Probability

Nick Climaco

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

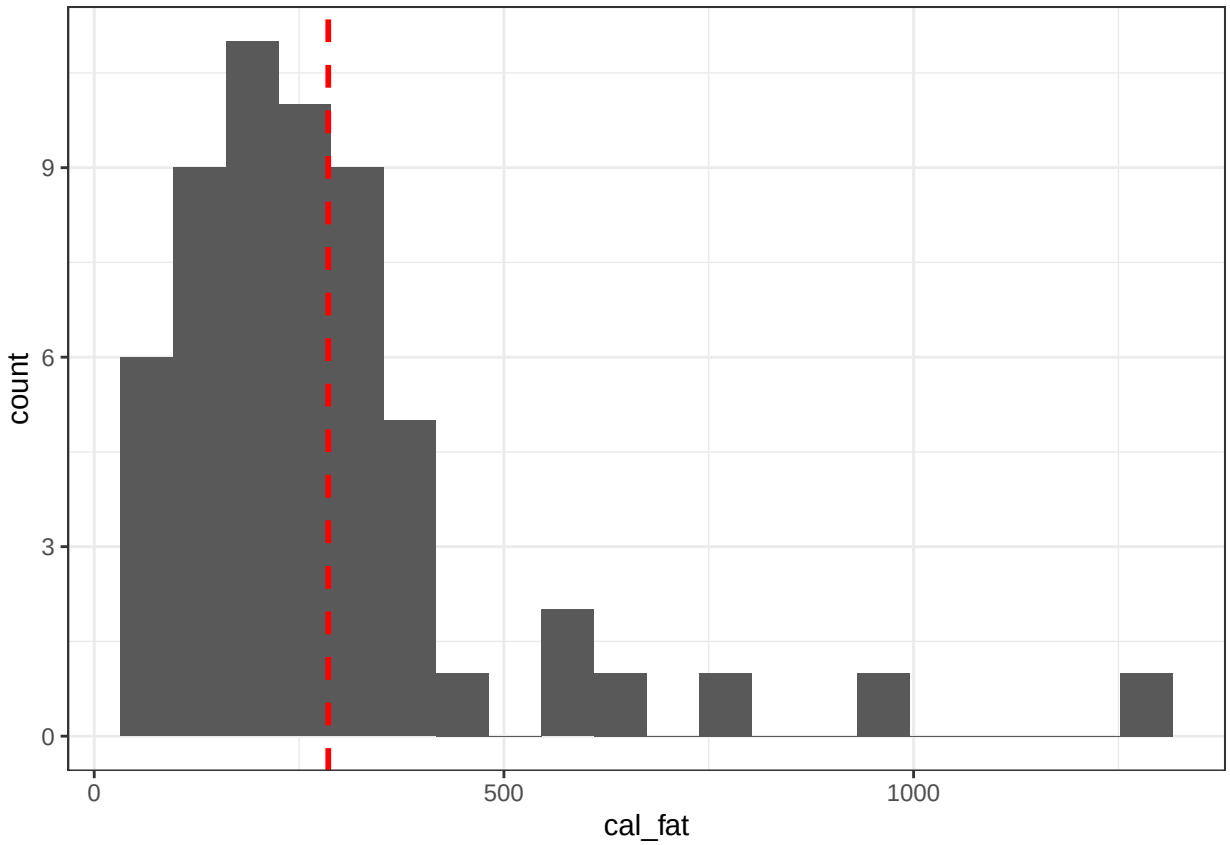
```
library(tidyverse)
library(openintro)
data("fastfood", package='openintro')
head(fastfood)
```

```
## # 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
```

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

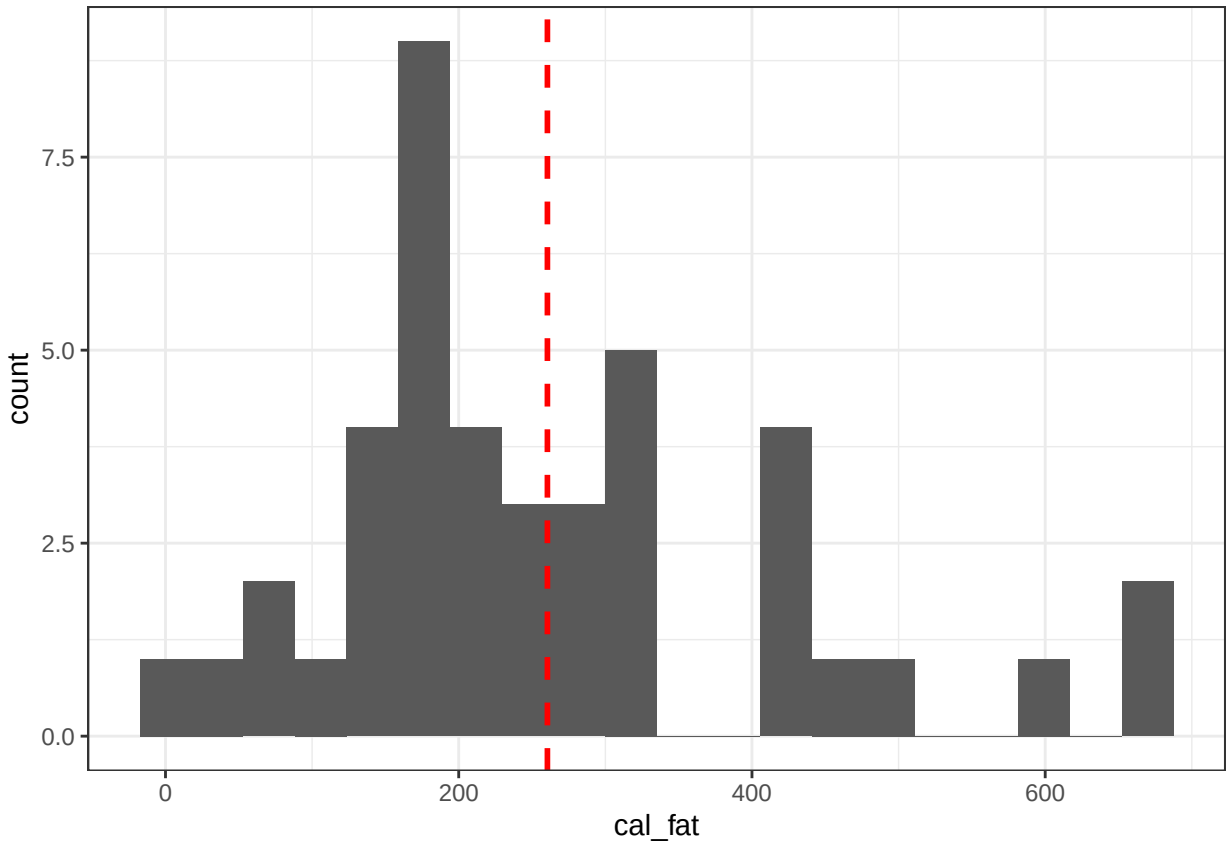
1. Make a plot (or plots) to visualize the distributions of the amount of calories from fat of the options from these two restaurants. How do their centers, shapes, and spreads compare?

```
mcdonalds %>%
    ggplot(aes(x = cal_fat)) +
    geom_histogram(position = "identity", bins = 20) +
    geom_vline(aes(xintercept=mean(cal_fat)),
               color="red",
               linetype="dashed",
               size=1) +
    theme_bw()
```



We can observe from the graph above that it is skewed to the right with the its highest frequency isj ust below the mean. With that, we can infer that most of mcdonalds' menu items have less than 500 calories from fat. At a glance, the graph looks more like it follows a poisson distribution.

