Statistics Review Visualizations

Zachary Himmelberger

1/8/2021

```
library(knitr)
library(effsize)
library(psych)
library(tidyverse)

set.seed(161925)

# simulate data
df <- data.frame(
  condition = c(rep("control", 10), rep("experimental", 10)),
  statistical.knowledge = c(rbinom(10, 100, .75), rbinom(10, 100, .8))
)
```

In this review, we will be working with a single example. All data is simulated. Therefore, we cannot make conclusions about the content, only the methods.

The psychology department at Maryville College notices that many students struggle to independently analyze data for their Senior study. They decide to conduct a randomized controlled experiment to determine whether offering a statistics review would be helpful.

Twenty students are randomly selected to be in the study. Ten of the students are randomly assigned to participate in a statistics review. This is called the **experimental group** because they are exposed to a treatment (or manipulation). The other ten students do not do anything different. This is called the control group. At the end of their Senior study, all twenty students are tested on their statistical knowledge using a measure that ranges from 0 to 100 (i.e., a score on an exam).

The **research question** is, Does the statistics review increase the students' statistical knowledge? We can also frame this as a **hypothesis**: It is hypothesized that students who take the statistics review will score higher on a measure of statistical knowledge than students who do not participate in the statistics review.

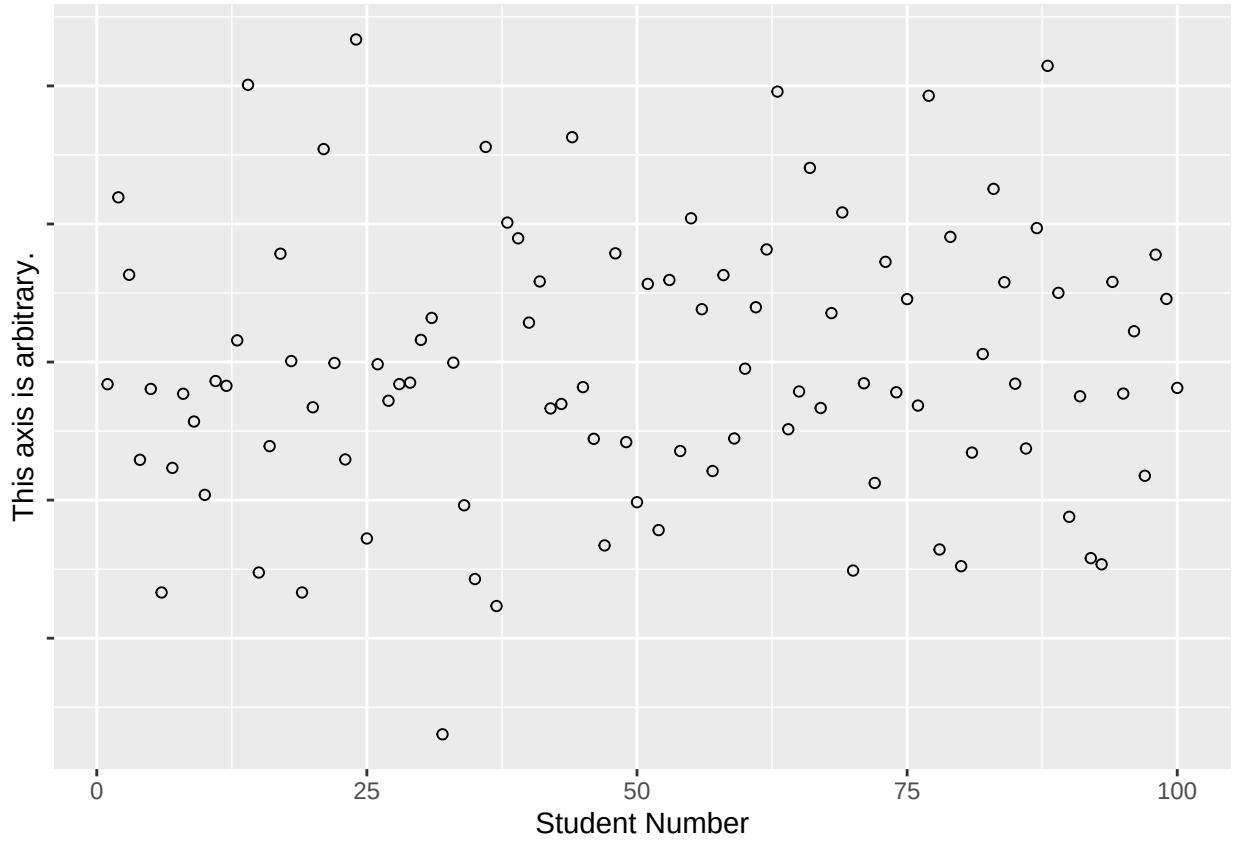
Population and Sample

A **population** is the collection of people (or things) about whom we wish to make conclusions. A **sample** is the subset of the population from whom we collected data. In our example, the population is all psychology Senior study students. The sample is the 20 students who participated in our study.

