Lab 4: The Normal Distribution

Code

Julian Adames-Ng

2022-02-27

Hide

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

```
## # A tibble: 6 × 17
##   restaurant item       calories cal_fat total_fat sat_fat trans_fat cholesterol
##   <chr>      <chr>         <dbl>   <dbl>     <dbl>   <dbl>     <dbl>       <dbl>
## 1 Mcdonalds  Artisan G…      380      60         7       2       0            95
## 2 Mcdonalds  Single Ba…      840     410        45      17       1.5         130
## 3 Mcdonalds  Double Ba…     1130     600        67      27       3           220
## 4 Mcdonalds  Grilled B…      750     280        31      10       0.5         155
## 5 Mcdonalds  Crispy Ba…      920     410        45      12       0.5         120
## 6 Mcdonalds  Big Mac         540     250        28      10       1            80
## # … with 9 more variables: sodium <dbl>, total_carb <dbl>, fiber <dbl>,
## #   sugar <dbl>, protein <dbl>, vit_a <dbl>, vit_c <dbl>, calcium <dbl>,
## #   salad <chr>
```

Hide

```
mcdonalds <- fastfood %>%
  filter(restaurant == "Mcdonalds")

mcdonalds
```

```
## # A tibble: 57 × 17
##    restaurant item      calories cal_fat total_fat sat_fat trans_fat cholesterol
##    <chr>      <chr>        <dbl>   <dbl>     <dbl>   <dbl>     <dbl>       <dbl>
##  1 Mcdonalds  Artisan …      380      60         7       2         0          95
##  2 Mcdonalds  Single B…      840     410        45      17         1.5       130
##  3 Mcdonalds  Double B…     1130     600        67      27         3         220
##  4 Mcdonalds  Grilled …      750     280        31      10         0.5       155
##  5 Mcdonalds  Crispy B…      920     410        45      12         0.5       120
##  6 Mcdonalds  Big Mac        540     250        28      10         1          80
##  7 Mcdonalds  Cheesebu…      300     100        12       5         0.5        40
##  8 Mcdonalds  Classic …      510     210        24       4         0          65
##  9 Mcdonalds  Double C…      430     190        21      11         1          85
## 10 Mcdonalds  Double Q…      770     400        45      21         2.5       175
## # … with 47 more rows, and 9 more variables: sodium <dbl>, total_carb <dbl>,
## #   fiber <dbl>, sugar <dbl>, protein <dbl>, vit_a <dbl>, vit_c <dbl>,
## #   calcium <dbl>, salad <chr>
```

Hide

```
dairy_queen <- fastfood %>%
  filter(restaurant == "Dairy Queen")

dairy_queen
```

```
## # A tibble: 42 × 17
##    restaurant  item     calories cal_fat total_fat sat_fat trans_fat cholesterol
##    <chr>       <chr>       <dbl>   <dbl>     <dbl>   <dbl>     <dbl>       <dbl>
##  1 Dairy Queen 1/2 lb.…     1000     660        74      26         2         170
##  2 Dairy Queen 1/2 lb.…      800     460        51      20         2         135
##  3 Dairy Queen 1/4 lb.…      630     330        37      13         1          95
##  4 Dairy Queen 1/4 lb.…      540     270        30      11         1          70
##  5 Dairy Queen 1/4 lb.…      570     310        35      11         1          75
##  6 Dairy Queen Origina…      400     160        18       9         1          65
##  7 Dairy Queen Origina…      630     310        34      18         2         125
##  8 Dairy Queen 4 Piece…     1030     480        53       9         1          80
##  9 Dairy Queen 6 Piece…     1260     590        66      11         1         120
## 10 Dairy Queen Bacon C…      420     240        26      11         1          60
## # … with 32 more rows, and 9 more variables: sodium <dbl>, total_carb <dbl>,
## #   fiber <dbl>, sugar <dbl>, protein <dbl>, vit_a <dbl>, vit_c <dbl>,
## #   calcium <dbl>, salad <chr>
```

Exercise 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?

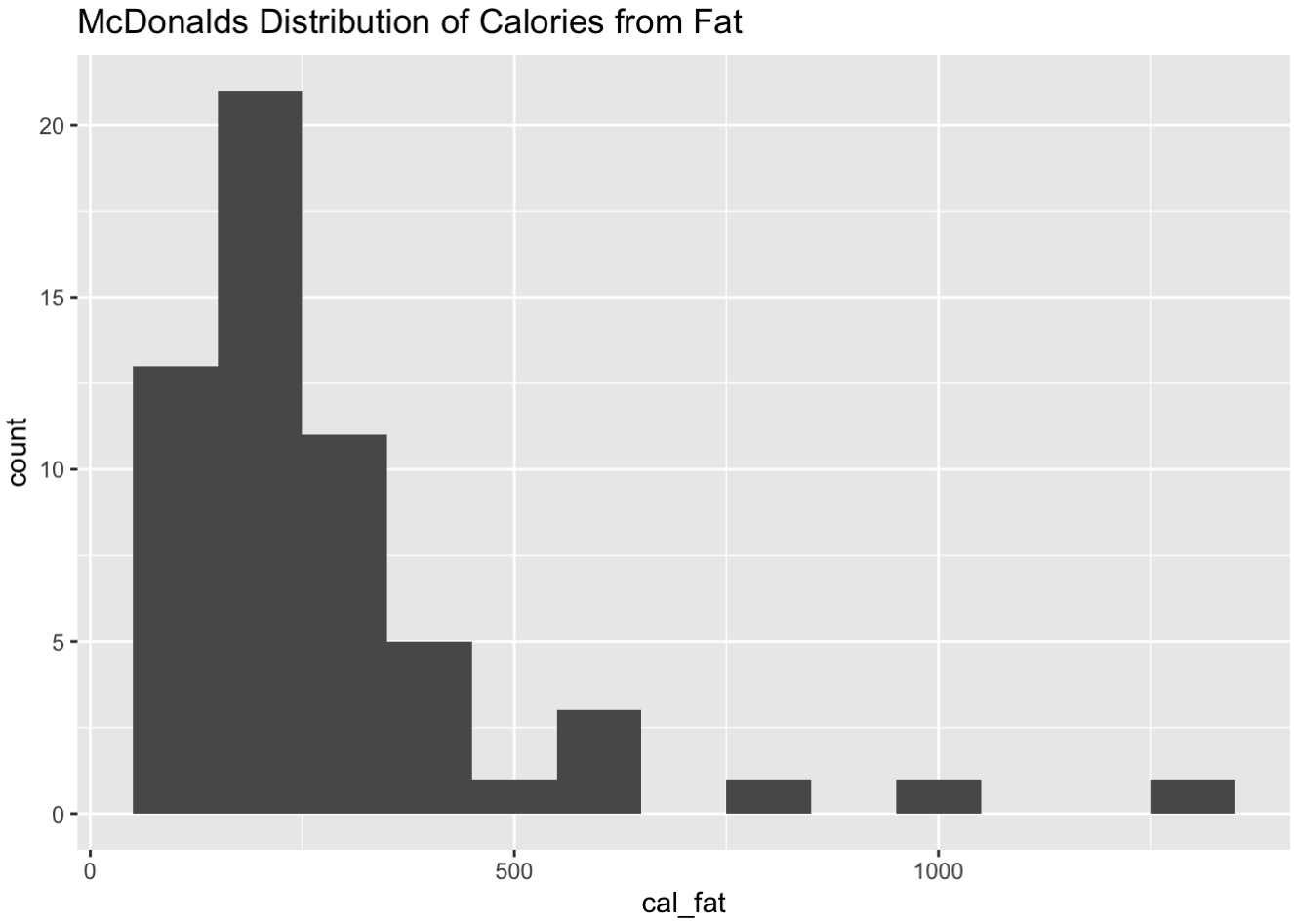
-McDonald's shows a wider range of data and seems to be centered around the 250 calorie mark. The shape of the distribution is skewed right.

-Diary Queen has a lesser range of data, but also seems more distributed within its own range compared to that of McDonald's. The data is centered around the 250-300 calorie mark. The shape of the distribution seems normal.

