Math 154 – Computational Statistics

Fall 2021

Jo Hardin

iClicker Questions

1. The reason to take random samples is:

(a) to make cause and effect conclusions

(b) to get as many variables as possible

(c) it’s easier to collect a large dataset

(d) so that the data are a good representation of the population

(e) I have no idea why one would take a random sample

2. The reason to allocate/assign explanatory variables is:

(a) to make cause and effect conclusions

(b) to get as many variables as possible

(c) it’s easier to collect a large dataset

(d) so that the data are a good representation of the population

(e) I have no idea what you mean by “allocate/assign” (or “explanatory variable” for that matter)

3. How big is a tweet?

(a) 0.01Kb

(b) 0.1Kb

(c) 1Kb

(d) 100Kb

(e) 1000Kb = 1Mb

4. R2 measures:

(a) the proportion of variability in vote margin as explained by tweet share.

(b) the proportion of variability in tweet share as explained by vote margin.

(c) how appropriate the linear part of the linear model is.

(d) whether or not particular variables should be included in the model.

5. R / R Studio

(a) all good

(b) started, progress is slow and steady

(c) started, very stuck

(d) haven’t started yet

(e) what do you mean by “R”?

6. GitHub

(a) all good

(b) started, progress is slow and steady

(c) started, very stuck

(d) haven’t started yet

(e) what do you mean by “GitHub”?

8. HW is due on Thursday

(a) in the mailbox in Liz’s office

(b) to Professor Hardin in class

(c) emailed to Professor Hardin

(d) on GitHub + Gradescope

(e) on Sakai

9. HW should be turned in as

(a) Markdown file

(b) pdf file

(c) Markdown file and a pdf file

(d) done by hand and scanned in electronically

10. Class participation / attendance is:

(a) optional

(b) the only part of the class that matters

(c) worth some of your grade

11. Which of the following includes talking to GitHub?

(a) changing your name (updating the YAML)

(b) committing the file(s)

(c) pushing the file(s)

(d) some of the above

(e) all of the above

12. What is the error?

ralph2 <-- “Hello to you!”

(a) poor assignment operator

(b) unmatched quotes

(c) improper syntax for function argument

(d) invalid object name

(e) no mistake

13. What is the error?

3ralph <- “Hello to you!”

(a) poor assignment operator

(b) unmatched quotes

(c) improper syntax for function argument

(d) invalid object name

(e) no mistake

14. What is the error?

ralph4 <- “Hello to you!

(a) poor assignment operator

(b) unmatched quotes

(c) improper syntax for function argument

(d) invalid object name

(e) no mistake

15. What is the error?

ralph5 <- date()

(a) poor assignment operator

(b) unmatched quotes

(c) improper syntax for function argument

(d) invalid object name

(e) no mistake

16. What is the error?

ralph <- sqrt 10

(a) no assignment operator

(b) unmatched quotes

(c) improper syntax for function argument

(d) invalid object name

(e) no mistake

17. Do you keep a calendar / schedule / planner?

1. Yes
2. No

18. Do you keep a calendar / schedule / planner? If you answered “Yes” …

1. Yes, on Google Calendar
2. Yes, on Calendar for macOS
3. Yes, on Outlook for Windows
4. Yes, in some other app
5. Yes, by hand

19. The goal of making a figure is:

(a) To draw attention to your work.

(b) To facilitate comparisons.

(c) To provide as much information as possible.

20. A good reason to make a particular choice of a graph is:

1. Because the journal / field has particular expectations for how the data are presented.
2. Because some variables naturally fit better on some graphs (e.g., numbers on scatter plots).
3. Because that graphic displays the message you want as optimally as possible.

21. The Snow figure was most successful at:

1. making the data stand out
2. facilitating comparison
3. putting the work in context
4. simplifying the story

22. The Challenger figure(s) was(were) least successful at:

1. making the data stand out
2. facilitating comparison
3. putting the work in context
4. simplifying the story

23. The biggest difference between Snow and the Challenger was:

1. The AMOUNT of information portrayed.
2. One was better at displaying CAUSE.
3. One showed the relevant COMPARISON better.
4. One was more ARTISTIC.

22. Caffeine and Calories. What was the biggest concern over the average value axes?

 (a) It isn’t at the origin.

