Homework #1 Due Thursday Sept 20th 5 pm

*Submit your homework on Canvas and to your TA's mailbox before the due date/time. The mailboxes are to the left as you enter the Medical Science Center (1300 University Ave.) from the main University Ave. entrance.

*No late homework will be accepted for credit.

*If a problem asks you to use R, include a copy of the code and output. Please edit your code and output to be only the relevant portions.

*If a problem does not specify how to compute the answer, you may use any appropriate method. I may ask you to use R or manual calculations on exam, so practice accordingly.

1. If you wanted to estimate the mean height of all the students at UW Madison, which one of the following sampling strategies would be best? Why? Note that none of the methods are true simple random samples.

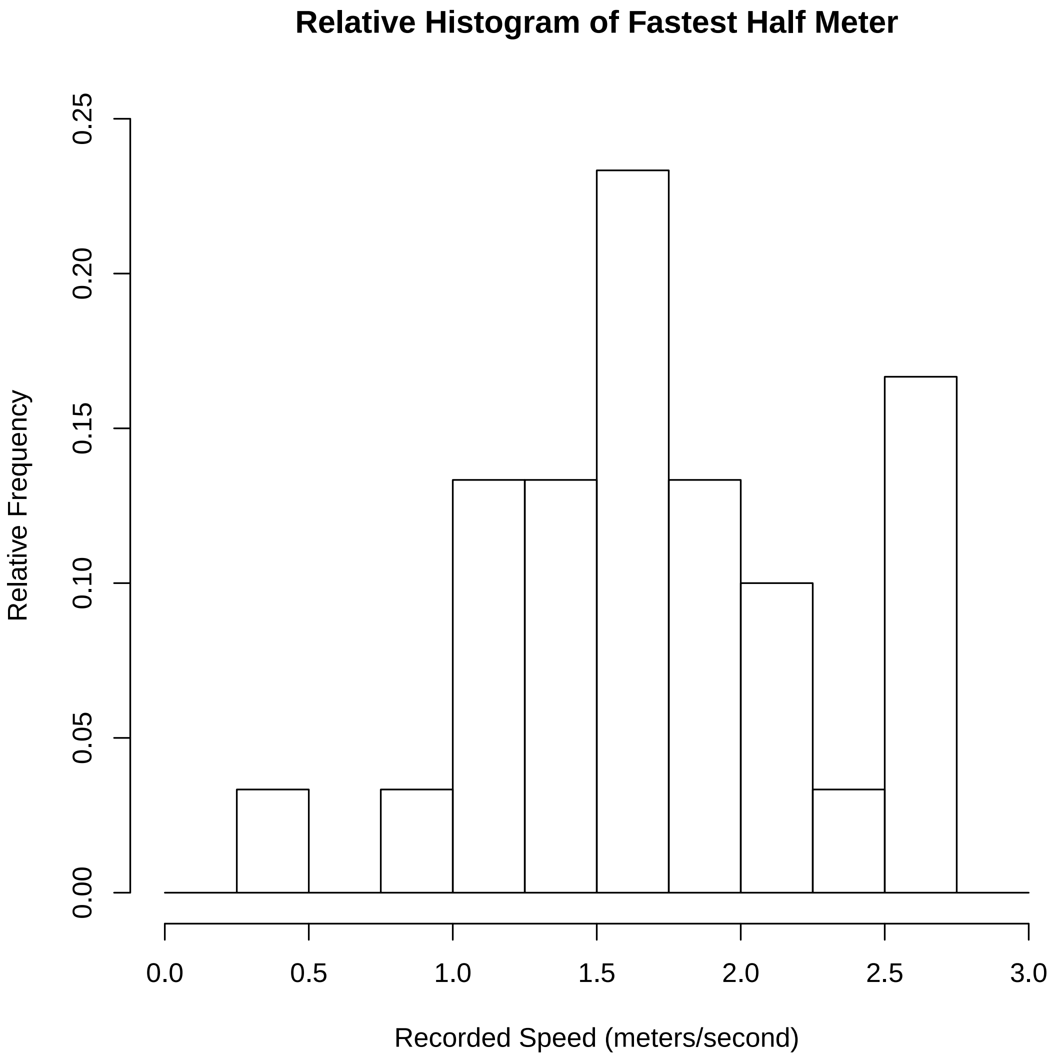
   (a) Measure the heights of 100 students found in the gym during basketball intramurals.

