Stat260: Bayesian Modeling and Inference **Lecture Date: March 8, 2010**

Confidence Intervals and Hypothesis Testing

Lecturer: Michael I. Jordan *Scribe: Charlie Gibbons*

1 Confidence intervals

Confidence intervals play important roles in both frequentist and Bayesian statistics. As we often do in this course, we first consider the frequentist approach and contrast that of the Bayesians.

1.1 Frequentist interpretation

As always, frequentists assume that there exists a singular true value of some unknown parameter θ. As a result, it is not logical to provide a distribution for its value. Additionally, for a given set of data, the confidence interval is fixed. Either the true θ is in this interval or it is not; the probability is trivially 1 or 0 if the true value is or is not contained by this interval (a probability, of course, that we can never observe). The probability of a confidence interval arises by considering repeated intervals, constructed using different data sources; the data becomes random and induces a probability on the intervals constructed in a prescribed manner.

The $100(1-\alpha)\%$ confidence interval $C(X)$, where X is a random set of data, is constructed such that

$$P_\theta\left(\theta \in C(X)\right) \geq 1-\alpha. \tag{1}$$

This interval is said to have *coverage* $1-\alpha$. It should be noted that $C(x)$ is not unique, but is typically chosen such that it is the most compact set satisfying this inequality.

1.2 Bayesian approach

Frequentist confidence intervals are based upon the notion of repeated trials and precise interpretation and description of these intervals is unintuitive and cumbersome; *e.g.*, intervals constructed in this manner are expected to contain the true parameter 95% of the time. Note that this interpretation doesn't answer the actual question of interest—what is the probability that the parameter estimated from a given data set lies in this range?

Bayesian confidence intervals do allow this natural interpretation; indeed, once a posterior distribution $\pi(\theta|x)$ has been found, a confidence interval $C(x)$ is constructed such that

$$\pi\left(\theta \in C(x)|x\right) \geq 1-\alpha, \tag{2}$$

where θ is taken to be random, as opposed to X in the frequentist framework. This process is comparatively simple because Bayesians naturally create a distribution of θ in its own space. As above, this interval is not unique; the region is often chosen such that it contains the most likely regions of the posterior distribution, perhaps by sacrificing contiguity. The *highest posterior density region* (HPD) is chosen to satisfy

$$C(x) = \left\{\theta:\ \pi(\theta|x) > K_\alpha\right\}, \tag{3}$$

where K_α is chosen such that this region satisfies Equation 2. Though this is a pleasing means to produce a unique confidence interval, this region is often difficult to calculate in practice.

We might wonder whether Bayesian confidence intervals also exhibit the coverage qualities of their frequentist counterparts.

2 An aside: Metaprinciples in statistics

There are two metaprinciples in statistics: *coherence* and *calibration*.

Coherence is, loosely, being consistent in methodology. For example, if you use the same data set, but perform your analysis in a different way, your results should be similar. This principle has guided Bayesian analysis; consider the relationship between this concept and the likelihood principle. Statisticians in this camp often criticize their frequentist brethren for failing to reach this standard. Traditional p-values, for example, are subject to "Dutch book" critiques, which state that a series of bets against these inferences have positive expected value.

Of course, you can be perfectly coherent, but perfectly wrong; simply answer "24" to all questions that you confront. To this end, frequentists rely upon calibration to assess the accuracy of their inferences—over repeated iterations, how well do these inferences fare? Confidence intervals are subject to the Dutch book claims along with p-values (as the two are intimately linked), but are calibrated in that, over repeated samples, they do attain correct coverage. Bayesian confidence intervals may not have correct coverage in the frequentist sense; they are coherent, but not calibrated. As with all frequentist analyses, calibration focuses on procedure and chance mechanisms, rather than underlying probability distributions.

How might we reconcile these two principles? A good approach may be to perform Bayesian analysis followed by frequentist calibration checks.

3 Hypothesis testing

We formulate a null hypothesis H_0 and an alternative H_a regarding $\theta \in \Theta$ according to

$$\begin{aligned} H_0 &: \theta \in \Theta_0 \\ H_a &: \theta \in \Theta_a, \end{aligned}$$

where $\Theta_0 \cup \Theta_a = \Theta$.

3.1 Frequentist approach

Again, we begin with the frequentist approach as a comparison. Here, we'll follow a decision-theoretic approach, which also could have been applied to the case of confidence intervals, though it is not especially revealing in that context.

We use *zero-one loss*, under which the loss function is defined as

$$l(\theta, \delta) = \begin{cases} \delta & \text{if } \theta \in \Theta_0, \\ 1 - \delta & \text{if } \theta \in \Theta_a, \end{cases}$$

where $\delta \in \{0, 1\}$ is the decision that is made; $\delta = 0$ implies that the null hypothesis is not rejected and it equals 1 if the null is rejected.

