DATA 101, Practice Final
Review session (May 1, 2015)

Part I. Multiple choice

1. Which of these functions is **not** related to plotting?

(a) `abline()`
(b) `dev.off()`
(c) `histogram()`
(d) `curve()`
(e) `lda()`

2. What would R say, if `sum(c(T,F,T))` is entered to the console?

(a) `[1] 0`
(b) `[1] 1`
(c) `[1] 2`
(d) `[1] 3`
(e) Something else.

3. Which of the following is a correct way to select 2 rows from the data frame `suarez`?

(a) `suarez(1:2)`
(b) `suarez[,c(1,2)]`
(c) `suarez[2,2]`
(d) `suarez[1:2,]`
(e) `suarez[2,]`

4. Suppose `w <- c(1,0,-3,-2,10)` is entered into the R console. What would R say if you enter `w[w == 0]`?

(a) `[1] FALSE TRUE FALSE FALSE FALSE`
(b) `[1] 1 -3 -2 10`
(c) `[1] 0`
(d) `[1] 1 0 -3 -2 10`
(e) Something else.

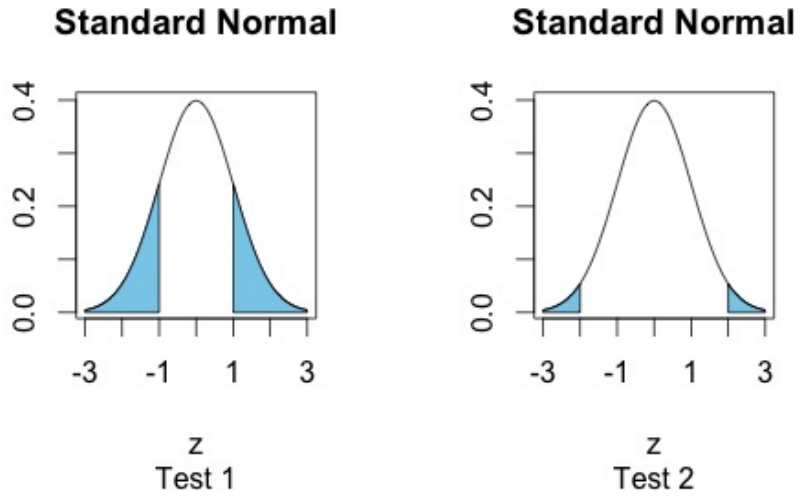
5. True or false: Permutation tests can only be used for continuous variables.

(a) True.
(b) False.

6. True or false: t-tests are appropriate for hypothesis testing problems where the standard deviation is *known*; z-tests are appropriate for hypothesis testing problems where the standard deviation is *unknown*.

(a) True.
(b) False.

7. Suppose that z-values are obtained from two hypothesis tests (Test 1 and Test 2), which are then used to compute p-values. The z-values and p-values are represented in the plots below.



Which of the following statements best describes the results of these tests?

(a) The p-value from Test 1 is smaller than that from Test 2. Thus, the null hypothesis from Test 1 is more likely to be rejected than the null hypothesis from Test 2.
(b) The p-value from Test 1 is smaller than that from Test 2. Thus, the null hypothesis from Test 2 is more likely to be rejected than the null hypothesis from Test 1.
(c) The p-value from Test 2 is smaller than that from Test 1. Thus, the null hypothesis from Test 1 is more likely to be rejected that the null hypothesis from Test 2.
(d) The p-value from Test 2 is smaller than that from Test 1. Thus, the null hypothesis from Test 2 is more likely to be rejected that the null hypothesis from Test 1.

8. Linear regression is a popular method for prediction problems, which easily implemented in R using the function `lm()`. Linear regression is most appropriate for problems where the goal is to predict a:

(a) Continuous outcome.
(b) Categorical outcome.
(c) Ordinal outcome.
(d) Multivariate outcome.

9. Which of the following best describes the relationship between decision trees and random forest?

(a) Random forest combines features of decision trees and linear regression.
(b) Decision trees are a more general version of random forest.
(c) Random forest is for random data; a decision tree is for non-random data.
(d) Random forest is a method for creating and combining many different decision trees in order to improve predictions.
(e) Random forests are endangered in the Amazon.

10. Decision trees and linear regression are different statistical methods that are easy to implement in R and can be used for prediction problems. Which of the following statements is FALSE?

(a) Decision trees can be used for prediction problems with either categorical or continuous outcomes.
(b) Linear regression tends to perform better when the outcome and predictors have a linear relationship.
(c) Decision trees and linear regression can be combined to create improved prediction rules.
(d) If used by themselves, linear regression almost always performs worse than decision trees.
(e) Decision trees are especially useful for detecting nonlinear patterns in the data.

