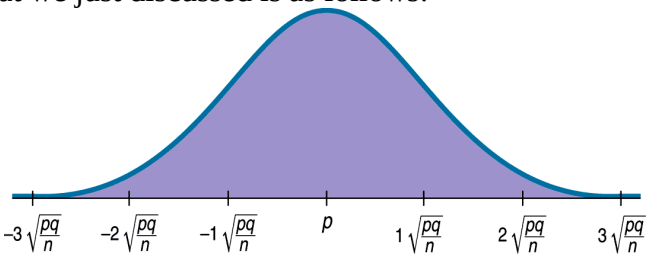
Chapter 18 Summary
Sampling Distribution Models

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!
- 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).

Modeling the Distribution of 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. What would the histogram of all the sample proportions look like?
- 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.
- To use a Normal model, we need to specify its mean and standard deviation. The mean of this particular Normal is at *p*.
- 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 $N\left(p, \sqrt{\frac{pq}{n}}\right)$
- A picture of what we just discussed is as follows:



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 sampled values must be independent of each other.
  2. The sample size, *n*, must be large enough.
- 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 10% Condition and the Success/Failure Condition.

1. 10% condition: If sampling has not been made with replacement, then the sample size, *n*, must be no larger than 10% of the population.
2. Success/failure condition:
   The sample size has to be big enough so that both $n\hat{p}$ and $n\hat{q}$ are greater than 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 quantity that has a 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.
- 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.
- Provided that the sampled values are independent and the sample size is large enough, the sampling distribution of 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?

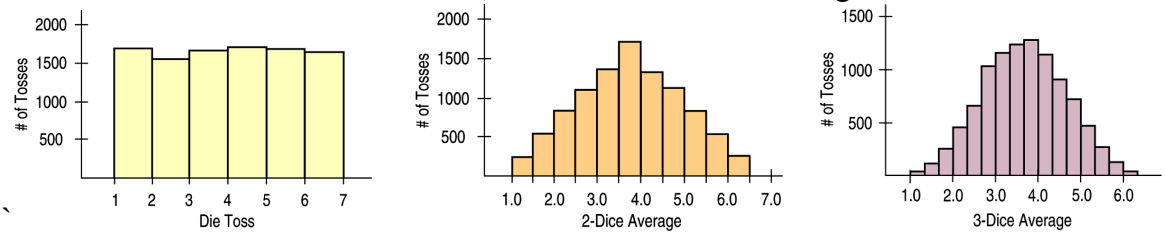
- Proportions summarize categorical variables.
- The Normal sampling distribution model looks like it will be very useful.
- Can we do something similar with quantitative data?
- We can indeed. Even more remarkable, not only can we use all of the same concepts, but almost the same model.

Simulating the Sampling Distribution of a Mean

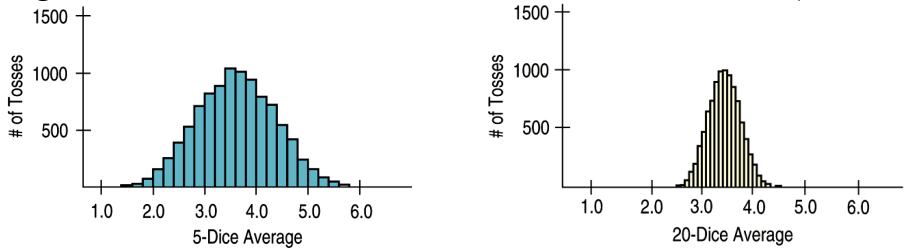
- Like any statistic computed from a random sample, a sample mean also has a sampling distribution.
- We can use simulation to get a sense as to what the sampling distribution of the sample mean might look like…

Means – The "Average" of One Die

- Let's start with a simulation of 10,000 tosses of a die. A histogram of the results is:



- Looking at the average of two dice after a simulation of 10,000 tosses (see above)
- The average of three dice after a simulation of 10,000 tosses looks like (see above)
- The average of 5 dice after a simulation of 10,000 tosses looks like (see below)
- The average of 20 dice after a simulation of 10,000 tosses looks like (see below)



Means – What the Simulations Show

- 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.

The Fundamental Theorem of Statistics

- The sampling distribution of *any* mean becomes 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 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 Fundamental Theorem of Statistics (cont.)
- **The Central Limit Theorem (CLT)** - 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.

