Stat 801A Notes

Table of contents

Course Goals for STAT 801A

STAT 801A is an introduction to research methods, and how statistical methods may be used to answer research questions. By the end of the course, you will:

- understand the role statistics plays in the research process, and how a statistical investigation works.
- understand statistical evidence, and what conclusions may be drawn based on the evidence and study design.
- be able to make simple probability calculations, and be able to differentiate a few different probability distributions based on the scenario.
- understand that variablility is natural, and commonly used statistics such as the mean, variance, and others have their own probability distributions. Such a probability distribution is called a sampling distribution.
- understand the underlying logic behind commonly used statistical inference techniques (hypothesis tests and confidence intervals).
- realize that the most appropriate statistical inference method changes based on the explanatory variable(s), response variable, and goals of the study.
- be able to calculate and interpret statistical analyses for studies in which there is one (or fewer) explanatory variables.
- be able to sketch a skeleton ANOVA table from a description of the study.
- use statistical software appropriately.
- be able to clearly write up the results of an analysis.

1 Introduction to Data and the Scientific Method

Sound scientific conclusions require evidence from data. Statistics is the science of collecting, analyzing, and drawing conclusions from data. The goal of STAT 801A is to introduce you to the statistical methods used to answer research questions.

The **scientific method** has been used for hundreds of years for discovering new knowledge, and can be summarized with the following diagram:

It's not coincidental that the steps in the scientific method are closely related to the steps in a statistical investigation. These steps appear in Tintle et al. (2021), but are not at all unique to this textbook.

- Step 1: Ask a research question
- Step 2: Design a study and collect data
- Step 3: Explore the data
- Step 4: Draw inferences beyond the data
- Step 5: Formulate conclusions
- Step 6: Look back and look ahead

How do you think the steps in a statistical investigation map to the scientific method? Can you map the baby study to either paradigm?

Each of these steps has a lot of moving parts, so we'll look at each step in more detail and introduce some concepts and introductory definitions as we do so.

1.1 Step 1: Ask a research question

Step 1 boils down to

This may involve

-
-
-

Let's consider the baby example.

Why is a well-stated research question so important?

Asking a research question is often the hardest part of the process, and requires technical information and experience in the discipline. A big reason why you are in graduate school is to gain this information and experience! A statistician can help you narrow your research question and state it precisely, but will not be able to formulate it for you.

1.2 Step 2: Design a study and collect data

Step 2 involves

There is so much going on here that is not evident from the simple statement of "collect data." Let's first think about why there are so many things to consider.

Let's think about the babies. What questions do you think the researchers had to address in their design and data collection?

We're trying answer a research question, and let's specifically think about evaluating hypotheses (though the same applies to estimating an unknown quantity). We can almost never absolutely accept or reject a research theory for two reasons:

1. Variability of experimental material

2. Sampling

Variability and sampling are probably the two most important ideas in statistics, but they are also some of the hardest to grasp. Let's lay out some basic concepts.

A researcher's major goal is to make general statements about their question as it applies to their **population of interest**.

Populations can be finite or infinite. Even if the population is finite, we typically can't measure all of the units in the population. So, to collect data, we must select a subset of the population, a **sample** and hope that the subset is representative of the population.

We'd really rather not rely on hope, and collect data in a way that ensures the sample represents the population. This is typically accomplished by **random sampling**

There are other considerations as well, typically driven by both the research question and practicality. These include:

Experiment or observational study?

If it's an experiment, what is the experimental unit?

What variable(s) will be measured?

How will the variables be measured? With how much precision?

If two or more variables are measured, can one be considered the response variable and the other(s) be considered explanatory?

Is it possible to employ random sampling, random assignment, or both?

How many observations should we collect?

1.3 Step 3: Explore the data

Exploring the data means

For example, consider the histogram below. It shows the percent of residents aged 65 years and over in the 50 US states and District of Columbia.

