as described in section 2.3. Since the  $n(d, a)$ 's are mutually independent with variance function (2.6), we find that

$$\begin{aligned} \text{Var}\{M(A_i) - \hat{M}(A_i)\} &= \left\{ \frac{N - R(A_i)}{NR(A_i)} \right\}^2 \sum_a \sum_{d=0}^{T-a} \sigma^2 N(d) \lambda(a) + \frac{1}{N^2} \sum_a \sum_{d=T-a+1}^T \sigma^2 N(d) \lambda(a) \\ &= \left\{ \frac{N - R(A_i)}{NR(A_i)} \right\}^2 \sigma^2 \Lambda(A_i) R(A_i) + \frac{1}{N^2} \sigma^2 \Lambda(A_i) \{N - R(A_i)\} \\ &= \left\{ \frac{N - R(A_i)}{NR(A_i)} \right\} \sigma^2 \Lambda(A_i). \end{aligned}$$

### ACKNOWLEDGEMENTS

This research was supported in part by grants from the Manufacturing Research Corporation of Ontario and the Natural Sciences and Engineering Research Council of Canada. The authors thank Claude Nadeau for suggestions and computational assistance.

### REFERENCES

- Breslow, N. (1990). Tests of hypotheses in overdispersed Poisson regression and other quasi-likelihood models. *J. Amer. Statist. Assoc.* **85**, 565-571.
- Kalbfleisch, J.D. and Lawless, J.F. (1988). Estimation of reliability in field performance studies (with discussion). *Technometrics* **30**, 365-388.
- Kalbfleisch, J.D. and Lawless, J.F. (1992). Some useful statistical methods for truncated data. *J. Quality Tech.* **24**, 145-152.
- Kalbfleisch, J.D., Lawless, J.F. and Robinson, J.A. (1991). Methods for the analysis and prediction of warranty claims. *Technometrics* **33**, 273-285.

- Kalbfleisch, J.D. and Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. New York: John Wiley and Sons.
- Lawless, J.F. (1982). *Statistical Models and Methods for Lifetime Data*. New York: John Wiley and Sons.
- Lawless, J.F. and Kalbfleisch, J.D. (1992). Some issues in the collection and analysis of field reliability data. In *Survival Analysis: State of the Art*, pp. 141-152, eds. J.P. Klein and P.K. Goel. Amsterdam: Kluwer.
- Lawless, J.F. and Nadeau, J.C. (1993). Some simple robust methods for the analysis of recurrent events. Manuscript.
- McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*. London: Chapman and Hall.
- Nelson, W.B. (1982). *Applied Life Data Analysis*. New York: John Wiley and Sons.
- Robinson, J.A. and McDonald, G.C. (1991). Issues related to field reliability and warranty data. In *Data Quality Control*, 69-90, eds. G.E. Liepins and V.R.R. Uppuluri. New York: Marcel Dekker.
- Suzuki, K. (1985). Estimation of lifetime parameters from incomplete field data. *Technometrics* **27**, 263-271.

