Sampling Distribution Models

Chapter 17
Objectives:

1. Sampling Distribution Model
2. Sampling Variability (sampling error)
3. Sampling Distribution Model for a Proportion
4. Central Limit Theorem
5. Sampling Distribution Model for a Mean
Introduction:

- Relationship between the sample and its population.
Terminology:

• Parameter
  – A numerical descriptive measure of a population.
  – In statistical practice, the value of a parameter is not normally known.

• Statistic
  – A numerical descriptive measure of a sample.
  – Use a statistic to estimate an unknown parameter.
Compare

- parameter
  - mean: $\mu$
  - standard deviation: $\sigma$
  - proportion: $p$

- Sometimes we call the parameters “true”; true mean, true proportion, etc.

- statistic
  - mean: $x$-bar
  - standard deviation: $s$
  - proportion: $p$-hat

- Sometimes we call the statistics “sample”; sample mean, sample proportion, etc.
# Notation:

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Statistic</th>
<th>Parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>$\bar{x}$</td>
<td>$\mu$</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>$s$</td>
<td>$\sigma$</td>
</tr>
<tr>
<td>Proportion</td>
<td>$\hat{p}$</td>
<td>$p$</td>
</tr>
</tbody>
</table>
Sampling Variability

- When a sample is drawn at random from a given population, each sample drawn will be different.
- These sample to sample differences are called **sampling variability**.
- Sometimes, unfortunately, called sampling error, sampling variability is no error at all, but just the natural result of random sampling.
- Sampling variability can be reduced by increasing the sample size.
Example: Sampling Variability

- Frequency distribution for the number of accidents by bus drivers.
- The population – 708 bus drivers employed by public corporations and the number of traffic accidents in which each bus driver was involved during a 4-year period.
- Four different samples of 50 observations from this population.
- The histograms of these five distributions (population and 4 samples) resemble one another in a general way, but there are obvious dissimilarities due to sampling variability.
The Population:

Population histogram

Relative frequency

Number of accidents
The Samples:
Sampling Distribution

• Is a distribution of a sample statistic in all possible samples of the same size from the same population.

• Note the “-ing” on the end of **Sample**. It looks and sounds similar to the Sample Distribution, but in reality the concept is much closer to a population model.
Continued

• It is a model of a distribution of scores, like the population distribution, except that the scores are not raw scores, but statistics.

• It is a thought experiment; “what would the world be like if a person repeatedly took samples of size N from the population distribution and computed a particular statistic each time?”

• The resulting distribution of statistics is called the sampling distribution.

• Every statistic has a sampling distribution.
Example: Sampling Distribution

• Westvaco is laying off workers. There are 10 workers who could have been laid off; their ages are \{25, 33, 35, 38, 48, 55, 55, 55, 56, 64\}.

• The ages of the three workers actually laid off were 55, 55, and 64.

• Did Westvaco discriminate against the three laid off workers due to age?
Solution:

• Randomly select three workers, without replacement because you can’t layoff the same worker twice, from the 10 possible for being laid off.

• The summary statistic is the mean age of the three workers chosen at random.

• Repeat this process many times in order to generate a sampling distribution of the summary statistic.

• From this sampling distribution, make a decision about whether it was reasonable to assume that workers were selected for layoff without respect to their age.
The actual mean age of the laid off workers was 58. As can be seen it is hard to get an mean age that large just by chance. Westvaco has some explaining to do.
Steps for Generating a Sampling Distribution

1. **Random Sample** – take a random sample of a fixed size $n$ from the population (may be simulated).

2. **Summary Statistic** – Compute a summary statistic.

3. **Repetition** – Repeat steps 1 and 2 many times.

4. **Distribution** – display the distribution of the summary statistics.
Describing Sampling Distributions

• Use the same tools of data analysis used to describe any distribution.
  1. Overall shape
  2. Outliers
  3. Center
  4. Spread
Example:

Display 5.7  A simulated sampling distribution of the sample mean \((n = 5)\) for the areas of rectangles.

- This sampling distribution is approximately normal, with a slight skew to the right, mean 7.31, standard deviation 2.39, and no outliers.
Bias of a Statistic

• How well does a statistic estimate the value of a parameter.

