Midterm Study Guide

Robert Horton

March 15, 2015

Lecture 01a

rmhorton Which symbol is used to collect a variable number of function arguments?

- `...`
- `???`
- `+++`
- `yada, yada, yada`

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

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

cpkaur In the following code what is u + v?

```
u <- c(2,3,4)
v <- c(3,4,5,6,7,8)
```

- `(5 7 9 8 10 12)`
- `(5 7 9 6 7 8)`
- `NaN`
- `(3 4 5 8 10 12)`

vchaudhuri

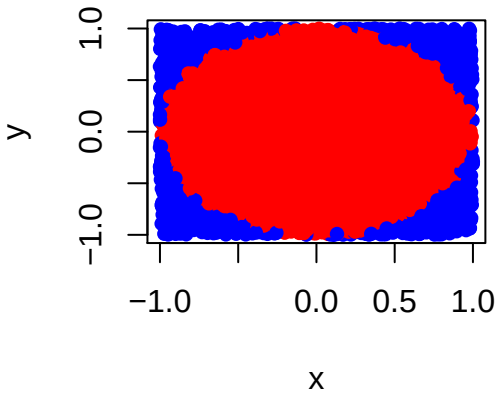
```
f <- sqrt
x <- 1:20
y <- f(x)
z <- x^2
d <- data.frame(x, y, z)
```

What is the typeof variable d ?

- list
- dataframe
- integer
- character

nsh87 The following R code generates the image below.

```r
N <- 5000
x <- runif(N, min=-1, max=1)
y <- runif(N, min=-1, max=1)
plot(x, y, pch=16, col=ifelse(x^2 + y^2 < 1, "red", "blue"))
```



How would you swap the colors in the plot?

- `plot(x, y, pch=16, col=ifelse(x^2 + y^2 > 1, "red", "blue"))`
- `plot(x, y, pch=16, col=ifelse(x^2 + y^2 > 1, "blue", "red"))`
- `plot(y, x, pch=16, col=ifelse(x^2 + y^2 < 1, "red", "blue"))`
- `plot(x, y, pch=16, col=ifelse(x^2 - x^2 < 1, "red", "blue"))`

lakarbatti If `x <- 1:4` and `y <- 5:8` what is the output of `x + y` ?

- A vector with values 6 8 10 12
- A numeric integer with value 6
- A numeric integer with values 6 8 10 12
- Running the statement gives an error

xxu26 Which symbol can be used for slicing and extracting data from a vector in R?

- `[ ]`
- `[[c( ) ]]`
- `$`
- `[, c( )]`

sneha-krishna Which of the following statements about R are true?

- All of these statements are true
- R is free!
- R can directly access and import data from a wide variety of sources, including text files, database management systems, statistical packages, web pages, and social media sites. It can write data out to these systems as well.
- R provides quite a bit of flexibility and control over where input comes from and where it goes.

Lecture 01b

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

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

- logical
- numeric
- integer
- character

cpkaur What data type does apply function return?

- All of these answers are correct
- Lists
- Vectors
- Matrices

vchaudhuri 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))` ?

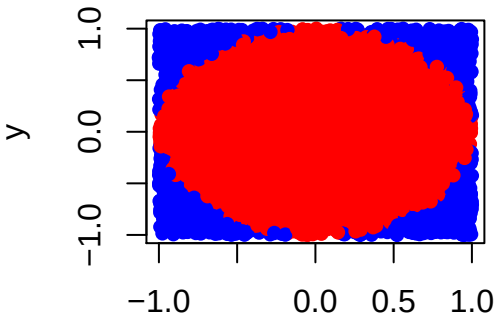
- `1 1 1 1 1 0`
- `TRUE TRUE TRUE TRUE TRUE FALSE`
- `1 -1 1 0 -1 0`
- `0.23, ,0.34, 0.12, 0.45, 0.55, 0.79`

johnedwardgreer Every data type is at least a _____

- vector
- matrix
- array
- factor

nsh87 The following R code generates the image below.

```
N <- 5000
x <- runif(N, min=-1, max=1)
y <- runif(N, min=-1, max=1)
plot(x, y, pch=16, col=ifelse(x^2 + y^2 < 1, "red", "blue"))
```



To plot just the the 3rd quadrant of the image, what modification would you make to the original code?

