Bayes and conjugate forms

Saad Mneimneh

1 Conjugate forms

Recall that

$$f(y|x) = \frac{f(x|y)f(y)}{f(x)}$$

and since $f(x)$ is not a function of y, we can write:

$$f(y|x) \propto f(x|y)f(y)$$

In expressing $f(y|x)$ as above, we can ignore any multiplicative constant in $f(x|y)$ and $f(y)$ that does not involve y, as long as we impose the constraint that:

$$\int_{-\infty}^{\infty} f(y|x)dy = 1$$

The prior $f(y)$ is called conjugate if, when multiplied by $f(x|y)$, produces a posterior $f(y|x)$ with the "same form" as $f(y)$. In other words, the prior and the posterior belong to the same class of distributions. This is particularly interesting for two reasons:

- it simplifies the math (in particular the integral in the denominator)
- it produces a posterior of a similar nature to the prior, with updated parameters based on the observed data

For the larger part of treating the subject of conjugate priors, the Bayesian approach is introduced mearly as a framework that generalizes the standard statistical techniques. In most cases, the prior can be set in trivial (and sometimes unrealistic) ways to mimic the non-Bayesian approach. It remains a philosophical question whether the knowledge of a better prior is really accessible or not (e.g. if the prior is estimated from the data itself), but in some cases, it is conceivable that such a knowledge exists, e.g. biological data.

2 Gaussian prior

If $X|\mu \sim N(\mu, \sigma^2)$ then $\mu \sim N(\beta, \tau^2)$ is a conjugate prior.

$$f(\mu|x) \propto e^{-\frac{(x-\mu)^2}{2\sigma^2}} e^{-\frac{(\mu-\beta)^2}{2\tau^2}}$$

With a little bit of rearrangement of terms, we find that:

$$\mu|x \sim N(\frac{\sigma^2\beta + \tau^2 x}{\sigma^2 + \tau^2}, \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2})$$

If we have n independent observations $x_1 \dots x_n$, and each $X_i \sim N(\mu, \sigma^2)$, then define $\bar{x} = \sum_i x_i/n$. Note that a linear combination of Gaussian random variables is Gaussian; therefore, $\bar{X} \sim N(\mu, \sigma^2/n)$, and:

$$\mu|\bar{x} \sim N(\frac{\sigma^2\beta/n + \tau^2\bar{x}}{\sigma^2/n + \tau^2}, \frac{\sigma^2\tau^2/n}{\sigma^2/n + \tau^2})$$

We can also show, however, that $f(\mu|\bar{x}) = f(\mu|x_1, \dots x_n)$ (in this case $\bar{x}$ is called a sufficient statistic for μ with respect to x).

$$f(\mu|x_1, \dots x_n) \propto e^{-\frac{(\mu-\beta)^2}{2\tau^2}} \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

But

$$\prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \propto e^{-\frac{(\sum_i x_i/n-\mu)^2}{2\sigma^2/n}}$$

Therefore,

$$f(\mu|x_1, \dots x_n) \propto e^{-\frac{(\mu-\beta)^2}{2\tau^2}} e^{-\frac{(\bar{x}-\mu)^2}{2\sigma^2/n}}$$

3 Does prior matter?

In finding $f(\mu|\bar{x})$ one might argue that $\lim_{n\to\infty} \bar{x} = \mu$ (in some probabilistic notion of limit). For instance, the Chebychev inequality states:

$$P(|\bar{x} - \mu| \leq \epsilon) \geq 1 - \frac{\sigma^2}{n\epsilon^2}$$

So why even bother finding the posterior distribution of μ. The answer is that n may not be large enough. But even if it is, the Bayesian analysis is in accordance with the above limit. For example, when $n \to \infty$, our Gaussian posterior for μ (from above) becomes

$$\mu|\bar{x} \sim N(\bar{x}, \sigma^2/n)$$

This is generally true. A large amount of data overcomes any prior that is positive everywhere. For instance, fix an small ϵ and assume that we observe $\bar{x} = z$ and $|y - z| \gg \epsilon$:

$$f(\mu = z|\bar{x} = z) \propto f(\bar{x} = z|\mu = z)f(\mu = z)$$

$$f(\mu = y|\bar{x} = z) \propto f(\bar{x} = z|\mu = y)f(\mu = y)$$

Assume that the prior satisfies $f(\mu = z)/\max_y f(\mu = y) = c$ for some constant $c > 0$, then:

$$\frac{f(\mu = z|\bar{x} = z)}{f(\mu = y|\bar{x} = z)} \geq c\frac{f(\bar{x} = z|\mu = z)}{f(\bar{x} = z|\mu = y)}$$

