Exploring Week 2 in R

Tuesday

Introduction to Today

The goal for today is to put into practice the lecture topics we have gone over thus far, namely, exploring and evaluationg assumptions in R.

R Tip of the Day: Google is your best friend. *"How to do X in R"* will usually return extremely helpful pages.

Loading Our Libraries & Data

First, we have to load our "libraries". A **library** in R, as a reminder, is an open-source package created by a very kind individual that contains functions (short cuts) to get things done in R.

```
library(car)
library(ggplot2)
library(pastecs)
library(psych)
library(gridExtra)
library(kableExtra)

dlf <- read.delim("DownloadFestival.dat", header=TRUE)
dlf[dlf$day1 >20,] <- NA #getting rid of outliers
```

Visually Exploring Normality: Q-Q Plot

A **Q-Q Plot** will check if our data came from a theoretically normal distribution. Data being normally distributed is an assumption of many many many statistical tests and it is important to always inspect your data.

It's just a visual check, not an air-tight proof, so it is somewhat subjective. But it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation.

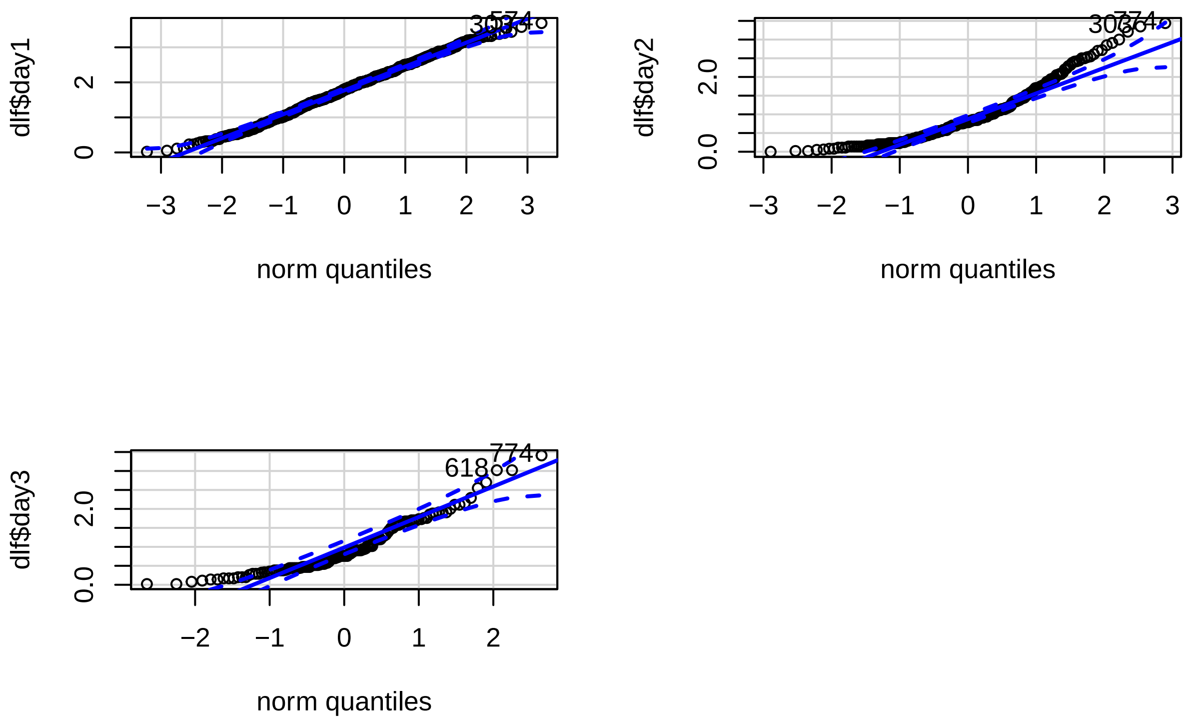
Q-Q Plot Example 1: Festival Data

If datapoints fall outside the dashed line (our cushion room aka Confidence Intervals) = not from a normal distribution.

```
par(mfrow=c(2,2)) # for non-ggplot objects, use par to adjust plot window
#Q-Q plot for day 1:
qqplot.day1 <-qqPlot(dlf$day1)

#Q-Q plot for day 2:
qqplot.day2 <- qqPlot(dlf$day2)

#Q-Q plot of the hygiene scores on day 3:
qqplot.day3 <- qqPlot(dlf$day3)
```



