STAT 513 fa 2020 Lec 09

Bayesian Inference

Karl B. Gregory

Parameters as random variables

- Until now we have focused on estimating and making inferences concerning the unknown value of a parameter θ lying in some parameter space Θ. We assumed that θ had a fixed, unknown value.
- In the Bayesian paradigm we regard the parameter θ as a random variable taking values in Θ.
- In the fixed-θ paradigm, within which we have so far been working, only the data were random, and we made careful studies about the random behavior of various sample statistics and about the performance of tests of hypotheses. Our inquiries were of the following sort:
  - With what probability will our estimator fall within some distance of θ?
  - With what probability will the data lead us to reject the null if θ takes a certain value?

  If we understand probability as how often an outcome of a statistical experiment will occur if the experiment is repeated a number of times approaching infinity, then we can rephrase the above questions as

  - With what frequency will our estimator fall within some distance of θ?
  - With what frequency will the data lead us to reject the null if θ takes a certain value?

  In light of the questions which arise when assuming that θ is fixed and only the data are random, the fixed-θ setting has come to be called the *frequentist paradigm*.
- If we regard θ as a random variable, as we do the *Bayesian paradigm*, we can make probability statements about θ, whereas in the frequentist paradigm we could only make probability statements about the data. For example, the Bayesian approach allows us to make a statement such as, "given the observed data, θ lies in the interval $(2,3)$ with probability 0.95." This statement is nonsense to the frequentist, who sees θ as a fixed constant.

Data distribution, prior, and posterior

- The Bayesian approach begins with a hierarchical model in which the data distribution is a distribution conditional on the the value of some parameters, which in turn have their own marginal distribution. In general, we have the following:

Let $X_1, \dots, X_n$ be a collection of random variables which are to be observed as data and let θ be a random variable taking values in $\Theta \subset \mathbb{R}$. Then we assume the hierarchical model

$$\begin{aligned} X_1, \dots, X_n | \theta &\sim f(x_1, \dots, x_n | \theta) \\ \theta &\sim \pi(\theta), \end{aligned}$$

where $f(\cdot|\theta)$ is the joint pmf or pdf of the random variables $X_1, \dots, X_n$, conditional on the value of θ, and $\pi(\cdot)$ is the marginal pmf or pdf of the parameter θ.

- We will call the distribution with pmf/pdf $f(x_1, \dots, x_n|\theta)$ the *data distribution*.
- We will call the distribution with pmf/pdf $\pi(\theta)$ the *prior distribution* of θ.
- The prior distribution may be chosen to reflect beliefs about which values θ is likely to take—beliefs which are held *before* any data are observed. The idea behind Bayesian inference is to use the observed data to update previously held ("prior") beliefs about the parameter θ. We do this by finding what is called the posterior distribution of θ.
- The *posterior distribution* of θ is the distribution of θ conditional on the data $X_1, \dots, X_n$. We have

$$\theta | X_1, \dots, X_n \sim \pi(\theta | X_1, \dots, X_n) = \begin{cases} \dfrac{f(X_1, \dots, X_n|\theta)\pi(\theta)}{\int_\Theta f(X_1, \dots, X_n|\tilde{\theta})\pi(\tilde{\theta})d\tilde{\theta}} & \text{if } \theta \text{ is continuous} \\ \dfrac{f(X_1, \dots, X_n|\theta)\pi(\theta)}{\sum_{\tilde{\theta}\in\Theta} f(X_1, \dots, X_n|\tilde{\theta})\pi(\tilde{\theta})} & \text{if } \theta \text{ is discrete,} \end{cases}$$

  so that $\pi(\theta|X_1, \dots, X_n)$ is the conditional pmf/pdf of θ given $X_1, \dots, X_n$.

- To see how we obtained the above expressions for $\pi(\theta|X_1, \dots, X_n)$, recall that for any two continuous random variables U and V such that

$$\begin{aligned} U|V &\sim f(u|v) \\ V &\sim f(v), \end{aligned}$$

  we have

$$V|U \sim f(v|u) = \frac{f(u,v)}{f(u)} = \frac{f(u,v)}{\int_{\mathbb{R}} f(u,\tilde{v})d\tilde{v}} = \frac{f(u|v)f(v)}{\int_{\mathbb{R}} f(u|\tilde{v})f(\tilde{v})d\tilde{v}},$$

  where the integrals change to sums over the support of V if V is discrete.

- The Bayesian approach is to use the posterior distribution of θ to estimate and make inferences about θ.
- **Bayesian estimation with the posterior mean:** A usual way to estimate the value of θ in the Bayesian paradigm is with the mean of its posterior distribution. That is, we use the estimator

$$\hat{\theta}_{\text{Bayes}} = \mathbb{E}[\theta|X_1, \dots, X_n] = \begin{cases} \int_\Theta \theta\pi(\theta|X_1, \dots, X_n)d\theta & \text{if } \theta \text{ is continuous} \\ \sum_{\theta\in\Theta} \theta\pi(\theta|X_1, \dots, X_n) & \text{if } \theta \text{ is discrete.} \end{cases}$$

- **Result:** The above estimator is given by the minimization

$$\hat{\theta}_{\text{Bayes}} = \underset{a}{\text{argmin}} \quad \mathbb{E}[(\theta - a)^2 | X_1, \ldots, X_n].$$

  Proof: We find the value of a which satisfies

$$\frac{\partial}{\partial a} \mathbb{E}[(\theta - a)^2 | X_1, \ldots, X_n] = 0.$$

  We have

$$\frac{\partial}{\partial a} \mathbb{E}[(\theta - a)^2 | X_1, \ldots, X_n] = \mathbb{E}[-2(\theta - a) | X_1, \ldots, X_n] = -2\mathbb{E}[\theta | X_1, \ldots, X_n] + 2a.$$

  Setting this equal to zero gives the result.