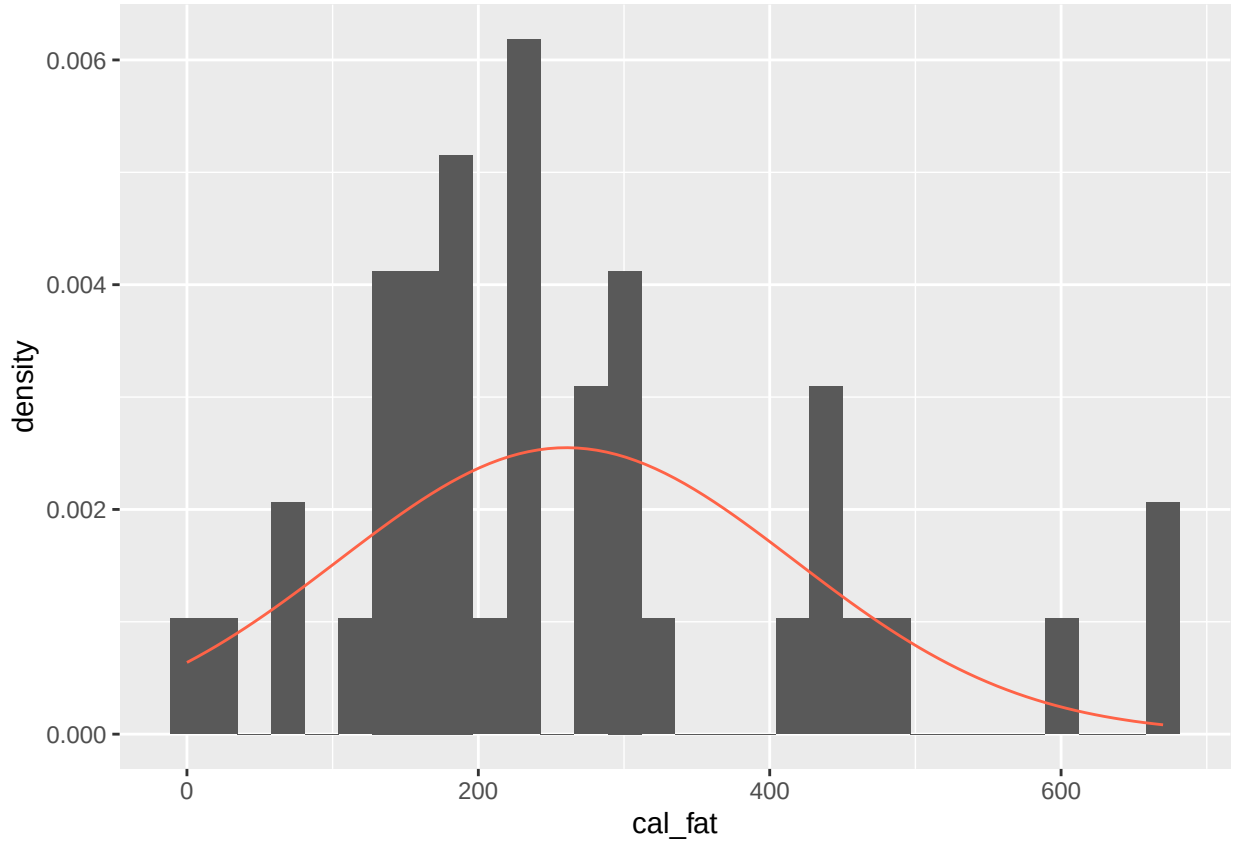
```
dairy_queen %>%
    ggplot(aes(x = cal_fat)) +
    geom_histogram(position = "identity", bins = 20) +
    geom_vline(aes(xintercept=mean(cal_fat)),
               color="red",
               linetype="dashed",
               size=1) +
    theme_bw()
```



The Dairy Queen's graph is more center-aligned compared to the mcdonalds data. DQ's graph are less spread out with the highest caloreies from just under 700 calories, whereas mcdonalds's data is more spread out reaching over 1000 calories from fat.

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

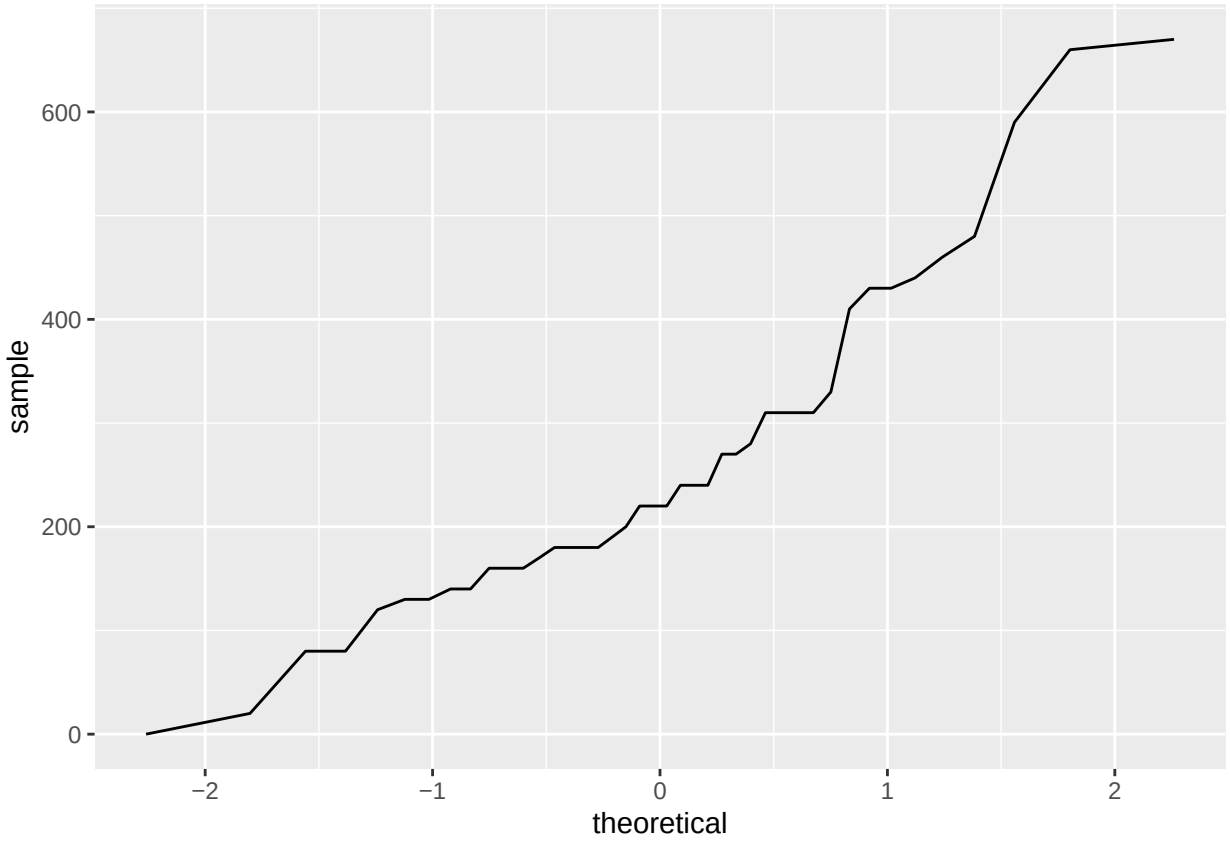
```
ggplot(data = dairy_queen, aes(x = cal_fat)) +
        geom_blank() +
        geom_histogram(aes(y = ..density..)) +
        stat_function(fun = dnorm, args = c(mean = dqmean, sd = dqsd), col = "tomato")
```



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

The data does appear to follow a near normal distribution with fatter tails at the end. We see that the majority of its frequencies near the mean which resembles a bell-shaped curve. Perhaps, a student-t distribution would better fit the data above due the fatter tails ends.

