Mathematics 1372/D552, Spring 2016 **Exam #1**
Instructor: Suman Ganguli **March 9, 2016** **Name:** _____

---

1. (20 points) The following data set lists the high temperatures (in degrees Fahrenheit) observed in Central Park on July 4th from 2003-2012 (from `http://www.erh.noaa.gov/okx/climate_cms.html`):

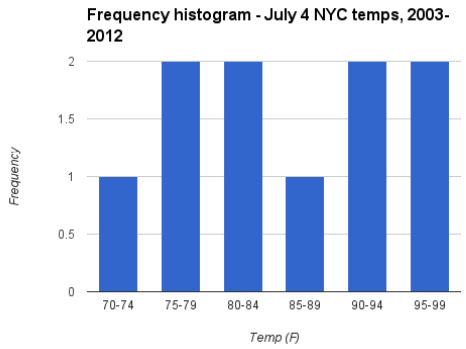
$$\{96, 92, 82, 83, 87, 71, 78, 79, 96, 92\}$$

(a) Complete the following frequency table for this data set:

Class	Frequency, f	Relative frequency
70-74	1	0.1
75-79	2	0.2
80-84	2	0.2
85-89	1	0.1
90-94	2	0.2
95-99	2	0.2

(b) Use your frequency table to sketch a frequency histogram:

Solution:



(c) Use a spreadsheet to find the following sample statistics:

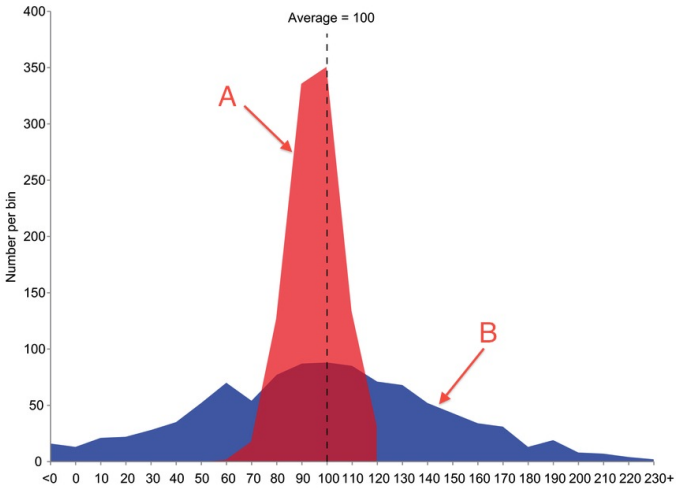
sample mean $\bar{x} =$

sample standard deviation $s =$

sample median $=$

Solution:
Using the spreadsheet functions, `=average`, `=stdev`, and `=median`:
sample mean $\bar{x} = 85.6$
sample standard deviation $s = 8.396$
sample median $= 85$

2. (10 points) The graph below shows frequency histograms for two different data sets, labelled A & B. As indicated on the graph, the two data sets have the same mean. Which one has the higher standard deviation? **Briefly explain why.**



Solution: The data set B clearly has the higher standard deviation, since the histograms show that the values of B are much more dispersed than those of A, i.e., on average the values of B are much further away from the mean than values of A, which are concentrated relatively close to the mean.

3. (20 points) Consider the probability experiment consisting of flipping a coin three times in a row. The sample space of the experiment is

$$\{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$

Consider the events

- A = "getting 3 heads in a row"
- B = "getting heads twice and tails once"

For each of the events, (i) list the outcomes that make up the event, and (ii) calculate the probability of the event (using the definition of "theoretical probability" discussed in class:

$$P(A) = \frac{\#\text{ of outcomes in the event } A}{\#\text{ of total outcomes in the sample space } S}$$

(a) **Solution:** $A = \{HHH\}, P(A) = \frac{1}{8} = 0.125$

(b) **Solution:** $B = \{HHT, HTH, THH\}, P(B) = \frac{3}{8} = 0.375$

(c) **Solution:** $A \cup B = \{HHH, HHT, HTH, THH\}, P(A \cap B) = \frac{4}{8} = 0.5$

(d) **Solution:** $A \cap B = \emptyset, P(A \cap B) = 0$

4. (10 points) A bag contains three balls–two of them are red and one is blue. Consider the probability experiment that consists of withdrawing two balls **without replacement** (i.e., the first ball is chosen, and then the second ball is chosen without putting the first ball back in the bag.)

