Asymptotic Expansions

4.1 Introduction

One of the techniques used in the latest case study (Section 3.7) is an asymptotic expansion of an estimator as well as that for a log-likelihood function. The most well-known asymptotic expansion is the Taylor expansion, which is a mathematical tool, rather than a statistical method. However, the method is used so extensively in both theoretical and applied statistics that its role in statistics can hardly be overstated. Several other expansions, including the Edgeworth expansion and Laplace approximation, can be derived from the Taylor expansion. It should be pointed out that some elementary expansions can also be very useful (see Section 4.5).

Asymptotic expansions are extremely helpful in cases where the quantities of interest do not have closed-form expressions. Sometimes, even when the quantity does have a closed form, an asymptotic expansion may still be useful in simplifying the expression and revealing the dominant factor(s). For example, consider the following.

**Example 4.1** (Variance estimation in linear regression). A multiple linear regression model may be expressed as

\[ Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \tag{4.1} \]

where \( i = 1, \ldots, n \), where \( x_{i1}, \ldots, x_{ip} \) are known covariate, \( \beta_0, \ldots, \beta_p \) are unknown regression coefficients, and \( \epsilon_i \) is a random error. Here, we assume that \( \epsilon_1, \ldots, \epsilon_n \) are independent and distributed as \( N(0, \sigma^2) \), where \( \sigma^2 \) is an unknown variance.

The error variance is typically estimated by the unbiased estimator

\[ \hat{\sigma}^2 = \frac{\text{RSS}}{n - p - 1}, \]

where RSS represents the residual sum of squares, \( \sum_{i=1}^{n} \hat{\epsilon}_i^2 \), where \( \hat{\epsilon}_i = Y_i - \hat{Y}_i \), as the fitted value given by \( \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip} \), and \( \hat{\beta} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)' \).
is the least squares estimator of $\beta = (\beta_0, \ldots, \beta_p)'$ given by

$$\hat{\beta} = (X'X)^{-1}X'Y.$$  

(4.2)

In (4.2), $Y$ is the vector of observations, $Y = (Y_1)_{1 \leq i \leq n}$, and $X$ the matrix of covariates, $X = (x_{ij})_{1 \leq i \leq n, 1 \leq j \leq p}$, with $x_{i0} = 1$, $1 \leq i \leq n$, corresponding to the intercept. Using this notation, (4.1) can be expressed as:

$$Y = X\beta + \epsilon,$$  

(4.3)

where $\epsilon = (\epsilon_i)_{1 \leq i \leq n}$. Furthermore, the residuals and RSS can be expressed in terms of a project matrix $P_{X^\perp} = I_n - P_X$, where $P_X = X(X'X)^{-1}X'$; that is, $\hat{\epsilon} = \epsilon_i_{1 \leq i \leq n} = P_{X^\perp}Y$ and RSS $= \|\hat{\epsilon}\|^2 = Y'P_{X^\perp}Y$. Here, we assume, for simplicity, that $X$ is of full (column) rank, but similar expressions can be obtained even without this restriction. See Appendix A.1 for the definition and properties of projection matrices. It follows that

$$\sigma^2 = \frac{Y'P_{X^\perp}Y}{n - p - 1}.$$  

Alternatively, since normality is assumed, $\sigma^2$ may be estimated by the MLE, which can be expressed as

$$\hat{\sigma}^2 = \frac{Y'P_{X^\perp}Y}{n}.$$  

Both estimators have closed-form expressions, although the expression for $\hat{\sigma}^2$ is even simpler. On the other hand, an asymptotic expansion shows the close relation between the two estimators in terms of decreasing orders. Note that

$$\frac{1}{1 - x} = 1 + x + x^2 + \cdots$$  

(4.4)

for any $0 \leq x < 1$. Then, for $n > p + 1$, we have

$$\left(1 - \frac{p + 1}{n}\right)^{-1} = 1 + \frac{p + 1}{n} + \frac{(p + 1)^2}{n^2} + \cdots$$

$$= 1 + \frac{p + 1}{n} + O\left(\frac{1}{n^2}\right),$$  

(4.5)

provided that $n \rightarrow \infty$ and $p$ is bounded.

Expansion (4.5) implies the following connection between $\sigma^2$ and $\hat{\sigma}^2$:

$$\sigma^2 = \left(1 - \frac{p + 1}{n}\right)^{-1} \hat{\sigma}^2$$

$$= \hat{\sigma}^2 + \left(\frac{p + 1}{n}\right) \hat{\sigma}^2 + O_P\left(\frac{1}{n^2}\right).$$  

(4.6)

The reason that the remaining terms is $O_P(n^{-2})$ is because $E(\|\hat{\epsilon}\|^2) = \sigma^2(n - p - 1)$. This follows from the unbiasedness of $\hat{\sigma}^2$; alternatively, it can also be derived using the following simple arguments:

$$E(\|\hat{\epsilon}\|^2) = E(Y'P_{X^\perp}Y)$$

$$= E\{Y - X\beta)'P_{X^\perp}(Y - X\beta)\}$$

$$= E\{[tr\{Y - X\beta)'P_{X^\perp}(Y - X\beta)\}]\}$$

$$= tr[E\{P_{X^\perp}(Y - X\beta)(Y - X\beta)'\}]$$

$$= tr[P_{X^\perp}E\{Y - X\beta)(Y - X\beta)'\}]$$

$$= \sigma^2 tr(P_{X^\perp})$$

$$= \sigma^2(n - p - 1).$$

Here, we used the facts that one can exchange the order of trace and expectation and that $E\{Y(X\beta)(Y - X\beta)'\} = Var(Y) = \sigma^2 I_n$. It follows from Theorem 3.1 that $\|\hat{\epsilon}\|^2 = O_P(n)$; hence, $\hat{\sigma}^2 = O_P(1)$. Equation (4.6) shows that $\hat{\sigma}^2$ is the leading $[O_P(1)]$ term in an expansion of $\sigma^2$. It also shows that the next term in the expansion is $(p + 1)/n)\sigma^2$, which is $O_P(n^{-1})$, and the next term is $O_P(n^{-2})$, and so forth. Even though (4.6) is derived as a large-sample approximation, assuming that $n \rightarrow \infty$ while $p$ remains fixed or bounded, it can also be useful under a finite-sample consideration. For example, it shows that if the number of covariates, $p$, is comparable to the sample size [i.e., if the ratio $(p + 1)/n$ is not very small], the difference between the two methods of variance estimation can be nontrivial. In fact, the latter is the reason for the failure of consistency of the MLE in Example 3 of the Preface. See Chapter 5 for a further discussion.

4.2 Taylor expansion

As the author's view that the Taylor expansion is the single most useful mathematical tool for a statistician. We begin by revisiting (4.4). There is more than one way to derive this identity, one of which is to use the Taylor expansion. First, compute the derivatives of $f(x) = (1 - x)^{-1}$. We have $f'(x) = (x - a)^{-2}$, $f''(x) = 2(1 - x)^{-3}$, $f'''(x) = 6(1 - x)^{-4}$, and so on. Thus, we obtain (4.4) as the Taylor series at $a = 0$

$$f(x) = f(a) + f'(a)(x - a) + \frac{f''(a)}{2}(x - a)^2 + \frac{f'''(a)}{6}(x - a)^3 + \cdots$$

$$= \sum_{k=0}^{\infty} \frac{f^{(k)}(a)}{k!}(x - a)^k,$$

(4.7)

where $f^{(k)}(x)$ represents the $k$th derivative with $f^{(0)}(x) = f(x)$.
The formal statement of the Taylor expansion is the following.

**Theorem 4.1** (Taylor’s theorem). Suppose that the $n$th derivative of $f(x)$ is continuous on $[a, b]$ and the $(l+1)$st derivative of $f(x)$ exists on $(a, b)$. Then for any $x \in [a, b]$, we have

$$f(x) = f(a) + f'(a)(x-a) + \cdots + \frac{f^{(l)}(a)}{l!}(x-a)^l + \frac{f^{(l+1)}(c)}{(l+1)!}(x-a)^{l+1}$$

$$= \sum_{k=0}^{l} \frac{f^{(k)}(a)}{k!}(x-a)^k + \frac{f^{(l+1)}(c)}{(l+1)!}(x-a)^{l+1},$$

where $a < c < x$; that is, $z = (1-t)a + tx$ for some $t \in [0, 1]$.

If, in particular, $f^{(k)}(x)$ exists for all $k$ and the last term on the right side of (4.8) $\to 0$ as $n \to \infty$, then (4.7) holds, which is called the Maclaurin’s series. In the special case of $a = 0$, the Taylor series is also known as the Maclaurin’s series. For example, in addition to (4.4), we have

$$\frac{1}{1+x} = 1 - x + x^2 - x^3 + x^4 - \cdots,$$

$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \cdots,$$

$$\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots,$$

$$\sin(x) = x - \frac{x^3}{6} + \frac{x^5}{120} - \frac{x^7}{7!} + \cdots,$$

$$\cos(x) = 1 - \frac{x^2}{2} + \frac{x^4}{24} - \frac{x^6}{720} + \cdots.$$

English mathematician Brook Taylor (1685–1731) published a general method for constructing the Taylor series (which are now named after him) in 1715, although various forms of special cases were known much earlier. Colin Maclaurin, a Scottish mathematician who was once a professor in Edinburgh, published the special case of the Taylor series in the 18th century.

It should be pointed out that the Taylor expansion is a local property of a function. This means that the closer $x$ is to $a$ the more accurate the approximation. We illustrate this with an example.

**Example 4.2.** Consider the accuracy of the Maclaurin expansion for the function $f(x) = e^x$. The Taylor (Maclaurin) series for $e^x$ is given above. Table 4.1 shows the approximations using the first $n$ terms in the series, where the relative error is computed as the absolute value of the approximation error divided by the true value. It is clear that the approximation is much more accurate for $x = 1$ than for $x = 5$. This is because the expansion is at $x = 0$, $z = 1$ is much closer to zero than $x = 5$. However, as long as $n$ is large enough, the same accuracy will be achieved for any $x$ (Exercise 4.1).

A multivariate extension of the Taylor expansion is perhaps even more useful in practice. To illustrate the results, we define a linear differential operator as follows. Let $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, and $\nabla$ denote the gradient operator, or vector differential operator, defined by $\nabla = (\partial/\partial x_1, \ldots, \partial/\partial x_n)$. Consider the linear differential operator

$$x' \nabla = \sum_{i=1}^{n} x_i \frac{\partial}{\partial x_i}.$$

Note that $(x' \nabla)^k$ can be operated in a similar way as the $k$th power of a sum. For example, with $s = 2$, we have

$$(x' \nabla)^2 = \left( x_1 \frac{\partial}{\partial x_1} + x_2 \frac{\partial}{\partial x_2} \right)^2$$

$$= x_1^2 \frac{\partial^2}{\partial x_1^2} + 2x_1x_2 \frac{\partial^2}{\partial x_1 \partial x_2} + x_2^2 \frac{\partial^2}{\partial x_2^2};$$

$$(x' \nabla)^3 = \left( x_1 \frac{\partial}{\partial x_1} + x_2 \frac{\partial}{\partial x_2} \right)^3$$

$$= x_1^3 \frac{\partial^3}{\partial x_1^3} + 3x_1^2x_2 \frac{\partial^3}{\partial x_1^2 \partial x_2} + 3x_1x_2^2 \frac{\partial^3}{\partial x_1 \partial x_2^2} + x_2^3 \frac{\partial^3}{\partial x_2^3};$$

and so on. The multivariate Taylor expansion, or the Taylor expansion in several variables, can be stated as follows.

**Theorem 4.2** (Multivariate Taylor expansion). Let $f : D \to \mathbb{R}$, where $D \subseteq \mathbb{R}^n$. Suppose that there is a neighborhood of $a$, $S_\delta(a) \subseteq D$ such that $f$ and its up to $(l+1)$st partial derivatives are continuous in $S_\delta(a)$. Then, for each $x \in S_\delta(a)$, we have

$$f(x) = f(a) + \sum_{k=1}^{l} \frac{1}{k!} (x-a)^{\nabla^k} f(a)$$

and
\[ f(x) = f(a) + \sum_{k=1}^{I} \frac{1}{k!} \{(x - a)^\ell \nabla\}^k f(a) + O(\varepsilon^{1/2}). \]

where \(z = (1 - t)a + tz\) for some \(t \in [0, 1]\).

In the author's experience, the (multivariate) Taylor expansion of second or third orders are most useful in practice. For such expansions, there is an alternative expression that may be more interpretable and easier to use. Let

\[ \frac{\partial^2 f(x)}{\partial x^j \partial x^k} = \left[ \frac{\partial^2 f(x)}{\partial x^i \partial x^j} \right]_{1 \leq i, j \leq s}, \]

(4.11)

Note that (4.10) is the transpose of the gradient vector, or \(\nabla f(x)\)', and (4.11) is the matrix of second derivatives, or Hessian matrix. Then the second-order Taylor expansion can be expressed as

\[ f(x) = f(a) + \frac{\partial f(a)}{\partial x'} (x - a) + \frac{1}{2} (x - a)' \frac{\partial^2 f(a)}{\partial x \partial x'} (x - a), \]

where \(z = (1 - t)a + tz\) for some \(t \in [0, 1]\). Equation (4.12) shows that, locally (i.e., in a neighborhood of \(a\)), \(f(x)\) can be approximated by a quadratic function. For example, suppose that \(a\) is a point such that \(\partial f(a)/\partial x = \partial f(a)/\partial x' = 0\). Furthermore, suppose that, locally, the Hessian matrix of \(f(x)\) is positive definite. It follows from (4.12) that \(f(x) > f(a)\) near \(a\) and, hence, has a unique local minimum at \(x = a\). Similarly, the third-order Taylor expansion can be expressed as

\[ \tilde{f}(x) = f(a) + \frac{\partial f(a)}{\partial x'} (x - a) + \frac{1}{2} (x - a)' \frac{\partial^2 f(a)}{\partial x \partial x'} (x - a) + \frac{1}{6} \left[ (x - a)' \frac{\partial^2 f(a)}{\partial x \partial x \partial x'} (x - a) \right]_1 \leq i, j \leq s, \]

(4.13)

where \(z = (1 - t)a + tz\) for some \(t \in [0, 1]\). Note that in a small neighborhood of \(a\), the third-order term (i.e., the last term) in (4.13) is dominated by the leading quadratic function.

The last term in the Taylor expansion (i.e., the term that involves \(z\), is called the remaining term. This term is sometimes expressed in terms of a small \(o\) or big \(O\). For example, suppose that all of the \((l + 1)st\) partial derivatives of \(f(x)\) are bounded in the neighborhood \(S_0(a)\). Then (4.9) can be written as

\[ f(x) = f(a) + \sum_{k=1}^{\ell} \frac{1}{k!} \{(x - a)^\ell \nabla\}^k f(a) + O(\varepsilon^{1/2}), \]

where \(|x - a| = \{\sum_{i=1}^{s} (x_i - a_i)^2\}^{1/2}\), or

In particular, (4.13) can be expressed as

\[ f(x) = f(a) + \frac{\partial f(a)}{\partial x'} (x - a) + \frac{1}{2} (x - a)' \frac{\partial^2 f(a)}{\partial x \partial x'} (x - a) + o(|x - a|^2), \]

(4.16)

and the \(o(|x - a|^2)\) can be replaced by \(O(|x - a|^3)\). However, caution should be paid when using such an expression for a large-sample approximation, in which the function \(f(x)\) may depend on the sample size \(n\).

\[ \text{Example 4.3. Suppose that } X_1, \ldots, X_n \text{ are i.i.d. observations from the Logistic}(\theta) \text{ distribution whose pdf is given by} \]

\[ f(x; \theta) = \frac{e^{x-\theta}}{(1 + e^{x-\theta})^2}, -\infty < x < \infty. \]

Consider the second-order Taylor expansion of the log-likelihood function,

\[ l(\theta) = \sum_{i=1}^{n} \log \{f(X_i; \theta)\} \]

at the true \(\theta\), which we assume, for simplicity, to be zero. Then the Taylor expansion can be expressed as

\[ l(\theta) = l(0) + l'(0)\theta + \frac{1}{2} l''(\hat{\theta})\theta^2, \]

(4.17)

where \(\hat{\theta}\) lies between zero and \(\theta\). It can be shown (Exercise 4.2) that

\[ l(0) = -\sum_{i=1}^{n} \{X_i + 2 \log(1 + e^{-X_i})\}, \]

\[ l'(0) = \sum_{i=1}^{n} \frac{1 - e^{-X_i}}{1 + e^{-X_i}}, \]

\[ l''(\hat{\theta}) = -2 \sum_{i=1}^{n} \frac{e^{\hat{\theta} - X_i}}{(1 + e^{\hat{\theta} - X_i})^2}. \]

Furthermore, it can be shown that \(l(0) = O_P(n)\), \(l'(0) = O_P(\sqrt{n})\), and \(l''(\hat{\theta}) = O_P(n)\). Now, suppose that one wishes to study the behavior of the log-likelihood near the true value \(\theta = 0\) by considering a sequence \(\theta = t/\sqrt{n}\), where \(t\) is a constant known as the local deviation (e.g., Bickel et al. 1993, p. 17). If one blindly uses (4.15) (with \(s = 1\) and \(l = 1\), one
would have \( l(\theta) = l(0) + l'(0)\theta + O_P(\theta^2) \) [here we use \( O_P(\theta^2) \) instead of \( O(\theta^2) \) because \( l(\theta) \) is random]; hence,
\[
    l(\theta_n) = l(0) + l'(0)\theta_n + O_P(\theta_n^2).
\]
(4.18)
The first term on the right side of (4.18) is \( O_P(n) \), the second term is \( O_P(n^{1/2}(t/\sqrt{n}) = O_P(1) \), and the third terms appears to be \( O_P(n^{-1}) \). This suggests that the third term is negligible because it is of lower order than the second term. However, this is not true because the more accurate expression (4.17) (with \( \theta = \theta_n \)) shows that the third term is \( O_P(n)(t^2/n) = O_P(1) \), which is of the same order as the second term.

