M358K, Spring 2006 (Smith)

REVIEW PROBLEMS FOR FINAL EXAM

Exam time/date/place: 2 – 5 pm, Monday, May 15, RLM 6.104

The exam will be comprehensive, but will focus mainly on material covered since the second exam

Suggestions for studying:

- Be sure to study items you missed on previous exams and homework.
- The following include examples of possible exam questions and other questions to help you review for the exam.
- Also review the review for the midsemester exams
- Remember that reasoning and giving clear, complete explanations are important.
- Especially if you did not do well on previous exams: Spend some time practicing writing out solutions to problems.

Exercises from the Book: 10.31, 10.32

1. An exam question in a previous class asked for what the following means (that is, what the definition of P-value for this test is):

> The P-value for a hypothesis test with hypotheses
> H_0: $\mu = 3$
> H_1: $\mu < 3$
> is 0.04.

Critique the following responses for clarity, completeness and correctness.

a. This means that the probability of getting our test statistic is .04.

b. This means that the probability of getting a test statistic at least as extreme as ours is .04.

c. This means that if the null hypothesis is true, the probability of getting a test statistics at least as extreme as ours is .04

d. This means that if the null hypothesis is true, the probability of getting a test statistic less than or equal to the one we got is .04

e. This means that it is very unlikely that the result that was used to compute this P-value would have happened by pure chance alone, assuming that H_0 is true. Therefore we could conclude that the evidence is against the Null Hypothesis , and H_0 is probably not true.

f. The sentence means that assuming the population average is equal to three, the likelihood of getting an average as large or larger than we got for our sample is about 4 percent.

g. The p-value is the probability that the data will be as extreme or more extreme as the alternate hypothesis suggests.

2. Another exam question asked for an explanation of what the following sentence means:

(2.25, 2.75) is a 99% confidence interval for the mean GPA of UT students having between 45 and 60 credit hours.

Critique the following responses for clarity and correctness.

a. A 99% confidence interval is used to show that 99% of the time when you pick a sample from the population (students having between 45 and 60 credit hours) you will find a mean GPA in the interval (2.25, 2.75).

b. This means that 99% of the confidence intervals we compute using the same method used here will contain the true population men. Or, there is a 99% chance that $2.25 \le \mu \le 2.75$.

c. This means that if we took many, many simple random samples and constructed a confidence interval based on each sample, 99% of the resulting confidence intervals would contain the true mean.

3. In developing the two-sample test comparing proportions, we found that the variance of the difference $\hat{p}_1 - \hat{p}_2$ of sample proportions had variance $\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$. The reasoning proceeded in two steps:

Step 1: $\text{Var}(\hat{p}_1 - \hat{p}_2) = \text{Var}(\hat{p}_1) + \text{Var}(\hat{p}_2)$

Step 2: $\text{Var}(\hat{p}_1) + \text{Var}(\hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$

Give the reason for each of these steps.

4. For a certain set of data collected for regression, you know x-bar = 1, $s_x = 2$, y-bar = 3, and $s_y = 4$. From this information, you know one point on the least squares regression line. What is it?

5. A study classified stock funds as winners or losers depending on whether their rate of return was less than or greater than the median for all funds. This was done for two years in a row. The following two-way table shows the results. You are interested in seeing whether funds that performed well the first year also performed well the second year.

a. Is one variable explanatory and the other response? If so, which is which?

b. Rephrase the question of interest in terms of the distributions of a categorical variable in two populations.

c. Rephrase the question in terms of comparing proportions of two populations.

d. Which type(s) of bar graph would be appropriate for studying this question? (a bar graph of row counts ? column counts? row percents? column percents?) Are your answers to part (b) or (c) relevant here?

	Second year	
First year	Winner	Loser
Winner	85	35
Loser	37	83

6. What kind of analysis (hypothesis test, confidence interval, prediction interval, correlation, regression; one or two sample procedure; z, t, or chi-squared statistic; other; none of the above) should be used in each of the following situations? Be as precise as possible (give all that apply – e.g., "hypothesis test in regression using t-statistic" and give reasons for your choices.

a. You have collected data on the amount of sodium in hot dogs with varying numbers of calories. You want to predict the average sodium content of hot dogs having 100 calories.

b. The following table shows the number of students in an introductory statistics class who received possible combinations of first exam letter grades and second exam letter grades. Is there a relationship between letter grade received on the first exam and letter grade received on the second exam?