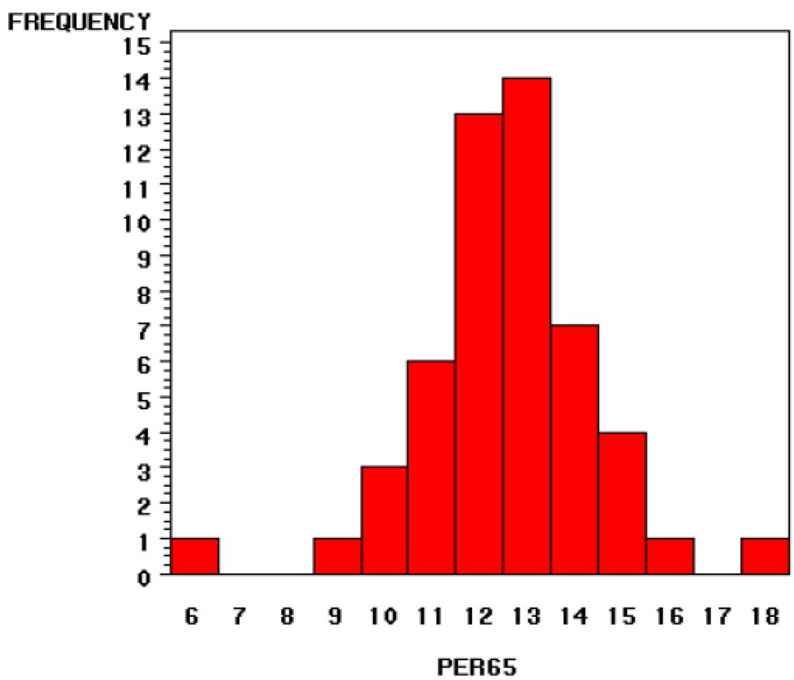

Figure 1.1: Histogram of percent of residents aged 65 and over.

Do you think these outliers are the result of a recording error?

However, exploring the data goes beyond looking for unexpected outcomes, it also encompasses **exploratory data analysis** (EDA). EDA includes both numerical exploration and graphical exploration. Our textbook does a great job summarizing both numerical and graphical summaries of data (pages 30-73), including walking through how EDA can be used in several case studies.

We won't spend a lot of time here, since these are mostly very familiar concepts (mean, median, etc.) However, we'll go through a small example as a preview of coming attractions.

Example: The Gettysburg Address is comprised of 268 words, with word lengths varying from 1 ("a") to 11 ("consecrated") letters. Supposed we're interested in the average word length.

The population of interest is

We're going to take a random sample of $n = 9$ words. The sample is

Table 1.1: Random sample of 9 words from the Gettysburg Address

Word ID	Word	Length
53	long	4
31	Now	3
120	brave	5
263	shall	5
264	not	3
249	of	2
221	full	4
144	note	4
209	take	4

Using our sample, we can easily find the sample mean and sample median.

These values are **statistics**.

We typically use statistics to estimate **parameters**.

In this case, we can actually calculate the parameters, because we have access to the entire population.

This is a very artificial situation. Most of the time, we only have the data in the sample and we want to use the statistics to make some statements about the parameters.

We may also be interested in how much variability there is among word lengths. There are a few ways we could quantify variability. Again, let's consider the sample of 9 words.

Again, these are statistics because they're calculated only from our sample of $n = 9$ observed words. In this case, we can get the parameters.

If we were to take a different sample of size $n = 9$, we'd likely get different statistics. Let's try it.

So the sample mean, a **statistic**, is itself a **random variable**. There is uncertainty associated the outcome.

What happens if we draw samples that are bigger than $n = 9$?

So, the sample mean has its own variance, which depends on the sample size. Specifically,

But, in real life, we only observe one sample–which means we get one mean and one variance. We need to understand the underlying behavior/variability of the sample statistics to be able to use them to make statements about the population parameters. This is why variability and sampling are such important concepts in statistics. We need to know how our sample statistic behaves in order to ...

1.4 Step 4: Draw inferences beyond the data

The general idea in drawing inferences beyond the data

Basically, we're trying to see what the sample data tells us about the population of interest.

Let's go back to the babies. If the babies really can't tell right from wrong, how likely is a baby to pick the good character?

We haven't even seen the data yet, but we can think about how a sample statistic should behave. What was measured? What is the sample statistic of interest? Once we get a handle on how the sample statistic should behave, we can assess how unusual the observed data actually are, if the babies really can't tell right from wrong.

1.5 Step 5: Formulate conclusions

Here, our conclusions must consider the scope of inference made in Step 4.

It's important to keep in mind the population of interest, and whether we employed random assignment, random sampling, both, or neither.

1.6 Step 6: Look back and ahead

This step involves

As we progress through the semester, Step 4 is where we'll spend most of our time. We'll consider different types of variables, different research goals, different study designs, and how we can use the data to draw inferences to a larger population.

As we saw earlier, in order to draw those inferences we need to understand and be able to quantify how much variability we expect to see in the sample statistic. We also need more precise definitions and rules around the uncertainty associated with data. In the next section, we'll discuss the basics of probability and probability distributions.

2 Probability Basics and Probability Distributions

Probability is the language we use to talk about chance and quantify uncertainty. A probability is a number between 0 and 1, where an event is more likely the closer the probability is to 1.

We've already seen a probability! Back to the babies–when we considered how unusual it was to see 13/16 babies pick the good puppet, we calculated:

The value we calculated is a **p-value**: the (empirical) probability of observing what we did in the data (or something even more extreme), under the assumption that the null hypothesis is true. For better or worse, science runs on p-values.

