

1 Estimation, Unbiasedness and Asymptotic Theory

Last Updated: February 18, 2009

Suppose that X has some distribution $f(x; \theta)$ where θ is an unknown parameter. For example, we may have that $f(x; \theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ where $\theta = (\mu, \sigma^2)$. The question that we want to focus on is how we can gain some idea of what θ is given observations from the distribution of X . Suppose that we observe an independently and identically distributed or *i.i.d.* sample from $f(x; \theta)$ which we will call (x_1, \dots, x_n) . How can we estimate θ given observation of (x_1, \dots, x_n) ?

Now, suppose that we are interested in the mean and the variance of the distribution:

$$\mu = E[X] = \int x dF(x)$$

and

$$\sigma^2 = E[(X - \mu)^2] = \int (x - \mu)^2 dF(x).$$

One of the most common estimation methods is to use the **analogy principle** or to replace the above population moments with their sample analogues:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

and

$$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2.$$

We now want to investigate the properties of our estimators.

1.1 Unbiasedness

Suppose that there is an unknown parameter θ that we want to estimate and suppose we have an *i.i.d* sample, (x_1, \dots, x_n) , from some distribution $f(x; \theta)$. Let $\hat{\theta} = h(x_1, \dots, x_n)$ be some function of the data that we are using to estimate θ . We call $\hat{\theta}$ a **statistic**. We say that $\hat{\theta}$ is an **unbiased estimator** of θ if $E[\hat{\theta}] = \theta$.

Example 1 Consider the sample mean \bar{x} . This is an unbiased estimator because

$$E[\bar{x}] = E\left[\frac{1}{n} \sum_{i=1}^n x_i\right] = \frac{1}{n} * n * E[x_i] = \mu.$$

Note that we used the fact that the x_i 's have identical distributions, but did not use independence.

Example 2 Consider the sample variance s^2 . This is a biased estimator because

$$E[s^2] = E\left[\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2\right] = E[x_i^2] - E[\bar{x}^2] = E[x_i^2] - \frac{1}{n}\sigma^2 - \mu^2 = \left(\frac{n-1}{n}\right)\sigma^2.$$

However, $\tilde{s}^2 = \left(\frac{n}{n-1}\right)s^2$ is an unbiased estimator of σ^2 . This is why we often use $\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ to estimate the variance of a distribution. Also, note that we used the fact that the sample is distributed both identically and independently.

1.2 Convergence in Probability

Let $\{X_n\}$ be a sequence of random variables and let X be a random variable. We say that the sequence $\{X_n\}$ **converges in probability** to X if for all $\varepsilon > 0$ we have that

$$P(|X_n - X| \geq \varepsilon) \rightarrow 0$$

or, equivalently, that

$$P(|X_n - X| < \varepsilon) \rightarrow 1.$$

If X_n converges in probability to X then we say that

$$X_n \xrightarrow{p} X.$$

Now, let's consider the sample mean once again, \bar{x} , as defined above. A very powerful result is that the sample mean will converge in probability to the population mean or that

$$\bar{x} \xrightarrow{p} \mu.$$

This result is called the **Weak Law of Large Numbers**. To prove this, recall Chebyshev's Inequality which says that

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

Consequently, we will have that

$$P(|\bar{X} - \mu| \geq \varepsilon) = P\left(|\bar{X} - \mu| \geq \frac{\varepsilon\sqrt{n}}{\sigma} * \frac{\sigma}{\sqrt{n}}\right) \leq \frac{\sigma^2}{\varepsilon^2 n} \rightarrow 0.$$

Another important result is that if $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$ then $X_n + Y_n \xrightarrow{p} X + Y$. To see

why this statement is true, note the following:

$$\begin{aligned} P(|X_n + Y_n - (X + Y)| \geq \varepsilon) &\leq P(|X_n - X| + |Y_n - Y| \geq \varepsilon) \\ &\leq P\left(|X_n - X| \geq \frac{\varepsilon}{2}\right) + P\left(|Y_n - Y| \geq \frac{\varepsilon}{2}\right) \rightarrow 0. \end{aligned}$$

Note that we used the triangle inequality for the first inequality which states that

$$|X + Y| \leq |X| + |Y|.$$

Another important result which is trivial to prove is that for any constant a , we will have that $aX_n \xrightarrow{p} aX$ whenever $X_n \xrightarrow{p} X$.