- The posterior mean $\hat{\theta}_{\text{Bayes}}$ is the value from which θ is expected, conditional on the data, to have the smallest squared distance. This makes it a natural choice of estimator for θ, although there are other ways to define estimators in the Bayesian world (one can use the median or the mode of the posterior distribution, for example).

- **Exercise:** Suppose we have

$$\begin{aligned} Y | p &\sim \text{Binomial}(n, p) \\ p &\sim \text{Beta}(\alpha, \beta), \end{aligned}$$

  so that the pmf $f(y|p)$ of $Y|p$ is given by

$$f(y|p) = \binom{n}{y} p^y (1-p)^{n-y}, \qquad \text{for } Y \in \{0, 1, \ldots, n\}$$

  and the prior pdf $\pi(p)$ of p is given by

$$\pi(p) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha - 1} (1-p)^{\beta - 1}, \qquad \text{for } p \in (0, 1).$$

  i) Find the posterior distribution of $p|Y$.

  ii) Find an expression for $\hat{p}_{\text{Bayes}} = \mathbb{E}[p | X_1, \ldots, X_n]$.

  iii) Suppose $n = 10$ and $Y = 3$ is observed. Find the posterior mean of p under following choices of the prior parameters:

  (a) $\alpha = 1$, $\beta = 1$

  (b) $\alpha = 4$, $\beta = 4$

  (c) $\alpha = 4$, $\beta = 10$

  Answers:

i) The posterior density of p is given by

$$\pi(p|Y) = \frac{\binom{n}{Y}p^Y(1-p)^{n-Y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta-1}}{\int_0^1\binom{n}{Y}\tilde{p}^Y(1-\tilde{p})^{n-Y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\tilde{p}^{\alpha-1}(1-\tilde{p})^{\beta-1}d\tilde{p}}$$

for $p \in (0,1)$. The denominator is given by

$$\begin{aligned}
&\int_0^1\binom{n}{Y}\tilde{p}^Y(1-\tilde{p})^{n-Y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\tilde{p}^{\alpha-1}(1-\tilde{p})^{\beta-1}d\tilde{p}\\
&\quad= \binom{n}{y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1\tilde{p}^{Y+\alpha-1}(1-\tilde{p})^{n-Y+\beta-1}d\tilde{p}\\
&\quad= \binom{n}{Y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(Y+\alpha)\Gamma(n-Y+\beta)}{\Gamma(n+\alpha+\beta)}\int_0^1\frac{\Gamma(n+\alpha+\beta)}{\Gamma(Y+\alpha)\Gamma(n-Y+\beta)}\tilde{p}^{Y+\alpha-1}(1-\tilde{p})^{n-Y+\beta-1}d\tilde{p}\\
&\quad= \binom{n}{y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(Y+\alpha)\Gamma(n-Y+\beta)}{\Gamma(n+\alpha+\beta)},
\end{aligned}$$

for $Y \in \{0,1,\dots,n\}$. Plugging this into the expression for $\pi(p|Y)$, we have

$$\pi(p|Y) = \frac{\Gamma(n+\alpha+\beta)}{\Gamma(Y+\alpha)\Gamma(n-Y+\beta)}p^{Y+\alpha-1}(1-p)^{n-Y+\beta-1},$$

for $p \in (0,1)$, which we recognize as the pdf of the $\text{Beta}(Y+\alpha, n-Y+\beta)$ distribution.

ii) The Bayes estimator of p is therefore

$$\hat{p}_{\text{Bayes}} = \frac{Y+\alpha}{n+\alpha+\beta},$$

which is obtained by using a formula for the mean of the Beta distribution. It is interesting to note that the Bayes estimator $\hat{p}_{\text{Bayes}}$ can be written as

$$\hat{p}_{\text{Bayes}} = \frac{Y}{n}\left(\frac{n}{n+\alpha+\beta}\right) + \frac{\alpha}{\alpha+\beta}\left(\frac{\alpha+\beta}{n+\alpha+\beta}\right),$$

which is a weighted average of the data-based estimator Y/n of p and the mean $\alpha/(\alpha+\beta)$ of the prior distribution of p. What happens if the sample size is very large?