Now for the small increment ϵ, $f(\bar{x} = z|\mu)2\epsilon \approx P(|\bar{x} - z| \leq \epsilon|\mu)$. Therefore,

$$\frac{f(\mu = z|\bar{x} = z)}{f(\mu = y|\bar{x} = z)} \geq c\frac{P(|\bar{x} - z| \leq \epsilon|\mu = z)}{P(|\bar{x} - z| \leq \epsilon|\mu = y)}$$

By Chebychev inequality, the numerator goes to 1 and the denominator goes to 0 when n goes to infinity. Therefore, if the data is large, given $\bar{x} = z$, the posterior $\mu = z$ is far more likely than any other value y away from z, regardless of the chosen prior.

4 Deciding on the mean of two groups

Assume we have two groups of statistically independent values $x_1 \dots x_{n_x}$, and $y_1, \dots y_{n_y}$, and that $X_i \sim N(\mu_x, \sigma_x^2)$ and $Y_j \sim N(\mu_y, \sigma_y^2)$. Assume further that μ_x and μ_y are unknown while σ_x^2 and σ_y^2 are known. We would like to decide whether $\mu_x = \mu_y$.

A classical approach is to compute the following statistic:

$$z = \frac{\bar{x} - \bar{y}}{\sqrt{\sigma_x^2/n_x + \sigma_y^2/n_y}}$$

where $\bar{x}$ and $\bar{y}$ represent the sample means, i.e. $\bar{x} = \sum_i x_i/n_x$ and $\bar{y} = \sum_j y_j/n_y$. Note that $Z \sim N(\frac{\mu_x - \mu_y}{\sqrt{\sigma_x^2/n_x + \sigma_y^2/n_y}}, 1)$. Therefore, $\mu_x = \mu_y$ iff $Z \sim N(0, 1)$.

Given the computed value for z, we then check how "extreme" it is, i.e. compute (the P-value) $P(Z > |z|) = 1 - P(Z \leq |z|) = \Phi(-|z|)$. Typically, a value of $\Phi(-|z|) \leq 0.05$ (arbitrary) suggests that z is extreme and, therefore, $\mu_x \neq \mu_y$. However, this approach is not very informative; for instance, if $\Phi(-|z|) > 0.05$, then we have no reason to reject that $\mu_x = \mu_y$, but we have not strong enough reason to believe it either.

Regardless of the pros and cons of such an approach, we will show that it represents a special case of the more general Bayesian approach. We will adopt a simplified model to illustrate this fact, so assume that μ_x and μ_y are independent and that:

$$\mu_x \sim \mu_y \sim N(\beta, \tau^2)$$

In this case,

$$f(\mu_x, \mu_y|x_1 \dots x_{n_x}, y_1 \dots y_{n_y}) \propto e^{-\frac{(\bar{x}-\mu_x)^2}{2\sigma_x^2/n_x}} e^{-\frac{(\bar{y}-\mu_y)^2}{2\sigma_x^2/n_y}} e^{-\frac{(\mu_x-\beta)^2}{2\tau^2}} e^{-\frac{(\mu_y-\beta)^2}{2\tau^2}}$$

We get,

$$\mu_x|\bar{x} \sim N(\frac{\sigma_x^2\beta/n_x + \tau^2\bar{x}}{\sigma_x^2/n_x + \tau^2}, \frac{\sigma_x^2\tau^2/n_x}{\sigma_x^2/n_x + \tau^2})$$

$$\mu_y|\bar{y} \sim N(\frac{\sigma_y^2\beta/n_y + \tau^2\bar{y}}{\sigma_y^2/n_y + \tau^2}, \frac{\sigma_y^2\tau^2/n_y}{\sigma_y^2/n_y + \tau^2})$$

For simplicity of illustration, assume that $n_x = n_y = n$ and $\sigma_x^2 = \sigma_y^2 = \sigma^2$, then:

$$\mu_x - \mu_y|\bar{x}, \bar{y} \sim N(\frac{\tau^2(\bar{x} - \bar{y})}{\sigma^2/n + \tau^2}, \frac{2\sigma^2\tau^2/n}{\sigma^2/n + \tau^2})$$

Denoting $\frac{\tau^2(\bar{x}-\bar{y})}{\sigma^2/n+\tau^2}$ by m and $\frac{2\sigma^2\tau^2/n}{\sigma^2/n+\tau^2}$ by v^2, then

$$1 - \Phi(\frac{b-m}{v}) = \Phi(\frac{m-b}{v})$$

gives $P(\mu_x - \mu_y > b|\bar{x}, \bar{y})$, which can be obtained for several values of b. This is now more informative as it gives the probability that μ_x will exceed μ_y by a certain amount. In particular, we can find $P(\mu_x > \mu_y|\bar{x}, \bar{y})$ by setting b to 0.