-An analysis of the mean and median for each data set confirms the predicted measures of center.

Hide

```
#mcdonalds histogram
ggplot(data = mcdonalds, aes(x = cal_fat)) + geom_histogram(binwidth = 100) + ggtit
        le("McDonalds Distribution of Calories from Fat")
```



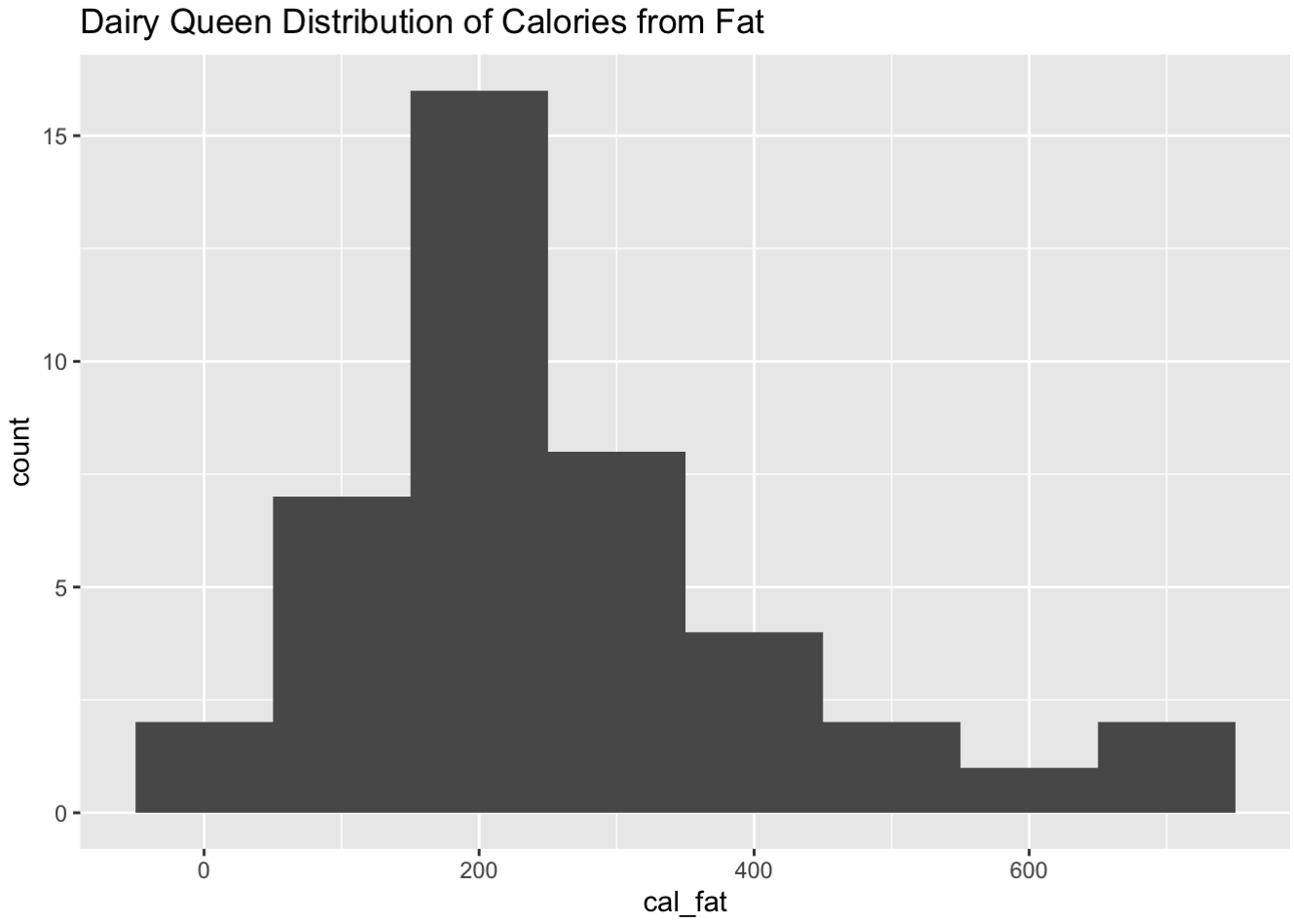
Hide

```
#summary of Mcdonald's measure of center
mcdonalds %>% summarise(mean_MD = mean(cal_fat),
median_MD = median(cal_fat),
n = n())
```

```
## # A tibble: 1 × 3
##   mean_MD median_MD     n
##     <dbl>     <dbl> <int>
## 1    286.       240    57
```

Hide

```
#dairy queen histogram
ggplot(data = dairy_queen, aes(x = cal_fat)) + geom_histogram(binwidth = 100) + ggt
        itle("Dairy Queen Distribution of Calories from Fat")
```



Hide

```
#summary of Dairy Queen's measure of center
dairy_queen %>% summarise(mean_DQ = mean(cal_fat),
median_DQ = median(cal_fat),
n = n())
```

```
## # A tibble: 1 × 3
##   mean_DQ median_DQ     n
##     <dbl>     <dbl> <int>
## 1    260.       220    42
```

Exercise 2

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

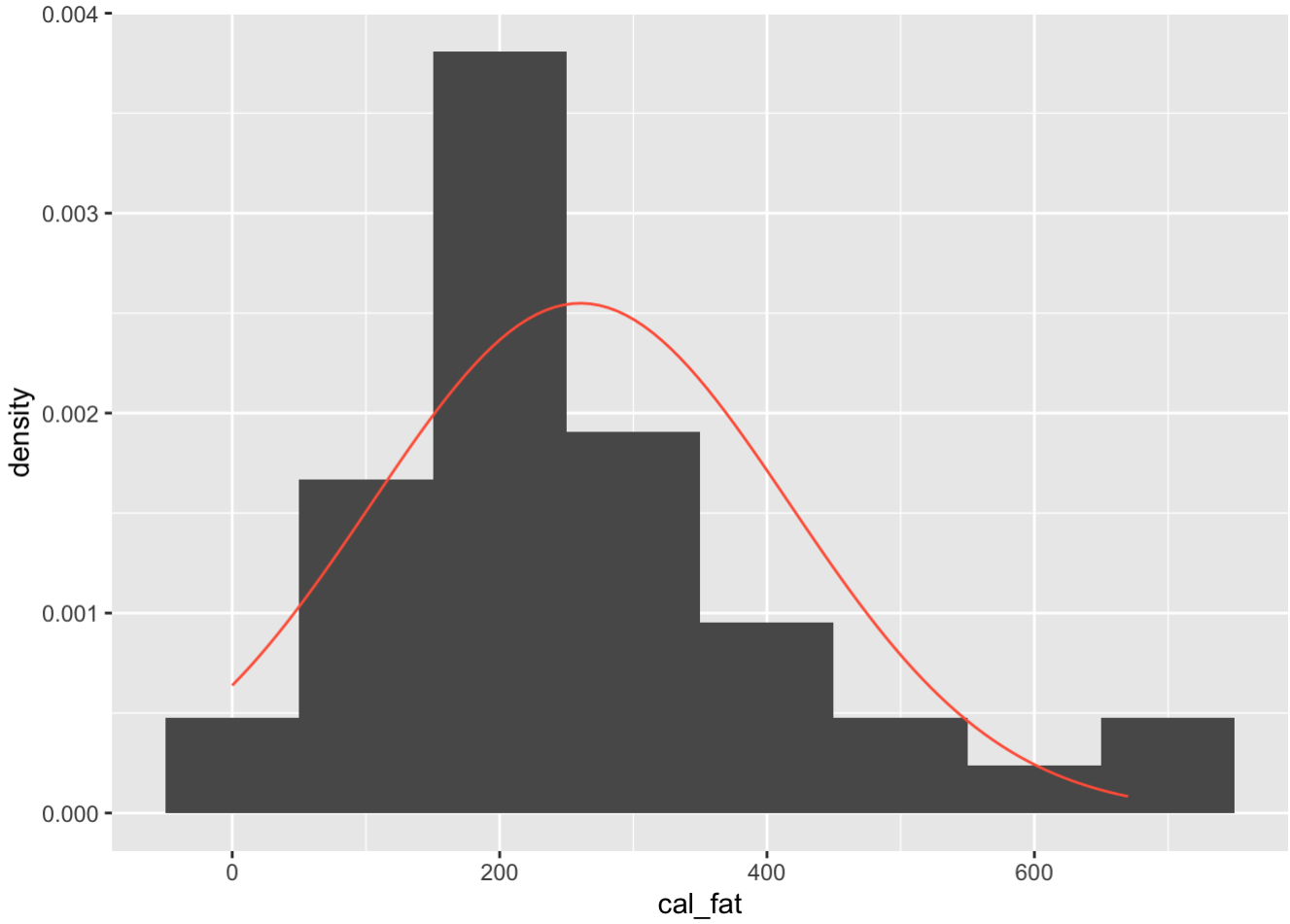
-The distribution seems nearly normal. The center of the data mostly fits the curves center.

