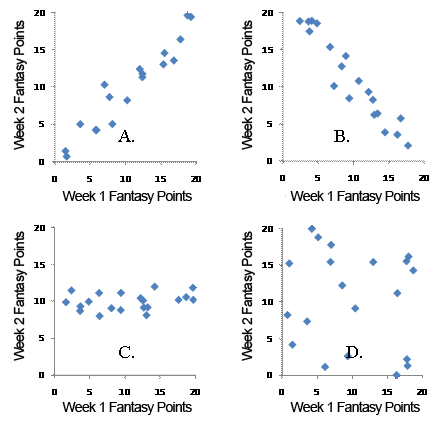
C105 Prediction Probability and Pigskin Fall 2014 Final Exam

*Good luck! You got this!*

1. Which of the following is *not* a normative statement?
   1. The optimal 15-player fantasy team has 5 running backs and 4 wide receivers.
   2. Fantasy managers should select a quarterback and running back in the first two rounds of the draft.
   3. **Fantasy managers tend to keep under-performing players on their roster if they picked them early in the draft.**
   4. NFL teams should never punt, because of the substantial opportunity costs associated with surrendering possession.
2. Human behavior deviates from the \_\_\_\_\_\_\_\_\_\_ statement in \_\_\_\_\_\_\_\_\_\_ ways.
   1. Descriptive / random
   2. Descriptive / systematic
   3. Normative / random
   4. **Normative / systematic**
3. According to classical decision theory, when you choose a starting quarterback, you are:
   1. Making a random choice between two options.
   2. **Carefully weighing the probabilities and values of each of your options.**
   3. Going with gut instincts about which player will earn the most points.
   4. Influenced by the risk of losing, and the consequences for smack-talk on the league’s message board.
4. Because people make decisions based on \_\_\_\_\_\_ risk and not \_\_\_\_\_\_\_ risk, NFL Coach Belichick’s (Patriots) decisions to go-for-it on 4th down more often, are \_\_\_\_\_\_\_ what other coaches would do in similar situations.
   1. actual / perceived / the same as
   2. perceived / actual / different from
   3. **actual / perceived / different from**
   4. perceived / actual / the same as
5. An NFL coach’s decision to attempt an onside kick may be affected by the way the decision is framed. Which of the following describes “framing” in this situation?
   1. **Whether the coach is considering the potential benefits or considering the potential losses associated with the onside kick.**
   2. Whether the coach is thinking algorithmically or heuristically.
   3. Careful consideration of the precise likelihood of recovering the onside kick, multiplied by the expected value of having possession at the 40 yard line.
   4. Whether the coach would go to Michael’s or Hobby Lobby to get the decision framed.

The following two questions refer to this scenario: As discussed in class, Kevin Kelly’s high school football team never punts. Coach Kelly’s non-traditional approach has been validated by game theorists, and has yielded great success for his program. It seems to be the optimal decision, and there’s strong empirical evidence that his refusal to punt has resulted in more wins. But despite the fact that many college and professional coaches are aware of his win record, and also aware of this quantitative research, nearly everyone, from high school through the NFL, still punts.

1. If coaches rely on punting because it just feels like the best option in these situations, regardless of what the stats show, what heuristic are they probably using?
   1. Availability heuristic
   2. Anchoring and adjustment heuristic
   3. Confirmation bias
   4. **Representativeness heuristic**
2. If coaches find reasons to punt because they can easily remember situations when punting seemed like a good idea, regardless of what the stats show, what heuristic are they probably using?
   1. **Availability heuristic**
   2. Anchoring and adjustment heuristic
   3. Confirmation bias
   4. Representativeness heuristic
3. The correlation coefficient (“r”) is a number that ranges from \_\_\_\_\_\_\_\_\_\_.
   1. 0 to 1
   2. 0 to 158.3
   3. **-1 to 1**
   4. 0 to 100
4. Imagine that researchers found a very strong correlation between the length of a customer’s driveway and the size of the tips the customer gave Papa John’s delivery people during Sunday Night Football. The longer the driveway, the smaller the tip the delivery person received. The correlation coefficient that most likely represents this relationship would be:
   1. + .90
   2. - .45
   3. + .45
   4. **- .90**
5. Which of the following diagrams shows a positive correlation between week 1 fantasy points and week 2 fantasy points?  
     
