Stat 102 Notes

Table of contents

Course Goals for STAT 102

STAT 102 is an introduction to formal statistical inference. We will carry out inference using both simulation-based approaches and classical, theory-based methods. By the end of the course, you will:

- Read an example where the research question is explicitly stated, and then translate what's stated into a statistical statement involving parameters or other simple distributional characteristics.
- Identify whether the ideal data collection strategy would involve random assignment, random sampling, or both and explain why.
- Work with an example where the research question is explicitly stated, along with an existing data set, and propose and carry out an appropriate analysis to answer the research question.
- Explain the terms/components of a given statistical model, and connect those terms to the research question at hand.
- Check basic assumptions of various (simple) analysis methods and justify the use of the method.
- Apply existing functions and point-and-click software for implementing basic data analyses.
- Use tactile simulation to carry out a simple resampling procedure.
- Identify the steps and perform the calculations required for routine statistical procedures to address a given problem.
- Calculate simple analyses (t-test, chi-squared test for proportions) by hand, to verify the validity of the computational algorithm.
- Recognize when computational results do not make sense in the context of the problem.

1 Philosophy of Statistical Inference (Chapters 11-14)

In STAT 101, you focused on Exploratory Data Analysis. Exploratory data analysis aims to investigate the characteristics of a data set through visualizations and numerical summaries. Visualizations may include:

- box plots
- histograms
- bar charts
- pie charts
- scatterplots
- heat maps
-
-
-

Numerical summaries used to explore a data set may include:

- sample mean
- sample variance/standard deviation
- five number summary, and other order statistics
- sample proportions
- calculated regression slope and intercept
-
-
-

More often than not, the data were collected to answer a research question about a larger population for which the data collected are a (hopefully) representative sample. This notion of drawing conclusions beyond the data collected is at the heart of statistical inference.

Example: Bred in the Bone

- If each baby is really guessing/choosing blindly, what proportion would you expect to choose the good guy? Why?

- Based on this, what randomizing device could we use to model this experiment?

- Experiment! Add your results to the plot on the board.

- Applet

- What do you observe in the plot?

- Real experiment

Take away:

In exploratory data analysis, the visualizations and numerical summaries you choose are driven by the type of data at hand. This is true for statistical inference as well. The type of data will drive the appropriate inference techniques. However, the goal of the research study will also impact the selected method, as will the underlying assumptions of the technique (we'll talk **a lot** more about this). That said, there are some overarching approaches to quantifying variability, and thus drawing conclusions beyond the data set at hand.

Approaches to quantifying variability

- Randomization methods (Chapter 11)
- Bootstrap methods (Chapter 12)
- Mathematical models (Chapter 13)

We'll start the semester by talking about these three approaches fairly generally. For (most of) the rest of the semester, we'll see how these approaches fit with different types of data.

1.1 Randomization Tests (Chapter 11)

The goal of hypothesis tests is to use an **observed** data set to answer a yes/no question about a characteristic of a larger population from which the observed data set was drawn. For example, is swimming with dolphins therapeutic for patients with clinical depression? That is, we want to assess whether or not the explanatory variable causes changes in the response variable.

To answer this question, Antonioli and Reveley (2005) recruited 30 subjects with a clinical diagnosis of mild to moderate depression. The subjects were required to stop all other treatments (therapy and/or pharmaceuticals) 4 weeks prior the experiment, and the 30 subjects were all taken to an island off the coast of Honduras. The subjects were randomly assigned to one of two groups. Both groups spent one hour swimming and snorkeling each day, but one group did so in the presence of dolphins and the other group did not. At the end of two weeks, each subject's level of depressions was evaluated, and whether or not the subjects had a substantial improvement in their depression was recorded.

Explanatory variable:

Response variable:

Is this an observational study or an experiment? What does that imply about inference?

The question we will answer is whether the resulting data provide convincing evidence that subjects who swam with dolphins were more likely to see depression improvement than subjects who swam without dolphins.

If there really is no impact of swimming with dolphins, what does this imply about the explanatory and response variables?

If swimming with dolphins does improve depression, what does this imply about the explanatory and response variables?

