Sampling distributions and Estimation

Suppose we have a population about which we want to know some characteristic, e.g. height, income, voting intentions. If it is a large population, it may be difficult to look at every individual. We therefore take a sample. For example, if we want to know the average height of the population, we might sample 100 people and take the average height of the sample. But how good an estimate will this be? Is it likely to be biased upwards or downwards from the population average? How inaccurate is it likely to be? How big a sample do we need to be confident of getting a figure close to the true figure? These are the sorts of questions answered in this and subsequent sections.

We discuss sampling with and without replacement, and from finite and infinite populations. We shall start by assuming that samples are randomly selected from the population, and that they are large, e.g. 30 observations or larger. Small samples will be dealt with later.

Each type of sampling leads to a different sampling distribution. The sampling distribution of a parameter, such as sample mean or sample proportion is either a theoretical distribution, like the normal, or is obtained by taking many samples of the same size from a population and constructing a frequency distribution.

Distribution of the sample mean

Let us assume that we take a random sample from an infinite population or from a finite population with replacement. If a random sample of size n is taken from a population with mean μ and variance σ^2, the sample mean is given by

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + ... + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Where X_1 is the value of the first observation, etc. Before each observation is chosen, it could take any value in the population. In other words, $X_1 \ldots X_n$ are random variables having the same distribution as the population. Hence $\overline{X}$, as a linear combination of random variables, is also a random variable. Moreover, as the population is infinite or, if finite, we sample with replacement, the observations are independent. Thus we can easily find the expected value of $\overline{X}$:

$$E(\bar{X}) = E\left\{\frac{1}{n}(X_1 + ... + X_n)\right\} = \frac{1}{n}E(X_1 + ... + X_n)$$

$$= \frac{1}{n}\left[E(X_1) + E(X_2) + ... + E(X_n)\right] = \frac{1}{n}(\mu + \mu + ... + \mu)$$

$$= \frac{1}{n}n\mu = \mu$$

Using results from the previous section. Thus, on average, the sample mean will be equal to the population mean. Later we shall see that this means that the sample mean is an unbiased estimator of the population mean.

We may also find the variance of the sample mean:

$$Var(\bar{X}) = Var\left\{\frac{1}{n}(X_1 + ...X_n)\right\} = \frac{1}{n^2}Var(X_1 + ... + X_n)$$

$$= \frac{1}{n^2}[Var(X_1) + ... + Var(X_n)] = \frac{1}{n^2}(\sigma^2 + ... + \sigma^2)$$

$$= \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$$

Hence, the variance of the sample mean declines as n increases. We call the resulting distribution the sampling distribution of $\bar{X}$, and the standard deviation of this distribution ($\frac{\sigma}{\sqrt{n}}$) is called the standard error of the sampling distribution of $\bar{X}$. When n is large, the standard error will be very small, so that there will be hardly any sampling error involved in using $\bar{X}$ as an estimate for the population mean. Note however that, because of the square root sign, if we want to halve the standard error, we must quadruple the sample size.

In general, this result means that, if σ is known, we can choose n to achieve any desired degree of accuracy of our estimate $\bar{X}$ for the population mean. (That is, any desired standard error).

If the distribution of X in the population is normal, then the distribution of $\bar{X}$ will also be normal, as it is a linear combination of normal variables. What is more, even if X is not normally distributed, there is an extremely powerful theorem called the Central Limit Theorem (CLT) that says that provided the sample size is sufficiently large (say $n \geqslant 30$), that

$\bar{X}$ will nonetheless have an approximately normal distribution – the higher the sample size, the closer the distribution of $\bar{X}$ to a normal distribution.

Distribution of the sample variance

What if we don't know either the mean or the standard deviation of a variable X in a population, and wish to try to estimate it by looking at the sample variance and standard deviation?

Let the variable X be distributed with E(X)=μ and Var(X)=σ^2 (which are both unknown). We take a sample of size n, $X_1,\ldots,X_n$. So these are independent r.v.s, with the same distribution as X.

The sample variance is equal to $S^2=\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n}$ - that is, the average squared deviation from the mean. Let us call this quantity S^2.

It can be shown* that $E(S^2) = \frac{n-1}{n}\sigma^2$ - in other words, the sample variance is on average smaller than the population variance – it is a biased estimate. Intuitively, this is because the first observation gives us no idea of the variance – we need at least two to see a difference between them. Hence we 'lose an observation' when estimating the variance.

This is not too much of a problem, as we can multiply our sample variance by (n-1)/n to compensate. Let $\hat{S}^2 = \frac{n}{n-1}S^2 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}$.

Then $E(\hat{S}^2) = \frac{n}{n-1}E(S^2) = \sigma^2$.

We shall not for now consider the variance of this statistic.

Distribution of sample proportion

Suppose in a given population, there is a proportion P with a certain attribute. So selecting a random individual from that population to see if they have that attribute is a Bernoulli trial. If we select n people at random from the population (with replacement if the population is not

infinite), then the number X of individuals in the sample with the attribute will have a binomial distribution, X~B(n,P).

We saw in the last session that we can work out the sample proportion, p=X/n, which we can use as an estimate for the population proportion, or probability P of 'success' in each trial.

We also saw that this sample proportion p has E(p)=P, and Var(P)=P(1-P)/n. In other words, the sample proportion is an unbiased estimate for the population proportion, and the standard error of the estimate, that is the standard deviation of the distribution, which is equal to $\sqrt{\frac{P(1-P)}{n}}$, decreases as n increases, in other words the estimate becomes more accurate in a larger sample.

What is more, it can be shown that as n increases, the distribution of the sample proportion tends towards a normal distribution, in other words

$p \rightarrow N(P,P(1-P)/n))$.