In this section, we'll see some basic probability theory and calculations, as well as probability distributions.

2.1 Probability Basics

When we are uncertain about an outcome's occurrence (e.g., whether a coin will come up heads or tails, the number of dots observed on the roll of a die, whether or not the bus will be late), we typically quantify this uncertainty with a probability. Probability is the foundation upon which all of statistics is built, and it a provides a framework for modeling populations, experiments, and almost anything that could be considered a random phenomenon.

A **sample space**, denoted by S, is comprised of all possible outcomes of a random phenomenon.

An **event** is a collection of possible outcomes. Each event A is a subset of S.

We want to formalize the idea of the "chance" that event A occurs. We will do this by defining the **probability** of each A, which we denote $P(A)$.

Probabilities are calculated by defining functions on sets, and should be defined for all possible events. One thing that must be true:

$$0 \leq P(A) \leq 1$$

More formally, a probability function is defined as follows.

Given a sample space S, a **probability function** is a function $\mathrm{P}(\cdot)$ that satisfies

-
-
-

Any function $\mathrm{P}(\cdot)$ that satisfies these three requirements is called a probability function.

If we let S be a sample space with associated probability function P, we can state some basic facts. Let A, B be events in S.

1.
2.
3.
4.
5.
6.

We'll use these facts when calculating probabilities. First, however, we need to figure out how to assign probabilities to specific events. There are several ways we can do this.

1. **Equally likely outcomes**

2. **Relative frequencies**

3. **Making assumptions**

However we arrive at probabilities for a given scenario, we can use them to construct a **probability distribution.** There are several flavors of probability distribution. The simplest is a list of all possible outcomes and their associated probabilities, and it must satisfy three rules:

1.
2.
3.

Any probability distribution that can be written this way corresponds to a discrete variable or one that we have discretized.

We'll see some other (more common, but more complicated) flavors of probability distributions in a bit, after some facts and definitions.

Consider the following table:

	Survived	Did Not Survive
First Class	201	123
Second Class	118	166
Third Class	181	528

The counts in the table are the number of Titanic passengers that fell into each of the categories. From this table, we can calculate some probabilities.

Sometimes we have partial information about a certain event and wish to know how this affects the probablities of other events, if at all. For example, we might be interested in the probability a passenger survived, given they were in First Class. This is called **conditional probability.**

Definition:

Example: Toss a fair die. Let $A = \{1\}$ and let $B = \{1, 3, 5\}$. What is the probability of throwing a 1, given an odd number was thrown?

This definition of conditional probability leads to:

Let $A_1, A_2, ...$ be a collection of mutually exclusive and exhaustive events. What does this mean?

Suppose we want the probability of an event B.

This leads to the general form of Bayes' Theorem:

Example: (Problem 2.18) A genetic test is used to determine if people have a predisposition for thrombosis, which is a formation of a blood clot inside a blood vessel that obstructs the flow of blood through the circulatory system. It is believed that 3% of people actually have this predisposition. The genetic test is 99% accurate if a person actually has the predisposition. The test is 98% accurate if a person does not have the predisposition.

What is the probability a randomly selected person who tests positive for the predisposition by the test actually has the predisposition?

Consider the following table, which summarizes all flights arriving at an airport in a single day:

	Late	On Time
Domestic	12	109
International	6	53

What is the probability a randomly selected flight on this day was on time?

What is the probability a randomly selected flight was on time, given it was a domestic flight?

What do you notice?

Does this make sense in the context of this scenario? What do you think it means?

Sometimes the occurrence of one event, B, will have no effect on the probability of another event, A. If A and B are unrelated, then intuitively it should be the case

Also, it follows that

Definition:

How is independence used? Let's do a pretty famous example. We'll use a few of the rules we've seen so far.

2.2 Random Variables and Probability Distributions

Typically we are interested in a numerical measurement of the outcome of a random experiment. For example, we might want to know the number of insects treated with a dose of a new insecticide that are killed. In this case, the outcome is the survival status of each dosed insect and the numerical measurement we're interested in is the number that died. However, the observed number varies depending on the actual result of the experiment. This type of variable is called a **random variable.**

Definition: A **random variable** is a function that associates a real number with each element in the sample space. That is, a random variable is a function from a sample space, S, into the real numbers.

Example: Suppose we roll two dice and we're interested in the number of 1s that are thrown.

Random variables can also be defined on a continuous range.

Example: Take a 1 gram soil sample and measure the amount of phosphorus in the sample (in g).

We've already seen one flavor of **probability distribution**: a list of possible outcomes for the random variable, and the associated probabilities.

We can define probability distribution more generally.

Definition: A probability distribution is a function that is used to assign probability to each value the random variable can take on.

