A Convergence Concepts

This chapter is about convergence of random variables. If $X_1, X_2, \ldots$ is a sequence of random variables, in what sense can we say that the sequence converges? And how do we use that information to do statistics?

A.1 Convergence in Distribution

Convergence in Distribution

If $X_1, X_2, \ldots$ is a sequence of random variables with $X_n$ having c. d. f. $F_n$, we say the sequence $X_n$ converges in distribution to a random variable $X$ having c. d. f. $F$ if

$$F_n(x) \to F(x), \quad \text{as } n \to \infty$$

for every real number $x$ that is a continuity point of $F$. We indicate $X_n$ converging in distribution to $X$ by writing

$$X_n \xrightarrow{D} X, \quad \text{as } n \to \infty.$$

“Continuity point” means a point $x$ such that $F$ is continuous at $x$ (a point where $F$ does not jump). If the limiting random variable $X$ is continuous, then every point is a continuity point. If $X$ is discrete or of mixed type, then $F_n(x) \to F(x)$ must hold at points $x$ where $F$ does not jump but it does not have to hold at the jumps.

From the definition it is clear that convergence in distribution is a statement about distributions not variables. Though we write $X_n \xrightarrow{D} X$, what this means is that the distribution of $X_n$ converges to the distribution of $X$. We could dispense with the notion of convergence in distribution and always write $F_{X_n}(x) \to F_X(x)$ for all continuity points $x$ of $F$ in place of $X_n \xrightarrow{D} X$, but that would be terribly cumbersome.

How does one establish that a sequence of random variables converges in distribution? By writing down the c. d. f.’s and showing that they converge? No. In the common applications of convergence in distribution in statistics, convergence in distribution is a consequence of the central limit theorem or the law of large numbers.
A.2 The Central Limit Theorem

**Central Limit Theorem (CLT)**

If $X_1, X_2, \ldots$ is a sequence of independent, identically distributed random variables having mean $\mu$ and variance $\sigma^2$, and

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

is the sample mean for sample size $n$, then

$$\frac{\overline{X}_n - \mu}{\sigma / \sqrt{n}} \xrightarrow{d} Z, \quad \text{as } n \to \infty, \quad (1)$$

where $Z \sim \mathcal{N}(0,1)$.

The only requirement is that the variance $\sigma^2$ exist (this implies that the mean $\mu$ also exists). No other property of the distribution of the $X_i$ matters.

The left hand side of (1) always has mean zero and variance one for all $n$, regardless of the distribution of the $X_i$ so long as the variance exists, because $\overline{X}_n$ always has mean $\mu$ and variance $\sigma^2 / n$ (box on p. 337 in the textbook), and standardizing always produces a variable with mean zero and variance one by linearity of expectation (p. 192 in the textbook). Thus the central limit theorem doesn’t say anything about means and variances, rather it says that the shape of the distribution of $\overline{X}_n$ approaches the bell-shaped curve of the normal distribution as $n \to \infty$.

A sloppy way of rephrasing (1) is

$$\overline{X}_n \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

for “large $n$.” Most of the time the sloppiness causes no harm and no one is confused. For an explanation of why this is sloppy, see Example A.2.4 below.

**Example A.2.1 Normal Distributions**

Suppose now $X_1, X_2, \ldots$ are i. i. d. $\mathcal{N}(\mu, \sigma^2)$. Then, since a linear combination of normals is normal (box on p. 234 in the textbook), the left hand side of (1) has distribution $\mathcal{N}(0,1)$ for all $n$.

This is the trivial case. We don’t need the central limit theorem to tell us that a constant sequence converges. If $F_n = F$ for all $n$, then trivially $F_n(x) \to F(x)$ for all $x$. Note that the sequence $X_1, X_2, \ldots$ is not constant. The $X_i$, being continuous random variables, are all different. But (just a repetition to drive the point home) the central limit is not a statement about convergence of random variables, but about convergence of distributions of random variables, and the distributions of the variables on the left hand side of (1) are constant, to wit $\mathcal{N}(0,1)$ for all $n$. 

2
Example A.2.2 Blah-blah Distributions

It’s hard to produce specific nontrivial examples, because almost everything is an example. Suppose $X_1, X_2, \ldots$ are i. i. d. blah-blah distributed (fill in your choice, the only requirement is finite variance). Then (1) holds.

Example A.2.3 Distributions with an Addition Rule

By addition rules for distributions we mean rules like

If $X_i \sim \begin{cases} \text{Bin}(n_i, p) \\ \text{Negbin}(r_i, p) \\ \text{Poi}(m_i) \\ \text{Gam}(\alpha_i, \lambda) \end{cases}$ then $X_1 + \cdots + X_k \sim \begin{cases} \text{Bin}(n_1 + \cdots + n_k, p) \\ \text{Negbin}(r_1 + \cdots + r_k, p) \\ \text{Poi}(m_1 + \cdots + m_k) \\ \text{Gam}(\alpha_1 + \cdots + \alpha_k, \lambda) \end{cases}$

This means that for large $k$ the distribution of the sum is approximately normal, that is

\begin{align*}
\text{Bin}(n, p) & \approx \mathcal{N}(np, npq), & \text{for large } n \\
\text{Negbin}(r, p) & \approx \mathcal{N} \left( \frac{r}{p}, \frac{r}{p^2} \right), & \text{for large } r \\
\text{Poi}(m) & \approx \mathcal{N}(m, m), & \text{for large } m \\
\text{Gam}(\alpha, \lambda) & \approx \mathcal{N} \left( \frac{\alpha}{\lambda}, \frac{\alpha}{\lambda^2} \right), & \text{for large } \alpha
\end{align*}

Of course, the normal distribution also has an “addition rule,” but we did it in the preceding example. Also, of course, since the chi-square distribution is a special case of the gamma distribution

\begin{equation}
\chi_k^2 \approx \mathcal{N}(k, 2k), \quad \text{for large } k
\end{equation}

where $\chi_k^2$ denotes a chi-square random variable with $k$ degrees of freedom.

These normal approximations to the binomial and the chi-square are described in the textbook (sections 4.5 and 6.4). The other two are not. But all four hold for the same reason, they are special cases of the CLT.

Example A.2.4 A Bogus Proof that Poisson is Normal

The Poisson distribution has a curious property called infinite divisibility. The parameter $m$ can be any positive real number, hence for any integer $n$ we can represent a Poisson random variable $X \sim \text{Poi}(m)$ as the sum $X = X_1 + \cdots + X_n$ where the $X_i$ are i. i. d. Poi$(m/n)$. As $n \to \infty$, $X$ converges to normal by the CLT. Not!

This isn’t quite the way convergence in distribution works. You can’t change the definition of the $X_i$ as $n$ increases. What’s the point? There is a tendency to state the CLT sloppily as “the sum of many i. i. d. random variables is approximately normal.” This example shows why that statement is sloppy. We can represent $X$ as the sum of $n$ i. i. d. random variables for $n$ as large as you please, but $X$ is always the same and not getting any closer to normally distributed.

Example A.2.5 A Symmetric Bimodal Distribution

Let us take a look at how the CLT works in practice. How large does $n$ have to be before the distribution of $X_n$ is approximately normal?
On the left is a severely bimodal distribution. The p. d. f. of $X$ is the sum of two normal densities with different means but the same standard deviations. The p. d. f. of $X_{10}$ on the right still shows the effects of the bimodality. It is plotted over the normal p. d. f. with the same mean and variance. The two curves are not very close. The CLT doesn’t provide a good approximation at $n = 10$.

At $n = 20$ and $n = 30$ we have much better results. The p. d. f. of $X_{30}$ is almost indistinguishable from the normal p. d. f. with the same mean and variance. There is a bit of wiggle at the top of the curve, but everywhere else the fit is terrific. It is this kind of behavior that leads to the rule of thumb propounded in elementary statistics texts that $n > 30$ is “large sample” territory, thirty is practically infinity.