- `x <- runif(N, min=-1, max=0); y <- runif(N, min=-1, max=0)`
- `x <- runif(N, min=-0.5, max=0); y <- runif(N, min=-0.5, max=0)`
- `x <- runif(N, min=0, max=1); y <- runif(N, min=-1, max=0)`
- `x <- runif(N, min=0, max=1); y <- runif(N, min=0, max=1)`

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

- logical
- integer
- character
- list

xxu26 In the following code, what is the type of the variable returned?

```
y <- c(5, 6, 7, 8, NA)
is.na(y)
```

- logical
- numeric
- integer
- character

sneha-krishna The runif(n) function in R:

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

Lecture 02a

rmhorton The standard normal distribution has a mean of 0 and a standard deviation of 1, and the area under this curve over all possible x-values is one. What is the area under the curve of a normal probability distribution function with a standard deviation of 2?

- 1

- 2
- 2 pi
- 4

catterbu Which equation represents Bayes Theorem?

- $P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$
- $P(A|B) = \frac{P(B|A)\,P(B)}{P(A)}$
- $P(A|B) = \frac{P(A|B)\,P(B)}{P(A)}$
- $P(A|B) = \frac{P(A|B)\,P(A)}{P(B)}$

cpkaur The cbind() function accepts what type of inputs?

- Vectors, matrices and data frames
- Vectors and matrices
- Data frames only
- Vectors only

vchaudhuri What does the Central Limit Theorem state ?

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

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

- sensitivity = number of true positives / number with disease
- sensitivity = number of true negatives / number without disease
- sensitivity = number with disease / total population
- sensitivity = number of true positives / number of true negatives

xxu26 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"))
```

- m=-1; n=0
- m=-1.0; n=1.0
- m=-2.0; n=2.0
- m=-3.0; n=3.0

Lecture 02b

rmhorton 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?

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

rmhorton Consider the following code:

```
coinflips <- strsplit('TTTHTHTTTH','')[[1]]
flip10 <- sapply(1:10000, function(i) paste(sample(coinflips),collapse=''))
length(unique(flip10))
```

What does the code produce?

- An estimate of the number of possible permutations of the given sequence
- Ten thousand random permutations, each with 3 heads and 7 tails
- A function for flipping ten coins
- A vector of 10 'H' and 'T' characters

cpkaur Indentify the distribution type in the following code:

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

- Normal
- Poisson
- Unified constant
- Binomial

vchaudhuri True or false: "probability mass function" means the same thing as "probability density function"; what type of functions are they?

- False, the probability mass function is a discrete distribution and the probability density function is a continuous distribution
- True, they both are continuous distribution
- True, they both are discrete distribution
- False, the probability mass function is a continuous distribution and the probability density function is a discrete distribution

nsh87 "Setting the seed", e.g. `set.seed(42)`, in R...

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

lakarbatti The Poisson Distribution is a type of

- Discrete Probability Distribution
- Continuous Probability Distribution
- Cumulative distibution
- Random number generation

xxu26 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)
```

- 202
- 441
- 40
- 200

sneha-krishna A Poisson distribution is defined as:

- The probability of a given number of events occurring in a fixed interval of time and/or space if these events occur with a known average rate and independently of the time since the last event.
- The discrete probability distribution of the number of successes in a sequence of n independent yes/no experiments, each of which yields success.
- Is a very commonly occurring continuous probability distribution—a function that tells the probability that any real observation will fall between any two real limits or real numbers, as the curve approaches zero on either side.
- None of the above. These answers are terrible.

Lecture 03a

rmhorton Which command will create a multiplication table for the numbers from 1 to 10? Assume `v` is a row vector defined like this: `v <- matrix(1:10,1)`

- t(v) %*% v
- v %*% t(v)
- v^2
- t(v)^2

catterbu What kind of matrix is this?

```
##      [,1] [,2] [,3]
## [1,]    1    5    4
## [2,]    0    1    2
## [3,]    0    0    1
```

- Upper Triangular
- Lower Triangular
- Identity

cpkaur 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)
```

- `byrow = TRUE`
- `bycol = TRUE`
- `byrow = FALSE`
- No change required in the code

vchaudhuri What is the result of the following code?

