GPH-GU2372/3372
Applied Bayesian Analysis in Public Health

Lecture 2: One-parameter Models

Hai Shu, PhD

09/19/2022

Binomial model: Topics

- Binomial model:
    - Conjugate prior and posterior distribution
    - Confidence region
    - Sensitivity to prior distribution
    - Prediction
    - Jeffreys' prior

Binomial model: Conjugate prior and posterior distribution

- Data: $Y_1, \ldots, Y_n$ are exchangeable binary random variables. For the uniform prior and the Bernoulli sampling model, we have obtained: $p(\theta|y_1, ..., y_n) = \text{Beta}(1 + \sum_{i=1}^n y_i, 1 + n - \sum_{i=1}^n y_i)$
- Now, let's suppose we want to compare the posterior odds of two values θ_a and θ_b of θ in light of the data: $y_1, ..., y_n$.
- We compute $\frac{p(\theta_a|y_1,...,y_n)}{p(\theta_b|y_1,...,y_n)}$ and we obtain:

$$\begin{aligned}
\frac{p(\theta_a|y_1, ..., y_n)}{p(\theta_b|y_1, ..., y_n)} &= \frac{\frac{\Gamma(n+2)}{\Gamma(1+\sum_{i=1}^n y_i)\Gamma(1+n-\sum_{i=1}^n y_i)}\theta_a^{\sum_{i=1}^n y_i}(1-\theta_a)^{n-\sum_{i=1}^n y_i}}{\frac{\Gamma(n+2)}{\Gamma(1+\sum_{i=1}^n y_i)\Gamma(1+n-\sum_{i=1}^n y_i)}\theta_b^{\sum_{i=1}^n y_i}(1-\theta_b)^{n-\sum_{i=1}^n y_i}} \\
&= \frac{\theta_a^{\sum_{i=1}^n y_i}(1-\theta_a)^{n-\sum_{i=1}^n y_i}}{\theta_b^{\sum_{i=1}^n y_i}(1-\theta_b)^{n-\sum_{i=1}^n y_i}} \\
&= \left(\frac{\theta_a}{\theta_b}\right)^{\sum_{i=1}^n y_i} \left(\frac{1-\theta_a}{1-\theta_b}\right)^{n-\sum_{i=1}^n y_i}
\end{aligned}$$

Binomial model: Conjugate prior and posterior distribution

- Since $\frac{p(\theta_a|y_1,\ldots,y_n)}{p(\theta_b|y_1,\ldots,y_n)} = \left(\frac{\theta_a}{\theta_b}\right)^{\sum_{i=1}^n y_i} \left(\frac{1-\theta_a}{1-\theta_b}\right)^{n-\sum_{i=1}^n y_i}$, the posterior odds depend on the data only through $\sum_{i=1}^n y_i$.
- This is true not only for the posterior odds, but for the posterior probability in general since $p(\theta|y_1,\ldots,y_n) = \text{Beta}(1+\sum_{i=1}^n y_i, 1+n-\sum_{i=1}^n y_i)$.
- We interpret this by saying that $\sum_{i=1}^n Y_i$ is a sufficient statistic.
- Definition: If $Y_1,\ldots,Y_n$ are random variables, a statistic $T(Y_1,\ldots,Y_n)$ is a sufficient statistic if it captures all the information about θ that is contained in the sample.
- If $Y_1,\ldots,Y_n$ are binary random variables that are conditionally independent and identically distributed with $p(y|\theta)$ a Bernoulli distribution with parameter θ, then $\sum_{i=1}^n Y_i$ is a Binomial random variable and its density is:

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

Binomial model: Conjugate prior and posterior distribution

- What would be the posterior distribution of θ if instead of working directly with the random variables $Y_1, ..., Y_n$, we work with the sufficient statistic $Y = \sum_{i=1}^{n} Y_i$?

- Now we have the following:
  - $p(y|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}, \qquad y \in \{0, 1, \dots, n\}$
  - $\theta \in \Theta = [0, 1]$ with prior distribution $p(\theta)$

- Let's use again a uniform prior $p(\theta) = 1$ for θ. Then:

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

$$\propto \theta^y (1-\theta)^{n-y}$$

Binomial model: Conjugate prior and posterior distribution

- $p(\theta|y) \propto \theta^y(1-\theta)^{n-y}$ for $\theta \in [0,1]$.
- We can determine again the proportionality constant by using the fact that $p(\theta|y)$ is a density and thus it needs to integrate 1.
- We can see that the part of $p(\theta|y)$ that depends on θ (i.e., the kernel of $p(\theta|y)$) has the same functional form as the kernel of a Beta distribution

$$\text{Beta}(x; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, \qquad x \in [0,1]$$

  where $a, b > 0$.

