Bayesian Statistics
(Hogg Chapter Eleven)

STAT 406-01: Mathematical Statistics II [*]

Spring Semester 2016

Contents

Tuesday 26 January 2016 – Refer to Section 11.1 of Hogg, Chapters 1 and 2 of Jaynes and Van Horn 2003

As a supplement/alternative to the "Dutch Book" presentation in Section 11.1 of Hogg, we'll be using the following references to develop probability theory as an extension of logical reasoning:

- E. T. Jaynes, *Probability Theory: the Logic of Science*, chapters one and two. The first three chapters of the book are available free online at `http://bayes.wustl.edu/etj/prob/book.pdf`
- K. S. Van Horn, *International Journal of Applied Reasoning*, **38**, 3 (2003). available on line at `http://ksvanhorn.com/bayes/Papers/rcox.pdf`

[*] Copyright 2016, John T. Whelan, and all that

0 Preliminaries

0.1 Administrata

- Syllabus
- Text: Hogg, McKean, and Craig, *Introduction to Mathematical Statistics*, 7th edition.
- Other useful books:
  - Casella and Berger, *Statistical Inference*, 2nd edition. This is a standard first-year graduate text in statistics. It covers roughly the same material, but with a little more sophistication (more possible pathologies are mentioned) but also more of a practical philosophy.
  - Jaynes, *Probability Theory: the Logic of Science.* This is a sort of Bayesian manifesto and as such doesn't overlap much with the traditional approach, but it's got a lot of interesting bits in it, such as a demonstration that you can derive probability as an obvious extension of logic.
- Course website: `http://ccrg.rit.edu/~whelan/STAT-406/`
  - Will contain links to notes and problem sets; course calendar is probably the most useful.
  - Course calendar: *tentative* timetable for course.
- Course work:
  - Please read the relevant sections of the textbook *before* class so as to be prepared for class discussions.
  - There will be quasi-weekly homeworks. Collaboration is allowed an encouraged, but please turn in your own work, as obviously identical homeworks may not receive credit.
  - There will be two prelim exams, in class, and one cumulative final exam.
- Grading:

5%	Class Participation
20%	Problem Sets
20%	First Prelim Exam
20%	Second Prelim Exam
35%	Final Exam

  You'll get a separate grade on the "quality point" scale (e.g., 3.1667–3.5 is the B+ range) for each of these five components; course grade is weighted average.

0.2 Outline

1. Bayesian Statistics (Chapter Eleven)
2. Maximum Likelihood (Chapter Six)
3. Sufficiency (Chapter Seven)
4. Optimal Tests of Hypotheses (Chapter Eight)

Warning: the material in this course is even more advanced and abstract than in Math Stat I.

1 The Bayesian Approach to Probability

Unlike classical frequentist inference, which assigns probabilities only to the outcomes of repeatable experiments, Bayesian inference allows the assignment of probabilities to any statement which can be true or false. Classical probabilities are defined in terms of the frequencies of outcomes in the limit of many repetitions of the experiment, and are manipulated using set theory.

Bayesian probabilities are an expression of uncertainty rather than intrinsic randomness, and are better understood in terms of logic rather than set theory. The remarkable thing is that these two interpretations are described by the same mathematical operations, as we'll see in a moment.

Section 11.1 of Hogg contains a standard description of Bayesian reasoning as "subjective probabilities" using explanations related to gambling. Rather than this "Dutch book" approach, we'll motivate the Bayesian approach to probability as an extension to the formal logic encoded in Boolean algebra. The basic ingredient of deductive reasoning is the *syllogism*:

- A implies B
- A is true
- therefore B is true

For example, proposition A may be "it is raining" and proposition B "there are clouds in the sky". Plausible reasoning, on the other hand, includes the "weak syllogism"

- A implies B
- B is true
- therefore A is more plausible

Knowledge that there are clouds in the sky doesn't mean that it's definitely raining, but the proposition of rain is more plausible with the information about clouds than without it.

1.1 Logic and Deductive Reasoning

To understand how probability arises out of extended logic, we first need to define a few of the basic ingredients of symbolic logic. We write propositions as A, B, etc., and we can combine them with basic logical operations:

- Negation: $\overline{A}$ is true if A is false, and vice-versa. (This is the direct analogue of the complement A^c of set theory.) This can be thought of as "not A", and is sometimes written $\neg A$.
- Conjunction: $A \wedge B$ is true if A and B are both true, and false otherwise. (This is the direct analogue of the intersection $A \cap B$ of set theory.) This is also known as "A and B", and for compactness, is often written as A, B or even AB.
- Disjunction: $A \vee B$ is true if either A or B, or both, are true, and false otherwise. (This is the direct analogue of the union $A \cup B$ of set theory.) The meaning is "A or B" (inclusive or), and is sometimes (somewhat confusingly) written $A + B$. It's actually not necessary to define this operation separately, because it can be written in terms of negation and conjunction as $A \vee B = \overline{\overline{A} \wedge \overline{B}}$.

We can check logical statements like that by generating a *truth table* with entries for each possible combination of truth and falsehood for the atomic propositions A and B.

A	B	$A \wedge B$	$A \vee B$	$\overline{A}$	$\overline{B}$	$\overline{A} \wedge \overline{B}$	$\overline{\overline{A} \wedge \overline{B}}$
T	T	T	T	F	F	F	T
T	F	F	T	F	T	F	T
F	T	F	T	T	F	F	T
F	F	F	F	T	T	T	F

We can also represent the statement "A implies B" symbolically as $A \Rightarrow B$ as the conditional proposition $B|A$ (i.e., "$B|A$ is true" means B is true if A is true). In the real world, all of our statements of logicial (as well as plausible) reasoning are understood in the context of some background information, so in some sense all propositions are conditional. If we want to be pedantic, we can represent this background information with the proposition I and write $A|I$ and $B|I$ rather than just A and B. Some of the basic rules of logic include:

- Idempotence: $A \vee A = A$ and $A \wedge A = A$
- Commutativity: $A \vee B = B \vee A$ and $A \wedge B = B \wedge A$
- Associativity: $A \vee (B \vee C) = (A \vee B) \vee C$ and $A \wedge (B \wedge C) = (A \wedge B) \wedge C$
- Distributivity: $A \wedge (B \vee C) = (A \vee B) \wedge (A \vee C)$

These all also apply to conditional propositions, e.g, $(A \vee A)|I \equiv AA|I = A|I$

1.2 Plausible Reasoning and Cox's Theorem

If we want to consider propositions like $A|I$ to be "more plausible" or "less plausible" rather than "definitely true" or "definitely false", we need an extension of logic which assigns a "plausibility value" to $A|I$. In a confusing bit of notation, most references use the notation $(A|I)$ to refer to this plausibility value. In the interest of clarity, I'll call it $\pi(A|I)$ instead. (It doesn't matter which letter we use, because the notation will become obsolete in a moment.) There is a remarkable result that if the plausibility obeys a few basic conditions, this construction is equivalent to standard probability theory. Those conditions can be written as[1]

1. The plausibility $\pi(A|I)$ of any proposition is a real number, with definitely true propositions being the most plausible and definitely false propositions the least plausible.
2. Plausible reasoning should not contradict "common sense" and in particular should reduce to deductive reasoning when propositions are known to be true or false.
3. The formalism should be consistent, specifically
   (a) If there are two ways to calculate the same quantity, they should give the same answer
   (b) The formalism should use all relevant information
   (c) Equivalent propositions should be represented by the same plausibility value

The full proof of Cox's theorem is somewhat long and tedious, but here are some of the highlights. The basic results are

1. the development of a "product rule" expressing $\pi(A \wedge B|I) \equiv \pi(A, B|I)$ in terms of combinations of $\pi(A|I)$, $\pi(A|B, I)$, $\pi(B|I)$, and $\pi(B|A, I)$;
2. a statement about the plausibility values corresponding to truth and falsehood;
3. the development of a "sum rule" relating $\pi(A|I)$ to $\pi(\overline{A}|I)$.

