STA 2023

Module 6
The Sampling Distributions

Module Objectives

In this module, we will learn the following:

1. Define sampling error and explain the need for sampling distributions.
2. Recognize that sampling variability may be unavoidable, but it is also predictable.
3. State and apply the central limit theorem.
4. Describe the behavior of sample proportions when our sample is random and large enough to expect at least 10 successes and failures.
5. Describe the behavior of sample means when our sample is random and normally distributed or the sample size is relatively large (if our data come from a population that's not roughly unimodal and symmetric).

Why Sampling?

We have seen that using a sample to acquire information about a population is often preferable to conducting a census.

Generally, sampling is less costly and can be done more quickly than a census; it is often the practical way to gather information.

What is Sampling Error?

However, because a sample provides data for only a portion of an entire population, we cannot expect the sample to yield perfectly accurate information about the population. Thus, we should anticipate that a certain amount of error - called sampling error - will result simply because we are sampling.

Sampling Error

Sampling error is the error resulting from using a sample to estimate a population characteristic.

Let's Look at The Central Limit Theorem for 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?

Distribution of Sample Proportions

We would expect the histogram of the sample proportions to center at the true proportion, *p*, in the population. [Note that a true proportion is also known as population proportion, it is simply the percentage of a population that has a specified attribute.]

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.

Modeling the Distribution for 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 μ, the mean of the Normal, at p.

Modeling the Distribution for Sample Proportions (Cont.)

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

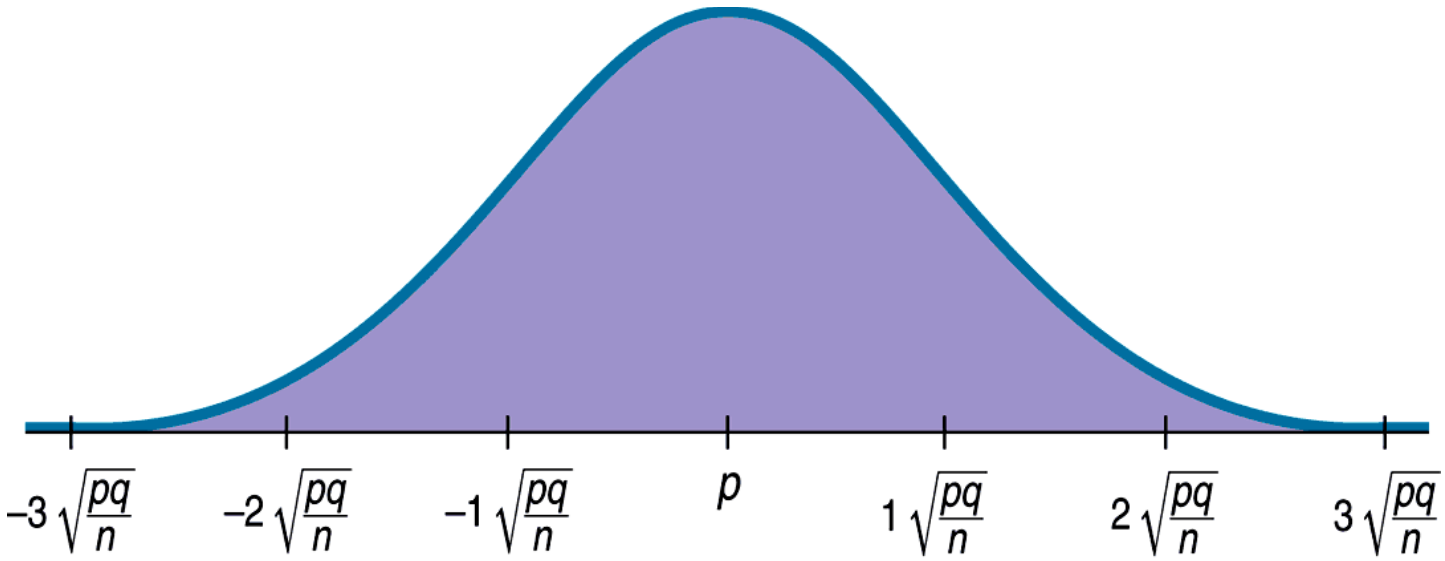
$$\sqrt{\frac{pq}{n}}$$

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

$$N\left(p, \sqrt{\frac{pq}{n}}\right)$$

Modeling the Distribution for Sample Proportions (Cont.)