We conclude this section with an example of a well-known application of the Taylor expansion. More examples will be discussed in the sequel.

Example 4.4 (The delta-method). Let \( \xi_n, n = 1, 2, \ldots \) be a sequence of \( s \)-dimensional random vectors such that
\[
a_n(\xi_n - c) \xrightarrow{d} \eta
\]
(4.19)
as \( n \to \infty \), where \( c \) is a constant vector, \( a_n \) is a sequence of positive constants such that \( a_n \to \infty \) as \( n \to \infty \), and \( \eta \) is an \( s \)-dimensional random vector. Then for any continuously differentiable function \( g(x): \mathbb{R}^s \to \mathbb{R} \), we have
\[
a_n \{ g(\xi_n) - g(c) \} \xrightarrow{d} \frac{\partial g(c)}{\partial x'} \eta
\]
(4.20)
as \( n \to \infty \). To establish (4.20), use the first-order Taylor expansion to get
\[
g(\xi_n) = g(c) + \frac{\partial g(c)}{\partial x'}(\xi_n - c),
\]
where \( \xi_n \) lies between \( c \) and \( \xi_n \). It follows that \( |\xi_n - c| \leq |\xi_n - c| \). (4.19) and the fact that \( a_n \to \infty \) implies that \( \xi_n - c = o_P(1) \); hence, \( \xi_n - c = o_P(1) \). It follows that (Exercise 4.3)
\[
    \frac{\partial g(\xi_n)}{\partial x'} - \frac{\partial g(c)}{\partial x'} = o_P(1).
\]
Therefore, we have, by Theorem 2.13,
\[
a_n \{ g(\xi_n) - g(c) \} = \frac{\partial g(c)}{\partial x'} a_n(\xi_n - c) + \left\{ \frac{\partial g(c)}{\partial x'} - \frac{\partial g(c)}{\partial x'} \right\} a_n(\xi_n - c)
\]
\[
    = \frac{\partial g(c)}{\partial x'} a_n(\xi_n - c) + o_P(1) O_P(1)
\]
\[
    = \frac{\partial g(c)}{\partial x'} a_n(\xi_n - c) + o_P(1)
\]
\[
    \xrightarrow{d} \frac{\partial g(c)}{\partial x'} \eta.
\]

In particular, let \( \theta \) be a \( s \)-dimensional parameter vector and \( \hat{\theta} \) be an estimator of \( \theta \) based on i.i.d. observations \( X_1, \ldots, X_n \). We say the estimator \( \hat{\theta} \) is asymptotically normal if
\[
    \sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \Sigma),
\]
(4.21)
where \( \Sigma \) is called the asymptotic covariance matrix. It follows that for any differentiable function \( g(x): \mathbb{R}^s \to \mathbb{R} \), we have
\[
    \sqrt{n} \{ g(\hat{\theta}) - g(\theta) \} \xrightarrow{d} N(0, \sigma^2),
\]
(4.22)
where
\[
    \sigma^2 = \frac{\partial g(\theta)}{\partial x'} \Sigma \frac{\partial g(\theta)}{\partial x},
\]
(4.23)
where \( \partial g(\theta)/\partial x = (\partial g(\theta)/\partial x')' \). For example, suppose that \( X_1, \ldots, X_n \) are i.i.d. with \( \text{E}(X_i) = \mu \) and \( \text{var}(X_i) = \sigma^2 \in (0, \infty) \). Then, according to the CLT, we have \( \sqrt{n}(\bar{X} - \mu) \xrightarrow{d} N(0, \sigma^2) \), where \( \bar{X} \) is the sample mean. It follows that the following hold (Exercise 4.4):
\[
    \sqrt{n}(e^{\bar{X}} - e^\mu) \xrightarrow{d} N(0, e^{2\mu} \sigma^2),
\]
\[
    \sqrt{n} \{ \log(1 + \bar{X}^2) - \log(1 + \mu^2) \} \xrightarrow{d} N \left\{ 0, \frac{4\mu^2 \sigma^2}{(1 + \mu^2)^2} \right\},
\]
\[
    \sqrt{n} \left( \frac{\bar{X}}{1 + \bar{X}^2} - \frac{\mu}{1 + \mu^2} \right) \xrightarrow{d} N \left\{ 0, \frac{(1 - \mu^2)^2 \sigma^2}{(1 + \mu^2)^4} \right\}.
\]

Obviously, many more such results can be derived.

### 4.3 Edgeworth expansion; method of formal derivation

The central limit theorem (CLT) states that, subject to a mild condition (that the second moment is finite), the sample mean \( \bar{X} \) of i.i.d. observations \( X_1, \ldots, X_n \) is asymptotically normal in the sense that
\[
    \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \xrightarrow{d} N(0, 1)
\]
(4.23)
as \( n \to \infty \), where \( \mu = \text{E}(X_1) \) and \( \sigma^2 = \text{var}(X_1) \). Over the years, this astonishing result has amazed, surprised, or even confused its users. For example, it says that no matter what the population distribution (of \( X_1 \)) is, the limiting distribution on the right side of (4.23) is always the same: the standard normal distribution. Imagine how different the population distribution can be in terms of its shape: symmetric, skewed, bimodal, continuous, discrete, and so forth. Yet, they do not make a difference as long as the CLT is concerned.
Nevertheless, the CLT is correct from a theoretical point of view—and this has been confirmed by countless empirical studies. Here, from a theoretical point of view it means that \( n \to \infty \) or at least is very large. However, in a finite-sample situation, it can well be a different story. For example, suppose that \( n = 30 \). It can be shown that in this case the shape of the population distribution makes a difference (Exercise 4.5). This raises an issue about the convergence rate of the CLT. In particular, two characteristic measures of the shape of the population distribution are the skewness and kurtosis, defined as

\[
\kappa_3 = \frac{E(X_1 - \mu)^3}{\sigma^3}, \tag{4.24}
\]

\[
\kappa_4 = \frac{E(X_1 - \mu)^4}{\sigma^4} - 3, \tag{4.25}
\]

respectively. One would expect these characteristics to have some impact on the convergence rate of the CLT. For example, the celebrated Berry–Esseen theorem, discovered by Berry (1941) and Esseen (1942), states that if the third moment of \( X_1 \) is bounded, then

\[
\sup_x |F_n(x) - \Phi(x)| \leq \frac{c E(|X_1|^3)}{\sqrt{n}}, \tag{4.26}
\]

where \( F_n(x) \) is the cdf of \( \xi_n = (\sqrt{n}/\sigma)(X - \mu) \), \( \Phi(x) \) is the cdf of \( N(0,1) \), and \( c \) is an absolute constant (i.e., a constant that does not depend on the distribution of \( X_1 \)). The Edgeworth expansion, named in honor of the Irish mathematician Francis Ysidro Edgeworth (1845–1926), carries the approximation in (4.26) to higher orders.

Like the Taylor expansion, the Edgeworth expansion can be expressed up to \( k + 1 \) terms plus a remaining term. The difference is that, in the Taylor expansion the terms are in decreasing orders of \( |x - a| \) [see (4.9)]; and in the Edgeworth expansion, the terms are in decreasing orders of \( n^{-1/2} \). For the sake of simplicity, we mainly focus on the case \( k = 2 \), which can be expressed as

\[
F_n(x) = \Phi(x) + \frac{\kappa_3 p_1(x)}{6 \sqrt{n}} \phi(x) + \frac{\kappa_4 p_2(x) - \kappa_3^2 p_4(x)}{24n} \phi(x) + O(n^{-3/2}), \tag{4.27}
\]

where \( \phi(x) \) is the pdf of \( N(0,1) \); that is, \( \phi(x) = (1/\sqrt{2\pi}) e^{-x^2/2} \),

\[
p_1(x) = 1 - x^2,
\]

\[
p_2(x) = x(3 - x^2),
\]

\[
p_3(x) = \frac{x}{3}(15 - 10x^2 + x^4).
\]

Expansion (4.27) is known as the two-term Edgeworth expansion (rather than three-term Edgeworth expansion). Note that \( \phi(x) \) does not count as a "term" (or it may be counted as the zeroth term), so the first term of the expansion is \( O(n^{-1/2}) \), the second term is \( O(n^{-1}) \), and so on. We see that the first and second terms of the Edgeworth expansion involve the skewness, \( \kappa_3 \), and kurtosis, \( \kappa_4 \), confirming our earlier speculation that these quantities may influence the convergence rate of the CLT.

To derive the Edgeworth expansion we introduce a method called formal derivation, which will be used repeatedly in this book. Note that the validity of the Taylor expansion is not with conditions. For example, for (4.14) to hold, it is necessary that the remaining term is really \( o(|x - a|^k) \), which requires certain conditions. Furthermore, according to the results of Chapter 2, convergence in probability does not necessarily imply convergence in expectation. So, for example, it is not necessarily true that \( E(\phi(p)) = o(1) \). However, such arguments as the above will be used in the derivation of the Edgeworth expansion as well as many other asymptotic results in the sequel. So what should we do? Should we verify the necessary conditions for every single step of the derivation or should we go ahead with the derivation without having to worry about the conditions? The answer depends on at what stage you are in during the development of a method. Science will not advance if we have to watch our steps for every tiny little move. In the development of most statistical methods there is an important first step—that is, to propose the method. After the method is proposed, the next step is to study the performance of the method, which includes theoretical justification, empirical studies, and applications. At the first stage of the development (i.e., propose the method), one may not need to worry about the conditions. In other words, the first step does not have to wait for the second step to follow immediately. This is what we called formal derivation.

More specifically, in the first step, one derives the formula (or procedure), assuming that all of the necessary conditions are satisfied or that the formula or procedure will hold under certain conditions. Quite often, the first step is done by some researcher(s) and later justified by others. For example, Efron (1979) proposed the bootstrap method without establishing its theoretical properties. General accounts of theory for the bootstrap were later given by Nickell and Freedman (1981) and Beran (1984), among others. In conclusion, the conditions are important, but they should not tie our hands. This is a moral we learned, among other things, from the development of bootstrap (see Chapter 14) and many other statistical methods.

Going back to the Edgeworth expansion, we use the method of formal derivation. Recall \( \xi_n = (\sqrt{n}/\sigma)(X - \mu) = \sum_{j=1}^{n} Z_j \), where \( Z_j = (X_j - \mu)/\sigma \) and \( n \). Then the cdf of \( \xi_n \) can be expressed as

\[
c_n(t) = E(\exp(it\xi_n)) = \prod_{j=1}^{n} E(\exp(itZ_j)) = E\{\exp(itZ_1)\}^n. \tag{4.28}
\]
4.3 Edgeworth expansion: method of formal derivation

Furthermore, by the Taylor expansion, we have
\[ \exp(iz) = \sum_{k=0}^{\infty} \frac{(iz)^k}{k!} + O(n^{-5/2}). \]

Note that the Taylor expansion also holds for functions of complex variables
\( i = \sqrt{-1} \) is a complex number). Also note that \( Z_1 = O(n^{-1/2}) \). Thus,
\[
E\{\exp(iz)\} = \sum_{k=0}^{\infty} \frac{(iz)^k}{k!} E(Z_1^k) + O(n^{-5/2})
\]
\[
= 1 + \frac{z^2}{2n} + \sum_{k=3}^{\infty} \frac{(iz)^k}{k!} E(Z_1^k) + O(n^{-5/2})
\]
because \( E(Z_1) = 0 \) and \( E(Z_1^2) = 1/n \). Another Taylor expansion gives
\[
\log(E\{\exp(iz)\}) = -\frac{z^2}{2n} + \sum_{k=3}^{\infty} \frac{(iz)^k}{k!} E(Z_1^k) + O(n^{-5/2})
\]
\[
= -\frac{z^2}{2n} + \frac{(iz)^3}{3!} + \frac{(iz)^4}{4!} + \frac{(iz)^5}{5!} + \frac{(iz)^6}{6!} + O(n^{-5/2})
\]
because \( E(Z_1^3) = \kappa_3/n^{3/2} \) and \( E(Z_1^4) = \kappa_4 + 3/n^2 \). Therefore, we have
\[
n \log E\{\exp(iz)\} = -\frac{z^2}{2n} + \frac{(iz)^3}{3!} + \frac{(iz)^4}{4!} + O(n^{-3/2});
\]
hence, by (4.28) and the Taylor expansion of \( f(x) = e^x \) at \( x = -t^2/2 \),
\[
c_n(t) = \exp \left\{ \frac{t^2}{2} + \frac{(iz)^3}{3!} + \frac{(iz)^4}{4!} + n^{-3/2} \right\}
\]
\[
= e^{-t^2/2} - e^{-t^2/2} \left\{ \frac{(iz)^3}{3!} + \frac{(iz)^4}{4!} + n^{-3/2} \right\}
\]
\[
+ e^{-t^2/2} \left\{ \frac{(iz)^3}{3!} + \frac{(iz)^4}{4!} + \ldots \right\} + O(n^{-3/2})
\]
\[
= e^{-t^2/2} + n^{-1/2} r_1(z)e^{-t^2/2} + n^{-1} r_2(z)e^{-t^2/2} + O(n^{-3/2}),
\]
where \( r_1(z) = (\kappa_3/6)z^3 \) and \( r_2(z) = (\kappa_4/24)z^4 + (\kappa_3^2/72)z^6 \). Note that
\[
c_n(t) = \int e^{itz} dF_n(x)
\]
is the Fourier–Stieltjes transform of \( F_n(x) = P(\xi_n \leq x) \), which has the asymptotic expansion (4.29). An inversion of the transform then gives
\[
F_n(x) = \Phi(x) + n^{-1/2} R_1(x) + n^{-1/2} R_2(x) + O(n^{-3/2}),
\]
where \( R_j(x) \) is a function that satisfies
\[
\int e^{itz} dR_j(x) = r_j(it)e^{-t^2/2}, \quad j = 1, 2, \ldots
\]
Note that \( \int e^{itz} d\Phi(x) = e^{-t^2/2} \). It can be shown that (e.g., Hall 1992, p. 44)
\[
R_1(x) = -\frac{\kappa_3}{6} (x^2 - 1) \phi(x),
\]
\[
R_2(x) = -\left\{ \frac{\kappa_4}{24} (x^2 - 3) + \frac{\kappa_3^2}{72} (x^4 - 10x^2 + 15) \right\} x \phi(x).
\]
Thus, by (4.30) we obtain the expansion (4.27). We consider some examples.

Example 4.5. Let \( X_1, \ldots, X_n \) be independent with the Beta(\( \alpha, \beta \)) distribution. We consider two special cases: (i) \( \alpha = \beta = 2 \) and (ii) \( \alpha = 2, \beta = 6 \).

The skewness and kurtosis of the Beta(\( \alpha, \beta \)) distribution are given by
\[
\kappa_3 = \frac{2(\beta - \alpha)\sqrt{\alpha + \beta + 1}}{\alpha + \beta + 2 \sqrt{\alpha \beta}},
\]
\[
\kappa_4 = \frac{6\alpha^2 - 2\alpha^2(\beta - 1) + \beta^2(\beta + 1) - 2\alpha\beta(\beta + 2)}{(\alpha + \beta + 2)\alpha(\alpha + \beta + 3)},
\]
respectively. Therefore, in case (i), we have \( \kappa_3 = 0 \) and \( \kappa_4 = -6/7 \). Thus, the Edgeworth expansion (4.27) becomes
\[
F_n(x) = \Phi(x) - \frac{p_2(x) \phi(x)}{28n} + O(n^{-3/2}).
\]

Note that in case (i), the distribution of \( X_i \) is symmetric. As a result, the Edgeworth expansion has a simpler form. On the other hand, in case (ii), we have \( \kappa_3 = 2\sqrt{3}/5 \) and \( \kappa_4 = 6/55 \). Thus, (4.27) becomes
\[
F_n(x) = \Phi(x) + \frac{\sqrt{3} p_1(x) \phi(x)}{15 \sqrt{n}} + \frac{5p_2(x) - 22p_3(x) \phi(x)}{11000n} + O(n^{-3/2}).
\]

Comparing (4.31) and (4.32), one would expect the convergence in CLT to be faster for case (i) than for case (ii) (Exercise 4.6).