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

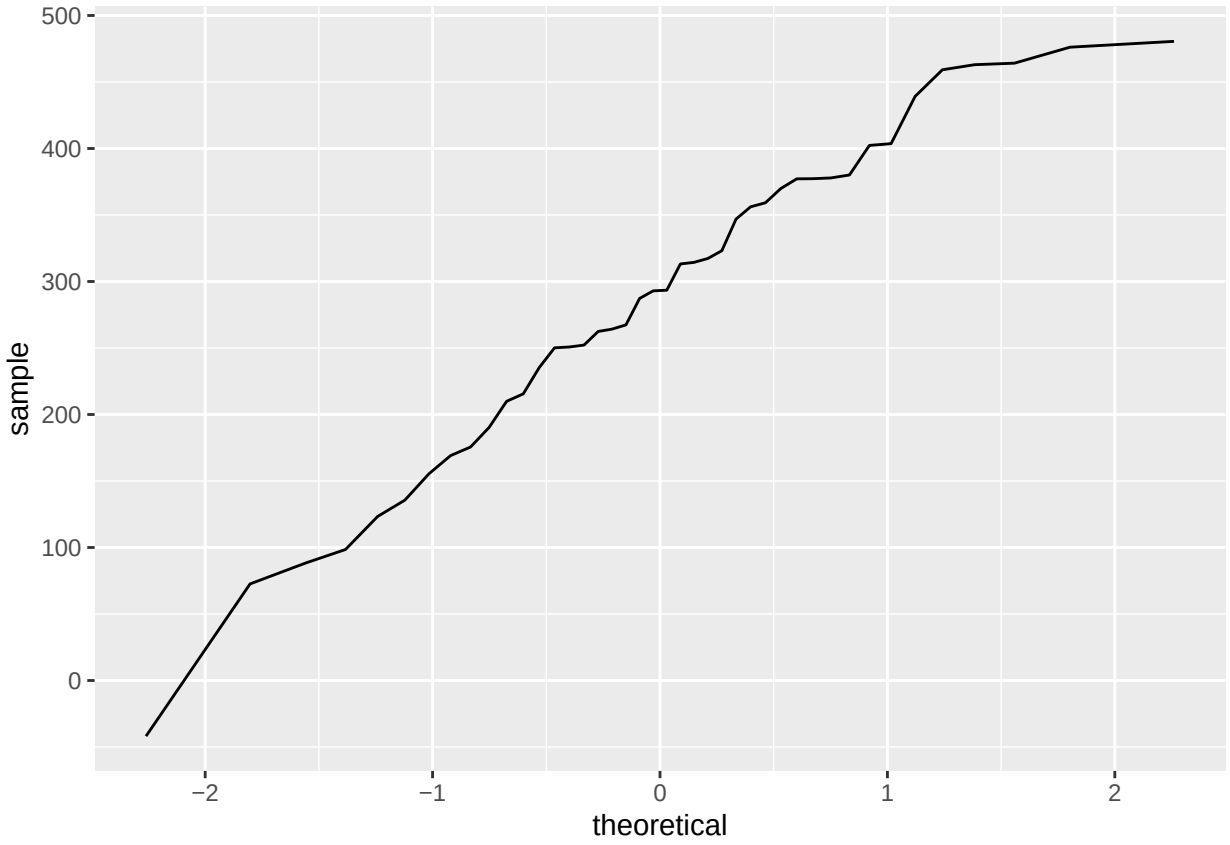


3. Make a normal probability plot of `sim_norm`. Do all of the points fall on the line? How does this plot compare to the probability plot for the real data? (Since `sim_norm` is not a data frame, it can be put directly into the `sample` argument and the `data` argument can be dropped.)

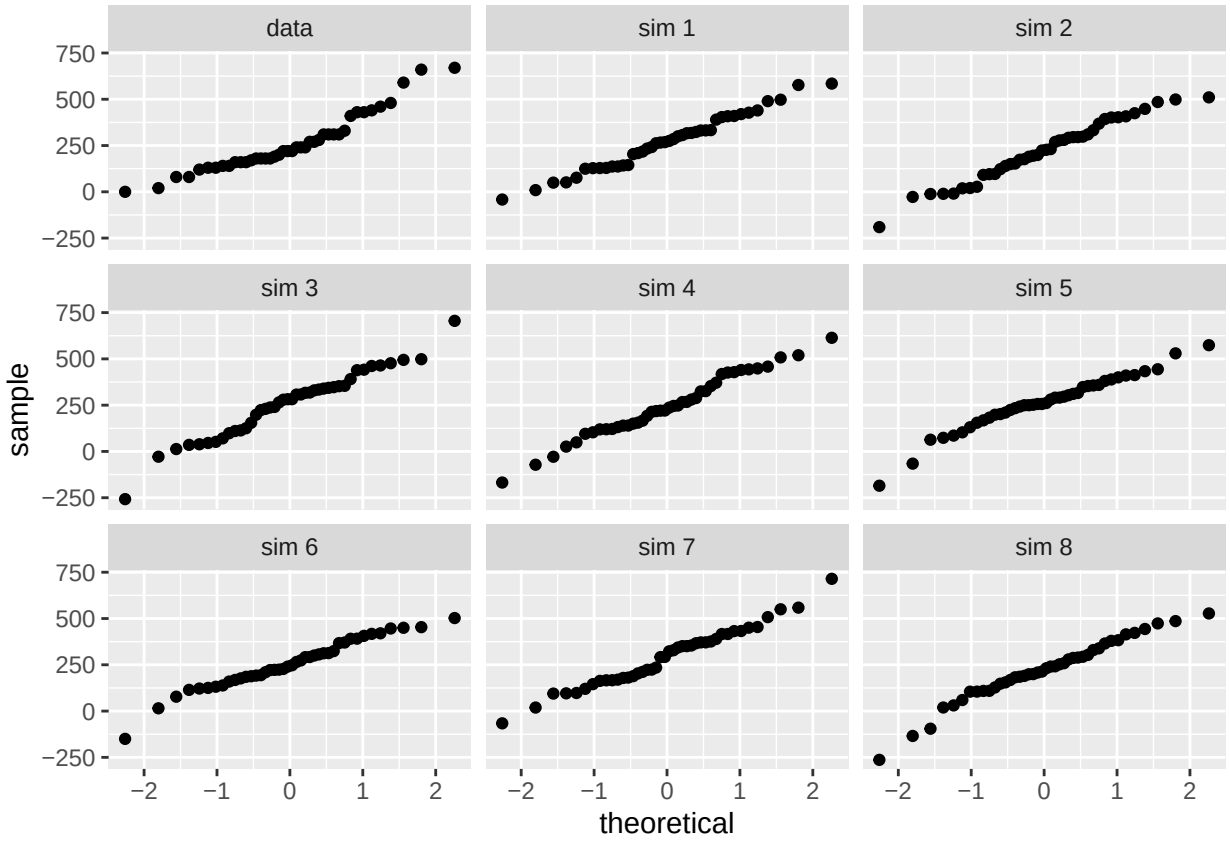
Comparing the real data plot versus the probability plot. We observed that both does look similar but at closer inspection we can see that there are differences especially at comparing intervals in the graphs that the slopes of each graphs are different. Moreover, the range in the real data plot spreads from 0 -700 while the probability plot -100 to 500.