A result that we will get a lot of mileage out of is that for a continuous function $g(\cdot)$, we will have that $g(X_n) \xrightarrow{p} g(X)$ whenever $X_n \xrightarrow{p} X$. We will prove this result for the special case in which X is a constant a . The more general case can be found in a more advanced text. We call this result the **continuous mapping theorem**.

First, we choose $\varepsilon > 0$. By continuity, we can produce some $\delta > 0$ such that

$$|x - a| < \delta \Rightarrow |g(x) - g(a)| < \varepsilon.$$

Then, we will have that

$$P(|g(X_n) - g(a)| \geq \varepsilon) \leq P(|X_n - a| \geq \delta) \rightarrow 0.$$

Consequently, this theorem implies that if $X_n \xrightarrow{p} a$ then we will have that $X_n^2 \xrightarrow{p} a^2$ and $\frac{1}{X_n} \xrightarrow{p} \frac{1}{a}$ (provided that $a \neq 0$).

Another useful result is that if $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$ then $X_n Y_n \xrightarrow{p} XY$. To see this, note that

$$X_n Y_n = \frac{1}{2} X_n^2 + \frac{1}{2} Y_n^2 - \frac{1}{2} (X_n - Y_n)^2 \xrightarrow{p} \frac{1}{2} X^2 + \frac{1}{2} Y^2 - \frac{1}{2} (X - Y)^2 = XY.$$

Note that we employed the continuous mapping theorem in the calculation above.

1.3 Convergence in Distribution

The notion of convergence in probability is useful in that it tells us that as the sample size grows a statistic “gets close” to a population parameter. However, it does not tell us how close. To determine this, we will need to work with a new concept: convergence in distribution. This concept will be necessary when we discuss the Central Limit Theorem.

Let $\{X_n\}$ be a sequence of random variables and let X be a random variable. Let the respective CDF’s be given by F_n and F . We say that X_n **converges in distribution** to X if

$$\lim_{n \rightarrow \infty} F_n(x) = F(x)$$

for all points of continuity of $F(x)$. If X_n converges in distribution to X , we say that $X_n \xrightarrow{d} X$.

It is important to bear in mind that convergence in distribution does not imply convergence in probability. To see this, let X be a continuous random variable with a symmetric PDF around

0 so that $f(x) = f(-x)$. If we define the sequence,

$$X_n = \begin{cases} X & \text{for } n \text{ odd} \\ -X & \text{for } n \text{ even} \end{cases}$$

then we will have that $F_n(z) = P(X_n \leq z) = P(X \leq z)$ for n odd and

$$F_n(z) = P(X_n \leq z) = P(X \geq -z) = F(z)$$

where the last step followed by symmetry. However, it should be obvious that X_n does not converge in probability to X .

One useful result that I will not prove is that if $X_n \sim t_n$ then $X_n \xrightarrow{d} N(0, 1)$. If you recall what you know of basic statistics, this is why we can use the z table rather than the t table when the sample size is sufficiently large. So, really most of you have already been using asymptotic approximations of finite sample distributions while not knowing it. Also, it turns out that the sample size usually only needs to be larger than 30 observations which really is not that large. This is a testament to the power of these asymptotic results.

Example 3 Let (X_1, \dots, X_n) be a random sample from a uniform distribution on the interval $[0, \theta]$. Define $Y_n = \max \langle X_1, \dots, X_n \rangle$ and let $Z_n = n(\theta - Y_n)$. Then, we will have that

$$P(Z_n \leq t) = P(n(\theta - Y_n) \leq t) = P\left(Y_n \geq \theta - \frac{t}{n}\right) = 1 - \left(1 - \frac{t}{\theta n}\right)^n \rightarrow 1 - e^{-t/\theta}.$$

Now, we were going to introduce a new concept. Consider a sequence $\{a_n\}$ and define

$$b_n = \sup \{a_n, a_{n+1}, \dots\}$$

and

$$c_n = \inf \{a_n, a_{n+1}, \dots\}.$$

Clearly, the sequence $\{b_n\}$ is non-increasing and $\{c_n\}$ is non-decreasing. Next, we are going to define

$$\lim b_n = \limsup a_n$$

and

$$\lim c_n = \liminf a_n.$$

The limit of a sequence may not always exist, but the \limsup and \liminf of the sequences will always exist. Clearly, if the sequence converges, then we will have that \limsup and \liminf are the same. To illustrate, these concepts consider the following example.