Maybe that function can be written in tabular form, as above, maybe it's a function in the mathematical function sense (we'll see some of these later in this section). We can have probability distributions for discrete random variables and continuous random variables.

Discrete probability distributions

- Probabilities are denoted $\mathrm{P}(X = x)$ for the realized value x of random variable X
- $\sum_i \mathrm{P}(X = x_i) = 1$.

Example: We have two seeds in a Petri dish, and will observe how many germinate. We assume the seeds germinate independently, and the probability a randomly selected seed germinates is 0.80.

Continuous probability distributions

- This distribution is called a probability density function (pdf) and denoted $f(x)$.
- The area bounded by $f(x)$, the horizontal axis, and the values a and b is $\mathrm{P}(a \leq X \leq b)$.
- The total area under the pdf is 1.

Example: Let X = phosphorus in a 1 gram soil sample. Suppose we assume the pdf is

$$f(x) = \begin{cases} 1 & 0 \leq x \leq 1 \\ 0 & x < 0, x > 1 \end{cases}$$

Joint probability distributions: We've already seen some of these! A joint probability distribution can be used to study the relationship between two variables, X and Y, simultaneously. We're going to restrict our attention to discrete joint probability distributions, and summarize them as two-way tables.

Let's go back to the Titanic example:

	Survived	Did Not Survive
First Class	201	123
Second Class	118	166
Third Class	181	528

If we know the probability distribution for a random variable, we can use it to calculate things like the "true" mean and variance for that variable.

Expected value: The expected value (or mean) of a discrete random variable is defined as

There are some rules that come along with expected values (discrete or continuous):

1. If X is a random variable and c is a constant, then

2. If X is a random variable, b and c are constants, and $Y = bX + c$, then

3. If X and Y are random variables, b and c are constants, and $W = bX + cY$, then

Example: Let X = number of 1s thrown when rolling two dice.

Variance: The variance of a discrete random variable is defined as

There are also rules that come along with variance (discrete or continuous):

1. For any random variable X and any constant c,

2. If X is a random variable, b and c are constants, and $Y = bX + c$, then

3. If X and Y are **independent** random variables, and b and c are constants, then

4. If X and Y are any two random variables, and b and c are constants, then

Example: In Mendel's experiments on pea plants, he found the trait of being tall is dominant over being short. His theory indicates that if pure-line tall and pure-line short plants are cross-pollinated and then the hybrids in the next generation are cross-pollinated, in the resulting population approximately 3/4 of the plants will appear tall and 1/4 will appear short. If four plants are chosen at random from such a population, the best model (i.e., probability distribution) for the number of tall plants out of the four is

y	0	1	2	3	4
$P(Y = y)$	1/256	12/256	54/256	108/256	81/256

- Find the expected number of tall plants

- Find the variance of number of tall plants

- Find the standard deviation of number of tall plants

- What is the probability that the value of Y will be more than 2 standard deviations below the expected value?

Example: Three patients receive injections to desensitize them from an allergen. The serum used is said to be 90% effective. Let X denote the number of patients who become desensitized.

- Find the probability distribution of X.

- Find the expected number of patients that will become desensitized.

- Find the variance and standard deviation of the number of patients who become desensitized.

- If a patient does not become desensitized, the insurance company will spend \$50 on additional treatment. How much should the insurance company expect to pay in additional costs for these three patients?

Example: A forester is studying a population of trees that are known to have a mean height of 23.4 ft with a variance of 256 ft^2. A tree is randomly selected from the population and its height is measured in feet. Let X represent the height of the randomly selected tree.

- What is the selected tree's expected height in meters? (there are 0.3048 meters in a foot)

- What is the variance of the height of the selected tree in meters?

Example: Contracts for two construction jobs are randomly assigned to one or more of three firms: A, B, and C. Let Y_1 denote the number of contracts assigned to firm A and Y_2 the number of contracts assigned to firm B. The joint probability distribution for this scenario is

- Find the expected number of contracts awarded to Firm A.

- Find the expected number of contracts awarded to Firm B.

- Find the variance of number of contracts awarded to Firm A.

- Find the variance of number of number of contracts awarded to Firm B.

- Find the expected number of contracts awarded to either Firm A or Firm B.

- Find the variance of the number of contracts award to either Firm A or Firm B.

What now? What is this Cov?

Covariance is a measure of the linear relationship between two random variables. It can be positive or negative. A positive covariance indicates that as the value of one RV increases, so does the other. A negative covariance indicates that as the value of RV increases, the other decreases.

For discrete RVs, the covariance is calculated as

If two random variables are independent, the covariance is 0.

For our example, do you think covariance will be positive, negative, or 0?

Let's calculate it, and find the variance above.

