Math 10 – Homework 1 **ANSWER KEY**

Instructions:

- Type your answers and paste images directly into this document.
- Answers are usually short, with 1-3 sentences.
- Print out and hand in homework in class on Tuesday.
- Please use the online histogram making tool here: http://www.shodor.org/interactivate/activities/Histogram/ (select "My Data" in the drop down menu and enter the data into the box).
- You may collaborate on the homework but you must write it up yourselves.

Problem 1 - Inferential Statistics

(*Modified from a question in "Statistics" by Freedman, Pisani and Purves.*)

California is evaluating a new program to rehabilitate prisoners before their release; the object is to reduce the recidivism rate – the percentage who will be back in prison within two years of release. The program involves several months of "boot camp" – military style basic training with very strict discipline. Admission to the program is voluntary. According to a prison spokesman, "Those who complete boot camp is less likely to return to prison than other inmates".

1. What is the treatment group? What is the control group? (*2 points*)

   Treatment = prisoners who took part in boot camp.
   Control = prisoners who did not take part.

2. True or false: the data show that boot camp worked. Explain your answer. (*2 points*)

   False. Those who are motivated enough to volunteer for boot camp might tend to be the ones who will not be back in prison. The assignment is biased.
   Alternative answer: the data only showed that those who volunteered for boot camp tend to not be back in prison.

3. Suggest one modification that would have improved the evaluation of the boot camp program. (*2 points*)

   Assign prisoners randomly into the treatment and control group, do not allow volunteering. Optional but correct to add: select a random or stratified sample of all prisoners for this assignment.

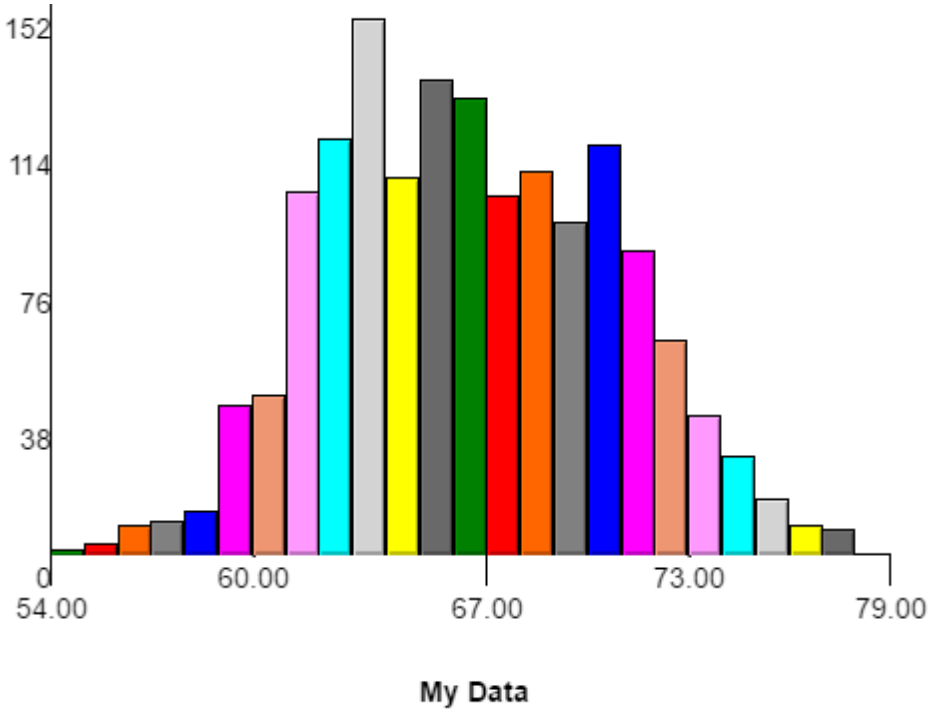
   Alternative answer: allow volunteering but randomly assign volunteers into treatment and control group.

Problem 2 - Histograms, Subpopulations, Distribution Shape

"height.txt" contains heights in inches of 1490 subjects. This data was part of the National Longitudinal Survey of Youth (1999).