Q-Q Plot Example 2: Tooth Growth Data

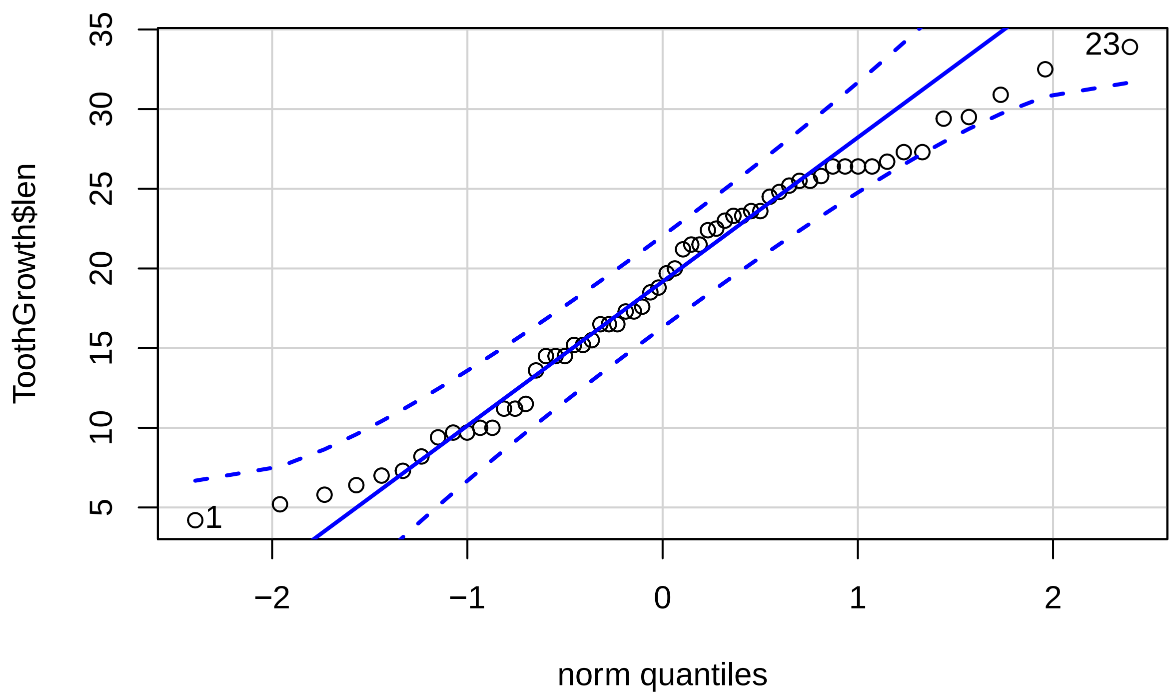
Peek at the first 6 rows of data with *head()*

```
head(ToothGrowth)
```

```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

Q-Q Plot of Tooth Length

