TRGN 599 Midterm Spring 2019 – 03/07/2019 - Answers

Name: _____

INSTRUCTIONS: Please answer the following questions. One answer per question. Questions 61-63 are extra points. Please ask your instructor if you have any questions. (* = Correct Answer)

Question 1
Which skills have become an inherent component of life-science research, particularly 'omic'-based research?
*A) Bioinformatics
B) Marketing
C) Artificial Intelligence
D) Clinics

Question 2
Which essential skills are essential to many research projects today?
A) Clinical consultation
B) Marketing consultation
*C) Data analysis and interpretation, and especially in data management
D) Artificial intelligence

Question 3
Usually statistical analyses begin with which of the following?
A) ANOVA table
*B) Data
C) Confusion matrix
D) Graphs

Question 4
Usually statistical analysis end with which of the following?
A) Confusion matrix
B) Data management
C) Data analysis
*D) Reports

Question 5
Which of the following R functions create a graph?
A) bio()
B) data()
*C) plot()
D) length()

Question 6
The R function summary() provides which of the following information?
A) ANOVA table
*B) mean, median, 25th and 75th quartiles, min, max of the variables in a dataset
C) mode, mean, median, 25th and 75th quartiles, min, max of the variables in a dataset
D) T-test

Question 7
Which of the following is a definition of “Continuous Data”?
A) Data analysis and interpretation
B) Reports of variables
C) Data that can take on only a specific set of values representing a set of possible categories
*D) Data that can take on any value in an interval

Question 8
Which of the following are more likely examples of “Discrete Data” ?
*A) number of students in a class, results of rolling 2 dice: 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 and 12
B) mean, median, 25th and 75th quartiles, min, max of the variables in a dataset
C) 0/1, true/false
D) Reports

Question 9
Which of the following are more likely examples of “Binary Data” ?
A) Data that can take on only a specific set of values representing a set of possible categories
B) Race, sex, educational level
*C) 0/1, true/false
D) Data that can take on any value in an interval

Question 10
Which of the following is a definition of “Categorical Data”?
A) Number of students in a class, results of rolling 2 dice: 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 and 12
*B) Data that can take on only a specific set of values representing a set of possible categories
C) Data analysis and interpretation, and especially in data management
D) Data that can take on any value in an interval

Question 11
Which of the following is a basic data structure for statistical and machine learning models also called “Rectangular data” (like a spreadsheet)?
A) 0/1, true/false
B) Array data
*C) Dataframe
D) Date data

Question 12
Which of the following plot is frequently used as a quick way to visualize the distribution of data?
A) Number of students in a class, results of rolling 2 dice: 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 and 12
B) Data that can take on only a specific set of values representing a set of possible categories
*C) Boxplot
D) Data that can take on any value in an interval

Question 13
What is the advantage of a violin plot over a boxplot?
A) Use binary data such as 0/1, true/false
*B) It can show nuances in the distribution that aren't perceptible in a boxplot
C) It is similar than a dataframe
D) It can show data that can take on any value in an interval that isn't perceptible in a boxplot

Question 14
Which of the following is a definition of "Frequency table"?
A) Number of students in a class, results of rolling 2 dice: 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 and 12
B) Data that can take on only a specific set of values representing a set of possible categories
*C) A tally of the count of numeric data values that fall into a set of intervals
D) Table with data that just can take on any value in a dataset

Question 15
What is the Mode of the following set of numbers: 3, 3, 2, 1, 4, 5, 3, 3, 4, 9,7?
A) 4
B) 7
*C) 3
D)9

Question 16
Which of the following is a definition of "Density plot"?
*A) A smoothed version of the histogram often based on a kernel density estimate
B) Data that can take on only a specific set of values representing a set of possible categories
C) A tally of the count of numeric data values that fall into a set of intervals
D) A smoothed version of the histogram with data that can take on any value in an interval

Question 17
Which of the following is a familiar way of dealing with differing survival times (times-to-event), especially when not all the subjects continue in the study?
A) Confusion matrix
B) Linear regression output
*C) Kaplan-Meier curves
D) ANOVA table

Question 18
Which of the following is a graphical representation of data where the individual values contained in a matrix are represented as colors?
A) Kaplan-Meier curves
*B) Heatmap
C) Linear regression output
D) Confusion matrix

