The normal distribution - Lab 4

Avery Davidowitz

Load packages

```
library(tidyverse)
library(openintro)
```

Load data

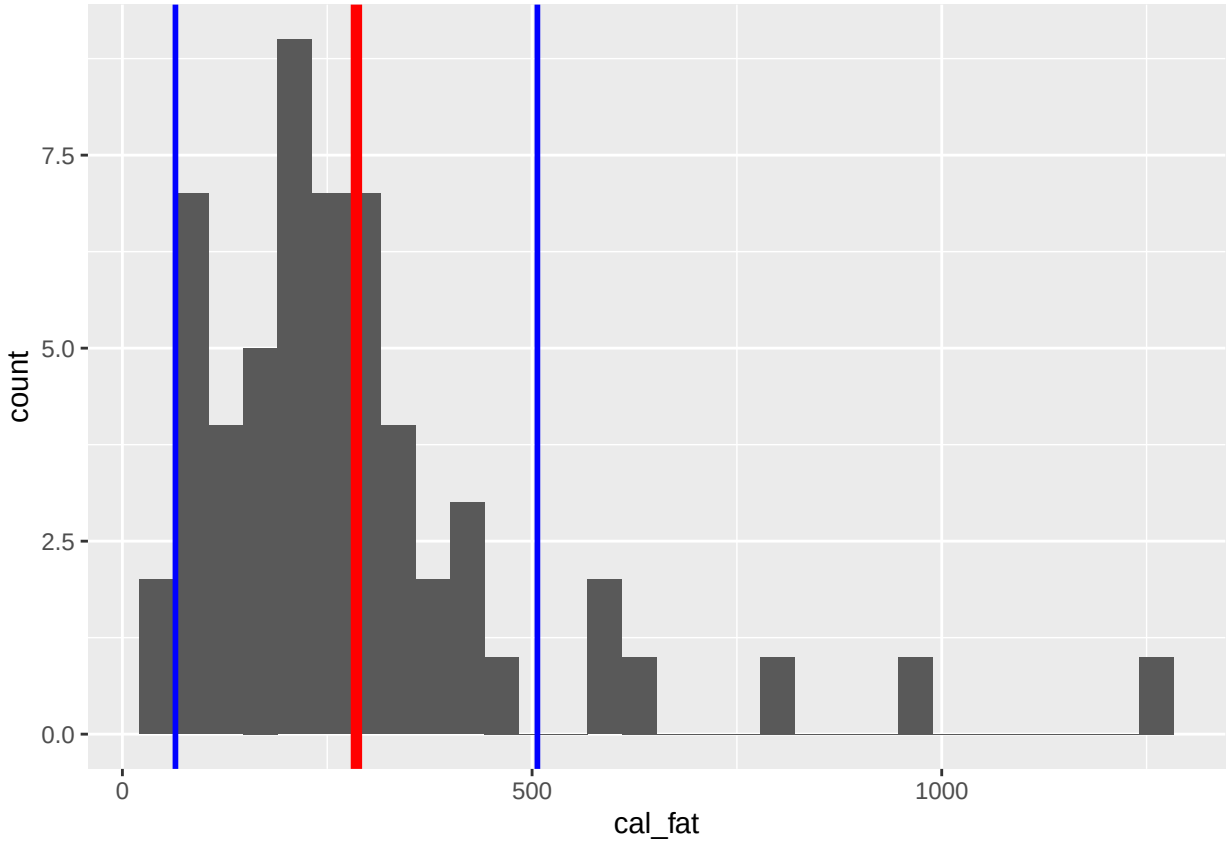
```
data("fastfood", package='openintro')
head(fastfood)
```

```
## # A tibble: 6 x 17
##   restaur~1 item  calor~2 cal_fat total~3 sat_fat trans~4 chole~5 sodium total~6
##   <chr>     <chr>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>  <dbl>   <dbl>
## 1 Mcdonalds Arti~     380      60       7       2     0        95   1110      44
## 2 Mcdonalds Sing~     840     410      45      17     1.5     130   1580      62
## 3 Mcdonalds Doub~    1130     600      67      27     3       220   1920      63
## 4 Mcdonalds Gril~     750     280      31      10     0.5     155   1940      62
## 5 Mcdonalds Cris~     920     410      45      12     0.5     120   1980      81
## 6 Mcdonalds Big ~     540     250      28      10     1        80    950      46
## # ... with 7 more variables: fiber <dbl>, sugar <dbl>, protein <dbl>,
## #   vit_a <dbl>, vit_c <dbl>, calcium <dbl>, salad <chr>, and abbreviated
## #   variable names 1: restaurant, 2: calories, 3: total_fat, 4: trans_fat,
## #   5: cholesterol, 6: total_carb
```

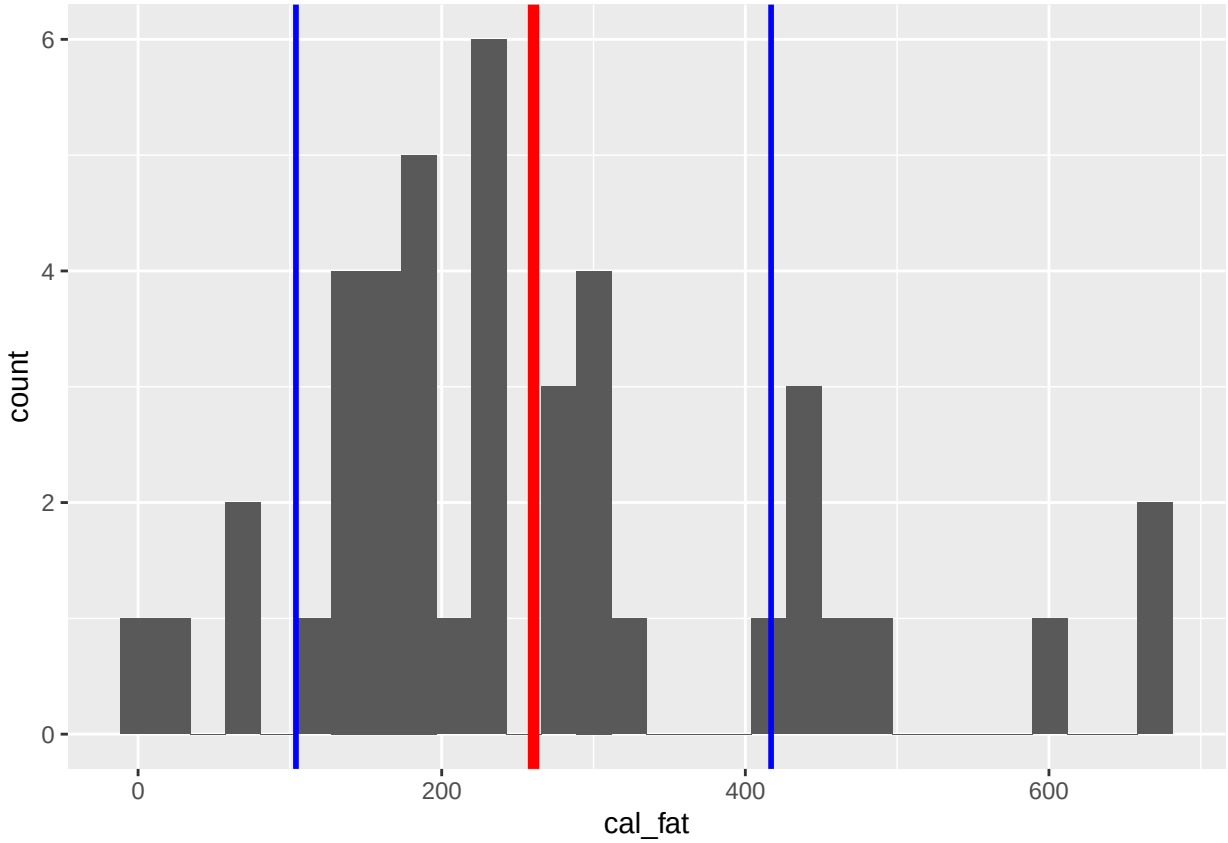
Exercise 1

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?

```
mcdonalds <- fastfood %>%
  filter(restaurant == "Mcdonalds")
dairy_queen <- fastfood %>%
  filter(restaurant == "Dairy Queen")
ggplot(mcdonalds, aes(x=cal_fat)) + geom_histogram() +
  geom_vline(aes(xintercept = mean(cal_fat)),col='red',size=2) +
  geom_vline(aes(xintercept = mean(cal_fat) - sd(cal_fat)),col='blue',size=1) +
  geom_vline(aes(xintercept = mean(cal_fat) + sd(cal_fat)),col='blue',size=1)
```




```
ggplot(dairy_queen, aes(x=cal_fat)) + geom_histogram() +
  geom_vline(aes(xintercept = mean(cal_fat)),col='red',size=2) +
  geom_vline(aes(xintercept = mean(cal_fat) - sd(cal_fat)),col='blue',size=1) +
  geom_vline(aes(xintercept = mean(cal_fat) + sd(cal_fat)),col='blue',size=1)
```



The histograms have lines added for the mean in red and +/- one standard deviation displayed in blue. The McDonald's distribution of fat calories per dish is skewed right. The distribution is unimodal but has a long right tail of outliers. Without those outliers above 500 calories, the McDonalds data is relatively bell shaped.

