Data 606- Lab 4

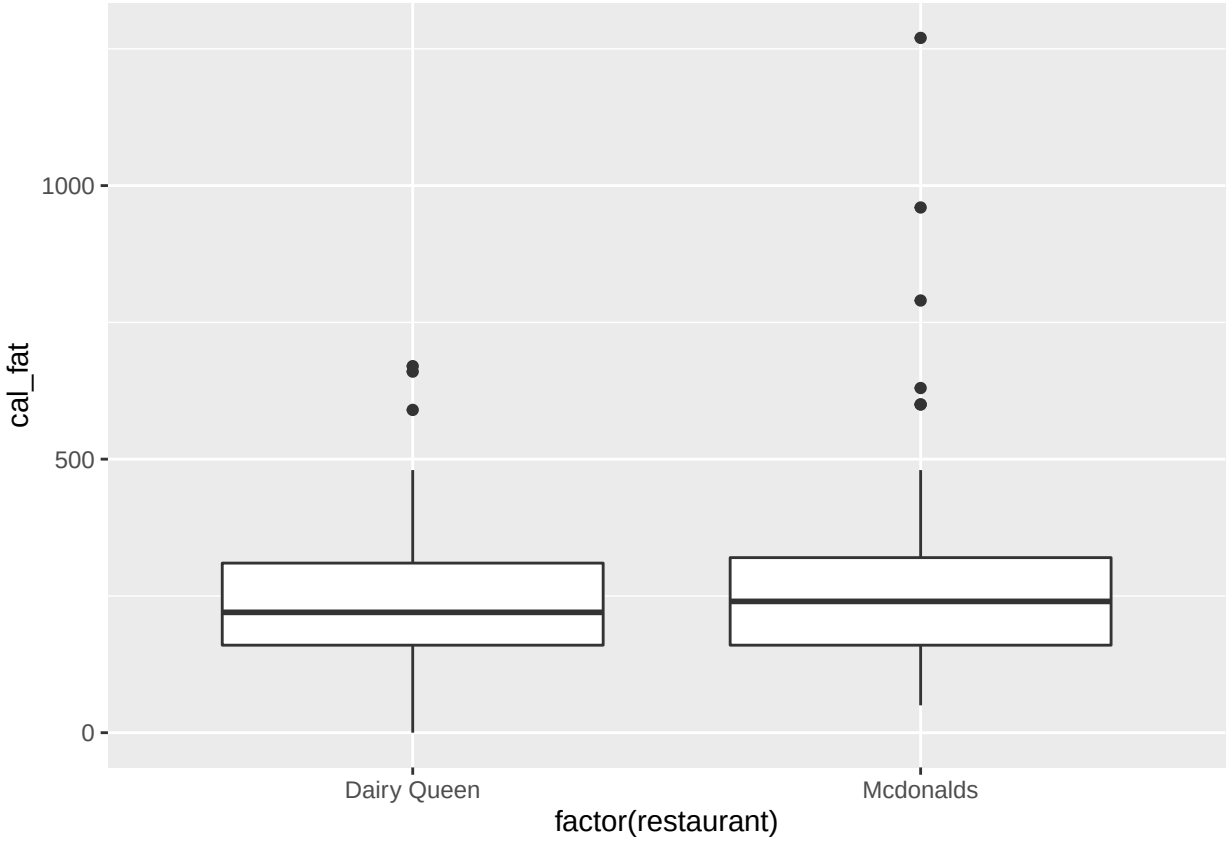
Jack Wright

9/20/2020

```
mcdonalds <- fastfood %>%
  filter(restaurant == "Mcdonalds")
dairy_queen <- fastfood %>%
  filter(restaurant == "Dairy Queen")
calfat<-rbind(mcdonalds,dairy_queen)
```

1. Make a plot (or plots) to visualize the distributions of the amount of calories from fat of the options from these two restaurants. How do their centers, shapes, and spreads compare?

```
ggplot(calfat, aes(factor(restaurant),cal_fat))+geom_boxplot()
```



The centers seem to be about the same, the spread of the interquartile range are about the same, but McDonalds is more right skewed, and has far more extremely high calorie choices.

