HW03 - Correlation

Stat 131A, Fall 2018

Due Sep-17

General Instructions

- This assignment assumes that you have read the tutorials `09-scatter-diagrams.pdf` and `10-correlation.pdf` available in the course's github repository: https://github.com/ucb-introstat/introstat-fall-2018/tree/master/tutorials
- Write your narrative and code in an `Rmd` (R markdown) file.
- Name this file as `hw03-first-last.Rmd`, where `first` and `last` are your first and last names (e.g. `hw03-gaston-sanchez.Rmd`).
- Please do not use code chunk options such as: `echo = FALSE`, `eval = FALSE`, `results = 'hide'`. All chunks must be visible and evaluated.
- Submit your Rmd and html files to bCourses.
- If you have questions/problems, don't hesitate to ask us for help in OH. Also, make use of piazza and seek advice from your peers.

---

1) The following lines of code define three pairs of x, y values:

```
x1 = c(1, 1, 1, 1, 2, 2, 2, 3, 3, 4)
y1 = c(5, 3, 5, 7, 3, 3, 1, 1, 1, 1)

x2 = c(1, 1, 1, 1, 2, 2, 2, 3, 3, 4)
y2 = c(1, 2, 1, 3, 1, 4, 1, 2, 2, 3)

x3 = c(1, 1, 1, 1, 2, 2, 2, 3, 3, 4)
y3 = c(2, 2, 2, 2, 4, 4, 4, 6, 6, 8)
```

For each pair of variables, use R, showing your code, to:

1. plot a scatter diagram, and
2. find the correlation coefficient using the procedure described in SticiGui (chapter 8: computing the correlation coefficient):

   Convert each variable to standard units. The average of the products gives the correlation coefficients.

   r = average of (x in standard units) $\times$ (y in standard units)

Note: to convert to standard units, you need to obtain SD (not SD^+). If you decide to use the function `sd()` you will need to adjust its output so that n is used instead of $n - 1$. You can use the function `cor()` only to verify that your computations are correct.

2) Consider the following two data sets shown below.

```
# set 1
x1 = c(1, 2, 3, 4, 5, 6, 7)
y1 = c(2, 1, 4, 3, 7, 5, 6)

# set 2
x2 = c(1, 2, 3, 4, 5, 6, 7)
y2 = c(5, 4, 7, 6, 10, 8, 9)
```

A computer program prints out r for the two data sets shown above: `cor(x1, y1) = 0.8214`, and `cor(x2, y2) = 0.7619`. Is the program working correctly? Answer yes or no, and explain briefly.

3) A number if missing in each of the data sets below. If possible, fill in the blank to make r equal to 1. If this is not possible, say why not.

```
# set a)
x1 = c(1, 2, 3, 4)
y1 = c(1, 3, 3, ?)

# set b)
x2 = c(1, 2, 3, 4)
y2 = c(1, 3, 4, ?)
```

4) The table below shows per capita consumption of cigarettes in various countries in 1930. and the death rates from lung cancer for men in 1950. (In 1930, hardly any women smoked; and a long period of time is needed for the effects of smoking to show up.)

country	consumption	deaths
Australia	480.00	180.00
Canada	500.00	150.00
Denmark	380.00	170.00
Finland	1100.00	350.00
Great Britain	1100.00	460.00
Iceland	230.00	60.00
Netherlands	490.00	240.00
Norway	250.00	90.00
Sweden	300.00	110.00
Switzerland	510.00	250.00
USA	1300.00	200.00

- `consumption` = cigarette consumption
- `deaths` = deaths per million

a. Plot a scatter diagram for these data.

b. True or False. The higher cigarette consumption was in 1930 in one of these countries, on the whole the highet the death rate from lung cancer in 1950. Or can this be determined from the data?

c. True or False. Death rates from lung cancer tend to be higher among those persons whi smoke more. Or can this be determined from the data?