Bayesian Estimation

STAT422 Al Parker

March 3, 2017

1 What's under the hood when using Bayesian methods?

The objective of Chapter 16 is to demonstrate the Bayesian approach to statistical inference, which can be summed up in a word: *posterior*.

Bayesian statistical conclusions about k parameters of interest, $\theta \in \Re^k$, are made in terms of probability statements which are conditional on the observed value of $y \in \Re^n$, the data [2, 4]. This probability is called the *posterior density*, written $p(\theta|y)$. The posterior is found by using the definition of conditional probability,

$$p(\theta|y_1, ..., y_n) = \frac{p(y_1, ..., y_n, \theta)}{p(y_1, ..., y_n)}.$$

Applying this definition once more yields *Bayes Rule*,

$$p(\theta|y_1, ..., y_n) = \frac{p(y_1, ..., y_n|\theta)p(\theta)}{p(y_1, ..., y_n)}, \tag{1}$$

interpreted as the distribution of the parameters given the data. The *likelihood* $p(y_1, ..., y_n|\theta) = L(y_1, ..., y_n|\theta)$ is the probability model which one assumes yielded the data $y_1, ..., y_n$ actually observed. When the data are a SRS, then the likelihood is

$$p(y_1, ..., y_n|\theta) = \Pi_{i=1}^n p(y_i|\theta).$$

The *prior*, $p(\theta)$, is the density over the population of possible values for θ based on the modeler's knowledge of θ. The marginal density of $y_1, ..., y_n$ is $p(y_1, ..., y_n) = \int p(y_1, ..., y_n, \theta)d\theta = \int p(y_1, ..., y_n|\theta)p(\theta)d\theta$.

Equation (1) is the engine driving the Bayesian perspective. This equation illuminates the three main differences between the Bayesian perspective and the non-Bayesian or *frequentist* perspective.

- From a Bayesian point of view, the parameter vector θ is assumed to be random. This is very different than the frequentist approach taught in many elementary statistics classes, where the parameter θ is assumed to be a fixed, unknown quantity.

- It can be argued that it is the perceived arbitrariness of the prior distribution imposed upon θ that truly divides statisticians into two camps. Although hard-core Bayesians would maintain that all priors ought to be built on knowledge about the parameters, it is convenient to choose a *conjugate prior*, which yields a posterior of a known parametric type, explained below. On the other hand, non-Bayesians may agree (grudgingly) only with the use of *non-informative priors*.

- The interpretation of even markedly similar numerical results are different depending on the perspective of the statistician.

2 Calculating the posterior

Using equation (1) to calculate the posterior, two components need to be specified in order to (attempt to) calculate the posterior: the likelihood $p(y_1, ..., y_n|\theta)$, which you are already familiar with (e.g. when finding sufficient statistics or MLEs); and the prior $p(\theta)$. In this section, we'll look at some common choices for a prior, and then explain why the marginal $p(y_1, ..., y_n)$ (usually) does not need to be explicitly calculated.

Note that for the same data set generated by the same likelihood model, two different analysts can get different posteriors by using different priors.

2.1 The marginal $p(y_1, ..., y_n)$ does not need to be calculated

In equation (1), notice that

$$p(\theta|y_1, ..., y_n) \propto p(y_1, ..., y_n|\theta)p(\theta),$$

the posterior is proportional to the product of the likelihood and the prior. The marginal

$$p(y_1, ..., y_n) = \int p(y_1, ..., y_n, \theta)d\theta = \int p(y_1, ..., y_n|\theta)p(\theta)d\theta$$

is a constant with respect to θ. Thus, it is a normalizing constant (so that the posterior is a valid pdf or pmf), and does not need to calculated directly. This simplifies the calculation of the posterior tremendously. Section 16.5 of your textbook [5] discusses this more.

