ST414: Advanced Topics in Statistics

Asymptotic Statistics

An Introduction

Ioannis Kosmidis
http://go.warwick.ac.uk/kosmidis

May 13, 2010

Summary

The aim of this topics course is to introduce students to basic asymptotic methods that are used in Statistics and to highlight their importance. Special attention will be paid to likelihood based asymptotics for parametric models, focusing on the asymptotic properties of the maximum likelihood estimator.

The course will combine the development of general asymptotic tools with their application to some much-used results in statistics (such as, for example, the derivation of the asymptotic distribution of the log likelihood-ratio statistic, the 1/2 adjustment to the binomial counts for bias reduction in log-odds estimation etc.).

Advantages and shortcomings of different asymptotic methods will be explored using computer simulation.

For the development of the current set of notes, the main references that have been consulted are Pace and Salvan (1997) and Young and Smith (2005) (mainly chapters 8 and 9). The textbook by Brazzale et al. (2007) contains numerous real-data illustrations on how asymptotic results can be used to draw accurate inferences in complex situations. Some other textbooks that have been consulted are Cox and Hinkley (1974), Barndorff-Nielsen and Cox (1989) and van der Vaart (1998) and Cox (2006).

Chapter 1

Introduction

1.1 Inference for a scalar parameter

Consider observations $y_1, \ldots, y_n$ from n independent random variables $Y_1, \ldots, Y_n$ with density functions $f_{Y_i}(y_i; \beta)$ depending on a scalar parameter β.

The log-likelihood for β is defined as $l(\beta; \{y_i\}) = \sum_{i=1}^{n} \log f_{Y_i}(y_i; \beta)$. The maximum likelihood estimator $\hat{\beta}$ is defined as the value of β that maximizes $l(\beta) \equiv l(\beta; \{Y_i\})$. The observed information function is defined as $j(\beta) = -\mathrm{d}^2 l(\beta)/\mathrm{d}\beta^2$ and the expected information as $i(\beta) = E_\beta(j(\beta))$. Both $1/j(\hat{\beta})$ and $1/i(\hat{\beta})$ provide an estimator for the standard error of the asymptotic distribution of the maximum likelihood estimator.

The log-likelihood, the maximum likelihood estimator and the observed/expected information can serve as the basic ingredients for likelihood based inferences about β. For this, these quantities can be combined to provide a *pivotal quantity*:

Definition 1.1. A function $T(S, \beta)$ of $S \equiv S(Y_1, \ldots, Y_n)$ and the parameter β, is said to be a pivotal quantity for inferences about β (or simply a pivot), if its distribution (pivotal distribution) does not depend on β and if for any value s of S the function $T(s, \beta)$ is monotone decreasing in β.

If we were able to find a pivotal quantity for a given problem, then we could easily draw inferences about β by constructing *confidence intervals* and calculating *p-values*. In particular, if we knew that $T(S, \beta)$ is a pivotal quantity then we could find constants c_1 and c_2 such that

$$P(c_1 \leq T(S, \beta) \leq c_2) = 1 - \alpha \,, \quad \text{for all} \quad \beta \,,$$

by merely using the quantiles of the pivotal distribution. The above relationship can be rewritten as

$$P(L(S) \leq \beta \leq U(S)) = 1 - \alpha \,, \quad \text{for all} \quad \beta \,. \tag{1.1}$$

Hence for an observed value s of S, $(L(s), U(s))$ is a $100(1-\alpha)\%$ confidence interval[1] for β. Equivalently, for testing the hypothesis $H_0 : \beta \leq \beta_0$ against the alternative $H_1 : \beta > \beta_0$ we could calculate the p-value $P(Z \geq T(s, \beta_0))$, where Z is distributed according to the pivotal distribution, and reject H_0 for small p-values.

[1]Expression (1.1) should not be seen as assigning probabilities to the unknown parameter β. It is merely specifying a hypothetical long run of statements about β a proportion $1-\alpha$ of which are correct. If we observe s for S, $\beta \in (T(s), U(s))$ is one of this long run of statements.

Example 1.1. (Inferences about the mean of a Normal population with known variance) Suppose that $Y_1, \ldots, Y_n$ are independent and identically distributed random variables from a Normal distributed with mean μ and known variance σ^2. The maximum likelihood estimator for μ is the sample average $\bar{Y}$ which is distributed according to a Normal distribution with mean μ and variance σ^2/n. Thus $T(\bar{Y}, \mu) = \sqrt{n}(\bar{Y} - \mu)/\sigma$ is a pivot for μ and the pivotal distribution is $N(0, 1)$. Hence upon observing $y_1, \ldots, y_n$, a $100(1 - a)\%$ equi-tailed confidence interval for μ is

$$\left(\bar{y} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \, , \, \bar{y} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right) ,$$

where $z_{1-\alpha/2}$ is the $(1 - \alpha/2)$th quantile of the standard Normal. □

However, in many cases it is difficult or even impossible to construct pivotal quantities. In those cases an *approximate pivot* could be used instead to obtain approximate inferences about β. Approximate pivots are defined again as in (1.1) with the difference that $T(S, \beta)$ has *asymptotically* a distribution not depending on β and, possibly, is a monotone decreasing function in β only in a region of the observed value s. Asymptotically here means as $n \to \infty$, where n is the sample size, or more generally some other measure of how information about β accumulates. The most well-used approximate pivots are the signed likelihood root, the score pivot and the Wald pivot, which are defined as

$$\begin{aligned} r(\beta) &\equiv r(\hat{\beta}, \beta) = \mathrm{sign}(\hat{\beta} - \beta) \left[2 \left\{ l(\hat{\beta}) - l(\beta) \right\} \right]^{1/2} , \\ s(\beta) &\equiv s(\hat{\beta}, \beta) = j(\hat{\beta})^{-1/2} \mathrm{d}l(\beta)/\mathrm{d}\beta \, , \\ t(\beta) &\equiv t(\hat{\beta}, \beta) = j(\hat{\beta})^{1/2} (\hat{\beta} - \beta) \, , \end{aligned}$$

respectively, and, as we will see in later chapters, asymptotically all have a standard Normal distribution. The same limiting distribution applies if $j(\beta)$ is replaced by either $j(\hat{\beta})$ or $i(\beta)$ or $i(\hat{\beta})$. Despite the fact that all $r(\theta)$, $s(\theta)$ and $t(\theta)$ have the same asymptotic distributions, the accuracy of the $N(0, 1)$ approximation for finite n has to justified, most often by simulation.

1.2 A simple example

As an example consider the illustrative setting in Brazzale et al. (2007, Section 2.2). Assume n independent and identically distributed random variables from the exponential distribution with mean $1/\lambda$, that is

$$f(y_i; \lambda) = \lambda \exp\left\{-\lambda y_i\right\} , \quad \lambda > 0 \, , y_i > 0 \quad (i = 1, \ldots, n) \, . \tag{1.2}$$

A simple analysis gives that $l(\lambda) = n(\log \lambda - \lambda \bar{Y})$ is unimodal with a unique maximum at $\hat{\lambda} = 1/\bar{y}$ and that $j(\lambda) = i(\lambda) = n/\lambda^2$. Then, the approximate pivots of the previous section have the forms

$$\begin{aligned} r(\lambda) &= \mathrm{sign}\left(1 - \bar{Y}\lambda\right) \sqrt{2n \left\{ \bar{Y}\lambda - \log(\bar{Y}\lambda) - 1 \right\}} \, , \\ s(\lambda) &= \frac{\sqrt{n}\left(1 - \bar{Y}\lambda\right)}{\bar{Y}\lambda} \\ t(\lambda) &= \sqrt{n}\left(1 - \bar{Y}\lambda\right) \, . \end{aligned}$$

Brazzale et al. (2007) argue that the evidence on some unknown parameter β can be appropriately summarized by plotting the log-likelihood for β and the so-called *significance function*, which is defined as the p-value for testing the hypothesis $H_0 : \beta \geq \beta_0$ versus the alternative $H_1 : \beta < \beta_0$. For example, a significance function based on $r(\beta)$ is $\Phi(r(\beta))$, where $\Phi(.)$ denotes the distribution function of a standard Normal random variable.

In this particular case of an exponential sample, an exact pivot can be found by noting that $X = n\lambda\bar{Y}$ is distributed according to a gamma distribution with shape n and scale 1. That is the distribution function of X is

$$f(x) = \frac{x^{n-1}\exp(-x)}{\Gamma(n)}, \quad x > 0\,.$$

Then, $e(\lambda) = z_{1-F(n\lambda\bar{Y})}$, $F(x) = \int_0^x f(t)dt$ is an exact pivot with $N(0,1)$ pivotal distribution and can be used to examine the performance of inferences based on approximate pivots. The $100(1-a)\%$ confidence interval based on the exact pivot is

$$\left(\frac{g_{n,a/2}}{n\bar{y}}, \frac{g_{n,1-a/2}}{n\bar{y}}\right),$$

where $g_{n,a/2}$ is the $a/2$th quantile of the gamma distribution with shape n and scale 1.

Assume that we observe $\bar{y} = 1$ from a sample of size $n = 1$. Substituting to the latter confidence interval, the 95% confidence interval for λ based on the exact pivot is $(0.025, 3.689)$.

The log-likelihood and the significance functions for the pivots can be found in Figure 1.1. The log-likelihood is quite asymmetric in this example and hence we expect confidence intervals based on the asymptotic normality of the Wald pivot not to perform well. Indeed, a simple calculation, as the one of Example 1.1, shows that the $100(1-a)\%$ approximate confidence interval based on the Wald pivot has the form

$$\left(\frac{1}{\bar{y}} - \frac{z_{1-a/2}}{\sqrt{n}\bar{y}}, \frac{1}{\bar{y}} + \frac{z_{1-a/2}}{\sqrt{n}\bar{y}}\right),$$

and so for $n = 1$ and $\bar{y} = 1$ and at the 95% nominal level, the Wald confidence interval for λ is $(-0.96, 2.96)$, which is quite unsatisfactory for a positive parameter. Furthermore, note that in the bottom plot of Figure 1.1 the score pivot does not intersect $z_{0.025}$ (the bottom grey line) for the plotted values of λ. Actually, for $n = 1$ and $\bar{y} = 1$ the score pivot never intersects that line as it has an asymptote at -1. Hence, based on such data, according to the score pivot, we do not have sufficient evidence to reject $H_0 : \lambda \geq \lambda_0$ for any value of λ_0 and at any sensible significance level. Furthermore, in that case, the 95% approximate confidence interval has an infinite upper limit. On the other hand, the signed likelihood root seems to perform well, even for $n = 1$, giving $(0.057, 4.403)$ for a 95% approximate confidence interval for λ.

Of course, we should not be strict in judging the performance of the above approximate pivots, because here we used them (inappropriately) for inferences for λ when $n = 1$ and they are designed to perform well as n grows without limit. A more careful examination results in Table 1.1 which contains 95% confidence intervals based on the four aforementioned pivots for $\bar{y} = 1$ and several values of n. Comparing the approximate confidence intervals with the exact one we clearly see that, in this case, the likelihood root performs considerably well for all values of n, while the score and Wald pivots give satisfactory results for $n = 5$ and start

Table 1.1: 95% equi-tailed confidence intervals for λ based on the exact, Wald, score and likelihood root pivots for $\bar{y} = 1$ and various values of n.

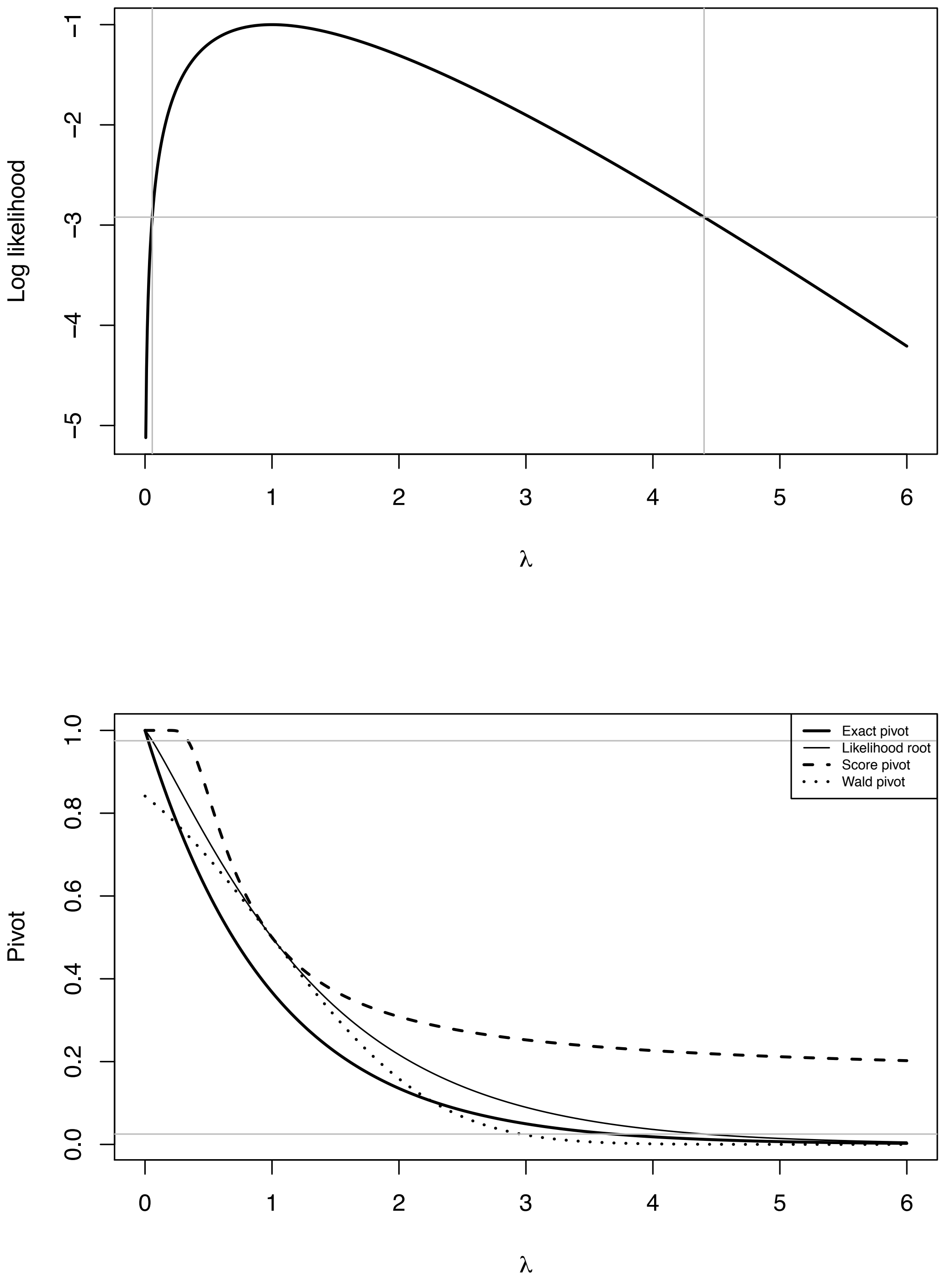
Pivot	$n = 1$		$n = 5$		$n = 10$		$n = 20$		$n = 50$	
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
$e(\lambda)$	0.025	3.689	0.325	2.048	0.480	1.708	0.611	1.484	0.742	1.296
$t(\lambda)$	-0.960	2.960	0.124	1.877	0.380	1.620	0.562	1.438	0.723	1.277
$s(\lambda)$	0.338	∞	0.533	8.099	0.617	2.630	0.695	1.780	0.783	1.383
$r(\lambda)$	0.057	4.403	0.359	2.149	0.501	1.754	0.623	1.504	0.748	1.303

giving good approximations to the exact confidence interval for $n \geq 10$, being very accurate as soon as $n = 50$.

Hence, even for small sample sizes inferences for λ can be based on approximate pivots with small error.

As already mentioned, exact pivots are not generally available and thus the need for asymptotic approximations of good performance is apparent. The discussion of the next chapters focuses on developing the asymptotic arguments that result in the aforementioned asymptotic pivots, developing corrections to the moments of the maximum likelihood estimator that can result in estimators with improved properties, and constructing more accurate pivots.

Figure 1.1: Log-likelihoog of λ (top) with the grey vertical lines showing the 95% approximate confidence interval for λ based on the likelihood root. Significance functions for the exact and approximate pivots (bottom). The grey lines correspond to probabilities 0.975 and 0.025.



Chapter 2

Some basic tools for asymptotics

The current chapter concerns the exposition of the main tools that are used for the development of asymptotic approximations in Statistics. The presentation begins with Taylor's theorem, which enables the approximation of any sufficiently smooth function by an appropriate polynomial. A review of the basic modes of convergence of sequences of random variables is given, with parallel review of well-known limiting results in Probability and Statistics, such as the weak law of large numbers and the central limit theorem. The moments and cumulants of a distribution function are thoroughly studied as they provide the basic ingredients of the approximations that are considered in the current course. In conclusion, the general form of a (stochastic) asymptotic expansion is presented.

2.1 Taylor series

Theorem 2.1. (Taylor) *Let f be a real function on $[a, b]$ and suppose that the mth derivative of f, denoted $f^{(m)}$, is continuous on $[a, b]$ and $f^{(m+1)}(u)$ exists for every $u \in (a, b)$. If x and x_0 are distinct points of $[a, b]$, then there exists c between x and x_0 such that*

$$f(x) = \sum_{k=0}^{m} (x - x_0)^k \frac{f^{(k)}(x_0)}{k!} + (x - x_0)^{m+1} \frac{f^{(m+1)}(c)}{(m+1)!} . \tag{2.1}$$

Theorem 2.1 is a core result of mathematical analysis, which allows the approximation of any sufficiently smooth function by a polynomial of finite degree.

Note that Taylor's theorem provides the error encountered when $f(x)$ is approximated by the first term of the right hand side of (2.1). More importantly, if $R_m(x)$ is that error of approximation and given that $\left|f^{(m+1)}(x)\right| \le N$, $N > 0$, for $x \in (a, b)$, then

$$|R_m(x)| \le \frac{N \left|x - x_0\right|^{m+1}}{(m+1)!} .$$

Furthermore, if f is infinitely differentiable in a neighbourhood of x_0 and if $|x - x_0|^k \left|f^{(k)}(c)\right| / k! \to 0$ as $k \to \infty$, we can write

$$f(x) = \sum_{k=0}^{\infty} (x - x_0)^k \frac{f^{(k)}(x_0)}{k!} . \tag{2.2}$$

Expression (2.2) is called the *Taylor series expansion of f around x_0*.

If f is a function of d variables then the Taylor series expansion of f around a d-vector $a = (a_1, \ldots, a_d)$ is

$$f(x_1, \ldots, x_d) = \sum_{k_1=1}^{\infty} \cdots \sum_{k_d=1}^{\infty} \frac{(x_1 - a_1)^{k_1} \ldots (x_d - a_d)^{k_d}}{k_1! \ldots k_d!} \left[\frac{\partial^{k_1+\ldots+k_d} f(x_1, \ldots x_d)}{\partial x_1^{k_1} \ldots \partial x_d^{k_d}} \right]_{x=a} . \quad (2.3)$$

Some important Taylor series expansions around 0 (also called MacLaurin series) along with their convergence range are

$$\exp(x) = \sum_{k=0}^{\infty} \frac{x^k}{k!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots , \quad x \in \Re , \quad (2.4)$$

$$\log(1+x) = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{x^k}{k} = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \ldots , \quad |x| \leq 1 , x \neq -1 , \quad (2.5)$$

$$\log(1-x) = -\sum_{k=1}^{\infty} \frac{x^k}{k} = -(x + \frac{x^2}{2} + \frac{x^3}{3} + \frac{x^4}{4} + \ldots) , \quad |x| \leq 1 , x \neq 1 . \quad (2.6)$$

2.2 Modes of convergence and some limiting results

2.2.1 Convergence of sequences of random variables

Definition 2.1. The sequence of independent and identically distributed random variables $\{Y_1, Y_2, \ldots\}$ is said to *converge in probability* to $c \in \Re$ if, given $\epsilon > 0$ and $\delta > 0$, there exists an $n_0(\delta, \epsilon)$ such that, for all $n > n_0(\delta, \epsilon)$,

$$P(|Y_n - c| > \epsilon) < \delta .$$

We write $Y_n \xrightarrow{p} c$[1].

Definition 2.2. The sequence of independent and identically distributed random variables $\{Y_1, Y_2, \ldots\}$ is said to *converge in distribution* if there exists a distribution function F such that

$$\lim_{n \to \infty} P(Y_n \leq y) = F(y) ,$$

for all y that are continuity points of the limiting distribution. If F is the distribution function of the random variable Y, we write $Y_n \xrightarrow{d} Y$; if Y has a distribution with a standard code, such as $N(0, 1)$, then we can also write $Y_n \xrightarrow{d} N(0, 1)$.

Particular examples of convergence of sequences of random variables are the weak law of large numbers and the central limit theorem. Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with mean μ and finite variance σ^2. Under these assumptions, the *weak law of large numbers* asserts that $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i \xrightarrow{p} \mu$ and the *central limit theorem* says that $Z_n = \sqrt{n}(\bar{X}_n - \mu)/\sigma \xrightarrow{d} N(0, 1)$.

The extension of the above results to the case of sequences of random vectors in $\Re^p$ (random p-vectors) is straightforward: In the definition of convergence in probability let Y_n

[1] The general definition of convergence in probability refers to convergence to a random variable X and essentially results by replacing c with X in the definition. Nevertheless, the only concept of convergence in probability used in the current notes is convergence to a constant.

be a random p-vector, $c \in \Re^p$ and define $|Y_n - c|$ as a distance metric in $\Re^p$ (for example, the Euclidean distance). Similarly, for the weak law of large numbers and the central limit theorem assume that $X_1, \ldots, X_n$ are independent and identically distributed random p-vectors with mean μ and variance Σ with finite entries. Then $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i \xrightarrow{p} \mu$ and $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N_p(0, \Sigma)$.

