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.

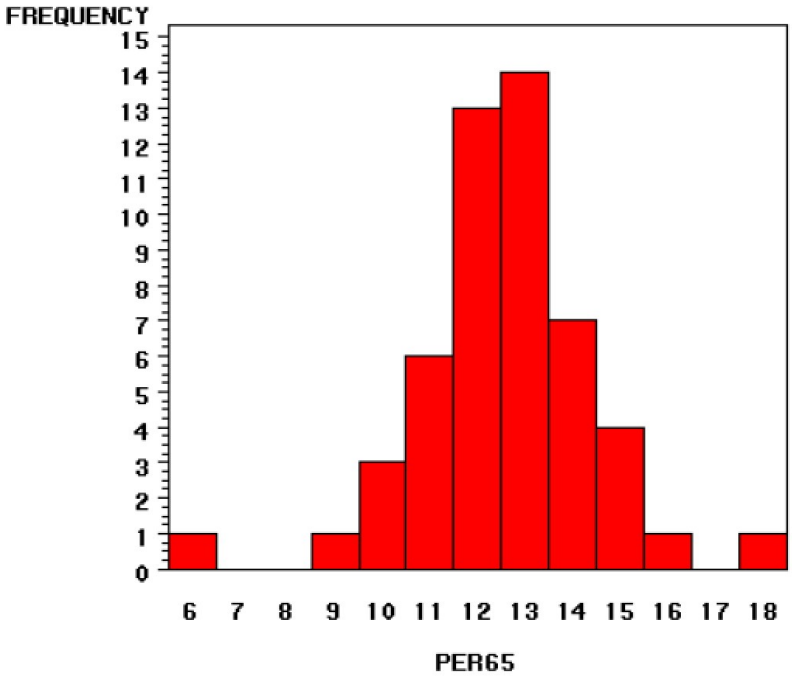


Figure 1.1: Histogram of percent of residents aged 65 and over.

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
$\mathrm{P}(Y = y)$	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.

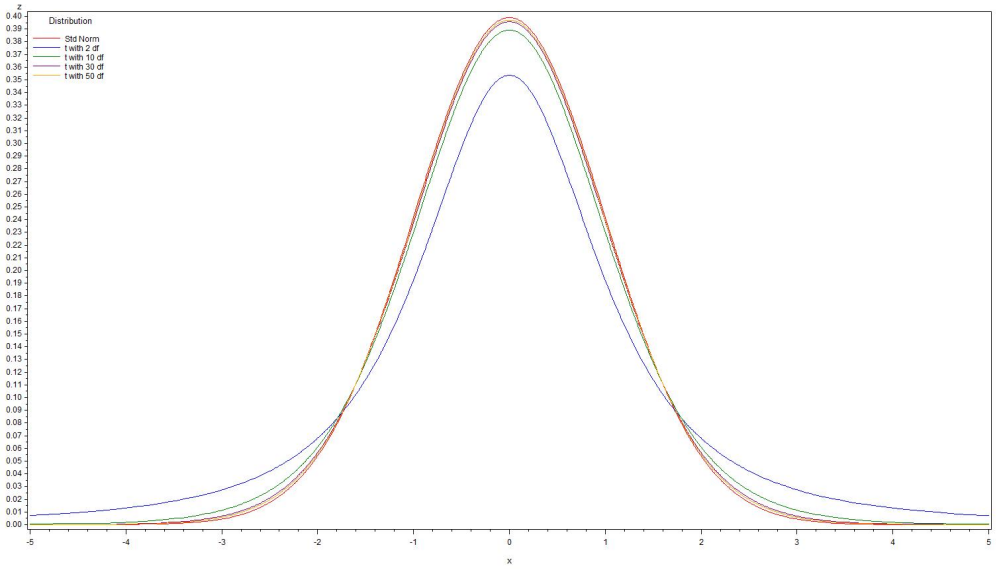


Figure 3.1: Standard normal compared to the t distribution with various df

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.

4 One Predictor/Explanatory Variable, Two Levels

As mentioned at the end of Chapter 3, we'll now move on to cases in which we have a single explanatory variable and a single response variable. In this section, we'll cover the case where the explanatory variable is categorical with only two levels, and the response variable is either categorical or numeric. This means that in this chapter, we'll be focusing on comparing two groups.

Data like these may show up in a spreadsheet like

4.1 Categorical Response, Two Levels

First, we'll consider situations in which two categorical variables are measured on each unit in the sample, and each variable has two possible values. In cases like these, typically one variable is considered the response and one variable is considered explanatory. The explanatory variable may be randomly assigned (like whether a subject was assigned to a treatment or control) or it may be merely observed (like smoking status).

The two possible values of the explanatory variable lead to two groups, and we're interested in comparing the population proportions that arise from these two groups. We'll focus on the function of parameters $p_1 - p_2$. The natural estimate of this is $\hat{p_1} - \hat{p_2}$: the difference in the sample proportions. We'll be constructing hypothesis tests to compare p_1 to p_2 and finding confidence intervals to estimate $p_1 - p_2$. To demonstrate these methods, we'll use an example.

Example: Biologists studying crows will capture a crow, tag it, and release it. These crows seem to remember the scientists who caught them and will scold them later. A study to examine this effect had several scientists wear a caveman mask while they trapped and tagged 7 crows. A control group did not tag any crows and wore a different mask. The two masks did not elicit different reactions from the crows before tagging. Volunteers then strolled around town wearing one or the other of the two masks.The crows scolded a person wearing a caveman mask in 158 out of 444 encounters with crows, whereas crows scolded a person in a neutral mask in 109 out of 922 encounters. Suppose we want to find a confidence interval for the difference in proportion of crow scoldings between volunteers wearing the caveman mask and those wearing the neutral mask.

For a single proportion, we needed two conditions to be met to ensure the sampling distribution of $\hat{p}$ is approximately normal:

-
-

If these conditions are met, then

We must meet similar conditions to ensure the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is approximately normal:

-
-

If these conditions are met, then

Like before we don't know p_1 and p_2, so we'll use our best guess. And, like before, our best guess will change depending on whether we're constructing a confidence interval or carrying out a hypothesis test.

How is this going to play out in a confidence interval?

Let's go back to the crows.

How is this going to play out in a hypothesis test?

Again, let's go back to the crows.

We can also do this is R or SAS, but either program will use a different (but also not really) approach. We'll start with R.

```
prop.test(x=c(158,109), n=c(444,922))
```

```
    2-sample test for equality of proportions with continuity correction

data:  c(158, 109) out of c(444, 922)
X-squared = 106.11, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 0.1867976 0.2884716
sample estimates:
   prop 1    prop 2
0.3558559 0.1182213
```

From this output, what looks familiar?

What doesn't look familiar?

But is this what we actually tested?

```
prop.test(x=c(158,109), n=c(444,922),alternative="greater",correct=FALSE)
```

```
    2-sample test for equality of proportions without continuity correction

data:  c(158, 109) out of c(444, 922)
X-squared = 107.62, df = 1, p-value < 2.2e-16
alternative hypothesis: greater
95 percent confidence interval:
 0.196371 1.000000
sample estimates:
   prop 1    prop 2
0.3558559 0.1182213
```

What do you notice?

In SAS, we can use `proc freq`. First, we'll need to read in the data.

```
data crows;
  input mask $ NumScold Total;
  response="Scold"; Count=NumScold;  output;
  response="NoScold"; Count=Total-NumScold; output;
  datalines;
Caveman 158 444
Neutral 109 922
;

proc print data=crows; run;
```

Here's the data set

Obs	mask	Num Scold	Total	response	Count
1	Caveman	158	444	Scold	158
2	Caveman	158	444	NoSco	286
3	Neutral	109	922	Scold	109
4	Neutral	109	922	NoSco	813

Now, to get the test

```
proc freq data=crows;
  weight Count;
  table mask*response/chisq;
run;
```

which gives

The FREQ Procedure

Frequency Percent Row Pct Col Pct	Table of mask by response		
	response		
mask	NoSco	Scold	Total
Caveman	286 20.94 64.41 26.02	158 11.57 35.59 59.18	444 32.50
Neutral	813 59.52 88.18 73.98	109 7.98 11.82 40.82	922 67.50
Total	1099 80.45	267 19.55	1366 100.00

Statistics for Table of mask by response

Statistic	DF	Value	Prob
Chi-Square	1	107.6155	<.0001
Likelihood Ratio Chi-Square	1	101.6008	<.0001
Continuity Adj. Chi-Square	1	106.1097	<.0001
Mantel-Haenszel Chi-Square	1	107.5368	<.0001
Phi Coefficient		-0.2807	
Contingency Coefficient		0.2702	
Cramer's V		-0.2807	

Fisher's Exact Test	
Cell (1,1) Frequency (F)	286
Left-sided Pr <= F	<.0001
Right-sided Pr >= F	1.0000
Table Probability (P)	<.0001
Two-sided Pr <= P	<.0001

Sample Size = 1366

Figure 4.1: Output from proc freq

Why do both R and SAS carry out the procedure like this? This is a method that can be used in situations where both the explanatory and response variable have any number (≥ 2) possible values. We'll see examples like this in the next chapter.

But! The methods are actually doing the same thing. Let's look at the test we carried out by hand.

Example: Do metal tags on penguins harm them? Scientists trying to tell penguins apart have several ways to tag the birds. One method involves wrapping metal strips with ID numbers around the penguin's flipper, while another involves electronic tags. Neither tag seems to physically harm the penguins. However, since tagged penguins are used to study **all** penguins, scientists wanted to determine whether the tagging method has any effect. Data were collected over a 10-year time span from a sample of 100 penguins that were randomly given either metal or electronic tags. Information collected includes number of chicks, survival over the decade, and length of time on foraging trips. Let's first consider survival. We're interested in estimating the difference in survival rate between penguins with metal tags and penguins with electronic tags.

What parameters are of interest here?

What kind of research question are we trying to answer? What does this imply about the analysis method?

What next?

Let's do the analysis in R.

4.2 Quantitative Response

Example: Data were collected over a 10-year time span from a sample of 100 penguins that were randomly given either metal or electronic tags. Information collected includes number of chicks, survival over the decade, and length of time on foraging trips. Now let's focus on length of foraging trips. Longer foraging trips can jeopardize both breeding success and survival of chicks waiting for food. Suppose we're interested in estimating the difference in mean trip length between penguins with metal tags and those with electronic tags.

What are the parameters?

What kind of research question are we trying to answer? What does this imply about the analysis method?

What's different from the crows example?

This means we will have to change our analysis approach.

Just like with the t methods for single means, we need to check conditions to determine whether we can the t-distribution to construct tests and form confidence intervals for the difference in means.

- Independence–both between and within groups
- Check normality of each group separately (basically checking for extreme outliers)
- If these are both met, then the standard error of $\bar{x}_1 - \bar{x}_2$ is $SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ with $df =$ really complicated (you'll see we get non-integers in R/SAS–it's doing the complicated calculation). We'll use $\min(n_1 - 1, n_2 - 1)$ if we're not using R/SAS. We won't know σ_1^2 and σ_2^2, so we'll approximate the standard error using $SE \approx \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

As with tests for a single mean (and one proportion, and two proportions), our test statistic will have the usual form:

$$\text{test statistic} = \frac{\text{observed value - hypothesized value}}{SE}$$

In the case of two means, this is

$$T = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

When the null hypothesis is true and the conditions are met, T has a t-distribution with $df = \min(n_1 - 1, n_2 - 1)$.

Confidence intervals will also have the same form:

$$\text{observed statistic} \pm \text{multiplier} \times SE$$

For this specific situation of comparing two independent means, this is

$$(\bar{x}_1 - \bar{x}_2) \pm t^*_{df} \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

and we'll again use $df = \min(n_1 - 1, n_2 - 1)$ (or let R/SAS calculate it for us).

With two proportions, our SE changed depending on whether we were doing a hypothesis test or calculating a confidence interval. Here, it doesn't. Any guesses why?

Example: There were 344 foraging trips made by penguins with a metal tag, and those trips had a sample mean of $\bar{x}_M = 12.70$ days with standard deviation $s_M = 3.71$ days. For those penguins with electronic tags, the mean was $\bar{x}_E = 11.60$ days with standard deviation $s_E = 4.53$ days over 512 trips.

Example: Another variable measured was the date penguins arrive at the breeding site, with later arrivals hurting breeding success. Arrival date is measured as the number of days after 1 November. The researchers are interested in whether metal tagged penguins arrive later than electronic tagged penguins.

What are the parameters?

What kind of research question are we trying to answer? What does this imply about the analysis method?

Mean arrival date for the 167 times metal tagged penguins arrived was 7 December (37 days after 1 November) with standard deviation $s_M = 38.77$ days, while mean arrival date for the 189 times electronic tagged penguins arrived was 21 November (21 days after 1 November) with standard deviation $s_E = 27.50$

We can easily carry out t tests and confidence intervals in R and SAS. But, we can't for the penguin data. R and SAS both require the whole data set, as opposed to summary statistics.

Example: The data set may be found in Canvas: 'NutritionStudy.csv'. This data set gives nutrition levels in people's blood as well as information about their eating habits, and comes from a random sample of 315 US adults. Suppose we are interested in estimating the difference in mean beta carotene blood level between smokers and non-smokers. Let's start by reading the data into R.

```
NutritionStudy<-read.csv("NutritionStudy.csv",header=TRUE)

head(NutritionStudy)
```

```
  ID Age Smoke Quetelet Vitamin Calories  Fat Fiber Alcohol Cholesterol
1  1  64    No  21.4838       1   1298.8 57.0   6.3     0.0       170.3
2  2  76    No  23.8763       1   1032.5 50.1  15.8     0.0        75.8
3  3  38    No  20.0108       2   2372.3 83.6  19.1    14.1       257.9
4  4  40    No  25.1406       3   2449.5 97.5  26.5     0.5       332.6
5  5  72    No  20.9850       1   1952.1 82.6  16.2     0.0       170.8
6  6  40    No  27.5214       3   1366.9 56.0   9.6     1.3       154.6
  BetaDiet RetinolDiet BetaPlasma RetinolPlasma    Sex VitaminUse PriorSmoke
1     1945         890        200           915 Female    Regular          2
2     2653         451        124           727 Female    Regular          1
3     6321         660        328           721 Female Occasional          2
4     1061         864        153           615 Female         No          2
5     2863        1209         92           799 Female    Regular          1
6     1729        1439        148           654 Female         No          2
```

If I wanted to calculate the confidence interval by hand, I could use R to get the summary statistics

```
NutMeanNS<-mean(NutritionStudy$BetaPlasma[NutritionStudy$Smoke=="No"])
NutMeanNS
```

```
[1] 200.7316
```

```
NutSDNS<-sd(NutritionStudy$BetaPlasma[NutritionStudy$Smoke=="No"])
NutSDNS
```

```
[1] 192.2929
```

```
size_NS<-sum(with(data=NutritionStudy, Smoke=="No"))
size_NS
```

```
[1] 272
```

```
NutMeanS<-mean(NutritionStudy$BetaPlasma[NutritionStudy$Smoke=="Yes"])
NutMeanS
```

```
[1] 121.3256
```

```
NutSDS<-sd(NutritionStudy$BetaPlasma[NutritionStudy$Smoke=="Yes"])
NutSDS
```

```
[1] 78.81163
```

```
size_S<-sum(with(data=NutritionStudy, Smoke=="Yes"))
size_S
```

```
[1] 43
```

So now we have all the components we need to calculate the confidence interval.