The first step is to consider what functional dependencies are possible for $\pi(A, B|I)$. There are eleven possible choices of two or more arguments out of the four options $\pi(A|I)$, $\pi(A|B, I)$, $\pi(B|I)$, and $\pi(B|A, I)$, but all but two of them can be excluded by appeals to "common sense" (see Van Horn for details and commentary). For example, we would not expect to be able to write $\pi(A, B|I)$ as a function only of $\pi(A|I)$ and $\pi(B|I)$, because it ignores all possible relationships between the plausibility of A and B in the light of I. For example, if the proposition A is that a person is a man and A' is that they are a woman, we might have $\pi(A|I) = \pi(A'|I)$. But if B is the proposition that they have a beard, we would expect that $P(A, B|I) \neq P(A', B|I)$. The only possibilities that survive are a function of $\pi(A|I)$ and $\pi(B|A, I)$, or a function of $\pi(B|I)$ and $\pi(A|B, I)$. But since A, B is the same as B, A, both expressions must be valid and represented by the same functional form:

$$\pi(A, B|I) = F[\pi(A|I), \pi(B|A, I)] = F[\pi(B|I), \pi(A|B, I)] \tag{1.1}$$

[1] You may notice that some of these (e.g., "common sense") are not rigorously defined. See Van Horn for a more rigorous statement of the conditions.

Next one applies the condition of consistency to show that

$$\begin{aligned} \pi(A,B,C|I) &= F[\pi(C|I), F[\pi(B|C,I), \pi(A|B,C,I)]] \\ &= F[F[\pi(C|I), \pi(B|C,I)], \pi(A|B,C,I)] \end{aligned} \tag{1.2}$$

The equation $F[x, F[y,z]] = F[F[x,y],z]$ is known as the associativity equation, and its general solution is

$$F[x,y] = w^{-1}(w(x)w(y)) \tag{1.3}$$

where $w(x)$ is some positive, continuous, monotonic function. This means the product rule for plausibilities is

$$w(\pi(A,B|I)) = w(\pi(B|I))\, w(\pi(A|B,I)) \tag{1.4}$$

Next, we consider the value of $w(\pi(A|I))$ in the cases where $A|I$ is definitely true or definitely false.

- If $A|I$ is definitely true, then $A,B|I$ is an equivalent state of knowledge to $B|I$. Information about the truth of A is already encoded in I. Thus $\pi(A,B|I) = \pi(B|I)$ by the consistency condition. Likewise, I already tells us all there is to know about A, so $A|B,I$ is equivalent to $A|I$, and $p(A|B,I) = p(A|I)$. This means

  $$w(\pi(B|I)) = w(\pi(B|I))w(\pi(A|I)) \qquad \text{if } A|I \text{ definitely true} \tag{1.5}$$

  which implies that truth corresponds to $w(\pi(A|I)) = 1$.
- If $A|I$ is definitely false, then $A,B|I$ is an equivalent state of knowledge to $A|I$. The proposition A is already false given I, without adding the further restriction of B. Thus $\pi(A,B|I) = \pi(A|I)$ by the consistency condition. As before, I already gives us the full knowledge about A, so $A|B,I$ is equivalent to $A|I$, and $p(A|B,I) = p(A|I)$ again. This means

  $$w(\pi(A|I)) = w(\pi(B|I))w(\pi(A|I)) \qquad \text{if } A|I \text{ definitely false} \tag{1.6}$$

  There are two ways this can be true for arbitrary B. Either $w(\pi(A|I)) = 0$ for a false statement, in which case $w(x)$ is a monotonically increasing function with $0 \le w(x) \le 1$, or $w(\pi(A|I))$ is infinite for a false statement, in which case $w(x)$ is a monotonically decreasing function with $1 \le w(x) < \infty$. In the latter case, we can just replace $w(x)$ with a new function $w_{\text{new}}(x) = 1/w_{\text{old}}(x)$ and we once again have $w_{\text{new}}(x)$ monotonically increasing from 0 for definite falsehood to 1 for definite truth. We'll assume we've already done this.

Finally, we consider the relationship of $\pi(A|I)$ to $\pi(\overline{A}|I)$, or equivalently $w(\pi(A|I))$ to $w(\pi(\overline{A}|I))$. The argument goes that, since the propositions A and $\overline{A}$ are logically related, in particular $A \wedge \overline{A} \equiv A\overline{A}$ is definitely false and $A \vee \overline{A}$ is definitely true, the plausibilities must be related by

$$w(\pi(\overline{A}|I)) = S(w(\pi(A|I))) \tag{1.7}$$

where $S(x)$ some monotonically decreasing function $S(x)$ with $S(0) = 1$ and $S(1) = 0$. Careful consideration of this function in light of the product rule (see Jaynes for the gory details) allows one to conclude

$$[w(\pi(A|I))]^m + [w(\pi(\overline{A}|I))]^m = 1 \tag{1.8}$$

where m is some positive real number. Since we can also write the product rule as

$$[w(\pi(A,B|I))]^m = [w(\pi(A|I))]^m\, [w(\pi(B|A,I))]^m \tag{1.9}$$

we can define, in lieu of the plausibility $\pi(A|I)$, a number

$$P(A|I) = [w(\pi(A|I))]^m \tag{1.10}$$

which is a monotonically increasing function of the plausibility, and which obeys the fundamental rules of probability:

1. $P(A,B|I) = P(A|I)P(B|A,I)$
2. $0 \leq P(A|I) \leq 1$ with 0 corresponding to certain falsehood and 1 to certain truth
3. $P(A|I) + P(\overline{A}|I) = 1$

From these rules we can derive all of the other results of probability theory. For instance, if A and B are mutually exclusive in the context of of I, so that $A, B|I$ is impossible, we can work out

$$\begin{aligned} P(A \vee B|I) &= 1 - P(\overline{A \vee B}|I) = 1 - P(\overline{A}, \overline{B}|I) \\ &= 1 - P(\overline{A}|I)\, P(\overline{B}|\overline{A}, I) = 1 - P(\overline{A}|I)[1 - P(B|\overline{A}, I)] \\ &= 1 - P(\overline{A}|I) + P(\overline{A}|I)\, P(B|\overline{A}, I) \\ &= P(A|I) + P(\overline{A}, B|I) \end{aligned} \tag{1.11}$$

Now, since A and B are mutually exclusive given I, $\overline{A}, B|I$ is equivalent to $B|I$, which means

$$P(A,B|I) = P(A|I) + P(B|I) \quad \text{if } A, B|I \text{ impossible} \tag{1.12}$$

which is the standard restricted sum rule for mutually exclusive propositions.