Finally, two intuitive but powerful results are presented, which are widely used for the development of asymptotic arguments.

Theorem 2.2. (Continuous mapping) *Let $g : \Re^p \to \Re^k$ be continuous at every point of a set C such that $P(Y \in C) = 1$.*

i) If $Y_n \xrightarrow{d} Y$, then $g(Y_n) \xrightarrow{d} g(Y)$;

ii) If $Y_n \xrightarrow{p} c$, then $g(Y_n) \xrightarrow{p} g(c)$.

Example 2.1. From the central limit theorem $Z_n = \sqrt{n}(\bar{X}_n - \mu)/\sigma \xrightarrow{d} N(0,1)$. Then, by the continuous-mapping theorem $\exp(Z_n) \xrightarrow{d} U$, where a simple change of variable gives that U has a log-Normal distribution with density

$$\frac{1}{u\sqrt{2\pi}} \exp\left\{-\frac{(\log u)^2}{2}\right\}, \quad u > 0\,.$$

□

Lemma 2.1. (Slutzky) *Let X_n, X and Y_n be random variables. If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, $c \in \Re$ then*

i) $X_n + Y_n \xrightarrow{d} X + c$;

ii) $Y_n X_n \xrightarrow{d} cX$;

iii) $X_n / Y_n \xrightarrow{d} X/c$, if $c \neq 0$.

Slutzky's lemma says that, if $Y_n \xrightarrow{p} c$ then variation in Y_n can be omitted when examining the limiting distributions of $X_n + Y_n$, $Y_n X_n$ or X_n/Y_n. The result also extends to the case where X_n is a sequence of random p-vectors converging in distribution to a random p-vector X and Y_n is a sequence of random matrices which converges in probability (entry-wise) to a fixed matrix C of appropriate dimension; then $Y_n X_n \xrightarrow{d} CX$.

Example 2.2. (t-test for the mean) Consider observations on independent and identically distributed random variables $X_1, X_2, \ldots, X_n$ from a Normal population with unknown mean μ and unknown variance σ^2. Interest is on testing the null hypothesis $H_0 : \mu = \mu_0$. A test statistic for this null hypothesis is $T_n = \sqrt{n}(\bar{X}_n - \mu_0)/\hat{\sigma}_n$ (t-statistic), where $\hat{\sigma}_n^2 = (n-1)^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ is the sample variance. Large absolute values of the statistic indicate departures from the null hypothesis. To formally perform the test we need to know the distribution of the test statistic, at least asymptotically.

By the weak law of large numbers $n^{-1}\sum_{i=1}^n X_i^2 \xrightarrow{p} E(X_1^2)$ and $\bar{X}_n \xrightarrow{p} \mu$, and hence by the continuous-mapping theorem $\bar{X}_n^2 \xrightarrow{p} \mu^2$. Another application of the continuous-mapping theorem gives

$$\hat{\sigma}_n^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}_n^2\right) \xrightarrow{p} 1 \cdot \left\{E(X_1^2) - \mu^2\right\} = \sigma^2\,,$$

and thus $\hat{\sigma}_n \xrightarrow{p} \sigma$. Under the null hypothesis, $X_i \sim N(\mu_0, \sigma^2)$ and so by the central limit theorem $\sqrt{n}(\bar{X}_n - \mu_0) \xrightarrow{d} Y$ with $Y \sim N(0, \sigma^2)$. Finally, Slutzky's lemma gives that $\sqrt{n}(\bar{X}_n - \mu_0)/\hat{\sigma}_n \xrightarrow{d} Y/\sigma$ and so $T_n \xrightarrow{d} N(0,1)$. Hence, T_n is an approximate pivot and p-values can be calculated based on its asymptotic normality.

Note that, in the present set up it is known that the exact sampling distribution of T_n is the t distribution with $n-1$ degrees of freedom. □

For brevity, in what follows, a sequence $\{a_1, a_2, \ldots\}$ will be often denoted simply as $\{a_n\}$.

2.2.2 Stochastic order symbols

The stochastic order symbols O_p and o_p are the most commonly used symbols for describing the asymptotic order of random quantities and are defined as follows:

Definition 2.3. Consider a sequence of random variables $\{X_n\}$ and a sequence of constants $\{a_n\}$. We write $X_n = o_p(a_n)$ if $X_n/a_n \xrightarrow{p} 0$.

Definition 2.4. Consider a sequence of random variables $\{X_n\}$ and a sequence of constants $\{a_n\}$. We write $X_n = O_p(a_n)$ if for every $\epsilon > 0$ there exists $K(\epsilon) > 0$ and $n_0(\epsilon)$ such that, for all $n > n_0(\epsilon)$,

$$P\left(\left|\frac{X_n}{a_n}\right| \leq K(\epsilon)\right) > 1 - \epsilon\,.$$

The statement $X_n = O_p(1)$ is equivalent to saying that $\{X_n\}$ is *bounded in probability.* If $X_n \xrightarrow{d} X$, where X has a distribution not depending on n, then $X_n = O_p(1)$. The converse is not generally true, but is almost true; by a result called *Prohorov's theorem*, if $X_n = O_p(1)$ then there is a subsequence $\{X_{n_1}, X_{n_2}, \ldots\}$ with $X_{n_j} \xrightarrow{d} X$ as $j \to \infty$, for some X.

The above definitions extend to the case of sequences of random p-vectors, by understanding $X_n = o_p(a_n)$ as if $||X_n|| = o_p(a_n)$ and $X_n = O_p(a_n)$ as if $||X_n|| = O_p(a_n)$, where $||.||$ denotes some norm in $\Re^p$. (for definitions and examples on the various types of stochastic convergence the reader is referred to van der Vaart, 1998, Chapter 2).

These symbols first appeared in Mann and Wald (1943) among several other symbols denoting different kinds of stochastic relationships, and they are generalizations of their deterministic counterparts $o(.)$ and $O(.)$ (usually referred to as the Landau symbols).

Definition 2.5. Consider two sequences of real constants $\{a_n\}$ and $\{b_n\}$. We write $b_n = o(a_n)$ if $\lim_{n\to\infty} |b_n/a_n| = 0$.

Definition 2.6. Consider two sequences of real constants $\{a_n\}$ and $\{b_n\}$. We write $b_n = O(a_n)$ if there exists $\epsilon > 0$ and positive integer $N(\epsilon)$ such that

$$\text{if} \quad n \geq N(\epsilon) \quad \text{then} \quad |b_n| < \epsilon |a_n|$$

or, alternatively, $\limsup_{n\to\infty} |b_n|/|a_n| < \infty$.

By the above definitions, for any real constant c, $O_p(a_n)$, $o_p(a_n)$, $O(a_n)$, $o(a_n)$ are equivalent to $ca_nO_p(1)$, $ca_no_p(1)$, $ca_nO(1)$, $ca_no(1)$, respectively. Also, while $X_n = O_p(n^c)$ implies $X_n = O_p(n^{c+1})$, $X_n = O_p(n^{c+1})$ does not necessarily mean that $X_n = O_p(n^c)$, and the same is true when O_p is replaced with either O or o or o_p. So, in expressions like $X_n = o_p(a_n)$ the use of the equality symbol is a slight abuse of notation. However, its use is convenient and it denotes the assignment of the property in the right hand side to the quantities of the left hand side.

Some of the properties of stochastic and deterministic order symbols are given below and they are used without comment throughout the notes. If a, b are real numbers and $k = \max\{a, b\}$ then

Products	Sums
$o(n^a)o(n^b) = o(n^{a+b})$	$o(n^a) + o(n^b) = o(n^k)$
$O(n^a)O(n^b) = O(n^{a+b})$	$O(n^a) + O(n^b) = O(n^k)$
$O(n^a)o(n^b) = o(n^{a+b})$	$O(n^a) + o(n^b) = O(n^k)$
$o_p(n^a)o_p(n^b) = o_p(n^{a+b})$	$o_p(n^a) + o_p(n^b) = o_p(n^k)$
$O_p(n^a)O_p(n^b) = O_p(n^{a+b})$	$O_p(n^a) + O_p(n^b) = O_p(n^k)$
$O_p(n^a)o_p(n^b) = o_p(n^{a+b})$	$O_p(n^a) + o_p(n^b) = O_p(n^k)$
$o_p(n^a)o(n^b) = o_p(n^{a+b})$	Compositions
$O_p(n^a)o(n^b) = o_p(n^{a+b})$	$O_p(O(n^a)) = O_p(n^a)$
$O(n^a)o_p(n^b) = o_p(n^{a+b})$	$o(O_p(n^a)) = o_p(n^a)$
$O(n^a)O_p(n^b) = O_p(n^{a+b})$	$o_p(O_p(n^a)) = o_p(n^a)$

The compositions above represent the effect of a linear function $f(x)$ on a quantity of known stochastic or deterministic order. For example, the equivalence $o(O_p(n^a)) = o_p(n^a)$ is interpreted as if $f(x) = o(x)$ and $X_n = O_p(n^a)$, then $f(X_n) = o_p(n^a)$. Lastly, we mention without proof a useful result that describes when an O_p quantity is o_p and gives the appropriate order for the convergence in probability.

Theorem 2.3. (A connection between O_p and o_p) *If $X_n = O_p(n^{-a})$ with $a > 0$ then $X_n = o_p(n^{-a+t})$, for every $t > 0$.*

The above result is used in asymptotic expansions to formally justify the omission of lower order terms, ensuring that under repeated sampling they converge in probability to zero faster than the included terms.

A complete treatment of stochastic order symbols and illustrative examples of their use is given in Bishop et al. (1975, Section 14.4).

2.3 Moments, cumulants and their generating functions

2.3.1 The moment generating function

Let Y be a random variable with density $f_Y(y)$ and let $\mu_r = E(Y^r)$ denote the rth moment of Y and $\bar{\mu}_r = E\{(Y - \mu)^r\}$ the rth central moment of Y (moment about the mean), where $\mu = \mu_1$.

All moments of Y can be derived from the moment generating function of Y,

$$M_Y(t) = E\{\exp(tY)\}, \quad t \in \Re .$$

If the moment generating function of Y exists (i.e. if $M_Y(t) < \infty$ for $|t| < t_0$, for some $t_0 > 0$) then

i) Y has finite moments of any order, and

ii) the moment generating function may be expanded as a power series with some convergence range $R \geq t_0$:

$$M_Y(t) = 1 + t\mu_1 + \frac{t^2}{2!}\mu_2 + \frac{t^3}{3!}\mu_3 + \frac{t^4}{4!}\mu_4 + \dots .$$

Then, differentiating the above expression with respect to t, the rth moment of Y is

$$\mu_r = \left.\frac{\mathrm{d}^r}{\mathrm{d}t^r} M_Y(t)\right|_{t=0} \quad (r = 1, 2, \dots)\,,$$

and the rth central moment of Y is

$$\bar{\mu}_r = \left.\frac{\mathrm{d}^r}{\mathrm{d}t^r} M_{Y-\mu}(t)\right|_{t=0} \quad (r = 1, 2, \dots)\,.$$

There is a correspondence between convergence in distribution of a sequence of random variables and pointwise convergence of the sequence of their moment generating functions. Specifically, if $\{Y_n\}$ is a sequence of random variables with corresponding sequence of moment generating functions $\{M_{Y_n}(t)\}$, and if Y is a random variable with moment generating function $M_Y(t)$, then

i) if $Y_n \xrightarrow{d} Y$ and $M_n(t) < c < \infty$ for $|t| < t_0$, $t_0 > 0$, with c a constant not depending on n, then

$$\lim_{n\to\infty} M_{Y_n}(t) = M_Y(t)\,, \quad |t| < t_0\,;$$

ii) if $\lim_{n\to\infty} M_{Y_n}(t) = M_Y(t)$, $|t| < t_0$, $t_0 > 0$ then $Y_n \xrightarrow{d} Y$, with Y a random variable with moment generating function $M_Y(t)$;

iii) under the same assumptions as either in i) or in ii), the rth moment of Y_n converges to the rth moment of Y $(r = 1, 2, \dots)$.

Example 2.3. (Normal distribution) Consider a normally distributed random variable Y with mean μ and variance σ^2. The moment generating function is

$$\begin{aligned}
M_Y(t) &= E\left\{\exp(tY)\right\} \\
&= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} \exp\left\{ty - \frac{(y-\mu)^2}{2\sigma^2}\right\} dy \\
&= \exp\left\{\frac{-\mu^2 + (t\sigma^2+\mu)^2}{2\sigma^2}\right\} \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} \exp\left\{-\frac{y^2 - 2(t\sigma^2+\mu)y + (t\sigma^2+\mu)^2}{2\sigma^2}\right\} dy \\
&= \exp\left(t\mu + \frac{t^2\sigma^2}{2}\right)\,.
\end{aligned}$$

Hence, the first four moments are

$$\begin{aligned}
\mu_1 &= \mu\,, \\
\mu_2 &= \mu^2 + \sigma^2\,, \\
\mu_3 &= \mu^3 + 3\mu\sigma^2\,, \\
\mu_4 &= \mu^3 + 6\mu^2\sigma^2 + 3\sigma^4\,,
\end{aligned}$$

and the first four central moments are

$$
\begin{aligned}
\bar{\mu}_1 &= 0\,, \\
\bar{\mu}_2 &= \sigma^2\,, \\
\bar{\mu}_3 &= 0\,, \\
\bar{\mu}_4 &= 3\sigma^4\,.
\end{aligned}
$$

Actually, all the odd numbered central moments of a normally distributed random variable are 0. □

Example 2.4. (log-Normal distribution) Consider a random variable X that has the log-Normal distribution with parameters μ and σ^2, that is

$$
f_X(x) = \frac{1}{x\sqrt{2\pi}\sigma}\exp\left\{-\frac{(\log x - \mu)^2}{2\sigma^2}\right\}\,, \quad x > 0\,.
$$

Noting that $X = \exp(Y)$ where $Y \sim N(\mu, \sigma^2)$, the rth moment of X is

$$
\mu_r = E\left(X^r\right) = E\left\{\exp(rY)\right\} = \exp\left(r\mu + \frac{r^2\sigma^2}{2}\right) \quad (r = 1, 2, \ldots)\,.
$$

Hence, all the moments of X are well-defined. Nevertheless, the moment generating function of X does not exist; if $t > 0$,

$$
\begin{aligned}
E\left\{\exp(tX)\right\} &= E\left[\exp\left\{t\exp(Y)\right\}\right] \\
&= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{t\exp(y) - \frac{(y-\mu)^2}{2\sigma^2}\right\} dy \\
&\geq \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{t\left(1 + y + \frac{y^2}{2} + \frac{y^3}{6}\right) - \frac{(y-\mu)^2}{2\sigma^2}\right\} dy = \infty\,,
\end{aligned}
$$

because the term in the exponent is a third-degree polynomial in y for which the y^3 term has positive coefficient. □

2.3.2 The cumulant generating function

Let $M_Y(t)$ be finite for $|t| < t_0$. The cumulant generating function of Y is defined as

$$
K_Y(t) = \log M_Y(t)\,, \quad |t| < t_0\,.
$$

Like $M_Y(t)$, $K_Y(t)$ determines the distribution of Y. The function $K_Y(t)$ can also be expanded in power series, with range of convergence $R \geq t_0$, as

$$
K_Y(t) = t\kappa_1 + \frac{t^2}{2!}\kappa_2 + \frac{t^3}{3!}\kappa_3 + \frac{t^4}{4!}\kappa_4 + \ldots\,.
$$

The coefficient κ_r of $t^r/r!$ is called the rth cumulant (or cumulant of order r) of Y. Clearly,

$$
\kappa_r \equiv \kappa_r(Y) = \left.\frac{\mathrm{d}^r}{\mathrm{d}t^r}K_Y(t)\right|_{t=0} \quad (r = 1, 2, \ldots)\,.
$$

Example 2.5. (One-parameter natural exponential families) Consider a random variable Y that has a density (or probability mass) function of the form

$$f_Y(y;\theta) = \exp\{y\theta - k(\theta) + a(y)\} \,, \tag{2.7}$$

where y take values in $C \subseteq \Re$, θ is a scalar parameter, $k : \Re \to \Re$ and $a : C \to \Re$. We say that $f_Y(y)$ is the density for a natural exponential family with *natural parameter* θ. Because $f_Y(y;\theta)$ is a density function

$$\int_C \exp\{y\theta - k(\theta) + a(y)\}\, dy = 1 \,, \tag{2.8}$$

where the integral is replaced by summation over all $y \in C$ when $f_Y(y)$ is a probability mass function. By (2.8), the moment generating function of Y is

$$\begin{aligned} M_Y(y) &= \int_C \exp\{yt + y\theta - k(\theta) + a(y)\}\, dy \\ &= \exp\{k(\theta+t) - k(\theta)\} \int_C \exp\{y(\theta+t) - k(\theta+t) + a(y)\}\, dy \\ &= \exp\{k(\theta+t) - k(\theta)\} \,. \end{aligned}$$

The above result is unchanged when $f_Y(y)$ is a probability mass function.

Consequently, the cumulant generating function has the form

$$K_Y(t) = k(\theta+t) - k(\theta) \,,$$

and hence the rth cumulant of Y is $\kappa_r = \mathrm{d}^r k(\theta)/\mathrm{d}\theta^r$ $(r = 1, 2, \ldots)$. This is the reason that $k(\theta)$ is often called the *cumulant transform* of the family.

Some well-known one-parameter natural exponential families are given in Table 2.1 along with their density or probability mass functions, natural parameters and cumulant transforms.

For example, in the case of a $N(\mu, \sigma^2)$ random variable with σ^2 fixed, direct differentiation of the cumulant transform with respect to θ and substitution of θ by μ/σ^2 gives $\kappa_1 = \mu$, $\kappa_2 = \sigma^2$ and $\kappa_r = 0$ $(r = 3, 4, \ldots)$. Importantly, cumulants of order higher than two are all zero if, and only if, Y has the normal distribution. In the case of a Poisson random variable, all cumulants are equal to μ. □

Location and scale changes to the range of Y

Consider a new random variable $Z = \alpha + \beta Y$ for real constants α and β. Then

$$M_{\alpha+\beta Y}(t) = E\left[\exp\{t(\alpha + \beta Y)\}\right] = \exp(\alpha t) M_Y(\beta t) \,.$$

Thus, $K_{\alpha+\beta Y}(t) = \alpha t + K_Y(\beta t)$, which reveals two useful properties of the cumulants:

i) Cumulants of order 2 and higher are invariant to location changes, that is

$$\begin{aligned} \kappa_1(\alpha + Y) &= \alpha + \kappa_1(Y) \,, \\ \kappa_r(\alpha + Y) &= \kappa_r(Y) \quad (r = 2, 3, \ldots) \,. \end{aligned}$$

ii) By the definition of the rth cumulant of Y as the rth derivative of $K_Y(t)$ at $t = 0$,

$$\kappa_r(\beta Y) = \left.\frac{\mathrm{d}^r}{\mathrm{d}t^r} K_Y(\beta t)\right|_{t=0} = \beta^r \kappa_r(Y) \quad (r = 1, 2, \ldots) \,.$$

That is, all the cumulants of Y are affected by a scale change $(\beta \neq 1)$ and the greater the order of the cumulant the greater the effect of a change in scale is.

Table 2.1: Some well-known distributions from the one-parameter exponential family.

	Normal $N(\mu, \sigma^2)$	Poisson $P(\mu)$	Binomial $B(n, p)$	Gamma $G(\nu, \phi)$
Parameter	$\mu \in \Re$ $\sigma^2 > 0$ fixed	$\mu > 0$	$p \in [0, 1]$ $n \in \{1, 2, \ldots\}$ fixed	$\phi > 0$ $\nu > 0$ fixed
Range of y	$y \in \Re$	$y \in \{0, 1, 2, \ldots\}$	$y \in \{0, 1, \ldots, n\}$	$y > 0$
$f_Y(y)$	$\frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}$	$\frac{\exp(-\mu)\mu^y}{y!}$	$\binom{n}{y} p^y (1-p)^{n-y}$	$\frac{y^{\nu-1}\exp(-y/\phi)}{\phi^\nu \Gamma(\nu)}$
θ	$\frac{\mu}{\sigma^2}$	$\log \mu$	$\log\left(\frac{p}{1-p}\right)$	$-\frac{1}{\phi}$
$k(\theta)$	$\frac{\theta^2\sigma^2}{2}$	$\exp(\theta)$	$m \log\{1 + \exp(\theta)\}$	$-\nu \log(-\theta)$

Cumulants in terms of moments

The expression of cumulants in terms of moments can be explicitly done using expansion (2.5). For example, for the first four cumulants, ignoring powers of t greater than four, we get

$$\begin{aligned} K_Y(t) &= \log\left[1 + \left\{t\mu_1 + \frac{t^2}{2}\mu_2 + \frac{t^3}{3!}\mu_3 + \frac{t^4}{4!}\mu_4 + O(t^5)\right\}\right] \\ &= \left(t\mu_1 + \frac{t^2}{2}\mu_2 + \frac{t^3}{3}\mu_3 + \frac{t^4}{4}\mu_4\right) - \frac{1}{2!}\left(t\mu_1 + \frac{t^2}{2}\mu_2 + \frac{t^3}{3}\mu_3\right)^2 \\ &+ \frac{1}{3!}\left(t\mu_1 + \frac{t^2}{2}\mu_2\right)^3 - \frac{1}{4!}t^4\mu_1^4 + O\left(t^5\right) . \end{aligned}$$

Expanding the latter expression, moving terms of order $O(t^5)$ to the remainder and collecting terms in increasing powers of t gives:

$$K_Y(t) = t\mu_1 + \frac{t^2}{2}\left(\mu_2 - \mu_1^2\right) + \frac{t^3}{3!}\left(\mu_3 - 3\mu_1\mu_2 + 2\mu_1^3\right) + \frac{t^4}{4!}\left(\mu_4 - 3\mu_2^2 - 4\mu_1\mu_3 + 12\mu_1^2\mu_2 - 6\mu_1^4\right) + O\left(t^5\right) .$$

Thus,

$$\begin{aligned} \kappa_1 &= \mu_1 , \\ \kappa_2 &= \mu_2 - \mu_1^2 = \bar{\mu}_2 , \\ \kappa_3 &= \mu_3 - 3\mu_1\mu_2 + 2\mu_1^3 = \bar{\mu}_3 , \\ \kappa_4 &= \mu_4 - 3\mu_2^2 - 4\mu_1\mu_3 + 12\mu_1^2\mu_2 - 6\mu_1^4 = \bar{\mu}_4 - 3\bar{\mu}_2^2 . \end{aligned} \tag{2.9}$$

The expressions of cumulants in terms of central moments, above, can be obtained in two alternative ways:

i) Note that $K_Y(t) = t\mu + \log M_{Y-\mu}(t)$ and expand $\log M_{Y-\mu}(t)$.

ii) Express the rth moment in terms of μ_1 and central moments by noting that $M_Y(t) = \exp(t\mu)M_{Y-\mu}(t)$. Thus,

$$\mu_r = \frac{\mathrm{d}^r}{\mathrm{d}t^r}\left\{\exp(t\mu)M_{Y-\mu}(t)\right\}\bigg|_{t=0} \quad \text{with} \quad \frac{\mathrm{d}^r}{\mathrm{d}t^r}M_{Y-\mu}(t)\bigg|_{t=0} = \bar{\mu}_r \quad (r = 1, 2, \ldots)\,.$$

Note that, expressions (2.9) suggest that any cumulant of order 2 and higher can be expressed solely in terms of central moments. That is true and is a direct consequence of the invariance of cumulants of order 2 or higher to location changes of the random variable.

