MSDS 6371-405 Analysis Guide

David Josephs

October 21, 2018

Contents

VII Workflow for testing hypotheses 205

List of Codes

Part I

Drawing Statistical Conclusions

Chapter 1

Problem 1: Randomized Experiment vs Random Sample

Question 1

What is the difference between a randomized experiment and a random sample? Under what type of study/sample can a causal inference be made?

Answer to Question 1

A randomized experiment is when the the application of the experimental variable ("treatment") is applied to subjects chosen randomly. So for example, in a study with 400 subjects, and treatments A, B, and a control group, each subject would randomly be assigned into either the control group, group A, or group B. This is done to eliminate confounding variables, as well as possible bias. In a random sample, subjects are randomly chosen from the population. This is done so that the subjects of the study can be assumed to be representative of the population as a whole. [1]. We can make causal inferences from a randomized experiment, but not from a random sample.

Score: 20/20. Explanation: This answer gets full marks because it covers all of the points made in the key, it defines both random sampling and randomization in the same manner as the key. However in the future it should be less wordy.

Chapter 2

Problem 2: Identifying Confounding Variables

Question 2

In 1936, the Literary Digest polled 1 out of every 4 Americans and concluded that Alfred Landon would win the presidential election in a landon-slide. Of course, history turned out dramatically different (see http://historymatters.gmu.edu/d/5168/ for further details). The magazine combined three sampling sources: subscribers to its magazine, phone number records, and automobile registration records. Comment on the desired population of interest of the survey and what population the magazine actually drew from.

Answer To Question 2

The magazine had hoped to get a random sample, or a dichotomy of the voting population, which would be representative of the entire voting population of the country as a whole. Instead, they only polled subscribers to the magazine, phone number records, and automobile registration records. 1936 was in the height of the great depression, which means that the average American was struggling to survive. Therefore, while in the past this sampling techique had worked, this time around they ended up only sampling the wealthiest people, those who could afford phones, cars, and magazine subscriptions, and the results were not representative of the population. Without truly random sampling, "the statistical results only apply to [those] sampled", and cannot be representative of the entire population. [2]. Therefore, itis just chance that in the previous years, the polls worked.

Score: 10/10. Explanation: This answer gets full marks because it states that the poll wanted to cover all of the voters (5 points), and it identifies the actual group polled with some explanation (affluent people) (5 points).

Chapter 3

Problem 3: Identifying a Scope of Inference

Question 3

3. Suppose we have developed a new fertilizer that is supposed to help corn yields. This fertilizer is so potent that a small vial of it sprayed over an entire field is a sufficient dose. We find that the new fertilizer results in an average yield of 60 more bushels over the old fertilizer with a p-value of 0.0001. Write up a scope of inference under the following study designs that generated this data.

1. We offer the new fertilizer at a discount to customers who have purchased the old fertilizer along with a survey for them to fill out. Some farmers send in the survey after the growing season, reporting their crop yield. From our records, we know which of these farmers used the new fertilizer and which used the old one.
2. When a customer makes an order, we randomly send them either the old or new fertilizer. At the end of the season, some of the farmers send us a report of their yield. Again, from our records, we know which of these farmers used the new fertilizer and which used the old.
3. When a customer makes an order, we randomly send them either the old or new fertilizer. At the end of the season, we sub-select from the fertilizer orders and send a team out to count those farmers' crop yields.
4. We offer the new fertilizer at a discount to customers who have purchased the old fertilizer. At the end of the season, we sub-select from the fertilizer orders and send a team out to count those farmers' crop yields. From our records, we know which of these farmers used the new fertilizer and which used the old one.

Answer

1. We cannot make causal inferences or inferences about the population, as it was not randomized or a random sample. Available units from distinct groups were selected, however the treatment was not assigned randomly, which may mean only farmers who needed a change in fertilizer or were struggling and could not afford the old fertilizer decided to go for the discount, and then the study is also only representative of those who submitted reports, as no random sampling was done

   Score: 8/8. Explanation: This answer gets full credit because it states that causal inferences cannot be made and that population inferences cannot be made, which agrees with the key

2. We can make causal inferences but not inferences about the population. The treatment was applied at random to the subjects, but no random sampling was done. Therefore this study only speaks to the effect of the treatment on farmers who submitted reports, which may mean that they had noteably different yields.

   Score: 8/8. Explanation: This answer receives full credit because it states that causal inferences can be made, and that population statements cannot be made, with explanations, all agreeing with the key

3. We can make causal inferences and inferences about the population. The farmers were randomly assigned different treatments, which allows us to make causal inferences, and then the farmers were randomly selected for the yield to be counted, which means that the selected farmers should be representative of the entire population. With these experimental parameters, we can decide whether the new fertilizer worked better, worse, or the same.

Score: 7/8. Explanation: This answer loses a point because the problem does not explicitly state that the sub sample was random. I assumed it was a random sample, and with that assumption, the answer is entirely correct, however the randomness is not explicitly stated. Therefore a point is taken away. The rest of the answer agrees entirely with the key, therefore no more points will be lost

4. We can make inferences about the population but not causal inferences. The treatment was not supplied randomly, so maybe only farmers who needed a discount or the old fertilizer wasnt working for chose the new fertilizer. However, they were randomly sampled, which means we can make inferences about the population to some degree but we definitely cannot make causaul inferences.

   Score: 7/8. Explanation: This answer loses a point because the problem does not explicitly state that the sub sample was random. I assumed it was a random sample, and with that assumption, the answer is entirely correct, however the randomness is not explicitly stated. Therefore a point is taken away. The rest of the answer agrees entirely with the key, therefore no more points will be lost.

Chapter 4

Problem 4: Visual comparison of population means and a permutation test

Question 4

4. A Business Stats class here at SMU was polled, and students were asked how much money (cash) they had in their pockets at that very moment. The idea was to see if there was evidence that those in charge of the vending machines should include the expensive bill / coin acceptor or if the machines should just have the credit card reader. Also, a professor from Seattle University polled her class last year with the same question. Below are the results of the polls. SMU 34, 1200, 23, 50, 60, 50, 0, 0, 30, 89, 0, 300, 400, 20, 10, 0 Seattle U 20, 10, 5, 0, 30, 50, 0, 100, 110, 0, 40, 10, 3, 0

1. Use SAS to make a histogram of the amount of money in a student's pocket from each school. Does it appear there is any difference in population means? What evidence do you have? Discuss your thoughts.
2. Use the following R code to reproduce your histograms. Simply cut and paste the histograms into your HW. SMU = c(34, 1200, 23, 50, 60, 50, 0, 0, 30, 89, 0, 300, 400, 20, 10, 0) Seattle = c(20, 10, 5, 0, 30, 50, 0, 100, 110, 0, 40, 10, 3, 0) hist(SMU) hist(Seattle)
3. Run a permutation test to test if the mean amount of pocket cash from students at SMU is different than that of students from Seattle University. Write up a statistical conclusion and scope of inference (similar to the one from the PowerPoint). (This should include identifying the Ho and Ha as well as the p-value.)

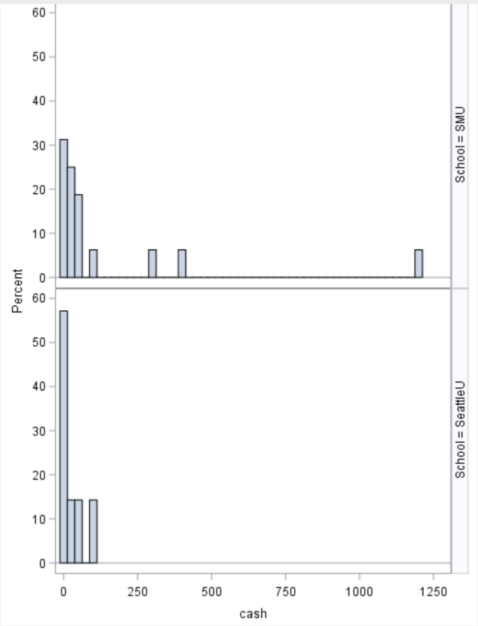
Answer

1. Code (see Appendix 1) for the SAS histogram (Figure 1) was inspired by [3]. The code used to produce this histogram is as follows:

Code 4.1. Creating Paneled histograms in SAS

```
proc sgpanel data=CashMoney;
panelby School / rows=2 layout=rowlattice;
histogram cash / binwidth = 25;
run;
```

Figure 4.0.1. Distribution of Cash by School, produced in SAS



It appears that for the sample means, the SMU sample has a slighly higher mean, however I do not believe that means that the population of SMU has a higher mean than Seattle U, as this was not a random sample, it was just of business students. It appears that the SMU cash distribution is wider, with higher values, but again it is hard to tell if it is indicative of the entire population, I believe, based off of where the majority of the distributions lie, both populations would have similar means, with SMU having a slightly higher mean. SMU is a private school and Seattle U is one of the best value schools in the country, so it is possible that SMU students might have in general, more money than students at Seattle U, and therefore more cash.

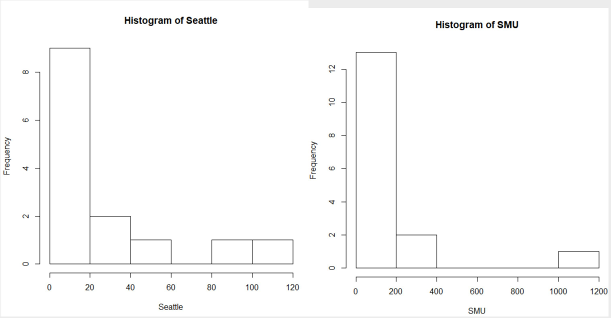
Score: 5/5. Explanation: This receives full marks, the histograms are correct and the conclusions are similar to the key, and are very logical. The code is included in the appendix.

2. The code used to generate the R histograms (Figure 2) was given in the homework and is presented below

Code 4.2. Producing histograms in R

```
SMU = c(34, 1200, 23, 50, 60, 50, 0, 0, 30, 89, 0, 300, 400, 20, 10, 0)
Seattle = c(20, 10, 5, 0, 30, 50, 0, 100, 110, 0, 40, 10, 3, 0)
par(mfrow=c(1,2))
hist(SMU)
hist(Seattle)
```

Figure 4.0.2. Cash Distributions at SMU and Seattle U, Produced using R

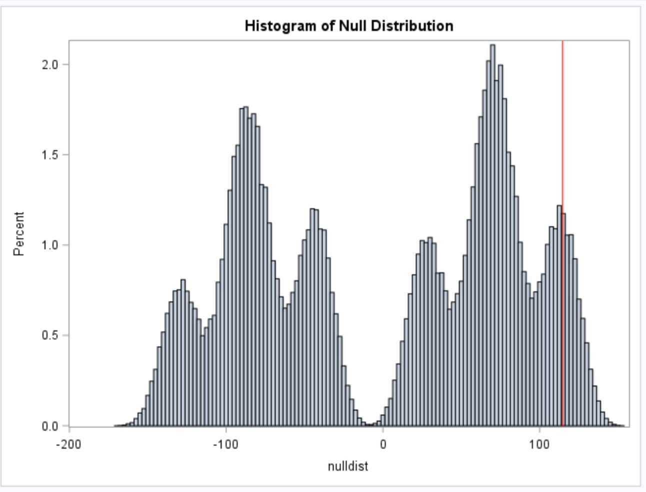


he code used to generate the permutation test (Appendix 2), using SAS, is given in [4]. The results of the permutation test, with 999999 permutations can be seen in Figure 3 Below is SAS and R code for permutation tests:

Code 4.3. Two Tailed permutation test in SAS, using manually input groups

```
proc iml;
G1 = {/*SMU student data*/};
G2 = {/*Seattle U student data*/};
obsdiff = mean(G1) - mean(G2); /*difference in the means of the two data sets*/
print obsdiff;
call randseed(12345); /* set random number seed */
alldata = G1 // G2; /* stack data in a single vector */
N1 = nrow(G1); N = N1 + nrow(G2);
NRepl = 999999; /* number of permutations, I did ~ 1 million just because I thought the
nulldist = j(NRepl,1); /* allocate vector to hold results */
do k = 1 to NRepl;
x = sample(alldata, N, "WOR"); /* permute the data */
nulldist[k] = mean(x[1:N1]) - mean(x[(N1+1):N]); /* difference of means */
end;
title "Histogram of Null Distribution";
refline = "refline " + char(obsdiff) + " / axis=x lineattrs=(color=red);";/*build a nic
call Histogram(nulldist) other=refline;
pval = (1 + sum(abs(nulldist) >= abs(obsdiff))) / (NRepl+1); print pval;/*calculate the
/*https://blogs.sas.com/content/iml/2014/11/21/resampling-in-sas.html*/
```

Figure 4.0.3. Results of Permutation Tests



And some R code: In this test, the null hypothesis is that there is no difference between the mean amount of cash in a student's pocket in the two groups, while the alternative hypothesis is that there is a meaningful difference between the two[4]. The permutations were used to generate the null distribution of differences, and the red line shows where the experimental difference lies. Further calculation shows that the p value of the experimental mean was 0.149, meaning about 15% of the null distribution is greater than our mean[5]. With a 5 or 10 % confidence interval, we cannot reject the null hypothesis, and therefore we cannot say there is any difference between the two means. The SMU students and Seattle U students have more or less the same amount of cash in their pockets, the result of the study does not bear statistical inference. As for scope of inference, this was not a randomized experiment or random sample, and therefore we cannot make any causal inferences (there was no treatment applied, and we definitely cannot say going to SMU makes you have more or less money in your pocket than going to Seattle U), and we cannot make any inferences about the student bodies as a whole (population inferences). The sample is only representative of the students sampled, so we have very little scope of inference.

Code 4.4. Two Tailed permutation test in R, using manually input groups

```r
school1 <- rep('SMU', 16)
school2 <- rep('Seattle', 14)
school <- as.factor(c(school1, school2))
all.money <- data.frame(name=school, money=c(SMU, Seattle))

t.test(money ~ name, data=all.money)
number_of_permutations <- 1000
xbarholder <- numeric(0)
counter <- 0
observed_diff <- mean(subset(all.money, name == "SMU")\$money)-mean(subset(all.money, name == "Seattle")\$money)

set.seed(123)
for(i in 1:number_of_permutations)
{
  scramble <- sample(all.money\$money, 30)
  smu <- scramble[1:16]
  seattle <- scramble[17:30]
  diff <- mean(smu)-mean(seattle)
  xbarholder[i] <- diff
  if(abs(diff) > abs(observed_diff))
  counter <- counter + 1
}
hist(xbarholder, xlab='Permuted SMU - Seattle', main='Histogram of Permuted Mean Differences')
box()
pvalue <- counter / number_of_permutations
pvalue
observed_diff
```

Score: 15/15. Explanation: This receives full marks, 5 points for running the test, 5 points for the p value, and 5 points for mentioning the null and alternative hypotheses and getting the correct conclusion. The code is included in the Appendix.

Chapter 5

Unit 1 Lecture Slides

MSDS 6371: Lecture 1

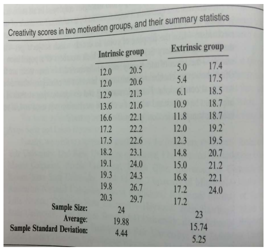
DRAWING STATISTICAL CONCLUSIONS
RANDOMIZED EXPERIMENTS V. OBSERVATIONAL STUDIES

RANDOM SAMPLES V. SELF-SELECTION

Symbols!

	Mean	Standard Deviation	Variance
Sample	$\bar{x}$	s	s^2
Population	μ	σ	σ^2

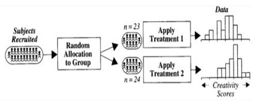
Creativity Scores: Intrinsic vs. Extrinsic Motivation



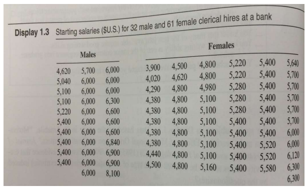
Creativity scores in two motivation groups, and their summary statistics

	Intrinsic group		Extrinsic group	
	12.0	20.5	5.0	17.4
	12.0	20.6	5.4	17.5
	12.9	21.3	6.1	18.5
	13.6	21.6	10.9	18.7
	16.6	22.1	11.8	18.7
	17.2	22.2	12.0	19.2
	17.5	22.6	12.3	19.5
	18.2	23.1	14.8	20.7
	19.1	24.0	15.0	21.2
	19.3	24.3	16.8	22.1
	19.8	26.7	17.2	24.0
	20.3	29.7	17.2	
Sample Size:	24		23	
Average:	19.88		15.74	
Sample Standard Deviation:	4.44		5.25	

Subjects volunteered for the study.
Then, treatments were randomly assigned.



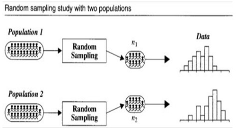
Starting Salaries: Female vs. Male



Display 1.3 Starting salaries ($U.S.) for 32 male and 61 female clerical hires at a bank

Males			Females					
4,620	5,700	6,000	3,900	4,500	4,800	5,220	5,400	5,640
5,040	6,000	6,000	4,020	4,620	4,800	5,220	5,400	5,700
5,100	6,000	6,000	4,290	4,800	4,980	5,280	5,400	5,700
5,100	6,000	6,300	4,380	4,800	5,100	5,280	5,400	5,700
5,220	6,000	6,600	4,380	4,800	5,100	5,280	5,400	5,700
5,400	6,000	6,600	4,380	4,800	5,100	5,400	5,400	5,700
5,400	6,000	6,600	4,380	4,800	5,100	5,400	5,400	6,000
5,400	6,000	6,840	4,380	4,800	5,100	5,400	5,520	6,000
5,400	6,000	6,900	4,440	4,800	5,100	5,400	5,520	6,120
5,400	6,000	6,900	4,500	4,800	5,160	5,400	5,580	6,300
	6,000	8,100						6,300

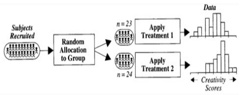
Subjects were NOT randomly chosen by the researcher (all employees at a bank were included), and the group assignments were not random either.
If a random sample of the employees had been used…



Random sampling study with two populations

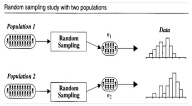
Types of Studies

Creativity Study



Randomized Experiment

Salary Study



Observational Study

Causal Inference: Randomized vs. Observational Study

- Causal inferences **can** be drawn from randomized experiments
- Causal inferences **cannot** be drawn from observational studies due to **CONFOUNDING**

CONFOUNDING VARIABLE: Related to both group membership and to the outcome

Example: Since 2000, the U.S. median wage…

- has overall increased about 1%
- has decreased for high school (or below) dropouts and high school graduates (no college)

- Is this a paradox?

No, more people are going to college.

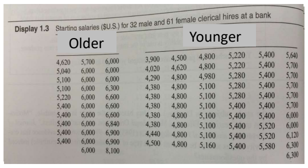
Causal Inference: Randomized vs. Observational Study

- Causal inferences **can** be drawn from randomized experiments
- Causal inferences **cannot** be drawn from observational studies due to **CONFOUNDING**

What are some possible confounding variables in the gender/salary study?

In the starting salaries study, maybe males have

- more education
- more seniority
- more age (older)
- more willingness to negotiate starting salary



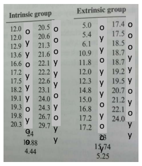
Display 1.3 Starting salaries ($U.S.) for 32 male and 61 female clerical hires at a bank

Older

4,620	5,700	6,000
5,040	6,000	6,000
5,100	6,000	6,000
5,100	6,000	6,300
5,220	6,000	6,600
5,400	6,000	6,600
5,400	6,000	6,600
5,400	6,000	6,840
5,400	6,000	6,900
5,400	6,000	6,900
	6,000	8,100

Younger

3,900	4,500	4,800	5,220	5,400	5,640
4,020	4,620	4,800	5,220	5,400	5,700
4,290	4,800	4,980	5,280	5,400	5,700
4,380	4,800	5,100	5,280	5,400	5,700
4,380	4,800	5,100	5,280	5,400	5,700
4,380	4,800	5,100	5,400	5,400	5,700
4,380	4,800	5,100	5,400	5,400	6,000
4,380	4,800	5,100	5,400	5,520	6,000
4,440	4,800	5,100	5,400	5,520	6,120
4,500	4,800	5,160	5,400	5,580	6,300
					6,300



Intrinsic group		Extrinsic group	
12.0	20.5	5.0	17.4
12.0	20.6	5.4	17.5
12.9	21.3	6.1	18.5
13.6	21.6	10.9	18.7
16.6	22.1	11.8	18.7
17.2	22.2	12.0	19.2
17.5	22.6	12.3	19.5
18.2	23.1	14.8	20.7
19.1	24.0	15.0	21.2
19.3	24.3	16.8	22.1
19.8	26.7	17.2	24.0
20.3	29.7	17.2	
24		23	
19.88		15.74	
4.44		5.25	

In a randomized experiment, variables like age are also randomly distributed to each group, removing the confounding effect.

Why do an observational study?

- Establishing causation not always the goal
  - Predict whether or not an email is spam
- Randomization may not be ethical
  - Assign subjects of a clinical trial of a cancer drug to treatment or placebo
- May be arguable scientifically that a confounder is "unlikely"
  - 6 month smoking ban in Helena, MT coinciding with 40% reduction in heart attacks
- Might have an incidentally observed dataset
  - Walmart collects petabytes of data/day. Should this data be discarded because it is observational?

Inference to Populations: Random Sample vs. Self-Selection

- Inference to populations **can** be drawn from a **RANDOM SAMPLE** FROM THAT POPULATION.
- Inference to populations **cannot** be drawn if units are self-selected. In this creativity example, inference can only be drawn to the subjects in the sample that was taken.

RANDOM SAMPLE: Experimental units selected via a "chance mechanism" from a well defined population

Example: call randomly selected phone numbers for a survey.

- What is the population from which the sample is taken? If drawing from a physical phone book, is it the people who live in the city?
- Would this sampling method result in inferences to different populations if it were used in 1950? 1990? Present day?

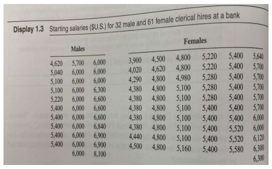
SIMPLE RANDOM SAMPLE: Every subset of size *n* is equally likely

Example: I'll assign everyone in this class a random integer 17, 200, -3, 472, … and survey the *n* people (units) with smallest numbers

Inference to Populations: Random Sample vs. Self-Selection

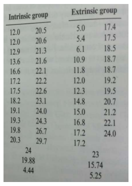
- Inference to populations **can** be drawn from a **RANDOM SAMPLE**
- Inference to populations **cannot** be drawn if units are self-selected
- **WHICH OF THE STUDIES USES RANDOM SAMPLING?**

- Neither study uses random sampling
  - Creativity study: units are volunteers
  - Bank study: units are the entire staff
- No inference about a larger population is possible
- Does not mean the results are not interesting or compelling!



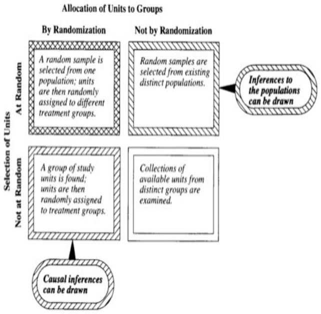
Display 1.3 Starting salaries ($U.S.) for 32 male and 61 female clerical hires at a bank

Males			Females					
4,620	5,700	6,000	3,900	4,500	4,800	5,220	5,400	5,640
5,040	6,000	6,000	4,020	4,620	4,800	5,220	5,400	5,700
5,100	6,000	6,000	4,290	4,800	4,980	5,280	5,400	5,700
5,100	6,000	6,300	4,380	4,800	5,100	5,280	5,400	5,700
5,220	6,000	6,600	4,380	4,800	5,100	5,280	5,400	5,700
5,400	6,000	6,600	4,380	4,800	5,100	5,400	5,400	5,700
5,400	6,000	6,600	4,380	4,800	5,100	5,400	5,400	6,000
5,400	6,000	6,840	4,380	4,800	5,100	5,400	5,520	6,000
5,400	6,000	6,900	4,440	4,800	5,100	5,400	5,520	6,120
5,400	6,000	6,900	4,500	4,800	5,160	5,400	5,580	6,300
	6,000	8,100						6,300

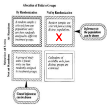


Intrinsic group		Extrinsic group	
12.0	20.5	5.0	17.4
12.0	20.6	5.4	17.5
12.9	21.3	6.1	18.5
13.6	21.6	10.9	18.7
16.6	22.1	11.8	18.7
17.2	22.2	12.0	19.2
17.5	22.6	12.3	19.5
18.2	23.1	14.8	20.7
19.1	24.0	15.0	21.2
19.3	24.3	16.8	22.1
19.8	26.7	17.2	24.0
20.3	29.7	17.2	
24		23	
19.88		15.74	
4.44		5.25	

Statistical Inferences Permitted by Study Design



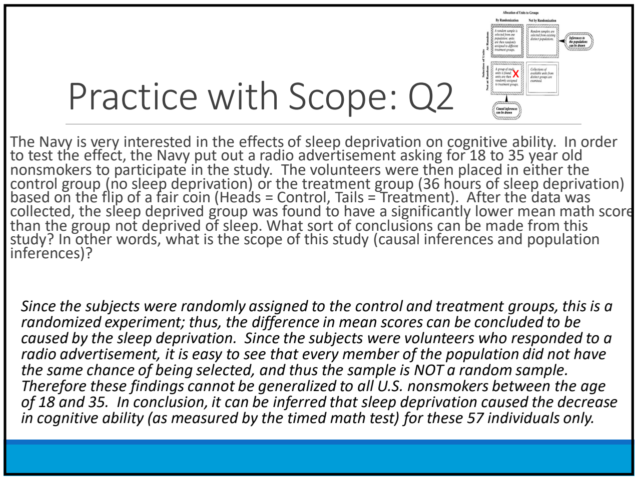
Practice with Scope: Q1



A particular study focused on high school freshman and seniors and their GPAs in a required economics class. The study consisted of enumerating every freshman and senior in the school and randomly selecting them from that sampling frame. Their scores in the economics class were then recorded, and a hypothesis test for the difference of means was conducted. The seniors were found to have a significantly greater mean score in the class than the freshman. What sort of conclusions can be made from this study? In other words, what is the scope of this study? In this class, scope typically constitutes both the causal inferences and populations inferences.

Since the subjects cannot be randomly assigned to be freshman or seniors, this is an observational study, and thus the difference in mean scores is only associated with the freshman / senior status. We can't tell if the class (freshman or senior) caused the difference or not.

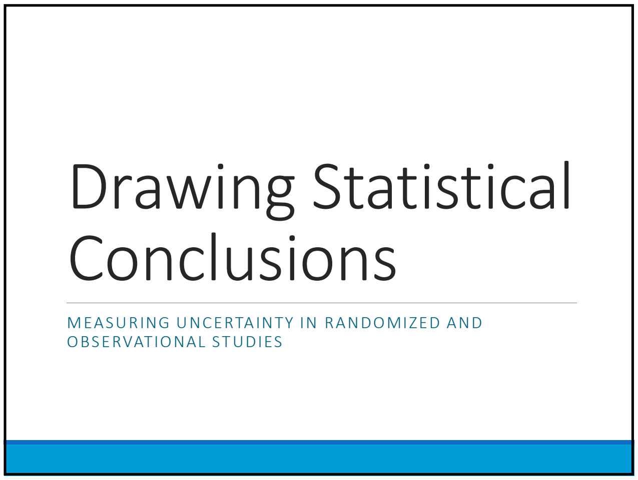
The sample was a random sample from the school; therefore, these findings can be generalized to all freshman and seniors in the school. In conclusion, it can be inferred that the mean economics score of the seniors in the school is greater than that of the freshman although the cause of this difference cannot be determined from this study.



Practice with Scope: Q2

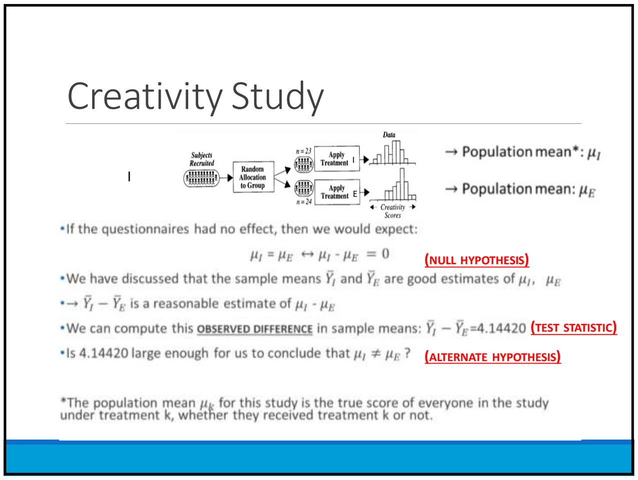
The Navy is very interested in the effects of sleep deprivation on cognitive ability. In order to test the effect, the Navy put out a radio advertisement asking for 18 to 35 year old nonsmokers to participate in the study. The volunteers were then placed in either the control group (no sleep deprivation) or the treatment group (36 hours of sleep deprivation) based on the flip of a fair coin (Heads = Control, Tails = Treatment). After the data was collected, the sleep deprived group was found to have a significantly lower mean math score than the group not deprived of sleep. What sort of conclusions can be made from this study? In other words, what is the scope of this study (causal inferences and population inferences)?

Since the subjects were randomly assigned to the control and treatment groups, this is a randomized experiment; thus, the difference in mean scores can be concluded to be caused by the sleep deprivation. Since the subjects were volunteers who responded to a radio advertisement, it is easy to see that every member of the population did not have the same chance of being selected, and thus the sample is NOT a random sample. Therefore these findings cannot be generalized to all U.S. nonsmokers between the age of 18 and 35. In conclusion, it can be inferred that sleep deprivation caused the decrease in cognitive ability (as measured by the timed math test) for these 57 individuals only.



Drawing Statistical Conclusions

MEASURING UNCERTAINTY IN RANDOMIZED AND OBSERVATIONAL STUDIES



Creativity Study

→ Population mean*: μ_I

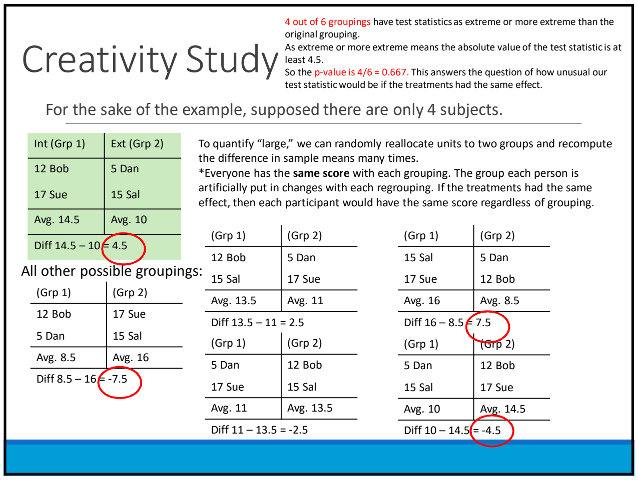
→ Population mean: μ_E

- If the questionnaires had no effect, then we would expect:

$$\mu_I = \mu_E \leftrightarrow \mu_I - \mu_E = 0$$ **(NULL HYPOTHESIS)**

- We have discussed that the sample means $\bar{Y}_I$ and $\bar{Y}_E$ are good estimates of μ_I, μ_E
- → $\bar{Y}_I - \bar{Y}_E$ is a reasonable estimate of μ_I - μ_E
- We can compute this **OBSERVED DIFFERENCE** in sample means: $\bar{Y}_I - \bar{Y}_E$=4.14420 **(TEST STATISTIC)**
- Is 4.14420 large enough for us to conclude that $\mu_I \neq \mu_E$? **(ALTERNATE HYPOTHESIS)**

*The population mean μ_k for this study is the true score of everyone in the study under treatment k, whether they received treatment k or not.



Creativity Study

4 out of 6 groupings have test statistics as extreme or more extreme than the original grouping.
As extreme or more extreme means the absolute value of the test statistic is at least 4.5.
So the p-value is 4/6 = 0.667. This answers the question of how unusual our test statistic would be if the treatments had the same effect.

For the sake of the example, supposed there are only 4 subjects.

Int (Grp 1)	Ext (Grp 2)
12 Bob 17 Sue	5 Dan 15 Sal
Avg. 14.5	Avg. 10
Diff 14.5 – 10 = 4.5	

To quantify "large," we can randomly reallocate units to two groups and recompute the difference in sample means many times.
*Everyone has the **same score** with each grouping. The group each person is artificially put in changes with each regrouping. If the treatments had the same effect, then each participant would have the same score regardless of grouping.

All other possible groupings:

(Grp 1)	(Grp 2)
12 Bob 5 Dan	17 Sue 15 Sal
Avg. 8.5	Avg. 16
Diff 8.5 – 16 = -7.5	

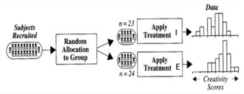
(Grp 1)	(Grp 2)
12 Bob 15 Sal	5 Dan 17 Sue
Avg. 13.5	Avg. 11
Diff 13.5 – 11 = 2.5	

(Grp 1)	(Grp 2)
5 Dan 17 Sue	12 Bob 15 Sal
Avg. 11	Avg. 13.5
Diff 11 – 13.5 = -2.5	

(Grp 1)	(Grp 2)
15 Sal 17 Sue	5 Dan 12 Bob
Avg. 16	Avg. 8.5
Diff 16 – 8.5 = 7.5	

(Grp 1)	(Grp 2)
5 Dan 15 Sal	12 Bob 17 Sue
Avg. 10	Avg. 14.5
Diff 10 – 14.5 = -4.5	

Creativity Study: all 47 subjects



→ Population mean: μ_I

→ Population mean: μ_E

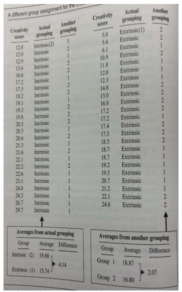
- To quantify “large,” we can randomly reallocate units to two groups and recompute the difference in sample means many times
- We say that a recomputed difference is <u>**MORE EXTREME (OR AS EXTREME)**</u> provided

  abs(recomputed difference) ≥abs($\bar{Y}_I - \bar{Y}_E$)
- Suppose that $\frac{number\ of\ more\ extreme\ recomputed\ differences}{total\ number\ of\ random\ reallocations} = p - value$ <u>**(P-VALUE)**</u>
- If *p-value* is very small (say 0.01), this provides evidence that the intrinsic/extrinsic group result would be very unusual if the questionnaire had no effect
- If *p-value* is very big (say 0.2), this provides little evidence that the intrinsic/extrinsic group result would be very unusual if the questionnaire had no effect

Creativity Study:
Testing the Hypothesis

Number of random regroupings: 1.6 x 10^{13}

Half a year with a computer that can perform a million calculations per second!


A different group assignment for the cre

Creativity score	Actual grouping	Another grouping
12.0	Intrinsic(2)	1
12.0	Intrinsic	2
12.9	Intrinsic	1
13.6	Intrinsic	2
16.6	Intrinsic	2
17.2	Intrinsic	1
17.5	Intrinsic	2
18.2	Intrinsic	2
19.1	Intrinsic	1
19.3	Intrinsic	2
19.8	Intrinsic	2
20.3	Intrinsic	2
20.5	Intrinsic	1
20.6	Intrinsic	2
21.3	Intrinsic	1
21.6	Intrinsic	2
22.1	Intrinsic	1
22.2	Intrinsic	2
22.6	Intrinsic	1
23.1	Intrinsic	1
24.0	Intrinsic	1
24.3	Intrinsic	1
26.7	Intrinsic	1
29.7	Intrinsic	1

Creativity score	Actual grouping	Another grouping
5.0	Extrinsic(1)	2
5.4	Extrinsic	2
6.1	Extrinsic	1
10.9	Extrinsic	2
11.8	Extrinsic	1
12.0	Extrinsic	1
12.3	Extrinsic	1
14.8	Extrinsic	2
15.0	Extrinsic	2
16.8	Extrinsic	2
17.2	Extrinsic	2
17.2	Extrinsic	1
17.4	Extrinsic	2
17.5	Extrinsic	2
18.5	Extrinsic	2
18.7	Extrinsic	1
18.7	Extrinsic	1
19.2	Extrinsic	1
19.5	Extrinsic	1
20.7	Extrinsic	1
21.2	Extrinsic	1
22.1	Extrinsic	2
24.0	Extrinsic	2

Averages from actual grouping

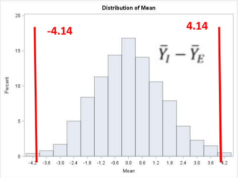
Group	Average	Difference
Intrinsic (2)	19.88	4.14
Extrinsic (1)	15.74	

Averages from another grouping

Group	Average	Difference
Group 1	18.87	2.07
Group 2	16.80	

$H_{0:}\ \mu_I - \mu_E = 0$
$H_{A:}\ \mu_I - \mu_E \neq 0$

1000 different groupings (relabelings)*



*Everyone has the **same score** with each grouping. What group each person is artificially put in changes with each regrouping. If the treatments had the same effect, then each participant would have the same score regardless of grouping.

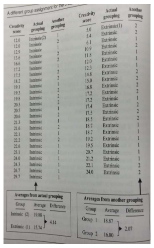
Creativity Study

(go to SAS code)

The TTEST Procedure
Variable: score

treatment	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		19.8833	18.0087	21.7580	4.4395	3.4504	6.2276
1		15.7391	13.4677	18.0105	5.2526	4.0623	7.4343
Diff (1-2)	Pooled	4.1442	1.2914	6.9970	4.8541	4.0261	6.1138
Diff (1-2)	Satterthwaite	4.1442	1.2776	7.0108			

Creativity Study


A different group assignment for the cre

Creativity score	Actual grouping	Another grouping
12.0	Intrinsic(2)	1
12.0	Intrinsic	2
12.9	Intrinsic	1
13.6	Intrinsic	2
16.6	Intrinsic	2
17.2	Intrinsic	1
17.5	Intrinsic	2
18.2	Intrinsic	2
19.1	Intrinsic	1
19.3	Intrinsic	2
19.8	Intrinsic	2
20.3	Intrinsic	2
20.5	Intrinsic	1
20.6	Intrinsic	2
21.3	Intrinsic	1
21.6	Intrinsic	2
22.1	Intrinsic	1
22.2	Intrinsic	2
22.6	Intrinsic	1
23.1	Intrinsic	1
24.0	Intrinsic	1
24.3	Intrinsic	1
26.7	Intrinsic	1
29.7	Intrinsic	1

Creativity score	Actual grouping	Another grouping
5.0	Extrinsic(1)	2
5.4	Extrinsic	2
6.1	Extrinsic	1
10.9	Extrinsic	2
11.8	Extrinsic	1
12.0	Extrinsic	1
12.3	Extrinsic	1
14.8	Extrinsic	2
15.0	Extrinsic	2
16.8	Extrinsic	2
17.2	Extrinsic	2
17.2	Extrinsic	1
17.4	Extrinsic	2
17.5	Extrinsic	2
18.5	Extrinsic	2
18.7	Extrinsic	1
18.7	Extrinsic	1
19.2	Extrinsic	1
19.5	Extrinsic	1
20.7	Extrinsic	1
21.2	Extrinsic	1
22.1	Extrinsic	2
24.0	Extrinsic	2

Averages from actual grouping

Group	Average	Difference
Intrinsic (2)	19.88	4.14
Extrinsic (1)	15.74	

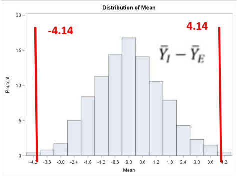
Averages from another grouping

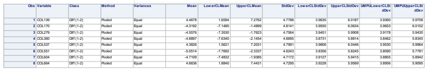
Group	Average	Difference
Group 1	18.87	2.07
Group 2	16.80	

$H_{0:}\ \mu_I - \mu_E = 0$
$H_{A:}\ \mu_I - \mu_E \neq 0$

1000 different groupings (relabelings)

P-value = 8/1000
= 0.008





Obs	Variable	Class	Method	Variances	Mean	LowerCLMean	UpperCLMean	StdDev	LowerCLStdDev	UpperCLStdDev	UMPULowerCLStdDev	UMPUUpperCLStdDev
1	COL139	Diff (1-2)	Pooled	Equal	4.4678	1.6594	7.2762	4.7786	3.9635	6.0187	3.9360	5.9708
2	COL170	Diff (1-2)	Pooled	Equal	-4.3192	-7.1485	-1.4899	4.8141	3.9930	6.0634	3.9653	6.0152
3	COL279	Diff (1-2)	Pooled	Equal	-4.5576	-7.3530	-1.7623	4.7564	3.9451	5.9908	3.9178	5.9430
4	COL360	Diff (1-2)	Pooled	Equal	-4.8897	-7.6340	-2.1454	4.6695	3.8731	5.8814	3.8462	5.8345
5	COL537	Diff (1-2)	Pooled	Equal	4.3826	1.5621	7.2031	4.7991	3.9806	6.0446	3.9530	5.9964
6	COL551	Diff (1-2)	Pooled	Equal	-5.0514	-7.7692	-2.3337	4.6243	3.8356	5.8245	3.8090	5.7781
7	COL604	Diff (1-2)	Pooled	Equal	-4.7109	-7.4832	-1.9385	4.7172	3.9127	5.9415	3.8855	5.8942
8	COL664	Diff (1-2)	Pooled	Equal	4.6636	1.8840	7.4431	4.7295	3.9228	5.9569	3.8956	5.9095

There is strong evidence to suggest that the mean score of those who receive intrinsic motivation is not equal to those who receive the extrinsic motivation (p-value = .008). The burden to reject the null hypothesis is lower under a one-sided test, so we can say that the evidence supports the claim that the intrinsic mean is higher than the extrinsic mean.
Since this was a randomized experiment, we can conclude that the intrinsic motivation caused this increase. In addition, since these were volunteers, this inference can only be assumed to apply to these 47 subjects, although the findings are very intriguing.

From Randomized to Observational Studies

- In the Creativity study, the Intrinsic/Extrinsic groups were randomly assigned to subjects
- This motivated comparing the observed difference to re-randomized difference to test a hypothesis about the questionnaire having no effect
- This is known as a **RANDOMIZATION TEST**

- In observational studies, the groups are not randomly assigned
- Though not technically the same test, we can still apply exactly the same re-randomization idea to observational data
- However, now it is called a **PERMUTATION TEST**

Appendix

Age Discrimination

In the United States, it is illegal to discriminate against people based on various attributes. One such attribute is age. An active lawsuit, filed August 30, 2011, in the Los Angeles District Office is a case against the American Samoa Government for systematic age discrimination by preferentially firing older workers.

Is there evidence for age discrimination in this study?

Data sampled at random from all American Samoa government workers:

Fired

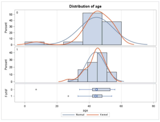
34 37 37 38 41 42 43 44 44 45 45 45 46 48 49 53 53 54 54 55 56

Not fired

27 33 36 37 38 38 39 42 42 43 43 44 44 44 45 45 45 45 46 46 47 47 48 48 49 49 51 51 52 54

Age Discrimination (Two Sided)

Fired
34 37 37 38 41 42 43 44 44 45 45
45 46 48 49 53 53 54 54 55 56
Not fired
27 33 36 37 38 38 39 42 42 43 43 44
44 44 45 45 45 45 46 46 47 47 48 48
49 49 51 51 52 54



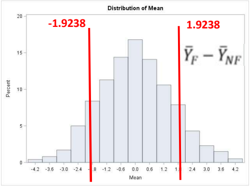
$\bar{Y}_F - \bar{Y}_{NF} = 45.8571 - 43.9333 = 1.9238$

$H_{0:}\ \mu_F - \mu_{NF} = 0$

$H_{A:}\ \mu_F - \mu_{NF} \neq 0$

P-value =204/1000
= 0.204

1000 different
groupings
(relabelings)



200	COL991	Diff (1-2)	Pooled	Equal	3.0714	0.0988	6.0439	5.0579	4.1952	6.3705	4.1661	6.3198
201	COL994	Diff (1-2)	Pooled	Equal	2.1518	-0.8927	5.1963	5.1803	4.2968	6.5248	4.2670	6.4728
202	COL995	Diff (1-2)	Pooled	Equal	-3.1357	-6.1022	-0.1692	5.0476	4.1867	6.3576	4.1576	6.3069
203	COL998	Diff (1-2)	Pooled	Equal	-3.5444	-6.4691	-0.6197	4.9765	4.1277	6.2680	4.0991	6.2181
204	COL1000	Diff (1-2)	Pooled	Equal	-2.5652	-5.5807	0.4503	5.1310	4.2558	6.4626	4.2263	6.4111

There is not sufficient evidence to suggest that the mean age of those who were fired is different from the mean age of those who were not fired (p-value = 0.204). The p-value is so high that even the null hypothesis of a one-sided test cannot be rejected. (There is insufficient evident to claim that the mean age of fired employees is greater than that of not fired employees.)
Since this was a random sample of government employees in Samoa, we can generalize this inference to all government-employed people in Samoa.
Note: since we FTR (fail to reject) Ho, there is no need to discuss causation or association.

Part II

Inferences Using the t-distribution

Chapter 6

Problem 1: A one sample t test

Question 1

The world's smallest mammal is the bumblebee bat, also known as the Kitti's hog nosed bat. Such bats are roughly the size of a large bumblebee! Listed below are weights (in grams) from a sample of these bats. Test the claim that these bats come from the same population having a mean weight equal to 1.8 g. (Beware: This data is NOT the same as in the lecture slides!) Sample: 1.7 1.6 1.5 2.0 2.3 1.6 1.6 1.8 1.5 1.7 1.2 1.4 1.6 1.6 1.6

1. Perform a complete analysis using SAS. Use the six step hypothesis test with a conclusion that includes a statistical conclusion, a confidence interval and a scope of inference (as best as can be done with the information above … there are many correct answers given the vagueness of the description of the sampling mechanism.)
2. Inspect and run this R Code and compare the results (t statistic, p-value and confidence interval) to those you found in SAS. To run the code, simply copy and paste the below code into R.

Code 6.1. One sample t test in R with manual data input

```
        sample = c(1.7, 1.6, 1.5, 2.0, 2.3, 1.6, 1.6, 1.8, 1.5, 1.7, 1.2, 1.4, 1.6, 1.6, 1.6)
        t.test(x=sample, mu = 1.8, conf.int = "TRUE", alternative = "two.sided")
```

Answer

6.1 Complete Analysis

Hypothesis definition

$$H_0 : \mu = 1.8 \tag{6.1.1}$$
$$H_1 : \mu \neq 1.8 \tag{6.1.2}$$

Identification of a critical value and drawing a shaded t distribution

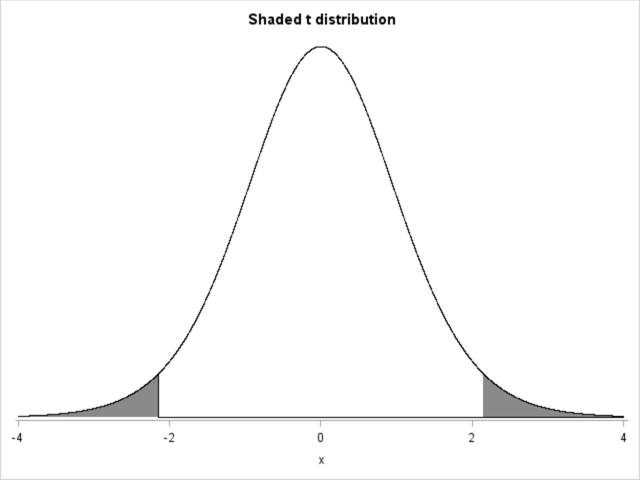
We have that $n = 15 \rightarrow df = n - 1 = 14,\ \ \alpha = 0.05$. We input this into SAS and get our lovely shaded distribution and critical value with the following code: This gives us a critical t value of ±2.14479, as seen in the following figures:

Figure 6.1.1. Critical t value

Obs	p
1	2.14479

Code 6.2. Critical value and two sided shaded t distribution using SAS

```
data critval;
p = quantile("T",.975,14); /*two sided test*/;
proc print data=critval;
run;

data pdf;
do x = -4 to 4 by .001;
pdf = pdf("T", x, 14);
if x <= quantile("T",.025,14) then lower = pdf;
else lower = 0;
if x >= quantile("T",.975,14) then upper = pdf;
else upper = 0;
output;
end;
run;
title 'Shaded t distribution';
proc sgplot data=pdf noautolegend noborder;
yaxis display=none;
band x = x lower = lower upper = upper / fillattrs=(color=gray8a);
series x = x y = pdf / lineattrs = (color = black);
series x = x y = lower / lineattrs = (color = black);
run;
```



Value of Test Statistic

The t statistic was calculated using the following SAS code

Code 6.3. One sample t test in SAS

```
proc ttest data=bats h0=1.8
sides=2 alpha=0.05;
run;
```

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \approx \frac{1.65 - 1.8}{\frac{0.25}{15}} = -2.35$$

P value

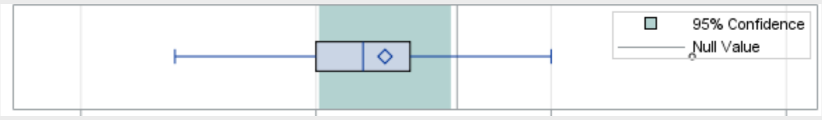
This gives us a p-value of $p = 0.0342$

Assessment of the Hypothesis test

From here we can see that $p = .0342 < \alpha = .05$, indicating that we REJECT the null hypothesis, which claims that $\mu = 1.8$

Conclusion and scope of inference

We cannot say that this sample of bats comes from a population with a mean weight of 1.8 grams (p value = 0.0242 from a two sided t test). Below is a graph produced with the code from step 4 which shoes a 95% confidence interval on the distribution of the data (green) vs the null hypothesis(gray bar)



The mean of 1.8 lies outside the reasonable range of the data from the sample, and as our hypothesis test showed, vice versa is also true. We cannot say that our sample of bats has a mean weight of 1.8, and it is difficult to say that it came from a population of mean 1.8. However, we cannot make any conclusions about the population this sample came from, because it is not a random sample (we also clearly cant make any causal inferences), We only know, with 95% confidence, that our sample does not have a mean of 1.8 grams, and that is about all we can say.

Some R code

Code 6.4. one sample t test in r

```
        sample <- c(1.7, 1.6, 1.5, 2.0, 2.3, 1.6, 1.6,
        1.8, 1.5, 1.7, 1.2, 1.4, 1.6, 1.6, 1.6)
        t.test(x=sample, mu = 1.8,
        conf.int = "TRUE", alternative = "two.sided")
```

Chapter 7

Problem 2: Two sample one sided t test

Question

2. In the United States, it is illegal to discriminate against people based on various attributes. One example is age. An active lawsuit, filed August 30, 2011, in the Los Angeles District Office is a case against the American Samoa Government for systematic age discrimination by preferentially firing older workers. Though the data and details are currently sealed, suppose that a random sample of the ages of fired and not fired people in the American Samoa Government are listed below: Fired 34 37 37 38 41 42 43 44 44 45 45 45 46 48 49 53 53 54 54 55 56 Not fired 27 33 36 37 38 38 39 42 42 43 43 44 44 44 45 45 45 45 46 46 47 47 48 48 49 49 51 51 52 54

a. Perform a permutation test to test the claim that there is age discrimination. Provide the Ho and Ha, the p-value, and full statistical conclusion, including the scope (inference on population and causal inference). Note: this was an example in Live Session 1. You may start from scratch or use the sample code and PowerPoints from Live Session 1.

b. Now run a two sample t-test appropriate for this scientific problem. (Use SAS.) (Note: we may not have talked much about a two-sided versus a one-sided test. If you would like to read the discussion on pg. 44 (Statistical Sleuth), you can run a one-sided test if it seems appropriate. Otherwise, just run a two-sided test as in class. There are also examples in the Statistics Bridge Course.) Be sure to include all six steps, a statistical conclusion, and scope of inference.

c. Compare this p-value to the randomized p-value found in the previous sub-question.

d. The jury wants to see a range of plausible values for the difference in means between the fired and not fired groups. Provide them with a confidence interval for the difference of means and an interpretation.

f. Inspect and run this R Code and compare the results (t statistic, p-value, and confidence interval) to those you found in SAS. To run the code, simply copy and paste the code below into R.

Answers

7.1 Permutation test

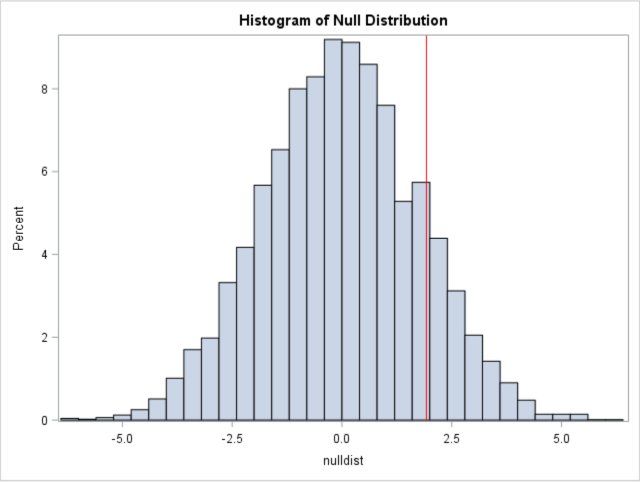
First, a permutation test is ran using $n = 9999$, using the code I wrote in homework one, inspired by [2]. The code used to run the permutation test is shown below: In this scenario, we have that:

$$\begin{aligned} H_0 &: \mu_f - \mu_{uf} \leq 0 \\ H_1 &: \mu_f - \mu_{uf} > 0 \end{aligned}$$

where the null hypothesis is that the average age of the unfired individuals is the same as the average age of the fired individuals, and the alternative is that the average age of the individuals who were fired is higher. The results of the permutation test are as follows:

Code 7.1. A one sided permutation test in SAS

```
obsdiff = mean(G1) - mean(G2); /*G1 and G2 represent the two groups*/
print obsdiff;
call randseed(12345);             /* set random number seed */
alldata = G1 // G2;             /* stack data in a single vector */
N1 = nrow(G1);
N = N1 + nrow(G2);
NRepl = 9999;               /* number of permutations */
nulldist = j(NRepl,1);    /* allocate vector to hold results */
do k = 1 to NRepl;
x = sample(alldata, N, "WOR");      /* permute the data */
nulldist[k] = mean(x[1:N1]) - mean(x[(N1+1):N]);
/* difference of means */
end;
title "Histogram of Null Distribution";
refline = "refline " + char(obsdiff) + " / axis=x lineattrs=(color=red);";
call Histogram(nulldist) other=refline;
pval = (1 + sum(abs(nulldist) >= (obsdiff))) / (NRepl+1);
print pval;
```



In the above figure, the red line represents the mean of the difference between the two samples, and the rest of the bars represent our null distribution. SAS tells us that the P-value is 0.2812, meaning 28.12 percent of the null distribution is greater than our sample mean. Therefore, with a 5%, or even a 10% confidence interval, we cannot reject the null hypothesis. We cannot say whether or not there was age discrimination in the firing of workers with the given sample. With this procedure, we can make generalizations about the population, and generalize about all of the government-employed people in Samoa, as we did a random sample, however, we cannot make causal inferences, as there may be confounding variables in the system, and we did not run a randomized experiment. There is also no need to discuss causal problems, because we failed to reject the null hypothesis.

7.2 Two sample T test, full analysis

This time we will conduct a t test on the two data sets to determine whether age discrimination occured or not. Because we believe the older workers may have been fired, we are going to perform a one sided t-test.

Hypothesis definition

First we construct our hypotheses:

$$
\begin{aligned}
H_0 :& \mu_f - \mu_{uf} \leq 0 \\
H_1 :& \mu_f - \mu_{uf} > 0
\end{aligned}
$$

critval and distribution

Next we draw and shade our distribution:

In a two sample t-test, we have that:

$$df = n_f + n_{nf} - 2$$

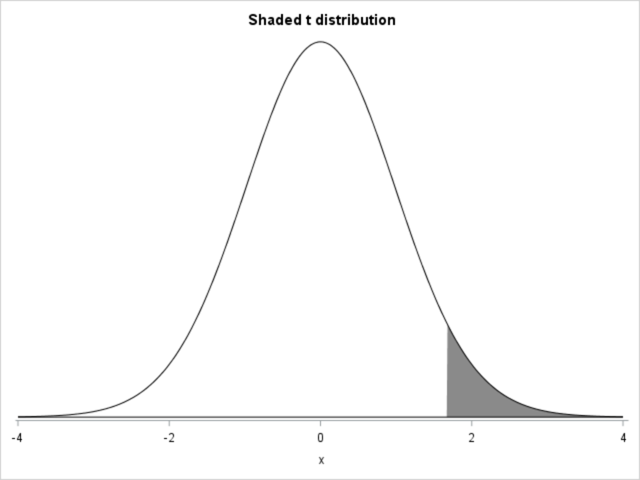
where in our case, $df = 21 + 30 - 2 = 49,\ \alpha = 0.05$

Now we input this information into SAS to draw our distribution[1]:

Code 7.2. One sided shaded t distribution in SAS and Critval

```
data pdf;
do x = -4 to 4 by .01;
pdf = pdf("T", x, 49);
lower = 0;
if x >= quantile("T",0.95,49) then upper = pdf;/*one sided*/     else upper = 0;
output;
end;
run;
title 'Shaded t distribution';
proc sgplot data=pdf noautolegend noborder;
yaxis display=none;
band x = x
lower = lower
upper = upper / fillattrs=(color=gray8a);
series x = x y = pdf / lineattrs = (color = black);
series x = x y = lower / lineattrs = (color = black);
run;

data critval;
p = quantile("T",.95,49); /*one sided test*/;
proc print data=critval;
run;
```

Giving us this lovely graph:



Next we find a number for the critical value, using the same code as problem 1:

Obs	p
1	1.67655

This gives us a critical t value of 1.67655.

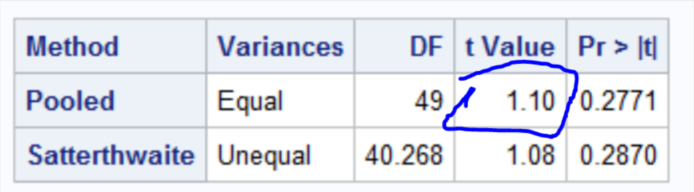
Calculation of the T statistic

Next we calculate our two sample t statistic using SAS:

Code 7.3. Two sample t test using SAS

```
proc ttest data=samoa
alpha=.05 test=diff
sides=U;
class fired;
var age;
run;
```

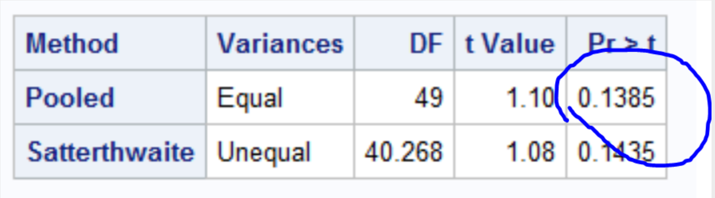
Which tells us that our t statistic is 1.10



Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	49	1.10	0.2771
Satterthwaite	Unequal	40.268	1.08	0.2870

P value

With the code from the previous step, we also see the p value:



Method	Variances	DF	t Value	Pr > t
Pooled	Equal	49	1.10	0.1385
Satterthwaite	Unequal	40.268	1.08	0.1435

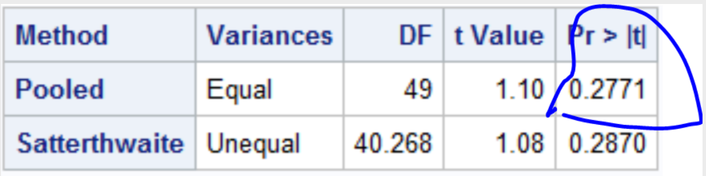
$p = 0.1385$

hypothesis assement

$p = 0.1385 > \alpha = 0.05$ for the one tailed hypothesis test, indicating that we CANNOT REJECT the null hypothesis

conclusion

The p value for the t test was about half of the p value for the random test, I believe this is because I ran a one-sided t test. It is interesting to note that if you do a two sided t-test in SAS, you get roughly the same value for p as in the permutation test:



Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	49	1.10	0.2771
Satterthwaite	Unequal	40.268	1.08	0.2870

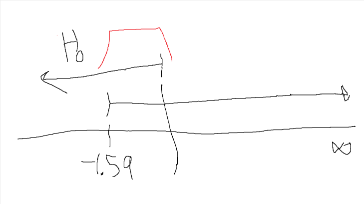
This means that maybe a permutation test is a good estimator of the two-sided t-test.

We cannot reject the null hypothesis, meaning we cannot say that older workers were fired from the samoan government. Note that we used a one tailed hypothesis test in this scenario, as we wanted to deternine if the fired group was OLDER than the nonfired group. As a result of this test, we cannot say that the fired group was older than the unfired group, and since this sample was random, we can say the same thing about the entire samoan government. However, we cannot make causal inferences and there is no need to because we did not reject the null hypothesis

We can provide a lot of confidence intervals for the jury. I think the most telling is the one sided confidence interval, which would tell us what difference in the means constitutes age discrimination. This was produced using the following SAS code:

```
proc ttest data=samoa
alpha=.05 test=diff
sides=U; /*an upper tailed test*/
class fired;
var age;
run;
```

which gives us a confidence interval of $[-1.0107, \infty)$. This confidence interval represents the upper difference of means at a 95% confidence level. We can interpret this as follows: if the confidence interval contains the null hypothesis, then we cannot reject it. However if it does not contain the null hypothesis, we must reject it. As we can see in this beautifully drawn figure, the null hypothesis, $\mu_f - \mu_{nf} \leq 0$ is contained within our CI:



. This means we cannot reject the null hypothesis, we cannot say there was age discrimination. It is plausible that the mean differnence of the entire population of samoan government employees is less than or equal to zero, as it is within the 95% confidence interval, which means we cannot, as objective jurors, claim there was age discrimination.

Incorrect calculations

The pooled sample standard deviation, s_p, is defined as

$$s_p^2 = \frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{\sum_{i=1}^{k}(n_i - 1)}$$

which for us is:

$$s_p = \sqrt{\frac{(21-1)(6.5214)^2 + (30-1)(5.8835)^2}{20+29}} = 6.152$$

The equation for standard error in the difference of means is given as

$$\sigma_{\bar{x_1} - \bar{x_2}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Which gives us that

$$\sigma_{\bar{x_1} - \bar{x_2}} = \sqrt{\frac{6.5214^2}{21} + \frac{5.8835^2}{30}} = 1.811$$

7.3 Rcode

The following code (supplied in the homework) was put into R: returning this:

Code 7.4. two sample t test in R

```
      Fired = c(34, 37, 37, 38, 41, 42, 43,
      44, 44, 45, 45, 45, 46, 48, 49, 53,
      53, 54, 54, 55, 56)
      Not_fired = c(27, 33, 36, 37, 38, 38,
      39, 42, 42, 43, 43, 44, 44, 44, 45,
      45, 45, 45, 46, 46, 47, 47, 48, 48,
      49, 49, 51, 51, 52, 54)
      t.test(x = Fired, y = Not_fired, conf.int = .95, var.equal = TRUE, alternative = "greater
          ")
```

```
  Two Sample t-test
  data:  Fired and Not_fired
  t = 1.0991,
  df = 49,
  p-value = 0.1385 alternative hypothesis: true difference in means is greater than 0
  95 percent confidence interval:  -1.010728       Inf sample estimates: mean of x mean of y   45.85714  43.93333
```

The results are near identical, I cannot tell which one is better but I imagine R is more accurate as well, but just a very small difference between the results in all regards . The var.Equal statement is important because it uses the pooled test.

Chapter 8

Problem 3: two sample two sided t test

Question

3. In the last homework, it was mentioned that a Business Stats professor here at SMU polled his class and asked students them how much money (cash) they had in their pockets at that very moment. The idea was that we wanted to see if there was evidence that those in charge of the vending machines should include the expensive bill / coin acceptor or if it should just have the credit card reader. However, another professor from Seattle University was asked to poll her class with the same question. Below are the results of our polls.

SMU 34, 1200, 23, 50, 60, 50, 0, 0, 30, 89, 0, 300, 400, 20, 10, 0 Seattle U 20, 10, 5, 0, 30, 50, 0, 100, 110, 0, 40, 10, 3, 0 a. Run a two sample t-test to test if the mean amount of pocket cash from students at SMU is different than that of students from Seattle University. Write up a complete analysis: all 6 steps including a statistical conclusion and scope of inference (similar to the one from the PowerPoint). (This should include identifying the Ho and Ha as well as the p-value.) Also include the appropriate confidence interval. FUTURE DATA SCIENTIST'S CHOICE!: YOU MAY USE SAS OR R TO DO THIS PROBLEM! b. Compare the p-value from this test with the one you found from the permutation test from last week. Provide a short 2 to 3 sentence discussion on your thoughts as to why they are the same or different.

Answer

8.1 Full Analysis

Hypothesis Definition

Hypothesis set up:

$$\begin{aligned} H_0 :& \mu_1 - \mu_2 = 0 \\ H_1 :& \mu_1 - \mu_2 \neq 0 \end{aligned}$$

Critical value and shaded distribution

Next we draw and shade our distribution: In a two sample t-test, we have that:

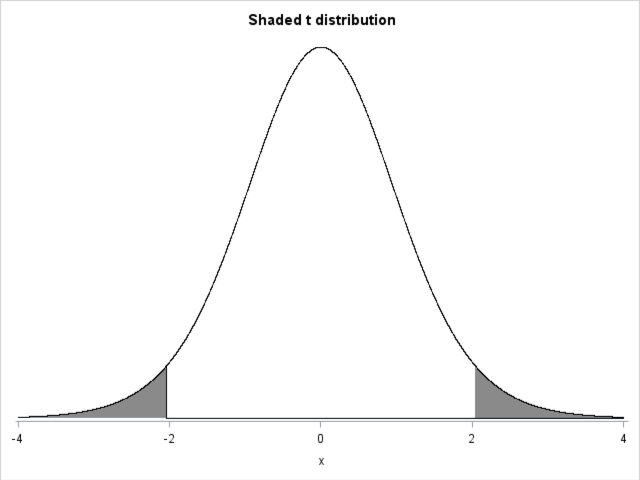
$$df = n_1 + n_2 - 2$$

where in our case, $df = 16 + 14 - 2 = 28,\ \alpha = 0.05$. In this case we are performing a two tailed test. Now we input this information into SAS to draw our distribution[1]:

```
data pdf;
do x = -4 to 4 by .001;
pdf = pdf("T", x, 14);
/*here it is important to set up a two sided test*/
if x <= quantile("T",.025,28) then lower = pdf;
else lower = 0;
if x >= quantile("T",.975,28) then upper = pdf;
```

```
else upper = 0;
output;    end; run;
title 'Shaded t distribution';
proc sgplot data=pdf noautolegend noborder;
yaxis display=none;
band x = x lower = lower upper = upper / fillattrs=(color=gray8a);
series x = x y = pdf / lineattrs = (color = black);
series x = x y = lower / lineattrs = (color = black);
run;
```

With this bit of code, we have produced our shaded two tailed PDF:



This critical value, where the bands start, is calculated using the following SAS code:

```
data critval;
p = quantile("T",.975,28); /*two sided test*/;
proc print data=critval;
run;
```

This gives us a critical t value of ±2.04841

Obs	p
1	2.04841

T statistic

the t stat is calculated using the following code:

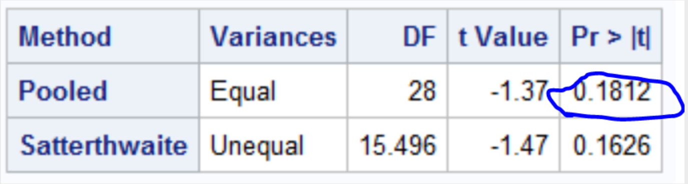
Code 8.1. Two sided two sample t test in SAS

```
proc ttest data=wallet
alpha=.05 test=diff
sides=2; /*an upper tailed test*/
class school;
var cash;
run;
```

which tells us that our t statistic is -1.37

P value

With the code from the previous step, we also see the p value, $p = 0.1812$:



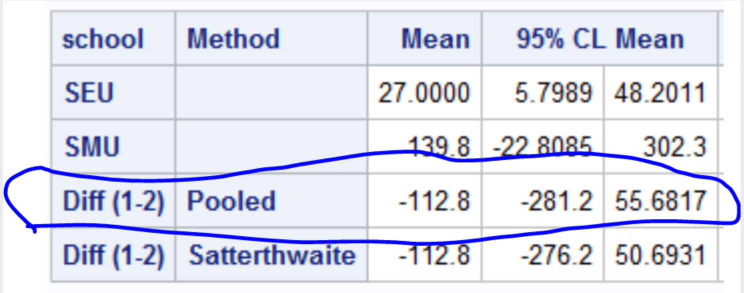
Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	28	-1.37	0.1812
Satterthwaite	Unequal	15.496	-1.47	0.1626

Hypothesis Assessment

$p = 0.1812 > \alpha = 0.05$ for the one tailed hypothesis test, indicating that we CANNOT REJECT the null hypothesis

Conclusion and Scope of inference

We cannot reject the null hypothesis, meaning we cannot say that the mean amount of cash in an SMU student's wallet is any different than the mean amount of cash in a Seattle U student's wallet. The following figure is a good reference for the results of this test:



school	Method	Mean	95% CL Mean	
SEU		27.0000	5.7989	48.2011
SMU		139.8	-22.8085	302.3
Diff (1-2)	Pooled	-112.8	-281.2	55.6817
Diff (1-2)	Satterthwaite	-112.8	-276.2	50.6931

The circled area tells us the difference between the mean amount of cash in a Seattle student's wallet and an SMU student's wallet. We can see that the average student from the seattle sample had about 112 dollars less in his wallet than the average SMU student. This may sound like a lot, however it is not significant. For this result to be statistically significant, and the mean amount of cash in a Seattle U student's wallet to be considered different than the mean amount of cash in an SMU student's wallet, the difference of the two means would have to fall outside of the 95% confidence interval. The confidence interval is highlighted, and is $(-281.2,\ 55.6817)$, which tells us that for the means to be considered truly different, the seattle student should have either 281 dollars less than the SMU student, or 55 dollars more. Our p value of 0.1812 tells us a similar story. It tells us that there is an 18% chance that a greater difference in the means would occur, which, at a 5 or 10 percent confidence interval, is not statistically significant at all. As for scope of inference, we cannot make inferences about the greater population of either university, because these were not random samples. We also cannot make causal inferences (eg going to SMU makes you have money in your wallet!), as this is not a randomized experiment either. Something about outliers!

Chapter 9

Problem 4: power

Question

4. A. Calculate the estimate of the pooled standard deviation from the Samoan discrimination problem. Use this estimate to build a power curve. Assume we would like to be able to detect effect sizes between 0.5 and 2 and we would like to calculate the sample size required to have a test that has a power of .8. Simply cut and paste your power curve and SAS code. HINT: USE THE CODE FROM DR. McGEE's lecture. Instead of using groupstddevs, use stddev since we are using the pooled estimate. B. Now suppose we decided that we may be able to live with slightly less power if it means savings in sample size. Provide the same plot as above but this time calculate curves of sample size (y-axis) vs. effect size (.5 to 2) (x axis) for power = 0.8, 0.7, and 0.6. There should be three plots on your final plot. Simply cut and paste your power curve and SAS code. HINT: USE THE CODE FROM DR. McGEE's lecture. Instead of using groupstddevs, use stddev since we are using the pooled estimate. The effect size here refers to a difference in means, though there are many effect size metrics, such a Cohen's D. C. Using similar code, estimate the savings in sample size from a test aimed at detecting an effect size of 0.8 with a power of 80% versus a power of 60%. Note: You will learn how to do this in R in a future HW!

