The normal distribution

Leticia Salazar

2021-09-25

In this lab, you'll investigate the probability distribution that is most central to statistics: the normal distribution. If you are confident that your data are nearly normal, that opens the door to many powerful statistical methods. Here we'll use the graphical tools of R to assess the normality of our data and also learn how to generate random numbers from a normal distribution.

Getting Started

Load packages

In this lab, we will explore and visualize the data using the **tidyverse** suite of packages as well as the **openintro** package.

Let's load the packages.

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

The data

This week you'll be working with fast food data. This data set contains data on 515 menu items from some of the most popular fast food restaurants worldwide. Let's take a quick peek at the first few rows of the data.

Either you can use `glimpse` like before, or `head` to do this.

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

```
## # A tibble: 6 x 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>
```

You'll see that for every observation there are 17 measurements, many of which are nutritional facts.

You'll be focusing on just three columns to get started: restaurant, calories, calories from fat.

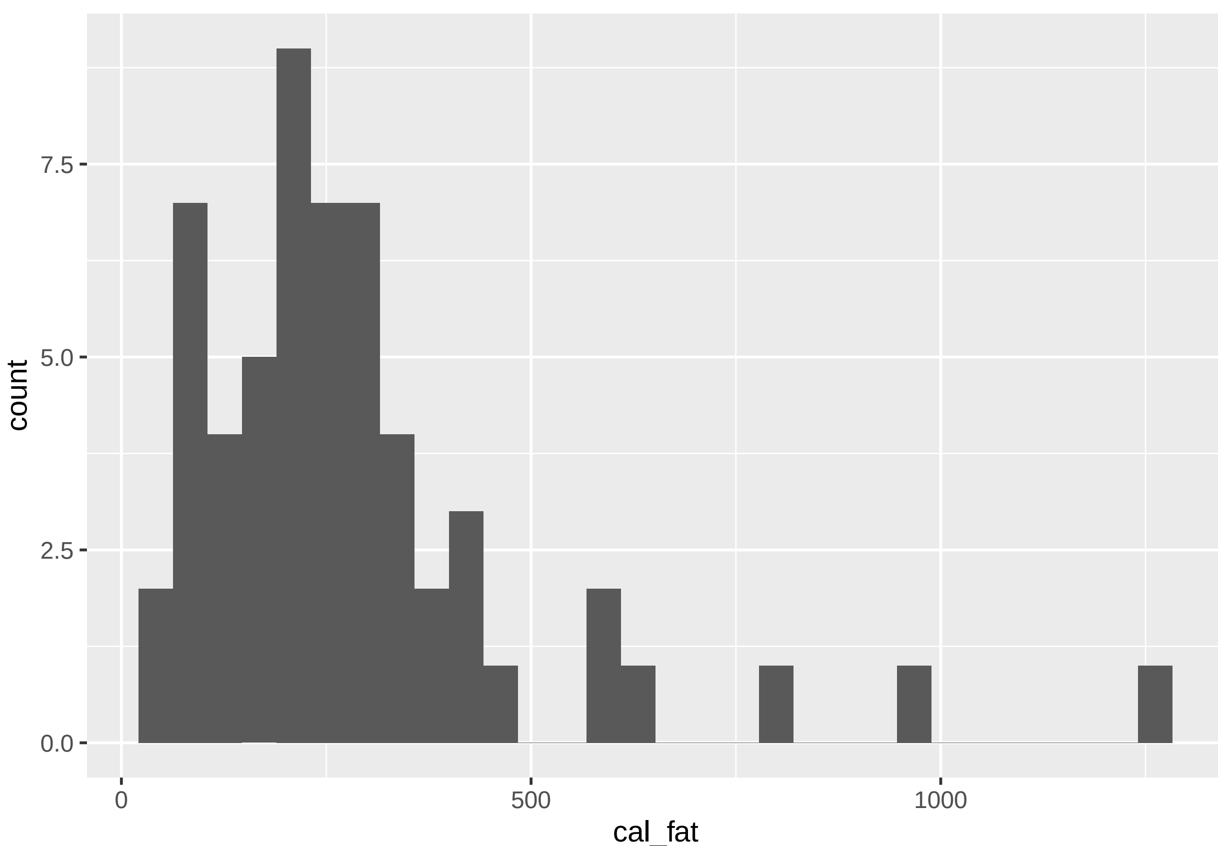
Let's first focus on just products from McDonalds and Dairy Queen.

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

Both histograms appear to be unimodal and right skewed. When finding the summary for both, McDonald's mean is of 285.6 with a max IQR of 1270. Meanwhile, Dairy Queen's mean is 260.5 with a max IQR of 670. There are plenty more outliers in McDonald's than there is of Dairy Queen. In terms of their shape, they the Dairy Queen histogram appears to be more normally distributed than the McDonald's histogram since it has less larger outliers.

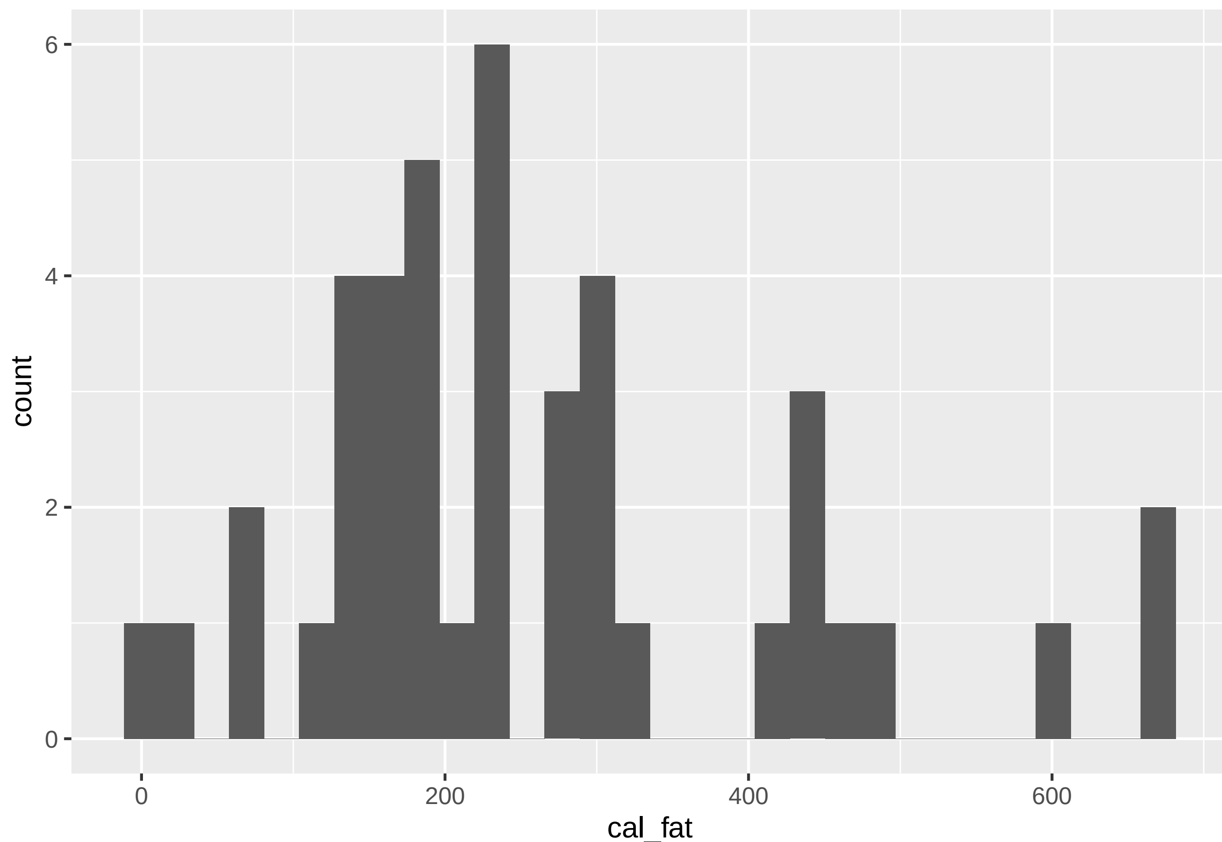
```
#McDonald's Histogram
McDonalds <- ggplot(mcdonalds, aes(x=cal_fat)) +
  geom_histogram()
McDonalds
```



```r
#Summary of data
summary(mcdonalds$cal_fat)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    50.0   160.0   240.0   285.6   320.0  1270.0
```

```r
#Dairy Queen Histogram
Dairy_Queen <- ggplot(dairy_queen, aes(x=cal_fat)) +
  geom_histogram()
Dairy_Queen
```




```r
#Summary of data
summary(dairy_queen$cal_fat)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   160.0   220.0   260.5   310.0   670.0
```