We can also let R calculate the confidence interval for us, using `t.test`:

```
t.test(BetaPlasma~Smoke,data=NutritionStudy)
```

```
    Welch Two Sample t-test

data:  BetaPlasma by Smoke
t = 4.7421, df = 139.15, p-value = 5.175e-06
alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
95 percent confidence interval:
  46.29873 112.51335
sample estimates:
 mean in group No mean in group Yes
         200.7316          121.3256
```

We can change the confidence level easily

```
t.test(BetaPlasma~Smoke,data=NutritionStudy, conf.level=0.90)
```

```
    Welch Two Sample t-test

data:  BetaPlasma by Smoke
t = 4.7421, df = 139.15, p-value = 5.175e-06
alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
90 percent confidence interval:
  51.67854 107.13353
sample estimates:
 mean in group No mean in group Yes
         200.7316          121.3256
```

Let's carry out a test by hand, to see how it compares to the output.

4.3 Comparing Paired Means

Everything we've done so far has assumed independence among observations. If we only had one group, it was just independence among observations. If we had two or more groups, it was independence between and within groups. Now, we'll turn our attention to a common situation: dependence between groups. Specifically, a particular dependency–pairing. This occurs in before/after studies, other studies in which subjects are matched. For example, considering the price of a item purchased from two different retailers.

In these situations, we generally take the difference between the two values, and consider the difference as our observation. So, for example, if we want to compare cost of textbooks between the campus bookstore and Amazon, we'd randomly select a set of book titles, and find their price at both the bookstore and Amazon. We'd find the difference in price, and use those differences as our observations.

Note that we're distinguishing between **difference in means** and **mean difference**.

- Parameters:

- Observed Statistics:

Good news: we've already seen how to construct tests and confidence intervals here! We just use the same techniques we used for a single mean (Chapter 3), but on the differences. The changes come in the form of the hypotheses and interpretation of the confidence interval.

Example: Long distance runners contend that moderate exposure to ozone increases lung capacity. In investigate this possibility, a researcher exposed 12 rats to ozone at the rate of 2 ppm for a period of 30 days. The lung capacity of the rats was determined at the beginning of the study and again after 30 days of ozone exposure. The lung capacities (in mL) are in the file 'ozone.csv'.

```
ozone<-read.csv("ozone.csv",header=TRUE)

head(ozone)
```

```
  Rat Before After
1   1    8.7   9.4
2   2    7.9   9.8
3   3    8.3   9.9
4   4    8.4  10.3
5   5    9.2   8.9
6   6    9.1   8.8
```

The first thing we'll do is calculate the change in lung capacity.

```r
ozone$diff<-ozone$Before - ozone$After

head(ozone)
```

```
  Rat Before After diff
1   1    8.7   9.4 -0.7
2   2    7.9   9.8 -1.9
3   3    8.3   9.9 -1.6
4   4    8.4  10.3 -1.9
5   5    9.2   8.9  0.3
6   6    9.1   8.8  0.3
```

What is the parameter?

What research question are we trying to answer?

What does this imply about the analysis method we should use?

```r
t.test(ozone$diff)
```

```
    One Sample t-test

data:  ozone$diff
t = -3.885, df = 11, p-value = 0.002541
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -1.8928932 -0.5237735
sample estimates:
mean of x
-1.208333
```

```r
t.test(ozone$Before,ozone$After,paired=TRUE)
```

```
    Paired t-test

data:  ozone$Before and ozone$After
t = -3.885, df = 11, p-value = 0.002541
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -1.8928932 -0.5237735
sample estimates:
mean difference
      -1.208333
```

What is incorrect about this analysis in R? How can we fix it?

```r
t.test(ozone$Before,ozone$After,paired=TRUE,alternative="less")
```

```
    Paired t-test

data:  ozone$Before and ozone$After
t = -3.885, df = 11, p-value = 0.001271
alternative hypothesis: true mean difference is less than 0
95 percent confidence interval:
       -Inf -0.6497695
sample estimates:
mean difference
      -1.208333
```

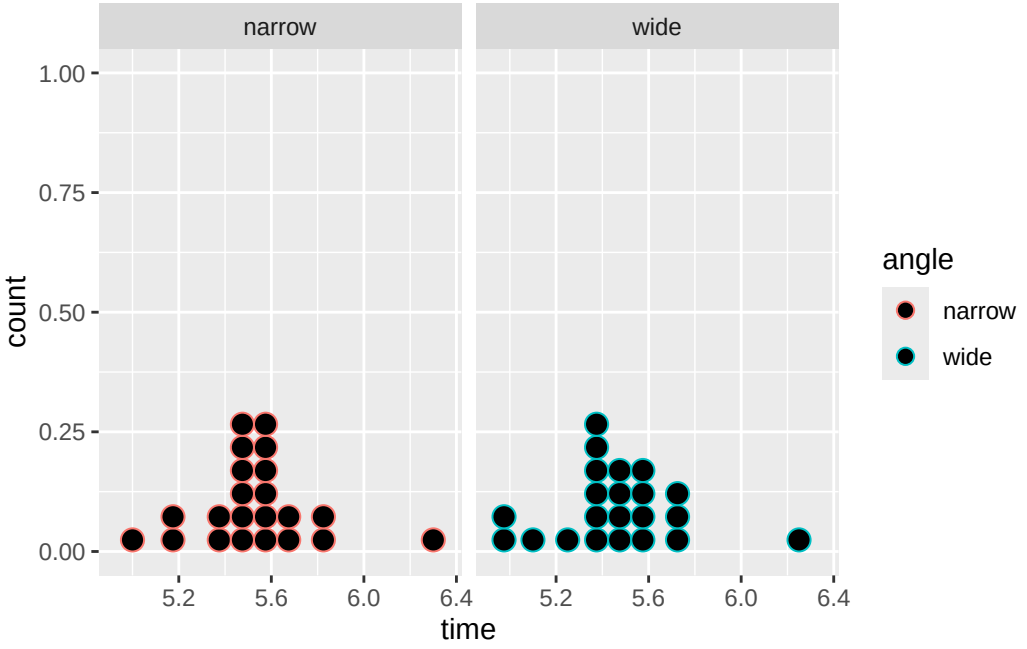
Let's write a couple of conclusions here.

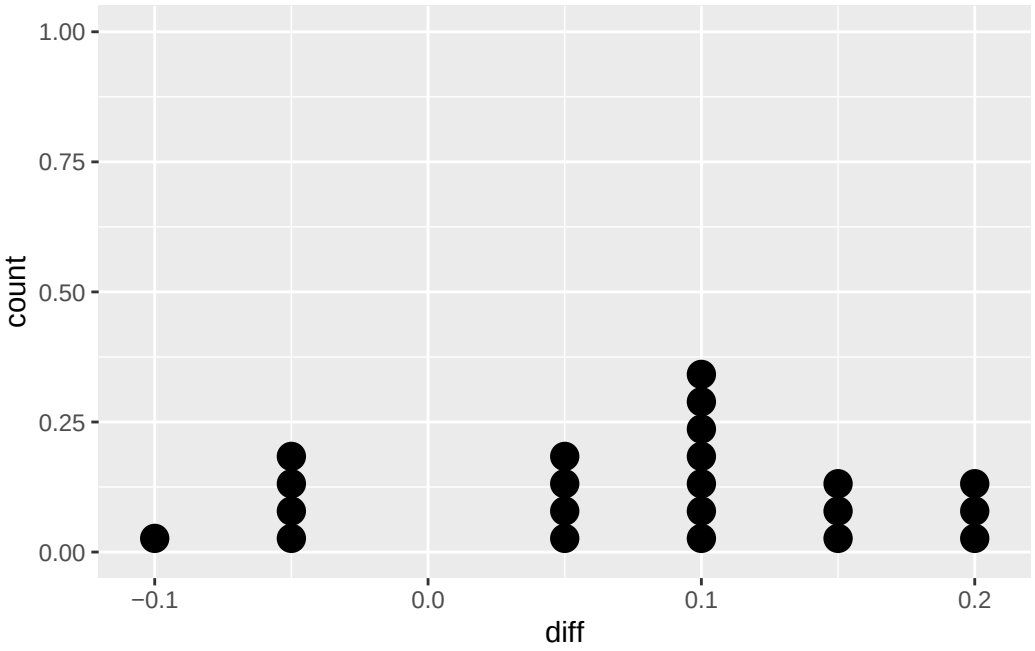
What are the consequences of ignoring pairing? Let's look at a different example.

Example: Suppose you are playing baseball and hit a hard line drive. You want to turn a single into a double. Does the path you take to round first base make a difference? A masters thesis way back in 1970 considered the difference between a "narrow angle" and a "wide angle" around first base. Suppose we have 22 baseball players who have volunteered to participate. There are a couple ways we could design an experiment to see if there is a difference.

- Randomly assign 11 players to run a wide angle and 11 players to run a narrow angle. Problems: some players may be faster than others. Ideally, randomization will equally distribute the speedy runners between the two groups, but there is no guarantee. Speed could be a confounding variable.
- Have each of the 22 runners run both angles, with the angle run first randomized using a coin. This allows each player to serve as their own control.

The second option is what the thesis writer did–he randomly determined the angle the player would take first. He then used a stopwatch the time the run from going from a spot 35 feet past home to a spot 15 feet before 2nd base. After a rest period, the runner then ran the second angle. This controls for runner-to-runner variability. It's important to randomize the order of the treatments, where possible! (This isn't possible in before-and-after type studies.)






```
[1] 0.075
```

Parameter of interest:

Hypotheses of interest:

Observed statistic:

Like before, we're trying to determine if it's surprising to see such a large difference as $\bar{x}_d = 0.075$ just by chance, if running strategy has no effect on running time.

```
t.test(bases$diff)
```

```

    One Sample t-test

data:  bases$diff
t = 3.9837, df = 21, p-value = 0.0006754
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.03584814 0.11415186
sample estimates:
```

```
mean of x
    0.075
```

```r
t.test(time~angle,data=bases2)
```

```
    Welch Two Sample t-test

data:  time by angle
t = 0.93383, df = 41.899, p-value = 0.3557
alternative hypothesis: true difference in means between group narrow and group wide is not equal to
95 percent confidence interval:
 -0.08709334  0.23709334
sample estimates:
mean in group narrow   mean in group wide
            5.534091             5.459091
```

4.4 Comparing Variances

Often it is useful to test if variances from independent populations are different. For example,

- a geneticist wants to test equality of the genotypic variances of kernel weight of two different corn populations
- an engineer is interested in comparing the process variance of two different types of production systems used to make a electronic component
- the two-sample t-test is based on the assumption that the variances of the two populations are equal

Assume the data from both populations follow a normal distribution with different means and possibly different variances. We want to test

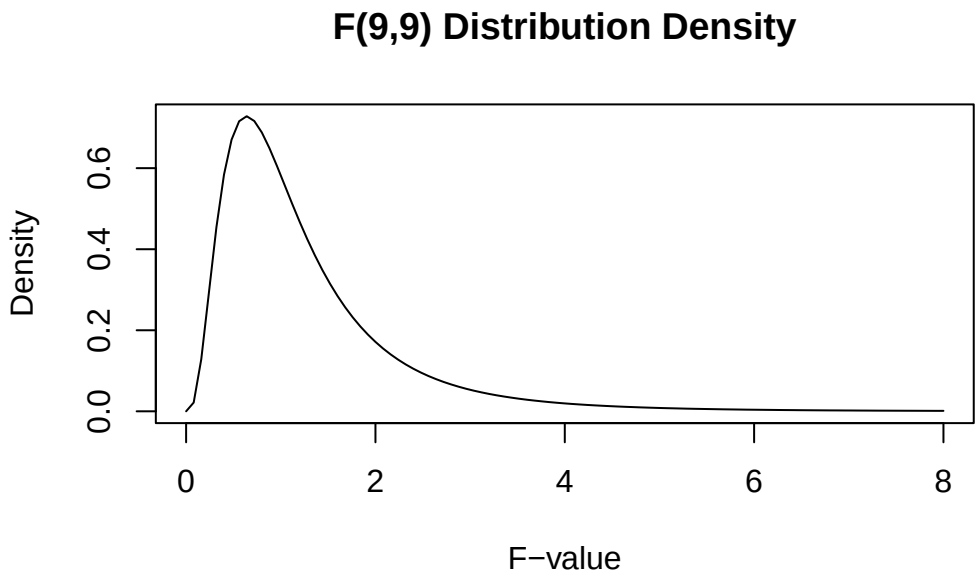
A natural approach would be to take samples of n_1 and n_2 observations from the two populations, and compute s_1^2 and s_2^2. We could then take the ratio s_1^2/s_2^2 and reject H_0 if the ratio is very different from 1. But, we need to know the sampling distribution of the ratio S_1^2/S_2^2. Recall

Sir R. A. Fisher showed that the ratio of two independent χ^2 distributions has an F distribution with $(n_1 - 1)$ and $(n_2 - 1)$ degrees of freedom. Specifically,

Under $\mathrm{H}_0 : \sigma_1^2 = \sigma_2^2$ then

The F distribution

- is non-negative, unimodal, and right skewed



- the shape of the distribution depends on the numerator and denominator degrees of freedom

So, to test $\mathrm{H}_0 : \sigma_1^2 = \sigma_2^2$ versus $\mathrm{H}_a : \sigma_1^2 \neq \sigma_2^2$, we can

- Assume that S_1^2 is the larger of the two sample variaces
- Use S_1^2/S_2^2 as a test statistic. Under H_0, this ratio will follow an F distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom
- Use the F distribution to see if s_1^2/s_2^2 is enough bigger than 1 to convince us the null hypothesis is not true (always a right-tail test!)