Thursday 28 January 2016 – Read Section 11.2.1 of Hogg

2 Bayesian Parameter Estimation

The most common paradigm of Bayesian inference can be summarized in terms of probabilities involving the propositions

- $D \equiv$ "we collected some specific data or results in an experiment or observations"
- $H \equiv$ "some specific hypothesis is true"

as well as the usual background information/knowledge I we include in our calcuations. D will typically include the values of quantities with some "randomness" or "noise" included, while H may include values or ranges of values for model parameters. The fundamental calculational tool is *Bayes's theorem*, which uses the product rule to write

$$P(D|I)\, P(H|D, I) = P(H, D|I) = P(H|I)\, P(D|H, I) \tag{2.1}$$

and then solves it for

$$P(H|D, I) = \frac{P(H|I)\, P(D|H, I)}{P(D|I)} \tag{2.2}$$

This relationship is useful because we can usually write down $P(D|H, I)$ fairly easily. It's the basic ingredient in frequentist statistics: a sampling distribution that tells us how probable a particular realization of random data is given a hypothesis H and background information I. However, what we usually intuitively want to do in statistical inference is say how plausible a particular model or hypothesis or parameter range is, given that we've observed certain data. This is quantified by $P(H|D, I)$. The different parts of (2.2) have names:

- $P(D|H, I)$ is the *likelihood* (i.e., the sampling distribution viewed with an eye towards its dependence on H).
- $P(H|I)$ is the *prior probability* which we assign to H based only on the background information I. (This is obviously a tricky thing to do sometimes.)
- $P(H|D, I)$ is the *posterior probability* for H given both I and the observed data D.
- $P(D|I)$ is, in other contexts, called the *evidence*. It's basically a normalization factor; if $\{H_i\}$ is an exhaustive set

of mutually exclusive hypotheses, so that $\sum_i P(H_i|I) = 1$, then $P(D|I) = \sum_i P(H_i, D|I) = \sum_i P(H_i|I)\, P(D|H_i, I)$ ensures that $\sum_i P(H_i|D, I) = 1$. We can often avoid the need to calculate it exactly by considering $P(D|I)$ to be a proportionality constant independent of H and writing

$$P(H|D, I) \propto P(H|I)\, P(D|H, I) \tag{2.3}$$

2.1 Prior and Posterior Distributions

The simplest manifestation of this idea is a model with one parameter and one measurement. Because we want to assign probabilities to the values of both, we write the parameter as a "random variable" Θ (it's not random in the usual sense, but since the uncertainty associated with is described by a probability distribution, we can treat it as a random variable in the formalism), and the data as another random variable X. Then our background information I typically assigns prior probabilities to different values of Θ, as well as to values of X, given a value for Θ. If the distributions are discrete, this looks like probability mass functions

$$p_\Theta(\theta) = P(\Theta = \theta|I) \tag{2.4}$$

and

$$p_{X|\Theta}(x|\theta) = P(X = x|\Theta = \theta, I) \tag{2.5}$$

For the sake of notational simplicity in what follows, we'll generally suppress the reference to background information I, but keep in mind that all probabilities are conditional upon our general state of knowledge. Bayes's theorem becomes

$$p_{\Theta|X}(\theta|x) = \frac{p_\Theta(\theta)\, p_{X|\Theta}(x|\theta)}{p_X(x)} \propto p_\Theta(\theta)\, p_{X|\Theta}(x|\theta) \tag{2.6}$$

If the distributions are continuous, there's an analogous form in terms of probability density functions

$$f_{\Theta|X}(\theta|x) = \frac{f_\Theta(\theta)\, f_{X|\Theta}(x|\theta)}{f_X(x)} \propto f_\Theta(\theta)\, f_{X|\Theta}(x|\theta) \tag{2.7}$$

There are straightforward generalizations to the case where X and/or Θ are replaced with multivariate random vectors. One particularly interesting case is where $\mathbf{X}$ is a random sample drawn from a distribution $f_{X|\Theta}(x|\theta)$, i.e,

$$f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta) = \prod_{i=1}^{n} f_{X|\theta}(x_i|\theta) \tag{2.8}$$

and then the posterior pdf for Θ becomes

$$f_{\Theta|\mathbf{X}}(\theta|\mathbf{x}) = \frac{f_{\mathbf{X},\Theta}(\mathbf{x}, \theta)}{f_{\mathbf{X}}(\mathbf{x})} = \frac{f_\Theta(\theta)\, f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta)}{f_{\mathbf{X}}(\mathbf{x})} \propto f_\Theta(\theta)\, f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta) \tag{2.9}$$

For some reason, Hogg invents a lot of notation, using different letters for $f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta)$ [which he calls $L(\mathbf{x}|\theta)$], $f_\Theta(\theta)$ [$h(\theta)$], $f_{\mathbf{X}}(\mathbf{x})$ [$g_1(\mathbf{x})$], and $f_{\mathbf{X},\Theta}(\mathbf{x}, \theta)$ [$g(\mathbf{x}, \theta)$]. In practice, the subscripts on the pdfs or pmfs ought to tell you which random variables (observations and/or unknown parameters) are in question. One can (and sometimes does) go even further and rely upon the names of the arguments to tell the reader what quantities are meant[2], e.g.,

$$f(\theta|\mathbf{x}) = \frac{f(\theta)\, f(\mathbf{x}|\theta)}{f(\mathbf{x})} \tag{2.10}$$

Note that the denominator of (2.9) can be calculated as

$$f_{\mathbf{X}}(\mathbf{x}) = \int_{-\infty}^{\infty} f_\Theta(\theta)\, f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta)\, d\theta \tag{2.11}$$

[2] This is sometimes considered an abuse of notation in mathematical circles, though, and is related to the language gap between mathematicial and physical scientists, exemplified by the question: if $T(x, y) = kx^2 + ky^2$, what is $T(r, \phi)$?

2.1.1 Example: Gaussian Likelihood

To give a very simple example of the calculation of a posterior pdf, suppose we have a model with one parameter Θ and make a single measurement of a quantity X, with a Gaussian likelihood

$$f_{X|\Theta}(x|\theta) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(x-\theta)^2/(2\sigma^2)} \tag{2.12}$$

and a prior pdf of

$$f_\Theta(\theta) = \frac{1}{\gamma\sqrt{2\pi}}\, e^{-\theta^2/(2\gamma^2)} \tag{2.13}$$

Bayes's theorem tells us that the posterior pdf is

$$f_{\Theta|X}(\theta|x) = \frac{f_\Theta(\theta)\, f_{X|\Theta}(x|\theta)}{f_X(x)} \tag{2.14}$$

The numerator is

$$f_{X,\Theta}(x,\theta) = f_\Theta(\theta)\, f_{X|\Theta}(x|\theta) = \frac{1}{2\pi\sigma\gamma}\exp\left(-\frac{1}{2}[\sigma^{-2}(x-\theta)^2 + \gamma^{-2}\theta^2]\right) \tag{2.15}$$

This still has a Gaussian dependence on θ, which we see by completing the square on the expression in square brackets:

$$\begin{aligned}\sigma^{-2}(x-\theta)^2 + \gamma^{-2}\theta^2 &= (\sigma^{-2}+\gamma^{-2})\theta^2 - 2\sigma^{-2}x\theta + \sigma^{-2}x^2 \\ &= (\sigma^{-2}+\gamma^{-2})\left(\theta - \frac{x}{1+\sigma^2/\gamma^2}\right)^2 + \frac{x^2}{\sigma^2+\gamma^2}\end{aligned} \tag{2.16}$$

If we define $\sigma' = (\sigma^{-2}+\gamma^{-2})^{-1/2}$ and $\theta_0(x) = \frac{x}{1+\sigma^2/\gamma^2}$, we have

$$f_{X,\Theta}(x,\theta) = \frac{e^{-x^2/[2(\sigma^2+\gamma^2)]}}{2\pi\sigma\gamma} e^{-[\theta-\theta_0(x)]^2/(2\sigma'^2)} \tag{2.17}$$