The normal distribution

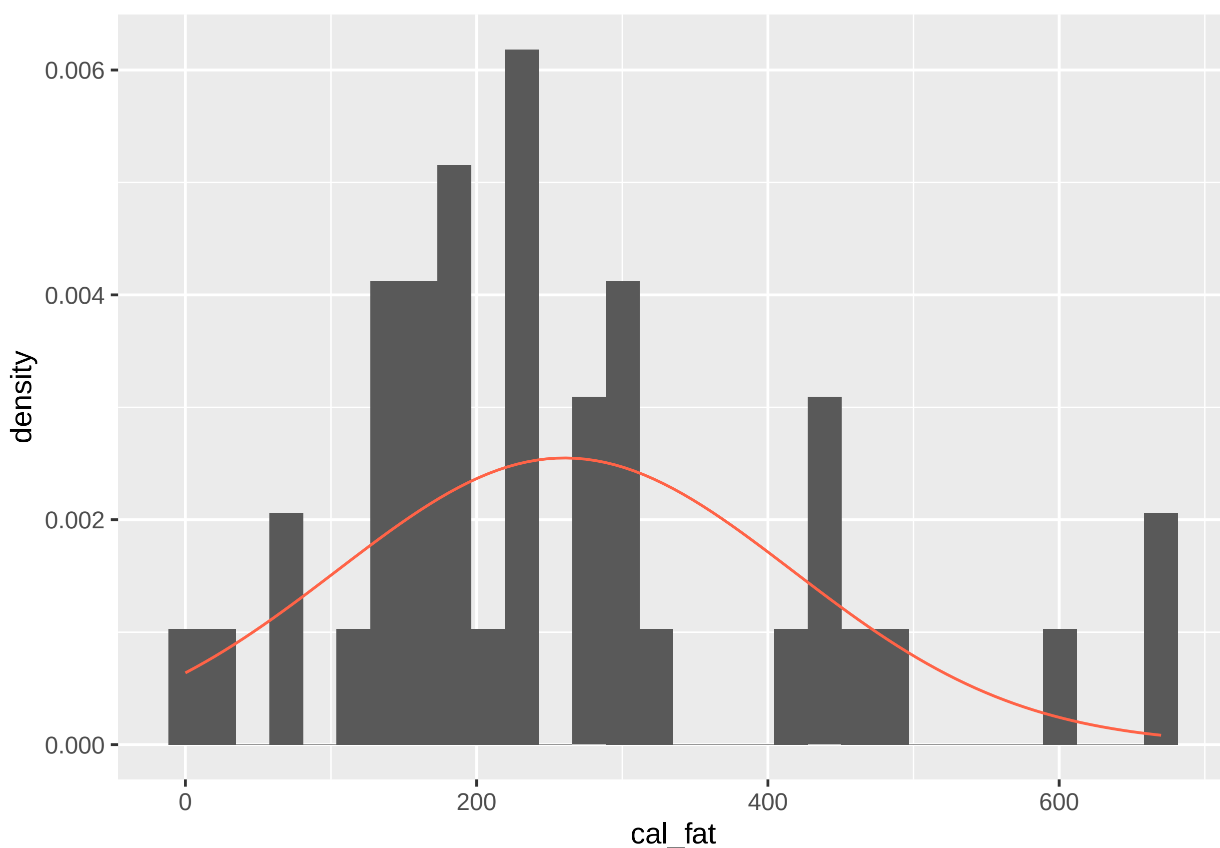
In your description of the distributions, did you use words like *bell-shaped*or *normal*? It's tempting to say so when faced with a unimodal symmetric distribution.

To see how accurate that description is, you can plot a normal distribution curve on top of a histogram to see how closely the data follow a normal distribution. This normal curve should have the same mean and standard deviation as the data. You'll be focusing on calories from fat from Dairy Queen products, so let's store them as a separate object and then calculate some statistics that will be referenced later.

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

Next, you make a density histogram to use as the backdrop and use the `lines` function to overlay a normal probability curve. The difference between a frequency histogram and a density histogram is that while in a frequency histogram the *heights* of the bars add up to the total number of observations, in a density histogram the *areas* of the bars add up to 1. The area of each bar can be calculated as simply the height *times* the width of the bar. Using a density histogram allows us to properly overlay a normal distribution curve over the histogram since the curve is a normal probability density function that also has area under the curve of 1. Frequency and density histograms both display the same exact shape; they only differ in their y-axis. You can verify this by comparing the frequency histogram you constructed earlier and the density histogram created by the commands below.

```
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")
```

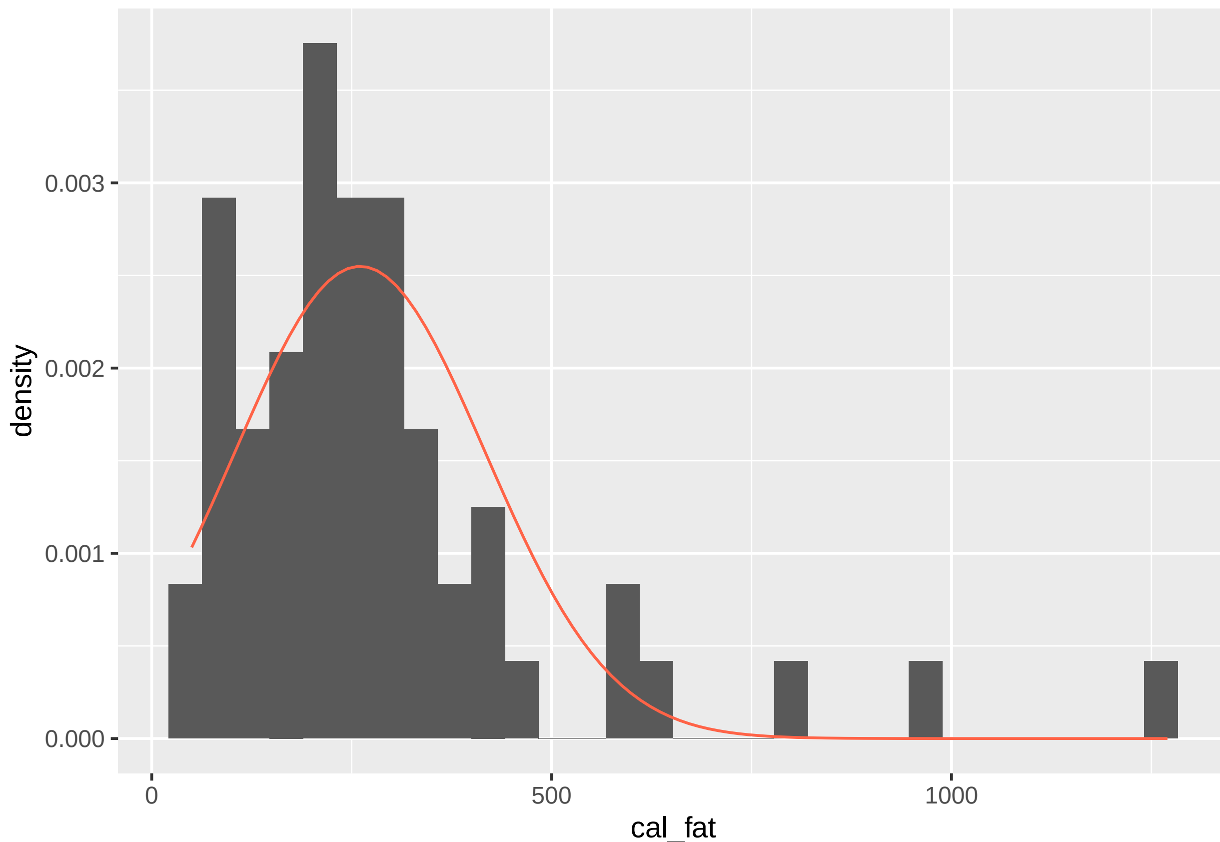


After initializing a blank plot with `geom_blank()`, the `ggplot2` package (within the `tidyverse`) allows us to add additional layers. The first layer is a density histogram. The second layer is a statistical function – the density of the normal curve, `dnorm`. We specify that we want the curve to have the same mean and standard deviation as the column of fat calories. The argument `col` simply sets the color for the line to be drawn. If we left it out, the line would be drawn in black.

