Chapter 18: Sampling Distribution Models

Suppose I randomly select 100 seniors in Scott County and record each one’s GPA.

1.95 1.98 1.86 2.04 2.75 2.72 2.06 3.36 2.09 2.06
2.33 2.56 2.17 1.67 2.75 3.95 2.23 4.53 1.31 3.79
1.29 3.00 1.89 2.36 2.76 3.29 1.51 1.09 2.75 2.68
2.28 3.13 2.62 2.85 2.41 3.16 3.39 3.18 4.05 3.26
1.95 3.23 2.53 3.70 2.90 2.79 3.08 2.79 3.26 2.29
2.59 1.36 2.38 2.03 3.31 2.05 1.58 3.12 3.33 2.04
2.81 3.94 0.82 3.14 2.63 1.51 2.24 2.22 1.85 1.96
2.05 2.62 3.27 1.94 2.01 1.68 2.01 3.15 3.44 4.00
2.33 3.01 3.15 2.25 3.34 2.22 3.29 3.90 2.96 2.61
3.01 2.86 1.70 1.55 1.63 2.37 2.84 1.67 2.92 3.29

These 100 seniors make up one possible______________. All seniors in Scott County make up the
______________________.

The sample mean (_____) is 2.5470 and the sample standard deviation (_____) is 0.7150. The
population mean (____) and the population standard deviation (____) are unknown.

We can use _____ to estimate _____ and we can use _____ to estimate ____. These estimates may or
may not be reliable.

A number that describes the population is called a _______________. Hence, μ and σ are both
_______________. A parameter is usually represented by _____.

A number that is computed from a sample is called a _______________. Therefore, \( \bar{x} \) and \( s \) are both
_______________. A statistic is usually represented by _____.

If I had chosen a different 100 seniors, then I would have a different sample, but it would still represent
the same population. A different sample almost always produces different statistics.

**Example**: Let _____ represent the proportion of seniors in a sample of 100 seniors whose GPA is 2.0 or
higher.

\[ \hat{p}_1 = .78 \quad p_3 = .81 \quad \hat{p}_5 = .68 \quad \hat{p}_7 = .79 \quad \hat{p}_9 = .83 \]
\[ \hat{p}_2 = .72 \quad \hat{p}_4 = .70 \quad \hat{p}_6 = .75 \quad \hat{p}_8 = .72 \quad \hat{p}_{10} = .76 \]
If I compare many different samples and the statistic is very similar in each one, then the _______ is low. If I compare many different samples and the statistic is very different in each one, then the _______ is high.

The __________________________ of a statistic is a model of the values of the statistic from all possible samples of the same size from the same population.

**Example:** Suppose the sampling model consists of the samples \( \hat{p}_1, \hat{p}_2, \ldots, \hat{p}_n \). (Note: There are actually many more than ten possible samples.) This sampling model has mean 0.754 and standard deviation 0.049.

The statistic used to estimate a parameter is ______ _________ if the mean of its ______________ is equal to the true value of the parameter being estimated.

**Example:** Since the mean of the sampling model is 0.754, then _____ is an unbiased estimator of _____ if the true value of _____ (the proportion of all seniors in Scott County with a GPA of 2.0 or higher) equals 0.754.

A statistic can be _______ and still have high _______. To avoid this, increase the size of the sample. Larger samples give smaller spread.

<table>
<thead>
<tr>
<th>Sample Proportions:</th>
</tr>
</thead>
</table>

The parameter _____ is the population proportion. In practice, this value is always unknown. (If we know the population proportion, then there is no need for a sample.)

The statistic _____ is the sample proportion. We use _____ to estimate the value of ____. The value of the statistic _____ changes as the sample changes.

How can we describe the sampling model for _____?
1. shape?
2. center?
3. spread?
If our sample is an SRS of size $n$, then the following statements describe the sampling model for ____:

1. The shape is ______________________________.
   
   **ASSUMPTION:** Sample size is sufficiently large.
   **CONDITION:** $np \geq 10$ and $nq \geq 10$

2. The _______________is $p$.

3. The ______________________________is $\sqrt{\frac{pq}{n}}$.
   
   **ASSUMPTION:** Sample size is sufficiently large.
   **CONDITION:** The population is at least 10 times as large as the sample.

If we have _____________________data, then we must use ______________________________to construct a sampling model.

**Example:**
Suppose we want to know how many seniors in Kentucky plan to attend college. We want to know how many seniors would answer, “YES” to the question, “Do you plan to attend college?” These responses are _____________________.

So _____ (our parameter) is the proportion of all seniors Kentucky who plan to attend college. Let _____ (our statistic) be the proportion of Kentucky students in an SRS of size 100 who plan to attend college. To calculate the value of _____, we divide the number of “Yes” responses in our sample by the total number of students in the sample.

If I graph the values of _____ for all possible samples of size 100, then I have constructed a ______________________________. What will the sampling model look like? It will be _____________________normal. In fact, the larger my sample size, the closer it will be to a normal model. It can never be perfectly normal, because our data is discrete, and normal distributions are continuous.

So how large is large enough to ensure that the sampling model is close to normal?? Both _____and _____should be at least 10 in order for normal approximations to be useful. Furthermore...

The mean of the sampling model will equal the true population proportion, _____.

The standard deviation (if the population is at least 10 times as large as the sample) will be __________.
Sample Means:

If, on the other hand, we have _____________________data, then we can use ___________________ _________________to construct a sampling model.