```
qqPlot(ToothGrowth$len)
```




```
## [1] 23  1
```

Q-Q Plot Example 3: Tree Data

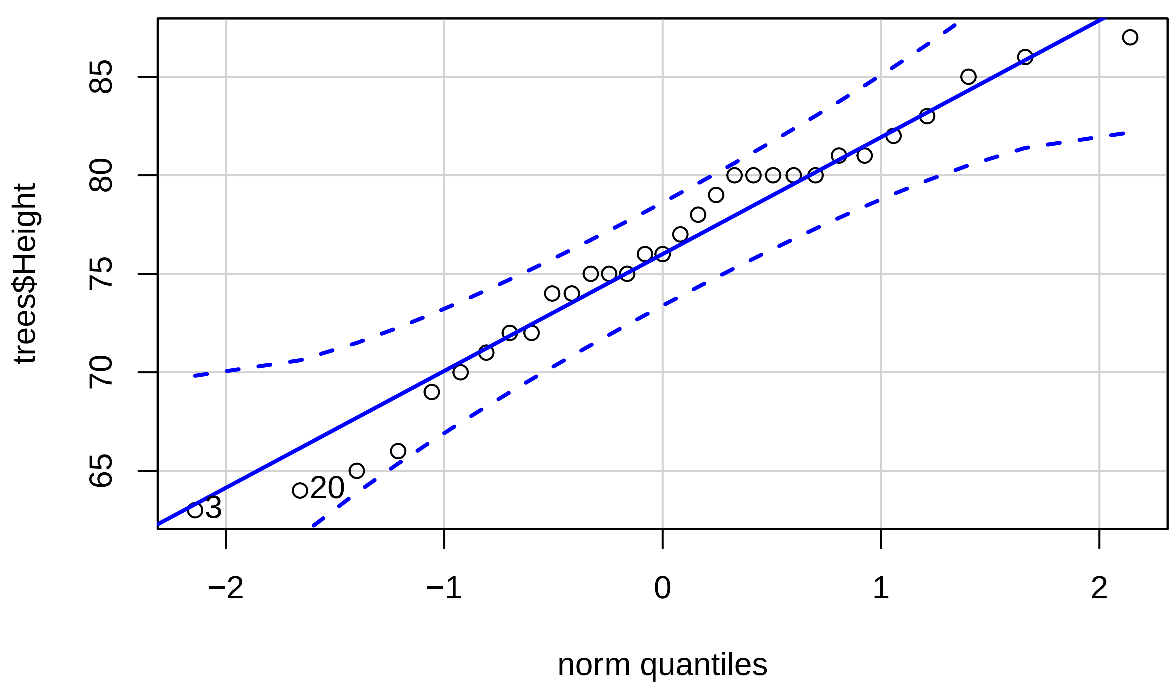
Peek at the first 6 rows of data with *head()*

```r
head(trees)
```

```
##   Girth Height Volume
## 1   8.3     70   10.3
## 2   8.6     65   10.3
## 3   8.8     63   10.2
## 4  10.5     72   16.4
## 5  10.7     81   18.8
## 6  10.8     83   19.7
```

Q-Q Plot of Tree Height

```r
qqPlot(trees$Height)
```




```
## [1]  3 20
```

Q-Q Plot Example 4: Beaver Data

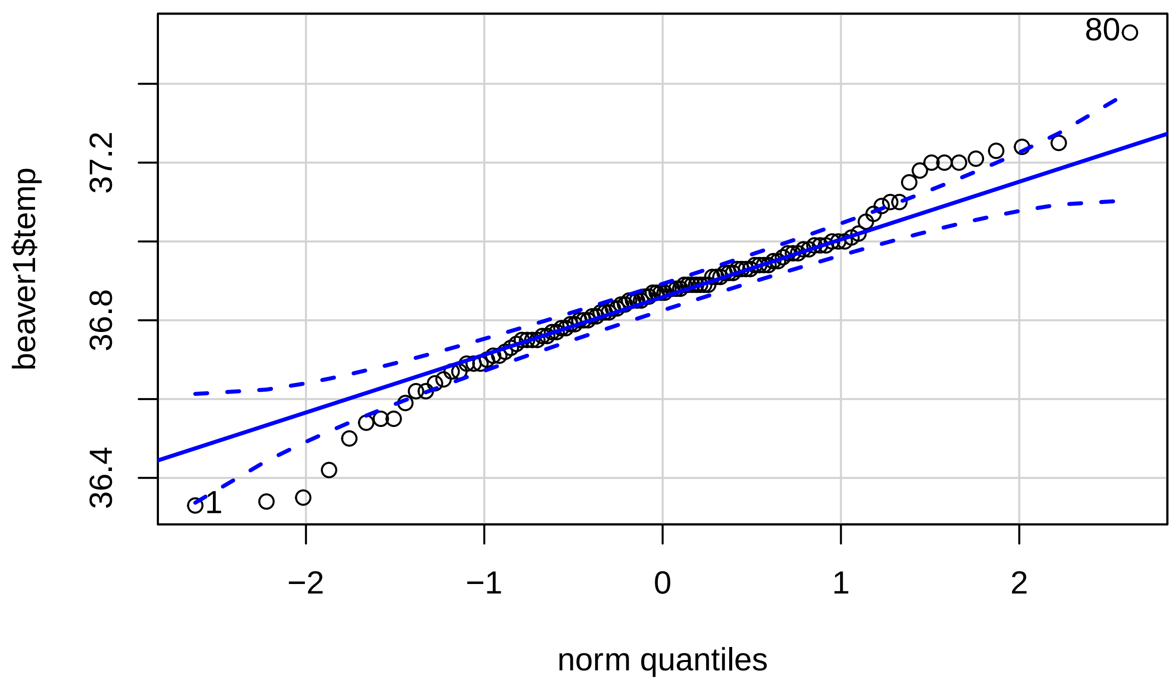
Peek at the first 6 rows of data with *head()*