   (a) What is the sample space of this experiment? (Hint: There are six outcomes in the sample space. Label one red ball R_1, the other red ball R_2, and the blue ball B. Now list all the possible outcomes of the experiment.)

   Solution: $\{R_1B, R_1R_2, R_2B, R_2R_1, BR_1, BR_2\}$

   (b) What is the probability of withdrawing the two red balls?

   Solution: There are two outcomes in the given event: $A = \{R_1R_2, R_2R_1\}$. Since there are six outcomes in the sample space, $P(A) = \frac{2}{6} = \frac{1}{3}$

5. (10 points) Suppose at the end of the semester you have created the following frequency table summarizing the commute time data you collected:

Commute time (minutes)	Frequency	(Time)*(Frequency)
28	3	28*3 = 84
29	5	29*5 = 145
30	8	30*8 = 240
31	7	31*7 = 217
32	4	32*4 = 128
33	3	33*3 = 99
34	1	34*1 = 34
35	0	35*0 = 0
36	2	36*2 = 72
37	0	37*0 = 0
38	1	38*1 = 38
39	0	39*0 = 0
40	1	40*1 = 40
Totals:	35	1097

   (a) What was your sample size, i.e., how many commute times did you record?

   Solution: Summing up frequencies shows that the sample size is 35.

   (b) Compute the sample mean of your commute times.
   (Hint: You should use a spreadsheet to do the necessary calculations using the numbers in the frequency table–but also write in the results of those calculations in the table above.)

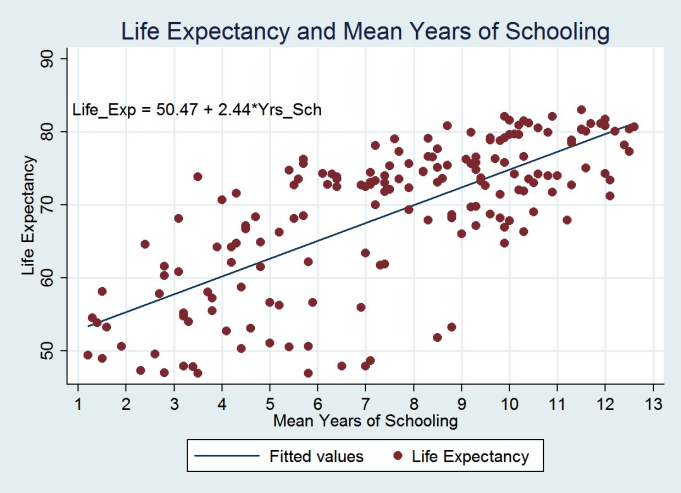
   Solution: To compute the sample mean, we need to sum up all the commute time observations. To do this, compute (time*frequency) for each line of the frequency table, as shown in the table above (all these calculations are most easily done in spreadsheet). This gives the sum of the observed commutes for that particular line (e.g., the frequency of the commute time 28 (minutes) is 3, i.e., there were 3 such observations, so summing up the the times for those 3 observations is $28 + 28 + 28 = 28 * 3 = 84$.
   Now sum up that column of (time*frequency) and divide by the sample size to compute the sample mean:

$$\bar{x} = \frac{(28*3) + (29*5) + (30*8) + (31*7) + (32*4) + (33*3) + (34*1) + (36*2) + (38*1) + (40*1)}{35}$$

$$= \frac{1097}{35} \approx 31.34$$

6. (25 points) The following scatterplot and text is taken from a blog called "Decisions Based on Evidence" (`http://www.decisionsonevidence.com/2012/11/life-expectancy-and-infant-mortality-the-key-role-of-education/`):

   Here is a graph with a linear fit depicting the relationship between 175 country life expectancies and corresponding mean years of schooling in each country. Clearly, a direct relationship exists.



(a) What is the nature of the "direct relationship" between life expectancy and average years of schooling? Are they positively or negatively correlated? Explain briefly both in terms of the given scatterplot and in terms of the variables involved (years of schooling and life expectancy).