Example 4.6. Suppose that \( X_1, \ldots, X_n \) are independent and distributed as Exponential(1). Then it is easy to verify that \( \kappa_3 = 2 \) and \( \kappa_4 = 6 \). Thus, the Edgeworth expansion (4.27) becomes
4 Asymptotic Expansions

\[ F_n(x) = \Phi(x) + \frac{p_1(x)\phi(x)}{3\sqrt{n}} + \frac{3p_2(x) - 2p_3(x)}{12n}\phi(x) + O(n^{-5/2}). \]

Now comes the second step, the justification part. A rigorous treatment of the Edgeworth expansion, including sufficient conditions, can be found in Hall (1992, Section 2.4). One of the key conditions is known as Cramér’s condition, which states the following:

\[ \lim_{t \to \infty} \sup \{E(e^{itX_1})\} < 1. \]  \hfill (4.33)

In other words, the cf of \( X_1 \) is bounded strictly by 1. It holds, in particular, if \( X_1 \) has a pdf with respect to the Lebesgue measure (see Appendix A.2).

In fact, the Edgeworth expansion is not limited to the sample mean \( \bar{X} \), as has been discussed so far. Let \( \hat{\theta} \) be an estimator of \( \theta \) such that \( \sqrt{n}(\hat{\theta} - \theta) \) is asymptotically normal with mean 0 and variance \( \sigma^2 > 0 \). Then the cdf of \( \xi_n = (\sqrt{n}/\sigma)(\hat{\theta} - \theta) \), \( F_n(x) \), may be expanded as (4.30), where \( R_1(x) = p_j(x)\phi(x) \) and \( p_j(x) \) is a polynomial of degree no more than \( 3j - 1 \). We consider an example below and refer more details to Hall (1992, Section 2.3).

Example 4.7. The t-test and confidence interval are associated with the random variable \( \xi_n = \sqrt{n}(\bar{X} - \mu)/\sigma \), where \( \sigma^2 = n^{-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \). Here, \( X_1, \ldots, X_n \) are assumed to be i.i.d. with a finite fourth moment. The Edgeworth expansion for \( F_n(x) = P(\xi_n \leq x) \) is given by

\[ F_n(x) = \Phi(x) + \frac{P_1(x)\phi(x)}{\sqrt{n}} + \frac{P_2(x)\phi(x)}{n} + o(n^{-1}), \]

where \( P_1(x) = (-\kappa_3/6)(2x^2 + 1) \) and

\[ P_2(x) = x \left\{ \frac{\kappa_4}{12}(x^2 - 3) - \frac{\kappa_3^2}{18}(x^4 + 2x^2 - 3) - \frac{x^2 + 3}{4} \right\}. \]

4.4 Other related expansions

4.4.1 Fourier series expansion

The Fourier–Stieljes transform in the previous section is an extension of the Fourier transform, which has had a profound impact in the mathematical world. The Fourier series may be regarded as a discrete version of the inversion of Fourier transform, which is defined as

\[ \hat{f}(k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t)e^{-ikt} \, dt, \]  \hfill (4.34)

for \( k = 0, \pm 1, \pm 2, \ldots \), where \( i = \sqrt{-1} \). Here, \( f \) denotes an integrable function. Given the Fourier transform, one may recover \( f \) via the Fourier series

\[ \sum_{k=-\infty}^{\infty} \hat{f}(k)e^{ikt}. \]  \hfill (4.35)

Here, the series is understood as convergent in some sense (see below). Note that one may express (4.35) as

\[ \int f(x)e^{ixz} \mu(dx), \]  \hfill (4.36)

where \( Z = \{0, \pm 1, \pm 2, \ldots \} \) and \( \mu \) represents the counting measure on \( Z \) [i.e., \( \mu\{k\} = 1 \) for any \( k \in Z \)]. Comparing (4.36) with (4.35), we see that the Fourier series (4.35) actually corresponds to an inversion formula of the Fourier transform. Note that we have not written (4.35) as

\[ f(t) = \sum_{k=-\infty}^{\infty} \hat{f}(k)e^{ikt}. \]  \hfill (4.37)

The question is whether (4.37) actually holds, or holds in some sense. Before we answer this question, let us point out the following facts.

First, if (4.37) does hold, say, at a certain point \( t \), then by truncating the series after a given number of terms, one obtains the Fourier expansion

\[ f(t) = \sum_{k=-N}^{N} \hat{f}(k)e^{ikt} + o(1), \]  \hfill (4.38)

where \( N \) is a positive integer and the remaining term is \( o(1) \) as \( N \to \infty \). Expansion (4.38) is more useful from a practical point of view because, realistically, one can only evaluate a finite number of terms. Unlike the Taylor expansion, there is no general result on the order of the remaining term, so (4.38) is all one can say at this point. This is because the Fourier series applies to a much broader class of functions than the Taylor expansion. For a function to have a Taylor series, it must be infinitely differentiable, or at least have the same order(s) of derivatives in an interval, if one uses Taylor expansion that involves a finite number of terms. On the other hand, the Fourier series may be used to approximate not only nondifferentiable functions; “they even do a good job in the wilderness of the wildly discontinuous” (Bachman et al. 2000, 139). Therefore, it is difficult to evaluate the order of the remaining term because it depends on, among other things, the degree of “smoothness” of the function. For example, for a noncontinuous function, the convergence of the Fourier series may not be in the sense of (4.37) (see below).

Second, the Fourier series (4.35) is expressed in the form of an exponential series or, more generally,
\[ \sum_{k=-\infty}^{\infty} c_k e^{ikt}. \]  
(4.39)

Alternatively, it may be expressed in the form of a trigonometric series,

\[ \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos(kt) + \sum_{k=1}^{\infty} b_k \sin(kt), \]  
(4.40)

through the simple transformation \( a_k = c_k + c_{-k} \) and \( b_k = i(c_{-k} - c_k) \).

The following theorem, known as Dirichlet's pointwise convergence theorem, states sufficient conditions for the convergence of the Fourier series as well as to what value the series converges. For any function \( f \) defined on \([-\pi, \pi]\), its 2\(\pi\)-periodic extension is defined by \( f(t+2k\pi) = f(t) \), \( k \in \mathbb{Z} \), \( t \in \mathbb{R} \). Furthermore, the left (right) limit of a function \( g \) at a point \( t \) is defined as \( g(t-) = \lim_{t \to t^-} g(s) \) (\( g(t+) = \lim_{t \to t^+} g(u) \)). Here, \( s \to t^- \) (\( u \to t^+ \)) means that \( s \) (\( u \)) approaches \( t \) from the left (i.e., \( s < t \)) (right (i.e., \( u > t \))).

**Theorem 4.4.** Let \( f \) be the 2\(\pi\)-periodic extension of an integrable function on \([-\pi, \pi]\). If \( f'(t-) \) and \( f'(t+) \) exist for all \( t \), then the Fourier series (4.39) or (4.40), where \( c_k = \hat{f}(k) \), \( k \in \mathbb{Z} \), converges to

\[ f(t-) + f(t+) \]

\[ \frac{1}{2} \]

at every \( t \). In particular, at any continuity point \( t \), the series converges to \( f(t) \).

An alternative to pointwise convergence is \( L^2 \)-convergence. A function \( f \in L^2[-\pi, \pi] \) if \( \int_{-\pi}^{\pi} |f(t)|^2 \, dt < \infty \). For any \( f \in L^2[-\pi, \pi] \), its \( n \)th-order Fourier approximation is defined as

\[ S_n f(t) = \sum_{k=-n}^{n} \hat{f}(k) e^{ikt}. \]

Here, we use the term “Fourier approximation” instead of “Fourier expansion,” the difference between the two being that the latter is the former plus a remaining term, which may be expressed in terms of big \( O \) or small \( o \). In fact, \( L^2[-\pi, \pi] \) constitutes a Hilbert space if we define the inner product of any \( f, g \in L^2[-\pi, \pi] \) by \( \langle f, g \rangle = (2\pi)^{-1} \int_{-\pi}^{\pi} f(t) \overline{g(t)} \, dt \), where the bar denotes complex conjugation. It follows that the \( n \)th-order Fourier approximation is simply the projection of \( f \) onto the subspace of \( L^2[-\pi, \pi] \) spanned by \( \{e_k, |k| \leq n\} \), where \( e_k(t) = e^{ikt} \) and the coefficients in the approximation are the inner products, \( \hat{f}(k) = \langle f, e_k \rangle \), \( |k| \leq n \). A sequence \( f_n \) in \( L^2[-\pi, \pi] \) converges in \( L^2 \) to a limit \( f \in L^2[-\pi, \pi] \) if

\[ \int_{-\pi}^{\pi} |f_n(t) - f(t)|^2 \, dt \longrightarrow 0 \]

as \( n \to \infty \). We have the following result.

**Theorem 4.5.** For any \( f \in L^2[-\pi, \pi] \), its Fourier approximation \( S_nf \) converges in \( L^2 \) to \( f \) as \( n \to \infty \).

Modern harmonic analysis treats Fourier series as a special case of orthonormal series for the representation or approximation of functions or signals. Let \( S \) be a subset of \( \mathbb{R} \), the real line. An orthonormal system on \( S \) is defined as a sequence of functions \( \phi_k, k \in I \), on \( S \) such that

\[ \int_S \phi_k(t) \phi_l(t) \, dt = 0, \quad k \neq l, \]  
(4.41)

\[ \int_S |\phi_k(t)|^2 \, dt = 1. \]  
(4.42)

Here, \( I \) represents an index set. Give an orthonormal system \( \phi_k, k \in I \), one may consider the following series expansion of a function \( f \):

\[ f(t) = \sum_{k \in I} c_k \phi_k(t), \]  
(4.43)

where \( c_k = \int_S f(t) \overline{\phi_k(t)} \, dt \). Again, the series expansion may be interpreted in terms of projection in a Hilbert space, with the coefficients being the inner products. Below are some examples.

**Example 4.8.** If we let \( \phi_k(t) = e^{ikt}/\sqrt{2\pi} \), then it is easy to verify that \( \phi_k \), \( k \in \mathbb{Z} \), is an orthonormal system (Exercise 4.8). This orthonormal system on \([-\pi, \pi]\) corresponds to the Fourier series (4.39) with \( c_k = \hat{f}(k) \), \( k \in \mathbb{Z} \).

**Example 4.9.** Similarly, the sequence

\[ \frac{1}{\sqrt{2\pi}}, \quad \frac{\sin(kt)}{\sqrt{2\pi}}, \quad k = 1, 2, \ldots, \frac{\cos(kt)}{\sqrt{2\pi}}, \quad k = 1, 2, \ldots, \]

forms an orthonormal system (Exercise 4.9). This orthonormal system on \([0, \pi]\) corresponds to the Fourier series (4.40) with \( c_k = \hat{f}(k) \), \( k \in \mathbb{Z} \).

**Example 4.10** (Orthogonal polynomials). Consider polynomial approximation to a function \( f \) on \([0, 1]\). A polynomial is a linear combination of the powers \( 1, x, x^2, \ldots \). Unfortunately, the power functions themselves are not orthonormal. A general procedure for constructing an orthonormal system is called the Gram-Schmidt orthonormalization. The procedure is described as follows. Starting with \( \phi_0(x) = 1 \), let

\[ \phi_1(x) = \frac{x - f_0^1 u \, du}{\int_0^1 (v - f_0^1 u \, du)^2 \, dv} \]  
(4.44)

\[ = \sqrt{3}(2x - 1). \]  
(4.45)
In general, the sequence $\phi_k(x)$ is defined recursively by

$$\phi_k(x) = \frac{x^k - \sum_{j=0}^{k-1} \left( \int_0^1 u^k \phi_j(u) \, du \right) \phi_j(x)}{\left( \int_0^1 u^k \, du - \sum_{j=0}^{k-1} \left( \int_0^1 u^k \phi_j(u) \, du \right)^2 \, du \right)^{1/2}}.$$ 

$k = 1, 2, \ldots$. This defines an orthonormal system on $[0, 1]$. In particular, it is fairly straightforward to compute the first few orthonormal polynomials and verify that they are orthonormal (Exercise 4.10).

**Example 4.11 (Haar functions).** This system is a special case of wavelets. It is a sequence of discontinuous functions defined through transformations of the indicator function of $[0, 1)$: $I_{[0,1)}(t) = 1$ if $0 \leq t < 1$ and 0 otherwise. Let $\phi_0(t) = I_{[0,1)}(2t) - I_{[0,1)}(2t-1)$. $\phi_0$ is called the Haar mother wavelet and it can be expressed more explicitly as

$$\phi_0(t) = \begin{cases} 
1, & 0 \leq t < 1/2 \\
-1, & 1/2 \leq t < 1 \\
0, & \text{otherwise}
\end{cases}$$

(see Figure 4.1). The subsequent Haar functions are defined as

$$\phi_{j,k}(t) = 2^{j/2} \phi_0(2^j t - k), \quad j = 0, 1, 2, \ldots, \quad k = 0, 1, \ldots, 2^j - 1, \quad (4.44)$$

where $\phi_{0,0} = \phi_0$, the mother wavelet. It can be shown that the Haar functions defined by (4.44) together with $I_{[0,1)}$ constitute an orthonormal system on $(-\infty, \infty)$ (Exercise 4.11).

### 4.4.2 Cornish–Fisher expansion

The Edgeworth expansion discussed in Section 4.3 can be inverted, leading to a useful expansion for the quantiles of $\xi_n$. This is known as the Cornish–Fisher expansion. For any $\alpha \in (0, 1)$, define $g_n(\alpha) = \inf \{ x : F_n(x) \geq \alpha \}$, which is called the upper $\alpha$th quantile of $F_n$. Here, as in Section 4.3, $F_n$ denotes the cdf of $\xi_n = (\sqrt{n}/\sigma)(X - \mu)$ and $X$ is the sample mean of i.i.d. observations $X_1, \ldots, X_n$. Let $z_\alpha$ denote the upper $\alpha$th quantile of $N(0, 1)$ [i.e., $\Phi(z_\alpha) = \alpha$]. Then the two-term Cornish–Fisher expansion may be expressed as

$$g_n(\alpha) = z_\alpha + \frac{(z_\alpha^2 - 1)\kappa_3}{6\sqrt{n}} + \frac{1}{12n} \left( \frac{(z_\alpha^3 - 3z_\alpha)\kappa_4}{2} - \frac{(2z_\alpha^3 - 5z_\alpha)\kappa_3^2}{3} \right) + \mathcal{O}(n^{-3/2}), \quad (4.45)$$

where $\kappa_3$ and $\kappa_4$ are defined by $(4.24)$ and $(4.25)$, respectively.

The Cornish–Fisher expansion is useful in obtaining more accurate confidence intervals and critical values for tests. Note that the CLT approximation to $g_n(\alpha)$ would be $z_\alpha$, which is the leading term on the right side of (4.45). The following example shows how much more accuracy the expansion (4.45) may bring compared to the CLT approximation.

**Example 4.12.** Barndorff-Nielsen and Cox (1989, p. 119) reported the results of the Cornish–Fisher approximation in the situation where the $X_i$'s are distributed as $\chi_1^2$. In this case, we have $\kappa_3 = 2\sqrt{2}$ and $\kappa_4 = 12$, so the two-term expansion (4.45) becomes

$$g_n(\alpha) = z_\alpha + \frac{\sqrt{2}(z_\alpha^2 - 1)}{3\sqrt{n}} + \frac{z_\alpha^3 - 7z_\alpha}{18n} + \mathcal{O}(n^{-3/2}).$$

Note that the mean and variance of $\chi_1^2 = \sum_{i=1}^n X_i$ are $n$ and $2n$, respectively. Thus, the $\alpha$th quantile of $\chi_1^2$ is...
\[ n + \sqrt{2n} q_n(\alpha) = n + z_\alpha \sqrt{2n} + \frac{2(z_\alpha^2 - 1)}{3} + z_\alpha^3 - 7z_\alpha \frac{1}{9\sqrt{2n}} + O(n^{-1}). \]

Of course, the quantiles of \( \chi_n^2 \) can be calculated exactly. Table 4.2, extracted from Table 4.5 of Barndorff-Nielsen and Cox (1989), compares the approximations by the two-term Cornish–Fisher expansion as well as by the CLT with the exact quantiles for \( \alpha = 0.1 \), where C-F refers to the two-term Cornish–Fisher expansion. The results showed astonishing accuracy of the C-F approximation even with very small sample size (\( n = 5 \)).