Thus the distribution of the sampling proportion is fully specified, and can be used to obtain an estimate of the population proportion of any desired degree of accuracy, i.e. of standard error, given a sufficiently large sample.

Estimation

An important use of sampling theory is to estimate parameters of the population and to assess accuracy of such estimates. A point estimate of a parameter of a population is a single-valued estimate derived from sample information. A parameter of a population may be a mean, or variance of some variable, or a correlation coefficient between two variables, or generally any other statistic relating to measurable variables in the population.

For example, we may use sample mean as an estimate for the mean of a variable, sample proportion as an estimate of the proportion of the population with a certain attribute, etc. In econometrics, we see how we derive estimates of the regression coefficients of the relationship between two or more variables.

Formally, an estimator of a population parameter θ, relating to a variable X, based on a on sample $X_1,\ldots,X_n$, is a function $\hat{\theta}(X_1,...,X_n)$. For example, if θ is the mean of a variable X, then we may use $\hat{\theta} = \frac{X_1 + ... + X_n}{n}$. The definition can be extended to involve more than one variable, whereupon the estimator can be a function of the values of each variable in the sample.

Of course, not any function of the data will serve as a good estimator. We look for certain desirable properties of estimators:

Unbiasedness

An estimator $\hat{\theta}$ of a parameter θ is said to be unbiased if $E(\hat{\theta})=\theta$. Otherwise it is biased.

In other words, on average the estimator should equal the 'true' value.

Efficiency

As well as the property of unbiasedness, we are interested in the *accuracy* of an estimator, that is, how far it is likely to be from the population parameter. This is measured by the Standard Error of the estimator, given by

$SE(\hat{\theta})=E(\hat{\theta}-\theta)^2$. If $\hat{\theta}$ is unbiased, then this simplifies to;

$$SE(\hat{\theta})= E(\hat{\theta}^2) - 2E(\theta\hat{\theta}) + E(\theta^2) = E(\hat{\theta}^2) - 2\theta E(\hat{\theta}) + \theta^2$$

Since θ is a constant

$$= E(\hat{\theta}^2) - 2\theta^2 + \theta^2 = E(\hat{\theta}^2) - \theta^2$$

Since θ is unbiased, so $E(\theta)=\hat{\theta}$

$$= E(\hat{\theta}^2) - E(\theta)^2 = Var(\theta)$$

Again since θ is unbiased.

If we have two *unbiased* estimators, θ_1 and θ_2, we say θ_1 is more efficient than θ_2 if $SE(\theta_1)<SE(\theta_2)$.

Consistency

An estimator $\hat{\theta}$ is said to be a consistent estimator of θ if as n tends to ∞, $E(\hat{\theta})$ tends to θ, and $Var(\hat{\theta})$ tends to 0.

Thus, a biased estimator can be nonetheless consistent. For example, we defined above the estimator s^2, the sample variance, as an estimator for the population variance σ^2. This is biased as $E(s^2)=(n-1)\sigma^2/n$. However, as n tends to ∞, (n-1)/n tends to 1, so $E(s^2)$ tends to σ^2. It can also be shown that $Var(s^2)$ tends to 0 as n increases. Hence s^2 is a consistent estimator for σ^2.

Biased but consistent estimators may therefore be used with large samples. (say, at least n=30).

Interval estimates (confidence intervals)

An alternative to a point estimator is an interval estimate, also called a confidence interval. Such an interval specifies a range which is likely to contain the population parameter with a given probability. The calculation of a confidence interval is based on the sampling distribution, which for large samples (n>30) can be assumed to be approximately normal.

Using $\bar{X}$ as an example of an estimator (for $\mu=E(X)$, for some r.v. X), note that we know from the Central Limit Theorem that

$\bar{X} \rightarrow N(\mu,\sigma^2/n)$. We know that $P[-1.96<z<1.96] = 0.95$, where $z\sim N(0,1)$. Hence by the linear properties of the normal distribution,

$$P[-1.96<\frac{\bar{X}-\mu}{(\sigma/\sqrt{n})}<1.96] = 0.95$$

That is, the probability that any normal variable is within 1.96 standard deviations of its mean is 0.95, or 95%. Rearranging this inequality gives

$$P[(\bar{X}-1.96(\sigma/\sqrt{n})) < \mu < (\bar{X}+1.96(\sigma/\sqrt{n}))] = .95$$

This result gives us the 95% confidence interval for the population mean, μ, with ($\bar{X}$ -1.96σ/√n) being the lower limit, and ($\bar{X}$ +1.96σ/√n) being the upper limit of the interval.

If we wanted a 90% confidence interval instead, we would replace 1.96 In the inequality with 1.645, since P(-1.645<z<1.645)=0.9, where z~N(0,1).

Note that μ is not a variable which falls into the calculated confidence interval with 0.95 probability. It is a constant which is either inside the confidence interval or outside it. The random variable here is not μ, but the confidence interval itself, which will be different each time we take a sample. The 95% probability tells us that, 95% of the time, the confidence interval we calculate from our random sample will contain the actual population parameter μ.

In the case of the population proportion related to a binomial distribution, if P is the population proportion and p the sample proportion, we know that for n large, $p \rightarrow N(P,P(1-P)/n)$. Hence, we can derive a 95% confidence interval for P, given by:

$$P[(p - 1.96\sqrt{P(1-P)/n} < P < (p + 1.96\sqrt{P(1-P)/n})] = 0.95.$$

Again, this is based on the fact that 95% of the time, a normal variable (in this case p), will fall within 1.96 standard deviations of its mean. In this case the mean is P, and the SD is √(P(1-P))/n. Note that the letter P here is used both as the population proportion, and to denote the probability operator.