Example: The writings of different authors can be partially characterized by the variablity in the lengths of their sentences. Two manuscripts, A and B, are found by a historian and they want to know whether they have the same author. Fifteen sentences from each are chosen at random, and word counts per sentence are recorded. The historian finds $s_A^2 = 0.114$ and $s_B^2 = 0.143$.

We can use `var.test()` in R, but must have the whole data set.

Example: Earlier, we used a data set with nutrition levels in people's blood as well as information about their eating habits that came from a random sample of 315 US adults.

```
NutritionStudy<-read.csv("NutritionStudy.csv",header=TRUE)

head(NutritionStudy)
```

```
  ID Age Smoke Quetelet Vitamin Calories  Fat Fiber Alcohol Cholesterol
1  1  64    No  21.4838       1   1298.8 57.0   6.3     0.0       170.3
2  2  76    No  23.8763       1   1032.5 50.1  15.8     0.0        75.8
3  3  38    No  20.0108       2   2372.3 83.6  19.1    14.1       257.9
4  4  40    No  25.1406       3   2449.5 97.5  26.5     0.5       332.6
5  5  72    No  20.9850       1   1952.1 82.6  16.2     0.0       170.8
6  6  40    No  27.5214       3   1366.9 56.0   9.6     1.3       154.6
  BetaDiet RetinolDiet BetaPlasma RetinolPlasma    Sex VitaminUse PriorSmoke
1     1945         890        200           915 Female    Regular          2
2     2653         451        124           727 Female    Regular          1
3     6321         660        328           721 Female Occasional          2
4     1061         864        153           615 Female         No          2
5     2863        1209         92           799 Female    Regular          1
6     1729        1439        148           654 Female         No          2
```

The Quetelet index is a measure of body mass (BMI). Suppose we are interested in whether smokers and nonsmokers have the same variability of BMI scores.

```
var.test(Quetelet~Smoke,data=NutritionStudy)
```

```
    F test to compare two variances

data:  Quetelet by Smoke
F = 1.563, num df = 271, denom df = 42, p-value = 0.08157
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.9438906 2.3908039
sample estimates:
ratio of variances
          1.563047
```

Now that we've covered one predictor variable with two levels, we can move on to one predictor variable with more than 2 levels.

5 One Predictor/Explanatory Variable, More Than Two Levels

As mentioned at the end of Chapter 4, we'll now move on to cases in which we have a single explanatory variable and a single response variable. In this section, we'll cover the case where the explanatory variable is categorical with more than two levels, and the response variable is either categorical or numeric. This means that in this chapter, we'll be focusing on comparing more than two groups

Data like these may show up in a spreadsheet like

5.1 Categorical Response, More Than Two Levels

First, we'll consider situations in which two categorical variables are measured on each unit in the sample, and each variable has potentially more than two possible values. Many categorical variables have more than two possible outcomes, so we can't easily define the proportion of "successes." Instead, we'll summarize categorical data with more than two levels using two-way tables. In this class, we're still going to restrict ourselves to only two variables (often explanatory and response, but not necessarily), both with two or more levels. However, there are certainly statistical methods for more complicated situations.

Typically, research questions focus on how the proportions of the possible outcomes in the response variable change (or don't) across the levels of the explanatory variable. However, we can also consider questions about a single variable with more than two outcomes (are the possible outcomes all equally likely? do the possible outcomes follow a particular pattern? We've already seen these!) or just whether the two categorical variables are independent or dependent without assigning an explanatory/response relationship. Due to the structure of the variable(s), there really isn't a population parameter of interest. We can't (usually) make a function of proportion of successes that makes sense to estimate, like we can with $p_1 - p_2$. That means we'll be considering only tests, not confidence intervals.

Example: When surveys are administered, we hope that the respondents give accurate answers. Does the mode of survey delivery affect this? Schober et al (2015) investigated this question. They had 147 people who agreed to be interviewed on an iPhone, and they were randomly assigned to one of three interview modes: human voice, automated voice, text. One question asked was whether they exercise less than once per week during a typical week (a yes is mostly likely considered socially undesirable). The explanatory variable here is survey mode and the response is whether or not the respondent said yes. Here are the data:

Table 5.1: Survey Mode Data

	Text	Human Voice	Automated Voice	Total
Exercise Yes	34	21	20	75
Exercise No	124	139	139	402
Total	158	160	159	477

Based on these data, it looks like the answer to the question does change depending on survey mode, with respondents more likely to say yes via text. However, we don't know if this result could have happened by chance.

We saw expected counts when we did χ^2 goodness of fit tests. We'll need to find them again here. We don't expect the proportion of 'yes' to be exactly the same across all survey modes, but we want to know if these vary enough to convince us that survey mode and answer are not independent. To do this, we need to find **expected counts** for each cell in the table.

So,

$$\text{Expected Count}_{\text{row } i,\text{col } j} = \frac{(\text{row } i \text{ total})(\text{col } j \text{ total})}{\text{table total}}$$

Table 5.2: Survey Model Data with Expected Counts

	Text	**Human Voice**	**Automated Voice**	**Total**
Exercise Yes	34 (_____)	21 (_____)	20 (_____)	75
Exercise No	124 (_____)	139 (_____)	139 (_____)	402
Total	158	160	159	477

So just like with the goodness-of-fit test, the key question is whether the observed and expected cell counts are different enough.

- Cell(1,1) obs - exp = 34 -
- Cell(1,2) obs - exp = 21 -
- Cell(1,3) obs - exp = 20 -
- Cell(2,1) obs - exp = 124 -
- Cell(2,2) obs - exp = 139 -
- Cell(2,3) obs - exp = 139 -

Our χ^2 test statistic gets just a little more complicated:

In our example:

- Cell(1,1) (obs - exp)^2/exp $= (34 - 24.84)^2/(24.84) = 9.16^2/24.84 = 3.3778$
- Cell(1,2) (obs - exp)^2/exp $= (21 - 25.16)^2/(25.16) = (-4.16)^2/25.16 = 0.6878$
- Cell(1,3) (obs - exp)^2/exp $= (20 - 25)^2/(25) = (-5)^2/25 = 1$
- Cell(2,1) (obs - exp)^2/exp $= (124 - 133.16)^2/(133.16) = (-9.16)^2/133.16 = 0.6301$
- Cell(2,2) (obs - exp)^2/exp $= (139 - 134.84)^2/(134.84) = 4.16^2/134.84 = 0.1283$
- Cell(2,3) (obs - exp)^2/exp $= (139 - 134)2/(134) = 52/134 = 0.3731$

We already know this test statistic will follow a χ^2 distribution, but now

Again, we have conditions that need to be met for the χ^2 distribution to work:

-
-

Example: First, we'll need to check the conditions:

-
-

To find the p-value, we can use `pchisq(6.1971,df=2,lower.tail=FALSE)` =

We can also do the test directly in R:

```
surveymodetable<-read.csv("surveymodetable.csv",row.names=1)
surveymodetable
```

```
    Text Hvoice Avoice
Yes   34     21     20
No   124    139    139
```

```
chisq.test(surveymodetable)
```

```
    Pearson's Chi-squared test

data:  surveymodetable
X-squared = 6.0069, df = 2, p-value = 0.04962
```

Example: Integrated Pest Management (IPM) adopters apply significantly less insecticides and fungicides than nonadopters among grape producers. A 2008 paper published in *Agricultural Economics* gave data on IPM adoption rates for the six states that accounted for most of the US grape production. The data are in the file 'ipmtable.csv'.

```
ipmtable<-read.csv("ipmtable.csv",row.names=1)
ipmtable
```

```
          Cal Mich NewYork Oregon Penn Wash
Adopted    39   55      19     22   24   30
NotAdopt   92   69     114     88   83   77
```

```
chisq.test(ipmtable)
```

```
    Pearson's Chi-squared test

data:  ipmtable
X-squared = 34.59, df = 5, p-value = 1.816e-06
```

Example: A study of drinking habits of college students at a particular college produced the two-way table found in 'drinkingtable.csv' and shown below. Students were randomly selected to participate in the survey.

Table 5.3: Drinking habits of college students

	Reside on campus	**Reside off campus, not with parents**	**Reside off campus, with parents**	**Total**
Abstain from drinking	46	17	43	106
Light or moderate drinking	126	72	68	266
Heavy drinking	130	52	32	214
Total	302	141	143	586

```
drinktable<-read.csv("drinkingtable.csv",row.names=1)
drinktable
```

```
             OnCampus OffNoParents OffWithParents
Abstain            46           17             43
LightModerate     126           72             68
Heavy             130           52             32
```

What's different about this example?

This leads to hypotheses

```
chisq.test(drinktable)
```

```
    Pearson's Chi-squared test

data:  drinktable
X-squared = 28.949, df = 4, p-value = 8.007e-06
```

5.2 Quantitative Response

We're going to start this section by considering an example. The data are in the file 'mice.csv.'

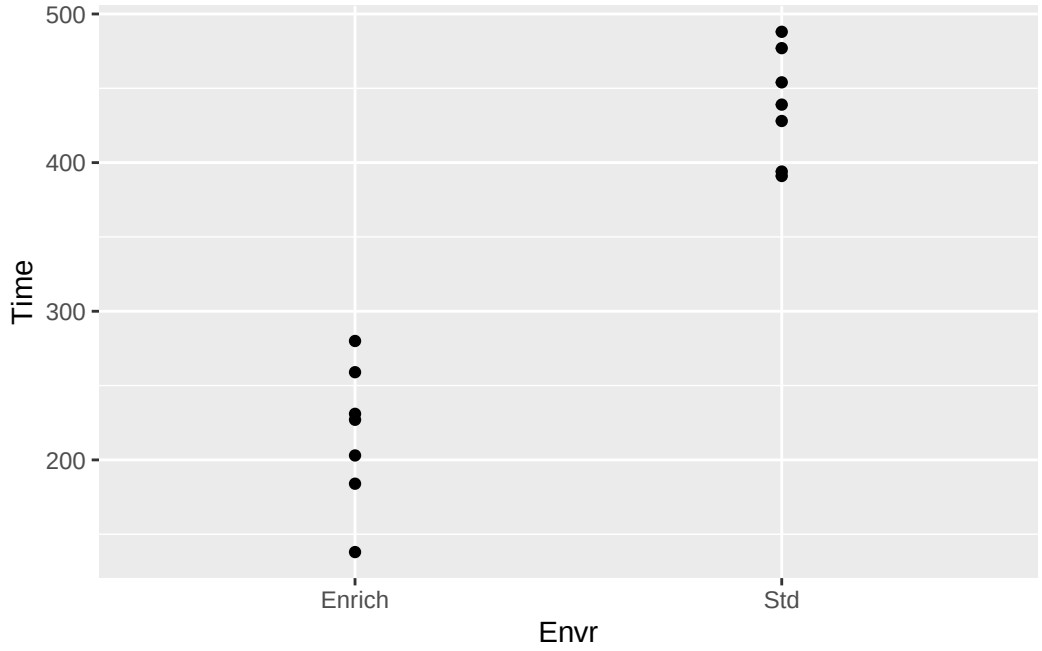
Example: These data come from an experiment to determine if exercise confers some resilience to stress. Mice were randomly assigned to either an enriched environment (exercise wheel) or standard environment, and spent three weeks there. After that time, they were exposed for five minutes per day for two weeks to a "mouse bully"–a mouse very strong, aggressive, and territorial. After those two weeks, anxiety in the mice was measured, as amount of time hiding in a dark compartment. Mice that are more anxious spend more time in darkness. We want to determine if there is a difference in time spent in darkness for the two groups of mice.

```
mice<-read.csv("mice.csv",header=TRUE)
head(mice)
```

```
    Envr Time
1 Enrich  259
2 Enrich  280
3 Enrich  138
4 Enrich  227
5 Enrich  203
6 Enrich  184
```

We already know how to answer this research question!

Let's first plot the data



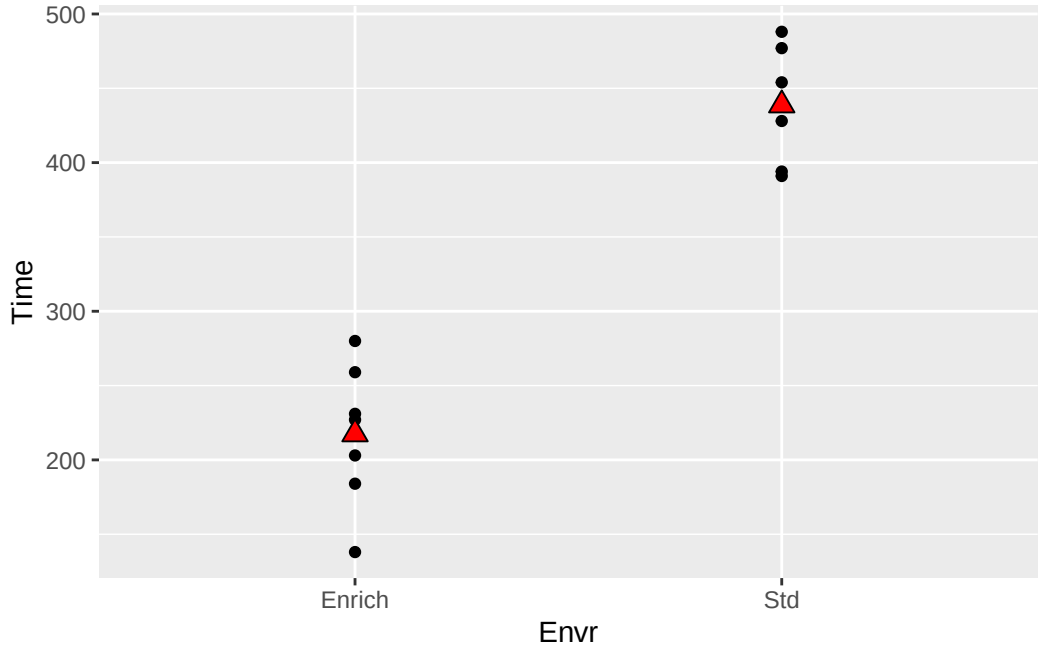
It definitely looks like there's a difference between the groups! We can find the group means and standard deviations. We'll also add the sample means to the plot.

```
aggregate(mice$Time, by=list(mice$Envr), FUN=mean)
```

```
  Group.1        x
1  Enrich 217.4286
2     Std 438.7143
```

```
aggregate(mice$Time, by=list(mice$Envr), FUN=sd)
```

```
  Group.1        x
1  Enrich 47.52844
2     Std 37.68162
```

