

Random Sampling, Statistics, and Estimators

statistics.tex Feb. 11, 2003

In general, we assume the rv of interest, x , has some known distribution, $f_X(x)$ in the population of interest, but we do not know the values of the parameters in that distribution.

- That is, we assume that in the population of interest, X has some **known** distribution
- This is a very strong assumption.
- Given that we know, by assumption, the distribution, the problem is to estimate the values of the parameters.

We do this by taking a sample from the population of interest, and use information in the sample to estimate the values of the population parameters.

Hereafter, I will use the term *population* to refer to the *population of interest*. Estimation always starts by defining the population of interest.

If one wants to emphasize the parameters in $f_X(x)$, one might write it $f_X(x; \theta)$, where θ is the vector of parameters.

Samples and random samples

We would like our sample to be a random sample from $f_X(x)$.

Definition: Define a sample of size n as (x_1, x_2, \dots, x_n) where x_j is the j th observation in the sample (MGB 223)

Note that a sample is a vector of random variables with some joint distribution.

Definition: The sample (x_1, x_2, \dots, x_n) is a random sample from $f_X(x)$ if

$$f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) = f_X(x_1)f_X(x_2)\dots f_X(x_n)$$

where $f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n)$ is the joint distribution of the sample (MGB 223 and G74).

In explanation, each variable in $f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n)$ is a random variable; that is, observation j can take different values, so observation j is a rv. Denote this random variable X_j , and the specific value it takes x_j .

$f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n)$ is therefore a joint density function for the n random variables in the sample.

Said in words, $f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n)$ is a random sample from $f_X(x)$ if each observation is an independent draw from $f_X(x)$.

Just to be clear, let me write out the above in a little more detail. The sample (x_1, x_2, \dots, x_n) is a random sample from $f_X(x)$ if

$$\begin{aligned} f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) &= f_{X_1}(x_1) f_{X_2}(x_2) \dots f_{X_n}(x_n) \\ &= f_X(x_1) f_X(x_2) \dots f_X(x_n) \\ &= f(x_1) f(x_2) \dots f(x_n) \end{aligned}$$

because

$$f_{X_i}(x_i) = f_X(x_i)$$

that is, each observation in the sample has the same distribution.

Often we say (G74) a sample is random if the observations in it are independent, identically distribution - I.I.A.

That is, a sample is random if each observation in the sample is independently drawn from the same (identical) distribution.

Said loosely, the sample is random if for each observation, each value of the rv in the population has an equal chance of appearing as the j th observation, and this is true for all j .

Give me an example of a nonrandom sample. Start by identifying the population you are sampling from.

Can one tell, by observation, whether a sample is a random sample?

Note that *random* does not mean *representative*.

However as n increases, the sample will likely become more representative of the population

Because of sampling variation, samples differ. That is, any two random samples of size n from the same population are likely to not exhibit the same values of the n random variables X_1, X_2, \dots, X_n .

Remember that the n random variables X_1, X_2, \dots, X_n have some joint density function

$$f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n)$$

We call this the *distribution of the samples*. Each sample is a draw from this distribution. Picture a sample with two observations; that is $n = 2$.

The central problem in *statistics*

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).
This statement describes most 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 estimate of some element of θ , call that element θ_k

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*

That is, a statistic is just a function of the observed data (a function of the observed values of the n random variables in the sample).

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

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

Consider some *estimator*

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

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

Estimating the population mean

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 .

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

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

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 have proposed three candidates for estimators for μ_x

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

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

and

$$X_3$$

Do they have any desirable properties?

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

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?

Consider another estimator for μ_x

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

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

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

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

Estimating σ_x^2 , the population variance

Consider the following two statistics as estimators for σ_x^2 :

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

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 one had to choose between these two estimators, at first blink, I might 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, $E[(X - E[X])^2]$. 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 do this, consider the following algebra, which will turn out to be 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

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

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

Solve this for $\sum_{i=1}^n (x_i - \bar{x})^2$ 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 will prove useful. We will use 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 σ_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.