BIOS 721 Final Project Fall 2013

Introduction

The final project will consist of performing a statistical simulation study and drafting a written report describing the study and its major findings. Students should work INDEPENDENTLY on the final project. If they have any questions about the final project, these should be discussed with Megan only and NOT with other students or other faculty.

- Due **Saturday December 14, 2013** by midnight in your Sakai Drop Box.
- Please draft your written report following the guidelines outlined below in the Report Section of this assignment. You may draft report using Word or LaTeX. Save the file using the following naming convention:
    - LastName_FirstName_FinalProject.docx (or pdf)

Background

Genetic association studies search for correlations between genetic variations and trait values (phenotype). Genetic markers that are often used in these studies are single nucleotide polymorphisms (SNPs) in which the genetic variation involves a single base change. SNPs are diallelic markers taking two forms, usually denoted by A and a, which results in three possible genotypes (combination of an individual's two alleles, one from each chromosome, at a locus), usually denoted by AA, Aa, and aa. By convention, 'A' usually refers to the allele that is less frequent in the population (i.e. the 'risk' allele) and its frequency is known as the minor allele frequency. When using SNP-based tests in genetic association studies, each marker under consideration is tested for association between its genotypes and the phenotype. If the phenotype is binary (e.g. disease status), two commonly used procedures for detecting association between SNPs and the outcome are Pearson's chi-square test and logistic regression.

Objective

To compare the power of Pearson's chi-square test and logistic regression to detect association signals between a single SNP and a binary phenotype in a case-control study.

Under the Pearson' chi-square test framework, the null hypothesis of no association is rejected if and only if

- Note: The p-value for the Pearson's chi-square test can be computed in R using `chisq.test(y,x)$p.value` where y is a vector indicating case-control status and x is vector indicating genotype group.

Under the logistic regression framework, a subject's probability of having the trait is modeled as function of their genotype. That is, the following model is fit using the sample

been generated. Under a case-control design, it is assumed that the prevalence of the trait is rare and that cases will need to be 'oversampled' in order to achieve the a:b ratio (here, 1:1) of cases to controls desired in the sample. Samples under this design can be created by randomly selecting subjects from the population until the requisite number of cases in selected and discarding the additional controls that were selected during the process.

