Chapter 4

Inference with Numeric Data

This ActiveStats document contains a set of activities for Introduction to Statistics, MA 207 at Carroll College. This is a non-calculus based statistics class which serves many majors on campus. This document is intended for the classroom teacher to support students in active engagement with statistics on a daily basis. This document is not designed to be given to students as is. Rather, it is a teacher resource.

The activity set is designed to work alongside the OpenIntro *Introductory Statistics with Randomization and Simulation* textbook by Diez, Barr, and Cetinkaya-Rundel. The chapters in ActiveStats are numbered to align with OpenIntro, though the subsections may differ. OpenIntro is an open source curriculum with accompanying data sets. OpenIntro is the textbook resource to direct students to for out-of-class reading assignments and review. We also use the Cartoon Guide to Statistics as a supplement for assigned reading.

Data sets for ActiveStats can be found at mathquest.carroll.edu/activestats/data/ or on the class Moodle page.

4.1 Shifting to numerical data

Reading Assignment 4.1.

- Read about the sampling distribution of the mean and the t-distribution in the Cartoon Guide to Statistics, pages 104 to 109 and pages 131-136

▲

Preview Activity 4.2.

- True or False: The t distribution has wider tails for larger samples.
- True or False: As the sample size increases, the t distribution becomes more and more normal.
- Presume that you are given the size of a sample, n, as well as the mean for the sample. If you wanted to make up a data set that had that particular mean, how many of the numbers could simply be random choices?

  (a) n, (b) n-1, (c) n+1, (d) all of them, (e) none of them

- When building a sampling distribution for either proportions or for means, the standard error for the sampling distribution is always proportional to:

  (a) $\frac{1}{n}$, (b) $\frac{1}{n^2}$, (c) n, (d) n^2, (e) $\frac{1}{\sqrt{n}}$, (f) $\sqrt{n}$

- Based on your answer to the previous question, if you want to make the sampling distribution narrower, you should

  (a) take a larger sample, (b) take a smaller sample, (c) take more samples, (d) take fewer samples.

Return to TinkerPlots or use an online resource like StatsKey to engage in repeated sampling to create sampling distributions for quantitative data.

Activity 4.3. 100 Calorie Snack Packs

Teacher Note: Whole class, teacher-led activity, to introduce concept and techniques.

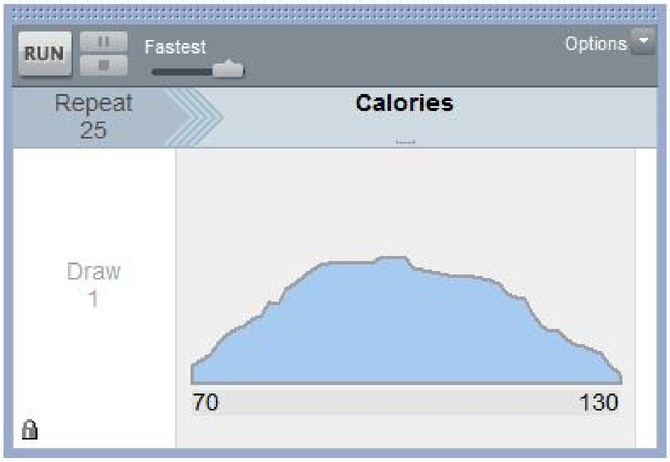
In recent years, the 100 calorie snack pack has become a popular phenomenon. The packages claim that the snack contains 100 calories. However, we know that we should expect some level of variability in the actual calories in each pack. If a pack contained 105 calories or 92 calories, that would not be terribly unexpected. However, if the **average** calorie count was significantly off from 100, that would be a false advertising problem.

A consumer advocacy group has decided to test the 100 calorie pack claim for a local cookie manufacturor. They suspect that there really are more than 100 calories per pack.

$$\text{Null Hypothesis: } H_0 : \mu_{calories} = 100$$
$$\text{Alternate Hypothesis: } H_A : \mu_{calories} > 100$$

- How might the consumer advocacy group test the advertised claim that the average number of calories is 100? Design a reasonable study.

- If the cookie company is telling the truth, the population of cookie packs might look something like this:



Download the 100CalorieSnackPackSimulator TinkerPlots file. Collect a random sample of 25 snack packs and find the average $\bar{x}_{calorie}$.

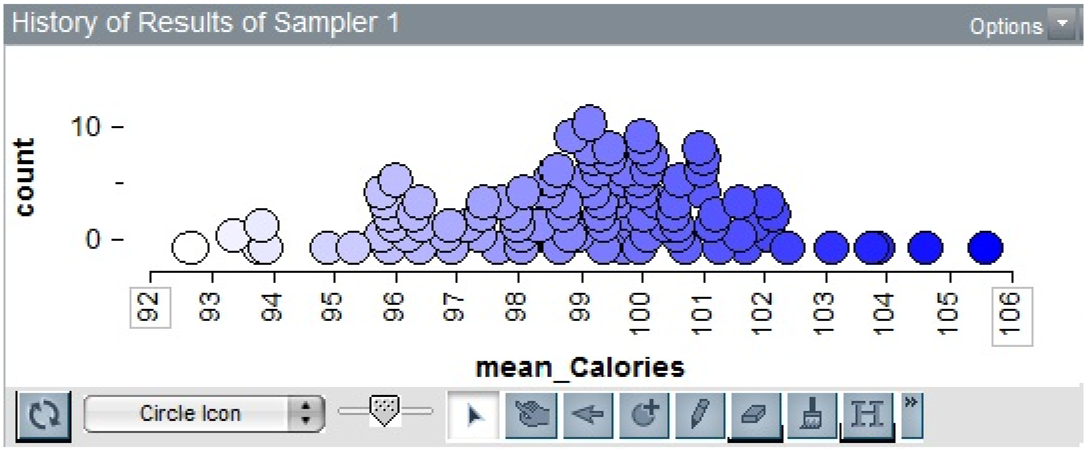
Teacher Note: We generally don't have access to population data like this. One reason is that we would have to measure every cookie pack (likely thousands or millions). Even more problematic, more cookies are being manufactured each day. Yikes!

If we actually have data about the whole population, we don't need to use hypothesis testing to verify a mean. We can just analyze the whole population and find the real mean. We are using this as an example to help make sense of the mechanics underlying sampling distributions for means. This population distribution serves the same role as the spinner when working with sampling distributions for proportions.

- Share your sample mean $\bar{x}_{calorie}$ with the class. Did everyone get the same sample mean? If so, why? If not, how much variability do you see between sample means?
- The consumer advocacy group took a random sample of 25 snack packs and found that the average calorie count was $\bar{x}_{calorie} = 105$ calories.
- How far is this sample mean from the hypothesized 100 calories? (In addition to a numeric answer, indicate whether you think that's a large difference or a small difference.)
- If we want to conduct a hypothesis test, what information do we need to know?
  - What is the standard error for the sampling distribution for sample means?
  - How far is the sample mean from the expected mean, when measured in standardized units? *test statistic* $= \dfrac{\bar{x} - \mu}{SE}$
  - How rare is it to find data as extreme as the sample data? (i.e. What's the p-value?)

▲

If 100 students each collected a sample of size 25, we could plot all of these results to see a sampling distribution like the one below. What do you notice about the shape of the distribution?



It's bell shaped!
Three big ideas:

- Point estimates from a sample are useful for estimating population parameters.
- Point estimates are not exact. We expect them to vary between samples.
- We can quantify the variability of point estimates. The Central Limit Theorem describes how.

Definition 4.4. Central Limit Theorem For a population with mean μ and population standard deviation σ, the sampling distribution of the mean approaches a normal distribution. Moreover, we also know what the mean and standard error of this sampling distribution will be.

$$\text{Mean of the sampling distribution } = \mu$$
$$\text{Standard error of the sampling distribution } = \text{SE} = \frac{\sigma}{\sqrt{n}}$$

Fine print - Check the following conditions before assuming that the sampling distribution is nearly normal

- The sample data must be independent.
- The underlying population distribution is not strongly skewed.
- The sample size is large enough. Often $n \geq 30$ is the guideline, but sometimes you can get away with lower, if your underlying population is nicely behaved.

Activity 4.5. Practicing with the Central Limit Theorem

Teacher Note: Good opportunity for Nearpod open response questions

An IQ test is designed to have a mean of $\mu = 100$ and standard deviation of $\sigma = 15$. Use the norm.dist() and norm.inv() commands in Excel to answer the following.