One may argue, however, that this was obtained by assuming a prior for μ_x and μ_y which, in principle, may be unknown. In that case, as Bayes himself argued, one could assume a uniform prior to indicate that no value is more likely than another, i.e. $f(\mu) = c$ (a constant). Such prior, of course, does not constitute a valid density function because $\int f(\mu)d\mu = \infty$. We call it "improper". We will justify this later. For now, if $f(\mu_x) \propto f(\mu_y) \propto 1$, we get:

$$f(\mu_x, \mu_y|x_1 \dots x_{n_x}, y_1 \dots y_{n_y}) \propto e^{-\frac{(\bar{x}-\mu_x)^2}{2\sigma_x^2/n}} e^{-\frac{(\bar{y}-\mu_y)^2}{2\sigma_x^2/m}}$$

which implies that:

$$\mu_x|\bar{x} \sim N(\bar{x}, \sigma_x^2/n_x)$$

$$\mu_y|\bar{y} \sim N(\bar{y}, \sigma_y^2/n_y)$$

$$\mu_x - \mu_y|\bar{x}, \bar{y} \sim N(\bar{x} - \bar{y}, \sigma_x^2/n_x + \sigma_y^2/n_y)$$

Therefore, by making $z(b) = \frac{b-(\bar{x}-\bar{y})}{\sqrt{\sigma_x^2/n_x+\sigma_y^2/n_y}}$, $1 - \Phi(z(b)) = \Phi(-z(b))$ gives $P(\mu_x - \mu_y > b|\bar{x}, \bar{y})$. In particular, $z = -z(0)$ and, therefore, when $b = 0$ we are looking at $\Phi(z)$.

$$\Phi(z) = \begin{cases} \Phi(-|z|) = \text{P-Value} & z \leq 0 \\ 1 - \Phi(-|z|) = 1 - \text{P-value} & z \geq 0 \end{cases}$$

Of particular interest is the second case ($z \geq 0$ i.e. $\bar{x} \geq \bar{y}$), which shows that a small P-value implies a higher probability that $\mu_x > \mu_y$ (thus leading to reject the hypothesis that $\mu_x = \mu_y$). This gives a better interpretation of the P-value.

We conclude that the P-value approach is a special case of a Bayesian approach with no prior information. Note, however, that the terminology/interpretation of the Bayesian approach is more precise, as the P-value is not the probability that $\mu_x = \mu_y$.

Now, how can we justify the use of an improper uniform prior? One way is to simply let $\tau^2 \to \infty$ in the Gaussian prior (note that this limiting procedure yields the same expression for the uniform prior). Another way is to observe that a uniform density $f(\mu) = c$ is valid over an interval of length $1/c$. Therefore, when $c \to 0$, the uniform prior covers the entire range from $-\infty$ to ∞. This suggests that if c is small enough, the improper density becomes an approximation of a valid uniform density. This is justified by the following result.

Consider n independent random samples $x_1 \ldots x_n$ taken from $N(\mu, \sigma^2)$ where σ^2 is known. Suppose that there exist positive constants α, ϵ, M and c, such that

$$(1-\epsilon)c \le f(\mu) \le (1+\epsilon)c \quad \mu \in [\bar{x} - \tfrac{\lambda\sigma}{\sqrt{n}}, \bar{x} + \tfrac{\lambda\sigma}{\sqrt{n}}]$$

$$f(\mu) \le Mc \qquad \text{otherwise}$$

where

$$2\Phi(-\lambda) = \alpha$$

Then for $\mu \in [\bar{x} - \frac{\lambda\sigma}{\sqrt{n}}, \bar{x} + \frac{\lambda\sigma}{\sqrt{n}}]$,

$$\frac{1-\epsilon}{(1+\epsilon)(1-\alpha) + M\alpha}\left[\frac{\sqrt{n}}{\sigma}\phi(\frac{\mu-\bar{x}}{\sigma/\sqrt{n}})\right] \le$$

$$f(\mu|x_1, \ldots x_n)$$

$$\le \frac{1+\epsilon}{(1-\epsilon)(1-\alpha)}\left[\frac{\sqrt{n}}{\sigma}\phi(\frac{\mu-\bar{x}}{\sigma/\sqrt{n}})\right]$$

