Statistics 250, Section 701, Final (150 Points)

June 18, 1998, Dr. Jürgen Symanzik

Your Name: _____

Question 1: Short Answers (**44 Points**)

1. For $x_1 = 1, x_2 = 4, x_3 = -10, x_4 = 0, x_5 = 3, x_6 = 2$, and $n = 6$, determine the following summary statistics. Indicate which formula you use: (**15 Points**)

   mean:

   median:

   range:

   sample variance:

   standard deviation:

2. For the numbers given in (1) above, determine the following sums: (**9 Points**)

$$\sum_{i=3}^{n} x_{(i)} =$$

$$\sum_{i=3}^{n-2} i \cdot x_i =$$

$$\sum_{i=1}^{n-3} \frac{x_i}{x_{(n-i)}} =$$

3. Similar to the median (when compared to the mean), there exists a more robust measure of variability (when compared to the variance) that is not affected too easily by unusual large or small values. This value is called the *median absolute deviation* (MAD) and it is defined as **the median of the absolute deviations from the median**, in symbols

$$\text{MAD} = median\{\ |x_i - \tilde{x}|,\quad i = 1, \ldots, n\}.$$

Calculate the MAD for the numbers given in (1) above. (**6 Points**)

4. Determine the slope and the y–intercept of the lines whose equations are given as: (**4 Points**)

(a) $5x - 4y = 20$

(b) $x + 7y = 14$

5. The Spanish lottery has a game called "8 out of 34". You make a selection of 8 numbers between 1 and 34 and you win the big prize if exactly these numbers are drawn in the weekly drawing. How many different combinations are possible to select 8 out of 34 numbers? Calculate this value! (**4 Points**)

6. Do you like this graphic as much as I do? Isn't it a perfect example from Huff's "How to Lie with Statistics"? In addition to the fact that we don't know what the numbers stand for, there are several design violations with respect to how to create statistical graphics. Name at least **TWO** of the violations in this graphic. (**6 Points**)

Question 2: Binomial Probability Distributions (**35 Points**)

In a Stat 250 quiz there are 5 true–false questions. Since the instructor got suspicious that students identified his strategy in asking questions, he flipped a coin for each question and then formulated the question according to the outcome of the coin toss (where we assume that head relates to true and tail to false). Let us also assume that students are only guessing when answering these questions.

1. Indicate in how many different ways these questions can be answered. Do **NOT** list the individual cases, just provide a (correct) number. (**3 Points**)

2. Indicate the following probabilities: (**4 Points**)

   $P(\text{Answer to Q1 is true}) =$

   $P(\text{Student answers Q1 correctly}) =$

3. Are the events "Answer to Q1 is true" and "Student answers Q1 correctly" dependent or independent. Explain why. (**5 Points**)

4. Let X be the random variable that describes the number of correct guesses when answering 5 true–false questions. Complete the follow formula that relates to the probability distribution of X: (**3 Points**)

   $X \sim$

5. Calculate the mean and the variance of X. Indicate which formula you use. (**6 Points**)

6. Determine the following probabilities. Recall that $0! = 1$. Indicate which formula you use. Do **NOT** round your results. (**6 Points**)

$$P(X = 4) =$$

$$P(X = 5) =$$

7. In this particular Stat 250 class, all 34 students randomly guess their answers to these 5 true–false questions. What is the expected number of students in this class that answer 4 or all 5 questions correctly? You may want to introduce a new random variable Z to answer this question. (**8 Points**)

Question 3: Probability (**17 Points**)

Instead of studying for a quiz, one clever (?) Stat 250 student (in a class of 34 students) decided to believe in his neighbors' solutions. Unfortunately, this poor student wasn't aware that the question has been formulated in such a way that only 2 out of 33 students (which does not include the answer of this particular student) will answer the question correctly while all other 31 students will answer the question incorrectly. In this classroom, all students have been randomly assigned to a seat and our particular student doesn't know anything about the performance of his/her neighbors.