   
6. \_\_\_\_\_\_\_ tells you the strength and direction of a relationship between two variables, while \_\_\_\_\_\_\_ tells you the proportion of variability in one variable explained by another.
   1. r2 ; r
   2. r2; p
   3. r2; adjusted-r2
   4. **r ; r2**
7. If \_\_\_\_\_\_\_\_ is less than \_\_\_\_\_\_\_\_\_, most people will accept that the result probably wasn’t observed due to chance.
   1. a p-value; 1
   2. **a p-value; 0.05**
   3. a negative correlation coefficient; -0.05
   4. a standard deviation; 0.05

For the next 4 questions, consider the output of a regression analysis using the Excel Data Analysis Toolpak, below:



1. Which of the following is the equation for the best-fitting regression line?
   1. ŷ = 2.103 + 0.913x
   2. ŷ = 2.158 + 12.36x
   3. **ŷ = 12.36 + 2.158x**
   4. ŷ = 0.396 + 0.157x
2. The best-fitting regression line explains what proportion of variability in the “y” variable?
   1. 39.6%
   2. **15.7%**
   3. 12.9%
   4. 2.16%
3. What is the percent likelihood that the relationship between the x-variable and y-variable was observed due to chance?
   1. 39.6%
   2. 15.7%
   3. 2.16%
   4. **2.47%**
4. Does the x-variable have a statistically-significant effect on the y-variable?
   1. **Yes**
   2. No
5. Why is it misleading to use r2 to assess the strength of a linear model with multiple predictor variables?
   1. **Because when you add additional predictor x-variables, r2 will always increase, even if the predictor variables aren’t really related to the y-variable.**
   2. Because when you add additional predictor x-variables, r2 will always decrease, even if the predictor variables aren’t really related to the y-variable.
   3. Because it would be more appropriate to use the product of the coefficients’ p-values.
   4. Because you should use the adjusted-p-value, instead.



1. Which is the most likely equation for the regression line shown in the graph to the right?
   1. ŷ = -10 + 2x
   2. ŷ = 20 - 2x
   3. ŷ = 2 + 20x
   4. **ŷ = 20 + 2x**
2. Which of the following represents the cumulative distance between each data point (y) and a regression line (ŷ)?
   1. Correlation coefficient
   2. Coefficient of determination
   3. Intercept
   4. **Sum of squared error**
3. How is regression analysis different from correlation analysis?
   1. **Regression is for predicting values; correlation measures strength of a relationship**
   2. Correlation is for predicting values; regression measures strength of a relationship
   3. Regression determines cause-and-effect; correlation just measures the presence of a relationship.
   4. Regression tells you the direction of a relationship; correlation only measures the strength of a relationship.
4. If we do a linear regression analysis with multiple predictor variables, how would this change the equation for the output variable (ŷ)?
   1. There'd be more than one intercept.
   2. **There'd be more than one coefficient.**
   3. There'd be more than one solution.
   4. The r-squared would be higher.

For the following two questions, consider the correlation matrix shown to the right. V1 thru V6 are different variables.

