STAT 201
Midterm Examination

Richard Lockhart 21 October 2004

Instructions:

1. This is a closed book exam.
2. You may use a calculator (with no wireless communications ability).
3. You may bring one sheet of notes.
4. You may also bring the tear out sheet of tables and notes from the text.
5. Sometimes, to do the problem, you will need to make assumptions. You should be clear and explicit about what assumptions you need to make the technique you are using reasonable.
6. The exam is out of 30.
7. **DON'T PANIC.**

1. In a study of dietary fat intake 1000 father-son pairs were examined. (Both father and son were adults in all pairs.) For each person the percentage of dietary calories received in the form of fat is measured. The fathers received an average of 35% of their dietary calories in the form of fat with an SD of 6 percentage points. The sons received an average of 30% of their dietary calories in the form of fat with an SD of 8 percentage points. The correlation between father and son was r=0.4.

   (a) About what percentage of the fathers receive more than 40% of their daily caloric intake in the form of fat? Be clear about any assumption you must make to do the problem. (4 marks)

   A: *in order to do this problem it is necessary to assume that the histogram of father's daily caloric intake is roughly normal. It is NOT enough to write the word 'normal' on the page without using a full clear sentence to say what has to be normal.* **1.5 marks**. *Convert 40 to standard units (***1.5 marks***) to get*

$$z = \frac{40 - 35}{6} = 0.83$$

   *Then look up the area to the left of z (***1 mark***) and get*

$$0.7967 \approx 79.7\%$$

   The desired area is the area to the right of z which is 1-0.7967 or about 20.3%. I am not concerned with how many digits students round off to.

   Related Problems: *Homework 1, # 5,6,7,8.*

(b) If a father receives 28% of his calories in the form of fat about what percentage should you predict for the son? (5 marks)

A: *This is a regression problem. Students may proceed in one of two ways:*

i) The estimated slope is

$$b = rs_y/s_x = 0.4\frac{8}{6} = 0.5333$$

The intercept is

$$a = \bar{y} - b\bar{x} = 30 - 0.5333 \times 35 = 11.33$$

Then plug in $x = 28$ *to the equation* $y = a = bx$ *to get the predicted value*

$$\hat{y} = 26.26$$

*ii) Alternatively: convert 28 to standard units to get (28-35)/6 = -1.17. Then multiply by r to get -0.467. Convert back to standard units for sons: 30-0.4667*8 =26.27.*

The differences are unimportant round off errors.

Take significant marks away for confusing the two standard deviations or the two averages. This problem is easy except for figuring out which number is which. Any mistake of this sort should give no more than 2.5 marks out of 5.

Related Problems: *Homework 2, # 5, 6, 7, 8.*

(c) Suppose we select 36 families at random from this group of 1000 for a more detailed dietary assessment. You may assume that they are selected with replacement so that the selections are independent. What is the chance that the 36 selected fathers have an average between 34 and 36% of their daily dietary calories in the form of fat. (4 marks)

A: *This is a question about the sampling distribution of the mean of a sample of* $n = 36$. *Students should indicate that they need to assume that 36 is not too large compared to 100 or that they assumed we were sampling with replacement.*

The mean of the sampling distribution of $\bar{x}$ *is* $\mu = 354$. *The SD is* $\sigma/\sqrt{n} = 6/6 = 1$.

Now convert 34 and 36 to standard deviation units to get -1 and 1. The area between these is 0.6826 or roughly 68.3%

A continuity correction is not appropriate. (Take off only 1 mark.)

Students who use $n = 25$ *and so get 1.2 instead of 1 for standard deviation will have -0.83 and 0.83 so get an area of 0.5935.*

Related Problems: *Homework 3, # 17, 18, 19, 20.*

(d) In 50 of the families the father received less than 15% of his daily caloric intake in the form of fat. If we eliminate this group of 50 father-son pairs from our study will the correlation coefficient go up or down; that is, is the correlation coefficient for the other 950 pairs more than 0.4, less than 0.4, or still about 0.4? Explain with a graph. (2 marks)

A: *The correlation will go down if you chop off the left part of the scatterplot; the new plot will be less bunched up around the regression line relative to the spread in the x direction. This question is like the discussion I had in class about combining data for men and women on height versus weight. The correlation coefficient in that case for both sexes put together is higher than in the individual sexes because putting the two together makes the overall plot look more like a long thin oval (the shape of a scatterplot with a high correlation).*

Students' answers should include a graph.

