Lab 4: Normal Distribution

Joseph Annand

2023-06-04

library(tidyverse)  
library(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  
## # ℹ 9 more variables: sodium <dbl>, total\_carb <dbl>, fiber <dbl>, sugar <dbl>,  
## # protein <dbl>, vit\_a <dbl>, vit\_c <dbl>, calcium <dbl>, salad <chr>

mcdonalds <- fastfood %>%  
 filter(restaurant == "Mcdonalds")  
dairy\_queen <- fastfood %>%  
 filter(restaurant == "Dairy Queen")

normal\_plot <- function(food\_data, mean = 0, sd = 1, …) { # X grid for non-standard normal distribution x <- food\_data

# Density function

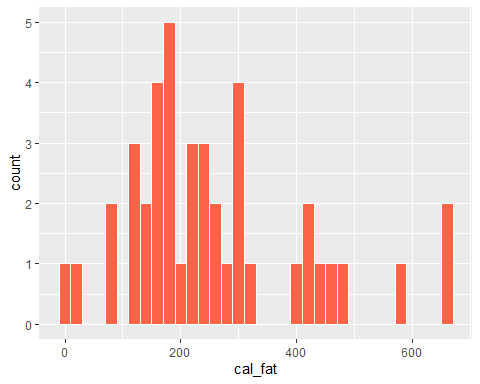
f <- dnorm(x, mean, sd)

plot(x, f, type = “l”, lwd = 2, col = “blue”, ylab = ““, xlab =”cal\_fat”, …) abline(v = mean) # Vertical line on the mean }

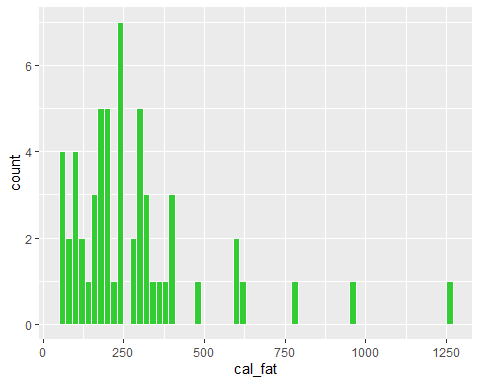
### Exercise 1

The Dairy Queen distribution centers around 200, while the McDonald’s distribution centers around 250. The spread for the McDonald’s distribution is much larger, with a longer right tail that goes to 1250. Both plot show a bell shape in their distributions.

ggplot(dairy\_queen,  
 mapping = aes(x = cal\_fat),  
) + geom\_histogram(binwidth = 20.0, colour = 'white',  
 fill = 'tomato')



ggplot(mcdonalds,  
 mapping = aes(x = cal\_fat),  
) + geom\_histogram(binwidth = 20.0, colour = 'white',  
 fill = 'limegreen')

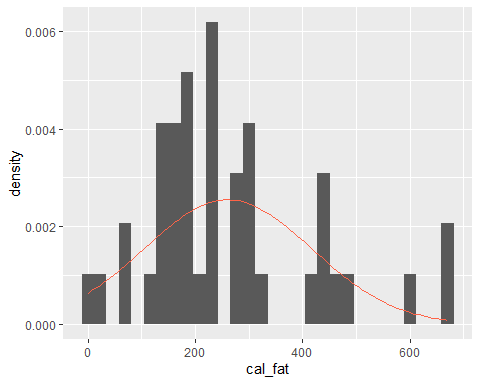


### Exercise 2

Both curves are bell-shaped and unimodal. The Dairy Queen curve is fairly symmetric and appears close to normal. The McDonald’s curve appears to be right-skewed and not very normal.