Answers

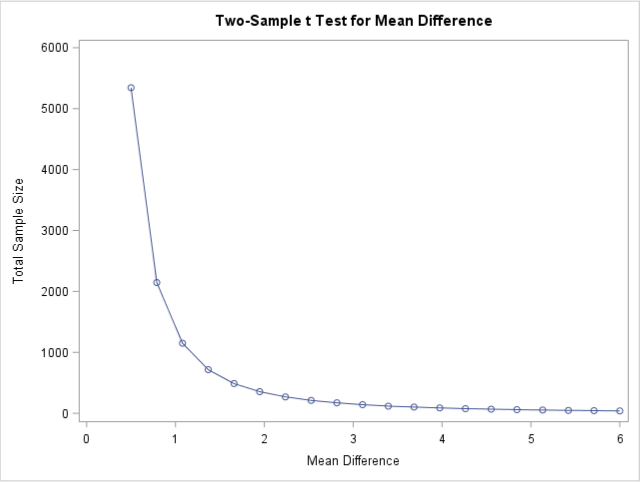
9.1 Single power curve

he pooled standard deviation, calculated in Problem 2, part e, part 1, is $s_p = 6.5215$. The difference of the means of the two groups, meandiff in the code, is just set to the difference between the means of our two populations, calculated using the R-generated means in Problem 2, Part f, $\mu_f - \mu_{uf} = 1.924$. The value of meandiff is not important, because by plotting the effect size, we are cycling through mean differences between 0.5 and 6, so the meandiff parameter only really matters if you want to know a sample size for a specific difference of means. When building a power curve it is not important at all, but you need it to get proc power to work. The SAS code used to build the power curve is shown below:

Code 9.1. Proc power single with pooled variance

```
proc power;
twosamplemeans
/*test=diff not diffsatt bc pooled variance*/
test=diff
stddev=6.5215
/*meandiff is a dummy variable in this case*/
meandiff=1.924
power=.8
ntotal = .;
plot x=effect min=.5 max=6;
run;
```

And the power curve:



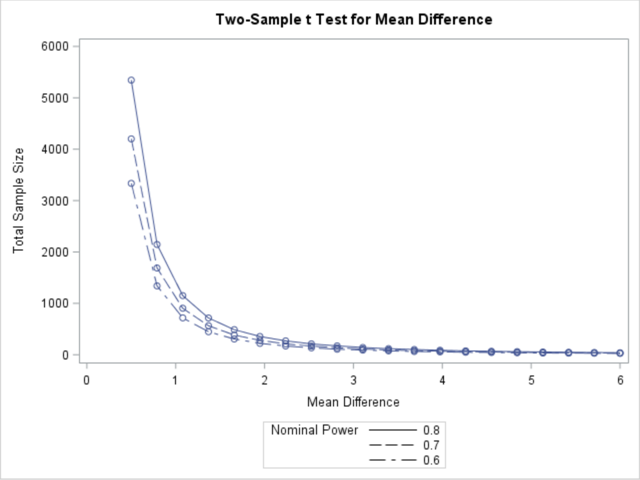
9.2 Multiple power curves

The same notes as above apply here, this time we used the SAS code to generate multiple power curves:

Code 9.2. Producing several curves with proc power

```
proc power;
twosamplemeans
/*test=diff not diffsatt bc pooled variance*/
test=diff
stddev=6.5215
/*meandiff is a dummy variable in this case*/
meandiff=1.924
power=.8 .7 .6
ntotal = .;
plot x=effect min=.5 max=6;
run;
```

And the curves:

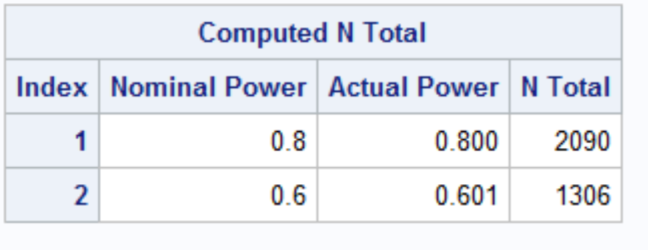


9.3 Calculating change in N

It is important to remember that the “effect size” calculated in this SAS code is the exact same thing as the “mean difference”. Therefore we can write our SAS code as follows:

```
proc power;
twosamplemeans
test=diff /*diff not diffsatt bc pooled variance*/
stddev=6.5215
meandiff= 0.8 /*this represents the effect size*/
power=.8 .6
ntotal = .;
run;
```

Which gives us our sample size savings:



Computed N Total			
Index	Nominal Power	Actual Power	N Total
1	0.8	0.800	2090
2	0.6	0.601	1306

As we see from the figure above, by raising the power from 0.6 to 0.8, we actually have to nearly double the sample size to meet the test parameters. By using a power of 0.6, we save 784 N’s (or sample size units)

Chapter 10

Unit 2 Lecture Slides

Inference Using t-Distributions

MEASURING UNCERTAINTY IN RANDOMIZED AND OBSERVATIONAL STUDIES

-DISTRIBUTION OF THE SAMPLE AVERAGE

-USING T-DISTRIBUTION FOR ONE SAMPLE INFERENCE

-STARTING TO EXPLORE T-DISTRIBUTION FOR TWO SAMPLE PROBLEMS

Central Limit Theorem

Distribution of Sample Average

- If $Y_1, Y_2, \ldots, Y_n$ is the sample, then

$$\bar{Y} = \frac{(Y_1 + Y_2 + \ldots + Y_n)}{n}$$

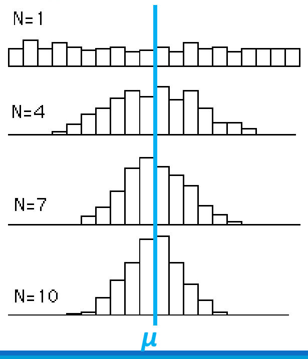
- The idea: $\bar{Y}$ is a point estimate for the population mean μ
- The sample mean is an unbiased estimator for the population mean.
  - $E(\bar{Y}) = \mu$ because $E(Y_i) = \mu$*

  *See proof in appendix.

Distribution of Sample Average

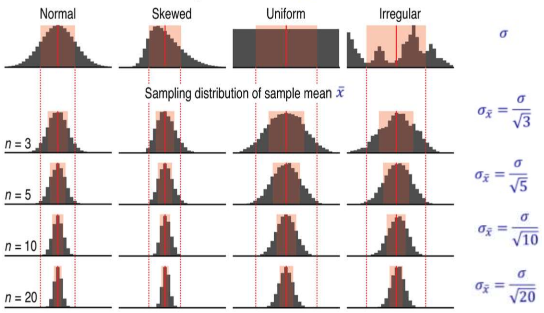
- We can say more about $\bar{Y}$ than that!
- It turns out that
  1. $\bar{Y}$ is unbiased.
  2. Variance$(\bar{Y}) = \frac{\sigma^2}{n}$, where σ^2 is the variance of the population
  3. $\bar{Y}$ distribution is approximately normal if n is larger than 30
- This last fact is due to the **CENTRAL LIMIT THEOREM (CLT)**

The more data you pick for each sample, the more normal (and tighter) the distribution of the sample mean is.

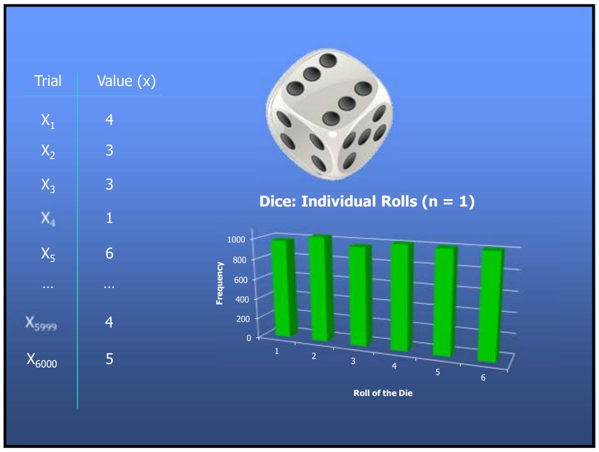


Note that the distribution of the original data is the distribution of a sample mean of size 1.

The more data you pick for each sample, the more normal (and tighter) the distribution of the sample mean is.
If original data is approx. normal, then the distribution of the sample mean will be approx. normal, regardless of sample size.

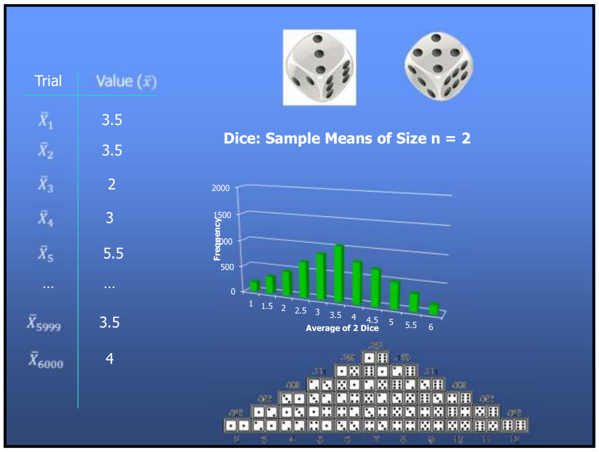


http://onlinestatbook.com/stat_sim/sampling_dist/



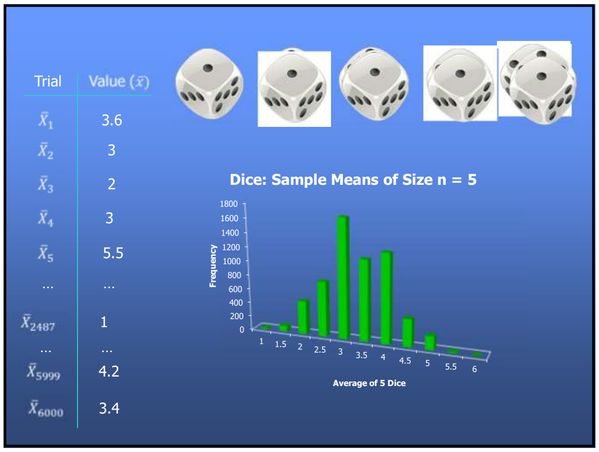
Trial	Value (x)
X_1	4
X_2	3
X_3	3
X_4	1
X_5	6
…	…
X_{5999}	4
X_{6000}	5

Dice: Individual Rolls (n = 1)



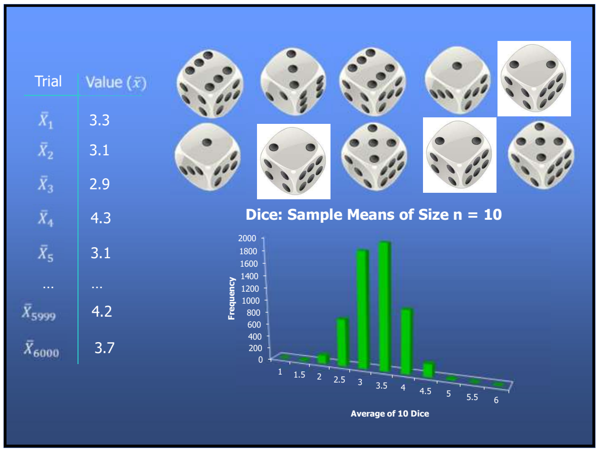
Trial	Value ($\bar{x}$)
$\bar{X}_1$	3.5
$\bar{X}_2$	3.5
$\bar{X}_3$	2
$\bar{X}_4$	3
$\bar{X}_5$	5.5
…	…
$\bar{X}_{5999}$	3.5
$\bar{X}_{6000}$	4

Dice: Sample Means of Size n = 2



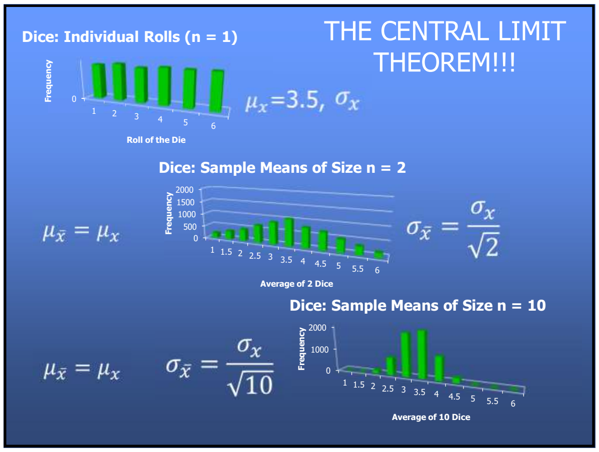
Trial	Value ($\bar{x}$)
$\bar{X}_1$	3.6
$\bar{X}_2$	3
$\bar{X}_3$	2
$\bar{X}_4$	3
$\bar{X}_5$	5.5
…	…
$\bar{X}_{2487}$	1
…	…
$\bar{X}_{5999}$	4.2
$\bar{X}_{6000}$	3.4

Dice: Sample Means of Size n = 5



Trial	Value ($\bar{x}$)
$\bar{X}_1$	3.3
$\bar{X}_2$	3.1
$\bar{X}_3$	2.9
$\bar{X}_4$	4.3
$\bar{X}_5$	3.1
…	…
$\bar{X}_{5999}$	4.2
$\bar{X}_{6000}$	3.7

Dice: Sample Means of Size n = 10



THE CENTRAL LIMIT THEOREM!!!

Dice: Individual Rolls (n = 1)

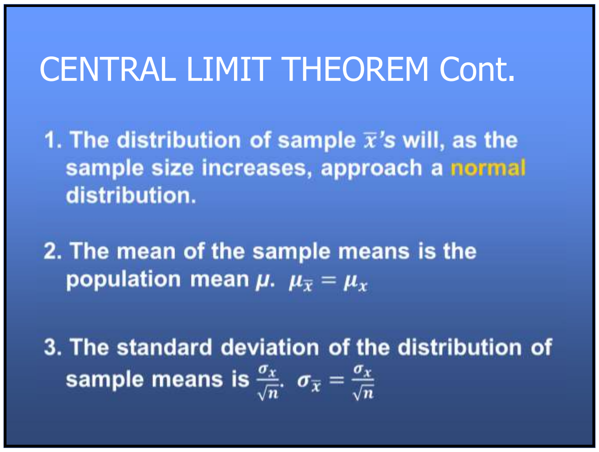
μ_x=3.5, σ_x

Dice: Sample Means of Size n = 2

$\mu_{\bar{x}} = \mu_x$ $\quad \sigma_{\bar{x}} = \dfrac{\sigma_x}{\sqrt{2}}$

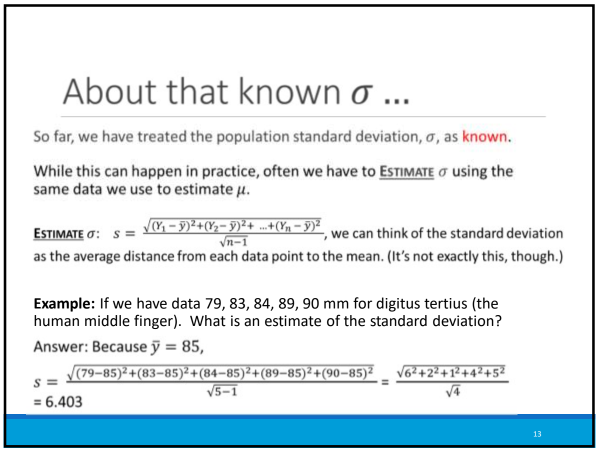
Dice: Sample Means of Size n = 10

$\mu_{\bar{x}} = \mu_x$ $\quad \sigma_{\bar{x}} = \dfrac{\sigma_x}{\sqrt{10}}$



CENTRAL LIMIT THEOREM Cont.

1. **The distribution of sample $\bar{x}$'s will, as the sample size increases, approach a normal distribution.**
2. **The mean of the sample means is the population mean μ. $\mu_{\bar{x}} = \mu_x$**
3. **The standard deviation of the distribution of sample means is $\frac{\sigma_x}{\sqrt{n}}$. $\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$**



About that known σ …

So far, we have treated the population standard deviation, σ, as known.

While this can happen in practice, often we have to ESTIMATE σ using the same data we use to estimate μ.

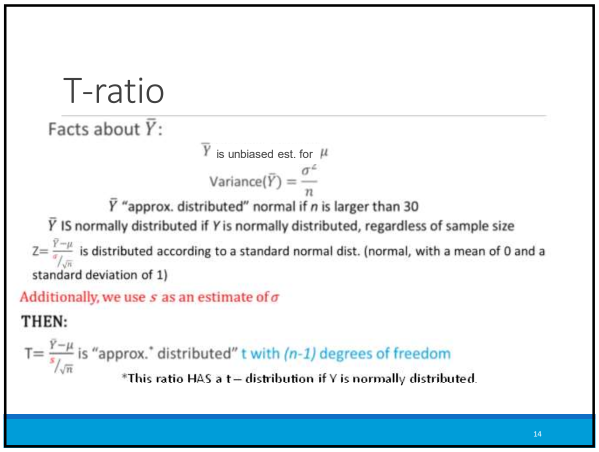
ESTIMATE σ: $s = \frac{\sqrt{(Y_1-\bar{y})^2+(Y_2-\bar{y})^2+\ \ldots+(Y_n-\bar{y})^2}}{\sqrt{n-1}}$, we can think of the standard deviation as the average distance from each data point to the mean. (It's not exactly this, though.)

Example: If we have data 79, 83, 84, 89, 90 mm for digitus tertius (the human middle finger). What is an estimate of the standard deviation?

Answer: Because $\bar{y} = 85$,

$$s = \frac{\sqrt{(79-85)^2+(83-85)^2+(84-85)^2+(89-85)^2+(90-85)^2}}{\sqrt{5-1}} = \frac{\sqrt{6^2+2^2+1^2+4^2+5^2}}{\sqrt{4}}$$

= 6.403



T-ratio

Facts about $\bar{Y}$:

$\bar{Y}$ is unbiased est. for μ

$\text{Variance}(\bar{Y}) = \frac{\sigma^2}{n}$

$\bar{Y}$ "approx. distributed" normal if n is larger than 30

$\bar{Y}$ IS normally distributed if Y is normally distributed, regardless of sample size

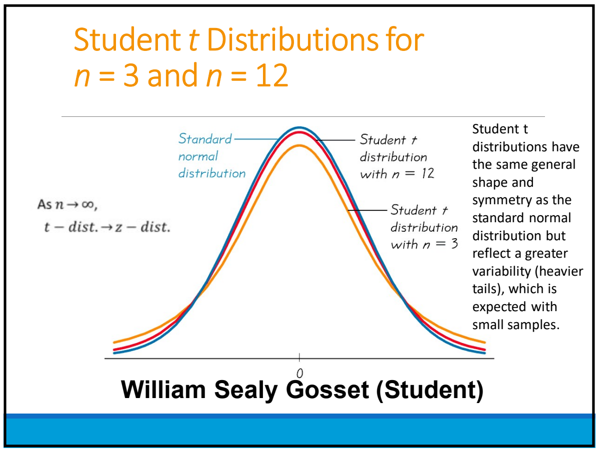
$Z = \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}$ is distributed according to a standard normal dist. (normal, with a mean of 0 and a standard deviation of 1)

Additionally, we use s as an estimate of σ

THEN:

$T = \frac{\bar{Y}-\mu}{s/\sqrt{n}}$ is "approx.* distributed" t with *(n-1)* degrees of freedom

*This ratio HAS a t – distribution if Y is normally distributed.

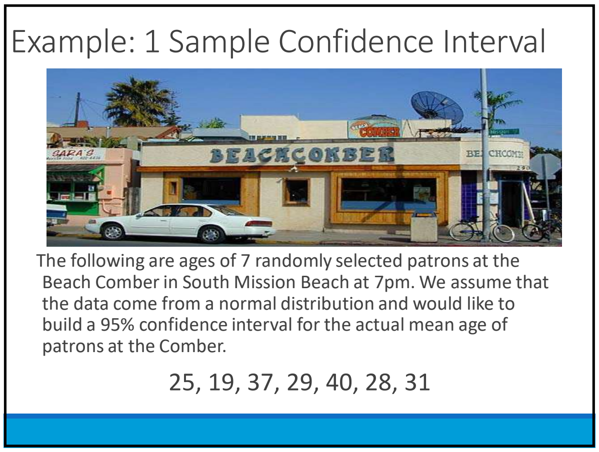


Student *t* Distributions for *n* = 3 and *n* = 12

As $n \to \infty$,
$t-dist. \to z-dist.$

Student t distributions have the same general shape and symmetry as the standard normal distribution but reflect a greater variability (heavier tails), which is expected with small samples.

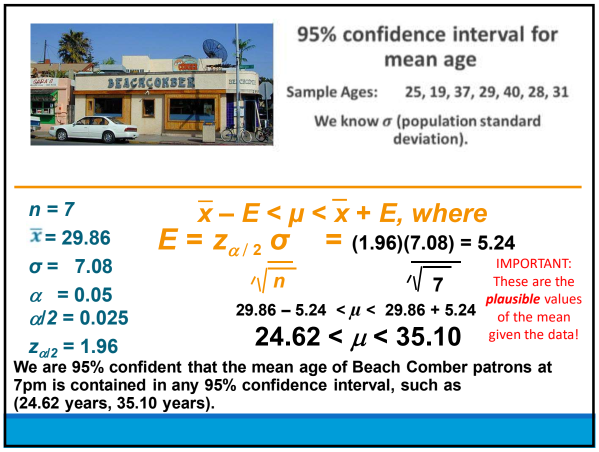
William Sealy Gosset (Student)



Example: 1 Sample Confidence Interval

The following are ages of 7 randomly selected patrons at the Beach Comber in South Mission Beach at 7pm. We assume that the data come from a normal distribution and would like to build a 95% confidence interval for the actual mean age of patrons at the Comber.

25, 19, 37, 29, 40, 28, 31



95% confidence interval for mean age

Sample Ages: 25, 19, 37, 29, 40, 28, 31

We know σ (population standard deviation).

$n = 7$
$\bar{x} = 29.86$
$\sigma = 7.08$
$\alpha = 0.05$
$\alpha/2 = 0.025$
$z_{\alpha/2} = 1.96$

$$\bar{x} - E < \mu < \bar{x} + E, \text{ where}$$

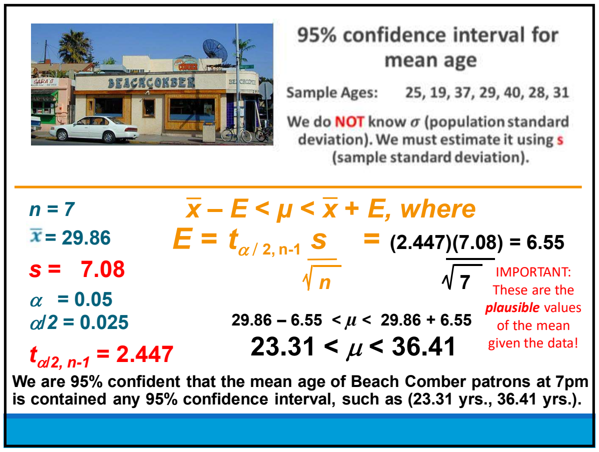
$$E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = \frac{(1.96)(7.08)}{\sqrt{7}} = 5.24$$

$$29.86 - 5.24 < \mu < 29.86 + 5.24$$

$$24.62 < \mu < 35.10$$

IMPORTANT: These are the ***plausible*** values of the mean given the data!

We are 95% confident that the mean age of Beach Comber patrons at 7pm is contained in any 95% confidence interval, such as (24.62 years, 35.10 years).



95% confidence interval for mean age

Sample Ages: 25, 19, 37, 29, 40, 28, 31

We do NOT know σ (population standard deviation). We must estimate it using s (sample standard deviation).

$n = 7$
$\bar{x} = 29.86$
$s = 7.08$
$\alpha = 0.05$
$\alpha/2 = 0.025$
$t_{\alpha/2,\, n-1} = 2.447$

$$\bar{x} - E < \mu < \bar{x} + E, \text{ where}$$

$$E = t_{\alpha/2,\, n-1}\frac{s}{\sqrt{n}} = \frac{(2.447)(7.08)}{\sqrt{7}} = 6.55$$

$$29.86 - 6.55 < \mu < 29.86 + 6.55$$

$$23.31 < \mu < 36.41$$

IMPORTANT: These are the ***plausible*** values of the mean given the data!

We are 95% confident that the mean age of Beach Comber patrons at 7pm is contained any 95% confidence interval, such as (23.31 yrs., 36.41 yrs.).

Comparison of z to t

$n = 7$
$\bar{x} = 29.86$
$\sigma = 7.08$
$\alpha = 0.05$
$\alpha/2 = 0.025$
$z_{\alpha/2} = 1.96$

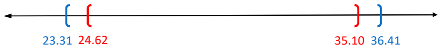
$$E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = \frac{(1.96)(7.08)}{\sqrt{7}} = 5.24$$

$$\bar{x} - E < \mu < \bar{x} + E$$

$$29.86 - 5.24 < \mu < 29.86 + 5.24$$

$$24.62 < \mu < 35.10$$

We are 95% confident that the mean age of Beach Comber patrons at 7pm is contained in the interval (24.62 years, 35.10 years).



$n = 7$
$\bar{x} = 29.86$
$s = 7.08$
$\alpha = 0.05$
$\alpha/2 = 0.025$
$t_{\alpha/2,\, n-1} = 2.447$

$$E = t_{\alpha/2,\, n-1} \frac{s}{\sqrt{n}} = \frac{(2.447)(7.08)}{\sqrt{7}} = 6.55$$

$$\bar{x} - E < \mu < \bar{x} + E$$

$$29.86 - 6.55 < \mu < 29.86 + 6.55$$

$$23.31 < \mu < 36.41$$

We are 95% confident that the mean age of Beach Comber patrons at 7pm is contained in the interval (23.31 years, 36.41 years).

1 Sample Hypothesis Testing: The 6 Steps

1. Identify Ho and Ha.
2. Find the Critical Value(s) and Draw and Shade.
3. Calculate the Test – Statistic. (The evidence!)
4. Calculate the P-value.
5. Make a decision… Reject Ho or FTR Ho.
6. Write a clear conclusion in the context of the problem…. Use mostly non statistical terms but always report the p-value! Add a confidence interval if appropriate. End this conclusion with a statement about the scope.

Example: 1 Sample t-test

The following are ages of 7 randomly chosen patrons seen leaving the Beach Comber in South Mission Beach at 7pm. We assume that the data come from a normal distribution and would like to test the claim that the mean age of the distribution of Comber patrons is different than 21.

25, 19, 37, 29, 40, 28, 31

Let's Formalize This Test Into 6 Steps!

We would like to test the claim that the population mean is different than 21.

Step 1: Identify the null (Ho) and alternative (Ha) hypothesis.

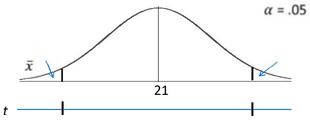
$$Ho: \mu = 21$$
$$Ha: \mu \neq 21$$

Let's Formalize This Test Into 6 Steps!

We would like to test the claim that the population mean is different from 21.
To do this, we take a sample of size n = 7.

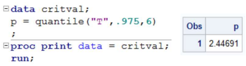
Step 1: Identify the null (Ho) and alternative (Ha) hypothesis.

Step 2: Draw and Shade and Find the Critical Value.



$\alpha = .05$ = significance level.

df = 7 – 1 = 6



```
data critval;
 p = quantile("T",.975,6)
 ;
proc print data = critval;
 run;
```

Obs	p
1	2.44691

Let's Formalize This Test Into 6 Steps!

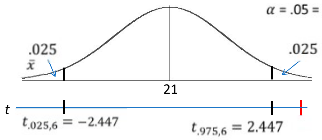
We would like to test the claim that the population mean is not equal to 21. To do this, we take a sample of size n = 7 and find that $\bar{x}$ = 29.86 years and s = 7.08 years.

Step 1: Identify the null (Ho) and alternative (Ha) hypothesis.

Step 2: Draw and Shade and Find the Critical Value.

$\alpha = .05$ = significance level.

df = 7 – 1 = 6



Step 3: Find the test statistic. (The t value for the data.)

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

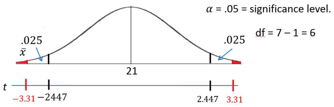
Let's Formalize This Test Into 6 Steps!

We would like to test the claim that the population mean is not equal to 21. To do this, we take a sample of size n = 8 and find that $\bar{x}$ = 29.86 years and s = 7.09 years.

Step 1: Identify the null (Ho) and alternative (Ha) hypothesis. Ho: $\mu = 21$

Ha: $\mu \neq 21$

Step 2: Draw and Shade and Find the Critical Value.



Step 3: Find the test statistic. (The t value for the data.) $t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}} = \frac{29.86-21}{\frac{7.09}{\sqrt{7}}} = 3.31$

Step 4: Find the p-value: *The probability of observing by random chance something* ***as extreme or more extreme*** *than what was observed under the assumption that the null hypothesis is true.* (Usually found with software.) The red shaded region above is 0.0162 (sum of both red areas)

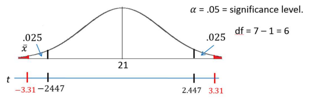
Let's Formalize This Test Into 6 Steps!

We would like to test the claim that the population mean is not equal to 21. To do this, we take a sample of size n = 8 and find that $\bar{x}$ = 29.86 years and s = 7.09 years.

Step 1: Identify the null (Ho) and alternative (Ha) hypothesis. Ho: $\mu = 21$

Ha: $\mu \neq 21$

Step 2: Draw and Shade and Find the Critical Value.



Step 3: Find the test statistic. (The t value for the data.)

$$t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}} = \frac{29.86-21}{\frac{7.09}{\sqrt{7}}} = 3.31$$

Step 4: Find the p-value: P-value 0.0162< .05

Step 5: Key! The sample mean we found is very unusual under the assumption that the true mean age is 21. So we Reject the assumption that the true mean age is 21. That is, we REJECT Ho.

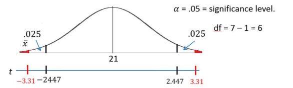
Let's Formalize This Test Into 6 Steps!

We would like to test the claim that the population mean is not equal to 21. To do this, we take a sample of size n = 8 and find that $\bar{x}$ = 29.86 years and s = 7.09 years.

Step 1: Identify the null (Ho) and alternative (Ha) hypothesis. Ho: $\mu = 21$

Ha: $\mu \neq 21$

Step 2: Draw and Shade and Find the Critical Value.



Step 3: Find the test statistic. (The t value for the data.) $t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}} = \frac{29.86-21}{\frac{7.09}{\sqrt{7}}}$ $= 3.31$

Step 4: Find the p-value: P-value 0.0162 < .05

Step 5: REJECT Ho

Step 6: There is sufficient evidence to conclude that the true mean age of patrons at the Comber at 7pm is not equal to 21 (p-value =0.0162 from a t-test). We could also say that there is sufficient evidence to conclude that the true mean is greater than 21. (Consider the red area in the right most tail.) This was not a random sample of all times, only at 7pm; thus, the result cannot be applied to the bar at all times. The results are nevertheless intriguing.

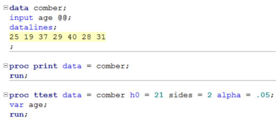
Finding the P-value – more detail

Step 4: Find the p-value: p-value < .05

You could use Stat Trek / or the t-table.

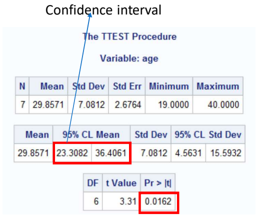
OR

Software like SAS:



```
data comber;
input age @@;
datalines;
25 19 37 29 40 28 31
;

proc print data = comber;
run;

proc ttest data = comber h0 = 21 sides = 2 alpha = .05;
var age;
run;
```

Confidence interval



The TTEST Procedure

Variable: age

N	Mean	Std Dev	Std Err	Minimum	Maximum
7	29.8571	7.0812	2.6764	19.0000	40.0000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
29.8571	23.3082	36.4061	7.0812	4.5631	15.5932

DF	t Value	Pr > |t|
6	3.31	0.0162

One-Sided Test + Two-Sided CI Demonstration

Suppose we would like to test the claim that the mean age of patrons is greater than 24.

Step 1: State the null and alternative hypotheses.

- $H_o: \mu \leq 24\ (or\ \mu = 24)\ vs.\ H_a: \mu > 24$

One-Sided Test + Two-Sided CI Demonstration

Suppose we would like to test the claim that the mean age of patrons is greater than 24.

Skipping to the most important stuff…

Critical value, $t_{0.95,6} = \pm 1.943$

Test statistic, $t = 2.1884$

P-value, $p = 0.036$

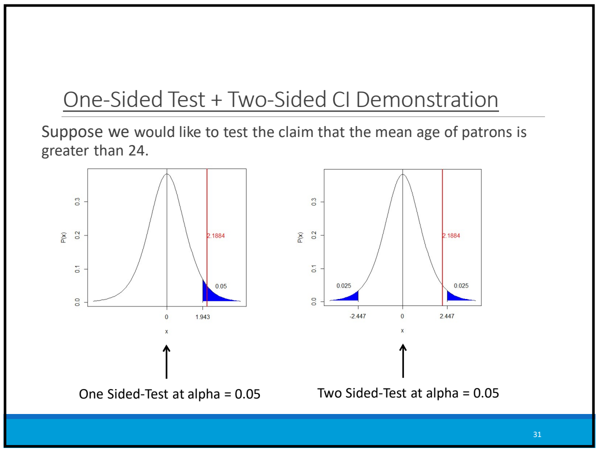
Conclusion: reject H_o

i.e. conclude that the mean is greater than 24.

1-sided 95% CI: $[24.7, \infty]$

2-sided 95% CI: $[23.3, 36.4]$

- But… wait! 24 is in the CI, implying it is a 'plausible' value – i.e. we would fail to reject the null.

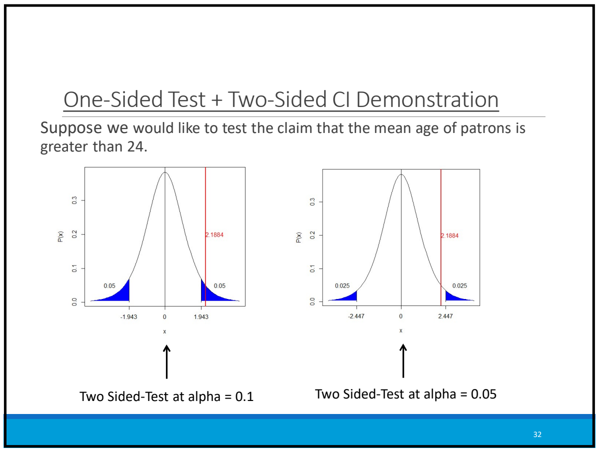


One-Sided Test + Two-Sided CI Demonstration

Suppose we would like to test the claim that the mean age of patrons is greater than 24.

One Sided-Test at alpha = 0.05

Two Sided-Test at alpha = 0.05

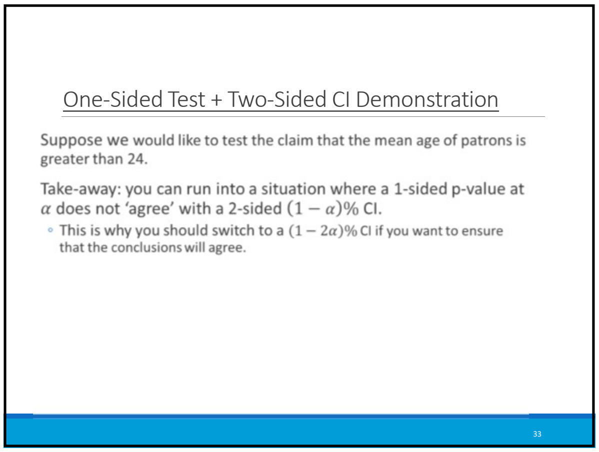


One-Sided Test + Two-Sided CI Demonstration

Suppose we would like to test the claim that the mean age of patrons is greater than 24.

Two Sided-Test at alpha = 0.1

Two Sided-Test at alpha = 0.05

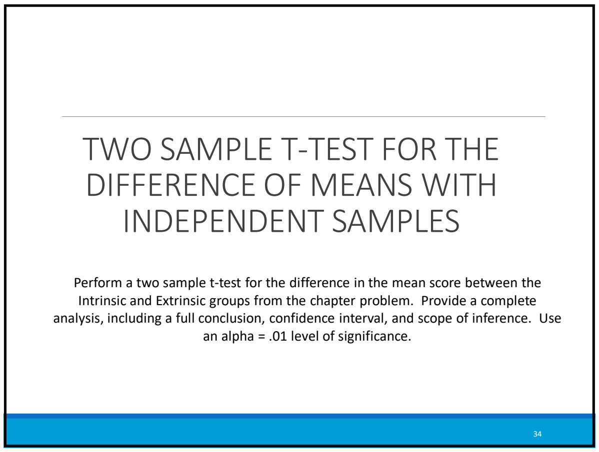


One-Sided Test + Two-Sided CI Demonstration

Suppose we would like to test the claim that the mean age of patrons is greater than 24.

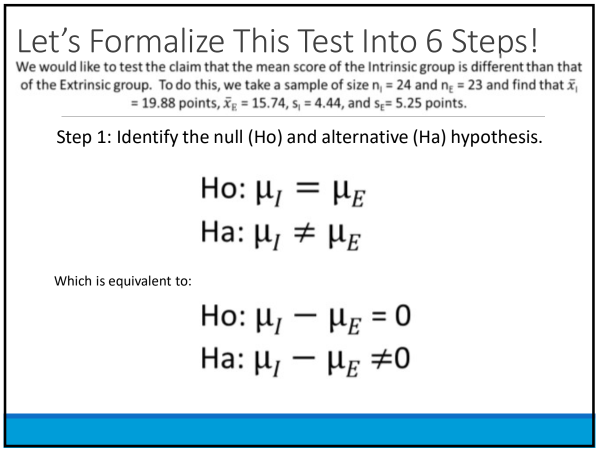
Take-away: you can run into a situation where a 1-sided p-value at α does not 'agree' with a 2-sided $(1-\alpha)\%$ CI.

- This is why you should switch to a $(1-2\alpha)\%$ CI if you want to ensure that the conclusions will agree.



TWO SAMPLE T-TEST FOR THE DIFFERENCE OF MEANS WITH INDEPENDENT SAMPLES

Perform a two sample t-test for the difference in the mean score between the Intrinsic and Extrinsic groups from the chapter problem. Provide a complete analysis, including a full conclusion, confidence interval, and scope of inference. Use an alpha = .01 level of significance.



Let's Formalize This Test Into 6 Steps!

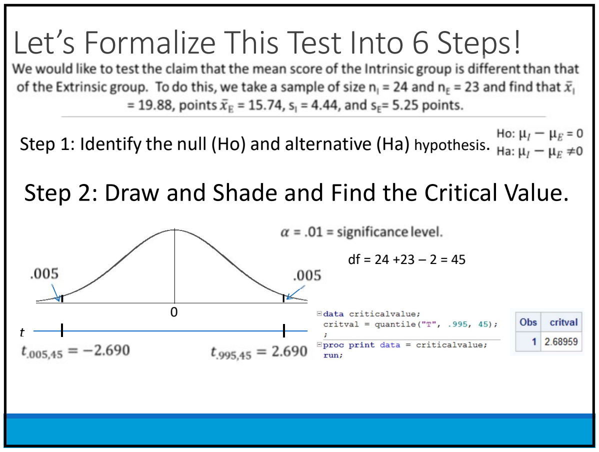
We would like to test the claim that the mean score of the Intrinsic group is different than that of the Extrinsic group. To do this, we take a sample of size $n_I = 24$ and $n_E = 23$ and find that $\bar{x}_I = 19.88$ points, $\bar{x}_E = 15.74$, $s_I = 4.44$, and $s_E = 5.25$ points.

Step 1: Identify the null (Ho) and alternative (Ha) hypothesis.

$$\text{Ho: } \mu_I = \mu_E$$
$$\text{Ha: } \mu_I \neq \mu_E$$

Which is equivalent to:

$$\text{Ho: } \mu_I - \mu_E = 0$$
$$\text{Ha: } \mu_I - \mu_E \neq 0$$



Let's Formalize This Test Into 6 Steps!

We would like to test the claim that the mean score of the Intrinsic group is different than that of the Extrinsic group. To do this, we take a sample of size $n_I = 24$ and $n_E = 23$ and find that $\bar{x}_I = 19.88$, points $\bar{x}_E = 15.74$, $s_I = 4.44$, and $s_E = 5.25$ points.

Step 1: Identify the null (Ho) and alternative (Ha) hypothesis. Ho: $\mu_I - \mu_E = 0$ Ha: $\mu_I - \mu_E \neq 0$

Step 2: Draw and Shade and Find the Critical Value.

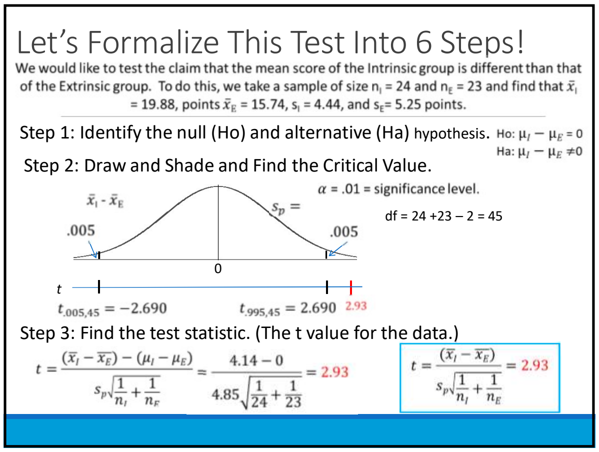
$\alpha = .01$ = significance level.

df = 24 +23 – 2 = 45

$t_{.005,45} = -2.690$ $t_{.995,45} = 2.690$

```
data criticalvalue;
 critval = quantile("T", .995, 45);
 ;
proc print data = criticalvalue;
 run;
```

Obs	critval
1	2.68959



Let’s Formalize This Test Into 6 Steps!

We would like to test the claim that the mean score of the Intrinsic group is different than that of the Extrinsic group. To do this, we take a sample of size $n_I = 24$ and $n_E = 23$ and find that $\bar{x}_I = 19.88$, points $\bar{x}_E = 15.74$, $s_I = 4.44$, and $s_E = 5.25$ points.

Step 1: Identify the null (Ho) and alternative (Ha) hypothesis. Ho: $\mu_I - \mu_E = 0$
Ha: $\mu_I - \mu_E \neq 0$

Step 2: Draw and Shade and Find the Critical Value.

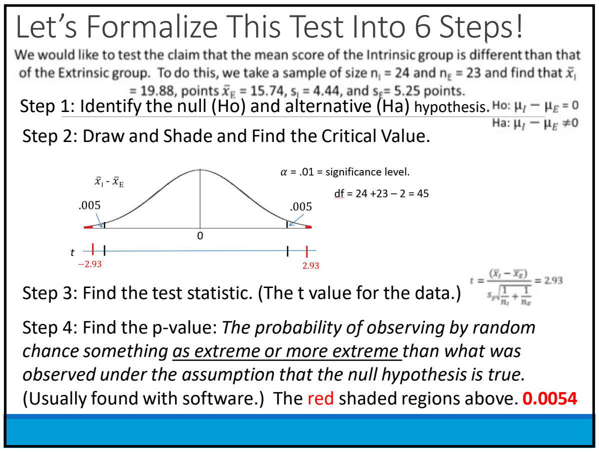
$\alpha = .01$ = significance level.

df = 24 +23 – 2 = 45

Step 3: Find the test statistic. (The t value for the data.)

$$t = \frac{(\bar{x}_I - \bar{x}_E) - (\mu_I - \mu_E)}{s_p\sqrt{\frac{1}{n_I} + \frac{1}{n_E}}} = \frac{4.14 - 0}{4.85\sqrt{\frac{1}{24} + \frac{1}{23}}} = 2.93$$

$$\boxed{t = \frac{(\bar{x}_I - \bar{x}_E)}{s_p\sqrt{\frac{1}{n_I} + \frac{1}{n_E}}} = 2.93}$$



Let’s Formalize This Test Into 6 Steps!

We would like to test the claim that the mean score of the Intrinsic group is different than that of the Extrinsic group. To do this, we take a sample of size $n_I = 24$ and $n_E = 23$ and find that $\bar{x}_I = 19.88$, points $\bar{x}_E = 15.74$, $s_I = 4.44$, and $s_E = 5.25$ points.

Step 1: Identify the null (Ho) and alternative (Ha) hypothesis. Ho: $\mu_I - \mu_E = 0$
Ha: $\mu_I - \mu_E \neq 0$

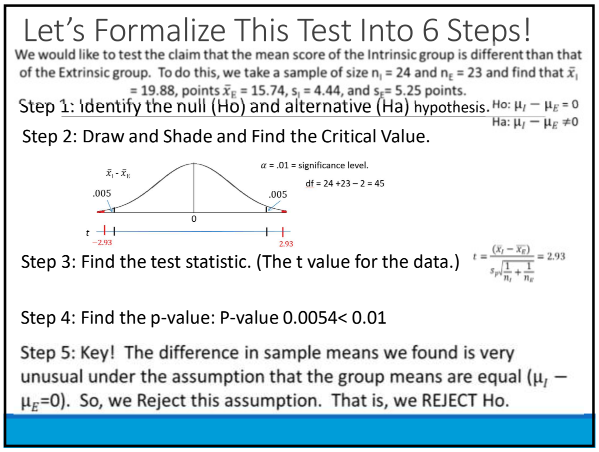
Step 2: Draw and Shade and Find the Critical Value.

$\alpha = .01$ = significance level.

df = 24 +23 – 2 = 45

Step 3: Find the test statistic. (The t value for the data.) $t = \frac{(\bar{x}_I - \bar{x}_E)}{s_p\sqrt{\frac{1}{n_I} + \frac{1}{n_E}}} = 2.93$

Step 4: Find the p-value: *The probability of observing by random chance something* *as extreme or more extreme* *than what was observed under the assumption that the null hypothesis is true.* (Usually found with software.) The red shaded regions above. **0.0054**



Let’s Formalize This Test Into 6 Steps!

We would like to test the claim that the mean score of the Intrinsic group is different than that of the Extrinsic group. To do this, we take a sample of size $n_I = 24$ and $n_E = 23$ and find that $\bar{x}_I = 19.88$, points $\bar{x}_E = 15.74$, $s_I = 4.44$, and $s_E = 5.25$ points.

Step 1: Identify the null (Ho) and alternative (Ha) hypothesis. Ho: $\mu_I - \mu_E = 0$
Ha: $\mu_I - \mu_E \neq 0$

Step 2: Draw and Shade and Find the Critical Value.

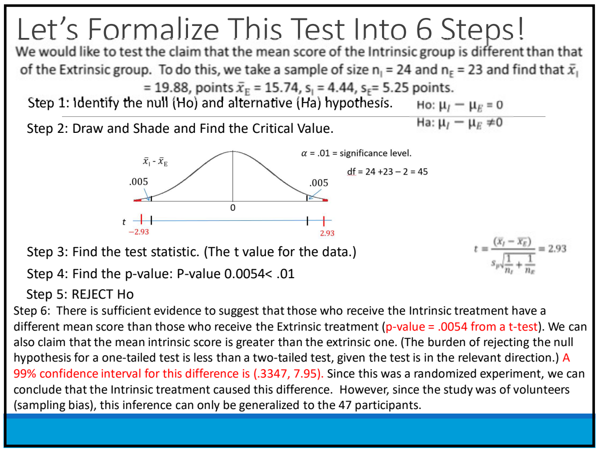
$\alpha = .01$ = significance level.

df = 24 +23 – 2 = 45

Step 3: Find the test statistic. (The t value for the data.) $t = \frac{(\bar{x}_I - \bar{x}_E)}{s_p\sqrt{\frac{1}{n_I} + \frac{1}{n_E}}} = 2.93$

Step 4: Find the p-value: P-value 0.0054< 0.01

Step 5: Key! The difference in sample means we found is very unusual under the assumption that the group means are equal ($\mu_I - \mu_E = 0$). So, we Reject this assumption. That is, we REJECT Ho.



Let’s Formalize This Test Into 6 Steps!

We would like to test the claim that the mean score of the Intrinsic group is different than that of the Extrinsic group. To do this, we take a sample of size $n_I = 24$ and $n_E = 23$ and find that $\bar{x}_I = 19.88$, points $\bar{x}_E = 15.74$, $s_I = 4.44$, $s_E = 5.25$ points.

Step 1: Identify the null (Ho) and alternative (Ha) hypothesis. Ho: $\mu_I - \mu_E = 0$
Ha: $\mu_I - \mu_E \neq 0$

Step 2: Draw and Shade and Find the Critical Value.

$\alpha = .01$ = significance level.

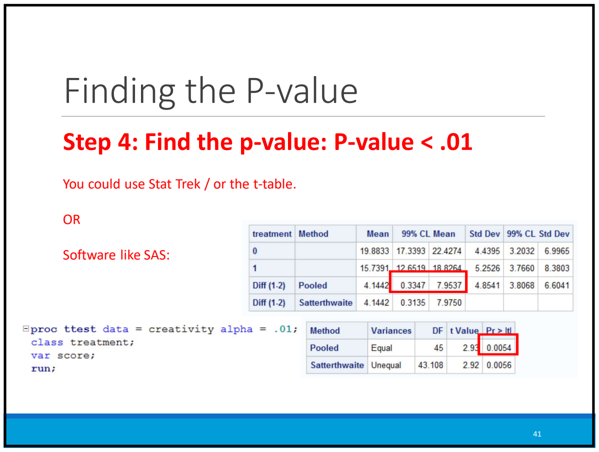
df = 24 +23 – 2 = 45

Step 3: Find the test statistic. (The t value for the data.) $t = \frac{(\bar{x}_I - \bar{x}_E)}{s_p\sqrt{\frac{1}{n_I} + \frac{1}{n_E}}} = 2.93$

Step 4: Find the p-value: P-value 0.0054< .01

Step 5: REJECT Ho

Step 6: There is sufficient evidence to suggest that those who receive the Intrinsic treatment have a different mean score than those who receive the Extrinsic treatment (p-value = .0054 from a t-test). We can also claim that the mean intrinsic score is greater than the extrinsic one. (The burden of rejecting the null hypothesis for a one-tailed test is less than a two-tailed test, given the test is in the relevant direction.) A 99% confidence interval for this difference is (.3347, 7.95). Since this was a randomized experiment, we can conclude that the Intrinsic treatment caused this difference. However, since the study was of volunteers (sampling bias), this inference can only be generalized to the 47 participants.



Finding the P-value

Step 4: Find the p-value: P-value < .01

You could use Stat Trek / or the t-table.

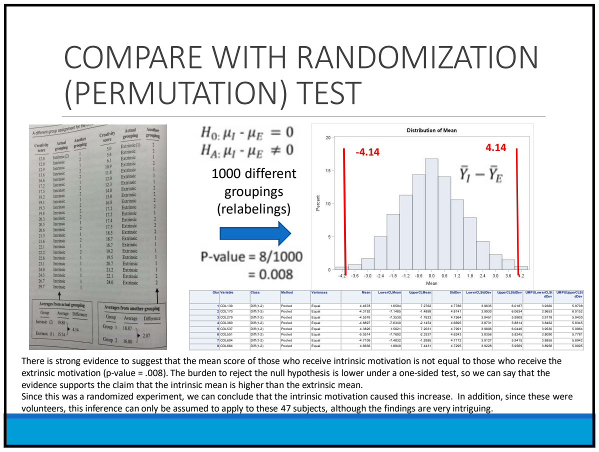
OR

Software like SAS:

treatment	Method	Mean	99% CL Mean		Std Dev	99% CL Std Dev	
0		19.8833	17.3393	22.4274	4.4395	3.2032	6.9965
1		15.7391	12.6519	18.8264	5.2526	3.7660	8.3803
Diff (1-2)	Pooled	4.1442	0.3347	7.9537	4.8541	3.8068	6.6041
Diff (1-2)	Satterthwaite	4.1442	0.3135	7.9750			

```
proc ttest data = creativity alpha = .01;
 class treatment;
 var score;
 run;
```

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	45	2.93	0.0054
Satterthwaite	Unequal	43.108	2.92	0.0056



COMPARE WITH RANDOMIZATION (PERMUTATION) TEST

$H_0: \mu_I - \mu_E = 0$
$H_A: \mu_I - \mu_E \neq 0$

1000 different groupings (relabelings)

P-value = 8/1000 = 0.008

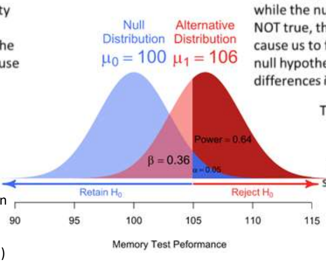
Obs	Variable	Class	Method	Variances	Mean	LowerCLMean	UpperCLMean	StdDev	LowerCLStdDev	UpperCLStdDev	UMPULowerCLStdDev	UMPUUpperCLStdDev
1	COL139	Diff (1-2)	Pooled	Equal	4.4678	1.6594	7.2762	4.7786	3.9635	6.0187	3.9360	5.9708
2	COL170	Diff (1-2)	Pooled	Equal	-4.3192	-7.1485	-1.4899	4.8141	3.9930	6.0634	3.9653	6.0152
3	COL279	Diff (1-2)	Pooled	Equal	-4.5576	-7.3530	-1.7623	4.7564	3.9451	5.9908	3.9178	5.9430
4	COL360	Diff (1-2)	Pooled	Equal	-4.8897	-7.6340	-2.1454	4.6695	3.8731	5.8814	3.8462	5.8345
5	COL537	Diff (1-2)	Pooled	Equal	4.3826	1.5621	7.2031	4.7991	3.9806	6.0446	3.9530	5.9964
6	COL551	Diff (1-2)	Pooled	Equal	-5.0514	-7.7692	-2.3337	4.6243	3.8356	5.8245	3.8090	5.7781
7	COL604	Diff (1-2)	Pooled	Equal	-4.7109	-7.4832	-1.9385	4.7172	3.9127	5.9415	3.8855	5.8942
8	COL664	Diff (1-2)	Pooled	Equal	4.6636	1.8840	7.4431	4.7295	3.9228	5.9569	3.8956	5.9095

There is strong evidence to suggest that the mean score of those who receive intrinsic motivation is not equal to those who receive the extrinsic motivation (p-value = .008). The burden to reject the null hypothesis is lower under a one-sided test, so we can say that the evidence supports the claim that the intrinsic mean is higher than the extrinsic mean.
Since this was a randomized experiment, we can conclude that the intrinsic motivation caused this increase. In addition, since these were volunteers, this inference can only be assumed to apply to these 47 subjects, although the findings are very intriguing.

Let's Talk Power!!!

α = Type I error
This is the probability that while the null hypothesis is true, the data in the study cause us to reject the null hypothesis.

Effect size basically measures the difference between the population mean (106) and the null mean(100). (It's not exactly this, though.)



β = Type II error
This is the probability that while the null hypothesis is NOT true, the data in the study cause us to fail to reject the null hypothesis (fail to detect differences in the means).

Power = $1 - \beta$
This is the probability that while the null hypothesis is NOT true, the data in the study correctly cause us to reject the null hypothesis (detect differences in the means).

Explore power!

Here is an applet that will show you what happens to the power/beta when you change the sample size, alpha, standard deviation, or effect size (measure of the difference between null mean and actual (alternative) mean).

http://shiny.stat.tamu.edu:3838/eykolo/power/

(Go to break out)
Consider the following options.

A. The probability of rejecting Ho when the null is true.
B. The probability of accepting Ho when the null is true.
C. The probability of rejecting Ho when the null is false.
D. The probability of FTR Ho when the null is true.
E. The probability of FTR Ho when the null is false.

WHICH IS POWER? C
WHICH IS ALPHA? A
WHICH IS BETA? E

Pick all that are true.
The power increases when:

A. The sample size decreases.
B. The sample size increases.
C. The standard deviation / standard error decreases.
D. The effect size increases.
E. The effect size decreases.

Pick all that are true.
The power increases when:

A. The sample size decreases.
B. The sample size increases.
C. The standard deviation / standard error decreases.
D. The effect size increases.
E. The effect size decreases.

Appendix

Distribution of Sample Average

Proof that $E(\bar{Y}) = \mu$:

$E(\bar{Y}) = E\left(\frac{Y_1 + Y_2 + \ldots + Y_n}{n}\right)$ by the definition of $\bar{Y}$.

$E(\bar{Y}) = \frac{1}{n}E(Y_1 + Y_2 + \ldots + Y_n)$ because n is a constant.

$E(\bar{Y}) = \frac{1}{n}[E(Y_1) + E(Y_2) + \ldots + E(Y_n)]$ because the expected value of a sum of random variables is equal to the sum of the expected values of the random variables.

$E(\bar{Y}) = \frac{1}{n}[\mu + \mu + \ldots + \mu]$ because $E(Y_i) = \mu$.

$$E(\bar{Y}) = \frac{1}{n}[n\mu] = \mu. \blacksquare$$

ANOTHER EXAMPLE FOR PRACTICE

World's Smallest Mammal The world's smallest mammal is the bumblebee bat, also known as the Kitti's hog-nosed bat (or *Craseonycteris thonglongyai*). Such bats are roughly the size of a large bumblebee. Listed below are weights (in grams) from a sample of these bats. Test the claim that these bats come from the same population having a mean weight equal to 1.8 g.

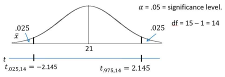
1.7 1.6 1.5 2.0 2.3 1.6 1.6 1.8 1.5 1.7 2.2 1.4 1.6 1.6 1.6

H_0: $\mu = 1.8$
H_1: $\mu \neq 1.8$
$\alpha = 0.05$
$\bar{x} = 1.713$
$s = .2588$

Critical Values $t = \pm 2.145$

```
data critval;
p = quantile("T",.975,14)
;
proc print data = critval;
run;
```

Obs	p
1	2.14479



α = .05 = significance level.

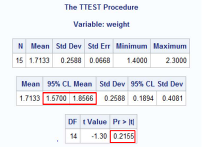
df = 15 – 1 = 14

H_0: $\mu = 1.8$
H_1: $\mu \neq 1.8$
$\alpha = 0.05$
$\bar{x} = 1.713$
$s = .2588$

World's Smallest Mammal The world's smallest mammal is the bumblebee bat, also known as the Kitti's hog-nosed bat (or *Craseonycteris thonglongyai*). Such bats are roughly the size of a large bumblebee. Listed below are weights (in grams) from a sample of these bats. Test the claim that these bats come from the same population having a mean weight equal to 1.8 g.

1.7 1.6 1.5 2.0 2.3 1.6 1.6 1.8 1.5 1.7 2.2 1.4 1.6 1.6 1.6

```
data bats;
 input weight @@;
 datalines;
 1.7 1.6 1.5 2.0 2.3 1.6 1.6 1.8 1.5 1.7 2.2 1.4 1.6 1.6 1.6
 ;

proc print data = bats;
 run;

proc ttest data = bats h0 = 1.8 sides = 2 alpha = .05;
 var weight;
 run;
```



The TTEST Procedure

Variable: weight

N	Mean	Std Dev	Std Err	Minimum	Maximum
15	1.7133	0.2588	0.0668	1.4000	2.3000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1.7133	1.5700	1.8566	0.2588	0.1894	0.4081

DF	t Value	Pr > |t|
14	-1.30	0.2155

On the basis of this test, there is not enough evidence to reject the claim that the mean weight of bumblebee bats is equal to 1.8g (p-value = .2155 from a t-test). A 95% confidence interval is (1.57 g, 1.8566 g). The problem was ambiguous on the randomness of the sample; thus, we will assume that it was not a random sample, which makes inference to all bats strictly speculative.

Part III

A Closer look at Assumptions

Chapter 11

Problem 1: Two Sample T test with assumptions

Question

1. In the United States, it is illegal to discriminate against people based on various attributes. One example is age. An active lawsuit, filed August 30, 2011, in the Los Angeles District Office is a case against the American Samoa Government for systematic age discrimination by preferentially firing older workers. Though the data and details are currently sealed, suppose that a random sample of the ages of fired and not fired people in the American Samoa Government are listed below: Fired 34 37 37 38 41 42 43 44 44 45 45 45 46 48 49 53 53 54 54 55 56 Not fired 27 33 36 37 38 38 39 42 42 43 43 44 44 44 45 45 45 45 46 46 47 47 48 48 49 49 51 51 52 54 a. Check the assumptions (with SAS) of the two-sample t-test with respect to this data. Address each assumption individually as we did in the videos and live session and make sure and copy and paste the histograms, q-q plots or any other graphic you use (boxplots, etc.) to defend your written explanation. Do you feel that the t-test is appropriate? b. Check the assumptions with R and compare them with the plots from SAS. c. Now perform a complete analysis of the data. You may use either the permutation test from HW 1 or the t-test from HW 2 (copy and paste) depending on your answer to part a. In your analysis, be sure and cover all the steps of a complete analysis: 1. State the problem. 2. Address the assumptions of t-test (from part a). 3. Perform the t-test if it is appropriate and a permutation test if it is not (judging from your analysis of the assumptions). 4. Provide a conclusion including the p-value and a confidence interval. 5. Provide the scope of inference.

Answer

11.1 Complete Analysis

Assmuption checking in SAS

The assumptions were tested using proc ttest, which outputs histograms, box plots, QQ-plots, and performs an F-test on the variances. The code used to produce all information in this section is presented below:

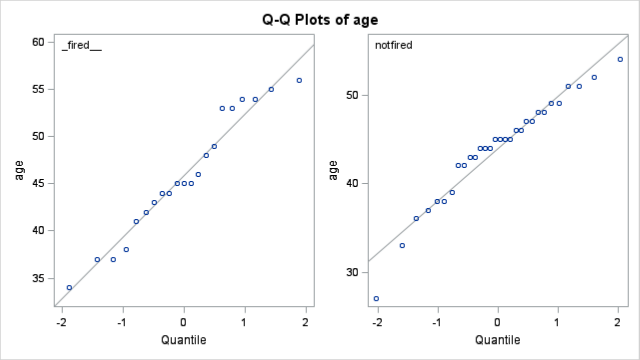
Code 11.1. Checking the assumptions of a t test in SAS

```
proc ttest data=samoa
alpha=.05 test=diff
sides=U; /*an upper tailed test*/
class fired;
var age;
run;
```

Normality

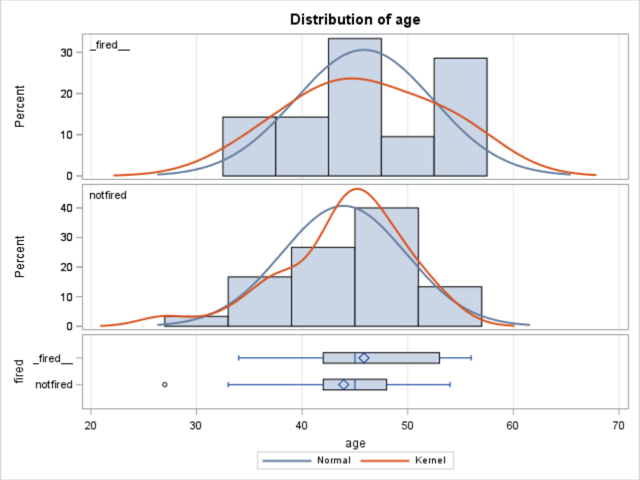
The normality of the data is checked using a QQ plot, a boxplot, and a histogram. First we will examine the QQ plot:

Figure 11.1.1. Q-Q Plot for Normality



In Figure 1.1, the y axis represents the data set, and the x axis the theoretical normal quantile. The line represents what a normal data set should look like, a 1-1 ratio between the data variable and the theoretical normal quantile. The data set follows the normal line pretty well, so in this case on a visual inspection, we can say both samples are normal. We can double check this using Figure 1.2, a histogram and boxplot:

Figure 11.1.2. Histogram and Boxplot for Normality

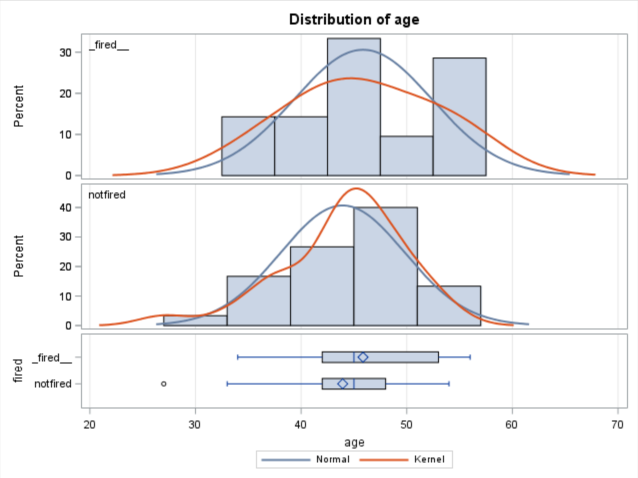


It is a bit harder to assess the normality using the histogram and boxplot, but SAS gives us useful kernel lines which show the distribution of the data in the histogram (the red line is the data and the blue line is normal). As we can see, the data loosely follows the normal distribution, it is a bit different but it is pretty close. The box plot tells the same story, as in both cases the mean is very near the medium (in a normal distribution the mean and median are the same), with slight left and right skewing, but overall we can assume the data is normal.

Equal Variances

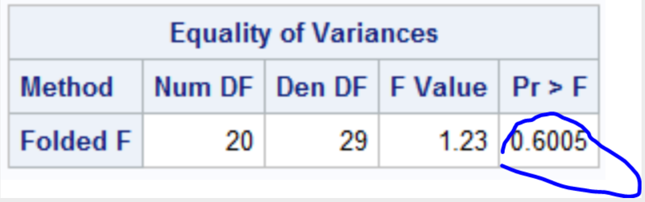
In order to assess the equality of the variances visually, we can again use the histogram and boxplot, this time displayed in Figure 1.3 (for ease of grading):

Figure 11.1.3. Histogram and Boxplot for Variance Equality



As we can see from the bounds of the histogram, the range of each data set is more or less the same size, with their means more or less in the center. This hints that the two data sets would have near equal variances. This is confirmed when looking at the box plot, the distance from the mean to the far left whisker and far right whisker is more or less the same for both data sets, which indicates again the variances are equal. This is confirmed by examining the F test for equal variances, the results of which are displayed below:

Figure 11.1.4. F Test for Equal Variances



Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	20	29	1.23	0.6005

The F test is valid here, because the data is normal and the sample size is large $(n \sim 30)$, and we see that the probability the variance difference is greater than what it is in our case is 60%, or a p value of 0.6 At a 5, 10, 15 or 20 percent confidence interval, the f test will tell us the variances are equal. Therefore, we can assume equal variances.

Independence

In this case, we can assume independence, the two data sets do not relate to each other. Any dependence that exists we will assume away, for the sake of the problem

Conclusion

In my opinion, we can use a t-test for this data set, based on the fact that all the assumptions are true.

Assumption Checking in R

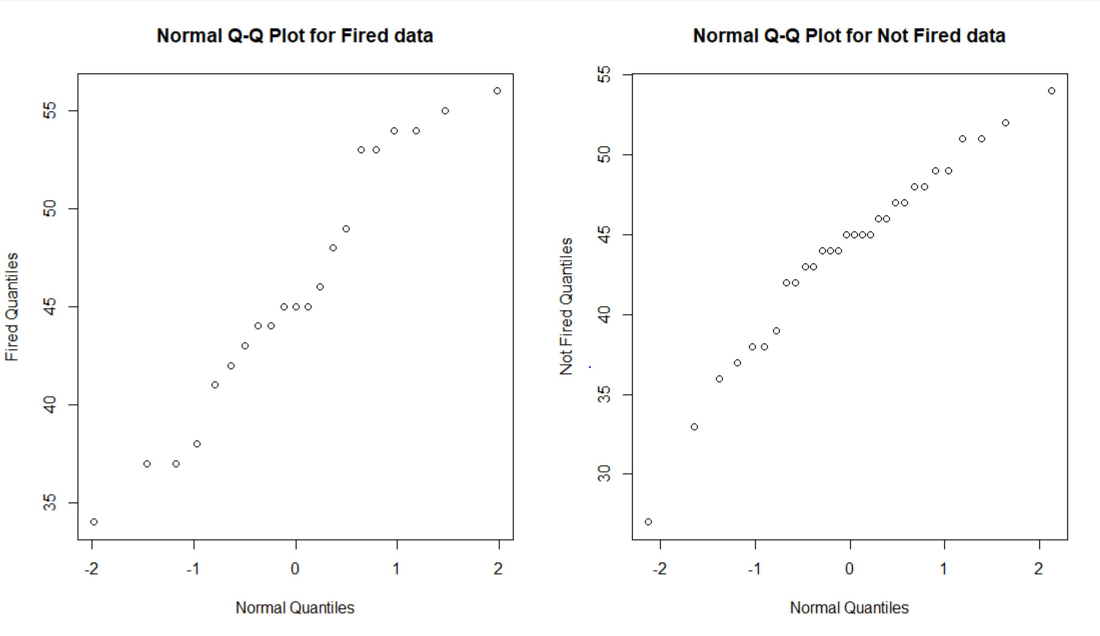
Normality test

To test for normality, we are going to again use the Q-Q plot and the histogram. To produce the Q-Q plots, the following code was used: The plots produced are shown below:

Code 11.2. t test Assumption checking in R, Q-Q plot

```
        #producing adjacent Q-Q plots
        par(mfrow=c(1,2))
        qqnorm(Fired,main="Normal Q-Q Plot for Fired data",
        xlab = "Normal Quantiles",
        ylab = "Fired Quantiles")
        qqnorm(Not_fired,main="Normal Q-Q Plot for Not Fired data",
        xlab = "Normal Quantiles",
        ylab = "Not Fired Quantiles")
```

Figure 11.1.5. Q-Q plots for Normality in R

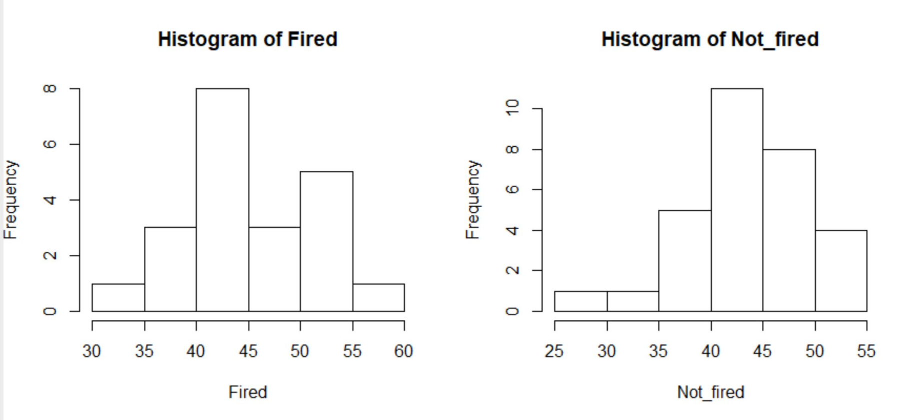


From the linearity of the data points in this figure, we can see that the data follows a more or less normal ditribution. The Q-Q plot produced in R is almost exactly the same as the Q-Q plot produced using SAS, however it is different in that it does not have a lovely line representing perfect normality, and the size of the boxes changes with window size, as does the aspect ratio, which is a bit of a pain. The following code is used to produce a histogram, further examining normality: This produces the following figure:

Code 11.3. t test Assumption checking in R, Histogram

```
        #producing the adjacent histograms
        par(mfrow=c(1,2))
        hist(Fired)
        hist(Not_fired)
```

Figure 11.1.6. Histogram for Normality in R



As can be seen in the figure, the distribution of these two data sets is again more or less normal, with what appears to be the mean and median lying in the center, however there is a bit of a bump in the fired data set, but again it is loosely normal in appearance. The graphs again look the same as in SAS more or less, other than formatting differences. We can identify numbers better in R. In this case, we can ASSUME NORMAL

Equality of Variances

Looking at the histogram in Figure 1.6, we can see that the fired data has a mean of about 45 years old, spanning from 30 to 60, and the not fired data has a mean of about 40 years old, spanning from 25 to 55. The spread of the two means is more or less the same in this case, therefore we can ASSUME EQUAL VARIANCEs

Independence

We can again assume independence.

Conclusion:

The t-test is appropriate

Complete Analysis:

Problem statement:

We would like to test the claim that the mean age of the individuals who were fired is greater than the mean age of the individuals who were not fired.

Assumptions:

We can assume normality, independence, and equal variances and therefore we can use the student t test, as proven in sections 1.a and 1.b.

t-test

Statement of the Hypotheses:

$$H_0 : \mu_f - \mu_{uf} \leq 0$$
$$H_1 : \mu_f - \mu_{uf} > 0$$

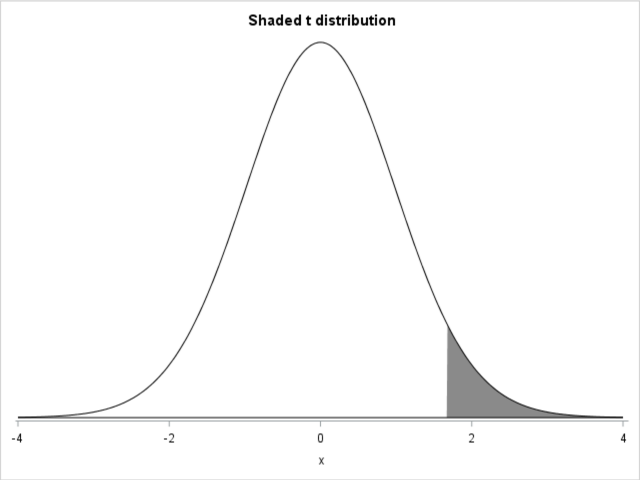
Shaded Distribution and Critical Values: In a two sample t-test, we have that:

$$df = n_f + n_{nf} - 2$$

where in our case, $df = 21+30-2 = 49,\ \alpha = 0.05$ Now we input this information into SAS to draw our distribution[1]:

```
data pdf;
do x = -4 to 4 by .01;
pdf = pdf("T", x, 49);
lower = 0;
if x >= quantile("T",0.9,49) then upper = pdf;/*one sided*/
else upper = 0;
output;
end;
run;
title 'Shaded t distribution';
proc sgplot data=pdf noautolegend noborder;
yaxis display=none;
band x = x
lower = lower
upper = upper / fillattrs=(color=gray8a);
series x = x y = pdf / lineattrs = (color = black);
series x = x y = lower / lineattrs = (color = black);
run;
```

Giving us this lovely graph:



Next we find a number for the critical value, using the same code as problem 1:

```
data critval;
p = quantile("T",.95,49); /*one sided test*/;
proc print data=critval;
run;
```

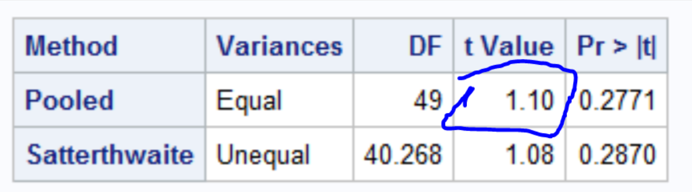
Obs	p
1	1.67655

This gives us a critical t value of 1.67655.

