

1 The central problem in *statistics: estimation*

erm: (draft) 11/01/2010

Typically

1. We desire to study a population which has density $f_X(x; \theta)$ where the form of $f_X(x; \theta)$ is known but θ is unknown (MGB 226). We want to estimate θ .

This statement describes much of the econometrics you will ever do.

2. We take a random sample from $f_X(x; \theta)$ of size n , x_1, x_2, \dots, x_n
3. We then assume some function $T = t(X_1, X_2, \dots, X_n)$ is an estimator for some element of θ ; call that element θ_k

(Note that 1–3 assumes knowledge of $f_X(x; \theta)$ by either fact or fiction. One can also talk about estimation and estimators when there is no knowledge of $f_X(x; \theta)$, but that is not what we doing now. Or we can talk about estimation where there is knowledge of $f_X(x; \theta)$ but the sample is not random.)

Given 1–3, the issue is whether $t(X_1, X_2, \dots, X_n)$ is a good estimator for θ_k

definition: A function of X_1, X_2, \dots, X_n is called *a statistic* (a *statisticator*)

That is, a statistic is a function of the observed data (a function of the observed values of the n random variables in the sample) and a function of nothing else.¹

$t(X_1, X_2, \dots, X_n)$ is what we mean by a statistic

Note that *statistics* is just the plural of *statistic*, so *statistics* is the study of functions of observed values of random variables, or, said another way, statistics is the study of functions of the data

¹It is a function of only observables. If a variable is a function of the sample and population parameters it is not a statistic.

For example, if one takes a random sample of size n from $f_X(x; \theta)$, the following are all statistics:

- X_1 (the first X drawn)
- the smallest (or largest) X drawn
- $(e^{3X_1} + e^{6X_2} e^{1X_4} + e^{3X_{17}})$
- $\frac{1}{n} \sum_{i=1}^n X_i$

Each of these statistics might or might not be a good *estimator* of some element of θ

Each estimator has a sampling distribution; that is, each of the above examples is a rv that will vary across random samples, have *sampling variability*.

Consider some *estimator*

$$T = t(X_1, X_2, \dots, X_n)$$

If we use $t = t(x_1, x_2, \dots, x_n)$ to denote an estimate of θ , we say that $T = t(X_1, X_2, \dots, X_n)$ is an estimator of θ and $t(x_1, x_2, \dots, x_n)$ is an estimate of θ .

1.1

1.2 Estimating the population mean

(a quick review of stuff we already know)

One population that we often want to estimate is the population mean. Let μ_x represent the population mean such that

$$f_X(x; \mu_x, \sigma_x^2)$$

We want an estimator for μ_x .²

The sample mean, from a random sample drawn from $f_X(x; \mu_x, \sigma_x^2)$, is an estimate of μ_x . That is,

$$\frac{1}{n} \sum_{i=1}^n X_i = t(X_1, X_2, \dots, X_n)$$

is an estimator of μ_x and

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

is an estimate of μ_x from the specific sample (x_1, x_2, \dots, x_n) .

Let

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

Notationally \bar{X} indicates the estimator, and \bar{x} is the mean from a specific sample.

As an alternative estimators of μ_x consider $\min(X_1, X_2, \dots, X_n)$ and X_3 . These are also both statistics and estimators for μ_x .

In a general sense, every statistic from a sample is an estimator for each of the population parameters, maybe a bad estimator, but an estimator never-the-less.

I now propose three candidates for estimators for μ_x

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

²Given my assumption that $f_X(x; \mu_x, \sigma_x^2)$, I have implicitly assumed that the density has two parameters and those two parameters are the mean and variance - the parameters of the distribution and the moments of the distribution are one and the same. Note that this is not always the case. One, for example, might want to estimate the mean of the distribution where the mean is not a parameter, or one might want to estimate the parameters of a distribution although the parameter is not the mean or variance of the distribution.

$$\min(X_1, X_2, \dots, X_n)$$

and

$$X_3$$

Do they have any desirable properties? If the sample is a random sample,³

$$\begin{aligned} E[\bar{X}] &= E\left[\frac{1}{n} \sum_{i=1}^n X_i\right] \\ &= \frac{1}{n} E\left[\sum_{i=1}^n X_i\right] \\ &= \frac{1}{n} (E[X_1] + E[X_2] + \dots + E[X_n]) \\ &= \frac{1}{n} (\mu_x + \mu_x + \dots + \mu_x) \\ &= \frac{1n\mu_x}{n} \\ &= \mu_x \end{aligned}$$

That is, $E[\bar{X}] = \mu_x$, which seems like a nice property for \bar{X} to have?

Does X_3 have this property? Yes

$$E[X_3] = \mu_x$$

How about $\min(X_1, X_2, \dots, X_n)$? No. Can you prove it?

