Lab4 - Normal Distribution

Keith Colella

2023-02-28

Setup

```
library(tidyverse)
library(openintro)
library(cowplot)
library(EnvStats)
library(nortest)
set.seed(101112)
```

This data set contains data on 515 menu items from some of the most popular fast food restaurants worldwide.

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

```
## # A tibble: 6 × 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 ¹restaurant, ²calories, ³total_fat, ⁴trans_fat,
## #   ⁵cholesterol, ⁶total_carb
```

```
mcdonalds <- fastfood %>%
  filter(restaurant == 'Mcdonalds')
dairy_queen <- fastfood %>%
  filter(restaurant == 'Dairy Queen')
```

Exercise 1

Question

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?

Response

```
cat('No. observations for Mcdonalds:',nrow(mcdonalds),
    '\nNo. observations for Diary Queen:',nrow(dairy_queen),'\n')
```

```
## No. observations for Mcdonalds: 57
## No. observations for Diary Queen: 42
```

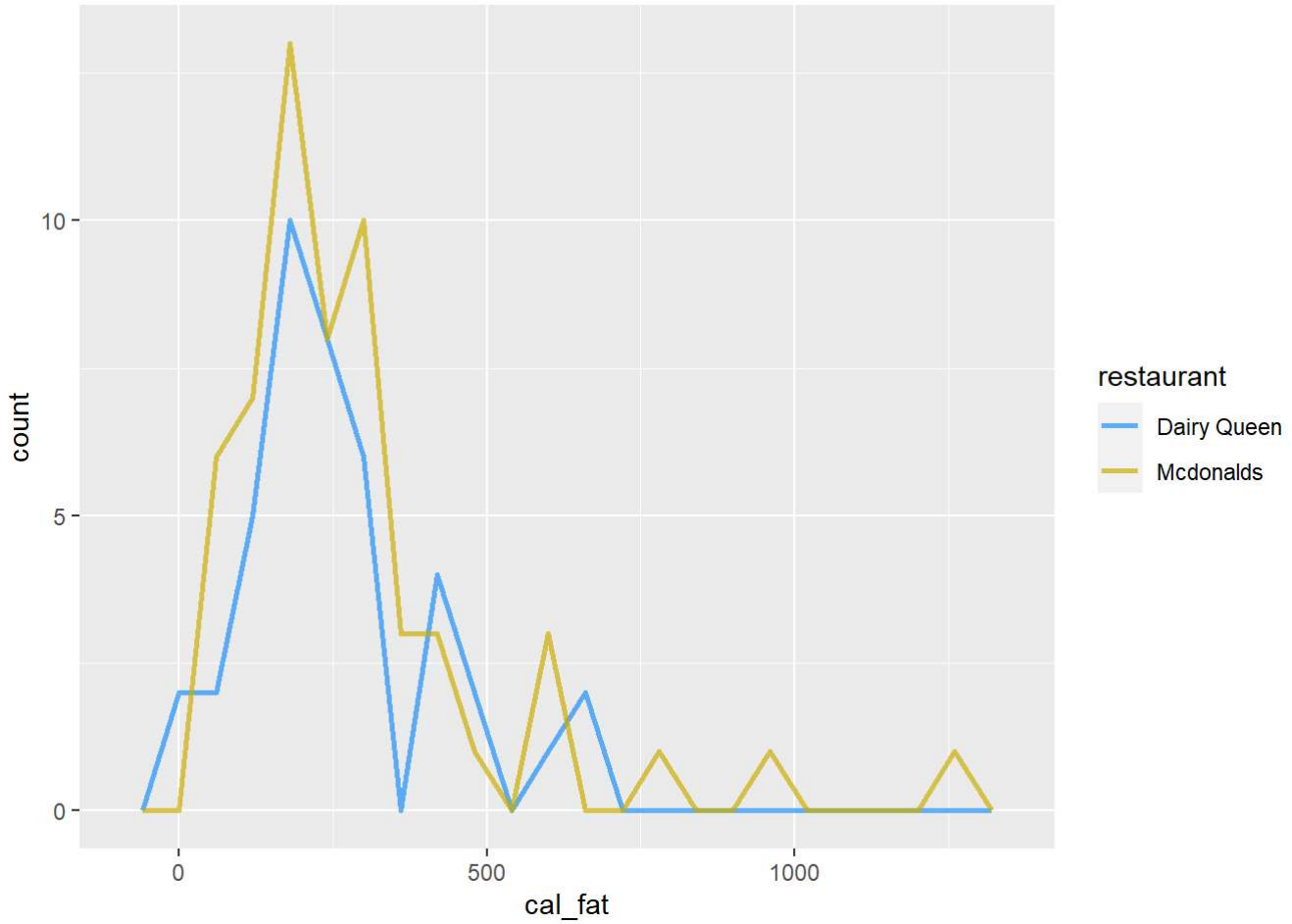
```
fastfood %>%
  filter(restaurant == 'Mcdonalds' | restaurant == 'Dairy Queen') %>%
  select(cal_fat) %>%
  summary()
```

```
##     cal_fat
##  Min.   :   0.0
##  1st Qu.: 160.0
##  Median : 240.0
##  Mean   : 274.9
##  3rd Qu.: 320.0
##  Max.   :1270.0
```

The number of observations in for each restaurant is relatively low (~50 each), and calories from fat covers a wide range of values, from 0 to 1270. As a result, the distributions can appear “bumpy” if the binwidth is set too low.

So, I’ve included two plots, using the freqpoly geom, which allows us to view both restaurants together. The first plot uses wide bins to reduce “noise” and give a broad sense of the distribution. From this first plot, the center of each distribution appears roughly aligned, centering around ~200 calories.

```
fastfood %>%
  filter(restaurant == 'Mcdonalds' | restaurant == 'Dairy Queen') %>%
  ggplot(aes(x = cal_fat, color = restaurant)) +
  geom_freqpoly(binwidth = 60, linewidth = 1, alpha = 0.7) +
  scale_color_manual(values=c('dodgerblue','gold3'))
```



We can check this against the actual mean for each distribution.

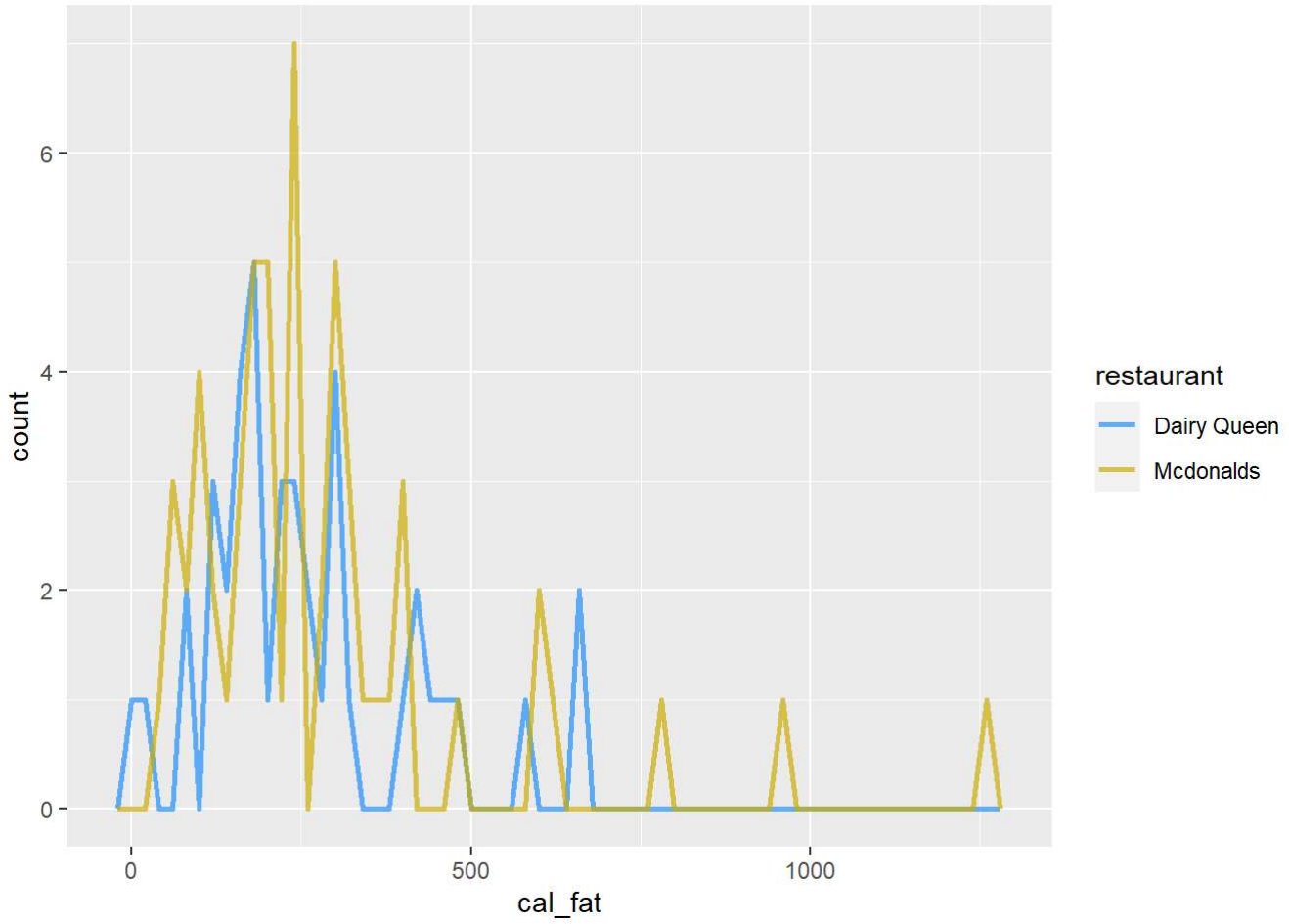
```
cat('Mean calories from fat for Mcdonalds:',
    mean(mcdonalds$cal_fat),
    '\nMean calories from fat for Dairy Queen:',
    mean(dairy_queen$cal_fat))
```

```
## Mean calories from fat for Mcdonalds: 285.614
## Mean calories from fat for Dairy Queen: 260.4762
```