```
A <- matrix(1:4, nrow=1)
A %*% t(A)
```

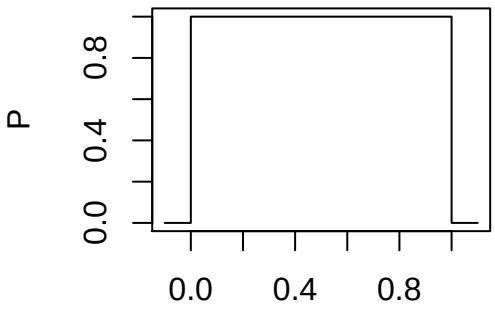
- 30, a single integer
- a 2X2 matrix of type double
- A^2, a matrix of type double
- inv(A), an integer matrix

tdenatale Which of the following cannot be done by the `diag()` function?

- diagnose errors in a linear regression
- extract a diagonal given a matrix
- construct a diagonal matrix
- replace the diagonal of a matrix

nsh87 There is more than one way to "multiply" vectors. In R, `A * B` performs elementwise multiplication. What operator would you use to get the dot product of vectors `A` and `B`?

- `A %*% B`
- `A & B`
- `A %>% B`
- `A ** B`



xxu26 What is the name of the following distribution?

- Uniform distribution
- Normal distribution
- Poisson distribution
- Binominal distribution

sneha-krishna The inverse of matrix A (of size 3x3) is called A_inverse (of size 3x3). Which is the following is FALSE?

- 'A * A_inverse' returns an identity matrix (of size 3x3)
- 'A %*% A_inverse' returns an identity matrix (of size 3x3)
- 'solve(A)' returns A_inverse
- 'solve(A_inverse)' returns A

Lecture 03b

rmhorton Given a 2 by 2 matrix `A <- matrix(c(2, 5, 3, 8), 2, byrow=TRUE)`, which command performs Gaussian elimination to put A in upper triangular form?

- A[2,] <- A[2,] - A[1,] * 3/2
- A[3,] <- A[3,] - A[2,]
- A[2,] <- A[1,] * 3/8
- A[2,] <- A[2,] + A[1,] + A[1,]/2

catterbu What is an eigenvector?

- a vector which, when multiplied by a square matrix, generates the same value as when the vector is multiplied by its eigenvalue.
- a vector that Dr. Richard Eigen designed to find a series of velocities relevant in physics.
- a vector with determinant zero.

cpkaur `A` is an n by n square matrix. Identify the correct code used for augumenting the matrix A by binding an identity matrix on the right?

- `cbind(A, diag(n))`
- `rbind(A, diag(n))`
- `outer(A, diag(n), "+")`
- `cbind(A, diag(1))`

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

- lamba is equal to 1
- lambda is infinity
- lambda is zero
- lamda doesn't exist for an identity matrix

tdenatale For a Markov Matrix, which is true?

- all entries of the matrix are nonnegative and the sum of each column vector is equal to 1
- The sum of the rows is the square root of pi

- Markov Matrices can be N X M in size, where N does not equal M
- A Markov matrix has no real eigenvectors.

johnedwardgreer Generate a sequence from 122 to 154 by intervals of 2.

- seq(122,154,2)
- seq(154,122,2)
- seq(122,2,154)
- seq(2,154,122)

nsh87 Which matrix results from the command `matrix(c(1, 2, 3, 4, 5, 6, 7, 8, 9), ncol=3)`?

-
```
1 4 7
2 5 8
3 6 9
```
-
```
1 2 3
4 5 6
7 8 9
```
-
```
9 8 7
6 5 4
3 2 1
```
-
```
9 6 3
8 5 2
7 4 1
```

sneha-krishna Which of following statements about a Markov chain is FALSE?

- a transition matrix can be used to describe the transitions of a Markov chain. the sum of each row OR the sum of each column will add up to 0.
- a Markov chain describes a random process
- transitions within a Markov chain from one state to the next depend only on the current state and not on the sequence of events that preceded it.
- each entry of the transition matrix represent a probability.

Lecture 04a

rmhorton Here is code to add normally distributed noise to an input vector:

```
addRandom <- function(i) i + rnorm(1)
y <- sapply(0:10, addRandom)
```

How would you change `addRandom` to a vectorized version that could be used like this: `y <- addRandom(0:10)`? (Choose the best solution)

- `addRandomVectorized <- function(v) v + rnorm(length(v))`
- `addRandomVectorized <- function(i) vapply(i, function(x) x+rnorm(1), 1)`
- `addRandomVectorized <- function(x) sapply(x, addRandom)`
- `addRandomVectorized <- Vectorize(addRandom)`

cpkaur What is the correct way to vectorize the following code:

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

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

tdenatale Consider this R code showing two ways of calculating the cost of daily medicine, and select the true statement.

