# EDAV Fall 2020 PSet 1

### Junzhi Ge(jg4281), Yifei Zhang(yz3925), Group 17

Read *Graphical Data Analysis with R*, Ch. 3

Grading is based both on your graphs and verbal explanations. Follow all best practices as discussed in class. *Do not expect the assignment questions to spell out precisely how the graphs should be drawn. Sometimes guidance will be provided, but the absense of guidance does not mean that all choices are ok.*

You must use R for this assignment, either base R or tidyverse (https://www.tidyverse.org), which includes **ggplot2**. If you're new to R, I recommend learning **ggplot2** rather than base R graphics. Instructions for getting started in R are available in this chapter (https://edav.info/basics.html) of edav.info.

The datasets in this assignment are from the **alr4** package which can be installed from CRAN:

## 1. lakes

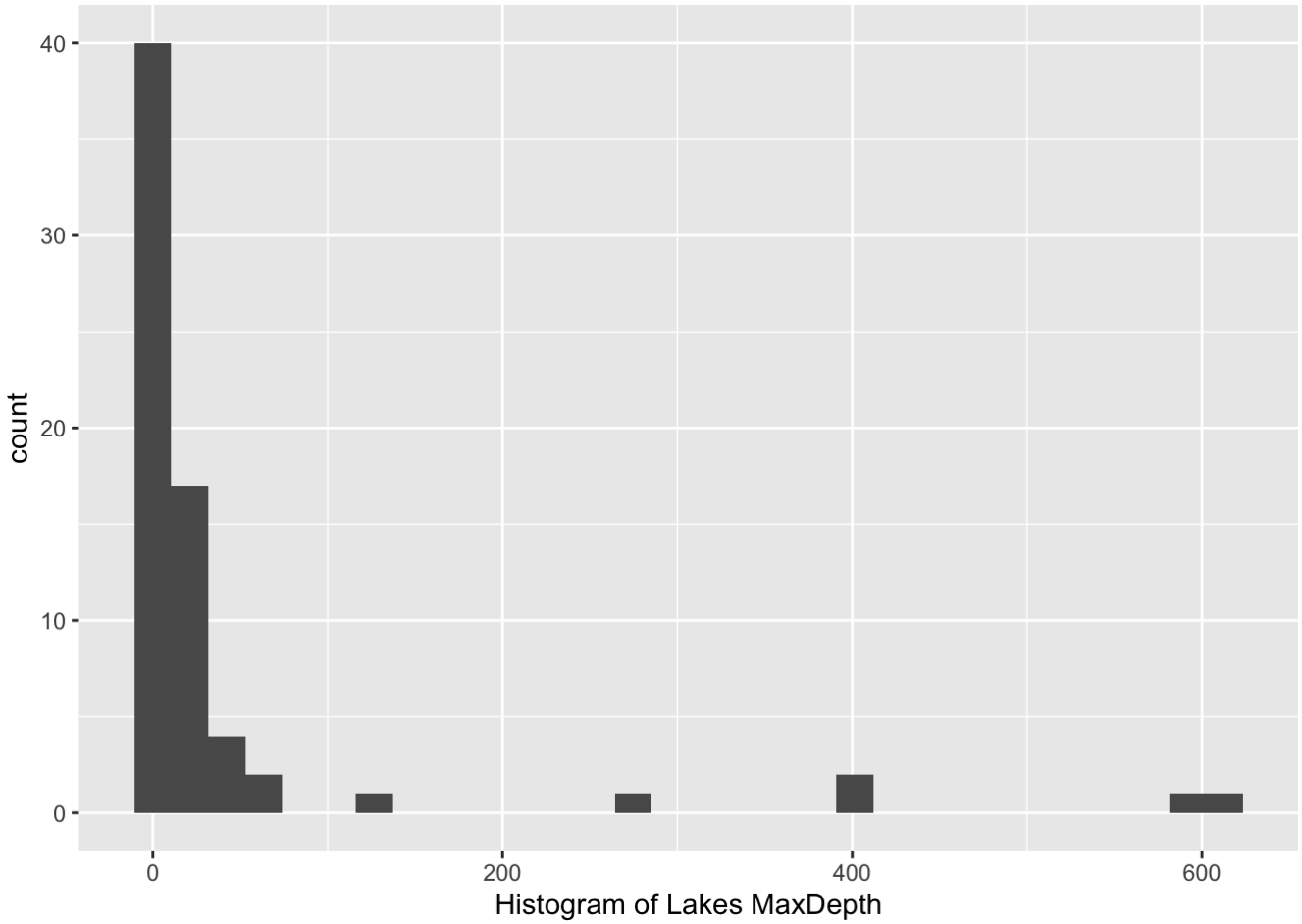
Choose one of the numeric variables in the `lakes` dataset.

```
library(rmarkdown)
any(grepl("alr4",installed.packages()))
```

```
## [1] TRUE
```

a. Plot a histogram of the variable.

```
help(package = "alr4")
data(lakes, package = 'alr4')
library(ggplot2)
a1 <- ggplot(lakes, aes(MaxDepth)) + geom_histogram() +
xlab('Histogram of Lakes MaxDepth')
a1
```



