Sample Final Exam

Statistics 280

Spring 1998

Directions

1. This is an Open Book, Open Notes, Open homework solutions, open whatever Exam.

2. You have 3 hours.

3. There is to be no sharing of calculators, books, or anything else during the exam.

4. The exam is worth 200 points. The value of each question is in square brackets after the problem number.

5. All plots appear on the final pages of the exam.

1. [50 points] For each of the statements below, circle **T** or **F** for "True" or "False," respectively. (5 pts. each)

T F : Probabilities are usually between 0 and 1, but can be any number.

T F : For inferences in a regression model with $p = 3$ predictor variables and an intercept, use the t distribution with $n - 1$ degrees of freedom.

T F : A P-value is a parameter.

T F : When testing a null hypothesis $H_0 : p = p_0$ about a population proportion p, use the standard error $\sqrt{p_0(1 - p_0)/n}$.

T F : With "before and after" data, one should use the two sample methods described in section 7.2 of the text, with one sample being the before data and the other being the after data.

T F : If the correlation is high, then it is not necessary to check the validity of a regression with scatterplots and residual plots.

T F : A *statistic* is an unknown quantity associated with the population.

T F : The probability of exactly 2 heads in 4 independent flips of a fair coin is 0.5.

T F : The P-value for a 2 sided test is $1 - C$ where C is the confidence level of a confidence interval.

T F : For the simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, the ANOVA F-statistic is the square of the t-statistic for testing the Null Hypothesis that the slope is 0.

SPECIAL EXTRA CREDIT T/F QUESTION:

T F : Statistics is incredibly easy and loads of fun.

2. [30 points] (Possible Final Project.) A local radio station KWHY claims they play more music than another station KNOT. A statistician decides to test this claim. He picks 50 random times during the week for each station and switches to that station and determines if they are playing music at exactly the time he switches the station on. He finds that KWHY is playing music at 22 out of the 50 times, and KNOT is playing music 28 out of the 50 times.

(a) What null and alternative hypotheses should the statistician test? Explain.

(b) Compute the appropriate test statistic and P-value, and determine if there is significant evidence against the claim made by KWHY.

3. [30 points] The director of information services at a cool school in the south central US believes that using computerized foreign language instruction is superior to the old method with human instructors. After his method is implemented, he makes his case to the administration with results on a standardized test of fluency for a particular language. On the year before his method was implemented, there were 49 students who completed Mongolian 101 and 102, and their average score on the test was 38 with a standard deviation of 14. In the year that the computerized method was implemented, 16 students completed both semesters of introductory Mongolian. Their average score on the standardized test was 50 with a standard devation of 24.

The Information Services Director says that the data show a statistically significant increase in scores for the computer taught students vs. the human taught students.

(a) Assuming that the data are independent samples from the populations of human taught and computer taught foreign language students, verify the IS Director's claim of statistical significance.

(b) What's wrong with this picture? Comment from the point of view of proper statistical experimental design and validity of assumptions.

4. [35 points] A random sample of 199 married British women are asked their height (in mm.) and age of marriage. (Note from D. Cox: I am not making this up. This is real data from a real sample.) A few refuse to reply to one or the other question, leaving 195 for which we have data. Below are given

1) the output of a regression analysis of these data with age of marriage as the dependent or response variable and height as the explanatory or predictor variable;

2) a plot of the residuals vs. x_i's.

3) a normal quantile plot of the residuals.

See last page for the plots. Use this information to answer the questions that follow.

Output from Stat Package:

```
Residual Standard Error = 5.6351,  Multiple R-Square = 0.0062
N = 195,  F-statistic = 1.2114 on 1 and 193 df, p-value = 0.2724

              coef std.err  t.stat p.value
Intercept  36.7218 10.3271  3.5559  0.0005
        X  -0.0071  0.0064 -1.1006  0.2724
```

(a) What is the predicted age of marriage for a woman who is 2000 mm. tall?

(b) According to the fitted regression model, do taller women tend to marry earlier or later?

(c) Comment on how well or poorly the regression model fits these data. Use all available information.

(d) A social scientist claims these data indicate there is no evidence that a woman's height has any bearing on the age at which she marries. A skeptic criticizes this conclusion, claiming, "There is evidence the assumptions of the regression model are violated." Discuss the pros and cons of each point of view.

