Fundamental Inference Concepts, Chapter 6

Elizabeth Purdom

This document has last been compiled on Sep 05, 2024.

Contents

1 Statistical Models

Recall, we have a sample of data $X = (X_1, \ldots, X_n)$ from an unknown distribution F (F can refer to the distribution or the CDF). Statistical inference will use the sample X to make conclusions about F.

To make this concrete, we will generate a **statistical model**. In generality, we say that a statistical model $\mathcal{F}$ is a set of possible distributions, and we assume that one of the distributions in $\mathcal{F}$ actually generated the data X (i.e. $F \in \mathcal{F}$).

This sounds like we are listing a discrete set $\mathcal{F}$ of possible data generating distributions. However, that's not (generally) the case. Usually $\mathcal{F}$ describes a continum of distributions. The true F is a single point on the continuum, which we are unlikely to guess exactly, but we hope to be close to the true F.

Website Example In our website example, we said we might have records of visits to the website and want to determine the actual rate λ. Here we have

- $X =$ the records of visits
- $\mathcal{F} = \{F : F_\lambda = Poisson(\lambda|t), \lambda > 0\}$. Notice how this is actually a continuum of distributions, where the continuum corresponds to differing λ

Types of statistical procedures Then the types of statistical inference procedures can be described as

- choosing the best guess for F amongst those models (**estimation**)
- describing how confident we are in that guess (and which other models in $\mathcal{F}$ might be also plausible) – these will take the form of **confidence sets or credible regions** showing a range within the continuum $\mathcal{F}$ that are plausible distributions for having generating X.
- We might have a particular default theory for how the data was generated, $F_0 \in \mathcal{F}$, called a *null distribution.* We might want to test whether F_0 might have generated our data X (**hypothesis testing**). F_0 is often a simple, not very interesting mechanism, and we would like to show that F_0 is insufficient to explain the data.

These three types of procedures form the core of statistical inference. All or some of these tasks can be done on the same data – it often goes hand in hand. For example, in the website data, we might want to *estimate* our best guess of F_λ, but we might also want to give a range of other λ that are plausible.

In this class we will try to understand not just how to do these core tasks (for which there are often many possible options), but how to evaluate which options for these tasks are likely to perform well. And, indeed, defining what "perform well" means is part of this class.

1.1 Parametric and Non-parametric models

There are two main types of statistical models to keep in mind.

Parametric models refer to a set $\mathcal{F}$ that can be described by a finite number of parameters.

For example:

- $X_i \sim N(\mu, \sigma^2), \quad i = 1, \ldots, n$
  $\mathcal{F}$ can be described by the 2 parameters (μ, σ^2)
- Website example: 1 parameter (λ) describes $\mathcal{F}$

We generally use θ to indicate an arbitrary vector of parameters. We also use θ as a subscript for calculations, e.g.

$$P_\theta(X \in A),$$

or

$$E_\theta(X)$$

which emphasizes that these are values that depend on θ. This is an important notation to be comfortable with; these say that the result of these calculations are functions of θ.

Nonparametric models Nonparametric models require an infinite number of parameters to describe $\mathcal{F}$. These models are sometimes called "distribution free" to indicate that we make few restrictions on the family of distributions.

For example,

- Most obvious: $\mathcal{F} = \{\text{all possible distributions}\}$!
- For estimating a density, you would restrict yourself to distributions with a smooth density

$$\mathcal{F} = \{\text{all distributions with a density } f\text{, and } f' \text{ is defined}\}$$

1.2 Parameters of interest

Often we might be interested in a particular aspect(s) of F, such as the mean, variance, quantiles, etc. Such aspects are called **parameters** of F. A parameter is simply any function of the distribution F, and can be univariate (like mean, median, variance, IQR, ...) but can also be multivariate. We will often use the symbol θ to represent our parameter of interest.

Focusing on θ simplifies our problem – rather than figuring out the entire distribution of F, we only need to focus inference on θ. In this case, rather than refering to our set of possible models $\mathcal{F}$, we will often refer to the set of possible parameters θ, often given by Θ.

Notice that simplifying the problem to estimating parameters does *not* require that we have to rely on parametric models. For example, suppose θ is the mean of the distribution F,

$$\theta = E(F).$$

We could assume either a parametric or non parametric model

- Parametric example:

  $$X_i \sim N(\mu, \sigma^2), \quad i = 1, \ldots, n$$

  In this case estimating θ would be equivalent to estimating μ, the parameters of our distribution F.

  But it's not required that θ correspond to a parameter of the distribution. For example, if instead

  $$X_i \sim Gamma(\alpha, \beta), \quad i = 1, \ldots, n,$$

  where α and β are the shape and rate parameters, then the mean is $\theta = \alpha/\beta$.

- Nonparametric example: all distributions with finite mean,

  $$\mathcal{F} = \{F : E(F) < \infty\}.$$

Note that in both cases, knowing just θ does not tell us F, the unknown distribution.

2 (Point) Estimation

Estimation is the process of selecting the best F among our models $\mathcal{F}$ based on the data we observe (X), or the best θ among our possible Θ values if we are focused on a parameter of interest. So our **estimate** is going to be some function of our data.

