Code ▾

DS606 - Lab 4

George Cruz

2020-10-19

Hide

```
library(tidyverse)
library(openintro)
knitr::kable(head(fastfood))
```

restaurant	item	calories	cal_fat	total_fat	sat_fat	trans_fat	cholesterol	sodium	total_carb	fiber	sugar	protein	vit_a	vit_c	calcium	sal
Mcdonalds	Artisan Grilled Chicken Sandwich	380	60	7	2	0.0	95	1110	44	3	11	37	4	20	20	Oth
Mcdonalds	Single Bacon Smokehouse Burger	840	410	45	17	1.5	130	1580	62	2	18	46	6	20	20	Oth
Mcdonalds	Double Bacon Smokehouse Burger	1130	600	67	27	3.0	220	1920	63	3	18	70	10	20	50	Oth
Mcdonalds	Grilled Bacon Smokehouse Chicken Sandwich	750	280	31	10	0.5	155	1940	62	2	18	55	6	25	20	Oth
Mcdonalds	Crispy Bacon Smokehouse Chicken Sandwich	920	410	45	12	0.5	120	1980	81	4	18	46	6	20	20	Oth
Mcdonalds	Big Mac	540	250	28	10	1.0	80	950	46	3	9	25	10	2	15	Oth

Hide

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

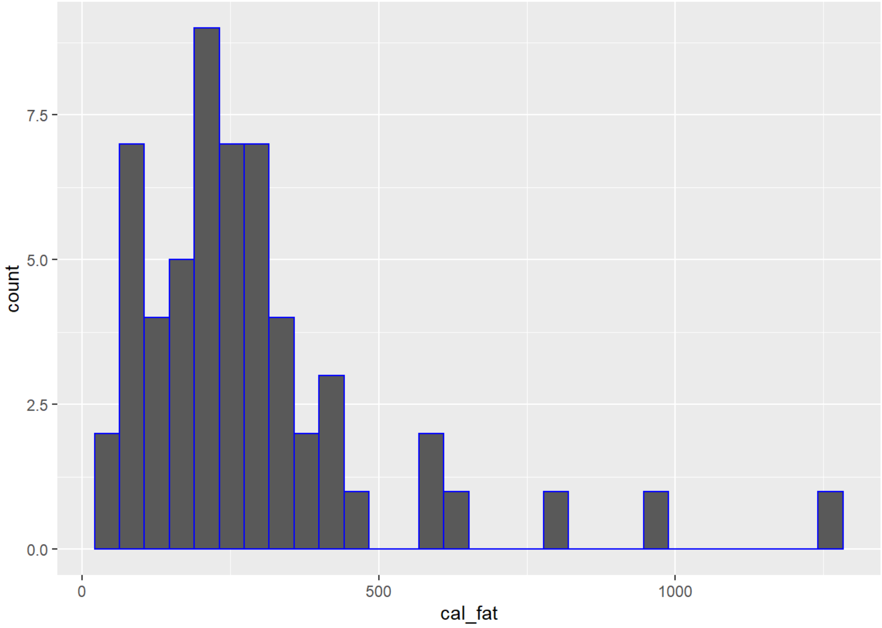
Exercise 1

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?

Hide

```
ggplot(data = mcdonalds, aes(x = cal_fat)) +
        geom_blank() +
        geom_histogram(color="blue")
```