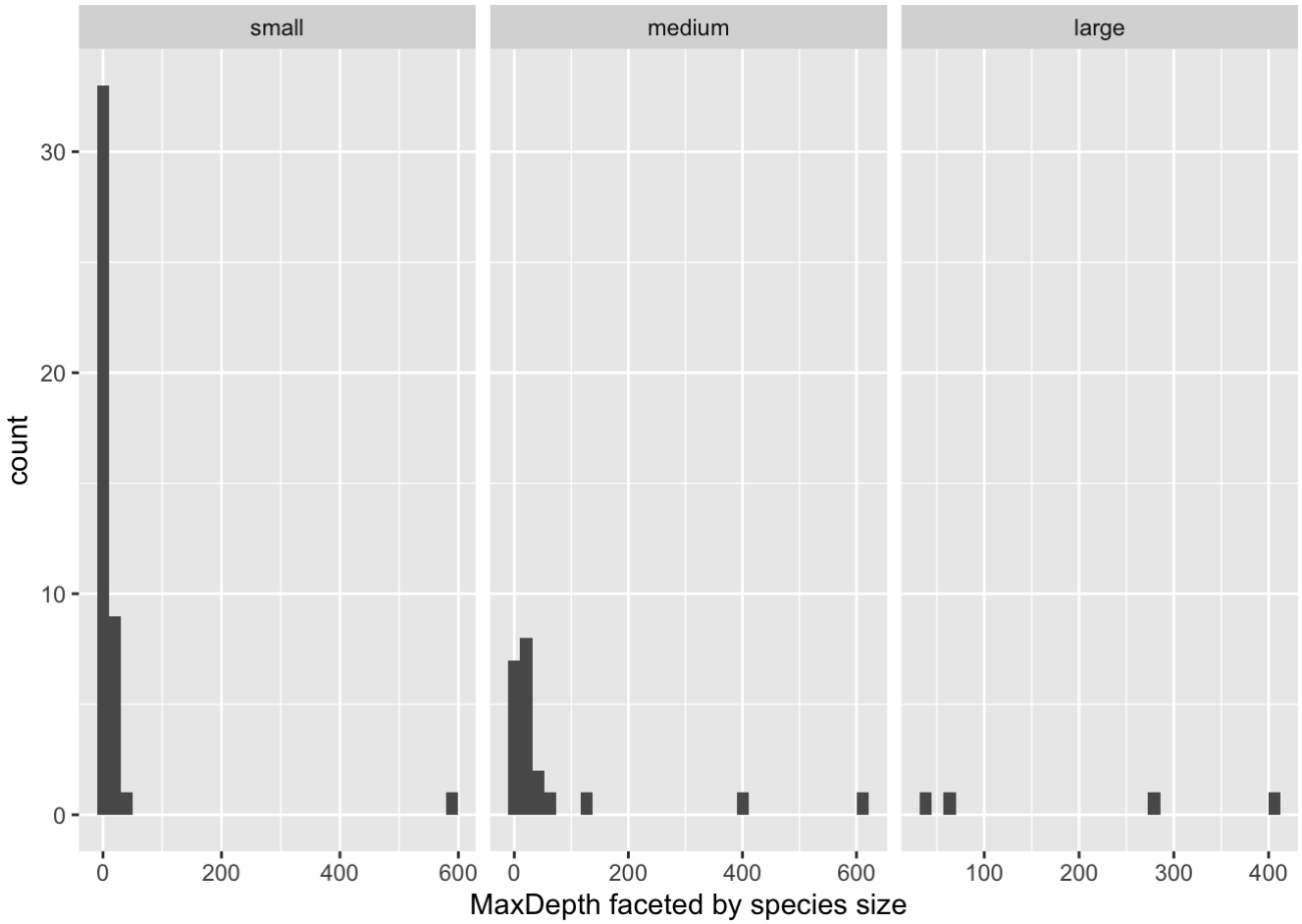

b. Create a new factor variable called `Species_size` from the `Species` column that divides the `Species` variable into three groups of equal range with levels `small`, `medium` and `large`. (Hint: use the `cut()` function.) Display the first 10 values of `Species_size`.

```
lakes$Species_size <- cut(x = lakes$Species, breaks = 3, labels = c('small', 'medium',
'large'))
head(lakes$Species_size, n = 10L)
```

```
##  [1] large  large  medium large  large  medium medium medium medium medium
## Levels: small medium large
```

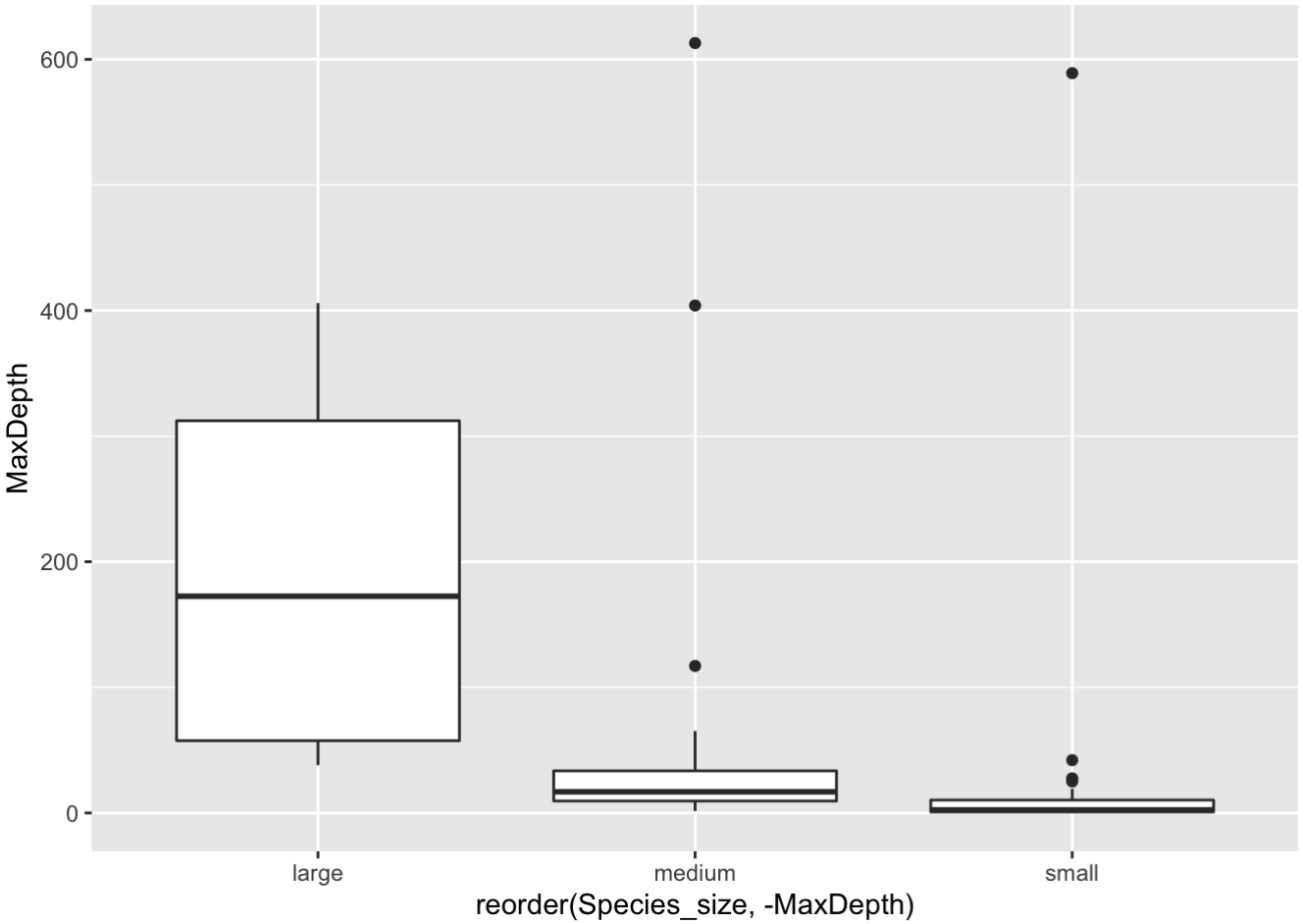
c. Plot histograms, faceted by `Species_size`, for the same variable.

```
a1 + facet_grid(cols = vars(lakes$Species_size), scale = 'free_x' ) + xlab('MaxDepth fac
eted by species size')
```



