Chapter 7: Univariate and Descriptive Statistics

Answers to Exercises

Brian Fogarty

Contents

Answers to Exercise 1

```
vf_england %>%
  count(vote2017_dum1)
```

```
# A tibble: 3 x 2
  vote2017_dum1     n
  <chr>         <int>
1 Did Not Vote    358
2 Loser           919
3 Winner          757
```

Since this variable is nominal, we can only provide the mode. The *mode* is "Loser".

Answers to Exercise 2

```
vf_england %>%
  count(vfalter1)
```

```
# A tibble: 7 x 2
  vfalter1                      n
  <fct>                     <int>
1 Strongly disagree            95
2 Disagree                    253
3 Slightly disagree           266
4 Neither agree nor disagree  769
5 Slightly agree              364
6 Agree                       166
7 Strongly agree              121
```

This variable is ordinal and thus we can provide the mode and median. But, since the variable has 7 categories, we can consider it a "high" ordinal variable and provide all of the descriptive statistics.

We see from the output above, the *mode* is "Neither agree nor disagree".

Let's use the `freq()` and `ordered()` functions to find the median.

```
freq(ordered(vf_england$vfalter1), plot=FALSE)
```

```
ordered(vf_england$vfalter1)
                          Frequency Percent Cum Percent
Strongly disagree                95   4.671       4.671
Disagree                        253  12.439      17.109
Slightly disagree               266  13.078      30.187
Neither agree nor disagree      769  37.807      67.994
Slightly agree                  364  17.896      85.890
Agree                           166   8.161      94.051
Strongly agree                  121   5.949     100.000
Total                          2034 100.000
```

We see that the *median* is "Neither agree nor disagree". We could also use the `median()` and `as.numeric()` functions to find the median.

```
median(as.numeric(vf_england$vfalter1))
```

```
[1] 4
```

Here, the *median* is "4", which is the numeric category of "Neither agree nor disagree".

Let's calculate the mean, variance, and standard deviation at the same time.

```
mean(as.numeric(vf_england$vfalter1))
```

```
[1] 4.000983
```

```
var(as.numeric(vf_england$vfalter1))
```

```
[1] 2.090506
```

```
sd(as.numeric(vf_england$vfalter1))
```

```
[1] 1.445858
```

We see the *mean* is 4.00, the *variance* is 2.09, and the *standard deviation* is 1.45.

Interestingly, we see that the mode, median, and mean are all the same value (4 or "Neither agree nor disagree").