2. Based on the this plot, does it appear that the data follow a nearly normal distribution?

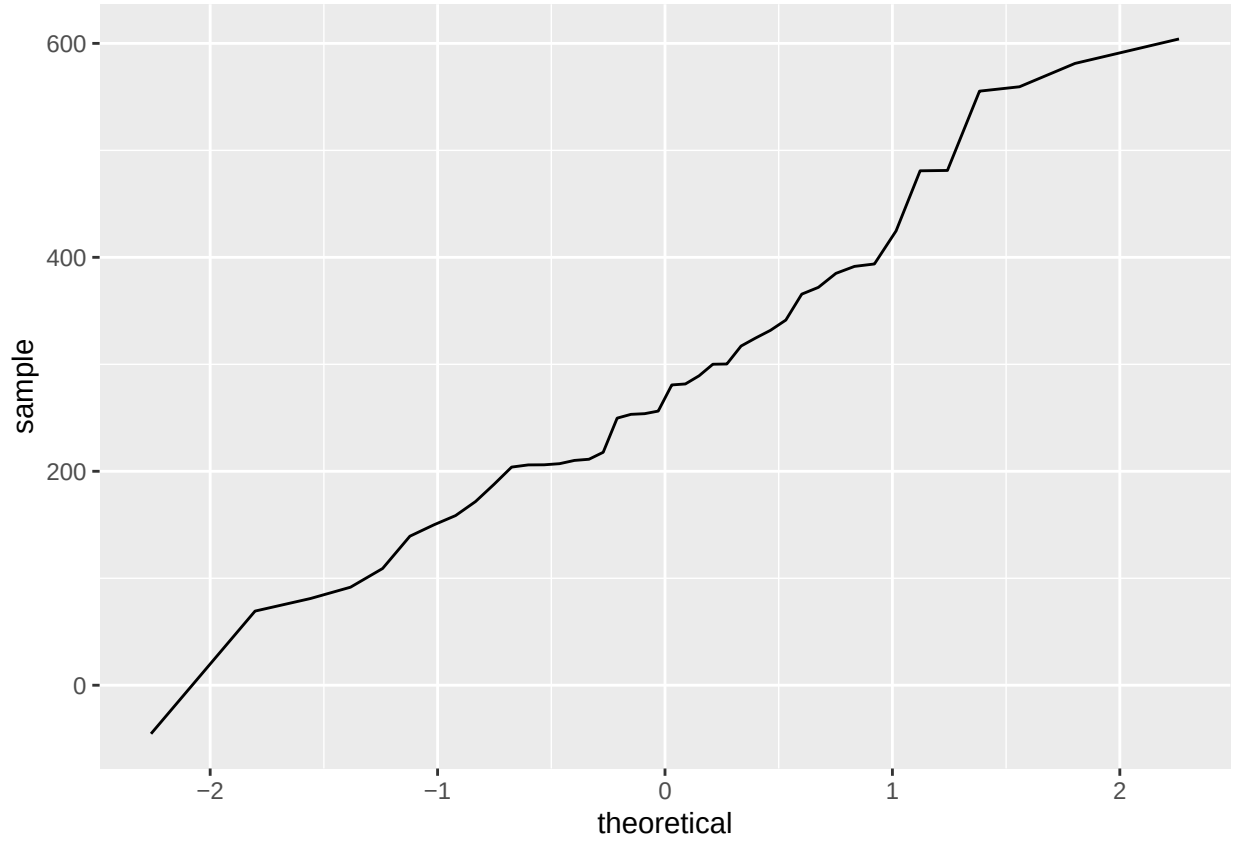
No it does not look like a normal distribution.

3. Make a normal probability plot of `sim_norm`. Do all of the points fall on the line? How does this plot compare to the probability plot for the real data? (Since `sim_norm` is not a dataframe, it can be put directly into the `sample` argument and the `data` argument can be dropped.)

```
dqmean <- mean(dairy_queen$cal_fat)
dqsd   <- sd(dairy_queen$cal_fat)


sim_norm <- rnorm(n = nrow(dairy_queen), mean = dqmean, sd = dqsd)

ggplot(, aes(sample = sim_norm)) +
  geom_line(stat = "qq")
```



