Single Parameter Models

2.1 - The Binomial Model

Consider a model for y as the number of successes out of n trials, with $P(\text{success}) = \theta$. In this case we have the following **likelihood**

$$p(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$$

We want to find a posterior distribution for θ. In order to calculate a posterior distribution we need two things:

1. **Likelihood**: $\implies p(y|\theta)$
2. **Prior Distribution**: $\implies p(\theta)$

We begin with models that have a binomial likelihood, as stated above. Our prior distribution describes θ, which in this case, is a probability and as such is defined on the range $[0, 1]$. The simplest prior distribution we can use is a *uniform prior*.

Uniform Prior :

In the case that we do not have any prior information about what we think θ may be. The uniform distribution reflects this by having all values of θ be equally likely $p(\theta) = 1$. This is called an **uninformative prior**. Thus, we have all of the pieces we need to derive our first posterior distribution:

$$p(\theta|y) \propto p(y|\theta)p(\theta) \qquad\qquad p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

$$\begin{aligned} p(y) = &\int p(y,\theta)d\theta = \int p(y|\theta)p(\theta)d\theta = \int \binom{n}{y}\theta^y(1-\theta)^{n-y} \times (1)\ d\theta \\ = &\binom{n}{y}\int \theta^y(1-\theta)^{n-y}\ d\theta = \binom{n}{y}\int \theta^{(y+1)-1}(1-\theta)^{(n-y+1)-1}\ d\theta \\ = &\binom{n}{y}\frac{\Gamma(y+1)\Gamma(n-y+1)}{\Gamma(n+2)} = p(y) \end{aligned}$$

$$\begin{aligned} p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{\binom{n}{y}\theta^y(1-\theta)^{n-y}}{\binom{n}{y}\frac{\Gamma(y+1)\Gamma(n-y+1)}{\Gamma(n+2)}} \\ &= \frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)}\theta^{y+1-1}(1-\theta)^{n-y+1-1} \\ &= \mathbf{Beta(y+1,\ n\text{-}y+1))} \end{aligned}$$

Thus it is clear that a binomial likelihood with a uniform prior yields a beta posterior distribution.

Binomial Model with Uniform Prior	**Prior:** $\theta \sim \text{Beta}(\theta|\ 1,\ 1)$ **Likelihood:** $y|\theta \sim \text{Binomial}(n, \theta)$ **Posterior:** $p(\theta|y) = \text{Beta}(\alpha + 1,\ n - y + 1)$

Informative Priors :

The posterior distribution resulting from a uniform prior is a beta distribution. Conveniently enough, the uniform distribution is just a special case of the beta distribution when $\alpha = 1$ & $\beta = 1$

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

if α =1& $\beta = 1$:

$$p(\theta|1,1) = \frac{\Gamma(2)}{\Gamma(1)\Gamma(1)}\theta^0(1-\theta)^0 = 1$$

$\Gamma(n)$ is the **Gamma Function**. If n is a positive integer then $\Gamma(n) = (n-1)!$. Otherwise, for any positive value of n we get $\Gamma(n) = \int_0^\infty x^{n-1}e^{-x}dx$

Above, we used a uninformative prior, but since the uniform distribution is just a special case of the beta distribution, if we want an **informative prior** we can simply use a different beta distribution! This is due to the fact that the beta distribution is defined on $[0,1]$ which lines up with the domain of θ, since θ is a proportion.

Binomial Model with Beta Prior	**Prior:** $\theta \sim \text{Beta}(\theta|\ \alpha,\ \beta)$ **Likelihood:** $y|\theta \sim \text{Binomial}(n,\theta)$ **Posterior:** $p(\theta|y) = \text{Beta}(\alpha+y,\ \beta+n-y)$

Conjugacy :

For the binomial model with a beta prior, the posterior distribution itself will *always* be a beta distribution. When the prior and posterior both follow the same distribution we say there is **conjugacy**. In particular, a prior is **conjugate** when the posterior distribution follows the same parametric form as the prior. This occurs when the prior follows the same functional form as the likelihood. This is similar to the idempotent from linear algebra.

In other words, multiplying a distribution by a conjugate prior produces a distribution with the same kernel as the conjugate distribution. The **kernel** of a distribution is a reduced version of its probability density function which drops any terms that do not contain the variable(s) of interest (i.e. with the pdf with no normalizing constants). Conjugate priors make the process of calculating the posterior distribution much easier.

For the binomial likelihood with a beta prior, determine the posterior mean. How does this posterior mean relate to the prior mean the mean from the likelihood?

$$\theta|y \sim \text{Beta}(y+\alpha,\ n-y+\beta) \ \Rightarrow \mathbb{E}[\text{Beta}(13,9)] = \frac{\alpha}{\alpha+\beta} = \frac{13}{22} = 0.59$$

$$\begin{aligned}
\mathbb{E}[\theta|y] &= \text{“}\frac{\alpha}{\alpha+\beta}\text{”} = \frac{y+\alpha}{\cancel{y}+\alpha+n-\cancel{y}+\beta} = \frac{y+\alpha}{n+\alpha+\beta} = \frac{y}{n+\alpha+\beta} + \frac{\alpha}{n+\alpha+\beta} \\
&= \frac{n}{n+\alpha+\beta}\frac{y}{n} + \frac{\alpha+\beta}{n+\alpha+\beta}\frac{\alpha}{\alpha+\beta} \\
&= \left(\frac{n}{n+\alpha+\beta}\right)(\text{data mean}) + \left(\frac{\alpha+\beta}{n+\alpha+\beta}\right)(\text{prior mean})
\end{aligned}$$

From this, we can loosely interpret $\alpha - 1$ as the prior number of successes, $\beta - 1$ as the prior number of failures, and a total prior sample size of $\alpha + \beta - 2$.