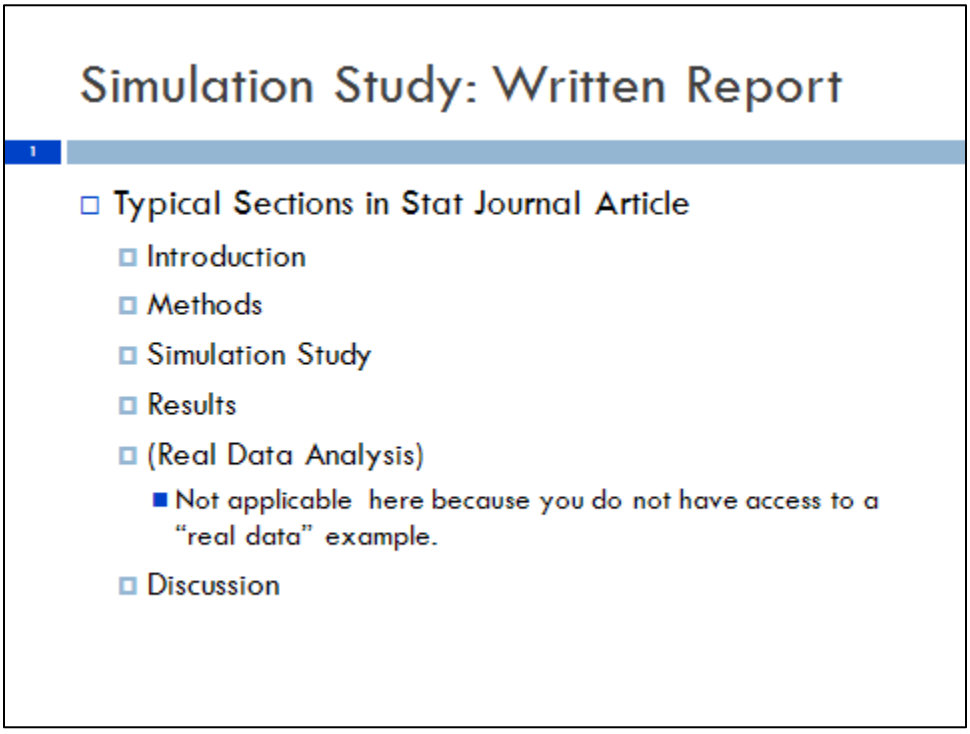
Simulation Design

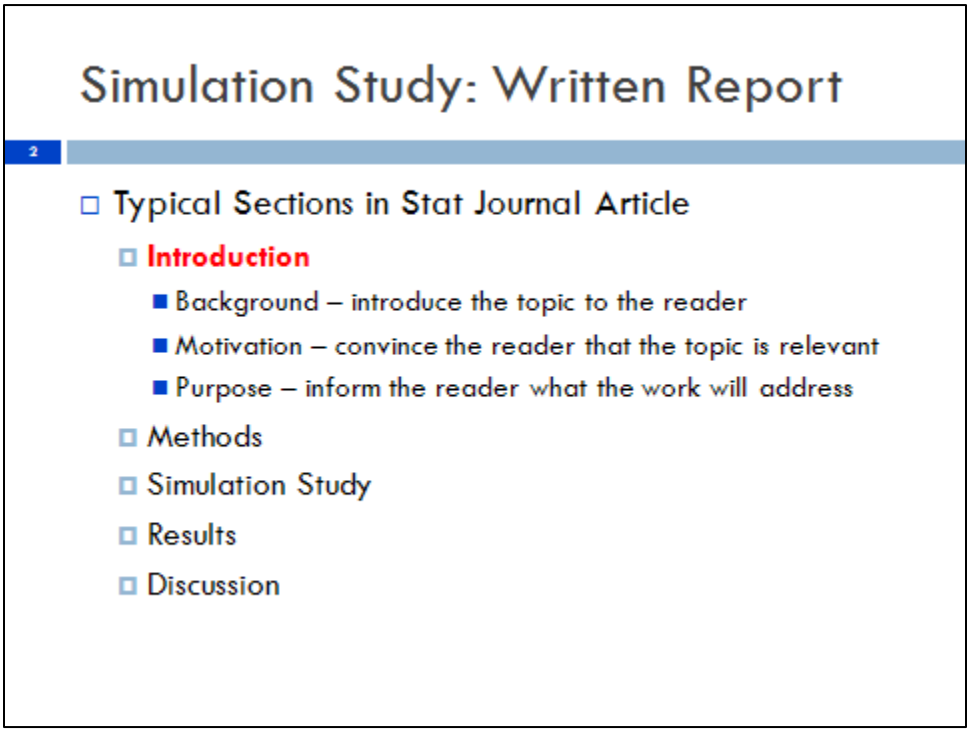
Data will be generated under the four genetic models listed below. The genetic model determines how the number of copies of minor alleles influences a subject's probability of being a case. The genetic models are described below:

1. General Risk increases with the number the copies of minor alleles

For each combination of your selected simulation parameters, you should generate 1,000 case-control samples with a 1:1 ratio under each genetic model listed above and perform Person's chi-square and logistic regression tests of association. The intercept in all models should be set to

Appendix: Tips for Writing Final Report





Simulation Study: Written Report

3

- Typical Sections in Stat Journal Article
  - Introduction
  - **Methods**
    - Describe each method under study precisely and clearly
    - This usually includes a discussion of the type of data the methods apply and the assumptions made by each method
  - Simulation Study
  - Results
  - Discussion