...

Hide

```
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..), binwidth = 100) +
  stat_function(fun = dnorm, args = c(mean = dqmean, sd = dqsd), col = "tomato")
```



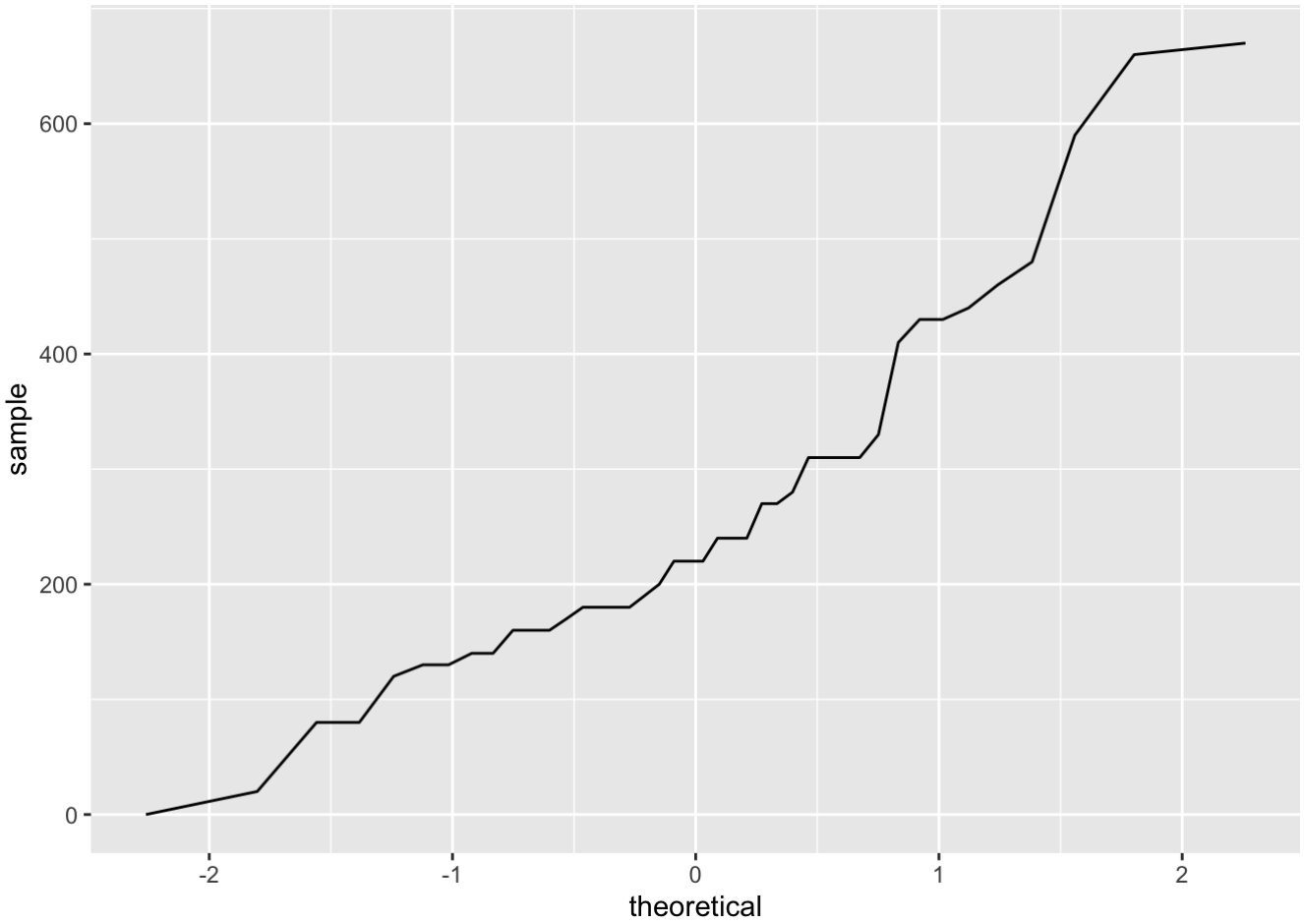
Exercise 3

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.)

-Most of the points do fall on the line, however the probability plot seems a little more jagged.

Hide

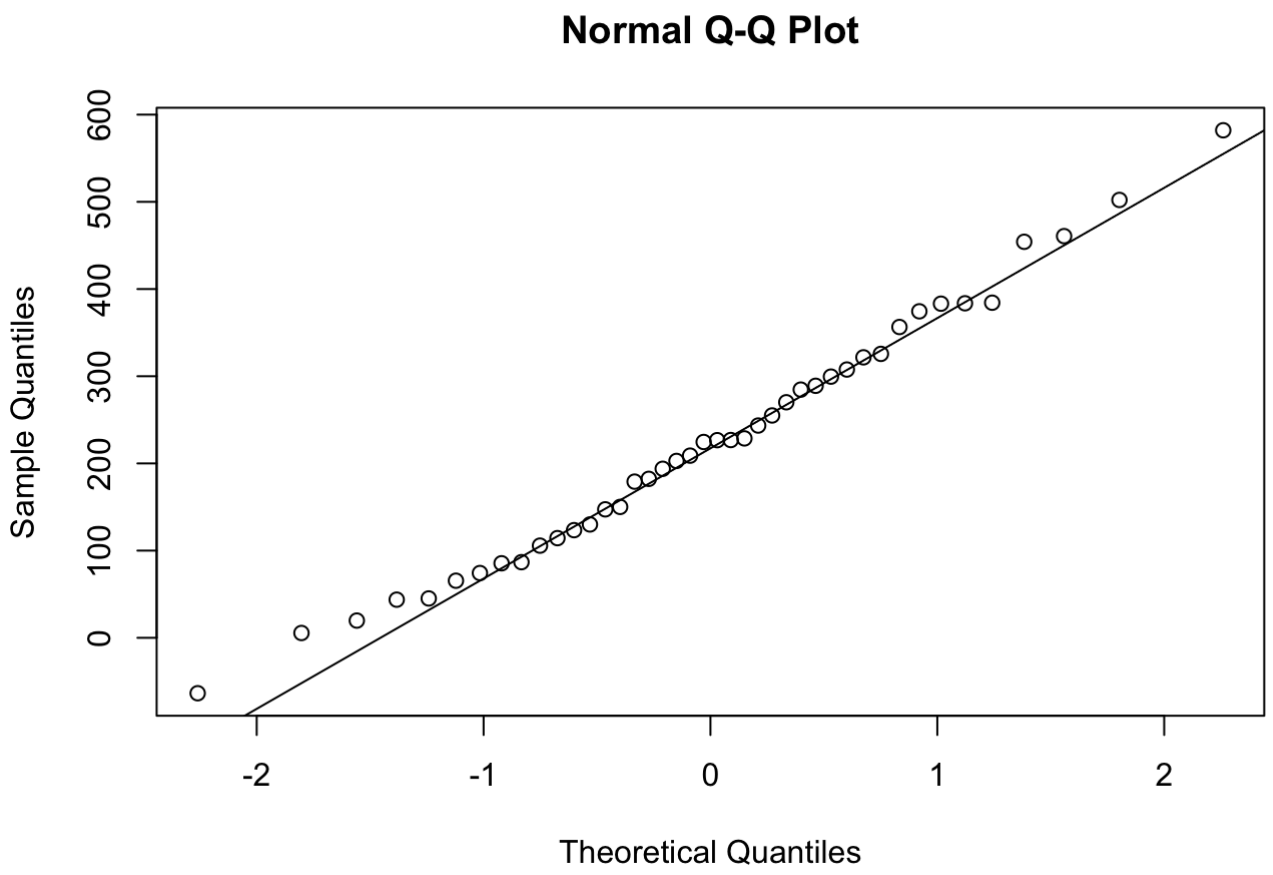
```
ggplot(data = dairy_queen, aes(sample = cal_fat)) +
  geom_line(stat = "qq")
```



