

UNIT 5 HW

This class allows you to practice preparing professional looking reports. Make sure all reports are typed and all graphs (unless otherwise noted) are computer generated and copied and pasted into your report. If you would like help with Word or Excel please don't hesitate to ask.

1. Read Chapter 4 from Statistical Sleuth and answer the conceptual problems at the end of the chapter. Note: You do not need to type these up and turn them in. The answers are at the very end of the chapter.
2. When wildfires ravage forests, the timber industry argues that logging (**actual logging of trees ... not the mathematical log!**) the burned trees enhance forest recovery; the EPA argues the opposite. The 2002 Biscuit Fire in southwest Oregon provided a test case. Researchers selected 16 fire-affected plots in 2004, before any logging was done and counted tree seedlings along a randomly located transect pattern in each plot. They returned in 2005, after nine of the plots had been logged, and counted the tree seedlings along the same transects. The percent of seedlings lost from 2004 to 2005 is recorded in the file **logging.csv** for logged (L) and unlogged (U) plots: Test the EPA's assertion (and thus the opposite of the logging industries assertion) that logging (**again.... not the mathematical log! :) actually, increases the percentage of seedlings lost from 2004 to 2005.**
 - a. Perform a complete analysis using a rank sum test in SAS. (Logging data).

Step 1: Hypotheses:

H_0 : The median of the percent of lost seedlings in the logged plots is equal to that of the unlogged plots. $\text{Median}_U = \text{Median}_L$

H_a : The median percentage of lost seedlings in the logged plots is more than that of the unlogged plots. $\text{Median}_U < \text{Median}_L$

Step 2 – Identify the Critical Value: critical value for Normal approximation:

p-value = -1.645

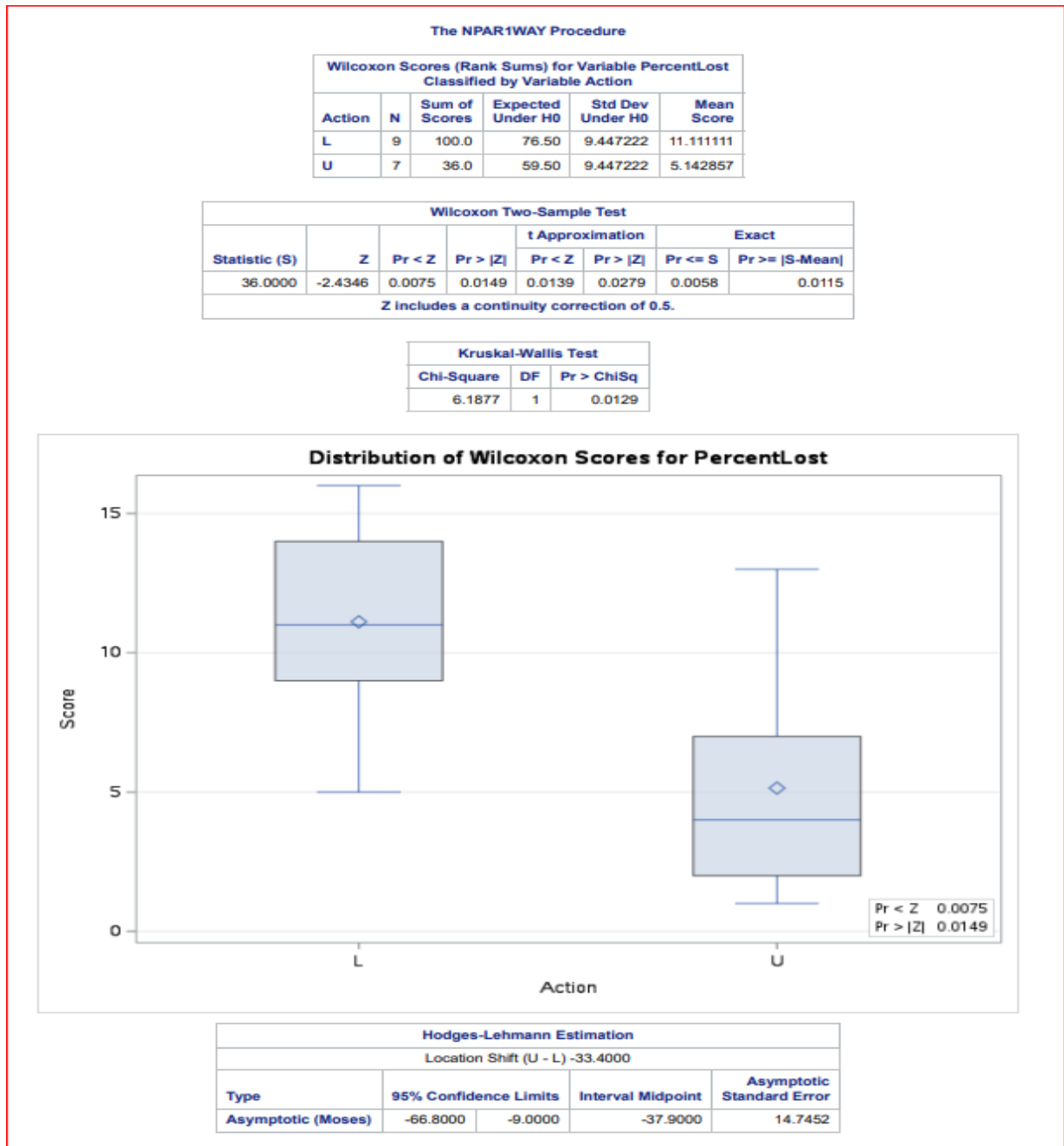
1-sided, $\alpha = 0.05$

Step 3 – Value of Test Statistic: $z = -2.4346$

Step 4 – Give p-value: *p-value* 0.0075 (1 sided)

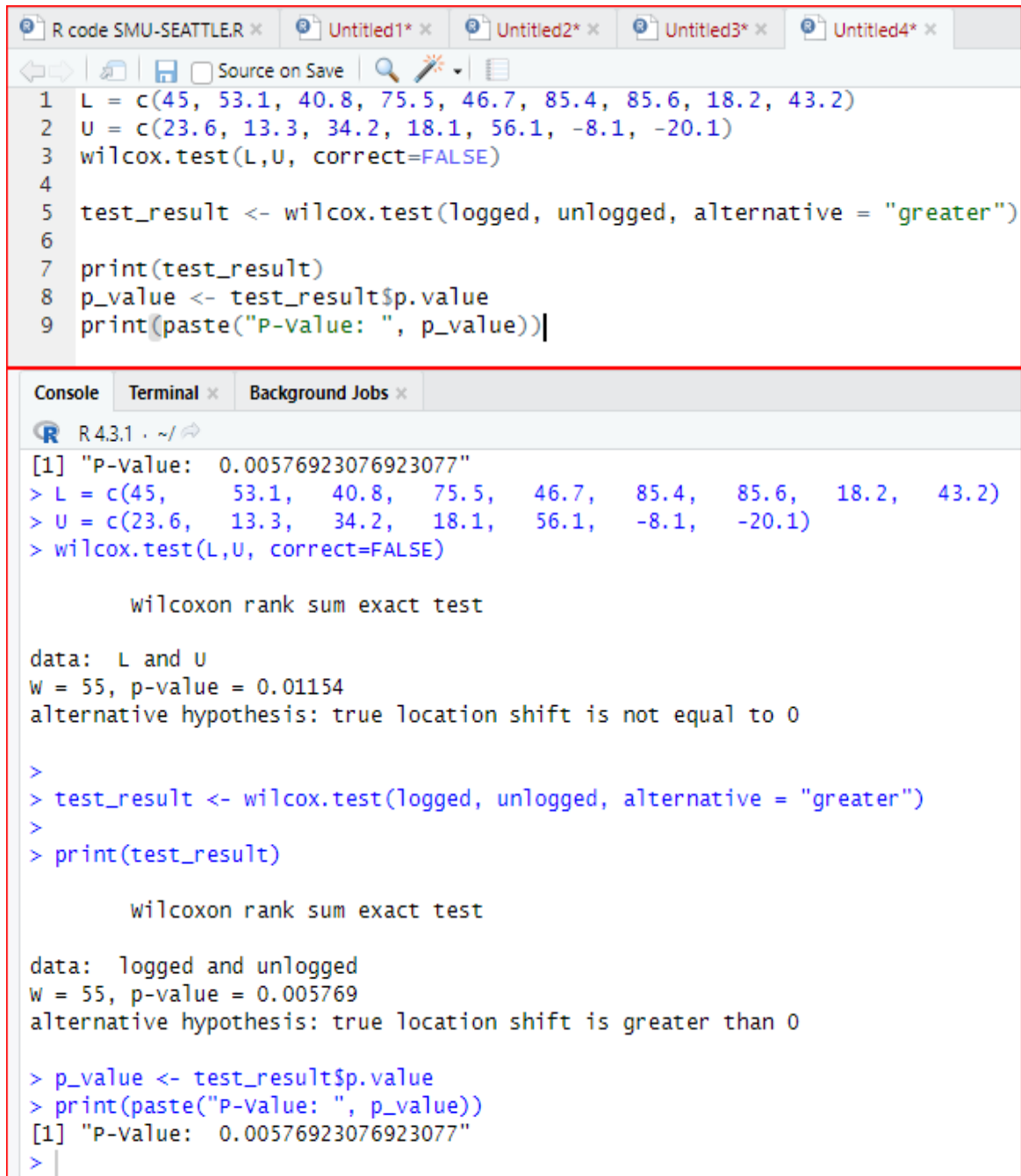
Step 5 – Decision: Reject H_0 (assuming $\alpha = 0.05$)