This calculation reveals that the first plot is a bit misleading. In using wide bins to “smooth” the distribution, we lose some sense of the finer details.

This brings us to the second plot, which uses much smaller bins.

```
fastfood %>%
  filter(restaurant == 'Mcdonalds' | restaurant == 'Dairy Queen') %>%
  ggplot(aes(x = cal_fat, color = restaurant)) +
  geom_freqpoly(binwidth = 20, linewidth = 1, alpha = 0.7)+
  scale_color_manual(values=c('dodgerblue','gold3'))
```



This plot gives a better sense of centrality, showing central tendency closer to ~250 calories, but it's quite difficult to read, given the "bumpy", multi-modal appearance.

In both plots, however, it's clear that the Mcdonalds distribution has some significant outliers.

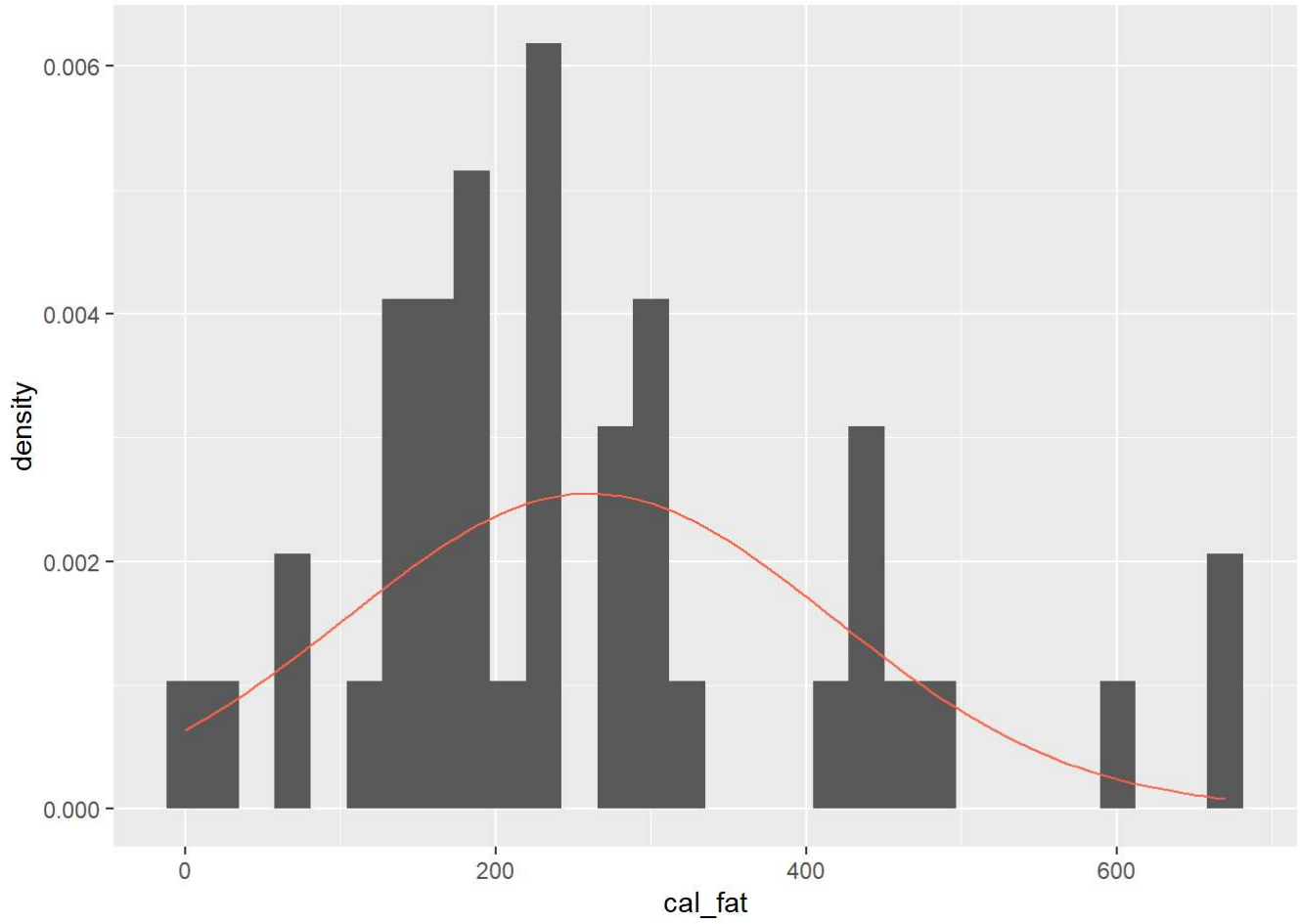
Exercise 2

Question

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 = after_stat(density)), bins = 30) +
        stat_function(fun = dnorm, args = c(mean = dqmean, sd = dqsd), col = "tomato")
```



Response

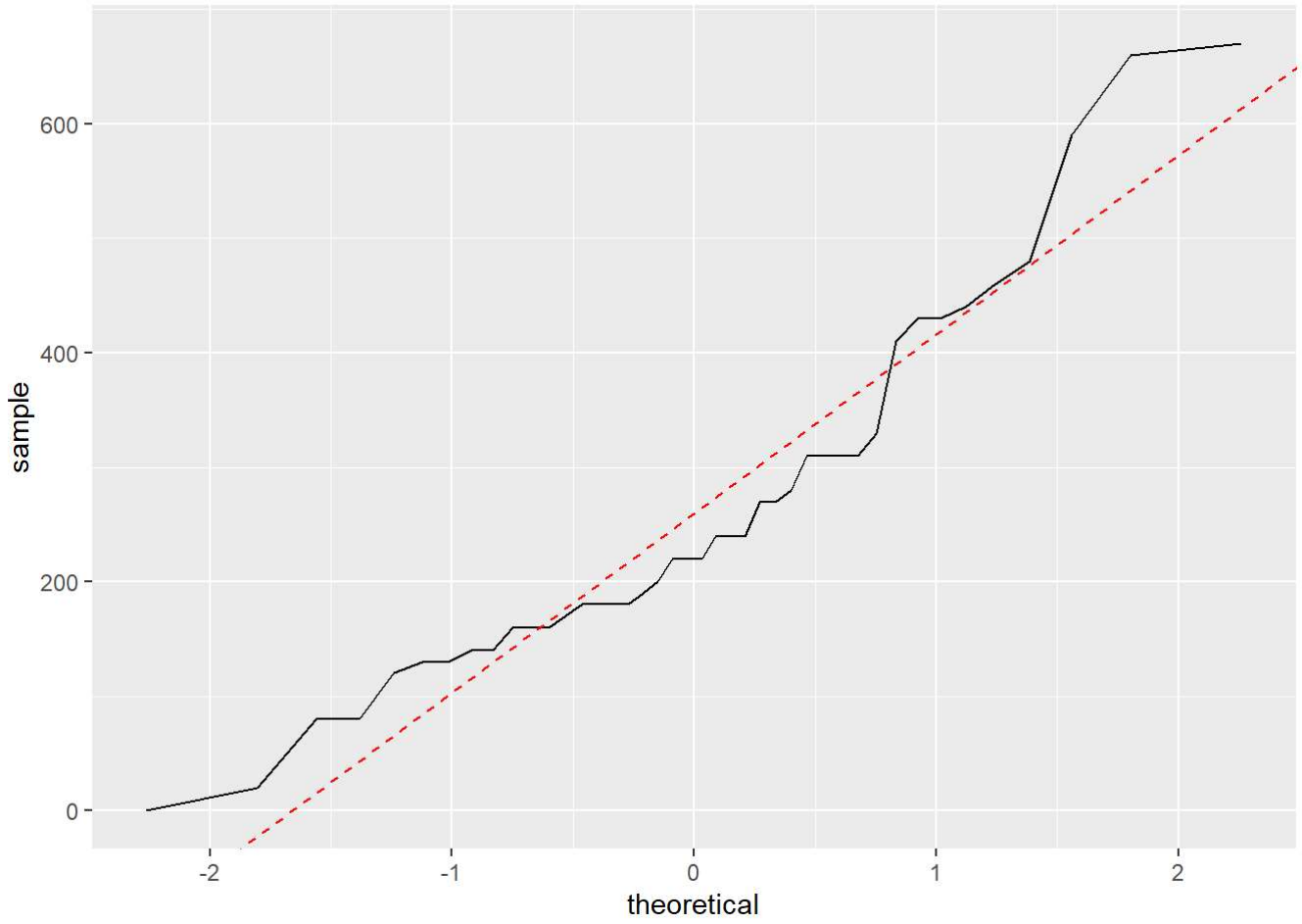
I would not call this distribution normal. There are a significant number of observations in the far right tail, and considerably more spread in general than would be expected from an approximately normal distribution.

Exercise 3

Question

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?

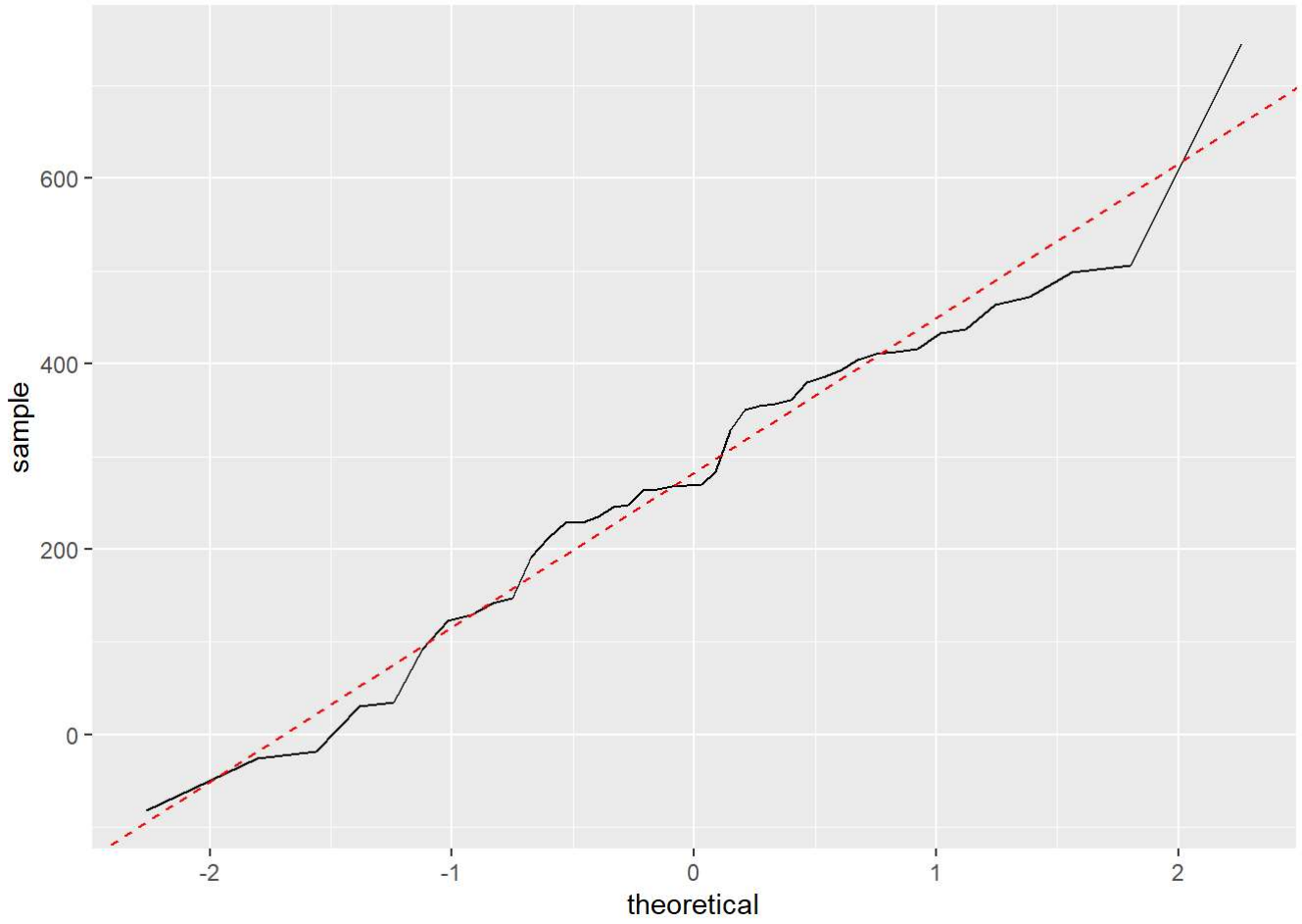
```
ggplot(data = dairy_queen) +
  geom_line(aes(sample = cal_fat), stat = 'qq') +
  geom_abline(intercept = dqmean, slope = dqsd,
              color = 'red', linetype = 'dashed')
```



Response

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

ggplot() +
  geom_line(aes(sample = sim_norm), stat = 'qq') +
  geom_abline(intercept = mean(sim_norm), slope = sd(sim_norm),
              color = 'red', linetype = 'dashed')
```

