7.3 - Sampling Distribution and the Central Limit Theorem

A population parameter (ex. $\mu$, $\sigma$) is always constant, but a sample statistics (ex. $\bar{x}$, $s$) is always a random variable, because it will depend on what elements are included in the sample. That is, different samples from the same population can have different means for instance.

The Sampling Distribution of $\bar{x}$ is the probability distribution of all possible values of $\bar{x}$, when all possible samples of the same size $n$ are taken from the same population.

There are $\frac{N}{n}C_n$ different samples of size $n$ that we can pick from a population of size $N$.

The mean and SD of the sampling distribution of $\bar{x}$ are called the mean and standard deviation of $\bar{x}$ and denoted by $\mu_\bar{x}$ and $\sigma_\bar{x}$ respectively.

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

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

The standard deviation $\sigma_\bar{x}$ is sometimes called the standard error of the mean.

Note that $\sigma_\bar{x}$ is smaller than $\sigma$ and as the sample size increases, $\sigma_\bar{x}$ decreases. That is, the sample means gets closer and closer to $\mu$.

When the population from which samples are drawn is normally distributed, then the shape of the sampling distribution of $\bar{x}$ is also normally distributed.

According to the Central Limit Theorem, the sampling distribution of $\bar{x}$ is approximately normal for a large sample size, regardless of the shape of its population distribution.

The approximation becomes more accurate as the sample size increases.

A sample is generally considered large if $n > 30$.

The $z$-value for a value of $\bar{x}$ is calculated as

$$z = \frac{\bar{x} - \mu_\bar{x}}{\sigma_\bar{x}}$$
Applications of the Central Limit Theorem

Using our calculators to find areas under the normal curve, we can use the central limit theorem to make statements as follows:

1. If we take all possible samples of the same (large) size from a population, then about 68.26% of the sample means will be within one standard deviation of the population mean.

2. If we take one large sample from a population, the probability that this sample mean will be within one standard deviation of the population mean is 0.6826.

This last statement is what we find more useful, since we in real life never look at ALL possible samples, but instead we want to select ONE sample and find the probability that the value of $\bar{x}$ from this sample falls within a given interval.

ex. The print on the package of 100-watt General Electric soft-white light-bulbs says that these bulbs have an average life of 750 hours. Assume that the lives of all such bulbs have a normal distribution with a mean of 750 hours and a standard deviation of 55 hours. Find the probability that the mean life of a random sample of 25 such bulbs will be less than 725 hours.

ex. The annual per capita (average per person) chewing gum consumption in the United States is 200 pieces. Suppose that the standard deviation of per capita consumption of chewing gum is 145 pieces per year.

(a) Find the probability that the average annual chewing gum consumption of 84 randomly selected Americans is more than 220 pieces.

(b) Find the probability that the average annual chewing gum consumption of 84 randomly selected Americans is within 100 pieces of the population mean.

(c) Find the probability that the average annual chewing gum consumption of 16 randomly selected Americans is less than 100 pieces.
7.4 - Population and Sample Proportion

The Population Proportion, denoted by \( p \), is the ratio of the number of elements in a population with a specific characteristic to the total number of elements in the population.

The Sample Proportion, denoted by \( \hat{p} \) (read as "p hat"), is the same ratio but for a sample.

\[
p = \frac{X}{N} \quad \text{and} \quad \hat{p} = \frac{x}{n}
\]

Note that the relative frequency of a category or class gives the proportion that belongs to that category or class, and the probability of success in a binomial experiment also represents a proportion.

ex. In a random sample of 1000 subjects, 640 possess a certain characteristic. A sample of 40 subjects selected from this population has 24 subjects who possess the same characteristic. What are the values of the population and sample proportions?

\( \hat{p} \) is a random variable (since it will vary depending on which elements are included in the sample), thus it has a probability distribution.

The Sampling Distribution of the Sample Proportion, \( \hat{p} \), is the probability distribution of \( \hat{p} \), which gives all different values that \( \hat{p} \) can assume and their probabilities.