Step 6 - Conclusion: Robust evidence indicates that logged plots experience a higher median percentage of seedling loss compared to unlogged plots, as supported by a one-sided rank sum test ($p=0.0075$). The 95% confidence interval for the elevated median percentage of seedlings lost in logged areas ranges from 10.8 to 65.1. It's worth noting that a 90% confidence interval, ranging from 18.9 to 62.0, might be more congruent with the one-sided test at a significance level of $\alpha = 0.05$, ensuring alignment between the hypothesis test results and the confidence intervals



```
proc nparway data = log wilcoxon;
class action;
var percentlost;
exact HL Wilcoxon;
run;
```

- b. Verify the p-value and confidence interval by running the rank sum test in R (using R function `Wilcox.test`). (You do not need to repeat the complete analysis ... simply cut and paste a screen shot of your code and the output.) You may use: <https://www.r-bloggers.com/wilcoxon-mann-whitney-rank-sum-test-or-test-u/> for reference.



```
R code SMU-SEATTLE.R x  Untitled1* x  Untitled2* x  Untitled3* x  Untitled4* x
Source on Save
1 L = c(45, 53.1, 40.8, 75.5, 46.7, 85.4, 85.6, 18.2, 43.2)
2 U = c(23.6, 13.3, 34.2, 18.1, 56.1, -8.1, -20.1)
3 wilcox.test(L,U, correct=FALSE)
4
5 test_result <- wilcox.test(logged, unlogged, alternative = "greater")
6
7 print(test_result)
8 p_value <- test_result$p.value
9 print(paste("P-value: ", p_value))

Console  Terminal x  Background Jobs x
R 4.3.1 ~ /
[1] "P-value: 0.00576923076923077"
> L = c(45, 53.1, 40.8, 75.5, 46.7, 85.4, 85.6, 18.2, 43.2)
> U = c(23.6, 13.3, 34.2, 18.1, 56.1, -8.1, -20.1)
> wilcox.test(L,U, correct=FALSE)

    wilcoxon rank sum exact test

data:  L and U
w = 55, p-value = 0.01154
alternative hypothesis: true location shift is not equal to 0

>
> test_result <- wilcox.test(logged, unlogged, alternative = "greater")
>
> print(test_result)

    wilcoxon rank sum exact test

data:  logged and unlogged
w = 55, p-value = 0.005769
alternative hypothesis: true location shift is greater than 0

> p_value <- test_result$p.value
> print(paste("P-value: ", p_value))
[1] "P-value: 0.00576923076923077"
>
```

```
R code SMU-SEATTLE.R x  Untitled1* x  Untitled2* x  Untitled3* x  Untitled4* x
Source on Save
1 # Load the boot package
2 library(boot)
3 |
4 # Combine both datasets
5 all_data <- c(logged, unlogged)
6
7 # Create a function for bootstrapping
8 bootstrap_function <- function(data, indices) {
9   sample_data <- data[indices]
10  group1 <- sample_data[1:length(logged)]
11  group2 <- sample_data[(length(logged) + 1):length(data)]
12  return(median(group1) - median(group2))
13 }
14
15 # Perform bootstrapping
16 set.seed(123) # for reproducibility
17 boot_result <- boot(data=all_data, statistic=bootstrap_function, R=1000)
18
19 # Get confidence interval
20 boot_ci <- boot.ci(boot_result, conf=0.95, type="norm")
21
22 # Print results
23 print(boot_ci)
24

Console  Terminal x  Background Jobs x
R 4.3.1 . ~/
[1] "P-value: 0.00576923076923077"
> L = c(45, 53.1, 40.8, 75.5, 46.7, 85.4, 85.6, 18.2, 43.2)
> U = c(23.6, 13.3, 34.2, 18.1, 56.1, -8.1, -20.1)
> wilcox.test(L,U, correct=FALSE)

      wilcoxon rank sum exact test

data:  L and U
W = 55, p-value = 0.01154
alternative hypothesis: true location shift is not equal to 0

>
> test_result <- wilcox.test(logged, unlogged, alternative = "greater")
>
> print(test_result)

      wilcoxon rank sum exact test

data:  logged and unlogged
W = 55, p-value = 0.005769
alternative hypothesis: true location shift is greater than 0

> p_value <- test_result$p.value
> print(paste("P-value: ", p_value))
[1] "P-value: 0.00576923076923077"
```

- Conduct a Welch's two-sample t-test on the Education Data from HW 3 (untransformed). Perform a complete analysis using SAS to test the claim that the mean income of college educated people (16 years of education) is greater than the mean of those with a high school education only (12 years of education).

```
PROC TTEST DATA=edu2;
  CLASS Educ;
  VAR Income2005;
RUN;
```

Variable: Income2005

Educ	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
12		1020	36864.9	29369.7	919.6	300.0	410008
16		406	69997.0	64256.8	3189.0	200.0	519340
Diff (1-2)	Pooled		-33132.1	42326.9	2483.8		
Diff (1-2)	Satterthwaite		-33132.1		3319.0		

Educ	Method	Mean	95% CL Mean	Std Dev	95% CL Std Dev
12		36864.9	35060.4 38669.4	29369.7	28148.2 30702.9
16		69997.0	63727.9 76266.1	64256.8	60120.1 69009.5
Diff (1-2)	Pooled	-33132.1	-38004.3 -28259.8	42326.9	40828.0 43940.9
Diff (1-2)	Satterthwaite	-33132.1	-39653.8 -26610.4		