Table 1. Estimation of  $\Lambda(a)$  from Data at  $T = 364$  Days

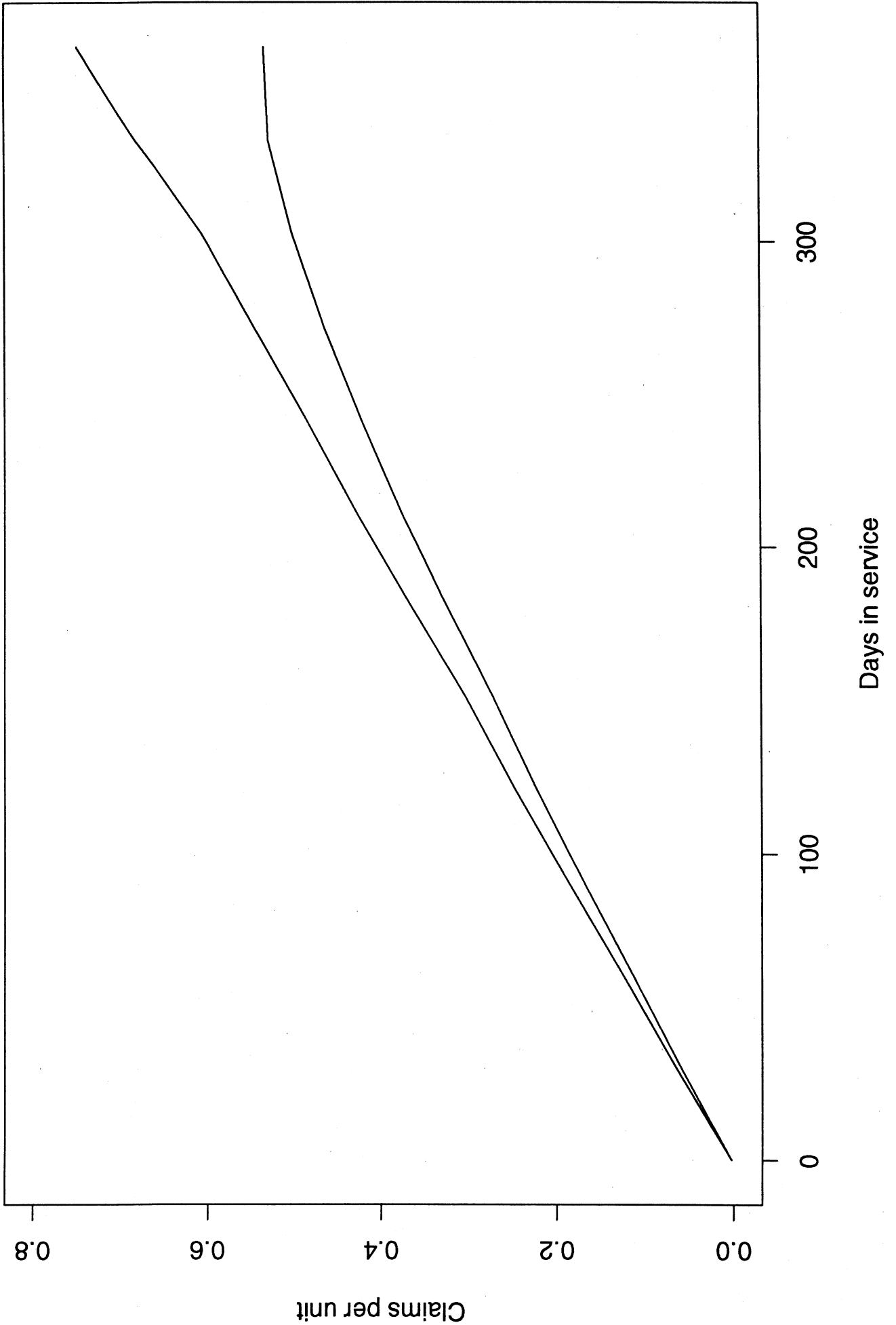
$a$	$R_T(a)$	$\hat{\Lambda}(a)$	$A_i = (a_{i-1}, a_i)$	$n(A_i)$	$R(A_i)$	$\hat{\Lambda}(A_i)$	$\hat{\Lambda}(a_i)$
0	33,550	.002					
30	30,550	.064	(0,30)	2,064	32,050	.0644	.064
60	27,550	.124	(31,60)	1,726	29,000	.0595	.124
90	24,550	.185	(61,90)	1,586	26,000	.0610	.185
121	21,450	.247	(91,121)	1,416	22,950	.0617	.247
151	18,450	.304	(122,151)	1,132	19,900	.0569	.304
181	15,450	.366	(152,181)	1,053	16,900	.0623	.366
211	12,450	.428	(182,211)	870	13,900	.0626	.428
242	9,450	.488	(212,242)	648	10,850	.0597	.488
272	6,352	.548	(243,272)	470	7,800	.0603	.548
303	3,250	.610	(273,303)	294	4,750	.0619	.610
333	635	.684	(304,333)	136	1808.5	.0752	.686
364	.83	.750	(334,364)	10	169.7	.0589	.744