• Sampling distributions allow us to describe *bias* more precisely by speaking of the bias of a statistic rather than bias in a sampling method.

• Bias concerns the center of the sampling distribution.
Unbiased Statistic/Unbiased Estimator

- A statistic used to estimate a parameter is *unbiased* if the mean of its sampling distribution is equal to the true value of the parameter being estimated.
- The statistic is then called an *unbiased estimator* of the parameter.
Example: Unbiased Statistic

- The approximate sampling distributions for sample proportions for SRS’s of two sizes drawn from a population with \( p = 0.37 \).

- (a) Sample size 100
- (b) Sample size 1000

Both statistics are unbiased because the means of the distributions equal the true population value \( p = 0.37 \). The statistic from the larger sample is less variable.
Variability of a Statistic

• Is described by the spread of its sampling distribution.

• This spread is determined by the sampling design and the size of the sample. Larger samples give smaller spread.

• As long as the population is much larger than the sample (at least 10 times as large), the spread of the sampling distribution is approximately the same for any population size.
Examples: Bias and Variability

• Think of the true value of the population parameter as the bull’s-eye on a target.
• Bias means our aim is off and we consistently miss the bull’s-eye in the same direction. Our sample values do not center on the population value.
• Variability means how widely scattered our shots are on the target.
Determine Bias & Variability
Determine Bias & Variability
Determine Bias & Variability
Determine Bias & Variability
Summary Bias and Variability

High bias, low variability
(a)

Low bias, high variability
(b)

High bias, high variability
(c)

The ideal: low bias, low variability
(d)
PRACTICE: STATE BIAS AND VARIABILITY

Hi Bias and Hi Variability

Population parameter

(a)

Hi Bias and Hi Variability
Low Bias and Low Variability
PRACTICE: STATE BIAS AND VARIABILITY

Low Bias and Hi Variability

(c) Population parameter
Hi Bias and Low Variability
School Note...

Sample Proportions
Sample Proportions

• Rather than showing real repeated samples, *imagine* what would happen if we were to actually draw many samples.

• Now imagine what would happen if we looked at the sample proportions for these samples.

• The histogram we’d get if we could see *all the proportions from all possible samples* is called the *sampling distribution* of the proportions.

• What would the histogram of all the sample proportions look like?

School Note...
Sample Proportions

- We would expect the histogram of the sample proportions to center at the true proportion, $p$, in the population.
- As far as the shape of the histogram goes, we can simulate a bunch of random samples that we didn’t really draw.
- It turns out that the histogram is unimodal, symmetric, and centered at $p$.
- More specifically, it’s an amazing and fortunate fact that a Normal model is just the right one for the histogram of sample proportions.
Sample Proportion

• The symbol for the sample proportion is \( \hat{p} \) (read as “p-hat”).

\[
\hat{p} = \frac{\# \text{ successes}}{\text{sample size}}
\]

• Example – Suppose your sample of 40 automobile drivers contains 26 who use seat belts. Then

\[
\hat{p} = \frac{26}{40} = .65
\]
Example

• A Gallup Poll found that 210 out of a random sample of 501 American teens age 13 to 17 knew the answer to this question:

• “What year did Columbus “discover” America?
Interpretation

• The sample proportion is $\frac{210}{501} = .42$
• Is .42 a parameter or a statistic?
• Does this mean that only 42% of American teens know this fact?
• What is the proper notation for this statistic?
  – $\hat{p} = .42$
Modeling the Distribution of Sample Proportions

• Modeling how sample proportions vary from sample to sample is one of the most powerful ideas we’ll see in this course.

• A **sampling distribution model** for how a sample proportion varies from sample to sample allows us to quantify that variation and how likely it is that we’d observe a sample proportion in any particular interval.

