**PH 538: Biostatistical Methods I**

**HOMEWORK 1 (Sampling and Descriptive Statistics) [Due September 15, 2016]**

**Dr. Fares Qeadan: fqeadan@salud.unm.edu**

|  |
| --- |
|  |

**Question 1 [4 points]:**

Please identify the population, sample, sampling frame, and sampling method:

A group of investigators are interested in conducting a clinical trial to determine whether taking bio-identical treatments prevents postmenopausal osteoporosis in women. They obtained a list of physicians in the region and recruited their patients by mailing out response cards to all eligible women; 60% of the cards were returned and 75% of those respondents entered the study. They were equally divided into the treatment group and placebo group, and were followed for 5 years to determine if they develop osteoporosis.

**Answer:**

*Population*:

*Sample*:

*Sampling frame*:

*Sampling method*:

**Question 2[4 points]:**

Please identify the population, sample, sampling frame, and sampling method:

Some researchers, in the city of Albuquerque, waited outside five bars they had randomly selected from a list of such establishments. For each selected bar, they stopped every 10th person who came out of the bar and asked whether he or she thought drinking and driving was a serious problem.

**Answer:**

*Population*:

*Sample*:

*Sampling frame*:

*Sampling method*:

**Question 3[2 points]:**

A 2-year study was conducted to investigate bicycle safety at a city’s 10 most congested intersections. The primary variable of interest was the number of accidents that involved a bicycle. Explain how one could use the number of accidents involving a bicycle either as a numerical or ordinal variable (recall that numerical variables could be categorized).

**Answer:**

**Question 4[2 points]:**

Briefly describe the difference between a bar graph and a histogram. Consider your own area of interest. Provide an example where a bar graph would be most appropriate. Provide an example where a histogram would be most appropriate.

**Answer:**

**Question 5[6 points]:**

Twenty fibromyalgia patients are asked to register their pain on a visual analog scale (VAS), where 0 represents no pain and 5 represents the worst pain imaginable. The responses are {0, 0, 0, 1, 1, 2, 2, 3, 3, 3, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5}.

a. What are the frequency distribution and probability distribution of this sample?

b. What are the mean and median for the VAS?

c. Use STATA to verify your answers in (a) and (b) by showing the STATA output?

**Answer:**

**Question 6[2 points]:**

Consider a study where the majority of the study subjects have high body mass index (BMI) and only a few have low BMI. When a histogram is constructed to describe the distribution of BMI, what shape of the distribution is most likely to be observed (right-skewed, left-skewed, or symmetric)?

**Answer:**

**Question 7[2 points]:**

You suspect that the salary of study participants might be skewed. A statistician told you that the skewed data should not be described by a mean. Why?

**Answer:**

**Question 8[2 points]:**

John waited 30 minutes to be treated in an emergency room. A 30-minute wait is in the 20th percentile of the wait time. Did he have a comparatively long- or short wait time? Interpret this percentile.

**Answer:**

**Question 9[1 points]:**

When are the mean and median the same?

a. If the distribution is symmetric.

b. If the distribution is not symmetric.

c. The mean and median can never be the same.

d. The sample size must be small.

**Answer:**

**Question 10[2 points]:**

You are given a large data set that includes patients from six major hospitals. You are interested in estrogen levels among the women in the data set. You find that most of the measurements are similar, but a few are rather large. If you plotted a histogram, what shape would it take on? How would the mean and median compare?

**Answer:**

**Question 11[4 points]:**

An investigator collected HbA1c levels on a sample of 10 patients who are suspected to be at risk for diabetes. The values for the 10 patients are 2, 2, 2, 4, 4, 6, 7, 7, 9, and 9. The mean level was 5.2 with a standard deviation of 2.78. The median was 5 with an interquartile range of 5 (Q1=2 and Q3=7). The investigator adds two more patients with values 1 and 12.

1. Explain what will happen to the mean, standard deviation, median and IQR of this distribution?
2. Use STATA to verify your answers in (a) by showing the STATA output?

**Answer:**

**Question 12[1 points]:**

Which of the following is NOT a good descriptive statistic for categorical variables?

a. proportions c. counts

b. percentages d. means

**Answer:**

**Question 13[7 points]:**

Identify the following variables as nominal, ordinal, or dichotomous [if the variable is categorical with only two levels then call it dichotomous]. Provide a justification for your choice.

a. Types of diabetes: type 1 (insulin-dependent), type 2 (non-insulin-dependent), gestational diabetes mellitus.

b. Time to recurrence: 0–2 years, 2–4 years, more than 4 years.

c. Clinic sites: Lexington, Covington, Frankfort, Nashville, Bowling Green.

d. Smoking status: smoker, nonsmoker.

e. Levels of body mass index (BMI): normal, overweight, obese.

f. Types of vaccine administration: oral, injection, puncture.

g. Types of residency: urban, rural.

**Question 14[2 points]:**

Can a variable be regarded as both ordinal and dichotomous? Why or why not?

**Answer**

**Question 15[4 points]:**

For the first cell (Low Income and White Race) in the following contingency table, interpret the frequency, percent, row percent, and column percent.



**Answer:**

**Question 16[1 points]:**

 ................... statistics involves summarizing and describing important features of the data.

(a) Descriptive

(b) Inferential

**Answer**

**Question 17[1 points]:**

Techniques for using sample information to draw dependable conclusions about the population are gathered within the branch of statistics called …………………………. statistics.