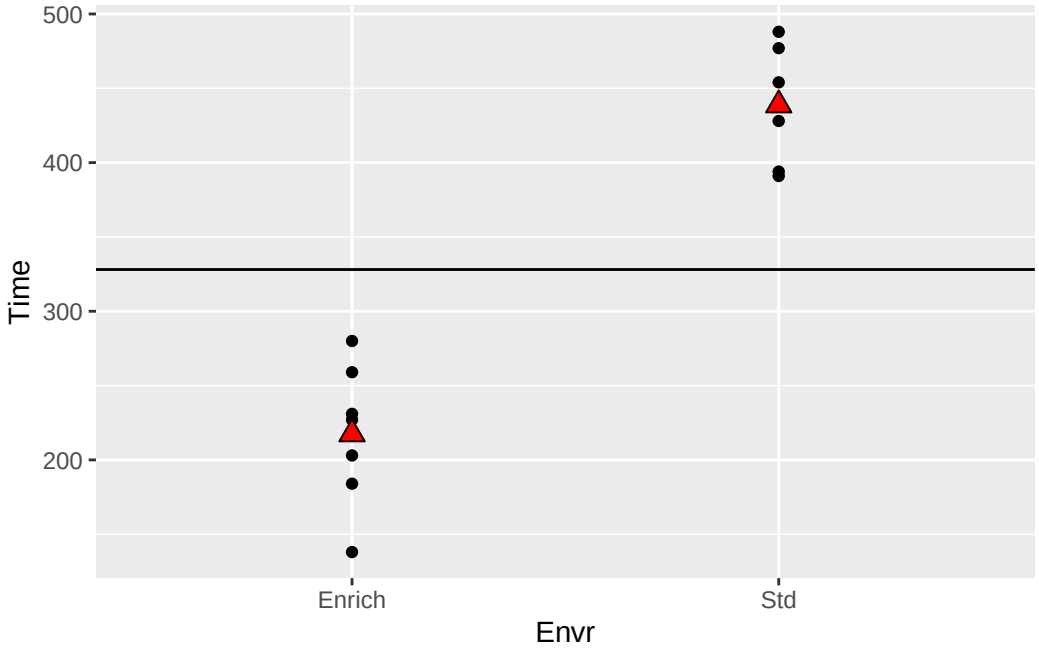


We're testing $H_0 : \mu_1 = \mu_2$, and assume this is true to construct the test. The overall common sample mean is $\bar{x} = 328.07$.

```
t.test(Time~Envr,data=mice)
```

```
    Welch Two Sample t-test

data:  Time by Envr
t = -9.6526, df = 11.407, p-value = 7.885e-07
alternative hypothesis: true difference in means between group Enrich and group Std is not equal to
95 percent confidence interval:
 -271.5245 -171.0470
sample estimates:
mean in group Enrich    mean in group Std
            217.4286             438.7143
```



It turns out the difference between the two groups will also manifest itself in the variances. There will be variation between the group means and the overall mean, as well as variation between the data points and their group means.

Remember how sample variance is calculated:

We're exploring how far, on average, observations are from the mean (squared). So, variance has to be positive. If there is a difference between the group means, the first kind of variation (between the group means and the overall mean) will be much greater than the second kind of variance (between the data points and their group mean). We can test whether the first variance is bigger than the second using an F statistic, just like we did in the last section when we were comparing two variances:

$$F = \frac{\text{variance between group means and overall mean}}{\text{variance between the data points and their group mean}}$$

If the variances are about equal, there's no evidence of a difference between the group means–they vary as much from the overall mean as data points vary from their group mean. This will result in an F statistic of about 1. If there is a difference between the group means, the first kind of variation (between the group means and the overall mean) will be much greater than the second kind of variance (between the data points and their group mean). This will result in an F statistic greater than 1.

For the mice data:

Notice!

We made some assumptions to carry out the t-test:

- approximate normality (no extreme outliers, no strong skew)
- independence between groups and between observations
- constant variance (we didn't make a big deal of this one, but mentioned it)

We can summarize these assumptions very succinctly, and to do so we're going to introduce some new notation.

Consider a random sample of observations from a normal distribution with mean μ and variance σ^2. If we let $Y_1, Y_2, \ldots, Y_n$ represent our data points we can summarize this as:

Or another way:

This is a **statistical model** with 2 parameters: μ and σ^2.

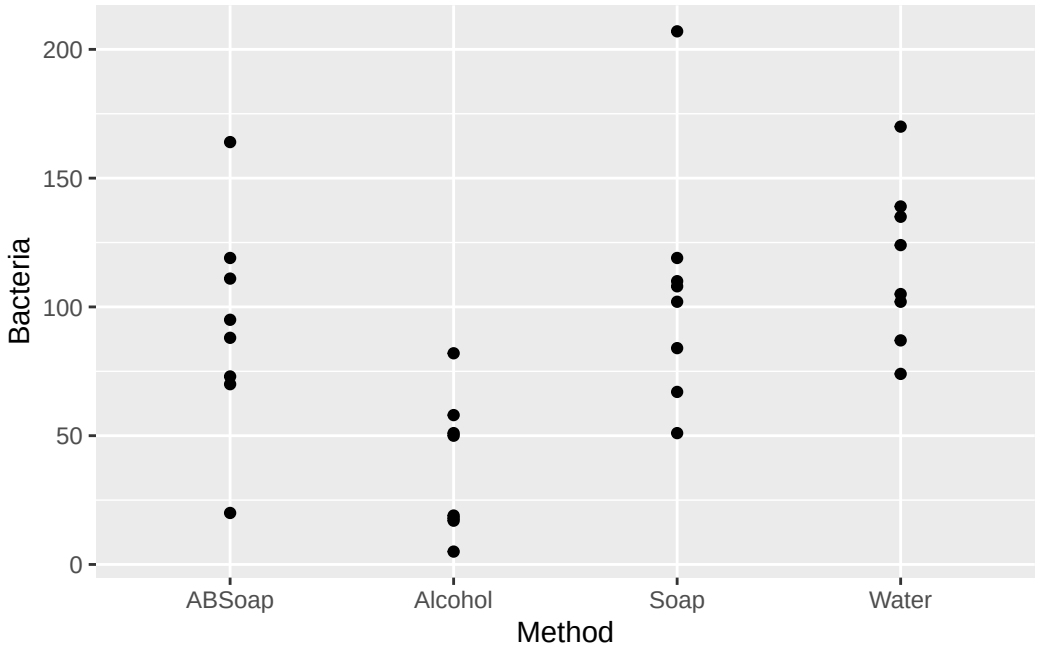
If we have two samples:

If we have more than two samples:

Let's start with some summary statistics

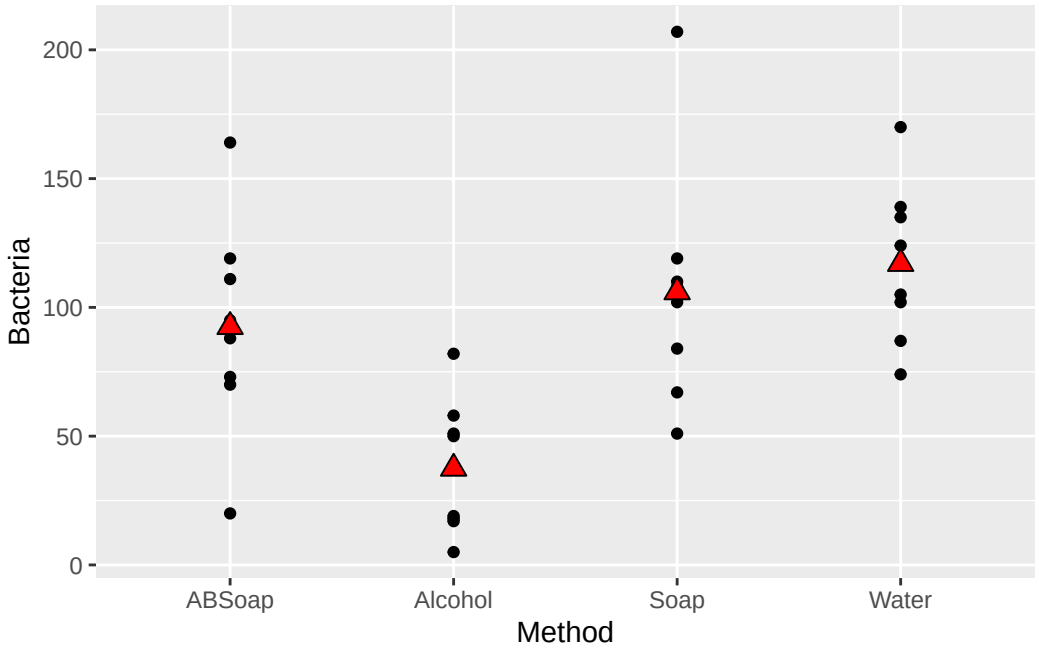
$$
\begin{aligned}
Y_{i\cdot} &= \sum_{j=1}^{n_i} Y_{ij} = i^{th} \text{ sample total} \\
\bar{Y}_{i\cdot} &= \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} = i^{th} \text{ sample mean} \\
Y_{\cdot\cdot} &= \sum_{i=1}^{t} \sum_{j=1}^{n_i} Y_{ij} = \text{ grand total} \\
\bar{Y}_{\cdot\cdot} &= \frac{1}{N} \sum_{i=1}^{t} \sum_{j=1}^{n_i} Y_{ij} = \text{ grand mean } (N = \sum_{i=1}^{n_i} n_i)
\end{aligned}
$$

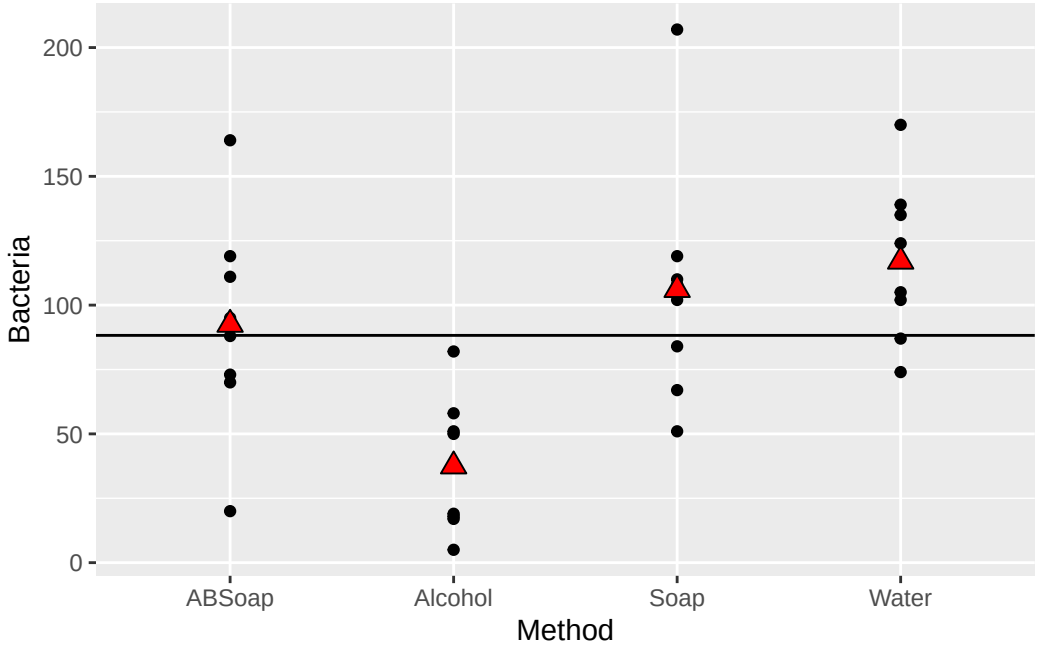
Example: A student carried out an experiment to investigate handwashing methods: water only, regular soap, antibacterial soap, and alcohol spray. Each treatment was replicated 8 times, and bacteria count was observed. The data are in 'handwash.csv'.




```
  Group.1     x
1  ABSoap  92.5
2 Alcohol  37.5
3    Soap 106.0
4   Water 117.0
```

```
  Group.1        x
1  ABSoap 41.96257
2 Alcohol 26.55991
3    Soap 46.95895
4   Water 31.13106
```


200
150
100
50
0
Bacteria
ABSoap
Alcohol
Soap
Water
Method


200
150
100
50
0
Bacteria
ABSoap
Alcohol
Soap
Water
Method

Remember how to calculate the sample variance, $S^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$. We're going to look at three difference variances. Let's assume for simplicity that $n_i = n$ (all groups have equal sample size, this is not really necessary, it's just to make it easier to look at notation):

1. Total Variance. Another name for the numerator is total sum of squares.

2. Error (Within-Group) Variance. Another name for the numerator is the error sum of squares.

3. Model (Between-Group) Variance. Another name for the numerator is the treatment (model) sum of squares.

To see what this is measuring, first consider the 'inside' sum:

This is still an estimate of variance, but it's an estimate of σ^2/n, because these are means. In order to be able to compare fairly to the error variance we must multiply by n (only works with equal sample sizes) or, equivalently, take the sum from $j = 1$ to n:

We can't lose sight of what we're interested in here: testing $H_0 : \mu_1 = \mu_2$. If H_0 is true, $\bar{y}_1$ and $\bar{y}_2$ should not be different from $\bar{y}_{\cdot\cdot}$. This means that error variance should be about equal to model variance (both would estimate σ^2). If H_0 is not true, model variance will be larger because of the deviations of the group averages from the grand average. If it's much larger, this gives us evidence against H_0.