**Example A.2.6 A Skewed Distribution**

The $30 = \infty$ rule does not hold for skewed distributions. Consider $X$ having the chi-square distribution with one degree of freedom.

The p. d. f. of $X$ is shown on the left. It is extremely skewed going to infinity at zero. On the right is the p. d. f. of $X_{30}$ and the normal p. d. f. with the same
mean and variance. The fit is not good. The p. d. f. of $\bar{X}_{30}$, a rescaled chi(30) density, is still rather skewed and so cannot be close to a normal density, which of course is symmetric.

The fit is better at $n = 100$ and $n = 300$, but still not as good as our bimodal example at $n = 30$. The moral of the story is that skewness slows convergence in the central limit theorem.

What happens to convergence in distribution when there is a change of variable? If $X_n \xrightarrow{D} X$, what happens to $g(X_n)$?

**Continuous Mapping Theorem**

If $X_n \xrightarrow{D} X$ and $g$ is a continuous function, then $g(X_n) \xrightarrow{D} g(X)$.

The CLT is often written in slightly different forms. Moving the $\sqrt{n}$ from the denominator of the denominator to the numerator gives

$$\sqrt{n} \frac{X_n - \mu}{\sigma} \xrightarrow{D} Z$$

(3)

Since $g(x) = \sigma x$ is a continuous function of $x$, the continuous mapping theorem says that $X_n \xrightarrow{D} X$ implies $\sigma X_n \xrightarrow{D} \sigma X$. In other words, we can multiply both sides of (3) by $\sigma$ getting

$$\sqrt{n} (X_n - \mu) \xrightarrow{D} \sigma Z$$

(4)

In both (3) and (4) $Z$ is still standard normal. On the right hand side of (4), however, we now have a normal random variable $Y = \sigma Z$ with mean zero and variance $\sigma^2$ so we can also write (4) as

$$\sqrt{n} (X_n - \mu) \xrightarrow{D} Y$$

(5)

where $Y$ is $N(0, \sigma^2)$.

It simplifies notation if we are allowed to write a distribution on the right hand side of a statement about convergence in distribution. For example, we can simplify (5) and the rest of the sentence following it to

$$\sqrt{n} (X_n - \mu) \xrightarrow{D} N(0, \sigma^2)$$
A.3 Convergence in Probability

A special case of convergence in distribution is convergence in distribution to a degenerate random variable concentrated at one point, \( X_n \xrightarrow{D} a \) where \( a \) is a constant. This notion is so important that it has a special name.

**Convergence in Probability**

If \( X_1, X_2, \ldots \) is a sequence of random variables and \( a \) is a constant, we say the sequence \( X_n \) **converges in probability** to \( a \) if for every \( \epsilon > 0 \)

\[
P(|X_n - a| > \epsilon) \to 0, \quad \text{as } n \to \infty.
\]

We indicate \( X_n \) converging in probability to \( a \) by writing

\[
X_n \xrightarrow{P} a, \quad \text{as } n \to \infty.
\]

Convergence in probability to a constant and convergence in distribution to a constant are the same thing, so we could write \( X_n \xrightarrow{D} a \) instead of \( X_n \xrightarrow{P} a \), but the latter is traditional. There is also a more general notion of convergence in probability to a random variable, but it has no application in statistics and we shall ignore it.

Why are convergence in probability to a constant and convergence in distribution to a constant the same thing? Another way to write convergence in probability to \( a \) is, for every \( \epsilon > 0 \),

\[
P(X_n \leq a - \epsilon) + P(X_n \geq a + \epsilon) \to 0, \quad \text{as } n \to \infty
\]

since both terms on the left are nonnegative, this occurs if and only if both go to zero, that is, for every \( \epsilon > 0 \),

\[
P(X_n \leq a - \epsilon) \to 0, \quad \text{as } n \to \infty
\]

\[
P(X_n \geq a + \epsilon) \to 0, \quad \text{as } n \to \infty
\]

Now we translate this into the language of c. d. f.'s. As usual, let \( F_n \) be the c. d. f. of \( X_n \). Then (6) becomes

\[
F_n(a - \epsilon) \to 0, \quad \text{as } n \to \infty
\]

Translating (7) is a bit trickier because of the possibility of discontinuities in \( F_n \). If \( F_n \) has a jump at \( a + \epsilon \) then

\[
P(X_n \geq a + \epsilon) = 1 - F_n^-(a + \epsilon)
\]

where \( F_n^- \) denotes the left continuous analog of the c. d. f.

\[
F_n^-(x) = \lim_{y \downarrow x} F_n(x)
\]
With this notation we can translate (7) to

$$1 - F_n^{-1}(a + \epsilon) \to 0, \quad \text{as } n \to \infty$$

or

$$F_n^{-1}(a + \epsilon) \to 1, \quad \text{as } n \to \infty \quad (9)$$

The $F_n^{-1}$ here is a nuisance, what we want is the same with $F_n$

$$F_n(a + \epsilon) \to 1, \quad \text{as } n \to \infty \quad (10)$$

(9) and (10) are not equivalent considered for one particular $\epsilon$, but since

$$F_n^{-1}(a + \epsilon_1) \leq F_n(a + \epsilon_1) \leq F_n^{-1}(a + \epsilon_2) \leq F_n(a + \epsilon_2), \quad \text{if } \epsilon_1 \leq \epsilon_2$$

(9) holds for all $\epsilon > 0$ if and only if (10) also holds for all $\epsilon > 0$.

Thus we have established that $X_n \xrightarrow{P} a$ if and only if

$$F_n(x) \xrightarrow{n \to \infty} \begin{cases} 0, & x < a \\ 1, & x > a \end{cases}$$

(11)

If $X$ is a random variable concentrated at $a$, its c. d. f. is

$$F(x) = \begin{cases} 0, & x < a \\ 1, & x \geq a \end{cases}$$

So (11) says $F_n(x) \to F(x)$ except possibly for $x = a$ which is where $F$ jumps, and this is the same as saying $X_n \xrightarrow{P} X$.

Because convergence in probability to a constant is the same as convergence in distribution to a constant, the continuous mapping theorem also applies to it. In this case, the function only needs to be continuous at one point.

**Continuous Mapping Theorem** (for Convergence in Probability)

If $X_n \xrightarrow{P} a$ and $g$ is a function continuous at $a$, then $g(X_n) \xrightarrow{P} g(a)$.

### A.4 The Law of Large Numbers

One place convergence in probability appears is in the law of large numbers. The general form of the law, mentioned on p. 379 where it is called *Kintchine’s theorem*, is
Law of Large Numbers (LLN)

If \(X_1, X_2, \ldots\) is a sequence of independent, identically distributed random variables having mean \(\mu\), and

\[
\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i
\]

is the sample mean for sample size \(n\), then

\[
\overline{X}_n \xrightarrow{P} \mu, \quad \text{as } n \to \infty. \tag{12}
\]

The only requirement is that the mean \(\mu\) exist. No other property of the distribution of the \(X_i\) matters.

The law of large numbers stated here is known technically as the weak law of large numbers. There is another version known as the strong law of large numbers that says under the same condition (existence of the mean) that a stronger form of convergence than convergence in probability actually obtains. This stronger form of convergence is called almost sure convergence. The distinction between convergence in probability and almost sure convergence is little used in statistics, and we shall ignore it.

Example A.4.1 Blah-blah Distributions

Again almost everything is an example. Suppose \(X_1, X_2, \ldots\) are i. i. d. blah-blah distributed (fill in your choice, now the only requirement is finite mean). Then (12) holds.

