175 Possible Points.

Elements of Statistics (Math 106) - Exam 1
Spring 2009 - Brad Hartlaub

Name ___*Brad Hartlaub - Solutions*___

Directions: Please answer all of the questions below and <u>show your work</u>. The point values for each problem are indicated in parentheses. You may use one sheet of formulas, our course web page, and Minitab during the exam. Don't spend too much time on any one part of the exam.

1. Thirty percent of all automobiles undergoing an emissions inspection at a certain inspection station fail the inspection. $X = $ # of automobiles that fail the inspection. $X \sim B(n=15, P=.3)$

   a. Among 15 randomly selected cars, what is the probability that at most 5 fail the inspection? (5)

$$P(X \le 5) \underset{MTB. \ cumulative}{=} .721621 \quad (or .722 \ Graph > Prob \ Dist \ Plot)$$

   b. Among 15 randomly selected cars, what is the probability that exactly 5 fail the inspection? (5)

$$P(X=5) \underset{MTB \ prob}{=} .206130$$

Trouble.   c. Among 15 randomly selected cars, what is the probability that at least 5 fail the inspection? (5)

$$P(X \ge 5) = 1 - P(X < 5) = 1 - P(X \le 4) = 1 - .515491 = .484509$$
$$(.485 \ Graph > Prob \ Dist \ plot)$$

   d. Among 25 randomly selected cars, what is the mean value of the number of cars that pass inspection, and what is the standard deviation of the number that pass inspection? (10)

$$Y \sim B(25, .7)$$
$$\hookrightarrow \text{# that pass.}$$
$$\mu_Y = nP = 25(.7) = 17.5$$
$$\sigma_Y = \sqrt{25(.7)(.3)} = 2.2913$$

   e. What is the probability that among 25 randomly selected cars, the number than pass is within 1 standard deviation of the mean? (10)

$$17.5 - 2.2913 = 15.2087 \qquad P(15.2087 \le Y \le 19.7913)$$
$$17.5 + 2.2913 = 19.7913 \qquad = P(16 \le Y \le 19) = .617$$

2. What type of graph or graphs would you plan to make in a study of each of the following issues? (18 – 3 each question)

   a. What makes of cars do students drive? ___*bar graph (or pie chart)*___
   b. How old are their cars? ___*histogram, dotplot, stem-and-leaf plot*___
   c. How many hours per week do students study? ___*histogram, dotplot, stem-and-leaf plot*___
   d. How does the number of study hours change during the semester? ___*time series plot or scatterplot*___
   e. Which radio stations are the most popular with students? ___*bar graph (or pie chart)*___

3. You are assigned to direct a study at Kenyon College to discover factors that are associated with strong academic performance. You decide to identify 20 students who have perfect GPAs of 4.0, and then measure explanatory variables for them that you think may be important, such as high school GPA and average amount of time spent studying per day.

   a. Explain what is wrong with this study design. (10)

Restricting the range of GPAs (to perfect 4.0) will not allow us to explore the relationship with the explanatory variables.

b. Describe a study design that would provide more useful information. (10)

Several responses are possible. For example, a simple random sample of students could be selected or a stratified sample to obtain equal numbers of students for each year could be obtained. Randomization must be used and the association between college GPA and the other explanatory variables should be examined.

4. An interviewer stands at the street corner and conducts interviews until obtaining a quota in various groups representing the relative sizes of the groups in the population. For instance, the quota might be 50 factory workers, 100 housewives, 60 elderly people, 30 Hispanics, and so forth. This is called **quota sampling**. Is this a random sampling method? Explain, and discuss the potential advantages and disadvantages of this method. (15)

No, this is not a random sampling method. The advantage is that the "quota" is attempting to represent the population. However, people who approach the street corner are interviewed as they arrive and agree to be interviewed so may not be representative of the population.

5. Raw scores on behavioral tests are often transformed for easier comparison. A test of reading ability has mean 75 and standard deviation 10 when given to third graders. Sixth graders have mean score 82 and standard deviation 11 on the same test. To provide separate "norms" for each grade, we want scores in each grade to have mean 100 and standard deviation 20.

a. What linear transformation will change <u>third-grade</u> scores x into new scores $x_{new} = a + bx$ that have the desired mean and standard deviation? (Use $b > 0$ to preserve the order of the scores.) (10)

$$\overline{X}_{new} = a + b\overline{X} \qquad 20 = b \cdot 10 \qquad a = \overline{X}_{new} - b\overline{X}$$
$$S_{new} = bS. \qquad \Rightarrow \boxed{b = 2} \qquad a = 100 - 2(75) \qquad \boxed{X_{new} = -50 + 2X}$$
$$\boxed{a = -50}$$

b. The linear transformation that will change <u>sixth-grade</u> scores x into new scores that have the desired mean and standard deviation is $x_{new} = -49.0924 + 1.8182x$. Nancy is a sixth-grade student who scores 78 on the test. What is her transformed score? David is a third-grade student who scores 78. Who scores higher within his or her grade? (10)

$$\boxed{6^{th}} \quad X_{Nancy} = -49.0924 + 1.8182(78) = 92.7272$$

$$\boxed{3^{rd}} \quad X_{David} = -50 + 2(78) = 106 \qquad * \text{ David has the highest score within his grade.}$$