- Since the kernel of $p(\theta|y)$ matches that of a $\text{Beta}(a,b)$ with $a = 1+y$ and $b = 1+n-y$, the constant needed in $p(\theta|y)$ is $\frac{\Gamma(n+2)}{\Gamma(1+y)\Gamma(1+n-y)}$.

$$\left.\begin{array}{l} \theta \sim p(\theta) = 1, \theta \in [0,1] \\ Y|\theta \sim p(y|\theta) = \text{Binomial}(n;\theta) \end{array}\right\} \Longrightarrow p(\theta|y) = \text{Beta}(1+y, 1+n-y)$$

Binomial model: Conjugate prior and posterior distribution

- Notice that
  - $Y_1, ..., Y_n$ exchangeable binary random variables
  - $p(\theta) = 1$ with $\theta \in \Theta = [0, 1]$
  - $p(y_1, ..., y_n|\theta) = \prod_{i=1}^n \theta^{y_i}(1-\theta)^{1-y_i}$

  $\Longrightarrow p(\theta|y_1, ..., y_n) = \text{Beta}(1 + \sum_{i=1}^n y_i, 1 + n - \sum_{i=1}^n y_i)$

  while
  - $Y = \sum_{i=1}^n Y_i$
  - $p(\theta) = 1$ with $\theta \in \Theta = [0, 1]$
  - $p(y|\theta) = \text{Binomial}(n; \theta)$

  $\Longrightarrow p(\theta|y) = \text{Beta}(1 + y, 1 + n - y)$

- Since $y = \sum_{i=1}^n y_i$ and $1 + n - y = 1 + n - \sum_{i=1}^n y_i$, it follows that $p(\theta|y_1, ..., y_n) = p(\theta|y)$ confirming that $y = \sum_{i=1}^n y_i$ contains the same information about θ than the entire sample $\{y_1, ..., y_n\}$.

- If we want to derive the posterior distribution of a parameter θ, we can identify a sufficient statistic $T(Y_1, ..., Y_n)$ and derive $p(\theta|t)$ where $t = T(y_1, ..., y_n)$.

Binomial model: Conjugate prior and posterior distribution

- $Y_1, ..., Y_n$ exchangeable binary random variables and $Y = \sum_{i=1}^{n} Y_i$
- $p(\theta) = 1$ with $\theta \in \Theta = [0, 1]$
- $p(y|\theta) = \text{Binomial}(n; \theta)$

  $\Longrightarrow p(\theta|y) = \text{Beta}(1 + y, 1 + n - y)$

- If we set $a = 1$ and $b = 1$ in a $\text{Beta}(a, b)$ pdf, we obtain:

  uniform

$$\text{Beta}(x; 1, 1) = \frac{\Gamma(2)}{\Gamma(1)\Gamma(1)} x^0 (1 - x)^0 = 1, \qquad x \in [0, 1]$$

since if r is a positive integer, $\Gamma(r) = (r - 1)!$ and by definition $\Gamma(1) = 0! = 1$.
So, $\text{Uniform}[0, 1] = \text{Beta}(1, 1)$ distribution and we have:

$$\left.\begin{array}{l}\theta \sim p(\theta) = \text{Beta}(1, 1) \\ Y|\theta \sim p(y|\theta) = \text{Binomial}(n; \theta)\end{array}\right\} \Longrightarrow p(\theta|y) = \text{Beta}(1 + y, 1 + n - y)$$

Binomial model: Conjugate prior and posterior distribution

- Does the results shown above hold in general?
  I.e., does $p(\theta) = \text{Beta}(a, b)$ and $p(y|\theta) = \text{Binomial}(n; \theta)$ imply that $p(\theta|y) = \text{Beta}(\tilde{a}, \tilde{b})$ for some $\tilde{a}$ and $\tilde{b}$?