   (b) Measure the heights of the engineering majors.

   (c) Measure the heights of the students selected by choosing the first name on each page of a list of students enrolled that semester.

   2 points, +1 for selection, +1 for reasonable explanation *The third option seems to have the most randomness in terms of who is included in the sample (could change year to year, publication to publication). The two other samples may introduce more bias- systematically over or under estimate heights since they may result in samples that are less representative of the population of interest.*

2. A zoologist collected wild lizards in the Southwestern United States. Thirty lizards from the genus *Phrynosoma* were placed on a treatmill and their speed measured. The recorded speeds (meters/second) (the fastest time to run a half meter) for the thirty lizards are summarized in the relative histogram below. (Data Courtesy of K. Bonine *)



(a) Is the percent of lizards with recorded speed below 1.25 closest to: 25%, 50%, or 75%? *25%*

(b) In which interval are there more speeds recorded: 1.5-1.75 or 2-2.5? *1.5-1.75 looks to have about 23%, while 2-2.5 has about 13%*

(c) About how many lizards had recorded speeds above 1 meters/second? *28 since it looks like there were 2 below 1 meters/second*

3. In a sample of 30 men, the mean height was 179 cm with standard deviation of 6 cm. In a sample of 25 women, the mean height was 163 cm with standard deviation of 6 cm. If both samples were combined into one larger group...

   (a) What is the mean height for the combined group?

   (b) The standard deviation for the combined group would be

      i. Less than 6 cm

      ii. Greater than 6 cm

iii. Equal to 6 cm

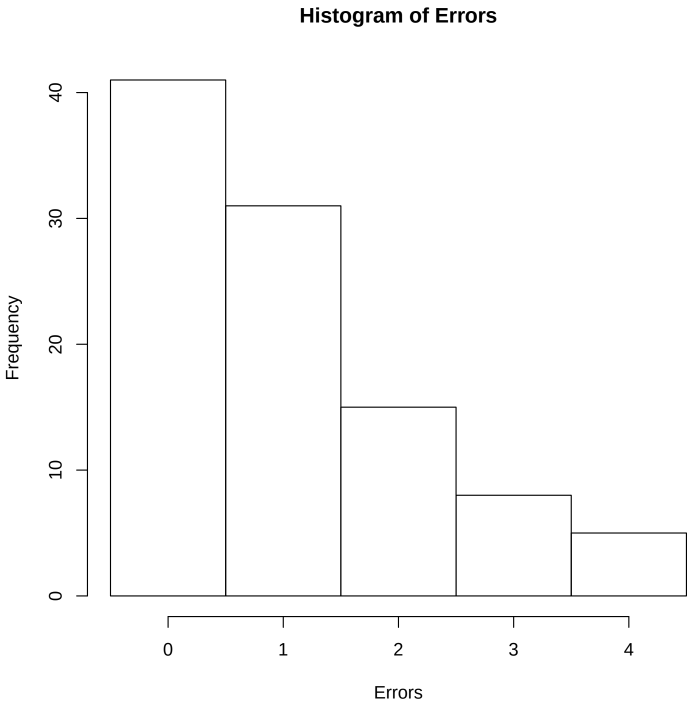
iv. Can not tell from the information given

Mean height: (30(179)+25(163))/55=171.73. SD for the combined group will be larger since there is more variability (bimodal distribution)

4. After manufacture, computer disks are tested for errors. The table below gives the number of errors detected on a random sample of 100 disks.

Number of Defects	0	1	2	3	4
Frequency	41	31	15	8	5

(a) Construct a frequency histogram of the information by creating bins at [-.5, .5), [.5, 1.5), etc. **3 points +1 rel height of bars ok; +1 bars centered around 0,1,2,3,4; +1 axis labeled**



(b) What is the shape of the histogram for the number of defects observed in this sample?
1 point *Right Skewed*

(c) Calculate the mean and median number of errors detected on the 100 disks. How do these values compare and is that consistent with what we would guess based on the shape?
3 point +1 for each value, +1 comparison*mean: 1.05, median: 1. The mean is slightly above the median which is consistent with the right skew. Median is closer to higher frequency lower values*