• To use a Normal model, we need to specify its mean and standard deviation. We’ll put \( \mu \), the mean of the Normal, at \( p \).
Modeling the Distribution of Sample Proportions

- When working with proportions, knowing the mean automatically gives us the standard deviation as well—the standard deviation we will use is

\[ \sqrt{\frac{pq}{n}} \]

- So, the distribution of the sample proportions is modeled with a probability model that is

\[ N\left(p, \sqrt{\frac{pq}{n}}\right) \]
Modeling the Distribution of Sample Proportions

- A picture of the Distribution of Sample Proportions is as follows:
Characteristics of the Sampling Distribution of a Sample Proportion

1. The sampling distribution of $\hat{p}$ is approximately normal and is closer to a normal distribution when the sample size $n$ is large.
2. The **mean** of the sampling distribution of $\hat{p}$ is exactly $p$ (the population parameter).

3. The **standard deviation** of the sampling distribution of $\hat{p}$ is:

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1 - p)}{n}}$$
Summary: Sampling Distribution of a Sample Proportion
Formulas for Sampling Proportions

\[ \mu_{\hat{p}} = p \]

\[ \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{pq}{n}} \]
Sample Proportions – Sampling Variability

- Because we have a Normal model, for example, we know that 95% of Normally distributed values are within two standard deviations of the mean.
- So we should not be surprised if 95% of various polls gave results that were near the mean but varied above and below that by no more than two standard deviations.
- This is what we mean by **sampling error**. It’s not really an error at all, but just variability you’d expect to see from one sample to another. A better term would be **sampling variability**.
Behavior of $\hat{p}$

- Because the mean of the sampling distribution of $\hat{p}$ is always equal to the parameter $p$, the sample proportion $\hat{p}$ is an **unbiased estimator** of $p$.
- The standard deviation of $\hat{p}$ gets smaller as the sample size $n$ increases because $n$ is in the denominator of the formula for standard deviation. That is, $\hat{p}$ is less variable in larger samples.
- $\hat{p}$ is less variable in larger samples.
Assumptions and Conditions

• Most models are useful only when specific assumptions are true.
• There are two assumptions in the case of the model for the distribution of sample proportions:
  1. The Independence Assumption: The sampled values must be independent of each other.
  2. The Sample Size Assumption: The sample size, \( n \), must be large enough.
Assumptions and Conditions

• Assumptions are hard—often impossible—to check. That’s why we assume them.
• Still, we need to check whether the assumptions are reasonable by checking conditions that provide information about the assumptions.
• The corresponding conditions to check before using the Normal model to model the distribution of sample proportions are the Randomization Condition, the 10% Condition and the Success/Failure Condition.
Assumptions and Conditions

1. **Randomization Condition**: The sample should be a simple random sample of the population.
2. **10% Condition**: the sample size, $n$, must be no larger than 10% of the population.
3. **Success/Failure Condition**: The sample size has to be big enough so that both $np$ (number of successes) and $nq$ (number of failures) are at least 10.

...So, we need a large enough sample that is not too large.
Condition Thumb Rules for Sample Proportions

1. Use the standard deviation of \( \hat{p} \)

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

_only_ when the population is at least 10 times as large as the sample, that is, when \( N \geq 10n \).

2. To use the Normal approximation to the sampling distribution of \( \hat{p} \), must satisfy \( np \geq 10 \) and

\[ n(1-p) \geq 10. \]
Problem:

• Records show that 65% of visitors will spend money in the gift shop at a local attraction. If there are 525 visitors on a given day, what is the probability that at least 70% of these visitors will purchase something in the gift shop?
Solution:

- State what we want to know.
- Check the Thumb Rules

We want to find the probability that in a group of 525 visitors, 70% or more would make a purchase in the gift shop.

- \( N \geq 10n \) - The 525 visitors can be considered a random sample of visitors and it is reasonable to expect the attraction will draw at least 5250 visitors. Can use;

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

\( np \geq 10 \) and \( n(1-p) \geq 10 \)

\( np = 525(.65) = 341.25 > 10 \)

\( n(1-p) = 525(.35) = 183.75 > 10 \)

Can use the Normal approximation.
• State the parameters and the sampling distribution model.

• The population proportion is $p = .65$. The mean of the normal model for $\hat{p}$ is .65 (i.e. the mean of the sampling distribution of $\hat{p}$ equals $p$). The standard deviation is

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{(.65)(.35)}{525}} \approx .0208.$$ 

The model for $\hat{p}$ is $N(.65,.0208)$.

• Make a picture
• State the problem in terms of \( \hat{p} \).