11. Which of the following statements about *cross-validation* is FALSE?

(a) Cross-validation is useful for estimating the error-rate of a prediction rule, before you have access to the actual test data.
(b) Cross-validation is designed to help minimize problems with overfitting.
(c) When comparing many prediction rules, the rule with the lowest cross-validation error-rate always has the lowest actual error-rate on the test data.
(d) If you perform cross-validation several times, using the same data and prediction rules, you might get different results.

12. Suppose that you conduct 1,000 hypothesis tests and obtain a single p-value from each test. Suppose that *all* of the null hypotheses are true. About how many of the p-values would you expect to be less that 0.05?

(a) 100
(b) 50
(c) 25
(d) 5
(e) 0

13. Suppose that you conduct 10,000 hypothesis tests and obtain a single p-value from each test. The Bonferroni method is a simple method to control for multiple comparisons. Which of the following statements best describes how to implement the Bonferroni method in this setting?

(a) Divide each p-value by 10,000.
(b) Multiply each p-value by 10,000.
(c) Multiply each p-value by 500.
(d) Only reject tests with p-value less that 0.001.
(e) Reject tests with the 500 smallest p-values.

14. Consider the following modified version of the Monty Hall problem:

> Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1. The host, who knows what's behind the doors, opens another door, say No. 3 – door No. 3 might have a goat behind it *or* it might have the car behind it. If door No. 3 has the car behind it, then you lose. If door No. 3 has the goat behind it, then you have the opportunity to switch doors, i.e. change your choice to door No. 2. Is it to your advantage to switch your choice, if the host reveals a goat behind door No. 3?

(a) Yes.
(b) No.
(c) It depends.

15. Consider the data frame in R `dd`:

```
> dd
  u v   w  y
1 5 0 100  0
2 4 0   0 10
3 3 0   0 20
4 2 0   0 30
5 1 0   0 40
6 0 1   0 50
```

If you run a linear regression model in R with the command `lm(y~.,data=dd)`, what variable will be used as the outcome (to be predicted) and what variables will be used as the predictors (to predict the outcome)?

(a) Outcome, `y`; predictors, `u`, `v`, `w`, `y`.
(b) Outcome, `y`; predictors, `u`, `v`, `w`.
(c) Outcome, `y`; predictors, none.
(d) Outcome `dd`; predictors, `u`, `v`, `w`, `y`.
(e) Outcome `u`; predictors, `v`, `w`, `y`.

16. The R function `randomForest()` is used to create a random forest for prediction problems. One option for the `randomForest()` function is `ntree`. What does this option control?

(a) The size of each tree in the random forest.
(b) The number of trees in the random forest.
(c) The number of possible values that the random forest predictions may take.
(d) The minimum number of times to run cross-validation for determining the best random forest.
(e) The minimum number of leaf (terminal) nodes in the random forest.

17. Consider the Marriage data, with variables STATUS, GAGE, BAGE, BP, and SP. Suppose we wish to predict STATUS using the other variables in the dataset, and use the rpart() function in R to create a prediction rule. Suppose that the R output from rpart() is as follows:

```
1) root 986 464 Married (0.47058824 0.52941176)
  2) GAGE< 58.5 797 389 Married (0.48808030 0.51191970)
    4) BAGE>=52.5 130  35 Divorced (0.73076923 0.26923077)
      8) GAGE< 48.5 79   5 Divorced (0.93670886 0.06329114) *
      9) GAGE>=48.5 51  21 Married (0.41176471 0.58823529) *
    5) BAGE< 52.5 667 294 Married (0.44077961 0.55922039)
     10) GAGE>=52.5 75  25 Divorced (0.66666667 0.33333333)
       20) BAGE< 44 38   3 Divorced (0.92105263 0.07894737) *
       21) BAGE>=44 37  15 Married (0.40540541 0.59459459) *
     11) GAGE< 52.5 592 244 Married (0.41216216 0.58783784)
       22) GAGE< 18.5 9   2 Divorced (0.77777778 0.22222222) *
       23) GAGE>=18.5 583 237 Married (0.40651801 0.59348199)
         46) BAGE>=49.5 32  14 Divorced (0.56250000 0.43750000)
           92) GAGE< 42 14   1 Divorced (0.92857143 0.07142857) *
           93) GAGE>=42 18   5 Married (0.27777778 0.72222222) *
         47) BAGE< 49.5 551 219 Married (0.39745917 0.60254083)
           94) GAGE>=49.5 29  13 Divorced (0.55172414 0.44827586)
            188) BAGE< 43.5 12   1 Divorced (0.91666667 0.08333333) *
            189) BAGE>=43.5 17   5 Married (0.29411765 0.70588235) *
           95) GAGE< 49.5 522 203 Married (0.38888889 0.61111111)
            190) GAGE< 48.5 494 197 Married (0.39878543 0.60121457)
              380) GAGE>=33.5 269 117 Married (0.43494424 0.56505576)
                760) BAGE< 31.5 64  17 Divorced (0.73437500 0.26562500) *
                761) BAGE>=31.5 205  70 Married (0.34146341 0.65853659)
                 1522) GP=A 95  41 Married (0.43157895 0.56842105)
                   3044) BP=A 50  18 Divorced (0.64000000 0.36000000) *
                   3045) BP=B 45   9 Married (0.20000000 0.80000000) *
                 1523) GP=B 110  29 Married (0.26363636 0.73636364) *
              381) GAGE< 33.5 225  80 Married (0.35555556 0.64444444)
                762) BAGE>=34.5 45  11 Divorced (0.75555556 0.24444444) *
                763) BAGE< 34.5 180  46 Married (0.25555556 0.74444444) *
            191) GAGE>=48.5 28   6 Married (0.21428571 0.78571429) *
  3) GAGE>=58.5 189  75 Married (0.39682540 0.60317460)
    6) BAGE< 51.5 41   8 Divorced (0.80487805 0.19512195) *
    7) BAGE>=51.5 148  42 Married (0.28378378 0.71621622) *
```