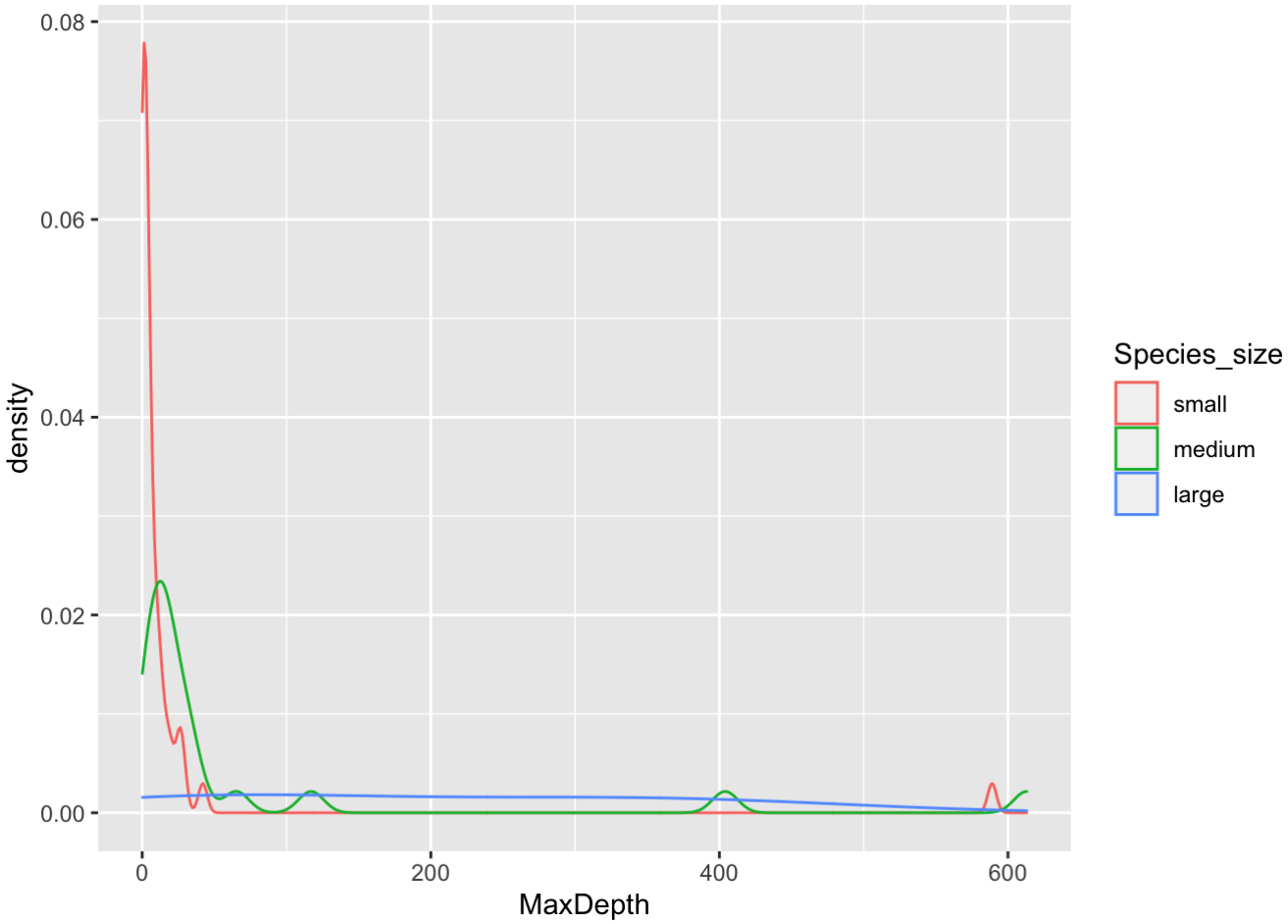

d. Plot multiple boxplots, grouped by `Species_size` for the same variable. The boxplots should be ordered by decreasing median from left to right.

```
a2 <- ggplot(lakes, aes(x = reorder(Species_size, -MaxDepth), y = MaxDepth)) +
        geom_boxplot()
a2
```




e. Plot overlapping density curves of the same variable, one curve per factor level of `Species_size`, on a single set of axes. Each curve should be a different color.

```
ggplot(data = lakes, aes(x = MaxDepth, color = Species_size)) +
geom_density()
```




f. Summarize the results of c), d) and e): what unique information, *specific to this variable*, is provided by each of the three graphical forms? In other words, describe one by one what you learn from the histograms, boxplots, and density curves that the others don't show as well or at all.

Ans: c) We can get the exact value of histogram plot. Specificly, small has highest count of more than 30 at the range of 0 to 10.
d) The bigger the species size of lakes, the larger the median and interquartile range of the Maxdepth. e) The Maxdepth of lakes with small species size is mainly located on value slightly larger than 0, medium around 10, large uniformly distributed.