Summarizing Posterior Inference :

We can summarize the procedure we have established as follows:

1. Look at data to determine likelihood
2. Determine appropriate conjugate prior
3. Calculate posterior distribution
4. Summarize posterior distribution
   → mean
   → credible interval
   → plot distributions

Posterior Predictive Distribution :

Now that we have determined the posterior distribution, we are interested in making predictions about future values of y. Which means we need to calculate the **posterior predictive distribution**, $p(\tilde{y}|y)$. Let's compute the probability that a future observation will be a success:

$m =$ new sample size $\qquad n =$ original sample size

$\tilde{y} =$ new number of successes $\qquad y =$ original number of successes

$$P(\tilde{y}=1|y) = \int P(\tilde{y}=1,\theta|y)d\theta = \int P(\tilde{y}=1|\theta,y)P(\theta|y)d\theta$$

$$= \int_0^1 P(\tilde{y}=1|\theta)P(\theta|y)d\theta \qquad \Longrightarrow \quad P(\tilde{y}=1|\theta) = \text{Binomial}(1,\theta) \implies \text{Likelihood}$$

$$P(\theta|y) = \text{Beta}(y+\alpha,\ n-y+\beta) \implies \text{Posterior Distribution}$$

$$= \int_0^1 \cancel{\binom{1}{1}}\theta^1\cancel{(1-\theta)^{1-1}}\frac{\Gamma(n+\alpha+\beta)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)}\theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}d\theta$$

$$= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)}\int_0^1 \theta^{y+\alpha+1-1}(1-\theta)^{n-y+\beta-1}d\theta$$

$$= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)}\frac{\Gamma(y+\alpha+1)\Gamma(n-y+\beta)}{\Gamma(n+\alpha+\beta+1)}\underbrace{\int_0^1 \frac{\Gamma(n+\alpha+\beta+1)}{\Gamma(y+\alpha+1)\Gamma(n-y+\beta)}\theta^{(y+\alpha+1)-1}(1-\theta)^{(n-y+\beta)-1}d\theta}_{\text{Beta density function, so it must integrate to 1}}$$

$$= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(y+\alpha)\cancel{\Gamma(n-y+\beta)}}\frac{\Gamma(y+\alpha+1)\cancel{\Gamma(n-y+\beta)}}{\Gamma(n+\alpha+\beta+1)}$$

$\downarrow$ **Recall:** $\Gamma(c) = (c-1)!$

$$= \frac{(n+\alpha+\beta-1)!}{(y+\alpha-1)!}\frac{(y+\alpha)!}{(n+\alpha+\beta)!} = \frac{\cancel{(n+\alpha+\beta-1)!}}{\cancel{(y+\alpha-1)!}}\frac{(y+\alpha)\cancel{(y+\alpha-1)!}}{(n+\alpha+\beta)\cancel{(n+\alpha+\beta-1)!}}$$

$$= \frac{y+\alpha}{n+\alpha+\beta} = \frac{\text{Total Successes}}{\text{Total Sample Size}} = \mathbb{E}[\theta|y]$$

$$\tilde{y} \sim \text{Binomial}(m=1,\ \theta) \iff \text{Bernoulli}(\theta)$$

$$p(\tilde{y}=1|y) \overset{\text{Property of the Bernoulli distribution}}{=} \mathbb{E}(\tilde{y}|y)$$

$$\overset{\text{Law of Total Expectation}}{=} \mathbb{E}\left[\mathbb{E}(\tilde{y}|\theta,y)|y\right]$$

$$= \mathbb{E}\left[\mathbb{E}(\tilde{y}|\theta)|y\right]$$

$$= \mathbb{E}(\theta|y) = \frac{y+\alpha}{n+\alpha+\beta}$$

Example :

In 2024, the top spot on the Billboard Hot 100 chart was held by female lead artists 12 times (with the remaining 40 weeks being held by male artists). Overall, we are interested in determining the probability that a female artist holds the top spot on the Bill -board Hot 100. Is a binomial model appropriate for this scenario?

⇒ **No!** Since the weeks are not independent with one another. Which means the probability of success changes with each week, and as such these observations are *not* exchangeable!

290 Rochesterians received an Instagram advertisement for Restaurant Good Luck. 44 of the 290 people clicked on the ad. We are interested in determining the proportion of people who will clock on the Good Luck link. Is a Binomial model, as we have discussed, appropriate for this scenario?

⇒ **Yes!** Since each observation is binary, independent of the other observations, and has the same probability of success, the binomial model is appropriate for our analysis.

Let θ be the proportion of people who click on the Good Luck ad. *Assuming a uniform prior distribution for* θ.

What is the posterior distribution for θ?

$$y|\theta \sim \text{Binomial}(n = 290,\ \theta) \qquad \theta = P(\text{clicking on an ad})$$
$$\theta \sim \text{Beta}(1, 1)$$
$$\theta|y \sim \text{Beta}(y + \alpha,\ n - y + \beta) = \text{Beta}(44 + 1,\ 290 - 44 + 1)$$
$$= \boxed{\text{Beta}(45,\ 247)}$$

What is the posterior mean and standard deviation of θ?

$$\mathbb{E}[\theta|y] = \text{“}\frac{\alpha}{\alpha+\beta}\text{”} = \frac{45}{45+247} = \frac{45}{292} \approx 0.1541$$
$$SD[\theta|y] = \sqrt{\mathbb{V}(\theta|y)} = \sqrt{\text{“}\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}\text{”}} = \sqrt{\frac{(45)(247)}{(45+247)^2(45+247+1)}} \approx 0.0211$$

Find a 95% posterior interval for θ using normal approximation.

$$\mathbb{E}[\theta|y] \pm 1.96 \times \sqrt{\mathbb{V}[\theta|y]}$$
$$0.1541 \pm 1.96(0.0211) = (0.1127,\ 0.1955)$$

Find a 95% posterior interval for θ using the quantiles.

$$\textbf{qbeta}(c(0.025, 0.975),\ 45,\ 247) = (0.1151,\ 0.1976)$$

“We are 95% confident that the true value of θ is between 0.1151 and 0.1955.”