 (b) They should have used all the data possible to find averages.

 (c) There wasn’t a random sample.

 (d) There wasn’t a label explaining why the axes were where they were.

22. What are the visual cues on this plot?



(a) position

(b) length

(c) shape

(d) area/volume

(e) shade/color

23. What are the visual cues on this plot?



(a) position

(b) length

(c) shape

(d) area/volume

(e) shade/color

24. What are the visual cues on this plot?



(a) position

(b) length

(c) shape

(d) area/volume

(e) shade/color

25. Setting vs. Mapping. If I want information to be passed to all data points (not variable):

(a) map the information **inside** the aes (aesthetic) command

(b) set the information **outside** the aes (aesthetic) command

26. What is wrong with the following statement?

Result <- %>% filter(babynames,

 name== “Sydney”)

(a) should only be one =

(b) Sydney should be lower case

(c) name should not be in quotes

(d) use mutate instead of filter

(e) babynames in wrong place

27. Which is the best format for ggplot/dplyr?

A

|  |  |  |  |
| --- | --- | --- | --- |
| Year | Algeria | Brazil | Columbia |
| 2000 | 7 | 12 | 16 |
| 2001 | 9 | 14 | 18 |

B

|  |  |  |
| --- | --- | --- |
| Country | Y2000 | Y2001 |
| Algeria | 7 | 9 |
| Brazil | 12 | 14 |
| Columbia | 16 | 18 |

C

|  |  |  |
| --- | --- | --- |
| Country | Year | Value |
| Algeria | 2000 | 7 |
| Algeria | 2001 | 9 |
| Brazil | 2000 | 12 |
| Columbia | 2001 | 18 |
| Columbia | 2000 | 16 |
| Brazil | 2001 | 14 |

(a) A (b) B (c) C

28. Each of the statements except one will accomplish the same calculation. Which one does not match?

(a) babynames %>%

group\_by(year,sex) %>% summarize(totalBirths=sum(num))

(b) group\_by(babynames,year,sex) %>% summarize(totalBirths=sum(num))

(c) group\_by(babynames,year,sex) %>% summarize(totalBirths=mean(num))

(d) Tmp<-group\_by(babynames,year,sex)

summarize(Tmp,totalBirths = sum(num))

(e) summarize(group\_by(babynames,year,sex), totalBirths = sum(num))

(And what does it do?)

29.

result <- babynames %>%

Q1(name %in% c("Jane", "Mary")) %>%

# just the Janes and Marys

group\_by(Q2, Q2) %>%

summarize(total = Q3)

(a) filter

(b) arrange

(c) select

(d) mutate

(e) group\_by

30.

result <- babynames %>%

Q1(name %in% c("Jane", "Mary")) %>%

group\_by(Q2, Q2) %>%

# for each year for each name

summarize(total = Q3)

(a) (year, sex)

(b) (year, name)

(c) (year, num)

(d) (sex, name)

(e) (sex, num)

31.

result <- babynames %>%

Q1(name %in% c("Jane", "Mary")) %>%

group\_by(Q2, Q2) %>%

# number of babies (each year, each name)

summarize(total = Q3)

(a) n\_distinct(name)

(b) n\_distinct(n)

(c) sum(name)

(d) sum(n)

(e) mean(n)

(num = variable of count)

babynames %>%

filter(name %in% c("Jane","Mary")) %>%

# just the Janes and Marys

group\_by(name, year) %>%

# for each year for each name

summarize(total = sum(num))

 name year total

 (chr) (dbl) (int)

1 Jane 1880 215

2 Jane 1881 216

3 Jane 1882 254

4 Jane 1883 247

5 Jane 1884 295

.. ... ... ...

babynames %>%

 filter(name %in% c("Jane","Mary")) %>%

 group\_by(name, year) %>%

 summarize(number = sum(num))

babynames %>%

 filter(name %in% c("Jane","Mary")) %>%

 group\_by(name, year) %>%

 summarize(n\_distinct(name))

babynames %>%

 filter(name %in% c("Jane","Mary")) %>%

 group\_by(name, year) %>%

 summarize(n\_distinct(num))

babynames %>%

 filter(name %in% c("Jane","Mary")) %>%

 group\_by(name, year) %>%

 summarize(sum(name))

babynames %>%

 filter(name %in% c("Jane","Mary")) %>%

 group\_by(name, year) %>%

 summarize(mean(num))

babynames %>%

 filter(name %in% c("Jane","Mary")) %>%

 group\_by(name, year) %>%

 summarize(median(num))

32.