Question 19
What is statistical bias?
A) A Statistical Analysis
B) A Categorical variable
*C) A Systematic error
D) A hypothesis testing

Question 20
Which of the following is a subset from a larger data set?
*A) Sample
B) Heatmap
C) Hypothesis testing
D) Frequency table

Question 21
What specifies the probability distribution of a random variable?
*A) Its possible values (i.e., its range) and their corresponding probabilities
B) Systematic error
C) Kaplan-Meier curves
D) Hypothesis testing

Question 22
Which of the following distributions is famously represented as a bell curve?
*A) Normal distribution
B) Bernoulli distribution
C) Binomial distribution
D) Residuals distribution

Question 23
Which of the following is a definition of Poison Distribution?
A) Is a Normal distribution
B) It can show nuances in the distribution that aren't perceptible in a boxplot
*C) The frequency distribution of the number of events in sampled units of time or space
D) The frequency distribution of a Binomial distribution

Question 24
A sequence of binary random variables X1, X2, . . . , Xn is called Bernoulli trials if:
A) A number of students in a class, results of rolling 2 dice: 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 and 12
B) Data that can take on only a specific set of values representing a set of possible categories
*C) They all have the same Bernoulli distribution (i.e., the same probability θ for the outcome of interest) and are independent (i.e., not affecting each other's probabilities)
D) The frequency distribution is a Normal distribution

Question 25
Which of the following is a definition of "Data distribution"?
A) Is 0/1 and true/false like data distributed in a bell form
B) It can show nuances in the distribution that aren't perceptible in a boxplot
*C) The frequency distribution of individual values in a data set
D) The frequency table data that can take on any value in an interval

Question 26
Which of the following probability functions is used to specify the distribution of continuous random variables?
A) The frequency table data that can take on any value in an interval
B) The frequency distribution of individual values in a data set
*C) Probability density function (pdf)
D) Probability mass function (pmf)

Question 27
Which of the following probability functions provides the probability of each possible value in discrete random variables?
A) The frequency table data that can take on any value in an interval
B) The frequency distribution of individual values in a data set
C) Probability density functions (pdf)
*D) Probability mass function (pmf)

Question 28
Which of the following is true regarding a normal distribution?
A) A number of students in a class, results of rolling 2 dice: 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 and 12.
*B) Is fully specified by the mean μ and variance σ2
C) Probability density functions (pdf)
D) Probability mass function (pmf)

Question 29
Why we say that the probability distribution is unimodal?
A) Because the height of the density curves reduces to one symmetrically as we move away from the center, we say that the probability distribution is non-symmetric.
*B) Because the height of the density curves reduces to zero symmetrically as we move away from the center, we say that the probability distribution is symmetric.

C) Because the probability density functions (pdf)
D) Because the probability mass function (pmf)

Question 30
Which calculation you need to do to convert data to z-scores?
A) You add the number of students in a class, or the results of rolling 2 dice: 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 and 12
*B) You subtract the mean of the data and divide by the standard deviation
C) You sum the probability density functions (pdf)
D) You multiply the probability mass function (pmf)

Question 31
Which of the following is a definition of Central limit theorem?
*A) The tendency of the sampling distribution to take on a normal shape as sample size rises
B) The tendency to sum the height of the density curves that reduces to one symmetrically as we move away from the center
C) Is the probability density function (pdf)
D) Is the probability mass function (pmf)

Question 32
Which of the following plots could help to visualize how close a sample distribution is to a normal distribution?
A) Pie plot
*B) QQ-Plot
C) Probability density function (pdf)
D) Probability mass function (pmf)

Question 33
Which of the following is any statistical association, though in common usage, it most often refers to how close two variables are to having a linear relationship with each other?
A) The tendency of the sampling distribution to take on a normal shape as sample size rises
*B) Correlation
C) ANOVA table
D) Probability mass function (pmf)

Question 34
Which of the following correlations are appropriate whenever there is a concern that the data is not “normally distributed”?
*A) Spearman Correlation
B) Pearson Correlation
C) Probability density functions (pdf)
D) ANOVA table

Question 35
Which of the following are statistical measures of the performance of a binary classification test or classification function?
A) Spearman Correlation
B) Correlation
*C) Sensitivity and Specificity
D) Probability mass function (pmf)