This leads to two competing claims:

- **Null hypothesis**: H_0

- **Alternative hypothesis**: H_a

If the null hypothesis is true, how would this manifest in the observed data?

If the alternative hypothesis is true, how would this manifest in the observed data?

We will choose between the competing claims by assessing whether the data conflict so much with H_0 that the null hypothesis cannot be considered reasonable. If this happens, we'll reject the notion of H_0 and conclude that H_a must be true.

Up to now, we haven't seen the data! Here's a summary:

	Dolphin Therapy	Control Group	Total
Showed Improvement			
No Improvement			
Total			

We can see that

-
-

So,

The question remains…is this enough different from what we would expect under the null hypothesis to conclude that swimming with dolphins does make a difference in depression?

So far, nothing we've laid out is unique to a randomization test. Where does randomization come in?

Let's visualize these observations as a set of cards. Each card denotes a subject in the study. The color indicates the response: red for substantial improvement and black for no substantial improvement.

Any difference we see in the simulation is due to chance–the cards were randomly dealt into the dolphin/control groups.

It's not realistic to keep shuffling and dealing by hand…we need to turn to technology to do the randomization for us: Applet

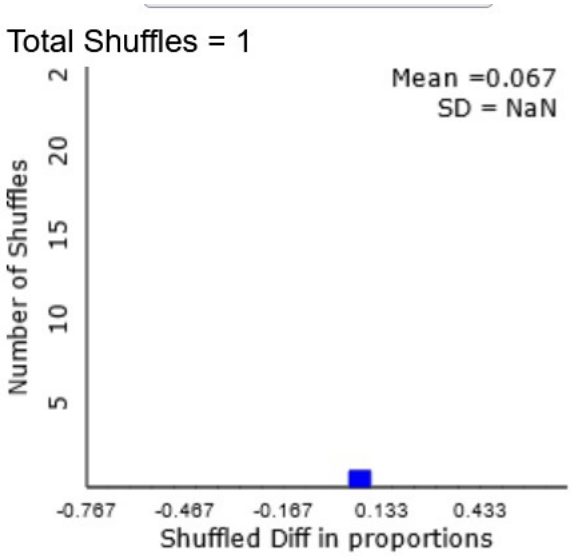

Figure 1.1: One Shuffle

We can do this over and over again to build up a **null distribution**. This distribution shows how we expect the variability to behave under the null hypothesis:

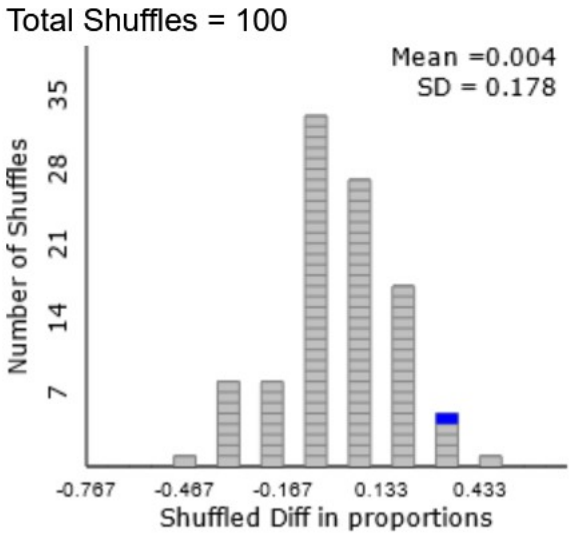

Figure 1.2: 100 Shuffles

What do you notice about this null distribution?

How rare is it to see our observed statistic **0.467** in this distribution? What does this imply?

So, we've just carried out a **statistical inference** technique! We might be wrong in our conclusions (more on this in Chapter 14), but we've made the best decision we could with the data available.

In summary:

Randomization Test Procedure:

- Frame the research question in terms of hypotheses
- Collect data from an observational study or experiment
- Model randomness that would occur if H_0 is true
- Analyze the data by comparing the observed data to the simulated distribution
- Conclusion

Now let's go to R!