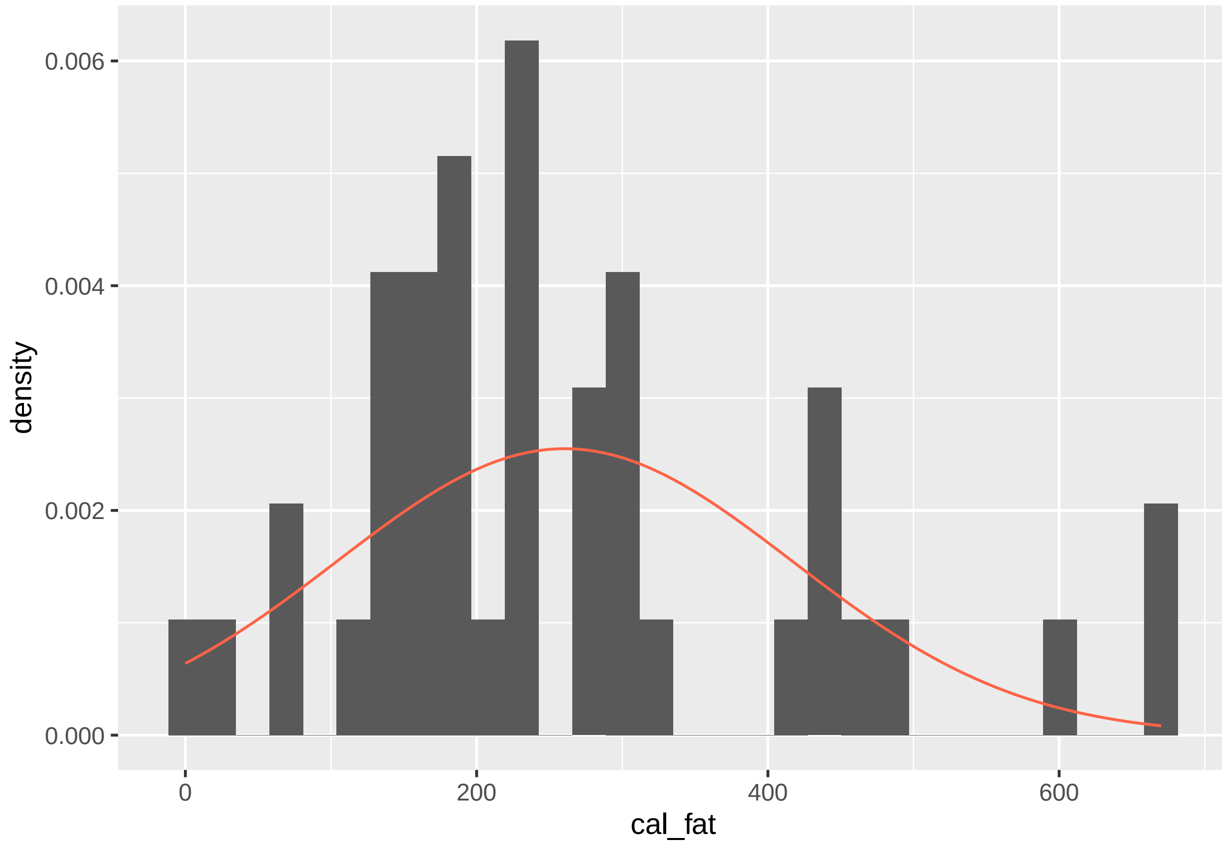
2. Based on the this plot, does it appear that the data follow a nearly normal distribution?

Based on the plots below, both plots appear to have a near normal distribution until it reaches 250, that's when the appearance changes. In the McDonald's histogram you can see the right skewness even more as opposed to the Dairy Queen histogram where there is more of an appearance to a normal distribution even at it's peak.

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



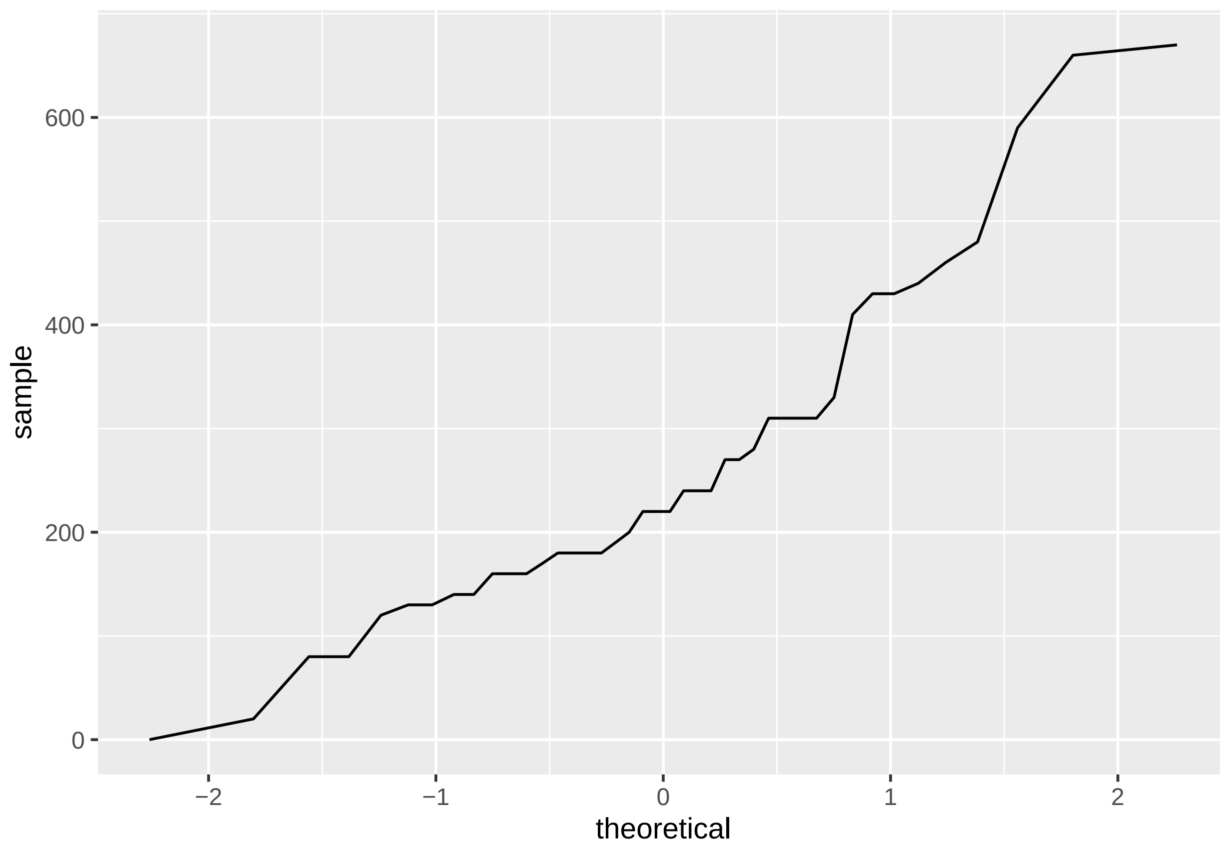

```
#Dairy Queen density histogram
ggplot(data = dairy_queen, aes(x = cal_fat)) +
        geom_blank() +
        geom_histogram(aes(y = ..density..)) +
        stat_function(fun = dnorm, args = c(mean = dqmean, sd = dqsd), col = "tomato")
```



Evaluating the normal distribution

Eyeballing the shape of the histogram is one way to determine if the data appear to be nearly normally distributed, but it can be frustrating to decide just how close the histogram is to the curve. An alternative approach involves constructing a normal probability plot, also called a normal Q-Q plot for "quantile-quantile".