Why do we worry about three variances when we only use two (error and model) to get the F stat? It turn out that:

$$\text{Total SS} = \text{Model SS} + \text{Error SS}$$

For the mice data:

$$\begin{aligned}
\text{Total SS} &= (259-328.07)^2 + \cdots + (231-328.07)^2 + (394-328.07)^2 + \cdots + (454-328.07)^2 = 193459 \\
\text{Error SS} &= (259-217.43)^2 + \cdots + (231-217.43)^2 + (394-438.71)^2 + \cdots + (454-438.71)^2 = 22073 \\
\text{Model SS} &= 6(217.43-328.07)^2 + 6(438.71-328.07)^2 = 171386
\end{aligned}$$

To convert these sums of squares into variances (which we call mean squares), they must be divided by denominators noted above. These are degrees of freedom, and have the same relationship as the sums of squares do:

$$\text{Total } df = \text{ Model } df + \text{ Error } df$$

In our mice example, we have

$$\text{Total } df = \text{ Model } df + \text{ Error } df$$

We often summarize our calculations in a table (df assuming equal sample sizes):

Source	df	SS	MS
Model	$t-1$	SSModel	MSModel
Error	$t(n-1)$	SSError	MSError
Total	$nt-1$	SSTotal	

The MSError (usually called MSE) is our estimate of σ^2. In our mice example, we get the table:

Source	df	SS	MS
Model	1	171386	171386
Error	12	22073	1839
Total	13	193459	

To test $H_0 : \mu_1 = \mu_2$ we use the F stat:

$$F = \frac{\text{MSModel}}{\text{MSError}} = \frac{171386}{1839} = 93.2$$

and we can add this to the table:

Source	df	SS	MS	F
Model	1	171386	171386	93.2
Error	12	22073	1839	
Total	13	193459		

What we've just done is called an **Analysis of Variance (ANOVA)**, and the resulting table is called an ANOVA table. It's a single hypothesis test to check whether the means across many groups are equal. Specifically, it's testing:

We still have assumptions: - Independence between and among groups - Responses/errors are approximately normal - Variability across groups is about equal

We already know how to determine if $F = 93.2$ is enough greater than 1 to determine there's a difference–the F distribution we used to test the equality of two variances in the last section. Our numerator and denominator degrees of freedom will be the Model df and Error df, respectively:

```
pf(93.2,df1=1,df2=12,lower.tail=FALSE)
```

```
[1] 5.232224e-07
```

The p-value typically gets added to the table as well:

Source	df	SS	MS	F	p-value
Model	1	171386	171386	93.2	0.0000005
Error	12	22073	1839		
Total	13	193459			

This is the only time we'll do an ANOVA by hand! Let's do the same in R.

```
anova(lm(Time~Envr, data=mice))
```

```
Analysis of Variance Table

Response: Time
          Df Sum Sq Mean Sq F value   Pr(>F)
Envr       1 171386  171386  93.173 5.24e-07 ***
Residuals 12  22073    1839
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example: Let's now carry out the ANOVA on the handwashing data. We'll start by writing the model and sketching the ANOVA table.

```
anova(lm(Bacteria~Method,data=handwash))
```

```
Analysis of Variance Table

Response: Bacteria
          Df Sum Sq Mean Sq F value   Pr(>F)
Method     3  29882  9960.7  7.0636 0.001111 **
Residuals 28  39484  1410.1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We could also use SAS:

```
PROC IMPORT OUT= WORK.mice
DATAFILE= "C:\Users\Erin\OneDrive - University of Nebraska-Lincoln\STAT 801\Book Notes\mice.csv"
DBMS=CSV REPLACE;
GETNAMES=YES;
DATAROW=2;
RUN;

proc glimmix data=mice;
  class Envr;
  model Time=Envr;
run;
```

SAS `proc glimmix` uses a different numerical method to calculate the ANOVA, and so the SSTrt/SSError don't exist in the same way.

The GLIMMIX Procedure

Model Information	
Data Set	WORK.MICE
Response Variable	Time
Response Distribution	Gaussian
Link Function	Identity
Variance Function	Default
Variance Matrix	Diagonal
Estimation Technique	Restricted Maximum Likelihood
Degrees of Freedom Method	Residual

Class Level Information		
Class	Levels	Values
Envr	2	Enrich Std

Number of Observations Read	14
Number of Observations Used	14

Dimensions	
Covariance Parameters	1
Columns in X	3
Columns in Z	0
Subjects (Blocks in V)	1
Max Obs per Subject	14

Optimization Information	
Optimization Technique	None
Parameters	3
Lower Boundaries	1
Upper Boundaries	0
Fixed Effects	Not Profiled

Fit Statistics	
-2 Res Log Likelihood	128.15
AIC (smaller is better)	134.15
AICC (smaller is better)	137.15
BIC (smaller is better)	135.61
CAIC (smaller is better)	138.61
HQIC (smaller is better)	133.61
Pearson Chi-Square	22073.14
Pearson Chi-Square / DF	1839.43

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
Envr	1	12	93.17	<.0001

For the handwash data:

```
PROC IMPORT OUT= WORK.handwash
DATAFILE= "C:\Users\Erin\OneDrive - University of Nebraska-Lincoln\STAT 801\Book Notes\handwash.csv'
DBMS=CSV REPLACE;
GETNAMES=YES;
DATAROW=2;
RUN;

proc glimmix data=handwash;
  class Method;
  model Bacteria=Method;
run;
```

The GLIMMIX Procedure

Model Information	
Data Set	WORK.HANDWASH
Response Variable	Bacteria
Response Distribution	Gaussian
Link Function	Identity
Variance Function	Default
Variance Matrix	Diagonal
Estimation Technique	Restricted Maximum Likelihood
Degrees of Freedom Method	Residual

Class Level Information		
Class	Levels	Values
Method	4	ABSoap Alcohol Soap Water

Number of Observations Read	32
Number of Observations Used	32

Dimensions	
Covariance Parameters	1
Columns in X	5
Columns in Z	0
Subjects (Blocks in V)	1
Max Obs per Subject	32

Optimization Information	
Optimization Technique	None
Parameters	5
Lower Boundaries	1
Upper Boundaries	0
Fixed Effects	Not Profiled

Fit Statistics	
-2 Res Log Likelihood	290.82
AIC (smaller is better)	300.82
AICC (smaller is better)	303.55
BIC (smaller is better)	307.48
CAIC (smaller is better)	312.48
HQIC (smaller is better)	302.86
Pearson Chi-Square	39484.00
Pearson Chi-Square / DF	1410.14

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
Method	3	28	7.06	0.0011

The reason I like SAS for ANOVA is because we can easily add fanciness:

```
proc glimmix data=handwash;
  class Method;
  model Bacteria=Method;
  lsmeans Method/pdiff cl;
run;
```

Method Least Squares Means

Method	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
ABSoap	92.5000	13.2766	28	6.97	<.0001	0.05	65.3041	119.70
Alcohol	37.5000	13.2766	28	2.82	0.0086	0.05	10.3041	64.6959
Soap	106.00	13.2766	28	7.98	<.0001	0.05	78.8041	133.20
Water	117.00	13.2766	28	8.81	<.0001	0.05	89.8041	144.20

Differences of Method Least Squares Means

Method	_Method	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
ABSoap	Alcohol	55.0000	18.7759	28	2.93	0.0067	0.05	16.5392	93.4608
ABSoap	Soap	-13.5000	18.7759	28	-0.72	0.4781	0.05	-51.9608	24.9608
ABSoap	Water	-24.5000	18.7759	28	-1.30	0.2026	0.05	-62.9608	13.9608
Alcohol	Soap	-68.5000	18.7759	28	-3.65	0.0011	0.05	-106.96	-30.0392
Alcohol	Water	-79.5000	18.7759	28	-4.23	0.0002	0.05	-117.96	-41.0392
Soap	Water	-11.0000	18.7759	28	-0.59	0.5627	0.05	-49.4608	27.4608

Next, we'll add some more details and formality to the ANOVA.

5.3 The Completely Randomized Design

Way back in the first section, we talked about the steps in a statistical investigation

- 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

As part of Step 2, we noted that we need to consider questions like 'what variable(s) wil be measured'. This basically involves identifying the response variable as well as any explanatory variable(s). Now, let's introduce some new terminology that really only becomes relevant once we are doing ANOVA.

Example: Handwash, again. The student considered four treatments: water only, regular soap, antibacterial soap, and alcohol spray.

In this example, there is one **factor.**

Definition:

In order to study the effect of the factor on the response, two or more values of the factor are considered. These values are called **levels**.

In some cases, there is more than one factor.

Example: Two students at Queensland University of Technology, as a project for their statistics class, carried out an experiment to test the effect certain factors such as refrigeration, stem length, and water content have on the life of a cut rose. The students considered

- Stem length (15 cm or 25 cm)
- Water content (tap water or tap water + citric acid)
- Temperature (refrigerated or room temperature)

The response measured was the number of days until death, and the goal was to determine the conditions that will extend rose life.

In this example, there are 3 factors:

Each factor has 2 levels:

In multifactor experiments like this, we define a **treatment** as a combination of factor levels.

- Factors:

- Levels:

-Treatments:

We also have to consider the **treatment design** and the **experimental design.**

Definition: The **treatment design**

Definition: The **experimental design**

The experimental design should address the three basic principles underlying formal experimentation:

- Replication: a repetition of the basic experiment

- Randomization: both allocation of experimental material and order in which individual runs/trials are performed

- Control: control the effect of extraneous variables

The first **experimental design** we'll consider is the **completely randomized design** or CRD. The CRD is an experimental design because

The CRD is characterized by

The CRD may be combined with several different **treatment designs**. To explore the CRD in more detail, we'll start with the simplest treatment design, the **one-way design**. The one-way design is so named because

Within one-way designs there are four basic treatment structures:

1. Unstructured
2. Control versus other treatments
3. Quantitative
4. Other structure

Example: Handwash, again. The student considered four treatments: water only, regular soap, antibacterial soap, and alcohol spray. The student replicated each treatment 8 times.

- **Treatment Design:**
  - Factor:
  - Levels:
- **Experimental Design:**
  - Run 8 trials in a Completely Randomized Design

Here's one possible sequence of trials:

AL	RS	AB	W	W	AL	AB	RS	RS	RS	RS	AL	RS	AB	AB	AB	W	W	AB	AB	AL	W	RS	RS	AB	AL	W	W	AL	W	AL	AL

5.3.1 CRD Model and Basic Analysis

The **CRD Model** can be written in two different ways.

- y_{ij} = bacteria count for the j^{th} trial after handwashing the i^{th} method
- μ = overall mean bacteria count
- τ_i = treatment effect of method i = additional amount of bacteria observed using handwashing method i
- ϵ_{ij} = random error = additional amount of bacteria in the j^{th} trial using handwashing method i

Example: A donut manufacturer wants to see if the type of fat used to fry the donuts has any impact on the amount of fat absorbed by the donuts. The manufacturer has two types of animal fat and two types of vegetable fat that they would like to compare. They also have available 4 fryers, which can each fry 1 batch of 18 donuts at a time. They plan to measure the amount of fat absorbed in each batch.They have the resources to test 24 total batches of donuts.

- **Treatment Design:**
  - Factor:
  - Levels:
- **Experimental Design:**
  - Run 6 batches of each fat in a Completely Randomized Design

For this particular treatment design, there are several hypothesis tests that may be of interest. Write out in the symbols the null and alternative hypotheses for the following specified objectives. Reminder: Fats 1 and 2 are animal fats and Fats 3 and 4 are vegetable fats.

1. Are there differences among the four fats with respect to the amount of fat absorbed?

2. Do the vegetable fats differ from the animal fats in the amount of fat absorbed?

3. Are there differences between the two animal fats? Are there differences between the two vegetable fats?

We've already seen how to fit the basic ANOVA in R and SAS.

```
data donut;
  do type=1 to 4;
    input absorb @@;
    output;
  end;
   datalines;
   164 178 175 155
   172 191 193 166
   168 197 178 149
   177 182 171 164
   156 185 163 170
   195 177 176 168
   ;

proc glimmix data=donut;
  class type;
  model absorb=type;
  lsmeans type/pdiff cl;
run;
```

Fit Statistics	
-2 Res Log Likelihood	156.21
AIC (smaller is better)	166.21
AICC (smaller is better)	170.49
BIC (smaller is better)	171.19
CAIC (smaller is better)	176.19
HQIC (smaller is better)	167.18
Pearson Chi-Square	2018.00
Pearson Chi-Square / DF	100.90

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
type	3	20	5.41	0.0069

type Least Squares Means								
type	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
1	172.00	4.1008	20	41.94	<.0001	0.05	163.45	180.55
2	185.00	4.1008	20	45.11	<.0001	0.05	176.45	193.55
3	176.00	4.1008	20	42.92	<.0001	0.05	167.45	184.55
4	162.00	4.1008	20	39.50	<.0001	0.05	153.45	170.55

Differences of type Least Squares Means									
type	_type	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
1	2	-13.0000	5.7994	20	-2.24	0.0365	0.05	-25.0974	-0.9026
1	3	-4.0000	5.7994	20	-0.69	0.4983	0.05	-16.0974	8.0974
1	4	10.0000	5.7994	20	1.72	0.1001	0.05	-2.0974	22.0974
2	3	9.0000	5.7994	20	1.55	0.1364	0.05	-3.0974	21.0974
2	4	23.0000	5.7994	20	3.97	0.0008	0.05	10.9026	35.0974
3	4	14.0000	5.7994	20	2.41	0.0255	0.05	1.9026	26.0974

5.3.2 Treatment Comparisons and Contrasts

We can see in the results above that we may reject the overall hypothesis that the four treatments produce the same mean fat absorption ($F = 5.41$, p-value$= 0.0069$). But, this doesn't address the hypotheses you constructed earlier. Remember, we also considered:

- Do the vegetable fats differ from the animal fats in the amount of fat absorbed?
- Are there differences between the two animal fats?
- Are there differences between the two vegetable fats?

The output above allows us to address some of these questions, but not the one regarding vegetable fats versus animal fats. Let's look at a more general way to construct treatment comparisons.

Contrasts

A well-thought-out treatment design's objectives can usually be stated in terms of a set of **contrasts**. This is usually an important goal in planning the design, and contrasts are constructed before data are collected.

A **contrast** is

Estimates of the contrast are obtained by substituting in the sample means

We may also obtain standard errors of the contrast estimate

Standard errors may then be used to carry out tests and construct confidence intervals.

The contrasts of interest depend on the basic treatment design structure and the goals of the experiment. Remember, the four basic structures are

1. Unstructured
2. Control versus other treatments
3. Quantitative
4. Other structure

Let's first consider Unstructured designs, because these are the simplest.