Calculation of t statistic: Next we calculate our two sample t statistic using SAS:

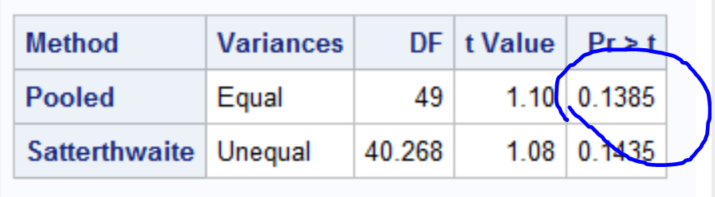
```
proc ttest data=samoa
alpha=.05 test=diff
sides=U;
class fired;
var age;
run;
```

Which tells us that our t statistic is 1.10



Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	49	1.10	0.2771
Satterthwaite	Unequal	40.268	1.08	0.2870

Calculation of P-value With the code from the previous step, we also see the p value:



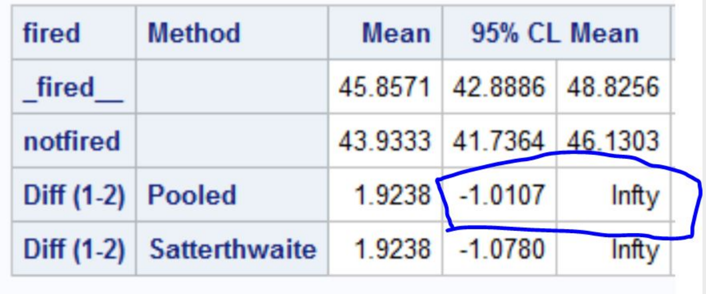
Method	Variances	DF	t Value	Pr > t
Pooled	Equal	49	1.10	0.1385
Satterthwaite	Unequal	40.268	1.08	0.1435

$p = 0.1385$

Discussion of the Null Hypothesis $p = 0.1385 > \alpha = 0.05$ for the one tailed hypothesis test, indicating that we CANNOT REJECT the null hypothesis

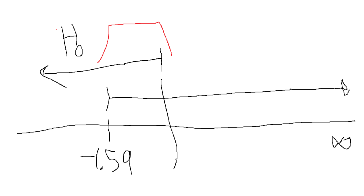
Conclusion:

We cannot reject the null hypothesis, meaning we cannot say that older workers were fired from the Samoan government. Note that we used a one tailed hypothesis test in this scenario, as we wanted to deternine if the fired group was OLDER than the nonfired group. With a one-sided p-value of 0.1385, there is a nearly 14% chance that there be a greater difference in mean ages given the distribution. At a critical p-value of .05 (5%), we can say that this data fails to reject the null hypothesis. Using the code that calculated the t statisitic, we produce the following one sided confidence interval:



fired	Method	Mean	95% CL Mean	
_fired__		45.8571	42.8886	48.8256
notfired		43.9333	41.7364	46.1303
Diff (1-2)	Pooled	1.9238	-1.0107	Infty
Diff (1-2)	Satterthwaite	1.9238	-1.0780	Infty

The confidence interval is: $[-1.0107, \ \infty)$. This confidence interval represents the upper difference of means at a 95% confidence level. We can interpret this as follows: if the confidence interval contains the null hypothesis, then we cannot reject it. However if it does not contain the null hypothesis, we must reject it. As we can see in this beautifully drawn figure, the null hypothesis, $\mu_f - \mu_{nf} \leq 0$ is contained within our CI:



. This means we cannot reject the null hypothesis, we cannot say there was age discrimination. It is plausible that the mean differnence of the entire population of samoan government employees is less than or equal to zero, as it is within the 95% confidence interval, which means we cannot, as objective jurors, claim there was age discrimination.

Scope of Inference:

Since this sample was random, we can make generalizations about the Samoan Government as a whole, however, we cannot make causal inferences, as this was not a randomized experiment.

Chapter 12

Outliers and Logarithmic Transformations

As an example, consider the hypothetical sample: 10, 20, 30, 50, 70. The sample average is 36, and the sample median is 30. Now change the 70 to 700, and what happens? The sample average becomes 162, but the sample median remains 30. The sample average is not a resistant statistic because it can be severely influenced by the change in a single observation. The median, however, is resistant.

Resistance is a desirable property. A resistant procedure is insensitive to outliers. A nonresistant one, on the other hand, may be greatly influenced by one or two outlying observations.

3.3.2 Resistance of *t*-Tools

Since t-tools are based on averages, they are not resistant. A small portion of the data can potentially have a major influence on the results. In particular, one or two outliers can affect a confidence interval or change a p-value enough to completely alter a conclusion.

If the outlier is due to contamination from another population, it can lead to false impressions about the population of interest. If the outlier does come from the population of interest, which happens to be long-tailed, the outcome is still undesirable for the following reason. In statistics, the goal is to describe *group* characteristics. An estimate of the center of a distribution should represent the typical value. The estimate is a good one if it represents the typical values possessed by the great majority of subjects; it is a bad one if it represents a feature unique to one or two subjects. Furthermore, a conclusion that hinges on one or two data points must be viewed as quite fragile.

3.4 PRACTICAL STRATEGIES FOR THE TWO-SAMPLE PROBLEM

Armed with information about the broad set of conditions under which the t-tools work well and the effect of outliers, the challenge to the data analyst is to size up the actual conditions using the available data and evaluate the appropriateness of the t-tools. This involves thinking about possible cluster and serial effects; evaluating the suitability of the t-tools by examining graphical displays; and considering alternatives.

In considering alternatives it is important to realize that even though the t-tools may still be valid when the ideal assumptions are not met, an alternative procedure that is more *efficient* (i.e., makes better use of the data) may be available. For example, another procedure may provide a narrower confidence interval.

Consider Serial and Cluster Effects

To detect lack of independence, carefully review the method by which the data were gathered. Were the subjects selected in distinct groups? Were different groups of subjects treated differently in a way that was unrelated to the primary treatment? Were different responses merely repeated measurements on the same subjects? Were observations taken at different but proximate times or locations? Affirmative answers to any of these questions suggest that independence may be lacking.

The principal remedy is to use a more sophisticated statistical tool. Identifiable clusters, which may be planned or unplanned, can be accounted for through analysis of variance (Chapters 13 and 14) or possibly through regression analysis (Chapters 9–12). Serial effects require time series analysis, the topic of Chapter 15.

Evaluate the Suitability of the t-Tools

Side-by-side histograms or box plots of the two groups of data should be examined and departures from the ideal model should be considered in light of the robustness properties of the t-tools. It is important to realize that the conditions of interest, which are those of the populations, must be investigated through graphical displays of the samples.

If the conditions do not appear suitable for use of the t-tools, then some alternative is necessary. A transformation should be considered if the graphical displays of the transformed data appear to be closer to the ideal conditions. (See Section 3.5.) Alternative tools for analyzing two independent samples are the rank-sum procedure, which is resistant and does not depend on normality (Section 4.2); other permutation tests (Section 4.3.1); and the Welch procedure for comparing normal populations that have unequal standard deviations (Section 4.3.2).

A Strategy for Dealing with Outliers

If investigation reveals that an outlying observation was recorded improperly or was the result of contamination from another population, the solution is to correct it if the right value is known or to leave it out. Often, however, there is no way to know how the outliers arose. Two statistical approaches for dealing with this situation exist. One is to employ a resistant statistical tool, in which case there is no compelling reason to ponder whether the offending observations are natural, the result of contamination, or simply blunders. (The rank-sum procedure in Section 4.2 is resistant.) The other approach is to adopt the careful examination strategy shown in Display 3.6. An important aspect of adopting this procedure is that an outlier does not get swept under the rug simply because it is different from the other observations. To warrant its removal, an explanation for why it is different must be established.

Example—Agent Orange

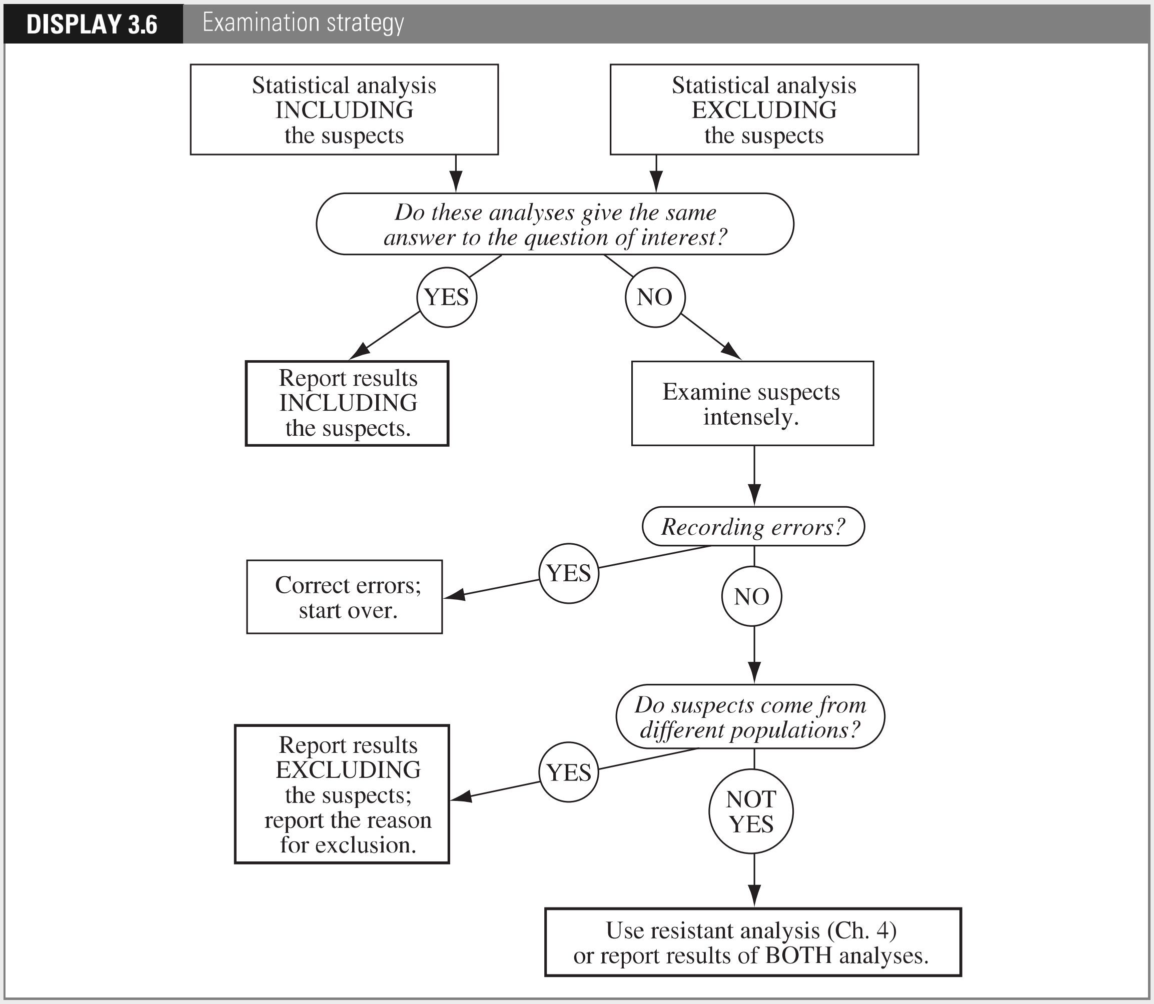
Box plots of dioxin levels in Vietnam and non–Vietnam veterans (Display 3.3) appear again in Display 3.7. The distributions have about the same shape and spread. Although the shape is not normal, the skewness is mild and unlikely to cause any problems with the t-test or the confidence interval. Two Vietnam veterans (#645 and #646) had considerably higher dioxin levels than the others.

From the results listed in Display 3.7 it is evident that the comparison of the two groups is changed very little by the removal of one or both of these outliers. Consequently, there is no need for further action. Even so, it is useful to see what else can be learned about these two, as indicated at the bottom of the display.

Notes

1. It is not useful to give a precise definition for an *outlier*. Subjective examination is the best policy. If there is any doubt about whether a particular observation deserves further examination, give it further examination.

DISPLAY 3.6 Examination strategy



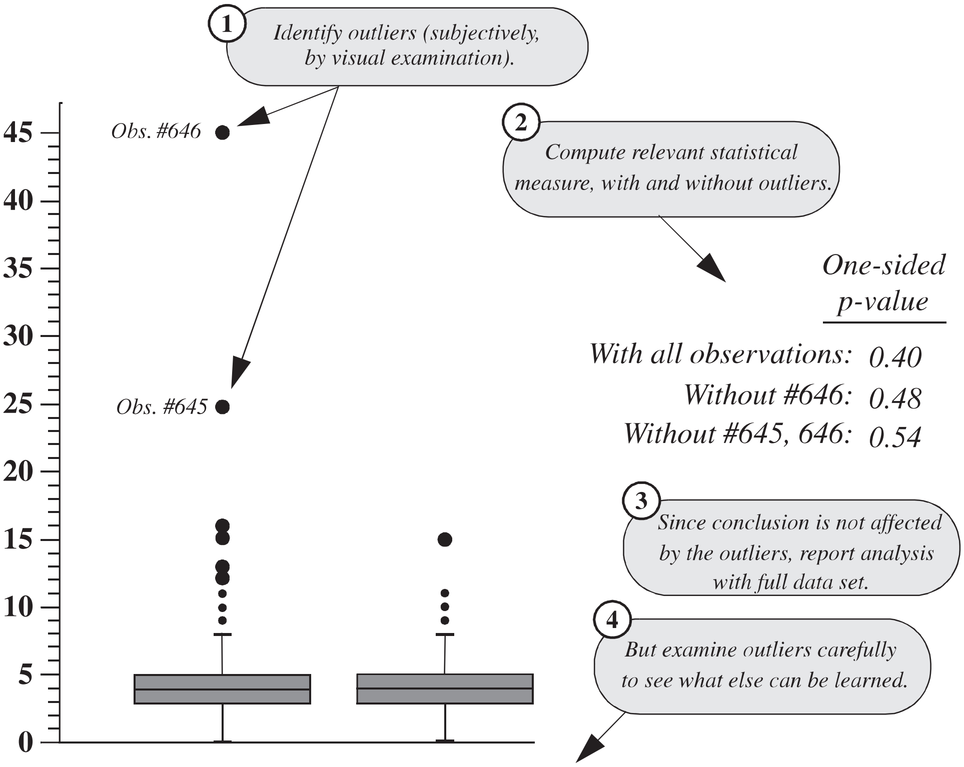
2. It is not surprising that the outliers in the Agent Orange example have little effect, since the sample sizes are so large.
3. The apparent difference in the box plots may be due to the difference in sample sizes. If the population distributions are identical, more observations will appear in the extreme tails from a sample of size 646 than from a sample of size 97.

3.5 TRANSFORMATIONS OF THE DATA

3.5.1 The Logarithmic Transformation

The most useful transformation is the *logarithm* (log) for positive data. The common scale for scientific work is the *natural* logarithm (ln), based on the number

DISPLAY 3.7 Outlier analysis for Agent Orange data: effect of outliers on the p-value, for equal population means



① *Identify outliers (subjectively, by visual examination).*

② *Compute relevant statistical measure, with and without outliers.*

	One-sided p-value
With all observations:	*0.40*
Without #646:	*0.48*
Without #645, 646:	*0.54*

③ *Since conclusion is not affected by the outliers, report analysis with full data set.*

④ *But examine outliers carefully to see what else can be learned.*

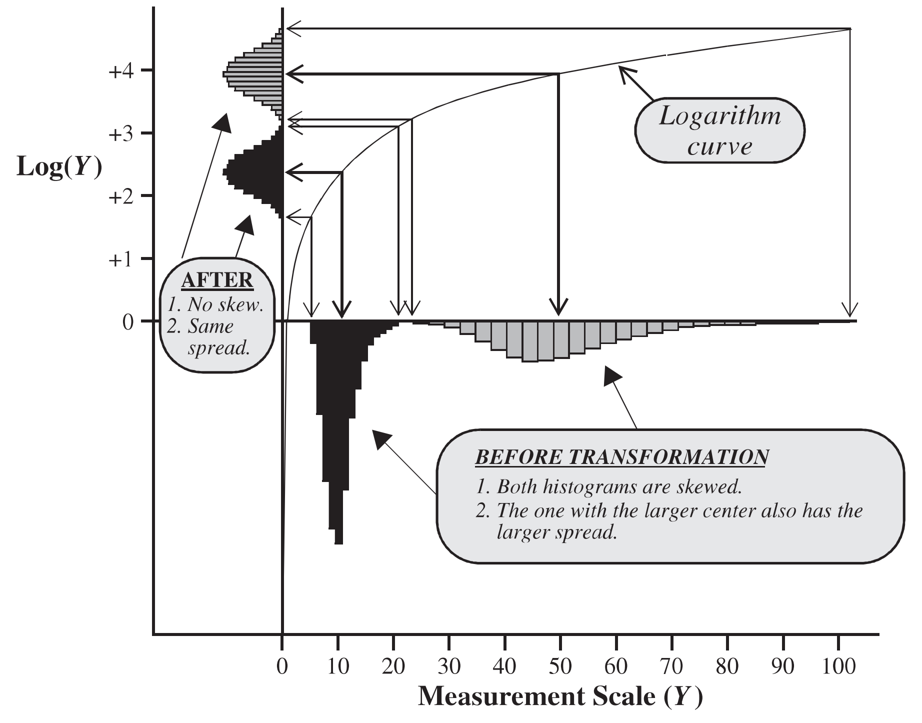
Veteran # 645: reported 180 days of indirect military exposure to herbicides.
Veteran # 646: reported no exposure (military or civilian) to herbicides.

$e = 2.71828\ldots$. The logarithm of e is unity, denoted by $\log(e) = 1$. Also, the log of 1 is 0: $\log(1) = 0$. The general rule for using logarithms is that $\log(e^x) = x$. Another choice is the *common* logarithm based on the number 10, rather than e. Common logs are defined by $\log_{10}(10^x) = x$. Unless otherwise stated, *log* in this book refers to the natural logarithm.

Recognizing the Need for a Log Transformation

The data themselves usually suggest the need for a log transformation. If the ratio of the largest to the smallest measurement in a group is greater than 10, then the data are probably more conveniently expressed on the log scale. Also, if the graphical displays of the two samples show them both to be skewed and if the group with the larger average also has the larger spread (see Display 3.2), the log transformation is likely to be a good choice.

DISPLAY 3.8 The logarithmic transformation used to arrive at favorable conditions for the two-sample t-analysis



Display 3.8 illustrates the behavior of the log transformation. On the scale of measurement Y the two groups have skewed distributions with longer tails in the positive direction. The group with the larger center also has the larger spread. The measurements on the transformed scale have the same ordering, but small numbers get spread out more, while large numbers are squeezed more closely together. The overall result is that the two distributions on the transformed scale appear to be symmetric and have equal spread—just the right conditions for applying the t-tools.

3.5.2 Interpretation After a Log Transformation

For some measurements, the results of an analysis are appropriately presented on the transformed scale. Most users feel comfortable with the Richter scale for measuring earthquake strength, even though it is a logarithmic scale. Similarly, pH as a measure of acidity is the negative log of ion concentration. In other cases, however, it may be desirable to present the results on the original scale of measurement.

Randomized Experiment Model: Multiplicative Treatment Effect

If the randomized experiment model with additive treatment effect is thought to hold for the log-transformed data, then an experimental unit that would respond to treatment 1 with a logged outcome of $\log(Y)$ would respond to treatment 2 with a logged outcome of $\log(Y) + \delta$. By taking antilogarithms of these two quantities, one finds that an experimental unit that would respond to treatment 1 with an outcome of Y would respond to treatment 2 with an outcome of Ye^{δ}. Thus, e^{δ} is the *multiplicative treatment effect* on the original scale of measurement. To test whether there is any treatment effect, one performs the usual t-test for the hypothesis that δ is zero with the log-transformed data. To describe the multiplicative treatment effect, one back-transforms the estimate of δ and the endpoints of the confidence interval for δ.

Interpretation After Log Transformation (Randomized Experiment)

Suppose $Z = \log(Y)$. *It is estimated that the response of an experimental unit to treatment 2 will be* $\exp(\overline{Z}_2 - \overline{Z}_1)$ *times as large as its response to treatment 1.*

Example—Cloud Seeding

Display 3.2 shows that the log-transformed rainfalls have distributions that appear satisfactory for using the t-tools; so in Display 3.9 a full analysis is carried out on the log scale. Tests and confidence intervals are constructed in the usual way but on the transformed data. The estimate of the additive treatment effect on log rainfall is back-transformed to an estimate of the multiplicative effect of cloud seeding on rainfall.

Population Model: Estimating the Ratio of Population Medians

The t-tools applied to log-transformed data provide inferences about the difference in means of the logged measurements, which may be represented as $\text{Mean}[\log(Y_2)] - \text{Mean}[\log(Y_1)]$, where $\text{Mean}[\log(Y_2)]$ symbolizes the mean of the logged values of population 2. A problem with interpretation on the original scale arises because the mean of the logged values is not the log of the mean. Taking the antilogarithm of the estimate of the mean on the log scale does *not* give an estimate of the mean on the original scale.

If, however, the log-transformed data have symmetric distributions, the following relationships hold:

$$\text{Mean}[\log(Y)] = \text{Median}[\log(Y)]$$

(and since the log preserves ordering)

$$\text{Median}[\log(Y)] = \log[\text{Median}(Y)],$$

where $\text{Median}(Y)$ represents the *population median* (the 50th percentile of the population). In other words, the 50th percentile of the logged values is the log of the 50th percentile of the untransformed values. Putting these two equalities together,

DISPLAY 3.9 Two-sample t-analysis and statement of conclusions after logarithmic transformation—cloud seeding example

① **Transform the data.**

Unseeded		Seeded	
Y (acre-ft)	log (Y)	Y (acre-ft)	log (Y)
1202.6	7.092	2745.6	7.918
830.1	6.722	1697.8	7.437
372.4	5.920	1656.0	7.412
345.5	5.845	978.0	6.886
321.2	5.772	703.4	6.556
244.3	5.498	489.1	6.193
163.0	5.094	430.0	6.064
147.8	4.996	334.1	5.811
95.0	4.554	302.8	5.713
87.0	4.466	274.7	5.616
81.2	4.397	274.7	5.616
68.5	4.227	255.0	5.541
47.3	3.857	242.5	5.491
41.1	3.716	200.7	5.302
36.6	3.600	198.6	5.291
29.0	3.367	129.6	4.864
28.6	3.353	119.0	4.779
26.3	3.270	118.3	4.773
26.1	3.262	115.3	4.748
24.4	3.195	92.4	4.526
21.7	3.077	40.6	3.704
17.3	2.851	32.7	3.487
11.5	2.446	31.4	3.447
4.9	1.589	17.5	2.862
4.9	1.589	7.7	2.041
1.0	0.000	4.1	1.411

② **Use the two-sample t-tools on the log rainfall.**

Difference in averages = 1.1436 (SE = 0.4495).

Test of the hypothesis of no effect of cloud seeding on log rainfall: one-sided p-value from two-sample t-test = 0.0070 (50 d.f.).

95% confidence interval for additive effect of cloud seeding on log rainfall: 0.2406 to 2.0467.

③ **Back-transform estimate and confidence interval.**

Estimate = $e^{1.1436} = 3.1382$

Lower confidence limit = $e^{0.2406} = 1.2720$.

Upper confidence limit = $e^{2.0467} = 7.7425$.

④ **State the conclusions on the original scale.**

Conclusion: *There is convincing evidence that seeding increased rainfall (one-sided p-value = 0.0070). The volume of rainfall produced by a seeded cloud is estimated to be 3.14 times as large as the volume that would have been produced in the absence of seeding (95% confidence: 1.27 to 7.74 times).*

it is evident that the antilogarithm of the mean of the log values is the median on the original scale of measurements.

If $\overline{Z}_1$ and $\overline{Z}_2$ are used to represent the averages of the logged values for samples 1 and 2, then $\overline{Z}_2 - \overline{Z}_1$ estimates $\log[Median(Y_2)] - \log[Median(Y_1)]$, and therefore

$$\overline{Z}_2 - \overline{Z}_1 \text{ estimates } \log\left[\frac{\text{Median}(Y_2)}{\text{Median}(Y_1)}\right]$$

and, therefore,

$$\exp(\overline{Z}_2 - \overline{Z}_1) \text{ estimates } \left[\frac{\text{Median}(Y_2)}{\text{Median}(Y_1)}\right].$$

The point of this is that a very useful multiplicative interpretation emerges in terms of the ratio of population medians. This is doubly important because the median is a better measure of the center of a skewed distribution than the mean. The multiplicative nature of this relationship is captured with the following wording:

Interpretation After Log Transformation (Observational Study)

It is estimated that the median for population 2 is $\exp(\overline{Z}_2 - \overline{Z}_1)$ *times as large as the median for population 1.*

In addition, back-transforming the ends of a confidence interval constructed on the log scale produces a confidence interval for the ratio of medians.

Example (Sex Discrimination)

Although the analysis of the sex discrimination data of Section 1.1.2, was suitable on the original scale of the untransformed salaries, graphical displays of the log-transformed salaries indicate that analysis would also be suitable on the log scale. The average male log salary minus the average female log salary is 0.147. Since $e^{0.147} = 1.16$, it is estimated that the median salary for males is 1.16 times as large as the median salary for females. Equivalently, the median salary for males is estimated to be 16% more than the median salary for females. Since a 95% confidence interval for the difference in means on the log scale is 0.100 to 0.194, a 95% confidence interval for the ratio of population median salaries is 1.11 to 1.21 ($e^{0.100}$ to $e^{0.194}$). With 95% confidence, it is estimated that the median salary for males is between 11% and 21% greater than the median salary for females.

3.5.3 Other Transformations for Positive Measurements

There are other useful transformations for positive measurements with skewed distributions where the means and standard deviations differ between groups. The *square root* transformation $\sqrt{Y}$ applies to data that are counts—counts of bacteria clusters in a dish, counts of traffic accidents on a stretch of highway, counts of red giants in a region of space—and to data that are measurements of area. The *reciprocal* transformation $1/Y$ applies to data that are waiting times—times to failure of lightbulbs, times to recurrence for cancer patients treated with radiation, reaction times to visual stimuli, and so on. The reciprocal of a time measurement can often be interpreted directly as a rate or a speed. The *arcsine square root* transformation, $\text{arcsine}(\sqrt{Y})$, and the *logit* transformation, $\log[Y/(1-Y)]$, apply when the measurements are proportions between zero and one—proportions of trees infested by a wood-boring insect in experimental plots, proportions of weight lost as a side effect of leukemia therapy, proportions of winning lottery tickets in clusters of a certain size, and so forth.

Only the log transformation, however, gives such ease in converting inferences back to the original scale of measurement. One may estimate the difference in means of $\sqrt{Y_2}$ and $\sqrt{Y_1}$, but the square of this difference does not make much sense on the original scale.

Choosing a Transformation

Formal statistical methods are available for selecting a transformation. Nevertheless, it is recommended here that a trial-and-error approach, with graphical analysis, be used instead. For positive data in need of a transformation, the logarithm should almost always be the first tried. If it is not satisfactory, the reciprocal or the square root transformations might be useful. Keep in mind that the primary goal is to establish a scale where the two groups have roughly the same spread. If several transformations are similar in their ability to accomplish this, think carefully about which one offers the most convenient interpretation.

Caveat About the Log Transformation

Situations arise where presenting results in terms of population medians is not sufficient. For example, the daily emissions of dioxin in the effluent from a paper mill have a very skewed distribution. An agency monitoring the emissions will be interested in estimating the total dioxin load released during, say, a year of operation. The total dioxin load would be the population mean times the population size, and therefore is estimated by the sample average times the population size. It cannot be estimated directly from the median, unless more specific assumptions are made.

3.6 RELATED ISSUES

3.6.1 Prefer Graphical Methods Over Formal Tests for Model Adequacy

Formal tests for judging the adequacy of various assumptions exist. Tests for normality and tests for equal standard deviation are available in most statistical computer programs, as are tests that determine whether an observation is an outlier. Despite their widespread availability and ease of use, these diagnostic tests are not very helpful for model checking. They reveal little about whether the data meet the broader conditions under which the tools work well. The fact that two populations are not exactly normal, for example, is irrelevant. Furthermore, the formal tests themselves are often not very robust against their own model assumptions. Graphical displays are more informative, if less formal. They provide a good indication of whether or not the data are amenable to t-analysis and, if not, they often suggest a remedy.

3.6.2 Robustness and Transformation for Paired *t*-Tools

The one-sample t-test, of which the paired t-test is a special case, assumes that the observations are independent of one another and come from a normally distributed population. P-values and confidence intervals remain valid for moderate and large sample sizes for nonnormal distributions. For smaller sample sizes skewness can be a problem. When cluster or serial effects are present (see Section 3.2.4), the t-tools may give misleading results. When the observations within each pair are positive, either an apparent multiplicative treatment effect (in an experiment) or a tendency for larger differences in pairs with larger average values suggests the use of a log transformation. The transformation is applied before taking the difference, which is equivalent to forming a ratio within each pair and performing a one-sample analysis on the logarithms of the ratios. If there are n pairs, let $Z_i = \log(Y_{1i}) - \log(Y_{2i})$, which is the same as $\log(Y_{1i}/Y_{2i})$. In an observational study, $\exp(\overline{Z})$ is an estimate of the median of the ratios, Y_1/Y_2. (This is not the same as the ratio of the medians [see Exercise 20].) In a randomized, paired experiment, $\exp(\overline{Z})$ estimates a multiplicative treatment effect on the original scale. In both cases, the statistical work of testing and constructing a confidence interval is done on the log scale. The estimate and associated interval are transformed back to the original scale.

3.6.3 Example—Schizophrenia

In the schizophrenia example of Section 2.1.2, Z_i represents the logarithm of the left hippocampus volume of the unaffected twin divided by the left hippocampus volume of the affected twin in pair i. The average of the 15 log ratios is 0.1285. A one-sample analysis gives a p-value of 0.0065 for the test that the mean is zero and a 95% confidence interval from 0.0423 to 0.2147 for the mean itself. Taking antilogarithms of the estimate and the endpoints of the confidence interval yields the following conclusion: It is estimated that the median of the unaffected-to-affected volume ratios is 1.137. A 95% confidence interval for the median ratio is from 1.043 to 1.239.

3.7 SUMMARY

Cloud Seeding and Rainfall Study

The box plots of the rainfalls for seeded and unseeded days reveal that the two distributions of rainfall are skewed and that the distribution with the larger mean also has the larger variance. This is the situation where log-transformed data behave in accordance with the ideal model. A plot of the data after transformation confirms the adequacy of the transformation. The two-sample t-test can be used as an approximation to the randomization test, and the difference in averages (of log rainfall) can be back-transformed to provide a statement about a multiplicative treatment effect. In the example, it is estimated that the rainfall is 3.1 times as much when a cloud is seeded as when it is left unseeded.

Since randomization is used, the statistical conclusion implies that the seeding causes the increase in rainfall. Since the decision about whether to seed clouds is determined (in this case) by a random mechanism, and since the airplane crew is *blind* to which treatment they are administering, human bias can have had little influence on the result.

Agent Orange Study

Graphical analysis focuses attention on the possibly undue influence of two outliers, but analyses with and without the outliers reveal no such influence, so the t-tools are used on the entire data set. The form of the sampling from the populations of living Vietnam veterans and of other veterans is a major concern in accepting the reliability of the statistical analysis. Protocols for obtaining the samples have not been discussed here, except to note that random sampling is not being used. Conclusions based on the two-sample t-test are supplied, along with the caveat that there may be biases due to the lack of random sampling.

3.8 EXERCISES

Conceptual Exercises

1. Cloud Seeding. What is the experimental unit in the cloud seeding experiment?

2. Cloud Seeding. Randomization in the cloud seeding experiment was crucial in assessing the effect of cloud seeding on rainfall. Why?

3. Cloud Seeding. Why was it important that the airplane crew was unaware of whether seeding was conducted or not?

4. Cloud Seeding. Why would it be helpful to have the date of each observed rainfall?

5. Agent Orange. How would you respond to the comment that the box plots in Display 3.3 indicate that the dioxin levels in the Vietnam veterans tend to be larger since their values appear to be larger?

6. Agent Orange. (a) What course of action would you propose for the statistical analysis if it was learned that Vietnam veteran #646 (the largest observation in Display 3.6) worked for several years, after Vietnam, handling herbicides with dioxin? (b) What would you propose if this was learned instead for Vietnam veteran #645?

7. Agent Orange. If the statistical analysis had shown convincing evidence that the mean dioxin levels differed in Vietnam veterans and other veterans, could one conclude that serving in Vietnam was responsible for the difference?

8. Schizophrenia. In the schizophrenia study in Section 2.1.2, the observations in the two groups (schizophrenic and nonschizophrenic) are not independent since each subject is matched with a twin in the other group. Did the researchers make a mistake?

9. True or false? A statistical computer package will only print out a p-value or confidence interval if the conditions for its validity are met.

10. True or false? A sample histogram will have a normal distribution if the sample size is large enough.

The permutation test was performed using the following code: We will now perform the same procedure on the assumptions without an outlier, as well as some other comparisons. Unless otherwise noted, the following code was used to produce the results and to remove outliers:

Code 12.1. Automatically input permutation test in SAS

```
/*Permutation test*/
data Wallet;
INFILE 'file location';
INPUT school $ cash;
run;
proc iml;
use Wallet var {school cash};
/*making two groups in IML*/
read all var {cash} where(school='SMU') into g1;
read all var {cash} where(school='SEU') into g2;
obsdiff = mean(g1) - mean(g2);
print obsdiff;
call randseed(12345);          /* set random number seed */
alldata = g1 // g2;            /* stack data in a single vector */
N1 = nrow(g1);
N = N1 + nrow(g2);
NRepl = 9999;             /* number of permutations */
nulldist = j(NRepl,1);      /* allocate vector to hold results */
do k = 1 to NRepl;
x = sample(alldata, N, "WOR");     /* permute the data */
nulldist[k] = mean(x[1:N1]) - mean(x[(N1+1):N]);  /* difference of means */
end;
title "Histogram of Null Distribution";
refline = "refline " + char(obsdiff) + " / axis=x lineattrs=(color=red);";
call Histogram(nulldist) other=refline;
pval = (1 + sum(abs(nulldist) >= abs(obsdiff))) / (NRepl+1);
/*this means two sided test*/
print pval;
run;
```

Code 12.2. Outlier removal in SAS

```
data Wallet;
INFILE 'file location';
INPUT school \$ cash;
run;
data CleanCash;
set Wallet;
/*we are going to remove all the really high values*/
if cash >150 then delete;
run;
proc ttest data=CleanCash
alpha=.05 test=diff
sides=2; /*a 2 tailed test*/
class school;
var cash;
run;
```

Chapter 13

Log Transformed data

13.1 Full Analysis

Problem Statement:

We would like to test the claim that the distribution of incomes for those who have 16 years of education is greater than those who have 12 years of education.

Assumptions

We first produce the plots for our assumption analysis using the following bit of code:

```
proc import
/*to use proc import first we specify the file*/
datafile='genericfilepath/genericname.csv'
/*then we specify the name of the output dataset*/
out=edudata /*then we specify the data type*/
dbms=CSV;
run;
proc sort data=edudata;
by descending educ;
run;
proc ttest data=edudata
order=DATA /*This changes theorder of the groups you are using to the one you set*/
sides=U; /*an Upper tailed test*/
class Educ;
var Income2005;
run;
```

Producing the following figures:

Figure 13.1.1. Q-Q plot of sample

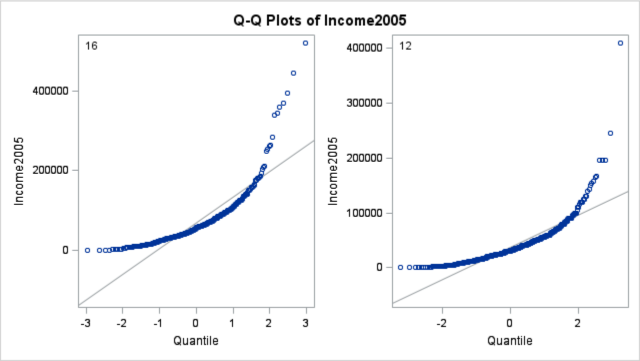
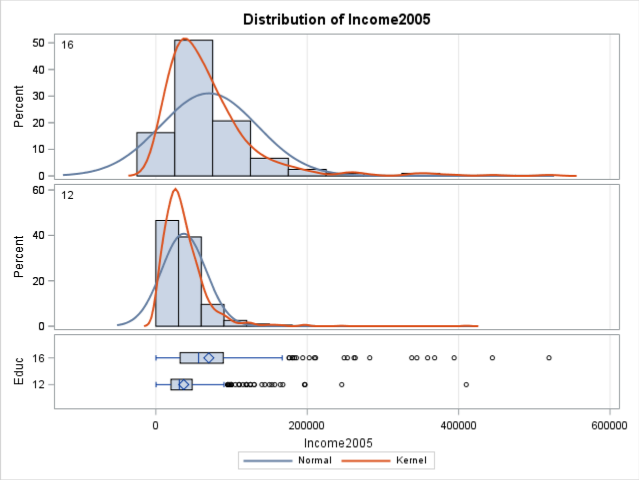


Figure 13.1.2. Histogram and Boxplot of the sample



Normality assumption:

Looking at the Q-Q plot(Figure 3.1), it is clear to see that the data is not normal at all. To investigate further, we will look at the histograms and box plots in Figure 3.2. These paint a more complete picture, we see that the data is skewed to the right, and that the higher values are much greater than the lower values (hundreds of thousands of times). To combat this, lets perform a natural log transformation with this bit of code and see whatthe data looks like:

Code 13.1. log transform in SAS

```
data edudata2;
set edudata;
lincome=log(Income2005);
run;
proc ttest data=edudata2
order=DATA sides=U; /*an Upper tailed test*/
class Educ;
var lincome;
run;
```

Producing the following figures:

Figure 13.1.3. Q-Q plot of logs

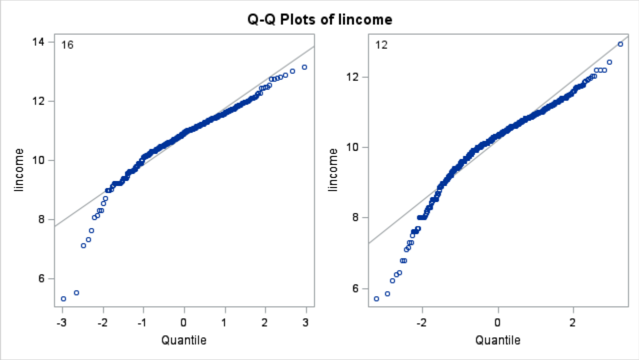
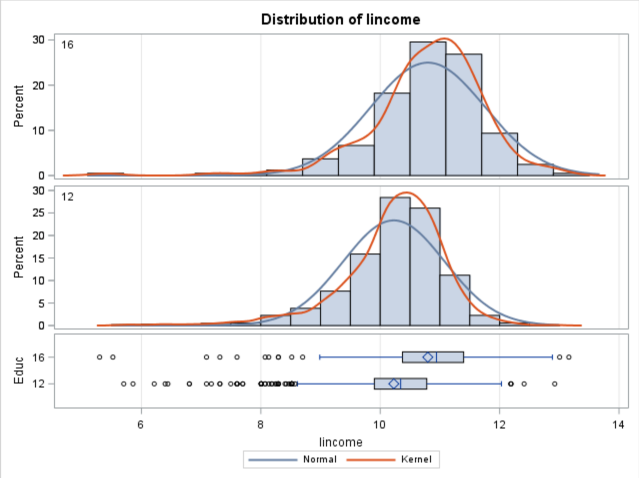


Figure 13.1.4. Histogram and Boxplot of Logs



With this transformation, we first look at the Q-Q plot (Figure 3.3), and we see that the data is mostly normal! Looking at the histograms (Figure 3.4) this is confirmed, just in their shape and the shape of the kernel density plots. The nearness of the median to the mean is also a telltale sign the data is normal. Therefore, we can assume the log-transformed data is normal.

Equality of Variances

Since we cannot assume normality with the untransformed data, it makes little sense to analyze the equality of variances of that data set. We will look at the log transformed data for the equality of variances. Looking at figure 3.4, we see that the spread of the two data sets is pretty similar, just in the histograms, they are of similar length, where the 12 year data set is a bit narrowerthan the 16 year set. The Boxplot confirms this, the distance from the means to the end of the whiskers is roughly the same for both plots, as well as within the IQRS. The one with the larger mean also has a larger variance, Therefore, we can assume the log transformed data has equal variances.

Independence

We can assume the data is independent in this scenario.

3.3 Hypothesis testing

We will be using a one tailed pooled t test of the log transformation of the data in this scenario, so that we can do a t test

Statement of Hypotheses:

Note that since we are dealing with a pooled t-test of a log transformation, we are dealing in medians rather than means, the medians should tell us whether or not the distribution of the people with 16 years of education exceeds that of those with 12 years of education

$$\begin{aligned} H_0 :& Median_{16} = Median_{12} \\ H_1 :& Median_{16} > Median_{12} \end{aligned}$$

$$\begin{aligned} H_0 :& \ distribution_{16} = distribution_{12} \\ H_1 :& \ distribution_{16} > distribution_{12} \end{aligned}$$

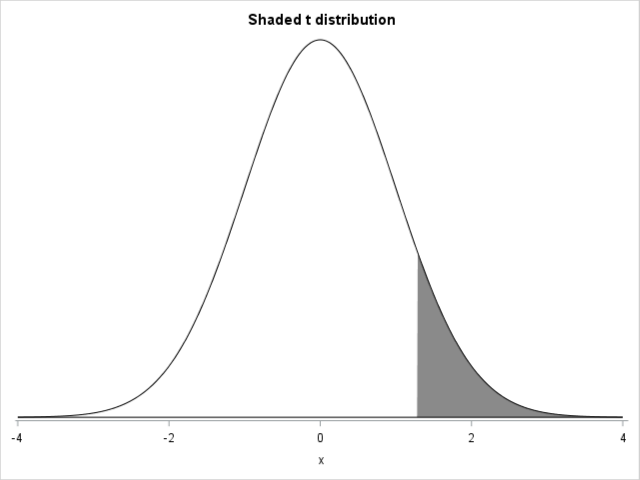
Critical Value

In this scenario, $\alpha = 0.1$ and $df = 1424$, and from that we can shade a one sided distribution and find a critical value, using the code below:

```
data pdf;
do x = -4 to 4 by .01;
pdf = pdf("T", x, 1424);
lower = 0;
if x >= quantile("T",0.9,1424) then upper = pdf;/*one sided*/
else upper = 0;
output;
end; run;
title 'Shaded t distribution';
proc sgplot data=pdf noautolegend noborder;
yaxis display=none;
band x = x
lower = lower
upper = upper / fillattrs=(color=gray8a);
series x = x y = pdf / lineattrs = (color = black);
series x = x y = lower / lineattrs = (color = black);
run;
data critval;
p = quantile("T",.9,1424); /*one sided test*/;
proc print data=critval; run;
```

This produces the shaded distribution:

Figure 13.1.5. Shaded t distribution

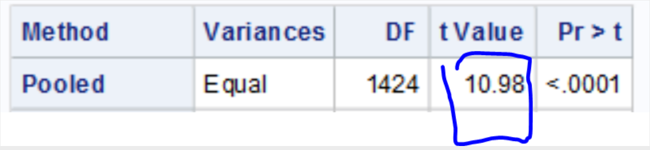


and a critical value of $t = 1.28215$

Obs	p
1	1.28215

Calculation of the t statistic:

Now we calculate our t statististic using the code from Section 3.2.1, which tells us that $t = 10.98$, which is an astounding value!



Method	Variances	DF	t Value	Pr > t
Pooled	Equal	1424	10.98	<.0001

Calculation of the p-value:

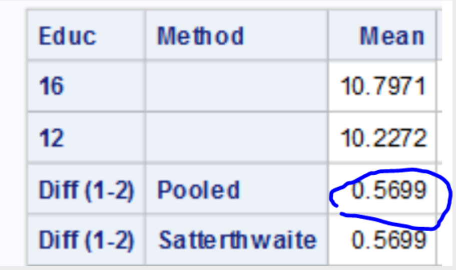
$p < 0.0001$, see the figure above!

3.3.5 Discussion of the Null hypothesis

We REJECT the null hypothesis, $p \approx 0 < 0.1 = \alpha$

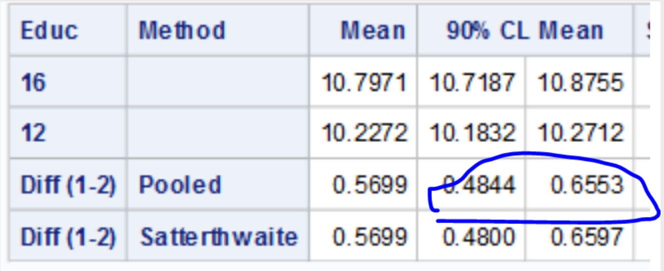
Conclusion

We Reject the null hypothesis which states that the two distributions are equal. We have convincing evidence that the income distribution of the people with 16 years of education is greater than those with 12. With a one-sided p value of ~0, the distributions are very different, the median income of the people with a 16 year education is evidently greater than the median income of people with a 12 year education. The figure below shows the difference between the natural logarithm of the two medians:



Educ	Method	Mean
16		10.7971
12		10.2272
Diff (1-2)	Pooled	0.5699
Diff (1-2)	Satterthwaite	0.5699

This tells us that the median income of people with 16 years education is $e^{0.5699} = 1.77$ times greater than those with 12 years of education. A 90% confidence interval for this multiplicative effect is 1.62 to 1.93 times.



Educ	Method	Mean	90% CL Mean	
16		10.7971	10.7187	10.8755
12		10.2272	10.1832	10.2712
Diff (1-2)	Pooled	0.5699	0.4844	0.6553
Diff (1-2)	Satterthwaite	0.5699	0.4800	0.6597

We cannot make causal inferences in this scenario, as there was no random experimentation, and we cannot make population inferences either, as there was no random sampling

Chapter 14

Unit 3 Lecture slides

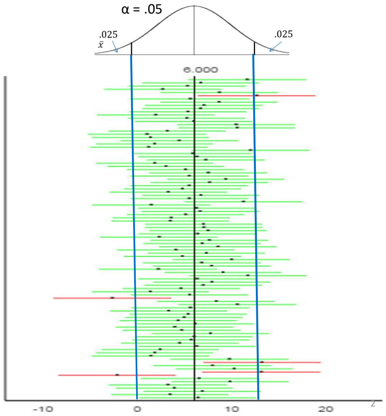
Chapter 3

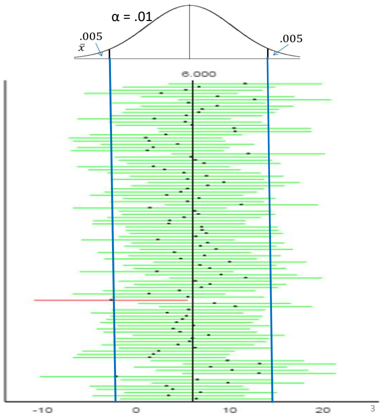
A Closer Look at Assumptions!

Confidence Intervals and Hypothesis Tests

95% CI
Vs.
α = .05 Hyp Test

For the corresponding alpha, a (1-alpha)% CI will contain mu_0 when the test of Ho: mu = mu_0 fails to reject Ho and will not contain mu_0 when the test rejects Ho.





Confidence Intervals and Hypothesis Tests

99% CI
Vs.
α = .01 Hyp Test

The Take Away

Two-Sided 100(1-α)% Confidence Intervals are Equivalent to Two-Tailed Hypothesis Tests that have an α level of significance.

“Equivalent” here means that if we test any specific value in the interval, the test will FTR Ho. And if we test any specific value outside the interval, the test will Reject Ho.

Example:
95% confidence interval for the mean is equivalent to an α = .05 hypothesis test.

Example:
99% confidence interval for the mean is equivalent to an α = .01 level hypothesis test.

So we can evaluate hypothesis tests through the evaluation of confidence intervals!

Assumptions of one sample T-Tests

1. Samples are drawn from a **normally** distributed population.
2. The observations in the sample are independent of one another.

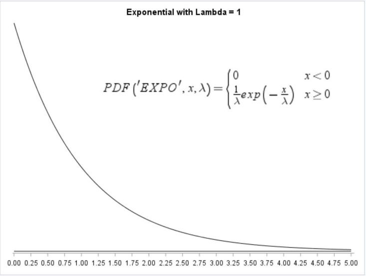
Robustness of One Sample T-test / CI

When the original (population) distribution is not normal, the one sample t-test is still valid with a large enough sample size. (Central Limit Theorem)

That is, the one sample t-test is robust to the normality assumption when the sample size is large enough.

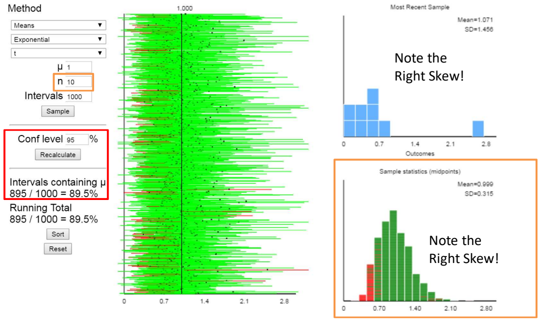
Assume the population distribution is Exponential.

With λ= 1.



$$PDF('EXPO', x, \lambda) = \begin{cases} 0 & x < 0 \\ \frac{1}{\lambda} exp\left(-\frac{x}{\lambda}\right) & x \geq 0 \end{cases}$$

1000 CIs for the Mean of an Exponential(1) Distribution: n = 10



Method
Means
Exponential
t
μ 1
n 10
Intervals 1000

Conf level 95 %

Intervals containing μ
895 / 1000 = 89.5%

Running Total
895 / 1000 = 89.5%

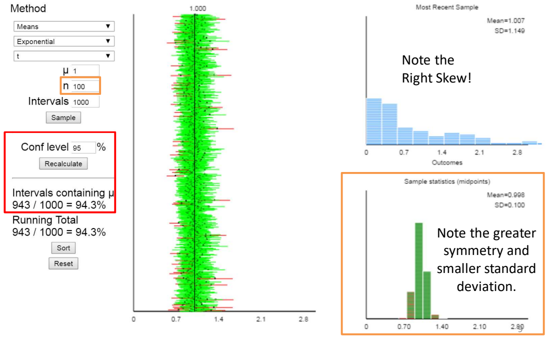
Mean=1.071
SD=1.456

Note the Right Skew!

Mean=0.999
SD=0.315

Note the Right Skew!

1000 CIs for the Mean of an Exponential(1) Distribution: n = 100



Method
Means
Exponential
t
μ 1
n 100
Intervals 1000

Conf level 95 %

Intervals containing μ
943 / 1000 = 94.3%

Running Total
943 / 1000 = 94.3%

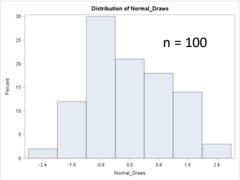
Mean=1.007
SD=1.149

Note the Right Skew!

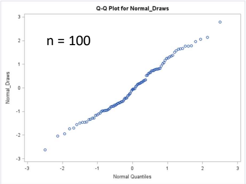
Mean=0.998
SD=0.100

Note the greater symmetry and smaller standard deviation.

Given Data, How Do We Check the Normality Assumption? Visually!



Histogram



q-q Plot

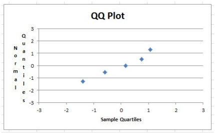
```
/* Generate Normal Random Draws */
data Normal(keep = Normal_Draws);
call streaminit(14);
do i = 1 to 100;
Normal_Draws = rand("Normal");
output;
end;
run;
```

```
proc univariate data = Normal;
var Normal_Draws;
histogram Normal_Draws;
qqplot Normal_Draws;
run;
```

Normal q-q Plot

DATA
41.2
76.6
109.3
134.5
148.6

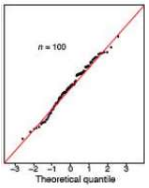
data	rank	middle = (rank + previous rank)/2n	standard normal hypothetical value based on middle	hypothetical data if data were perfectly normal	z-score of data = (data -xbar)/s
41.2	1	0.1	-1.28	46.09	-1.39
76.6	2	0.3	-0.52	79.15	-0.58
109.3	3	0.5	0.00	102.04	0.17
134.5	4	0.7	0.52	124.93	0.74
148.6	5	0.9	1.28	157.99	1.07
102.04	=xbar				
43.65459	=s				
5	=n				

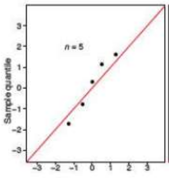


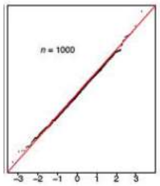
Q-Q plots are constructed differently depending on the software or textbook, but usually include some combination of the above columns. If the graph plots green vs. green or orange vs. orange, if the data is normal, then points should fall close to the line y=x. If one green and one orange are used, if the data is normal, the points should fall along a straight line, but not necessarily one with slope=1. Different software will calculate this line differently.

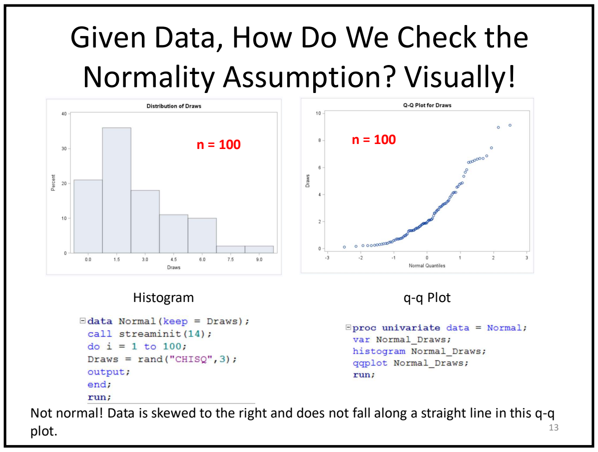
Normal q-q Plot

Data (z)	Rank (*i*)	Middle of the *i*th Interval	Normal *(z)*
-1.96	1	.1	-1.28
-.78	2	.3	-0.52
.31	3	.5	0.00
1.15	4	.7	0.52
1.62	5	.9	1.28









Given Data, How Do We Check the Normality Assumption? Visually!

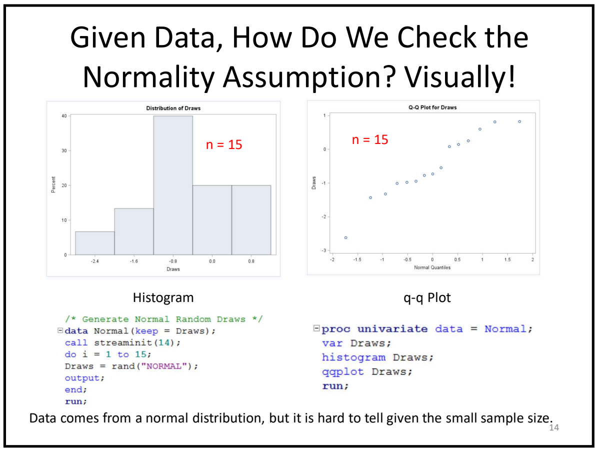
Histogram

q-q Plot

```
data Normal(keep = Draws);
call streaminit(14);
do i = 1 to 100;
Draws = rand("CHISQ",3);
output;
end;
run;
```

```
proc univariate data = Normal;
var Normal_Draws;
histogram Normal_Draws;
qqplot Normal_Draws;
run;
```

Not normal! Data is skewed to the right and does not fall along a straight line in this q-q plot.



Given Data, How Do We Check the Normality Assumption? Visually!

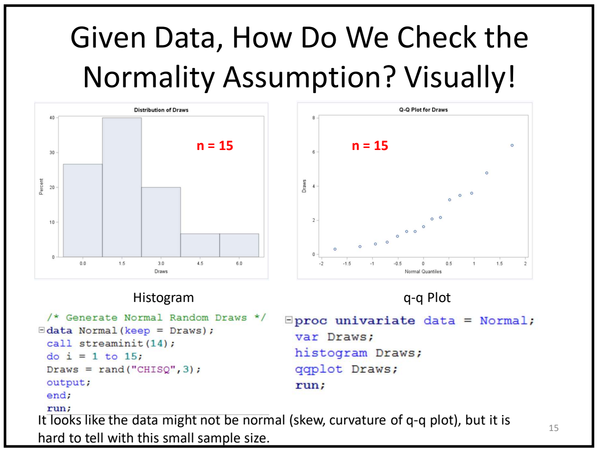
Histogram

q-q Plot

```
/* Generate Normal Random Draws */
data Normal(keep = Draws);
call streaminit(14);
do i = 1 to 15;
Draws = rand("NORMAL");
output;
end;
run;
```

```
proc univariate data = Normal;
var Draws;
histogram Draws;
qqplot Draws;
run;
```

Data comes from a normal distribution, but it is hard to tell given the small sample size.



Given Data, How Do We Check the Normality Assumption? Visually!

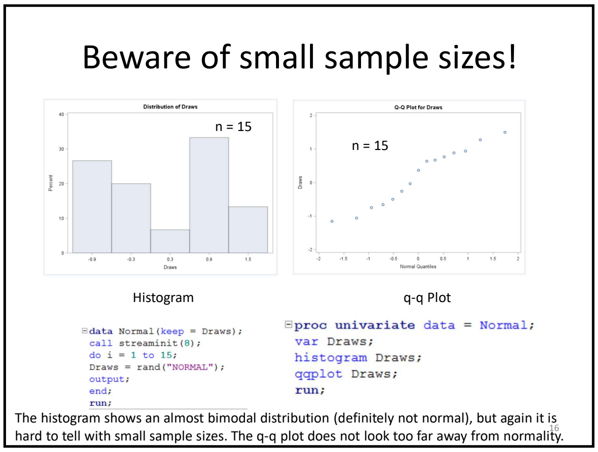
Histogram

q-q Plot

```
/* Generate Normal Random Draws */
data Normal(keep = Draws);
call streaminit(14);
do i = 1 to 15;
Draws = rand("CHISQ",3);
output;
end;
run;
```

```
proc univariate data = Normal;
var Draws;
histogram Draws;
qqplot Draws;
run;
```

It looks like the data might not be normal (skew, curvature of q-q plot), but it is hard to tell with this small sample size.



Beware of small sample sizes!

Histogram

q-q Plot

```
data Normal(keep = Draws);
call streaminit(8);
do i = 1 to 15;
Draws = rand("NORMAL");
output;
end;
run;
```

```
proc univariate data = Normal;
var Draws;
histogram Draws;
qqplot Draws;
run;
```

The histogram shows an almost bimodal distribution (definitely not normal), but again it is hard to tell with small sample sizes. The q-q plot does not look too far away from normality.

A Way to Decide:

	Small Sample Size	Large Sample Size
Little to no Evidence Against Normality	No Problem if you feel Normality is a safe assumption … run the T-Test. (You may want to be “conservative” here and run a test with fewer assumptions.)	No Problem! Run the T-Test
Significant Evidence Against Normality	Assumptions are not met and test is not robust here … Try a transformation and, if appropriate, run a t-test. If not appropriate, do NOT run the T-Test and proceed to a test with fewer / different assumptions.	No Problem .. You have the Central Limit Theorem. Run the T-Test.

A Complete Analysis:

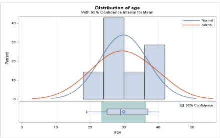
- Statement of the Problem
- Address the Assumptions
- Perform the Appropriate Test (5 Steps)
- Step 6: Provide a conclusion that a non statistician can understand, include a p-value and confidence interval.
- Scope of Inference

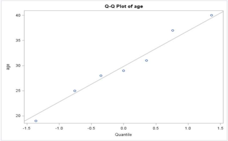
Example: Beach Comber

The following are ages of 7 randomly chosen patrons seen leaving the Beach Comber in South Mission Beach at 7pm! We assume that the data come from a normal distribution and would like to test the claim that the mean age of the distribution of Comber patrons is different than 21.

25, 19, 37, 29, 40, 28, 31

Example: Comber





PROBLEM STATEMENT:
Test the claim that the mean age of Beach Comber patrons at 7pm is different from 21.
ASSUMPTIONS:
Normal Population Distribution: Judging from the histogram and q-q plots, there is little to no evidence that the population distribution of patron ages at the Comber at 7pm is not normal. We will assume that this distribution is normal and proceed.
Independence: These subjects were randomly selected from the population; thus, we will assume that the observations are independent.

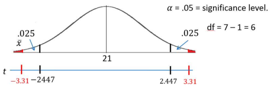
Revised Write Up!

We would like to test the claim that the population mean is different from 21. To do this, we take a sample of size n = 7 and find that $\bar{x}$ = 29.86 years and s = 7.09 years.

Step 1: Identify the null (Ho) and alternative (Ha) hypothesis.

Ho: $\mu = 21$
Ha: $\mu \neq 21$

Step 2: Draw and Shade and Find the Critical Value.



$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{29.86 - 21}{\frac{7.09}{\sqrt{7}}}$$

$$= \mathbf{3.31}$$

Step 3: Find the test statistic. (The t value for the data.)

Step 4: Find the p-value: P-value = .0162 < .05

Step 5: REJECT Ho

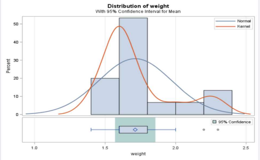
Step 6: There is sufficient evidence to conclude that the true mean age of patrons at the Comber at 7pm is different from 21 (p-value =.0162 from a t-test). A 95% confidence interval for the mean age is (23.3, 36.4) years. Scope: Since this was a random sample, we can generalize these findings to the entire population of Comber patrons at 7pm. Note that we have evidence to support the claim that the mean age is greater than 21 as well.

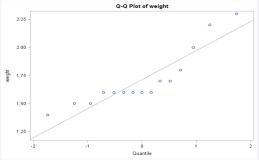
Example: Bats

World's Smallest Mammal The world's smallest mammal is the bumblebee bat, also known as the Kitti's hog-nosed bat (or *Craseonycteris thonglongyai*). Such bats are roughly the size of a large bumblebee. Listed below are weights (in grams) from a sample of these bats. Test the claim that these bats come from the same population having a mean weight equal to 1.8 g.

1.7 1.6 1.5 2.0 2.3 1.6 1.6 1.8 1.5 1.7 2.2 1.4 1.6 1.6 1.6

Example: Bats





PROBLEM STATEMENT:
Test the claim that the mean weight of the bumble bee bat is different from 1.8 g.

ASSUMPTIONS:
Normal Population Distribution: Judging from the histogram and q-q plots, there is some visual evidence of a departure from normality. With a sample size of 15 and no extreme outliers, we will assume the distribution of sample means is decently approximated by a normal distribution via the CLT and proceed with caution.
Independence: Not much is known about the sampling scheme used to obtain this sample. We will assume the observations are independent.

World's Smallest Mammal The world's smallest mammal is the bumblebee bat, also known as the Kitti's hog-nosed bat (or *Craseonycteris thonglongyai*). Such bats are roughly the size of a large bumblebee. Listed below are weights (in grams) from a sample of these bats. Test the claim that these bats come from the same population having a mean weight equal to 1.8 g.

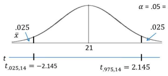
1.7 1.6 1.5 2.0 2.3 1.6 1.6 1.8 1.5 1.7 2.2 1.4 1.6 1.6 1.6

H_0: μ= 1.8
H_1: $\mu \neq$ 1.8
α = 0.05
$\bar{x}$= 1.713
s = .2588

Critical Values $t = \pm 2.145$

```
data critval;
p = quantile("T",.975,14)
;
proc print data = critval;
run;
```

```
Obs       p
  1 2.14479
```



Test Statistic

$t = -1.297$

P-value: .2155 > .05

Fail to Reject H_0

On the basis of this test, there is not enough evidence to reject the claim that the mean weight of bumblebee bats is equal to 1.8 g (p-value = .2155 from a t-test). A 95% confidence interval is (1.57, 1.8566) grams. The problem was ambiguous on the randomness of the sample; thus, we will assume that it was not a random sample, which makes inference to all bats strictly speculative.

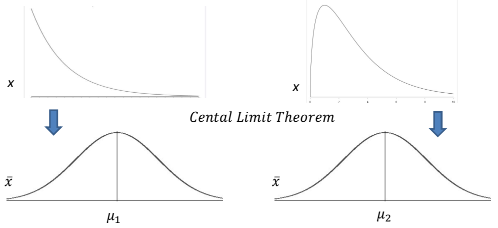
Assumptions of one and two sample T-Tests

1. Samples are drawn from a **normally** distributed population.
2. If it is a two sample test, both populations are assumed to have the same standard deviation (same shape).
3. The observations in the sample are independent of one another.

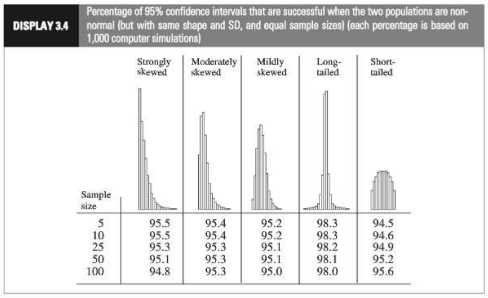
What happens if the normality assumption is broken?

Many times ….

NO PROBLEM!!!



When data is not normal

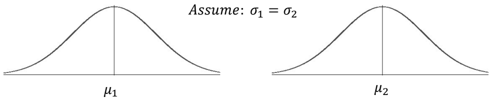


DISPLAY 3.4 Percentage of 95% confidence intervals that are successful when the two populations are non-normal (but with same shape and SD, and equal sample sizes) (each percentage is based on 1,000 computer simulations)

Sample size	Strongly skewed	Moderately skewed	Mildly skewed	Long-tailed	Short-tailed
5	95.5	95.4	95.2	98.3	94.5
10	95.5	95.4	95.2	98.3	94.6
25	95.3	95.3	95.1	98.2	94.9
50	95.1	95.3	95.1	98.1	95.2
100	94.8	95.3	95.0	98.0	95.6

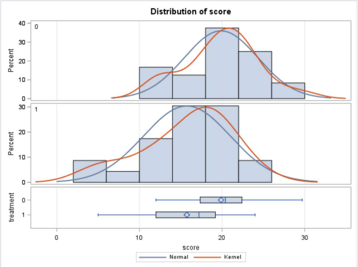
2. In a two sample test, both populations are assumed to have the same standard deviation (same shape).

$Assume:\ \sigma_1 = \sigma_2$



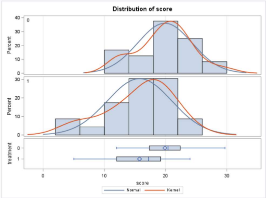
$We\ want\ inference\ on:\ \mu_2 - \mu_1$

Evidence of Inequality of Variance: VISUAL

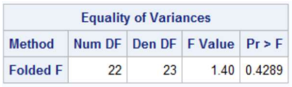


Little visual evidence against equal standard deviations (variances).

Evidence of Inequality of Variance: F-Test for Equal Variance



Ho: population variances are equal
Ha: population variances are not equal

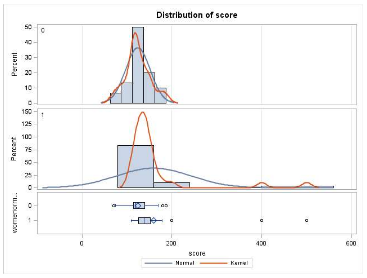


Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	22	23	1.40	0.4289

There is not sufficient evidence to conclude the variances are different (p-value = .4289 from a F-Test.)

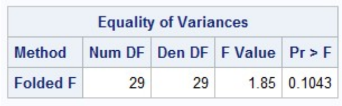
Evidence of Inequality of Variance: <u>VISUAL</u>



Strong visual evidence against equal standard deviations (variances).

Evidence of Inequality of Variance: F-<u>Test</u> for Equal Variance

Ho: population variances are equal
Ha: population variances are not equal

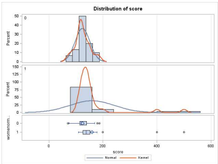


Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	29	29	1.85	0.1043

There is not sufficient evidence to conclude the variances are different (p-value = .1043 from a F-Test.)

Evidence of Inequality of Variance: F-Test / VISUAL



The F-test has a <u>strong assumption</u> that the two populations that it is testing the variances of must be normal. It is not robust to this assumption. Since the second distribution has strong evidence of right skew, the F-test for Equal Variance **is not appropriate here.**

For this example, the visual evidence is so strong that we would not need to consult a hypothesis test to test this assumption of equal variances.

However, later in the semester we will study a test of spread/dispersion that does not have this assumption and can be used in a wider range of statistical environments.

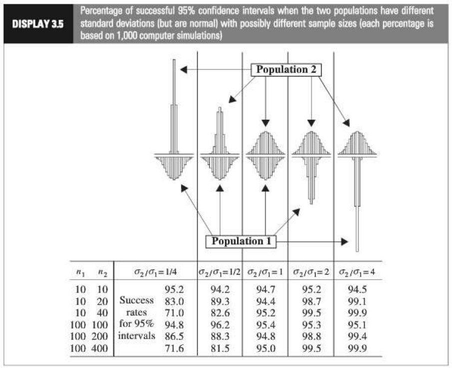
What happens if the assumption of equal variances (standard deviations) is broken?

In some circumstances ….

This could be serious …. In others…..

No Problem!

When variances are not equal



DISPLAY 3.5 Percentage of successful 95% confidence intervals when the two populations have different standard deviations (but are normal) with possibly different sample sizes (each percentage is based on 1,000 computer simulations)

n_1	n_2		$\sigma_2/\sigma_1=1/4$	$\sigma_2/\sigma_1=1/2$	$\sigma_2/\sigma_1=1$	$\sigma_2/\sigma_1=2$	$\sigma_2/\sigma_1=4$
10	10	Success rates for 95% intervals	95.2	94.2	94.7	95.2	94.5
10	20		83.0	89.3	94.4	98.7	99.1
10	40		71.0	82.6	95.2	99.5	99.9
100	100		94.8	96.2	95.4	95.3	95.1
100	200		86.5	88.3	94.8	98.8	99.4
100	400		71.6	81.5	95.0	99.5	99.9

The Take Away

What you will find in practice will most likely not fit exactly into the scenarios identified here. There will be some judgment involved … this is the "art" of statistics.

Here are some general <u>rules of thumb</u> that we will assume this semester.

1. If sample sizes are the same and sufficiently large, the t tools (tests and confidence intervals) are valid … since they are robust to the violation of normality.
2. If the two populations have the same standard deviation, then the t tests are valid … given sufficient sample sizes.
3. If the standard deviations are different and the sample sizes are different then the t tools are not valid and another procedure should be used. (Ch. 4)

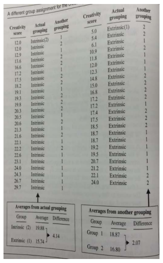
A Complete Analysis:

- Statement of the Problem
- Address the Assumptions
- Perform the Appropriate Test (5 Steps)
- Step 6: Provide a conclusion that a non statistician can understand. Include a p-value and confidence interval
- Scope of Inference

FULL EXAMPLE: CREATIVITY STUDY!

We would like to test the claim that the mean score of the Intrinsic group is different than that of the Extrinsic group. To do this we take a sample of size n_I = 24 and n_E = 23 and find that $\bar{x}_I$ = 19.88 points, $\bar{x}_E$ = 15.74, s_I = 4.44, and s_E= 5.25 points.

Step 1: Identify the null (Ho) and alternative (Ha) hypothesis.



A different group assignment for the cre...

Creativity score	Actual grouping	Another grouping
12.0	Intrinsic(2)	1
12.0	Intrinsic	2
12.9	Intrinsic	1
13.6	Intrinsic	2
16.6	Intrinsic	2
17.2	Intrinsic	1
17.5	Intrinsic	2
18.2	Intrinsic	2
19.1	Intrinsic	1
19.3	Intrinsic	2
19.8	Intrinsic	2
20.3	Intrinsic	2
20.5	Intrinsic	1
20.6	Intrinsic	2
21.3	Intrinsic	1
21.6	Intrinsic	2
22.1	Intrinsic	1
22.2	Intrinsic	2
22.6	Intrinsic	1
23.1	Intrinsic	1
24.0	Intrinsic	1
24.3	Intrinsic	1
26.7	Intrinsic	1
29.7	Intrinsic	1

Creativity score	Actual grouping	Another grouping
5.0	Extrinsic(1)	2
5.4	Extrinsic	2
6.1	Extrinsic	1
10.9	Extrinsic	2
11.8	Extrinsic	1
12.0	Extrinsic	1
12.3	Extrinsic	1
14.8	Extrinsic	2
15.0	Extrinsic	2
16.8	Extrinsic	2
17.2	Extrinsic	2
17.2	Extrinsic	1
17.4	Extrinsic	2
17.5	Extrinsic	2
18.5	Extrinsic	2
18.7	Extrinsic	1
18.7	Extrinsic	1
19.2	Extrinsic	1
19.5	Extrinsic	1
20.7	Extrinsic	1
21.2	Extrinsic	1
22.1	Extrinsic	2
24.0	Extrinsic	2

Averages from actual grouping

Group	Average	Difference
Intrinsic (2)	19.88	4.14
Extrinsic (1)	15.74	

Averages from another grouping

Group	Average	Difference
Group 1	18.87	2.07
Group 2	16.80	

Ho: $\mu_I = \mu_E$

Ha: $\mu_I \neq \mu_E$

Which is equivalent to:

Ho: $\mu_I - \mu_E = 0$

Ha: $\mu_I - \mu_E \neq 0$

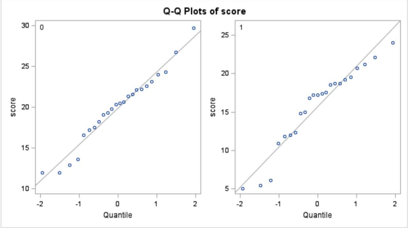
Full Example: Creativity Data

State the Problem: We would like to test the claim that the mean score of the Intrinsic group is different than that of the Extrinsic group.

Check Assumptions:

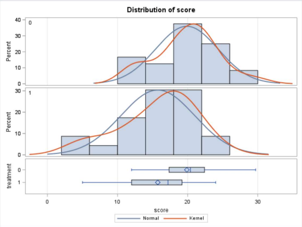
1. Normally Distributed Populations

First Check …. q-q Plot



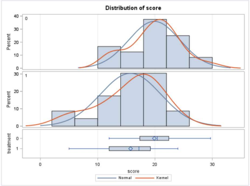
The q-q plots for both populations look sufficiently normal. We look at the histograms as well … but there is not sufficient evidence here to suggest that they are not normal.