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



This time, you can use the `geom_line()` layer, while specifying that you will be creating a Q-Q plot with the `stat` argument. It's important to note that here, instead of using `x` instead `aes()`, you need to use `sample`.

The x-axis values correspond to the quantiles of a theoretically normal curve with mean 0 and standard deviation 1 (i.e., the standard normal distribution). The y-axis values correspond to the quantiles of the original unstandardized sample data. However, even if we were to standardize the sample data values, the Q-Q plot would look identical. A data set that is nearly normal will result in a probability plot where the points closely follow a diagonal line. Any deviations from normality leads to deviations of these points from that line.

The plot for Dairy Queen's calories from fat shows points that tend to follow the line but with some errant points towards the upper tail. You're left with the same problem that we encountered with the histogram above: how close is close enough?

A useful way to address this question is to rephrase it as: what do probability plots look like for data that I *know* came from a normal distribution? We can answer this by simulating data from a normal distribution using `rnorm`.

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

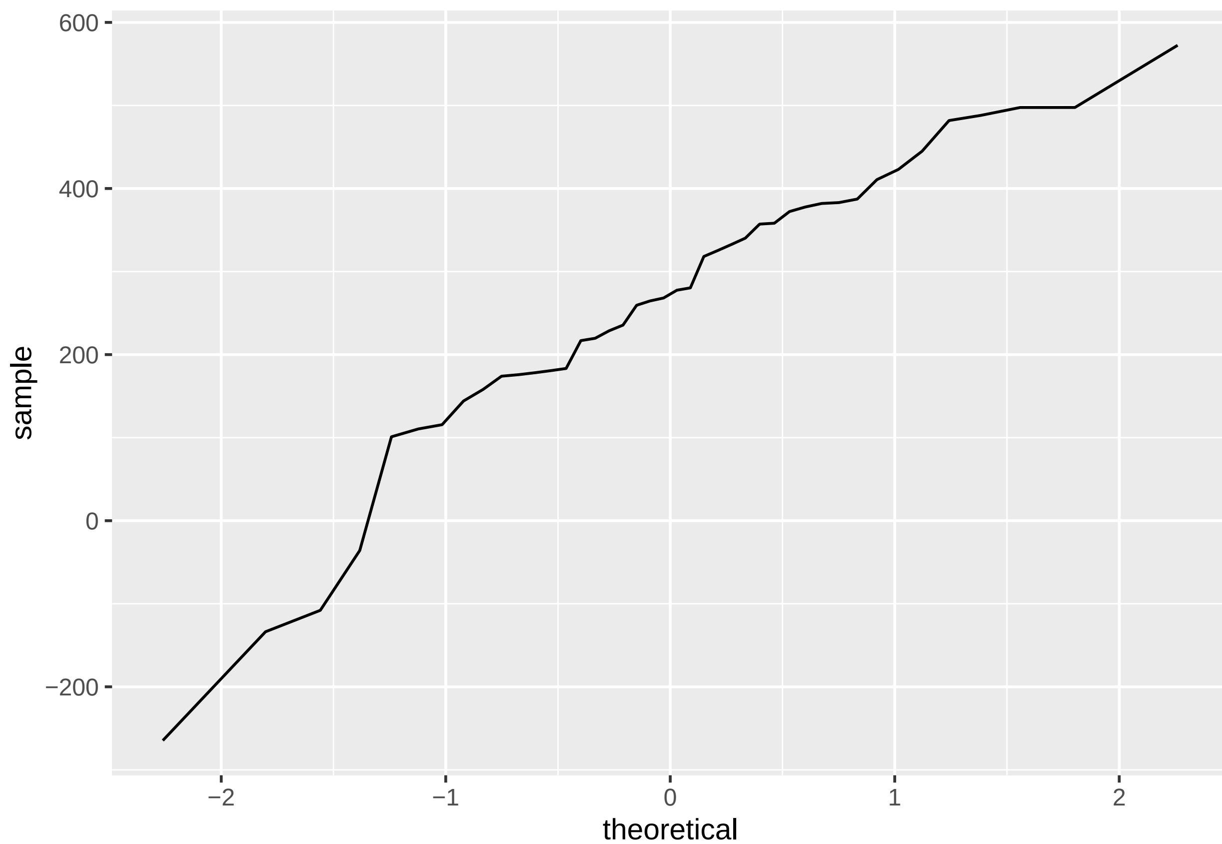
The first argument indicates how many numbers you'd like to generate, which we specify to be the same number of menu items in the `dairy_queen` data set using the `nrow()` function. The last two arguments determine the mean and standard deviation of the normal distribution from which the simulated sample will be generated. You can take a look at the shape of our simulated data set, `sim_norm`, as well as its normal probability plot.

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.)

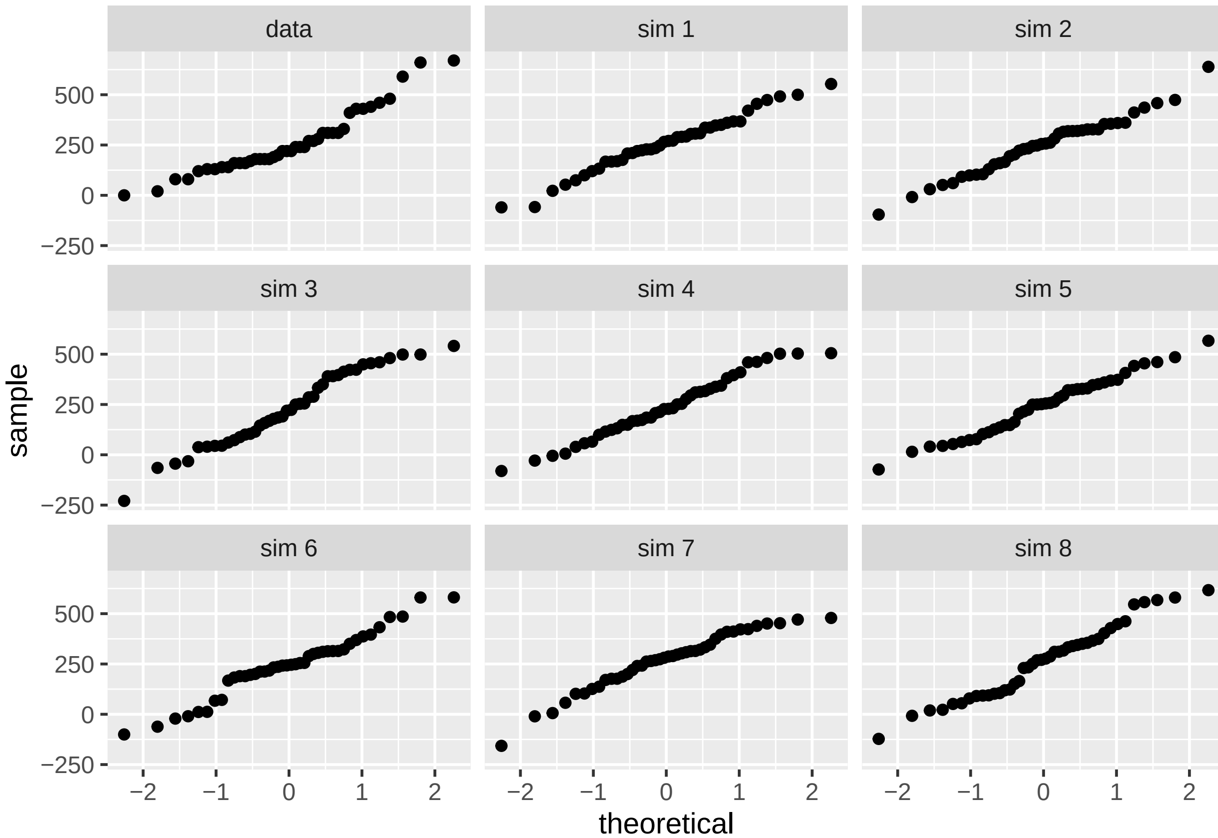
Basing the Q-Q plot above from Dairy Queen to the Q-Q plot from sim_norm, the points .

```
#Plot of normal probability data 'sim_norm'
ggplot(data = NULL, aes(sample = sim_norm)) +
  geom_line(stat = 'qq')
```