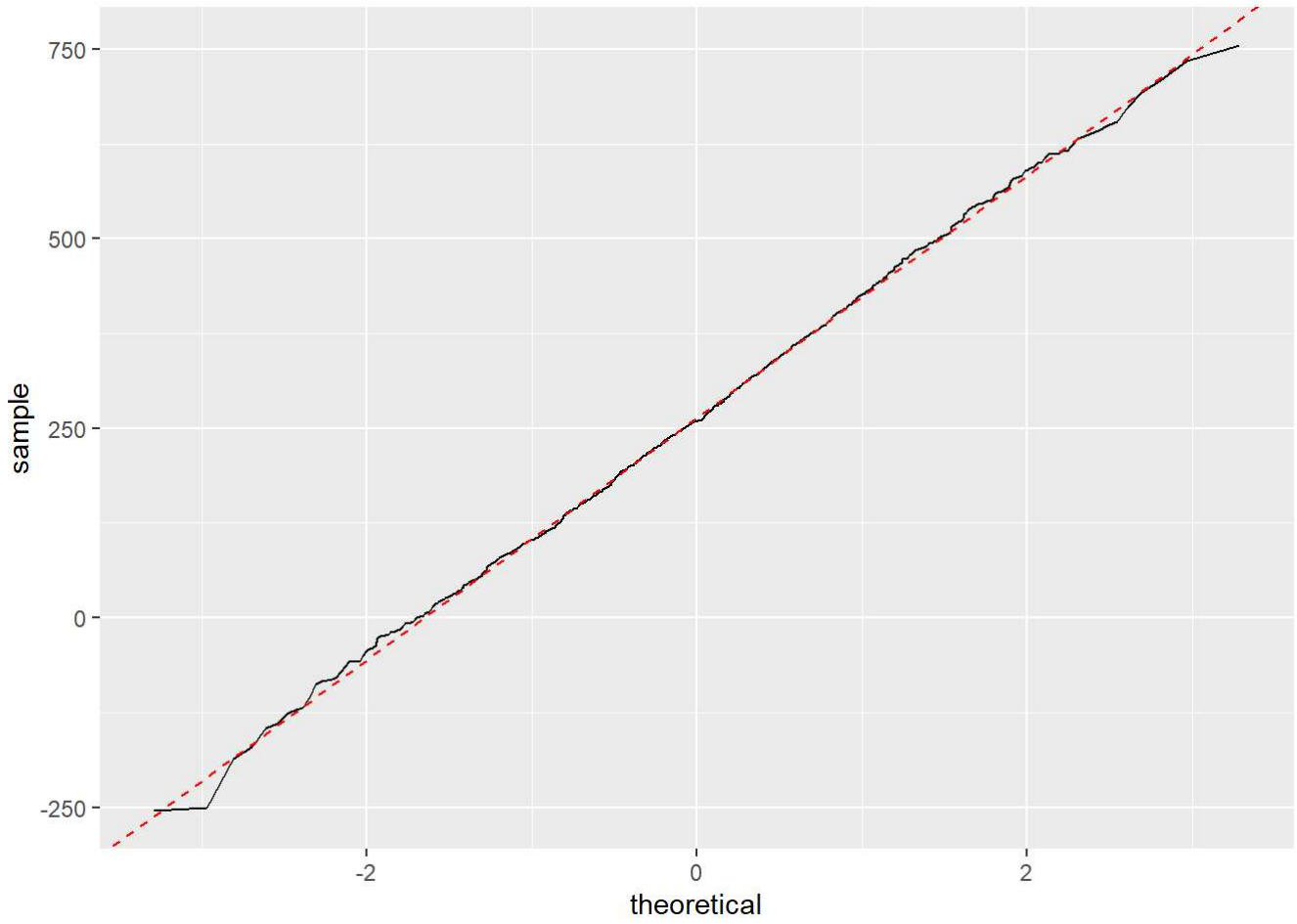


The simulated normal Q-Q plot fit to the theoretical diagonal better than the observed data. However, significant differences appear, especially in the right tail. This deviation is presumably because of the relatively small sample size.

If we were to increase the sample size of our simulation, we should expect a greater fit, as shown in the plot below, created with 1000 simualed samples.

```
sim_norm2 <- rnorm(n = 1000, mean = dqmean, sd = dqsd)

ggplot() +
  geom_line(aes(sample = sim_norm2), stat = 'qq') +
  geom_abline(intercept = mean(sim_norm2), slope = sd(sim_norm2),
              color = 'red', linetype = 'dashed')
```

