Practice Exam Math 10 Spring 2006

1. (10 points) The scatter plot shows five data points to the right given by dots. Not suprisingly the correlation for these five points is $r = 0$. Suppose *one* additional data point is added at one of the five positions indicated by a triangle. Match each of (a-e) wit the correct new correlation form the list given.

   (a) -0.87 (Answer: E)

   (b) -0.05 (Answer: B)

   (c) 0 (Answer: C)

   (d) -0.28 (Answer: A)

   (e) 0.84 (Answer: D)

   (Comments: By symmetry C is very near zero and the only candidate to actually be 0. The sign of the others should be clear by sketching an intuitive line of best fit. By thinking in terms of point leverage, D will be should influential result in a comparatively large $|r|$ value which is positive, while E should be influential result in a comparatively large $|r|$ value that is negative. B has very little leverage with regards the points and should have result in a very small $|r|$ value that is negative probably -0.05. This leaves A, which by elimination must be -0.28. Certainly the difference between A and B is by far the hardest to determine with the point choices here. The easiest way to convince on self that A should have smaller r value is perhaps to sketch a standardized picture. (This one was slightly harder than I intended.)

2. (10 points) Tell what each plot of the residual plots (Problem 2 figures 1-3) indicates about the appropriateness of the linear model.

   (Answers:

   Figure 1: A linear model may be appropriate, but one should be careful! The reason is that the the residuals have a variable variance, and hence we cannot easily use the variance of the errors as estimated by the correlation coefficient $(1 - R^2)$ to help us interpret a value predicted by our model. In this case, the predictive power of the model would be much better near the origin.

   Figure 2: We should not use a linear model. These residuals tell us that the relationship has a strong non-linear association.

   Figure 3: A linear model looks extremely appropriate and we can easily interpret the content a value predicted by our model.)

3. (10 points) Every normal model is defined by its parameters, the mean and the standard deviation. For each model described below, find the missing parameter. (You would be given the table on page A-98 if this problem were on the actual exam).

(a) $\mu = 80$, 30% are below 70, $\sigma =$? (Answer: Draw a picture! Using the table on page A-98 we have $70 = 80 + \sigma z_{.3} = 80 - 0.525\sigma$. So $\sigma = \frac{-10}{-5.25} \approx 19.0$.)

(b) $\sigma = 12$, 10% are below 210, $\mu =$?. (Answer: Draw a picture! $210 = \mu + 12z_{0.1} = \mu - (1.28)12$, or $\mu = 210 + (1.28)12 \approx 225.4$.)

4. (15 points) How fast do squid swim? Atlantic Derby winners were all greater than 5 nautical miles per hour, as shown in the graph. In fact, this graph shows the percentage of Derby winners that have swam slower than a given speed.

   (a) Estimate the median winning speed. (Answer: $\approx 6.75nmph$)

   (b) Estimate the quartiles. (Answer: $Q_1 \approx 5.75nmph$ while $Q_3 \approx 7.75nmph$).

   (c) Estimate the range and the IQR. (Answer: $Range \approx 11-5 = 6nmph$ while $IQR = 7.75 - 5.75 = 2nmph$

   (d) (5 points) Is the mean bigger or small than the median. Why? (Answer: Scketch and histogram and note that their is a right tail, so the mean is bigger than the median.)

5. (10 points) The figure Problem 5 figure 1 is a histogram of assets (in million of dollars) of 79 companies chosen from the Forbes list of the nation's top corporations.

   (a) What aspect of this distribution makes it difficult to summarize, or to discuss, center or spread?

   (Answer: Well the histogram is extremely skewed to the right complicating a meaningful discussion of center and spread.)

   (b) In figures Problem 5 figures 2 and 3 are the same data after re-expressing as the square root of assets and the the logarithm of assets. Which re-expresion "should" you prefer and why?

   (Answer: The more uni-model and symmetric the easy to discuss, hence Figure 3 the logarithm would be the preferred re-expression.)

   (c) In the square root re-expression, what does the value 50 actually indicate about the company's assets?

   (Answer: Typo there was no 50! However if there were then this value would correspond to $(50)^2 = 2500$ million of dollars.)

   (d) In the logarithm re-expression, what does the value 3 actually indicate about the company's assets?

   (Answer: This depend on which logarithm we intended. I suppose in high log would mean the base 10 logarithm, so it corresponds to $10^3 = 1000$ millions of dollars.)

6. (10 points) A 1975 article in the magazine *Science* examined the graduate admissions process at Berkeley for evidence of gender bias. The table below shows the number of applicants accepted to each of four graduate programs.

$Program$	$MalesAccepted$	$FemalesAccepted$
1	511 of 825	89 of 108
2	352 of 560	17 of 25
3	137 of 407	132 of 375
4	22 of 373	24 of 341
$Totals$	1022 of 2165	262 of 849

(a) What percent of the total applicants were admitted? (Answer: $\approx$ 42.6%)

(b) Overall, were a higher percentage of males or females admitted?
(Answer: The percent males admitted was $\approx 47.2\%$, while the percent Females admitted $\approx 30.9\%$. So a significantly higher percentage males were admitted.)

