Notes on Statistical Inference

ASTP 611-01: Statistical Methods for Astrophysics[*]

Fall Semester 2017

Contents

[*]

Tuesday, October 31, 2017

1 Methods of Inference

Our studies of probability theory have primarily shown us how to predict the outcome of experiments given some model and/or set of parameters, to calculate $\mathrm{P}(D|H,I)$ where D represents the data, H the hypothesis (possibly including parameter values) and I represents any other background information. The goal of statistical inference is to take the outcome of an experiment, and say something about the validity of one or more hypotheses.

From the Bayesian point of view, this is as simple as using Bayes's Theorem to construct

$$\mathrm{P}(H|D,I) = \frac{\mathrm{P}(D|H,I)\,\mathrm{P}(H|I)}{\mathrm{P}(D|I)} \tag{1.1}$$

In the frequentist approach, we're not allowed to assign probabilities to hypotheses, so instead we have to use $\mathrm{P}(D|H,I)$ to say something about the hypothesis H once we know the value of D. In practice, this often involves dividing up the space of possible values of D into a region $\mathcal{D}$ which is in some sense "consistent" with H, i.e., likely given H, so that $\sum_{D\in\mathcal{D}} \mathrm{P}(D|H,I)$ is above some threshold. But it's a bit arbitrary to choose such regions. After all, even if a coin is fair, the exact sequence HTTTTH-HTHTHTT is very unlikely (one in 2^{13}), but it's somehow more consistent with a fair coin than a string of 12 heads and a tail would be. So we often find ourselves constructing a statistic, a single function of the data which we can use for a simple threshold. So for example, to check if a coin is fair, given that we flipped k heads in n tries, we could take $(k-n/2)^2$. If this is small, it the number of heads is close to what we'd expect from a fair coin. This is a goodness-of-fit statistic, and the chi-square statistics we've constructed so far are examples.

Another example of a statistic would be if we want to estimate the probability parameter associated with a binomial distribution. Given k successes in n trials, we'd expect k/n to be a sensible estimate of this parameter. This discards any information about the order of successes and failures, and just retains that one number.

1.1 Statistics Constructed from Data: Two Approaches

To see how a preferred statistic might arise, let's consider the case where we have n data points $\{x_i\}$, drawn from independent distributions with the same unknown mean μ and different unknown variances $\{\sigma_i^2\}$. We are thus basically making n independent measurements of some unknown quantity μ, each with its own error of standard deviation σ_i. How can we use the values $\{x_i\}$ to say something about μ?

1.1.1 Bayesian Approach: Posterior pdf

The Bayesian answer to that question is straightforward: construct the posterior pdf

$$f(\mu|\{x_i\},\{\sigma_i\},I) = \frac{f(\{x_i\}|\mu,\{\sigma_i\},I)f(\mu|\{\sigma_i\},I)}{f(\{x_i\}|\{\sigma_i\},I)} \tag{1.2}$$

(From here on, we'll suppress the implicit conditional dependence on $\{\sigma_i\}$ and I in the interest of compactness of notation.)

To do this construction, we need to know the form of the joint pdf

$$f(\{x_i\}|\mu) = \prod_{i=1}^{n} f(x_i|\mu) \tag{1.3}$$

so let's add the additional assumption that the errors are Gaussian, so

$$f(x_i|\mu) = \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma_i^2}\right) \tag{1.4}$$

and

$$\begin{aligned} f(\{x_i\}|\mu) &= \prod_{i=1}^{n} \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma_i^2}\right) \\ &= \frac{1}{(2\pi)^{n/2}\prod_{i=1}^{n}\sigma_i} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma_i^2}\right) \end{aligned} \tag{1.5}$$

Although this is a pdf for the $\{x_i\}$, when we substitute this likelihood function into (1.2), we will end up with a pdf for μ, so we're most interested in the μ dependence, which we can see is Gaussian, since the sum in the exponential is quadratic in μ. We can write this in a transparent way by completing the square and writing

$$\chi^2(\{x_i\};\mu) = \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma_i^2} = \frac{[\mu-\mu_0(\{x_i\})]^2}{\sigma_\mu^2(\{x_i\})} + \chi_0^2(\{x_i\}) \tag{1.6}$$

and solving for $\mu_0(\{x_i\})$, $\sigma_\mu^2(\{x_i\})$, and $\chi_0^2(\{x_i\})$ (the names of which have been deliberately somewhat provocatively chosen). Expanding both sides gives us

$$\begin{aligned} &\mu^2\sum_{i=1}^{n}\frac{1}{\sigma_i^2} - 2\mu\sum_{i=1}^{n}\frac{x_i}{\sigma_i^2} + \sum_{i=1}^{n}\frac{x_i^2}{\sigma_i^2} \\ &\quad = \frac{\mu^2}{\sigma_\mu^2(\{x_i\})} - 2\mu\frac{\mu_0(\{x_i\})}{\sigma_\mu^2(\{x_i\})} + \frac{[\mu_0(\{x_i\})]^2}{\sigma_\mu^2(\{x_i\})} + \chi_0^2(\{x_i\}) \end{aligned} \tag{1.7}$$

so we can solve for

$$\frac{1}{\sigma_\mu^2(\{x_i\})} = \sum_{i=1}^{n}\frac{1}{\sigma_i^2} \tag{1.8}$$

(which we see is actually independent of the data $\{x_i\}$ so we will just write σ_μ from now on)

$$\mu_0(\{x_i\}) = \sigma_\mu^2(\{x_i\})\sum_{i=1}^{n}\frac{x_i}{\sigma_i^2} = \frac{\sum_{i=1}^{n}\sigma_i^{-2}x_i}{\sum_{i=1}^{n}\sigma_i^{-2}} \tag{1.9}$$

and

$$\chi_0^2(\{x_i\}) = \sum_{i=1}^{n}\frac{x_i^2}{\sigma_i^2} - \frac{[\mu_0(\{x_i\})]^2}{\sigma_\mu^2(\{x_i\})} \tag{1.10}$$

While $\chi_0^2(\{x_i\})$ is useful for some applications to come later in the semester, it will turn out to be irrelevant right now, so we don't bother to work out the explicit form.

We can rewrite the likelihood function to stress its μ dependence:

$$f(\{x_i\}|\mu) = \frac{e^{-\chi_0^2(\{x_i\})/2}}{(2\pi)^{n/2}\prod_{i=1}^{n}\sigma_i} \exp\left(-\frac{[\mu-\mu_0(\{x_i\})]^2}{2\sigma_\mu^2}\right) \tag{1.11}$$

Because the posterior pdf $f(\mu|\{x_i\})$ is guaranteed to be normalized:

$$\int_{-\infty}^{\infty} f(\mu|\{x_i\})\,d\mu = \frac{\int_{-\infty}^{\infty} f(\{x_i\}|\mu)\,f(\mu)\,d\mu}{f(\{x_i\})} = 1 \tag{1.12}$$

we can write

$$f(\mu|\{x_i\}) \propto \exp\left(-\frac{[\mu-\mu_0(\{x_i\})]^2}{2\sigma_\mu^2}\right) f(\mu) \tag{1.13}$$

where the $\{x_i\}$-dependent proportionality constant can be worked out from the normalization. Explicitly,

$$f(\mu|\{x_i\}) = \mathcal{C}(\{x_i\}) \exp\left(-\frac{[\mu - \mu_0(\{x_i\})]^2}{2\sigma_\mu^2}\right) f(\mu) \tag{1.14}$$

where

$$\begin{aligned}\mathcal{C}(\{x_i\}) &= \frac{e^{-\chi_0^2(\{x_i\})/2}}{f(\{x_i\})(2\pi)^{n/2}\prod_{i=1}^n \sigma_i} \\ &= \left(\int_{-\infty}^{\infty} \exp\left(-\frac{[\mu - \mu_0(\{x_i\})]^2}{2\sigma_\mu^2}\right) f(\mu)\, d\mu\right)^{-1}\end{aligned} \tag{1.15}$$

In particular, if the prior pdf $f(\mu)$ is constant[1], the posterior is

$$f(\mu|\{x_i\}) = \frac{1}{\sigma_\mu^2\sqrt{2\pi}} \exp\left(-\frac{[\mu - \mu_0(\{x_i\})]^2}{2\sigma_\mu^2}\right) \tag{1.16}$$

In any event, no matter what the prior on μ, the essential information about the outcome of the experiment is encoded in the weighted average $\mu_0(\{x_i\})$.

1.1.2 Frequentist Approach: Optimal Estimator

Now let's shift to the frequentist perspective, where we have n random variables $\{X_i\}$ with unknown mean $E[X_i] = \mu$ and known variances $\operatorname{Cov}(X_i, X_j) = \delta_{ij}\operatorname{Var}(X_i) = \delta_{ij}\sigma_i^2$. We want to say something about μ, and the simplest thing we can do is try to estimate its value. So we construct a statistic $\widehat{\mu}(\{X_i\})$. This is a random variable, and for any data realization it is our guess for the value of μ. Since it's a random variable, it has an expectation value. We say that $\widehat{\mu}$ is an *unbiased estimator* of μ if $E[\widehat{\mu}(\{X_i\})] = \mu$. There are a lot of possible statistics which satisfy this requirement. For instance, we could just take X_1 and throw away the rest of the data. Or we could take the sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$. Either one of these is an unbiased estimator (since they both have expectation value μ), but we'd expect $\overline{X}$ to do a better job of estimating μ. On the other hand, it won't be the best in all cases; for example, if σ_2 is much less than all the other $\{\sigma_i\}$, i.e., the second measurement is good and the others are all lousy, we'd like to pay more attention to X_2 than the other random variables.

We'd like to consider what is the best estimator $\widehat{\mu}$ to use. For simplicity, let's restrict ourselves to linear combinations of the random variables, i.e., estimators of the form

$$\widehat{\mu}(\{X_i\}) = \sum_{i=1}^n a_i X_i \tag{1.17}$$

the estimator will be unbiased if

$$E[\widehat{\mu}(\{X_i\})] = \sum_{i=1}^n a_i\mu = \mu\sum_{i=1}^n a_i \tag{1.18}$$

is equal to μ, i.e., if $\sum_{i=1}^n a_i = 1$. The variance of the estimator is

$$\operatorname{Var}(\widehat{\mu}(\{X_i\})) = \sum_{i=1}^n\sum_{j=1}^n a_i a_j \operatorname{Cov}(X_i, X_j) = \sum_{i=1}^n a_i^2\sigma_i^2 \tag{1.19}$$

The *optimal estimator* is the unbiased estimator with the lowest variance, i.e., it minimizes $\sum_{i=1}^n a_i^2\sigma_i^2$ subject to the constraint

[1] In practice, to have a normalizable prior, we need something like

$$f(\mu) = \begin{cases} \frac{1}{\mu_{\max}-\mu_{\min}} & \mu_{\min} < \mu < \mu_{\max} \\ 0 & \text{otherwise} \end{cases}$$

but in the limit $\mu_{\min} \ll \mu_0 - \sigma$ and $\mu_{\max} \gg \mu_0 + \sigma$ we get the simpler result given here.

$\sum_{i=1}^n a_i = 1$. We can find this with the method of Lagrange multipliers, by minimizing

$$\sum_{i=1}^n a_i^2 \sigma_i^2 + \lambda(\sum_{i=1}^n a_i - 1) \tag{1.20}$$

with respect to $\{a_i\}$ and λ. Taking $\frac{\partial}{\partial a_i}$ gives

$$2a_i\sigma_i^2 + \lambda = 0 \tag{1.21}$$

so

$$a_i = -\frac{\lambda}{2}\sigma_i^{-2} \tag{1.22}$$

Taking $\frac{\partial}{\partial\lambda}$ gives the constraint $\sum_{i=1}^n a_i = 1$ so

$$-\frac{\lambda}{2}\sum_{i=1}^n \sigma_i^{-2} = 1 \tag{1.23}$$

i.e.,

$$-\frac{\lambda}{2} = \frac{1}{\sum_{i=1}^n \sigma_i^{-2}} \tag{1.24}$$

and

$$a_i = \frac{\sigma_i^{-2}}{\sum_{j=1}^n \sigma_j^{-2}} \tag{1.25}$$

which makes the optimal estimator

$$\widehat{\mu}_{\text{opt}}(\{X_i\}) = \frac{\sum_{i=1}^n \sigma_i^{-2} X_i}{\sum_{i=1}^n \sigma_i^{-2}} \tag{1.26}$$

and its variance

$$\operatorname{Var}(\widehat{\mu}_{\text{opt}}(\{X_i\})) = \sum_{i=1}^n a_i^2\sigma_i^2 = \frac{\sum_{i=1}^n \sigma_i^{-4}\sigma_i^2}{(\sum_{j=1}^n \sigma_j^{-2})^2} = \frac{1}{\sum_{i=1}^n \sigma_i^{-2}} \tag{1.27}$$

but we see that this optimal estimator is just the same weighted average that showed up in the posterior pdf for μ in the Bayesian approach:

$$\widehat{\mu}_{\text{opt}}(\{x_i\}) = \mu_0(\{x_i\}) \tag{1.28}$$

and its variance is the width of the posterior on μ in the case where the prior is uniform and the sampling distribution is Gaussian.

$$\operatorname{Var}(\widehat{\mu}_{\text{opt}}(\{X_i\})) = \sigma_\mu^2 \tag{1.29}$$

Wednesday, November 1, 2017

2 Parameter Estimation

Our preceding example considered two related questions about unknown parameters

- What posterior distribution do we assign to an unknown parameter in light of observed data, in the Bayesian framework?
- How can we estimate an unknown parameter given observed data?