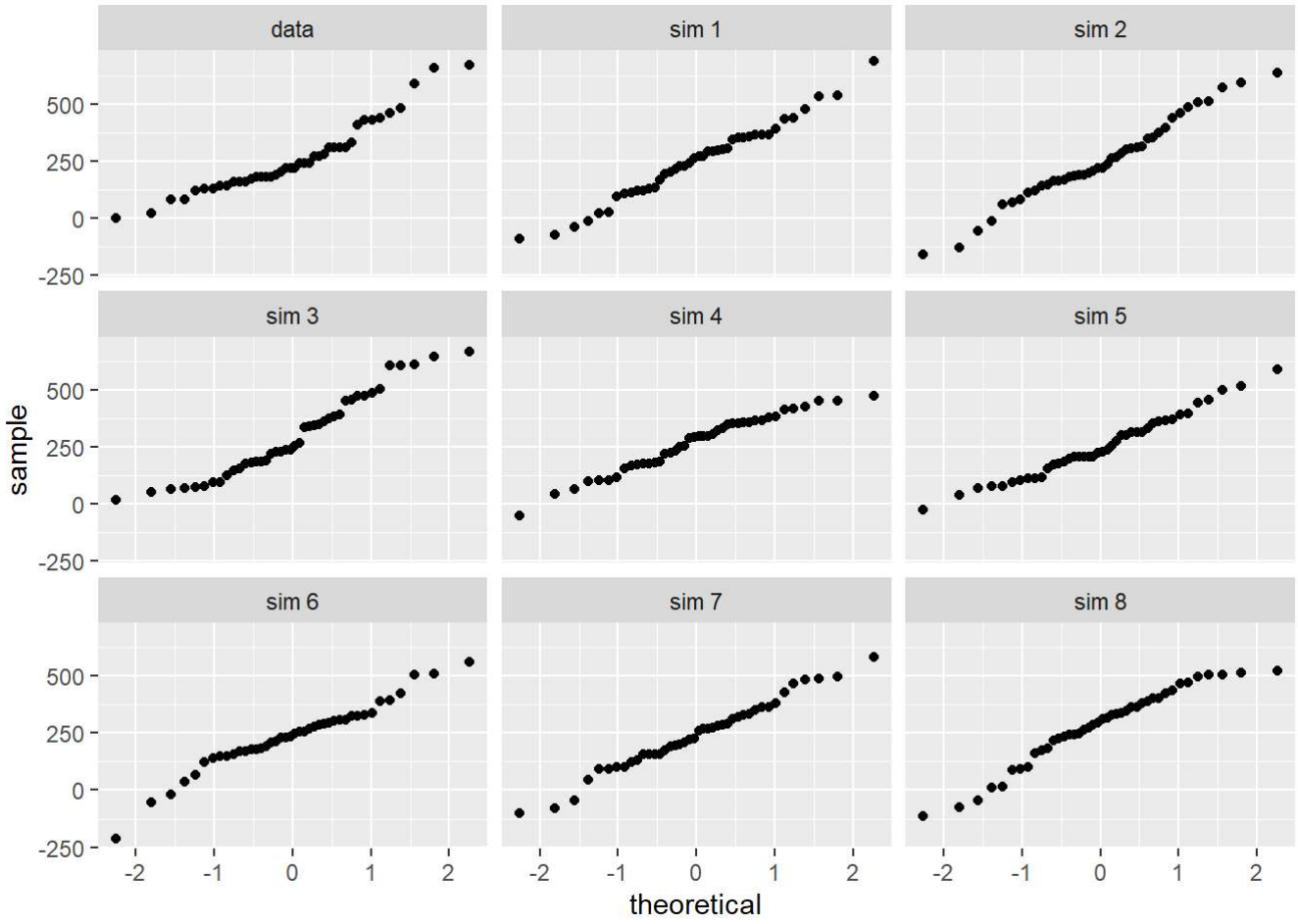


Exercise 4

Question

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)
```



Response

All of these simulations appear slightly more closely aligned to the diagonal than our observed data. As noted in response #3, however, we do see some notable deviation from the diagonal due to the small sample size. So, while the observed data does seem less aligned to a normal distribution, it is difficult to make definitive conclusions based on the limited sample available.

Exercise 5

Question

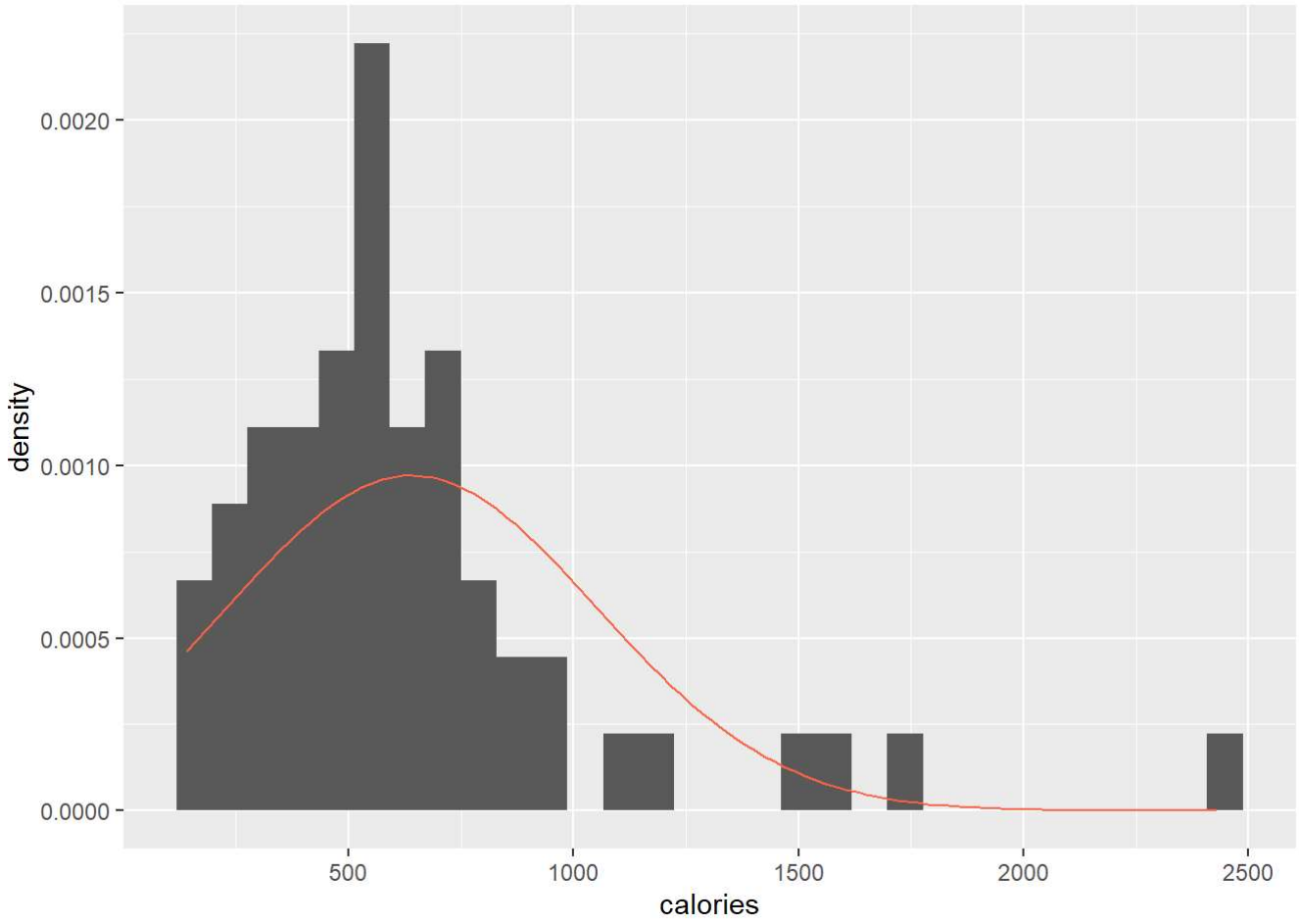
Using the same technique, determine whether or not the calories from McDonald's menu appear to come from a normal distribution.

Response

The question mentions only calories, rather than calories from fat, so we'll focus on that variable for this response.

```
mdmean <- mean(mcdonalds$calories)
mdsd <- sd(mcdonalds$calories)

ggplot(mcdonalds, aes(x = calories)) +
  geom_blank() +
  geom_histogram(aes(y = after_stat(density)), bins = 30) +
  stat_function(fun = dnorm, args = c(mean = mdmean, sd = mdsd), color = 'tomato')
```



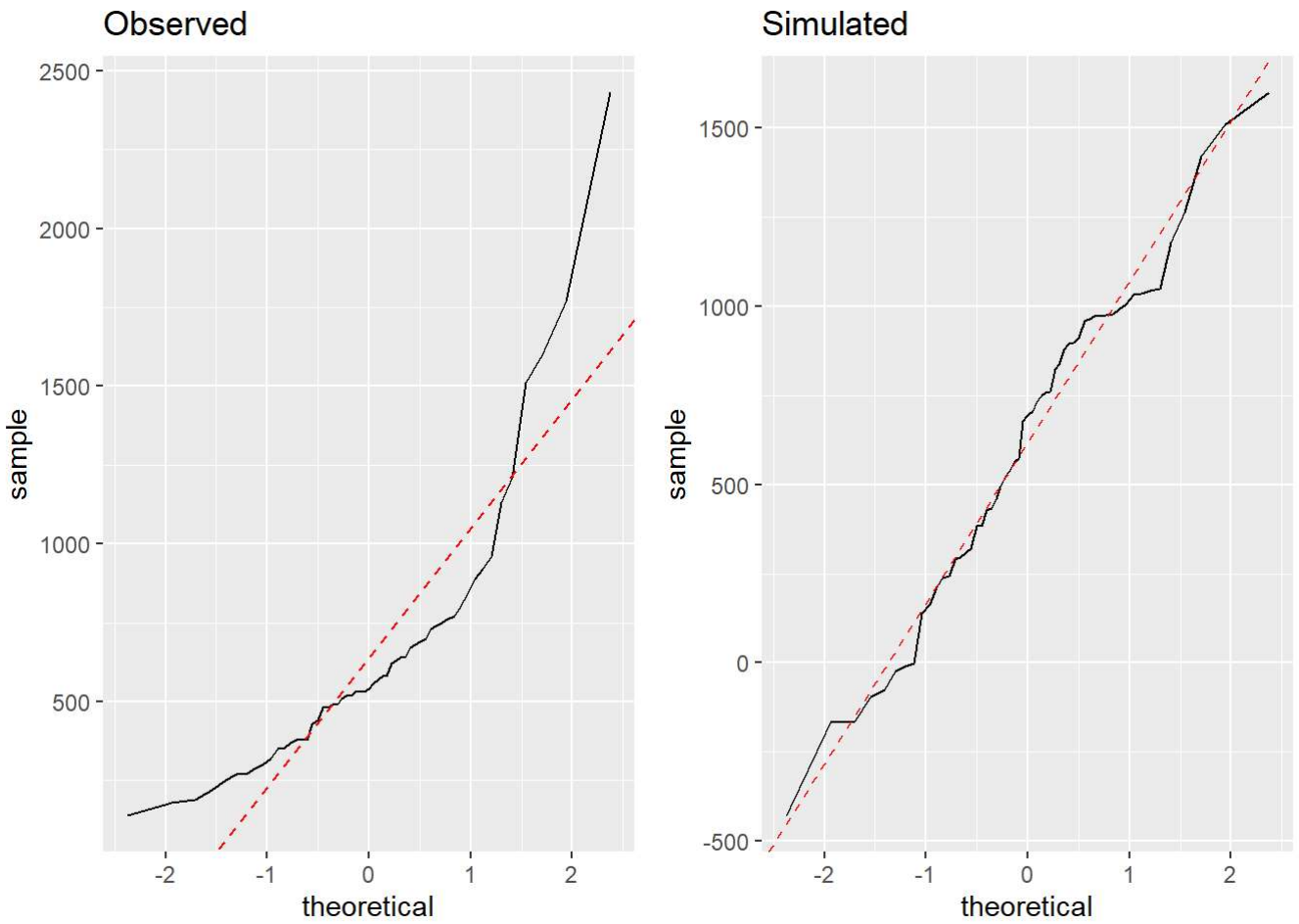