We calculate the *frequentist risk*

$$\begin{aligned} R(\theta, \delta) &= \int l(\theta, \delta(x)) P(x|\theta)\, dx \\ &= \begin{cases} P_\theta(\delta(X) = 1) & \text{if } \theta \in \Theta_0 \quad \text{(Type I Error)}, \\ P_\theta(\delta(X) = 0) & \text{if } \theta \in \Theta_a \quad \text{(Type II Error)}. \end{cases} \end{aligned}$$

Recall that frequentists condition upon the true value of θ and consider various possibilities for X. A Type I Error occurs when you reject the null though it is true; a Type II Error arises when you fail to reject a false null hypothesis. Neyman and Pearson proposed fixing the Type I Error at some level α and finding the test that minimized Type II Error or, equivalently, maximized power.

3.2 Bayesian approach

We continue to use zero-one loss to evaluate our null hypothesis. In the Bayesian context, however, we integrate the risk over θ, rather than X:

$$\begin{aligned} &\int l(\theta, \delta) \pi(\theta|x)\, d\theta \\ &= \begin{cases} \pi(\theta \in \Theta_a|x) & \text{if } \delta = 0. \\ \pi(\theta \in \Theta_0|x) & \text{if } \delta = 1, \end{cases} \end{aligned}$$

The optimal decision rule becomes

$$\delta^*(x) = \begin{cases} 0 & \text{if } \pi(\theta \in \Theta_0|x) \geq \pi(\theta \in \Theta_a|x), \\ 1 & \text{if } \pi(\theta \in \Theta_0|x) < \pi(\theta \in \Theta_a|x). \end{cases}$$

Simply put, you choose the more probable hypothesis under the posterior distribution of θ. Note that this implies that you always reject a point null, which has probability 0.

4 Bayes factors

Let's consider the ratio of the posterior probabilities of the null and alternative hypotheses:

$$\underbrace{\frac{\pi\,(H_0|x)}{\pi\,(H_a|x)}}_{\text{Posterior odds}} = \underbrace{\frac{P(x|H_0)}{P(x|H_a)}}_{\text{Bayes factor}} \underbrace{\frac{P(H_0)}{P(H_a)}}_{\text{Prior odds}}.$$

The *marginal likelihood* is defined by

$$P(x|H_i) = \int P(x|\theta,\, H_i) P(\theta|H_i)\, d\theta, \tag{4}$$

where $H_i \in \{H_0,\, H_a\}$.

The *Bayes factor* is a summary of the evidence from the data regarding the null. It is separate from the priors used for the null and alternative hypotheses, making it a more objective assessment of the likelihoods of the hypotheses. Note that the Bayes factor does use the "subjective" assessment of the prior of the hypotheses on the parameter space, $P(\theta|H_i)$ in the marginal likelihood above, but it does remove the prior odds of the hypotheses themselves. By reporting the Bayes factor, the reader can use his own prior odds ratio of the hypotheses to compute the posterior odds for that case.

Jeffreys provided guidelines for evaluating the null hypothesis using the Bayes factor. Specifically, he proposes the following interpretations for the $\log_{10}$ of Bayes factors falling within certain ranges:

Between:	0.0 and 0.5	Weak evidence against the null
	0.5 and 1.0	Substantial evidence against the null
	1.0 and 2.0	Strong evidence against the null
	2.0 and above	Decisive evidence against the null.

In calculating the Bayes factor, we must consider what prior to use for θ. Unfortunately, flat or non-informative priors do not work in this situation. To see this, note that under flat priors, the Bayes factor becomes

$$\frac{c_0 \int P(x|\theta)\pi_0(\theta)\, d\theta}{c_a \int P(x|\theta)\pi_a(\theta)\, d\theta}, \tag{5}$$

where $\frac{c_0}{c_a}$ is an arbitrary constant that illustrate that the Bayes factor is an arbitrary result. As an example, a prior that is flat in the null space could cover fewer dimensions than a flat prior over the space of the alternative hypothesis. It may be appropriate to use a flat prior if θ is "the same" under both hypotheses; σ^2 across two regressions with restrictions on β, for example.

5 Linear regression

We now turn to linear regression. The standard model is

$$Y = X\beta + \epsilon, \tag{6}$$

where Y and ϵ are N dimensional column vectors, X is an $N \times K$ matrix, and β is a K dimensional column vector of parameters. The error ϵ is assumed to have a $N(0,\, \sigma^2 I)$ distribution.

A flat prior on β is fine for estimation and it returns the MLE estimator. Unfortunately, this procedure does not provide desirable properties for hypothesis testing and confidence intervals. Conjugate priors are a natural alternative, but this would require forming an appropriate $K \times K$ covariance matrix. Instead, we use the *g prior*:

$$\begin{aligned} \beta &\sim N\left(\beta_0,\, g\sigma^2 \left(X'X\right)^{-1}\right) \\ \pi_J\left(\sigma^2\right) &\propto \frac{1}{\sigma^2}, \end{aligned}$$

with scalars σ^2, a nuisance parameter taking the Jeffreys' prior, and g. As we shall see, a choice of g that is independent of N yields poor properties, as does an empirical Bayes approach. Instead, we put a prior on g, leading to *hierarchical models*.