# Stats 10 Lec 2: Midterm 1

**Version 1**

Introduction to Statistical Reasoning
UCLA, Fall 2016

**Instructions:** You have 40 minutes to complete the following questions. This exam is closed book. You may use only one sheet of paper with handwritten notes. Calculators are allowed, but no other electronic devices are allowed. There are 21 (19 multiple choice and 2 short answer) questions worth a total of 50 points. Good luck!

**Academic Misconduct:** Any potential violation of UCLA's policy on academic integrity will be reported to the Office of the Dean of Students. All work on this exam must be your own.

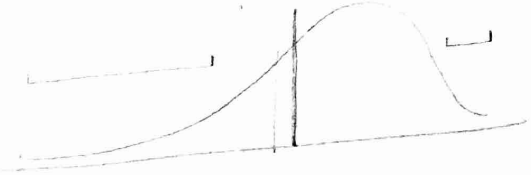
**In the Special Codes section of your scantron, fill in a 1 in the K column.**

**(Turn over when exam starts)**

## Multiple Choice Questions

**Mark your answers to all multiple choice questions on the scantron provided. Any answers marked on these pages will not be scored. Each question is worth 2 points.**

**Problem 1** Say we have the five-number summary of a data set. We see that a larger gap exists between the lowest value and Q1 than between Q3 and the highest value. What does this say about the shape of the data set?



(a) The distribution of the data is symmetric.

(b) The distribution of the data is right-skewed.

(c) The distribution of the data is left-skewed.

(d) The shape of the data set cannot be determined from the information given.

**Problem 2** The regression line to predict variable $y$ from variable $x$ is computed to be

$$\text{Predicted } y = a + bx,$$

for some values $a$ and $b$. The regression line to predict $x$ from $y$ is computed to be

$$\text{Predicted } x = c + dy,$$

for some values $c$ and $d$. Is it possible for the slope $b$ to be the same as the slope $d$?

(a) Yes, always

(b) Yes, sometimes

(c) No, never

(d) Cannot be determined from the information given

**Problem 3** Many medical researchers have suggested that the incidence of asthma is higher in city-dwelling children than among those living in the suburbs. Some reasons provided are that city-dwellers are exposed to more pollution, more indoor smoke, and other possible contributors to asthma. However, researchers at Johns Hopkins published a study that found that there was no difference in asthma rates between city children and suburban children. Which of the following most likely describes this study?

(a) This is an observational study, because researchers can't control the factors that affect asthma.

(b) This is an observational study, because researchers can't control whether the children live in cities or suburbs.

(c) This is a controlled experiment, because the researchers selected the children they wished to study.

(d) This is a controlled experiment, because the researchers selected the factors that might lead to asthma.

**Problem 4** If the correlation between two quantitative variables is exactly $-1$, which of the following are true?

I. The scatterplot could be a random cloud of points.

II. The scatterplot could be nonlinear.

III. All the points would lie along a perfect straight line, with no deviation at all.

(a) I only

(b) II only

(c) III only

(d) I and II only

(e) I, II, and III

**The following information is used in Problems 4 and 5.**

A class has 10 students. The scores on the latest test, in order, were:

88 79 81 81 101 82 83 84 90 79

**Problem 5** What is the interquartile range of the test scores?

(a) 22

(b) 11

(c) 7

(d) 3

79 79 81 81 82 83 84 88 90 101

82.5

**Problem 6** What is the median test score?

(a) 83.5

(b) 83

(c) 82.5

(d) 82

(e) 81

**Problem 7** Greg Pikitis played on his high school's soccer team and basketball team last year. He scored 18 goals throughout the soccer season and made 28 baskets throughout the basketball season. The following summarize the total number of goals and baskets by each of his team members throughout the season.

| | Mean | SD |
|---|---|---|
| Soccer Goals | 10 | 4 |
| Basketball Baskets | 25 | 5 |

Based on these statistics alone, is Greg a better soccer player or basketball player?

