Decision Theory and Bayesian Analysis

Dr. Vilda Purutcuoglu

[1]

[1]Edited by Anil A. Aksu based on lecture notes of STAT 565 course by Dr. Vilda Purutcuoglu

Contents

LECTURE 1
Bayesian Paradigm

1.1. Bayes theorem for distributions

If A and B are two events,

$$P(A \mid B) = \frac{P(A)P(B \mid A)}{P(B)}. \tag{1.1}$$

This is just a direct consequence of the multiplication law of probabilities that says we can express $P(A \mid B)$ as either $P(A)P(B \mid A)$ or $P(B)P(A \mid B)$. For discrete distributions, if Z, Y are discrete random variables

$$P(Z = z \mid Y = y) = \frac{P(Z = z)P(Y = y \mid Z = z)}{P(Y = y)}. \tag{1.2}$$

- How many distributions do we deal with here?

We can express the denominator in terms of the distribution in the numerator[**1**].

$$P(Y = y) = \sum_z P(Y = y, Z = z) = \sum_z P(Z = z)P(Y = y \mid Z = z). \tag{1.3}$$

- This is sometimes called the law of total probability

In this context, it is just an expression of the fact that as z ranges over the possible values of Z , the probabilities on the left hand-side of equation 1.2 make up the distribution of Z given $Y = y$, and so they must add up to one. The extension to continuous distribution is easy. If Z, Y are continuous random variable,

$$f(Z \mid Y) = \frac{f(Z)f(Y \mid Z)}{f(Y)}. \tag{1.4}$$

where the denominator is now expressed as an integral:

$$f(Y) = \int f(Z)f(Y \mid Z)\mathrm{d}Z. \tag{1.5}$$

$$f = \begin{cases} \mathit{continous} \quad \mathit{name}? \\ \mathit{discrete} \quad \mathit{name}? \end{cases} \tag{1.6}$$

1.2. How Bayesian Statistics Uses Bayes Theorem

Theorem 1.7 (Bayes' theorem)**.**

$$P(A \mid B) = \frac{P(A)P(B \mid A)}{P(B)}$$

$P(B)$=if we are interested in the event B, $P(B)$ is the initial or prior probability of the occurence of event B. Then we observe event A $P(B \mid A)$ = How likely B is when A is known to have occurred is the posterior probability $P(B \mid A)$.

Bayes' theorem can be understood as a formula for updating from prior to posterior probability, the updating consists of multiplying by the ratio $P(B \mid A)/P(A)$. It describes how a probability changes as we learn new information. Observing the occurrence of A will increase the probability of B if $P(B \mid A) > P(A)$. From the law of total probability,

$$P(A) = P(A \mid B)P(B) + P(A \mid B^c) + P(A \mid B^c)P(B^c). \tag{1.8}$$

where $P(B^c) = 1 - P(B)$.

Lemma 1.9.

$$P(A \mid B) - P(A) = \frac{P(A) - P(A \mid B^c)P(B^c)}{1 - P(B^c)} - P(A)$$

Proof.

$$P(A \mid B) - P(A) = \frac{P(A) - P(A \mid B^c)P(B^c) - P(A) + P(A)P(B^c)}{P(B)}$$

$$P(A \mid B) - P(A) = \frac{P(B^c)(P(A) - P(A \mid B^c))}{P(B)}$$

$$P(A \mid B) - P(A) = P(B^c)(\frac{P(B)P(A \mid B) + P(B^c)P(A \mid B^c)}{P(B)} - \frac{P(A \mid B^c)}{P(B)})$$

$$P(A \mid B) - P(A) = P(B^c)(P(A \mid B) - \frac{P(A \mid B^c)(1 - P(B^c))}{P(B)})$$

$$P(A \mid B) - P(A) = P(B^c)(P(A \mid B) - P(A \mid B^c))$$

□

1.2.1. Generalization of the Bayes' Theorem

Let $B_1, ..., B_n$ be a set of mutually exclusive events. Then

$$P(B_r \mid A) = \frac{P(B_r)P(A \mid B_r)}{P(A)} = \frac{P(B_r)P(A \mid B_r)}{\sum_{i=1}^{n} P(B_r)P(A \mid B_r)}. \tag{1.10}$$

- Assuming that $P(B_r) > 0$,$P(A \mid B) > P(A)$ if and only if $P(A \mid B) > P(A \mid B^c)$.

