Chapter 7: Univariate and Descriptive Statistics

Exercises

Brian Fogarty

15 May 2018

Contents

EXERCISE I

Using the six variables recoded in Exercise I of Chapter 5 from the abbreviated version of 2015 UK Millennium Cohort survey dataset (`mcs.dta`), provide the mode, median, mean, standard deviation, and variance (where appropriate). Note: you need to use the `haven` package to read-in the data.

ANSWERS FOR EXERCISE I

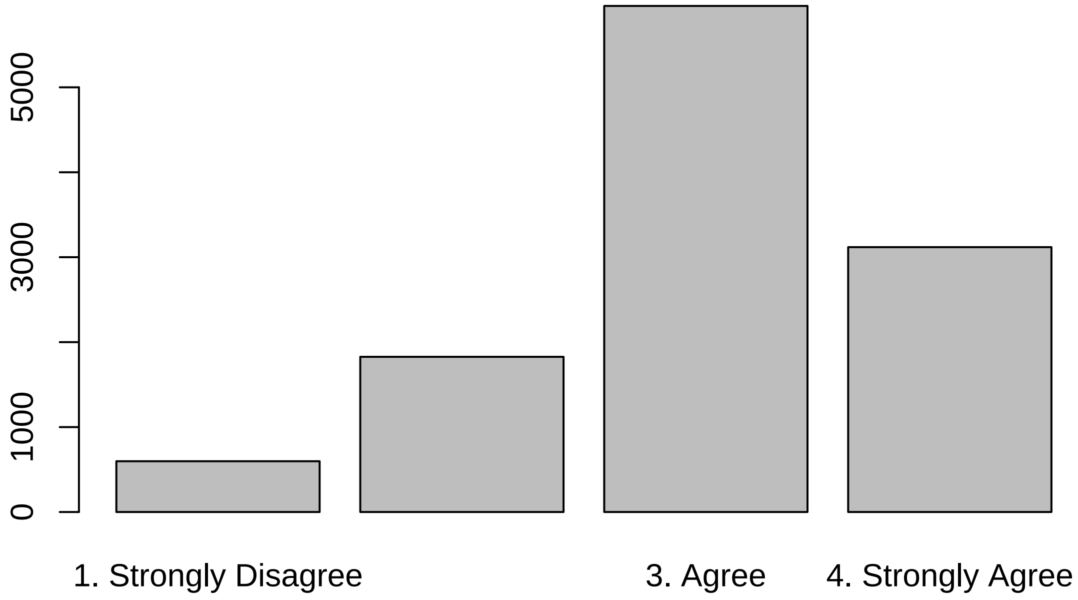
Question 1.1

Since `maths` is an ordinal-level variable, we can only look at the mode and median.

```r
library(descr)
```

```
Warning: package 'descr' was built under R version 3.4.3
```

```r
freq(mcs$maths)
```



```
mcs$maths
                       Frequency Percent Valid Percent
1. Strongly Disagree         598   5.037          5.20
2. Disagree                 1827  15.389         15.89
3. Agree                    5958  50.185         51.80
4. Strongly Agree           3118  26.263         27.11
NA's                         371   3.125              
Total                      11872 100.000        100.00
```

```
mcs$maths.num <- as.numeric(mcs$maths)
median(mcs$maths.num, na.rm=TRUE)
```