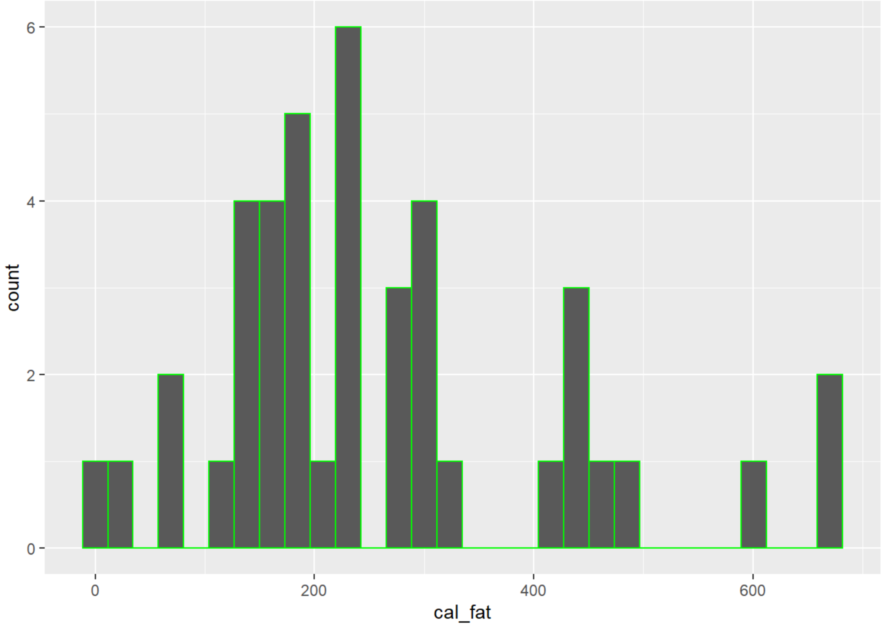
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Hide

```
ggplot(data = dairy_queen, aes(x = cal_fat)) +
        geom_blank() +
        geom_histogram(color="green")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Both plots are unimodal, the Dairy Queen histogram is slightly right-skewed whereas the McDonalds is right-skewed. The Dairy Queen one appears almost normal.

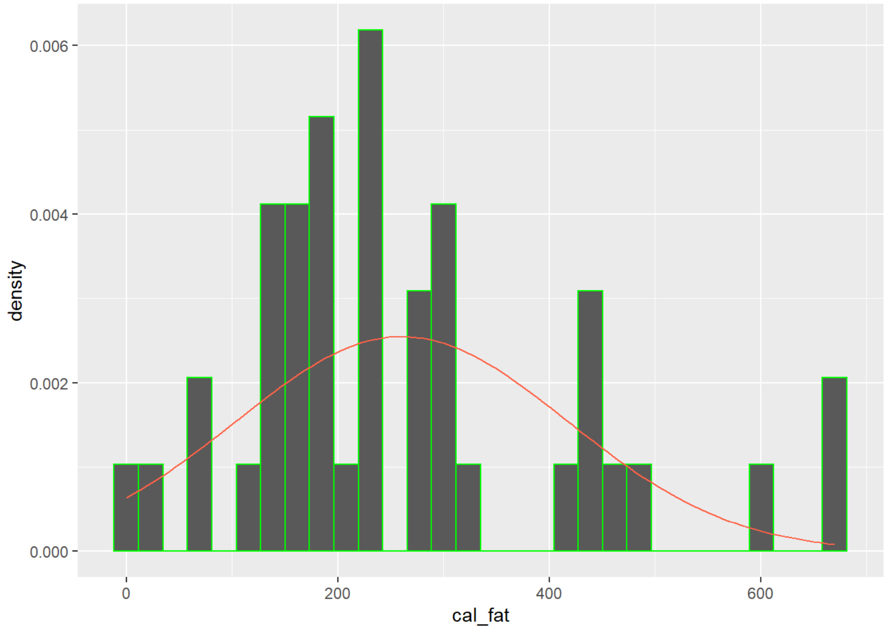
Exercise 2

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

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

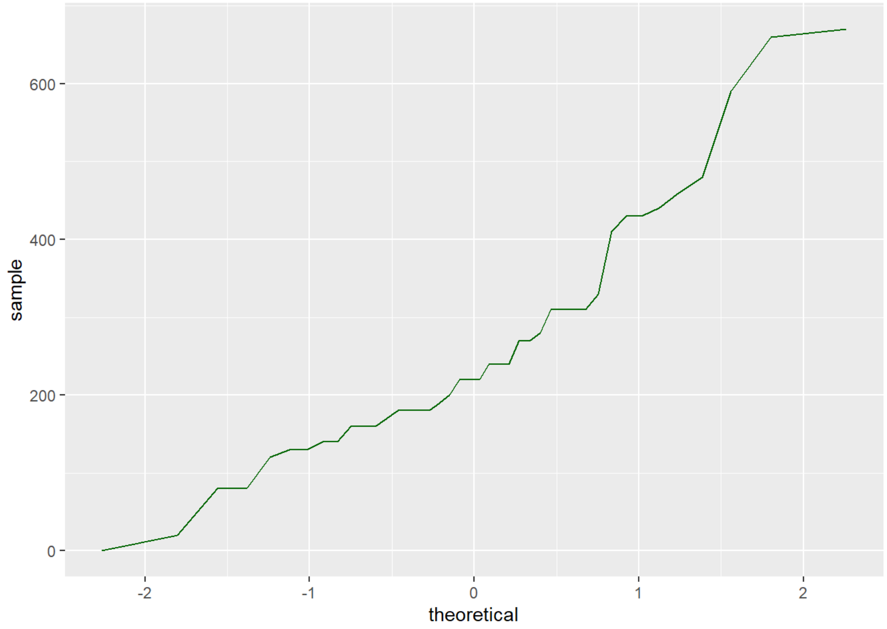
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Based on this plot, we can say that this distribution is nearly normal.