1.2 Bootstrap Methods (Chapter 12)

Bootstrap methods are a relatively new statistical technique (proposed in 1979 by Efron), but they are based on a very simple idea. The goal is to characterize the variability of the statistic across many samples. One way we could do this is take lots and lots of samples from the population, and get a picture of how much variance there is among the samples. This is almost always impossible. So, rather than resample from the population, we could try resampling from the sample. This is the basic idea behind the bootstrap.

Bootstrapping is used in many different applications. For this general introduction to the approach, we're going to consider a confidence interval for a proportion.

A **confidence interval** is

Note the goal of the confidence interval is different from the goal of a hypothesis test!

However, like with hypothesis tests, we need to understand the variability inherent to the statistic. To figure out how wide the range of plausible values should be, we need to know how a statistic varies from sample to sample in the population.

For example, let's think back to the Baby scenario and suppose our goal is to estimate the population parameter

The researchers collected one sample of 16 babies, and found that 14 picked the good guy. This is our observed data. What do you think would happen if we took a sample of 16 different babies? And then a different sample of 16 babies?

Idea of the bootstrap:

Infinite populations are pretty tough to work with, though. However, we can produce an equivalent bootstrap distribution by

So, we'll repeatedly draw bootstrap samples of size 16 (why 16?) and calculate the proportion of successes in each bootstrap samples. After we do this many many times, we'll have an idea of a range of plausible values for the population parameter. We'll set the **confidence level** by opting for a wider or a narrower interval, based on how certain we need to be in the results.

Bootstrap Process

- Frame the research question in terms of a parameter to estimate
- Collect data using an observational study or an experiment
- Model the randomness by using the observed data as a proxy for the population
- Create the interval (in future chapters we'll see there are multiple ways to do this)
- Conclusion

Let's go to R!

1.3 Inference with Mathematical Models (Chapter 13)

So far, we've seen computational methods like randomization and bootstrapping to characterize the variability of a statistic. The use of computational methods is relatively recent, due to the increase in computing power. In pre-computing days, re-sampling and randomization was very difficult. As a result, mathematical approximations were used and are still pervasive. If you took AP Statistics or a different intro statistics course, you employed mathematical models. However, to be clear, all of the methods we'll talk about (randomization, bootstrap, mathematical models) are techniques to get a **sampling distribution**.

The sampling distributions we've seen so far have been (mostly):

This isn't coincidence…it's guaranteed by a very important theorem, the **Central Limit Theorem**.

Central Limit Theorem

What are the requirements here?

- Independence:

- "Large enough":

Normal Distribution: Nothing followings it exactly–it's a mathematical construct. But, a lot of things follow it approximately, either:

- naturally:

- created to follow it:

The normal distribution depends on two parameters, μ = mean (where the distribution is centered) and σ = standard deviation (how spread out it is). μ shifts the distribution up and down the number line, σ stretches and contracts the curve. The **standard normal** distribution has $\mu = 0$ and $\sigma = 1$ (this is the distribution tabulated in normal tables in textbooks).

The standard normal gives us a convenient way to compare observations, and any normal distribution can be transformed into a standard normal. The **Z-score** is

If the Z-score is positive

If the Z-score is negative

Z-scores can be used to

- gauge the unusualness of an observation

- find probabilities

Helpful R functions:

- `pnorm(x, mean=0, sd=1)`
- `normTail(m=0,s=1,L=x)` or `normTail(m=0,s=1,U=x)` will draw pretty pictures–need to use the OpenIntro library
- `qnorm(prob, mean=0, sd=1)` gives a Z-score with area to the left

Pictures are super-helpful!

Example: Full-term birth weights for single babies are normally distributed with a mean of 7.5 pounds and a standard deviation of 1.1 pounds.

1. A baby is born weighing 9.1 pounds. What is the weight percentile for this baby?

2. Babies that weigh less than 5.5 pounds are considered low birth weight. What proportion of babies are low birth weight?

3. What weight would make a baby at the 25th percentile?

4. What is the probability a randomly selected baby weighs between 7 and 8 pounds?

The **Empirical Rule** (aka the 68-95-99.7 Rule) presents a general rule for the probability of falling within one, two, and three standard deviations of the mean in a normal distribution.

