Chapter 4

Basic models and simple applications

We use models to represent reality in simpler ways, so we can understand, describe, predict, or control it better. In statistics, the reality is the 'variable phenomena,' without which there is no statistics (and life will be rather dull). In modelling we typically separate a systematic pattern from the stochastic variation. The systematic pattern may describe a relationship between variables of interest, while the stochastic variation usually represents the unaccounted part of the total variation. Probability models are used to represent this stochastic variation.

From a statistical point of view, any probability model has a potential for data analysis, although we tend to concentrate on models that are convenient from analytical and computational aspects, and have parameters that are easy to interpret. The models covered in this chapter are the basic building block for real data analytic models. Only the simplest systematic structure is present, mostly in terms of one-sample or one-group structure. The exceptions are the comparison of two binomial proportions and two Poisson means.

4.1 Binomial or Bernoulli models

The Bernoulli model is useful for experiments with dichotomous outcomes. Each experimental unit is thought of as a trial; in its simplest form the Bernoulli trials are assumed to be independent, each with probability θ for the success outcome. Some simple examples: a coin tossing experiment, the sex of children in a family, the success or failure of business enterprises, the success or failure of a medical procedure, etc. In twin studies it is common to collect data where at least one of the twins has a particular condition; in these studies a pair of twins is the 'trial' and the 'success' occurs if both twins have the condition. A twin study usually tries to establish if there is a genetic factor in the condition under study.

Suppose we observe $x \equiv (x_1, \ldots, x_n)$, which are a realization of a Bernoulli experiment with $P(X_i = 1) = \theta$ and $P(X_i = 0) = 1 - \theta$. Then

$$L(\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \tag{4.1}$$

$$= \theta^{\sum x_i}(1-\theta)^{n-\sum x_i} \tag{4.2}$$

Example 4.1: For example, we observe a sequence 0111110111, then $n = 10$, $\sum x_i = 8$, and the likelihood function for θ is given in Figure 4.1. It is easy to show the maximum likelihood estimate is $\widehat{\theta} = \sum x_i/n = 0.8$. The likelihood interval at 15% cutoff (computed numerically) is (0.50,0.96); since the likelihood is reasonably regular this is an approximate 95% confidence interval (CI) for θ. (From hereon, we will only report one interval or the other depending on the regularity of the likelihood, and we will err on the regularity side.)

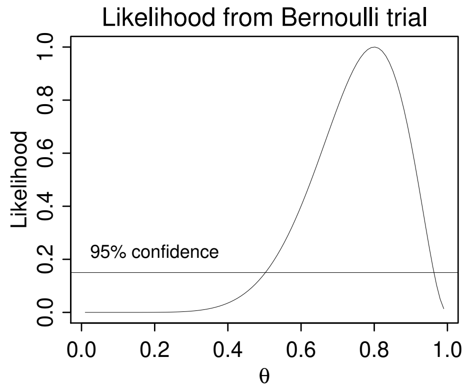

Figure 4.1: *Likelihood function of θ from a Bernoulli experiment with outcome 0111110111, so $n = 10$ and $\sum x_i = 8$. This is the same likelihood from a binomial experiment with $n = 10$ trials and $x = 8$ successes. The approximate 95% CI is (0.50,0.96).*

We have described the likelihood construction for a simple binomial observation in Section 2.1. For completeness, suppose $X \sim \text{binomial}(n, \theta)$ and we observe $X = x$, then

$$L(\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}. \tag{4.3}$$

If we observe $x = 8$ successes in $n = 10$ trials, the likelihood function and other quantities are exactly as obtained above. This is sensible since knowing the order of the 0-1 outcomes in an independent Bernoulli experiment should not add any information about θ; this would not be the case if, for example, the sequence is dependent and the strength is dependence is also a function of θ.

Discrete data are usually presented in grouped form. For example, suppose $x_1, \ldots, x_N$ are an iid sample from binomial(n, θ) with n known. We first summarize the data as

k	0	1	$\cdots$	n
n_k	n_0	n_1	$\cdots$	n_n

where n_k is the number of x_i's equal to k, so $\sum_k n_k = N$. We can now think of the data $(n_0, \ldots, n_n)$ as having a multinomial distribution with probabilities $(p_0, \ldots, p_n)$

given by the binomial probabilities

$$p_k = \binom{n}{k} \theta^k (1-\theta)^{n-k}.$$

The log-likelihood is given by

$$\log L(\theta) = \sum_{k=0}^{n} n_k \log p_k.$$

We can show that the MLE of θ is

$$\widehat{\theta} = \frac{\sum_k k n_k}{Nn}.$$

with standard error

$$\mathrm{se}(\widehat{\theta}) = \sqrt{\frac{\widehat{\theta}(1-\widehat{\theta})}{Nn}}.$$

Example 4.2: In a classic study of human sex ratio, based on the hospital records in Saxony, in 1889 Geissler reported the distribution of the number of boys per family. Among 6115 families with 12 children we observe:

No. boys k	0	1	2	3	4	5	6	7	8	9	10	11	12
No. families n_k	3	24	104	286	670	1033	1343	1112	829	478	181	45	7

The estimated proportion of boys is

$$\widehat{\theta} = \frac{\sum_k k n_k}{6115 \times 12} = \frac{38,100}{6115 \times 12} = 0.5192,$$

with standard error $\mathrm{se}(\widehat{\theta}) = 0.0018$. (For comparison, in Ireland in 1994 the proportion of boys among 48,255 births is 0.5172.)

Negative binomial model

In the so-called negative or inverse binomial experiment we continue a Bernoulli trial with parameter θ until we obtain x successes, where x is fixed in advance. Let n we the number of trials needed, then the likelihood function is

$$\begin{aligned} L(\theta) &= P_\theta(N=n) & (4.4) \\ &= \binom{n-1}{x-1} \theta^x (1-\theta)^{n-x} & (4.5) \end{aligned}$$

Again here we find the same likelihood function as the one from the binomial experiment, even though the sampling property is quite different. This is an example where the likelihood function ignores the sampling scheme. Further uses of the negative binomial model are given in Section 4.5, which connects this model with the Poisson model.

4.2 Binomial model with under- or overdispersion

The binomial model specifies a rigid relationship between the mean and the variance, which is a weakness as a modelling tool.

Example 4.2: continued Using the estimate $\widehat{\theta} = 0.5192$ we can compute the expected frequencies

$$e_k = N\widehat{p}_k$$

and obtain the goodness-of-fit statistic

$$\chi^2 = \sum_k \frac{(n_k - e_k)^2}{e_k} = 110.5,$$

highly significant at 11 degrees of freedom. To see the nature of the model violation, define a residual

$$r_k = \frac{n_k - e_k}{\sqrt{e_k}}$$

such that $\chi^2 = \sum_k r_k^2$.

No. boys k	0	1	2	3	4	5	6	7	8	9	10	11	12
Observed n_k	3	24	104	286	670	1033	1343	1112	829	478	181	45	7
Expected e_k	1	12	72	258	628	1085	1367	1266	854	410	133	26	2
Residual r_k	2.1	3.4	3.8	1.7	1.7	−1.6	−0.7	−4.3	−0.9	3.4	4.2	3.7	3.0

Hence the observed frequencies tend to be larger at the edges, and lower in the middle of the distribution, indicating greater variability than expected under the binomial model. □

We now describe theoretically how the standard binomial model might fail. Suppose $X_1, \ldots, X_n$ are independent Bernoulli trials with probabilities $p_1, \ldots, p_n$. Let $X = \sum_i X_i$, then

$$E(X) = \sum p_i \equiv n\theta$$

where $\theta = \sum p_i/n$, but

$$\begin{aligned}
\text{var}(X) &= \sum_i \text{var}(X_i) \\
&= \sum_i p_i(1 - p_i) \\
&= \sum_i p_i - \sum_i p_i^2 \\
&= \sum_i p_i - (\sum_i p_i)^2/n - \{\sum_i p_i^2 - (\sum_i p_i)^2/n\} \\
&= n\theta - n\theta^2 - n\sigma_p^2 \\
&= n\theta(1 - \theta) - n\sigma_p^2
\end{aligned}$$

where we have defined the variance among p_i's as

$$\sigma_p^2 = \frac{1}{n}\left\{\sum_i p_i^2 - \frac{(\sum_i p_i)^2}{n}\right\}.$$

So, allowing individual Bernoulli probabilities to vary produces less-than-standard binomial variance.

Now, suppose $X_1, \ldots, X_m$ are independent binomial(n, p_1), ..., binomial(n, p_m), and let $X = X_I$ be a random choice from one of these X_i's, i.e., the random index $I = i$ has probability $1/m$. This process produces a mixture of binomials with marginal probability

$$\begin{aligned} P(X = x) &= E\{P(X_I = x|I)\} \\ &= \frac{1}{m}\sum_{i=1}^{m} P(X_i = x) \\ &= \frac{1}{m}\sum_i \binom{n}{x} p_i^x (1-p_i)^{n-x}, \end{aligned}$$

which does not simplify further, but

$$\begin{aligned} E(X) &= E\{E(X_I|I)\} \\ &= \frac{1}{m}\sum_i E(X_i) \\ &= \frac{n}{m}\sum_i p_i \equiv n\theta \end{aligned}$$

where we set $\theta = \sum p_i/m$. And the variance is

$$\begin{aligned} \text{var}(X) &= E\{\text{var}(X_I|I)\} + \text{var}\{E(X_I|I)\} \\ &= \frac{1}{m}\sum_i \text{var}(X_i) + \frac{1}{m}\sum_i E^2(X_i) - (\sum_i E(X_i)/m)^2 \\ &= \frac{1}{m}\sum_i np_i(1-p_i) + \frac{1}{m}\sum_i (np_i)^2 - (n\theta)^2 \\ &= n\theta(1-\theta) + n(n-1)\sigma_p^2 \end{aligned}$$

where σ_p^2 is the variance among p_i's as defined above. So, here we have greater-than-standard binomial variation.

The equal sampling probability simplifies the previous evaluation of mean and variance, but the extra-binomial variance is always observed when we sample from a mixed population. For example, if families in the population have different probabilities p_i's, then a random sample of families will exhibit values with extra-binomial variance.

Wrong binomial assumption can adversely impact inference by producing wrong standard errors. Section 4.10 describes a general model that allows under- or overdispersion within the general exponential family models. In random effects modelling (Section 17.1) the individual estimates of p_i's maybe of interest, in which case we typically put more structure on how they vary. Lindsey and Altham (1998) analyse Geissler's (complete) data taking the overdispersion into account; see also Exercise 4.4.

4.3 Comparing two proportions

Comparing two binomial proportions is probably the most important statistical problem in epidemiology and biostatistics.

Example 4.3: A geneticist believes he has located a new gene that controls the spread or metastasis of breast cancer. He analyzed the presence of such gene in the cells of 15 patients whose cancer had spread (metastasized) and 10 patients with localized cancer. The first group had 5 patients with the gene, while one patient in the second group had the gene. The data are usually presented in a 2×2 table. Is there evidence to justify his belief?

	spread	localized	total
present	5(33%)	1(10%)	6
absent	10	9	19
total	15	10	25

It is instructive to start with the (large-sample) frequentist solution. First put some names on the elements of the 2×2 table

	spread	localized	total
present	x	y	$t = x + y$
absent	$m - x$	$n - y$	$N - t$
total	m	n	$N = m + n$

The standard test of equality of proportions is the famous χ^2-test given by

$$\chi^2 = \frac{N\{x(n-y) - y(m-x)\}^2}{mnt(N-t)},$$

which, under the null hypothesis, has χ^2_1 distribution.

For the observed data

$$\chi^2 = \frac{25(5 \times 9 - 1 \times 10)^2}{15 \times 10 \times 6 \times 19} = 1.79,$$

producing P-value=0.18; this is a two-sided P-value for the hypothesis of equal proportion. This method has an appealing simplicity, but note that in small samples its validity is doubtful, and it does not give full information about the parameter of interest (e.g., it is not clear how to get a confidence interval).

In small sample situations, the so-called Fisher's exact test is commonly used. Under the null hypothesis and conditional on the observed margins, the probability of an observed table is a hypergeometric probability

$$p(x) = \frac{\binom{m}{x}\binom{n}{y}}{\binom{m+n}{x+y}}.$$

Fisher's exact P-value is then computed as the probability of the observed or more-extreme tables. For the above example, there is only one more-extreme table, i.e., when we got 0 'present' out of 10 localized cases. The one-sided P-value is

$$\text{P-value} = \frac{\binom{15}{5}\binom{10}{1}}{\binom{25}{6}} + \frac{\binom{15}{6}\binom{10}{0}}{\binom{25}{6}} = 0.17 + 0.03 = 0.20.$$

To proceed with a likelihood analysis, suppose the number of successes X in the first group is binomial $B(m, \pi_x)$, and, independently, Y in the second group is $B(n, \pi_y)$. On observing x and y the joint likelihood of (π_x, π_y) is

$$L(\pi_x, \pi_y) = \pi_x^x (1-\pi_x)^{m-x} \pi_y^y (1-\pi_y)^{n-y}.$$

Comparison of two proportions can be expressed in various ways, for example using the difference $\pi_x - \pi_y$, the relative risk π_x/π_y or the log odds ratio θ defined by

$$\theta = \log \frac{\pi_x/(1-\pi_x)}{\pi_y/(1-\pi_y)}.$$

In terms of θ the null hypothesis of interest H_0: $\pi_x = \pi_y$ is equivalent to H_0: $\theta = 0$. Each parameterization has its own advantage/disadvantage in terms of interpretation and statistical implications.