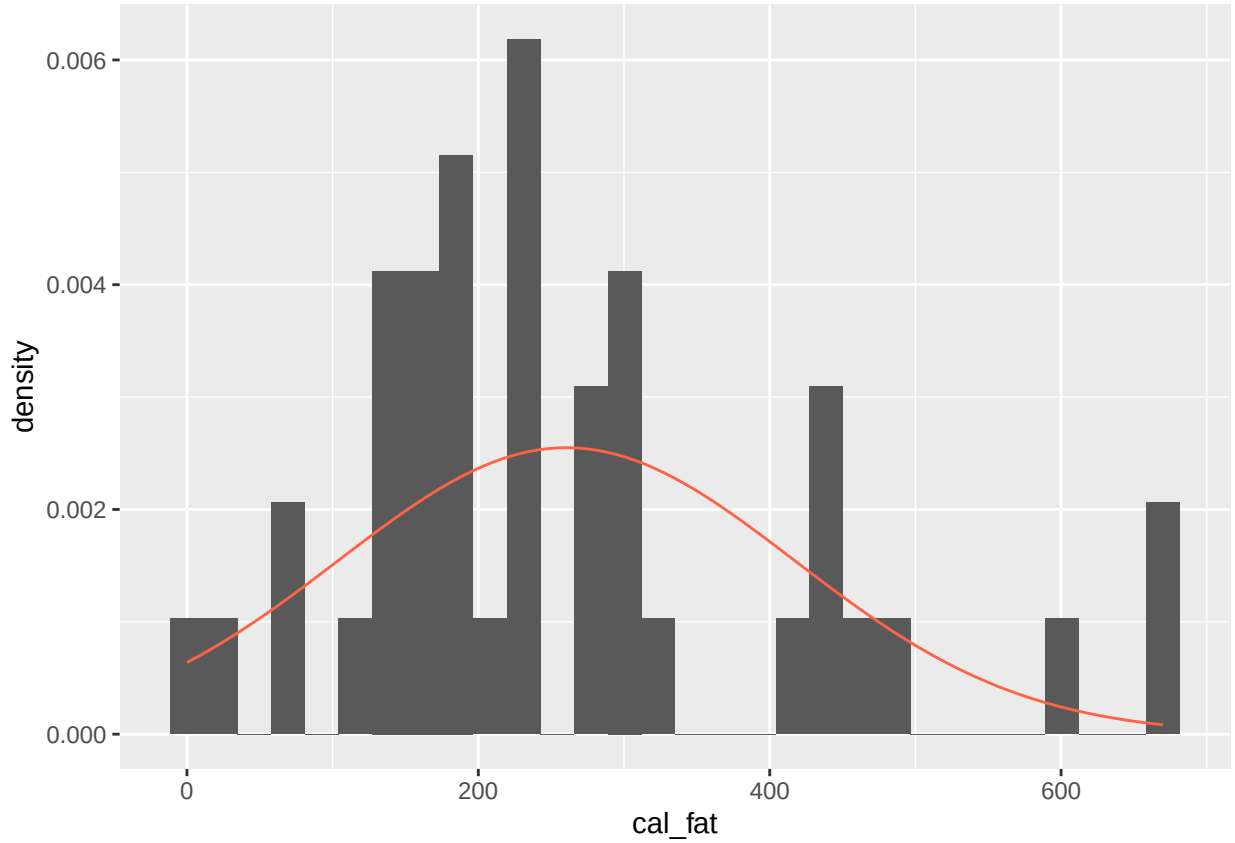
The Dairy Queen plot for calories from fat is much more symmetric about the mean. The Dairy Queen data is also unimodal having a clear mode close to the mean.

Both data sets have very similar spread with standard deviations of 160 and 156.