Hide

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



Hide

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

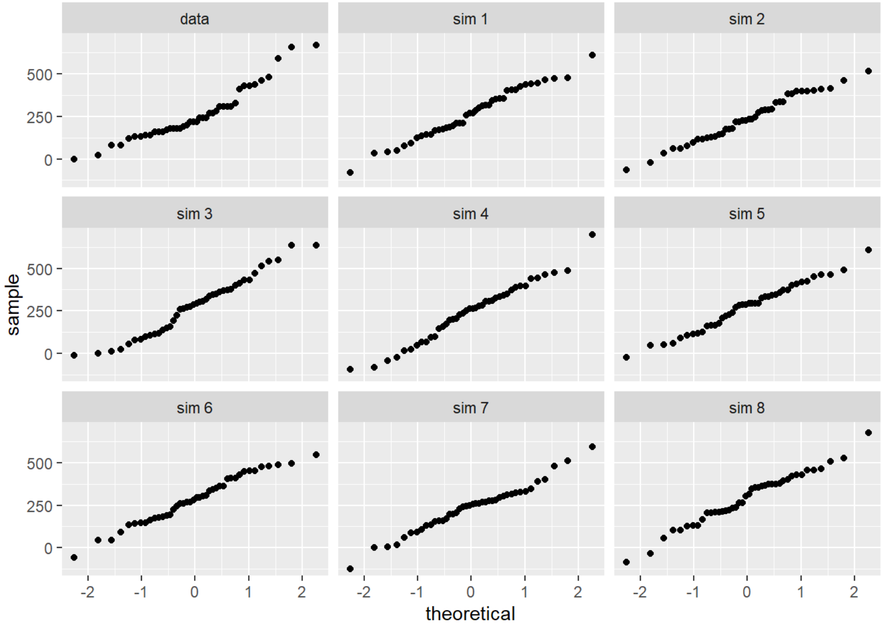
Exercise 3

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

Looks very similar

Hide

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



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 distribution is nearly normal?

Yes

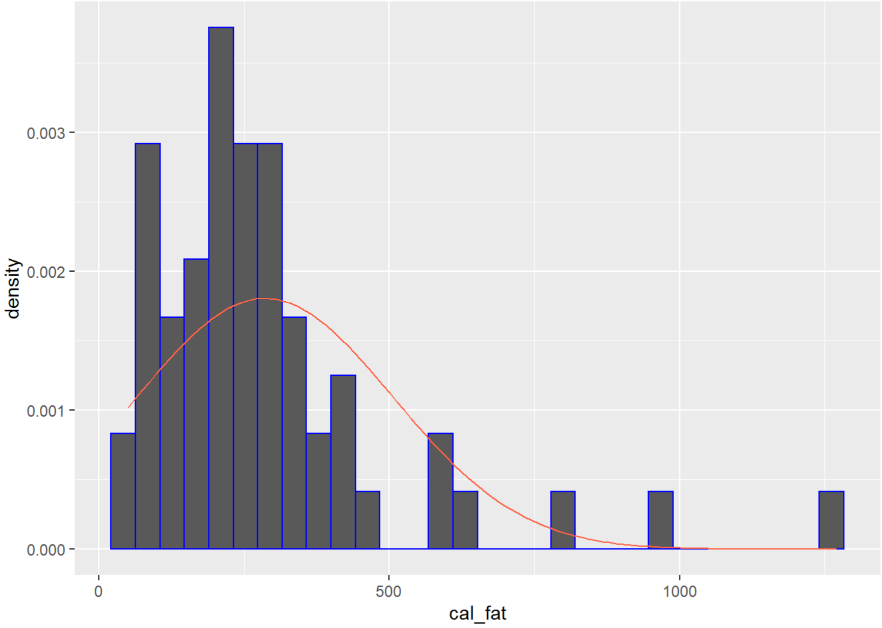
Exercise 5

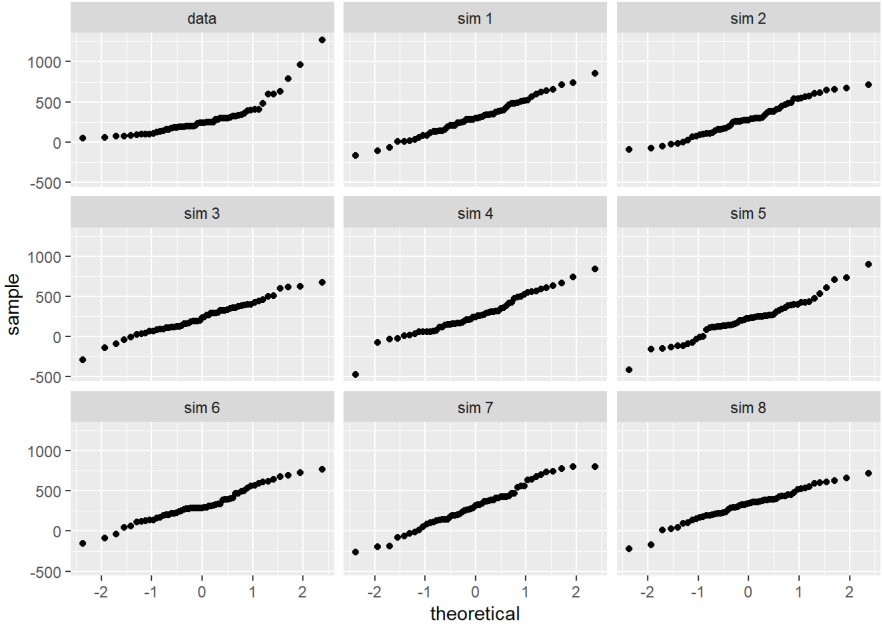
Using the same technique, determine whether or not the calories from McDonald's menu appear to come from a normal distribution.

Hide