2.2 Choosing a prior $p(\theta)$

2.2.1 Noninformative prior

When the modeler has little prior knowledge about θ, then it is prudent to use a flat or non-informative distribution for $p(\theta)$. For example, for values of $\theta \in [0, 4]$, a noninformative prior is $p(\theta) = \frac{1}{4}$. Using non-informative priors can yield parameter estimates that are similar to the estimates obtained by non-Bayesian methods.

Choosing a noninformative prior over an infinite domain results in an *improper* prior since the integral over the domain is infinite. For example, if

$$p(\theta) = 1 \text{ for } \theta \in (0, \infty) \text{ then } \int_0^\infty p(\theta)d\theta = \int_0^\infty 1d\theta = \infty.$$

Remarkably, using equation (1) to calculate the $p(\theta|y)$ (usually) still results in a posterior which is a valid pdf or pmf.

2.2.2 Conjugate priors

A conjugate prior (see Definition 16.1 in your textbook [5]) is one where the posterior distribution is of the same distribution type or class as the distribution type of the prior. See the examples given in sections 4.1.2 and 4.2.2 in these notes. A critic might point out that such a choice of prior is due to convenience (i.e. to get an analytic expression of the posterior that is well behaved), not necessarily due to a priori knowledge of the data.

2.2.3 Priors from previous sets of data

A satisfactory way to obtain a prior distribution that is driven by data is to set a prior equal to the posterior found in a previous data analysis. Let $p(\theta|x_1, ..., x_n)$ be the posterior given that the data $x_1, ..., x_n \sim p(x_1, ..., x_n|\theta)$ were observed, and suppose that a new data vector $y_1, ..., y_n$ is observed, such that $y_1, ..., y_n \sim p(y_1, ..., y_n|\theta)$. The posterior $p(\theta|y_1, ..., y_n)$ is found by setting $p(\theta) := p(\theta|x_1, ..., x_n)$,

$$\begin{aligned} p(\theta|y_1, ..., y_n) &= \frac{p(y_1, ..., y_n|\theta)p(\theta)}{p(y_1, ..., y_n)} \\ &= \frac{p(y_1, ..., y_n|\theta)p(\theta|x_1, ..., x_n)}{p(y_1, ..., y_n)} \end{aligned}$$

When not using conjugate priors, this methodology can yield posteriors that are not of a known parametric family, and so numerical techniques are necessary to estimate these posteriors. In these notes, one simple numerical technique is illustrated.

3 Bayesian Estimators

3.1 Point estimators

We have seen in Chapters 8 and 9 of your textbook [5] how to calculate MOM estimators, MLEs, and MVUEs. Thus, not surprisingly, there is no one single estimator in the Bayesian framework. One common estimator, which your book calls the *posterior Bayes estimator* in Definition 16.2, is the mean of the posterior,

$$\hat{\theta}_B = E(\theta|y_1, ..., y_n) = \int_{-\infty}^{\infty} \theta p(\theta|y_1, ..., y_n) d\theta.$$

The last equation only holds when θ is continuous - otherwise, expectation is calculated using a sum. The Bayes estimator for $t(\theta)$, a function of θ, is

$$\widehat{t(\theta)}_B = E(t(\theta)|y_1, ..., y_n) = \int_{-\infty}^{\infty} t(\theta) p(\theta|y_1, ..., y_n) d\theta.$$

Another common way to estimate θ is by maximizing the posterior (analagous to how we maximized the likelihood in Chapter 9), called the the *maximum a posteriori*, or MAP estimator,

$$\hat{\theta}_{MAP} = \text{argmax}_\theta p(\theta|y_1, ..., y_n).$$

Just as we did when finding MLEs, it is often helpful to log transform the posterior before taking derivatives with respect to θ.

For uni-model symmetric posteriors such as the normal and t distributions, $\hat{\theta}_B = \hat{\theta}_{MAP}$. For flat, uniform priors,

$$\hat{\theta}_{MAP} = \hat{\theta}_{MLE}.$$

Generally, Bayesian estimator are biased, although they are always functions of a sufficient statistic.