(a) Greg is better at basketball.

(b) Greg is better at soccer.

(c) Greg is equally good at soccer and basketball.

(d) Cannot be determined from the information given.

**The following information is used in Problems 8 and 9.**

The manager of an ice cream shop recorded information about his customers over a one week period. For each customer, the manager recorded the ice cream flavor ordered (flavor preference) and whether the person wanted it in a cone of cup (container preference).

| | Chocolate | Vanilla | Strawberry |
|---|---|---|---|
| Cone | 120 | 160 | 20 |
| Cup | 45 | 40 | 15 |

**Problem 8** What proportion of customers ordered a cup?

(a) 0.73

(b) 0.40

(c) 0.75

(d) 0.25

**Problem 9** What proportion of customers who ordered a cone wanted chocolate ice cream?

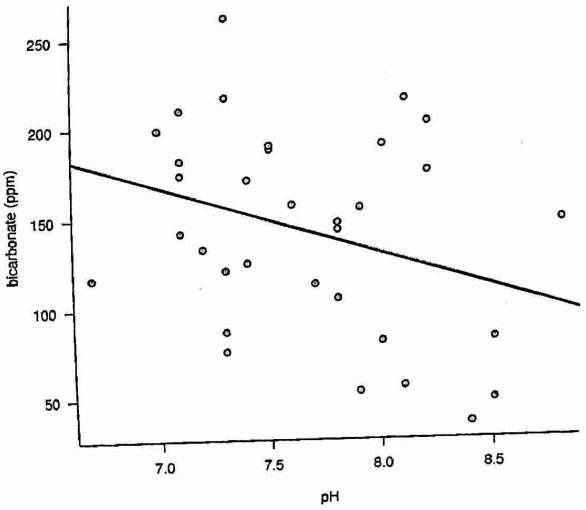
(a) 0.73

(b) 0.40

(c) 0.75

(d) 0.25

**Problem 10** The pH is used to measure the acidicity or basicity of aqueous solutions. Neutral solutions (like water) have a pH of 7. Acidic solutions (like lemon juice) have a pH less than 7, basic solutions (like bleach) have a pH greater than 7. For a technical report from the Union Carbide Corporation, data on the basicity of ground water was collected on a random sample of wells in Northwest Texas. Below is a scatterplot that illustrates the observed relationship between pH and bicarbonate in the well water.




The report uses the pH to predict the level of bicarbonate (measured in parts per million) in the well water. The regression line has the equation

$$\text{Predicted bicarbonate} = 432.15 - 37.77\ \text{pH}$$

with an $r^2$ of about 11.5%. Choose the best interpretation of the value of $r^2$.

(a) Predictions of bicarbonate based on the regression line will be correct in about 11.5% of the cases.

(b) About 11.5% of of the data lie within one standard deviation of the regression line.

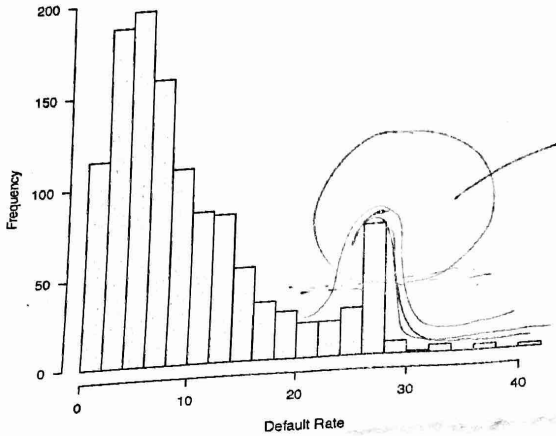
(c) About 11.5% of the observations fall on the regression line.

(d) About 11.5% of the total variation in bicarbonate is explained by the regression line.

(e) About 11.5% of the total variation in pH is explained by the regression line.