- In Bayesian inference we use Bayes' theorem in a particular way.
- Z is the parameter (vector) θ.
- Y is the data (vector) X.

So we have

$$f(\theta \mid X) = \frac{f(\theta)f(X \mid \theta)}{f(X)} \tag{1.11}$$

$$f(X) = \int f(\theta)f(X \mid \theta)\mathrm{d}\theta. \tag{1.12}$$

$$f(\theta) = \tag{1.13}$$

$$f(\theta \mid X) = \tag{1.14}$$

$$f(X \mid \theta) = \tag{1.15}$$

1.2.2. Interpreting our sense

How do we interpret the things we see, hear, feel, taste or smell?

Example 1.2.1. I hear a song on the radio I identify the singer as Robbie Williams. Why do I think it's Robbie Williams?. Because he sounds like that. Formally, $P($ What I hear Robbie Williams $) >> P($What I hear someone else $)$

Example 1.2.2. I look out of the window and see what appears to be a tree. It has a big, dark coloured part sticking up out of the ground that branches into thinner sticks and on the ends of these are small green things. Clearly, $P(view \mid tree)$ is high and $P(view \mid car)$ or $P(view \mid Robbie \quad Williams)$ are very small. But $P(view \mid carboard \quad cutout \quad cunningly \quad painted \quad to \quad look \quad like \quad a \quad tree)$ is also very high. Maybe even higher than $P(view \mid tree)$ in the sense that what I see looks almost like a tree.
Does this mean I should now believe that I am seeing a cardboard cut-out cunningly painted to look like a tree? No because it is much less likely to begin with than a red tree.

In statistical terms, consider some data X and some unknown parameter θ. The first step in any statistical analysis is to build a model that links the data to unknown parameters and the main function of this model is to allow us to state the probability of observing any data given any specified values of the parameters. That is the model defines $f(x \mid \theta)$.

When we think of $f(x \mid \theta)$ as a function of θ for fixed observed data X, we call it likelihood function and it by $L(\theta, X)$.

- So how can we combine this with our example?

This perspective underlies the differences between the two main theories of statistical inference.

- Frequentist inference essentially uses only the likelihood, it does not recognize $f(\theta)$.
- Bayesian inference uses both likelihood and $f(\theta)$.

The principal distinguishing feature of Bayesian inference as opposed to frequentist inference is its use of $f(\theta)$.

1.3. Prior to Posterior

We refer to $f(\theta)$ as the prior distribution of θ. It represents knowledge about θ prior to observing the data X. We refer to $f(\theta \mid X)$ as the posterior distribution of θ and it represents knowledge about θ after observing X.

- So we have two sources of information about θ.
- Here $f(x)$ does not depend on θ. Thus $\int f(\theta \mid x)\mathrm{d}\theta = 1$. Since $f(x)$ is a constant within the integral, we can take it outside to get $1 = f^{-1}(x) \int f(\theta) f(x \mid \theta)\mathrm{d}\theta$.
- $f(\theta \mid x) = \propto f(\theta) f(x \mid \theta) \propto f(\theta) L(\theta; x)$ (the posterior is proportional to the prior times the likelihood).
- The constant that we require to scale the right hand side to integrate to 1 is usually called the normalizing constant. If we haven't dropped any constants form $f(\theta)$ or $f(x \mid \theta)$, then the normalising constant is just $f^{-1}(x)$, otherwise it also restores any dropped constants.

1.4. Triplot

If for any value of θ, we have either $f(\theta) = 0$ or $f(x \mid \theta) = 0$, then we will also have $f(\theta \mid x) = 0$. This is called the property of zero preservation. So if either:

- the prior information says that this θ value is impossible
- the data say that this value of θ is impossible because if it were the true value, then the observed data would have been impossible, then the posterior distribution confirms that this value of θ is impossible.

Definition 1.16. Crowwell's Rule: If either information source completely rules out a specific θ, then the posterior must rule it out too.