We can marginalize over θ to get the denominator

$$f_X(x) = \frac{e^{-x^2/[2(\sigma^2+\gamma^2)]}}{2\pi\sigma\gamma}\int_{-\infty}^{\infty} e^{-[\theta-\theta_0(x)]^2/(2\sigma'^2)}\, d\theta = \frac{e^{-x^2/[2(\sigma^2+\gamma^2)]}}{2\pi\sigma\gamma}\sigma'\sqrt{2\pi} \tag{2.18}$$

so the posterior is

$$f_{\Theta|X}(\theta|x) = \frac{f_{X,\Theta}(x,\theta)}{f_X(x)} = \frac{1}{\sigma'\sqrt{2\pi}} e^{-[\theta-\theta_0(x)]^2/(2\sigma'^2)} \tag{2.19}$$

i.e., a Gaussian distribution with mean $\theta_0(x) = \frac{x}{1+\sigma^2/\gamma^2}$ and variance $\sigma'^2 = (\sigma^{-2}+\gamma^{-2})^{-1}$. Note that we didn't actually have to keep track of the factor out front which cancelled. It would have been sufficient to write

$$\begin{aligned}f_{\Theta|X}(\theta|x) &\propto f_\Theta(\theta)\, f_{X|\Theta}(x|\theta) \\ &\propto \exp\left(-\frac{(\sigma^{-2}+\gamma^{-2})}{2}\left[\theta - \frac{x}{1+\sigma^2/\gamma^2}\right]^2\right)\end{aligned} \tag{2.20}$$

and then deduce the form of the proportionality constant from the requirement that $\int_{-\infty}^{\infty} f_{\Theta|X}(\theta|x)\, d\theta = 1$.

Note that if the prior is very broad, $\gamma \gg \sigma$, then $\theta_0(x) \approx x$ and $\sigma' \approx \sigma$, and the posterior is

$$f_{\Theta|X}(\theta|x) \approx \frac{1}{\sigma\sqrt{2\pi}} e^{-[\theta-x]^2/(2\sigma^2)} \tag{2.21}$$

i.e., it has the same shape as the likelihood function, but is normalized as a pdf in θ.

On the other hand, if the likelihood is broad compared to the prior, $\sigma \gg \gamma$, then $\sigma' \approx \gamma$, and $\theta_0(x) \approx \gamma^2/\sigma^2\, x$. If we further have $\gamma^2/\sigma^2\, x \ll \sigma$, i.e., $x \ll \sqrt{\sigma\gamma}$, we have

$$f_{\Theta|X}(\theta|x) \approx \frac{1}{\gamma\sqrt{2\pi}} e^{-\theta^2/(2\gamma^2)} \tag{2.22}$$

I.e., the posterior looks just like the prior. What's happening here is that the prior information constrains the value of Θ much more tightly than the observation of X.

Tuesday 2 February 2016 – Read Section 11.2.2 of Hogg

2.2 Bayesian point estimation

When we do Bayesian inference, e.g., from a sample $\mathbf{X}$ drawn from a distribution $f_{X|\Theta}(x|\theta)$, the full information is contained in the posterior probability distribution

$$f_{\Theta|\mathbf{X}}(\theta|\mathbf{x}) \propto f_\Theta(\theta)\prod_i f_{X|\Theta}(x_i|\theta) \ . \tag{2.23}$$

For example, suppose we observe a Poisson process with an unknown rate of Θ events per observation interval, so that the likelihood function is

$$p_{X|\Theta}(x|\theta) = \frac{\theta^x}{x!}\,e^{-\theta} \tag{2.24}$$

with a prior pdf which is a gamma distribution with parameters α and β:

$$f_\Theta(\theta) \propto \theta^{\alpha-1}e^{-\theta/\beta} \qquad 0<\theta<\infty \tag{2.25}$$

Note that we might want to try something like a uniform distribution on θ, restricted to $\theta > 0$. That wouldn't be normalizable for $0<\theta<\infty$, so it wouldn't be an allowed pdf. However, we could consider it as a limit of the gamma family with $\alpha = 1$ as $\beta\to\infty$. Now suppose we observe n intervals, i.e., collect s sample of size n from this Poisson distribution, with observed numbers of events $\{x_1, x_2, \dots, x_n\}\equiv\mathbf{x}$. The likelihood function is then

$$p_{\mathbf{X}|\Theta}(\mathbf{x}|\theta) = \prod_{i=1}^n \frac{\theta^{x_i}}{x_i!}\,e^{-\theta} = \theta^{\sum_{i=1}^n x_i}\,e^{-n\theta}\prod_{i=1}^n\frac{1}{x_i!} \tag{2.26}$$

For the sake of constructing the posterior pdf, we're only interested in the dependence of the likelihood on the parameter θ, so

$$p_{\mathbf{X}|\Theta}(\mathbf{x}|\theta) \propto \theta^{\sum_{i=1}^n x_i}\,e^{-n\theta} \tag{2.27}$$

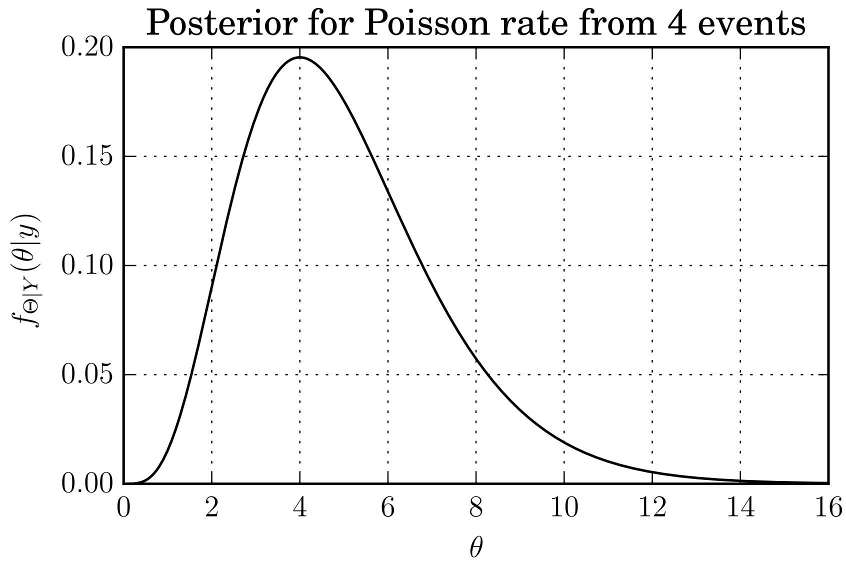
Note that this only depends on the total number of events $y = \sum_{i=1}^n x_i$ (and the number of observation intervals n). We thus say that $Y = \sum_{i=1}^n X_i$ is a *sufficient statistic* for the parameter θ, i.e., the only combination of the random sample $\mathbf{X}$ which matters for constructing the likelihood, up to a θ-independent constant. The posterior is then

$$f_{\Theta|\mathbf{X}}(\theta|\mathbf{x}) \propto f_\Theta(\theta)p_{\mathbf{X}|\Theta}(\mathbf{x}|\theta) \propto \theta^{\alpha+y-1}e^{-(\beta+n)\theta} \tag{2.28}$$

i.e., a gamma distribution with parameters $\alpha' = \alpha + y$ and $\beta' = (\beta^{-1}+n)^{-1}$. If we go with the uniform prior limit, i.e., $\alpha = 1$ and $\beta^{-1}\ll n$, we have

$$f_{\Theta|\mathbf{X}}(\theta|\mathbf{x}) \propto \theta^y e^{-n\theta} \tag{2.29}$$

For example, if $y = 4$ and $n = 1$, the posterior is a Gamma(5,1) distribution:



If you want to know what we have to say about the unknown rate Θ after observing three events in an observation interval, given a uniform prior, that plot tells the whole story. Still, sometimes we want to boil the posterior down to a single estimate. Hogg refers to this as $\delta(\mathbf{x})$, using some notation which we haven't developed yet. Three obvious choices of a single number $\delta(\mathbf{x})$ are