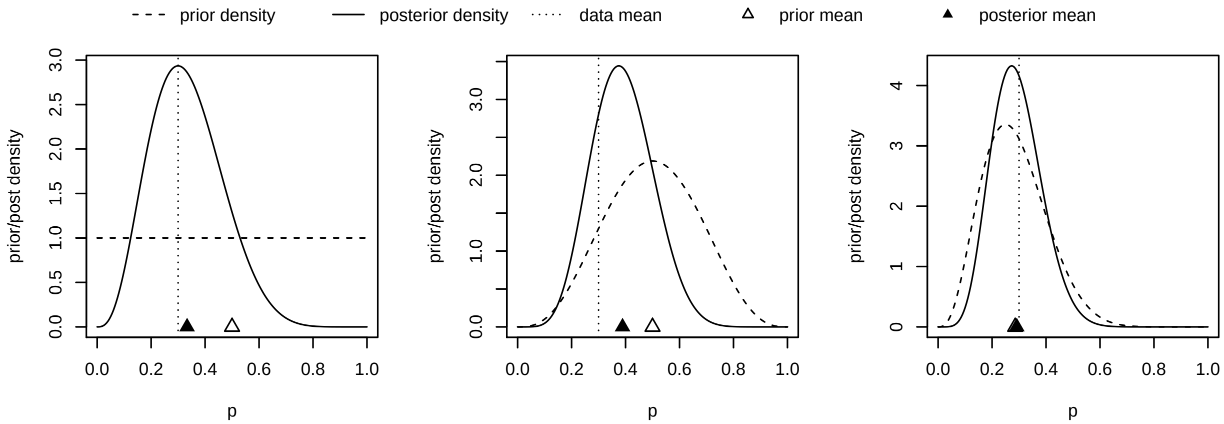
iii) We get the following with $n = 10$ and $Y = 3$:

(a) Under $\alpha = 1$, $\beta = 1$ the posterior mean is $1/3$.

(b) Under $\alpha = 4$, $\beta = 4$ the posterior mean is $7/18$.

(c) Under $\alpha = 4$, $\beta = 10$ the posterior mean is $7/24$.

The following plots depict under each setting the prior and posterior distributions as well as the data mean Y/n and the prior and posterior means:



Note that the posterior mean always lies between the data mean and the prior mean. This is the effect of the data on our beliefs about the parameter; we do not abandon our prior beliefs (the prior mean), but we allow the observed data to update our beliefs.

- **Strategy for finding the posterior pdf/pmf:** Suppose we have

$$\begin{aligned} X_1, \dots, X_n | \theta &\sim f(x_1, \dots, x_n | \theta) \\ \theta &\sim \pi(\theta), \end{aligned}$$

and suppose that θ is continuous so that

$$\pi(\theta | X_1, \dots, X_n) = \frac{f(X_1, \dots, X_n | \theta)\pi(\theta)}{\int_\Theta f(X_1, \dots, X_n | \tilde{\theta})\pi(\tilde{\theta}) d\tilde{\theta}}.$$

A very helpful strategy for simplifying $\pi(\theta | X_1, \dots, X_n)$ is to separate the part of it which is a function of θ from the part of it which is constant with respect to θ:

The quantity in the denominator does not involve θ (θ has been integrated out), so we may write

$$\pi(\theta | X_1, \dots, X_n) = C_1 f(X_1, \dots, X_n | \theta)\pi(\theta),$$

where $C_1 = [\int_\Theta f(X_1, \dots, X_n | \tilde{\theta})\pi(\tilde{\theta}) d\tilde{\theta}]^{-1}$ is a constant. We can furthermore separate the part of $f(X_1, \dots, X_n | \theta)\pi(\theta)$ which is a function of θ from the part of it which is constant with respect to θ; let $g(\theta)$ be a function of θ such that we may write

$$f(X_1, \dots, X_n | \theta)\pi(\theta) = C_2 g(\theta),$$

where C_2 is some quantity which does not involve θ. Then we can write

$$\pi(\theta | X_1, \dots, X_n) = C_1 f(X_1, \dots, X_n | \theta)\pi(\theta) = C_1 C_2 g(\theta) = C g(\theta),$$

where $C = C_1 C_2$. To avoid writing explicit expressions for these constants, Bayesians often make use of the "proportional to" symbol "$\propto$": We write $a \propto b$ if there exists a constant C such that $a = Cb$. Using this symbol allows us to write

$$\pi(\theta | X_1, \dots, X_n) \propto f(X_1, \dots, X_n | \theta)\pi(\theta) \propto g(\theta),$$

for some $g(\theta)$, where the proportionality symbol absorbs the constants.

Our time-saving strategy for finding the posterior pdf $\pi(\theta|X_1,\dots,X_n)$ is thus the following:

1. *Find a function $g(\theta)$ such that $f(X_1,\dots,X_n|\theta)\pi(\theta)\propto g(\theta)$.*
2. *Then $\pi(\theta|X_1,\dots,X_n)=Cg(\theta)$, where C is the normalizing constant $C=[\int_\Theta g(\theta)d\theta]^{-1}$.*

Sometimes we can recognize that $g(\theta)$ is proportional to a familiar pdf, in which case the constant C can be pulled from the full expression for the pdf.

Note: If θ is discrete, the above is the same, but with integrals replaced by sums.