1. What are the numbers inside the table?
   1. Eigenvalues
   2. **r-values**
   3. Distances to cluster centers
   4. Standard deviations
2. Which of the following statements are probably NOT true?
   1. **There’s a latent factor that accounts for variance in V3 and V4.**
   2. There’s a latent factor that accounts for variance in V1, V2, and V3.
   3. There’s a latent factor that accounts for variance in V4, V5, and V6.
   4. There’re at least two latent factors that account for variance in this dataset.
3. If a measure has low \_\_\_\_\_\_\_\_\_\_\_\_\_ variance, then it won’t be useful in a factor analysis.
   1. Specific
   2. Error
   3. **Common**
   4. Partial
4. What is a latent factor?
   1. **A hypothetical unobserved variable that is correlated with a number of observed variables.**
   2. The outcome of a Monte Carlo analysis.
   3. A specific observed variable that is correlated with other variables.
   4. An average of observed variables that are correlated with one another.
5. What is “specific variance”?
   1. The systematic ways that human decisions deviate from expected utility.
   2. The variance that is specifically due to measurement error.
   3. The variance that is common to all measures in a dataset.
   4. **The variance that is limited to a single measure/variable.**
6. Generally speaking, a factor loading of 0.80 means that:
   1. The variable is poorly related to the factor.
   2. That factor explains 20% of the variance in the dataset.
   3. **The variable is strongly associated with the factor.**
   4. That variable should be excluded from the factor analysis because the loading is less than 1.
7. In class, we ran factor analysis on all team defenses for the last 7 years, and found one factor that seems to do a very good job of explaining differences between defenses. The factor loadings are shown in the table to the right. Which of the following would apply to a defense that gets a large positive score for this factor?
   1. The defense is relatively good.
   2. **The defense is relatively poor.**
   3. The defense is able to limit opposing running backs.
   4. The defense earns a lot of turnovers.
8. Consider a hypothetical dataset like the one shown on the right, with four variables shown for each team: passYds, passTds, passAtts, and passComp. The numbers have been excluded so that no one gets distracted by details, but imagine that the numbers are really there. Given your understanding of football and factor analysis, which of the following is probably *NOT* true about this dataset?
   1. This dataset probably has high communality.
   2. There is probably at least one latent factor that describes variance in this dataset.
   3. **All values in the correlation matrix would probably be close to zero.**
   4. The eigenvalue of the first principal component is probably at least 1.
9. Which of the following is most likely a latent factor?
   1. A quarterback’s percent pass completion for a season.
   2. **A team defense’s “toughness.”**
   3. A running back’s yards-per-carry during away games.
   4. The number of fumbles lost under different weather conditions.
10. Some people just look at a scree plot to determine the number of components to include in a factor analysis, but there’s another way to select the number of components. Which of the following is it?
    1. Select components that give you the results you want. (this is not the correct answer)
    2. **Select components with eigenvalues over 1.0.**
    3. Select components with eigenvalues over 100%.
    4. Select components with error variance beneath 1.0.
11. Why on earth would anyone want to use factor analysis on a dataset?
    1. Because it’ll be really interesting to write about on Twitter. (hint: you’ll get unfollowed)
    2. Because it’s a really easy way to sound intelligent. (hint: it’s not easy)
    3. To increase sex appeal. (hint: it won’t)
    4. **To reduce the number of dimensions (or variables) describing some observations.**
12. What does factor analysis do that cluster analysis *can’t* do?
    1. **Factor analysis tells you precisely how your observations differ from one other.**
    2. Factor analysis provides a way to classify your observations into different groups.
    3. Factor analysis takes a large number of variables as input.
    4. Factor analysis can eat more hot dogs than Joey Chestnut.
13. In addition to the data for cluster analysis, which of the following needs to be provided as input for k-means clustering?
    1. A threshold for the eigenvalue of each cluster
    2. A threshold for the communality of all variables
    3. **The number of clusters**
    4. The number of observed variables to include in the final cluster analysis
14. In k-means cluster analysis, which of the following will always *decrease* the average distance to cluster center?
    1. Adding more variables to the input dataset.
    2. Adding more observations to the input dataset.
    3. **Adding more clusters to the analysis.**
    4. Adding more outliers to the dataset.
15. How does the k-means clustering algorithm initialize?
    1. It computes the optimal locations for the initial centroids.
    2. It automatically determines the correct number of clusters by analyzing variance in the dataset.
    3. With a “k”.
    4. **It selects random initial locations for the cluster centroids.**
16. When has the k-means clustering algorithm converged?
    1. When the average distance to cluster center has been maximized.
    2. When the distances between centroids have been maximized.
    3. When the initial clusters combine into a single cluster.
    4. **When the centroids don’t move anymore.**
17. Cluster analysis can be used to…
    1. … determine how running backs are different from one another.
    2. **… classify running backs into different categories.**
    3. … find the best running back.
    4. … determine the optimal lineup for an auction league’s draft.
18. In cluster analysis, observations with larger distances between them are more similar to each other than those at smaller distances.
    1. **False**
    2. True
19. If I ran cluster analysis on NFL quarterbacks, it'll assign each individual QB into "groups" or "clusters" by…
    1. **…determining which centroid is closest to each individual QB.**
    2. …creating a correlation matrix and determining which QBs are best-correlated with each other.
    3. …drawing a regression line.
    4. ...inserting random observations of each individual quarterback and running the analysis 10,000 times.



initialization

1. Consider the initialization of the k-means clustering algorithm shown above (dots are observations, stars are centroids). Which of the four options (labeled A-D) most likely depict the next step in the algorithm?

For the next three questions, consider the following: Imagine that you create a simple multiple regression model of quarterback performance:

expected fantasy points = 0.987 X + 0.015 Y + 2.45 Z + 5

… where X represents past fantasy points per game, Y represents passing yards allowed by opponent per game, and Z represents passing touchdowns allowed by opponent per game. You want to use this model in a Monte Carlo simulation to estimate Tom Brady’s performance in an upcoming game against the Houston Texans.

