Chapter 11 : Sampling Distributions

We only discuss part of Chapter 11, namely the sampling distributions, the Law of Large Numbers, the (sampling) distribution of \bar{X} and the Central Limit Theorem.

Parameter and Statistic

A parameter is a number that describes the population. Typically that is a number that is of interest but in a statistical problem it is unknown.

A statistic is a number that can be calculated from a given data set.

Example 1 : We are interested in the population of all grades on the first term test. Some parameters of interest are the population mean and the population standard deviation. A random sample of size n (for example n = 20) of grades is given. We can calculate the sample mean, but typically this is not the population mean. In fact the sample mean is a random variable which has its own distribution. This is because if we were to repeat the experiment of taking another random sample of size n we would typically get a different value.

Example 2 : We are interested in the annual growth rate in manufacturing production. A parameter in this example is the *true but unknown* value of the manufacturing growth rate. Statistics Canada publishes an estimate of this based on a sample of manufacturing companies. This estimate is calculated from the random sample and is a statistic.

Example 3 : At a given time the population of Canadian voters has a population proportion that supports a given political party. This is a parameter. An opinion poll takes a random sample of voters and calculates the observed proportion of voters who support the given political party. This sample proportion is a statistic since it is calculated from the sample and not the whole population.

Statistical Estimation and the Law of Large Numbers

The idea of the Law of Large Numbers was illustrated in Chapter 10. Recall the sample proportion of "1" in a game of rolling a single die. There we saw that the running average

or proportion of outcomes eventually settled down to a value of $\frac{1}{6}$. That graph is given here again in Figure 1. This gave us one way of thinking about the idea of probability.

Law of Large Numbers Draw a random sample of size n from a population with mean μ . As the number of observations n becomes large (tends to infinity), the sample mean \bar{x} of the observed data values gets closer to μ (\bar{x} converges to μ).

This is what we see in the graph for the die rolling game above. Thus the idea of long run proportion as probability is the same idea as the Law of Large Numbers.



Running Sample Proportions 1 and 2 for Die Rolls

Figure 1: Running Means for 6000 Fair Rolls of a Die

Sampling Distributions

Suppose we the population of grades for the first term test. As a student you do not know the grades, but I can give you a random sample of n grades. For example n = 10. The Law of Large Numbers says that as n becomes large the value of \bar{x} tends to μ , the true but unknown (to you) value.

 \bar{x} is a random variable with its own distribution. What does this distribution look like? In fact we can use a *computer* simulation game to see what it is.

- 1. Take a random sample of size n = 10.
- 2. Calculate \bar{x} for the sample
- 3. Repeat 1 and 2 a large number of times; say M = 500 times.
- 4. For the M values of \bar{x} make a relative frequency histogram

This relative frequency histogram is a simulation approximation to the probability distribution of \bar{x} .

Aside : We could get a better simulation approximation by taking M larger, for example M = 1000 or M = 10,000. In some uses of this idea, for example in Finance, one might even take M = 50,000 or M = 100,100. In some models in Health Sciences one might take M = 500, and in geography M = 100 or M = 1000.

This histogram tells us something about the shape of the distribution, whether outliers tend to happen, if the shape is symmetric or skewed, or is bell shaped like a normal distribution. One can even use simulation to see what happens to the shape as the sample size n changes.

Example : Grades on Test 1

Figure 2 gives the relative frequency histogram for the grades on Test 1. We can see this distribution is not symmetric and is slightly skewed towards the lower tail, that is



Grades and Normal(mean = 23.6, sd = 4.04)

Figure 2: Test 1 Histogram with Normal Approximation

skewed to the left. Overlaid on this histogram is a normal approximation. This normal approximation is not very good, since it greatly overestimates the proportion of data that would be greater than 30, which is the maximum score possible on the test. Specifically if have X a normal random variable with mean 23.6 and standard deviation 4.04 then

$$P(X > 30) = P\left(\frac{X - 30}{4.04} > \frac{30 - 23.6}{4.04}\right)$$
$$= P(Z > 1.58) = 0.058$$

Thus the normal approximation says that about 6% of the data should be greater than 30, which is not possible.

Suppose we did not have such a histogram or even some information about the population, other than the fact that a mean value is a useful way of describing the centre of the data. How could we learn what is the population mean? From our statistical ideas so far we could take a random sample of the data from the population. From this sample we could then calculate \bar{x} the sample mean.

Some natural question arise that will help our understanding of properties of the sample mean

- How is this sample mean related to the population mean?
- Is \bar{x} a random variable?
- The answer to question above is yes (why?) and so we can ask something about the *distribution* of \bar{x} . What is the mean of the distribution of \bar{x} , the so called sampling distribution of \bar{x} . What is its variance? How is the variance and more generally the sampling distribution of \bar{x} related to the sample size n?