Note that for the same data set generated by the same likelihood model, two different analysts can get different Bayesian posterior means and MAPs by using different priors.

3.2 Interval estimators

Once a posterior $p(\theta|y_1, ..., y_n)$ has been calculated, then one has the distribution of the parameter θ given the data that we actually observed. From the posterior, one can calculate an interval estimator, called a *credible interval*, of the parameter θ. The construction of the interval is similar to how we calculated an interval estimator (called a confidence interval) from the sampling distribution of a statistic in Chapters 7 and 8. That is, we will find lower and upper limits, $\hat{\theta}_L$ and $\hat{\theta}_U$ respectively, such that

$$P(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U) = 1 - a$$

for some specified a. The interval

$$[\hat{\theta}_L, \hat{\theta}_U]$$

is called a $100(1-a)\%$ credible interval for θ.

The extreme difference between credible intervals and confidence intervals is the interpretation of the intervals. With credible intervals, one can say that, with probability $1-a$, the parameter θ is in the interval calculated from data. With confidence intervals, one can only make probability statements about the fixed value of the parameter θ being in the random interval before an explicit interval is calculated from data. After the confidence interval is calculated from data, only confidence statements can be made about the interval. According to a frequentist, the parameter θ is in a confidence interval calculated from data with probability either 0 or 1. It can be argued that the introduction of this new measure of uncertainty, called "confidence" is weird. In some sense, since a probability is already a measure of uncertainty, the Bayesian notion of using a probability to measure the uncertainty of whether the interval contains the parameter is more satisfactory.

Note that for the same data set generated by the same likelihood model, two different analysts can get different credible intervals by using different priors.

4 A few simple Bayesian models

For the binomial and normal likelihoods, we'll choose a few different priors, then use equation (1) to get the posterior.

4.1 Estimating a population proportion p from binary data

4.1.1 Binomial likelihood and an uninformative prior

Consider a SRS $y_1, ..., y_n$ of Bernoulli trials, so that $p(y_i|p) = p^{y_i}(1-p)^{1-y_i}$. The likelihood is $p(y_1, ..., y_n|p) = \Pi_{i=1}^n \left(p^{y_i}(1-p)^{1-y_i}\right) = p^{\sum y_i}(1-p)^{n-\sum y_i}$, which is Binomial($n$,$p$). Using the non-informative prior $p(p) = \text{Uniform}(0, 1)$, the posterior is

$$\begin{aligned} p(p|y_1, ..., y_n) &\propto p(y_1, ..., y_n|p) \times p(p) \\ &= p^{\sum y_i}(1-p)^{n-\sum y_i} \times 1 \\ &\propto \text{Beta}(\alpha^\star = \sum y_i + 1, \beta^\star = n - \sum y_i + 1) \end{aligned} \tag{2}$$

([4] p. 29). The Bayesian posterior mean, an estimator of p, is

$$\hat{p}_B = E(p|y_1, ..., y_n) = \frac{\alpha^\star}{\alpha^\star + \beta^\star} = \frac{\sum y_i + 1}{n + 2}.$$

This is different than the MVUE and MLE $\hat{p}_{MLE} = \frac{\sum y_i}{n}$ for p we found in Chapter 9 due to the addition of the constant 1 in the numerator and the constant 2 in the denominator, which

makes $\hat{p}_B$ biased for p. These constants have been referred to as "small sample corrections" in the literature [1], and advocated in STAT216 texts ([3] p. 471). As the sample size increases, $\hat{p}_B \approx \hat{p}_{MLE}$; in other words $\lim_{n\to\infty} Bias(\hat{p}_B) = 0$. Coupled with the fact that

$$\lim_{n\to\infty} Var(\hat{p}_B) = 0,$$

then $\hat{p}_B$ is consistent for p.