A picture of what we just discussed is as follows:



The Central Limit Theorem for Sample Proportions (Cont.)

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**.

How Good is the Normal Model?

The Normal model gets better as a good model for the distribution of sample proportions as the sample size gets bigger.

Just how big of a sample do we need? This will soon be revealed…

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 (Cont.)

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 to model the distribution of sample proportions are the Randomization Condition, the 10% Condition and the Success/Failure Condition.

Assumptions and Conditions (Cont.)

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.

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.

A Sampling Distribution Model for a Proportion (Cont.)

Still, sampling distribution models are important because

- they act as a bridge from the real world of data to the imaginary world of the statistic and
- enable us to say something about the population when all we have is data from the real world.

The Sampling Distribution Model for a Proportion (Cont.)

Provided that the sampled values are independent and the sample size is large enough, the sampling distribution of $\hat{p}$ is modeled by a Normal model with

Mean: $\mu(\hat{p}) = p$

Standard deviation: $SD(\hat{p}) = \sqrt{\frac{pq}{n}}$

What About Quantitative Data?

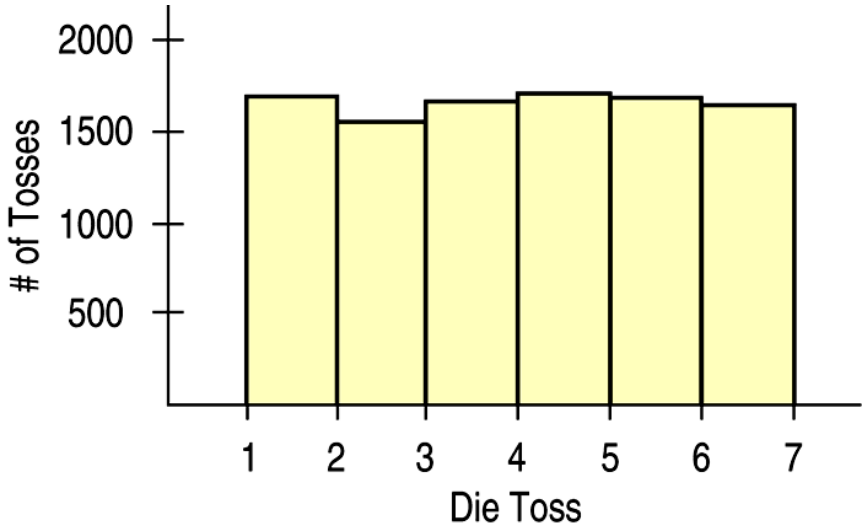
- Proportions summarize categorical variables.
- The Normal sampling distribution model looks like it will be very useful.
- Can we do something similar with quantitative data?
- We can indeed. Even more remarkable, not only can we use all of the same concepts, but almost the same model.