The reader can check that, in small samples, the likelihood of the log odds-ratio is more regular than the likelihood of the other parameters. So, we will consider the log odds-ratio parameter θ as the parameter of interest. Any other parameter can be used as a nuisance parameter, but for convenience we will use the log odds η defined by

$$\eta = \log \frac{\pi_y}{1-\pi_y}.$$

Some simple algebra would show that

$$\begin{aligned} \pi_y &= \frac{e^\eta}{1+e^\eta} \\ \pi_x &= \frac{e^{\theta+\eta}}{1+e^{\theta+\eta}}. \end{aligned}$$

Therefore, we get the joint likelihood

$$\begin{aligned} L(\theta, \eta) &= \left(\frac{\pi_x}{1-\pi_x}\right)^x (1-\pi_x)^m \left(\frac{\pi_y}{1-\pi_y}\right)^y (1-\pi_y)^n \\ &= \left(\frac{\pi_x/(1-\pi_x)}{\pi_y/(1-\pi_y)}\right)^x \left(\frac{\pi_y}{1-\pi_y}\right)^{x+y} (1-\pi_x)^m (1-\pi_y)^n \\ &= e^{\theta x} e^{\eta(x+y)} (1+e^{\theta+\eta})^{-m} (1+e^\eta)^{-n}. \end{aligned}$$

The MLE of θ is available directly from the invariance property:

$$\widehat{\theta} = \log \frac{x/(m-x)}{y/(n-y)}.$$

The standard error has an interesting formula

$$\mathrm{se}(\widehat{\theta}) = \left(\frac{1}{x} + \frac{1}{y} + \frac{1}{m-x} + \frac{1}{n-y} \right)^{1/2}.$$

Its derivation is left as an exercise.

To get the profile likelihood of θ we can compute the MLE of η at each fixed value of θ, but there is no closed form formula for the MLE. The profile likelihood has to be computed numerically according to the definition:

$$L(\theta) = \max_{\eta} L(\theta, \eta).$$

Example 4.3: continued. Figure 4.2(a) shows the contours of the joint likelihood at 10%, 30%, 50%, 70% and 90% cutoff. The profile likelihood in Figure 4.2(b) shows little evidence of the gene associated with the spread of cancer. The likelihood-based 95% CI for θ is (−0.54,4.52), with the corresponding 95% CI (0.58,92.22) for the odds ratio. The MLE of θ is

$$\widehat{\theta} = \log \frac{5/10}{1/9} = 1.50$$

with $\mathrm{se}(\widehat{\theta}) = 1.19$.

Now consider the analysis of the more extreme table: $x = 6$ out of $m = 15$, versus $y = 0$ out of $n = 10$. The MLE of θ is $\widehat{\theta} = \infty$, and the likelihood is irregular. Figure 4.2(c) shows the joint likelihood, and Figure 4.2(d) the profile likelihood. The 15% likelihood interval for θ is now a naturally one-sided interval $\theta > 0.93$. The likelihood of H_0: $\theta = 0$ is around 2.5%, indicating a stronger evidence of association. □

A series of 2×2 tables

In practice, we often stratify the study groups according to some risk factor if the risk factor is not balanced. For example, the evidence of cancer depends on age, and the two groups being compared have different age distributions. In epidemiology, age is called a confounding variable. Stratifying by age reduces the comparison bias due to confounding. Thus we divide the study subjects to young and old strata, and for each stratum we construct a 2×2 table.

Assuming the tables have a common odds-ratio parameter θ, we can compute the profile likelihoods $L_i(\theta)$'s from each table, and combine them using

$$\log L(\theta) = \sum_i \log L_i(\theta).$$

This method works if the number of strata is small, otherwise there can be a serious bias problem; the proper method to combine information from many sparse tables is given in Section 10.5.

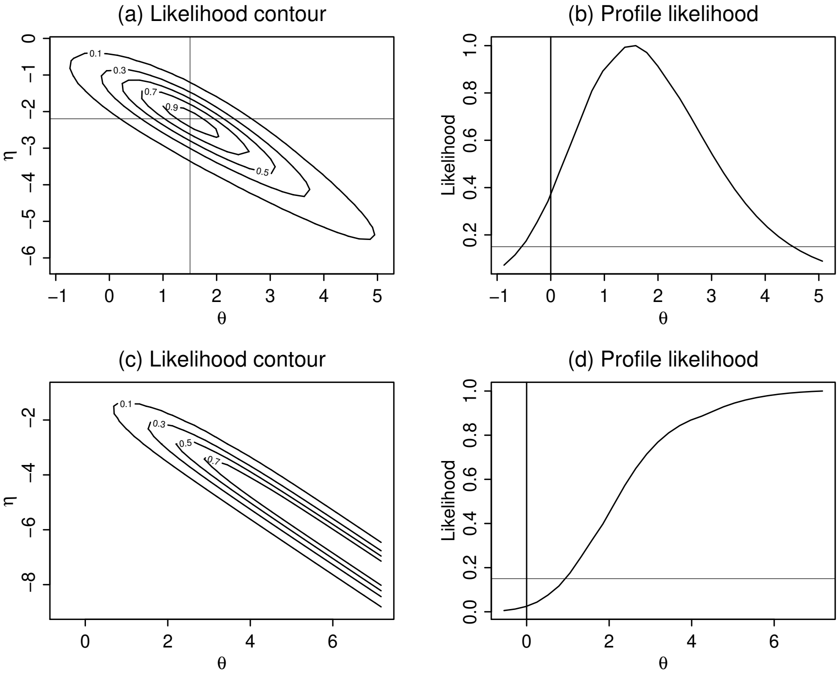

Figure 4.2: *(a) Joint likelihood of θ and η for the original data; (b) Profile likelihood of θ; (c) For extreme data: 6 out of 15 versus 0 out of 10; (d) Profile likelihood from (c).*

4.4 Poisson model

A discrete random variable X has a Poisson distribution with parameter θ if

$$P(X = x) = e^{-\theta}\frac{\theta^x}{x!}.$$

This model is extremely versatile for modelling any data involving counts: the number of accidents in a freeway each year, the number of deaths due to a certain illness per week, the number of insurance claims in a region each year, the number of typographical errors per page in a book, etc. The main difference with the binomial model is that there is no obvious notion of a total of n independent units, each having a fixed probability.

For modelling purposes, it is fruitful to remember that the Poisson model with mean θ is an approximation of the binomial model with large n and a small success probability $\pi = \theta/n$:

$$\begin{aligned} P(X = x) &= \binom{n}{x}\pi^x(1-\pi)^{n-x} \\ &= \frac{n!}{(n-x)!x!}\left(\frac{\theta}{n}\right)^x\left(1-\frac{\theta}{n}\right)^{n-x} \\ &= \frac{n\cdot(n-1)\cdots(n-x+1)}{n^x}\times\frac{\theta^x}{x!}\left(1-\frac{\theta}{n}\right)^{n-x} \end{aligned}$$

$$\rightarrow \frac{\theta^x}{x!} e^{-\theta}.$$

On observing an iid sample $x_1, \ldots, x_n$ from Poisson(θ), the likelihood of θ is given by

$$L(\theta) = e^{-n\theta} \theta^{\sum x_i}.$$

The maximum likelihood function is $\widehat{\theta} = \overline{x}$ with standard error $\text{se}(\widehat{\theta}) = \sqrt{\overline{x}/n}$. We can check that the likelihood is quite regular if $\sum x_i$ is large enough; this is true even for $n = 1$.

Example 4.4: For each year in the past 5 years, a town recorded 3, 2, 5, 0 and 4 earthquakes (of at least a certain magnitude). Assuming a Poisson model, the likelihood function of the earthquake frequency θ is given in Figure 4.3. There was an average of $\widehat{\theta} = 2.8$ (se = 0.83) earthquakes per year, with approximate 95% CI $1.6 < \theta < 4.5$.

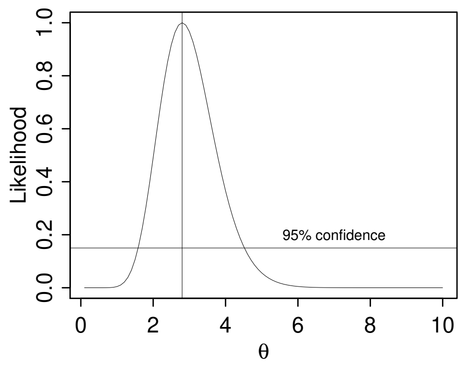

Figure 4.3: *Likelihood function of the earthquake frequency.*

Example 4.5: The Poisson assumption can be checked if we have grouped data. Jenkins and Johnson (1975) reported 76 incidents of international terrorism in the USA between January 1968 and April 1974. The data are categorized according to the monthly number of incidents as follows

Number of incidents k	Number of months n_k
0	38
1	26
2	8
3	2
4	1
12	1

The idea of a Poisson plot (Hoaglin and Tukey 1985) is to plot k against a function of n_k such that it is expected to be a straight line for Poisson data. Since $p_k = P(X = k) = e^{-\theta}\theta^k/k!$, then

$$\log p_k + \log k! = -\theta + k \log \theta,$$

indicating that we should plot $\log(p_k k!)$ versus k, where p_k is estimated by $n_k/\sum n_k$. Figure 4.4 shows that the months with $k = 0, \ldots, 3$ incidents follow the Poisson prescription, but

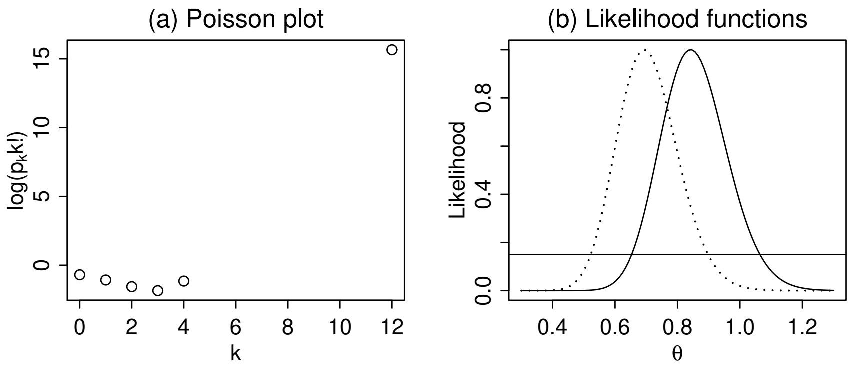

Figure 4.4: *(a) The Poisson plot of the incidents of international terrorism in the USA. The months with 4 and 12 incidents are unusual; (b) Likelihood function of the mean number of incidents including (solid) and excluding (dotted) the month with* $k = 12$ *incidents.*

the month with 4 incidents is rather unusual and the one with 12 incidents is extremely unusual. (It turns out that 11 of those 12 incidents were carried out by an anti-Castro group. This raises inductive questions as discussed in Section 1.3: how should we defined an event? Should we instead count the groups involved?)

In general, for a Poisson model with mean θ the likelihood based on observing n_k, for $k = 0, 1, \ldots$ is

$$L(\theta) = \prod_k p_k^{n_k},$$

where $p_k = P(X = k) = e^{-\theta}\theta^k/k!$, and the log-likelihood is

$$\log L(\theta) = -\theta \sum_k n_k + \sum_k k n_k \log \theta.$$

We can show that the MLE $\widehat{\theta} = \sum_k k n_k / \sum_k n_k$. Figure 4.4 shows the likelihood functions including and excluding the month with $k = 12$. It is clear that the likelihood is sensitive to the outlier. The mean number of incidents is 0.84 and 0.69, with approximate 95% CI's given by $0.65 < \theta < 1.06$ and $0.53 < \theta < 0.96$.

4.5 Poisson with overdispersion

Like the binomial model, the Poisson model imposes a strict relationship between the mean and variance that may not be appropriate for the data.

Example 4.6: The following table summarizes the number of accidents among factory workers (Greenwood and Yule 1920). For example, 447 workers did not have any accident.

Number of accidents k	Number of workers n_k	Fitted number Poisson	Negative binomial
0	447	406.3	446.2
1	132	189.0	134.1
2	42	44.0	44.0
3	21	6.8	14.9
4	3	0.8	5.1
> 4	2	0.1	2.7

Assuming Poisson model, the average accident rate is

$$\sum_k kn_k / \sum_k n_k = 0.47.$$

To simplify the computation, assume that the last category is represented by 5 accidents. Comparing the observed and fitted frequencies, the Poisson model is clearly inadequate. The χ^2 statistic is 107.3, with 4 degrees of freedom. The observed data show an overdispersion: there are more accident-free and accident-prone workers than predicted by the Poisson model.