Hide

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

qqnorm(sim_norm)
qqline(sim_norm)
```



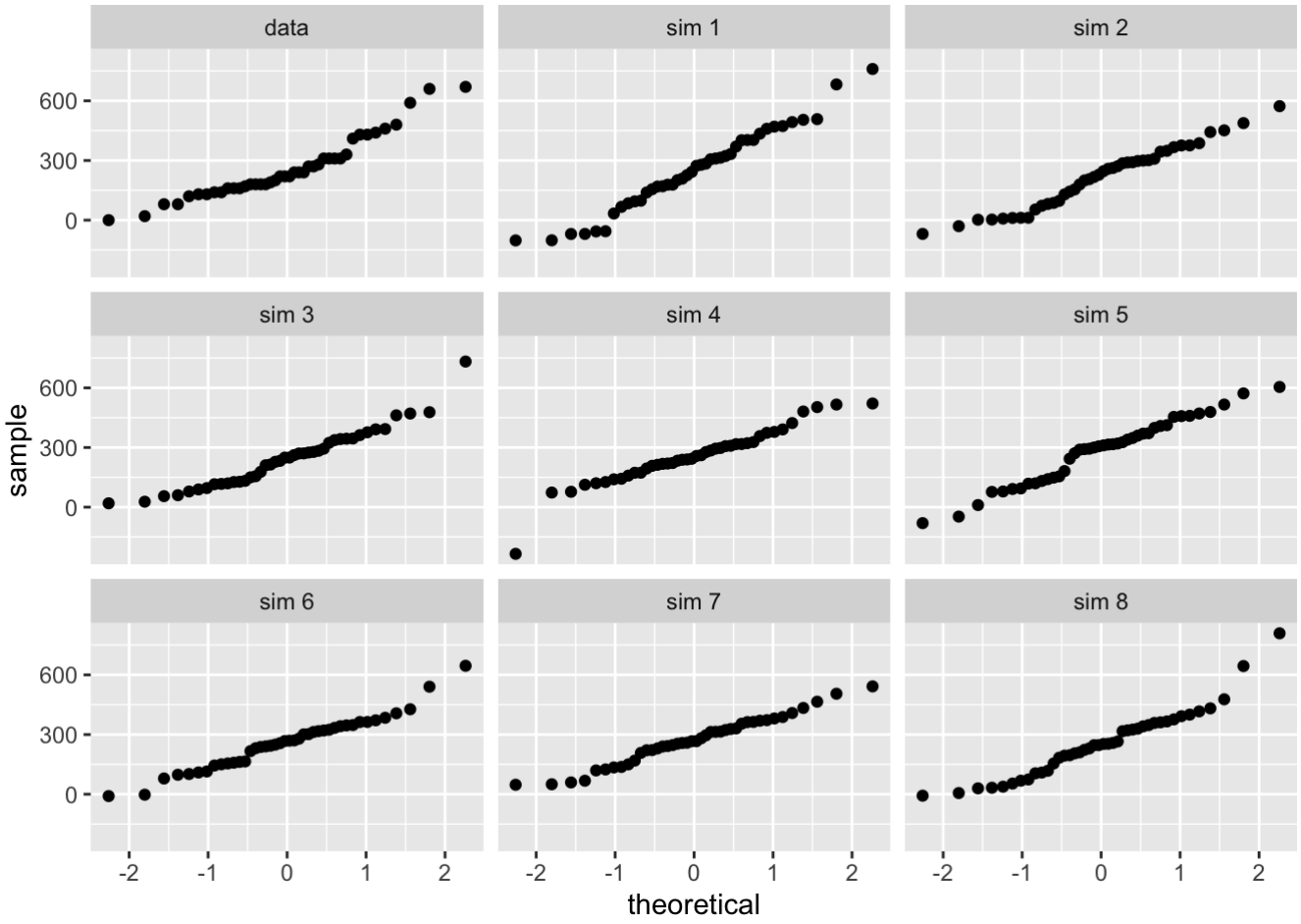
Exercise 4

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?

-Although the actual plot differs a bit from the simulated data, it doesn't differ by much. The plots seem to provide enough evidence that the calories are nearly normal.

Hide

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



Exercise 5

