DS BMED 200, Fall 2024
Problem Set 3: Statistics
Due Dec 3, 2024 at 11:59pm PST

Submission instructions

- Submit your solutions electronically on the course Gradescope site as PDF files.
- Submit through the BruinLearn course website under the Gradescope tab on the left. You can add yourself to the course Gradescope site by going to gradescope.com, clicking "Add a course" and entering the following entry code: GPBBKD.
- Please provide short and concise answers. Long, cumbersome, or unclear answers will not be checked.
- If you plan to typeset your solutions, please use the LaTeX solution template. If you plan to submit scanned handwritten solutions, please use a black pen on blank white paper and a high-quality scanner app.

1 Estimating Poisson parameter with MLE [16 pts]

RNA sequencing is a powerful technology widely used for quantifying the number of RNA transcripts (i.e., gene expression) of different genes. Sequencing is a stochastic process where, conceptually, the machine (sequencer) repeatedly traverses the genome hundreds of thousands of times and each time chooses a genomic location at random to produce a "read". The expression activity of a gene can then be approximated by counting the total number of reads that originated from that gene. A natural approach is to model the number of such reads with a Poisson distribution: Poisson(λ), where the rate parameter λ denotes the average expression level of that gene.

Suppose we collected gene expression from n independent samples, and let X_i denote the number of reads mapped to gene A from samples i. Thus, $X_1, X_2, \ldots, X_n \sim$ Poisson(λ). We want to estimate the underlying parameter λ that best explains the data we observe.

(a) [2 pts] Write down the log-likelihood of the data $\ell(\lambda)$.

(b) [4 pts] Calculate the maximum likelihood estimator (MLE) $\hat{\lambda}_{\text{MLE}} = \arg\max_\lambda \ell(\lambda)$

(c) [4 pts] Calculate the bias of the MLE.

(d) [4 pts] Calculate the standard error of the MLE estimator.

(e) [2 pts] Assume we observe the following number of “reads” for gene A in a set of $n = 5$ samples $\{x_1 = 12, x_2 = 0, x_3 = 5, x_4 = 3, x_5 = 0\}$. Calculate $\hat{\text{se}}(\hat{\lambda}_{\text{MLE}})$, the estimated standard error of the MLE.

2 Estimating Poisson parameter with the method of moments [6 pts]

Let $X_1, X_2, \ldots, X_n \sim \text{Poisson}(\lambda)$ be IID random variables.

(a) [2 pts] Use the method of moments to calculate an estimator for λ by matching the first moment.

(b) [4 pts] Use the method of moments to calculate an estimator for λ by matching the **second** moment (i.e., a single equation based on the second moment).

3 Estimating Gaussian parameters [14 pts]

Let $X_1, X_2, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ be IID random variables. Let $\bar{X}$ denote $\frac{1}{n}\sum_{i=1}^{n} X_i$. Recall that the MLEs of the mean and variance are given by $\hat{\mu}_{\text{MLE}} = \bar{X}$ and $\hat{\sigma^2}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

(a) [4 pts] Calculate the expected value of the MLE of the mean: $\mathbb{E}(\hat{\mu}_{\text{MLE}}) = \mathbb{E}(\bar{X})$. Conclude whether the MLE is an unbiased estimator of the mean μ.

(b) [4 pts] Calculate the variance of the MLE of the mean.

(c) [2 pts] Calculate $\mathbb{E}(\bar{X}^2)$. Hint: use the identity $\text{Var}(Y) = \mathbb{E}(Y^2) - \mathbb{E}^2(Y)$

(d) [4 pts] Derive the method of moments estimators of μ and σ^2 by matching the first two moments with the sample moments. Show that $\hat{\sigma^2}_{\text{MoM}}$, the method of moments estimator of σ^2, is equivalent to the MLE counterpart.

4 Hypothesis testing[9 pts]

Suppose we test whether a specific gene is differentially expressed between a patient who contracted a rare disease and healthy individuals (controls). We have collected RNA sequencing data on our patient and a large cohort of controls. Similarly to Q1, we will use the number of sequenced "reads" from individual i, X_i, as an approximation of the expression level of individual i and model it using a Poisson distribution parameterized by λ: Poisson(λ). Suppose we estimated the rate parameter for the group of controls to be $\lambda_0 = 12$.

Our null hypothesis is that the patient who contracted the disease shares the same gene expression distribution as those from the control group (i.e. $H_0 : \lambda_1 = \lambda_0 = 10$). Our alternative hypothesis is $H_1 : \lambda_1 \neq \lambda_0$. A specialist suggests that we reject the null hypothesis if the absolute difference between λ_0 and the observed number of reads from the patient is larger than or equal to 4.

(a) [3 pts] Based on the description of the experiment, what is the test statistic $T(x)$, the critical value c, and the rejection region R?

