Homework #2 Solutions APPM 4570/5570, Statistical Methods, Fall 2017

Due in class on Friday September 15, 2017. Covers exploratory data analysis and intro to probability. *Instructions for "theoretical" questions: Answer all of the following questions. The theoretical problems should be neatly numbered, written out, and solved. Please do not turn in messy work. Working in small groups is allowed, but it is important that you make an effort to master the material and hand in your own work.*

Theoretical Questions

1. (a) Let a and b be constants and let $y_i = ax + b$ for $i = 1, ..., n$. What are the relationships between $\bar{x}$ and $\bar{y}$ and between s_x^2 and s_y^2?

Given the equation for the mean of x,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

the mean for y is,

$$\begin{aligned}\bar{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i \\ &= \frac{1}{n}\sum_{i=1}^{n}(a \cdot x_i + b) \\ &= a\frac{1}{n}\sum_{i=1}^{n} x_i + b \\ &= a\bar{x} + b\end{aligned}$$

Thus, the relationship between the means is that $\bar{y}$ is a multiple of $\bar{x}$ with an offset.

Given the equation for the variance of x,

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

the variance for y is,

$$\begin{aligned}
s_y^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(a \cdot x_i + b - (a \cdot \bar{x} + b))^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(a \cdot x_i - a \cdot \bar{x})^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(a(x_i - \bar{x}))^2 \\
&= a^2(\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2) \\
&= a^2 s_x^2.
\end{aligned}$$

(b) A sample of temperatures for initiating a certain chemical reaction yielded a sample average of 87.3 degrees Celsius and a sample standard deviation of 1.04. What are the sample average and standard deviation measured in Fahrenheit [HINT: F = 9/5C + 32]?

Given the results from the previous part and that,

$$\begin{aligned}
F &= 9/5C + 32 \\
\bar{C} &= 87.3 \\
s_C &= 1.04
\end{aligned}$$

the sample average in Fahrenheit is calculated as,

$$\begin{aligned}
\bar{F} &= 9/5 \cdot \bar{C} + 32 \\
&= 9/5 \cdot 87.3 + 32 \\
&= \underline{189.14}
\end{aligned}$$

using the variance, the sample standard deviation in Fahrenheit is calculated as,

$$\begin{aligned}
s_F^2 &= (9/5)^2 \cdot \sigma_C^2 \\
&= (9/5)^2 \cdot 1.04^2 \\
&= 3.504384 \\
s_F &= \sqrt{s_F^2} = \sqrt{3.504384} = \underline{1.872}
\end{aligned}$$

2. Let $\bar{x}_n$ and s_n^2 denote the sample mean and variance for the sample $x_1, ..., x_n$ and let $\bar{x}_{n+1}$ and s_{n+1}^2 denote these quantities when an additional observation x_{n+1} is added to the sample.

(a) Show that $\bar{x}_{n+1} = \bar{x}_n + \dfrac{x_{n+1} - \bar{x}_n}{n+1}$.

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

(b) Show that $s_{n+1}^2 = \frac{(n-1)}{n} s_n^2 + \frac{1}{n+1}(x_{n+1} - \bar{x}_n)^2$.

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

Evaluating the (n+1)-th term,

$$\begin{aligned}
\left(x_{n+1} - \bar{x}_n + \frac{\bar{x}_n - x_{n+1}}{n+1}\right)^2 &= \left(\frac{nx_{n+1} - n\bar{x}_n}{n+1}\right)^2 \\
&= \frac{n^2}{(n+1)^2}(x_{n+1} - \bar{x}_n)^2
\end{aligned}$$