1. If you were going to use this model in a Monte Carlo simulation, what might you do?
   1. Plug-in Tom Brady’s average weekly fantasy points for X, and plug-in the Texans’ average allowed passing yards and TDs for Y and Z, and run the model 1,000 times.
   2. Plug-in Tom Brady’s average weekly fantasy points for X, and plug-in the Texans’ average allowed passing yards and TDs for Y and Z, and compute the outcome.
   3. **Plug-in Tom Brady’s fantasy points in a randomly chosen week for X, and plug-in the Texans’ allowed passing yards and TDs in a randomly chosen week for Y and Z, and run the model 1,000 times (using different randomly-chosen values each time).**
   4. Take it to Las Vegas.
2. What’ll be the output of this Monte Carlo simulation?
   1. **A bunch of numbers: A range of different fantasy points that Tom Brady might be expected to earn.**
   2. A single number: Tom Brady’s expected fantasy points.
   3. A bunch of numbers: A range of standard errors; the model’s accuracy at predicting past performance.
   4. A single number: The standard deviation of expected performances.
3. Why would the Monte Carlo approach yield better results than simply using the model?
   1. Because the Monte Carlo approach gives you one single expected value for Tom Brady’s expected points, while modeling has some error associated with its predictions.
   2. Because the Monte Carlo approach is faster to compute.
   3. This is a trick question: They’d both yield the same results.
   4. **Because the Monte Carlo approach uses more realistic observations as inputs.**
4. There’s something called “the Monte Hall problem,” based on a game show challenge. In this proble, Monte Hall, the host of “Let’s Make a Deal,” gives you an opportunity to win a new car. The car is behind one of three doors, and there’re goats behind the other two doors. You pick a door, but then Monte Hall opens one of the *other* two doors, showing you a goat. You have the option of keeping your original door (which is still unopened), or switching to the other unopened door. How would you “solve” the Monte Hall problem (knowing whether to stay or switch doors) using simulation?
   1. Create a regression model with three inputs: car location, tendency to stay, and tendency to switch, and plug-in average outcomes.
   2. **Attempt the situation 1,000 times when you switch doors, and 1,000 times when you stay on your original door, and see which results in more wins.**
   3. Create a linear regression model that takes two inputs: likelihood to switch and likelihood to stay, and have Excel optimize the outcome.
   4. This situation is not conducive to simulation, because the Monte Hall problem is not solvable. There’s no difference in outcomes when you switch doors or stay on your original door.
5. Using just your league’s initial draft results, the Monte Carlo simulation of fantasy football that we explored in the Week 15 lab (NFL week 13) might tell you which of the following:
   1. **That your fantasy team performed better than 85% of modeled seasons.**
   2. The precise number of fantasy points you should’ve scored by the 13th week of the NFL season.
   3. The best regression model for predicting player performance.
   4. Who to start and drop in the subsequent (14th) week of the NFL season.
6. According to the article in the *North Carolina Journal of Law & Technology*, when determining whether a game should be considered gambling, “[t]he test of the character of the game is not whether it contains an element of chance or an element of skill, but \_\_\_\_\_\_\_\_\_\_.”
   1. …whether the game is considered to be moral or immoral.
   2. …whether society would be improved if the game were outlawed.
   3. **…whether chance or skill is the dominating element that determines the result of the game.**
   4. …how much money the economy would lose if the industry were to be outlawed.
7. Betting on your team’s fantasy success is legal in Indiana because:
   1. …gambling is legal in Indiana.
   2. **…the outcome of a fantasy season can be determined by player skill.**
   3. …the game is usually conducted online, which is not under the state’s jurisdiction.
   4. Trick question: betting on your team’s fantasy success is gambling, which is illegal in Indiana.

Happy Hannukah! Merry Chrismas! Joyous End of the Semester! Happy Kwanzaa! Your holiday gift is two bonus points on this final exam:

1. Which is correct?
   1. Incorrect
   2. Incorrect
   3. **Correct**
   4. Incorrect
2. Which is true?
   1. False
   2. **True**

It's been a real pleasure having you in C105 this semester. I hope you learned a lot in the class and maybe even had some fun. If there's ever anything that I can do to help you out in the future, don't hesitate to ask. Have a restful and relaxing holiday break.