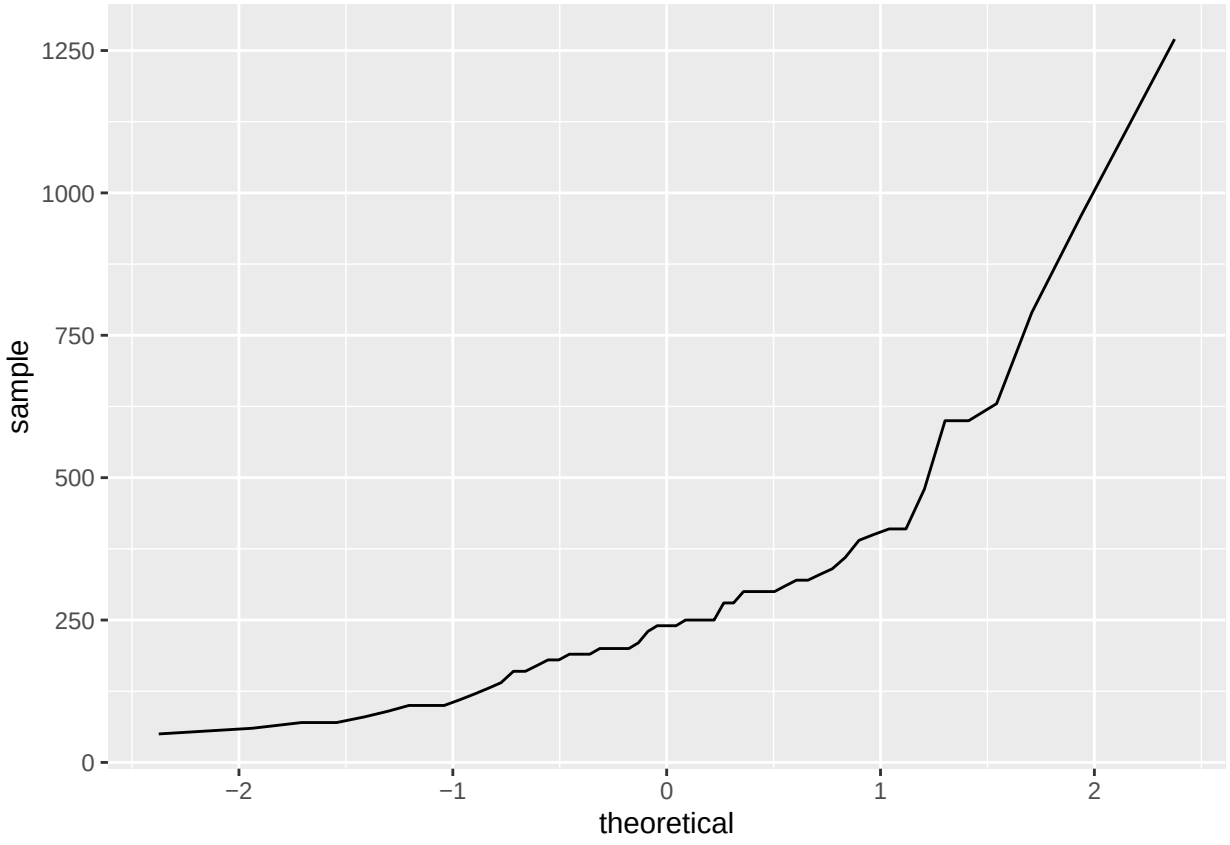
4. Does the normal probability plot for the calories from fat look similar to the plots created for the simulated data? That is, do the plots provide evidence that the female heights are nearly normal?

Yes they do look similar. I believe this does provide evidence that the fat calories are nearly normal.

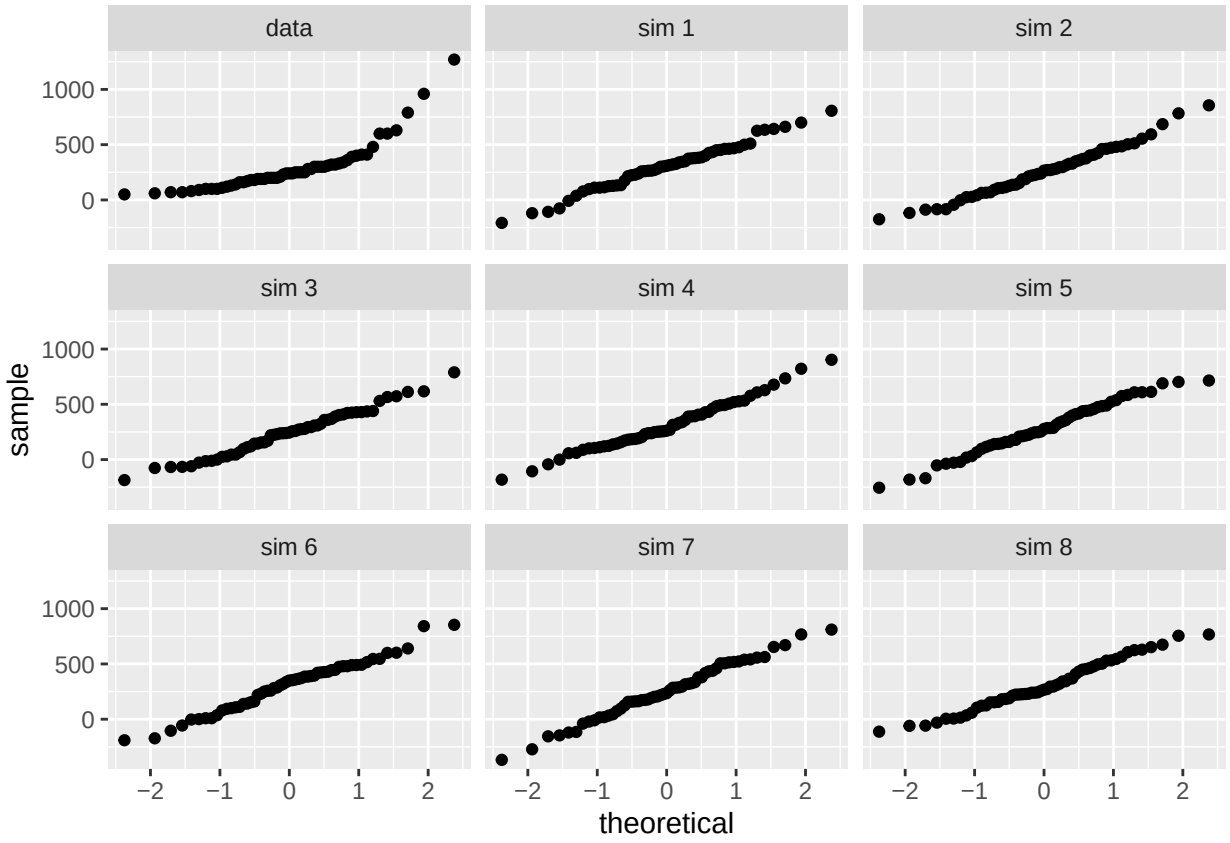
5. Using the same technique, determine whether or not the calories from McDonald's menu appear to come from a normal distribution.