(c) Compare the percentage of males and females admitted in each program.
(Answer:

$Program$	$MalesPercent$	$FemalePercent$
1	$\frac{511}{825}100\% \approx 61.9\%$	82.4%
2	62.9%	68.0%
3	33.7%	35.2%
4	5.9%	7%

Females were accepted at a higher rate in every program. Though only in program 1 does this difference rather significant.)

(d) Which of the comparisons you made do you consider to be the most valid? Why?
(Answer: This depends on the point being made. Certainly the original average is misleading since there is a confounding factor, namely departments. If one was claiming the people making admission decisions were gender biased then clearly the original percentage is very misleading and one would need to use the controlled percentage to support such an argument. However, given only the percentages computed after controlling for the confounding factor would also be misleading and not really "valid". For example, it is clear that the "supply" offered by the university to satisfy the "demand" by male applicants is much larger than the "supply" offered by the university to satisfy the "demand" by female applicants. Without presenting both percentages this fact becomes obscured, and if this were the point being made then the controlled percentage would obscure the

reality of the situation. So in isolation neither the single percentage nor the collection of departmental percentage can give us the whole story. Hence, neither can be viewed as entirely "valid", and the validity of each will be very context dependent.)

7. (15 points) To study the perception of age among college students you had 67 Dartmouth students participate in a version of our *Picture Experiment* that we performed in class. Each subject was presented with 10 pictures of people whose actual ages in years are gioven by

$$Actual = [17, 25, 29, 36, 42, 47, 50, 58, 63, 67].$$

To entice people to participate you advertise that the best guess will receive a gift certificate for a VerMonster ™ at Ben and Jerry's. Foolishly, you forgot to account for the fact that their might be more than one "best guess", and two participants (Viridian and Insomnia) tied for your assessment of the best guess:

$$Viridian = [19, 22, 32, 36, 41, 44, 53, 55, 60, 65]$$

$$Insomnia = [17, 24, 30, 35, 42, 45, 58, 54, 66, 70]$$

(a) You had intend to use the "simplest" ranking system and it resulted in this tie. What system did you use?

(Answer: Well it would probably have used total absolute error. The errors are given by $Viridian - Actual = [2, -3, 3, 0, -1, -3, 3, -3, -3, -2]$ while $Insomnia - Actual = [0, -1, 1, -1, 0, -2, 8, -4, 3, 3]$ Hence the absolute errors are both 23 and we would have a tie.)

(b) Describe a second ranking system that separates Viridian's and Insomnia's guesses. Based on our class data who will probably win and why?

(Answer: As we saw, some pictures were much harder to guess the ages of than others. Hence we might want to re-scale our absolute errors to compensate for how difficult a given picture was to guess. A reasonable choice would be to re-scale by the standard deviation of the guesses. From our experience the older samples will have a higher standard deviation, hence Insomnia would probably win.)

(c) Construct (fake!) summary statistics for the whole population that confirm (or deny) your suspicions in part (b).

(Answer: For simplicity, imagine that the standard deviation of the guesses are $Sd = [3, 3, 3, 3, 3, 6, 6, 6, 6, 6]$. In this case Insomnia's rescaled error is 4.3 while Viridian's is 5.3 and Insomnia indeed wins.

(d) You use your new ranking system to break ties and you contact the winner, tell her the Actual ages, and give her the Ben Jerry's gift certificate. Unbeknownst to you Viridian and Insomnia are roommates!

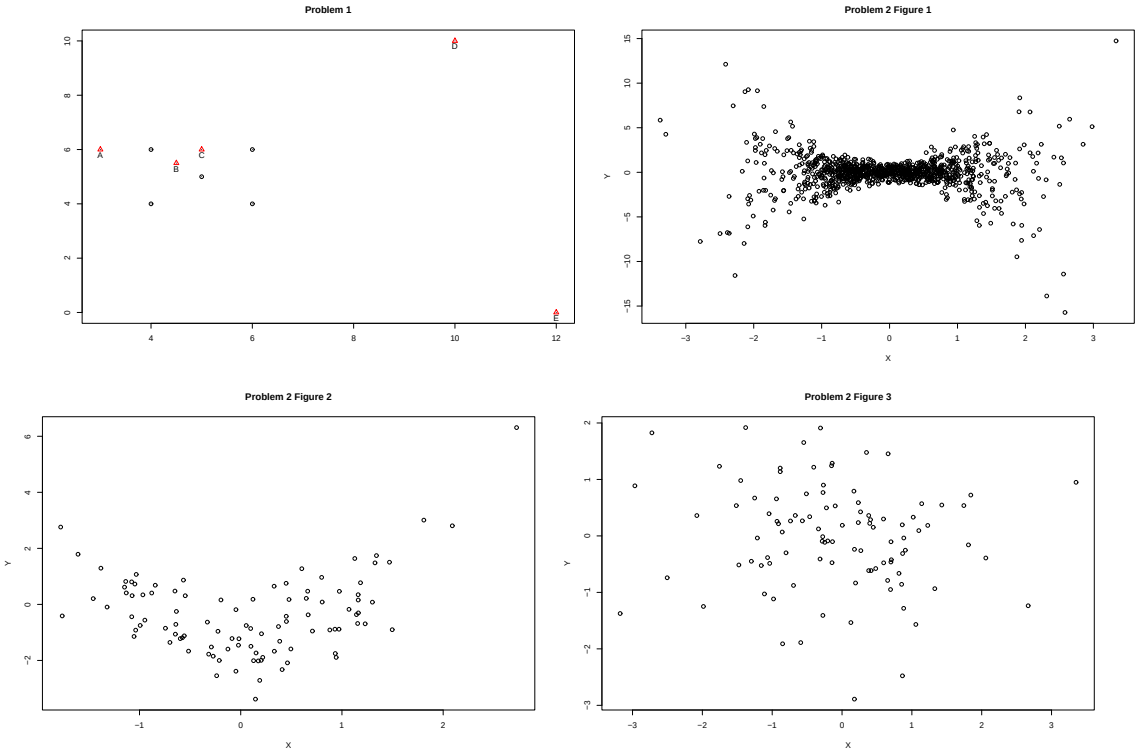
The loser is confused, upset, and without ice cream. So she contacts to you. How do you explain your system to her?