- If $p(\theta) = \text{Beta}(a, b)$,then:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{\left\{\binom{n}{y}\theta^y(1-\theta)^{n-y}\right\}\left\{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}\right\}}{p(y)}$$

$$\propto \theta^{a+y-1}(1-\theta)^{b+n-y-1}$$

$\Longrightarrow$ the kernel of $\text{Beta}(\tilde{a} = a + y, \tilde{b} = b + n - y)$

- So, if $p(y|\theta) = \text{Binomial}(n; \theta)$ and $p(\theta) = \text{Beta}(a, b)$, then $p(\theta|y) = \text{Beta}(\tilde{a}, \tilde{b})$.
  $\Longrightarrow$ The Beta distribution is conjugate for the binomial sampling model!

beta binomial because both prior is beta & outcome is binomial

Binomial model: Conjugate prior and posterior distribution

- Still considering:

~: the posterior numbers

$$\left.\begin{aligned} p(\theta) &= \text{Beta}(a, b) \\ p(y|\theta) &= \text{Binomial}(n; \theta) \end{aligned}\right\} \Longrightarrow p(\theta|y) = \text{Beta}(\tilde{a} = a + y, \tilde{b} = b + n - y)$$

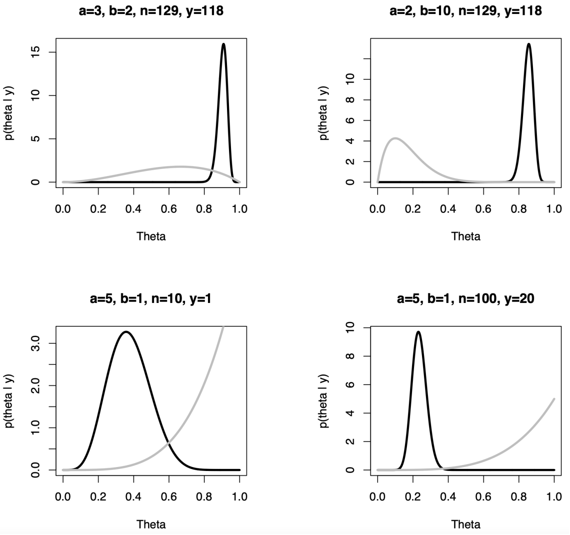
- Notice that:
    - the parameter $\tilde{a}$ of the posterior distribution can be obtained by adding the parameter a of the prior distribution to the number of 1's y
    - the parameter $\tilde{b}$ of the posterior distribution can be obtained by adding the parameter b of the prior distribution to the number of 0's n-y
    - the numerator in the posterior distribution has $\Gamma(\tilde{a} + \tilde{b}) = \Gamma(a + y + b + n - y) = \Gamma(a + b + n)$

In light of this we can interpret:

- a as the prior number of 1's
- b as the prior number of 0's
- $a + b$ as the prior sample size

Binomial model: Conjugate prior and posterior distribution

Different forms of the posterior distribution $p(\theta|y)$ for different choices of (a, b) and different data observed: prior (gray), posterior (black)



Binomial model: Conjugate prior and posterior distribution

- The posterior distribution $p(\theta|y)$ is an entire distribution. What if we want to report some summaries of the posterior distribution? Or what if we want to give just an estimate of θ? point estimation
  - Summaries: we can report quantiles of the posterior distribution, the mean of the posterior distribution (posterior mean), the mode of the posterior distribution (posterior mode), the variance (posterior variance) or credible intervals (Bayesian version of confidence intervals).
  - Estimates: as point estimates we can use the posterior mean, the posterior median, or the posterior mode. The usual limitations of those summary statistics apply also here.

Binomial model: Conjugate prior and posterior distribution

- The posterior variance of θ is $\text{Var}(\theta|y)$. Now remember the following result.

  Conditional variance identity: If θ and Y are two random variables, then:

  $$\text{Var}(\theta) = \text{E}[\text{Var}(\theta|Y)] + \text{Var}[\text{E}(\theta|Y)]$$

  provided that the expectations exist.

- Since Variance ≥ 0, this means that $\text{E}[\text{Var}(\theta|Y)] \leq \text{Var}(\theta)$.

  In other words, the posterior variance is, on average, smaller than the prior variance. Notice that this is on average: there might be instances, datasets, for which this does not happen.

- The amount by which the mean posterior variance is smaller than the prior variance depends on the variance of the posterior mean.

Binomial model: Conjugate prior and posterior distribution

- What about the posterior mean? Is there a relationship between the prior mean and the posterior mean?
- The conditional mean identity states that if θ and Y are two random variables then:

  $$E(\theta) = E[E(\theta|Y)]$$

  provided that the expectations exist.
- Therefore, the prior mean is the average of the posterior expectation averaged over the distribution of all possible data.

