RMarkdown STAT 2218 Problem Set 3

Jack Ambery

Friday, Nov 11, 2023

Data: AntPreference.csv

A study (in Australia) was done to see what type of sandwiches ants preferred. There were four types of bread, three types of sandwich fillings (vegemite, peanut butter, and ham & pickles). Butter was put on some of the sandwiches. The study looked at different sandwiches, each was made using one type of bread, one type of filling and it may have had butter on it.

To conduct the experiment a randomly chosen sandwich was selected and a piece of the sandwich was left on the ground near an ant hill. After 3 minutes a jar was placed over the sandwich piece and the number of ants was counted. The process was repeated with a different type of sandwich, allowing time for ants to return to the hill after each trial.

The data set has 48 observations on the following 5 variables. This data set will be used for all 3 problems.

* Trial: Trial number
* Bread: Type of bread (Multigrain, Rye, White, or Wholemeal)
* Filling: Type of filling (HamPickles, PeanutButter, or Vegemite)
* Butter: Whether there was butter on the sandwich? no or yes
* Ants: The number of ants on the sandwich piece after 3 minutes

# Input data file below.

Import data into RStudio using the base option, and then cut and paste YOUR file directory in the read.csv command below. It may be different from what I have listed below since your files are located in a different directory/folder. Then you can attach the file.

Ant <- read.csv("C:/Users/jacka/OneDrive - fairfield.edu/Fall 2023/Statistics II/datasets/AntPreference.csv")  
attach(Ant)

# Problem 1 (40 points)

In this problem, we wish to examine if there is a difference in average number of ants based on type of bread.

# Part 1a: Exploratory Analysis

Get numerical and graphical summaries of the number of ants for the different types of bread. Based on this exploratory analysis does there appear to be a relationship between bread type and number of ants? Explain. [Include all code used.]

## Part 1a Code

# pairwise summaries  
tapply(Ants, Bread, summary)

## $MultiGrain  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 19.00 27.50 40.00 43.00 59.25 76.00   
##   
## $Rye  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 18.00 30.00 44.50 42.25 51.00 67.00   
##   
## $White  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 21.00 36.25 48.00 44.50 54.00 66.00   
##   
## $WholeWheat  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 21.00 35.50 42.00 44.25 57.25 65.00

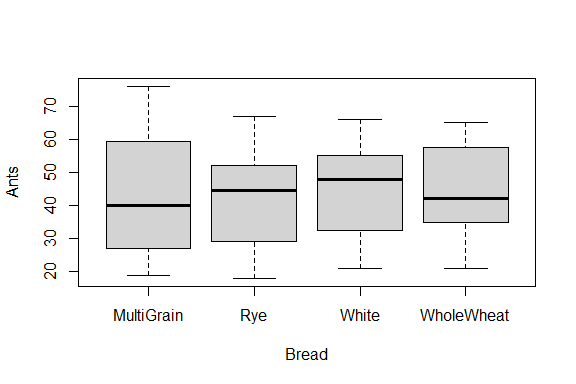
# pairwise standard deviations  
tapply(Ants, Bread, sd)

## MultiGrain Rye White WholeWheat   
## 18.23084 15.86377 14.60697 13.39691

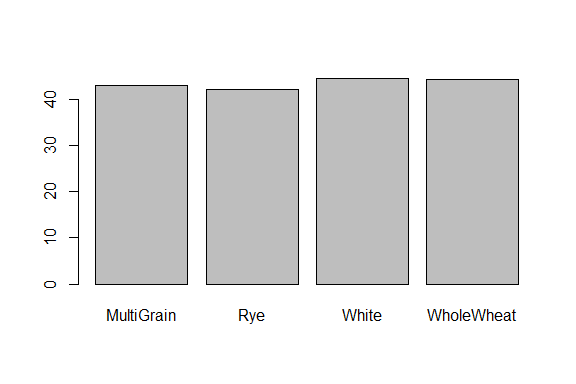
#pairwise length  
tapply(Ants, Bread, length) #all have the same length

## MultiGrain Rye White WholeWheat   
## 12 12 12 12

#plots  
boxplot(Ants~Bread)



barplot(tapply(Ants, Bread, mean))



## Part 1a Analysis

Based on this exploratory analysis does there appear to be a relationship between bread type and number of ants? Explain.

Upon this initial exploratory analysis, it does not immediately seem like there is a significant relationship between bread type and number of ants.

The summaries of each bread type indicate very similar means with White bread having a slightly higher mean and much higher median.

The pairwise standard deviations show MultiGrain bread having the highest standard deviation.

The pairwise lengths show 12 trials in each bread-type group.

The pairwise boxplots help show that there may be a slightly higher relationship between the number of ants and White bread, although it is hard to tell if it will be significant.

It does not seem likely that there will be a relationship between bread type and number of ants.

# Part 1b: Hypothesis Test

At the 5% level, is there evidence of a difference in the average number of ants based on type of bread? Set up the null and alternative hypotheses and define the unknown parameters and summarize the results of the test in a sentence.

## Part 1b Code

results = aov(Ants ~ Bread)  
summary(results)

## Df Sum Sq Mean Sq F value Pr(>F)  
## Bread 3 40 13.5 0.055 0.983  
## Residuals 44 10746 244.2

## Part 1b Analysis

Be sure to give the null and alternative hypotheses and define the unknown parameters and summarize the results of the test in a sentence.

= Average Number of Ants found on MultiGrain bread.

= Average Number of Ants found on Rye bread.

= Average Number of Ants found on White bread.

= Average Number of Ants found on WholeWheat bread.

H: = = = ; The average number of ants found on each type of bread is equal.

H: At least one type of bread has a different average number of ants compared to the others.

F = 0.055 df = 3 p-value = 0.983

Since the one-way ANOVA test yields a p-value of 0.983, we do not reject H0.

At the 5% level, there is no evidence that any of the mean number of ants is different for any particular type of bread.

# Part 1c: Checking the Assumptions

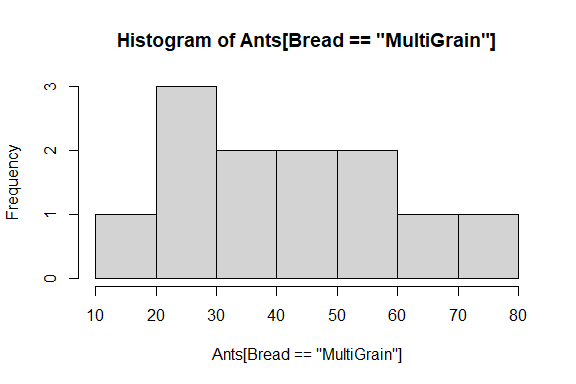
Check each of the assumptions (independence, normality and homogeneity of variance for this test) and state whether each condition is met or not. Use a significance level of 5%. Is it appropriate to use this test? Include relevant R codes.

## Part 1c Code

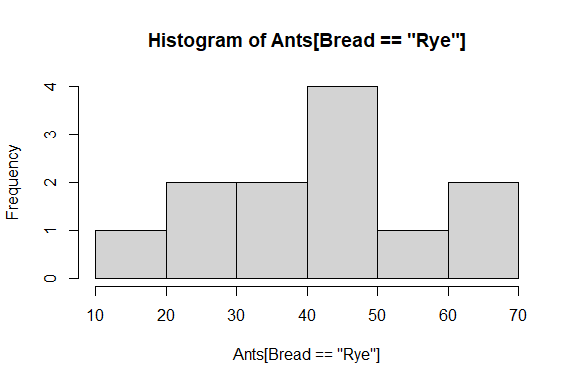
# Independence  
# no code needed, given by the design of the experiment  
  
# Normality  
tapply(Ants, Bread, shapiro.test)

## $MultiGrain  
##   
## Shapiro-Wilk normality test  
##   
## data: X[[i]]  
## W = 0.94815, p-value = 0.6101  
##   
##   
## $Rye  
##   
## Shapiro-Wilk normality test  
##   
## data: X[[i]]  
## W = 0.96239, p-value = 0.8174  
##   
##   
## $White  
##   
## Shapiro-Wilk normality test  
##   
## data: X[[i]]  
## W = 0.92831, p-value = 0.3625  
##   
##   
## $WholeWheat  
##   
## Shapiro-Wilk normality test  
##   
## data: X[[i]]  
## W = 0.96217, p-value = 0.8143

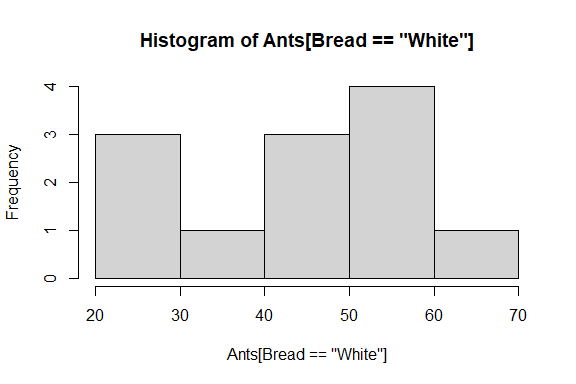
hist(Ants[Bread == 'MultiGrain'])



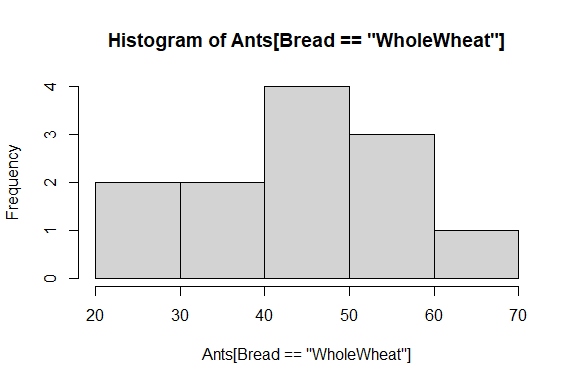
hist(Ants[Bread == 'Rye'])



hist(Ants[Bread == 'White'])



hist(Ants[Bread == 'WholeWheat'])



# Homogeneity of Variance   
library(car)

## Loading required package: carData

leveneTest(Ants, Bread)

## Warning in leveneTest.default(Ants, Bread): Bread coerced to factor.

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 3 0.5266 0.6663  
## 44

## Part 1c Analysis

State whether each condition is met or not. Use a significance level of 5%. Is it appropriate to use this test?

Independence

By design of the experiment, this assumption is met because the trials are not related to each other and each trial’s bread type was picked randomly.

Normality

The pairwise Shapiro-Wilk test tests the normality of each Bread type. Below are the null and alternative hypotheses.

H: The number of ants across the certain Bread type is normally distributed.

H: The number of ants across the certain Bread type is not normally distributed.

The results above show that the p-values of each pair were:

MultiGrain: 0.61

Rye: 0.82

White: 0.36

WholeWheat: 0.81

Since each group has a p-value above , 0.05, it is evident that at the 5% level, each group of Bread is normally distributed.

Some of the histograms of each bread type above can also help us visually confirm this, although some of them do not appear as normal due to the small size of the data sets.

Homogeneity of Variance:

For the next assumption, we check the equality of variances for each treatment. Below are the hypotheses.

= Variance of number of Ants found on MultiGrain bread.

= Variance of number of Ants found on Rye bread.

= Variance of number of Ants found on White bread.

= Variance of number of Ants found on WholeWheat bread.

H: = = = ; The variance of number of ants found on each type of bread is equal.

H: At least one type of bread has a different variance of number of ants compared to the others.

F = 0.527 df = 3 p-value = 0.67

Since the p-value is greater than 0.05, we have evidence that the variance of each group is equal.

It is appropriate to use this test because all assumptions are met.

# Part 1d: Post-hoc tests

Run both post-hoc tests, if necessary and interpret the results. Be sure to write out your conclusions explicitly. Use alpha = 5%. If it is not necessary to run the post-hoc tests explain why.

## Part 1d Code

# no code necessary

## Part 1d Analysis

Interpret the results of the post-hoc tests. Be sure to write out your conclusions explicitly. Use alpha = 5%. If it is not necessary to run the post-hoc tests explain why.

There is no need to run post-hoc tests for Problem #1 because our one-way ANOVA test indicates no differences in the average number of ants found on a certain bread type.

# Problem 2 (40 points)

In this problem, we wish to examine if there is a difference in the average number of ants based on type of filling.

# Part 2a: Exploratory Analysis

Get numerical and graphical summaries of the number of ants for the different types of filling. Based on this exploratory analysis does there appear to be a relationship between filling type and number of ants? Explain. [Include all code used.]

## Part 2a Code

# pairwise summaries  
tapply(Ants, Filling, summary)

## $HamPickles  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 34.0 47.0 58.5 55.5 65.0 76.0   
##   
## $PeanutButter  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 19.00 26.25 44.50 40.38 50.25 60.00   
##   
## $Vegemite  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 18.00 25.75 34.50 34.62 42.00 57.00

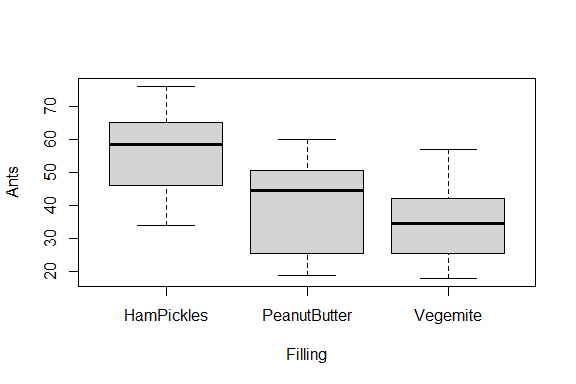
# pairwise standard deviations  
tapply(Ants, Filling, sd)

## HamPickles PeanutButter Vegemite   
## 12.05543 14.18391 11.15870

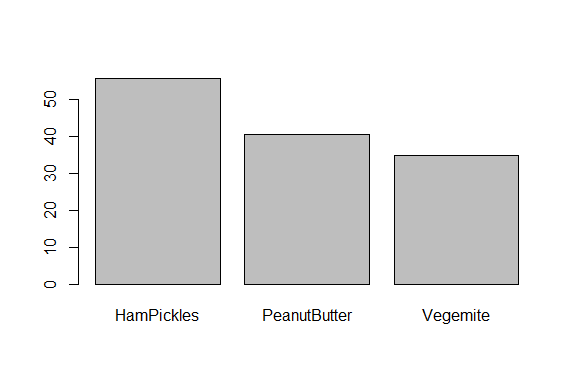
#pairwise length  
tapply(Ants, Filling, length)

## HamPickles PeanutButter Vegemite   
## 16 16 16

#plots  
boxplot(Ants~Filling)



barplot(tapply(Ants, Filling, mean))



## Part 2a: Analysis

Based on this exploratory analysis does there appear to be a relationship between filling type and number of ants? Explain.

Upon this initial exploratory analysis, it is extremely evident that there is a significant relationship between filling type and number of ants observed.

Right away, we can see in the numerical summary that Ham and Pickles has a mean of 55.5, PeanutButter has a mean of 40.2, and Vegemite has a mean of 34.6. The large difference in these means will likely prove to be significant in testing.

The pairwise standard deviations show PeanutButter filling having the highest standard deviation.

The pairwise lengths show 16 trials in each sandwich filling group.

The pairwise boxplots help show the significantly different average number of ants observed on each type of sandwich filling. The trial of the highest number of ants seen on PeanutButter and Vegemite are only about the average number of ants seen on HamPickles sandwiches.

It seems very likely that there will be a relationship between filling type and number of ants.

# Part 2b: Hypothesis Test

At the 5% level, is there evidence of a difference in the average number of ants based on type of filling? Set up the null and alternative hypotheses and define the unknown parameters and summarize the results of the test in a sentence.

## Part 2b Code

results = aov(Ants ~ Filling)  
summary(results)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Filling 2 3721 1860 11.85 7.35e-05 \*\*\*  
## Residuals 45 7065 157   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Part 2b Analysis

Be sure to give the null and alternative hypotheses and define the unknown parameters and summarize the results of the test in a sentence.

= Average Number of Ants found on HamPickles filling.

= Average Number of Ants found on PeanutButter filling.

= Average Number of Ants found on Vegemite filling.

H: = = = ; The average number of ants found on each type of filling is equal.

H: At least one type of filling has a different average number of ants compared to the others.

F = 11.85 df = 2 p-value = 7.35e-05

Since the one-way ANOVA test yields a p-value of 7.35e-05, we reject H0.

At the 5% level, there is evidence that at least one type of filling has a different mean number of ants observed compared to the other types of filling.

# Part 2c: Checking the Assumptions

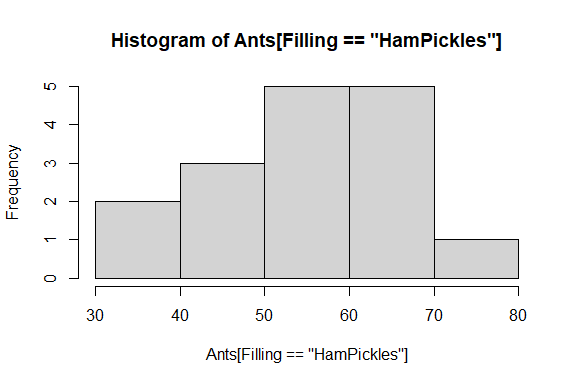
Check each of the assumptions (independence, normality and homogeneity of variance for this test) and state whether each condition is met or not. Use a significance level of 5%. Is it appropriate to use this test? Include relevant R codes.

## Part 2c Code

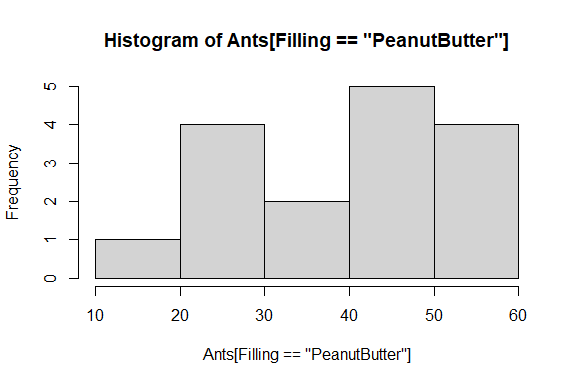
# Independence  
# no code needed, given by the design of the experiment  
  
# Normality  
tapply(Ants, Filling, shapiro.test)

## $HamPickles  
##   
## Shapiro-Wilk normality test  
##   
## data: X[[i]]  
## W = 0.95346, p-value = 0.5463  
##   
##   
## $PeanutButter  
##   
## Shapiro-Wilk normality test  
##   
## data: X[[i]]  
## W = 0.91595, p-value = 0.1451  
##   
##   
## $Vegemite  
##   
## Shapiro-Wilk normality test  
##   
## data: X[[i]]  
## W = 0.95393, p-value = 0.5545

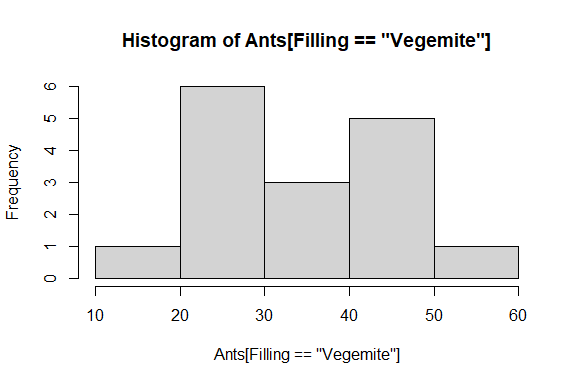
hist(Ants[Filling == 'HamPickles'])



hist(Ants[Filling == 'PeanutButter'])



hist(Ants[Filling == 'Vegemite'])



# Homogeneity of Variance   
library(car)  
leveneTest(Ants, Filling)

## Warning in leveneTest.default(Ants, Filling): Filling coerced to factor.

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 2 0.6345 0.5349  
## 45

## Part 2c Analysis

State whether each condition is met or not. Use a significance level of 5%. Is it appropriate to use this test?

Independence

By design of the experiment, this assumption is met because the trials are not related to each other and each trial’s filling type was picked randomly.

Normality

The pairwise Shapiro-Wilk test tests the normality of each Filling type. Below are the null and alternative hypotheses.

H: The number of ants across the certain Filling type is normally distributed.

H: The number of ants across the certain Filling type is not normally distributed.

The results above show that the p-values of each pair were:

HamPickles: 0.55

PeanutButter: 0.15

Vegemite: 0.55

Since each group has a p-value above , 0.05, it is evident that at the 5% level, each group of Filling is normally distributed.

The histograms of each Filling type also show that each distribution is approximately normal.

Homogeneity of Variance:

For the next assumption, we check the equality of variances for each treatment. Below are the hypotheses.

= Variance of Ants found on HamPickles filling.

= Variance of Ants found on PeanutButter filling.

= Variance of Ants found on Vegemite filling.

H: = = ; The variance of number of ants found on each type of filling is equal.

H: At least one type of filling has a different variance of number of ants compared to the others.

F = 0.635 df = 2 p-value = 0.53

Since the p-value is greater than 0.05, we have evidence that the variance of each group is equal.

Since every assumption has been met, it is appropriate to use this test.

# Part 2d: Post-hoc tests

Run both post-hoc tests, if necessary and interpret the results. Be sure to write out your conclusions explicitly. Use alpha = 5%. If it is not necessary to run the post-hoc tests explain why.

## Part 2d Code

# pairwise t tests  
pairwise.t.test(Ants, Filling, p.adjust.method = "bonferroni")

##   
## Pairwise comparisons using t tests with pooled SD   
##   
## data: Ants and Filling   
##   
## HamPickles PeanutButter  
## PeanutButter 0.0041 -   
## Vegemite 7.2e-05 0.6028   
##   
## P value adjustment method: bonferroni

#tukey's honest significant difference test  
TukeyHSD(results, conf.level = 0.95)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = Ants ~ Filling)  
##   
## $Filling  
## diff lwr upr p adj  
## PeanutButter-HamPickles -15.125 -25.86202 -4.387985 0.0038299  
## Vegemite-HamPickles -20.875 -31.61202 -10.137985 0.0000698  
## Vegemite-PeanutButter -5.750 -16.48702 4.987015 0.4037366

## Part 2d Analysis

Interpret the results of the post-hoc tests. Be sure to write out your conclusions explicitly. Use alpha = 5%. If it is not necessary to run the post-hoc tests explain why.

The post-hoc tests help us determine exactly which means are different. The first method I used was the pairwise t-test. Below are the hypotheses for this test. There are 3 hypothesis tests that occur. They are listed after H and H, respectively.

= Average Number of Ants found on HamPickles filling.

= Average Number of Ants found on PeanutButter filling.

= Average Number of Ants found on Vegemite filling.

H: = ; = ; =

H: ; ;

These are the resulting p-values:

HamPickles & PeanutButter: 0.0041

HamPickles & Vegemite: 7.2e-05

PeanutButter & Vegemite: 0.6028

P-values above mean the two means are not significantly different. Therefore, at the 5% level, there is evidence that the average number of ants observed on PeanutButter and Vegemite fillings is not significantly different. There is also evidence at the 5% level that the average number of ants observed on HamPickles and PeanutButter and the average number of ants observed on HamPickles and Vegemite are significantly different.

The second way to conduct the post-hoc tests is by Tukey’s Honest Significant Difference Test. This test, conducted above, produced the same conclusions as the pairwise t-tests. Tukey’s test uses the same exact hypotheses as the pairwise t-tests, stated above. Below are the resulting p-values:

HamPickles & PeanutButter: 0.0038299

HamPickles & Vegemite: 0.0000698

PeanutButter & Vegemite: 0.4037366

Like the pairwise t-tests, p-values above mean the two means are not significantly different. Therefore, at the 5% level, there is evidence that the average number of ants observed on PeanutButter and Vegemite fillings is not significantly different. There is also evidence at the 5% level that the average number of ants observed on HamPickles and PeanutButter and the average number of ants observed on HamPickles and Vegemite are significantly different.

# Problem 3 (30 points)

In this problem, we wish to examine if there is a difference in average number of ants based on whether the sandwich is buttered or not.

# Part 3a: Exploratory Analysis

Get numerical and graphical summaries of the number of ants for buttered and unbuttered sandwiches. Based on this exploratory analysis does there appear to be a relationship between number of ants and whether the sandwich is buttered? Explain. [Include all code used.]

## Part 3a Code

# pairwise summaries  
tapply(Ants, Butter, summary)

## $no  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 18.00 24.75 37.00 38.12 48.00 65.00   
##   
## $yes  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 22.00 40.75 49.00 48.88 59.25 76.00

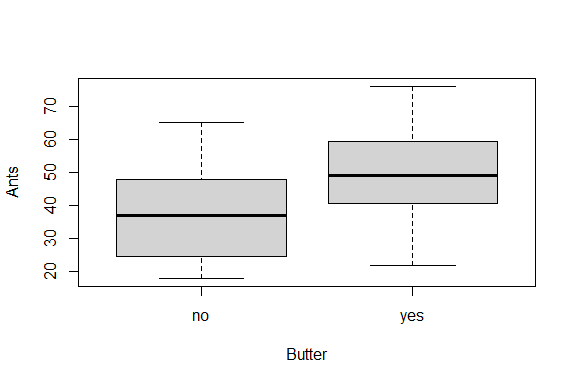
# pairwise standard deviations  
tapply(Ants, Butter, sd)

## no yes   
## 14.07221 14.51330

#pairwise length  
tapply(Ants, Butter, length)

## no yes   
## 24 24

#plots  
boxplot(Ants~Butter)



barplot(tapply(Ants, Butter, mean))



## Part 3a Analysis

Based on this exploratory analysis does there appear to be a relationship between the number of ants and whether the bread is buttered? Explain.

Upon this initial exploratory analysis, it is seems very likely that there is a significant relationship between the sandwiches having butter or not and number of ants observed.

The summary tells us that the mean number of ants for butter is 48.9 and the mean number of ants for no butter is 38.1.

The pairwise standard deviations show similar standard deviations between butter and no butter.

The pairwise lengths show 24 trials in each butter and no butter.

The pairwise boxplots help show the large difference in average number of ants observed on buttered sandwiches and non buttered sandwiches.

It seems very likely that there will be a relationship between butter or no butter and number of ants observed.

# Part 3b: Hypothesis Test

At the 5% level, is there evidence of a difference in the average number of ants based on whether the sandwich is buttered? Set up the null and alternative hypotheses and define the unknown parameters and summarize the results of the test in a sentence.

## Part 3b Code

# seeing whether to use a pooled or unpooled two sample t-test  
var.test(Ants ~ Butter)

##   
## F test to compare two variances  
##   
## data: Ants by Butter  
## F = 0.94014, num df = 23, denom df = 23, p-value = 0.8836  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 0.4066981 2.1732654  
## sample estimates:  
## ratio of variances   
## 0.9401398

# pooled two sample t-test  
t.test(Ants~Butter, mu = 0, var.equal = TRUE, alternative = "two.sided")

##   
## Two Sample t-test  
##   
## data: Ants by Butter  
## t = -2.6051, df = 46, p-value = 0.01233  
## alternative hypothesis: true difference in means between group no and group yes is not equal to 0  
## 95 percent confidence interval:  
## -19.056122 -2.443878  
## sample estimates:  
## mean in group no mean in group yes   
## 38.125 48.875

## Part 3b Analysis

Be sure to give the null and alternative hypotheses and define the unknown parameters and summarize the results of the test in a sentence. Explain what test you used. (Think carefully about this!)

To test if there is a difference in ants observed on buttered sandwiches and non buttered sandwiches, I used a two sample t-test, since an ANOVA test is not necessary. I first ran var.test to determine whether to use a pooled or unpooled test. Below is the variance test:

= Variance of Ants found on buttered sandwiches.

= Variance of Ants found on non buttered sandwiches.

H: = ; The variance of number of ants found on each type of sandwich is equal.

H: ; The variances are different.

F = 0.94 df = 23 p-val = 0.88

Since the p-value is 0.88 and above , we have evidence that the variances are equal and know to use a pooled two sample t-test. Below are the hypotheses for the test:

= Average Number of Ants found on buttered sandwiches.

= Average Number of Ants found on non buttered sandwiches.

H: = ; The average number of ants found on buttered sandwichs is equal to the average number of ants observed on non buttered sandwiches.

H: ; The average number of ants observed on each sandwich type is different.

t = -2.61 df = 46 p-val = 0.012

At the 5% level, there is evidence to support that or the average number of ants observed on buttered sandwiches differs than that observed on non buttered sandwiches.

# Part 3c: Checking the Assumptions

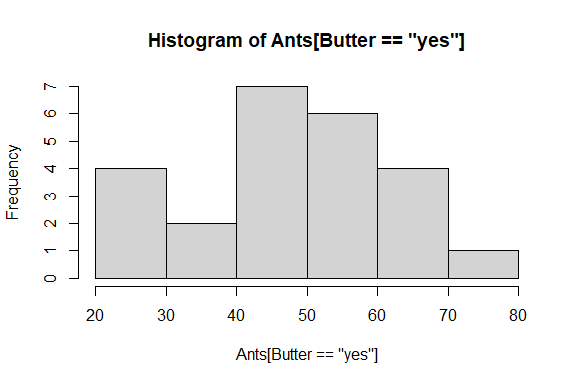
Check each of the assumptions for the test. Use a significance level of 5%. Is it appropriate to use this test? Include relevant R codes.

## Part 3c Code

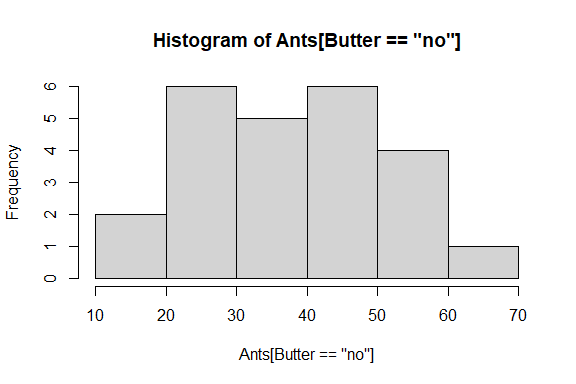
# 5 - Each pop distribution is approximately normal  
tapply(Ants, Butter, shapiro.test)

## $no  
##   
## Shapiro-Wilk normality test  
##   
## data: X[[i]]  
## W = 0.95174, p-value = 0.2953  
##   
##   
## $yes  
##   
## Shapiro-Wilk normality test  
##   
## data: X[[i]]  
## W = 0.97004, p-value = 0.6679

hist(Ants[Butter == 'yes'])



hist(Ants[Butter == 'no'])



## Part 3c Analysis

State whether each condition of the test is met or not. Use a significance level of 5%. Is it appropriate to use this test?

1 - Two independent populations with unknown means and .

This assumption is given by the problem statement. We know that the populations are independent and have unknown means.

2 - Assumed the two population standard deviations are normal

This condition was proved correct earlier in the problem when I used var.test.

3 - Samples from each sample are randomly selected

This condition is also given by the problem statement. Random anthills were assigned random sandwiches.

4 - and are < 10% of their respective population sizes.

It is safe to assume that and , which are each 24, are less than the number of anthills in the world.

5 - Each population distribution is approximately normal.

This condition is proved using the Shapiro-Wilk Test above as well as looking at histograms which are above as well.

Since the Shapiro test yields p-values of 0.3 and 0.67, there is evidence at the 5% level that each population is normally distributed. The histograms for butter and no butter also support a normal distribution in each.

Since each condition of the test is met, it can be said that this is an appropriate test to use.

End of Problem Set