CM226, Fall 2021
Problem Set 1: Statistics and Multiple Testing
Due Oct 18, 2021 at 11:59pm

Submission instructions

- Submit your solutions electronically on the course Gradescope site as PDF files.
- If you plan to typeset your solutions, please use the LaTeX solution template. If you must submit scanned handwritten solutions, please use a black pen on blank white paper and a high-quality scanner app.

1 Bias-Variance decomposition [8 pts]

(a) Given n iid samples from a distribution P: $X_i \overset{iid}{\sim} P, i \in \{1, \ldots, n\}$. Let $\hat{\theta} = t(X_1, \ldots, X_n)$ be an estimator of θ (which is a function of the distribution P). Show that the mean-squared error $mse(\hat{\theta}) = (bias(\hat{\theta}))^2 + \text{Var}\left[\hat{\theta}\right]$.

(b) We would like to estimate the variance σ^2 given n iid samples $X_i \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$. Two commonly used estimators are $S^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$, $\hat{\sigma^2} = \frac{n-1}{n} S^2$.

i. Compute the bias of the two estimators.

ii. Using the identity that $\text{Var}\left[S^2\right] = \frac{2\sigma^4}{n-1}$, compute the variance of $\hat{\sigma^2}$.

iii. Which estimator has a smaller mean-squared error?

2 Normal distribution[10 pts]

X_1 and X_2 are bivariate normal with mean (μ, μ) and covariance matrix

$$\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

(a) For what values of ρ is this a valid probability distribution?

(b) What is the marginal distribution of X_1?

(c) What is the conditional distribution of X_2 given X_1?

(d) Let $X = a + bZ$ where $Z \sim \mathcal{N}(0, 1)$. Show that X is normally distributed. What is the mean and variance of X?

3 Testing Mendel's first law [10 pts]

We set up an experiment to test Mendel's first law *i.e.*, the two copies of an individual's genome are equally likely to be transmitted to the offspring. Choose a SNP at which the individual carries different alleles. Denote the two alleles at this SNP as 0 and 1.

The state of the allele in a gamete (offspring) i is given by a Bernoulli random variable $X_i \overset{iid}{\sim} \text{Ber}(p), i \in \{1, \ldots, n\}$. Here p is the probability that a gamete inherits a 1 allele. Mendel's law implies $p = \frac{1}{2}$.

(a) Write the likelihood of p. Show that the maximum likelihood estimator of p, $\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x}$.

(b) The Fisher Information is defined as

$$I_n(p) = -\mathbb{E}\left[\frac{\partial^2 \log P(X_1, \ldots, X_n|p)}{\partial p^2}\right]$$

Calculate the Fisher information for the above model.

Assuming n is large so that $\hat{p}$ is asymptotically normal, what is the distribution of $\hat{p}$?

(c) Write the likelihood-ratio test statistic for testing $H_0 : p = \frac{1}{2}$ vs $H_1 : p \neq \frac{1}{2}$.

(d) If we observe all alleles of type 1 across 5 gametes what is the exact p-value of the LRT statistic? Using this p-value, would you reject the null hypothesis at a significance level of 0.05?

(e) Use the asymptotic distribution of the likelihood-ratio test statistic to compute the asymptotic p-value of the LRT statistic when we observe all allele type 1 for $n = 5$. Using this p-value, would you reject the null hypothesis at a significance level of 0.05?

4 Multiple hypothesis testing [10 pts]

(a) We want to find which of m genes differ in their expression level between individulas who belong to one of two groups (for example, disease vs healthy). We are testing m hypotheses: $H_{0,i} : \mathbb{E}[X_i] = 0, i \in \{1, \ldots, m\}$. Here X_i is a random variable that measures the difference in expression of gene i across the two groups. Our data consists of the expression level for each gene in each individual. Assume we have chosen a statistic as well as a procedure to compute a p-value for this statistic under each of the null hypotheses $H_{0,i}$. We compute m p-values, $p_1, \ldots, p_m$ and decide to reject all hypotheses with p-value $\leq \alpha$. We would like to estimate the FWER and the FDR for this procedure.

One approach to do so relies on using permutations to estimate the distribution of the test statistics (or their p-values) under the null. Specifically, we permute the groups to which each individual is assigned. For each permutation $t = \{1, \ldots, B\}$, we compute p-values $p_1^{(t)}, \ldots, p_m^{(t)}$.

How do we estimate the FWER and FDR from this data?

(b) A study examines $m = 3226$ genes across two conditions and finds 51 genes to differ. Of these 51 genes, 9 are known to be truly null. Among genes found to not differ, 2000 are known to be truly null.

i. What are the false positives, false negatives, true positives and true negatives for these tests?

ii. What is the false discovery proportion?

iii. The sensitivity or power is defined as the fraction of true non-null hypotheses that are predicted to be non-null. The specificity is the fraction of null hypotheses that are predicted to be null.What are the sensitivity and specificity?

(c) Show that FDR $\leq$ FWER.

5 Data analysis [15 pts]

In this problem, you will test the assocation of SNPs to a phenotype using permutations as well as asymptotic approximations and compare the two approaches.

Provided is a data set of simulated phenotypes (Y) for 250 individuals and a corresponding matrix of genotypes at 10 SNPs (G). We are interested in testing whether the genotype is associated with the phenotype in this data. To determine this, we will use the following test statistic.

$$T_i = N\rho^2(Y, G_i)$$

Here $\rho^2(Y, G_i)$ refers to the squared Pearson correlation coefficient between phenotype and genotype at the i'th SNP and N is the number of individuals.

i. Design and implement a permutation test for the first SNP (column 1 of the genotype matrix G). Plot the observed test statistic T_1 as well as a histogram of the test statistic from $B = 100,000$ permutations.

ii. What is the p-value of T_1 estimated by permutations?

iii. The test statistic T_1 asymptotically follows a χ^2 with one degree of freedom under the null. Plot the probability density function for a chi squared with one degree of freedom. What is the p-value of T_1 based on the chi-square approximation?

iv. We would now like to test each of the 10 SNPs but want to control the FWER at level 0.05. One approach to control FWER is the Bonferroni procedure. What is the p-value threshold at which we should reject each of the 10 p-values using the Bonferroni procedure?

v. Extend the permutation test to do multiple testing adjustment for all 10 SNPs. Run at least $B = 10,000$ permutations. What is the p-value threshold needed to control FWER at 0.05? Is there any evidence of association? Explain the apparent discrepancy in results, if any, compared to the above question.