```
head(beaver1)
```

```
##   day time  temp activ
## 1 346  840 36.33     0
## 2 346  850 36.34     0
## 3 346  900 36.35     0
## 4 346  910 36.42     0
## 5 346  920 36.55     0
## 6 346  930 36.69     0
```

Q-Q Plot of Tree Height

```
qqPlot(beaver1$temp)
```




```
## [1] 80  1
```

Q-Q Plot Example 5: Plant Growth Data

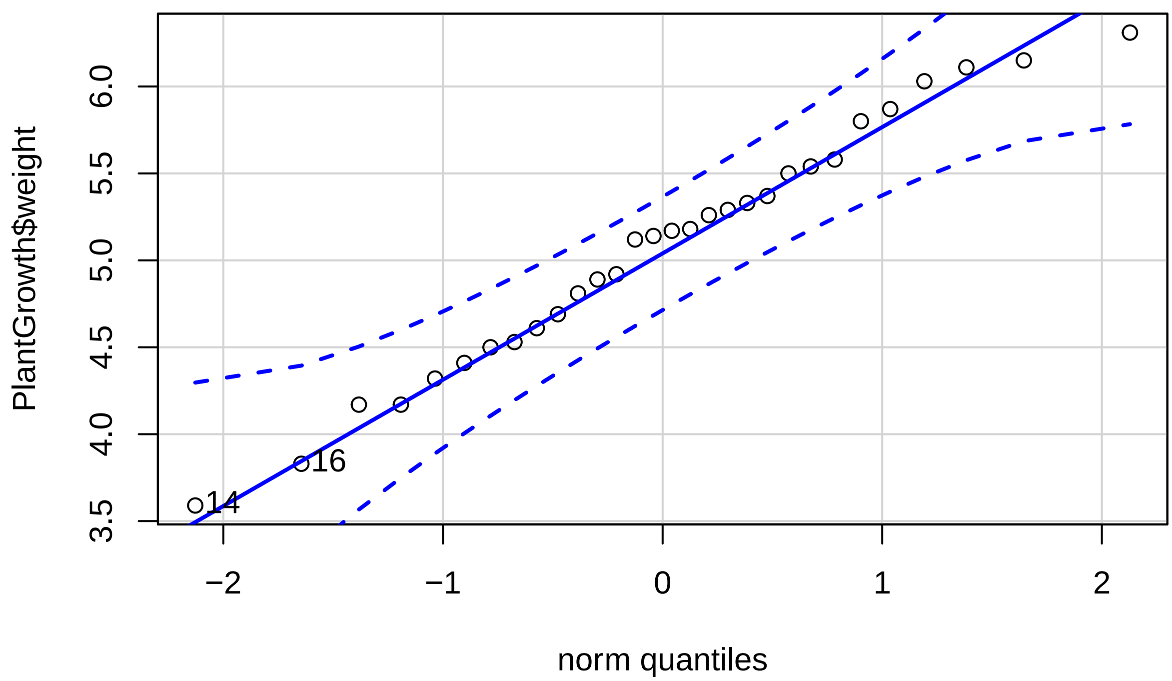
Peek at the first 6 rows of data with *head()*