Example A.4.2 The Cauchy Distribution

The textbook gave the Cauchy distribution as the canonical example of a distribution for which the mean does not exist. Here we give it as the canonical example of failure of the law of large numbers. The Cauchy distribution also has an addition rule: if \(X_1, \ldots, X_n\) are i. i. d. Cauchy, then \(\overline{X}_n\) is Cauchy. (This is easily proved using characteristic functions, but that goes beyond the tools developed in this course). Since \(\overline{X}_n\) has the same distribution for all \(n\), this is another trivial example of convergence in distribution: \(\overline{X}_n \xrightarrow{D} X\), where \(X\) is Cauchy. So \(\overline{X}_n\) doesn’t converge in probability to a constant, and the law of large numbers does not hold for the Cauchy distribution.
A.5 Slutsky’s Theorem

**Slutsky’s Theorem**

If $X_n \xrightarrow{D} X$ and $Y_n \xrightarrow{P} a$, then as $n \to \infty$

\[
X_n + Y_n \xrightarrow{D} X + a, \\
Y_n X_n \xrightarrow{D} aX,
\]

and if $a \neq 0$

\[
X_n/Y_n \xrightarrow{D} X/a.
\]

In other words, we have all the nice properties we expect of limits, the limit of a sum is the sum of the limits, and so forth. The point of the theorem is this is not true unless one of the limits is a constant. If we only had $X_n \xrightarrow{D} X$ and $Y_n \xrightarrow{P} Y$, we couldn’t say anything about the limit of $X_n + Y_n$ without knowing about the joint distribution of $X_n$ and $Y_n$. When $Y_n$ converges to a constant, Slutsky’s theorem tells us that we don’t need to know anything about joint distributions.

A special case of Slutsky’s theorem involves two sequences converging in probability. If $X_n \xrightarrow{P} a$ and $Y_n \xrightarrow{P} b$, then $X_n + Y_n \xrightarrow{P} a + b$, and so forth. This is a special case of Slutsky’s theorem because convergence in probability to a constant is the same as convergence in distribution to a constant.

A.6 Comparison of the LLN and the CLT

When $X_1, X_2, \ldots$ is an i.i.d. sequence of random variables having a variance, both the law of large numbers and the central limit theorem apply. In this case the CLT tells us much more than the LLN. The LLN says $\overline{X}_n \xrightarrow{P} \mu$ or, applying the continuous mapping theorem to subtract $\mu$ from both sizes, $\overline{X}_n - \mu \xrightarrow{P} 0$. If we think of $\overline{X}_n$ as an estimator of $\mu$, then $\overline{X}_n - \mu$ is the estimation error, and the LLN says this error converges in probability to zero.

The CLT tells us how fast this error converges to zero. It says $\sqrt{n}(\overline{X}_n - \mu)$ converges in distribution to a random variable $Y$, specifically $Y \sim N(0, \sigma^2)$ but that doesn’t matter for the point we are trying to make here.

Speaking a bit sloppily, this says that the estimation error goes to zero like $1/\sqrt{n}$. A more precise statement is that if we multiply by anything that goes to infinity slower than $\sqrt{n}$, say $a_n = n^{0.49}$,

\[a_n(\overline{X}_n - \mu) \xrightarrow{P} 0, \quad \text{as } n \to \infty\]

This follows from Slutsky’s theorem, $a_n/\sqrt{n} \to 0$ so

\[a_n(\overline{X}_n - \mu) = \frac{a_n}{\sqrt{n}} \sqrt{n}(\overline{X}_n - \mu) \xrightarrow{D} 0 \cdot Y = 0\]
So why do we even care about the law of large numbers? Is it because there are lots of important probability models having a mean but no variance (so the LLN holds but the CLT does not)? No, not any used for real data. The point is that sometimes we don’t care about the information obtained from the central limit theorem. When the only fact we want to use is \( \overline{X}_n \overset{p}{\to} \mu \), we refer to the law of large numbers as our authority. Its statement is simpler, and there is no point in dragging an unnecessary assumption about variance when it’s not needed.

### A.7 Consistency

Using the notion of convergence in probability we can restate the definition of consistency given in the textbook on p. 378 as

**Consistency**

A sequence \( T_1, T_2, \ldots \) of estimators is **consistent** as an estimate of \( \theta \) if

\[
T_n \overset{p}{\to} \theta, \quad \text{as } n \to \infty.
\]

**Example A.7.1 Exponential Distributions**

Example 9.10f in the textbook says that the maximum likelihood estimator of \( \lambda \) given an i. i. d. sample of size \( n \) from an \( \text{Exp}(\lambda) \) distribution is \( \hat{\lambda}_n = 1/\overline{X}_n \).

The law of large numbers tells us that

\[
\overline{X}_n \overset{p}{\to} \mu = \frac{1}{\lambda}
\]

Since \( g(x) = 1/x \) is a continuous function, the continuous mapping theorem tells us that

\[
\hat{\lambda}_n = \frac{1}{\overline{X}_n} \overset{p}{\to} \lambda
\]

or in other words, that \( \hat{\lambda}_n \) is a consistent estimator of \( \lambda \).

**Example A.7.2 Consistent Estimators of \( \sigma^2 \) and \( \sigma \)**

Here we carefully establish the consistency of the popular estimators of \( \sigma \) and \( \sigma^2 \). First consider

\[
V_n = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2
\]

defined on p. 289 in the textbook. Note that \( \overline{X}_n \) is the mean and \( V_n \) is the variance of the empirical distribution that puts probability \( 1/n \) at each of the \( n \) data points \( X_i \). The parallel axis theorem (p. 189 in the textbook) applies to the empirical distribution just like any other distribution and gives

\[
V_n = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2 - (\overline{X}_n - \mu)^2
\]
The first term on the right hand side is a sample mean $Y_n$ of random variables

$$Y_i = (X_i - \mu)^2$$

The $Y_i$ are i. i. d. random variables with mean

$$\mu_Y = E(Y_i) = E\left\{ (X_i - \mu)^2 \right\} = \sigma^2$$

The law of large numbers applied to $Y_n$ tells us that $Y_n \xrightarrow{P} \sigma^2$. That takes care of the first term on the right hand side of (14).

The law of large numbers applied to $X_n$ tells us that $X_n \xrightarrow{P} 0$, hence the continuous mapping theorem tells us that $(X_n - \mu)^2 \xrightarrow{P} 0$, because $g(u) = u^2$ is a continuous function. That takes care of the second term on the right hand side of (14).

By Slutsky’s theorem, the limit of a sum is the sum of the limits, so putting all this together, we have

$$V_n = Y_n - (X_n - \mu)^2 \xrightarrow{P} \sigma^2 - 0 = \sigma^2$$

The continuous mapping theorem now tells us that $\sqrt{V_n} \xrightarrow{P} \sigma$, because $g(u) = \sqrt{u}$ is a continuous function.

And how about our other popular estimator of $\sigma^2$

$$S_n^2 = \frac{n}{n - 1} V_n \quad (15)$$

defined on p. 290 in the textbook? Since $n/(n - 1) \to 1$ and $V_n \xrightarrow{P} \sigma^2$, Slutsky’s theorem tells us that $S_n^2 \xrightarrow{P} \sigma^2$ as well. Then the continuous mapping theorem tells us that $S_n \xrightarrow{P} \sigma$. In summary

If $X_1, X_2, \ldots$ is a sequence of independent, identically distributed random variables having variance $\sigma^2$, and $V_n$ and $S_n^2$ are defined by (13) and (15), then $V_n$ and $S_n^2$ are consistent estimators of $\sigma^2$ and $\sqrt{V_n}$ and $S_n$ are consistent estimators of $\sigma$.

Problems

A-1. If $X_1, X_2, \ldots$ is a sequence of independent, identically distributed random variables having mean $\mu$ and variance $\sigma^2$, what does $n (X_n - \mu)^2$ converge to?

A-2. Use the central limit theorem to find a normal approximation to $P(X > 25)$ when $X \sim \text{Gam}(100, 5)$.
A-3. Suppose $X_1, X_2, \ldots$ is a sequence of i.i.d. random variables having the Gam$(\alpha, \lambda)$ distribution defined in the box on p. 244 in the textbook. On p. 399 in the textbook, the estimators
\[
\hat{\alpha}_n = \frac{\sum^d X_i}{V_n} \quad \text{and} \quad \hat{\lambda}_n = \frac{\sum^d X_i}{V_n}
\]
are proposed as estimators of $\alpha$ and $\lambda$, respectively. Show that $\hat{\alpha}_n$ is a consistent estimator of $\alpha$ and $\hat{\lambda}_n$ is a consistent estimator of $\lambda$. [Use Example A.7.2, the law of large numbers, the continuous mapping theorem, and Slutsky’s theorem. Don’t try to do everything from scratch.]