The term **statistic** refers to any function of the data X. So our estimate will be a statistic (of course when we do confidence intervals or hypothesis testing later, these will also rely on our data, i.e. be functionals of our data, so they will also use statistics.)

Terminology If we are interested in only a particular parameter θ, then we will be estimating just θ and not the entire distribution F. In this case we will refer to our estimation as a **point estimator**. The function of the data that estimates it will be

$$\hat{\theta}(x_1, \dots, x_n),$$

i.e. a function of the n data points.

We call $\hat{\theta}(X_1, \dots, X_n)$ (the r.v.) an *estimator*, while we call $\hat{\theta}(x_1, \dots, x_n)$ (the realization) an *estimate.*

We use $\hat{\theta}_n$ or $\hat{\theta}$ for both.

Note that θ is just a generic parameter, but in particular instances we might be interested in parameters and use other greek letters in the same way. For example, for our example of $\mathcal{F}$ be limited to $N(\mu, \sigma^2)$, if we want to estimate the mean, we would be interested in estimating μ and our estimate of μ would be called $\hat{\mu}$.

Sampling Distribution Any statistic is of course a random variable, since it is a function of our random data. We call the probability distribution of a statistic its **sampling distribution** to emphasize that it is dependent on the distribution of the sample of data.

Warning! Be careful not to confuse the distribution of X with the sampling distribution of $\hat{\theta}_n$. For example

$$X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$$

implies a sampling distribution of

$$\hat{\theta}_n = \bar{X}_n \sim N(\mu, \sigma^2/n)$$

The sampling distribution is *derived* from the data distribution.

The *standard error* of $\hat{\theta}$ refers to the standard deviation of the sampling distribution of $\hat{\theta}$ and can be written as

$$se(\hat{\theta}_n).$$

This terminology is to help distinguish the standard deviation of $\hat{\theta}$ from the standard deviation of an individual data point (or the data distribution) which is generally quite different. For example, in the above example, each data point has standard deviation σ, while the estimate of the mean has standard error $\sigma/\sqrt{n}$.

The standard deviation of both the sampling distribution and the data distribution are unknown parameters.

An important part of frequentist theory are the many ways to evaluate and compare estimators, which we'll discuss more formally when we come to decision theory. For now, we will consider some basics of point estimators, and we will revisit these properties more carefully when we return to a deep dive into parametric models.

2.1 Finite Properties of Estimators

Some properties we would consider of estimators are what we call *finite properties* of the estimator. They are exactly true for any size n.

- Bias:
$$bias(\hat{\theta}_n) = E_\theta[\hat{\theta}_n] - \theta$$
We say $\hat{\theta}_n$ is **unbiased** if its bias is zero, $E_\theta[\hat{\theta}_n] = \theta$
- Standard error:
$$se(\hat{\theta}_n) = \sqrt{var_\theta(\hat{\theta}_n)}$$
If we are comparing variance of two estimators, we will generally consider their ratio call this their *relative efficiency*,
$$eff(\hat{\theta}_1, \hat{\theta}_2) = \frac{var(\hat{\theta}_1)}{var(\hat{\theta}_2)}.$$
- Mean squared error:
$$MSE(\hat{\theta}_n) \quad = \quad E_\theta[(\hat{\theta}_n - \theta)^2]$$

Example Let $X_1, \ldots, X_n \overset{iid}{\sim} Poisson(\lambda)$ and let $\hat{\lambda}_n = \bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$.

Find the bias, standard error, and MSE of this estimator.

More about MSE Notice the difference between variance and MSE:

$$MSE(\hat{\theta}_n) = E_\theta[(\hat{\theta}_n - \theta)^2]$$
$$var(\hat{\theta}_n) = E_\theta[(\hat{\theta}_n - E(\hat{\theta}_n))^2]$$

We can write

$$MSE = bias^2(\hat{\theta}_n) + V_\theta[\hat{\theta}_n]$$

showing that MSE takes into account both kinds of performance of an estimator.

3 Confidence Sets

Point estimation is usually the first step, but a single value doesn't give any idea the precision of this "guess", and in particular the range of other reasonable guesses. The most common method for giving more context is a confidence interval.

A $1-\alpha$ **confidence interval** for θ is an interval $C_n = C_n(X)$ computed from the data such that $P_\theta(\theta \in C_n) \geq 1-\alpha$ for all θ.

- $1-\alpha$ is called the coverage of the interval.
- For confidence intervals, θ is considered fixed. It is C_n that is random. To emphasize this, we could write $P(C_n \ni \theta) \geq 1-\alpha$ for all θ.
- Common values for α are 0.01 or 0.05

Assuming θ is univariate, we need two bounds that are functions of the data so that:

$$1-\alpha = P(L_\alpha(X) \leq \theta \leq U_\alpha(X))$$

Why will the bounds depend on α?

How do we find CI for θ? Generally we find these from considering the distribution of our estimator of θ, so that we have

$$1-\alpha = P(L_\alpha(\hat{\theta}) \leq \theta \leq U_\alpha(\hat{\theta}))$$

Example: Normal data distribution We can illustrate this with an example you have probably seen before. Suppose we have that $X_i \overset{i.i.d}{\sim} N(\mu, 1)$ and estimate μ with $\hat{\mu} = \bar{X}$. Then $se(\hat{\mu}) = 1/\sqrt{n}$.