In addition to the Bayesian vs frequentist issues, there are also differences between trying to get a single point estimate of a parameter, and saying something about the uncertainty associated with that estimate.

2.1 Maximum likelihood and maximum a posteriori estimates

2.1.1 Maximum likelihood estimation

As we saw previously, there are many different estimators that could conceivably be used to try to gain information about an unknown parameter θ. One way, in the frequentist picture, to pick an estimate is the so-called maximum likelihood method, which chooses the value that maximizes the likelihood function $f(\{x_i\}|\theta)$ where $\{x_i\}$ are the observed data.

In the previous example, where the parameter was μ, the likelihood function was

$$\begin{aligned} f(\{x_i\}|\mu) &= \frac{1}{(2\pi)^{n/2}\prod_{i=1}^{n}\sigma_i} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma_i^2}\right) \\ &= \frac{e^{-\chi_0^2(\{x_i\})/2}}{(2\pi)^{n/2}\prod_{i=1}^{n}\sigma_i} \exp\left(-\frac{[\mu-\mu_0(\{x_i\})]^2}{2\sigma_\mu^2}\right) \end{aligned} \tag{2.1}$$

where $\frac{1}{\sigma_\mu^2} = \sum_{i=1}^{n}\frac{1}{\sigma_i^2}$, $\mu_0(\{x_i\}) = \frac{\sum_{i=1}^{n}\sigma_i^{-2}x_i}{\sum_{i=1}^{n}\sigma_i^{-2}}$, and $\chi_0^2(\{x_i\}) = \sum_{i=1}^{n}\frac{x_i^2}{\sigma_i^2} - \frac{[\mu_0(\{x_i\})]^2}{\sigma_\mu^2(\{x_i\})}$. We can see by inspection that the likelihood function (which happens to be a Gaussian) is maximized when $\mu = \mu_0(\{x_i\})$.

As another example, consider a random sample $\{X_i\}$ of size n drawn from an exponential distribution with rate parameter θ. Since each random variable X_i is drawn from the pdf $f(x_i|\theta) = \theta e^{\theta x_i}$, the likelihood function is

$$f(\{x_i\}|\theta) = \prod_{i=1}^{n} f(x_i|\theta) = \theta^n e^{-\theta\sum_{i=1}^{n}x_i} \tag{2.2}$$

It's actually easiest to find the θ that maximizes the likelihood by considering the log-likelihood

$$\ell(\theta) = \ln f(\{x_i\}|\theta) = n\ln\theta - \theta\sum_{i=1}^{n}x_i \tag{2.3}$$

whose derivative is

$$\ell'(\theta) = \frac{n}{\theta} - \sum_{i=1}^{n}x_i \tag{2.4}$$

so the maximum-likelihood rate is

$$\widehat{\theta} = \frac{n}{\sum_{i=1}^{n}x_i} = \frac{1}{\overline{x}} \tag{2.5}$$

2.1.2 MAP estimation

Note that in the Bayesian approach we could simply find the value of θ which maximizes $f(\theta|\{x_i\}) \propto f(\{x_i\}|\theta)f(\theta)$, which is known as the maximum a posteriori (MAP) estimate. If the prior $f(\theta)$ is uniform, the MAP estimate is the same as the maximum likelihood estimate. Note, though, that if we do a change of variables on the parameter, the maximum-likelihood point won't change, but the maximum-posterior point will. For instance, if we parametrize the exponential distribution in terms of a rate parameter $\beta = \theta^{-1}$, the likelihood function is

$$f(\{x_i\}|\beta) = \beta^{-n}e^{-\sum_{i=1}^{n}x_i/\beta} \tag{2.6}$$

and the derivative of the log-likelihood is

$$\frac{d}{d\beta}\ln f(\{x_i\}|\beta) = \frac{-n}{\beta} + \frac{\sum_{i=1}^{n}x_i}{\beta^2} \tag{2.7}$$

which is zero when

$$\beta = \frac{\sum_{i=1}^{n}x_i}{n} = \widehat{\theta}^{-1} \tag{2.8}$$

The reason this doesn't work for the maximum-posterior point is that $f(\theta|\{x_i\})$ is a density in θ, while $f(\{x_i\}|\theta)$ is not. On the one hand,

$$f_{X|\mathrm{B}}(x|\beta) = f_{X|\Theta}(x|\beta^{-1}) \tag{2.9}$$

because the condition $\mathrm{B} = \beta$ is the same as the condition $\Theta = \beta^{-1}$, but if we transform the pdf,

$$f_{\mathrm{B}|X}(\beta|x) = \frac{dP}{d\beta} = \left|\frac{d\theta}{d\beta}\right|\frac{dP}{d\theta} = \beta^{-2}f_{\Theta|X}(\beta^{-1}|x) \tag{2.10}$$

(By the same token, the statement "the prior is uniform in the parameter" depends on what the parameter is. If $f_\Theta(\theta)$ is a constant, $f_{\mathrm{B}}(\beta) = \beta^{-2} f_\Theta(\beta^{-1})$ can't be.

2.1.3 Expansion about the MAP point

Presuming we've described the parameter space in a convenient set of coördinates, we can ask how the posterior $f(\boldsymbol{\theta}|\mathbf{x})$ behaves in the vicinity of the parameter value which maximizes it. Consider first the case of a single parameter θ, with posterior $f(\theta|\mathbf{x})$. One trick would be to Taylor expand the function near its maximum, but this could cause trouble if we extrapolate it too far, since we know $f(\theta|\mathbf{x}) \geq 0$. So instead, we Taylor expand the logarithm $\ell(\theta) = \ln f(\theta|\mathbf{x})$. For convenience we'll call the MAP point $\widehat{\theta}$ and the log-posterior $\ell(\theta)$, respectively, even though these usually refer to the maximum likelihood point and the log-likelihood, respectively. The expansion looks like

$$\ell(\theta) = \ell(\widehat{\theta}) + (\theta - \widehat{\theta})\ell'(\widehat{\theta}) + \frac{(\theta - \widehat{\theta})^2}{2}\ell''(\widehat{\theta}) + \cdots \tag{2.11}$$

Now, since $\widehat{\theta}$ maximizes $\ell(\theta)$, we know $\ell'(\widehat{\theta}) = 0$ and $\ell''(\widehat{\theta}) < 0$. If we truncate the expansion at the first non-trivial order, we have

$$\ell(\theta) \approx \ell(\widehat{\theta}) - \frac{(\theta - \widehat{\theta})^2}{2}[-\ell''(\widehat{\theta})] \tag{2.12}$$

or

$$f(\theta|\mathbf{x}) \approx f(\widehat{\theta}|\mathbf{x}) \exp\left(-\frac{(\theta - \widehat{\theta})^2}{2}[-\ell''(\widehat{\theta})]\right) \tag{2.13}$$

which is a Gaussian with width $[-\ell''(\widehat{\theta})]^{-1/2}$.

In the case where there are multiple parameters, $\boldsymbol{\theta} \equiv \{\theta_i | i = 1, \ldots, m\}$, the Taylor expansion of $\ell(\boldsymbol{\theta}) = \ln f(\boldsymbol{\theta}|\mathbf{x})$ is

$$\ell(\boldsymbol{\theta}) \approx \ell(\widehat{\boldsymbol{\theta}}) + \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \frac{\partial^2 \ell}{\partial\theta_i \partial\theta_j}(\theta_i - \widehat{\theta}_i)(\theta_j - \widehat{\theta}_j) \tag{2.14}$$

so that

$$f(\boldsymbol{\theta}|\mathbf{x}) \approx f(\widehat{\boldsymbol{\theta}}|\mathbf{x}) \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^{\mathrm{T}}\mathbf{H}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})\right) \tag{2.15}$$

where $\mathbf{H}$ is the Hessian matrix, which has elements

$$H_{ij}(\mathbf{x}) = -\left.\frac{\partial^2 \ln f(\boldsymbol{\theta}|\mathbf{x})}{\partial\theta_i \partial\theta_j}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}(\mathbf{x})} \tag{2.16}$$

The Hessian matrix gives an estimate of uncertainties of the parameters; $\mathbf{H}$ is just the inverse of the variance-covariance matrix for the approximate multivariate Gaussian posterior:

$$\mathrm{Cov}(\boldsymbol{\theta}) = E\left[(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^{\mathrm{T}}\right] \approx \mathbf{H}^{-1} \tag{2.17}$$

In particular, the width of the marginal pdf for a particular parameter is

$$\sqrt{\mathrm{Var}(\theta_i)} = E\left[(\theta_i - \widehat{\theta}_i)^2\right] \approx \sqrt{[\mathbf{H}^{-1}]_{ii}} \tag{2.18}$$

This is one justification for the practice of quoting $\sqrt{[\mathbf{H}^{-1}]_{ii}}$ as the one-sigma uncertainty for the parameter θ_i.

Note that if the Hessian matrix has off-diagonal elements, it's important to take the diagonal elements of the inverse Hessian matrix rather than one over the diagonal elements of the Hessian matrix, since

$$\left[\mathbf{H}^{-1}\right]_{ii} \neq \frac{1}{H_{ii}} \tag{2.19}$$

In general $(H_{ii})^{-1/2}$ will be an underestimate of the correct error $([\mathbf{H}^{-1}]_{ii})^{1/2}$, as you showed in your consideration of the bivariate Gaussian distribution on the homework.

2.1.4 The Fisher matrix

The Hessian matrix (2.16) $\mathbf{H}(\mathbf{x})$ is closely related to the Fisher Information Matrix $\mathbf{F}(\boldsymbol{\theta})$ of classical statistics, which is defined as

$$F_{ij}(\boldsymbol{\theta}) = E\left[-\frac{\partial^2 \ln f(\mathbf{X}|\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}\right] \tag{2.20}$$

They differ in that

1. The Fisher matrix is constructed from the likelihood $f(\mathbf{x}|\boldsymbol{\theta})$ rather than the posterior $f(\boldsymbol{\theta}|\mathbf{x})$.
2. While the Hessian is a function of the observed data $\mathbf{x}$, evaluated at the MAP point $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}(\mathbf{x})$, the Fisher matrix is a function of the parameter space point $\boldsymbol{\theta}$, and any $\mathbf{x}$-dependence in the second derivative is handled via an expectation value.

These differences are irrelevant, and the Fisher matrix is identical to the Hessian, if the prior $f(\boldsymbol{\theta})$ is uniform so that the log-posterior and log-likelihood only differ by a constant, and the second derivative is independent of both $\boldsymbol{\theta}$ and $\mathbf{x}$.

Thursday, November 2, 2017

2.2 Interval estimation

Beyond finding some "most likely" parameter value and describing the shape of either the likelihood function or the posterior around that value, an important task in parameter estimation is to provide an interval that we associate quantitatively with likely values of the parameter. This can be extended to a region in a multidimensional parameter space. The biggest difference between the Bayesian and frequentist versions of these intervals turns out to be the interpretation.

2.2.1 Bayesian plausible intervals

We start with the Bayesian version, which is considerably more straightforward. Given a posterior pdf $f(\theta|\mathbf{x})$, we can construct a *plausible interval* in which we think θ is likely to lie with some probability $1-\alpha$, defined by

$$\mathrm{P}(\theta_\ell < \theta < \theta_u) = \int_{\theta_\ell}^{\theta_u} f(\theta|\mathbf{x})\, d\theta = 1-\alpha \tag{2.21}$$

So this means the area under the posterior pdf, between θ_ℓ and θ_u, is $1-\alpha$. This does leave the freedom to choose where the interval begins. Some convenient choices are

- A lower limit (one-sided plausible interval), so $\mathrm{P}(\theta_\ell < \theta) = 1-\alpha$.
- An upper limit (one-sided plausible interval), so $\mathrm{P}(\theta < \theta_u) = 1-\alpha$.
- A symmetric two-sided plausible interval, so $\mathrm{P}(\theta < \theta_\ell) = \alpha/2 = \mathrm{P}(\theta_u < \theta)$.
- A plausible interval centered on the mode $\widehat{\theta}$ of the posterior, so $\mathrm{P}(\widehat{\theta} - \frac{\Delta\theta}{2} < \theta < \widehat{\theta} + \frac{\Delta\theta}{2}) = 1-\alpha$.
- The narrowest possible plausible interval, i.e., of all of the intervals with $\mathrm{P}(\theta_\ell < \theta < \theta_u) = 1-\alpha$, pick the one that minimizes $\theta_u - \theta_\ell$. You can show that a necessary condition for this is $f(\theta_\ell|\mathbf{x}) = f(\theta_u|\mathbf{x})$.

