Stat 5102 Notes: More on Confidence Intervals

Charles J. Geyer

February 24, 2003

1 The Pivotal Method

A function $g(\mathbf{X}, \theta)$ of data and parameters is said to be a *pivot* or a *pivotal quantity* if its distribution does not depend on the parameter. The primary example of a pivotal quantity is

$$g(\mathbf{X}, \mu) = \frac{\overline{X}_n - \mu}{S_n / \sqrt{n}} \tag{1.1}$$

which has the distribution $t(n-1)$, when the data $X_1, \ldots, X_n$ are i. i. d. Normal(μ, σ^2) and

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{1.2a}$$

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 \tag{1.2b}$$

Pivotal quantities allow the construction of *exact* confidence intervals, meaning they have exactly the stated confidence level, as opposed to so-called "asymptotic" or "large-sample" confidence intervals which only have approximately the stated confidence level and that only when the sample size is large. An exact confidence interval is valid for any sample size. An asymptotic confidence interval is valid only for sufficiently large sample size (and typically one does not know how large is large enough).

Exact intervals are constructed as follows.

- Find a pivotal quantity $g(\mathbf{X}, \theta)$.
- Find upper and lower confidence limits on the pivotal quantity, that is, numbers c_1 and c_2 such that

$$\Pr\{c_1 < g(\mathbf{X}, \theta) < c_2\} = \gamma \tag{1.3}$$

  where γ is the desired confidence coefficient.

- Solve the inequalities: the confidence set (usually an interval) is

$$\{ \theta \in \Theta : c_1 < g(\mathbf{X}, \theta) < c_2 \} \tag{1.4}$$

The point is that the probability (1.3) does not depend on the parameter θ by definition of "pivotal quantity." If $g(\mathbf{X}, \theta)$ were not pivotal, then the probability (1.3) would depend on the unknown true parameter value and could not be calculated.

The constants c_1 and c_2 in (1.3) are called *critical values*. They are obtained from a table for the distribution of the pivotal quantity or from a computer program.

1.1 Pivot for Normal Mean

For the t pivotal quantity (1.1) we usually choose symmetric critical values: for some positive number c, we choose $c_2 = c$ and $c_1 = -c$. Then the inequalities in (1.4) become

$$-c < \frac{\overline{X}_n - \mu}{S_n / \sqrt{n}} < c$$

which when solved for μ are equivalent to

$$\overline{X}_n - c \frac{S_n}{\sqrt{n}} < \mu < \overline{X}_n + c \frac{S_n}{\sqrt{n}}$$

the usual t confidence interval.

1.2 Pivot for Exponential Rate

For the t interval, we just relearned what we already knew. Here's another example. Suppose $X_1, \ldots, X_n$ are i. i. d. Exponential(λ). Then we know from the addition rule for the exponential that

$$\sum_{i=1}^{n} X_i \sim \text{Gamma}(n, \lambda).$$

Then because the second parameter of the gamma distribution is a "rate" parameter (reciprocal scale parameter) multiplying by a constant gives another gamma random variable with the same shape and rate divided by that constant (DeGroot and Schervish, Problem 1 of Section 5.9). We choose to multiply by λ / n giving

$$\lambda \overline{X}_n \sim \text{Gamma}(n, n) \tag{1.5}$$

Since the distribution here does not depend on the parameter λ, we see that

$$g(\mathbf{X}, \lambda) = \lambda \overline{X}_n$$

is a pivotal quantity. Hence we can apply the pivotal method.

When the distribution of the pivotal quantity is not symmetric, there is no reason to choose symmetric critical values (plus and minus some number). In this case it is impossible to choose symmetric critical values. Negative critical values make no sense when the pivotal quantity is almost surely positive.

A convenient choice (which specializes to symmetric critical values when the distribution of the pivotal quantity is symmetric) is to use *equal-tailed* critical values. We choose c_1 and c_2 to be the $\alpha/2$ and $1 - \alpha/2$ quantiles of the distribution of the pivotal quantity, where $\alpha = 1 - \gamma$ and γ is the confidence coefficient.