Our initial plot compares the histogram of calories for McDonald's items to a normal curve with the same mean and standard deviation. We see significant divergence from the normal curve, especially in the left and right tails. It seems from this plot that the distribution of calories does not follow a normal, but we'll check the Q-Q plot to confirm.

```
p1 <- ggplot(mcdonalds) +
  geom_line(aes(sample = calories), stat = 'qq') +
  geom_abline(intercept = mdmean, slope = mdsd,
              color = 'red', linetype = 'dashed') +
  ggtitle('Observed')

sim_norm3 <- rnorm(n = nrow(mcdonalds), mean = mdmean, sd = mdsd)

p2 <- ggplot() +
  geom_line(aes(sample = sim_norm3), stat = 'qq') +
  geom_abline(intercept = mean(sim_norm3), slope = sd(sim_norm3),
              color = 'red', linetype = 'dashed') +
  ggtitle('Simulated')

plot_grid(p1, p2)
```



Comparing Q-Q plots for observed and simulated data, it becomes clear that the calories of Mcdonald's items deviate significantly from a normal distribution. As seen in the plot above, both tails deviate quite a bit. Overall, it seems quite clear that a normal distribution is not a good fit for the data.

Exercise 6

Question

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?

Response

Question 1: What is the probability of choosing an item from Taco Bell with more than 800 calories?

