Design of Experiments Final Project

Alexis Athens

4/23/2018

### Final Project: Waste-Less UF

## A Study of UF Students’ Carbon Footprints on Campus

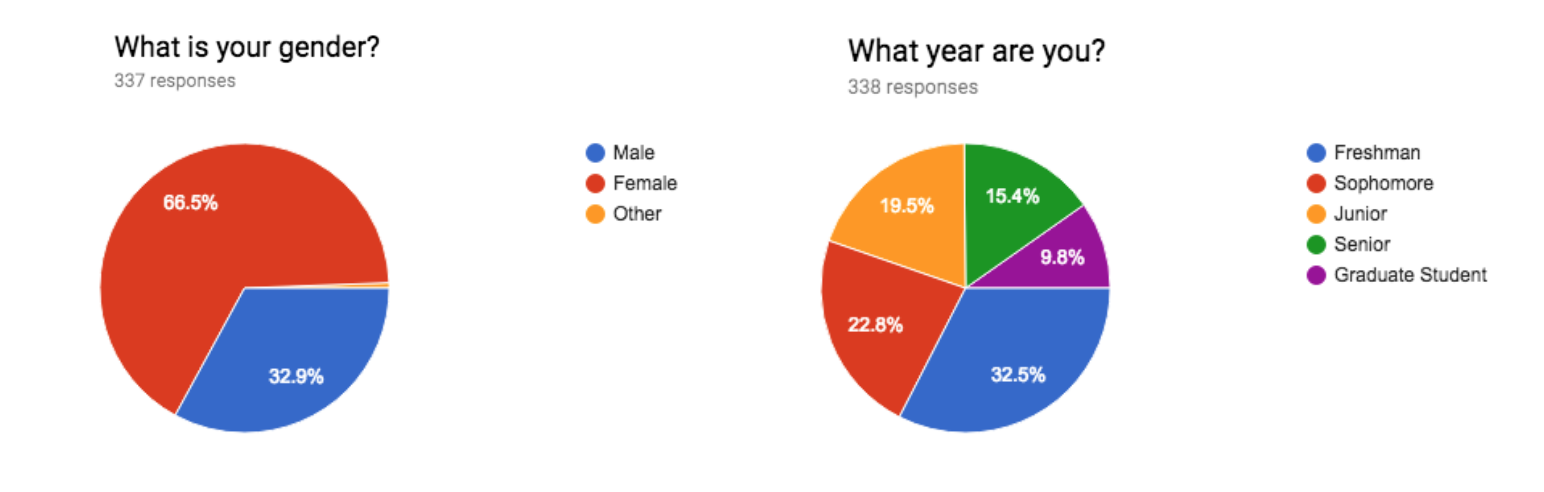
# Motivation for the project

For my climate change science final project, I proposed an system that incentivized students to bring their own reusable items to campus in place of plastic/paper consumables in attemp to reduce UF’s carbon footprint. In order to quantify the predicted decrease in UF’s footprint, I polled UF students for the average amount of plastic/paper consumables they throw away or recycle weekly, which was used as the response variable in this study.

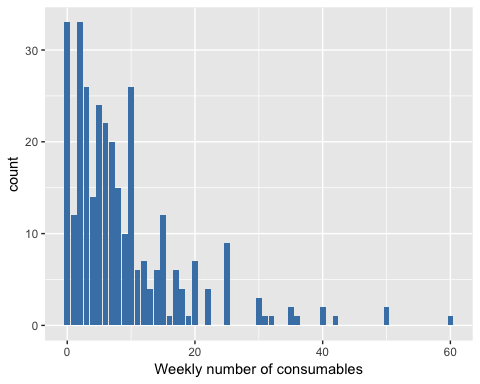
# Objective of the study

For this project, I asked the respondents demographic questions (gender, year in school) to try to find if these factors affect the amount of consumables used weekly. I will test whether these are significant predictors of the amount of consumables used weekly. RQ1: Does gender affect the amount of plastic and paper consumables thrown away on campus? RQ2: Does year in school affect the amount of plastic and paper consumables thrown away on campus?

# Method of data collection

The study was conducted by anonymous polling via social media over a few days with 338 total responses. Data was cleaned upon initial collection. 

#import and further clean data  
rm(list=ls(all=TRUE))  
  
library(plyr); library(dplyr)  
library(ggplot2)  
library(car)  
  
dat <- read.csv("project/CCDataApr10.csv",  
 col.names=c("gender","year","plastic.b","paper.b","willing","plastic.a","paper.a","credit"))  
#View(dat)  
  
#add additional columns  
dat <- mutate(dat, waste.b=plastic.b+paper.b) #total weekly waste before  
dat <- mutate(dat, waste.a=plastic.a+paper.a) #total weekly waste after  
dat <- mutate(dat, diff=waste.b-waste.a) #total weekly waste after  
  
#clean data  
#remove after values larger than before values  
dat <- dat[-which(dat$diff<0),]  
  
#total sample size  
nt=dim(dat)[1] #316 valid responses  
  
#rephrase credit variable  
dat$credit <- mapvalues(dat$credit, from = c("Redeem additional reusable items for personal use",   
 "Fund an on-campus organization of your choice"),   
 to = c("Personal", "Club"))  
  
attach(dat)  
  
#order factors  
gender <- factor(gender, levels=c("Male","Female"))  
year <- factor(year, levels=c("Freshman","Sophomore","Junior","Senior","Graduate Student"))  
willing <- factor(willing, levels=c("Yes","Sometimes","No"))  
credit <- factor(credit,levels=c("Personal","Club"))  
  
a <- length(levels(gender)) #2 levels  
b <- length(levels(year)) #5 levels  
  
ggplot(data=dat,aes(x=waste.b,color=waste.b)) + geom\_bar(fill="steelblue") + labs(x="Weekly number of consumables")

 See back page for data set.

# Model and assumptions

For my study, I used an unbalanced 2-factor model with levels A=gender and B=year with a transormed response: and for , , and

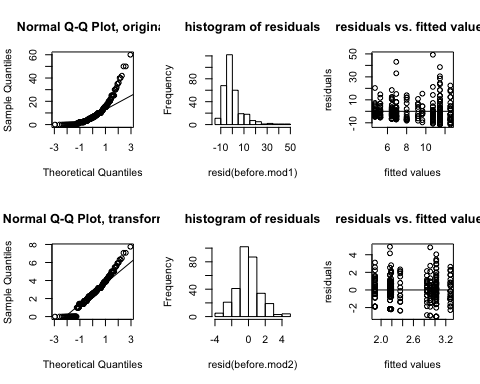
After checking the normal probability plot and the histogram of residuals for the non-transformed model, it was immediately apparent that the residuals are skewed right. To account for this, I decided to use a square root transformation on the response; I found this transformation to be most appopriate since there are zero values in the data set and the transformation had only a moderate effect of the distribution shape (square root transformation preserved the shape of the oringal response). I also considered a cubed root transformation, but ultimately decided against it because it seemed too heavy-handed and appeared to worsen the normality of variances assumption.

Even after the transformation, the normal probability plot still reflects that the error terms are not normally distributed due to the zero-inflated data. Using the Shapiro-Wilk normality of error test, we can conclude that the terms are not normally distributed since we had a p-value of approximately zero.

Aside from this large caveat, I concluded that there is a homogeneity of variances between groups by the high p-value of Levene’s test (and supported visually by the horizontal band of data on the residuals vs fitted plot). Furthermore, the transformation seemed to help reduce the appearance of outliers and there was no trend apparent in the residuals vs fitted plot, which supports the claim of independence between error terms.

I decided to continue with this model despite the non-normality of error terms, but to further improve this model, I would use a zero-inflated Poisson distribution since we are dealing with non-negative count data.

#model with interaction term and amount of waste as a response  
before.mod1 <- aov(waste.b ~ gender \* year)  
  
#check model assumptions  
  
#before response transformation...  
#normal probability plot of residuals  
par(mfrow=c(2,3))  
qqnorm(waste.b, main="Normal Q-Q Plot, original")  
qqline(waste.b)  
  
#histogram of residuals  
hist(resid(before.mod1),main="histogram of residuals")  
  
#residuals vs fits plot  
e.mod1 <- resid(before.mod1)  
yhat.mod1 <- predict(before.mod1)  
plot(yhat.mod1,e.mod1,main="residuals vs. fitted values",xlab="fitted values",ylab="residuals")  
abline(h=0)  
  
  
  
#after response transformation...  
sqrt.waste.b <- sqrt(waste.b) #square root transformation to reduce right skew  
before.mod2 <- aov(sqrt.waste.b ~ gender \* year) #with interaction and transformed y  
  
#normal probability plot of residuals  
qqnorm(sqrt.waste.b, main="Normal Q-Q Plot, transformed") #helps but zero-inflation still apparent  
qqline(sqrt.waste.b)  
  
#histogram of residuals  
hist(resid(before.mod2),main="histogram of residuals")  
  
#residuals vs fits plot  
e.mod2 <- resid(before.mod2)  
yhat.mod2 <- predict(before.mod2)  
plot(yhat.mod2,e.mod2,main="residuals vs. fitted values",xlab="fitted values",ylab="residuals")  
abline(h=0)



#plot indicates a bit of variation in variances (among other things), test using levene  
  
  
#Levene homogeneity of variance test  
leveneTest(before.mod2)

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 9 0.7949 0.6213  
## 306

#H0: group variances are equal. high p-value -> can conclude variances equal  
#this test is robust to non-normality :-)  
  
#Shapiro-Wilk normality of error test  
shapiro.test(sqrt.waste.b) #p-value of 0... error terms not normally distributed

##   
## Shapiro-Wilk normality test  
##   
## data: sqrt.waste.b  
## W = 0.96798, p-value = 1.856e-06

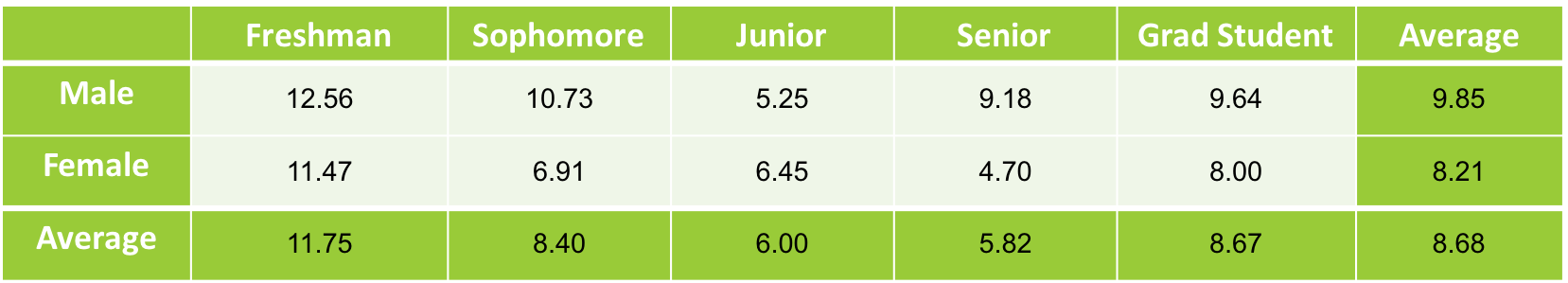
# Method of analysis

As an exploratory analysis, I looked at group means as well as box plots of the data. The response observed was the number of plastic and paper food-related consumables used on campus weekly. These were grouped by gender and year in school.

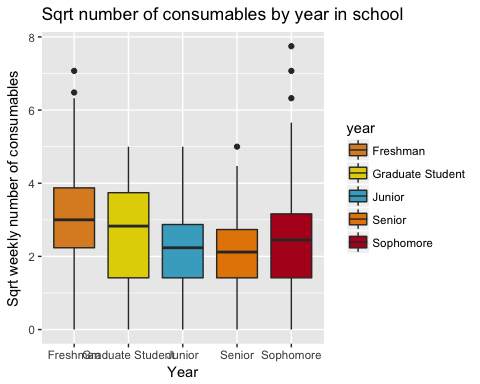
It can be seen that freshman seem to waste more on average compared to other years and that males appear to waste slightly more than females.

ybar.groups.b <- round(tapply(waste.b,list(gender,year), mean),2) #group means  
ybar.dot.j <- round(tapply(waste.b,list(year), mean),2) #sum over all a levels  
ybar.i.dot <- round(tapply(waste.b,list(gender), mean),2) #sum over all b levels  
overall.mean <- mean(waste.b)  
  
groups.size <- tapply(waste.b,list(gender, year), length) #group sample sizes  
groups.size #smallest group sample size of 11

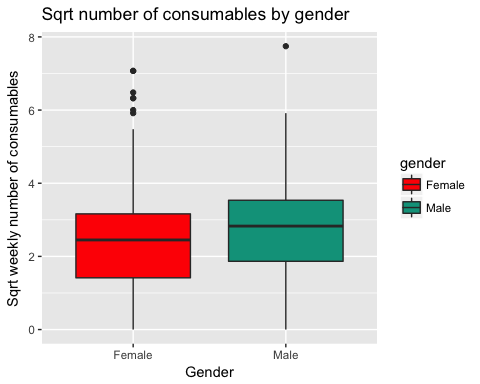
## Freshman Sophomore Junior Senior Graduate Student  
## Male 27 30 24 11 11  
## Female 77 47 40 33 16



library(wesanderson)  
ggplot(data=dat,aes(x=year,y=sqrt.waste.b,fill=year)) + geom\_boxplot() +   
 scale\_fill\_manual(values=wes\_palette(n=5, name="FantasticFox")) +   
 labs(x = "Year") + labs(y="Sqrt weekly number of consumables") +   
 labs(title="Sqrt number of consumables by year in school")



ggplot(data=dat,aes(x=gender,y=sqrt.waste.b,fill=gender)) + geom\_boxplot() +   
 scale\_fill\_manual(values=wes\_palette(n=2, name="Darjeeling")) +   
 labs(x = "Gender") + labs(y="Sqrt weekly number of consumables") +   
 labs(title="Sqrt number of consumables by gender")



Now that I have decided on using a square root transformation, I will test for whether or not the main effects and interaction term are statistically significant.

Firstly, I ran a type I ANOVA on the full model containing the interaction term. Since the interaction term was not significant, I decided to drop it and proceed with a type III ANOVA for the reduced model due to the unbalanced sample sizes.

After running the ANOVA, we can reject the null hypothesis that the group means are all equal for both main effects because of the low p-values. Thus there is a statistically significant difference in means between at least two groups of the groups.

before.mod2 <- aov(sqrt.waste.b ~ gender \* year) #model with interaction term  
before.mod3 <- aov(sqrt.waste.b ~ gender + year) #model without interaction term  
anova(before.mod3, before.mod2)

## Analysis of Variance Table  
##   
## Model 1: sqrt.waste.b ~ gender + year  
## Model 2: sqrt.waste.b ~ gender \* year  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 310 624.35   
## 2 306 611.81 4 12.547 1.5689 0.1825

library(car)  
Anova(before.mod3,type="III") #reduced model

## Anova Table (Type III tests)  
##   
## Response: sqrt.waste.b  
## Sum Sq Df F value Pr(>F)   
## (Intercept) 642.51 1 319.0142 < 2.2e-16 \*\*\*  
## gender 9.62 1 4.7743 0.02964 \*   
## year 56.52 4 7.0156 2.035e-05 \*\*\*  
## Residuals 624.35 310   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

To further explore these differences, we will use Tukey’s post-hoc comparison of means. It can be seen that there are significant differences between jr-fr and sr-fr and so-fr, but the Tukey test did not produce a significant difference between genders.

TukeyHSD(before.mod3, c("gender","year")) #pairwise mean comparisons

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = sqrt.waste.b ~ gender + year)  
##   
## $gender  
## diff lwr upr p adj  
## Female-Male -0.3070986 -0.6422309 0.02803384 0.072351  
##   
## $year  
## diff lwr upr p adj  
## Sophomore-Freshman -0.67529695 -1.2607339 -0.08986001 0.0146075  
## Junior-Freshman -1.05214913 -1.6708074 -0.43349082 0.0000447  
## Senior-Freshman -0.96020344 -1.6605141 -0.25989281 0.0018714  
## Graduate Student-Freshman -0.57988332 -1.4209685 0.26120183 0.3239008  
## Junior-Sophomore -0.37685217 -1.0355361 0.28183176 0.5179044  
## Senior-Sophomore -0.28490649 -1.0208150 0.45100207 0.8256582  
## Graduate Student-Sophomore 0.09541364 -0.7755345 0.96636177 0.9982158  
## Senior-Junior 0.09194569 -0.6706572 0.85454853 0.9974058  
## Graduate Student-Junior 0.47226581 -0.4213517 1.36588337 0.5957047  
## Graduate Student-Senior 0.38032012 -0.5716511 1.33229134 0.8084379

# Conclusions

Due to the nature of the non-negative and zero-inflated count data I collected, I think that a zero-inflated Poisson distribution would have been more appropriate. Since that model was beyond the scope of this class, I tried to make the normal regression work to the best of my ability.

Using this model, I found that both main effects (gender and year) were significant and the interaction term was not. In specific, freshmen seemed to have a higher amount of consumables used weekly compared to most other groups. With respect to the inconsistency between the ANOVA for the gender effect and the Tukey comparison between male and female, I am inclined to side with the ANOVA and conclude that there is a statistically significant difference between males and females using this model. Females seem to consistently produce lower amounts of waste between the different groups for year in school and the low p-value indicates that the factor does illustrate a statistically significant effect on the dependent variable, amount of consumable waste.

In conclusion, we can reject both null hypotheses since both main effects were significant!