1. How rare is it for a randomly selected person to have a score of 95 or lower?
2. How rare is it for a randomly selected group of 30 people to have an average score of 95 or lower?
3. How rare is it for a randomly selected group of 100 people to have an average score of 95 or lower?
4. How rare is it for a randomly selected person to have a score of 110 or higher?
5. How rare is it for a randomly selected group of 30 people to have an average score of 110 or higher?
6. How rare is it for a randomly selected group of 100 people to have an average score of 110 or higher?

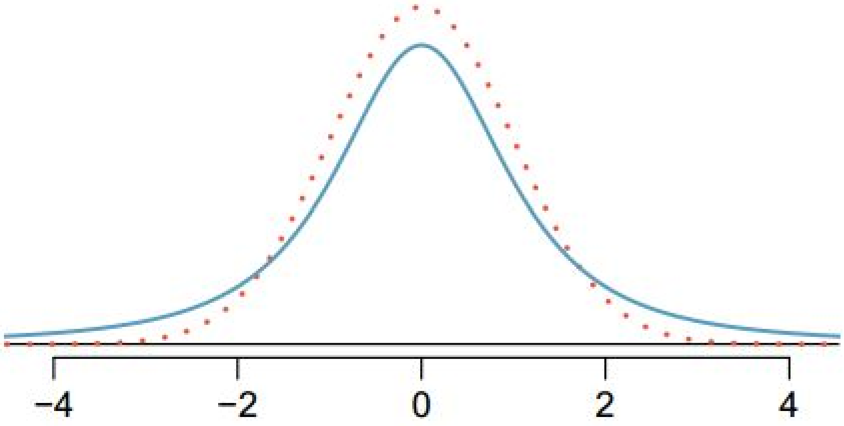
7. Create the interval that contains the middle 95% of individual test takers.
8. Create the interval that contains the middle 95% of means for randomly selected groups of 30.
9. Create the interval that contains the middle 95% of means for randomly selected groups of 100.
10. In a small town, a random selection of 100 citizens are given the IQ test. Their average score is 108. Is that unusual?

▲

In the previous activity, we were told what the population standard deviation σ was, along with the population mean μ. In most real world situations we do not know this information. In fact, it is often unreasonable or impossible to find the mean and standard deviation of a full population. For example, if we want to know the mean and standard deviation for weights of all newborn babies in Montana, we would have the challenge of gathering all of the data from all prior babies born. And even if we managed to do that, new babies would be added to our data set every day. But don't worry. We have statistical tools that allow us to work with a randomly selected sample from our population and make generalizations to the whole population.

Definition 4.6. Central Limit Theorem - Modified

The Central Limit Theorem requires that we know the population mean μ and population standard deviation σ in order to claim that the sampling distribution will be a normal distributoin with mean μ and SE $= \dfrac{\sigma}{\sqrt{n}}$. If we do not have those pieces of information, then the sampling distribution takes on slightly different shape. The sampling distribution is a t-distribution with mean μ and $SE \approx \dfrac{s}{\sqrt{n}}$



The blue curve is a t-distribution and the red curve is the standard normal distribution. Image from OpenIntro Ch 4.

While the t-distribution may look like a normal distribution, it is actually slightly different in shape. Although it is still shaped like a bell, the t-distribution is lower in the center and has slightly thicker tails.

The t-distribution is a family of mound shaped distributions that depends on the degrees of freedom of the scenario (degrees of freedom is related to sample size). As the sample size increases, the t-distribution approaches a standard normal distribution.

In order to choose which member of the t-distribution family we need, we introduce a new vocab term: Degrees of freedom (df)

- For a sample of size n for one mean problems, the degrees of freedom is df=n-1. There are other formulas for degrees of freedom for other scenarios.
- The degrees of freedom determine the shape of the t-distribution.
- The larger the degrees of freedom, the more closely the distribution approximates the normal model.
- Some older textbooks say that when $n > 30$, the t-distribution can be replaced with the normal distribution. This choice is related to limitations on calculations prior to computers. We will stick with the t-distribution when σ is unknown because it is more accurate than the normal distribution.

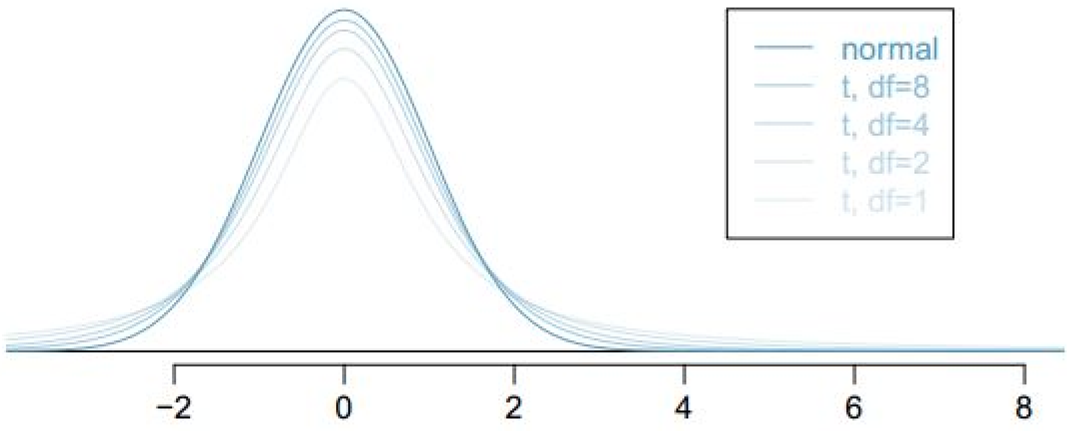


Image from OpenIntro Ch 4.

Working with the t-distribution is very similar to working with the normal distribution. The Excel commands t.dist(test stat, df,1) and t.inv(probability, df) serve the role parallel to norm.s.dist(test stat, 1) and norm.s.inv(probability).

Activity 4.7. Practicing with the t-distribution

In order to find the proportion of data to the left of the test statistic x, use the command

=T.DIST(test statistic x, DF, 1).

DF refers to degrees of freedom and the final 1 refers to cumulative. The test statistic is computed $\frac{\bar{x}-\mu}{SE}$, the number of standard errors the sample data is from the hypothesized mean.

The command T.INV(probability, df) is the inverse of T.DIST. A probability between 0 and 1 is entered as the first argument and the output of the funciton is the t-statistic which corresponds with that probability to the left.

Example 4.8. Excel commands

In order to find the area to the left of -2.50, with df = 12, use Excel command

=T.DIST(-2.5,12,true)

In order to find the top 5% of data with df = 12, use the Excel command

=T.INV(0.95,12)

0.95 is used because that is the proportion of data below our desired cutoff. Remember, always measure from the left side.

1. Find the portion of the area to the left of -1.30, with df = 12.
2. Find the portion of the area to the left of -1.30, with df = 24.
3. Find the portion of the area to the right of 2.30, with df = 10.
4. Find the cutoffs for the middle 80% of data, with df = 15.
5. Find the cutoffs for the middle 95% of data, with df = 20.

▲

4.2 One sample means - Hypothesis Testing

Reading Assignment 4.9. Read section 4.1 in the OpenIntro textbook (pages 163-170) ▲

Preview Activity 4.10. Questions related to the reading

1. Will a larger standard deviation, s, lead to a larger standard error (SE)? Choose all that are correct.

   (a) No, if the standard deviation is larger, then the standard error is smaller because they are inversely related.

   (b) Yes, if the standard deviation is larger this means there is more volatility in the data and we will be less certain of where the true mean is.

   (c) It depends. Sometimes a larger standard deviation gives a larger standard error, and sometimes it gives a smaller one.

   (d) Standard error depends only on sample size, so it doesn't matter at all what the standard deviation is.

2. For the Cherry Blossom Run, the standard deviation was given as 15.78 for a sample of size 100. If the study had used a sample of size 500, what would the standard error (SE) be, according to the formula provided?

3. When you use the t-distribution, you need to compute the degrees of freedom for the scenario you are using. What degrees of freedom are appropriate for the following scenario:

   You have conducted a study on 24 wild grizzly bears, investigating the average running speed of grizzly bears. You have computed your sample average and now you are building a 95% confidence interval for the average running speed. What degree of freedom corresponds with this scenario?