Since Poisson model is a limit of the binomial, models for under- or overdispersion we describe in Section 4.2 also apply to this case. General modelling of Poisson-type data with extra variation is described under the general exponential family models in Section 4.10.

Overdispersion can occur if, conditionally on μ, an outcome X_μ is Poisson with mean μ, and μ is random with mean $E\mu$ and variance σ^2. For example, individuals vary in their propensity to accidents, so even if the number of accidents per individual is Poisson, the marginal distribution will show some overdispersion. In this setup, the marginal distribution of X_μ has

$$\begin{aligned} E(X_\mu) &= E\{E(X_\mu|\mu)\} = E\mu \\ \mathrm{var}(X_\mu) &= E\{\mathrm{var}(X_\mu|\mu)\} + \mathrm{var}\{E(X_\mu|\mu)\} \\ &= E\mu + \mathrm{var}(\mu) \\ &= E\mu + \sigma^2, \end{aligned}$$

showing an extra variability compared to the standard Poisson model.

If μ has a gamma distribution we will get a closed form formula for the the marginal probabilities. Specifically, let X_μ be Poisson with mean μ, where μ has density

$$f(\mu) = \frac{1}{\Gamma(\alpha)} \mu^{\alpha-1} \lambda^\alpha e^{-\lambda\mu}.$$

A random sample of X from the the mixture of X_μ for all μ has mean

$$E(X) = E\mu = \frac{\alpha}{\lambda}$$

and variance

$$\begin{aligned} \mathrm{var}(X) &= E\mu + \mathrm{var}(\mu) \\ &= \frac{\alpha}{\lambda} + \frac{\alpha}{\lambda^2}. \end{aligned}$$

For likelihood modelling we need to compute the marginal probability of $x = 0, 1, \ldots$:

$$\begin{aligned} P(X=x) &= E\{P(X_\mu = x|\mu)\} \\ &= E\left(e^{-\mu}\frac{\mu^x}{x!}\right) \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)x!}\int e^{-\mu}\mu^x\mu^{\alpha-1}e^{-\lambda\mu}\,d\mu \\ &= \frac{\lambda^\alpha\Gamma(x+\alpha)}{(\lambda+1)^{x+\alpha}\Gamma(\alpha)x!} \\ &\equiv \binom{x+\alpha-1}{\alpha-1}\left(\frac{\lambda}{\lambda+1}\right)^\alpha\left(1-\frac{\lambda}{\lambda+1}\right)^x. \end{aligned} \tag{4.6}$$

using as a definition $(x+\alpha-1)! \equiv \Gamma(x+\alpha)$. For integer α we have a negative binomial distribution: the outcome x is the number of failures recorded when we get exactly α successes, and the probability of success is

$$\pi = \frac{\lambda}{\lambda+1}.$$

Note, however, that as a parameter of the gamma distribution α does not have to be an integer. Given the probability formula, and on observing data $x_1, \ldots, x_n$, we can construct the likelihood of (α, λ) or (α, π).

Example 4.6: continued To fit the negative binomial model to the data, the log-likelihood of (α, π) is

$$\log L(\alpha, \pi) = \sum_{k=0}^{4} n_k \log P(X=k) + n_{[>4]}\log\{1 - P(X>4)\}$$

where $P(X=k)$ is given by (4.6). Numerical optimization of the log-likelihood yields

$$\widehat{\alpha} = 0.84, \quad \widehat{\pi} = 0.64.$$

The fitted frequencies shown in the previous table are now much closer to the observed; the χ^2 statistic is 3.7, with 3 degrees of freedom, indicating a very good fit. The estimated accident rate is $0.84(1-0.64)/0.64 = 0.47$; this is the same as in Poisson model, but has different profile likelihood and standard error (Exercise 4.16). □

4.6 Traffic deaths example

In our prototypical problem suppose the number of traffic deaths increases from $x = 170$ last year to $y = 190$ this year. Is this a significant increase? Let assume that the number of deaths X and Y are independent Poisson with parameters λ_x and λ_y. Then the likelihood function is

$$L(\lambda_x, \lambda_y) = e^{-(\lambda_x+\lambda_y)}\lambda_x^x\lambda_y^y.$$

Here we are only interested in comparing the two rates, not in the absolute level.

Define the parameter of interest $\theta = \lambda_y/\lambda_x$, and consider λ_x as the nuisance parameter. (One might also consider λ_y as nuisance). We have reparameterized (λ_x, λ_y) as $(\lambda_x, \theta\lambda_x)$, so using the invariance principle we obtain

$$L(\theta, \lambda_x) = e^{-\lambda_x(1+\theta)}\lambda_x^{x+y}\theta^y.$$

For each fixed θ the MLE for λ_x is $\widehat{\lambda}_x(\theta) = (x+y)/(1+\theta)$, so the profile likelihood of θ is

$$\begin{aligned} L(\theta) &= e^{-\widehat{\lambda}_x(\theta)(1+\theta)}\widehat{\lambda}_x(\theta)^{x+y}\theta^y \\ &= \text{constant} \times \left(\frac{\theta}{1+\theta}\right)^y \left(\frac{1}{1+\theta}\right)^x \end{aligned}$$

The MLE of θ is $\widehat{\theta} = y/x = 1.12$. Figure 4.5 shows the profile likelihood and the approximate 95% CI of θ is (0.91,1.37). So, there is not enough evidence to claim that there is an increase.

Suppose the death count is now redefined as the number of drivers involved in the accidents (several deaths maybe associated with one driver), and this is split according to the age of drivers. The following table summarizes the data. The profile

Age group	Last year	Current year	$\widehat{\theta}$	Approx. 95% CI
Under 20	20	35	1.75	(1.05,3.05)
20–30	40	54	1.35	(0.93,2.03)
40–50	75	65	0.87	(0.65,1.20)
Over 50	10	2	0.20	(0.05,0.73)

Table 4.1: *Traffic deaths data according to the number of drivers involved, categorized by the drivers' age, and the summaries from the likelihood functions.*

likelihood functions for the four age groups are shown in Figure 4.5. We see here the need to summarize the plot: it is too busy and contains too much information. In the table, each likelihood is represented by the MLE and the 95% CI interval. The data show that the accident rate ratio θ is a function of age and it appears that the accident rate has increased among the 'under 20' group, but has not increased for other age groups and in fact significantly dropped for the 'over 50' group.

4.7 Aspirin data example

The analysis of the aspirin data example (Section 1.1) is similar with the traffic death example. Suppose that the number of heart attacks in the active group X_a is binomial(n_a, θ_a) and that in the placebo group X_p is binomial(n_p, θ_p). We observed $x_a = 139$ from a total $n_a = 11,037$ subjects, and $x_p = 239$ from a total of $n_p = 11,034$. The parameter of interest is $\theta = \theta_a/\theta_p$. Since n_a and n_p are large, while the event

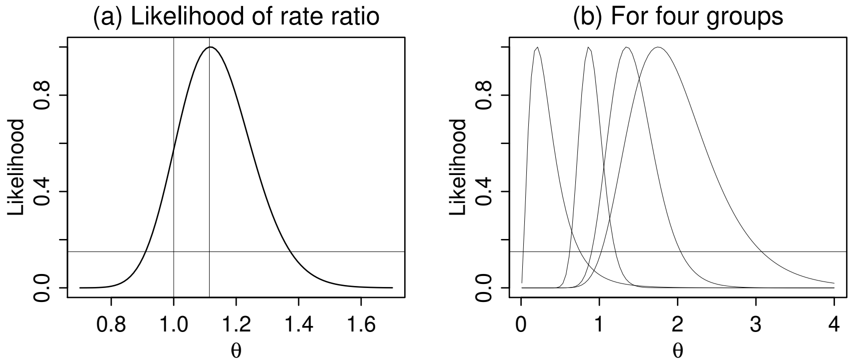

Figure 4.5: *(a) Profile likelihood of rate ratio θ. (b) Profile likelihood of θ for the four age groups shown in Table 4.1.*

rates are small, X_a and X_p are approximately Poisson with parameter $n_a\theta_a$ and $n_p\theta_p$. The analysis can be further simplified by using $n_a \approx n_p$, though the simplification is minor. The likelihood of θ_p and $\theta = \theta_a/\theta_p$ is

$$\begin{aligned} L(\theta, \theta_p) &= e^{-(n_a\theta_a + n_p\theta_p)}(n_a\theta_a)^{x_a}(n_p\theta_p)^{x_p} \\ &= \text{constant} \times e^{-\theta_p(n_a\theta + n_p)}\theta^{x_a}\theta_p^{x_a+x_p} \end{aligned}$$

and one proceeds in the usual way to produce the profile likelihood of θ (Exercise 4.19)

$$L(\theta) = \text{constant} \times \left(\frac{n_a\theta}{n_a\theta + n_b}\right)^{x_a}\left(1 - \frac{n_a\theta}{n_a\theta + n_b}\right)^{x_p}.$$

Exactly the same theory also applies for the number of strokes. The profile likelihood of θ for heart attacks and stroke are shown in Figure 4.6. The approximate 95% CIs for θ are (0.47,0.71) and (0.93,1.59), respectively. There is a significant benefit of aspirin in reducing heart attacks, but the evidence for increased rate of stroke is not significant.

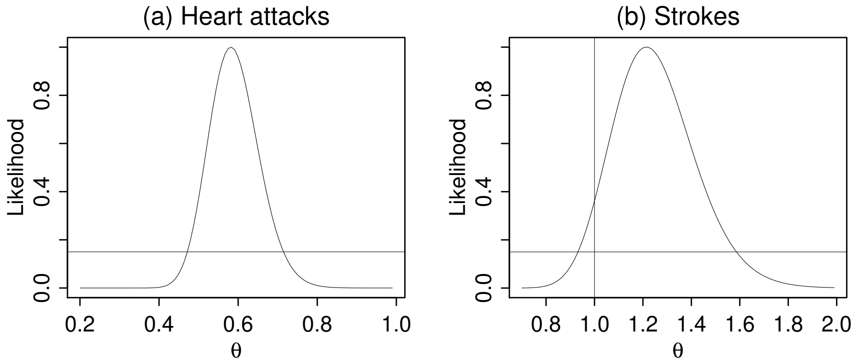

Figure 4.6: *(a) Profile likelihood of θ for the number of heart attacks. The approximate 95% CI is (0.47,0.71). (b) Profile likelihood of θ for the number of strokes. The approximate 95% CI is (0.93,1.59).*

4.8 Delta method

As a comparison we will analyze the aspirin data example in a more ad hoc way, without using any likelihood. We will use the Delta method, one of the most useful classical tools to derive distributions of statistical estimates.

Assuming that X_a and X_p are Poisson with rate $n_a\theta_a$ and $n_p\theta_p$, we can show (Exercise 4.20) that the conditional distribution of X_a given X_a+X_p is binomial with parameters X_a+X_p and $\pi = n_a\theta_a/(n_a\theta_a + n_b\theta_b)$. Since $n_a = 11,307 \approx n_p = 11,304$, we approximately have $\pi = \theta/(\theta+1)$ or $\theta = \pi/(1-\pi)$. On observing $x_a = 139$ and $x_a + x_p = 378$, we get

$$\begin{aligned} \widehat{\pi} &= \frac{139}{378} = 0.368 \\ \widehat{\theta} &= \frac{\widehat{\pi}}{1-\widehat{\pi}} = 0.58. \end{aligned}$$

We then use the following theorem to get a confidence interval for θ.

Theorem 4.1 (Delta Method) *Let $\widehat{\psi}$ be an estimate of ψ based on a sample of size n such that*

$$\sqrt{n}(\widehat{\psi}-\psi) \to N(0,\sigma^2).$$

Then, for any function $h(\cdot)$ that is differentiable around ψ and $h'(\psi) \neq 0$, we have

$$\sqrt{n}(h(\widehat{\psi}) - h(\psi)) \to N(0, \sigma^2|h'(\psi)|^2).$$

In view of the Central Limit Theorem, the Delta method applies to functions of the sample mean. Informally we say that $h(\widehat{\psi})$ is approximately normal with mean $h(\psi)$ and variance $|h'(\psi)|^2\text{var}(\widehat{\psi})$.