Standardized cumulants

It is useful to define the standardized cumulants

$$\rho_r = \frac{\kappa_r}{\kappa_2^{r/2}}\,, \quad r = 3, 4, \ldots\,,$$

which are invariant under scale (and inherently location) changes to the range of Y. The first four cumulants are measures of location, variability, skewness (asymmetry) and kurtosis, respectively. The quantities ρ_3 and ρ_4 are the well-known dimensionless indices of skewness and kurtosis.

2.3.3 Generating functions for sums

Let $S_n = \sum_{i=1}^n Y_i$ where $Y_1, \ldots, Y_n$ are n independent random variables. The density of S_n is given by

$$f_{S_n}(s) = \int f_{Y_1}(y_1)\ldots f_{Y_{n-1}}(y_{n-1})f_{Y_n}(s - y_{n-1} - \ldots - y_1)dy_1\ldots dy_{n-1}\,,$$

with a corresponding sum for the mass function, in the discrete case. Except for small values of n, however, the above expression is of little or no use either for analytical or numerical purposes.

In most cases where sums of independent random variables are involved, it turns out that the calculation of the above integral (or sum) is unnecessary for accessing features of the distribution of the sum.

For the moment generating functions of S_n,

$$M_{S_n}(t) = E\left\{\exp\left(t\sum_{i=1}^n Y_i\right)\right\} = \prod_{i=1}^n E\left\{\exp(tY_i)\right\} = \prod_{i=1}^n M_{Y_i}(t)\,.$$

Note that, if $Y_1, \ldots, Y_n$ are also identically distributed copies of a random variable Y, then $M_{S_n}(t) = \{M_Y(t)\}^n$. In that case the moments of S_n can be obtained by the *binomial theorem*. Specifically,

$$M_{S_n}(t) = \left\{1 + \left(t\mu_1 + \frac{t^2}{2!}\mu_2 + \frac{t^3}{3!}\mu_3 + \frac{t^4}{4!}\mu_4 + \ldots\right)\right\}^n = \sum_{k=0}^n \binom{n}{k}\left(\sum_{r=1}^\infty \frac{t^r}{r!}\mu_r\right)^k\,.$$

The first N moments of S_n can now be obtained using similar arguments to the ones that were used earlier for expressing the cumulants in terms of moments (i.e. consider only the relevant terms of the infinite series, expand and then match the coefficients of $t^r/r!$).

For the cumulant generating function,

$$K_{S_n}(t) = \log M_{S_n}(t) = \sum_{i=1}^{n} K_{Y_i}(t)\,.$$

Thus, by the definition of a cumulant, $\kappa_r(S_n) = \sum_{i=1}^n \kappa_r(Y_i)$, and if $Y_1, \ldots, Y_n$ are identically distributed, $\kappa_r(S_n) = n\kappa_r(Y)$ $(r = 1, 2, \ldots)$. The simplicity of the latter expressions motivate the use of cumulants instead of moments when dealing with sums of independent random variables.

Example 2.6. (Sum of i.i.d exponential random variables with mean ϕ) Consider independent and identically distributed random variables $X_1, \ldots X_n$ from an exponential distribution with density

$$\frac{1}{\phi}\exp\left(-\frac{x}{\phi}\right)\,, \quad x > 0\,, \phi > 0\,.$$

Interest is on finding the distribution of $Y_n = \sum_{i=1}^n X_i$ by using the properties of the cumulant generating function. The moment generating function of X_i is

$$\begin{aligned} M_{X_i}(t) &= \frac{1}{\phi}\int_0^\infty \exp\left\{tx - \frac{x}{\phi}\right\} \\ &= \frac{1}{1-t\phi}\frac{1-t\phi}{\phi}\int_0^\infty \exp\left\{-\frac{x(1-t\phi)}{\phi}\right\} = \frac{1}{1-t\phi}\,. \end{aligned}$$

Hence, by the properties of the cumulant generating function $K_{Y_n}(t) = -n\log(1-t\phi)$ which is the cumulant generating function of $G(n, \phi)$ (to see this use the results in Table 2.1 to derive the cumulant generating function for the Gamma distribution).

Note that the mean and variance of Y_n are $\mu = n\phi$ and $\sigma^2 = n\phi^2$. Thus, according to the central limit theorem, $Z_n = (Y_n - n\phi)/(\sqrt{n}\phi) \xrightarrow{d} N(0,1)$. This can also be verified by direct use the of the cumulant generating function. Because Z_n results by changing the location and scale of Y_n,

$$K_{Z_n}(t) = -t\sqrt{n} - n\log\left(1 - \frac{t}{\sqrt{n}}\right)\,.$$

Using (2.6), we can expand $\log(1 - t\sqrt{n})$ to get

$$K_{Z_n}(t) = \frac{t^2}{2} + \frac{t^3}{3n^{1/2}} + \frac{t^4}{4n} + \ldots = \frac{t^2}{2} + O\left(n^{-1/2}\right)\,.$$

Hence, $\lim_{n\to\infty} K_{Z_n}(t) = t^2/2$ which is the cumulant generating function of a $N(0,1)$ random variable. □

2.3.4 Multivariate extensions

The above definitions and results can be extended to the case of a random d-vector $Y = (Y_1, \ldots, Y_d)$, $d > 1$. A generic joint moment of order r of Y is given by

$$\mu_{i_1,\ldots,i_r} = E(Y_{i_1}\ldots Y_{i_r}) \quad (i_1, \ldots, i_r = 1, \ldots, d)\,,$$

and the generic joint central moment of order r is

$$\bar{\mu}_{i_1,\ldots,i_r} = E\{(Y_{i_1} - \mu_{i_1})\ldots(Y_{i_r} - \mu_{i_r})\} \quad (i_1, \ldots, i_r = 1, \ldots, d)\,,$$

The moment generating function of Y is defined as

$$M_Y(t) = E\{\exp(t \cdot Y)\} = E\{\exp(t_1 Y_1 + \ldots + t_d Y_d)\}\,, \quad t = (t_1, \ldots, t_d) \in \Re^d\,,$$

and is said to exist if $M_Y(t)$ is finite for $||t|| < t_o$, $t_0 > 0$. In that case the moment generating function can be expanded using (2.3) into a multivariate power series with some convergence radius $R \geq t_0$ and then as in the univariate case

$$\mu_{i_1,\ldots,i_r} = \left.\frac{\partial^r M_Y(t)}{\partial t_{i_1} \ldots \partial t_{i_r}}\right|_{t=0} \qquad (i_1, \ldots, i_r = 1, \ldots, d)\,.$$

If $M_Y(t)$ exists, the cumulant generating function is again defined as $K_Y(y) = \log M_Y(t)$ and it can be expanded in multivariate power series in a neighbourhood of the origin. The coefficients of this expansion define the cumulants of Y. The generic joint cumulant of order r is

$$\kappa_{i_1,\ldots,i_r} = \left.\frac{\partial^r K_Y(t)}{\partial t_{i_1} \ldots \partial t_{i_r}}\right|_{t=0} \qquad (i_1, \ldots, i_r = 1, \ldots, d)\,.$$

Given that the moment generating function exists, both $M_Y(t)$ and $K_Y(t)$ characterize the, multivariate in this case, distribution of Y. Furthermore, note that the notation we used for moments and cumulants in the multivariate case is not consistent with the notation in the univariate case. Nevertheless, the multivariate notation will be rather useful in later chapters.

The relations in (2.9) can be extended to refer to joint cumulants, joint moments and joint central moments. In particular, for $i, j, k, l = 1, \ldots, d$,

$$\begin{aligned} \kappa_{i,j} &= \mu_{i,j} - \mu_i \mu_j = \bar{\mu}_{i,j}\,, \\ \kappa_{i,j,k} &= \mu_{i,j,k} - \mu_i \mu_{j,k}[3] + 2\mu_i \mu_j \mu_k = \bar{\mu}_{i,j,k}\,, \\ \kappa_{i,j,k,l} &= \mu_{i,j,k,l} - \mu_i \mu_{j,k,l}[4] - \mu_{i.j}\mu_{k,l}[3] + 2\mu_i \mu_j \mu_{k,l}[6] - 6\mu_i \mu_j \mu_k \mu_l = \bar{\mu}_{i,j,k,l} - \bar{\mu}_{i,j}\bar{\mu}_{k,l}[3]\,, \end{aligned} \tag{2.10}$$

where the symbol $[c]$ denotes the sum of all c distinct terms obtained by permutations of indices between the factors involved in the product. For example,

$$\mu_i \mu_{j,k,l}[4] = \mu_i \mu_{j,k,l} + \mu_j \mu_{i,k,l} + \mu_k \mu_{i,j,l} + \mu_l \mu_{i,j,k}\,.$$

Example 2.7. (Exponential families) Consider a random q-vector Y with density (or probability mass) function of the form

$$\exp\left\{\sum_{i=1}^d s_i(y)\theta_i(\beta) - k(\theta(\beta)) + a(y)\right\}\,, \quad \beta \in B \subseteq \Re^p\,, \tag{2.11}$$

where $s(y) = (s_1(y), \ldots, s_d(y))$ is a d-vector of real-valued functions of y called the *natural statistics*, $\theta(\beta) = (\theta_1(\beta), \ldots, \theta_d(\beta))$ is a d-vector of real-valued functions of β called the *natural parameters*, and $a(.)$ and $k(.)$ are real-valued functions of d-vectors and q-vectors, respectively. Note that the value of d can be further reduced if either $s(y)$ or $\theta(\beta)$ satisfies a linear constraint, so we assume that d is as small is possible. If $\dim(B) = d$ then the family is called *full* while if $\dim(B) < d$ the family is called *curved*. The case $\dim(B) > d$ is not interesting, as then the parameter β is not identifiable. The one-parameter natural exponential family results for $d = q = p = 1$ and $s(y) = y$.

For simplicity, consider the case of discrete Y. Then, for some $x = (x_1, \ldots, x_d)$, if $T_x = \{y : s_1(y) = x_1, \ldots, s_d(y) = x_d\}$,

$$\begin{aligned} P(s_1(Y) = x_1, \ldots, s_d(Y) = x_d) &= \sum_{y \in T_x} \exp\left\{\sum_{i=1}^d s_i(y)\theta_i - k(\theta) + a(y)\right\} \\ &= \exp\left\{\sum_{i=1}^d x_i\theta_i - k(\theta) + b(x)\right\} , \end{aligned}$$

where $b(x) = \log \sum_{y \in T_x} \exp\{a(y)\}$. This is again of the form (2.11) with natural statistics x_i and and natural parameters θ_i $(i = 1, \ldots, d)$, respectively. The above expression remains unchanged for continuous Y. Now, the same argument as in Example 2.5, but now in d-dimensions gives $K(t) = k(\theta + t) - k(\theta)$. Thus,

$$\kappa_{i_1,\ldots,i_r}(S) = \frac{\partial^r k(\theta)}{\partial\theta_{i_1} \ldots \partial\theta_{i_r}} \quad (i_1, \ldots, i_r = 1, \ldots, d) .$$

In this way, we can access features of the joint distribution of the natural statistics without having to explicitly derive it. For example, the Beta density

$$f(y; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1 - y)^{\beta-1} \quad 0 < y < 1 \, , \alpha > 0 \, , \beta > 0 \, ,$$

is of the exponential form with $\theta_1 = \alpha$, $\theta_2 = \beta$, $s_1(y) = \log y$, $s_2(y) = \log(1 - y)$ and $k(\theta) = \log \Gamma(\theta_1) + \log \Gamma(\theta_1) - \log \Gamma(\theta_1 + \theta_2)$. Thus, $E(\log Y) = \partial k(\theta)/\partial\theta_1 = \psi(\theta_1) - \psi(\theta_1 + \theta_2)$ and

$$\mathrm{cov}\{\log Y, \log(1 - Y)\} = \frac{\partial^2 k(\theta)}{\partial\theta_1\partial\theta_2} = -\psi'(\theta_1 + \theta_2) \, ,$$

where $\psi(z) = \mathrm{d}\log\Gamma(z)/\mathrm{d}z$ and $\psi'(z) = \mathrm{d}\psi(z)/\mathrm{d}z$ are the digamma and trigamma functions, respectively.

For a Normal density with unknown mean μ and variance σ^2, $s_1(y) = y$, $s_2(y) = -y^2/2$, $\theta_1 = \mu/\sigma^2$, $\theta_2 = 1/\sigma^2$ and $k(\theta) = -\log\theta_2/2 + \theta_1^2/(2\theta_2)$. Thus, for example,

$$\mathrm{cov}\left(Y, -\frac{Y^2}{2}\right) = \frac{\partial^2 k(\theta)}{\partial\theta_1\partial\theta_2} = -\frac{\theta_1}{\theta_2^2} = -\mu\sigma^2 \, .$$

□

2.4 Asymptotic expansions and their inversion

2.4.1 The general form of an asymptotic expansion

Assume that interest lies in a sequence of functions $f_1(x), f_2(x), \ldots$ and that the general term $f_n(x)$ of this sequence can be written as

$$f_n(x) = \gamma_0(x)c_{0,n} + \gamma_1(x)c_{1,n} + \gamma_2(x)c_{2,n} + \ldots + \gamma_k(x)c_{k,n} + O\left(c_{k+1,n}\right) , \tag{2.12}$$

where $\{c_{0,n}, c_{1,n}, \ldots\}$ is a sequence of real constants, such as $\{1, n^{-1/2}, n^{-1}, \ldots\}$ or $\{1, n^{-1}, n^{-2}, \ldots\}$, and $\{\gamma_0(x), \ldots \gamma_k(x)\}$ is a sequence of functions of x with terms not depending on n. Note that an essential condition for (2.12) to be valid, is that $c_{r+1,n} = o(c_{r,n})$ as $n \to \infty$ $(r = 1, 2, \ldots)$, because otherwise some terms in the sum might have to be absorbed in the remainder $O(c_{k+1,n})$. Expression (2.12) for $f_n(x)$ is called an *asymptotic expansion* for $f_n(x)$.

Another type of expansion is the *stochastic asymptotic expansion*. For a sequence of random variables $\{Y_n\}$, a stochastic asymptotic expansion for Y_n is expressed as

$$Y_n = X_0 c_{0,n} + X_1 c_{1,n} + X_2 c_{2,n} + \ldots + X_k c_{k,n} + O_p(c_{k+1,n}) \,, \tag{2.13}$$

where $\{c_{0,n}, c_{1,n}, \ldots\}$ are as before and $\{X_0, X_1, \ldots\}$ are random variables having distributions not depending on n.

Example 2.8. (Haldane-Anscombe correction) Consider an observation y from a binomial random variable $Y \sim B(n,p)$. The obvious estimator (which is also the maximum likelihood estimator) of the log-odds $\beta = \log\{p/(1-p)\}$ is $\hat\beta = \log\{Y/(n-Y)\}$. Nevertheless, $\hat\beta$ has the disadvantage that it is infinite if $y = 0$ or $y = n$ is observed. One possible way to overcome this is to define the new estimator (Haldane, 1955; Anscombe, 1956)

$$\tilde\beta_a = \log\frac{Y+a}{n-Y+a}\,, \quad a \in \Re\,, \tag{2.14}$$

essentially adding a and $2a$ to y and n, respectively, and then replacing y and n in the expression for $\hat\beta$. An appropriate value for a can be chosen so that the expectation of $\tilde\beta_a$ is as close as possible to β.

Firstly, note that Y can be expressed as the sum of n independent Bernoulli random variables all with probability of success p and that $E(Y) = np$, $\mathrm{var}(Y) = np(1-p)$ and $E\{(Y-np)^3\} = np(1-p)(1-2p)$. Hence, if $U = (Y-np)/\sqrt{n}$ then by the central limit theorem $U/\sqrt{p(1-p)} \xrightarrow{d} N(0,1)$ which implies $U = O_p(1)$. Replacing $Y = np + U\sqrt{n}$ in (2.14) and subtracting β from both sides, simple algebra gives

$$\tilde\beta_a - \beta = \log\left[1 + \left\{\frac{U}{\sqrt{np}} + \frac{a}{np}\right\}\right] - \log\left[1 - \left\{\frac{U}{\sqrt{nq}} - \frac{a}{nq}\right\}\right]\,,$$

where $q = 1-p$. The first and the second terms of the right hand side of the above expression can be expanded using to (2.5) and (2.6), respectively, up to order $O_p(n^{-2})$. This gives the stochastic asymptotic expansion

$$\begin{aligned}\tilde\beta_a - \beta = &\frac{U}{\sqrt{np}} + \frac{a}{np} - \frac{U^2}{2np^2} - \frac{aU}{n^{3/2}p^2} + \frac{U^3}{3n^{3/2}p^3} \\ &+ \frac{U}{\sqrt{nq}} - \frac{a}{nq} + \frac{U^2}{2nq^2} - \frac{aU}{n^{3/2}q^2} + \frac{U^3}{3n^{3/2}q^3} + O_p(n^{-2})\,.\end{aligned}$$

Taking expectations in both sides and noting that $E(U) = 0$, $E(U^2) = pq$ and $E(U^3) = n^{-1/2}pq(1-2p)$ gives the asymptotic expansion

$$E(\tilde\beta_a - \beta) = \frac{(1-2p)(2a-1)}{2npq} + O(n^{-2})\,.$$

Thus for $a = 1/2$, $\tilde\beta_{1/2}$ has bias of order $O(n^{-2})$, having the additional property of being finite for every value of y. □

In this example the asymptotic orders were apparent because powers of n were explicitly involved in each term. This need not be the case in general, and the explicit involvement on n might be dropped, as done in the next chapter, for the sake of convenience; that is we will be using random variables with distributions depending on n and moments and cumulants of those. Furthermore, the expansions given in the previous example do not necessarily refer to convergent series because (2.5) and (2.6) converge only for for $-1 \le x < 1$ and $-1 < x \le 1$, respectively. In general, taking more terms in an asymptotic expansion and keeping n finite will not necessarily improve the approximation.

2.4.2 Inversion of asymptotic series

In most cases, while we are interested in an asymptotic expansion for a quantity x, the asymptotic expansion of a function $f(x)$ is much more convenient to obtain. Hence, a formal device is required for inverting the expansion for $f(x)$ so as to obtain an expansion for x.

For the sake of simplicity, let $y = f(x)$, $x \in \Re$ be the generic component of a sequence of real smooth functions which admit a power series expansion

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \dots \,,$$

where $a_r = O(1)$, $r = 0, 1, \dots$ and $x = O(n^{-t})$, $t > 0$. This is an asymptotic expansion for y with $\gamma_r(x) = a_r n^{rt} x^r$ and $c_{r,n} = n^{-rt}$ in (2.12) ($r = 0, \dots, n$). Define a new variable $z = (y - a_0)/a_1$ ($a_1 \neq 0$). Then the power series expansion can be written as

$$z = x + b_2 x^2 + b_3 x^3 + \dots \,, \tag{2.15}$$

where $b_i = a_i/a_1$, $i = 2, 3, \dots$.

Note that, as a first approximation $x = z + O\left(n^{-2t}\right)$. Because $f'(0) = a_1 \neq 0$, the function $z = (f(x) - a_0)/a_1$ can be inverted in a neighbourhood of $x = 0$, with inverse $x = g(z)$. We wish to express $g(z)$, in a neighbourhood of $z = 0$, as a power series in the form

$$x = z + d_2 z^2 + d_3 z^3 + \dots \,,$$

with d_2 and d_3 to be determined as functions of b_2 and b_3 ignoring terms of order $O\left(n^{-4t}\right)$.

This can be done by a procedure called the *iterative substitution method*. As the name suggests, the iterative substitution method proceeds by writing (2.15) as

$$x = z - b_2 x^2 - b_3 x^3 + \dots \,, \tag{2.16}$$

and then iteratively substituting all instances of x on the right hand side, according to its expression in (2.16), until the only terms that cannot be neglected are functions only of z. The first substitution gives

$$\begin{aligned} x &= z - b_2 (z - b_2 x^2)^2 - b_3 z^3 + O\left(n^{-4t}\right) \\ &= z - b_2 z^2 + 2 b_2^2 x^2 z - b_3 z^3 + O\left(n^{-4t}\right) \,. \end{aligned}$$

A second step is required, where x^2 is substituted by z^2 on the right hand side of the above expression. This gives

$$\begin{aligned} x &= z - b_2 z^2 + 2 b_2^2 z^3 - b_3 z^3 + O\left(n^{-4t}\right) \\ &= z - b_2 z^2 + (2 b_2^2 - b_3) z^3 + O\left(n^{-4t}\right) \,, \end{aligned}$$

which is the desired form with $d_2 = -b_2$ and $d_3 = 2b_2^2 - b_3$.