• Convert to a \( z \)-score.

• Find the resulting probability.

• Discuss the probability in the context of the question.

• \( P(\hat{p} \geq .70) \)

\[
P \left( z \geq \frac{.70 - .65}{.0208} \right) = P(z \geq 2.40)
\]

• \( P(z \geq 2.40) = .0082 \)

• There is a probability of about .0082 that 70% or more visitors will buy something in the gift shop.
Your Turn:

• The Census Bureau reports that 40% of the 50,000 families in a particular region have more than one color TV in their household. What is the probability that a simple random sample of size 100 will indicate 45% or more households with more than one color TV?
Solution:

- Verify thumb rules:
  
  \[ N \geq 10n \quad 50,000 \geq 10(100) \]
  
  \[ np \geq 10 \text{ and } n(1-p) \geq 10 \]
  
  \[(100)(.4) = 40; (100)(.6) = 60\]

- Therefore, the sampling distribution of is

  \[ N\left(p, \sqrt{\frac{p(1-p)}{n}}\right) = N(.4,.049)\]
• Draw a picture.

• State the problem in terms of $\hat{p}$.
  
  \[ P(\hat{p} \geq 0.45) \]

• Convert to a z-score.
  
  \[ P \left( z \geq \frac{0.45 - 0.4}{0.049} \right) = P(z \geq 1.02) \]

• Find the resulting probability.
  
  \[ P(z \geq 1.02) = 0.1537 \]
Different Example:

- An SRS of 1500 first-year college students were asked whether they applied for admission to any other college. In fact, 35% of all first-year students applied to colleges beside the one they are attending.

- What is the probability that the poll will be within 2 percentage points of the true p?
Solution:

- State what we want to know.

- Check the Thumb Rules

- We want to find the probability that the poll shows between 33% and 37% of first-year students applied to colleges beside the one they are attending.

- \( N \geq 10n - 10n = 15000 \) and it is reasonable to assume there are more than 15000 first-year students. Can use;

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

\( np \geq 10 \) and \( n(1-p) \geq 10 \)

\( np = 1500(.35) = 525 > 10 \)

\( n(1-p) = 1500(.65) = 975 > 10 \)

Can use the Normal approximation.
- State the parameters and the sampling distribution model.

- The population proportion is $p = .35$. The mean of the normal model for $\hat{p}$ is .35 (ie. the mean of the sampling distribution of $\hat{p}$ equals $p$). The standard deviation is

$$
\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{.35(.65)}{1500}} = .01232
$$

The model for $\hat{p}$ is $N(.35,.01232)$.

- Make a picture

- School note...
• State the problem in terms of \( \hat{p} \).

• Convert to a z-score.

• Find the resulting probability.

• Conclusion

\[ P(0.33 \leq \hat{p} \leq 0.37) \]

\[ P(-1.626 \leq z \leq 1.626) \]

\[ 0.8968 \]

• About 90% of the samples fall within 2% of the real \( p \).
A Sampling Distribution Model for a Proportion

- A proportion is no longer just a computation from a set of data.
  - It is now a random variable quantity that has a probability distribution.
  - This distribution is called the sampling distribution model for proportions.

- Even though we depend on sampling distribution models, we never actually get to see them.
  - We never actually take repeated samples from the same population and make a histogram. We only imagine or simulate them.
Properties of the Sampling Distribution of the Sample Proportion

If a random sample of size $n$ is selected from a population with proportion of successes $p$, then the sampling distribution of $\hat{p}$ has these properties:

- The mean of the sampling distribution is equal to the mean of the population, or $\mu_{\hat{p}} = p$.
- The standard error of the sampling distribution is equal to the standard deviation of the population divided by the square root of the sample size:

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

- As the sample size gets larger, the shape of the sampling distribution gets more normal and will be approximately normal if $n$ is large enough.

As a guideline, if both $np$ and $n(1-p)$ are at least 10, then using the normal distribution as an approximation for the shape of the sampling distribution will give reasonably accurate results.
School Note...

Sample Means
What About Quantitative Data?