Binomial model: Conjugate prior and posterior distribution

- Suppose that we have: $\theta \sim p(\theta) = \text{Beta}(a, b)$
  $$Y|\theta \sim p(y|\theta) = \text{Binomial}(n; \theta)$$
  and we want to report a point estimate for θ based on the data. We can then report the posterior mean.

- Since $p(\theta|y) = \text{Beta}(\tilde{a} = a + y, \tilde{b} = b + n - y)$, using the formula for the mean of a $\text{Beta}(a, b)$ random variable, we have that:
  $$\begin{aligned}\text{posterior mean} &= E(\theta|y) = \tfrac{\tilde{a}}{\tilde{a}+\tilde{b}} = \tfrac{a+y}{(a+y)+(b+n-y)} = \tfrac{a+y}{a+b+n} \\ &= \tfrac{a+b}{a+b+n} \cdot \tfrac{a}{a+b} + \tfrac{n}{a+b+n} \cdot \tfrac{y}{n} \\ &= \tfrac{a+b}{a+b+n} \cdot \text{prior mean} + \tfrac{n}{a+b+n} \cdot \text{sample average}\end{aligned}$$
  The posterior mean is a weighted average of the prior mean and the sample average.

- The weight of the prior mean, $\frac{a+b}{a+b+n}$, is a ratio of the prior sample size to the sum of the prior sample size and the sample size of the data.

Binomial model: Conjugate prior and posterior distribution

- $E(\theta|y) = \frac{a+b}{a+b+n} \cdot$ prior mean $+ \frac{n}{a+b+n} \cdot$ sample average.
- If $n \gg a + b =$ prior sample size, then the weight of the prior mean becomes rather small and the majority of the information about θ comes from the data and not from the prior. In this case:

$$E(\theta|y) \approx \frac{y}{n} = \bar{y}$$

- The mode of the posterior distribution, $p(\theta|y) = \text{Beta}(a + y, b + n - y)$ is given by:

$$\text{Mode}(\theta|y) = \frac{a + y - 1}{a + b + n - 2}$$

Also here note that if the sample size n is very large, the posterior mode is approximately equal to $\frac{y}{n}$, the sample average.

Binomial model: Conjugate prior and posterior distribution

- What about the posterior variance under our choice of prior $p(\theta)$ and sampling model $p(y|\theta)$?
- Since $p(\theta|y) = \text{Beta}(a+y, b+n-y)$, it follows:

$$\text{Var}(\theta|y) = \frac{E(\theta|y) \cdot E(1-\theta|y)}{a+b+n+1}$$

and if $n \gg a+b =$ "prior sample size" then

$$\text{Var}(\theta|y) = \frac{E(\theta|y) \cdot E(1-\theta|y)}{a+b+n+1} \approx \frac{1}{n} \cdot \frac{y}{n} \cdot (1 - \frac{y}{n})$$

Binomial model: Confidence region

- The posterior distribution $p(\theta|y)$ can also be summarized by providing intervals for θ of the form $[l(y), u(y)]$ that have a high posterior probability of containing θ.
- E.g., we say that an interval $[l(y), u(y)]$ has a 95% Bayesian coverage or is a 95% credible interval for θ if

$$P(\theta \in [l(y), u(y)]|y) = 0.95$$

- Note the difference in interpretation of the intervals in the frequentist and Bayesian paradigms:
  - The frequentist coverage of a confidence interval refers to the performance of the interval **prior** to observing the data.
  - The Bayesian coverage gives the probability that an interval contains the true parameter **after** the data has been observed.

Binomial model: Confidence region

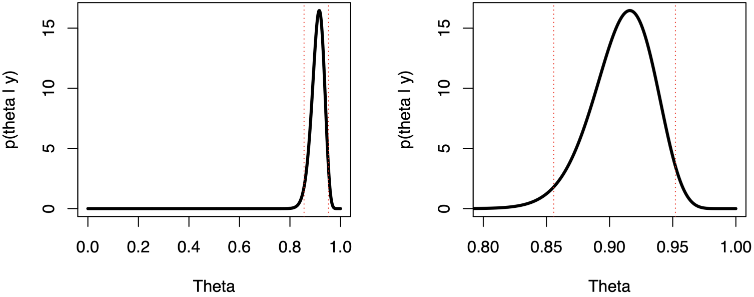
- How do we obtain a 95% credible interval in the Beta-binomial example that we have considered?

