IES302 2011/2 Part II.2 Dr.Prapun

12 Sampling Distributions

Definition 12.1. The link between the probability concepts in the earlier sections and the data is made as follows. Each numerical value in the data is the observed value of a random variable. Furthermore, the random variables are usually assumed to be generated by a distribution (pmf or pdf). These random variables are known as a ***random sample***.

If there are n observations, say, $X_1, X_2, \ldots, X_n$, we say that we have a random sample of size n.

Definition 12.2. A (sample) ***statistic*** is any function of the observations $X_1, X_2, \ldots, X_n$. in a random sample.

12.3. A statistic is also a random variable.

Since a statistic is a random variable, it has a probability distribution (pmf if discrete; pdf if continuous). We call the probability distribution of a statistic a ***sampling distribution***.

Example 12.4. We have seen that we can use the sample data to compute the sample mean that is in some sense a reasonable value (a good guess) of the true population mean.

Do you expect the sample mean $\overline{X}$ to be exactly equal to the value of the population mean, μ? Your answer should be no. We do not expect the means to be identical, but we will be satisfied with our sample results if the sample mean is "close" to the value of the population mean.

Lets consider a second question: If a second sample is taken, will the second sample have a mean equal to the population mean? Equal to the first sample mean? Again, no, we do not expect the sample mean to be equal to the population mean, nor do we expect the second sample mean to be a repeat of the first one. We do, however, again expect the values to be "close." (This argument should hold for any other sample statistic and its corresponding population value.)

12.5. The sample mean is a statistic and it is a random variable.

The probability distribution of $\overline{X}$ is called the ***sampling distribution of the mean*** or the ***distribution of sample means***.

Example 12.6. Introduction to the concept of the sampling distribution of the mean: Consider a population of four members. Each of the four members of the population has a given number of bottles of Diet Pepsi in his or her refrigerator. Bill has 1, Carl has 1, Denise has 3, and Ed has 5.

The population mean of the number of bottles of Diet Pepsi in the refrigerator is

$$\mu = \frac{1+1+3+5}{4} = 2.5 \text{ bottles of Diet Pepsi.}$$

Suppose you can only observe the refrigerator of two persons. Then, you will need to estimate the population mean from the sample of size $n = 2$. There are $\binom{4}{2} = 6$ possible (simple[20]) random samples from this population. The probability that each possible sample might be selected is $\frac{1}{6}$ (each sample has the same probability).

For each sample, we can calculate the sample mean $\overline{X}$ as shown in the following table:

[20]In the simple random sample, every person or element in the population has an equal chance of being included in the sample.

Sample	Mean of this sample
Bill, Carl	$\overline{x} = (1+1)/2 = 1.0$
Bill, Denise	$\overline{x} = (1+3)/2 = 2.0$
Bill, Ed	$\overline{x} = (1+5)/2 = 3.0$
Carl, Denise	$\overline{x} = (1+3)/2 = 2.0$
Carl, Ed	$\overline{x} = (1+5)/2 = 3.0$
Denise, Ed	$\overline{x} = (3+5)/2 = 4.0$

Therefore, the probability distribution of the sample means (i.e., the sampling distribution of the mean), is given by

Remark: the mean of the sample means is 2.5 (the same mean as the original population from which the samples were drawn).

12.7. To summarize what we learned from Example 12.6, Figure 14 shows how the sampling distribution of sample means is formed.