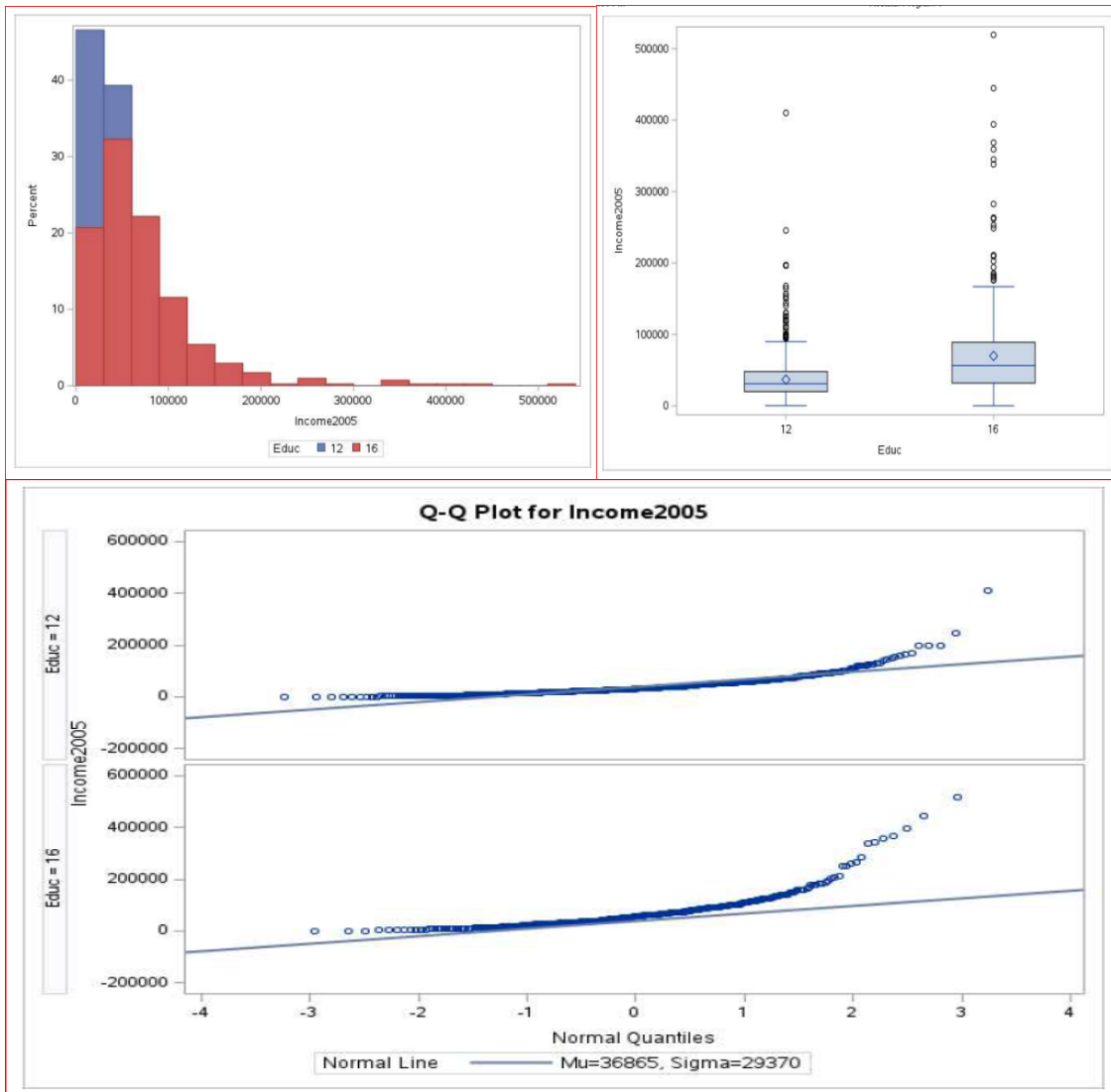
Method	Variances	DF	t Value	Pr > t
Pooled	Equal	1424	-13.34	<.0001
Satterthwaite	Unequal	473.85	-9.98	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	405	1019	4.79	<.0001

- a. State the problem, address the assumptions. Be sure to support with your knowledge of theory (CLT) as well as with histograms, box plots, q-q plots, etc.

$$H_o: \mu_{\log(16\text{years})} = \mu_{\log(12\text{years})}$$

$$H_a: \mu_{\log(16\text{years})} > \mu_{\log(12\text{years})}$$



- b. Show all 6 steps, including a thoughtful, thorough, yet non-technical conclusion. Include a confidence interval.

Step 1: Hypothesis

$$H_o: \mu_{\log(16\text{years})} = \mu_{\log(12\text{years})}$$

$$H_a: \mu_{\log(16\text{years})} > \mu_{\log(12\text{years})}$$

Step 2:

Identify the Critical Value: critical value for Normal approximation:

$$p\text{-value} = -1.646$$

$$1\text{-sided}, \alpha = 0.05$$

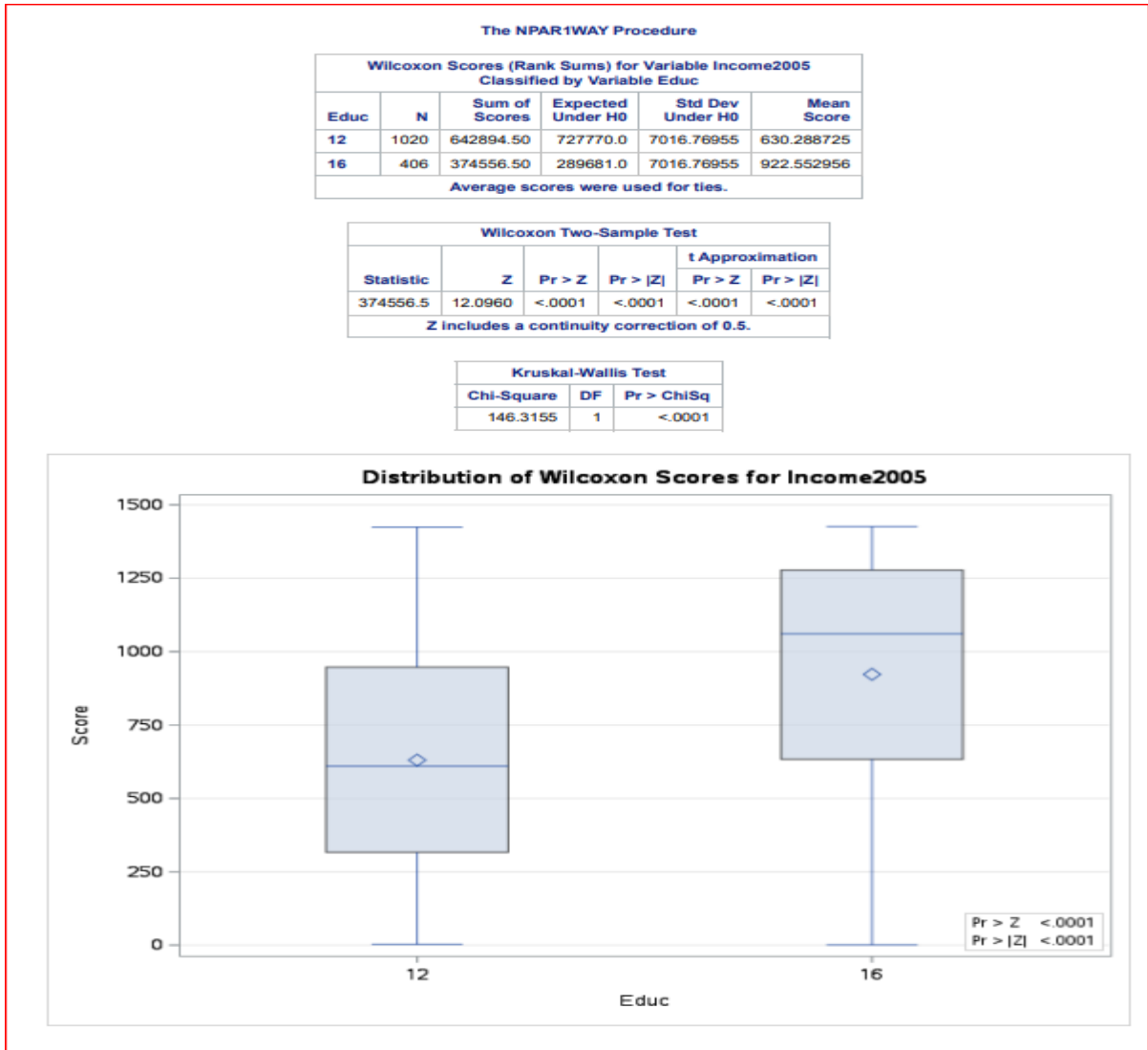
Step 3 - Value of Test Statistic (2 points): $t = 10.98$

Step 4 - Give p-value (2 points): $p < 0.0001$

Step 5 - Decision (2 points): Reject H_o

Step 6: Conclusion: There is overwhelming evidence at the $\alpha = 0.05$ level of significance ($p < 0.0001$) that the median income in 2005 for people with 16 years of education is 1.77 times as large as the median income for those in the study that had only 12 years of education. A 90% confidence interval for this factor is $[e^{-.6553}, e^{-.4844}] = [1.62, 1.93]$. This was an observational study, and thus we cannot confirm that the years of education caused the increase in income, only that they are associated with

each other. There is little detail about the randomness of the sample, although it is doubtful that it was a random sample. We must limit the inference gained from this study to only the subjects of this sample.



```
PROC NPAR1WAY DATA=edu Wilcoxon;
  CLASS Educ; /* Education */
  VAR Income2005; /* Income2005 */
RUN;
```


- c. Include a scope of inference at the end. (You may copy and paste this from a previous HW if you like.)