5.3.2.1 Unstructured Treatment Designs and All Pairwise Comparisons

Example: Handwashing, again. The student considered four treatments: water only, regular soap, antibacterial soap, and alcohol spray. The student replicated each treatment 8 times.

- **Treatment Design:**
  - Factor:
  - Levels:
- **Experimental Design:**
  - Run 8 trials in a Completely Randomized Design

In designs like this without structure, we are typically interested in **all pairwise comparisons**.

There are multiple methods for making such comparisons. The simplest is the **least significant difference** (LSD), also called the unprotected LSD. It's easy, but the Type I error rate can be badly inflated (we'll talk more about this in a bit).

A (slightly) more conservative option is **Fisher's protected LSD**.

We've already seen these, but let's add even more fanciness!

```
proc glimmix data=handwash;
  class Method;
  model Bacteria=Method;
  lsmeans Method/pdiff cl lines plot=diffplot;
run;
```

Type III Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
Method	3	28	7.06	0.0011

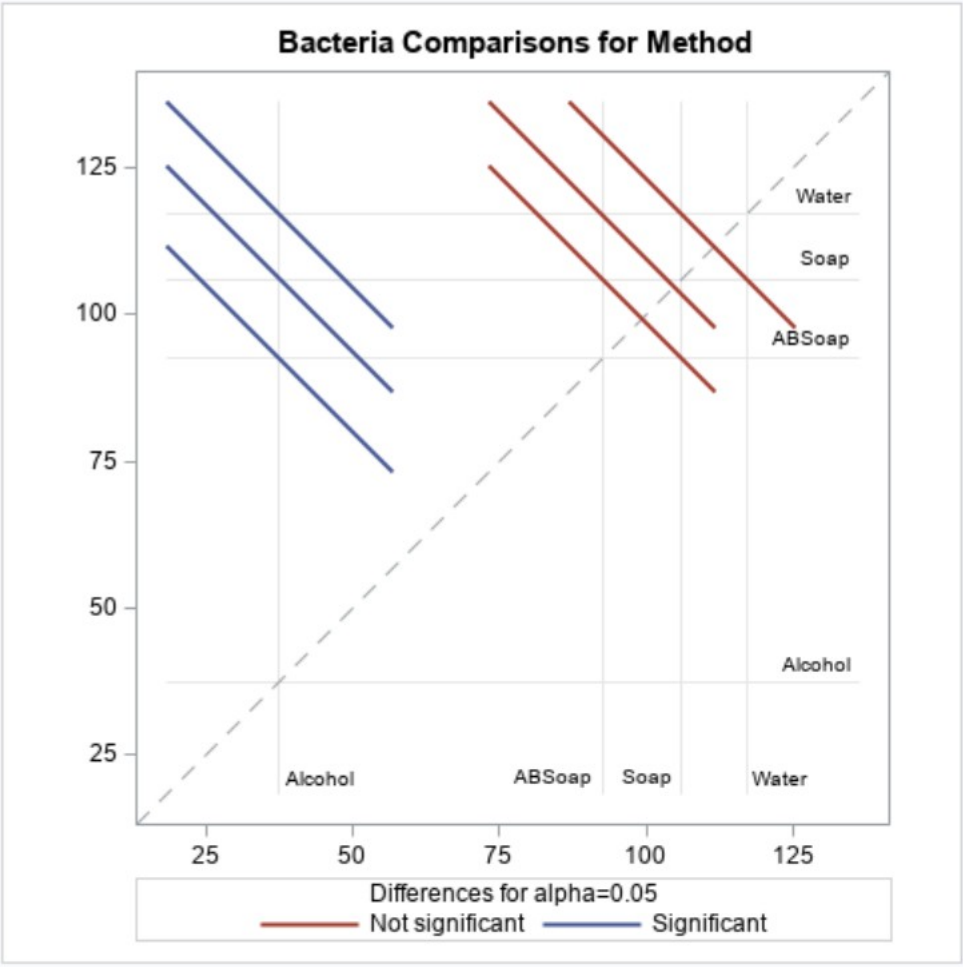
Method Least Squares Means

Method	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
ABSoap	92.5000	13.2766	28	6.97	<.0001	0.05	65.3041	119.70
Alcohol	37.5000	13.2766	28	2.82	0.0086	0.05	10.3041	64.6959
Soap	106.00	13.2766	28	7.98	<.0001	0.05	78.8041	133.20
Water	117.00	13.2766	28	8.81	<.0001	0.05	89.8041	144.20

Differences of Method Least Squares Means

Method	_Method	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
ABSoap	Alcohol	55.0000	18.7759	28	2.93	0.0067	0.05	16.5392	93.4608
ABSoap	Soap	-13.5000	18.7759	28	-0.72	0.4781	0.05	-51.9608	24.9608
ABSoap	Water	-24.5000	18.7759	28	-1.30	0.2026	0.05	-62.9608	13.9608
Alcohol	Soap	-68.5000	18.7759	28	-3.65	0.0011	0.05	-106.96	-30.0392
Alcohol	Water	-79.5000	18.7759	28	-4.23	0.0002	0.05	-117.96	-41.0392
Soap	Water	-11.0000	18.7759	28	-0.59	0.5627	0.05	-49.4608	27.4608

T Grouping for Method Least Squares Means (Alpha=0.05)		
LS-means with the same letter are not significantly different.		
Method	**Estimate**	
Water	117.00	A
		A
Soap	106.00	A
		A
ABSoap	92.5000	A
Alcohol	37.5000	B



This plot is called a **diffogram** and is a way to visualize differences among the treatments.

So these plots are awesome, and the output is easy to interpret! Why do we care about anything other than the LSD? The big issue is Type I error rate, and it can be a concern for pairwise comparisons as well as more complicated contrasts.

Multiple Comparisons

If more than one comparison is made among the treatment means, then we have multiple comparisons which can lead to the problem of **multiplicity**.

Definition: Multiplicity is

For a single test, the significance level of a Type I error is called a **comparison-wise** error rate. This means

But, if we have multiple tests, the Type I errors for these tests accumulate. This accumulated rate is the called the **experiment-wise** error rate. This is

But, the errors don't just add up. They accumulate in a power-type relationship. Consider a situation with a comparison-wise error rate of α and c independent comparisons. Then, the experiment-wise error rate is

For example, consider a situation with $\alpha = 0.05$ and 5 independent comparisons (there are as many independent comparisons as there are df for treatment). In that case:

We can control the experiment-wise error rate by setting it to a pre-specified value α (maybe 0.05) and then solving for the comparison-wise error rate, assuming c independent comparisons. So, for example, if $\alpha = 0.05$ and $c = 5$,

We'd then use this as the critical value (cut-off) value for our independent treatment comparisons.

But here's another issue. If the comparisons are not independent (which they aren't in all pairwise-comparisons, and often aren't in pre-planned contrasts of interest), then the experiment-wise error rate is actually even bigger than we see above. What can we do?

There are a multitude of multiple comparison procedures which control the overall experiment-wise error rate, which have different pros and cons. We're only going to the talk about a few.

Tukey's HSD: Tukey's Honestly Significant Difference (HSD) procedure is based on the studentized range statistic. To get this HSD from SAS:

```
proc glimmix data=handwash;
  class Method;
  model Bacteria=Method;
  lsmeans Method/pdiff cl adjust=tukey;
run;
```

Differences of Method Least Squares Means
Adjustment for Multiple Comparisons: Tukey

Method	_Method	Estimate	Standard Error	DF	t Value	Pr > |t|	Adj P	Alpha	Lower	Upper	Adj Lower	Adj Upper
ABSoap	Alcohol	55.0000	18.7759	28	2.93	0.0067	0.0320	0.05	16.5392	93.4608	3.7358	106.26
ABSoap	Soap	-13.5000	18.7759	28	-0.72	0.4781	0.8887	0.05	-51.9608	24.9608	-64.7642	37.7642
ABSoap	Water	-24.5000	18.7759	28	-1.30	0.2026	0.5676	0.05	-62.9608	13.9608	-75.7642	26.7642
Alcohol	Soap	-68.5000	18.7759	28	-3.65	0.0011	0.0056	0.05	-106.96	-30.0392	-119.76	-17.2358
Alcohol	Water	-79.5000	18.7759	28	-4.23	0.0002	0.0012	0.05	-117.96	-41.0392	-130.76	-28.2358
Soap	Water	-11.0000	18.7759	28	-0.59	0.5627	0.9355	0.05	-49.4608	27.4608	-62.2642	40.2642

We could also request lines and the diffogram; they would be adjusted as well.

The other multiple comparison procedures we'll discuss are used with other treatment design structures. The three other one-way treatment design structures are:

1. Control versus other treatments
2. Quantitative (we'll put a pin in this one for now)
3. Other structure

5.3.2.2 Control versus other treatments

In some scenarios, one of the factor levels acts as a control treatment for some or all of the remaining levels. Often, we are interested in comparing all of the treatments against the control but not against each other. This means there are

Dunnett's procedures is a modification to the two-sample t test that is used when comparing all treatments against a control.

Example: Sections of tomato plant tissue were grown in culture with differing amounts and types of sugars with five replications of four treatments. The treatments were: control, 3% glucose, 3% fructose, and 3% sucrose.

- **Treatment Design:**
  - Factor:
  - Levels:
- **Experimental Design:**
  -

In a situation like this, we may be interested in comparing each of the sugar treatments to the control.

```
data tomato;
  input trt $ growth @@;
  datalines;
  control 45 glucose 25 fructose 28 sucrose 31
  control 39 glucose 28 fructose 31 sucrose 37
  control 40 glucose 30 fructose 24 sucrose 35
  control 45 glucose 29 fructose 28 sucrose 33
  control 42 glucose 33 fructose 27 sucrose 34
;
 proc glimmix data=tomato;
   class trt;
   model growth=trt;
   lsmeans trt/diff=control('control') cl adjust=dunnett plot=controlplot;
 run;
```

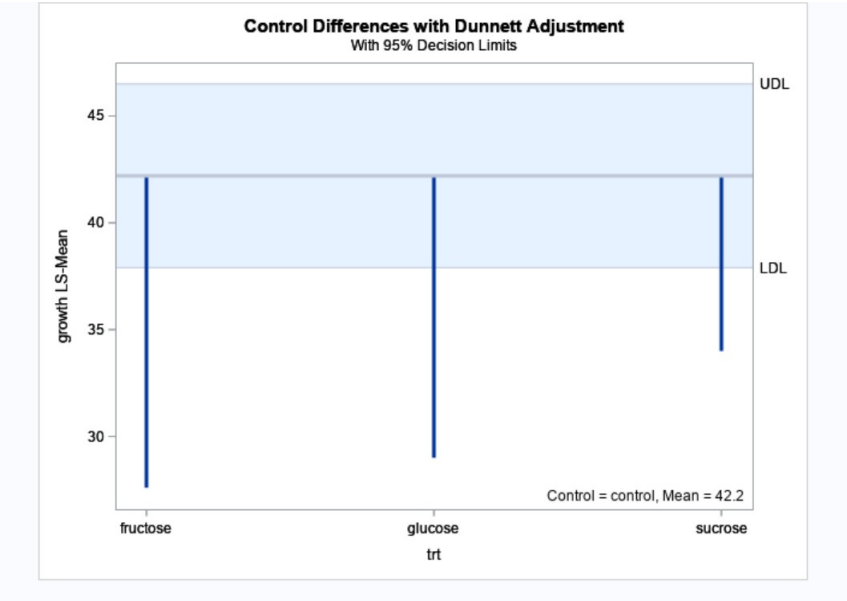
Note that unless otherwise specified, SAS will assume the first treatment level (alphabetically or numerically) is the control.

trt Least Squares Means

trt	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
control	42.2000	1.1726	16	35.99	<.0001	0.05	39.7142	44.6858
fructose	27.6000	1.1726	16	23.54	<.0001	0.05	25.1142	30.0858
glucose	29.0000	1.1726	16	24.73	<.0001	0.05	26.5142	31.4858
sucrose	34.0000	1.1726	16	29.00	<.0001	0.05	31.5142	36.4858

Differences of trt Least Squares Means
Adjustment for Multiple Comparisons: Dunnett

trt	_trt	Estimate	Standard Error	DF	t Value	Pr > |t|	Adj P	Alpha	Lower	Upper	Adj Lower	Adj Upper
fructose	control	-14.6000	1.6583	16	-8.80	<.0001	<.0001	0.05	-18.1155	-11.0845	-18.8990	-10.3010
glucose	control	-13.2000	1.6583	16	-7.96	<.0001	<.0001	0.05	-16.7155	-9.6845	-17.4990	-8.9010
sucrose	control	-8.2000	1.6583	16	-4.94	0.0001	0.0004	0.05	-11.7155	-4.6845	-12.4990	-3.9010



5.3.2.3 Treatment Designs with (other) Structure

This is where the donut example fits in. There isn't a true control, but we also may not care about all pairwise comparisons. Instead, we had some specific, pre-planned comparisons of interest:

- Do the vegetable fats differ from the animal fats in the amount of fat absorbed?
- Are there differences between the two animal fats?

- Are there differences between the two vegetable fats?

Why pre-plan comparisons?

Earlier, we wrote out the hypotheses of interest corresponding to these comparisons:

There are three options available in SAS to test these hypotheses and/or construct confidence intervals:

- `contrast` statement
- `estimate` statement
- `lsmestimate` statement

All three statements involve specifying the coefficients of the treatment effects/treatment means. Let's look at the comparison of vegetable and animal fats.

and two different contrast statements we could write:

```
proc glimmix data=donut;
  class type;
  model absorb=type;
  contrast "animal vs veg" type 1 1 -1 -1;
  contrast "animal vs veg 2" type 0.5 0.5 -0.5 -0.5;
run;
```

Both give the same results!

```
                      Contrasts

Num      Den
Label                 DF       DF    F Value    Pr > F

animal vs veg          1       20       5.37    0.0313
animal vs veg 2        1       20       5.37    0.0313
```

Let's try them with `estimate` statements, and add a third option:

```
estimate "animal vs veg" type 1 1 -1 -1;
estimate "animal vs veg 2" type 0.5 0.5 -0.5 -0.5;
estimate "animal vs veg 3" type 1 1 -1 -1/divisor=2;
```

```
                              Estimates

                              Standard
Label              Estimate       Error       DF    t Value    Pr > |t|

animal vs veg       19.0000      8.2016       20       2.32      0.0313
animal vs veg 2      9.5000      4.1008       20       2.32      0.0313
animal vs veg 3      9.5000      4.1008       20       2.32      0.0313

                  type Least Squares Means

                    Standard
type    Estimate       Error       DF    t Value    Pr > |t|

1         172.00      4.1008       20      41.94      <.0001
2         185.00      4.1008       20      45.11      <.0001
3         176.00      4.1008       20      42.92      <.0001
4         162.00      4.1008       20      39.50      <.0001
```

What's going on?