Suppose that you see 10 of your friends have Good Luck ads, and you expect about 20% of them to click on it. Use this to get an informative prior distribution.

Three Methods: 20%clicked $\Rightarrow$ expected 2 successes and 8 failures

$$\underbrace{\text{Beta}(2, 8)\begin{cases}\alpha = 2\\ \beta = 8\end{cases}}_{\bigstar \text{ This method is preferable } \bigstar} \qquad \text{Beta}(3, 9)\begin{cases}\alpha - 1 = 2 \Rightarrow \alpha = 3\\ \beta - 1 = 8 \Rightarrow \beta = 9\end{cases} \qquad \text{Beta}(2.4, 9.6)\begin{cases}0.2 = \frac{\alpha}{12} \Rightarrow \alpha = 2.4\\ \mathbb{E}[\theta] = \frac{2.4}{2.4+\beta} \Rightarrow \beta = 9.6\end{cases}$$

If these three priors produce drastically different posterior that may be indicative that you need to collect more data.

Using the uniform prior distribution, what would be predict for the probability of a future Instagram user clicking on the Good Luck ad?

$$p(\theta|y) = \text{Beta}(45,\ 247) \quad m = 1$$
$$p(\tilde{y} = 1|y) = \mathbb{E}[\theta|y] = \frac{45}{292} \approx 0.1541$$

Suppose from a previous study, 3 out of 12 people click on the ad. How can we use this information to obtain an informative prior?

$$\left.\begin{array}{rl}\alpha - 1 \Rightarrow \text{prior number of successes} & \alpha - 1 = 3 \Rightarrow \alpha = 4\\ \beta - 1 \Rightarrow \text{prior number of failures} & \beta - 1 = 9 \Rightarrow \beta = 10\end{array}\right\} \Rightarrow \text{Beta}(4, 10)$$

2.2 Exponential Family Distributions

The binomial distribution is an example of an **exponential family** distribution. Generally, exponential family distributions take the following form:

$$p(y_i|\theta) = f(y_i)g(\theta)e^{\phi(\theta)^T u(y_i)}$$

$\phi(\theta)$ is called the **natural parameter**. $t(y) = \sum_{i=1}^{n} u(y_i)$ is called the **sufficient statistic** for θ, meaning that it contains all the information we need to make inference about θ. Generally, exponential family distributions are the only ones to have natural conjugate priors.

2.3 Poisson Model

Suppose we have some measurement that has whole number values, such as number of siblings, or number of courses taken during college. This data could be model by a Poisson distribution.

$$p(y|\theta) = \frac{\theta^y e^{-\theta}}{y!}$$

If we have n measurements, all of which are Poisson distributed, we can calculate the joint distribution as follows:

$$\begin{aligned} p(y_1, ..., y_n|\theta) &= \prod_{i=1}^{n} \frac{1}{y_i!}\theta^{y_i} e^{-\theta} \\ &\propto \theta^{n\bar{y}} e^{-n\theta} \end{aligned}$$

For this Poisson setup, $\sum_i y_i$ is a sufficient statistic for θ, and, $\sum_i y_i|\theta \sim \text{Poisson}(n\theta)$. What form must the prior have to be conjugate?

$$p(y|\theta) \propto \left(e^{-\theta}\right)^n \theta^{t(y)\cdot \ln(\theta)}$$

If we have $n = 1$ measurement, show that the posterior distribution is $\theta|y \sim \text{Gamma}(\alpha + y,\ \beta + 1)$.

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{\frac{1}{y!}\theta^y e^{-\theta}\frac{1}{\Gamma(\alpha)}\beta^\alpha\theta^{\alpha-1}e^{-\beta\theta}}{\frac{1}{y!}\beta^\alpha\int\theta^{y+\alpha-1}e^{-(1+\beta)\theta}d\theta}$$

$$= \frac{\theta^{y+\alpha-1}e^{-\theta(1+\beta)} \times \frac{(\beta+1)^y}{\Gamma(y+\alpha)}}{\int\theta^{y+\alpha-1}e^{-(1+\beta)\theta}d\theta \times \frac{(\beta+1)^y}{\Gamma(y+\alpha)}} = \frac{(\beta+1)^y}{\Gamma(y+\alpha)}\theta^{y+\alpha-1}e^{-\theta(1+\beta)}$$

$$= \text{Gamma}(y+\alpha,\ \beta+1)$$

Let's now consider the more general case of n observations. In this case, and using a $\text{Gamma}(\alpha + y,\ \beta)$ prior, we have the following posterior distribution:

$$p(\theta|y_1, \ldots, y_n) = \frac{p(y_1, \ldots, y_n|\theta)p(\theta)}{p(y_1, \ldots, y_n)} = \frac{\left(\prod_{i=1}^n \frac{1}{y_i!}\theta^{y_i}e^{-\theta}\right)\left(\frac{1}{\Gamma(\alpha)}\beta^\alpha\theta^{\alpha-1}e^{-\beta\theta}\right)}{\frac{1}{\Gamma(\alpha)\prod_{i=1}^n y_i!}\beta^\alpha\int\theta^{\sum_{i=1}^n y_i+\alpha-1}e^{-(n+\beta)\theta}d\theta}$$