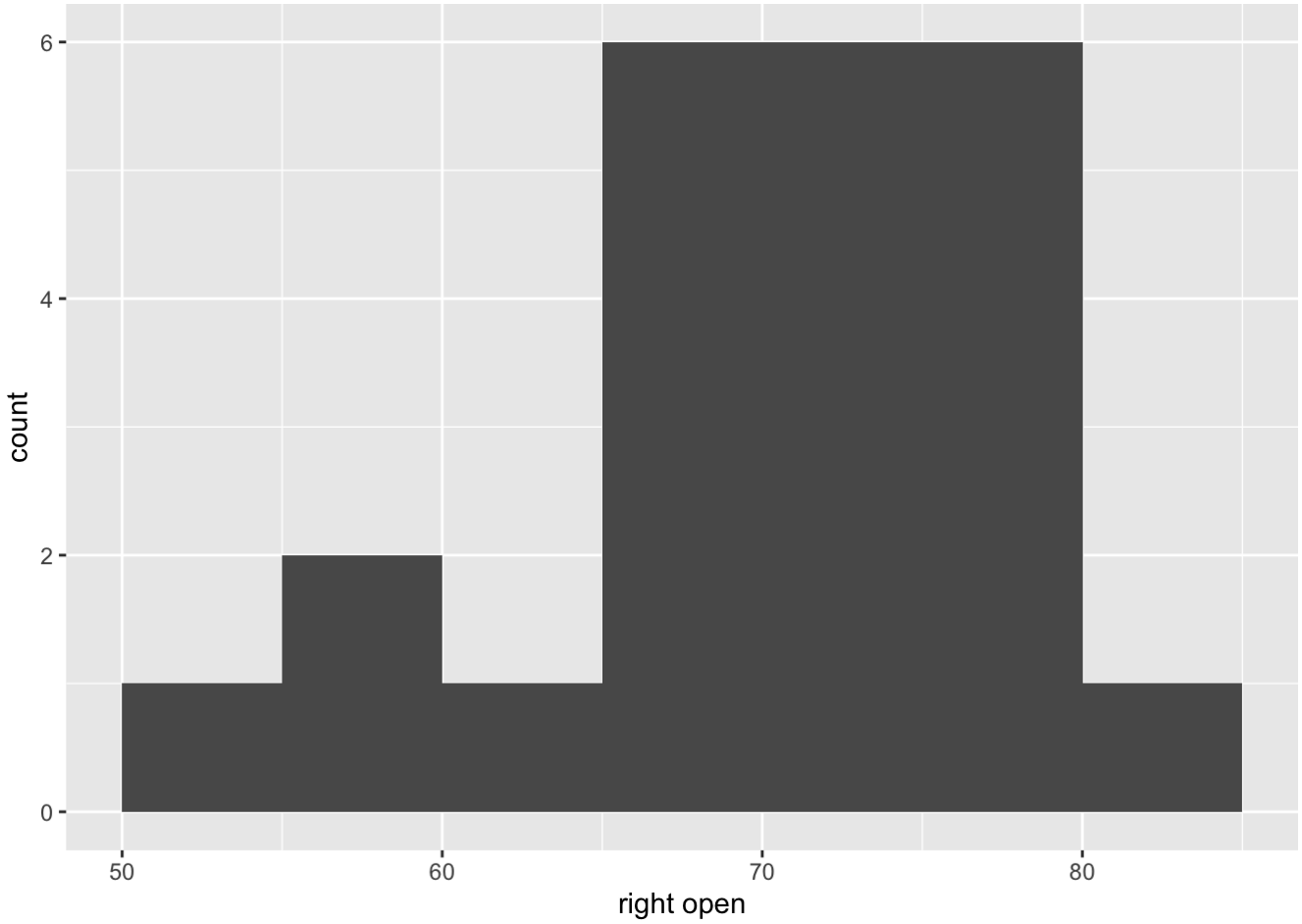
g. Briefly research the lake with the highest number of `Species`. What factors do you think contribute to this? Is this also the largest lake?

Lake Tanganyika is an African Great Lake. It is the second-oldest freshwater lake in the world, the second-largest by volume, and the second-deepest, in all cases after Lake Baikal in Siberia. [https://en.wikipedia.org/wiki/Lake_Tanganyika (https://en.wikipedia.org/wiki/Lake_Tanganyika)] We can conclude that the depth and number of species are positively correlated.