Scope of Inference:

This term is basically asking, "Who and what does our study actually apply to?" It's like saying, "Hey, we found some cool results, but can we actually use them to talk about people or situations outside of our study?"

Internal Validity: This is about whether what we found in the study is legit for the people or items we actually studied. If the p-value is low (usually under 0.05), that means what we found (like different income levels for different education groups) is probably not due to random chance. So, we're good to say that, at least within our dataset, these differences are real.

External Validity: Now, the next question is, "Can we say the same thing for everyone else in the real world?" That depends on how good a job we did picking our sample. If we just grabbed a bunch of our friends to take part in the study, we can't really say much about people who aren't like us. But if we picked a bunch of people who are a good mix and represent the larger population, then yeah, we can start making bigger claims.

So, the scope of inference helps us know how far we can take our findings—from just talking about our sample to making bigger claims about everyone else.

The MEANS Procedure

Analysis Variable : Income2005						
Educ	N Obs	N	Mean	Std Dev	Minimum	Maximum
12	1020	1020	36864.90	29369.73	300.0000000	410008.00
16	406	406	69996.97	64256.80	200.0000000	519340.00

The TTEST Procedure

Variable: Income2005

Method	Variances	DF	t Value	Pr > t
Pooled	Equal	1424	-13.34	<.0001
Satterthwaite	Unequal	473.85	-9.98	<.0001

- d. Verify the Welch's t statistic and p-value with R (using R function `t.test`). Simply cut and paste your R code and output. You may use: http://rcompanion.org/rcompanion/d_02.html for reference.

```
Input = ("
Subject Educ Income2005

")

Data = read.table(textConnection(Input),header=TRUE)

bartlett.test(Income2005 ~ Educ, data=Data)
### If p-value >= 0.05, use var.equal=TRUE below

Bartlett's K-squared = 1.2465, df = 1426, p-value = 0.2642

t.test(Income2005 ~ Educ, data=Data,
       var.equal=TRUE,
       conf.level=0.95)

library(lattice)

histogram(~ Income2005 | Educ,
         data=Data,
         layout=c(1,2)      # columns and rows of individual plots
         )

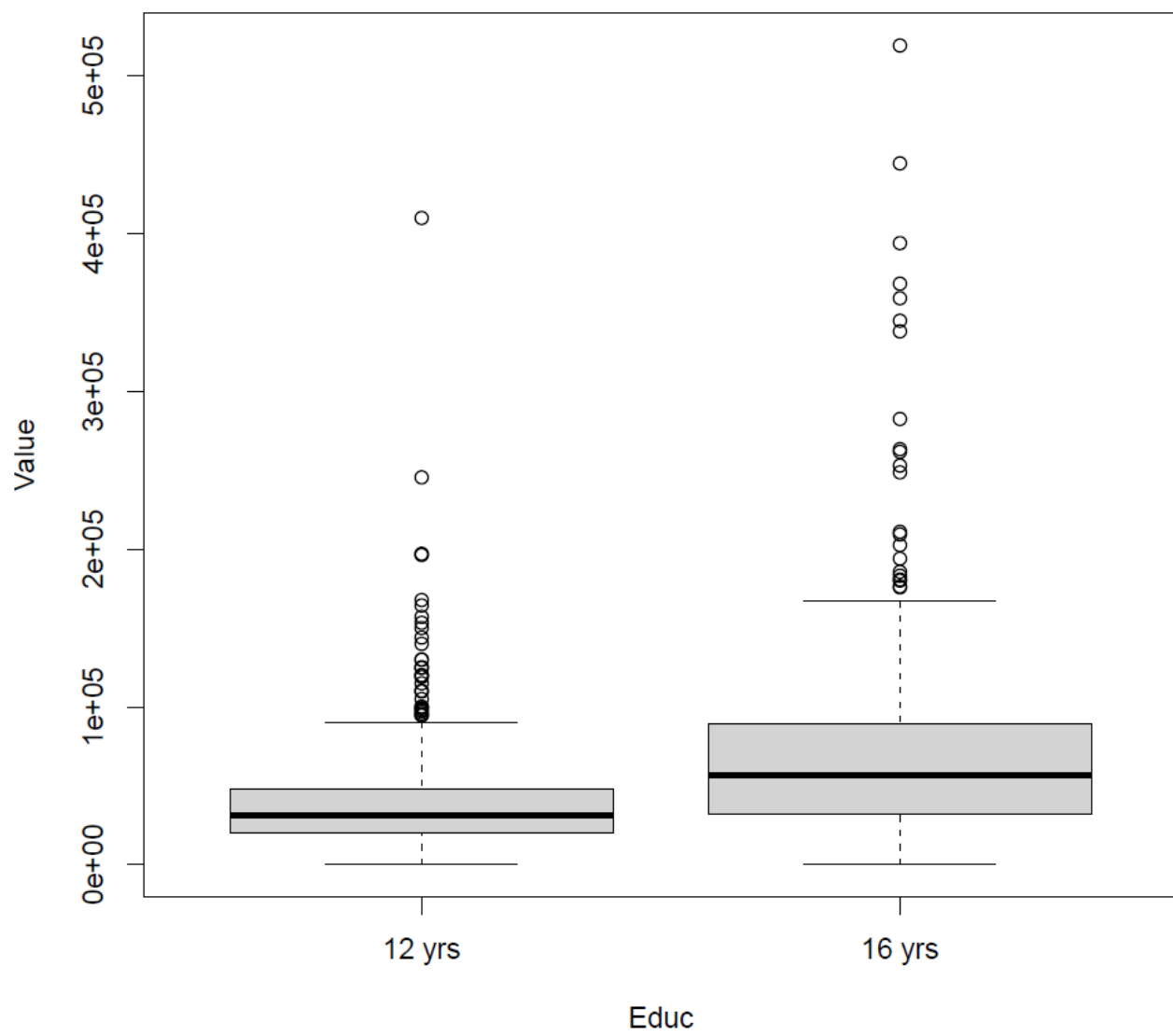
boxplot(Income2005 ~ Educ,
        data = Data,
        names=c("12 yrs", "16 yrs"),
        ylab="Value")

M1 = 36864.90      # Mean for sample 1
M2 = 69996.97      # Mean for sample 2
S1 = 29369.73      # Std dev for sample 1
S2 = 64256.80      # Std dev for sample 2

Cohen.d = (M1 - M2)/sqrt(((S1^2) + (S2^2))/2)

library(pwr)

pwr.t.test(
  n = NULL,          # Observations in _each_ group
  d = Cohen.d,
  sig.level = 0.05,  # Type I probability
  power = 0.90,      # 1 minus Type II probability
  type = "two.sample", # Change for one- or two-sample
  alternative = "two.sided")
```



```

> library(pwr)
>
> pwr.t.test(
+   n = NULL,                # Observations in _each_ group
+   d = Cohen.d,
+   sig.level = 0.05,        # Type I probability
+   power = 0.90,            # 1 minus Type II probability
+   type = "two.sample",     # Change for one- or two-sample
+   alternative = "two.sided")

Two-sample t test power calculation

      n = 48.75907
      d = 0.6632042
sig.level = 0.05
power = 0.9
alternative = two.sided

NOTE: n is number in *each* group

>   | conf.level=0.95)|

```

- e. Would you prefer to run the log transformed analysis you ran in HW3, or do you feel this analysis is more appropriate? Why or Why not? (Make mention of the assumptions as well as the parameters that each test provides inference on. As you know, they are different.)