Going back to the initial equation,

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

$$= \sum_{i=1}^{n}\left[(x_i - \bar{x}_n)^2 + \left(\frac{\bar{x}_n - x_{n+1}}{n+1}\right)^2 + 2(x_i - \bar{x}_n)\left(\frac{x_n - x_{n+1}}{n+1}\right)(x_i - \bar{x}_n)\right] + \frac{n^2}{(n+1)^2}(x_{n+1} - \bar{x}_n)^2$$

$$= \underbrace{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}_{=(n-1)s_n^2} + \frac{n}{(n+1)^2}(\bar{x}_n - x_{n+1})^2 + 2\left(\frac{x_n - x_{n+1}}{n+1}\right)\underbrace{\sum_{i=1}^{n}(x_i - \bar{x}_n)}_{=o} + \frac{n^2}{(n+1)^2}(x_{n+1} - \bar{x}_n)^2$$

$$\begin{aligned}
&= (n-1)s_n^2 + \frac{n^2+n}{(n+1)^2}(x_{n+1} - \bar{x}_n)^2 \\
&= (n-1)s_n^2 + \frac{n(n+1)}{(n+1)^2}(x_{n+1} - \bar{x}_n)^2 \\
&= (n-1)s_n^2 + \frac{n}{n+1}(x_{n+1} - \bar{x}_n)^2 \\
\implies s_{n+1}^2 &= \frac{n-1}{n}s_n^2 + \frac{1}{n+1}(x_{n+1} - \bar{x}_n)^2
\end{aligned}$$

(c) Why might these results be useful?

These results are useful in that when a new term is added to the sample, we do not need to go through all the steps of recalculating the variance and mean. Instead, we can use these equalities to update the new variance and mean of the new sample set.

(d) In both (a) and (b), describe what happens as $n \to \infty$.

As $n \to \infty$, we find that,

$$\lim_{x\to\infty} \bar{x}_{n+1} = \lim_{x\to\infty}\left[\bar{x}_n + \frac{x_{n+1} - \bar{x}_n}{n+1}\right] = \bar{x}_n, \text{ and}$$

$$\lim_{x\to\infty} s_{n+1}^2 = \lim_{x\to\infty}\left[\frac{n-1}{n}s_n^2 + \frac{1}{n+1}(x_{n+1} - \bar{x}_n)^2\right] = s_n^2.$$

3. Three fair dice are thrown. What is the probability that a sum of 8 appears on the faces? What is the probability that a sum of 10 appears?

Given that three dice are thrown, the probability that a sum of 8 appears can be calculated by summing up the possibilities of rolling a 8 and dividing by the total number of possible outcomes. Looking at Table 1, there is a total of 21 possibilities of rolling a 8. The total number of possible outcomes is $6^3 = 216$. Thus the probability or rolling a 8 is $P = \frac{21}{216} \approx 0.0972$.

First Die	Second Die	Third Die
1	1	6
1	2	5
1	3	4
1	4	3
1	5	2
1	6	1
2	1	5
2	2	4
2	3	3
2	4	2
2	5	1
3	1	4
3	2	3
3	3	2
3	4	1
4	1	3
4	2	2
4	3	1
5	1	2
5	2	1
6	1	1

Table 1: Rolling 8 Possibilities

First Die	Second Die	Third Die
1	3	6
1	4	5
1	5	4
1	6	3
2	2	6
2	3	5
2	4	4
2	5	3
2	6	2
3	1	6
3	2	5
3	3	4
3	4	3
3	5	2
3	1	6
4	1	5
4	2	4
4	3	3
4	4	2
4	5	1
5	1	4
5	2	3
5	3	2
5	4	1
6	1	3
6	2	2
6	3	1

Table 2: Rolling 10 Possibilities

Likewise, the probability that a sum of 10 appears can be calculated by summing up the possibilities of rolling a 10 and dividing by the total number of possible outcomes. Looking at Table 2, there is a total of 27 possibilities of rolling a 10. The total number of possible outcomes is $6^3 = 216$. Thus the probability or rolling a 10 is $P = \frac{27}{216} = 0.125$.

4. Consider randomly selecting a student at CU Boulder. Let A denote the event that the selected student has a Venmo account and let B be the event that the selected student has a Paypal account. Suppose that $P(A) = 0.6$ and $P(B) = 0.5$.

   (a) Is it possible that $P(A \cap B) = 0.55$? Cite a theorem of probability in your answer.