Example 4 *Define*

$$x_n = (-1)^n + \frac{1}{n}.$$

This sequence does not converge, but we do have that

$$\limsup x_n = 1$$

and that

$$\liminf x_n = -1.$$

We will now use these concepts in the proof of the following theorem.

Theorem 5 If $X_n \xrightarrow{p} X$ then we will have that $X_n \xrightarrow{d} X$

Proof. First, note that

$$\begin{aligned} F_{X_n}(x) &= P(X_n \leq x) = P(\{X_n \leq x\} \cap \{|X_n - X| < \varepsilon\}) + P(\{X_n \leq x\} \cap \{|X_n - X| \geq \varepsilon\}) \\ &\leq P(X \leq x + \varepsilon) + P(|X_n - X| \geq \varepsilon). \end{aligned}$$

Therefore, we will have that

$$\limsup F_{X_n}(x) \leq F_X(x + \varepsilon).$$

Similar reasoning gives us that

$$\liminf F_{X_n}(x) \geq F_X(x - \varepsilon).$$

Thus, we will have the following string of inequalities

$$F_X(x - \varepsilon) \leq \liminf F_{X_n}(x) \leq \limsup F_{X_n}(x) \leq F_X(x + \varepsilon)$$

Letting ε go to zero, we will have that

$$\lim F_{X_n}(x) = F(x)$$

and the result is proven. ■

Although, it is not true in general that convergence in distribution implies convergence in probability, it is true that if a random variable converges in distribution to a constant then this is equivalent to saying that this random variable converges in probability to a constant.

Theorem 6 If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} 0$ then we will have that $X_n + Y_n \xrightarrow{d} X$.

Proof. We will bypass the formalities of \limsup and \liminf throughout this proof. Let $\varepsilon > 0$ and choose z so that $F_X(z)$ is continuous at z . First, note that

$$\begin{aligned} \lim P(X_n + Y_n \leq z) &= \lim P(X_n + Y_n \leq z \cap \{|Y_n| < \varepsilon\}) + \lim P(X_n + Y_n \leq z \cap \{|Y_n| \geq \varepsilon\}) \\ &= \lim P(X_n + Y_n \leq z \cap \{|Y_n| < \varepsilon\}) \\ &\leq \lim P(X_n \leq z + \varepsilon) = F_X(z + \varepsilon). \end{aligned}$$

This calculation taken together with a similar calculation gives us that

$$F_X(z - \varepsilon) \leq \lim P(X_n + Y_n \leq z) \leq F_X(z + \varepsilon)$$

and if we let ε tend to zero, the result obtains. ■

Here are some other important results that we will not prove. First, if $g(\cdot)$ is a continuous function, then we will have that $g(X_n) \xrightarrow{d} g(X)$ whenever $X_n \xrightarrow{d} X$. Second, if $X_n \xrightarrow{d} X$, $A_n \xrightarrow{p} a$ and $B_n \xrightarrow{p} b$, then we will have that $A_n + B_n X_n \xrightarrow{d} a + bX$. This result is known as the **Slutsky Theorem**.

Now, let's assume that $g(\cdot)$ is a differentiable function and that

$$\sqrt{N}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \sigma^2).$$

Then, we will have that

$$g(\hat{\theta}) = g(\theta) + g'(\theta) (\hat{\theta} - \theta) + 0.5g''(\tilde{\theta}) (\hat{\theta} - \theta)^2$$

for $\tilde{\theta}$ between θ and $\hat{\theta}$. Rearranging of terms gives us that

$$\sqrt{N} (g(\hat{\theta}) - g(\theta)) = g'(\theta) \sqrt{N} (\hat{\theta} - \theta) + \underbrace{0.5 g''(\tilde{\theta}) \sqrt{N} (\hat{\theta} - \theta)^2}_{\xrightarrow{p_0} 0} \xrightarrow{d} N(0, g'(\theta) \sigma^2 g'(\theta)).$$

This result is known as the **delta method**.