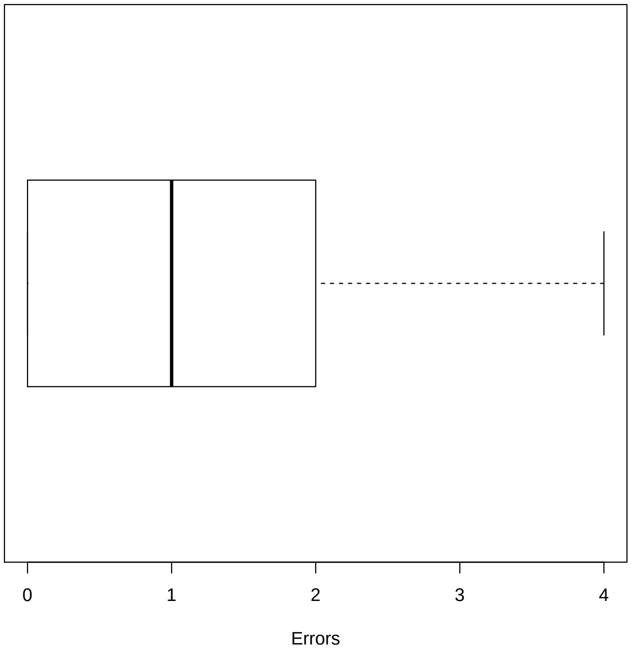
(d) Calculate the sample standard deviation with your calculator and R. Are the values consistent between the two methods? Explain what this value means in the context of the problem.
2 point, +1 for values, +1 comparison *sd= 1.157976. Yes, the values are consistent. The number of defaults on a computer disk in the sample of 100 were about 1.16 deviations away from the mean of 1.05 errors, on average.*

(e) Calculate the first and third quartiles and IQR by hand and with R. Are the values consistent betweeen the two methods? Explain what the three values mean in the context of the problem.
2 point +1 for values by hand, +1 for values on R *Since there are 100 data values, Median is the average of the 50 and 51st values. Median=1. Q1 is then the median of the first 50 observations, so the average of 25th and 26th observations, which are both 0: Q1=0. Q3 is the median of the last 50 observations, so the average of the 75th and 76th observations which are both 2, so Q3 is 2. These are the same values calculated by R. The IQR is 2-0=2. Q1 tells us the number of errors below which approximately 25% of the disks experienced. Q3 tells us the number of errors below which approximately 75% of the disks experienced. IQR tells us the range of the middle 50% of errors.*

(f) What proportion of the computer disks had a number of errors greater than the mean number of errors?
1 point *Since mean was 1.05, those with 2, 3 or 4 had a number greater than the mean so that is 15+8+5=28/100 or 28% of the disks*

(g) What range of values for this sample data are not considered outliers using the [Q1-1.5IQR, Q3+1.5IQR] designation (using the IQR you calculated by hand)?.
1 point *1.5*IQR=3, so Q1-1.5*IQR=-3, Q3+1.5*IQR=5. Any values between 0 and 5 are not considered outliers for this sample. Saying -3 is ok.*

(h) Sketch a boxplot of the data by hand (using the relevant values you calculated by hand). **5 points for box drawn at Q1(+1), Med(+1), Q3(+1) hand-values, (+1) for whiskers to min and max, (+1) label**



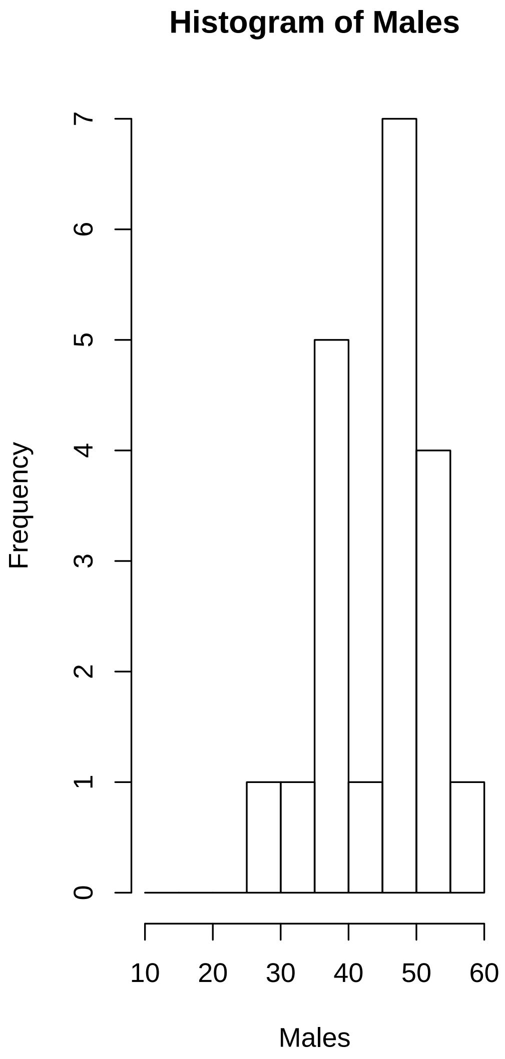
(i) Compare and contrast (briefly) the information about the data given by the histogram in part a and the boxplot in part h. *Both graphs show the right skew; The histogram more clearly shows the discrete nature of the data; the boxplot more clearly identifies the 5 number summary (Min, Q1, Med, Q3, Max)*

