Math 154 – Computational Statistics

Fall 2017

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 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” (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”?

7. Professor Hardin’s office hours are:

(a) Tues & Thurs mornings

(b) Tues & Thurs afternoons

(c) Mon & Wed 1:15-4

(d) Tues morning and Thurs afternoon

(e) Tues afternoon and Thurs morning

8. HW is due

(a) in the mailbox in Kathy’s office

(b) to Professor Hardin in class

(c) emailed to Professor Hardin

(d) on GitHub

(e) on Sakai

9. HW should be turned in as

(a) Markdown or Sweave file

(b) pdf file

(c) Markdown or Sweave file and a pdf file

(d) done by hand and scanned in electronically

(Also, put your name in the HW file, but keep the prefix as Ma154-HWX-…)

10. Participation / attendance is:

(a) optional

(b) the only part of the class that matters

(c) worth some of your grade

11. 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

12. 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

13. 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

14. 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

15. 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

16. 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.

17. 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.

18. What are the visual cues on this plot?



(a) position

(b) length

(c) shape

(d) area/volume

(e) shade/color

19. What are the visual cues on this plot?



(a) position

(b) length

(c) shape

(d) area/volume

(e) shade/color

20. What are the visual cues on this plot?



(a) position

(b) length

(c) shape

(d) area/volume

(e) shade/color

21. Setting vs. Mapping (again). 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

16. What is wrong with the following statement?

Result <- %>% filter(babynames,

 name== “Prince”)

(a) should only be one =

(b) Prince should be lower case

(c) name should not be in quotes

(d) use mutate instead of filter

(e) babynames in wrong place

17. 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

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

(a) babynames %>% group\_by(year,sex) %>% summarise(totalBirths=sum(num))

(b) group\_by(babynames,year,sex) %>% summarise(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?)

19. Result <- babynames %>%

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

# just the Janes and Marys

group\_by(Q2, Q2) %>%

summarise(count = Q3)

(a) filter

(b) arrange

(c) select

(d) mutate

(e) group\_by

20. Result <- babynames %>%

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

group\_by(Q2, Q2) %>%

# for each year for each name

summarise(count = Q3)

(a) (year, sex)

(b) (year, name)

(c) (year, n)

(d) (sex, name)

(e) (sex, n)

21. Result <- babynames %>%

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

group\_by(Q2, Q2) %>%

# number of babies (each year, each name)

summarise(count = Q3)

(a) n\_distinct(name)

(b) n\_distinct(num)

(c) sum(name)

(d) sum(num)

(e) mean(num)

num=n

(what is n???)

babynames %>%

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

# just the Janes and Marys

group\_by(name, year) %>%

# for each year for each name

summarise(count = sum(n))

 name year count

 (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) %>%

 summarise(count = sum(n))

babynames %>%

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

 group\_by(name, year) %>%

 summarise(n\_distinct(name))

babynames %>%

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

 group\_by(name, year) %>%

 summarise(n\_distinct(n))

babynames %>%

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

 group\_by(name, year) %>%

 summarise(sum(name))

babynames %>%

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

 group\_by(name, year) %>%

 summarise(sum(n))

babynames %>%

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

 group\_by(name, year) %>%

 summarise(median(n))

22. gdp <- gdp %>%

select(country = starts\_with("Income"), starts\_with("1")) %>%

gather(Q1, Q2, Q3)

Q1:

(a) gdp

(b) year

(c) gdpval

(d) country

(e) –country

23. gdp <- gdp %>%

select(country = starts\_with("Income"), starts\_with("1")) %>%

gather(Q1, Q2, Q3)

Q2:

(a) gdp

(b) year

(c) gdpval

(d) country

(e) –country

24. gdp <- gdp %>%

select(country = starts\_with("Income"), starts\_with("1")) %>%

gather(Q1, Q2, Q3)

Q3

(a) gdp

(b) year

(c) gdpval

(d) country

(e) –country

25. The last problem on HW3 asks about the significance of weekday for average visibility:

> summary(aov(visib~dayofweek, data=weather4))

 Df Sum Sq Mean Sq F value Pr(>F)

dayofweek 6 149 24.878 5.691 6.47e-06

Residuals 8711 38079 4.371

 dayofweek mean(visib)

1 Sun 9.261475

2 Mon 8.989968

3 Tues 9.222133

4 Wed 9.102059

5 Thurs 9.380714

6 Fri 9.236156

7 Sat 9.379123

(a) visib definitely different

(b) visib not different b/c pvalue

(c) visib not different b/c sampling

(d) visib unknown b/c pvalue

(e) visib unknown b/c sampling

26. 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 more than 17

(e) whenever she wants to

27. 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

28. 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

29. 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.

30. 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.

31. We typically compare means 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.

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

(a) none

(b) normal data

(c) $n \geq 30$

(d) random sampling / random allocation for appropriate conclusions

33. 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

Follow up: do the assumptions change based on whether the statistic is the mean, median, proportion, etc.?

34. 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

35. 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

36. 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.

37. 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.

38. 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.