We can visualize the population as all possible students who could have participated in the study. The y-axis on this plot is arbitrary (but it helps us separate the points). The x-axis is just counting the students.

```
# Visualize the population.
population.df <- data.frame(
  student.number=1:100,
  value = rnorm(100)
)
```

```
ggplot(data=population.df, aes(x=student.number, y=value)) +
  geom_point(shape=1) +
  scale_y_continuous(name="This axis is arbitrary.") +
  scale_x_continuous(name="Student Number") +
  theme(axis.text.y=element_blank())
```

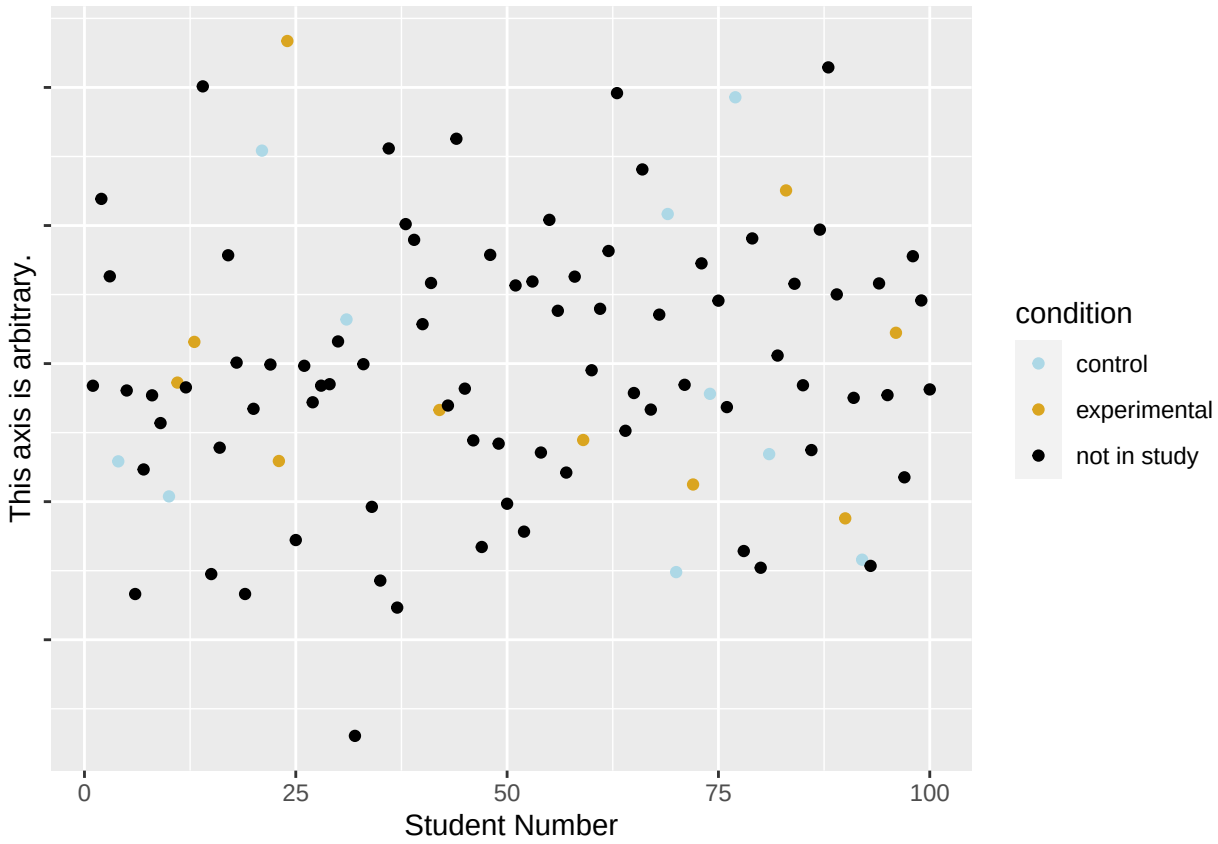


Now we can plot the same group of students, but select twenty at random to participate in our study. Of these students, ten will be in the experimental group and 10 will be in the control group.

```
# Randomly select 20 students and assign ten of them to the experimental group.
participants <- sample(1:100, size=20)
experimental.group <- sample(participants, size=10)

# Add a new column to the dataframe, then identify the twenty participants, and finally identify the te
population.df$condition <- 'not in study'
population.df$condition[population.df$student.number %in% participants] <- 'control'
population.df$condition[population.df$student.number %in% experimental.group] <- 'experimental'

# Visualize the population and sample.
ggplot(data=population.df, aes(x=student.number, y=value, color=condition)) +
  geom_point() +
  scale_color_manual(values=c("lightblue", "goldenrod", "black")) +
  scale_y_continuous(name="This axis is arbitrary.") +
  scale_x_continuous(name="Student Number") +
  theme(axis.text.y=element_blank())
```