- Since $p(\theta|y) = \text{Beta}(a+y, b+n-y)$, we can obtain a 95% credible interval for θ by taking the 2.5-th percentile and the 97.5-th percentile of the $\text{Beta}(a+y, b+n-y)$ distribution. got some problems but easier to address

- Example: For our example from the General Survey with $n = 129, y = 118$, we have the following 95% credible intervals for θ for different choices of a and b:

Prior parameters (a, b)	95% Credible Interval for θ
(1,1)	[0.845,0.951]
(2,1)	[0.855,0.952]
(3,1)	[0.856,0.952]
(1,2)	[0.845,0.946]
(2,2)	[0.847,0.947]

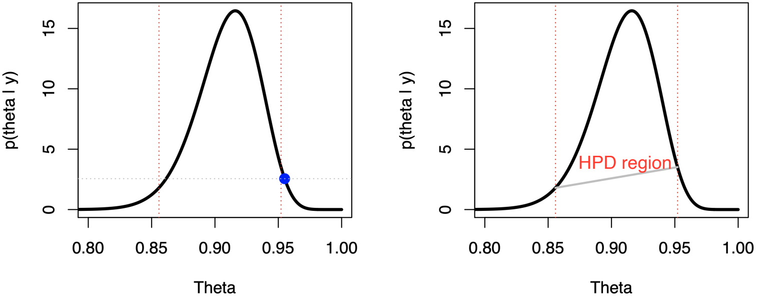
Binomial model: Confidence region



- Posterior distribution obtained when the prior distribution for θ is $p(\theta) = \text{Beta}(3, 1)$.
- The dashed red lines indicate the 95% credible interval for θ.

Binomial model: Confidence region

this lower point has a lower prob. (parallel to blue dot)
-> drawback of creating CI in this way

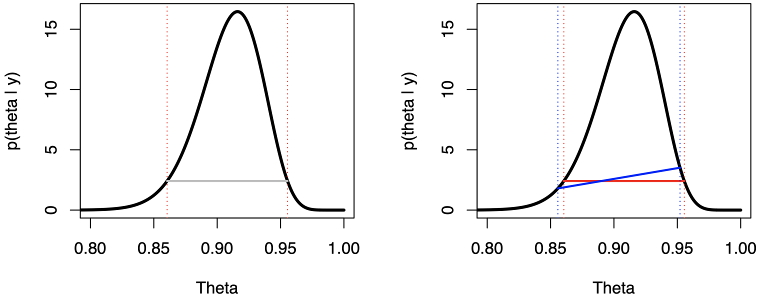


- Posterior distribution obtained when the prior distribution for θ is $p(\theta) = \text{Beta}(3, 1)$.
- The dashed red lines indicate the 95% credible interval for θ.
- There are values of θ that fall outside of the credible interval $[l(y), u(y)]$ for which the posterior density has higher values than those inside the interval. This might seem a non-desirable property.

Binomial model: Confidence region

- A $100 \cdot (1-\alpha)\%$ highest posterior density (HPD) region is a subset $s(y)$ of the parameter space Θ such that:
  - $P(\theta \in s(y)|y) = 100 \cdot (1-\alpha)\%$
  - If $\theta_a \in s(y)$ and $\theta_b \notin s(y)$, then $P(\theta_a|y) > P(\theta_b|y)$
- The second condition ensures that all points within the HPD region have larger posterior probability than those outside the region.
- To find a HPD region by eye, draw the posterior density $p(\theta|y)$ of θ, draw an horizontal line and take as elements of the HPD region all values of θ that fall above the horizontal line. Move the horizontal line down across the density until the region so delimited has a posterior probability of $100 \cdot (1-\alpha)\%$.
- Note that a HPD region might not yield an interval if the posterior density is not unimodal.

Binomial model: Confidence region



- 95% High posterior density (HPD) region for θ obtained when the prior distribution for θ is: $p(\theta) = \text{Beta}(3, 1)$.
- 95% credible interval for θ compared with the 95% HPD region.

Binomial model: Sensitivity to prior distribution

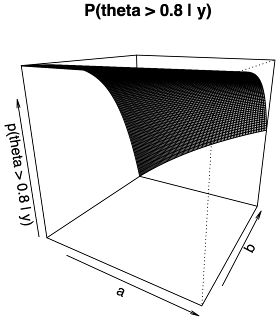
- We have seen that in the Beta-binomial model the prior $p(\theta) = \text{Beta}(a, b)$ yields the posterior density $p(\theta|y) = \text{Beta}(a + y, b + n - y)$.