6. The usual way to study the brain's response to sounds is to have subjects listen to "pure tones." The response to recognizable sounds may differ. To compare responses, researchers anesthetized macaque monkeys. They fed pure tones and also monkey calls directly to their brains by inserting electrodes. Response to the stimulus was measured by firing rate (electrical spikes per second) of neurons in various areas of the brain. The file p:\data\math\stats\monkey.mtw contains the responses for 37 neurons.

a. One notable finding is that responses to monkey calls are generally stronger than responses to pure tones. Give a numerical measure that supports this finding. (5)

	Mean	Median	Q1	Q3
Pure Tone	106.2	72	38	155.5
Monkey Call	176.6	141	91	205.5

b. Find the least squares line for predicting monkey call response from pure tone response. (5)

$$\hat{Call} = 93.92 + .7783 \, Tone$$

c. Identify and interpret the value of r^2. (10)

R-squared = .408 or 40.8%
40.8% of the variability in ~~aldatt~~ Monkey calls can be explained using linear regression and the explanatory variable pure tones.

d. Identify the point with the largest residual. (5)

The 3rd point (241, 485) has the largest residual (203.5045)
tone   call

e. One point is an outlier in the x direction. Identify this point. How influential is this point on the correlation coefficient? (10)

The 1st point (474, 500) is an outlier in the x direction. The value of r = .639 with this point, but r drops to .479 when this point is removed. This point obviously has a major impact on r.

f. Would you be willing to use your least squares regression line from part (b) to predict the monkey call response when the pure tone response is 550? Explain. (5)

No, the tones range from 19 to 474 (the outlier) in the data set so this would go well beyond the range of available data. The regression line should not be used for extrapolation.

7. Different varieties of the tropical flower *Heliconia* are fertilized by different species of hummingbirds. Over time, the lengths of the flowers and the form of the hummingbirds' beaks have evolved to match each other. The data on the lengths in millimeters of three varieties of these flowers on the island of Dominca are in the file p:\data\math\stats\heliconia.mtw. Use visual displays and descriptive statistics to compare the three distributions. Your comparison should address center, spread, and shape of the three distributions. What are the most important differences among the three varieties of flowers? (30)

Descriptive Stats for Length (in mm)

Variety	N	Mean	s	Min	Q1	Median	Q3	Max	IQR.
bihai	16	47.598	1.213	46.34	46.69	47.12	48.293	50.26	1.6025
red	23	39.711	1.799	37.40	38.07	39.16	41.69	43.09	3.62
yellow	15	36.180	.975	34.57	35.45	36.11	36.82	38.13	1.37

Boxplot.
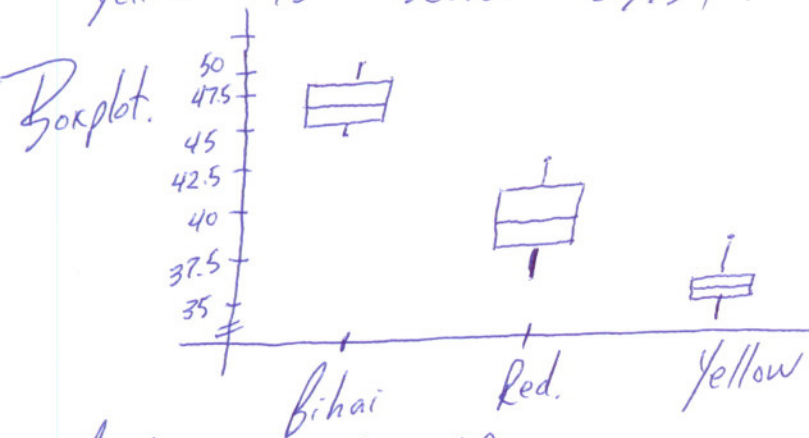
Students do NOT need to provide this rough sketch but they must comment on center, shape, & spread.

Most important differences:

1. The different varieties have different lengths. — Yellow is the shortest, red is next, and bihai is the longest. (The means, medians, or any other measures of location clearly show this fact.)

2. The variability (as measured by IQR — ht of box — or std deviation) is highest for the red variety. The yellow has the lowest/smallest variability and bihai is in between.

3. The distribution of lengths for the yellow variety is symmetric. The other two distributions (red and bihai) are skewed to the right.