XDAS2020 Final Exam Study Questions

Distributions

1) The shape of the normal distribution is defined by two parameters, mean and standard deviation. They control the center and spread of the normal curve. What parameters control the shape of the following distributions? How do the distributions change as these parameters change?
    a. Poisson distribution
    b. Binominal distribution
    c. Negative binomial distribution
    d. Geometric distribution
    e. Hypergeometric distribution

2) How are the exponential and Poisson distributions related conceptually? Give an example of two related questions that can be answered using these two distributions.

3) The binomial and negative binomial are discrete distributions that are related in some way. Describe the difference between these, and outline in broad terms an illustrative case study (in biology) where each would be applied.

4) What is a log-normal distribution and when is it sometimes useful? Give an example.

5) Why is the negative binomial a better model than the Poisson for RNA-seq data? (This relates to noise in gene expression studies as a function of gene expression levels and something called “overdispersion”.)

Hypothesis Testing

6) Hypotheses, Error and Power
    a. What is a "null hypothesis"? What “alternative” hypotheses can be tested?
    b. Define Type I and Type II errors and clearly explain the difference between them
    c. What is power, and what’s the tradeoff between error and power?

7) Parametric vs Nonparametric
    a. What is the main difference between parametric and nonparametric tests?
    b. What are the advantages of a nonparametric test? What are the advantages of a parametric test?
    c. Which nonparametric test compares values between two independent populations to find if one is greater than the other? What is the test statistic for this nonparametric test? How is it calculated?

8) P-values
    a. What is a p-value?
    b. What are the shortcomings of p-values?
    c. Is it possible for something to be significant but not important? Explain.

9) Confidence Intervals
    a. What is a confidence interval? What, specifically, does a 95% CI mean?
    b. Why and how are confidence intervals useful? In particular, how do confidence intervals complement p-values?

10) T-tests
    a. What is the purpose of the t-test?
    b. What are some assumptions about that data that need to be true in order for someone to use the t-test?
    c. What is the formal, mathematical definition of the t-statistic?
    d. What is the difference between a one-sided and two-sided t-test? What are the null and alternative hypotheses for each?
    e. What does a significant p-value of such a t-test mean?
    f. How are confidence intervals for t-tests determined for two-sample comparisons?

11) Multiple Hypothesis Testing
    a. Why is multiple hypothesis testing important for high-dimensional data?
    b. How does controlling for False Discovery Rate (FDR) work? Outline the general framework for controlling the FDR to 5%.

12) Gene Ontology Enrichment
    a. What is Gene Ontology (GO) and why is testing enrichment of Gene Ontologies in a subset of genes often useful?
    b. Explain which statistical test is most frequently used for GO enrichment testing and why.

Statistical Modeling

13) Model Formulae
    a. In the formula Y ~ X, what is another name for Y and for X?
    b. How do you write a formula if you are interested in an interaction term?
    c. How would you determine whether an interaction term should be included in your model or not?
    d. Describe a hypothetical experimental scenario when an interaction term might be significant.

14) ANOVA models
    a. What does ANOVA test?
    b. What types of values (continuous, discrete, or categorical) are the Response and Predictor variables?
    c. Why is it useful to consider interaction terms instead of just marginal effects?
    d. For the image of an ANOVA result below:
        i. What is the Df column describing?
        ii. What is the Sum Sq column describing?
        iii. How is the F distribution created? How is the F-statistic calculated?
        iv. What is the null hypothesis of the F-test?

```
summary(aov(SQBLOOMS ~ WATER + BED + SHADE, data=blooms))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## WATER        2  3.715  1.8577   8.503 0.001302 **
## BED          2  4.132  2.0661   9.457 0.000728 ***
## SHADE        3  1.646  0.5488   2.512 0.078945 .
## Residuals   28  6.117  0.2185
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

15) Regression models
    a. When performing a linear regression, what type of values (continuous, discrete, categorical) are the Response variable and the Predictor variable?
    b. Why would someone want to create a regression model?
    c. When looking at the results, what does the R^2 value represent? Provide a definition for R^2 and describe the concept behind this measure.
    d. What does the Estimate mean? Describe the basic idea of how the estimate is calculated.
    e. What is the null hypothesis of the test that provides the p-value for the predictor?

16) Planned vs Unplanned Experimental design
    a. What is the difference between planned and unplanned experimental designs? Give an example of each.
    b. How does one perform an anova analysis in R of a planned experiment? How do you determine the effect size?
    c. How does one perform an ANOVA analysis in R of an unplanned experiment? How do you determine the effect size?

17) Logistic Regression
    a. For what kind of question is logistic regression used? Give an example.
    b. What types of values (continuous, discrete, or categorical) are the Response and Predictor variables?
    c. How would you decide between ANOVA, linear regression, and logistic regression?

18) Bayesian Models
    a. What is the fundamental conceptual difference between Bayesian statistics and "frequentist" statistics?
    b. Outline the basic framework for Bayesian analysis.
    c. What is a prior?
    d. Give an example (e.g. from class) to which you could apply a Bayesian model and discuss how your estimates might change with more data.

Descriptive Statistics

19) Distance
    a. Explain the difference between Euclidean distance and Manhattan distance.
    b. Explain the relationship between covariance and Pearson correlation. What are the similarities? What are the differences? How does Pearson correlation differ from R^2?
    c. When is it more appropriate to use Euclidean Distance vs. Correlation to cluster genes (and vice versa)? Why?

20) Principal Components Analysis
    a. What's the basic idea behind PCA, and how are principal components identified?
    b. Why is it useful to use dimensional reduction methods like PCA?
    c. How many principal components can be calculated for FACS data that has 2500 observed cells and eight features (six fluorescent data channels, side scatter, and forward scatter)? Why?
    d. Give an example of an application for PCA and what you would gain from it.

21) Clustering
    a. What are the steps for Hierarchical clustering?
    b. What are the steps for K-means clustering?
    c. What are the advantages and disadvantages of each method?

Tabular Statistics

22) Describe a simple scenario in which you would use a contingency table.

23) How do you calculate the Chi-Square test?

24) When is it NOT OK to use the Chi-Square test?

25) What distribution is the Fisher's Exact Test based on? Is there a model design for which an alternative test might be preferred?

26) How do you calculate the p-value for Fisher's Exact test?

Resampling methods

27) Why might someone want to use resampling instead of a t-test?

28) How can someone determine if the difference of the means from two samples is significant using the resampling method? Describe the steps in detail.

29) Explain what the bootstrap is and why it is often useful in practice.

30) What is the difference between the bootstrap and Monte Carlo simulation?