Table 2. Estimated Probability  $F(r)$  that Reporting Delay is  $\leq r$  Days

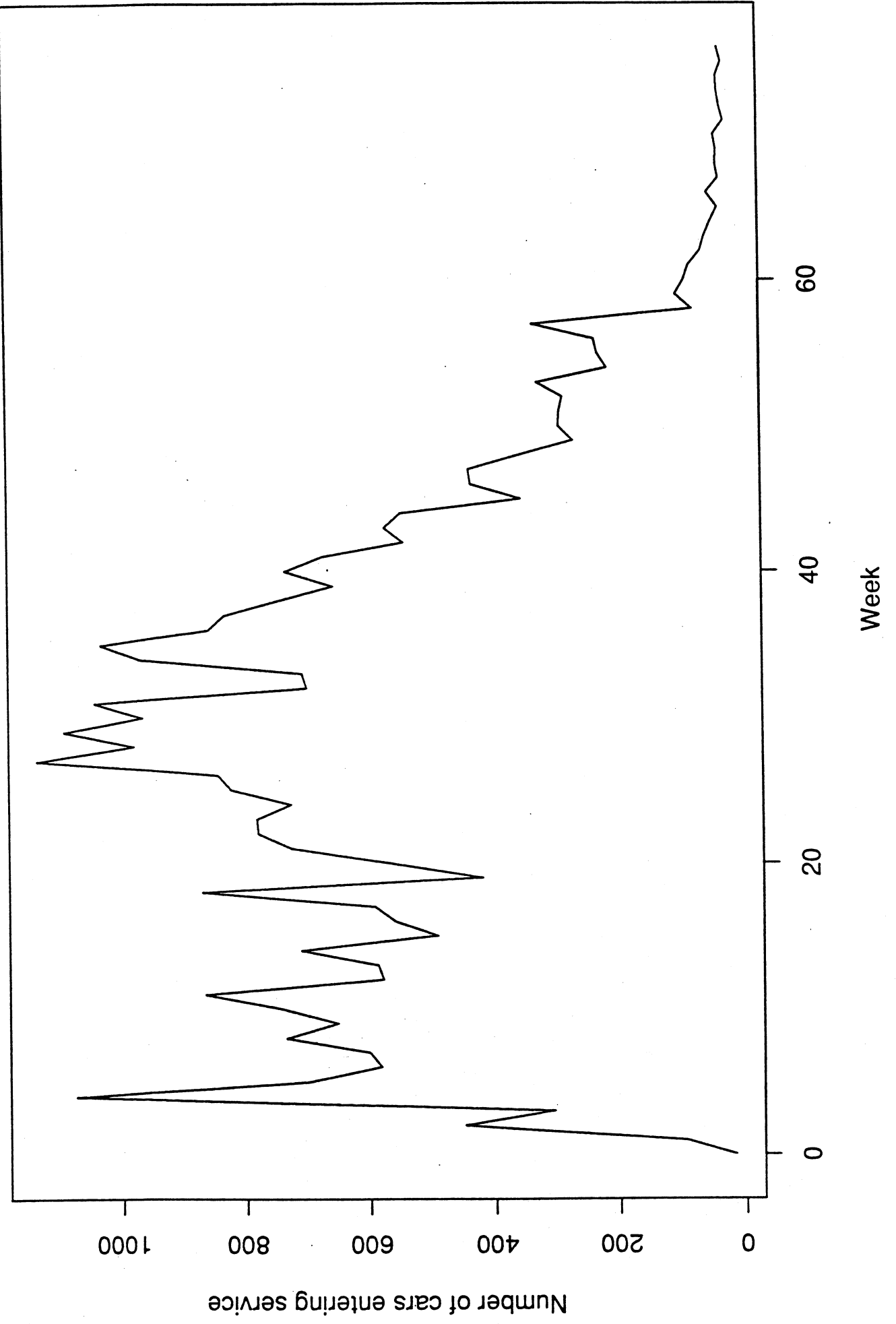
$r$	$F(r)$
10	.058
20	.553
30	.811
40	.903
50	.934
60	.960
80	.981
100	.989
120	.992