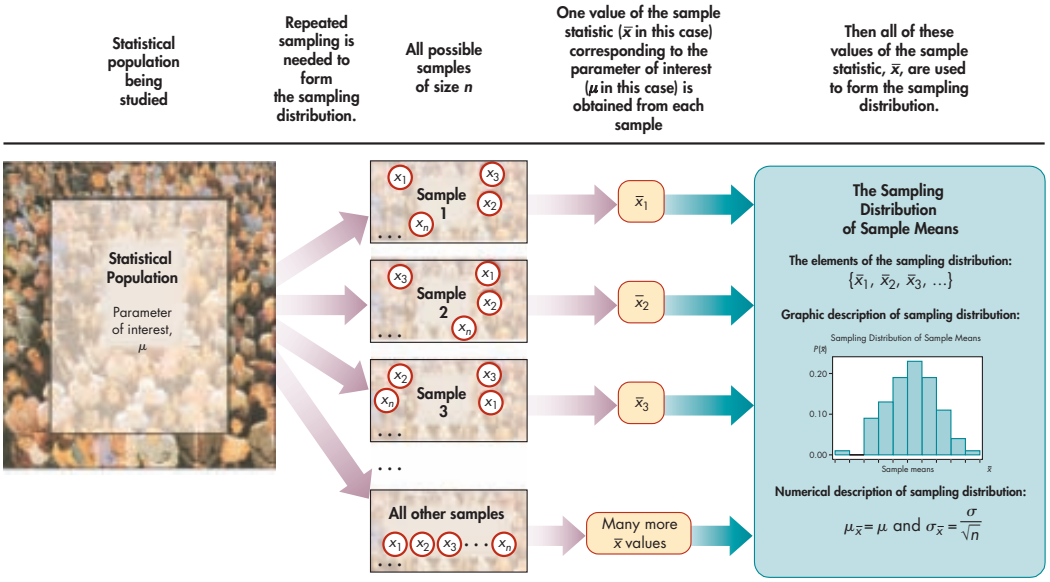

Figure 14: The Sampling Distribution of Sample Means

Theorem 12.8. When a (simple) random sample of size n is from a population that is normally distributed with mean μ, the sample means $\overline{X}$ will also be normally distributed with the same mean. This will be true regardless of the sample size.

Moreover, if the population standard deviation is σ, then the standard deviation of $\overline{X}$ is $\sigma/\sqrt{n}$. Here, the standard deviation of the sample mean is usually referred to as the standard error.

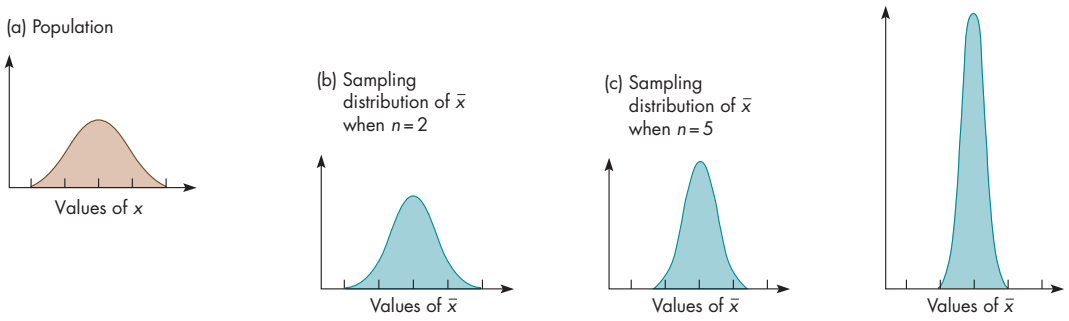

Figure 15: Normally distributed population and three sampling distributions ($n = 2, 5, 30$)

The assumption of normality for a population isn't always realistic, since in many cases the population is either not normally distributed or we have no knowledge about its actual distribution. However, provided that the sample size is large (i.e., $n \geq 30$), the sampling distribution of the mean can still be assumed to be normal. This is because of what is known as the central limit theorem:

Theorem 12.9. ***Central Limit Theorem***: If $X_1, X_2, \ldots, X_n$ is a (simple) random sample of size n taken from a population with mean μ and variance σ^2, and if $\overline{X}$ is the sample mean, the limiting form of the distribution of

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

as $n \to \infty$, is the standard normal distribution.

12.10. Normal approximation: When the random sample size n is large enough, the distribution of the sample means $\overline{X}$ is approx-

imated reasonably well by the normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$.

- As the sample size (n) is increased, the sampling distribution of the mean will more closely approach the normal distribution.
- In many cases of practical interest, if $n \geq 30$, the normal approximation will be satisfactory regardless of the shape of the population.