Activity 4.11. For this activity and several other activities going forward, we will explore a random sample from census data for Montana, North Dakota, Oregon, Utah, and Washington. This abridged sample data is taken from the 2013 data set available at www2.census.gov and was found at www.kaggle.com/datasets. In order to work with this data set, you may find Pivot Tables to be a helpful tool. With Pivot Tables, you can find counts, as well as averages and standard deviations. These can be sorted by states, gender, marital status, and other attributes.

Census data dictionary - abridged

Variable Name	Description
STATE	MT, ND, OR, UT, WA
AGE	age of participant
ADULT	child: age $<$ 18, adult: age $\geq$ 18
MARRIED	current marital status - 5 categories
GENDER	Male or Female
INCOME	annual income per individual
BirthState	State abbreviation or number
HINS1 to HINS7	insurance related
TRAVEL TIME	travel time to work in minutes

1. A recent newspaper article reported that the average age for people living in the US is 38 years old. You might anticipate that this average could vary by state, with some states being "older" or "younger" in their general demographics.

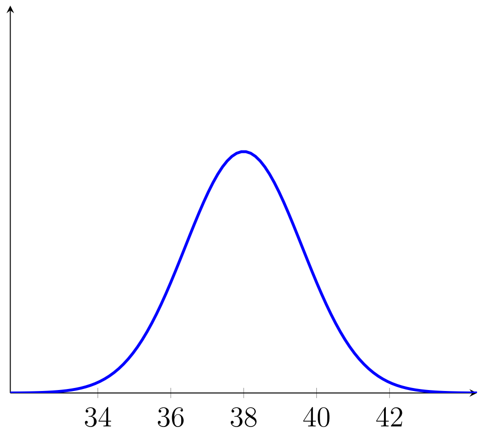
   First investigate the average age in Montana to determine whether it matches the average age of 38 indicated by the article.

   (a) State the null and alternative hypotheses

   (b) Find and label the sample average, $\bar{x}_{MT}$, sample standard deviation, s_{MT}, and sample size, n_{MT}, using a pivot table in Excel. Be sure to focus your attention on only Montana data. You can sort the data by state.

   (c) Find the standard error and label this is Excel. Then compute the test statistic, $\frac{\bar{x}-\mu}{SE}$, and the degrees of freedom.

   (d) Sketch a bell curve of the situation, including the hypothesized population average, μ, and sample average, $\bar{x}_{MT}$, clearly labeled. While you have not yet computed the p-value, shade in the part of the curve that corresponds with the p-value. Make a mental estimation of whether you think it will be large (over 0.10) or small.



(e) Calculate the p-value using Excel.

(f) State your conclusions in a complete sentence related to the context of the problem.

(g) If you rejected the null hypothesis, what should you do to follow-up? (We don't yet have the tools for this. That will appear in the next section.)

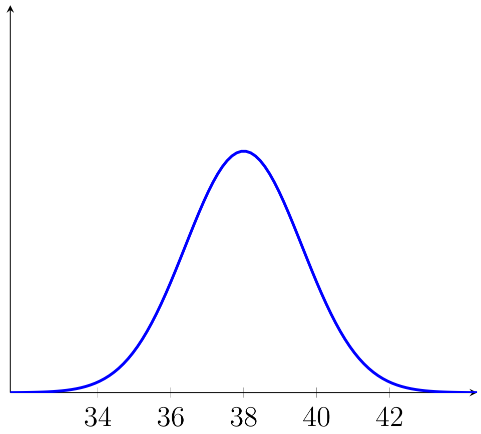
2. Now investigate the average age in Utah to determine whether it matches the average age of 38 indicated by the article.

(a) State the null and alternative hypotheses

(b) Find and label the sample average, $\bar{x}_{Utah}$, sample standard deviation, s_{Utah}, and sample size, n_{Utah}, using a pivot table in Excel.

(c) Find the standard error and label this is Excel. Then compute the test statistic, $\frac{\bar{x}-\mu}{SE}$, and the degrees of freedom.

(d) Sketch a bell curve of the situation, including the hypothesized population average, μ, and sample average, $\bar{x}_{Utah}$, clearly labeled. Shade in the part of the curve that corresponds with the p-value. Make a mental estimation of whether you think it will be large (over 0.10) or small.



(e) Calculate the p-value using Excel.

(f) State your conclusions in a complete sentence related to the context of the problem.

▲

Activity 4.12. Hypothesis testing with one mean, weight examples

1. According to the CDC, the average weight for adult men in the US is 196 lbs. A recent survey of 45 men in technology fields found an average weight of 189 lbs and a standard deviation of $s = 41$ lbs. Conduct a hypothesis test to determine whether men in technology fields tend to weight less than the national average.

2. The average weight of adult women in the US is 170 lbs. A recent survey of 50 women in the financial field found an average weight of 158 lbs and standard deviation of $s = 28$ lbs. Conduct a hypothesis test to determine whether women in the financial field tend to weight less than the national average.

Solution:

Men in Tech		Women in Finance	
μ	196	μ	170
average $\bar{x}$	189	average $\bar{x}$	158
s	41	s	28
n	45	n	50
SE	6.11192	SE	3.9598
test stat	-1.1453	test stat	-3.0305
p-value	0.129138	p*value	0.001946

▲

4.3 One sample means -Confidence Intervals

Preview Activity 4.13. For these questions, use the UsedCar data set on Moodle.

1. A recent ad claimed that the average price for a Buick is $20,000. Is that correct? The UsedCar data set includes 80 randomly selected Buicks, among many other cars. Use this data to test the hypothesis that the average Buick price is $20,000. Once you have conducted your hypothesis test, record your p-value below. Select one:

   (a) The p-value is 0.0069 and so we reject the null hypothesis. The average price is not $20,000.

   (b) The p-value is 0.9931 and so we fail to reject the null hypothesis. The average price is near $20,000.

   (c) The p-value is 0.0138 and so we reject the null hypothesis. The average price is not $20,000.

   (d) The p-value is 2.519 and so we reject the null hypothesis. The average price is not $20,000.

2. A local dealership claims that the average price for a Cadillac is $40,000. Test this claim with a hypothesis test at the $\alpha = 0.05$ level. (Use the same data set). What SE do you calculate for this problem?

3. A local dealership claims that the average price for a Cadillac is $40,000. Test this claim with a hypothesis test at the $\alpha = 0.05$ level. (Use the same data set) Select the appropriate p-value below. Select one:

   (a) The p-value is 0.203 and so we reject the null hypothesis.

   (b) The p-value is 0 and so we should fail to reject the null hypothesis.

   (c) The p-value is 0 and so we reject the null hypothesis.

   (d) The p-value is 0.837 and so we reject the null hypothesis.

   (e) The p-value is 0.203 and so we should fail to reject the null hypothesis.

   (f) The p-value is 0.405 and so we reject the null hypothesis.

   (g) The p-value is 0.797 and so we should fail to reject the null hypothesis.

   (h) The p-value is 0.405 and so we should fail to reject the null hypothesis.

4. A local dealership claims that the average mileage for a used Saturn is 25,000. Test this claim with a hypothesis test at the $\alpha = 0.05$ level. (Use the same data set as the earlier preview questions.)

   (a) What SE do you calculate for this problem?

   (b) State your p-value.

Background information

1. If you conduct a hypothesis test and reject the null hypothesis, use a confidence interval to provide a new estimate for the population mean.

2. In exploratory research, you may not yet have a theory about the population mean you are investigating. In this situation, you may create a confidence interval without first working through a hypothesis test.

3. Walk through Excel commands for calculating the t-critical score, using t.inv(lower tail) or t.inv(upper tail)

Activity 4.14. Return to the census data from the previous section.

Census data dictionary

Variable Name	Description
STATE	MT, ND, OR, UT, WA
AGE	age of participant
ADULT	child: age < 18, adult: age ≥ 18
MARRIED	current marital status - 5 categories
GENDER	Male or Female
INCOME	annual income per individual
BirthState	State abbreviation or number
HINS1 to HINS7	insurance related
TRAVEL TIME	travel time to work in minutes

1. We often use confidence intervals when we want to estimate a parameter (age, weight, income, etc.) and we do not have a prior theory we are evaluating. Create a 95% confidence interval for average individual income for adults (18 and over) in each of the 5 states.