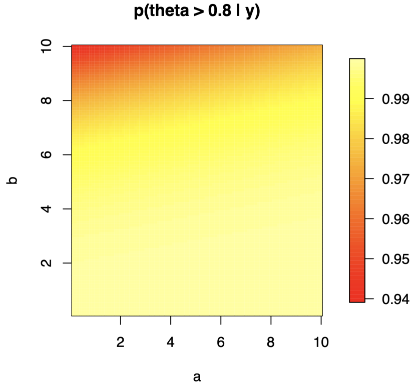
- We want to investigate the sensitivity of the posterior density on the choice of the prior: we can look for example at how the posterior mean changes as we vary a and b.

- More interestingly, we can look at how different choices of the prior parameters a and b affect our posterior belief that θ lies in a certain region of the parameter space Θ.

- For example, for the General Survey data considered where $n = 129, y = 118$ and $n - y = 11$, we can look at how the posterior probability $P(\theta > 0.8|y)$ varies as a and b change. We can plot this.

Binomial model: Sensitivity to prior distribution

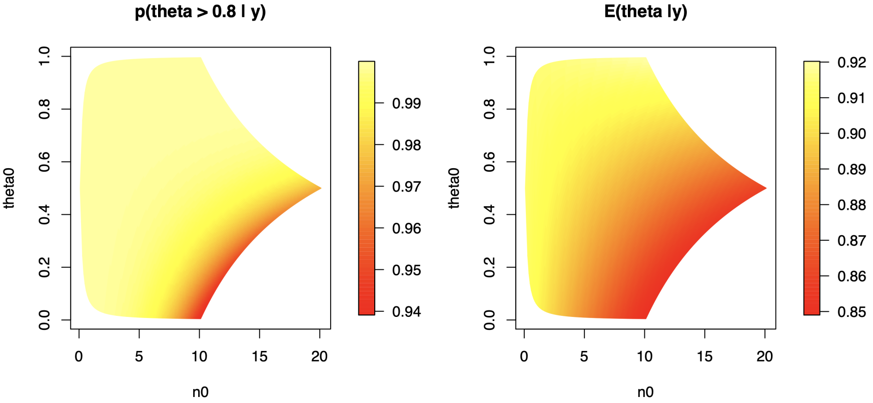

P(theta > 0.8 | y)


p(theta > 0.8 | y)

- 3D plot of posterior probability $p(\theta > 0.8|y)$ for a and b varying between 0 and 10.
- 2D plot of posterior probability $p(\theta > 0.8|y)$ for a and b varying between 0 and 10.

Binomial model: Sensitivity to prior distribution

- Since a and b determine the prior mean, thus our prior belief of θ, we can change the axes of the previous plots to these new axes:
  - $\theta_0 = \frac{a}{a+b}$: the prior mean
  - $n_0 = a + b$: the prior sample size, which represents our confidence in θ_0



Binomial model: Prior and posterior predictive distribution

only assume conditional independent: independent when given theta: Y_1, Y_2… | \theta \sim iid p(y|\theta)

- Let's go back to our original data in the form of binary random variable.
- $Y_1, ..., Y_n$ are exchangeable binary random variables. Suppose that we want to predict a new observation.
- We distinguish two ways to make predictions:
  - predicting a new observation **before** having observed the data: this is done by deriving the marginal distribution p(y).
  - predicting a new observation **after** having observed data $y_1, ..., y_n$: this is done by sampling from the conditional distribution $p(y|y_1, ..., y_n)$

therefore, not unconditionally independent

- In the first case, we are using the prior predictive distribution, in the latter we are using the posterior predictive distribution.

Binomial model: Prior predictive distribution

- Let's derive the prior predictive distribution $p(y)$ in the case Y is a binary random variable for which we have set up a Bernoulli sampling model, that is, $p(y|\theta) = \text{Bernoulli}(\theta)$:

$$\begin{aligned} p(y) &= \int_{\theta\in\Theta} p(y,\theta)d\theta = \int_{\theta\in\Theta} p(y|\theta)p(\theta)d\theta \\ &= \int_{\theta\in\Theta} \theta^y(1-\theta)^{1-y}p(\theta)d\theta \end{aligned}$$

- Now, we can compute explicitly $p(1)$. This is:

$$p(1) = \int_{\theta\in\Theta} \theta p(\theta)d\theta = E(\theta)$$

So, prior to seeing any observation, the probability of $Y = 1$, that is, the probability that a patient will not have any complication 90 days after taking the drug is equal to the prior mean of θ, our prior belief of θ.