The value of p which maximizes the Beta$(\alpha^\star, \beta^\star)$ distribution is the MAP estimator for p,

$$\hat{p}_{MAP} = \frac{\alpha^\star - 1}{\alpha^\star + \beta^\star - 2} = \frac{\sum y_i}{n} = \hat{p}_{MLE}.$$

Thus, when there is a uniform prior, the estimator $\hat{p}_{MAP}$ is the MVUE for p.

A $100(1-a)\%$ credible interval for p is

$$[B_{a/2}, B_{1-a/2}]$$

where B_p is the p^{th} percentile from a Beta$(\alpha^\star = \sum y_i + 1, \beta^\star = n - \sum y_i + 1)$.

In R, use the following commands to generate Bayesian point and interval estimates of p.

```
% Here is some data
> y=c(0, 0, 1, 0, 0, 0, 1)
> n=length(y)
> a=.05

% The estimate from the Bayesian posterior mean is
> (sum(y) + 1)/(n+2)
[1] 0.3333333
% The MAP and MVUE estimates
> phat = sum(y)/n
> phat
[1] 0.2857143

% A 95% credible interval for p
> qbeta(c(a/2,1-a/2),sum(y)+1,n-sum(y)+1)
[1] 0.08523341 0.65085579

% PROPER CONCLUSION FOR CREDIBLE INTERVALS:
% Thus, the evidence suggests that, with probability .95, p is between .085 and 0.651.
% In other words, we do not have a precise interval estimate of p!
%

% Lets compare the credible interval to a BINOMIAL CI
% (also called a Clopper-Pearson interval), which I found to be [.099, .71]
% by trial and error:
%
% To get the lower confidence limit, find a p such that
% U ~ Bin(n,p) with Prob(U<=2)>= a/2
> 1-pbinom(sum(y),n,.09)
[1] 0.01933348
> 1-pbinom(sum(y),n,.099)
[1] 0.02500795
```

```
% Thus, the lower confidence limit is estimated to 0.099
%
% To get the upper confidence limit, find a p such that
% U ~ Bin(n,p) with Prob(U<=sum(y))>= a/2
> pbinom(sum(y),n,.7)
[1] 0.0287955
> pbinom(sum(y),n,.71)
[1] 0.02484208
% Thus, the upper confidence limit is estimated to 0.71
%


% Lets compare the credible interval to a LARGE SAMPLE CI
% (even though n=7 is not large)
> phat+c(-1,1)*qnorm(1-a/2)*sqrt(phat*(1-phat)/n)
[1] -0.04894358  0.62037215
```

4.1.2 Binomial likelihood and a beta prior

Now consider a SRS $y_1, ..., y_n$ of Bernoulli trials, so that likelihood is $p(y_1, ..., y_n|p) = \Pi_{i=1}^n \left(p^{y_i}(1-p)^{1-y_i}\right) = p^{\sum y_i}(1-p)^{n-\sum y_i} = \text{Binomial}(n,p)$, but now the prior is $p(p) = \text{Beta}(\alpha, \beta)$ where α and β are known. The posterior in this case is

$$\begin{aligned} p(p|y_1, ..., y_n) &\propto p(y_1, ..., y_n|p) \times p(p) \\ &= p^{\sum y_i}(1-p)^{n-\sum y_i} \times \text{Beta}(\alpha, \beta) \\ &\propto \text{Beta}(\alpha^\star = \sum y_i + \alpha, \beta^\star = n - \sum y_i + \beta) \end{aligned}$$

(see Example 16.1 in your textbook [5]). Observe that Beta(1,1)=Uniform(0,1), so that setting the prior $p(p) = \text{Beta}(\alpha = 1, \beta = 1)$ is the non-informative prior that we investigated in the last section and resulted in the posterior given in equation (2).