Suppose for some reason we wanted to test whether fats 1-3 (collectively) were different from fat 4.

The way we write the `estimate` statement really matters here:

```
estimate "first 3 vs last" type 0.33 0.33 0.33 -1;
estimate "first 3 vs last" type 1 1 1 -3/divisor=3;
```

```
                              Estimates

                              Standard
Label              Estimate       Error       DF    t Value    Pr > |t|

first 3 vs last     Non-est           .        .        .         .
first 3 vs last     15.6667      4.7352       20       3.31      0.0035
```

We do still have a multiplicity issue, because we are interested in three pre-planned contrasts. We can use the **Sidak** adjustment to control experiment-wise error rate:

```
estimate "1 vs 2" type 1 -1 0 0,
         "3 vs 4" type 0 0 1 -1,
         "animal vs veg" type 0.5 0.5 -0.5 -0.5/adjust=sidak;
```

```
                              Estimates

                                Standard
Label                Estimate       Error       DF    t Value    Pr > |t|

animal vs veg 2        9.5000      4.1008       20       2.32      0.0313
```

```
                              Estimates
                    Adjustment for Multiplicity: Sidak

                              Standard
Label              Estimate       Error       DF    t Value    Pr > |t|     Adj P

1 vs 2            -13.0000      5.7994       20      -2.24      0.0365    0.1055
3 vs 4             14.0000      5.7994       20       2.41      0.0255    0.0745
animal vs veg       9.5000      4.1008       20       2.32      0.0313    0.0909
```

Finally, we can use the `lsmestimate` statement. `lsmestimate` basically does the same thing as **estimate** but it allows for more complicated models than we have yet encountered. For a CRD, the output of the two should be identical, though `lsmestimate` does have some additional options (and slightly different syntax).

```
lsmestimate type "1 vs 2" 1 -1 0 0,
                 "3 vs 4" 0 0 1 -1,
                 "animal vs veg"  0.5 0.5 -0.5 -0.5/joint;
```

The `joint` option gives a joint test for whether the LSMeans are the same, which is the same as the overall test in the simple designs like the CRD. There are also multiple comparison adjustments available in `lsmestimate`.

What happens if you don't pre-plan? Ideally, comparisons are set up ahead of time based on specific research questions. If comparisons are selected after examining the data, most researchers construct tests that correspond to large differences in the means. These differences could be due to a real treatment effect, or they could be due to random error. Picking the largest differences to compare will inflate Type I error. If you do want to look at comparisons suggested by the data (post hoc comparisons), then you should replace the t test with a VERY conservative test called the **Scheffé** test. Scheffè works for pairwise comparisons or contrasts. We request it by adding the `adjust=scheffe` option.

To see how conservative Scheffè is, let's look at the comparison of Fats 1 vs 2 (and pretend that Fat 1 is a control, just for illustration.

Adjustment Type	p-value	Lower CL	Upper CL
Unadjusted	0.0365	-25.0974	-0.9026
Tukey	0.1462	-29.2320	3.2320
Dunnett	0.0908	-27.7326	1.7326
Scheffè	0.2044	-30.6813	4.6813

What do you notice?

Which one to use? It depends. Is it more important to control the comparison-wise error rate or experiment-wise error rate? That will depend on the situation. Keep in mind that the more conservative the adjustment, the lower the power. That is, the more likely you are to make a Type II error.

Example: A study is being planned to study the ability of a liberty ship artificial reef to attract and hold macrobenthic epifauna. One of the variables of interest is the density of oysters, and researchers are interested in comparing different locations on the artificial reef. There are 6 locations

- Floors of holds
- Sides of holds
- Starboard deck
- Starboard side
- Port side
- Port deck

and 12 observations were randomly sampled at each location.

- **Treatment Design:**
  - Factor:
  - Levels:
- **Experimental Design:**
- **Model:**

The researcher is particularly interested in some specific comparisons. We'll write the contrasts to address each one.

1. Floors versus sides of holds

2. Port versus starboard

3. Deck versus sides, except for holds

4. Port sides versus starboard sides

5. Port decks versus starboard decks

6. Port side versus port deck

7. Starboard side versus starboard deck

5.3.3 Model Adequacy

Everything we've done so far is based on the assumptions that the observations are adequately described by the model

If these assumptions are not valid, then the estimates of the treatment means and tests of significance from the ANOVA will be affected. We typically use **residuals** as a basis of our diagnostic tools.

The **residual** for observation j in treatment i is defined as:

Examining residuals should be an automatic part of the analysis of variance, and can be used to check the assumptions of common variance and normality of the error term. The assumptions can be checked using a visual inspection or formally through tests, and SAS makes it very easy to do so.

There's a lot of code here, but we'll examine it piece-by-piece.

```
proc glimmix data=donut plot=residualpanel;
  class type;
  model absorb=type;
  random _residual_/group=type;
  covtest homogeneity;
  output out=donutout pred=pred residual=resid;
run;
```

Here's what the options are doing:

- `plot=residualpanel` produces a set of residual plots
- `random _residual_/group=type` tells SAS you want to estimate a residual variance for each treatment group (i.e., get separate estimates of σ^2 from each treatment group)
- `covtest` produces a hypothesis test for comparing variances, and `homogeneity` says you want to test whether they are all equal
- `output` produces a new data set (called `donutout`) which contains the observed residuals (`resid`) and predicted values (`pred`)

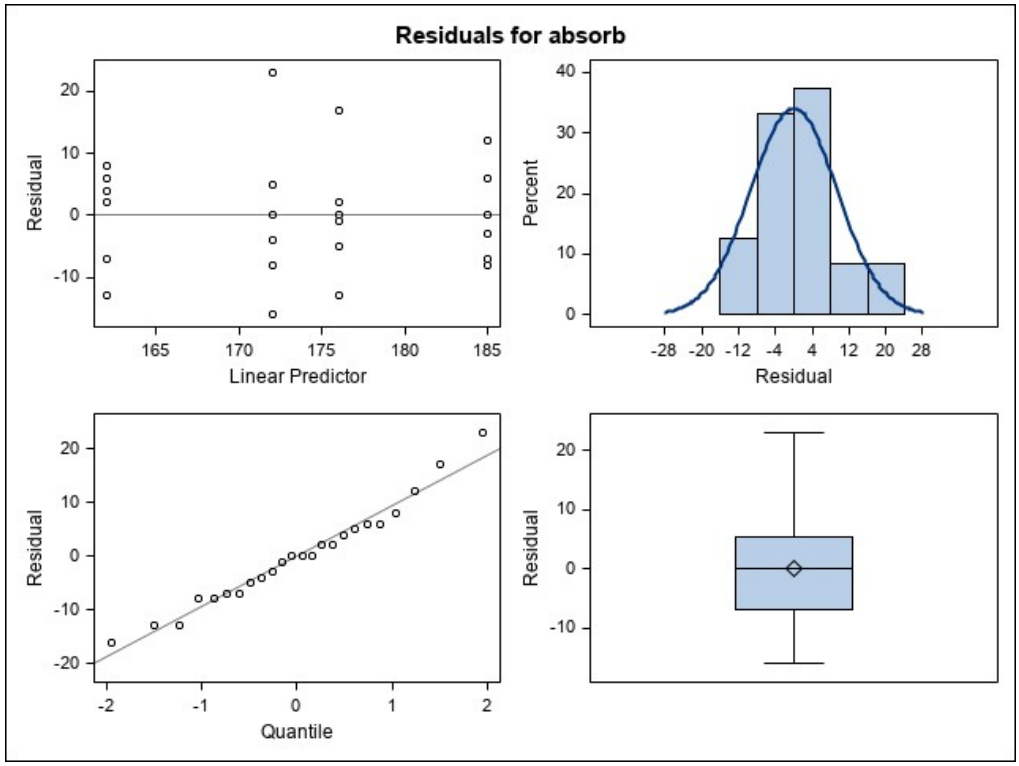


Figure 5.1: Residual panel for the donut data

The upper left hand plot shows

The other three plots all deal with the normality assumption.

We can also use `proc univariate` to check normality, using the `donoutout` data set we created above.

```
proc univariate data=donutout plot normal;
  var resid;
run;
```

Here's part of the output

```
                        Tests for Normality

Test                  --Statistic---    -----p Value------

Shapiro-Wilk          W     0.972165    Pr < W      0.7205
```

The Shapiro-Wilk test is the most commonly used test for normality. A highly significant p-value would indicate there may be a problem with non-normality.

What happens if we do see a large departure from normality?

The other assumption we can check with residuals is the constant variance assumption, also called the assumption of homogeneous variances. From the SAS output

Covariance Parameter Estimates

Cov Parm	Group	Estimate	Standard Error
Residual (VC)	type 1	178.00	112.58
Residual (VC)	type 2	60.4000	38.2003
Residual (VC)	type 3	97.6000	61.7277
Residual (VC)	type 4	67.6000	42.7540

Type III Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
type	3	20	8.39	0.0008

Tests of Covariance Parameters
Based on the Restricted Likelihood

Label	DF	-2 Res Log Like	ChiSq	Pr > ChiSq	Note
Homogeneity	3	156.21	1.90	0.5942	DF

DF: P-value based on a chi-square with DF degrees of freedom.

The covariance parameter estimates are the estimates of the variances for each of the four treatments, along with their standard errors. What do you notice?

The Tests of Covariance Parameters is testing the null hypothesis that the four variances are equal, versus the alternative that at least one is different.

In general, moderate departures from normality are of little concern, especially with the CRD. Nonconstant variance can be a bigger issue, but there are things we can do (like transformations) to stabilize the variance.

5.3.4 Power for the Completely Randomized Design

With multiple comparisons, we talked about Type I error and its probability. We defined Type I error as rejecting the null hypothesis when it is, in fact, true. If P(Type I error) = α, then P(no Type I error) = $1 - \alpha$.

There is another kind of error – Type II error. A Type II error occurs when H_0 is not rejected, but H_0 is actually false. Earlier this semester, we summarized these two types of errors in a table:

Earlier in STAT 801 we've said that we can "set" the probability of a Type I error. Anytime we say we'll reject H_0 if the p-value $< \alpha$, we're setting P(Type I error) = α. What about Type II error? We generally call the probability of a Type II error β, P(Type II error) = β. The problem is we can't "set" both β and α without some other complications.

We are typically interested in the **power** of a test:

Let's explore this via simulation. Consider a two-sample t-test. In Canvas, there is an R file called `simulation example.R`.

- In the program, we're generating $n_1 = 20$ normal random variables with $\mu_1 = 10$, $\sigma^2 = 25$ and $n_2 = 20$ normal random variables with $\mu_2 = 10$, $\sigma^2 = 25$.
- Carry out the t-test (code already included) and observe the p-value. Using $\alpha = 0.10$, what is your decision?
- Run the program 9 more times, so you have a total of 10 p-values. How many times out of 10 did you reject H_0?

  What does this estimate?

- Edit the program to generate data with $\mu_1 = 10$ and $\mu_2 = 12$ (still with $n_1 = n_2 = 20$ and $\sigma^2 = 25$). Run the program 10 times total. How many times did you reject H_0, using $\alpha = 0.10$?
- Edit the program to generate data with $\mu_1 = 10$ and $\mu_2 = 15$ (still with $n_1 = n_2 = 20$ and $\sigma^2 = 25$). Run the program 10 times total. How many times did you reject H_0, using $\alpha = 0.10$?
- Edit the program to generate data with $\mu_1 = 10$ and $\mu_2 = 20$ (still with $n_1 = n_2 = 20$ and $\sigma^2 = 25$). Run the program 10 times total. How many times did you reject H_0, using $\alpha = 0.10$?

What do you observe as μ_1 and μ_2 get further apart?

Now let's try the following:

- Edit the program to generate data with $\mu_1 = 10$ and $\mu_2 = 12$, but with $\sigma^2 = 1$ (still with $n_1 = n_2 = 20$). Run the program 10 times total. How many times did you reject H_0, using $\alpha = 0.10$?
- Edit the program to generate data with $\mu_1 = 10$ and $\mu_2 = 20$, but with $\sigma^2 = 625$ (still with $n_1 = n_2 = 20$). Run the program 10 times total. How many times did you reject H_0, using $\alpha = 0.10$?

What do observe as σ^2 gets larger or smaller?

Power is the probability of rejecting H_0 when it is really false. It is a function of several quantities:

-
-
-
-

Power analyses usually focus on calculating the sample size required to achieve a particular power. What do you think would happen if instead of using $n_1 = n_2 = 20$ we used $n_1 = n_2 = 10$?

What do you think would happen if instead of using $n_1 = n_2 = 20$ we used $n_1 = n_2 = 40$?

What do you think would happen if instead of using $n_1 = n_2 = 20$ we used $n_1 = 10$ and $n_2 = 30$?

In some simple situations, we can SAS procs to do power calculations. There are two: `PROC POWER` and `PROC GLMPOWER`. We'll use `PROC POWER`. This proc will do power calculations for two sample t tests and ANOVA.

For a two-sample t test, the basic code is:

```
proc power;
  twosamplemeans test=diff
  alpha=
  stddev=
  meandiff=
  npergroup=
  power=               ;
run;
```

We'll need to supply values for alpha, stddev, and meandiff. We can either supply a value for npergroup and use `power=.` or supply a value for power and use `npergroup=.`

We could also add the lines

- `plot x=power min=0.5 max=0.95;` (for `ntotal=.`)
- `x=n min= max= ;` (for `power=.`)

Let's try this, going back to our example with $\mu_1 = 10$, $\mu_2 = 12$, and $\sigma^2 = 25$.

We can also use `PROC POWER` for ANOVA and contrasts. This time, the basic code is:

```
proc power;
  onewayanova
  alpha=
  stddev=
  groupmeans=  |   |
  ntotal=
  power=
  contrast= (     );
run;
```

Let's go back to the donut data. In that example, the MSE was 100.90, so we'll use $\sigma = 10$ as a guess for future experiments. We observed sample means of $\overline{y}_{1\cdot} = 172$, $\overline{y}_{2\cdot} = 185$, $\overline{y}_{3\cdot} = 176$, and $\overline{y}_{4\cdot} = 162$. We can certainly use these as guesses for future experiments. We'll consider the contrast testing animal fats versus vegetable fats. We could also look at the overall test.