```
mcmean <- mean(mcdonalds$cal_fat)
mcsd   <- sd(mcdonalds$cal_fat)

ggplot(data = mcdonalds, aes(x = cal_fat)) +
        geom_blank() +
        geom_histogram(aes(y = ..density..), color="blue") +
        stat_function(fun = dnorm, args = c(mean = mcmean, sd = mcsd), col = "tomato")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```





Line does not seem to follow the same slope, does not appear normal.

Hide

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

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

Hide

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

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

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?

a) What is the percentage of McDonald’s menu that is over 600 calories from fat?

Hide

```
1 - pnorm(q = 600, mean = mcmean, sd = mcsd)
```

```
## [1] 0.07733771
```

8% of McDonald’s menu is over 600 calories from fat.

Hide

```
sum(mcdonalds$cal_fat > 600) / length(mcdonalds$cal_fat)
```

```
## [1] 0.07017544
```

7%

b) What is the pecentage of McDonald’s menu that is over 600 calories?

Hide

```
mcmean2 <- mean(mcdonalds$calories)
mcsd2   <- sd(mcdonalds$calories)

1 - pnorm(q = 600, mean = mcmean2, sd = mcsd2)
```

```
## [1] 0.5391331
```

54% of the theoretical normal.

Hide

```
sum(mcdonalds$calories > 600) / length(mcdonalds$calories)
```

```
## [1] 0.4210526
```

42% of the real data.

Calories from fat are closer than overall calories.