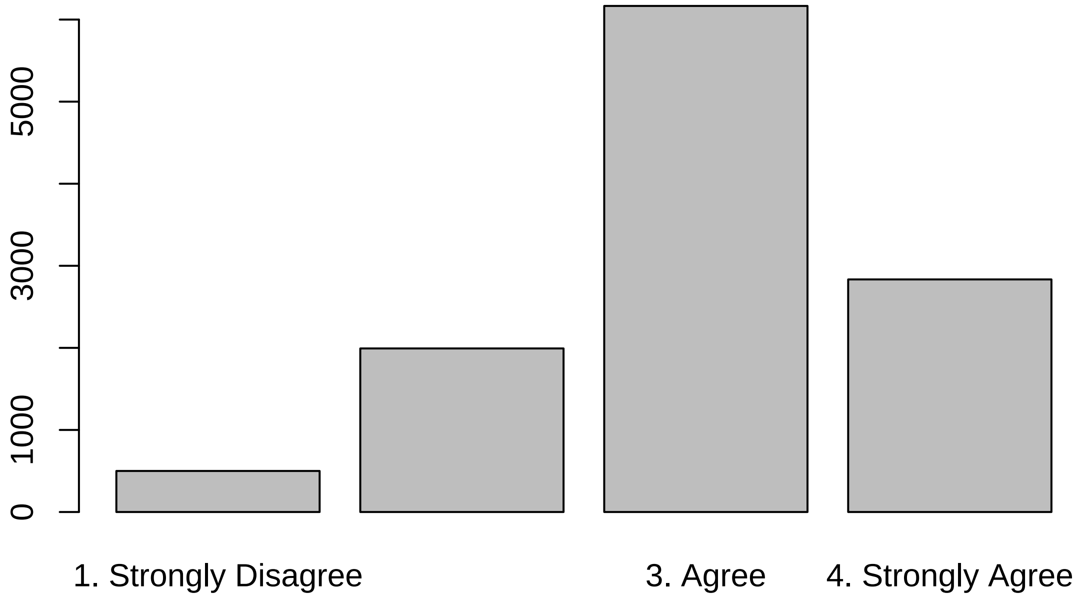
```
[1] 3
```

The mode is "agree" and the median is "agree". If we want to use the `median()` function, we need to convert `maths` to a numeric variable and run the function without missing values (`na.rm=TRUE`). Notice that the `median()` function gives us the numeric value, "3", and not the label. We need to simply check what the label is for "3" to see it is for "agree".

Question 1.2

Since `science` is an ordinal-level variable, we can only look at the mode and median.

```
freq(mcs$science)
```



```
mcs$science
                     Frequency Percent Valid Percent
1. Strongly Disagree       500   4.212          4.35
2. Disagree               1993  16.787         17.34
3. Agree                  6166  51.937         53.65
4. Strongly Agree         2834  23.871         24.66
NA's                       379   3.192
Total                    11872 100.000        100.00
```

```
mcs$science.num <- as.numeric(mcs$science)
median(mcs$science.num, na.rm=TRUE)
```

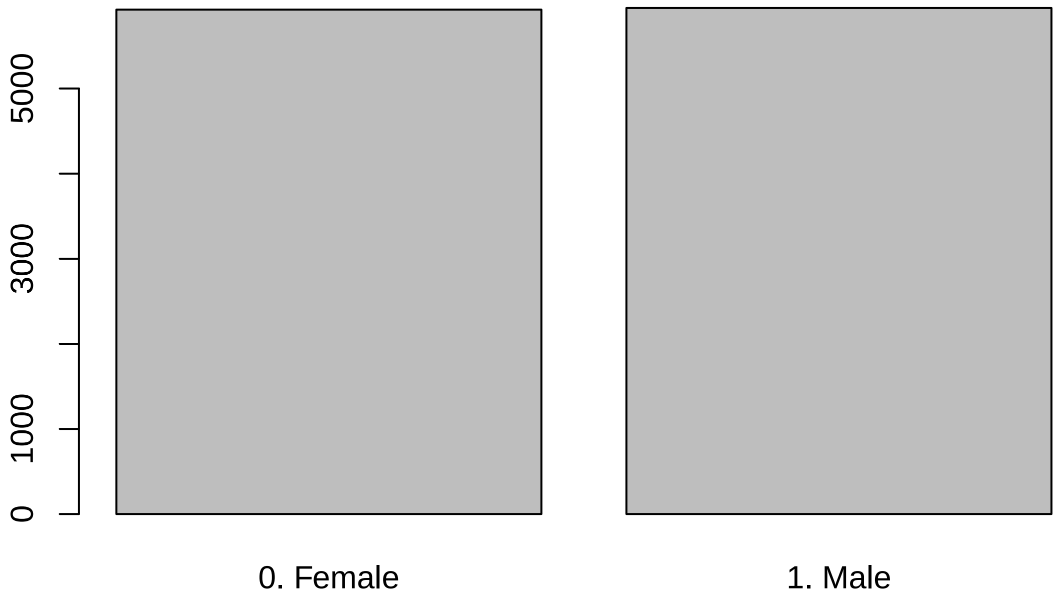
```
[1] 3
```

The mode is "agree" and the median is "agree". If we want to use the `median()` function, we need to convert `science` to a numeric variable.