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? (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.
- The Normal model for the sampling distribution of the proportion has a standard deviation equal to $SD(\hat{p}) = \sqrt{\frac{pq}{n}}$

Assumptions and Conditions

- The CLT requires remarkably few assumptions, so there are few conditions to check:
  1. Random Sampling Condition: The data values must be sampled randomly or the concept of a sampling distribution makes no sense.
  2. Independence Assumption: The sample values must be mutually independent. (When the sample is drawn without replacement, check the 10% condition…)
  3. Large Enough Sample Condition: There is no one-size-fits-all rule.

Diminishing Returns

- The standard deviation of the sampling distribution declines *only* with the square root of the sample size.
- 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 Error

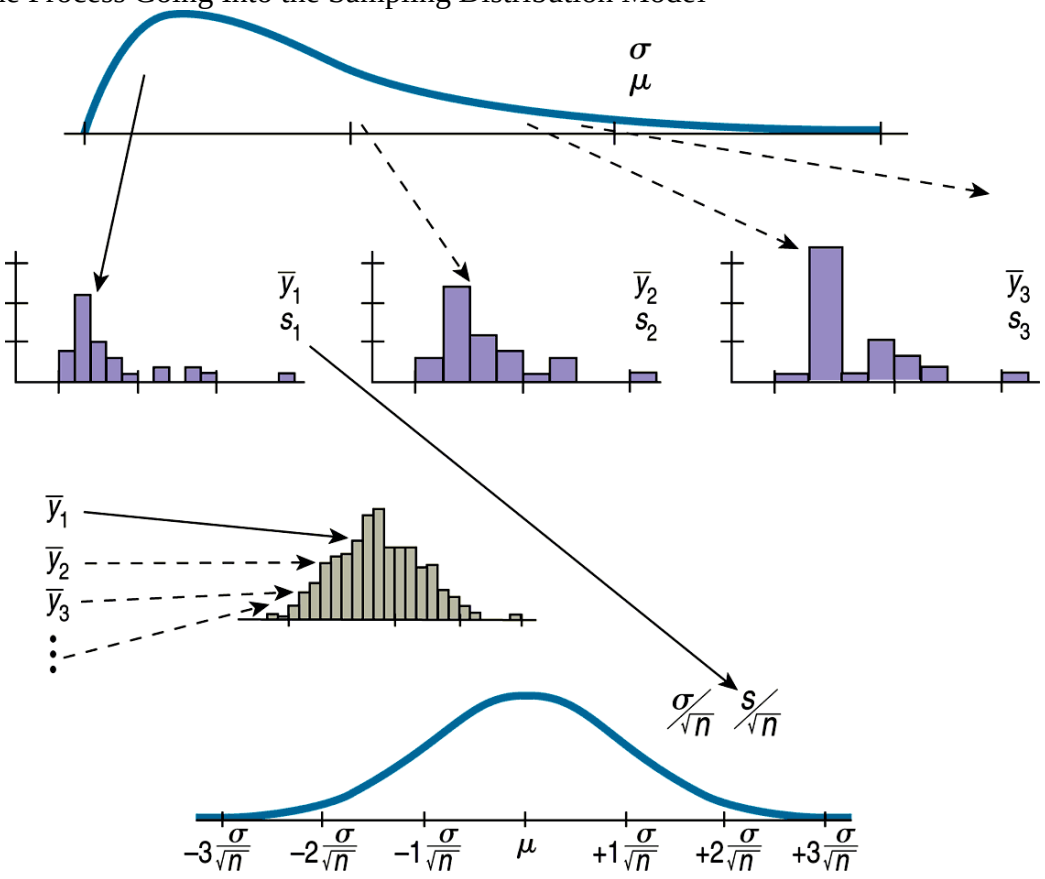
- Both of the sampling distributions we've looked at are Normal.
- For proportions $SD(\hat{p}) = \sqrt{\frac{pq}{n}}$ For means $SD(\bar{y}) = \frac{\sigma}{\sqrt{n}}$
- When we don't know p or σ, we're stuck, right?
  - Nope. We will use sample statistics to estimate these population parameters.
  - Whenever we estimate the standard deviation of a sampling distribution, we call it a standard error.
- For a sample proportion, the standard error is $SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}}$
- For the sample mean, the standard error is $SE(\bar{y}) = \frac{s}{\sqrt{n}}$

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 (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



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.