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

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




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



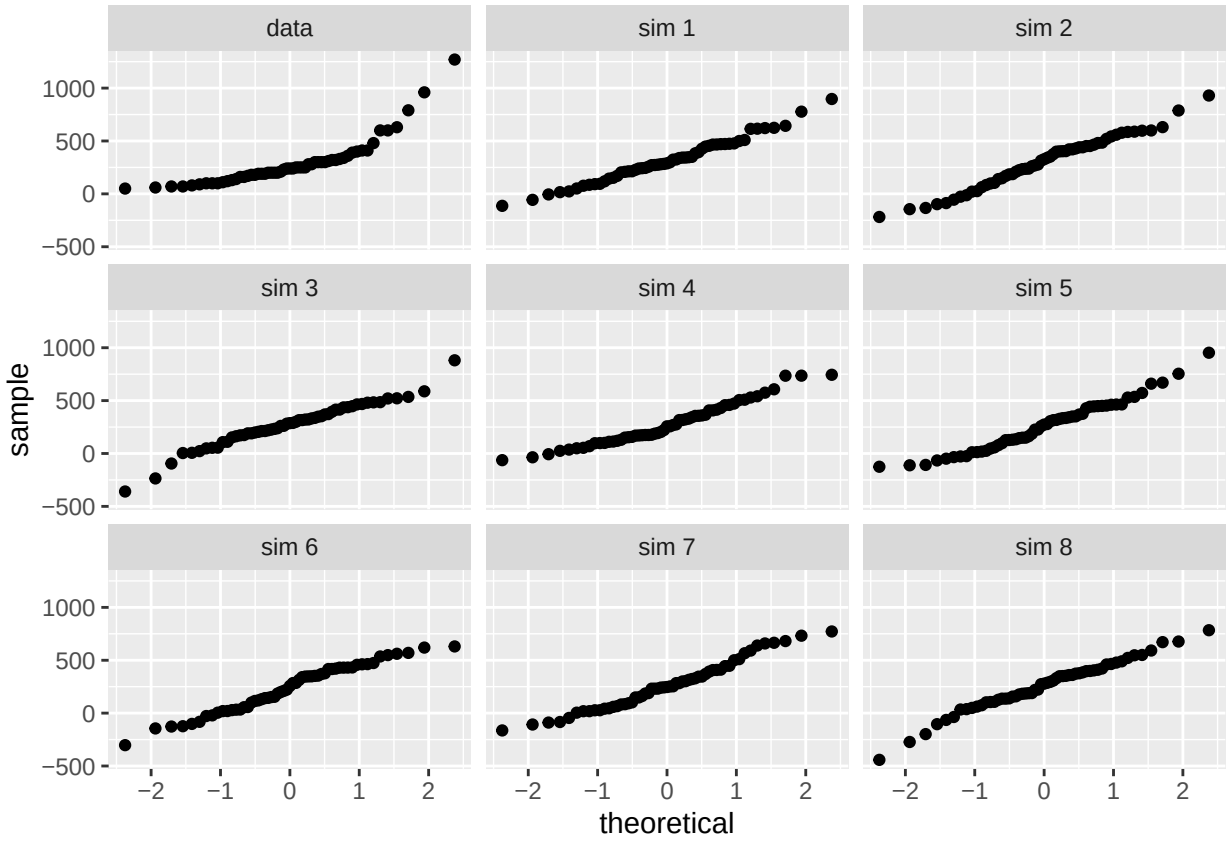
4. Does the normal probability plot for the calories from fat look similar to the plots created for the simulated data? That is, do the plots provide evidence that the calories are nearly normal?

Yes the plots look similar, the differences between is probability due to randomness. It does suggests that the calories nearly follow a normal distribution.

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

It appears that real data graph is skewed and does not follow as closely to the normal distribution relative to Dairy Queen's data. Although, the simulations are looks more normal than the actual data set.

```
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 any given item is greater than 400 calories from fat in Dairy Queen ? What is P(X > 400) ?

```
1 - pnorm(400, mean = dqmean, sd = dqsd)
```

```
## [1] 0.1863007
```

```
mean(dairy_queen$cal_fat > 400)
```

```
## [1] 0.2142857
```

Question 2: What is the probability for calories from fat in Dairy Queens that is between the values of P(50 < X < 200) ?