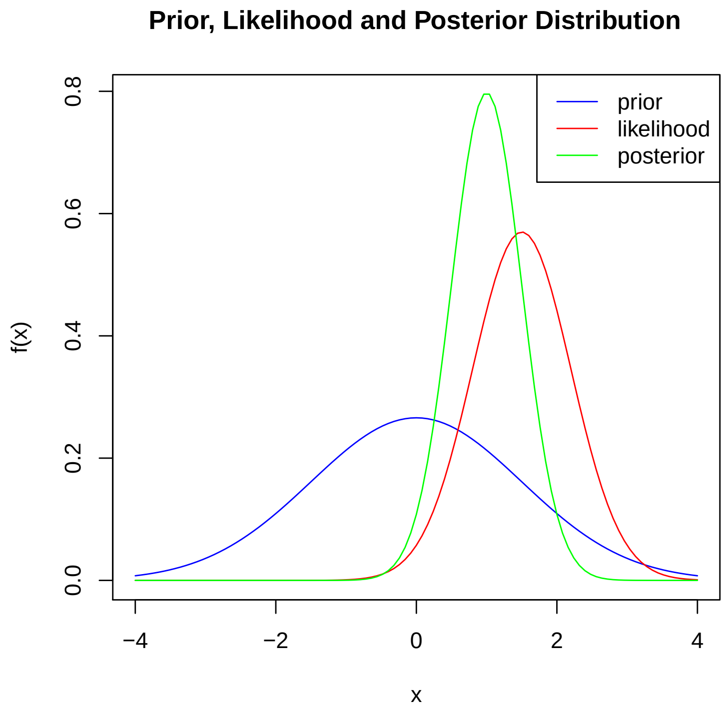

Figure 1. Triplot of prior, likelihood and posterior.

This means that we should be very careful about giving zero probability to something unless it is genuinely impossible. Once something has zero probability then no amount of further evidence can cause it to have a non-zero posterior probability.

- More generally, $f(\theta \mid x)$ will be low if either $f(\theta)$ is very small. We will tend to find that $f(x \mid \theta)$ is large when both $f(\theta)$ and $f(x \mid \theta)$ are relatively large, so that this θ value is given support by both information sources.

When θ is a scalar parameter, a useful diagram is the triplot, which shows the prior, likelihood and posterior on the same graph. An example is in Figure 1.[1]

A strong information source in the triplot is indicated by a crve that is narrow (and therefore, because it integrates to one, also has a high peak). A narrow curves concentrates on a small range of θ values, and thereby "rules out" all values of θ outside that range.

[1]All plots are generated in R, relevant codes are provided in Appendix R Codes

- Over the range $\theta < -1$, the likelihood:
- Over the range $\theta > 3$,the likelihood:
- Values of θ between -1 and 3, the likelihood:
- The maximum value of the posterior at:
- The MLE of θ is:

1.4.1. Normal Mean

For example, suppose that $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$ and σ^2 is known. Then the likelihood is :

$$\begin{aligned} f(x \mid \mu) = \prod_{i=1}^{n} f(x_i \mid \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x_i-\mu)^2\right) \\ \propto \exp\left(-\frac{1}{2\sigma^2}(x_i-\mu)^2\right). \end{aligned} \tag{1.17}$$

As,

$$\begin{aligned} \sum(x_i - \bar{x} + \bar{x} - \mu)^2 = \sum(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2 + 2(\bar{x}-\mu)\sum(x_i-\bar{x}) \\ = \sum(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2 \\ \propto \exp\left(-\frac{1}{2\sigma^2}n(\bar{x}-\mu)^2\right). \end{aligned} \tag{1.18}$$

Note that $2(\bar{x}-\mu)\sum(x_i-\bar{x}) = 0$ as $\sum(x_i-\bar{x}) = 0$. Suppose the prior distribution for μ is normal:

$$\mu \sim \mathcal{N}(m, v). \tag{1.19}$$

Then applying Bayes' theorem we have:

$$\begin{aligned} f(\mu \mid x) \propto \underbrace{\exp\left(-\frac{1}{2\sigma^2}n(\bar{x}-\mu)^2\right)}_{f(x\mid\mu)} \underbrace{\exp\left(-\frac{1}{2\sigma^2}n(\mu-m)^2\right)}_{f(\mu)} \\ = \exp\left(-\frac{\theta}{2}\right). \end{aligned} \tag{1.20}$$