The Bayesian posterior mean, an estimator of p, is

$$\hat{p}_B = E(p|y_1, ..., y_n) = \frac{\alpha^\star}{\alpha^\star + \beta^\star} = \frac{\sum y_i + \alpha}{n + \alpha + \beta} = \frac{n\hat{p}_{MLE} + \alpha}{n + \alpha + \beta}.$$

The last equation gives $\hat{p}_B$ as a function of the MVUE and MLE $\hat{p}_{MLE} = \frac{\sum y_i}{n}$. Thus, $\hat{p}_B$ is biased for p. As the sample size increases, $\hat{p}_B \approx \hat{p}_{MLE}$, so $\lim_{n\to\infty} Bias(\hat{p}_B) = 0$ and $\lim_{n\to\infty} Var(\hat{p}_B) = 0$, which show that $\hat{p}_B$ is consistent for p.

Your textbook ([5] (see p. 802) rewrites the posterior mean as the weighted average

$$\hat{p}_B = \left(\frac{n}{n + \alpha + \beta}\right)\hat{p}_{MLE} + \left(\frac{\alpha + \beta}{n + \alpha + \beta}\right)\left(\frac{\alpha}{\alpha + \beta}\right).$$

This shows that the Bayesian estimator $\hat{p}_B$ is a compromise between the sample proportion $\hat{p}_{MLE} = \frac{\sum y_i}{n}$ (the MVUE and MLE for p), and $\frac{\alpha}{\alpha+\beta}$, the mean that the prior distribution suggests for p. As the sample size increases, $\hat{p}_{MLE}$ is given more weight over the prior mean.

The MAP estimator for p is

$$\hat{p}_{MAP} = \frac{\alpha^\star - 1}{\alpha^\star + \beta^\star - 2} = \frac{\sum y_i + \alpha - 1}{n + \alpha + \beta - 2}.$$

Thus, for arbitrary choice of α and β, $\hat{p}_{MAP}$ is biased yet consistent for p. However, if the prior Beta($\alpha = \sum_j x_j + 1$, $\beta = n_x - \sum_j x_j + 1$) is chosen, based on a previous SRS $x_1, .., x_{n_x}$, then $\hat{p}_{MAP} = \frac{\sum_i y + \sum_j x_j}{n_x + n_y} = \hat{p}_{MLE}$ is the MVUE.

A $100(1-a)\%$ credible interval for p is

$$[B_{a/2}, B_{1-a/2}]$$

where B_p is the p^{th} percentile from a Beta($\alpha^\star = \sum y_i + \alpha, \beta^\star = n - \sum y_i + \beta$).

An example in R:

```
% Suppose we have some prior knowledge that p is around .25.
% Thus, if we choose as a prior Beta(a=1,b=3) yields
% E(p) = alpha/(alpha+beta) = .25
> alpha=1
> beta=3
> a=.05

% 95% Credible interval for p using a Beta(1,3) prior
> qbeta(c(a/2,1-a/2),sum(y)+alpha,n-sum(y)+beta)
[1] 0.06673951 0.55609546

% This interval is more precise (narrow) than when we used a non-informative
% (or Beta(alpha=1,beta=1)) prior

% For larger values of alpha and beta, Beta(alpha,beta) has smaller variance.
% For a Beta prior with larger values of alpha and beta, we are
% saying that we have more confident in our prior information
% about p being close to E(p) = alpha/(alpha + beta) = .25
> alpha=10
> beta=30
> qbeta(c(a/2,1-a/2),sum(y)+alpha,n-sum(y)+beta)
[1] 0.1426685 0.3876689

% This interval is more precise (narrow) than when we used a Beta(alpha=1,beta=3) prior
```

4.2 Estimating a population mean μ from quantitative data

4.2.1 Normal likelihood when σ^2 is known and an uninformative prior

Now we'll assume that $p(y_i|\mu) = N_{y_i}(\mu, \sigma^2)$ when σ^2 is known. We'll use the uninformative prior $p(\mu) = 1$, so that for a SRS $y_1, ..., y_n$, the posterior is

$$\begin{aligned} p(\mu|y_1, ..., y_n) &\propto \Pi_{i=1}^n p(y_i|\mu, \sigma^2) \times 1 \\ &= \Pi_{i=1}^n N_{y_i}(\mu, \sigma^2) \\ &= N\left(\bar{y}, \sigma^2/n\right) \end{aligned} \tag{3}$$

