1 Sampling Distributions

In this chapter we will be developing the mathematical models for the populations under investigation in statistical studies. We will see that statistics (a quantity computed from values in a sample) can be used to ESTIMATE the unknown parameter characterizing a population. In order to evaluate how close the estimate is to the true value we need to study the distribution of the statistic.

In practical situations the investigator might be able to determine the type of distributions to use as a model, but the values of the parameters (mean and standard deviation or the probability for "success") that specify its exact form are unknown.

1.1 Statistics and Sampling Distributions

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

Definition:
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 sample 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.

Definition:
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

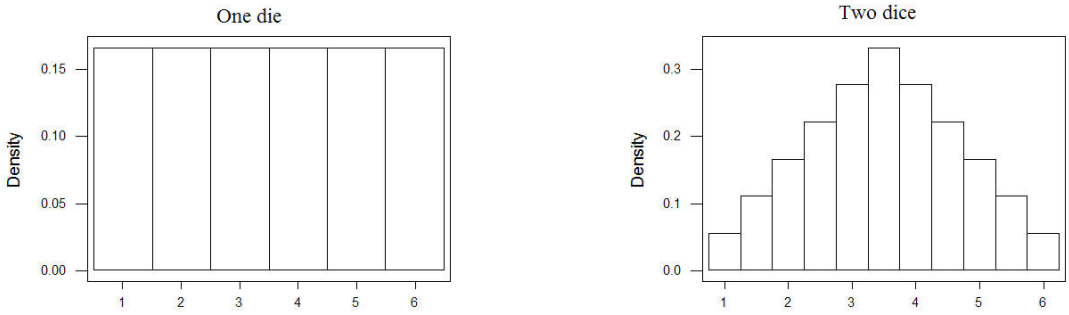
This section explains why the normal distribution is so important in statistics.
The result is surprising. The Central Limit Theorem states, that under rather general conditions, means of random samples drawn from one population tend to have an approximately normal distribution. We find that it does 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.

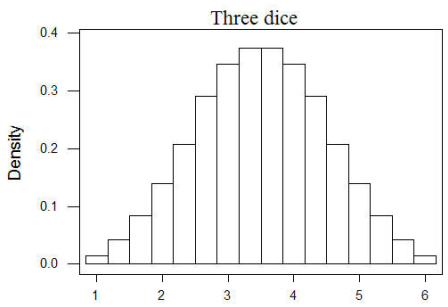
Central Limit Theorem
If random samples of n observations are drawn from **any** population with finite mean μ and standard deviation σ, then, when n is large, the sampling distribution of the mean $\bar{X}$ is approximately normal distributed, with mean μ and standard deviation $\sigma/\sqrt{n}$.

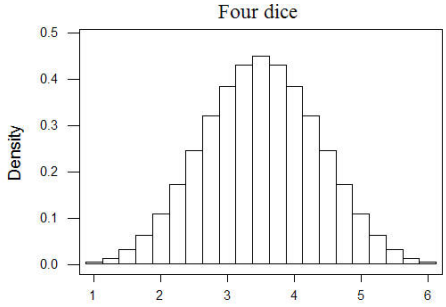
Remarks:

- 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:
The duration of Alzheimer's disease from the onset of symptoms until death ranges from 3 to 20 years. The mean is 8 years and the standard deviation is 4 years.
Looking at the average duration for 30 randomly selected Alzheimer patients:
What is the probability that the average duration of those 30 patients is less than 7 years?

$$\begin{aligned} P(\bar{X} < 7) &= P\left(\frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{x}}} < \frac{7 - \mu_{\bar{X}}}{\sigma_{\bar{x}}}\right) \quad \text{standardize} \\ &= P\left(z < \frac{7 - 8}{4/\sqrt{30}}\right) \\ &= P(z < -1.37) \\ &= 0.0853 \qquad\qquad \text{Table} \end{aligned}$$

1.4 The Sampling Distribution of the Sample Proportion

Look now at a binomial distributed random variable X. That is the experiment is the result of n trials, in which the event of interest occurs with probability p. X is the random variable that gives the number out of n trials in which the event of interest occurred.
For example:

- flip n coins and observe, how often tail was tossed.
- look at n people and survey how many have an IQ above 120.
- check for n students, how many are "nonresidents".
- check how many out of n patients survived at least five years, after a specific cancer treatment.
- count how many out of n clinical tests are positive.

Looking at a random sample of the size n, the probability p can be estimated by calculating the sample proportion of Successes. If X gives the count of Successes out of n trials

$$\hat{p} = \frac{X}{n}.$$

The random variable X is binomial distributed with mean pn and standard deviation equals $\sqrt{p(1-p)}$. Since $\hat{p}$ is simply the value of X expressed as an proportion, the sampling distribution of $\hat{p}$ is identical to the probability distribution of X, except that it has a new measurement scale.

Now rewrite the sample proportion in the following way.

First let single observations be encoded in the following way.

$$X_i = \begin{cases} 0 & \text{if individual is labelled F} \\ \\ 1 & \text{if individual is labelled S} \end{cases}$$

Then is $\hat{p} = \sum X_i/n = \bar{X}$. So that the Central Limit Theorem applies to $\hat{p}$.

Result:

- $\mu_{\hat{p}}$ the mean of the sampling distribution of $\hat{p}$ equals p: $\mu_{\hat{p}} = p$
- The population standard deviation is

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

- The sampling proportion $\hat{p}$ is for large n approximately normal distributed. (Central Limit Theorem)

A rule of thumb states that the Central Limit Theorem can be applied if:

$$np > 5 \text{ and } n(1-p) > 5$$

Example:
A study showed, that the proportion of people in the 20 to 34 age group with an IQ (on the Wechsler Intelligence Scale) of over 120 is about 0.35.

Calculate the probability for the event that in a sample of 50 there are more than 30 people with an IQ of at least 120.
First obtain $\mu_{\hat{p}}$ and $\sigma_{\hat{p}}$ for the distribution of the sample proportion $\hat{p}$.

- $\mu_{\hat{p}} = p = 0.35$
- $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.35(1-0.35)}{50}} = 0.06745.$

$$
\begin{aligned}
P(\hat{p} > \tfrac{30}{50}) &= P\left(\frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} > \frac{0.6 - \mu_{\hat{p}}}{\sigma_{\hat{p}}}\right) && \text{standardize} \\
&= P\left(\frac{\hat{p} - 0.35}{0.06745} > \frac{0.6 - 0.35}{0.06745}\right) && \text{use above results} \\
&= 1 - P\left(\frac{\hat{p} - 0.35}{0.06745} \leq 3.706\right) && \text{Rule for Compliments} \\
&= 1 - 0.9999 && \text{CLT, np} = 17.5 > 5, \text{ Table3} \\
&= 0.0001
\end{aligned}
$$

We calculated that the probability that more than 30 out of 50 people (between 20 and 34) have an IQ greater than 120 is 0.0001. It is highly unlikely!