DSB205, Winter 2025
Problem Set 1: Statistics and Multiple Testing
Due Jan 31, 2025 at 11:59pm PST

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 Testing Mendel's first law [12 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. We are examining a SNP at which the individual carries different alleles on their maternal and paternal genome. Denote the two alleles at this SNP as 0 and 1.

We observe the allele carried by n offspring. The state of the allele in an 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 first law hypothesizes that $p = \frac{1}{2}$.

(a) Write the likelihood of p (Hint: to write the likelihood, we need the probability of a Bernoulli random variable X: $P(x) = p^x(1-p)^{(1-x)}$).

$$L(\theta|x) = \prod_{i=1}^{n} P(x_i|\theta) = \prod p^x(1-p)^{1-x}$$

(b) Write the log-likelihood of p.

$$LL(\theta|x) = log \prod p^x(1-p)^{1-x} = \sum_{i=1}^{n} log[p^x(1-p)^{1-x}] = \sum_{i=1}^{n} xlog(p) + (1-x)log(1-p)$$

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

$$MLE = argmax_p LL(p|x) \implies \frac{d}{dp} LL(p|x) = 0 \implies \frac{d}{dp} \sum log[p^x(1-p)^{1-x}] = 0$$

distributing the log, we get:

$$\frac{\partial}{\partial p} \sum log(p^x) + log(1-p)^{1-x} = \frac{\partial}{\partial p} \sum xlog(p) + (1-x)log(1-p) = 0$$

treating the log as Ln (to simplify derivation and because the base of log doesn't matter)

$$= \sum_{i=1}^{n} (\frac{x}{p} + -1(\frac{1-x}{1-p})) \implies \sum_{i=1}^{n} (\frac{x}{p} - \frac{1-x}{1-p}) = 0$$

$$\sum_{i=1}^{n} \frac{x}{p}(\frac{1-p}{1-p}) - \frac{1-x}{1-p}(\frac{p}{p}) = \sum_{i=1}^{n} \frac{(x-px) - (p-px)}{p(1-p)} = \sum_{i=1}^{n} \frac{x-p}{p(1-p)} = 0$$

multiplying both sides by denominator:

$$\sum_{i=1}^{n} x_i - p = 0 \implies \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} p$$

$$\sum_{i=1}^{n} x_i = np \quad p = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x}$$

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

$$LRT\lambda = \frac{L(\hat{\theta_0})}{L(\hat{\theta})} = \frac{\prod 0.5^x (0.5)^{1-x}}{\prod \bar{x}^x (1-\bar{x})^{1-x}}$$

(e) 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?

$$\{1,1,1,1,1\} \implies LRT\lambda = \frac{\prod 0.5^x (0.5)^{1-x}}{\prod \bar{x}^x (1-\bar{x})^{1-x}}$$

$$\prod_{i=1}^{n=5} 0.5^1 (0.5)^0 \implies 0.5^5 \quad \prod \bar{x}^x (1-\bar{x})^{1-x} = 1^5 (0)^0$$

$$LRT\lambda = \frac{0.5^5}{1} = 0.03125 = \frac{1}{32}$$

To get p value, let's see where this LRT falls under all other possible LRTs.
Since the alleles are iid, the Likelihood of $\{1,0,0,0,0\}$ is the same as the Likelihood of any other combination that results in 1 total allele. So, calculating all LRTs based on # of 1s:

$$1 : LRT = \frac{0.5^5}{\prod 0.2^1 (0.8)^0} = \frac{0.5^5}{0.2^1 (0.8)^4} = 0.3814$$

$$2 : LRT = \frac{0.5^5}{\prod 0.4^1 (0.6)^0} = \frac{0.5^5}{0.4^2 (0.6)^3} = 0.9042$$

$$3 : LRT = \frac{0.5^5}{\prod 0.6^1 (0.4)^0} = \frac{0.5^5}{0.6^3 (0.4)^2} = 0.9042$$

$$4 : LRT = \frac{0.5^5}{\prod 0.8^1 (0.2)^0} = \frac{0.5^5}{0.8^4 (0.2)^1} = 0.3814$$

As we can see, every other LRT is higher than the LRT of five type 1 alleles. There is one additional LRT that shares the same value (a set of five 0s). Thus, our probability of getting this LRT or lower is 2 divide by the total number of combinations possible, which is $2^5 = 32$. 2/32 leads to p value of 0.0625. So, we would not reject.

(f) 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?

$$-2ln(\lambda) \sim \chi_1^2$$

$$-2ln(\lambda) = -2ln(\frac{1}{32}) = 6.93$$

At significance level 0.05, we reject the null because $6.93 > 3.841$ (from Chi squared table)

2 Multiple hypothesis testing [6 pts]

A study examines $m = 3226$ genes across two conditions. For each gene, the study aims to test the null hypothesis that the expression level of the gene is the same across the two conditions. Applying a hypothesis test, the study finds 51 genes to differ in their expression level across the conditions. Of these 51 genes, 9 are known to be truly null. Among genes found to not differ, 2000 are known to be truly null.

(a) What are the false positives, false negatives, true positives and true negatives for these tests?

FP: 9, FN: 3226-(51+2000) = 1175, TP: (51-9) = 42, TN: 2000

(b) 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?