No i dont believe that the McDonalds fat calories are normally distributed

```
ggplot(data = mcdonalds, aes(sample = cal_fat)) +
  geom_line(stat = "qq")
```




```
qqnormsim(sample = cal_fat, data = mcdonalds)
```



6. Write out two probability questions that you would like to answer about any of the restaurants in this dataset. Calculate those probabilities using both the theoretical normal distribution as well as the empirical distribution (four probabilities in all). Which one had a closer agreement between the two methods?

Question 1

What is the probability that a randomly chosen Dairy Queen item is below 300 calories?

```
dqmean <- mean(dairy_queen$cal_fat)
dqsd   <- sd(dairy_queen$cal_fat)
 a1<- pnorm(300,mean=dqmean,sd=dqsd)
 a1
```

```
## [1] 0.5997007
```

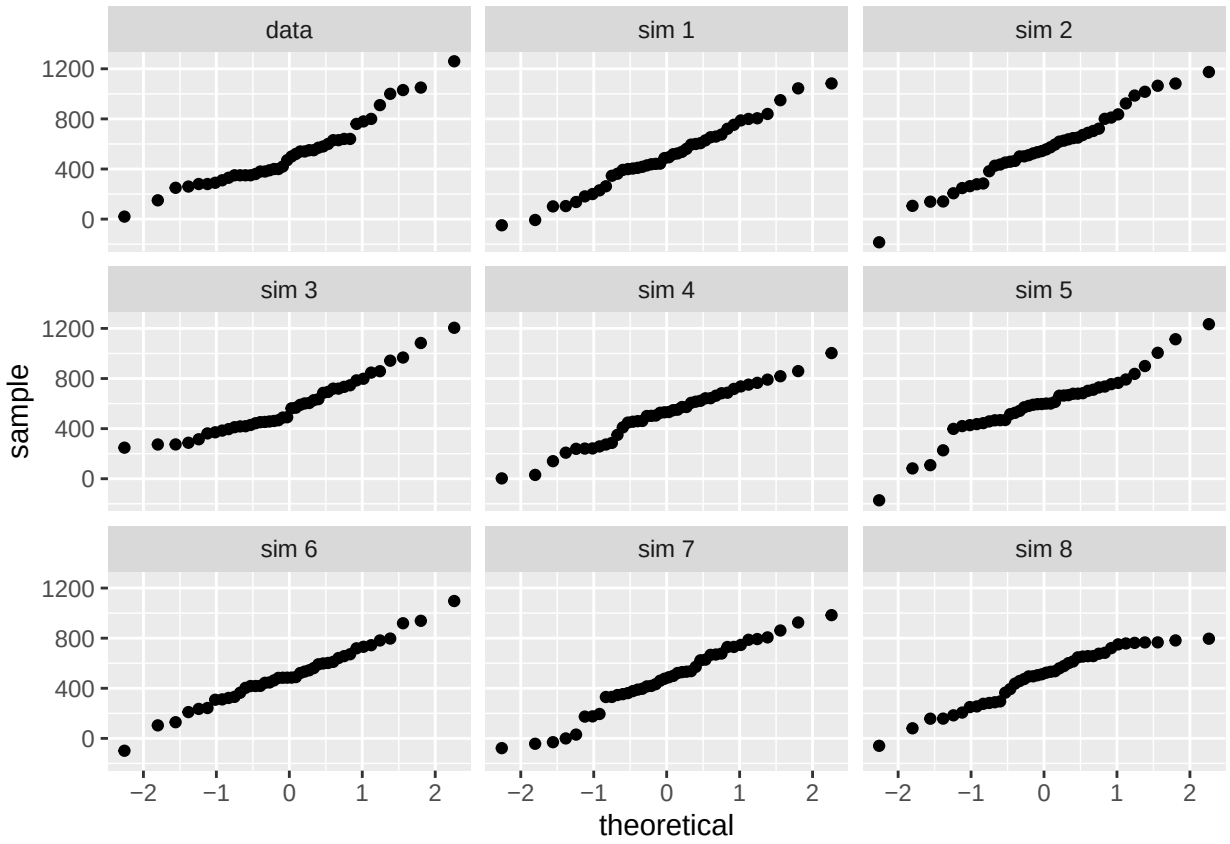
```
a2<-dairy_queen %>%
  filter(cal_fat < 300) %>%
  summarise(percent = n() / nrow(dairy_queen))
a2
```

```
## # A tibble: 1 x 1
##   percent
##     <dbl>
## 1   0.667
```