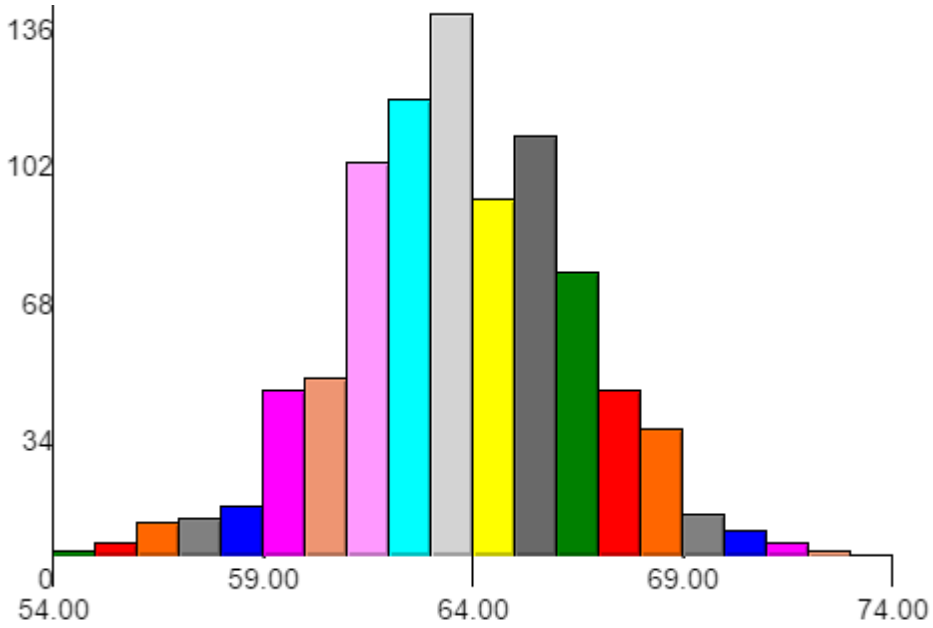
1. Open the text file "heights.txt". Cut and paste the data into the histogram making tool here: http://www.shodor.org/interactivate/activities/Histogram/

   Change "Interval Size" to 1 and "X Min" to 54. Right click to save the image of the histogram and drag it into this document so that it appears below. (*1 point*)

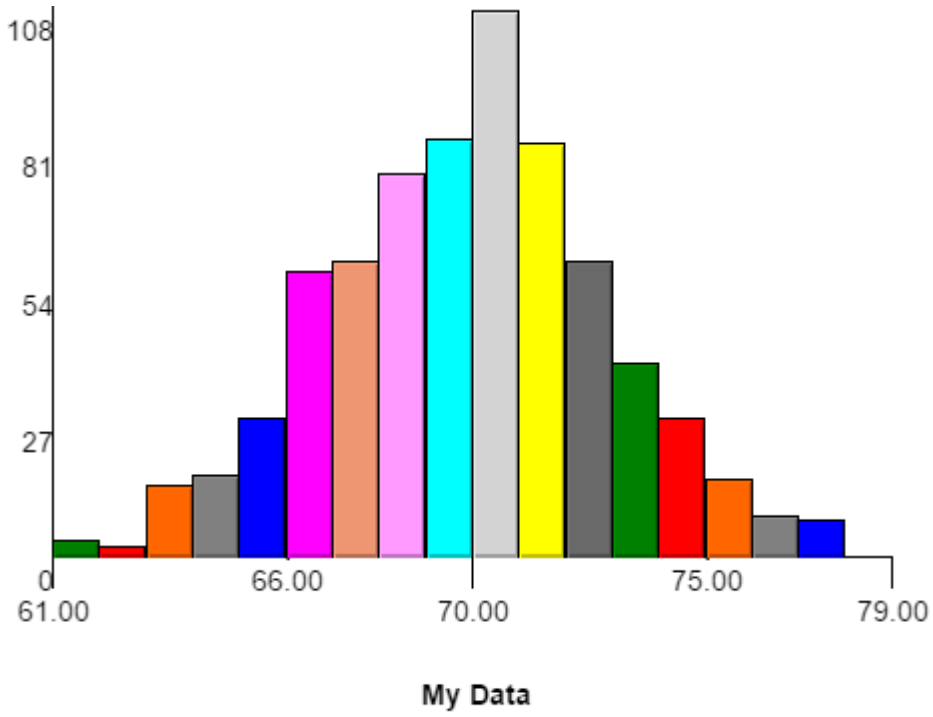


The data above contained two subpopulations: 19 year old males and 16 year old females. These are now split into two files “male.txt” and “female.txt”.