For the pivotal quantity (1.5), the following R statements find these critical values, assuming the sample size `n` and confidence coefficient `alpha` are already defined

```
qgamma(alpha / 2, shape = n, rate = n)
qgamma(1 - alpha / 2, shape = n, rate = n)
```

Using the vectorizing property of R functions, we can get both with one statement

```
qgamma(c(alpha / 2, 1 - alpha / 2), shape = n, rate = n)
```

For a concrete example, suppose $n = 10$ and $\bar{x}_n = 23.45$ and we want a 95% confidence interval, so $\gamma = 0.95$ and $\alpha = 0.05$. The the code above gives

```
Rweb:> n <- 10
Rweb:> alpha <- 0.05
Rweb:> qgamma(c(alpha / 2, 1 - alpha / 2), shape = n, rate = n)
[1] 0.4795389 1.7084803
```

So our confidence interval found by the pivotal method is

$$0.4795389 < \lambda \overline{X}_n < 1.708480$$

solved for λ, that is

$$\frac{0.4795389}{\overline{X}_n} < \lambda < \frac{1.708480}{\overline{X}_n}$$

or using the computer

```
Rweb:> n <- 10
Rweb:> alpha <- 0.05
Rweb:> xbar <- 23.45
Rweb:> crit <- qgamma(c(alpha / 2, 1 - alpha / 2),
+     shape = n, rate = n)
Rweb:> print(crit / xbar)
[1] 0.02044942 0.07285630
```

1.3 Pivot for Exponential Rate, Part 2

If you don't have a computer and must use tables in the back of some book, then you probably don't find tables of the gamma distribution. What you do find is tables of the chi-square distribution, which is a gamma distribution with integer or half-integer degrees of freedom and rate parameter 1/2. Integer degrees of freedom is what we need for estimating the rate parameter of the exponential, and the rate parameter can be adjusted to what we want by multiplying by a constant

$$2n\lambda\overline{X}_n \sim \text{Chi-Square}(2n) \tag{1.6}$$

Note that the degrees of freedom becomes $2n$ because that makes the shape parameter of the gamma distribution n.

Now we find critical values for an equal-tailed 95% confidence interval from the table on pp. 774–775 in DeGroot and Schervish. For $2n = 20$ degrees of freedom the 0.025 and 0.975 quantiles are 9.591 and 34.17, and the corresponding confidence interval for λ is

$$\frac{9.591}{20\overline{X}_n} < \lambda < \frac{34.17}{20\overline{X}_n}$$

and for $\bar{x}_n = 23.45$ this works out to (0.02044989, 0.07285714), which agrees with the answer we got before to four significant figures (the accuracy of the table in the book).

1.4 Philosophy

It must be admitted that exactness depends crucially on assumptions. The t pivotal quantity (1.1) has the $t(n-1)$ distribution it is supposed to have if (a very big if!) the assumptions (i. i. d. normal data) are true. If the assumptions are incorrect, then the so-called "exact" confidence isn't actually exact. So this "exact" language must be taken with a grain of salt. It all depends on whether the required assumptions are true, and in reality they are never true. Real data are messy, never exactly obeying a simple theoretical model.

So exactness is best thought of as just another kind of approximation, one that isn't critically dependent on large sample size but is critically dependent on other distributional assumptions.

2 The Asymptotic Method

2.1 Convergence in Distribution

In order to understand the asymptotic method for confidence intervals (and later for hypothesis tests), we need to a better understanding of the central limit theorem than we can get from reading DeGroot and Schervish.

First we define (which DeGroot and Schervish don't, at least not this precisely) the notion of convergence in distribution. A sequence of random variables

X_n having distribution functions F_n *converges in distribution* to a random variable X having distribution function F if

$$F_n(x) \to F(x), \qquad \text{for every } x \text{ at which } F \text{ is continuous.}$$

This is denoted

$$X_n \xrightarrow{\mathcal{D}} X.$$

2.2 The Central Limit Theorem (CLT)

As DeGroot and Schervish note (p. 288) the central limit theorem uses this notion.

Theorem 2.1 (Central Limit Theorem). *If X_1, X_2, ... are independent and identically distributed random variables having mean μ and variance σ^2 and $\overline{X}_n$ is defined by* (1.2a), *then*

$$\sqrt{n}\left(\overline{X}_n - \mu\right) \xrightarrow{\mathcal{D}} Y, \qquad \text{as } n \to \infty, \tag{2.1}$$

where $Y \sim \text{Normal}(0, \sigma^2)$.