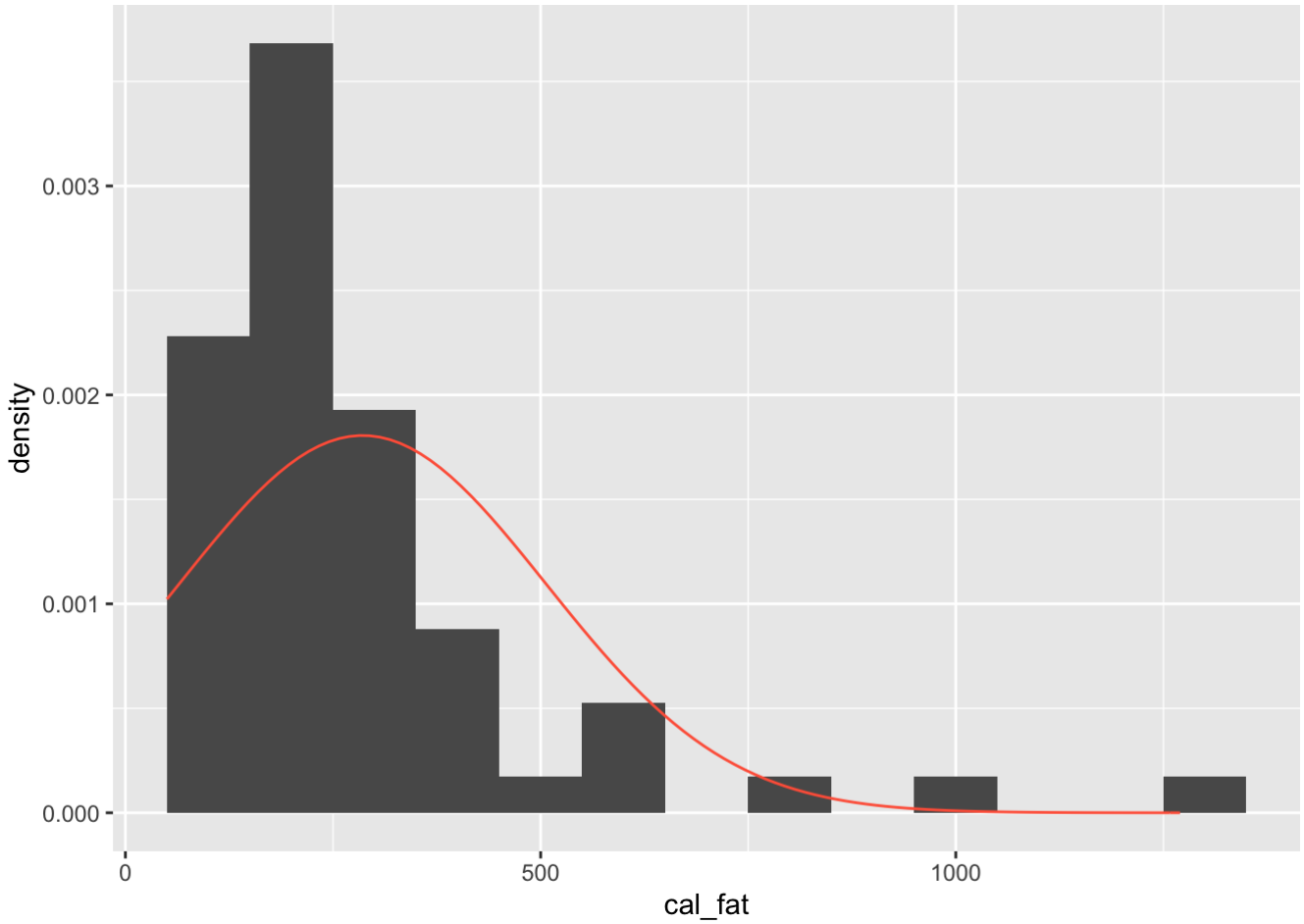
Using the same technique, determine whether or not the calories from McDonald's menu appear to come from a normal distribution.

-According to the probability plot, the calories from McDonald's menu do not seem to come from a normal distribution. The plot shows significantly more curvature than that of the simulations.

Hide

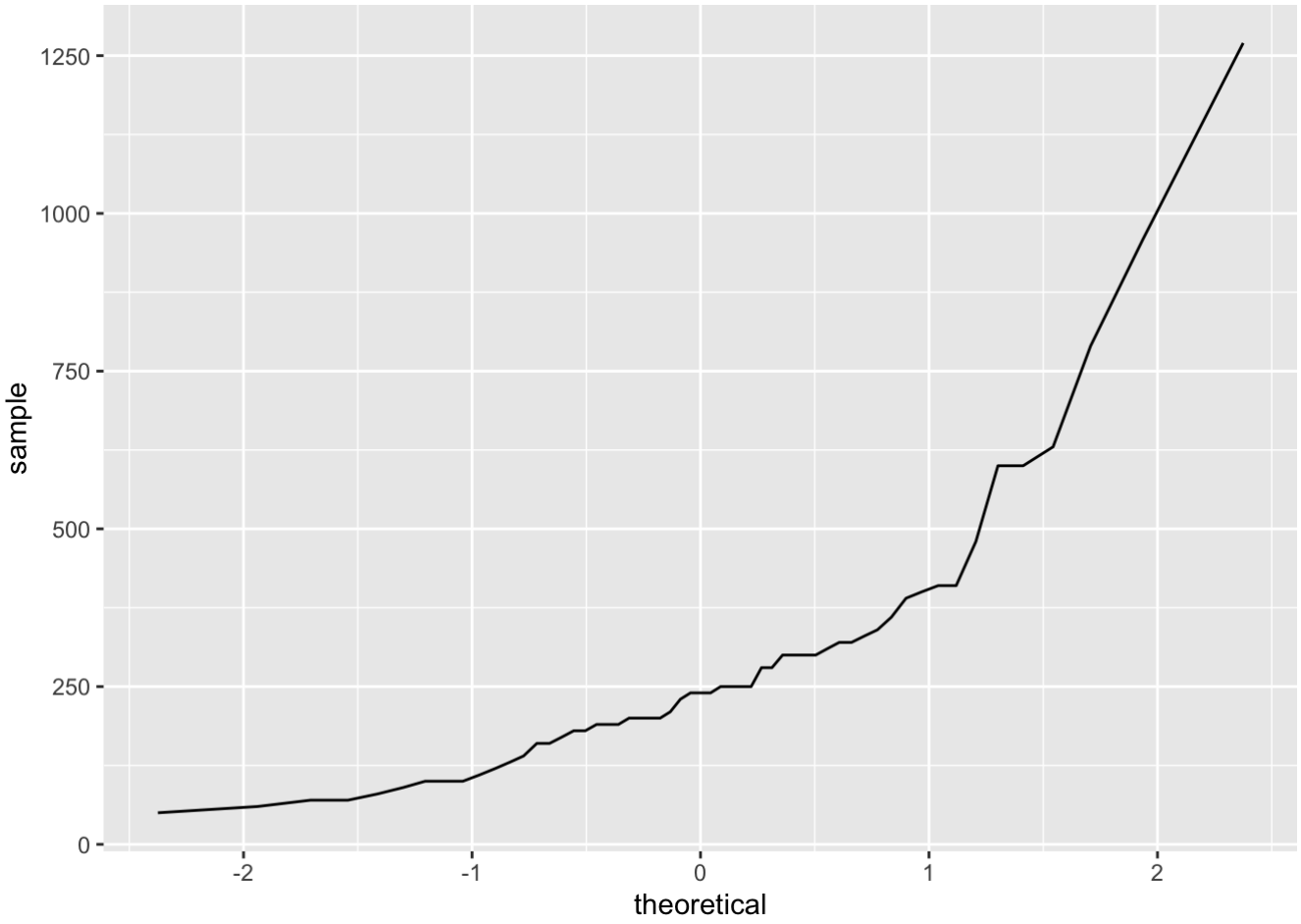
```
mDmean <- mean(mcdonalds$cal_fat)
mDsd <- sd(mcdonalds$cal_fat)