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…

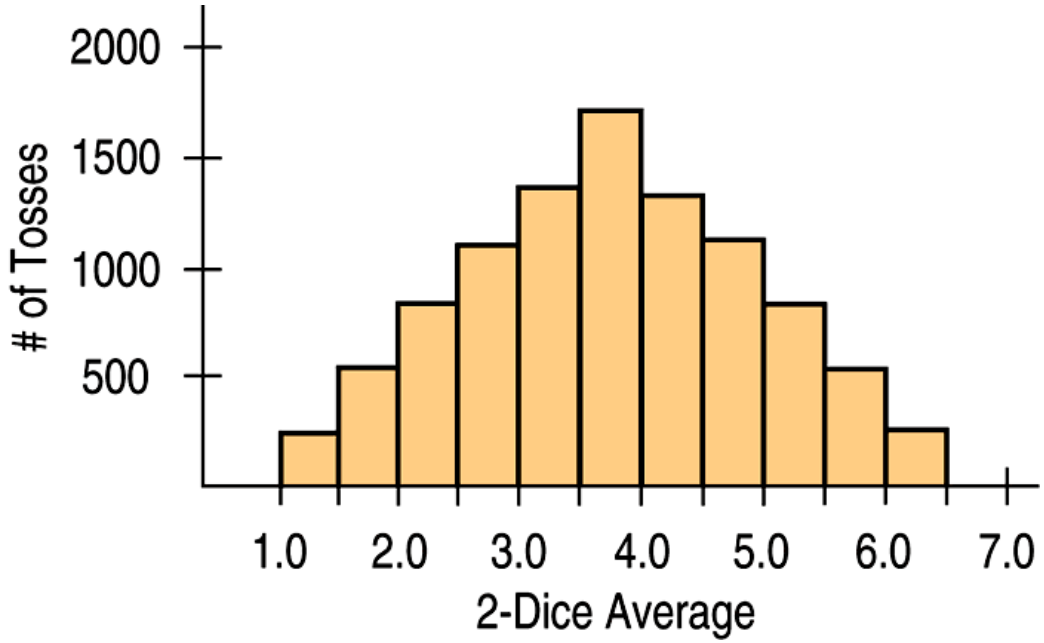
Means – The "Average" of One Die

- Now, let's go through a simulation of 10,000 tosses of a die. A histogram of the results is:



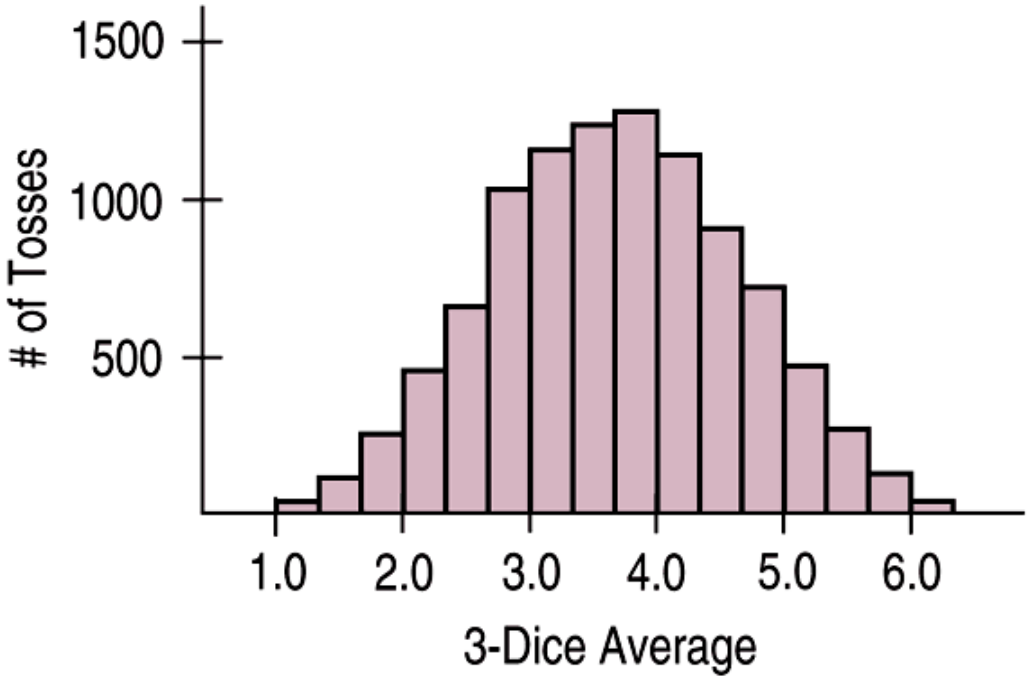
Means – Averaging Two Dice

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



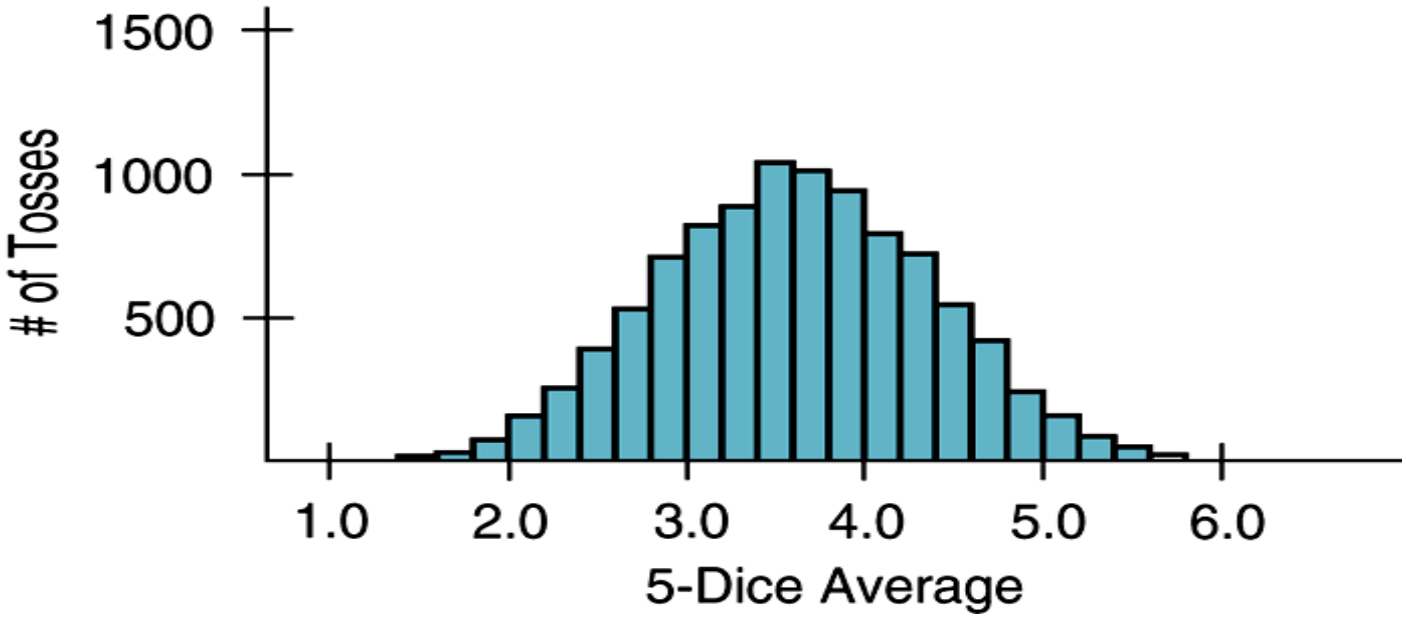
Means – Averaging Three Dice

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



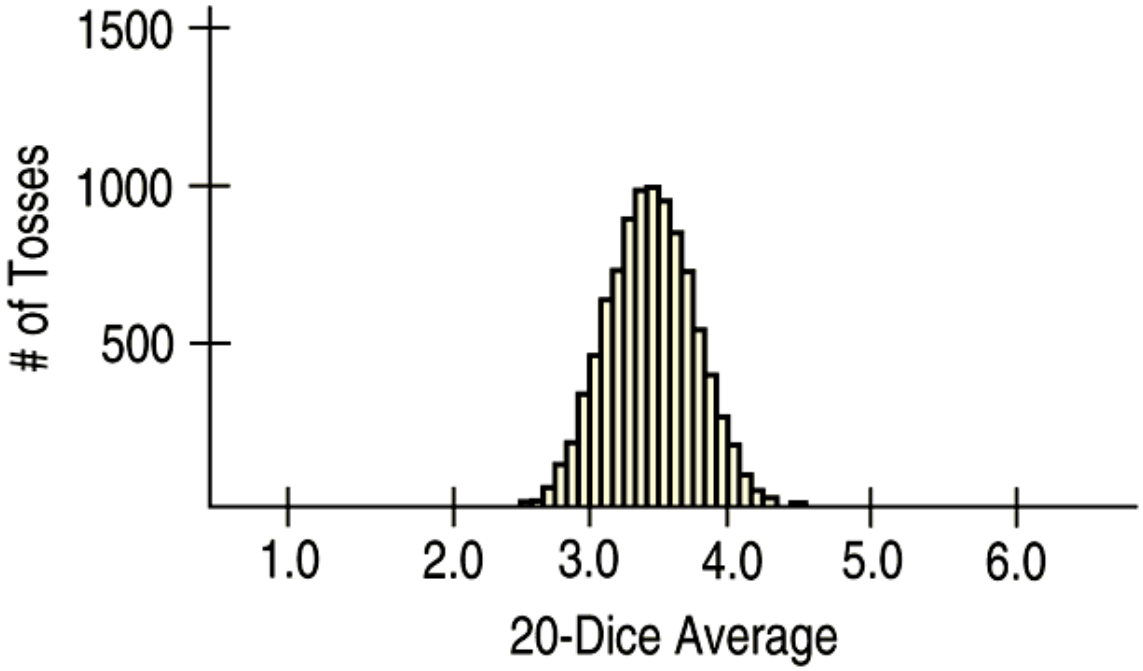
Means – Averaging Five Dice

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



Means – Averaging Twenty Dice

- The average of 20 dice after a simulation of 10,000 tosses looks like:



What the Simulations Show?