- **Beta-Binomial exercise redone with time-saving strategy:** We have

$$\pi(p|Y)\propto f(Y|p)\pi(p)\propto p^Y(1-p)^{n-Y}p^{\alpha-1}(1-p)^{\beta-1}=p^{Y+\alpha-1}(1-p)^{n-Y+\beta-1}.$$

We recognize that this is proportional to the $\text{Beta}(Y+\alpha,n-Y+\beta)$ distribution. Multiplying by the appropriate normalizing constant gives

$$\pi(p|Y)=\frac{\Gamma(n+\alpha+\beta)}{\Gamma(n+\alpha)\Gamma(n-Y+\beta)}p^{Y+\alpha-1}(1-p)^{n-Y+\beta-1}.$$

- **Exercise:** Suppose

$$\begin{aligned}X_1,\dots,X_n|\lambda &\overset{\text{ind}}{\sim}\text{Poisson}(\lambda)\\ \lambda&\sim\text{Gamma}(\alpha,\beta).\end{aligned}$$

i) Find the posterior distribution of $\lambda|X_1,\dots,X_n$.

ii) Find an expression for $\hat{\lambda}_{\text{Bayes}}=\mathbb{E}[\lambda|X_1,\dots,X_n]$.

iii) Under $\alpha=4$ and $\beta=5$ and a sample size of $n=3$, compute the posterior mean of $\lambda|X_1,\dots,X_n$ when $\bar{X}_n=10,15,30$.

Answers:

i) The joint pmf of the data $X_1,\dots,X_n|\lambda$ is given by

$$f(X_1,\dots,X_n|\lambda)=\prod_{i=1}^n\frac{e^{-\lambda}\lambda^{X_i}}{X_i!}=\frac{e^{-n\lambda}\lambda^{n\bar{X}}}{\prod_{i=1}^n X_i!},$$

and the prior pdf of λ is given by

$$\pi(\lambda)=\frac{1}{\Gamma(\alpha)\beta^\alpha}\lambda^{\alpha-1}\exp\left[-\frac{\lambda}{\beta}\right],$$

so the posterior pdf of $\lambda | X_1, \dots, X_n$ is given by

$$\begin{aligned}\pi(\lambda | X_1, \dots, X_n) &= \frac{f(X_1, \dots, X_n | \lambda)\pi(\lambda)}{\int_0^\infty f(X_1, \dots, X_n | \tilde{\lambda})\pi(\tilde{\lambda})d\tilde{\lambda}} \\ &\propto f(X_1, \dots, X_n | \lambda)\pi(\lambda) \\ &= \frac{e^{-n\lambda}\lambda^{n\bar{X}}}{\prod_{i=1}^n X_i!}\frac{1}{\Gamma(\alpha)\beta^\alpha}\lambda^{\alpha-1}\exp\left[-\frac{\lambda}{\beta}\right] \\ &\propto \lambda^{n\bar{X}_n+\alpha-1}\exp\left[-\lambda\left(\frac{1}{\beta}+n\right)\right] \\ &= \lambda^{n\bar{X}_n+\alpha-1}\exp\left[-\lambda\left(\frac{\beta}{1+n\beta}\right)^{-1}\right],\end{aligned}$$

which is proportional to the pdf of the $\text{Gamma}(n\bar{X}_n + \alpha, \beta/(1+n\beta))$ distribution. So we have

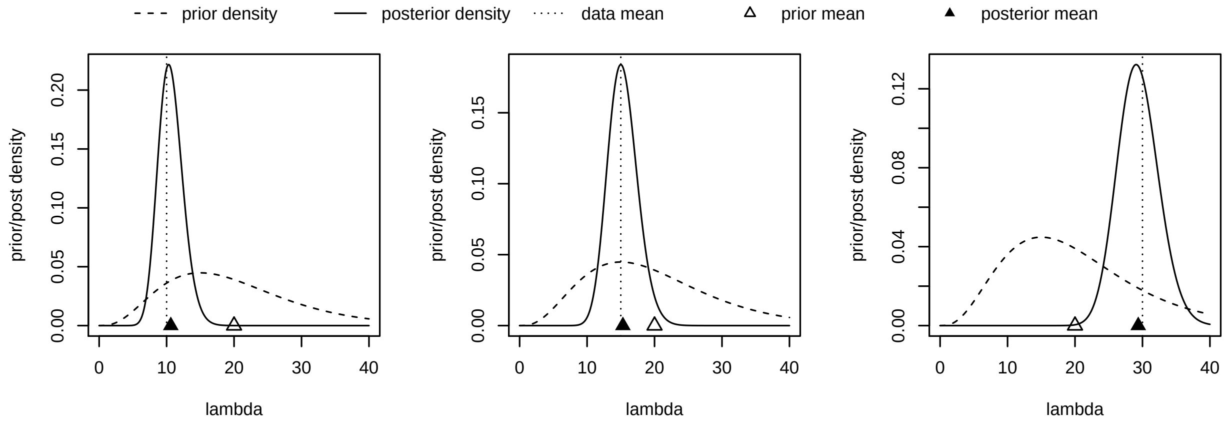
$$\lambda | X_1, \dots, X_n \sim \text{Gamma}\left(n\bar{X}_n + \alpha, \frac{\beta}{1+n\beta}\right).$$

ii) The posterior mean of $\lambda | X_1, \dots, X_n$ is

$$\hat{\lambda}_{\text{Bayes}} = (n\bar{X}_n + \alpha)\left(\frac{\beta}{1+n\beta}\right) = \bar{X}_n\left(\frac{n\beta}{1+n\beta}\right) + \alpha\beta\left(\frac{1}{1+n\beta}\right),$$

which is a weighted average of the data mean $\bar{X}_n$ and the mean of the prior distribution $\alpha\beta$.

iii) Under $\alpha = 4$ and $\beta = 5$ and $n = 3$, the posterior means of $\lambda | X_1, \dots, X_n$ when $\bar{X}_n = 10, 15, 30$ are 10.625, 15.3125, and 29.375, respectively. Note that the prior mean is $\alpha\beta = 4(5) = 20$. The plots below show the prior density of λ as well as the posterior density of $\lambda | X_1, \dots, X_n$ at each of these values of $\bar{X}_n$.



