STAT 350

Assignment 1

NOTE: Due by 3 PM Friday in my mailbox in the Statistics and Actuarial Science Department or by email to lhzhao@cs.sfu.ca by 8PM Friday. I have postponed questions 9 and 10 to assignment 2 due 21 May.

The first few problems are review. I want to see lots of words in your answers.

1. An environmental agency sets a standard of 200 ppb for the concentration of cadmium in a lake. The concentration of cadmium in one lake is measured 17 times. The measurements average 211 parts per billion with an SD of 15 parts per billion. Could the real concentration of cadmium be below the standard of 200 ppb?

2. Consider a population of 200 million people of whom 200 thousand have a certain condition. A test is available with the following properties. Assuming that a person has the condition the probability that the test detects the condition is 0.9. Assuming that a person does not have the condition the test detects (incorrectly) the condition with probability 0.001.

   A person is picked at random from the 200 million people and the test is administered.

   (a) What is the chance that the test detects the condition for this randomly selected person?

   (b) Assuming that the condition is detected by the test for this randomly selected person what is the chance that the person has the condition?

   (c) A mandatory testing program is contemplated. If all 200 million are tested about how many positive results should be expected? Of these about how many will not have the condition?

3. 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.

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

(c) Suppose we select 25 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 average percentage of daily dietary calories taken in the form of fat for the 25 selected fathers have comes out between 34 and 36%?

(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.

(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?

4. 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.

5. 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.

6. Suppose $Z_1, \ldots, Z_{10}$ are independent random variables each having a $N(0, 6)$ distribution. Let $\bar{Z} = \sum Z_i/10$, $U = \sum_{i=1}^{4} Z_i^2/6$, $V = \sum_{i=5}^{10} Z_i^2/6$, $X = Z_1/\sqrt{V}$ and $Y = (3U)/(2V)$. Give the names for the distributions of each of $\bar{Z}$, U, V, X and Y and use tables to find $P(|\bar{Z}| > 1)$, $P(U \leq 9.49)$, $P(-1.2 \leq X \leq 1.2)$, $P(Y > 6.23)$, $P(U \leq V)$, $P(Z_1 - 2Z_2 \geq 10)$.

7. A new process for measuring the concentration of a chemical in water is being investigated. A total of n samples are prepared in which the concentrations are the known numbers x_i for $i = 1, \ldots, n$; the new process is used to measure the concentrations for these samples. It is thought likely that the concentrations measured by the new process, which we denote Y_i, will be related to the true concentrations via

$$Y_i = \beta x_i + \epsilon_i$$

where the ϵ_i are independent, have mean 0 and all have the same variance σ^2 which is unknown.

(a) If this model is fitted by least squares, (that is by minimizing $\sum_i (Y_i - \beta x_i)^2$) show that the least squares estimate of β is

$$\hat{\beta} = \frac{\sum x_i Y_i}{\sum x_i^2}\,.$$

(b) Show that the estimator in part (a) is unbiased.

(c) Compute (give a formula for) the standard error of $\hat{\beta}$.

(d) The error sum of squares for this model is $\sum(Y_i - \hat{\beta}x_i)^2$ which may be shown to have $n-1$ degrees of freedom. If the x_i are the numbers 1, 2, 3 and 4, $\hat{\beta} = 1$ and the error sum of squares is 0.12 find a 95% confidence interval for β and explain what further assumptions you must make to do so.

(e) Show that the estimator

$$\tilde{\beta} = \frac{\sum Y_i}{\sum x_i}$$

is also unbiased.

(f) Compute (give a formula for) the standard error of $\tilde{\beta}$. Which is bigger, the standard error of $\hat{\beta}$ or that of $\tilde{\beta}$?

(g) Show that the mle of β in this model is $\hat{\beta}$, the least squares estimate, if the ϵ_i have normal distributions.

8. Consider the two-way layout without replicates. We have data Y_{ij} for $i = 1, \ldots I$ and $j = 1, \ldots, J$. We generally fit a so-called additive model

$$Y_{ij} = \mu + \rho_i + \gamma_j + \epsilon_{ij}$$

In the following questions consider the case $I = 2$ and $J = 3$.

(a) If we treat μ, ρ_1, ρ_2, γ_1, γ_2 and γ_3 as the entries in the parameter vector β what is the design matrix $X = X_a$ and what is the rank of X_a?

(b) What is the determinant of the matrix $X_a^T X_a$? Is this matrix invertible? How many solutions do the normal equations have?

(c) Usually we impose the restrictions $\rho_1 + \rho_2 = 0$ and $\gamma_1 + \gamma_2 + \gamma_3 = 0$. Use these restrictions to eliminate ρ_2 and γ_3 from the model equation and, for the parameter vector $\beta^T = (\mu, \rho_1, \gamma_1, \gamma_2)$ find the design matrix X_b.

(d) An alternate set of restrictions is called corner point coding where we assume $\rho_1 = \gamma_1 = 0$. With this restriction and the parameter vector $\beta^T = (\mu, \rho_2, \gamma_2, \gamma_3)$ what is the design matrix X_c?

(e) Show that the three design matrices have the same column space by finding a matrix A such that $X_a = X_b A$ and similarly for X_b and X_c and for X_a and X_c.

(f) Use the previous part to show that the vectors of fitted values $\hat{Y}$ will be the same for any solution of the normal equations for any of the three design matrices.

The next two questions require the use of computing software. Those of you who have not done any computing should come to the tutorial next week in the PC lab – room details to follow. You may use any statistical package you like. In class I will use SAS, JMP or R as I see fit.

9. From the text questions 1.19 and 1.23. The data are available on the disk accompanying the text; email me if you can't get it. In addition: write a short paragraph discussing the difficulties in using data on admitted students to study the relation between ACT score and first year GPA.

10. From the text questions 2.13 a and b and 2.23 a, b and c. In 2.23 c give a P-value and interpret this P-value.

DUE: Friday, 16 May.