State	Point estimate $\bar{x}$	Margin of Error	95% confidence interval
MT			
ND			
OR			
UT			
WA			

2. Split the data for male and female adults. Find a 98% confidence interval for the average income for adult men.

3. Find a 98% confidence interval for the average income for adult women.

4. At this point, you might be curious whether there is a significant difference between the average incomes for men and women. In the next section, we will learn how how to compare means for two samples. In the meanwhile, based on your confidence intervals, do you anticipate we will find a statistically significant difference in average incomes for men and women? Why?

▲

Activity 4.15. Following up with hypothesis testing

1. The consumer advocacy group who was investigating the 100 calorie snack packs (in Sect 4.1) has conducted an investigation of the Preztels-R-Us 100 calorie packs. They collected a random sample of 20 packs and found an average of $\bar{x}_{calories} = 118$ and a standard deviation of $s = 23$.

   (a) Conduct a hypothesis test.

   (b) Follow up with a 95% confidence interval for the average calorie count for the Pretzels-R-Us bags.

2. In Sect 4.2, you determined that the average age in Montana does not match the national average of 38 years old. Follow up with a 95% confidence interval for the average age of Montanans.

3. Likewise, find a 95% confidence interval for the average age of people in Utah.

▲

4.4 Paired Tests

Reading Assignment 4.16. The goal of this reading is to make sense of a situation involved a paired comparison. In the Cartoon Guide to Statistics, read pages 174 - 180 about Paired Comparisons. You may want to skim the pages just before this so you get the context of the examples. ▲

Preview Activity 4.17. OK. Now that you've read the pages, download the Excel File PairedGasPreview4-4 (linked on Moodle). This is another data set for the gas problem from OpenIntro page 175.

1. Find the standard error for the paired differences: **Solution:** 2.22982248
2. Find the t-score for the hypothesis test: **Solution:** -1.11993976
3. You should find that the t-score is negative. This means:
   (a) Gas B has a larger sample average than Gas A
   (b) Gas B has a smaller sample average than Gas A
   (c) Gas B has a larger population average than Gas A
   (d) Gas B has a smaller population average than Gas A
4. Why do we run both types of gas through the same taxis?
   (a) To eliminate variability between the taxis
   (b) To eliminate the variability between the gasses
   (c) To eliminate the need for statistics
   (d) Because the stats-gods said so
   (e) Shhh. It is a secret.
5. If we were to create a plot like the bottom one on page 177, would the lines lean more left or more right in our data set?

Activity 4.18. Find the data set *textbooks.xlsx* on the Moodle page. This data set is from OpenIntro Stats. In this data set we have 73 textbooks that are sold in the UCLA bookstore with their course abbreviations, course numbers, ISBN numbers, UCLA bookstore price, and the price of the same book if purchased on Amazon.com.

- Discuss: What do you notice? What do you wonder?
- Write a research question for this data set.
- What would your null and alternative hypotheses be?
- Would your hypothesis test require a t or z distribution? Why?

1. Discuss which student-generated questions are appropriate.
2. Perform the most appropriate hypothesis test and write the conclusion in plain language.
3. Create a 95% confidence interval for the average price difference between books at the UCLA bookstore and books on Amazon.com.

▲

Activity 4.19. A set of voting questions about paired differences.

1. Air quality measurements were collected in a random sample of 25 country capitals in 2013, and then again in the same cities in 2014. We would like to use these data to compare air quality between the two years.

   (a) We should use a t test.

   (b) We should use a p test.

   (c) We should use a z test.

   (d) We cannot use a statistical test on this data.

2. Air quality measurements were collected in a random sample of 25 country capitals in 2013, and then again in the same cities in 2014. We would like to use these data to compare air quality between the two years.

   (a) We should use a one-tailed paired t test.

   (b) We should use a one-tailed non-paired t test.

   (c) We should use a two-tailed paired t test.

   (d) We should use a two-tailed non-paired t test.

3. True / False: In a paired analysis we first take the difference of each pair of observations, and then we do inference on these differences.

   (a) True and I am very confident

   (b) True and I am not very confident

   (c) False and I am very confident

   (d) False and I am not very confident

4. True / False: Two data sets of different sizes cannot be analyzed as paired data.

   (a) True and I am very confident

   (b) True and I am not very confident

   (c) False and I am very confident

   (d) False and I am not very confident

   Teacher Note: Generally the answer is no, because the difference in sizes of data usually indicates that the data is not paired. If, however, you know the data is paired (such as pre-post data for individual participants) and a few data points are missing their partners, those data points could be treated as NA and excluded. Then an analysis could be run on the remaining parts of the data.

5. True / False: Each observation in one data set has a natural correspondence with exactly one observation from the other data set in a paired t test.
   (a) True and I am very confident
   (b) True and I am not very confident
   (c) False and I am very confident
   (d) False and I am not very confident

6. True / False: Each observation in one data set has a natural correspondence with the mean from the other data set in a paired t test.
   (a) True and I am very confident
   (b) True and I am not very confident
   (c) False and I am very confident
   (d) False and I am not very confident

7. True / False: Each observation is subtracted from the average of the other data set's observations in a paired t test.
   (a) True and I am very confident
   (b) True and I am not very confident
   (c) False and I am very confident
   (d) False and I am not very confident

▲

Activity 4.20. **Teacher Note:** May want to use this as a handout / exit slip where students submit their work at the end to get a bit of formative feedback before the exam. Also, knowing they have to submit something might be motivating.

Is there strong evidence of climate change? Let's consider a small scale example, comparing how temperatures have changed in the US from 1968 to 2008. The daily high temperatures on Jan. 1 was collected in 1968 and 2008 for 100 randomly selected location in the continental US. Then the difference between the two readings (2008 temp minus 1968 temp) was calculated for each of the 100 different locations. The average of these 100 values was 1.1 degrees with a standard deviation of 4.9 degrees.

1. Is this a case for a paired t test?

2. Write appropriate hypotheses.

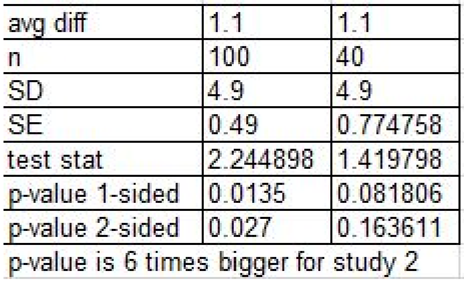
3. Conduct your hypothesis test. Label all important pieces in your Excel sheet.

4. What do you conclude?

Follow-up. Another researcher attempts to duplicate this study using 40 randomly selected US cities. Conveniently, they happen to get the same average difference of 1.1 degrees with a standard deviation of 4.9 degrees.

1. Explain why this second p-value will not be significant by explaining what changes in the formulas used to calculate the p-value.
2. Explain why two studies with the same standard deviation and the same average difference could give you different conclusions.
3. If both of these studies and p-values were presented to you, what would you conclude about evidence for climate change?

Solution: Because the sample size is smaller in the second version of the study, our evidence becomes weaker. However, this doesn't nullify the results from the earlier study. The study that fails to show significance does not actually prove that the null hypothesis is true. It just isn't strong enough to cause us to reject our null hypothesis. The first study is strong enough to cause us to reject the null.



avg diff	1.1	1.1
n	100	40
SD	4.9	4.9
SE	0.49	0.774758
test stat	2.244898	1.419798
p-value 1-sided	0.0135	0.081806
p-value 2-sided	0.027	0.163611
p-value is 6 times bigger for study 2		

▲

4.5 Review

Activity 4.21. Review practice questions based on the Used Car data set
Download the UsedCar data set (adapted from OpenIntro stats).

1. Create a 99% confidence interval for the average price for a Pontiac.

2. Create a 95% confidence interval for the average milage for a SAAB.

3. Conduct a hypothesis test for the null hypothesis that the average mileage of a Cadillac is 20,000 miles. Clearly state your results, including your p-value.

4. A used car salesman says that about 50% of Buicks have leather seats. A consumer group thinks that number is lower. Conduct a hypothesis test and clearly state your results, including your p-value.

5. Is there a difference in the proportion of cars with leather seats, depending on whether the car is a Buick or a Saturn? Conduct a hypothesis test and clearly state your results, including your p-value.

6. Hank argues that 70% of all cars have sound systems. Using your full set of car data, conduct a hypothesis test about Hank's claim.