   It is a theorem that $P(A \cap B) \leq P(B)$; but $P(B) = 0.5$. Therefore, $P(A \cap B) \neq 0.55$

   (b) For the remaining questions, let $P(A \cap B) = 0.3$. Compute the probability that the selected individual has at least one of the two types of accounts.

The probability that the selected individual has at least one of the two types of cards can be calculated by using the specific addition rule.

$$\begin{aligned} P(\text{`at least one'}) = P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= 0.6 + 0.5 - 0.3 \\ &= 0.8 \end{aligned}$$

(c) What is the probability that the selected individual has neither type of account?

The probability that the selected individual has neither type of card can be calculated by taking the complement of the answer from part B,

$$P = 1 - 0.8 = 0.2$$

(d) Describe in terms of A and B the event that the selected student has Venmo but not Paypal, and then calculate the probability of this event.

The event that the selected student has Venmo (A), but not Paypal (B), can be written as,

$$P(A \cap B^C)$$

Solving for this probability,

$$\begin{aligned} P(A \cap B^C) &= P(A) - P(A \cap B) \\ &= 0.6 - 0.3 \\ &= 0.3 \end{aligned}$$

5. (**APPM 5570 Only**) Prove that if one event A is contained in another event B (i.e., A is a subset of B) then $P(A) \leq P(B)$. [HINT: you might consider the set $B \cap A^c$ in your computation, where A^c is the complement of A.]

Let $A \subset B$. Note that $B = A \cup \left(A^c \cap B\right)$, and note that A and $A^c \cap B$ are disjoint. So, $P(B) = P(A) + P(A^cB)$. Since $P(A^cB) \geq 0$, it must be true that $P(B) \geq P(A)$.

Computational Questions

Instructions for "computational" questions: Your work should be neatly done and include all graphs, code, and comments, labeled and in order based on the problem you are addressing. Do not put graphs in at the end, stick code in random locations, or do anything else that will make this homework difficult to read and grade. If you turn in something that is messy or out of order, it will be returned to you with a zero. Working in small groups is allowed, but it is important that you make an effort to master the material and hand in your own work.

1. Some claim that the final hours aboard the Titanic were marked by class warfare; other claim it was characterized by male chivalry. The data frame `TITANIC3` from the `PASWR2` package contains information pertaining to class status `pclass`, survival of passengers `survived`, and gender `sex`, among others. Based on the information in the dataframe:

   (a) Determine the fraction of survivors from each passenger class.

   First Class = .62
   Second Class = .42
   Third Class = .26

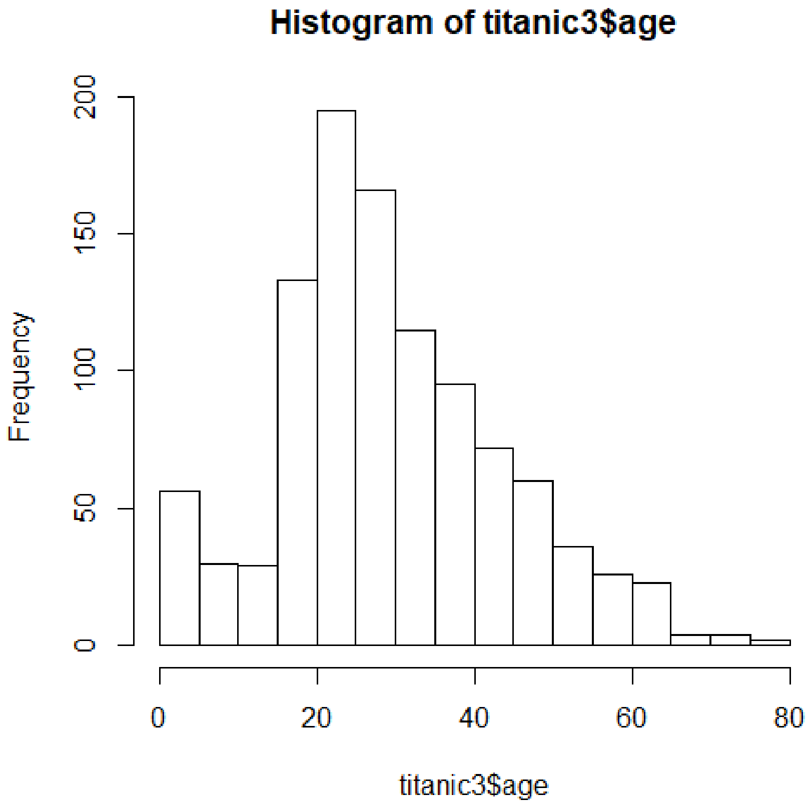
   (b) Compute the fraction of survivors according to class and gender. Did men in the first class or women in the third class have a higher survival rate?