- The mode $\hat{\theta} = \operatorname{argmax}_\theta f_{\Theta|\mathbf{X}}(\theta|\mathbf{x})$ of the posterior pdf.
- The mean or expectation value

$$\bar{\theta} = E(\Theta|\mathbf{x}) = \int_{-\infty}^{\infty} \theta \, f_{\Theta|\mathbf{X}}(\theta|\mathbf{x}) \, d\theta \tag{2.30}$$

  Choosing $\delta(\mathbf{x}) = E(\Theta|\mathbf{x})$ minimizes the mean square error

$$E([\Theta - \delta(\mathbf{x})]^2|\mathbf{x}) = \int_{-\infty}^{\infty} [\theta - \delta(\mathbf{x})]^2 \, f_{\Theta|\mathbf{X}}(\theta|\mathbf{x}) \, d\theta \tag{2.31}$$

  Note that you showed this on the extra credit problem of problem set 2 in STAT405 last semester.
- The median $\tilde{\theta}$ defined by

$$\int_{-\infty}^{\tilde{\theta}} \theta \, f_{\Theta|\mathbf{X}}(\theta|\mathbf{x}) \, d\theta = \frac{1}{2} \tag{2.32}$$

  Choosing $\delta(\mathbf{x}) = \tilde{\theta}$ minimizes the mean absolute error

$$E(|\Theta - \delta(\mathbf{x})| \, |\mathbf{x}) = \int_{-\infty}^{\infty} |\theta - \delta(\mathbf{x})| \, f_{\Theta|\mathbf{X}}(\theta|\mathbf{x}) \, d\theta \tag{2.33}$$

  This was also part of the extra credit problem of problem set 2 in STAT405 last semester.

For the example at hand, we can see from the plot that the mode is $\hat{\theta} = 4$. We can also show this in general for the Gamma distribution with parameters α' and β'; if

$$f(\theta) \propto \theta^{\alpha'-1} e^{-\theta/\beta'} \tag{2.34}$$

we can maximize the pdf by maximizing its logarithm

$$\ln f(\theta) = (\alpha' - 1) \ln \theta - \theta/\beta' + \text{const} \tag{2.35}$$

so

$$\frac{\partial}{\partial\theta} \ln f(\theta) = \frac{(\alpha' - 1)}{\theta} - \frac{1}{\beta'} \tag{2.36}$$

setting this to zero gives $\hat{\theta} = (\alpha' - 1)\beta'$ or in our case, where $\alpha' = y + 1$ and $\beta' = 1/n$, $\hat{\theta} = \frac{y+1-1}{n} = y/n$.

On the other hand, we can see that the distribution is not symmetric about its mode $\theta = 4$, so the mean and median will be different from the mode, in this case somewhat larger. We know the expectation value of a gamma distribution is $\alpha'\beta'$, which in this case is $\bar{\theta} = \frac{y+1}{n}$ or specifically $4 + 1 = 5$.

There's no closed-form expression for the median of a gamma distribution, but we can use scipy or R to find that it is $\tilde{\theta} \approx 4.67$ for the specific example we've considered,

2.2.1 Change of variables

Both the mean and median can be described as minimizing the expectation value of a *loss function* (square error in the former case and absolute error in the latter), and Hogg refers to them as different choices of a *Bayes estimator*. There is still something appealing about quoting the mode of the posterior, since it reflects the parameter space value near which we most expect to find the parameter given our observations. It has some additional pitfalls, though. Suppose in the example we've be considering that we change variables from Θ to $\Gamma = \frac{1}{\Theta}$, i.e., instead of the rate we use the inverse rate, the expected interval between observed events. We don't need to start the problem over, but rather note that the posterior pdf $f_{\Theta|\mathbf{X}}(\theta|\mathbf{x})$ is a probability density in θ. Thus we can change variables with the usual formalism which we recall as

$$\frac{dP}{d\gamma} = \left|\frac{d\theta}{d\gamma}\right| \frac{dP}{d\theta} = \frac{1}{\gamma^2}\frac{dP}{d\theta} \tag{2.37}$$

so

$$f_{\Gamma|\mathbf{x}}(\gamma|\mathbf{x}) = \gamma^{-2} f_{\Theta|\mathbf{x}}(\gamma^{-1}|\mathbf{x}) \propto \gamma^{-y-2} e^{-n/\gamma} \tag{2.38}$$

we can find the maximum

$$\frac{\partial}{\partial\gamma} \ln f_{\Gamma|\mathbf{x}}(\gamma|\mathbf{x}) = \frac{-y-2}{\gamma} + \frac{n}{\gamma^2} \tag{2.39}$$

setting this to zero gives

$$\hat{\gamma} = \frac{n}{y+2} \neq \frac{1}{\hat{\theta}} = \frac{n}{y} \tag{2.40}$$

Note that the mean has the same problem, since

$$E\left(\frac{1}{\Theta}\middle|\mathbf{x}\right) \neq \frac{1}{E(\Theta|\mathbf{x})} \tag{2.41}$$

Specifically, when

$$f_{\Theta|\mathbf{x}}(\theta|\mathbf{x}) = \frac{1}{\Gamma(\alpha')\beta'^{\alpha'}} \theta^{\alpha'-1}\, e^{-\theta/\beta'}\ , \tag{2.42}$$

we get

$$\begin{aligned} E\left(\frac{1}{\Theta}\middle|\mathbf{x}\right) &= \int_{-\infty}^{\infty} \theta^{-1} f_{\Theta|\mathbf{x}}(\theta|\mathbf{x})\, d\theta = \frac{1}{\Gamma(\alpha')\beta'^{\alpha'}} \int_0^\infty \theta^{\alpha'-2}\, e^{-\theta/\beta'}\, d\theta \\ &= \frac{1}{\Gamma(\alpha')\beta'} \int_0^\infty u^{\alpha'-2}\, e^{-u}\, du = \frac{\Gamma(\alpha'-1)}{\Gamma(\alpha')\beta'} \\ &= \frac{1}{(\alpha'-1)\beta'} \neq \frac{1}{\alpha'\beta'} \end{aligned} \tag{2.43}$$

Specifically, if $\alpha' = y+1$ and $\beta' = 1/n$, $E\left(\frac{1}{\Theta}\middle|\mathbf{x}\right) = \frac{n}{y}$ while $\frac{1}{E(\Theta|\mathbf{x})} = \frac{n}{y+1}$.

Thursday 4 February 2016 – Read Section 11.2.3 of Hogg

2.3 Bayesian Interval Estimation

2.3.1 Posterior mean and standard deviation

A single point estimate of an unknown parameter Θ only tells one piece of the story contained in the posterior pdf $f_{\Theta|\mathbf{x}}(\theta|\mathbf{x})$. Often we want to express not only an estimate of a parameter but also the uncertainty associated with that estimate. One way to do this would be to report both the mean

$$\bar{\theta} = E(\Theta|\mathbf{x}) = \int_{-\infty}^{\infty} \theta\, f_{\Theta|\mathbf{x}}(\theta|\mathbf{x})\, d\theta \tag{2.44}$$

of the posterior and its variance

$$\sigma_\theta^2 = \operatorname{Var}(\Theta|\mathbf{x}) = E([\Theta - \bar{\theta}]^2|\mathbf{x}) = \int_{-\infty}^{\infty} (\theta - \bar{\theta})^2\, f_{\Theta|\mathbf{x}}(\theta|\mathbf{x})\, d\theta \tag{2.45}$$

and quote something like $\theta \sim \bar{\theta} \pm \sigma_\theta$. So for example, if $f_{\Theta|\mathbf{x}}(\theta|\mathbf{x})$ is a Gamma(α',β') distribution as in our previous example, $\bar{\theta} = \alpha'\beta'$ and $\sigma_\theta = \beta'\sqrt{\alpha'}$. If we've assumed a uniform prior on the rate Θ and observed y events in n standard intervals, so that $\alpha' = y+1$ and $\beta' = 1/n$, we would estimate $\theta \sin \frac{y+1}{n} \pm \frac{\sqrt{y+1}}{n}$. In the specific example where $y = 4$ and $n = 1$, where we know that $\bar{\theta} = 5$, we'd conclude $\theta \sim 5 \pm \sqrt{5} \approx 5 \pm 2.24$.