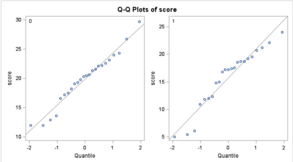
Histograms



- Keeping in mind the relative small sample size from each population, we do not observe any extreme outliers and observe a pretty strong bell shape which lends evidence to support normality of the populations.

Normality Assumption





Visual inspection of the histograms and q-q plots of each population are consistent with the normality of each population. We assume normality and move on to the second assumption.

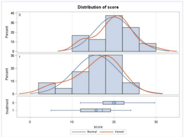
Full Example: Creativity Data

State the Problem: We would like to test the claim that the mean score of those with intrinsic motivation is the same for those with extrinsic motivation.

Check Assumptions:

1. Normally Distributed Populations
2. Equal Standard Deviations

Equality of Variances



A visual check was done by looking at the histograms, which reveal similar shapes and support the equal variances assumption. You can assume equal variances here.

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	29	29	1.85	0.1043

Since we are able to assume normal population distributions, we can use the F-Test to provide secondary evidence if the visual is inconclusive. Since the p-value is greater than our significance level of alpha = 0.05, we fail to reject the null hypothesis of equality (p-value = 0.1043) and conclude that there is not enough evidence to suggest the variances are different.

Full Example: Creativity Data

State the Problem: We would like to test the claim that the mean score of those with intrinsic motivation is the same for those with extrinsic motivation.

Check Assumptions:

1. Normally Distributed Populations
2. Equal Standard Deviations
3. Independent Observations

Independent Observations

The sample consisted of volunteers and thus subjects may not be independent of one another. However, we will assume independence and proceed with caution.

Full Example: Creativity Data

State the Problem: We would like to test the claim that the mean intrinsic score is the same as the extrinsic score.

Check Assumptions:

1. Normally Distributed Populations
2. Equal Standard Deviations
3. Independent Observations

Run the Test:

1. First 5 steps.

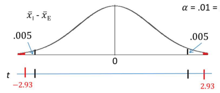
Let’s Formalize This Test Into 6 Steps!

We would like to test the claim that the mean score of the Intrinsic group is different than that of the Extrinsic group. To do this we take a sample of size $n_I = 24$ and $n_E = 23$ and find that $\bar{x}_I$ = 19.88 points, $\bar{x}_E$ = 15.74, s_I = 4.44, and s_E= 5.25 points.

Step 1: Identify the null (Ho) and alternative (Ha) hypothesis.

Ho: $\mu_I - \mu_E = 0$
Ha: $\mu_I - \mu_E \neq 0$

Step 2: Draw and Shade and Find the Critical Value.



α = .01 = significance level.

df = 24 +23 – 2 = 45

Step 3: Find the test statistic. (The t value for the data.)

$$t = \frac{(\bar{x}_I - \bar{x}_E)}{s_p\sqrt{\frac{1}{n_I}+\frac{1}{n_E}}} = 2.93$$

Step 4: Find the p-value: P-value 0.0054< .01

Step 5: Key! The sample mean we found is very unusual under the assumption that the group means are equal ($\mu_I - \mu_E$). So we Reject this assumption. That is, we REJECT Ho.

Full Example: Creativity Data

State the Problem: We would like to test the claim that the mean intrinsic score is the same as the extrinsic score.

Check Assumptions:

1. Normally Distributed Populations
2. Equal Standard Deviations
3. Independent Observations

Run the Test:

1. First 5 steps.

State the Scope and Conclusion.

Let's Fill in the P-value (and add a CI)!

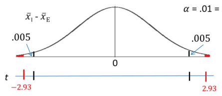
We would like to test the claim that the mean score of the Intrinsic group is different than that of the Extrinsic group. To do this we take a sample of size $n_I = 24$ and $n_E = 23$ and find that $\bar{x}_I = 19.88$ points, $\bar{x}_E = 15.74$, $s_I = 4.44$, and $s_E = 5.25$ points.

Step 1: Identify the null (Ho) and alternative (Ha) hypothesis.

Ho: $\mu_I - \mu_E = 0$

Ha: $\mu_I - \mu_E \neq 0$

Step 2: Draw and Shade and Find the Critical Value.



α = .01 = significance level.

df = 24 +23 – 2 = 45

Step 3: Find the test statistic. (The t value for the data.)

$$t = \frac{(\bar{x}_I - \bar{x}_E)}{s_p\sqrt{\frac{1}{n_I} + \frac{1}{n_E}}} = 2.93$$

Step 4: Find the p-value: P-value = .0054

Step 5: REJECT Ho

Step 6:

Conclusion: There is sufficient evidence to suggest that those who receive the Intrinsic treatment have a higher mean score than those who receive the Extrinsic treatment (p-value = .0054 from a two sided t-test). A 99% confidence interval for this difference is (1.29, 7.00).

SCOPE: Since this was a randomized experiment, we can conclude that the Intrinsic treatment caused this difference. However, since the study was of volunteers, this inference can only be generalized to the 47 participants.

LET'S TRY SOME!

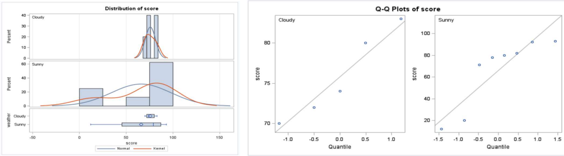
For each of these data sets, write up the assumption statement with respect to checking the assumptions for a one or two sample t-test. You may assume the data to be independent.

Happiness Data Set

Mice Experiment Data Set

All data sets can be found in one file in this week's materials.
You will need to add the proc ttest statement for each.
However, you will not need the data for this exercise.

Happiness Study

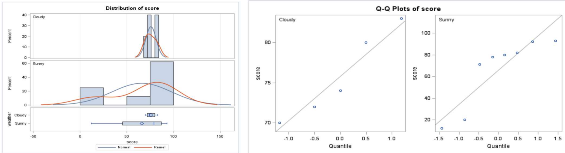


5 randomly selected people were asked to rate their happiness on a scale from 1 – 100 on a cloudy day and 8 randomly selected people were asked the same question on a sunny day.

QOI: Is the mean happiness of individuals different on a cloudy day than a sunny day? If possible, can we test if cloudy weather causes a change in happiness?

Address each assumption of the two sample t-test and then decide if the two-sample t-test is appropriate to answer this QOI with this data.

Happiness Study



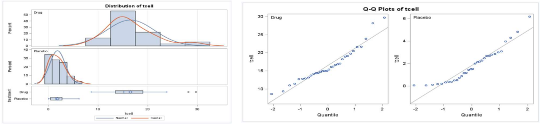
Normality of Distributions: Judging from the histograms and q-q plots, there is evidence of outliers in both the Cloudy and Sunny sets. The most pronounced outlier seems to be in the Sunny data set; thus, there is significant visual evidence against these data being normally distributed. In addition, we are not satisfied that the t-test will be robust to this assumption since the sample sized are so small.

Equal Standard Deviations: Judging from the histograms, q-q plots and box plots, there is significant visual evidence that the standard deviations are different. In addition, since the sample sizes are different we know that the t-test is not robust to this assumption.

Independence: We will assume that these data are independent.

The two sample t-test is not appropriate here. We should look for a different test.

Mice Study

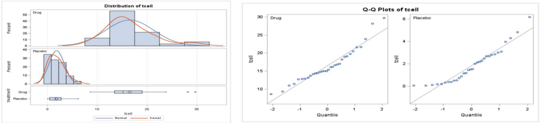


A large sample of mice were randomly assigned to receive a drug or a placebo (sample size $n_D = 32$ and $n_P = 32$). The mice's tcell counts were then taken and histograms and q-q plots are displayed above.

QOI: **Is the mean tcell count of mice that receive the drug greater than that of the mice that receive the placebo?**
Can we draw draw evidence of causality from this study?

Address each assumption of the two sample t-test and then decide if the two-sample t-test is appropriate to answer this QOI with this data.

Mice Study



Normality of Distributions: Judging from the histograms and q-q plots, there is significant visual evidence to suggest the data come from right skewed distributions. However, since the sample size is large $n_D = 32$ and $n_P = 32$ the t-test is robust to this assumption violation.

Equal Standard Deviations: There is strong visual evidence to suggest that the data come from distributions with different standard deviations. However, since we have the same sample size in each group, the t-test is robust to this assumption violation, by a previous "rule of thumb".

Independence: We will assume that these data are independent.

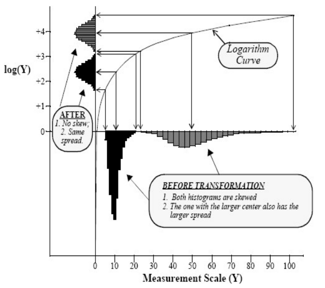
The two sample t-test is appropriate here.

Transformations

Log Transformation

Display 3.8 p. 69

The logarithmic transformation used to arrive at favorable conditions for the two-sample t-analysis



Appropriate Interpretations After a Log Transformation – Example Write Ups….

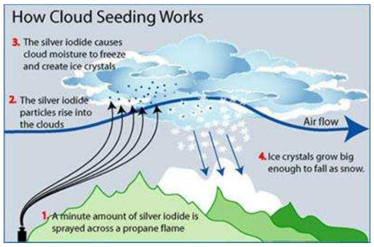
Observational Study:

"It is estimated that the median for population X is exp(mean(log(x)) – mean(log(y))) times as large as the median for population Y."

Randomized Experiment:

"It is estimated that the median response of an experimental unit to treatment x will be exp(mean(log(x)) – mean(log(y))) times as large as its response to treatment y."

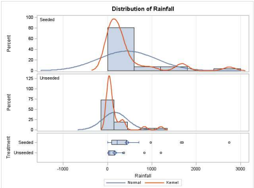
Cloud Seeding!

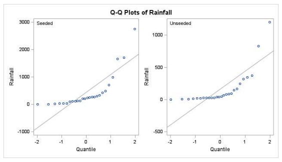


Does Cloud Seeding Work?

On days that were deemed suitable for cloud seeding, a random mechanism was used to decide whether to seed the target cloud on that day or to leave it unseeded as a control. Precipitation was measured as the total rain volume falling from the cloud base following the airplane seeding run, as measured by radar. We would like to test at the alpha = .05 level of significance whether cloud seeding is effective in increasing precipitation.

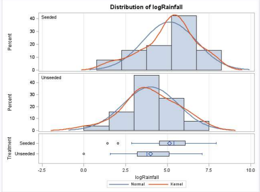
Cloud Seeding: Original Data

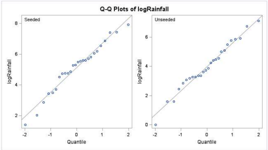




```
proc ttest data = cloud sides = u;
class Treatment;
var rainfall;
run;
```

After Log Transformation





```
data lcloud;
set cloud;
lograin = log(rainfall);
run;
```

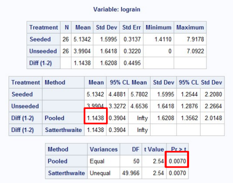
```
proc ttest data = lcloud sides = u;
class Treatment;
var lograin;
run;
```

T Test and Confidence!!!

H_0: Cloud Seeding does not work.
H_1: Cloud Seeding does work.

H_0: $Median_{seeded}$ = $Median_{unseeded}$
H_1: $Median_{seeded}$ > $Median_{unseeded}$

$e^{0.3904}$ = 1.5,
$e^{1.8972}$ = 6.7



Variable: lograin

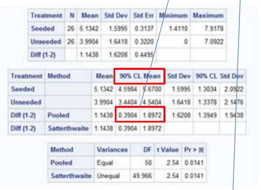
Treatment	N	Mean	Std Dev	Std Err	Minimum	Maximum
Seeded	26	5.1342	1.5995	0.3137	1.4110	7.9178
Unseeded	26	3.9904	1.6418	0.3220	0	7.0922
Diff (1-2)		1.1438	1.6208	0.4495		

Treatment	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Seeded		5.1342	4.4881	5.7802	1.5995	1.2544	2.2080
Unseeded		3.9904	3.3272	4.6536	1.6418	1.2876	2.2664
Diff (1-2)	Pooled	1.1438	0.3904	Infty	1.6208	1.3562	2.0148
Diff (1-2)	Satterthwaite	1.1438	0.3904	Infty			

Method	Variances	DF	t Value	Pr > t
Pooled	Equal	50	2.54	0.0070
Satterthwaite	Unequal	49.966	2.54	0.0070

For the one sided test.

```
proc ttest data = lcloud sides = u;
class Treatment;
var lograin;
run;
```



Treatment	N	Mean	Std Dev	Std Err	Minimum	Maximum
Seeded	26	5.1342	1.5995	0.3137	1.4110	7.9178
Unseeded	26	3.9904	1.6418	0.3220	0	7.0922
Diff (1-2)		1.1438	1.6208	0.4495		

Treatment	Method	Mean	90% CL Mean		Std Dev	90% CL Std Dev	
Seeded		5.1342	4.5984	5.6700	1.5995	1.3034	2.0922
Unseeded		3.9904	3.4404	4.5404	1.6418	1.3378	2.1476
Diff (1-2)	Pooled	1.1438	0.3904	1.8972	1.6208	1.3949	1.9438
Diff (1-2)	Satterthwaite	1.1438	0.3904	1.8972			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	50	2.54	0.0141
Satterthwaite	Unequal	49.966	2.54	0.0141

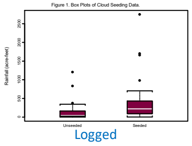
For confidence interval.

```
proc ttest data = lcloud sides = 2 alpha = .1;
class Treatment;
var lograin;
run;
```

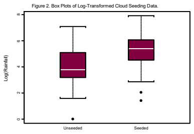
It is estimated that the median volume of rainfall on days when clouds were seeded was $e^{1.1438}$=3.1 times as large as when not seeded (p-value = .007). A 90% confidence interval for this multiplicative effect on the median is 1.5 to 6.7 times. Since randomization was used to determine whether any particular suitable day was seeded or not, it is safe to interpret this as evidence that the seeding caused the larger median rainfall.

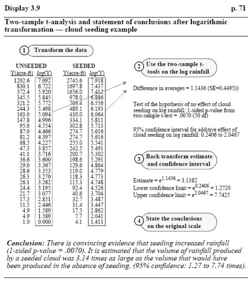
Cloud Seeding Book Example

Original



Logged





Display 3.9 **p. 71**

Two-sample t-analysis and statement of conclusions after logarithmic transformation — cloud seeding example

① Transform the data

UNSEEDED Y(acre-ft)	 log(Y)	SEEDED Y(acre-ft)	 log(Y)
1202.6	7.092	2745.6	7.918
830.1	6.722	1697.8	7.437
372.4	5.920	1656.0	7.412
345.5	5.845	978.0	6.886
321.2	5.772	703.4	6.556
244.3	5.498	489.1	6.193
163.0	5.094	430.0	6.064
147.8	4.996	334.1	5.811
95.0	4.554	302.8	5.713
87.0	4.466	274.7	5.616
81.2	4.397	274.7	5.616
68.5	4.227	255.0	5.541
47.3	3.857	242.5	5.491
41.1	3.716	200.7	5.302
36.6	3.600	198.6	5.291
29.0	3.367	129.6	4.864
28.6	3.353	119.0	4.779
26.3	3.270	118.3	4.773
26.1	3.262	115.3	4.748
24.4	3.195	92.4	4.526
21.7	3.077	40.6	3.704
17.3	2.851	32.7	3.487
11.5	2.446	31.4	3.447
4.9	1.589	17.5	2.862
4.9	1.589	7.7	2.041
1.0	0.000	4.1	1.411

② Use the two-sample t-tools on the log rainfall

Difference in averages = 1.1436 (SE=0.4495))

Test of the hypothesis of no effect of cloud seeding on log rainfall: 1-sided p-value from two-sample t-test = .0070 (50 df)

95% confidence interval for additive effect of cloud seeding on log rainfall: 0.2406 to 2.0467

③ Back transform estimate and confidence interval

Estimate = $e^{1.1436}$ = 3.1382
Lower confidence limit = $e^{0.2406}$ = 1.2720
Upper confidence limit = $e^{2.0467}$ = 7.7425

④ State the conclusions on the original scale

Conclusion: *There is convincing evidence that seeding increased rainfall (1-sided p-value = .0070). It is estimated that the volume of rainfall produced by a seeded cloud was 3.14 times as large as the volume that would have been produced in the absence of seeding. (95% confidence: 1.27 to 7.74 times).*

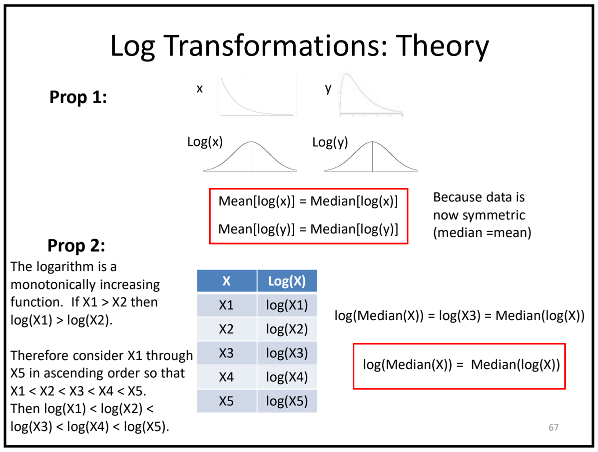
Recap: The Take Away

What you will find in practice will most likely not fit exactly into the scenarios we identified here. There will be some judgment involved … this is the "art" of statistics.

Here are some general rules of thumb that we will assume this semester.

1. If sample sizes are the same and sufficiently large, the t tools (tests and confidence intervals) are valid … since they are robust to the violation of normality.
2. If the two populations have the same standard deviation then the t tests are valid … given sufficient sample sizes.
3. If the standard deviations are different and the sample sizes are different then the t tools are not valid and another procedure should be used. (Ch. 4)

Appendix



Log Transformations: Theory

Prop 1:

Mean[log(x)] = Median[log(x)]

Mean[log(y)] = Median[log(y)]

Because data is now symmetric (median =mean)

Prop 2:

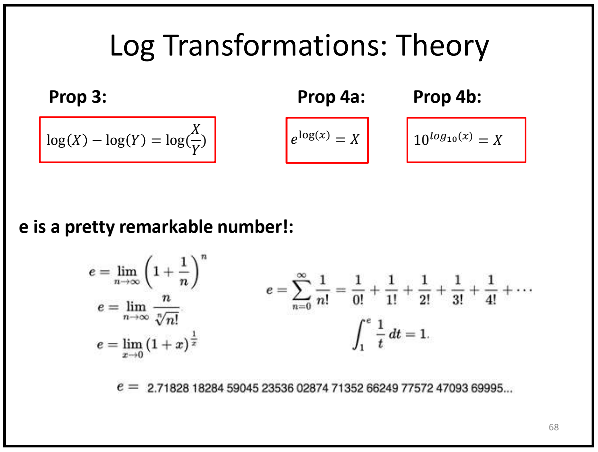
The logarithm is a monotonically increasing function. If X1 > X2 then log(X1) > log(X2).

Therefore consider X1 through X5 in ascending order so that X1 < X2 < X3 < X4 < X5. Then log(X1) < log(X2) < log(X3) < log(X4) < log(X5).

X	Log(X)
X1	log(X1)
X2	log(X2)
X3	log(X3)
X4	log(X4)
X5	log(X5)

log(Median(X)) = log(X3) = Median(log(X))

log(Median(X)) = Median(log(X))



Log Transformations: Theory

Prop 3:

$\log(X) - \log(Y) = \log(\frac{X}{Y})$

Prop 4a:

$e^{\log(x)} = X$

Prop 4b:

$10^{log_{10}(x)} = X$

e is a pretty remarkable number!:

$$e = \lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n$$

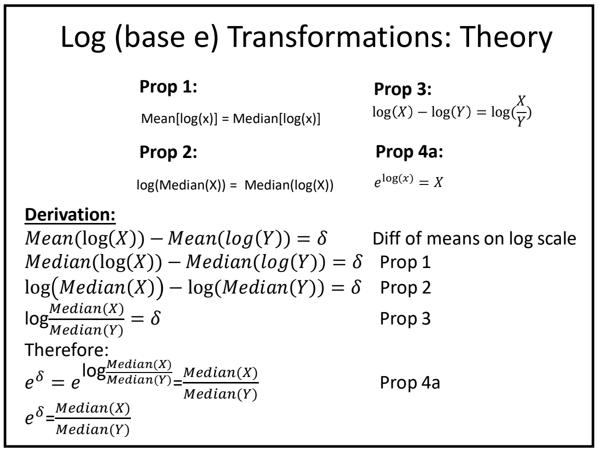
$$e = \lim_{n\to\infty}\frac{n}{\sqrt[n]{n!}}$$

$$e = \lim_{x\to 0}(1+x)^{\frac{1}{x}}$$

$$e = \sum_{n=0}^{\infty}\frac{1}{n!} = \frac{1}{0!}+\frac{1}{1!}+\frac{1}{2!}+\frac{1}{3!}+\frac{1}{4!}+\cdots$$

$$\int_1^e \frac{1}{t}\,dt = 1.$$

$e =$ 2.71828 18284 59045 23536 02874 71352 66249 77572 47093 69995...



Log (base e) Transformations: Theory

Prop 1:

Mean[log(x)] = Median[log(x)]

Prop 2:

log(Median(X)) = Median(log(X))

Prop 3:

$\log(X) - \log(Y) = \log(\frac{X}{Y})$

Prop 4a:

$e^{\log(x)} = X$

Derivation:

$Mean(\log(X)) - Mean(log(Y)) = \delta$ Diff of means on log scale

$Median(\log(X)) - Median(log(Y)) = \delta$ Prop 1

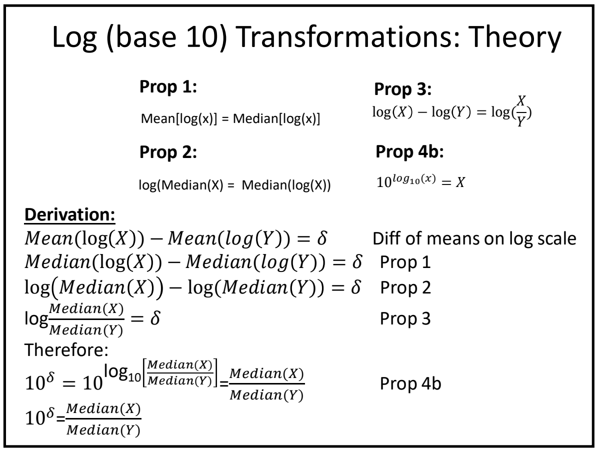
$\log\big(Median(X)\big) - \log(Median(Y)) = \delta$ Prop 2

$\log\frac{Median(X)}{Median(Y)} = \delta$ Prop 3

Therefore:

$e^{\delta} = e^{\log\frac{Median(X)}{Median(Y)}} = \frac{Median(X)}{Median(Y)}$ Prop 4a

$e^{\delta} = \frac{Median(X)}{Median(Y)}$



Log (base 10) Transformations: Theory

Prop 1:

Mean[log(x)] = Median[log(x)]

Prop 2:

log(Median(X) = Median(log(X))

Prop 3:

$\log(X) - \log(Y) = \log(\frac{X}{Y})$

Prop 4b:

$10^{log_{10}(x)} = X$

Derivation:

$Mean(\log(X)) - Mean(log(Y)) = \delta$ Diff of means on log scale

$Median(\log(X)) - Median(log(Y)) = \delta$ Prop 1

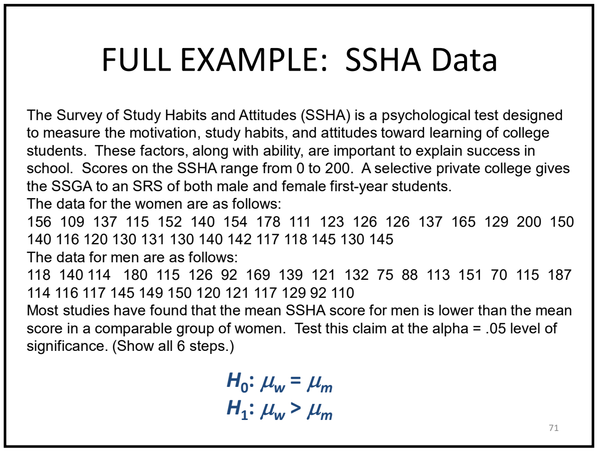
$\log\big(Median(X)\big) - \log(Median(Y)) = \delta$ Prop 2

$\log\frac{Median(X)}{Median(Y)} = \delta$ Prop 3

Therefore:

$10^{\delta} = 10^{\log_{10}\left[\frac{Median(X)}{Median(Y)}\right]} = \frac{Median(X)}{Median(Y)}$ Prop 4b

$10^{\delta} = \frac{Median(X)}{Median(Y)}$



FULL EXAMPLE: SSHA Data

The Survey of Study Habits and Attitudes (SSHA) is a psychological test designed to measure the motivation, study habits, and attitudes toward learning of college students. These factors, along with ability, are important to explain success in school. Scores on the SSHA range from 0 to 200. A selective private college gives the SSGA to an SRS of both male and female first-year students.

The data for the women are as follows:

156 109 137 115 152 140 154 178 111 123 126 126 137 165 129 200 150 140 116 120 130 131 130 140 142 117 118 145 130 145

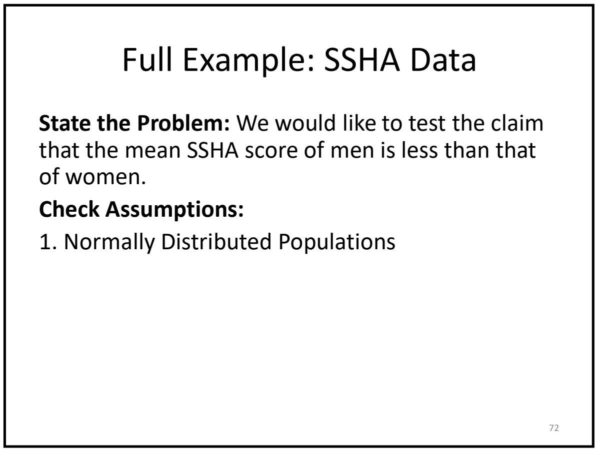
The data for men are as follows:

118 140 114 180 115 126 92 169 139 121 132 75 88 113 151 70 115 187 114 116 117 145 149 150 120 121 117 129 92 110

Most studies have found that the mean SSHA score for men is lower than the mean score in a comparable group of women. Test this claim at the alpha = .05 level of significance. (Show all 6 steps.)

H_0: $\mu_w = \mu_m$

H_1: $\mu_w > \mu_m$



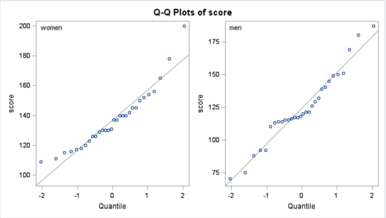
Full Example: SSHA Data

State the Problem: We would like to test the claim that the mean SSHA score of men is less than that of women.

Check Assumptions:

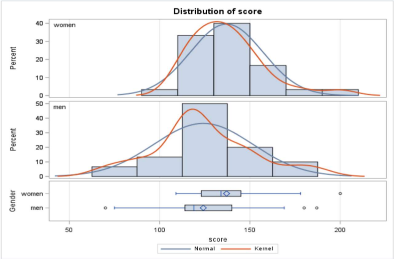
1. Normally Distributed Populations

First Check …. q-q Plot



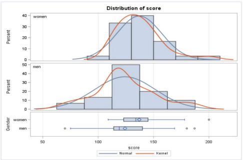
The q-q plots for both populations look sufficiently normal. We look at the histograms as well … but there is not sufficient evidence here to suggest that they are not normal.

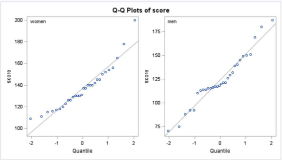
Histograms



- Keeping in mind the relative small sample size from each population, we do not observe any extreme outliers and observe a pretty strong bell shape which lends evidence to support normality of the populations.

Normality Assumption





Visual inspection of the histograms and q-q plots of each population is consistent with the normality of each population. We assume normality and move on to the second assumption.

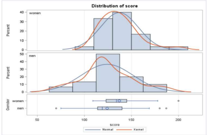
Full Example: SSHA Data

State the Problem: We would like to test the claim that the mean SSHA score of men is less than that of women.

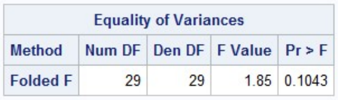
Check Assumptions:

1. Normally Distributed Populations
2. Equal Standard Deviations

Equality of Variances



A visual check was done by looking at the histograms which reveal similar shapes and support the equal variances assumption. You can assume equal variances here.



Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	29	29	1.85	0.1043

Since we are able to assume normal population distributions, we can use the F-Test to provide secondary evidence if the visual is inconclusive. Since the p-value is greater than our significance level of alpha = 0.05, we fail to reject the null hypothesis of equality (p-value = 0.1043) of variances and conclude that there is not enough evidence to suggest the variances are different.

Full Example: SSHA Data

State the Problem: We would like to test the claim that the mean SSHA score of men is less than that of women.

Check Assumptions:

1. Normally Distributed Populations
2. Equal Standard Deviations
3. Independent Observations

Independent Observations

The sample was indeed a SRS (simple random sample) from the population of the selective private college, therefore we assume the observations are independent of one another.

Full Example: SSHA Data

State the Problem: We would like to test the claim that the mean SSHA score of men is less than that of women.

Check Assumptions:

1. Normally Distributed Populations
2. Equal Standard Deviations
3. Independent Observations

Run the Test:

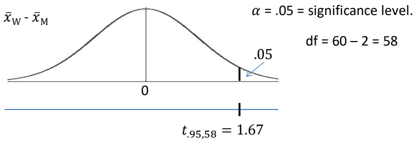
1. First 5 steps.

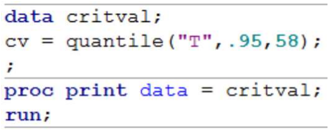
Run The Two Sample T-Test!!!

- There is no reason to pair these observations and we have two samples …. Therefore we should use the two sample t-test with pooled standard deviation since we are assuming the population standard deviations are equal. We are testing here:

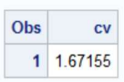
$$H_0: \mu_W = \mu_M$$
$$H_1: \mu_W > \mu_M$$

Critical Value





```
data critval;
cv = quantile("T",.95,58);
;
proc print data = critval;
run;
```



Obs	cv
1	1.67155

Two Sample T-Test … SAS Output

Gender	N	Mean	Std Dev	Std Err	Minimum	Maximum
women	30	137.1	20.1528	3.6794	109.0	200.0
men	30	124.2	27.3837	4.9996	70.0000	187.0
Diff (1-2)		12.9000	24.0416	6.2075		

Gender	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
women		137.1	129.5	144.6	20.1528	16.0498	27.0916
men		124.2	113.9	134.4	27.3837	21.8086	36.8123
Diff (1-2)	Pooled	12.9000	2.5238	Infty	24.0416	20.3521	29.3778
Diff (1-2)	Satterthwaite	12.9000	2.5089	Infty			

Method	Variances	DF	t Value	Pr > t
Pooled	Equal	58	2.08	0.0211
Satterthwaite	Unequal	53.288	2.08	0.0213

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	29	29	1.85	0.1043

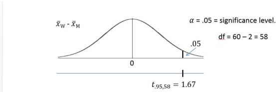
Let's Formalize This Test Into 6 Steps!

We would like to test the claim that the mean SSHA score of the men is less than the mean score of women. To do this we take a sample of size n_M = 30 and n_W = 30 and find that $\bar{x}_M$ = 124.2 points, $\bar{x}_W$ = 137.1 and s_M = 27.2 s_W= 20.2 points.

Step 1: Identify the null (Ho) and alternative (Ha) hypothesis.

Ho: $\mu_W - \mu_M = 0$
Ha: $\mu_W - \mu_M > 0$

Step 2: Draw and Shade and Find the Critical Value.



Step 3: Find the test statistic. (The t value for the data.)

$$t = \frac{(\bar{x}_W - \bar{x}_M)}{s_p\sqrt{\frac{1}{n_I} + \frac{1}{n_E}}} = 2.08$$

Step 4: Find the p-value: P-value = .0211

Step 5: REJECT Ho.

Full Example: SSHA Data

State the Problem: We would like to test the claim that the mean SSHA score of men is less than that of women.

Check Assumptions:

1. Normally Distributed Populations
2. Equal Standard Deviations
3. Independent Observations

Run the Test:

1. First 5 steps.

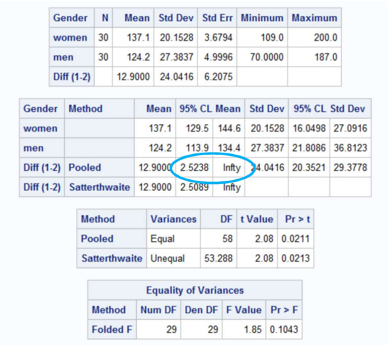
State the Scope and Conclusion.

Scope

Since the study is between women and men, the subjects cannot be randomly assigned to the two groups, and we have an observational study. For this reason, we cannot make any causal inference and must limit our conclusions to differences of group means.

However, the sample was an SRS and thus any results can be inferred back to the population of students at this particular private college.

Two Sample T-Test … SAS Output



Gender	N	Mean	Std Dev	Std Err	Minimum	Maximum
women	30	137.1	20.1528	3.6794	109.0	200.0
men	30	124.2	27.3837	4.9996	70.0000	187.0
Diff (1-2)		12.9000	24.0416	6.2075		

Gender	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
women		137.1	129.5	144.6	20.1528	16.0498	27.0916
men		124.2	113.9	134.4	27.3837	21.8086	36.8123
Diff (1-2)	Pooled	12.9000	2.5238	Infty	24.0416	20.3521	29.3778
Diff (1-2)	Satterthwaite	12.9000	2.5089	Infty			

Method	Variances	DF	t Value	Pr > t
Pooled	Equal	58	2.08	0.0211
Satterthwaite	Unequal	53.288	2.08	0.0213

Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	29	29	1.85	0.1043

Conclusion

There is sufficient evidence to support the claim at the α=.05 level of significance (p-value = .0211) that the mean SSHA score is lower for men than for women at this college. A 95% one side confidence interval for this difference is (2.5238 points, ∞.)

Scope of Inference: Since the study is between women and men, the subjects cannot be randomly assigned to the two groups, and we have an observational study. For this reason, we cannot make any causal inference and must limit our conclusions to differences of group means.

However, the sample was an SRS, and thus any results can be inferred back to the population of students at this particular private college.

ANOTHER FULL EXAMPLE

FULL EXAMPLE: Promotion Data

The Revenue Commissioners in Ireland conducted a contest for promotion. The ages of the unsuccessful and successful applicants are given below. Some of the applicants who were unsuccessful in getting the promotion charged that the competition involved discrimination based on age. Treat the data as samples from larger populations and use a .05 significance level to test the claim that the unsuccessful applicants are from a population with a greater mean age than the mean age of successful applicants. Based on the result, does there appear to be discrimination based on age? (Show all 6 steps.) Assume all data comes from a normally distributed population.

Unsuccessful Applicants:

34 37 37 38 41 42 43 44 44 45
45 60 46 65 49 65 53 54
62 55 56 70 64

Successful Applicants

27 33 36 37 38 38 39 42 42 43
43 44 44 44 45 70 71 72
80 46 47 75 48 72 49 49
51 51 52 54

H_0: $\mu_U = \mu_S$
H_1: $\mu_S < \mu_U$

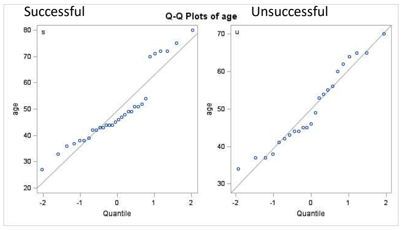
Full Example: Promotion Data

State the Problem: We would like to test the claim that the mean of the successful group is less than the mean of the unsuccessful group.

Check Assumptions:

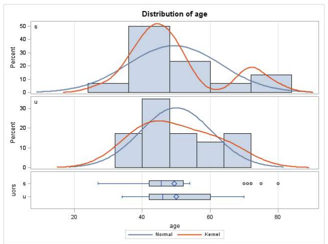
1. Normally Distributed Populations

First Check …. q-q Plot



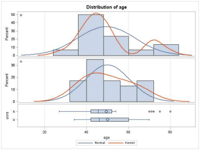
The q-q plot for the successful data provides some evidence of non normality, while the q-q plot for the unsuccessful data looks consistent with normally distributed data.

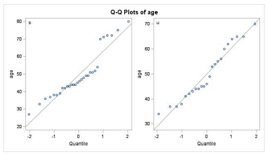
Histograms



- The successful group (top) has a clear right skew to the data, while the unsuccessful group shows a possible mild right skew. This suggests that both sets of data may be from right skewed populations. We know that the t-tools are robust to non normality for these types of distributions so we proceed with the t test…. We will readdress these concerns when we talk about the standard deviation.

Normality Assumption





Visual Inspection of the histograms and q-q plots indicates the both data sets may be from a right skewed distribution. We know that the t-tests are robust to violations of the normality assumption when the data are from a right skewed distribution (when the sample size is sufficient), so we proceed with the t-test.

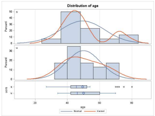
Full Example: Promotion Data

State the Problem: We would like to test the claim that the mean of the successful group is less than the mean of the unsuccessful group.

Check Assumptions:

1. Normally Distributed Populations
2. Equal Standard Deviations

Equality of Variances



A visual check was done by looking at the histograms, which reveal similar shapes and support the equal variances assumption. We will assume equal variances here.

Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	29	22	1.65	0.2286

As secondary evidence of the visual is inconclusive, given that the p-value is greater than our significance level of alpha = 0.05, we fail to reject the null hypothesis of equality of variances (p-value = 0.2286) and conclude that there is not enough evidence to suggest the variances are different.

Full Example: Promotion Data

State the Problem: We would like to test the claim that the mean of the successful group is less than the mean of the unsuccessful group.

Check Assumptions:

1. Normally Distributed Populations
2. Equal Standard Deviations
3. Independent Observations

Independent Observations

The sample was indeed a SRS (simple random sample) from the population of the selective private college, therefore we assume the observations are independent of one another.

Full Example: Promotion Data

State the Problem: We would like to test the claim that the mean of the successful group is less than the mean of the unsuccessful group.

Check Assumptions:

1. Normally Distributed Populations
2. Equal Standard Deviations
3. Independent Observations

Run the Test:

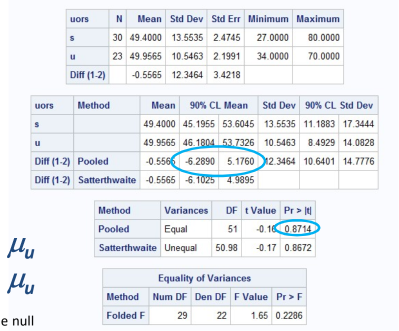
1. First 5 steps.

Run The Two Sample T-Test!!!

- There is no reason to pair these observations, and we have two samples. Therefore, we should use the two sample t-test with a pooled standard deviation, since we are assuming the population standard deviations are equal. We are testing here:

$$H_0\text{: } \mu_s = \mu_u$$
$$H_1\text{: } \mu_s < \mu_u$$

Two Sample T-Test … SAS Output



uors	N	Mean	Std Dev	Std Err	Minimum	Maximum
s	30	49.4000	13.5535	2.4745	27.0000	80.0000
u	23	49.9565	10.5463	2.1991	34.0000	70.0000
Diff (1-2)		-0.5565	12.3464	3.4218		

uors	Method	Mean	90% CL Mean		Std Dev	90% CL Std Dev	
s		49.4000	45.1955	53.6045	13.5535	11.1883	17.3444
u		49.9565	46.1804	53.7326	10.5463	8.4929	14.0828
Diff (1-2)	Pooled	-0.5565	-6.2890	5.1760	12.3464	10.6401	14.7776
Diff (1-2)	Satterthwaite	-0.5565	-6.1025	4.9895			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	51	-0.16	0.8714
Satterthwaite	Unequal	50.98	-0.17	0.8672

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	29	22	1.65	0.2286

$$H_0\text{: } \mu_s = \mu_u$$
$$H_1\text{: } \mu_s < \mu_u$$

Fail to reject the null hypothesis at 0.05 level.

Full Example: Promotion Data

State the Problem: We would like to test the claim that the mean of the successful group is less than the mean of the unsuccessful group.

Check Assumptions:

1. Normally Distributed Populations
2. Equal Standard Deviations
3. Independent Observations

Run the Test:

1. First 5 steps.

State the Scope and Conclusion.

SCOPE

Since the study is between successful and unsuccessful candidates for a promotion, subjects cannot be randomly assigned to the two groups, and we have an observational study. For this reason we cannot make any causal inference and must limit our conclusions to differences of group means.

However, the sample was an SRS and thus any results can be inferred back to candidates for promotion from the population that the Revenue Commissioners of Ireland sampled.

Conclusion

There is not sufficient evidence to support the claim at the α=.05 level of significance (p-value = .4357) that the mean age of those who were given a promotion is lower than those who were not given the promotion in this . A 90% confidence interval for this difference is (-6.3 points, 5.2 points.)

Part IV

Alternatives to the t tools

Chapter 15

Problem 2: Logging problem

We are doing rank sum analysis

15.1 Complete Rank-Sum Analysis Using SAS

Problem Statement

We would like to test the claim that logging burned trees increased the percentage of seedlings lost in the Biscuit Fire region from 2004 to 2005.

Assumptions

Independence

The two-sample Wilcoxon Rank-Sum test assumes that the samples are independent. In this case, the two sets of tree plots are independent of each other, the amount of tree seedlings in one plot is not directly related to the amount of tree seedlings in another, if it is, it is not a tangible amount of dependence. Therefore, we can assume independence. We can also assume ordinality with numericla data

Statement of the Hypothesis

Our null hypothesis, H_0, is that the distribution of percent of saplings lost in the logged plots is less than or equal to the distribution of percent of saplings lost in the unlogged plots. Our alternative hypothesis, H_1, is that the distribution of percent of saplings lost in the logged plots is greater than the distribution of percent of saplings lost in the unlogged plots. Mathematically speaking, we have:

$$H_0 : meanRank_{logged} - meanRank_{unlogged} \leq 0 \tag{15.1.1}$$
$$H_1 : meanRank_{logged} - meanRank_{unlogged} > 0 \tag{15.1.2}$$

The significance level, α, is:

$$\alpha = 0.05 \tag{15.1.3}$$

Calculation of the P-value

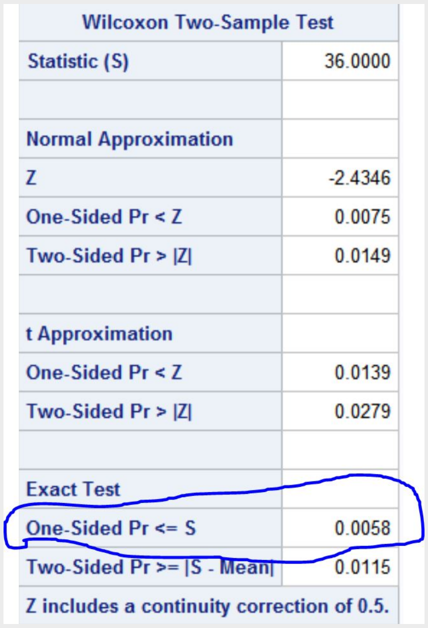
To find the p value, I performed a Wilcoxon Rank-Sum test. Because the sample size is small, an exact test was used, as there is no need for a normal approximation. The code used to perform the test is as follows:

Code 15.1. Exact rank sum test using SAS

```
/* We want the wilcoxon test and the Hodges-Lehman Confidence Interval*/
proc NPAR1WAY data=loggingData Wilcoxon HL;
class Action;
Var PercentLost;
/* Because our sample size is small, we want to do an Exact test*/
Exact;
run;
```

The output of this code is displayed in Figure 2.1:

Figure 15.1.1. Results of the Rank-Sum Test on the Logging Data



Wilcoxon Two-Sample Test	
Statistic (S)	36.0000
Normal Approximation	
Z	-2.4346
One-Sided Pr < Z	0.0075
Two-Sided Pr > |Z|	0.0149
t Approximation	
One-Sided Pr < Z	0.0139
Two-Sided Pr > |Z|	0.0279
Exact Test	
One-Sided Pr <= S	0.0058
Two-Sided Pr >= |S - Mean|	0.0115
Z includes a continuity correction of 0.5.	

The calculated p value is

$$p = 0.0058 \tag{15.1.4}$$

Results of the Hypothesis Test

We have that:

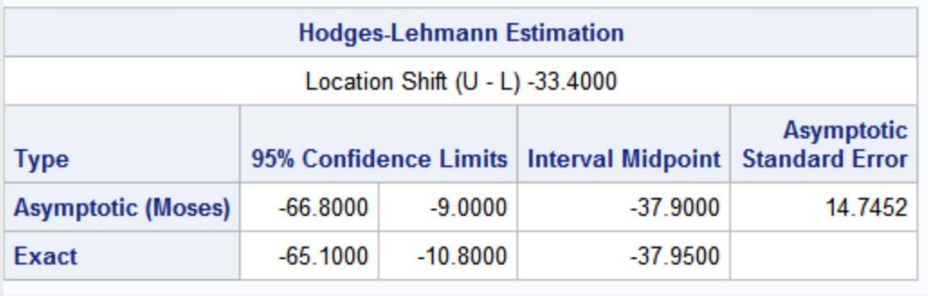
$$p = 0.0058 < \alpha = .05 \tag{15.1.5}$$

Therefore, we Reject the Null Hypothesis There is sufficient evidence at the $\alpha = 0.5$ significance level ($p-value = 0.0058$ for the exact test) to suggest that the distribution of percentages of saplings lost in the logged plots was greater than the distribution of percentages of saplings lost.

Statistical Conclusion

MEDIANS FOR NONPAR The data provides convincing evidence that forest recovery is decreased in areas where burned trees were logged. At a significance level of .05 (or even .01), the distribution/MEDIAN of the percentage of saplings lost in the logged plots was greater than that of the unlogged areas. This was done with a one sided, exact p-value of 0.0058. A range of plausible values (95 % confidence interval) for how much greater the median loss of saplings was for the logged trees is [10.8,65.1], as displayed in Figure 2.2

Figure 15.1.2. 95% Confidence Interval



Hodges-Lehmann Estimation

Location Shift (U - L) -33.4000

Type	95% Confidence Limits		Interval Midpoint	Asymptotic Standard Error
Asymptotic (Moses)	-66.8000	-9.0000	-37.9000	14.7452
Exact	-65.1000	-10.8000	-37.9500	

Note that the negative of these values was taken, because this figure shows $Unlogged - Logged$.

Scope of Inference

This study was a random sample of trees in the plots, therefore we can make generalizations about all of the trees in the 16 plots, and say that the areas which were logged had a greater loss of saplings and therefore recovered more poorly than the unlogged areas. However, this was not a randomized experiment, and therefore we cannot make causal inferences. That is, we cannot say that the logging of burnt trees caused the greater percent loss of saplings.

Since the plots were not randomized to receive either the logging or not logging treatment, no causation can be implied here. Since the transect patterns were randomly selected, this inference can be generalized to the 16 larger plots.

Confirmation Using R

In this section we confirm our findings using R. The R code input is shown below:

Code 15.2. wilcoxon rank sum test using R

```
        loggingData <- read.csv("Data/Logging.csv",header=TRUE, sep=",")
        wilcox.test(PercentLost ~ Action,
        data = loggingData,
        exact = TRUE,
        alternative = "greater")
```

And the output:

```
      Wilcoxon rank sum test

      data:  PercentLost by Action
      W = 55, p-value = 0.005769
      alternative hypothesis: true location shift is greater than 0
```

The results of the two programs are identical!

Chapter 16

Problem 3: Welch's Two Sample T-Test with Education Data

16.1 Problem Statement and Assumptions

Problem Statement

We would like to examine the claim that the mean income of college educated people (16 years of education) is greater than the mean income of people with only a high school education (12 years of education)

Assumptions

The code used to produce everything in this section is shown below:

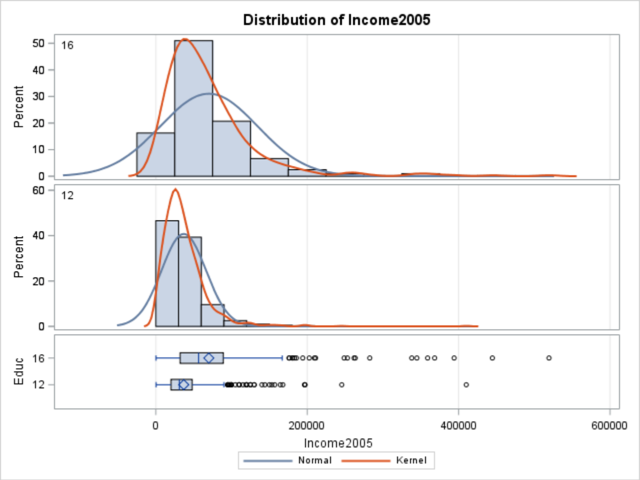
Code 16.1. welch's t test

```
proc ttest data=edudata order=DATA
sides=U; /*an Upper tailed test*/
class Educ;
var Income2005;
run;
```

Normality

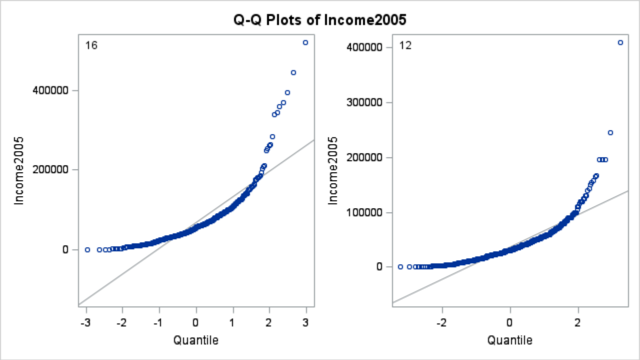
Figure 3.1 shows histograms and Box plots relating to the data:

Figure 16.1.1. Histograms and Box plots



As we can see from the figure, the data is not normal, it is heavily right skewed in both cases. Both the histograms and the Box plots show this, as the histograms are way taller on the left side than on the right, while the box plots show that there is a bunch of data on the left with a ton of outliers, clearly not normal. We examine this further with the Q-Q plot in Figure 3.2

Figure 16.1.2. Q-Q Plot



The Q-Q plot conifrims our findings that the data is not very normal. However, the sample sizes are 400 and 1000, which means that we can definitely apply the central limit theorem. This means that we can treat the data as normal, we will assume normality.

Independence

We will assume independence in this case.

16.2 Complete Analysis Using SAS

Statement of Hypotheses

$$H_0 : \mu_{16yeareduc} - \mu_{12yeareduc} \leq 0 \tag{16.2.1}$$

$$H_1 : \mu_{16yeareduc} - \mu_{12yeareduc} > 0 \tag{16.2.2}$$

Critical t Value

With $\alpha = .05$ and a one sided test, the critical t value (with the appropriate degrees of freedom) is calculated using the code shown below.

```
data critval;
p = quantile("T",.95,473.85); /*one sided test*/;
proc print data=critval;
run;
```

The critical t value is shown in Figure 3.3:

Figure 16.2.1. Critical t-value

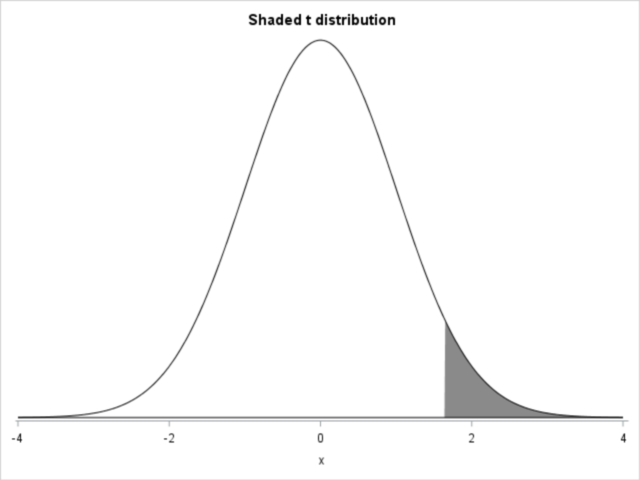
Obs	p
1	1.64808

The critical t value is $t = 1.64$. This is illustrated using the following bit of SAS code:

```
data pdf;
do x = -4 to 4 by .01;
pdf = pdf("T", x, 473.85);
lower = 0;
if x >= quantile("T",0.95,473.85) then upper = pdf;/*one sided*/
else upper = 0;
output;
end;
run;
title 'Shaded t distribution';
proc sgplot data=pdf noautolegend noborder;
yaxis display=none;
band x = x
lower = lower
upper = upper / fillattrs=(color=gray8a);
series x = x y = pdf / lineattrs = (color = black);
series x = x y = lower / lineattrs = (color = black);
run;
```

This produces Figure 3.4

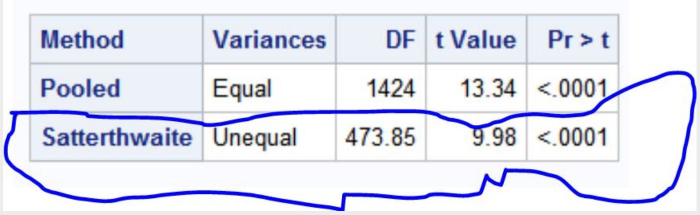
Figure 16.2.2. Shaded t Distribution



Calculation of the t Statistic

To calculate Welch's t Statistic, we use the code seen in Section 3.a.2, giving us a t value of $t = 9.98$, as seen in Figure 3.5

Figure 16.2.3. Results of Welch's t-test



Method	Variances	DF	t Value	Pr > t
Pooled	Equal	1424	13.34	<.0001
Satterthwaite	Unequal	473.85	9.98	<.0001

We see that in this case, we have a t-value of 9.98

Calculation of the p Value

We also see from Figure 3.5 that $p = 0$

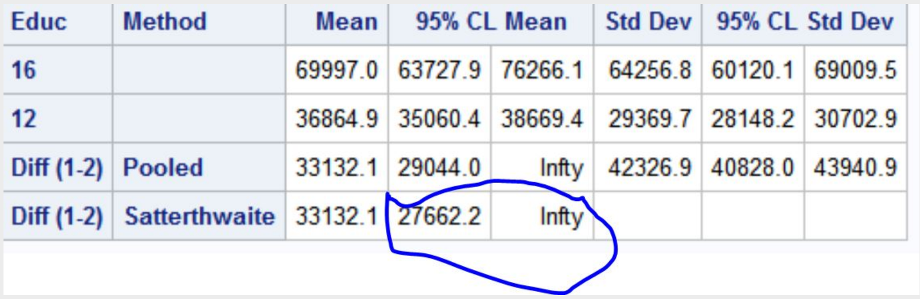
Results of Hypothesis Test

We have that $p = 0 < \alpha = .05$ and therefore we reject the null hypothesis

Conclusion

We have convincing evidence that the mean income of people with an education of 16 years is greater than the mean income of people with an education of 12 years. A one sided p-value of zero shows us that the means are truly different. The figure below shows a one sided 95% confidence interval on our data:

Figure 16.2.4. Confidence Interval on the Difference of Means



Educ	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
16		69997.0	63727.9	76266.1	64256.8	60120.1	69009.5
12		36864.9	35060.4	38669.4	29369.7	28148.2	30702.9
Diff (1-2)	Pooled	33132.1	29044.0	Infty	42326.9	40828.0	43940.9
Diff (1-2)	Satterthwaite	33132.1	27662.2	Infty			

The confidence interval on the difference of means is $[27662.2, \infty)$. This estimates what is a plausible difference between the means of the two samples. As we can see, the distribution of income of the sample with a 16-year education is at least $27,000 greater than the distribution of income of the sample with a 12-year education.

Scope of Inference

This was an observational study; therefore, we cannot conclude that the extra education caused the change (increase) in mean incomes. Households were selected from a random sample of a previously selected "area of the United States" and the subjects in this study are the members of those households. Therefore, since every member of the "area" had the same chance of being selected, it is a random sample of the "areas." However, no indication is given on how the "areas" were selected. In conclusion, the association between education and income above can be generalized to all the members of the "areas" that were selected for this study, but not generalized to the U.S. as a who

Verification using R

The following R code was used to verify the analysis

```
eduData <- read.csv("Data/EducationData.csv",header=TRUE, sep=",")
t.test(Income2005 ~ Educ,
data = eduData,
# we use less because R is doing 12 - 16 #
alternative = "less")
```

This gives the following output:

```
Welch Two Sample t-test

data:  Income2005 by Educ
t = -9.9827, df = 473.85, p-value < 2.2e-16
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
-Inf -27662.19
sample estimates:
mean in group 12 mean in group 16
36864.90         69996.97
```

Note that R is telling us that the distribution of income of the sample with a 12 year education is at least 27,000 less than those with a 16 year education

Preferences

I prefer the log transformed analysis, they both assume normality, however the log transformed analysis has the more actually normal data to start with, and the variances are roughly equal. It also speaks more to the medians, instead of the means, which is much more robust to the huge number of outliers. I think because of the outliers, I definitely prefer the log method, as the mean is not such a good measurement with these crazy outliers.

Chapter 17

Problem 4: Trauma and Metabolic Expenditure rank sum

17.1 Hand-Written Calculations

To summarize, $T = 82$, $\mu(T) = 56$, $sd(T) = 8.632$ The handwritten work was done before the author understood continuity correction, the continuity corrected Z and P values were calculated as follows:

$$Z = \frac{(T - 0.5) - mean(T)}{SD(T)} = 2.95 \tag{17.1.1}$$

$$\rightarrow p = .001568 \tag{17.1.2}$$

With a continuity correction of 0.5

#4

A)

Y	Group	order	Rank
18.8	N	1	1
20	N	2	2
20.1	N	3	3
20.9	N	4	4.5
20.9	N	5	4.5
21.4	N	6	6
22.0	Tr	7	7
22.7	N	8	8
22.9	N	9	9
23	Tr	10	10
24.5	Tr	11	11
25.3	Tr	12	12
30	Tr	13	13
37.6	Tr	14	14
38.5	Tr	15	15

Group 1 = Tr ~~smaller sample size~~

3) $T = \sum \text{Rank}(Tr) = \underline{82}$

→

Setup: H_0: Distribution(nontrauma) − Distribution(trauma) ≥ 0

H_1: Dist(nontrauma) − Dist(trauma) < 0

C)

$$\text{Mean}(T) = \overset{7}{n_{Tr}} \cdot \overset{8}{\overbrace{\frac{\sum R}{n_{Tr}+n_N}}} = 7 \cdot 8 = \underline{\underline{56}}$$

$$SD(T) = \underbrace{\left(\sqrt{\frac{n_{Tr} \cdot n_N}{(n_{Tr}+n_N)}}\right)}_{\underbrace{\sqrt{\frac{7 \cdot 8}{7+8}}}_{1.932}} \cdot \underbrace{\sqrt{\frac{\sum(R_i-\bar{R})^2}{(n_{Tr}+n_N-1)}}}_{4.468} = \underline{8.632}$$

$$Z \overset{\text{w/ cont correction}}{=} \frac{T - \text{mean}\, T}{SD(T)} = \frac{82-56}{8.632} = \boxed{3.012}$$

D)

$\boxed{P = 0.001298}$

$H_0: \text{Dist}(Tr) - \text{Dist}(N) \leq 0$

E) other useful values: $H_1: \text{Dist}(Tr) - \text{Dist}(N) > 0$

~~critical value:~~

α: .05

1-sided

critical: 1.64485

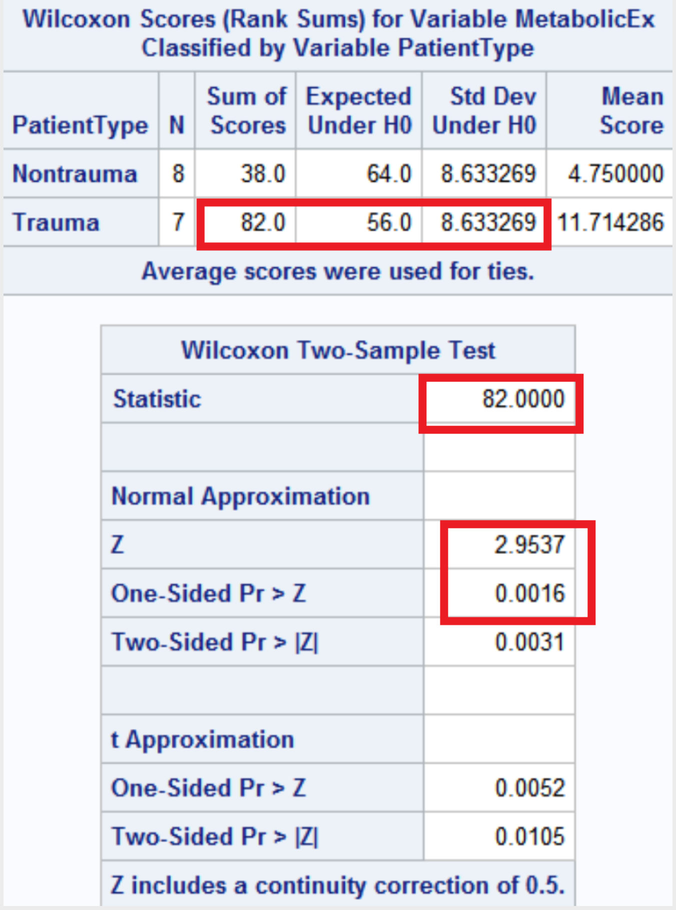
17.2 SAS verification

To verify the Z and p values calculated in Section 4.a, the following SAS code was run:

```
proc NPAR1WAY data=TraumaStudy Wilcoxon HL;
class PatientType;
Var MetabolicEx;
run;
```

The results of this code are shown in Figure 4.1

Figure 17.2.1. Continuity Corrected Wilcoxon Test Using SAS



Wilcoxon Scores (Rank Sums) for Variable MetabolicEx Classified by Variable PatientType

PatientType	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Nontrauma	8	38.0	64.0	8.633269	4.750000
Trauma	7	82.0	56.0	8.633269	11.714286

Average scores were used for ties.

Wilcoxon Two-Sample Test	
Statistic	82.0000
Normal Approximation	
Z	2.9537
One-Sided Pr > Z	0.0016
Two-Sided Pr > |Z|	0.0031
t Approximation	
One-Sided Pr > Z	0.0052
Two-Sided Pr > |Z|	0.0105

Z includes a continuity correction of 0.5.

The Results of the two tests are the same! Note that if you add the phrase "correct=no" to the proc NPAR1WAY statement, you get the same values as the non corrected ones in the handwritten work

17.3 Full Statistical Analysis

Problem Statement

We would like to test the claim that the Trauma patients had higher metabolic expenditures/

Assumptions

The Wilcoxon Rank-Sum test only assumes the data are independent, which in this case we will assume independence because the patients were not related to each other in any way, or at least their metabolic expenditures aren't dependent on the other people's metabolic expenditures. ALSO obviously normal

Hypothesis definitions

$$H_0 : meanRank_{Trauma} - meanRank_{NonTrauma} \leq 0 \tag{17.3.1}$$
$$H_1 : meanRank_{Trauma} - meanRank_{NonTrauma} > 0 \tag{17.3.2}$$

In other words, the null hypothesis is that the nontrauma and trauma patients have equal distributions of metabolic expenditures, while the alternative hypothesis claims that the distribution of the trauma patients' metabolic expenditures is higher. We are using a one sided hypothesis test because that is what the book calls for. In this scenario, we will say $\alpha = 0.05$

Critical Value

The critical value was calculated using the following chink of SAS code:

```
data critval;
p = quantile("Normal",.95); /*one sided test*/;
proc print data=critval;
run;
```

Producing a critical t value of $t = 1.64485$

Figure 17.3.1. Critical Value

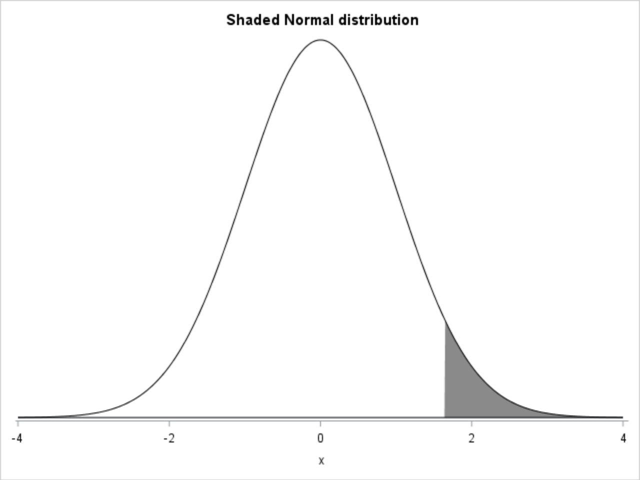
Obs	p
1	1.64485

The critical value is shown on a normal distribution using the following bit of SAS code

```
data pdf;
do x = -4 to 4 by .01;
pdf = pdf("Normal", x);
lower = 0;
if x >= quantile("Normal",0.95) then upper = pdf;/*one sided*/
else upper = 0;
output;
end;
run;
title 'Shaded Normal distribution';
proc sgplot data=pdf noautolegend noborder;
yaxis display=none;
band x = x
lower = lower
upper = upper / fillattrs=(color=gray8a);
series x = x y = pdf / lineattrs = (color = black);
series x = x y = lower / lineattrs = (color = black);
run;
```

The shaded distribution is displayed in Figure 4.3

Figure 17.3.2. Shaded Normal Distribution



Calculation of the z statistic

Our z statistic, calculated in Sections 4.a and 4.b is 2.95.

Calculation of the p value

Our p-value, calculated in Sections 4.a and 4.b is 0.0016

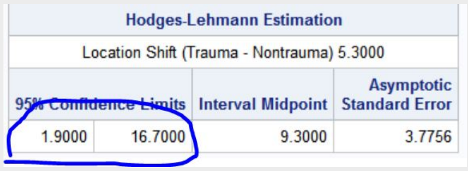
Discussion of the hypothesis

We Reject the null hypothesis, $p = .0016 < 0.5 = \alpha$

Conclusion

We have convincing evidence that the distribution of metabolic expenditure of trauma patients is than the non-trauma patients (p=0.0016 on a one sided Wilcoxon rank-sum test). The figure below shows a 95% Hodges-Lehmann confidence interval on the difference of the two distributions:

Figure 17.3.3. 95% Confidence Interval



Hodges-Lehmann Estimation			
Location Shift (Trauma - Nontrauma) 5.3000			
95% Confidence Limits		Interval Midpoint	Asymptotic Standard Error
1.9000	16.7000	9.3000	3.7756

This tells us that a plausible difference between the two distributions is between 1.9 and 16.7. As we can see this does not include the null hypothesis which says their difference is less than or equal to zero. This cannot give us causal or population inferences because it was neither a randomized experiment nor a random sample ALSO MEDIANS DUH

Chapter 18

Problem 5: Autism and Yoga signed rank

18.1 Hand-Written Calculations

The results of the calculations are as follows: $S = 41$, $\mu_S = 22.5$, $SD_S = 8.4409$, The Z value on the paper is incorrect, as it does not correct for continuity. So, here we will aplply the continuity correction:

$$z = \frac{S - 0.5 - \bar{S}}{SD_S} \tag{18.1.1}$$

$$z = \frac{40.5 - 22.5}{8.4409} = 2.13 \rightarrow p_{oneTail} = .0166 p_{twoTail} = .033 \tag{18.1.2}$$

#5

A Child	~~Be~~fore Before	After	Differ.	order mag	Ranks	+Ranks	-Ranks
1	85	75	10 ✓	5(-)	1		1
2	70	50	20 ✓	10	3	3 ~~8~~	
3	40	50	-10 ✓	10	3	3	
4	65	40	25 ✓	~~10~~ (-10)	3		3
5	80	20	60	15	5	5	
6	75	65	10 ✓	20	6	6	
7	55	40	15 ✓	25	7	7	
8 ~~9~~	20	25	-5 ✓	40	8	8	
9	70	30	40	60	9	9	

$S = 41$

$$\text{Mean}(S) = \frac{n(n+1)}{4} = \frac{90}{4} = 22.5$$

$$SD(S) = \sqrt{\frac{n(n+1)(2n+1)}{24}} = 8.4409$$

$$= \sqrt{\frac{9 \cdot 10 \cdot 19}{24}}$$

$$Z = \frac{41 - 22.5}{8.4409} = \boxed{2.19}$$

18.2 Verification in SAS and R

Verification in SAS

To verify this, the following bit of SAS code was employed: Producing:

Code 18.1. Signed Rank test in SAS

```
data Autismdiff;
set Autism;
diff= Before-After;
run;
proc univariate data=Autismdiff;
var diff;
run;
```

Figure 18.2.1. Signed Rank Test In SAS

Signed Rank	S	18.5	Pr >= |S|	0.0313

This two sided p value of 0.0313 is the same as a one sided p value of .01565, and a z value of 2.15. It is slightly different with my calculations and SAS's because they didnt use a normal approximation, I did.

Verification in R

This R code was employed for the same purposes:

```
      AutismData <- read.csv("Data/Autism.csv",header=TRUE, sep=",")
      wilcox.test(AutismData\$Before, AutismData\$After,
      paired = TRUE,
      alternative = "greater",
      conf.int=TRUE)
```

Yielding:

```
      Wilcoxon signed rank test with continuity correction

      data:  AutismData\$Before and AutismData\$After
      V = 41, p-value = 0.01618
      alternative hypothesis: true location shift is greater than 0
      95 percent confidence interval:
      4.999993      Inf
      sample estimates:
      (pseudo)median
      17.49993
```

The R code applied a continuity correction, instead of doing the exact permutation like SAS. Their P value corresponds with a Z score of 2.139

18.3 6 step Sign Rank test using SAS

Statement of Hypothesis

$$H_0 : Median_{Before} - Median_{After} \leq 0 \tag{18.3.1}$$

$$H_1 : Median_{Before} - Median_{After} > 0 \tag{18.3.2}$$

We will say that $\alpha = .05$ and we are doing a one sided test

Critical Values

The critical value was calculated using the following chunk of SAS code:

```
data critval;
p = quantile("Normal",.95); /*one sided test*/;
proc print data=critval;
run;
```

Producing a critical t value of $t = 1.64485$

Figure 18.3.1. Critical Value

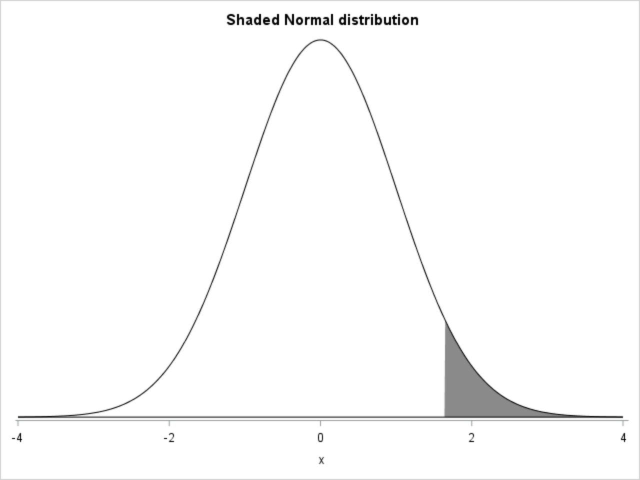
Obs	p
1	1.64485

The critical value is shown on a normal distribution using the following bit of SAS code

```
data pdf;
do x = -4 to 4 by .01;
pdf = pdf("Normal", x);
lower = 0;
if x >= quantile("Normal",0.95) then upper = pdf;/*one sided*/
else upper = 0;
output;
end;
run;
title 'Shaded Normal distribution';
proc sgplot data=pdf noautolegend noborder;
yaxis display=none;
band x = x
lower = lower
upper = upper / fillattrs=(color=gray8a);
series x = x y = pdf / lineattrs = (color = black);
series x = x y = lower / lineattrs = (color = black);
run;
```

The shaded distribution is displayed in Figure 5.3

Figure 18.3.2. Shaded Normal Distribution



Calculation of a Z statistic

We will use the Z statistic calculated using R/by hand,Z = 2.13, however it will not have a huge effect on the outcome of the test

Calculation of a p value

For our z value, a one sided p value is $p = 0.016$.

Assessment of hypothesis

$p = .016 < \alpha = .05 \rightarrow$We reject the null hypothesis.

Conclusion

We have conclusive evidence that the median time to complete the puzzle for Autistic children is greater before 20 minutes of Yoga than after 20 minutes of Yoga. We cannot infer causality becuase this was not a randomized experiment, and we cannot infer anything about the population because this was not a random sample. The median time for the children was at least 5 seconds longer before Yoga as compared to after Yoga, as seen by the confidence interval displayed in the R output.

18.4 Paired t test in SAS

Statement of Hypothesis

$$H_0 : \mu_{before-after} \leq 0 \tag{18.4.1}$$

$$H_1 : \mu before - after > 0 \tag{18.4.2}$$

We will say that $\alpha = .05$ and we are doing a one sided test.