GDP <- GDP %>%

select(country = starts\_with("Income"),

 starts\_with("1")) %>%

 pivot\_longer(cols = Q1,

 names\_to = Q2,

 values\_to = Q3)

Q1 =

(a) gdp

(b) year

(c) gdpval

(d) country

(e) –country

33.

GDP <- GDP %>%

select(country = starts\_with("Income"),

 starts\_with("1")) %>%

 pivot\_longer(cols = Q1,

 names\_to = Q2,

 values\_to = Q3)

Q2 =

(a) gdp

(b) year

(c) gdpval

(d) country

(e) –country

34.

GDP <- GDP %>%

select(country = starts\_with("Income"),

 starts\_with("1")) %>%

 pivot\_longer(cols = Q1,

 names\_to = Q2,

 values\_to = Q3)

Q3 =

(a) gdp

(b) year

(c) gdpval

(d) country

(e) –country

36. In R the ifelse function takes the arguments:

(a) question, yes, no

(b) question, no, yes

(c) statement, yes, no

(d) statement, no, yes

(e) option1, option2, option3

37. What is the output of the following:

data <- c(1, 30, 5, NA, 6)

ifelse(data > 5, “cat”, data)

1. “cat”, 30, “cat”, “cat”, 6
2. “cat”, “30”, “cat”, “cat”, “6”
3. 1, “cat”, 5, “cat”, “cat”
4. 1, “cat”, 5, NA, “cat”
5. “1”, “cat”, “5”, NA, “cat”

38. In R, the set.seed function

(a) makes your computations go faster

(b) keeps track of your computation time

(c) provides an important parameter

(d) repeats the function

(e) makes your results reproducible

39. In Blackjack, the dealer gets another card (“hits”) if:

(a) you have at least 15

(b) you have less than 15

(c) she has less than 17

(d) she has 17 or more

(e) whenever she wants to

40. In the outcome code, what does the following line do:

any(cards == 1)

1. Checks to see if the sum is 1
2. Finds the cards that are 1
3. TRUE / FALSE depending on if a card is 1
4. TRUE / FALSE depending on if the first card is 1
5. TRUE / FALSE depending on if the last card is 1

41. In the outcome code, what is the 21.5?

1. Adding 21 to the number of cards
2. The total sum of cards is 21
3. The total sum of cards is 21 and you only have two cards
4. Sum of cards is 21.5

42. `winnings` is

1. A variable
2. A TRUE / FALSE question
3. A package
4. A built in R function
5. A user created function

43. What does the following code output?

sum( c( TRUE, TRUE, FALSE, TRUE) )

1. 1
2. 3
3. 4
4. TRUE
5. FALSE

44. A 95% confidence interval:

(a) is created in such a way that 95% of samples will produce intervals that capture the parameter.

(b) has a probability of 0.95 of capturing the parameter after the data have been collected.

(c) has a probability of 0.95 of capturing the parameter before the data have been collected.

45. The p-value

(a) is the probability H0 is true given the observed data (or more extreme).

(b) is the probability of the observed data (or more extreme) given H0 is true.

46. We typically compare means (across two groups) instead of medians because

(a) we don’t know the SE of the difference of medians

(b) means are inherently more interesting than medians

(c) permutation tests don’t work with medians

(d) the Central Limit Theorem doesn’t apply for medians.

47. What are the technical assumptions for a t-test?

1. none
2. normal data
3. $n \geq 30$
4. random sampling / random allocation for appropriate conclusions

48. What are the technical conditions for permutation tests?

(a) none

(b) normal data

(c) $n \geq 30$

(d) random sampling / random allocation for appropriate conclusions

49. Follow up to permutation test: the assumptions change based on whether the statistic used is the mean, median, proportion, etc.