```
pnorm(200, mean = dqmean, sd = dqsd) - pnorm(50, mean = dqmean, dqsd)
```

```
## [1] 0.260267
```

```r
mean(dairy_queen$cal_fat > 50 & dairy_queen$cal_fat < 200)
```

```
## [1] 0.3809524
```

The difference is probabilities suggests that the actual data may not be as close to the normal distribution as initially thought especially on the interval of $50 < X < 200$.

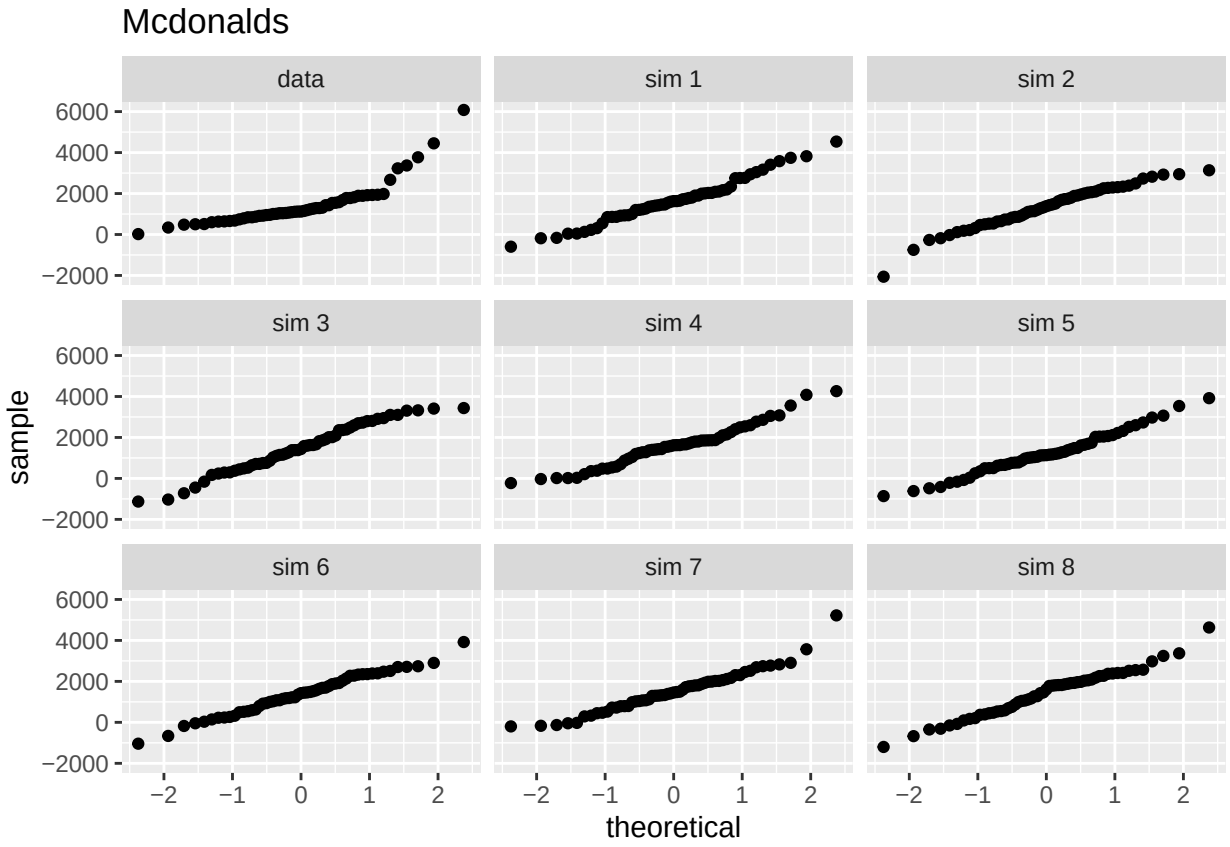
More Practice

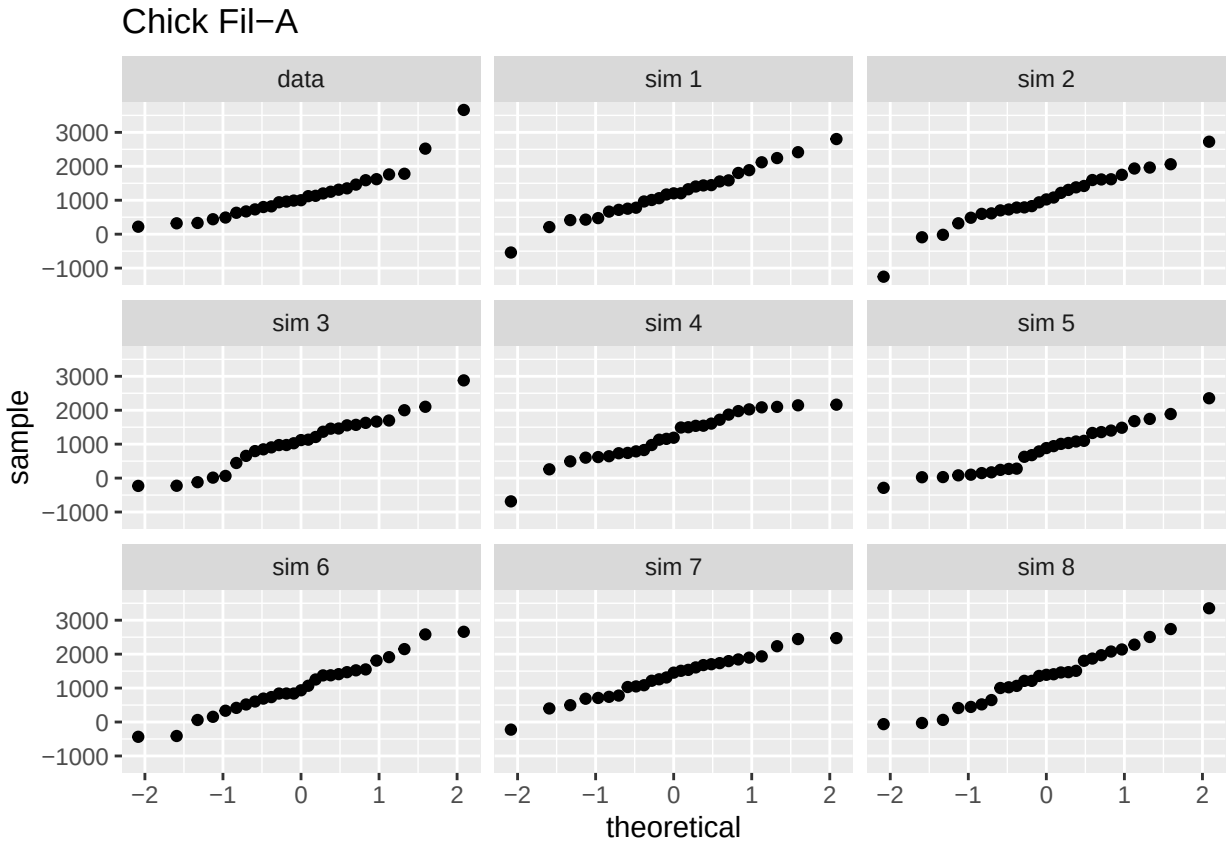
7. Now let's consider some of the other variables in the dataset. Out of all the different restaurants, which ones' distribution is the closest to normal for sodium?

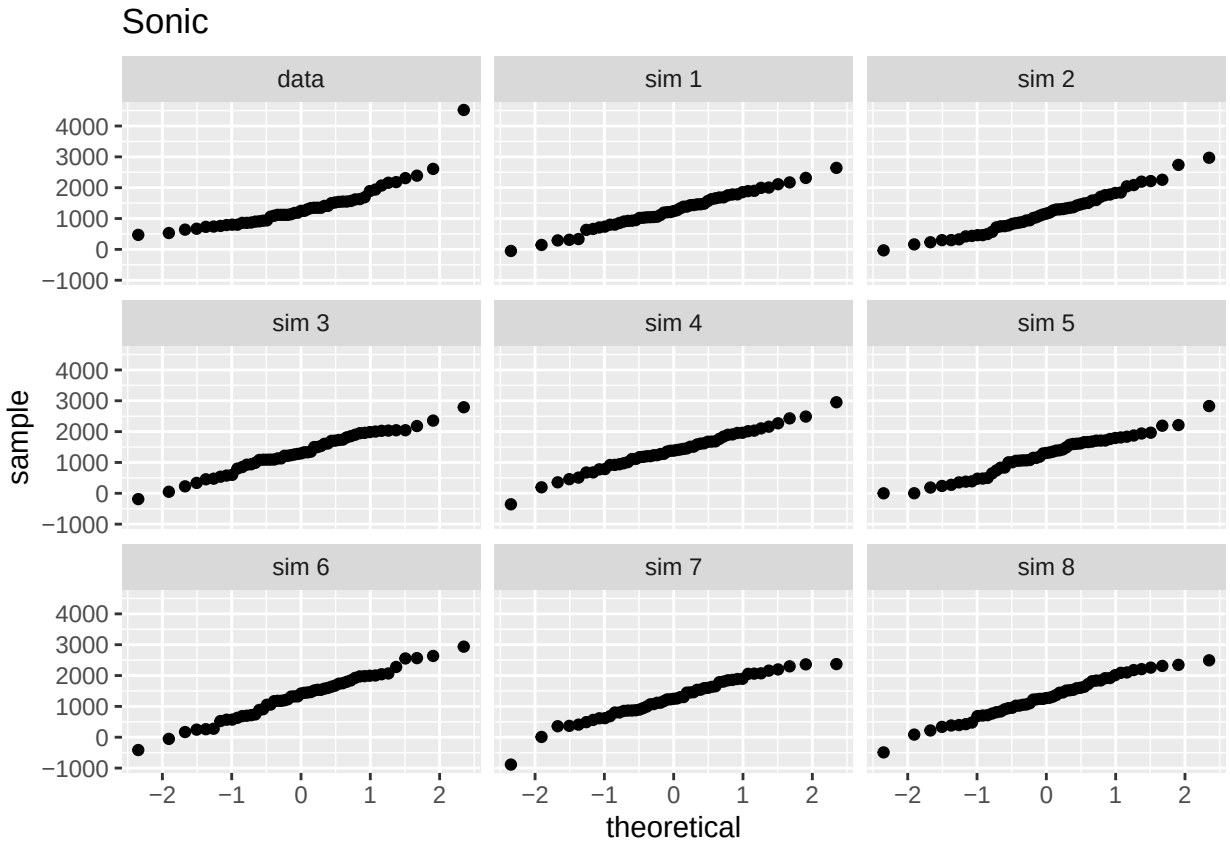
```r
# function that creates subset data for each restaurant
filter_restaurant <- function(restaurants, fastfood) {