Question 36
Which of the following terms is the proportion of a population who have a specific characteristic in a given time period?
A) Spearman Correlation
B) Pearson Correlation
*C) Prevalence
D) Probability mass function (pmf)

Question 37
Which of the following is a statistical measure of the performance that measures the proportion of actual positives that are correctly identified as such?
A) The tendency of the sampling distribution to take on a normal shape as sample size rises
*B) Sensitivity
C) Specificity
D) Probability mass function (pmf)

Question 38
Which of the following is a statistical measure of performance that measures the proportion of actual negatives that are correctly identified as such ?
A) The tendency of the sampling distribution to take on a normal shape as sample size rises
B) Sensitivity
*C) Specificity
D) Pearson correlation

Question 39
Regarding an Application to screening study: Imagine a study evaluating a new test that screens people for a disease. Each person taking the test either has or does not have the disease. The test outcome can be positive (classifying the person as having the disease) or negative (classifying the person as not having the disease). The test results for each subject may or may not match the subject's actual status. Which of the following is false?
A) True positive: Sick people correctly identified as sick
B) False positive: Healthy people incorrectly identified as sick
C) True negative: Healthy people correctly identified as healthy
*D) False negative: Healthy people incorrectly identified as healthy

Question 40
In a Confusion Matrix the True positive is also named:
A) The tendency of the sampling distribution to take on a normal shape as sample size rises
B) Sensitivity
*C) Power
D) Probability mass function (pmf)

Question 41
In a Confusion Matrix the False negative is also called:
A) True positive
B) Power
C) Type I error
*D) Type II error

Question 42
In a Confusion Matrix the sum of false positives divided by the sum of the predictive condition positive is called:
*A) False discovery rate (FDR)
B) Sensitivity
C) Power
D) Probability mass function (pmf)

Question 43
Which of the following denote the null hypothesis:
*A) H0
B) HA
C) True negative: Healthy people correctly identified as healthy
D) False negative: Healthy people incorrectly identified as healthy

Question 44
The procedure for evaluating a hypothesis is called:
*A) Hypothesis testing
B) Sensitivity
C) Power
D) Probability mass function (pmf)

Question 45
Which of the following occurrences generate a Type I error?
A) H0
B) HA
*C) We reject H0 when it is true and should not be rejected
D) We fail to reject H0 when it is false and should be rejected

Question 46
Which of the following occurrences generate a Type II error?
A) H0
B) HA
C) We reject H0 when it is true and should not be rejected
*D) We fail to reject H0 when it is false and should be rejected

Question 47
Now suppose that we have a hypothesis testing procedure that fails to reject the null hypothesis when it should be rejected with probability β. This means that our test correctly rejects the null hypothesis with probability 1 − β. (Note that the two events are complementary.) We refer to this probability (i.e., 1 − β) as:
A) H0
B) HA
*C) The power of the test
D) We fail to reject H0 when it is false and should be rejected

Question 48
In a z-test, instead of comparing the observed sample mean $\bar{x}$ to the population mean according to the null hypothesis, we compare the:
A) H0
B) HA
*C) Z-score to 0
D) We fail to reject H0 when it is false and should be rejected

Question 49
Which of the following is a definition of p-value?
A) Is the H0
*B) Is the conditional probability of extreme values (as or more extreme than what has been observed) of the test statistic assuming that the null hypothesis is true
C) Is the power of the test
D) Is when we fail to reject H0 when it is false and should be rejected

Question 50
About interpreting the p-value, what means when the p-value is small, say 0.01 for example?
*A) It means it is rare to find values as extreme as what we have observed
B) It means the HA
C) It means the the conditional probability of extreme values (as or more extreme than what has been observed) of the test statistic assuming that the null hypothesis is true
D) It means the fail to reject H0 when it is false and should be rejected

Question 51
In order to use the p-value to decide whether we should reject the null hypothesis, a convenient approach is to prespecify a cutoff for the p-value and reject the null hypothesis if pobs is below the cutoff. This cutoff is called:
*A) Significance level or the size of the test
B) The conditional probability of extreme values (as or more extreme than what has been observed) of the test statistic assuming that the null hypothesis is true
C) The power of the test
D) The variance

