Stat 131A Fall 2021 **ORAL FINAL EXAM QUESTIONS**

Random variables and their distributions, data distributions

1. Interpreting EDA, such as histograms and box plots, deciding how many breaks looks best. Interpreting the histograms and box plots (we will show you plots and ask you questions about them).
   You should be able to read a box plot and tell us about the connections between box plots and histograms etc. and generally be familiar with numerical and graphical summaries.

2. Transforming data. When would you do it, and why? Specifically, when do you want to take the log transformation of variables? What are the limitations of log and square root transformations? How can we work around this?

3. Effect of transformations of data on numerical summaries such as mean, median, sd , IQR

4. What are density histograms? How do we compute the height of a rectangle for a given bin? Recognizing heavy vs light tailed histograms, estimating median, percentiles from a histogram.

5. Probability density function and cumulative distribution functions, connection between them, statement of central limit theorem, , Law of large numbers, the (continuous) distributions: uniform, binomial, normal distributions.

6. Estimating the pdf and cdf from a sample, kernel density estimators, show that $\hat{p}(x) = \frac{1}{n} f(x, x_i)$ is a density function, that is, it integrates to 1. How do we go from density histograms to kernel density functions?

Hypothesis testing

7. Given a situation, decide if a hypothesis test is appropriate, and set up the test. Define H_0 and H_1, compute the test statistic. Say I give you the P-value, tell me whether to accept or reject. Tell me how to compute the power and how I might increase it.

8. Definition and interpretation of P-values and power. Definition of type I and type II errors. How might assumptions be violated? What are the consequences of violating the assumptions? How do you check the assumptions?
   What is the duality between Hypothesis tests and Confidence Intervals? Specifically, given a confidence interval of some statistic, θ, calculated from an observed value of θ, what are the conclusions for $H_0 : \theta = 0$ and $H_a : \theta \neq 0$?

9. What is a t-test? What are the assumptions? What is the test statistic? What does significance level mean? What is the connection to Type I error? How are power and significance level related?

10. How do we compute bootstrap confidence intervals? Can you explain the set up of a bootstrap confidence interval? What are the assumptions?

11. What is a permutation test? What are the assumptions? Why would we use it? Describe the different resampling processes of a permutation test and computing a bootstrap CI.

12. What is multiple testing? What is the Bonferroni correction? What is the family-wise error rate?

13. How do we compute parametric confidence intervals? Given output from a test, write down a confidence interval and vice versa (given a CI, can you say what the result of a test would be)?

Curve Fitting

14. Assumptions of simple linear regression, and how to estimate the line. Meaning of "least squares fit", definition of loss function, Estimate of slope and intercept of the regression line.

15. Residuals of regression, inference on coefficients

16. Give the pros and cons (and assumptions) of bootstrap CIs and parametric CIs.

17. How would you implement a polynomial model in R? What is the LOESS model?

Multivariate Visualization

18. How might you visualize given data? Given a plot, can you critique the visualization? How do you visualize categorical/numerical data? Read the mosaic and alluvial plots, or bar plots, and interpret them.

19. Given a contingency table, can you compute the conditional distribution?

20. Given data, you should be able to describe how you might visualize it, and be able to read alluvial plots and mosaic plots.

21. When would we use principal components? What are the two ways to think about PCA? How is a PC represented? Given the first PC, how do you find the second PC? How do we decide how many PCs to use? How many are there?

22. What is a heat map for correlation? What do the colors of a heat map for data represent? In general, what does a heat map show? What do the dendrograms show on heatmaps? Whats the difference between the dendrograms on the columns and rows of a heatmap?

Multiple regression

23. You should know what the assumptions of linear regression are, and the consequences of those assumptions (and violating them). Given regression output, interpret coefficients, including those of categorical variables.

24. How do you interpret the coefficient β_j if we take the log transformation of the explanatory variable x_j? What about if we take the log of the response variable y? What if we take the log of both?

25. Show that the sum of the residuals in a linear regression must be 0.
Is TSS always greater than RSS?

26. True or false & explain (this is to give you the *flavor* of what we might ask):

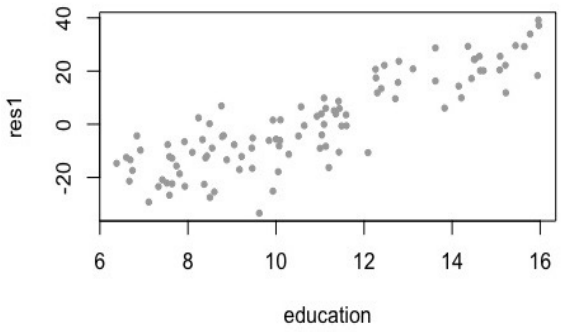
    (a) If the P-value for an estimated coefficient in the output from lm is small, this means that there must be a linear relationship between the response and the corresponding variable.

    (b) R^2 *always* increases if we add an explanatory variable and therefore, so does adjusted R^2.

    (c) We can perform the F-test on two models if the size of one model is strictly smaller than the other.

    (d) It is possible that none of the predictors is significant but the F-statistic is.

27. Install the package "car" and look up the dataset `Prestige` (Prestige of Canadian Occupations). Consider the following plot, which shows the **residuals** from the fit of `prestige` (prestige score) to the variable `women` (percentage of female incumbents), plotted against the variable `education` (average education of incumbents). Use this plot to decide if the following statements are true or false.



(a) **True/False** The variable `education` should be included in the model.

(b) **True/False** The estimate of `women` in the full model will have a large SE.

28. If I give you regression output (summary of the output from lm()), with some of the values missing, you should be able to fill out the missing values.

29. A dataset has two numeric variables and one categorical variable with 4 unique categories. How many coefficient, in total, does lm() output, if you want a model with different intercepts and different slopes for each category?

30. F-test for global fit; Definitions TSS, RSS, F, Given n = number of observations, and p = number of variables, and the value of R^2, you can test global fit, that is $H_0 : \beta_j = 0$?

31. How do we perform a permutation test? What are the assumptions? Bootstrap CI: how do we compute this? What are the assumptions here? Describe the different resampling processes of a permutation test and computing a bootstrap CI

32. What is the difference between the prediction interval and the confidence interval?

33. What does regsubsets() do in R? Why would you want to use cross validation on the result of regsubsets()? Why not just pick the model with smallest error or AIC?

34. Why should we do variable selection? What is the best criterion to use for model selection? Why? Should you use step-wise regression using a criterion or regression subsets with cross-validation? What are the pros and cons of each?

35. Given the output from lm(), you should be able to say which variable you would want to remove, based on p-value, and also tell us why, in general, this might not be a good approach when you want to reduce the number of variables of the current model?

Logistic regression

36. How did we set up a logistic regression? Why not just use the usual least squares model? What is the response? Why bring in log-odds? How is randomness accounted in the logistic regression model?

37. Given a logistic regression output from glm, can you give us the log-odds of the response? Given the log-odds, can you tell us the probability of the response for a given value of the explanatory variable?

38. I will give you the results from a logistic regression, and ask you to compute odds or log-odds, or probability, and also might ask you what you can infer. That is, what is the scope of your inference? Can you use your results as a proof?