2.2.2 Frequentist confidence intervals

In the frequentist picture we can't assign a probability to the statement that a particular interval contains or doesn't contain an unknown parameter. It either does or it doesn't. So instead we can define a procedure to generate an interval such that if you collect many random data sets and make such an interval from each, some fraction of those intervals will contain the true parameter value. This is known as a (frequentist) confidence interval. It's a pair of statistics $L = L(\mathbf{X})$ and $U = U(\mathbf{X})$ chosen so that the probability that the parameter θ lies between them is $1 - \alpha$ (e.g., if $\alpha = 0.10$, it is 90%):[2]

$$\mathrm{P}(L < \theta < U) = 1 - \alpha \tag{2.22}$$

It's important to note that the probabilities here refer to the randomness of L and U, and not to the unknown θ. From the frequentist perspective, we can't talk about probabilities for different values of θ; it has some specific value, even if it's unknown. What's random is the sample $\mathbf{X}$ and the statistics L and U created from it.

Given a particular realization $\mathbf{x}$ of the sample $\mathbf{X}$, we have a specific confidence interval between $\ell = L(\mathbf{x})$ and $u = U(\mathbf{x})$. Note that the probabilistic statements do not actually refer to the properties of a particular confidence interval (ℓ, u) but to the procedure used to construction of the confidence interval.

One method to construct the confidence interval is to choose a statistic $T = T(\mathbf{X}; \theta)$, known as a *pivot variable*, whose probability distribution is a known function of the parameters, and construct an interval using the percentiles of the distribution

$$\mathrm{P}(a < T(\mathbf{X}; \theta) < b) = 1 - \alpha \tag{2.23}$$

By algebraically solving the inequalities $a < T(\mathbf{X}; \theta)$ and $T(\mathbf{X}; \theta) < b$ for θ, we should be able to write

$$\mathrm{P}(L(\mathbf{X}) < \theta < U(\mathbf{X})) = 1 - \alpha \tag{2.24}$$

Note that this construction is not unique; different choices for the pivot variable will give different confidence intervals with the same confidence.

2.2.3 Example: Mean of a Normal Distribution

To illustrate the pivot variable method, consider the case where $\mathbf{X}$ is a sample of size n drawn from a $N(\mu, \sigma)$ distribution with both μ and σ unknown, where we want a confidence interval on μ. The pivot variable should depend on μ and $\mathbf{X}$ but not σ, so

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \tag{2.25}$$

will not work, even though we know it obeys as $N(0, 1)$ distribution (because $\overline{X}$ obeys a normal distribution with $E(\overline{X}) = \mu$ and $\mathrm{Var}(\overline{X}) = \sigma/\sqrt{n}$. Fortunately, we know from Student's theorem that

$$T = \frac{\overline{X} - \mu}{\sqrt{S^2/n}} \tag{2.26}$$

obeys a t distribution with $n - 1$ degrees of freedom. This will work as a pivot variable, since

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2 \tag{2.27}$$

depends only on the sample, and requires no knowledge of μ or σ. Having identified a pivot variable which obeys a t distribution is useful not so much because we know the precise form of the pdf

$$f_T(t; \nu) = \frac{\Gamma([\nu + 1]/2)}{\sqrt{\nu\pi}\,\Gamma(\nu/2)} \left(1 + \frac{t^2}{\nu}\right)^{-[\nu+1]/2} \tag{2.28}$$

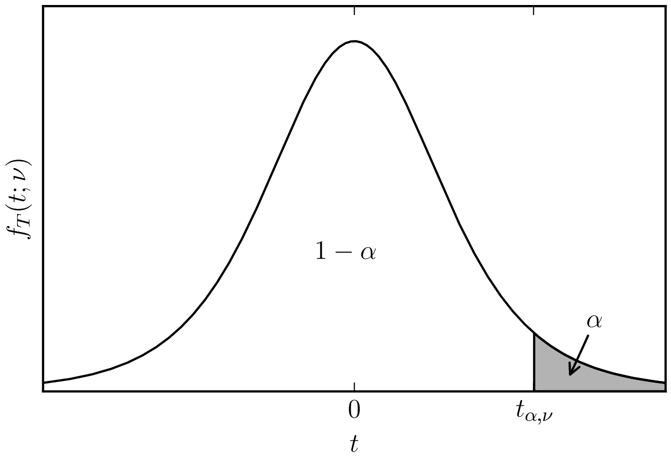
but because it's a standard distribution for which the percentiles are tabulated in various books or available in R, scipy, etc. The 90th percentile, for example, of a t distribution with ν degrees of freedom is written $t_{0.1,\nu}$; in general, the $(1-\alpha) \times 100$th percentile $t_{\alpha,\nu}$ is defined by

$$1 - \alpha = \mathrm{P}(T \le t_{\alpha,\nu}) = \int_{-\infty}^{t_{\alpha,\nu}} f_T(t; \nu)\, dt \tag{2.29}$$

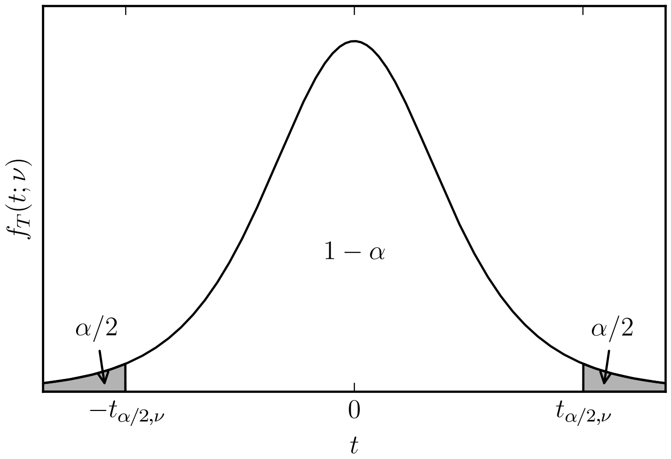
[2] We're implicitly considering a *two-sided* confidence interval, so we also have $\mathrm{P}(\theta < L) = \alpha/2$ and $\mathrm{P}(U < \theta) = \alpha/2$.

or equivalently by

$$\int_{t_{\alpha,\nu}}^{\infty} f_T(t;\nu)\,dt = \alpha \tag{2.30}$$



Since we want a two-sided confidence interval, we actually need $t_{\alpha/2,\nu}$ and $t_{1-\alpha/2,\nu}$. Since the t distribution is symmetric, though, we can take advantage of the fact that $t_{1-\alpha/2,\nu} = -t_{\alpha/2,\nu}$, e.g., the 5th percentile is minus the 95th:



Thus, returning to the case of the pivot variable T, which is t-distributed with $n-1$ degrees of freedom,

$$\begin{aligned} 1-\alpha &= \mathrm{P}(-t_{\alpha/2,n-1} < T < t_{\alpha/2,n-1}) \\ &= P\left(-t_{\alpha/2,n-1} < \frac{\overline{X}-\mu}{\sqrt{S^2/n}} < t_{\alpha/2,n-1}\right) \end{aligned} \tag{2.31}$$

Doing a bit of algebra, we can see that

$$\frac{\overline{X}-\mu}{\sqrt{S^2/n}} < t_{\alpha/2,n-1} \tag{2.32}$$

is equivalent to

$$\overline{X} - t_{\alpha/2,n-1}\sqrt{\frac{S^2}{n}} < \mu \tag{2.33}$$

and

$$-t_{\alpha/2,n-1} < \frac{\overline{X}-\mu}{\sqrt{S^2/n}} \tag{2.34}$$

is equivalent to

$$\mu < \overline{X} + t_{\alpha/2,n-1}\sqrt{\frac{S^2}{n}} \tag{2.35}$$

so

$$P\left(\overline{X} - t_{\alpha/2,n-1}\sqrt{\frac{S^2}{n}} < \mu < \overline{X} + t_{\alpha/2,n-1}\sqrt{\frac{S^2}{n}}\right) = 1-\alpha \tag{2.36}$$

which defines a confidence interval for μ.

Tuesday, November 7, 2017

3 Model Selection

3.1 Frequentist hypothesis testing

See Gregory, Chapter 7

Often want to evaluate hypothesis $\mathcal{H}$ in light of observed data $\mathbf{x}$, or compare hypotheses

In Bayesian picture, can define $P(\mathcal{H}|\mathbf{x})$ and evaluate e.g., $\frac{P(\mathcal{H}_1)}{P(\mathcal{H}_2)}$

So far, we've considered methods to get a handle on the unknown parameter(s) θ of a probability distribution $f(x;\theta)$ given that we draw a sample $\mathbf{X}$ from that distribution, with joint pdf

$$f_{\mathbf{X}}(x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i;\theta) \tag{3.1}$$

and find a particular realization $\mathbf{X} = \mathbf{x}$. Now we want to consider how to use the realization of the sample to distinguish between two competing hypotheses about what the underlying distribution $f(x)$ is. In principle the differences could be qualitative, but for simplicity we'll assume that there is one family $f(x;\theta)$ parametrized by θ which lies somewhere in a region Ω and then take the hypotheses to be:

- $\mathcal{H}_0$: the distribution is $f(x;\theta)$ where $\theta \in \omega_0$.
- $\mathcal{H}_1$: the distribution is $f(x;\theta)$ where $\theta \in \omega_1$.

Typically, $\mathcal{H}_0$ represents the absence of the effect we're looking for, and is known as the *null hypothesis*, while $\mathcal{H}_1$ represents the presence of the effect, and is known as the *alternative hypothesis*.

For example, suppose someone claims to have extrasensory perception, and to be able to use their telepathic powers to determine the suits of cards drawn from a deck. For simplicity, assume we shuffle the deck after each draw. Then the data $\{X_i\}$ are a sample drawn from a Bernoulli distribution, with each X_i having some probability θ of being correct. The null hypothesis $\mathcal{H}_0$ is that the person does not have ESP, and has a 25% chance of guessing each suit correctly, so $\theta = 0.25$. The alternative hypothesis $\mathcal{H}_1$ is that they can determine the suit more accurately than by random chance (but perhaps not perfectly), so $\theta > 0.25$.

As another example, suppose that someone claims that when twins are born, the birth weight of the first twin is on average greater than that of the second. We could take the data $\{X_i\}$ to be the difference between the birth weights of the two twins, and assume that the weights are normally distributed with unknown variance. Then the null hypothesis $\mathcal{H}_0$ is that $f(x)$ is a normal distribution with mean $\mu = 0$ and standard deviation $\sigma > 0$, while the alternative hypothesis $\mathcal{H}_1$ is that $f(x)$ is a normal distribution with mean $\mu > 0$ and standard deviation $\sigma > 0$. (In this case there is a vector of parameters $\boldsymbol{\theta} = (\mu, \sigma)$.

A hypothesis test is simply a rule for choosing between the two hypotheses depending on the realization $\mathbf{x}$ of the sample $\mathbf{X}$. Stated most generally, we construct a critical region C which is a subset of the n-dimensional sample space $\mathcal{D}$. If $\mathbf{X} \in C$, we "reject the null hypothesis $\mathcal{H}_0$", i.e., we favor $\mathcal{H}_1$. If $\mathbf{X} \notin C$, i.e., $\mathbf{X} \in C^c$ we "accept the null hypothesis $\mathcal{H}_0$", i.e., we favor $\mathcal{H}_0$ over $\mathcal{H}_1$. Now of course, since $\mathbf{X}$ is random, there will be some probability $P(\mathbf{X} \in C;\theta)$ that we'll reject the null hypothesis, which depends on the value of θ. If the test were perfect, that probability would be 0 if $\mathcal{H}_0$ were true, i.e., for any $\theta \in \omega_0$, and 1 if $\mathcal{H}_1$ were true, i.e., for any $\theta \in \omega_1$, but then we wouldn't be doing statistics. So instead there is some chance we will choose the "wrong" hypothesis, i.e., some probability that, given a value of $\theta \in \omega_0$ associated with $\mathcal{H}_0$, the realization of our data will cause us to reject $\mathcal{H}_0$, and some probability that, given a value of $\theta \in \omega_1$ associated with $\mathcal{H}_1$, the realization of our data will cause us to accept $\mathcal{H}_0$. As a bit of nomenclature,

- If $\mathcal{H}_0$ is true and we reject $\mathcal{H}_0$, this is called a *Type I Error* or a false positive.
- If $\mathcal{H}_1$ is true and we reject $\mathcal{H}_0$, we have made a correct decision (true positive).
- If $\mathcal{H}_0$ is true and we accept $\mathcal{H}_0$, we have made a correct decision (true negative).

- If $\mathcal{H}_1$ is true and we accept $\mathcal{H}_0$, this is called a *Type II Error* or a false negative.

Typically, a false positive is considered worse than a false negative, so usually we decide how high a false positive probability we can live with and then try to find the test which gives us the lowest false negative probability.