Question 52
The most common significant levels of a p-value are?
A) 1, 5, 10
*B) 0.01, 0.05, and 0.1
C) The conditional probability of extreme values above one
D) 100, 200, 300

Question 53
Which hypothesis testing procedure is used when the population variance σ2 is unknown (need to be estimated from the data)?
A) Normal distribution
B) Binomial distribution
C) The power of the test
*D) T-test

Question 54
Which of the following testing procedures can be used to evaluate the appropriateness of the normality assumption?
A) ANOVA table
*B) Shapiro–Wilk test of normality
C) Confusion Matrix
D) Survival analysis

Question 55
Which regression model is used when we are interested in the relationship between the response variable and multiple explanatory variables?
A) 1, 5, 10
B) Linear regression model with one continuous explanatory variable
C) Linear regression model with one binary explanatory variable
*D) Multiple Linear Regression

Question 56
Which of the following datasets has available a vast of microarray data?
A) Pubmed

*B) Gene Expression Omnibus (GEO)
C) Google cloud
D) Norris Medical Library (NML)

Question 57
Which of the following is NOT an normalization approach used in Microarray Analyses?
A) Locally weighted scatterplot smoothing (LOWESS normalization)
B) Robust multi-array average (RMA)
C) Variance stabilizing normalization (VSN)
*D) T-test

Question 58
Which of the following is the ultimate goal of gene expression microarray experiments?
A) To identify the structure of the proteins
B) To identify data in the Gene Expression Omnibus (GEO)
*C) To identify genes that are significantly over- or under-expressed in groups of samples
D) To identify proteins that are significantly in groups of samples

Question 59
Which of the following is a statistical procedure that uses an orthogonal transformation (Linear transformation) to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables?
A) Variance stabilizing normalization (VSN)
B) Robust multi-array average (RMA)
*C) Principal component analysis (PCA)
D) t-test

Question 60
A hypothesis test that is used to compare the means of two populations is called?
A) Robust multi-array average (RMA).
B) Gene Expression Omnibus (GEO)
*C) T-test
D) Principal component analysis (PCA)

Question 61
A hypothesis test that is used to compare the means of more than two populations is called?
*A) Analysis of Variance or ANOVA
B) Gene Expression Omnibus (GEO)
C) Shapiro–Wilk test of normality
D) Binomial distribution

Question 62
Regarding the interpretation of ANOVA, lets suppose that we would like to investigate the effectiveness of various feed supplements (feed) on the growth rate (weight) of chickens. We use ANOVA examine the effectiveness of feed supplements. Which of the following interpretations is correct of the attached R output?

```
            Df Sum Sq Mean Sq F value   Pr(>F)
feed         5 231129   46226   15.37 5.94e-10 ***
Residuals   65 195556    3009
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

*A) The observed value of F-statistic is f = 15.37, and the corresponding p-value is quite small. Therefore, we can reject the null hypothesis and conclude that based on this experiment, various feed supplements have quite different effects on the growth rate and the relationship between the two variables (feed type and weight) is statistically significant.
B) The observed value of F-statistic is f = 46226, and the corresponding p-value is quite small. Therefore, we can reject the null hypothesis and conclude that based on this experiment, various feed supplements have quite different effects on the growth rate and the relationship between the two variables (feed type and weight) is not statistically significant.
C) The observed value of F-statistic is f = 15.37, and the corresponding p-value is quite large. Therefore, we cannot reject the null hypothesis and conclude that based on this experiment, various feed supplements have quite different effects on the growth rate and the relationship between the two variables (feed type and weight) is not statistically significant.
D) The observed value of F-statistic is f = 15.37, and the corresponding p-value is quite small. Therefore, we cannot reject the null hypothesis and conclude that based on this experiment, various feed supplements have quite different effects on the growth rate and the relationship between the two variables (feed type and weight) is not significant.

Question 63
Which of the following is a definition of the Bayes' theorem?
*A) Describes the probability of an event, based on prior knowledge of conditions that might be related to the event
B) Is a Robust multi-array average (RMA)
C) Describes the orthogonal transformation (Linear transformation), converting a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables
D) Is a hypothesis testing method, based on prior knowledge of conditions