	Second Exam Letter Grade					
First Exam Letter Grade	A	B	C	D	F	Total
A	11	7	5	0	2	25
B	9	16	4	8	2	39
C	0	6	6	2	1	15
D	1	2	3	2	2	10
F	0	2	0	1	5	18
Total	21	33	18	13	12	97

c. A randomized, double blind, six-month trial was conducted on the effectiveness and side effects of the drug prednisone on patients with Duchenne's muscular dystrophy (DMD). There were three treatment groups: placebo, low dose of prednisone, and high dose of prednisone. One possible side effect was weight gain. The table below shows the observed counts of weight gain by type of treatment and range of percent of weight gained.

	Weight Gain (percent)				
Treatment	<5	5-10	11-20	>20	Total
Placebo	18	10	5	2	35
Low dose	3	3	13	11	30
High dose	3	5	14	10	32
Total	24	18	32	23	97

Is there any relationship between treatment and percent of weight gain?

d. You are interested in whether wearing Brand A of running shoes will help runners run faster than Brand B. You find twelve runners to volunteer to help you test the shoes. Each runs a 10 kilometer race in each of two consecutive weeks. In one of the races the runners wear one brand of shoe and in the other race the other brand. Which brand they wear in which race is determined at random.

e. You want to find the average change in sales per additional $1000 spent on advertising by retail department stores.

f. Bags of a certain brand of tortilla chips claim to have a net weight of 14 ounces. A representative of a consumer advocate group wishes to see if there is any evidence that the mean net weight is less than advertised. She selects 16 bags of this brand at random and determines the net weight of each.

g. A committee of the American Academy of Pediatrics was interested in whether the proportion of family practice physicians prescribing tetracycline to children under eight varied according to whether their practice was in an urban, rural, or intermediate county. They found that 65 of 214 physicians in urban counties prescribed the drug, 90 of 226 in intermediate counties, and 172 of 330 in rural counties.

h. Researchers are comparing two methods of measuring the actual amount of a certain antibiotic in tablets. Two tablets were randomly selected from each of fifteen batches of tablets; one was analyzed by the first method, the other by the second method.

i. A graduate student working as a calculus TA was surprised to observe that many of her students who had had calculus in high school did not do well in calculus at the University. To study this further, she collected data from her classes. She asked students who had had calculus in high school what their high school calculus grades were and compared these with the grades they received in calculus at UT. Her results are summarized in the following table:

		UT Calculus Grade					
		A	B	C	D	F	Total
High School Calculus Grade	A	1	4	3	2	2	12
	B	2	4	5	1	0	12
	C	0	1	1	1	2	5
	D	0	0	1	1	2	4
	F	0	0	1	2	2	5
	Total	3	9	11	7	8	38

j. A school district has a voluntary busing program giving students from primarily minority schools the opportunity to be bussed to primarily majority schools. A researcher is interested in seeing if busing of minority students to majority schools influences their school achievement. At the end of the first year of the program, he selects a random sample of students who were bussed and a random sample of students who stayed at the primarily minority schools in order to compare the average score of the bussed students on the state end-of-year exams with those of the non-bussed students.

k. Two researchers believe that sending people a "prenotification letter" saying they will be receiving a survey before actually sending them the survey will increase the response rate to the survey. They test out their theory by randomly selecting two groups of people and sending those in the first group a prenotification letter before sending surveys to people in both groups. The results are shown:

Response	Letter	No letter
Yes	2570	2645
No	2448	2384
Total	5018	5029

l. In 1994, the US senate consisted of 7 women and 93 men. Find a 90% confidence interval for the proportion of women in the senate in 1994.

m. A teacher compares the scores of her students on a test before a certain method of instruction and after the instruction.

n. A teacher compares the scores of students using a computer-based method of instruction with the scores of other students using a traditional method of instruction.

7. For all the situations in Problem 5 that require hypothesis testing, state the null and alternate hypotheses.

8. For all the situations in Problem 5 that involve two or more variables, determine which (if any) are response variables and which are explanatory variables.

9. True or false. Explain.

a. In a matched pairs test, a P-value of .013 means the results of the two procedures being compared are equal in 13 out of every 1000 samples.

b. A 99% confidence interval is wider than a 95% confidence interval for the same parameter.

c. A larger population gives a larger margin of error in a confidence interval.

d. A larger sample gives a larger margin of error in a confidence interval.

