Practice Midterm

HS616

March 24, 2015

Question 1

The Poisson Distribution is a type of

- A : Cumulative distibution
- B : Continuous Probability Distribution
- C : Discrete Probability Distribution
- D : Random number generation

Question 2

What data type does apply function return?

- A : Lists
- B : Vectors
- C : All of these answers are correct
- D : Matrices

Question 3

Which of the following equations represents the sensitivity of a test?

- A : sensitivity = number of true positives / number with disease
- B : sensitivity = number of true positives / number of true negatives
- C : sensitivity = number of true negatives / number without disease
- D : sensitivity = number with disease / total population

Question 4

The command `tidyr::gather(df, var, val)` produced the following result:

```
  var val
1   a   1
2   a   2
3   a   3
4   b   1
5   b   2
6   b   3
```

Which answer correctly defines the dataframe `df`?

- A : `df <- data.frame(var=letters[1:3], val=letters[1:3])`
- B : `df <- data.frame(a=var[1:3], b=val[1:3])`
- C : `df <- data.frame(var=rep(c('a','b'), each=3), val=rep(1:3, times=2))`
- D : `df <- data.frame(a=1:3, b=1:3)`

Question 5

What does the following function return?

```
f <- function(x) {
  f <- function(x) {
    f <- function(x) {
      x ^ 2
    }
    f(x) + 1
  }
  f(x) * 2
}
f(10)
```

- A : 441
- B : 202
- C : 200
- D : 40

Question 6

What is the correct code for subtracting two dates from one another and then cast the difference to a numeric value?

- A : `as.numeric %>% (as.Date("2014-10-10" - "2014-10-1" ))`
- B : `as.Date("2014-10-10") - as.Date("2014-10-1") %>% as.numeric`
- C : `(as.Date("2014-10-10") - as.Date("2014-10-1")) %>% as.numeric`
- D : `as.Date %>% ("2014-10-10") - as.Date %>% ("2014-10-1") >%> as.numeric`

Question 7

Simulated coin-tossing can be done using different methods. Which of the following will NOT work?

- A : rbinom(10, 1, .5)
- B : coin <- sample(c("H", "T"), 10, replace = F)
- C : ifelse(rbinom(10, 1, .5) == 1, "H", "T")
- D : c("H", "T") [1 + rbinom(10, 1, .5)]

Question 8

Which characteristics describe "tidy" data?

- A : Multiple variables are stored in one column. Each observation forms a row. Column headers are values, not variable names.
- B : Column headers are values, not variable names. Variables are stored in both rows and columns.
- C : As many observational units as possible are stored in the same table. Do not store a single observational unit in a single table.
- D : Each variable forms a column. Each observation forms a row. Each type of observational unit forms a table.

Question 9

In the following code, what values of m and n will produce a plot showing a quarter of a circle?

```
N <- 10000
x <- runif(N, min=m, max=n)
y <- runif(N, min=m, max=n)
plot(x, y, pch=16, col=ifelse(x^2 + y^2<1, "red", "blue"))
```

- A : m=-3.0; n=3.0
- B : m=-1; n=0
- C : m=-2.0; n=2.0
- D : m=-1.0; n=1.0

Question 10

Indentify the distribution type in the following code:

```
x <- seq(0, 4, 0.1)
plot(x, dnorm(x, 2, 0.5), type = "l")
```

- A : Normal
- B : Unified constant
- C : Binomial
- D : Poisson

Question 11

In the following code, what is the type of the variable `v`?

```
v <- runif(10) < 0.5
```

- A : character
- B : integer
- C : numeric
- D : logical

Question 12

Consider the following code:

```
N <-10000
x <- runif(N)
y <- runif(N)
vlength <- sqrt(x^2 +y^2)
in_circle <- vlength < 1
```

Which of the following could be the output of `head(as.integer(in_circle))` ?

- A : 1 1 1 1 1 0
- B : 0.23, ,0.34, 0.12, 0.45, 0.55, 0.79
- C : 1 -1 1 0 -1 0
- D : TRUE TRUE TRUE TRUE TRUE FALSE

Question 13

In database management, what is meant by "Data Aggregation"?

- A : Normalizing the data in a database table
- B : Using an inner join to extract data from a table
- C : Finding the mean of columns in a database table
- D : The process by which data is gathered and summarized for further statistical analyses

Question 14

The function head() does this:

- A : summarizes the data in a table
- B : creates a header in the data frame
- C : displays the first few observations of a data frame

Question 15

Every data type is at least a _____

- A : matrix
- B : vector
- C : array
- D : factor

Question 16

sqldf is a fantastic tool for data scientists. Which of the following statements are true?

- A : All of these
- B : sqldf operates on dataframes
- C : Right and full outer joins, which are unavailable in sqldf, can be accomplished with the "merge" function of base R
- D : sqldf is a useful tool for manipulation data with such statements such as: sqldf::sqldf("SELECT * FROM A JOIN B ON a=b")

Question 17

What is the correct way to vectorize the following code:

```
for(i in 1:3) x[i] <- i+i
```

- A : `for(i in range(1,4)) x+= [i+i]`
- B : `x <- c(1,2,3) + c(1,2,3)`
- C : `for(i<4) x[i] <- 2i`
- D : `while(i<4) x+= [2i]`

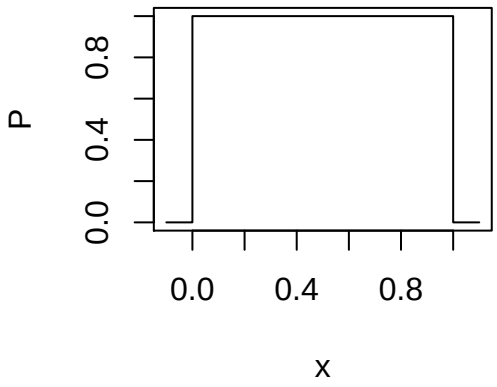
Question 18

What needs to be changed in the following code for values to be arranged row wise in ascending order?

```
m <- matrix(1:20, nrow=5, ncol=4)
```

- A : `byrow = TRUE`
- B : No change required in the code
- C : `byrow = FALSE`
- D : `bycol = TRUE`

Question 19



What is the name of the following distribution?