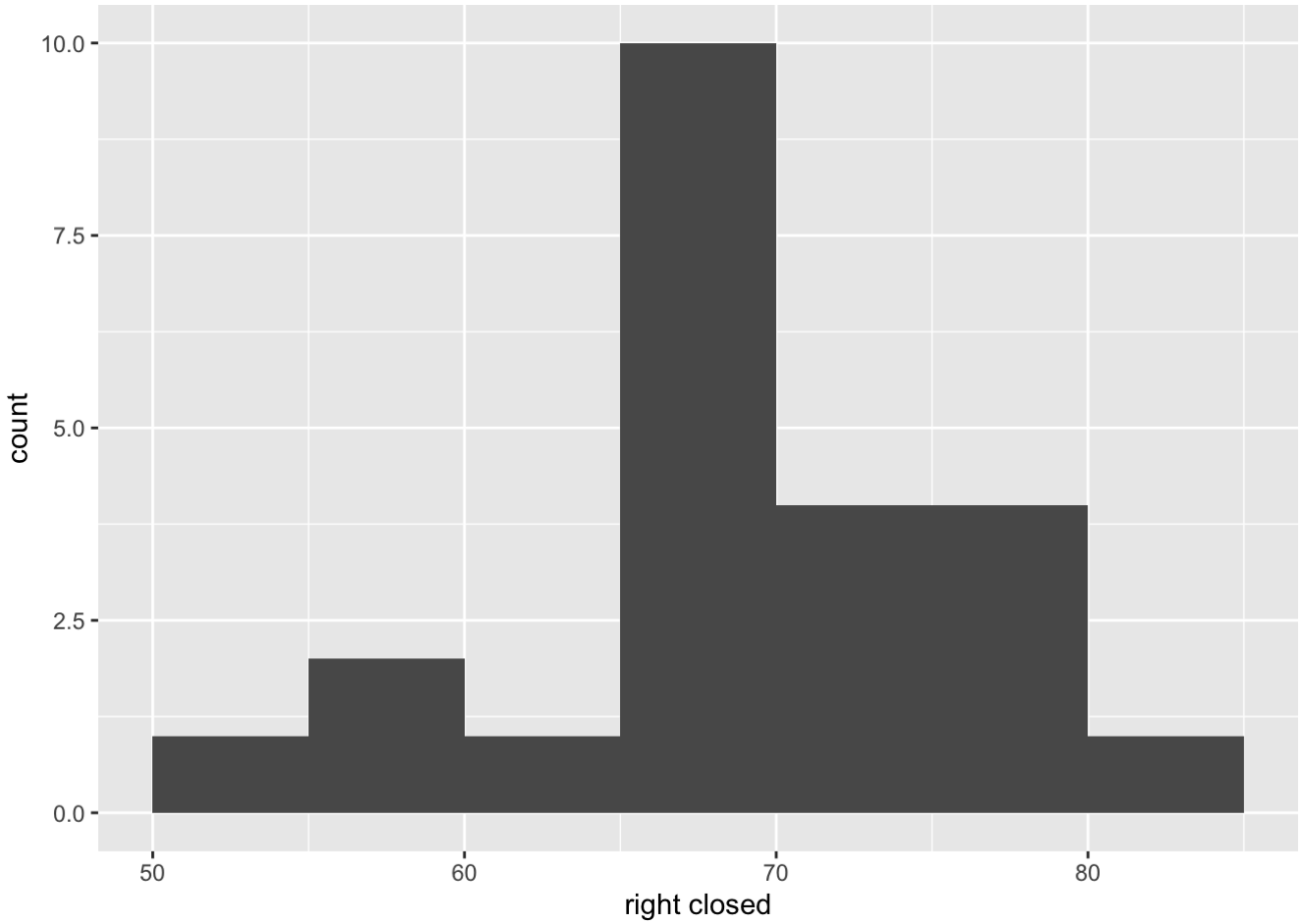
# 2. Challeng

a. Draw two histograms of the `temp` variable in the `Challeng` dataset in the **alr4** package, with binwidths of 5 years and `boundary = 0`, one right open and one right closed. How do they compare?

```
data(Challeng, package = 'alr4')
ggplot(data = Challeng, mapping = aes(x = temp)) +
  geom_histogram(binwidth = 5, boundary = 0, closed = 'left') +
  xlab('right open')
```



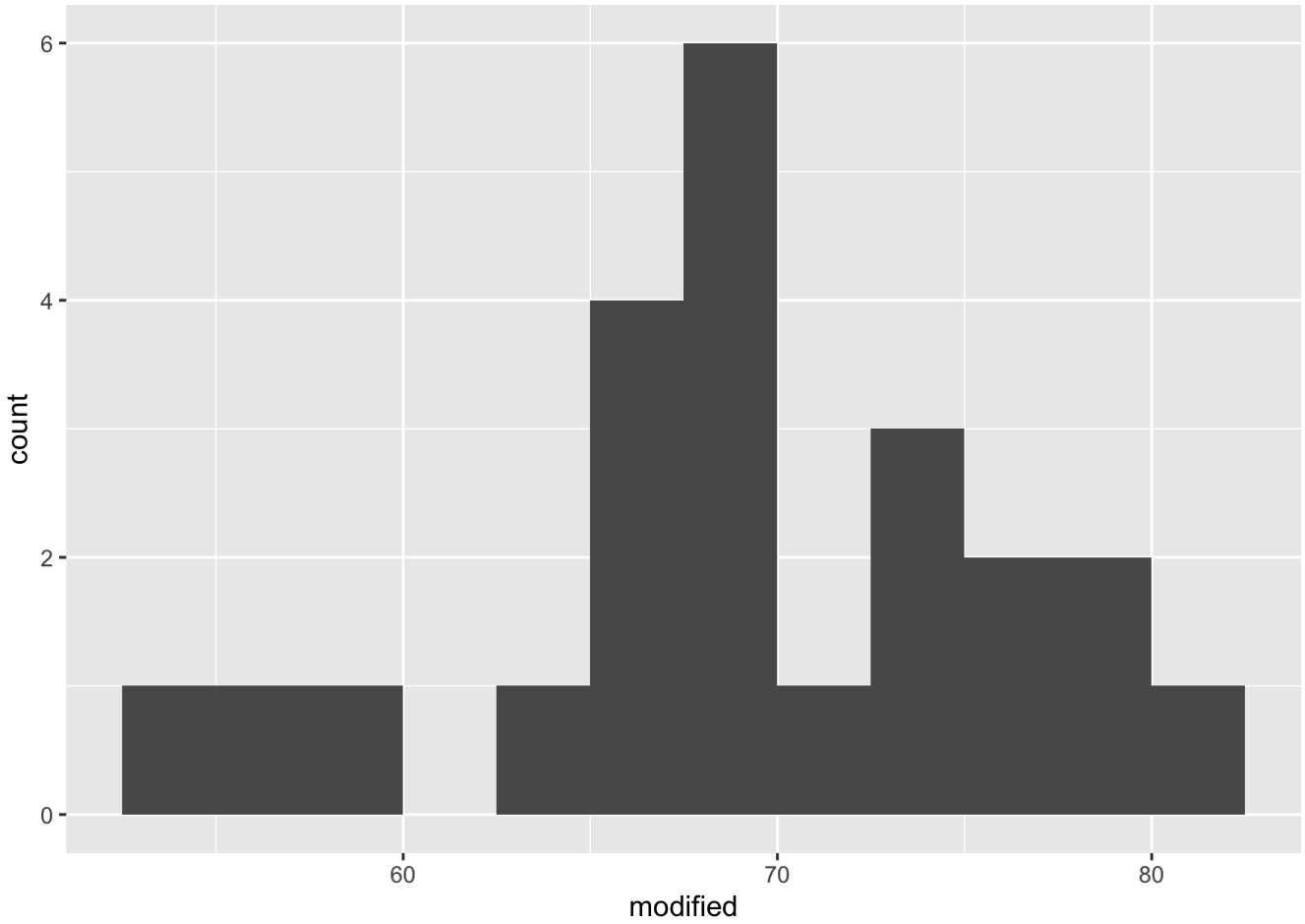

```
ggplot(data = Challeng, mapping = aes(x = temp)) +
  geom_histogram(binwidth = 5, boundary = 0, closed = 'right') +
  xlab('right closed')
```




As we can see the histogram of right open one and right closed one different.

b. Redraw the histogram using the parameters that you consider most appropriate for the data. Explain why you chose the parameters that you chose.

```
ggplot(data = Challeng, mapping = aes(x = temp)) +
  geom_histogram(binwidth = 2.5, boundary = 0, closed = 'right') +
  xlab('modified')
```




When we set the value of binwidth = 5, we can see that the histogram with right open bin and right closed bin are dramatically different. This is not what we want. So we changed the binwidth to be smaller (2.5). The difference is not obvious.