It simplifies notation if we are allowed to write a distribution on the right hand side of a statement about convergence in distribution, simplifying (2.1) and the rest of the sentence following it to

$$\sqrt{n}\left(\overline{X}_n - \mu\right) \xrightarrow{\mathcal{D}} \text{Normal}(0, \sigma^2), \qquad \text{as } n \to \infty. \tag{2.2}$$

There's nothing wrong with this mixed notation because convergence in distribution is a statement about distributions of random variables, not about the random variables themselves. So when we replace a random variable with its distribution, the meaning is still clear.

A sloppy way of rephrasing (2.2) is

$$\overline{X}_n \approx \text{Normal}\left(\mu, \frac{\sigma^2}{n}\right) \tag{2.3}$$

for "large n." Most of the time the sloppiness causes no harm and no one is confused. The mean and variance of $\overline{X}_n$ are indeed μ and σ^2/n and the shape of the distribution is approximately normal if n is large. What one cannot do is say $\overline{X}_n$ converges in distribution to Z, where $Z \sim \text{Normal}(\mu, \sigma^2/n)$. Having an n in the supposed limit of a sequence is mathematical nonsense. To make mathematical sense, all of the n's must be on the left hand side of the limit statement, as they are in (2.1) and (2.2).

2.3 Convergence in Probability to a Constant

A special case of convergence in distribution is convergence in distribution to a degenerate random variable concentrated at one point, which we denote by a small letter a to show that it is a constant random variable. It turns out this is a concept we have met before. Convergence in probability to a constant was defined in Section 4.8 of DeGroot and Schervish. It is denoted $X_n \xrightarrow{P} a$.

Theorem 2.2. *If X_1, X_2, ... is a sequence of random variables and a a constant, then $X_n \xrightarrow{P} a$ if and only if $X_n \xrightarrow{\mathcal{D}} a$.*

In view of the theorem, we could dispense with the notion of convergence in probability to a constant and just use convergence in distribution to a constant. The reason we don't is social. For some reason, the social convention among statisticians is to always write $X_n \xrightarrow{P} a$ and never write $X_n \xrightarrow{\mathcal{D}} a$ when a is a constant and the two mean the same thing. We will go along with the rest of the herd.

2.4 The Law of Large Numbers (LLN)

As DeGroot and Schervish note (p. 234) the central limit theorem uses this notion (although their statement of the theorem has an irrelevant condition).

Theorem 2.3 (Law of Large Numbers). *If X_1, X_2, ... is a sequence of independent, identically distributed random variables having mean μ, and $\overline{X}_n$ is defined by* (1.2a)*, then*

$$\overline{X}_n \xrightarrow{P} \mu, \qquad \text{as } n \to \infty. \tag{2.4}$$

The difference between our statement and the one in DeGroot and Schervish is that they add the condition that the X_i have second moments to the LLN. But then the conditions of the LLN and the CLT are the same. Since the conclusion of the CLT is much stronger than the conclusion of the LLN (see Section 2.6 below), there is no point to the LLN if this irrelevant condition is added. (The reason they impose this irrelevant condition is to get a simple proof. The theorem as stated here is too difficult to prove with the techniques developed in this course).

An example of a theorem for which the LLN as stated here holds but the CLT does not is Student's t distribution with two degrees of freedom. The mean exists (and is zero) but the variance does not exist. Thus we have (2.4) with $\mu = 0$, but we do not have (2.1).

2.5 Slutsky's Theorem

Theorem 2.4 (Slutsky). *If $g(x, y)$ is a function jointly continuous at every point of the form (x, a) for some fixed a, and if $X_n \xrightarrow{\mathcal{D}} X$ and $Y_n \xrightarrow{P} a$, then*

$$g(X_n, Y_n) \xrightarrow{\mathcal{D}} g(X, a).$$

Corollary 2.5. *If* $X_n \xrightarrow{\mathcal{D}} X$ *and* $Y_n \xrightarrow{P} a$*, then*

$$X_n + Y_n \xrightarrow{\mathcal{D}} X + a,$$
$$Y_n X_n \xrightarrow{\mathcal{D}} aX,$$

and if $a \neq 0$

$$X_n / Y_n \xrightarrow{\mathcal{D}} X/a.$$

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 $X_n \xrightarrow{\mathcal{D}} X$ and $Y_n \xrightarrow{\mathcal{D}} Y$, 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.