This rule is useful in a wide range of settings when trying to make quick estimate (we'll use it with bootstraps too!).

Some more definitions we'll use throughout the semester:

- **Standard error**:

- **Margin of error**:

Example (13.11): In 2013, the Pew Research Foundation reported that "45% of US adults report that they live with one ore more chronic conditions." However, this value was based on a sample, so it may not be a perfect estimate for the population parameter of interest on its own. The study reported a standard error of about 1.2%, and a normal model may reasonably be used. Create a 95% confidence interval for the proportion of US adults who live with one or more chronic conditions. Interpret the confidence interval in the context of the study.

1.4 Decision Errors (Chapter 14)

Anytime we're using sample data to make decisions about a larger population we can potentially make a mistake. We can make an incorrect decision in a hypothesis test or calculate a confidence interval that does not capture the true population parameter. In a hypothesis test, there are four possible outcomes:

Type I error:

Type II error:

Examples:

- Doping in the Olympics

- Criminal trial

- Diagnostic test for a serious disease

Errors require a balancing act. We want to reduce the chance of making a Type I error but this will necessarily increase the chance of making a Type II error. The best we can do is to set the probability of a Type I error. We can do through setting the **significance level**.

Significance level:

Another consideration that will impact the chance of making an error is the whether the test is one- or two-sided.

Two-sided hypotheses:

Example: Standard anticoagulant therapy to prevent blood clots requires frequent (expensive) lab monitoring. A new procedure called riva was tested because it did not require frequent monitoring. A randomized trial was conducted in 2012, with standard therapy randomly assigned to 2416 patients and riva randomly assigned to 2416 patients. A bad result was a recurrence of a blood clot in a vein. We want to know if the likelihood of a bad result is different between the two therapies.

Here are the results of the randomized trial

	Riva	Standard	Total
Clot	44	60	104
No Clot	2372	2356	4728
Total	2416	2416	4832

For two-sided tests, the p-value is the probability that we observe a result as least as favorable to the alternative hypothesis as the result we observe. That is, that we observe a result as extreme or more extreme in either direction.

When in doubt, use a two-sided test! Use a one-sided test only if you truly have interest in only one direction.

So, how can we control Type I error?

- Set up tests before seeing the data.
- Collect enough data that the test has sufficient **power**. We'll talk more about power later (and LOTS more in an experimental design course), but power is the probability of correctly rejecting a false null hypothesis. It's a function of how big the true difference is (which we don't know and can't control) and the sample size (which we can control).

2 Philosophy of Statistical Inference (Chapters 11-14)

In STAT 101, you focused on Exploratory Data Analysis. Exploratory data analysis aims to investigate the characteristics of a data set through visualizations and numerical summaries. Visualizations may include:

- box plots
- histograms
- bar charts
- pie charts
- scatterplots
- heat maps
-
-
-

Numerical summaries used to explore a data set may include:

- sample mean
- sample variance/standard deviation
- five number summary, and other order statistics
- sample proportions
- calculated regression slope and intercept
-
-
-

More often than not, the data were collected to answer a research question about a larger population for which the data collected are a (hopefully) representative sample. This notion of drawing conclusions beyond the data collected is at the heart of statistical inference.

Example: Bred in the Bone

- If each baby is really guessing/choosing blindly, what proportion would you expect to choose the good guy? Why?

- Based on this, what randomizing device could we use to model this experiment?

- Experiment! Add your results to the plot on the board.

- Applet

- What do you observe in the plot?

- Real experiment

Take away:

In exploratory data analysis, the visualizations and numerical summaries you choose are driven by the type of data at hand. This is true for statistical inference as well. The type of data will drive the appropriate inference techniques. However, the goal of the research study will also impact the selected method, as will the underlying assumptions of the technique (we'll talk **a lot** more about this). That said, there are some overarching approaches to quantifying variability, and thus drawing conclusions beyond the data set at hand.