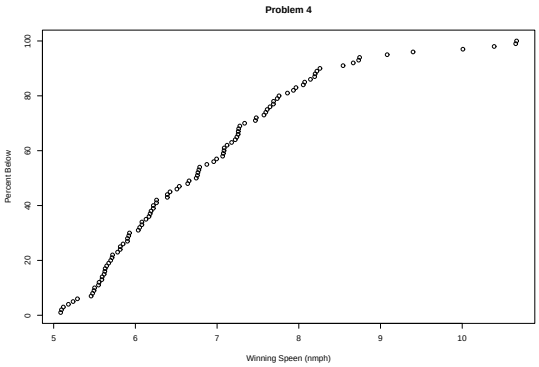
(Answer: Dear Viridian,

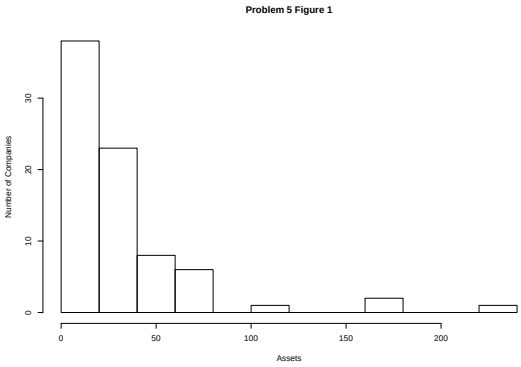
We wanted to use a fair ranking system, so we obviously did not simply use the total absolute errors. This would make no sense since the pictures were NOT equally hard to guess. It would be like ranking college football teams with out looking at their schedules. For example, in this case missing the second picture by 3 years was a rather terrible guess since in general this was a very easy picture to guess. The fact that you did poorly on the easy to guess pictures is why your ranking was worse than that of acquaintance Insomnia. We apologize for the stress this has caused, and in our experience we have found that 20 scoops of ice cream, hot fudge, banana, cookies, brownies, and all of your favorite toppings is too much for a single person to eat in one sitting - so perhaps Insomnia will choses to share her well earned VerMonster with you.

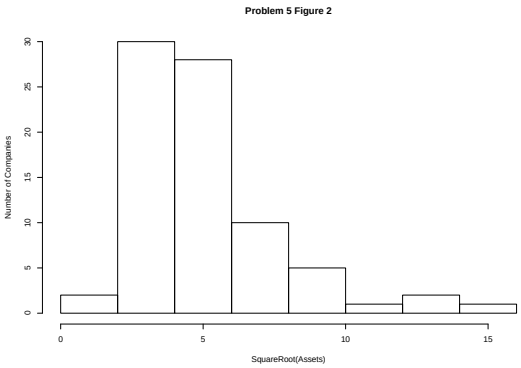
Sincerely,

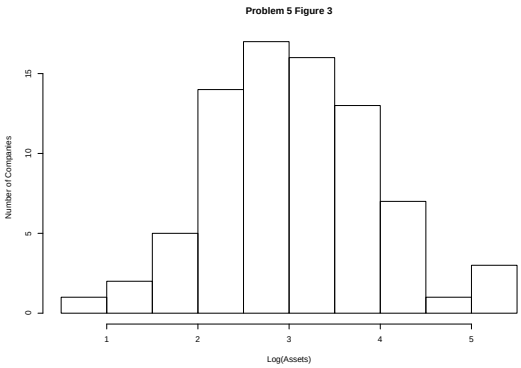
Picture Guess Ranking Committee)




Problem 4
Percent Below
Winning Speen (nmph)


Problem 5 Figure 1
Number of Companies
Assets


Problem 5 Figure 2
Number of Companies
SquareRoot(Assets)


Problem 5 Figure 3
Number of Companies
Log(Assets)