```
tbell <- filter(fastfood, restaurant == 'Taco Bell')

normal <- 1 - pnorm(q = 800,
                    mean = mean(tbell$calories),
                    sd = sd(tbell$calories))

empirical <- tbell %>%
  filter(calories > 800) %>%
  summarize(percentile = n() / nrow(tbell))

cat('Probability from normal distr.:',normal,
    '\nProbability from empirical distr.:',empirical[[1]])
```

```
## Probability from normal distr.: 0.02661504
## Probability from empirical distr.: 0.02608696
```

Question 2: What is the probability of choosing an item from Taco Bell with less than 30g of protein?

```
normal <- pnorm(q = 30,
                mean = mean(tbell$protein),
                sd = sd(tbell$protein))

empirical <- tbell %>%
  filter(protein < 30) %>%
  summarize(percentile = n() / nrow(tbell))

cat('Probability from normal distr.:',normal,
    '\nProbability from empirical distr.:',empirical[[1]])
```

```
## Probability from normal distr.: 0.9610869
## Probability from empirical distr.: 0.9304348
```

For both questions, the parametric and empirical distributions produce very similar results. However, the probabilities for the first question (related to sampling an item with >800 calories) are slightly closer than those for the second (related to items with <30g of protein). Still, while further testing would be needed, I would suspect the distributions of both items are roughly normal, at least in the tails.

Finally, I wanted to run a quick test on a package that can estimate the probability of a given quantile from an empirical distribution. That package is `EnvStats`, and the related function is `pemp`.

```
# library(EnvStats)

emp_calories <- 1 - pemp(800, tbell$calories, discrete = TRUE)
emp_protein <- pemp(30, tbell$protein, discrete = TRUE)

cat('Calories >800 probability from EnvStats:',emp_calories,
    '\nProtein <30g probability from EnvStats:',emp_protein)
```

```
## Calories >800 probability from EnvStats: 0.02608696
## Protein <30g probability from EnvStats: 0.9304348
```

In both cases, the probabilities match the results for our “manual” approach above!

Exercise 7

Question

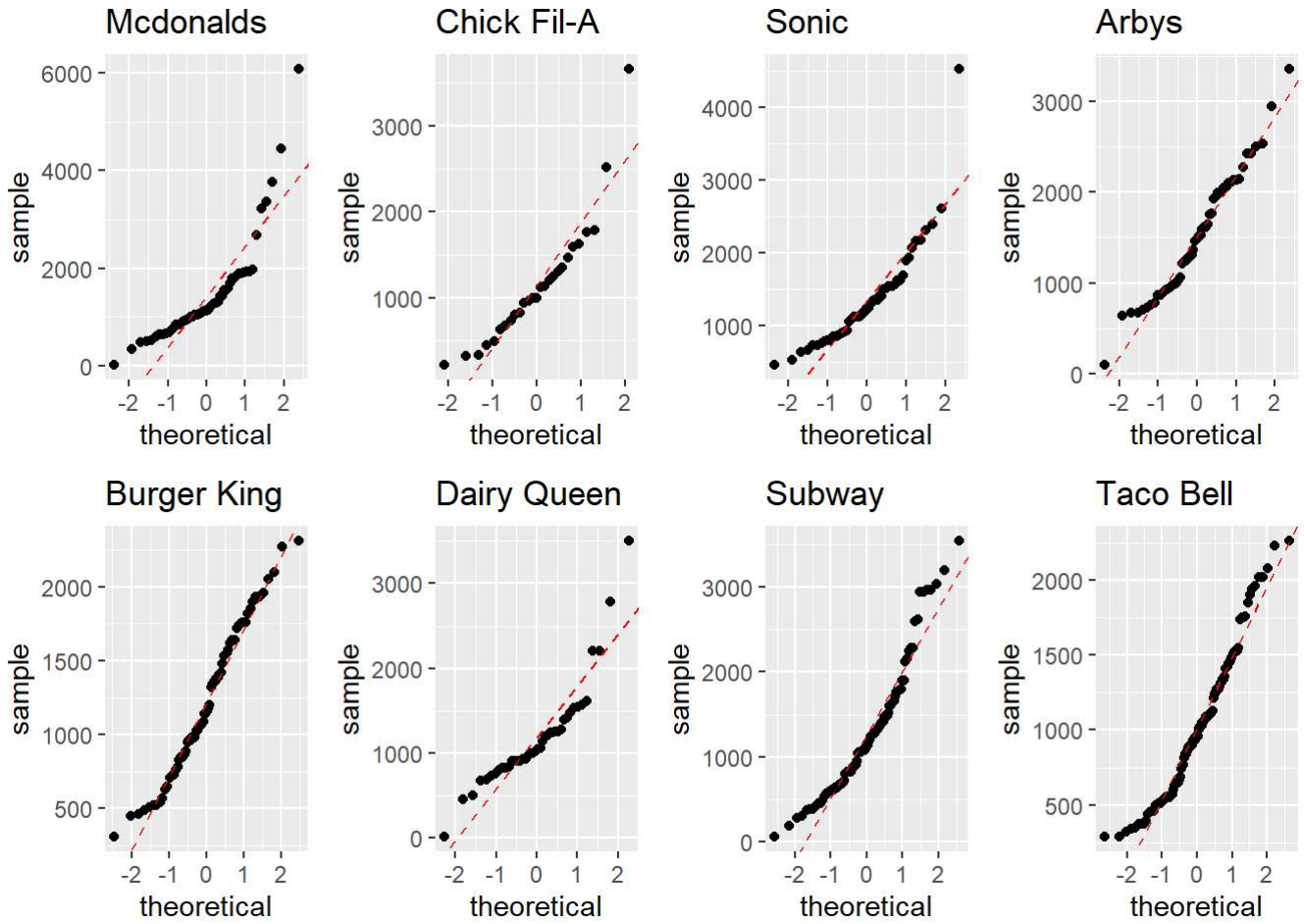
Now let’s consider some of the other variables in the dataset. Out of all the different restaurants, which ones’ distribution is the closest to normal for sodium?

Response

```
restaurants <- unique(fastfood$restaurant)
plot_list <- list()

for (i in restaurants) {
  subset <- fastfood %>%
    filter(restaurant == i)
  p <- ggplot() +
    geom_point(data = subset, aes(sample = sodium), stat = 'qq') +
    geom_abline(intercept = mean(subset$sodium), slope = sd(subset$sodium),
                color = 'red', linetype = 'dashed') +
    ggtitle(i)
  plot_list[[i]] <- p
}

plot_grid(plotlist = plot_list, nrow = 2, ncol = 4)
```



Based on the above plots, the distribution of sodium in Burger King products appears the most close to normal, but Taco Bell and Arby's come in a close second.

WARNING: A wild tangent appears!

As an experiment, I'd like to compare the above plots to results from the Anderson-Darling goodness-of-fit test. The test provides an assessment of the goodness-of-fit for a set of observations to some parametric distribution (in this case a normal). The Null Hypothesis is that the data are sampled from a normal distribution, and the Alternative Hypothesis is that they are not.

We can use the `ad.test()` function in the `nortest` package to assess goodness-of-fit to a normal distribution.