Critical Values

The critical value was calculated using the following chunk of SAS code:

```
data critval;
p = quantile("T",.95,8); /*one sided test*/;
proc print data=critval;
run;
```

With the following output:

Figure 18.4.1. Critical Value

Obs	p
1	1.85955

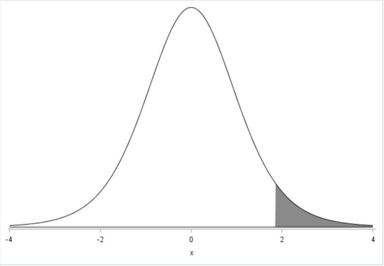
With a critical t value of t=1.86. This is demonstrated in a shaded t distribution with the following chunk of code:

```
data pdf;
do x = -4 to 4 by .01;
pdf = pdf("T", x,8);
lower = 0;
if x >= quantile("T",0.95,8) then upper = pdf;/*one sided*/
else upper = 0;
output;
end;
run;
```

```
title 'Shaded Normal distribution';
proc sgplot data=pdf noautolegend noborder;
yaxis display=none;
band x = x
lower = lower
upper = upper / fillattrs=(color=gray8a);
series x = x y = pdf / lineattrs = (color = black);
series x = x y = lower / lineattrs = (color = black);
run;
```

The shaded distribution is displayed in Figure 5.5

Figure 18.4.2. Shaded T Distribution



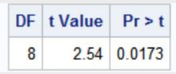
Calculation of a t statistic

The T statistic was calculated using the following SAS code: The t value is shown in Figure 5.6

Code 18.2. Paired T test in SAS

```
proc ttest data=Autism alpha = .05 sides=U;
paired Before*After;
run;
```

Figure 18.4.3. Paired t statistic



DF	t Value	Pr > t
8	2.54	0.0173

We have a t value of 2.54.

Calculation of a P value

The p value can be seen in Figure 5.6: $p = .0173$

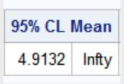
Assessment of Hypothesis

$p = .0173 > \alpha = .05$ →we reject the null hypothesis.

Conclusion

We have conclusive evidence that the mean of the differences of times before and after the yoga is greater than zero (p=.0173 on a one sided paired t test). A confidence interval for the mean of the difference of time for the children to finish the puzzle before and after yoga is shown in Figure 5.7:

Figure 18.4.4. 95% Confidence interval



95% CL Mean	
4.9132	Infty

This means that the mean of the differences was at least 4.9 seconds. We cannot infer causality because this was not a randomized experiment, and we cannot make inferences about the population because this was not a random sample. We also cannot make causal inferences with a paired t test

18.5 Confirmation with R

The R code below was used to verify the results of the previous section:

```
        t.test(AutismData\$Before, AutismData\$After,
        paired = TRUE,
        alternative = "greater",
        conf.int=TRUE)
```

The output is presented below:

```
        Paired t-test

        data:  AutismData\$Before and AutismData\$After
        t = 2.5403, df = 8, p-value = 0.01735
        alternative hypothesis: true difference in means is greater than 0
        95 percent confidence interval:
        4.913201       Inf
        sample estimates:
        mean of the differences
        18.33333
```

18.6 Complete Statistical Analysis

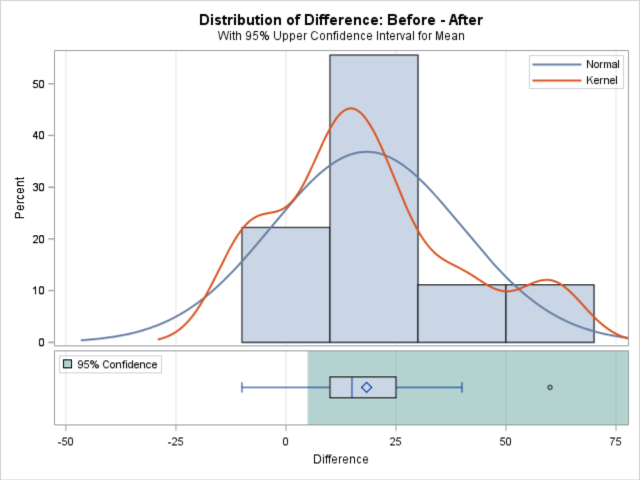
In this section, I will be using a paired t-test, because the data is pretty normal, as we will see in the following section. When both are possible, I believe the paired t test is better because it doesnt mess with the data in any way, we can see the magnitudes etc.

Assumptions

We can assume the differences are independent because the children did not affect the other children.

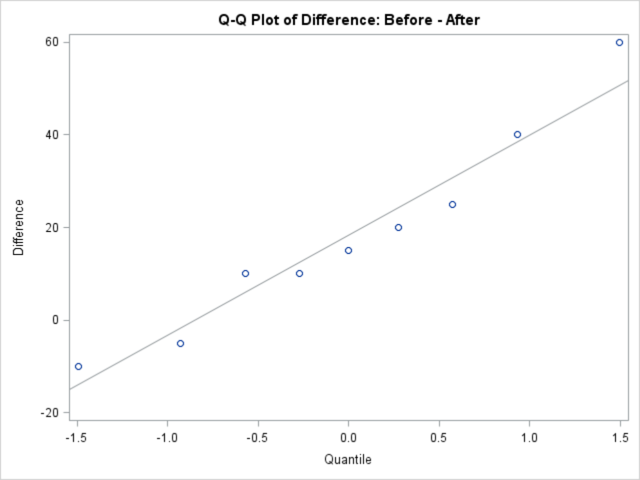
To check for normality we examine the following figure:

Figure 18.6.1. Histogram and Box Plot



As we see from Figure 5.8, the data is fairly normally distributed. The histogram is heavier in the center than on the edges, and the mean is near the median on the Box plot. We will examine this further in Figure 5.9

Figure 18.6.2. Q-Q Plot



As we can see, the data follows the line of normality closely, and therefore we can assume normality. This means that a paired t test is appropriate.

Statement of Hypothesis

$$H_0 :\mu_{before-after} \leq 0 \tag{18.6.1}$$

$$H_1 :\mu before - after > 0 \tag{18.6.2}$$

We will say that $\alpha = .05$ and we are doing a one sided test.

Critical Values

The critical value was calculated using the following chunk of SAS code:

```
data critval;
p = quantile("T",.95,8); /*one sided test*/;
proc print data=critval;
run;
```

With the following output:

Figure 18.6.3. Critical Value

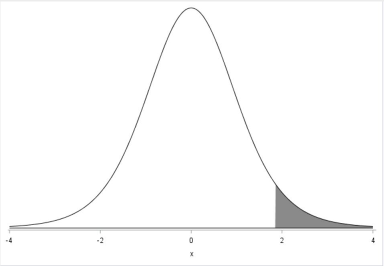
Obs	p
1	1.85955

With a critical t value of t=1.86. This is demonstrated in a shaded t distribution with the following chunk of code:

```
data pdf;
do x = -4 to 4 by .01;
pdf = pdf("T", x,8);
lower = 0;
if x >= quantile("T",0.95,8) then upper = pdf;/*one sided*/
else upper = 0;
output;
end;
run;
title 'Shaded Normal distribution';
proc sgplot data=pdf noautolegend noborder;
yaxis display=none;
band x = x
lower = lower
upper = upper / fillattrs=(color=gray8a);
series x = x y = pdf / lineattrs = (color = black);
series x = x y = lower / lineattrs = (color = black);
run;
```

The shaded distribution is displayed in Figure 5.11

Figure 18.6.4. Shaded T Distribution



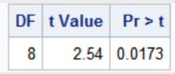
Calculation of a t statistic

The T statistic was calculated using the following SAS code:

```
proc ttest data=Autism alpha = .05 sides=U;
paired Before*After;
run;
```

The t value is shown in Figure 5.12

Figure 18.6.5. Paired t statistic



DF	t Value	Pr > t
8	2.54	0.0173

We have a t value of 2.54.

Calculation of a P value

The p value can be seen in Figure 5.6: $p = .0173$

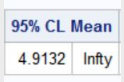
Assessment of Hypothesis

$p = .0173 > \alpha = .05 \rightarrow$we reject the null hypothesis.

Conclusion

We have conclusive evidence that the mean of the differences of times before and after the yoga is greater than zero (p=.0173 on a one sided paired t test). A confidence interval for the mean of the difference of time for the children to finish the puzzle before and after yoga is shown in Figure 5.13:

Figure 18.6.6. 95% Confidence interval



95% CL Mean	
4.9132	Infty

This means that the mean of the differences was at least 4.9 seconds. We cannot infer causality because this was not a randomized experiment, and we cannot make inferences about the population because this was not a random sample. We also cannot make causal inferences with a paired t test

Chapter 19

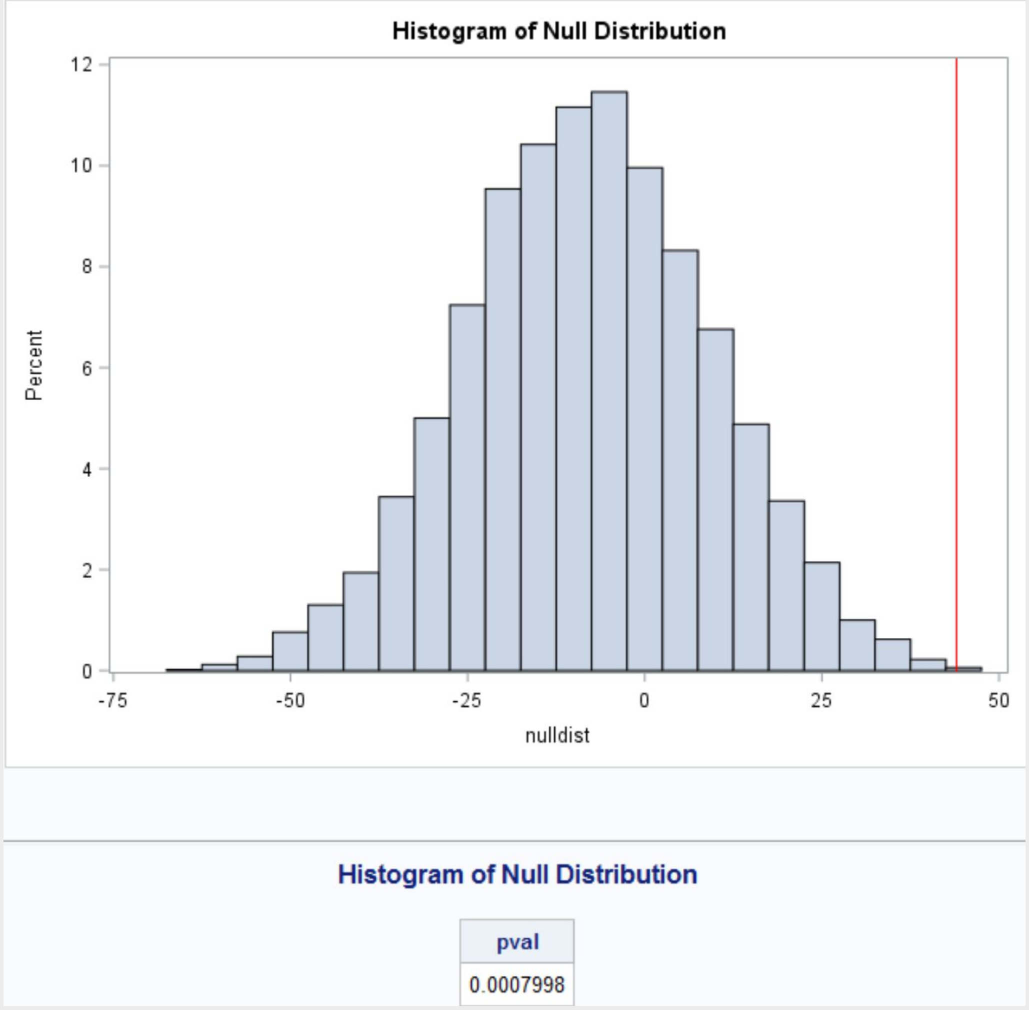
sexy ranked permutation test

Here is the SAS code I designed to conduct a Ranked permutation test I did not have time to add a normal curve to my figure, however, the p value is more or less the same as the wilcoxon test however it is a more reasonable number.

Code 19.1. handcrafted rank sum test

```
proc import
datafile='c:\Users\david\Desktop\MSDS\MSDS6371\Homework\Week4\Data\Trauma.csv'
out=TraumaStudy
DBMS=CSV;
run;
proc rank data=TraumaStudy out=Ranked ties=mean;
var MetabolicEx;
ranks rank;
run;
proc print data=Ranked;
run;


proc iml;
use Ranked var {PatientType rank};
/*making two groups in IML*/
read all var {rank} where(PatientType='Nontrauma') into g2;
read all var {rank} where(PatientType='Trauma') into g1;
obsdiff = sum(g1) - sum(g2);
print obsdiff;
call randseed(12345);                              /* set random number seed */
alldata = g1 // g2;                         /* stack data in a single vector */
N1 = nrow(g1);  N = N1 + nrow(g2);
NRepl = 5000;                                      /* number of permutations */
nulldist = j(NRepl,1);                    /* allocate vector to hold results */
do k = 1 to NRepl;
x = sample(alldata, N, "WOR");                        /* permute the data */
nulldist[k] = sum(x[1:N1]) - sum(x[(N1+1):N]);   /* difference of sums */
end;

title "Histogram of Null Distribution";
refline = "refline " + char(obsdiff) + " / axis=x lineattrs=(color=red);";
call Histogram(nulldist) other=refline ;

pval = (1 + sum((nulldist) >= (obsdiff))) / (NRepl+1); /*this means one sided test, no
print pval;
quit;
```

Figure 19.0.1. Permutation Test



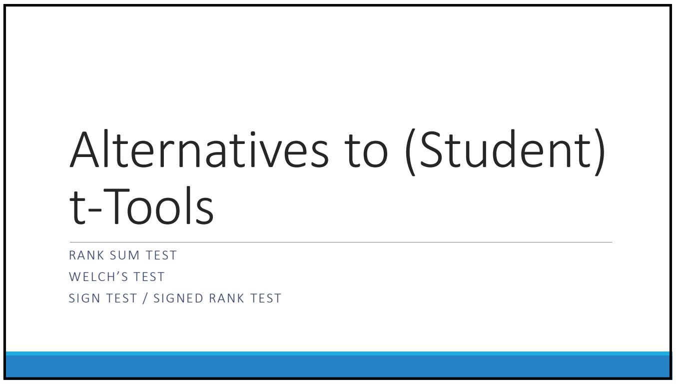
Histogram of Null Distribution

pval
0.0007998

Chapter 20

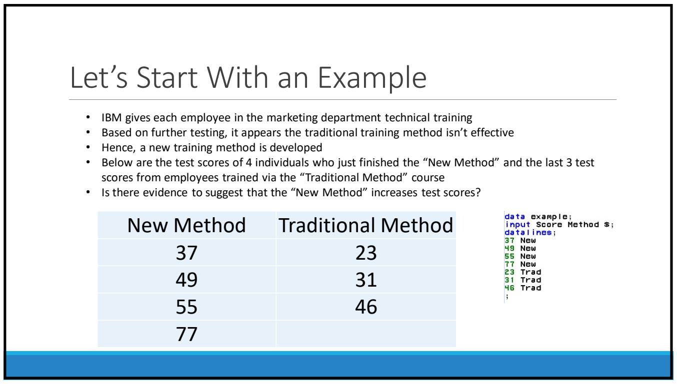
Unit 4 lecture slides

Here it is



Alternatives to (Student) t-Tools

RANK SUM TEST
WELCH'S TEST
SIGN TEST / SIGNED RANK TEST

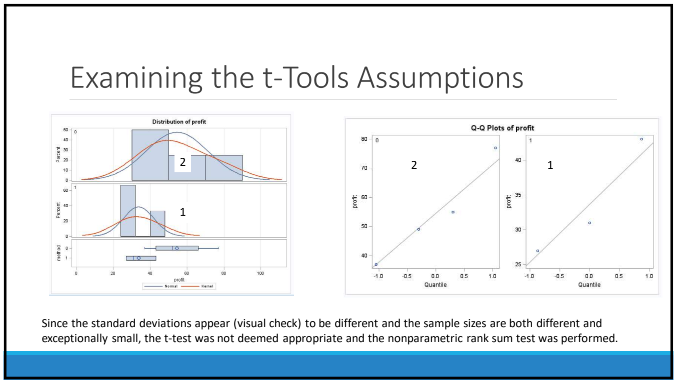


Let's Start With an Example

- IBM gives each employee in the marketing department technical training
- Based on further testing, it appears the traditional training method isn't effective
- Hence, a new training method is developed
- Below are the test scores of 4 individuals who just finished the "New Method" and the last 3 test scores from employees trained via the "Traditional Method" course
- Is there evidence to suggest that the "New Method" increases test scores?

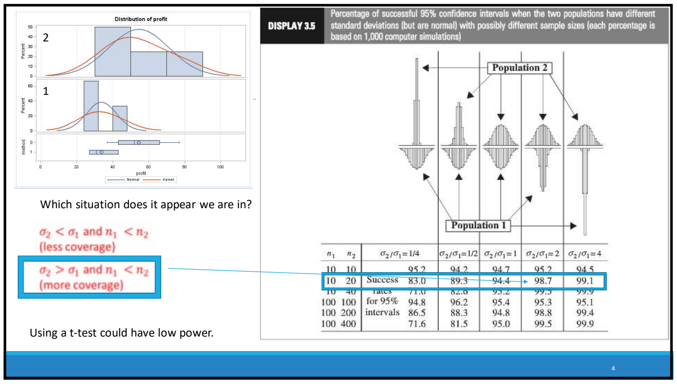
New Method	Traditional Method
37	23
49	31
55	46
77	

```
data example;
input Score Method $;
datalines;
37 New
49 New
55 New
77 New
23 Trad
31 Trad
46 Trad
;
```



Examining the t-Tools Assumptions

Since the standard deviations appear (visual check) to be different and the sample sizes are both different and exceptionally small, the t-test was not deemed appropriate and the nonparametric rank sum test was performed.



Which situation does it appear we are in?

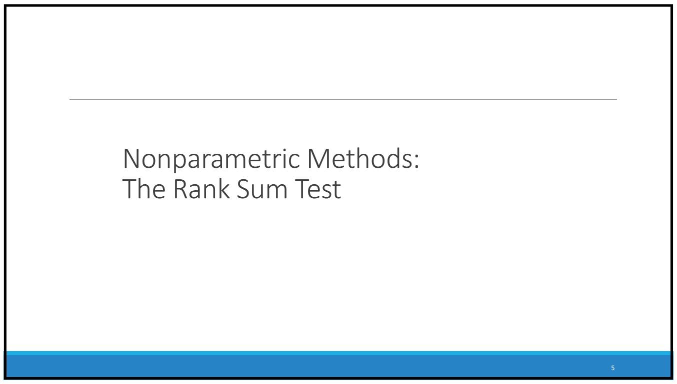
$\sigma_2 < \sigma_1$ and $n_1 < n_2$
(less coverage)

$\sigma_2 > \sigma_1$ and $n_1 < n_2$
(more coverage)

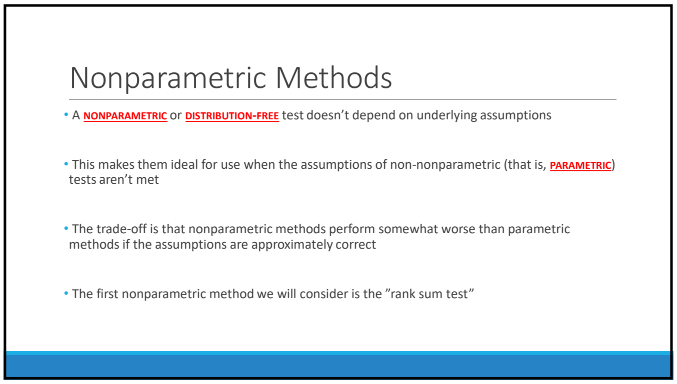
Using a t-test could have low power.

DISPLAY 3.5 Percentage of successful 95% confidence intervals when the two populations have different standard deviations (but are normal) with possibly different sample sizes (each percentage is based on 1,000 computer simulations)

n_1	n_2		$\sigma_2/\sigma_1=1/4$	$\sigma_2/\sigma_1=1/2$	$\sigma_2/\sigma_1=1$	$\sigma_2/\sigma_1=2$	$\sigma_2/\sigma_1=4$
10	10		95.2	94.2	94.7	95.2	94.5
10	20	Success	83.0	89.3	94.4	98.7	99.1
10	40	rates	71.0	82.6	95.2	99.5	99.9
100	100	for 95%	94.8	96.2	95.4	95.3	95.1
100	200	intervals	86.5	88.3	94.8	98.8	99.4
100	400		71.6	81.5	95.0	99.5	99.9



Nonparametric Methods: The Rank Sum Test



Nonparametric Methods

- A **NONPARAMETRIC** or **DISTRIBUTION-FREE** test doesn't depend on underlying assumptions
- This makes them ideal for use when the assumptions of non-nonparametric (that is, **PARAMETRIC**) tests aren't met
- The trade-off is that nonparametric methods perform somewhat worse than parametric methods if the assumptions are approximately correct
- The first nonparametric method we will consider is the "rank sum test"

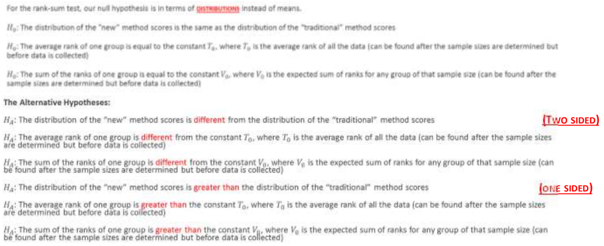
Rank Sum Test: Advantages

- No distributional assumptions
- Resistant to outliers
- Performs nearly as well as the t-test when the two populations are normal and considerably better when there are extreme outliers
- Works well with **ORDINAL** (as opposed to interval data)
- Works with censored values

- It still requires some assumptions:
  1. All observations are independent
  2. The *Y* values are ordinal

59 patients with arthritis who participated in a clinical trial were assigned to two groups, active and placebo. The response status:
(excellent=5, good=4, moderate=3, fair=2, poor=1)
of each patient was recorded.

The Hypothesis Test



For the rank-sum test, our null hypothesis is in terms of DISTRIBUTIONS instead of means.

H_0: The distribution of the "new" method scores is the same as the distribution of the "traditional" method scores

H_0: The average rank of one group is equal to the constant T_0, where T_0 is the average rank of all the data (can be found after the sample sizes are determined but before data is collected)

H_0: The sum of the ranks of one group is equal to the constant V_0, where V_0 is the expected sum of ranks for any group of that sample size (can be found after the sample sizes are determined but before data is collected)

The Alternative Hypotheses:

H_A: The distribution of the "new" method scores is different from the distribution of the "traditional" method scores (TWO SIDED)

H_A: The average rank of one group is different from the constant T_0, where T_0 is the average rank of all the data (can be found after the sample sizes are determined but before data is collected)

H_A: The sum of the ranks of one group is different from the constant V_0, where V_0 is the expected sum of ranks for any group of that sample size (can be found after the sample sizes are determined but before data is collected)

H_A: The distribution of the "new" method scores is greater than the distribution of the "traditional" method scores (ONE SIDED)

H_A: The average rank of one group is greater than the constant T_0, where T_0 is the average rank of all the data (can be found after the sample sizes are determined but before data is collected)

H_A: The sum of the ranks of one group is greater than the constant V_0, where V_0 is the expected sum of ranks for any group of that sample size (can be found after the sample sizes are determined but before data is collected)

The Rank Sum test

- We can compute the rank sum test statistic using the following steps:
  1. List all observations from both groups in increasing order
  2. Assign each observation a rank, from 1 to *n*
  3. If there are any ties, assign each tied observation's rank to be the average of their ranks.
  4. Identify each observation by its group

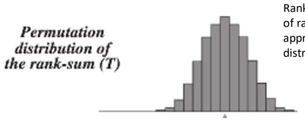
Note: *n* is the total # of observations

- The test statistic, *T*, is the sum of the ranks in one of the groups.

- We can find a p-value in two ways:
  - Normal approximation
  - Re-randomization (exact or approximate)

The Sampling Distribution of …

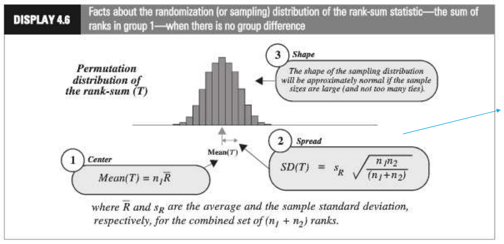
The Rank Sum Statistic!



Rank Sum test statistic (sum of ranks of one group) is approximately normally distributed!

Rank-Sum Test: Normal Approximation

DISPLAY 4.6 Facts about the randomization (or sampling) distribution of the rank-sum statistic—the sum of ranks in group 1—when there is no group difference



$$Mean(T) = n_1\bar{R}$$

$$SD(T) = s_R\sqrt{\frac{n_1 n_2}{(n_1+n_2)}}$$

where $\bar{R}$ and s_R are the average and the sample standard deviation, respectively, for the combined set of $(n_1 + n_2)$ ranks.

$$Z = \frac{T - Mean(T)}{SD(T)}$$

Rank Sum Test: randomly assign ranks

Name	Order #	Group	Rank
Bob	1	New	5
Sue	2	New	7
Fred	3	New	2
Jim	4	New	1
Pam	5	Trad	3
Tim	6	Trad	4
Zac	7	Trad	6

Name	Order #	Group	Rank
Sue	1	New	7
Bob	2	New	5
Fred	3	New	2
Jim	4	New	1
Pam	5	Trad	3
Tim	6	Trad	4
Zac	7	Trad	6

…

Name	Order #	Group	Rank
Pam	1	New	3
Tim	2	New	4
Sue	3	New	7
Zac	4	New	6
Fred	5	Trad	2
Bob	6	Trad	5
Jim	7	Trad	1

Record sum of ranks of one group (e.g. "Trad.") for all 7! permutations of ranks. (7!=7*6*5*4*3*2*1=5040)
P-value is the number of permutations with a sum equal to or more extreme than the one in the original data set divided by the total number of permutations.

*Could also do an approximate p-value by randomly choosing, say, 1000 orderings of the data.

Rank-Sum Test:
Normal Approximation

Common interpretation:
H_0: The distribution of New Method Scores = The distribution of the Traditional Method Scores
H_1:The distribution of New Method Scores > The distribution of the Traditional Method Scores

Technical mathematical interpretation:
H_0: Average *rank of New Method Scores* = Average *rank of all Scores (constant)*
H_1: Average rank of New Method Scores > Average rank of all Scores (constant)

```
proc npar1way data =  example Wilcoxon;
class Method;
var Score;
run;
```

There is mild evidence (alpha = 0.1) to suggest that the *distribution* of scores from the “New” method is greater than the *distribution* of the “Traditional” method (normal approximation to rank-sum test p-value = 0.0558).

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable Score Classified by Variable Method

Method	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
New	4	21.0	16.0	2.828427	5.250000
Trad	3	7.0	12.0	2.828427	2.333333

Wilcoxon Two-Sample Test	
Statistic	7.0000
Normal Approximation	
Z	-1.5910
One-Sided Pr < Z	0.0558
Two-Sided Pr > |Z|	0.1116
t Approximation	
One-Sided Pr < Z	0.0814
Two-Sided Pr > |Z|	0.1627
Z includes a continuity correction of 0.5.	

Rank-Sum Test:
Normal Approximation

Common interpretation:
H_0: The distribution of New Method Scores = The distribution of the Traditional Method Scores
H_1:The distribution of New Method Scores > The distribution of the Traditional Method Scores

There is mild evidence (alpha = 0.1) to suggest that the *distribution* of scores from the “New” method is greater than the *distribution* of the “Traditional” method (normal approximation to rank-sum test p-value = 0.0558).

```
proc npar1way data =  example Wilcoxon;
class Method;
var Score;
run;
```

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable Score Classified by Variable Method

Method	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
New	4	21.0	16.0	2.828427	5.250000
Trad	3	7.0	12.0	2.828427	2.333333

Wilcoxon Two-Sample Test	
Statistic	7.0000
Normal Approximation	
Z	-1.5910
One-Sided Pr < Z	0.0558
Two-Sided Pr > |Z|	0.1116
t Approximation	
One-Sided Pr < Z	0.0814
Two-Sided Pr > |Z|	0.1627
Z includes a continuity correction of 0.5.	

Permutation Test
(Exact P-value)

```
data example;
input Score Method $;
datalines;
37 New
49 New
55 New
77 New
23 Trad
31 Trad
46 Trad
;
```

```
proc npar1way data =  example Wilcoxon;
class Method;
var Score;
exact;
run;
```

Normal approximation p-values

Exact p-values

Wilcoxon Scores (Rank Sums) for Variable Score Classified by Variable Method

Method	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
New	4	21.0	16.0	2.828427	5.250000
Trad	3	7.0	12.0	2.828427	2.333333

Wilcoxon Two-Sample Test	
Statistic (S)	7.0000
Normal Approximation	
Z	-1.5910
One-Sided Pr < Z	0.0558
Two-Sided Pr > |Z|	0.1116
t Approximation	
One-Sided Pr < Z	0.0814
Two-Sided Pr > |Z|	0.1627
Exact Test	
One-Sided Pr <= S	0.0571
Two-Sided Pr >= |S - Mean|	0.1143
Z includes a continuity correction of 0.5.	

Rank Sum Test (Wilcoxon)

H_0: *The distribution of New Method Scores = The distribution of the Traditional Method Scores*

H_1: *The distribution of New Method Scores > The distribution of the Traditional Method Scores*

```
proc npar1way data = example Wilcoxon;
 class Method;
 var score;
 exact;
 run;
```

There is sufficient evidence at the alpha = 0.1 level of significance (p-value = .0571 for the exact test) to suggest that the *distribution* of scores from four IBM employees that were given the New Method is greater than the *distribution* of the 3 employees that took the test having had the Traditional Method of instruction.

Wilcoxon Scores (Rank Sums) for Variable Score Classified by Variable Method

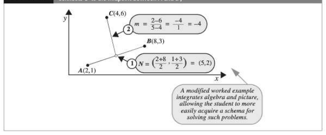
Method	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
New	4	21.0	16.0	2.828427	5.250000
Trad	3	7.0	12.0	2.828427	2.333333

Wilcoxon Two-Sample Test	
Statistic (S)	7.0000
Normal Approximation	
Z	-1.5910
One-Sided Pr < Z	0.0558
Two-Sided Pr > |Z|	0.1116
t Approximation	
One-Sided Pr < Z	0.0814
Two-Sided Pr > |Z|	0.1627
Exact Test	
One-Sided Pr <= S	0.0571
Two-Sided Pr >= |S - Mean|	0.1143
Z includes a continuity correction of 0.5.	

Cognitive Load Experiment

- Researchers compared the effectiveness of conventional textbook examples to modified ones
- They selected 28 ninth-year students who had no previous exposure to coordinate geometry
- The students were randomly assigned to one of two self study instructional groups, using conventional and modified instructional materials
- After instruction, they were given a test and the time to complete one of the problems was recorded.

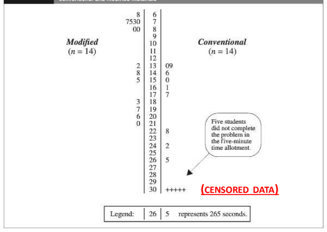
DISPLAY 4.3 Cognitive load experiment: modified method of instruction (for finding the slope of the line that connects *C* to the midpoint between *A* and *B*)



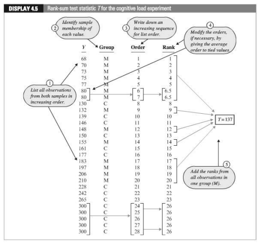
Is there sufficient evidence to suggest that the cognitive load theory (modified instruction) shortened response times?

Cognitive Load Experiment

DISPLAY 4.4 Number of seconds to solution of a problem in coordinate geometry, for students instructed with conventional and modified materials



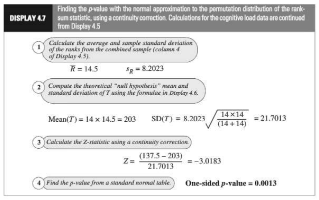
Cognitive Load Experiment


DISPLAY 4.5 Rank-sum test statistic *T* for the cognitive load experiment

Y	Group	Order	Rank
68	M	1	1
70	M	2	2
73	M	3	3
75	M	4	4
77	M	5	5
80	M	6	6.5
80	M	7	6.5
130	C	8	8
132	M	9	9
139	C	10	10
146	C	11	11
148	M	12	12
150	C	13	13
155	M	14	14
161	C	15	15
177	C	16	16
183	M	17	17
197	M	18	18
206	M	19	19
210	M	20	20
228	C	21	21
242	C	22	22
265	C	23	23
300	C	24	26
300	C	25	26
300	C	26	26
300	C	27	26
300	C	28	26

With ties, the ranks are averaged.

Cognitive Load Experiment: Normal Approximation


DISPLAY 4.7 Finding the *p*-value with the normal approximation to the permutation distribution of the rank-sum statistic, using a continuity correction. Calculations for the cognitive load data are continued from Display 4.5

① *Calculate the average and sample standard deviation of the ranks from the combined sample (column 4 of Display 4.5).*

$$\bar{R} = 14.5 \qquad s_R = 8.2023$$

② *Compute the theoretical "null hypothesis" mean and standard deviation of T using the formulae in Display 4.6.*

$$\text{Mean}(T) = 14 \times 14.5 = 203 \qquad \text{SD}(T) = 8.2023\sqrt{\frac{14 \times 14}{(14+14)}} = 21.7013$$

③ *Calculate the Z-statistic using a continuity correction.*

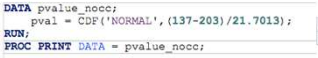
$$Z = \frac{(137.5 - 203)}{21.7013} = -3.0183$$

④ *Find the p-value from a standard normal table.* **One-sided *p*-value = 0.0013**

(CONTINUITY CORRECTION)

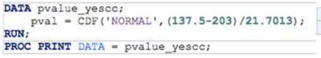
Statistical Conclusion: The data provide convincing evidence that a student could solve the problem more quickly after the "modified" rather than the the "conventional" method (one-sided, normal approximation w/ C.C. p-value = 0.0013, from the rank-sum test).

Cognitive Load Experiment: Using SAS

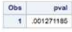


```
DATA pvalue_nocc;
    pval = CDF('NORMAL',(137-203)/21.7013);
RUN;
PROC PRINT DATA = pvalue_nocc;
```

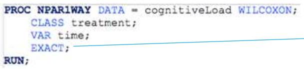
Obs	pval
1	.001177825



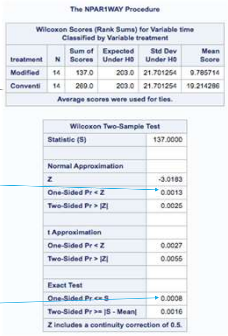
```
DATA pvalue_yescc;
    pval = CDF('NORMAL',(137.5-203)/21.7013);
RUN;
PROC PRINT DATA = pvalue_yescc;
```



Obs	pval
1	.001271185



```
PROC NPAR1WAY DATA = cognitiveLoad WILCOXON;
    CLASS treatment;
    VAR time;
    EXACT;
RUN;
```


The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable time Classified by Variable treatment

treatment	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Modified	14	137.0	203.0	21.701254	9.785714
Conventi	14	269.0	203.0	21.701254	19.214286
Average scores were used for ties.					

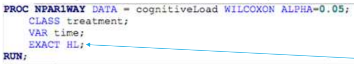
Wilcoxon Two-Sample Test	
Statistic (S)	137.0000
Normal Approximation	
Z	-3.0183
One-Sided Pr < Z	0.0013
Two-Sided Pr > |Z|	0.0025
t Approximation	
One-Sided Pr < Z	0.0027
Two-Sided Pr > |Z|	0.0055
Exact Test	
One-Sided Pr <= S	0.0008
Two-Sided Pr >= |S - Mean|	0.0016
Z includes a continuity correction of 0.5.	

Confidence Interval for the Location Parameter (Median): Hodges Lehmann Confidence Interval

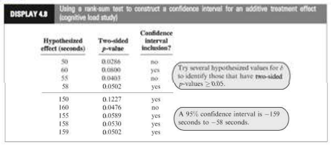
https://en.wikipedia.org/wiki/Hodges%E2%80%93Lehmann_estimator

*We will look at an example later

Cognitive Load Experiment



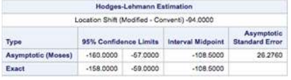
```
PROC NPAR1WAY DATA = cognitiveLoad WILCOXON ALPHA=0.05;
    CLASS treatment;
    VAR time;
    EXACT HL;
RUN;
```


DISPLAY 4.8 Using a rank-sum test to construct a confidence interval for an additive treatment effect (cognitive load study)

Hypothesized effect (seconds)	Two-sided *p*-value	Confidence interval inclusion?
50	0.0286	no
60	0.0800	yes
55	0.0403	no
58	0.0502	yes
150	0.1227	yes
160	0.0476	no
155	0.0589	yes
158	0.0530	yes
159	0.0502	yes

Try several hypothesized values for δ to identify those that have two-sided *p*-values ≥ 0.05.

A 95% confidence interval is −159 seconds to −58 seconds.

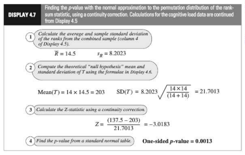

Hodges-Lehmann Estimation

Location Shift (Modified - Conventi) -94.0000

Type	95% Confidence Limits		Interval Midpoint	Asymptotic Standard Error
Asymptotic (Moses)	-160.0000	-57.0000	-108.5000	26.2760
Exact	-158.0000	-59.0000	-108.5000	

Statistical Conclusion (continued): A range of plausible values for how much smaller the "modified" distribution is than the "traditional" (treatment effect) is [-158, -59] s. (95% confidence interval based on a rank-sum test) with a point-estimate of 108.5 s.

Cognitive Load Experiment (All Together)


DISPLAY 4.7 Finding the *p*-value with the normal approximation to the permutation distribution of the rank-sum statistic, using a continuity correction. Calculations for the cognitive load data are continued from Display 4.5

① *Calculate the average and sample standard deviation of the ranks from the combined sample (column 4 of Display 4.5).*

$$\bar{R} = 14.5 \qquad s_R = 8.2023$$

② *Compute the theoretical "null hypothesis" mean and standard deviation of T using the formulae in Display 4.6.*

$$\text{Mean}(T) = 14 \times 14.5 = 203 \qquad \text{SD}(T) = 8.2023\sqrt{\frac{14 \times 14}{(14+14)}} = 21.7013$$

③ *Calculate the Z-statistic using a continuity correction.*

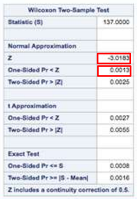
$$Z = \frac{(137.5 - 203)}{21.7013} = -3.0183$$

④ *Find the p-value from a standard normal table.* **One-sided *p*-value = 0.0013**

```
PROC NPAR1WAY DATA = cognitiveLoad WILCOXON ALPHA=0.05;
    CLASS treatment;
    VAR time;
    EXACT HL;
RUN;
```

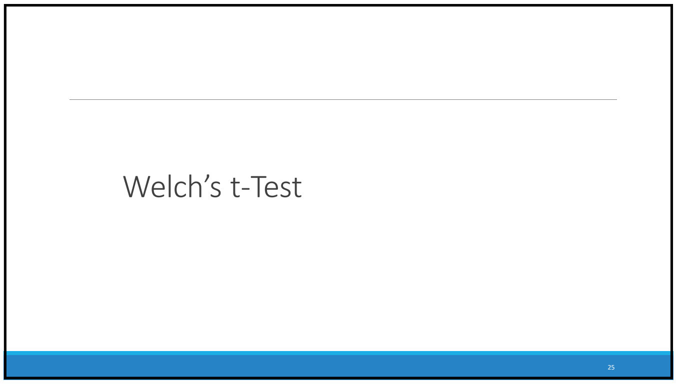
Ho: Distribution of Modified and Conventional Scores are equal
Ha: Distribution of Modified Scores is less than that of Conventional

Critical Value (left sided): -1.645 (alpha = .05)
Test Statistic: z-stat = -3.0183
P-value (left sided)= .0013
Reject Ho

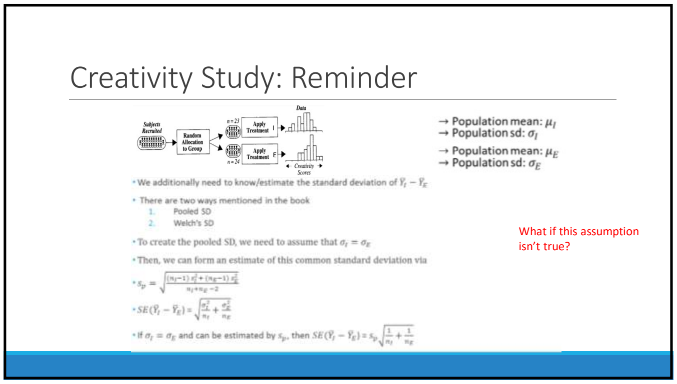


Wilcoxon Two-Sample Test	
Statistic (S)	137.0000
Normal Approximation	
Z	-3.0183
One-Sided Pr < Z	0.0013
Two-Sided Pr > |Z|	0.0025
t Approximation	
One-Sided Pr < Z	0.0027
Two-Sided Pr > |Z|	0.0055
Exact Test	
One-Sided Pr <= S	0.0008
Two-Sided Pr >= |S - Mean|	0.0016
Z includes a continuity correction of 0.5.	

Statistical Conclusion (continued): The data provide convincing evidence that a student could solve the problem more quickly after the "modified" rather than the "conventional" method (one-sided, normal approximation w/ C.C. p-value = 0.0013, from the rank-sum test). A range of plausible values for how much smaller the "modified" distribution is than the "traditional" (treatment effect) is [-158, -59] sec. (95% confidence interval based on a rank-sum test) with a point-estimate of 108.5 sec.



Welch's t-Test



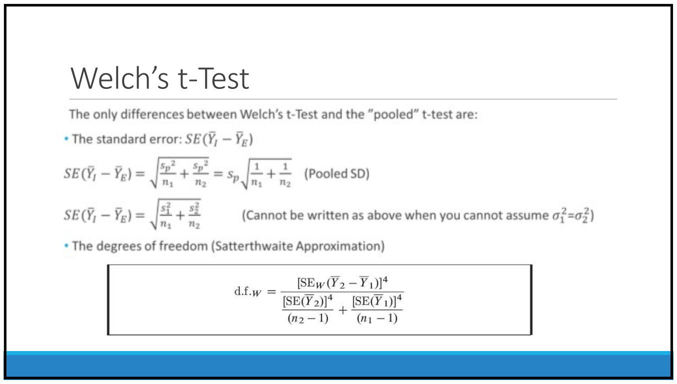
Creativity Study: Reminder

→ Population mean: μ_I
→ Population sd: σ_I

→ Population mean: μ_E
→ Population sd: σ_E

- We additionally need to know/estimate the standard deviation of $\bar{Y}_I - \bar{Y}_E$
- There are two ways mentioned in the book
  1. Pooled SD
  2. Welch's SD
- To create the pooled SD, we need to assume that $\sigma_I = \sigma_E$
- Then, we can form an estimate of this common standard deviation via
- $s_p = \sqrt{\frac{(n_I-1)\,s_I^2 + (n_E-1)\,s_E^2}{n_I+n_E-2}}$
- $SE(\bar{Y}_I - \bar{Y}_E) = \sqrt{\frac{\sigma_I^2}{n_I} + \frac{\sigma_E^2}{n_E}}$
- If $\sigma_I = \sigma_E$ and can be estimated by s_p, then $SE(\bar{Y}_I - \bar{Y}_E) = s_p\sqrt{\frac{1}{n_I} + \frac{1}{n_E}}$

What if this assumption isn't true?



Welch's t-Test

The only differences between Welch's t-Test and the "pooled" t-test are:

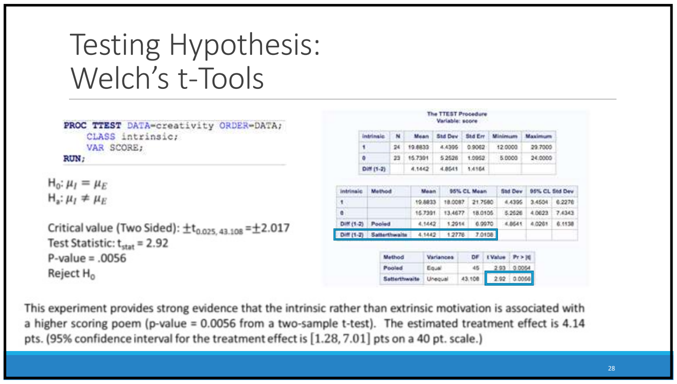
- The standard error: $SE(\bar{Y}_I - \bar{Y}_E)$

$$SE(\bar{Y}_I - \bar{Y}_E) = \sqrt{\frac{s_p{}^2}{n_1} + \frac{s_p{}^2}{n_2}} = s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad \text{(Pooled SD)}$$

$$SE(\bar{Y}_I - \bar{Y}_E) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad \text{(Cannot be written as above when you cannot assume } \sigma_1^2 = \sigma_2^2\text{)}$$

- The degrees of freedom (Satterthwaite Approximation)

$$\boxed{\text{d.f.}_W = \frac{[\text{SE}_W(\overline{Y}_2 - \overline{Y}_1)]^4}{\frac{[\text{SE}(\overline{Y}_2)]^4}{(n_2-1)} + \frac{[\text{SE}(\overline{Y}_1)]^4}{(n_1-1)}}}$$



Testing Hypothesis:
Welch's t-Tools

```
PROC TTEST DATA=creativity ORDER=DATA;
    CLASS intrinsic;
    VAR SCORE;
RUN;
```

$H_0: \mu_I = \mu_E$
$H_a: \mu_I \neq \mu_E$

Critical value (Two Sided): $\pm t_{0.025,\,43.108} = \pm 2.017$
Test Statistic: $t_{stat} = 2.92$
P-value = .0056
Reject H_0

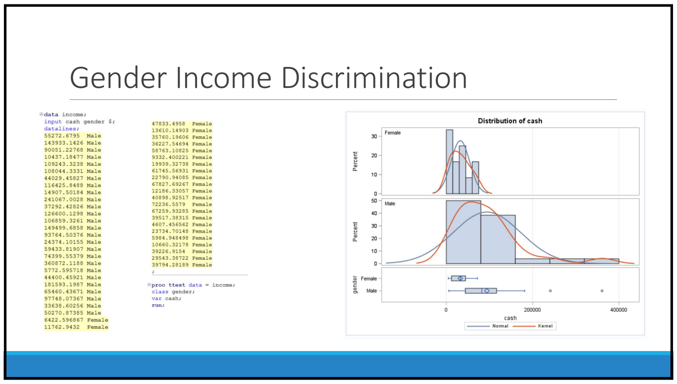
The TTEST Procedure
Variable: score

intrinsic	N	Mean	Std Dev	Std Err	Minimum	Maximum
1	24	19.8833	4.4395	0.9062	12.0000	29.7000
0	23	15.7391	5.2526	1.0952	5.0000	24.0000
Diff (1-2)		4.1442	4.8541	1.4164		

intrinsic	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1		19.8833	18.0087	21.7580	4.4395	3.4504	6.2270
0		15.7391	13.4677	18.0105	5.2526	4.0623	7.4343
Diff (1-2)	Pooled	4.1442	1.2914	6.9970	4.8541	4.0261	6.1138
Diff (1-2)	Satterthwaite	4.1442	1.2776	7.0108			

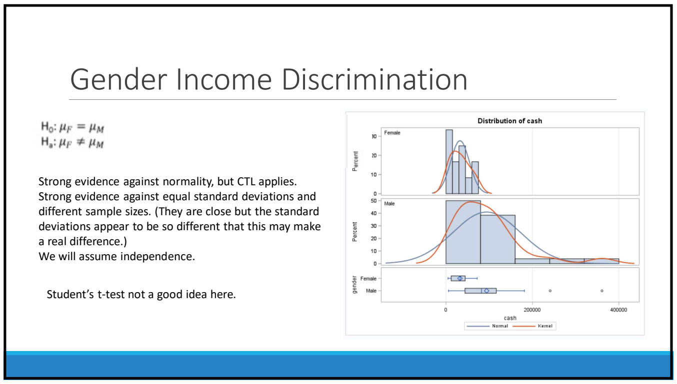
Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	45	2.93	0.0054
Satterthwaite	Unequal	43.108	2.92	0.0056

This experiment provides strong evidence that the intrinsic rather than extrinsic motivation is associated with a higher scoring poem (p-value = 0.0056 from a two-sample t-test). The estimated treatment effect is 4.14 pts. (95% confidence interval for the treatment effect is [1.28, 7.01] pts on a 40 pt. scale.)



Gender Income Discrimination

```
data income;
 input cash gender $;
 datalines;
 55272.6795  Male
 143933.1426 Male
 90051.22768 Male
 10437.18477 Male
 109243.3238 Male
 108044.3331 Male
 44029.45827 Male
 116425.8488 Male
 14907.50184 Male
 241067.0028 Male
 37292.42826 Male
 126600.1298 Male
 106859.3261 Male
 149499.6858 Male
 93764.50376 Male
 24374.10155 Male
 59433.81907 Male
 74399.55379 Male
 360872.1188 Male
 5772.595718 Male
 44400.45921 Male
 181593.1987 Male
 65460.43671 Male
 97748.07367 Male
 33638.60256 Male
 50270.87385 Male
 6422.596867 Female
 11762.9432  Female
 47833.4958  Female
 13610.14903 Female
 35760.19606 Female
 36227.54694 Female
 58763.10825 Female
 9332.400221 Female
 19939.32738 Female
 61745.56931 Female
 22790.94085 Female
 67827.69267 Female
 12186.33057 Female
 40898.92517 Female
 72236.5579  Female
 67259.93285 Female
 39517.38315 Female
 4607.456562 Female
 23734.70148 Female
 5984.948498 Female
 10660.32178 Female
 39226.9154  Female
 29543.38722 Female
 39794.28189 Female
 ;

proc ttest data = income;
 class gender;
 var cash;
 run;
```



Gender Income Discrimination

$H_0: \mu_F = \mu_M$
$H_a: \mu_F \neq \mu_M$

Strong evidence against normality, but CTL applies.
Strong evidence against equal standard deviations and different sample sizes. (They are close but the standard deviations appear to be so different that this may make a real difference.)
We will assume independence.

Student's t-test not a good idea here.

Gender Income Discrimination!

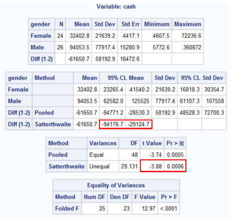
Ho: $\mu_F = \mu_M$
Ha: $\mu_F \neq \mu_M$

Critical value (Two Sided): $\pm t_{0.025,\,29.131} = \pm 2.045$

Test Statistic: $t_{stat} = -3.88$
P-value = .0006
Reject H_0

Conclusion: There is strong evidence to suggest that the mean income of the female group is different from the mean income of the male group (p-value = .0006). A 95% confidence interval for this difference is ($29,124, $94,176) in favor of the males.

That is quite a difference!



Variable: cash

gender	N	Mean	Std Dev	Std Err	Minimum	Maximum
Female	24	32402.8	21639.2	4417.1	4607.5	72236.6
Male	26	94053.5	77917.4	15280.9	5772.6	360872
Diff (1-2)		-61650.7	58192.9	16472.6		

gender	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Female		32402.8	23265.4	41540.2	21639.2	16818.3	30354.7
Male		94053.5	62582.0	125525	77917.4	61107.3	107558
Diff (1-2)	Pooled	-61650.7	-94771.2	-28530.3	58192.9	48528.3	72700.3
Diff (1-2)	Satterthwaite	-61650.7	-94176.7	-29124.7			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	48	-3.74	0.0005
Satterthwaite	Unequal	29.131	-3.88	0.0006

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	25	23	12.97	<.0001

Rank Sum versus Welch's … the Take Away

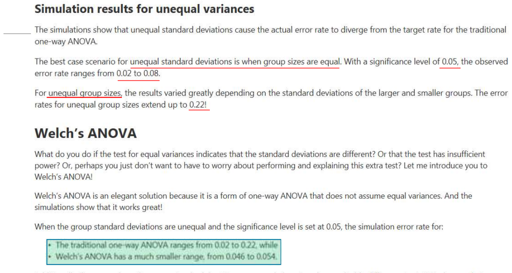
If you wish to make inference on the difference of means and you have the sample size to invoke the CLT, Welch's t-test is preferred by most statisticians, and it is robust to different standard deviations even when the sample size is not equal.

Often, especially in skewed distributions, the median is a better measure of center. For this reason, one may prefer the rank sum test even when Welch's t-test is available.

If you have small sample sizes, you may not be very confident about the normality assumption even if the histograms and q-q plots look okay. For this reason, one may wish to be "conservative" and run the rank sum test and obtain inference on the median.

If there are outliers or censored values, the rank sum test is often the most appropriate as the t-test is not resistant to outliers and has no way of using censored data.

Performance of Welch's t-test



Simulation results for unequal variances

The simulations show that unequal standard deviations cause the actual error rate to diverge from the target rate for the traditional one-way ANOVA.

The best case scenario for unequal standard deviations is when group sizes are equal. With a significance level of 0.05, the observed error rate ranges from 0.02 to 0.08.

For unequal group sizes, the results varied greatly depending on the standard deviations of the larger and smaller groups. The error rates for unequal group sizes extend up to 0.22!

Welch's ANOVA

What do you do if the test for equal variances indicates that the standard deviations are different? Or that the test has insufficient power? Or, perhaps you just don't want to have to worry about performing and explaining this extra test? Let me introduce you to Welch's ANOVA!

Welch's ANOVA is an elegant solution because it is a form of one-way ANOVA that does not assume equal variances. And the simulations show that it works great!

When the group standard deviations are unequal and the significance level is set at 0.05, the simulation error rate for:

- The traditional one-way ANOVA ranges from 0.02 to 0.22, while
- Welch's ANOVA has a much smaller range, from 0.046 to 0.054.

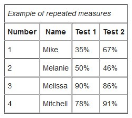
Additionally, for cases where the group standard deviations are equal, there is only a negligible difference in statistical power between these two procedures.

Paired T-Test

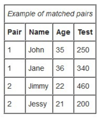
Paired T-Test

Known alternatively as Matched Pairs or Dependent t-Test

Assumptions
- Data are either:
  - From one sample that has been tested twice (example pre- and post-test or repeated measures)
  - From a group of subjects that are thought to be similar and can thus be matched or paired (example from same family, or twins)
- Differences are normally distributed, independent between observations (but dependent from one group to the next).



Example of repeated measures			
Number	Name	Test 1	Test 2
1	Mike	35%	67%
2	Melanie	50%	46%
3	Melissa	90%	86%
4	Mitchell	78%	91%



Example of matched pairs			
Pair	Name	Age	Test
1	John	35	250
1	Jane	36	340
2	Jimmy	22	460
2	Jessy	21	200

A Look at the Variance

- Suppose Y_1 and Y_2 are variables for two groups
- Fact: $Variance(Y_1 - Y_2) = \sigma_1^2 + \sigma_2^2 \; - \; 2\,Covariance(Y_1, Y_2)$
- If the data in each group is **independent** between groups, then $\boldsymbol{Covariance(Y_1, Y_2) = 0}$
- For independent groups, $Variance(Y_1 - Y_2) = \sigma_1^2 + \sigma_2^2$
- If Y_1 and Y_2 are before and after variables for the same subject (or otherwise logically **paired**, **dependent** data), the variables are usually positively correlated ($\boldsymbol{Covariance(Y_1, Y_2) > 0}$)
- For dependent (paired) groups, $Variance(Y_1 - Y_2) = \sigma_1^2 + \sigma_2^2 \; - \; 2\,Covariance(Y_1, Y_2) < \sigma_1^2 + \sigma_2^2$

- If data can be paired, the variance can be reduced.

Example:
Medical Reasoning Test

- The AMA has a diagnostic test for medical reasoning
- On average, people score about 500 points on this test
- We have data from 10 subjects who took the medical reasoning test. These subjects were randomly selected from St. Paul Hospital in Dallas

- **Not fatigued:** is the baseline, taking the test before a shift
- **Fatigued:** is after the treatment; working for 12 operational hours prior to re-taking the test.

Subject #	Not Fatigued	Fatigued
1	567	530
2	512	492
3	509	510
4	593	580
5	588	600
6	491	483
7	520	512
8	588	575
9	529	530
10	508	490

(Lower numbers = worse score)

Example:
Keith’s Medical Reasoning Test

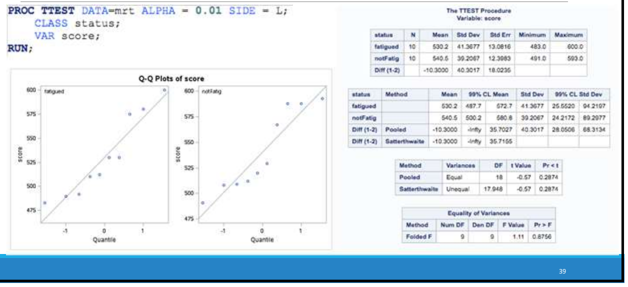
We can try to test whether the DIFFERENCE OF THE MEANS between the fatigued scores and the not fatigued scores is less than zero.

$$H_A: \mu_{fatigued} - \mu_{not\ fatigued} < 0$$

Example:
Medical Reasoning Test

If we did this, we would be wrong! Why?

A fundamental assumption is violated: independence



```
PROC TTEST DATA=mrt ALPHA = 0.01 SIDE = L;
    CLASS status;
    VAR score;
RUN;
```

The TTEST Procedure
Variable: score

status	N	Mean	Std Dev	Std Err	Minimum	Maximum
fatigued	10	530.2	41.3677	13.0816	483.0	600.0
notFatig	10	540.5	39.2067	12.3983	491.0	593.0
Diff (1-2)		-10.3000	40.3017	18.0235		

status	Method	Mean	99% CL Mean		Std Dev	99% CL Std Dev	
fatigued		530.2	487.7	572.7	41.3677	25.5520	94.2197
notFatig		540.5	500.2	580.8	39.2067	24.2172	89.2977
Diff (1-2)	Pooled	-10.3000	-Infty	35.7027	40.3017	28.0506	68.3134
Diff (1-2)	Satterthwaite	-10.3000	-Infty	35.7155			

Method	Variances	DF	t Value	Pr < t
Pooled	Equal	18	-0.57	0.2874
Satterthwaite	Unequal	17.948	-0.57	0.2874

Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	9	9	1.11	0.8756

Assumption Check Failure

We need to account for the dependence between the two groups

Example:
Keith’s Medical Reasoning Test

Instead of testing the DIFFERENCE OF THE MEANS:

$$\xcancel{\begin{aligned} H_o&: \mu_{fatigued} - \mu_{not\ fatigued} = 0 \\ H_A&: \mu_{fatigued} - \mu_{not\ fatigued} < 0 \end{aligned}}$$

We should test the MEAN OF THE DIFFERENCES:

$$H_o: \mu_{fatigued - not\ fatigued} = 0$$
$$H_A: \mu_{fatigued - not\ fatigued} < 0$$

Subject	Fatigued	Not Fatigued	Difference
1	530	567	-37
2	492	512	-20
3	510	509	1
4	580	593	-13
5	600	588	12
6	483	491	-8
7	512	520	-8
8	575	588	-13
9	530	529	1
10	490	508	-18

Paired t-test reduces to a one-sample t-test

Subject	Fatigued	Not Fatigued	Difference (d_i)
1	530	567	-37
2	492	512	-20
3	510	509	1
4	580	593	-13
5	600	588	12
6	483	491	-8
7	512	520	-8
8	575	588	-13
9	530	529	1
10	490	508	-18

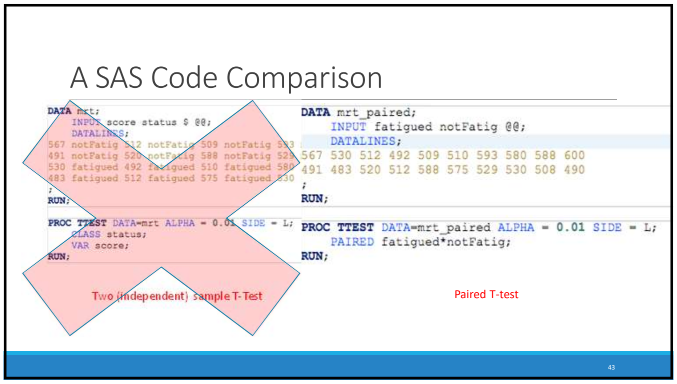
H_0: d = 0
H_a: d < 0

$$\bar{d} = \frac{d_1 + d_2 + \ldots + d_{10}}{10}$$

s_d is the sample std. dev.

$$SE(\bar{d}) = \frac{s_d}{\sqrt{10}}$$

$$T = \frac{\bar{d} - 0}{SE(\bar{d})} = \frac{\bar{d}}{SE(\bar{d})}$$



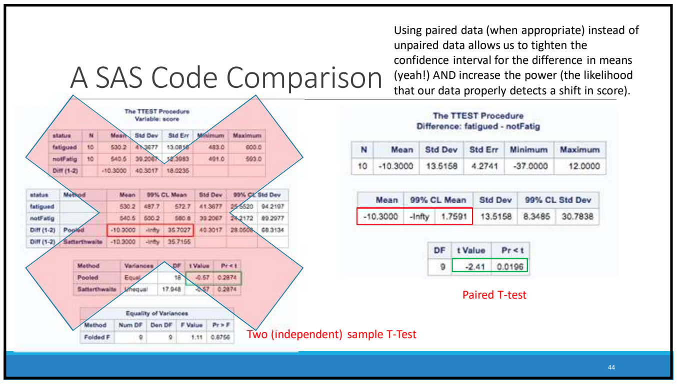
A SAS Code Comparison

```
DATA mrt;
    INPUT score status $ @@;
    DATALINES;
567 notFatig 512 notFatig 509 notFatig 593
491 notFatig 520 notFatig 588 notFatig 529
530 fatigued 492 fatigued 510 fatigued 580
483 fatigued 512 fatigued 575 fatigued 530
;
RUN;

PROC TTEST DATA=mrt ALPHA = 0.01 SIDE = L;
    CLASS status;
    VAR score;
RUN;
```

Two (independent) sample T-Test

```
DATA mrt_paired;
    INPUT fatigued notFatig @@;
    DATALINES;
567 530 512 492 509 510 593 580 588 600
491 483 520 512 588 575 529 530 508 490
;
RUN;

PROC TTEST DATA=mrt_paired ALPHA = 0.01 SIDE = L;
    PAIRED fatigued*notFatig;
RUN;
```

Paired T-test



A SAS Code Comparison

Using paired data (when appropriate) instead of unpaired data allows us to tighten the confidence interval for the difference in means (yeah!) AND increase the power (the likelihood that our data properly detects a shift in score).

The TTEST Procedure
Variable: score

status	N	Mean	Std Dev	Std Err	Minimum	Maximum
fatigued	10	530.2	41.3677	13.0818	483.0	600.0
notFatig	10	540.5	39.2067	12.3983	491.0	593.0
Diff (1-2)		-10.3000	40.3017	18.0235		

status	Method	Mean	99% CL Mean		Std Dev	99% CL Std Dev	
fatigued		530.2	487.7	572.7	41.3677	26.6520	94.2197
notFatig		540.5	500.2	580.8	39.2067	24.2172	89.2977
Diff (1-2)	Pooled	-10.3000	-Infty	35.7027	40.3017	28.0506	68.3134
Diff (1-2)	Satterthwaite	-10.3000	-Infty	35.7155			

Method	Variances	DF	t Value	Pr < t
Pooled	Equal	18	-0.57	0.2874
Satterthwaite	Unequal	17.948	-0.57	0.2874

Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	9	9	1.11	0.8756

Two (independent) sample T-Test

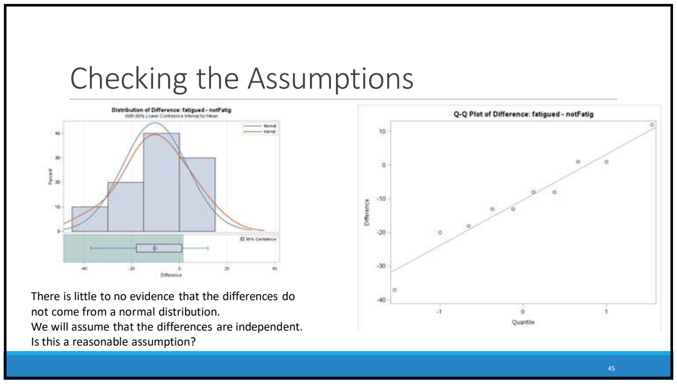
The TTEST Procedure
Difference: fatigued - notFatig

N	Mean	Std Dev	Std Err	Minimum	Maximum
10	-10.3000	13.5158	4.2741	-37.0000	12.0000

Mean	99% CL Mean		Std Dev	99% CL Std Dev	
-10.3000	-Infty	1.7591	13.5158	8.3485	30.7838

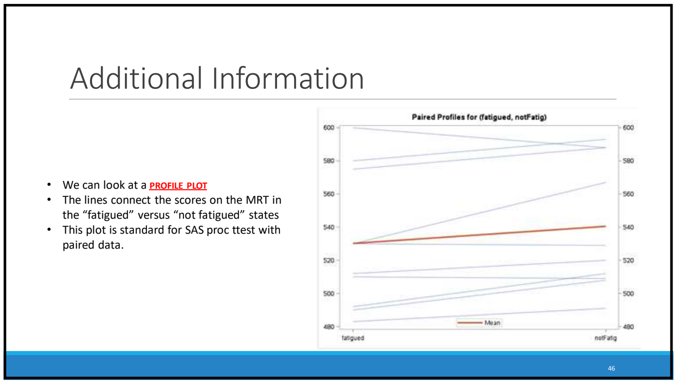
DF	t Value	Pr < t
9	-2.41	0.0196

Paired T-test



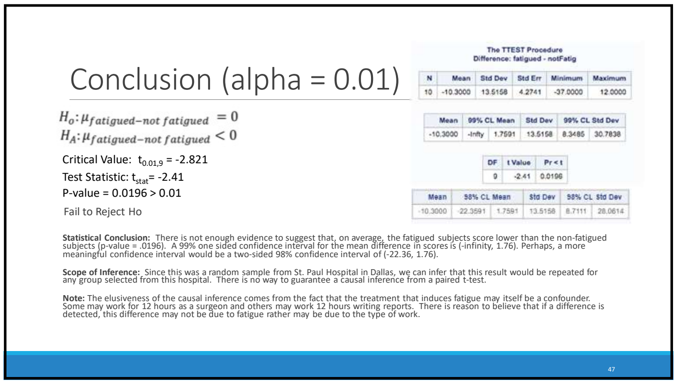
Checking the Assumptions

There is little to no evidence that the differences do not come from a normal distribution.
We will assume that the differences are independent.
Is this a reasonable assumption?



Additional Information

- We can look at a **PROFILE PLOT**
- The lines connect the scores on the MRT in the “fatigued” versus “not fatigued” states
- This plot is standard for SAS proc ttest with paired data.



Conclusion (alpha = 0.01)

$H_o: \mu_{fatigued-not\ fatigued} = 0$
$H_A: \mu_{fatigued-not\ fatigued} < 0$

Critical Value: $t_{0.01,9} = -2.821$
Test Statistic: $t_{stat} = -2.41$
P-value = 0.0196 > 0.01
Fail to Reject Ho

The TTEST Procedure
Difference: fatigued - notFatig

N	Mean	Std Dev	Std Err	Minimum	Maximum
10	-10.3000	13.5158	4.2741	-37.0000	12.0000

Mean	99% CL Mean		Std Dev	99% CL Std Dev	
-10.3000	-Infty	1.7591	13.5158	8.3485	30.7838

DF	t Value	Pr < t
9	-2.41	0.0196

Mean	98% CL Mean		Std Dev	98% CL Std Dev	
-10.3000	-22.3591	1.7591	13.5158	8.7111	28.0614

Statistical Conclusion: There is not enough evidence to suggest that, on average, the fatigued subjects score lower than the non-fatigued subjects (p-value = .0196). A 99% one sided confidence interval for the mean difference in scores is (-infinity, 1.76). Perhaps, a more meaningful confidence interval would be a two-sided 98% confidence interval of (-22.36, 1.76).

Scope of Inference: Since this was a random sample from St. Paul Hospital in Dallas, we can infer that this result would be repeated for any group selected from this hospital. There is no way to guarantee a causal inference from a paired t-test.

Note: The elusiveness of the causal inference comes from the fact that the treatment that induces fatigue may itself be a confounder. Some may work for 12 hours as a surgeon and others may work 12 hours writing reports. There is reason to believe that if a difference is detected, this difference may not be due to fatigue rather may be due to the type of work.

Appendix

Alternatives to the t-Test for Paired Data

Example: Nerve Data

```
/* Sign Test and Signed Rank Test */

data horse;
input horse      site1      site2;
datalines;
6       14.2      16.4
4       17      19
8       37.4      37.6
5       11.2      6.6
7       24.2      14.4
9       35.2      24.4
3       35.2      23.2
1       50.6      38
2       39.2      18.6
;
```

For each of the 9 horses, a veterinary anatomist measured the density of nerve cells at specified sites in the intestine.

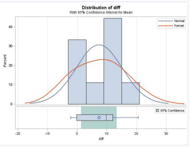
horse	site1	site2
6	14.2	16.4
4	17	19
8	37.4	37.6
5	11.2	6.6
7	24.2	14.4
9	35.2	24.4
3	35.2	23.2
1	50.6	38
2	39.2	18.6

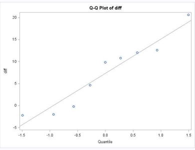
Using the paired t-Test

N	Mean	Std Dev	Std Err	Minimum	Maximum
9	7.3333	7.7929	2.5976	-2.2000	20.6000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
7.3333	1.3431	13.3235	7.7929	5.2638	14.9295

DF	t Value	Pr > |t|
8	2.82	0.0224





The sample size is rather small, hence the normality assumption is somewhat suspect.

The Hypothesis Test

The hypotheses will be in terms of **MEDIANS** instead of means

H_0: The **MEDIAN** difference in nerve cell count between "site 1" and "site 2" is zero

The Alternative Hypotheses:

H_A: The **MEDIAN** difference in nerve cell count between "site 1" and "site 2" is not zero (TWO SIDED)

H_A: The **MEDIAN** difference in nerve cell count between "site 1" and "site 2" is greater than zero (ONE SIDED)

Sign Test: Horse Data

H_A: The **MEDIAN** difference in nerve cell count between "site 1" and "site 2" is > 0

$$z = \frac{K - .5 - {}^{n}/_{2}}{\sqrt{{}^{n}/_{4}}}$$

$$= \frac{6 - .5 - {}^{9}/_{2}}{\sqrt{{}^{9}/_{4}}} = .6666$$

$$P(Z > .6666) = 0.2527$$

(ONE SIDED, CC P-VALUE)

horse	site1	site2	diff	Sign
8	37.4	37.6	-0.2	-
4	17	19	-2	-
6	14.2	16.4	-2.2	-
5	11.2	6.6	4.6	+
7	24.2	14.4	9.8	+
9	35.2	24.4	10.8	+
3	35.2	23.2	12	+
1	50.6	38	12.6	+
2	39.2	18.6	20.6	+

K = 6

Test and Conclusion

H_0: The **MEDIAN** difference in nerve cell count between "site 1" and "site 2" is zero
H_A: The **MEDIAN** difference in nerve cell count between "site 1" and "site 2" is positive.

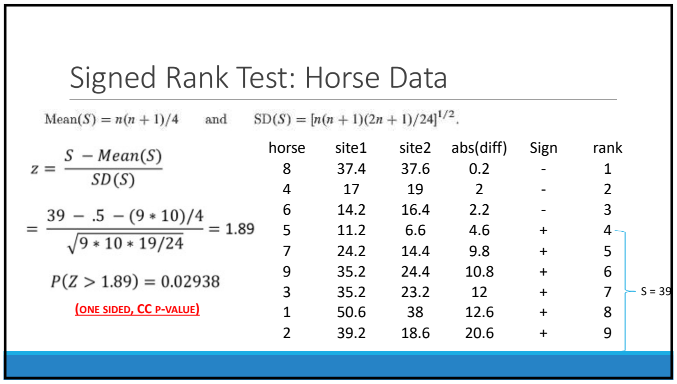
Critical Value (right sided): $z_{0.05}=1.645$

P-value (one sided) = .2527

t statistic: $t_{stat} = 0.666$

Fail to Reject H_0.

Statistical Conclusion: There is not enough evidence that the median nerve density at site 1 is greater than the median nerve density at site 2 (Wilcoxon sign test one-sided p-value of 0.2527).



Signed Rank Test: Horse Data

$\text{Mean}(S) = n(n+1)/4 \quad \text{and} \quad \text{SD}(S) = [n(n+1)(2n+1)/24]^{1/2}.$

$$z = \frac{S - Mean(S)}{SD(S)}$$

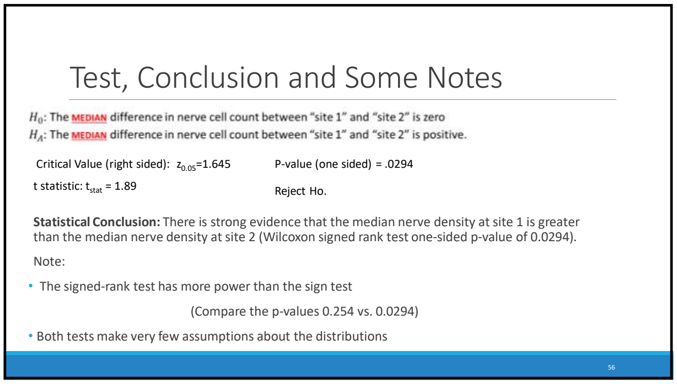
$$= \frac{39 - .5 - (9*10)/4}{\sqrt{9*10*19/24}} = 1.89$$

$$P(Z > 1.89) = 0.02938$$

(ONE SIDED, CC P-VALUE)

horse	site1	site2	abs(diff)	Sign	rank
8	37.4	37.6	0.2	-	1
4	17	19	2	-	2
6	14.2	16.4	2.2	-	3
5	11.2	6.6	4.6	+	4
7	24.2	14.4	9.8	+	5
9	35.2	24.4	10.8	+	6
3	35.2	23.2	12	+	7
1	50.6	38	12.6	+	8
2	39.2	18.6	20.6	+	9

$\Big\}$ S = 39



Test, Conclusion and Some Notes

H_0: The MEDIAN difference in nerve cell count between "site 1" and "site 2" is zero
H_A: The MEDIAN difference in nerve cell count between "site 1" and "site 2" is positive.

