CSE 312: Foundations of Computing II

Section 9: MAP, Hypothesis Testing, Confidence Intervals

1. Posterior

Let $\mathbf{x} = (x_1, \ldots, x_n)$ be iid samples from $Exp(\Theta)$ where Θ is a random variable (not fixed).

(a) Using the prior $\Theta \sim Gamma(r, \lambda)$ (for some arbitrary but known parameters $r, \lambda > 0$), show that the posterior distribution $\Theta|\mathbf{x}$ also follows a Gamma distribution and identify its parameters (by computing $\pi_\Theta(\theta|\mathbf{x})$). Then, explain this sentence: "The Gamma distribution is the conjugate prior for the rate parameter of the Exponential distribution". Hint: This can be done in just a few lines!

(b) Now derive the MAP estimate for Θ. The mode of a $Gamma(s, \nu)$ distribution is $\frac{s-1}{\nu}$. Hint: This should be just one line using your answer to part (a).

(c) Explain how this MAP estimate differs from the MLE estimate (recall for the Exponential distribution it was just the inverse sample mean $\frac{n}{\sum_{i=1}^n x_i}$, and provide an interpretation of r and λ as to how they affect the estimate.

2. Do you have the confidence?

Imagine you are polling a population to estimate the true proportion p of individuals that support putting pineapple on pizza. You do this by sampling n people from the population with replacement and asking whether they do or don't (these are the only two choices) and using the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ as your estimate for p (where each X_i is 1 if person i supports putting pineapple on pizza, and 0 otherwise). At least how many samples n do you need to perform such that 98% of the time, the estimate $\bar{X}$ is within 5% of the true p?

3. Tree Hypothesis

Suppose you live on a tree farm with a large field. You've always used Fertilizer Y, but your friend recently recommended you to use Fertilizer X. You plant 545 trees, and you give $n = 254$ of them Fertilizer X and $m = 291$ Fertilizer Y, and measure their height after three years.

Now you have iid samples (assume trees grow independently) $x_1, x_2, ..., x_n$ which measure the height of the n trees given fertilizer X, and iid samples $y_1, y_2, ..., y_m$ which measure the height of the m trees given fertilizer Y. The data you are given has the following statistics:

Fertilizer	Number of samples	Sample Mean	Sample Variance
X	$n = 254$	$\bar{x} = 6.99$	$s_x^2 = 28.56^2$
Y	$m = 291$	$\bar{y} = 4.21$	$s_y^2 = 23.97^2$

Perform a hypothesis test using the procedure in 8.4, and report the exact p-value for the observed difference in means. In other words: assuming that the heights of trees which had been given fertilizer X and fertilizer Y has the same mean μ_X, μ_Y, what is the probability that you could have sampled two groups of tress such that you could have observed that the difference of means between Fertilizer Y and Fertilizer X was as extreme, or more extreme, than the one observed (which is $\bar{x} - \bar{y} = 2.78$)?