Note the units of measurement on covariance.

This makes covariance less intuitive as a measure of dependence–its value depends on the scale of measurement. A measure of dependence that is not dependent on scale is the **correlation**:

The correlation is unitless, and must be $-1 \leq \rho \leq 1$. Just like covariance, if two random variables are independent, their correlation will be 0.

2.3 Special Probability Distributions

Earlier, we mentioned that some probability distributions can be written as mathematical functions. We're going to discuss some probability distributions that commonly arise in data analysis.

2.3.1 The Binomial Distribution

In some studies, the variable of interest only has two potential outcomes: success and failure. These could be died/survived, yes/no, occurred/did not occur, picked the good puppet/picked the bad puppet. Under some very specific conditions, variables like these follow a theoretical probability distribution called the **binomial distribution.**

Here are the conditions we need:

1.

2.

3.

4.

5.

If these conditions are met, the probability distribution of X = number of "successes" observed in n trials is

If X follows a binomial distribution with **parameter** p, then

Right now, we'll use the binomial distribution to calculate some probabilities assuming a specific value for p, but inference for scenarios like this typically focuses on testing hypotheses about p (like the babies!) and estimating p.

Example: A new variety of turfgrass has been developed for use on golf courses, with the goal of obtaining a germination rate of 85%. To evaluate the grass, 20 seeds are planted in a greenhouse so that each seed will be exposed to identical conditions. If the 85% germination rate is correct, what is the probability that 18 or more seeds will germinate?

How many seeds do we expect to germinate? What is the variance of the number of germinated seeds?

2.3.2 The Poisson Distribution

The **Poisson distribution** models count data, typically the number of events observed for a particular unit of time or space. For example, the Poisson can be used to model variables like:

- the number of hits to a website per minute
- the number of PCB particles in a liter of water
- the number of insects in a square meter
- the number of cars passing through an intersection in 5 minutes
- the number of flaws in a yard of fabric

Like the Binomial, the Poisson has some requirements:

1.

2.

3.

The probability distribution for the Poisson is

The Poisson distribution has a couple of interesting features:

Example: Suppose grasshoppers are distributed at random in a large field according to a Poisson distribution with $\lambda = 2$ grasshoppers per square meter.

- Find the probability that no grasshoppers will be found in a randomly selected square meter.

- Find the probability that 2 or fewer grasshoppers will be found in 2 square meters.

- Find the expected number of grasshoppers in 10 square meters.

- Find the expected number of grasshoppers in 0.5 square meters.

2.3.3 The Normal Distribution

The most commonly used continuous distribution (maybe the most commonly used distribution, period) is the **normal distribution**. It's commonly used because

-
-
-

The normal distribution is bell-shaped, symmetric, and unimodal. In fact, we shouldn't call it **the** normal distribution, there are an infinite number of different normal distributions, depending on the **parameters** of the distribution, μ and σ^2.

- μ represents the mean of the distribution

- σ^2 represents the variance of the distribution

The normal distribution does has a mathematical function (a pdf) that governs its shape:

We denote random variables following the normal as

and the normal with mean $\mu = 0$ and variance $\sigma^2 = 1$ is called the **standard normal** distribution.

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

Some helpful R functions:

- `pnorm(x, mean=0, sd=1)`
- `qnorm(prob, meam=0, sd=1)`

- `normTail(m=0,s=1, L=x)` or `normTail(m=0,s=1,U=x)` (does require the OpenIntro library)

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.

- A randomly selected newborn weighs 9.1 pounds. What is the weight percentile for this baby?

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

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

- 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 a quick estimate.

The normal distribution is useful because it can be used to approximate other distributions, such as the binomial.

Let's see what happens with $p = 0.15$ as we change the sample size.

Recall the binomial distribution has

If n is sufficiently large, the binomial can be well-approximated with a normal distibution with $\mu = np$ and $\sigma^2 = np(1 - p)$.

What's sufficiently large?

Example: (problem 3.33) Suppose a university announced that it admitted 2500 students for the incoming first year class. However, the univesity has dorm room spots for only 1786 first year students. If there is a 70% chance an admitted student will enroll at the university, what is the probability the university will not have enough dorm room spots?

3 Sampling Distributions and Foundations of Statistical Inference

As we've seen in the last two chapters, variability is natural and expected. We expect to see variability in observations, which implies there will also be variability in summary statistics. We've seen this already:

If we want to use a summary statistic (like $\bar{X}$ or $\hat{p}$) calculated from our sample to draw inferences about the population, we have to understand how the summary statistic behaves.

This means, we need to know the **sampling distribution** of the statistic.

3.1 Sampling Distributions

As a refresher, the goal of statistical inference is to use an observed data set to answer questions about the overall population from which the sample data set was drawn. Typically, those questions may be answered using some **parameter(s)** of the population distribution.