5. [10 points] A random sample of 100 students at an exclusive, snobby, elitist private college on the east coast are asked their beliefs about whether or not sexual harassment is prevalent at their school. The results are summarized in the table below.

	Belief on sexual harassment:			
	Prevalent	Not Prevalent	Don't Know	Total
Male	8	30	12	50
Female	12	10	28	50
Total	20	40	40	100

(a) What is the value of Pearson's Chi-squared statistic for this table?

(b) Is there a statistically significant difference of opinion on the sexual harassment issue between the two genders?

(c) Extra Credit. Do you believe these are real data?

6. [20 points] Below is the 5 number summary of the age of marriage of the women in the data set described in the previous problem.

```
Five Number Summary:
```

Min	Q1	Med	Q3	Max
16	22	24	27	52

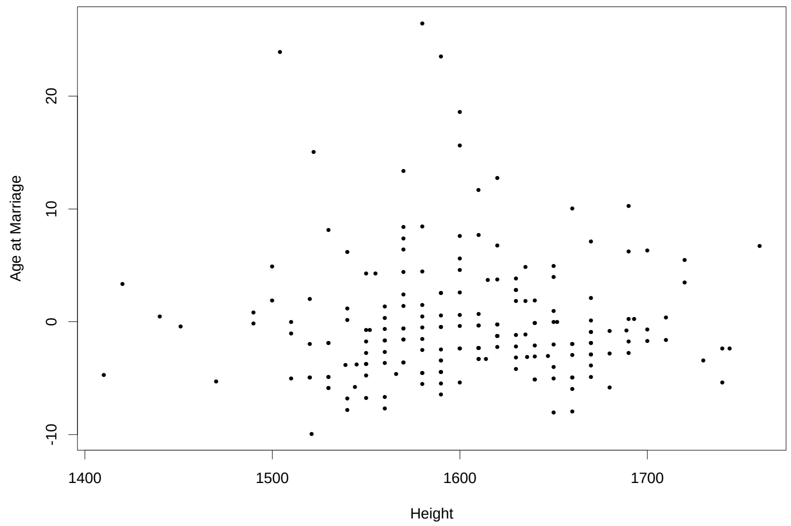
(a) Sketch a (density) histogram of the age of marriage using the available information.

(b) Which of the following do you think is probably true about these data: the mean and median are about the same; the mean is somewhat greater than the median; the mean is somewhat less than the median. Explain your answer.

7. [25 points] A baseball player has a lifetime record of making hits in 30% of his "at-bats" (that is, his batting average is .300). In the first playoff game, he has 7 at-bats (the game went 19 innings) but makes only 1 hit. The player is depressed about this and fears he is in a slump but the coach says it is just chance variation.

(a) State some more or less reasonable assumptions that will allow you to compute a probability that the player makes 1 or fewer hits in 7 attempts, and compute that probability.

(b) Comment on the validity of the coach's claim that the player's performance in the first playoff game is chance variation.



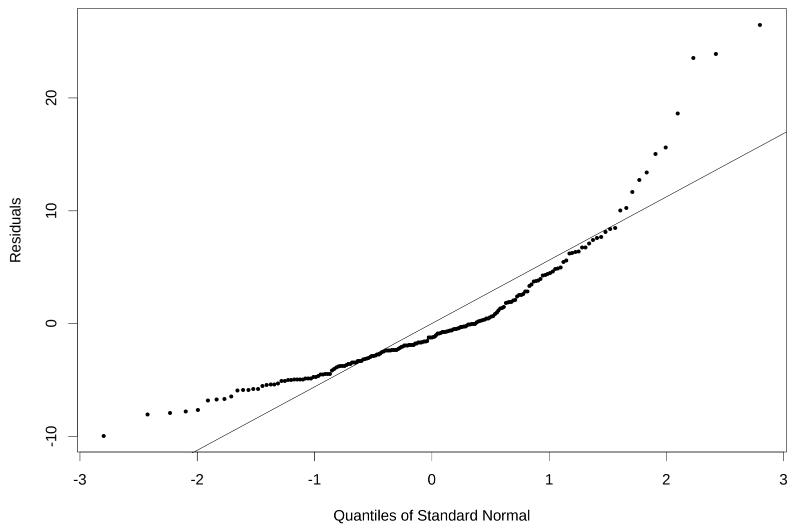


Figure 1: Plot of residuals vs. x values and Normal Quantile Plot of Residuals.