# 3. drugcost

a. Use **tidyr** functions to convert the numeric columns in the `drugcost` dataset in the **alr4** package to two columns: `variable` and `value`. The first few rows should be:

```
    variable      value
1       COST       1.34
2       COST       1.34
3       COST       1.38
4       COST       1.22
```

```
data(drugcost, package = 'alr4')
head(drugcost, 10)
```

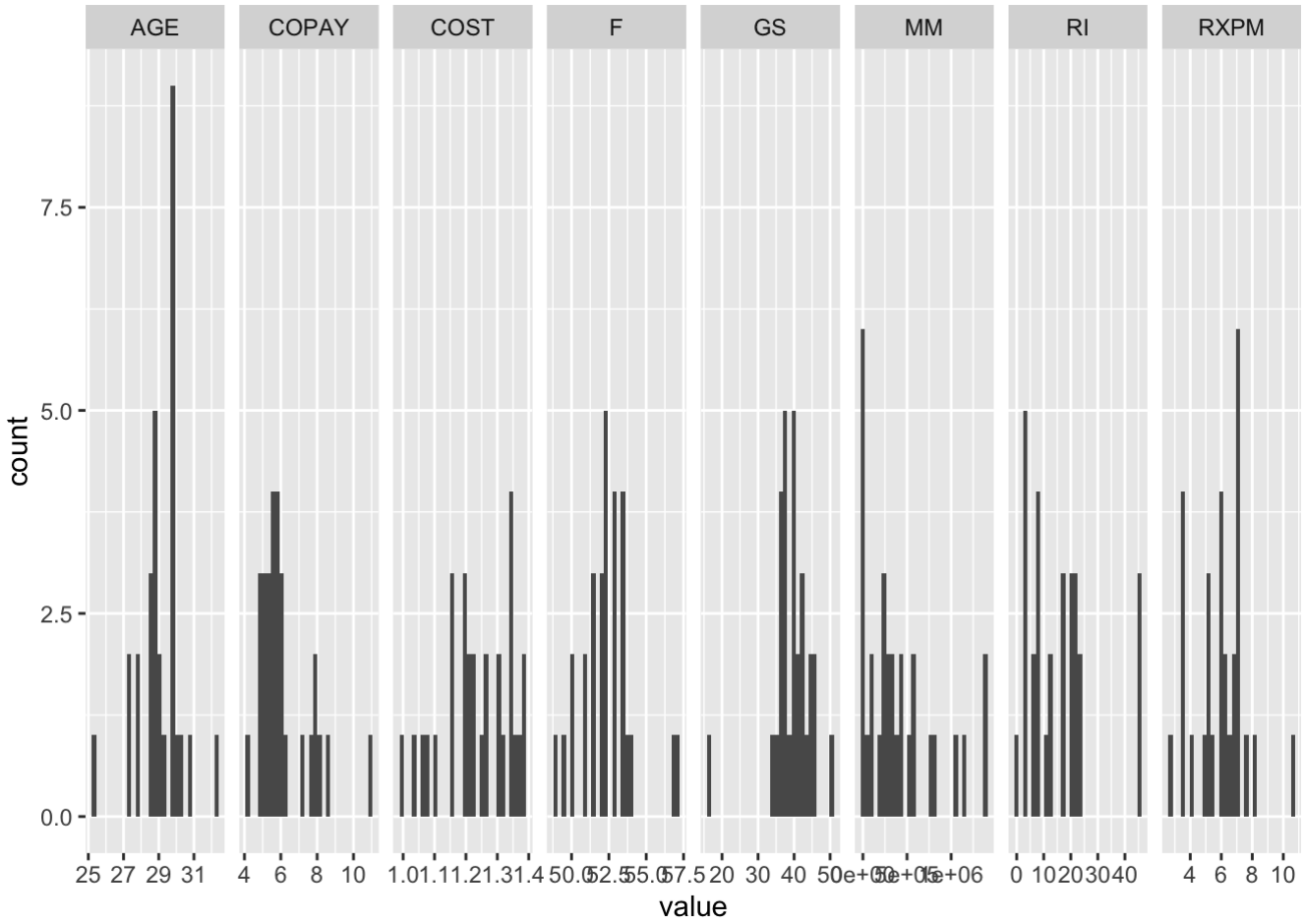
```
##           COST RXPM GS   RI COPAY  AGE    F      MM
## MN1       1.34  4.2 36 45.6 10.87 29.7 52.3 1158096
## MN2       1.34  5.4 37 45.6  8.66 29.7 52.3 1049892
## MN3       1.38  7.0 37 45.6  8.12 29.7 52.3   96168
## GA        1.22  7.1 40 23.6  5.89 28.7 53.4  407268
## GA2       1.08  3.5 40 23.6  6.05 28.7 53.4   13224
## AZ1       1.16  7.2 46 22.3  5.05 29.1 52.2  303312
## AZ2       1.25 10.7 40 22.3  4.96 29.1 52.2     720
## TN        1.20  7.6 43 21.3  7.59 29.8 51.6   73380
## San_Diego 1.10  7.2 45 20.0  6.01 32.4 50.8  513266
## NCa       1.04  6.6 42 20.0  5.79 29.8 50.0 1388605
```

Use this form to plot histograms of all of the variables in one plot by faceting on `variable`. What patterns do you observe?

```
library(tidyr)
q3a <- drugcost %>% gather(variable, value, COST:MM)
head(q3a, 4)
```

```
##   variable value
## 1     COST  1.34
## 2     COST  1.34
## 3     COST  1.38
## 4     COST  1.22
```

```
ggplot(data = q3a, mapping = aes(value)) +
  geom_histogram(bin = 6) +
  facet_grid( . ~ variable,scale = 'free_x')
```

