Exercise 1 - How Deep is the Ocean? September 17, 2018.

EPIB607 - Inferential Statistics[a]

[a]Fall 2018, McGill University

This version was compiled on September 17, 2018

This in-class exercise will introduce you to sampling distributions for means and proportions.

Sampling distribution | Means | Proportions | Standard error | Standard deviation | Parameter | Statistic

1. What percentage of the world's surface is covered by water?

The data provided by the Scripps Institution of Oceanography can provide an answer, but some work is required on your part. James Hanley (JH) has randomly sampled $n = 5$ and $n = 20$ latitudes and longitudes for every student in the class. This document contains unique latitudes and longitudes for

```
[1] "Bhatnagar, Sahir"
```

and was sent in an email (using the gmailr package) as a pdf attachment to the following address:

```
[1] "sahir.bhatnagar@mcgill.ca"
```

A sample of 5.

```
Latitude.n.5 <- c(8.042,45.329,37.743,-12.237,21.552)
```

```
Longitude.n.5 <- c(35.241,125.735,-134.383,-23.623,-5.339)
```

A sample of 20.

```
Latitude.n.20 <- c(40.428,59.851,-38.35,-8.862,-26.312,-49.721,-4.02,44.028,-36.308,22.071,
44.765,-18.818,37.185,41.768,-5.255,14.199,61.165,10.874,-27.083,15.792)
```

```
Longitude.n.20 <- c(57.814,40.026,131.012,-145.304,-130.685,48.011,59.257,111.303,41.137,-61.05,
159.666,-110.195,-87.961,-118.8,99.592,-102.281,159.827,-166.667,-2.171,-6.664)
```

1.1. Determine the proportion of water from the sample of 5.

1. Using the sample of 5, manually enter the latitudes and longitudes in Google maps and record if you land on water or land. Figure1 shows how to enter them.

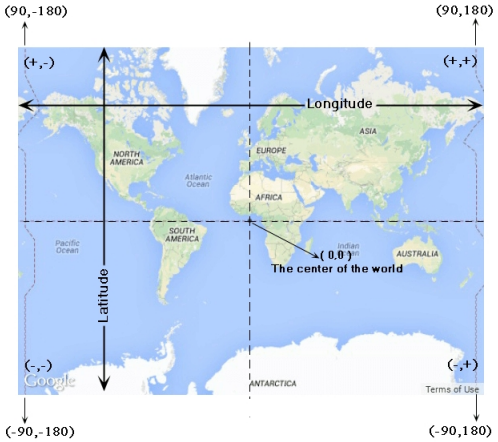

Fig. 1. Latitude and longitude in Google maps. Latitudes range from (-90,90) and longitudes range from (-180, 180). Latitude is entered first followed by longitude and separated by a comma.

2. In R, store your results in a binary vector of length 5. For example, if your sample landed on water 3 times out of 5, then you would enter the following in R (1 for water and 0 for land):

```
landed_in_water <- c(1,1,1,0,0)
```

3. Using R, estimate the percentage of the world's surface covered by water from your sample of 5. This can be done by simply taking the mean of the binary vector created in Step 2: `mean(landed_in_water)`.

4. Enter your estimate in this Google sheet next to your name and in the column titled `PropnW.5.locations`.

1.2. Determine the proportion of water from the sample of 20.

1. Repeat the above steps for the sample of 20. **Before you do**, take a moment to think about how tedious a process it can be to manually enter 20 latitudes and longitudes into Google maps. JH and SB hope you can appreciate the parallels between this toy exercise and that of collecting data for your research projects, i.e., it becomes increasingly difficult (effort and money!) to collect more and more samples.

Now that you appreciate the amount of work it takes to estimate the proportion of water from 20 samples, JH and SB think it is sufficient for you to use some automatic procedures to complete this task, which are further described in step 2.

2. Create an R script and copy the following index vector into it:

```
index.n.20 <- c(2131,2132,2133,2134,2135,2136,2137,2138,2139,2140,
2141,2142,2143,2144,2145,2146,2147,2148,2149,2150)
```

3. Load a function into your R session that JH and SB created to automate the process using the following command:

```
source("https://github.com/sahirbhatnagar/EPIB607/raw/master/exercises/water/automate_water_task.R")
```

4. Now that the function `automate_water_task` has been loaded into your environment, you can use it to automatically determine which of the locations in your sample of 20 are on water. This function requires the unique index vector shown in Step 2 above as input and returns a binary vector of length 20 (1 for water, 0 for land). You can call the function as follows in R:

```
landed_in_water <- automate_water_task(index = index.n.20)
```

As before, enter your estimate of the proportion of the earth's surface covered by water in this Google sheet next to your name, but in the column titled `PropnW.20.locations`.

2. What is the average depth of the ocean?

We will now turn to estimating the average depth of the ocean. You will again make use of the `automate_water_task` function.

2.1. Determine the average depth of the ocean from the sample of 5.

3. Plot the sampling distribution of the proportion and mean

```
# read in the results from the Google sheet
m <- read.csv("~/Dropbox/mcgill/teaching/EPIB607/Link to Jim2018/Water/filled.sheet.csv",
              as.is=TRUE)

# count the number of students who provided a mean and proportion
N.r <- nrow(m)

par(mfrow=c(2,2),mar = c(4,4,2,1))
plot(table(m[,3]),xlim=c(0,1),
     xlab="Students' Estimates of Proportion Covered by Water",
     main="n=5",ylim=c(0,N.r/1.5),ylab="Frequency")

d.BREAKS=seq(1000,6000,500)
hist(m[,5],xlim=c(0,6000),
     ylim=c(0, N.r/1.5), breaks=d.BREAKS,
     xlab="Students' Estimates of Mean Ocean Depth (m)",
     main="n=5")
```

```r
plot(table(m[,4]),xlim=c(0,1),
     main="n=20",ylim=c(0,N.r/2),ylab="Frequency")

hist(m[,6],xlim=c(0,6000),
     ylim=c(0, N.r/1.5),breaks=d.BREAKS,
     main="n=20",xlab="")

library(mosaic)
library(tidyr)
m.melt <- m %>% tidyr::gather(key = "type", value = "value", -X, -Student)
m.melt.means <- subset(m.melt, type %in% c("Mean.25.depths","Mean.5.depths"))
m.melt.props <- subset(m.melt, type %in% c("PropnW.25.locations","PropnW.5.locations"))
gf_density(~ value, data = m.melt.means, fill = ~ type) + theme_bw()
gf_histogram(~ value, data = m.melt.props, fill = ~ type, position = "dodge") + theme_bw()
```