Solution: Strong positive correlation: life expectancy tends to increase as average years of schooling increases

(b) Which of the following is closest to the correlation coefficient for this paired data set?
(I) 0.7 (II) 0.1 (III) -0.1 (IV) -0.7

Solution: Since the scatterplot shows a strong positive correlation, the correlation coefficient must be positive and relatively close to 1, so the only reasonable choice is **(I) = 0.7**.

(c) The graph includes the linear regression line and its equation: "Life_Exp = 50.47 + 2.44*Yrs_Sch". What do the numbers 50.47 and 2.44 represent in terms of the line and its equation? What do the numbers represent in terms of the relationship between the variables, i.e., in terms of life expectancy and average years of schooling (according to the linear regression model)?

Solution: 50.47 is the y-intercept of the regression line, which represents the linear regression model's predicted life expectancy in a country with 0 average years of schooling.

2.44 is the slope of the regression line, which represents the gain in a country's life expectancy if its average years of schooling increases by 1.

(d) Use the equation of the regression line to predict the life expectancy in countries with the following mean years of schooling (i.e., plug in the given values of "Yrs_Sch" into the given equation of the linear regression line to get the corresponding (predicted) value for "Life_Exp")

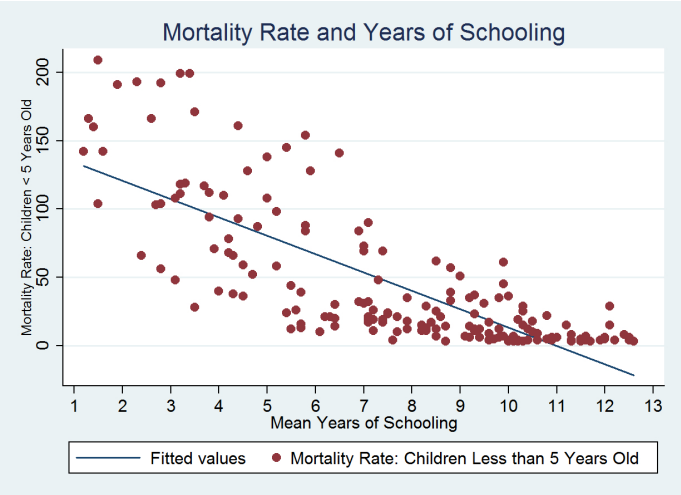
(i) 1 year:

Solution: Life_Exp = 50.47 + 2.44*(1) = 52.91

(ii) 10 years:

Solution: Life_Exp = 50.47 + 2.44*(10) = 50.47 + 24.4 = 74.87

The blog post also has the following scatterplot, for Child Mortality Rate vs. Mean Years of Schooling:



(e) What is the relationship between child mortality rate and average years of schooling? Are they positively or negatively correlated? Again, explain why in terms of the given scatterplot and in terms of the variables involved (in this case years of schooling and child mortality rate).

Solution: Strong negative correlation: a country's child mortality rate tends to decrease as its average years of schooling increases.

(f) Extra credit: Use the graph of the linear regression line to estimate the linear regression parameters.
(Hint: Extend the line to the left to the y-axis in order to estimate the y-intercept, and then use that value together with the fact that the line appears to cross the x-axis at $x = 11$ in order to calculate the slope.)

The regression line is: "Mortality_Rate $= \alpha + \beta *$Yrs_Sch" where

$$\alpha \approx$$

$$\beta \approx$$

Solution:
The y-intercept $\alpha \approx 140$, since that's the approximate y-value where the regression line intersects the y-axis.

To calculate the slope, we need the coordinates of two points on the line, so that we can calculate (rise/run). From our estimate of the y-intercept, we know the point $(0, 140)$ is on the line, and since the regression line appears to cross the x-axis at $x = 11$, the point (11,0) is also on the line. Hence we can compute the slope as

$$\beta \approx \frac{0 - 140}{11 - 0} \approx= -12.7$$

We can interpret the regression parameters in the usual way: the y-intercept $\alpha \approx 140$ represents the regression model's prediction of child mortality rate in a country with 0 years average schooling, and the slope $\beta \approx= -12.7$ represents that a country's child mortality rate will decrease by 12.7 years for each additional year of average schooling.