• Proportions summarize categorical variables.
• The Normal sampling distribution model looks like it will be very useful.
• Can we do something similar with quantitative data?
• We can indeed. Even more remarkable, not only can we use all of the same concepts, but almost the same model.
Simulating the Sampling Distribution of a Mean

• Like any statistic computed from a random sample, a sample mean also has a sampling distribution.

• We can use simulation to get a sense as to what the sampling distribution of the sample mean might look like…
Means – The “Average” of One Die

- Let’s start with a simulation of 10,000 tosses of a die. A histogram of the results is:
Means – Averaging More Dice

• Looking at the average of two dice after a simulation of 10,000 tosses:

• The average of three dice after a simulation of 10,000 tosses looks like:
Means – Averaging Still More Dice

- The average of 5 dice after a simulation of 10,000 tosses looks like:

- The average of 20 dice after a simulation of 10,000 tosses looks like:
• As the sample size (number of dice) gets larger, each sample average is more likely to be closer to the population mean.
  – So, we see the shape continuing to tighten around 3.5
• And, it probably does not shock you that the sampling distribution of a mean becomes Normal.
Sampling Distribution of the Sample Means

• For a finite population, the sampling distribution of the sample mean is the set of means from all possible samples of a specific size.

• 2 reasons for the popularity of sample means in statistical inference;
  1. Averages are less variable than individual observations.
  2. Averages are more normal than individual observations.
The Fundamental Theorem of Statistics

• The sampling distribution of any mean becomes more nearly Normal as the sample size grows.
  – All we need is for the observations to be independent and collected with randomization.
  – We don’t even care about the shape of the population distribution!

• The Fundamental Theorem of Statistics is called the Central Limit Theorem (CLT).
Central Limit Theorem

• Draw an SRS of size $n$ from any population whatsoever with mean $\mu$ and finite standard deviation $\sigma$. When $n$ is large, the sampling distribution of the sample mean $\bar{x}$ is close to the Normal Distribution $N(\mu, \sigma/\sqrt{n})$ with mean $\mu$ and standard deviation $\sigma/\sqrt{n}$. 
Central Limit Theorem
Central Limit Theorem

As sample size gets large enough (n ≥ 30) ...
Central Limit Theorem

As sample size gets large enough ($n \geq 30$) ... sampling distribution becomes almost normal.
Central Limit Theorem

As sample size gets large enough ($n \geq 30$) ...

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

sampling distribution becomes almost normal.

$\mu_{\bar{x}} = \mu$
Central Limit Theorem

• The CLT is surprising and a bit weird:
  – Not only does the histogram of the sample means get closer and closer to the Normal model as the sample size grows, but this is true regardless of the shape of the population distribution.

• The CLT works better (and faster) the closer the population model is to a Normal itself. It also works better for larger samples.
What This Means:

• For a large sample size $n$, the sampling distribution of $\bar{x}$ is approximately normal for any shape population distribution.

• If the population is Normal then any size sample is “large enough.”

• If the population is not normal, the sample size should be at least 30 for the CLT to apply.
Example CLT

- The sampling distribution of sample means from a strongly non-normal population, as the sample size increases. (a) $n=1$, (b) $n=2$, (c) $n=10$, (d) $n=25$
- As $n$ increases, the shape becomes more normal.
- The mean remains at $\mu = 1$.
- The standard deviation deceases taking the value $1/\sqrt{n}$.
Example of the CLT for (a) normal, (b) reverse-J-shaped, (c) uniform, Population Distributions
The Central Limit Theorem (CLT)

The mean of a random sample is a random variable whose sampling distribution can be approximated by a Normal model. The larger the sample, the better the approximation will be.
How large a sample is needed for the CLT?

- It depends on the population distribution.
- The more normal the population distribution is, the smaller the sample size needed.
- The less normal the population distribution is, the larger the sample size needed.
Assumptions and Conditions

• The CLT requires essentially the same assumptions we saw for modeling proportions:
  ▪ Independence Assumption: The sampled values must be independent of each other.
  ▪ Sample Size Assumption: The sample size must be sufficiently large.
Assumptions and Conditions

• We can’t check these directly, but we can think about whether the Independence Assumption is plausible. We can also check some related conditions:
  – Randomization Condition: The data values must be sampled randomly.
  – 10% Condition: When the sample is drawn without replacement, the sample size, $n$, should be no more than 10% of the population.
  – Large Enough Sample Condition: The CLT doesn’t tell us how large a sample we need. For now, you need to think about your sample size in the context of what you know about the population.
Characteristics of the Sampling Distribution of a Sample Mean

1. The sampling distribution of $\bar{x}$ is approximately normal and is closer to a normal distribution as the sample size $n$ gets larger.