Chapter 3

Likelihood-based asymptotics

Maximum likelihood is one of the central principles for estimation in Statistics mainly because of the neat asymptotic properties of the maximum likelihood estimator (consistency, asymptotic normality and asymptotic efficiency). In this chapter we briefly show how these properties result using the basic tools of the previous chapter. Asymptotic expansions for the moments of the maximum likelihood estimator are derived whose form suggests the possibility of simple corrections to the maximum likelihood estimator. It is also shown how the approximate pivots of Chapter 1 result. Special attention is paid to exponential family models. For simplicity, the derivations are made only for one-parameter models but the extension of some of the results to the case of multidimensional parameters is also given.

3.1 Maximum likelihood estimation

Suppose that y is the observed value of a random n-vector $Y = (Y_1, \ldots, Y_n)$ from a parametric family of distributions with density $f_Y(y; \beta)$, where the parameter β is generally multidimensional, $\beta = (\beta_1, \ldots, \beta_p) \in B \subseteq \Re^p$, for some $p \geq 1$. The *likelihood function* is defined by

$$L(\beta; y) = f_Y(y; \beta)\,,$$

and is viewed as a function of β for the fixed data y. The *maximum likelihood estimate* is the value of β for which $L(\beta; y)$ is maximized or, more conveniently, where the *log-likelihood* $l(\beta; y) = \log L(\beta; y)$ is maximized.

In most cases $l(\beta; y)$ is differentiable and hence the maximum can be located by solving the *likelihood equations*,

$$\nabla l(\beta; y) = 0\,,$$

where $\nabla l(\beta; y) = (\partial l(\beta; y)/\partial\beta_1, \ldots, \partial l(\beta; y)/\partial\beta_p)^T$ and then checking negative definiteness of the matrix of second derivatives evaluated at the solution to establish that a maximum has been located.

We mainly focus on the case where $y_1, \ldots, y_n$ are observations on independent and identically distributed random variables $Y_1, \ldots, Y_n$ but the following results extend far beyond this case (using appropriate extensions of the definitions and theorems of the previous chapter). In the case of independent and identically distributed random variables, $l(\beta; y) = \sum_{i=1}^n \log f(y_i; \beta)$, where $f(y_i; \beta)$ is the density of Y_i.

3.2 Log-likelihood related quantities

3.2.1 Log-likelihood derivatives and Bartlett relations

The building blocks of asymptotic arguments in likelihood theory are the log-likelihood derivatives and moments and cumulants of their products. In the current section we introduce notation for the aforementioned quantities in the case of models with a scalar parameter β. Assume that the log-likelihood is infinitely differentiable and denote the rth log-likelihood derivative by

$$l_r(\beta) \equiv l_r(\beta; Y) = \frac{\mathrm{d}l(\beta; Y)}{\mathrm{d}\beta} \quad (r = 1, 2, \ldots) .$$

We call $l_1(\beta)$ the efficient score function (or simply score function). Furthermore, denote by

$$\begin{aligned} \nu_r(\beta) &= E_\beta\{l_r(\beta)\} , \\ \nu_{r,s}(\beta) &= E_\beta\{l_r(\beta) l_s(\beta)\} , \\ \nu_{r,s,t}(\beta) &= E_\beta\{l_r(\beta) l_s(\beta) l_t(\beta)\} , \\ &\vdots \end{aligned}$$

the joint null moments of the log-likelihood derivatives (or more precisely, the joint null moments of the random variable $U = (l_1, l_2, \ldots, l_d)$, for appropriate $d > 1$), where the word null is used to indicate that the operations of differentiation and expectation take place at the same value β of the parameter.

By definition,

$$\int_{\mathcal{Y}} f_Y(y; \beta) dy = 1 , \tag{3.1}$$

and hence, assuming that the parameter space does not depend on the sample space, we can differentiate both sides of (3.1) with respect to β and exchange the order of differentiation and integration over the sample space $\mathcal{Y}$ (with integration replaced by summation in the case of discrete random variables). Hence,

$$\begin{aligned} 0 &= \frac{\mathrm{d}}{\mathrm{d}\beta} \int_{\mathcal{Y}} f_Y(y; \beta) dy \\ &= \int_{\mathcal{Y}} \frac{\mathrm{d} f_Y(y; \beta)}{\mathrm{d}\beta} \frac{1}{f_Y(y; \beta)} f_Y(y; \beta) dy \\ &= \int_{\mathcal{Y}} \frac{\mathrm{dlog}\, f_Y(y; \beta)}{\mathrm{d}\beta} f_Y(y; \beta) dy = E_\beta\{l_1(\beta)\} = \nu_1(\beta) . \end{aligned}$$

Further differentiation gives

$$\begin{gathered} \nu_{1,1}(\beta) + \nu_2(\beta) = 0 , \\ \nu_3(\beta) + 3\nu_{2,1}(\beta) + \nu_{1,1,1}(\beta) = 0 , \\ \nu_4(\beta) + 3\nu_{2,2}(\beta) + 4\nu_{1,3}(\beta) + 6\nu_{1,1,2}(\beta) + \nu_{1,1,1,1}(\beta) = 0 , \\ \vdots \end{gathered} \tag{3.2}$$

The above equations are called *Bartlett relations* (Bartlett, 1953, Section 2). There is a simple rule for direct differentiation of moments of log-likelihood derivatives (Skovgaard, 1986).

Theorem 3.1. (Skovgaard)

$$\frac{\mathrm{d}}{\mathrm{d}\beta}\nu_{r_1,\ldots,r_d}(\beta) = \sum_{j=1}^{d} E_\beta\left\{l_{r_1}(\beta)\ldots l_{r_j+1}(\beta)\ldots l_{r_d}(\beta)\right\} + E_\beta\left\{l_{r_1}(\beta)\ldots l_{r_d}(\beta)l_1(\beta)\right\} .$$

For example, $\mathrm{d}\nu_{1,3,6}(\beta)/\mathrm{d}\beta = \nu_{2,3,6}(\beta) + \nu_{1,4,6}(\beta) + \nu_{1,3,7}(\beta) + \nu_{1,1,3,6}(\beta)$. Hence, using Skovgaard's theorem, each Bartlett relation is obtained by the previous one by direct differentiation (exercise with the first three). Skovgaard's theorem is also applicable for the differentiation of joint null cumulants of log-likelihood derivatives.

The first identity in (3.2) is usually referred to as the information identity and shows that the variance of the score function is equal to the expectation of minus the second log-likelihood derivative. The name occurs because the quantity $i(\beta) = \nu_{1,1}(\beta)$ is called the expected (or Fisher) information matrix for β and the quantity $j(\beta) = -l_2(\beta)$ is called the observed information for β.

3.2.2 Asymptotic orders

Below, it will be essential to be able to identify the asymptotic order of likelihood related quantities in a systematic way. Firstly, note that $l_r(\beta)$ is the sum of n independent $O_p(1)$ terms and hence $l_r(\beta) = O_p(n)$ and $\nu_r(\beta) = O(n)$ $(r = 1, 2, \ldots)$. In particular for the score function a more refined result can be obtained because $\nu_1(\beta) = 0$ and thus by the central limit theorem $l_1(\beta)/\sqrt{i(\beta)} \xrightarrow{d} N(0,1)$ resulting $l_1(\beta) = O_p\left(n^{1/2}\right)$. Furthermore, all joint null cumulants of log-likelihood derivatives will also be of order $O(n)$. However, the order of moments and central moments is not so straightforward to obtain.

Define the centered log-likelihood derivatives $H_r(\beta) = l_r(\beta) - \nu_r(\beta)$ $(r = 1, 2, \ldots)$. A similar argument as for the score function gives $H_r(\beta) = O_p\left(n^{1/2}\right)$. Furthermore, define the joint null central moments as $\bar\nu_{r,s}(\beta) = E_\beta\{H_r(\beta)H_s(\beta)\}$, $\bar\nu_{r,s,t}(\beta) = E_\beta\{H_r(\beta)H_s(\beta)H_t(\beta)\}$, and so on. Then[1],

$$\bar\nu_{r_1,\ldots,r_d}(\beta) = \begin{cases} O\left(n^{d/2}\right) & \text{if d is even} \\ O\left(n^{(d-1)/2}\right) & \text{if d is odd} \end{cases} \qquad (r_1, \ldots, r_d = 1, 2, \ldots)\,. \tag{3.3}$$

Example 3.1. Consider the joint null central moment $\bar\nu_{r,s,t,u}(\beta) = E_\beta\left\{H_r(\beta)H_s(\beta)H_t(\beta)H_u(\beta)\right\}$ $(r, s, t, u = 1, 2, \ldots)$. A direct application of the rule in (3.3) gives $\bar\nu_{r,s,t,u}(\beta) = O(n^2)$. Assume we want to discard any $O(n)$ terms of $\bar\nu_{r,s,t,u}(\beta)$ keeping only the $O\left(n^2\right)$ terms. On the basis of (2.10) we get

$$\kappa_{r,s,t,u}(\beta) = \bar\nu_{t,s,t,u}(\beta) - \bar\nu_{r,s}(\beta)\bar\nu_{t,u}(\beta) - \bar\nu_{r,t}(\beta)\bar\nu_{s,u}(\beta) - \bar\nu_{r,u}(\beta)\bar\nu_{s,t}(\beta)\,,$$

and because $\kappa_{r,s,t,u} = O(n)$,

$$\bar\nu_{r,s,t,u}(\beta) = \bar\nu_{r,s}(\beta)\bar\nu_{t,u}(\beta) + \bar\nu_{r,t}(\beta)\bar\nu_{s,u}(\beta) + \bar\nu_{r,u}(\beta)\bar\nu_{s,t}(\beta) + O(n)\,.$$

For simplicity, we temporarily omit the argument β of the functions in the expressions. Noting that $\bar\nu_{r,s} = \nu_{r,s} - \nu_r\nu_s$ (the covariance of l_r and l_s) gives

$$\begin{aligned}\bar\nu_{r,s,t,u} &= \nu_{r,s}\nu_{t,u} + \nu_{r,t}\nu_{s,u} + \nu_{r,u}\nu_{s,t} \\ &\quad - \nu_{r,s}\nu_t\nu_u - \nu_{r,t}\nu_s\nu_u - \nu_{r,u}\nu_s\nu_t - \nu_{s,t}\nu_r\nu_u - \nu_{s,u}\nu_r\nu_t - \nu_{r,u}\nu_s\nu_t \\ &\quad + 3\nu_r\nu_s\nu_t\nu_u + O(n)\ .\end{aligned}$$

[1]This result is a direct consequence of the *exlog relations* (see, for example, Barndorff-Nielsen and Cox 1989, section 5.4 and Pace and Salvan 1997, section 9.2.2).

The above expression gives an evaluation of the central moment $\bar{\nu}_{r,s,t,u}(\beta)$ in terms of joint null moments of orders up to 2, ignoring contributions of order $O(n)$.

For example, noting that $\nu_1 = 0$, we get

$$\begin{aligned} \bar{\nu}_{1,1,1,r}(\beta) &= 3\nu_{1,1}(\beta)\nu_{1,r}(\beta) + O(n) \quad (r = 1, 2, \ldots)\,, \\ \bar{\nu}_{1,1,r,r}(\beta) &= 2\nu_{1,r}(\beta)^2 + \nu_{1,1}(\beta)\{\nu_{r,r}(\beta) - \nu_r(\beta)^2\} + O(n) \quad (r = 2, 3, \ldots)\,. \end{aligned} \tag{3.4}$$

□

3.3 Properties of the maximum likelihood estimator

3.3.1 Consistency

Let $\hat{\beta}$ be the maximum likelihood estimator of the parameter β based on a random sample of size n and denote by β_0 the true but unknown value of β. Then, under fairly general conditions, it can be shown that $\hat{\beta} \xrightarrow{p} \beta_0$, that is $\hat{\beta}$ is a consistent (or asymptotically consistent) estimator. The result follows by the Jensen's inequality (that is, for concave $g : \Re \to \Re$, $E\{g(X)\} \leq g\{E(X)\}$) and the defining property of the maximum likelihood estimator. More specifically, if $\mathcal{Y}$ is the sample space and β is any point of B, by Jensen's inequality,

$$\begin{aligned} E_{\beta_0}\{l(\beta; Y) - l(\beta_0; Y)\} &= \int_{\mathcal{Y}} \left\{\log \frac{f_Y(y;\beta)}{f_Y(y;\beta_0)}\right\} f_Y(y,\beta_0)dy \\ &\leq \log \int_{\mathcal{Y}} \frac{f_Y(y;\beta)}{f_Y(y;\beta_0)} f_Y(y,\beta_0)dy = 0\,. \end{aligned}$$

This implies $E_{\beta_0}\{l(\beta;Y)\} \leq E_{\beta_0}\{l(\beta_0;Y)\}$ where the inequality is strict unless $f(y;\beta)/f(y;\beta_0) = 1$ almost everywhere in $\mathcal{Y}$. On the other hand, by definition, $n^{-1}l(\hat{\beta};y) \geq n^{-1}l(\beta;y)$ for any $\beta \in \beta$. Heuristically, the two inequalities are incompatible unless $\hat{\beta}$ converges to β_0.

A rigorous proof of the consistency of the maximum likelihood estimator is given in Cox and Hinkley (1974, Section 9.2) and is due to Wald (1949). The interested reader can also see van der Vaart (1998, Section 5.2) for an alternative consistency proof in the more general context of M-estimation. For our purposes, we shall merely note that because $\hat{\beta} \xrightarrow{p} \beta_0$, $\hat{\beta} - \beta_0 = o_p(1)$.

3.3.2 Asymptotic normality

For simplicity, we only consider the case where β_0 is scalar. Furthermore we assume that the log-likelihood is twice continuously differentiable on a neighbourhood of β_0 and that $\left|\mathrm{d}^3 \log f(y_i;\beta)/\mathrm{d}\beta^3\right| \leq g(y_i)$ uniformly for some β in the neighbourhood of β_0, with $E_{\beta_0}\{g(Y_i)\} < \infty$ $(i = 1, \ldots, n)$.

By Taylor's theorem (Theorem 2.1), the first derivative of the log-likelihood can be written as

$$0 = l_1(\hat{\beta}) = l_1(\beta_0) + (\hat{\beta} - \beta_0)l_2(\alpha)\,,$$

where α is between $\hat{\beta}$ and β_0. Thus, $\hat{\beta} - \beta_0 = -l_1(\beta_0)/l_2(\alpha)$. Multiplying and dividing by $\sqrt{i(\beta_0)}$ gives

$$\sqrt{i(\beta_0)}(\hat{\beta} - \beta_0) = \frac{A(\beta_0)C(\beta_0)}{B(\beta_0)}\,, \tag{3.5}$$

where $A(\beta) = l_1(\beta)/\sqrt{i(\beta)}$, $B(\beta) = -l_2(\beta)/i(\beta)$ and $C(\beta) = l_2(\beta)/l_2(\alpha)$. By the central limit theorem and the law of large numbers,

$$A(\beta_0) \xrightarrow{d} N(0,1)\,, \tag{3.6}$$

$$B(\beta_0) \xrightarrow{p} 1\,, \tag{3.7}$$

respectively (write $i(\beta) = ni^*(\beta)$ to see those, where $i^*(\beta)$ is called the Fisher information per observation and is independent of n). Now, by assumption the third derivative of the log-likelihood is bounded and a simple Taylor expansion of $l_2(\alpha)$ around β_0 gives

$$\left|\frac{l_2(\alpha) - l_2(\beta_0)}{n}\right| \leq |\alpha - \beta_0| \frac{\sum_{j=1}^n g(Y_j)}{n}\,. \tag{3.8}$$

The second factor on the right hand side tends to some finite constant, while the first tends to 0 in probability by the consistency of $\hat{\beta}$ and because $|\alpha - \beta_0| < |\hat{\beta} - \beta_0|$. Thus the left hand side of (3.8) tends in probability to 0 and it follows that

$$C(\beta_0) = \frac{l_2(\alpha) - l_2(\beta_0)}{n} \cdot \left\{\frac{l_2(\beta_0)}{n}\right\}^{-1} + 1 \xrightarrow{p} 0 \cdot \left\{-\frac{1}{i^*(\beta_0)}\right\} + 1 = 1\,. \tag{3.9}$$

By (3.6), (3.7) and (3.9), an application of Slutzky's lemma to (3.5) gives that

$$\sqrt{i(\beta_0)}(\hat{\beta} - \beta_0) \sim N(0,1)\,. \tag{3.10}$$

Expression (3.10) implies that $\sqrt{n}(\hat{\beta} - \beta_0) = O_p(1)$ and thus we can refine the statement $\hat{\beta} - \beta_0 = o_p(1)$ to

$$\hat{\beta} - \beta_0 = O_p(n^{-1/2})\,.$$

Furthermore, it can be shown that $i(\beta_0)$ can be replaced by either $i(\hat{\beta})$, $j(\beta_0)$ or $j(\hat{\beta})$ in (3.10) without affecting the limiting distribution. Lastly, for a p dimensional parameter β the same reasoning applies and the result is $\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N_p(0, \{i^*(\beta_0)\}^{-1})$, where $i^*(\beta_0)$ is an $p \times p$ matrix with (j,k)th entry $n^{-1}E(-\partial^2 l(\beta)/\partial\beta_j\partial\beta_k)$. It is part of the assumption that $i^*(\beta_0)$ is invertible.

3.3.3 Asymptotic efficiency

Let $\hat{\beta} \equiv \hat{\beta}(Y)$ be *any* estimator of β which is a sufficiently smooth function of the data and let $m(\beta) = E_\beta\{\hat{\beta}(Y)\}$. Because the correlation of two random variables is strictly between -1 and 1,

$$\left[\mathrm{cov}_\beta\left\{\hat{\beta}(Y), l_1(\beta)\right\}\right]^2 \leq i(\beta)\mathrm{var}_\beta\left\{\hat{\beta}(Y)\right\}\,. \tag{3.11}$$

But under the usual regularity conditions,

$$\mathrm{cov}_\beta\left\{\hat{\beta}(Y), l_1(\beta)\right\} = \int_{\mathcal{Y}} \hat{\beta}(y)\left\{\frac{\mathrm{d}}{\mathrm{d}\beta}\log f(y;\beta)\right\} f(y;\beta)dy = m'(\beta)\,.$$

Hence, by (3.11),

$$\mathrm{var}_\beta\left\{\hat{\beta}(Y)\right\} \geq \frac{\{m'(\beta)\}^2}{i(\beta)}\,. \tag{3.12}$$

The above inequality is known as the *Cramér-Rao lower bound*. In the case of *unbiased* estimators (that is $m(\beta) = \beta$) the Cramér-Rao lower bound takes the simpler form

$$\text{var}_\beta \left\{ \hat{\beta}(Y) \right\} \geq \frac{1}{i(\beta)} \, .$$

Thus any unbiased estimator that achieves the above lower bound is immediately seen to be a *minimum variance unbiased estimator*.

Now, by (3.10) the maximum likelihood estimator $\hat{\beta}$ is asymptotically normally distributed with mean β_0 and variance $1/i(\beta_0)$. Hence, even though there might not be an estimator that achieves the Cramér-Rao lower bound for finite n, the maximum likelihood estimator achieves it as $n \to \infty$. In this sense the maximum likelihood estimator is *asymptotically efficient* (or, more accurately, *first-order efficient*).

3.4 Fundamental asymptotic expansions

3.4.1 Expansion of $\hat{\beta} - \beta_0$

Assume that the log-likelihood is four times continuously differentiable and that the fifth derivative exists in some neighbourhood of the true value β_0.