Random Sampling and Genearlization

We said that the sample was chosen at random from the population. This is called **random sampling**, and occurs when every person in the population has an equal chance of being selected for the sample. Random samples are the gold standard because they allow us to **generalize** our results to the population. By this, we mean that we can use what we have learned from the sample to estimate how people that were not in our sample would respond. For our example, random sampling allows us to assume that the effect of participating in the statistics review will generalize to all psychology majors that are completing a Senior study.

Quantitive and Categorical Variables

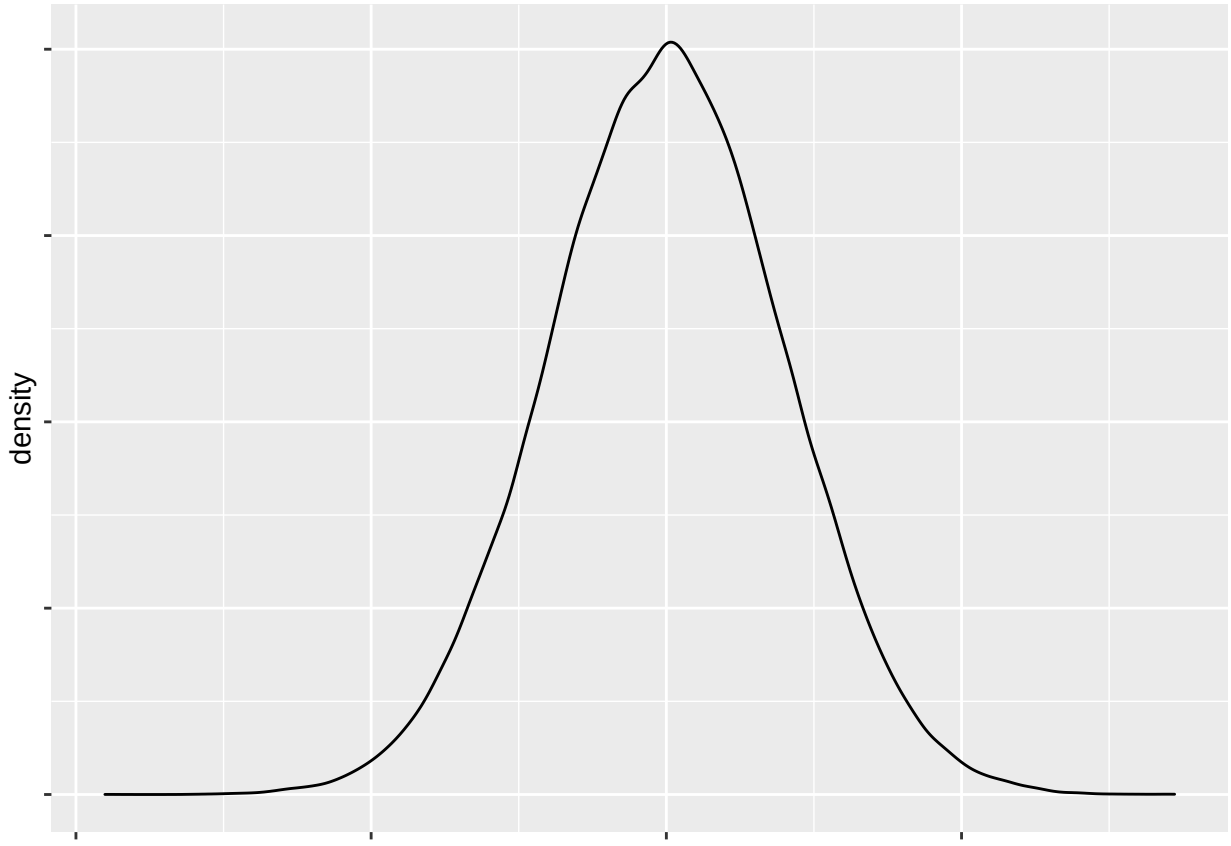
We want to distinguish between different types of variables. **Quantitative variables** are inherently numeric, which means that we can add and multiply them. Quantitative variables are either ratio or interval. The only difference is that a zero means an absence of that variable in ratio variables. For example, zero inches tall is the absence of height, so the variable is ratio; however, zero degrees Fahrenheit is not the absence of temperature (or heat), so the variable is interval. **Categorical variables** also come in two varieties. Ordinal variables have an inherent order to them, though there is not a consistent difference between points. For example, runners in a race can finish first, second, and third. These placings have a natural order, though the runner who finished the race 10 seconds faster than the person in second place, while the person in second place finished 30 seconds faster than the person in third place. Nominal variables are categories that have no inherent numerical meanings, such as favorite beer. As we will see, the type of variable dictates how we treat it in our statistical analysis.

