Experiences from teaching Bayesian inference to students familiar with frequentist statistics

Bertil Wegmann

Division of Statistics and Machine Learning, IDA, Linköping University

April 19, 2017

Background

- Target group: third-year students at the Bachelor programme in Statistics and Data Analysis, Linköping University
- Students have a two-year background in frequentist statistics
- **Challenge**: Introduce Bayesian statistics as another way of doing statistical inference.
- **Challenge**: What difference does it make? What can we gain? Can we learn more about the parameters of interest with Bayesian inference?
- Students need motivated examples of the benefits with Bayesian statistics. Why do we need other tools for statistical inference than frequentist statistics?

Bayesian and frequentist inference

- **Frequentists** think of parameters, such as μ in a normal population, as fixed constants. They would **never** assign a probability distribution to μ.

- **Bayesians** may also think of parameters as constants, but may nevertheless assign a probability distribution to a parameter *if they do not know* its (constant) value.

- **Bayesians** add extra (prior) information to statistical inference:

$$\textit{Frequentist}: \ \textit{DATA}$$

$$\textit{Bayesian}: \ \textit{PRIOR} + \textit{DATA} \rightarrow \textit{POSTERIOR BELIEF}$$

- To a **Bayesian**, **probability is subjective**.

The prior - the missing piece of the puzzle

- Bayes theorem for continuous variables

$$p(\theta|x_1,...x_n) = \frac{f(x_1,...,x_n|\theta)p(\theta)}{f(x_1,...,x_n)}$$

- It is the prior $p(\theta)$ that helps to convert the likelihood function $f(x_1,...,x_n|\theta)$ into a (posterior) density for θ.

- Ignoring the prior is just as wrong as ignoring $P(A_i)$ in Bayes' theorem for events

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)}$$

Questioning the prior

- Student: 'But if the prior is subjective, then statistical science is subjective. That can not be right.'
- Bayesian: 'We all have different prior experience and expertise and the only honest thing is a subjective prior'.
- Bayesian: '**The objective part** of statistical inference is the **updating** from the prior to posterior, which is always done by **Bayes' theorem**'.
- Bayesian: 'A prior can be made minimally informative' ("Objective"). However, "non-informative priors do not exist".
- Bayesian: 'Non-Bayesian inference is also subjective. The class of entertained models, choice of significance level, etc are all subjective choices'.

Questioning similar results

- Student: 'Frequentist and Bayesian inference may give similar numerical results in a given problem. Why bother about Bayesian inference?'

- Bayesian: 'Interpretations of the results are always different. All information about probable values of an unknown parameter θ is included in the posterior distribution.'

- Bayesian: 'The **likelihood function** is the expression $f(X_1 = x_1, ..., X_n = x_n|\theta)$ **considered as function of** θ. This is NOT a pdf for θ.' Hence,
$$\int f(x|\theta)dx = 1$$
but in general
$$\int f(x|\theta)d\theta \neq 1$$

- Bayesian: 'So, Likely $\neq$ Probable. What does Likely mean?'

Credible interval vs confidence interval

- Student: 'My 95 % credible and 95 % confidence interval are essentially the same. Why bother about credible interval?'

- Bayesian: '95 % credible interval: the probability that the unknown parameter θ lies in the interval is 0.95.'

- Bayesian: '95 % confidence interval: the interval is stochastic and does not give any information about the probability of certain intervals for θ.'

Motivating example: Bernoulli model

- Model

$$X_1, ..., X_n|\theta \overset{iid}{\sim} Bern(x|\theta)$$

- Prior-Posterior mapping for Bernoulli model with Beta prior

$$\theta \sim Beta(\alpha, \beta)$$

$$\overset{x_1,\ldots,x_n}{\Longrightarrow}$$

$$\theta|x_1, ..., x_n \sim Beta(\alpha + s, \beta + f),$$

where $s = \sum_{i=1}^{n} x_i$ is the number of successes and $f = n - s$.

- George has gone through a part of his company's storage of soft drink bottles. He observed 6230 type A bottles out of 10000.
- Let $X_i = 1$ if the i:th bottle is a type A bottle. Posterior

$$\theta|x \sim Beta(\alpha + 6230, \beta + 3770)$$

- Elicit the prior by asking the expert George probabilistic questions: $E(\theta) =?$, $SD(\theta) =?$ or $Pr(\theta < c) =?$.

Motivating example: Normal prior and data, known variance

- Prior-Posterior mapping for normal model with normal prior

$$\theta \sim N(\mu_0, \tau_0^2) \overset{x_1,\ldots,x_n}{\Longrightarrow} \theta|x \sim N(\mu_n, \tau_n^2).$$

- Posterior precision:

$$\frac{1}{\tau_n^2} = \frac{n}{\sigma^2} + \frac{1}{\tau_0^2}.$$

Data precision + Prior precision

- Posterior mean:

$$\mu_n = w\bar{x} + (1-w)\mu_0,$$

and

$$w = \frac{\frac{n}{\sigma^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}}.$$

$$\frac{\text{Data precision}}{\text{Posterior precision}}(\text{Data mean}) + \frac{\text{Prior precision}}{\text{Posterior precision}}(\text{Prior mean})$$

- Prior-Posterior mapping for Poisson model with gamma prior

$$\text{Model:} \quad Y_1, ..., Y_n|\theta \overset{iid}{\sim} Pois(\theta)$$

$$\text{Prior:} \quad \theta \sim Gamma(\alpha, \beta)$$

$$\text{Posterior: } \theta|y_1, ..., y_n \sim Gamma(\alpha + \textstyle\sum_{i=1}^n y_i, \beta + n).$$

- **Data** of number of bomb hits in London: $n = 576$, $\sum_{i=1}^n y_i = 537$.
- The **posterior** distribution

$$p(\theta|y) \propto \theta^{\alpha+537-1} \exp[-\theta(\beta + 576)]$$

- **Posterior summaries**

$$E(\theta|y) = \frac{\alpha + \sum_{i=1}^n y_i}{\beta + n} \approx \bar{y} \approx 0.9323,$$

and

$$SD(\theta|y) = \left(\frac{\alpha + \sum_{i=1}^n y_i}{(\beta + n)^2}\right)^{\frac{1}{2}} = \frac{(\alpha + \sum_{i=1}^n y_i)^{\frac{1}{2}}}{(\beta + n)} \approx \frac{(537)^{\frac{1}{2}}}{576} \approx 0.0402.$$

if α and β are small compared to $\sum_{i=1}^n y_i$ and n.

Visualizing prior-posterior mapping: bomb hits in London