2. The mean of the sampling distribution of $\bar{x}$ is exactly $\mu$ (the population parameter).

3. The standard deviation of the sampling distribution of $\bar{x}$ is:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$
Summary: Sampling Distribution of a Sample Mean

Population
Mean $\mu$
Std. dev. $\sigma$

Values of $\bar{x}$
Behavior of $\bar{x}$

- The sample mean $\bar{x}$ is an unbiased estimator of the population mean $\mu$.
- The values of $\bar{x}$ are less spread out for larger samples.
- Use the equation $\sigma_{\bar{x}} = \sigma / \sqrt{n}$ for the standard deviation of $\bar{x}$ only when the population is at least 10 times as large as the sample ($N \geq 10n$).
Thumb Rules for Sample Means

1. Use the standard deviation of $\bar{x}$

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

only when the population is at least 10 times as large as the sample, that is, when $N \geq 10n$.

2. To use the Normal approximation to the sampling distribution of $\bar{x}$, must satisfy the Central Limit Theorem (normally, $n \geq 30$).
Problem:

• Of 500 people attending an international convention, it was determined that the average distance traveled by the conventioneers was 1917 miles with a standard deviation 2500 miles. What is the probability that a random sample of 40 of the attendees would have traveled an average distance of 900 miles or less to attend this convention?
Solution:

- State what we want to know.
- Check the conditions (thumb rules).
- Find the probability that in a group of 40 attendees, the average distance traveled to a convention is 900 miles or less.

\[ N \geq 10n \quad N=500, \quad 10n=10(40)=400 \]
\[ 500 \quad 400. \text{ Can use; } \sigma_{\bar{x}} = \sigma/\sqrt{n} \]

\[ n = 40 \quad \text{(large, } n>30) \quad \text{CLT tells use the sampling distribution will be normal.} \]
• State the parameters and the sampling distribution model.

\[ \mu_{\bar{x}} = \mu = 1917 \]
\[ \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2500}{\sqrt{40}} \approx 395.28 \]
therefore, N(1917,395.28)

• Make a picture.

• Write the problem in terms of \( \bar{x} \).

\[ P(\bar{x} \leq 900) \]
• Convert to a z-score.

• Find the resulting probability.

• Discuss the probability in the context of the problem.

\[ P \left( z \leq \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \right) = P \left( z \leq \frac{900 - 1917}{395.28} \right) = -2.57 \]

\[ P(z \leq -2.57) \approx .005 \]

• In a sample of size 40, the probability that the average distance traveled is less than 900 miles is only .005.
Your Turn:

• The census bureau has established that the mean income of heads of household in a particular city is $41,500 with a standard deviation of $18,700. A simple random sample of 100 heads of households in that city indicates that the mean income of the sample individuals is $45,510 with a standard deviation of $23,156. Find the probability that a sample of this size will indicate a mean of $45,510 or more?
Solution:

- Given $\mu = 41,500, \sigma = 18,700, n = 100, \bar{x} = 45,510$

- Verify thumb rules – $N \geq 10n$, $N$ is the population of a city, it is reasonable to assume it is greater than 1,000. Therefore can use $\sigma_{\bar{x}} = \sigma / \sqrt{n}$.

  - Sample size is large ($n \geq 30$), therefore by the CLT the sampling distribution will be normal.

- The sampling distribution model is therefore;

  $$N \left( \mu, \frac{\sigma}{\sqrt{n}} \right) = N \left( 41500, \frac{18700}{\sqrt{100}} \right) = N(41500, 1870)$$
• Write the problem in terms of $\bar{x}$. \( P(\bar{x} \geq 45510) \)

• Convert to a z-score.

\[
P \left( z \geq \frac{\bar{x} - \mu}{\sigma} \right) = P \left( z \geq \frac{45510 - 41500}{1870} \right) = P(z \geq 2.144)
\]

• Draw a picture.

