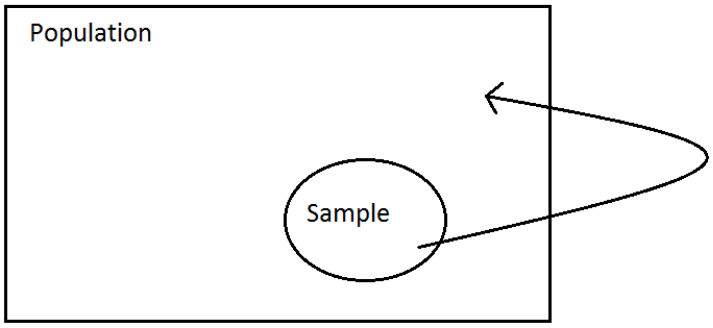
1 Sampling Distributions

Remember in inferential statistics sample data is used to learn about a population.



In this chapter we present how statistics (quantities computed from sample data like $\bar{x}$) can be used to LEARN about unknown parameters characterizing a population (e.g.: mean income of Canadians, mean tuition fee per course at MacEwan, mean lifetime of a car, etc.). In particular statistics will be used to estimate those unknown parameters. In order to evaluate how close we can expect the estimate to be to the true value we need to study the distribution of the statistic.

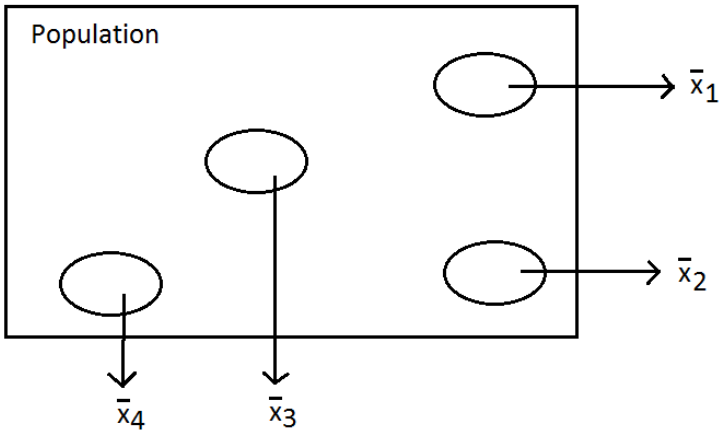
In practical situations the investigator might be able to choose a distribution to use as a model for the population (e.g. mean age at retirement is normally distributed), but the values of the parameters, mean and standard deviation, which specify the distribution are unknown.

1.1 Statistics and Sampling Distributions

When you select a random sample the descriptive measures you calculate are called statistics. These statistics vary or change for each different random sample you select; they are random variables.

Any quantity computed from values in a sample is called a *statistic*.

The value of a statistic varies from sample to sample this is called sampling variability. Since the sampling is done randomly, the value of a statistic is random.



In conclusion: Statistics are random variables.

Since statistics are random variables, they have a distribution, which gives the possible values and their probability.

Wording:
The distribution of a statistic is called a *sampling distribution*.

It provides the following information:

- What values of the statistic can occur.
- What is the probability of each value to occur.

Example:
The population is all students in this section of Stat 151. Let μ be the population mean of the height in this population.

Select a random sample of size 5 and observe the height.

For every random sample the sample mean $\bar{x}$ is different, this is called the sample variability.

Now suppose you look at every possible random sample of 5 students from this class and the corresponding sample mean. From these numbers you can create the sampling distribution.

You will find that

1. The value of $\bar{x}$ differs from one random sample to another (sample variability).
2. Some samples produced $\bar{x}$ values larger than μ, whereas other produce $\bar{x}$ smaller than μ.
3. They can be fairly close to the mean μ, or also quite far off the population mean μ.

The sampling distribution of $\bar{x}$ provides important information about the behavior of the statistic $\bar{x}$ and how it relates to the population mean μ.

Considering how many different samples of size five (for 60 students it's C_5^{60}) there are in this class this process is very cumbersome. Fortunately, there are mathematical theorems that help us to obtain information about the sampling distributions.

1.2 The Sampling Distribution of a Sample Mean

$\bar{x}$ based on a large sample tends to be closer to μ than does $\bar{x}$ based on a small n. This can be explained by the following theoretical results.

Lemma: Suppose $X_1, \ldots, X_n$ are random variables with the same distribution with mean μ and population standard deviation σ.
Now look at the random variable $\bar{X}$.

1. The population mean of $\bar{X}$, denoted $\mu_{\bar{X}}$, is equal to μ.

2. The population standard deviation of $\bar{X}$, denoted $\sigma_{\bar{X}}$, is

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

This means that the sampling distribution of $\bar{x}$ is always centered at μ and the second statement gives the rate the spread of the sampling distribution (sampling variability) decreases as n increases.

Definition:
The standard deviation of a statistic is called the standard error of the statistic (abbreviated SE).

The standard error gives the precision of statistic for estimating a population parameter. The smaller the standard error, the higher the precision.

The standard error of the mean $\bar{X}$ is $SE(\bar{X}) = \sigma/\sqrt{n}$.

Now that we learned about the mean and the standard deviation of the sampling distribution of the sample mean, we might ask, if there is anything we can tell about the shape of the density curve of this distribution.