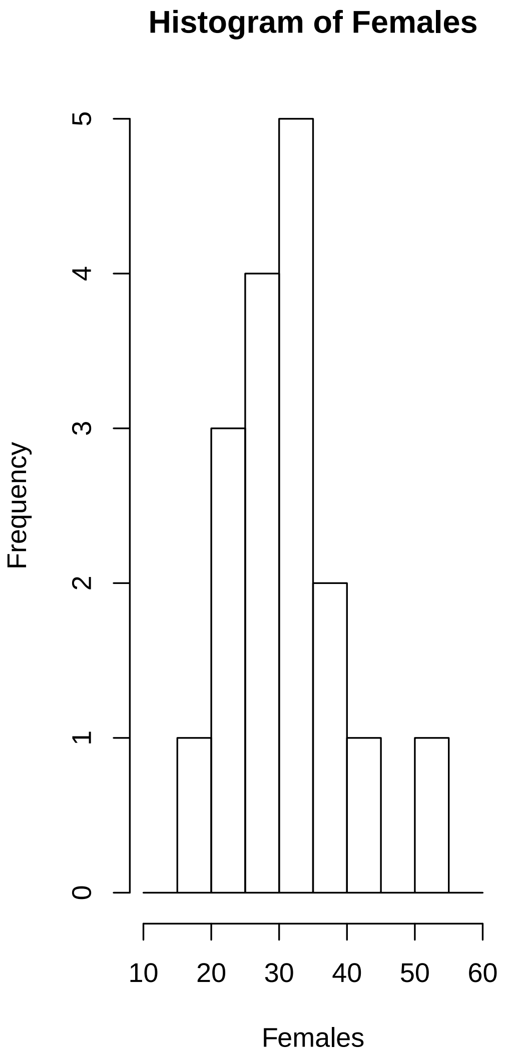
5. Physical education researchers interested in the development of the overarm throw measured the horizontal velocity of a thrown ball at the time of release. The results for first-grade children (in feet/sec) (courtesy of L. Halverson and M. Roberton*) are:

   Males: 54.2, 39.6, 52.3, 48.4, 35.9, 30.4, 25.2, 45.4, 48.9, 48.9, 45.8, 44.0, 52.5, 48.3, 59.9, 51.7, 38.6, 39.1, 49.9, 38.3

   Females: 30.3, 43.0, 25.7, 26.7, 27.3, 31.9, 53.7, 32.9, 19.4, 23.7, 23.3, 23.3, 37.8, 39.5, 33.5, 30.4, 28.5

(a) Use R to create a histogram for the males and a histogram for the females (any kind of histogram that you want). Adjust the x axis scale so the two groups are more easily compared.


Histogram of Males


Histogram of Females

(b) Compare the shape of the throws from the male and female students observed in this sample.

The girl velocities were slightly right skewed while the boy's were more symmetric. The girls' had 1 prominent peak, while the boys' had two.