(b) [3 pts] Assume the gene is not differentially expressed. What is the false positive rate (FPR; i.e., probability of a false positive) of this test? What is the true negative rate (TNR)? Express your answer as a function of F, the cumulative density function of the Poisson distribution.

(c) [3 pts] If $\lambda_1 = 18$, what is the power and the false negative rate (FNR) of this test? Express your answer as a function of F, the cumulative density function of the Poisson distribution.

5 Implementation: testing differentially methylated sites [20 pts]

DNA methylation of the genome is an epigenetic mechanism by which our cells trigger the adjustment of gene expression activity via the addition or removal of a methyl molecule. DNA methylation profiles (i.e., whether a given location in the DNA has a methyl molecule) change across the life course, and the methylated/unmethylated status of methylation sites (CpGs) is often associated with disease states. The methylation status of an individual in a given genomic location is commonly quantified by a ratio between 0 and 1, indicating the fraction of methyl molecules identified in that genomic location out of the total number of DNA copies evaluated for that individual.

We assume that the methylation fractions at a given CpG site are approximately normally distributed across samples. With this assumption, we have simulated the methylation profiles for 10 CpG sites in a control cohort with 100 healthy individuals and a case cohort with 25 patients. Your task is to detect CpG sites that are differentially methylated between conditions (i.e., either elevated or decreased methylation fractions in the case group).

Data provided with this assignment: the samples by features (CpG sites) data matrix for the control group is provided in the file `control.tsv`, and similarly for cases: `case.tsv`. The ground truth parameters used for generating the control group are provided in the file `control_params.tsv`.

(a) [4 pts] Implement the MLE of the mean and variance for data coming from IID normal distribution. Apply these estimators to the methylation profiles from the control group. Report the estimated mean and variance of methylation fractions across samples for each CgG site separately.

(b) [2 pts] Evaluate your mean and variance estimates by calculating the (sample) correlation between the estimates and the ground truth parameters (provided in the file `control_params.tsv`), across all 10 features. Implement the calculation of the sample correlations explicitly (i.e., without using existing correlation functions) and report the correlation scores (one for the means and one for the variances).

(c) [6 pts] We next test the CpG sites for differential methylation between the controls and cases. For each CpG site j under test, we will use the difference between the (estimated) sample mean of methylation fraction in the case cohort and the provided ground truth mean methylation fraction of the control population as our test statistic. Put differently, let x_{ij} be the methylation fraction of individual i in CpG j, our test statistic for CpG j is

$$T(x_j) = \frac{1}{n^{\text{case}}} \sum_i x_{ij}^{\text{case}} - \mu_j \tag{1}$$

where n^{case} is the number of cases and μ_j is the known mean of CpG j of the controls. Our null assumption is that the methylation profiles of the cases and the controls come from the same distribution; thus, no differential methylation between conditions. Under the null, and given our assumption that methylation fractions are normally distributed, one can show that $T(x_j) \sim \mathcal{N}\left(0, \frac{\sigma_j^2}{n^{\text{cases}}}\right)$, where σ_j^2 is the known variance of the controls in CpG j.

Calculate a p-value for each CpG site using the above-mentioned parametric assumption. Report those that are *nominally significant* at level 0.05 (i.e., methylation sites with p-value $<$ 0.05), and report those that are significant at level 0.05 after controlling for multiple hypothesis testing using the Bonferroni correction.

(d) [6 pts] In this section, you will implement permutation testing, a non-parametric approach to derive p-values.

For every given CpG j, implement permutation testing using the same test statistic as in Eq (1); only this time instead of using the known mean of the control group: μ_j, consider the sample estimate of the mean based on the available control samples. Compare the observed test statistics to a null distribution and calculate an *empirical* p-value (i.e., p-value based on permutation testing).

To create a null distribution of the test statistic for a given CpG j, you should first pool all the methylation fractions of CpG j together (i.e., from both the case and the control cohort). Then, consider 10,000 permutations for CpG j. To create a permutation, randomly shuffle the group assignment (i.e., case/control labels); put differently, each methylation fraction of CpG j is randomly assigned with the case/control label of one of the individuals in the data. This will allow you to calculate the test statistic under each permutation and calculate empirical p-values.

Report CpGs with nominally significant p-values at level 0.05, and report CpGs with p-values that are significant at level 0.05 after controlling for multiple hypothesis testing using the Bonferroni correction.

(e) [2 pts] Plot histograms of methylation fractions for the CpG sites that are deemed differentially methylated (after multiple hypothesis correction) in either the parametric testing or the permutation testing. If you observe any inconsistencies between the parametric and non-parametric significant results, speculate what could explain these inconsistencies. (Hint: the difference between the sample median and the sample mean of the methylated fraction in the cases can be informative.)