A **parameter** is

For example,

Parameters are generally fixed, unknown constants. We want to use our sample data to answer a question about the parameter (hypothesis test) or estimate the parameter (confidence interval). We may also be interested in functions of parameters.

Often, the **statistic** we'll use to estimate the underlying parameter is pretty intuitive.

But, if we want to use a statistic, we have to understand its behavior.

The **sampling distribution** is

We've can study sampling distributions empirically, through simulation. We've already done this!

We can also quantify sampling distributions theoretically. We've already done this too!

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:

But wait, the sample mean? Weren't we also considering sample proportions?

Let's think more about these requirements:

- Independence
- "Large enough"

If the Central Limit Theorem holds, the underlying parameters of the resulting approximate normal distribution will depend on the population from which the original data were drawn.

Other statistics will have sampling distributions that do not follow an approximate normal. For example, the sample variance is a natural estimate for the population variance. But, the CLT does not apply to variances. We'll need a different distribution.

Once we can articulate the sampling distribution, we can use it to do statistical inference.

3.2 Foundations of Statistical Inference

In Chapter One, we talked about framing a research question. Many (but not all) research questions can be answered using **statistical inference**. We'll now lay out the basic logic of statistical inference, illustraing the different methods for the case in which we have a single response variable (quantitative or categorical) and no explanatory variable. The framework for statistical inference will not change as we move to more complicated scenarios.

Statistical inference is a collection of techniques which use information from a sample to make precise statements about the entire population. In STAT 801A, the general statements about populations will be expressed in terms of the parameters, or functions of parameters, of probability distributions. Because we know the sampling distribution, we can use probability to precisely quantify the accuracy of our general statements.

Statistical inference is broken into two broad categories: estimation and testing. These map back to the types of research questions we outlined in Chapter One.

3.2.1 Estimation

This category of statistical inference is concerned with using sample information to estimate one or more parameters, or functions of parameters, of the probability distribution for a population. For example, we may be interested in estimating the mean of a population, or the difference in means between two populations. There are two types of estimation, point estimation and interval estimation.

Point estimation

But a single value is not very meaningful without some way of telling how close our estimate comes to the true value.

Interval estimation

We'll illustrate how interval estimation works with an example.

Example: An entomologist is studying a new tick species that may be the carrier of the pathogen associated with lyme disease. They design a study to estimate prevalence of the pathogen in the tick. They examine 200 ticks randomly selected in the study region during a period of the year when ticks have been known to be infected with the pathogen in other regions of the country. They find 18 ticks that are infected with the pathogen.

- Parameter of interest:

- Sample statistic:

Are the requirements for the Central Limit Theorem met?

The Central Limit Theorem tells us

Now let's consider the Empirical Rule.

Most (but not all) confidence intervals have the form:

So, to calculate a confidence interval of this form we'll need the **margin of error**, which is calculated based on the **standard error of the statistic** and the **sampling distribution of the statistic**. We'll also need to specify how much certainly we want in our interval estimate.

Back to the ticks.

What happens if we change our level of confidence?

What if we want a confidence level that isn't 68, 95, or 99.7?

Let's think more carefully about what this confidence level means. A confidence interval is a probability statement, but not the probability statement that is intuitive. Suppose we are interested in an interval estimate for a parameter θ.

It's super important to understand that this probability statement is only valid for as long as L and U are unknown. Once we use the data to estimate L and U, and get $\hat{L}$ and $\hat{U}$, the interval is no longer random. The interval either contains the parameter or it doesn't. This means statements like

are **incorrect**, as tempting as they are to write. Rather, the statement of probability is about the method used to obtain the confidence interval.

Let's look at the applet to explore what that confidence level really means: Applet

So, let's find a 98% confidence interval for the proportion of ticks that are infected by the pathogen.

Example: (4.17, sort of) The nutrition label on a bag of potato chips says that a one ounce serving has 130 calories and 10 grams of fat. A random sample of 35 bags yielded a sample mean of $\bar{x} = 134$ calories with a sample standard deviation of $s = 17$ calories. Assume the distribution of bags is relatively symmetric. We want a 95% confidence interval for the true mean calorie count of a bag of potato chips.

What's different about this example, compared to the tick example?

Let's state the Central Limit Theorem again.

This presents a few complications:

-

-

The natural fix is to use s (the sample standard deviation) in place of σ, so the standard error is

But this leads to yet another complication: the normal distribution isn't quite right. Instead, we end up with a distribution that has heavier tails than the normal. Instead, we use the t distribution. The t distribution has a single parameter, the degrees of freedom (df). The degrees of freedom determines the shape of the t, with the distribution getting closer and closer to the normal as the df increase.