Even better than comparing the original plot to a single plot generated from a normal distribution is to compare it to many more plots using the following function. It shows the Q-Q plot corresponding to the original data in the top left corner, and the Q-Q plots of 8 different simulated normal data. It may be helpful to click the zoom button in the plot window.

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



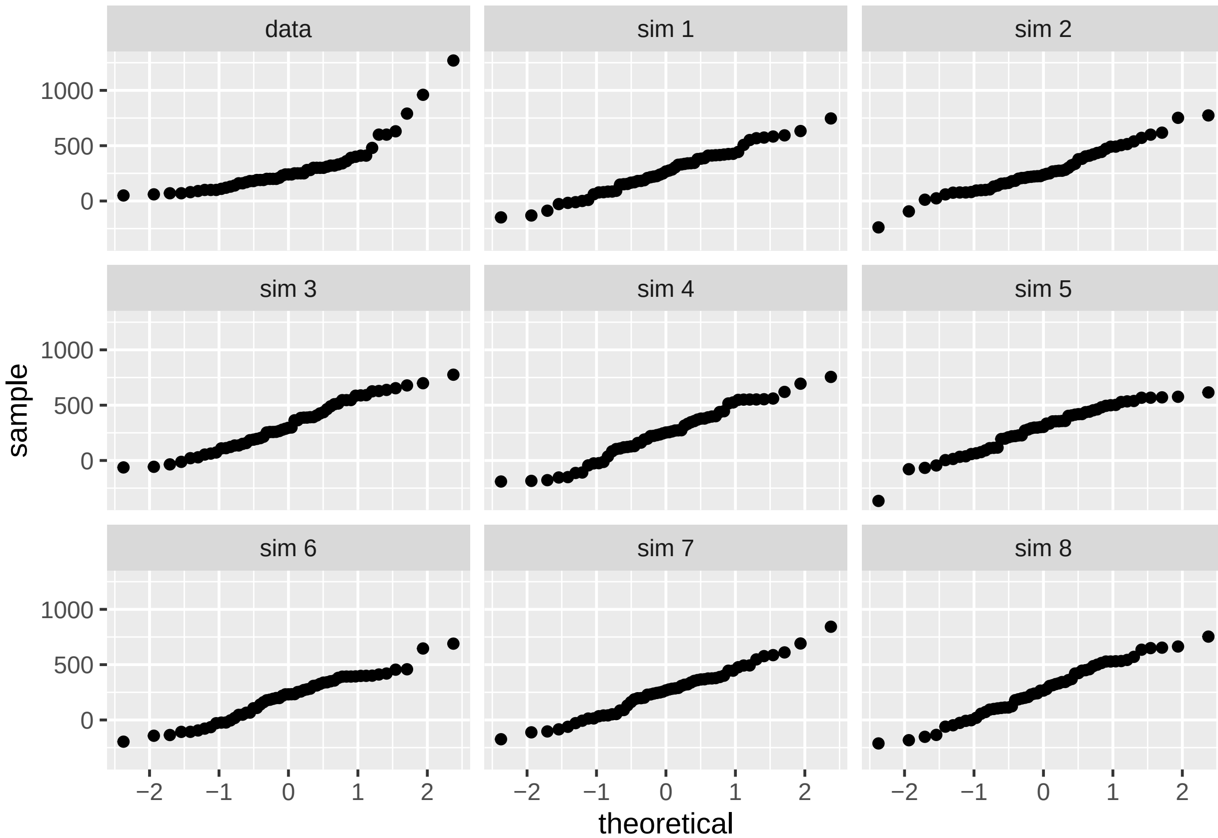
4. Does the normal probability plot for the calories from fat look similar to the plots created for the simulated data? That is, do the plots provide evidence that the calories are nearly normal?

The plots for the calories from fat and the sim_norm do look almost normal. They all have very similar upward trend.

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

The calories from McDonald's do appear to come from a normal distribution. Similarly to the Dairy Queen distribution, there's an upward linear trend showing with a couple of exceptions, the outliers.

```r
#Q-Q Plot of McDonald's
qqnormsim(sample = cal_fat, data = mcdonalds)
```



Normal probabilities

Okay, so now you have a slew of tools to judge whether or not a variable is normally distributed. Why should you care?

It turns out that statisticians know a lot about the normal distribution. Once you decide that a random variable is approximately normal, you can answer all sorts of questions about that variable related to probability. Take, for example, the question of, "What is the probability that a randomly chosen Dairy Queen product has more than 600 calories from fat?"

If we assume that the calories from fat from Dairy Queen's menu are normally distributed (a very close approximation is also okay), we can find this probability by calculating a Z score and consulting a Z table (also called a normal probability table). In R, this is done in one step with the function `pnorm()`.

```
1 - pnorm(q = 600, mean = dqmean, sd = dqsd)
```

```
## [1] 0.01501523
```

Note that the function `pnorm()` gives the area under the normal curve below a given value, `q`, with a given mean and standard deviation. Since we're interested in the probability that a Dairy Queen item has more than 600 calories from fat, we have to take one minus that probability.

Assuming a normal distribution has allowed us to calculate a theoretical probability. If we want to calculate the probability empirically, we simply need to determine how many observations fall above 600 then divide this number by the total sample size.

```
dairy_queen %>%
  filter(cal_fat > 600) %>%
  summarise(percent = n() / nrow(dairy_queen))
```

```
## # A tibble: 1 x 1
##   percent
##     <dbl>
## 1  0.0476
```

Although the probabilities are not exactly the same, they are reasonably close. The closer that your distribution is to being normal, the more accurate the theoretical probabilities will be.

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 Subway has items in their menu that are less than 250 calories? Fiber content >2?

```
#Filter out Subway restaurant
subway <- fastfood %>%
  filter(restaurant == "Subway")
subway
```

```
## # A tibble: 96 x 17
##    restaurant item      calories cal_fat total_fat sat_fat trans_fat cholesterol
##    <chr>      <chr>        <dbl>   <dbl>     <dbl>   <dbl>     <dbl>       <dbl>
##  1 Subway     "6\" B.L~      320      80         9     4           0          20
##  2 Subway     "Footlon~      640     160        18     8           0          40
##  3 Subway     "6\" BBQ~      430     160        18     6           0          50
##  4 Subway     "Footlon~      860     320        36    12           0         100
##  5 Subway     "6\" Big~      580     310        31    11           0          85
##  6 Subway     "Footlon~     1160     620        62    22           0         170
##  7 Subway     "6\" Big~      500     150        17     9           1          85
##  8 Subway     "Footlon~     1000     300        34    18           2         170
##  9 Subway     "Kids Mi~      180      20         3     0.5         0          10
## 10 Subway     "6\" Bla~      290      40         5     1           0          20
## # ... with 86 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>
```