Setup to ask Question 2

Are the calories in Dairy queen food normally distributed?

```
qqnormsim(sample = calories, data = dairy_queen)
```



Yes... So

Question 2

What is the probability that a randomly chosen Dairy Queen item is within 400 to 500 calories?

```
dqmean<-mean(dairy_queen$calories)
dqsd<-sd(dairy_queen$calories)
b1<-pnorm(500,mean=dqmean,sd=dqsd)-pnorm(400,mean=dqmean,sd=dqsd)
b1
```

```
## [1] 0.1474445
```

```
b2<-dairy_queen %>%
  filter(calories >= 400 & calories<=500) %>%
  summarise(percent = n() / nrow(dairy_queen))
b2
```

```
## # A tibble: 1 x 1
##   percent
##     <dbl>
## 1   0.119
```

comparison

```
output<-c("first question"=(abs(a1-a2)/a2),"second question"=(abs(b1-b2)/b2))
output
```

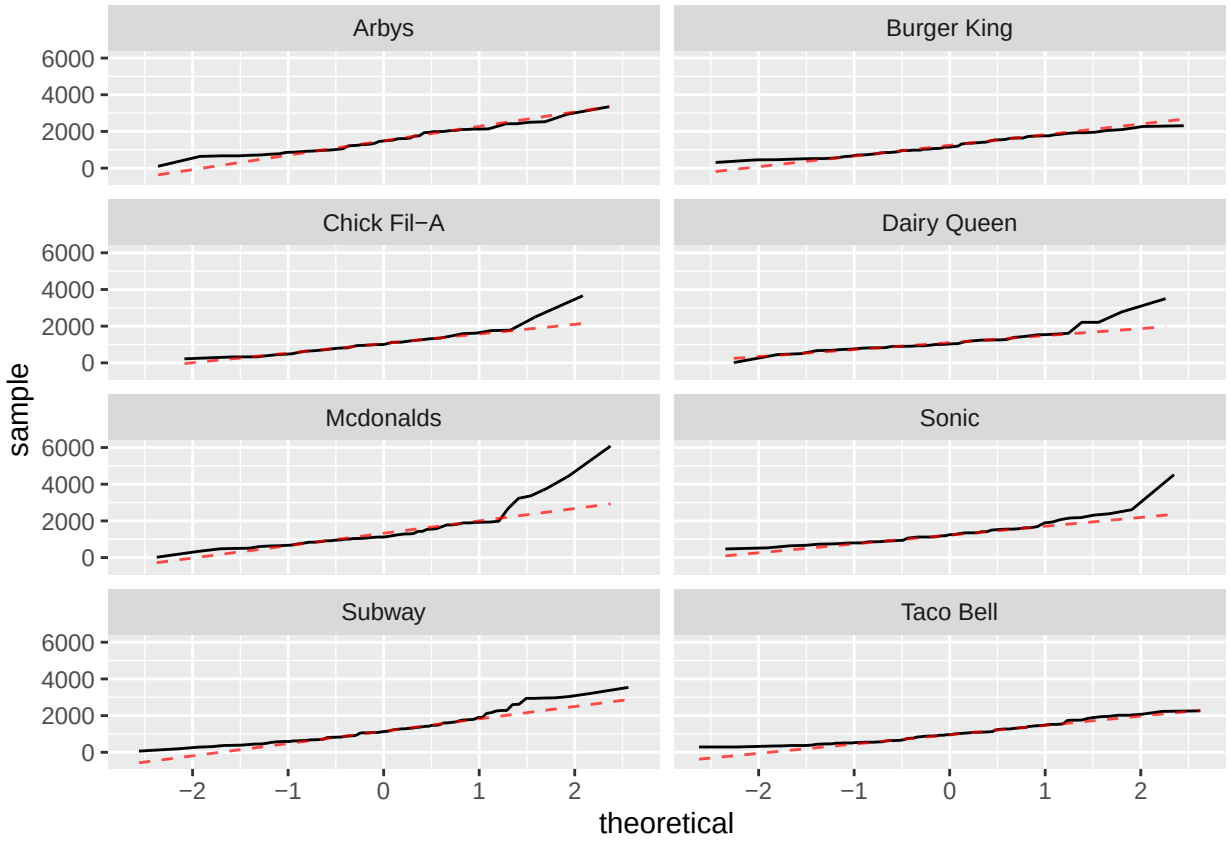
```
## $`first question.percent`
## [1] 0.100449
##
## $`second question.percent`
## [1] 0.2385339
```