```
head(PlantGrowth)
```

```
##   weight group
## 1   4.17  ctrl
## 2   5.58  ctrl
## 3   5.18  ctrl
## 4   6.11  ctrl
## 5   4.50  ctrl
## 6   4.61  ctrl
```

Q-Q Plot of Plant weight

```
qqPlot(PlantGrowth$weight)
```




```
## [1] 14 16
```

Q-Q Plot Example 6: Motor Trend Car Road Tests

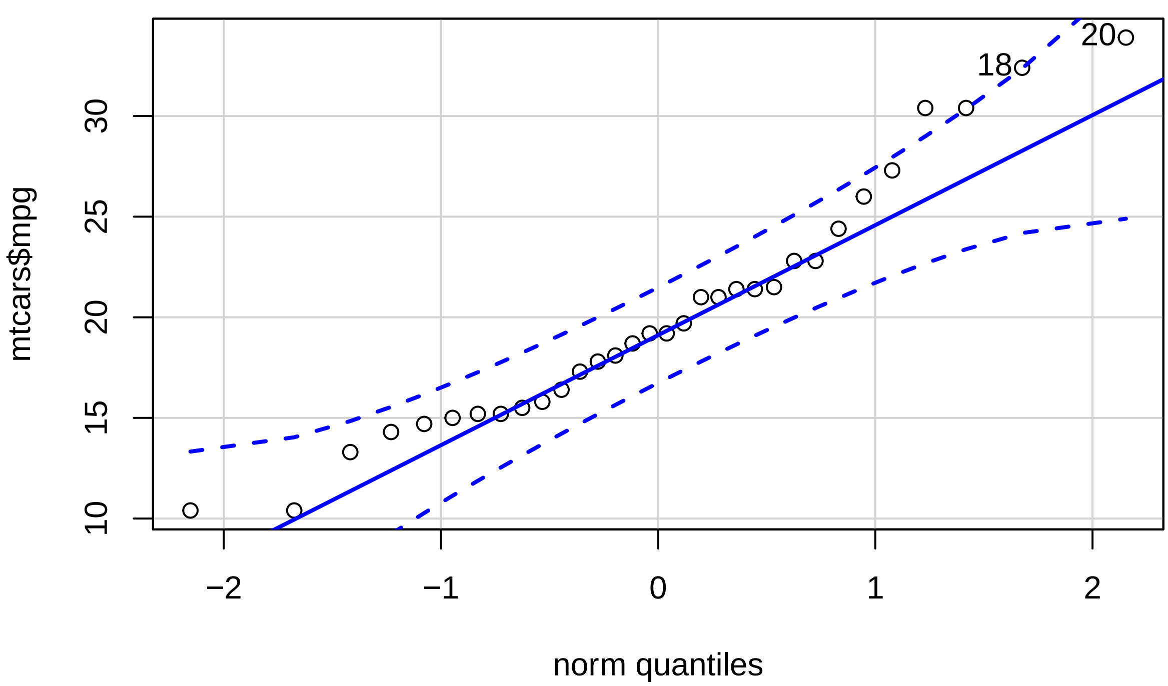
Peek at the first 6 rows of data with *head()*

```
head(mtcars)
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Q-Q Plot of mpg

```
qqPlot(mtcars$mpg)
```




```
## [1] 20 18
```