In this case, we have that $\hat{\mu} \sim N(\mu, 1/n)$, which is another way of saying that

$$\frac{\hat{\mu} - \mu}{1/\sqrt{n}} = \frac{\hat{\mu} - \mu}{se(\hat{\mu})} \sim N(0, 1)$$

This gives us a function of μ and $\hat{\mu}$ that we know the distribution of, and we can backtrack from there

$$\begin{aligned}
1 - \alpha &= P(-z_{\alpha/2} \leq \frac{\hat{\mu} - \mu}{1/\sqrt{n}} \leq z_{\alpha/2}) \\
&= P(-z_{\alpha/2}/\sqrt{n} \leq \hat{\mu}_n - \mu \leq z_{\alpha/2}/\sqrt{n}) \\
&= P(-z_{\alpha/2}/\sqrt{n} - \hat{\mu}_n \leq -\mu \leq z_{\alpha/2}/\sqrt{n} - \hat{\mu}_n) \\
&= P(z_{\alpha/2}/\sqrt{n} + \hat{\mu}_n \geq \mu \geq -z_{\alpha/2}/\sqrt{n} + \hat{\mu}_n) \\
&= P(\hat{\mu}_n - 1.96/\sqrt{n} \leq \mu \leq \hat{\mu}_n + 1.96/\sqrt{n})
\end{aligned}$$

where $z_{\alpha/2}$ is chosen such that $P(Z > z_{\alpha/2}) = \alpha/2$ for $Z \sim N(0, 1)$.

This gives us a CI

$$C_n = (\hat{\mu}_n - 1.96/\sqrt{n}, \hat{\mu}_n + 1.96/\sqrt{n}) = (\bar{X} - 1.96/\sqrt{n}, \bar{X} + 1.96/\sqrt{n})$$

for μ. Notice how this is a random inverval, since it depends on our random statistic $\bar{X}$.

Of course I could have replaced $X_i \sim N(\mu, 1)$ with $X_i \sim N(\mu, \sigma)$; this would have made $se(\hat{\mu}) = \sigma/\sqrt{n}$. So we would simply replaced $1/n$ with σ^2/n so long as σ^2 is a fixed known variance,

$$(\hat{\mu}_n - 1.96\frac{\sigma}{\sqrt{n}}, \hat{\mu}_n + 1.96\frac{\sigma}{\sqrt{n}}).$$

Estimating σ Notice that it is not a valid CI if it contains unknown parameters, so we could not use the above if we do not know σ^2. And we do not usually actually know the variance σ^2.

Usually we have an estimate of σ^2, like our standard estimate

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2.$$

Then we can estimate $se(\hat{\mu}) = \sigma/\sqrt{n}$ with

$$\hat{se}(\hat{\mu}) = \hat{\sigma}/\sqrt{n}.$$

But we can't actually say that

$$\frac{\bar{X}-\mu}{\hat{se}(\bar{X})} \sim N(0,1),$$

so the above CI calculation isn't exactly correct anymore.

However, we do know that

$$\frac{\bar{X}-\mu}{\hat{se}(\bar{X})} \overset{D}{\approx} N(0,1)$$

which means the same probability statements above are approximately correct. So we can use the above CI calculations, replacing σ with $\hat{\sigma}$, to get an *approximate* confidence interval,

$$(\hat{\mu}_n - 1.96\frac{\hat{\sigma}}{\sqrt{n}}, \hat{\mu}_n + 1.96\frac{\hat{\sigma}}{\sqrt{n}}).$$

We will see that this is a common exercise: to find an approximately normal distribution for our estimator, and be able to calculate confidence intervals.

4 Asymptotic Analysis of Estimators

There are many times where we cannot calculate finite properties of estimators, particularly for more complicated problems, but we can say something about them when there is a large sample size. These are stated as properties of estimators in the limit – i.e. as $n \to \infty$.

4.1 The general idea of convergence

To talk about these properties, we need to consider what do we mean about "asymptotics". In probability, this is dealt with by considering a random sequence of variables W_n, i.e.

$$W_1, W_2, W_3, \ldots$$

A concrete example would be to let $W_n = \sqrt{n}\bar{X}_n$, where $\bar{X}_n$ is the mean of n i.i.d random variables $X_i \sim F$, with F having mean 0 and standard deviation 1.

The point of creating such a sequence is that the distribution of the random variable W_n can be described by its index n, and generally as $n \to \infty$ that distribution becomes more stable or predictable, such that the distribution of the random variable can be easily described by something that doesn't even depend on n anymore.[1] The above example W_n is an example of a sequence of random variables where we can say many things about W_n as n gets large: we know that W_n will be close to 0, the mean of F and that its distribution will get close to a normal distribution.

This ideas are all loosely described as "convergence", but there are actually several different ways of discussing what happens as $n \to \infty$.

In statistics, the sequence of random variables that is of interest is practically always our estimates as our data sample gets larger, so that

$$W_n = \hat{\theta}(X_1, \ldots, X_n),$$

but the probabilistic definition can be used for sequences defined more arbitrarily.