- **0.1849 or 18.49% or calories less than 250 and 0.2936 or 29.36% for fiber content greater than 2**

```
#Find mean and standard deviation for <250 calories and fiber content >2
s_mean <- mean(subway$calories)
s_sd   <- sd(subway$calories)
fiber_mean <- mean(subway$fiber)
fiber_sd <- mean(subway$fiber)

calories <- pnorm(q = 250, mean = s_mean, sd = s_sd)
calories
```

```
## [1] 0.1849837
```

```r
fiber <- pnorm(q = 3, mean = fiber_mean, sd = fiber_sd)
fiber
```

```
## [1] 0.2936141
```

- **0.219 or 21.9% for calories less than 250 and 1% for fiber content greater than 2**

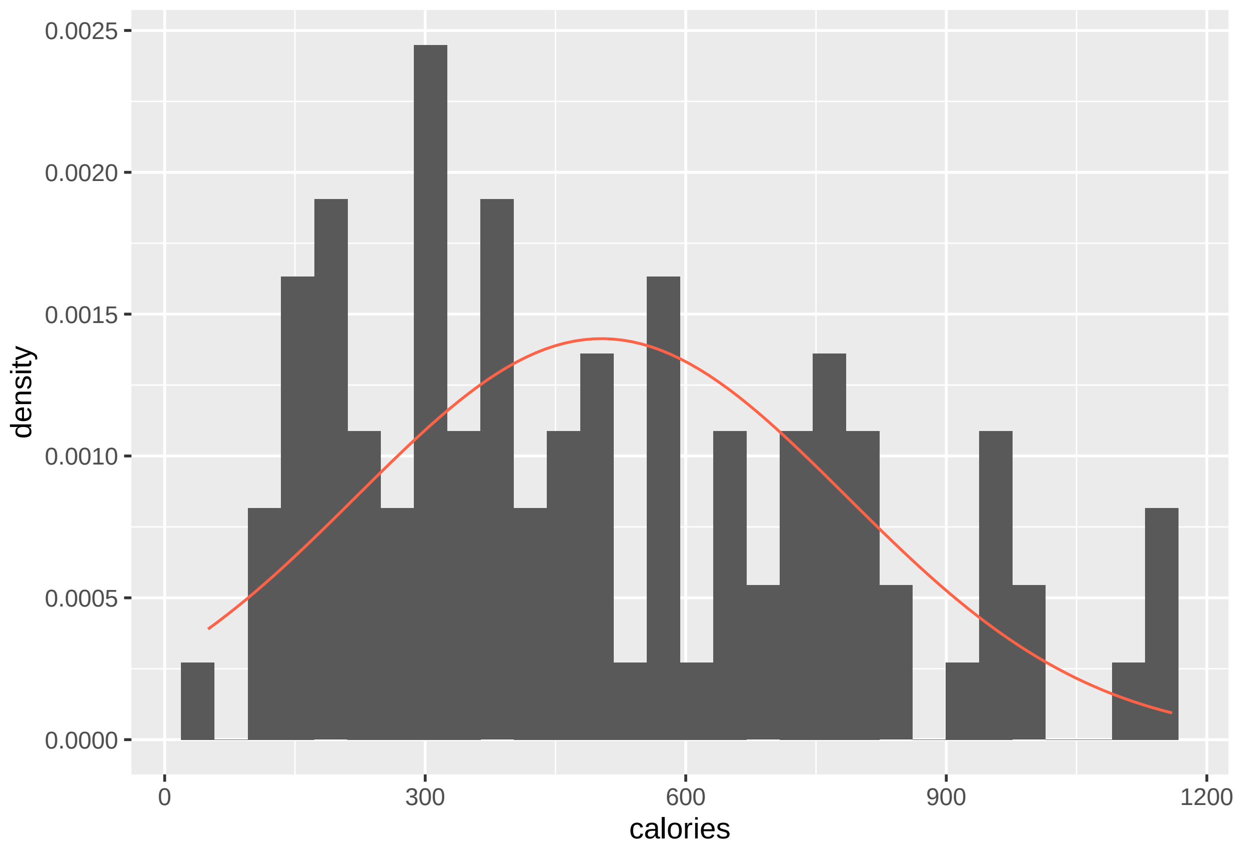
```r
#Filtering calories less than 250
calories <- subway %>%
  filter(calories < 250) %>%
  summarise(percent = n() / nrow(subway))
calories
```

```
## # A tibble: 1 x 1
##   percent
##     <dbl>
## 1   0.219
```

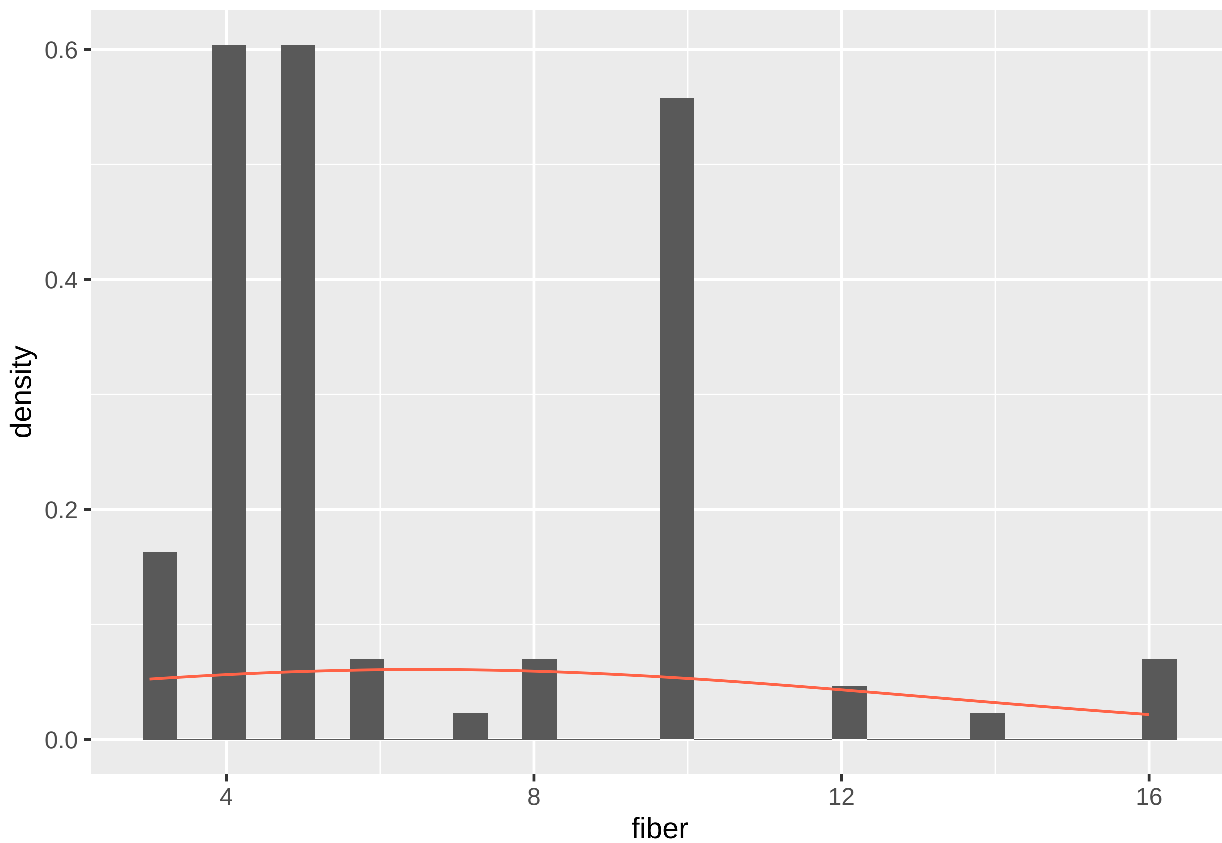
```r
#Filtering calories less than 250
fiber <- subway %>%
  filter(fiber > 2) %>%
  summarise(percent = n() / nrow(subway))
fiber
```