39. 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

40. The player is more worried about

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

41. The coach is more worried about

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

42. Increasing your sample size

1. Increases your power
2. Decreases your power

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

1. Increases your power
2. Decreases your power

44. A more extreme alternative

1. Increases your power
2. Decreases your power

45. 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

46. 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

47. The best way to access the stuff on the Course Materials repo is:

a. Use your browser to open the GitHub site and look at the solutions and datasets using the browser.

b. Clone the Course Materials repo onto your computer locally and access the files locally via your computer (and pull every time the repo gets updated).

c. What is the Course Materials repo?

47. 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

48. 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

49. The bootstrap distribution is centered around the

(a) population parameter

(b) sample statistic

(c) bootstrap statistic

(d) bootstrap parameter

50.



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)

51. 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

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

(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.

53. 

(a) As little data as possible?

(b) Alphabetical sorting?

(c) Wrong / changing scales?

(d) One-dim info in two- or three-dim?

RA Fisher (1929)

“… An observation is judged significant, if it would rarely have been produced, in the absence of a real cause of the kind we are seeking. It is a common practice to judge a result significant, if it is of such a magnitude that it would have been produced by chance not more frequently than once in twenty trials. This is an arbitrary, but convenient, level of significance for the practical investigator, but it does not mean that he allows himself to be deceived once in every twenty experiments. The test of significance only tells him what to ignore, namely all experiments in which significant results are not obtained. He should only claim that a phenomenon is experimentally demonstrable when he knows how to design an experiment so that it will rarely fail to give a significant result. Consequently, isolated significant results which he does not know how to reproduce are left in suspense pending further investigation.”

George Cobb (2014)

Q: Why do so many colleges and grad schools teach p = .05?

A: Because that's still what the scientific community and journal editors use.

Q: Why do so many people still use p = 0.05?

A: Because that's what they were taught in college or grad school.

Basic and Applied Social Psychology (2015)

With the banning of the NHSTP (null hypothesis significance testing procedures) from BASP, what are the implications for authors?

**Question 3.** *Are any inferential statistical procedures required?*

**Answer to Question 3.** No, because the state of the art remains uncertain. However, BASP will require strong descriptive statistics, including effect sizes. We also encourage the presentation of frequency or distributional data when this is feasible. Finally, we encourage the use of larger sample sizes than is typical in much psychology research, because as the sample size increases, descriptive statistics become increasingly stable and sampling error is less of a problem. However, we will stop short of requiring particular sample sizes, because it is possible to imagine circumstances where more typical sample sizes might be justifiable.

American Statistical Association’s Statement on p-values (2016)

<http://amstat.tandfonline.com/doi/pdf/10.1080/00031305.2016.1154108>

**Statisticians issue warning over misuse of *P* values (Nature, March 7, 2016)**

http://www.nature.com/news/statisticians-issue-warning-over-misuse-of-p-values-1.19503

54. P-values can indicate how incompatible the data are with a specified statistical model.

(a) TRUE

(b) FALSE

55. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.

(a) TRUE

(b) FALSE

56. Scientific conclusions and business or policy decisions should not be based only on whether a p- value passes a specific threshold.

(a) TRUE

(b) FALSE

57. Proper inference requires full reporting and transparency.

(a) TRUE

(b) FALSE

58. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.

(a) TRUE

(b) FALSE

59. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

(a) TRUE

(b) FALSE

Dance of the p-values:

<https://www.youtube.com/watch?v=5OL1RqHrZQ8>

Your own p-value:

<https://www.openintro.org/stat/why05.php?stat_book=os>

60. Small k in kNN will help reduce the risk of overfitting.

(a) True

(b) False

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

(a) TRUE

(b) FALSE

62. Generally, the kNN algorithm can take any distance measure.

(a) True

(b) False

63. In R, the knn function can use any distance measure.

(a) True

(b) False

64. the “k” in kNN refers to

(a) k groups

(b) k partitions

(c) k neighbors

65. the “k” in k-fold CV refers to

(a) k groups

(b) k partitions

(c) k neighbors

66. All of the following are true for the use of a regression tree 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

66. Simplifying the decision tree by pruning peripheral branches will cause overfitting.

(a) TRUE

(b) FALSE

67. 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

68. 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

69. CART suffers from

(a) high variance

(b) high bias

70. bagging uses bootstrapping on:

(a) the variables

(b) the observations

(c) both

(d) neither

71. 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

72. 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

73. 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

74. 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

75. 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.

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

(a) TRUE

(b) FALSE

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

(a) TRUE

(b) FALSE

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

(a) TRUE

(b) FALSE

79. 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

80. 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

81. If the data are not linearly separable, there exists a “widest street”.

(a) yes

(b) no

(c) up to a constant

(d) with the alpha values “tuned” appropriately

82. For larger values of C, the model is expected to

(a) overfit the training data more

(b) overfit the training data less

(c) not related to overfitting the data

83. What is the behavior of the width of the margin (2/||w||) as C goes to zero?

(a) behaves like hard margin (without C)

(b) width goes to zero

(c) width goes to infinity

(d) width goes to outside edges of data

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

(a) TRUE

(b) FALSE

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

(a) True

(b) False

86. 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.

87. 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

88. \_\_\_\_\_ 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

89. \_\_\_\_\_ 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

90. 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

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

(a) TRUE

(b) FALSE

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

a. Partitioning clustering

b. Hierarchical clustering

c. Divisive clustering

d. Agglomerative clustering

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

(a) TRUE

(b) FALSE