4.2 Consistent estimators

First let's define an important notion of convergence of a sequence of random variables: convergence in probability

Definition (Convergence in Probability)**.** A sequence of random variables W_n converges in probability to W if for every ϵ,

$$\lim_{n\to\infty} P(|W_n - W| \geq \epsilon) = 0.$$

We write $W_n \xrightarrow{P} W$.

Example: Weak Law of Large Numbers If we have X_i i.i.d with the same mean $E(X_i) = \mu < \infty$, then

$$\bar{X}_n \xrightarrow{P} \mu$$

We will use this definition with $W_n = \hat{\theta}$ and $W = \theta$, i.e. a constant (not random!). When an estimator $\hat{\theta}_n$ converges in probability to θ, we call it a consistent estimator:

Definition (Consistency)**.** An estimator $\hat{\theta}_n$ is consistent for θ if

$$\hat{\theta}_n \xrightarrow{P} \theta$$

for all choice of $\theta \in \Theta$.[2]

[1] Sometimes this requires that we carefully define W_n to get rid of the dependence on n. For example, instead of defining $W_n = \bar{X}_n$, sometimes we find it convenient to multiply $\bar{X}_n$ by $\sqrt{n}$ or subtract off the mean, particularly for distribution convergence.

[2] This is called *weak* consistency. Strong consistency would be replacing this definition with almost sure convergence, another notion of convergence we will not discuss.

Consistency tells us that if n is sufficiently large, the probability $\hat{\theta}_n$ is far from θ is arbitrarily small.

4.2.1 Relationship of Consistency to Convergence in moments

Consistency can seem like a strange choice compared to our finite properties. Why not look at quantities like

$$\lim bias_\theta \hat{\theta}_n \text{ or } \lim var_\theta \hat{\theta}_n \text{ ?}$$

One reason is that dealing with convergence of moments is messy, because you can run into all kinds of technicalities for complicated estimators: you can concoct examples with errant amount of mass in the tails that makes working with moments, in particular, problematic (in addition to the fact that moments might be undefined).[3]

Consistency is easier to work with mathematically, and as a concept embraces some of the same goals as MSE in the sense that it ensures that the probability that $\hat{\theta}_n$ is far from θ is small.

However, while consistency in some ways "takes the place" of MSE in asymptotics, consistency is not equivalent to having the variance and bias approach zero.

We have the following relationships:

- If the MSE tends to zero then the estimator is consistent.

  This is by Markov's inequality,

  $$P_\theta(|\hat{\theta}_n - \theta| \geq \epsilon) = P_\theta(|\hat{\theta}_n - \theta|^2 \geq \epsilon^2) \leq \frac{E_\theta[(\hat{\theta}_n - \theta)^2]}{\epsilon^2} = \frac{MSE(\hat{\theta}_n)}{\epsilon^2}.$$

- This further means that if

  $$\lim var_\theta \hat{\theta}_n = 0 \text{ and } \lim bias_\theta \hat{\theta}_n = 0$$

[3] Note in fact that the term "asymptotically unbiased" actually is sometimes defined as something different than the bias approaching zero. See e.g. Lehmann, *Theory of Point Estimate, 2nd Edition*, p. 438 where it is defined as an estimator θ_n such that

$$r_n(\hat{\theta}_n - \theta) \overset{D}{\to} H$$

for some sequence r_n and distribution H which has expectation zero. Focusing on this definition allows the author to avoid the issue of convergent moments. This author makes the same distinction between "asymptotic variance" versus the "limit of the variance". In common place discussion, we often think of these interchangeable and (like my introductory lecture) do not overly distinguish between these.And I'm not sure that this is a universal definition. But regardless, this points to the fact that an author who is trying to be precise, as this book is, overcomes this difficulty with a different definition to distinguish the two concepts.

then $\hat{\theta}_n$ is consistent (Casella and Berger, p 469).

- However, consistency of an estimator does NOT mean that the bias goes to zero

  We can see this from the following example: let $\hat{\theta}_n$ take the value θ with probability $(n-1)/n$ and take the value n with probability $1/n$. Then $\hat{\theta}_n$ is a consistent sequence of estimators for θ:

  $$P(|\hat{\theta}_n - \theta| > \epsilon) = \frac{I(|n-\theta| > \epsilon)}{n} \to 0$$

  but the bias is not approaching 0:

  $$E(\hat{\theta}_n) = \theta\frac{n-1}{n} + n\frac{1}{n} = \frac{(\theta+1)n - \theta}{n} \to \theta + 1$$

  Counter examples for these types of situations generally involve a distribution that has a small amount of probability mass very far in the tails where the probability of such a value decreases with n. Sometimes these counter examples are fairly artificial, but the point is to create a simple counter-example. The heart of the counter-examples is that there continues to be the chance of wildly extreme values of the estimator, and generally are situations where the variance of the estimator is not tending to zero.

- If *in addition* to consistency, we assume the variance of $\hat{\theta}$ is uniformly bounded,

  $$var(\hat{\theta}_n) < C \quad \forall n$$

  then this implies the bias of $\hat{\theta}$ is approaching 0 as well.

4.3 Asymptotically Normal Estimators

Another important asymptotic property of an estimator is it's asymptotic sampling distribution.

In particular, we often will see that for large sample sizes, the sampling distribution of many reasonable estimators $\hat{\theta}_n$ is roughly a normal distribution. Which normal distribution should it converge to (i.e. what are its parameters)?