([4] p. 45 and 67). Since the normal posterior is uni-model and symmetric, then the Bayesian posterior mean estimator and the MAP are the same

$$\hat{\mu}_B = \hat{\mu}_{MAP} = \bar{Y},$$

which is also the MOM, MLE, and MVUE estimators we found in Chapters 8 and 9 of your textbook [5] for the normal likelihood. Thus, $\hat{\mu}_B = \hat{\mu}_{MAP}$ is the MVUE for μ when σ is known, with minimum variance when compared to any other estimator of the normal mean μ,

$$Var(\hat{\mu}_B) = Var(\hat{\mu}_{MAP}) = \sigma^2/n.$$

The posterior in equation (3) shows that when σ is fixed, a $100(1-a)$ credible interval for μ is

$$\bar{y} \pm z_{1-a/2}\frac{\sigma}{\sqrt{n}}, \tag{4}$$

where $z_{1-a/2}$ is the $1-a/2$ percentile from $N(0,1)$.

Does this look familiar? It is the same as the $100(1-a)$ confidence interval for μ when σ is known. This illustrates the point made earlier that, even with the same interval estimate of μ, the Bayesian interpretation of the interval is different than the frequentist's. A Bayesian would say that (4) contains the parameter μ with probability $1-a$. The frequentist would say that once $\bar{y}$ is computed from data and substituted into (4), then one is $100(1-a)\%$ *confident* that μ is in the interval. No statements about probability are appropriate, since the frequentist maintains that μ is either in the interval or it is not, and the level of confidence now measures the uncertainty of whether μ is in the interval.

4.2.2 Normal likelihood when σ^2 is known and a normal prior

Consider the case when σ^2 is known, with $p(y_i|\mu) = N_{y_i}(\mu, \sigma^2)$, and a normal prior $p(\mu) = N(\eta, \delta^2)$ with η and δ^2 known. Then, when $y_1, ..., y_n$ are a SRS, the posterior is

$$\begin{aligned} p(\mu|y_1, ..., y_n) &\propto (\Pi_{i=1}^n p(y_i|\mu)) \times p(\mu) \\ &= \left(\Pi_{i=1}^n N(\mu, \sigma^2)\right) \times N(\eta, \delta^2) \\ &= N\left(\frac{\frac{\eta}{\delta^2} + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\delta^2} + \frac{n}{\sigma^2}}, \frac{1}{\frac{1}{\delta^2} + \frac{n}{\sigma^2}}\right) \end{aligned} \tag{5}$$

(see Example 16.4 in your textbook [5]). In the expressions for the posterior mean and variance, notice the balance between the data $y_1, ..., y_n$ and the prior. The term $\frac{1}{\delta^2}$ is called the *prior precision* since δ^2 is the variance of the prior, which, if large, implies that we have little knowledge of θ. The term $\frac{n}{\sigma^2}$ is called the *data precision* since $\frac{\sigma^2}{n}$ is the variance of the sample mean, which depends on the data. As $\delta \to \infty$, then the prior precision goes to zero, essentially giving a noninformative prior for θ, so that the posterior in this instance simplifies to

$$p(\mu|y_1, ..., y_n) = N(\bar{y}, \sigma^2/n)$$

which is the posterior we saw in equation (3).

