

Name: \_\_\_\_\_

## Statistics 528 – Practice Midterm Solutions Winter 2004

### PART 1: MULTIPLE CHOICE

The following 5 problems are worth 4 points each. Circle the response that best answers each of the questions. There is **no** partial credit for these problems so there is no need to show your work.

1. The least-squares regression line is
  - a) the line that best splits the data in half such that half of the data points fall above the line and half fall below the line.
  - b) the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible.
  - c) the line that makes the square of the correlation in the data as large as possible.
  - d) all of the above.

**Answer: b**

2. For a density curve, which of the following statements is correct?
  - a) The median is 0.
  - b) It is symmetric about its median.
  - c) The total area under the curve is 1.
  - d) None of the above

**Answer: c**

3. Which of the following statements about the normal distribution is correct?
  - a) The mean is always equal to 0.
  - b) The density curve is skewed to the right if there are outliers.
  - c) 95 percent of the area under the curve is within two standard deviations of 0.
  - d) It is completely described by its median and standard deviation.

**Answer: d**

4. Researchers studying acid rain measured the acidity of precipitation in a Colorado wilderness area for 150 consecutive weeks. Acidity is measured in pH. Lower pH values show higher acidity. The acid rain researchers observed a linear pattern over time. They reported the least-squares regression

$$\text{pH} = 5.43 - (0.0053 \times \text{Weeks})$$

This implies that the correlation between pH and time is

- a) -0.0053
- b) 1/0.0053
- c) It is impossible to say anything about the correlation from the information given.
- d) negative.

**Answer: d**

5. Which sampling scheme is **not** a form of probability sampling?

- a) a simple random sample
- b) a stratified random sample
- c) drawing people at random from the telephone book
- d) a voluntary response survey

**Answer: d**

## PART 2: SHORT ANSWER

Show your work on all problems in this section to receive full credit.

### Problem 1 (16 points)

A sample was taken of the salaries of 20 employees who work at a large company. The following are the salaries (in thousands of dollars) for this year. For convenience, the data are ordered.

28    31    34    35    37    41    42    42    42    47  
49    51    52    52    60    61    67    72    75    77

Some numerical summaries of the data (computed using MINITAB) are given below.

### Descriptive Statistics: salary

Variable	N	Mean	StDev	Minimum	Maximum	Q1	Q3
salary	20	49.75	14.75	28.00	77.00	38.00	60.75

a) Find the median salary of the 20 employees.

**(47+49)/2=48. The median salary is 48,000 dollars.**

b) Find the observation that has the smallest squared deviation from the mean.

**Since the sample mean is 49.75, the value 49 has the smallest squared deviation.**

c) If the maximum salary, 77 (thousand dollars) was mistakenly recorded as 97, which of the following statistics would be different?

Circle all the statistics listed below that would change.

Mean / Median / Standard deviation / IQR / Q1 / Q3

d) Suppose each employee in the company receives a \$3,000 raise for next year (each employee's salary is increased by \$3,000).

Circle the statistics that will be changed because of this salary increase.

Mean / Median / Standard deviation / Variance / IQR / Min / Max

**Problem 2** (18 points)

The weights of boxes of cookies produced by a certain manufacturer approximately follow a normal distribution with a mean of 202 g and a standard deviation of 3 g.

a) Between what values do the middle 95% of the weights of boxes of cookies lie?

**Between  $202-2*3=196$  g and  $202+2*3=208$  g by the 68-95-99.7 rule.**

b) If the manufacturer stamps 205 g on all the boxes, what percent of boxes of cookies are overweight?

**Let's call the weight of a randomly chosen box of cookies X.**

**$P(X>205)=P(Z>(205-202)/3)=P(Z>1)=1-P(Z<1)=1-.8413=.1587$ . So, about 16% of the boxes are overweight.**

c) What weight should be stamped on the boxes so that only 1% of the boxes are underweight?

**Let x be the weight to be stamped on the box. In order to have only 1% of underweight boxes, the z-score of x should be equal to -2.33, which gives 1% of values to the left in the standard normal distribution. (From the standard normal table,  $P(Z<-2.33)=.0099$ .)**

**So,  $(x-202)/3=-2.33$ . Solving for x, we have  $-2.33*3 +202=195.01$  g.**

**Problem 3** (20 points)

A researcher is proposing a new diet that limits the number of grams of carbohydrates people eat in the morning. He plans to test his diet by asking a group of subjects to reduce the number of carbohydrates they consume for breakfast and observe whether or not they lose weight over time. The researcher is worried, however, that people using his diet might not actually know or be able to find out how many grams of carbohydrates they are consuming. He hypothesizes that perhaps there is a linear relationship between the number of calories and the number of grams of carbohydrates in breakfast foods. If this were the case, he could place a restriction on the number of calories his subjects could eat and then use a linear model to predict the number of carbohydrates his subjects consume. The researcher collects data on the calories and number of grams of carbohydrates in one serving of each of 12 breakfast cereals. Summary statistics of these data are given below.