Having established that $\hat{\beta} - \beta_0 = O_p\left(n^{-1/2}\right)$ and how the asymptotic orders of likelihood-related quantities can be obtained, an application of Taylor's theorem leads to the following stochastic asymptotic expansion

$$0 = l_1(\hat{\beta}) = l_1(\beta_0) + (\hat{\beta} - \beta_0) l_2(\beta_0) + \frac{1}{2}(\hat{\beta} - \beta_0)^2 l_3(\beta_0) + \frac{1}{6}(\hat{\beta} - \beta_0)^3 l_4(\beta_0) + O_p\left(n^{-1}\right) . \quad (3.13)$$

For notational simplicity, write $\delta = \hat{\beta} - \beta_0$ and suppress the argument whenever a function is evaluated at β_0, that is $l_r \equiv l_r(\beta_0)$, $H_r \equiv H_r(\beta_0)$, $\nu_{r,s} \equiv \nu_{r,s}(\beta_0)$, etc. $(r, s = 1, 2, \ldots)$. Then (3.13) can be written as

$$0 = l_1 + \delta l_2 + \frac{1}{2}\delta^2 l_3 + \frac{1}{6}\delta^3 l_4 + O_p\left(n^{-1}\right) . \quad (3.14)$$

Now, each l_r is the sum of a $O_p\left(n^{1/2}\right)$ stochastic term H_r and a $O(n)$ deterministic term ν_r (see Subsection 3.2.2). Substituting in (3.15) leads to

$$0 = l_1 + \delta\nu_2 \dot{+} \delta H_2 + \frac{1}{2}\delta^2\nu_3 \dot{+} \frac{1}{2}\delta^2 H_3 + \frac{1}{6}\delta^3\nu_4 \dot{+} O_p\left(n^{-1}\right) , \quad (3.15)$$

where the symbol $\dot{+}$ denotes a drop in asymptotic order by $n^{-1/2}$ and $\delta^3 H_4/6 = O_p\left(n^{-1/2}\right)$ and hence absorbed by the remainder. Noting that $-\nu_2 = i = O(n)$, with $i \equiv i(\beta_0)$, and after some algebra we get

$$\delta = \frac{l_1}{i} \dot{+} \delta\frac{H_2}{i} + \delta^2\frac{\nu_3}{2i} \dot{+} \delta^2\frac{H_3}{2i} + \delta^3\frac{\nu_4}{6i} + O_p(n^{-2}) . \quad (3.16)$$

The above expression is the implicit form of the expansion sought and allows direct application of the iterative substitution method. Noting that

$$\begin{aligned} \delta &= \frac{l_1}{i} \dot{+} \delta\frac{H_2}{i} + \delta^2\frac{\nu_3}{2i} \dot{+} O_p(n^{-3/2}) , \\ \delta^2 &= \frac{l_1^2}{i^2} \dot{+} 2\delta\frac{H_2 l_1}{i^2} + \delta^2\frac{\nu_3 l_1}{i^2} \dot{+} O_p(n^{-2}) \quad \text{and} \\ \delta^3 &= \frac{l_1^3}{i^3} \dot{+} O_p(n^{-2}) , \end{aligned}$$

two steps of the iterative substitution method on (3.16) and some rearrangement of terms leads to

$$\delta = \frac{l_1}{i} \dot{+} \frac{H_2 l_1}{i^2} + \frac{\nu_3 l_1^2}{2i^3} \dot{+} \frac{H_2^2 l_1}{i^3} + \frac{3\nu_3 H_2 l_1^2}{2i^4} + \frac{\nu_3^2 l_1^3}{2i^5} + \frac{H_3 l_1^2}{2i^3} + \frac{\nu_4 l_1^3}{6i^4} \dot{+} O_p\left(n^{-2}\right) . \quad (3.17)$$

Expression (3.17) is a basic ingredient for the subsequent discussion as it can be used for obtaining asymptotic expansions for the bias, variance and higher order cumulants of the maximum likelihood estimator.

3.4.2 Asymptotic bias of $\hat{\beta}$

For an asymptotic expansion of the bias one merely has to take expectations in both sides of (3.17). The current form of (3.17) simplifies this process as the stochastic parts of the summands are products of centered log-likelihood derivatives and hence their expectations are joint null centered moments, whose asymptotic order can be easily calculated using rule (3.3). Noting that $E_{\beta_0}(l_1/i) = 0$, $\bar{\nu}_{1,1} = i$, $\bar{\nu}_{1,2} = \nu_{1,2}$ and that all terms of order $O_p\left(n^{-3/2}\right)$ have expectations of order $O_p\left(n^{-2}\right)$, the asymptotic bias of $\hat{\beta}$ admits the expansion

$$E_{\beta_0}(\hat{\beta} - \beta_0) = 0 \ddot{+} \frac{\nu_3 + 2\nu_{1,2}}{2i^2} \ddot{+} O\left(n^{-2}\right) , \tag{3.18}$$

where the symbol $\ddot{+}$ denotes a drop in asymptotic order by n^{-1}.

3.4.3 Bias correction

Using (3.18), we can define a *bias-corrected estimator* as $\hat{\beta}_{BC} = \hat{\beta} - b(\hat{\beta})$, where

$$b(\beta) = \frac{\nu_3(\beta) + 2\nu_{1,2}(\beta)}{2\{i(\beta)\}^2} , \tag{3.19}$$

or by a use of the Bartlett relations (3.2),

$$b(\beta) = -\frac{\nu_{1,1,1}(\beta) + \nu_{1,2}(\beta)}{2\{i(\beta)\}^2} . \tag{3.20}$$

The bias-corrected estimator will have bias

$$\begin{aligned} E_{\beta_0}(\hat{\beta}_{BC} - \beta_0) &= E_{\beta_0}(\hat{\beta} - \beta_0) - E_{\beta_0}\{b(\hat{\beta})\} \\ &= b(\beta_0) - E_{\beta_0}\{b(\beta_0) + (\hat{\beta} - \beta_0)b'(\beta_0)\} + O\left(n^{-2}\right) \\ &= b'(\beta_0)b(\beta_0) + O\left(n^{-2}\right) . \end{aligned}$$

A direct application of Skovgaard's theorem (Theorem 3.1) gives

$$b'(\beta_0) = \left.\frac{\mathrm{d}b(\beta)}{\mathrm{d}\beta}\right|_{\beta=\beta_0} = \frac{\nu_3^2 + 2\nu_{1,2}^2 + 3\nu_3\nu_{1,2}}{i^3} + \frac{\nu_4 + 3\nu_{1,3} + 2(\nu_{2,2} + \nu_{1,1,2})}{2i^2} = O\left(n^{-1}\right) , \tag{3.21}$$

and thus $\hat{\beta}_{BC}$ has bias of order $O\left(n^{-2}\right)$ as opposed to that of $\hat{\beta}$ which is of order $O(n^{-1})$.

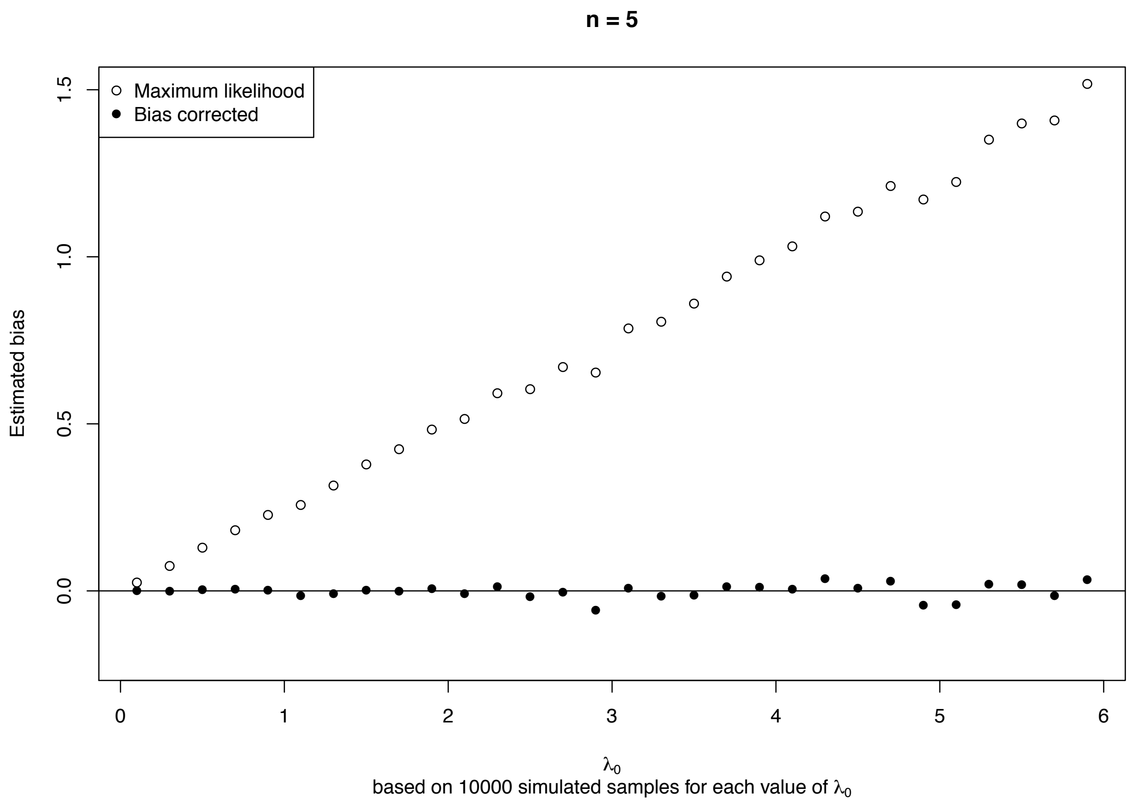
Example 3.2. (Exponential distribution with mean $1/\lambda$) Consider independent and identically distributed random variables with densities as in (1.2). Then, as we have seen in Chapter 1, the maximum likelihood estimator for λ is $\hat{\lambda} = 1/\bar{Y}$. The first three derivatives of the log-likelihood are $l_1(\lambda) = n/\lambda - n\bar{Y}$, $l_2(\lambda) = -n/\lambda^2 = -i(\lambda)$ and $l_3(\lambda) = 2n/\lambda^3 = \nu_3(\lambda)$. Furthermore, as $\nu_1(\lambda) = 0$, in this case we get $\nu_{1,2}(\lambda) = 0$ and substituting in (3.19), $b(\lambda) = \lambda/n$. Hence the bias corrected estimator is $\hat{\lambda}_{BC} = (n-1)/\sum_{i=1}^n Y_i$.

Table 3.1 gives the estimated biases and mean squared errors of $\hat{\lambda}$ and $\hat{\lambda}_{BC}$ for various values of λ_0, based on 10000 simulated samples of size $n = 5$. The superior performance of $\hat{\lambda}_{BC}$ in terms of bias and mean squared error is apparent for any value of λ_0. A more careful

Table 3.1: Estimated bias and estimated mean squared error (MSE) of $\hat{\lambda}$ and $\hat{\lambda}_{BC}$ for a random sample of size $n = 5$ from the exponential distribution with mean $1/\lambda$. The estimates are calculated based on 10000 simulated sampled for each value of λ_0.

λ_0	Estimated bias		Estimated MSE	
	$\hat{\lambda}$	$\hat{\lambda}_{BC}$	$\hat{\lambda}$	$\hat{\lambda}_{BC}$
0.1	0.02504	0.00003	0.00587	0.00336
0.7	0.18160	0.00528	0.29559	0.16810
1.3	0.31539	-0.00769	0.97714	0.56177
1.9	0.48286	0.00629	2.11375	1.20362
2.5	0.60376	-0.01699	3.31455	1.88830
3.3	0.80564	-0.01549	6.22776	3.57062
3.9	0.98943	0.01154	9.20051	5.26192
4.5	1.13511	0.00809	12.52054	7.18859
5.1	1.22392	-0.04086	14.57797	8.37287
5.9	1.51747	0.03397	21.93058	12.56300

Figure 3.1: Estimated bias of $\hat{\lambda}$ and $\hat{\lambda}_{BC}$ for a random sample of size $n = 5$ from the exponential distribution with mean $1/\lambda$.



examination results in Figure 3.1. It appears that the bias of $\hat{\lambda}$ increases linearly with the value of λ and that this bias is eliminated when $\hat{\lambda}_{BC}$ is used.

In fact, in this particular case, noting that $\sum_{i=1}^n Y_i \sim G(n, 1/\lambda)$, a change of variable gives that $\hat{\lambda}$ has density

$$g(\hat{\lambda}; \lambda) = \frac{(n\lambda)^n \exp\left(-n\lambda/\hat{\lambda}\right)}{\hat{\lambda}^{n+1}\Gamma(n)}, \quad \lambda > 0 .$$

A calculation of the expected value gives that $E_{\lambda_0}(\hat{\lambda}) = n\lambda_0/(n-1)$, so that the estimator $(n-1)/\sum_{i=1}^n Y_i$ is exactly unbiased. Hence, in that case correction of the first-order bias results in an exactly unbiased estimator. □

Example 3.3. (Sampling from exponential families) Consider independent random variables $Y_1, \ldots, Y_n$ each having density (or probability mass) functions of the form (2.11) with $d = 1$, natural statistic $t(y_i)$, natural parameter $\theta_i(\beta)$ $(i = 1, \ldots, n)$ and common cumulant transform $k(.)$. Then the log-likelihood for β has the form

$$l(\beta) = \sum_{i=1}^n t(Y_i)\theta_i(\beta) - \sum_{i=1}^n k\{\theta_i(\beta)\}, \quad \beta \in B \subseteq \Re^p , \tag{3.22}$$

up to some additive constant not depending on β. Note that, according to the definition in Example 2.7, if $\theta_i(\beta) = c_{i0} + \sum_{t=1}^p \beta_t c_{it}$, for constants c_{it} $(i = 1, \ldots, n\,; t = 0, \ldots, p)$ then the joint distribution of $Y_1, \ldots, Y_n$ is a full exponential family with natural statistics $\sum_{i=1}^n t(y_i)c_{i1}, \ldots, \sum_{i=1}^n t(y_i)c_{ip}$ and natural parameters $\beta_1, \ldots, \beta_p$.

Expression (3.22) is the general form of the log-likelihood for many well-used generalized linear models. For example, in binomial regression models, $t(y_i) = y_i$, $Y_i \sim B(m_i, \pi_i)$, $\theta_i = \log\{\pi_i/(1-\pi_i)\}$, $k(\theta_i) = m_i \log\{1+\exp(\theta_i)\}$ and $g(\pi_i) = \sum_{t=1}^p \beta_t x_{it}$, for some monotone function g where x_{it} is the (i,t)th entry of a matrix X with full rank p.

For $p = 1$, the first two log-likelihood derivatives are

$$\begin{aligned} l_1(\beta) &= \sum_{i=1}^n \theta_i'(\beta)\left[t(Y_i) - k'\{\theta_i(\beta)\}\right] , \\ l_2(\beta) &= \sum_{i=1}^n \theta_i''(\beta)\left[t(Y_i) - k'\{\theta_i(\beta)\}\right] - \sum_{i=1}^n \{\theta_i'(\beta)\}^2 k''\{\theta_i(\beta)\} , \end{aligned} \tag{3.23}$$

where $k'(\theta) = \mathrm{d}k(\theta)/\mathrm{d}\theta$, $k''(\theta) = \mathrm{d}^2k(\theta)/\mathrm{d}\theta^2$ etc. Noting that $E_\beta\{t(Y_i)\} = k'\{\theta_i(\beta)\}$ and $\mathrm{var}_\beta\{t(Y_i)\} = k''\{\theta_i(\beta)\}$,

$$\begin{aligned} i(\beta) &= \sum_{i=1}^n \{\theta_i'(\beta)\}^2 \mathrm{var}_\beta\{t(Y_i)\} , \\ \nu_{1,2}(\beta) &= \sum_{i=1}^n \theta_i'(\beta)\theta_i''(\beta)\mathrm{var}_\beta\{t(Y_i)\} , \\ \nu_{1,1,1}(\beta) &= \sum_{i=1}^n \{\theta_i'(\beta)\}^3 \mathrm{cum}_{3,\beta}\{t(Y_i)\} , \end{aligned}$$

where $\mathrm{cum}_{3,\beta}\{t(Y_i)\} = k'''\{\theta_i(\beta)\}$ is the third-order cumulant of $t(Y_i)$ $(i = 1, \ldots, n)$. Then, a substitution of the above expressions in (3.20) gives

$$b(\beta) = -\frac{\sum_{i=1}^n \left[\theta_i'(\beta)\theta_i''(\beta)\mathrm{var}_\beta\{t(Y_i)\} + \{\theta_i'(\beta)\}^3 \mathrm{cum}_{3,\beta}\{t(Y_i)\}\right]}{2\left[\sum_{i=1}^n \{\theta_i'(\beta)\}^2 \mathrm{var}_\beta\{t(Y_i)\}\right]^2} . \quad (3.24)$$

Now, if $\theta_i(\beta) = c_{i0} + c_i\beta$, for constants c_{i0} and c_i $(i = 1, \ldots, n)$ (i.e. when the family is full with natural statistic $\sum_{i=1}^n t(Y_i)c_i$), then $\theta_i'(\beta) = c_i$ and $\theta_i''(\beta) = 0$. Hence, (3.24) simplifies to

$$b(\beta) = -\frac{\sum_{i=1}^n c_i^3 \mathrm{cum}_{3,\beta}\{t(Y_i)\}}{2\left[\sum_{i=1}^n c_i^2 \mathrm{var}_\beta\{t(Y_i)\}\right]^2} . \quad (3.25)$$

Thus, the expression for the first-order bias in Example 3.2 can be derived by noting that, in that case, $t(y_i) = y_i$, $c_i = -1$, $\mathrm{var}_\lambda(Y_i) = 1/\lambda^2$ and $\mathrm{cum}_{3,\lambda}(Y_i) = 2/\lambda^3$.

3.4.4 Bias reduction

An alternative estimator with bias of order $O\left(n^{-2}\right)$ results as the root of the adjusted score equation

$$l_1(\beta) - i(\beta)b(\beta) = 0 .$$

This result was derived in Firth (1993) for regular families with $p \geq 1$ parameters, but it can be easily derived in the one-parameter case: let $\hat\beta_{BR}$ be the estimator that results from the solution of $l_1^*(\beta) = l_1(\beta) + A(\beta) = 0$, where $A(\beta)$ is some function of the parameter which is $O(1)$ as $n \to \infty$. Then, exactly as was done for the maximum likelihood estimator, expand $l_1(\hat\beta_{BR})$ around β_0 and discard terms of order $O_p\left(n^{-1}\right)$. Inverting the resultant expansion in terms of $\hat\beta_{BR} - \beta_0$ and taking expectations reveals that a bias-reducing adjustment to the score function is $A(\beta) = -i(\beta)b(\beta)$.

In the case of repeated sampling under a full exponential family with natural parameter β, the first term of the right hand side of expression (3.23) for $l_2(\beta)$ is zero and thus $\nu_{2,1}(\beta) = 0$ because $l_2(\beta)$ does not depend on the data. Hence, an application of Skovgaard's theorem gives

$$A(\beta) = i(\beta)\frac{\nu_{1,1,1}(\beta)}{2\{i(\beta)\}^2} = \frac{\mathrm{d}}{\mathrm{d}\beta}\log\{i(\beta)\}^{1/2} .$$

Thus, $l_1^*(\beta)$ is the derivative of the logarithm of the penalized likelihood $L^*(\beta) = L(\beta)\{i(\beta)\}^{1/2}$; for a full exponential family model, the posterior mode when using the Jeffreys invariant prior (Jeffreys, 1946) has second-order bias in terms of its frequentist properties.

Example 3.4. (Logistic regression) Consider independent binomial random variables $Y_1, \ldots, Y_n$ with probabilities of "success" $\pi_1, \ldots, \pi_n$, respectively, and totals m each, and suppose that the log-odds of success satisfy the relationship

$$\log\frac{\pi_i}{1-\pi_i} = \beta x_i \quad (i = 1, \ldots, n) ,$$

where β is a scalar parameter. This is a logistic regression model and is an one-parameter exponential family with natural statistic $\sum_{i=1}^n y_i x_i$ and natural parameter β (see Example 3.3). The likelihood equation $l_1(\beta) = 0$ is directly obtained by (3.23); substituting $\theta_i(\beta) = \beta x_i$ gives

$$\sum_{i=1}^n Y_i x_i = \sum m\pi_i(\beta)x_i , \quad (3.26)$$

that is the maximum likelihood estimate results by equating the natural statistic to its expected value. The latter is true for *any* full exponential family model.

Differentiating the cumulant transform for the binomial distribution in Table 2.1 gives $\mathrm{var}_\beta(Y_i) = m\pi_i(1-\pi_i)$ and $\mathrm{cum}_{3,\beta}(Y_i) = m\pi_i(1-\pi_i)(1-2\pi_i)$. Thus, by (3.25) the bias-corrected estimator for β is

$$\hat\beta_{BC} = \hat\beta + \frac{\sum_{i=1}^n \left\{\hat\pi_i(1-\hat\pi_i)(1-2\hat\pi_i)x_i^3\right\}}{2m\left\{\sum_{i=1}^n \hat\pi_i(1-\hat\pi_i)x_i^2\right\}^2}\,,$$

where $\hat\pi_i = \exp(\hat\beta x_i)/\{1+\exp(\hat\beta x_i)\}$ $(i = 1,\ldots,n)$.

Furthermore, because the family is full, the bias-reduced estimate $\hat\beta_{BR}$ results by maximizing

$$l(\beta) + \frac{1}{2}\log\left[\sum_{i=1}^n m\pi_i(\beta)\{1-\pi_i(\beta)\}x_i^2\right]$$

For $n = 1$, and $x = 1$, the setting of Example 2.8 results with $\hat\beta_{BC} = \log\{Y/(m-Y)\} - (m-2Y)/\{2Y(m-Y)\}$ and $\hat\beta_{BR} = \log\{(Y+1/2)/(m-Y+1/2)\}$. Thus $\hat\beta_{BR}$ reproduces the Haldane-Anscombe correction, while $\hat\beta_{BC}$ naively subtracts the estimated first-order bias from $\hat\beta$, and thus it is undefined if either $Y = 0$ or $Y = m$ is observed.

Now, let $n = 5$, $m = 2$ and $x = (-2,-1,0,1,2)$. In this case the natural statistic can only take the values $-6,-5,\ldots,0,\ldots,5,6$ and so according to (3.26), there are at most 13 possible values for the estimators $\hat\beta$, $\hat\beta_{BC}$ and $\hat\beta_{BR}$.

Those values are given in Table 3.2. When the natural statistic is either -6 or 6, $\hat\beta$ is infinite and hence $\hat\beta_{BC}$ is undefined. In contrast, the bias-reduced estimate is always finite.

Note that if we give a certain value for the true parameter β_0, then we can calculate the true $\pi_1,\ldots,\pi_5$ and hence the probability that the natural statistic takes a particular value. These sampling probabilities are given in Table 3.2 for $\beta_0 = 0.5$ and $\beta_0 = 1$. The probability of observing a sample that results in infinite maximum likelihood estimate — or, equivalently the probability that $\hat\beta_{BC}$ is undefined — is not negligible; it is about 0.04 for $\beta_0 = 0.5$ and 0.17 for $\beta_0 = 1$. The bias and mean squared error of the $\hat\beta_{BR}$ are well-defined quantities and are 0.007 and 0.264, respectively for $\beta_0 = 0.5$, and -0.029 and 0.328 for $\beta_0 = 1$. These seem satisfactory given the small sample-size and given that the corresponding quantities for $\hat\beta$ and $\hat\beta_{BC}$ are undefined. □

In previous examples, we have seen how we can improve estimation by using estimators with $O\left(n^{-2}\right)$ bias. However, such estimators have the big disadvantage of not being parameterization invariant. For example, the unbiased estimator $\hat\sigma^2$ of σ^2 (see Example 2.2) does not deliver an unbiased estimator of σ. Furthermore, there are cases where bias correction/reduction can inflate the variance (see Subsection 3.4.6). The use of those estimators is suggested only after a specific parameterization for the problem has been chosen via a parameterization invariant method (like maximum likelihood). Nevertheless, even then their properties should be carefully examined either analytically or via simulation.