The first question posed is closer together

More Practice

7. Now let's consider some of the other variables in the dataset. Out of all the different restaurants, which ones' distribution is the closest to normal for sodium?

```
ggplot(data = fastfood, aes(sample = sodium)) +
  geom_line(stat = "qq")+
    stat_qq_line(alpha = 0.7, color = "red", linetype = "dashed") +
  facet_wrap(~restaurant, nrow=4)
```



Either Burger King or Taco Bell seem to have the sodium most normally distributed

8. Note that some of the normal probability plots for sodium distributions seem to have a stepwise pattern. why do you think this might be the case?

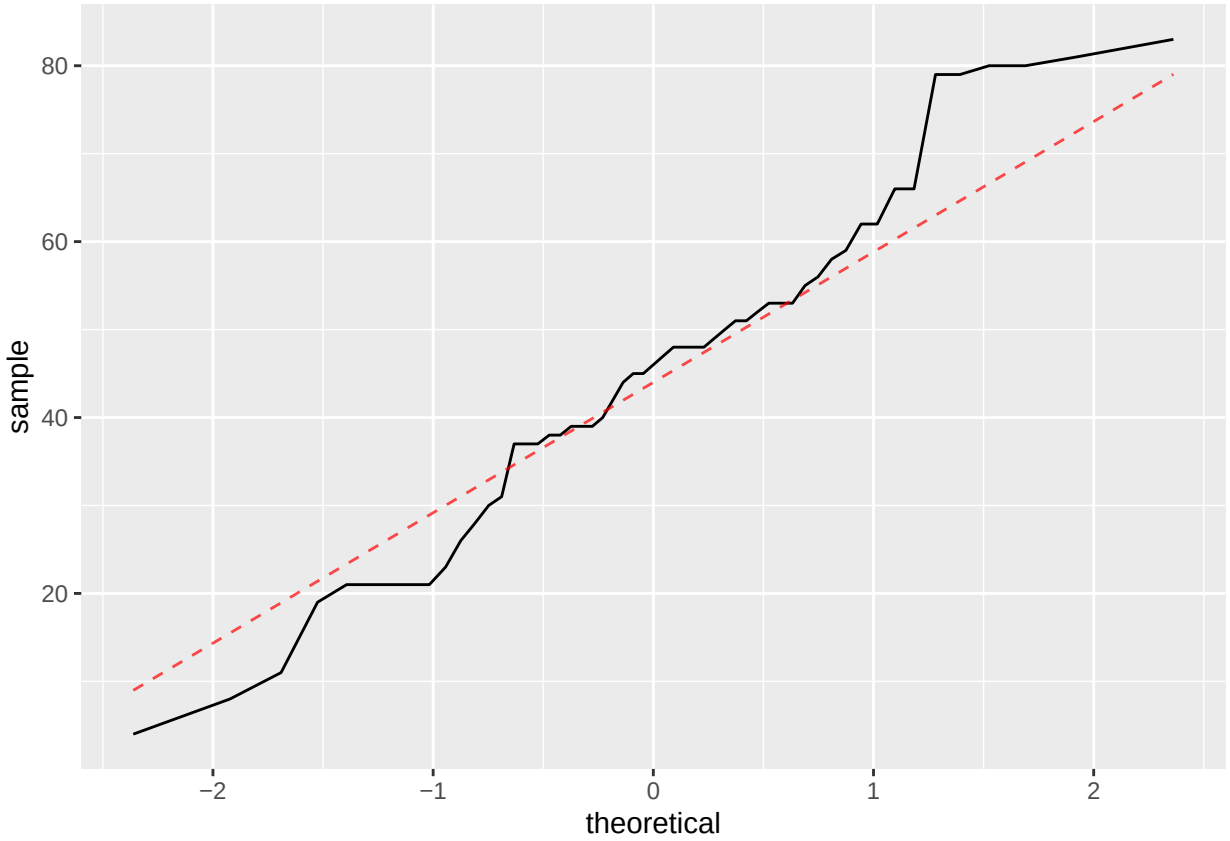
I am not sure what this is referring to. Maybe if you did each plot as a histogram you would get binning, and then you would have steps between the groups?