Exercise 2

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

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

ggplot(data = dairy_queen, aes(x = cal_fat)) +
        geom_blank() +
        geom_histogram(aes(y = ..density..)) +
        stat_function(fun = dnorm, args = c(mean = dqmean, sd = dqsd), col = "tomato")
```

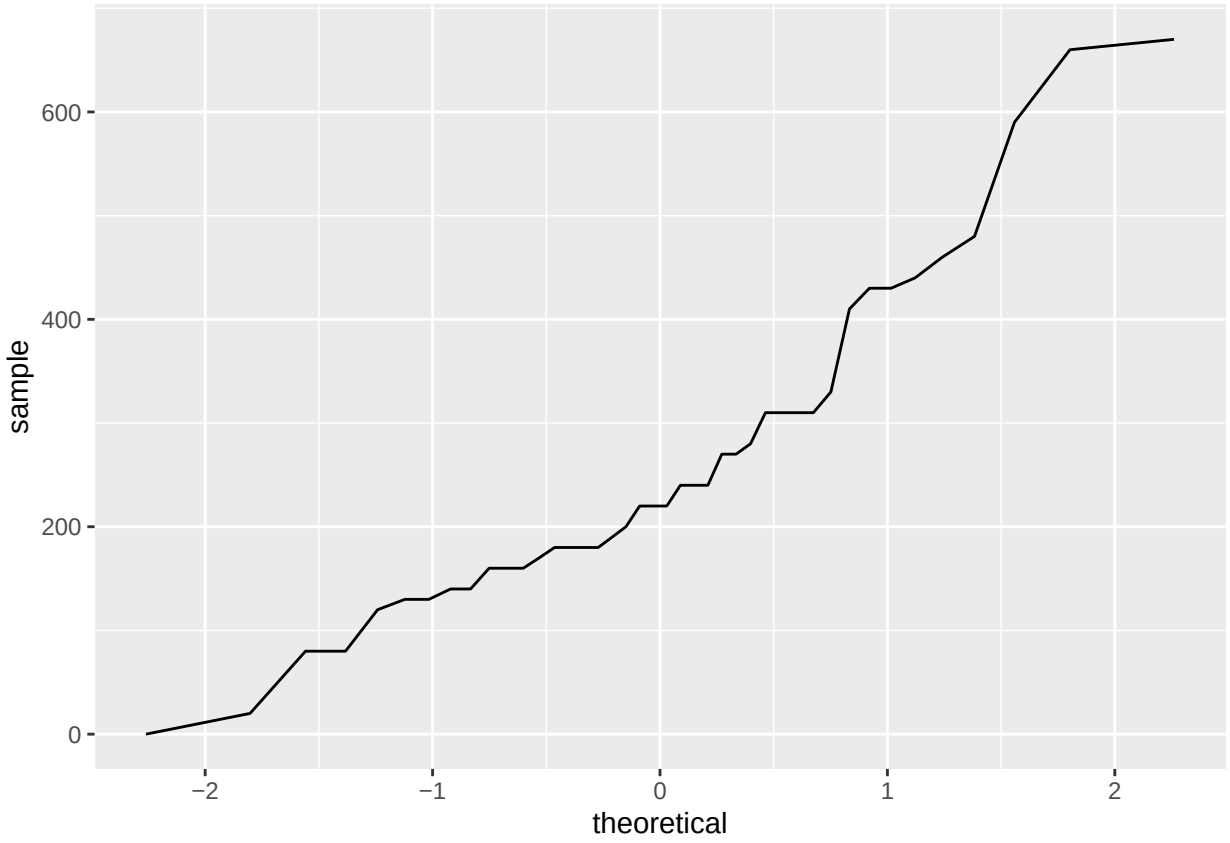


Although not perfectly centered, the plot strongly resembles a normal curve.

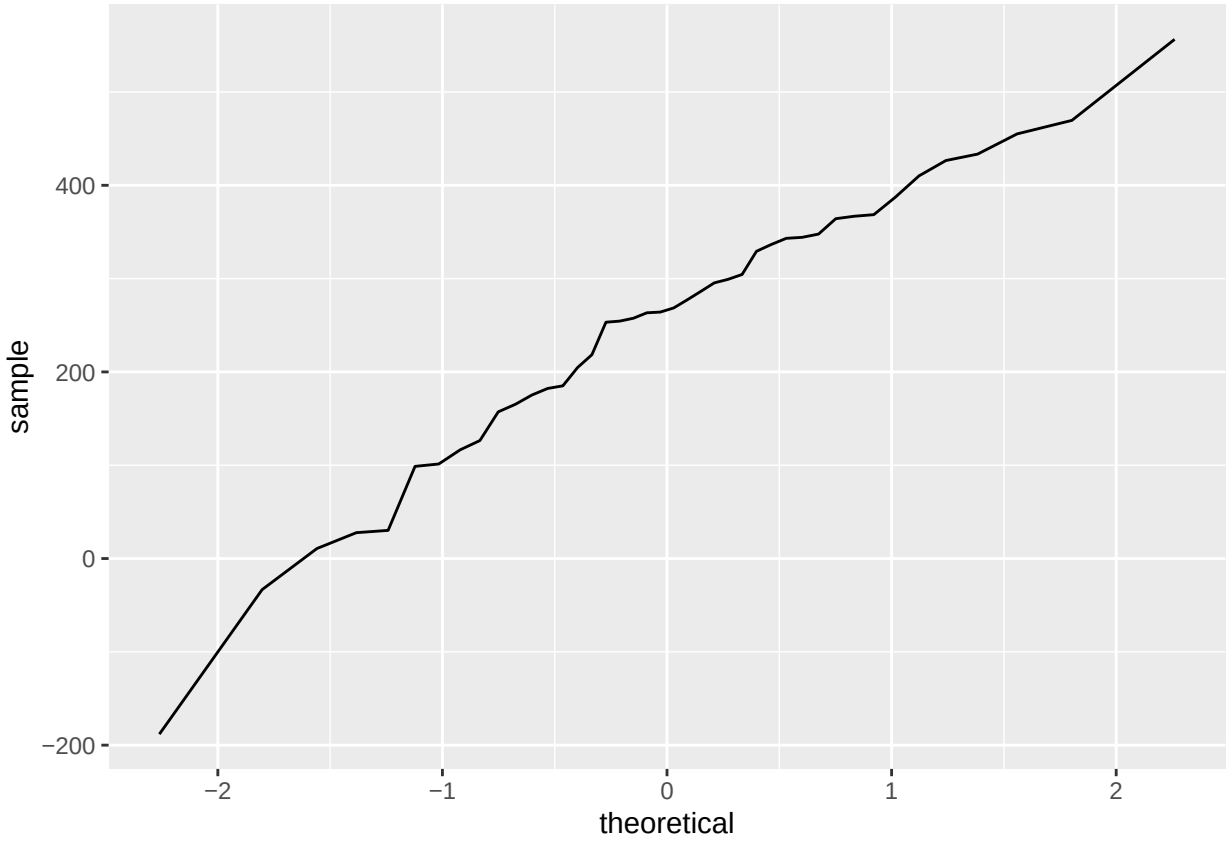
Exercise 3

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 data frame, it can be put directly into the `sample` argument and the `data` argument can be dropped.)

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

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




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

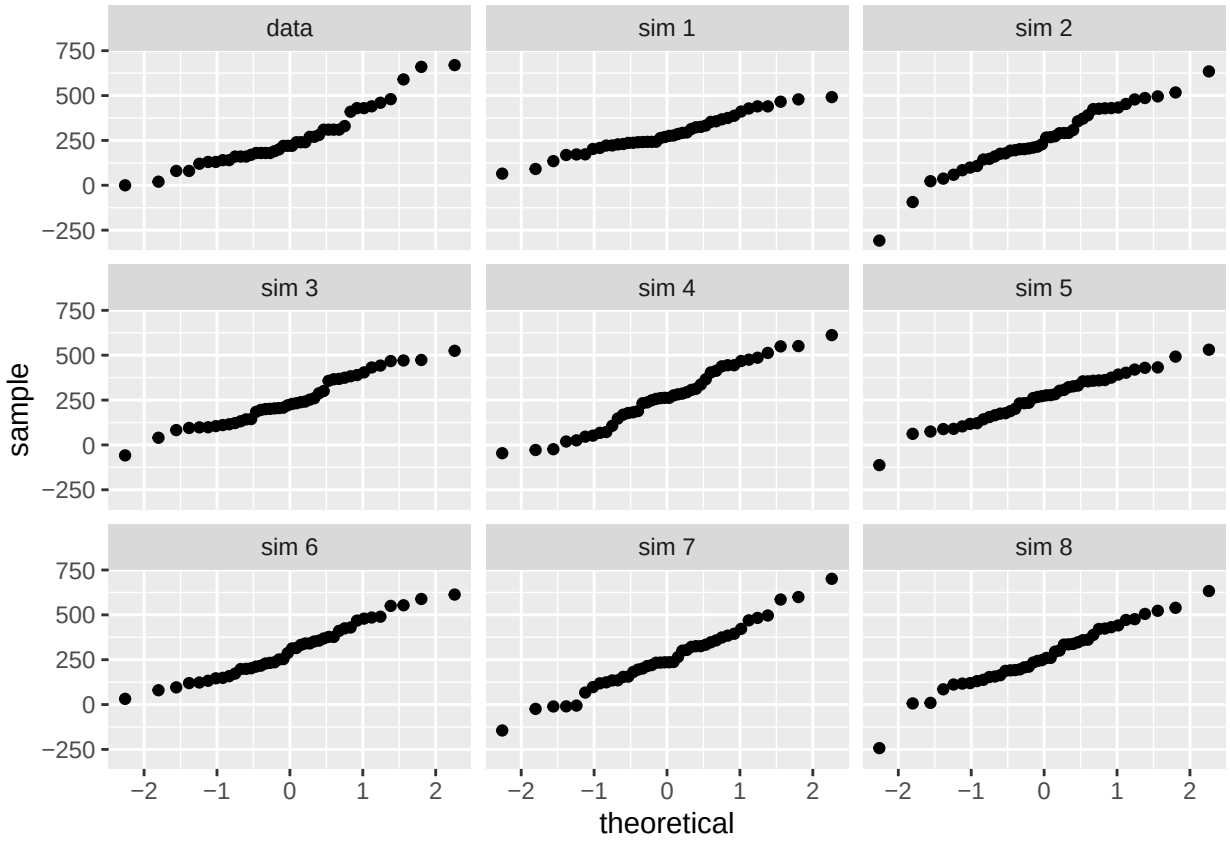


The simulated sample data resembles the Dairy Queen data and deviates similarly from the straight line expected from a normal distribution QQ plot.

Exercise 4

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 calories are nearly normal?

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

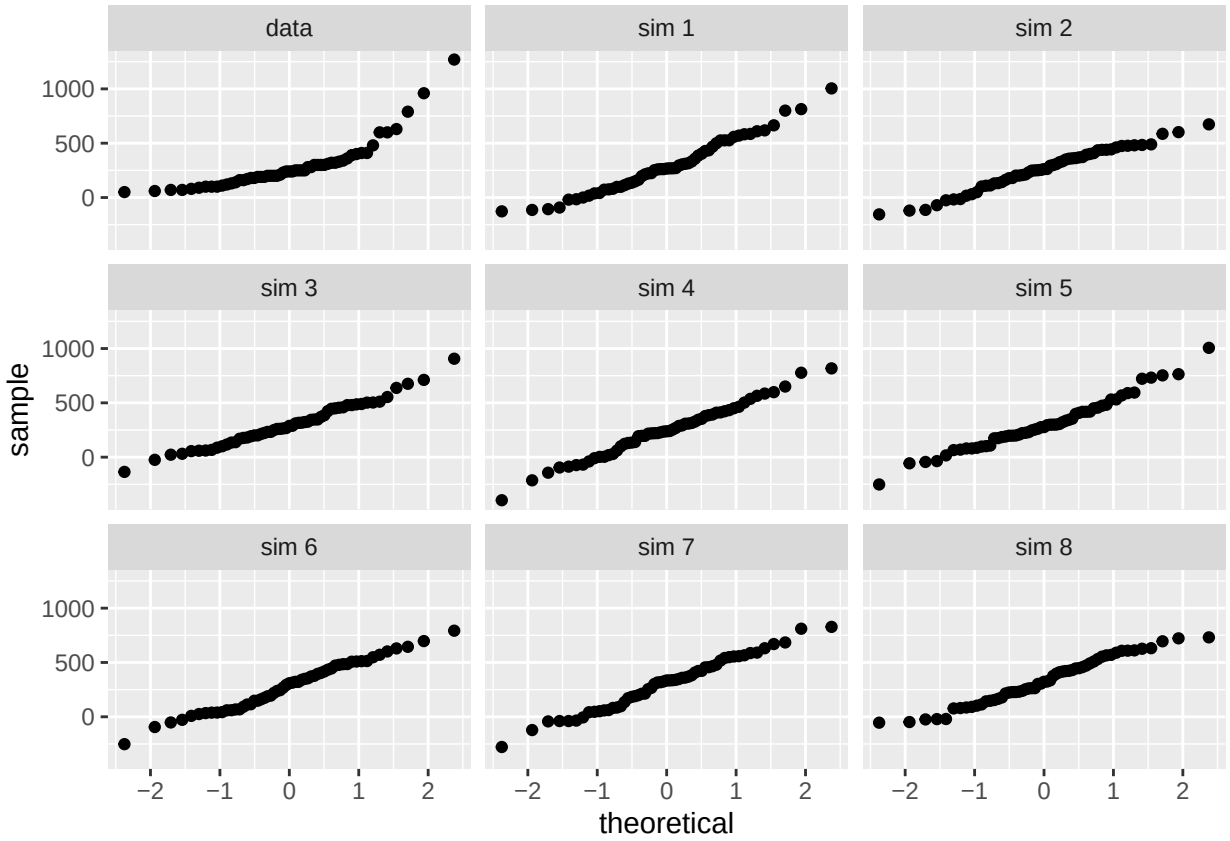


The Dairy Queen plot strongly resembles the overall trend of the simulations. The plots provide moderately strong evidence of a case for a nearly normal distribution.

Exercise 5

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

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



As suspected, the McDonald's data does not resemble any of the normal simulation QQ plots.

Exercise 6

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?

What is the probability that a McDonald's item has less than 2 grams of fiber? What is the probability that a McDonald's item has more than 4 grams of fiber?

Using the assumed theoretical normal distribution:

```
mcdmean <- mean(mcdonalds$fiber)
mcdsd   <- sd(mcdonalds$fiber)

fiber_under2_norm <- pnorm(2, mean = mcdmean, sd = mcdsd)
fiber_over4_norm <- 1 - pnorm(4, mean = mcdmean, sd = mcdsd)
```

Using the empirical distribution:

```
fiber_qty2 <- mcdonalds %>% dplyr::filter(fiber < 2) %>% nrow()
fiber_under2_freq <- fiber_qty2 / nrow(mcdonalds)
```

```
fiber_qty4 <- mcdonalds %>% dplyr::filter(fiber > 4) %>% nrow()
fiber_over4_freq <- (fiber_qty4 / nrow(mcdonalds))
```

Comparing both methods yields a minimal difference of:

```
fiber_under2_norm - fiber_under2_freq
```

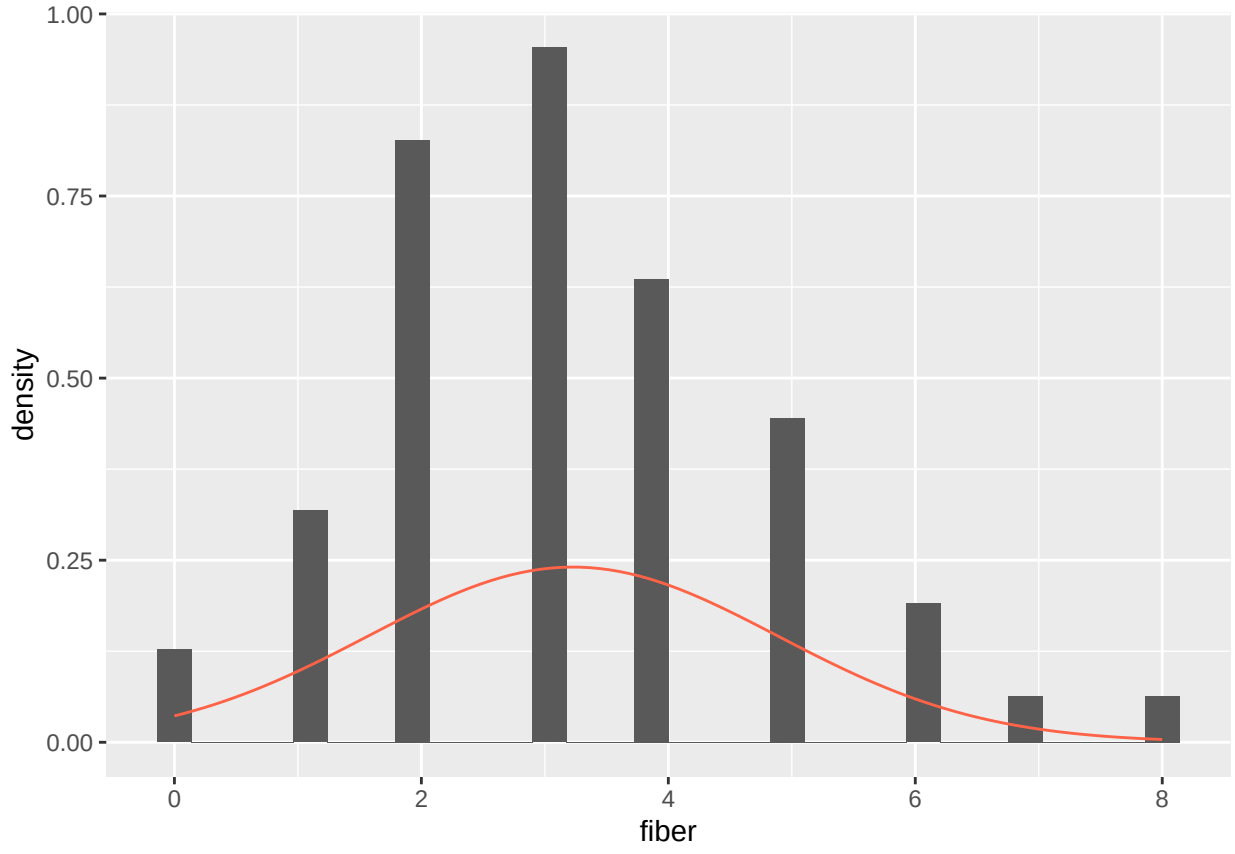
```
## [1] 0.1067005
```