Given a critical region C, we'd like to talk about the associated false positive probability α and false negative probability $1-\gamma$, but we have to be a bit careful, since $\mathcal{H}_0$ and $\mathcal{H}_1$ are in general *composite hypotheses*. This means that each of them corresponds not to a single parameter value θ and thus a single distribution, but rather to a range of values $\theta \in \omega_0$ or $\theta \in \omega_1$. So both α and γ may depend on the value of θ. We take the false alarm probability α to be the worst-case scenario within the null hypothesis

$$\alpha = \max_{\theta\in\omega_0} \mathrm{P}(\mathbf{X} \in C; \theta) \tag{3.2}$$

This is also called the *size* of the critical region C. Somewhat confusingly, it's also referred to as the *significance* of the test. This is a bit counter intuitive, since a low value of α means the probability of a false positive is low, which means a positive result is *more* significant than if α were higher. It is the probability that we'll falsely reject the null hypothesis $\mathcal{H}_0$, maximized over any parameters within the range associated with $\mathcal{H}_0$. On the other hand, since the alternative hypothesis almost always has a parameter θ associated with it, we define the probability of correctly rejecting the null hypothesis (which is one minus the probability of a false negative) as a function of θ:

$$\gamma_C(\theta) = \mathrm{P}(\mathbf{X} \in C; \theta), \qquad \theta \in \omega_1 \tag{3.3}$$

We explicitly consider this as a function of the critical region C, since we might want to compare different tests with the same false alarm probability α (critical regions with the same size α) to see which is more powerful.

3.2 Example: Binomial Proportion

To give a concrete example, consider the ESP test described above. We let the would-be psychic predict the suit of n cards, count the total number of successes $Y = \sum_{i=1}^{n} X_i$, and reject the null hypothesis if $Y > k$ where k is some integer we've chosen, with $k > n/4$. For both of the hypotheses, Y is a binomial random variable, so

$$\mathrm{P}(Y > k) = \sum_{i=k+1}^{n} \binom{n}{i} \theta^i (1-\theta)^{n-i} = 1 - F(k;\theta) \tag{3.4}$$

where

$$F(k;\theta) = \sum_{i=0}^{k} \binom{n}{i} \theta^i (1-\theta)^{n-i} \tag{3.5}$$

is the cdf of a binomial distribution $b(n,\theta)$. For the null hypothesis $\theta = 0.25$ and for the alternative hypothesis $0.25 < \theta < 1$. Thus the false alarm probability is

$$\alpha = 1 - F(k; 0.25) \tag{3.6}$$

and the power of the test is

$$\gamma_k(\theta) = 1 - F(k;\theta) \tag{3.7}$$

If we make the threshold k higher, we get a lower false alarm probability α, but we also get a less powerful test.

As a concrete example, suppose that $n = 20$, and we set a threshold of $k = 8$. We can use scipy, invoked by

```
ipython --pylab
```

to calculate the false alarm probability

```
In [1]: from scipy.stats import binom

In [2]: n = 20

In [3]: k = 8

In [4]: alpha = 1 - binom.cdf(k,n,0.25); alpha
Out[4]: 0.04092516770651855
```

So $\alpha \approx 0.041 = 4.1\%$. The power $\gamma(\theta)$ depends on the strength of the ESP effect, but suppose $\theta = 0.50$, that the psychic has a 1 in 2 chance rather than 1 in 4 of picking the right suit. Then we can calculate the power:

```
In [5]: gamma_50 = 1 - binom.cdf(k,n,0.50); gamma_50
Out[5]: 0.74827766418457031
```

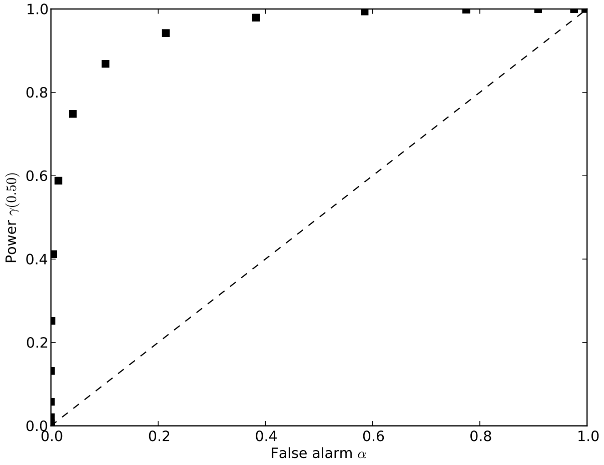
so $\gamma(0.50) \approx 0.748 = 74.8\%$.

3.2.1 Aside: ROC Curves

We could make the test more powerful by lowering the threshold k, but then we would also increase the false alarm probability α. A useful construction is the *receiver operating characteristic* curve, or ROC curve for short. Given a value of θ, we plot α versus $\gamma(\theta)$ for a range of threshold values k. We can do this with matplotlib as well, using the `arange` function to generate an array of integer values for k between 0 and 19:

```
In [6]: k = arange(20)

In [7]: alpha = 1 - binom.cdf(k,n,0.25)

In [8]: gamma_50 = 1 - binom.cdf(k,n,0.50)

In [9]: plot(alpha,gamma_50,'ks');

In [10]: xlabel(r'False alarm $\alpha$');

In [11]: ylabel(r'Power $\gamma(0.50)$');

In [12]: plot([0,1],[0,1],'k--');

In [13]: savefig('roc.eps');
```

The plot looks like this:



The diagonal line is $\gamma = \alpha$; we don't expect any sensible test to lie below this line, since it would mean that we were more likely to reject $\mathcal{H}_0$ when it's true than when $\mathcal{H}_1$ is true!

3.3 Example: Mean of a Normal Distribution

Consider the second example, where **X** is a random sample of size n from a normal distribution, where the null hypothesis

$\mathcal{H}_0$ is $\mu = 0$ and the alternative hypothesis $\mathcal{H}_1$ is $\mu > 0$. For simplicity, let's assume that the variance σ^2 is actually known. (If the sample is large enough, we can use the sample variance s^2 as an estimate.) From our work on confidence intervals, we know that

$$P\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} > z_\alpha\right) = \alpha \tag{3.8}$$

So if we define a critical region

$$C \equiv \frac{\overline{X}}{\sigma/\sqrt{n}} > z_\alpha \tag{3.9}$$

this will correspond to a test with false alarm rate α. The power of the test for a given true value of μ is

$$\begin{aligned}\gamma(\mu) &= P\left(\frac{\overline{X}}{\sigma/\sqrt{n}} > z_\alpha\right) = P\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} > z_\alpha - \frac{\mu}{\sigma/\sqrt{n}}\right) \\ &= 1 - \Phi\left(z_\alpha - \frac{\mu}{\sigma/\sqrt{n}}\right) = \Phi\left(\frac{\mu}{\sigma/\sqrt{n}} - z_\alpha\right)\end{aligned} \tag{3.10}$$

3.3.1 p-Values

In this example, as in the last one, we actually have a family of tests, parametrized by a threshold which we could imagine varying. Given a data realization $\mathbf{x}$, and in particular a sample mean $\overline{x}$, we will reject $\mathcal{H}_0$ if $\overline{x} > z_\alpha\sigma/\sqrt{n}$. This means there will be some values of the false alarm probability α for which we reject $\mathcal{H}_0$, and some for which we do not. One convenient way to report which tests would indicate a positive result (reject the null hypothesis) is to quote the α of the most stringent test for which $\mathcal{H}_0$ would be rejected. Put another way, we ask, given a measurement (in this case $\overline{x}$), how likely is it that we would find a measurement at least this extreme, just by accident, if the null hypothesis were true. This is known as the p-value, and in this case it is defined as

$$p = \mathrm{P}(\overline{X} \geq \overline{x}; \mu = 0) = 1 - \Phi\left(\frac{\overline{x}}{\sigma/\sqrt{n}}\right) = \Phi\left(-\frac{\overline{x}}{\sigma/\sqrt{n}}\right) \tag{3.11}$$

A lower p value means that the results were less likely to have occurred by chance in the absence of a real effect (i.e., if the null hypothesis $\mathcal{H}_0$ were true). Typically if $p < 0.05$, the result is considered interesting and worth future study.[3]

Note that the p value is often misinterpreted. It does *not* represent the probability that the null hypothesis is true (we cannot evaluate such a probability in frequentist inference). A p value of 0.01 simply means, for the statistic we decided to measure, if we repeated the test on many systems for which the null hypothesis was true, we'd get a measurement as extreme, or more, as the one we got, one percent of the time.

Thursday, November 9, 2017
Review for Second Prelim Exam

Tuesday, November 14, 2017
Second Prelim Exam

Thursday, November 16, 2017

3.4 Odds ratio and Bayes factor

See Gregory, Section 3.5 and Sivia, Chapter 4

One of the problems about using a frequentist test like a chi-squared test to assess the validity of a model is that you can

[3] However, if we test for many different effects, or test many different data sets, and only report the result with the lowest p value, we can greatly overstate the significance of our results. See `http://xkcd.com/882/`.

always make the fit better by adding more parameters to the model. In the extreme case, if you have as many model parameters as data points, you can make the fit perfect. But clearly a model which is "overtuned" in this way is scientifically unsatisfying.

Bayesian statistics offers a natural way to compare models, which automatically penalizes models that use too many parameters to fine-tune themselves to match a data set. This is known as the odds ratio.

Consider Bayes's theorem in the context of a model $\mathcal{M}$ with parameters $\boldsymbol{\theta}$. Given an observation $\mathbf{x}$, we can construct the posterior pdf for the parameters $\boldsymbol{\theta}$ as follows

$$f(\boldsymbol{\theta}|\mathbf{x},\mathcal{M}) = \frac{f(\mathbf{x}|\boldsymbol{\theta},\mathcal{M})f(\boldsymbol{\theta}|\mathcal{M})}{f(\mathbf{x}|\mathcal{M})} \tag{3.12}$$

which is sometimes abbreviated as

$$(\text{posterior}) = \frac{(\text{likelihood})(\text{prior})}{(\text{evidence})} \tag{3.13}$$

So far we've just treated the denominator as a normalization factor

$$f(\mathbf{x}|\mathcal{M}) = \int d\theta\, f(\mathbf{x}|\boldsymbol{\theta},\mathcal{M})f(\boldsymbol{\theta}|\mathcal{M}) \tag{3.14}$$

but we will now see how it gets the name "evidence". Note that it is the overall probability to get the observed result $\mathbf{x}$ given the model $\mathcal{M}$, marginalizing over the parameters $\boldsymbol{\theta}$.

Now, consider the case where $\mathcal{M}$ is one of a number of possible models, and we'd like to construct a posterior probability $\mathrm{P}(\mathcal{M}|\mathbf{x})$ that $\mathcal{M}$ is the correct model. Well, since we have a way to calculate $f(\mathbf{x}|\mathcal{M})$, we can try using Bayes's theorem:

$$\mathrm{P}(\mathcal{M}|\mathbf{x}) = \frac{f(\mathbf{x}|\mathcal{M})\,\mathrm{P}(\mathcal{M})}{f(\mathbf{x})} \tag{3.15}$$

The right-hand side has a couple of things that are harder to get a handle on: the prior probability $\mathrm{P}(\mathcal{M})$ of $\mathcal{M}$ being the correct model, and the overall pdf $f(\mathbf{x})$ which requires somehow marginalizing over all possible models. The usual way around this is to consider two competing models $\mathcal{M}_1$ and $\mathcal{M}_2$, and to calculate the ratio of their posteriors, known as the odds ratio

$$\begin{aligned}\mathcal{O}_{12} = \frac{\mathrm{P}(\mathcal{M}_1|\mathbf{x})}{\mathrm{P}(\mathcal{M}_2|\mathbf{x})} &= \frac{f(\mathbf{x}|\mathcal{M}_1)\,\mathrm{P}(\mathcal{M}_1)}{f(\mathbf{x}|\mathcal{M}_2)\,\mathrm{P}(\mathcal{M}_2)} = \left(\frac{f(\mathbf{x}|\mathcal{M}_1)}{f(\mathbf{x}|\mathcal{M}_2)}\right)\left(\frac{\mathrm{P}(\mathcal{M}_1)}{\mathrm{P}(\mathcal{M}_2)}\right) \\ &= \left(\frac{\mathrm{P}(\mathcal{M}_1)}{\mathrm{P}(\mathcal{M}_2)}\right)\mathcal{B}_{12}\end{aligned} \tag{3.16}$$

So the factor of $f(\mathbf{x})$ has cancelled out, and the odds ratio $\mathcal{O}_{12}$ is the ratio of prior probabilities for each model times something known as the *Bayes factor*

$$\mathcal{B}_{12} = \frac{f(\mathbf{x}|\mathcal{M}_1)}{f(\mathbf{x}|\mathcal{M}_2)} \tag{3.17}$$

which is the ratio of the "evidence" in each of the models. It represents how our relative confidence in the two probabilities has changed with the measurement $\mathbf{x}$. If each model has some parameters, the Bayes factor can be written as

$$\mathcal{B}_{12} = \frac{\int d\theta_1\, f(\mathbf{x}|\boldsymbol{\theta}_1,\mathcal{M}_1)\, f(\boldsymbol{\theta}_1|\mathcal{M}_1)}{\int d\theta_2\, f(\mathbf{x}|\boldsymbol{\theta}_2,\mathcal{M}_2)\, f(\boldsymbol{\theta}_2|\mathcal{M}_2)} \tag{3.18}$$

To see how the Bayes factor penalizes modes for over-tuning, consider a simple case where there are two models: $\mathcal{M}_0$, which has no parameters and $\mathcal{M}_1$, which has a parameter θ. If we measure data $\mathbf{x}$, the Bayes factor comparing the two models is

$$\mathcal{B}_{10} = \frac{\int_{-\infty}^{\infty} d\theta\, f(\mathbf{x}|\theta,\mathcal{M}_1)\, f(\theta|\mathcal{M}_1)}{f(\mathbf{x}|\mathcal{M}_0)} \tag{3.19}$$