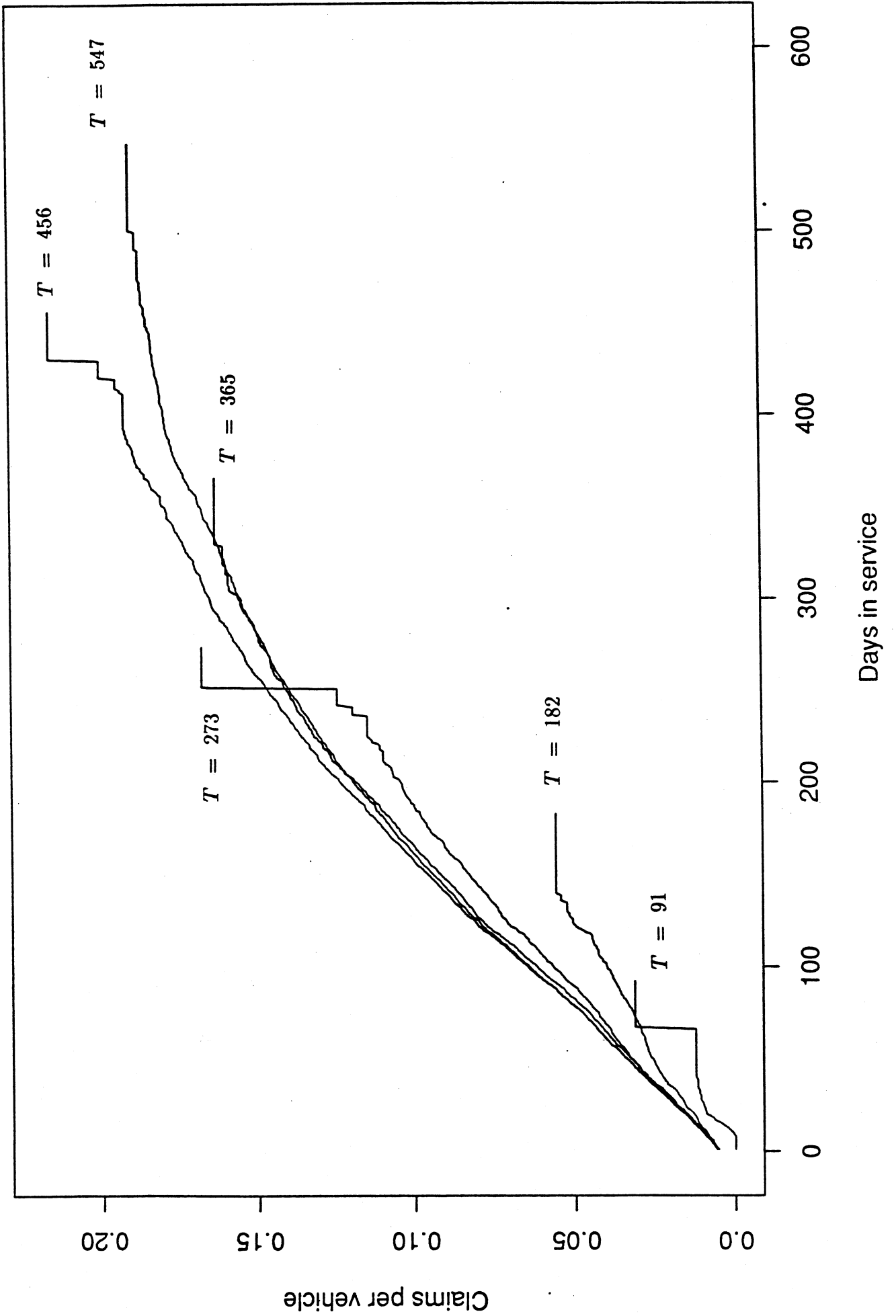
## FIGURE LEGENDS

- Figure 1. Estimated Average Cumulative Claims per Unit,  $\hat{\Lambda}(a)$ , Based on Artificial Data. (The lower curve ignores reporting delays and so is biased.)
- Figure 2. Number of Cars Entering Service (Sold), by Week.
- Figure 3. Estimated Average Cumulative Claims per Car,  $\hat{\Lambda}(a)$ , Based on Claims Reported to Various Days  $T$ .
- Figure 4. Estimated Average Cumulative Claims per Car,  $\hat{\Lambda}(a)$ , Stratified According to Production Period.









# Warranty Claim Frequencies