Sensitivity = TP/(TP+FN) = 42/1217 = 3.45%
Specificity = TN/(TN+FP) = 2000/2009 = 99.6%

3 Data analysis [15 pts]

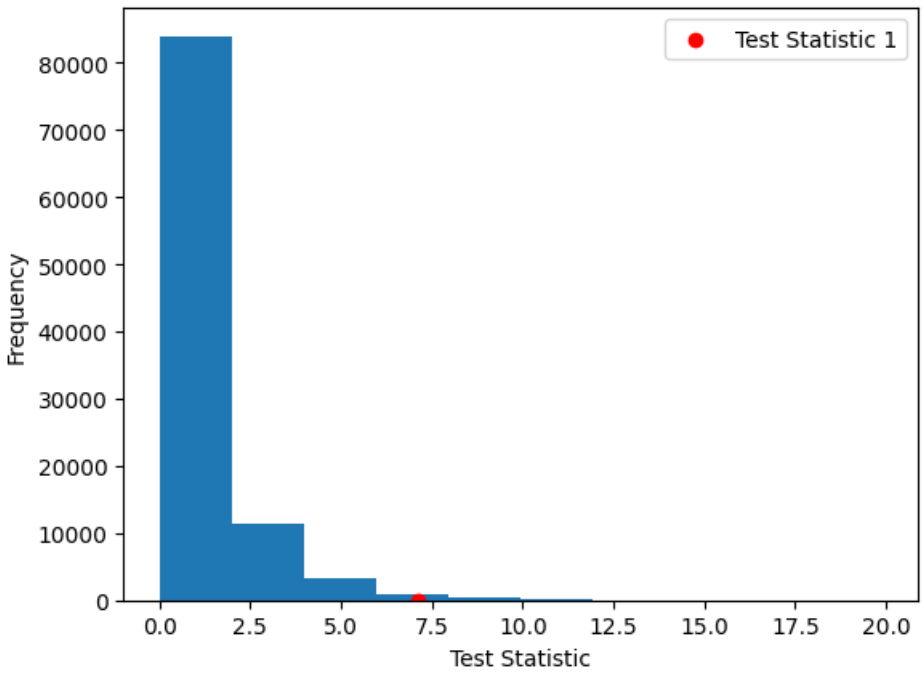
In this problem, you will test the association of SNPs to a phenotype using permutations as well as asymptotic approximations and compare the two approaches.

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.

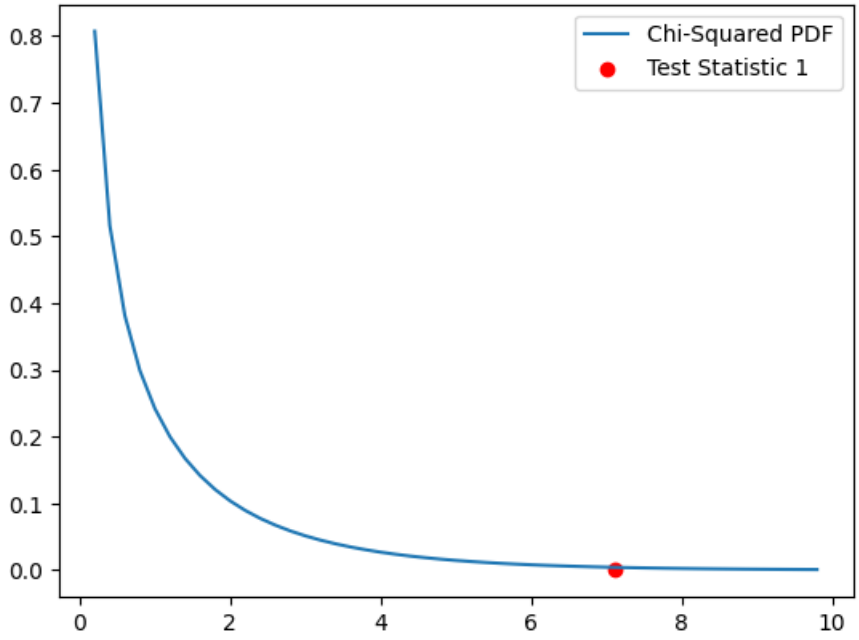
(a) Design and implement a permutation test for the first SNP (column 1 of the genotype matrix G). Plot the observed test statistic T_1 as well as a histogram of the test statistic from $B = 100,000$ permutations.



(b) What is the p-value of T_1 estimated by permutations?

p value is about 0.00747. In other words, the probability of observing that Test Statistic or higher, is less than 1%. We would reject the null, and claim there is some correlation.

(c) 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?



p-value = 0.004243

(d) We would now like to test each of the 10 SNPs for association with the phenotype but want to control the FWER at level 0.05. To do this, we first compute test statistics $(T_1, T_2, \ldots, T_{10})$ for each of the 10 SNPs and then compute p-values $(p_1, p_2, \ldots, p_{10})$ assuming the chi-squared distribution as in part (c). We propose to reject the null hypothesis that SNP i is not associated with phenotype if $p_i < t$. Our goal is to pick t so that the FWER is < 0.05. One approach to control FWER is the Bonferroni procedure. What is the p-value threshold t at which we should reject each of the 10 p-values using the Bonferroni procedure?

Using the Bonferonni correction, the t threshold would be α/m, which in this case would simply be 0.005.