Weekly Quiz 1

The **due date** for this quiz is **Sun 27 Jan 2013 8:30 PM PST**.

Question 1

A web administrator is examining the web log file which provides information about people who visited his site. In the log file, the administrator sees a value for the internet protocol (IP) address of "127.0.0.1", a user ID of "frank" for the individual accessing the file and "2326", measured in bytes, for the size of the file returned to the user. Which of the following are true? Here is a description of a web log file in common log format: http://en.wikipedia.org/wiki/Common_Log_Format. Here is a description of an IP address: http://en.wikipedia.org/wiki/IP_address

○ The IP address, user ID, and bytes returned are all qualitative variables.

○ The IP address, user ID, and bytes returned are all quantitative variables.

○ The IP address and bytes returned are quantitative variables and the user ID is a qualitative variable.

● The IP address and the user ID are qualitative variables and the number of bytes returned is quantitative.

Question 2

Suppose that random variable X follows a Poisson distribution with rate parameter L. If we increase the value of L, which of the following is true?

● The spread increases but the center remains unchanged.

○ The spread increases but the center decreases.

○ The center increases but the spread remains unchanged.

○ Both the spread and the center increase.

Question 3

Run the following commands to create a data frame in R with measurements for 30 men describing their height in centimeters, weight in kilograms, and a logical indicator for whether they have a daughter or not.

```
set.seed(31);
heightsCM = rnorm(30,mean=188, sd=5);
weightsK = rnorm(30,mean=84,sd=3);
hasDaughter = sample(c(TRUE,FALSE),size=30,replace=T);
dataFrame = data.frame(heightsCM,weightsK,hasDaughter);
```

Subset the data frame to only the individuals that are greater than 188 centimeters tall. Assign this subset to a data frame called dataFrameSubset. Then run this command: mean(dataFrameSubset$weightsK) to get the average weight among this subset of men in the data. What is the value that is produced?

○ 82.48989

○ 83.77462

○ 83.13611

● 82.40639

Question 4

Run a command to generate 100 Cauchy random variables with default parameters and assign them to a vector cauchyValues immediately after running the command

```
set.seed(41)
```

Then run a command to sample 10 values with replacement from cauchyValues immediately after running the command

```
set.seed(415)
```

What are the first three values of the resulting sample? Note: It is critical that you run the set.seed commands immediately before the commands to perform the data generation and sampling or you will get the wrong answer.

○ 0.8084719, 1.7312325, 0.3716671

○ -0.05093145, 6.66059126, -12.09755185

○ The answer is none of the other options.

● 0.8084719, -1.1122863, 0.3716671

Question 5

We take a random sample of individuals in a population and identify whether they smoke and if they have cancer. We observe that there is a strong relationship between whether a person in the sample smoked or not and whether they have lung cancer. We claim that the smoking is related to lung cancer in the larger population. We explain we think that the reason for this relationship is because cigarette smoke contains known carcinogens such as arsenic and benzene, which make cells in the lungs become cancerous.

● This is an example of an inferential data analysis.

○ This is an example of an descriptive data analysis.

○ This is an example of a causal data analysis.

○ This is an example of a mechanistic data analysis.

Question 6

Suppose that we collect data on every goal scored in the Spanish Primera Division (http://soccernet.espn.go.com/stats/scorers/_/league/esp.1/spanish-primera-division?cc=5901) in the 2011/2012 and 2012/2013 seasons. We use the data from 2011/2012 to build a model to predict the number of goals scored in 2012/2013. What is the complete list of labels that apply to this data set?

● Census, prediction, longitudinal

○ Prediction and longitudinal

○ Only longitudinal

○ Prediction, inferential, and longitudinal

Question 7

What are the three characteristics of tidy data?

○ 1. Each variable forms a column, 2. Each table is tab delimited, 3. Quantitative and qualitative variables are stored in separate files/tables.

○ 1. Each variable forms a column, 2. Each observation forms a row, 3. Quantitative and qualitative variables are stored in separate files/tables.

○ 1. Each observation forms a row, 2. No variable has missing values 3. Quantitative and qualitative variables are stored in separate files/tables.

● 1. Each variable forms a column, 2. Each observation forms a row, 3. Each table/file stores data about one kind of observation.

Question 8

Which of the following are components of data processing that should be recorded for use in later data analyses?

○ Which missing values were removed

○ Which variables were omitted because they were not relevant to the analysis

○ That a variable was log transformed to make it easier to analyze

● All of the above

Question 9

When writing about data, what does it mean when we write: X | Y = y?

● We are referring to the random variable X when we know the random variable Y has value y. The distribution of this variable may be different than the distribution of the variable X when Y is also random.

○ We are referring to the random variable X when we know the random variable Y has value y. The distribution of this variable is the same as the distribution of the variable X when Y is also random.

○ We are referring to the random variable X when we know the random variable Y has a random value y.

○ We are referring to the random variable X when we know the random variable Y has value y. In this case, we know X also has value y.

Question 10

Suppose we take a sample of people in Baltimore and observe that younger people have taken more Coursera courses. We use an inferential data analysis to show a relationship between age and Coursera courses. Which of the following statements are true based only on our analysis?

○ The reason that younger people take more Coursera courses is that they understand technology better.

○ When comparing two people in Baltimore, the one who is younger will always have taken more Coursera courses.

● If we took a census of all people in Baltimore, we would expect to see that the younger a person was, the more likely they were to have taken a Coursera course.

○ If we took a census of all people in Boston, we would expect to see that the younger a person was, the more likely they were to have taken a Coursera course.

☑ In accordance with the Honor Code, I certify that my answers here are my own work.

Submit Answers Save Answers