Figure 3 shows the sampling distribution of \bar{x} when n = 10. This plot was produced using M = 500 simulation replicates or runs of a random sample of size n = 10 to produce M = 500 simulated values of \bar{x} . There is an extra dashed vertical line in the centre of the histogram, which indicates the value of the population mean.



n = 10, Normal(mean = 23.6, sd = 1.31)

Figure 3: Test 1 Sampling Distribution of \bar{x} with n = 10 and Normal Approximation

There are several interesting features about histogram.

- the distribution is centred about the same value as the population mean; 23.6. This is highlighted by the dashed vertical line in the centre of the histogram. This line goes through grade = 23.6, the true population mean.
- the spread of the possible values of \bar{x} is much smaller than the spread of the original grades population data.
- The shape of the sampling distribution of \bar{x} is much closer to normal. In fact it is quite good even though n is only 10. The normal approximation used is one with mean = average of \bar{x} and standard deviation = standard deviation of \bar{x} for the M = 500 simulations.

Sampling Distribution of \bar{x}

Mean and Standard Deviation of \bar{x}

Suppose that a simple random sample (SRS) of size n is taken from a large population with mean μ and standard deviation σ (equivalently variance σ^2). The random variable \bar{x} has mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

Since \bar{x} has mean μ , no matter what the real population mean μ might actually be, and no matter what the sample size n might be, we say that \bar{x} is an *unbiased* estimator of μ .

An unbiased estimator is *correct on average*. The variance or equivalently the standard deviation of an unbiased estimator is a measure of how spread out the distribution of the estimator is about its mean.

 \bar{x} has mean μ no matter what is the value of n. However the standard deviation is $\frac{\sigma}{\sqrt{n}}$. Thus if n increases the standard deviation of \bar{x} decreases.

Central Limit Theorem

Suppose that a simple random sample (SRS) of size n is taken from a large population with mean μ and standard deviation σ (equivalently variance σ^2). Suppose that the sample size n is also large. The random variable \bar{x} then has sampling distribution that is approximately Normal with mean = μ , standard deviation $\frac{\sigma}{\sqrt{n}}$.

Remarks

- The shape of the population distribution is not required to have a normal shape, nor even be symmetric. For example is could be skewed.
- The size of the sample does not have to be very big before a normal approximation for the sampling distribution of \bar{x} is quite reasonable. Typically even for a skewed distribution this approximation is quite good for n = 30, while for a distribution that is not too skewed or even symmetric this normal approximation works quite well for sample sizes of 15 or 20, and often for sample sizes of n = 10.

We have seen this in our Grades example for n = 10. There our population distribution in Figure 2 is a little skewed, but the sampling distribution of \bar{x} for n = 10 is well described by a normal distribution as shown in Figure 3. Return to Test 1 grades example.

Figure 4 shows the same idea with several different values of n.

• the distribution is centred about the same value as the population mean; 23.6. This is highlighted by the dashed vertical line in the centre of the histogram. This happens for each n.

Recall that this property is generally stated as the estimator (\bar{x} in this case) is an unbiased estimator of the parameter $\mu =$ population parameter (mean in this case).

• The shape of the sampling distribution of \bar{x} is much closer to normal. This approximation improves as n gets larger.

The normal approximation used is one with mean = average of \bar{x} and standard deviation = standard deviation of \bar{x} for the M = 500 simulations.

There is also a way of finding a normal approximation that can be used when there is only 1 sample taken, and not M = 500 simulation replicates.

• the spread of the possible values of \bar{x} is much smaller than the spread of the original grades population data, and the spread is less as n increases.

Recall the spread is actually described by the standard deviation $\frac{\sigma}{\sqrt{n}}$ where σ is the population standard deviation.



Histograms of Grades Sampling Distribution with Normal Overlays

Figure 4: Test 1 Sampling Distribution of \bar{x} with several different n and Normal Approximation

gr.samp.mean

gr.samp.mean

Remark These properties of the sampling distribution for a given sample size n are determined by the shape of the population distribution. It does not matter if the population has 150 individuals (as in our case), 15,000 individuals of 15 million individuals. This sampling method for estimating a population mean from a sample of size n = 30 for example is especially important when the population size is very large, or possibly not even known exactly other than it is large.

Remark From our knowledge of distributions, and normal distributions in particular we can now state various things. For example based the sampling distribution for n = 10we can say is virtually impossible for the true population mean to be less than 18. Why can we state this? If the population mean were in fact 18, the sampling distribution of \bar{x} would have a normal shape and centred at 18. As we can see from Figure 3 this is clearly not the case. Such an assumption about the population mean μ would be inconsistent with the types of values of \bar{x} that we are nearly guaranteed to observe. We cannot so definitively state that the true population mean is for example 22.5 as this is not very far from the centre of the sampling distribution.