- A : Uniform distribution
- B : Binominal distribution
- C : Normal distribution
- D : Poisson distribution

Question 20

What is typically the fastest way to analyze and manipulate data using R?

- A : With recursion
- B : With vectorized functions
- C : Using iteration
- D : With loops

Question 21

Which symbol can be used for slicing and extracting data from a vector in R?

- A : `[, c( )]`
- B : `[[c( ) ]]`
- C : `$`
- D : `[ ]`

Question 22

"Setting the seed", e.g. `set.seed(42)`, in R...

- A : has nothing to do with random number generation.
- B : ensures that the outcome of random number generators is *not* repeated upon re-execution of your code.
- C : ensures that someone else who runs your code does not get the same random numbers you do.
- D : ensures that the outcome of random number generators will be repeated upon re-execution of your code.

Question 23

Consider the following profiling results:

```
             self.time self.pct
"function_A"    278.39    86.46
"function_B"     29.32     9.10
"function_C"     14.29     4.44
```

If you make `function_B` 100 times faster, how much faster would you expect the program be?

- A : 100 times as fast
- B : twice as fast
- C : no faster
- D : less than 10% faster

Question 24

Which of these is not a problem with messy data

- A : Multiple variables stored in a single column
- B : Values stored in table format
- C : Multiple types of entities in the same table
- D : Variables stored in both rows and columns

Question 25

Explain what the first line of code does in making a table or dataframe named "less_toxic"

```
less_toxic <- read.csv("toxic_text.csv", na.strings=c("UNK", "?"))
knitr::kable(data.frame(
  toxic = sapply(toxic, class),
  less_toxic = sapply(less_toxic, class)
)
```

- A : reads a csv file and from the knitr library kables or knocksout table entries, hence the acronym kable in the knock out table
- B : reads a csv file named ("toxic_test.csv") and puts "NA" for those entries that are marked 'UNK' or with a question mark.
- C : reads a csv file and halts if a missing or unknown character string is encountered
- D : writes a csv file to toxic_test.csv and invokes an Excel workbook session after making the dataframe

Question 26

The runif(n) function in R:

- A : is similar to ifelse(); it only runs if 'n' is TRUE.
- B : doesn't really do anything
- C : returns a vector of 'n' uniformly distributed random numbers
- D : always generates numbers in the range from 0 to 100

Question 27

Which keyword is used in a SQL select statement to eliminate duplicate values within a column?

- A : DISTINCT
- B : ONLY
- C : DIFFERENT
- D : can use '*'

Question 28

What does the Central Limit Theorem state ?

- A : The distribution of the means of a set of random samples is approximately Normal
- B : The area under the normal denisty curve is one
- C : Measures of central tendency should always be computed with and without outliers
- D : Confidence intervals have zero margin of error for large sample sizes.

Question 29

In the statement `var <- runif (10) < 0.5`, what is the class() of the vector 'var' ?

- A : integer
- B : character
- C : logical
- D : list

Question 30

Which of these addresses cannot be read by the built-in `url()` function?

- A : `http://rseek.org/`
- B : `http://ftp.ics.uci.edu/pub/machine-learning-databases/`
- C : `file:///usr/share/dict/words`
- D : `https://connect.usfca.edu`

Question 31

Consider a sequence of 10 coin flips, represented by the string `TTTHTHTTTH`. Which statement gives the total number of different sequences of 10 coin flips that could result in this number of heads?

- A : `sapply(3:10, function(x) factorial(x))`
- B : `integrate(dnorm, -Inf, 0)`
- C : `choose(10,3)`
- D : `factorial(10)/(factorial(4)*factorial(7))`

Question 32

Which command opens a connection to an SQLite database?

- A : dsets <- dbConnect(RSQLite::SQLite(), “datasets.sqlite”)
- B : res <- dbSendQuery(dsets, “select * from iris limit 10”)
- C : sqliteCopyDatabase(dsets, “datasets.sqlite”)
- D : dbListTables(dsets)

Question 33

How does an ellipsis behave as a function parameter in R?

- A : Each period acts as an anonymous parameter in the function.
- B : It takes an undefined number of arguments and applies them wherever the ellipsis is used in the function, similar to a normal parameter.
- C : It takes each argument passed in by the user and applies them to undefined variables in the function based on order.

Question 34

What does the selectorGadget do?

- A : Allows you to interactively click on a web page to generate CSS selectors
- B : Generates data for a linear model
- C : Helps to select and time profiler functions
- D : Selects the best function in a given program

Question 35

Consider the equation $Av == \lambda v$. If A is the identity matrix, what is lambda?

- A : lambda is infinity
- B : lamda doesn't exist for an identity matrix
- C : lamba is equal to 1
- D : lambda is zero