Statistically Exploring Normality: Shapiro-Wilks tests

If p-value > 0.05, it implies that the distribution of the data are not significantly different from normal distribution. In other words, we can assume the normality (good). In other words, **we want the p value (significance level) to be greater than 0.05 in other to have a statistically normal distrobution.**

Shapiro-Wilks Example 1: Festival Data

We can use the *shapiro.test()* function in R. All we have to do is tell the function the dataset and variable we want to look at.

```r
#Shapiro-Wilks day 1:
shapiro.test(dlf$day1)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dlf$day1
## W = 0.99591, p-value = 0.03184
```

```r
#Shapiro-Wilks day 2:
shapiro.test(dlf$day2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dlf$day2
## W = 0.908, p-value = 1.291e-11
```

```r
#Shapiro-Wilks day 3:
shapiro.test(dlf$day3)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dlf$day3
## W = 0.90775, p-value = 3.804e-07
```

While Day 1 is the most normal, it does not pass the statistical test for normality.

Shapiro-Wilks Example 2: Tooth Growth Data

```
shapiro.test(ToothGrowth$len)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ToothGrowth$len
## W = 0.96743, p-value = 0.1091
```

Normal!

Shapiro-Wilks 3: Tree Data

```
shapiro.test(trees$Height)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  trees$Height
## W = 0.96545, p-value = 0.4034
```

Normal!

Shapiro-Wilks Example 4: Beaver Data

```
shapiro.test(beaver1$temp)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  beaver1$temp
## W = 0.97031, p-value = 0.01226
```

Not Normal.

Shapiro-Wilks Example 5: Plant Growth Data

```
shapiro.test(PlantGrowth$weight)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  PlantGrowth$weight
## W = 0.98268, p-value = 0.8915
```

Normal!

Shapiro-Wilks Example 6: Motor Trend Car Road Tests

```
shapiro.test(mtcars$mpg)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mtcars$mpg
## W = 0.94756, p-value = 0.1229
```

Normal!

ExtRa PRactice

Now, with our remaining time, I'd like to walk you through best-practices/the steps I take when I have a dataset to analyze. For now, I'll just show you the steps I take prior to data analysis, i.e. loading and describing data.

Load Libraries

```
library(psych)
library(car)
library(apaTables)
```

Load in data

```
goggles <- read.csv("goggles.csv", stringsAsFactors = TRUE)
```

Make sure the data looks okay

```
head(goggles)
```

```
##   gender alcohol attractiveness
## 1 Female 4 Pints             55
## 2 Female 4 Pints             65
## 3 Female 4 Pints             70
## 4 Female 4 Pints             55
## 5 Female 4 Pints             55
## 6 Female 4 Pints             60
```

My Research Question

Drinking alcohol impacts accuracy on how attractive someone is

My Hypothesis

I hypothesize that drinking alcohol makes people less accurate at accuracy ratings of attractiveness. Specifically, I believe that this effect will be more pronounced for men.

Research Design

This research design is called a 2 x 2 between-subjects design. This means that we have two different independent variables with two levels each. We have two independent variables (gender and alcohol) and that is why we have two different numbers. They both are 2's because each variable has two levels. In long form, we can say this is a 2 (Gender: Female vs Male) x 2 (Alcohol: 4 Pints vs 0 Pints) between-subjects design.

It is between-subjects because all participants are independent (i.e. they *either* drink 4 or 0 pints, not both). Let's explore our variables and their structures.
Note: I picked this because this is the exact same design your final project will have

Independent Variable 1: Gender

Look at the structure with *str()*

```r
str(goggles$gender)
```

```
##  Factor w/ 2 levels "Female","Male": 1 1 1 1 1 1 1 1 2 2 ...
```

```r
kable(table(goggles$gender))
```

Var1	Freq
Female	16
Male	16

Independent Variable 2: Alcohol

```r
str(goggles$alcohol)
```

```
##  Factor w/ 2 levels "4 Pints","None": 1 1 1 1 1 1 1 1 1 1 ...
```

```r
kable(table(goggles$alcohol))
```