A-4. Suppose $X_n \sim \text{Bin}(n, p)$, and $\hat{p}_n = X_n/n$, the sample proportion, is considered as an estimator of $p$.

(a) Use the law of large numbers to show that $\hat{p}_n$ is a consistent estimator of $p$.

(b) The logit function is defined by
\[
\logit(p) = \begin{cases} 
 \log \left( \frac{p}{1-p} \right), & 0 < p < 1 \\
 +\infty, & p = 1 \\
 -\infty, & p = 0 
\end{cases}
\]
If $\hat{\theta}_n = \logit(\hat{p}_n)$ and $\theta = \logit(p)$, show that $\hat{\theta}_n$ is a consistent estimator of $\theta$.

A.8 Large-Sample Confidence Intervals

The results of the preceding sections allow us to derive large-sample confidence intervals (Sec 9.4 in the textbook) without handwaving. The central limit theorem (1) tells us that
\[
\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{D} N(0, 1)
\]
from which we get
\[
P\left(-k < \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} < k\right) \rightarrow \Phi(k) - \Phi(-k)
\]
so if $k$ is chosen so that (for example) $\Phi(k) - \Phi(-k) = .95$, that is, $k = 1.96$, then
\[
P\left(-k < \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} < k\right) = P\left(\bar{X}_n - k\frac{\sigma}{\sqrt{n}} < \mu < \bar{X}_n + k\frac{\sigma}{\sqrt{n}}\right) \rightarrow .95
\]
so $\bar{X}_n \pm k\sigma/\sqrt{n}$ is a large-sample 95% confidence interval for $\mu$. 

12
Great! Only one problem: we almost never know \( \sigma \), so this is useless. The book tells us to substitute a consistent estimator \( S_n \) for \( \sigma \) and proceed as usual. Why is that valid? We know from Section A.7 that \( S_n \xrightarrow{P} \sigma \). The continuous mapping theorem tells us we can divide both sides by \( \sigma \) giving \( S_n/\sigma \xrightarrow{P} 1 \). Now applying Slutsky’s theorem gives

\[
\frac{\overline{X}_n - \mu}{S_n / \sqrt{n}} = \frac{\sigma}{S_n} \cdot \frac{\overline{X}_n - \mu}{\sigma / \sqrt{n}} \xrightarrow{d} 1 \cdot Z = Z
\]

where \( Z \) is standard normal.

### Large-Sample Confidence Intervals for Means

If \( X_1, X_2, \ldots \) is a sequence of independent, identically distributed random variables having mean \( \mu \) and variance \( \sigma^2 \), and \( S_n \) is any consistent estimator of \( \sigma \) then

\[
\frac{\overline{X}_n - \mu}{S_n / \sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1)
\]

Hence

\[
\overline{X}_n \pm k \frac{S_n}{\sqrt{n}}
\]

is a 100(1 - \( \alpha \))% confidence interval for \( \mu \) if \( \Phi(-k) = 1 - \Phi(k) = \frac{\alpha}{2} \).

### Example A.8.1 The Binomial Distribution

The box on p. 383 in the textbook giving the large-sample confidence interval for a population proportion is a special case of the result given above if \( \sqrt{\hat{p}_n(1 - \hat{p}_n)} \) is a consistent estimator of \( \sqrt{p(1 - p)} \). Why is that? The law of large numbers implies \( \hat{p}_n \xrightarrow{P} p \) (Problem 1.2(a) is to prove this). Then, since \( g(p) = \sqrt{p(1 - p)} \) is a continuous function, the continuous mapping theorem implies \( \sqrt{\hat{p}_n(1 - \hat{p}_n)} \xrightarrow{P} \sqrt{p(1 - p)} \).

### Problems

A-5. Suppose 100 widget failure times have been observed, and the failure times are assumed to be i. i. d. exponentially distributed. Suppose \( \overline{X} \) is 543 days. Give an approximate large-sample 95% confidence interval for the population mean failure time \( \mu \).

A.9 The Delta Method

The subject of this section, with its particularly unenlightening name is what used to be called “propagation of errors.” If we know the large sample behavior of a statistic \( X \), what can we say about the behavior of \( Y = g(X) \)? What happens to large-sample behavior when we transform the variable?
If $T_n$ is a consistent estimator of $\theta$ (that is, $T_n \xrightarrow{P} \theta$) and $g$ is a continuous function, then the continuous mapping theorem tells us that $g(T_n) \xrightarrow{P} g(\theta)$, or in other words that $g(T_n)$ is a consistent estimator of $g(\theta)$. An example of this was Example A.7.1 above, where we saw that $\hat{\lambda}_n = 1/X_n$ was a consistent estimator of the parameter $\lambda$ of the exponential distribution.

But this is not enough to construct confidence intervals. For that we need something like the central limit theorem.

## The Delta Method

Suppose $T_n$ is an asymptotically normal estimator of $\theta$, that is

$$\sqrt{n}(T_n - \theta) \xrightarrow{D} Y$$

(16)

where $Y \sim N(0, \sigma^2)$, and suppose $g$ is differentiable at $\theta$, then

$$\sqrt{n}[g(T_n) - g(\theta)] \xrightarrow{D} g'(\theta)Y$$

(17)

The sloppy way of saying this is

$$g(T_n) \approx N \left( g(\theta), \frac{g'(\theta)^2 \sigma^2}{n} \right)$$

The “delta” in the name of the method is supposed to remind you of derivatives, $\Delta y/\Delta x$ and all that.

Note please, that the delta method does not produce a statement about asymptotic normality out of nowhere. It turns one statement about asymptotic normality (16) into another statement about asymptotic normality (17). In order to use the delta method, you have to start with a statement about asymptotic normality. Typically we get that from the central limit theorem; compare (16) with (5).

Why does the delta method work? By definition, $g$ has derivative $g'(\theta)$ at the point $\theta$ if

$$\frac{g(\theta + h) - g(\theta)}{h} \rightarrow g'(\theta), \quad \text{as } h \rightarrow 0$$

Another way to say this is that the function

$$o(h) = \frac{g(\theta + h) - g(\theta) - g'(\theta)}{h}$$

(18)

go to zero as $h$ goes to zero, which is the same as saying that $o(h)$ is continuous at zero. Thus by the continuous mapping theorem

$$o(T_n - \theta) \xrightarrow{P} 0, \quad \text{as } n \rightarrow \infty$$

(19)

because $T_n - \theta \xrightarrow{P} 0$. We can rewrite (18) as

$$g(t) - g(\theta) = [g'(\theta) + o(t-\theta)](t-\theta)$$
\[
\sqrt{n}[g(T_n) - g(\theta)] = [g'(\theta) + o(T_n - \theta)] \sqrt{n}(T_n - \theta) = W_n Y_n
\]

where
\[
W_n = g'(\theta) + o(T_n - \theta) \\
Y_n = \sqrt{n}(T_n - \theta)
\]

Then by (19) \( W_n \xrightarrow{P} g'(\theta) \) and by assumption \( Y_n \xrightarrow{D} Y \), so by Slutsky’s theorem \( W_n Y_n \xrightarrow{D} g'(\theta) Y \), which is what the delta method asserts.