Using the standard deviation of the posterior to set a "one sigma errorbar" is somewhat arbitrary and best suited to cases where the posterior is Gaussian. In practice, it can have some pitfalls. for example, if we decided to put three sigma errorbars on the estimate we've just considered, they'd stretch from $5 - 3\sqrt{5} \approx -1.71$ to $5 + 3\sqrt{5} \approx 11.71$. But we know the posterior pdf for Θ is zero for negative values of theta, so the lower limit of that interval doesn't make a lot of sense.

2.3.2 Credible intervals

Instead, we can use the probabilistic meaning of the posterior pdf to define an interval with lower and upper ends $\ell(\mathbf{x})$ and $u(\mathbf{x})$ such that there is some posterior probability $1-\alpha$ (e.g., 90% or 95%) that the unknown value of Θ lies between $\ell(\mathbf{x})$ and $u(\mathbf{x})$.

$$1-\alpha = P(\ell(\mathbf{x}) < \Theta < u(\mathbf{x})|\mathbf{X}=\mathbf{x}) = \int_{\ell(\mathbf{x})}^{u(\mathbf{x})} f_{\Theta|\mathbf{X}}(\theta|\mathbf{x})\,d\theta \tag{2.46}$$

$(\ell(\mathbf{x}), u(\mathbf{x}))$ is known as a *credible interval* or plausible interval for the parameter Θ. Note that this is closer to most people's intuition than the corresponding frequentist confidence interval, defined by

$$1-\alpha = P(\ell'(\mathbf{X}) < \theta < u'(\mathbf{X})) \tag{2.47}$$

where the probability in question applies not to the uncertainty in the parameter θ, but to the random variation of the endpoints of the interval under hypothetical repetitions of the experiment.

We saw that the construction of a frequentist confidence interval was somewhat arbitrary, depending on e.g., the choice of a pivot variable. On the other hand, the definition of a credible interval follows straightforwardly from the construction of the posterior pdf. The only ambiguity is how much of the "leftover" probability falls on each side of the interval. For instance, a 90% credible interval could be between the 5th and 95th percentiles of the posterior pdf, or the 1st and 91st, or everything above the 10th, etc. One simple choice is to make it symmetric, by requiring

$$P(\Theta < \ell(\mathbf{x})|\mathbf{X}=\mathbf{x}) = \frac{\alpha}{2} = P(u(\mathbf{x}) < \Theta|\mathbf{X}=\mathbf{x}) \tag{2.48}$$

E.g., if we want a symmetric 90% credible interval for the pdf considered last time, which is a gamma distribution with $\alpha' = y+1 = 5$ and $\beta' = \frac{1}{n} = 1$, we can use scipy to get it:

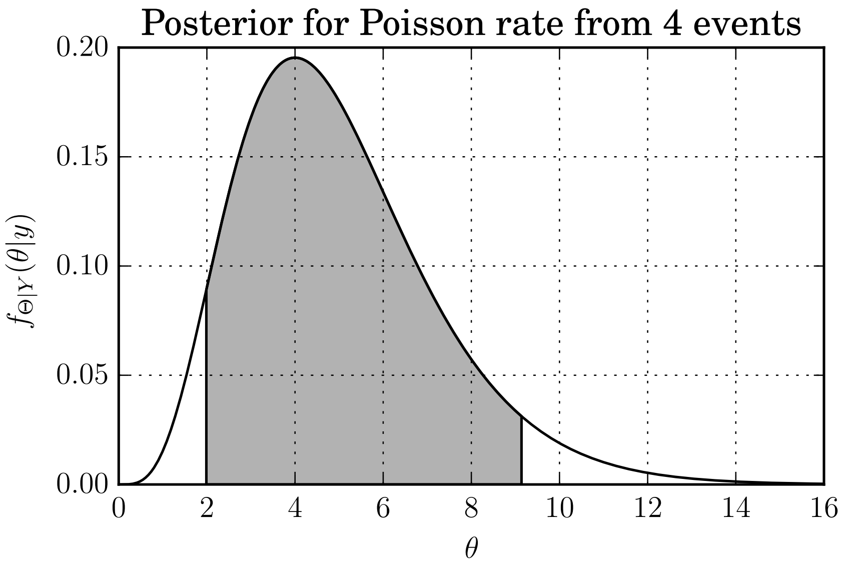
```
In [1]: from scipy.stats import gamma as gammadist

In [2]: y=4;

In [3]: rv = gammadist(y+1)

In [4]: print ( "90 pct symmetric credible interval is"
   ...: + " %g to %g" % (rv.ppf(0.05),rv.ppf(0.95)) )
90 pct symmetric credible interval is 1.97015 to 9.15352
```

We can shade this in on the pdf:

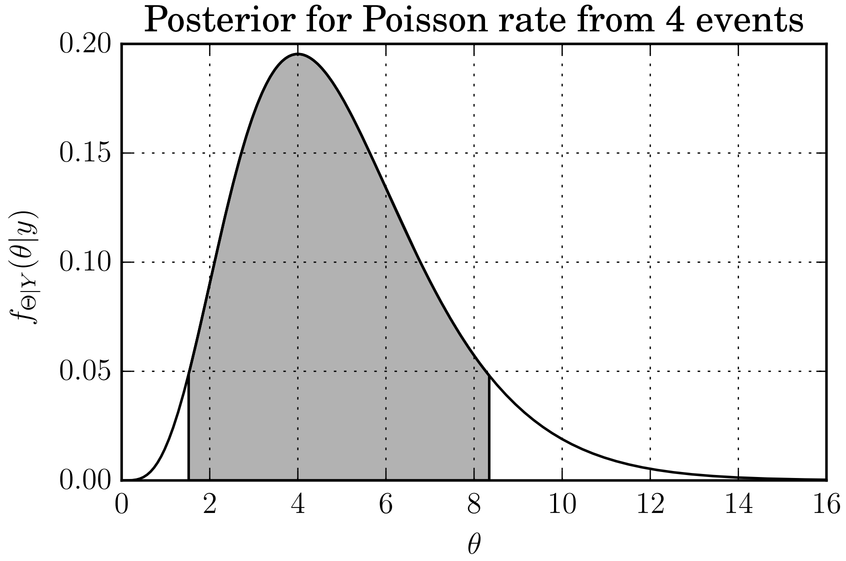


The white region on the left contains 5% of the area under the posterior pdf; the white region on the left contains another 5%, and the remaining 90% is in the shaded region in the middle.

2.3.3 Highest-density regions

Note: this is actually covered in the second half of section 11.3 of Hogg.