```
fiber_over4_norm - fiber_over4_freq
```

```
## [1] 0.1102825
```

As shown in the histogram below, the distribution for fiber among McDonald's items is nearly normal. As a result both methods are only off by 10% and 11% by using an assumed theoretical normal distribution.

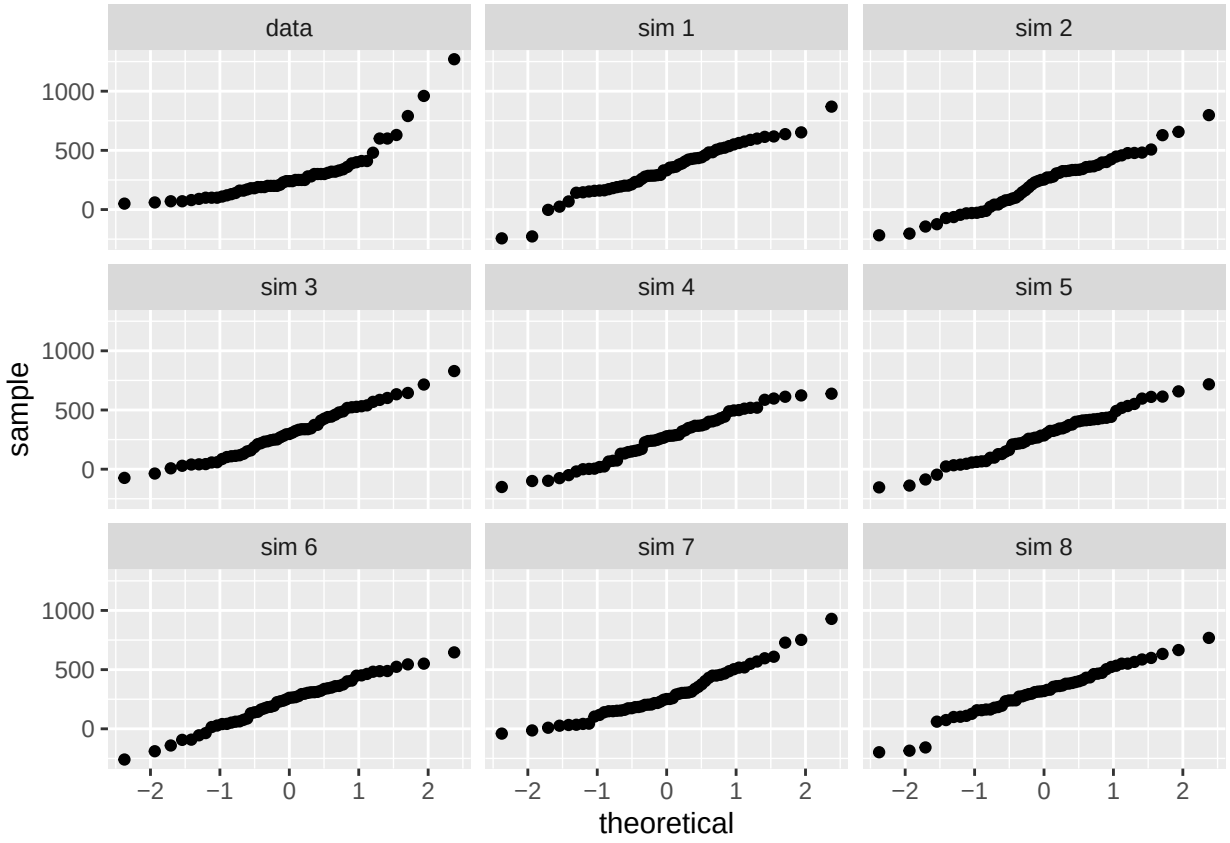
```
ggplot(data = mcdonalds, aes(x = fiber)) +
        geom_blank() +
        geom_histogram(aes(y = ..density..)) +
        stat_function(fun = dnorm, args = c(mean = mcdmean, sd = mcdsd), col = "tomato")
```



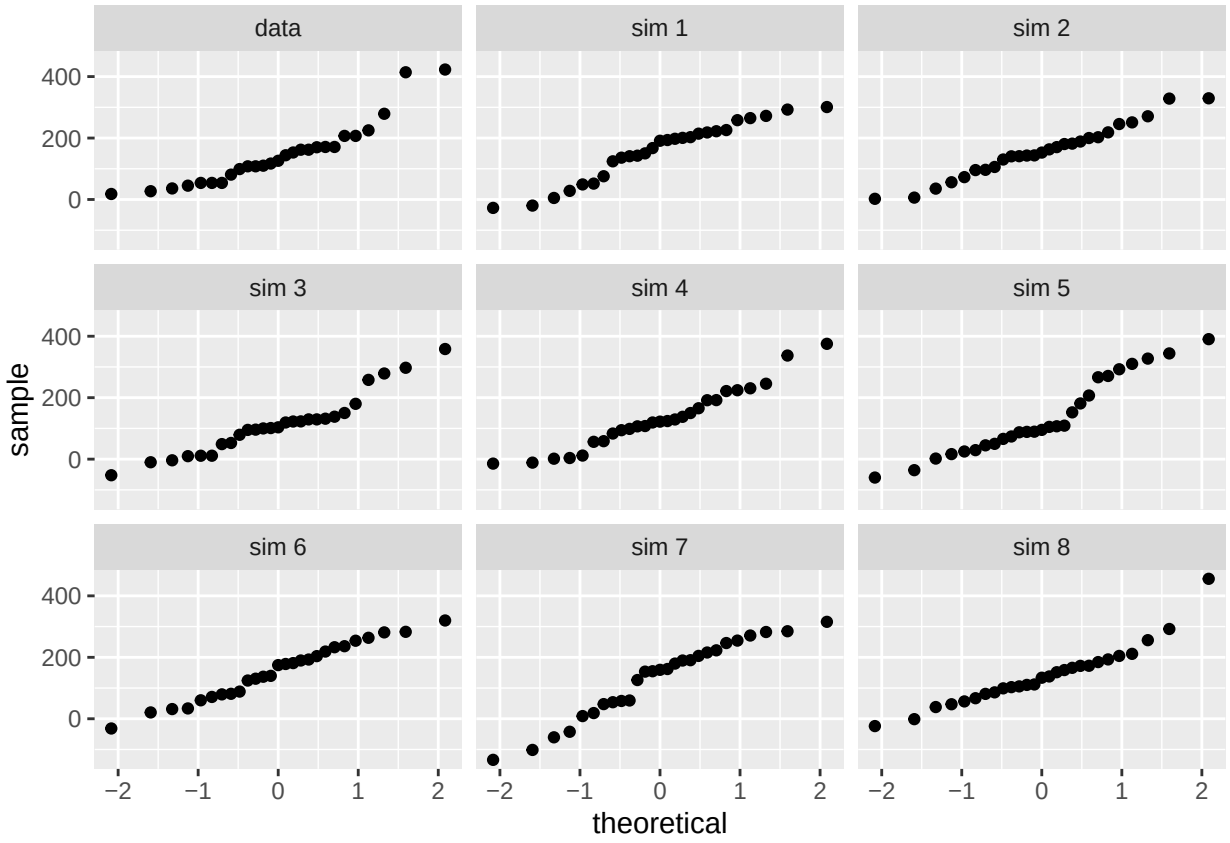
Exercise 7

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?

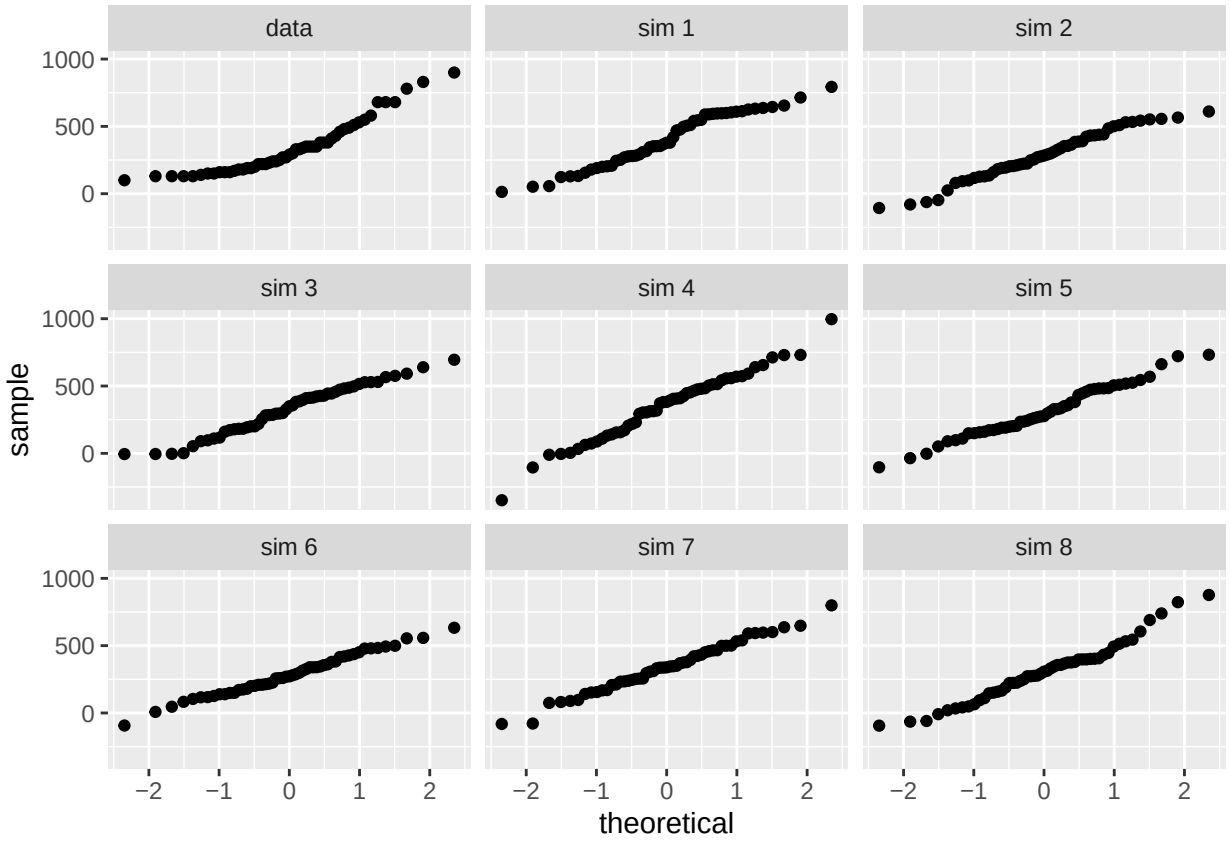
```
taco_bell <- fastfood %>%
  filter(restaurant == "Taco Bell")
subway <- fastfood %>%
  filter(restaurant == "Subway")
burger_king <- fastfood %>%
  filter(restaurant == "Burger King")
arbys <- fastfood %>%
  filter(restaurant == "Arbys")
sonic <- fastfood %>%
  filter(restaurant == "Sonic")
chick <- fastfood %>%
  filter(restaurant == "Chick Fil-A")