3.4.5 Asymptotic variance of $\hat\beta$

By definition, the variance of an estimator $\hat\beta$ can be written as

$$\mathrm{var}_{\beta_0}(\hat\beta) = E_{\beta_0}\left\{(\hat\beta-\beta_0)^2\right\} - \left\{E_{\beta_0}(\hat\beta-\beta_0)\right\}^2\,. \qquad (3.27)$$

Table 3.2: The values of $\hat{\beta}$, $\hat{\beta}_{BC}$ and $\hat{\beta}_{BR}$ for each value of the natural statistic for $n = 5$, $m = 2$ and $x = (-2, -1, 0, 1, 2)$. The sampling probabilities for $\beta_0 = 0.5$ and $\beta_0 = 1$ are also given.

$\sum_{i=1}^{5} y_i x_i$	$\hat{\beta}$	$\hat{\beta}_{BC}$	$\hat{\beta}_{BR}$	Sampling probabilities $\beta_0 = 0.5$	$\beta_0 = 1$
-6	$-\infty$	–	-2.009	< 0.001	< 0.001
-5	-1.587	-1.047	-1.183	0.001	< 0.001
-4	-1.012	-0.767	-0.815	0.003	< 0.001
-3	-0.674	-0.538	-0.561	0.010	< 0.001
-2	-0.420	-0.343	-0.355	0.024	0.002
-1	-0.202	-0.167	-0.173	0.052	0.006
0	0	0	0	0.094	0.019
1	0.202	0.167	0.173	0.141	0.046
2	0.420	0.343	0.355	0.180	0.098
3	0.674	0.538	0.561	0.191	0.171
4	1.012	0.767	0.815	0.158	0.233
5	1.587	1.047	1.183	0.104	0.253
6	∞	–	2.008	0.043	0.172

Hence, if we obtain an asymptotic expansion for the mean squared error $E_{\beta_0}\{(\hat{\beta} - \beta_0)^2\}$ and substitute the asymptotic expansion (3.18) for $E_{\beta_0}(\hat{\beta} - \beta_0)$ in the above expression gives an asymptotic expansion for the variance of $\hat{\beta}$.

Taking squares in both sides of (3.17) and disregarding terms of order $O_p(n^{-5/2})$ (this order is the best we can do in terms of approximation as it is the order of the product of the first term and the remainder in (3.17)), gives

$$
\begin{aligned}
(\hat{\beta} - \beta_0)^2 = {} & \frac{l_1^2}{i^2} \dot{+} \frac{2H_2 l_1^2}{i^3} + \frac{\nu_3 l_1^3}{i^4} \\
& \dot{+} \frac{3H_2^2 l_1^2}{i^4} + \frac{4\nu_3 H_2 l_1^3}{i^5} + \frac{5\nu_3^2 l_1^4}{4i^6} + \frac{H_3 l_1^3}{i^4} + \frac{\nu_4 l_1^4}{3i^5} \dot{+} O_p\left(n^{-5/2}\right) .
\end{aligned} \tag{3.28}
$$

Taking expectations in both sides of the above expression gives

$$
\begin{aligned}
E\left\{(\hat{\beta} - \beta_0)^2\right\} = {} & \frac{1}{i} \ddot{+} \frac{2\bar{\nu}_{1,1,2}}{i^3} + \frac{\nu_3 \bar{\nu}_{1,1,1}}{i^4} + \frac{3\bar{\nu}_{1,1,2,2}}{i^4} + \frac{4\nu_3 \bar{\nu}_{1,1,1,2}}{i^5} \\
& + \frac{5\nu_3^2 \bar{\nu}_{1,1,1,1}}{4i^6} + \frac{\bar{\nu}_{1,1,1,3}}{i^4} + \frac{\nu_4 \bar{\nu}_{1,1,1,1}}{3i^5} \ddot{+} O\left(n^{-3}\right) .
\end{aligned} \tag{3.29}
$$

Now, by (3.4) derived in Example 3.1 and by the relation $v_2 = -i$,

$$
\begin{aligned}
\bar{\nu}_{1,1,1,1} &= 3i^2 + O(n) , \\
\bar{\nu}_{1,1,1,2} &= 3i\nu_{1,2} + O(n) , \\
\bar{\nu}_{1,1,1,3} &= 3i\nu_{1,3} + O(n) , \\
\bar{\nu}_{1,1,2,2} &= 2\nu_{1,2}^2 + i(\nu_{2,2} - i^2) + O(n) .
\end{aligned} \tag{3.30}
$$

Furthermore, by the definition of H_1 and H_2 and the Bartlett identities,

$$
\begin{aligned}
\bar{\nu}_{1,1,1} &= \nu_{1,1,1} = -\nu_3 - 3\nu_{1,2} , \\
\bar{\nu}_{1,1,2} &= \nu_{1,1,2} + i^2 .
\end{aligned} \tag{3.31}
$$

Substituting relations (3.30) and (3.31) in (3.29) and collecting terms, an asymptotic expansion for the mean squared error of $\hat{\beta}$ is

$$E\left\{(\hat{\beta}-\beta_0)^2\right\} = \frac{1}{i} \ddot{+} \frac{2\nu_{1,1,2} + (\nu_{2,2} - i^2) + 2\nu_{2,2} + \nu_4 + 3\nu_{3,1}}{i^3} + \frac{11\nu_3^2 + 36\nu_3\nu_{1,2} + 24\nu_{1,2}^2}{4i^4} \ddot{+} O\left(n^{-3}\right) . \tag{3.32}$$

Taking squares at both sides of (3.18) results in the asymptotic expansion

$$\left\{E_{\beta_0}(\hat{\beta}-\beta_0)\right\}^2 = \frac{\nu_3^2 + 4\nu_{1,2}^2 + 4\nu_3\nu_{1,2}}{4i^4} \ddot{+} O\left(n^{-3}\right) . \tag{3.33}$$

Hence, a simple substitution of (3.33) and (3.32) in (3.27) gives

$$\mathrm{var}_{\beta_0}(\hat{\beta}) = \frac{1}{i} \ddot{+} \frac{2\nu_{1,1,2} + (\nu_{2,2} - i^2) + 2\nu_{2,2} + \nu_4 + 3\nu_{3,1}}{i^3} + \frac{5\nu_3^2 + 16\nu_3\nu_{1,2} + 10\nu_{1,2}^2}{2i^4} \ddot{+} O\left(n^{-3}\right) . \tag{3.34}$$

3.4.6 Bias correction/reduction and second-order efficiency

Using (3.33) and (3.21), (3.34) can be re-written in the simple form

$$\mathrm{var}_{\beta_0}(\hat{\beta}) = \frac{1}{i} \ddot{+} \frac{2b'}{i} + 2b^2 + \frac{\gamma^2}{i} \ddot{+} O\left(n^{-3}\right) , \tag{3.35}$$

where $b \equiv b(\beta_0)$, $b' \equiv b'(\beta_0)$. The quantity

$$\gamma \equiv \gamma(\beta_0) = i^{-3/2}\{i(\nu_{2,2} - i^2) - \nu_{1,2}^2\}^{1/2} ,$$

is called the *statistical curvature* (Efron, 1975) and is zero for full exponential families and positive else.

Furthermore, it can be shown that the variance of the bias-corrected estimator admits the expansion

$$\mathrm{var}_{\beta_0}(\hat{\beta}_{BC}) = \frac{1}{i} \ddot{+} 2b^2 + \frac{\gamma^2}{i} \ddot{+} O\left(n^{-3}\right) , \tag{3.36}$$

and that the same expansion is also valid for $\mathrm{var}_{\beta_0}(\hat{\beta}_{BR})$. Ignoring the $O\left(n^{-3}\right)$ terms, expressions (3.35) and (3.36) are the same except for the term $2b'/i$ which does not appear in (3.36). While all other terms outside the remainder are non-negative, the sign of $2b'/i$ depends on the particular model we consider and on β_0. Hence, we cannot say anything definite on which estimator has smaller variance, and, as already mentioned, there can be cases where bias correction/reduction leads to estimators with larger variance than $\mathrm{var}_{\beta_0}(\hat{\beta})$.

Nevertheless, the $O\left(n^{-2}\right)$ term in (3.36) is non-negative, and is zero only when $\gamma = 0$ and $b = 0$ that is for full exponential families in mean-value parameterization (i.e. families for which $E_\beta\{T(Y_1, \ldots, Y_n)\} = \beta$, where $T(Y_1, \ldots, Y_n)$ is the natural statistic). In fact, it can be shown that this is the case where the maximum likelihood estimator is exactly unbiased and the Cramér-Rao lower bound is attained.

Finally, it can be shown (Efron, 1975) that for *any* estimator $\tilde{\beta}$ with bias of order $O\left(n^{-2}\right)$,

$$\mathrm{var}_{\beta_0}(\tilde{\beta}) = V + \Delta + O(n^{-3}) ,$$

where V is the right hand side of (3.36) and $\Delta \equiv \Delta(\beta_0)$ is non-negative being zero when $\tilde{\beta}$ is either $\hat{\beta}_{BC}$ or $\hat{\beta}_{BR}$. Hence, in the class of estimators with bias of order $O\left(n^{-2}\right)$, $\hat{\beta}_{BC}$ and $\hat{\beta}_{BR}$ are both *second order efficient* estimators of β.

3.5 Approximate pivots

3.5.1 Expansion of $l(\hat\beta) - l(\beta_0)$

We now turn our attention to the log-likelihood ratio and the construction of asymptotic pivots based on it. Letting $\delta = \hat\beta - \beta_0$, an expansion of $l(\hat\beta)$ around β_0 gives

$$l(\hat\beta) - l(\beta_0) = \delta l_1 + \frac{1}{2}\delta^2 l_2 \dot{+} \frac{1}{6}\delta^3 l_3 \dot{+} \frac{1}{24}\delta^4 l_4 + O_p\left(n^{-3/2}\right) .$$

Replacing $l_r = \nu_r + H_r$ $(r = 1, \ldots, n)$ we get

$$l(\hat\beta) - l(\beta_0) = \delta l_1 + \frac{1}{2}\delta^2 \nu_2 \dot{+} \frac{1}{2}\delta^2 H_2 + \frac{1}{6}\delta^3 \nu_3 \dot{+} \frac{1}{6}\delta^3 H_3 + \frac{1}{24}\delta^4 \nu_4 + O_p\left(n^{-3/2}\right) . \tag{3.37}$$

Now, consider random variables X, Y and Z such that $X = O_p\left(n^{-1/2}\right)$, $Y = O_p\left(n^{-1}\right)$ and $Z = O_p\left(n^{-3/2}\right)$. Then,

$$(X + Y + Z)^3 = X^3 \dot{+} 3X^2 Y \dot{+} 3XY^2 + 3X^2 Z \dot{+} O_p\left(n^{-3}\right) .$$

Substituting each of X, Y and Z with the $O_p\left(n^{-1/2}\right)$, $O_p\left(n^{-1}\right)$ and $O_p\left(n^{-3/2}\right)$ terms in (3.17) we get an expansion for δ^3. Now, if we use that expansion in (3.37) along with the expansions (3.17) and (3.28), some algebra gives that

$$\begin{aligned} l(\hat\beta) - l(\beta_0) = & \frac{l_1^2}{2i} \dot{+} \frac{\nu_3 l_1^3 + 3iH_2 l_1^2}{6i^3} \dot{+} \frac{3\nu_3^2 l_1^4 + i\nu_4 l_1^4}{24 i^5} \\ & + \frac{H_3 l_1^3 + 3\nu_3 H_2 l_1^3}{6i^3} + \frac{H_2^2 l_1^2}{2i^3} \dot{+} O_p\left(n^{-3/2}\right) . \end{aligned} \tag{3.38}$$

According to the above expansion a first-order approximation of the log-likelihood ratio is $l_1^2/2i$. Nevertheless, by the central limit theorem $l_1/\sqrt{i} \xrightarrow{d} N(0,1)$ and an application of the continuous mapping theorem gives that $l_1^2/i \xrightarrow{d} \chi^2_1$ (chi-squared distribution with one degree of freedom). Hence, by Slutzky's lemma $w(\beta_0) = 2\{l(\hat\beta) - l(\beta_0)\} \xrightarrow{d} \chi^2_1$. Furthermore, an expansion of $\{l_1^2/i + a\}^{1/2}$ around $a = 0$ gives that $r(\beta)$ of Chapter 1 has a $N(0,1)$ limiting distribution.

Finally, expanding $l(\beta_0)$ around $\hat\beta$ we can write

$$-2\{l(\beta_0) - l(\hat\beta)\} = (\hat\beta - \beta_0)^2 j(\hat\beta) \dot{+} O_p\left(n^{-1/2}\right) . \tag{3.39}$$

The above expression reveals that the quantity $t^2(\beta_0) = (\hat\beta - \beta_0)^2 j(\hat\beta)$ can be viewed as a quadratic approximation of $w(\beta_0)$ at $\hat\beta$. Then, as before, an expansion of $\{(\hat\beta - \beta_0)^2 j(\hat\beta) + a\}^{1/2}$ around $a = 0$ gives that the approximate pivot $t(\beta)$ is a linearized version of $r(\beta)$ at $\hat\beta$. Note that by the asymptotic normality of the maximum likelihood estimator $t(\beta)$ has a $N(0,1)$ limiting distribution. This provides an alternative justification for the limiting distribution of $w(\beta)$.

3.6 Bartlett correction

If we take expectations in both sides of (3.38) and multiply by 2, then

$$E_{\beta_0}\{w(\beta_0)\} = 1 \ddot{+} d(\beta_0) \ddot{+} O\left(n^{-2}\right) .$$

The above expression suggests that if we define

$$w^*(\beta) = \frac{w(\beta)}{1 + d(\beta)} \,,$$

then $E_\beta\{w^*(\beta)\} = 1 + O\left(n^{-2}\right)$ which gives a better approximation to the expectation of the limiting χ^2_1 distribution. This simple adjustment is known as Bartlett correction (Bartlett, 1937), and for continuous models it has the remarkable effect of correcting not only the expectation but simultaneously all the cumulants — and hence, the distribution — of $w(\beta)$ towards those of the χ^2_1 distribution.

In particular, if $G(c)$ is the distribution function of a χ^2_1 random variable, it can be shown that

$$P\{w(\beta) \le c; \beta\} = G(c)\left\{1 + O(n^{-1})\right\} \,,$$

while

$$P\{w^*(\beta) \le c; \beta\} = G(c)\left\{1 + O(n^{-2})\right\} \,,$$

Example 3.5. (Sampling from exponential families - continued) Continuing Example 3.3 we consider a more special case where $Y_1, \ldots, Y_n$ are independent and *identically* distributed copies of a random variable Y which is distributed according to an one-parameter exponential family of distributions with natural parameter θ. According to (3.22) the log-likelihood for the natural parameter θ is

$$l(\theta) = \theta \sum_{i=1}^{n} t(Y_i) - nk(\theta) \,.$$

Then,

$$\begin{aligned} l_1(\theta) &= \sum_{i=1}^{n} t(Y_i) - nk'(\theta) \,, \\ l_r(\theta) &= -n \frac{\mathrm{d}^r k(\theta)}{\mathrm{d}\theta^r} \quad (r = 2, 3, \ldots) \,, \end{aligned} \tag{3.40}$$

hence $l_r(\theta) = \nu_r(\theta)$ and $H_r(\theta) = 0$ for $r = 2, 3, \ldots$. Furthermore, by Example 2.7, the cumulants of $t(Y)$ are $\kappa_r = \mathrm{d}^r k(\theta)/\mathrm{d}\theta^r$ $(r = 1, 2, \ldots)$ and so $i(\theta) = n\kappa_2$. Hence, in that case the terms in the second row of expression (3.38) are all zero and thus

$$w(\theta) = \frac{1}{n\kappa_2}\{l_1(\theta)\}^2 \dot{-} \frac{\kappa_3}{3n^2\kappa_2^3}\{l_1(\theta)\}^3 \dot{-} \frac{1}{12n^3\kappa_2^4}\left[\kappa_4 - \frac{3\kappa_3^2}{\kappa_2}\right]\{l_1(\theta)\}^4 \dot{+} O_p\left(n^{-3/2}\right) \tag{3.41}$$

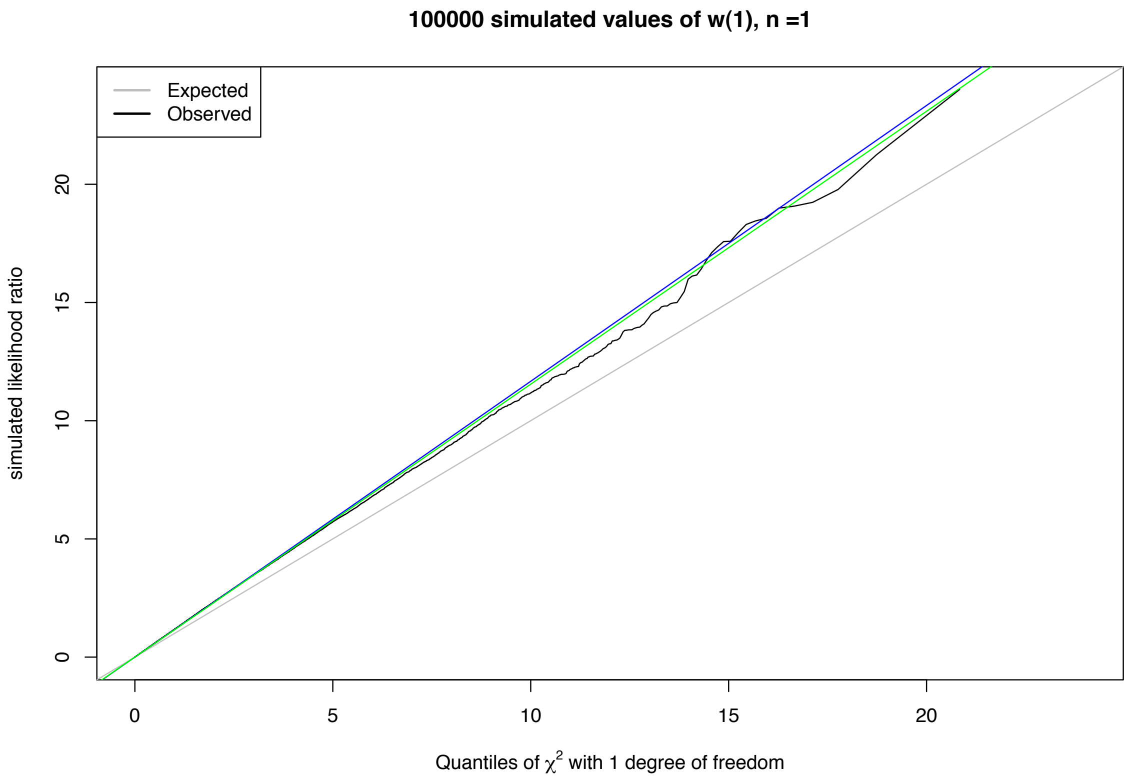
Noting that $\nu_{1,r} = 0$ for $r = 2, 3, \ldots$ and by (3.31) and (3.30),

$$\begin{aligned} \nu_{1,1,1}(\theta) &= n\kappa_3 \,, \\ \nu_{1,1,1,1}(\theta) &= 3n^2\kappa_2^2 + O(n) \,. \end{aligned}$$

Thus, if we take expectations in both sides of (3.41), we get the expansion

$$E_\theta\{w(\theta)\} = 1 \ddot{+} \frac{3\kappa_3^2}{4n\kappa_2^3} - \frac{\kappa_3^2}{3n\kappa_2^3} - \frac{\kappa_4}{4n\kappa_2^2} \dot{+} O_p\left(n^{-2}\right) \,.$$

Figure 3.2: Q-Q plot of 100000 simulated values of $w(1)$ against the quantiles of a χ^2_1 distribution for $n = 1$. The grey line is expected relationship between the quantiles of the χ^2_1 distribution and the values of $w(1)$. The blue and the green lines are through the origin and have slopes the expectation of $w(1)$ up to order $O(n^{-2})$ and the exact expectation of $w(1)$, respectively.