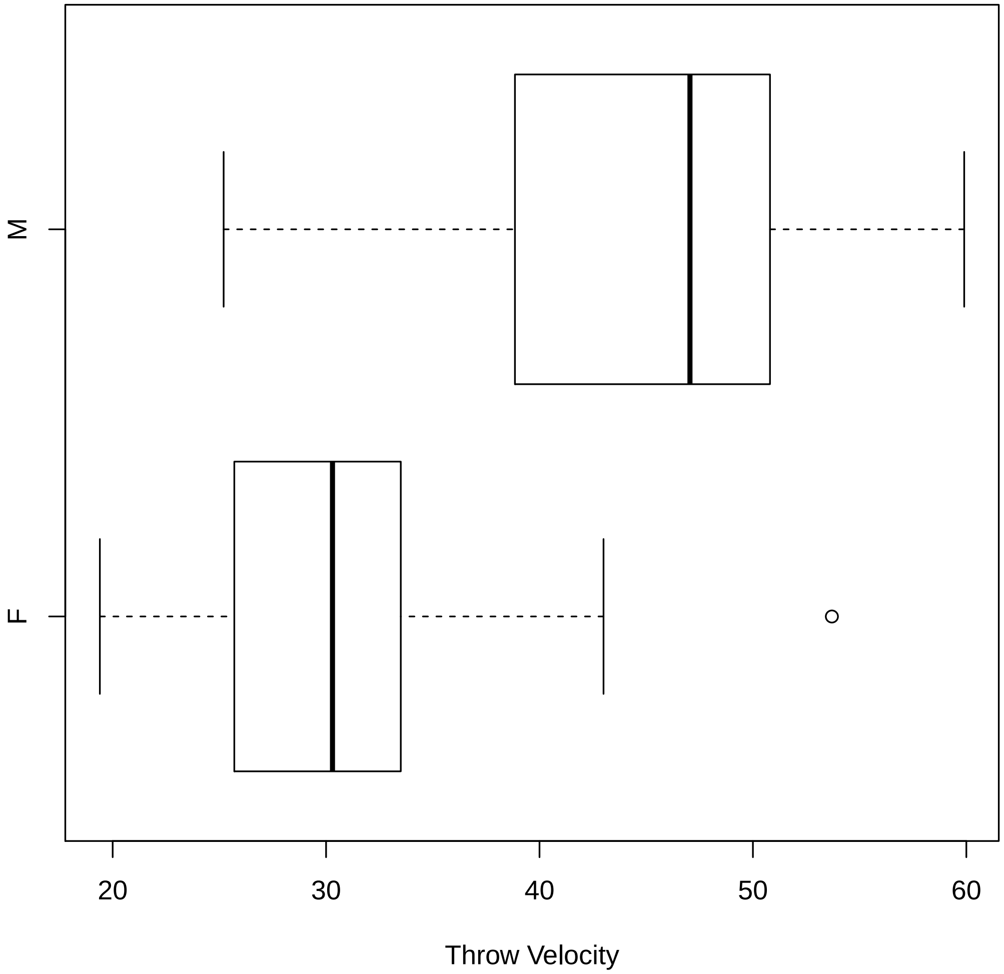
(c) Compute and compare the mean and median throw velocities observed in the male and famale students across gender.

Female Mean: 31.23, Male Mean:44.87, Female Median: 30.3, Male Median: 47.05. The male mean and median throw velocities were higher than that observed in the female students.

(d) Compute and compare the standard deviation in throw velocities observed in the male and famale students.

Female SD:8.519666 Male SD:8.513845 The standard deviation within the male and female samples are very similar.

(e) Use R (or by hand if you prefer) to help you create a boxplots of the two sets so they are easily comparable.



(f) Which, if any values were identified as outliers? Would this value have been identified as an outlier if it were thrown by the opposite gender?

The maximum female throw of 53.7 was identified as an outlier. This would not be an outlier had it been thrown by a male (since higher values thrown by males were not designated as outliers.)

6. There are 12 numbers on a list, and the mean is 24. The smallest number on the list is changed from 11.9 to 1.19.

(a) Is it possible to determine the direction in which (increase/decrease) the mean changes? Or how much the mean changes? If so, by how much does it change? If not, why not? ***2 points +1 new mean, +1 for amount of change*** *Yes. Original mean is 24, which means original total is* $12 * 24 = 288$.