Since the posterior is normal, it is uni-modal and symmetric, so

$$\hat{\mu}_B = \hat{\mu}_{MAP} = \frac{\frac{\eta}{\delta^2} + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\delta^2} + \frac{n}{\sigma^2}},$$

which is a weighted average of the sample mean $\bar{y}$ and the mean of the prior η,

$$\hat{\mu}_B = \frac{\frac{1}{\delta^2}}{\frac{1}{\delta^2} + \frac{n}{\sigma^2}}\eta + \frac{\frac{n}{\sigma^2}}{\frac{1}{\delta^2} + \frac{n}{\sigma^2}}\bar{y}.$$

It is straightforward to show that $\hat{\mu}_B$ is biased when $E(\eta) \neq \mu$. As the sample size increases, then the sample mean is weighed more, so that $\lim_{n\to\infty} \text{Bias}(\hat{\mu}_B) = 0$. Furthermore, since

$$\lim_{n\to\infty} Var(\hat{\mu}_B) = 0,$$

$\hat{\mu}_B$ is consistent for μ.

Thus, when σ^2 is known, a $100(1-a)\%$ credible interval for μ is

$$\frac{\frac{\eta}{\delta^2} + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\delta^2} + \frac{n}{\sigma^2}} \pm \frac{z_{1-a/2}}{\sqrt{\frac{1}{\delta^2} + \frac{n}{\sigma^2}}}$$

where $z_{1-a/2}$ is the $1 - a/2$ percentile from $N(0, 1)$.

4.3 Estimating a population mean μ and variance σ^2

4.3.1 Normal likelihood when σ^2 is unknown and an uninformative prior

Now we let $p(y_i|\mu) = N_{y_i}(\mu, \sigma^2)$ with both μ and σ^2 unknown, and assume a non-informative prior $p(\mu, \sigma^2)$ for $\theta = [\mu, \sigma^2]$. Then the posterior is

$$\begin{aligned} p(\mu, \sigma^2|y_1, ..., y_n) &\propto (\Pi_{i=1}^n p(y_i|\mu)) \times p(\mu, \sigma^2) \\ &= \left(\Pi_{i=1}^n N(\mu, \sigma^2)\right) \times p(\mu, \sigma^2) \\ &= N(\bar{y}, \sigma^2/n) \times (1/\chi^2_{n-1}) \end{aligned}$$

([4] p. 67) where $\bar{y}$ is the sample mean (the MVUE for μ), and s^2 is the usual sample variance (the MVUE for σ^2).

The marginal posteriors are

$$\begin{aligned} p(\mu|y) &= t_{n-1}(\bar{y}, s^2/n) \\ p(\sigma^2|y) &= 1/\chi^2_{n-1}. \end{aligned}$$

The posterior $p(\mu|y)$ shows that a $100(1-a)\%$ credible interval for μ is

$$\bar{y} \pm t_{1-a/2}\frac{s}{\sqrt{n}}$$

where $t_{1-a/2}$ is the $1 - a/2$ percentile from a t distribution with $n - 1$ degrees of freedom. The posterior $p(\sigma^2|y)$ shows that a $100(1-a)\%$ credible interval for σ^2 is

$$\left[\frac{(n-1)s^2}{\chi^2_{1-a/2}}, \frac{(n-1)s^2}{\chi^2_{a/2}}\right]$$

where χ^2_p is the p^{th} percentile from the χ^2 distribution with $n - 1$ degrees of freedom.

Do these intervals look familiar? These are the same as the $100(1-a)\%$ confidence intervals for μ and σ^2 that we derived in Chapter 8! This makes me wonder why we beat up on STAT216 students for making conclusions of confidence intervals that include the word "probability."

References

[1] Agresti and Coull. Approximate is better than exact for interval estimation of binomial proportions. *The American Statistician*, 52:119–129, 1998.

[2] G. Casella and R. L. Berger. *Statistical Inference*. Belmont CA: Duxbury Press, 1990.

[3] R. DeVeaux, P. Velleman, and D. Bock. *Stats: Data and Models*. Pearson Addison Wesley, 2nd edition, 2008.

[4] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Batesian Data Analysis*. New York: Chapman and Hall, 2000.

[5] Wackerly, Mendenhall, and Scheaffer. *Mathematical Statistics with Applications*. New York: Chapman and Hall, 7th edition, 2008.