Exercise 7

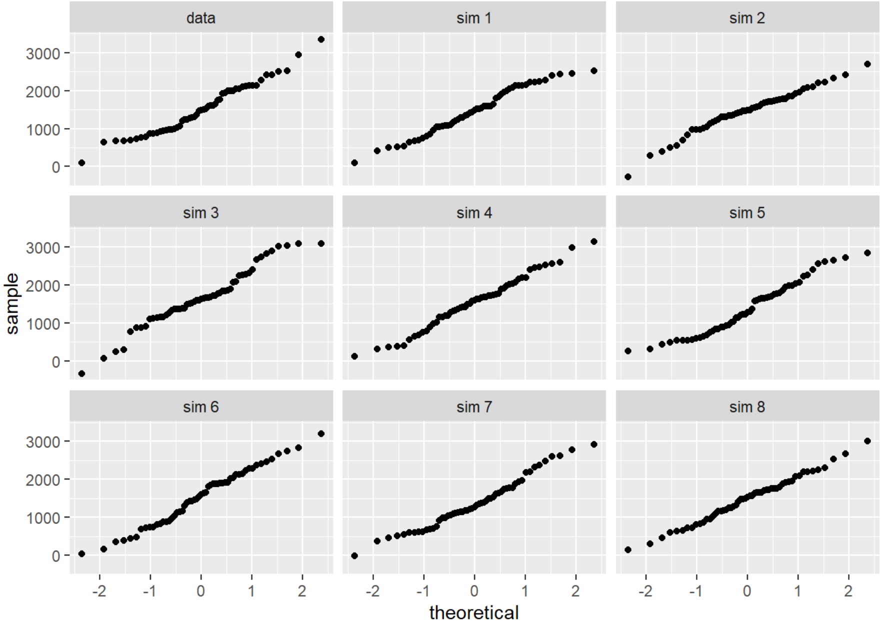
Now let’s consider some of the other variables in the dataset. Out of all the different restaurants, which ones’ distribution is the closest to normal for sodium?

Hide

```
restaurants <- fastfood %>%
  group_by(restaurant) %>%
  select(restaurant, sodium) %>%
  summarise( avg = mean(sodium), sd_sodium= sd(sodium), .groups = "drop_last")
```

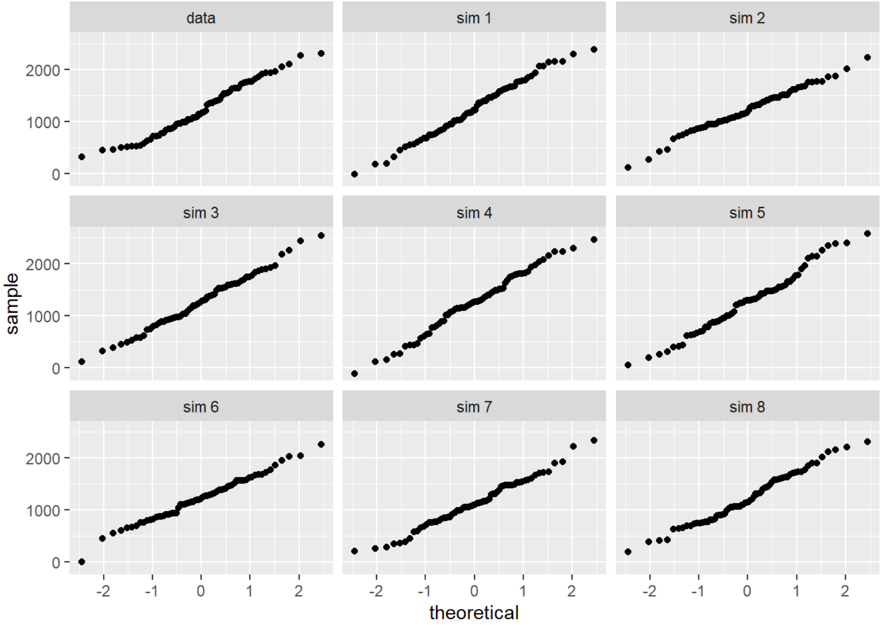
Hide

```
datar <- fastfood%>%
    filter(restaurant == "Arbys")
dmean = restaurants$avg[1]
dsd = restaurants$sd_sodium[1]

qqnormsim(sample = sodium, data = datar)
```