$$= \frac{\cancel{\left(\prod_{i=1}^n \frac{1}{y_i!}\theta^{y_i}e^{-\theta}\right)}\cancel{\left(\frac{1}{\Gamma(\alpha)}\beta^\alpha\theta^{\alpha-1}e^{-\beta\theta}\right)}}{\cancel{\prod_{i=1}^n \frac{1}{y_i!}\frac{1}{\Gamma(\alpha)}\beta^\alpha}\int\theta^{\sum_{i=1}^n y_i+\alpha-1}e^{-(n+\beta)\theta}d\theta} = \underbrace{\frac{\theta^{\sum_{i=1}^n y_i+\alpha-1}e^{-\theta(n+\beta)} \times \frac{(\beta+n)^{n\bar{y}}}{\Gamma(n\bar{y}+\alpha)}}{\int\theta^{\sum_{i=1}^n y_i+\alpha-1}e^{-(n+\beta)\theta}d\theta \times \frac{(\beta+n)^{n\bar{y}}}{\Gamma(n\bar{y}+\alpha)}}}_{\text{Gamma pdf; must integrate to one}}$$

$$= \frac{(\beta+n)^{n\bar{y}+\alpha}\theta^{n\bar{y}+\alpha-1}e^{-n\theta-\beta\theta}}{\Gamma(n\bar{y}+\alpha)} = \frac{(\beta+n)^{n\bar{y}+\alpha}}{\Gamma(n\bar{y}+\alpha)}\theta^{n\bar{y}+\alpha-1}e^{-\theta(n+\beta)}$$

$$= \mathbf{Gamma}(\boldsymbol{n\bar{y}+\alpha,\ \beta+n})$$

What is the posterior expected value of θ?

Recall: If $X \sim \text{Gamma}(\alpha,\ \beta)$

$$\mathbb{E}[X] = \frac{\alpha}{\beta} \qquad \Longrightarrow \qquad \mathbb{E}[\theta|y] = \frac{n\bar{y}+\alpha}{\beta+n}$$

β can be interpreted as the prior number of observations. $\alpha - 1$ can be interpreted as the the sum of counts from β prior observations.

Poisson Model with Gamma Prior	**Prior:** $\theta \sim \text{Gamma}(\alpha,\ \beta)$ **Likelihood:** $y|\theta \sim \text{Poisson}(\theta)$ **Posterior:** $p(\theta|y) = \text{Gamma}(\alpha + n\bar{y},\ \beta + n)$

If we are interested in making predictions based on this model, we then would need to calculate the posterior predictive distribution.

$$p(\tilde{y}=1|y) = \int_0^\infty p(\tilde{y}=1,\theta|y)d\theta = \int_0^\infty \underbrace{p(\tilde{y}=1|\theta,\ y)}_{\text{"likelihood"}}\underbrace{p(\theta|y)}_{\text{"prior"}}d\theta \quad \leftarrow \text{of the posterior predictive distribution}$$

$$= \int_0^\infty \frac{\theta^y e^{-\theta}}{y!}\frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}d\theta = \frac{\beta^\alpha}{y!\Gamma(\alpha)}\int_0^\infty \underbrace{\theta^{y+\alpha-1}e^{-(\beta+1)\theta}}_{\substack{\text{Kernel of Gamma}\\ \text{Distribution}}}d\theta$$

$$= \frac{\beta^\alpha}{y!\Gamma(\alpha)}\frac{\Gamma(\alpha+y)}{(\beta+1)^{\alpha+y}}\underbrace{\int_0^\infty \frac{(\beta+1)^{y+\alpha}}{\Gamma(\alpha+y)}\theta^{y+\alpha-1}e^{-(\beta+1)\theta}d\theta}_{\substack{\text{pdf of Gamma}(\alpha+y,\ \beta+1);\\ \text{Must Integrate to 1}}} = \frac{\beta^\alpha}{y!\Gamma(\alpha)}\frac{\Gamma(\alpha+y)}{(\beta+1)^{\alpha+y}}$$

$$= \binom{y+\alpha-1}{y}\left(\frac{\beta}{\beta+1}\right)^\alpha\left(\frac{1}{\beta+1}\right)^y = \textbf{Negative Binomial}(\boldsymbol{\alpha},\ \boldsymbol{\beta})$$

This is a negative binomial distribution with parameters $(\alpha+\Sigma y_i,\ \beta+n)$.

The mean and variance of this distribution are:

$$\text{Negative Binomial}(\alpha,\ \beta):$$

$$\mathbb{E}[y] = \frac{\alpha}{\beta} \qquad\qquad \mathbb{V}[y] = \frac{\alpha(1+\beta)}{\beta^2}$$

Example : Suppose we want to model the number of people who use a Netflix account. You know that for your personal account, there are 5 people who use it to watch Netflix. You then conduct a study where you examine $n=22$ premium Netflix accounts. There are, on average $\bar{y}=4.5$ people using an account. We want to draw inference on the average number of people using an account. It is reasonable to assume that $y_i|\theta \sim \text{Poisson}(\theta)$.

