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] "Cho, Minsoo "
```

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

```
[1] "minsoo.cho@mail.mcgill.ca"
```

A sample of 5.

```
Latitude.n.5 <- c(-26.237,-54.927,12.057,82.539,70.009)
```

```
Longitude.n.5 <- c(-115.959,-0.298,151.502,44.736,12.873)
```

A sample of 20.

```
Latitude.n.20 <- c(-50.847,-16.098,-32.15,-63.081,62.072,28.384,3.14,38.064,-38.406,42.467,
42.521,-45.406,-9.08,-78.756,29.575,-4.618,20.456,-27.836,32.37,-49.45)
```

```
Longitude.n.20 <- c(-98.168,140.487,-117.061,132.87,18.162,-87.788,-93.103,-6.614,177.316,-90.846,
158.807,-91.242,-61.646,-80.135,-72.092,52.206,-96.787,-41.55,156.635,-11.911)
```

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.

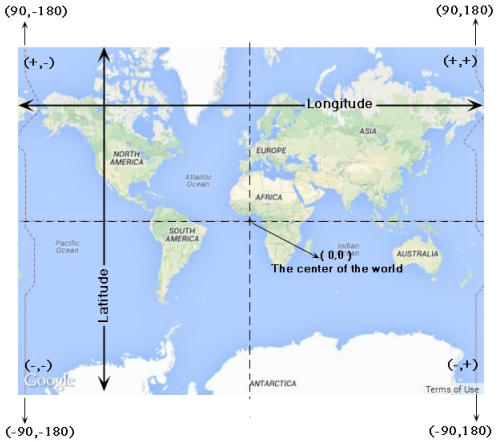

Fig. 1. Latitude and longitude in Google maps. Latitudes range from (-90,90) and longitudes range from (-180, 180). Latitude is entered first followed by longitude and separated by a comma.

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(331,332,333,334,335,336,337,338,339,340,
341,342,343,344,345,346,347,348,349,350)
```

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 and 20.

1. In the same R script as before, copy the following index vector into it:

```
index.n.5 <- c(326,327,328,329,330)
```

2. Use the `automate_water_task` function to get a sample of 5 depths. Note: some of the returned samples will not correspond to the same latitudes and longitudes provided to you earlier. This is because we need to restict our sample to locations on water only in order to estimate the mean depth of the ocean. Here we show some example code and its output. You need to specify the `type` and `student_id` argument:

```
# be sure to provide your own student id
depths.n.5 <- automate_water_task(index = index.n.5, student_id = 222333444, type = "depth")
```

The `alt` column gives the depth in meters:

	X	lon	lat	alt	water
3	3	-134.38257	37.742717	5028	1
4	4	-23.62332	-12.237161	5358	1
17391	17391	-124.01387	11.596521	4577	1
46573	46573	-167.19004	-9.532013	3850	1
23312	23312	83.01109	7.885076	3895	1

3. Repeat step 2 for a sample of 20 using the `index.n.20` vector specified above.
4. Calculate an estimate of the mean depth of the ocean from your samples of 5 and 20 using the `mean` function, e.g., `mean(depths.n.5$alt)`.
5. Enter your estimates of the mean depth of the ocean from your samples of 5 and 20 in this Google sheet next to your name, in the columns titled `Mean.5.depths` and `Mean.20.depths`, respectively.

3. Plot the sampling distribution of the proportion and mean

It is now time to plot the sampling distributions of the proportions and means. Once everyone has filled in the Google sheet, export the sheet as a `.csv` file by clicking on `File --> Download as --> Comma-separated values (.csv, current sheet)`.

1. Read in the data:

```
# read in the results from the Google sheet
water_results <- read.csv("EPIB607_FALL2018_water_exercise - water.csv", as.is=TRUE)

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

2. Plot the students' estimates of the proportion covered by water for samples of size 5. You may use the following code or run your own:

```
plot(table(water_results[,"PropnW.5.locations"]),
     xlim = c(0,1),
     xlab = "Students' Estimates of Proportion Covered by Water",
     main = "n = 5",
     ylim = c(0, N.r/1.5),
     ylab = "Frequency")
```

a) Comment on this graph. Does this shape look sensible to you?

3. Now plot the students' estimates of the mean depth of the ocean for samples of size 5. You may use the following code or run your own:

```
d.BREAKS <- seq(1000,6000,500)
hist(water_results[,"Mean.5.depths"],
     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")
```

a) Calculate the mean and the standard error of the mean depth for samples of size 5
b) Comment on this graph (e.g. range, variability)

4. Repeat Steps 2 and 3 for samples of size 20.

a) Compare the two graphs for proportions, and the two graphs for means. What do you notice? You might find it helpful to overlay the distributions on the sample plot. You may use the following code or run your own:

```r
library(mosaic)
library(tidyr)

# first 'melt' the data to get it in plotting form
m.melt <- water_results %>% tidyr::gather(key = "type", value = "value", -ID, -student)

# subset for means
m.melt.means <- subset(m.melt, type %in% c("Mean.20.depths","Mean.5.depths"))

# plot for means
gf_density(~ value, data = m.melt.means, fill = ~ type) + theme_bw()

# subset for proportions
m.melt.props <- subset(m.melt, type %in% c("PropnW.20.locations","PropnW.5.locations"))

# plot for proportions
gf_histogram(~ value, data = m.melt.props, fill = ~ type, position = "dodge") + theme_bw()
```