New total is $288 - 11.9 + 1.19 = 277.29$. *New mean is* $277.29/12 = 23.1075$, *so mean decreased by* $24 - 23.1075 = 0.8925$

(b) Is it possible to determine the direction in which the median changes? Or how much the median changes? If so, by how much does it change? If not, why not? ***2 points +1 new median, +1 for amount of change****Yes. Since median is related to a middle value and the smallest number is not a middle number in a list of 12, so the median will not be affected. Change=0.*

(c) Is it possible to determine the direction in which the standard deviation changes? ~~Or how much the standard deviation changes? If so, by how much does it change?~~ If not, why not? *No. We do not know the original standard deviation (so we're unable to calculate the original squared deviations away from the mean. We do know standard deviation increases since the data is more spread out than before.*

7. The UW Statisics Department is trying to determine what day of the week to hold their annual fall festival. Assume that the weekdays, Monday through Friday, are equally likely and that each weekend day, Saturday and Sunday, is three times as likely as a weekday to be selected.

   (a) Assign probabilities to the seven outcomes.
   Mon:1x, Tues:1x, Wed:1x, Thurs:1x, Fri:1x, Sat: 3x, Sun: 3x, Total=11x=1, x=1/11. P(Mon)=..P(Fri)=1/11, P(Sat)=P(Sun)=3/11

   (b) Find the probability a weekday will be selected.
   *P(weekday)=5*1/11=5/11*

8. (*)Suppose you are eating at a pizza parlor with two friends. You have agreed to the following rule to decide who will pay the bill. Each person will toss a coin. The person who gets a result that is different from the other two will pay the bill. If all three tosses yield the same result, the bill will be shared by all. (It may be helpful to list the outcomes in the sample space)

   HHH, HTT, THT, TTH, HHT, THH, HTH, TTT

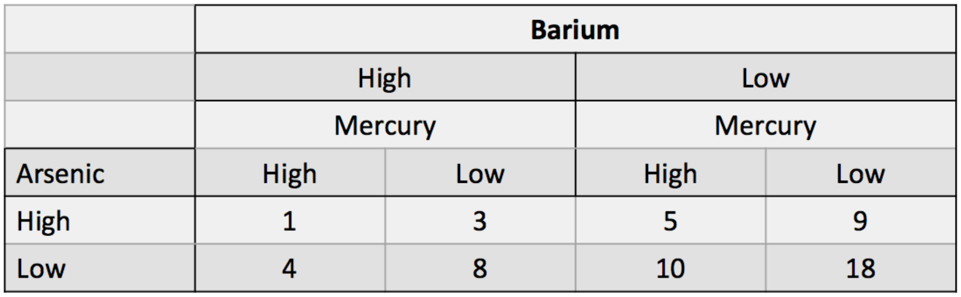
   (a) Find the probability that only you have to pay.

   2 points *I'm going to assume you is the first in the list. If only you have to pay that means, the first in the list is different from the other two: HTT or THH, so you have 2/8=1/4=0.25 probability of paying the bill by yourself.*

   (b) Find the probability that all three will share.

   2 points *If all three have to share payment, that means all three yeild the same result: TTT or HHH, so you have 2/8=1/4=0.25 probability of splitting the bill.*

9. (*) The following frequency table shows the classification of 58 landfills in a state according to their concentration of the three hazardous chemicals arsenic, barium, and mercury.



	Barium			
	High		Low	
	Mercury		Mercury	
Arsenic	High	Low	High	Low
High	1	3	5	9
Low	4	8	10	18

If a landfill is selected at random, find the probability that it has:

(a) A high concentration of barium. ***1 point****16/58*

(b) A high concentration of mercury and low concentrations of both arsenic and barium. ***1 points****10/58*

(c) A high concentration of any one of the chemicals and low concentrations of the other two.***2 point****P(Mba or mBa or mbA)=(10+8+9)/58=27/58*

(d) A high concentration of Mercury given it has a high concentration of Arsenic. Is having high concentrations of Mercury and high concentrations of Arsenic independent in this set of landfills? ***3 points, +1 conditional, +1 unconditional, +1 not indep****P(high Merc—high Arsenic)=6/18=.33333 P(high Merc)=20/58=.3448. The Prob of high Merc within high Arsenic landfills is slightly lower than that in general, so not ind. in this sample*

(e) A high concentration of Barium given it has a low concentration of Mercury. Is having high concentrations of Barium and low concentrations of Mercury independent in this set of landfills? *P(high Barium — low Mercury)=11/38=0.289, P(High Barium)=16/58=0.276. No, because these are not equal. Low Mercury group has higher prob of high Barium than general.*

*Johnson and Bhattacharyya, *Statistics- Principles and Methods* data and some prompts