What would the predicted STATUS be for a couple with BAGE = 45, GAGE = 52, BP = A, and BP = B?

(a) Married
(b) Divorced

18. Consider the Professor Moody data, with variables SCORE, GRADE, ATTENDANCE, ALERT, ASKQUESTIONS. Suppose we wish to predict GRADE using the other variables in the dataset, and use the rpart() function in R to create a prediction rule. Suppose that the R output from rpart() is as follows:

```
 1) root 215 112 F (0.17209302 0.20930233 0.13953488 0.47906977)
   2) SCORE>=49 111  66 B (0.33333333 0.40540541 0.24324324 0.01801802)
     4) SCORE>=63.5 80  40 B (0.46250000 0.50000000 0.03750000 0.00000000)
       8) SCORE>=88 19   1 A (0.94736842 0.05263158 0.00000000 0.00000000) *
       9) SCORE< 88 61  22 B (0.31147541 0.63934426 0.04918033 0.00000000)
        18) ASKQUESTIONS=Frequently 18   5 A (0.72222222 0.27777778 0.00000000 0.00000000) *
        19) ASKQUESTIONS=Never,Rarely 43   9 B (0.13953488 0.79069767 0.06976744 0.00000000) *
     5) SCORE< 63.5 31   7 C (0.00000000 0.16129032 0.77419355 0.06451613) *
   3) SCORE< 49 104   3 F (0.00000000 0.00000000 0.02884615 0.97115385) *
```

What would the predicted GRADE be for a student with SCORE=75, ATTENDANCE=Perfect, ALERT=Texting, and ASKQUESTIONS=Never?

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

19. Consider a version of the Marriage dataset, with variables STATUS, AGEDIFF, and PERSONALITY. Suppose we wish to predict STATUS using the other variables, and use the rpart() function in R to create a prediction rule. Suppose that the R output from rpart() is as follows:

```
 1) root 986 464 Married (0.4705882 0.5294118)
   2) AGEDIFF< -10.5 142  15 Divorced (0.8943662 0.1056338) *
   3) AGEDIFF>=-10.5 844 337 Married (0.3992891 0.6007109)
     6) AGEDIFF>=10.5 127  17 Divorced (0.8661417 0.1338583) *
     7) AGEDIFF< 10.5 717 227 Married (0.3165969 0.6834031)
      14) PERSONALITY=match 365 175 Divorced (0.5205479 0.4794521) *
      15) PERSONALITY=diff 352  37 Married (0.1051136 0.8948864) *
```

How many leaf nodes does this tree have?

(a) 2
(b) 3
(c) 4
(d) 5
(e) 6

20. Consider the Loan dataset, with variables LOAN, ZODIAC, FIVEELEMENTS, and MOON. Suppose we wish to predict LOAN using the other variables, with the randomForest() function in R. Additionally suppose that the training data is stored in the data frame train and that the testing data is stored in the data frame test. In the following R code and output, randomForest() is used to generate predictions for the test data *twice*, and the test error rate is computed each time:

```
> loan.forest <- randomForest(LOAN ~ ZODIAC + FIVEELEMENTS + MOON,data=train)
> pred.forest <- predict(loan.forest,newdata=test,type="class")
> mean(pred.forest != test$LOAN)
[1] 0
> loan.forest <- randomForest(LOAN ~ ZODIAC + FIVEELEMENTS + MOON,data=train)
> pred.forest <- predict(loan.forest,newdata=test,type="class")
> mean(pred.forest != test$LOAN)
[1] 0.002696456
```

Note that the test error rates are different for the two sets of predictions. Which statement below best applies to this code and the R output?