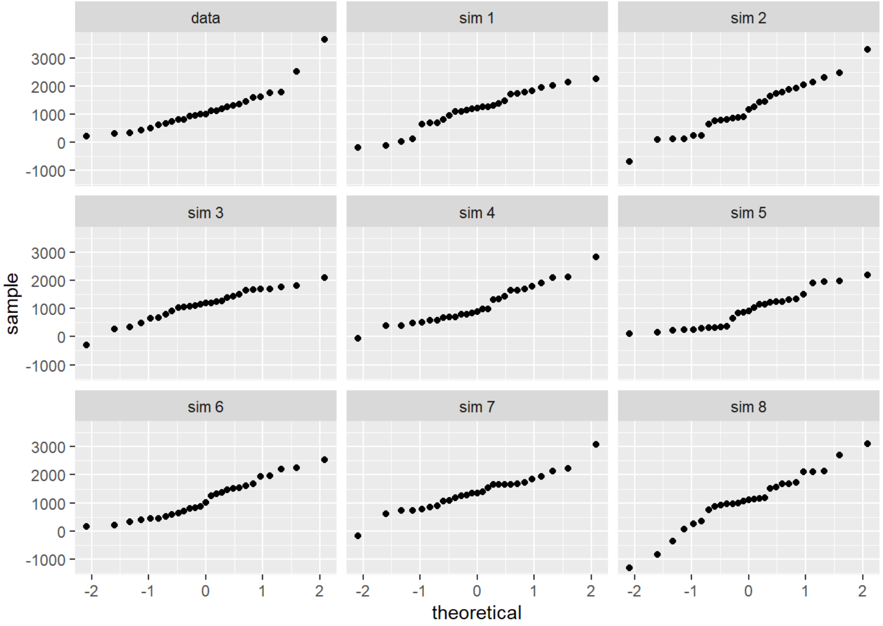
Hide

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

dmean = restaurants$avg[2]
dsd = restaurants$sd_sodium[2]

qqnormsim(sample = sodium, data = datar)
```



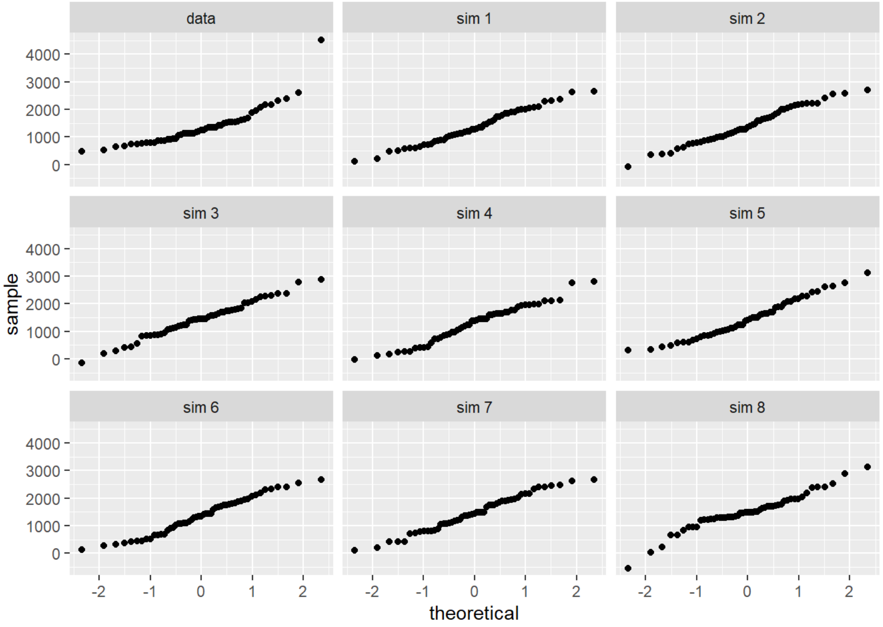
Hide