- **Exercise:** Let σ^2 be a known constant and suppose

$$\begin{aligned}Y_1, \dots, Y_n | \mu &\overset{\text{ind}}{\sim} \text{Normal}(\mu, \sigma^2) \\ \mu &\sim \text{Normal}(\mu_0, \tau^2),\end{aligned}$$

where $Y_1, \dots, Y_n$ are conditionally independent given μ.

i) Find the posterior distribution of $\mu | Y_1, \dots, Y_n$.

ii) Find an expression for $\hat{\mu}_{\text{Bayes}} = \mathbb{E}[\mu | Y_1, \dots, Y_n]$.

iii) Let $\sigma = 10$, $\tau = 20$, and $\mu_0 = 100$. Suppose that under sample sizes of $n = 1, 5, 15$, the sample mean $\bar{Y}_n = 120$ is observed. Compute the posterior mean of $\mu | Y_1, \dots, Y_n$ in each case.

Answer:

i) The pdf of $Y_1, \dots, Y_n$ conditional on μ is

$$\begin{aligned} f(Y_1, \dots, Y_n | \mu) &= \prod_{i=1}^{n} (2\pi)^{-1/2} (\sigma^2)^{-1/2} \exp\left[-\frac{1}{2\sigma^2} (Y_i - \mu)^2 \right] \\ &= (2\pi)^{-n/2} (\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \mu)^2 \right] \end{aligned}$$

and the pdf of the prior distribution of μ is

$$\pi(\mu) = (2\pi)^{-1/2} (\tau^2)^{-1/2} \exp\left[-\frac{1}{2\tau^2} (\mu - \mu_0)^2 \right]$$

The pdf of the posterior distribution of μ is thus

$$\begin{aligned} \pi(\mu | Y_1, \dots, Y_n) &= \frac{f(Y_1, \dots, Y_n | \mu) \pi(\mu)}{\int_{-\infty}^{\infty} f(Y_1, \dots, Y_n | \tilde{\mu}) \pi(\tilde{\mu}) d\tilde{\mu}} \\ &\propto f(Y_1, \dots, Y_n | \mu) \pi(\mu) \\ &\propto \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \mu)^2 \right] \exp\left[-\frac{1}{2\tau^2} (\mu - \mu_0)^2 \right] \\ &= \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \mu)^2 - \frac{1}{2\tau^2} (\mu - \mu_0)^2 \right] \\ \text{(expand sums)} &= \exp\left[-\frac{1}{2\sigma^2} \left(\sum_{i=1}^{n} Y_i^2 - 2 \sum_{i=1}^{n} Y_i \mu + n\mu^2 \right) + \frac{1}{2\tau^2} (\mu^2 - 2\mu\mu_0 + \mu_0^2) \right] \\ \text{(take out const)} &= \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} Y_i^2 + \frac{1}{\tau^2} \mu_0^2 \right] \exp\left[-\frac{1}{2\sigma^2} \left(-2n\bar{Y}_n \mu + n\mu^2 \right) + \frac{1}{2\tau^2} (\mu^2 - 2\mu\mu_0) \right] \\ &\propto \exp\left[-\frac{1}{2} \left(-2\mu \left[\frac{\bar{Y}_n}{\sigma^2/n} + \frac{\mu_0}{\tau^2} \right] + \mu^2 \left[\frac{1}{\sigma^2/n} + \frac{1}{\tau^2} \right] \right) \right] \\ \text{(isolate } \mu^2) &= \exp\left[-\frac{1}{2} \left(-2\mu \underbrace{\left[\frac{\bar{Y}_n}{\sigma^2/n} + \frac{\mu_0}{\tau^2} \right] \left[\frac{1}{\sigma^2/n} + \frac{1}{\tau^2} \right]^{-1}}_{=\mu_*} + \mu^2 \right) \Big/ \underbrace{\left[\frac{1}{\sigma^2/n} + \frac{1}{\tau^2} \right]^{-1}}_{=\sigma_*^2} \right]. \end{aligned}$$

Defining

$$\sigma_*^2 = \left[\frac{1}{\sigma^2/n} + \frac{1}{\tau^2}\right]^{-1} = \frac{(\sigma^2/n)\tau^2}{\sigma^2/n + \tau^2}$$

and

$$\mu_* = \left[\frac{\bar{Y}_n}{\sigma^2/n} + \frac{\mu_0}{\tau^2}\right]\left[\frac{1}{\sigma^2/n} + \frac{1}{\tau^2}\right]^{-1} = \left[\frac{\bar{Y}_n}{\sigma^2/n} + \frac{\mu_0}{\tau^2}\right]\frac{(\sigma^2/n)\tau^2}{\sigma^2/n + \tau^2} = \frac{\tau^2\bar{Y}_n + (\sigma^2/n)\mu_0}{\sigma^2/n + \tau^2},$$

we have

$$\begin{aligned}
\pi(\mu|Y_1,\dots,Y_n) &\propto \exp\left[-\frac{1}{2\sigma_*^2}\left(-2\mu\mu_* + \mu^2\right)\right] \\
\text{(complete the square)} &= \exp\left[-\frac{1}{2\sigma_*^2}\left(\mu_*^2 - 2\mu\mu_* + \mu^2\right)\right]\exp\left[-\frac{1}{2\sigma_*^2}(-\mu_*^2)\right] \\
&\propto \exp\left[-\frac{1}{2\sigma_*^2}(\mu - \mu_*)^2\right],
\end{aligned}$$

which is proportional to the pdf of the $\text{Normal}(\mu_*, \sigma_*^2)$ distribution. So we have