```
price <- c( lisonopril=106/30, crestor=204.00/30,
             clorthiazide=12.10/15, fibrosol=160/30)
dosage_day <- c( lisonopril=3, crestor=0.5,
            clorthiazide=0.5, fibrosol=1)
cost_day_1  = sum(price * dosage_day)
cost_day_1a = price %*% dosage_day
```

- The Dot product of 2 vectors equals the sum of the element-wise products of the vectors
- A vector times a vector is a scalar
- A diagonal times a vector of that diagonal results in a squared value
- R is fun only for statisticians

nsh87 To plot variables `x` and `y` along the x-axis and y-axis, respectively, one could use `plot(x, y)`. What is an alternative command that generates the same plot?

- `plot(y ~ x)`
- `plot(x ~ y)`
- `plot(x %>% y)`
- `plot(y % x)`

lakarbatti To find the square of each number from 1 to N, which of the following is the fastest approach

- `x <- 1:N; y <- x^2`
- `y <- numeric(); for (i in 1:N) y[i] <- i^2`
- `y <- numeric(N); for (i in 1:N) y[i] <- i^2`
- `y <- sapply(1:N, function(i) i^2)`

sneha-krishna Given `A <- matrix(1:5)` and `B <- matrix(6:10)` and `C <- rbind(A, B)`, what does matrix `C` look like?

-
```
.    [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,]    1    2    3    4    5    6    7    8    9    10
```

-
```
.    [,1]
[1,]    1
[2,]    2
[3,]    3
[4,]    4
[5,]    5
[6,]    6
[7,]    7
[8,]    8
[9,]    9
[10,]   10
```

-
```
.    [,1] [,2]
[1,]    1    6
[2,]    2    7
[3,]    3    8
[4,]    4    9
[5,]    5   10
```

-
```
.    [,1] [,2] [,3] [,4] [,5]
[1,]    1    2    3    4    5
[2,]    6    7    8    9   10
```

Lecture 04b

rmhorton 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?

- less than 10% faster
- twice as fast
- 100 times as fast
- no faster

cpkaur Which statement is true if Q is a 2 by 2 orthogonal matrix?

- `all.equal(solve(Q), t(Q))`
- `all.equal(Q %*% diag(2), diag(2) %*% t(Q))`
- `all.equal( solve(Q) %*% Q, Q )`
- `all.equal( Q %*% diag(2), diag(2) )`

tdenatale Given the following runtime data, the goal is for a 10 times improvement in run time. Which function(s) must be improved and in what order should be chosen to reach the goal most quickly?

```
             self.time self.pct
"function_A"    278.39    84.87
"function_B"     39.32    11.99
"function_C"     10.30     3.14
------------------------------
Total           322.00   100.00
```

- Must improve both (function_A and function_B), and improve function_C only if close to target
- Always improve all functions
- Must improve function_C only
- Must improve function_A only

johnedwardgreer Create a rectangular matrix with 16 entries of numbers 1 through 16. The matrix should have 4 rows and have the numbers increasing across each row. What would this command look like in R?

- matrix(1:16, byrow=TRUE, nrow=4)
- matrix(1:16)
- matrix(1:16, byrow =FALSE)
- matrix(1:16, byrow = TRUE, nrow=2)

nsh87 What is typically the fastest way to analyze and manipulate data using R?

- With vectorized functions
- Using iteration
- With loops
- With recursion

lakarbatti Which of the following function keeps track of the function stack and tabulates how much time is spent on each function?

- RProf()
- runif()
- system.time()
- rnorm()

xxu26 The following code will produce a warning in R. Please explain why?

```
x <- 1:10
if (x > 5) {
        x <- 0
}
```

- 'x' is a vector of length 10 and 'if' can only test a single logical statement.
- use 'x' is a vector and 0 is a scalar.
- There are no elements in 'x' that are greater than 5
- The expression uses curly braces.

sneha-krishna Given x,f,y,l,z below, which of the following are equivalent?

```
x <- 1:10
f <- function(n) n^2
y <- sapply(x, f)
l <- lapply(x, f)
z <- vapply(x, f, numeric(1))
```

- y and z
- y and l
- z and l
- None

Lecture 05a

rmhorton Which of the regular expressions below matches this sentence exactly once? "The key, the whole key, and nothing but the key."

- "\.$"
- "[Tt]he\s"
- "(and|not)"
- "?key?"

cpkaur What is the correct code to find mean of the available numbers in the following vector?

```
age <- c(12,15,16, NA, 18, 30, NA)
```

- mean(age, na.rm=TRUE)
- mean(age)
- mean cannot be found
- mean (age, rm(any.na))

tdenatale 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)
)
```

- reads a csv file named ("toxic_test.csv") and puts "NA" for those entries that are marked 'UNK' or with a question mark.