In our example, we have two variables: review and statistics knowledge. Review is a categorical (or nominal) variable with two conditions: statistics review and no statistics review. Statistical knowledge is a quantitative (or ratio) variable.

Distributions, the Mean, and Variability

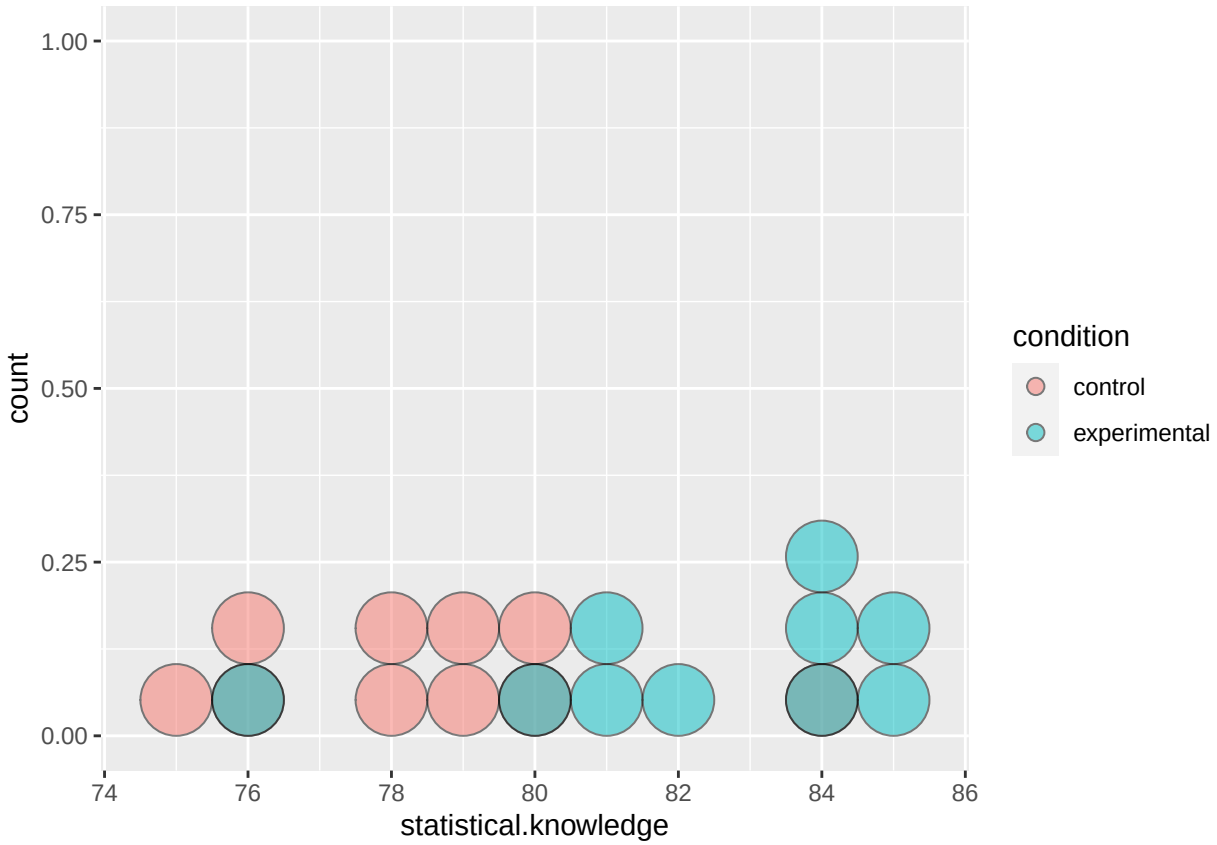
A **distribution** refers to the shape that the data take when plotted. In this class, we will be using a set of statistics called linear models. These statistical models rely on the normal distribution (shown below).

```
ggplot(data=NULL, aes(x=rnorm(100000))) +
  geom_density() +
  theme(axis.text.x=element_blank(),
        axis.text.y=element_blank(),
        axis.title.x=element_blank())
```



It is helpful to know the specific distribution that the data follows. If we know that the data is normally distributed, then all we need to know is the mean and variance to know everything about the distribution. I will say more on that momentarily. First, let's plot the observed data from our experiment.

```
# graph distribution
ggplot(data=df, aes(x=statistical.knowledge, fill=condition)) +
geom_dotplot(binwidth=1, position='identity', alpha=.5)
```



We can see that the students in the experimental group tend to do better on the measure of statistical knowledge, though there is also some overlap. Before we quantify the pattern, it will be helpful to explore a few more concepts.