Note that

$$\theta = n\sigma^{-2}(\bar{x}-\mu) + v^{-1}(\mu-m)^2 = (v^*)^{-1}(\mu-m^*)^2 + R \tag{1.21}$$

and

$$v^* = (n\sigma^{-2} + v^{-1})^{-1} \tag{1.22}$$

$$m^* = (n\sigma^{-2} + v^{-1})^{-1}(n\sigma^{-2}\bar{x} + v^{-1}m) = a\bar{x} + (1-a)m \tag{1.23}$$

where $a = n\sigma^-2/(n\sigma^{-2} + v^{-1})$

$$R = (n^{-1}\sigma^2 + v)(\bar{x} - m)^2 \tag{1.24}$$

Therefore,

$$f(\mu \mid x) \propto \exp\left(-\frac{1}{2\sigma^2}n(\mu - m)^2\right) \tag{1.25}$$

and we have shown that the posterior distribution is normal too: $\mu \mid x \sim \mathcal{N}(m^*, v^*)$

- m^* = weighted average of the mean m and the usual frequentist data-only estimate $\bar{x}$.
  The weights $\propto$:
- Bayes' theorem typically works in this way. We usually find that posterior estimates are compromises between prior estimates and data based estimates and tend to be closer whichever information source is stronger. And we usually find that the posterior variance is smaller than the prior variance.

1.4.2. Weak Prior Information

It is the case where the prior information is much weaker that the data. This will occur, for instance, if we do not have strong information about Q before seeing the data, and if there are lots of data. Then in triplot, the prior distribution will be much broader and flatter that the likelihood. So the posterior is approximately proportional to the likelihood.

Example 1.4.1. In the normal mean analysis, we get weak prior information by letting the prior precision of v^{-1} become small. Then $m^* \to \bar{x}$ and $v^* \to \sigma^2/n$ so that the posterior distribution of μ corresponds very closely with standard frequentist theory.

LECTURE 2
Some Common Probability Distributions

Binomial on $Y \in \{0, 1, ..., n\}$ with parameters $n \in \{1, 2, 3, ...\}$ and $p \in (0, 1)$ is denoted by $Bi(n, p)$ and

$$f(y \mid n, p) = \binom{n}{y} p^y (1-p)^{n-y} \tag{2.1}$$

for $y = 0, 1, ..., n$. The mean is given as:

$$E(y) = np. \tag{2.2}$$

Also the variance is given as:

$$v(y) = np(1-p). \tag{2.3}$$

Beta on $Y \in \{0, 1\}$ with parameters $a, b > 0$ is denoted by $Beta(p, q)$ and the density function is:

$$f(y \mid p, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1} \tag{2.4}$$

for $y \in (0, 1)$. The mean is given as:

$$E(y) = \frac{p}{p+q}, \tag{2.5}$$

Also the variance is given as:

$$v(y) = \frac{pq}{(p+q)^2(p+q+1)}. \tag{2.6}$$

$B(p, q) = \int_0^1 y^{p-1}(1-y)^{q-1}\mathrm{d}y$ is the beta function and defined to be the normalizing constant for this density.

- In beta distribution, p and q change the shape of the distribution. Discuss!

Uniform distribution on $Y \in \{l, r\}$ where $-\infty < l < r < \infty$ is denoted by uniform (l, r) and its pdf is:

$$f(y \mid l, r) = \frac{1}{r-l} \tag{2.7}$$

for $y \in \{l, r\}$. The mean is given as:

$$E(y) = \frac{l+r}{2}, \tag{2.8}$$

Also the variance is given as:

$$v(y) = \frac{(r-l)^2}{12}. \tag{2.9}$$

Poisson distribution on $Y \in \{0, 1, 2, ..., \}$ with parameter $\theta > 0$ is denoted by $Poisson(\theta)$ and its pdf is:

$$f(y \mid \theta) = \frac{\exp(-\theta)\theta^y}{y!} \tag{2.10}$$

for $y = 0, 1, 2, ...$ The mean and the variance are given as[**2**]:

$$E(y) = v(y) = \theta. \tag{2.11}$$

Gamma distribution on $Y > 0$ with shape parameter $\alpha > 0$ and rate parameter $\lambda > 0$ is denoted by $Gamma(\alpha, \lambda)$ and the corresponding density is:

$$f(y \mid \alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} y^{\alpha-1} \exp(-\lambda y) \tag{2.12}$$

for $y > 0$. The mean is given as:

$$E(y) = \frac{\alpha}{\lambda}, \tag{2.13}$$

Also the variance is given as:

$$v(y) = \frac{\alpha}{\lambda^2}. \tag{2.14}$$

Note that

$$\exp(\lambda) = Gamma(1, \lambda). \tag{2.15}$$

Univariate normal distribution on $Y \in \mathbb{R}$ with $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$ is denoted by $\mathcal{N}(\mu, \sigma^2)$ and its pdf is:

$$f(y \mid \mu, \sigma^2) = \frac{1}{\sigma} \left(\frac{1}{2\pi}\right)^{1/2} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}. \tag{2.16}$$

The mean is given as:

$$E(y) = \mu, \tag{2.17}$$

Also the variance is given as:

$$v(y) = \sigma^2. \tag{2.18}$$

<u>**K-variate normal distribution**</u> on $Y \in \mathbb{R}^k$ with vector $\mathbf{b} \in \mathbb{R}^k$ and positive definite symmetric (PDS) covariance matrix $\mathbf{C}$ is denoted by $\mathcal{N}_k(\mathbf{b}, \mathbf{C})$ and the corresponding density function is:

$$f(y \mid \mathbf{b}, \mathbf{C}) = \frac{1}{\underbrace{|\mathbf{C}|^{1/2}}_{determinant}} \frac{1}{(2\pi)^{k/2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mathbf{b})^T \mathbf{C}^{-1}(y-\mathbf{b})\right\}. \tag{2.19}$$

The mean is given as:

$$E(y) = \mathbf{b}, \tag{2.20}$$

And the covariance matrix is given as:

$$Cov(y) = \mathbf{C}. \tag{2.21}$$

2.1. Posterior

Not only the beta distributions are the simplest and the most convenient distributions for a random variable confined to $[0, 1]$, they also work very nicely as prior distribution for a binomial observation. If $\theta \sim Be(p, q)$ then

$$f(x \mid \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}. \tag{2.22}$$

for $x = 1, 2, ..., n$.

$$f(\theta) = \frac{1}{Be(p, q)} \theta^{p-1} (1-\theta)^{q-1} \tag{2.23}$$

where $0 \leq \theta \leq 1$ and $p, q > 0$.

$$\begin{aligned} f(x) = \int f(\theta) f(x \mid \theta) \mathrm{d}\theta &= \binom{n}{r} \frac{1}{Be(p, q)} \int_0^1 \theta^{p+x-1} (1-\theta)^{q+n-x-1} \mathrm{d}\theta \\ &= \binom{n}{r} \frac{Be(p+x, q+n-x)}{Be(p, q)}. \end{aligned} \tag{2.24}$$

From

$$f(\theta \mid x) = \frac{f(\theta) f(x \mid \theta)}{f(x)}. \tag{2.25}$$

(2.26)
$$f(\theta \mid x) = \frac{\theta^{p+x-1}(1-\theta)^{q+n-x-1}}{Be(p+x, q+n-x)} \propto \underbrace{\theta^{p-1}(1-\theta)^{q-1}}_{Beta \quad part} \underbrace{\theta^{x}(1-\theta)^{n-x}}_{Binomial \quad part}.$$

So $(\theta \mid x) \propto Beta(p+x, q+n-x)$. The posterior mean is:

$$E(\theta \mid x) = \frac{p+x}{p+q+n} = \frac{p+q}{p+q+n}E(\theta) + \frac{n}{p+q+n}\hat{\theta} \tag{2.27}$$

where $\hat{\theta} = x/n$. The posterior variance is:

$$\begin{aligned} v(\theta \mid x) &= \frac{(p+x)(q+n-x)}{(p+q+n)^2(p+q+n+1)} \\ &= \frac{E(\theta)(1-E(\theta \mid x))}{p+q+n+1} \end{aligned} \tag{2.28}$$

