Introduction to R
Session 3 exercises

Statistical Consulting Centre

1 March, 2017

1 Missing values

(i) In question 3(ii) of exercise 2 you identified the "Can't choose" cases in `q1a`. Now, replace these cases by `NA`.

(ii) Repeat 1(i) for `q1b` – `q1e`, so that all cases of "Can't choose" are replaced by `NA`.

(iii) Produce a one-way frequency table of `ethnicity`.

(iv) Repeat **1**(iii) after replacing all cases of "NA, dont know" with `NA`.

(v) There are only two possible values for `partner`: `Yes` and `No`. Replace any values which are not `Yes` or `No` with `NA`.

2 Factor

(i) Produce a two-way frequency table of `q1a` versus `gender`.

(ii) Table 1 shows the appropriate ordering of the levels of the values in `q1a` – `q1e`.

Table 1: The right levels for `q1a` to `q1e`

q1a	Factor(q1a)
Daily	1
Several times a week	2
Several times a month	3
Several times a year or less often	4

Convert `q1a` – `q1e` into factors with their levels ordered as shown in Table 1. Then generate two-way frequency tables between `q1a` to `q1e`, respectively, versus `gender` to check that you've appropriately ordered these factors' levels.

(iii) Create a new variable which categorises all participants into one of three age groups: "Under 40", "41 to 60" and "Over 61".

(iv) Convert the variable created in **2**(iii) into factors with appropriate levels.

(v) Add the factor into `sports.df` and name it `age.group`

3 Challenge

We mentioned in Exercise 2 that the function `mystder` calculates the standard error of the mean (SEM), i.e.

```
mystder <- function(x){
       mysd <- sd(x, na.rm = T)
       n <- length(x)
       mysd/sqrt(n)
}
```

This function only calculates the standard error correctly if the input does NOT contain missing values. This is because the `length()` function counts the number of elements in the variable, including missing values. For example:

```
test <- c(1, 2, 3, 4, NA)
length(test)

[1] 5
```

So, `length(test)` returns 5 instead of 4. Suppose you repeat an experiment 5 times, resulting in one missing value; your real/valid sample size is 4. Thus, when you calclate your standard error, use $n = 4$ instead of 5. For example,

```
mysd <- sd(test, na.rm = T)
mysd

[1] 1.290994

n <- 4
n

[1] 4

mysd/sqrt(n)

[1] 0.6454972
```

The real SEM for `test` should be 0.6454972; however, if we use `mystder()` to calculate it we get:

```
mystder(test)

[1] 0.5773503
```

Thus, calculating the sample size using `length()` will lead to an incorrect solution when there are missing values in the data.

(i) Now that you know what is wrong with `mystder()`, modify it so it gives the correct SEM even if the input contains missing values.

(ii) Apply your modified `mystder` function to `test` to see whether it returns the correct answer, i.e. 0.6454972.

(iii) Create `test2`, as shown below, and test your function on this new variable.

```r
test2 <- c(1:100, rep(NA, 30))
```

The correct value for the SEM should be 2.9011492.