Critical Value (right sided): $z_{0.05}=1.645$

P-value (one sided) = .0294

t statistic: $t_{stat} = 1.89$

Reject Ho.

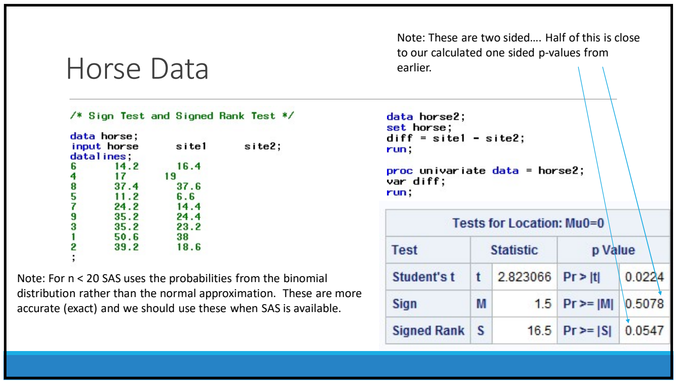
Statistical Conclusion: There is strong evidence that the median nerve density at site 1 is greater than the median nerve density at site 2 (Wilcoxon signed rank test one-sided p-value of 0.0294).

Note:

- The signed-rank test has more power than the sign test

(Compare the p-values 0.254 vs. 0.0294)

- Both tests make very few assumptions about the distributions



Horse Data

Note: These are two sided…. Half of this is close to our calculated one sided p-values from earlier.

```
/* Sign Test and Signed Rank Test */

data horse;
input horse      site1      site2;
datalines;
6       14.2      16.4
4       17       19
8       37.4      37.6
5       11.2      6.6
7       24.2      14.4
9       35.2      24.4
3       35.2      23.2
1       50.6      38
2       39.2      18.6
;
```

```
data horse2;
set horse;
diff = site1 - site2;
run;

proc univariate data = horse2;
var diff;
run;
```

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	2.823066	Pr > |t|	0.0224
Sign	M	1.5	Pr >= |M|	0.5078
Signed Rank	S	16.5	Pr >= |S|	0.0547

Note: For n < 20 SAS uses the probabilities from the binomial distribution rather than the normal approximation. These are more accurate (exact) and we should use these when SAS is available.

Part V

ANOVA

Chapter 21

Problem 1: Plots and Logged Data

We begin our work looking at raw and transformed data.

21.1 Plots and Transformations

Raw Data Analysis

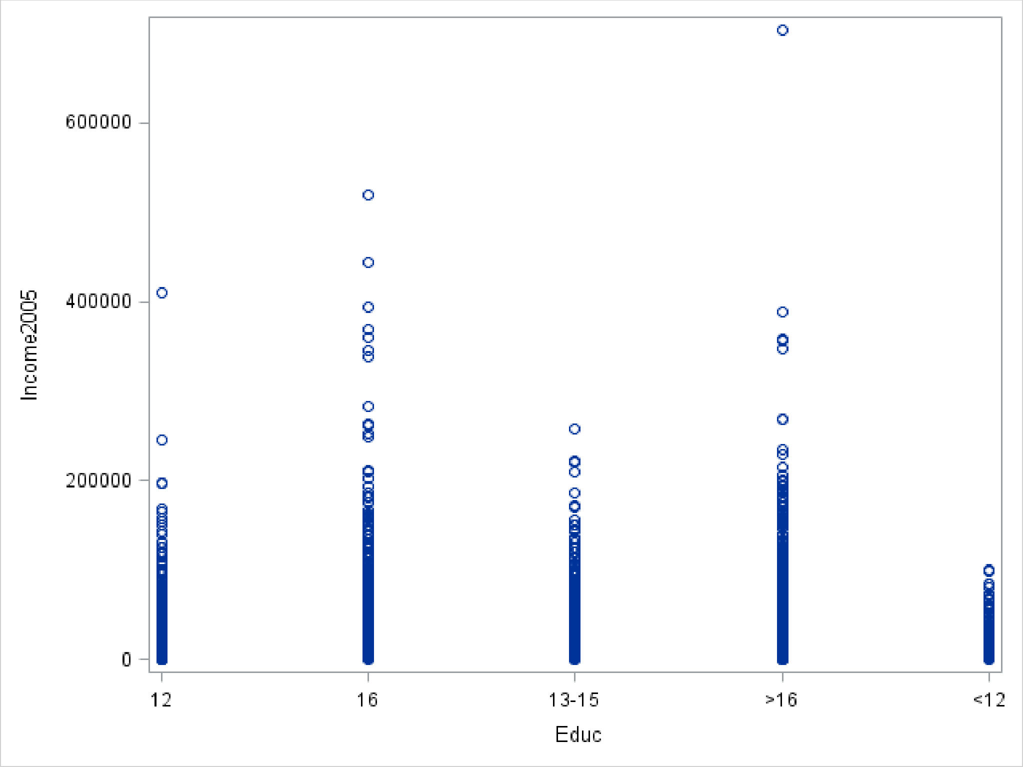
First, we will look at the raw data. To check if the raw data fits the assumptions, we will first look at a scatter plot. The scatter plot of the raw data was produced by the following bit of SAS code:

Code 21.1. Scatterplot of Raw Data Using SAS

```
proc sgplot data=EduData;
scatter x=educ y=Income2005;
run;
```

This results in the following plot21.1:

Figure 21.1.1. Scatter Plot of the Raw Data



Looking at Figure 21.1.1, we see that the raw data is very heavy in between 0 and 20,000 for all categories, but some groups spread further and wider than others, which suggests the variances may not be equal. The heaviness of the lower end of each group may also suggest a lack of normality. We will examine this further with some Box plots. These were produced using the following chunk of SAS code: This results in the following plot:

Code 21.2. Boxplot of Raw Data Using SAS

```
proc sgplot data=EduData;
vbox Income2005 / category=educ
dataskin=matte
;
xaxis display=(noline noticks);
yaxis display=(noline noticks) grid;
run;
```

Figure 21.1.2. Box Plot of the Raw Data

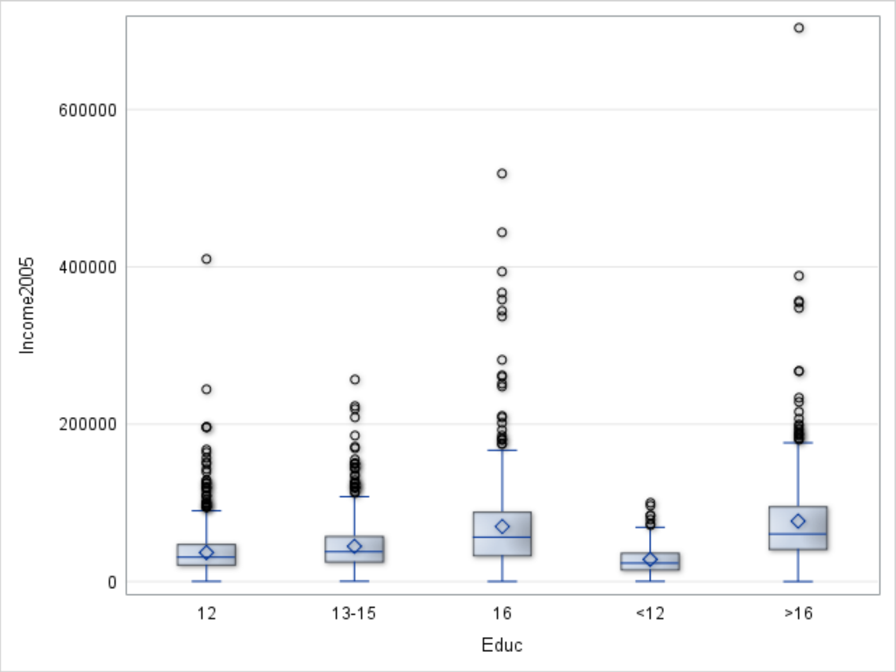


Figure 21.1.2 tells us a lot about our data. We see from the size and shape of the boxes that the variances of our data are by no means homogeneous. Note that there are a lot of outliers while the distribution is heavily weighted towards the bottom, this suggests our data may have departed from normality. We will examine this phenomenaa further using histograms.

To produce histograms of the raw data, the following SAS code was used: This results in the following plot:

Code 21.3. Histogram of Raw Data Using SAS

```
proc sgpanel data=EduData;
panelby educ / rows=5 layout=rowlattice;
histogram Income2005;
run;
```

Figure 21.1.3. Histogram of the Raw Data

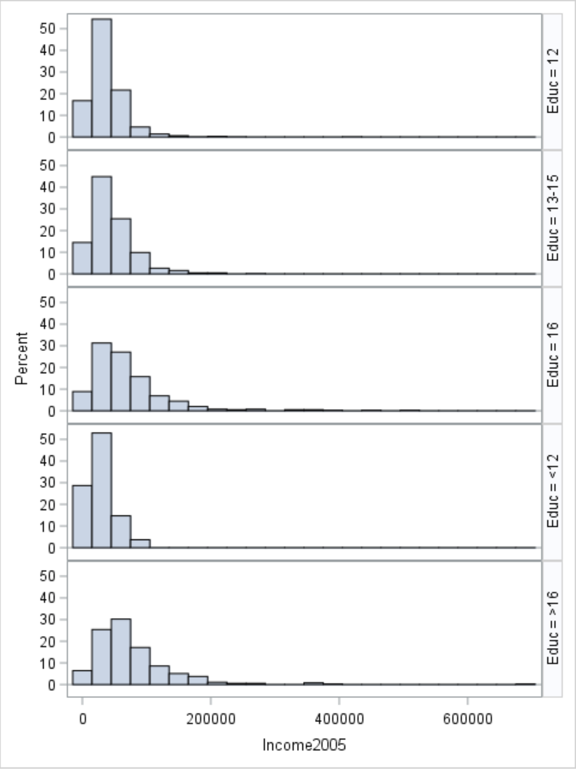


Figure 21.1.3 confirms our suspicions, the variances of the data are likely unequal, but more importantly, the data is clearly skewed to the right. We will confirm this using Q-Q plots.

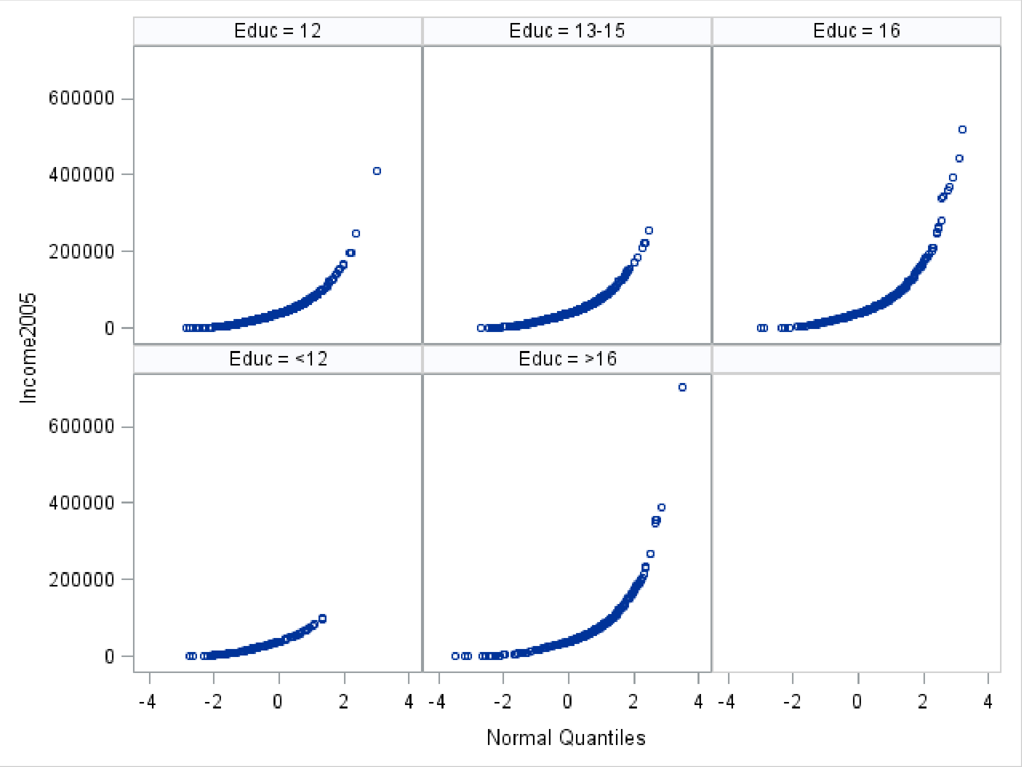
To produce Q-Q plots of the raw data, the following SAS code was used:

Code 21.4. Q-Q of Raw Data Using SAS

```
/* Normal = blom produces normal quantiles from the data */
/* To find out more, look at the SAS documentation!*/
proc rank data=EduData normal=blom out=EduQuant;
var Income2005;
/* Here we produce the normal quantiles!*/
ranks Edu_Quant;
run;
proc sgpanel data=EduQuant;
panelby educ;
scatter x=Edu_Quant y=Income2005 ;
colaxis label="Normal Quantiles";
run;
```

This results in the following plot:

Figure 21.1.4. Q-Q Plot of the Raw Data



The Q-Q plots in Figure 21.1.4 tell us what we already know: The raw data is not normal, and does not have equal variances. The ANOVA test is not super robust to highly skewed, long tailed data, and it relies entirely on equal variances, so we absolutely cannot use the raw data

Transformed Data Analysis

Now we will perform a log transformation on the data and see if that helps it meet our assumptions better. To do a log transformation, we will employ the following SAS code: We will begin our analysis of the transformed data

Code 21.5. Logging of Raw Data Using SAS

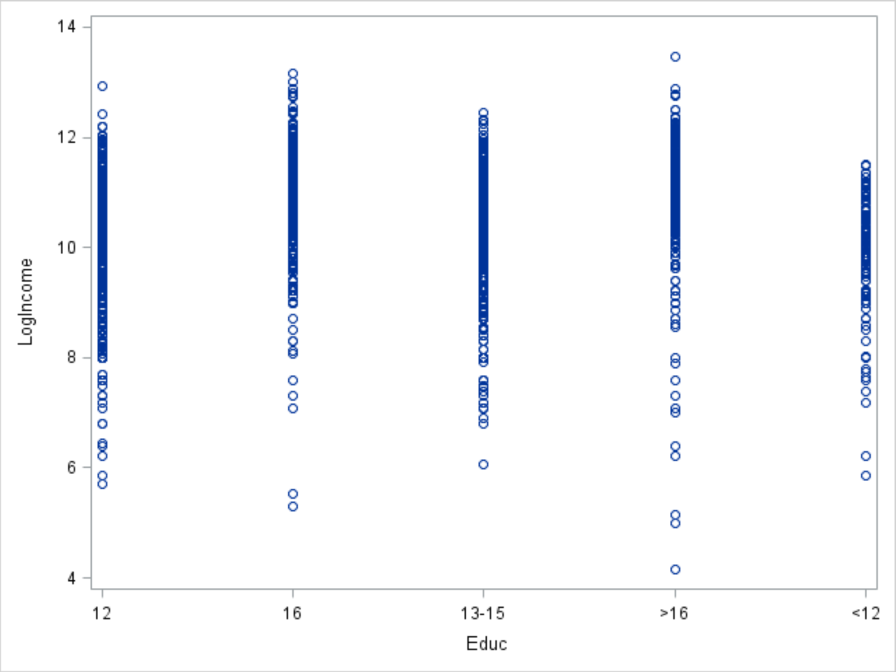
```
data LogEduData;
set EduData;
LogIncome=log(Income2005);
run;
```

with a scatter plot, produced with the following SAS code: This results in the following plot:

Code 21.6. Scatterplot of Logged Data Using SAS

```
proc sgplot data=LogEduData;
scatter x=educ y=LogIncome;
run;
```

Figure 21.1.5. Scatter Plot of the Log-Transformed Data



As we can see in Figure 21.1.5, the groups have a much more similar size, suggesting similar variances, and the heavy part of the scatter plot is closer to the center, in between the outliers, which tells us the log transformation may have done a good deal towards normalizing our data. We can examine this further using Box plots.

To produce Box plots of the transformed data, the following SAS code was used: This gives us the following plot:

Code 21.7. Boxplot of Logged Data Using SAS

```
proc sgplot data=LogEduData;
vbox LogIncome / category=educ
dataskin=matte
;
xaxis display=(noline noticks);
yaxis display=(noline noticks ) grid;
run;
```

Figure 21.1.6. Box Plot of the Log-Transformed Data

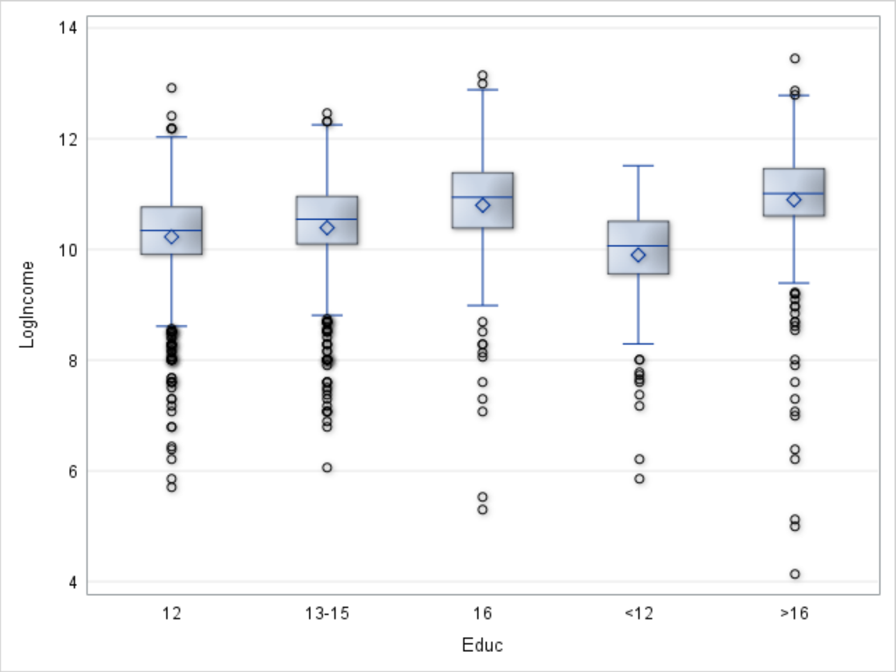
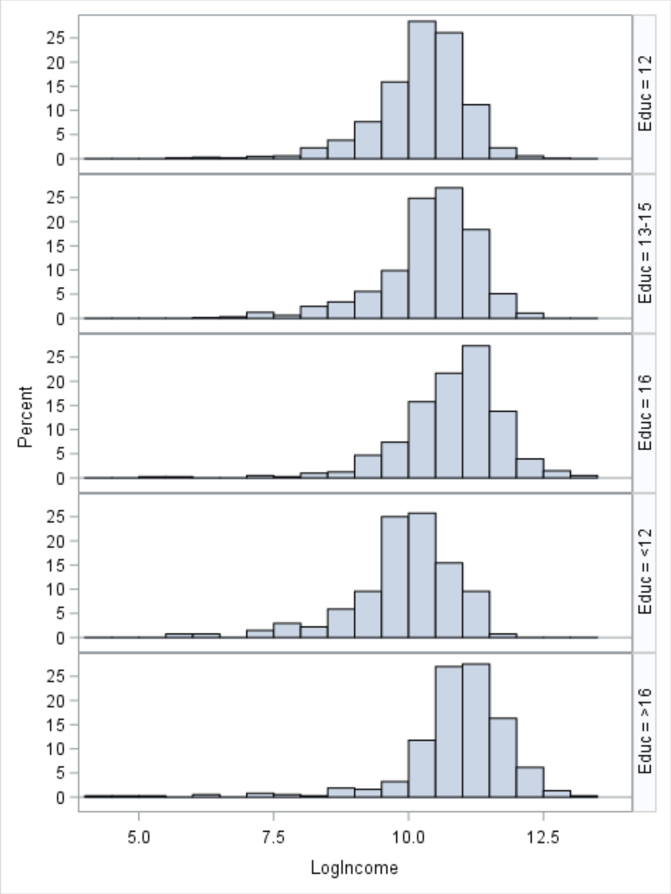


Figure 21.1.6 gives us some useful information about our data. We see the boxes and whiskers are of similar size, which tells us the variances are likely homogeneous. Furthermore, the medians and means are near each other, and the boxes are near the center of the distribution, which suggests that the data may be normal. We will examine these two phenomena further with histograms. To produce histograms of the log-transformed data, the following SAS code was used: This results in the following plot:

Code 21.8. Histogram of Logged Data Using SAS

```
proc sgpanel data=LogEduData;
panelby educ / rows=5 layout=rowlattice;
histogram LogIncome;
run;
```

Figure 21.1.7. Histogram of the Log-Transformed Data



From the spread of the histograms in Figure 21.1.7, we see two things. First, the similar width of the histograms confirms that variances are roughly equal. Second, the shape of the histograms, and their location near the center suggests that the data is very nearly normal. We will further examine the normality of the data using Q-Q plots.

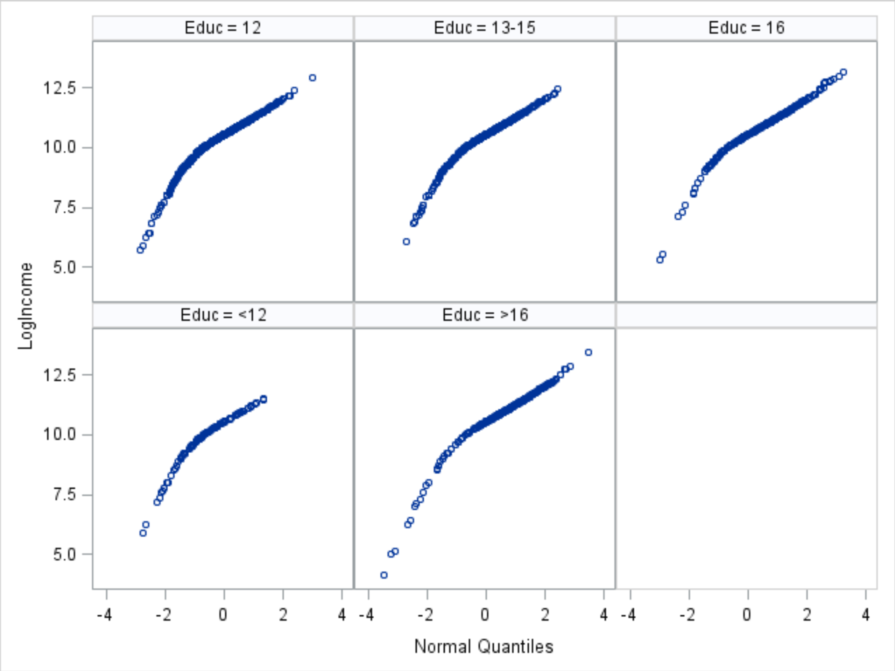
To produce the Q-Q plots of the transformed data, the following SAS code was used: This results in the following

Code 21.9. Q-Q of Logged Data Using SAS

```
proc rank data=LogEduData normal=blom out= LogEduQuant;
var LogIncome;
ranks LogEduQuant;
run;
proc sgpanel data=LogEduQuant;
panelby educ;
scatter x=LogEduQuant y=LogIncome ;
colaxis label="Normal Quantiles";
run;
```

plot:

Figure 21.1.8. Q-Q Plot of the Log-Transformed Data



Examining Figure 21.1.8, we see a confirmation of our beliefs: The log-transformed data, when plotted against normal quantiles, is fairly normal. This means, with the log transformed data, we can reasonably assume normality and homogeneity of variances.

21.2 Complete Analysis

We will now perform a complete analysis of our data, using Pure ANOVA.

Problem Statement

We would like to determine whether or not at least one of the five population distributions (corresponding to different years of education) is different from the rest.

Assumptions

As seen in Section 21.1, the raw data does not meet the assumption of normality nor of homogeneity of variance. However, in Section 21.1, we proved that after a log transformation, the data does meet both of these assumptions. The ANOVA test is fairly robust to the slight departure from normality presented by the log transformed data, and the variances are equal. The data is clearly independent, so that assumption is met. Therefore, all assumptions of ANOVA are met by the log transformed data.

Hypothesis Definition

In this problem, our Null (Reduced Model) Hypothesis, H_0, is that all the groups have the same distribution and our Alternative (Full Model) Hypothesis, H_1 is that the distributions are different. Mathematically, that is written as:

$$H_0 : median_{grand} \quad median_{grand} \quad median_{grand} \quad median_{grand} \quad median_{grand} \tag{21.2.1}$$

$$H_1 : median_{<12} \quad median_{12} \quad median_{13-15} \quad median_{16} \quad median_{>16} \tag{21.2.2}$$

We will consider our confidence level, α to be 0.05

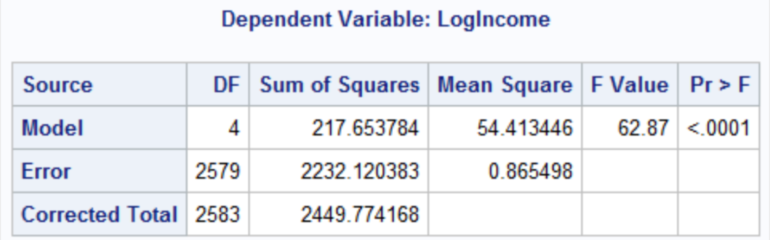
F Statistic

To conduct this hypothesis test, the following SAS code was used: This results in the following ANOVA Output:

Code 21.10. ANOVA Test Using SAS

```
proc glm data = LogEduData;
class educ;
model LogIncome = educ;
run;
```

Figure 21.2.1. ANOVA Table



Dependent Variable: LogIncome

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	217.653784	54.413446	62.87	<.0001
Error	2579	2232.120383	0.865498		
Corrected Total	2583	2449.774168			

Figure 21.2.1 tells us what our F statistic is. We see that

$$F = 62.87 \tag{21.2.3}$$

P-value

Figure 21.2.1 also tells us our p-value. In this case,

$$p < .0001 \tag{21.2.4}$$

Hypothesis Assessment

In this scenario, we have that $p < .0001 < \alpha = .05$ and therefore we reject the null hypothesis.

Conclusion

There is substantial evidence ($p < 0.0001$) that at least one of the distributions is different from the others. To further examine this, we will see if the distribution varies within similar levels of schooling. We will compare <12 and 12 years of school, 12 and 13-15 years of school, 13-15 and 16 years of school, and 16 and >16 years of school. To do this, we will compare medians, using the following SAS code: This results in the following Table:

Code 21.11. Comparison of distributions using SAS

```
proc sort data=LogEduData;
by educ;
run;
proc means data = LogEduData  median order=data;
by educ;
var LogIncome;
run;
```

Table 21.1. Comparison of Logged Means

Education	μ
<12	9.9
12	10.22
13-15	10.39
16	10.79
>16	10.89

From Table 21.1, we can calculate the differences of the means for our log transformed groups, and see how much the distributions differ, shown in the following table:

Table 21.2. Comparison of Distributions

Pair	Difference	Multiplicative Effect ($e^{\mu_1-\mu_2}$)	% Increase
<12 and 12	0.32	1.38	38
12 and 13-15	0.17	1.19	19
13-15 and 16	.4	1.49	49
16 and >16	.1	1.11	11

Table 21.2 shows us how many times greater the distribution of the income of the larger education in each pair is than the lower education level.

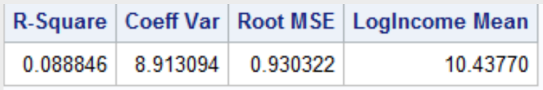
Scope of Inference

As this was a random sample, we can make inferences about the population, however, we cannot make causal inferences, as this was not a randomized experiment. That means, we can say that in general, people with X years of education make Y many times as people with Z years of education, but we cannot say it is due to the education itself.

21.3 Extra Values

The extra values were produced with the same code as in Section 28.1. They can be found in Figure 21.2.1, and in the figure below:

Figure 21.3.1. Extra Values



R-Square	Coeff Var	Root MSE	LogIncome Mean
0.088846	8.913094	0.930322	10.43770

Value of R^2

Figure 21.3.1 tells us R^2 is 0.0888

Mean Square Error and Degrees of Freedom

The Mean Square Error, shown in Figure 21.2.1, is 2232.12, with 2579 degrees of freedom

ANOVA in R!

Here is the R code and output to do ANOVA in R on the log transformed data:

Code 21.12. ANOVA in R

```
##################### Anova in R ######################
edudata <- read.csv(file='data/ex0525.csv', header=TRUE, sep = ",")
edudata$logincome <- log(edudata$Income2005)

# http://www.sthda.com/english/wiki/one-way-anova-test-in-r
anovatest <- aov(logincome~Educ,data =edudata)
summary(anovatest)

######################### Results #####################

Df Sum Sq Mean Sq F value Pr(>F)
Educ           4   217.7   54.41   62.87 <2e-16 ***
Residuals   2579 2232.1    0.87
```

Chapter 22

Problem 2: Build Your Own Anova!

In this section we will be building an ANOVA table to determine whether or not the distribution of income of people with > 16 years is different than the distribution of income of people with exactly 16 years of education. To build this ANOVA table, we need two preliminary ANOVA analyses. First, is the ANOVA analysis seen in Section 21.2. This has the null hypothesis that all the distributions are the same, and the alternative hypothesis that the distributions differ. Next, we build a second ANOVA table, which will have a null hypothesis that all the distributions are the same, and an alternative hypothesis that all the distributions are different, except the group with 16 years and the group with >16 years are still the same. This is done by grouping the two into one group, with the following SAS code: Next, to compute important parameters, an ANOVA test is conducted on the grouped, logged, data, with

Code 22.1. Regrouping data using SAS

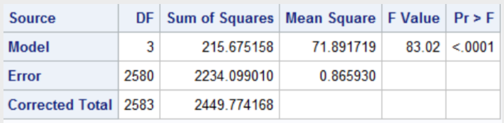
```
data EduGroupData;
set LogEduData;
Others = educ;
if educ eq "16"  educ = ">16" then Others="a";run;
```

the following bit of code: This results in the following intermediate ANOVA table:

Code 22.2. Secondary ANOVA using SAS

```
proc glm data = EduGroupData;
class Others;
model LogIncome = Others;
run;
```

Figure 22.0.1. Grouped ANOVA Table



Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	215.675158	71.891719	83.02	<.0001
Error	2580	2234.099010	0.865930		
Corrected Total	2583	2449.774168			

22.1 Building the Extra Sum of Squares Anova Table

Using the data from 22.0.1 and the data from 21.2.1, we can make our own ANOVA table, which has a null hypothesis that all the distributions different and (except 16 and >16, which are the same), and an alternative hypothesis that all the distributions are different. Since both hypotheses have the same prediction about the data for <12, 12, and 13-15, the null hypothesis of our custom-made ANOVA table is that 16 and >16 have the same distribution, and the alternative is that they have different distributions. We will now construct our new, extra sum of squares ANOVA table.

First, for our full model (the ”Error” row in the ANOVA table), we will use the full model (alternative hypothesis, or the ”Error” row), from Figure 21.2.1. This represents our alternative hypothesis, where the distribution of 16 and >16 are different. Next, we will construct our reduced model (The ”Total” row in the ANOVA table) using the full model (alternative hypothesis, or the ”Error”) from 22.0.1. This represents our null hypothesis, where 16 and >16 have the same distribution. To generate our Model, or Extra Sum of Squares, which will allow us to find our F statistic and p value, we need to take a couple of steps. To determine the number of degrees of freedom of our model, we subtract the number of degrees of freedom from the Error row from the number of degrees of freedom of the Total row. To calculate the extra sum of squares, we subtract the residual sum of squares of the full model (error) from the residual sum of squares of the reduced model (total). Then, to find the mean square, we divide the extra sum of squares by the number of degrees of freedom in our model. Our F statistic is then produced by normalizing the Extra Sum of Squares, dividing it by the Mean Square Error (in the Error row). To get a p value from the F statistic, we examine an F distribution with degrees of freedom $= \frac{df_{model}}{df_{full}}$. The results of these computations are displayed in the following table:

Table 22.1. Homemade ANOVA Table

Source	DF	Sum of Squares	Mean Square	F Value	Pr>F
Model (Extra SS)	1	1.98	1.98	2.3	0.129
Error (Full)	2579	2232.12	.86		
Total (Reduced)	2580	2234.1			

22.2 Complete Analysis

Problem Statement

We would like to determine whether or not people with a college degree or a graduate degree have different distributions of incomes.

Assumptions

There are three assumptions of ANOVA: normality, homogeneity of variance, and independence. We have shown, in Section 21.1 that while the raw data does not meet the first two assumptions, the log transformed data does. Both the transformed and raw data meet the assumption of independence. We will proceed with our ANOVA test.

Hypothesis Definition

Our null hypothesis states that the distribution of the >16 and 16 groups is the same, and our alternative hypothesis states that the distribution of the >16 and 16 groups is different. We proved this in Section 22.1, and this is written mathematically as:

$$H_0 : median_{<12} \quad median_{12} \quad median_{13-15} \quad median_{16,>16} \quad median_{16,>16} \tag{22.2.1}$$

$$H_1 : median_{<12} \quad median_{12} \quad median_{13-15} \quad median_{16} \quad median_{>16} \tag{22.2.2}$$

OR:

$$H_0 : median_{16} = median_{>16} \tag{22.2.3}$$

$$H_1 : median_{16} \neq median_{>16} \tag{22.2.4}$$

We will consider our confidence level, α to be 0.05

F Statistic

The F statistic is calculated with the following equation:

$$F = \frac{\left(\frac{SS_{extra}}{DF_{extra}}\right)}{\sigma^{\hat{2}}_{full}} = \frac{\left(\frac{SS_{extra}}{DF_{extra}}\right)}{MSE} \tag{22.2.5}$$

The results of this calculation can be seen in Table 22.1, we have that F $=$ 2.3 This is a small F statistic, which is likely indicative of weak evidence.

P-value

The P value is calculated using F, the Extra degrees of freedom, and the Full (Error) degrees of freedom. Using the values calculated in Table 22.1, we have that p $=$ 0.129

Hypothesis Assessment

At a confidence level $\alpha = 0.05$, we have that p $=$ .0129 $> \alpha =$ .05. Therefore, we cannot reject the null hypothesis.

Conclusion

There is not enough evidence to suggest that the distribution of income of people with a college only (16 years) is different from the distribution of income of people with a postgraduate education (>16 years).

Scope of Inference

It is not necessary to write a scope of inference as we did not reject the null hypothesis, however this is a random sample, so we can make inferences about the population as whole, but we cannot infer causality, as this was not a random experiment.

22.3 Degrees of Freedom and Comparison to T-Test

This test had 2579 degrees of freedom (as seen in Table 22.1). This is a lot more than than the t test, which is a lot more than the number of degrees of freedom in the t test. Therefore, this ANOVA test has more power than the t test!.

Chapter 23

Problem 3: Nonhomogeneous Standard Deviations

23.1 Complete Analysis

Problem Statement

We would like to determine whether or not at least one of the five population distributions (corresponding to different years of education) is different from the rest.

Assumptions

As seen in Section 21.1, the raw data does not meet the assumption of normality nor of homogeneity of variance. However, in Section 21.1, we proved that after a log transformation, the data is at least normal. The ANOVA test is fairly robust to the slight departure from normality presented by the log transformed data, so we can safely assume normality. However, we cannot assume homogeneity variances. Therefore, pure ANOVA is not appropriate. Since the data is to some extent normal, we should try and use a parametric test, as they have more power in general than their nonparametric analogs. Therefore, the Kruskal-Wallis test is not the most appropriate test. We will instad use Welch's ANOVA Test, which assumes normality but does not assume homogeneity of variance, on the log transformed data. We can assume the data is independent.

Hypothesis Definition

In this problem, our Null (Reduced Model) Hypothesis, H_0, is that all the groups have the same distribution and our Alternative (Full Model) Hypothesis, H_1 is that the distributions are different. Mathematically, that is written as:

$$H_0 : median_{grand} \quad median_{grand} \quad median_{grand} \quad median_{grand} \quad median_{grand} \tag{23.1.1}$$

$$H_1 : median_{<12} \quad median_{12} \quad median_{13-15} \quad median_{16} \quad median_{>16} \tag{23.1.2}$$

We will consider our confidence level, α to be 0.05

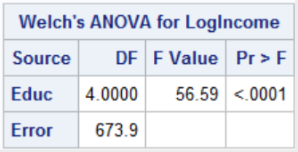
F Statistic

To conduct this hypothesis test, the following SAS code was used: This results in the following table:

Code 23.1. Welch's ANOVA in SAS

```
proc glm data = LogEduData;
class educ;
model LogIncome = educ;
means educ / welch;
run;
```

Figure 23.1.1. Welch's ANOVA Table



Welch's ANOVA for LogIncome			
Source	**DF**	**F Value**	**Pr > F**
Educ	4.0000	56.59	<.0001
Error	673.9		

From Figure 23.1.1, we have that $F = 56.59$. This is a pretty large F statistic, which means that we probably have some good evidence in favor of the alternative hypothesis.

P-value

Figure 23.1.1 Also tells us that the p-value associated with the F statistic, which is given as $p < 0.0001$.

Hypothesis Assessment

We have that $p < 0.0001 < \alpha = .05$ and therefore we Reject the null hypothesis

Conclusion

There is convincing evidence ($p < 0.0001$) that at least one of the distributions is different from the others.

Scope of Inference

As this was a random sample, we can make inferences about the population, however, we cannot make causal inferences, as this was not a randomized experiment. That means, we can say that in general, people with X years of education make Y many times as people with Z years of education, but we cannot say it is due to the education itself.

Chapter 24

unit 5 lecture slides

More slides

UNIT 5: Chapter 5

ANOVA

ANOVA

1. Make a Scatterplot of the data in the table below. “Level” is the Explanatory Variable (X=1, 2, or 3).

	Level i=1	Level i=2	Level i=3
Y_1|X=i	3	10	20
Y_2|X=i	5	12	22
Y_3|X=i	7	14	24
$\hat{\mu}_{Y\|X=i}$			

2. Find the Grand Mean … this is the mean of all the Ys together … regardless of Level.

$\hat{\mu} = \bar{\bar{x}} =$

3. Find the Conditional (Level) Means … this is the mean of the Ys per Level. Example: The Conditional mean $\hat{\mu}(Y|X=1) = 5$.

ANOVA

1. Make a Scatterplot of the data in the table below. “Level” is the Explanatory Variable (X=1, 2, or 3).

	Level i=1	Level i=2	Level i=3
Y_1|X=i	3	10	20
Y_2|X=i	5	12	22
Y_3|X=i	7	14	24
$\hat{\mu}_{Y\|X=i}$	5	12	22

2. Find the Grand Mean … this is the mean of the sample means. If the sample size is the same in each group, then this is the mean of all the Ys together … regardless of Level.

$\hat{\mu} = \bar{\bar{x}} = 13$

3. Find the Conditional (Level) Means … this is the mean of the Ys per Level. Example: The Conditional mean $\hat{\mu}(Y|X=1) = 5$.

Pure ANOVA

	Level i=1	Level i=2	Level i=3
Y_1|X=i	3	10	20
Y_2|X=i	5	12	22
Y_3|X=i	7	14	24
$\hat{\mu}_{Y\|X=i}$	5	12	22

4. Now we need to find the Sum of the Squared Residuals for the **Equal** Means Model.

$((Y_i|X) - \hat{\mu})^2$

$\hat{\mu} = \bar{\bar{x}} = 13$

	Level i=1	Level i=2	Level i=3
$((Y_1\|X=i) - \hat{\mu})^2$			
$((Y_2\|X=i) - \hat{\mu})^2$			
$((Y_3\|X=i) - \hat{\mu})^2$			
*Total Sum of Squared Residuals for **Equal** Means Model:*			

5. Now we need to find the Sum of the Squared Residuals for the **Separate** Means Model, where $\hat{\mu}_i = \hat{\mu}(Y|X=i)$.

$((Y_i|X=i) - \hat{\mu}_i)^2$

	Level i=1	Level i=2	Level i=3
$((Y_1\|X=i) - \hat{\mu}_i)^2$			
$((Y_2\|X=i) - \hat{\mu}_i)^2$			
$((Y_3\|X=i) - \hat{\mu}_i)^2$			
*Total Sum of Squared Residuals for **Separate** Means Model:*			

6. Compare the Total Sum of Squares for each model. Which do you think “fits” better?

Pure ANOVA

	Level i=1	Level i=2	Level i=3
Y_1|X=i	3	10	20
Y_2|X=i	5	12	22
Y_3|X=i	7	14	24
$\hat{\mu}_{Y\|X=i}$	5	12	22

4. Now we need to find the Sum of the Squared Residuals for the **Equal** Means Model.

$((Y_i|X) - \hat{\mu})^2$

$\hat{\mu} = \bar{\bar{x}} = 13$

	Level i=1	Level i=2	Level i=3
$((Y_1\|X=i) - \hat{\mu})^2$	$(3-13)^2 = 100$	$(10-13)^2 = 9$	49
$((Y_2\|X=i) - \hat{\mu})^2$	$(5-13)^2 = 64$	1	81
$((Y_3\|X=i) - \hat{\mu})^2$	36	1	121
*Total Sum of Squared Residuals for **Equal** Means Model:* 462			

5. Now we need to find the Sum of the Squared Residuals for the **Separate** Means Model, where $\hat{\mu}_i = \hat{\mu}(Y|X=i)$.

$((Y_i|X=i) - \hat{\mu}_i)^2$

	Level i=1	Level i=2	Level i=3
$((Y_1\|X=i) - \hat{\mu}_i)^2$			
$((Y_2\|X=i) - \hat{\mu}_i)^2$			
$((Y_3\|X=i) - \hat{\mu}_i)^2$			
*Total Sum of Squared Residuals for **Separate** Means Model:*			

6. Compare the Total Sum of Squares for each model. Which do you think “fits” better?

Pure ANOVA

	Level i=1	Level i=2	Level i=3
Y_1|X=i	3	10	20
Y_2|X=i	5	12	22
Y_3|X=i	7	14	24
$\hat{\mu}_{Y\|X=i}$	5	12	22

4. Now we need to find the Sum of the Squared Residuals for the **Equal** Means Model.

$((Y_i|X) - \hat{\mu})^2$

$\hat{\mu} = \bar{\bar{x}} = 13$

	Level i=1	Level i=2	Level i=3
$((Y_1\|X=i) - \hat{\mu})^2$	$(3-13)^2 = 100$	9	49
$((Y_2\|X=i) - \hat{\mu})^2$	64	1	81
$((Y_3\|X=i) - \hat{\mu})^2$	36	1	121
*Total Sum of Squared Residuals for **Equal** Means Model:* 462			

5. Now we need to find the Sum of the Squared Residuals for the **Separate** Means Model, where $\hat{\mu}_i = \hat{\mu}(Y|X=i)$.

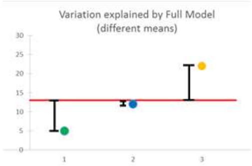
$((Y_i|X=i) - \hat{\mu}_i)^2$

	Level i=1	Level i=2	Level i=3
$((Y_1\|X=i) - \hat{\mu}_i)^2$	$(3-5)^2 = 4$	$(10-12)^2 = 4$	$(20-22)^2 = 4$
$((Y_2\|X=i) - \hat{\mu}_i)^2$	0	0	0
$((Y_3\|X=i) - \hat{\mu}_i)^2$	4	4	4
*Total Sum of Squared Residuals for **Separate** Means Model:* 24			

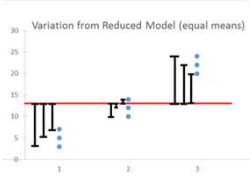
6. Compare the Total Sum of Squares for each model. Which do you think “fits” better?

Sum of Squares in ANOVA

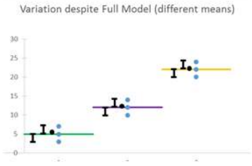
Between group variation (top row)



Total variation (bottom row)



Within group variation (middle row)



*To compute the sum of squares column for the ANOVA table, square each distance (lines in black) and then add.

The sum of squared* distances (black lines) for left two graphs = the sum of squared distances (black lines) for the right graph.

*Each distance squared for the top left graph is multiplied by the number in each group.

Pure ANOVA

	Level i=1	Level i=2	Level i=3
$Y_1 \vert X=i$	3	10	20
$Y_2 \vert X=i$	5	12	22
$Y_3 \vert X=i$	7	14	24
$\hat{\mu}_{Y \vert X=i}$			

7. Now we would like to make an ANOVA table to test the alternative hypothesis!

Formally write the H_o and H_a and fill in the table.

	df	SS	MS	F	Pr > F
Model / Extra SS					
Error / Residual/Full Model					
Total (Reduced)					

Extra Sum of Squares = Residual Sum of Squares Reduced – Residual Sum of Squares Full

Pure ANOVA

7. Now we would like to make an ANOVA table to test the alternative hypothesis!

Formally write the Ho and Ha and fill in the table.

H_o: $\mu_1 = \mu_2 = \mu_3$ **(Equal Means Model μ μ μ)**
H_a: At least 1 pair are different **(Separate Means Model μ_1 μ_2 μ_3)**

	df	SS	MS	F	Pr > F
Model / Extra SS					
Error / Residual/Full Model	**6**	**24**	**4**		
Total (Reduced)	**8**	**462**			

Extra Sum of Squares = Residual Sum of Squares Reduced – Residual Sum of Squares Full

Pure ANOVA

7. Now we would like to make an ANOVA table to test the alternative hypothesis!

Formally write the Ho and Ha and fill in the table.

H_o: $\mu_1 = \mu_2 = \mu_3$ **(Equal Means Model μ μ μ)**
H_a: At least 1 pair are different **(Separate Means Model μ_1 μ_2 μ_3)**

	df	SS	MS	F	Pr > F
Model / Extra SS	**8-6=2**	**462-24=438**			
Error / Residual/Full Model	**6**	**24**	**4**		
Total (Reduced)	**8**	**462**			

Extra Sum of Squares = Residual Sum of Squares Reduced – Residual Sum of Squares Full

Pure ANOVA

7. Now we would like to make an ANOVA table to test the alternative hypothesis!

Formally write the Ho and Ha and fill in the table.

H_o: $\mu_1 = \mu_2 = \mu_3$ **(Equal Means Model μ μ μ)**
H_a: At least 1 pair are different **(Separate Means Model μ_1 μ_2 μ_3)**

	df	SS	MS	F	Pr > F
Model / Extra SS	**2**	**438**	**438/2=219**		
Error / Residual/Full Model	**6**	**24**	**4**		
Total (Reduced)	**8**	**462**			

Extra Sum of Squares = Residual Sum of Squares Reduced – Residual Sum of Squares Full

Pure ANOVA

7. Now we would like to make an ANOVA table to test the alternative hypothesis!

Formally write the Ho and Ha and fill in the table.

H_o: $\mu_1 = \mu_2 = \mu_3$ **(Equal Means Model μ μ μ)**
H_a: At least 1 pair are different **(Separate Means Model μ_1 μ_2 μ_3)**

	df	SS	MS	F	Pr > F
Model / Extra SS	**2**	**438**	**219**	**219/4=54.75**	
Error / Residual/Full Model	**6**	**24**	**4**		
Total (Reduced)	**8**	**462**			

Extra Sum of Squares = Residual Sum of Squares Reduced – Residual Sum of Squares Full

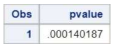
Pure ANOVA

7. Now we would like to make an ANOVA table to test the alternative hypothesis!

Formally write the H_o and H_a and fill in the table.

H_o: $\mu_1 = \mu_2 = \mu_3$ **(Equal Means Model μ μ μ)**
H_a: At least 1 pair are different **(Separate Means Model μ_1 μ_2 μ_3)**

```
data pval;
pvalue = 1-probf(54.75, 2, 6);
run;
proc print data = pval;
run;
```



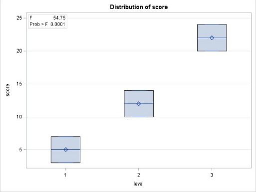
Obs	pvalue
1	.000140187

	df	SS	MS	F	Pr > F
Model / Extra SS	**2**	**438**	**219**	**54.75**	**.0001**
Error / Residual/Full Model	**6**	**24**	**4**		
Total (Reduced)	**8**	**462**			

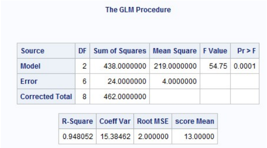
Extra Sum of Squares = Residual Sum of Squares Reduced – Residual Sum of Squares Full

F -Test of Different Means …

H_o: $\mu_1 = \mu_2 = \mu_3$ (Equal Means Model)
H_a: At least 1 pair are different (Separate Means Model)



```
data AnovaData;
input score level;
datalines;
3 1
5 1
7 1
10 2
12 2
14 2
20 3
22 3
24 3
;

proc glm data = AnovaData;
class level;
model score = level;
run;
```



The GLM Procedure

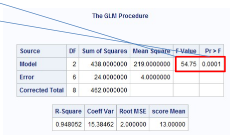
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	438.0000000	219.0000000	54.75	0.0001
Error	6	24.0000000	4.0000000		
Corrected Total	8	462.0000000			

R-Square	Coeff Var	Root MSE	score Mean
0.948052	15.38462	2.000000	13.00000

6 Steps for ANOVA F Test (diff means)!

1. H_o: $\mu_1 = \mu_2 = \mu_3$ (Equal Means Model)
   H_a: At least 1 pair are different (Separate Means Model)
2. Critical value: You can skip this step for ANOVA.
3. F statistic = 54.75
4. P-value = .0001
5. Reject Ho.
6. The evidence suggests that at least 1 pair of the group means are different. (P-value < .0001 from an ANOVA.)

```
proc glm data = AnovaData;
class level;
model score = level;
run;
```

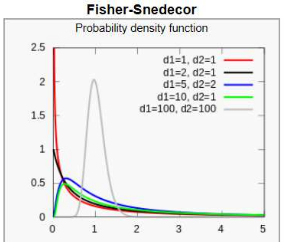


The GLM Procedure

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	438.0000000	219.0000000	54.75	0.0001
Error	6	24.0000000	4.0000000		
Corrected Total	8	462.0000000			

R-Square	Coeff Var	Root MSE	score Mean
0.948052	15.38462	2.000000	13.00000

F-Distribution

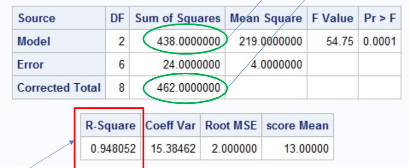


$$F - \text{Statistic} = \frac{\frac{Extra\ Sum\ of\ Squares}{Extra\ Degress\ of\ Freedom}}{\hat{\sigma}^2_{Full}} = \frac{MS\ Between}{MS\ Within} = \frac{Variation\ Explained\ by\ Full\ Model}{Variation\ Left\ to\ be\ Explained}$$

R-Squared!

R =correlation coefficient
R^2 = coefficient of determination

$$R - \text{Squared} = \frac{Variation\ Explained\ by\ Full\ Model}{Total\ Variation} = \frac{Extra\ Sum\ of\ Squares}{Total\ Sum\ of\ Squares}$$



Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	438.0000000	219.0000000	54.75	0.0001
Error	6	24.0000000	4.0000000		
Corrected Total	8	462.0000000			

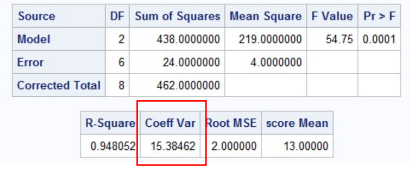
R-Square	Coeff Var	Root MSE	score Mean
0.948052	15.38462	2.000000	13.00000

$$R - \text{Squared} = \frac{438}{462} = 0.948052$$

*Rho (ρ) is the parameter for which r is an estimate (just like μ and $\bar{x}$ or σ and s). A hypothesis test of whether ρ =0 is equivalent to the basic ANOVA test of whether all the means are the same (try it!).

Coefficient of Variation

$$\text{Coefficient of Variation} = \frac{square\ root\ of\ the\ unexplained\ variation}{grand\ mean} x\ 100\%$$



Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	438.0000000	219.0000000	54.75	0.0001
Error	6	24.0000000	4.0000000		
Corrected Total	8	462.0000000			

R-Square	Coeff Var	Root MSE	score Mean
0.948052	15.38462	2.000000	13.00000

$$\text{Coefficient of Variation} = \frac{\sqrt{MSE}}{\bar{X}} x\ 100 = \frac{2}{13} x\ 100 = 15.38462$$

Coefficient of Variation is also called the relative standard deviation.

ANOVA: Assumptions and Robustness

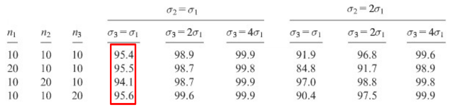
1. Normality: Similar to t-tools hypothesis testing, ANOVA is robust to this assumption. Extremely long-tailed distributions (outliers) or skewed distributions, coupled with different sample sizes (especially when the sample sizes are small) present the only serious distributional problems.
2. Equal Standard Deviations: This assumption is crucial, paramount, and VERY important.
3. The assumptions of independence within and across groups are critical. If lacking, different analysis should be attempted.

Samples drawn from Normal Distributions

- Same visual checks as with t-tools, just for more groups.
  - Histograms
  - Q-Q plots

More on Constant SD

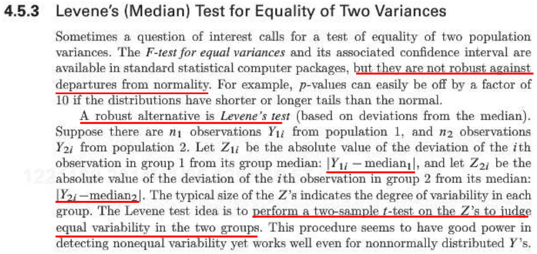
95% confidence interval accuracy with different sample sizes and standard deviations for three groups.



			$\sigma_2=\sigma_1$			$\sigma_2=2\sigma_1$		
n_1	n_2	n_3	$\sigma_3=\sigma_1$	$\sigma_3=2\sigma_1$	$\sigma_3=4\sigma_1$	$\sigma_3=\sigma_1$	$\sigma_3=2\sigma_1$	$\sigma_3=4\sigma_1$
10	10	10	95.4	98.9	99.9	91.9	96.8	99.6
20	10	10	95.5	98.7	99.8	84.8	91.7	98.9
10	20	10	94.1	98.7	99.9	97.0	98.8	99.8
10	10	20	95.6	99.6	99.9	90.4	97.5	99.9

Levene's Test (Median)

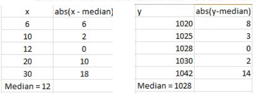
$H_o: \sigma_1 = \sigma_2$
$H_a: \sigma_1 \neq \sigma_2$



4.5.3 Levene's (Median) Test for Equality of Two Variances

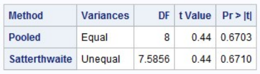
Sometimes a question of interest calls for a test of equality of two population variances. The *F-test for equal variances* and its associated confidence interval are available in standard statistical computer packages, but they are not robust against departures from normality. For example, p-values can easily be off by a factor of 10 if the distributions have shorter or longer tails than the normal.

A robust alternative is *Levene's test* (based on deviations from the median). Suppose there are n_1 observations Y_{1i} from population 1, and n_2 observations Y_{2i} from population 2. Let Z_{1i} be the absolute value of the deviation of the ith observation in group 1 from its group median: $|Y_{1i} - \text{median}_1|$, and let Z_{2i} be the absolute value of the deviation of the ith observation in group 2 from its median: $|Y_{2i} - \text{median}_2|$. The typical size of the Z's indicates the degree of variability in each group. The Levene test idea is to perform a two-sample t-test on the Z's to judge equal variability in the two groups. This procedure seems to have good power in detecting nonequal variability yet works well even for nonnormally distributed Y's.



x	abs(x - median)
6	6
10	2
12	0
20	10
30	18
Median = 12	

y	abs(y-median)
1020	8
1025	3
1028	0
1030	2
1042	14
Median = 1028	

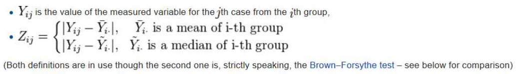


Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	8	0.44	0.6703
Satterthwaite	Unequal	7.5856	0.44	0.6710

But … proc ttest does not have Levene's Test!!!

Proc GLM Has Levene's Test

```
proc glm data = Spock_ttest;
class judge;
model percentage = judge
means judge / HOVTEST = Levene;
run;
```

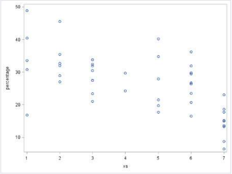


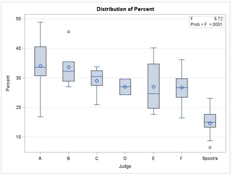
- Y_{ij} is the value of the measured variable for the jth case from the ith group,
- $Z_{ij} = \begin{cases} |Y_{ij} - \bar{Y}_{i\cdot}|, & \bar{Y}_{i\cdot} \text{ is a mean of i-th group} \\ |Y_{ij} - \tilde{Y}_{i\cdot}|, & \tilde{Y}_{i\cdot} \text{ is a median of i-th group} \end{cases}$

(Both definitions are in use though the second one is, strictly speaking, the Brown–Forsythe test – see below for comparison)

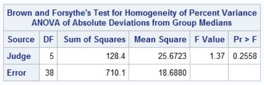
```
proc glm data = Spock_ttest;
class judge;
model percentage = judge;
means judge / HOVTEST = BF;
run;
```

Check of Assumptions: Constant SD





```
/* Generates Scatterplot */
proc sgplot data=Spock1;
scatter x=xs y=percentage;
run;
```



Brown and Forsythe's Test for Homogeneity of Percent Variance ANOVA of Absolute Deviations from Group Medians					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Judge	5	128.4	25.6723	1.37	0.2558
Error	38	710.1	18.6880		

There is some visual evidence against equal standard deviations. The Brown-Forsythe test was used as secondary evidence and does not provide significant evidence against equal standard deviations. (p-value = .2558)

```
proc glm data = spock;
class judge;
model percent = judge;
means judge / hovtest = bf;
run;
```

Archeology in New Mexico

An archeological dig in New Mexico yielded four sites with lots of artifacts. The depth (cm) that each artifact was found was recorded along with which site it was found in.

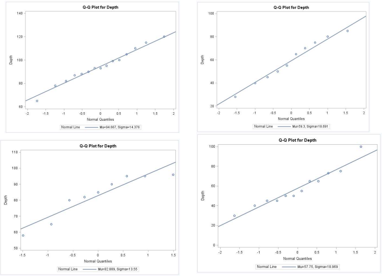
The researcher has reason to believe that sites 1 and 4 and sites 2 and 3 may be similar in age. In theory, the deeper the find, the older the village.

Is there any evidence that sites 1 and 4 have a mean depth that is different than the mean depth of artifacts from sites 2 and 3?

Archaeology Example

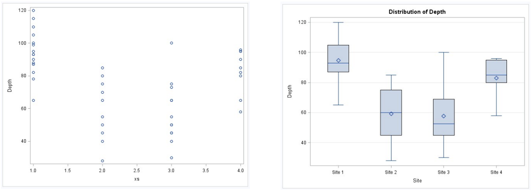
Depth	Site		Depth	Site		Depth	Site		Depth	Site
93	1		85	2		100	3		96	4
120	1		45	2		75	3		58	4
65	1		80	2		65	3		95	4
105	1		28	2		40	3		90	4
115	1		75	2		73	3		65	4
82	1		70	2		65	3		80	4
99	1		65	2		50	3		85	4
87	1		55	2		30	3		95	4
100	1		50	2		45	3		82	4
90	1		40	2		50	3			
78	1					45	3			
95	1					55	3			
93	1									
88	1									
110	1									

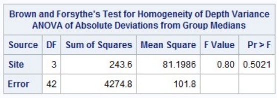
Archeology Example
Assumptions: Normality



Histograms will be helpful as well!

Archeology Example
Assumptions: Homogeneity (Equal SD)





Brown and Forsythe's Test for Homogeneity of Depth Variance
ANOVA of Absolute Deviations from Group Medians

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Site	3	243.6	81.1986	0.80	0.5021
Error	42	4274.8	101.8		

Archeology Example
Assumption: Independence

The discovered artifacts associated with the depths were randomly selected from the log (book of recordings … not logarithms!) of discoveries.

Since the artifacts and, thus, the depths are associated with completely different sites, it is assumed that the data are independent between sites.

Question of Interest:

1. Are any of the means different?

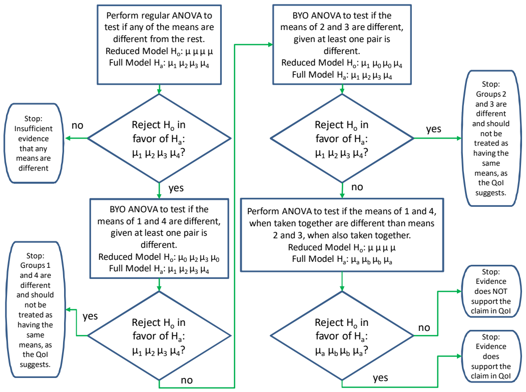
2. Are the means of sites 1 and 4 different?

3. Are the means of sites 2 and 3 different?

4. Satisfactory results of questions 1 and 2 will allow us to ask the third question: are sites 1 and 4 different than 2 and 3?

Are sites 1 and 4 different from 2 and 3?

*Assumes ANOVA assumptions are met



Perform regular ANOVA to test if any of the means are different from the rest.
Reduced Model H_o: $\mu\ \mu\ \mu\ \mu$
Full Model H_a: $\mu_1\ \mu_2\ \mu_3\ \mu_4$

Reject H_o in favor of H_a: $\mu_1\ \mu_2\ \mu_3\ \mu_4$?

no → Stop: Insufficient evidence that any means are different

yes → BYO ANOVA to test if the means of 1 and 4 are different, given at least one pair is different.
Reduced Model H_o: $\mu_0\ \mu_2\ \mu_3\ \mu_0$
Full Model H_a: $\mu_1\ \mu_2\ \mu_3\ \mu_4$

Reject H_o in favor of H_a: $\mu_1\ \mu_2\ \mu_3\ \mu_4$?

yes → Stop: Groups 1 and 4 are different and should not be treated as having the same means, as the QoI suggests.

no → BYO ANOVA to test if the means of 2 and 3 are different, given at least one pair is different.
Reduced Model H_o: $\mu_1\ \mu_0\ \mu_0\ \mu_4$
Full Model H_a: $\mu_1\ \mu_2\ \mu_3\ \mu_4$

Reject H_o in favor of H_a: $\mu_1\ \mu_2\ \mu_3\ \mu_4$?

yes → Stop: Groups 2 and 3 are different and should not be treated as having the same means, as the QoI suggests.

no → Perform ANOVA to test if the means of 1 and 4, when taken together are different than means 2 and 3, when also taken together.
Reduced Model H_o: $\mu\ \mu\ \mu\ \mu$
Full Model H_a: $\mu_a\ \mu_b\ \mu_b\ \mu_a$

Reject H_o in favor of H_a: $\mu_a\ \mu_b\ \mu_b\ \mu_a$?

no → Stop: Evidence does NOT support the claim in QoI

yes → Stop: Evidence does support the claim in QoI

First Ask: Is there reason to believe any of them are different?

The reduced and full models are associated with H_o and H_a, respectively, although they are not exactly equal to the hypotheses.

(H_o) Reduced Model: $\mu\ \mu\ \mu\ \mu$

(H_a) Full Model: $\mu_1\ \mu_2\ \mu_3\ \mu_4$

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	12397.34082	4132.44694	15.14	<.0001
Error	42	11464.57222	272.96601		
Corrected Total	45	23861.91304			

There is evidence to suggest that at the alpha = .05 level of significance (p-value < .0001) that at least 2 of the sites have different mean depths.

Question of Interest:
2. Are the means of sites 1 and 4 different?

*Recode the variables into three groups: 2, 3, and 1/4 combined and perform ANOVA to get the first table.

(H_o) Reduced Model: $\mu_o\ \mu_2\ \mu_3\ \mu_o$

(H_a) Full Model: $\mu_1\ \mu_2\ \mu_3\ \mu_4$

Compare this model against equal means model ($\mu\ \mu\ \mu\ \mu$)

Compare this model against equal means model ($\mu\ \mu\ \mu\ \mu$)

(H_o) Reduced: $\mu\ \mu\ \mu\ \mu$
(H_a) Full*: $\mu_o\ \mu_2\ \mu_3\ \mu_o$

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	11617.06304	5808.53152	20.40	<.0001
Error	43	12244.85000	284.76395		
Corrected Total	45	23861.91304			

(H_o) Reduced: $\mu\ \mu\ \mu\ \mu$
(H_a) Full: $\mu_1\ \mu_2\ \mu_3\ \mu_4$

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	12397.34082	4132.44694	15.14	<.0001
Error	42	11464.57222	272.96601		
Corrected Total	45	23861.91304			

Source	DF	SS	MS	F	Pr>F
Model (Full)	1	780.3	780.3	2.86	.098
Error (From Full)	42	11464.6	273.0		
Total (From Reduced*)	43	12244.9			

There is not enough evidence to suggest (alpha = .05, p-value = .098) that site 1 and site 4 have different mean depths.

Question of Interest: (try it!)
3. Are the means of sites 2 and 3 different?

*Recode the variables into three groups: 1, 4, and 2/3 combined and perform ANOVA to get the first table.

(H_o) Reduced Model: $\mu_1\ \mu_o\ \mu_o\ \mu_4$

(H_a) Full Model: $\mu_1\ \mu_2\ \mu_3\ \mu_4$

(H_o) Reduced: $\mu\ \mu\ \mu\ \mu$
(H_a) Full*: $\mu_1\ \mu_o\ \mu_o\ \mu_4$

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	12384.23628	6192.11814	23.20	<.0001
Error	43	11477.67677	266.92272		
Corrected Total	45	23861.91304			

(H_o) Reduced: $\mu\ \mu\ \mu\ \mu$
(H_a) Full: $\mu_1\ \mu_2\ \mu_3\ \mu_4$

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	12397.34082	4132.44694	15.14	<.0001
Error	42	11464.57222	272.96601		
Corrected Total	45	23861.91304			

Source	DF	SS	MS	F	Pr>F
Model (Full)					
Error (From Full)					
Total (From Reduced*)					

Question of Interest: (try it!)
3. Are the means of sites 2 and 3 different?

*Recode the variables into three groups: 1, 4, and 2/3 combined and perform ANOVA to get the first table.

(H_o) Reduced Model: $\mu_1\ \mu_o\ \mu_o\ \mu_4$

(H_a)Full Model: $\mu_1\ \mu_2\ \mu_3\ \mu_4$

(H_o) Reduced: $\mu\ \mu\ \mu\ \mu$
(H_a) Full*: $\mu_1\ \mu_o\ \mu_o\ \mu_4$

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	12384.23628	6192.11814	23.20	<.0001
Error	43	11477.67677	266.92272		
Corrected Total	45	23861.91304			

(H_o) Reduced: $\mu\ \mu\ \mu\ \mu$
(H_a) Full: $\mu_1\ \mu_2\ \mu_3\ \mu_4$

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	12397.34082	4132.44694	15.14	<.0001
Error	42	11464.57222	272.96601		
Corrected Total	45	23861.91304			

Source	DF	SS	MS	F	Pr>F
Model (Full)					
Error (From Full)	42	11464.6	273		
Total (From Reduced)	43	11477.7			

Question of Interest: (try it!)
3. Are the means of sites 2 and 3 different?

*Recode the variables into three groups: 1, 4, and 2/3 combined and perform ANOVA to get the first table.

(H_o) Reduced Model: $\mu_1\ \mu_o\ \mu_o\ \mu_4$

(H_a) Full Model: $\mu_1\ \mu_2\ \mu_3\ \mu_4$

(H_o) Reduced: $\mu\ \mu\ \mu\ \mu$
(H_a) Full*: $\mu_1\ \mu_o\ \mu_o\ \mu_4$

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	12384.23628	6192.11814	23.20	<.0001
Error	43	11477.67677	266.92272		
Corrected Total	45	23861.91304			

(H_o) Reduced: $\mu\ \mu\ \mu\ \mu$
(H_a) Full: $\mu_1\ \mu_2\ \mu_3\ \mu_4$

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	12397.34082	4132.44694	15.14	<.0001
Error	42	11464.57222	272.96601		
Corrected Total	45	23861.91304			

Source	DF	SS	MS	F	Pr>F
Model (Full)	1	13.1	13.1	.048	.828
Error (From Full)	42	11464.6	273		
Total (From Reduced)	43	11477.7			

There is not enough evidence to suggest (alpha = .05, p-value = .828) that site 2 and site 3 have different mean depths.

Question of Interest:
4. Are sites 1 and 4 different than 2 and 3?

*Recode the
variables into two
groups 1/4 and 2/3
and perform ANOVA
to get the table.

(H_o) Reduced: μ μ μ μ

(H_a) Full: μ_b μ_a μ_a μ_b

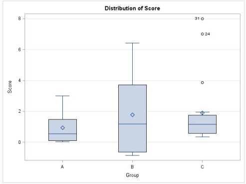
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	11603.95850	11603.95850	41.65	<.0001
Error	44	12257.95455	278.58988		
Corrected Total	45	23861.91304			

There is sufficient evidence to suggest (alpha = .05, p-value < .0001) that sites 1 and 4 have different mean depths than sites 2 and 3.

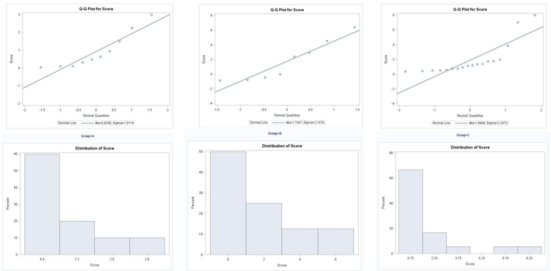
A Small Example

```
data Example;
input Group $ Score xs;
datalines;
A       0.626613219      1
A       2.257851416      1
A       0.094928519      1
A       0.110304406      1
A       1.487169637      1
A       0.472428856      1
A       0.939020998      1
A       3       1
A       0.038837703      1
A       0.308524833      1
B       6.434402812      2
B       -0.808322517      2
B       -0.459055808      2
B       2.385018878      2
B       2.931932653      2
B       4.525187028      2
B       -0.031221249      2
B       -0.859967193      2
C       1.211803482      3
C       0.91899087      3
C       1.745859656      3
C       1.386666032      3
C       0.742693027      3
C       7.012310251      3
C       0.51000385      3
C       1.955086042      3
C       0.424691139      3
C       1.344755353      3
C       1.715018431      3
C       3.875406227      3
C       8       3
C       0.496604384      3
C       0.750848548      3
C       0.340235176      3
C       0.577707141      3
C       1.133029678      3
```

Level of Group	N	Score	
		Mean	Std Dev
A	10	0.93356796	1.01157431
B	8	1.76474683	2.74781436
C	18	1.89676163	2.20726331



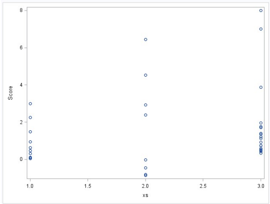
Normality Assumption



```
proc univariate data = Example;
by group;
histogram score;
qqplot score;
run;
```

There is strong evidence against these data coming from a normal distribution and the sample size is small. ANOVA? WELCH'S ANOVA?

Homogeneity of Variance Assumption

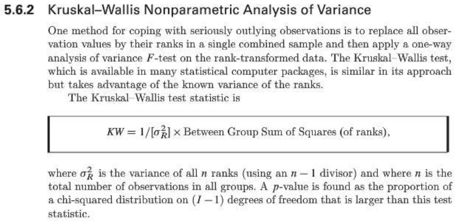


Brown and Forsythe's Test for Homogeneity of Score Variance ANOVA of Absolute Deviations from Group Medians					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Group	2	11.3518	5.6759	2.26	0.1207
Error	33	83.0246	2.5159		

There is some (weak) evidence in support of these data coming from distributions with different standard deviations. If the standard deviation assumption and normality assumption are both violated, what should we do?

```
proc glm data = Example;
class group;
model score = group;
means group / hovtest = bf;
run;
```

So …. NONPARAMETRIC!!!!



5.6.2 Kruskal–Wallis Nonparametric Analysis of Variance

One method for coping with seriously outlying observations is to replace all observation values by their ranks in a single combined sample and then apply a one-way analysis of variance F-test on the rank-transformed data. The Kruskal–Wallis test, which is available in many statistical computer packages, is similar in its approach but takes advantage of the known variance of the ranks.

The Kruskal–Wallis test statistic is

$$\boxed{KW = 1/[\sigma_R^2] \times \text{Between Group Sum of Squares (of ranks)},}$$

where σ_R^2 is the variance of all n ranks (using an $n-1$ divisor) and where n is the total number of observations in all groups. A p-value is found as the proportion of a chi-squared distribution on $(I-1)$ degrees of freedom that is larger than this test statistic.

Kruskal-Wallis Test

Ho: $\text{Median}_{\text{Group1}} = \text{Median}_{\text{Group2}} = \text{Median}_{\text{Group3}}$
Ha: At least 1 pair of medians are different.