ggplot(data = mcdonalds, aes(x = cal_fat)) +
  geom_blank() +
  geom_histogram(aes(y = ..density..), binwidth = 100) +
  stat_function(fun = dnorm, args = c(mean = mDmean, sd = mDsd), col = "tomato")
```



Hide

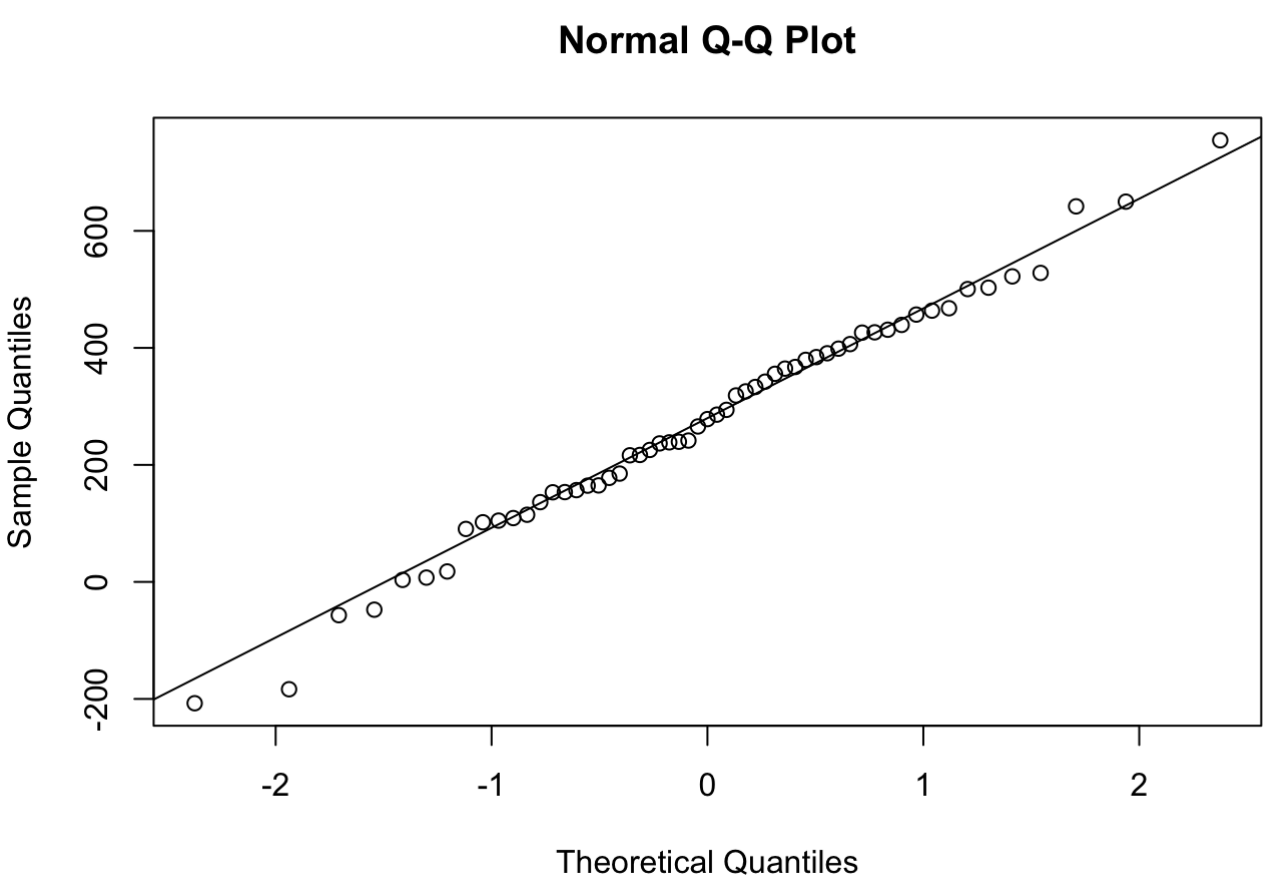
```
ggplot(data = mcdonalds, aes(sample = cal_fat)) +
  geom_line(stat = "qq")
```



Hide

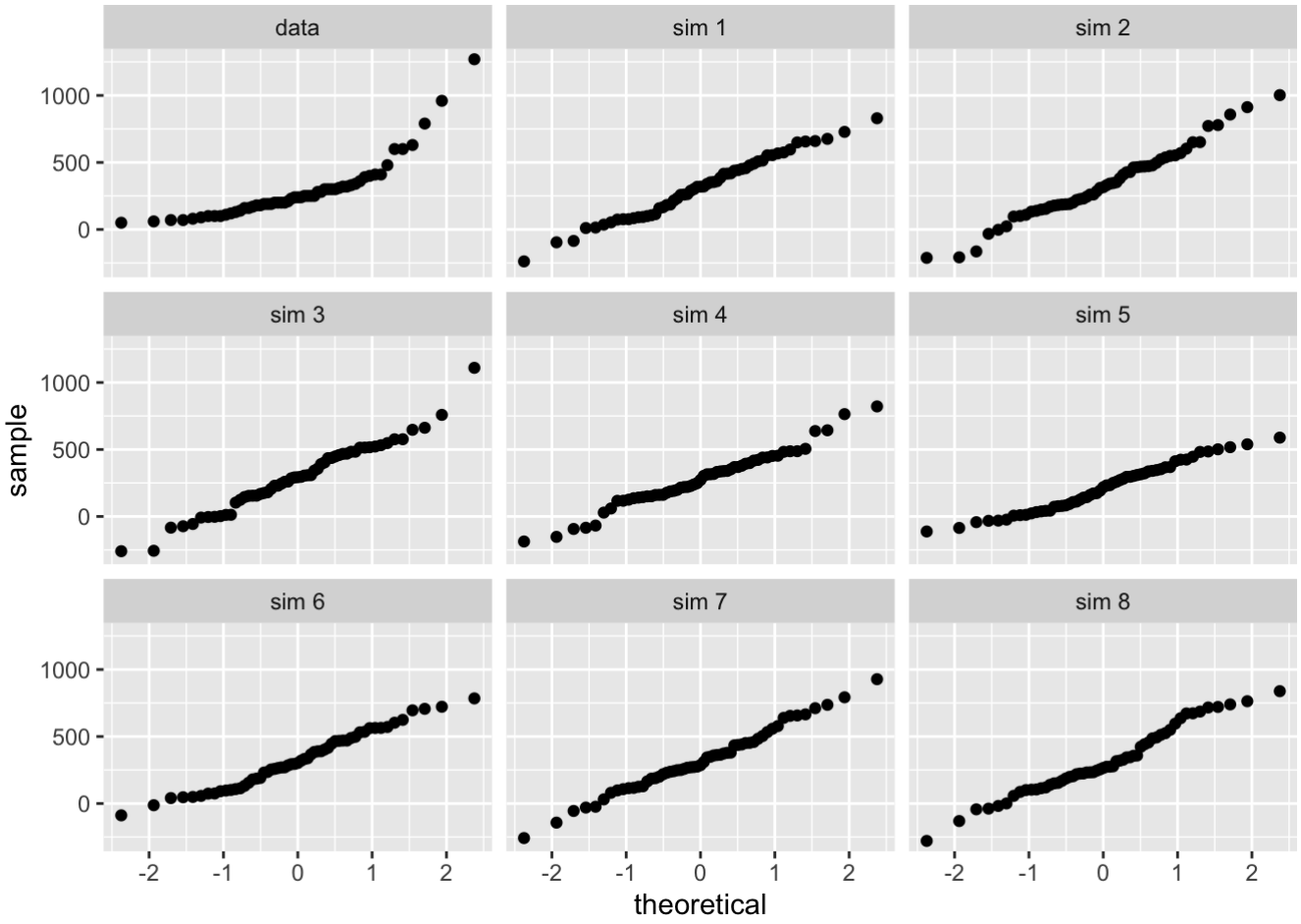
```
sim_norm1 <- rnorm(n = nrow(mcdonalds), mean = mDmean, sd = mDsd)