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

```r
#Density histogram of calories from Subway
ggplot(data = subway, aes(x = calories)) +
        geom_blank() +
        geom_histogram(aes(y = ..density..)) +
        stat_function(fun = dnorm, args = c(mean = s_mean, sd = s_sd), col = "tomato")
```




```r
#Density histogram of fiber content > 2 from Subway
ggplot(data = subway, aes(x = fiber)) +
        geom_blank() +
        geom_histogram(aes(y = ..density..)) +
        stat_function(fun = dnorm, args = c(mean = fiber_mean, sd = fiber_sd), col = "tomato")
```

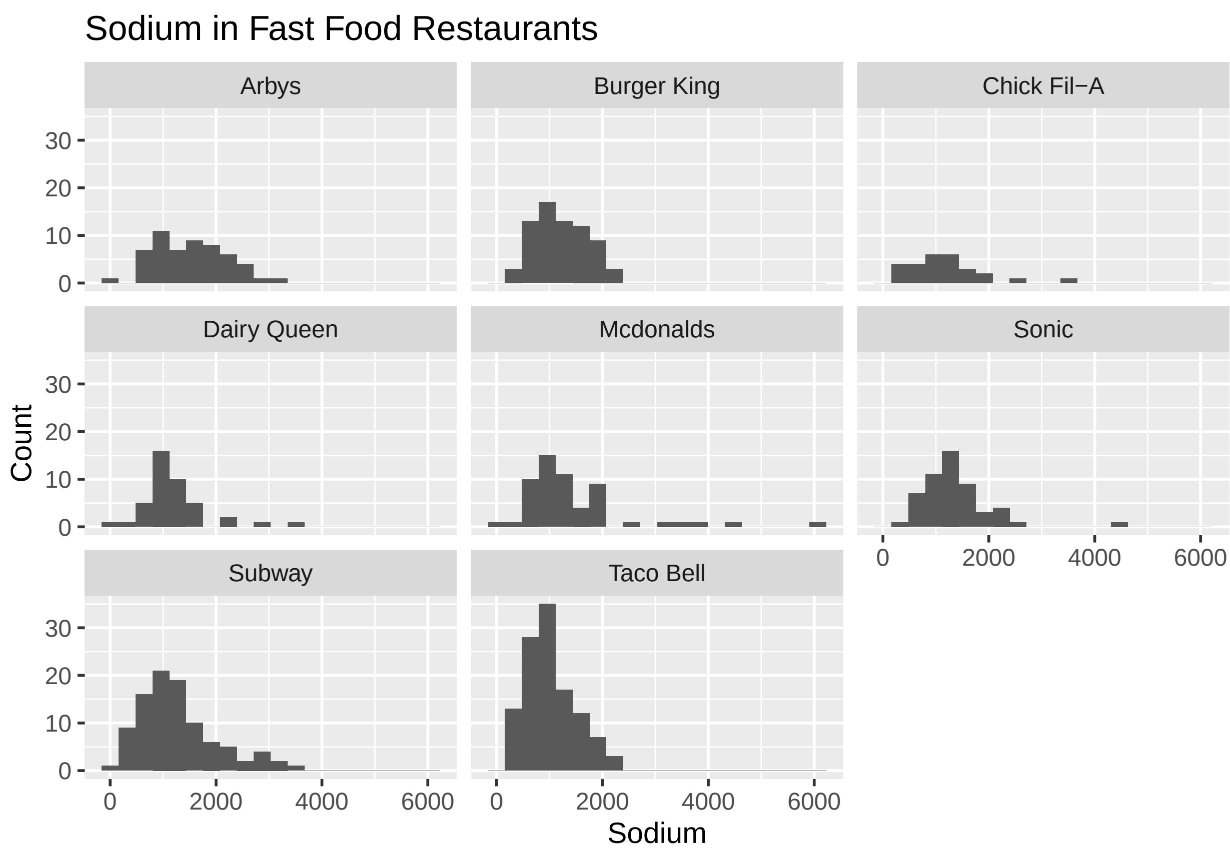


More Practice

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?

Based on the histograms below there are Taco Bell, Arby's, Burger King and Sonic all show close to normal distribution in sodium content.

```
#Histograms Sodium in all Restaurants
fastfood %>%
  group_by(restaurant) %>%
  ggplot() +
  geom_histogram(aes(x = sodium), bins = 20) +
  ggtitle("Sodium in Fast Food Restaurants") +
  xlab("Sodium") +
  ylab("Count") +
  facet_wrap(. ~restaurant)
```



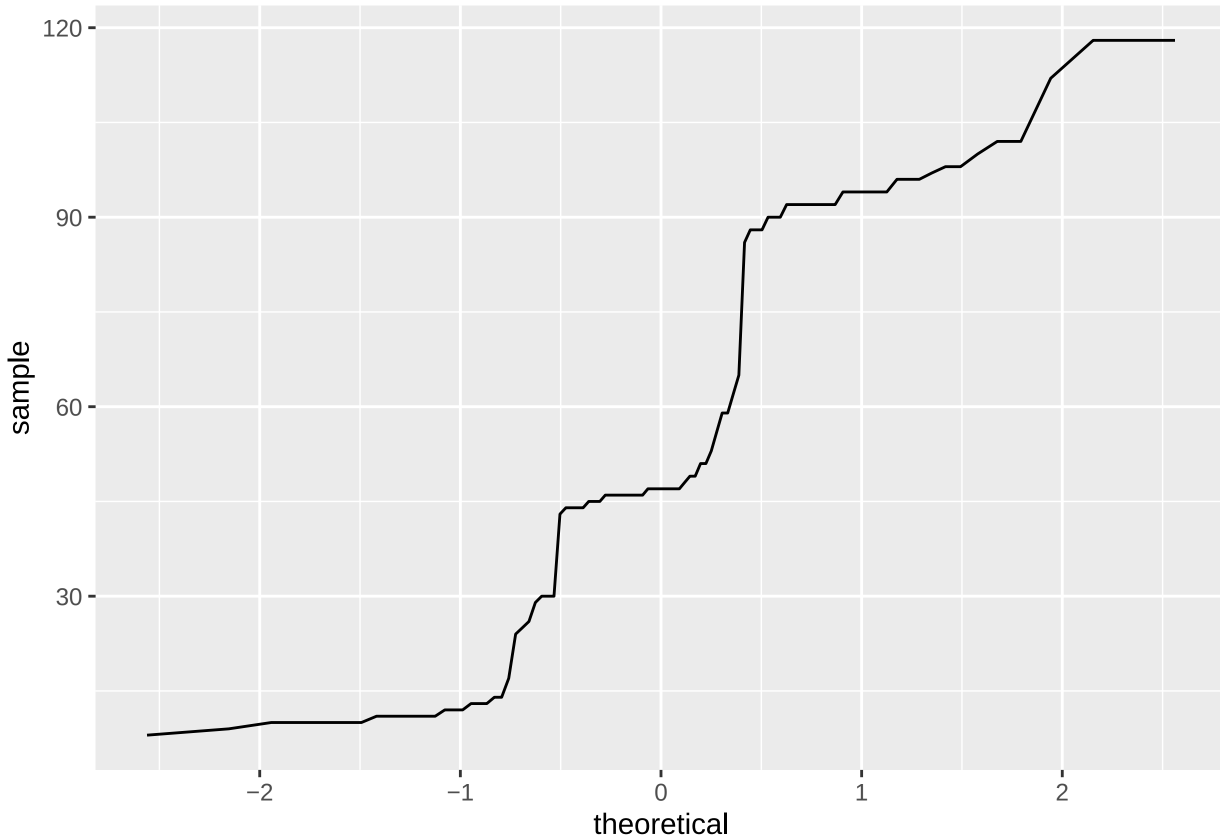
8. Note that some of the normal probability plots for sodium distributions seem to have a stepwise pattern. why do you think this might be the case?

These restaurants often serve a variety of foods that do not require sodium such as salads, fruits, etc so this can be why some of the distributions seem to have a stepwise pattern.