Slutsky's theorem can be extended by mathematical induction to many sequences converging in probability. If $X_n \xrightarrow{\mathcal{D}} X$, $Y_n \xrightarrow{P} a$, and $Z_n \xrightarrow{P} b$ then $X_n + Y_n + Z_n \xrightarrow{\mathcal{D}} X + a + b$, and so forth. Note that this induction argument cannot be applied when we have several sequences converging in distribution. Slutsky's theorem only allows one sequence converging in distribution.

2.6 Comparison of the LLN and the CLT

When X_1, X_2, ... is an i. i. d. sequence of random variables having second moments, both the law of large numbers and the central limit theorem apply, but the CLT tells us much more than the LLN.

It could not tell us less, because the CLT implies the LLN. By Slutsky's theorem, the CLT (2.1) implies

$$\overline{X}_n - \mu = \frac{1}{\sqrt{n}} \cdot \sqrt{n}\left(\overline{X}_n - \mu\right) \xrightarrow{\mathcal{D}} 0 \cdot Y = 0$$

where $Y \sim \text{Normal}(0, \sigma^2)$. Another application of Slutsky's theorem, this time with the sequence converging in probability being a constant sequence, shows that $\overline{X}_n - \mu \xrightarrow{\mathcal{D}} 0$ implies $\overline{X}_n \xrightarrow{\mathcal{D}} \mu$. By Theorem 2.2 $\overline{X}_n \xrightarrow{\mathcal{D}} \mu$ is the same as $\overline{X}_n \xrightarrow{P} \mu$. Thus the conclusion of the CLT

$$\sqrt{n}\left(\overline{X}_n - \mu\right) \xrightarrow{\mathcal{D}} Y$$

implies the conclusion of the LLN

$$\overline{X}_n \xrightarrow{P} \mu.$$

And this tells us that the CLT is the stronger of the two theorems.

2.7 Consistent Estimation

A sequence of point estimators T_n of a parameter θ is *consistent* if

$$T_n \xrightarrow{P} \theta, \qquad \text{as } n \to \infty.$$

Generally we aren't so pedantic as to emphasize that consistency is really a property of a sequence. We usually just say T_n is a consistent estimator of θ.

Consistency is not a very strong property, since it doesn't say anything about how fast the errors go to zero nor does it say anything about the distribution of the errors. So we generally aren't interested in estimators that are merely consistent unless for some reason consistency is all we want. The most important instance when consistency is all we want is explained in Section 2.9.1 below.

By the law of large numbers, if X_1, X_2, $\ldots$ are i. i. d. from a distribution with mean μ, then the sample mean $\overline{X}_n$ is a consistent estimator of μ. The only requirement is that the expectation μ exist.

More generally, any moment that exists can be consistently estimated. First consider ordinary moments

$$\alpha_k = E(X^k). \tag{2.5}$$

If X_1, X_2, $\ldots$ are i. i. d. with the same distribution as X, then the natural estimator of the theoretical moment (2.5) is the *sample moment*

$$\hat{\alpha}_{k,n} = \frac{1}{n} \sum_{i=1}^{n} X_i^k. \tag{2.6}$$

A very simple argument shows that $\hat{\alpha}_{k,n}$ is a consistent estimator of α_k. Define $Y_i = X_i^k$. Then $\overline{Y}_n = \hat{\alpha}_{k,n}$ and $E(Y_i) = \mu_Y = \alpha_k$ and the consistency statement

$$\hat{\alpha}_{k,n} \xrightarrow{P} \alpha_k$$

is just the law of large numbers

$$\overline{Y}_n \xrightarrow{P} \mu_Y$$

in different notation.

For central moments, the argument becomes a bit more complicated. The k-th central moment of a random variable X with mean μ is defined by

$$\mu_k = E\{(X - \mu)^k\} \tag{2.7}$$

assuming the expectation exists. If X_1, X_2, $\ldots$ are i. i. d. with the same distribution as X, then the natural estimator of the theoretical moment (2.7) is the *sample central moment*

$$\hat{\mu}_{k,n} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^k. \tag{2.8}$$

The LLN does not apply directly to sample central moments because (2.8) is not the average of i. i. d. terms, because $\overline{X}_n$ appears in each term and involves all the X_i. Nevertheless, a few applications of Slutsky's theorem shows that $\hat{\mu}_{k,n}$ is a consistent estimator of μ_k.