- As the sample size (number of dice) gets larger, each sample average (sample mean) is more likely to be closer to the true mean (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 Mean

Sampling Distribution of the Sample Mean

For a variable x and a given sample size, the distribution of the variable $\bar{x}$ is called the **sampling distribution of the sample mean.**

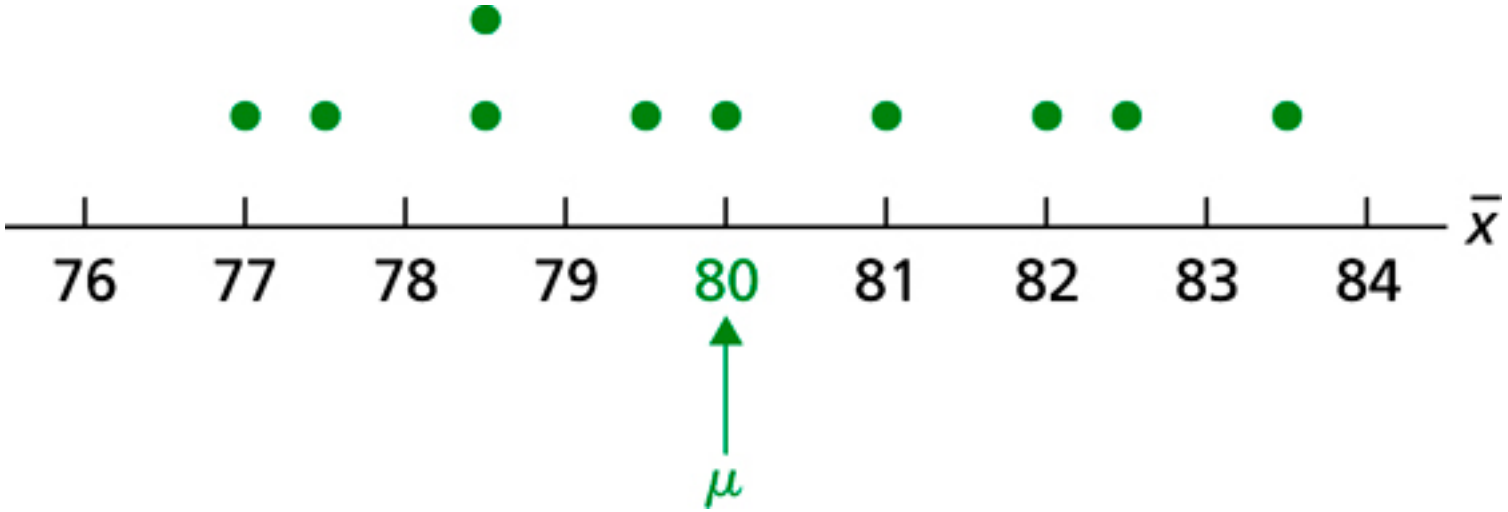
What does it mean?

The sampling distribution of the sample mean is the distribution of all possible sample means for samples of a given size.