**The following information is used in Problems 11, 12, and 13.**

The Wall Street Journal collected data on a large number of colleges and universities in the United States in order to examine factors that affect defaults on student loans. The histogram below shows the student loan default rates (in percentage points) for primarily bachelor's degree granting colleges and universities in the United States. The bin width is 2 percentage points.




why is this bimodal?

**Problem 11** What is the shape of the histogram?

(a) Symmetric and unimodal

(b) Right-skewed and unimodal

(c) Right-skewed and bimodal

(d) Left-skewed and unimodal

(e) Left-skewed and bimodal

outliers??

**Problem 12** Which summary statistics would you use to measure the center and spread for the histogram above?

(a) Median and standard deviation

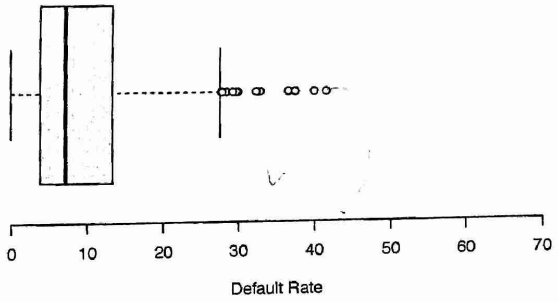
(b) Median and interquartile range

(c) Mean and standard deviation

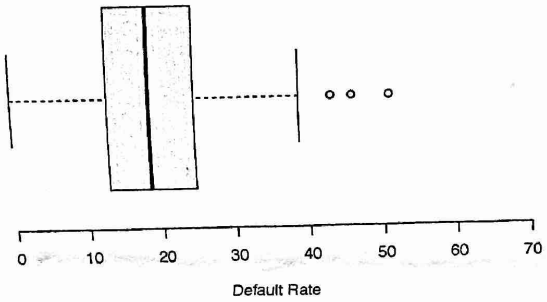
(d) Mean and interquartile range

**Problem 13** Which of the following boxplots corresponds to this distribution?

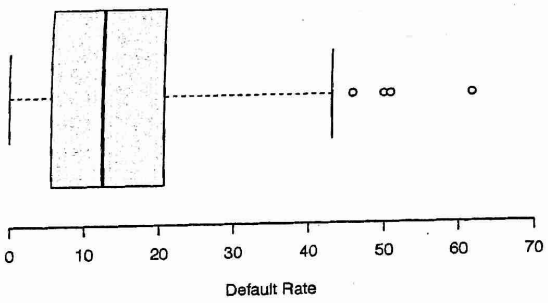
(a)



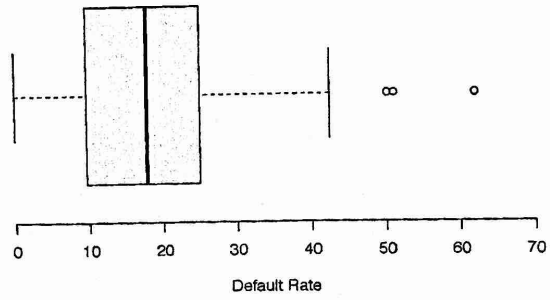

(b)




(c)




(d)

