ST102/ST109 Exercise 1, Outline solutions

1. (a) Insert your guess here. Chances are that you got the average correct, without even thinking that it could possibly be anything else. The standard deviation was probably less obvious.

   (b) i. Using the rules of summation:

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\sum_{i=1}^{n}(aX_i+b) = \frac{1}{n}\left(\sum_{i=1}^{n} aX_i + \sum_{i=1}^{n} b\right) = \frac{a}{n}\sum_{i=1}^{n} X_i + \frac{nb}{n} = a\bar{X}+b.$$

   ii. We have:

$$\begin{aligned} s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{Y})^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(aX_i+b-(a\bar{X}+b))^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(a(X_i-\bar{X})+b-b)^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n} a^2(X_i-\bar{X})^2 \\ &= \frac{1}{n-1}a^2\sum_{i=1}^{n}(X_i-\bar{X})^2 \\ &= a^2 s_X^2. \end{aligned}$$

   iii. From above, we immediately get $s_Y = \sqrt{s_Y^2} = |a|s_X$. Note that the absolute value is there to deal with cases of negative a.

   It is worth noting two special cases where the standard deviation does not change.

   - $a = -1$ and $b = 0$, i.e. $Y_i = -X_i$. Therefore, $s_Y = |-1|s_X = s_X$. Simply changing the sign of each X_i does not change the standard deviation.
   - $a = 1$, i.e. $Y_i = X_i + b$. Therefore, $s_Y = s_X$. Adding the same constant to each observation does not change the standard deviation.

   (c) In (a), X is the temperature in Celsius, Y is the temperature in Fahrenheit, $a = 9/5$ and $b = 32$. Using (b), $\bar{Y} = (9/5) \times 14 + 32 = 57.2$, and $s_Y = (9/5) \times 7.7 = 13.86$.

2. We have:

$$\begin{aligned} \sum_{i=1}^{n}(X_i-\bar{X})^2 = \sum_{i=1}^{n}(X_i^2 - 2X_i\bar{X}+\bar{X}^2) &= \sum_{i=1}^{n} X_i^2 - 2\bar{X}\overbrace{\sum_{i=1}^{n} X_i}^{=n\bar{X}} + \overbrace{\sum_{i=1}^{n}\bar{X}^2}^{=n\bar{X}^2} \\ &= \sum_{i=1}^{n} X_i^2 - n\bar{X}^2. \end{aligned}$$

3. & 4. The output from R is shown below. In Question 3, it is easiest to use the table of frequencies, and the rules of calculating the statistics from grouped data, with the frequencies, f_j, of distinct values of X.

- For the mean:

$$\bar{X} = \frac{\sum f_j X_j}{\sum f_j} = \frac{\sum f_j X_j}{n} = \frac{1 \times 8 + 1 \times 9 + \cdots + 1 \times 40}{1 + 1 + \cdots + 1} = \frac{705}{44} = 16.02.$$

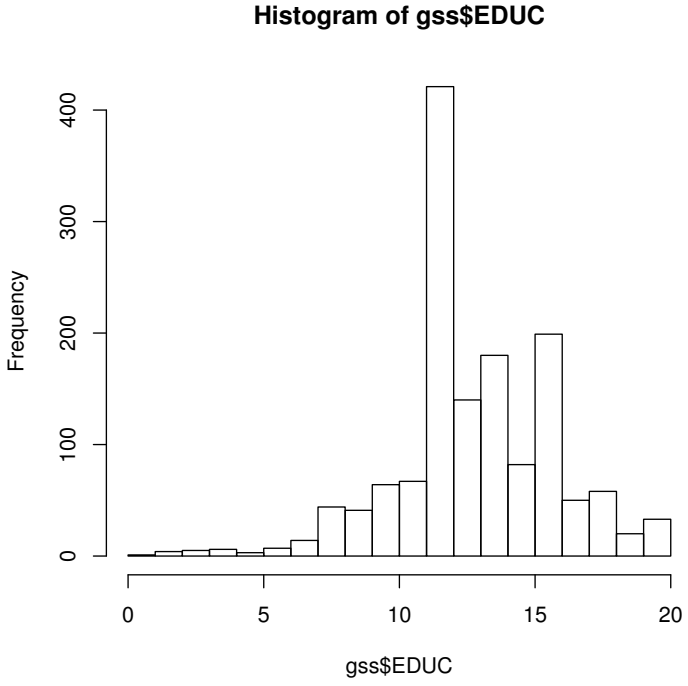
- For the standard deviation:

$$s = \sqrt{\frac{\sum f_j X_j^2 - n\bar{X}^2}{n-1}} = 7.14.$$

– For the median, find the value where the cumulative percentage passes 50%. Here it is 15.

```
> summary(lengths$Guess)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   8.00   11.00   15.00   16.02   17.00   40.00
> mean(lengths$Guess)
[1] 16.02273
> median(lengths$Guess)
[1] 15
> sd(lengths$Guess)
[1] 7.144647
```