(a) Descriptive

(b) Inferential

**Answer**

**Question 18[1 points]:**

Which of the following statements is false?

(a) A variable is any characteristic whose value may change from one object to another in the population.

(b) All drug companies in the USA is an example of a sample.

(c) Sampling frame is a listing of the individuals or objects to be sampled.

(d) ”All possible yields (in grams) from a certain chemical reaction carried out in a laboratory” is an example of conceptual or hypothetical population.

**Answer:**

**Question 19[1 points]:**

Which of the following statements regarding histograms are correct?

(a) A unimodal histogram is one that rises to a single peak and then declines, whereas a bimodal histogram is one that has two different peaks.

(b) A unimodal histogram is positively skewed if the right or upper tail is stretched out compared to the left or lower tail.

(c) A unimodal histogram is negatively skewed if the left or lower tail is stretched out compared to the right or upper tail.

(d) A histogram is symmetric if the left half is a mirror image of the right half.

(e) All of the above.

**Answer:**

**Question 20[1 points]:**

 Which of the following statements are false?

(a) A physical interpretation of the sample mean $\overbar{x} $demonstrates how it measures the location (center) of a sample.

(b) The sample median is the middle value when the observations are ordered from smallest to largest.

(c) The sample median is very sensitive to extremely small or extremely large data values (outliers).

(d) The sample mean is very sensitive to extremely small or extremely large data values (outliers).

**Answer:**

**Question 21[16 points (2 points for each part)]:**

Consider the Diabetes and obesity, cardiovascular risk factors data set we have used in Lab 1 (link: <http://www.mathalpha.com/lab1/diabetes.dta>) to answer the following questions:

1. The diabetes status variable ***diab*** is an example of what type of data?

**Answer**

1. Use the **count** command in STATA to show how many observations (records) are in the data set? (show output)

**Answer**

1. Use the **nmissing** command in STATA to show how many subjects in the study don’t have diabetes status? (show output) [if you get an error then use:

**“ssc install nmissing**” to install it first]

**Answer**

1. Use the **tab** command in STATA to calculate the relative frequency of diabetic subjects in the study? (show output)

**Answer**

1. Is the relative frequency of diabetes higher in males or females? (show output)

**Answer**

1. Were there more males or females (absolute count) with diabetes? [use the output from part (e) to answer this question]

**Answer**

1. Use STATA to display the contingency table for the diabetes variable (***diab***) cross-tabbed with body frame variable (***frame***)? Do you observe any association between diabetes status and body frame? Explain

**Answer**

1. Please run the following two lines of code and explain what does each one of them do? (show graph)

ssc install catplot

catplot diab frame, percent(frame) blabel(bar)

**Answer**

**Question 22[14 points (2 points for each part)]:**

Consider the Diabetes and obesity, cardiovascular risk factors data set we have used in Lab 1 (link: <http://www.mathalpha.com/lab1/diabetes.dta>) to answer the following questions:

1. The First Systolic Blood Pressure variable (***bp\_1s***) is an example of what type of data?

**Answer**

1. Give the five numbers summary for this variable? [Show output]

**Answer**

1. Create a histogram for the First Systolic Blood Pressure variable (***bp\_1s***) and describe its distribution, being sure to consider symmetry, skewness, presence of outliers, and modality? [make sure you give appropriate labeling to your graph]

**Answer**

1. How could you support your answer in part (c) regarding the skewness of the distribution sing the numerical summary statistics? Explain.

**Answer**

1. What measures of location (center tendency) and variability should we use for numerically describing the First Systolic Blood Pressure variable? Please give their values. [show output]

**Answer**

1. Use:

 tabstat bp\_1s, by(gender) statistics(n mean sd median min max p25 p75 iqr)

to fill in the following table:

|  |  |
| --- | --- |
|  | 1st SBP (***bp\_1s***) |
|  | Males | Females |
| Number of non-missing observations |  |  |
| Mean 1st SBP:  |  |  |
| St. Dev. of 1st SBP |  |  |
| Median 1st SBP |  |  |
| Range of 1st SBP |  |  |
| IQR of 1st SBP |  |  |

1. Please draw a box plot for 1st SBP for each sex category in one figure and make comments on the distribution by sex?

**Answer**

**Question 23[6 points (2 points for each part)]:**

The numeric difference between your systolic and diastolic blood pressure is called your pulse pressure (PP).

1. Use the **gen** command in STATA to create the new variable PP and then make a histogram for the new generated variable.

**Answer**

b) Create a categorical PP variable and call it PP\_cat such that PP\_cat is 1 if PP is larger or equal than 60 and it’s 0 if PP is less than 60. Please provide the frequency ad relative frequency distribution for the new generated categorical variable?

**Answer**

c) Is the relative frequency of diabetes higher in subjects with PP>=60 compared to subjects with PP<60? [Show output for the contingency table of PP\_cat and diab]

**Answer**

**Question 24: [14 points (4 points for (a) and 10 points for (b))]:**

1. Do you think that our data set from lab 1 has enough information to create a figure identical or similar to figure 3 in the following article? Explain

*Diabetes, pulse pressure and cardiovascular mortality: the Hoorn Study:* <http://dspace.ubvu.vu.nl/bitstream/handle/1871/22224/263589.pdf?sequence=1>

1. If your answer in (a) was yes then please use STATA to construct the plot. If the answer was no then construct a plot as similar as you could, given the available data.

**Answer**

**Question 25 [EXTRA CREDIT] [up to 3 points]:**

Suggest a way to calculate the mode and range in STATA?