What is an appropriate conjugate prior for θ?

$$\theta \sim \text{Gamma}(6,\ 1)$$
$$\alpha - 1 \Rightarrow \text{Prior number of events}$$
$$\beta \Rightarrow \text{Prior number of intervals}$$

What is the posterior distribution?

$$y_i \Rightarrow \text{number of people using a particular account}$$
$$\theta \Rightarrow \text{expected number of people using an account}$$

$$\theta|y \sim \text{Gamma}(\alpha+n\bar{y},\ \beta+n)$$
$$= \text{Gamma}(6+22(4.5),\ 1+22) = \text{Gamma}(105,\ 23)$$

What is the posterior mean?

$$\mathbb{E}[\theta|y] = \frac{n\bar{y}+\alpha}{\beta+n} = \frac{105}{23} \approx 4.57$$

Example : Suppose that causes of death are reviewed in detail for a city in the United States for a single year. It is found that 3 persons, out of a population of 200,000, died of asthma, giving a crude estimated asthma mortality rate in the city of

$$\frac{3}{200,000} \times 100,000 = 1.5 \text{ cases per } 100,000 \text{ persons per year.}$$

A Poisson sampling model is often used for epidemiological data of this form. The Poisson model derives from an assumption of exchangeability among all small intervals of exposure. Under the Poisson model, the sampling distribution of y, the number of deaths in a city of 200,000 in one year, may be expressed as $y \sim \text{Poisson}(2.0\theta)$, where θ represents the true underlying long-term asthma mortality rate in our city (measured in cases per 100,000 persons per year). In this notation, $y = 3$ is a single observation with exposure $x = 2.0$ (since θ is defined in units of 100,000 people) and unknown rate θ. We can use knowledge about asthma mortality rates around the world to construct a prior distribution for θ and then combine the datum $y = 3$ with that prior distribution to obtain a posterior distribution.

Setting up a Prior Distribution: Reviews of asthma mortality rates around the world suggest that mortality rates above 1.5 per 100,000 people are rare in Western countries, with typical asthma mortality rates around 0.6 per 100,000. Trial-and-error exploration of the properties of the gamma distribution, the conjugate prior family for this problem, reveals that a $\theta \sim \text{Gamma}(3.0, 5.0)$ density provides a plausible prior density for the asthma mortality rate in this example if we assume exchangeability between this city and other cities and this year and other years. The mean of this prior distribution is

$$\mathbb{E}[\theta] = \frac{\alpha}{\beta} = \frac{3.0}{5.0} = 0.6,$$

Furthermore, 97.5% of the mass of the density lies below 1.44. In practice, specifying a prior mean sets the ratio of the two gamma parameters, and then the shape parameter can be altered by trial and error to match the prior knowledge about the tail of the distribution.

Posterior Distribution: As shown above, the posterior distribution of θ for a $\text{Gamma}(\alpha, \beta)$ prior distribution is $\theta|y \sim \text{Gamma}(\alpha + y, \beta + x)$.

With the prior distribution and data described, the posterior distribution for θ is $\theta|y \sim \text{Gamma}(6.0, 7.0)$, which has mean

$$\mathbb{E}[\theta|y] = \frac{6.0}{7.0} = 0.86.$$

Substantial shrinkage has occurred toward the prior distribution. The posterior probability that the long-term death rate from asthma in our city is more than 1.0 per 100,000 per year, computed from the gamma posterior density, is $P(\theta > 1.0) \approx 0.30$.

Posterior Distribution with Additional Data: To consider the effect of additional data, suppose that ten years of data are obtained for the city in our example, instead of just one, and it is found that the mortality rate of 1.5 per 100,000 is maintained. That is, we observe

$$y = 30 \text{ deaths over 10 years.}$$

Assuming the population remains constant at 200,000, and assuming the outcomes in the ten years are independent with a constant long-term rate θ, the posterior distribution of θ is then

$$\theta|y \sim \text{Gamma}(3.0 + 30, 5.0 + 20) = \text{Gamma}(33.0, 25.0).$$

Figure 2.5b displays 1000 draws from this distribution. The posterior distribution is much more concentrated than before, and it still lies between the prior distribution and the data. After ten years of data, the posterior mean of θ is

$$\mathbb{E}[\theta|y] = \frac{33.0}{25.0} = 1.32.$$

The posterior probability that θ exceeds 1.0 is $P(\theta > 1.0) \approx 0.93$.

This analysis shows that with additional data, the posterior estimate becomes more concentrated, reflecting increased certainty in the inferred long-term mortality rate.

2.4 Normal Model

The normal distribution is perhaps the most widely used distribution, describing the general distributional shape for many situations. The normal distribution is used to model continuous random variables, and it has two parameters: the mean, μ, and variance σ^2.

For this distribution, we know that there is symmetry about μ, and the mean, median, and mode all equal μ. Additionally, 95% of the population lies within 1.96 standard deviations of the mean. With the normal distribution, there are nice properties regarding the combination of random variables.

In particular, if $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$ are independent, then $aX + bY$ follows a normal distribution with mean $a\mu_X + b\mu_Y$ and variance $a^2\sigma_X^2 + b^2\sigma_Y^2$. This property makes the normal distribution particularly useful in statistical modeling and inference.

Unknown Mean and Known Variance: Suppose that we want to draw some conclusion regarding the mean for a population. This was the primary goal of inference when conducting t-tests in frequentist statistics. When we have a normally distributed population and want to draw a conclusion about its mean, we can determine a Bayesian approach to alternatively meet this goal. We begin by assuming that $Y \sim N(\theta, \sigma^2)$, where σ^2 is known.

Likelihood: If we have one observation of y, then we can write the normal distribution as follows:

$$p(y|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(y-\theta)^2}{2\sigma^2}\right\} \propto \exp\left\{\frac{-(y-\theta)^2}{2\sigma^2}\right\}$$

Prior: What would be a conjugate prior for this normal likelihood?

$$\propto \exp\left\{\frac{-(\theta-\theta_0)^2}{2\tau^2}\right\} \quad \Longrightarrow \quad \theta \sim \mathcal{N}(\theta_0, \tau^2)$$

Considered as a function of θ, the likelihood is an exponential of a quadratic form in θ. Since we reaaaaaallllly want conjugacy this can be expressed as a normal distribution with mean μ_0 and variance τ_0^2. This is a normal distribution with mean μ_0 and variance τ_0^2.