1.3 Central Limit Theorem

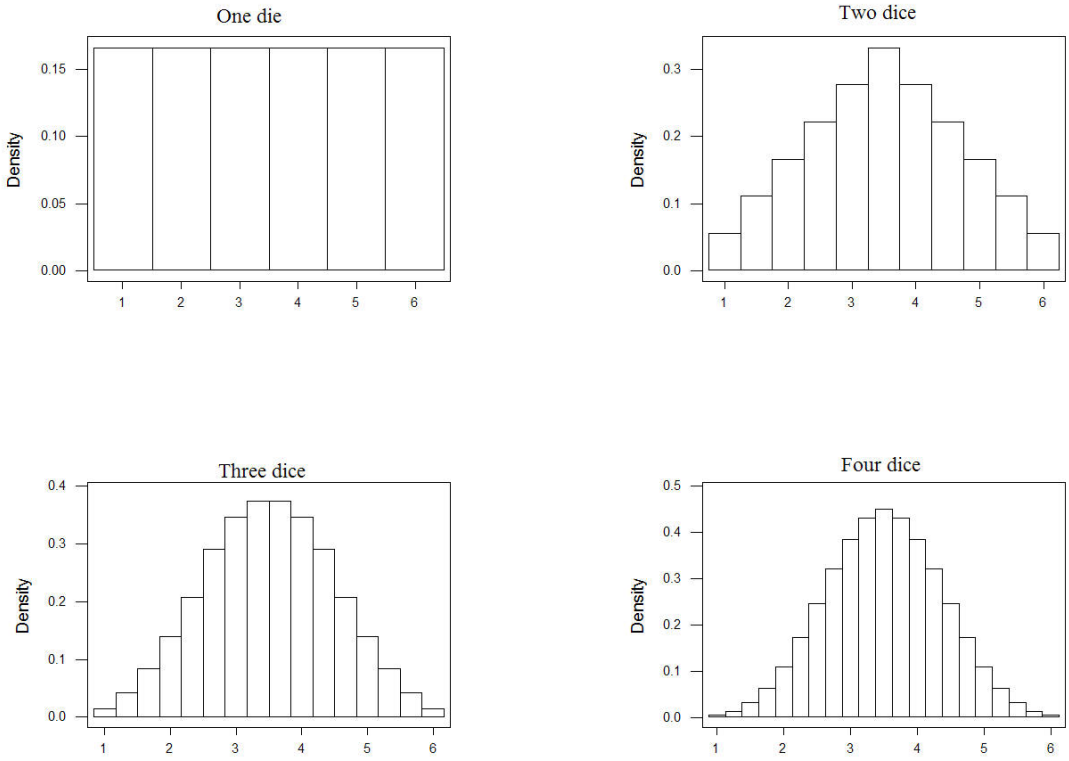
This section explains why the normal distribution is so important in statistics.
The result is surprising. The Central Limit Theorem states, that under rather general conditions, means of random samples drawn from one population tend to have an approximately normal distribution. We find that it does not matter which kind of distribution we find in the population. It even can be discrete or extremely skewed. But if n is *large enough* the distribution of the mean is approximately normal distributed.
That is under all the possible distributions we find one family of distributions, that describes approximately the distribution of a sample mean, if only n is large enough.

Central Limit Theorem
If random samples of n observations are drawn from **any** population with finite mean μ and standard deviation σ, then, when n is large, the sampling distribution of the mean $\bar{X}$ is approximately normal distributed, with mean μ and standard deviation $\sigma/\sqrt{n}$.

Remarks:

- If the population itself is normal $\bar{X}$ is normal distributed for all n, so that n does not have to be large.
- When the sampled population has a symmetric distribution, the sampling distribution of $\bar{X}$ becomes quickly normal. Compare the example below for $n = 3$.
- If the distribution is skewed, usually for $n = 30$ the sampling distribution is already close to a normal distribution.

Example:
Consider tossing n unbiased dice and recording the average number of the upper faces.
The graphs display the sampling distribution for $\bar{X}$ for $n = 1, 2, 3, 4$.



Looking at only $n = 4$ dice leads to a distributions that is very close to a normal distribution.

Summary: Assume that the measurements in a population follow all the same distribution with finite mean μ and standard deviation σ. Then

- The mean of the sampling distribution of the mean $\bar{X}$ of n observations holds

$$\mu_{\bar{x}} = \mu$$

- The standard deviation of the sampling distribution of the mean $\bar{X}$ of n observations holds

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

- The sampling distribution of the mean $\bar{X}$ of n observations is approximately normal distributed, if n is large enough.

Example: According to Stats Canada in 2011 the average commuting time in Canada was 25.4 minutes with a standard deviation of 13 minutes (I made up the standard deviation). (Does the use of the mean make sense here? Let's pretend it does!)

What is the probability that the commuting time of a randomly chosen person is greater than 35 minutes?

$$\begin{aligned} P(X > 35) &= P\left(\frac{X-\mu}{\sigma} > \frac{35-\mu}{\sigma}\right) \quad \text{standardize} \\ &= P\left(Z > \frac{35-25.4}{13}\right) \\ &= P(Z > 0.74) \\ &= 1 - 0.7704 = 0.2296 \qquad \text{Table} \end{aligned}$$

What is the probability that the **average** commuting time of 5 people is greater 35 minutes?

$$\begin{aligned} P(\bar{X} > 35) &= P\left(\frac{\bar{X}-\mu_{\bar{X}}}{\sigma_{\bar{x}}} > \frac{35-\mu_{\bar{X}}}{\sigma_{\bar{x}}}\right) \quad \text{standardize} \\ &= P\left(Z > \frac{35-25.4}{13/\sqrt{5}}\right) \\ &= P(z > 1.65) \\ &= 1 - 0.9505 = 0.0495 \qquad \text{Table} \end{aligned}$$