- **Mean** Well, clearly we would want the mean to be θ. Why?

- **Variance** The variance should match the variance of $\hat{\theta}_n$, i.e. $se(\hat{\theta}_n)^2$.

So when we say an estimator is $\hat{\theta}_n$ is asymptotically normal, we mean

$$\hat{\theta}_n \overset{D}{\approx} N(\theta, se(\hat{\theta}_n)^2)$$

I write $\overset{D}{\approx}$ to emphasize that this approximation is refering to its sampling distribution. (This is only an intuitive notation; we will give more precise definition later.)

4.3.1 Confidence Intervals from Asy. Normal Estimators

One example of why we would want to know the asymptotic distribution of an estimator would be to construct a confidence interval for θ.

The basic strategy above for normal confidence intervals can be extended to statistics that are asymptotically normal, even though the data is not. Suppose $\hat{\theta}_n$ is asymptotically normal. Then we can use the above rationale to have a confidence interval for θ based on $\hat{\theta}_n$, i.e.

$$\frac{\hat{\theta}_n - \theta}{se(\hat{\theta}_n)} \overset{D}{\approx} N(0,1)$$

so that we can develop a confidence interval

$$(\hat{\theta}_n - 1.96 se(\hat{\theta}_n), \hat{\theta}_n + 1.96 se(\hat{\theta}_n)).$$

Unfortunately, this is generally not a valid confidence interval, since $se(\hat{\theta}_n)$ is usually unknown (and is a function of the unknown θ).

However, we usually can estimate $se(\hat{\theta}_n)$. If have an estimate $\hat{se}(\hat{\theta}_n)$ of $se(\hat{\theta}_n)$ so that

$$\hat{\theta}_n \overset{D}{\approx} N(\theta, \hat{se}(\hat{\theta}_n)^2)$$

then we have

$$\frac{\hat{\theta}_n - \theta}{\hat{se}(\hat{\theta}_n)} \overset{D}{\approx} N(0,1).$$

Then we can form an approximate $1-\alpha$ confidence interval for θ of

$$C_n = \hat{\theta}_n \pm z_{\alpha/2} \hat{se}_n.$$

For $\alpha = 0.05, z_{\alpha/2} = 1.96$, so this gives us the commonly seen confidence interval of roughly ± 2 standard deviations.

Warning! Notice that $\hat{se}_n$ is an estimate of $se(\hat{\theta}_n)$, not the standard deviation of the individual X_i.

4.3.2 Practical Usage

Utility Notice this is a powerful reason why asymptotic normality is useful. If we can show an estimator is asymptotically normal (and *many* estimators are!), then we have an out-of-the box confidence interval at our disposal for large sample sizes.

This is far easier than working with difficult finite distributions to figure out confidence intervals; and there are many very complicated estimators whose finite distribution we cannot describe *at all*, yet we can describe their asymptotic distribution much more easily.

We can of course work with other asymptotic distributions if our estimator is not asymptotically normal, but we will see that many standard approaches to developing reasonable estimators result in asymptotically normal estimators.

How do you know these properties? Powerful theorems, like the Central Limit Theorem (CLT) and other similar theorems provide asymptotic normality for many estimators. Specifically, the CLT tells us that if X_i are i.i.d with mean μ and variance σ^2, then

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \overset{D}{\approx} N(0, 1).$$

When we discuss standard estimators like the MLE, we will see that these are asymptotically normal. So many estimators are asymptotically normal.

Consistent estimators of $se(\hat{\theta})$ can be more specific (and as we said $\hat{\theta}$ sometimes will not have a closed form solution, much less $se(\hat{\theta})$) but we will see similar theorems that allow for estimating $se(\hat{\theta})$ for classes of estimators, like MLEs.

What will we do in this class In this class, you will not be expected to prove these types of big theorems; they are beyond the expectations of the class. You will usually be told the theorems (like CLT) that allow you to figure out if an estimator is asymptotically normal. But you will need to put the results together.

You might be expected to construct a reasonable estimator $\hat{\theta}_n$ that is asymptotically normal, meaning

$$\hat{\theta}_n \overset{D}{\approx} N(\theta, se(\hat{\theta}_n)^2).$$

And you might need to then figure out what is an estimate $\hat{se}(\hat{\theta}_n)$ that is an estimate of the standard error.

Then you might need to go further to use other results that allow you to say

$$\hat{\theta}_n \overset{D}{\approx} N(\theta, \hat{se}(\hat{\theta}_n)^2)?$$

We will discuss what conditions for this are (Section 4.6.3).

Example (Wasserman) Let $X_1, \ldots, X_n \overset{iid}{\sim} Bernoulli(p)$. A reasonable estimate of p is the proportion of $X_i = 1$, $\hat{p}$. We have that

$$\hat{p}_n = \bar{X}_n,$$

is an unbiased estimator of p and

$$[se(\hat{p}_n)]^2 = var(\hat{p}_n) = \frac{p(1-p)}{n}$$

We know by the CLT that $\bar{X}_n$ is asymptotically normal, so $\hat{p}_n \overset{D}{\approx} N(p, \frac{p(1-p)}{n})$.