As with all Bayesian approaches, we want to incorporate data with our prior information to get a better understanding of what we believe θ to be. Thus, our goal is to determine the posterior distribution for θ. Recall that:

$$p(\theta|y) \propto p(y|\theta)p(\theta). \qquad \Rightarrow \qquad p(\theta|y, \sigma^2) \propto p(y|\theta, \sigma^2)p(\theta).$$

Since we have two parameters in the normal model, with one of them (σ^2) being known.

Let's derive the posterior distribution:

$$p(\theta|y) \propto p(y|\theta)p(\theta) \quad \Longrightarrow \quad \mathcal{N}(\theta, \sigma^2) \times \mathcal{N}(\theta_0, \tau^2)$$

$$\propto \exp\left\{\frac{-(y-\theta)^2}{2\sigma^2}\right\} \times \exp\left\{\frac{-(\theta-\theta_0)^2}{2\tau^2}\right\} \propto \exp\left\{\frac{-(y-\theta)^2}{2\sigma^2} + \frac{-(\theta-\theta_0)^2}{2\tau^2}\right\}$$

$$\propto \exp\left\{\frac{-(y-\theta)^2\tau^2 - (\theta-\theta_0)^2\sigma^2}{2\sigma^2\tau^2}\right\} \quad \longleftarrow \quad \text{complete the squares}$$

$$\propto \exp\left\{\frac{\boxed{-y^2\tau^2} - \boxed{2\theta y\tau^2} + \boxed{\theta^2\tau^2} + \boxed{\theta^2\sigma^2} + \boxed{2\theta\theta_0\sigma^2} - \boxed{\theta_0^2\sigma^2}}{2\sigma^2\tau^2}\right\}$$

$$\propto \exp\left\{\frac{\boxed{\theta^2(\tau^2+\sigma^2)} - \boxed{2\theta(y\tau^2+\theta_0\sigma^2)} + \boxed{y^2\tau^2+\theta_0^2\sigma^2}}{2\sigma^2\tau^2}\right\}$$

$$\propto \exp\left\{\frac{-(\tau^2+\sigma^2)}{2\sigma^2\tau^2}\left(\theta^2 - 2\theta\left(\frac{y\tau^2+\theta_0\sigma^2}{\tau^2+\sigma^2}\right) + \cancel{\frac{y^2\tau^2+\theta_0^2\sigma^2}{\tau^2+\sigma^2}}\right)\right\}$$

since we are using proportionality, this can be treated as a constant & thus, it can be dropped from the kernel

$$\propto \exp\left\{\frac{-(\tau^2+\sigma^2)}{2\sigma^2\tau^2}\left(\theta - \left(\frac{y\tau^2+\theta_0\sigma^2}{\tau^2+\sigma^2}\right)\right)^2\right\} \quad \Rightarrow \quad P(\theta|y) \sim \mathcal{N}\left(\frac{\tau^2 y + \sigma^2\theta_0}{\tau^2+\sigma^2}, \frac{\sigma^2\tau^2}{\tau^2+\sigma^2}\right)$$

We will continue exploring this distribution in a homework assignment to better conceptualize the meaning of the mean and variance and to determine the posterior predictive distribution.

Normal Model with Normal Prior

Prior: $\theta \sim \mathcal{N}(\theta_0,\ \tau^2)$

Likelihood: $y|\theta \sim \mathcal{N}(\theta,\ \sigma^2)$

Posterior: $p(\theta|y) = \mathcal{N}\left(\frac{\frac{\theta_0}{\tau^2} + \frac{n\bar{y}}{\sigma^2}}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}},\ \frac{1}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}}\right)$

Known Mean and Unknown Variance: In some situations, we may know the mean of a distribution but not know what the variance is for it. This would be similar to conducting an F-test for equal variance in the frequentist realm. Ultimately, we are going to want a model where both the mean and variance are unknown, since that is realistic of situations we mainly encounter, but for now, we will consider the case of the mean known and variance unknown as a stepping stone to the main inferential question.

If we take a sample of n exchangeable observations from the normal distribution with known mean μ but unknown variance σ^2, the likelihood function is:

$$\begin{aligned} p(y_1,\ldots,y_n|\sigma^2) &= \prod_{i=1}^{n} p(y_i|\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_i-\mu)^2}{2\sigma^2}\right\} \\ &\propto (\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right\} \\ &= \left(\sigma^2\right)^{\frac{n}{2}} e^{\frac{-n\nu}{2\sigma^2}} \text{ where } \nu = \underbrace{\frac{1}{n}\sum_{i=1}^{n}(y_i-\mu)^2}_{\text{Constant}} \end{aligned}$$

The conjugate prior for this distribution is the inverse gamma distribution, which has two parameters: a shape parameter and a scale parameter. We can choose a prior for σ^2 as $\text{IG}(a,b)$. Here, a can be thought of as $1/2$ the approximate prior sample size, and b can be thought of as $1/2$ of the prior sum of squared residuals. The inverse gamma distribution is the distribution of the inverse of a gamma-distributed random variable.

$$p(\sigma^2) = \frac{\beta^\alpha}{\Gamma(\alpha)}(\sigma^2)^{-(\alpha+1)} e^{-\frac{\beta}{\sigma^2}}$$