10. A random sample of 79 companies from the Forbes 500 list (which actually consists of nearly 800 companies) was selected. For each company in the sample, data on sales (in hundreds of thousands of dollars) and profits (in hundreds of thousands of dollars) were collected. You wish to a) predict the mean profits for all companies that had sales of 500 thousand dollars, and b) predict the profits for a company that had sales of 500 thousand dollars. Explain the difference in these two problems and how you would solve them.

11. A college professor has developed a new freshman course that he hopes will teach students good learning skills. To evaluate the success of the course, he waits until a class of freshmen who have taken his course have graduated from college. He then computes the correlation between their scores on the final exam in his freshman course and their cumulative college GPA's. This correlation turns out to be 0.92. Thus students who did well in his course on average got high GPA's whereas those who did poorly in his course got low GPA's. Can he conclude that his course had a positive effect on students' learning skills?

12. A study of the popularity of a particular brand of aspirin was made by interviewers stopping people as they came out of drugstores and supermarkets where the brand was stocked. Despite the fact that those interviewed covered most ages, men and women, and people in a wide variety of occupations, the overall results showed clear bias: few people in the sample said that they used aspirin of any kind, whereas the known sales figures indicate that aspirin is widely bought in the population as a whole. What faults in sample design and execution might have caused this discrepancy?

13. If we regress x on y and y on x (for the same data), do we get the same least squares regression line? Explain.

14. The British Post Office claims that 94% of all first class letters posted for delivery in England and Wales are delivered the next day (counting Monday as the day following Saturday). In answer to an advertisement for a teaching position in Southwest England, 295 applicants from various parts of the country sent first class letters. A total of 253 of theses letters arrived on the day after mailing. A hypothesis test using theses data show that the proportion of the letters arriving on the day after mailing is significantly different from 0.94 at the 0.1% significance level. Explain why this result does not necessarily invalidate the Post Office's claim, and briefly describe an experiment which could be carried out the test this claim properly.

15. The table at the right shows the percent of sugar in some popular breakfast cereals.

a. Make a stemplot of these data. Describe the overall shape of this distribution. Are there any clear outliers?

b. Based on your stemplot, which numerical summary (mean and standard deviation or five-number summary) is more appropriate for this distribution? Why? Calculate this summary.

Product	% Sugar
All Bran	19
Alpha Bits	38
Cap'n Crunch	40
Cheerios	3
Corn Flakes	5
Golden Grahams	30
Grape Nuts Flakes	13
Post Toasties	5
Product 19	10
Raisin Bran (General Mills)	48
Raisin Bran (Kellogg)	29
Rice Krispies	8
Special K	5
Sugar Smacks	56
Wheaties	8

16. If you asked each student in a class how many pets they had ever had (in their entire life) and then made a graph (histogram) of the information you obtained, which of the following would it most resemble? Explain why.

a.							b.							c.							d.				
		X						X		X				X										X	X
		X						X		X				X	X								X	X	X
	X	X	X					X	X	X				X	X	X						X	X	X	X
X	X	X	X	X			X	X	X	X	X			X	X	X	X	X			X	X	X	X	X

17. When we compare proportions of two groups that have some characteristic, we can summarize our data as follows:

Population	Population proportion	Sample size	Count of successes	Sample proportion
1	p_1	n_1	X_1	$\hat{p}_1 = X_1/n_1$
2	p_2	n_2	X_2	$\hat{p}_2 = X_2/n_2$

If we are finding a confidence interval for the difference $p_1 - p_2$, we use one formula for the standard error (SE) of $\hat{p}_1 - \hat{p}_2$, but if we are doing a hypothesis test involving the null hypothesis H_0: $p_1 = p_2$, we use a different formula for the standard error.

a. What are these two formulas?
b. The formula used in the hypothesis test contains a term $\hat{p}$. Explain what $\hat{p}$ is, and why we use it in the second formula but not the first.

18. In the situation of Problem 6 (c):
a. If you were asked to use bar graphs to show the conditional distributions of weight gain by treatment group, would you need three bar graphs or four?
b. What is the marginal distribution of weight gain?