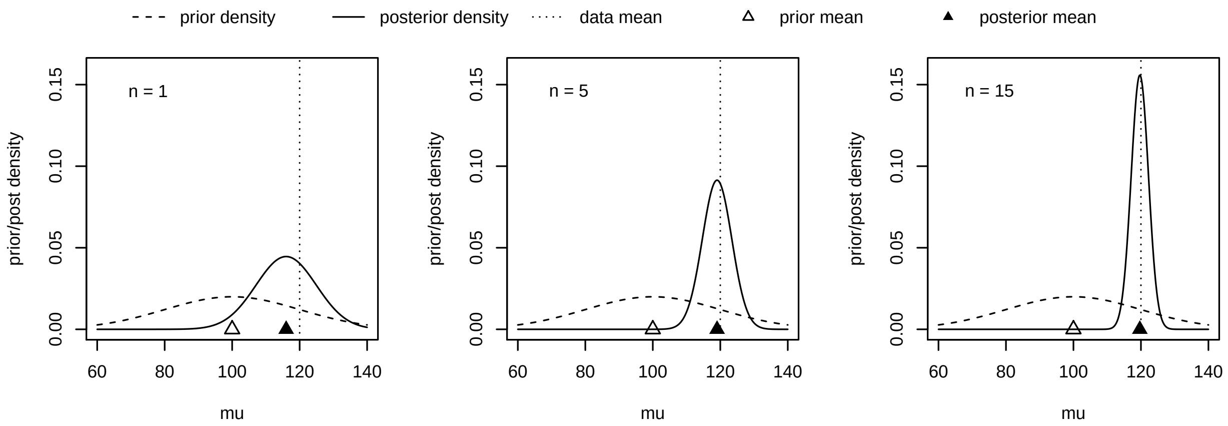
$$\mu|Y_1,\dots,Y_n \sim \text{Normal}\left(\frac{\tau^2\bar{Y}_n + (\sigma^2/n)\mu_0}{\sigma^2/n + \tau^2}, \frac{(\sigma^2/n)\tau^2}{\sigma^2/n + \tau^2}\right).$$

ii) The posterior mean of $\mu|X_1,\dots,X_n$ is given by

$$\hat{\mu}_{\text{Bayes}} = \frac{\tau^2\bar{Y}_n + (\sigma^2/n)\mu_0}{\sigma^2/n + \tau^2} = \bar{Y}_n\left(\frac{\tau^2}{\sigma^2/n + \tau^2}\right) + \mu_0\left(\frac{\sigma^2/n}{\sigma^2/n + \tau^2}\right),$$

which is a weighted average of the data-based estimator $\bar{Y}_n$ and the prior mean μ_0. What happens as $n \to \infty$?

iii) For $\bar{Y}_n = 120$ under $n = 1, 5, 15$, we get $\hat{\mu}_{\text{Bayes}} = 116, 119.0476, 119.6721$, respectively. The plots below show the prior density of μ as well as the posterior density of $\mu|X_1,\dots,X_n$ for $n = 1, 5, 15$.



- **Conjugacy:** In the three examples so far, the posterior distribution of the parameter belonged to the same family of distributions as the prior distribution. For the Beta-Binomial model, the posterior distribution was a Beta distribution; for the Gamma-Poisson model, the posterior distribution was a Gammma distribution; for the Normal-Normal model, the posterior distribution was a Normal distribution. In each case, the parameters of the prior distributions were updated by the data to give the parameters of the posterior distribution. In the Bayesian paradigm, a prior distribution is called a *conjugate prior* if the posterior distribution resulting from it belongs to the same family of distributions.

  We do not need to use conjugate priors! They just happen to be convenient, because they result in very simple posterior distributions which allow us to compute quantities like the posterior mean with simple formulas. The choice of a non-conjugate prior may result in a posterior distribution with a form we do not recognize, which may make it difficult to compute the posterior quantities in which we are interested. For example, we may not be able to compute the posterior mean $\mathbb{E}\theta|X_1, \ldots, X_n$ directly using any formula; in such a case we would have to search for it algorithmically. Much effort has gone into devising these kinds of algorithms, one class of which are Markov Chain Monte Carlo methods, which we will not cover in this class.

Bayesian credible intervals

- The Bayesian paradigm provides a very simple way to construct something like a confidence interval for a parameter.

- Recall that a frequentist $(1-\alpha)100\%$ confidence interval for a parameter θ based on the data $X_1, \ldots, X_n$ takes the form (L_n, U_n), where $L_n = L_n(X_1, \ldots, X_n)$ and $U_n = U_n(X_1, \ldots, X_n)$ are functions of the data such that

  $$P(L_n < \theta < U_n) = 1 - \alpha.$$

  Here θ is a fixed constant and the endpoints of the interval are random; if you sampled data 1000 times and computed the endpoints of the interval 1000 times, you would expect the interval to capture the true (fixed) value of θ about $(1-\alpha)1000$ times.