The posterior distribution for σ^2 is then:

$$\begin{aligned} p(\sigma^2|y_1,\ldots,y_n) &\propto p(y_1,\ldots,y_n|\sigma^2)p(\sigma^2) \\ &\propto \underbrace{\left(\sigma^2\right)^{\frac{n}{2}} e^{\frac{-n\nu}{2\sigma^2}}}_{\text{Likelihood}} \times \underbrace{\left(\sigma^2\right)^{-(\alpha+1)} e^{-\frac{\beta}{\sigma^2}}}_{\text{Prior}} \\ &\propto \left(\sigma^2\right)^{-(\frac{n+2\alpha}{2}+1)} e^{-\left(\frac{n\nu+2\beta}{2\sigma^2}\right)} \\ &\propto \textbf{Inv-Gamma}\left(\frac{n+2\alpha}{2},\ \frac{n\nu+2\beta}{2}\right) \quad \text{where } \nu = \frac{1}{n}\sum_{i=1}^{n}(y_i-\mu)^2 \end{aligned}$$

Normal Model with Inverse-Gamma Prior

Prior: $\sigma^2 \sim \text{Inv-Gamma}(\alpha_0,\ \beta_0)$

Likelihood: $y|\sigma^2 \sim \mathcal{N}(\mu,\ \sigma^2)$

Posterior: $p(\sigma^2|y) \sim \text{Inv-Gamma}\left(\frac{n+2\alpha}{2},\ \frac{n\nu+2\beta}{2}\right)$
where $\nu = \frac{1}{n}\sum_{i=1}^{n}(y_i-\mu)^2$

Normal Model for Non-Normal Data

In statistics, we like to use normal distributions for many different situations, particularly when drawing inference on the mean of a distribution. Generally, because of the central limit theorem, we can say that regardless of the distribution of y, $\bar{y}$ will have a normal distribution with mean θ and variance $\frac{\sigma^2}{n}$ so long as the sample size is large enough.

This motivation for doing normal-based frequentists tests. We can similarly use sampling distributions as Bayesians to apply a normal model.So long as the sample size n is large enough, we can say that $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is normally distributed (by the central limit theorem). Thus, we can use the following likelihood:

$$p(\bar{y}|\theta) = \frac{1}{\sqrt{2\pi\frac{\sigma^2}{n}}}\exp\left\{-\frac{(\bar{y}-\theta)^2}{2\frac{\sigma^2}{n}}\right\}$$

From this, if we want to make inference for the mean θ, we can use a normal prior and get the following posterior:

$$p(\theta|\bar{y}) \sim \mathcal{N}\left(\frac{\frac{n\bar{y}}{\sigma^2}+\frac{\theta_0}{\tau^2}}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}, \frac{1}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}\right)$$

When it comes to drawing inference for the variance, things get a bit more complicated, since we do need the distribution of y itself to be normal in order to apply the methods as previously described.

2.5 Chosing Priors

We often want to use a prior that will play a minimal role in the posterior distribution. We colloquially call such priors "noninformative" or "uninformative."

In doing this, we are letting the data speak for themselves, which is particularly useful if we have no a priori information regarding θ.For the binomial model, we discussed the uninformative flat prior of Beta$(1, 1)$. With the Poisson model, an "uninformative prior" is usually chosen as Gamma$(0.001, 0.001)$. For the gamma distribution, we cannot create a completely flat prior, but by having both hyperparameters α and β be close to 0, their effect on the posterior is minimal.

Let's revisit the normal model. Suppose we have $y \sim \mathcal{N}(\theta, \sigma^2)$ for some known σ^2. A conjugate prior is $\theta \sim \mathcal{N}(\theta_0, \tau^2)$ What values of θ_0 and τ^2 are uninformative? If we use this uninformative prior, what effect does this have on the posterior distribution?

Let's now consider the case of $y \sim \mathcal{N}(\theta, \sigma^2)$ for known θ. In this case, we used a conjugate prior of $sigma^2 \sim$ Inv-Gamma$(\alpha_0,\ \beta_0)$ What values of α_0 and β_0 are uninformative?We can have the issue of an uninformative prior being improper. Recall that a prior is improper if it does not integrate to 1 (or, more general, to a constant). There isn't necessarily anything wrong with using an improper prior so long as the posterior is still proper. If an improper prior leads to an improper posterior, then we would not want to use that prior.

There are other methods for creating an uninformative prior. Perhaps the most standardized way is to use what we call the **Jeffreys' prior**. The Jeffreys prior is defined as $p(\theta) \propto \sqrt{\mathcal{I}(\theta)}$, where $\mathcal{I}(\theta)$ is the **Fisher information for** θ. The **Fisher information** is a measure of the sensitivity of a maximum likelihood estimator, defined as being the negative expectation of the second derivative of the log-likelihood, or:

$$\mathcal{I}(\theta) = -\mathbb{E}_{y|\theta}\left[\frac{\partial^2}{\partial\theta^2}\big(\ln p(y|\theta)\big)\right] \qquad \Rightarrow \qquad J(\theta) = \sqrt{\mathcal{I}(\theta)}$$

Considering the binomial model, what is the Jeffreys' prior?

$$
\begin{aligned}
&p(y|\theta) = \text{Binomial}(n,\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y} \\
&\mathcal{I}(\theta) = -\mathbb{E}_{y|\theta}\left[\frac{\partial^2}{\partial\theta^2}\big(\ln p(y|\theta)\big)\right] \\
&\ln p(y|\theta) = \ln\binom{n}{y} + y\ln\theta + (n-y)\ln(1-\theta) \\
&\frac{\partial}{\partial\theta}\Big[\ln p(y|\theta)\Big] = \frac{y}{\theta} - \frac{n-y}{1-\theta} \\
&\frac{\partial}{\partial\theta}\Big[\frac{y}{\theta} - \frac{n-y}{1-\theta}\Big] = -\frac{y}{\theta^2} - \frac{n-y}{(1-\theta)^2} \\
&-\mathbb{E}_{y|\theta}\left[-\frac{y}{\theta^2} - \frac{n-y}{(1-\theta)^2}\right] = \frac{n\theta}{\theta^2} + \frac{n-n\theta}{(1-\theta)^2} = \frac{n}{\theta} + \frac{n}{1-\theta} \\
&\mathcal{I}(\theta) = \frac{n}{\theta} + \frac{n}{1-\theta} = \frac{n}{\theta(1-\theta)} \\
&J(\theta) = \sqrt{\frac{n}{\theta(1-\theta)}} = \theta^{-n/2}(1-\theta)^{-n/2} \\
&= \theta^{(n/2)-1}(1-\theta)^{(n/2)-1} \propto \text{Beta}\left(\frac{n}{2},\frac{n}{2}\right).
\end{aligned}
$$