An estimator of $se(\hat{p}_n)$ is

$$\hat{se}_n = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

We will show (Section 4.6.3) that this estimator satisfies

$$\hat{p}_n \overset{D}{\approx} N(p, [\hat{se}_n]^2)$$

Putting this together, we have 95% CI

$$C_n = \hat{p}_n \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Exercise 1. Let $X_1, \ldots, X_n \overset{iid}{\sim} Poisson(\lambda)$. Show how we can use this sample to construct an approximate 95% confidence interval for λ (see exercises at end of Section 4.6.3 as well).

4.4 Definition of Asymptotic Normality

We have written

$$\hat{\theta}_n \overset{D}{\approx} N(\theta, se(\theta_n)^2).$$

But what does does that actually mean? First let's start with a definition from probability:

Definition (Convergence in Distribution)**.** Let F_n be the distribution of a random variable Z_n, and Z a RV with distribution F. Then the sequence of random variables Z_n converges in distribution to Z if

$$\lim_{n\to\infty} F_n(t) = F(t)$$

for all t for which $F(t)$ is continuous.

We write $Z_n \Rightarrow Z$ or $Z_n \overset{D}{\to} Z$. We will often write this as $Z_n \Rightarrow F$, as abbreviation for saying Z_n converges in distribution to a random variable Z with distribution F.

We will use this definition to define more precisely what we mean by an estimator being asymptotically normal.[4]

Definition (Asymptotically Normal)**.** We say that $\hat{\theta}_n$ is **asymptotically normal** if

$$Z_n = \frac{\hat{\theta}_n - \theta}{se(\hat{\theta}_n)} \Rightarrow N(0,1)$$

i.e. for sufficiently large n, the distribution of Z_n is close to that of $N(0,1)$.

Example: Central Limit Theorem (CLT) If we have X_i i.i.d with mean μ and variance σ^2, define $Z_n = \sqrt{n}(\bar{X} - \mu)$. Then the CLT tells us that

$$Z_n \Rightarrow Z$$

where $Z \sim N(0, \sigma^2)$.

Thus the CLT tells us that $\hat{\mu}$ is an asymptotically normal estimator of μ.

4.5 Relationship of Consistency and Asymptotic Normality

These asymptotic properties are two different concepts. In particular,

- **Consistency doesn't imply Asy. Normality** Just because an estimator is getting close to the truth, doesn't tell you *anything* about what the distribution of the estimator is for large n.

[4] Note that we are only considering estimators that have finite variance. You could consider a more general estimator.

- **Does Asy. Normality imply Consistency?** In general, no.

  However, typically, a "good" estimator would have $se(\hat{\theta}_n) \xrightarrow{P} 0$. Why?

  So if we *also* know that $se(\hat{\theta}_n) \xrightarrow{P} 0$, then $\hat{\theta}_n$ being asymptotically normal implies that $\hat{\theta}_n$ is a consistent estimator.

  This makes sense. If $\hat{\theta}$ is asymptotically normal, i.e. we have

  $$Z_n = \frac{\hat{\theta}_n - \theta}{se(\hat{\theta}_n)} \Rightarrow N(0,1)$$

  and $se(\hat{\theta}_n) \xrightarrow{P} 0$, then clearly $\hat{\theta}_n - \theta$ must be getting small so that Z_n converges to a $N(0,1)$

  But if we don't make any assumptions about how $se(\hat{\theta}_n)$ behaves for large n, $\hat{\theta}_n$ may not be getting closer to θ with larger sample size, and so we do not have consistency.

Counter-Example For example, suppose $X_1, \ldots, X_n \overset{i.i.d}{\sim} N(\theta, 1)$, and $\hat{\theta}_n = X_1$, i.e. you always just use the first data point collected as the estimate no matter how much you collect. Yes, it's a idiotic estimator, but it is unbiased ($E\hat{\theta}_n = \theta$). Then $se(\hat{\theta}_n) = 1$ and

$$\frac{\hat{\theta}_n - \theta}{se(\hat{\theta}_n)} = X_1 - \theta \sim N(0,1)$$

(no asymptotics needed). So clearly $\hat{\theta}_n$ is asymptotically normal, but isn't consistent for θ.

$$P(|\hat{\theta} - \theta| > \epsilon) = \Phi^{-1}(1 - \epsilon)$$

i.e. the probability that a single normal is more than ϵ from its mean is a fixed quantity, not getting smaller with n.

Greater Generality We can even generalize the definition of asymptotic normality further and replace $1/se(\hat{\theta}_n)$ in the definition of asymptotic normality with a sequence of positive numbers r_n. In other words, assume only that

$$r_n(\hat{\theta}_n - \theta) \xrightarrow{D} N(0,1)$$

In this case, if you also have $r_n \to \infty$ then you can show that this implies that $\hat{\theta}_n$ is consistent for θ (using Slutsky's Theorem).[5]

[5] This is true even if you replace r_n with the positive random variables $R_n \xrightarrow{P} \infty$, i.e. for all $\epsilon > 0$

$$P(R_n > \epsilon) \to 1 \quad n \to \infty.$$

4.5.1 Other variations

Assume $\hat{\theta}_n$ is asymptotically normal. Here are other related results would follow depending on what you know about $se(\hat{\theta}_n)$.