But,

$$v(\theta) = \frac{E(\theta)(1-\theta)}{p+q+1} \tag{2.29}$$

So the posterior has higher relative precision that the prior.
<u>SPECIAL NOTE:</u>
The classical theory of estimation regards an estimator as good if it is unbiased and has small variance, or more generally if its mean-square-error is small. The MSE is an average squared error where the error is the difference between θ, i.e. y in previous notation , and the estimate t. In accordance with classical theory, the average is taken with respect to the sampling distribution of the estimator.
In Bayesian inferenc, θ is a random variable and it is therefore appropriate to average the squared error with respect to the posterior distribution of θ. Consider

$$\begin{aligned} E\left\{(t-\theta)^2 \mid x\right\} &= E(t^2 \mid x) - E(2t\theta \mid x) + E(\theta^2 \mid x) \\ &= t^2 - E(2t\theta \mid x) + E(\theta^2 \mid x) \\ &= \{t - E(\theta \mid x)\}^2 + v(\theta \mid x). \end{aligned} \tag{2.30}$$

Therefore the estimate t which minimizes posterior expected square error is $t = E(\theta \mid x)$, the posterior mean. The posterior mean can therefore be seen as an estimate of θ which is the best in the sense of minimizing expected squared error. This is distinct from, but clearly related to, its more natural role as a useful summary of location of the posterior distribution.

2.2. Weak Prior

If we reduce the prior relative precision to zero by setting $p = q = 0$, we get $\theta \mid x \sim Be(x, n-x)$. Then $E(\theta \mid x) = \hat{\theta}$ and $v(\theta \mid x) = \hat{\theta}(1-\hat{\theta})/(n+1)$ results which nicely parallel standard frequentist theory.

- Notice that we are not really allowed to let either parameter of the beta distribution be zero. However, by making p and q extremely small, we get as close to these results as we like. We can think of $p = q = 0$ as a defining limiting (if strictly improper) case of weak prior information.

Example 2.2.1. A doctor proposes a new treatment protocol for a certain kind of cancer. With current methods about 40% of patients with this cancer survive six months after diagnosis. After one year of using the new protocol, 15 patients with diagnosis, of whom 6 survived. After two years a further 55 patients have been followed to teh six months mark, of whom 28 survived. So in total we have 34 patients surviving out of 70.
Let θ be the true success rate of the new treatment protocol, i.e. the true proportion of patients who survive 6 months and we wish to make comparison of θ with the current survival rate 40%.
Suppose that the doctor in charge has prior information leading her to assign a prior distribution with expectation $E(\theta) = 0.45$, i.e. expects a slight improvement over the existing protocol, form 40% to 45%, however her prior standard deviation is 0.07, $v(\theta) = 0.07^2 = 0.0049$

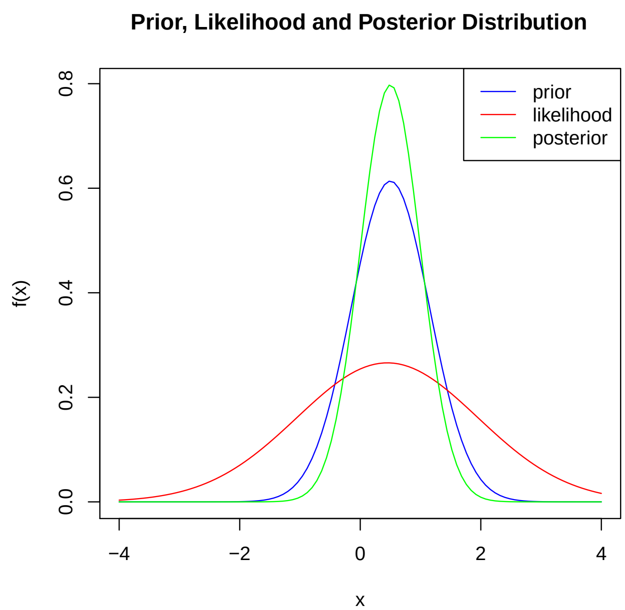

Figure 1. The triplot from first year's data.

