Arjun Biddanda
Statistical Genetics Reading Group

Problem Set 1
6/1/2025

Probability Distributions (Review)

Discrete Distributions

We want to model the counts of two *independent* processes described by the random variables X and Y, with the following distributions:

$$X \sim Binomial(n, p)$$
$$Y \sim Poisson(\lambda)$$

1. What are the following expected values $\mathbb{E}[X], \mathbb{E}[Y], \mathbb{E}[X+Y]$?
2. What is the probability of $X < Y$ when $n = 10, p = 0.5, \lambda = 5$?
3. What is the *conditional* probability distribution of $\mathbb{P}(X+Y|Y)$?

Continuous Distributions

We have a density function with the following form for $x \in (-\infty, \infty)$:

$$f(x) \propto e^{-x^2}$$

1. Estimate the proportionality constant for the density function to be a valid probability density. You are allowed to use symbolic algebra tools (e.g. `sympy`) for this task if needed (Hint: a probability density should integrate to 1). Numerical answers are also acceptable.
2. What is the expected value of the random variable X with the above density function?
3. What is the variance of X? (Hint: use the relationship $Var(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$)

Likelihoods and P-values

Likelihood of Linear Regression

Linear regression is a central model for relating predictor variables ($\mathbf{X}$) to response variables ($\mathbf{y}$). The N (x_i, y_i) identically and independently distributed (i.i.d) observations are related through the following simplified probability model:

$$y_i \sim \mathcal{N}(x_i\beta, \sigma)$$

Using this definition:

- Write the expression for the log-likelihood of N datapoints in terms of β, σ using the normal density function.
- Using the data provided in `data/pset1/linregA.tsv`, calculate the log-likelihood of the data with $\beta = 0, \sigma = 1$. Calculate the log-likelihood at the value of $\hat{\beta}$ which maximizes the log-likelihood.
- For computational estimation, many approaches use "least-squares", which tries to find the β which minimizes $\sum_i (y_i - x_i\beta)^2$. Calculate the least-squares estimate of β based on the current data. Comment on how this is similar (both mathematically and empirically) to the maximum-likelihood estimator.

Mixture Models

We want to model how p-values are distributed in an experiment, and from some early looks it appears that some portion of them appear to be non-uniformly distributed (like we would expect under the null). You choose to model each p-value as coming from a mixture of beta distributions:

$$\overrightarrow{\mathbf{p}} \overset{\text{iid}}{\sim} \pi_0 \text{Beta}(1, \beta) + (1 - \pi_0)\text{Beta}(1, 1)$$

,
where the fraction π_0 can be treated as an estimate of the proportion of non-uniform p-values (e.g. proportion of rejections).

1. Write down the log-likelihood function for p-values (e.g. $\log(\mathbb{P}(\overrightarrow{\mathbf{p}}|\pi_0, \beta))$)
2. Using the data in `data/pset1/sumstats_testA.tsv`, plot the log-likelihood as a function of π_0 when $\beta = 4$. Provide an estimate of the value of π_0 that maximizes the log-likelihood. (Hint: use `logsumexp` to avoid numerical underflow.)
3. Plot the log-likelihood as a function of π_0 with $\beta = 4$ for the data in `data/pset1/sumstats_testB.tsv`. How does the curve appear relative to the first dataset?
4. ♦♦ The dataset also contains estimates of the standardized effect-size (or Z-score) corresponding to the p-values. Using the following mixture model for the Z-scores:

   $$\overrightarrow{\mathbf{Z}} \overset{\text{iid}}{\sim} \pi_0 \mathcal{N}(0, 1) + (1 - \pi_0)\mathcal{N}(0, \sigma)$$

   repeat the above steps to estimate $\hat{\pi_0}$ using the Z-scores. Comment on the estimated confidence intervals for $\hat{\pi_0}$ (using 2 log-likelihood units) between the effect-size and p-value mixture models (using $\beta = 4, \sigma = 3$ for example).

Multiple Testing Correction and False Discovery Rates

There are two error rates relevant to statistical genetics, the family-wise error rate (FWER) and the false-discovery rate (FDR):

$$FWER = P(R > 0|H_0)$$
$$FDR = P(R/n \leq r|H_0)$$

, where R is a random variable representing the number of rejections of the null hypothesis and n is the total number of tests.

1. Assuming that you are conducting $n = 10^6$ independent hypothesis tests, at a marginal testing threshold of $\alpha = 0.05$ what is the FWER and FDR respectively? (Hint: consider the distribution of null p-values as Uniform and properties of the Bernoulli distribution to show this analytically).
2. At what marginal threshold of α does the FWER drop below 0.05 for the first time? What about the FDR? How is this related to the Bonferroni-threshold of 5×10^{-8}?
3. Plot the FWER and FDR as a function of the marginal p-value threshold α and indicate the Bonferroni threshold on the plot as a vertical line.