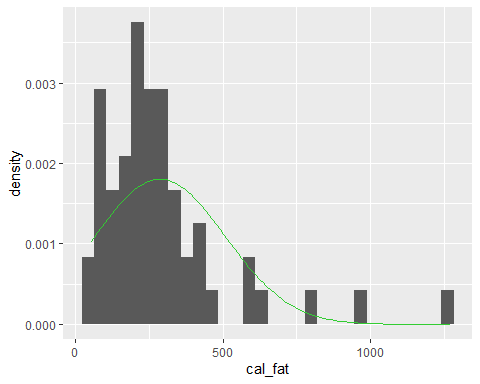
mcmean <- mean(mcdonalds$cal\_fat)  
mcsd <- sd(mcdonalds$cal\_fat)  
  
dqmean <- mean(dairy\_queen$cal\_fat)  
dqsd <- sd(dairy\_queen$cal\_fat)  
  
ggplot(data = dairy\_queen, aes(x = cal\_fat)) +  
 geom\_blank() +  
 geom\_histogram(aes(y = after\_stat(density))) +  
 stat\_function(fun = dnorm, args = c(mean = dqmean, sd = dqsd), col = "tomato")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



ggplot(data = mcdonalds, aes(x = cal\_fat)) +  
 geom\_blank() +  
 geom\_histogram(aes(y = after\_stat(density))) +  
 stat\_function(fun = dnorm, args = c(mean = mcmean, sd = mcsd), col = "limegreen")

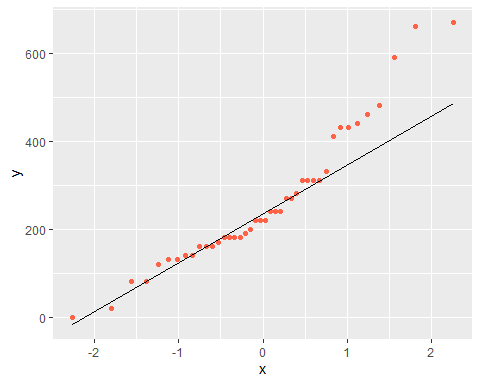
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



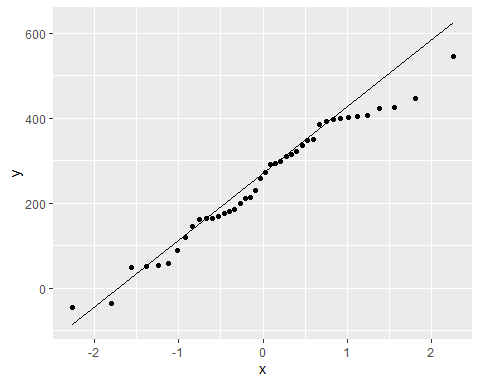
### Exercise 3

All the points in the Q-Q plot of the simulated data do not fall on the line. Compared to probability plot for the real data, the simulated data probability plot has a similar range of points that are on or close to the line around the center. At the higher end of the graph, the points drift above the line. While the points on the real data plot remain near the line on the lower end of the graph, the points on the simulated data plot drift below the line in the same area.

ggplot(data = dairy\_queen, aes(sample = cal\_fat)) +  
 # used stat\_qq instead of geom\_line(stat = "qq")  
 stat\_qq(colour = "tomato") +  
 stat\_qq\_line()



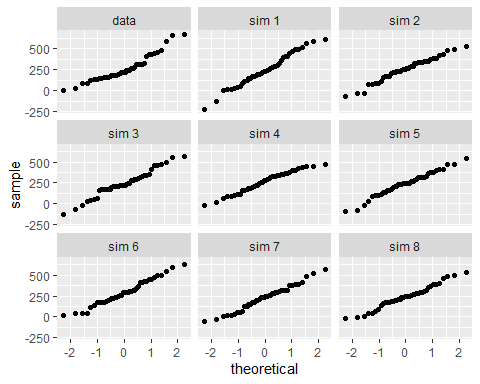
sim\_norm <- rnorm(n = nrow(dairy\_queen), mean = dqmean, sd = dqsd)  
  
ggplot(mapping = aes(sample = sim\_norm)) +  
 stat\_qq() +  
 stat\_qq\_line()



### Exercise 4

The real data plot appears very similar to all of the simulated data plots. The comparisons provide evidence that the real data for Dairy Queen’s calories form fat are nearly normal.