To apply the Delta method here, recall that $\widehat{\pi}$ is approximately $\text{N}(\pi, \sigma^2/378)$, where

$$\sigma^2 = \pi(1-\pi).$$

So, from $h(\pi) = \pi/(1-\pi)$, we have

$$\begin{aligned} h'(\pi) &= \frac{1}{(1-\pi)^2} \\ \text{var}(\widehat{\theta}) &= \frac{1}{(1-\pi)^4}\frac{\pi(1-\pi)}{378} \\ &= \frac{\pi}{378(1-\pi)^3} \\ &\approx \frac{139/378}{378(1-139/378)^3} = 0.3849 \times 10^{-2}. \end{aligned}$$

So, the approximate 95% confidence interval for θ, given by $\widehat{\theta} \pm 1.96\sqrt{\text{var}(\widehat{\theta})}$, is

$$0.46 < \theta < 0.70,$$

shifted slightly to the left of the likelihood-based interval. (Of course in practice the difference is not important, what matters for discussion here is the manner to arrive the intervals.)

We may interpret the Delta Method this way: we obtain a single observation $\widehat{\theta} = 0.58$ from $\mathrm{N}(\theta, 0.3849 \times 10^{-2})$. This produces an approximate likelihood of θ, differing slightly from the likelihood we derived before; see Figure 4.7. Hence the Delta method can be seen to produce a quadratic approximation of the likelihood function.

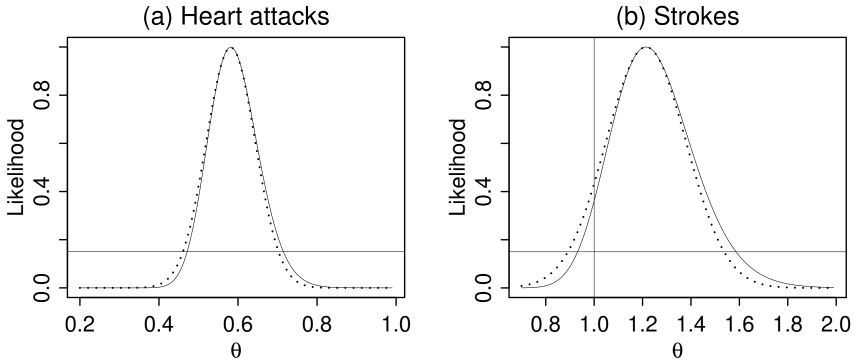

Figure 4.7: *(a) Profile likelihood of θ for the number of heart attacks in the aspirin data example and its normal approximation (dotted line). (b) The same as (a) for strokes.*

4.9 Continuous data

Normal models

We have discussed the normal model $N(\mu, \sigma^2)$ in Section 2.5 as the ideal or exactly regular case for likelihood inference. It is one of the most commonly used models for analysing continuous outcomes. Many results in classical statistics are derived for normal data. Let $x_1, \ldots, x_n$ are an iid sample from $N(\mu, \sigma^2)$ with both parameters unknown. Ignoring irrelevant constant terms, we can write

$$\log L(\mu, \sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_i (x_i - \mu)^2,$$

so maximizing the likelihood function for μ is equivalent to finding the *least-squares* estimate. The maximum likelihood estimate is given by

$$\begin{aligned}\widehat{\mu} &= \overline{x}\\ \widehat{\sigma}^2 &= \frac{1}{n}\sum_i (x_i - \overline{x})^2.\end{aligned}$$

Note the $1/n$ divisor for the variance estimate, which is different from the usual $1/(n-1)$ divisor for the sample variance S^2.

Example 4.7: The following are 100 measurements of the speed of light (in km/sec, minus 299,000) made by Michelson between June 5 and July 2, 1879 (Stigler, 1977). The data were based on 5 experiments, each with 20 runs. The first two lines of data (read by row) are from the first experiment, etc.

850	740	900	1070	930	850	950	980	980	880
1000	980	930	650	760	810	1000	1000	960	960
960	940	960	940	880	800	850	880	900	840
830	790	810	880	880	830	800	790	760	800
880	880	880	860	720	720	620	860	970	950
880	910	850	870	840	840	850	840	840	840
890	810	810	820	800	770	760	740	750	760
910	920	890	860	880	720	840	850	850	780
890	840	780	810	760	810	790	810	820	850
870	870	810	740	810	940	950	800	810	870

The MLEs of the normal parameters are $\widehat{\mu} = 852.4$, and $\widehat{\sigma}^2 = 6180.24$. The sample variance is $S^2 = 6242.67$; it is common practice to use $S/\sqrt{n}$ as the standard error for $\widehat{\mu}$. The 95% confidence interval for μ is

$$\widehat{\mu} \pm 1.96\ S/\sqrt{n},$$

yielding $803.4 < \mu < 901.4$; this does not cover the currently accepted $\widehat{\mu} = 734.5$. There are several problems: The means of the 5 experiments are 909, 856, 845, 820.5, 831.5, so there seems to be a problem with the first experiment (Exercise 4.22). The average from the last 4 experiments is 838.3 (se = 7.2). Also, there is a correlation of 0.54 between consecutive measurements, so they are not independent and the standard error is wrong.

A normal assumption can be checked using the QQ-plot; it is a general method to see the shape and texture of a distribution. If X has a continuous distribution $F(x)$, then

$$P\{F(X) \leq u\} = P\{X \leq F^{-1}(u)\} = F(F^{-1}(u)) = u.$$

This means $F(X)$ is a standard uniform variate. Conversely, if U is standard uniform, then $F^{-1}(U)$ is a random variable with distribution function $F(x)$. If $X_1, \ldots, X_n$ are an iid sample from $F(x)$, then $F(X_1), \ldots, F(X_n)$ are an iid sample from $U(0,1)$.

Suppose $x_1, \ldots, x_n$ have a hypothesized distribution F. Intuitively the ordered values $x_{(1)}, \ldots, x_{(n)}$ should behave like order statistics $s_1, \ldots, s_n$ simulated from F. We can plot s_i against $x_{(i)}$; if the data do come from F, we expect to see a straight line. The randomness of $s_1, \ldots, s_n$ is a nuisance; we might use the expected order statistic $EX_{(i)}$ instead. Denoting by $U_{(i)}$ the i'th order statistic from $U_1, \ldots, U_n$, we have

$$\begin{aligned} \text{median}\{X_{(i)}\} &= \text{median}\{F^{-1}(U_{(i)})\} \\ &= F^{-1}(\text{median}\{U_{(i)}\}) \\ &\approx F^{-1}\left(\frac{i - 1/3}{n + 1/3}\right). \end{aligned}$$

The approximate median of $U_{(i)}$ is accurate even for very small n (Hoaglin 1985); for example, for $n = 5$ the median of $U(1)$ is 0.129, and the approximation is (1-1/3)/(5+1/3)= 0.125. The QQ-plot is a plot of

$$F^{-1}\left(\frac{i-1/3}{n+1/3}\right) \text{ versus } x_{(i)}.$$

As a visual aid, we can draw a line going through the first and third quartiles of the distribution.

If $X = \sigma X_0 + \mu$ for some standard variate X_0, and some unknown μ and σ, we do not need to estimate μ and σ. We can use the distribution of X_0 as the basis of QQ-plot. Here we expect

$$EX_{(i)} \approx \sigma F_0^{-1}\left(\frac{i}{n+1}\right) + \mu$$

where F_0 is the distribution of X_0. For example, to check normality we use the standard normal distribution function $\Phi(x)$ in place of $F(x)$.

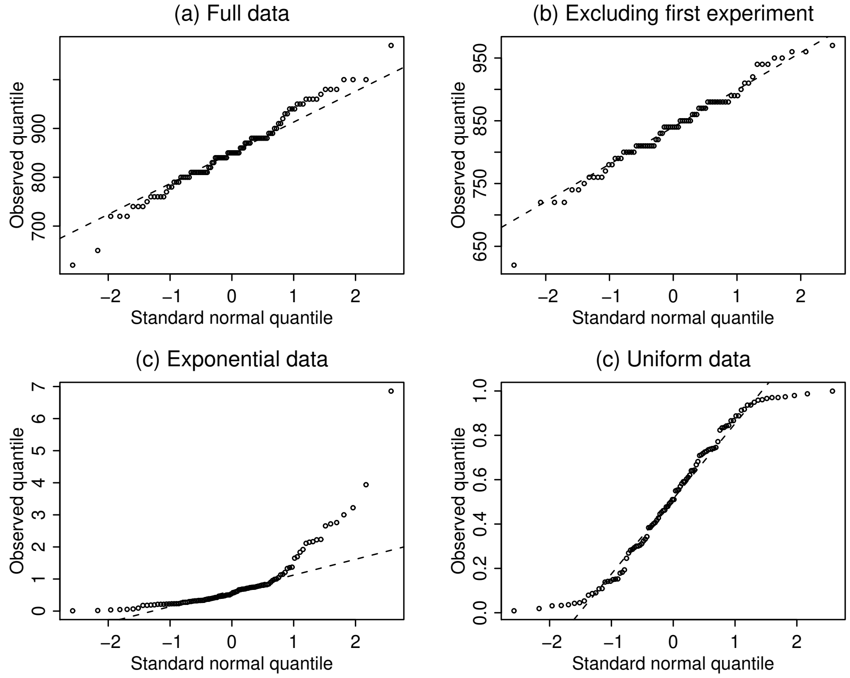

Figure 4.8: *(a) The normal QQ-plot indicates that Michelson's data are not normal; (b) Excluding data from the first experiment, the rest of the data appears normal; (c) Simulated exponential data, showing skewed distribution; (d) Simulated uniform data, showing short-tailed distribution.*

The QQ-plot of Michelson's data in Figure 4.8(a) indicates some nonnormality: notice the deviation from the guideline, especially of the upper half. This occurs because the first experiment is different from the rest. As shown in Figure 4.8(b), excluding data from the first experiment, the remaining data look normal, except for possibly a single outlier.

Even for nonnormal data, the normal QQ-plot is a useful exploratory tool. The plot shows skewness, heavy-tailed or short-tailed behaviour, digit preference, or outliers and other unusual values. Figure 4.8(c) shows the QQ-plot of simulated exponential data (skewed to the right) and uniform data (short tail).

The case $n = 1$

In the extreme case $n = 1$ the likelihood function becomes unbounded at $\mu = x_1$, so $\widehat{\mu} = x_1$, $\widehat{\sigma}^2 = 0$, and the estimated distribution is degenerate at x_1. The maximized likelihood is $L(\widehat{\mu}, \widehat{\sigma}^2) = \infty$. This has been cited as a weakness of the likelihood method: the likelihood suggests there is a very strong evidence that $\widehat{\sigma}^2 = 0$, which is 'surprising' as we feel a single observation cannot tell us anything about variability.

The infinite likelihood can be avoided by using a finite precision model and the original definition of likelihood as probability. (Although we will end up with the same result that the best model is 'degenerate' at x_1, where 'x_1' represents an interval $(x - \epsilon, x + \epsilon)$.) However, this example is a valid caveat about over-interpretation of the (observed maximized) likelihood. Unlike the likelihood, our sense of surprise is affected not just by the data, but also by our prior expectation (e.g., the data are continuous). The likelihood result will not feel as surprising if we know beforehand that a discrete point model $\mu = x_1$ and $\sigma^2 = 0$ is a real possibility.

Two-sample case

Suppose $x_1, \ldots, x_m$ are an iid sample from $N(\mu, \sigma_x^2)$, and $y_1, \ldots, y_n$ from $N(\mu+\delta, \sigma_y^2)$. If we are only interested in the mean difference δ, there are three nuisance parameters. Computation of the profile likelihood of δ is left as an exercise.

In practice it is common to assume that the two variances are equal. Under this assumption, the classical analysis is very convenient; the t-statistic

$$t = \frac{\overline{y} - \overline{x} - \delta}{S_p\sqrt{\frac{1}{m} + \frac{1}{m}}}$$

where S_p^2 is the pooled variance

$$S_p^2 = \frac{(m-1)S_x^2 + (n-1)S_y^2}{m+n-2},$$

has t distribution with $m + n - 2$ degrees of freedom. We can use this for testing hypotheses or setting confidence intervals on δ.

For Michelson's data, suppose we want to compare the means of the first experiment with the remaining data. We obtain

$$\begin{aligned} m = 20, \quad \overline{x} = 909, \quad S_x^2 = 11009.47 \\ n = 80, \quad \overline{y} = 840.5, \quad S_y^2 = 4604.8 \end{aligned}$$

and $t = -3.58$ to test $\delta = 0$, so there is a strong evidence that the first experiment differs from the subsequent ones.