To get a handle on what the marginalization of the parameter θ does, as compared with the maximization done by the frequentist method, let's make some simplifying assumptions. First let's assume the likelihood $f(\mathbf{x}|\theta,\mathcal{M}_1)$, seen as a function of θ, can be approximated as a Gaussian about the maximum likelihood value $\widehat{\theta}$:

$$f(\mathbf{x}|\theta,\mathcal{M}_1) \approx f(\mathbf{x}|\widehat{\theta},\mathcal{M}_1)\, e^{-(\theta-\widehat{\theta})/2\sigma_\theta^2} \tag{3.20}$$

We'll also assume that this is sharply peaked compared to the prior $f(\theta|\mathcal{M}_1)$ and therefore we can replace θ in the argument of the prior with $\widehat{\theta}$, and

$$\begin{aligned}\int_{-\infty}^{\infty} d\theta\, f(\mathbf{x}|\theta,\mathcal{M}_1)\, f(\theta|\mathcal{M}_1) &\approx f(\mathbf{x}|\widehat{\theta},\mathcal{M}_1)\, f(\widehat{\theta}|\mathcal{M}_1) \int_{-\infty}^{\infty} d\theta\, e^{-(\theta-\widehat{\theta})/2\sigma_\theta^2} \\ &= f(\mathbf{x}|\widehat{\theta},\mathcal{M}_1)\, f(\widehat{\theta}|\mathcal{M}_1)\, \sigma_\theta\sqrt{2\pi}\end{aligned} \tag{3.21}$$

We can then approximate the Bayes factor as

$$\mathcal{B}_{10} = \frac{f(\mathbf{x}|\widehat{\theta},\mathcal{M}_1)}{f(\mathbf{x}|\mathcal{M}_0)} \frac{\sigma_\theta\sqrt{2\pi}}{[f(\widehat{\theta}|\mathcal{M}_1)]^{-1}} \tag{3.22}$$

The first factor is the ratio of the likelihoods between the best-fit version of model $\mathcal{M}_1$ and the parameter-free model $\mathcal{M}_0$. That's basically the end of the story in frequentist model comparison, and we can see that if $\mathcal{M}_0$ is included as a special case of $\mathcal{M}_1$, this ratio will always be greater or equal to one, i.e., the tunable model will always be able to find a higher likelihood than the model without that tunable parameter. But in Bayesian model comparison, there is also the second factor:

$$\frac{\sigma_\theta\sqrt{2\pi}}{[f(\widehat{\theta}|\mathcal{M}_1)]^{-1}} \qquad \text{"Occam factor"} \tag{3.23}$$

This is called the *Occam factor* because it implements Occam's razor, the principle that, all else being equal, simpler explanations will be favored over more complicated ones. Because the prior $f(\theta|\mathcal{M}_1)$ is normalized, $[f(\widehat{\theta}|\mathcal{M}_1)]^{-1}$ is a measure of the width of the prior, i.e., how much parameter space the tunable model has available to it. In particular, if the prior is uniform over some range:

$$f(\theta|\mathcal{M}_1) = \begin{cases} \frac{1}{\theta_{\text{max}}-\theta_{\text{min}}} & \theta_{\text{min}} < \theta < \theta_{\text{max}} \\ 0 & \text{otherwise} \end{cases} \tag{3.24}$$

then the Occam factor becomes

$$\frac{\sigma_\theta\sqrt{2\pi}}{\theta_{\text{max}} - \theta_{\text{min}}} \tag{3.25}$$

because we assumed the likelihood function was narrowly peaked compared to the prior, the Occam factor is always less than one, and the tunable model must have a large enough increase in likelihood over the simpler model in order to overcome this.

3.4.1 Caveats About the Bayes Factor

A number of statisticians, even Bayesian ones, are skeptical about the Bayes factor; e.g., in Chapter Six of Gelman et al *Bayesian Data Analysis*, it takes the authors a while to get around to even talk about the Bayes factor, and by the time they does they mostly has negative things to say about it. There are two major shortcomings that come to mind: First, the Bayes factor only compares the evidences for two models, rather than considering whether either of them is really appropriate in light of the data. Second: if one of the models being compared has one or more continuous parameters, the Bayes factor can depend sensitively on the prior range you assign to the parameter(s), and as a corollary is typically undefined if you try to use a non-informative prior.

The first is in some sense a feature rather than a bug. Bayesian analysis is not designed to ask, in the abstract, how likely the data are given the model; the data have been observed, and we want to use them to evaluate the model. But this is only meaningful in the context of other models which could have produced the same data. Even classical methods which claim to check if data are consistent with a model have to make a choice of test statistic into which to combine the data in order to do quantitative hypothesis tests. (There is also role for such tests in Bayesian analysis.) Still, clues that something is not right with the model can cause us to examine our prior knowledge more carefully and look for the alternative models which were considered unlikely enough to neglect and see which of them might be promoted by the data. So we should definitely not limit ourselves to a Bayes factor between assumed models, which would amount to wearing blinders.

The problem with prior ranges is a serious technical limitation. We saw last time that with a Gaussian likelihood of width $H^{-1/2}$ and maximum likelihood point $\widehat{\theta}$, the Bayes factor between a model M_1 with a tunable parameter θ given a uniform prior from θ_{min} to θ_{max} and a model M_0 with no parameter was (assuming $\theta_{\text{max}} - \theta_{\text{min}} \gg H^{-1/2}$ and $\theta_{\text{min}} < \widehat{\theta} < \theta_{\text{max}}$)

$$\mathcal{B}_{10} \approx \frac{p(\mathbf{y}|\widehat{\theta}, M_1, I)}{p(\mathbf{y}|M_0, I)} \frac{\sqrt{2\pi/H}}{\theta_{\text{max}} - \theta_{\text{min}}} \tag{3.26}$$

The second ratio is the “Occam factor” penalizing M_1 for having a tunable parameter. But we see that the prior range for that parameter is part of the Bayes factor, and if we tried to go to the limit of a non-informative prior by taking $\theta_{\text{min}} \to -\infty$ and $\theta_{\text{max}} \to \infty$, the Occam factor, and therefore the Bayes factor, would go to zero. This is indeed a serious problem, and indicates that we should be careful about assigning too much meaning to a Bayes factor of say 10 or so.

There are a couple of saving graces that can come into play, however. First, we’re assuming the likelihood function is a Gaussian, and in general probability distributions tend to fall off exponentially once you get far from their peaks. One reasonable pair of evidence functions would look like (keeping in mind that $\widehat{\theta}$ is a function of the data $\mathbf{y}$)

$$p(\mathbf{y}|M_0, I) = \sqrt{\frac{H}{2\pi}} \exp\left(-\frac{H}{2}\widehat{\theta}^2\right) \tag{3.27a}$$

$$p(\mathbf{y}|\theta, M_1, I) = \sqrt{\frac{H}{2\pi}} \exp\left(-\frac{H}{2}(\widehat{\theta} - \theta)^2\right) \tag{3.27b}$$

Then the Bayes factor will be

$$\mathcal{B}_{10} \approx e^{H\widehat{\theta}^2/2} \times \frac{\sqrt{2\pi/H}}{\theta_{\text{max}} - \theta_{\text{min}}} \tag{3.28}$$

If $H\widehat{\theta}^2$ is large, it may not matter much what the prior range for θ was. One often quotes Bayes factors on a log scale as well, and the log Bayes factor will be

$$\ln \mathcal{B}_{10} \approx \frac{H\widehat{\theta}^2}{2} \frac{\sqrt{2\pi/H}}{\theta_{\text{max}} - \theta_{\text{min}}} \tag{3.29}$$

We may not know the precise range of reasonable parameter values for a model, but we will usually know it to a couple of orders of magnitude. If, for example, $\left|\widehat{\theta}\right|$ is $8/\sqrt{H}$, the part of the log Bayes factor coming from the likelihood ratio is 32, which means the increase in relative plausibility for M_1, not considering the Occam factor, is[4] $e^{32} \approx 8 \times 10^{13}$. The Occam factor (which

[4] We can see the awkwardness in interpreting the natural log scale, even though it’s simpler mathmatically. One number to keep in mind is $\ln 10 = 1/(\log_{10} e) \approx 2.303$). Thirty-two e-foldings is $32/2.303 \approx 13.9$ orders of magnitude or 139 dB.

is more or less the ratio of the widths of the likelihood and the prior) is almost certainly nowhere near 10^{-13}, and we can say this with confidence even if we don't know the reasonable prior range to better than one or two orders of magnitude.

The other reason why an undefined scale for the Bayes factor may not be a big deal is that we don't always need to look at the numerical value of the Bayes factor itself. We can also use it as a statistic in decision theory, for example preferring $\mathcal{H}_1$ if $\mathcal{B}_{12} > c$ for some threshold k, and $\mathcal{H}_2$ if $\mathcal{B}_{12} < k$. (It is the Bayesian analogue of the likelihood ratio statistic specified by the Neyman-Pearson lemma.) But typically we'll choose k to obtain some specified value of the efficiency $\mathrm{P}(\mathcal{B}_{12} > c|\mathcal{H}_1, I)$ or the false alarm probability $\mathrm{P}(\mathcal{B}_{12} > c|\mathcal{H}_2, I)$, and in practice the threshold k can be tied to the prior parameter range in a way that makes things like efficiency as a function of false alarm probability remain constant in the limit of a noninformative prior.

3.4.2 The Neyman-Pearson Lemma

There is a theorem, usually known as the Neyman-Pearson lemma, that shows how the most powerful test to of one point hypothesis $\mathcal{H}_0$ against another $\mathcal{H}_1$ can be constructed from the likelihood ratio

$$\Lambda(\mathbf{x}) = \frac{f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_0)}{f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_1)} = \frac{L(\theta_0;\mathbf{x})}{L(\theta_1;\mathbf{x})} \tag{3.30}$$

We define C so that $\mathbf{x} \in C$ if and only if $\Lambda(\mathbf{x}) \le k$ where k is defined by

$$P(\Lambda(\mathbf{X}) \le k|\mathcal{H}_0) = \int\limits_C f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_0)\, d^n x = \alpha \tag{3.31}$$

which ensures that the critical region C is of size α. I.e., we reject $\mathcal{H}_0$ if and only if $\Lambda(\mathbf{x}) \le k$.

The Neyman-Pearson lemma states that the power of this test

$$\gamma = P(\Lambda(\mathbf{X}) \le k|\mathcal{H}_1) = \int\limits_C f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_1)\, d^n x \tag{3.32}$$

is greater than or equal to the power of any other test with the same significance. I.e., if A is some other critical region with size α, so that

$$\int\limits_A f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_0)\, d^n x = \alpha \tag{3.33}$$

the Neyman-Pearson lemma says that

$$\gamma(C) = \int\limits_C f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_1)\, d^n x \ge \int\limits_A f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_1)\, d^n x \tag{3.34}$$

The demonstration of the Neyman-Pearson lemma involves breaking up the regions C and A in terms of their overlap $C \cap A$. Evidentally, we can write

$$C = (C \cap A^c) \cup (C \cap A) \tag{3.35a}$$
$$A = (C^c \cap A) \cup (C \cap A) \tag{3.35b}$$

The contribution to both α and γ from $C \cap A$ cancel out of any comparison between C and A. So the Neyman-Pearson lemma is equivalent to the condition that

$$\begin{aligned}\gamma(C) - \int\limits_{C\cap A} & f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_1)\, d^n x \\ &= \int\limits_{C\cap A^c} f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_1)\, d^n x \ge \int\limits_{C^c\cap A} f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_1)\, d^n x\end{aligned} \tag{3.36}$$

If we can prove that, we've proved the lemma

Now, by definition

$$\frac{f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_0)}{f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_1)} \le k \qquad \text{for } \mathbf{x} \in C \tag{3.37}$$

so

$$\int\limits_{C\cap A^c} f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_1)\, d^n x \ge \frac{1}{k} \int\limits_{C\cap A^c} f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_0)\, d^n x \tag{3.38}$$

while

$$\frac{f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_0)}{f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_1)} \ge k \qquad \text{for } \mathbf{x} \in C^c \tag{3.39}$$

so

$$\int\limits_{C^c\cap A} f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_1)\, d^n x \le \frac{1}{k} \int\limits_{C^c\cap A} f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_0)\, d^n x \tag{3.40}$$

But since the tests defined by C and A both have the same significance α, the integrals of $f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_0)$ over the non-overlapping regions must be the same:

$$\begin{aligned} \alpha - \int\limits_{C\cap A} & f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_0)\, d^n x \\ &= \int\limits_{C\cap A^c} f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_0)\, d^n x = \int\limits_{C^c\cap A} f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_0)\, d^n x \end{aligned} \tag{3.41}$$

so the right-hand sides of (3.40) and (3.38) must be equal, which means

$$\int\limits_{C\cap A^c} f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_1)\, d^n x \ge \int\limits_{C^c\cap A} f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_1)\, d^n x \tag{3.42}$$

which, as we've argued above, means that the power of the likelihood ratio test defined by C is greater than or equal to that defined by A.

