17
3
13
13
50

STAT:2010/4200, Statistical Methods and Computing
Spring 2020, Instructor: Cowles
Midterm 1

Show your work on any problems that involve calculations.

Name: Solutions

1. What is the data type of each of the following variables? Circle one choice for each.

   (a) types of kitchen appliances manufactured by a company (microwave, pressure cooker, blender, etc.)

   i. Binary
   ii. Nominal
   iii. Ordinal
   iv. Discrete quantitative
   v. Continuous quantitative

2 ii circled

   (b) the ratings awarded to high school bands in a state competition (Superior, Excellent, Very Good, Fair)

   i. Binary
   ii. Nominal
   iii. Ordinal
   iv. Discrete quantitative
   v. Continuous quantitative

2 iii circled

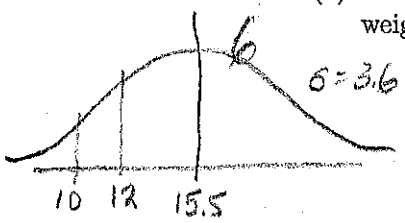
   (c) the number of cats housed by the Iowa City animal shelter each year

   i. Binary
   ii. Nominal
   iii. Ordinal
   iv. Discrete quantitative
   v. Continuous quantitative

2 iv circled

2. The brain weight of sperm whales follows a normal distribution with mean 15.5 pounds and standard deviation 3.6 pounds.

   (a) What is the probability that a randomly selected sperm whale has a brain that weighs between 10 and 12 pounds? (Numeric answer; show your work.)



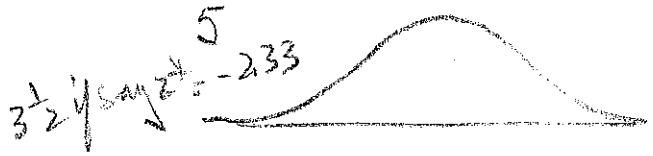
$z_{10} = \frac{10 - 15.5}{3.6} = -1.53$ $\quad z_{12} = \frac{12 - 15.5}{3.6} = -0.97$

$Pr(Z < -1.53) = .0630$ $\quad Pr(Z < -0.97) = .1660$

$.1660 - .0630 = \boxed{.1030}$

   (b) Only 10% of sperm whales have brains that weigh less than 10.89 pounds. (Fill in the blank with a numeric answer; show your work.)

$15.5 - 1.28(3.6) = 10.89$



17

2½ pt 7

3. In lectures at the beginning of the semester, we used a dataset containing nutritional information on different kinds of cereals. One of the variables in the dataset is:

   - `fiber`: grams of fiber per serving

   Refer to the SAS output provided in answering the following questions about these data.

   (a) The distribution of the `fiber` variable is (circle one):

   i. right skewed
   ii. left skewed
   iii. roughly symmetric
   iv. no way to tell from output provided

   (b) Give the range of the `fiber` variable (numeric answer). Show your work, and tell what part of the SAS output you used to get it.

   14.0 - 0.0 = 14

   (c) Would the mean and standard deviation provide a good numeric summary of the `fiber` variable? (yes / no) Why or why not?

   Mean and standard deviation are appropriate only for data with a symmetric distribution and no extreme outliers.

Fiber

Stem	Leaf	#	Boxplot
14	0	1	*
13			
12			
11			
10	0	1	*
9	0	1	0
8			
7			
6	0	1	|
5	0000	4	|
4	0000	4	|
3	000000000000000	15	+-----+
2	000000000057	12	*--+--*
1	0000000000000000555	19	+-----+
0	0000000000000000000	19	|

----+----+----+----+

4. This set of questions is based on a dataset described as follows:

Data on Vocabulary and Education from the 1989 General Social Survey

[1] Observation Index

[2] Education, in years

[3] Vocabulary Test Score, 10-Item Test

Source: 1989 General Social Survey, National Opinion Research Center. Distributed by the Inter-University Consortium for Political and Social Research.

(a) What direction of relationship would you expect between education and vocabulary score? Circle the **one** best answer.

i. positive
ii. negative
iii. no relationship

(b) Before running correlation, we should (circle **all** of the statements that apply):

i. Make a pie chart of each variable individually.
ii. Make sure that both variables are quantitative.
iii. Check for extreme outliers in both variables.
iv. Make a scatterplot to determine whether the relationship is linear.

(c) The coefficient of variation (R^2) is 0.248 for a linear regression with education as the predictor variable and vocabulary score as the response variable. Circle **all** of the following statements that this implies.

i. There is a very strong linear relationship between the two variables.
ii. About 25% of the variability in vocabulary score is explained by education.
iii. The correlation coefficient r between the two variables is about 0.5.
iv. There is a curved relationship between education and vocabulary score.

(d) The slope b from the regression model is 0.374. Explain what this means in terms of years of education and vocabulary test score.

For each one-year increase in education, we expect on average a 0.374-point increase in vocabulary score.

3½ if leave out "expected" or "on average"

13

5. Investigators suspected that Benzo(a)pyrene, or BaP, from a pipe foundry in Phillipsburg, NJ, might be contaminating household air. This dataset presents data from 14 different days on samples of indoor air from a house near the foundry and samples of outdoor air collected at the same times. The measures are concentrations of BaP-containing particles no larger than 10 micrograms.

The two variables are: indoor air BaP outdoor air BaP

Reference: Lioy, PL, Walman, JM, Greenberg, A, Harkov, R and Pietarninen, C (1988). The total human environmental exposure study (THEES) to Benzo(a)pyrene: Comparison of the inhalation and food pathways. Archives of Environmental Health, 43: 304-312.

The following are the indoor values.

10 10 25 40 40 45 45 | 55 55 70 75 90 220 285

(a) Find the five-number summary for the distribution of indoor air BaP.

8

Min 10
Q1 40
Median 50
Q3 75
Max 285

by get Q1 = Q3

(b) Which types of plots would be appropriate for representing the distribution of this variable? (Circle **all** that apply.)

5

i. bar graph
ii. boxplot
iii. histogram
iv. pie chart
v. stem and leaf plot

13

2 | 2 9
1 |
0 | 1 1 3 4 4 5 5 5 5 7 8 9

stem-leaf by 10^2

multiply