Hence writing, $\rho_3 = \kappa_3/\kappa_2^{3/2}$ and $\rho_4 = \kappa_4/\kappa_2^2$ for the standardized cumulants (see Subsection 2.3.2),

$$E_\theta\{w(\theta)\} = 1 - \frac{1}{12n}\left(3\rho_4 - 5\rho_3^2\right) + O\left(n^{-2}\right) .$$

Thus, the Bartlett corrected version of $w(\beta)$ is

$$w^*(\theta) = \frac{12n}{12n + 5\rho_3^2 - 3\rho_4} w(\theta) .$$

For the setting of Section 1.2, $w(\lambda) = \{r(\lambda)\}^2 = 2n(\bar{Y}\lambda - \log(\bar{Y}\lambda) - 1)$ and the cumulants of $t(Y) = Y$, in this case, are $\kappa_1 = 1/\lambda$, $\kappa_2 = 1/\lambda^2$, $\kappa_3 = 2/\lambda^3$ and $\kappa_4 = 6/\lambda^4$. Thus $\rho_3 = 2$ and $\rho_4 = 6$. Hence, $E_\lambda\{w(\lambda)\} = 1 + 1/(6n) + O\left(n^{-2}\right)$ and the Bartlett corrected version of $w(\lambda)$ is

$$w^*(\lambda) = \frac{12n(\bar{Y}\lambda - \log(\bar{Y}\lambda) - 1)}{6n + 1} .$$

Figure 3.2 is a Q-Q plot of 100000 simulated values of $w(1)$ against the quantiles of a χ^2_1 distribution for $n = 1$. The expected relationship between the quantiles of the χ^2_1 distribution and the values of $w(1)$ is the grey 45^o line. Nevertheless, the relationship departs from what is expected. Importantly, the relationship seems to agree with a line through the origin with slope $1+1/(6n)$ (blue line). This is the Bartlett correction and in this case, even for $n = 1$, is very close to the exact expectation of $w(\lambda)$ (green line) which is $E_\lambda\{w(\lambda)\} = 2n\log n - 2n\psi(n)$ (to see that this is the exact expectation of $w(\lambda)$ note that $\sum_{i=1}^n Y_i \sim G(n, 1/\lambda)$, find the natural statistics when both natural parameters $\theta_1 = n$ and $\theta_2 = \lambda$ are unknown and find $E(\log \sum_{i=1}^n Y_i)$ via the cumulant transform). $\square$

In models for discrete random variables, Bartlett correction does not necessarily lead to an improved χ^2_1 approximation. For example, the simulation studies in Frydenberg and Jensen (1989) suggest that for a model with multinomial responses the Bartlett correction of $w(\beta)$ does not lead to any significant improvement to the χ^2_1 approximation.

3.7 The case of more than one parameters

To obtain the asymptotic expansions of previous sections in the case of more than one parameters would require the application of the multivariate Taylor's theorem and the multivariate extensions of the definitions and results of Chapter 2. As for the notation, there is a set of notational rules called *index notation* which along with the *Einstein's summation convention* provide an elegant way of obtaining the corresponding expansions, treating the terms as if they were one-dimensional and avoiding algebraic considerations when taking products of multidimensional arrays. For the interested reader, two excellent textbooks focusing on multi-parameter expansions and generally statistical asymptotics are McCullagh and Nelder (1989) and Pace and Salvan (1997). Here we shall merely provide the extensions of the results we have seen so far in many dimensions.

We need to distinguish between two different setting in the multi-parameter case: 1) all p parameters are of interest, and 2) there are k parameters of interest and $p-k$ *nuisances*, that is parameters which are of secondary scientific importance, though essential for realistic modelling.

3.7.1 All p parameters of interest

Score function and information

For problems with a p-dimensional parameter β (no nuisances), the score function $u(\beta) \equiv u(\beta; Y)$ has p components and is defined as

$$u(\beta) = \nabla l(\beta) \, ,$$

where $l(\beta) \equiv l(\beta; Y)$ is the log-likelihood function. Arguing component by component, a similar argument as in the single-parameter case gives $E_\beta\{u(\beta)\} = 0$.

Furthermore the observed information $j(\beta)$ is now a $p \times p$ matrix and is defined as

$$j(\beta) = -\nabla\nabla^T l(\beta) \, .$$

The expected information matrix is

$$i(\beta) = E_\beta\{j(\beta)\} \, . \tag{3.42}$$

Similar arguments as in the case of a single parameter result in $i(\beta) = \mathrm{cov}_\beta\{u(\beta)\}$, which along with (3.42) provide a generalization of the information identity in (3.2). In fact, all Bartlett identities can be written in the multi-parameter settings with appropriate extensions in notation (for example, in the case of p parameters, the quantity to ν_4 corresponds to a four-way $p \times p \times p \times p$ array).

Asymptotic bias and asymptotic variance of the maximum likelihood estimator

The first-order bias term of the maximum likelihood estimator $\hat\beta$ in the multi-parameter case takes the form

$$b(\beta_0) = -i(\beta_0)^{-1} A(\beta_0) \, .$$

The function $A(\beta_0)$ has tth component

$$A_t = \frac{1}{2} \mathrm{tr}\left[i^{-1}\{P_t + Q_t\}\right] \quad (t = 1, \ldots, p) \, ,$$

where $\mathrm{tr}(B)$, denotes the trace of a matrix B and $P_t = E_{\beta_0}\{uu^T u_t\}$, $Q_t = -E_{\beta_0}\{ju_t\}$ are joint null moments of log-likelihood derivatives with $u \equiv u(\beta_0)$, $j \equiv j(\beta_0)$, etc. The bias corrected estimator is $\hat{\beta}_{BC} = \hat{\beta} - b(\hat{\beta})$ while the bias-reduced estimator results by the solution of the adjusted-score equations

$$u(\beta) - i(\beta)b(\beta) = 0\,.$$

Furthermore, the nice penalized likelihood interpretation of bias-reduction in Subsection 3.4.4 extends in the multi-parameter case; in *full exponential families with natural parameters* β, if we penalize the likelihood by Jeffreys invariant prior, that is

$$l^*(\beta) = l(\beta) + \frac{1}{2}\log|i(\beta)|\;,$$

then the posterior mode has bias of order $O\left(n^{-2}\right)$. In the above expression $|i(\beta)|$ denotes the determinant of $i(\beta)$.

A similar expansion as (3.34), but now in the multi-parameter case gives

$$\mathrm{cov}_{\beta_0}(\hat{\beta}) = \{i(\beta_0)\}^{-1} \ddot{+} O\left(n^{-2}\right)\;,$$

so that the asymptotic variance-covariance matrix of the maximum likelihood estimator is $\{i(\beta_0)\}^{-1}$, which is also the Cramér-Rao lower bound. All the considerations on first and second order efficiency in previous sections directly extend to the case of many parameters.

Confidence regions

As already mentioned in Subsection 3.3.2 if β is p-dimensional then

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N_p(0, \{i^*(\beta_0)\}^{-1})\,, \tag{3.43}$$

where $i^*(\beta_0)$ is the expected information per observation ($i^*(\beta_0) = i(\beta_0)/n$). Thus, if we define,

$$t^2(\beta) = (\hat{\beta} - \beta)^T i(\beta)(\hat{\beta} - \beta)\,, \tag{3.44}$$

then, by the continuous mapping theorem, we can construct approximate *confidence regions* at level $1 - \alpha$ as

$$\{\beta\,:\, t^2(\beta) \le \chi^2_{p,1-\alpha}\}\,, \tag{3.45}$$

where $\chi^2_{p,1-\alpha}$ is the $(1-a)$th quantile of the chi-squared distribution with p degrees of freedom. The expected information $i(\beta)$ in expression (3.44) can be replaced by either $i(\hat{\beta})$, $j(\beta)$, or $j(\hat{\beta})$ without affecting the limiting distribution.

Furthermore, the generalization of the expansion (3.38) of the log-likelihood ratio to many dimensions gives

$$w(\beta_0) = u(\beta_0)^T\{i(\beta_0)\}^{-1}u(\beta_0) + O_p\left(n^{-1/2}\right)\;, \tag{3.46}$$

and so, by an application of the multivariate version of the central limit theorem on $u(\beta)$, the quantities

$$s^2(\beta) = u(\beta)^T\{i(\beta)\}^{-1}u(\beta)\,, \tag{3.47}$$

$$w(\beta) = 2\{l(\hat{\beta}) - l(\beta)\}\,, \tag{3.48}$$

both have a limiting chi-squared distribution with p degrees of freedom. Hence, approximate confidence regions can be constructed as in (3.45) by replacing $t^2(\beta)$ with either $w(\beta)$ or $s^2(\beta)$. Again, $i(\beta)$ in expression (3.47) can be replaced by either $i(\hat{\beta})$, $j(\beta)$, or $j(\hat{\beta})$ without affecting the limiting distribution.

Example 3.6. (Weibull distribution) Consider $Y_1, \ldots, Y_n$ independent and identically distributed copies of a random variable Y from the Weibull distribution with parameters λ and ν, that is

$$f_Y(y; \lambda, \nu) = \nu\lambda y^{\nu-1}\exp\{-\lambda y^\nu\}\,, \quad \lambda, \nu > 0\,, y > 0$$

Then the log-likelihood function for λ and ν is

$$l(\lambda, \nu) = -\lambda\sum_{i=1}^n Y_i^\nu + \nu\sum_{i=1}^n \log Y_i + n\log\lambda + n\log\nu\,.$$

The score function $u(\lambda, \nu)$ has components

$$\begin{aligned}\frac{\partial l(\lambda,\nu)}{\partial\lambda} &= \frac{n}{\lambda} - \sum_{i=1}^n Y_i^\nu\,,\\ \frac{\partial l(\lambda,\nu)}{\partial\nu} &= \frac{n}{\nu} + \sum_{i=1}^n \log Y_i - \lambda\sum_{i=1}^n Y_i^\nu \log Y_i\,.\end{aligned}$$

Hence the likelihood equation $\partial l(\lambda,\nu)/\partial\lambda = 0$ gives $\hat\lambda_\nu = n/\sum_{i=1}^n Y_i^\nu$, and replacing for λ in $\partial l(\lambda,\nu)/\partial\nu = 0$, gives

$$\frac{n}{\hat\nu} + \sum_{i=1}^n \log Y_i - \frac{n\sum_{i=1}^n Y_i^{\hat\nu}\log Y_i}{\sum_{i=1}^n Y_i^{\hat\nu}} = 0\,,$$

which needs to be solved numerically. Furthermore, a calculation of the components of the matrix of second derivatives gives that the observed information on λ and ν is

$$j(\lambda,\nu) = \begin{bmatrix} \dfrac{n}{\lambda^2} & \displaystyle\sum_{i=1}^n Y_i^\nu \log Y_i \\ \displaystyle\sum_{i=1}^n Y_i^\nu \log Y_i & \dfrac{n}{\nu^2} + \lambda\displaystyle\sum_{i=1}^n Y_i^\nu (\log Y_i)^2 \end{bmatrix}\,,$$

whose inverse at $(\hat\lambda, \hat\nu)$ gives and estimate of the asymptotic variance-covariance matrix of the maximum likelihood estimator.

Figure 3.3 shows the confidence regions based on $w(\lambda,\nu)$, $s^2(\lambda,\nu)$, $t^2(\lambda,\nu)$ for various confidence levels based on a simulated data set of size 20 from the Weibull distribution with $\lambda = 1$ and $\nu = 1$. For convenience, we use the versions of $s^2(\lambda,\nu)$ and $t^2(\lambda,\nu)$ which are based on the observed information. The confidence regions based on $t^2(\lambda,\nu)$ are concentric ellipsoids around the maximum likelihood estimates and their shape approximates the shape of the corresponding regions based on $w(\theta)$. Confidence regions based on $s^2(\lambda,\nu)$ have quite weird shapes for confidence levels greater or equal to 0.9. This behaviour is mainly due to the small sample size and is comparable to the behaviour of the approximate pivot $s(\beta)$ in the example of Section 1.2. The estimated coverages of the confidence regions in Table 3.3 confirm that both $w(\lambda,\nu)$ and $t^2(\lambda,\nu)$ perform considerably well, resulting in confidence regions being quite close to the nominal levels, while $s^2(\lambda,\nu)$ performs poorly undercovering the true values.

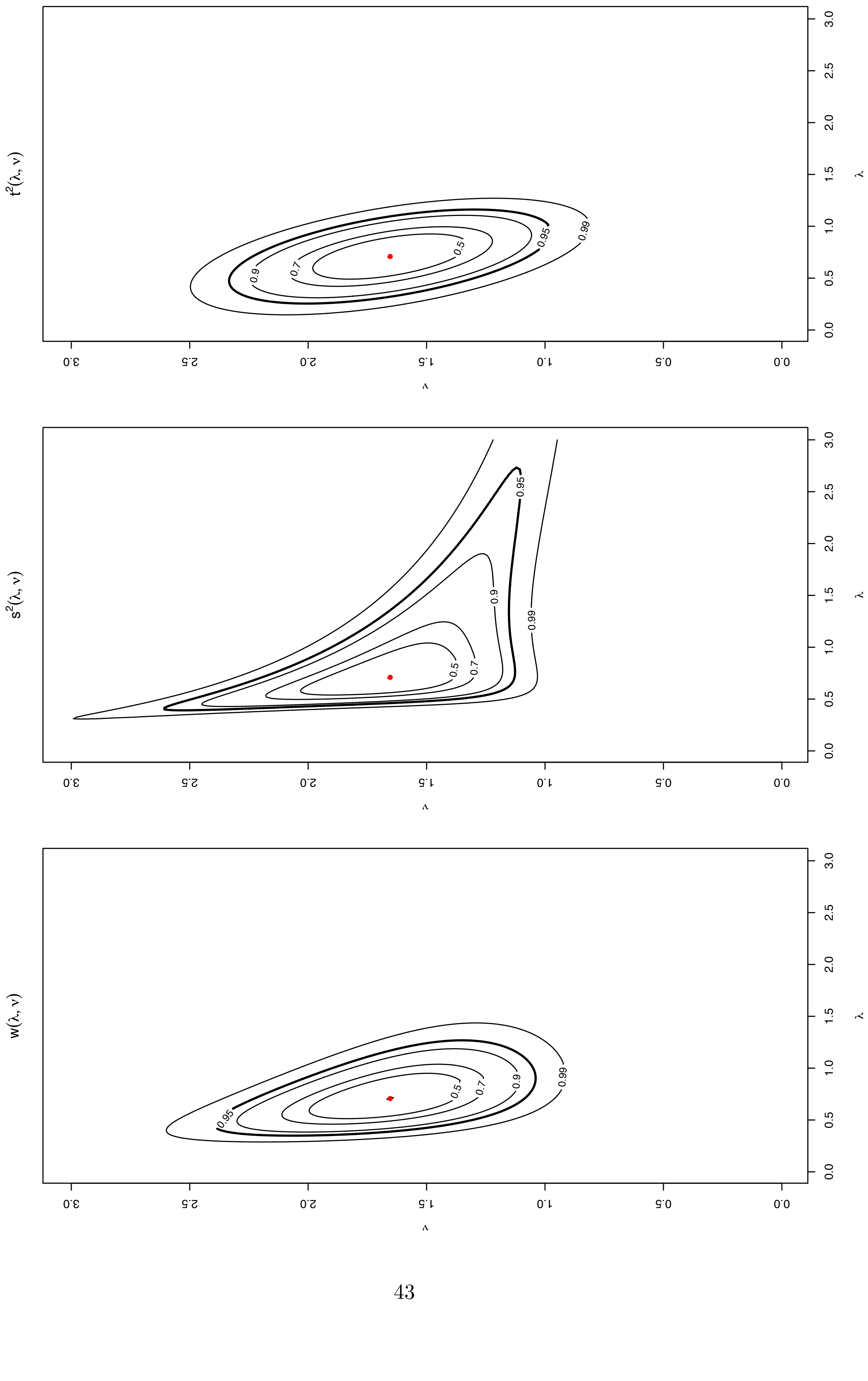


Figure 3.3: Confidence regions for λ and μ based on $w(\lambda, \nu)$, $s^2(\lambda, \nu)$, $t^2(\lambda, \nu)$ (left to right) for a simulated data set of size 20.

Table 3.3: Estimated coverage of confidence regions based on $w(\lambda,\nu)$, $s^2(\lambda,\nu)$ and $t^2(\lambda,\nu)$ when $n = 20$. The estimates are based on a simulation of size 10000 when the true parameter values are $\lambda_0 = 1$ and $\nu_0 = 1$.

Nominal	$w(\lambda,\nu)$	$t^2(\lambda,\nu)$	$s^2(\lambda,\nu)$	Simulation s.e.
0.90	0.8900	0.8938	0.8538	0.0030
0.95	0.9455	0.9428	0.9041	0.0022
0.99	0.9875	0.9816	0.9573	0.0010

3.7.2 Changes in parameterization

Suppose we transform from β by a one-to-one smooth transformation to a new parameter $\phi = g(\beta)$. Then, the inverse transformation is $\beta = g^{-1}(\phi)$ and the Jacobian of the transformation is $\partial\phi/\partial\beta$, in which the rows correspond to the components of ϕ and the columns to those of β. The log-likelihood for ϕ is $l^{(\phi)}(\phi) = l^{(\beta)}\{g^{-1}(\phi)\}$, where $l^{(\beta)}(\beta)$ is the log-likelihood on β, and $\hat\phi = g(\hat\beta)$. Nevertheless, for the score functions and the expected information matrix, direct differentiation gives the relationships

$$u^{(\phi)}(\phi) = \left(\frac{\partial\beta}{\partial\phi}\right)^T u^{(\beta)}\{g^{-1}(\phi)\}\,,$$
$$i^{(\phi)}(\phi) = \left(\frac{\partial\beta}{\partial\phi}\right)^T i^{(\beta)}\{g^{-1}(\phi)\}\left(\frac{\partial\beta}{\partial\phi}\right)\,,$$

where $u^{(\beta)}$, $i^{(\beta)}$ and $u^{(\phi)}$, $i^{(\phi)}$ are the score and the expected information in the β and ϕ parameterization, respectively. Hence, the score and the information matrix are generally affected by changes in parameterization.

Because $l^{(\phi)}(\phi) = l^{(\beta)}\{g^{-1}(\phi)\}$, the quantity $w(\beta)$ in expression (3.48) is parameterization invariant. For example, if $p = 1$ and $(-2, 3)$ is a 95% approximate confidence interval for β based on the log-likelihood ratio, then $(-1/2, 1/3)$ is a confidence interval for $\phi = 1/\beta$ based on the log-likelihood ratio. The same is not true for $t^2(\beta)$ and is true for $s^2(\beta)$ only when the expected information is used in its expression (as is done in (3.47)).

Hence, for example, from the approximate pivots of Chapter 1, $r(\beta)$ is parameterization invariant, $s(\beta)$ is parameterization invariant only if either $i(\beta)$ of $i(\hat\beta)$ are used in place of $j(\hat\beta)$ and $t(\beta)$ is not parameterization invariant.

Nevertheless, there might be a case where we are interested on constructing a confidence interval for a scalar parameter $\phi = g(\beta)$ where, for example, $g : \Re^p \to \Re$. There is a simple result called the *delta method* for this. A Taylor expansion of $g(\hat\beta)$ around β gives

$$g(\hat\beta) = g(\beta) + \{\nabla g(\beta)\}^T(\hat\beta - \beta) + O_p\left(n^{-1}\right)\,.$$

Thus,

$$\sqrt{n}\left\{g(\hat\beta) - g(\beta)\right\} = \{\nabla g(\beta)\}^T\sqrt{n}(\hat\beta - \beta) + O_p\left(n^{-1/2}\right)\,.$$

Using (3.43), an application of Slutzky's lemma for multivariate random variables gives that

$$\sqrt{n}\left\{g(\hat\beta) - g(\beta)\right\} \xrightarrow{d} N\left(0, \sigma^2(\beta)\right)\,,$$

Table 3.4: The effect of various doses of carbon disulfide on beetles.

logDose	1.691	1.724	1.755	1.784	1.811	1.837	1.861	1.884
killed	6	13	18	28	52	53	61	60
total	59	60	62	56	63	59	62	60

where $\sigma^2(\beta)/n = \{\nabla g(\beta)\}^T \{i(\beta)\}^{-1} \nabla g(\beta)$. Hence, an approximate $100(1-\alpha)\%$ confidence interval for ϕ is

$$\left\{ g(\hat\beta) - z_{1-a/2}\sqrt{\frac{\sigma^2(\hat\beta)}{n}}\, , \, g(\hat\beta) + z_{1-a/2}\sqrt{\frac{\sigma^2(\hat\beta)}{n}} \right\} . \tag{3.49}$$

Corresponding results can be obtained when $g : \Re^m \to \Re^p$.

Example 3.7. (A confidence interval for ED-50) Consider $Y_1, \ldots, Y_N$ independent binomial random variables with totals $m_1, \ldots, m_N$ and probabilities $\pi_1, \ldots, \pi_N$, respectively. Furthermore, suppose that the probabilities are linked to parameters α and γ through the relationship

$$h(\pi_i) = \alpha + \gamma x_i \quad (i = 1, \ldots, N)\, ,$$

where $h : (0, 1) \to \Re$ and invertible. Models of this form are usually used to model the effectiveness of a given drug where x_i is a given dose (or some function of a given dose) for which a binomial experiment took place. A parameter of interest in such experiments is the ED-50 (effective dose 50) which is the dose for which the drug is effective in 50% of the cases according to the binomial regression model. If $\hat\alpha$ and $\hat\gamma$ are the maximum likelihood estimators of α and γ then the estimator of ED-50 is $\hat\phi = (h(0.5) - \hat\alpha)/\hat\gamma$. In this case,

$$g(\alpha, \gamma) = \frac{h(0.5) - \alpha}{\gamma}\, ,$$

and

$$\nabla g(\alpha, \gamma) = \left(-\frac{1}{\gamma}, -\frac{h(0.5) - \alpha}{\gamma^2} \right) .$$

Hence by the delta method, a simple substitution in (3.49) gives an approximate $100(1-\alpha)\%$ confidence interval for ϕ.

Suppose that $h(\pi) = \log(-\log(1 - \pi))$ (complementary log-log link) and that we wish to estimate the ED-50 for the beetles mortality data given in Table 3.4 (Agresti, 2002, Table 6.14). According to the above, a simple calculation gives that $\hat\phi = 1.779$ with estimated standard error 0.004. Now, suppose that the $\hat\alpha$ and $\hat\gamma$ are the true parameter values, so that $\hat\phi = 1.779$ is also the true value of ED-50. From a simulation of size 10000, the estimated coverages of the approximate confidence interval (3.49) at nominal levels 0.90, 0.95 and 0.99 are 0.9017, 0.9512 and 0.9906 (the simulation standard errors are 0.0030, 0.0022 and 0.0030, respectively), which illustrates the usefulness of the delta method.