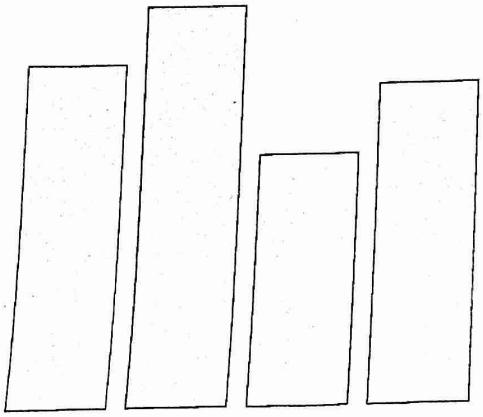


**The following plots are used in Problems 14 and 15.**

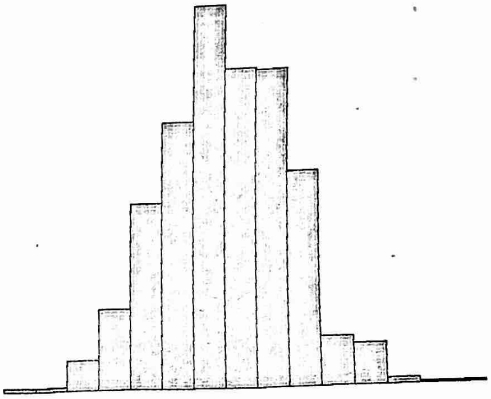
**Problem 14** Which of the plots below could describe the distribution of class levels of undergraduate students at UCLA taking an introductory statistics class?

**Problem 15** Which of the plots below could describe the distribution of heights of self identified female students in a large UCLA statistics class?

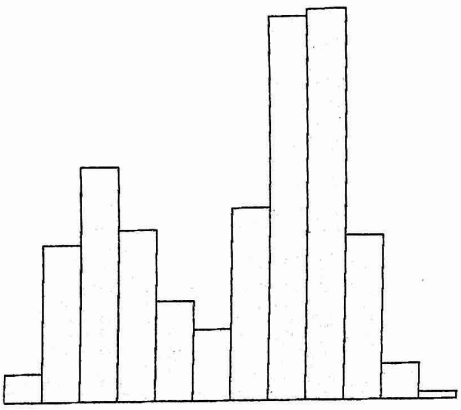
(a)



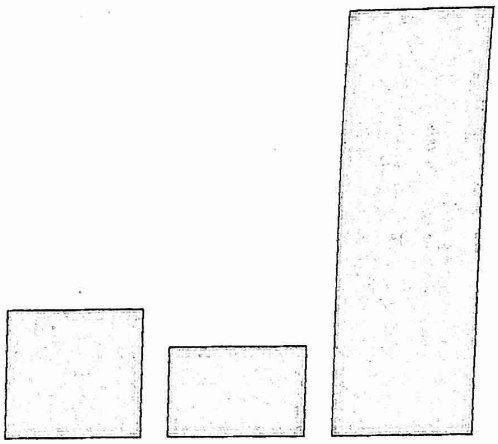
(b)



(c)

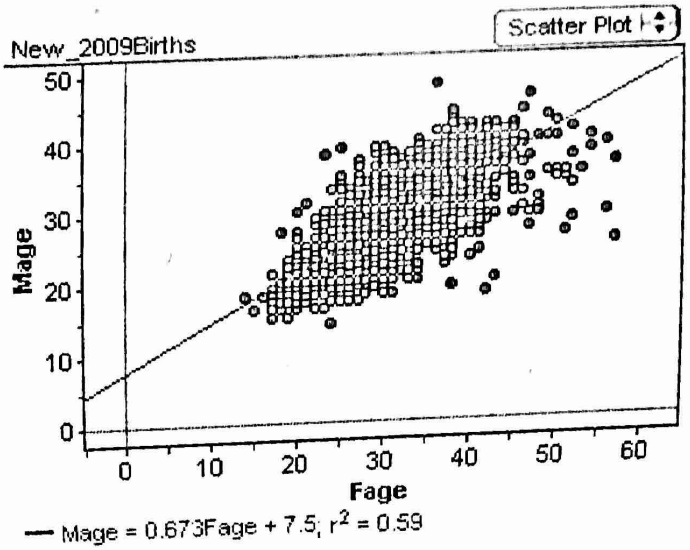


(d)



**The following information is used in Problems 16, 17, and 18.**

The data *2009Births.ftm* that you analyzed in lab contains information about a random sample of births in North Carolina. The scatterplot below shows the association between the mother's age (Mage) and the father's age (Fage).