   First Class Men = .34
   Second Class Men = .14
   Third Class Men = .15

   First Class Women= .96
   Second Class Women = .88
   Third Class Women = .49

   Comparing the fraction of survivors between men in the first class and women in the third class, men in the first class had a lower survival rate.

   (c) How would you characterize the distribution of `age` (e.g., is it symmetric, positively/negatively skewed, unimodal, multimodal)?



hist.png

The distribution appears to be unimodal and slightly positively skewed towards youngers ages.

(d) Were the median and mean ages for females who survived higher or lower than for females who did not survive? Report the median and mean ages as well as an appropriate measure of spread for each statistic.

Mean age of women survivors: 29.8
Median age of women survivors: 28.5
Mean age of women who died: 25.3
Median age of women who died: 24.5

Standard deviation of women survivors: 14.8
Standard deviation of women who died: 13.5
Range of women survivors: 2 months - 76 years old
Range of women who died: 1 - 63 years old

From this it is easy to see that the mean and median age of women who survived is higher than of those who died.

(e) Were the median and mean ages for males who survived higher or lower than for males who did not survive? Report the median and mean ages as well as an appropriate measure of spread for each statistic.

Mean age of men survivors: 27.0
Median age of men survivors: 27.0
Mean age of men who died: 31.5
Median age of men who died: 29.0

Standard deviation of men survivors: 15.6
Standard deviation of men who died: 13.8
Range of men survivors: 5 months - 80 years old
Range of men who died: 3 months - 74 years old

For men we see the opposite, that the mean and median age of men who survived is lower than of those who died.

(f) What was the age of the youngest female in the first class who survived?

The youngest first class female to survive was 14 years old.

(g) Do the data suggest that the final hours aboard the Titanic were characterized by class warfare, male chivalry, some combination of both, or neither? Justify your answer based on computations above, or based on other explorations of the data.

While the first class tended to have slightly higher rates of survival than the lower classes, the disparity of rates of survivors is much greater between genders. So, the finals hours on the Titanic had some class favoritism, but male chivalry was also very prominent.

2. Conduct a simulation in R to numerically illustrate the results from theoretical question 2 (a) and (b).

There are many ways to generate a sample set of data. One is to use the “runif” command.

```
> #Generate random sample of data
> n=100
> x=runif(n, 0, 50)
>
> #Add the n+1 term
> xn1=append(x,40,after=length(x))
>
> #Calculate the mean of each
> xbar=mean(x)
> xn1bar=mean(xn1)
>
> #Calculate the variance of each
> varx=var(x)
> varxn1=var(xn1)
>
> #Show that the quantities are equal
> xn1bar
 [1] 23.10433
> xbar+(sum(xn1-xbar))/(n+1)
 [1] 23.10433
>
> varxn1
 [1] 224.1348
> ((n-1)/n)*varx+(1/(n+1))*(sum(xn1-xbar))^2
 [1] 224.1348
```

This creates a random sample of n objects, adds an n+1 term, and finds the mean and standard deviation for both samples. Then, we can confirm our result from part 2a and 2b by printing both quantities and showing they are equal. Thus:

$$\bar{x}_{n+1} = \bar{x} + \frac{x_{n+1} - \bar{x_n}}{n+1}$$

$$s_{n+1}^2 = \frac{n-1}{n} s_n^2 + \frac{1}{n+1}(x_{n+1} - \bar{x}_n)^2$$