Arjun Biddanda
Statistical Genetics Reading Group

Problem Set 1
7/1/2025

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, $\mathbb{E}[X]$, 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: 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:

1. Write the expression for the log-likelihood of N datapoints in terms of β, σ using the normal density function.
2. 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.
3. Calculate the log-likelihood when $\beta = 0$, and evaluate the log-likelihood ratio (Λ). Comment on if the
4. 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.

Multiple Testing Correction and False Discovery Rates

There are two relevant error rates in 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/Binomial 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.

Mixture Models

We want to model how genetic effects β are distributed in an experiment, and from some early looks it appears that some portion of them appear to be much higher or lower than you might expect under the standard $\mathcal{N}(0,1)$. You choose to model each β as coming from a mixture of normal distributions:

$$\overrightarrow{\beta} \overset{\text{iid}}{\sim} (1-\pi_0)\mathcal{N}(0,1) + \pi_0\mathcal{N}(0,\sigma^2)$$

,
where the fraction π_0 can be treated as an estimate of the proportion of effects that are not "null".

1. Write down the log-likelihood function for the observed effects (e.g. $\log(\mathbb{P}(\overrightarrow{\beta}|\pi_0,\beta))$)
2. Using the data in `data/pset1/sumstats_testA.tsv`, plot the log-likelihood as a function of π_0 when $\sigma^2 = 2$. 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 $\sigma^2 = 2$ for the data in `data/pset1/sumstats_testB.tsv`. How does the curve appear relative to the first dataset? Would you estimate there are greater or fewer *non-null* signals in this second dataset?
4. ♦ Fixing the value of π_0 at 0.25, what is the value of σ^2 which maximizes the log-likelihood? Is this different than the assumed value of 2? Plot the two-dimensional log-likelihood surface for all appropriate values of π_0, σ^2 (Note: limit $\sigma^2 < 10$). Comment on the likelihood-surface.
5. ♦♦ Up to now we have assumed that the non-null variance parameter (σ^2) is known, but in practice it is not. If we knew the latent component assignment parameters ($Z_i \in \{0,1\}$) per β_i, we could estimate the following two parameters straightforwardly:

$$\pi_0 = \frac{1}{n}\sum \delta(Z_i = 1)$$
$$\sigma^2 = \frac{1}{N_k}\sum (\beta_i^2)\delta(Z_i = 1)$$

   , where N_k is the number of $Z_i = 1$ out of the whole dataset. One decent guess of Z_i, might be $P(Z_i = 1|\beta_i, \pi_0, \sigma^2)$! Show the form of this probability distribution for a single outcome β_i (Hint: use Bayes Rule ...)
6. ♦♦♦ Using the estimates of $\gamma_i = P(Z_i = 1|\beta_i, \pi_0, \sigma^2)$, we can actually revise the parameter estimates:

$$\pi_0 = \frac{1}{n}\sum_i \gamma_i$$
$$\sigma^2 = \frac{1}{N_k}\sum (\beta_i^2)\delta(Z_i = 1)$$

   , where $N_k = \sum_i^n \gamma_i$. Starting from initial parameter guesses of $\pi_0 = 0.5, \sigma^2 = 5$, calculate what the updated parameters (π_0‘, σ^{2}‘) would be (using the initial guesses to calculate γ_i for `data/pset1/sumstats_testA.tsv`). Calculate the log-likelihood using the initial parameters and the updated parameters, which one is higher? What if you repeat the process one more step (using π_0‘, σ^{2}‘ to re-calculate γ_i and update the parameters again), does the likelihood increase?[1]

[1] If you've gotten to this step, you've implemented your first EM-algorithm!