  restaurant_data <- list()
  for (r in restaurants) {
    restaurant_data[[r]] <- subset(fastfood, restaurant == r)
  }
  return(restaurant_data)
}
```

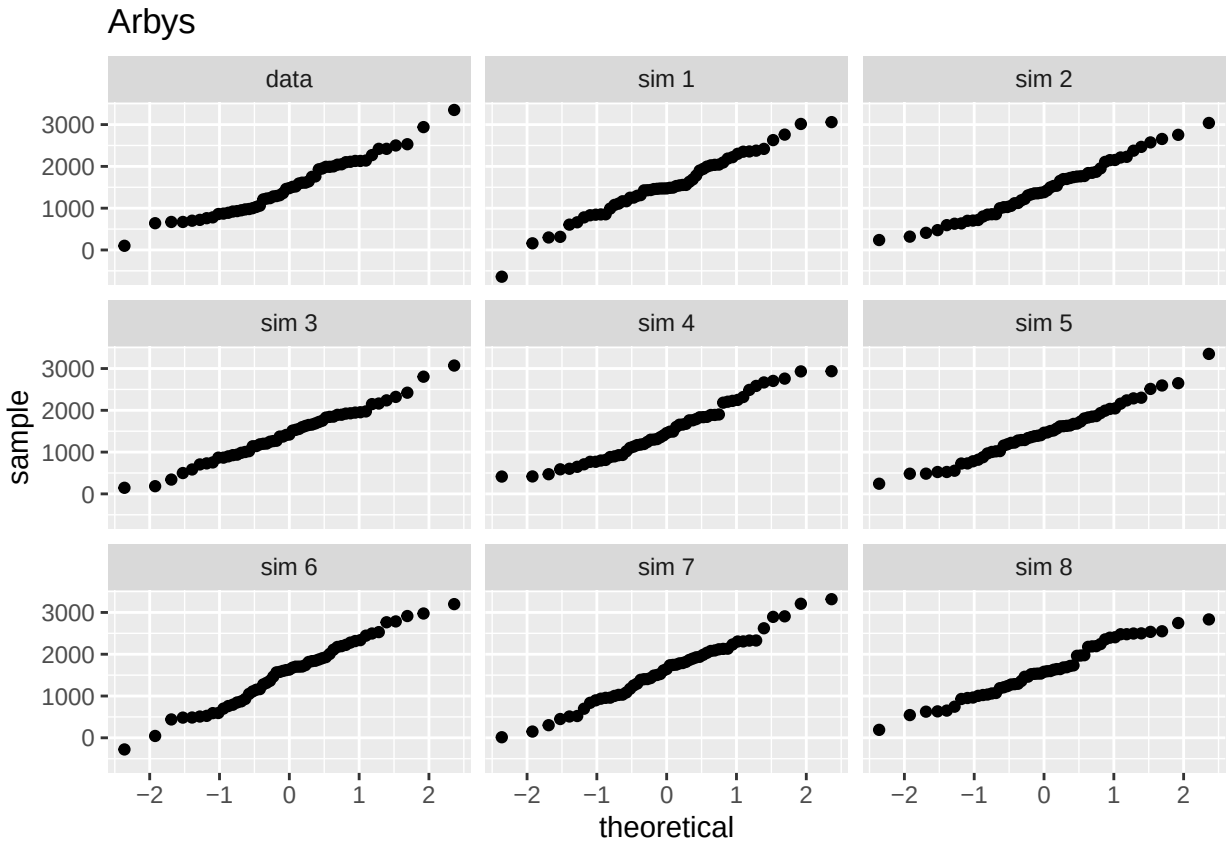
```r
restaurants <- as.character(unique(fastfood$restaurant))
subset_data <- filter_restaurant(restaurants, fastfood)
```

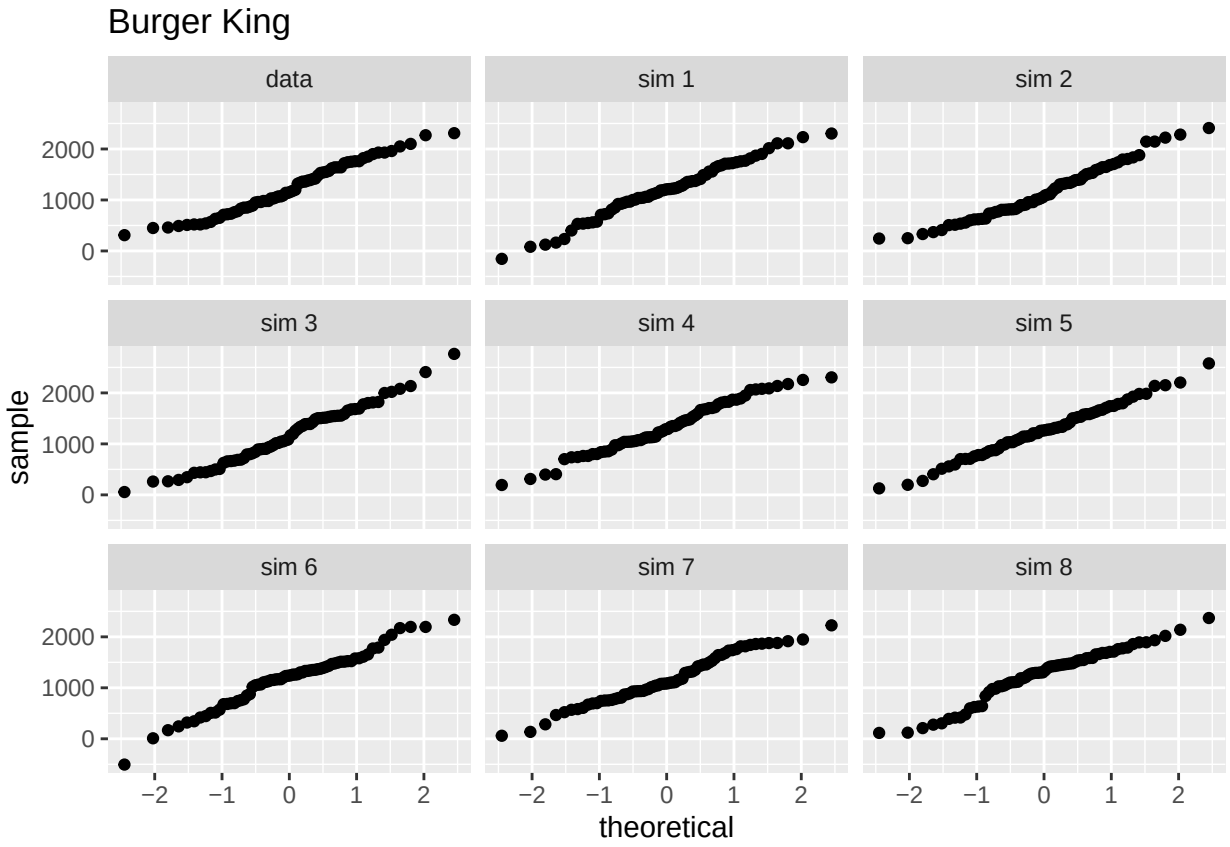
```r
#loops through the subset_data and generate the qq norm simulations
for (r in names(subset_data)) {
    graph <- qqnormsim(sample = sodium, data = subset_data[[r]])
    print(graph + ggtitle(r))
}
```

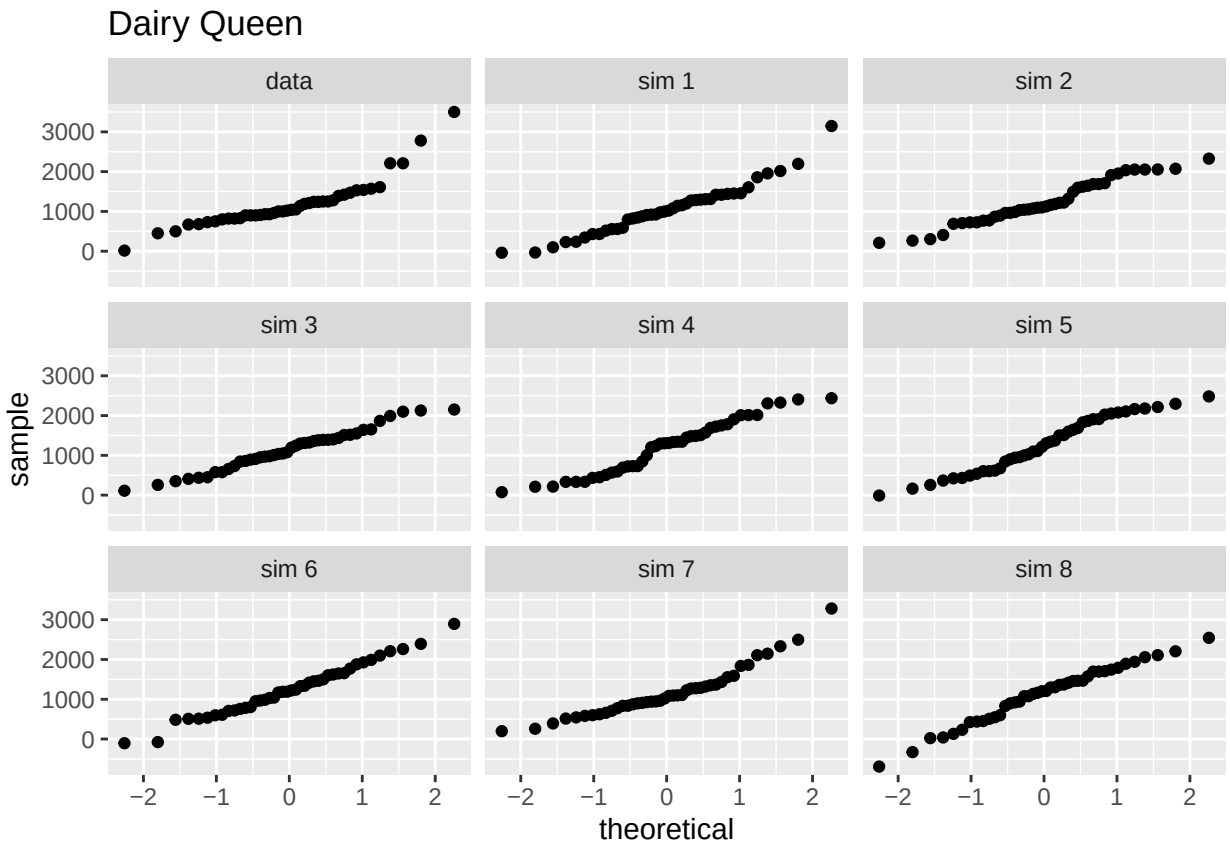

Mcdonalds
data
sim 1
sim 2
sim 3
sim 4
sim 5
sim 6
sim 7
sim 8
sample
theoretical
6000
4000
2000
0
−2000
−2
−1
0
1
2

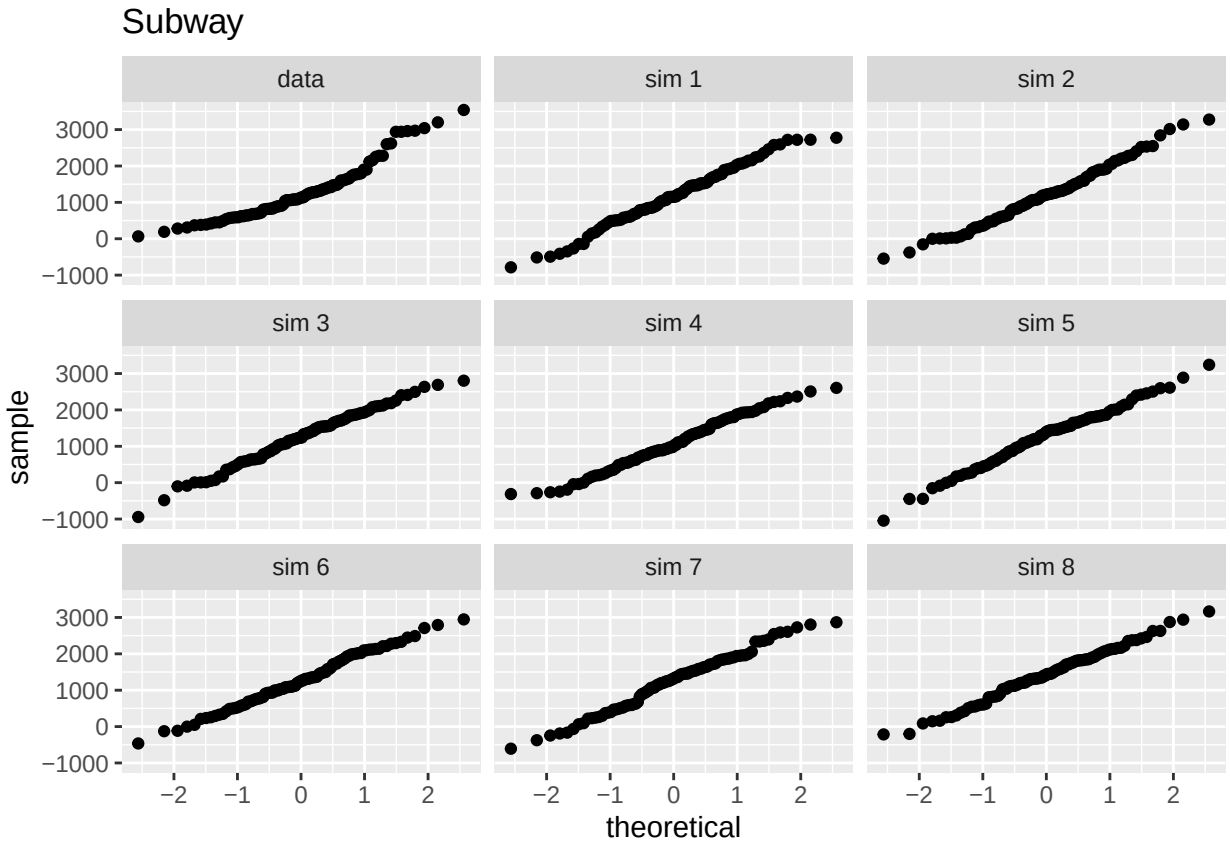

Chick Fil−A
data
sim 1
sim 2
sim 3
sim 4
sim 5