The **mean**, or arithmetic average, is one measure for the center of a distribution. If the data is distributed in a normal distribution, then the mean is the exact center of the distribution. We will look at the formula below.

Most of the scores will differ from the mean. The **standard deviation** refers to the average (mean) difference between each score and the mean. Again, we will look at the formula below.

There are several properties of the normal distribution that are very important. All of these properties are true of perfect normal distributions. Although data doesn't follow these exactly, they provide excellent approximations.
1. The center of the distribution is the mean. Half of all scores will fall below the mean and half above. 2. We can describe the probability that a score will be at a certain point of a normal distribution, if we know the mean and standard deviation. For example, about 68% of scores will fall within one standard deviation above and below the mean and about 95% of scores will fall within two standard deviations of the mean. 3. The normal distribution is symmetrical.

We will see how we can use these facts to conduct statistical tests.

Calculating the Mean and Standard Deviation

A bit of mathematics will help us better understand the mean and standard deviation. Understanding the formulas to calculate these values will provide insight into their interpretation. It is also worth noting that we can only calculate the mean and standard deviation for quantitative variables. If our variable is categorical, we should instead look at proportions.

The mean is calculated by summing all of the scores and dividing by the number of scores. As a formula, this can be written as

$$\bar{x} = \frac{\Sigma x_i}{N}$$

where i refers to the ith individual. It is now helpful to define a **deviation score** as a difference between a person's score (notated with x_i) and the mean ($\bar{x}$). Thus, the formula for a deviation score is $x_i - \bar{x}$, which tells us how much a given score differs from the mean. Note that scores above the mean will give a positive deviation score and scores below the mean will give a negative deviation score.

We defined standard deviation to be the mean deviation score. To calculate the mean, we start by adding up all of the deviation scores. However, we will run into an issue: The sum of all deviations scores will be zero, for any distribution. In fact, the mean is a special number because half of the deviation is above it and half is below it.

To get around this problem, we will square all of the deviation scores (which makes them all positive, so they no longer sum to zero), calculate the mean, then take the square root (which brings the number back to the original unit). Let's see an example.

x	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
8	8 - 10 = -2	$-2^2 = 4$
12	2	4
6	-4	16
10	0	0
14	4	16
$\bar{x} = 10$		$SS = \Sigma(x_i - \bar{x})^2 = 40$ $s^2 = \frac{SS}{N} = \frac{40}{5} = 8$ $s = \sqrt{8} \approx 2.83$

Figure 1: Calculating standard deviation.

Exploratory Data Analysis

Now that we better understand the basics, we can start to conduct our data analysis. A good place to start is with exploratory analyses. Our focus here will be on descriptive statistics and graphs.

```
describe(df$statistical.knowledge)
```

```
##    vars  n  mean   sd median trimmed  mad min max range  skew kurtosis   se
## X1    1 20 80.35 3.25     80   80.38 4.45  75  85    10 -0.04    -1.38 0.73
```

We can see that the mean score on the statistical knowledge test was 80.35 with a standard deviation of 4.7. We can interpret this as saying that, on average, each point differs from the mean by about 5 points. Because we want to compare two groups, it will be helpful to see the mean and standard deviation separately for each group.

```
describeBy(df$statistical.knowledge, group=df$condition)
```

```
##
##  Descriptive statistics by group
## group: control
##    vars  n mean   sd median trimmed  mad min max range skew kurtosis   se
## X1    1 10 78.5 2.59   78.5   78.25 2.22  75  84     9 0.57    -0.45 0.82
## ------------------------------------------------------------
## group: experimental
```