Question 1.3

Since `gender` is a nominal-level variable, we can only look at the mode.

```
freq(mcs$gender)
```



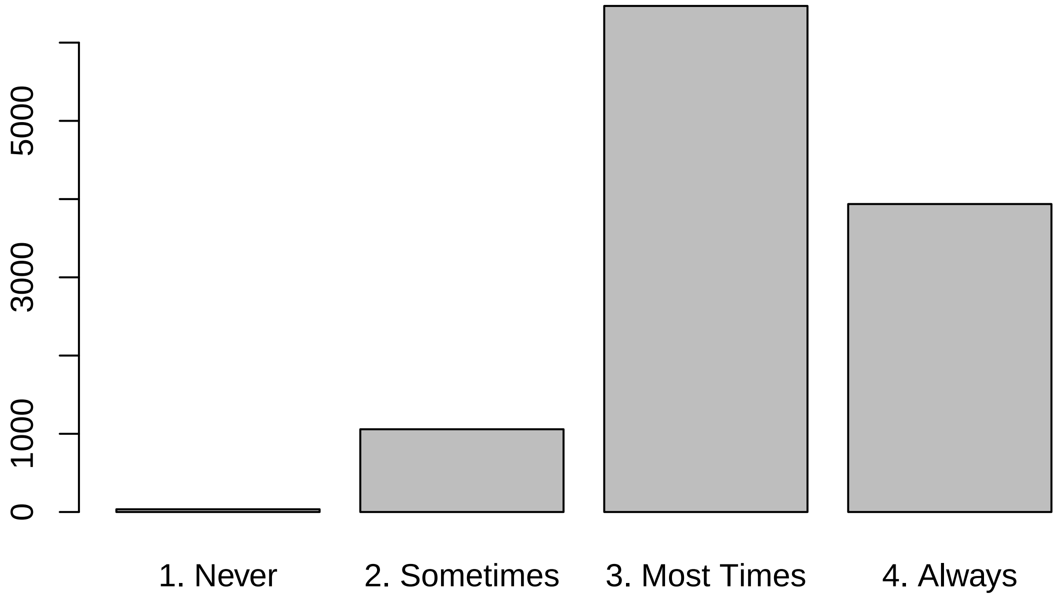
```
mcs$gender
          Frequency Percent
0. Female      5926   49.92
1. Male        5946   50.08
Total         11872  100.00
```

The mode is “male”.

Question 1.4

Since `bestsch` is an ordinal-level variable, we can only look at the mode and median.

```
freq(mcs$bestsch)
```



```
mcs$bestsch
              Frequency  Percent Valid Percent
1. Never             35   0.2948        0.3044
2. Sometimes       1058   8.9117        9.2008
3. Most Times      6469  54.4896       56.2571
4. Always          3937  33.1621       34.2378
NA's                373   3.1418
Total             11872 100.0000      100.0000
```

```
mcs$bestsch.num <- as.numeric(mcs$bestsch)
median(mcs$bestsch.num, na.rm=TRUE)
```

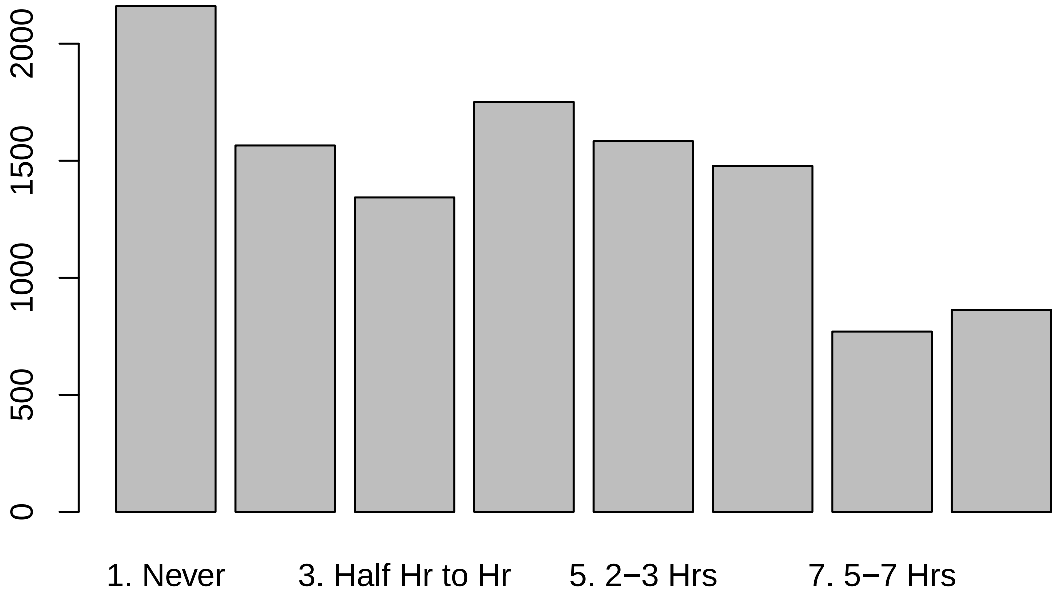
```
[1] 3
```

The mode is "most times" and the median is "most times". If we want to use the `median()` function, we need to convert `bestsch` to a numeric variable.