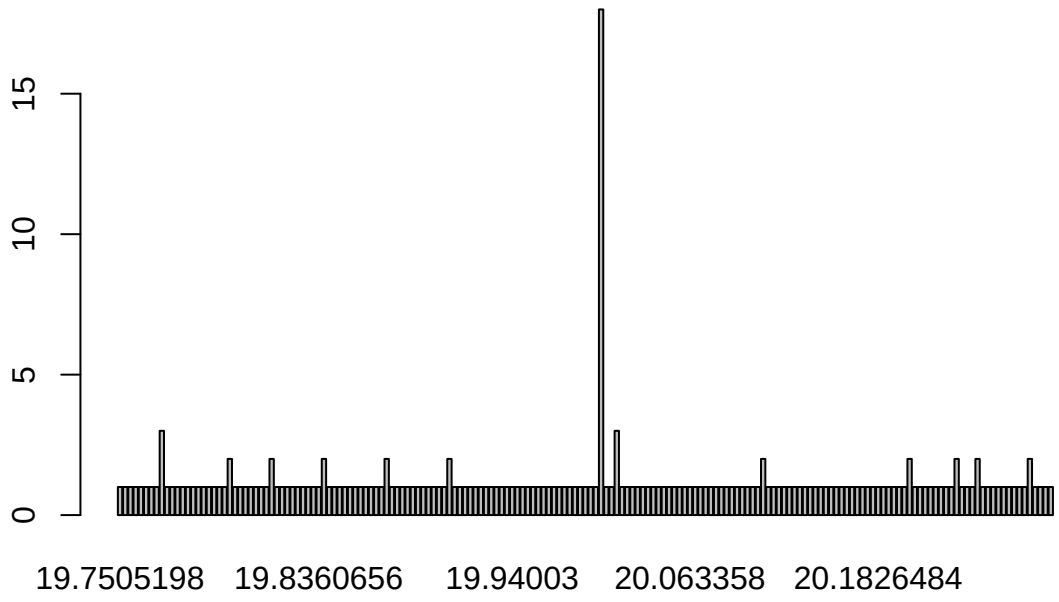
Answers to Exercise 3

Since `pct_depress` is a ratio-level variable, we can provide all of the descriptive statistics.

Finding the mode is a bit wonky as `pct_depress` has many unique values due to the decimals. We could write a function to find the mode, but a simple approach is to use the `freq()` function and look for the tallest bar in the frequency distribution plot. After doing this once, we can narrow the range of where the mode likely exists and we re-run the `freq()` function. (Output is hidden below.)

```
simd1 <- simd %>%
  filter(pct_depress > 19.75 & pct_depress < 20.25)

freq(simd1$pct_depress)
```



We see that the *mode* is 20.

Since `pct_depress` has so many values, we'll use the `median()` function and include `na.rm=TRUE` to find the median.

```
median(simd$pct_depress, na.rm=TRUE)
```

```
[1] 18.67786
```

The *median* is 18.68.

Let's calculate the mean, variance, and standard deviation at the same time.

```
mean(simd$pct_depress, na.rm=TRUE)
```

```
[1] 19.06647
```

```
var(simd$pct_depress, na.rm=TRUE)
```

```
[1] 28.87691
```

```
sd(simd$pct_depress, na.rm=TRUE)
```

```
[1] 5.373724
```

We find that the *mean* is 19.07, the *variance* is 28.88, and the *standard deviation* is 5.37. Since the mean is slightly larger than the median, we know that there are some datazones with higher values of `pct_depress` pulling the distribution to the right.

Answers to Exercise 4

```
simd %>%
  count(pct_depress_cat)
```

```
# A tibble: 4 x 2
  pct_depress_cat     n
  <fct>           <int>
1 Low               175
2 Medium           3926
3 High             2874
4 <NA>                1
```

This recoded version of `pct_depress` is ordinal with 3 values and thus we can only look at the mode and median.

```
freq(ordered(simd$pct_depress_cat), plot=FALSE)
```

```
ordered(simd$pct_depress_cat)
       Frequency   Percent Valid Percent Cum Percent
Low          175   2.50860         2.509       2.509
Medium      3926  56.27867        56.287      58.796
High        2874  41.19839        41.204     100.000
NA's           1   0.01433
Total       6976 100.00000       100.000
```

The *mode* and *median* is "Medium".

Answers to Exercise 5

a.

$$z = \frac{60 - 35}{15} = 1.67$$

```
pnorm(1.67)
```

```
[1] 0.9525403
```

The probability is .953.

We interpret this as *During lockdown, Mama Llama smoked more or the same number of cigarettes per week than* 95.3% *of the Glasgow llama population.* This can be phrased differently by using the .047 probability - 4.7% *of the Glasgow llama population smoked more or the same number of cigarettes per week than Mama Llama during lockdown.*

b.

$$z = \frac{30 - 25}{12} = 0.417$$

```
pnorm(0.417)
```

```
[1] 0.6616608
```

The probability is .662.

We interpret this as *Currently, Mama Llama smokes more or the same number of cigarettes per week than* 66.2% *of the Glasgow llama population.* This can be phrased differently by using the .338 probability -

currently, 33.8% *of the Glasgow llama population smokes more or the same number of cigarettes per week than Mama Llama.*