We thus have now have two different beta priors that are uninformative: Beta$(1,1)$ and Beta$(\frac{1}{2},\frac{1}{2})$ In both cases, we are saying that there are an equal number of success and failures a priori. A third option is to use the Beta$(0,0)$ prior, which is improper. Why is it improper?

$$
\begin{gathered}
\text{Beta}(0,0) \propto \theta^{-1}(1-\theta)^{-1} \\
\lim_{\theta\to 0}\left(\theta^{-1}\right) \to \infty \qquad \& \qquad \lim_{\theta\to 1}\left((1-\theta)^{-1}\right) \to \infty \\
\int_0^1 \theta^{-1}(1-\theta)^{-1}\,d\theta \neq 1.0
\end{gathered}
$$

While this is an reasonable choice to use, it may lead to the posterior distribution being improper if the observered sample does not contain at least one success and one failure. This is because the posterior distribution will be proportional to the product of the likelihood and the prior, and if the likelihood is 0, then the posterior will be 0. This is why we generally prefer to use the Beta$(1,1)$ or Beta$(\frac{1}{2},\frac{1}{2})$ priors.

What is the Jeffreys prior for the normal distribution with unknown variance?

$$
\begin{aligned}
&p(y|\theta) = \mathcal{N}(\mu,\theta) = \frac{1}{\sqrt{2\pi\theta}}\exp\left\{-\frac{(y-\mu)^2}{2\theta}\right\} \\
&\mathcal{I}(\theta) = -\mathbb{E}_{y|\theta}\left[\frac{\partial^2}{\partial\theta^2}\big(\ln p(y|\theta)\big)\right] \\
&\ln(p(y|\theta)) = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\theta) - \frac{(y-\mu)^2}{2\theta} \\
&\frac{\partial}{\partial\theta}\Big[\ln(p(y\mid\theta))\Big] = -\frac{1}{2\theta} + \frac{(y-\mu)^2}{2\theta^2} \\
&\frac{\partial}{\partial\theta}\Big[-\frac{1}{2\theta} + \frac{(y-\mu)^2}{2\theta^2}\Big] = \frac{1}{2\theta^2} - \frac{(y-\mu)^2}{\theta^3} \\
&-\mathbb{E}_{y|\theta}\left[\frac{1}{2\theta^2} - \frac{(y-\mu)^2}{\theta^3}\right] = \frac{1}{2\theta^2} - \frac{\mathbb{E}_{y|\theta}[(y-\mu)^2]}{\theta^3} \\
&\mathcal{I}(\theta) = \frac{1}{2\theta^2} - \frac{\theta}{\theta^3} = \frac{1}{2\theta^2} - \frac{1}{\theta^2} \\
&J(\theta) = \sqrt{\mathcal{I}(\theta)} = \sqrt{\frac{1}{2\theta^2} - \frac{1}{\theta^2}} = \frac{1}{\theta} = \frac{1}{\sigma^2} \propto \text{Inv-Gamma}(0,0)
\end{aligned}
$$

What is Jeffreys' prior for the normal distribution with unknown mean?

$$p(\bar{y}|\theta) = \mathcal{N}(\theta, \frac{\sigma^2}{n}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(\bar{y}-\theta)^2}{2\frac{n}{\sigma^2}}\right\}$$

$$\ln p(\bar{y}|\theta) = \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(\bar{y}-\theta)^2}{2\frac{n}{\sigma^2}}$$

$$\frac{\partial}{\partial\theta}\Big(\ln p(\bar{y}|\theta)\Big) = \frac{\bar{y}-\theta}{\frac{n}{\sigma^2}}$$

$$\frac{\partial}{\partial\theta}\Big(\frac{y-\theta}{\frac{n}{\sigma^2}}\Big) = -\frac{1}{\frac{n}{\sigma^2}} \quad \text{Therefore...}$$

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\Big(\ln p(y|\theta)\Big)\right] = -\mathbb{E}\left[-\frac{1}{\frac{n}{\sigma^2}}\right] = \frac{n}{\sigma^2} \quad \text{and}$$

$$\sqrt{I(\theta)} = J(\theta) = \sqrt{\frac{n}{\sigma^2}} \propto \frac{1}{\sqrt{\sigma^2}}$$

In some scenarios, we may want to have a "weakly informative prior." For a weakly informative prior, we can either start with an uninformative prior and add enough information so inference is reasonable. Or, you can start with an informative prior and broaden it to account for uncertainty in your prior beliefs.

Example: Suppose $y|\theta \sim \text{Binomial}(n, \theta)$, and $\theta \sim \text{Beta}(1, 1)$. Let $y_i = 1$ if subject i has a disease, and suppose that this disease is very rare (assume the prevalence is known 1 in 10,000).
If we collect $n = 100$ observations and $y = 0$ among these subjects, then the posterior mean will be $\frac{1}{102} \approx 0.01$. The "uninformative" prior is thus having a relatively large effect on the posterior. The better option would be to incorporate previous information to better decide on α and β in the beta prior.