**Example A.9.1 Exponential Failure Rate**

Continuing what was begun in Example A.7.1, let us consider the asymptotic distribution of \( \tilde{\lambda}_n \) for the exponential distribution. The central limit theorem tells us that
\[
\sqrt{n} \left( \overline{X}_n - \mu \right) \xrightarrow{D} Y
\]
with \( Y \sim \mathcal{N}(0, \sigma^2) \) and \( \sigma^2 = 1/\lambda \). By definition
\[
\tilde{\lambda}_n = \frac{1}{\overline{X}_n} = g(X_n)
\]

where \( g(x) = 1/x \). Then \( g'(x) = -1/x^2 \). So the delta method says that
\[
\sqrt{n} \left( \tilde{\lambda}_n - \lambda \right) = \sqrt{n} \left[ g(\overline{X}_n) - g(\mu) \right] \xrightarrow{D} g'(\theta) Y = -\frac{1}{\mu^2} Y = -\lambda^2 Y
\]

or that \( \tilde{\lambda}_n \) is asymptotically normal with mean \( \lambda \) and variance
\[
\frac{1}{n} (-\lambda^2)^2 \sigma^2 = \frac{1}{n} \lambda^4 \left( \frac{1}{\lambda} \right)^2 = \frac{\lambda^2}{n}
\]

**Example A.9.2 Probability of No Failures**

Continuing the preceding example, suppose we are interested in the probability of an item not failing in a time interval of length \( t \), which is \( P(X > t) = e^{-\lambda t} \). Write \( g(x) = e^{-x} \), then \( g'(x) = -te^{-x} \). So starting with the asymptotic normality derived in the preceding example
\[
\sqrt{n} \left( \tilde{\lambda}_n - \lambda \right) \xrightarrow{D} Y
\]

where \( Y \sim \mathcal{N}(0, \lambda^2) \) and applying the delta method one more time, we get
\[
\sqrt{n} \left[ g(\tilde{\lambda}_n) - g(\lambda) \right] \xrightarrow{D} g'(\theta) Y = -te^{-\lambda t} Y
\]

or that \( e^{-\tilde{\lambda}_n t} \) is asymptotically normal with mean \( e^{-\lambda t} \) and variance
\[
\left[ te^{-\lambda t} \right]^2 \frac{\lambda^2}{n} = \frac{\lambda^2}{n} \left( e^{-2\lambda t} - e^{-\lambda t} \right)
\]
Example A.9.3 Square-Root Transformation of Chi-Square
On p. 246 the textbook recommends using the normal approximation for $\sqrt{2\chi^2}$ rather than for $\chi^2$ itself. Let us apply the delta method to figure out why this has the stated distribution. In Example A.2.3 we found that $\chi_n^2 \approx N(n, 2n)$. In order to apply the delta method we must rewrite this as a statement about a sample mean rather than a sum. By definition a chi-square random variable with $n$ degrees of freedom is

$$\chi_n^2 = Z_1^2 + \cdots + Z_n^2$$

where the $Z_i$ are i.i.d. standard normal. Writing $Y_i = Z_i^2$, we have

$$\chi_n^2 = \sum_{i=1}^n Y_i = n \bar{Y}$$

Then the CLT says

$$\sqrt{n} (\bar{Y} - 1) \overset{D}{\to} W$$

where $W \sim N(0, 2)$. Applying the delta method to the transformation $g(x) = \sqrt{2x}$, which has derivative $g'(x) = 1/\sqrt{2x}$, we get

$$\sqrt{n} (g(\bar{Y}) - g(1)) = \sqrt{2n\bar{Y} - \sqrt{2n} \overset{D}{\to} \bar{Y} - 1)W = \frac{1}{\sqrt{2}} W$$

The right hand side is normal with mean zero and variance $(\frac{1}{\sqrt{2}})^2 2 = 1$. So

$$\sqrt{2\chi^2} - \sqrt{2n} \approx N(0, 1)$$

which is almost what the textbook says on p. 246. The only difference is that the textbook has $\sqrt{2n - 1}$ where we have $\sqrt{2n}$. This difference is negligible for large $n$ and cannot be derived from the delta method.

In the same way as in confidence intervals for means, an application of Slutsky’s theorem is required to get useful confidence intervals. We are allowed to plug in consistent estimators of parameters when estimating the asymptotic variance of $g(T_n)$, appealing to Slutsky’s theorem for justification.

Example A.9.4 The Geometric Distribution
Suppose $X_1, X_2$ is a sequence of i.i.d. Geo($p$) random variables. From the box on p. 131 in the textbook, $E(X_i) = \mu = 1/p$, so the method of moments estimator of $p$ is $\hat{p}_n = 1/\bar{X}_n$. What is the asymptotic distribution of this estimator?

Since $\sigma^2 = 1/p^2$ (same box in the textbook), the central limit theorem says

$$\sqrt{n} (\bar{X}_n - \mu) \overset{D}{\to} Y$$

with $Y \sim N(0, \sigma^2)$. We have $\hat{p}_n = g(\bar{X}_n)$ with $g(x) = 1/x$ so $g'(x) = -1/x^2$. Thus the delta method says

$$\sqrt{n} (\hat{p}_n - p) \overset{D}{\to} g'(\mu)Y = -\frac{1}{\mu^2} Y = -p^2 Y$$
and the variance of the right hand side is

\[ (p^2)^2 \sigma^2 = \frac{p^4q}{p^2} = p^2q \]

This result looks a little strange. Why \( p^2q \) when we get \( pq \) for the binomial? To make the anomaly clear, the sum of the \( X_i \) is negative binomial. The binomial and negative binomial distributions have the same likelihood. For the binomial, when we observe \( x \) successes in \( n \) trials the likelihood is

\[ L(p) = \binom{n}{x} p^x q^{n-x} \]

and the factor \( \binom{n}{x} \) can be thrown away, since it doesn’t contain the parameter \( p \). For the negative binomial, when we take \( k \) trials to observe \( r \) successes, the likelihood is

\[ L(p) = \binom{k-1}{r-1} p^r q^{k-r} \]

and the factor \( \binom{k-1}{r-1} \) can be thrown away, since it doesn’t contain the parameter \( p \). In the binomial the number of trials \( n \) is fixed, and the number \( x \) of success is random, in the negative binomial the number of successes \( r \) is fixed and the number of trials \( k \) is random, but when the number of successes and the number of trials are the same (that is \( x = r \) and \( k = n \)), both have the same likelihood. When both likelihoods are the same, the likelihood principle requires that all inferences be the same. Are they?

The point estimators of \( p \) are the same. In the binomial case \( \hat{p}_n = x/n \) and in the negative binomial case \( \hat{p}_r = r/k \), but these are the same number when \( x = r \) and \( k = n \). Note that we have changed notation from the first part of this example. The number of trials, now \( k \), was there \( \sum_i X_i \), and the number of successes, now \( r \), was there \( n \), since that is the number of geometric random variables observed (waiting time until a success). Hence what was \( X_n \) in first part is now \( k/r \) and \( \hat{p} = 1/X_n = r/k \).

Confidence intervals for \( p \) should also be the same, so we should get the same standard errors from the delta method. For the binomial, the delta method gives

\[ \sqrt{n} (\hat{p}_n - p) \xrightarrow{D} N(0, pq) \]

For the negative binomial, the delta method gives

\[ \sqrt{r} (\hat{p}_r - p) \xrightarrow{D} N(0, p^2q) \]

So, at first sight we have an anomaly. We are using the same estimator, but we get different expressions \( pq/n \) and \( p^2q/r \) for the asymptotic variances. The resolution of the anomaly comes when we estimate these quantities. For the binomial

\[ \frac{\hat{p} \hat{q}}{n} = \frac{x(n - x)}{n^3} \]
For the negative binomial
\[ \frac{\hat{p}^2 \hat{q}}{r} = \frac{r(k - r)}{k^3} \]
and these are the same when \( x = r \) and \( k = n \).

Before we did this analysis, we had no idea whether inverse sampling made a difference. Now we see that it does not (at least for large sample sizes). This is a special property of maximum likelihood estimators (\( \hat{p} \) is also the MLE in both cases). In general, inverse sampling does make a difference.

Problems

A-6. Continuing the widget failure time problem (A-5), where 100 widget failures were observed with a mean failure time of 543 days. Give approximate large-sample 95\% confidence intervals

(a) for the population failure rate \( \lambda \), and

(b) for the probability of no failure in 365 days.