	<b>Calories</b>	<b>Grams of Carbohydrates</b>
<b>Mean</b>	105	23
<b>Standard Deviation</b>	17.84	3.95
<b>Correlation (r)</b>	0.838	

a) Find the least-squares regression line for predicting the number of grams of carbohydrates from the number of calories in breakfast cereals.

**Grams of carbo. = 3.47 + 0.186calories**

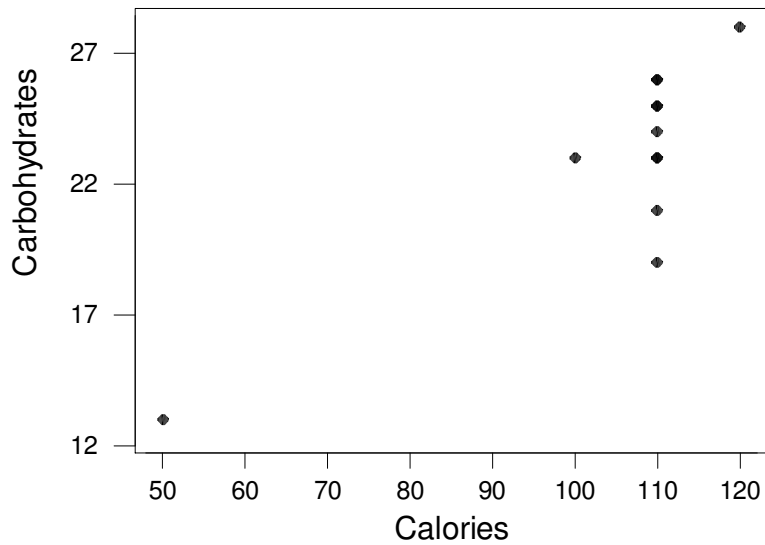
b) Use your regression line to predict the number of grams of carbohydrates in a cereal that has 75 calories per serving.

**17.42 grams of carbohydrates**

c) Do you think your regression line can be used to predict the number of grams of carbohydrates in typical lunches. Why?

**No. The relationship should not be generalized.**

d) A scatterplot of the data is shown below. Identify a possibly influential observation and explain what it means for an observation to be influential. Describe how you might determine whether the point you identified is actually an influential observation.



**The observation in the bottom left corner of the figure (carbs.  $\approx$  13, calories  $\approx$  50) is potentially an influential observation. Influential points are observations that are outliers with respect to the value of the x variable and have strong influence on the position of the regression line. One can determine whether the point is influential by refitting the regression line without the observation in question and seeing how much the regression line changes.**

**Problem 4** (18 points)

Can herbal supplements reduce the length and severity of the common cold? An experiment administered a cold nasally to 300 paid volunteers. Volunteers received either a 100 milligram dose of an herbal supplement, a 50 milligram dose of the supplement, or a placebo to take two times a day over the next week. Volunteers were evaluated daily by weighing the tissues they used, being asked how they felt, and having their sinuses rinsed so that levels of the cold germ could be measured. The researchers found that differences in length and severity of the cold in the different groups were not statistically significant.

a) How many factors are there in this experiment? What are the levels?

**1 factor**

**3 levels – 100 milligrams, 50 milligrams, placebo (0 milligrams)**

b) Explain what the placebo effect is, which subjects experience the placebo effect, and why a placebo was used in this experiment.

**The placebo effect is a positive response to any treatment. It is experienced by the patients in the treatment groups and the placebo (control) group. A placebo is needed to control for the placebo effect.**

c) What does it mean to say that the observed differences are not statistically significant?

**It means that the observed differences are not meaningful – they are due to chance variation.**

**Problem 5** (8 points)

Does regular exercise reduce the risk of a heart attack? Here are two ways to study this question. Which design will produce more trustworthy data? **Explain your answer.**

- a) A researcher finds 2000 men over 40 who exercise regularly and have not had heart attacks. She matches each man with a similar man who does not exercise regularly, and she follows both groups for 5 years.
- b) A researcher finds 4000 men over 40 who have not had heart attacks. She randomly 2000 of the men to a regular program of supervised exercise. The other 2000 men continue their usual habits. The researcher follows both groups for 5 years.

**Design B will produce more trustworthy data. In order to control for possible confounding factors such as diet, weight, and other indicators of a healthy lifestyle, the subjects should be randomly assigned to the different treatment groups.**