STA 250: Statistics

Notes 13. Classical Interval Reporting: Confidence Intervals

Book chapters: 8.5

1 Cross-connecting classical testing and Bayesian prediction

In the first two modules, we got introduced to the classical and the Bayesian approach to statistical inference. Not only did we learn about the differences between the two, we saw them being applied to two disjoint inference problems: classical inference for hypotheses testing and Bayesian inference for prediction. In a certain sense, we saw the two approaches in their natural playing fields (i.e., contexts which contributed most to the philosophical development of the two approaches).

We saw that the guiding light for classical approach is "performance under repeated experiments". For classical hypothesis testing, we construct a test rule by choosing a test statistic and a cut-off so that we can control the size of the rule (maximum type I error probability) while ensuring large power at the alternative (small type II error probabilities). Testing is carried out by applying a "good rule" to the data at hand.

The Bayesian mantra is "quantifying uncertainty", which begins with a prior pdf/pmf to describe pre-data uncertainty that is either context based or has some mathematical notion of being uniform or non-informative. The data at hand, through Bayes rule, updates the prior belief to posterior belief. Prediction of a future observable is made by producing a range that reflects large certainty under the posterior (predictive). The same idea applies to produce a (posterior) range for a model parameter.

In this week's lectures we will probe the classical theory of producing a range for a parameter or a future observable and the Bayesian theory of hypotheses testing. We will see that when subjected to the un-natural playing field, either approach will be faced with some additional issues, but will eventually overcome them while retaining their signature flavor. How these issues are overcome will provide better insight into the foundations of classical and Bayesian thinking.

2 Classical theory of producing a range for parameter

For today's class we will look at producing a range of values for a model parameter or a future observable under the classical approach. We will treat these two cases separately, starting with the task of producing a range for parameter.

Consider the model $X \sim f(x|\theta)$, $\theta \in \Theta$ and a future observable modeled as $X^* \sim f^*(x^*|\theta)$. In order to report an interval for θ, the classical approach will require a "rule of interval construction", i.e., a rule or algorithm that will take any plausible data x as input and produce an interval $A = A(x)$ of Θ as output.

Example (Normal model)**.** Consider $X = (X_1, \cdots, X_n)$ modeled as $X_i \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu, \sigma^2)$, $\mu \in (-\infty, \infty)$, σ known. An interval rule for μ is "for input x produce output $A(x) = \bar{x} \mp 1.96\sigma/\sqrt{n}$".

A trivial interval rule is "for input x produce output $A(x) = \Theta$", i.e., always report the whole parameter space Θ as the plausible range for θ. With this rule, you never make a wrong statement. But it's a useless rule, because you make no use of the data to narrow down the range of θ from Θ to a smaller subset. From classical viewpoint, a good interval rule is one that produces an interval $A(x)$ containing the true value of θ with large probability but is small in size relative to the whole parameter space.

Definition 1 (Coverage and Confidence). For any interval rule "for input x produce $A(x)$" its coverage at any $\theta \in \Theta$ is defined as $P_{[X|\theta]}(\theta \in A(X))$. Its confidence coefficient is defined as its minimum coverage across all $\theta \in \Theta$, i.e.,

$$\text{confidence coefficient} = \min_{\theta \in \Theta} P_{[X|\theta]}(\theta \in A(X)).$$

A rule with 95% confidence coefficient is referred to as a 95% confidence interval rule; a rule with 99% confidence coefficient is referred to as a 99% confidence interval rule and so on.

Definition 2 (Efficiency). Suppose two interval rules have the same confidence coefficient. Then one rule is said to be more efficient than the other if its average size is smaller than the average size of the other for every θ.

Example (Normal model (contd.))**.** The interval rule with $A(x) = \bar{x} \mp 1.96\sigma/\sqrt{n}$ has 95% confidence coefficient. This is because, for any μ,

$$P_{[X|\mu]}(\mu \in \bar{X} \mp 1.96\sigma/\sqrt{n}) = P_{[X|\mu]}\left(-1.96 \le \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \le 1.96\right) = 0.95,$$

i.e., the rule has coverage 95% at every value of μ, and hence the minimum coverage is also 95%.

Example (Normal model (contd.))**.** The interval rule with $A(x) = \bar{x} \mp z(\alpha)\sigma/\sqrt{n}$ has $100(1-\alpha)$% confidence. Same calculation as before.

Example (Normal model (contd.))**.** The rule with $A(x) = x_{\text{med}} \mp z(\alpha)\sigma\sqrt{\pi/(2n)}$ has confidence (approximately) $100(1-\alpha)$% (for large n) [Homework 2 #5]. The average size of the mean based rule is $2z(\alpha)\sigma/\sqrt{n}$ whereas the average size of the median based rule is $2z(\alpha)\sigma\sqrt{\pi/(2n)} = 2z(\alpha)\sigma/\sqrt{n} \times \sqrt{\pi/2}$. Because $\pi > 2$, the median based rule has a bigger average size than the mean based rule. So the mean based rule is more efficient.

3 ML interval rules: confidence and efficiency

We did ML testing by first deriving ML intervals $A_b(x)$ of the form:

$$A_b(x) = \{\theta : L_x(\theta) \ge bL_x(\hat{\theta}_{\text{MLE}}(x))\}.$$

An ML interval rule is "for input x produce $A_b(x)$ where a cut-off $b > 0$ has been preselected. The confidence coefficient of this rule depends on the choice of b. In Notes 6, we went through the same calculations that are needed to calculate the confidence coefficient for such rules and to choose b to guarantee an ML interval rule of a given confidence. The following table gives the results for some standard models.