3.4.3 Composite Hypothesis Testing with Priors

Consider now a slightly different situation. Suppose that the null hypothesis $\mathcal{H}_0$ is a point hypothesis, but the alternative hypothesis $\mathcal{H}_1$ is a composite hypothesis which allows for a range of values of the model parameter(s) $\boldsymbol{\theta}$, but comes with a prior distribution $f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}|\mathcal{H}_1)$ on those parameters. (We include the possibility of a p-dimensional parameter space, but it may also be that $p=1$.) If we define a test which rejects $\mathcal{H}_0$ when

$$\frac{f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_0)}{f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_1)} = \frac{L(\boldsymbol{\theta}_0;\mathbf{x})}{\int L(\boldsymbol{\theta};\mathbf{x})\, f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}|\mathcal{H}_1)\, d^p\theta} \le k \tag{3.43}$$

the Neyman-Pearson lemma tells us this is the most powerful test of $\mathcal{H}_0$ versus $\mathcal{H}_1$. This is "most powerful" in the sense of maximizing the power function

$$\begin{aligned} \gamma(\mathcal{H}_1) = P(\mathbf{X}\in C|\mathcal{H}_1) &= \int_C f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_1)\, d^n x \\ &= \int_C \int L(\boldsymbol{\theta};\mathbf{x})\, f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}|\mathcal{H}_1)\, d^p\theta\, d^n x = \int \gamma(\boldsymbol{\theta})\, f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}|\mathcal{H}_1)\, d^p\theta \end{aligned} \tag{3.44}$$

A few points to note:

- The test statistic in (3.43) is just the Bayes factor which we've already motivated using Bayes's theorem in the form

$$P(\mathcal{H}_i|\mathbf{x}) = \frac{f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_i)\, P(\mathcal{H}_i)}{f_{\mathbf{X}}(\mathbf{x})} \tag{3.45}$$

  to write

$$\frac{P(\mathcal{H}_0|\mathbf{x})}{P(\mathcal{H}_1|\mathbf{x})} = \frac{f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_0)}{f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_1)}\frac{P(\mathcal{H}_0)}{P(\mathcal{H}_1)} \tag{3.46}$$

- If there is a uniformly most powerful test which covers any $\boldsymbol{\theta}$ in the support of $f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}|\mathcal{H}_1)$, this will also be the most powerful test of $\mathcal{H}_0$ against $\mathcal{H}_1$ for any prior distribution.

- A possible objection is that the frequentist hypothesis testing formalism only applies to the outcomes of repeated experiments, and isn't supposed to know about any prior distribution. Searle http://arxiv.org/abs/0804.1161 applies this to the outcome of a Monte Carlo experiment, where $\boldsymbol{\theta}$ is also randomly generated along with the realization of $\mathbf{X}$, so the relevant joint distribution is $f_{\mathbf{X}\boldsymbol{\Theta}}(\mathbf{x},\boldsymbol{\theta}|\mathcal{H}_1) = f_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta})\, f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}|\mathcal{H}_1$. In that context, the Bayes factor constructed using the same prior as the Monte Carlo simulation gives the most powerful test, when attention is restricted to tests which define C using only $\mathbf{X}$ and not $\boldsymbol{\Theta}$.

Tuesday, November 21, 2017

4 Estimating Rates from Counting Experiments

See Gregory, Chapter 14

A common experiment in Physics and Astronomy involves counting observed events (including, in principle, photons within a spectral channel) and trying to estimate the rate associated with the underlying process. This is often complicated by the presence of *background events* which are not produced by the process in question (as opposed to the *foreground events* we're interested in). Three common scenarios, of increasing complexity are:

1. We observe k events in a time T and want to infer the rate r associated with those events.
2. We know the background rate b and want to infer the foreground (or signal) rate $s = r - b$ from the observation.
3. Both the foreground rate s and the background rate b are unknown, and we make observations both on-source (where the rate will be $s+b$) and off-source (where only background events will be present, and the rate will be b).

In all of these cases, the number of events observed should obey a Poisson with a mean equal to the rate times the observation time,

$$p(k|r,I) = \frac{(rT)^k}{k!}e^{-rT} \tag{4.1}$$

We'll mostly consider Bayesian approaches to these problems, but also keep the frequentist prescriptions in mind.

4.1 Case 1: No background

4.1.1 Frequentist approaches

Frequentist statistics doesn't allow us to define probabilities for the rate r to lie in an interval, but it does allow constructions like the maximum likelihood estimate, which turns out to be

$$\widehat{r} = \frac{k}{T} \tag{4.2}$$

or a confidence interval at confidence level α, defined by

$$P(R_\ell \le r \le R_u) = \alpha \tag{4.3}$$

where $R_\ell = \ell(K)$ and $R_u = u(K)$ are statistics constructed from the random variable K. The measured confidence interval is then $[\ell(k), u(k)]$ For example, if we are simply interested in an upper limit, so that $r_\ell = 0$, we want

$$\alpha = P(r \le u(K)) = P(u^{-1}(r) \le K) = \sum_{j=u^{-1}(r)}^{\infty} \frac{(rT)^j}{j!}e^{-rT} \tag{4.4}$$

This looks like a pretty confusing way to define the function $u^{-1}(r)$, but remember, we're interested in $u(k)$ for the actual measured k, so it means that if we evaluate (4.4) for $r = u(k)$ (which we can do since it's supposed to be true for any r), we get

$$\alpha = \sum_{j=k}^{\infty} \frac{(u(k)T)^j}{j!} e^{-u(k)T} = 1 - \sum_{j=0}^{k-1} \frac{(u(k)T)^j}{j!} e^{-u(k)T} \tag{4.5}$$

which is now an equation which can be solved for any $k > 0$. For example, for $k = 1$ it gives us

$$\alpha = 1 - e^{-u(1)T} \tag{4.6}$$

so

$$u(1) = \frac{\ln \frac{1}{1-\alpha}}{T} \; ; \tag{4.7}$$

for $k = 2$ it says

$$\alpha = 1 - [1 + u(2)T]e^{-u(2)T} \tag{4.8}$$

which is a transcendental equation, but we can solve it numerically for $u(2)T$, given α.

4.1.2 Bayesian Approach

As usual, the Bayesian approach to the problem is more straightforward; if we want to know about r given that we've seen k events, we just construct a posterior using Bayes's theorem:

$$f(r|k, I) = \frac{p(k|r, I) f(r|I)}{p(k|I)} \propto p(k|r, I) f(r|I) \tag{4.9}$$

The main subtlety is choosing the prior distribution $f(r|I)$. An obvious simple choice is a uniform prior

$$f(r|I_0) = \begin{cases} \frac{1}{r_{\max}} & 0 < r < r_{\max} \\ 0 & \text{otherwise} \end{cases} \tag{4.10}$$

where we will find our calculations simplify greatly if $r_{\max}$ is large enough that $k \ll r_{\max} T$. There are some conceptual problems with a uniform prior, though. For example, if we replaced the rate parameter in question with a scale parameter $\beta = \frac{1}{r}$ we would find that the prior on β is no longer uniform, but instead $f(\beta|I_0) \propto \beta^{-2}$.

An alternative is to use the prior

$$f(r|I_1) = \begin{cases} \frac{1}{\ln(r_{\max}/r_{\min})} \frac{1}{r} & r_{\min} < r < r_{\max} \\ 0 & \text{otherwise} \end{cases} \tag{4.11}$$

This is often referred to as a Jeffreys prior[5] and you can show that $f(\beta|I_1) \propto \beta^{-1}$. We can also call this "uniform in log rate" because if you do a change of variables to $\lambda = \ln r$ you'll find $f(\lambda|I_1)$ is uniform over the allowed range.

Physically, the $\frac{1}{r}$ prior is appropriate when the rate is uncertain over many orders of magnitude, so e.g., it's as likely to be between 10^{-3} Hz and 10^{-2} Hz as between 10^{-5} Hz and 10^{-4} Hz. More likely, we have a sense of what the order of magnitude of the rate should be, so a uniform prior, in addition to being simpler, may actually reflect our knowledge better.

So let's move ahead with the assumption that $p(r|I)$ is constant, so Bayes's theorem tells is that

$$f(r|k, I) \propto p(k|r, I) \propto (rT)^k \, e^{-rT} \tag{4.12}$$

We can get the proportionality constant from normalization, so

$$f(r|k, I) = \frac{(rT)^k \, e^{-rT}}{\int_0^{r_{\max}} (r'T)^k \, e^{-r'T} \, dr'} \tag{4.13}$$

[5] This is a slight misnomer, since the Jeffreys prior is defined by a mathematical formula using the likelihood, and for some distributions the uniform prior *is* the Jeffreys prior. To make things more confusing, the Jeffreys prior for the rate parameter in an exponential distribution is proportional to r^{-1} as above, but for a Poisson distribution, it's actually proportional to $r^{-1/2}$.

If $k \ll r_{\max}T$, the denominator becomes

$$\int_0^{r_{\max}} (r'T)^k \, e^{-r'T} \, dr' = \frac{1}{T} \int_0^{r_{\max}T} u^k \, e^{-u} \, du \approx \frac{1}{T} \int_0^{\infty} u^k \, e^{-u} \, du = \frac{\Gamma(k+1)}{T} = \frac{k!}{T} \tag{4.14}$$

so

$$f(r|k,I) \approx \begin{cases} \frac{T}{k!} \, (rT)^k \, e^{-rT} & 0 < r < r_{\max} \\ 0 & \text{otherwise} \end{cases} \tag{4.15}$$

Note that this is a Gamma distribution with shape parameter $k+1$ and scale parameter T.

4.2 Case 2: Known Background

Now we have a case where the actual event rate is the unknown quantity of interest, s (the signal or foreground rate) plus a known background rate b, i.e., $r = s + b$. Now, if we knew the exact number of background events, we could subtract that, but as it is, all that's known is the event rate, so there's also randomness in the background, so estimating $s = r - b$ doesn't work out the same as estimating r.

4.2.1 Frequentist approach and issues

We can proceed mostly as before, for example we have a maximum likelihood estimate of

$$\widehat{s} = \widehat{r} - b = \frac{k}{T} - b \tag{4.16}$$

and likewise our confidence interval could be defined using

$$P(R_\ell - b \le s \le R_u - b) = \alpha \tag{4.17}$$

But if we happen to get a small number of events, the results can look weird. For instance, if $k < bT$, the maximum likelihood estimate $\widehat{s}$ would be negative. Similar pathological things can happen with the confidence intervals. This is one of the problems addressed in Feldman and Cousins, "Unified approach to the classical statistical analysis of small signals", *Phys. Rev. D* **57**, 3873 (1998).

4.2.2 Bayesian method

The construction of the posterior proceeds as before, but now we have

$$f(s|k,I) = \frac{([s+b]T)^k \, e^{-[s+b]T}}{\int_0^{s_{\max}} ([s'+b]T)^k \, e^{-[s'+b]T} \, ds'} = \frac{([s+b]T)^k \, e^{-sT}}{\int_0^{s_{\max}} ([s'+b]T)^k \, e^{-s'T} \, ds'} \tag{4.18}$$

where the constant e^{-bT} cancels out. The denominator can be evaluated as

$$\begin{aligned} &\int_0^{s_{\max}} ([s'+b]T)^k \, e^{-s'T} \, dr' \\ &= \frac{1}{T} \sum_{j=0}^{k} \frac{k!}{j!(k-j)!} \int_0^{s_{\max}T} u^{k-j} \, (bT)^j \, e^{-u} \, du \\ &\approx \frac{1}{T} \sum_{j=0}^{k} \frac{k!}{j!(k-j)!} \underbrace{\int_0^{\infty} u^{k-j} \, (bT)^j \, e^{-u} \, du}_{\Gamma(k-j+1)=(k-j)!} = \frac{k!}{T} \sum_{j=0}^{k} \frac{(bT)^j}{j!} \end{aligned} \tag{4.19}$$

Tuesday, November 28, 2017

4.3 Case 3: Unknown/estimated background

We move now to the general case where the foreground and background rates are both unknown. In order to estimate the foreground rate s and disentangle it from the background rate b, we conduct two sets of observations:

- an *OFF-source* observation where the rate of events is b, of duration T_{OFF}, in which k_{OFF} events are observed
- an *ON-source* observation where the rate of events is $s+b$, of duration T_{ON}, in which k_{ON} events are observed

The probability mass functions associated with the on- and off-source distributions are

$$p(k_{\text{ON}}|s,b,I) = \frac{([s+b]T_{\text{ON}})^{k_{\text{ON}}}}{k_{\text{OFF}}!} e^{-[s+b]T_{\text{ON}}} \tag{4.20}$$

and

$$p(k_{\text{OFF}}|b,I) = \frac{(bT_{\text{OFF}})^{k_{\text{OFF}}}}{k_{\text{OFF}}!} e^{-bT_{\text{OFF}}} \tag{4.21}$$

Our goal is to make an inference about the rate r given the on- and off-source observations; in the Bayesian approach this means working out the posterior pdf $f(r|k_{\text{ON}},k_{\text{OFF}},I)$, where the information I includes things like the duration of the observations, but not a specific value for b.