- reads a csv file and halts if a missing or unknown character string is encountered
- reads a csv file and from the knitr library kables or knocksout table entries, hence the acronym kable in the knock out table
- writes a csv file to toxic_test.csv and invokes an Excel workbook session after making the dataframe

johnedwardgreer Body Mass Index is a measure of body fat based on height(m) and weight(kg). Patient heights and weights are held in vectors `HEIGHT` and `WEIGHT`, respectively. Knowing that BMI = Weight(kg)/Height$^{2(m}$2), what function in R would produce a vector of BMI values?

- BMI <- WEIGHT/HEIGHT^2
- BMI <- HEIGHT/WEIGHT^2
- It cannot be performed

xxu26 x is a data frame and z is a column of x. Which of the following commands is equivalent to with(x, f(z))?

- f(x$z)
- x$f(z)
- f(z)
- It depends.

sneha-krishna Which of the following packages in R will allow you to easily 'scrape' (ie. download, then manipulate, both html and xml?

- rvest
- ggplot
- microbenchmark
- scraping is never easy

Lecture 05b

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

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

catterbu What does the `magrittr` library do in `R`?

- It makes the operator `%>%` in an `R` script act a lot like the pipe character `|` , in the Unix terminal, though it is not the pipe character.
- It allows one to use the pipe character, `|`, in an R script in the same way that it is used in the terminal.
- It offers a more confusing alternative to the typical passing of arguments into a function, using the `$` character. For this reason, it is becoming less popular among programmers.

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

- `(as.Date("2014-10-10") - as.Date("2014-10-1")) %>% as.numeric`

- `as.Date("2014-10-10") - as.Date("2014-10-1") %>% as.numeric`
- `as.numeric %>% (as.Date("2014-10-10" - "2014-10-1" ))`
- `as.Date %>% ("2014-10-10") - as.Date %>% ("2014-10-1") >%> as.numeric`

vchaudhuri Consider the following code, then select the correct statement regarding it.

```
maxMinusMin <- function(v, ...) max(v, ...) - min(v, ...)
apply(A, 1, maxMinusMin, na.rm=TRUE)
```

- If additional parameters are given to the function, they will be passed to `max` and `min`
- Function is invalid and cannot be executed
- It's an invalid function that will need more parameters
- Typing error

johnedwardgreer A dataframe called CDC has columns representing patient name, age, height, and weight – Which R command allows the selection of all entries within the weight category?

- `CDC$weight`
- `Weight$CDC`
- `CDC(weight)`
- `Weight(CDC)`

nsh87 The magrittr operator `%>%` is a useful tool for:

- piping data into a function
- getting the remainder after division
- redirecting results to standard input
- getting the dot product of two matrices

lakarbatti What does the selectorGadget do?

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

xxu26 Simulated coin-tossing can be done using different methods. Which of the following will NOT work?

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

sneha-krishna The 'stingsAsFactors = FALSE' option is useful when reading a data file because:

- all choices are correct
- it allow us to keep character variables as they are rather than convert to factors
- the default in R is for columns with character data to be made into factors
- even if stingsAsFactors= F, it is easy to convert character data to factors using as.factor()

Lecture 06a

rmhorton How may rows are returned by the following query?

```
A <- data.frame(a=1:10)
B <- data.frame(b=5:15)
sqldf::sqldf("SELECT * FROM A JOIN B ON a==b")
```

- 6
- 10
- 8
- none

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

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

nsh87 When working with databases through R on your local computer, what is the advantage of working with SQLite instead of MySQL?

- SQLite uses a flat file, as opposed to requiring a database connection.
- There isn't an advantage because there is no way to connect to a SQLite database in R.
- SQLite is also suitable for a multi-user environment where hundreds of users connect to the database simultaneously.
- There are no packages to connect to a MySQL database in R.

lakarbatti Which statement below best describes "natural join"?

- "natural join" uses an obviously similar column for join.
- "natural join" keeps only records in first table
- "natural join" keeps only the information from second table if available
- SQL does not support natural join

xxu26 A vector x <- 1:10, which of the following choice will NOT insert 1.23 between x[7] and x[8]?

- z <- rbind(x, 1.23, after = 7)
- z <- append(x, 1.23, after = 7)
- z <- c(x[1:7], 1.23, x[8:10])
- v <- 1.23; k <- 7; i <- seq(along = x); z <- c(x[i <= k], v, x[i > k])

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

- DISTINCT
- ONLY
- DIFFERENT
- can use '*'

Lecture 06b

rmhorton Which command opens a connection to an SQLite database?

