CSE 312: Foundations of Computing II

Section 9: MAP, Hypothesis Testing, Confidence Intervals Solutions

1. Posterior

Let $\mathbf{x} = (x_1, \dots, x_n)$ 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 $\dfrac{s-1}{\nu}$. 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 $\dfrac{n}{\sum_{i=1}^n x_i}$, and provide an interpretation of r and λ as to how they affect the estimate.

Solution:

(a) Remember the posterior is proportional to likelihood times prior:

$$\begin{aligned}
\pi_\Theta(\theta|x) &\propto L(x|\theta)\pi_\Theta(\theta) && \text{[def of posterior]}\\
&= \left(\prod_{i=1}^n \theta e^{-\theta x_i}\right) \cdot \frac{\lambda^r}{(r-1)!}\theta^{r-1}e^{-\lambda\theta} && \text{[def of likelihood, Gamma pdf]}\\
&\propto \theta^n e^{-\theta\sum x_i}\theta^{r-1}e^{-\lambda\theta} && \text{[algebra]}\\
&= \theta^{(n+r)-1}e^{-(\lambda+\sum x_i)\theta}
\end{aligned}$$

Therefore $\Theta|\mathbf{x} \sim Gamma(n+r, \lambda + \sum x_i)$, since the final line above is proportional to the pdf for the gamma distribtution.

(b) Just citing the mode of a Gamma, we get

$$\frac{n+r-1}{\sum x_i + \lambda}$$

(c) We see how the estimate changes from the MLE: pretend we saw $r-1$ extra events over λ units of time. (Instead of waiting for n events, we waited for $n+r-1$, and instead of $\sum x_i$ as our total time, we now have $\lambda + \sum x_i$ units of time).

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?

Solution:

First, we define the probability of a "bad event". In this case, that means that $\bar{X}$ deviates from p by 0.05 or more. Thus, we write this as:

$$\mathbb{P}(|\bar{X} - p| > 0.05)$$

To use the CLT to approximate this, we first have to find the expected value and variance of $\bar{X}$. We do this by leveraging that the sum of the X_is is distributed according the binomial distribution. Thus:

$$\mathbb{E}[X] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}np = p$$

We find the variance similarly:

$$Var(\bar{X}) = Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}Var(\sum_{i=1}^{n} X_i) = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}$$

Thus, we can approximate $\bar{X}$ using the CLT:

$$\bar{X} \sim N(p, \frac{p(1-p)}{n})$$

We then standardize the earlier statement. We only have to divide both sides by the standard deviation, because $\bar{X}$ already has its mean (p), being subtracted from it:

$$\mathbb{P}(|\bar{X} - p| > 0.05) = \mathbb{P}(\frac{|\bar{X} - p|}{\sqrt{\frac{p(1-p)}{n}}} > \frac{0.05}{\sqrt{\frac{p(1-p)}{n}}}) = \mathbb{P}(|Z| > \frac{0.05\sqrt{n}}{\sqrt{p(1-p)}})$$

Since p is a probability, the value of $p(1-p)$ is at most $\frac{1}{4}$. We can use this to upper bound our probability (this is fine since ultimately we just need *at least* how many samples we need, i.e. we need a lower bound on n). Thus:

$$\frac{0.05\sqrt{n}}{\sqrt{p(1-p)}} \geq \frac{0.05\sqrt{n}}{\sqrt{\frac{1}{4}}} = 0.1\sqrt{n}$$

This means we can say:

$$\mathbb{P}(|Z| > \frac{0.05\sqrt{n}}{\sqrt{p(1-p)}}) \leq \mathbb{P}(|Z| > 0.1\sqrt{n})$$

We want this probability to be lower than 0.02, since we want a 98% confidence interval. We can then break it up using the absolute value, to get our probability in a place where can use the Z-table:

$$\mathbb{P}(|Z| > 0.1\sqrt{n}) = \mathbb{P}(Z > 0.1\sqrt{n}) + \mathbb{P}(Z < -0.1\sqrt{n}) < 0.02$$

Due to the symmetry of the Normal, this becomes:

$$\mathbb{P}(Z > 0.1\sqrt{n}) + \mathbb{P}(Z < -0.1\sqrt{n}) = 2\mathbb{P}(Z > 0.1\sqrt{n}) = 2(1 - \mathbb{P}(Z \leq 0.1\sqrt{n})) = 2(1 - \Phi(0.1\sqrt{n})) < 0.02$$

We then rearrange the equation further:

$$2(1 - \Phi(0.1\sqrt{n})) < 0.02 \rightarrow 1 - \Phi(0.1\sqrt{n}) < 0.01 \rightarrow 0.99 < \Phi(0.1\sqrt{n})$$

Using the Z-table, we then see that the input to Φ that satisfies this is ≥ 2.33, so we can solve for n:

$$0.1\sqrt{n} \geq 2.33 \rightarrow \sqrt{n} \geq \frac{2.33}{0.01} \rightarrow n \geq 543$$

Thus, we need to sample at least 543 times from the population to get an estimate $\bar{X}$ of the proportion p that is within 5% 98% of the time.

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 $\bar{x} - \bar{y} = 2.78$)?

Solution:

Our null and alternative are (since your friend claims that Fertilizer X is better than Y):

$$H_0 : \mu_X = \mu_Y \qquad H_A : \mu_X > \mu_Y$$

Let's choose our significance level $\alpha = 0.05$. By the CLT, $\bar{X} \sim \mathcal{N}(\mu_x, s_x^2/n)$ and $\bar{Y} \sim \mathcal{N}(\mu_y, s_y^2/m)$. By closure properties of the normal and our null hypothesis (under this, $\mu_X = \mu_Y \to \mu_X - \mu_Y = 0$),

$$\bar{X} - \bar{Y} \sim \mathcal{N}\left(\mu = 0, \sigma^2 = \frac{28.56^2}{254} + \frac{23.97^2}{291} = 5.18575\right)$$

Then, we are asking

$$\begin{aligned} P(\bar{X} - \bar{Y} \geq \bar{x} - \bar{y}) &= P\left(\frac{(\bar{X} - \bar{Y}) - (\mu_{\bar{X}} - \mu_{\bar{Y}})}{\sqrt{5.18575}} \geq \frac{(\bar{x} - \bar{y}) - (\mu_{\bar{X}} - \mu_{\bar{Y}})}{\sqrt{5.18575}}\right) \\ &= P\left(\frac{(\bar{X} - \bar{Y}) - 0}{\sqrt{5.18575}} \geq \frac{(2.78) - 0}{\sqrt{5.18575}}\right) \\ &= P(Z \geq 1.22) = 0.1112 \end{aligned}$$

Since our p-value of 0.1112 is $> \alpha = 0.05$, we fail to reject the null hypothesis. There is insufficient evidence to show that Fertilizer X is better than Fertilizer Y.