In order to apply the LLN, we define

$$\tilde{\mu}_{k,n} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^k. \tag{2.9}$$

This is the average of i. i. d. terms and the expectation of each term is μ_k. Hence

$$\tilde{\mu}_{k,n} \xrightarrow{P} \mu_k$$

by the LLN, the argument being just the same as the argument for ordinary moments.

Now we rewrite $\hat{\mu}_{k,n}$ in terms of $\tilde{\mu}_{k,n}$ using the binomial theorem

$$\begin{aligned}
\hat{\mu}_{k,n} &= \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^k \\
&= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=0}^{k} \binom{k}{j} (-1)^j (\overline{X}_n - \mu)^j (X_i - \mu)^{k-j} \\
&= \sum_{j=0}^{k} \binom{k}{j} (-1)^j (\overline{X}_n - \mu)^j \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^{k-j} \\
&= \sum_{j=0}^{k} \binom{k}{j} (-1)^j (\overline{X}_n - \mu)^j \tilde{\mu}_{k-j,n}
\end{aligned}$$

Now as we already saw in Section 2.6

$$\overline{X}_n - \mu \xrightarrow{P} 0$$

(this already involved one application of Slutsky's theorem). Another application of the theorem implies

$$(\overline{X}_n - \mu)^j \tilde{\mu}_{k-j,n} \xrightarrow{P} 0^j \mu_{k-j}$$

and $0^j = 0$ for all terms except the $j = 0$ term, where $0^j = 1$. Then a final application of Slutsky's theorem gives us that the limit of the sum is the sum of the limits, and since all terms but the first are zero, we get

$$\hat{\mu}_{k,n} \xrightarrow{P} \binom{k}{0}(-1)^0 0^0 \mu_k = \mu_k$$

which is the consistency statement we wanted to prove.

The most important example of consistency of sample central moments is the second moment

$$\hat{\mu}_{2,n} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \xrightarrow{P} \mu_2 = \sigma^2$$

It is a nuisance that the sample central moment $\hat{\mu}_{2,n}$ is not commonly used. Rather the slightly different (1.2b) is widely used, because of its appearance in the usual formulation of the t pivotal quantity. Asymptotically, there is no difference between the two estimators. By Slutsky's theorem

$$S_n^2 = \frac{n-1}{n}\hat{\mu}_{2,n} \xrightarrow{P} 1 \cdot \sigma^2 = \sigma^2$$

Another application of Slutsky's theorem, using the fact that the square root function is continuous, gives us

$$S_n \xrightarrow{P} \sigma.$$

In summary.

- S_n^2 is a consistent estimator of σ^2.
- S_n is a consistent estimator of σ.

2.8 Asymptotic Normality

Mere consistency is a fairly uninteresting property, unless it just happens to be all we want. A much more important property is asymptotic normality. We say an estimator T_n is consistent for a parameter θ if

$$T_n \xrightarrow{P} \theta,$$

or equivalently if

$$T_n - \theta \xrightarrow{P} 0.$$

The estimator T_n is supposed to estimate the parameter θ, so $T_n - \theta$ is the error of estimation. Consistency says the error goes to zero. We would like to know more than that. We would like to know how about big the error is, more specifically we would like an approximation of its sampling distribution.

It turns out that almost all estimators of practical interest are not just consistent but also *asymptotically normal*, that is,

$$\sqrt{n}(T_n - \theta) \xrightarrow{\mathcal{D}} \text{Normal}(0, \tau^2) \tag{2.10}$$

holds for some constant τ^2, which may depend on the true distribution of the data. We say an estimator T_n that satisfies (2.10) is *consistent and asymptotically normal* (that is, asymptotically normal when centered at θ).

The simplest example of this phenomenon is when T_n is the sample mean $\overline{X}_n$ and the parameter θ is the population mean μ. Asymptotic normality of this estimator is just the CLT (2.2). But many other estimators of many other parameters are consistent and asymptotically normal. For example, every sample moment is a consistent and asymptotically normal estimator of the corresponding theoretical moment if enough theoretical moments exist. In order for sample k-th moments to be consistent and asymptotically normal we need existence of theoretical moments of order $2k$.