```
datar <- fastfood%>%
    filter(restaurant == "Chick Fil-A")

qqnormsim(sample = sodium, data = datar)
```



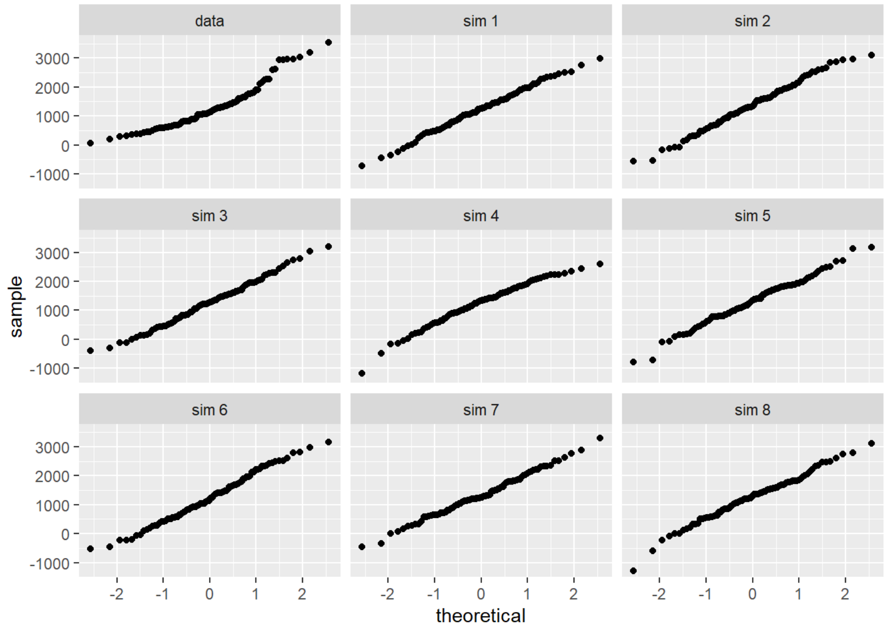
Hide

```
datar <- fastfood%>%
    filter(restaurant == "Sonic")

qqnormsim(sample = sodium, data = datar)
```



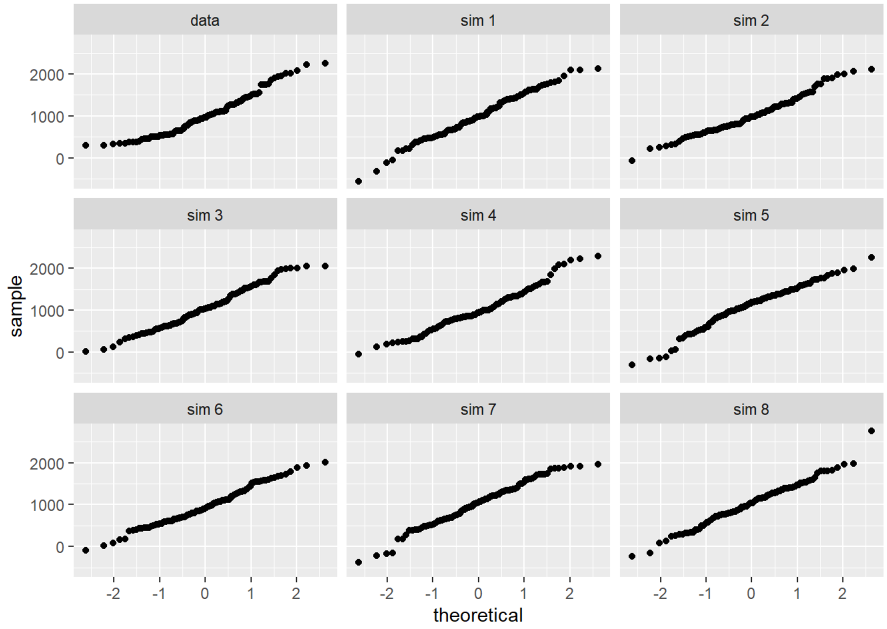
Hide