³If you need to, review expectations.

Consider another estimator for μ_x

$$s(X_1, X_2) = .5X_1 + .25X_2$$

$$\begin{aligned} E[s(X_1, X_2)] &= E[.5X_1 + .25X_2] \\ &= .5E[X_1] + .25E[X_2] \\ &= .5\mu_x + .25\mu_x \\ &= .75\mu_x \\ &< \mu_x \end{aligned}$$

$s(X_1, X_2)$ systematically underestimated μ_x .

definition: $\{t(X_1, X_2, \dots, X_n)$ is an unbiased estimator of $\theta_k \Leftrightarrow E[t(X_1, X_2, \dots, X_n)] = \theta_k$

1.3 Estimating σ_x^2 , the population variance

Consider the following two statistics as estimators for σ_x^2 , where the sample is random:

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

where $\bar{X} \equiv \frac{1}{n} \sum_{i=1}^n X_i$ and

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

If I had to choose between these two estimators, at first blink, I would go for the first one because it is the average of the squared deviations. Remember that σ_x^2 is the expectation of the squared deviations in the population, $E[(X - E[X])^2]$. \tilde{s}_x^2 is called the method of moment estimator of σ_x^2 .

Note that

$$\lim_{n \rightarrow \infty} \tilde{s}_x^2 = \lim_{n \rightarrow \infty} s_x^2$$

Consider the expectation of each:

However, before we look at the expected values, the following algebra will prove useful.

$$\begin{aligned} \sum_{i=1}^n (x_i - \mu_x)^2 &= \sum_{i=1}^n (x_i - \bar{x} + \bar{x} - \mu_x)^2 \\ &= \sum_{i=1}^n [(x_i - \bar{x}) + (\bar{x} - \mu_x)]^2 \\ &= \sum_{i=1}^n [(x_i - \bar{x})^2 + (\bar{x} - \mu_x)^2 + 2(x_i - \bar{x})(\bar{x} - \mu_x)] \\ &= \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu_x)^2 + 2(\bar{x} - \mu_x) \sum_{i=1}^n (x_i - \bar{x}) \end{aligned}$$

but since

$$\sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n x_i - n\bar{x} = n\bar{x} - n\bar{x} = 0$$

the last expression in the expansion of $\sum_{i=1}^n (x_i - \mu_x)^2$ is zero and

$$\sum_{i=1}^n (x_i - \mu_x)^2 = \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu_x)^2$$

Solve this for $\sum_{i=1}^n (x_i - \bar{x})$ to obtain

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

This bit of algebra will prove useful. We will use it in our derivation of $E[s_x^2]$

$$\begin{aligned} E[s_x^2] &= E\left[\frac{1}{(n-1)} \sum_{i=1}^n (X_i - \bar{X})^2\right] \\ &= \frac{1}{(n-1)} E\left[\sum_{i=1}^n (X_i - \bar{X})^2\right] \end{aligned}$$

Substituting, the algebraic relationship we just derived

$$\begin{aligned} E[s_x^2] &= \frac{1}{(n-1)} E\left[\sum_{i=1}^n (X_i - \mu_x)^2 - n(\bar{X} - \mu_x)^2\right] \\ &= \frac{1}{(n-1)} \sum_{i=1}^n E[(X_i - \mu_x)^2] - nE[(\bar{X} - \mu_x)^2] \\ &= \frac{1}{(n-1)} \left\{ \sum_{i=1}^n \sigma_x^2 - n\text{var}[\bar{X}] \right\} \\ &= \frac{1}{(n-1)} [n\sigma_x^2 - n\text{var}[\bar{X}]] \\ &= \frac{1}{(n-1)} \left[n\sigma_x^2 - n\left[\frac{\sigma_x^2}{n}\right] \right] \\ &= \frac{1}{(n-1)} [n\sigma_x^2 - \sigma_x^2] \\ &= \frac{(n-1)\sigma_x^2}{(n-1)} = \sigma_x^2 \end{aligned}$$

What did we just show?

$$E[s_x^2] = \sigma_x^2$$

That is, s_x^2 is an unbiased estimate of σ_x^2 .

Therefore \tilde{s}_x^2 is a biased estimate of σ_x^2 . Note that the degree of bias in \tilde{s}_x^2 decreases as n increases.

That is why we prefer s_x^2 , over \tilde{s}_x^2 , as an estimator for σ_x^2

What is the intuition? If one has a sample of n observation, once \bar{X} is determined there are only $(n-1)$ independent $(X_i - \bar{X})^2$. That is, if one knows \bar{X} and $X_1, X_2, \dots, X_{n-1}, X_n$ is completely determined.