ENGRD 2700: Basic Engineering Probability and Statistics
Fall 2019

Homework 1 Solutions

Due Friday, September 13 by 11:59 pm. Submit to Gradescope by clicking the name of the assignment. See `https://people.orie.cornell.edu/yudong.chen/engrd2700_2019fa.html#homework` for detailed submission instructions.

When completing this assignment (and all subsequent ones), keep in mind the following:

- You must complete the homework individually and independently.
- Provide evidence for each of your answers. If a calculation involves only very minor computation then explain the computation you performed and give the results. If a calculation involves more complicated steps on many many records then hand in the calculations and formulas for the first few records only.
- Write clearly and legibly. You are encouraged to *type* your work although you do not have to. We may deduct points if your answers are difficult to read or disorganized.
- For questions that you answer using R, attach any code that you write, along with the relevant plots. You may use other software, but the same condition applies.
- Submit your homework a single pdf file on Gradescope.

1. The file `Quartet.csv` contains four datasets of x and y values, side by side.

   Import the dataset either using the GUI (the "Heading" option should be turned on), or through the command `quartet = read.csv("Quartet.csv")`. Attach the data frame (so that columns can be accessed by their names) with the command `attach(quartet)`.

   (a) Compute the sample mean, sample median, and sample standard deviation for each column of the dataset.

   Below is the result rounded to two decimal places.

	x_1	x_2	x_3	x_4	y_1	y_2	y_3	y_4
Mean	9.00	9.00	9.00	9.00	7.50	7.50	7.50	7.50
Median	9.00	9.00	9.00	8.00	7.58	8.14	7.11	7.04
St. Dev	3.32	3.32	3.32	3.32	2.03	2.03	2.03	2.03

   (b) Based solely upon the summary statistics you computed in part (a), how do the four datasets compare?

   All four x-vectors have identical sample means and sample variances, which are 9 and 11, respectively. The y-vectors have sample means and sample variances which match to two decimal places (7.501 and 4.123, respectively). The medians of the y-vectors are roughly equal, with the exception of $y2$ (slightly higher). A similar statement can be made for the medians of the x-vectors, with the possible exception of $x4$ (slightly lower). Based on the information from part (a) alone, the datasets appear to be very similar.

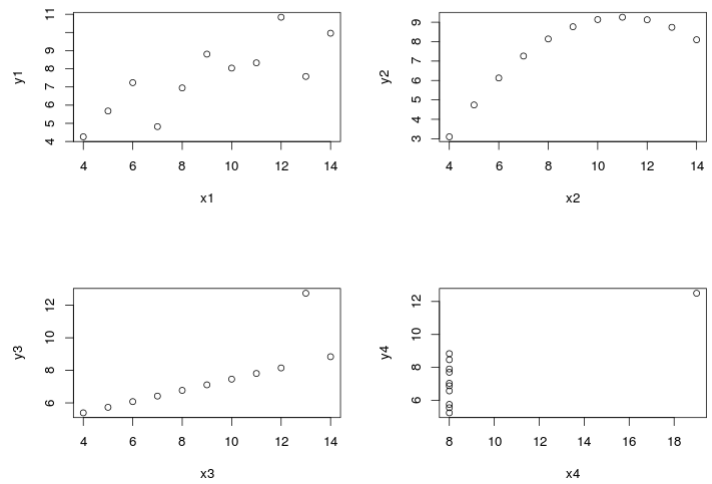
   (c) Construct scatterplots for each of the four datasets. (Hint: In R, you can use the command `par(mfrow=c(2,2))` to combine multiple plots into a single 2-by-2 graph in R. If you do, this command should precede any code that you use to generate plots.)

   We generate the following plots using the `R` code:

```
par(mfrow=c(2,2))
plot(x1,y1)
```

```
plot(x2,y2)
plot(x3,y3)
plot(x4,y4)
```

and get the following:



(d) Based solely upon the plots you generated in part (c), how do the four datasets compare?

The distribution of points varies across datasets. This is most clearly illustrated in dataset 4. While the other datasets have points more or less evenly distributed across the x-axis, dataset 4 has points concentrated on a vertical line segment at $x = 8$ with a single outlier.