**Problem 16** The correlation coefficient is approximately

(a) 0.59

(b) 0.77

(c) 7.5

(d) 0.673

**Problem 17** What is the value of the intercept?

(a) 0.59

(b) 0.77

(c) 7.5

(d) 0.673

**Problem 18** What is the approximate predicted age of the mother when the father is 32 years old?

(a) 32

(b) 29

(c) 35

(d) 25

**Problem 19** A researcher reports that there is a "strong, negative, linear association" between the amount of money spent on neighborhood police patrols and the number of crimes committed. This implies that

(a) If we were to fit a regression line between the number of crimes in a neighborhood and the amount of money spent on police patrols in that neighborhood, the slope would be close to $-1$.

(b) If we were to compute the correlation coefficient between the number of crimes in a neighborhood and the amount of money spent on police patrols in that neighborhood, it would be close to $-1$.

(c) If we were to fit a regression line between the number of crimes in a neighborhood and the amount of money spent on police patrols in that neighborhood, the slope would be close to $+1$.

(d) If we were to compute the correlation coefficient between the number of crimes in a neighborhood and the amount of money spent on police patrols in that neighborhood, it would be close to $+1$.

## Short Answer Questions

**Write your answers to all short answer questions in the space provided on these pages. Show all your work. Each question is worth 6 points (2 points per part).**

**Problem 20** Census at School is an international program that asks students to contribute data by answering some survey questions. From this program, we gather the height and arm span measurements of a random sample of 500 students in the United States. The scatterplot shows a positive linear relationship between height and arm span. The corresponding coefficient of determination is 0.64. The mean arm span is 70 inches with a standard deviation of 13.5 inches. The mean height is 66 inches with a standard deviation of 12 inches.

2 (a) Calculate and interpret the slope of the linear model for predicting arm span from height.

$$\text{slope} = r\frac{s_y}{s_x} = 0.64 \cdot \frac{13.5}{12} = 0.72$$

the slope is 0.72, meaning for every approximately inch increase in height, there's a 0.72 inch increase in arm span

2 (b) Calculate and interpret the intercept of the linear model for predicting arm span from height.

$a = 0.72$

Sub in 66 for height, 70 for arm span

$y = ax + b$

$70 = 0.72 \cdot 66 + b$

$22.48 = b$

The linear model's intercept is 22.48. When combined with the slope, this means that the arm span will always be greater than height. However, it is illogical to say that the armspan for a person with height zero is 22.48 inches because a person must have positive height

2 (c) What is the equation for the regression line for predicting arm span from height?

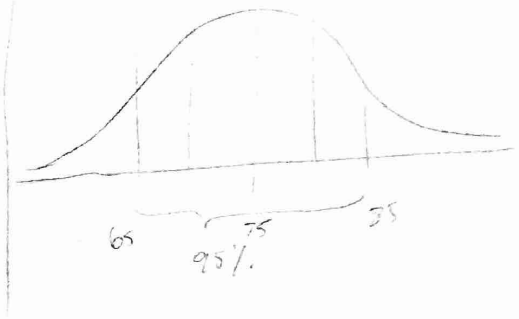
$y$ = armspan $x$ = height

$$y = 0.72x + 22.48$$

5

2 **Problem 21** The temperature in August in San Diego is unimodal and symmetric with a mean of 75 degrees (Fahrenheit) and standard deviation of 5 degrees.

(a) What percent of days are warmer than 65 degrees?




$\frac{75-65}{5} = 2$

with a z-score of 2, we can say that 95% of the days are between 65°F & 85°F. This means that $95\% + \frac{5\%}{2} = 97.5\%$ of days are warmer than 65°F.

(b) Between what two temperatures do 68% of the days in San Diego fall between?

2 Empirical Rule: 1 stdev = 5° = 68%.

$75 \pm 5 = (70, 80)$

∴ 68% of the days fell between 70°F & 80°F.

1 (c) We recorded a temperature of 57.5 degrees on August 2nd. Is this unusual? Justify your answer.

$\frac{75 - 57.5}{5} = 3.5$

with a z-score of 3.5 ✗ it seems that the temp was more than 3 stdevs away from the mean. This is highly unusual because it is apart from more than 99.7% of all the observations, meaning that the probability of it occuring was less than 0.3%.

10