1. TRUE
2. FALSE

50. Why do we care about the distribution of the test statistic?

1. Better estimator
2. So we can find rejection region
3. So we can control power
4. Because we love the CLT

51. Given a statistic T = r(X), how do we find a (reasonable) test?

(a) Maximize power

(b) Minimize type I error

(c) Control type I error

(d) Minimize type II error

(e) Control type II error

52. Type I error is

1. We give him a raise when he deserves it.
2. We don’t give him a raise when he deserves it.
3. We give him a raise when he doesn’t deserve it.
4. We don’t give him a raise when he doesn’t deserve it.

53. Type II error is

1. We give him a raise when he deserves it.
2. We don’t give him a raise when he deserves it.
3. We give him a raise when he doesn’t deserve it.
4. We don’t give him a raise when he doesn’t deserve it.

54. Power is the probability that:

1. We give him a raise when he deserves it.
2. We don’t give him a raise when he deserves it.
3. We give him a raise when he doesn’t deserve it.
4. We don’t give him a raise when he doesn’t deserve it.

55. Why don’t we always reject H0?

(a) type I error too high

(b) type II error too high

(c) level of sig too high

(d) power too high

54. The player is more worried about

1. A type I error
2. A type II error

56. The coach is more worried about

1. A type I error
2. A type II error

57. Increasing your sample size

1. Increases your power
2. Decreases your power

58. Making your significance level more stringent (α smaller)

1. Increases your power
2. Decreases your power

59. A more extreme alternative

1. Increases your power
2. Decreases your power

60. What is the primary reason to use a permutation test (instead of a test built on calculus)

(a) more power

(b) lower type I error

(c) more resistant to outliers

(d) can be done on statistics with unknown sampling distributions

61. What is the primary reason to bootstrap a CI (instead of creating a CI from calculus)?

(a) larger coverage probabilities

(b) narrower intervals

(c) more resistant to outliers

(d) can be done on statistics with unknown sampling distributions

62. You have a sample of size n = 50. You sample with replacement 1000 times to get 1000 bootstrap samples.

 What is the sample size of each bootstrap sample?

 (a) 50

 (b) 1000

63. You have a sample of size n = 50. You sample with replacement 1000 times to get 1000 bootstrap samples.

 How many bootstrap statistics will you have?

 (a) 50

 (b)1000

64. The bootstrap distribution is centered around the

(a) population parameter

(b) sample statistic

(c) bootstrap statistic

(d) bootstrap parameter

65.



95% CI for the difference in proportions:

1. (0.39, 0.43)
2. (0.37, 0.45)
3. (0.77, 0.81)
4. (0.75, 0.85)

66. Suppose a 95% bootstrap CI for the difference in means was (3,9), would you reject H0?

(uh…. What is the null hypothesis here???)

(a) yes

(b) no

(c) not enough information to know

67. Given the situation where Ha is TRUE. Consider 100 CIs (for true difference in means), the power of the test can be approximated by:

Each of the 100 CIs is created using a different dataset.

(a) The proportion that contain the true difference in means.

(b) The proportion that do not contain the true difference in means.

(c) The proportion that contain zero.

(d) The proportion that do not contain zero.

68 – 74 we didn’t cover

75. Small k in k-NN will help reduce the risk of overfitting.

(a) True

(b) False

76. The training error for 1-NN classifier is zero.

(a) TRUE

(b) FALSE

77. Generally, the k-NN algorithm can take any distance measure.

(a) True

(b) False

78. In R, the `kknn` method can use any distance measure.

(a) True

(b) False

(note: it uses Minkowski(p) distance which is Euclidean for p=2.)

79. the “k” in k-NN refers to

(a) k groups

(b) k partitions

(c) k neighbors

80. the “V” in V-fold CV refers to

(a) V groups

(b) V partitions

(c) V neighbors

81. All of the following are **TRUE** for the use of CART except for:

(a) Can deal with missing data

(b) Require the assumptions of statistical models

(c) Variable selection is automatic

(d) Produce rules that are easy to interpret and implement

82. Simplifying the decision tree by pruning peripheral branches will cause over fitting.

(a) TRUE

(b) FALSE

83. All are true with regards to PRUNING except:

(a) Multiple (sequential) trees are possible to create by pruning

(b) CART lets tree grow to full extent, then prunes it back

(c) Pruning generates successively smaller trees by pruning leaves

(d) Pruning is only beneficial when purity improvement is statistically significant