Another approach to choosing a credible interval is to collect our 90% (or whatever $1-\alpha$ is equal to) by taking the θ interval with the highest values of the pdf. We call this the *highest-density region* or HDR. For example, in the case just considered, this turns out to be 1.52 to 8.34:



We see, of course, that the pdf at the lower end of the interval is the same as at the upper end, which will always happen if the pdf is continuous. We also note that the width of this credible interval, $8.34-1.52=6.82$ is less than the width $9.15-1.97=7,18$ of the symmetric credible interval. It's not hard to see that the HDR will be the narrowest credible interval corresponding to a given probability. (If you want to integrate up to a given probability, the way to get it in the narrowest range of the integrand is to take the region with the highest probability density.)

A few note about highest-density credible intervals:

- That the HDR is easiest to construct if the posterior distribution is symmetric about its peak; in that case, the HDR will be the same as the symmetric credible interval. In general one has to find it numerically (which is what I did for the gamma distribution).
- The HDR is a nice way to transition between one-sided and two-sided credible intervals. For instance, if we have a truncated Gaussian posterior on θ, of the form

$$f_{\Theta|\mathbf{X}}(\theta|\mathbf{x}) \propto e^{(\theta-\mu)^2/(2\sigma^2)} \qquad 0<\theta<\infty \tag{2.49}$$

  then, if $\mu < 1.22\sigma$, the lower end of the HDR for θ will be at zero. If $\mu > 1.22\sigma$, both ends will be at positive θ, and we will set a two-sided credible interval.
- Like the mode, the HDR is not invariant under reparametrization.
- The HDR construction generalizes nicely to higher-dimensional parameter spaces, where the boundaries of the HDR are just level surfaces of the multi-dimensional posterior pdf.

Tuesday 9 February 2016 – Read Section 11.3 of Hogg

3 Prior Distributions in Bayesian Inference

3.1 Conjugate Prior Families

We now come to the recurring question about Bayesian inference: why do we choose a particular prior? The answer so far is not very satisfying: our priors so far have basically been chosen to make the math easier. To collect together some results from the last couple of weeks of lectures and homeworks:

- If the prior for the probability in a binomial experiement is a Beta(α,β) distribution, and we observe k successes in n trials, the posterior is a Beta($\alpha+k$,$\beta+n-k$) distribution.
- If the prior for the mean of a normal distribution with known variance σ^2 is a $N(\mu_0,\sigma_0^2)$ distribution and we collect a sample of size n with sample mean $\overline{x}$, the posterior is a $N\left(\frac{\mu_0+\sigma_0^{-2}n\sigma^{-2}\overline{x}}{\sigma_0^{-2}+n\sigma^{-2}},\frac{1}{\sigma_0^{-2}+n\sigma^{-2}}\right)$ distribution.
- If the prior for the rate per unit time of a Poisson process is a Gamma(α_0,β_0) distribution, and we observe y events in a time t, the posterior is a Gamma(α_0+y,$[\beta_0^{-1}+t]^{-1}$) distribution.

These are all examples of what is called a *conjugate prior family* of distributions for a given likelihood function and parameter(s) of interest. This means that if the prior distribution is a member of the family, the posterior will be a member as well, with parameters determined by the sufficient statistic(s) constructed from the data. This seems like an arbitrary way to choose a prior (which is after all supposed to reflect your knowledge going into the experiment, and shouldn't have to depend on what measurement you're planning to make), but it's often the case that a "realistic" prior is close to a member of the conjugate prior distribution, or at least a limiting form of the family. For instance, we noted last week that a uniform distribution from 0 to ∞ is the limit of a Gamma(1,β) distribution as $\beta\to\infty$.

3.2 Non-informative and Improper Priors

Often, we want to consider a problem using as little prior information as possible. For instance, when estimating the rate of a Poisson process, we considered a uniform prior for $0<\theta<\infty$. Of course, there is no normalized pdf of that form. The best we can do is take a family of distributions which becomes uniform in some limit, like the Gamma(1,β) family considered above, or

$$f_\Theta(\theta)=\begin{cases}\frac{1}{\theta_{\max}} & 0<\theta<\theta_{\max}\\ 0 & \text{otherwise}\end{cases}\tag{3.1}$$

If we try to take the limit $\theta_{\max}\to\infty$, we find that the normalizing constant $\theta_{\max}^{-1}$ goes to zero. But if we ignore normalizing constants and just write $f_\Theta(\theta)\propto 1$, $0<\theta<\infty$, we can write the posterior arising from a counting experiment with y events in a time t as

$$f_{\Theta|Y}(\theta|y)\propto f_\Theta(\theta)\,f_{Y|\Theta}(y|\theta)\propto\theta^y\,e^{-\theta t}\tag{3.2}$$

which is a Gamma($y+1$,$1/t$) distribution, and can be written in normalized form as

$$f_{\Theta|Y}(\theta|y)=\frac{t^{y+1}}{\Gamma(y+1)}\,\theta^y\,e^{-\theta t}\tag{3.3}$$

In this case, we call the prior distribution (which is not a true probability density function, because it can't be normalized) an *improper prior*, but note that the results of this experiment produce a posterior pdf which *is* normalized.

A prior which attempts to include no knowledge about the parameter is known as a *noninformative prior*. It's not always obvious what that should be, however. In the case of an event rate, we've used a uniform prior on the rate, which seemed pretty obvious, but note that this prior says that the rate is as likely to be between 1 and 2 events per hour as it is between 1001 and 1002 events per hour. It might seem more reasonable to choose a prior which gives the same probablity for the rate to be between 1 and 2 events per hour as between 1000 and 2000 events per hour. That would be uniform not in the rate θ but the logarithm $\lambda = \ln\theta$. If we write

$$\frac{dP}{d\theta} \sim \left|\frac{d\lambda}{d\theta}\right| \frac{dP}{d\lambda} = \frac{1}{\theta}\frac{dP}{d\lambda} \tag{3.4}$$

we see that $f_\Theta(\theta) = \frac{1}{\theta} f_\Lambda(\ln\theta)$ and therefore a uniform prior $f_\Lambda(\lambda) =$ constant corresponds to a prior $f_\Theta(\theta) \propto \frac{1}{\theta}$. Note that this is *also* a limiting form of the conjugate prior family Gamma(α,β), where $\alpha \rightarrow 0$ and $\beta \rightarrow \infty$. If we start with this prior and observe y events in a time t, our posterior is Gamma(y,$1/t$). This is then normalizable as long as $y > 0$, i.e., we observe at least one event.

3.3 Case study: Bernoulli trials

To illustrate the difficulty in defining a noninformative prior, consider the case of repeated Bernoulli trials with probability θ of success on each trial. We know that the Beta distribution forms a conjugate prior family for this parameter, i.e., if the prior is

$$f_\Theta(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{3.5}$$

and we see y successes in n trials, the posterior is

$$f_{\Theta|Y}(\theta|y) = \frac{\Gamma(\alpha+y)\Gamma(\beta+n-y)}{\Gamma(\alpha+\beta+n)}\theta^{\alpha+y-1}(1-\theta)^{\beta+n-y-1} \tag{3.6}$$

The seemingly obvious noninformative prior is uniform in θ, i.e., a Beta distribution with $\alpha = 1 = \beta$. But the expectation value of the resulting Beta($y+1$,$n-y+1$) posterior distribution is

$$\begin{aligned} E(\Theta|y) &= \frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)}\int_0^1 \theta^{y+1}(1-\theta)^{n-y}\,d\theta \\ &= \frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)}\frac{\Gamma(y+2)\Gamma(n-y+1)}{\Gamma(n+3)} = \frac{y+1}{n+2} \end{aligned} \tag{3.7}$$

which is probably different from the $\frac{y}{n}$ that we might have been expecting. But after a moment's reflection, it makes some sense. For example, if $y = 0$, a naïve $\frac{y}{n}$ estimate would give zero. But we know from the prior that there's some support for all $0 < \theta < 1$, so the mean of the posterior can't be at one end or the other of the range of permissible values. This probability is known as the Bayes-Laplace rule of succession, and this prior for the probability is the Bayes-Laplace prior.