Model	$100(1-\alpha)\%$ ML confidence interval
$X = (X_1, \cdots, X_n)$ $X_i \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu, \sigma^2)$ σ known	$\bar{x} \mp z(\alpha)\frac{\sigma}{\sqrt{n}}$
$X = (X_1, \cdots, X_n)$ $X_i \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu, \sigma^2)$ both unknown	$\bar{x} \mp z_{n-1}(\alpha)\frac{s_x}{\sqrt{n}}$
Regular models $X \sim f(x\|\theta)$ (binomial, Poisson, exponential)	$\hat{\theta}_{\text{MLE}}(x) \mp z(\alpha)/\sqrt{I_{\text{OBS}}(x)}$

It turns out that for many standard models, an ML interval rule of a given confidence is more efficient than any other interval rule with the same confidence. There is some theory that points out that all good interval rules are constructed by first finding a "good estimator" $T(X)$ of θ. An estimator is simply a summary statistic of data, and a good estimator is one satisfying $\sqrt{n}(T(X) - \theta) \approx \mathsf{Normal}(0, se^2(\theta))$ for some constant $se(\theta)$ at every $\theta \in \Theta$. One then looks for a "consistent" estimator $S(X)$ of $se(\theta)$, i.e., another summary statistic such that $S(X) \approx se(\theta)$ for large n. $100(1-\alpha)\%$ confidence interval rules are then constructed with intervals $A(x) = T(x) \mp z(\alpha) \cdot S(x)/\sqrt{n}$, with size $\approx 2z(\alpha)se(\theta)/\sqrt{n}$. There is also some general theory suggesting that for good estimates of the above kind, $se^2(\theta) \geq 1/I_1^F(\theta)$, the expected Fisher information at θ (Notes 9). ML interval rules use $T(X) = \hat{\theta}_{\text{MLE}}(X)$ which usually satisfies $\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta) \approx \mathsf{Normal}(0, 1/I_1^F(\theta))$ and a consistent estimator of $1/I_1^F(\theta)$ is $S(X) = n/I_{\text{OBS}}(X)$. So ML interval rules are generally the most efficient.

Example (NSW)**.** Consider the NSW study where earning differences $X = (X_1, \cdots, X_n)$ are modeled as $X_i \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu, \sigma^2)$, both unknown. Recorded data show $\bar{X} = 4253.57$, $s_X^2 = 8926.985$ and $n = 185$. So, a 95% ML confidence interval for μ is $4253.57 \mp z_{184}(0.05) \cdot 8926.985/\sqrt{185} = [2958.68, 5548.46]$.

It is important to remember that confidence is a performance guarantee across repeated trials (just like the size of a test). In the NSW study, a 95% confidence interval rule applies to the data at hand produces the interval [2958.68, 5548.46], which is colloquially referred to as a 95% interval for μ. But it is wrong to interpret this analysis as saying there is 95% chance that $\mu \in p2958.68, 5548.46]$. Such statements cannot be evaluated under the classical theory.

4 Classical theory of producing a range for a future observable

Conceptually this is same as producing a range for a model parameter, but some practical issues arise. Consider the model $X \sim f(x|\theta)$, $\theta \in \Theta$ and a future observable $X^* \sim f^*(x^*|\theta)$. Let S^* denote the set of values X^* can realize. A prediction interval rule is "for input x produce a subset $A^* = A^*(x) \subset S^*$". From classical viewpoint one would like $A^*(x)$ to have small average size at all θ while $P_{[X,X^*|\theta]}(X^* \in A^*(X))$ should be large across all θ.

Unfortunately, there is no universal theory on constructing good prediction interval rules. The general strategy is as follows. If we knew the value of θ, we could produce a 95% range for X^* from the corresponding model pdf/pmf $f^*(x^*|\theta)$. Call this range $S^*(\theta)$. Because we do not know θ, we instead plug-in an estimate $\hat{\theta}(x)$ of it based on data $X = x$. So our plug-in interval rule is: $A^*(x) = S^*(\hat{\theta}(x))$. Standard choice is to use the mle $\hat{\theta}_{\text{MLE}}(x)$ to plug in for θ. While this looks reasonable, the plug-in rule does not quite guarantee a 95% probability of capturing X^*. Here is an example.

Example (Normal model)**.** Consider $X = (X_1, \cdots, X_n)$ and X^* modeled as: $X_1, \cdots, X_n, X^* \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu, \sigma^2)$, σ known. If we knew μ, then a 95% interval for X^* is $\mu \mp 1.96\sigma$. So the 95% ML plug-in rule uses the interval $A^*(x) = \bar{x} \mp 1.96\sigma$, by plugging in $\hat{\mu}_{\text{MLE}}(x) = \bar{x}$ for μ. But for any μ, $P_{[X,X^*|\mu]}(X^* \in \bar{X} \mp 1.96\sigma)$ is strictly smaller than 95%. To see this, notice that $X^* - \bar{X} \sim \mathsf{Normal}(0, \sigma^2(1 + 1/n))$ and consequently the above probability is precisely $2\Phi(1.96 \cdot \sqrt{n/(n+1)}) - 1$ which is smaller than $2\Phi(1.96) - 1 = 0.95$ because $\sqrt{n/(n+1)} < 1$.

The problem with the plug-in rule is that it does not account for the uncertainty in plugging-in $\hat{\theta}_{\text{MLE}}(x)$ for θ. A little more calculation shows that for the normal model above, the correct 95% confidence prediction interval rule should use $A^*(x) = \bar{x} \mp 1.96\sigma\sqrt{1 + 1/n}$. Such fixes are available for normal models. But it is difficult to fix the problem for other models. Not accounting for the uncertainty about θ poses lesser problems when n is large, so the plug-in intervals are OK asymptotically.