- If $se(\hat{\theta}_n) = \sqrt{v(\theta)/n}$, then
$$\frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sqrt{v(\theta)}} \Rightarrow N(0,1)$$
or equivalently
$$\sqrt{n}(\hat{\theta}_n - \theta) \Rightarrow N(0, v(\theta))$$
This is simply a straight substitution of $\sqrt{v(\theta)/n}$ for $se(\hat{\theta}_n)$.

  Moreover, this a common property of the variance of many estimators, particularly of i.i.d. data – that their variance relies on n only through a constant divided by $\sqrt{n}$.

- If $\sqrt{n}se(\hat{\theta}_n) \xrightarrow{P} \sqrt{v(\theta)}$, then
$$\frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sqrt{v(\theta)}} \Rightarrow N(0,1)$$
or equivalently
$$\sqrt{n}(\hat{\theta}_n - \theta) \Rightarrow N(0, v(\theta))$$
This is an asymptotic version of the above property, since more informally our condition is that $se(\hat{\theta}_n) \approx \sqrt{v(\theta)/n}$ for large sample sizes.[6]

Notice that both of these assumptions about $se(\hat{\theta})$ also implies that $se(\hat{\theta}_n) \xrightarrow{P} 0$, which we've seen also implies consistency of the estimator.

- If you have an estimate $\hat{se}$ of $se(\hat{\theta}_n)$ so that
$$\frac{se}{\hat{se}} \xrightarrow{P} 1,$$

[6] Notice that we are not using the convergence in probability of $se(\hat{\theta}_n)$ but rather the convergence in probability of $\sqrt{n}se$. As we've said, we can't make statements like $se \xrightarrow{P} \sqrt{v(\theta)/n}$. It doesn't make sense to have n on the right hand side of a convergence statement. Why don't we use the convergence in probability of $se(\hat{\theta}_n)$? As we've seen, most estimators will have $se \xrightarrow{P} 0$, so this property doesn't help us in working with standard error. The convergence in probability of $\sqrt{n}$ times an estimator to a constant is another kind of convergence that is similar to what is often called $\sqrt{n}$-consistency ("root n consistency"); $\sqrt{n}$-consistency is weaker since it doesn't require convergence in probability, but just bounded in probability. $\sqrt{n}$-consistency is the kind of convergence we need to be able to say $se \approx \sqrt{v(\theta)/n}$.

then

$$Z_n = \frac{\hat{\theta}_n - \theta}{\hat{se}} \Rightarrow N(0,1).$$

This is the precise meaning of our earlier claim that if $\hat{se}$ a good estimator of se, you can replace se with $\hat{se}$ (e.g. for asymptotic confidence intervals).

- If $\sqrt{n} se(\hat{\theta}_n) \xrightarrow{P} \sqrt{v(\theta)}$, and $\hat{v}$ is a consistent estimator of $v(\theta)$ then

$$\frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\hat{v}} \Rightarrow N(0,1)$$

This is the same principle as the previous result.

4.6 Tools for working with convergences

If you are wanting more tools to understand when you make some of the various substitutions and manipulations described above, there are various probability rules that can be helpful for working with asymptotic properties. You won't be explicitly asked to work with these rules, but it can help to be aware of these various substitutions.

4.6.1 The Continuous Mapping Theorem

The continuous mapping theorem allows us to carry over asymptotic properties of a random variable to *continuous functions* of the random variable.

Theorem (Continuous Mapping Theorem)**.** *Assume that g is a deterministic continuous function.*[7]

$$\textit{If } W_n \xrightarrow{P} W \textit{ then } g(W_n) \xrightarrow{P} g(W)$$
$$\textit{If } W_n \Rightarrow W \textit{ then } g(W_n) \Rightarrow g(W).$$

Note that this theorem also applies to random vectors and multivariate functions g.

Applying this to statistics, if $\hat{\theta}_n$ is a consistent estimator of θ and g is a continuous function, then $g(\hat{\theta}_n)$ is a consistent estimator of $g(\theta)$.

Simple Example 1: If $\hat{v}$ is a consistent estimator of $var(\hat{\theta})$, then $\sqrt{\hat{v}}$ is a consistent estimator of $se(\hat{\theta})$.

[7] Or more precisely is continuous on all regions of non-zero probability

Simple Example 2 Suppose

$$\hat{\theta}_n \Rightarrow W$$

where $W \sim N(0, s(\theta)^2)$. Then if we define the function

$$g(x) = x/s(\theta)$$

we have that

$$g(\hat{\theta}_n) = \frac{\hat{\theta}_n}{s(\theta)} \Rightarrow g(W),$$

where $g(W) = \frac{W}{s(\theta)}$. Of course if $W \sim N(0, s(\theta)^2)$ then $g(W) = \frac{W}{s(\theta)} \sim N(0, 1)$.

We would normally write all of this in short hand as

$$\hat{\theta}_n \Rightarrow N(0, s(\theta)^2)$$

implies that

$$\frac{\hat{\theta}_n}{s(\theta)} \Rightarrow N(0, 1),$$

as we have done above.