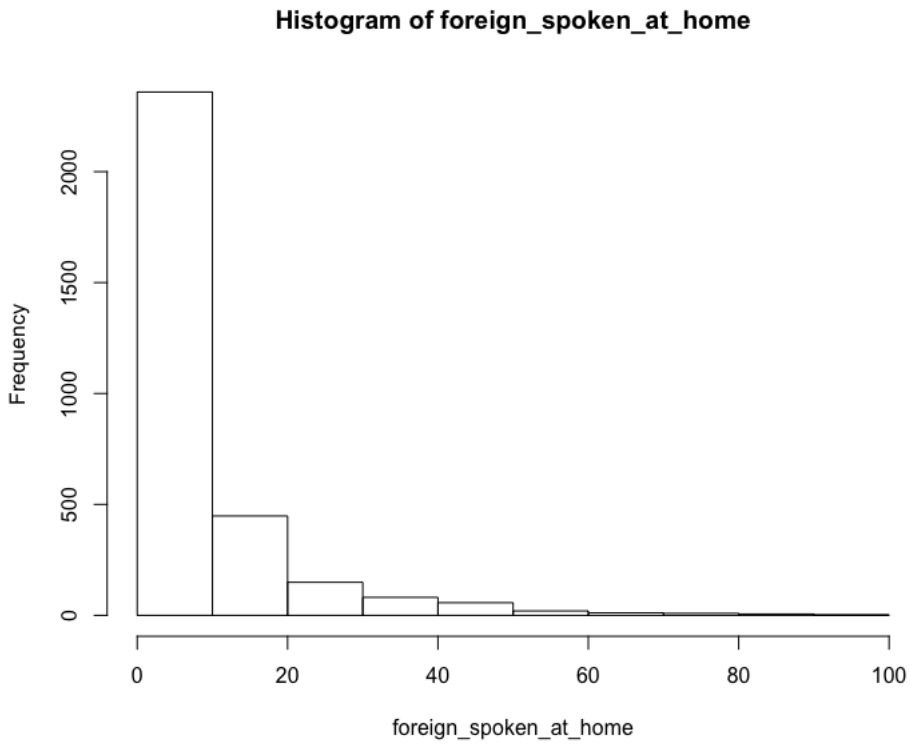
(e) What's the moral of the story? (That is, what does this example suggest about what should be done when analyzing data?)

The dataset demonstrates that data sometimes cannot be adequately described using summary statistics (such as the sample mean and the sample variance), particularly when attempting to understand the relationship between two variables. The four plots from part (b) arose from vectors that, statistically, appeared very similar, but exhibited vastly different statistical trends. We would have only seen this by plotting the data.

2. Answer the questions below about the dataset `CountyData.csv` from the U.S. Census Bureau, performing any data analysis you deem appropriate. The dataset consists of 3143 observations on 53 variables, which are described in the file `CountyData_Info.pdf`.

(a) Provide a histogram of the per-county percentage of residents who speak a foreign language at home during 2006-2010.

Using the `R` command `hist(foreign_spoken_at_home)` gives us the following histogram:



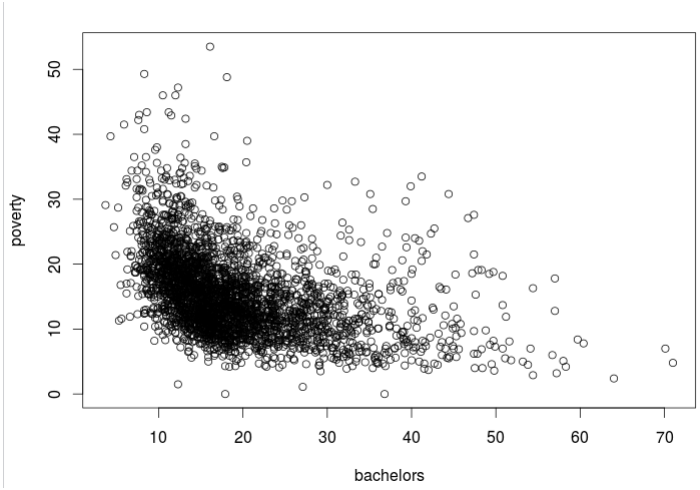
(b) What was the median per-county amount of federal spending in 2009?

Just need to notice that there is 4 data not available for federal spending in a county in 2009. Then we can easily have that the median is 214994.
The `R` code is as follows.

```
median(fed_spending, na.rm = "TRUE")
```

(c) Create a scatter plot of the percentage of residents below the poverty level (y-axis) versus the percentage of the population with a bachelors degree. Comment on what you see.

Using the command `plot(bachelors,poverty)`, we get the following histogram:



We observe that there seems to be a negative correlation between percentage with a bachelors degree and percentage below the poverty level.

(d) What fraction of counties have a population whose percentage under the age of 18 is above 30%?

There are 114 counties having more than 30% people under 18 years old, while the total number of counties is 3143. So $\frac{114}{3143} = 3.63\%$ of the counties have a population whose percentage under the age of 18 is above 30%.
The `R` code is as follows.

```
nrow(subset(CountyData,age_under_18 > 30.0))/length(age_under_18)
```

3. A subdivision of 24 houses has a mean price of \$500,000, a median of \$440,000, and a standard deviation of \$30,000. A new house is then built in the subdivision that has a price of \$700,000.

(a) What is the new mean house price?