The Bayes-Laplace rule of succession looks almost like we've added a "prior" two trials, one success and one failure, to the actual trials observed. We could ask what would happen if we left out those two trials. I.e., what is the prior such that if we conduct two trials–one success and one failure–the posterior is uniform in θ? Since a uniform distribution is just a Beta(1,1) distribution, we see that the desired prior is Beta(α,β) in the limit that α and β go to zero. This is not normalizable, but forms an improper prior known as the Haldane prior:

$$f_\Theta(\theta) \propto \theta^{-1}(1-\theta)^{-1} \tag{3.8}$$

Starting with the Haldane prior, we end with a posterior which is Beta(y,$n-y$), whose mean is y/n. However, the posterior will not be normalizable unless $0 < y < n$.

Thursday 11 February 2016 – Read Sections 11.2.4-11.2.5 of Hogg

4 Bayesian Hypothesis Testing

So far we've talked mostly about parameter estimation in a Bayesian framework. We've been able to write a posterior distrubtion for a parameter using Bayes's theorem

$$P(\Theta=\theta|\mathbf{X}=\mathbf{x},I) = \frac{P(\Theta=\theta|I)\,P(\mathbf{X}=\mathbf{x}|\Theta=\theta,I)}{P(\mathbf{X}=\mathbf{x}|I)} \tag{4.1}$$

where we've implicitly assumed discrete distributions to make the notation simpler, and put back in the conditioning on background information I. Built into that background information is a model that defines the prior probability distribution for the parameter Θ and even what parameter(s) determine the sampling distribution for the data. We can explicitly note that by replacing I with M, I, where M indicates the proposition that the model in question is the correct one. The denominator

$$P(\mathbf{X} = \mathbf{x}|M, I) = \sum_{\theta} P(\Theta = \theta|I)\, P(\mathbf{X} = \mathbf{x}|\Theta = \theta, I) \tag{4.2}$$

or, in the case of continuous distributions,

$$f_{\mathbf{X}}(\mathbf{x}|M, I) = \int_{-\infty}^{\infty} f_{\Theta}(\theta|M, I)\, f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta, M, I)\, d\theta \tag{4.3}$$

can be treated as a largely ignorable normalization factor if we're interested in possible values of Θ within the context of the model M. But we could turn this around and ask about the plausibility of the model itself in light of the data

$$P(M|\mathbf{X} = \mathbf{x}, I) = \frac{P(M|I)\, P(\mathbf{X} = \mathbf{x}|M, I)}{P(\mathbf{X} = \mathbf{x}|I)} \tag{4.4}$$

Now, literally working out this probability is typically impossible, since you have to know the prior probability $P(M|I)$ of the model, and to get the denominator you actually need to know about *every* feasible model given I. But what one can consider is the relative probabilities for two models M_1 and M_2 to be correct. This is

$$\frac{P(M_1|\mathbf{X} = \mathbf{x}, I)}{P(M_2|\mathbf{X} = \mathbf{x}, I)} = \left(\frac{P(M_1|I)}{P(M_2|I)}\right) \left(\frac{P(\mathbf{X} = \mathbf{x}|M_1, I)}{P(\mathbf{X} = \mathbf{x}|M_2, I)}\right) \tag{4.5}$$

and is known as the *odds ratio* between the two models. The troublesome denominator has cancelled out, and we see further that even the prior probability ratio $\frac{P(M_1|I)}{P(M_2|I)}$ factors out, and if we want to see how to update the prior ratio to the posterior one, we just need to calculate the second fraction, known as the *Bayes factor*

$$\frac{P(\mathbf{X} = \mathbf{x}|M_1, I)}{P(\mathbf{X} = \mathbf{x}|M_2, I)} = \frac{f_{\mathbf{X}}(\mathbf{x}|M_1, I)}{f_{\mathbf{X}}(\mathbf{x}|M_2, I)} \tag{4.6}$$

The Bayes factor is an indicator of how much we've shifted our relative assessment of the two models given the data. This is the reason why

$$f_{\mathbf{X}}(\mathbf{x}|M, I) = \int_{-\infty}^{\infty} f_{\Theta}(\theta|M, I)\, f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta, M, I)\, d\theta \tag{4.7}$$

is sometimes called the *evidence* for model M.

Note that one conceptually interesting choice is to define $M_1 = M$ and $M_2 = \overline{M}$, so that

$$\frac{P(M|\mathbf{X} = \mathbf{x}, I)}{1 - P(M|\mathbf{X} = \mathbf{x}, I)} = \left(\frac{P(M|I)}{1 - P(M|I)}\right) \left(\frac{P(\mathbf{X} = \mathbf{x}|M, I)}{P(\mathbf{X} = \mathbf{x}|\overline{M}, I)}\right) \tag{4.8}$$

but again this is difficult in practice, because in order to know about $\overline{M}$, you need to know about all of the other possible models.

4.1 Gaussian example

Suppose in both models the sampling distribution is a Gaussian, but in M_1 it's $N(0, \sigma^2)$, where σ is known while in M_2 the sampling distribution is $N(\theta, \sigma^2)$ and the prior on θ is $N(0, \sigma_0^2)$. For simplicity, assume we have a sample of size 1. (We could make this a sample of size n by replacing X with $\overline{X}$ and σ^2 with σ^2/n.) The evidence for M_1 is

$$f_X(x|M_1) = f_{X|\Theta}(x|0) = \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/(2\sigma^2)} \tag{4.9}$$

while for M_2 it is

$$\begin{aligned} f_X(x|M_2) &= \int_{-\infty}^{\infty} f_\Theta(\theta|M_2)\, f_{X|\Theta}(x|\theta)\, d\theta \\ &= \frac{1}{2\pi\sigma\sigma_0} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\left[\frac{\theta}{\sigma_0^2} + \frac{(x-\theta)^2}{\sigma^2}\right]\right) d\theta \end{aligned} \tag{4.10}$$

Completing the square gives

$$\frac{\theta}{\sigma_0^2}+\frac{(x-\theta)^2}{\sigma^2} = (\sigma_0^{-2}+\sigma^{-2})\left(\theta - \frac{\sigma^{-2}x}{\sigma_0^{-2}+\sigma^{-2}}\right)^2 + \frac{x^2}{\sigma^2} - \frac{\sigma^{-4}x^2}{\sigma_0^{-2}+\sigma^{-2}} \tag{4.11}$$

so

$$f_X(x|M_2) = \frac{e^{-x^2/(2\sigma^2)}}{\sigma\sigma_0\sqrt{2\pi(\sigma_0^{-2}+\sigma^{-2})}} \exp\left(\frac{x^2}{2\sigma^2(1+\sigma^2/\sigma_0^2)}\right) \tag{4.12}$$

and the Bayes factor is

$$\frac{f_X(x|M_2)}{f_X(x|M_1)} = \frac{1}{\sigma_0\sqrt{\sigma_0^{-2}+\sigma^{-2}}} \exp\left(\frac{x^2}{2\sigma^2(1+\sigma^2/\sigma_0^2)}\right) \tag{4.13}$$

Note that if $\sigma_0 \gg \sigma$, we can write the Bayes factor as

$$\frac{f_X(x|M_2)}{f_X(x|M_1)} \approx \frac{\sigma}{\sigma_0}\, e^{-x^2/(2\sigma^2)} \tag{4.14}$$

The second factor shows that the more detailed model M_2 can fit the observed data better, while the first factor acts as an “Occam factor” which penalizes it for having a tunable parameter.