<table>
<thead>
<tr>
<th>( n )</th>
<th>Exact</th>
<th>CLT</th>
<th>C-F</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>9.24</td>
<td>9.65</td>
<td>9.24</td>
</tr>
<tr>
<td>10</td>
<td>15.99</td>
<td>15.73</td>
<td>15.99</td>
</tr>
<tr>
<td>50</td>
<td>63.17</td>
<td>62.82</td>
<td>63.16</td>
</tr>
<tr>
<td>100</td>
<td>118.50</td>
<td>118.12</td>
<td>118.50</td>
</tr>
</tbody>
</table>

It should be pointed out that, like the Edgeworth expansion, the Cornish–Fisher expansion requires certain regularity conditions in order to hold, and one of the key conditions is (4.33). If the condition fails, the Cornish–Fisher expansion may not improve over the CLT. The following is an example.

**Example 4.13.** Suppose that \( X_1, \ldots, X_n \) are i.i.d. from the Bernoulli(\( p \)) distribution with \( p = 0.5 \). It is easy to show that the distribution does not satisfy (4.33) (Exercise 4.12). If one blindly applies the Cornish–Fisher expansion (4.45), then since in this case \( \kappa_3 = 0 \) and \( \kappa_4 = -2 \), one would get

\[ q_n(\alpha) = z_\alpha + \frac{3z_\alpha - z_\alpha^3}{12} + O(n^{-3/2}). \]

Despite the simple form, (4.46) may not give a better approximation than the CLT. To see this, note that \( X \sim \text{Binomial}(n, p) \), so the exact cth quantile of \( X \) can be calculated. On the other hand, the cth quantile of \( X \) is given by

\[ n \left\{ p + q_n(\alpha) \sqrt{\frac{p(1-p)}{n}} \right\}. \]

Table 4.3 compares the approximations to the cth quantiles, where \( \alpha = P(X \leq k) \) for \( n = 15 \) and \( k = 3, 6, 9, 12 \), by C-F [i.e., (4.47) with \( q_n(\alpha) \) given by (4.46)] as well as by the CLT [i.e., (4.47) with \( q_n(\alpha) = z_\alpha \)], with the exact quantiles. It is seen that inappropriate use of C-F sometimes makes things worse.

**Table 4.3. Approximation of quantiles**

<table>
<thead>
<tr>
<th>( k )</th>
<th>( n = 3 )</th>
<th>( n = 6 )</th>
<th>( n = 9 )</th>
<th>( n = 12 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha )</td>
<td>0.0176</td>
<td>0.0356</td>
<td>0.0491</td>
<td>0.0995</td>
</tr>
<tr>
<td>CLT</td>
<td>3.4207</td>
<td>6.5046</td>
<td>9.4998</td>
<td>12.6878</td>
</tr>
<tr>
<td>C-F</td>
<td>3.4533</td>
<td>6.4895</td>
<td>9.5212</td>
<td>12.5674</td>
</tr>
</tbody>
</table>

**4.4 Other related expansions**

**4.4.3 Two time series expansions**

A time series is a set of observations, each recorded at a specified time \( t \). In this subsection we consider a stationary (complex-valued) time series, denoted by \( \{X_t, t = 0, \pm 1, \pm 2, \ldots\} \), or simply \( X_t \). This means that \( E(\|X_t\|^2) < \infty \) and \( E(X_t) \) and \( E(X_{t+k}, X_t) \) do not depend on \( t \). We can then define the autocovariance function of \( X_t \) as

\[ \gamma(k) = E(\{X_{t+k} - \mu\})(X_t - \mu), \quad (4.48) \]

\( k = 0, \pm 1, \pm 2, \ldots \), where \( \mu = E(X_t) \). One special stationary time series is called a white noise, for which \( \mu = 0 \) and \( \gamma(k) = \sigma^2 1(\kappa = 0) \). In other words, \( W_t \) is a white noise if \( E(W_t) = 0 \), \( E(W_t^2) = \sigma^2 \), and \( E(W_t W_s) = 0 \), \( s \neq t \). The following conditions (i) and (ii) are both necessary and sufficient for \( \gamma \) to be the autocovariance function of a stationary time series \( X_t \).

(i) \( \gamma(k) = \int_{-\pi}^{\pi} e^{ik\lambda} dF(\lambda) \), where \( F \) is a right-continuous, nondecreasing and bounded function on \( [-\pi, \pi] \) with \( F(-\pi) = 0 \).

(ii) \( \sum_{k=-1}^{n-1} \gamma(i-j) e^{i\lambda j} \geq 0 \) for any positive integer \( n \) and \( a_1, \ldots, a_n \in C \).

Here, \( C \) denotes the set of complex numbers and \( F \) is right-continuous at \( F(\nu) \to F(\lambda) \) as \( \nu \to \lambda \) from the right (i.e., \( \nu > \lambda \)). The function \( F \) is called the spectral distribution function of \( \gamma \) or \( X_t \). In particular, if \( F \) is absolutely continuous such that \( F(\lambda) = \int_{-\pi}^{\lambda} f(\nu) d\nu \), \( -\pi \leq \lambda \leq \pi \), \( f \) is called the spectral density of \( \gamma \) or \( X_t \). Note that the properties of \( F \) imply that \( \lambda \geq 0, \lambda \in [-\pi, \pi] \). If \( \gamma \) is the autocovariance function of a stationary time series \( X_t \) that is absolutely summable, that is,

\[ \sum_{k=-\infty}^{\infty} |\gamma(k)| < \infty. \]

there exists a function \( f \) such that \( f(\lambda) \geq 0, \lambda \in [-\pi, \pi] \) and

\[ \gamma(k) = \int_{-\pi}^{\pi} e^{ik\lambda} f(\lambda) d\lambda, \quad k = 0, \pm 1, \pm 2, \ldots. \]

In other words, \( f \) is the spectral density of \( X_t \). Furthermore, we have

\[ f(\lambda) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \gamma(k) e^{-ik\lambda}, \]

\[ f(\lambda) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \gamma(k) e^{-ik\lambda}, \]

\[ f(\lambda) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \gamma(k) e^{-ik\lambda}, \]
\[ f(\lambda) = \frac{1}{2\pi} \sum_{k=-n}^{n} \gamma(k)e^{-ik\lambda} + o(1), \quad (4.52) \]

\( \lambda \in [-\pi, \pi], \) where \( o(1) \to 0 \) as \( n \to \infty. \) Equation (4.51) or (4.52) can be established using the results of Fourier expansion (see Section 4.4.1) or verified directly using Fubini’s theorem (see Exercise 4.12).

Another well-known expansion in time series is called the Wold decomposition. For simplicity, assume that \( X_t \) is real-valued. Consider the space \( \mathcal{H} \) of all random variables \( X \) satisfying \( \mathbb{E}(X^2) < \infty. \) Then \( \mathcal{H} \) is a Hilbert space with the inner product \( X^{'}, Y = \mathbb{E}(XY). \) Let \( \mathcal{H}_t \) denote the subspace of \( \mathcal{H} \) spanned by \( \{X_s, s \leq t\}. \) Let \( P_{\mathcal{H}_t}, X_t \) denote the projection of \( X_t \) onto \( \mathcal{H}_t \) (called the one-step predictor). See Chapter 9 for more details. Also, define \( \mathcal{H}_\infty = \bigcap_{t=1}^{\infty} \mathcal{H}_t. \) A time series \( X_t \) is said to be deterministic of \( \mathcal{H}_t = \mathcal{H}_\infty \) for all \( t. \) The Wold decomposition states that if \( \sigma^2 = \mathbb{E}(X_t - P_{\mathcal{H}_t}X_t)^2 > 0, \) then \( X_t \) can be expressed as

\[ X_t = \sum_{k=0}^{\infty} \psi_k Z_{t-k} + V_t, \quad (4.53) \]

where \( \psi_0 = 1 \) and \( \sum_{k=0}^{\infty} \psi_k^2 < \infty; Z_t \) is a white noise with variance \( \sigma^2 \) and \( Z_t \in \mathcal{H}_t; V_t \) and \( Z_t \) are uncorrelated (i.e., \( \mathbb{E}(Z_t V_u) = 0, \forall t, u \)) and \( V_t \in \mathcal{H}_\infty \) and is deterministic. In fact, (4.53) and the above properties uniquely determine \( \psi_k, Z_t, \) and \( V_t. \) We consider an example.

Example 4.14. Consider the real-valued function

\[ \gamma(k) = \begin{cases} 1, & k = 0 \\ \rho^{|k|}, & k = \pm 1 \\ 0, & \text{otherwise.} \end{cases} \]

It is easy to show that \( \gamma \) is an autocovariance function if \( |\rho| \leq 1/2 \) (Exercise 4.13). Since (4.49) is obviously satisfied, it follows by the spectral representation (4.51) that

\[ f(\lambda) = (2\pi)^{-1} \sum_{k=\infty}^{\infty} \gamma(k)e^{-ik\lambda} = (2\pi)^{-1}\{1 + 2\rho \cos(\lambda)\}. \]

Clearly, we have \( f(\lambda) \geq 0, \lambda \in [-\pi, \pi] \) provided that \( |\rho| \leq 1/2. \) In fact, this is the spectral density of an MA(1) process defined by \( X_t = Z_t + \theta Z_{t-1}, \) where \( Z_t \) is a white noise with variance \( \sigma^2 > 0 \) and \( \theta = \rho/\sigma^2. \) See Chapter 9 for more details. Clearly, the Wold decomposition holds for this \( X_t \) with \( \psi_0 = 1, \psi_1 = \theta, \psi_k = 0, k > 1, \) and \( V_t = 0. \)

4.5 Some elementary expansions

The asymptotic expansions encountered so far are well known in the mathematical or statistical literature, and their derivations involve (much) more than just a few lines of algebra. However, these are not the only ways to come up with an asymptotic expansion. In this section, we show that one can derive some useful asymptotic expansions oneself using some elementary approaches that involve nothing more than a few lines of simple algebra.

Let us begin with a simple problem. Suppose that one wishes to expand the function \( f(x) = x^{-1} \) at \( x = a. \) Most people would immediately think that, well, let’s try the Taylor expansion. Surely one can do so without a problem. However, here is an alternative approach. Write

\[ \frac{1}{x} = \frac{1}{a} + \frac{a-x}{a} = \frac{x}{a} - \frac{a}{x} \quad (4.54) \]

Equation (4.54) suggests an iterative procedure so that we have

\[ \frac{1}{x} = \frac{1}{a} - \frac{x-a}{a} \left( \frac{1}{a} - \frac{x-a}{a} \right) \]

\[ = \frac{1}{a} - \frac{x-a}{a} \cdot \frac{(x-a)^2}{a^2} \quad \frac{1}{x} = \frac{1}{a} - \frac{x-a}{a} \cdot \frac{(x-a)^2}{a^2} \cdot \frac{(1 - \frac{x-a}{a})}{a}. \]

In general, one has the asymptotic expansion

\[ \frac{1}{x} = \sum_{k=1}^{l} (-1)^{k-1} \frac{(x-a)^{k-1}}{a^k} + (-1)^l \frac{(x-a)^l}{a^l}, \quad (4.55) \]

For \( l = 1, 2, \ldots \) If one instead uses the Taylor expansion, then since \( f^{(k)}(x) = (-1)^{k+1}k!(x^{-k-1}), \) we have, by (4.8),

\[ \frac{1}{x} = \sum_{k=1}^{l} (-1)^{k-1} \frac{(x-a)^{k-1}}{a^k} + (-1)^l \frac{(x-a)^l}{a^{l+1}}, \quad (4.56) \]

Here \( \xi \) lies between \( a \) and \( x. \) Comparing the Taylor expansion with (4.55), which we call elementary expansion, the only difference is that (4.55) is more precise in terms of the remaining term than (4.56). In other words, in the Taylor expansion, we only know that \( \xi \) is somewhere between \( a \) and \( x, \) whereas the elementary expansion has no such uncertainty. Here is another look at the difference. If we drop the remaining term in the Taylor expansion (4.8) with \( l \) replaced by \( l + 1, \) we can write

\[ \frac{1}{x} \approx \sum_{k=1}^{l} (-1)^{k-1} \frac{(x-a)^{k-1}}{a^k} + (-1)^l \frac{(x-a)^l}{a^{l+1}}. \quad (4.57) \]
Comparing (4.55) with (4.57), the difference is that the elementary expansion is exact (characterized by =), whereas the Taylor expansion is approximate (characterized by \(\approx\)).

In fact, it is not just that the results are (slightly) different. The elementary expansion is derived using very simple algebras—no results of calculus such as derivatives are involved. This is important because such an elementary expansion is easier to extend to situations beyond real numbers, such as matrices. For example, suppose that one wishes to approximate the inverse of matrix \(B\) by that of matrix \(A\). Then, by a similar derivation, we have

\[
B^{-1} = A^{-1} + B^{-1} - A^{-1} = A^{-1} + A^{-1}(A - B)B^{-1} \\
= A^{-1} + A^{-1}(A - B)(A^{-1} + A^{-1}(A - B)B^{-1}) \\
= A^{-1} + A^{-1}(A - B)A^{-1} + (A^{-1}(A - B)B^{-1})^2B^{-1} \\
= A^{-1} + A^{-1}(A - B)A^{-1} + (A^{-1}(A - B)B^{-1})^2A^{-1} + (A^{-1}(A - B)B^{-1})^3B^{-1} \\
= \ldots.
\]

In general, we have the matrix asymptotic expansion

\[
B^{-1} = \sum_{l=0}^{\infty} (A^{-1}(A - B))^l A^{-1} + (A^{-1}(A - B))^l B^{-1}
\]

for \(l = 0, 1, 2, \ldots\) (e.g., Das et al. 2004, Lemma 5.4).

For example, expansions such as (4.55) and (4.58) are useful in situations where \(\mathbf{X}(\mathbf{B})\) is a random variable (matrix) and \(a(A)\) is its expectation. We consider an example.

**Example 4.15.** Suppose that \(X_1, \ldots, X_n\) are i.i.d. \(p\)-dimensional standard normal random vectors; that is, the \(X_i\)'s are independent \(\sim N(0, I_p)\), where \(I_p\) is the \(p\)-dimensional identity matrix. Let \(B = I_p + \bar{X}\bar{X}'\), where \(\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i\), and suppose that one wishes to evaluate \(E(B^{-1})\). Note that \(\bar{X} \sim N(0, n^{-1} I_p)\); hence, \(n^{-1/2} \bar{X} \sim N(0, I_p)\). It follows that \(\bar{X} = O_p(n^{-1/2})\).

Let \(A = E(B) = ((n+1)/n) I_p\). Then we have \(E((A - B)^2) = E(XX'X\bar{X}') - n^{-2} I_p\). Write \(\xi = n^{1/2} \bar{X} = (\xi_1, \ldots, \xi_p)' \sim N(0, I_p)\). Then the \((i, j)\) element of \(\xi\xi'\xi\xi' = \eta_{ij}\) is \(\eta_{ij} = \xi_i \xi_j \sum_{k=1}^{n} \xi_k^2\). It is easy to show (Exercise 4.15) that \(E(\xi_{ij}) = (p+1)I_{(i=j)}\). It follows that \(E(\xi\xi'\xi\xi') = (p+2)I_p\); hence, \(E(\{\{A - B\}^2\} = n^{-2} E(\xi\xi'\xi\xi') - n^{-2} I_p = n^{-2}(p+1) I_p\). Now, by (4.58) with \(l = 2\), we have

\[
B^{-1} = I_p + A^{-1}(A - B) + (A^{-1}(A - B))^2 A^{-1} - O_p(n^{-3}).
\]

Here, we used Theorem 3.2 to argue that \(B^{-1} = O_p(1)\), and note that \(A^{-1} = \{n/(n+1)\} I_p = O(1)\), and \(A - B = n^{-1} I_p - \bar{X}\bar{X}' = O(n^{-1}) + O_p(n^{-1}) = O_p(n^{-1})\). Thus, by the method of formal derivation (see Section 4.3), we have

\[
E(B^{-1}) = A^{-1} + I_p + \left( \frac{n}{n+1} \right)^3 E((A - B)^2) + O(n^{-3})
\]

\[
= \frac{n}{n+1} I_p + \left( \frac{n}{n+1} \right)^3 \left\{ 1 + \frac{p+1}{(n+1)^2} \right\} I_p + O(n^{-3})
\]

\[
= \left\{ 1 - \frac{1}{n} + \frac{p+2}{n^2} \right\} I_p + O(n^{-3}).
\]

The last equality in (4.59) is because, by (4.55) with \(l = 3\), we have \(n^{-1} = n^{-1} - n^{-2} + n^{-3} + O(n^{-4})\); hence, \(n/(n+1) = 1 - n^{-1} + n^{-2} + O(n^{-3})\) and \(n^{-1} = n^{-1} - n^{-2} + O(n^{-3})\). Therefore, \(\{n/(n+1)\} \{1 + (p+1)/(n+1)^2\} \{1 - n^{-1} + n^{-2} + O(n^{-3})\} \{1 - n^{-1} + n^{-2} + O(n^{-3})\} = 1 - n^{-1} + (p+2)n^{-2} + O(n^{-3})\).