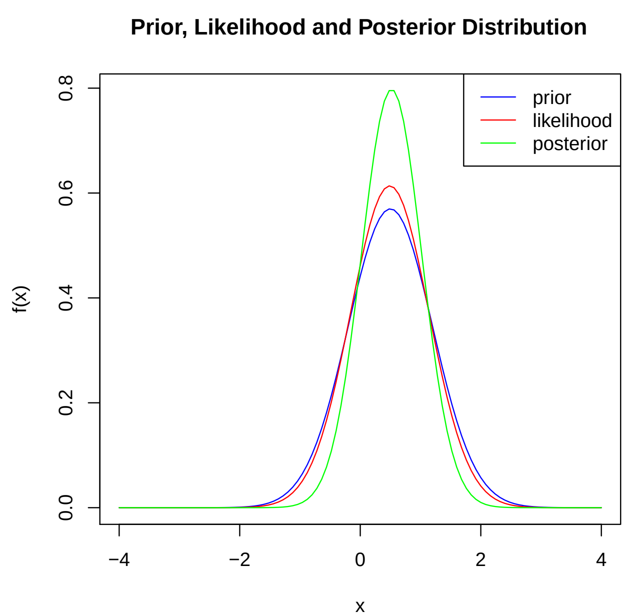

Figure 2. The triplot from two years' data.

2.3. Sequential Updating

In the last example we pooled the data from the two years and went back to the original prior distribution to use Bayes' theorem. We did not need to do this. A nice feature of Bayes' theorem is the possibility of updating sequentially, incorporating data as they arrive. In this case, consider the data to be just the new patients observed to a six months follow-up during the second year. These comprise 55 patients, of whom 28 had survived. The doctor could consider these as the data x with $n = 55$ and $r = 28$. What would the prior information be?
Clearly, the prior distribution should express her information prior to obtaining these new data, i.e. after the first years' data, so her prior for this second analysis is her posterior distribution form the first. This was $Be(28.28, 36.23)$. Combining this prior with the new data gives the same posterior $Be(28.28+28, 36.23+27) = Be(56.28, 63.23)$ as before. This simply confirms that we can get to the posterior distribution.

- In a single step, combining all the data with a prior distribution representing information available before any of the data were obtained.
- Sequentially, combining each item or block of new data with a prior distribution representing information available just before the new data were obtained (but after getting data previously received).

2.4. Normal Sample

Let $X_1, X_2, ..., X_n$ be from $\mathcal{N}(\mu, \sigma^2)$. $\theta = (\mu, \sigma^2) \rightarrow$ unknown parameters. The likelihood is:

$$\begin{aligned} f(x \mid \mu, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x_i - \mu)^2\right\} \\ &\propto \sigma^{-n} \exp\left[-\frac{1}{2\sigma^2}\left\{n(\bar{x}-\mu)^2 + S^2\right\}\right] \end{aligned} \tag{2.31}$$

where $S^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$

2.5. NIC distributions

For the prior distribution, we now need a joint distribution for μ and σ^2

Definition 2.32. The normal-inverse-chi-squared distribution(NIC) has density:

$$f(x \mid \mu, \sigma^2) \propto \sigma^{-(d+3)/2} \exp\left[-\frac{1}{2\sigma^2}\left\{v^{-1}(\mu-m)^2 + a\right\}\right] \tag{2.33}$$

where $a > 0$, $d > 0$ and $v > 0$.

The following facts are easy to derive about $NIC(m, v, a, d)$ distribution.