7. Is a car with leather seats more likely to also have a sound system? Use your knowledge of statistics to address this question.

   Following-up: If any of your hypothesis tests result in rejection of the null, be sure to follow-up with a 95% confidence interval.

8. Create 5 more statistical questions. Be sure to include (a) one proportion, (b) two proportions, and (c) one mean (t-test), as well as confidence intervals and both one tailed and two tailed hypothesis tests. The Used Car data set does not lend itself to paired means statistical questions, though it provides a rich data set for unpaired means tests. Check with your instructor to find out if paired means and unpaired means are on your upcoming exam.

   Teacher Note: Depending on your timing, it might be better to delay the paired and unpaired means for exam 3, after students have completed the next set of labs. Unpaired means is the next section.

▲

4.6 Test 2

4.7 Difference of means

Reading Assignment 4.22. Read pages 168 and 169 in The Cartoon Guide to Statistics. Take careful note of the formulas in this section. Then in the OpenIntro Stats book, read pages 176-183 (Section 4.3) ▲

Preview Activity 4.23. Now that you've read about how to compare the means between two populations, let's put some of that reading to practice.

- If we have two samples with the following information, what is the standard error for the sampling distribution of the difference between the means?

Sample 1:	size = 200	mean=7	sd=1.5
Sample 2:	size = 250	mean=6.5	sd = 2

  Standard Error = **Solution:** 0.1245

- In the samples listed above we see a difference of 0.5 between the means. What is the test statistic for this difference? **Solution:** 4.016

- Based only on the test statistic, are these sample means significantly different?

At this point in the course, the basic structure of a hypothesis test should be well engrained. Once you have your research question, you engage in something like the following plan:

1. Look at your data (if you start with raw data rather than summary statistics). Create plots. TinkerPlots is helpful for this. What do you notice? Is there anything odd or suprising about the data? Are there missing or bizarre data points?

2. Write appropriate null and alternative hypotheses.

3. Make a list of the descriptive statistics that you would need for this test.

4. Look up the appropriate standard error formula for your test.

5. If you are using the t distribution, look up the formula for the degrees of freedom.

6. Run the test, find a p-value, and state your conclusion.

7. If you reject your null hypothesis, create a confidence interval (usually 95%) for your parameter of interest.

So far we have conducted hypothesis tests for one proportion and two proportion problems, as well as one mean and paired means problems. In this section we investigate scenarios where we compare two means (not paired). To do this, we will need formulas for standard error and degrees of freedom.

Confidence Interval:

- The sample means are: $\bar{x}_1$ and $\bar{x}_2$
- We are interested in: $\bar{x}_1 - \bar{x}_2$
- Degrees of Freedom: $df = \min\{n_1 - 1, n_2 - 1\}$
- Standard Error: $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
- Confidence Interval: $(\bar{x}_1 - \bar{x}_2) \pm t^*_{df} \times SE$

Hypothesis Test:

- Null Hypothesis: $H_0 : (\mu_1 - \mu_2) = 0$
- Alternate Hypothesis: $H_A : (\mu_1 - \mu_2) \neq, <, or > 0$
- Standard Error: $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
- Test Statistic: $t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{SE}$

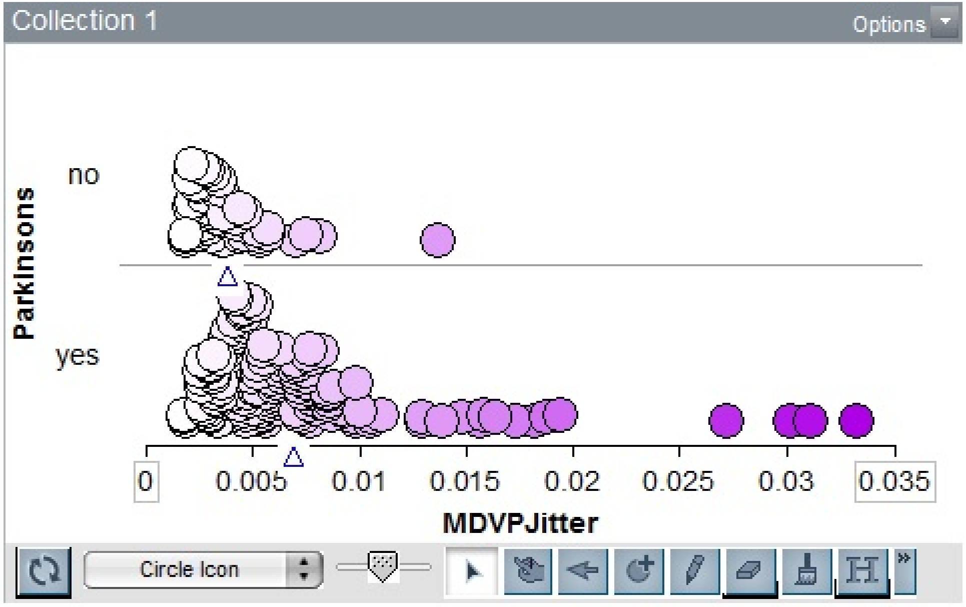
Activity 4.24. Listening for Parkinson's

Parkinson's is a disease of the nervous system which affects muscle control, movement, and speech. In a study at the University of Oxford, Dr. Max Little investigated the vocal patterns of research participants with and without Parkinson's disease. He measured attributes including vocal frequency (in Hz), jitter (variations in frequency), and shimmer (variations in amplitude). A subset of this data is included in the file ParkinsonsSpeech (data set from UCI Machine Learning Repository).

Source: Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection', Little MA, McSharry PE, Roberts SJ, Costello DAE, Moroz IM. BioMedical Engineering OnLine 2007, 6:23 (26 June 2007)

1. Open up the data set in Excel and copy the columns into TinkerPlots. Explore the data. What do you notice about frequency, jitter, and shimmer for the Parkinson's and non-Parkinson's patients?

Teacher Note:

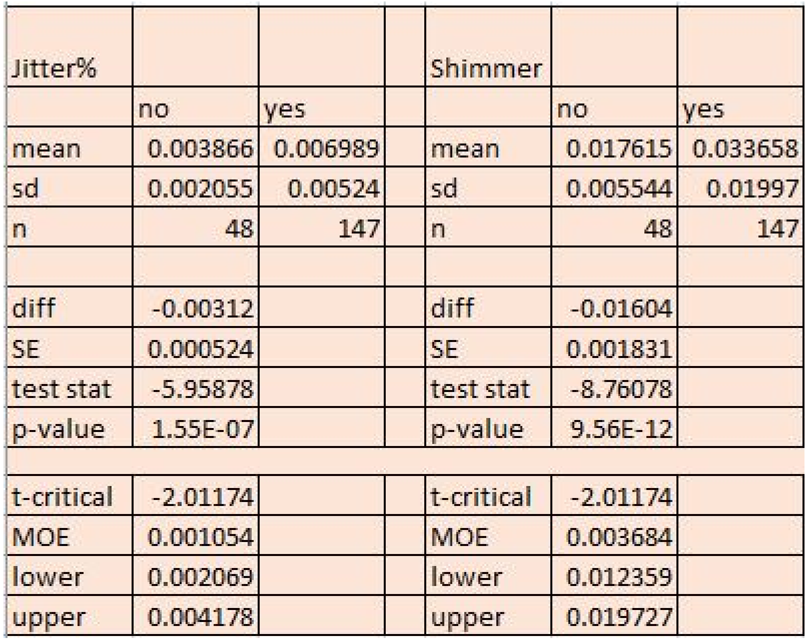


2. From your graphs, you might notice that vocal jitter looks somewhat higher on the

Parkinson's group than the non-Parkinson's group. Conduct a hypothesis test to determine if the Parkinson's group has a higher average level of jitter in speech. Note: This is a one tailed test because of our prior knowledge of Parkinson's as a disease that affects muscle control in speech.