84. Regression trees are invariant to monotonic transformations of:

(a) the explanatory (predictor) variables

(b) the response variable

(c) both types of variables

(d) none of the variables

85. CART suffers from

(a) high variance

(b) high bias

86. bagging uses bootstrapping on:

(a) the variables

(b) the observations

(c) both

(d) neither

87. oob samples

(a) are in the test data

(b) are in the training data and provide independent predictions

(c) are in the training data but do not provide independent predictions

88. oob samples are great because

(a) oob is “boo” spelled backwards

(b) oob samples allow for independent predictions

(c) oob samples allow for more predictions than a “test group”

(d) oob data frame is always bigger than the test sample data frame

(e) some of the above

89. bagging is random forests with:

(a) m = # predictor variables

(b) all the observations

(c) the most important predictor variables isolated

(d) cross validation to choose m

80 training observations, 20 test observations

to get the test MSE

1. 20 predictions from all trees
2. 20 predictions from oob trees
3. 80 predictions form all trees
4. 80 predictions from oob trees

90. With random forests, the value for m is chosen

(a) using OOB error rate

(b) as p/3

(c) as sqrt(p)

(d) using cross validation

91. A tuning parameter:

(a) makes the model fit the training data as well as possible.

(b) makes the model fit the test data as well as possible.

(c) allows for a good model that does not overfit the data.

92. With binary response and X1 and X2 continuous, kNN (k=1) creates a linear decision boundary.

(a) TRUE

(b) FALSE

93. With binary response and X1 and X2 continuous, a classification tree with one split creates a linear decision boundary.

(a) TRUE

(b) FALSE

94. With binary response and X1 and X2 continuous, a classification tree with one split creates the best linear decision boundary.

(a) TRUE

(b) FALSE

95. If the data are linearly separable, there exists a “widest street”.

(a) yes

(b) no

(c) up to a constant

(d) with the alpha values “tuned” appropriately

96. In the case of linearly separable data, the SVM:

(a) has a tuning parameter of α

(b) has a tuning parameter of dimension

(c) has no tuning parameters

97. If the data have a complex boundary, the value of gamma in RBF kernel should be:

1. Big
2. Small

98. If the data have a simple boundary, the value of gamma in RBF kernel should be:

1. Big
2. Small

99. If I am using all features of my dataset and I achieve 100% accuracy on my training set, but ~70% on validation set, what should I look out for?

(a) Underfitting

(b) Nothing, the model is perfect

(c) Overfitting

100. For a large value of C, the model is expected to

(a) overfit the training data more as compared to a small C

(b) overfit the training data less as compared to a small C

(c) not related to overfitting the data

101. Suppose you have trained an SVM with linear decision boundary after training SVM, you correctly infer that your SVM model is under fitting.

Which of the following option would you more likely to consider iterating SVM next time?

A) You want to increase your data points

B) You want to decrease your data points

C) You will try to calculate more variables

D) You will try to reduce the features

102. Suppose you have trained an SVM with linear decision boundary after training SVM, you correctly infer that your SVM model is under fitting.

Suppose you gave the correct answer in previous question. What do you think that is actually happening?

1. We are lowering the bias

2. We are lowering the variance

3. We are increasing the bias

4. We are increasing the variance

A) 1 and 2

B) 2 and 3

C) 1 and 4

D) 2 and 4

103. Suppose you are using SVM with linear kernel of polynomial degree 2, Now think that you have applied this on data and found that it perfectly fit the data that means, Training and testing accuracy is 100%.

You increase the complexity(degree of polynomial of this kernel). What will happen?

A) Increasing the complexity will overfit the data

B) Increasing the complexity will underfit the data

C) Nothing will happen since your model was already 100% accurate

D) None of these

104. Suppose you are using SVM with linear kernel of polynomial degree 2, Now think that you have applied this on data and found that it perfectly fit the data that means, Training and testing accuracy is 100%.

In the previous question after increasing the complexity you found that training accuracy was still 100%. According to you what is the reason behind that?