9. As you can see, normal probability plots can be used both to assess normality and visualize skewness. Make a normal probability plot for the total carbohydrates from a restaurant of your choice. Based on this normal probability plot, is this variable left skewed, symmetric, or right skewed?
Use a histogram to confirm your findings.

This QQ plot looks like the carbs in arbys food could be normally distributed

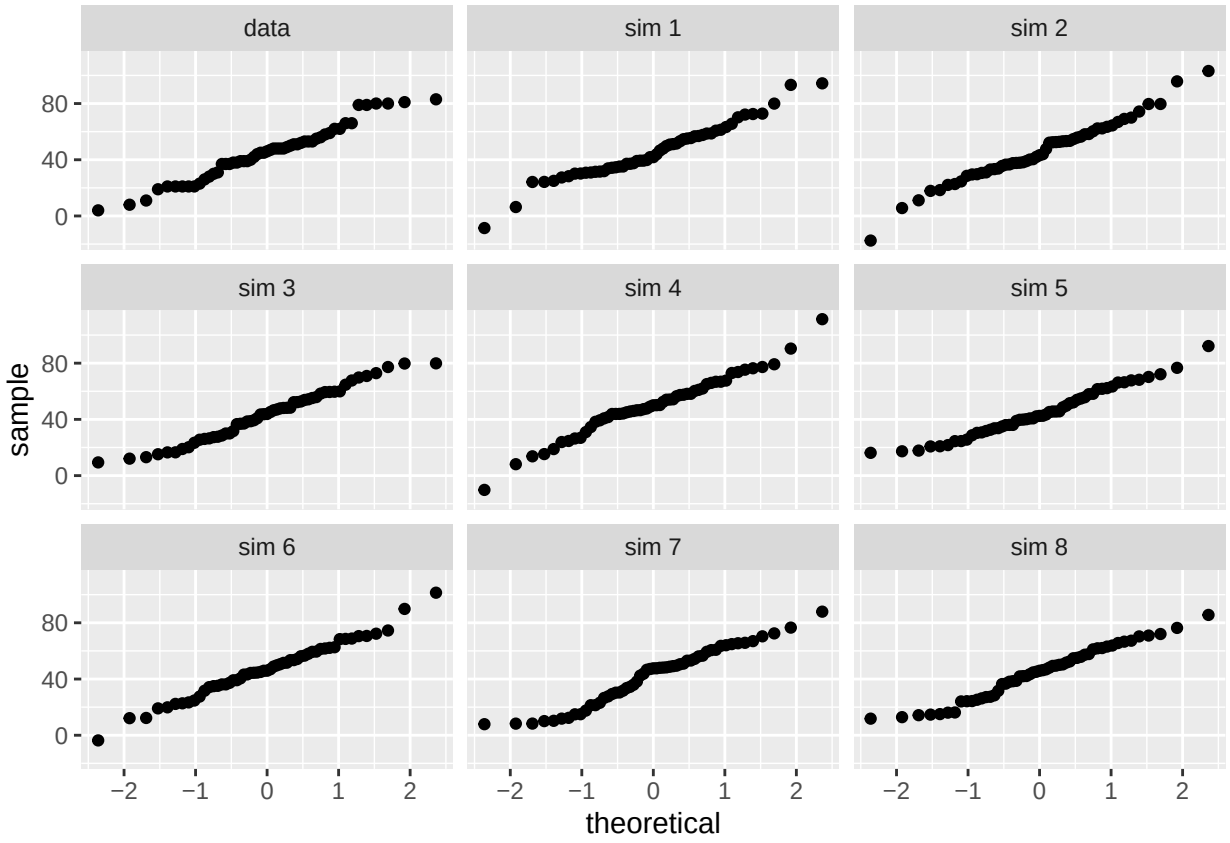
```
arbys <- fastfood %>%
  filter(restaurant == "Arbys")

ggplot(data = arbys, aes(sample = total_carb)) +
  geom_line(stat = "qq")+
  stat_qq_line(alpha = 0.7, color = "red", linetype = "dashed")
```