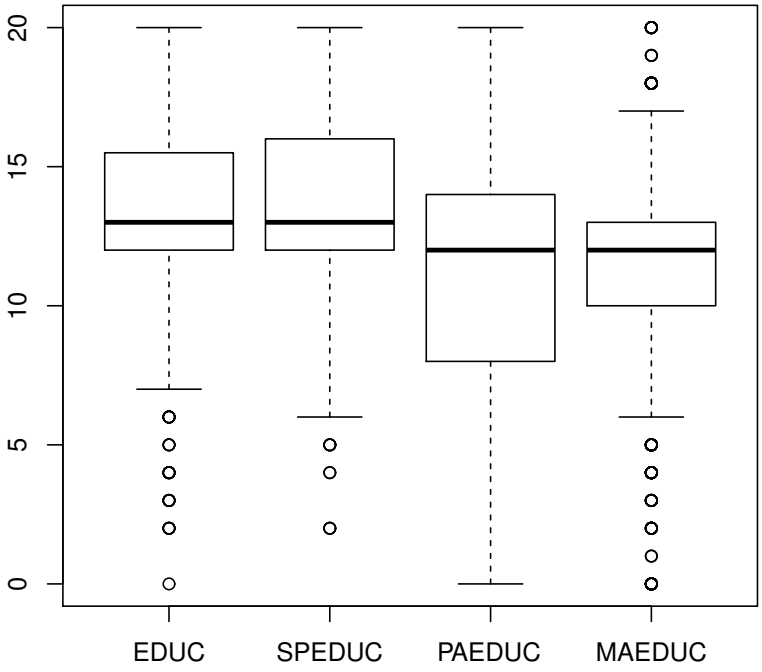
5. (a) The histogram is shown below. Note the two high bars at around 12 years (people who finished high school, but no more) and 16 years (4 years of education after high school, such as an undergraduate college degree). Another point to note is that a relatively low proportion of the respondents has less than high-school education.



(b) The statistics and plot are shown below. Some observations are the following.

- The respondents' and spouses' distributions are very similar. This is not surprising, because the two groups are effectively two random samples from nearly the same population, containing both men and women (if the respondent is a man then the spouse is a woman, and vice versa). The only difference between the two is that the spouses are by definition all married, while the respondents are not (note the different sample sizes: `SPEDUC` is missing for respondents who are not married).
- The averages for the mothers and fathers are about two years, and medians one year, lower than for the respondents and spouses. The average length of education has, therefore, increased that much between these generations. Note that here 'generation' is relative to the respondent, and does not correspond to constant age. For example, if the respondent is 30 years old, then the parents are typically 50–60, but the parents of a respondent who is 60 years old are 80 or older.
- The middle 50% of the distribution of the mothers' education (the box of the boxplot) is tightly concentrated around 12 years, with relatively few observations far from that (this is also why R stops the whiskers quite quickly and declares the rest of the observations to be 'outliers'). The fathers' distribution, on the other hand, has higher dispersion, and also includes a fairly large proportion of men with less than 10 years of education.

```
> summary(gss$EDUC)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.00   12.00   13.00   13.23   15.50   20.00       6
> sd(gss$EDUC, na.rm=TRUE)
[1] 2.949263
> summary(gss$SPEDUC)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   2.00   12.00   13.00   13.33   16.00   20.00     779
> sd(gss$SPEDUC, na.rm=TRUE)
[1] 2.854709
> summary(gss$PAEDUC)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.00    8.00   12.00   11.26   14.00   20.00     423
> sd(gss$PAEDUC, na.rm=TRUE)
[1] 4.298578
> summary(gss$MAEDUC)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
    0.0    10.0    12.0    11.4    13.0    20.0     236
> sd(gss$MAEDUC, na.rm=TRUE)
[1] 3.557759
```



(c) The plot is shown below. It is not completely easy to read, because the years of education are recorded as integers. This means that only pairs of integers are possible in the plot, and most of the points in it correspond to several individual observations. Nevertheless, it is clear that there is a positive association of the kind discussed in the question. For example, note that all of the married respondents who themselves have more than 16 years of education have spouses with at least 12 years of education, and that no respondents with less than 10 years of education have spouses with more than 14 years of education.