The proof is straight forward using Bayes' rule, but let us consider a numerical example before proving it. Suppose $\lambda \approx 2$ so that $\alpha = 0.05$. Consider the values of μ within the interval extending $2\sigma/\sqrt{n}$ either side of $\bar{x}$. It may be judged that prior to taking the sample, no one value of μ in the interval was more probable than any other so that $f(\mu) = c$ within the interval, and we can put $\epsilon = 0$. Consider the values of μ outside the interval. It may be judged that prior to taking the sample, no value of μ there is more than twice as probable as any value of μ in the interval, i.e. $M = 2$. Then the posterior density for μ lies between $(1.05)^{-1}$ and $(0.95)^{-1}$ of the normal density within the interval.

The proof is as follows, let I be the interval $[\bar{x} - \frac{\lambda\sigma}{\sqrt{n}}, \bar{x} + \frac{\lambda\sigma}{\sqrt{n}}]$:

$$f(\mu|x_1 \ldots x_n) = \frac{\frac{1}{\sqrt{2\pi/n}\sigma} e^{-\frac{(\mu-\bar{x})^2}{2\sigma^2/n}} f(\mu)}{\int_{\mu\in I} \frac{1}{\sqrt{2\pi/n}\sigma} e^{-\frac{(\mu-\bar{x})^2}{2\sigma^2/n}} f(\mu)d\mu + \int_{\mu\notin I} \frac{1}{\sqrt{2\pi/n}\sigma} e^{-\frac{(\mu-\bar{x})^2}{2\sigma^2/n}} f(\mu)d\mu}$$

$$= \frac{\frac{1}{\sqrt{2\pi/n}\sigma} e^{-\frac{(\mu-\bar{x})^2}{2\sigma^2/n}} f(\mu)}{\int_{z\in[-\lambda,\lambda]} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} f(\mu) dz + \int_{z\notin[-\lambda,\lambda]} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} f(\mu) dz}$$

Now note that $(1-\epsilon)c \leq f(\mu) \leq (1+\epsilon)c$ when $z \in [-\lambda, \lambda]$ and $f(\mu) \leq Mc$ when $z \notin [-\lambda, \lambda]$, and that $\int_{-\lambda}^{\lambda} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = 1 - \alpha$ by definition of λ. Combining these facts yields the result.

An interpretation of this result is as follows: As long as the prior for μ is uniform in a large enough interval centered around $\bar{x}$, the posterior for μ is almost normal. This type of reasoning involving an "improper" uniform prior can, of course, be applied for densities other than normal.

5 Lindley's paradox

Lindley's paradox arises when we combine knowledge and lack of knowledge in a prior. This is usually done by using a prior with a mixture of discrete mass and continuous density. For instance, suppose we want to decide on the mean μ of a sampe $x_1 \ldots x_n$, and that prior opinion about μ is a mixture of a point mass p at a specific value μ_0 ($P(\mu = \mu_0) = p$), and the remaining probability $1 - p$ distributed uniformly (lack of knowledge) over an interval of length L centered around μ_0. Therefore,

$$f(\mu) = p\delta(\mu - \mu_0) + (1-p)\frac{1}{L} \qquad \mu \in [\mu_0 - \frac{L}{2}, \mu_0 + \frac{L}{2}]$$

where $\delta(x)$ is the Dirac function, i.e. $\delta(x) = 0$ for $x \neq 0$, and

$$\int_a^b f(x)\delta(x)dx = \begin{cases} f(0) & 0 \in [a,b] \\ 0 & 0 \notin [a,b] \end{cases}$$

This is not really a function, but rather a shorthand for a limiting process. For instance, consider the function δ_n defined as $1/n$ is the interval $[-n/2, n/2]$. Think of the delta function as the limit of δ_n, as n goes to 0. Obviously, $\delta(x) = \lim_{n\to 0} \delta_n(x) = 0$ for every $x \neq 0$. Moreover,

$$\lim_{n\to 0} \int_a^b f(x)\delta_n(x)dx$$

If $0 \in [a, b]$, this is

$$= \lim_{n\to 0} \int_{-n/2}^{n/2} \frac{f(x)}{n} dx$$

$$= \lim_{n\to 0} \frac{1}{n} \int_{-n/2}^{n/2} f(x)dx$$

$$= \lim_{n\to 0} \frac{1}{n} f(nc) \int_{-n/2}^{n/2} dx = f(0)$$

for some $-1/2 < c < 1/2$ (by the mean value theorem).