4.3.1 Qualitative

Roughly speaking, the off-source observation will serve as a sort of calibration and allow us to estimate b, albeit with some residual uncertainty. We can then estimate r from the on-source observation, subject to the uncertainty in subtracting the background rate. So the result will look something like

$$b \sim \widehat{b} \pm \delta b \sim \frac{k_{\text{OFF}}}{T_{\text{OFF}}} \pm \frac{\sqrt{k_{\text{OFF}}}}{T_{\text{OFF}}} \tag{4.22}$$

and

$$s \sim \widehat{s} \pm \delta s \sim \frac{k_{\text{ON}}}{T_{\text{ON}}} - \widehat{b} \pm \sqrt{\frac{k_{\text{ON}}}{{T_{\text{ON}}}^2} + (\delta b)^2} \tag{4.23}$$

but this back of the envelope calculation will fail if the numbers of events are small.

4.3.2 Bayesian method

We want to work out the posterior pdf $f(s|k_{\text{ON}},k_{\text{OFF}},I)$ for the foreground rate, given the on- and off-source observations, which we've marginalized over the background rate b. We assume that the priors on the foreground and background rates are uniform, i.e.,

$$f(s|I_0) = \begin{cases} \frac{1}{s_{\max}} & 0 < s < s_{\max} \\ 0 & \text{otherwise} \end{cases} \tag{4.24}$$

and

$$f(b|I_0) = \begin{cases} \frac{1}{b_{\max}} & 0 < b < b_{\max} \\ 0 & \text{otherwise} \end{cases} \tag{4.25}$$

and we'll assume $k_{\text{OFF}} \ll b_{\max}T_{\text{OFF}}$, $k_{\text{ON}} \ll s_{\max}T_{\text{ON}}$ and $k_{\text{ON}} \ll b_{\max}T_{\text{ON}}$ to make the integrals simpler.

There are several equivalent ways to arrive at basically the same expression for the posterior. The first two use the fact that the on- and off-source measurements are independent to work in terms of $p(k_{\text{ON}},k_{\text{OFF}}|s,b,I) = p(k_{\text{ON}}|s,b,I)\,p(k_{\text{OFF}}|b,I)$.

1. Use Bayes's theorem to get the joint posterior

$$\begin{aligned} f(s,b|k_{\text{ON}},k_{\text{OFF}},I) &\propto p(k_{\text{ON}},k_{\text{OFF}}|s,b,I)\,f(s,b|I) \\ &= p(k_{\text{ON}},k_{\text{OFF}}|s,b,I)\,f(b|I)\,f(s|I) \end{aligned} \tag{4.26}$$

   and then marginalize over b to get

$$\begin{aligned} f(s|k_{\text{ON}},k_{\text{OFF}},I) &= \int_0^\infty f(s,b|k_{\text{ON}},k_{\text{OFF}},I) \\ &\propto \int_0^\infty p(k_{\text{ON}}|s,b,I)\,p(k_{\text{OFF}}|b,I)\,f(b|I)\,f(s|I)\,db \end{aligned} \tag{4.27}$$

2. Use Bayes's theorem to write

$$f(s|k_{\text{ON}}, k_{\text{OFF}}, I) \propto p(k_{\text{ON}}, k_{\text{OFF}}|s, I)\, f(s|I) \tag{4.28}$$

and get the marginalized likelihood by writing

$$p(k_{\text{ON}}, k_{\text{OFF}}|s, I) = \int_0^\infty p(k_{\text{ON}}, k_{\text{OFF}}|s, b, I)\, f(b|I)\, db \tag{4.29}$$

3. Consider $I' = k_{\text{OFF}}, I$ to be the state of information after the off-source experiment, and describe the observation in two steps: First, we get a pdf for b based on the off-source experiment

$$f(b|I') = f(b|k_{\text{OFF}}, I) \propto p(k_{\text{OFF}}|b, I)\, f(b|I) \tag{4.30}$$

and then use this posterior as the prior on b in the on-source experiment:

$$\begin{aligned} f(s|k_{\text{ON}}, I') &\propto \int_0^\infty p(k_{\text{ON}}|s, b, I')\, f(b|I')\, db\, f(s|I') \\ &\propto \int_0^\infty p(k_{\text{ON}}|s, b, I)\, p(k_{\text{OFF}}|b, I)\, f(b|I)\, db\, f(s|I) \end{aligned} \tag{4.31}$$

where we use the fact that neither the pmf for the on-source experiment nor the pdf for the signal rate depend on the outcome are directly affected by the results of the off-source experiment, so $p(k_{\text{ON}}|s, b, I') = p(k_{\text{ON}}|s, b, I)$ and $f(s|I') = f(s|I)$.

Of course, it's not surprising that all three approaches give the same expression, since they all just follow the rules of probability. The last approach gives us a head start to constructing the posterior on s, since we know the pdf of the background rate after the off-source experiment will be a Gamma distribution

$$f(b|k_{\text{OFF}}, I) \propto (bT_{\text{OFF}})^{k_{\text{OFF}}} e^{-bT_{\text{OFF}}} \tag{4.32}$$

Note that this distribution has a mean of $\frac{k_{\text{OFF}}+1}{T_{\text{OFF}}}$ and a width of $\frac{\sqrt{k_{\text{OFF}}+1}}{T_{\text{OFF}}}$, so in the limit of a long off-source observation with many events, we get a more and more sharply-peaked distribution in b, which makes the estimation of s tend towards the case of a know background.

Moving on to the construction of the posterior on the foreground rate,

$$\begin{aligned} f(s|k_{\text{ON}}, k_{\text{OFF}}, I) &\propto \int_0^\infty ([s{+}b]T_{\text{ON}})^{k_{\text{ON}}} e^{-[s+b]T_{\text{ON}}} (bT_{\text{OFF}})^{k_{\text{OFF}}} e^{-bT_{\text{OFF}}}\, db \\ &\propto e^{-sT_{\text{ON}}} \int_0^\infty (s+b)^{k_{\text{ON}}} b^{k_{\text{OFF}}} e^{-b(T_{\text{ON}}+T_{\text{OFF}})}\, db \\ &\propto e^{-sT_{\text{ON}}} \sum_{j=0}^{k_{\text{ON}}} \frac{k_{\text{ON}}!}{(k_{\text{ON}}-j)!j!} (s[T_{\text{ON}}+T_{\text{OFF}}])^j \int_0^\infty u^{k_{\text{ON}}+k_{\text{OFF}}-j} e^{-u}\, du \\ &\propto \sum_{j=0}^{k_{\text{ON}}} \frac{(k_{\text{ON}}+k_{\text{OFF}}-j)!}{(k_{\text{ON}}-j)!j!} \left(1+\frac{T_{\text{OFF}}}{T_{\text{ON}}}\right)^j (sT_{\text{ON}})^j\, e^{-sT_{\text{ON}}} \end{aligned} \tag{4.33}$$

Now, it's pretty straightforward to do the integral over s and work out that normalization constant to get

$$f(s|k_{\text{ON}}, k_{\text{OFF}}, I) = \frac{T_{\text{ON}} \sum_{j=0}^{k_{\text{ON}}} \frac{(k_{\text{ON}}+k_{\text{OFF}}-j)!}{(k_{\text{ON}}-j)!j!} \left(1+\frac{T_{\text{OFF}}}{T_{\text{ON}}}\right)^j (sT_{\text{ON}})^j\, e^{-sT_{\text{ON}}}}{\sum_{j'=0}^{k_{\text{ON}}} \frac{(k_{\text{ON}}+k_{\text{OFF}}-j')!}{(k_{\text{OFF}}-j')!} \left(1+\frac{T_{\text{OFF}}}{T_{\text{ON}}}\right)^{j'}} \tag{4.34}$$

although in practice the shape of the pdf is more interesting. You will investigate this on the homework.

Thursday, November 30, 2017

5 Monte Carlo Methods

Monte Carlo, in general, refers to calculations carried out with random or pseudo-random elements. (The name refers to the Monte Carlo Casino in Monaco.) There are a number of different such methods, but we'll focus on two; in each case we'll assume the model includes a sampling distribution $f(\mathbf{x}|\boldsymbol{\theta}, I)$ that describes the joint pdf of the data $\mathbf{X}$ given a parameter vector $\boldsymbol{\theta}$.

1. Monte Carlo simulations to test the validity of a statistical method or estimate the sampling distribution of statistics.
2. Drawing samples from a posterior to simulate higher-dimensional integrals, including the Markov Chain Monte Carlo (MCMC) method for generating a sample.

Note: how to simulate a random variable with a specified pdf $f(\mathbf{x})$:

1. Easy way/cheating: use statistical package, e.g., `scipy.stats.norm(loc=mu,scale=sigma,size=n)`
2. General approach for univariate distribution $f(x)$: given cdf $F(x) = \int_{-\infty}^{x} f(x')\,dx'$, invert to get $x = F^{-1}(P)$ for $0 < P < 1$. Generate uniform random number α and take $x = F^{-1}(\alpha)$. Note that if you have a closed form cdf that you can invert, the distribution is probably already coded up in a statistical package
3. Grid approximation: cover the space of possible $\mathbf{x}$ values with a grid $\{\mathbf{x}_i\}$ of discrete points. (Ideally it should extend over the whole space over which $f(\mathbf{x})$ has support, and be closely enough spaced that $f(\mathbf{x}_i)$ and $f(\mathbf{x}_j)$ are not so different if $\mathbf{x}_i$ and $\mathbf{x}_j$ are adjacent points in the grid. You can then draw from the discrete distribution with pmf $p(\mathbf{x}_i) = f(\mathbf{x}_i)/\sum_j f(\mathbf{x}_j)$.
4. Markov Chain Monte Carlo. We will discuss this on Tuesday.

5.1 Monte Carlo Simulations

The frequentist definition of probability tells us that if we have a repeatable experiment described by random vector $\mathbf{X}$ with pdf $f(\mathbf{x}|\boldsymbol{\theta}, I)$, and we do N repetitions of the experiment, whose results are $\{\mathbf{x}_s|s = 1, \ldots, N\}$, probabilities constructed using $\mathbf{X}$ will correspond to frequencies of the corresponding events constructed from the $\{xvec_s\}$. Specifically, for some region C of the data space,

$$P(\mathbf{X} \in C|\boldsymbol{\theta}, I) \approx \frac{N(\mathbf{x}_s \in C)}{N} \tag{5.1}$$

when N is sufficiently large. So we can use this to either check or estimate things like the distribution of a statistic $\mathbf{Y} = \mathbf{g}(\mathbf{X})$ or the power $\gamma(\boldsymbol{\theta}) = P(\mathbf{X} \in C|\boldsymbol{\theta}, I)$ of a test definied by the critical region C. The latter can be estimated using the fraction of points in the sample $\{\mathbf{x}_s\}$ which are in C. For an example of the former, see the notebook `http://ccrg.rit.edu/~whelan/courses/2017_3fa_ASTP_611/data/notes_inference_montecarlo.ipynb`

5.2 Sampling from a Posterior

One of the ways in which the Bayesian interpretation of probability is more powerful than the frequentist one is that it allows us to assign probabilities to things that are uncertain, without requiring them to be the outcomes of repeatable experiments. But we can also turn this relationship on its head to our benefit. A posterior pdf $f(\boldsymbol{\theta}|\mathbf{x}, I)$ doesn't describe the relative frequencies of a bunch of $\boldsymbol{\theta}$ values, but the mathematics of probability are the same. So if we can use the pdf $f(\boldsymbol{\theta}|\mathbf{x}, I)$ to synthetically generate a sample $\{\boldsymbol{\theta}_s\}$ (where $\mathbf{x}$ is the one actual observed data vector), relative frequencies in the sample will correspond approximately to probabilities in the posterior. We now illustrate this in the notebook `http://ccrg.rit.edu/~whelan/courses/2017_3fa_ASTP_611/data/notes_inference_sampling.ipynb`

Tuesday, December 5, 2017

6 Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) methods execute a random walk through parameter space, and are designed to visit points in parameter space in proportion to their probabilities under the target distribution $q(\boldsymbol{\theta})$ which is proportional to the posterior from which we're trying to sample, $p(\boldsymbol{\theta}|\mathbf{x}, I) \propto q(\boldsymbol{\theta})$. The "Markov" part means that the probability of reaching a point $\boldsymbol{\theta}^t$ at step t of the chain depends only the previous point $\boldsymbol{\theta}^{t-1}$ and not any earlier points in the chain.

6.1 Metropolis Algorithm

To carry out an MCMC we need a rule for moving from one point to another in parameter space. Typically there is a proposal distribution $J(\boldsymbol{\theta}'|\boldsymbol{\theta})$ from which a jump is considered, and then a rule for deciding whether to jump to that new point. The Metropolis algorithm requires that the rule be symmetric, so $J(\boldsymbol{\theta}'|\boldsymbol{\theta}) = J(\boldsymbol{\theta}|\boldsymbol{\theta}')$ but doesn't otherwise restrict it.[6] The rule is then as follows: draw a point $\boldsymbol{\theta}^*$ from $J(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{t-1})$ and calculate the ratio

$$r = \frac{f(\boldsymbol{\theta}^*|\mathbf{x}, I)}{f(\boldsymbol{\theta}^{t-1}|\mathbf{x}, I)} = \frac{q(\boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^{t-1})} \; ; \tag{6.1}$$

If $r \geq 1$, make the jump and let $\boldsymbol{\theta}^t = \boldsymbol{\theta}^*$. If $r < 1$, generate a Uniform$[0, 1]$ random number u; if $u \leq r$, make the jump and $\boldsymbol{\theta}^t = \boldsymbol{\theta}^*$. If $u > r$, don't make the jump and let $\boldsymbol{\theta}^t = \boldsymbol{\theta}^{t-1}$. I.e., make the jump with probability r.