Nonnormal models

As an example of a nonnormal continuous model, the gamma model is useful for positive outcome data such as measures of concentration, weight, lifetime, etc. The density of the gamma(α, λ) model is

$$f(x) = \frac{1}{\Gamma(\alpha)} \lambda^\alpha x^{\alpha-1} e^{-\lambda x}, \quad x > 0.$$

The parameter α is called the *shape* parameter and λ^{-1} is the *scale* parameter. One important property of the gamma model is that the mean $E(X) = \alpha/\lambda$ is quadratically related to the variance $\text{var}(X) = \alpha/\lambda^2 = E^2(X)/\alpha$. It is sometimes useful to reparameterize using $\mu = \alpha/\lambda$ and $\phi = 1/\alpha$, so that we have Gamma(μ, ϕ) with mean $E(X) = \mu$, variance $\phi\mu^2$ and coefficient of variation $\sqrt{\phi}$. The density can be written as

$$f(x) = \frac{1}{x\Gamma(1/\phi)} \left(\frac{x}{\phi\mu}\right)^{1/\phi} \exp(-\frac{x}{\phi\mu}), \quad x > 0.$$

If the shape parameter $\alpha = 1$ we obtain the exponential model with density

$$f(x) = \mu^{-1} e^{-x/\mu}.$$

Example 4.8: The following data are the duration of service (in minutes) for 15 bank customers. Of interest is the average length of service μ.

```
23.91 27.33  0.15  3.65  5.99  0.88  0.93  0.53  0.17 14.17
 6.18  0.05  3.89  0.24  0.08.
```

Figure 4.9 shows the joint likelihood of μ and ϕ; the contour lines are drawn at 90%, 70%, 50%, 30% and 10% cutoffs. The profile likelihood for μ can be computed numerically from the joint likelihood. Also shown are the gamma likelihood assuming ϕ is known at $\widehat{\phi} = 2.35$ and the exponential likelihood by assuming $\phi = 1$. From Figure 4.9 (a) it is clear that the exponential model ($\phi = 1$) is not well supported by data.

4.10 Exponential family

The exponential family is a very wide class of models that includes most of the commonly used models in practice. It is an unfortunate family name, since it is mixed up with the standard exponential model, which is a member of the exponential family. However, we will stick with the terminology as no one has come up with a better name, and usually what we mean is clear from the context. A general p-parameter exponential family has $\theta \equiv (\theta_1, \ldots, \theta_p)$ and its log-density is of the form

$$\log p_\theta(x) = \sum_{i=1}^{p} \eta_i(\theta) T_i(x) - A(\theta) + h(x) \tag{4.7}$$

for known functions $A(\theta)$ and $h(x)$, and $\eta_i(\theta)$ and $T_i(x)$ for each i. Furthermore the support of the distribution must not involve the unknown parameter θ.

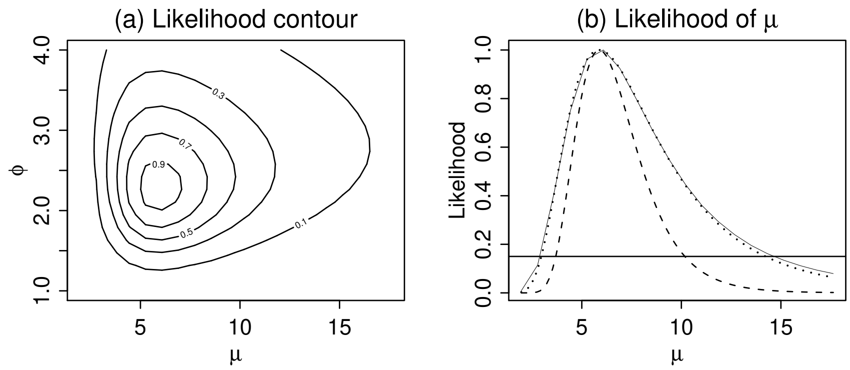

Figure 4.9: *(a) Joint likelihood of μ and ϕ using the gamma model; (b) The profile likelihood of μ from the gamma model (solid line), the estimated likelihood using the gamma model assuming $\phi = \widehat{\phi} = 2.35$ (dotted line) and the likelihood using an exponential model (dashed line), which is equivalent to the gamma model with $\phi = 1$.*

The parameters η_i's are called the natural parameters of the family, and T_i's are the natural statistics. T_i's can be shown to be minimal sufficient (Exercise 4.25). The family is called *regular* if the natural parameter space contains an open set (e.g., a 2D-square in 2D-space contains an open set, but a curve in 2D-space does not); this is typically the case when the number of unknown parameters is the same as the number of natural sufficient statistics.

If the number of natural sufficient statistics is greater than the number of unknown parameters, the family is called a *curved* family. The distinction is important, since many theoretical results are true only for regular families. Compared to those of a regular family, problems involving curved exponential family are somewhat 'more nonlinear.'

The exponential family includes both discrete and continuous random variables. The normal, binomial, Poisson or gamma models are in the exponential family, but the Cauchy and t-distributions are not.

Example 4.9: For the normal model with $\theta = (\mu, \sigma^2)$, with both parameters unknown,

$$\log p_\theta(x) = \frac{\mu x}{\sigma^2} - \frac{x^2}{2\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)$$

is a 2-parameter regular exponential family model with natural parameters $\eta_1 = \mu/\sigma^2$ and $\eta_2 = -1/(2\sigma^2)$, and natural statistics $T_1(x) = x$ and $T_2(x) = x^2$.

If x has a known coefficient of variation c, such that it is $N(\mu, c^2\mu^2)$, with unknown $\mu > 0$, then x is a curved exponential family. Even though there is only one unknown parameter, the natural statistics are still $T_1(x) = x$ and $T_2(x) = x^2$. These statistics are minimal sufficient.

Example 4.10: For the Poisson model with mean μ we have

$$\log p_\mu(x) = x \log \mu - \mu - \log x!,$$

therefore it is a regular 1-parameter exponential family.

If x is a truncated Poisson in the sense that $x = 0$ is not observed, we have, for $x = 1, 2, \ldots,$

$$P(X = x) = \frac{e^{-\mu}\mu^x/x!}{1 - e^{-\mu}},$$

so

$$\log P(X = x) = x \log \mu - \mu - \log(1 - e^{-\mu}) - \log x!.$$

This also defines a regular 1-parameter exponential family with the same canonical parameter and statistic, but a slightly different $A(\mu)$ function.

The joint distribution of an iid sample from an exponential family is also in the exponential family. For example, let $x_1, \ldots, x_n$ be an iid sample from $N(\mu, \sigma^2)$. The joint density is

$$\log p_\theta(x_1, \ldots, x_n) = \frac{\mu \sum_i x_i}{\sigma^2} - \frac{\sum_i x_i^2}{2\sigma^2} - \frac{n\mu^2}{2\sigma^2} - \frac{n}{2}\log(2\pi\sigma^2).$$

This has the same natural parameters $\eta_1 = \mu/\sigma^2$ and $\eta_2 = -1/(2\sigma^2)$, and natural statistics $T_1 = \sum_i x_i$ and $T_2 = \sum_i x_i^2$.

To illustrate the richness of the exponential family and appreciate the special role of the function $A(\theta)$, suppose X is any random variable with density $\exp\{h(x)\}$ and moment generating function

$$m(\theta) = Ee^{\theta X}.$$

Let $A(\theta) \equiv \log m(\theta)$, usually called the cumulant generating function. Then

$$\int e^{\theta x + h(x)} dx = e^{A(\theta)}$$

or

$$\int e^{\theta x - A(\theta) + h(x)} dx = 1$$

for all θ. So

$$p_\theta(x) \equiv e^{\theta x - A(\theta) + h(x)} \tag{4.8}$$

defines an exponential family with parameter θ. Such a construction is called an 'exponential tilting' of the random variable X; the original X corresponds to $\theta = 0$.

For any exponential model of the form (4.8) we can show (Exercise 4.27) that

$$\mu = E_\theta(X) = A'(\theta)$$

and

$$\text{var}_\theta(X) \;=\; A''(\theta) = \frac{\partial}{\partial\theta}E_\theta(X) = V(\mu).$$

Therefore the choice of $A(\theta)$ implies a certain relationship between the mean and variance.

Exponential dispersion model

For statistical modelling it is often adequate to consider a two-parameter model known as the exponential dispersion model (Jorgensen 1987). Based on an observation x the log-likelihood of the scalar parameters θ and ϕ is of the form

$$\log L(\theta, \phi) = \frac{x\theta - A(\theta)}{\phi} + h(x, \phi), \tag{4.9}$$

where $A(\theta)$ and $h(x, \phi)$ are assumed known functions. In this form the parameter θ is called the canonical parameter and ϕ the *dispersion parameter.* Since $h(x, \phi)$ or $A(\theta)$ can be anything there are infinitely many submodels in this exponential family, though of course the density must satisfy

$$\sum_x \exp\left\{\frac{x\theta - A(\theta)}{\phi} + h(x, \phi)\right\} = 1,$$

which forces a certain relationship between $A(\theta)$ and $h(x, \phi)$.

The dispersion parameter allows the variance to vary freely from the mean:

$$\mu = E_\theta(X) = A'(\theta)$$

and

$$\begin{aligned} \mathrm{var}_\theta(X) &= \phi A''(\theta) \\ &= \phi \frac{\partial}{\partial\theta} E_\theta(X) = \phi V(\mu). \end{aligned}$$

This is the biggest advantage of the exponential dispersion model over the more rigid model (4.8).

In practice the form of $A(\theta)$ in (4.9) is usually explicitly given from the standard models, while $h(x, \phi)$ is left implicit. This is not a problem as far as estimation of θ is concerned, since the score equation does not involve $h(x, \phi)$. However, without explicit $h(x, \phi)$, a likelihood-based estimation of ϕ and a full likelihood inference on both parameters are not available. One might compromise by using the method of moments to estimate ϕ, or the approximation of the likelihood.

Example 4.11: For the normal model $N(\mu, \sigma^2)$ we have

$$\log L(\mu, \sigma^2) = \frac{x\mu - \mu^2/2}{\sigma^2} - \frac{1}{2}\log\sigma^2 - \frac{1}{2}x^2/\sigma^2,$$

so the normal model is in the exponential family with a canonical parameter $\theta = \mu$, dispersion parameter $\phi = \sigma^2$, and $A(\theta) = \theta^2/2$ and $h(x, \phi) = -\frac{1}{2}\log\phi - \frac{1}{2}x^2/\phi$. This is a rare case where $h(x, \phi)$ is available explicitly.

Example 4.12: For the Poisson model with mean μ

$$\log L(\mu) = x\log\mu - \mu - \log x!$$

so we have a canonical parameter $\theta = \log\mu$, dispersion parameter $\phi = 1$ and $A(\theta) = \mu = e^\theta$.

Keeping $A(\theta) = e^\theta$, but varying the dispersion parameter would generate a Poisson model with under- or overdispersion: with the same $EX = \mu$ we have

$$\mathrm{var}(X) = \phi A''(\theta) = \phi\mu.$$

The function $h(x, \phi)$ has to satisfy

$$\sum_{x=0}^{\infty} \exp\left[\left\{\frac{x\theta - e^\theta}{\phi}\right\} + h(x,\phi)\right] = 1,$$

for all θ and ϕ. To compute the likelihood we need $h(x, \phi)$ explicitly, but finding the solution $h(x, \phi)$ that satisfies such an equation is not trivial.

Example 4.13: The following data shows the number of paint defects recorded in a sample of 20 cars using an experimental process:

```
0 10  1  1  1  2  1  4 11  0  5  2  5  2  0  2  0  1  3  0
```

The sample mean is $\overline{x} = 2.55$ and sample variance is $s^2 = 9.84$, indicating overdispersion. Using a Poisson-type model above $A(\theta) = e^\theta$, so

$$\mathrm{var}_\theta(X) = \phi E_\theta(X),$$

and the method of moments estimate of ϕ is $\widehat{\phi} = 9.84/2.55 = 3.86$. Obviously there is some variability in $\widehat{\phi}$; is it significantly away from one? A likelihood analysis is difficult, instead we can do a Monte Carlo experiment to test $\phi = 1$ (Exercise 4.30):

1. generate $x_1^*, \ldots, x_{20}^*$ as an iid sample from the Poisson with mean $\mu = \overline{x} = 2.55$
2. compute $\widehat{\phi}^* = (s^2)^*/\overline{x}^*$ from data in part 1
3. repeat (1) and (2) a large number of times and consider the $\widehat{\phi}^*$'s as a sample from the distribution of $\widehat{\phi}$.
4. Compute P-value= proportion of $\widehat{\phi}^* >$ observed $\widehat{\phi}$.

Figure 4.10 shows the QQ plot of 500 $\widehat{\phi}^*$'s and we can see that P-value ≈ 0, confirming that $\phi > 1$.

An approximate likelihood inference for μ can proceed based on assuming that ϕ is known at $\widehat{\phi}$; this is made particularly simple since the term $h(x_i, \phi)$ is constant relative to μ or θ. Figure 4.10 compares the standard Poisson likelihood with the estimated likelihood using $\widehat{\phi} = 3.86$. For small samples this approach is not satisfying as the uncertainty in $\widehat{\phi}$ is not accounted for in the inference for μ.

Approximate likelihood*

While an exact likelihood of the exponential dispersion model (4.9) might not be available, there is a general approximation that can be computed easily. The main advantage of the approximate likelihood is that it puts the whole estimation within the likelihood context. It also facilitates further modelling of the dispersion parameter.