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



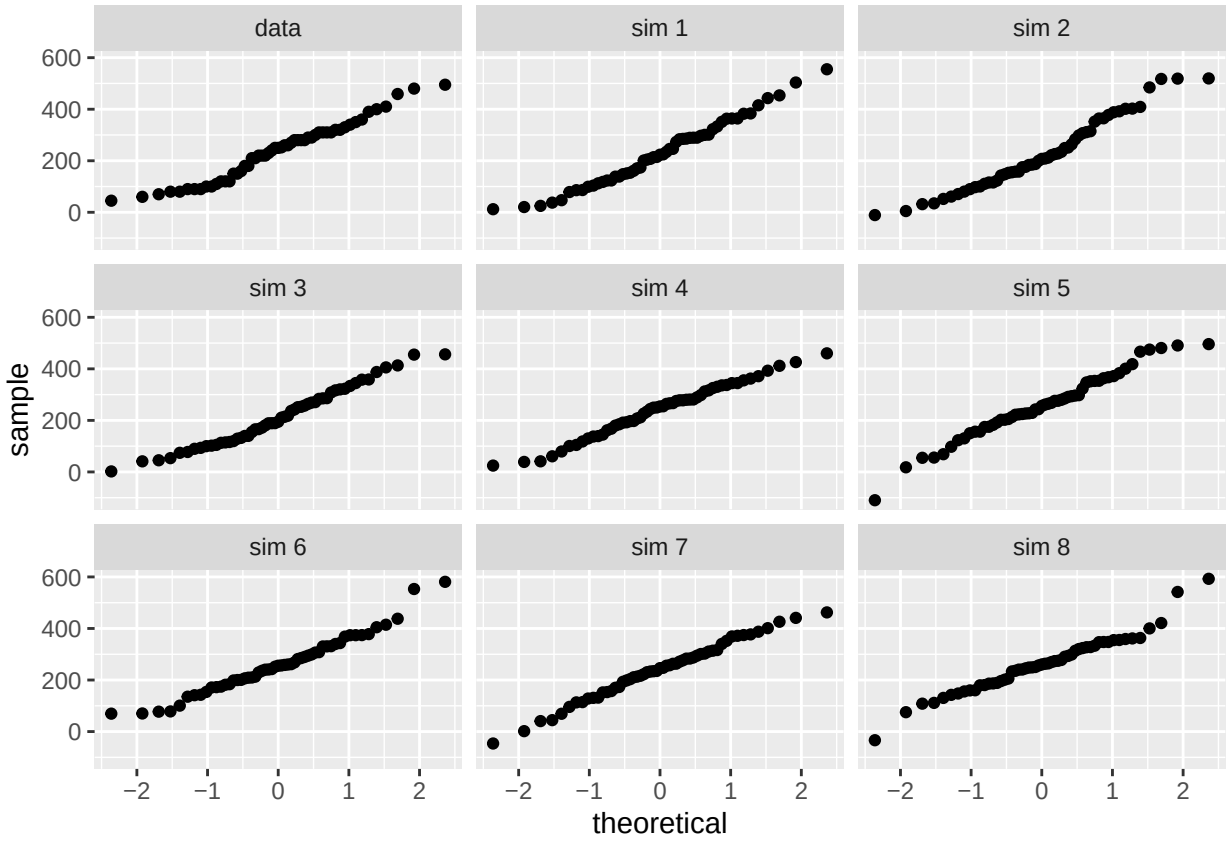

```
qqnormsim(sample = cal_fat, data = chick)
```



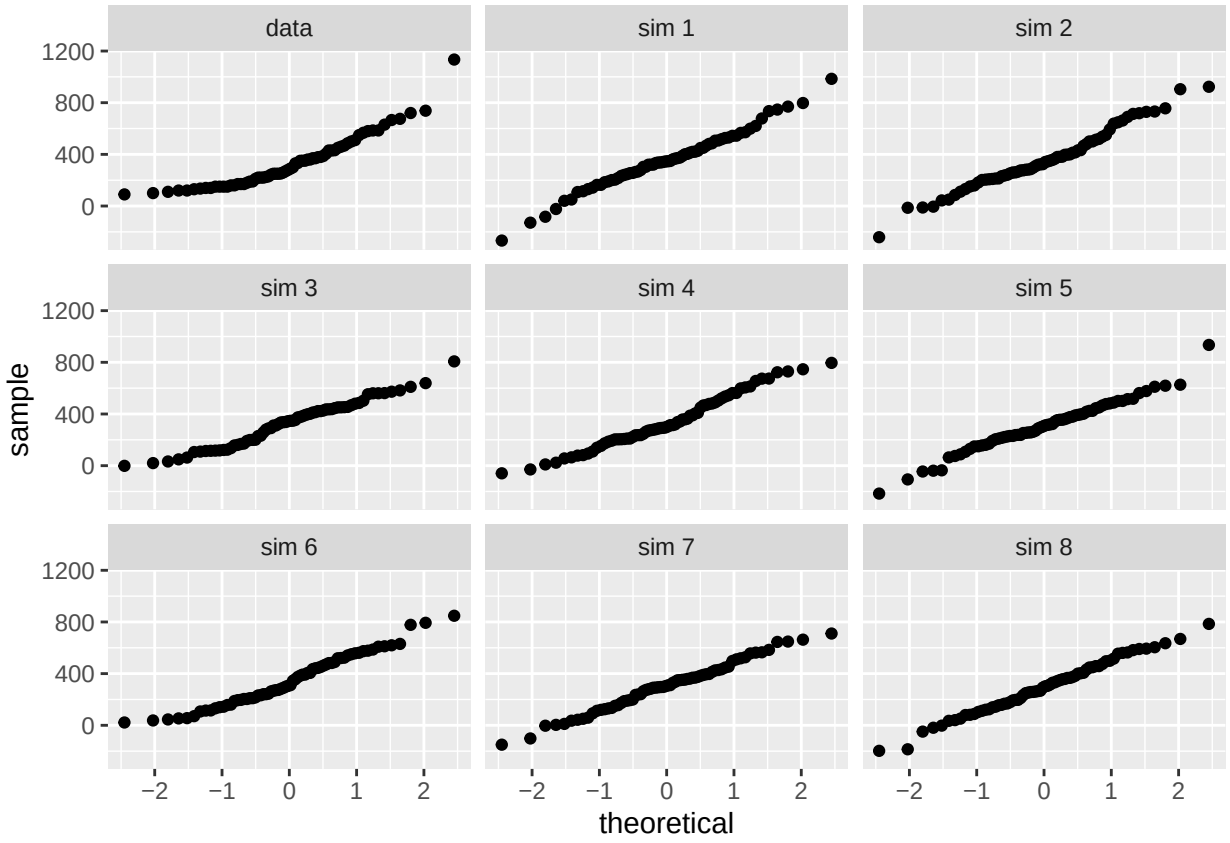

```
qqnormsim(sample = cal_fat, data = sonic)
```