Related Problems: *Problem set 1, # 10 and the discussion in class of the effect on correlation between height and weight of putting men and women together.*

(e) Consider the families where the father receives about 28% of his calories in the form of fat. Approximately what would be the standard deviation of the sons' percentage of daily caloric intake in the form of fat in these families? (1 mark)

A: $\sqrt{1-r^2}\sigma$ *or* $\sqrt{1-0.4^2} \times 8$ *or 7.33.*

Related Problems: *this problem was intended to be a little one mark throw away based on my discussion in class but few students succeeded because there were no problems related to this concept. I should not have put this problem on the exam.*

2. An executive of a large supermarket chain discovers that the correlation between the total amount of overtime worked by cashiers at a store and the total number of bad cheques accepted at a store is 0.7. He recommends that a ban be placed on overtime, arguing that cashiers at the end of a long day are less careful. What is wrong with this thinking; your answer should include an alternative explanation of the observed correlation. (2 marks)

A: *This correlation is probably largely produced by the fact that if the rate at which bad checks are presented is constant then workers doing more hours will see more cheques and get more bad ones. You need to compare acceptance rates (cheques per hour worked) with hours worked by individual cashiers to see if any change in overtime hours might be useful.*

Related Problems: *Problem set 2, #9.*

3. A large bank has loans outstanding on 100,000 pieces of real estate. At the last audit the average assessed value of the pieces of real estate was \$150,000. The bank suspects that recent economic events mean that the real estate values may have fallen suddenly. A simple random sample of 400 of the outstanding loans shows an average present value of \$139,600 with an SD of \$160,000.

   Looking more closely at the data collected the bank president goes through the files to find the assessed values of the 400 sampled pieces of real estate at the time of the last audit. The figures average \$145,000 with an SD of \$165,000. He discovers, however, that those properties valued at over \$400,000 at the last audit have decreased in value \$20,000 each on average while those valued at under \$100,000 have not decreased in value at all on average. He develops the following explanation of this observation. Owners of expensive properties have had to sell them. They have then taken the

proceeds of the sales and bought less expensive properties thereby keeping up the prices of these properties. Identify a pitfall in the executive's reasoning. (2 marks)

A: *The effect noted here is the regression effect. The executive picked out properties which are above average in value at last audit. In essentially any scatterplot with a positive relation between x and y the units on the right hand side (above average in x will be above average in y, but not so much in terms of standard units.*

Related Problems: *There were no homework problems directly about the regression effect but there was substantial discussion in class.*

4. A forester lays out 10 small plots of land scatterd over a large area of recently logged forest land. In each plot 20 seedlings are planted. The forester returns a year later and counts the total number of the 200 seedlings which are still alive. S/he plans to treat the number alive as having a Binomial distribution. Is this wise? Why or why not? (2 marks)

   A: *No, this is not wise. Seedlings in the same plot are likely to have soil conditions and so on which are more similar than when comparing two seedlings in different plots. You may think of this either as dependence or as a probability of success which depends on the plot the seedling is in.*

   Related problems: *Problem set 3, # 10, 11, 12.*

5. A company manufactures steel rods. Each day it produces 50000 rods. The day's production is judged acceptable if fewer than 10% of the 50000 rods have a breaking strength less than some specified level. In order to check this a simple random sample of 100 rods is tested at the end of each day. Assuming that in fact the day's production is just barely acceptable–that is, that 5000 of the 50000 rods have too low a breaking strength–what is the chance that 14 or more of the 100 sampled rods fail the strength test? I am expecting an approximate, not exact, answer. (5 marks)

   A: *If the rods could have been selected with replacement then the number of defective rods would have been Binomial with $n = 100$ and $p = 0.10$. Students should note the fact that we are probably sampling without replacement or at least comment on the issue somehow; deduct 1 mark if not. Then compute*

$$\mu = np = 100 \times 0.1 = 10$$

   and

$$\sigma = \sqrt{np(1-p)} = \sqrt{100 \times 0.1 \times 0.9} = 3.$$

   Now convert 13.5 to standard units:

$$\frac{13.5 - 10}{3} = 1.17$$

   and look up the area to the right of 1.17 by taking 1-0.8790=0.1210 or about 0.12.

   Using 14 not 13.5: deduct 1 mark.

   Using 14.5 not 13.5: deduct 1 mark.

Ask Richard if the wierd typo (caused by the % sign in LaTeX) caused visible difficulty for the student.

Related Problems*: Problem set 3, # 13, 15, 16.*

6. I toss a fair coin and throw a thumbtack. The coin is equally likely to land either heads (H) or tails (T). The thumbtack has chance 1/3 of landing point up (U); otherwise it lands tipped over (O). Make a list of all the outcomes in the sample space for this experiment and show the probability of each outcome. Explain the rules you used to get these probabilities. (3 marks)

   A*: The sample space has 4 outcomes:*

$$\{HU, HO, TU, TO\}$$

   The event that the coin lands heads has probability 1/2 while the event that the tack lands up has probability 1/3. These two events are independent to the probability that both happen is $1/2 \times 1/3 = 1/6$ *by the product rule. The other chances are computed similarly to get*

$$P(HU) = 1/6 \quad P(HO) = 2/6 \quad P(TU) = 1/6 \quad P(TO) = 2/6$$

   If the student used a tree diagram correctly: no problem – that is acceptable.

   Related Problems*: Problem set 3 # 6, 8, 9.*