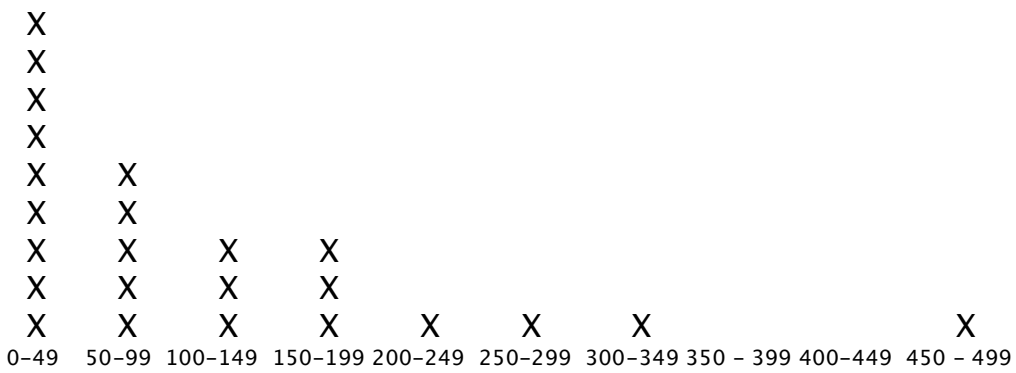
19. Which of the following stemplots shows the larger variance? Explain.

a.
```
1 | 4
1 | 8
2 | 2 4
2 | 6 8
3 | 0 1 2 3 4
3 | 5 6 7 8 9
4 | 2 4
4 | 6 8
5 | 4
5 | 8
```

b.
```
1 | 4 8
2 | 2 4 6 8
3 | 0 1 2 3 4 5 6 7 8 9
4 | 2 4 6 8
5 | 4 8
```

20. For each part, draw a scatter plot satisfying the conditions given, or else explain why the conditions are impossible:
a. Regression line has small positive slope and correlation is high and positive.
b. Regression line has large positive slope and correlation is high and positive.
c. Regression line has small positive slope and correlation is low and positive.
d. Regression line has large positive slope and correlation is low and positive.
e. Regression line has positive slope and correlation is negative.

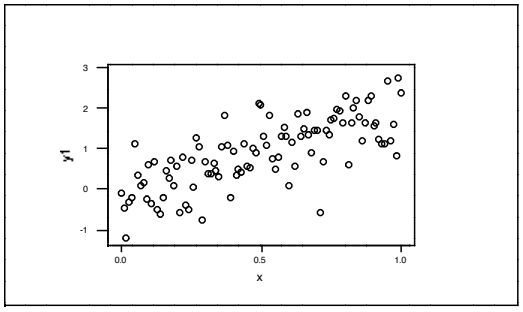
21. The students in a class were asked how many music CD's they own. The following graph shows the data collected:



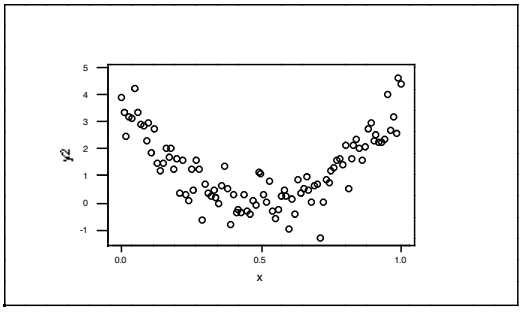
Would the mean or the median better describe the typical number of music CD's a student in the class owns? Why? How would the mean compare in this example? Why?

22. For which of the data sets shown in the following plots is the correlation coefficient a good descriptive statistics? Why?

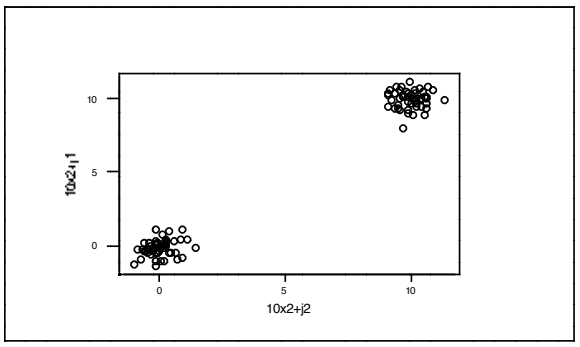
a.



b.



c.



23. Explain the difference between a parameter and a statistic, and give several examples of each.