We now derive (4.59) using a different method—this time by the Taylor expansion. To do so, we first make use of the following matrix identity (e.g., Sen and Srivastava 1990, p. 275): For any \(p \times p\) matrix \(P, P \times q\) matrix \(U\) and \(q \times p\) matrix \(V\), we have

\[
(P + UV)^{-1} = P^{-1} - P^{-1} U (I_q + VP^{-1} U)^{-1} V P^{-1},
\]

provided that the inverses involved exist. By letting \(P = I_p, U = \bar{X}\), and \(V = \bar{X}'\) in (4.60), we have

\[
B^{-1} = (I_p + \bar{X}\bar{X}')^{-1}
\]

\[
= I_p - \frac{\bar{X}\bar{X}'}{1 + \bar{X}'\bar{X}}
\]

\[
= I_p - \frac{\xi\xi'}{n + \xi\xi'}
\]

where \(\xi\) is defined as above. Note that the \((i, j)\) element of \(\xi = (n + \xi\xi')^{-1} \xi\xi' = (n + \xi\xi')^{-1} \xi\xi'\xi\xi'\). If \(i \neq j\), then \(E(\xi_{ij}) = 0\); if \(i = j\), then \(E(\xi_{ii}) = E((n + \xi\xi')^{-1} \xi\xi'\xi\xi')\). It does not depend on \(i\) (Exercise 4.16). Thus,

\[
E(\xi_{ii}) = \frac{1}{p} \sum_{i=1}^{p} E \left( \frac{\xi_{ii}^2}{n + \sum_{k=1}^{n} \xi_k^2} \right)
\]

\[
= \frac{1}{p} \left( \frac{\sum_{i=1}^{p} \xi_{ii}^2}{n + \sum_{k=1}^{n} \xi_k^2} \right)
\]

\[
= \frac{1}{p} \left( \frac{\bar{X}'\bar{X}}{n + \bar{X}'\bar{X}} \right)
\]

where \(\chi^2_p\) represents a random variable with a \(\chi^2\)-distribution. By the Taylor expansion, we have
Now, again, use the method of formal derivation (Section 4.3), the facts that $E(\chi^2_p) = p$ and $E(\chi^4_p) = p(p + 2)$, and (4.61) to get

\[
E(\xi_k) = \frac{p}{n} - \frac{p(p + 2)}{n^2} + O(n^{-3})
\]

Thus, we have

\[
E(\mathcal{B}^{-1}) = I_p - E(\zeta)
\]

\[
= I_p - \left\{ \frac{1}{n} + \frac{p + 2}{n^2} + O(n^{-3}) \right\}
\]

\[
= \left\{ 1 - \frac{1}{n} + \frac{p + 2}{n^2} \right\} I_p + O(n^{-3}),
\]

which is the same as (4.59).

Note that in the latest derivation using the Taylor expansion we actually benefited from the identity (4.60) of matrix inversion and results on moments of the $\chi^2$-distribution (otherwise, the derivation could be even more tedious).

### 4.6 Laplace approximation

Suppose that one wishes to approximate an integral of the form

\[
\int e^{-q(z)} dz,
\]

where $q(\cdot)$ is a "well-behaved" function in the sense that it achieves its minimum value at $x = \tilde{x}$ with $q'(\tilde{x}) = 0$ and $q''(\tilde{x}) > 0$. Then we have, by the Taylor expansion,

\[
q(x) = q(\tilde{x}) + \frac{1}{2} q''(\tilde{x})(x - \tilde{x})^2 + \cdots,
\]

which yields the following approximation (Exercise 4.18):

\[
\int e^{-q(z)} dz \approx \sqrt{\frac{2\pi}{q''(\tilde{x})}} e^{-q(\tilde{x})},
\]

Approximations such as (4.63) are known as the Laplace approximation, named after the French mathematician and astronomer Pierre-Simon Laplace.

There is a multivariate extension of (4.63), which is often more useful in practice. Let $q(x)$ be a well-behaved function that attains its minimum value at $x = \tilde{x}$ with $q'(\tilde{x}) = 0$ and $q''(\tilde{x}) > 0$ (positive definite), where $q'$ and $q''$ denote the gradient vector and Hessian matrix, respectively. Then we have

\[
\int e^{-q(x)} dx \approx c |q''(\tilde{x})|^{-1/2} e^{-q(\tilde{x})},
\]

where $c$ is a constant depending only on the dimension of the integral (Exercise 4.13) and $|A|$ denotes the determinant of matrix $A$.

Approximations (4.63) or (4.64) are derived using the second-order Taylor expansion. This is called the first-order Laplace approximation. If one uses the higher order Taylor expansions, the results are the higher order Laplace approximations, which are more complicated in their forms (e.g., Barndorff-Nielsen and Cox 1989, Section 3.3; Lin and Breslow 1996). For a fixed-order (e.g., first order) Laplace approximation, its accuracy depends on the behavior of the function $q$. Roughly speaking, the more "concentrate" the function is far the more accurate; and the more normal-look-like the function is the more accurate. For example, consider the following.

**Example 4.16 (t-distribution).** Consider the function

\[
q(x) = \frac{\nu + 1}{2} \log \left( 1 + \frac{x^2}{\nu} \right), \quad -\infty < x < \infty,
\]

where $\nu$ is a positive integer. Note that, subject to a normalizing constant, $q(x)$ corresponds to the pdf of the $t$-distribution with $\nu$ degrees of freedom. It is easy to verify (Exercise 4.20) that, in this case, the exact value of (4.62) given by $n^{\nu/2} \Gamma(\nu/2)/\Gamma((\nu + 1)/2)$, where $\Gamma$ is the gamma function; the Laplace approximation (4.63) is $2\nu^{\nu/2}/(\nu + 1)$. Table 4.4 shows the numerical values (up to the fourth decimal) for a number of different $\nu$'s, where Relative Error is defined as Exact minus Approximate divided by Exact. It is seen that the accuracy improves as $\nu$ increases. This is because as $\nu$ increases, the $t$-distribution becomes more and more concentrate at $x = 0$. In the extreme case where $\nu \to \infty$, the $t$-distribution becomes the standard normal distribution, of which the Laplace approximation is exact (Exercise 4.20).

So, if $q$ is a fixed function, there is a limit for how accurate one can approximate (4.62) with a fixed-order Laplace approximation. Note that, in practice, the first-order Laplace approximation is by far the most frequently used and is more accurate than the second order Laplace approximation is rarely even considered. This is because as the order increases, the formula for the approximation quickly becomes complicated, especially in the multivariate case. Therefore, theoretically, increasing the order of Laplace approximation may not be an op-
The accuracy of Laplace approximation

<table>
<thead>
<tr>
<th>ν</th>
<th>Exact Approximate Relative Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.31416</td>
<td>1.7725</td>
</tr>
<tr>
<td>3.6284</td>
<td>2.0467</td>
</tr>
<tr>
<td>3.7207</td>
<td>2.1708</td>
</tr>
<tr>
<td>4.6067</td>
<td>2.2420</td>
</tr>
<tr>
<td>5.8343</td>
<td>2.3882</td>
</tr>
<tr>
<td>10.5700</td>
<td>2.9500</td>
</tr>
<tr>
<td>50.1592</td>
<td>2.4819</td>
</tr>
<tr>
<td>100.1592</td>
<td>2.4942</td>
</tr>
<tr>
<td>250.1592</td>
<td>2.5016</td>
</tr>
</tbody>
</table>

The function $q$ in (4.62) is not a fixed function but rather depends on $n$, the sample size. In other words, the sample size $n$ may play a role in the accuracy of Laplace approximation, which so far has not been taken into account. To see why the sample size may help, let us consider a simple example. Suppose that the function $q$ in (4.62), or, more precisely, $e^{-q(x)}$, corresponds to the pdf of a sample mean $\bar{X}$ of i.i.d. random variables $X_1, \ldots, X_n$. According to the law of large numbers (LLN), as $n$ increases, $\bar{X}$ becomes more and more concentrated near the population mean $\bar{x} = E(X_1)$. Therefore, the Laplace approximation is expected to become more accurate as $n$ increases. To show this more precisely, let us first consider a simple case.

Suppose that in (4.62), $q(x) = nx$, and another function $p(x)$ is added in front of $dx$. More specifically, we consider

$$I_n = \int_0^\infty e^{-nx}p(x) \, dx.$$  \hspace{1cm} (4.55)

Suppose that $p^{(k)}(x)e^{-nx} \rightarrow 0$ as $x \rightarrow \infty$ for $k = 0, 1, 2, \ldots$. Then, by integration by parts, we have

$$I_n = \frac{p(0)}{n} + \frac{1}{n} \int_0^\infty e^{-nx}p'(x) \, dx$$

$$= \frac{p(0)}{n} + \frac{p'(0)}{n^2} + \frac{1}{n} \int_0^\infty e^{-nx}p''(x) \, dx$$

$$= \frac{p(0)}{n} + \frac{p'(0)}{n^2} + \frac{p''(0)}{n^3} + \cdots.$$  \hspace{1cm} (4.56)

In other words, we have an asymptotic expansion in terms of increasing powers of $n^{-1}$. Now, let us consider a more general case by replacing the function $x$ in (4.65) by $g(x)$ and assuming that $g'(0) \neq 0$. It is also assumed that as $x \rightarrow \infty$, $e^{-ng(x)}p(x)g'(x) \rightarrow 0$, $e^{-ng(x)}\left\{p(x)g'(x) - p(x)g''(x) / (g'(x))^2\right\} \rightarrow 0$, and so forth, so that we get, by integration by parts,

$$I_n = \int_0^\infty e^{-ng(x)}p(x) \, dx$$

$$= e^{-ng(0)} \frac{p(0)}{g'(0)} + \frac{1}{n} \int_0^\infty e^{-ng(x)} \left\{\frac{p(x)}{g'(x)}\right\}' \, dx$$

$$= e^{-ng(0)} \frac{p(0)}{g'(0)} + e^{-ng(0)} \frac{p'(0)}{g'(0)} g'(0) - p(0) g''(0) / (g'(0))^2 + \cdots.$$  \hspace{1cm} (4.66)

So, again, the expansion is in increasing powers of $n^{-1}$.

The assumption that $g'(0) \neq 0$ makes a big difference in the approximation. To see this, consider the integral

$$I_n = \int_0^\infty e^{-ng(x)}p(x) \, dx.$$  \hspace{1cm} (4.67)

Suppose that $g(x)$ attains its minimum at $\bar{x}$ such that $g'(\bar{x}) = 0$, $g''(\bar{x}) > 0$ and $p(\bar{x}) \neq 0$. Then, under regularity conditions, we have, by the Taylor expansion,

$$g(x) = g(\bar{x}) + g'(\bar{x})(x - \bar{x}) + \frac{1}{2}g''(\bar{x})(x - \bar{x})^2 + \cdots$$

$$= g(\bar{x}) + \frac{1}{2}g''(\bar{x})(x - \bar{x})^2 + \cdots.$$  \hspace{1cm} (4.68)

So, if we make a change of variable $y = x - \bar{x}$, we have

$$ng(x) = ng(\bar{x}) + \frac{n}{2}g''(\bar{x})(x - \bar{x})^2 + \cdots$$

$$= ng(\bar{x}) + \frac{y^2}{2} + \cdots.$$  \hspace{1cm} (4.69)

On the other hand, again by the Taylor expansion, we have

$$p(x) = p\left(\bar{x} + \frac{y}{\sqrt{ng''(\bar{x})}}\right)$$

$$= p(\bar{x}) + p'(\bar{x}) \frac{y}{\sqrt{ng''(\bar{x})}} + \frac{1}{2}p''(\bar{x}) \frac{y^2}{ng''(\bar{x})} + \cdots.$$  \hspace{1cm} (4.70)

If we ignore the $\cdots$ in both expansions, we obtain the following Laplace approximation of $I_n$ in (4.67):

$$I_n = \int_{-\infty}^\infty \exp\left\{-ng(\bar{x}) - \frac{y^2}{2} - \cdots\right\}$$

$$\times\left\{p(\bar{x}) + \frac{p'(\bar{x})}{\sqrt{ng''(\bar{x})}} y + \frac{p''(\bar{x})}{2ng''(\bar{x})} y^2 + \cdots\right\} \frac{dy}{\sqrt{ng''(\bar{x})}}$$

$$\approx \frac{e^{-ng(\bar{x})}}{\sqrt{ng''(\bar{x})}} \int_{-\infty}^\infty e^{-y/2} \left\{p(\bar{x}) + \frac{p'(\bar{x})}{\sqrt{ng''(\bar{x})}} y + \frac{p''(\bar{x})}{2ng''(\bar{x})} y^2\right\} \frac{dy}{\sqrt{ng''(\bar{x})}}.$$  \hspace{1cm} (4.71)
4.7 Case study: Asymptotic distribution of the MLE

A classical application of Taylor series expansion is the derivation of the asymptotic distribution of the MLE. Let us begin with the i.i.d. case with the same setup as in Section 1.4; that is, \( X_1, \ldots, X_n \) are i.i.d. observations with PDF \( f(x|\theta) \), where \( \theta \) is a real-valued unknown parameter with the parameter space \( \Theta = (-\infty, \infty) \). Let \( \hat{\theta} \) denote the MLE of \( \theta \). We assume that \( \hat{\theta} \) is consistent (see Section 1.4). Let \( l(\theta|X) \) denote the log-likelihood function; that is, \( l(\theta|X) = \sum_{i=1}^{n} \log \left[ f(X_i|\theta) \right] \). Here, \( X = (X_1, \ldots, X_n)' \) represents the vector of observations. Then under regularity conditions we have, by the Taylor expansion,

\[
0 = \left. \frac{\partial}{\partial \theta} l(\theta|X) \right|_{\theta = \hat{\theta}},
\]

\[
= \left. \frac{\partial^2}{\partial \theta^2} l(\theta|X) \right|_{\theta = \hat{\theta}} (\hat{\theta} - \theta) + \left. \frac{1}{2} \left\{ \frac{\partial^3}{\partial \theta^3} l(\theta|X) \right\} (\hat{\theta} - \theta)^2,
\]

(4.70)

where \( \hat{\theta} \) lies between \( \theta \) and \( \hat{\theta} \). Before we continue, let us note the following facts:

(i) We have

\[
\left. \frac{\partial^2}{\partial \theta^2} l(\theta|X) \right|_{\theta = \hat{\theta}} = \text{E} \left\{ \left( \frac{\partial}{\partial \theta} l(\theta|X) \right)^2 \right\} + \left( \frac{\partial}{\partial \theta} l(\theta|X) \right) \text{E} \left\{ \left( \frac{\partial}{\partial \theta} l(\theta|X) \right)^2 \right\} - \left( \text{E} \left\{ \frac{\partial}{\partial \theta} l(\theta|X) \right\} \right)^2.
\]

(ii) We have

\[
\left( \frac{\partial}{\partial \theta} l(\theta|X) \right) \left( \frac{\partial^3}{\partial \theta^3} l(\theta|X) \right) = \text{E} \left\{ \frac{\partial^3}{\partial \theta^3} l(\theta|X) \right\} + \text{E} \left\{ \left( \frac{\partial}{\partial \theta} l(\theta|X) \right)^2 \right\} - \text{E} \left\{ \left( \frac{\partial}{\partial \theta} l(\theta|X) \right)^2 \right\} \text{E} \left\{ \left( \frac{\partial}{\partial \theta} l(\theta|X) \right)^2 \right\}.
\]

(iii) We have

\[
\left. \frac{\partial^3}{\partial \theta^3} l(\theta|X) \right|_{\theta = \hat{\theta}} = \text{E} \left\{ \frac{\partial^3}{\partial \theta^3} l(\theta|X) \right\} + \text{E} \left\{ \left( \frac{\partial}{\partial \theta} l(\theta|X) \right)^2 \right\} - \text{E} \left\{ \left( \frac{\partial}{\partial \theta} l(\theta|X) \right)^2 \right\} \text{E} \left\{ \left( \frac{\partial}{\partial \theta} l(\theta|X) \right)^2 \right\}.
\]

These facts are used in the asymptotic normality of the MLE.
4.7 Case study: Asymptotic distribution of the MLE

using the consistency of \( \hat{\theta} \). Thus, we have

\[
\sqrt{n}(\hat{\theta} - \theta) = -\frac{1}{E(Z_i^2)} + o_P(1) \frac{1}{\sqrt{n}} \frac{\partial}{\partial \theta} l(\theta; X) \xrightarrow{d} N \left( 0, \frac{1}{E(Z_i^2)} \right) \tag{4.74}
\]

using Slutsky’s theorem (Theorem 2.13).