2.9 Asymptotically Pivotal Quantities

An estimator that is consistent and asymptotically normal can be used to construct, so-called *asymptotic* or *large-sample* confidence intervals if asymptotics can be arranged so that the asymptotic distribution does not depend on parameters. Then we say we have an asymptotically pivotal quantity.

2.9.1 The Plug-In Principle

The CLT does not at first sight produce an asymptotically pivotal quantity, because the asymptotic distribution $\text{Normal}(0, \sigma^2)$ depends on the unknown parameter σ. We can divide by σ giving (by Slutsky's theorem)

$$\frac{\overline{X}_n - \mu}{\sigma / \sqrt{n}} \xrightarrow{\mathcal{D}} \text{Normal}(0, 1). \tag{2.11}$$

Since the right hand side now contains no unknown parameter, we say the left hand side is an asymptotically pivotal quantity. But this is entirely useless because we don't know σ. If we were to try to use this to make a confidence interval for μ we would get the interval

$$\overline{X}_n - c \frac{\sigma}{\sqrt{n}} < \mu < \overline{X}_n + c \frac{\sigma}{\sqrt{n}}$$

where c is the critical value (more on that below). If we don't know σ, then we can't use this.

So we proceed a bit differently. Rewrite (2.11) as

$$\frac{\overline{X}_n - \mu}{\sigma / \sqrt{n}} \xrightarrow{\mathcal{D}} Z, \tag{2.12}$$

where Z is standard normal. Then by yet another application of Slutsky's theorem,

$$\frac{\overline{X}_n - \mu}{S_n / \sqrt{n}} = \frac{\sigma}{S_n} \cdot \frac{\overline{X}_n - \mu}{\sigma / \sqrt{n}} \xrightarrow{\mathcal{D}} \frac{\sigma}{\sigma} \cdot Z = Z \tag{2.13}$$

In summary

$$\frac{\overline{X}_n - \mu}{S_n / \sqrt{n}} \xrightarrow{\mathcal{D}} \text{Normal}(0, 1). \tag{2.14}$$

Thus (2.14) describes an asymptotically pivotal quantity, just like (2.11). But there is a world of difference between the two: (2.14) is useful when σ is unknown and (2.11) is useless.

The plug-in principle can be vastly generalized to any situation in which we have consistent and asymptotically normal estimators.

Theorem 2.6 (Plug-In). *Suppose* (2.10) *holds and* U_n *is any consistent estimator of* τ*, then*

$$\frac{T_n - \theta}{U_n / \sqrt{n}} \xrightarrow{\mathcal{D}} \text{Normal}(0, 1). \tag{2.15}$$

The argument is just like the argument leading to (2.14). Note that (2.14) is the special case of (2.15) with

- $T_n = \overline{X}_n$,
- $\theta = \mu$,
- $U_n = S_n$,
- and $\tau^2 = \sigma^2$.

2.9.2 Confidence Intervals Using Plug-In

The method of asymptotic pivotal quantities proceeds just like the method of exact pivotal quantities. We just replace exact probability statements by asymptotic (large-sample) approximate probability statements.

Suppose (2.15) holds. Choose c such that $\Pr(|Z| < c) = \gamma$, where Z is standard normal and γ is the desired confidence level. Then

$$T_n - c\frac{U_n}{\sqrt{n}} < \theta < T_n + c\frac{U_n}{\sqrt{n}}$$

is the desired (large-sample, asymptotic, approximate) confidence interval. In the particular case where (2.14) holds, this becomes

$$\overline{X}_n - c\frac{S_n}{\sqrt{n}} < \mu < \overline{X}_n + c\frac{S_n}{\sqrt{n}}$$

Most commonly, people use $\gamma = 0.95$ for which $c = 1.96$ or $\gamma = 0.90$ for which $c = 1.645$.