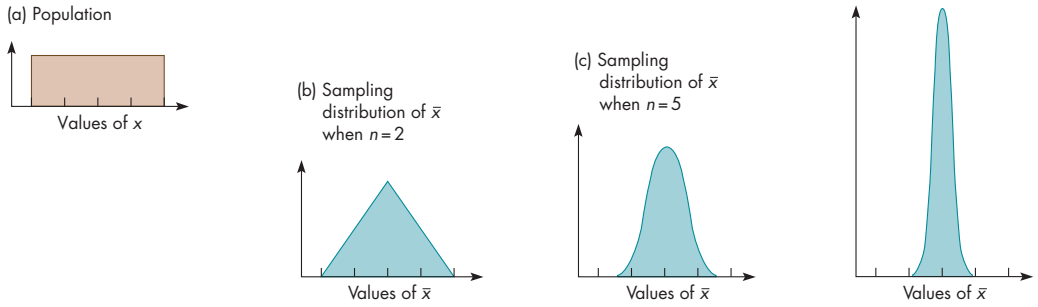

Figure 16: Uniform population distribution and three sampling distributions ($n = 2, 5, 30$)

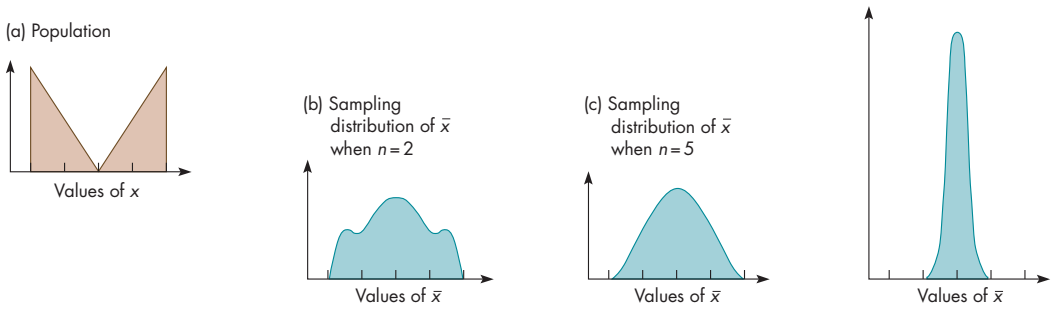

Figure 17: U-shaped population distribution and three sampling distributions ($n = 2, 5, 30$)

The central limit theorem is basic to the concept of statistical inference because it permits us to draw conclusions about the population based strictly on sample data, and without having any knowledge about the distribution of the underlying population.

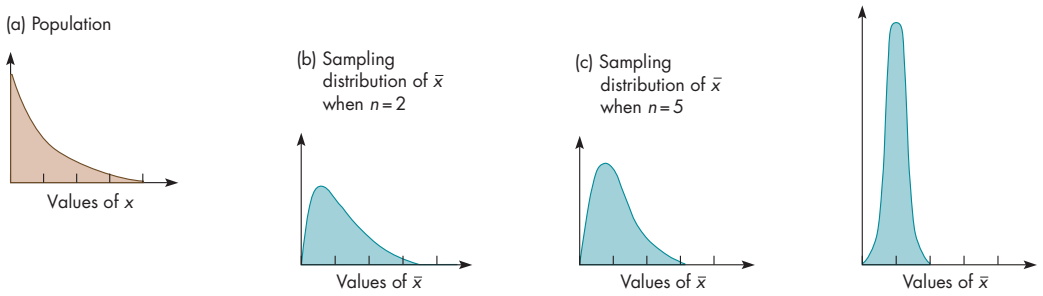

Figure 18: Exponentially distributed population and three sampling distributions ($n = 2, 5, 30$)

Example 12.11. Consider a normal population with $\mu = 100$ and $\sigma = 20$. If a random sample of size 16 is selected, what is the probability that this sample will have a mean value between 90 and 110? That is, what is $P\left[90 < \overline{X} < 110\right]$?