1. In a first attempt, our student decides to look at his/her left neighbor's quiz and copy the solution from this neighbor. What is the probability that the student with this test taking strategy #1 will answer the question correctly? (Note: To be able to answer part (3) below, you should work with fractions, e.g., 1/3, rather than with decimals, e.g., 0.333.) (**3 Points**)

2. Not really satisfied with his odds of answering the question correctly when following his/her test taking strategy #1, the student decides on the following test taking strategy #2 to increase his/her odds: Look at your left and right neighbors' solutions. If these are identical, then copy (either) solution. If they are different, then flip a coin to decide who's solution to copy. One can assume that the 2 correct answers are identical and all 31 incorrect answers are identical as well (but obviously they are different from the correct answers). What is the probability that the student with this test taking strategy #2 will answer the questions correctly? A tree diagram might be helpful in determining your answer. (**11 Points**)

3. So, which test taking strategy do you recommend? (**3 Points**)

Question 4: Linear Regression, Correlation & WebStat (**30 Points**)

10 students from Stat 250 wanted to determine the thickness of paper sheets used for the photocopying machines in S&T II. They grabbed a random number of sheets, counted the number of sheets, and measured the thickness of this pile of paper. Here is the data:

x = # Sheets	y = Thickness [mm]
111	7.3
149	9.2
117	7.2
165	10.0
238	14.1
145	8.8
84	4.8
155	9.5
94	6.2
413	24.8

Whenever you answer the following questions, indicate the formula you use to get full credit!!!

1. Draw a scatterplot of Sheets (horizontal x–axis) and Thickness (vertical y–axis). (**3 Points**)

2. Fit a least squares (linear regression) line to the data. Make clear what your variables stand for. Also, add this line to your plot in (1). It might help to know that:

$$\sum_{i=1}^{n} x_i = 1671, \sum_{i=1}^{n} x_i^2 = 363591, \sum_{i=1}^{n} y_i = 101.9, \sum_{i=1}^{n} y_i^2 = 1332.79, \sum_{i=1}^{n} x_i y_i = 22006.2$$

(**9 Points**)

3. Calculate Pearson's correlation coefficient r between x and y. How can we interpret this value for our given data set? (**5 Points**)

4. Based on your calculations in (2), what is the predicted thickness of 10 sheets, 150 sheets, 400 sheets, and 800 sheets. (**4 Points**)

5. Dr. S. used WebStat 1.0 to provide the answer to this question. The following WebStat 1.0 output has been produced for this sheet/thickness data set:

   Write down the regression equation based on this output. Make clear what your variables stand for! (**4 Points**)

6. Compare the WebStat 1.0 least squares regression equation from (5) above with your calculations from (2) above. Did you calculate exactly the same result? If you think the WebStat result is correct but your numbers are slightly different (most likely due to rounding) state **YES** to indicate that everything is OK. If you think something is wrong with the WebStat result, then state **NO** and indicate what is wrong. (**5 Points**)

Question 5: Scatterplot Matrix & Linked Brushing (**24 Points**)

The questions on the next page are based on the scatterplot matrix presented in William Cleveland's book "Visualizing Data". This scatterplot matrix has been reprinted below. It displays trivariate data that represents measurements of "Abrasion Loss", "Hardness", and "Tensile Strength" for 30 rubber specimens. "Abrasion Loss" is the dependent variable and "Hardness" and "Tensile Strength" are the independent variables. The goal of this study was to determine conditions that minimize the "Abrasion Loss".

1. Label the (individual) scatterplot that shows the "Tensile Strength" on the vertical (y–)axis and the "Abrasion Loss" on the horizontal (x–)axis with the letter "A". (**3 Points**)

2. What is the (approximate) range R of "Tensile Strength"? (**3 Points**)

3. We know an approximation of s through R. Indicate the correct formula and calculate s based on your answer in (2) above. (**3 Points**)

4. Which of these statements is correct/incorrect?

   (a) The brush is located in the scatterplot that shows the "Tensile Strength" on the vertical (y–)axis and the "Hardness" on the horizontal (x–)axis. (**3 Points**)

   (b) A straight line can be fit reasonably well to the data in any of the scatterplots that shows "Abrasion Loss" and "Hardness". Therefore, the correlation coefficient r between these two variables must be positive, i.e., $r > 0$. (**3 Points**)

   (c) The high values of "Tensile Strength", i.e., values in the range 200–240 have been brushed. (**3 Points**)

   (d) When "Hardness" is fixed at any given level, lower values of "Tensile Strength" result in a much higher value for "Abrasion Loss" than higher values of "Tensile Strength". (**3 Points**)

5. Decide on **ONE** of the following options. The best way to minimize "Abrasion Loss" is a combination of (**3 Points**)

   (a) high "Hardness" and high "Tensile Strength"

   (b) high "Hardness" and low "Tensile Strength"

   (c) low "Hardness" and high "Tensile Strength"

   (d) low "Hardness" and low "Tensile Strength"