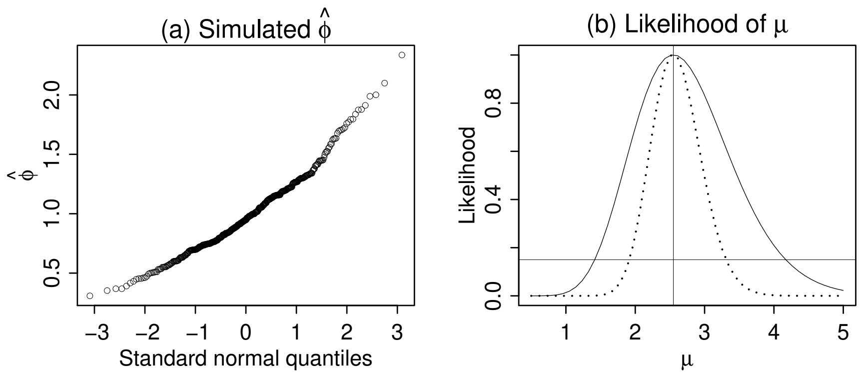

Figure 4.10: *(a) Monte Carlo estimate of the distribution of $\widehat{\phi}$ under standard Poisson model. (b) Likelihood of μ based on the standard Poisson model ($\phi = 1$, dotted line) and the model with extra Poisson variability ($\phi = \widehat{\phi} = 3.86$, solid line).*

At fixed ϕ the MLE of θ is the solution of

$$A'(\widehat{\theta}) = x.$$

Alternatively $\widehat{\mu} = x$ is the MLE of $\mu = A'(\theta)$. The Fisher information on θ based on x is

$$I(\widehat{\theta}) = A''(\widehat{\theta})/\phi.$$

From Section 9.8, at fixed ϕ, the approximate density of $\widehat{\theta}$ is

$$p(\widehat{\theta}) \approx (2\pi)^{-1/2} I(\widehat{\theta})^{1/2} \frac{L(\theta,\phi)}{L(\widehat{\theta},\phi)}.$$

Since $\widehat{\mu} = A'(\widehat{\theta})$, the density of $\widehat{\mu}$, hence of x, is

$$\begin{aligned} p(x) = p(\widehat{\mu}) &= p(\widehat{\theta}) \left| \frac{\partial \widehat{\theta}}{\partial \widehat{\mu}} \right| \\ &= p(\widehat{\theta}) \{A''(\widehat{\theta})\}^{-1} \\ &\approx \{2\pi\phi A''(\widehat{\theta})\}^{-1/2} \frac{L(\theta,\phi)}{L(\widehat{\theta},\phi)}. \end{aligned}$$

The potentially difficult function $h(x,\phi)$ in (4.9) cancels out in the likelihood ratio term, so we end up with something simpler. In commonly used distributions, the approximation is highly accurate for a wide range of parameter values.

Define the deviance function

$$\begin{aligned} D(x,\mu) &= 2 \log \frac{L(\widehat{\theta}, \phi = 1)}{L(\theta, \phi = 1)} \\ &= 2\{\widehat{\theta}x - \theta x - A(\widehat{\theta}) + A(\theta)\}, \end{aligned}$$

where $\mu = A'(\theta)$. The approximate log-likelihood contribution from a single observation x is

$$\log L(\theta,\phi) \approx -\frac{1}{2}\log\{2\pi\phi V(x)\} - \frac{1}{2\phi} D(x,\mu), \tag{4.10}$$

where $V(x) = A''(\widehat{\theta})$.

The formula is exact if x is $N(\mu, \sigma^2)$, in which case $\theta = \mu$, $\phi = \sigma^2$, $V(x) = 1$, and $D(x, \mu) = (x - \mu)^2$. In general it is called an extended quasi-likelihood formula by Nelder and Pregibon (1987). It is very closely related to the double exponential family (Efron 1986). The idea is that , given θ, the dispersion parameter also follows an exponential family likelihood.

Given an iid sample $x_1, \ldots, x_n$, at fixed ϕ, the estimate of μ is the minimizer of the total deviance

$$\sum_i D(x_i, \mu) = 0.$$

The approximate profile log-likelihood of ϕ is

$$\log L(\phi) \approx -\frac{1}{2}\log\{2\pi\phi V(x)\} - \frac{1}{2\phi}D(x, \widehat{\mu})$$

and the approximate MLE of ϕ is the average deviance

$$\widehat{\phi} = \frac{1}{n}\sum_i D(x_i, \widehat{\mu}).$$

Example 4.14: In Example 4.12 we cannot provide an explicit two-parameter Poisson log-likelihood:

$$\log L(\mu, \phi) = \frac{x\log\mu - \mu}{\phi} + h(x, \phi),$$

since $h(x, \phi)$ is not available. Using

$$\begin{aligned} \theta = \log\mu, \quad A(\theta) = e^{\theta} = \mu \\ \widehat{\theta} = \log x, \quad A(\widehat{\theta}) = x \\ D(x, \mu) = 2(x\log x - x\log\mu - x + \mu) \\ V(x) = A''(\widehat{\theta}) = e^{\widehat{\theta}} = x \end{aligned}$$

we get an approximate likelihood

$$\log L(\mu, \phi) \approx -\frac{1}{2}\log\{2\pi\phi x\} - \frac{1}{\phi}(x\log x - x\log\mu - x + \mu).$$

Nelder and Pegribon (1987) suggest replacing $\log\{2\pi\phi x\}$ by

$$\log\{2\pi\phi(x + 1/6)\}$$

to make the formula works for $x \geq 0$.

For the data in Example 4.13 the estimate of the mean μ is $\overline{x} = 2.55$, and

$$\widehat{\phi} = \frac{1}{20}\sum_i D(x_i, \overline{x}) = 3.21,$$

where we have used $0\log 0 = 0$; recall $\widehat{\phi} = 3.86$ using the method of moments. The approximate profile likelihoof of ϕ can also be computed from (4.10).

Minimal sufficiency and the exponential family*

There is a surprising connection between minimal sufficiency and the exponential family: Under mild conditions, a sufficient estimate exists, if and only if the model $p_\theta(x)$ is in the exponential family. The exponential family structure in turn implies that the likelihood function has a unique maximum and the MLE is sufficient. *Therefore, if a sufficient estimate exists, it is provided by the MLE.*

It is easy to show that the MLE of θ in a regular exponential family model is sufficient (Exercise 4.25). To show that the existence of a sufficient estimate implies the exponential family, we follow the development in Kendall *et al.* (1977) for the scalar case. Assume that we have an iid sample $x_1, \ldots, x_n$ from $p_\theta(x)$, $t(x)$ is a sufficient estimate (i.e., it is minimally sufficient for θ), and that the support of $p_\theta(x)$ does not depend on θ. By the factorization theorem,

$$\frac{\partial \log L(\theta)}{\partial \theta} = \sum_i \frac{\partial \log p_\theta(x_i)}{\partial \theta} = K(t, \theta)$$

for some $K(\cdot)$. Since this is true for any θ, choose one value of θ, so t must be of the form

$$t = M\{\sum_i k(x_i)\}$$

for some function $M(\cdot)$ and $k(\cdot)$. Defining $w(x) = \sum_i k(x_i)$, then $K(t, \theta)$ is a function of w and θ only, say $N(w, \theta)$. Now

$$\frac{\partial^2 \log L(\theta)}{\partial x_i \partial \theta} = \frac{\partial N}{\partial w} \frac{\partial w}{\partial x_i}.$$

Since $\partial^2 \log L(\theta)/\partial x_i \partial \theta$ and $\partial w/\partial x_i$ only depend on x_i and θ, it is clear that $\partial N/\partial w$ is also a function of θ and x_i. To be true for all x_i then $\partial N/\partial w$ must be a function of θ alone, which means

$$N(w, \theta) = w(x)u(\theta) + v(\theta)$$

for some function $u(\theta)$ and $v(\theta)$, or, in terms of the log-likelihood

$$\frac{\partial}{\partial \theta} \log L(\theta) = u(\theta) \sum_i k(x_i) + v(\theta).$$

This implies the model must be in the exponential family with log-density of the form

$$\log p_\theta(x) = \eta(\theta)T(x) - A(\theta) + h(x).$$

Now it is straightforward to show that the likelihood is unimodal, so it has a unique maximum.

4.11 Box-Cox transformation family

Box-Cox (1964) transformation family is another useful class of nonnormal models for positive-valued outcomes. Typical characteristics of such outcomes are skewness, and some relationship between the mean and variance. In Box-Cox transformation model it is assumed that there is $\lambda \neq 0$ such that a transformation of the observed data y according to

$$y_\lambda = \frac{y^\lambda - 1}{\lambda}$$

has a normal model $N(\mu, \sigma^2)$. As λ goes to zero $y_\lambda \to \log y$, so this family includes the important log transformation. For convenience we will simply write $\lambda = 0$ to represent the log-transformation. The log-likelihood contribution of a single observation y is

$$\begin{aligned} \log L(\lambda, \mu, \sigma^2) &= \log p(y_\lambda) + (\lambda - 1) \log y \\ &= -\frac{1}{2} \log \sigma^2 - \frac{(y_\lambda - \mu)^2}{2\sigma^2} + (\lambda - 1) \log y, \end{aligned}$$

where $p(y_\lambda)$ is the density of y_λ and $(\lambda - 1)\log y$ is the log of the Jacobian.

Note that if y is positive then y_λ cannot be strictly normal for any $\lambda \neq 0$. The possibility of a truncated normal may be considered, and we should check the normal plot of the transformed data. If $y_\lambda > 0$ has a truncated normal distribution, it has a density

$$p(y_\lambda) = \frac{\sigma^{-1}\phi\{(y_\lambda - \mu)/\sigma\}}{1 - \Phi(-\mu/\sigma)}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal density and distribution functions. So we only need to modify the log-likelihood above by adding $\log\{1 - \Phi(-\mu/\sigma)\}$. If μ/σ is quite large, $\Phi(-\mu/\sigma) \approx 0$, so we can simply ignore the effect of truncation.

Interpretation of the parameters is an important practical issue when using a transformation. After transformation μ and σ^2 may not have a simple interpretation in terms of the original data. However, since the Box-Cox transformation is monotone, *the median of y_λ is the transformed median of the original data y.* For example, if $\lambda = 0$ and the median of y_λ is $\mu = 0$, then the median of the original data y is $e^\mu = e^0 = 1$. It is sensible to limit the choice of λ to few meaningful values.

The main reason to consider the Box-Cox transformation family is to analyze the data in a more normal scale. Such a scale typically leads to models that are simpler and easier to interpret, have better-behaved residuals, and clearer inference.

To decide what transformation parameter λ to use in a dataset $y_1, \ldots, y_n$, it is convenient to compare a range of values or simply several sensible values such as $\lambda = -1,\ 0,\ 0.5,\ 1$ and 2. In practice, it will not be very sensible to transform data using, say, $\lambda = 0.5774$. The natural criterion to compare is the profile likelihood of λ, which is very simple to compute. At fixed λ we simply transform the data to get y_λ, and the MLEs of μ and σ^2 are simply the sample mean and variance of the transformed data:

$$\widehat{\mu}(\lambda) = \frac{1}{n} \sum_i y_{\lambda i}$$

and

$$\widehat{\sigma}^2(\lambda) = \frac{1}{n} \sum_i \{y_{\lambda i} - \widehat{\mu}(\lambda)\}^2.$$

The profile log-likelihood for λ is

$$\log L(\lambda) = -\frac{n}{2}\log\widehat{\sigma}^2(\lambda) - \frac{n}{2} + (\lambda - 1)\sum_i \log y_i.$$

There needs to be some care in the inference on μ based on the transformed data $y_{\widehat{\lambda}}$. Mathematically we expect to 'pay' for having to estimate λ (Bickel and Doksum, 1981), for example, by taking a profile likelihood over λ. However, such an approach can be meaningless. A serious problem in the use of transformation is the logical meaning of y_λ and the parameter μ. Here is a variation of a simple example from Box and Cox (1982): Suppose we have a well-behaved sample from a normal population with mean around 1000 and a small variance. Then there is an extremely wide range of λ for which the transformation is essentially linear and the transformed data are close to normal. This means λ is poorly determined from the data, and its profile likelihood is flat as it has a large uncertainty. Propagating this large uncertainty to the estimation of μ produces a very wide confidence interval for μ. That is misleading, since in fact the untransformed data themselves ($\lambda = 1$) provide the best information.

So, a rather flat likelihood of λ indicates that the Box-Cox transformation family is not appropriate; it warrants a closer look at the data. If we limit the use of the Box-Cox transformation model to cases where λ is well determined from the data, then it is reasonable to perform inference on μ assuming that λ is known at $\widehat{\lambda}$, i.e., we do not have to pay for having to estimate λ. This is the commonly adopted approach in practice.

Example 4.15: The following figures are the population of 27 countries already in or wishing to join the European Union:

```
82 59 59 57 39 38 22 16 10 10 10 10
10 9 9 8 5 5 5 4 4 2 2 1 0.7 0.4 0.4
```

Figure 4.11(a) shows the population data are skewed to the right. The profile log-likelihood of λ is maximized at $\widehat{\lambda} = 0.08$, pointing to the log transform $\lambda = 0$ as a sensible transformation. The QQ-plots of the square-root and the log-transform indicate we should prefer the latter. (The reader can verify that $\log(y + 1)$ is a better normalizing transform). Plotting the data in the log scale allows us to resolve the variability among low as well as high population counts.