- **Bayesian credible interval:** In the Bayesian framework, any interval (L_n, U_n), where $L_n = L_n(X_1, \ldots, X_n)$ and $U_n = U_n(X_1, \ldots, X_n)$ are functions of the data, such that

  $$P(L_n < \theta < U_n|X_1, \ldots, X_n) = 1 - \alpha$$

  is called a $(1-\alpha)100\%$ *Bayesian credible interval.* Inside the probability statement L_n and U_n are fixed due to conditioning on $X_1, \ldots, X_n$ and θ is random (note that this probability statement would make no sense in the frequentist setup). The following are two standard ways of choosing L_n and U_n:

  - *Equal tails interval:* Choose L_n and U_n such that

    $$P(\theta < L_n|X_1, \ldots, X_n) = P(\theta > U_n|X_1, \ldots, X_n) = \alpha/2.$$

- *Highest posterior density interval:* Choose L_n and U_n such that (L_n, U_n) is the smallest interval such that $P(L_n < \theta < U_n | X_1, \dots, X_n) = 1 - \alpha$.

If the posterior distribution of θ is symmetric, the two strategies above will produce the same interval.

- **Exercise:** As a continuation of the previous exercise, compute 95% Bayesian credible intervals for μ in the model

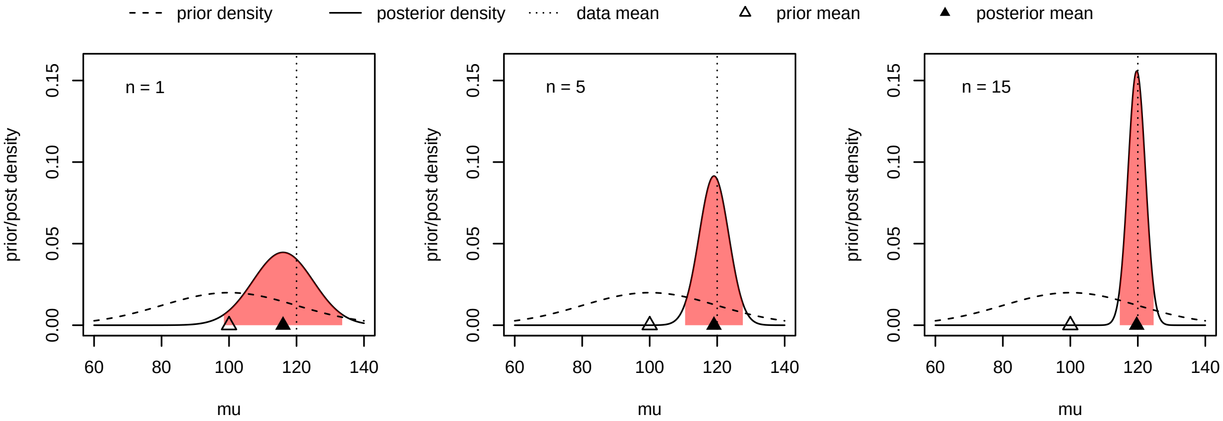
$$\begin{aligned} Y_1, \dots, Y_n | \mu &\overset{\text{ind}}{\sim} \text{Normal}(\mu, 10^2) \\ \mu &\sim \text{Normal}(100, 20^2), \end{aligned}$$

when $\bar{Y}_n = 120$ under the sample sizes $n = 1, 5, 15$.

Answer: We get the following:

- For $n = 1$, $\mu | Y_1 \sim \text{Normal}(116, 80)$. The upper and lower 0.025 quantiles of the posterior distribution are $U_n =$ `116+qnorm(.975)*sqrt(80)` $= 133.5305$ and $L_n =$ `116-qnorm(.975)*sqrt(80)` $= 98.46955$, so the 95% credible interval is $(98.46955, 133.5305)$.
- For $n = 5$ we get $(110.4936, 127.6016)$.
- For $n = 15$ we get $(114.6532, 124.6911)$.

The plots below depict the 95% credible intervals.



Bayesian hypothesis testing

- As before: For a parameter θ which takes values in some space Θ, we consider

$$H_0\colon\ \theta \in \Theta_0 \text{ versus } H_1\colon\ \theta \in \Theta_1,$$

where $\Theta_0 \cup \Theta_1 = \Theta$ and $\Theta_0 \cap \Theta_1 = \emptyset$.

- Frequentist tests of H_0 versus H_1 based on some data $X_1, \ldots, X_n$ depending on θ are of the form

  $$\text{Reject } H_0 \text{ iff } T(X_1, \ldots, X_n) \in \mathcal{R},$$

  where $T(X_1, \ldots, X_n)$ is a function of the data called the *test statistic* and $\mathcal{R}$ is a set called the *rejection region*. The frequentist chooses the rejection region $\mathcal{R}$ which makes the Type II error probability as small as possible for $\theta \in \Theta_1$ while ensuring that the Type I error probability is less than or equal to some value $\alpha \in (0, 1)$.