```
datar <- fastfood%>%
    filter(restaurant == "Subway")

qqnormsim(sample = sodium, data = datar)
```



Hide

```
datar <- fastfood%>%
    filter(restaurant == "Taco Bell")

qqnormsim(sample = sodium, data = datar)
```



I would say Taco Bell

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?

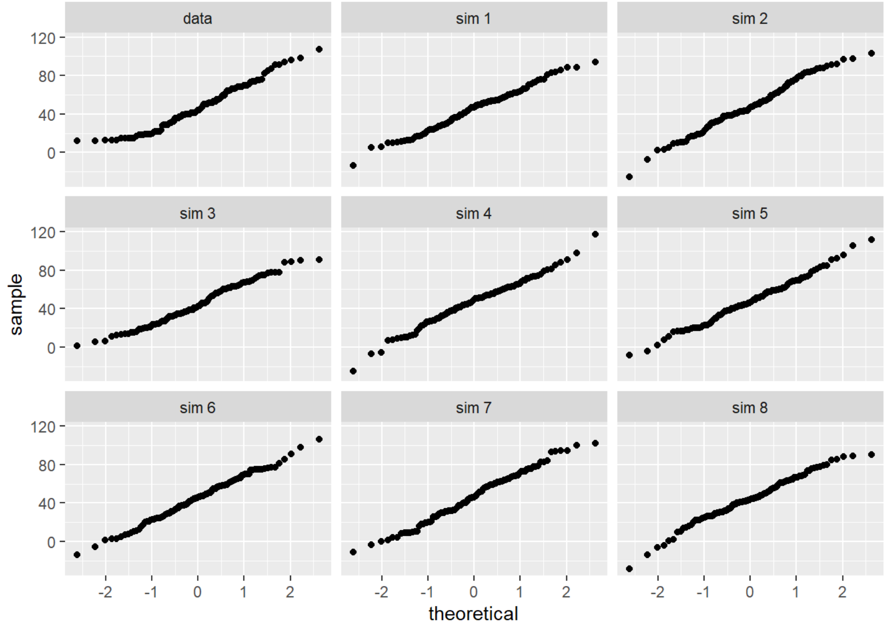
I would say the distributions are right or left-skewed.

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

Hide

```
datar <- fastfood%>%
    filter(restaurant == "Taco Bell")

qqnormsim(sample = total_carb, data = datar)
```

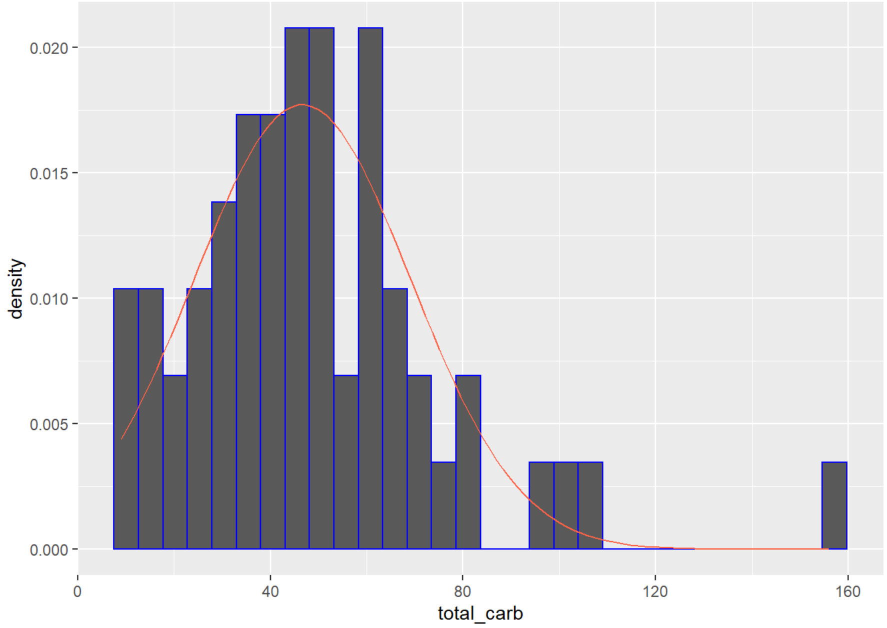


Based on this, I would say the histogram is almost normal and slightly right skewed.

Hide

```
dmean <- mean(datar$total_carb)
dsd   <- sd(datar$total_carb)

ggplot(data = mcdonalds, aes(x = total_carb)) +
        geom_blank() +
        geom_histogram(aes(y = ..density..), color="blue") +
        stat_function(fun = dnorm, args = c(mean = dmean, sd = dsd), col = "tomato")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



It appears the conclusions were correct. …