(a) The conditional distribution of μ given σ^2 is $\mathcal{N}(\mu, v\sigma^2)$ so $E(\mu \mid \sigma^2) = m$, $v(\mu \mid \sigma^2) = v\sigma^2$.
(b) The marginal distribution of σ^2 is such that $a\sigma^{-2} \sim \chi^2_d$. We say that σ^2 has the inverse-chi-square distribution $IC(a, d)$. We have $E(\sigma^2) = a/(d{-}2)$ if $d > 2$ and $v(\sigma^2) = 2a^2/\left\{(d-2)^2(d-4)\right\}$ if $d > 4$.
(c) The conditional distribution of σ^2 given μ is $IC(v^{-1}(\mu-m)^2 + a, d+1)$ and in particular $E(\sigma^2 \mid \mu) = (v^{-1}(\mu-m)^2+a)/(d-1)$ provided $d > 1$.
(d) The marginal distribution of μ is such that $(\mu-m)\sqrt{d}/\sqrt{av}\mu + d$. We say that μ has t-distribution $t_d(m, av/d)$. We have $E(\mu) = m$ if $d > 1$, and $v(\mu) = av/(d-2)$ if $d > 2$.

2.6. Posterior

Supposing then that the prior distribution is $NIC(m, v, a, d)$, we find

$$f(\mu, \sigma^2 \mid x) \propto \sigma^{d+n+3} \exp\left[-\frac{1}{2\sigma^2}\theta\right] \tag{2.34}$$

where $\theta = v^{-1}(\mu-m)^2 + a + n(\bar{x}-\mu) + s^2$ is a quadratic expression in μ. After completing the square, we see that $\mu, \sigma^2 \mid x \propto NIC(m^*, v^*, a^*, d^*)$ where $m^* = (v^{-1}m + n\bar{x})/(v^{-1}+n)$, $v^* = (v^{-1}+n)^{-1}$, $a^* = a + S^2 + (\bar{x}-m)^2/(n^{-1}+v)$, $d^* = d+n$. To interpret these results, note first that the posterior mean of μ is m^* which is a weighted average of the prior mean m and the usual data only-estimate $\bar{x}$ with weights v^{-1} and n.

The posterior mean of σ^2 is $a^*/(d^*-2)$ which is a weighted average of three terms: the prior mean $a/(d-2)$ with weight $(d-2)$, the usual data-only estimate $S^2/(n-1)$ with weight $(n-1)$ and $(\bar{x}-m)/(n^{-1}+v)$ with weight 1.

2.7. Weak prior

We clearly obtain weak prior information about μ by letting v go to infinity or $v^{-1} \to 0$. Then $m^* = \bar{x}$, $v^* = 1/n$, $a^* = a + S^2$, because the third term disappears.

To obtain weak prior information also about σ^2, if it is usual to set $a = 0$ and $d = 1$. Then $a^* = S^2$ and $d^* = n - 1$. The resulting inference match the standard frequentist results very closely with these parameters, since we have:

$$\frac{(\mu - \bar{x})\sqrt{n}}{S/\sqrt{n-1}} \propto t_{n-1}, \tag{2.35}$$

$$\frac{S^2}{\sigma^2} \propto \chi^2_{n-1} \tag{2.36}$$

Exactly the same distribution statements underlie standard frequentist inference in this problem.

LECTURE 3
Inference

Basic Statistics

R Codes

Listing 1. Triplot Code in R

```
###################################################
#                                              #
#      A Sample Triplot by Anil Aksu             #
#   It is developed to show some basics of R       #
#                                    #
###################################################

## the range of sampling
x=seq(-4,4,length=101)
## this function gets numbers from console
prior=dnorm(x, mean = 0.5, sd = 0.7, log = FALSE)
likelihood=dnorm(x, mean = 0.49, sd = 0.65, log = FALSE)
posterior=dnorm(x, mean = 0.52, sd = 0.5, log = FALSE)


## let's plot them
plot(range(x), range(c(likelihood,prior,posterior)), ...
    type='n', xlab="x", ylab="f(x)")
lines(x, prior, type='l', col='blue')
lines(x, likelihood, type='l', col='red')
lines(x, posterior, type='l', col='green')

title("Prior, Likelihood and Posterior Distribution")
legend(
  "topright",
  lty=c(1,1,1),
  col=c("blue", "red", "green"),
  legend = c("prior", "likelihood","posterior")
)
```

BIBLIOGRAPHY

1. Allen B. Dawney. *Think Bayes: Bayesian Statistics in Python.* O'REILLY, 2013.
2. Sheldon Ross. *Introduction to Probability Models.* Academic Press, Boston, 2014.