Var1	Freq
4 Pints	16
None	16

Combined Frequency Table

```
kable(table(goggles$gender, goggles$alcohol))
```

	4 Pints	None
Female	8	8
Male	8	8

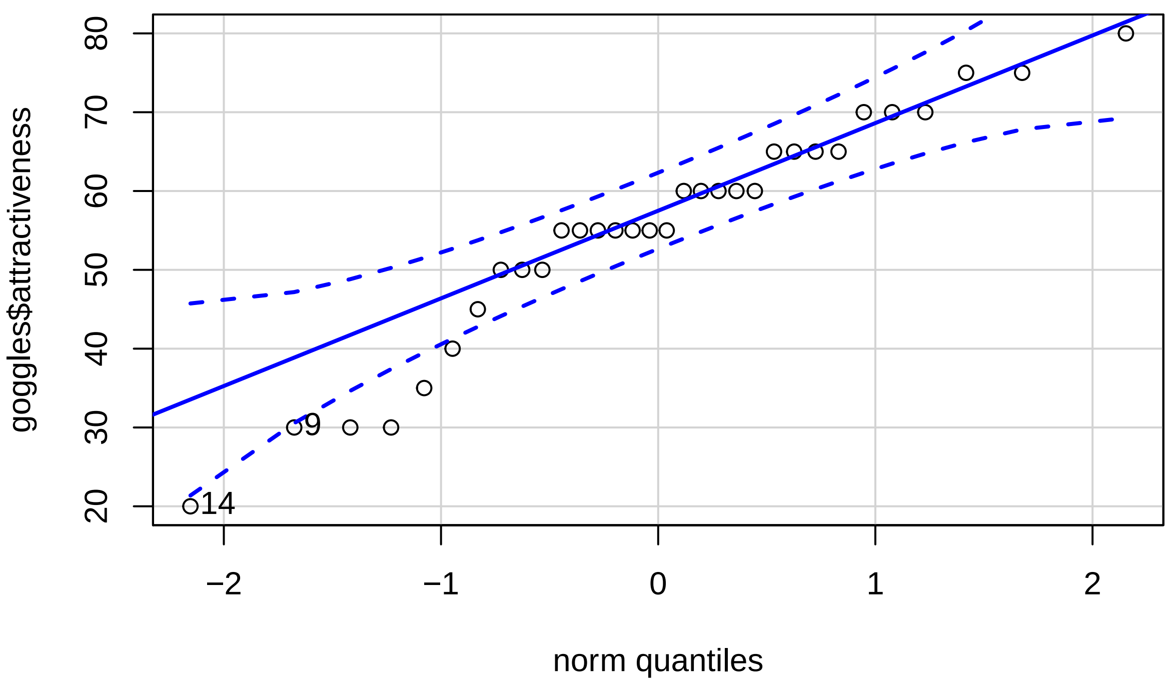
Dependent Variable: Attractiveness Accuracy

```
str(goggles$attractiveness)
```

```
##  int [1:32] 55 65 70 55 55 60 50 50 30 30 ...
```

Normality

```
qqPlot(goggles$attractiveness)
```




```
## [1] 14  9
```

```r
shapiro.test(goggles$attractiveness)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  goggles$attractiveness
## W = 0.94354, p-value = 0.09439
```

It's normally distributed!

Descriptives (Mean & SD) of our Research Design

apa.2way.table() will give us our descriptives broken down by each condition.

```r
apa.2way.table(gender, alcohol, attractiveness, data = goggles, table.number = 1,
  show.conf.interval = FALSE, show.marginal.means = FALSE,
  landscape = TRUE, filename = NA)
```

```
##
##
## Table 1
##
## Means and standard deviations for attractiveness as a function of a 2(gender) X 2(alcohol) design
##
##         alcohol
##         4 Pints        None
##  gender       M    SD     M    SD
##  Female   57.50  7.07 60.62  4.96
##    Male   35.62 10.84 66.88 10.33
##
## Note. M and SD represent mean and standard deviation, respectively.
```