1. Since data are fixed and we are fitting more polynomial terms, the algorithm starts memorizing everything in the data

2. Since data are fixed, SVM doesn’t need to search in high dimensional space

A) 1

B) 2

C) 1 and 2

D) None of these

105. The cost parameter in the SVM means:

A) The number of cross-validations to be made

B) The kernel to be used

C) The tradeoff between misclassification and simplicity of the model

D) None of the above

106. Suppose you have trained an SVM classifier with a Gaussian kernel, and it learned the following decision boundary on the training set:



You suspect that the SVM is underfitting your dataset. Should you try increasing or decreasing C? Increasing or decreasing gamma?

1. decreasing C and/or decrease gamma
2. decreasing C and/or increase gamma
3. increasing C and/or decrease gamma
4. increasing C and/or increase gamma

107. Suppose you have trained an SVM classifier with a Gaussian kernel, and it learned the following decision boundary on the training set:



When you measure the SVM’s performance on a cross validation set, it does poorly. Should you try increasing or decreasing C? Increasing or decreasing gamma?

1. decreasing C and/or decrease gamma
2. decreasing C and/or increase gamma
3. increasing C and/or decrease gamma
4. increasing C and/or increase gamma

108. Cross validation will guarantee that the model does not overfit.

(a) TRUE

(b) FALSE

109. The biggest problem with missing data is the resulting small sample size.

(a) True

(b) False

110. Which statement is not true about cluster analysis?

(a) Objects in each cluster tend to be similar to each other and dissimilar to objects in the other clusters.

(b) Cluster analysis is a type of unsupervised learning.

(c) Groups or clusters are suggested by the data, not defined a priori.

(d) Cluster analysis is a technique for analyzing data when the response variable is categorical and the predictor variables are continuous in nature.

111. A \_\_\_\_\_ or tree graph is a graphical device for displaying clustering results. Vertical lines represent clusters that are joined together. The position of the line on the scale indicates the distances at which clusters were joined.

(a) dendrogram

(b) scatterplot

(c) scree plot

(d) segment plot

112. \_\_\_\_\_ is a clustering procedure characterized by the development of a tree-like structure.

a. Partitioning clustering

b. Hierarchical clustering

c. Divisive clustering

d. Agglomerative clustering

113. \_\_\_\_\_ is a clustering procedure where all objects start out in one giant cluster.

Clusters are formed by dividing this cluster into smaller and smaller clusters.

a. Non-hierarchical clustering

b. Hierarchical clustering

c. Divisive clustering

d. Agglomerative clustering

114. The \_\_\_\_\_ method uses information on all pairs of distances, not merely the minimum or maximum distances.

a. single linkage

b. medium linkage

c. complete linkage

d. average linkage

115. Hierarchical clustering is deterministic, but k-means clustering is not.

(a) TRUE

(b) FALSE

116. k-means is a clustering procedure characterized referred to as \_\_\_\_\_\_\_\_.

a. Partitioning clustering

b. Hierarchical clustering

c. Divisive clustering

d. Agglomerative clustering

117. One method of assessing reliability and validity of clustering is to use different methods of clustering and compare the results.

(a) TRUE

(b) FALSE

118. The choice of k, the number of clusters to partition a set of data into,...

(a) is a personal choice that shouldn't be discussed in public

(b) depends on why you are clustering the data

(c) should always be as large as your computer system can handle

(d) has maximum 10

119. Which of the following is required by K-means clustering?

(a) defined distance metric

(b) number of clusters

(c) initial guess as to cluster centroids

(d) all of the above

(e) some of the above

120. For which of the following tasks might clustering be a suitable approach?

(a) Given sales data from many products in a supermarket, estimate future sales for each of these products.

(b) Given a database of information about your users, automatically group them into different market segments.

(c) From the user's usage patterns on a website, identify different user groups.

(d) Given historical weather records, predict if tomorrow's weather will be sunny or rainy.

121. K-means is an iterative algorithm, and two of the following steps are repeatedly carried out. Which two?

1. Assign each point to its nearest cluster

2. Test on the cross-validation set

3. Update the cluster centroids based the current assignment

4. Using the elbow method to choose K

1. 1 & 2
2. 1 & 3
3. 1 & 4
4. 2 & 3
5. 2 & 4
6. 3 & 4