

Name: _____

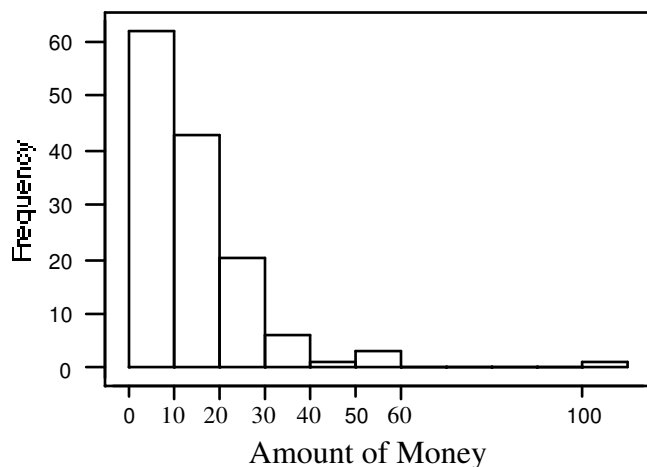
Statistics 528 – Midterm Solutions Winter 2004

INSTRUCTIONS

Show your work on all problems to receive full credit.

Problem 1 (15 points)

In a statistics class with 136 students, the professor recorded how much money each student had in his or her wallet during the first class of the quarter. The histogram below is of the data collected.



Indicate whether each of the statements is true or false. If false, **explain briefly why it is false**.

a) The percentage of students with under \$10.00 in their possession is about 50%. **True or False**

True

b) The distribution is skewed to the right. **True or False**

True

c) A normal density curve is a good approximation to the histogram. **True or False**

False, because the distribution is asymmetric (skewed to the right with outliers).

d) The median would be a better measure than the mean in describing the center of the distribution. **True or False**

True

e) The median amount of money students have is over \$20.00. **True or False**

False, because the number of students with under \$20.00 in their possession is over 100 (out of 136), thus the median can not be more than \$20.00.

Problem 2 (20 points)

The following are types of plots that we have discussed in class. For **TWO** of the types of plots, describe i) how to make the plot and ii) the purpose of the plot (what do you hope to learn about the distribution of a variable or the relationship between two variables) using the plot.

a) **BOXPLOT**

A boxplot is constructed using the five-number-summary of the distribution of a variable (min., 1st quartile, median, 3rd quartile, max.). A central box is drawn extending from the 3rd quartile of the distribution to the 1st quartile of the distribution with a horizontal line at the median. Lines extending from the center box are drawn up to the maximum value and down to the minimum value of the variable. Boxplots are used to graphically summarize the distribution of a variable and are good for summarizing skewed distributions. Side-by-side boxplots are also convenient for comparing the distributions of multiple variables that have similar ranges.

b) **NORMAL QUANTILE PLOT**

Normal quantile plots are constructed by plotting the data against the corresponding quantiles of the standard normal distribution. If the distribution of the data is normal, we would expect these points to fall on a straight line. Normal quantile plots are a good way to check if a normality assumption holds even in the tails of the distribution.

c) **RESIDUAL PLOT**

A residual plot is constructed by plotting the residual for each data point (predicted value – observed value) against the value of the explanatory variable. Residual plots are used to diagnose the fit of a linear regression model. If the assumptions of the regression model hold, we would expect that the residuals would be spread out on either side of the horizontal line with height zero and would have no systematic pattern.

Problem 3 (25 points)

A researcher wishes to study how the weight of children changes during the first year of life. He takes a sample of children under one year and records their weight Y (in kilograms) and their age X (in months). He computes the following quantities:

$$r = \text{correlation between } X \text{ and } Y = 0.9$$

$$\bar{x} = \text{mean of the values } X = 6.5$$

$$\bar{y} = \text{mean of the values of } Y = 6.6$$

$$s_x = \text{standard deviation of the values of } X = 3.6$$

$$s_y = \text{standard deviation of the values of } Y = 1.2$$

Assume that the distribution of the weight variable (Y) can be approximated by the normal distribution.

- a) What percentage of the children weigh between 4.2 and 9 kilograms?

4.2 is 2 standard deviations below the mean of Y and 9 is two standard deviations above the mean of Y so by the 68-95-99.9 percent rule, 95 percent of children fall in this range.

- b) How much do the heaviest 10 percent of the children weigh?

Find the point y under the $N(6.6, 1.2)$ curve such that 10 percent of the area under the curve is to the right of this point. The corresponding point under the $N(0, 1)$ curve is 1.285 so $1.285 = (y - 6.6)/1.2$. This implies $y = 8.142$ or that the heaviest 10 percent of one-year-olds weigh more than 8.142 kilograms.