![Diagram of a normal distribution]

• Find the resulting probability. \( P(z \geq 2.144) = .016 \)

• Discuss the probability in the context of the problem.

  – There is a probability of .016 that a sample of 100 will yield a sample mean of $45,510 or higher when the population mean is $41,500.
Review: Sample Means

Properties of the Sampling Distribution of the Sample Mean

If a random sample of size $n$ is selected from a population with mean $\mu$ and standard deviation $\sigma$, then

- the mean $\mu_{\bar{x}}$ of the sampling distribution of $\bar{x}$ equals the mean of the population $\mu$:

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

- the standard deviation $\sigma_{\bar{x}}$ of the sampling distribution of $\bar{x}$, sometimes called the **standard error of the mean**, equals the standard deviation of the population $\sigma$ divided by the square root of the sample size $n$:

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

- the shape of the sampling distribution will be approximately normal if the population is approximately normal; for other populations, the sampling distribution becomes more normal as $n$ increases. This property is called the **Central Limit Theorem**.
But Which Normal?

• The CLT says that the sampling distribution of any mean or proportion is approximately Normal.

• But which Normal model?
  – For proportions, the sampling distribution is centered at the population proportion.
  – For means, it’s centered at the population mean.

• But what about the standard deviations?
But Which Normal?

- The Normal model for the sampling distribution of the mean has a standard deviation equal to

$$SD(\bar{y}) = \frac{\sigma}{\sqrt{n}}$$

where $\sigma$ is the population standard deviation.
But Which Normal?

- The Normal model for the sampling distribution of the proportion has a standard deviation equal to

\[ SD(\hat{p}) = \sqrt{\frac{pq}{n}} = \frac{\sqrt{pq}}{\sqrt{n}} \]
About Variation

• The standard deviation of the sampling distribution declines *only* with the square root of the sample size (the denominator contains the square root of \( n \)).
• Therefore, the variability decreases as the sample size increases.
• While we’d always like a larger sample, the square root limits how much we can make a sample tell about the population. (This is an example of the Law of Diminishing Returns.)
The Real World and the Model World

Be careful! Now we have \textit{two} distributions to deal with.

- The first is the real world distribution of the sample, which we might display with a histogram.
- The second is the math world \textit{sampling distribution} of the statistic, which we model with a Normal model based on the Central Limit Theorem.

Don’t confuse the two!
Sampling Distribution Models

• Always remember that the statistic itself is a random quantity.
  – We can’t know what our statistic will be because it comes from a random sample.
• Fortunately, for the mean and proportion, the CLT tells us that we can model their sampling distribution directly with a Normal model.
Sampling Distribution Models

- There are two basic truths about sampling distributions:
  1. Sampling distributions arise because samples vary. Each random sample will have different cases and, so, a different value of the statistic.
  2. Although we can always simulate a sampling distribution, the Central Limit Theorem saves us the trouble for means and proportions.
The Process Going Into the Sampling Distribution Model
What Can Go Wrong?

• Don’t confuse the sampling distribution with the distribution of the sample.
  – When you take a sample, you look at the distribution of the values, usually with a histogram, and you may calculate summary statistics.
  – The sampling distribution is an imaginary collection of the values that a statistic *might* have taken for all random samples—the one you got and the ones you didn’t get.
What Can Go Wrong?

• Beware of observations that are not independent.
  – The CLT depends crucially on the assumption of independence.
  – You can’t check this with your data—you have to think about how the data were gathered.

• Watch out for small samples from skewed populations.
  – The more skewed the distribution, the larger the sample size we need for the CLT to work.
What have we learned?

- Sample proportions and means will vary from sample to sample—that’s sampling error (sampling variability).
- Sampling variability may be unavoidable, but it is also predictable!
What have we learned?

• We’ve learned to describe the behavior of sample proportions when our sample is random and large enough to expect at least 10 successes and failures.

• We’ve also learned to describe the behavior of sample means (thanks to the CLT!) when our sample is random (and larger if our data come from a population that’s not roughly unimodal and symmetric).