2. Produce the histogram for “female.txt” below, like you did for question 1. In the histogram making tool, change “Interval Size” to 1 and “X Min” to 54. (*1 point*)

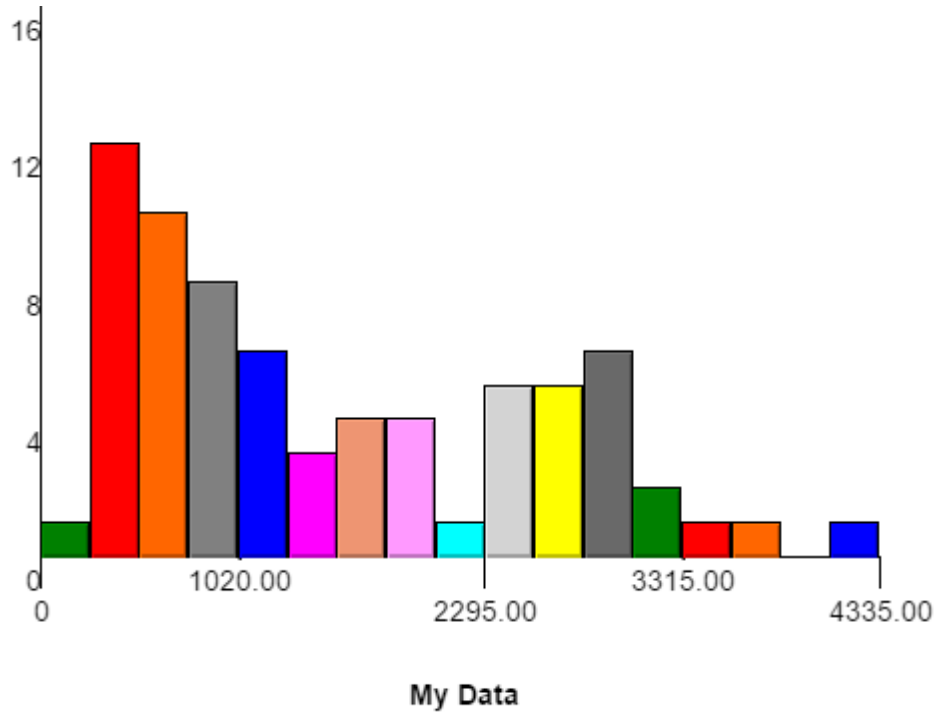


3. Produce the histogram for “male.txt” below, like you did for question 1. In the histogram making tool, change “Interval Size” to 1 and “X Min” to 61. (*1 point*)



"wages.txt" contains a list of average monthly wages for 70 different countries in 2009, adjusted for the cost of living. This data was compiled by the United Nations.

4. Produce the histogram for "wages.txt" below, like you did for the previous questions. In the histogram making tool, change "Interval Size" to 255 and "X Min" to 0. (*1 point*)



5. Describe the difference (if any) in the shapes between the distribution of average monthly wages and the distribution of 19 year old male heights. (*2 points*)

   Heights = looks symmetric. Wages = looks skewed. Specifics not required.

Problem 3 - Central Tendency, Variability, Percentiles

1. Included in the output of the histogram making tool is the mean and standard deviation. Please state them for the histogram of 19 year old male heights in "males.txt", up to three decimal places if possible. (*2 points*)

   Mean = 70.390
   Std. Dev = 2.971

2. Are the mean and standard deviation good summary statistics for this histogram? Explain your answer. (*2 points*)

   Yes. The mean and standard deviation are usually good summary statistics for histograms that are symmetric. (specifics not required)

3. State the mean and standard deviation for the histogram of wages in "wages.txt", up to three decimal places if possible. (*2 points*)

   Mean = 1517.157
   Std. Dev = 1021.509

4. Are the mean and standard deviation good summary statistics for this histogram? Explain your answer. (*2 points*)

   No, because the histogram is skewed. See chapter 3, section 11 for detailed reasons.

5. Calculate the 25^{th}, 50^{th} and 75^{th} percentile for the data in "wages.txt". Please show your work/calculation, no points will be given otherwise. (*3 points*)

   Note: the data in "wages.txt" has already been sorted. For this course, we will be using the textbook's definition of percentile in chapter 1, section 8, "third definition".

   Textbook formula: 0.25*(70+1) = 17.75. Then, 25^{th} percentile = 609 + 0.75*(656-609) = 644.25.

   0.50*(70+1) = 35.5. Then, 50^{th} percentile = 1109 + 0.50*(1135-1109) = 1122.

   0.75*(70+1) = 53.25. Then, 75^{th} percentile = 2445 + 0.25*(2522-2445) = 2464.25.

6. The interquartile range is the difference between the 75^{th} and 25^{th} percentile. Compute the interquartile range of the data in "wages.txt". (*1 point*)

   2464.25 – 644.25 = 1820. We will still give credit if this number is wrong due to computation errors in the previous part.

7. Find the median of the data in “wages.txt”. Show your work or explain your answer. (*1 point*)

   Median = (1135+1109)/2 = 1122. Full credit for stating that it is the same as 50th percentile.

8. Compare the median of the data in “wages.txt” to the mean computed earlier. Is there a difference? If there is, why is the median and the mean different? (*2 points*)

   Textbook chapter 3, section 11 : positive skew can cause the mean to be higher than the median.