confirming with qqnormsim

```
qqnormsim(sample = total_carb, data = arbys)
```

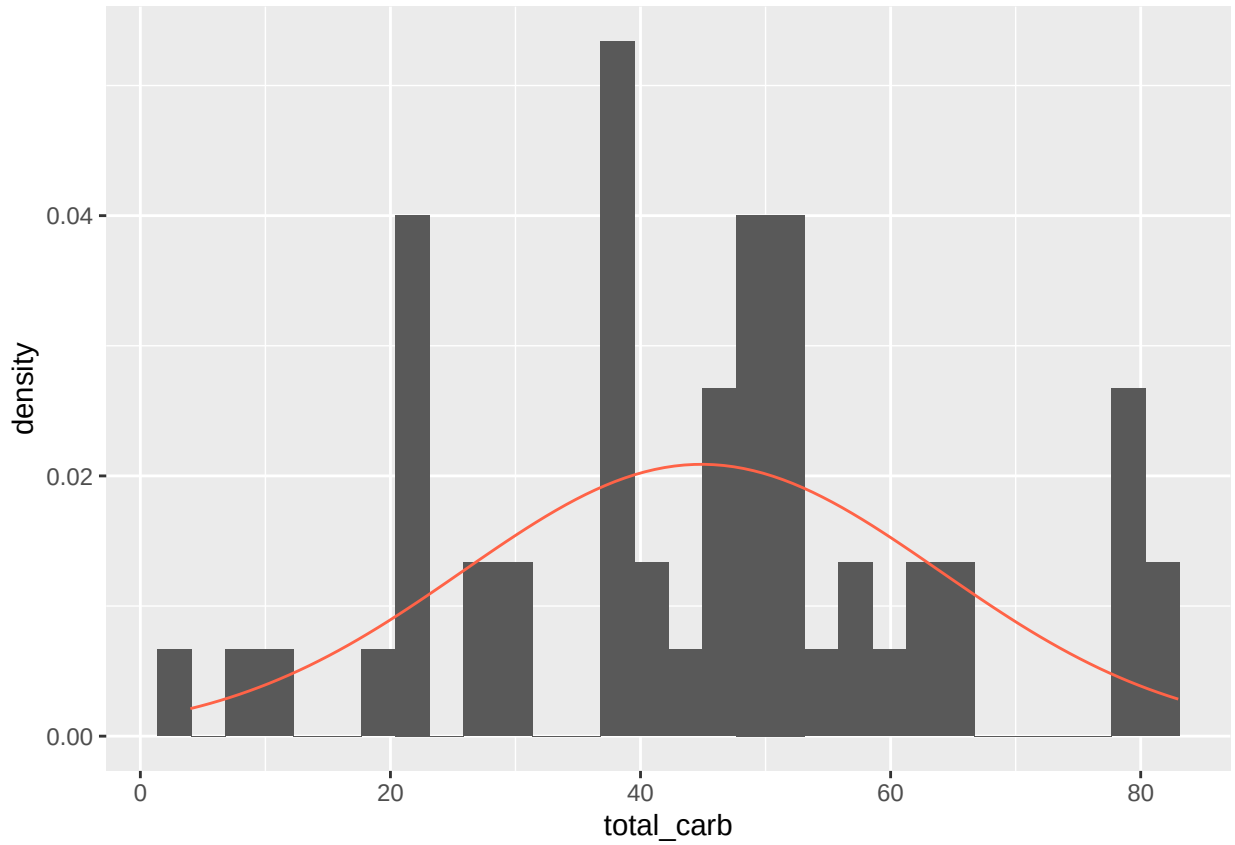


This looks normal so I will plot with a a histogram to check

```
arbmean<-mean(arbys$total_carb)
arbsd<-sd(arbys$total_carb)

ggplot(data = arbys, aes(x = total_carb)) +
        geom_blank() +
        geom_histogram(aes(y = ..density..)) +
        stat_function(fun = dnorm, args = c(mean = arbmean, sd = arbsd), col = "tomato")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


density
0.04
0.02
0.00
0
20
40
60
80
total_carb