Here's an example where we don't use S_n to estimate asymptotic variance. Suppose X_1, ..., X_n are i. i. d. Bernoulli(p) and we are to estimate the unknown success probability p. This is the same as saying that $Y = X_1 + \cdots + X_n$ is Binomial(n, p). Usually when this situation arises the problem is phrased in terms of Y rather than in terms of X_1, ..., X_n. But in order to see the relation to other asymptotic arguments, it is important to keep the i. i. d. Bernoulli variables in the picture. Then $\overline{X}_n$ is the same as Y/n the fraction of successes in n trials. It is commonly denoted $\hat{p}_n$ in this context, but is is a sample mean whether or not it is denoted $\overline{X}_n$. From the formulas for the Bernoulli distribution we know that

$$\sigma^2 = \operatorname{Var}(X_i) = p(1-p)$$

Thus for Bernoulli or binomial data the estimate of p can also be used to estimate σ^2 and σ. By Slutsky's theorem (yet again)

$$\sqrt{\hat{p}_n(1-\hat{p}_n)} \xrightarrow{P} \sqrt{p(1-p)}$$

So from the plug-in theorem we get

$$\frac{\hat{p}_n - p}{\sqrt{\hat{p}_n(1-\hat{p}_n)/n}} \xrightarrow{\mathcal{D}} \text{Normal}(0,1). \tag{2.16}$$

Solving the asymptotically pivotal quantity for p gives

$$\hat{p}_n - c\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}} < p < \hat{p}_n + c\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}$$

where c is the usual normal critical value (like 1.96 for 95% confidence).

2.9.3 Confidence Intervals Without Plug-In

Although plug-in is by far the most common way to make confidence intervals it is possible in special cases to avoid it.

Let us redo the binomial example avoiding plug-in. If we don't use plug-in, then the CLT gives

$$\frac{\hat{p}_n - p}{\sqrt{p(1-p)/n}} \xrightarrow{\mathcal{D}} \text{Normal}(0,1). \tag{2.17}$$

And the left hand side is an asymptotically pivotal quantity. Unlike the general case, this is a useful pivotal quantity because the only unknown parameter in it is p, which is what we are trying to estimate. Thus if c is the appropriate normal critical value, the set

$$\left\{ p : \frac{|\hat{p}_n - p|}{\sqrt{p(1-p)/n}} < c \right\} \tag{2.18}$$

is an asymptotic (large-sample, approximate) confidence interval for p. The "solution" here is not so easy. It involves solving a quadratic equation in p, but it is possible. The solutions for the endpoints of the confidence interval for p are

$$\frac{\hat{p}_n + \dfrac{c^2}{2n} \pm c\sqrt{\dfrac{c^2}{4n^2} + \dfrac{\hat{p}_n(1-\hat{p}_n)}{n}}}{\left(1 + \dfrac{c^2}{n}\right)} \tag{2.19}$$

For large n (2.19) is very close to

$$\hat{p}_n \pm c\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}} \tag{2.20}$$

which gives the endpoints of the interval obtained by the plug-in method. But for small n, the two kinds of intervals can be quite different. Simulation studies show the method of this section, generally works better, meaning the actual coverage probability of the interval is closer to its nominal coverage probability.

2.9.4 Confidence Intervals Without Plug-In, Again

Variance-stabilizing transformations provide another method that can be used to find asymptotically pivotal quantities. For the binomial distribution, this is Problem 7.5.11 in DeGroot and Schervish. The function

$$g(p) = 2\sin^{-1}(\sqrt{p})$$

is variance-stabilizing for the Bernoulli distribution (Problem 5.13.29 in DeGroot and Schervish), meaning

$$\sqrt{n}\big[g(\hat{p}_n) - g(p)\big] \xrightarrow{\mathcal{D}} \text{Normal}(0,1) \tag{2.21}$$

so the right hand side is an asymptotically pivotal quantity and can be used to make an asymptotic (large-sample, approximate) confidence interval.

The inverse function is given by

$$g^{-1}(\theta) = \sin^2(\theta/2)$$

If we write $\theta = g(p)$ and $\hat{\theta}_n = g(\hat{p}_n)$, then plugging these into (2.21) gives

$$\sqrt{n}\big(\hat{\theta}_n - \theta\big) \xrightarrow{\mathcal{D}} \text{Normal}(0,1)$$

Thus, if we wanted to estimate the parameter θ, the confidence interval would be

$$\hat{\theta}_n \pm \frac{c}{\sqrt{n}}$$

To estimate the original parameter p, we transform back to the original scale using $p = g^{-1}(\theta)$, that is, the confidence interval is

$$g^{-1}\left(\hat{\theta}_n \pm \frac{c}{\sqrt{n}}\right) = g^{-1}\left(g(\hat{p}_n) \pm \frac{c}{\sqrt{n}}\right) \tag{2.22}$$

For numerical example, consider binomial data with $n = 30$ and $x = 6$, and suppose we want a 95% confidence interval so $c = 1.96$. Then our three intervals are

Type	Interval
(2.20)	$(0.057, 0.343)$
(2.19)	$(0.095, 0.373)$
(2.22)	$(0.079, 0.359)$

Thus we see there is no single right way to do an asymptotic confidence interval. This is the nature of approximations. There are many approximations that are close to correct. Since none is exactly correct, there is no reason to say that one is without doubt better than the rest.