Another Example:
Sample Means for Samples of Size 2

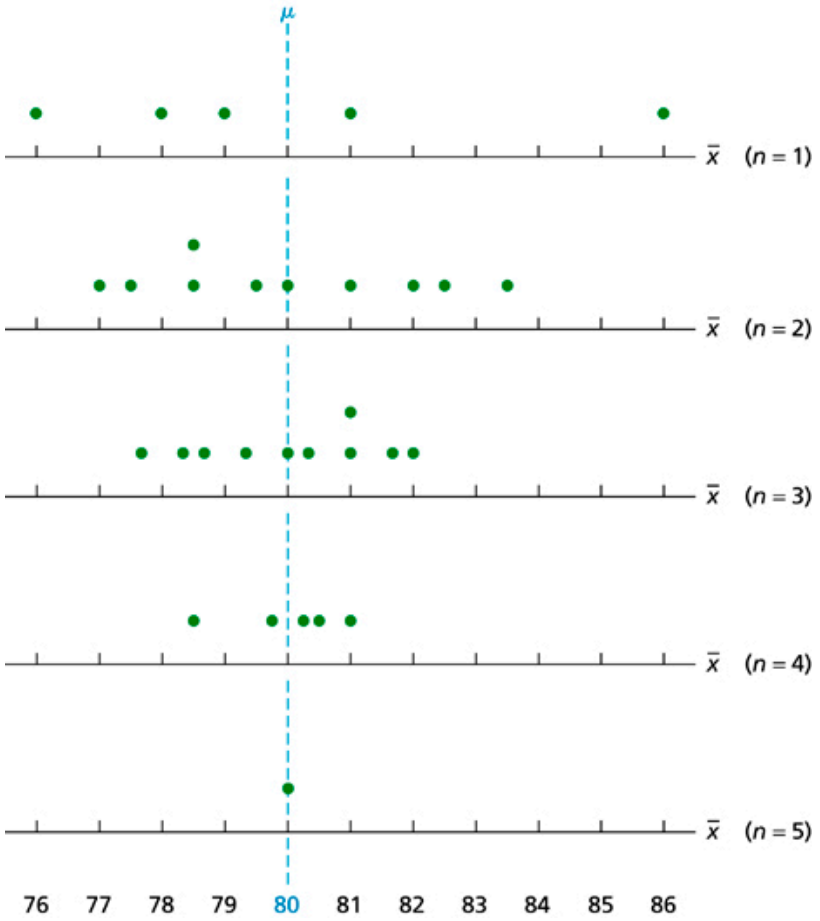
Sample	Heights	$\bar{x}$
A, B	76, 78	77.0
A, C	76, 79	77.5
A, D	76, 81	78.5
A, E	76, 86	81.0
B, C	78, 79	78.5
B, D	78, 81	79.5
B, E	78, 86	82.0
C, D	79, 81	80.0
C, E	79, 86	82.5
D, E	81, 86	83.5

Dotplot for the Sampling Distribution



This is a dotplot for the sampling distribution of the sample mean for samples of size 2 in the previous example. The dotplot shows that 3 out of 10 samples have means within 1 inch of the population mean of 80 inches.

Dotplot for the Sampling Distribution of the Sample Mean



As we can see here, the possible sample means cluster more closely around the population mean as the sample size increases. This suggests that sampling error tends to be smaller for large samples than for small samples.

Sample Size and Sampling Error

Sample size n	No. possible samples	No. within 1″ of μ	% within 1″ of μ	No. within 0.5″ of μ	% within 0.5″ of μ
1	5	2	40%	0	0%
2	10	3	30%	2	20%
3	10	5	50%	2	20%
4	5	4	80%	3	60%
5	1	1	100%	1	100%

As we can see here, the larger the sample size (first column), the percentage of sample means lie within half an inch from the population mean (last column) is getting larger. This means that the larger the sample size, the smaller the sampling error.

What is the Mean of the Sample Mean?

Mean of the Sample Mean

For samples of size n, the mean of the variable $\bar{x}$ equals the mean of the variable under consideration. In symbols,

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

In short, the mean of all possible sample means equal to the population mean.

What is the Standard Deviation of the Sample Mean?

