Assignment 2 - Histograms, Medians, Means, Boxplots and Standard Deviation. Due September 21, 11:59pm 2018

EPIB607 - Inferential Statistics[a]

[a]Fall 2018, McGill University

This version was compiled on September 11, 2018

The first step in understanding data is to hear what the data say, to "let the data speak for themselves". Numbers speak clearly only when we help them speak by organizing, displaying, and summarizing. In this assignment you will explore how to visualize your data and summarize it using summary statistics. All graphs and calculations are to be completed in an R Markdown document using the provided template. Please submit both the compiled HTML report and the source file (.Rmd) to myCourses by September 21, 2018, 11:59pm. Both HTML and .Rmd files should be saved as 'IDnumber_LastName_FirstName_EPIB607_A2'.

Histograms | Means | Medians | Boxplots | Standard Deviation | Summary statistics | mosaic package

Template

The `.Rmd` template for Assignment 2 is available here

The `mosaic` package (optional)

The `mosaic` package provides a consistent and user-friendly interface for descriptive statistics, plots and inference. You may find it useful to complete an interactive tutorial on its plotting functions. (note: this is optional and will not be counted for any marks). First install the following packages:

```
install.packages(c("learnr","mosaic"), dependencies = TRUE)
```

Then, from RStudio, run the following command which will open a new page in your web broswer:

```
learnr::run_tutorial("introduction", package = "ggformula")
```

An advanced tutorial on customizing plots is available also:

```
learnr::run_tutorial("refining", package = "ggformula")
```

1. Age-structures of Populations, then and now

The 1911 census of Ireland was taken on April 2nd 1911 and was released to the public in 1961. Follow this link for further details on the census. James Hanley (JH) has scrapped the data for Dublin, collected the age-frequency distribtion by gender and provided you with a three column .csv file which looks like this:

```
cens <- read.csv("https://github.com/sahirbhatnagar/EPIB607/raw/master/data/age_sex_frequencies_ireland.csv")
head(cens)
```

```
#    Gender Age Freq
#  1   Male   0 5332
#  2   Male   1 4570
#  3   Male   2 4979
#  4   Male   3 4789
#  5   Male   4 4884
#  6   Male   5 4787
```

The `Age` column represents the age in 1911. The `Freq` column gives the frequency of the number of people for a given age and `Gender`.

a) What was the ealiest year of birth for (i) males and (ii) females ?
b) Create a suitable visulaization of this data and then comment on any patterns you see and give reasons for these patterns. Your choice should leverage all the information provided in the data and be influeced by the message that you are trying to convey.
c) Calculate the mean age, the standard deviation (SD), and the quartiles: $Q_{25}, Q_{50}(median), Q_{75}$ separately for males and females.

2. Number of Authors

Fletcher et al. (1979, NEJM 301:180-183) studied the characteristics of 612 randomly selected articles published in the NEJM, JAMA, and the Lancet since 1946. Two characteristics examined were the number of authors per article and the number of students studied in each article:

Year	No. articles studied	No. Authors Mean (SD)	No. subjects Median
1946	151	2.0 (1.4)	25
1956	149	2.3 (1.6)	36
1966	157	2.8 (1.2)	16
1976	155	4.9 (7.3)	30

a) Why report median number of subjects (instead of average)?
b) In 1976, can the standard deviation of 7.3 really be larger than the mean of 4.9? Explain.
c) In light of (a) and (b), how would you present the data in this table?

3. Cancer Deaths in the US

The American Cancer Society (ACS) recently published age-adjusted cancer death rates by cancer site and gender at this link.

a) In the figure Trends in Age-adjusted Cancer Death Rates by Site, Males, US, 1930-2014 (PDF) list the scales being used and the variable that has been mapped onto them.
b) Briefly comment on what you like about the figure, and what could have been improved (e.g. aesthetics, labels, use of color)
c) The data used to make the figures has also been provided on the ACS website in Excel spreasheets: [males], [females]). Once you have downloaded the spreadsheets, the following code can be used to combine both datasets:

```
library(dplyr)
males <- readxl::read_xlsx("~/Downloads/age-adjusted-cancer-death-rates-males-1930-2014.xlsx",
                           skip = 1, n_max = 85) %>% mutate(Gender = "Male")
females <- readxl::read_xlsx("~/Downloads/age-adjusted-cancer-death-rates-females-1930-2014.xlsx",
                             skip = 1, n_max = 85) %>% mutate(Gender = "Female")
cancer_rates <- dplyr::bind_rows(list(females,males))
head(cancer_rates)
```

```
#  # A tibble: 6 x 13
#     Year Stomach `Colon and Rect~ `Pancreas‡` `Lung and Bronc~ Breast
#    <dbl>   <dbl>            <dbl>       <dbl>            <dbl>  <dbl>
#  1  1930    35.2             27.1        3.82             2.58   30.1
#  2  1931    33.8             27.7        4.12             2.63   30.6
#  3  1932    33.7             28.5        4.50             2.84   30.9
#  4  1933    32.5             28.7        4.27             2.96   30.8
#  5  1934    31.6             29.7        4.36             3.08   31.6
#  6  1935    31.4             30.2        4.67             3.54   31.3
#  # ... with 7 more variables: `Uterus†` <dbl>, `Liver‡` <dbl>,
#  #   Gender <chr>, `Liver†` <dbl>, `Pancreas†` <dbl>, Prostate <dbl>,
#  #   Leukemia <dbl>
```

In order to make the data ready for plotting, we need to *melt* it first:

```
library(tidyr)
cancer_rates_melt <- tidyr::gather(cancer_rates, key = "Site", value = "Rate", -Year, -Gender)
head(cancer_rates_melt)
```

```
#  # A tibble: 6 x 4
#     Year Gender Site     Rate
#    <dbl> <chr>  <chr>   <dbl>
#  1  1930 Female Stomach  35.2
#  2  1931 Female Stomach  33.8
#  3  1932 Female Stomach  33.7
#  4  1933 Female Stomach  32.5
#  5  1934 Female Stomach  31.6
```

```
#  6  1935 Female Stomach  31.4
```

Plot the data using a graph of your choice with the goal of comparing males vs. females. (you might consider using the function `gf_line` from the `ggformula` package). Briefly comment on your comparison.