Brown and Forsythe's Test for Homogeneity of Score Variance ANOVA of Absolute Deviations from Group Medians					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Group	2	11.3518	5.6759	2.26	0.1207
Error	33	83.0246	2.5159		

Welch's ANOVA for Score			
Source	DF	F Value	Pr > F
Group	2.0000	1.35	0.2885
Error	15.9313		

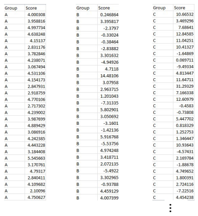
```
proc npar1way data = Example Wilcoxon;
class group;
var score;
run;
```

Kruskal-Wallis Test	
Chi-Square	1.9534
DF	2
Pr > Chi-Square	0.3766

There is not sufficient evidence at the alpha = .05 level of significance (p-value = .3766 from Kruskal-Wallis Test) to suggest that at least two of the medians are different.

Notice that each test failed to reject their respective H_o. The point isn't so much that one test will reject when the other will fail to reject. We must remember that as statisticians, we don't personally favor one outcome over the other. We just want the appropriate test: the one with the most power. Kruskal-Wallis Test is the ***appropriate*** test here.

Another Analysis!!!!

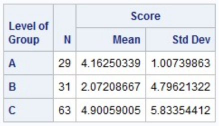


Group	Score
A	4.000308
A	3.958816
A	4.997734
A	4.638248
A	4.15137
A	2.831176
A	3.782846
A	4.238071
A	3.067494
A	4.531106
A	4.154173
A	2.847931
A	2.918759
A	4.770106
A	2.717302
A	4.239002
A	3.987699
A	4.889429
A	3.086916
A	4.242385
A	4.443228
A	3.184408
A	5.545663
A	3.170761
A	4.79317
A	2.840411
A	4.109682
A	2.10096
A	4.750627

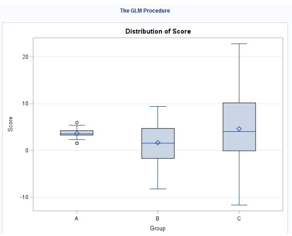
Group	Score
B	0.246864
B	3.395817
B	-2.3797
B	-0.33024
B	-0.38464
B	-2.83882
B	3.301632
B	-4.94926
B	4.7118
B	14.48106
B	3.07958
B	2.963715
B	1.201043
B	-7.31335
B	5.802901
B	3.050692
B	-3.1601
B	-1.42136
B	5.916768
B	-5.53756
B	4.974248
B	3.418711
B	2.072135
B	-5.4922
B	3.302965
B	-0.93788
B	4.459129
B	4.007399

Group	Score
C	10.66532
C	3.469296
C	7.68641
C	12.84585
C	11.04251
C	10.41327
C	-1.64869
C	0.069711
C	-9.49334
C	4.813447
C	11.64711
C	31.29329
C	7.166338
C	12.60979
C	-0.4583
C	-0.73808
C	5.447702
C	0.818329
C	1.252753
C	1.346447
C	10.93643
C	-4.57431
C	2.169784
C	-1.88678
C	4.749652
C	1.800391
C	2.724116
C	-7.22516
C	4.454238

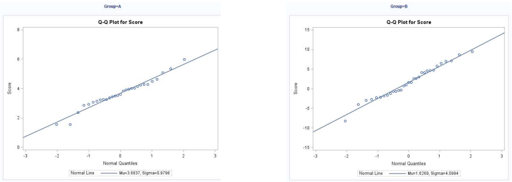
⋮

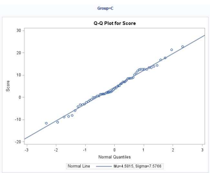


Level of Group	N	Score	
		Mean	Std Dev
A	29	4.16250339	1.00739863
B	31	2.07208667	4.79621322
C	63	4.90059005	5.83354412



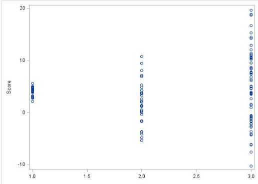
Normality Assumption

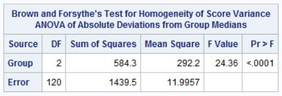




There is strong evidence in favor of these data coming from a normal distribution. We will proceed under this assumption.

Assumptions and Analysis:



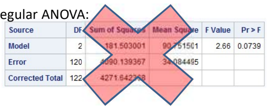


Brown and Forsythe's Test for Homogeneity of Score Variance
ANOVA of Absolute Deviations from Group Medians

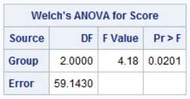
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Group	2	584.3	292.2	24.36	<.0001
Error	120	1439.5	11.9957		

There is strong evidence in support of these data coming from distributions with different standard deviations. We will proceed under this assumption and run the Welch's ANOVA.

Regular ANOVA:



Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	~~181.503001~~	~~90.751501~~	2.66	0.0739
Error	120	~~4090.139367~~	~~34.084495~~		
Corrected Total	122	~~4271.642368~~			



Welch's ANOVA for Score			
Source	DF	F Value	Pr > F
Group	2.0000	4.18	0.0201
Error	59.1430		

```
proc glm data = example;
class group;
model score = group;
means group / hovtest = bf Welch;
run;
```

There is sufficient evidence at the alpha = .05 level of significance (p-value = .0201 from Welch's ANOVA) to suggest that at least two of the means are different. However, remember caveat to any different SD's approach.

Fixed Effects vs. Random Effects

Quick answer:

- Do your groupings exhaust the data (e.g., data on four different machines and there are only four machines)? Fixed Effects! Use Proc GLM in SAS.
- Are your groupings a random sample of a larger population that could have been chosen to be a group (e.g., data on four different machines that were chosen from a random sample of 100 machines)? Random Effects! Use Proc Mixed in SAS.

Fixed or random effects

Measured the amount of liquid in twenty randomly selected cans of Coke and twenty randomly selected cans of Diet Coke at a regional bottling company. Coke and Diet Coke are bottled using different types of machines.

Scenario 1: There is only one machine of each type.

Fixed Effects

Scenario 2: There are several of each type of machine. The Coke samples all came from the same Coke bottling machine, and the Diet Coke samples all came from the same Diet Coke machine.

Random effects

APPENDIX

What does r^2 mean?

- r^2 is called the coefficient of determination, or square of the correlation coefficient
- $r^2 = \frac{SS_{model}}{SS_{total}}$

We can think of r^2 as the proportion of variability that is explained by the independent variables (grouping data).

What does r^2 mean?

While r^2 is gleaned from the data, the true parameter is referred to as ρ (rho).

The following two hypothesis tests are equivalent:

- 1

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$$
$$H_1: at\ least\ 1\ \mu_i\ is\ different$$

Test statistic:

$F = \frac{MS(model)}{MS(error)}$, where F is F −distributed with $k-1$, $n-k$ degrees of freedom

- 2

$$H_0: \rho = 0$$
$$H_1: \rho \neq 0$$

Test statistic:

$F = \frac{r^2(n-k)}{(1-r^2)(k-1)}$, where F is F −distributed with $k-1$, $n-k$ degrees of freedom

What does r^2 mean?

Let $F_1 = \frac{r^2(n-k)}{(1-r^2)(k-1)}$ and $F_2 = \frac{MS(model)}{MS(error)}$, where k is the number of groups and n is the total number of data points.

Recall that

$r^2 = \frac{SS_{model}}{SS_{total}} = \frac{SS_{total} - SS_{error}}{SS_{total}} = 1 - \frac{SS_{error}}{SS_{total}}$.

So, $1 - r^2 = \frac{SS_{error}}{SS_{total}}$.

Also remember that $MS(model) = \frac{SS(model)}{k-1}$ and $MS(error) = \frac{SS(error)}{n-k}$.

$$\frac{r^2(n-k)}{(1-r^2)(k-1)} = \frac{\frac{SS_{model}}{SS_{total}}(n-k)}{\frac{SS_{error}}{SS_{total}}(k-1)} = \frac{SS_{model}(n-k)}{SS_{error}(k-1)} = \frac{SS_{model}/(k-1)}{SS_{error}/(n-k)}$$
$$= \frac{MS(model)}{MS(error)}$$

Therefore, $F_1 = F_2$.

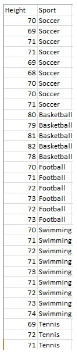
MSE vs. Variance in each group

MSE is a weighted average of the sample variances of each group. Let s_i^2 be the sample variance in group i.

$$MSE = s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + \cdots + (n_k-1)s_k^2}{(n_1-1)+(n_2-1)+\cdots+(n_k-1)}$$
$$MSE = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + \cdots + (n_k-1)s_k^2}{n-k}$$

Examples

Another example!

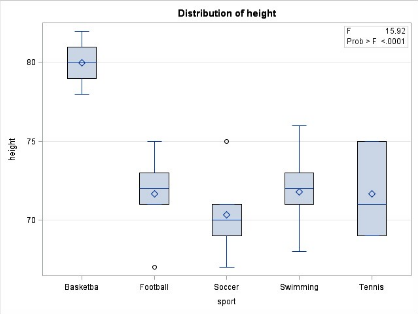


Height	Sport
70	Soccer
69	Soccer
71	Soccer
71	Soccer
69	Soccer
68	Soccer
70	Soccer
70	Soccer
71	Soccer
80	Basketball
79	Basketball
81	Basketball
82	Basketball
78	Basketball
70	Football
71	Football
72	Football
73	Football
72	Football
73	Football
70	Swimming
71	Swimming
72	Swimming
71	Swimming
73	Swimming
71	Swimming
72	Swimming
73	Swimming
74	Swimming
69	Tennis
72	Tennis
71	Tennis

5 different sports were analyzed to see if the average height of basketball players was greater than the average of all the other sports. We could, of course, compare each pairwise grouping of sports, but that would result in 4 tests. This would take a lot of time, and those tests would each have less power since they don't use all the data. Let's use ANOVA similarly to how we did in prior problems.

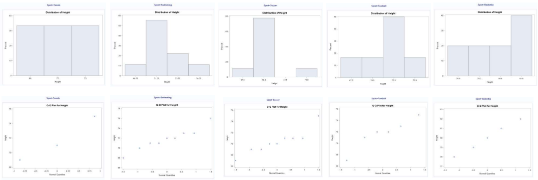
1. Make a side by side box plot of the data.
2. Run a basic ANOVA to test for any pairwise difference of means. Check the assumptions here, but no need to address them after this.
3. Test the model that keeps basketball by itself but groups the other sports as "others."
4. Use the previous two models to conduct an extra sum of squares F-Test:
   H_o: Reduced Model: μ_B μ_O μ_O μ_O μ_O
   H_a: Full Model: μ_B μ_F μ_{Soc} μ_{Swim} μ_T
5. Depending on the results of this test, test to see if there is evidence that basketball has a different mean than each of the sports. (Equivalent to testing basketball versus the others.)
   H_o: Reduced Model: μ_O μ_O μ_O μ_O μ_O
   H_a: Full Model: μ_B μ_O μ_O μ_O μ_O
6. Make sure and provide written conclusions for questions 2,3,4 and 5.

First … Plot the Data!



Plot the Data cont.

```
proc univariate data = basketball;
by sport;
histogram height;
qqplot height;
run;
```

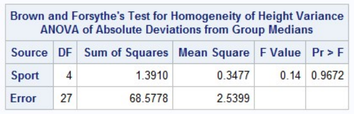


Normality: We have very small sample sizes here. There is not a lot of evidence against normality for each group, although there is not a lot of evidence to begin with. We will proceed with caution under the assumption of normal distributions for each sport.

Homogeneity of Variance: Judging from the box plots, there is some visual evidence against equal standard deviations, although the sample size is still small. A secondary test would be nice to lean on here.

We will assume the observations are independent both between and within groups.

Brown and Forsythe Test for Equality of Variance.



Brown and Forsythe's Test for Homogeneity of Height Variance
ANOVA of Absolute Deviations from Group Medians

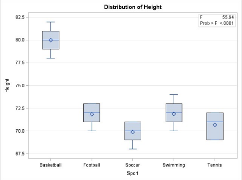
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Sport	4	1.3910	0.3477	0.14	0.9672
Error	27	68.5778	2.5399		

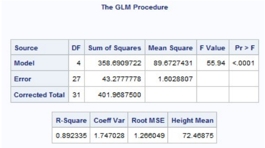
There is some visual evidence against equal standard deviations between sports. The Brown and Forsythe test was used as secondary evidence and does not provide significant evidence against equal standard deviations. (p-value = .9672)

---

1 Way ANOVA

H_o: $\mu_{Basketball} = \mu_{Football} = \mu_{Soccer} = \mu_{Swim} = \mu_{Tennis}$
H_a: At least one pair of means is different.





The GLM Procedure

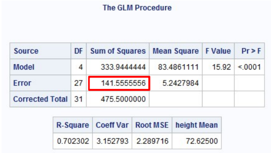
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	358.6909722	89.6727431	55.94	<.0001
Error	27	43.2777778	1.6028807		
Corrected Total	31	401.9687500			

R-Square	Coeff Var	Root MSE	Height Mean
0.892335	1.747028	1.266049	72.46875

There is strong evidence to suggest that the at least one of the sports has a mean height that is different than the others (p-value < .0001 from an ANOVA).

H_o: $\mu_{Basketball} = \mu_{Football} = \mu_{Soccer} = \mu_{Swim} = \mu_{Tennis}$

H_a: At least one pair of means are different.



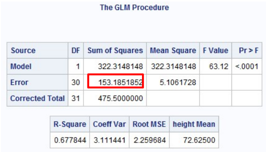
The GLM Procedure

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	333.9444444	83.4861111	15.92	<.0001
Error	27	141.5555556	5.2427984		
Corrected Total	31	475.5000000			

R-Square	Coeff Var	Root MSE	height Mean
0.702302	3.152793	2.289716	72.62500

H_o: $\mu_{Basketball} = \mu_{Football} = \mu_{Soccer} = \mu_{Swim} = \mu_{Tennis}$

H_a: $\mu_{Basketball}$ is different than the Others.



The GLM Procedure

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	322.3148148	322.3148148	63.12	<.0001
Error	30	153.1851852	5.1061728		
Corrected Total	31	475.5000000			

R-Square	Coeff Var	Root MSE	height Mean
0.677844	3.111441	2.259684	72.62500

F-TEST

H_o: The Others are equal. (Including Basketball)

H_a: The Others are different (Including Basketball)

$$F = \frac{\frac{Extra\ Sum\ of\ Squares}{ExtraDegress\ of\ Freedom}}{\hat{\sigma}^2{}_{Full}}$$

$$F = \frac{(153.19\ - 141.56)/(30 - 27)}{141.56/27}$$

$$F = .74$$

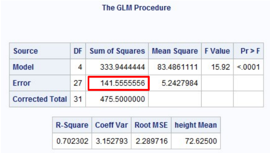
P-value = 0.5375

Fail to Reject Ho

There is not sufficient evidence at the alpha = .05 level of significance (p-value = 0.5375) to suggest that the mean heights of non-basketball sports are not equal. Therefore we will proceed as if they are equal.

Same Test as last slide ….
Different Notation

H_o: Reduced Model: $\mu\ \mu\ \mu\ \mu\ \mu$
H_a: Full Model: $\mu_B\ \ \mu_F\ \mu_{Soc}\ \ \mu_{Swim}\ \ \mu_T$

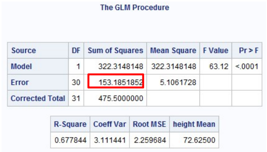


The GLM Procedure

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	333.9444444	83.4861111	15.92	<.0001
Error	27	141.5555556	5.2427984		
Corrected Total	31	475.5000000			

R-Square	Coeff Var	Root MSE	height Mean
0.702302	3.152793	2.289716	72.62500

H_o: Reduced Model: $\mu\ \mu\ \mu\ \mu\ \mu$
H_a: Full Model: $\mu_B\ \mu_O\ \mu_O\ \ \mu_O\ \ \mu_O$



The GLM Procedure

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	322.3148148	322.3148148	63.12	<.0001
Error	30	153.1851852	5.1061728		
Corrected Total	31	475.5000000			

R-Square	Coeff Var	Root MSE	height Mean
0.677844	3.111441	2.259684	72.62500

F-TEST

H_o: *Reduced* Model: $\mu_B\ \mu_O\ \mu_O\ \ \mu_O\ \ \mu_O$

H_a: Full Model: $\mu_B\ \ \mu_F\ \mu_{Soc}\ \ \mu_{Swim}\ \ \mu_T$

$$F = \frac{\frac{Extra\ Sum\ of\ Squares}{ExtraDegress\ of\ Freedom}}{\hat{\sigma}^2{}_{Full}}$$

$$F = \frac{(153.19\ - 141.56)/(30 - 27)}{141.56/27}$$

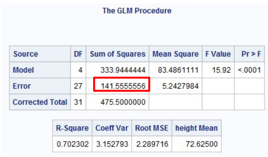
$$F = .74$$

Pvalue = 0.5375

Fail to Reject Ho

There is not sufficient evidence at the alpha = .05 level of significance (p-value = 0.5375) to suggest that the mean heights of non-basketball sports are not equal. Therefore we will proceed as if they are equal.

μ_B μ_F μ_{Soc} μ_{Swim} μ_T

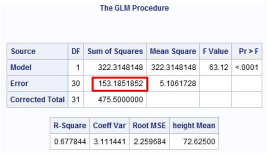


The GLM Procedure

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	333.9444444	83.4861111	15.92	<.0001
Error	27	141.5555556	5.2427984		
Corrected Total	31	475.5000000			

R-Square	Coeff Var	Root MSE	height Mean
0.702302	3.152793	2.289716	72.62500

μ_B μ_O μ_O μ_O μ_O



The GLM Procedure

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	322.3148148	322.3148148	63.12	<.0001
Error	30	153.1851852	5.1061728		
Corrected Total	31	475.5000000			

R-Square	Coeff Var	Root MSE	height Mean
0.677844	3.111441	2.259684	72.62500

F-TEST: Another Look

H_o: Reduced Model: μ_B μ_O μ_O μ_O μ_O

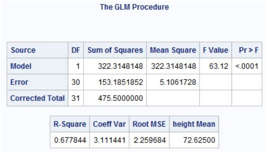
H_a: Full Model: μ_B μ_F μ_{Soc} μ_{Swim} μ_T

Source	DF	SS	MS	F	Pr > F
Model	3	11.63	3.87	.74	0.5375
Error	27	141.56	5.24		
Corrected Total	30	153.19			

Since we are proceeding under the assumption that the mean heights of the other sports (besides basketball) are equal, we can test whether basketball has a mean height different than the other sports by testing:

H_o: $\mu_{Basketball} = \mu_{Others}$

H_a: $\mu_{Basketball} \neq \mu_{Others}$



The GLM Procedure

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	322.3148148	322.3148148	63.12	<.0001
Error	30	153.1851852	5.1061728		
Corrected Total	31	475.5000000			

R-Square	Coeff Var	Root MSE	height Mean
0.677844	3.111441	2.259684	72.62500

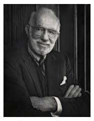
There is strong evidence at the alpha = .05 level of significance (p-value < .0001) that supports the claim that the mean height of basketball players is different than that of the other 4 sports.

Resources

www.itl.nist.gov/div898/handbook/prc/section4/prc433.htm

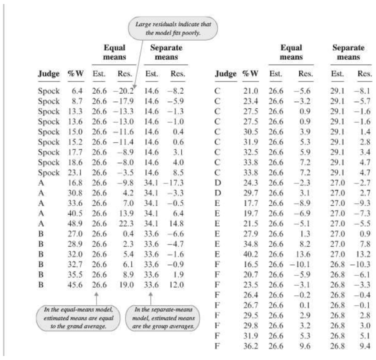
Spock Example

Spock Trial



- 1968: Dr. Ben Spock was accused of conspiracy to violate the Selective Service Act by encouraging young men to resist being drafted into military service for Vietnam.
- Jury Selection: A “venire” of 30 potential jurors is selected at random from a list of 300 names that were previously selected at random from citizens of Boston.
- A jury is then selected NOT at random by the attorneys trying the case.
- For this case, the venire consisted of only one woman, who was let go by the prosecution, thus resulting in an all male jury.
- There was reason to believe that women were more sympathetic to Dr. Spock’s actions due to his popular child rearing books.
- The defense argued that the judge in this case had a history of venires that underrepresented women, which is contrary to the law.
- Let’s see if there is any evidence for this claim!

The Raw Data



Large residuals indicate that the model fits poorly.

		Equal means		Separate means	
Judge	%W	Est.	Res.	Est.	Res.
Spock	6.4	26.6	−20.2	14.6	−8.2
Spock	8.7	26.6	−17.9	14.6	−5.9
Spock	13.3	26.6	−13.3	14.6	−1.3
Spock	13.6	26.6	−13.0	14.6	−1.0
Spock	15.0	26.6	−11.6	14.6	0.4
Spock	15.2	26.6	−11.4	14.6	0.6
Spock	17.7	26.6	−8.9	14.6	3.1
Spock	18.6	26.6	−8.0	14.6	4.0
Spock	23.1	26.6	−3.5	14.6	8.5
A	16.8	26.6	−9.8	34.1	−17.3
A	30.8	26.6	4.2	34.1	−3.3
A	33.6	26.6	7.0	34.1	−0.5
A	40.5	26.6	13.9	34.1	6.4
A	48.9	26.6	22.3	34.1	14.8
B	27.0	26.6	0.4	33.6	−6.6
B	28.9	26.6	2.3	33.6	−4.7
B	32.0	26.6	5.4	33.6	−1.6
B	32.7	26.6	6.1	33.6	−0.9
B	35.5	26.6	8.9	33.6	1.9
B	45.6	26.6	19.0	33.6	12.0

In the equal-means model, estimated means are equal to the grand average.

In the separate-means model, estimated means are the group averages.

		Equal means		Separate means	
Judge	%W	Est.	Res.	Est.	Res.
C	21.0	26.6	−5.6	29.1	−8.1
C	23.4	26.6	−3.2	29.1	−5.7
C	27.5	26.6	0.9	29.1	−1.6
C	27.5	26.6	0.9	29.1	−1.6
C	30.5	26.6	3.9	29.1	1.4
C	31.9	26.6	5.3	29.1	2.8
C	32.5	26.6	5.9	29.1	3.4
C	33.8	26.6	7.2	29.1	4.7
C	33.8	26.6	7.2	29.1	4.7
D	24.3	26.6	−2.3	27.0	−2.7
D	29.7	26.6	3.1	27.0	2.7
E	17.7	26.6	−8.9	27.0	−9.3
E	19.7	26.6	−6.9	27.0	−7.3
E	21.5	26.6	−5.1	27.0	−5.5
E	27.9	26.6	1.3	27.0	0.9
E	34.8	26.6	8.2	27.0	7.8
E	40.2	26.6	13.6	27.0	13.2
F	16.5	26.6	−10.1	26.8	−10.3
F	20.7	26.6	−5.9	26.8	−6.1
F	23.5	26.6	−3.1	26.8	−3.3
F	26.4	26.6	−0.2	26.8	−0.4
F	26.7	26.6	0.1	26.8	−0.1
F	29.5	26.6	2.9	26.8	2.8
F	29.8	26.6	3.2	26.8	3.0
F	31.9	26.6	5.3	26.8	5.1
F	36.2	26.6	9.6	26.8	9.4

Comparing Two Means From Many Groups.

$H_o: \mu_S = \mu_F$

$H_a: \mu_S \neq \mu_F$

Judge	N	Xbar	Sd
Spock	9	14.6	5.04
A	5	34.1	11.94
B	6	33.6	6.58
C	9	29.1	4.59
D	2	27.0	3.81
E	6	27.0	9.01
F	9	26.8	5.97

With 2 groups estimating the pooled SD.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$\mu = \bar{x}\ \boxed{-}\ 13$$

With all 7 groups estimating the pooled SD, bigger 'n' greater **df**! More POWER!!!

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_I - 1)s_I^2}{(n_1 - 1) + (n_2 - 1) + \cdots + (n_I - 1)}.$$

$$\text{Pvalue} = 0.5375$$

$s_p = 6.91$

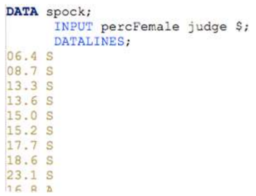
$CV = t_{.025\,\boxed{39}} = \pm 2.02$

$$t = \frac{14.6 - 26.8}{6.91\sqrt{\frac{1}{9} + \frac{1}{9}}} = \frac{-12.2}{3.25} = -3.75$$

P-value = .0006

Reject Ho

Spock Data Steps

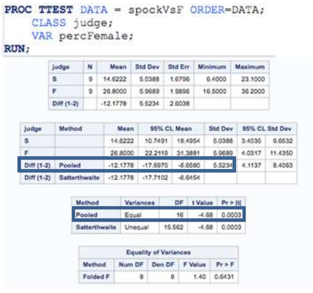


```
DATA spock;
      INPUT percFemale judge $;
      DATALINES;
06.4 S
08.7 S
13.3 S
13.6 S
15.0 S
15.2 S
17.7 S
18.6 S
23.1 S
```

```
DATA spockVsF;
    SET spock;
    if (judge NE 'S') & (judge NE 'F')  THEN DELETE;
RUN;
```

Question: Suppose we wish to test if the "S" judge's venires are different from the "F" judge's.

Two Judge Analysis w/ t-Tools



```
PROC TTEST DATA = spockVsF ORDER=DATA;
    CLASS judge;
    VAR percFemale;
RUN;
```

judge	N	Mean	Std Dev	Std Err	Minimum	Maximum
S	9	14.6222	5.0388	1.6796	6.4000	23.1000
F	9	26.8000	5.9689	1.9896	16.5000	36.2000
Diff (1-2)		-12.1778	5.5234	2.6038		

judge	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
S		14.6222	10.7491	18.4954	5.0388	3.4036	9.6532
F		26.8000	22.2119	31.3881	5.9689	4.0317	11.4350
Diff (1-2)	Pooled	-12.1778	-17.6975	-6.6580	5.5234	4.1137	8.4063
Diff (1-2)	Satterthwaite	-12.1778	-17.7102	-6.6454			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	16	-4.68	0.0003
Satterthwaite	Unequal	15.562	-4.68	0.0003

Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	8	8	1.40	0.6431

Statistical Conclusion: We find that there is substantial evidence that the difference in the mean percentage of females on judge S and judge F venires is not equal to zero.

Estimated Diff = -12.1778

S_p = 5.5234

Pooled Std. Error = 2.6038

t-Statistic = -4.68

Deg. of freedom = 16

Two Judge Analysis w/ Several-Groups

From PROC TTEST:

Estimated Diff = -12.1778

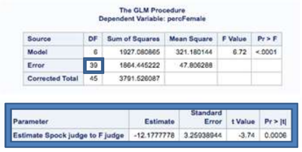
S_p = 5.5234

Pooled Std. Error = 2.6038

t-Statistic = -4.68

Deg. of freedom = 16

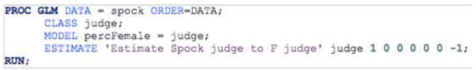
Deg. of freedom = 46 – 7 = 39



The GLM Procedure
Dependent Variable: percFemale

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	1927.080865	321.180144	6.72	<.0001
Error	39	1864.445222	47.806288		
Corrected Total	45	3791.526087			

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Estimate Spock judge to F judge	-12.1777778	3.25038944	-3.74	0.0006



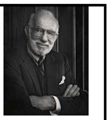
```
PROC GLM DATA = spock ORDER=DATA;
     CLASS judge;
     MODEL percFemale = judge;
     ESTIMATE 'Estimate Spock judge to F judge' judge 1 0 0 0 0 0 -1;
RUN;
```

Two Judge Analysis: Conclusion

Question: Suppose we wish to test if the "S" judge's venires are different from the "F" judge's.	**Answer:** There is evidence that the mean of the two groups is different.

- We can use regular t-Tools or several-group analysis.
- The several-group analysis allows us to use all of the available information → larger degrees of freedom → more power!

Spock Trial QOI 2



The defense argued that the judge in this case had a history of venires that underrepresented women, which is contrary to the law.

- QOI2: Is the percent of women on recent venires of Spock's judge (which we will call S) significantly lower than those of 6 other judges (which we notate A to F)?
- There are two key questions:
  1. Is there evidence that women are underrepresented on S's venires relative to A to F's?
  2. Is there evidence of a difference in women's representation on A to F's venires?
- The question of interest is addressed by 1
- The strength of the result in 1 would be substantially diminished if 2 is true

Spock: The Strategy

Since we found that there was evidence that at least one of the means was different than the others, we will first (Step 1) test to see if there is evidence that the other 6 judges have similar mean female representation in their venires. If there is no evidence their means are different then (Step 2) we have them share a mean (μ_O) and compare Spock's judge's (μ_S) mean with μ_O.

Step 1: Compare Judges A - F

H_o: All "other" means are equal (A, B, C, D, E, F)
H_a: At least 2 "other" means are different (A, B, C, D, E, F)

But … Let's use all the data to estimate the pooled standard deviation!

Reduced Model: $\mu_s\ \mu_o\ \mu_o\ \mu_o\ \mu_o\ \mu_o\ \mu_o$

Full Model: $\mu_s\ \mu_A\ \mu_B\ \mu_C\ \mu_D\ \mu_E\ \mu_F$

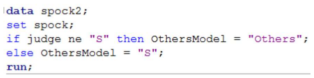
Different Models in SAS

At Least 2 are different (S, A, B, … F)

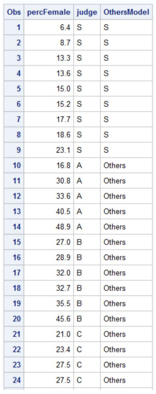
$\mu_s\ \mu_A\ \mu_B\ \mu_C\ \mu_D\ \mu_E\ \mu_F$

Spock is different than the Others

$\mu_s\ \mu_o\ \mu_o\ \mu_o\ \mu_o\ \mu_o\ \mu_o$



```
data spock2;
set spock;
if judge ne "S" then OthersModel = "Others";
else OthersModel = "S";
run;
```

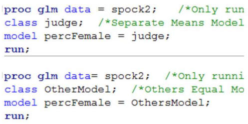


Obs	percFemale	judge	OthersModel
1	6.4	S	S
2	8.7	S	S
3	13.3	S	S
4	13.6	S	S
5	15.0	S	S
6	15.2	S	S
7	17.7	S	S
8	18.6	S	S
9	23.1	S	S
10	16.8	A	Others
11	30.8	A	Others
12	33.6	A	Others
13	40.5	A	Others
14	48.9	A	Others
15	27.0	B	Others
16	28.9	B	Others
17	32.0	B	Others
18	32.7	B	Others
19	35.5	B	Others
20	45.6	B	Others
21	21.0	C	Others
22	23.4	C	Others
23	27.5	C	Others
24	27.5	C	Others

Different Models in SAS

At Least 2 are different (S, A, B, … F)

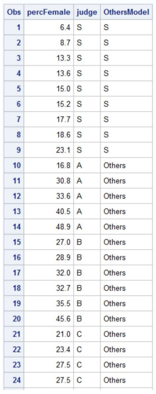
$\mu_s\ \mu_A\ \mu_B\ \mu_C\ \mu_D\ \mu_E\ \mu_F$



```
proc glm data = spock2;   /*Only run
class judge;  /*Separate Means Model
model percFemale = judge;
run;

proc glm data= spock2;  /*Only runni
class OtherModel;  /*Others Equal Mo
model percFemale = OthersModel;
run;
```

Spock is different than the Others

$\mu_s\ \mu_o\ \mu_o\ \mu_o\ \mu_o\ \mu_o\ \mu_o$



Obs	percFemale	judge	OthersModel
1	6.4	S	S
2	8.7	S	S
3	13.3	S	S
4	13.6	S	S
5	15.0	S	S
6	15.2	S	S
7	17.7	S	S
8	18.6	S	S
9	23.1	S	S
10	16.8	A	Others
11	30.8	A	Others
12	33.6	A	Others
13	40.5	A	Others
14	48.9	A	Others
15	27.0	B	Others
16	28.9	B	Others
17	32.0	B	Others
18	32.7	B	Others
19	35.5	B	Others
20	45.6	B	Others
21	21.0	C	Others
22	23.4	C	Others
23	27.5	C	Others
24	27.5	C	Others

Comparing Two Models:
Both are not Equal Means Model

SAS (proc glm) compares models to the equal means model. When you run proc glm, it always makes the "Corrected Total Row" the equal means model. However, we can build our own ANOVA table (BYOA) to compare two models, both of which are not the equal means model.

To do this we will need to identify the "full" model and the "reduced" model. The "full" model will be the model with the most parameters (means) in it while the "reduced model" will have fewer parameters. (Note that the equal means model (with one parameter) is the most reduced model you can have.)

Extra Sum of Squares
Test / BYOA

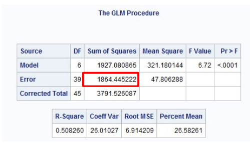
Separate Means Model (Full Model)

Equal Means Model (Reduced Model)

Source	DF	SS	MS	F	Pr > F
Model					
Error					
Corrected Total					

At least 2 are different (Spock, A, B, C … F)

$\mu_s\ \mu_A\ \mu_B\ \mu_C\ \mu_D\ \mu_E\ \mu_F$



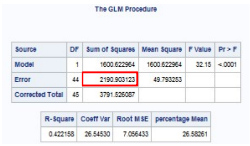
The GLM Procedure

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	1927.080865	321.180144	6.72	<.0001
Error	39	1864.445222	47.806288		
Corrected Total	45	3791.526087			

R-Square	Coeff Var	Root MSE	Percent Mean
0.508260	26.01027	6.914209	26.58261

Spock is different than others

$\mu_s\ \mu_o\ \mu_o\ \mu_o\ \mu_o\ \mu_o\ \mu_o$



The GLM Procedure

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1600.622964	1600.622964	32.15	<.0001
Error	44	2190.903123	49.793253		
Corrected Total	45	3791.526087			

R-Square	Coeff Var	Root MSE	percentage Mean
0.422158	26.54530	7.056433	26.58261

F-TEST: Another Look

H_o: μ_A, μ_B, μ_C …. μ_F are Equal
H_a: At least 2 are different (A,B,C …F)

Reduced : $\mu_s\ \mu_o\ \mu_o\ \mu_o\ \mu_o\ \mu_o\ \mu_o$

Full: $\mu_s\ \mu_A\ \mu_B\ \mu_C\ \mu_D\ \mu_E\ \mu_F$

	Source	DF	SS	MS	F	Pr > F
	Model	5	326.5	65.29	1.37	0.26
Full	Error	39	1864.4	47.81		
Reduced	Corrected Total	44	2190.9			

EXTRA SUMS OF SQUARES F TEST

F-TEST

H_o: $\mu_A - \mu_F$ are Equal

H_a: At least 2 are different (A,B, .. F)

$$F = \frac{\frac{Extra\ Sum\ of\ Squares}{ExtraDegress\ of\ Freedom}}{\hat{\sigma}^2{}_{Full}}$$

$$F = \frac{(2190.9 - 1864.4)/(44 - 39)}{1864.4/39}$$

$$F = 1.37$$

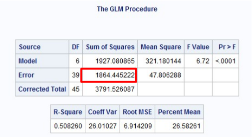
P-value = 0.26

Fail to Reject Ho

There is not sufficient evidence at the alpha = .05 level of significance (p-value = 0.26) to suggest that the means are not equal. Therefore, we will proceed as if they are equal.

H_o: All means are equal (Spock,A,B,C…,F)

H_a: At least 2 are different (Spock,A,B,….F)



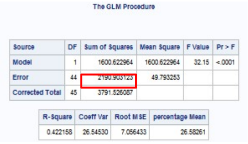
The GLM Procedure

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	1927.080865	321.180144	6.72	<.0001
Error	39	1864.445222	47.806288		
Corrected Total	45	3791.526087			

R-Square	Coeff Var	Root MSE	Percent Mean
0.508260	26.01027	6.914209	26.58261

H_o: Spock is equal to Others

H_a: Spock is diff from Others



The GLM Procedure

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1600.622964	1600.622964	32.15	<.0001
Error	44	2190.903123	49.793253		
Corrected Total	45	3791.526087			

R-Square	Coeff Var	Root MSE	percentage Mean
0.422158	26.54530	7.056433	26.58261

Step 1 Complete!

There is not sufficient evidence to suggest that the mean percent of women on judge's A-F venires are different from one another (p-value = .26 from an ANOVA). Therefore, we will now move on to Step 2 and compare Spock's judge's mean to the single mean that will represent the other judges.

F-TEST: Another Look

H_o: μ_A, μ_B, μ_C …. μ_F are Equal

H_a: At least 2 are different (A,B,C …F)

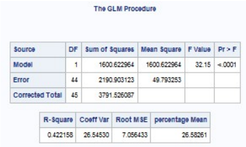
Source	DF	SS	MS	F	Pr > F
Model	5	326.5	65.29	1.37	0.26
Error	39	1864.4	47.81		
Corrected Total	44	2190.9			

Step 2!

Since we are proceeding under the assumption that the mean percentage of women in venires of the non-Spock judges are equal, we can test whether the Spock judge has a mean percentage different than the other judges by testing:

H_o: Mean of Spock is equal to the mean of the others.

H_a: Mean of Spock is different than the mean others.



The GLM Procedure

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1600.622964	1600.622964	32.15	<.0001
Error	44	2190.903123	49.793253		
Corrected Total	45	3791.526087			

R-Square	Coeff Var	Root MSE	percentage Mean
0.422158	26.54530	7.056433	26.58261

There is strong evidence at the alpha = .05 level of significance (p-value < .0001 from an ANOVA) to support the claim that the mean percentage of women in the Spock judge's venires is less than that of the other 6 judges and that there is no evidence that the other 6 judges have different mean percentages of women on their venires (p-value = .26 from an Extra Sum of Squares F Test). Spock's lawyer has evidence for a mistrial.

Part VI

Multiple comparisons and post hoc tests

Chapter 25

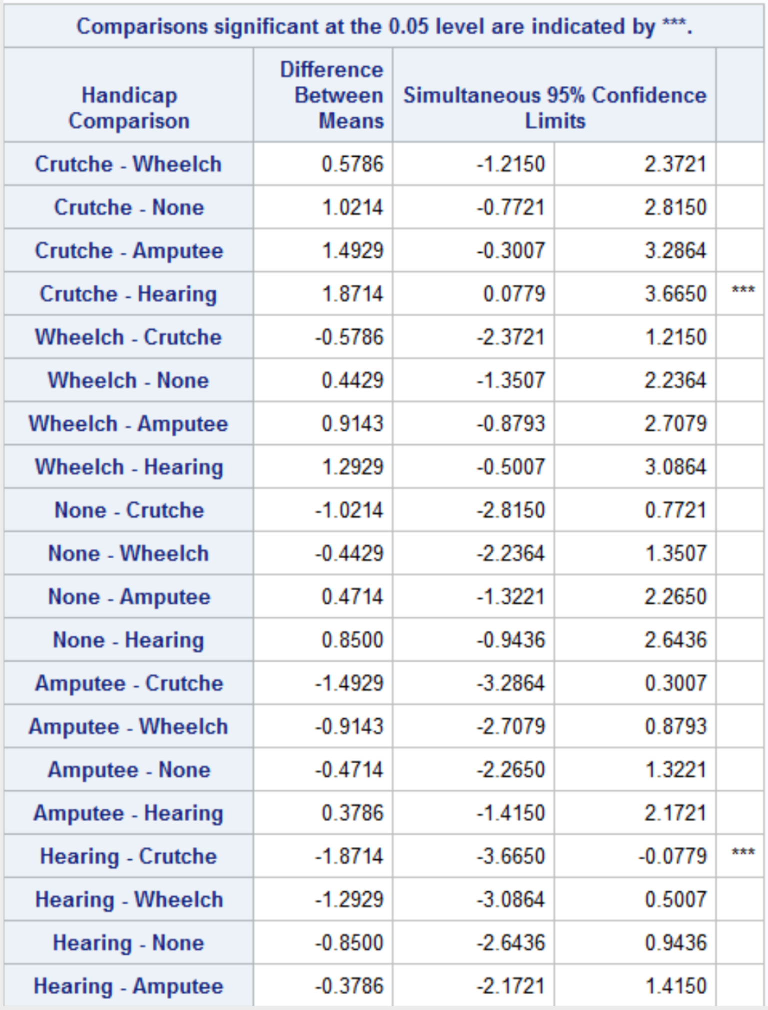
Problem 1: Bonferroni and the Handicap Study

The Bonferroni method was used to construct some simultaneous confidence intervals for $\mu_1 - \mu_2$, $\mu_2 - \mu_5$ and $\mu_3 - \mu_5$, to see whether there are differences in attitude toward the mobility type of handicaps. The Bonferroni CIs were calculated using the following SAS code: Note that lsmeans and means have the same results, because we are dealing with balanced data The result of this code is shown below:

Code 25.1. Bonferroni in SAS

```
proc glm data = handicap;
class handicap;
model score = handicap;
means handicap / hovtest =  bf bon cldiff;
lsmeans handicap / pdiff adjust = bon cl;
run;
```

Figure 25.0.1. Bonferroni Confidence Intervals

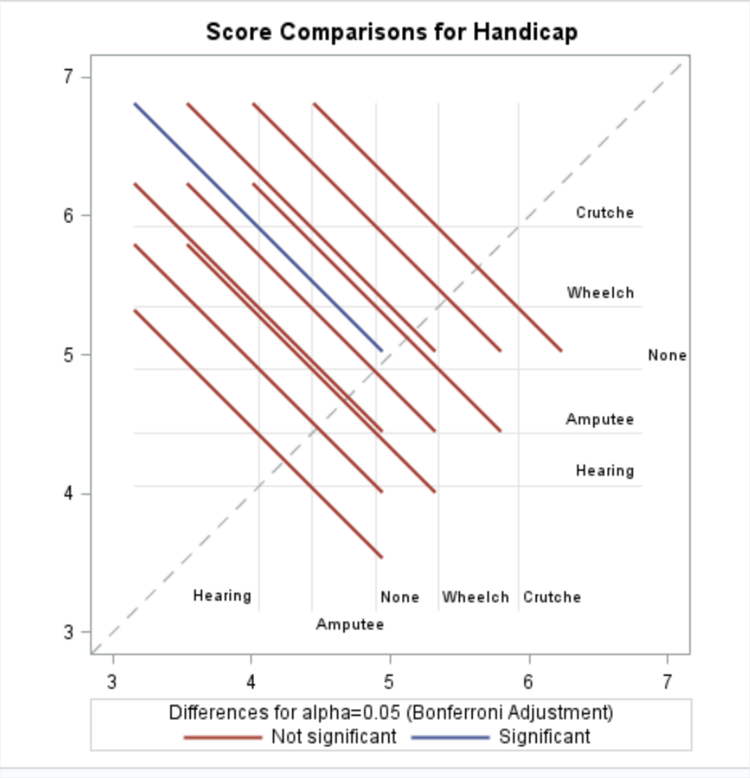


Comparisons significant at the 0.05 level are indicated by ***.

Handicap Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
Crutche - Wheelch	0.5786	-1.2150	2.3721	
Crutche - None	1.0214	-0.7721	2.8150	
Crutche - Amputee	1.4929	-0.3007	3.2864	
Crutche - Hearing	1.8714	0.0779	3.6650	***
Wheelch - Crutche	-0.5786	-2.3721	1.2150	
Wheelch - None	0.4429	-1.3507	2.2364	
Wheelch - Amputee	0.9143	-0.8793	2.7079	
Wheelch - Hearing	1.2929	-0.5007	3.0864	
None - Crutche	-1.0214	-2.8150	0.7721	
None - Wheelch	-0.4429	-2.2364	1.3507	
None - Amputee	0.4714	-1.3221	2.2650	
None - Hearing	0.8500	-0.9436	2.6436	
Amputee - Crutche	-1.4929	-3.2864	0.3007	
Amputee - Wheelch	-0.9143	-2.7079	0.8793	
Amputee - None	-0.4714	-2.2650	1.3221	
Amputee - Hearing	0.3786	-1.4150	2.1721	
Hearing - Crutche	-1.8714	-3.6650	-0.0779	***
Hearing - Wheelch	-1.2929	-3.0864	0.5007	
Hearing - None	-0.8500	-2.6436	0.9436	
Hearing - Amputee	-0.3786	-2.1721	1.4150	

Another nice way to visualize these confidence intervals is like this:

Figure 25.0.2. Diffogram of the Bonferroni Confidence Intervals



As we see from these two figures, the only statistically significant mean difference was the crutches vs the hearing, which means that the attitude towards the different mobility handicaps is the same ($\mu_1 - \mu_2$, $\mu_2 - \mu_5$ and $\mu_3 - \mu_5$ are not different)

Chapter 26

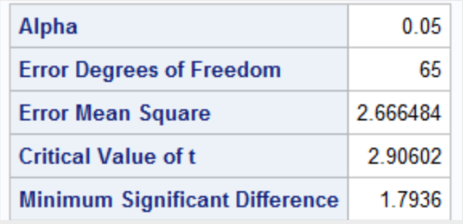
Multiple Comparison and the Handicap Study

To generate all the multiple comparisons, and the half widths, the follwoing SAS code was used: Here we see the

Code 26.1. all the multiple comparisons in SAS

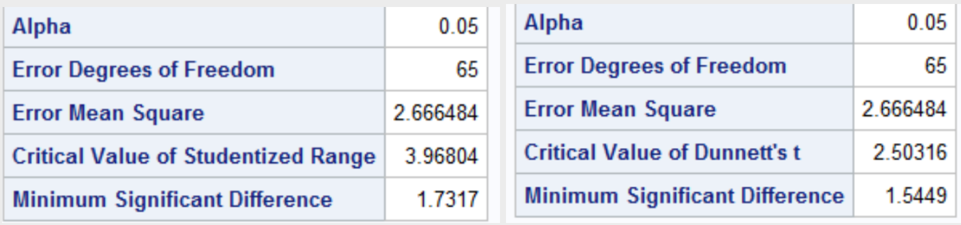
```
proc glm data = handicap;
class handicap;
model score = handicap;
means handicap / tukey bon scheffe LSD Dunnett('None');
run;
```

results of this



Alpha	0.05
Error Degrees of Freedom	65
Error Mean Square	2.666484
Critical Value of t	2.90602
Minimum Significant Difference	1.7936

(a) Bonferroni

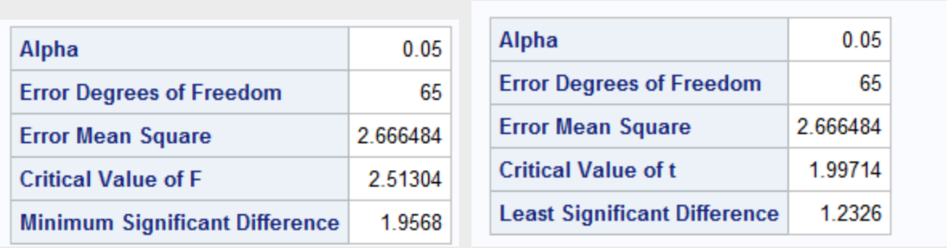


Alpha	0.05
Error Degrees of Freedom	65
Error Mean Square	2.666484
Critical Value of Studentized Range	3.96804
Minimum Significant Difference	1.7317

(b) Tukey

Alpha	0.05
Error Degrees of Freedom	65
Error Mean Square	2.666484
Critical Value of Dunnett's t	2.50316
Minimum Significant Difference	1.5449

(c) Dunnet



Alpha	0.05
Error Degrees of Freedom	65
Error Mean Square	2.666484
Critical Value of F	2.51304
Minimum Significant Difference	1.9568

(d) Scheffe

Alpha	0.05
Error Degrees of Freedom	65
Error Mean Square	2.666484
Critical Value of t	1.99714
Least Significant Difference	1.2326

(e) LSD

Figure 26.0.1. Half widths of different post hoc analyses in SAS

We did the same thing in R, with code and output shown below:

Code 26.2. Multiple comparisons with R

```
prob2 <- case0601
# we make none the first group so that dunnetts test behaves
prob2$Handicap<-factor(prob2$Handicap,levels=c('None', 'Amputee', 'Crutches', 'Hearing'
    , 'Wheelchair'))
aovmodel <- aov(Score ~ Handicap, data=Handi)
# Now we can begin our tests
# Tukey's test
tukey <- glht(aovmodel,linfct=mcp(Handicap="Tukey"))
confint(tukey) #Tukey



Simultaneous Confidence Intervals

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = Score ~ Handicap, data = Handi)

Quantile = 2.8066
95% family-wise confidence level


Linear Hypotheses:
Estimate lwr     upr
Amputee - None == 0         -0.4714  -2.2037  1.2608
Crutches - None == 0         1.0214  -0.7108  2.7537
Hearing - None == 0         -0.8500  -2.5822  0.8822
Wheelchair - None == 0       0.4429  -1.2894  2.1751
Crutches - Amputee == 0      1.4929  -0.2394  3.2251
Hearing - Amputee == 0      -0.3786  -2.1108  1.3537
Wheelchair - Amputee == 0    0.9143  -0.8179  2.6465
Hearing - Crutches == 0     -1.8714  -3.6037 -0.1392
Wheelchair - Crutches == 0 -0.5786  -2.3108  1.1537
Wheelchair - Hearing == 0    1.2929  -0.4394  3.0251

# Calculated by hand
half width = 1.73225

# bonferroni ##
confint(tukey,test=adjusted(type="bonferroni")) # bonferroni, we can just apply the
    bonferroni to whatever
# according to the documentation

Simultaneous Confidence Intervals

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = Score ~ Handicap, data = Handi)

Quantile = 2.8057
95% family-wise confidence level


Linear Hypotheses:
Estimate lwr     upr
Amputee - None == 0         -0.4714  -2.2031  1.2602
Crutches - None == 0         1.0214  -0.7102  2.7531
Hearing - None == 0         -0.8500  -2.5817  0.8817
Wheelchair - None == 0       0.4429  -1.2888  2.1745
Crutches - Amputee == 0      1.4929  -0.2388  3.2245
Hearing - Amputee == 0      -0.3786  -2.1102  1.3531
Wheelchair - Amputee == 0    0.9143  -0.8174  2.6459
Hearing - Crutches == 0     -1.8714  -3.6031 -0.1398
Wheelchair - Crutches == 0 -0.5786  -2.3102  1.1531
```

Chapter 27

Comparing groups: Education study

27.1 Assumptions

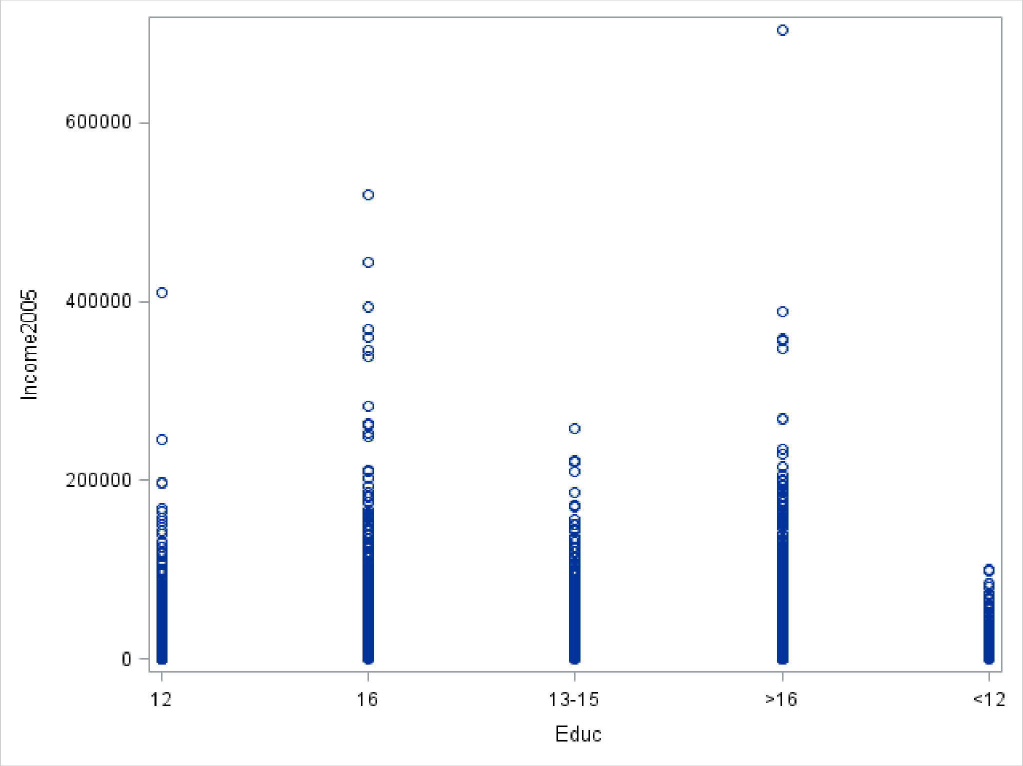
Raw Data Analysis

First, we will look at the raw data. To check if the raw data fits the assumptions, we will first look at a scatter plot. The scatter plot of the raw data was produced by the following bit of SAS code:

```
proc sgplot data=EduData;
scatter x=educ y=Income2005;
run;
```

This results in the following plot:

Figure 27.1.1. Scatter Plot of the Raw Data



Looking at Figure 27.1.1, we see that the raw data is very heavy in between 0 and 20,000 for all categories, but some groups spread further and wider than others, which suggests the variances may not be equal. The heaviness of the lower end of each group may also suggest a lack of normality. We will examine this further with some Box plots. These were produced using the following chunk of SAS code:

```
proc sgplot data=EduData;
vbox Income2005 / category=educ
dataskin=matte
;
xaxis display=(noline noticks);
yaxis display=(noline noticks) grid;
run;
```

This results in the following plot:

Figure 27.1.2. Box Plot of the Raw Data

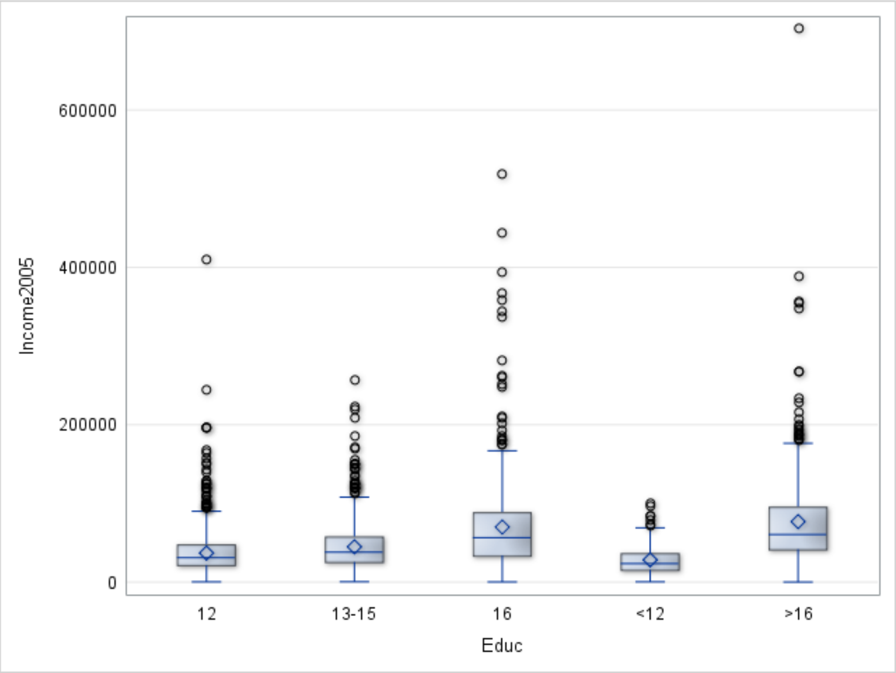


Figure 27.1.2 tells us a lot about our data. We see from the size and shape of the boxes that the variances of our data are by no means homogeneous. Note that there are a lot of outliers while the distribution is heavily weighted towards the bottom, this suggests our data may have departed from normality. We will examine this phenomenaa further using histograms. To produce histograms of the raw data, the following SAS code was used:

```
proc sgpanel data=EduData;
panelby educ / rows=5 layout=rowlattice;
histogram Income2005;
run;
```

This results in the following plot:

Figure 27.1.3. Histogram of the Raw Data

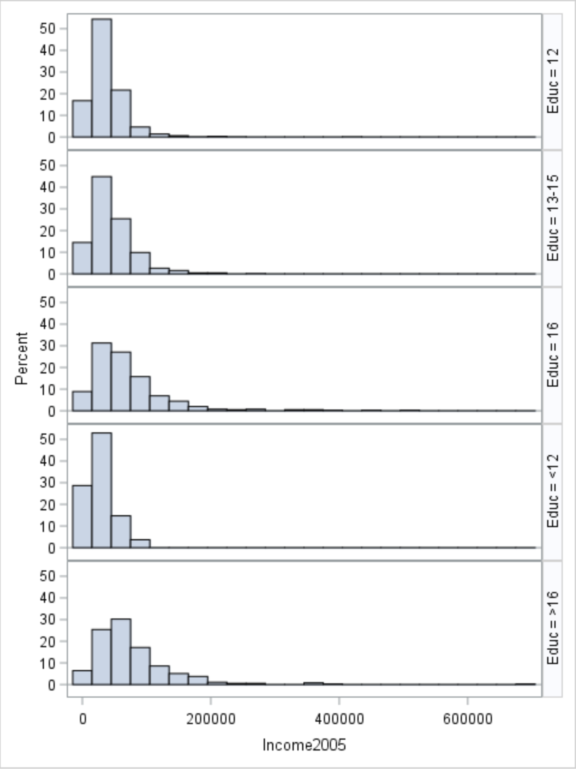
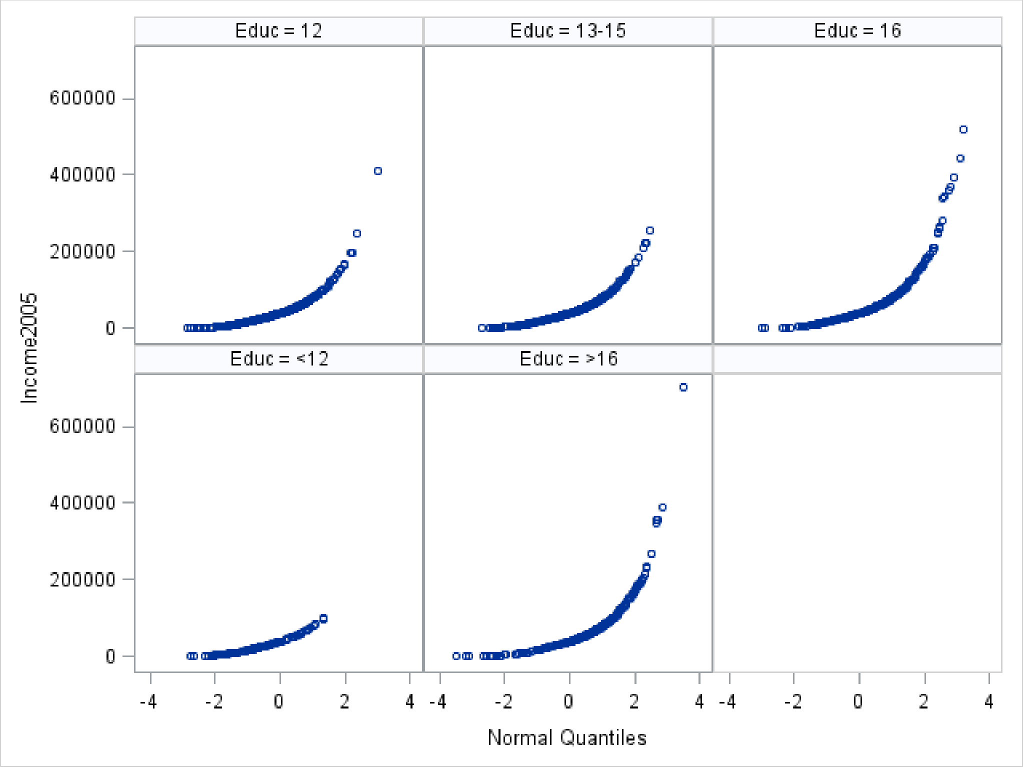


Figure 27.1.3 confirms our suspicions, the variances of the data are likely unequal, but more importantly, the data is clearly skewed to the right. We will confirm this using Q-Q plots. To produce Q-Q plots of the raw data, the following SAS code was used:

```
/* Normal = blom produces normal quantiles from the data */
/* To find out more, look at the SAS documentation!*/
proc rank data=EduData normal=blom out=EduQuant;
var Income2005;
/* Here we produce the normal quantiles!*/
ranks Edu_Quant;
run;
proc sgpanel data=EduQuant;
panelby educ;
scatter x=Edu_Quant y=Income2005 ;
colaxis label="Normal Quantiles";
run;
```

This results in the following plot:

Figure 27.1.4. Q-Q Plot of the Raw Data



The Q-Q plots in Figure 27.1.4 tell us what we already know: The raw data is not normal, and does not have equal variances. The ANOVA test is not super robust to highly skewed, long tailed data, and it relies entirely on equal variances, so we absolutely cannot use the raw data

Transformed Data Analysis

Now we will perform a log transformation on the data and see if that helps it meet our assumptions better. To do a log transformation, we will employ the following SAS code:

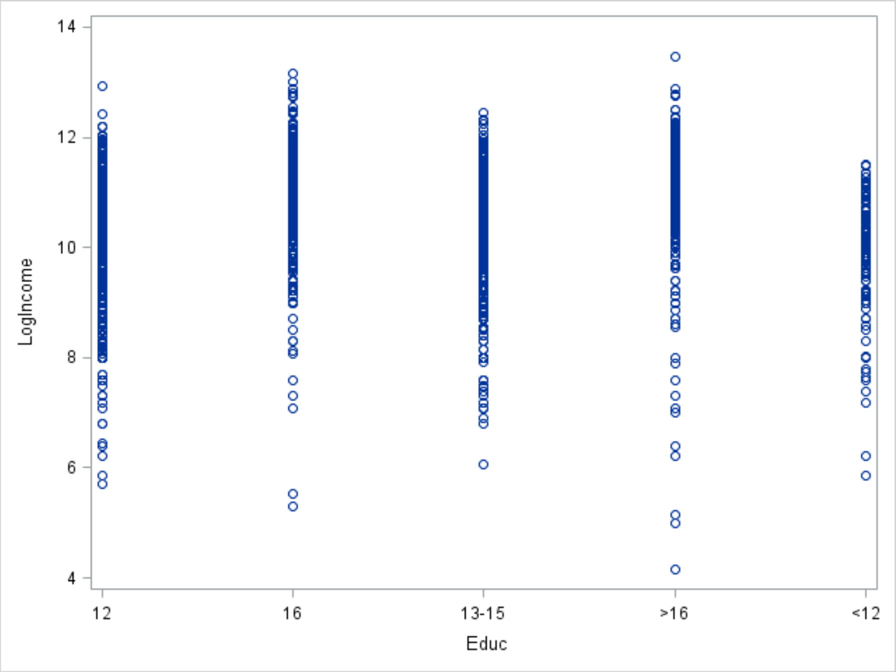
```
data LogEduData;
set EduData;
LogIncome=log(Income2005);
run;
```

We will begin our analysis of the transformed data with a scatter plot, produced with the following SAS code:

```
proc sgplot data=LogEduData;
scatter x=educ y=LogIncome;
run;
```

This results in the following plot:

Figure 27.1.5. Scatter Plot of the Log-Transformed Data



As we can see in Figure 27.1.5, the groups have a much more similar size, suggesting similar variances, and the heavy part of the scatter plot is closer to the center, in between the outliers, which tells us the log transformation may have done a good deal towards normalizing our data. We can examine this further using Box plots. To produce Box plots of the transformed data, the following SAS code was used:

```
proc sgplot data=LogEduData;
vbox LogIncome / category=educ
dataskin=matte
;
xaxis display=(noline noticks);
yaxis display=(noline noticks ) grid;
run;
```

This gives us the following plot:

Figure 27.1.6. Box Plot of the Log-Transformed Data

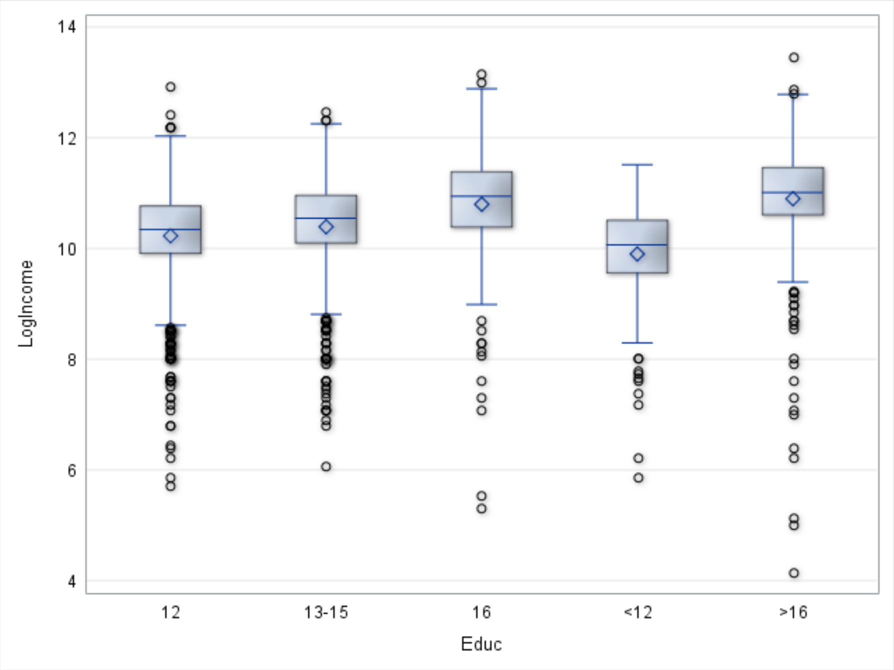
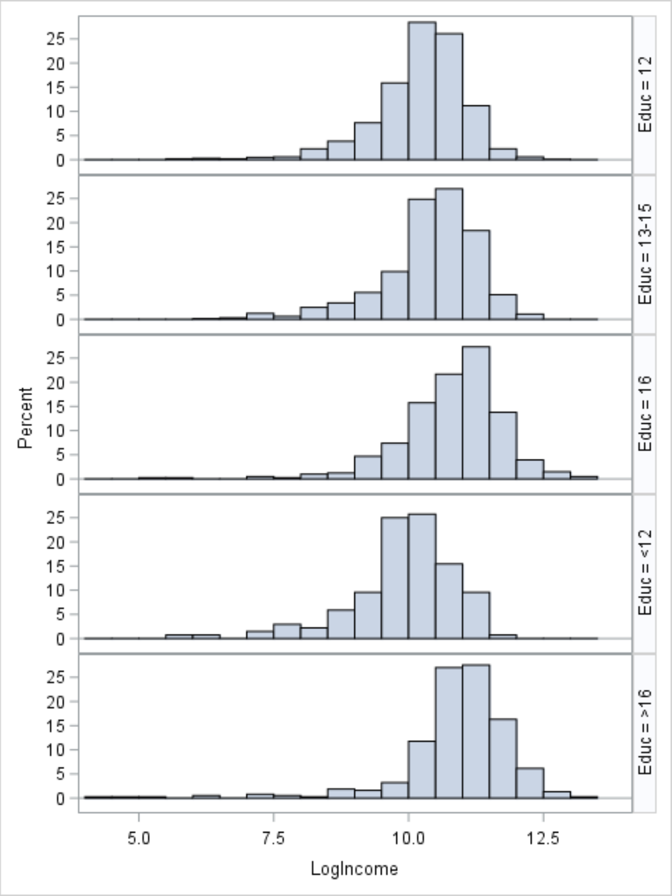


Figure 27.1.6 gives us some useful information about our data. We see the boxes and whiskers are of similar size, which tells us the variances are likely homogeneous. Furthermore, the medians and means are near each other, and the boxes are near the center of the distribution, which suggests that the data may be normal. We will examine these two phenomena further with histograms. To produce histograms of the log-transformed data, the following SAS code was used:

```
proc sgpanel data=LogEduData;
panelby educ / rows=5 layout=rowlattice;
histogram LogIncome;
run;
```

This results in the following plot:

Figure 27.1.7. Histogram of the Log-Transformed Data

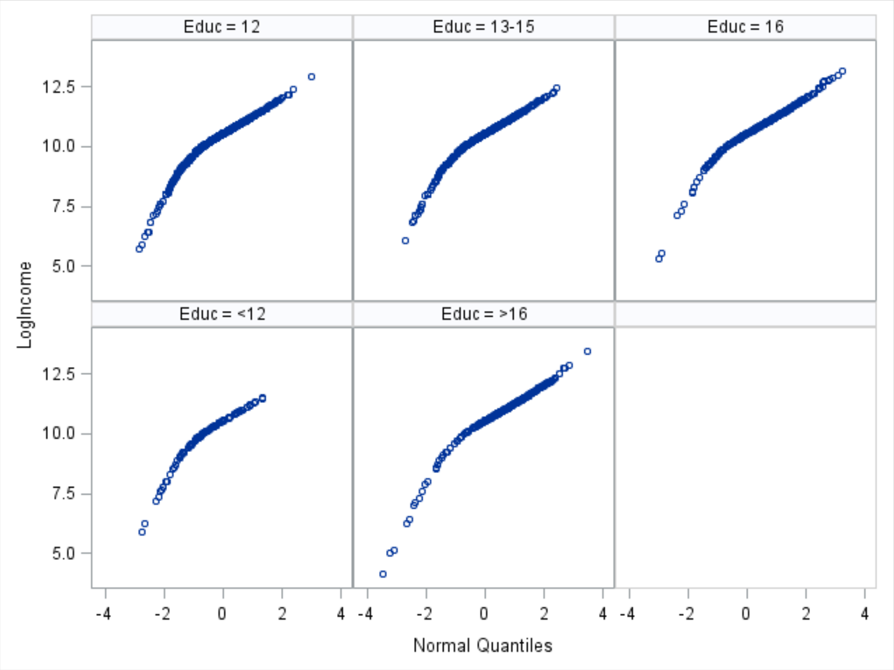


From the spread of the histograms in Figure 27.1.7, we see two things. First, the similar width of the histograms confirms that variances are roughly equal. Second, the shape of the histograms, and their location near the center suggests that the data is very nearly normal. We will further examine the normality of the data using Q-Q plots. To produce the Q-Q plots of the transformed data, the following SAS code was used:

```
proc rank data=LogEduData normal=blom out= LogEduQuant;
var LogIncome;
ranks LogEduQuant;
run;
proc sgpanel data=LogEduQuant;
panelby educ;
scatter x=LogEduQuant y=LogIncome ;
colaxis label="Normal Quantiles";
run;
```

This results in the following plot:

Figure 27.1.8. Q-Q Plot of the Log-Transformed Data



Examining the previous figure, we see a confirmation of our beliefs: The log-transformed data, when plotted against normal quantiles, is fairly normal. This means, with the log transformed data, we can reasonably assume normality and homogeneity of variances. We have fulfilled the assumptions of the ANOVA test and now we are ready to go!

Chapter 28

selection and execution

First, we run an f test to see if any of the means are different!

28.1 ANOVA

We will now perform a complete analysis of our data, using Pure ANOVA.

Problem Statement

We would like to determine whether or not at least one of the five population distributions (corresponding to different years of education) is different from the rest.

Assumptions

As seen in Section ??, the raw data does not meet the assumption of normality nor of homogeneity of variance. However, in Section 27.1, we proved that after a log transformation, the data does meet both of these assumptions. The ANOVA test is fairly robust to the slight departure from normality presented by the log transformed data, and the variances are equal. The data is clearly independent, so that assumption is met. Therefore, all assumptions of ANOVA are met by the log transformed data.

Hypothesis Definition

In this problem, our Null (Reduced Model) Hypothesis, H_0, is that all the groups have the same distribution and our Alternative (Full Model) Hypothesis, H_1 is that the distributions are different. Mathematically, that is written as:

$$H_0 : median_{grand} \quad median_{grand} \quad median_{grand} \quad median_{grand} \quad median_{grand} \tag{28.1.1}$$

$$H_1 : median_{<12} \quad median_{12} \quad median_{13-15} \quad median_{16} \quad median_{>16} \tag{28.1.2}$$

We will consider our confidence level, α to be 0.05

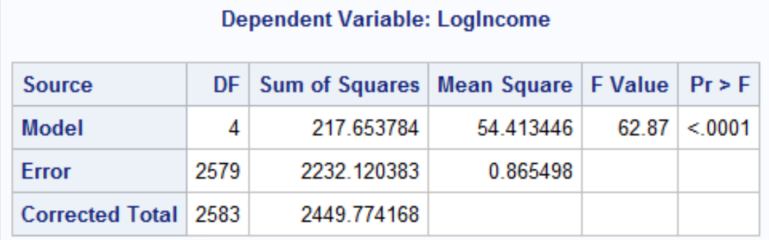
F Statistic

To conduct this hypothesis test, the following SAS code was used:

```
proc glm data = LogEduData;
class educ;
model LogIncome = educ;
run;
```

This results in the following ANOVA Output:

Figure 28.1.1. ANOVA Table



Dependent Variable: LogIncome

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	217.653784	54.413446	62.87	<.0001
Error	2579	2232.120383	0.865498		
Corrected Total	2583	2449.774168			

Figure 28.1.1 tells us what our F statistic is. We see that

$$F = 62.87 \tag{28.1.3}$$

P-value

Figure 28.1.1 also tells us our p-value. In this case,

$$p < .0001 \tag{28.1.4}$$

Hypothesis Assessment

In this scenario, we have that $p < .0001 < \alpha = .05$ and therefore we reject the null hypothesis.

Conclusion

There is substantial evidence ($p < 0.0001$) that at least one of the distributions is different from the others.