sim 6
sim 7
sim 8
sample
theoretical
3000
2000
1000
0
−1000
−2
−1
0
1
2

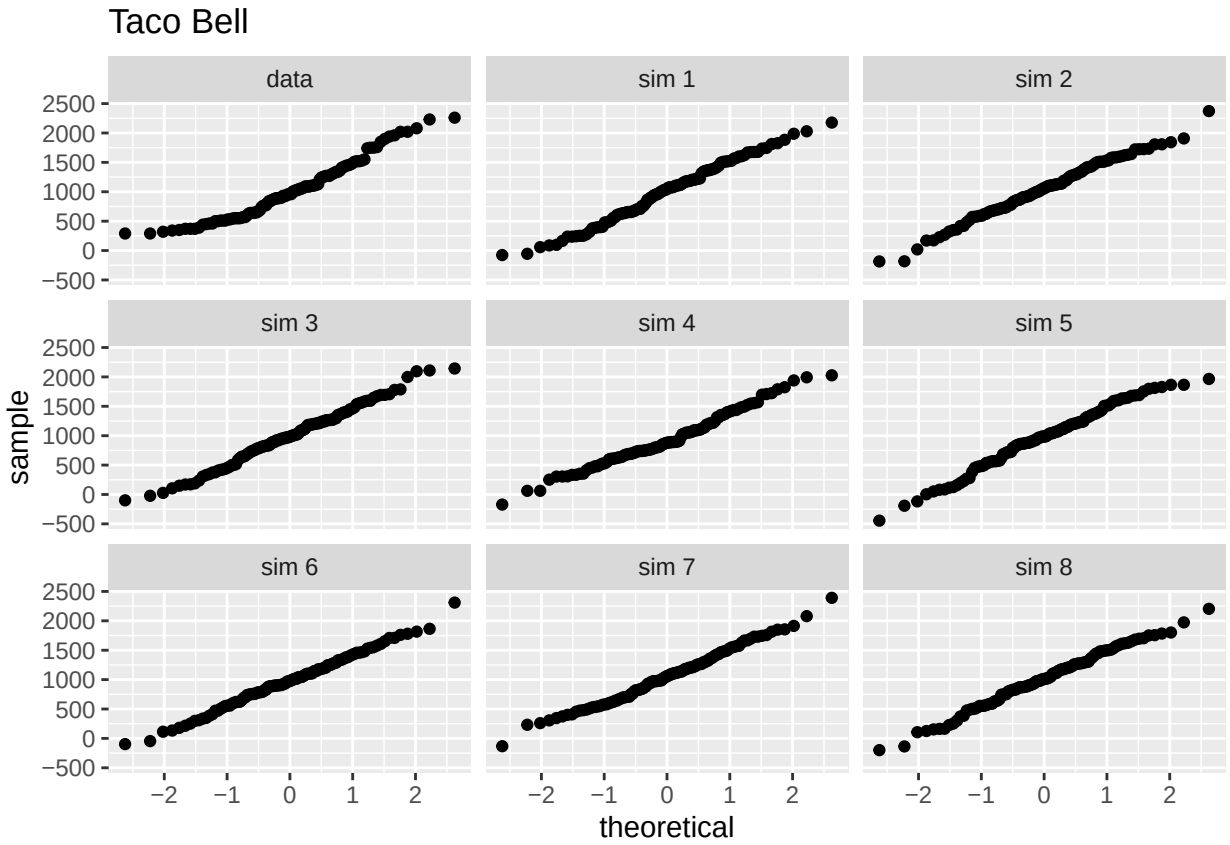

Sonic
data
sim 1
sim 2
sim 3
sim 4
sim 5
sim 6
sim 7
sim 8
sample
theoretical
4000
3000
2000
1000
0
−1000
−2
−1
0
1
2


Arbys
data
sim 1
sim 2
sim 3
sim 4
sim 5
sim 6
sim 7
sim 8
sample
theoretical
3000
2000
1000
0
−2
−1
0
1
2


Burger King
data
sim 1
sim 2
sim 3
sim 4
sim 5
sim 6
sim 7
sim 8
sample
theoretical
2000
1000
0
−2
−1
0
1
2


Dairy Queen
data
sim 1
sim 2
sim 3
sim 4
sim 5
sim 6
sim 7
sim 8
sample
theoretical
3000
2000
1000
0
−2
−1
0
1
2


Subway
data
sim 1
sim 2
sim 3
sim 4
sim 5
sim 6
sim 7
sim 8
sample
theoretical
3000
2000
1000
0
−1000
−2
−1
0
1
2



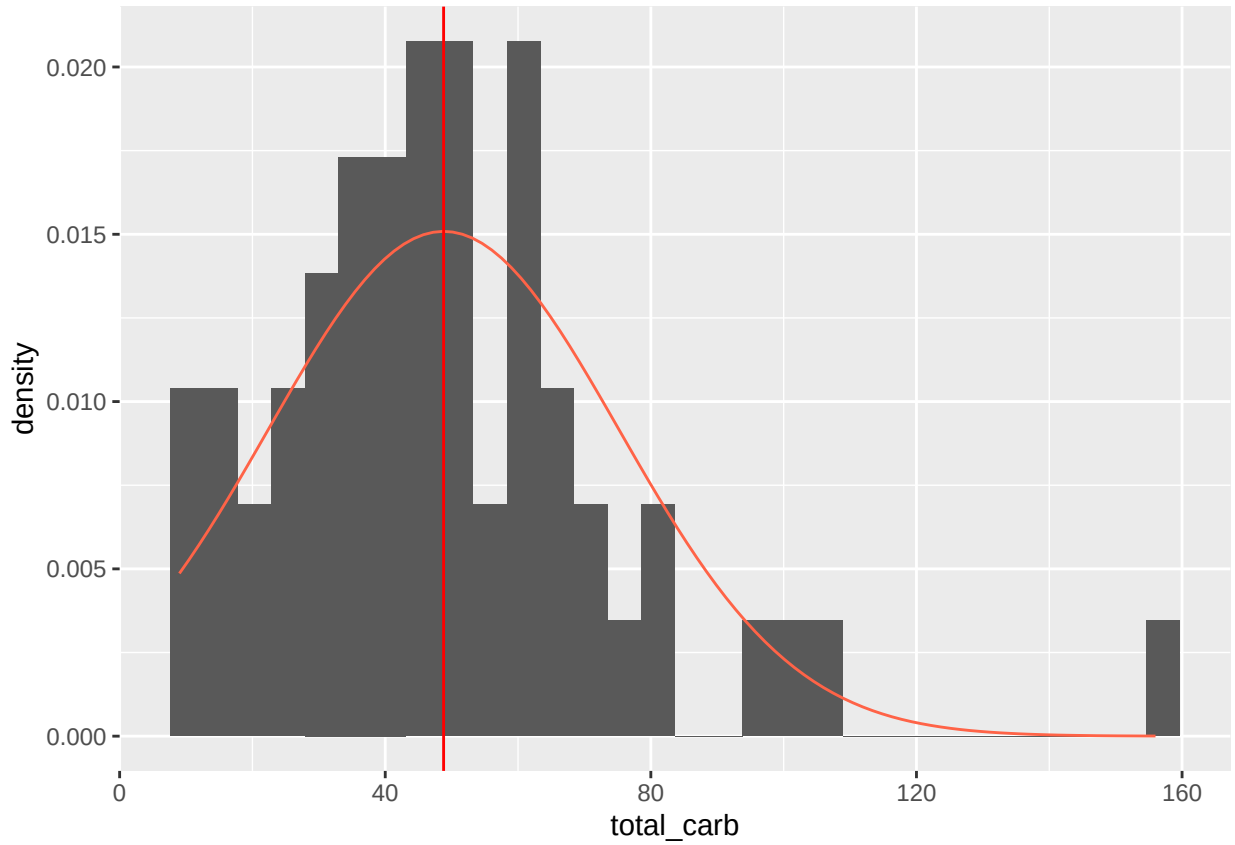
Based on the graphs above, Burger King, Taco Bell and Subway have the closest ditribution to a normal distrubition on their sodium count.

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?

Looking at the data, we can notice that the data for sodium is discrete. Perhaps, some the step size are small enough to approximate the normal distribution.

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 = mcdonalds, aes(x = total_carb)) +
        geom_blank() +
        geom_histogram(aes(y = ..density..)) +
        stat_function(fun = dnorm, args = c(mean = mean(mcdonalds$total_carb), sd = sd(mcdonalds$total_
    geom_vline(aes(xintercept = mean(mcdonalds$total_carb)), color = "red")
```



As we can observe, the data of total carbohydrates for Dairy Queen in left-skewed and the graph shows that most frequencies occur around the mean with a few outliers to the right. * * *