```r
# library(nortest)

results <- list()

for (i in restaurants) {
  subset <- fastfood %>%
    filter(restaurant == i)
  result <- ad.test(subset$sodium)
  result[[4]] <- i
  results[[i]] <- result
}

summary_results <- data.frame(matrix(nrow=0, ncol = 2))
colnames(summary_results) <- c('Restaurant','Pass-Fail')

for (i in 1:length(results)) {
  title <- results[[i]][[4]]
  pvalue <- results[[i]][[2]]
  threshold <- 0.05
  result <- if_else(pvalue < threshold, 'Reject Null', 'Accept Null')
  cat('--', title ,'--\n',
       'Test Stat:', results[[i]][[1]], '\n',
       'P-value:', pvalue, '\n',
       'Result:', result, '\n\n')
  summary_results[i,'Restaurant'] <- title
  summary_results[i,'Result'] <- result
}
```

```
## -- Mcdonalds --
##  Test Stat: 3.788831
##  P-value: 1.43533e-09
##  Result: Reject Null
##
## -- Chick Fil-A --
##  Test Stat: 0.8489713
##  P-value: 0.02501141
##  Result: Reject Null
##
## -- Sonic --
##  Test Stat: 1.738074
##  P-value: 0.0001633778
##  Result: Reject Null
##
## -- Arbys --
##  Test Stat: 0.5869544
##  P-value: 0.1201279
##  Result: Accept Null
##
## -- Burger King --
##  Test Stat: 0.5320359
##  P-value: 0.1680115
##  Result: Accept Null
##
## -- Dairy Queen --
##  Test Stat: 2.017902
##  P-value: 3.141413e-05
##  Result: Reject Null
##
## -- Subway --
##  Test Stat: 2.204952
##  P-value: 1.23248e-05
##  Result: Reject Null
##
## -- Taco Bell --
##  Test Stat: 1.252556
##  P-value: 0.002807607
##  Result: Reject Null
```

```
summary_results
```

```
##    Restaurant Pass-Fail      Result
## 1   Mcdonalds        NA Reject Null
## 2 Chick Fil-A        NA Reject Null
## 3       Sonic        NA Reject Null
## 4       Arbys        NA Accept Null
## 5 Burger King        NA Accept Null
## 6 Dairy Queen        NA Reject Null
## 7      Subway        NA Reject Null
## 8   Taco Bell        NA Reject Null
```

The results largely confirm what we inferred from the Q-Q plots above. For the two restaurants with very clear diagonal plots — Burger King and Arby's — the null of normality is not rejected. I was, however, surprised to see the null was rejected for Taco Bell, for which the Q-Q plot implied the distribution was approximately normal.

Frankly, I am not terribly familiar with the AD test, so it's quite possible there are susceptibilities for this test (e.g. sensitivity to certain data types or small sample sizes) that may have impacted results.

This will certainly be an area for study in the future!

NOTE: Tangent over

Exercise 8

Question

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?

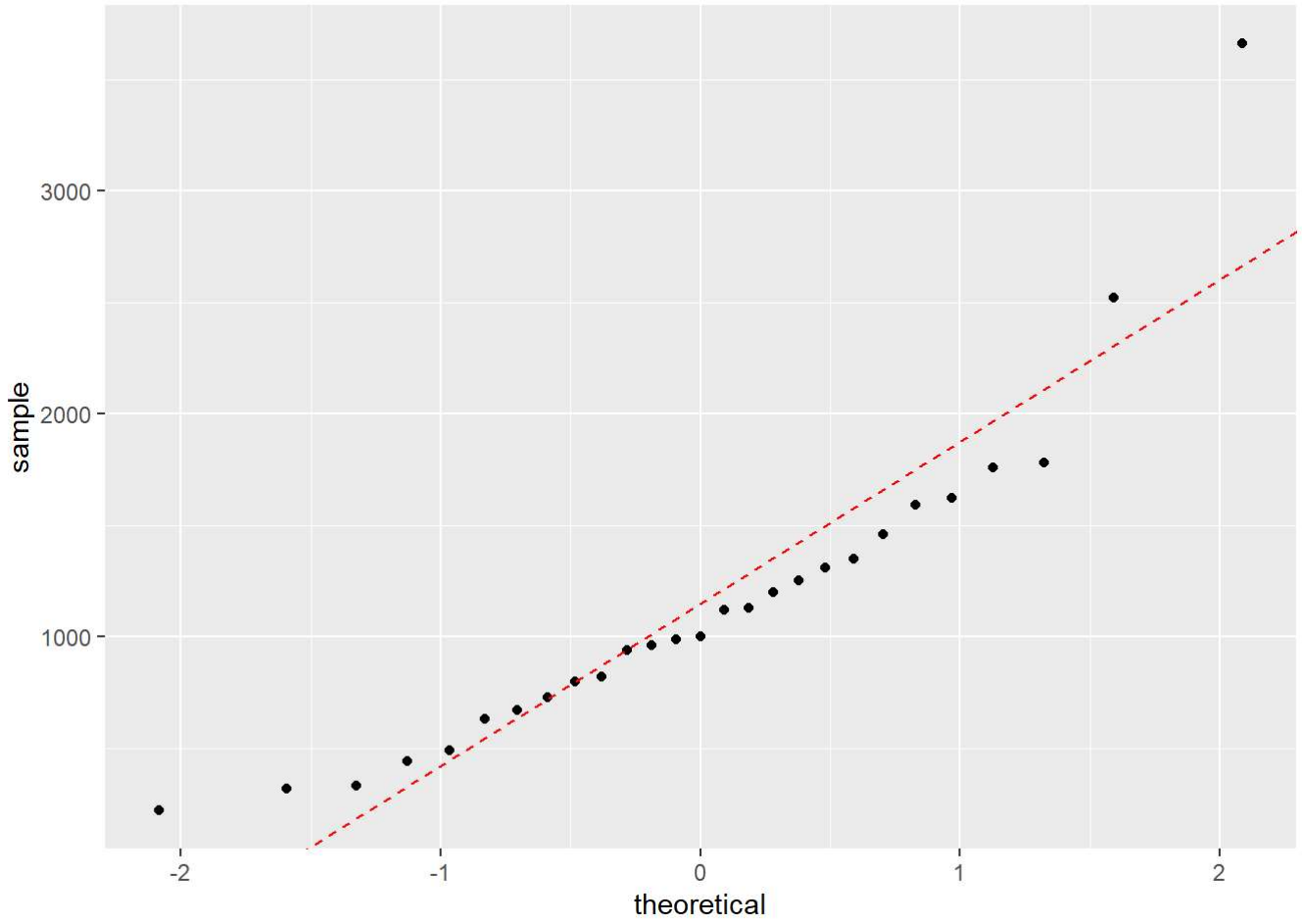
Response

The literature indicates that stepwise patterns in Q-Q plots are more prominent when dealing with discrete variables. While the amount of sodium is technically a continuous variable, if its measurement is discretized, then the normal probability plot may exhibit the stepwise pattern.

From the plots in Response 7, the plot for Chick Fil-A products appears to exhibit this pattern. The plot below provides a closer look. I've also provided a list of the unique values for sodium in Chick Fil-A products to assess the level of discretization.

```
chick_fila <- filter(fastfood, restaurant == 'Chick Fil-A')

ggplot(chick_fila) +
  geom_point(aes(sample = sodium), stat = 'qq') +
  geom_abline(intercept = mean(chick_fila$sodium), slope = sd(chick_fila$sodium),
              color = 'red', linetype = 'dashed')
```




```
chick_fila %>%
  arrange(sodium) %>%
  select(sodium) %>%
  unique() %>%
  as.list()
```