qqnorm(sim_norm1)
qqline(sim_norm1)
```



Hide

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



Exercise 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 product has less than 117 calories from fat? Answer #1: The probability that a randomly selected dairy product has less than 117 calories from fat is 0.1796.

Question #2: What is the probability that a randomly chosen dairy product has between 114 and 227 calories from fat? Answer #2: The probability that a randomly selected dairy product has between 114 and 227 calories from fat is .2407.

Hide

```
#Question 1
#Theoretical
pnorm(q = 117, mean = dqmean, sd = dqsd)
```

```
## [1] 0.1796058
```

Hide

```
#Empirical
dairy_queen %>%
  filter(cal_fat < 117) %>%
  summarise(percent = n() / nrow(dairy_queen))
```

```
## # A tibble: 1 × 1
##   percent
##     <dbl>
## 1  0.0952
```

Hide

```
#Question 2
#Theoretical
pnorm(q = 227, mean = dqmean, sd = dqsd) -
  pnorm(q = 114, mean = dqmean, sd = dqsd)
```

```
## [1] 0.240676
```

Hide

```
#Empirical
dairy_queen %>%
  filter(114 < cal_fat & cal_fat < 227) %>%
  summarise(percent = n() / nrow(dairy_queen))
```

```
## # A tibble: 1 × 1
##   percent
##     <dbl>
## 1   0.429
```

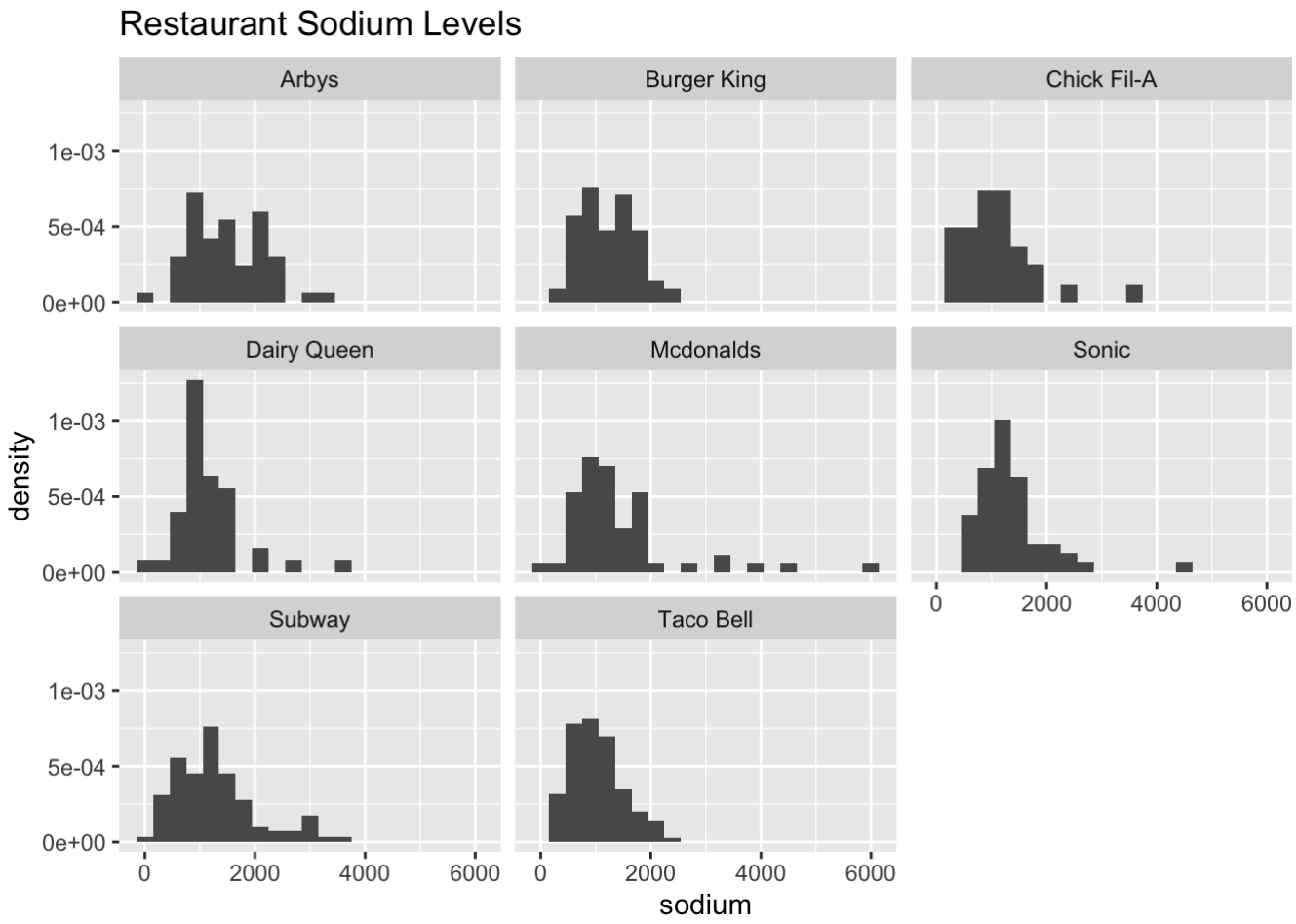
Exercise 7

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?

-The restaurant with the closest distribution to normal seems to be Burger King. (Or Arby's)

Hide

```
fastfood %>%
  group_by(restaurant) %>%
  ggplot() +
  geom_blank() +
  geom_histogram(aes(x = sodium, y = ..density..), binwidth = 300) +
  ggtitle("Restaurant Sodium Levels") +
  #stat_function(fun = dnorm, args = c(mean = bkmean, sd = dqsd), col = "tomato") +
  facet_wrap(. ~restaurant)
```



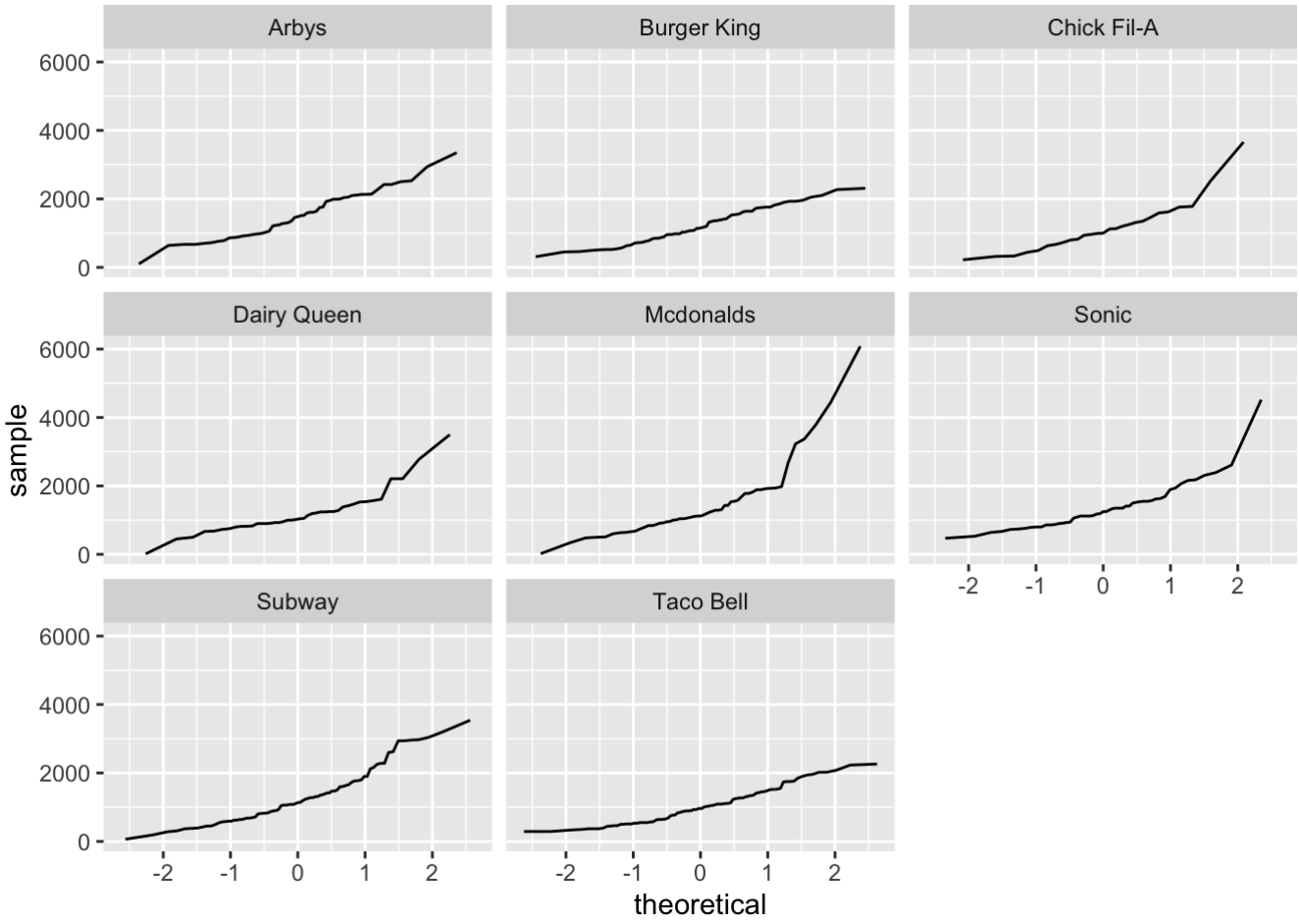
Exercise 8

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?

-A possible reason for this is that the data is discrete and many values are repeated so it's reflected in a stepwise fashion.

Hide