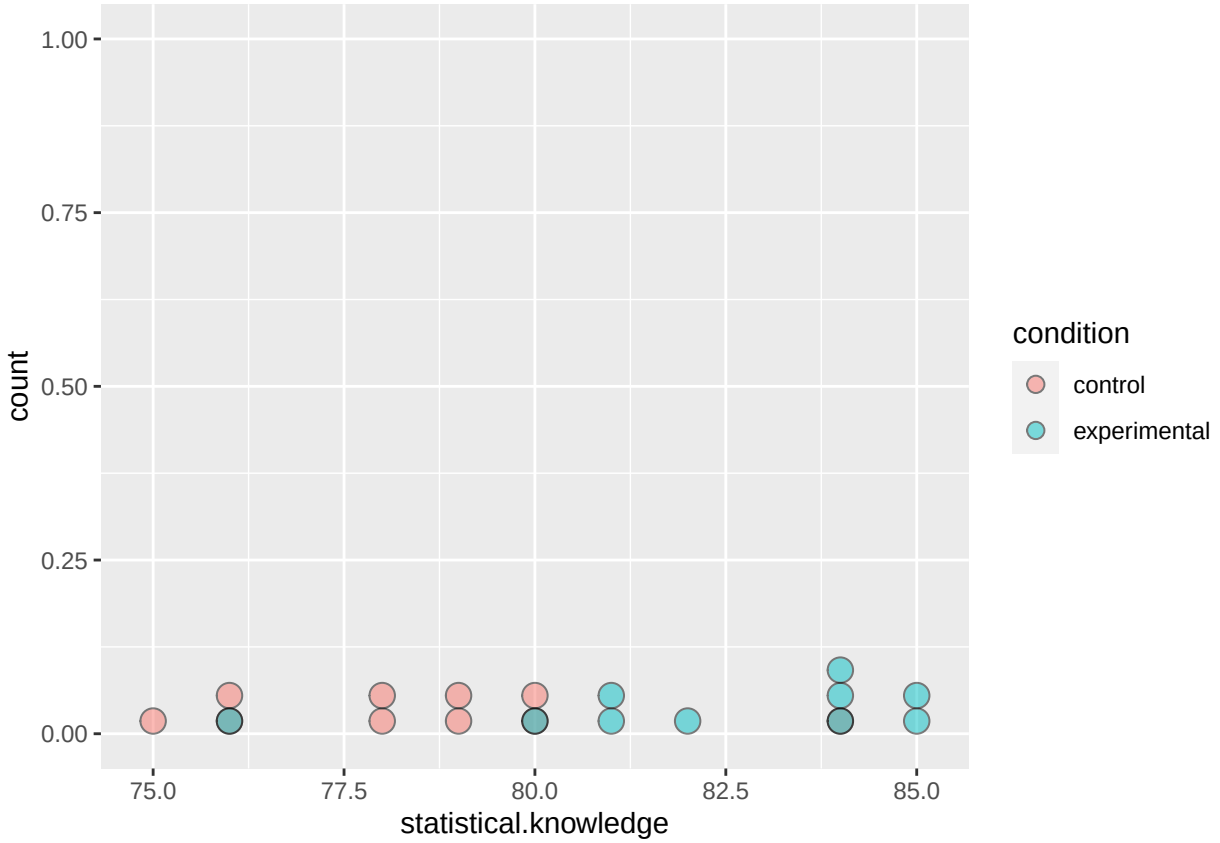
```
##    vars  n mean   sd median trimmed  mad min max range  skew kurtosis   se
## X1    1 10 82.2 2.82     83   82.62 2.97  76  85     9 -0.85    -0.38 0.89
```

We can see that the experimental condition (who received the stats review) did about five points better on the statistical knowledge test. There was also less deviation in their scores, as indicated by a smaller standard deviation.

Let's graph the results using several different methods. Each plot will show similar information, but help emphasize different aspects of the data. We won't worry too much about making the plots "publication quality" as we are just exploring relationships. We can add customization for the plots that we want to include in our final report.

