SDS/MTH 220: Badge Challenge 1

Name: _____

Instructions

- This badge challenge must be your own work entirely.
- This is an open mind, closed stats notebook, closed textbook, and closed fellow statistician badge challenge.
- All questions will be graded under the badge level grading system. On badge challenges, formatting and presentation guidelines are suspended, but you must show ***all*** work for computational answers and justify all claims for expository questions.
- Keep your explanations contextually meaningful and concise!
- You do not have to perform any long computations. For example, if the answer is 18.5, you will receive full credit for writing $2.5 * (4 + 3.5) - (1/2)^2$.
- Use the provided blank sheets of paper to write your answers. You may also use these pages for your scratch work.
- Please write on ***one side*** *of the blank pages* and staple any pages you want graded to the badge challenge. Your *answers should appear in* ***question order***.
- Put your name on the top right corner of each page that you submit and do not write where the staple will go.
- This is the first of four opportunities to demonstrate your current understanding of the first five badges for the course. You may complete as many badges as you would like. Questions left unattempted will not receive a badge level.

Name: _____

Question	Badge 1	Badge 2	Badge 3	Badge 4	Badge 5
1					
2					
3					
4					
5					
Recorded Score by Badge					

Question 1

Consider the Galton data, which contains the following variables among others:

`heights` - The child's height as an adult (inches)

`mother` - The child's mother's height as an adult (inches)

`sex` - The sex of the child (M or F)

Consider the following model:

```
heights = 43.15546 + 0.32655*mother + 1.96331*sexM + 0.05014*mother*sexM
```

(a) What is the reference level for this model? Justify your answer.

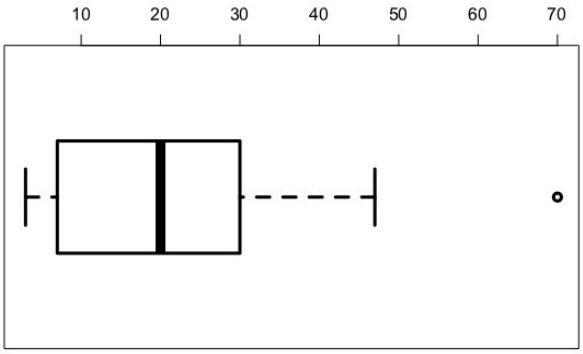
(b) Use the above model to construct a model formula between one's height and mother's height for females. Show or justify all work.

(c) Interpret the final coefficient in a contextually meaningful manner.

(d) Based on what you know about the above model, would you choose to keep the model as is or build a different trivariate model? Note that if you were to change the model, you must still use all of the three given variables and no others. Justify your choice.

Question 2

For this question consider the below box plot.



(a) For each of the below statements, state whether it is true or false and briefly ***explain*** your answer.

- The range of extremes is [3,47].
- According to the boxplot, the number of data points used to construct the box plot is 17.
- The 50% coverage interval is [7,30].
- The data is skew-left.
- According to the boxplot, the mean is 20.

(b) If you would like to know the above 5 summary statistics (i.e. range of extremes, skewness, mean, coverage intervals, and number of data points), is a box plot the best visual representation for your data? If not, what visualization would be better? Justify your answer.

Question 3

You are investigating if countries get a 'bump' in their gold medal count when they host the Winter Olympics. Recalling that the 2010 Winter Olympics were in Vancouver Canada, you decide to use Canada as a first test case. You found the following Canadian gold medal tallies for the past 9 Winter Olympics:

Year	1984	1988	1992	1994	1998	2002	2006	2010	2014
Canada's Gold Medals	2	0	2	3	6	7	7	26	10

(a) What is the median number of gold medals that Team Canada have won over the past 9 Winter Olympics? Show all work.

(b) Find the 50% coverage interval. Show all work.

(c) Based on the above table and results, do you think that Canada got a 'bump' for being the host country? Justify your answer.

(d) You double check the table's data and notice that for the 2010 games, the listed number of medals is the total number and not the just the gold medals. The correct number of gold medals is actually 13 gold medals. Does the median number of gold medals that Team Canada has won over the past 9 Winter Olympics change? Does the mean change? Justify ***both*** your answers ***without*** doing any ***additional*** computations.

Alternative

You have been asked to consult on a small study. Due to privacy concerns, all you have is a collection of values for one variable: Mystery $= \{8, 6, 12, 7, 6, 5, 12, 6, 10, 6, 12, 10, 7\}$. You do not know what the data is about.

a) Compute the *median* and the *mode* of Mystery. Show all work.

b) The head researcher found that the mean and the standard deviation of Mystery are 8.2. and 2.6, respectively. He says that 68% of the data falls between 5.6 and 10.8. Can he make such a claim about this data? Why or why not?

c) A postdoc on the study realizes that the data notebook for Mystery was incorrectly transcribed and that the true data is Mystery $= \{8, 6, 12, 7, 6, 5, 12, 6, 10, 0, 12, 10, 7\}$. In other words, the number recorded as a 6 is actually a 0. Does this discovery change the value of the standard deviation? Justify your answer ***without*** computing the new standard deviation.

d) Does the above discovery change the value of the 25^{th} percentile? Again, justify your answer ***without*** computing the new 25^{th} percentile.

Question 2

Australian researchers explored the association between depression and factors such as employment status, marital status, and satisfaction with their level of social interaction. You worked with a sociology researcher to build a model using the following variables:

`BDI`	Beck depression index (high `BDI` = high depression)
`employment`	`employed, govt assistance, other,` or `parental support`
`psisat`	satisfaction with positive social interaction (high `psisat` = high satisfaction)

Unfortunately, the original code got deleted! The below is all you have:

```
                           Estimate Std. Error
(Intercept)                 32.2064     4.7728
employmentgovt assistance    4.4003     2.5579
employmentother             -3.9285     2.4804
employmentparental support  -2.0387     2.2902
psisat                      -1.7314     0.3727
```

a) Write out the model formula associated to the above output.

b) Interpret the intercept term in a contextually meaningful way.

c) What is the reference level for `employment`? Justify your answer.

d) Does this model assume that the relationship between `employment` and `BDI` depends on `psisat`? Justify your answer.