The quantity \( E(Z_i^2) \) is known as the Fisher information, denoted by

\[
I(\theta) = \mathbb{E} \left[ \frac{\partial}{\partial \theta} \log \{ f(X_1 | \theta) \} \right]^2. \tag{4.75}
\]

In a suitable sense, \( I(\theta) \) represents the amount of information about \( \theta \) contained in \( X_1 \). The concept can be extended to multiple observations; that is, the amount of information contained in \( X_1, \ldots, X_n \) is

\[
\mathcal{I}(\theta) = \mathbb{E} \left[ \frac{\partial}{\partial \theta} \log \{ f(X_1, \ldots, X_n | \theta) \} \right]^2. \tag{4.76}
\]

Here, with a little abuse of the notation, \( f(x_1, \ldots, x_n | \theta) \) represents the joint pdf of \( X_1, \ldots, X_n \). Since, in the i.i.d. case, we have \( f(x_1, \ldots, x_n | \theta) = \prod_{i=1}^n f(x_i | \theta) \), it follows that, under regularity conditions, \( \mathcal{I}(\theta) = n I(\theta) \) (Exercise 4.22); that is, the amount of information contained in \( X_1, \ldots, X_n \) is \( n \) times that contained in \( X_1 \).

The result (4.74) on asymptotic distribution of the MLE may be generalized in many ways. First, the parameter \( \theta \) does not have to be univariate. Second, the observations do not have to be i.i.d. Let \( \theta \) be a multi-dimensional vector of parameters; that is, \( \theta \in \Theta \subset \mathbb{R}^p \, (p \geq 1) \). Let \( X_1, \ldots, X_n \) be observations whose joint pdf with respect to a measure \( \mu \) depends on \( \theta \), denoted by \( f(x | \theta) \), where \( x = (x_1, \ldots, x_n)' \). Then under some regularity conditions, the MLE of \( \theta \) satisfies the likelihood equation \( \partial l/\partial \theta = 0 \), where \( l(\theta; X) = \log \{ f(X | \theta) \} \) is the log-likelihood function with \( X = (X_1, \ldots, X_n)' \). By the multivariate Taylor expansion (4.12), we have

\[
0 = \frac{\partial l(\hat{\theta}; X)}{\partial \theta_i} = \frac{\partial l(\hat{\theta}; X)}{\partial \theta_i} + \left\{ \frac{\partial^2 l(\hat{\theta}; X)}{\partial \theta_i \partial \theta_i} \right\} (\hat{\theta} - \theta)
+ \frac{1}{2} (\hat{\theta} - \theta)' \left\{ \frac{\partial^3 l(\hat{\theta(i)}; X)}{\partial \theta_i \partial \theta_i} \right\} (\hat{\theta} - \theta),
\]

\( i = 1, \ldots, p \), where \( \theta_i \) is the \( i \)th component of \( \theta \) and \( \hat{\theta}(i) \) lies between \( \theta \) and \( \hat{\theta} \). Note that \( \hat{\theta}(i) \) depends on \( i \) (which is something that one might overlook).
4.8 Case study: The Prasad–Rao method

Surveys are usually designed to produce reliable estimates of various characteristics of interest for large geographic areas. However, for effective planning of health, social, and other services and for apportioning government funds, there is a growing demand to produce similar estimates for small geographic areas and subpopulations. The usual design-based estimator, which uses only the sample survey data for the particular small area of interest, is unreliable due to relatively small samples available from the area. In the absence of a reliable small-area design-based estimator, one may alternatively use a synthetic estimator (Rao 2003, Section 4.2), which utilizes data from censuses or administrative records to obtain estimates for small geographical areas or subpopulations. Although the synthetic estimators are known to have smaller variances compared to the direct survey estimators, they tend to be biased as they do not make use of the information on the characteristic of interest directly obtainable from sample surveys.

A compromise between the direct survey and the synthetic estimations is the method of composite estimation, which uses sample survey data in conjunction with different census and administrative data. Implicit or explicit models, which "borrow strength" from related sources, have been used in this latter approach. Research in this and related areas are usually called small area estimation. See Rao (2003) for a detailed account of different composite estimation and other techniques in small area estimation.
An explicit (linear) model for composite small area estimation may be expressed as follows:

\[ Y_i = X_i \beta + Z_i \upsilon_i + \epsilon_i, \quad i = 1, \ldots, m, \]  

(4.85)

where \( m \) is the number of small areas; \( Y_i \) represents the vector of observations from the \( i \)th small area; \( X_i \) is a matrix of known covariates for the \( i \)th small area, and \( \beta \) is a vector of unknown regression coefficients (the fixed effects); \( Z_i \) is a known matrix, and \( \upsilon_i \) is a vector of small-area specific random effects; and \( \epsilon_i \) represents a vector of sampling errors. It is assumed that \( Y_i \sim N \left( X_i \beta, \sigma^2 \right) \) and \( \epsilon_i \sim N \left( 0, \sigma^2 \right) \).

**Example 4.18 (The Fay–Herriot model).** Fay and Herriot (1979) proposed the following model for the estimation of per-capita income of small places with population sizes less than 1000:

\[ Y_i = x_i' \beta + \upsilon_i + \epsilon_i, \]  

(4.86)

\( i = 1, \ldots, m, \) where \( x_i \) is a vector of known covariates, \( \beta \) is a vector of unknown regression coefficients, \( \upsilon_i \) is area-specific random effects, and \( \epsilon_i \) represents sampling errors. It is assumed that \( \upsilon_i \) and \( \epsilon_i \) are independent with \( \upsilon_i \sim N \left( 0, D_i \right) \) and \( \epsilon_i \sim N \left( 0, D_i \right) \). The variance \( D \) is unknown, but the sampling variances \( D_i \) are assumed known. It is easy to show that the Fay–Herriot model is a special case of the general small-area model (4.85) (Exercise 4.25).

**Example 4.19 (The nested-error regression model).** Battese, Harter, and Fullerton (1988) presented data from 12 Iowa counties obtained from the 1978 June Enumerative Survey of the U.S. Department of Agriculture as well as data obtained from land observatory satellites on corn areas involving corn and soybeans. The objective was to predict the mean hectares of corn and soybeans per segment for the 12 counties using the satellite information. The authors introduced the following model, known as the nested-error regression model, for the prediction problem:

\[ Y_{ij} = x_{ij}' \beta + \upsilon_i + \epsilon_{ij}, \]  

(4.87)

\( i = 1, \ldots, m, \quad j = 1, \ldots, n_i \), where \( x_{ij} \) is a known vector of covariates, \( \beta \) is an unknown vector of regression coefficients, \( \upsilon_i \) is a random effect associated with the \( i \)th small area, and \( \epsilon_{ij} \) is the sampling error. It is assumed that the random effects are independent and distributed as \( N \left( 0, \sigma^2 \right) \), the sampling errors are independent and distributed as \( N \left( 0, \sigma^2 \right) \), and the random effects and sampling errors are uncorrelated. It can be shown that this is, again, a special case of the general small-area model (4.85) (Exercise 4.26).

The problem of main interest in the small-area estimation is usually the estimation, or prediction, of small-area means. A small-area mean may be expressed, at least approximately, as a mixed effect, \( \eta = \beta \beta' + \alpha \upsilon \), where \( \alpha \) and \( \beta \) are known vectors and \( \upsilon \) is the vector of fixed and random effects, respectively, in (4.85) (it is called a mixed effect because it is a combination of fixed and random effects). If \( \beta \) and \( \upsilon \) are both known, the best predictor (BP) for \( \eta \) is the conditional expectation \( E(\eta | Y) \). Furthermore, if the random effects \( \upsilon_i \) and errors \( \epsilon_i \) are normally distributed, this conditional expectation is given by

\[ \eta^* = \beta \beta' + \alpha E(\upsilon | Y) \]

\[ = \beta \beta' + \alpha GZ'V^{-1}(Y - X\beta), \]

where \( X = (X_{ij})_{1 \leq i \leq m} \), \( Y = (Y_{ij})_{1 \leq i \leq m} \), \( G = \text{diag}(G_1, \ldots, G_m) \), \( Z_i = \text{diag}(Z_{i1}, \ldots, Z_{in_i}) \), and \( V = \text{Var}(\upsilon) = \text{diag}(V_1, \ldots, V_m) \) with \( V_i = Z_iG_iZ_i' + R_i \).

In the absence of the normality assumption, \( \eta^* \) is the best linear predictor (BLP) of \( \eta \) in the sense that it minimizes the mean squared prediction error (MSPE) of a predictor that is linear in \( Y \) (e.g., Jiang 2007, Section 2.3). Of course, \( \beta \) is unknown in practice. It is then customary to replace \( \beta \) by

\[ \bar{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}Y, \]  

(4.88)

which is the MLE of \( \beta \) under the normality assumption, provided that \( \psi \) is known. The result is called the best linear unbiased predictor, or BLUP, denoted by \( \eta \). In other words, \( \eta \) is given by \( \eta^* \) with \( \beta \) replaced by \( \bar{\beta} \).

The expression of BLUP involves \( \psi \), the vector of variance components, which is typically unknown in practice. It is then customary to replace \( \psi \) by a consistent estimator, \( \psi \). The resulting predictor is often called the empirical BLUP, or EBLUP, denoted by \( \hat{\eta} \). To illustrate the EBLUP procedure, we consider a previous example.

**Example 4.18 (continued).** Consider the Fay–Herriot model. Let \( \eta \) denote the small-area mean for the \( i \)th area; that is, \( \eta = x_i' \beta + \upsilon_i \). Then the BP for \( \eta \) is given by (Exercise 4.25)

\[ \eta^* = (1 - B_i)Y_i + B_i x_i' \beta, \]

where \( B_i = D_i/(A + D_i) \). The BLUP is given by \( \hat{\eta} = \eta^* \) with \( \beta \) replaced by

\[ \bar{\beta} = \left( \sum_{i=1}^{m} \frac{x_i x_i'}{A + D_i} \right)^{-1} \left( \sum_{i=1}^{m} \frac{x_i y_i}{A + D_i} \right). \]

Finally, the EBLUP is given by

\[ \hat{\eta} = (1 - \hat{B}_i)Y_i + \hat{B}_i x_i' \bar{\beta}, \]
where $\tilde{B}_i$ and $\tilde{b}$ are $B_i$ and $b$, respectively, with $A$ replaced by $\tilde{A}$, a consistent estimator of $A$. One example of a consistent estimator of $A$ is the method of moments (MoM) estimator proposed by Prasad and Rao (1990), given by

$$\begin{equation}
\tilde{A} = \frac{Y'P_XY - \operatorname{tr}(P_XD)}{m - p},
\end{equation}$$

where $P_X = I - P_X$, with $P_X = X(X'X)^{-1}X'$, and $D = \operatorname{diag}(D_1, \ldots, D_m)$.

Although the EBLUP is fairly easy to obtain, assessing its uncertainty is quite a challenging problem. As mentioned, a measure of the uncertainty that is commonly used is the MSPE. However, unlike the BLUP, the MSPE of the EBLUP does not, in general, have a closed-form expression. This is because once the variance components $\psi$ are replaced by their (consistent) estimators, the predictor is no longer linear in $Y$. A naive approach to estimation of the MSPE of EBLUP would be to first obtain the MSPE of BLUP, which can be expressed in closed-form as a function of $\psi$ (see below), and then replace $\psi$ by $\hat{\psi}$ in the expression of the MSPE of BLUP, where $\hat{\psi}$ is the consistent estimator of $\psi$. However, as will be seen, this approach underestimate the MSPE of EBLUP, as it does not take into account the additional variation associated with the estimation of $\psi$.

Prasad and Rao (1990) proposed a method based on the Taylor series expansion to produce the second-order unbiased MSPE estimator for EBLUP. Here, the term "second-order unbiased" is with respect to the above naive MSPE estimator, which is first-order unbiased. The latter property is because, roughly speaking, the difference between the BLUP and EBLUP is of the order $O(m^{-1/2})$. To see this, note that the BLUP can be expressed as

$$\begin{equation}
\tilde{\eta} = \tilde{\eta}(\psi) = \tilde{b} + \tilde{a}'GZV^{-1}(Y - X\tilde{b}),
\end{equation}$$

where $\tilde{b}$ is given by (4.88). It follows that the EBLUP is simply $\tilde{\eta} = \tilde{\eta}(\hat{\psi})$. By the Taylor expansion, we have $\tilde{\eta}(\hat{\psi}) - \tilde{\eta}(\hat{\psi}) \approx (\partial \tilde{\eta}/\partial \hat{\psi})(\hat{\psi} - \psi) = O_p(m^{-1/2})$ under some regularity conditions. Therefore, $E(\tilde{\eta} - \hat{\eta})^2$ is typically of the order $O(m^{-1})$. On the other hand, Kacker and Harville (1984) showed that under the normality assumption,

$$\begin{equation}
\text{MSPE}(\tilde{\eta}) = E(\tilde{\eta} - \hat{\eta})^2 = E(\tilde{\eta} - \hat{\eta})^2 + E(\tilde{\eta} - \hat{\eta})^2 = \text{MSPE}(\hat{\eta}) + E(\tilde{\eta} - \hat{\eta})^2.
\end{equation}$$

Equation (4.90) clearly suggests that the naive MSPE estimator underestimates the true MSPE, because it only takes into account the first term on the right side. Furthermore, if one replaces $\psi$ by $\hat{\psi}$ in the expression of $\text{MSPE}(\tilde{\eta})$, it introduces a bias of the order $O(m^{-1})$ [not $O(m^{-1/2})$]. Thus, the bias of the naive MSPE estimator is $O(m^{-1})$. By second-order unbiasedness of the Prasad–Rao method, it means that

$$\begin{equation}
E\left(\text{MSPE} - \text{MSPE}\right) = o(m^{-1}),
\end{equation}$$

where $\text{MSPE} = \text{MSPE}(\hat{\eta})$ and $\text{MSPE}$ represents the Prasad–Rao estimator of MSPE. Furthermore, the following closed-form expression can be obtained (Exercise 4.27):

$$\begin{equation}
\text{MSPE}(\hat{\eta}) = a'(G - GZV^{-1}ZG)a + \tilde{d}'(X'V^{-1}X)'\tilde{d},
\end{equation}$$

where $\tilde{d} = b - X'V^{-1}ZG$. Here, we assume that $X$ is of full rank $p$. Note that, typically, the first term on the right side of (4.92) is $O(1)$ and the second-term is $O(m^{-1})$. An implication is that $\text{MSPE}(\hat{\eta}) = O(1)$. In view of (4.90) and (4.92), a main part of the Prasad–Rao method is therefore to derive an approximation to $E(\tilde{\eta} - \hat{\eta})^2$. Assume that the suitable regularity conditions are satisfied. Then we have, by the Taylor expansion and (4.89),

$$\begin{equation}
\tilde{\eta} - \hat{\eta} = \tilde{\eta}(\hat{\psi}) - \tilde{\eta}(\psi) = \frac{\partial \tilde{\eta}}{\partial \psi}(\hat{\psi} - \psi) + \frac{1}{2}(\hat{\psi} - \psi)\frac{\partial^2 \tilde{\eta}}{\partial \psi^2}(\hat{\psi} - \psi),
\end{equation}$$

where $\hat{\psi}$ lies between $\psi$ and $\hat{\psi}$. Suppose that $\hat{\psi}$ is a $\sqrt{m}$-consistent estimator in the sense that $\sqrt{m}(\hat{\psi} - \psi) = o_p(1)$ (see Section 3.4, above Example 3.7), and the following hold:

$$\frac{\partial \tilde{\eta}}{\partial \psi} = O_p(1), \quad \sup_{|\psi - \hat{\psi}| \leq |\hat{\psi} - \psi|} \frac{\partial^2 \tilde{\eta}}{\partial \psi^2} = O_p(1).$$

Then by the method of formal derivation (see Section 4.3), we have

$$\begin{equation}
E(\tilde{\eta} - \hat{\eta})^2 = E\left\{\frac{\partial \tilde{\eta}}{\partial \hat{\psi}}(\hat{\psi} - \psi)\right\}^2 + o(m^{-1}).
\end{equation}$$

Now, suppose the first term on the right side of (4.93) can be expressed as

$$\begin{equation}
E\left\{\frac{\partial \tilde{\eta}}{\partial \hat{\psi}}(\hat{\psi} - \psi)\right\}^2 = \frac{a(\psi)}{m} + o(m^{-1}),
\end{equation}$$

where $a(\cdot)$ is a known differentiable function. Also, let $b(\psi)$ denote the right side of (4.92). By (4.93) and (4.94), to obtain a second-order unbiased estimator of $E(\tilde{\eta} - \hat{\eta})^2$, all one needs to do is to replace $\psi$ by $b(\tilde{\psi})$ because the resulting bias is $O(m^{-1})$ (why?). However, one cannot use the same strategy to estimate $\text{MSPE}(\hat{\eta}) = b(\psi)$, because the resulting bias is $O(m^{-1})$ rather than $O(m^{-1})$. In order to reduce the latter bias to $O(m^{-1})$, we use the following bias correction procedure. Note that, by the Taylor expansion, we have
Asymptotic Expansions

$$b(\hat{\psi}) = b(\psi) + \frac{\partial b}{\partial \psi}(\psi - \hat{\psi}) + \frac{1}{2} (\psi - \hat{\psi})' \frac{\partial^2 b}{\partial \psi \partial \psi'}(\psi - \hat{\psi}) + o(m^{-1});$$

hence, by the method of formal derivation (Section 4.3),

$$E\{b(\hat{\psi})\} = b(\psi) + \frac{\partial b}{\partial \psi} E(\psi - \hat{\psi}) + \frac{1}{2} E \left\{ (\psi - \hat{\psi})' \frac{\partial^2 b}{\partial \psi \partial \psi'}(\psi - \hat{\psi}) \right\} + o(m^{-1})$$

$$= b(\psi) + \frac{c(\psi)}{m} + o(m^{-1}).$$

Here, we make the assumption that $E(\psi - \hat{\psi}) = O(m^{-1})$, which holds under regularity conditions. Now, we can apply the same plug-in technique used above for estimating $a(\psi)$ to the estimation of $c(\psi)$. In other words, we estimate $b(\psi)$ by $b(\psi) - c(\psi)/m$ because the bias of this estimator is

$$E\left\{ b(\hat{\psi}) - \frac{c(\hat{\psi})}{m} \right\} = b(\psi) + \frac{c(\psi)}{m} + o(m^{-1}) - \frac{E(c(\psi))}{m} - c(\psi)$$

$$= \frac{E(c(\psi) - c(\psi))}{m} + o(m^{-1})$$

$$= o(m^{-1}),$$

provided that $c(\cdot)$ is a smooth (e.g., differentiable) function.