Simulation Study: Written Report

4

- Typical Sections in Stat Journal Article
  - Introduction
  - Methods
  - **Simulation Study**
    - Describe the design of the study
      - What factors you studied and why
    - Describe the data model
      - How was the data generated (could be complicated)
    - Discuss any computational details
      - How many simulation runs, what software you used to fit the methods, what results you computed, etc.
  - Results
  - Discussion

Simulation Study: Written Report

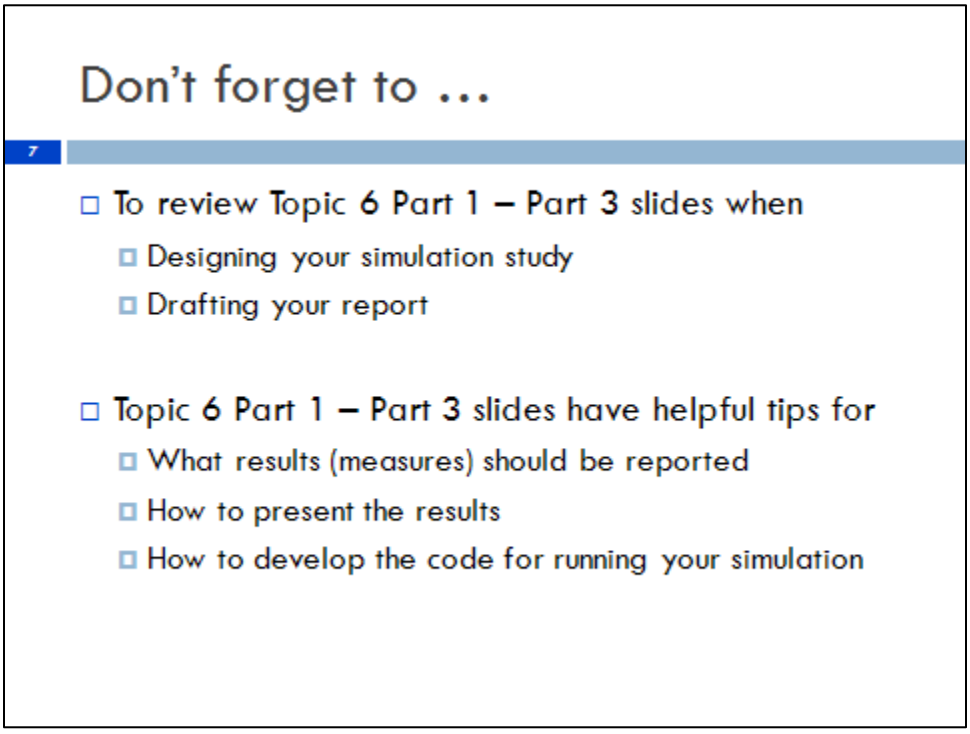
5

- Typical Sections in Stat Journal Article
  - Introduction
  - Methods
  - Simulation Study
  - **Results**
    - Describe the measures used compare the methods
      - Explain why these measure address the goals of the work
      - Explain to the reader how the results are presented
    - Clearly state the findings of EACH simulation setting
      - Start with the most interesting findings and then move onto those that are expected
      - After stating the facts of the findings, tell the reader what they mean in the context of the problem
      - Be careful when making "umbrella" statements
      - Use the same pattern of presenting the results for each simulation setting
  - Discussion

Simulation Study: Written Report

6

- Typical Sections in Stat Journal Article
  - Introduction
  - Methods
  - Simulation Study
  - Results
  - (Real Data Analysis)
  - **Discussion**
    - Restate the problem (i.e. motivation)
    - Concisely summarize what you found
      - Tell the reader what you recommend based on the findings
    - Describe any limitations or future work


Don't forget to …
7
To review Topic 6 Part 1 – Part 3 slides when
Designing your simulation study
Drafting your report
Topic 6 Part 1 – Part 3 slides have helpful tips for
What results (measures) should be reported
How to present the results
How to develop the code for running your simulation