28.2 Tukey's test

We want to compare all of the group means to see if they are different, so we do tukey's test! we do this with the following SAS code: With this we see that aside from the college and graduate school educations, they are all different. A confidence interval for these differences, the % change of the medians, is calculated by raising e to the confidence interval, and subtracting one from that and multiplying by 100. These are shown in the following figure:

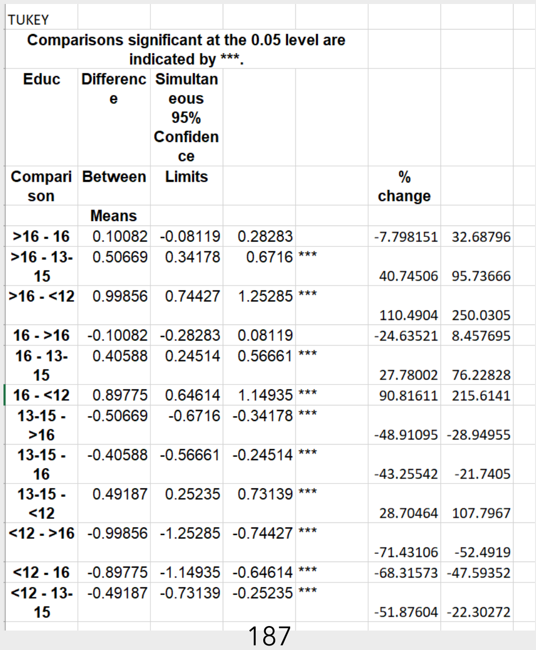
Code 28.1. Tukeys test in SAS and R

```
proc glm data = LogEduData;
class educ;
model LogIncome = educ;
lsmeans LogIncome / pdiff = ALL adjust=tukey cl;
run;
```

and the following R code (and output)

```
edudata <- read.csv(file='c:/Users/david/Desktop/MSDS/MSDS6371/Homework/Week6/Data/ex0525.csv', header=TRUE, sep = ",")
edudata$logincome <- log(edudata$Income2005)
prob3 <- edudata
aovmodel2 <- aov(logincome~Educ,data =prob3)
tukkey <- glht(aovmodel2,linfct=mcp(Educ="Tukey"))
summary(tukkey)

Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = logincome ~ Educ, data = prob3)

Linear Hypotheses:
Estimate Std. Error t value Pr(>|t|)
<12 - <<12 == 0   -0.32787    0.08493   -3.861   0.00101 **
>16 - <<12 == 0    0.67069    0.05624   11.926   < 0.001 ***
13-15 - <<12 == 0  0.16400    0.04674    3.509   0.00389 **
16 - <<12 == 0     0.56987    0.05459   10.439   < 0.001 ***
>16 - <12 == 0     0.99856    0.09316   10.719   < 0.001 ***
13-15 - <12 == 0   0.49187    0.08775    5.606   < 0.001 ***
16 - <12 == 0      0.89775    0.09217    9.740   < 0.001 ***
13-15 - >16 == 0  -0.50669    0.06041   -8.387   < 0.001 ***
16 - >16 == 0     -0.10082    0.06668   -1.512   0.54057
16 - 13-15 == 0    0.40588    0.05888    6.893   < 0.001 ***
---
```

Figure 28.2.1. Tukey CIs on percent increase in the median



TUKEY

Comparisons significant at the 0.05 level are indicated by *.**

Educ Comparison	**Difference Between Means**	**Simultaneous 95% Confidence Limits**			**% change**	
>16 - 16	0.10082	-0.08119	0.28283		-7.798151	32.68796
>16 - 13-15	0.50669	0.34178	0.6716	***	40.74506	95.73666
>16 - <12	0.99856	0.74427	1.25285	***	110.4904	250.0305
16 - >16	-0.10082	-0.28283	0.08119		-24.63521	8.457695
16 - 13-15	0.40588	0.24514	0.56661	***	27.78002	76.22828
16 - <12	0.89775	0.64614	1.14935	***	90.81611	215.6141
13-15 - >16	-0.50669	-0.6716	-0.34178	***	-48.91095	-28.94955
13-15 - 16	-0.40588	-0.56661	-0.24514	***	-43.25542	-21.7405
13-15 - <12	0.49187	0.25235	0.73139	***	28.70464	107.7967
<12 - >16	-0.99856	-1.25285	-0.74427	***	-71.43106	-52.4919
<12 - 16	-0.89775	-1.14935	-0.64614	***	-68.31573	-47.59352
<12 - 13-15	-0.49187	-0.73139	-0.25235	***	-51.87604	-22.30272

Dunnett's Test

To compare to a control, dunnets test is the best! We do this with the following SAS code: lets look at the SAS

Code 28.2. DUnnett's test

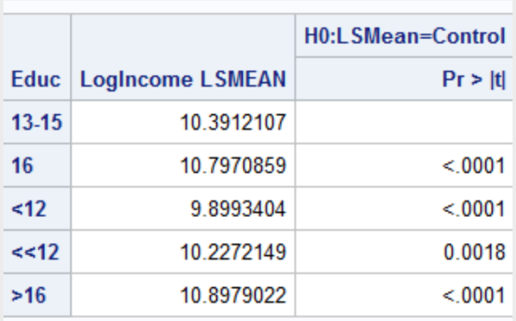
```
proc glm data = LogEduData;
class educ;
model LogIncome = educ;
lsmeans LogIncome / pdiff = ALL adjust=dunnett cl;
run;
```

and the following R code (and output!).

```
summary(dunnbett) #Dunnett

Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Dunnett Contrasts


Fit: aov(formula = logincome ~ Educ, data = prob3)

Linear Hypotheses:
Estimate Std. Error t value Pr(>|t|)
<12 - <<12 == 0    -0.32787     0.08493   -3.861 0.000461 ***
>16 - <<12 == 0     0.67069     0.05624   11.926  < 1e-04 ***
13-15 - <<12 == 0   0.16400     0.04674    3.509 0.001818 **
16 - <<12 == 0      0.56987     0.05459   10.439  < 1e-04 ***
---
```

output too!

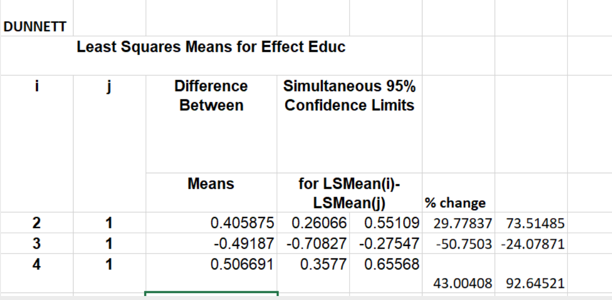
Figure 28.2.2. SAS p values



Educ	LogIncome LSMEAN	H0:LSMean=Control
		Pr > |t|
13-15	10.3912107	
16	10.7970859	<.0001
<12	9.8993404	<.0001
<<12	10.2272149	0.0018
>16	10.8979022	<.0001

We see that all of the groups are different from the control. We can calculate confidence intervals on how much percent different by raising e to the power of the CI, and then subtracting one and multiplying by 100, as seen in the next figure

Figure 28.2.3. Dunnett CIs on percent increase in the median



DUNNETT

Least Squares Means for Effect Educ

i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)		% change	
2	1	0.405875	0.26066	0.55109	29.77837	73.51485
3	1	-0.49187	-0.70827	-0.27547	-50.7503	-24.07871
4	1	0.506691	0.3577	0.65568	43.00408	92.64521

Chapter 29

Unit 6 lecture slides

lol

UNIT 6 Live Session

Contrasts

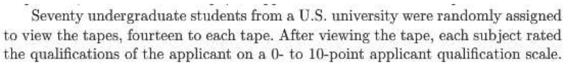
Multiple Comparison

Overview

- ANOVA provides an F-test for equality of several means
- The main weaknesses are
  - It doesn't tell us **which** means are different
  - It doesn't account for any **structure** in the groups

  (Example: Is the average treatment effect across 3 levels of treatments different from the placebo?)
- The downside to this more refined analysis is that we need to control for the number of comparisons we end up making

Example:
Handicap & Capability Study

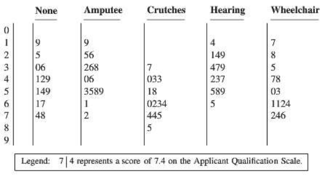


Seventy undergraduate students from a U.S. university were randomly assigned to view the tapes, fourteen to each tape. After viewing the tape, each subject rated the qualifications of the applicant on a 0- to 10-point applicant qualification scale.

- **Goal:** How do physical handicaps affect perception of employment qualification?
- (Cesare, Tannenbaum, and Dalessio "Interviewers' decisions related to applicant handicap type and rater empathy" (1990) *Human Performance)*
- The researchers prepared 5 video taped job interviews with same actors
- The tapes differed only in the handicap of the applicant:
  - No handicap (This is the control group)
  - One leg amputated
  - Crutches
  - Hearing Impaired
  - Wheelchair
- 14 students were randomly assigned to each tape to rate applicants: 0-10 pts (70 students total.)

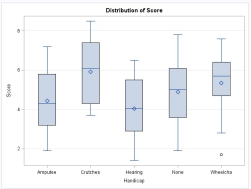
Example:
Handicap & Capability Study

- Do subjects systematically evaluate qualifications differently according to handicap?
- If so, which handicaps are evaluated differently?



	None	Amputee	Crutches	Hearing	Wheelchair
0					
1	9	9		4	7
2	5	56		149	8
3	06	268	7	479	5
4	129	06	033	237	78
5	149	3589	18	589	03
6	17	1	0234	5	1124
7	48	2	445		246
8			5		
9					

Legend: 7 | 4 represents a score of 7.4 on the Applicant Qualification Scale.



Distribution of Score

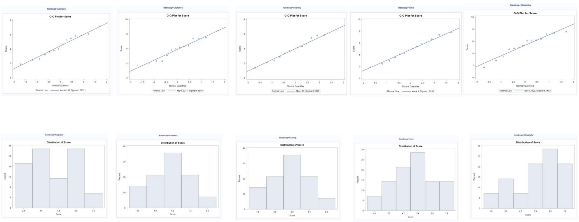
Is There Any Difference at All?

- We should begin any analysis involving several groups by using the ANOVA framework
- If there isn't any (statistically) significant difference in the population means, then there is no reason to address more refined questions
- The tapes differed only in the handicap of the applicant:
  - No handicap (This is the control group.) (μ_{None})
  - One leg amputated (μ_{Amp})
  - Crutches (μ_{Crutch})
  - Hearing Impaired (μ_{Hear})
  - Wheelchair (μ_{Wheel})

ANOVA: $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

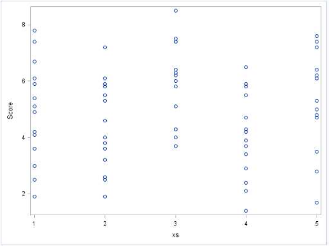
$H_A: \mu_j \neq \mu_k$ for some j, k

Handicap & Capability Study: Normality Assumption



There is NO visual evidence to suggest that the data are not normally distributed. We will proceed with the assumption of normally distributed groups.

Handicap & Capability Study: Equal Variances Assumption

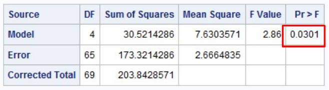


There is NO evidence to suggest variances are unequal.

Handicap & Capability Study: ANOVA results

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 \ (\mu)$

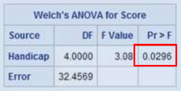
$H_A: \mu_j \neq \mu_k$ for some j, k



Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	30.5214286	7.6303571	2.86	0.0301
Error	65	173.3214286	2.6664835		
Corrected Total	69	203.8428571			

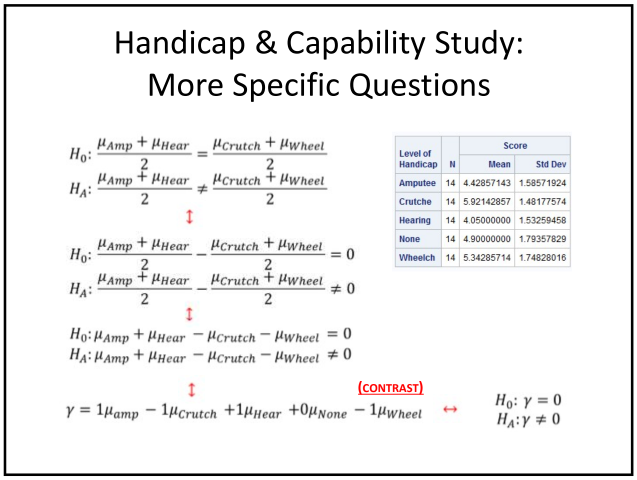
There is evidence to support the claim that at least two population means are different from each other (p-value of 0.0301 from a 1-way ANOVA).

Notice that since there is virtually no evidence of a difference in standard deviations, Welch's test is almost identical to the pure F ANOVA.



Welch's ANOVA for Score

Source	DF	F Value	Pr > F
Handicap	4.0000	3.08	0.0296
Error	32.4569		



Handicap & Capability Study: More Specific Questions

$$H_0: \frac{\mu_{Amp} + \mu_{Hear}}{2} = \frac{\mu_{Crutch} + \mu_{Wheel}}{2}$$
$$H_A: \frac{\mu_{Amp} + \mu_{Hear}}{2} \neq \frac{\mu_{Crutch} + \mu_{Wheel}}{2}$$

$$\updownarrow$$

$$H_0: \frac{\mu_{Amp} + \mu_{Hear}}{2} - \frac{\mu_{Crutch} + \mu_{Wheel}}{2} = 0$$
$$H_A: \frac{\mu_{Amp} + \mu_{Hear}}{2} - \frac{\mu_{Crutch} + \mu_{Wheel}}{2} \neq 0$$

$$\updownarrow$$

$$H_0: \mu_{Amp} + \mu_{Hear} - \mu_{Crutch} - \mu_{Wheel} = 0$$
$$H_A: \mu_{Amp} + \mu_{Hear} - \mu_{Crutch} - \mu_{Wheel} \neq 0$$

$$\updownarrow$$

(CONTRAST)

$$\gamma = 1\mu_{amp} - 1\mu_{Crutch} + 1\mu_{Hear} + 0\mu_{None} - 1\mu_{Wheel} \quad \leftrightarrow \quad \begin{matrix} H_0: \gamma = 0 \\ H_A: \gamma \neq 0 \end{matrix}$$

Level of Handicap	N	Score	
		Mean	Std Dev
Amputee	14	4.42857143	1.58571924
Crutche	14	5.92142857	1.48177574
Hearing	14	4.05000000	1.53259458
None	14	4.90000000	1.79357829
Wheelch	14	5.34285714	1.74828016

Linear Combinations & Contrasts

$$\gamma = C_1\mu_1 + C_2\mu_2 + \cdots + C_I\mu_I$$ (Constraint: $C_1 + C_2 + \cdots + C_I = 0$)

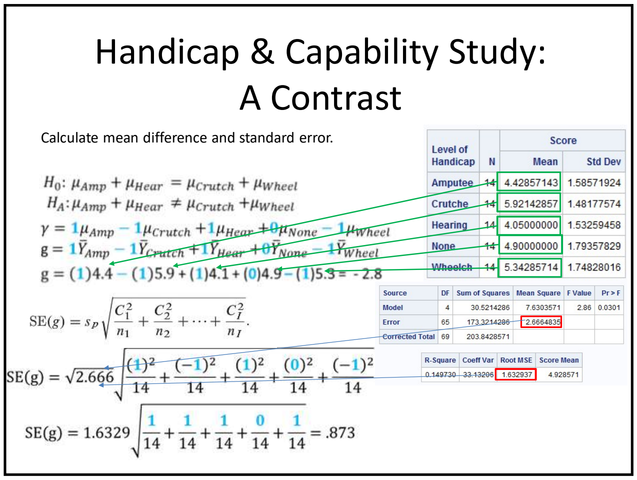
$$g = C_1\overline{Y}_1 + C_2\overline{Y}_2 + \cdots + C_I\overline{Y}_I.$$

$$\mathrm{SE}(g) = s_p\sqrt{\frac{C_1^2}{n_1} + \frac{C_2^2}{n_2} + \cdots + \frac{C_I^2}{n_I}}.$$ (this requires independence)

Example: $\gamma = 1\mu_{Amp} - 1\mu_{Crutch} + 1\mu_{Hear} + 0\mu_{None} - 1\mu_{Wheel}$

The test statistic t:

- $t = \frac{g-\gamma}{\mathrm{SE(g)}}$
- The test statistic has an approximate t-distribution w/ df = $n - I$
  - In this case, $n - I$ =#data points - #groups =70-5= 65



Handicap & Capability Study: A Contrast

Calculate mean difference and standard error.

$$H_0: \mu_{Amp} + \mu_{Hear} = \mu_{Crutch} + \mu_{Wheel}$$
$$H_A: \mu_{Amp} + \mu_{Hear} \neq \mu_{Crutch} + \mu_{Wheel}$$
$$\gamma = 1\mu_{Amp} - 1\mu_{Crutch} + 1\mu_{Hear} + 0\mu_{None} - 1\mu_{Wheel}$$
$$g = 1\bar{Y}_{Amp} - 1\bar{Y}_{Crutch} + 1\bar{Y}_{Hear} + 0\bar{Y}_{None} - 1\bar{Y}_{Wheel}$$
$$g = (1)4.4 - (1)5.9 + (1)4.1 + (0)4.9 - (1)5.3 = -2.8$$

$$\mathrm{SE}(g) = s_p\sqrt{\frac{C_1^2}{n_1} + \frac{C_2^2}{n_2} + \cdots + \frac{C_I^2}{n_I}}.$$

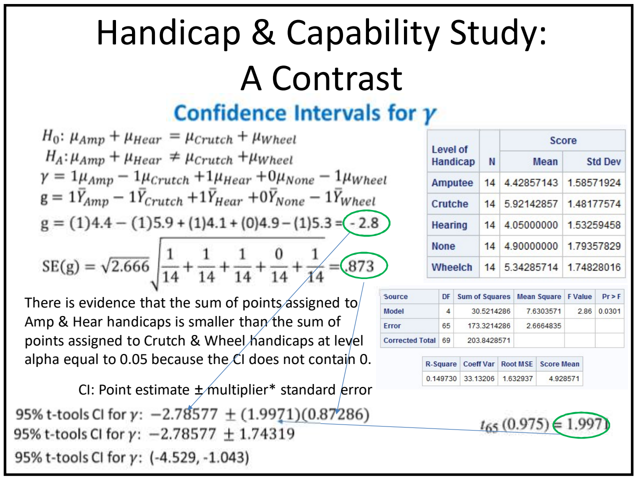
$$\mathrm{SE(g)} = \sqrt{2.666}\sqrt{\frac{(1)^2}{14} + \frac{(-1)^2}{14} + \frac{(1)^2}{14} + \frac{(0)^2}{14} + \frac{(-1)^2}{14}}$$

$$\mathrm{SE(g)} = 1.6329\sqrt{\frac{1}{14} + \frac{1}{14} + \frac{1}{14} + \frac{0}{14} + \frac{1}{14}} = .873$$

Level of Handicap	N	Score	
		Mean	Std Dev
Amputee	14	4.42857143	1.58571924
Crutche	14	5.92142857	1.48177574
Hearing	14	4.05000000	1.53259458
None	14	4.90000000	1.79357829
Wheelch	14	5.34285714	1.74828016

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	30.5214286	7.6303571	2.86	0.0301
Error	65	173.3214286	2.6664835		
Corrected Total	69	203.8428571			

R-Square	Coeff Var	Root MSE	Score Mean
0.149730	33.13206	1.632937	4.928571



Handicap & Capability Study: A Contrast

Confidence Intervals for γ

$$H_0: \mu_{Amp} + \mu_{Hear} = \mu_{Crutch} + \mu_{Wheel}$$
$$H_A: \mu_{Amp} + \mu_{Hear} \neq \mu_{Crutch} + \mu_{Wheel}$$
$$\gamma = 1\mu_{Amp} - 1\mu_{Crutch} + 1\mu_{Hear} + 0\mu_{None} - 1\mu_{Wheel}$$
$$g = 1\bar{Y}_{Amp} - 1\bar{Y}_{Crutch} + 1\bar{Y}_{Hear} + 0\bar{Y}_{None} - 1\bar{Y}_{Wheel}$$
$$g = (1)4.4 - (1)5.9 + (1)4.1 + (0)4.9 - (1)5.3 = -2.8$$

$$\mathrm{SE(g)} = \sqrt{2.666}\sqrt{\frac{1}{14} + \frac{1}{14} + \frac{1}{14} + \frac{0}{14} + \frac{1}{14}} = .873$$

There is evidence that the sum of points assigned to Amp & Hear handicaps is smaller than the sum of points assigned to Crutch & Wheel handicaps at level alpha equal to 0.05 because the CI does not contain 0.

CI: Point estimate ± multiplier* standard error

95% t-tools CI for γ: $-2.78577 \pm (1.9971)(0.87286)$
95% t-tools CI for γ: -2.78577 ± 1.74319
95% t-tools CI for γ: (-4.529, -1.043)

$t_{65}(0.975) = 1.9971$

Level of Handicap	N	Score	
		Mean	Std Dev
Amputee	14	4.42857143	1.58571924
Crutche	14	5.92142857	1.48177574
Hearing	14	4.05000000	1.53259458
None	14	4.90000000	1.79357829
Wheelch	14	5.34285714	1.74828016

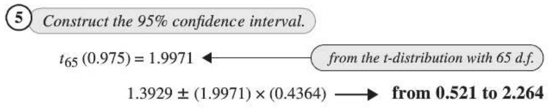
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	30.5214286	7.6303571	2.86	0.0301
Error	65	173.3214286	2.6664835		
Corrected Total	69	203.8428571			

R-Square	Coeff Var	Root MSE	Score Mean
0.149730	33.13206	1.632937	4.928571

Chapter 6: Compare with book!

$$Ho: \frac{\mu_{Amp}+\mu_{Hear}}{2} = \frac{\mu_{Crutch}+\mu_{Wheel}}{2}$$

$$Ha: \frac{\mu_{Amp}+\mu_{Hear}}{2} \neq \frac{\mu_{Crutch}+\mu_{Wheel}}{2}$$


⑤ *Construct the 95% confidence interval.*

$t_{65}(0.975) = 1.9971 \longleftarrow$ *from the t-distribution with 65 d.f.*

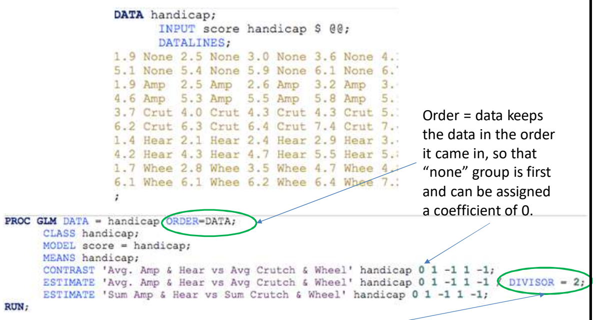
$1.3929 \pm (1.9971) \times (0.4364) \longrightarrow$ **from 0.521 to 2.264**

Note the sign switch and division by 2 of the coefficients.

$$\gamma = -0.5\mu_{Amp} + 0.5\mu_{Crutch} - 0.5\mu_{Hear} + 0\mu_{None} + 0.5\mu_{Wheel}$$

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
Compare Ave Amp and Hearing to Avg Crutch and Wheel	1	27.16071429	27.16071429	10.19	0.0022

Handicap & Capability Study: In SAS

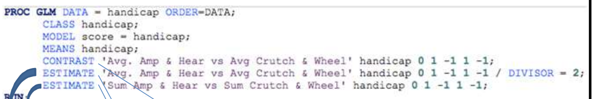


```
DATA handicap;
     INPUT score handicap $ @@;
     DATALINES;
1.9 None 2.5 None 3.0 None 3.6 None 4.
5.1 None 5.4 None 5.9 None 6.1 None 6.
1.9 Amp  2.5 Amp  2.6 Amp  3.2 Amp  3.
4.6 Amp  5.3 Amp  5.5 Amp  5.8 Amp  5.
3.7 Crut 4.0 Crut 4.3 Crut 4.3 Crut 5.
6.2 Crut 6.3 Crut 6.4 Crut 7.4 Crut 7.
1.4 Hear 2.1 Hear 2.4 Hear 2.9 Hear 3.
4.2 Hear 4.3 Hear 4.7 Hear 5.5 Hear 5.
1.7 Whee 2.8 Whee 3.5 Whee 4.7 Whee 4.
6.1 Whee 6.1 Whee 6.2 Whee 6.4 Whee 7.
;
PROC GLM DATA = handicap ORDER=DATA;
     CLASS handicap;
     MODEL score = handicap;
     MEANS handicap;
     CONTRAST 'Avg. Amp & Hear vs Avg Crutch & Wheel' handicap 0 1 -1 1 -1;
     ESTIMATE 'Avg. Amp & Hear vs Avg Crutch & Wheel' handicap 0 1 -1 1 -1 / DIVISOR = 2;
     ESTIMATE 'Sum Amp & Hear vs Sum Crutch & Wheel' handicap 0 1 -1 1 -1;
RUN;
```

Order = data keeps the data in the order it came in, so that “none” group is first and can be assigned a coefficient of 0.

Comes in handy when doing division by hand would result in the need to input a rounded number (example 0.33)

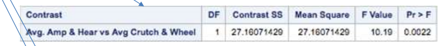
Handicap & Capability Study: In SAS



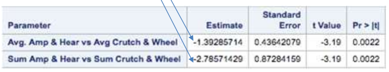
```
PROC GLM DATA = handicap ORDER=DATA;
     CLASS handicap;
     MODEL score = handicap;
     MEANS handicap;
     CONTRAST 'Avg. Amp & Hear vs Avg Crutch & Wheel' handicap 0 1 -1 1 -1;
     ESTIMATE 'Avg. Amp & Hear vs Avg Crutch & Wheel' handicap 0 1 -1 1 -1 / DIVISOR = 2;
     ESTIMATE 'Sum Amp & Hear vs Sum Crutch & Wheel' handicap 0 1 -1 1 -1;
RUN;
```

$$\gamma = 1\mu_{Amp} - 1\mu_{Crutch} + 1\mu_{Hear} + 0\mu_{None} - 1\mu_{Wheel}$$

$$\gamma = 0.5\mu_{Amp} - 0.5\mu_{Crutch} + 0.5\mu_{Hear} + 0\mu_{None} - 0.5\mu_{Wheel}$$



Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
Avg. Amp & Hear vs Avg Crutch & Wheel	1	27.16071429	27.16071429	10.19	0.0022



Parameter	Estimate	Standard Error	t Value	Pr > |t|
Avg. Amp & Hear vs Avg Crutch & Wheel	-1.39285714	0.43642079	-3.19	0.0022
Sum Amp & Hear vs Sum Crutch & Wheel	-2.78571429	0.87284159	-3.19	0.0022

Three different ways (contrast, estimate, estimate with divisor =2) to test for the same idea. (There are many more than three!)

Handicap & Capability Study: In SAS

Confidence Intervals

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Avg. Amp & Hear vs Avg Crutch & Wheel	-1.39285714	0.43642079	-3.19	0.0022
Sum Amp & Hear vs Sum Crutch & Wheel	-2.78571429	0.87284159	-3.19	0.0022

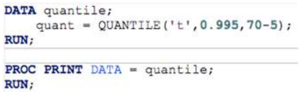
There is evidence that the average points assigned to Amp & Hear handicaps is smaller than the average points assigned to Crutch & Wheel handicaps (t-tools linear contrast p-value of 0.0022). We estimate that this difference is -1.39 pts with an associated 99% confidence interval of….

99% CI for the difference in averages of Amp and Hear vs. Crutch and Wheel:
Point estimate ± multiplier* standard error

-1.39±2.65*0.436

-1.39±1.155

(-2.55, -0.23), which of course does not include 0



```
DATA quantile;
     quant = QUANTILE('t',0.995,70-5);
RUN;

PROC PRINT DATA = quantile;
RUN;
```

Obs	quant
1	2.65360

Chapter 6

$$H_0: \frac{\mu_{Amp+}\mu_{Hear}}{2} = \frac{\mu_{Crutch+}\mu_{Wheel}}{2}$$
$$H_A: \frac{\mu_{Amp+}\mu_{Hear}}{2} \neq \frac{\mu_{Crutch+}\mu_{Wheel}}{2}$$

$$\gamma = 1\mu_{Amp} - 1\mu_{Crutch} + 1\mu_{Hear} + 0\mu_{None} - 1\mu_{Wheel}$$

```
proc glm data = Handicap;
class Handicap;
model Score = Handicap;
means Handicap / HOVTEST = BF Welch;
contrast 'Compare Ave Amp and Hearing to Avg Crutch and Wheel' Handicap 1 -1 1 0 -1;
run;
```

With no **Order = data** in the code, the contrasts are assigned in alphabetical order, so that "none" group is fourth.

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
Compare Ave Amp and Hearing to Avg Crutch and Wheel	1	27.16071429	27.16071429	10.19	0.0022

Let's Try Some from Spock Example!!

Groups: A, B, C, D, E, F, S

Write the statement (γ) for the population contrast below.
Then provide the contrast vector as you would input it in SAS. (Use alphabetical order of the subscripts.)

$$\text{H}_o\text{: } \mu_S - \frac{\mu_A + \mu_B + \mu_C + \mu_D + \mu_E + \mu_F}{6} = 0$$

$$\text{H}_a\text{: } \mu_S - \frac{\mu_A + \mu_B + \mu_C + \mu_D + \mu_E + \mu_F}{6} \neq 0$$

$\boldsymbol{\gamma} =$

Contrast vector (assume alphabetical order):

Answer on Next Slide ->

Let's Try Some from Spock Example!!

Groups: A, B, C, D, E, F, S

Write the statement (γ) for the population contrast below.
Then provide the contrast vector as you would input it in SAS. (Use alphabetical order of the subscripts.)

$$\text{H}_o\text{: } \mu_S - \frac{\mu_A + \mu_B + \mu_C + \mu_D + \mu_E + \mu_F}{6} = 0$$

$$\text{H}_a\text{: } \mu_S - \frac{\mu_A + \mu_B + \mu_C + \mu_D + \mu_E + \mu_F}{6} \neq 0$$

$$\boldsymbol{\gamma = -1\mu_A - 1\mu_B - 1\mu_C - 1\mu_D - 1\mu_E - 1\mu_F + 6\mu_S}$$

Contrast vector (assume alphabetical order): -1 -1 -1 -1 -1 -1 6

Let's Try ANOTHER (from Spock)!!

Groups: A, B, C, D, E, F, S

Write the statement (γ) for the population contrast below.
Then provide the contrast vector as you would input it in SAS. (Use alphabetical order of the subscripts.)

$$\text{H}_o\text{: } \frac{\mu_A + \mu_B + \mu_C}{3} - \frac{\mu_D + \mu_E + \mu_F}{3} = 0$$

$$\text{H}_a\text{: } \frac{\mu_A + \mu_B + \mu_C}{3} - \frac{\mu_D + \mu_E + \mu_F}{3} \neq 0$$

$\boldsymbol{\gamma} =$

Contrast vector (assume alphabetical order):

Let's Try ANOTHER (from Spock)!!

Groups: A, B, C, D, E, F, S

Write the statement (γ) for the population contrast below.
Then provide the contrast vector as you would input it in SAS. (Use alphabetical order of the subscripts.)

$$H_o: \frac{\mu_A + \mu_B + \mu_C}{3} - \frac{\mu_D + \mu_E + \mu_F}{3} = 0$$

$$H_a: \frac{\mu_A + \mu_B + \mu_C}{3} - \frac{\mu_D + \mu_E + \mu_F}{3} \neq 0$$

$$\boldsymbol{\gamma = 1\mu_A + 1\mu_B + 1\mu_C - 1\mu_D - 1\mu_E - 1\mu_F + 0\mu_S}$$

Contrast vector (assume alphabetical order): 1 1 1 -1 -1 -1 0

ADDITIONAL QUESTION:
Why is it better to include the Spock data in the calculation of the pooled SD (and thus the MSE) even though the hypothesis does not include it?

Let's Try ONE MORE (from Spock)!!

Groups: A, B, C, D, E, F, S

Write the statement (γ) for the population contrast below.
Then provide the contrast vector as you would input it in SAS. (Use alphabetical order of the subscripts.)

$$H_o: \frac{\mu_A + \mu_C}{2} - \frac{\mu_D + \mu_E + \mu_F}{3} = 0$$

$$H_a: \frac{\mu_A + \mu_C}{2} - \frac{\mu_D + \mu_E + \mu_F}{3} \neq 0$$

$\boldsymbol{\gamma =}$

Contrast vector (assume alphabetical order):

Answer on Next Slide ->

Let's Try ONE MORE (from Spock)!!

Groups: A, B, C, D, E, F, S

Write the statement (γ) for the population contrast below.
Then provide the contrast vector as you would input it in SAS. (Use alphabetical order of the subscripts.)

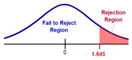
$$H_o: \frac{\mu_A + \mu_C}{2} - \frac{\mu_D + \mu_E + \mu_F}{3} = 0$$

$$H_a: \frac{\mu_A + \mu_C}{2} - \frac{\mu_D + \mu_E + \mu_F}{3} \neq 0$$

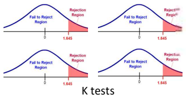
$$\boldsymbol{\gamma = 3\mu_A + 0\mu_B + 3\mu_C - 2\mu_D - 2\mu_E - 2\mu_F + 0\mu_S}$$

Contrast vector (assume alphabetical order): 3 0 3 -2 -2 -2 0

Multiple Comparison: Motivation



One Test:
P(Rejecting H_o | H_o is true) = $\alpha_{Individual}$



K tests

K Tests:
α_{Family} =P(Rejecting *at least 1* H_o | All H_o are true) $\neq \alpha_{Individual}$

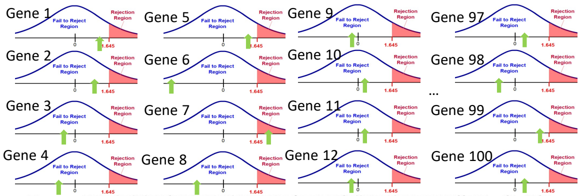
When all tests are independent and have the same alpha ($\alpha_{Individual}$),

$$\alpha_{Family} = 1 - (1 - \alpha_{Individual})^k$$

Regardless of independence, $\alpha_{Individual} \approx \frac{\alpha_{Family}}{k}$, the Bonferroni correction, where α_{Family} is typically controlled for, perhaps set at 0.05.

Multiple Comparison: Example k = 100

Familywise confidence level *is the success rate of a procedure for constructing a family of confidence intervals, where a "successful" usage is one in which all intervals in the family capture their parameters.*



α_{Family} =Probability (Reject at least 1 H_o | $\mu_i = 0$ (All n H_o's are true)) = $1 - (1- \alpha_i)^n$

$\alpha_i = .05$

$\alpha_{Family} = 1 - (1\text{-}0.05)^{100} = 1 - (0.95)^{100} = 0.994$......99% chance of a Type I error

$\alpha_i = .05/100$

$\alpha_{Family} = 1 - [1\text{-}(0.05/100)]^{100} = 1 - (.9995)^{100} = 0.0488$......5% chance of a Type I error

Confidence Intervals

Familywise confidence level *is the success rate of a procedure for constructing a family of confidence intervals, where a "successful" usage is one in which all intervals in the family capture their parameters.*

Interval half-width = (Multiplier) × (Standard error).

When we make a correction for multiple comparisons, it is the critical value in the hypothesis test and thus the multiplier in the confidence interval that is adjusted.

*The multiplier is usually the same as the critical value for a hypothesis test.

Planned & Post-hoc Tests

A planned test is one in which you know the comparisons (tests) you want to make before you look at the data.

If you have k planned comparisons then you need to correct for just those k comparisons.

When planned comparisons are not obvious, post hoc tests are conducted. In this case, we need to correct for all possible k comparisons between the m groups.

$$k = \frac{m(m-1)}{2}$$

Post-Hoc / Unplanned Tests

Post Hoc tests are appropriate when:

1. The researcher wants to examine all possible comparisons among pairs of group means (or a large number of comparisons).
2. Predictions about which groups will differ are not made prior to setting up the analysis.

Multiple Comparison: Bonferroni

If the confidence level for each of k individual comparisons is adjusted upward to $100\left(1-\frac{\alpha}{k}\right)\%$, the chance that all intervals succeed simultaneously is at least $100(1-\alpha)\%$

$$multiplier = t\_multiplier = t_{\left(1-\frac{\alpha}{2k}\right),df}$$

For a set of **Bonferroni adjusted t-tests**, (α/k) we must have normal distributions, equal spreads, and independence (same as typical t-tests).

However, the **Bonferroni correction** can be extended to tests that have no assumptions about distributions (e.g. rank sum test). For any set of independent parametric or non-parametric tests, the Bonferroni correction works the same.

This approach is very conservative, meaning that the intervals are much wider than the nominal level, particularly if the tests are not really independent.

Multiple Comparison: Tukey-Kramer

Tukey's HSD Procedure

- Makes use of the Studentized Range Statistic:

$$\text{Multiplier} = q = \frac{\bar{x}_{\text{largest}} - \bar{x}_{\text{smallest}}}{\sqrt{MS_w(1/n)}}$$

Studentized Range Statistic Table

- Obtains simultaneous confidence intervals for each pair of population means $(\mu_i - \mu_j)$

$$(\bar{x}_i - \bar{x}_j) \pm q_{\alpha;k,N-k}\sqrt{\frac{MS_w}{n}}$$

The Tukey-Kramer adjustment is a modification to this test to account for different sample sizes in the groups.

- $q_\alpha(k,N-k)$ is the upper-tail critical value of the Studentized range for comparing k populations.

Assumes normal distributions, equal spreads, independence (same as typical t-tests), and equal group sample sizes.

More consistent than Bonferroni with respect to Type I Error but not robust to its assumptions…. Bonferroni is a good alternative when the assumptions are violated.

Multiple Comparison: Dunnett
Many Groups to one Control

$$t_2 = \frac{\hat{\mu}_C - \hat{\mu}_2}{SE_{\hat{\mu}_C - \hat{\mu}_2}}$$

…

$$t_n = \frac{\hat{\mu}_C - \hat{\mu}_n}{SE_{\hat{\mu}_C - \hat{\mu}_n}}$$

Assumes normal distributions, equal spreads, and independence (same as typical t-tests).

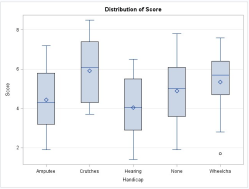
Replaces t-distribution with a multivariate t-distribution (n=# of groups versus control), where the tests are not independent.

Handicap / Capability Study: Data

Seventy undergraduate students from a U.S. university were randomly assigned to view the tapes, fourteen to each tape. After viewing the tape, each subject rated the qualifications of the applicant on a 0- to 10-point applicant qualification scale. Display 6.1 shows the results. The question is, do subjects systematically evaluate qualifications differently according to the candidate's handicap? If so, which handicaps produce the different evaluations?

	None	Amputee	Crutches	Hearing	Wheelchair
0					
1	9	9		4	7
2	5	56		149	8
3	06	268	7	479	5
4	129	06	033	237	78
5	149	3589	18	589	03
6	17	1	0234	5	1124
7	48	2	445		246
8			5		
9					

Legend: 7 | 4 represents a score of 7.4 on the Applicant Qualification Scale.



Handicap Data Analysis

Questions of Interest:

1. Is there any evidence that at least one pair of mean qualification scores are different from each other?
2. Let's say we are only interested in Amputee versus None. Test the claim the Amputee has a different mean score than the None group.
3. Now let's assume that we are interested in identifying specific differences between ***any two*** of the group means. Find evidence of any differences in the means between the groups.
4. Next, assume that we were interested in testing the means of the handicapped groups to the non-handicap group. Test this claim and identify any significant differences.

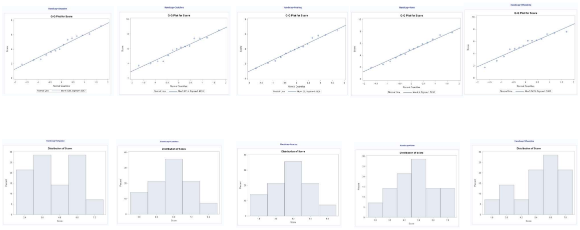
First Test!!!

H_o: *All Means are Equal*

H_a: *At least 2 means are different from eac*

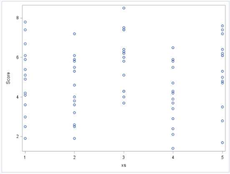
(or at least 1 mean is different from the rest)

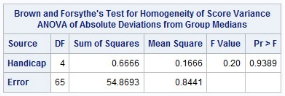
Normality: Handicap Data



There is no visual evidence to suggest that the data are not normally distributed. We will proceed with the assumption of normally distributed groups.

Homogeneity of SD Assumption





Brown and Forsythe's Test for Homogeneity of Score Variance
ANOVA of Absolute Deviations from Group Medians

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Handicap	4	0.6666	0.1666	0.20	0.9389
Error	65	54.8693	0.8441		

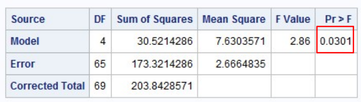
There is no evidence to suggest variances are unequal.

Independence may be violated here. We are going to proceed anyway for the sake of the example.

First QOI!!!

1. Is there any evidence that at least one pair of mean qualification scores are different from each other?

$$H_o: All\ Means\ are\ Equal$$
$$H_a: At\ least\ 2\ means\ are\ different\ from\ eacl$$
$$(or\ at\ least\ 1\ mean\ is\ different\ from\ the\ rest)$$



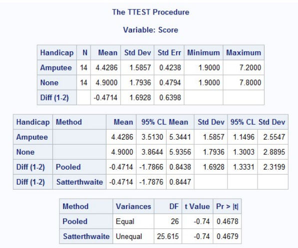
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	30.5214286	7.6303571	2.86	0.0301
Error	65	173.3214286	2.6664835		
Corrected Total	69	203.8428571			

There is sufficient evidence to suggest at the alpha = .05 level of significance (p-value = .0301) that at least 2 of the means are different from each other in this standard ANOVA.

Second QOI!!!

2. Let's say we are only interested in Amputee versus None. Test the claim the Amputee has a different mean score than the None group.

$$H_o: \mu_{Amputee} = \mu_{None}$$
$$H_a: \mu_{Amputee} \neq \mu_{None}$$



The TTEST Procedure

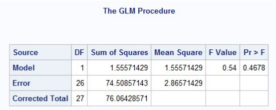
Variable: Score

Handicap	N	Mean	Std Dev	Std Err	Minimum	Maximum
Amputee	14	4.4286	1.5857	0.4238	1.9000	7.2000
None	14	4.9000	1.7936	0.4794	1.9000	7.8000
Diff (1-2)		-0.4714	1.6928	0.6398		

Handicap	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Amputee		4.4286	3.5130	5.3441	1.5857	1.1496	2.5547
None		4.9000	3.8644	5.9356	1.7936	1.3003	2.8895
Diff (1-2)	Pooled	-0.4714	-1.7866	0.8438	1.6928	1.3331	2.3199
Diff (1-2)	Satterthwaite	-0.4714	-1.7876	0.8447			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	26	-0.74	0.4678
Satterthwaite	Unequal	25.615	-0.74	0.4679

```
proc ttest data = handicap;
where handicap eq 'None' | handicap eq 'Amputee';
class handicap;
var score;
run;
```



The GLM Procedure

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1.55571429	1.55571429	0.54	0.4678
Error	26	74.50857143	2.86571429		
Corrected Total	27	76.06428571			

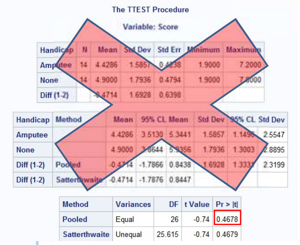
```
proc glm data = handicap;
where handicap eq 'None' | handicap eq 'Amputee';
class handicap;
model score = handicap;
means handicap / hovtest = bf bon cldiff;
run;
```

The results of these tests are equivalent! There is not sufficient evidence to suggest that the mean qualification rating of the amputee group is different than the group without handicap. (P-value = .4678 from a t-test and an ANOVA using only these two groups.)

Second QOI: Better approach!!!

2. Let's say we are only interested in Amputee versus None. Test the claim the Amputee has a different mean score than the None group.

$$H_o: \mu_{Amputee} = \mu_{None}$$
$$H_a: \mu_{Amputee} \neq \mu_{None}$$



The TTEST Procedure

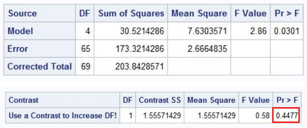
Variable: Score

Handicap	N	Mean	Std Dev	Std Err	Minimum	Maximum
Amputee	14	4.4286	1.5857	0.4238	1.9000	7.2000
None	14	4.9000	1.7936	0.4794	1.9000	7.8000
Diff (1-2)		-0.4714	1.6928	0.6398		

Handicap	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Amputee		4.4286	3.5130	5.3441	1.5857	1.1496	2.5547
None		4.9000	3.8644	5.9356	1.7936	1.3003	2.8895
Diff (1-2)	Pooled	-0.4714	-1.7866	0.8438	1.6928	1.3331	2.3199
Diff (1-2)	Satterthwaite	-0.4714	-1.7876	0.8447			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	26	-0.74	0.4678
Satterthwaite	Unequal	25.615	-0.74	0.4679

```
proc ttest data = handicap;
where handicap eq 'None' | handicap eq 'Amputee';
class handicap;
var score;
run;
```

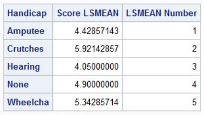


Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	30.5214286	7.6303571	2.86	0.0301
Error	65	173.3214286	2.6664835		
Corrected Total	69	203.8428571			

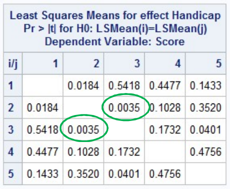
Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
Use a Contrast to Increase DF!	1	1.55571429	1.55571429	0.58	0.4477

```
proc glm data = handicap;
class handicap;
model score = handicap;
means handicap / hovtest = bf bon cldiff;
contrast "Use a Contrast to Increase DF!" handicap 1 0 0 -1 0;
run;
```

There is not sufficient evidence to suggest that the mean qualification rating of the amputee group is different than the group with no handicap (p-value = .4477 from a contrast using all available data). Even though the p-values for the two tests are only slightly different, it is better to use all available data (the procedure on the right).
Comparing a pair of means can be just a simple contrast.



Handicap	Score LSMEAN	LSMEAN Number
Amputee	4.42857143	1
Crutches	5.92142857	2
Hearing	4.05000000	3
None	4.90000000	4
Wheelcha	5.34285714	5



Least Squares Means for effect Handicap
Pr > |t| for H0: LSMean(i)=LSMean(j)
Dependent Variable: Score

i/j	1	2	3	4	5
1		0.0184	0.5418	0.4477	0.1433
2	0.0184		0.0035	0.1028	0.3520
3	0.5418	0.0035		0.1732	0.0401
4	0.4477	0.1028	0.1732		0.4756
5	0.1433	0.3520	0.0401	0.4756	

```
proc glm data = handicap;
class handicap;
model score = handicap;
means handicap / hovtest = bf;
lsmeans handicap / pdiff;
run;
```

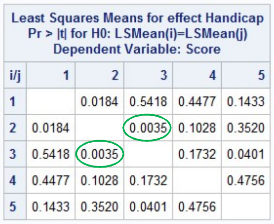
Third QOI!!!

Now let's assume that we are interested in identifying specific differences between ***any two*** group means. Find evidence of any differences in the means between the groups.

There are 10 different two sided tests conducted here; thus, we need to adjust alpha per test to be .05/10 = .005. With this adjustment, only one of the tests has a statistically significant result. Therefore, there is evidence (p-value = .0035 from a t-test) that the crutches and hearing groups have different mean qualification rating scores. We will provide a confidence interval in a few slides.

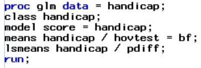
Bonferroni Adjusted P-Values

P-values not adjusted- compare to individual alpha



Least Squares Means for effect Handicap Pr > |t| for H0: LSMean(i)=LSMean(j) Dependent Variable: Score					
i/j	1	2	3	4	5
1		0.0184	0.5418	0.4477	0.1433
2	0.0184		0.0035	0.1028	0.3520
3	0.5418	0.0035		0.1732	0.0401
4	0.4477	0.1028	0.1732		0.4756
5	0.1433	0.3520	0.0401	0.4756	

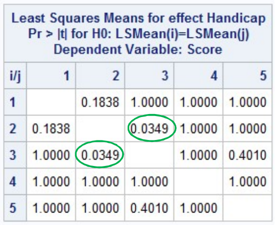
Compare to alpha = 0.005



```
proc glm data = handicap;
class handicap;
model score = handicap;
means handicap / hovtest = bf;
lsmeans handicap / pdiff;
run;
```

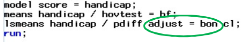
x 10, up to 1

P-values adjusted- compare to family-wise alpha



Least Squares Means for effect Handicap Pr > |t| for H0: LSMean(i)=LSMean(j) Dependent Variable: Score					
i/j	1	2	3	4	5
1		0.1838	1.0000	1.0000	1.0000
2	0.1838		0.0349	1.0000	1.0000
3	1.0000	0.0349		1.0000	0.4010
4	1.0000	1.0000	1.0000		1.0000
5	1.0000	1.0000	0.4010	1.0000	

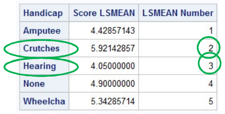
Compare to alpha = 0.05



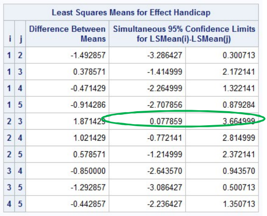
```
proc glm data = handicap;
class handicap;
model score = handicap;
means handicap / hovtest = bf;
lsmeans handicap / pdiff adjust = bon cl;
run;
```

Third QOI!!!

Now let's assume that we are interested in identifying specific differences between ***any two*** group means. Find evidence of any differences in the means between the groups.

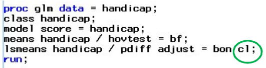


Handicap	Score LSMEAN	LSMEAN Number
Amputee	4.42857143	1
Crutches	5.92142857	2
Hearing	4.05000000	3
None	4.90000000	4
Wheelcha	5.34285714	5



Least Squares Means for Effect Handicap				
i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	-1.492857	-3.286427	0.300713
1	3	0.378571	-1.414999	2.172141
1	4	-0.471429	-2.264999	1.322141
1	5	-0.914286	-2.707856	0.879284
2	3	1.871429	0.077859	3.664999
2	4	1.021429	-0.772141	2.814999
2	5	0.578571	-1.214999	2.372141
3	4	-0.850000	-2.643570	0.943570
3	5	-1.292857	-3.086427	0.500713
4	5	-0.442857	-2.236427	1.350713

A 95% confidence interval for the difference in means of the crutches and hearing groups is (.0779, 3.66499).



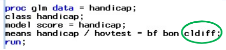
```
proc glm data = handicap;
class handicap;
model score = handicap;
means handicap / hovtest = bf;
lsmeans handicap / pdiff adjust = bon cl;
run;
```

Third QOI!!!

Now let's assume that we are interested in identifying specific differences between ***any two*** group means. Find evidence of any differences in the means between the groups.

Comparisons significant at the 0.05 level are indicated by ***.				
Handicap Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
Crutches - Wheelcha	0.5786	-1.2150	2.3721	
Crutches - None	1.0214	-0.7721	2.8150	
Crutches - Amputee	1.4929	-0.3007	3.2864	
Crutches - Hearing	1.8714	0.0779	3.6650	***
Wheelcha - Crutches	-0.5786	-2.3721	1.2150	
Wheelcha - None	0.4429	-1.3507	2.2364	
Wheelcha - Amputee	0.9143	-0.8793	2.7079	
Wheelcha - Hearing	1.2929	-0.5007	3.0864	
None - Crutches	-1.0214	-2.8150	0.7721	
None - Wheelcha	-0.4429	-2.2364	1.3507	
None - Amputee	0.4714	-1.3221	2.2650	
None - Hearing	0.8500	-0.9436	2.6436	
Amputee - Crutches	-1.4929	-3.2864	0.3007	
Amputee - Wheelcha	-0.9143	-2.7079	0.8793	
Amputee - None	-0.4714	-2.2650	1.3221	
Amputee - Hearing	0.3786	-1.4150	2.1721	
Hearing - Crutches	-1.8714	-3.6650	-0.0779	***
Hearing - Wheelcha	-1.2929	-3.0864	0.5007	
Hearing - None	-0.8500	-2.6436	0.9436	
Hearing - Amputee	-0.3786	-2.1721	1.4150	

A 95% confidence interval for the difference in means of crutches and hearing groups is (.0779, 3.66499).



```
proc glm data = handicap;
class handicap;
model score = handicap;
means handicap / hovtest = bf bon cldiff;
run;
```

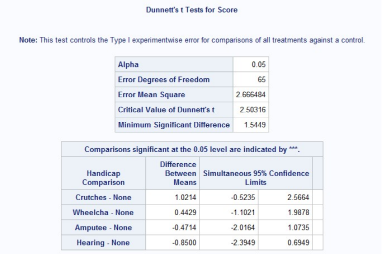
***Slightly different code from the last slide, producing slightly different output. Note the cl versus cldiff.**

4th QOI: Next, assume that we are interested in testing the means of the handicapped groups with the non-handicapped group. Test this claim and identify any significant differences. (Using CIs)

There is NOT sufficient evidence in this study to suggest that there are any differences between the average of the means of each handicap group and the mean of the group without handicap.

The 95% family-wise confidence intervals are constructed using Dunnett's procedure. All CIs contain zero, thus not providing sufficient evidence to conclude that the difference is not zero.

(The study results do not constitute sufficient evidence to support the claim that any means tested are individually different than the control.)

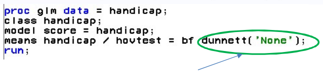


Dunnett's t Tests for Score

Note: This test controls the Type I experimentwise error for comparisons of all treatments against a control.

Alpha	0.05
Error Degrees of Freedom	65
Error Mean Square	2.666484
Critical Value of Dunnett's t	2.50316
Minimum Significant Difference	1.5449

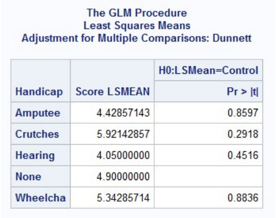
Comparisons significant at the 0.05 level are indicated by ***.				
Handicap Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
Crutches - None	1.0214	-0.5235	2.5664	
Wheelcha - None	0.4429	-1.1021	1.9878	
Amputee - None	-0.4714	-2.0164	1.0735	
Hearing - None	-0.8500	-2.3949	0.6949	



```
proc glm data = handicap;
class handicap;
model score = handicap;
means handicap / hovtest = bf dunnett('None');
run;
```

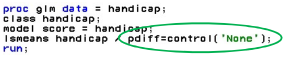
Specify the control group

4th QOI: Next, assume that we were interested in testing the means of the handicapped groups with the non-handicap group. Test this claim and identify any significant differences. (Using HTs)



The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Dunnett

Handicap	Score LSMEAN	H0:LSMean=Control Pr > |t|
Amputee	4.42857143	0.8597
Crutches	5.92142857	0.2918
Hearing	4.05000000	0.4516
None	4.90000000	
Wheelcha	5.34285714	0.8836

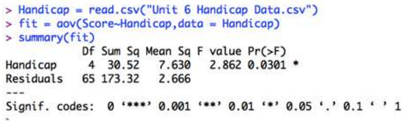


```
proc glm data = handicap;
class handicap;
model score = handicap;
lsmeans handicap / pdiff=control('None');
run;
```

Hypothesis tests also conclude that there is not sufficient evidence to suggest that there are any differences between the means of each handicapped group and the mean of the of the group without handicap. The above Dunnett adjusted p-values are all greater than alpha = .05, as is visible from the table above.

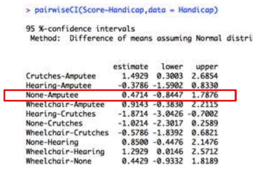
R Code for Handicap Example Question 1

Question 1: Reading in Data and ANOVA



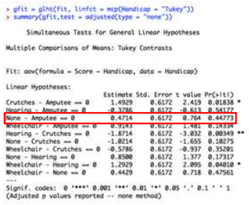
```
> Handicap = read.csv("Unit 6 Handicap Data.csv")
> fit = aov(Score~Handicap,data = Handicap)
> summary(fit)
            Df Sum Sq Mean Sq F value Pr(>F)
Handicap     4  30.52   7.630   2.862 0.0301 *
Residuals   65 173.32   2.666
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

R Code for Handicap Example Question 2



```
> pairwiseCI(Score~Handicap,data = Handicap)

95 %-confidence intervals
 Method:  Difference of means assuming Normal distribution, allowing unequal variances


                    estimate   lower   upper
Crutches-Amputee      1.4929  0.3003  2.6854
Hearing-Amputee      -0.3786 -1.5902  0.8330
None-Amputee          0.4714 -0.8447  1.7876
Wheelchair-Amputee    0.9143 -0.3830  2.2115
Hearing-Crutches     -1.8714 -3.0426 -0.7002
None-Crutches        -1.0214 -2.3017  0.2589
Wheelchair-Crutches  -0.5786 -1.8392  0.6821
None-Hearing          0.8500 -0.4476  2.1476
Wheelchair-Hearing    1.2929  0.0146  2.5712
Wheelchair-None       0.4429 -0.9332  1.8189
```

Note: Must Load pairwiseCI package

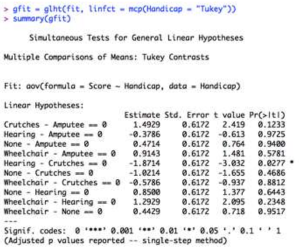


```
> gfit = glht(fit, linfct = mcp(Handicap = "Tukey"))
> summary(gfit,test = adjusted(type = "none"))

	 Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = Score ~ Handicap, data = Handicap)

Linear Hypotheses:
                           Estimate Std. Error t value Pr(>|t|)
Crutches - Amputee == 0      1.4929     0.6172   2.419  0.01838 *
Hearing - Amputee == 0      -0.3786     0.6172  -0.613  0.54177
None - Amputee == 0          0.4714     0.6172   0.764  0.44773
Wheelchair - Amputee == 0    0.9143     0.6172   1.481  0.14334
Hearing - Crutches == 0     -1.8714     0.6172  -3.032  0.00349 **
None - Crutches == 0        -1.0214     0.6172  -1.655  0.10275
Wheelchair - Crutches == 0  -0.5786     0.6172  -0.937  0.35201
None - Hearing == 0          0.8500     0.6172   1.377  0.17317
Wheelchair - Hearing == 0    1.2929     0.6172   2.095  0.04010 *
Wheelchair - None == 0       0.4429     0.6172   0.718  0.47561
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- none method)
```

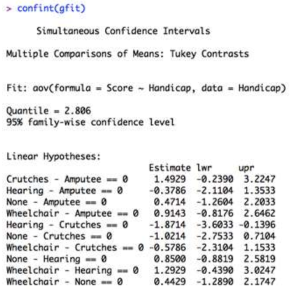
Note: Must Load multcomp package

R Code for Handicap Example Question 3

Note: Must Load multcomp package



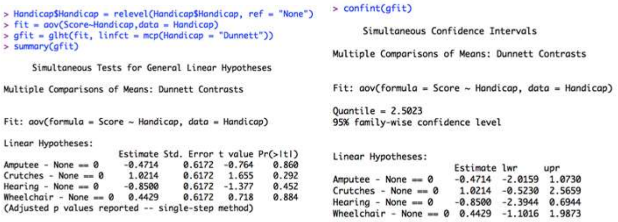
```
> gfit = glht(fit, linfct = mcp(Handicap = "Tukey"))
> summary(gfit)

	 Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = Score ~ Handicap, data = Handicap)

Linear Hypotheses:
                           Estimate Std. Error t value Pr(>|t|)
Crutches - Amputee == 0      1.4929     0.6172   2.419   0.1233
Hearing - Amputee == 0      -0.3786     0.6172  -0.613   0.9725
None - Amputee == 0          0.4714     0.6172   0.764   0.9400
Wheelchair - Amputee == 0    0.9143     0.6172   1.481   0.5781
Hearing - Crutches == 0     -1.8714     0.6172  -3.032   0.0277 *
None - Crutches == 0        -1.0214     0.6172  -1.655   0.4686
Wheelchair - Crutches == 0  -0.5786     0.6172  -0.937   0.8812
None - Hearing == 0          0.8500     0.6172   1.377   0.6443
Wheelchair - Hearing == 0    1.2929     0.6172   2.095   0.2348
Wheelchair - None == 0       0.4429     0.6172   0.718   0.9517
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```



```
> confint(gfit)

	 Simultaneous Confidence Intervals

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = Score ~ Handicap, data = Handicap)

Quantile = 2.806
95% family-wise confidence level


Linear Hypotheses:
                           Estimate lwr     upr
Crutches - Amputee == 0     1.4929  -0.2390  3.2247
Hearing - Amputee == 0     -0.3786  -2.1104  1.3533
None - Amputee == 0         0.4714  -1.2604  2.2033
Wheelchair - Amputee == 0   0.9143  -0.8176  2.6462
Hearing - Crutches == 0    -1.8714  -3.6033 -0.1396
None - Crutches == 0       -1.0214  -2.7533  0.7104
Wheelchair - Crutches == 0 -0.5786  -2.3104  1.1533
None - Hearing == 0         0.8500  -0.8819  2.5819
Wheelchair - Hearing == 0   1.2929  -0.4390  3.0247
Wheelchair - None == 0      0.4429  -1.2890  2.1747
```

R Code for Handicap Example Question 4

Note: Must Load multcomp package



```
> Handicap$Handicap = relevel(Handicap$Handicap, ref = "None")
> fit = aov(Score~Handicap,data = Handicap)
> gfit = glht(fit, linfct = mcp(Handicap = "Dunnett"))
> summary(gfit)

     Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Dunnett Contrasts


Fit: aov(formula = Score ~ Handicap, data = Handicap)

Linear Hypotheses:
                        Estimate Std. Error t value Pr(>|t|)
Amputee - None == 0      -0.4714     0.6172  -0.764    0.860
Crutches - None == 0      1.0214     0.6172   1.655    0.292
Hearing - None == 0      -0.8500     0.6172  -1.377    0.452
Wheelchair - None == 0    0.4429     0.6172   0.718    0.884
(Adjusted p values reported -- single-step method)
```

```
> confint(gfit)

     Simultaneous Confidence Intervals

Multiple Comparisons of Means: Dunnett Contrasts


Fit: aov(formula = Score ~ Handicap, data = Handicap)

Quantile = 2.5023
95% family-wise confidence level


Linear Hypotheses:
                       Estimate lwr     upr
Amputee - None == 0    -0.4714  -2.0159  1.0730
Crutches - None == 0    1.0214  -0.5230  2.5659
Hearing - None == 0    -0.8500  -2.3944  0.6944
Wheelchair - None == 0  0.4429  -1.1016  1.9873
```

Appendix

Bonferroni's Correction

- Let α_{Family} be the experiment-wise Type I error rate.
- Let k be the number of pairwise comparisons, where each pairwise comparison has an index *i* associated with it.
- Let $H_{o,i}$ be the event that the null hypothesis associated with pairwise comparison i is true, for $1 \leq i \leq k$.
- Let p_i be the p-value for hypothesis test i, for $1 \leq i \leq k$.
- Let $\alpha_{Individual} = \alpha_c$ be the same for all k hypothesis tests.
- By the def. of Type I error rate, $\alpha_c = P\left(p_i < \alpha_c | H_{o,i}\right)$ for all $1 \leq i \leq k$.
- Let T be the set of indices associated with all TRUE null hypotheses, and suppose$|T| = k_0$. That is, k_0 is the number of TRUE null hypotheses.
- Then, $\alpha_{Family} = P\{\cup_{i \in T}\left(p_i < \alpha_c | H_{o,i}\right)\}$.
- By Boole's inequality (i.e., $P(A \cup B) \leq P(A)+P(B)$),

$$P\left\{\bigcup_{i \in T}\left(p_i < \alpha_c | H_{o,i}\right)\right\} \leq \sum_{i \in T} P\left(p_i < \alpha_c | H_{o,i}\right)$$

Bonferroni's Correction

$$\sum_{i \in T} P\left(p_i < \alpha_c | H_{o,i}\right) = k_0 P\left(p_i < \alpha_c | H_{o,i}\right)$$

$$k_0 P\left(p_i < \alpha_c | H_{o,i}\right) = k_0 \alpha_c \leq k \alpha_c$$

Hence, $\alpha_{Family} \leq k\alpha_c$.

Now, if we have in mind a family-wise Type I error rate of α, we can set the Type I error of the individual hypothesis tests to $\frac{\alpha}{k}$. In doing so, we are assured that $\alpha_{Family} \leq k\frac{\alpha}{k} = \alpha$. Therefore, choosing an individual Type I error rate of $\frac{\alpha}{k}$ will ensure that the family-wise Type I error rate is less than α.

Bonferroni's Correction

We know that we can force α_{Family} to be less than a specified α, but with a lower α_{Family} comes a higher β (Type II error rate). So, we want to ensure that α_{Family} is not too low. How can we be sure that α_{Family} is really close to alpha, not just less than alpha?

When the k hypothesis tests are independent, $\alpha_{Family} = 1 - (1 - \alpha_c)^k$.

Remember from calculus that any differentiable function can be approximated by the elements in its Taylor Series expansion, with the approximation getting better and better the more terms you add to the series (because the terms of the series converge to zero).

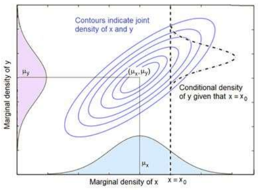
For the function $f(\alpha_c) = 1 - (1 - \alpha_c)^k$, here are the first two terms of the Taylor series approximation about the point 0 (which is reasonable as we expect to choose α_c near 0).

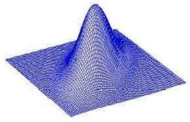
$$f(\alpha_c) \cong f(0) + f'(0)(\alpha_c - 0) = [1 - (1 - 0)^k] + k\,(1 - 0)^{k-1}(\alpha_c - 0) = [1 - (1)^k] + k(1)^{k-1}(\alpha_c)$$
$$= [1 - 1] + k\alpha_c = k\alpha_c$$

By setting $\alpha_c = \frac{\alpha}{k}$, $f(\alpha_c) \cong k\frac{\alpha}{k} = \alpha$. So, not only is α an upper bound on α_{Family}, but when the tests are independent, they are approximately equal. Even when the tests are not independent, simulations have shown that α_{Family} is pretty close to α.

Multivariate distribution

- A multivariate distribution is distribution of a vector of conditional random variables.
- Bivariate normal distribution can easily be shown graphically.

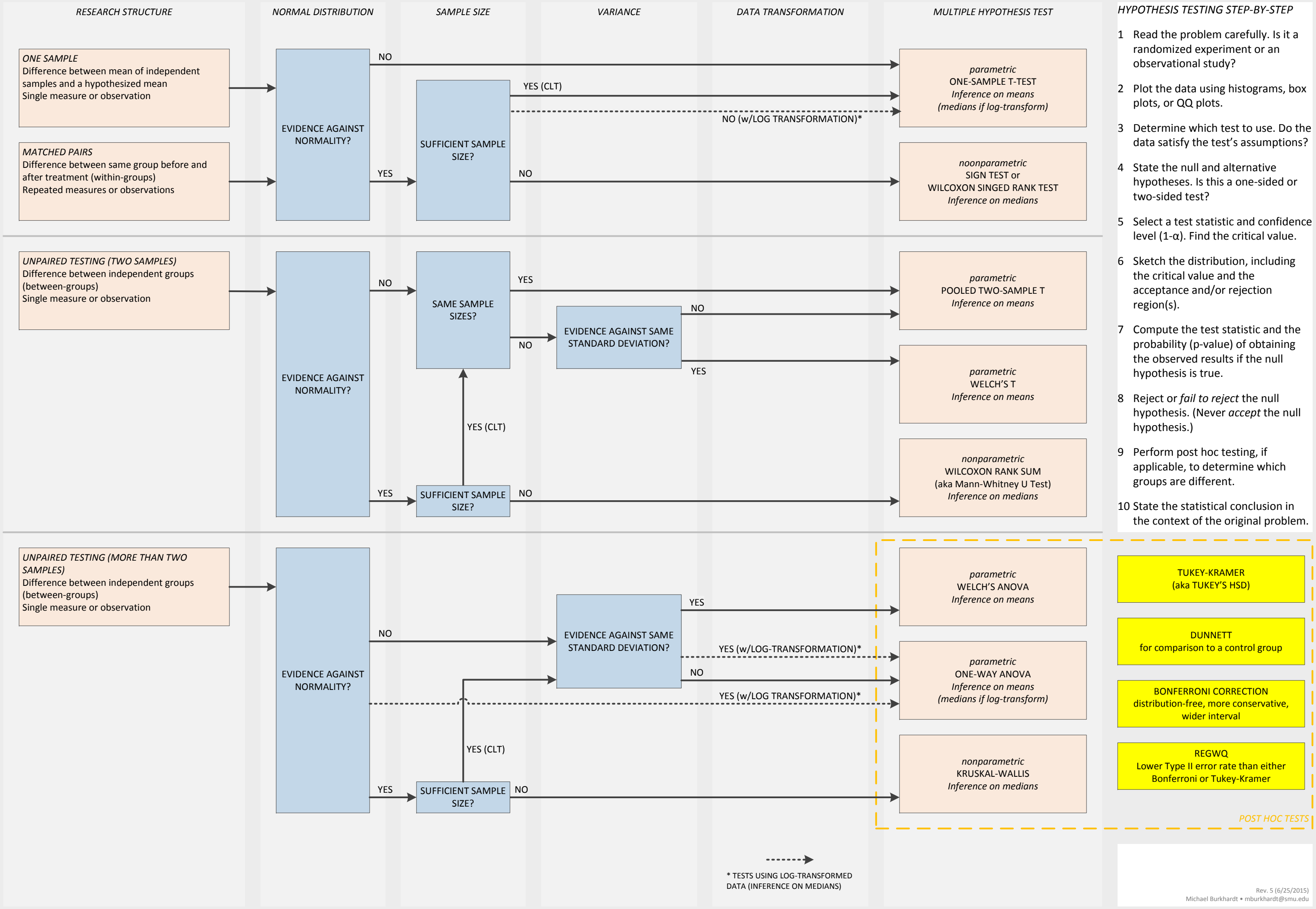




Part VII

Workflow for testing hypotheses

CHOOSING A HYPOTHESIS TEST



RESEARCH STRUCTURE

NORMAL DISTRIBUTION

SAMPLE SIZE

VARIANCE

DATA TRANSFORMATION

MULTIPLE HYPOTHESIS TEST

ONE SAMPLE
Difference between mean of independent samples and a hypothesized mean
Single measure or observation

MATCHED PAIRS
Difference between same group before and after treatment (within-groups)
Repeated measures or observations

EVIDENCE AGAINST NORMALITY?

NO

YES

SUFFICIENT SAMPLE SIZE?

YES (CLT)

NO (w/LOG TRANSFORMATION)*

NO

parametric
ONE-SAMPLE T-TEST
Inference on means
(medians if log-transform)

noonparametric
SIGN TEST or
WILCOXON SINGED RANK TEST
Inference on medians

UNPAIRED TESTING (TWO SAMPLES)
Difference between independent groups (between-groups)
Single measure or observation

EVIDENCE AGAINST NORMALITY?

NO

SAME SAMPLE SIZES?

YES

NO

EVIDENCE AGAINST SAME STANDARD DEVIATION?

NO

YES

YES (CLT)

YES

SUFFICIENT SAMPLE SIZE?

NO

parametric
POOLED TWO-SAMPLE T
Inference on means

parametric
WELCH'S T
Inference on means

nonparametric
WILCOXON RANK SUM
(aka Mann-Whitney U Test)
Inference on medians

UNPAIRED TESTING (MORE THAN TWO SAMPLES)
Difference between independent groups (between-groups)
Single measure or observation

EVIDENCE AGAINST NORMALITY?

NO

EVIDENCE AGAINST SAME STANDARD DEVIATION?

YES

YES (w/LOG-TRANSFORMATION)*

NO

YES (w/LOG TRANSFORMATION)*

YES (CLT)

YES

SUFFICIENT SAMPLE SIZE?

NO

parametric
WELCH'S ANOVA
Inference on means

parametric
ONE-WAY ANOVA
Inference on means
(medians if log-transform)

nonparametric
KRUSKAL-WALLIS
Inference on medians

TUKEY-KRAMER
(aka TUKEY'S HSD)

DUNNETT
for comparison to a control group

BONFERRONI CORRECTION
distribution-free, more conservative, wider interval

REGWQ
Lower Type II error rate than either Bonferroni or Tukey-Kramer

POST HOC TESTS

* TESTS USING LOG-TRANSFORMED DATA (INFERENCE ON MEDIANS)

HYPOTHESIS TESTING STEP-BY-STEP

1 Read the problem carefully. Is it a randomized experiment or an observational study?

2 Plot the data using histograms, box plots, or QQ plots.

3 Determine which test to use. Do the data satisfy the test's assumptions?

4 State the null and alternative hypotheses. Is this a one-sided or two-sided test?

5 Select a test statistic and confidence level (1-α). Find the critical value.

6 Sketch the distribution, including the critical value and the acceptance and/or rejection region(s).

7 Compute the test statistic and the probability (p-value) of obtaining the observed results if the null hypothesis is true.

8 Reject or *fail to reject* the null hypothesis. (Never *accept* the null hypothesis.)

9 Perform post hoc testing, if applicable, to determine which groups are different.

10 State the statistical conclusion in the context of the original problem.

Rev. 5 (6/25/2015)

Michael Burkhardt • mburkhardt@smu.edu

note that the nonparamteric ones do medians, kruskal is nonparametric for ANOVA