In conclusion, if we define the Prasad–Rao estimator as

$$\text{MSPE} = b(\psi) + \frac{a(\hat{\psi}) - c(\hat{\psi})}{m},$$

then we have

$$E(\text{MSPE}) = E\{b(\hat{\psi})\} + \frac{E\{a(\hat{\psi})\}}{m} - \frac{E\{c(\psi)\}}{m}$$

$$= b(\psi) + \frac{c(\psi)}{m} + o(m^{-1})$$

$$+ \frac{a(\psi)}{m} - \frac{E(a(\psi) - a(\psi))}{m}$$

$$- \frac{c(\psi) - E(c(\psi) - c(\psi))}{m}$$

$$= b(\psi) + \frac{a(\psi)}{m} + o(m^{-1})$$

$$= \text{MSPE}(\bar{\eta}) + E(\bar{\eta} - \bar{\eta})^2 + o(m^{-1})$$

$$= \text{MSPE} + o(m^{-1}),$$

using (4.90), (4.92), and (4.93) near the end. Therefore, (4.91) holds.

Prasad and Rao (1990) obtained a detailed expression of (4.94) and, hence, (4.95) for the two special cases discussed earlier—that is, the Fay–Herriot model (Example 4.18) and the nested-error regression model (Example 4.19), assuming normality and using the MoM estimators of $\psi$. Extensions of the Prasad–Rao method will be discussed in Chapters 12 and 13.

4.9 Exercises

4.1. Regarding Table 4.1, how large should $n$ be in order to achieve the same accuracy (in terms of the relative error) for $x = 5$? 4.2. This is regarding Example 4.3.

(a) Show that

$$l(0) = -\sum_{i=1}^n \{X_i + 2 \log(1 + e^{-X_i})\},$$

$$l'(0) = \sum_{i=1}^n \frac{1 - e^{-X_i}}{1 + e^{-X_i}},$$

$$l''(\theta) = -2 \sum_{i=1}^n \frac{e^{\theta - X_i}}{(1 + e^{\theta - X_i})^2}.$$

(b) Show that $n^{-1}l(\theta) \overset{P}{\to} a$ as $n \to \infty$, where $a$ is a positive constant.

(c) Show that $n^{-1/2}l'(\theta) \overset{d}{\to} N(0, \sigma^2)$ as $n \to \infty$, and determine $\sigma^2$.

(d) Show that there is a sequence of positive random variables $\xi_n$ and a constant $c > 0$ such that $\xi_n \overset{P}{\to} b$, where $b$ is a positive constant, and

$$\xi_n n \leq \sup_{\theta} |l''(\theta)| \leq cn.$$

4.3. In Example 4.4, show that

$$\frac{\partial g(c_n)}{\partial x} - \frac{\partial g(c)}{\partial x} = o_P(1),$$

where $\partial g/\partial x = (\partial g/\partial x')'$.

4.4. Let $X_1, \ldots, X_n$ be i.i.d. observations such that $E(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2$, where $0 < \sigma^2 < \infty$. Derive the (three) results at the end Section 4.2.

4.5. Let $X_1, \ldots, X_n$ be i.i.d. observations generated from the following distributions, where $n = 30$. Construct the histograms of the empirical distribution of $X$ based on 10,000 simulated values. Does the population distribution of $X$'s make a difference?

(i) $N(0, 1)$;

(ii) Uniform[0, 1];
4.14. Prove the (identity) expansions (4.55) and (4.58) by mathematical induction.
4.15. Show that in Example 4.15 we have $E(\eta_{ij}) = (p^2 + 1)i(\varepsilon_{ij})$, $1 \leq i, j \leq p$.
4.16. Show that in Example 4.15 we have $E(\zeta_{ij}) = 0$ if $i \neq j$ and $E(\zeta_{ii})$ does not depend on $i$.
4.17. Suppose that $X$ has a $\chi^2_\nu$-distribution, where $\nu > 2$. Use the elementary expansion (4.55) with $l = 4$ and without the remaining term to approximate $E(X^{-1})$. Note that closed-form expressions of moments of $X$, including $E(X^{-2})$, can be obtained, so that one can directly compare the accuracy of the approximation. Does the approximation improve as $\nu \to \infty$? (Hint: Consider the relative error of the approximation defined as approximate − exact)/exact.)
4.18. Derive the approximation (4.63) using the Taylor expansion

$$q(x) = q(\bar{x}) + \frac{1}{2}q''(\bar{x})(x - \bar{x})^2 + \cdots.$$ 

4.19. Derive the Laplace approximation (4.64). What is the constant $c$?
4.20. This exercise is related to Example 4.16.
(i) Show that in this case the exact value of (4.62) is given by

$$\frac{\sqrt{\pi} \Gamma\left(\frac{5}{2}\right)}{\Gamma\left(\frac{\nu + 1}{2}\right)},$$

and the Laplace approximation (4.63) is

$$\sqrt{\frac{2\nu\pi}{\nu + 1}}.$$

(ii) Show that if $q(x) = (x - \mu)^2/2\sigma^2$ for some $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, the Laplace approximation (4.63) is exact.
4.21. Show that the likelihood function in Example 4.17 can be expressed (4.66).
4.22. Show that in the i.i.d. case, the amount of information contained in $X_1, \ldots, X_n$ is $n$ times that contained in $X_1$; that is, $I(\theta) = nI(\theta)$ [see (4.75) and (4.76)]. The result requires some regularity conditions to hold. What regularity conditions?
4.23. Let $X_1, \ldots, X_n$ be i.i.d. observations with the pdf or pmf $f(x|\theta)$, where $\theta$ is a univariate parameter. Here, the pdf is with respect to the Lebesgue measure, whereas the pmf may be regarded as a pdf with respect to the counting measure [see below (4.36)]. Obtain the Fisher information (4.75) for the following cases:
(i) $X_1 \sim \text{Bernoulli}(\theta)$, so that

$$f(x|\theta) = \theta^x(1 - \theta)^{1-x}; \quad x = 0, 1,$$
where \( \theta \in (0, 1) \).

(ii) \( X_1 \sim \text{Poisson}(\theta) \), so that
\[
f(x | \theta) = e^{-\theta} \frac{\theta^x}{x!}, \quad x = 0, 1, \ldots,
\]
where \( \theta > 0 \).

(iii) \( X_1 \sim \text{Exponential}(\theta) \), so that
\[
f(x | \theta) = \frac{1}{\theta} e^{-x/\theta}, \quad x \geq 0,
\]
where \( \theta > 0 \).

(iv) \( X_1 \sim \text{Normal}(\theta_1, \theta_2^2) \), so that
\[
f(x | \theta) = \frac{1}{\sqrt{2\pi \theta^2}} \exp \left\{ -\frac{(x - \theta)^2}{2\theta^2} \right\}, \quad -\infty < x < \infty,
\]
where \( \theta \in (-\infty, \infty) \).

4.24. Let \( X_1, \ldots, X_n \) be i.i.d. with the following pdf or pmf depending on \( \theta = (\theta_1, \theta_2) \). Obtain the Fisher information matrix (4.78) in each case.

(i) \( X_1 \sim \text{Normal}(\mu, \sigma^2) \), where \( \mu \in (-\infty, \infty) \) and \( \sigma^2 > 0 \), so that \( \theta_1 = \mu \) and \( \theta_2 = \sigma^2 \).

(ii) \( X_1 \sim \text{Gamma}(\alpha, \beta) \), where \( \alpha > 0 \) and \( \beta > 0 \) are known as the shape and scale parameters, respectively, so that \( \theta_1 = \alpha \) and \( \theta_2 = \beta \).

(iii) \( X_1 \sim \text{Beta}(\alpha, \beta) \), whose pdf is given by
\[
f(x | \alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1}, \quad 0 < x < 1,
\]
where \( \alpha > 0 \) and \( \beta > 0 \), so that \( \theta_1 = \alpha \) and \( \theta_2 = \beta \).

4.25. Show that the Fay–Herriot model of Example 4.18 is a special case of the small-area model (4.85). Specify the matrices \( X_1, Z_1, G_i, \) and \( R_i \) in this case. Furthermore, show that the BP for \( \eta = x_i(\beta + v_i) \) is given by \( \bar{\eta} = (1 - B_1) \eta + B_1 x_i \beta \), where \( B_i = D_i / (A + D_i) \).

4.26. Show that the nested-error regression model of Example 4.19 is a special case of the small-area model (4.85). Specify the matrices \( X_1, Z_1, G_i, \) and \( R_i \) in this case.

4.27. Derive the expression (4.92) for \( \text{MSPE}(\bar{\eta}) \).

4.28. Consider a special case of the Fay–Herriot model (Example 4.18) in which \( D_i = D, 1 \leq i \leq m \). This is known as the balanced case. Without loss of generality, let \( D = 1 \). Consider the prediction of \( \eta_i = x_i(\beta + v_i) \). Let \( \bar{\eta}_i \) and \( \bar{\eta}_k \) denote the BLUP and EBLUP, respectively, where the MoM estimator of \( \bar{\eta}_i \) is used for the EBLUP [see Example 4.18 (continued) or Exercise 4.25].

(i) Show that
\[
\text{MSPE}(\bar{\eta}_k) = \frac{A}{A + 1} + \frac{x_i(X'X)^{-1}x_i}{A + 1},
\]
where \( X = (x_i'')_{1 \leq i \leq m} \).

(ii) Show that
\[
\text{MSPE}(\bar{\eta}_k) = \frac{A}{A + 1} + \frac{x_i'(X'X)^{-1}x_i + 2(1 - x_i'(X'X)^{-1}x_i)}{(A + 1)(m - p)} + \frac{4(1 - x_i'(X'X)^{-1}x_i)}{(A + 1)(m - p)(m - p - 2)}.
\]
[Hint: The moment of \( \chi_k^2 \) has a closed-form expression, where \( \chi_k^2 \) denotes a random variable with a \( \chi_k^2 \)-distribution. Find the expression.]

(iii) Let \( \eta = (\eta_i)_{1 \leq i \leq m} \) denote the vector of small-area means and \( \bar{\eta}_i \) denote the vector of EBLUPs. Define the overall MSPE of the EBLUP as \( \text{MSPE}(\bar{\eta}_i) = \mathbb{E}(\|\bar{\eta} - \eta\|^2) = \sum_{i=1}^m \mathbb{E}((\bar{\eta}_i - \eta_i)^2) = \sum_{i=1}^m \mathbb{E}(\bar{\eta}_i - \eta_i)^2 = \sum_{i=1}^m \text{MSPE}(\bar{\eta}_i) \). Show that
\[
\text{MSPE}(\bar{\eta}) = \frac{x_i(A + p + 2)}{A + 1} + \frac{4}{(A + 1)(m - p - 2)}.
\]

4.29 [Delta method (continued)]. In Example 4.4 we introduced the delta method for distributional approximations. The method can also be used for moment approximations. Let \( T_1, \ldots, T_k \) be random variables whose means and variances exist. Let \( g(t_1, \ldots, t_k) \) be a differentiable function. Then, by the Taylor expansion, we can write
\[
g(T_1, \ldots, T_k) \approx g(\mu_1, \ldots, \mu_k) + \sum_{i=1}^k \frac{\partial g}{\partial t_i} (T_i - \mu_i),
\]
where \( \mu_i = \mathbb{E}(T_i), 1 \leq i \leq k \), and \( \partial g / \partial t_i \) is evaluated as \( (\mu_1, \ldots, \mu_k) \). This leads to the following approximations:
\[
\mathbb{E}\{g(T_1, \ldots, T_k)\} \approx g(\mu_1, \ldots, \mu_k),
\]
\[
\text{var}\{g(T_1, \ldots, T_k)\} \approx \sum_{i=1}^k \left( \frac{\partial g}{\partial t_i} \right)^2 \text{var}(T_i) + 2 \sum_{i<j} \left( \frac{\partial g}{\partial t_i} \right) \left( \frac{\partial g}{\partial t_j} \right) \text{cov}(T_i, T_j).
\]

(i) Suppose that \( T \sim \text{Gamma}(\alpha, \beta) \) with the pdf given in Exercise 4.24(ii). Use the above delta method to approximate the mean and variance of \( T^{-1} \).
(ii) Note that the exact mean and variance of $T^{-1}$ can be obtained in this case, given a suitable range of $\alpha$. What is the range of $\alpha$ so that $E(T^{-1})$ exists? What is the range of $\alpha$ so that $E(T^{-2})$ exists?

(iii) Obtain the exact mean and variance of $T^{-1}$ given the suitable range of $\alpha$ and compare the results with the above delta-method approximations. How do the values of $\alpha$ and $\beta$ affect the accuracy of the approximations?

4.30. Let $X_1, \ldots, X_n$ be i.i.d. such that $E(X_1) = 0$, $E(X_i^2) = 1$, and $E(X_i^4) < \infty$. Consider approximation to the mean and variance of

$$Y = \frac{n}{n + \sum_{i=1}^{n} X_i^2}$$

using the delta method of Exercise 4.29.

(i) Let $g(x_1, \ldots, x_n) = n/(n + \sum_{i=1}^{n} x_i^2)$. What are the approximations to the mean and variance of $Y = g(X_1, \ldots, X_n)$?

(ii) If we let $g(t_1, \ldots, t_n) = n/(n + \sum_{i=1}^{n} t_i^2)$, and $T_i = X_i^2$, $1 \leq i \leq n$, what are the approximations to the mean and variance of $Y = g(T_1, \ldots, T_n)$?

(iii) How does the sample size $n$ affect the approximation to $E(Y)$? In other words, does the accuracy of the approximation improve as $n$ increases? [Hint: First use the dominated convergence theorem (Theorem 2.16) to show that $E(Y)$ converges to a limit as $n \to \infty$.]

(iv) Which approximation [(i) or (ii)] do you think is better? Any general comment(s) on the use of the delta method in moment approximations?

5

5.1 Introduction

It is said that high school algebra is characterized by equalities, whereas college and more advanced mathematics, inequalities. One may argue that, in a similar way, statistics is characterized by inequalities, too. For example, the words “margin of errors,” which nowadays come along routinely with survey results that are published, may be viewed as bounds for typical range of the sampling error. This may be expressed as (i) $P(|\varepsilon| \leq b) = 1 - \alpha$ or (ii) $P(|\varepsilon| \leq b) \geq 1 - \alpha$, where $\varepsilon$ represents the sampling error, $b$ is an upper bound, and $\alpha$ is a small positive number, such as 0.05. In (i), the event inside the probability is characterized by an inequality, whereas in (ii), both the event and the probability itself are characterized by inequalities.

Perhaps the simplest of all is the following basic triangle inequality:

$$|x + y| \leq |x| + |y|$$

(5.1)

for all $x$ and $y$. Inequalities such as (5.1) are called numerical inequalities, meaning that they hold for all real numbers. Many of the numerical inequalities can be extended beyond real numbers. For example, extensions of numerical inequalities to matrices have led to many of the matrix inequalities. However, not every numerical inequality has its matrix analogue. For example, if $A$ and $B$ are symmetric matrices such that $A \geq B$, meaning that $A - B$ is nonnegative definite, it is not necessarily true that $A^2 \geq B^2$. See Section 5.3.2 for more counterexamples. Numerical inequalities can be used to establish more sophisticated inequalities, such as moment and probability inequalities,
but this is not always the case. For example, the covariance inequality states that for any random variables $X$ and $Y$, one has
\[ \text{cov}(X, Y) \leq \sqrt{\text{var}(X) \text{var}(Y)}, \]  
(5.2)
or $E[(X - E(X)) \{Y - E(Y)\}] \leq \{\text{var}(X)\}^{1/2} \{\text{var}(Y)\}^{1/2}$, but this is not derived from $\{X - E(X)\} \{Y - E(Y)\} \leq \{\text{var}(X)\}^{1/2} \{\text{var}(Y)\}^{1/2}$, which, of course, does not always hold.