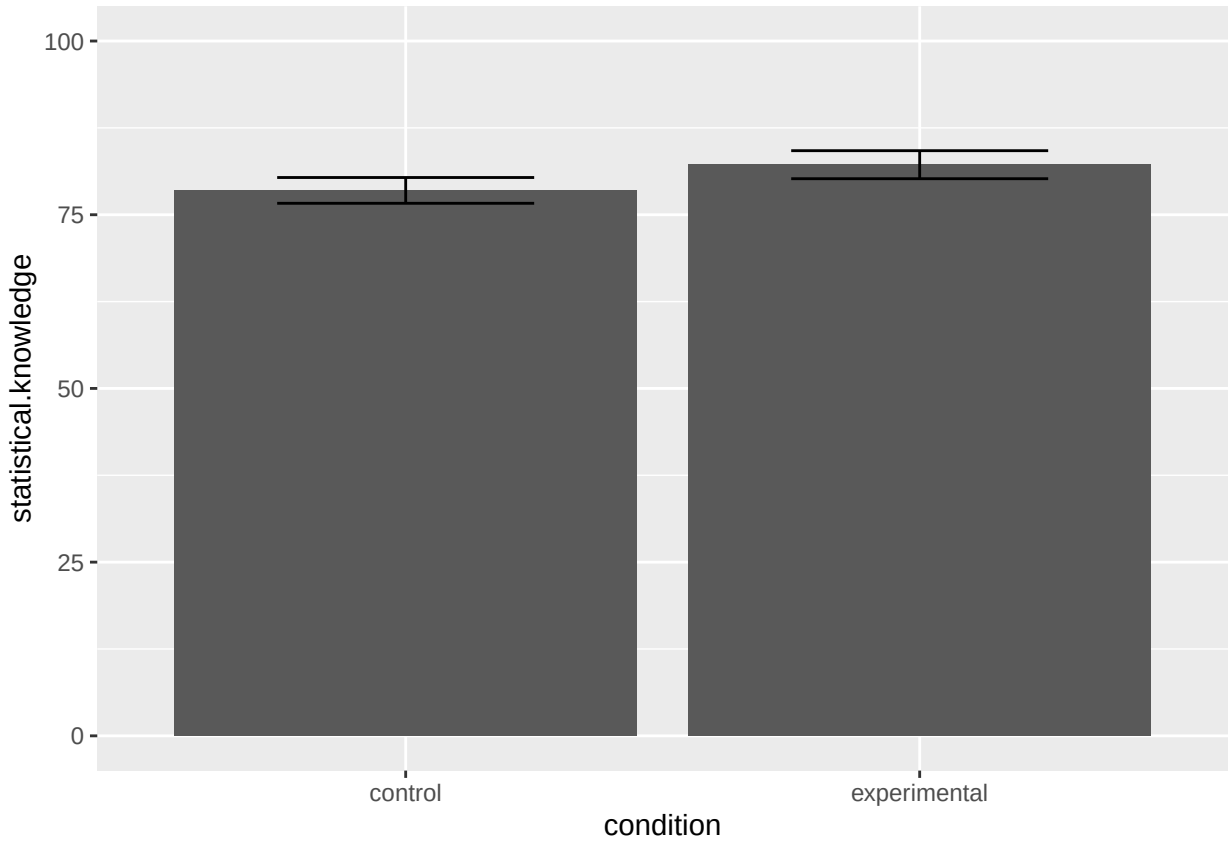
```
ggplot(data=df, aes(x=statistical.knowledge, fill=condition)) +
  geom_dotplot(alpha=.5)
```

```
## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`.
```



This dot plot helps show the full distribution of scores. We can immediately see that there is very little overlap in the two distributions.

```
ggplot(data=df, aes(x=condition, y=statistical.knowledge)) +
  stat_summary(fun='mean', geom='bar') +
  stat_summary(fun.data='mean_cl_normal', geom='errorbar', width=.5) +
  scale_y_continuous(limits=c(0,100))
```



This bar plot allows us to compare the mean performance. The error bars represent the 95% confidence limit (more on that later) for each group. Overall, we see a very small difference in means.

Statistical Model and Hypothesis Testing

During this class, you will learn to use different statistical models. I don't want you to worry too much about the details of this model. Instead, focus on the "bigger picture" ideas, such as inference and hypothesis testing.

One way to compare two means is to use a *t*-test. This test makes certain assumptions, listed below. 1. The errors have a mean of zero and are normally distributed. 2. The groups have the same population variance, call homogeneity of variance. 3. The errors are all independent.

For now, be aware that there are specific assumptions "hidden" in the statistical model. We will be considering each of these assumptions in more detail later in the course.

It will be helpful to stop here and think about our objective. Our research question was, Does the statistics review increase the students' statistical knowledge? However, we have two sets of students to consider: (1) the population of all psychology majors at Maryville College and (2) our sample from this population. **Our research question is asking about the population. Thus, we do not care if the sample means are different. We only care about the population means.**

The process of estimating a population value using samples is known as statistical inference. We are using information about the sample to *infer* information about the population. We distinguish between inferential statistics (like a *t*-statistic) from descriptive statistics (like a sample mean) for this very reason.

The basic process for making an inference about a population is as follows: 1. Define a null hypothesis and alternative hypothesis. 2. Calculate a test statistic. 3. Determine the probability of getting that test statistic if the null hypothesis was true. 4. Reject or fail to reject the null hypothesis.

Although a lot can be said about each of these steps, I am assuming you have some familiarity. Therefore, we will just go through the steps using our example.

Step 1. Define a null hypothesis and alternatie hypothesis.