Note that

$$P(\mu = \mu_0) = \int_{\mu_0}^{\mu_0} p\delta(\mu - \mu_0)d\mu + \int_{\mu_0}^{\mu_0} (1-p)\frac{1}{L}d\mu = p + 0 = p$$

Let us first consider Bayes' rule for a general mixture prior ($\alpha + \beta = 1$)

$$f(\theta) = \alpha g(\theta) + \beta h(\theta)$$

for which we are interested in computing the posterior $f(\theta|x)$.

$$f(\theta|x) = \frac{f(x|\theta)[\alpha g(\theta) + \beta h(\theta)]}{f(x)} = \frac{\alpha f(x|\theta)g(\theta)}{f(x)} + \frac{\beta f(x|\theta)h(\theta)}{f(x)}$$

Note that

$$f(x|\theta)g(\theta) = g(\theta|x)g(x) = g(\theta|x)\int f(x|\theta)g(\theta)d\theta$$

Similarly,

$$f(x|\theta)h(\theta) = h(\theta|x)h(x) = h(\theta|x)\int f(x|\theta)h(\theta)d\theta$$

Therefore,

$$f(\theta|x) = \frac{\alpha g(\theta|x)\int f(x|\theta)g(\theta)d\theta}{f(x)} + \frac{\beta h(\theta|x)\int f(x|\theta)h(\theta)d\theta}{f(x)} = \alpha(x)g(\theta|x) + \beta(x)h(\theta|x)$$

where

$$\frac{\alpha(x)}{\beta(x)} = \frac{\alpha\int f(x|\theta)g(\theta)d\theta}{\beta\int f(x|\theta)h(\theta)d\theta}$$

and

$$\alpha(x) + \beta(x) = 1$$

Let us apply this to our particular case for $\bar{x}$ and μ, we have:

$$\alpha = p$$
$$\beta = 1 - p$$
$$f(\bar{x}|\mu) = \frac{1}{\sqrt{2\pi/n}\sigma}e^{-\frac{(\bar{x}-\mu)^2}{2\sigma^2/n}}$$
$$g(\mu) = \delta(\mu - \mu_0)$$
$$h(\mu) = \frac{1}{L}$$

Let us compute $\alpha(\bar{x})$.

$$\frac{\alpha(\bar{x})}{1-\alpha(\bar{x})} = \frac{p\int_{\mu_0-L/2}^{\mu_0+L/2} f(\bar{x}|\mu)\delta(\mu-\mu_0)d\mu}{(1-p)\int_{\mu_0-L/2}^{\mu_0+L/2} f(\bar{x}|\mu)\frac{1}{L}d\mu} \geq \frac{pf(\bar{x}|\mu_0)}{(1-p)\frac{1}{L}\int_{-\infty}^{\infty} f(\bar{x}|\mu)d\mu} = \frac{pf(\bar{x}|\mu_0)}{(1-p)\frac{1}{L}}$$

Assume that we are observing $\bar{x} = \mu_0 + c\sigma/\sqrt{n}$ for some constant c, then:

$$\frac{\alpha(\bar{x})}{1-\alpha(\bar{x})} \geq \frac{Lpe^{-c^2/2}}{(1-p)\sqrt{2\pi/n}\sigma}$$

Note that the limit as n goes to infinity of the right hand side is infinity. Therefore, $\alpha(\bar{x}) \to 1$ for $\bar{x} = \mu_0 + c\sigma/\sqrt{n}$ as n goes to infinity, i.e. $f(\mu|\bar{x}) \to g(\mu|\bar{x})$. Now,

$$g(\mu|\bar{x}) = \frac{f(\bar{x}|\mu)g(\mu)}{\int f(\bar{x}|\mu)g(\mu)d\mu}$$

which is 0 when $\mu \neq \mu_0$ and $\delta(0)$ when $\mu = \mu_0$; therefore,

$$g(\mu|\bar{x}) = \delta(\mu - \mu_0)$$

This means that the posterior for μ is a point mass at $\mu = \mu_0$. In other words, given that we are observing $\bar{x} = \mu_0 + c\sigma/\sqrt{n}$, $P(\mu = \mu_0) \approx 1$. Note that this is true regardless of how small p is, and how large c is (as long as n is large enough). When c is large, however, a classical P-value approach will refute the hypothesis that $\mu = \mu_0$ since $\bar{x}$ is c standard deviations away from the mean. That's the paradox!