A-7. Continuing Problem A-4 about \( \theta = \logit(p) \) for the binomial distribution, find the asymptotic distribution of \( \hat{\theta}_n = \logit \hat{p}_n \).

A-8. Suppose radioactive decay forms a Poisson process and for a particular sample of a chemical isotope containing \( 10^20 \) atoms we observe 3568 counts in one minute.

(a) Give an approximate large-sample 95\% confidence interval for the rate parameter \( \lambda \) in counts per minute.

(b) The half-life of the isotope is

\[ T_{1/2} = \frac{N \ln 2}{\lambda} \]

where \( N \) is the number of atoms in the sample. Give an approximate large-sample 95\% confidence interval for \( T_{1/2} \).

B Asymptotics of Maximum Likelihood

B.1 Fisher Information

The asymptotics of maximum likelihood are intimately connected with expectations of derivatives of the log likelihood. So we start with this subject. Consider first a data set consisting of a single observation \( X \) from a distribution with density function \( f(x|\theta) \). The log likelihood is the function of \( \theta \)

\[ l(\theta) = \log f(X|\theta) \]
Despite the notation, the log likelihood $l(\theta)$ is also a function of $X$. Hence it is a random variable, and we can take expectations. Surprisingly we can learn a lot about the log likelihood and its derivatives by differentiating the identity

$$\int f(x|\theta) \, dx = 1$$

(20)

We assume throughout that it is permissible to take derivatives under the integral sign (this is usually, but not always valid). Differentiating (20) once, we get

$$\int f'(x|\theta) \, dx = 0$$

In this section a prime always denotes differentiation with respect to $\theta$ (not $x$). We want to rewrite this so that it is an expectation, an expression of the form

$$E_\theta \{ g(X, \theta) \} = \int g(x, \theta) f(x|\theta) \, dx$$

The subscript $\theta$ on the expectation operator reminds us that both $\theta$'s are the same, the one in $g(X, \theta)$ and the one in $f(x|\theta)$.

If we multiply and divide by $f(x|\theta)$ under the integral sign, we get

$$\int \frac{f'(x|\theta)}{f(x|\theta)} f(x|\theta) \, dx = E_\theta \left\{ \frac{f'(X|\theta)}{f(X|\theta)} \right\} = 0$$

Now we recognize the random variable in the expectation as the derivative of the log likelihood

$$l'(\theta) = \frac{\partial}{\partial \theta} \log f(X|\theta) = \frac{1}{f(X|\theta)} \frac{\partial}{\partial \theta} f(X|\theta) = \frac{f'(x|\theta)}{f(x|\theta)}$$

(21)

The first derivative of the log likelihood is called the score. What we have just established is that its expectation is zero,

$$E_\theta \{ l'(\theta) \} = 0$$

(22)

and hence its second moment is its variance

$$\text{var}_\theta \{ l'(\theta) \} = E_\theta \{ l'(\theta)^2 \}$$

(23)

Now we want to look at second derivatives. The notation becomes cumbersome if we don’t simplify it, so from here on we write just $f$, $f'$, and $f''$ instead of $f(x|\theta)$ and so forth. We also omit the $dx$ in integrals. Then differentiating (20) twice gives

$$\int f'' \, dx = 0$$

or multiplying and dividing by $f$ as before

$$\int \frac{f''}{f} f = E_\theta \left\{ \frac{f''}{f} \right\} = 0$$

(24)
Now we need to figure out what relationship this has to the second derivative of the log likelihood

\[ l'' = (\log f)' = \left( \frac{f'}{f} \right)' = \frac{f''}{f} - \frac{f' f'}{f^2} = \frac{f''}{f} - \left( \frac{f'}{f} \right)^2 \]

Taking expectations gives

\[ E_\theta \{ l''(\theta) \} = E_\theta \left\{ \frac{f''}{f} \right\} - E_\theta \left\{ \left( \frac{f'}{f} \right)^2 \right\} \]

From (24) we see that the first term on the right hand side is zero and from (21) and (23) we see that the second term on the right hand side is the variance of the score. Thus

\[ E_\theta \{ l''(\theta) \} = -\text{var}_\theta \{ l'(\theta) \} \quad (25) \]

### Fisher Information

The **Fisher information** at \( \theta \) for a sample of size one from \( f(x|\theta) \) is

\[ I_1(\theta) = \text{var}_\theta \{ l'(\theta) \} = E_\theta \left\{ \left( \frac{\partial}{\partial \theta} \log f(X|\theta) \right)^2 \right\} \quad (26) \]

If differentiation under the integral sign is permissible, then the Fisher information is also

\[ I_1(\theta) = -E_\theta \{ l''(\theta) \} = -E_\theta \left\{ \frac{\partial^2}{\partial \theta^2} \log f(X|\theta) \right\} \quad (27) \]

The equivalence of the two definitions is just (25). This concept is named after the English statistician R. A. Fisher (1890–1962) who invented maximum likelihood and stated its properties of consistency, asymptotic normality, and efficiency and first called this quantity “information.”

The log likelihood for an i. i. d. sample of size \( n \) is

\[ l_n(\theta) = \log f(X|\theta) = \log \left( \prod_{i=1}^{n} f(X_i|\theta) \right) = \sum_{i=1}^{n} \log f(X_i|\theta) \]

The Fisher information for an \( n \) sample is

\[ I_n(\theta) = \text{var}_\theta \{ l'_n(\theta) \} \]

\[ = \text{var}_\theta \left\{ \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(X_i|\theta) \right\} \]

\[ = \sum_{i=1}^{n} \text{var}_\theta \left\{ \frac{\partial}{\partial \theta} \log f(X_i|\theta) \right\} \]

\[ = n I_1(\theta) \]
The variance of the sum is the sum of the variances because the $X_i$ are independent. Each term in the sum is the same because the $X_i$ are identically distributed.

**Example B.1.1 Exponential Distributions**

For the Exp($\lambda$) family of distributions, the log likelihood is

$$l(\lambda) = \log f(x|\lambda) = \log (\lambda e^{-\lambda x}) = -\lambda x + \log \lambda$$

Hence the score is

$$l'(\lambda) = -x + \frac{1}{\lambda}$$

and the second derivative is

$$l''(\lambda) = -\frac{1}{\lambda^2}$$

Now let us check the identities (22) and (25) obtained by differentiating under the integral sign. These identities say that

- The expectation of (28) is zero. This checks because $E(X) = 1/\lambda$.
- The variance of (28) is minus the expectation of (29). The variance of (28) is the variance of $X$, which is $1/\lambda^2$. Since (29) does not contain $x$, it is a constant and thus its own expectation. So this checks too.

The Fisher information depends on the parameter used. The same family of distributions with a different parameterization will have different Fisher information.

**Example B.1.2 Exponential Distributions Again**

Continuing what was begun in Example B.1.1, let us consider the same Exp($\lambda$) family of distributions but take the parameter to be $\mu = 1/\lambda$, the mean of $X$. Now the log likelihood is

$$l(\mu) = -x \frac{1}{\mu} - \log \mu$$

The score is

$$l'(\mu) = +x \frac{1}{\mu^2} - \frac{1}{\mu}$$

and the second derivative is

$$l''(\mu) = -x \frac{2}{\mu^3} + \frac{1}{\mu^2}$$

Again let us check the identities (22) and (25).
The expectation of (30) is zero. This checks because $E(X) = \mu$ so

$$E[l'(\mu)] = \frac{E(X)}{\mu^2} - \frac{1}{\mu} = \frac{\mu}{\mu^2} - \frac{1}{\mu} = 0$$

The variance of (30) is minus the expectation of (31). This checks because $\text{var}(X) = 1/\lambda^2 = \mu^2$ so the variance of (30) is

$$\text{var}[l'(\mu)] = \text{var}\left(\frac{X}{\mu^2} - \frac{1}{\mu}\right) = \left(\frac{1}{\mu^2}\right)^2 \text{var}(X) = \frac{\mu^2}{\mu^4} = \frac{1}{\mu^2}$$

and the expectation of (31) is

$$E[l''(\mu)] = -\frac{2E(X)}{\mu^3} + \frac{1}{\mu^2} = -\frac{2\mu}{\mu^3} + \frac{1}{\mu^2} = -\frac{1}{\mu^2}$$

B.2 Asymptotics of Maximum Likelihood Estimates

In Section 9.10 the textbook defines maximum likelihood estimators (MLEs) and in Sections 9.10 and 9.12 the textbook states that MLEs are asymptotically normal and asymptotically efficient.