1.4 Central Limit Theorem

We have shown that if we observe a random sample (x_1, \dots, x_n) where $X_i \sim N(\mu, \sigma^2)$ then

$$\frac{\sqrt{n}(\bar{x} - \mu)}{\sigma} \sim N(0, 1).$$

This result allows us to use the sample mean to make a statement about the population mean.

However, this result requires each observation of X_i to have an exact Normal distribution. To

many students, this may appear to be a strong condition. As it turns out, even if the sample

does have an exact Normal distribution, the above condition will hold approximately. This

powerful result is known as the **Central Limit Theorem**.

Theorem 7 Suppose that we observe an i.i.d. sample (x_1, \dots, x_n) such that $E[X_i] = \mu$ and $V(X_i) = \sigma^2$. Then, we will have that

$$Y_n \equiv \frac{\sqrt{n}(\bar{x} - \mu)}{\sigma} \xrightarrow{d} N(0, 1)$$

Proof. Define $M(t) \equiv E[\exp(tX)]$ for $-h < t < h$. Then we will have that

$$E[\exp(t(X - \mu))] = \exp(-t\mu) M(t) \equiv m(t).$$

Note that

$$m(0) = 1$$

$$m'(0) = 0$$

$$m''(0) = E[(X_i - \mu)^2] = \sigma^2.$$

Next, we will take an expansion of $m(t)$ around zero. Doing this, we obtain

$$m(t) = m(0) + m'(0)t + \frac{m''(\xi)t^2}{2}$$

where ξ is between zero and t . Thus, we will have that

$$m(t) = 1 + \frac{m''(\xi)t^2}{2} = 1 + \frac{\sigma^2 t^2}{2} + \left[\frac{(m''(\xi) - \sigma^2)t^2}{2} \right]$$

The next step is to calculate the MGF of Y_n :

$$M_{Y_n}(t) = E \left[\exp \left(t * \frac{\sum x_i - n\mu}{\sqrt{n}\sigma} \right) \right] = E \left[\exp \left(t * \frac{x - \mu}{\sqrt{n}\sigma} \right) \right]^n = m \left(\frac{t}{\sqrt{n}\sigma} \right)^n$$

But we know that

$$m \left(\frac{t}{\sqrt{n}\sigma} \right) = 1 + \frac{t^2}{2n} + \left[\frac{(m''(\xi) - \sigma^2)t^2}{2n\sigma^2} \right]$$

and, thus, we can conclude that

$$M_{Y_n}(t) = \left(1 + \frac{t^2}{2n} + \psi(n) \right)^n$$

where $\psi(n) \equiv \frac{(m''(\xi) - \sigma^2)t^2}{2n\sigma^2}$. Note that ξ is now between zero and $\frac{t}{\sqrt{n}\sigma}$ and, thus, we will have that $\xi \rightarrow 0$ as $n \rightarrow \infty$. This then gives us that $\psi(n)$ will go to zero as the sample size approaches infinity. Accordingly, we will have

$$M_{Y_n}(t) \rightarrow \exp \left(\frac{t^2}{2} \right)$$

which is the MGF for a Standard Normal random variable. ■

We conclude with a simple example. Let $X_i \sim \text{binomial}(1, p)$. These are called Bernoulli Trials. Suppose that we observe a random sample of size n . Then, we will have that

$$\mu = p$$

and

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

Define

$$Y_n = X_1 + \dots + X_n.$$

Then, we will have that

$$\frac{\sqrt{n}(\hat{p} - p)}{\sqrt{p(1 - p)}} \xrightarrow{d} N(0, 1)$$

where $\hat{p} \equiv \frac{Y_n}{n}$. So, we will have that

$$\sqrt{n}(\hat{p} - p) \overset{A}{\sim} N(0, p(1 - p))$$

where $\overset{A}{\sim}$ means approximately distributed. The delta-method then gives us that

$$\sqrt{n}(\hat{p}(1 - \hat{p}) - p(1 - p)) \overset{A}{\sim} N(0, (1 - 2p)^2 p(1 - p)).$$