3. Conduct a hypothesis test to determine if the Parkinson's group has a higher average level of vocal shimmer.

   Solution:



Jitter%				Shimmer		
	no	yes			no	yes
mean	0.003866	0.006989		mean	0.017615	0.033658
sd	0.002055	0.00524		sd	0.005544	0.01997
n	48	147		n	48	147
diff	-0.00312			diff	-0.01604	
SE	0.000524			SE	0.001831	
test stat	-5.95878			test stat	-8.76078	
p-value	1.55E-07			p-value	9.56E-12	
t-critical	-2.01174			t-critical	-2.01174	
MOE	0.001054			MOE	0.003684	
lower	0.002069			lower	0.012359	
upper	0.004178			upper	0.019727	

4. Both vocal jitter and shimmer are statistically higher in the Parkinson's group, so it's appropriate to follow-up with a 95% confidence interval about the difference between the Parkinson's group and the non-Parkinson's group. State your confidence intervals as complete sentences related to the context.

5. Do individuals with Parkinson's disease always have higher levels of vocal jitter and shimmer than individuals without Parkinson's disease?

   Solution: Individuals with Parkinson's do not always have higher levels. As we see in the Tinkerplots graph, there are individuals with and without Parkinson's who have the same levels of Jitter and Shimmer. However, there is a difference in averages between the two groups. Reiterate that what holds for averages does not always hold for individuals. This is because of how variability in data works.

▲

Activity 4.25. Download the VideoGamesXBoxPS3 data set from Moodle.

Is there a difference in average sales per video game between the XBox360 and PS3 platforms? If you were a game designer, would one platform tend to lead to more profits than the other? We're going to explore this question for different geographic areas. For each of the following, conduct a hypothesis test using $\alpha = 0.05$. Note, the columns for sales are measured in millions of dollars.

1. Is there a difference in averages sales per video game between XBox360 and PS3 for the Global Sales column?

2. Is there a difference in averages sales per video game between XBox360 and PS3 for just North America (NA sales column)?

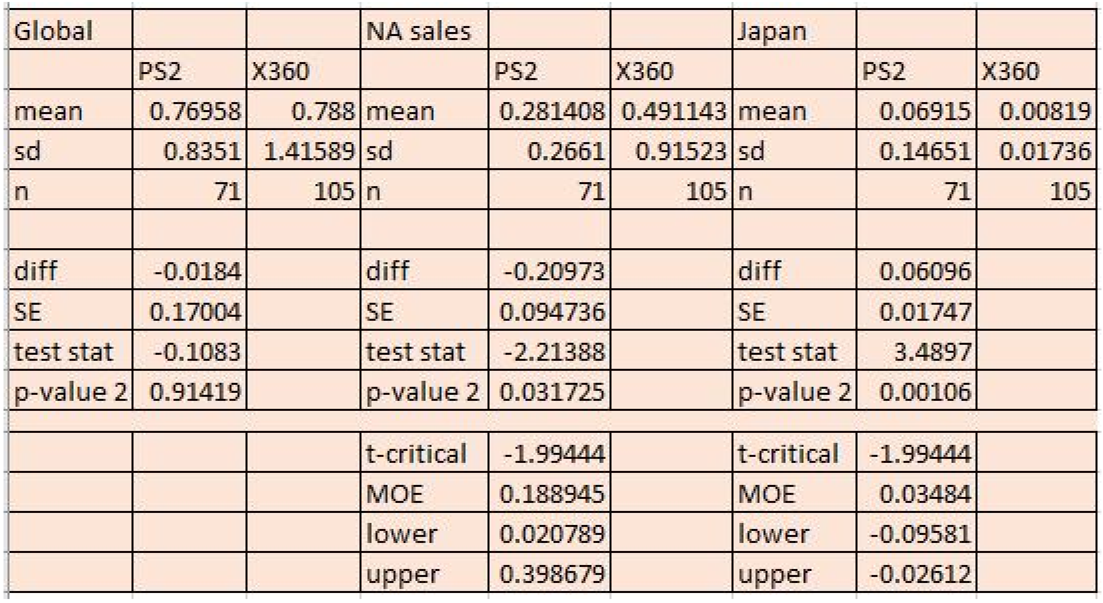
3. Is there a difference in averages sales per video game between XBox360 and PS3 for just Japan (JP sales column)?

4. Looking at your answers for the previous two questions, what do you notice about your results?

   Solution: In North America, XBox has higher sales and in Japan its reversed, PS3 does much better.

5. Concept check: Globally, XBox 360 games had a noticeably higher standard deviation than PS3 games. What does that mean in practical terms for video game sales?

Solution:



Global			NA sales			Japan		
	PS2	X360		PS2	X360		PS2	X360
mean	0.76958	0.788	mean	0.281408	0.491143	mean	0.06915	0.00819
sd	0.8351	1.41589	sd	0.2661	0.91523	sd	0.14651	0.01736
n	71	105	n	71	105	n	71	105
diff	-0.0184		diff	-0.20973		diff	0.06096	
SE	0.17004		SE	0.094736		SE	0.01747	
test stat	-0.1083		test stat	-2.21388		test stat	3.4897	
p-value 2	0.91419		p-value 2	0.031725		p-value 2	0.00106	
			t-critical	-1.99444		t-critical	-1.99444	
			MOE	0.188945		MOE	0.03484	
			lower	0.020789		lower	-0.09581	
			upper	0.398679		upper	-0.02612	

Recall the basic process for a hypothesis test:

1. Explore the data visually.

2. Write appropriate null and alternative hypotheses

3. Make a list of the descriptive statistics that you would need for this test. Organize these carefully in your Excel sheet.

4. Look up the appropriate standard error formula for your test

5. If you are using the t distribution, look up the formula for the degrees of freedom.

6. Run the test, find a p-value, and state your conclusion. Assume that $\alpha = 0.05$.

7. If you reject your null hypothesis, follow-up with a 95% confidence interval.

▲

Activity 4.26. **Teacher Note:** Find data sets relevant to your students, your local community, or your academic setting. Present the story of the data to your class. Then ask the class to generate questions. If you find good, compelling data sets that are free for sharing, please consider sending them to the Active Stats team at Carroll College. We are always looking for new and interesting data sets (with appropriate background story and data descriptions).

In the meanwhile, Cards Against Humanity data set (PulseOfTheNationSept2017) data is appropriate for this task.

Write three research questions that could be explored with the given data set. Make sure at least one involves the difference of two proportions.

Question 1:
Question 2:
Question 3:
Now answer your research questions.

1. Explore the data visually.

2. Write appropriate null and alternative hypotheses

3. Make a list of the descriptive statistics that you would need for this test. Organize these carefully in your Excel sheet.

4. Look up the appropriate standard error formula for your test

5. If you are using the t distribution, look up the formula for the degrees of freedom.

6. Run the test, find a p-value, and state your conclusion. Assume that $\alpha = 0.05$.

7. If you reject your null hypothesis, follow-up with a 95% confidence interval.

▲

4.8 Lab 7 Part 1 - Working with real world data sets - AirBnB

Reading Assignment 4.27. Read Lab 7 Part 1 - and complete problem 1 before class. ▲

Preview Activity 4.28. The preview activity asks about problem 1 of Lab 7 Part 1 (AirBnB data).

4.9 Lab 7 Part 2 - Working with real world data sets - Blackfoot River

Reading Assignment 4.29. Read Lab 7 Part 2 and complete problem 1 before class. ▲

Preview Activity 4.30. The preview activity asks about problem 1 of Lab 7 Part 2 (Fish data).

4.10 ANOVA - What is it and how do we interpret it?

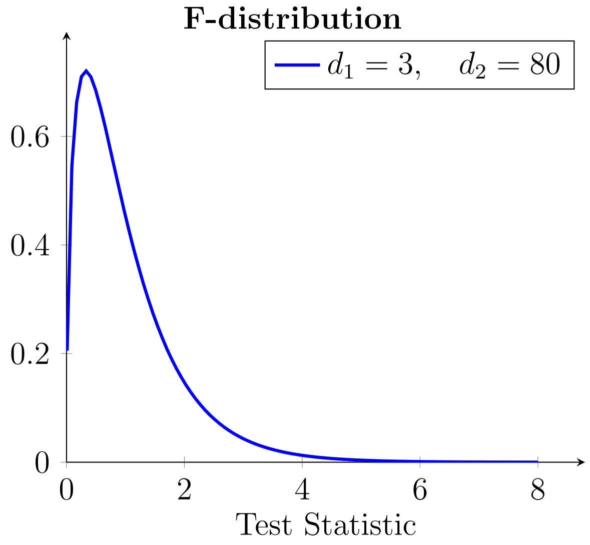
Teacher Note: Plan to use Excel's Analysis Toolpack or other pre-built resources. Focus attention on interpretation and follow-up.