As in the delta method (16) asymptotic normality of an estimator $T_n$ means

$$\sqrt{n}(T_n - \theta) \xrightarrow{D} \mathcal{N}(0, \sigma^2)$$

(33)

The estimator $T_n$ is efficient if its asymptotic variance $\sigma^2$ is as small as possible. The Cramér-Rao inequality (box on p. 410 in the textbook) states that if $T_n$ is unbiased, then the smallest possible variance it can have is $1/I_n(\theta)$. Hence the smallest possible variance the left hand side of (33) can have is $n/I_n(\theta) = 1/I_1(\theta)$. Thus the Cramér-Rao inequality implies that an unbiased estimator cannot have asymptotic variance less than $1/I_1(\theta)$.

The textbook remarks that maximum likelihood estimators (unbiased or not) are usually asymptotically efficient (p. 414) and asymptotically normal (p. 404). This means the asymptotic variance must be $1/I_1(\theta)$.

### Asymptotics of Maximum Likelihood

In most cases the maximum likelihood estimator $\hat{\theta}_n$ based on an i. i. d. sample of size $n$ is asymptotically normal with mean $\theta$, the true parameter value, and variance $1/I_n(\theta)$, the inverse Fisher information at the true parameter value

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{D} \mathcal{N}\left(0, \frac{1}{I_1(\theta)}\right)$$

The “in most cases” here reminds us that this is not always true. There do exist very simple examples in which these “standard asymptotics” fail, but they all involve one of the following pathologies: (1) the log likelihood is not
differentiable, (2) the support of the sampling distribution depends on $\theta$, (3) the log likelihood is unbounded and the maximum of the likelihood is achieved at $+\infty$, or (4) the sampling scheme is not i. i. d. It is remarkably difficult to find a practical sample without one of these four blemishes in which the standard asymptotics fail. Saying more than this gets into very deep mathematics, far beyond the scope of this course.

In calculating the asymptotic variance we may run into two problems. First we do not know the true parameter value $\theta$. Hence we cannot calculate $I_1(\theta)$. But if the Fisher information is a continuous function of $\theta$, then $I_1(\hat{\theta}_n)$ converges in probability to $\theta$ by the continuous mapping theorem, and we can use this estimate in constructing confidence intervals.

**Example B.2.1 Exponential Distributions Yet Again**

Continuing Examples B.1.1 and B.1.2, let us use the Fisher information calculations of those examples to derive confidence intervals. In Example B.1.1 we saw that the Fisher information for $\lambda$ is $I_1(\lambda) = 1/\lambda^2$, hence the asymptotic variance is $1/I_1(\lambda) = \lambda^2$, and a 95% confidence interval for $\lambda$ is

$$\hat{\lambda}_n \pm 1.96 \frac{\lambda_n}{\sqrt{n}}$$

Note that we got the same asymptotic variance $\lambda^2/n$ using the delta method in Example A.9.1 and using Fisher information in this example. This must be true, of course, whenever the usual asymptotics of maximum likelihood hold. After all, $\lambda_n$ is the same in both examples, so it must have the same asymptotic variance. From this we see that you don’t need the delta method to derive asymptotics for maximum likelihood. You can use Fisher information instead.

The second problem that may arise in calculating asymptotic variance is that we may not be able to explicitly calculate either of the expectations (26) or (27) defining the Fisher information. But all is not lost. We can still use (27) and the law of large numbers to estimate the Fisher information.

If we define random variables

$$Y_i = \frac{\partial^2}{\partial \theta^2} \log f(X_i | \theta)$$

then

$$\Gamma_n = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \log f(X_i | \theta) = \frac{1}{n} I_n(\theta)$$

By (27) $E_\theta(Y_i) = -I_1(\theta)$. Hence the law of large numbers applied to the variables $Y_i$ says $\Gamma_n \xrightarrow{P} -I_1(\theta)$. This gives yet another estimator of the Fisher information.
Observed Fisher Information

The observed Fisher information

\[ I_{\text{obs}}(\theta) = -\frac{1}{n} \frac{d^2}{d\theta^2} \log f(X_i|\theta) \] (34)

is a consistent estimator of the Fisher information \( I_1(\theta) \).

The observed Fisher information can be calculated whenever the likelihood can
be calculated, since it involves no expectations, only derivatives. \( I_1(\theta) \) is sometimes called the expected Fisher information for contrast with observed Fisher information, though strictly speaking the “expected” is redundant.

Example B.2.2 Exponential Distributions

In Example B.1.1 we found that \( l''(\lambda) \) was constant, hence equal to its expectation. So there was no difference between observed and expected Fisher information. In Example B.1.2 the observed and expected Fisher information were different: (31) is different from (32).

Example B.2.3 Exponential Families

The asymptotics of maximum likelihood are particularly simple for exponential families. Consider a one-parameter exponential family having p. d. f.

\[ f(x|\theta) = B(\theta)h(x)e^{Q(\theta)R(x)} \]

First we introduce the natural parameter \( \varphi = Q(\theta) \). Then in order that \( f(x|\theta) \) integrate to one, we must have

\[ f(x|\theta) = \frac{1}{C(\varphi)} h(x)e^{\varphi R(x)} \]

where

\[ C(\varphi) = \int h(x)e^{\varphi R(x)} \, dx \]

Then the log likelihood for \( \varphi \) for a sample of size one is

\[ l(\varphi) = \varphi R(x) - \log C(\varphi) \]

and the score is

\[ l'(\varphi) = R(x) - \frac{C'(\varphi)}{C(\varphi)} \] (35)

It is a fact about exponential families that the differentiation under the integral
sign involved in deriving the identities (22) and (25) is always valid. It is equivalent to a proof that differentiating a moment generating function always
gives the moments, something we also skipped. In order that the identity (22) hold we must have that the expectation of (35) is zero or

\[ E_\varphi R(X) = \frac{C''(\varphi)}{C(\varphi)} \]

Hence the likelihood equation is \( R(x) = E_\varphi R(X) \). This is true for any exponential family. The MLE is the parameter value that makes the expected value of the natural statistic equal to its observed value. This is shortened to “observed equals expected” to make a simple mnemonic.

Differentiating again we get

\[ I''(\varphi) = -\frac{d}{d \varphi} \frac{C''(\varphi)}{C(\varphi)} \]

which does not depend on the observation \( x \). Since \( I''(\varphi) \) does not depend on \( x \), it is its own expectation. Hence the observed and expected Fisher information for the natural parameter are the same. From identity (25) the Fisher information is the variance of the score, which we see from (35) is the variance of the natural statistic. Summarizing

**Maximum Likelihood in Exponential Families**
The MLE is the solution of “observed equals expected”

\[ R(x) = E_\varphi R(X) \]

and the Fisher information for the natural parameter is the variance of the natural statistic

\[ I_1(\varphi) = \text{var}_\varphi R(X) \]

What about a sample of size \( n \)? That just produces a different exponential family. The joint density becomes

\[ f(x|\varphi) = \frac{1}{C(\varphi)^n} \left( \prod_{i=1}^{n} h(x_i) \right) \exp \left( \varphi \sum_{i=1}^{n} R(x_i) \right) \]

This is also an exponential family with natural parameter \( \varphi \) and natural statistic \( \sum_i R(x_i) \). So everything holds as before. We just replace \( R(x) \) by \( \sum_i R(x_i) \).