Approaches to quantifying variability

- Randomization methods (Chapter 11)
- Bootstrap methods (Chapter 12)
- Mathematical models (Chapter 13)

We'll start the semester by talking about these three approaches fairly generally. For (most of) the rest of the semester, we'll see how these approaches fit with different types of data.

2.1 Randomization Tests (Chapter 11)

The goal of hypothesis tests is to use an **observed** data set to answer a yes/no question about a characteristic of a larger population from which the observed data set was drawn. For example, is swimming with dolphins therapeutic for patients with clinical depression? That is, we want to assess whether or not the explanatory variable causes changes in the response variable.

To answer this question, Antonioli and Reveley (2005) recruited 30 subjects with a clinical diagnosis of mild to moderate depression. The subjects were required to stop all other treatments (therapy and/or pharmaceuticals) 4 weeks prior the experiment, and the 30 subjects were all taken to an island off the coast of Honduras. The subjects were randomly assigned to one of two groups. Both groups spent one hour swimming and snorkeling each day, but one group did so in the presence of dolphins and the other group did not. At the end of two weeks, each subject's level of depressions was evaluated, and whether or not the subjects had a substantial improvement in their depression was recorded.

Explanatory variable:

Response variable:

Is this an observational study or an experiment? What does that imply about inference?

The question we will answer is whether the resulting data provide convincing evidence that subjects who swam with dolphins were more likely to see depression improvement than subjects who swam without dolphins.

If there really is no impact of swimming with dolphins, what does this imply about the explanatory and response variables?

If swimming with dolphins does improve depression, what does this imply about the explanatory and response variables?

This leads to two competing claims:

- **Null hypothesis**: H_0

- **Alternative hypothesis**: H_a

If the null hypothesis is true, how would this manifest in the observed data?

If the alternative hypothesis is true, how would this manifest in the observed data?

We will choose between the competing claims by assessing whether the data conflict so much with H_0 that the null hypothesis cannot be considered reasonable. If this happens, we'll reject the notion of H_0 and conclude that H_a must be true.

Up to now, we haven't seen the data! Here's a summary:

	Dolphin Therapy	Control Group	Total
Showed Improvement			
No Improvement			
Total			

We can see that

-
-

So,

The question remains…is this enough different from what we would expect under the null hypothesis to conclude that swimming with dolphins does make a difference in depression?

So far, nothing we've laid out is unique to a randomization test. Where does randomization come in?

Let's visualize these observations as a set of cards. Each card denotes a subject in the study. The color indicates the response: red for substantial improvement and black for no substantial improvement.

Any difference we see in the simulation is due to chance–the cards were randomly dealt into the dolphin/control groups.

It's not realistic to keep shuffling and dealing by hand…we need to turn to technology to do the randomization for us: Applet

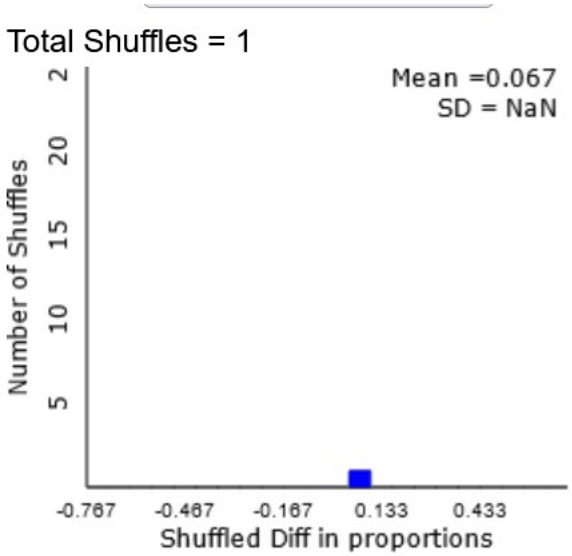

Figure 2.1: One Shuffle

We can do this over and over again to build up a **null distribution**. This distribution shows how we expect the variability to behave under the null hypothesis:

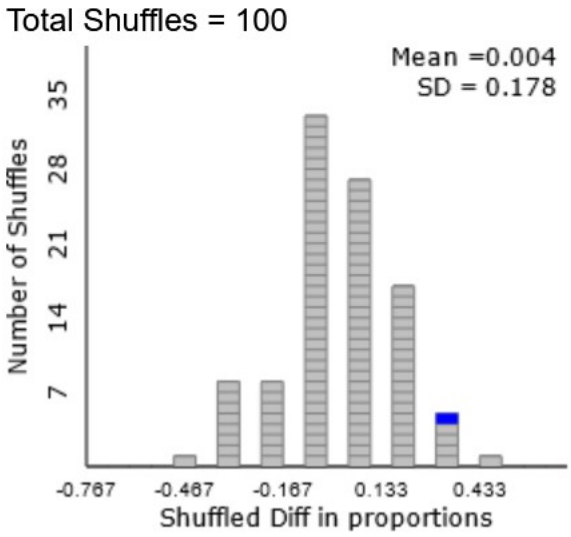

Figure 2.2: 100 Shuffles

What do you notice about this null distribution?

How rare is it to see our observed statistic **0.467** in this distribution? What does this imply?

So, we've just carried out a **statistical inference** technique! We might be wrong in our conclusions (more on this in Chapter 14), but we've made the best decision we could with the data available.

In summary:

Randomization Test Procedure:

- Frame the research question in terms of hypotheses
- Collect data from an observational study or experiment
- Model randomness that would occur if H_0 is true
- Analyze the data by comparing the observed data to the simulated distribution
- Conclusion

Now let's go to R!

2.2 Bootstrap Methods (Chapter 12)

Bootstrap methods are a relatively new statistical technique (proposed in 1979 by Efron), but they are based on a very simple idea. The goal is to characterize the variability of the statistic across many samples. One way we could do this is take lots and lots of samples from the population, and get a picture of how much variance there is among the samples. This is almost always impossible. So, rather than resample from the population, we could try resampling from the sample. This is the basic idea behind the bootstrap.

Bootstrapping is used in many different applications. For this general introduction to the approach, we're going to consider a confidence interval for a proportion.

A **confidence interval** is

Note the goal of the confidence interval is different from the goal of a hypothesis test!

However, like with hypothesis tests, we need to understand the variability inherent to the statistic. To figure out how wide the range of plausible values should be, we need to know how a statistic varies from sample to sample in the population.

For example, let's think back to the Baby scenario and suppose our goal is to estimate the population parameter

The researchers collected one sample of 16 babies, and found that 14 picked the good guy. This is our observed data. What do you think would happen if we took a sample of 16 different babies? And then a different sample of 16 babies?

Idea of the bootstrap:

Infinite populations are pretty tough to work with, though. However, we can produce an equivalent bootstrap distribution by

So, we'll repeatedly draw bootstrap samples of size 16 (why 16?) and calculate the proportion of successes in each bootstrap samples. After we do this many many times, we'll have an idea of a range of plausible values for the population parameter. We'll set the **confidence level** by opting for a wider or a narrower interval, based on how certain we need to be in the results.

Bootstrap Process

- Frame the research question in terms of a parameter to estimate
- Collect data using an observational study or an experiment
- Model the randomness by using the observed data as a proxy for the population
- Create the interval (in future chapters we'll see there are multiple ways to do this)
- Conclusion

Let's go to R!

2.3 Inference with Mathematical Models (Chapter 13)

So far, we've seen computational methods like randomization and bootstrapping to characterize the variability of a statistic. The use of computational methods is relatively recent, due to the increase in computing power. In pre-computing days, re-sampling and randomization was very difficult. As a result, mathematical approximations were used and are still pervasive. If you took AP Statistics or a different intro statistics course, you employed mathematical models. However, to be clear, all of the methods we'll talk about (randomization, bootstrap, mathematical models) are techniques to get a **sampling distribution**.

The sampling distributions we've seen so far have been (mostly):

This isn't coincidence…it's guaranteed by a very important theorem, the **Central Limit Theorem**.

Central Limit Theorem

What are the requirements here?