2.10 Sample Variance Again

The most important moment after the mean is the variance. If we have *exactly normal* data, then the sampling distribution of the sample variance is gamma (Theorem 7.3.1 in DeGroot and Schervish). But what if the data aren't normal? Then we don't have a brand name distribution for the sampling distribution of the sample variance. But we can use the asymptotic method.

As when we were discussing consistency of sample central moments, there is a problem that the terms in (1.2b) are not independent because all contain $\overline{X}_n$. Thus we first investigate the asymptotics of (2.9), with $k = 2$, because we now doing second moments. The central limit theorem applied to $\tilde{\mu}_{2,n}$ gives

$$\sqrt{n}\big(\tilde{\mu}_{2,n} - \mu_2\big) \xrightarrow{\mathcal{D}} \text{Normal}(0, \tau^2)$$

where

$$\begin{aligned}
\tau^2 &= \operatorname{Var}\{(X_i - \mu)^k\} \\
&= E\{(X_i - \mu)^{2k}\} - E\{(X_i - \mu)^k\}^2 \\
&= \mu_4 - \mu_2^2
\end{aligned}$$

Of course the second central moment is variance, $\mu_2 = \sigma^2$, but we use μ_2 where it makes the formulas make more sense.

We are just about done. We have actually found the asymptotic distribution of the sample variance, be we need several applications of Slutsky's theorem to see that. First

$$\begin{aligned}
\frac{n-1}{n} S_n^2 &= \frac{1}{n} \sum_{i=1}^n (X_i - \overline{X}_n)^2 \\
&= \frac{1}{n} \sum_{i=1}^n \big[(X_i - \mu)^2 - 2(X_i - \mu)(\overline{X}_n - \mu) + (\overline{X}_n - \mu)^2\big] \\
&= \tilde{\mu}_{2,n} + (\overline{X}_n - \mu)^2
\end{aligned}$$

So

$$\sqrt{n}\big(S_n^2 - \sigma^2\big) = \frac{n}{n-1} \cdot \sqrt{n}\big(\tilde{\mu}_{2,n} - \mu_2\big) + \frac{1}{n-1}\mu_2 + \frac{n}{n-1} \cdot \sqrt{n}(\overline{X}_n - \mu)^2 \quad (2.23)$$

The CLT says

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{\mathcal{D}} Y$$

where $Y \sim \text{Normal}(0, \sigma^2)$. Then by an argument similar to that in Section 2.6,

$$\frac{n}{n-1} \cdot \sqrt{n}(\overline{X}_n - \mu)^2 = \frac{\sqrt{n}}{n-1} \cdot \left[\sqrt{n}(\overline{X}_n - \mu)\right]^2 \xrightarrow{\mathcal{D}} 0 \cdot Y^2 = 0$$

Hence the third term on the right hand side in (2.23) converges in probability to zero. The second term on the right is nonrandom and converges to zero, hence it converges in probability to zero when considered a sequence of constant random variables. The factor $n/(n-1)$ in the first term on the right is asymptotically negligible. Hence by another application of Slutsky's theorem S_n^2 has the same asymptotic distribution as $\tilde{\mu}_{2,n}$, that is,

$$\sqrt{n}\big(S_n^2 - \sigma^2\big) \xrightarrow{\mathcal{D}} \text{Normal}(0, \mu_4 - \sigma^4).$$

When the *data are normally distributed*, then $\mu_4 = 3\sigma^4$ and we get

$$S_n^2 \approx \text{Normal}\left(\sigma^2, \frac{2\sigma^4}{n}\right)$$

But for *non-normal data*, we must keep the general form

$$S_n^2 \approx \text{Normal}\left(\sigma^2, \frac{\mu_4 - \sigma^4}{n}\right).$$