STAT 200 Midterm Exam I (Section 101)

(Time: **50 minutes**, *Friday, October 13, 2006, 9:00 - 9:50am*)

Name:

Student ID:

This exam is to be done with closed notes/books. One "cheat sheet" (8.5′ × 11′, two sided) is allowed. There are 5 problems, with a total of 100 points. Make sure you write your name and student ID. We prefer short answers without omitting key steps, but you must show your work for problems 3 - 5.

Problem 1 (15 pts). Circle the correct answer: true (T) or false (F).

(1) If there are too many outliers (say 40%), then these "outliers" should not be considered as outliers. Instead, they should be considered as a distinct feature of the distribution for all the data. (T) F

(2) Observational studies usually do not permit us to attribute cause. If the number of subjects is large enough, however, we can be 95 per cent certain (or more) of causality. T (F)

(3) A group of items may be a population to one researcher while the same group could be a sample to another. (T) F

(4) If $P(A) + P(B) > 1$, then the events A and B are independent. T (F)

(5) The least-squares regression line is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible. (T) F

Problem 2 (40 pts). Circle the answer which is the *closest* to the correct answer (please choose only one answer).

1). A Vancouver television station is interested in predicting whether voters are in favor of or against an immediate federal election. It asks those viewers who are eligible to vote to phone in with their views (for or against). Of the 100 viewers who called, 60 are opposed to the immediate election. The population of interest is

(A) Vancouver residents eligible to vote in a federal election
B) Regular viewers of the television station
C) The 100 viewers who called in
D) The population of viewers opposed to the immediate election

2). To study the relationship between sugar consumption and weight gain, a research randomly assigned 14 volunteers into two groups. The first group of 6 participants was put on a diet low in sugar and the second group of the remaining 8 participants received 25% of their calories from fat. After 1 month, weight gain was recorded from each participant. Circle one answer in the following questions. What type of study is this?

A) An observational study.
B) An experiment, but not a double-blind experiment.
C) A double-blind experiment.
D) A matched-pairs experiment.

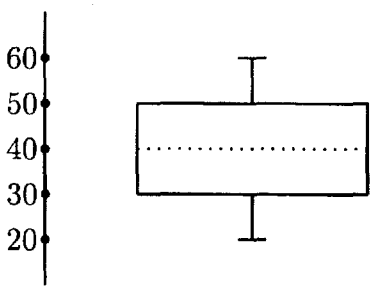
3). Referring to the previous problem (i.e., Problem 2)). Which of the following basic principles of statistical design was *not* used in the study?

A) control B) randomization C) repetition D) blinding

4). To determine if a new drug will reduce the symptoms of allergy, a researcher gave the new drug to 50 adult volunteers who suffer from allergies. Forty-five of these volunteers reported a significant reduction in their allergy symptoms. How could this study be improved?

A) Repeat the study with only the 45 volunteers who reported a significant reduction in their allergy symptoms, giving them a higher dosage this time.
B) Include people who do not suffer from allergies in the study in order to represent a more diverse population.
C) Give the spouses of the volunteers a placebo.
D) Use randomization to divide the volunteers into 2 groups – one to receive the new drug and one to receive the placebo.

5). The following boxplot is based on the salaries (in thousand dollars) of a sample of 160 programmers.



The number of programmers whose salaries are less than \$30,000 is approximately

A) 40 B) 50 C) 80 D) 100 E) 120

6). In a simple linear regression model, suppose that the independent variable x is the time spent on study and the dependent variable y is the exam score. The data show that no matter how much time is spent on study, the exam score is always 80. Which of the following statements is correct?

A) the intercept is zero.
B) it is impossible to fit a regression line.
C) both intercept and slope are zero.
D) the intercept is zero and the slope is 80.
E) the intercept is 80 and slope is zero.

7). A study found a correlation of $r = -0.70$ between the gender of a worker and his/her income. You may correctly conclude

A). women earn more than men on the average.

B). women earn less than men on the average.

C) this is incorrect because r makes no sense here.

D). an arithmetic mistake was made. Correlation must be positive.

8). There are three children ages 3, 4, and 5, in a room. If a 4-year-old child enters the room, the

A). mean age will stay the same but the variance will increase.

B) mean age will stay the same but the variance will decrease.

C). mean age and variance will stay the same.

D). mean age and variance will increase.

E). none of the above.

9). Here is a stem plot of scores in an upper year biology course.

```
11 | 6 8
12 | 1 4 8
13 | 3 7
14 | 2 6
15 |
16 |
17 | 9
```

If we discard the outlier,
A) both the mean and the standard deviation will be larger.
B) both the mean and the standard deviation will be smaller.
C) the mean will be larger but the standard deviation will be smaller.
D) the mean will be smaller but the standard deviation will be larger.