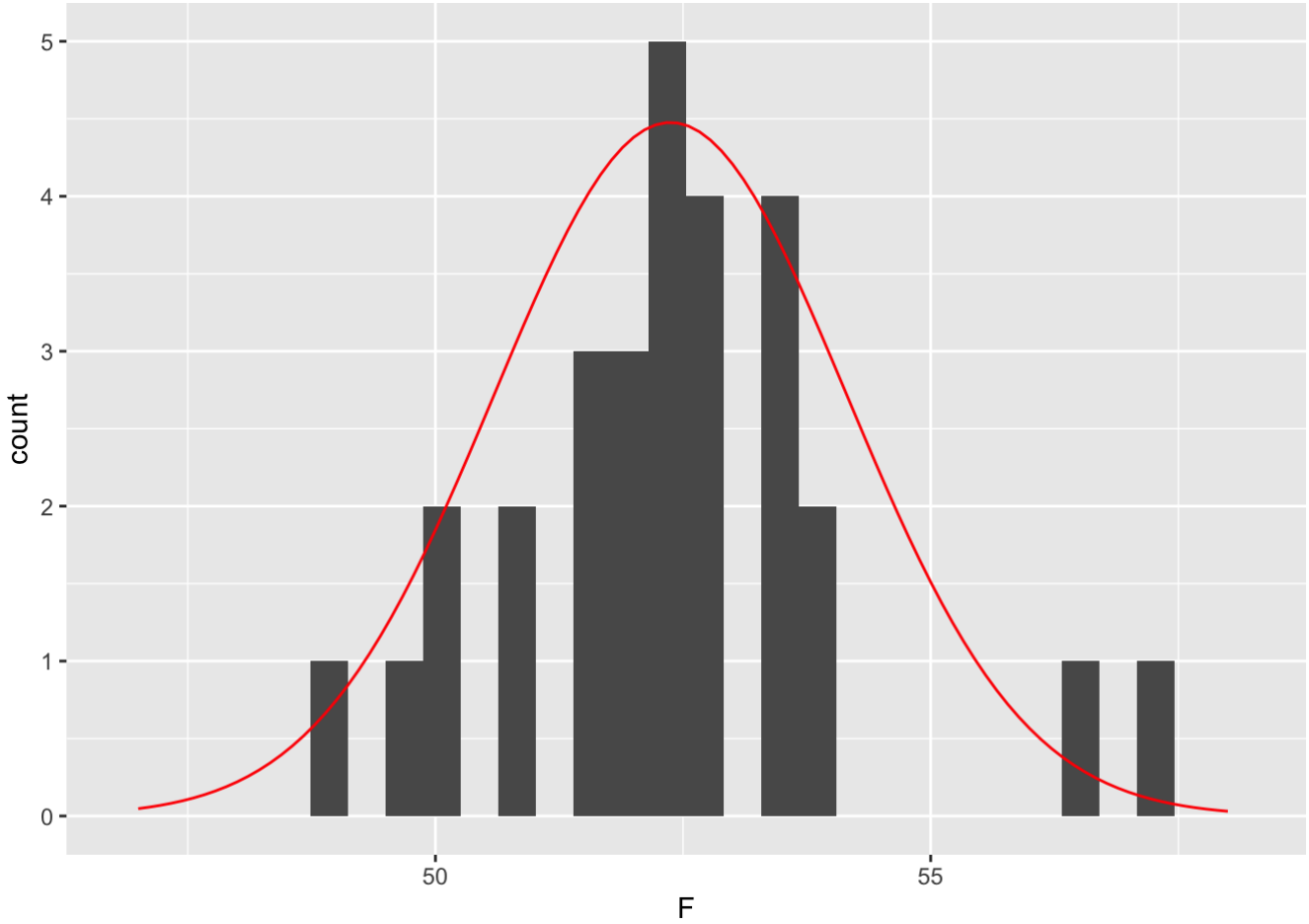



The distribution of features are quite different. For example, the age mainly focused between 28 and 30, MM is like a exponential distribution, GS focus around 40, and RI has two modals.

**For the remaining parts we will consider different methods to test for normality.**

b. Choose one of the variables with a unimodal shape histogram and draw a true normal curve on top on the histogram. How do the two compare?

```
x <- seq(47, 58, length.out=100)
dens <- with(drugcost, data.frame(x = x, y = 20 * dnorm(x, mean(F), sd(F))))
ggplot(data = drugcost, mapping = aes(F)) + geom_histogram(bin = 20) +
  geom_line(data = dens, aes(x = x, y = y), color = "red")
```




```
#stat_function(fun = dnorm,n = 29, args = list(mean = 52.5, sd = 1))
```

The histogram we had looks somewhat like normal distribution, but strictly they are not fit very well.

c. Perform the Shapiro-Wilk test for normality of the variable using the `shapiro.test()` function. What do you conclude?

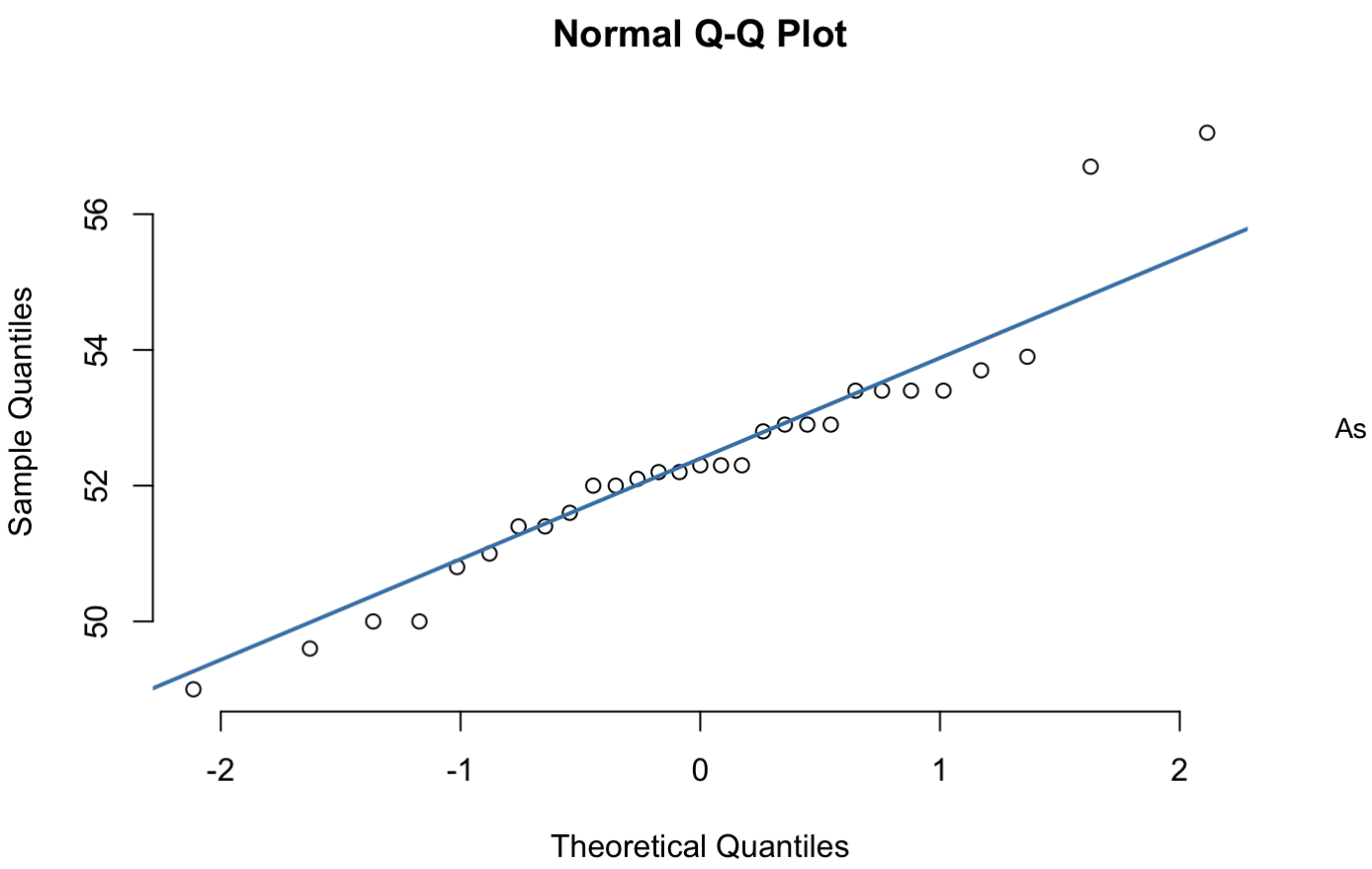
```
shapiro.test(drugcost$F)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  drugcost$F
## W = 0.92826, p-value = 0.04959
```

