ECON 3640–001

Problem Set 1

Marcio Santetti

Spring 2022

Problem 1

With your own words, define and give an example of each of the following statistical terms.

(a) Population;

(b) Sample;

(c) Parameter;

(d) Statistic.

Problem 2

A student received the following letter grades on the 14 quizzes she took during a semester: *A*, *A-*, *B+*, *A-*, *A*, *C*, *A*, *A*, *A-*, *B*, *B+*, *C*, *A*, and *A*.

(a) What is the `type` of these data? Explain.

(b) Use your best judgment to *illustrate* these data.

Problem 3

A sample of 12 people was asked how much change they had in their pockets and wallets. The responses (in cents) were:

52 25 15 0 104 44 60 30 33 81 40 5

(a) Determine the *mean*, *median* and *mode* for these data. Show your calculations.

(b) Determine the first, second, and third *quartiles* of these data.

(c) Compute the *variance* and *standard deviation* for these data. Show your calculations.

(d) Draw a *box plot* for these data.

Problem 4

From the `wooldridge` package, load the `mroz` data set. The data come from `Mroz (1987)`. Just as with any data set from this package, its documentation contains all variable names and descriptions. Simply google "wooldridge R package" and you will find it.

First, transform it into a `tibble`. Just follow the code below:

```
library(tidyverse)
library(wooldridge)

data("mroz")

mroz_tibble <- mroz %>% as_tibble()
```

Then, answer the following questions:

(a) What is the mode (i.e., most frequent) educational attainment— in years —in this sample?

(b) Select the first five rows of this data set and manually compute the covariance, correlation coefficient, and coefficient of determination (R^2) between educational attainment (`educ`) and hourly earnings (`wage`). Interpret these results. **Hint**: `head(5)`.

(c) Use the most appropriate visual descriptive technique to illustrate the association between the two variables from part (b). Do not forget to make your plot informative to a wide audience (i.e., label your axes and give it a nice title).

Problem 5

Still using the `mroz` data set, run the following:

```
mroz_tibble %>%
  select(inlf, hours, huswage) %>%
  head()
```

```
## # A tibble: 6 × 3
##    inlf hours huswage
##   <int> <int>   <dbl>
## 1     1  1610    4.03
## 2     1  1656    8.44
## 3     1  1980    3.58
## 4     1   456    3.54
## 5     1  1568   10
## 6     1  2032    6.71
```

Notice that the `inlf` variable is defined here as an integer (`int`), but what it is actually doing is serving as a *binary* indicator, which equals 1 if the woman interviewed is in the labor force, and 0 if not. Thus, if we want to use this variable for the upcoming plot, we should transform it into a factor (`fct`) class object.

Just do the following:

```
mroz_tibble <- mroz_tibble %>%
  mutate(inlf = as_factor(inlf))
```

and check out whether the variable is now a factor. The `as_factor()` function is part of the `tidyverse`.

(a) Draw a histogram of husband wages (`huswage`), comparing the difference between whether the interviewee is in the labor force or not. **Hint**: you may either use the `fill` argument within the `aes()` environment, or use the `facet_wrap()` function to do that. Interpret your results.

(b) In this sample, how many interviewees are in the labor force? How many aren't?

(c) From your answer to part (b), illustrate the relative frequencies (i.e., %) for each case either with a bar or pie chart.

Problem 6

During these pandemic times, you have probably come across the `moving average` term. It simply consists of calculating a mean that is adjusted over a specified time window. For instance, a 7-day moving (or rolling) average computes the mean value of a variable over the previous 7 days, and it gets adjusted as time moves on.

This application allows us to smooth out short-term oscillations in a data set. Applying this in R is very simple, and we'll get there soon.

(a) First, import the `covid-cases-22.csv` data set into your R environment. Call it `covid_cases`.

(b) Now, we need to convert the `period` column into a `date` object. It will be imported as a character (`chr`). You need to use the `lubridate` package, specifically built to deal with dates and times. Check out the code below:

```
library(lubridate) # make sure to have it installed first.

covid_cases <- covid_cases %>%
  mutate(period = mdy(period))
```

The `mdy` function simply converts a character string defined by `day-month-year` into a `date` object.

(c) Find out how to plot the `new_cases` variable over time using `ggplot2`.

(d) To calculate moving averages, one may use the `RcppRoll` package. Its `roll_mean()` function does the job. So, create a new column in your data set, called `new_cases_ma`, defined by the 14-day moving average for the `new_cases` series. You will use the `roll_mean()` function and 3 arguments: `n`, which is the number of days you want your moving-average window to be; `align`, which you will set equal to "right"; and `fill`, which you will set to `NA`. The latter guarantees that the first values (for which you will not be able to calculate the moving average) will be filled out with `NA` values.

(e) Lastly, plot this new variable from part (d) over time and compare it with your plot from part (c).