This could be one or two days. The content is adapted from the OpenIntro textbook. The example is different (BMI data was invented for this task). Read the Diez book for additional bits of knowledge you may need to work with ANOVA.

The following link has an explanation of how to use the ANOVA tools in Excel: http://www.excel-easy.com/examples/anova.html

When you would like to look for differences between three or more means, you need to upgrade from a difference of means hypothesis test to an ANOVA hypothesis test. ANOVA refers to ANalysis Of VAriance. ANOVA determines whether the variability between different groups is more notable than the variability within groups. In other words, we expect there to be variability within each subset of the data and we expect variability between subsets of data. If the variability between subsets is much higher than the variability inside subsets, that could indicate that the subsets are different from one another in important ways. Measuring these two types of variability is challenging and we will rely on technology for the gritty details.

In order to perform a hypothesis test with ANOVA, we must introduce a new sampling distribution, the F-distribution. Like the t-distribution, the F-distribution is actually a family of distributions, with differences depending on degrees of freedom. Unlike our previous distributions, F is not symmetric. We will use the F-distribution to compute our p-value, which is represented by the area in the upper tail, to the right of our test statistic.



Before performing ANOVA, check that

1. data within each group appears to be nearly normal
2. observations are independent
3. variability is similar across groups

We will explore the big ideas of ANOVA with an example comparing different exercise programs. Each program is considered a “treatment.”

Example 4.31. Four different exercise programs are under review. The programs vary in types of exercise, accountability, and intensity. We’ll refer to the programs as A, B, C, and D. Because the research is being conducted at a university, the researcher has selected 84 student volunteers and randomly assigned them to one of the four fitness programs. Students body mass index (BMI) is calculated before and after the 6 month study. Change in BMI is used as the key measure of impact of the exercise programs. We’ll use a significance level of $\alpha = 0.05$.

BMI data is in BMIAnoveData file.

An appropriate null hypothesis is to assume that the four exercise programs are all equally helpful (or equally unhelpful). Therefore, we would claim that all four averages are equal.

$H_0 : \mu_A = \mu_B = \mu_C = \mu_D$.

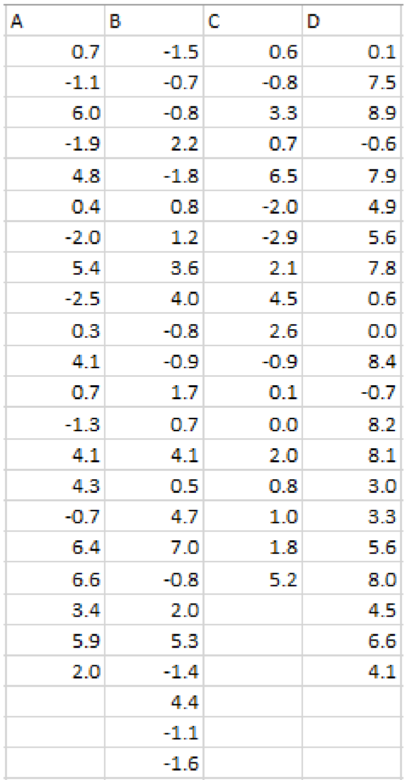
The alternate hypothesis is that at least one of the four population means is different from at least one another population mean.

Side note: It might be tempting to approach this problem with the strategies we used for comparing two means in our earlier sections. In that case, we would conduct a hypothesis test for teach of the following pairs:

$$\begin{array}{ccc} \mu_A = \mu_B & \mu_A = \mu_C & \mu_A = \mu_D \\ & \mu_B = \mu_C & \mu_B = \mu_D \\ & & \mu_C = \mu_D \end{array}$$

This would require 6 distinct hypothesis tests. Not only would this be annoying, it would be statistically problematic. Each hypothesis test has a certain probability (α) of a Type I error (finding a difference when none exists). Conducting 6 tests makes the likelihood of engaging in that error quite a bit higher. ANOVA will allow us to examine all 4 population means in a single test. If a difference is detected, then we can follow up with difference of means hypothesis tests.

When performing ANOVA in Excel, arrange your data in columns so that each group is in its own column. Be sure to provide labels, as the ANOVA tool in Excel will adopt those labels when it outputs the results.

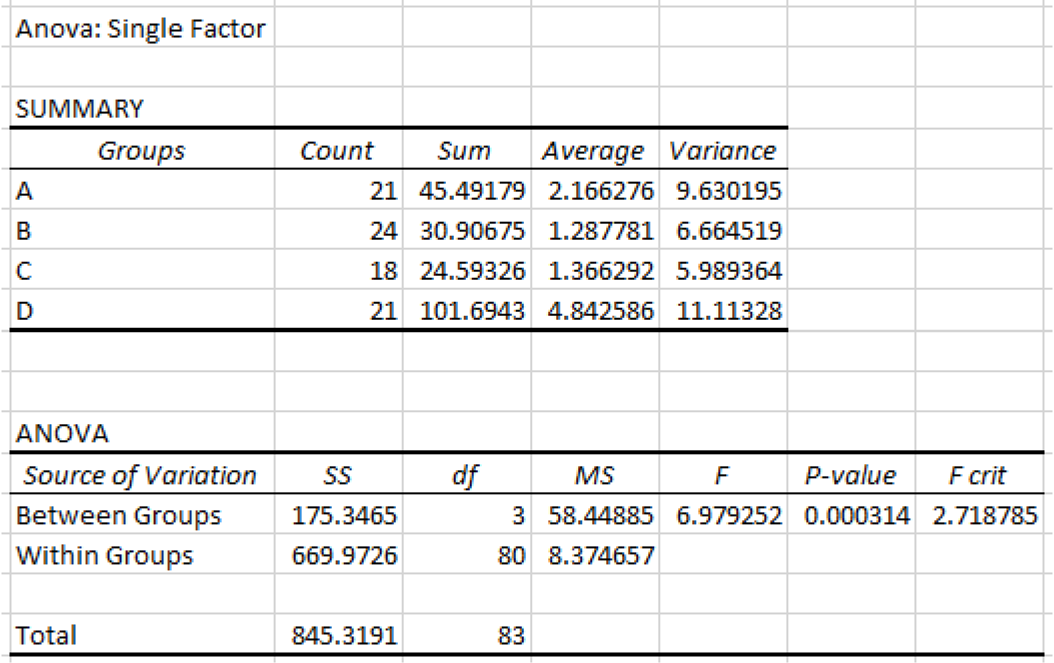


A	B	C	D
0.7	-1.5	0.6	0.1
-1.1	-0.7	-0.8	7.5
6.0	-0.8	3.3	8.9
-1.9	2.2	0.7	-0.6
4.8	-1.8	6.5	7.9
0.4	0.8	-2.0	4.9
-2.0	1.2	-2.9	5.6
5.4	3.6	2.1	7.8
-2.5	4.0	4.5	0.6
0.3	-0.8	2.6	0.0
4.1	-0.9	-0.9	8.4
0.7	1.7	0.1	-0.7
-1.3	0.7	0.0	8.2
4.1	4.1	2.0	8.1
4.3	0.5	0.8	3.0
-0.7	4.7	1.0	3.3
6.4	7.0	1.8	5.6
6.6	-0.8	5.2	8.0
3.4	2.0		4.5
5.9	5.3		6.6
2.0	-1.4		4.1
	4.4		
	-1.1		
	-1.6		

The following tables show typical output for an ANOVA test. The summary table tells us useful information about each of the subgroups. The ANOVA table below that contains our p-value (0.000314) and our test statistic F (6.979). This table also contains several other interesting pieces of information. The column labeled MS has two values: *mean square between groups* (MSG), which measures the variability between groups; and *mean square error* (MSE), also called *mean square within groups*, which measures the variability within groups. You may notice that the variability between groups (MSG) is almost 7 times higher than the variability within groups (MSE). This is what the F test statistic tells us.

$$F = \frac{MSG}{MSE} = \frac{58.44885}{8.374657} = 6.979$$

That is a lot more variability between groups than within groups. Our cutoff for a significance level of 0.05 would be the F critical value shown in the table as 2.718785.