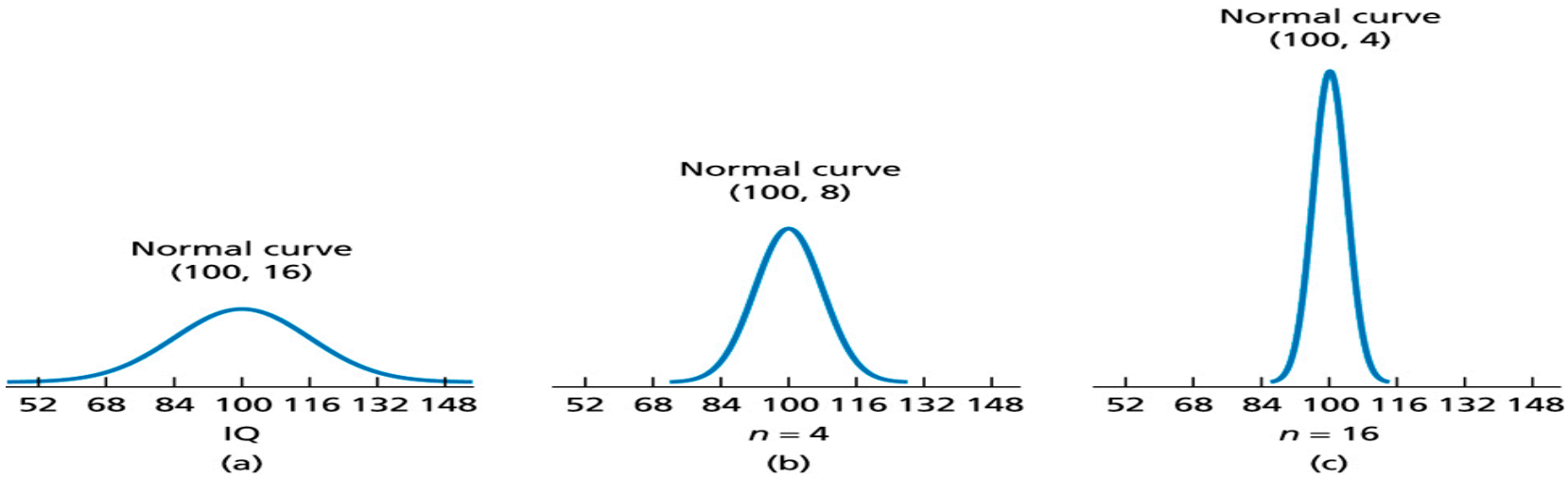
Standard Deviation of the Sample Mean

For samples of size n, the standard deviation of the variable $\bar{x}$ equals the standard deviation of the variable under consideration divided by the square root of the sample size. In symbols,

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

In short, for each sample size, the standard deviation of all possible sample means equals to the population standard deviation divided by the square root of the sample size.

Sampling Distribution of the Sample Mean for a Normally distributed Variable



The possible sample mean IQs for samples of four people have a normal distribution with mean 100 and standard deviation 8, whereas the possible sample mean IQs for samples of 16 people have a normal distribution with mean 100 and standard deviation 4.

Thus, the larger the sample size, the smaller the sampling error tends to be in estimating a population mean by sample mean.

The Central Limit Theorem:
The Fundamental Theorem of Statistics

- Thus, 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).

The Central Limit Theorem (cont.)

- 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.

The Central Limit Theorem (CLT)

In general, the mean of a random sample has a sampling distribution whose shape can be approximated by a Normal model. The larger the sample, the better the approximation will be.

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 (Cont.)

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.

Which Normal Model?

- 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 (true proportion).
  - For means, it's centered at the population mean (true mean).
- But what about the standard deviations?

Which Normal Model? (cont.)

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

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

where σ is the population standard deviation.

Which Normal Model? (cont.)

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

$$SD(\hat{p}) = \sqrt{\frac{pq}{n}} = \frac{\sqrt{pq}}{\sqrt{n}}$$

Standard Deviation and Standard Error

- 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.)

Standard Deviation and Standard Error (Cont.)

- When we don't know the population standard deviation σ, are we stuck?
- Nope! We can use sample statistics (like sample means or sample proportions) to estimate these population parameters (true mean or true proportions).
- Whenever we estimate the standard deviation of a sampling distribution, we call it a standard error.

The Real World and the Model World

Now we have *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 sampling distribution of any mean or proportion, which we model with a Normal model based on the Central Limit Theorem.

Don't confuse the two!