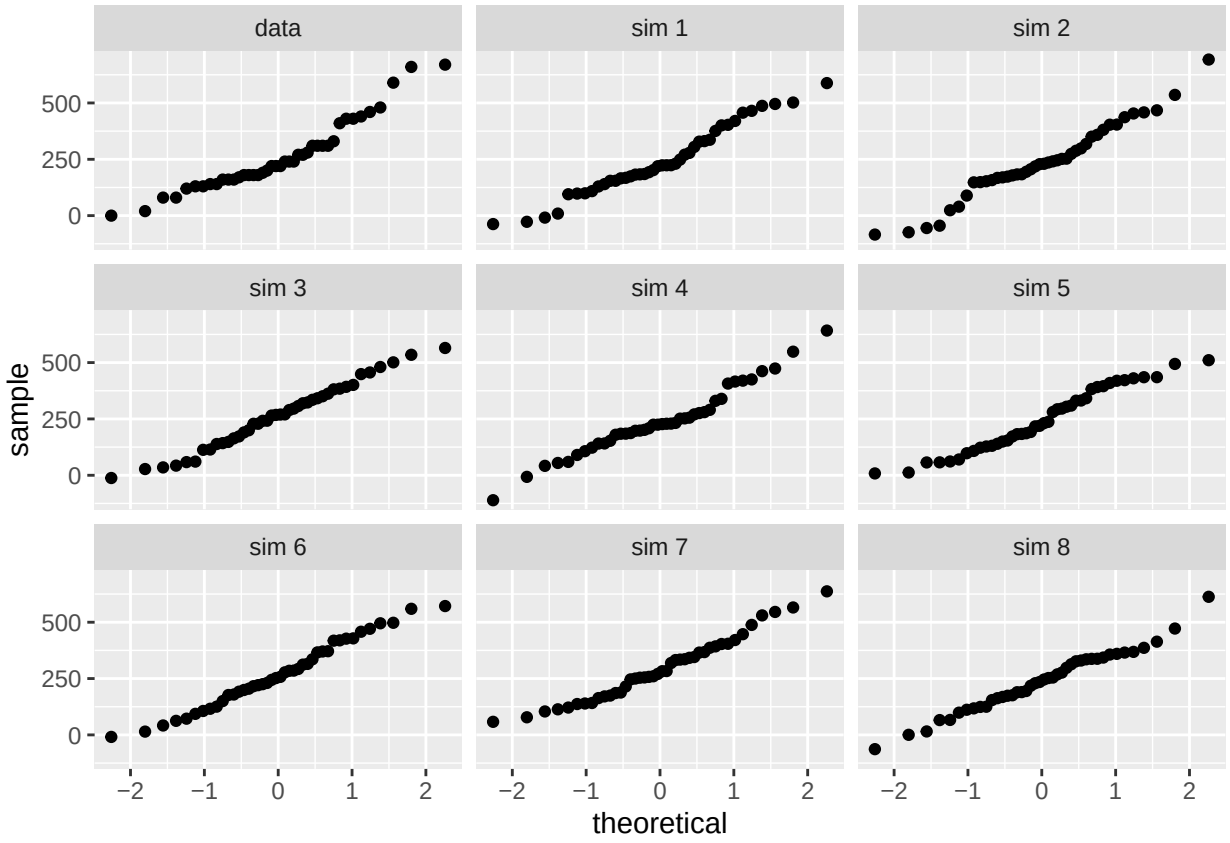

```
qqnormsim(sample = cal_fat, data = arbys)
```



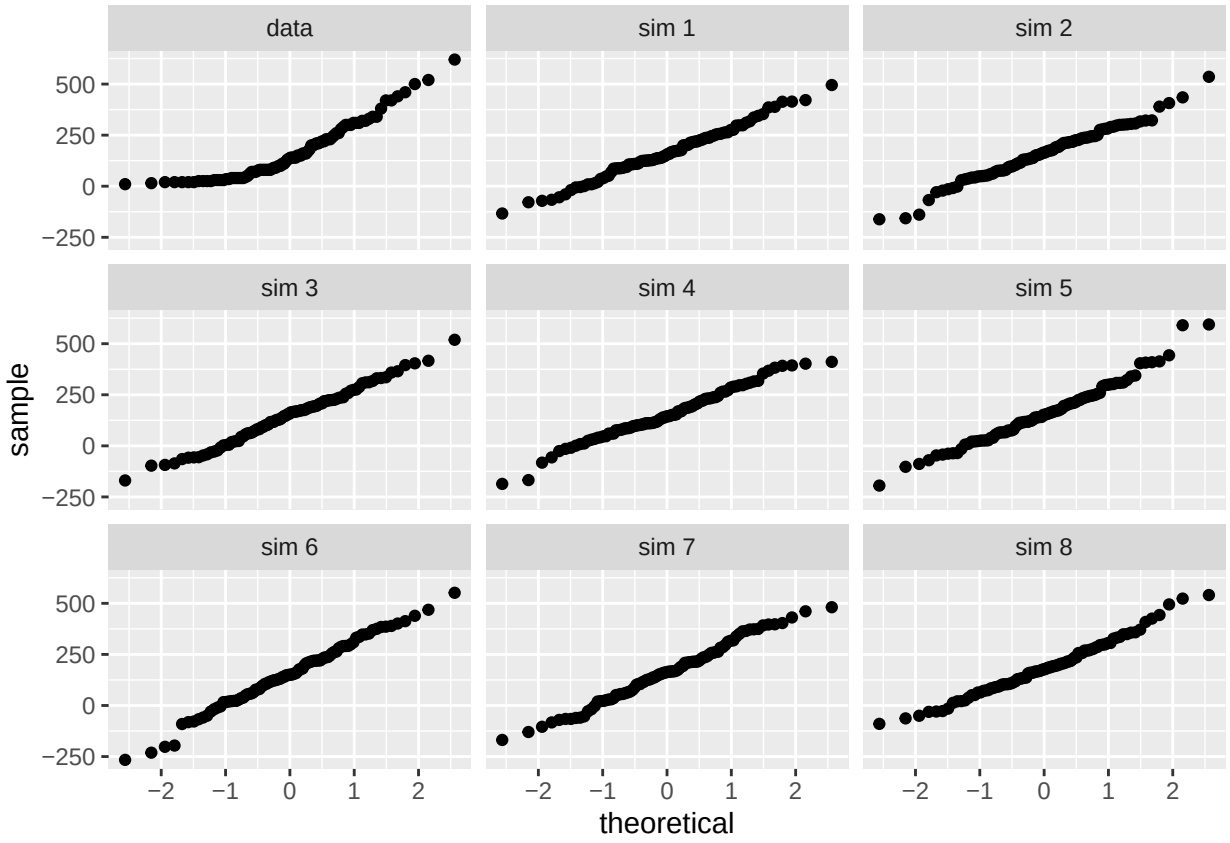

```
qqnormsim(sample = cal_fat, data = burger_king)
```



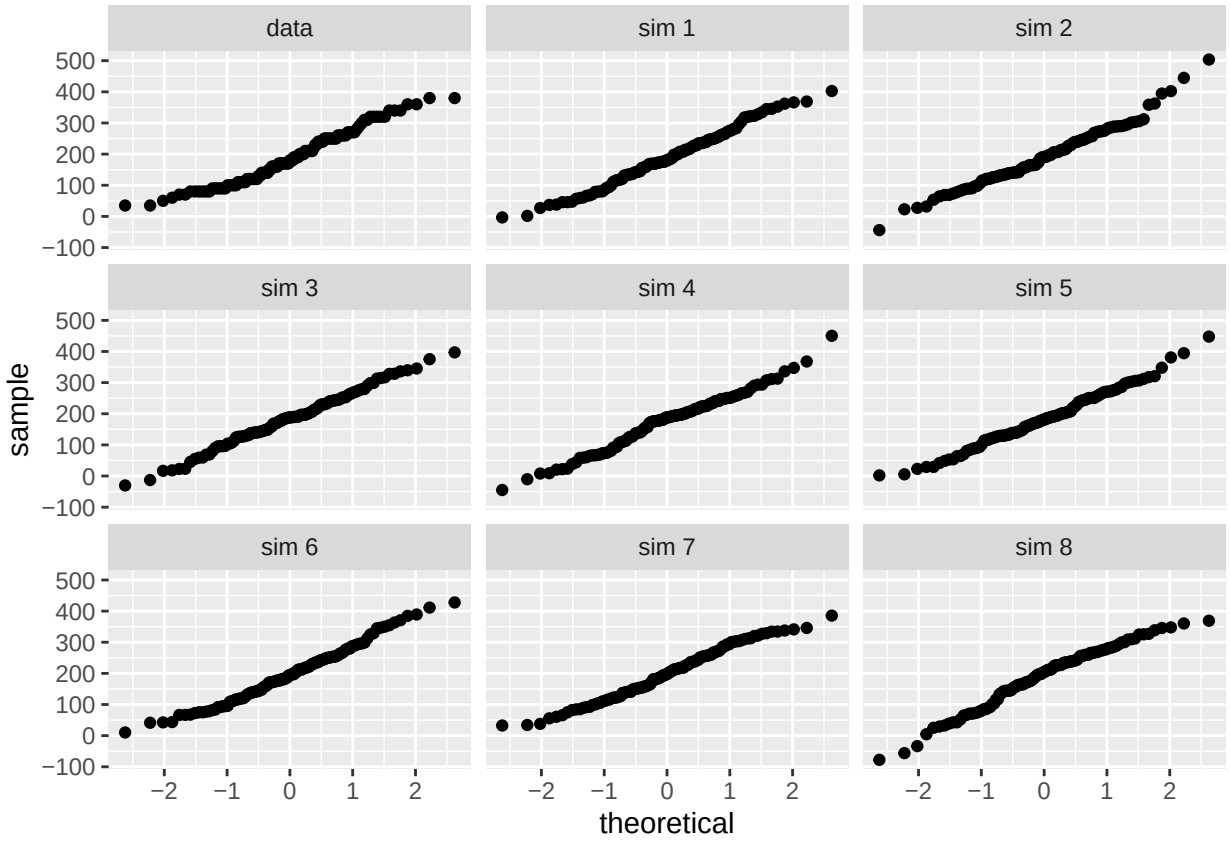

```
qqnormsim(sample = cal_fat, data = dairy_queen)
```




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




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



Based on the visual inspection of the straightness of the QQ plots, Taco Bell has the most nearly normal distribution of sodium in their dishes.