- dsets <- dbConnect(RSQLite::SQLite(), "datasets.sqlite")
- res <- dbSendQuery(dsets, "select * from iris limit 10")
- sqliteCopyDatabase(dsets, "datasets.sqlite")
- dbListTables(dsets)

catterbu What is the name of the R function that does the equialent of SQL joins?

- merge
- join
- sqlJoin
- aggregate

cpkaur What is the output of the following code?

```
x <- function(numRows=5, numCols=5, probZero=0.7, seed=NULL){
    if(!is.null(seed)) set.seed(seed)
    matrix( rnbinom(numRows * numCols, prob=probZero, size=1), nrow=numRows )
}
```

- Generates a random sparse matrix
- Generates a random vector
- Generates bionomial distribution values and stores them in x
- The code does not work

johnedwardgreer The function head() does this:

- displays the first few observations of a data frame
- creates a header in the data frame
- summarizes the data in a table

nsh87 Consider the following table called `patient`:

```
| id|name |sex |
|---|-----|----|
|  1|Alt  |F   |
|  2|Box  |M   |
|  3|Cox  |M   |
|  4|Dew  |F   |
|  5|Ely  |F   |
```

What would be the correct SQL query to get all females in this table?

- `SELECT * FROM patient WHERE sex='F'`
- `SELECT 'F' from COLUMN 'sex'`
- `SELECT sex='f' from patient`
- `SELECT sex='f' from patient where COLUMN='name'`

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

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

Lecture 07a

rmhorton Which command is equivalent to this pipeline? `myData %>% group_by(sex) %>% summarise(avg_price=mean(price))`

- `summarize(group_by(myData, sex), avg_price=mean(price))`
- `mean(avg_price, group_by(myData, sex), summarise)`
- `summarise(myData[, "sex"], avg_price=mean(price), group_by(myData))`
- `lapply(myData, function(sex){ group_by(sex); avg_price=mean(price)})`

catterbu What `SQL` command does matrix multiplication between matrices `A` and `B`?

- `SELECT A.row_num, B.col_num, SUM(A.value * B.value) AS value FROM A, B WHERE A.col_num = B.row_num GROUP BY A.row_num, B.col_num;`
- `SELECT A.row_num, B.col_num AS value FROM A, B WHERE A.col_num = B.row_num;`
- `SELECT SUM(A.value * B.value) FROM A, B WHERE A.col_num = B.row_num;`

cpkaur Which of these is not a problem with messy data

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

lakarbatti Which of the following in true with respect to relational database normalization?

- Normalization involves removing redundancies aross tables and defining keys
- Data in tables have a mean of zero and unit standard deviation
- It is a way of making sure the data is human readable
- It is basically converting binary data to text data

xxu26 What R function can be used to generate standard Normal random variables?

- rnorm
- pnorm
- dnorm
- qnorm

sneha-krishna xtab() does the following:

- all answers are correct
- crosstabulates variables
- is similar to table()
- can be used to easily generate a sparseMatrix

Lecture 07b

rmhorton 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`?

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

cpkaur Which of these lines of code cannot be used to generate a random data set?

- `qnorm(c(.05,.95))`
- `replicate( 100, runif(n=20))`
- `n <- rnorm(2500, mean=65, sd=4.58)`
- `z = rnorm(20, mean=10, sd=3)`

vchaudhuri In the following code what does the function xtabs do ?

```
T_shirts <- data.frame(
  sex=sample(c("M","F"), 100, replace=T),
  size=sample(c("L", "M", "S"), 100, replace=T)
)
table(T_shirts)
xtabs(~ sex + size, T_shirts)
```

- Crosstabulates variables with small numbers of unique values
- Introduces equally spaces tabs between columns in the output file
- Eliminates duplicate data in a table and merges data
- Breaks one data frame into separate dataframes depending on the arguments that are passed to to xtab

tdenatale What is TRUE of the following code?

```
T_shirts <- data.frame(
  sex=sample(c("M","F"), 100, replace=T),
  size=sample(c("L", "M", "S"), 100, replace=T)
)
```

- Only sometimes result in the same data, as the code does not identify a seed.
- Always result in females having more small sizes
- Always result in males having more large sizes

- Always result in the same data

nsh87 Which characteristics describe "tidy" data?

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

lakarbatti Which of the following is a common problem with messy datasets?

- One entity is stored in multiple tables
- Data is in the third norm form
- Primary and foreign keys are well defined
- Data is in human readable format