- Independence:

- "Large enough":

Normal Distribution: Nothing followings it exactly–it's a mathematical construct. But, a lot of things follow it approximately, either:

- naturally:

- created to follow it:

The normal distribution depends on two parameters, μ = mean (where the distribution is centered) and σ = standard deviation (how spread out it is). μ shifts the distribution up and down the number line, σ stretches and contracts the curve. The **standard normal** distribution has $\mu = 0$ and $\sigma = 1$ (this is the distribution tabulated in normal tables in textbooks).

The standard normal gives us a convenient way to compare observations, and any normal distribution can be transformed into a standard normal. The **Z-score** is

If the Z-score is positive

If the Z-score is negative

Z-scores can be used to

- gauge the unusualness of an observation

- find probabilities

Helpful R functions:

- `pnorm(x, mean=0, sd=1)`
- `normTail(m=0,s=1,L=x)` or `normTail(m=0,s=1,U=x)` will draw pretty pictures–need to use the OpenIntro library
- `qnorm(prob, mean=0, sd=1)` gives a Z-score with area to the left

Pictures are super-helpful!

Example: Full-term birth weights for single babies are normally distributed with a mean of 7.5 pounds and a standard deviation of 1.1 pounds.

1. A baby is born weighing 9.1 pounds. What is the weight percentile for this baby?

2. Babies that weigh less than 5.5 pounds are considered low birth weight. What proportion of babies are low birth weight?

3. What weight would make a baby at the 25th percentile?

4. What is the probability a randomly selected baby weighs between 7 and 8 pounds?

The **Empirical Rule** (aka the 68-95-99.7 Rule) presents a general rule for the probability of falling within one, two, and three standard deviations of the mean in a normal distribution.

This rule is useful in a wide range of settings when trying to make quick estimate (we'll use it with bootstraps too!).

Some more definitions we'll use throughout the semester:

- **Standard error**:

- **Margin of error**:

Example (13.11): In 2013, the Pew Research Foundation reported that "45% of US adults report that they live with one ore more chronic conditions." However, this value was based on a sample, so it may not be a perfect estimate for the population parameter of interest on its own. The study reported a standard error of about 1.2%, and a normal model may reasonably be used. Create a 95% confidence interval for the proportion of US adults who live with one or more chronic conditions. Interpret the confidence interval in the context of the study.

2.4 Decision Errors (Chapter 14)

Anytime we're using sample data to make decisions about a larger population we can potentially make a mistake. We can make an incorrect decision in a hypothesis test or calculate a confidence interval that does not capture the true population parameter. In a hypothesis test, there are four possible outcomes:

Type I error:

Type II error:

Examples:

- Doping in the Olympics

- Criminal trial

- Diagnostic test for a serious disease

Errors require a balancing act. We want to reduce the chance of making a Type I error but this will necessarily increase the chance of making a Type II error. The best we can do is to set the probability of a Type I error. We can do through setting the **significance level**.

Significance level:

Another consideration that will impact the chance of making an error is the whether the test is one- or two-sided.

Two-sided hypotheses:

Example: Standard anticoagulant therapy to prevent blood clots requires frequent (expensive) lab monitoring. A new procedure called riva was tested because it did not require frequent monitoring. A randomized trial was conducted in 2012, with standard therapy randomly assigned to 2416 patients and riva randomly assigned to 2416 patients. A bad result was a recurrence of a blood clot in a vein. We want to know if the likelihood of a bad result is different between the two therapies.

Here are the results of the randomized trial

	Riva	Standard	Total
Clot	44	60	104
No Clot	2372	2356	4728
Total	2416	2416	4832

For two-sided tests, the p-value is the probability that we observe a result as least as favorable to the alternative hypothesis as the result we observe. That is, that we observe a result as extreme or more extreme in either direction.

When in doubt, use a two-sided test! Use a one-sided test only if you truly have interest in only one direction.

So, how can we control Type I error?

- Set up tests before seeing the data.
- Collect enough data that the test has sufficient **power**. We'll talk more about power later (and LOTS more in an experimental design course), but power is the probability of correctly rejecting a false null hypothesis. It's a function of how big the true difference is (which we don't know and can't control) and the sample size (which we can control).