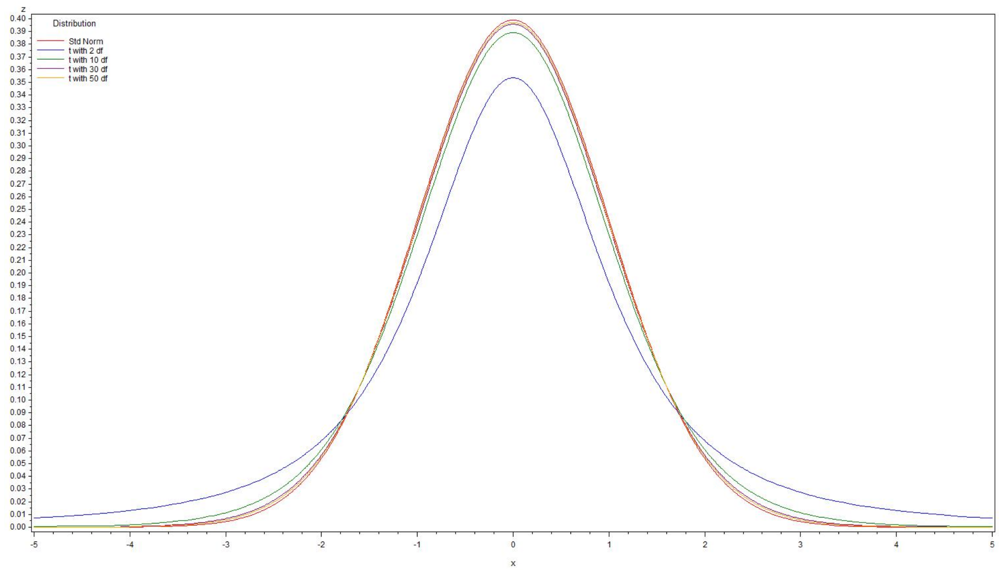

Figure 3.1: Standard normal compared to the t distribution with various df

In the scenario of a single mean, $df = n - 1$ but this will change as the scenario gets more complicated.

We can get t probabilities and quantiles using the R functions

- `pt(x, df=)`
- `qt(prob, df=)`

So, if we're interested in calculating an interval estimate for a mean

Back to the potato chips example. Are the conditions for the Central Limit Theorem met?

We'll calculate a 95% confidence interval.

Example: An ichthyologist is interested in estimating the variance of lengths of trout minnows in a very large tank at a fish hatchery. It is reasonable to assume that lengths are normally distributed. 15 minnows are randomly sampled from the tank and measured. The sample variance is $s^2 = 0.17$ inch^2.

What's different now?

What complications does this present?

We need a new distribution! We need the sampling distribution of S^2. It turns out that a function of S^2 follows the χ^2 distribution. The χ^2 has the following properties:

-

-

If our original observations come from a normal distribution, then

This gives us a straightforward way to find a confidence interval for σ^2.

We can find these $\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$ using the `qchisq(prob,df=)` function in R. For example,

This is one of the cases where the confidence interval does not have the estimate $\pm$ margin of error form. That's because the χ^2 isn't symmetric. But, we now have all the information we need to calculate the confidence interval for the variance in trout length.

A word of warning. This is not a robust procedure. If the assumption of normality is not met, this interval will give poor results. This is not true of the t interval for the mean.

3.2.2 Hypothesis Testing

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, let's consider the ticks again. The entomologist knows from a literature review that the prevalence of the lyme disease pathogen in the black-legged tick is 0.02. They are interested in whether the presence of the pathogen is more prevalent in the new tick species. The yes/no question we will answer is whether the resulting data provide convincing evidence that the pathogen is more prevalent in the new species. These questions lead to two competing claims, both stated in terms of parameters of a probability distribution

- **Null hypothesis**

- **Alternative hypothesis**

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. We will **NEVER** conclude that the null hypothesis is true.

Hypothesis tests work by assuming the null hypothesis is true, and assessing the plausibility of the observed data under that assumption.

The entomologist examined 200 ticks randomly selected from the study region. If we assume the null hypothesis is true, then we expect to see

In fact, 18/200 ticks were infected with the pathogen. The question then becomes

To see how unusual this sample result of 18/200 is, we again need the sampling distribution of the sample statistic. As a reminder, the Central Limit Theorem says

So we can use normal distribution to see how unusual 18/200 is, if the null hypothesis is true.

Example: Let's consider the potato chip example again. The bag claims that a serving contains 130 calories. We want to test whether this is true. This leads to the hypotheses

What's different here?

We can again appeal to the Central Limit Theorem and the t distribution to characterize the sampling distribution of $\bar{X}$, which leads to the **test statistic**

The random sample of 35 bags had a sample mean of $\bar{x} = 134$ and standard deviation $s = 17$.