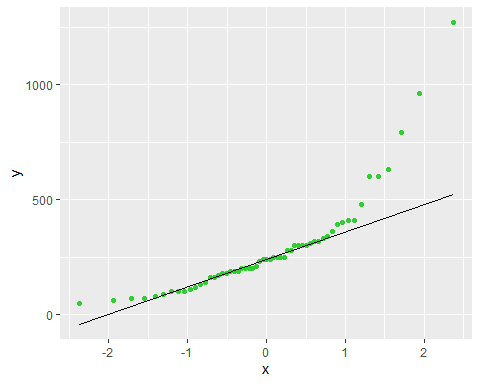
qqnormsim(sample = cal\_fat, data = dairy\_queen)



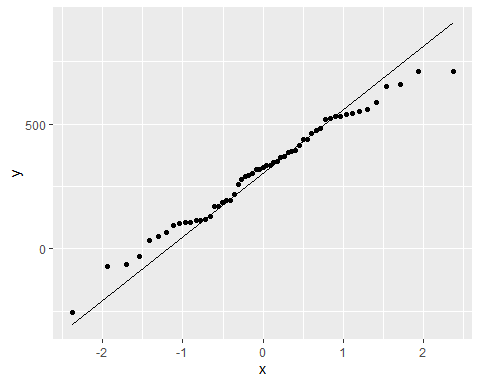
### Exercise 5

The Q-Q plot for the real data differs from that of the simulated data. The skewness of the real data can be seen in the higher end of the graph where several points deviate above the line. This skewness is not seen in the Q-Q plot for the simulated data, as the points in the same area are on or close to the line. The simulated data plots do not show evidence that the real data is nearly normal due to the difference at the higher end of the graph.

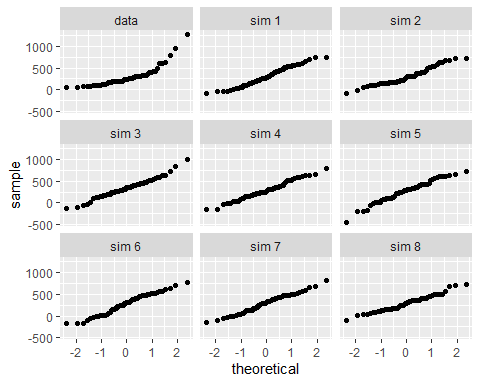
ggplot(data = mcdonalds, aes(sample = cal\_fat)) +  
 stat\_qq(colour = "limegreen") +  
 stat\_qq\_line()



mc\_sim\_norm <- rnorm(n = nrow(mcdonalds), mean = mcmean, sd = mcsd)  
  
ggplot(mapping = aes(sample = mc\_sim\_norm)) +  
 stat\_qq() +  
 stat\_qq\_line()



qqnormsim(sample = cal\_fat, data = mcdonalds)



### Exercise 6

What is the probability of a random Taco Bell product has more than 10 grams of protein?

pnorm method : 0.8507232 empirical method: 0.826087

What is the probability of a random Sonic product has less than 10 grams of sugar?

pnorm method: 0.8105874 empirical method: 0.7924528

The probabilities for a random Sonic product having less than 10 grams of sugar are closer than those for a Taco Bell product having more than 10 grams of protein.

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

## [1] 0.01501523

dairy\_queen %>%  
 filter(cal\_fat > 600) %>%  
 summarise(percent = n() / nrow(dairy\_queen))

## # A tibble: 1 × 1  
## percent  
## <dbl>  
## 1 0.0476

# What is the probability of a random Taco Bell product has more than 10 grams of protein?  
taco\_bell <- fastfood %>%  
 filter(restaurant == "Taco Bell")  
  
1 - pnorm(q = 10, mean = mean(taco\_bell$protein), sd = sd(taco\_bell$protein))

## [1] 0.8507232

taco\_bell %>%  
 filter(protein > 10) %>%  
 summarise(percent = n() / nrow(taco\_bell))

## # A tibble: 1 × 1  
## percent  
## <dbl>  
## 1 0.826

# What is the probability of a random Sonic product has less than 10 grams of sugar?  
sonic <- fastfood %>%  
 filter(restaurant == "Sonic")  
  
pnorm(q = 10, mean = mean(sonic$sugar), sd = sd(sonic$sugar))

## [1] 0.8105874

sonic %>%  
 filter (sugar < 10) %>%  
 summarise(percent = n() / nrow(sonic))

## # A tibble: 1 × 1  
## percent  
## <dbl>  
## 1 0.792