We can also add plot statements here.

But, we don't actually have to have guesses for the treatment means. We do have to have an idea of how large a difference we want to be able to detect. With our example data, we had a animal fat mean of 178.5 and a vegetable fat mean of 169. This is a difference of 9.5.

We could also the consider potential differences we might observe in pairwise differences.

5.4 Block Designs

Up to now, we've concentrated on analyzing data coming from various **treatment** designs. We've considered multiple flavors of one-way designs (unstructured, control vs others, regression, other structure). In all cases though, we've been using the same **experimental** design: the completely randomized design (CRD). Now, we change our focus to other experimental designs. The treatment designs will be those we've seen before, and we'll continue to analyze treatment effects using methods we've already discussed.

With a shift to experimental designs, we'll be considering

This also means

We're going to cover the simplest experimental design (besides the CRD): the **randomized complete block design** (RCBD) and leave more complicated block designs for STAT 802.

Consider an experiment in which we are interested in comparing six different lab activities for teaching the central limit theorem. Based on a power analysis, we believe four replications per treatment is sufficient, and so we need a total of 24 lab teams. We've got two options:

-
-

Which do you pick? Why? What are pros and cons of each?

Suppose you decide to use teams in different classes (or you don't have a choice). How will you assign treatments to teams?

Suppose we allocate treatments to teams completely at random (CRD), and by chance four out of six teams in one class are assigned to treatment 1. Would you be okay with this?

One of the main problems with the CRD is a possible 'conditional' bias. That is, treatment assignment is not balanced relative to any systematic variation/gradient. In this experiment, the gradient is

When this happens, treatment effect is confounded with gradient. Is any effect we observe really due to treatment, or is it due to the effect of the class? The other problem with the CRD is variance inflation. Suppose there is a gradient among the experimental units, with response increasing as you go up a gradient:

Even if all the same treatment is applied throughout, the variance among the experimental units (the residuals) will be composed of two quantities:

This means it will appear larger than it actually is. The most common solution to these problems of confounding and variance inflation is **blocking**.

The idea of blocking is:

Blocking allows us to reconcile two somewhat opposing aims of experimental design.

-
-

In summary, the general idea of blocking is to organize experimental units into groups that are as uniform as possible. We want to

Blocks usually represent naturally occurring differences not related to treatments. If we block 'correctly' then the design accounts for block variation, and allows us to pull it out and isolate the usual random error due to experimental units. If we block 'incorrectly' then we get a weaker experiment.

How Do We Block?

There are two basic steps in blocking an experiment:

1. Organize the experimental units into subsets (blocks) according to gradient
2. Restrict the randomization so that each treatment is assigned to one or fewer (zero) experimental units in each block.

5.4.1 The Randomized Complete Block Design

The simplest block design is the **randomized complete block design** (RCBD). In this design

In the RCBD, we carry out the two steps referenced on the previous page as

1. Divide the experimental units into blocks of homogeneous units.
2. Randomly assign treatments to units within blocks, using a separate randomization for each block. Every treatment will appear in every block.

Again, we carry out step 1 with the goal

The model for the RCBD helps point out some considerations for choosing blocks. The model (assuming a one-way treatment design) is:

The ANOVA table looks like

Example: An experiment was carried out to evaluate the effect of elevated CO_2 on rice grain yield. Four blocks of 2 rice paddies each (each block owned by a different farmer, who used different fertilizer regimes and management practices over the years) are available for the experiment. In each paddy there is a 12 m diameter circular plot. In one plot in each block there is a ring of tubing around the plot emitting CO_2 at a rate of 300 ppm above ambient level. In the other plot, no CO_2 is emitted. The grain yield is measured at 3 locations in each plot at the end of the season, and the response is the average of the 3 locations.

What is the experimental unit here?

What does the assumption of no block $\times$ treatment interaction mean in this example?

We can check it with an interaction plot. Here are the means for each plot

Block	Ambient CO_2	Elevated CO_2
1	6.21	6.41
2	6.25	6.42
3	6.10	6.26
4	6.14	6.30

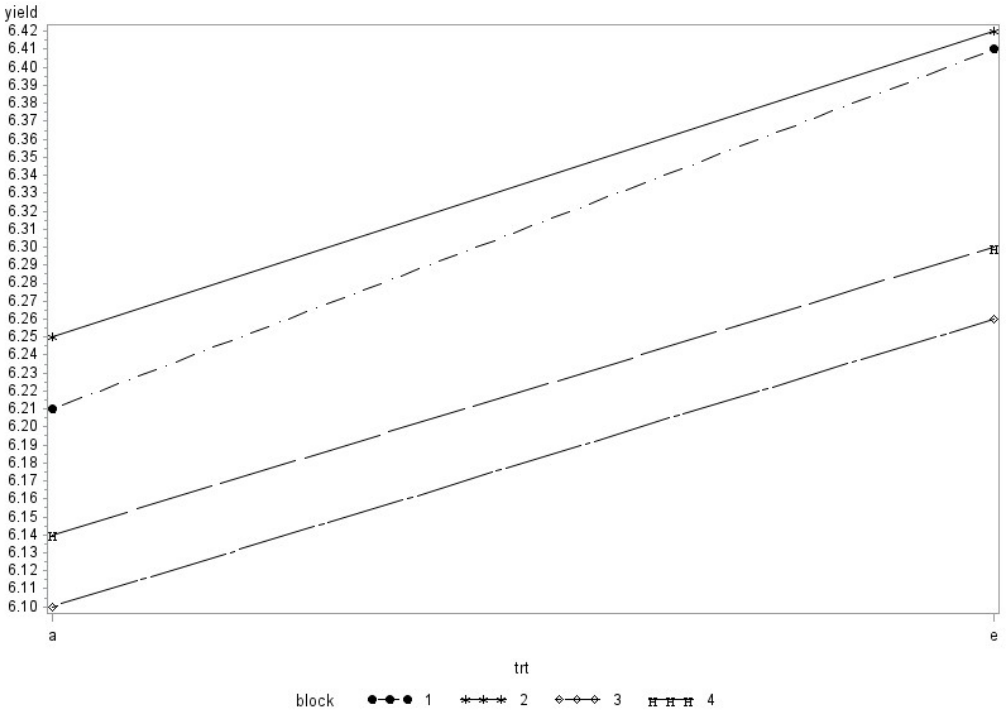


Figure 5.2: Interaction plot for treatment x block.

If you had been presented with this data earlier in the semester, how would you have analyzed it?

Block	Ambient CO_2	Elevated CO_2
1	6.21	6.41
2	6.25	6.42
3	6.10	6.26
4	6.14	6.30

5.4.2 Selecting Blocks

Remember that the RCBD is an **experimental** design, not a treatment design. It can be used with any treatment design. So, we might see RCBD layouts that look like:

Block 1	Block 2	Block 3
Control	Trt2	Trt4
Trt2	Control	Trt3
Trt3	Trt4	Trt2
Trt4	Trt3	Control

Block 1	Block 2	Block 3
20	60	80
60	40	20
80	80	40
40	20	60

Block 1	Block 2	Block 3
A1 & B1	A1 & B1	A2 & B2
A2 & B1	A2 & B1	A1 & B1
A2 & B2	A1 & B2	A1 & B2
A1 & B2	A2 & B2	A2 & B2

When you write a report, both the treatment design and the experimental design need to be described in the methods section.

Tips for Choosing Blocks:

- We want to maximize differences between blocks and minimize differences within blocks

- Block size should not be excessively large

- Keep in the mind the no block $\times$ treatment interaction assumption

Common Criteria for Blocking:

- gradients that occur in the field, in greenhouses, in growth chambers
- weight groups in animal experimentation, litters, cage positions in a room
- occasion (day, month, year)
- location (barn, different fields, different rooms, different states)
- subjects (each subject serves as their own control)

5.4.3 RCBD Model and Analysis

Let y_{ij} be

Earlier we stated the model

We do have another choice to make. We can consider the block effect to either be a **fixed effect** or a **random effect**.

Our choice will have implications in the standard errors of the cell means.

To estimate the difference between two treatment means $(\mu_{i\cdot} - \mu_{i'\cdot})$, we use $\overline{y}_{i\cdot} - \overline{y}_{i'\cdot}$. To figure out the variance (or estimate of the variance), let's look at what $\overline{y}_{i\cdot}$ is actually estimating:

Now let's explore the variance of this quantity.

Now let's consider $\overline{y}_{i\cdot} - \overline{y}_{i'\cdot}$:

and its variance

If we want to construct confidence intervals for treatment means or differences, they'll have the form

where the standard error is

For example, we use MSE as our estimate of σ^2, so confidence intervals for the difference between two means is

RCBDs in SAS

We can still use `PROC GLIMMIX` to fit the model if our experimental design is the RCBD. The basic program for **fixed blocks** is

```
proc glimmix data=dataset;
  class block trt;
  model y = block trt;
run;
```

Note:

The basic program for **random blocks** is

```
proc glimmix data=dataset;
  class block trt;
  model y = trt;
  random block;
run;
```

Note:

Example: This experiment is looking at the emergence rate of soybean seeds treated with four different chemical treatments and a control.

Treatment Number	Treatment Name
1	Control
2	Arasan
3	Spergon
4	Semesan
5	Fermate

Experimental Layout: The field is located on a slope, and blocks are formed based on elevation. There are five plots at each elevation, and five blocks.

Treatment Design:

100 seeds were planted in each plot, and the response is the number of plants that emerge out of the 100.

Model:

Analysis with Blocks Fixed:

If we assume blocks are fixed, we use the code

```
proc glimmix data=seeds;
  class block chem;
  model emerge=block chem;
run;
```

which gives

```
               Fit Statistics

Pearson Chi-Square / DF         5.41

     Type III Tests of Fixed Effects
               Num      Den
Effect          DF       DF    F Value    Pr > F

block            4       16       2.30    0.1032
chem             4       16       3.87    0.0219
```

The follow-up analyses don't change from what we've done so far. In this case, the treatment design is one-way treatment-versus-control, so comparing all treatments to the control is appropriate and we can use the Dunnett adjustment.

```
lsmeans chem/diff=control('Control') adjust=dunnett;

                    Chem Least Squares Means

                       Standard
Chem       Estimate       Error       DF    t Value    Pr > |t|

Arasan      93.8000      1.0402       16      90.18      <.0001
Control     89.2000      1.0402       16      85.75      <.0001
Fermate     94.2000      1.0402       16      90.56      <.0001
Semesan     93.4000      1.0402       16      89.79      <.0001
Spergon     91.8000      1.0402       16      88.25      <.0001
```

```
  Differences of Chem Least Squares Means
Adjustment for Multiple Comparisons: Dunnett

                                  Standard
Chem       _Chem      Estimate       Error       DF    t Value    Pr > |t|     Adj P

Arasan     Control      4.6000      1.4711       16       3.13      0.0065    0.0218
Fermate    Control      5.0000      1.4711       16       3.40      0.0037    0.0125
Semesan    Control      4.2000      1.4711       16       2.86      0.0115    0.0375
Spergon    Control      2.6000      1.4711       16       1.77      0.0962    0.2680
```

Analysis with Blocks Random:

If we assume blocks are random, we use the code

```
proc glimmix data=seeds;
  class block chem;
  model emerge=chem;
  random block;
  lsmeans chem/diff=control('Control') adjust=dunnett;
run;
```

which gives

```
          Fit Statistics
Gener. Chi-Square / DF          5.41

  Covariance Parameter Estimates
                        Standard
Cov Parm    Estimate       Error
block         1.4100      1.8032
Residual      5.4100      1.9127

       Type III Tests of Fixed Effects
              Num      Den
Effect         DF       DF    F Value    Pr > F
Chem            4       16       3.87    0.0219

                Chem Least Squares Means
                      Standard
Chem       Estimate       Error       DF    t Value    Pr > |t|
Arasan      93.8000      1.1679       16      80.31      <.0001
Control     89.2000      1.1679       16      76.38      <.0001
Fermate     94.2000      1.1679       16      80.66      <.0001
Semesan     93.4000      1.1679       16      79.97      <.0001
Spergon     91.8000      1.1679       16      78.60      <.0001
```

Differences of Chem Least Squares Means
Adjustment for Multiple Comparisons: Dunnett-Hsu

Chem	_Chem	Estimate	Standard Error	DF	t Value	Pr > |t|	Adj P
Arasan	Control	4.6000	1.4711	16	3.13	0.0065	0.0218
Fermate	Control	5.0000	1.4711	16	3.40	0.0037	0.0125
Semesan	Control	4.2000	1.4711	16	2.86	0.0115	0.0375
Spergon	Control	2.6000	1.4711	16	1.77	0.0962	0.2680

The results are the same whether we used fixed blocks or random blocks. This is because our data are **balanced**–we had the same number of observations in each block, and all treatments appear in all blocks. If our data had not been balanced, the results would be different.

5.4.4 Did Blocking Work?

When we treated blocks as fixed effects, we get a p-value associated with block but it is completely meaningless because there is no valid hypothesis test for evaluating the effect of block. However, we can check the **efficiency** of the block design relative to a competing design.

Suppose we have t treatments and rt experimental units available for our experiment. We have two possible experimental designs:

-
-

The only difference between these is whether we group the experimental units into blocks before randomly assigning the treatments. Efficiency gives us a way to compare the variance of two competing designs–we want to select the design that gives us the smaller variance of estimated treatment differences.

- CRD:

- RCBD

So the choice between these two designs comes down to a comparison of σ^2_{CRD} and σ^2_{RCBD}. We can compare variances using a ratio called the **relative efficiency**.

If RE > 1

Once we've conducted an RCBD experiment we can look and see whether we did the right thing when we used blocks.

Note there is a difference in the error degrees of freedom between the CRD and RCBD which can have an impact. We can adjust for this difference by calculating the **adjusted relative efficiency**:

The correction factor is always less than 1, and usually won't make much difference. It can make a difference if the number of treatments and reps is small.

Example: Rice paddies In this example, there were 4 blocks and two treatments. We'll fit the model both with blocks and without.

- RCBD:

- CRD:

Example: Seed Emergence In this example, there were five blocks and five treatments. Again, we'll fit the model both with blocks and without.

- RCBD:

- CRD: