4 Statistical Frameworks Study Guide (and other notes?)

EA30 2023 Students and Marc

2023-03-27

# Using R to Create R Study Guide

## Class Assignment

The study guide development will go for five weeks, so everyone can have a change to look at each of the for frameworks and then one session to integrate everything into one document. See sakai to see your group assignment.

## Developing Content

Each week, I intend to dedicate 30-45 minutes to develop content.

We’ll start with creating an example. But as we do, we want to be answering these following questions:

* What questions can be answered with the method?
* Are their limitations to the method?
* What are the assumptions?
* How can the assumptions be tested?

The example should include data to be analyzed – either imported or created in R, then the R codes to analyze the results and then a discussion about how to interpret the data.

After each week, please upload the version you have editted and I will harmonize the 4 groups work into one document for the following week.

# Getting Started with R

## R and Rstudio

### What is R?

TBD

### Getting started in R

TBD

If you havenâ€™t used R before, start by downloading R and R Studio.

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

summary(cars) # extract summary statistics from built-in cars dataset.

## speed dist   
## Min. : 4.0 Min. : 2.00   
## 1st Qu.:12.0 1st Qu.: 26.00   
## Median :15.0 Median : 36.00   
## Mean :15.4 Mean : 42.98   
## 3rd Qu.:19.0 3rd Qu.: 56.00   
## Max. :25.0 Max. :120.00

I added the comment line to help the reader remember what’s happening. These comments are useful to help everyone see the intentions of the author.

### Including Plots

You can also embed plots, for example:



Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot. But as a guide, we might want to show the code – which can be done with echo = TRUE, which is the default.

## Updated Template to Create Word.docx Study Guide

I created this document to give you a “head-start” in creating the study guide with a format that everyone can use. By using Rstudio, you will have the capacity to create an integrated approach that merges R commands and outputs and text.

I have defined the output as a word document to start with – but we can create a pdf or html output if we decide that is better. But by putting into Word, we can use Word to spot typos easier!

There are ways to share these documents via github.com – but frankly, it works great if the class is smaller, e.g. 10-12. So, we’ll use an assortment of sakai upload tools – not ideal, but I’ll make it work!

## Editing the Guide

Once you have both of these programs downloaded, open R Studio and click on File > Open File – and select this file :-).

## Run the Template

Make sure the template you downloaded runs! Nothing is more frustrating (hyperbole?) than trying to troubleshoot someone else’s code!

# Variables and Statistics

## Defining Variable Types

### Dependent vs. Independent Data (add text here…)

In some cases, we can manipulate an environmental setting or characteristic and test the impact on another characteritics. We generally call the one that we can maniupualte as the indpendent variable and the measured response as the dependent varaible.

Graphically, we often see the independent variable as the x-axis. The independant call might also be referred to as the prector variable. The response as the dependent (or response) variable.

In our titles, we often refer to the dependent first – “Response as a Function of X” or “Water Temperature vs. Air Temperature”

### Continuous vs. Count vs. Categorical

MLH: Describe…

NOTE: Spatial Data and Time Series data have special concerns… “auto-correlation”, see below…[Autocorrelation Section](#autocorrelation)

# Four Frameworks

## Linear Regression

MLH: What is linear Regression

MLH: What kinds of questions is the linear regression best suited for?

MLH: What are the “under the hood methods” used to analyze the data? Is this supported by theory? Is there any of the theory that’s useful for users?

## Linnear Regression Examples

### Example #1

We know surface water temperatures depend on the sunlight. But what about surface air temperature?

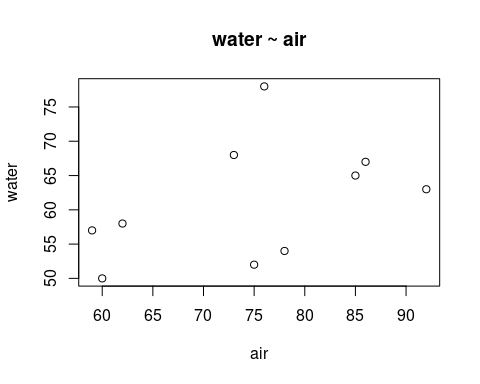
Is water and air temperature correlated? (whats the difference between correlation and regression?)

water = c(50, 52, 65, 67, 68, 78, 54, 63, 57, 58)  
air = c(60, 75, 85, 86, 73, 76, 78, 92, 59,62)  
lm(formula = water ~ air)

##   
## Call:  
## lm(formula = water ~ air)  
##   
## Coefficients:  
## (Intercept) air   
## 36.3333 0.3333

making a graph…

plot(x=air, y=water, main="water ~ air") # This was a scatterplot, but marc changed it to a "plot" on 4/1 because it wasn't clear what the model was in the curvy scatter plot approach.



#linearMod <- lm(water ~ air, data=temperature) # build linear regression model on full data  
#print(linearMod)

### Example #2: SRP and HABs

Following HABs and SRP as a predictor…

MLH: Describe why these data can be analyzed using regression

##### Define Data:

SRP <- c(.02, .03, 0.4, 0.04, .05, 0.033, 1.4)  
cyano<- c(22, 34, 23, 22, 23, 2, 122)

#### Create Linear Model using lm() function

The linear model coerces the data into a new object using a least squares method to calculate the slope and intercept (and other statistical date.)

HAB.lm = lm(cyano ~ SRP)  
HAB.lm

##   
## Call:  
## lm(formula = cyano ~ SRP)  
##   
## Coefficients:  
## (Intercept) SRP   
## 15.03 72.37

#### Model Summary (Results)

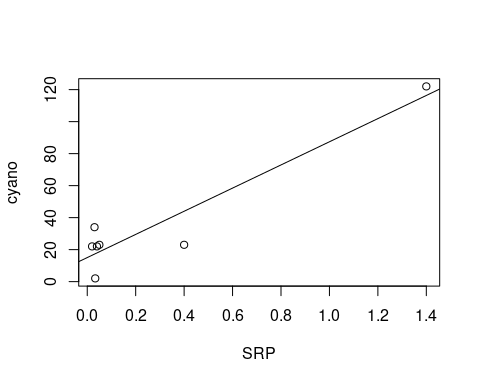
Let’s look at the model results with summary()

MLH: Describe results…

summary(HAB.lm)

##   
## Call:  
## lm(formula = cyano ~ SRP)  
##   
## Residuals:  
## 1 2 3 4 5 6 7   
## 5.523 16.799 -20.979 4.076 4.352 -15.418 5.647   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 15.029 6.398 2.349 0.06563 .   
## SRP 72.374 11.608 6.235 0.00155 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 14.55 on 5 degrees of freedom  
## Multiple R-squared: 0.886, Adjusted R-squared: 0.8632   
## F-statistic: 38.87 on 1 and 5 DF, p-value: 0.001554

plot(cyano ~ SRP)  
abline(HAB.lm)



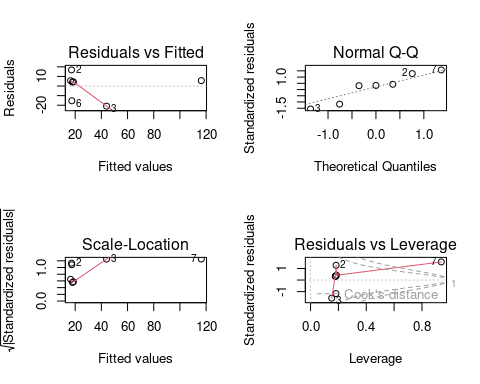
#### Model Interpretation (Discussion)

##### Linear Regression Assumptions

Let’s look at a default plot of model…

MLH: What do these mean?

par(mfrow=c(2,2))  
plot(HAB.lm)



## Logistic Regression

MLH: What is this used for?

MLH: What are good questions for this?

### Example #1 Gradaute School Admissions – MLH: Too Complicated…

This data predicts graduate school admission (discreet/binary) based on GRE score (continuous), GPA (continuous), and school prestige (discreet/binary, 1-4). This is a great resource to perform this: <https://stats.oarc.ucla.edu/r/dae/logit-regression/>

#### Name the data  
mydata <- read.csv("https://stats.idre.ucla.edu/stat/data/binary.csv")  
  
#### View the first few rows of the data  
head(mydata)

## admit gre gpa rank  
## 1 0 380 3.61 3  
## 2 1 660 3.67 3  
## 3 1 800 4.00 1  
## 4 1 640 3.19 4  
## 5 0 520 2.93 4  
## 6 1 760 3.00 2

#### Create a 5 number summary  
summary(mydata)

## admit gre gpa rank   
## Min. :0.0000 Min. :220.0 Min. :2.260 Min. :1.000   
## 1st Qu.:0.0000 1st Qu.:520.0 1st Qu.:3.130 1st Qu.:2.000   
## Median :0.0000 Median :580.0 Median :3.395 Median :2.000   
## Mean :0.3175 Mean :587.7 Mean :3.390 Mean :2.485   
## 3rd Qu.:1.0000 3rd Qu.:660.0 3rd Qu.:3.670 3rd Qu.:3.000   
## Max. :1.0000 Max. :800.0 Max. :4.000 Max. :4.000

MLH: I have simplified it to one variable… trying to stay focused on simple analyses!

MLH:What is the family=binomial for?

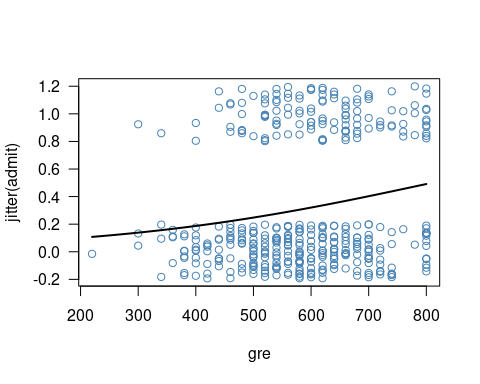
mydata$rank <- factor(mydata$rank)  
mylogit <- glm(admit ~ gre, data = mydata, family = "binomial")

summary(mylogit)

##   
## Call:  
## glm(formula = admit ~ gre, family = "binomial", data = mydata)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.1623 -0.9052 -0.7547 1.3486 1.9879   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.901344 0.606038 -4.787 1.69e-06 \*\*\*  
## gre 0.003582 0.000986 3.633 0.00028 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 499.98 on 399 degrees of freedom  
## Residual deviance: 486.06 on 398 degrees of freedom  
## AIC: 490.06  
##   
## Number of Fisher Scoring iterations: 4

#define new data frame that contains predictor variable  
newdata <- data.frame(gre=seq(min(mydata$gre), max(mydata$gre),len=100))  
  
#use fitted model to predict values of vs  
newdata$admit = predict(mylogit, newdata, type="response")

#plot logistic regression curve  
plot(jitter(admit) ~ gre, data=mydata, col="steelblue", las=1)  
lines(admit ~ gre, newdata, lwd=2)



### Example #2: Algae Concentration and Toxicity

Toxicity is a measure of toxity based on organism responses, often measured as LD50s, where 50% of the populations dies at a certain concentration.

Here we use the results of different concentrations of a toxin and the number of surviving Daphnia – a very common freshwater test organism.

MLH: Describe DATA…

## 'data.frame': 175 obs. of 2 variables:  
## $ Dose : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Result: num 1 1 1 1 1 1 1 1 1 1 ...

Create a nice table of results…

MLH: what are ones and zeros?

table(Results.df)

## Result  
## Dose 0 1  
## 0 7 18  
## 0.5 6 19  
## 1 13 12  
## 1.5 20 5  
## 2 19 6  
## 2.5 23 2  
## 3 24 1

#Results.df$Result <- as.factor(Results.df$Result)  
Results.logit <- glm(Result ~ Dose, data = Results.df, family = "binomial")

### Results

MLH: What do these results mean?

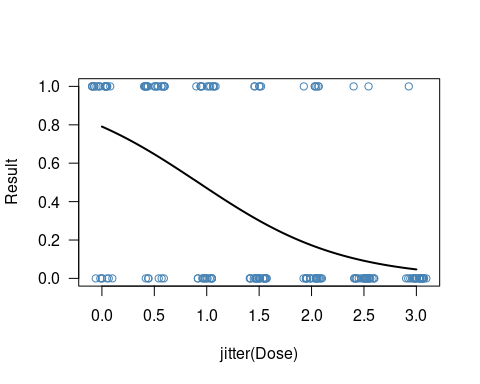
summary(Results.logit)

##   
## Call:  
## glm(formula = Result ~ Dose, family = "binomial", data = Results.df)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.7688 -0.8461 -0.4389 0.9334 2.4753   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.3296 0.3275 4.06 4.90e-05 \*\*\*  
## Dose -1.4485 0.2303 -6.29 3.18e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 228.70 on 174 degrees of freedom  
## Residual deviance: 171.04 on 173 degrees of freedom  
## AIC: 175.04  
##   
## Number of Fisher Scoring iterations: 4

#### Interpretation (Discussion)

#define new data frame that contains predictor variable  
newdata <- data.frame(Dose=seq(min(Results.df$Dose), max(Results.df$Dose),len=100))  
  
#use fitted model to predict values of vs  
newdata$Result = predict(Results.logit, newdata, type="response")

#plot logistic regression curve  
plot(Result ~ jitter(Dose), data=Results.df, col="steelblue", las=1)  
lines(Result ~ Dose, newdata, lwd=2)

 LD50 = dose at which Pr(dead | dose) = 0.5.

MLH: found this formula online… don’t know if it works…

Calculating Standard Error… MLH: Haven’t tried this yet

proportion dead 0.6 0.4 0.2 0.0 2 ˆ + SE(ˆ1) ˆ 1 LD50 2 cov(ˆ0, ˆ1) ˆ 0 ˆ1

## ANOVA

What is a one-way ANOVA?

What is the ANOVA used for?

### Example #1 –Fertilizer and Crop Yield MLH: Better for Multiway ANOVA

MLH – The example relies on external data, which adds a step that gets in the way. I suggest we developed example two to keep it simpler…

Research question: Which fertilizer produces the most crop yield? ##Independent variable/predictor: Fertilizer A, Fertilizer B, Fertilizer C ##Dependent variable/response: amount of crop yield

Question: Which fertilizer provides the best crop yield?

It is common for factors to be read as quantitative variables when importing a dataset into R. To avoid this, you can use the read.csv() command to read in the data, specifying within the command whether each of the variables should be quantitative (â€œnumericâ€) or categorical (â€œfactorâ€).

##Use the following code, replacing the path/to/your/file text with the actual path to your file:

# install.packages(c(“ggplot2”, “ggpubr”, “tidyverse”, “broom”, “AICcmodavg”))   
  
# Read Data  
# crop.data.csv = file.choose()  
  
crop.data.csv = "/home/mwl04747/beginnersluck/Stats/EA30\_Study\_Guide/ANOVA\_cropdata.csv"  
crop.data <- read.csv(crop.data.csv)  
  
names(crop.data)

## [1] "Density" "Block" "Fertilizer" "Yield"

crop.data.aov <- aov(Yield ~ Fertilizer, data=crop.data)  
  
summary(crop.data.aov)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Fertilizer 1 5.74 5.743 14.91 0.000207 \*\*\*  
## Residuals 94 36.21 0.385   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Example 2: Simulated HAB Data

Five mesocosms have been randomly selected to test the impact of zebra mussels on cyanobacteria concentrations, while five are controls. Water is circulate in each, and the experiment sampled the water column for bacteria counts every day for 10 days.

Here are the (fake) data:

Controls have a mean of 35 (and Standard Deviation of 8), Low Density treatment has a mean of 33 (SD = 8) and high density mean of 31 (SD=8)

## [1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 1 2 3 4 5 6 7 8 9 10  
## [26] 11 12 13 14 15 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 1 2 3 4 5  
## [51] 6 7 8 9 10 11 12 13 14 15 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15  
## [76] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 1 2 3 4 5 6 7 8 9 10  
## [101] 11 12 13 14 15 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 1 2 3 4 5  
## [126] 6 7 8 9 10 11 12 13 14 15 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

## [1] "Control" "Control" "Control" "Control" "Control" "Control" "Control"  
## [8] "Control" "Control" "Control" "Control" "Control" "Control" "Control"  
## [15] "Control" "Control" "Control" "Control" "Control" "Control" "Control"  
## [22] "Control" "Control" "Control" "Control" "Control" "Control" "Control"  
## [29] "Control" "Control" "Control" "Control" "Control" "Control" "Control"  
## [36] "Control" "Control" "Control" "Control" "Control" "Control" "Control"  
## [43] "Control" "Control" "Control" "Control" "Control" "Control" "Control"  
## [50] "Control" "LD" "LD" "LD" "LD" "LD" "LD"   
## [57] "LD" "LD" "LD" "LD" "LD" "LD" "LD"   
## [64] "LD" "LD" "LD" "LD" "LD" "LD" "LD"   
## [71] "LD" "LD" "LD" "LD" "LD" "LD" "LD"   
## [78] "LD" "LD" "LD" "LD" "LD" "LD" "LD"   
## [85] "LD" "LD" "LD" "LD" "LD" "LD" "LD"   
## [92] "LD" "LD" "LD" "LD" "LD" "LD" "LD"   
## [99] "LD" "LD" "HD" "HD" "HD" "HD" "HD"   
## [106] "HD" "HD" "HD" "HD" "HD" "HD" "HD"   
## [113] "HD" "HD" "HD" "HD" "HD" "HD" "HD"   
## [120] "HD" "HD" "HD" "HD" "HD" "HD" "HD"   
## [127] "HD" "HD" "HD" "HD" "HD" "HD" "HD"   
## [134] "HD" "HD" "HD" "HD" "HD" "HD" "HD"   
## [141] "HD" "HD" "HD" "HD" "HD" "HD" "HD"   
## [148] "HD" "HD" "HD"

## [1] 1 1 1 1 1 1 1 1 1 1

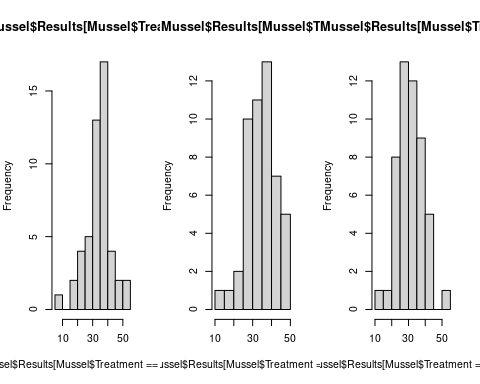
## [1] 30.7 36.2 34.1 32.4 38.9 48.2 30.4 36.6 31.4 27.1 31.8 35.2 25.6 39.2 20.6  
## [16] 36.7 23.9 32.5 36.2 38.6 39.4 17.3 26.8 39.4 42.5 42.8 54.6 40.2 37.8 25.0  
## [31] 24.3 33.1 35.4 46.0 29.9 38.6 16.6 31.1 34.4 38.9 28.8 30.9 44.4 35.3 9.7  
## [46] 36.3 39.9 30.9 50.1 34.7 38.6 25.7 31.9 45.7 41.1 34.0 43.5 49.3 36.0 34.2  
## [61] 37.1 26.1 25.2 36.4 33.9 27.4 43.2 32.5 37.2 32.2 47.7 33.7 27.2 23.8 14.7  
## [76] 27.3 36.3 48.6 43.5 41.3 26.8 26.3 41.1 34.9 39.1 30.2 28.1 39.2 34.6 21.6  
## [91] 19.4 34.5 37.6 36.1 36.4 36.7 45.1 35.8 26.5 42.7 43.4 33.4 35.4 39.2 27.9  
## [106] 29.5 38.8 26.0 34.6 21.3 24.1 22.2 26.6 34.7 32.4 22.6 40.1 29.8 26.6 30.7  
## [121] 38.4 23.1 35.7 32.2 19.3 29.4 35.0 28.9 29.5 36.5 21.4 34.1 26.6 29.6 41.2  
## [136] 28.2 31.7 32.9 43.9 23.2 42.5 35.7 39.8 10.7 26.7 51.5 20.9 33.0 32.7 39.1

### First six observations and histograms

head(Mussel)

## Mesocosm Treatment Day Results  
## 1 1 Control 1 30.7  
## 2 2 Control 1 36.2  
## 3 3 Control 1 34.1  
## 4 4 Control 1 32.4  
## 5 5 Control 1 38.9  
## 6 6 Control 1 48.2

par(mfrow=c(1,3))  
hist(Mussel$Results[Mussel$Treatment=="Control"])  
hist(Mussel$Results[Mussel$Treatment=="LD"])  
hist(Mussel$Results[Mussel$Treatment=="HD"])



### Create AOV Object (Analysis of Variance) Using Least Squares Method

Mussel.aov <- aov(results ~ treatment, data=Mussel)

### Results

MLH:Play with means or SD to see how it changes results

summary(Mussel.aov)

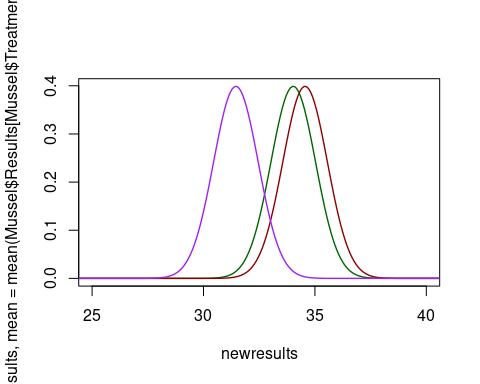
## Df Sum Sq Mean Sq F value Pr(>F)  
## treatment 2 276 137.96 2.19 0.116  
## Residuals 147 9260 62.99

### Some graphical Displays

Using a normal distribution, what do the data look like?

MLH: What do these mean?

newresults = seq(min(Mussel$Results), max(Mussel$Results), length=500)  
plot(newresults, dnorm(newresults, mean=mean(Mussel$Results[Mussel$Treatment=="Control"])), ty="l", xlim=c(25,40), col="darkgreen", lwd=1.4)  
lines(newresults, dnorm(newresults, mean=mean(Mussel$Results[Mussel$Treatment=="LD"])), ty="l", col="darkred", lwd=1.4)  
lines(newresults, dnorm(newresults, mean=mean(Mussel$Results[Mussel$Treatment=="HD"])), ty="l", col="purple", lwd=1.4)



## Tests for Association

MLH: What is this test used for?

MLH: Are their assumption to be concerned about?

# Example 2

Hospital asthema cases and HABs in Florida

HAB.tab = c(20, 15, 98, 223)  
sum(HAB.tab)

## [1] 356

dim(HAB.tab) = c(2,2)  
dimnames(HAB.tab) = list(c("Non-HAB", "HAB"), c("No Hospitalizations", "Hospitalizations"))  
  
HAB.tab

## No Hospitalizations Hospitalizations  
## Non-HAB 20 98  
## HAB 15 223

#HAB.mat = matrix(HAB.tab, nrow=2)

table(HAB.tab)

## HAB.tab  
## 15 20 98 223   
## 1 1 1 1

chisq.test(HAB.tab)

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: HAB.tab  
## X-squared = 8.9217, df = 1, p-value = 0.002818

# Other Types the Might be Important (Marc might discuss if we have time…)

## Count Data

## Multiple Regression

## Two-Way and Mult-Way ANOVAs

## Time Series Data

### Autocorrelation

## Count Data

## Mixed Effects Models