- Bayesian tests of hypotheses are constructed in a fundamentally different way because they treat the parameter as a random variable. Bayesian tests are based on the probability of the event $\theta \in \Theta_0$ according to the prior distribution of θ and according to the posterior distribution of θ conditional on the observed data $X_1, \ldots, X_n$. Let

  $$\pi_0 = P(\theta \in \Theta_0) \quad \text{and} \quad \pi_1 = P(\theta \in \Theta_1)$$

  denote the prior probabilities of H_0 and H_1, respectively, and let

  $$p_0 = P(\theta \in \Theta_0 | X_1, \ldots, X_n) \quad \text{and} \quad p_1 = P(\theta \in \Theta_1 | X_1, \ldots, X_n)$$

  denote the posterior probabilities of H_0 and H_1, respectively, conditional on $X_1, \ldots, X_n$.

- **Definitions:** The *prior odds* in favor of H_0 over H_1 are π_0/π_1 and the *posterior odds* in favor of H_0 over H_1 are p_0/p_1. The *Bayes factor* in favor of H_0 over H_1 is the ratio

  $$B = \frac{p_0/p_1}{\pi_0/\pi_1}.$$

- The Bayes factor reflects how much the data have changed our beliefs in favor of H_0 over H_1.

- A ____________(large/small) Bayes factor indicates that the data carry ____________(much/little) evidence in favor of H_0 over H_1.

- If the Bayes factor is ____________(less than/greater than) 1, the data have changed our prior beliefs in favor of ____________(H_0/H_1).

- **Exercise:** Suppose we have

  $$\begin{aligned} Y|p &\sim \text{Binomial}(n, p) \\ p &\sim \text{Beta}(\alpha, \beta) \end{aligned}$$

  and that we are interested in testing the hypotheses

  $$H_0\colon\ p \leq 1/2 \text{ versus } H_1\colon\ p > 1/2.$$

  Let $n = 100$ and let $Y = 55$ be observed, and set $\alpha = 10$ and $\beta = 10$.

  i) Find the posterior probability p_0 of H_0, that is of the event $p \leq 1/2$.

ii) Find $P(Y \geq 55|p = 1/2)$ and interpret this quantity from the frequentist perspective.

iii) Find the prior odds in favor of H_0 over H_1.

iv) Find the posterior odds in favor of H_0 over H_1.

v) Compute the Bayes factor of the data in favor of H_0 over H_1.

Answers:

i) Using our work from the first example of these notes for getting the posterior distribution of $p|Y$, the posterior probability of H_0 is

$$p_0 = P(p \leq p_0|Y = 55) = \texttt{pbeta(1/2,55+10,100-55+10)} = 0.1796791.$$

ii) We have

$$P(Y \geq 55|p = 1/2) = \texttt{1-pbinom(54,100,.5)} = 0.1841008.$$

This is the frequentist p-value. Interestingly, the frequentist p-value is close to the posterior probability p_0 of H_0. If the p-value is small, we consider the evidence against H_0 to be strong; likewise, if the posterior probability of H_0 in the Bayesian setup is small, we consider the evidence against H_0 to be strong. For these data, the evidence against H_0 is not very strong from either the frequentist or the Bayesian perspective.

iii) The prior probability of H_0 is

$$\pi_0 = P(p \leq p_0) = \texttt{pbeta(1/2,10,10)} = 0.5,$$

and $\pi_1 = 1 - 0.5 = 0.5$. So the prior odds in favor of H_0 are

$$\pi_0/\pi_1 = 0.5/0.5 = 1.$$

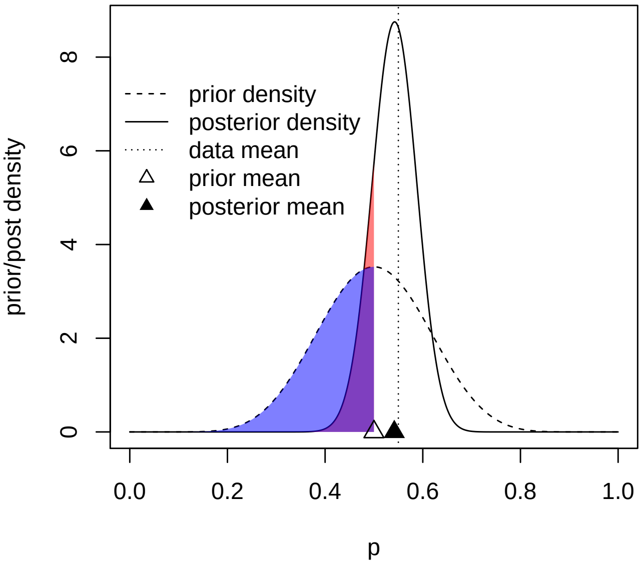
The prior distribution gives equal probabilities to $\theta \in \Theta_0$ and $\theta \in \Theta_1$. In the plot further below, we see that the prior density is symmetric around $1/2$.

iv) The posterior odds in favor of H_0 over H_1 are

$$p_0/p_1 = 0.1796791/(1 - 0.1796791) = 0.2190351.$$

v) The Bayes factor in favor of H_0 over H_1 is $B = 0.2190351/1 = 0.2190351$.

Below is a plot showing the prior and posterior densities with shaded areas corresponding to the prior and posterior probabilities of H_0.


prior density
posterior density
data mean
△ prior mean
▲ posterior mean
prior/post density
0
2
4
6
8
0.0
0.2
0.4
0.6
0.8
1.0
p