Note that g has to be deterministic. If you start including other random variables in your definition of the function g, then this wouldn't apply. For example, suppose I was interested in

$$\frac{\hat{\theta}_n}{\hat{s}(\theta)}$$

where $\hat{s}(\theta)$ is an estimate of $s(\theta)$. I couldn't define

$$g(x) = \frac{x}{\hat{s}(\theta)}$$

because $\hat{s}(\theta)$ is a random variable. So I need more information about the behavior of $\hat{s}(\theta)$ before I can make this transition. Slutsky's Theorem (below) deals with this situation and requires additional conditions on $\hat{s}(\theta)$.[8]

Multiple Estimators The following are some useful properties when working with estimators of different parameters, which are a result of the continuous mapping theorem:[9]

Lemma. *Suppose that the estimator $\hat{\theta}_n$ converges in probability to the parameter θ and the estimator $\hat{\beta}_n$ converges in probability to the parameter β. Then*

[8] the proof of Slutsky's theorem is simple and uses the continuous mapping theorem, but it only works because there are additional conditions on $\hat{s}(\theta)$.

[9] Let W_n be the random vector $(\hat{\theta}_n, \hat{\beta}_n)$ and then set, for example, $g(x, y) = xy$ or $g(x, y) = x + y$

- *$\hat{\theta}_n + \hat{\beta}_n$ converges in probability to $\theta + \beta$.*
- *$\hat{\theta}_n \hat{\beta}_n$ converges in probability to $\theta\beta$.*
- *If $\beta \neq 0$, $\hat{\theta}_n / \hat{\beta}_n$ converges in probability to θ/β.*

Note a simple result of this is that you can multiple or add constants in logical ways to a consistent estimator (if $\hat{\theta}_n$ converges in probability to θ then $c\hat{\theta}_n \xrightarrow{P} c\theta$).

4.6.2 Slutsky's Theorem

The following theorem is used a lot when we want to replace a parameter with its estimate[10]

Theorem (Slutsky's Theorem)**.** *Let X_n, Y_n be sequences of random variables. If $X_n \Rightarrow$ converges in distribution to X and $Y_n \xrightarrow{P} c$, c a constant, then*

$$X_n + Y_n \Rightarrow X + c$$
$$Y_n X_n \Rightarrow cX$$
$$X_n / Y_n \Rightarrow X/c$$

Using these tools, we can show the results given in Section 4.5.1.

Example If $Y_n = se/\hat{se} \xrightarrow{P} 1$, and

$$Z_n = \frac{\hat{\theta} - \theta}{se} \Rightarrow N(0,1)$$

then Slutsky's tells us that

$$Y_n Z_n = \frac{\hat{\theta} - \theta}{\hat{se}} \Rightarrow 1 \cdot N(0,1).$$

This would also work if we had two different estimators $\hat{se}_1$ and $\hat{se}_2$, and we only knew that

$$\frac{\hat{\theta} - \theta}{\hat{se}_1} \Rightarrow N(0,1)$$

but we knew that

$$\frac{\hat{se}_1}{\hat{se}_2} \xrightarrow{P} 1$$

[10] This theorem is actually a direct application of the continuous mapping theorem, combined with the fact that if X_n converges in distribution to X and Y_n converges in probability to a constant c, then the joint vector (X_n, Y_n) converges in distribution to (X, c). Then you can set $g(x, y) = x + y$ or $g(x, y) = x/c$.

Exercise 2. Show our previous result: if $\hat{\theta}$ is asymptotically normal and $\sqrt{n}se_n \overset{P}{\to} \sqrt{v(\theta)}$ then

$$\frac{\sqrt{n}(\hat{\theta}_n - \theta)}{v(\theta)} \Rightarrow N(0,1)$$

Exercise 3. Show our previous result: if $\hat{\theta}$ is asymptotically normal, $\sqrt{n}se_n \overset{P}{\to} \sqrt{v(\theta)}$, and $\hat{v} \overset{P}{\to} v(\theta)$ then

$$\frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\hat{v}} \Rightarrow N(0,1)$$

4.6.3 Common consistent estimators

Theorem. *Assume $X_1, \ldots, X_n$ are i.i.d with mean μ and variance σ^2. Then*

- *Mean*

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

  is a consistent estimator of μ (and unbiased)
- *$\bar{X}_n$ is an asymptotically normal estimator*
- *Variance*

  $$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

  is a consistent estimator of σ^2

Exercise 4. Use the above properties to show that if $X_1, \ldots, X_n$ are i.i.d with mean μ and variance σ^2, then

- s is a consistent estimator of σ.
- $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is a consistent estimator of σ^2
- $$\frac{(\bar{X}_n - \mu)}{s/\sqrt{n}} \Rightarrow N(0,1)$$

Now let's formalize our previous claims used for for making asymptotic confidence intervals.

Exercise 5. Let $X_1, \ldots, X_n \overset{iid}{\sim} Bernoulli(p)$, and $\hat{p}_n$ be the proportion of X_i that are equal to 1. Show that

$$\frac{\hat{p}_n - p}{\sqrt{\hat{p}_n(1-\hat{p}_n)/n}} \Rightarrow N(0,1)$$

Exercise 6. Let $X_1, \ldots, X_n \overset{iid}{\sim} Poisson(\lambda)$ and $\hat{\lambda}_n = \bar{X}_n$. Show that

$$\frac{\hat{\lambda}_n - \lambda}{\sqrt{\bar{X}_n/n}} \Rightarrow N(0,1)$$