Because the p-value of the test is 0.04959, which is less than the threshold, so we can conclude that the F column is approximate normal distribution.

d. Draw a quantile-quantile (QQ) plot of the variable. Does it appear to be normally distributed?

```
qqnorm(drugcost$F, pch = 1, frame = FALSE)
qqline(drugcost$F, col = "steelblue", lwd = 2)
```




As

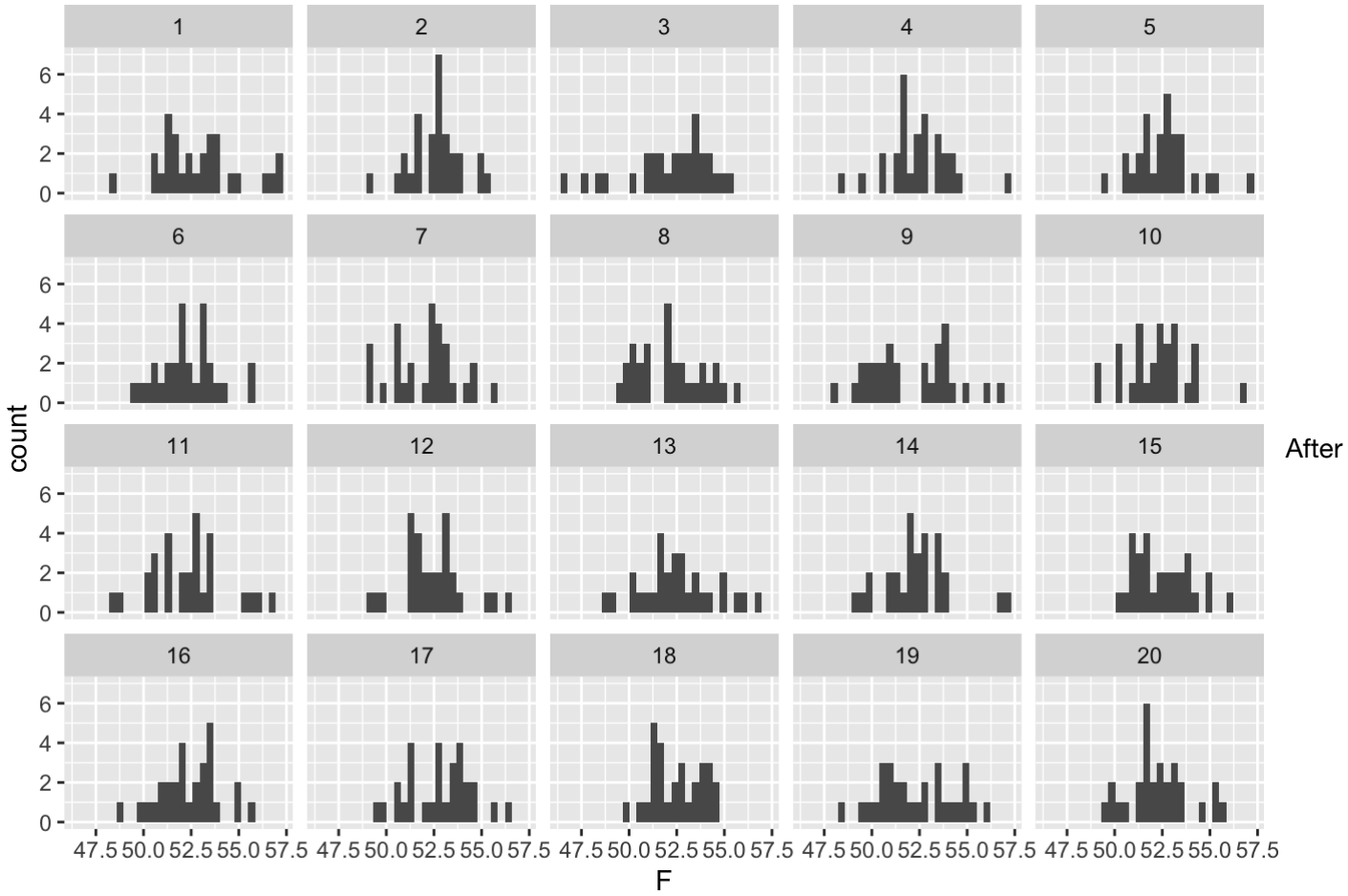
the points on the QQ plot distributes along the benchmark line with little fluctuation, it appears to be approximately normal distributed.

e. Use the **nullabor** package to create a lineup of histograms in which one panel is the real data and the others are fake data generated from a null hypothesis of normality. Can you pick out the real data? If so, how does the shape of its histogram differ from the others?

```
library(nullabor)
e3 <- lineup(null_dist("F", 'norm'), drugcost)
attr(e3, "pos")
```

```
## [1] 14
```

```
ggplot(data=e3, aes(x=F)) + geom_histogram() + facet_wrap(~ .sample)
```




we know the right answer, maybe we can see which is more normally distributed, it is hard to distinguish before knowing the righr answer.

f. Show the lineup to someone else, not in our class (anyone, no background knowledge required). Ask them which plot looks the most different from the others. Did they choose the real data? We asked two friends but only one can tell.

g. Briefly summarize your investigations. Did all of the methods produce the same result?

We used Shapiro-Wilk test, quantile-quantile (QQ) plot, and nullabor package to test if F column is normally distributed. Shapiro-Wilk test, quantile-quantile (QQ) plot, we can conclude that the F column is approximately normally distributed.But for the nullabor, the normality is not obvious.