XDAS 2019 Final Exam

Distributions

1) The binomial and Gaussian are 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.

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

3) How is the Poisson related to the exponential distribution? To what kind of processes / situations would the apply? Give an example of how the Poisson is relevant for single-cell experiments that use microfluidics.

4) 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

5) In terms of the power analysis, what kinds of errors can arise in comparing two samples? What's the tradeoff between error and power? Be sure to define error types and power in your answer.

6) P-values and confidence intervals

   a. What is a p-value, and what are some of its shortcomings?

   b. How can a test result be important but not significant? Give an example.

    c. What is a confidence interval? How do CIs complement p-values? Describe what a 95% CI means specifically.

7) Parametric and non-parametric tests

    a. What is a t-test for? What are some assumptions about that data that need to be true in order for someone to use the t-test?

    b. What does a significant p-value of such a t-test mean?

    c. List and describe two different alternative hypotheses for two-sample comparisons.

    d. What is the difference between parametric and non-parametric tests? When would you choose a rank-based test over a parametric one? Give an example.

8) Multiple Hypothesis Testing

    a. Why is multiple hypothesis testing important for high-dimensional data?

    b. What's the essential difference between the Bonferroni correction and FDR?

Statistical Modeling

9) ANOVA and regression

   a. What types of values (continuous, discrete, or categorical) are the Response and Predictor variables for ANOVA, linear, and logistic regression? Consider making a table to answer this question.

   b. How would you decide between ANOVA, linear regression, and logistic regression? Give an example scenario in which you would apply each of these tests.

c. Why and when is it useful to consider interaction terms instead of just marginal effects? Give an example.

d. In regression models, what does the R^2 value represent? Describe the concept behind this measure.

10) Bayesian Models

    a. What is the fundamental conceptual difference between Bayesian statistics and "frequentist" statistics?

    b. 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

11) Distance and Clustering

    a.   What's the essential difference between Euclidian and correlation-based distance? When is it more appropriate to use one or the other to cluster genes? Why?

    b.   What are three major differences between the mechanics of hierarchical and K-means clustering?

c.   What are advantages and disadvantages of each clustering method?

12) Principal Components Analysis

a.   Why is it useful to use dimensional reduction methods like PCA?

b.   What's the basic idea behind PCA, and how are principal components identified? How do PC's relate to the original variables measured?

c. How can you decide on the right number of principal components to use for your analysis?

d. Give an example of an application for PCA and what you would gain from it.

Tabular Statistics

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

14) When is it preferable to use Fisher's Exact Test vs. a Chi-Square test? Discuss the underlying probability distributions these two approaches are based on, and any important assumptions that must be considered.

Resampling methods

15) Why may someone want use resampling instead of a parametric test (e.g. t-test, ANOVA, Chi-squared)?

16) How can someone determine if the occurrence of a GO-term is enriched in a list of genes using the resampling method? Describe the steps in detail.

EXTRA SPACE IF NEEDED

EXTRA SPACE IF NEEDED

EXTRA SPACE IF NEEDED