We'll see in detail why it works on Thursday, but as a matter of formalism consider how the rules translate into statements about the probability distributions for the proposed jump position $\boldsymbol{\theta}^*$ and the next value $\boldsymbol{\theta}^t$ in terms of the current value $\boldsymbol{\theta}^{t-1}$. The proposal rule just tells us that

$$f(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{t-1}, \mathbf{x}, I) = J(\boldsymbol{\theta}^*|\boldsymbol{\theta}) \tag{6.2}$$

while the acceptance rule tells us that

$$f(\boldsymbol{\theta}^t|\boldsymbol{\theta}^*, \boldsymbol{\theta}^{t-1}, \mathbf{x}, I) = \begin{cases} \delta(\boldsymbol{\theta}^t - \boldsymbol{\theta}^*) & \text{if } r \geq 1 \\ r\delta(\boldsymbol{\theta}^t - \boldsymbol{\theta}^*) + (1-r)\delta(\boldsymbol{\theta}^t - \boldsymbol{\theta}^{t-1}) & \text{if } r < 1 \end{cases} \tag{6.3}$$

where $\delta(\boldsymbol{\theta} - \boldsymbol{\theta}')$ is the Dirac delta function defined so that

$$\int \delta(\boldsymbol{\theta} - \boldsymbol{\theta}')\, f(\boldsymbol{\theta})\, d\boldsymbol{\theta} = f(\boldsymbol{\theta}') \tag{6.4}$$

and r is the ratio defined in (6.1). Note that if $r < 1$, $f(\boldsymbol{\theta}^t|\boldsymbol{\theta}^*, \boldsymbol{\theta}^{t-1}, \mathbf{x}, I)$ is a *mixture distribution*, which is just a linear combination of other probability distributions (in this case degenerate ones). It's basically just a manifestation of the sum rule. If I says that $\boldsymbol{\theta}$ can be drawn from distributions D_1 or D_2, the probability distribution is

$$f(\boldsymbol{\theta}|I) = f(\boldsymbol{\theta}|D_1, I)\ \mathrm{P}(D_1|I) + f(\boldsymbol{\theta}|D_2, I)\ \mathrm{P}(D_2|I) \tag{6.5}$$

where $f(\boldsymbol{\theta}|D_1, I)$ and $f(\boldsymbol{\theta}|D_2, I)$ are separately normalized probability distributions, and $\mathrm{P}(D_1|I) + \mathrm{P}(D_2|I) = 1$.

[6] $J(\boldsymbol{\theta}'|\boldsymbol{\theta})$ should be a probability distribution which we can easily draw from for any $\boldsymbol{\theta}$ with $f(\boldsymbol{\theta}|\mathbf{x}, I) > 0$.

Returning to the distribution (6.3) we can actually combine the two cases by writing

$$f(\boldsymbol{\theta}^t|\boldsymbol{\theta}^*,\boldsymbol{\theta}^{t-1},\mathbf{x},I) = \min(1,r)\delta(\boldsymbol{\theta}^t-\boldsymbol{\theta}^*)+[1-\min(1,r)]\delta(\boldsymbol{\theta}^t-\boldsymbol{\theta}^{t-1}) \tag{6.6}$$

The product rule then gives the joint distribution

$$\begin{aligned} f(\boldsymbol{\theta}^t,\boldsymbol{\theta}^*|\boldsymbol{\theta}^{t-1},\mathbf{x},I) = J(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{t-1})\Big(&\min(1,r)\delta(\boldsymbol{\theta}^t-\boldsymbol{\theta}^*) \\ &+ [1-\min(1,r)]\delta(\boldsymbol{\theta}^t-\boldsymbol{\theta}^{t-1})\Big) \end{aligned} \tag{6.7}$$

We will show that the target distribution is the stable endpoint of the MCMC by showing that if the marginal sampling distribution for one step is the target distribution, than the marginal sampling distribution for the next step is as well:

$$f(\boldsymbol{\theta}^{t-1}|\text{MCMC}) = f(\boldsymbol{\theta}^{t-1}|\mathbf{x},I) \quad \text{implies} \quad f(\boldsymbol{\theta}^{t}|\text{MCMC}) = f(\boldsymbol{\theta}^{t}|\mathbf{x},I) \tag{6.8}$$

6.2 Why It Works

Now we turn to a demonstration of why the Metropolis algorithm produces, over the long term, samples from the target distribution. As a result of some mathematical theory that's beyond the scope of this course, any reasonable MCMC will settle down into a unique equilibrium, so all we need is to demonstrate that the target distribution is an equilibrium state. What that means is that if we (hypothetically) make a draw from the target distribution, and then take one MCMC step from that point and consider where we ended up, the probability distribution for the new point is also the target distribution. In the notation we've been using, this means that if we assume $f(\boldsymbol{\theta}^{t-1}|\text{MCMC}) = f(\boldsymbol{\theta}^{t-1}|\mathbf{x},I)$ we can show that $f(\boldsymbol{\theta}^{t}|\text{MCMC}) = f(\boldsymbol{\theta}^{t}|\mathbf{x},I)$. We'll do this by constructing the joint distribution $f(\boldsymbol{\theta}^t,\boldsymbol{\theta}^{t-1}|\text{MCMC})$ for the two points. Note that we're assuming

$$\int f(\boldsymbol{\theta}^t,\boldsymbol{\theta}^{t-1}|\text{MCMC}) = f(\boldsymbol{\theta}^{t-1}|\text{MCMC}) = f(\boldsymbol{\theta}^{t-1}|\mathbf{x},I) \tag{6.9}$$

so if the functional form of $f(\boldsymbol{\theta}^t,\boldsymbol{\theta}^{t-1}|\text{MCMC})$ is symmetric under interchange of the two arguments $\boldsymbol{\theta}^t$ and $\boldsymbol{\theta}^{t-1}$, it will be the case that the two marginal distributions $f(\boldsymbol{\theta}^t|\text{MCMC})$ and $f(\boldsymbol{\theta}^{t-1}|\text{MCMC})$ have the same form, and thus we'll have proved that $f(\boldsymbol{\theta}^t|\text{MCMC}) = f(\boldsymbol{\theta}^t|\mathbf{x},I)$.

To construct the joint distribution $f(\boldsymbol{\theta}^t,\boldsymbol{\theta}^{t-1}|\text{MCMC})$ we start with the joint distribution

$$\begin{aligned} f(\boldsymbol{\theta}^t,\boldsymbol{\theta}^*,\boldsymbol{\theta}^{t-1}|\text{MCMC}) &= f(\boldsymbol{\theta}^{t-1}|\mathbf{x},I)J(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{t-1})\Big(\min(1,r)\delta(\boldsymbol{\theta}^t-\boldsymbol{\theta}^*) \\ &\qquad + \max(1-r,0)\delta(\boldsymbol{\theta}^t-\boldsymbol{\theta}^{t-1})\Big) \\ &= \delta(\boldsymbol{\theta}^t-\boldsymbol{\theta}^*)J(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{t-1})\min(f(\boldsymbol{\theta}^{t-1}|\mathbf{x},I),f(\boldsymbol{\theta}^*|\mathbf{x},I)) \\ &\quad + \delta(\boldsymbol{\theta}^t-\boldsymbol{\theta}^{t-1})J(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{t-1})\max(f(\boldsymbol{\theta}^{t-1}|\mathbf{x},I)-f(\boldsymbol{\theta}^*|\mathbf{x},I),0) \end{aligned} \tag{6.10}$$

Next we marginalize over $\boldsymbol{\theta}^*$; in the first term, the delta function just sets $\boldsymbol{\theta}^*$ to $\boldsymbol{\theta}^t$. The second term is more complicated, but the result of the integral (factoring out the $\boldsymbol{\theta}^*$-independent delta function) is

$$\begin{aligned} &\int J(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{t-1})\max(f(\boldsymbol{\theta}^{t-1}|\mathbf{x},I)-f(\boldsymbol{\theta}^*|\mathbf{x},I),0)\,d\boldsymbol{\theta}^* \\ &\quad = \int\limits_{f(\boldsymbol{\theta}^*|\mathbf{x},I)<f(\boldsymbol{\theta}^{t-1}|\mathbf{x},I)} J(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{t-1})\left[f(\boldsymbol{\theta}^{t-1}|\mathbf{x},I)-f(\boldsymbol{\theta}^*|\mathbf{x},I)\right]\,d\boldsymbol{\theta}^* \\ &\quad = f(\boldsymbol{\theta}^{t-1},\text{reject}|\mathbf{x},I) \end{aligned} \tag{6.11}$$

I.e., it's the probability of drawing $\boldsymbol{\theta}^{t-1}$ from the target distribution and then rejecting the next jump. So the result of the marginalization is

$$f(\boldsymbol{\theta}^t, \boldsymbol{\theta}^{t-1}|\text{MCMC}) = J(\boldsymbol{\theta}^t|\boldsymbol{\theta}^{t-1}) \min(f(\boldsymbol{\theta}^{t-1}|\mathbf{x}, I), f(\boldsymbol{\theta}^t|\mathbf{x}, I)) \\ + \delta(\boldsymbol{\theta}^t - \boldsymbol{\theta}^{t-1}) f(\boldsymbol{\theta}^{t-1}, \text{reject}|\mathbf{x}, I) \quad (6.12)$$

The first term is symmetric under interchange $\boldsymbol{\theta}^t \longleftrightarrow \boldsymbol{\theta}^{t-1}$ as long as the proposal distribution $J(\boldsymbol{\theta}^t|\boldsymbol{\theta}^{t-1})$ is. The second term is also symmetric, because the Dirac delta function is even, and is only non-zero if $\boldsymbol{\theta}^t = \boldsymbol{\theta}^{t-1}$. Thus the joint distribution is symmetric and both marginal distributions are the same, i.e., the target distribution.

6.3 Metropolis-Hastings Algorithm

The Metropolis algorithm can be extended to situations where the proposal distribution is not symmetric, e.g., distributions which avoid boundaries of parameter space. The modification is to the acceptance rule, which now uses the value of the ratio

$$r' = \frac{f(\boldsymbol{\theta}^*|\mathbf{x}, I)\, J(\boldsymbol{\theta}^{t-1}|\boldsymbol{\theta}^*)}{f(\boldsymbol{\theta}^{t-1}|\mathbf{x}, I)\, J(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{t-1})} = \frac{q(\boldsymbol{\theta}^*)\, J(\boldsymbol{\theta}^{t-1}|\boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^{t-1})\, J(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{t-1})} \; ; \quad (6.13)$$

the rule is as before, i.e., accept the jump with probability $\min(1, r')$.

6.4 Examples and Properties

We can investigate the behavior of a simple MCMC using the notebook `http://ccrg.rit.edu/~whelan/courses/2017_3fa_ASTP_611/data/notes_inference_mcmc.ipynb` (see also the old notebook `http://ccrg.rit.edu/~whelan/courses/2014_1sp_ASTP_611/data/notes_inference_mcmctrinomial.ipynb` which draws from a discrete distribution.)

6.4.1 Choice of Proposal Distribution

One important piece of the puzzle is the proposal distribution $J(\boldsymbol{\theta}|\boldsymbol{\theta}')$. The typical jump size should be on the same scale as the features of the distribution. If the proposed jumps are too big, most of them will be to places where the target distribution is small, and will be likely to be rejected, so the chain will move slowly because it's sitting at the same point without jumping for many steps. Conversely, if the proposed jumps are too small, the chain will move very slowly and therefore take more steps to explore the full support of the posterior. As an extreme example of this, the proposal distribution $J(\boldsymbol{\theta}|\boldsymbol{\theta}') = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}')$ will of course satisfy the detailed balance requirement (since $\boldsymbol{\theta}^t = \boldsymbol{\theta}^{t-1}$, if $\boldsymbol{\theta}^{t-1}$ were drawn from the target distribution, $\boldsymbol{\theta}^t$ would be too), but the chain will clearly never expore the parameter space.

6.4.2 Diagnostics

- Plot the chains and make sure they revisit the same points in parameter space
- Plot the first and second halves of the same chain and make sure they look similar. Note that the very first part of the chain will be influenced by the starting point. It's standard to discard this "burn-in" (10-50% of the steps typically) before doing anything with the chain.
- Start chains in different parts of parameter space, discard the burn-in of each, and check that the two chains look similar otherwise.

6.4.3 Further Reading

There are a lot of MCMC examples in the notes I wrote up for the Bayesian course: `http://ccrg.rit.edu/~whelan/courses/2017_1sp_STAT_489/notes_comp.pdf`