Since we have that

$$\begin{aligned}
\bar{x}_{n+1} &= \frac{1}{n+1}\sum_{i=1}^{n+1} x_i \\
&= \frac{1}{n+1}\sum_{i=1}^{n} x_i + \frac{x_{n+1}}{n+1} \\
&= \frac{n}{n+1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) + \frac{x_{n+1}}{n+1} \\
&= \frac{n}{n+1}\bar{x}_n + \frac{x_{n+1}}{n+1}
\end{aligned}$$

The new mean house price is

$$\bar{x}_{25} = \frac{24}{25}(500{,}000) + \frac{1}{25}(700{,}000) = 508{,}000.$$

(b) What is the new standard deviation?

Rewriting s^2_{n+1} in terms of s^2_n, we find:

$$\begin{aligned}
s^2_{n+1} &= \frac{1}{n}\sum_{i=1}^{n+1}(x_i - \bar{x}_{n+1})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_{n+1})^2 + \frac{(x_{n+1} - \bar{x}_{n+1})^2}{n} \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n + \bar{x}_n - \bar{x}_{n+1})^2 + \frac{(x_{n+1} - \bar{x}_{n+1})^2}{n} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[(x_i - \bar{x}_n)^2 + 2(x_i - \bar{x}_n)(\bar{x}_n - \bar{x}_{n+1}) + (\bar{x}_n - \bar{x}_{n+1})^2\right] + \frac{(x_{n+1} - \bar{x}_{n+1})^2}{n} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[(x_i - \bar{x}_n)^2 + (\bar{x}_n - \bar{x}_{n+1})^2\right] + \frac{(x_{n+1} - \bar{x}_{n+1})^2}{n} \\
&= \frac{n-1}{n}s^2_n + \frac{1}{n}\sum_{i=1}^{n}(\bar{x}_n - \bar{x}_{n+1})^2 + \frac{(x_{n+1} - \bar{x}_{n+1})^2}{n} \\
&= \frac{n-1}{n}s^2_n + (\bar{x}_n - \bar{x}_{n+1})^2 + \frac{(x_{n+1} - \bar{x}_{n+1})^2}{n}.
\end{aligned}$$

where the fifth line follows because

$$\sum_{i=1}^{n}(x_i - \bar{x}_n) = \sum_{i=1}^{n} x_i - n\bar{x}_n = 0.$$

The new sample standard deviation is

$$\begin{aligned} s_{25} &= \sqrt{\frac{23}{24}(30{,}000)^2 + (500{,}000 \text{ - } 508{,}000)^2 + \frac{(700{,}000 - 508{,}000)^2}{24}} \\ &\approx 49{,}623.58. \end{aligned}$$

(c) Does the median increase, decrease, or stay the same after the new house is built? Or can no conclusion be made? Explain.

No conclusion can be made. The median price is currently given by the average of 12^{th} least expensive house and 13^{th} least expensive house in the subdivision, which we'll call P_{12} and P_{13} respectively. Since the new house has a price larger than the median, the new median would be P_{13}. If $P_{12} = P_{13}$, the median would not change. However, if $P_{12} < P_{13}$, then the median would increase.

4. Consider a data sample $x_1, x_2, \ldots, x_n$. Let $\bar{x}$ and s_x^2 denote its sample mean and sample variance.

(a) Suppose that you modify these data by adding a constant c to each observation in the sample, and then multiplying by another constant k, to obtain a modified sample $y_1, \ldots, y_n$. (That is, $y_i = (x_i + c) \times k$ for each i.)

What are $\bar{y}$ and s_y^2, the sample mean and variance of the modified data? Justify your answer mathematically, using the definition of the sample mean and variance.

For the sample mean, we have

$$\begin{aligned} \bar{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[(x_i + c) \times k\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}(kx_i + ck) \\ &= k\bar{x} + ck \end{aligned}$$

so the sample mean gets multiplied by k and shifted by ck. For the sample variance,

$$\begin{aligned} s_y^2 &= \frac{1}{n-1}\sum_{j=1}^{n}(y_j - \bar{y})^2 \\ &= \frac{1}{n-1}\sum_{j=1}^{n}\left[(x_j + c) \times k - (k\bar{x} + ck)\right]^2 \\ &= \frac{1}{n-1}\sum_{j=1}^{n}\left[k(x_j + c - \bar{x} - c)\right]^2 \\ &= k^2\frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2 = k^2 s_x^2, \end{aligned}$$

so the sample variance gets multiplied by k^2.

(b) Finally, suppose we built another modified data sample $z_1, \ldots, z_n$, where

$$z_i = \frac{x_i - \bar{x}}{s_x},$$

where s_x is the sample standard deviation of the x–data. This procedure *standardizes* the original data. What are $\bar{z}$ and s_z^2? Justify your answer.

Using part (a), if we substitute c with $-\bar{x}$, and k with $\frac{1}{s_x}$, we get,

$$\bar{z} = \frac{1}{s_x}\bar{x} - \bar{x}\frac{1}{s_x} = 0,$$
$$s_z^2 = (\frac{1}{s_x})^2 s_x^2 = 1.$$