4.12 Location-scale family

The location-scale family is a family of distributions parameterized by μ and σ, and a known density $f_0(\cdot)$, such that any member of the family has a density of the form

$$f(x) = \frac{1}{\sigma} f_0\left(\frac{x-\mu}{\sigma}\right).$$

The parameter μ is called the location parameter and σ the scale parameter, and $f_0(\cdot)$ is called the *standard density* of the family. For example the $N(\mu, \sigma^2)$ is a

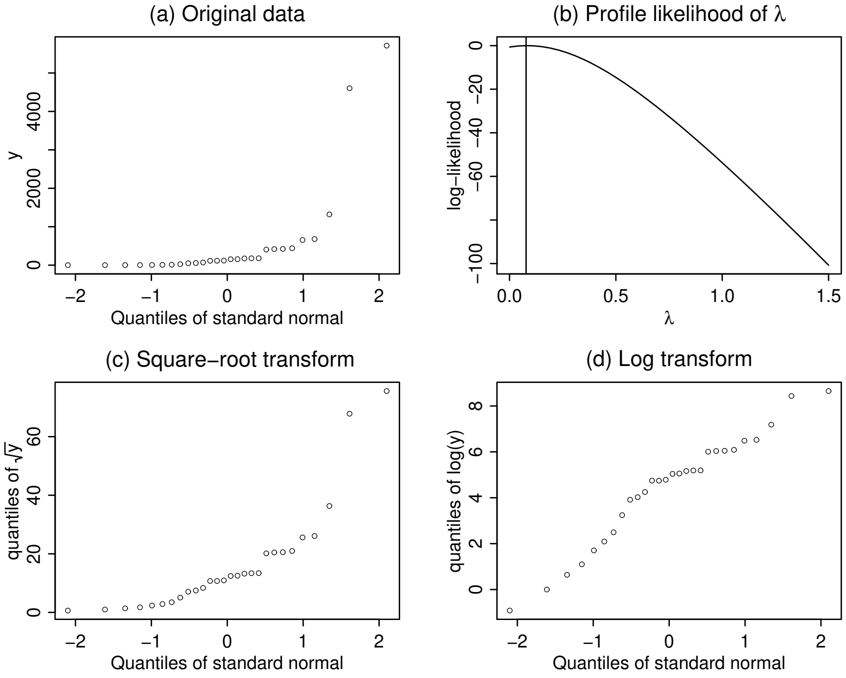

Figure 4.11: *(a) Normal plot of the population data; (b) the profile likelihood function of the transformation parameter λ; (c) and (d) QQplots of the log and square-root transform of the original data.*

location-scale family with standard density

$$f_0(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Another famous example is the Cauchy(μ, σ) with standard density

$$f_0(x) = \frac{1}{\pi(1+x^2)}.$$

The Cauchy model is useful as a model for data with heavy tail, characterized by the presence of outliers. Furthermore, in theoretical studies the Cauchy is a good representative of complex models.

Example 4.16: Recording of the difference in maximal sun radiation between two geographical regions over a time period produced the following (sorted) data

```
-26.8  -3.5  -3.4  -1.2   0.4   1.3   2.3   2.7   3.0   3.2   3.2
  3.5   3.6   3.9   4.2   4.4   5.0   6.5   6.7   7.1   8.1  10.5
 10.7  24.0  32.8
```

The normal plot in Figure 4.12(a) shows clear outliers or a heavy tail behaviour. The mean and median of the data are 3.6 and 4.5 respectively; they are not dramatically different since there are outliers on both ends of the distribution. The Cauchy model with location μ and scale σ has a likelihood function

$$L(\mu, \sigma) = \prod_i \frac{1}{\sigma} \left\{ 1 + \frac{(x_i - \mu)^2}{\sigma^2} \right\}^{-1},$$

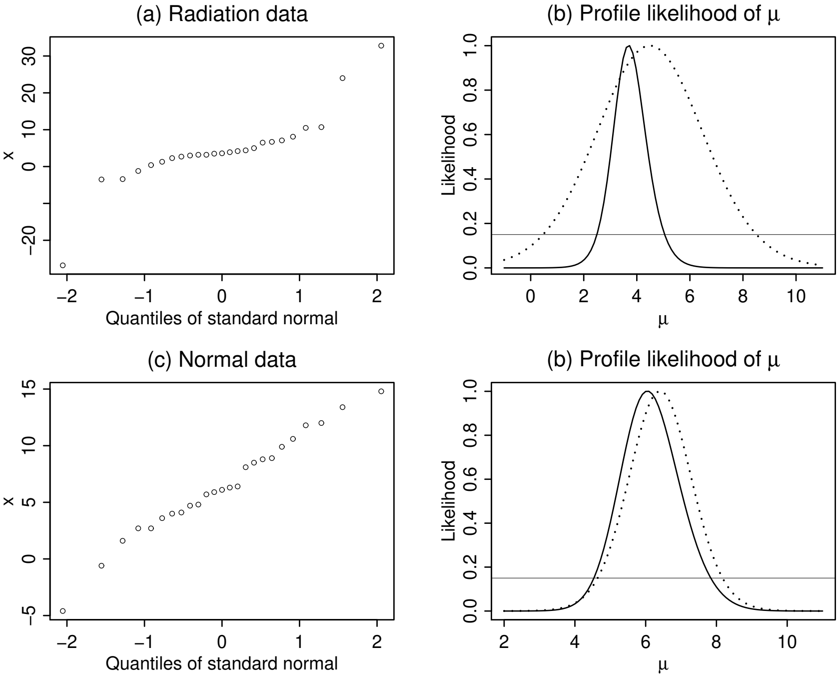

Figure 4.12: *(a) Normal plot of heavy-tailed data; (b) the profile likelihood function of the location parameter assuming the normal model (dashed line) and assuming the Cauchy model (solid line); (c)-(d) the same as (a)-(b) for normal data.*

from which we can compute a profile likelihood for μ, shown in Figure 4.12(b); there is no closed form solution to get the MLE of σ at fixed value of μ, so it needs to be found numerically. The MLE from the Cauchy model is $\widehat{\mu} = 3.7$, closer to the median, and $\widehat{\sigma} = 2.2$.

Illustrating the potential of robust modelling, the profile likelihood is compared with that assuming a normal model. The normal likelihood is centered at the sample mean. There is clearly a better precision for the location parameter μ using the Cauchy model than using the normal model.

(Note that in these two models the parameter μ is comparable as the median of the distributions, but the scale parameter σ in the Cauchy model does not have meaning as a standard deviation.)

What if the data are close to normal? Here is an example of a similar dataset:

```
-4.6  -0.6   1.6   2.7   2.7   3.6   4.0   4.1   4.7   4.8   5.7
 5.9   6.1   6.3   6.4   8.1   8.5   8.8   8.9   9.9  10.6  11.8
12.0  13.4  14.8
```

Figure 4.12(c) indicates that the data are quite normal, and Figure 4.12(d) shows that, in this case, the two models produce comparable results. Generally, there will be some loss of efficiency if we use the Cauchy model for normal data and vice versa. However, as we see in this example, the amount of loss is not symmetrical; it is usually larger if we wrongly assume a normal model, so it is not a bad strategy to err on a heavier-than-normal-tail model.

We can use the AIC (Section 3.5) to choose between the normal and Cauchy models. For the radiation data, the AIC of the normal model is

$$\text{AIC} = -2 \log L(\widehat{\mu}, \widehat{\sigma}^2) + 4$$

$$= n \log(2\pi\hat{\sigma}^2) + n + 4 = 189.7,$$

where $n = 25$, and $\hat{\sigma}^2 = 98.64$. For the Cauchy model, using the MLEs $\hat{\mu} = 3.7$ and $\hat{\sigma} = 2.2$, the AIC is 169.3. Therefore, as expected from the QQ-plot, the Cauchy model is preferred. For the second dataset, the AICs are 148.6 and 156.6, now preferring the normal model.

Another useful general technique is to consider a larger class of location models that includes both the Cauchy and the normal families. Such class is provided the t-family with ν degrees of freedom. The standard density is given by

$$f_0(x) = \frac{\Gamma\{(\nu+1)/2\}}{\Gamma(1/2)\Gamma(\nu/2)\nu^{1/2}}(1 + x^2/\nu)^{-(\nu+1)/2}.$$

If $\nu = 1$ we obtain the Cauchy model, while as $\nu \to \infty$ we get the normal model. So the parameter ν is allowed to vary we can adapt to the tail behaviour in the data.

To find which degree of freedom ν is appropriate for a particular data, one can maximize the profile likelihood of ν at several values such as $\nu = 1$, 2, 4, 8 and ∞. Inference on μ can be performed at the observed value of $\hat{\nu}$. As is commonly the case with complex models, all computations must be done numerically. The application of the t-family models on the above datasets is left as an exercise.

Other useful location-scale families of distributions include the logistic family, with the standard density

$$f_0(x) = \frac{e^x}{(1+e^x)^2}, \quad -\infty < x < \infty,$$

and the double exponential or Laplacian family, with density

$$f_0(x) = \frac{1}{2}e^{-|x|}, \quad -\infty < x < \infty.$$

Both of these distributions have heavier-than-normal tails, so they offer protection against outliers if we are interested in the location parameter. The MLE of the location parameter μ under the Laplacian assumption is the sample median, an estimate known to be resistant to outliers.

The existence of a wide variety of models tends to overwhelm a data analyst. It is important to reflect again that model selection is generally harder than a model fitting once a model is chosen.

4.13 Exercises

Exercise 4.1: The following table shows Geissler's data for families of size 2 and 6.

No. boys k	0	1	2	3	4	5	6
No. families n_k	42860	89213	47819				
No. families n_k	1096	6233	15700	22221	17332	7908	1579

(a) For each family size, assuming a simple binomial model(n, θ) for the number of boys, write down the likelihood of θ.

(b) Combining the data from families of size 2 and 6, draw the likelihood of θ; find the MLE of θ and its standard error.

(c) Evaluate if the binomial model is a good fit to the data. Describe the nature of the model violation and discuss what factors might cause it.

Exercise 4.2: Now suppose that (in another experiment on sex-ratio) the family information is based on questioning 100 boys from families of size k that attend a boys-only school. Explain what is unusual in this sampling scheme and what needs to be modified in the model.

Exercise 4.3: Simulate binomial-type data with under- and overdispersion. Compare the realizations with simulated binomial data. Identify real examples where you might expect to see each case.

Exercise 4.4: Another example of binomial mixture is the so-called beta-binomial distribution. Conditional on p let X_p be binomial(n, p), where p itself is random with beta distribution with density

$$f(p) = \frac{1}{B(\alpha, \beta)} p^{\alpha-1}(1-p)^{\beta-1}.$$

If $\alpha = \beta = 1$ then p has uniform distribution. Let X be a sample from the mixture distribution.

(a) Simulate data from the mixture for $n = 10$, $\alpha = 1$ and $\beta = 3$.

(b) Find the mean and variance formula for X, and show that there is an extra-binomial variation.

(c) Verify the marginal probability

$$P(X = x) = \frac{B(\alpha + x, \beta + n - x)}{(n+1)\ B(x+1, n-x+1)B(\alpha, \beta)},$$

for $x = 0, 1, \ldots, n$.

(d) Using Geissler's family data of size 6 in Exercise 4.1, find the joint likelihood of $\theta = (\alpha, \beta)$ as well as the profile likelihood of α, β and the mean of X.

(e) Compare the expected frequencies under the beta-binomial model with the observed frequency.

(f) Compare the profile likelihood of the mean of X in part (d) with the likelihood assuming the data are simple binomial. Comment on the result.

Exercise 4.5: Generalize the negative binomial distribution along the line of the beta-binomial distribution, where the success probability p is a random parameter.

Exercise 4.6: A total of 678 women, who got pregnant under planned pregnancies, were ask how many cycles it took them to get pregnant. The women were classified as smokers and non-smokers; it is of interest to compare the association between smoking and probability of pregnancy. The following table (Weinberg and Gladen, 1986) summarizes the data.

Cycles	Smokers	Non-smokers.
1	29	198
2	16	107
3	17	55
4	4	38
5	3	18
6	9	22
7	4	7
8	5	9
9	1	5
10	1	3
11	1	6
12	3	6
12	7	12

(a) Fit a geometric model to each group and compare the estimated probability of pregnancy per cycle.

(b) Check the adequacy of the geometric model (Hint: devise a plot, or run a goodness-of-fit test.)

(c) Allow the probability to vary between individuals according to a beta distribution (see Exercise 4.4), and find the beta-geometric probabilities.

(d) Fit the the beta-geometric model to the data and check its adequacy.