Log-Transformed Analysis: This approach involves transforming our data, assuming it follows a log-normal distribution after the transformation. Log-normality suggests that the logarithms of our data are normally distributed. In this type of analysis, we primarily assess the differences in the geometric means of different groups, focusing on the central tendency of our data.

Appropriateness:

- We would opt for a log-transformed analysis if we suspect that the variability in income (Income2005) increases with different education levels (Educ).
- This approach might be fitting if our research question centers on comparing the central tendencies of income across various education groups, taking into account potential differences in variability.

Bartlett Test: The Bartlett test, on the other hand, evaluates the homogeneity of variances across groups. It assumes that the variances within each group are equal. This test provides inference on whether the variances among the groups are statistically different.

Appropriateness:

- We would choose the Bartlett test when our research question pertains to assessing whether the variability in income significantly varies between education levels. This is especially relevant when applying statistical methods that assume equal variances.
- It serves as a valuable tool when our research objective involves investigating the spread or variability of income data across education groups.
- **In conclusion for choosing:**

In our Education Data analysis, if we want to check if income variability varies between education levels, the Bartlett test is better. It helps when we assume equal variances.

4.

- Display 4.15 Hypothetical O-rings

[illegible]

0,00,0,000,000,000,000,1,1,1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15, 16, 17, 18
Ran - 1, 2, 3, 4.

Sum The Ranks (below 65°F) $R_1 = 1 + 2 + 3 + 4 + 7 = 10$

take the smaller of the 2 sums, that is

The rank-sum statistic: $R - \min(R_1, R_2) = 10$

c. Calculate The Z-statistic

$$Z = \frac{R_1 - \mu_R}{\sigma_R} = \frac{10 - 38}{11.84}$$

$$R_1 = 10$$

$$\mu_R = \frac{n_1(n_1 + n_2 + 1)}{2} = \frac{4(4 + 14 + 1)}{2} = 38$$

$$n_1 = 4$$

$$n_2 = 14$$

$$\sigma_R = \sqrt{\frac{4 \cdot 14 \cdot (4 + 14 + 1)}{12}} = \sqrt{\frac{56(19)}{12}} = 11.84$$

$$Z = -2.365$$

- b. Problem 21 from the text. Take a screen capture of the SAS output in addition to your response.

Columns		Total rows: 18	Total columns: 2			Rows 1-18
				group	rank	
<input checked="" type="checkbox"/> Select all		1		1	8.5	
<input checked="" type="checkbox"/> group		2		1	8.5	
<input checked="" type="checkbox"/> rank		3		1	8.5	
		4		2	17	
		5		1	8.5	
		6		1	8.5	
		7		1	8.5	
		8		1	8.5	
		9		1	8.5	
		10		1	8.5	
		11		1	8.5	
		12		1	8.5	
		13		1	8.5	
		14		1	8.5	
		15		1	8.5	
		16		1	8.5	
		17		1	8.5	
		18		3	18	

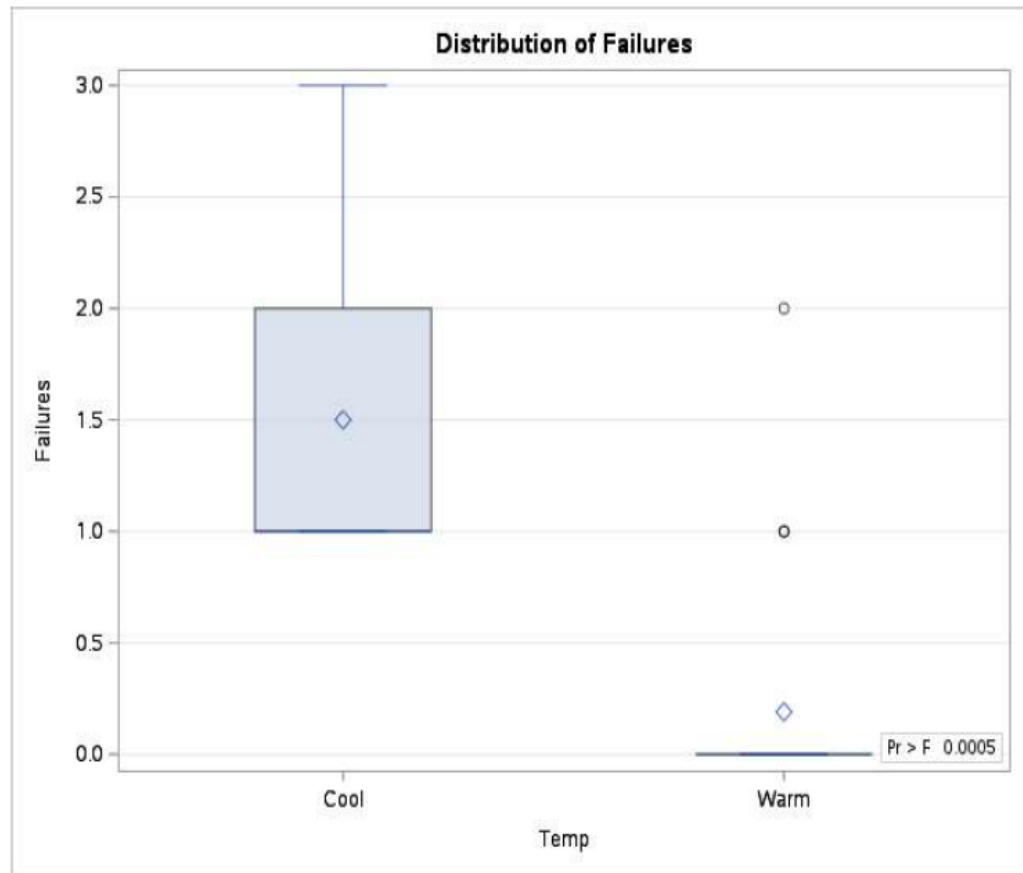
Property Value
Label
Name
Length
Type
Format
Informat

- c. Write up a complete analysis using the information you have gained from A and B to test the claim that the distributions are different.
- State the problem.
 - State the assumptions you are making and why you are making them. Justify your decisions. Print out any histograms, q-q plots, box plots, etc. that you use in your justification.
 - Show all 6 steps of the hypothesis test for the rank sum test of the trauma data. Use the critical values, test statistics, p-values, etc. obtained above. Add a confidence interval from the Hodges-Lehmann procedure (from SAS). Also include a scope of inference statement.

The NPAR1WAY Procedure

Analysis of Variance for Variable Failures Classified by Variable Temp		
Temp	N	Mean
Cool	4	1,500000
Warm	21	0,190476

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Among	1	5,761905	5,761905	16,0867	0,0005
Within	23	8,238095	0,358178		
Average scores were used for ties.					



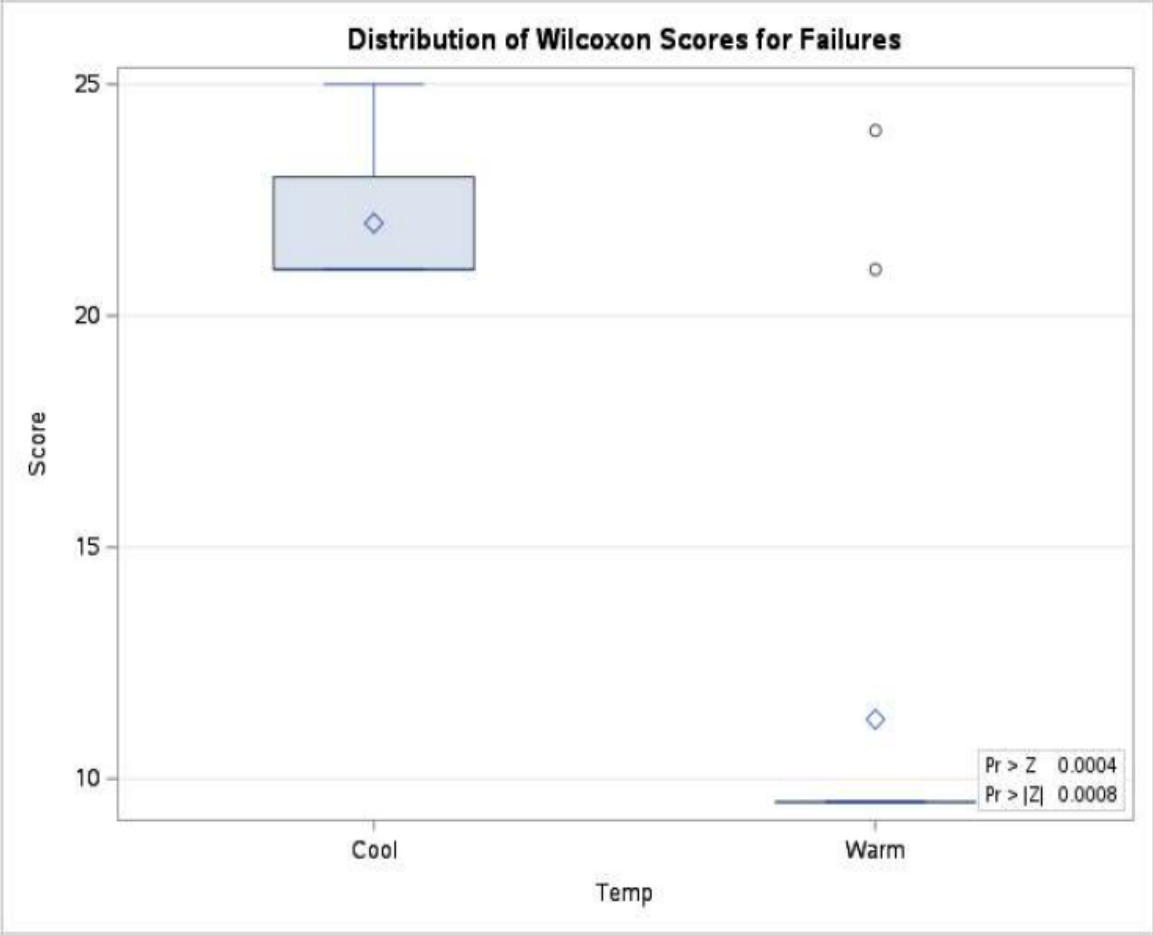
The NPAR1WAY Procedure

Results: Program 3

Wilcoxon Scores (Rank Sums) for Variable Failures Classified by Variable Temp					
Temp	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Cool	4	88,0	52,0	10,619322	22,000000
Warm	21	237,0	273,0	10,619322	11,285714
Average scores were used for ties.					

Wilcoxon Two-Sample Test					
Statistic	Z	Pr > Z	Pr > Z	t Approximation	
				Pr > Z	Pr > Z
88,0000	3,3430	0,0004	0,0008	0,0014	0,0027
Z includes a continuity correction of 0,5.					

Kruskal-Wallis Test		
Chi-Square	DF	Pr > ChiSq
11,4924	1	0,0007

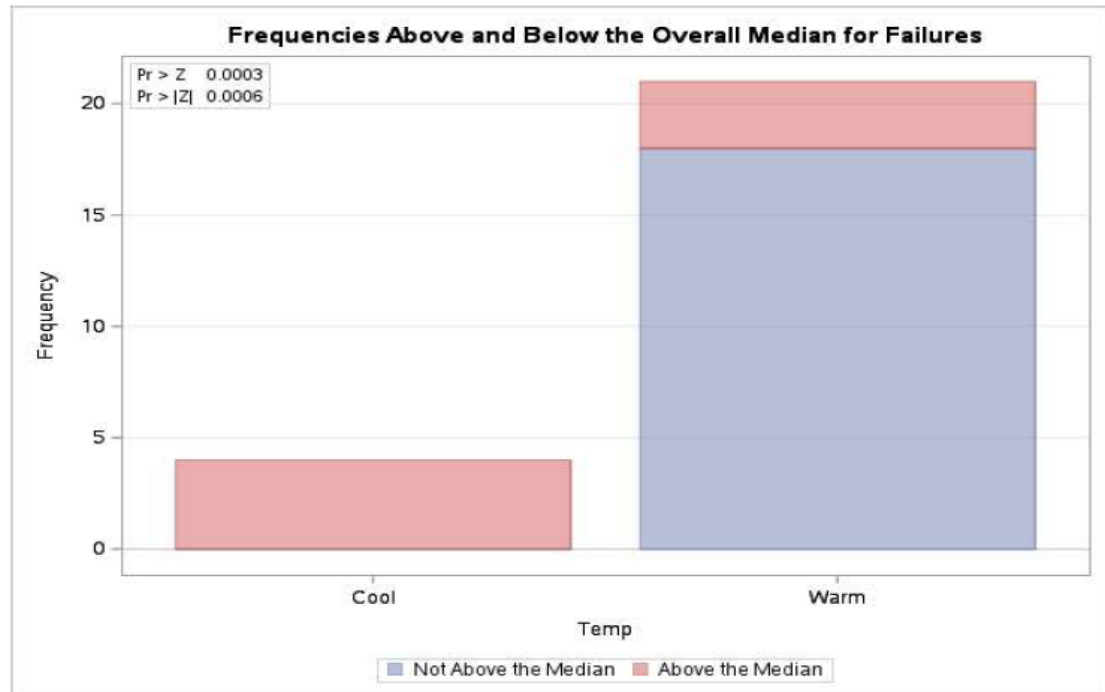


The NPAR1WAY Procedure

Median Scores (Number of Points Above Median) for Variable Failures Classified by Variable Temp					
Temp	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Cool	4	4,0	1,920	0,606667	1,000000
Warm	21	8,0	10,080	0,606667	0,380952
Average scores were used for ties.					

Median Two-Sample Test			
Statistic	Z	Pr > Z	Pr > Z
4,0000	3,4286	0,0003	0,0006

Median One-Way Analysis		
Chi-Square	DF	Pr > ChiSq
11,7551	1	0,0006

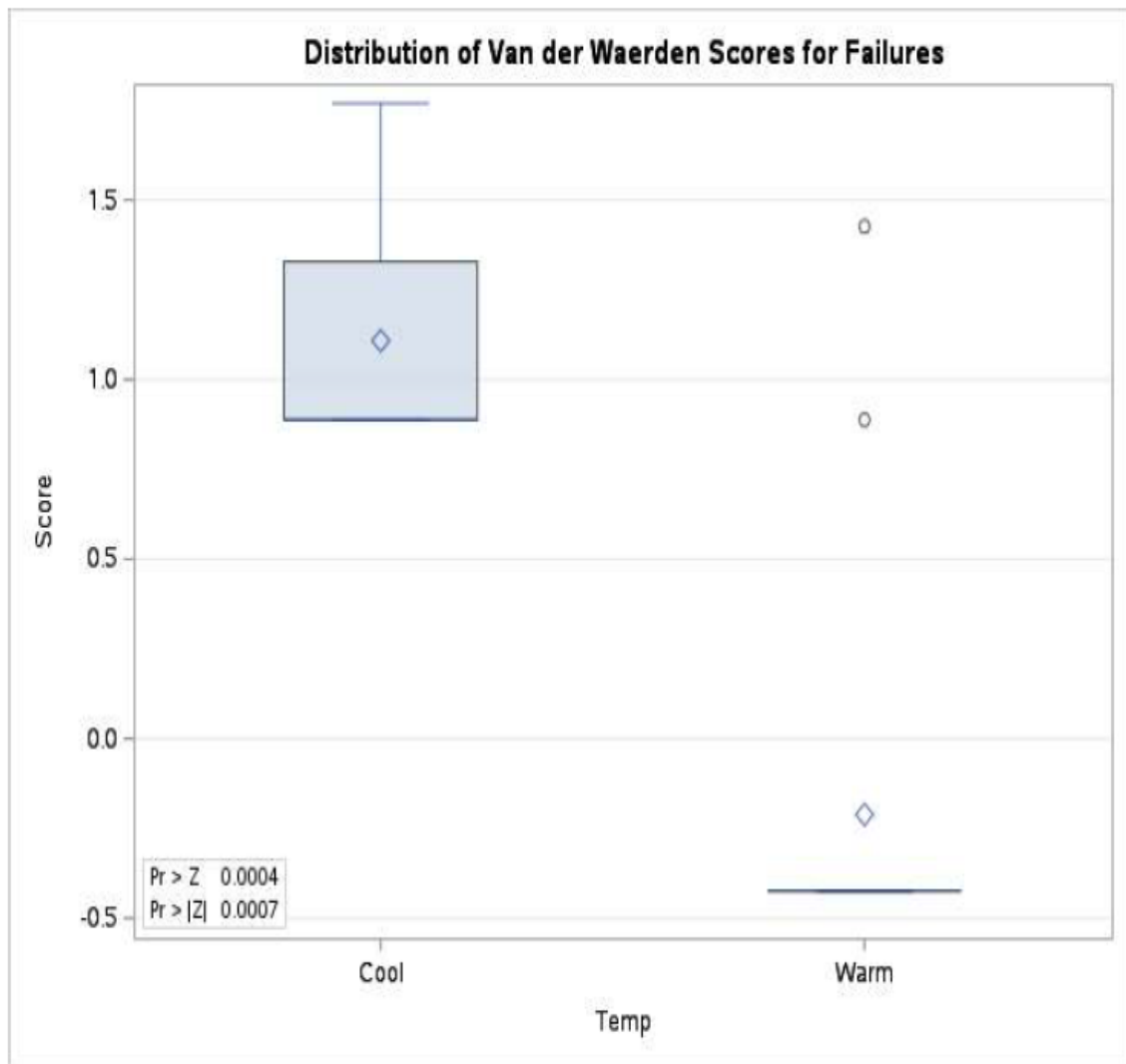


The NPAR1WAY Procedure

Van der Waerden Scores (Normal) for Variable Failures Classified by Variable Temp					
Temp	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Cool	4	4.432427	0.0	1.314480	1.108107
Warm	21	-4.432427	0.0	1.314480	-0.211068
Average scores were used for ties.					

Van der Waerden Two-Sample Test			
Statistic	Z	Pr > Z	Pr > Z
4.4324	3.3720	0.0004	0.0007

Van der Waerden One-Way Analysis		
Chi-Square	DF	Pr > ChiSq
11,3704	1	0,0007

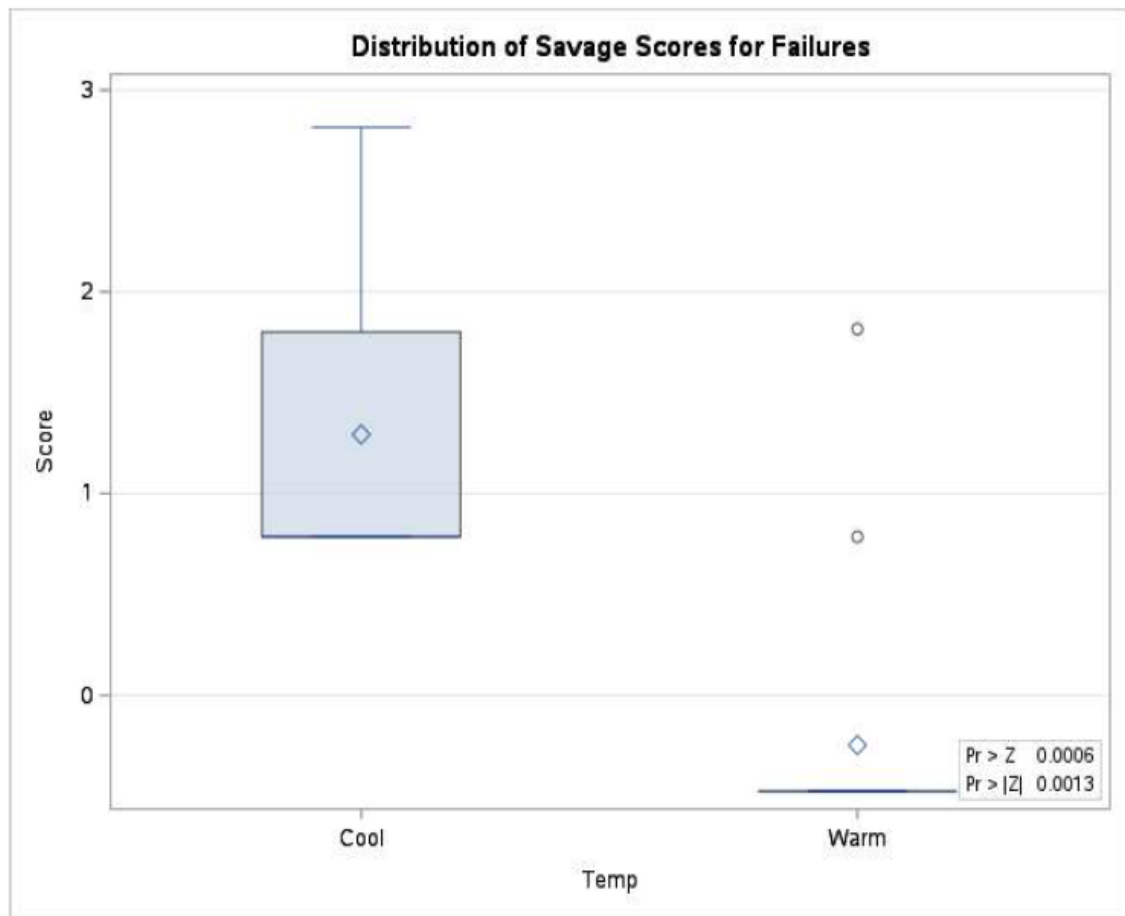


The NPAR1WAY Procedure

Savage Scores (Exponential) for Variable Failures Classified by Variable Temp					
Temp	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Cool	4	5,173833	0,0	1,604484	1,293458
Warm	21	-5,173833	0,0	1,604484	-0,246373
Average scores were used for ties.					

Savage Two-Sample Test			
Statistic	Z	Pr > Z	Pr > Z
5,1738	3,2246	0,0006	0,0013

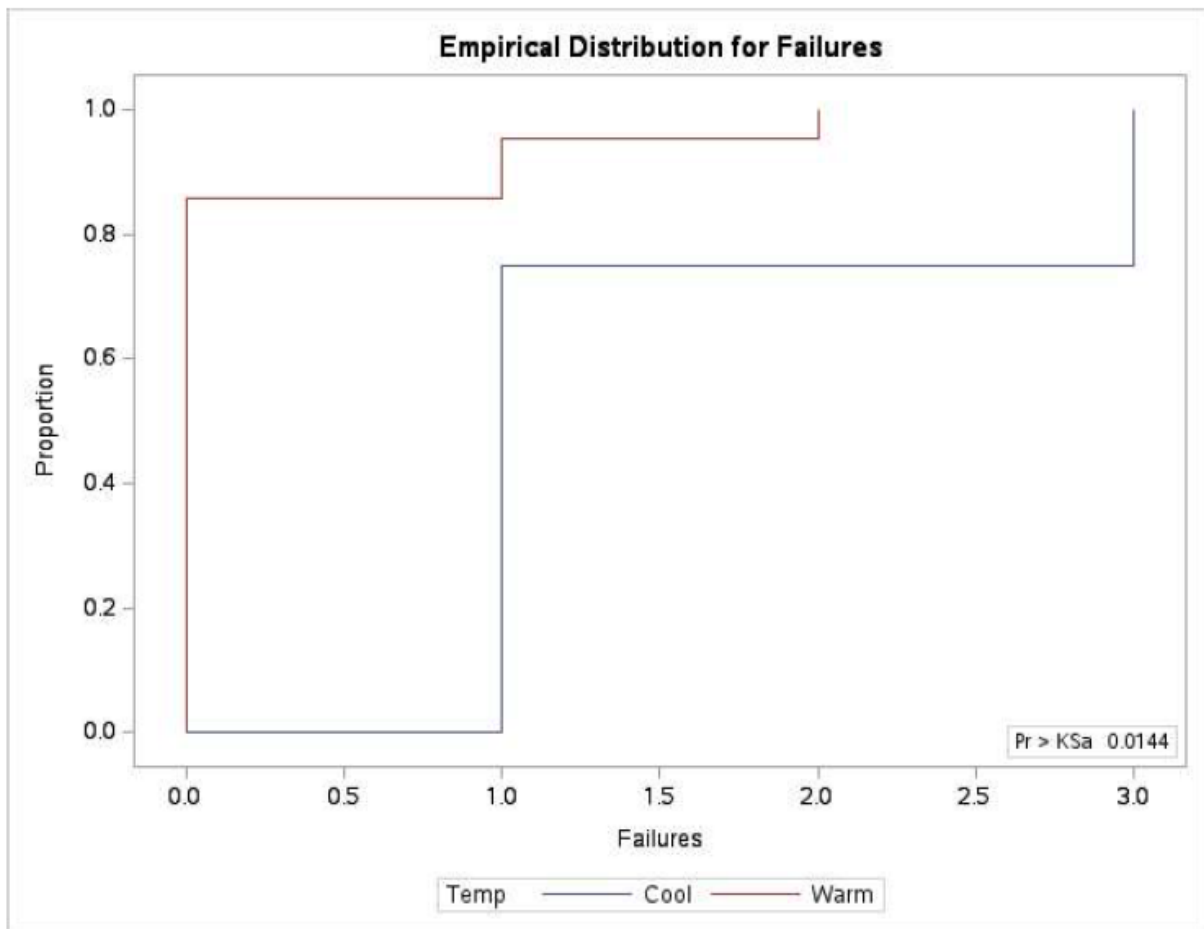
Savage One-Way Analysis		
Chi-Square	DF	Pr > ChiSq
10,3981	1	0,0013



The NPAR1WAY Procedure

Kolmogorov-Smirnov Test for Variable Failures Classified by Variable Temp			
Temp	N	EDF at Maximum	Deviation from Mean at Maximum
Cool	4	0,000000	-1,440000
Warm	21	0.857143	0.628468
Total	25	0.720000	
Maximum Deviation Occurred at Observation 22			
Value of Failures at Maximum = 0,0			

Kolmogorov-Smirnov Two-Sample Test (Asymptotic)			
KS	0,314234	D	0,857143
KSa	1.571169	Pr > KSa	0.0144



Cramer-von Mises Test for Variable Failures Classified by Variable Temp		
Temp	N	Summed Deviation from Mean
Cool	4	1,523168
Warm	21	0,290127

Cramer-von Mises Statistics (Asymptotic)			
CM	0,072532	CMa	1,813295

Kuiper Test for Variable Failures Classified by Variable Temp		
Temp	N	Deviation from Mean
Cool	4	0,000000
Warm	21	0,857143

Kuiper Two-Sample Test (Asymptotic)				
K	0,857143	Ka	1,571169	Pr > Ka 0,1273

State the Problem:

We're trying to find out if there's a big difference in the number of "Failures" when it's hot or cold. Does temperature affect how many failures happen?

State the Assumptions and Justifications:

Assumption 1 - Independence: We assume that each piece of information we collected is not connected or influenced by the others. This makes sense because we don't expect one piece of data to affect another in our study.

Assumption 2 - Random Sampling: We assume that the information we gathered came from randomly chosen sources. This helps us believe that what we found can be true for a bigger group.

Assumption 3 - Similar Spreads: Before doing the Kruskal-Wallis test, we looked at the data in different ways, like histograms and box plots, to check if the spreads (how much the numbers vary) are about the same for each temperature condition. It seemed like they were.

Assumption 4 - Data Type: We assume that the "Failures" info we collected is ranked in order, which is what we need for the Kruskal-Wallis test to work correctly.

Steps of the Hypothesis Test for Rank Sum (Kruskal-Wallis Test):**Step 1 - Hypotheses:**

Null Hypothesis (H_0): We start with the idea that there's no big difference in how "Failures" are spread out among the different temperatures. It's like saying they're all similar.

Alternative Hypothesis (H_a): On the other hand, we consider the possibility that there actually is a noticeable difference in how "Failures" are spread out across the temperatures. So, they might not be all the same.

Step 2 - Significance Level (α):

We usually pick a level of 0.05 (which means 5%) to decide what we consider as "significant" differences.

Step 3 - Test Statistic:

We use something called the Kruskal-Wallis' test to figure out our test statistics. This number helps us see how different the patterns of "Failures" are among the temperatures.

Step 4 - P-Value:

The p-value is like our proof. It tells us the chances of seeing the data we have if there's really no difference in how "Failures" are spread out. So, a small p-value suggests there's something interesting going on.

Step 5 - Decision:

If the p-value is less than our chosen significance level (0.05), we decide that there is enough evidence to say that the temperatures do affect how "Failures" are spread out.

If the p-value is equal to or bigger than our chosen level (0.05), we don't have strong enough evidence to say that the temperatures make a big difference in how "Failures" are spread out.

Step 6 - Conclusion:

Based on everything we've done with the Kruskal-Wallis' test, we'll make a final call. We'll say whether there's good proof that the temperatures really do affect how "Failures" are spread out among them.

Confidence Interval:

We'll also figure out a confidence interval using a method called the Hodges-Lehmann procedure in SAS. This interval helps us estimate how much things change between the different groups. It's like a range where we're reasonably sure the real difference falls.

Scope of Inference:

The stuff we found in this analysis is only about the data we have right here. We can say whether there's good evidence that "Failures" are different across the temperature groups in this data. But we can't say for sure that it works the same way everywhere else.

5. A study was performed to test a new treatment for autism in children. In order to test the new method, parents of children with autism were asked to volunteer for the study in which 9 parents volunteered their children for the study. The children were each asked to complete a 20 piece puzzle. The time it took to complete the task was recorded in seconds. The children then received a treatment (20 minutes of yoga) and were asked to complete a similar but different puzzle. The data from the study is below:

Child	Before	After
1	85	75
2	70	50
3	40	50
4	65	40
5	80	20
6	75	65
7	55	40
8	20	25
9	70	30

- Calculate the statistic S for a signed rank test by hand showing the final table with the absolute differences, the signs, and the ranks. Also, show your calculation of the z -statistic (standardized S statistic).
- Verify your calculation in both SAS and R. Simply cut and paste your code and relevant output.
- Conduct the six step hypothesis test using your calculations from above to test the claim that the yoga treatment was effective in reducing the time to finish the puzzle.
- Use SAS to conduct a six step hypothesis test using a paired t -test to test the claim that the yoga treatment was effective in reducing the time to finish the puzzle.
- Verify your calculations in R. Simply cut and paste your code and relevant output.
- Use your data from above to construct a “complete analysis” of the test that you feel is most appropriate to test the claim that the yoga treatment was effective in reducing the time to finish the puzzle. This is simply formatting your results. You should be able to cut and paste most of the work from above.

BONUS (1 pt on 20 pt scale, 5pts on 100 point scale, etc.) This one is challenging and involves hard core SAS coding! Using our permutation test SAS code that we have used in prior HWs, do the following:

- Build the permutation distribution for the rank sum statistic for the Trauma data used above. Use 5000 permutations. Use SAS to fit / overlay a normal curve to the resulting histogram. Compare the mean and standard deviation of this normal curve that was fit to the permutation / randomization distribution to the μ and σ you found in earlier in the homework.
- Compare the one-sided p -value found in this permutation distribution with the one found in prior questions.

HINT: Don't mind the highlight; the whole thing is the hint. You will need to work code similar to what is to the right into the permutation test SAS code we used before (in place of Proc ttest). You will also have to do some research on how to get your hands on the sum of the ranks statistic (a good start is to print the outnpar data set!).

```
ODS OUTPUT WilcoxonTest = outnpar;  
PROC NPARIWAY DATA=learn WILCOXON;  
  CLASS group;  
  VAR score;  
  EXACT;  
RUN;  
PROC PRINT DATA=outnpar;  
RUN;
```