But now what is "more unusual" assuming the null hypothesis is true?

When in doubt, use a two-sided test! Use a one-sided test only if you truly have interest in only one direction. Why? To fully answer this, we need to address **decision errors.**

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 at the outset of the study:

- **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 this through setting the **significance level**.

Significance level:

So how does this fit in with one- and two-sided hypotheses?

How else can we control Type I error?

- Set up tests before seeing the data.
- Collect enough data that the test has sufficient **power**. Power is the probability of correcting 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), the expected variability in our responses (also can't control, but might know), and the sample size (which we can control). We'll talk more about power later on in the semester.

The two examples we've seen have both utilized a test statistic with the form

With confidence intervals, we mentioned that many confidence intervals have the form estimate $\pm$ margin of error, but not all do. We saw an example, a confidence interval for a variance, that had a different form. Similarly, many tests have a test statistic of the form

$$\frac{\text{estimate - hypothesized value}}{\text{standard error of estimate}}$$

but not all do.

Example: The Poisson distribution is often a good model for scenarios in which we are counting occurrences over some specified time or space unit. However, the Poisson distribution has the characteristic that the population mean = population variance. In some scenarios, this may not be true, invalidating the Poisson as a possible model. We can use hypothesis testing to determine if the Poisson is a reaonable model for a data set. A scientist is interested in modeling the number of parasites found on a host, and believes the Poisson may be a feasible model.

The researcher examines 80 host organisms, and records the number of parasites found on each. The data are:

Number of Parasites	0	1	2	3	4	5
Number of hosts	20	28	19	9	3	1

There is not a single mean or proportion (or variance) we can calculate here that will summarize how closely these data follow a Poisson distribution. Instead, we'll need to come up with a new test statistic.

The first thing we'll need is an estimate of the Poisson parameter, λ.

Now, if we consider the Poisson distribution with $\lambda = 1.375$ we can calculate some probabilities:

X	Probability
0	0.2528
1	0.3477
2	0.2390
3	0.1095
4	0.0377
5	0.0104
over 5	0.0029

If the Poisson distribution is a realistic model, we would expect to see our data fall into these categories in about these proportions. So, we expect

Number of Parasites	0	1	2	3	4	5	>5
Number of hosts	20	28	19	9	3	1	0
Expected	20.224	27.816	19.12	8.76	3.016	0.832	0.232

and we can compare the observed counts to the expected counts.

Number of Parasites	0	1	2	3	4	5	>5
Number of hosts	20	28	19	9	3	1	0
Expected	20.224	27.816	19.12	8.76	3.016	0.832	0.232
Difference	-0.224	0.184	-0.12	0.24	-0.016	0.168	-0.232

But we've got another problem.

Again, our solution will be squaring! This time we'll also scale. The resulting test statistic is:

This is called the **chi-squared goodness-of-fit** test. Under the null hypothesis, this test statistic will follow a χ^2 distribution with $k-1$ degrees of freedom, where k is the number of categories. However, we also need a big enough sample so that all expected counts are at least 5. That's not true here. What now?

Number of Parasites	0	1	2	≥ 3
Number of hosts	20	28	19	13
Expected	20.224	27.816	19.12	12.84
Difference	-0.224	0.184	-0.12	0.16

So now,

We can also easily do this in R:

```
host<-c(20,28,19,9, 3, 1, 0)
chisq.test(host,p=c(0.2528, 0.3477, 0.2390, 0.1095, 0.0377, 0.0104, 0.0029))
```

```
Warning in chisq.test(host, p = c(0.2528, 0.3477, 0.239, 0.1095, 0.0377, :
Chi-squared approximation may be incorrect


    Chi-squared test for given probabilities

data:  host
X-squared = 0.27703, df = 6, p-value = 0.9996
```

So R is telling us our sample size isn't big enough for the χ^2 distribution to work. Like we did by hand, we can collapse some categories.

```
host<-c(20,28,19,13)
chisq.test(host,p=c(0.2528, 0.3477, 0.2390, 0.1605))
```

```

    Chi-squared test for given probabilities

data:  host
X-squared = 0.0064451, df = 3, p-value = 0.9999
```

So it appears we have no reason to doubt that the Poisson distribution is a good model for these data.

Now that we've seen the logic behind statistical inference, we can move on to more complicated situations. We'll consider cases in which we have a single explanatory variable and a single response variable. We'll first cover the case where the explanatory variable is categorical with only two levels, and the response variable is either categorical or numeric (comparing two groups). We'll then move on to the case where the the explanatory variable is categorical with more than two levels, and the response variable is categorical or numeric. Finally, we'll consider the case where the explanatory and response variable are both numeric.