```
fastfood %>%
group_by(restaurant) %>%
ggplot(aes(sample = sodium)) +
  geom_line(stat = "qq") +
  facet_wrap(.~restaurant)
```



Exercise 9

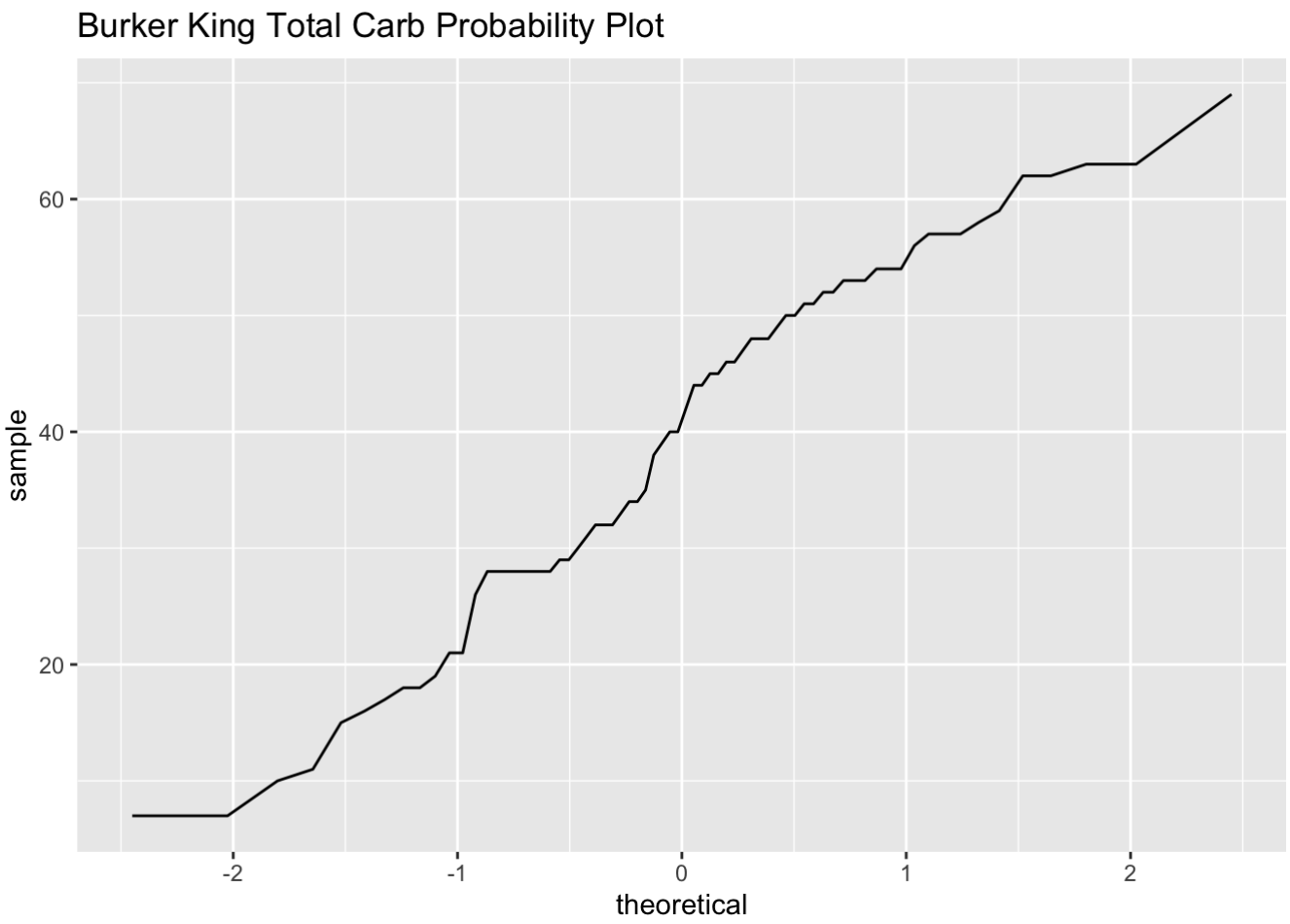
As you can see, normal probability plots can be used both to assess normality and visualize skewness. Make a normal probability plot for the total carbohydrates from a restaurant of your choice. Based on this normal probability plot, is this variable left skewed, symmetric, or right skewed? Use a histogram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

-Based on the probability plot, the data seems fairly normal. The plot of the histogram with the normal curve confirms a fairly normal distribution.

Hide

```
bK <- fastfood %>%
  filter(restaurant == "Burger King")

ggplot(data = bK, aes(sample = total_carb)) +
  geom_line(stat = "qq") +
  ggtitle("Burker King Total Carb Probability Plot")
```



Hide

```
bkmean <- mean(bK$total_carb)
bksd <- sd(bK$total_fat)

bkmean
```

```
## [1] 39.31429
```

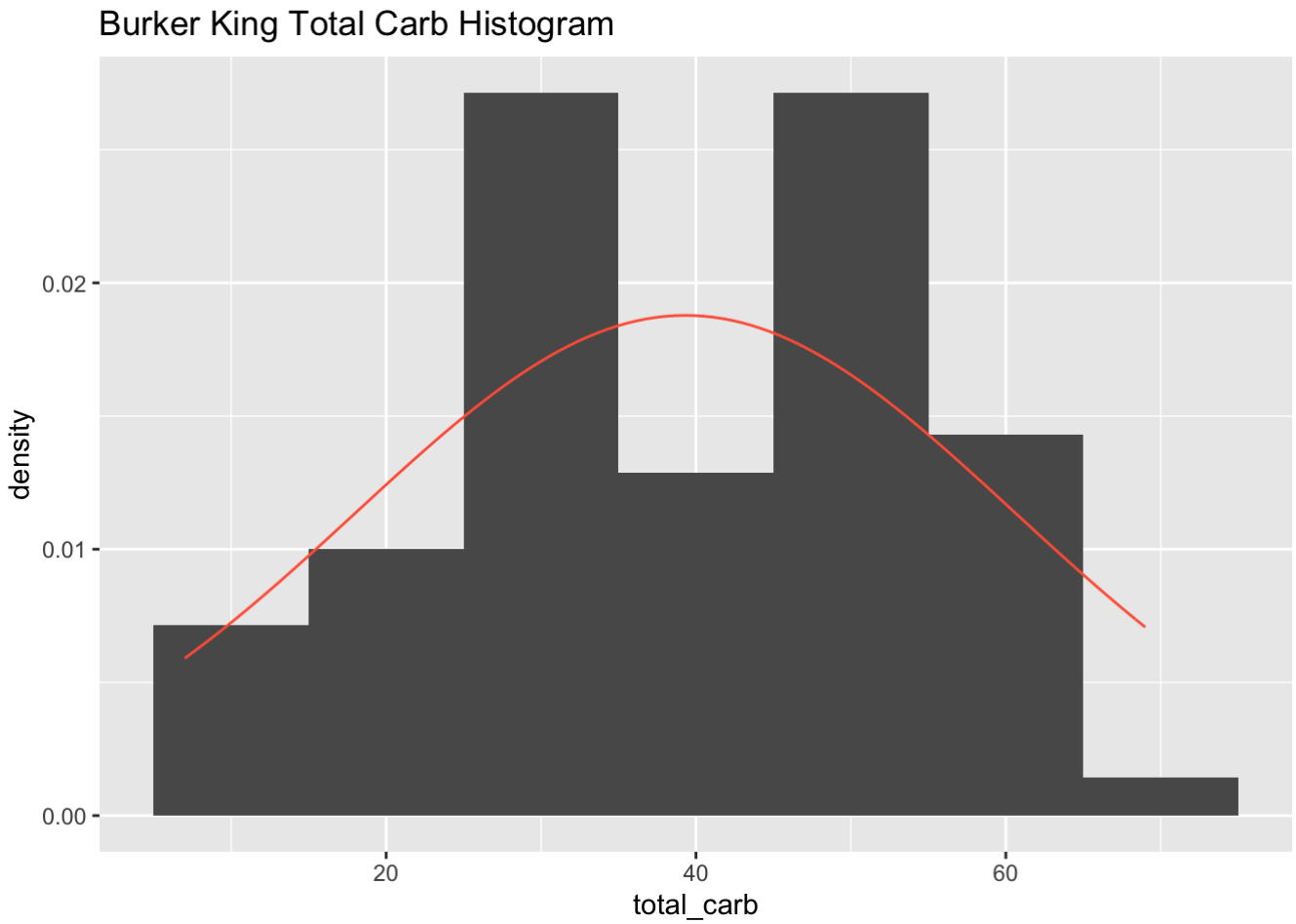
Hide

```
bksd
```

```
## [1] 21.24344
```

Hide

```
ggplot(data = bK, aes(x = total_carb)) +
  geom_blank() +
  geom_histogram(aes(y = ..density..), binwidth = 10) +
  ggtitle("Burker King Total Carb Histogram") +
  stat_function(fun = dnorm, args = c(mean = bkmean, sd = bksd), col = "tomato")
```


Burker King Total Carb Histogram
density
0.02
0.01
0.00
20
40
60
total_carb