The null and alternative hypotheses are determined by the research question or hypothesis. They are also about the population, not the sample (hence the use of Greek letters). In our case, we want to know whether the stats review helps students. Therefore, our null hypothesis is, On average, students in the experimental condition will perform no differently or worse than students in the control condition. Our alternative hypothesis is, On average, students in the experimental condition will perform better than the students in the control condition. We can also write this in symbols, where μ_E is the population mean in the experimental condition, μ_C is the population mean in the control condition, H_0 is the null hypothesis, and H_a is the alternative hypothesis.

$$H_0 : \mu_E <= \mu_C H_a : \mu_E > \mu_C$$

We call this a one-sided hypothesis because we only care if the experimental condition does better than the control condition. Most often, we will have two-sided hypothesis, which corresponds to $H_0 : \mu_E = \mu_C$.

Step 2. Calculate a test statistic.

As stated above, we will be conducting an independent samples t-test. This is easily done in R.

```
# The alternative = less is there because the t-stat is negative; we would flip this if the group order
t.test(statistical.knowledge ~ condition, data=df, var.equal=TRUE, alternative="less")
```

```
##
##  Two Sample t-test
##
## data:  statistical.knowledge by condition
## t = -3.054, df = 18, p-value = 0.003416
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -1.599149
## sample estimates:
##      mean in group control mean in group experimental
##                       78.5                       82.2
```

The results indicate that $t = -3.05$. It is important to note the arguments used in this t-test. Specifically, we assumed homogeneity of variance and a one-sided hypothesis test. It is also worth noting that the t-statistic is symmetrical. Specifically,

$$t = \frac{x_1 - x_2}{SE},$$

where SE stands for the standard error. Thus, if we had the experimental group as x_1 our statistic would be positive. Instead, R defaulted to the control group as being x_1, so our test statistic was negative.

3. Determine the probability of getting that test statistic if the null hypothesis was true.

This step requires the most background knowledge. I will simply remind you that the p-value is defined as the probability that we would observe the data (or something more extreme) given the null hypothesis is true. Thus, a low p-value indicates that it is unlikely that we would have observed this data if the null hypothesis were true.

In our example, the p-value is about .003.

4. Reject or fail to reject the null hypothesis.

In practice, when the p-value is less than .05, we reject the null hypothesis. Of course, there is nothing special about .05 and we can instead use .01 (or any other number). However, in psychology, we almost always use .05 as the cutoff (called alpha). If the p-value is greater than .05, we fail to reject the null hypothesis. This is *not* the same thing as accepting the null hypothesis is true.

Our p-value was .003. Thus, we can reject the null hypothesis. In words, this means that we have reason to believe that the population mean for the experimental condition is higher than the population mean for the control condition. If we used a null hypothesis of $H_0 : \mu_E = \mu_C$, then our p-value would have been twice as large, $p = .0068$, and our decision to reject the null hypothesis would stand.

I want to emphasize a key point. The sample means are different (82.2 v. 78.5). However, we don't care! All we care about is whether the population means are different. **An inferential statistical test is needed in order to determine whether the population means are different, thereby answering our research question.**

It is also worth noting that we are dealing with probabilities. The p-value can never equal 0. Thus, there is always some chance that you reject the null hypothesis, when the null is actually true. This is called a Type I error. Alternatively, we may fail to reject the null hypothesis (claiming that the population means are equal) when, in fact, the population means are actually different. This latter case is called a Type II error.

Effect Size

When we reject the null hypothesis, we can feel confident that there is an association between the variables (in the population). In our example, we feel confident that the statistics review had an effect on statistical knowledge. However, statistical significance does not indicate anything about whether the association is practically meaningful. An **effect size** describes the strength of the association. The most popular measure of effect size for the difference between two means is Cohen's D, which tells you the difference between the means in standard deviation units. Let's see an example.

```
effsize::cohen.d(df$statistical.knowledge ~ df$condition)
```

```
##
## Cohen's d
##
## d estimate: -1.365798 (large)
## 95 percent confidence interval:
##      lower      upper
## -2.4091660 -0.3224306
```

We can see that our effect size estimate is -1.37. We interpret this as saying that the means differ by 1.37 standard deviations (much like a t-test, we can ignore the sign, as it just depends on which variable is listed first in R). This is considered a large difference.

However, this still does not tell us about the meaningfulness of the difference. In terms of the original units of the exam, we would say that the conditions differ by about 4 points on the exam. Is this difference worth the investment of time and energy in holding a review session? That question cannot be answered using statistics.

This leads me to my final point. Statistics are tools to help us make decisions. They allow us to quantify our uncertainty. However, a statistic is just an equation. *You* are the expert. *You* have to think critically about how to appropriately apply a statistic. *You* must interpret the results. At the end of the day, statistics and statistical programming are tools that are useless without thoughtful application.