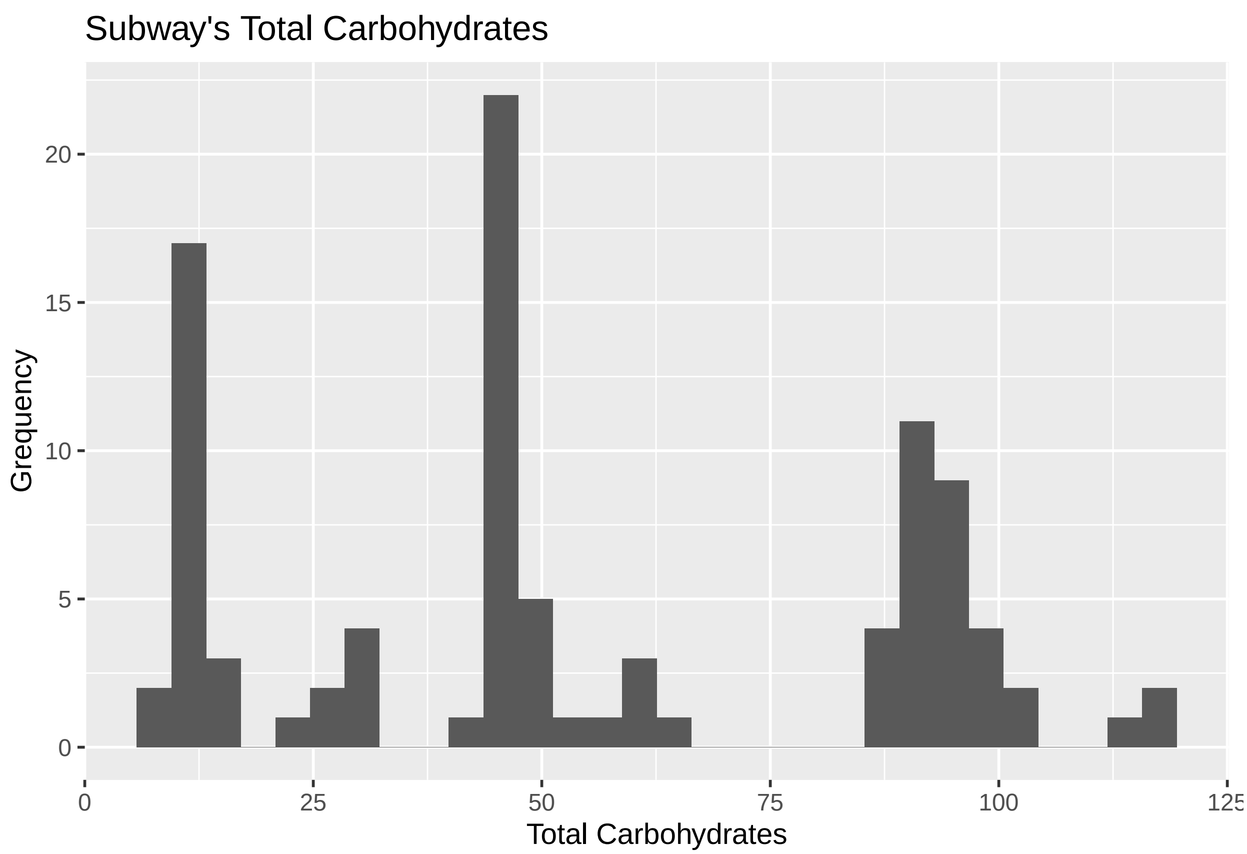
9. As you can see, normal probability plots can be used both to assess normality and visualize skewness. Make a normal probability plot for the total carbohydrates from a restaurant of your choice. Based on this normal probability plot, is this variable left skewed, symmetric, or right skewed? Use a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

Using the plots below, the normal probability plot appears to be symmetric.

```
#Probability Plot
Subway_pp <- ggplot(data = subway, aes(sample = total_carb)) +
  geom_line(stat = "qq")
Subway_pp
```

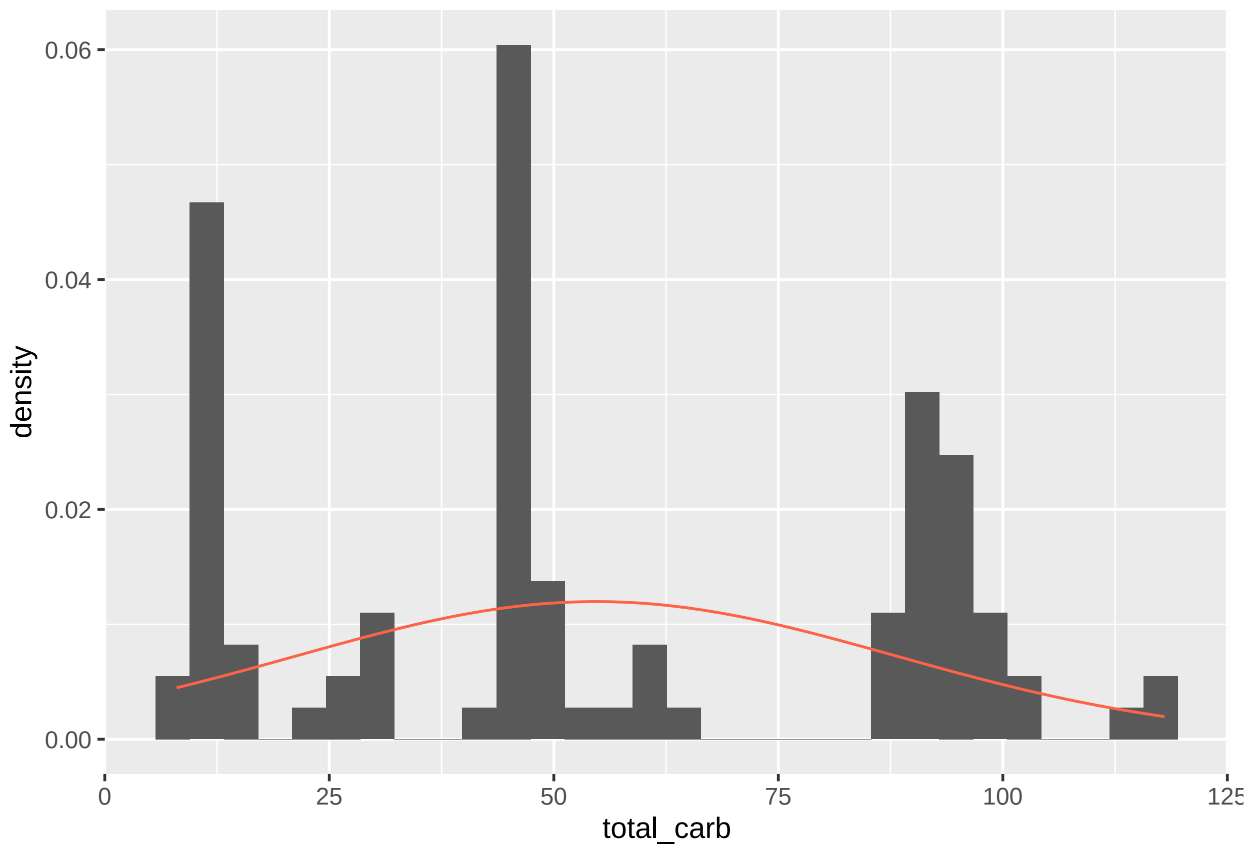



```r
#Subway's Histogram
Subway <- ggplot(subway, aes(x=total_carb)) +
  geom_histogram()  +
  xlab("Total Carbohydrates") +
  ylab("Grequency") +
  ggtitle("Subway's Total Carbohydrates")
Subway
```




```
#Mean and standard deviation for Subway
s_mean <- mean(subway$total_carb)
s_sd   <- sd(subway$total_carb)

#Density Histogram
ggplot(data = subway, aes(x = total_carb)) +
        geom_blank() +
        geom_histogram(aes(y = ..density..)) +
        stat_function(fun = dnorm, args = c(mean = s_mean, sd = s_sd), col = "tomato")
```


0.06
0.04
0.02
0.00
density
0
25
50
75
100
125
total_carb