Anova: Single Factor

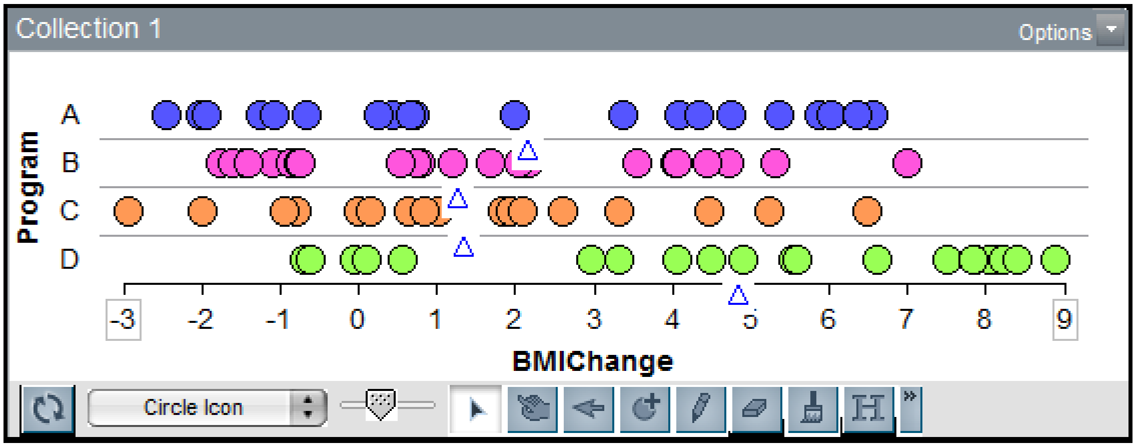
SUMMARY

Groups	Count	Sum	Average	Variance
A	21	45.49179	2.166276	9.630195
B	24	30.90675	1.287781	6.664519
C	18	24.59326	1.366292	5.989364
D	21	101.6943	4.842586	11.11328

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	175.3465	3	58.44885	6.979252	0.000314	2.718785
Within Groups	669.9726	80	8.374657			
Total	845.3191	83				

With a p-value of 0.000314, it does appear that at least one of the population means is different. That is, not all of the exercise programs have the same average results for change in BMI. In formal statistical language, we reject the null hypothesis that all of the exercise programs produce equal results. Now it would be appropriate to follow-up and determine which of the means are different.



Following up:

Looking at the plot above, is it clear which of the exercise program averages are different from one another in a statistically significant way? The sample averages $\bar{x}_i$ are indicated with blue triangles. At first glance, it may look like D and B have the biggest difference, but are D and C also different? What about D and A?

We need a method to decide which averages are different.

- Find the averages from each group
- Find the pooled standard deviation (s_{pooled}). This can be found by taking the square root of MSE, or by the following, more arduous formula, where k = number of groups and s_i indicates the standard deviation for group i.

$$s_{pooled} = \sqrt{MSE} \qquad \text{or}$$
$$s_{pooled} = \sqrt{\frac{n_1 \cdot s_1^2 + n_2 \cdot s_2^2 + ...n_k \cdot s_k^2}{n_1 + n_2 + ... + n_k}}$$

  In our example, $s_{pooled} = \sqrt{MSE} = \sqrt{8.37} = 2.89$

- Choosing two groups at a time, say exercise programs A and D, find the difference in means and determine the test statistic for that difference.

  $\bar{x}_A - \bar{x}_D = 2.166 - 4.843 = -2.676$

  $SE = \sqrt{\frac{s_{pooled}}{n_A} + \frac{s_{pooled}}{n_D}} = \sqrt{\frac{2.89}{21} + \frac{2.89}{21}} = 0.89$

  While we use the same s_{pooled} for any pair of means, the SE value can change if the sample sizes differ. As usual, bigger sample sizes lead to smalled SE.

  The test statistic is determined by dividing $t_{stat} = \frac{\bar{x}_A - \bar{x}_D}{SE} = \frac{-2.676}{0.89} = -2.997$

- Use the test statistic to find the p-value.

  This will be a two-tailed t-test, with degrees of freedom given by the“’within groups” $df = 80$ listed in the chart. This degrees of freedom comes from the total number of people involved (84) minus total number of groups (4).

  $p - value = 2 \times T.DIST(teststat, df, 1) = 2 \times T.DIST(-2.997, 80, 1) = 0.00363.$

- Interpret your p-value. With a p-value of 0.00363, there is definitely a significant difference between results for those in exercise program A and exercise program D.

  Looking at our chart of data, if A and D are significantly different, pairs B and D and pairs C and D will also be significantly different. What about A vs. B? Or A vs. C? For this we would conduct additional tests.

$$\bar{x}_A - \bar{x}_B = 2.166 - 1.288 = 0.878$$
$$SE = \sqrt{\frac{s_{pooled}}{n_A} + \frac{s_{pooled}}{n_B}} = \sqrt{\frac{2.89}{21} + \frac{2.89}{24}} = 0.865$$

$$t_{stat} = \frac{\bar{x}_A - \bar{x}_D}{SE} = \frac{0.878}{0.865} = 1.016$$
$$pvalue = 2 \times (1 - T.DIST(teststat, df, 1) = 2 \times T.DIST(1.016, 80, 1) = 0.313.$$

Warning Earlier we warned that if we conducted too many difference of means t-tests, the likelihood of a Type I Error would rise. To adjust for that, one option is the Bonferroni correction. We determine the number of pairwise tests (in the previous case K = 6) and divide our usual significance level α by this count. i.e. $\alpha^* = \alpha \div K$. For the prior problem, our $\alpha^* = 0.05/6 = 0.0083$.

In general, if there are r means to compare, then the number of tests is $K = \frac{r(r-1)}{2}$.

Activity 4.32. Revisiting data sets with ANOVA

1. Think back over the data sets we have been investigating this term. Which data sets had quantitative data for 3 or more subgroups? e.g. 4 types of fish, 3 groups of people, 5 types of cars, 4 categories of video games, etc. Which data sets did you find interesting?

2. Choose a data set and describe several statistical questions that you could approach with ANOVA. Then format your data appropriately and conduct an ANOVA test. Describe your results.

   Note: The Excel ANOVA tool assumes data will be given in column form, with one treatment group per column. To achieve this, you may need to sort your data and engage in a bit of cutting and pasting.

▲

4.11 IRB Lab

Teacher Note: The IRB lab can be moved later in the term as well. It may be appropriate to place it after regression, as the final lab of the year. Then students can imagine an even broader range of research studies.

Many of our students at Carroll engage in research while they are undergraduates. Psychology, health science, anthrozoology, and other programs offer opportunities for students to engage in research on human subjects. With this opportunity comes the responsibility to make ethical choices about protecting human research subjects. One tool that institutions of higher education use is the Institutional Review Board, or IRB. A researcher or research team puts together a proposal of their intended research and submits it to IRB for approval. The intention of this assignment is to prepare students to complete an IRB application. When creating an IRB application, you will have the opportunity to think about and articulate:

- The research goals of the study
- The types of data you will collect and how you will collect it
- The types of participants you will include and how you will recruit them
- The potential risks and benefits to the participants
- The statistical tests you will employ
- A consent form for the participants

Before you get started, spend a few minutes thinking about a research study that you would be interested in conducting, perhaps as part of a senior project in your major. For this lab, the study must involve humans and it must be a randomized controlled experiment, not an observation study. Plan to have a treatment and control group, or multiple treatment groups. Discuss your research study with your instructor before you get too far into the IRB sample application.

Many institutions (including Carroll College) have a template for the IRB application. These vary by institution, so contact your local IRB to see what they require. The template provided for this lab is an abridged version of the Carroll IRB application. When you are ready to conduct a human subjects research experiment, work with your research advisor to complete the appropriate application.

4.12 Summary

Summary 4.33. Student learning outcomes from Chapter 4

1. Students are able to make use of the Central Limit Theorem and have a conceptual understanding of sampling distributions for quantitative data.

2. Students are able to conduct hypothesis tests and create confidence intervals for one-mean, paired-mean, and unpaired-means scenarios. Students are able to interpret the results into everyday language.

3. Students appropriately choose between normal distributions and t-distributions. When working with t-dstributions, they are able to find the appropriate degrees of freedom.

4. Students are able to recognize scenarios that require ANOVA hypothesis testing. They are able to use software to conduct ANOVA testing and they are able to interpret the results and conduct appropriate follow-up tests.

5. Students are confident in working with real-world data sets, including somewhat messy data sets or data sets with missing entries.