The researcher decides to fit a least-squares regression line to the data with X as the explanatory variable and Y as the response variable. He computes the following quantities:

- c) What is the equation of the least-squares regression line of Y on X .

$$b = r \cdot (s_y/s_x) = 0.9 \cdot (1.2/3.6) = 0.3, a = \bar{y} - b \bar{x} = 6.6 - 0.3 \cdot 6.5 = 4.65$$

The least-squares regression line of Y on X is $Y = 0.3 X + 4.65$

d) What is the r^2 value for the regression?

$$r^2 = 0.9^2 = 0.81$$

e) After fitting the least-squares regression line, the researcher constructs a residual plot of the residual values versus the X values. Only one point is above the horizontal line at 0. What does this imply about the linear regression model and what should the researcher do?

This implies that the least-squares regression model is not appropriate; we would expect the residuals to be equally spread out on either side of the horizontal line at 0. Most likely there is one data point that is an outlier – it does not fit the linear pattern of the rest of the data. The researcher should check if there is a problem with this one point (misrecorded?) and possibly fit the regression model again without this point.

Problem 4 (15 points)

Studies of disease often ask people about their diet in years past in order to discover links between diet and disease. How well do people remember their past diet? Can we predict actual past diet as well or better from what subjects eat now as from their memory of past habits? Data on actual past diet are available for 91 people who were asked about their diet when they were 18 years old and again when they are 30. Researchers asked them at about age 55 to describe their eating habits at ages 18 and 30 and also their current diet. The study report says:

The first study aim, to determine how accurately this group of participants remembered past consumption, was addressed by correlations between recalled and historical consumption in each time period. To evaluate the second study aim, that is, whether recalled intake or current intake more accurately predicts historical intake of food groups at age 30 years, we performed regression analysis.

a) Explain in nontechnical language what “correlation” means, why correlation suits the first aim of the study, what “regression” means, and why regression fits the second aim of the study. Be sure to point out the distinction between correlation and regression.

Correlation measures the degree of linear association between two variables. Since the first aim of the study is to measure the strength of the relationship between recalled and historical consumption, correlation is appropriate assuming the relationship is linear. Regression is a model that can be used to predict one variable from another so it is more appropriate for the second aim of the study.

b) The authors used regression to predict intake of a number of foods at age 30 from current intake of those foods and from what the subjects now remember about this intake at age 30. They conclude that “recalled intake more accurately predicted historical intake at age 30 years than did current diet.” As evidence, they present r^2 values for the regressions. Explain why comparing r^2 values is one way to compare how well different explanatory variables predict a response.

r^2 measures the amount of variability in the response variable that is explained by the regression on the explanatory variable. As a result, the higher the r^2 value, the better the regression model is for prediction. Therefore, r^2 values is one way to compare how well different explanatory variables predict a response.

Problem 5 (25 points)

People who eat lots of fruits and vegetables have lower rates of colon cancer than those who eat little of these foods. Fruits and vegetables are rich in “antioxidants” such as vitamins A, C, and E. Will taking antioxidants help prevent colon cancer? A clinical trial studied this question with 864 people who were at risk for colon cancer. The subjects were equally divided into four groups: daily beta carotene, daily vitamin C and E, all three vitamins every day, and daily placebo. After four years, the researchers were surprised to find no significant difference in colon cancer among the groups.

a) What are the explanatory and response variables in this experiment?

Explanatory variable – antioxidants
Response variable – colon cancer

b) Outline the design of the experiment.

- i) randomly assign the 865 people to four groups of 216 people
- ii) give each of the subjects in the four groups a treatment – either daily beta carotene, daily vitamin C and E, all three vitamins, or a daily placebo – for 4 years
- iii) compare rates of colon cancer among the different groups

c) Assign labels to the 864 subjects and use Table B line 151 (shown below) to choose the first 5 subjects for the beta carotene group.

Table B (line 151): 03802 29341 29264 80198 12371 13121 54969 43912

38, 22, 129, 264, 801

d) What does “no significant difference” mean in describing the outcome of the study?

“No significant difference” means that there was no difference beyond what would be expected due to chance variation.

e) Suggest some lurking variables that could explain why people who eat lots of fruits and vegetables have lower rates of colon cancer. The experiment suggests that these variables, rather than the antioxidants, may be responsible for the observed benefits of fruits and vegetables.

Variables related to healthier lifestyle.