The mean and standard deviation of the sample proportion, \( \hat{p} \), is denoted by \( \mu_\hat{p} \), and \( \sigma_\hat{p} \), respectively.

\[
\mu_\hat{p} = p \quad \text{and} \quad \sigma_\hat{p} = \sqrt{\frac{p(1-p)}{n}}
\]

According to the Central Limit Theorem for Proportions, the sampling distribution of \( \hat{p} \) is approximately normal for a large sample size.

The approximation becomes more accurate as the sample size increases.

A sample is generally considered large if \( np \geq 10 \) and \( n(1-p) \geq 10 \).
ex. Gluten sensitivity affects approximately 15% of people in the US. Let \( \hat{p} \) be the proportion in a random sample of 800 individuals who have gluten sensitivity. Find the probability that the value of \( \hat{p} \) is

(a) within 0.02 of the population proportion

(b) not within 0.02 of the population proportion

(c) greater than the population proportion by 0.025 or more

ex. Seventy percent of adults favor some kind of government control on the prices of medicines. What is the probability that the proportion of adults in a random sample of 400 who favor some kind of government control is

(a) less than 0.65

(b) between 0.73 and 0.76

(c) within 0.06 of the population proportion
7.6 – Assessing Normality

When a sample is large, we have the Central Limit theorem to ensure that $\bar{x}$ is normally distributed. But when the sample is small we need to determine if the population, from which the sample was taken, is approximately normal in order to be able to use the techniques learnt in chapter 7.

We will NOT assume a population to be normal if
- The sample contains an outlier
- The sample exhibits a large degree of skewness
- The sample is multimodal

If a sample has none of these features, we can treat the population as being approximately normal.

We can use dotplots, boxplots, stem-and-leaf plots, and histograms to help us detect above features.

*Determine if it’s reasonable to treat the following as samples from approximately normal populations:*

ex. [Dotplot]

ex. [Boxplot]

ex. [Stem-and-leaf plot]

ex. [Histogram]
A more sophisticated way to assess normality is to use **Normal Quantile Plots**, also called **Normal Probability Plots**.

Below is an example of the distribution of systolic blood pressure, for a random group of healthy patients.

![Histogram and Normal Quantile Plot](image)

Looking at the histogram, you can see the sample is approximately normally distributed. But the bar heights for 120-122 and 122-124 make the distribution look slightly skewed, so it’s not perfectly clear.

The normal quantile plot to the right is clearer. It shows the *observations* on the X axis plotted against the *expected normal score* (Z-score) on the Y axis. It’s not necessary to understand what an *expected normal score* is, nor how it’s calculated, to interpret the plot. All you need to do is check that the points roughly follow a straight line. If the points roughly follow a line – as they do in this case – the sample has a normal distribution.

The steps for constructing a normal quantile plot on the TI-84 PLUS Calculator are:

1. **Step 1.** Enter the data into L1 in the data editor.
2. **Step 2.** Press 2nd, Y= to access the STAT PLOTS menu and select Plot1 by pressing 1.
3. **Step 3.** Select On and the normal quantile plot icon.
4. **Step 4.** For Data List, select L1, and for Data Axis, choose the X option.
5. **Step 5.** Press ZOOM and then 9: ZoomStat.

---

**ex.** A placement exam is given to each entering freshman at a large university. A simple random sample of 20 exam scores is drawn, with the following results.

<table>
<thead>
<tr>
<th>Score</th>
<th>Score</th>
<th>Score</th>
<th>Score</th>
<th>Score</th>
<th>Score</th>
<th>Score</th>
<th>Score</th>
<th>Score</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>61</td>
<td>60</td>
<td>60</td>
<td>68</td>
<td>63</td>
<td>63</td>
<td>94</td>
<td>66</td>
<td>65</td>
<td>98</td>
</tr>
<tr>
<td>61</td>
<td>71</td>
<td>74</td>
<td>63</td>
<td>66</td>
<td>61</td>
<td>61</td>
<td>65</td>
<td>72</td>
<td>85</td>
</tr>
</tbody>
</table>

Construct a normal probability plot to determine if the exam scores are approximately normal.