Exercise 4.7: The Steering Committee of the Physicians' Health Study (1989) also reported the following events during follow-up. Recall that 11,037 subjects were assigned to aspirin group, and 11,034 to placebo. For each endpoint, report the relative risk of aspirin

Events	Aspirin	Placebo
Total cardiac deaths	81	83
Fatal heart attacks	10	28
Sudden deaths	22	12

versus placebo, the profile likelihood, and the approximate 95% CI.

Exercise 4.8: To investigate the effect of race in the determination of death penalty Wolfgang and Riedel (1973) examined the convictions on 55 rapes in 19 Arkansas counties between 1945 and 1965.

(a) Out of 34 black defendants 10 received death penalty, compared to 4 out of 21 white defendants.

(b) Out of 39 cases where the victims were white, 13 led to death penalty, compared to 1 out 15 cases with black victims.

Find in each case the significance of race factor.

Exercise 4.9: We showed that the one-sided P-value from Fisher's exact test of the data in Example 4.3 is 0.20. What do you think is the exact two-sided P-value?

Exercise 4.10: Use the Delta method to show that the standard error of $\widehat{\theta}$ is

$$\mathrm{se}(\widehat{\theta}) = \left(\frac{1}{x} + \frac{1}{y} + \frac{1}{m-x} + \frac{1}{n-y}\right)^{1/2}$$

The Wald statistic $z = \widehat{\theta}/\mathrm{se}(\widehat{\theta})$ satisfies $z^2 \approx \chi^2$. Verify this for several 2×2 tables. This also means that, if the likelihood is regular, the approximate likelihood of H_0 is

$$e^{-\chi^2/2}.$$

Verify this for several 2×2 tables.

Exercise 4.11: To 'solve' the authorship question of 12 of the so-called 'Federalist' papers between Madison or Hamilton, two statisticians Mosteller and Wallace (1964) collated papers of already known authorship, and computed the appearance of some keywords in blocks of approximately 200 words. From 19 papers of Madison's, the appearance of the word 'may' is given in the following table.

Number of occurrences k	Number of blocks n_k
0	156
1	63
2	29
3	8
4	4
5	1
6	1

Check the Poisonness of the data and test whether the observed deviations are significant. Fit a Poisson model and compute a goodness-of-fit test.

Exercise 4.12: Modify the idea of Poisson plot to derive a binomial plot, i.e., to show graphically if some observed data are binomial. Try the technique on Geissler's data in Exercise 4.1.

Exercise 4.13: Another way to check the Poisson assumption is to compare consecutive frequencies (Hoaglin and Tukey, 1985). Under Poisson assumption with mean θ show that

$$kp_k/p_{k-1} = \theta$$

and in practice we can estimate p_k/p_{k-1} by n_k/n_{k-1}.

(a) Apply the technique to the datasets in Example 4.5 and Exercise 4.11.

(b) Develop a method to test whether an observed deviation is significant.

(c) Modify the technique for the binomial and negative binomial distributions, and apply them to the datasets in Exercise 4.11 and 4.1, respectively.

Exercise 4.14: This is an example of binomial-Poisson mixture. Suppose X_n is binomial with parameters n and p, but n itself is a Poisson variate with mean μ. Find the the marginal distribution of X_n. Discuss some practical examples where this mixture model is sensible.

Exercise 4.15: The annual report of a pension fund reported a table of the number of children of 4,075 widows, who were supported by the fund (Cramer, 1955).

Number of children	0	1	2	3	4	5	6
Number of widows	3062	587	284	103	33	4	2

(a) Fit a simple Poisson model and indicate why it fails.

(b) Fit a negative binomial model to the data, and check its adequacy using the χ^2 goodness-of-fit test.

(c) Consider a mixture model where there is a proportion θ of widows without children, and another proportion $(1-\theta)$ of widows with x children, where x is Poisson with mean λ. Show that the marginal distribution of the number of children X follows

$$P(X=0) = \theta + (1-\theta)e^{-\lambda}$$

and, for $k > 0$,

$$P(X=k) = (1-\theta)e^{-\lambda}\frac{\lambda^k}{k!}.$$

This model is also called 'zero-inflated' Poisson model.

(d) Fit the model to the data, and compare its adequacy with the negative binomial model.

Exercise 4.16: Refer to the data in Example 4.6.

(a) Check graphically the two candidate models, the Poisson and negative binomial models, using the technique described in Exercise 4.13.

(b) Verify the fit of Poisson and negative binomial models to the data.

(c) Compare the likelihood of the mean parameter under the two models. (It is a profile likelihood under the negative binomial model.) Comment on the results.

Exercise 4.17: Repeat the previous exercise for the dataset in Exercise 4.11.

Exercise 4.18: The following table (from Evans, 1953) shows the distribution of a species *Glaux maritima* in 500 contiguous areas of 20-cm^2 squares. Compare Poisson and negative

count	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
squares	1	15	27	42	77	77	89	57	48	24	14	16	9	3	1

binomial fits of the data. Interpret the result in terms of clustering of the species.

Exercise 4.19: Verify the profile likelihood in Section 4.7 for the parameter of interest $\theta = \theta_a/\theta_p$.

Exercise 4.20: Assuming that X and Y are Poisson with rate λ_x and λ_y, show that the conditional distribution of X given $X+Y$ is binomial with parameters $n = X+Y$ and $\pi = \lambda_x/(\lambda_x+\lambda_y)$.

Exercise 4.21: Prove Theorem 4.1.

Exercise 4.22: Verify that the first Michelson's experiment is out of line with the rest. If you know the analysis of variance technique, use it to analysis all experiments together, then another one excluding the first experiment. Otherwise, use a series of two-sample t-tests.

Exercise 4.23: Suppose $x_1, \ldots, x_m$ are an iid sample from $N(\mu, \sigma_x^2)$, and $y_1, \ldots, y_n$ from $N(\mu+\delta, \sigma_y^2)$. Define Michelson's first experiment as the x-sample, and the subsequent experiment as the y-sample.

(a) Compute the profile likelihood for δ assuming that $\sigma_x^2 = \sigma_y^2$.

(b) Compute the profile likelihood for δ without further assumption.

(c) Compute the profile likelihood for σ_x^2/σ_y^2.

Exercise 4.24: The following is the average amount of rainfall (in mm/hour) per storm in a series of storms in Valentia, southwest of Ireland. Data from two months are reported below.

```
January 1940
      0.15 0.25 0.10 0.20 1.85 1.97 0.80 0.20 0.10 0.50 0.82 0.40
      1.80 0.20 1.12 1.83 0.45 3.17 0.89 0.31 0.59 0.10 0.10 0.90
      0.10 0.25 0.10 0.90

July 1940
      0.30 0.22 0.10 0.12 0.20 0.10 0.10 0.10 0.10 0.10 0.10 0.17
      0.20 2.80 0.85 0.10 0.10 1.23 0.45 0.30 0.20 1.20 0.10 0.15
      0.10 0.20 0.10 0.20 0.35 0.62 0.20 1.22 0.30 0.80 0.15 1.53
      0.10 0.20 0.30 0.40 0.23 0.20 0.10 0.10 0.60 0.20 0.50 0.15
      0.60 0.30 0.80 1.10 0.20 0.10 0.10 0.10 0.42 0.85 1.60 0.10
      0.25 0.10 0.20 0.10
```

(a) Compare the summary statistics for the two years.

(b) Look at the QQ-plot of the data and, based on the shape, suggest what model is reasonable.

(c) Fit a gamma model to the data from each month. Report the MLEs and standard errors, and draw the profile likelihoods for the mean parameters. Compare the parameters from the two months.

(d) Check the adequacy of the gamma model using using gamma QQ-plot.

Exercise 4.25: In a regular p-parameter exponential family (4.7) show that the natural statistics $T_1(x), \ldots, T_p(x)$ are minimal sufficient. Explain why these statistics are the MLE of their own mean vector, which is a one-to-one function of θ, so the MLE of θ is sufficient.

Exercise 4.26: Let $x_1, \ldots, x_n$ be an iid sample from uniform$(0, \theta)$. The maximum order statistic $x_{(n)}$ is a sufficient estimate of θ, yet the uniform is not in the exponential family. This does not conform to the basic theorem about minimal sufficiency. What condition is not satisfied?

Exercise 4.27: In Section 4.10 verify the general mean and variance formula as derivatives of $A(\theta)$.

Exercise 4.28: Identify the exponential and gamma models as exponential family models; in particular, find the canonical and dispersion parameters, and the function $h(x, \phi)$.

Exercise 4.29: Identify the inverse-Gaussian model as an exponential family model. The density is

$$f(y) = \left(\frac{\lambda}{2\pi y^3}\right)^{1/2} \exp\left\{-\frac{\lambda}{2\mu^2}\frac{(y-\mu)^2}{y}\right\}, \qquad y > 0.$$

Verify that it has mean μ and variance μ^3/λ.

Exercise 4.30: Perform the Monte Carlo test described in Example 4.13 and verify the results.

Exercise 4.31: Compare the likelihood analysis of the mean parameter for the dataset in Exercise 4.16 using the negative binomial model and the general exponential family model. Discuss the advantages and disadvantages of each model.

Exercise 4.32: Compare the likelihood analysis of the mean parameter for the dataset in Exercise 4.4 using the beta-binomial model and the general exponential family model.

Exercise 4.33: Suppose x is in the exponential family with log-density

$$\log p_\mu(x) = x\theta - A(\theta) + h(x),$$

where $\mu = A'(\theta) = EX$ is used as the index; the variance is $V(\mu) = A''(\theta)$. The double exponential family (Efron 1986) is defined as

$$g(x) = c(\theta, \alpha)\alpha^{1/2}\{p_\mu(x)\}^\alpha \{p_x(x)\}^{1-\alpha}.$$

Efron shows that the normalizing constant $c(\theta, \alpha) \approx 1$, and the mean and variance are approximately μ and $V(\mu)/\alpha$.

(a) Starting with a standard Poisson model with mean μ, find the density of a double Poisson model.

(b) Show that the likelihood from part (a) is approximately the same as the extended quasi-likelihood formula in Example 4.14. Hint: use Stirling's approximation for the factorial:

$$n! = \sqrt{2\pi} e^{-n} n^{n+1/2}. \qquad (4.11)$$

Exercise 4.34: Starting with a standard binomial model, find the density function of a double binomial model.

Exercise 4.35: The following dataset is the sulphur dioxide (SO_2) content of air (in 10^{-6} gm/m^3) in 41 US cities averaged over 3 years:

10	13	12	17	56	36	29	14	10	24	110	28	17
8	30	9	47	35	29	14	56	14	11	46	11	23
65	26	69	61	94	10	18	9	10	28	31	26	29
31	16											

Use the Box-Cox transformation family to find which transform would be sensible to analyse the data. Plot the profile likelihood of the mean SO_2 content. Check the normal plot of the data before and after the transformation.

Exercise 4.36: Consider the Box-Cox tranformation family for the data from Michelson's fifth experiment:

890	840	780	810	760	810	790	810	820	850
870	870	810	740	810	940	950	800	810	870

(a) Plot the profile likelihood of λ.

(b) Plot the profile likelihood of μ assuming $\lambda = 1$.

(c) Compare the plot in (b) with the profile likelihood of μ by profiling over λ and σ^2. Comment on the result.

Exercise 4.37: Suppose $x_1, \ldots, x_n$ are an iid sample from the Laplacian distribution with parameters μ and σ. Find the MLE of the parameters and find the general formula for the profile likelihood of μ.

Exercise 4.38: Use the t-family model for a range of degrees of freedom ν for the datasets in Example 4.16, in particular plot the profile likelihood over ν.

Exercise 4.39: Repeat the likelihood analysis given for the datasets in Example 4.16 using logistic and Laplacian models and compare the results.

Exercise 4.40: In a study of physiologic effect of stress, Miralles et al. (1983) measured the level of beta endorphine in patients undergoing a surgery. (Beta endorphine is a morphine-like chemical found in the brain.) The measurements were taken at 12-14 hours and at 10 minutes before surgery,

Patient	12-14 hours	10 minutes	Patient	12-14 hours	10 minutes
1	10.0	6.5	11	4.7	25.0
2	6.5	14.0	12	8.0	12.0
3	8.0	13.5	13	7.0	52.0
4	12.0	18.0	14	17.0	20.0
5	5.0	14.5	15	8.8	16.0
6	11.5	9.0	16	17.0	15.0
7	5.0	18.0	17	15.0	11.5
8	3.5	42.0	18	4.4	2.5
9	7.5	7.5	19	2.0	2.0
10	5.8	6.0			

(a) Draw the QQ plot of the data at 12-14 hours and 10 minutes prior to surgery, and comment on the normality.

(b) Draw also the QQ plot of the change in the beta endorphine level. To get a measure of change note that beta endorphine level is a positive variate, so it maybe more meaningful to use a ratio or to consider a log transformation.

(c) Compare the likelihood of the location parameter assuming normal and Cauchy models for the three variables above.