```
## $sodium
##  [1]  220  320  330  440  490  630  670  730  800  820  940  960  990 1000 1120
## [16] 1130 1200 1250 1310 1350 1460 1590 1620 1760 1780 2520 3660
```

The plot confirms that there is a slight stepwise pattern in the Q-Q plot. However, while there appears to be some rounding (specifically, to the nearest ten), the sodium levels cover a range of specific values, so there is little evidence that the variable was discretized to a meaningful degree. The stepwise pattern therefore appears to have occurred moreso by chance.

Exercise 9

Question

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.

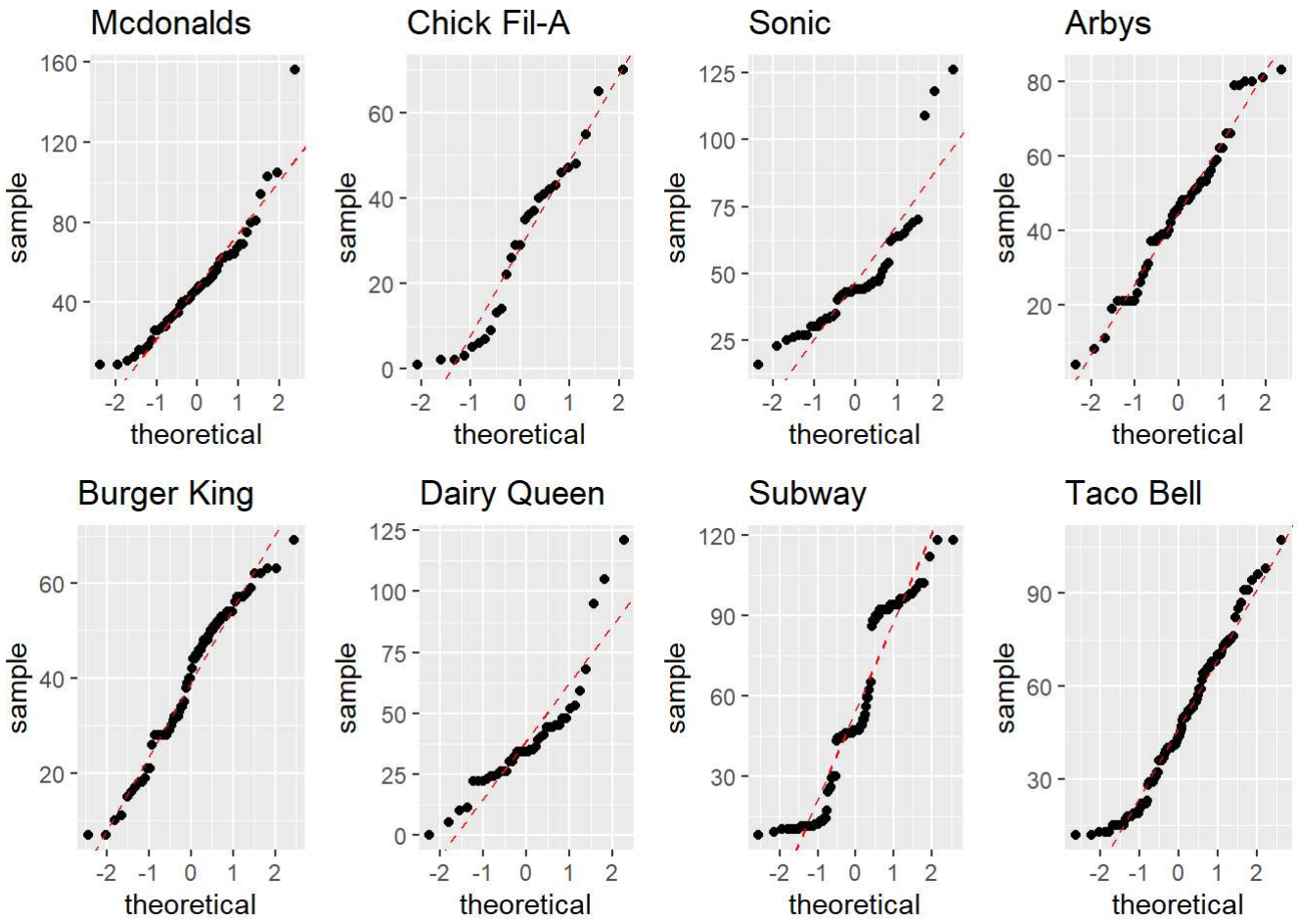
Response

I begin by again producing normal probability plots for all restaurants, but this time focusing on total carbohydrates.

```
plot_list <- list()

for (i in restaurants) {
  subset <- fastfood %>%
    filter(restaurant == i)
  p <- ggplot() +
    geom_point(data = subset, aes(sample = total_carb), stat = 'qq') +
    geom_abline(intercept = mean(subset$total_carb), slope = sd(subset$total_carb),
                color = 'red', linetype = 'dashed') +
    ggtitle(i)
  plot_list[[i]] <- p
}

plot_grid(plotlist = plot_list, nrow = 2, ncol = 4)
```



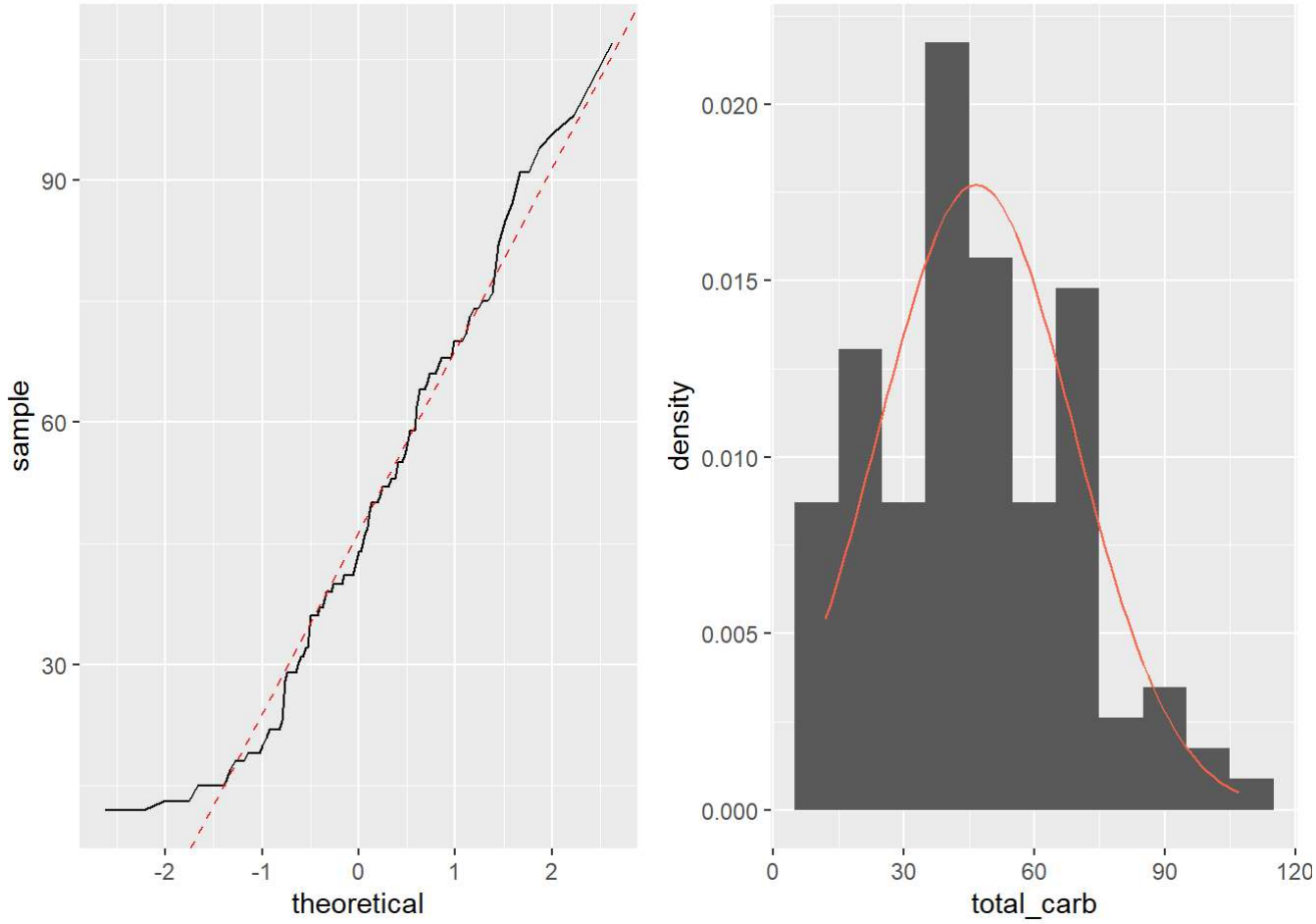
The appearance of near-horizontal segments in the line, especially towards zero on the x-axis, typically indicates skewness. A bunching of values to the left of the mean indicates right skewness, and vice versa for values bunched to the right of the mean. In the above plots, the distribution of carbs in Taco Bell products appears to give the clearest indication of skewness, as we see a horizontal segment appear to the left of the mean.

To investigate further, I plot a histogram showing the distribution of carbs in Taco Bell products, along with a closer look at the Q-Q plot.

```
p1 <- ggplot(tbell) +
  geom_line(aes(sample = total_carb), stat = 'qq') +
  geom_abline(intercept = mean(tbell$total_carb), slope = sd(tbell$total_carb),
              color = 'red', linetype = 'dashed')

p2 <- ggplot(tbell, aes(x = total_carb)) +
  geom_blank() +
  geom_histogram(aes(y = after_stat(density)), binwidth = 10) +
  stat_function(fun = dnorm,
                args = c(mean = mean(tbell$total_carb), sd = sd(tbell$total_carb)),
                color = 'tomato')

plot_grid(p1,p2)
```



The close-up Q-Q plot again shows a clear horizontal segment, and the histogram shows a slight right skewness.