Like the Taylor expansion, the value of inequalities to statistics cannot be overstated. There exist a huge number of inequalities: numerical inequalities, matrix inequalities, integral/moment inequalities, and probability inequalities. Instead of trying to come up with a list of all useful inequalities, which is impossible, we focus on developing the basic techniques for making use of existing inequalities and for developing new inequalities. We believe that these methods and techniques are even more important in solving the current and future problems than the inequalities themselves. We should also point out that although this chapter is entitled “Inequalities,” it by no means includes all of the inequalities introduced in this book. However, this is the only place that these materials are treated systematically as a single subject.

5.2 Numerical inequalities

5.2.1 The convex function inequality

The triangle inequality (5.1), of course, can be derived with an elementary argument. Since $x \leq |x|$ and $y \leq |y|$, we have $x + y \leq |x| + |y|$. Similarly, $-x - y \leq |x| + |y|$, or $x + y \geq -(|x| + |y|)$, which leads to (5.1). Alternatively, the (5.1) is a special case of the convex function inequality. A real-valued function $f(x)$ is convex if for any $x$, $y$, and $\lambda \in [0, 1]$, we have
\[ f((1 - \lambda)x + \lambda y) \leq (1 - \lambda)f(x) + \lambda f(y). \]  
(5.3)
Here, we did not specify the range of $x$, $y$. Typically, it is assumed that $x$, $y \in D$, where $D$ is a convex subset of $R$ in the sense that $x, y \in D$ implies $(1 - \lambda)x + \lambda y \in D$ for any $\lambda \in [0, 1]$.

To show that $f(x)$ is convex, one can, of course, verify (5.3) for any $x$, $y \in D$ and $\lambda \in [0, 1]$, but sometimes there are easier ways. For example, if $f'(x)$ exists, a necessary and sufficient condition for $f(x)$ to be a convex function is that $f''(x)$ is nondecreasing; if $f''(x)$ exists, then a necessary and sufficient condition for $f(x)$ to be convex is that $f''(x) \geq 0$.

A concave function may be thought of as a function that has the reversed properties of a convex function; that is, $f(x)$ is concave if and only if $f(-x)$ is concave. More generally, let $g(x)$ be a linear function of $x$; that is, $g(x) = ax + b$ for some constants $a$ and $b$. Then $f(x)$ is convex if and only if $g(x) - f(x)$ is concave. Therefore, any convex function inequality (see below) can be reversed for a concave function inequality.

The best know property of a convex function $f(x)$ is the following:
\[ f\left(\frac{x_1 + \cdots + x_n}{n}\right) \leq \frac{f(x_1) + \cdots + f(x_n)}{n}, \]  
(5.4)
for any $x_1, \ldots, x_n \in D$. To see that (5.1) is a special case of (5.4), note that $f(x) = |x|$ is a convex function; hence, by (5.4), we have
\[ \frac{x + y}{2} \leq \frac{|x| + |y|}{2}, \]
which is the same as (5.1). Note that, in this case, the convex function approach does not really make the derivation simpler if one takes into account that the verification of (5.3) takes about the same as the arguments right above it. However, in many other cases, the convex function approach is very effective. We consider some examples.

Example 5.1 (Arithmetic, geometric and harmonic means). The harmonic mean is bounded by the geometric mean, which, in turn, is bounded by the arithmetic mean. This string of fundamental inequalities can be expressed as
\[ \frac{x_1 + \cdots + x_n}{n} \leq \sqrt[n]{x_1 \cdots x_n} \leq \frac{x_1 + \cdots + x_n}{n}, \]
for any positive numbers $x_1, \ldots, x_n$. Both inequalities can be established by the convex function inequality. Let $f(x) = -\log(x)$. Then since $f''(x) = x^{-2} > 0$ for $x > 0$, the function is convex. Therefore, by (5.4), we have
\[ -\log\left(\frac{x_1 + \cdots + x_n}{n}\right) \leq -\frac{\log(x_1) + \cdots + \log(x_n)}{n}, \]

Inequalities (5.5) then follow by taking the negative and then exponential.

Example 5.2 (The sample $p$-norm). For any finite sequence $x_i$, $1 \leq i \leq n$, and $p \geq 1$, the sample $p$-norm of the sequence is defined as
\[ \|\{x_i\}\|_p = \left(\frac{|x_1| + \cdots + |x_n|^p}{n}\right)^{1/p} \]

The word “sample” corresponds to the case where the $x_1, \ldots, x_n$ are realized values of i.i.d. observations, say, $X_1, \ldots, X_n$, whose $p$-norm is defined as $\|X_i\|_p = E(\|X_i\|_p)$. Another look at the sample $p$-norm is to consider the empirical distribution of $x_1, \ldots, x_n$ defined as
\[ F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{(x_i \leq x)}. \]

It follows that the sample \( p \)-norm is simply \( \|X\|_p \), where \( X \) has the empirical distribution \( F_n \) (verify this). A property of the sample \( p \)-norm is that it is nondecreasing in \( p \). In other words, \( p \leq q \) implies \( \|x_i\|_p \leq \|x_i\|_q \). To show this, we may assume, without loss of generality, that the \( x_i \)'s are positive (why?). Consider \( f(x) = x^{q/p} \). Then since \( f''(x) = (q(p-q)/p^2)x^{q/p-2} \geq 0 \) for \( x > 0 \), \( f(x) \) is convex. It follows by (5.4) that

\[
\left( \frac{x_1^q + \cdots + x_n^q}{n} \right)^{q/p} \leq \frac{(x_1^{q/p} + \cdots + x_n^{q/p})}{n},
\]

\[
= \frac{x_1^q + \cdots + x_n^q}{n}.
\]

The claimed property is then verified by taking the \( q \)-th root. Given \( x_1, \ldots, x_n \), since the sequence \( \|x_i\|_k, k = 1, 2, \ldots \), is nondecreasing, according to §1.5.1.3, the limit \( \lim_{k \to \infty} \|x_i\|_k \) exists if the sequence has an upper bound. In fact, it is easy to show directly that the limit is equal to \( \|x_i\|_\infty = \max_{1 \leq i \leq n} |x_i| \), which is called the \( \infty \)-norm of the sequence (Exercise 5.1).

An extended property of (5.4) is the following. If \( f(x) \) is convex, then for any \( x_1, \ldots, x_n \in D \) and \( \lambda_1, \ldots, \lambda_n \geq 0 \) such that \( \lambda_1 + \cdots + \lambda_n = 1 \), we have

\[ f(\lambda_1 x_1 + \cdots + \lambda_n x_n) \leq \lambda_1 f(x_1) + \cdots + \lambda_n f(x_n). \]  

(5.6)

Clearly, inequality (5.3), which defines a convex function, is a special case of (5.6) with \( n = 2 \). We consider a well-known example as an application of (5.6):

**Example 5.3** (Cauchy–Schwarz inequality). For any real numbers \( x_1, \ldots, x_n \) and \( y_1, \ldots, y_n \), we have

\[ (x_1 y_1 + \cdots + x_n y_n)^2 \leq (x_1^2 + \cdots + x_n^2)(y_1^2 + \cdots + y_n^2). \]

(5.7)

To show (5.7), assume, without loss of generality, that \( \sum_{i=1}^{n} y_i^2 > 0 \) [because otherwise, both sides of (5.7) are zero]. Define \( u_i = x_i/y_i \) if \( y_i \neq 0 \) and \( u_i = 0 \) if \( y_i = 0 \). Then it is easy to verify that \( u_i y_i^2 = x_i y_i \) and \( u_i y_i^2 \leq x_i^2 \), \( 1 \leq i \leq n \), (Exercise 5.2). Now, let \( \lambda_i = y_i^2 / \sum_{i=1}^{n} y_i^2, 1 \leq i \leq n \). Note that the \( \lambda_i \)'s satisfy the requirements of (5.5). Using the fact that \( f(x) = x^2 \) is a convex function, we have, by (5.5), \( (\sum_{i=1}^{n} y_i^2)^{-2}(\sum_{i=1}^{n} x_i y_i)^2 \leq (\sum_{i=1}^{n} \lambda_i u_i)^2 \leq (\sum_{i=1}^{n} \lambda_i u_i^2)^{-1} \sum_{i=1}^{n} x_i^2 \), which leads to (5.7).

Hardy, Littlewood, and Pólya (1934) outlined a beautiful argument showing that if \( f(x) \) is continuous, the defining inequality (5.3) is actually equivalent to the following seemingly weaker one:

\[ f\left( \frac{x+y}{2} \right) \leq \frac{f(x) + f(y)}{2}. \]

(5.8)

for any \( x \) and \( y \). Another important result is regarding when the equality holds in (5.6). The same authors showed that if \( f(x) \) is continuous, then (5.6) holds with \( \leq \) replaced by \( < \) unless either (i) all of the \( x_i \)'s are equal or (ii) \( f(x) \) is linear in an interval that contains \( x_1, \ldots, x_n \). Based on these results, the authors called \( f(x) \) strictly convex if (5.8) holds with \( \leq \) replaced by \( < \) for any \( x \) and \( y \) unless \( x = y \). In particular, if \( f(x) \) is twice differentiable and \( f''(x) > 0 \), then (5.6) holds with \( \leq \) replaced by \( < \) unless all of the \( x_i \)'s are equal.

**Example 5.1 (continued).** Recall in this case the convex function is \( f(x) = \log(x) \) and \( f''(x) = x^{-2} > 0, x > 0 \). Thus, the strict inequalities in (5.5) hold unless all of the \( x_i \)'s are equal.

**Example 5.2 (continued).** Suppose that at least one of the \( x_i \)'s is positive. Then, as in Example 5.2, we may focus on the positive \( x_i \)'s. Recall that, in this case, \( f(x) = x^{1/k} \) with \( f''(x) = [(l(l-k))/k^2]x^{1/k-2} > 0 \) for \( x > 0 \) if \( k < l \). It follows that \( \|x_i\|_k < \|x_i\|_1 \) if \( k < l \), unless all of the positive \( x_i \)'s are equal.

Although we may use a similar argument to find out when equality occurs in the Cauchy–Schwarz inequality (Example 5.3), we would rather leave this for the next subsection, in which a different method will be used to derive the conditions for the equality.

### 5.2 Hölder’s and related inequalities

The celebrated Hölder’s inequality states the following. Let \( \alpha, \beta, \ldots, \gamma \) be positive numbers such that \( \alpha + \beta + \cdots + \gamma = 1 \). Then for any nonnegative numbers \( a_i, b_i, \ldots, g_i, 1 \leq i \leq n \), we have

\[ \sum_{i=1}^{n} a_i^\alpha b_i^\beta \cdots g_i^\gamma \leq \left( \sum_{i=1}^{n} a_i \right)^\alpha \left( \sum_{i=1}^{n} b_i \right)^\beta \cdots \left( \sum_{i=1}^{n} g_i \right)^\gamma. \]

(5.9)

Moreover, the strict inequality \( < \) holds in (5.9) unless either (i) one factor on the right side is zero (e.g., all of the \( a_i \)'s are zero); or (ii) \( a_i, b_i, \ldots, g_i \) are all proportional (i.e., \( a_i b_j = a_j b_i, \ldots, a_i g_j = a_j g_i \), for all \( i \) and \( j \)).

An alternative expression is probably more familiar to statisticians. It follows. Let \( p, q, \ldots, r \) be positive numbers such that

\[ \frac{1}{p} + \frac{1}{q} + \cdots + \frac{1}{r} = 1. \]

(5.10)

Then (5.10) implies that \( p, q, \ldots, r \) are all greater than one. Then for any nonnegative numbers \( x_i, y_i, \ldots, z_i, 1 \leq i \leq n \), we have

\[ \begin{align*}
  &f(\frac{x+y}{2}) \leq \frac{f(x) + f(y)}{2}, \\
  &f(\frac{x+y}{2}) \leq \frac{f(x) + f(y)}{2}.
\end{align*}
\[ \sum_{i=1}^{n} x_{i} y_{i} \leq \left( \sum_{i=1}^{n} x_{i}^{p} \right)^{1/p} \left( \sum_{i=1}^{n} y_{i}^{q} \right)^{1/q} \left( \sum_{i=1}^{n} z_{i}^{r} \right)^{1/r} . \]  

(5.11)

Moreover, the strict inequality holds if \( p, q > 0 \) and \( p^{-1} + q^{-1} = 1 \), then for any \( x_{i}, y_{i} \geq 0, 1 \leq i \leq n \), we have

\[ \sum_{i=1}^{n} x_{i} y_{i} \leq \left( \sum_{i=1}^{n} x_{i}^{p} \right)^{1/p} \left( \sum_{i=1}^{n} y_{i}^{q} \right)^{1/q} . \]  

(5.12)

The strict inequality holds in (5.12) unless either the \( x_{i} \)'s are all zero, or the \( y_{i} \)'s are all zero, or \( x_{i}^{p} y_{j}^{q} = x_{j}^{p} y_{i}^{q} \), \( 1 \leq i, j \leq n \) (in other words, \( x_{i} \) and \( y_{i} \) are proportional).

A special case of (5.11) is, by far, the most popular (in fact, this is called Hölder's inequality in most books). If \( p, q > 0 \) and \( p^{-1} + q^{-1} = 1 \), then for any \( x_{i}, y_{i} \geq 0, 1 \leq i \leq n \), we have

\[ \sum_{i=1}^{n} x_{i} y_{i} \leq \left( \sum_{i=1}^{n} x_{i}^{p} \right)^{1/p} \left( \sum_{i=1}^{n} y_{i}^{q} \right)^{1/q} . \]  

(5.14)

There is an interpretation of (5.14) in terms of inner product and norms in Hilbert space. Consider the space \( \mathbb{R}^{n} \) with the inner product \( \langle x, y \rangle = \sum_{i=1}^{n} x_{i} y_{i} \) for \( x = (x_{i})_{1 \leq i \leq n} \) and \( y = (y_{i})_{1 \leq i \leq n} \in \mathbb{R}^{n} \). If we define the p-norm \( \|x\|_{p} = (\sum_{i=1}^{n} |x_{i}|^{p})^{1/p} \) (note that this is slightly different from the sample p-norm defined in Example 5.2). Then (5.14) simply means that

\[ \|x + y\|_{p} \leq \|x\|_{p} + \|y\|_{p} . \]  

(5.15)

Hölder's inequality can be used to establish another famous inequality: the Minkowski's inequality. The result is better stated in terms of the p-norm (see above) as follows. If \( p > 1 \), then for any \( x, y, \ldots, z \in \mathbb{R}^{n} \), we have

\[ \|x + y + \cdots + z\|_{p} \leq \|x\|_{p} + \|y\|_{p} + \cdots + \|z\|_{p} . \]  

(5.16)

To prove (5.16), it suffices to show

\[ \|x + y\|_{p} \leq \|x\|_{p} + \|y\|_{p} . \]  

(5.17)

for any \( x \) and \( y \) (why?). We have, by (5.12),

\[ \sum_{i} (x_{i} + y_{i})^{p} = \sum_{i} x_{i}^{p} + \sum_{i} y_{i}^{p} + \sum_{i} (x_{i} + y_{i})^{p-1} y_{i} \]

\[ \leq \left( \sum_{i} x_{i}^{p} \right)^{1/p} \left( \sum_{i} y_{i}^{p} \right)^{1/q} \left( \sum_{i} (x_{i} + y_{i})^{p-1} y_{i} \right)^{1/q} \]

\[ = \left( \|x\|_{p} + \|y\|_{p} \right) \left( \sum_{i} (x_{i} + y_{i})^{p-1} \right)^{1/q} . \]

(Exercise 5.6). Thus far we have seen at least three proofs of the Cauchy–Schwarz inequality: by convex function, by Hölder’s inequality, and by (5.13).

Example 5.4. Let \( x_{1}, \ldots, x_{n} \) be positive numbers. If we replace \( x_{i} \) and \( y_{i} \) in the Cauchy–Schwarz inequality by \( \sqrt{x_{i}} \) and \( \sqrt{y_{i}} \), respectively, we obtain

\[ n \leq \left( \sum_{i=1}^{n} x_{i}^{1/2} \right) \left( \sum_{i=1}^{n} x_{i}^{-1/2} \right) . \]

which is equivalent to \( n \left( \sum_{i=1}^{n} x_{i}^{-1} \right)^{-1} \leq n^{-1} \sum_{i=1}^{n} x_{i} \). This is just the two ends of (5.5) implying that the harmonic mean is bounded by the arithmetic mean. Of course, (5.5) is a stronger result.