Example:
Suppose I randomly select 100 seniors in Kentucky and record each one’s GPA. I am interested in knowing the average GPA of a senior in Kentucky.

<table>
<thead>
<tr>
<th>1.95</th>
<th>1.98</th>
<th>1.86</th>
<th>2.04</th>
<th>2.75</th>
<th>2.72</th>
<th>2.06</th>
<th>3.36</th>
<th>2.09</th>
<th>2.06</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.33</td>
<td>2.56</td>
<td>2.17</td>
<td>1.67</td>
<td>2.75</td>
<td>3.95</td>
<td>2.23</td>
<td>4.53</td>
<td>1.31</td>
<td>3.79</td>
</tr>
<tr>
<td>1.29</td>
<td>3.00</td>
<td>1.89</td>
<td>2.36</td>
<td>2.76</td>
<td>3.29</td>
<td>1.51</td>
<td>1.09</td>
<td>2.75</td>
<td>2.68</td>
</tr>
<tr>
<td>2.28</td>
<td>3.13</td>
<td>2.62</td>
<td>2.85</td>
<td>2.41</td>
<td>3.16</td>
<td>3.39</td>
<td>3.18</td>
<td>4.05</td>
<td>3.26</td>
</tr>
<tr>
<td>1.95</td>
<td>3.23</td>
<td>2.53</td>
<td>3.70</td>
<td>2.90</td>
<td>2.79</td>
<td>3.08</td>
<td>2.79</td>
<td>3.26</td>
<td>2.29</td>
</tr>
<tr>
<td>2.59</td>
<td>1.36</td>
<td>2.38</td>
<td>2.03</td>
<td>3.31</td>
<td>2.05</td>
<td>1.58</td>
<td>3.12</td>
<td>3.33</td>
<td>2.04</td>
</tr>
<tr>
<td>2.81</td>
<td>3.94</td>
<td>0.82</td>
<td>3.14</td>
<td>2.63</td>
<td>1.51</td>
<td>2.24</td>
<td>2.22</td>
<td>1.85</td>
<td>1.96</td>
</tr>
<tr>
<td>2.05</td>
<td>2.62</td>
<td>3.27</td>
<td>1.94</td>
<td>2.01</td>
<td>1.68</td>
<td>2.01</td>
<td>3.15</td>
<td>3.44</td>
<td>4.00</td>
</tr>
<tr>
<td>2.33</td>
<td>3.01</td>
<td>3.15</td>
<td>2.25</td>
<td>3.34</td>
<td>2.22</td>
<td>3.29</td>
<td>3.90</td>
<td>2.96</td>
<td>2.61</td>
</tr>
<tr>
<td>3.01</td>
<td>2.86</td>
<td>1.70</td>
<td>1.55</td>
<td>1.63</td>
<td>2.37</td>
<td>2.84</td>
<td>1.67</td>
<td>2.92</td>
<td>3.29</td>
</tr>
</tbody>
</table>

These 100 seniors make up one possible _______________. The sample mean (_____ ) is 2.5470 and the sample standard deviation (_____ ) is 0.7150.

So _____ (our parameter) is the true mean GPA of a senior in Kentucky.

And _____ (our statistic) is the mean GPA of a senior in Kentucky in an SRS of size 100.

To calculate the value of _____, we find the mean of our sample (_____).

If we pick different samples, then the value of our statistic _____ changes:

\[
\hat{\mu}_1 = \bar{x}_1 = 2.5470 \quad \hat{\mu}_6 = \bar{x}_6 = 2.3962 \\
\hat{\mu}_2 = \bar{x}_2 = 2.4943 \quad \hat{\mu}_7 = \bar{x}_7 = 2.5019 \\
\hat{\mu}_3 = \bar{x}_3 = 2.6223 \quad \hat{\mu}_8 = \bar{x}_8 = 2.5621 \\
\hat{\mu}_4 = \bar{x}_4 = 2.5289 \quad \hat{\mu}_9 = \bar{x}_9 = 2.6083 \\
\hat{\mu}_5 = \bar{x}_5 = 2.4037 \quad \hat{\mu}_{10} = \bar{x}_{10} = 2.5667
\]

If I graph the values of _____ for all possible samples of size 100, then I have constructed a _________________ of sample means. What will the sampling model look like?
Remember that each _____ value is a mean. Means are __________________ than individual observations because if we are looking only at means, then we don’t see any extreme values, only average values. We won’t see GPA’s that are very low or very high, only average GPA’s.

The larger the sample size, the less variation we will see in the values of _____ . So the standard deviation decreases as the sample size increases.

**So what will the sampling model look like??**

If the sample size is large, it will be ______________________.

It can never be perfectly normal, because our data is discrete, and normal distributions are continuous.

Furthermore...

The mean of the sampling model will equal the true population mean _____.

And...

The standard deviation will be __________ (if the population is at least 10 times as large as the sample).

---

**Central Limit Theorem**

Draw an SRS of size \( n \) from any population whatsoever with mean \( \mu \) and standard deviation \( \sigma \).

When \( n \) is large, the sampling model of the sample means \( \bar{x} \) is close to the normal model \( N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \)

with mean \( \mu \) and standard deviation \( \frac{\sigma}{\sqrt{n}} \).

---

**Law of Large Numbers**

Draw observations at random from any population with mean \( \mu \). As the number of observations increases, the sample mean \( \bar{x} \) gets closer and closer to \( \mu \).