If we look back at the exponential families we identified in problem 8-10 we see that

- For the geometric family, the natural statistic is \( X \), hence the MLE is the \( p \) that satisfies \( x = E(X) = 1/p \) in the case of a single observation or \( \sum_i x_i = E(\sum_i X_i) = n/p \) in the case of a sample of size \( n \). The Fisher information for the natural parameter \( \varphi = \log p \) is \( \text{var}(X) = q/p^2 \).
• For the gamma family with known \( \alpha \), the natural statistic is \( X \), hence the MLE is the \( \lambda \) that satisfies \( x = E(X) = \alpha / \lambda \). The Fisher information for the natural parameter \( \lambda \) is \( \text{var}(X) = \alpha / \lambda^2 \).

• For the unnamed family of part (c), the natural statistic is \( \log X \) which it is not obvious how to integrate. However, it is easy to differentiate the log likelihood

\[
l(\theta) = \log(\theta) + \theta \log(x)
\]

obtaining

\[
l'(\theta) = \frac{1}{\theta} + \log(x)
\]

and

\[
l''(\theta) = -\frac{1}{\theta^2}
\]

And thus we see that it must be true that

\[
E_\theta(\log X) = -\frac{1}{\theta}
\]

and

\[
\text{var}_\theta(\log X) = \frac{1}{\theta^2}
\]

The identities discovered in this example are sometimes helpful, sometimes not. They only give the Fisher information for the natural parameter, not for other parameters. Sometimes the mean and variance of the natural statistic are known and immediately give the MLE and Fisher information without writing down the likelihood or differentiating. Sometimes, as in part (c), we can use the identities backwards to discover the mean and variance of the natural statistic by differentiating the log likelihood.

**Example B.2.4 A Problem in Genetics**

In his influential book *Statistical Methods for Research Workers* first published in 1925, R. A. Fisher described the following problem in genetics. The data have a trinomial distribution (Sec 4.8 in our textbook) with sample size \( n \) and three categories. The observed data form a vector \( X = (X_1, X_2, X_3) \) of the counts in the three categories. The joint p. d. f. is

\[
f(x) = \binom{n}{x_1, x_2, x_3} p_1^{x_1} p_2^{x_2} p_3^{x_3}
\]

The probabilities of the categories are all functions of a single parameter

\[
p_1 = \frac{1}{4}(2 + \theta)
\]

\[
p_2 = \frac{1}{4}(1 - \theta)
\]

\[
p_3 = \frac{1}{4}\theta
\]

The parameter \( \theta \) must satisfy \( 0 \leq \theta \leq 1 \) in order that these cell probabilities be between zero and one.
The log likelihood is
\[
l(\theta) = \log f(x|\theta)
\]
\[
= \sum_{i=1}^{3} x_i \log p_i(\theta)
\]
\[
= x_1 \log(2 + \theta) + x_2 \log(1 - \theta) + x_3 \log(\theta) + \text{terms not containing } \theta
\]
The “terms not containing \( \theta \)” do not affect derivatives with respect to \( \theta \). So the score is
\[
l'(\theta) = \frac{x_1}{2 + \theta} - \frac{x_2}{1 - \theta} + \frac{x_3}{\theta}
\]  
(36)
In order to find the maximum likelihood estimate, we need to solve the equation \( l'(\theta) = 0 \). Multiplying through by the denominators gives
\[
x_1 (1 - \theta) \theta - (2 + \theta) x_2 \theta + (2 + \theta)(1 - \theta) x_3 = 0
\]
or simplifying a bit
\[
2x_3 + (x_1 - 2x_2 - x_3) \theta - (x_1 + x_2 + x_3) \theta^2 = 0
\]
or
\[
n \theta^2 - (x_1 - 2x_2 - x_3) \theta - 2x_3 = 0
\]
since the sum of the category counts is the sample size \( n \). In the general case with all the \( x_i \) greater than zero, the MLE is a solution of the quadratic equation
\[
\hat{\theta} = \frac{x_1 - 2x_2 - x_3 \pm \sqrt{(x_1 - 2x_2 - x_3)^2 + 8nx_3}}{2n}
\]
Since the square root is larger than the first term of the numerator, choosing the minus sign always gives a negative solution, which is impossible. Hence the MLE is
\[
\hat{\theta} = \frac{x_1 - 2x_2 - x_3 + \sqrt{(x_1 - 2x_2 - x_3)^2 + 8nx_3}}{2n}
\]
Fisher gave a specific example with \( x_1 = 1997 \), \( x_2 = 1810 \), and \( x_3 = 32 \), so \( n = 3839 \). For these data
\[
\hat{\theta} = \frac{1997 - 2 \cdot 1810 - 32 + \sqrt{(1997 - 2 \cdot 1810 - 32)^2 + 8 \cdot 3839 \cdot 32}}{2 \cdot 3839}
\]
\[
= \frac{-1655 + \sqrt{1655^2 + 982784}}{7678}
\]
\[
= 0.0357123
\]
We now need to figure out the Fisher information. Finding the variance of the score (36) is a bit tricky because the \( x_i \) are correlated. Let’s instead calculate using the second derivative
\[
l''(\theta) = -\frac{x_1}{(2 + \theta)^2} - \frac{x_2}{(1 - \theta)^2} - \frac{x_3}{\theta^2}
\]
Since the marginal distribution of $X_i$ is Bin($n, p_i$) (p. 150 in our textbook) $E(X_i) = np_i$ and

$$I_n(\theta) = \frac{np_1}{(2 + \theta)^2} + \frac{np_2}{(1 - \theta)^2} + \frac{np_3}{\theta^2}$$

$$= n \left( \frac{1}{2 + \theta} + \frac{1}{1 - \theta} + \frac{1}{\theta} \right)$$

$$= n \left( \frac{1}{2 + \theta} + \frac{2}{1 - \theta} + \frac{1}{\theta} \right)$$

(and $I_1(\theta)$ is the same without the $n$). Plugging the data into these formulas gives

$$I_{\text{obs}}(\hat{\theta}) = \frac{1}{3839} \left( \frac{1997}{(2 + 0.0357123)^2} + \frac{1810}{(1 - 0.0357123)^2} + \frac{32}{0.0357123^2} \right)$$

$$= 7.16833$$

and

$$I_1(\hat{\theta}) = \frac{1}{4} \left( \frac{1}{2 + 0.0357123} + \frac{2}{1 - 0.0357123} + \frac{1}{0.0357123} \right)$$

$$= 7.64171$$

We may use either to construct confidence intervals. If we use the observed Fisher information, we get

$$\hat{\theta} \pm 1.96 \sqrt{\frac{1}{n I_{\text{obs}}(\hat{\theta})}}$$

The “plus or minus” is $1.96/\sqrt{3839} \cdot 7.16833 = 0.0118149$, so our 95% confidence interval is $0.036 \pm 0.012$. If we use expected Fisher information instead the “plus or minus” would be $0.0114431$, almost the same.

**Problems**

**B-1.** Given a random sample of size $n$ from Geo($p$), find the MLE of $p$, the Fisher information for $p$, and a 95% large-sample confidence interval for $p$ based on the Fisher information.

**B-2.** Given a random sample of $n$ pairs $(X_i, Y_i)$ from a bivariate population with joint p.d.f. $f(x, y) = \exp(-x\theta - y/\theta)$ for $x > 0$ and $y > 0$, where $\theta$ is a positive parameter. Find the MLE of $\theta$, the Fisher information for $\theta$, and a 95% large-sample confidence interval for $\theta$ based on the Fisher information.

**B-3.** Given a random sample of size $n$ from a $N(0, \sigma^2)$ where $\sigma$ is a positive parameter, find the MLE of $\sigma$, the Fisher information for $\sigma$ and a 95% large-sample confidence interval for $\sigma$ based on the Fisher information.
B-4. Given a random sample of size $n$ from a $N(\theta, \theta)$ where $\theta$ is a positive parameter, find the MLE of $\theta$, the Fisher information for $\theta$, and a 95\% large-sample confidence interval for $\theta$ based on the Fisher information.