3.8 q parameters of interest, $p - q$ nuisances

3.8.1 Profile likelihood

Cases like that of Example 3.7, where interest lies on inference for some function $\psi = g(\beta)$ rather than directly on β, are very common in statistical practice. While the delta method can be useful in this respect, it only allows for Wald-type inferences for ψ, which behave poorly for finite n when the distribution of $g(\hat\beta)$ is far from normal.

The concept of profile likelihood can be used to produce inferences for ψ from extensions of the quantities $w(\beta)$ and $s^2(\beta)$. The profile likelihood $L_p(\psi)$ for ψ is defined as

$$L_p(\psi) = \sup_{\{\beta : g(\beta) = \psi\}} L(\beta)\,,$$

that is the maximum is taken over all β which are consistent with a given value of ψ. In many cases and possibly after suitable reparameterization, ψ may be defined as a component of a given partitioning $\beta = (\psi, \lambda)$ of β into subvectors ψ and λ of dimensions q and $p - q$ respectively, where λ is the vector of nuisance parameters. In this case the profile log-likelihood for ψ is

$$l_p(\psi) = l(\psi, \hat\lambda_\psi)\,,$$

where $\hat\lambda_\psi$ denotes the *constrained* maximum likelihood estimator of λ for a fixed value of ψ.

The profile likelihood is not a genuine likelihood, that is it does not generally correspond to a probability distribution. Nevertheless, it has some special features that enable, to a considerable extend, to think of it as if it were a genuine likelihood:

- A simple calculation gives

  $$\sup_\psi l_p(\psi) = \sup_\psi \sup_\lambda l(\psi, \lambda) = l(\hat\psi, \hat\lambda)\,.$$

  Hence, the maximum profile likelihood estimator $\hat\psi$ of ψ is equal to the ψ partition of the maximum likelihood estimator $\hat\beta$.

- If we define

  $$w_p(\psi) = 2\{l_p(\hat\psi) - l_p(\psi)\}\,,$$

  then $w_p(\psi) = 2\{l(\hat\psi, \hat\lambda) - l(\psi, \hat\lambda)\}$. This latter form is used to perform a log-likelihood ratio test for a hypothesis on ψ when λ is unknown and similar arguments as in the previous sections can be used to show that $w_p(\psi) \xrightarrow{d} \chi^2_q$. Hence, profile confidence regions can be constructed as

  $$\{\psi : w_p(\psi) \le \chi^2_{q,1-\alpha}\}\,. \tag{3.50}$$

- The score function of $\beta = (\psi, \lambda)$ can be partitioned as $u(\beta) = \{l_\psi(\psi, \lambda), l_\lambda(\psi, \lambda)\}$, where $l_\psi(\psi, \lambda) = \nabla_\psi l(\psi, \lambda)$ and $l_\lambda(\psi, \lambda) = \nabla_\lambda l(\psi, \lambda)$. Correspondingly, the full observed information $j(\beta) = j(\psi, \lambda)$ and its inverse can be partitioned as

  $$j(\psi, \lambda) = \begin{bmatrix} j_{\psi\psi}(\psi, \lambda) & j_{\psi\lambda}(\psi, \lambda) \\ j_{\lambda\psi}(\psi, \lambda) & j_{\lambda\lambda}(\psi, \lambda) \end{bmatrix}$$
  $$\{j(\psi, \lambda)\}^{-1} = \begin{bmatrix} j^{\psi\psi}(\psi, \lambda) & j^{\psi\lambda}(\psi, \lambda) \\ j^{\lambda\psi}(\psi, \lambda) & j^{\lambda\lambda}(\psi, \lambda) \end{bmatrix}.$$

Then

$$\begin{aligned}\frac{\partial l_p(\psi)}{\partial \psi} &= l_\psi(\psi, \hat{\lambda}_\psi) + l_\lambda(\psi, \hat{\lambda}_\psi)\frac{\partial \hat{\lambda}_\psi}{\partial \psi} \\ &= l_\psi(\psi, \hat{\lambda}_\psi)\,,\end{aligned}$$

and

$$j_p(\psi) = -\frac{\partial l_p(\psi, \hat{\lambda}_\psi)}{\partial \psi \psi^T} = -l_{\psi\psi}(\psi, \hat{\lambda}_\psi) - l_{\psi\lambda}(\psi, \hat{\lambda}_\psi)\frac{\partial \hat{\lambda}_\psi}{\partial \psi}\,, \tag{3.51}$$

where $j_p(\psi)$ is the profile observed information and $l_{\psi\psi}(\psi, \lambda) = \nabla_\psi \nabla_\psi^T l(\psi, \lambda)$. On the other hand, differentiating both sides of $0 = l_\lambda(\psi, \hat{\lambda}_\psi)$ with respect to ψ gives

$$\frac{\partial \hat{\lambda}_\psi}{\partial \psi} = -\{l_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)\}^{-1} l_{\lambda\psi}(\psi, \hat{\lambda}_\psi)\,.$$

Substituting in (3.51) gives

$$j_p(\psi) = -\{l_{\psi\psi} - l_{\psi\lambda}(l_{\lambda\lambda})^{-1} l_{\lambda\psi}\}\,,$$

where all the derivatives are evaluated at $(\psi, \hat{\lambda}_\psi)$. By standard linear algebra rules for the inversion of blocked matrices, the later expression gives the important identity

$$\{j_p(\psi)\}^{-1} = j^{\psi\psi}(\psi, \hat{\lambda}_\psi)\,,$$

that is the inverse of the profile observed information is equal to the ψ-block of the inverse of the full information evaluated at $(\psi, \hat{\lambda}_\psi)$. Hence, as $\{j(\hat{\beta})\}^{-1}$ is an estimate of the asymptotic variance-covariance matrix of $\hat{\beta}$ (which is $\{i(\beta_0)\}^{-1}$) in the case with no nuisances, $\{j_p(\hat{\psi})\}^{-1}$ can be used as an estimate of the asymptotic variance-covariance matrix of $\hat{\psi}$. In fact, $\hat{\psi}$ has asymptotically a q-variate normal distribution with mean ψ_0 and variance-covariance matrix $i^{\psi\psi}(\psi_0, \lambda_{\psi_0})$, where $i^{\psi\psi}$ has corresponding meaning to $j^{\psi\psi}$. Thus, we can define the combinants

$$s_p^2(\psi) = \{l_\psi(\psi, \hat{\lambda}_\psi)\}^T j^{\psi\psi}(\psi, \hat{\lambda}_\psi) l_\psi(\psi, \hat{\lambda}_\psi)\,, \tag{3.52}$$

$$t_p^2(\psi) = (\hat{\psi} - \psi)^T \{j^{\psi\psi}(\psi, \hat{\lambda}_\psi)\}^{-1} (\hat{\psi} - \psi)\,, \tag{3.53}$$

which are both asymptotically distributed according to a χ^2_q distribution and can be used in the position of $w_p(\psi)$ in (3.50) for the construction of approximate confidence regions. Furthermore, in both 3.52 and 3.53 the inverse of the profile observed information $j^{\psi\psi}(\psi, \hat{\lambda}_\psi)$ can be replaced by either $j^{\psi\psi}(\hat{\beta})$ (since $\hat{\beta} = (\hat{\psi}, \hat{\lambda}_{\hat{\psi}})$), $i^{\psi\psi}(\psi, \hat{\lambda}_\psi)$ or $i^{\psi\psi}(\hat{\beta})$, without affecting the limiting χ^2_q distribution.

- Lastly, by similar arguments as in the case of no nuisances, $w_p(\psi)$ is invariant under interest-respecting reparameterizations (that is the new parameter of interest is an one-to-one function of ψ while the new nuisance parameters can be functions of both ψ and λ). The same is true for $s_p^2(\psi)$ only when $i^{\psi\psi}(\psi, \hat{\lambda}_\psi)$ or $i^{\psi\psi}(\hat{\beta})$ are used in its definition, and $t_p^2(\psi)$ is not invariant under reparameterization.

If the parameter of interest ψ is a scalar then we can define the approximate pivots $r_p(\psi)$ as the signed square root of $w_p(\psi)$ and $s_p(\psi)$, $t_p(\psi)$ are the square roots of $s_p^2(\psi)$ and $t_p^2(\psi)$.

Example 3.8. (Weibull distribution - continued) Continuing from Example 3.6, we have seen that $\hat{\lambda}_\nu = n/\sum_{i=1}^n Y_i^\nu$ and hence omitting any quantities that do not depend on the parameters, the profile log-likelihood for ν is

$$l_p(\nu) = n \log \nu - n \log \sum_{i=1}^n Y_i^\nu + \nu \sum_{i=1}^n \log Y_i \,.$$

On the other hand, if the parameter of interest is λ then the constraint maximum likelihood estimator $\hat{\nu}_\lambda$ of the nuisance ν is not defined explicitly and hence it has to be computed numerically for each value of λ. The profile log-likelihood for λ is

$$l_p(\lambda) = l(\lambda, \hat{\nu}_\lambda)\,.$$

The left plot of figure 3.4 shows he contours of $l(\lambda, \nu)$ for the simulated data set of Example 3.6. The traces of $l_p(\lambda)$ and $l_p(\nu)$ are also shown using the dashed blue and red lines respectively (the trace of $l_p(\lambda)$ is the curve $(\lambda, \hat{\nu}_\lambda)$). Both traces pass through the point $(\hat{\lambda}, \hat{\nu})$. This demonstrates that the maximum profile likelihood estimates are the same as the maximum likelihood estimates. The middle and right plots show $l_p(\lambda)$ and $l_p(\nu)$ and the dashed horizontal lines are at $l_p(\hat{\lambda}) - \chi^2_{1,0.95}/2$ and $l_p(\hat{\nu}) - \chi^2_{1,0.95}/2$, respectively. Those lines define the approximate 95% confidence intervals for λ and ν based on $w_p(\lambda)$ and $w_p(\nu)$, respectively. The confidence interval for ν is $(1.151, 2.240)$ and for λ is $(0.406, 1.139)$. $\square$

Despite the aforementioned nice properties of the profile likelihood, $l_p(\psi)$ is not a genuine log-likelihood and the profile score function $\partial l_p(\psi)/\partial\psi$ does not generally have zero null expectation. The use of $l_p(\psi)$ is the same as regarding λ known and equal to $\hat{\lambda}_\psi$ and this is certainly unreasonable if the data do not contain enough information on λ, which usually is the case when the dimension of λ is large. More formally, if the dimension of the nuisance parameters is fixed then the null expectation of the profile score for ψ is of order $O(1)$, whereas, when the dimension of λ increases with n, then that expectation is of order $O(n)$ and $\hat{\psi}$ is then inconsistent. Various modifications of the profile likelihood have been proposed in the literature that can, at varying degrees, compensate for the limited knowledge on ψ. The simplest of those is the approximate conditional log-likelihood (Cox and Reid, 1987) which takes the form

$$l_a(\psi) = l_p(\psi) - \frac{1}{2}\log\left|j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)\right| \,,$$

and has been derived for the case of *orthogonal parameterizations* as an approximation of the log-likelihood based on the distribution of Y given $\hat{\lambda}$. Unfortunately, $l_a(\psi)$ is not invariant under interest-respecting reparameterizations and the addition of an extra, in general rather complicated, term on the right hand side is required to achieve this. When the parameterization is orthogonal the contribution of this extra term becomes small. Nevertheless, plotting both $l_a(\psi)$ and $l_p(\psi)$ could be a first step to investigate the effect of the estimation of the nuisance parameters on inferences.

3.8.2 Orthogonal parameterization

If for a particular model, it was possible to write

$$l(\beta) = l_1(\psi) + l_2(\lambda)\,,$$

then $\hat\lambda_\psi = \hat\lambda$, $l_p(\psi) = l_1(\psi)$ and $j_p(\psi) = -\partial l_1(\psi)/\partial\psi\partial\psi^T$ and inferences on ψ would be very convenient. We can get close to this situation by using an *orthogonal parameterization.* The parameters ψ and λ are called orthogonal if $i_{\psi\lambda}(\psi, \lambda) = 0$.

The main consequence of parameter orthogonality is that $\hat\psi$ and $\hat\lambda$ are asymptotically closer to independence. To justify this statement consider the case $p = 2$ and $q = 1$. An expansion of $l(\psi, \lambda)$ around $(\hat\psi, \hat\lambda)$ gives

$$\begin{aligned} l(\psi, \lambda) &= l(\hat\psi, \hat\lambda) + (\psi - \hat\psi) l_\psi(\hat\psi, \hat\lambda) + (\lambda - \hat\lambda) l_\lambda(\hat\psi, \hat\lambda) \\ &\quad + \frac{1}{2}\left\{(\psi - \hat\psi)^2 l_{\psi\psi}(\hat\psi, \hat\lambda) + 2(\psi - \hat\psi)(\lambda - \hat\lambda) l_{\psi\lambda}(\hat\psi, \hat\lambda) + (\lambda - \hat\lambda)^2 l_{\lambda\lambda}(\hat\psi, \hat\lambda)\right\} \\ &\quad + O_p\left(n^{-1/2}\right) \end{aligned} \tag{3.54}$$

Now, noting that $l_\psi(\hat\psi, \hat\lambda) = 0$ and $l_\lambda(\hat\psi, \hat\lambda) = 0$ and denoting by $\hat j_{\psi\lambda} \equiv j_{\psi\lambda}(\hat\psi, \hat\lambda)$, $\hat j_{\lambda\lambda} \equiv j_{\lambda\lambda}(\hat\psi, \hat\lambda)$, etc., direct differentiation of (3.54) with respect to λ gives

$$l_\lambda(\psi, \lambda) = -(\psi - \hat\psi)\hat j_{\psi\lambda} - (\lambda - \hat\lambda)\hat j_{\lambda\lambda} + O_p(1) \, .$$

Because $l_\lambda(\psi, \hat\lambda_\psi) = 0$, we get

$$\hat\lambda_\psi - \hat\lambda = (\hat\psi - \psi)\frac{\hat j_{\psi\lambda}}{\hat j_{\lambda\lambda}} + O_p\left(n^{-1}\right) \, .$$

But $j(\psi, \lambda) = i(\psi, \lambda) + O_p\left(n^{1/2}\right)$ and so the above expression can be written as

$$\hat\lambda_\psi - \hat\lambda = (\hat\psi - \psi)\frac{\hat i_{\psi\lambda}}{\hat i_{\lambda\lambda}} + O_p\left(n^{-1}\right) \, ,$$

which becomes $\hat\lambda_\psi - \hat\lambda = O_p(n^{-1})$ when ψ and λ are orthogonal. Thus when ψ and λ are orthogonal $\hat\lambda_\psi$ varies only slowly in ψ in a neighbourhood of $\hat\psi$. Furthermore, by (3.54), when ψ and λ are orthogonal, we have

$$\begin{aligned} l(\psi, \lambda) &= c - \frac{1}{2}(\psi - \hat\psi)\hat j_{\psi\psi} - \frac{1}{2}(\lambda - \hat\lambda)\hat j_{\lambda\lambda} + O_p\left(n^{-1/2}\right) \\ &= l_1(\psi) + l_2(\lambda) + O_p\left(n^{-1/2}\right) \, , \end{aligned}$$

because $c = l(\hat\psi, \hat\lambda)$ is parameter constant and thus can be absorbed in either function $l_1(\psi)$ or the function $l_2(\psi)$. Hence if ψ and λ are orthogonal, for (ψ, λ) in a neighbourhood of $(\hat\psi, \hat\lambda)$ (of radius $O_p\left(n^{-1/2}\right)$) the log-likelihood is almost separable.

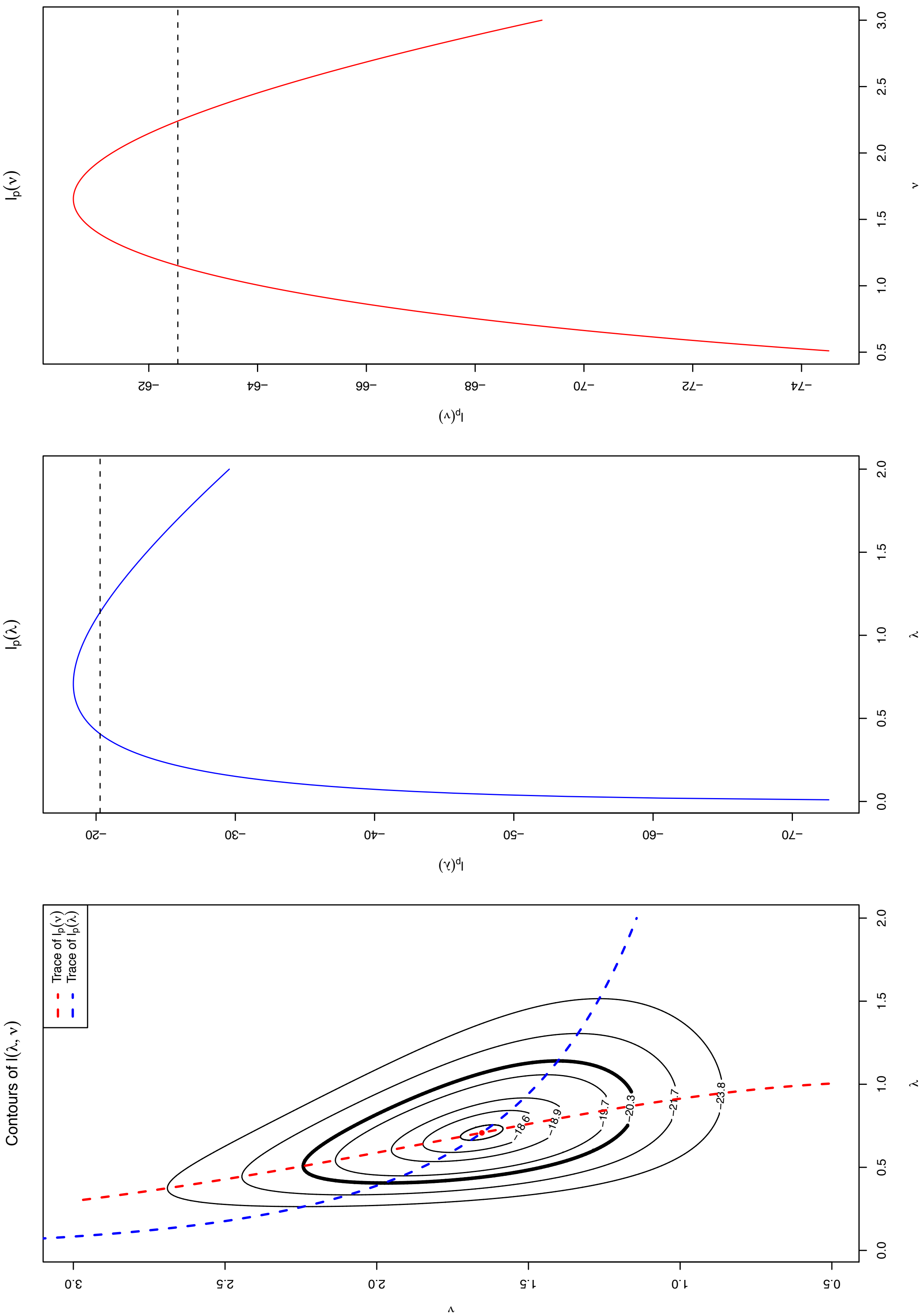


Figure 3.4: The contours of $l(\lambda, \nu)$ for a simulated data set of size 20 along with the traces of $l_p(\lambda)$ and $l_p(\nu)$ (left). The profile log-likelihood for λ (middle). The dashed horizontal line is at $l_p(\hat{\lambda}) - \chi^2_{1,0.95}/2$ and defines the approximate 95% confidence interval for λ based on $w_p(\lambda)$. The plot on the right is that of $l_p(\nu)$.

Bibliography

Agresti, A. (2002). *Categorical Data Analysis.* New York: Wiley.

Anscombe (1956). On estimating binomial response relations. *Biometrika 43*(3), 461–464.

Barndorff-Nielsen, O. E. and D. R. Cox (1989). *Asymptotic Techniques for Use in Statistics.* London: Chapman & Hall Ltd.

Bartlett, M. (1953). Approximate confidence intervals. ii. More than one uknown parameter. *Biometrika 40*, 306–317.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Journal of the Royal Statistical Society, Series A: General 160*, 268–282.

Bishop, Y. M., S. E. Fienberg, and P. W. Holland (1975). *Discrete Multivariate Analysis : Theory and Practice.* Cambridge, Massachutetts: M.I.T. Press.

Brazzale, A. R., A. C. Davison, and N. Reid (2007). *Applied asymptotics: case studies in small-sample statistics.* Cambridge University Press.

Cox, D. R. (2006). *Principles of Statistical Inference.* Cambridge University Press.

Cox, D. R. and D. V. Hinkley (1974). *Theoretical Statistics.* London: Chapman & Hall Ltd.

Cox, D. R. and N. Reid (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society, Series B: Methodological 49*, 1–18.

Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency) (with discussion). *The Annals of Statistics 3*, 1189–1217.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika 80*(1), 27–38.

Frydenberg, M. and J. L. Jensen (1989). Is the 'improved likelihood ratio statistic' really improved in the discrete case? *Biometrika 76*, 655–661.

Haldane, J. (1955). The estimation of the logarithm of a ratio of frequencies. *Annals of Human Genetics 20*, 309–311.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London 186*(1007), 453–461.

Mann, H. B. and A. Wald (1943, sep). On stochastic limit and order relationships. *The Annals of Mathematical Statistics 14*(3), 217–226.

McCullagh, P. and J. Nelder (1989). *Generalized Linear Models* (2nd ed.). London: Chapman and Hall.

Pace, L. and A. Salvan (1997). *Principles of Statistical Inference: From a Neo-Fisherian Perspective*. London: World Scientific.

Skovgaard, I. (1986). A note on the differentiation of cumulants of log-likelihood derivatives. *International Statistical Review 54*, 29–32.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics 20*(4), 595–601.

Young, A. G. and R. L. Smith (2005). *Essentials of Statistical Inference*. Cambridge, UK: Cambridge University Press.