Question 1.5

Since `vidgames` is an ordinal-level variable, we can look at the mode and median. But, we can also consider `vidgames` to be a high ordinal-level variable, which allows us to look at all the measures.

```
freq(mcs$vidgames)
```



```
mcs$vidgames
                 Frequency Percent Valid Percent
1. Never              2160  18.194        18.763
2. Less Half Hr       1565  13.182        13.595
3. Half Hr to Hr      1343  11.312        11.666
4. 1-2 Hrs            1751  14.749        15.210
5. 2-3 Hrs            1583  13.334        13.751
6. 3-5 Hrs            1478  12.449        12.839
7. 5-7 Hrs             770   6.486         6.689
8. More 7 Hrs          862   7.261         7.488
NA's                   360   3.032
Total                11872 100.000       100.000
```

```
mcs$vidgames.num <- as.numeric(mcs$vidgames)
median(mcs$vidgames.num, na.rm=TRUE)
```

```
[1] 4
```

The mode is “never” and the median is “1-2 hours”. If we want to use the `median()` function, we need to convert `vidgames` to a numeric variable.

For the high ordinal version, we can add in the mean, standard deviation, and variance. To do so, we need to use the numeric version of the variable.

```
mean(mcs$vidgames.num, na.rm=TRUE)
```

```
[1] 3.943016
```

```
sd(mcs$vidgames.num, na.rm=TRUE)
```

```
[1] 2.190292
```

```
var(mcs$vidgames.num, na.rm=TRUE)
```

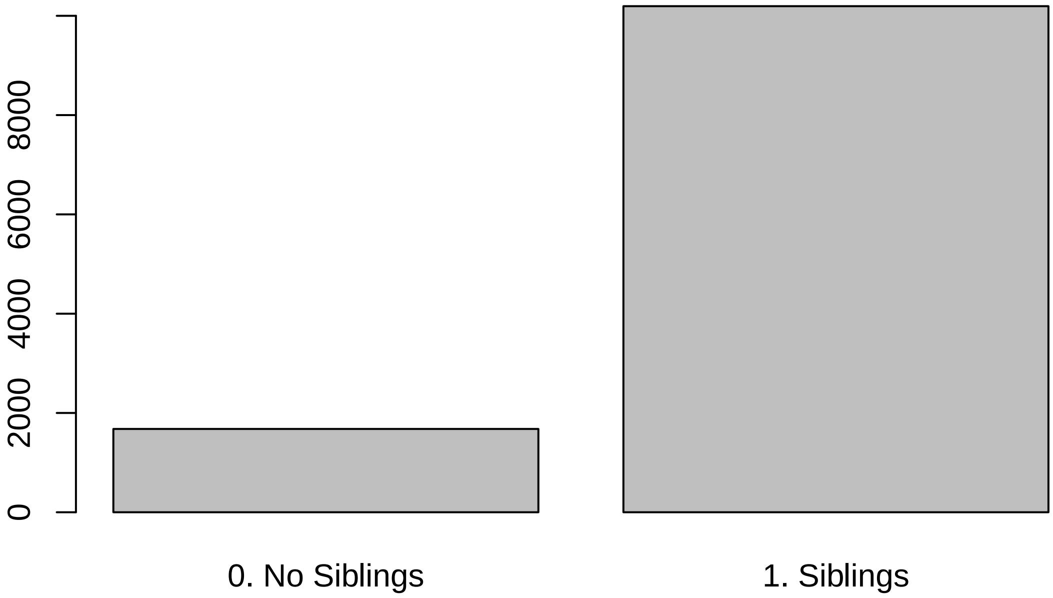
```
[1] 4.797378
```

The mean is 3.94, the standard deviation is 2.19, and the variance is 4.80.

Question 1.6

Since `siblings` is a nominal-level variable, we can only look at the mode.

```
freq(mcs$siblings)
```



```
Number of siblings
                  Frequency Percent
0. No Siblings         1678   14.13
1. Siblings           10194   85.87
Total                 11872  100.00
```

The mode is “siblings”.

EXERCISE II

Using the Scottish postcodes dataset (depdata.csv), provide the mode, median, mean, standard deviation, and variance (where appropriate) for the original version of pcnt_unemployed and the recoded version with labels from Exercise III in Chapter 5. Why is discussing the descriptive statistics for pcnt_unemployed likely more informative than for the recoded version?