(a) Running `randomForest()` on the data changes the datasets. Thus, the two test error rates are different.
(b) The random forest prediction rule relies on random subsampling of the data. Thus, there's no guarantee that the test error rates will be the same.
(c) There's an error in the code – the test error rates should be the same.

For Problems 21–25, assume that the following data frame has been entered into R:

```
d <- data.frame(t=c(0,10,100),u=c("c","b","a"),v=c("a","b","c"))
```

In each problem, what would R say if you entered the given command in the console? Choose from the following possible answers.

(a)
```
[1] 0 10 100
```
(b)
```
  u   t
1 c   0
2 b  10
3 a 100
```
(c)
```
 [1] a b c
Levels: a b c
```
(d)
```
   t u v
2 10 b b
```
(e) Something else.

21. `d["u",]`

22. `d$t`

23. `d[3:1,"u"]`

24. `d[2,]`

25. `d[,2:1]`

For Problems 26–30, determine which type of plot would be best to use for visualizing/plotting the decsribed data. Choose from the following answers:

(a) Box plot
(b) Bar plot
(c) Scatter plot
(d) Mosaic plot
(e) Histogram

26. College admission decision at Rutgers (accept/reject) and gender for 2000 applicants to the university.

27. Annual income (in dollars) and age for 1000 individuals from the census.

28. Average annual temperature and elevation (ft. above sea level) at 500 locations around earth.

29. The frequency of sunny, partly cloudy, and cloudy days in New Brunswick during the months January-June.

30. The distribution of height for the students in this class.

Part II. Short answer

31. You are given a table that describes session revenues for users accessing your web site (similar to Google Analytics). Each row of the table corresponds to a session and is characterized by the following attributes (variables).

- `COUNTRY` (user location)
- `ACQUISITION` (possible values: direct, Google adwords, Google search, weblink)
- `SESSION` (duration in seconds)
- `REVENUE` (in dollars)
- `GENDER` (user gender)

What types of plots would you use to visualize the following relationships?

(a) The impact of session duration on revenue for each country.
(b) Fraction of sessions with male users per country.

32. Using the dataset described in the previous problem, suppose that you want to come up with a rule for predicting `GENDER` based on the other variables.

(a) Name two methods you could use to come up with such a prediction rule.
(b) Suppose that 95% of all users are men. Describe a simple prediction rule that doesn't require any R libraries or complicated functions, and is guaranteed to have an error rate no worse than 5%.

33. You are given a table that describes wines. Each row of the table describes a wine. It has the following attributes (variables): `NAME` of the wine, `COUNTRY` of origin, average `PRICE`, `RATING` (scale from 0 to 100), `ALCOHOL` content, `TYPE` (e.g. Cabernet, Chablis) and `YEAR`.

Let's assume that average price of red wine is $14.88 and average price of white wine is $14.22. We wish to test the hypothesis that red wines are more expensive than white wines.

(a) What is the null hypothesis?
(b) Let $D = 14.88 - 14.22 = 0.66$ be the difference between the average price of red wines and average price of white wines. Suppose that you permute the dataset 10,000 times and each time you compute a corresponding value of D using the permuted dataset; so in total you have 10,000 values of D in addition to the original value $D = 0.66$. Suppose that among the D's based on the permuted datasets, the 200th largest value of D is 0.5. What can you say about the p-value of your test? In particular, is there an upper or lower bound on the p-value?

34. A coin has been tossed 100 times, and the results are 56 heads and 44 tails. We wish to know if the coin is fair (i.e. if heads and tails are equally likely). Suppose that we have another coin, which we *know* is fair, and that we repeat this experiment using the fair coin 10,000 times. In other words, we flip the fair coin 100 times and record the number of heads and number of tails; and we repeat this process 10,000 times (in total, we make $10,000 \times 100 = 1,000,000$ flips with the fair coin). Suppose that in the 10,000 experiments with the fair coin, we got less then 44 tails 2,512 times. Is the original coin fair? What is the p-value for the test we have just described?

35. How many leaf nodes are in the decision tree from problem 19 above (for the Marriage data)?