Sampling Distribution Models

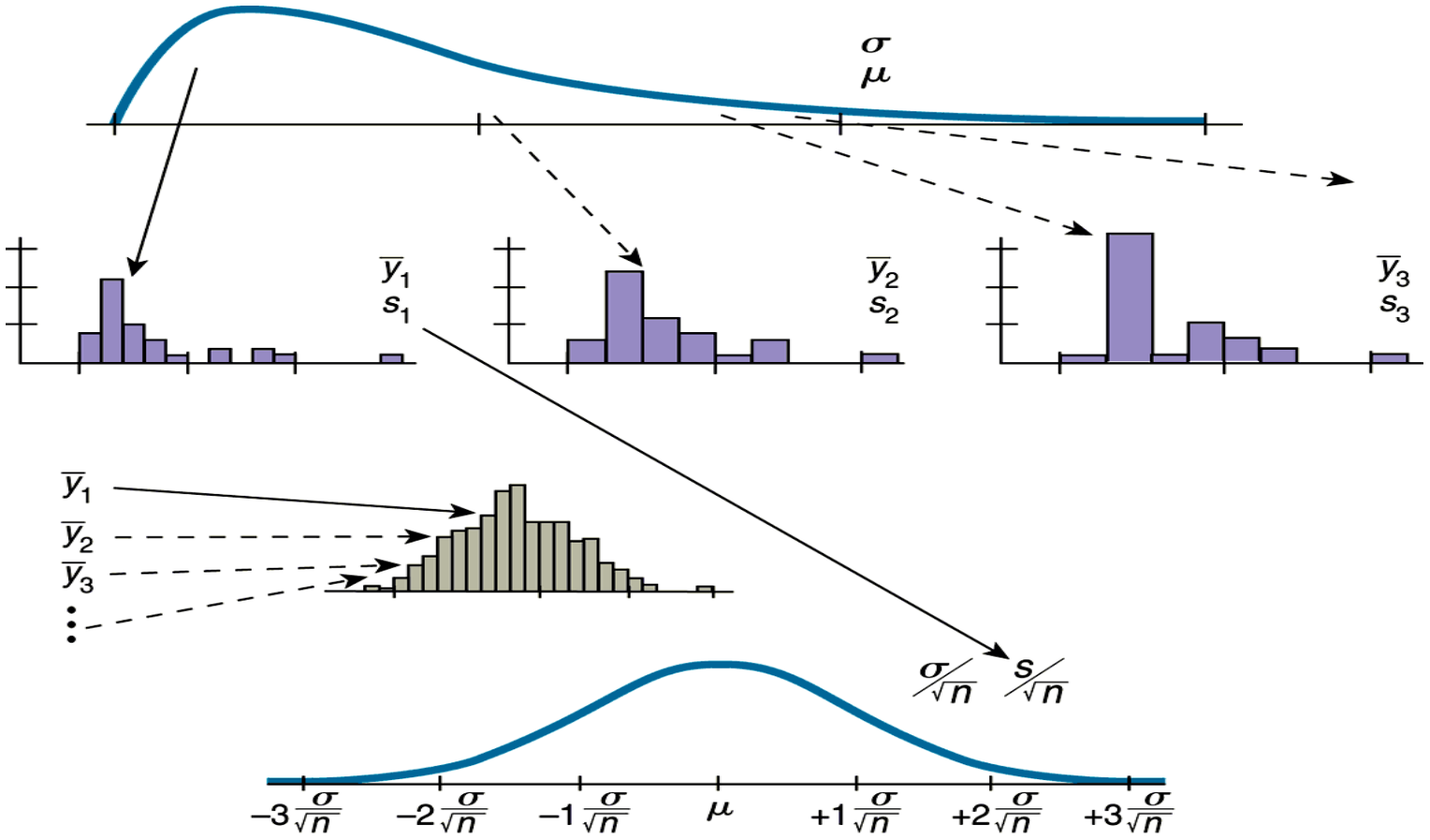
- Always remember that *the statistic itself is a random quantity.*
  - We can't know what our statistic will be because it comes from a random sample.
- Fortunately, for the mean, the CLT tells us that we can model their sampling distribution directly with a Normal model.

Sampling Distribution Models (cont.)

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



Key Fact

Sampling Distribution of the Sample Mean

Suppose that a variable x of a population has mean μ and standard deviation σ. Then, for samples of size n,

- the mean of $\bar{x}$ equals the population mean, or $\mu_{\bar{x}} = \mu$;
- the standard deviation of $\bar{x}$ equals the population standard deviation divided by the square root of the sample size, or $\sigma_{\bar{x}} = \sigma/\sqrt{n}$;
- if x is normally distributed, so is $\bar{x}$, regardless of sample size; and
- if the sample size is large, $\bar{x}$ is approximately normally distributed, regardless of the distribution of x.

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? (cont.)

- 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?

We have learned to:

1. Define sampling error and explain the need for sampling distributions.
2. Recognize that sampling variability may be unavoidable, but it is also predictable.
3. State and apply the central limit theorem.
4. Describe the behavior of sample proportions when our sample is random and large enough to expect at least 10 successes and failures.
5. Describe the behavior of sample means when our sample is random and normally distributed or the sample size is relatively large (if our data come from a population that's not roughly unimodal and symmetric).

Credit

Some of these slides have been adapted/modified in part/whole from the slides of the following textbooks.

- Weiss, Neil A., Introductory Statistics, 8th Edition
- Bock, David E., Stats: Data and Models, 3rd Edition