ANSWERS FOR EXERCISE II

Question 2.1

pcnt_unemployed is a ratio-level variable, so we can look at all the measures.

```
depdata <- read.csv("depdata.csv")

options(max.print=9999)
freq(depdata$pcnt_unemployed, plot=FALSE)
```

```
median(depdata$pcnt_unemployed)
```

```
[1] 5.609426
```

```
mean(depdata$pcnt_unemployed)
```

```
[1] 6.276225
```

```
sd(depdata$pcnt_unemployed)
```

```
[1] 2.858125
```

```
var(depdata$pcnt_unemployed)
```

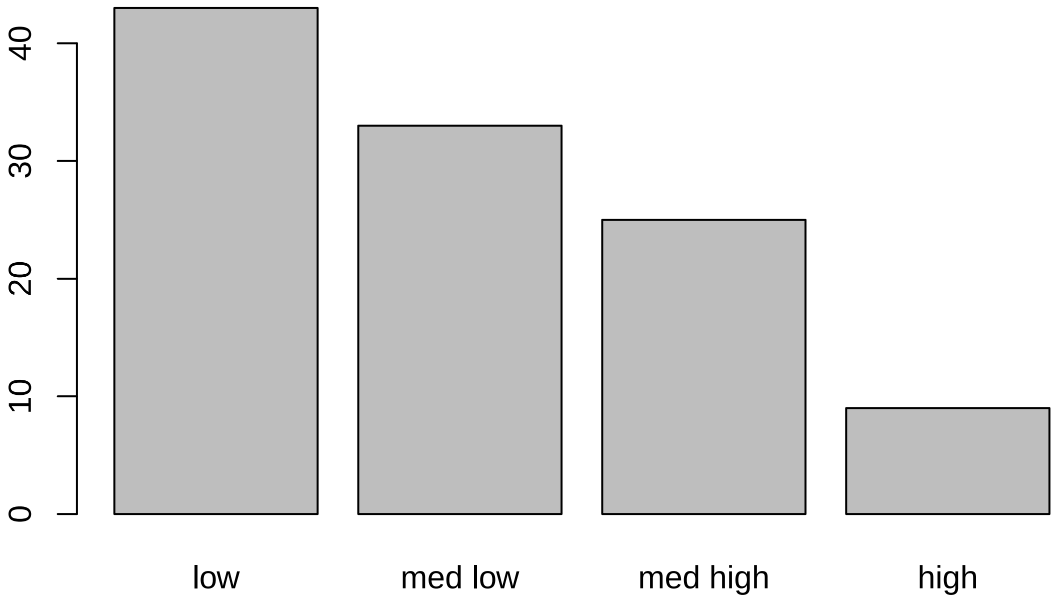
```
[1] 8.168876
```

There are 3 modes - 0, 4.615, and 7.222 percent (all the numbers correspond to percentages). The median is 5.61, the mean is 6.28, the standard deviation is 2.86, and the variance is 8.17.

Question 2.2

Since pnct_unemployed3 is an ordinal-level variable, we can only look at the mode and median.

```r
freq(depdata$pcnt_unemployed3)
```



```
depdata$pcnt_unemployed3
         Frequency  Percent Valid Percent
low             43   4.2490        39.091
med low         33   3.2609        30.000
med high        25   2.4704        22.727
high             9   0.8893         8.182
NA's           902  89.1304
Total         1012 100.0000       100.000
```

The mode is “low” and the median is “med low”.

The original `pcnt_unemployed` gives us a more precise understanding of unemployment across Scotland. We can say that the median unemployment percentage is 5.61%, which is more informative than saying the median unemployment is “med low”. The relatively small standard deviation suggests that most Scottish postcodes’ unemployment percentages are clustered near the mean of 6.28%; which is something we cannot determine from the recoded version.

EXERCISE III

Mama Llama wants to know whether her cigarette smoking is excessive in the Glasgow llama population. You need to help her figure it out.

1. Mama Llama smokes 40 cigarettes a week (x), while the mean llama smoking is 30 cigarettes a week (μ) and the standard deviation is 10 cigarettes a week (σ). Calculate the z-score.

2. Using the `pnorm()` function and the calculated z-score, find the probability.
3. Interpret the probability using plain language.

ANSWERS TO EXERCISE III

Question 3.1

$$z = \frac{40 - 30}{10} = 1.00$$

Question 3.2

```r
pnorm(1.00)
```

```
[1] 0.8413447
```

The probability is .841.

Question 3.3

We interpret this as *Mama Llama smokes more or the same number of cigarettes per week than* 84.1% *of the Glasgow llama population.* This can be phrased differently by using the .159 probability - 15.9% *of the Glasgow llama population smokes more or the same number of cigarettes per week than Mama Llama.*