Exercise 8

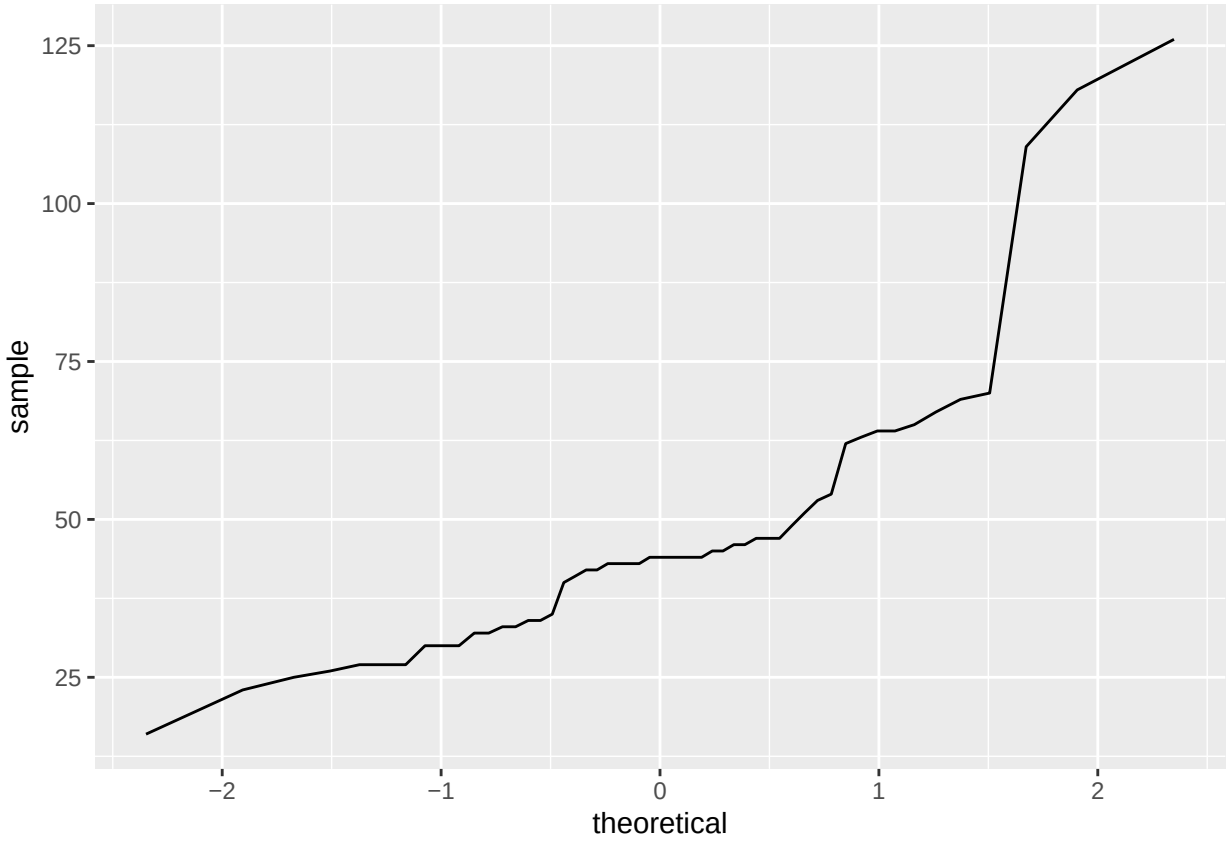
8. Note that some of the normal probability plots for sodium distributions seem to have a stepwise pattern. why do you think this might be the case?

The sample sizes across the various chains differ significantly. Taco Bell may appear smoother due to the larger quantity of 115 observations. Chick Fil-A appears more step wise in the sodium plots and that may be due to the fact that there are only 27 observations.

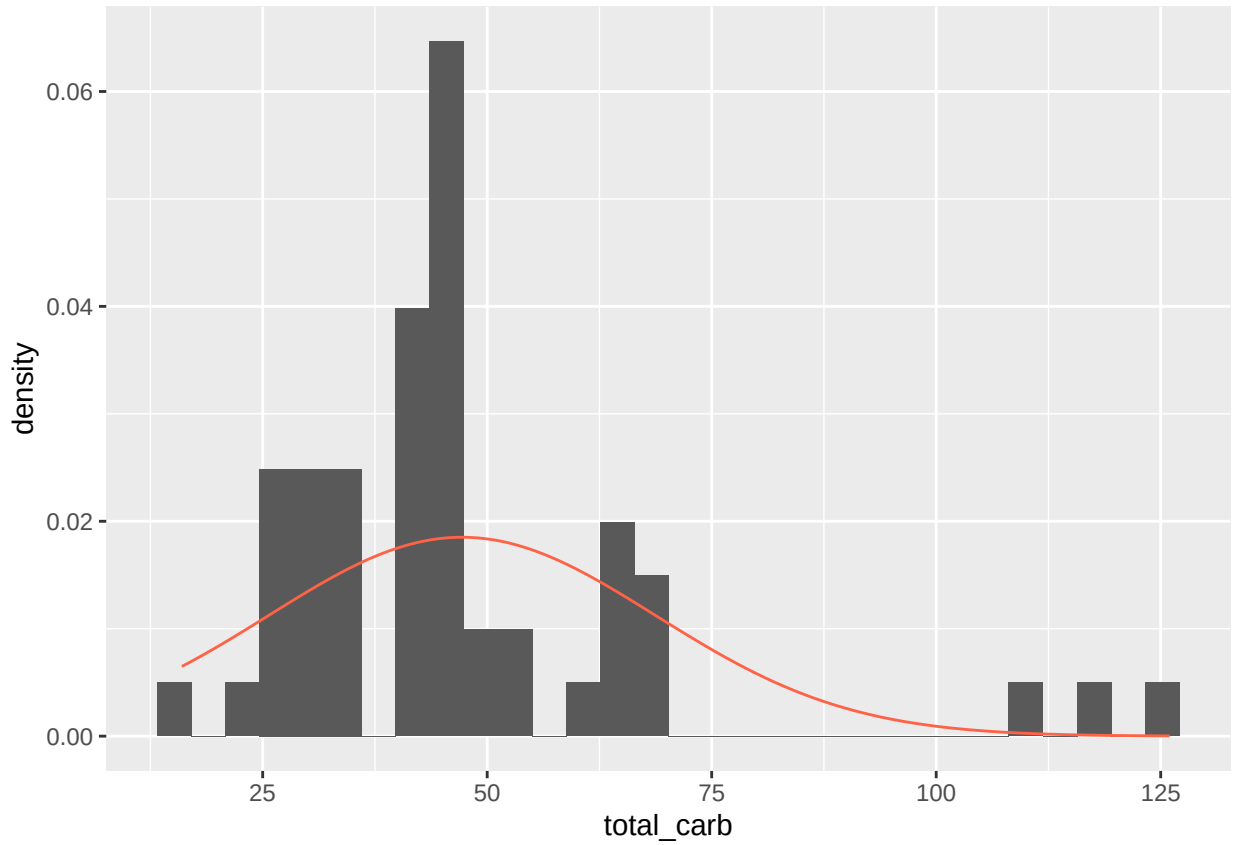
Exercise 9

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.

```
ggplot(data = sonic, aes(sample = total_carb)) +
  geom_line(stat = "qq")
```




```
ggplot(data = sonic, aes(x = total_carb)) +
        geom_blank() +
        geom_histogram(aes(y = ..density..)) +
        stat_function(fun = dnorm, args = c(mean = mean(sonic$total_carb), sd = sd(sonic$total_carb)),
```



The total carbohydrates in Sonic dishes is skewed right as shown in the Q-Q plot and histogram.