10). Which of the following sets has the largest possible standard deviation (no calculation is needed) A) 8, 8, 9, 9 B) 2, 2, 2, 2 C) 0, 0, 8, 8 D) 0, 1, 2, 3.

This table is for instructor use only

Problem 1	Problem 2	Problem 3	Problem 4	Problem 5	Total Mark

Problem 3 (10 pts). A sample of 50 students at UBC is randomly selected from those who have taken STAT200 and STAT305, and their scores on STAT200 and STAT305 are recorded. The summary statistics for the scores in the sample are:

```
STAT200:    mean = 68,    standard deviation = 15
STAT305:    mean = 71,    standard deviation = 16
```

The sample correlation between STAT200 and STAT305 scores equals 0.76.

(a) (8 pts) We want to predict STAT305 score from a student's STAT200 score. Find the equation of the least-squares regression line which can be used for this prediction.

Let $y = STAT305$, $x = STAT200$

(3 pts) $b = r \cdot \frac{S_y}{S_x} = 0.76 \times \frac{16}{15} = 0.81$

(3 pts) $a = \bar{y} - b\bar{x} = 71 - 0.81 \times 68 = 15.87$

2 pts Answer: $\hat{y} = 15.87 + 0.81x$

(2 pts) (b) The fraction of the variation in the STAT305 scores that can be explained by the linear relationship between STAT200 and STAT305 is $0.76^2 = 0.5776 = 57.76\%$

Problem 4 (18 pts). The weights of chocolate bars produced by a certain machine have a normal distribution with a mean of 10 oz. and a standard deviation of 0.4 oz.

(a) (6 pts) What proportion of chocolate bars weighs between 9 oz. and 11 oz.

(2 pts) $P(9 < X < 11) = P\left(\frac{9-10}{0.4} < Z < \frac{11-10}{0.4}\right)$

1 pts $= P(-2.5 < Z < 2.5)$

2 pts $= 0.9938 - 0.0062$

1 pt $= 0.9876$

Answer: 0.9876

(b) (6 pts) What is the weight that should be stamped on the chocolate bars so that only 5 % of chocolate bars are under-weight.

(2 pts) $z^* \approx -1.645$

(2 pts) $\frac{x-10}{0.4} = -1.645$

(2 pts) $x = 10 - 1.645 \times 0.4$

$= 9.342$

Answer: 9.342

(c) (6 pts) What is the third quartile for the weight distribution of these chocolate bars?

(2 pts) $z^* \approx 0.675$

(2 pts) $\frac{x-10}{0.4} = 0.675$

(2 pts) $x = 10 + 0.675 \times 0.4 = 10.27$

Answer: 10.27

Problem 5 (17 pts). A quiz consists of 2 independent multiple-choice questions. For each question, there are four choices, with only one being correct. There are 10 points for each question. Suppose student A and student B randomly guesses the correct answer to each question independently.

(a) (5 pts) What is the probability that at least one student get 10 points or more.

Let E = "at least one student get 10 points or more"

(2 pts) $P(A) = 1 - P(\bar{A}) = 1 - P(\text{both students get 0 points})$

(2 pts) $= 1 - \left(\frac{3}{4}\right)^2 \times \left(\frac{3}{4}\right)^2$

(1 pt) $= 0.68$

Answer: 0.68

(b) (6 pts) What is the expected score student A will get?

Let X = "score student A get"

(3 pts)

X	0	10	20
P	0.75^2	$2 \times 0.75 \times 0.25$	0.25^2

$\Rightarrow$

X	0	10	20
P	0.5625	0.375	0.0625

(3 pts) $E(X) = 0 \times 0.5625 + 10 \times 0.375 + 20 \times 0.0625$

$= 5$

Answer: 5

(c) (6 pts) If the instructor wants the probability of the event "neither student A nor student B will get 10 points or more by guessing" to be greater than 30%, how many multiple choice questions should the instructor put in the quiz?

Let D = "neither student A nor ...."

n = number of multiple choice questions

(2 pts) $P(D) = P(\text{"student A gets 0 pts" and "student B gets 0 pts"})$

$= 0.75^n \times 0.75^n > 0.30$

(2 pts) $0.75^{2n} > 0.30 \Rightarrow 2n \log(0.75) > \log(0.30)$

(2 pts) $\Rightarrow n < \dfrac{\log(0.30)}{2\log(0.75)} = 2.09$

Answer: $n \leq 2$