Hence, if our prior distribution is $p(\theta) = \text{Uniform}(0,1)$, $E(\theta) = 0.5$ and thus a priori, without having observed any data we claim that there is a 50% chance that a patient will not have any complication.

Binomial model: Posterior predictive distribution

- Now, suppose that we have observed data $y_1, ..., y_n$ and we want to predict a new observation y_{n+1} **given** what we have observed so far.

- The posterior predictive distribution is the conditional density:

$$
\begin{aligned}
p(y_{n+1}|y_1, ..., y_n) &= \int_{\theta\in\Theta} p(y_{n+1}, \theta|y_1, ..., y_n)d\theta \\
&= \int_{\theta\in\Theta} p(y_{n+1}|\theta, y_1, ..., y_n)p(\theta|y_1, ..., y_n)d\theta \\
&= \int_{\theta\in\Theta} p(y_{n+1}|\theta)p(\theta|y_1, ..., y_n)d\theta
\end{aligned}
$$

- The posterior predictive distribution depends on the data. Y_{n+1} is **NOT** independent of $Y_1, ..., Y_n$.

  In fact, $Y_1, ..., Y_n$ give information about θ which in turn gives information about Y_{n+1}!

  If Y_{n+1} was independent of $Y_1, ..., Y_n$ that would mean that we would never learn anything about the unsampled population from the sampled cases.

Binomial model: Posterior predictive distribution

- Let's derive the form of the posterior predictive distribution in the case $Y_1, ..., Y_n$ are binary random variables, $\theta \sim p(\theta) =$ Beta(a, b) and the sampling model is $p(y|\theta) =$ Bernoulli(θ).

- In this case, $p(\theta|y_1, ..., y_n) =$ Beta$(a + \sum_{i=1}^n y_i, b + n - \sum_{i=1}^n y_i)$. Thus:

$$
\begin{aligned}
&p(y_{n+1}|y_1, ..., y_n)\\
&= \int_{\theta \in \Theta} p(y_{n+1}|\theta) p(\theta|y_1, ..., y_n) d\theta\\
&= \int_{\theta \in [0,1]} \theta^{y_{n+1}} (1-\theta)^{1-y_{n+1}}\\
&\quad \cdot \tfrac{\Gamma(a+b+n)}{\Gamma(a+\sum_{i=1}^n y_i)\Gamma(b+n-\sum_{i=1}^n y_i)} \theta^{a+\sum_{i=1}^n y_i - 1} (1-\theta)^{b+n-\sum_{i=1}^n y_i - 1} d\theta\\
&= \tfrac{\Gamma(a+b+n)}{\Gamma(a+\sum_{i=1}^n y_i)\Gamma(b+n-\sum_{i=1}^n y_i)} \tfrac{\Gamma(a+\sum_{i=1}^n y_i + y_{n+1})\Gamma(b+n-\sum_{i=1}^n y_i + 1 - y_{n+1})}{\Gamma(a+b+n+1)}
\end{aligned}
$$

Binomial model: Posterior predictive distribution: example

- Therefore, after observing $y_1, ..., y_n$:
  - $p(0|y_1, ..., y_n) = \frac{b+n-\sum_{i=1}^{n} y_i}{a+b+n}$ a+b+n: posterior sample size
  - $p(1|y_1, ..., y_n) = \frac{a+\sum_{i=1}^{n} y_i}{a+b+n}$

a + \Sigma y_i = prior # of 1's + obs of 1's = posterior number of 1's

- Note that: $p(1|y_1, ..., y_n) = E(\theta|y_1, ..., y_n)$.
- Example: for the data on the drug for ischemic stroke, if the prior distribution has parameters $a = 1, b = 1$, then:

$$p(1|y_1, ..., y_n) = \frac{1 + 118}{1 + 1 + 129} = \frac{119}{131}$$

Binomial model: Predicting multiple observations

- The previous formulas for the posterior predictive distribution can be extended to predictions of multiple variables.
- Again, suppose that $Y_1, ..., Y_n$ are exchangeable binary random variables, $\theta \sim p(\theta) = \text{Beta}(a, b)$ and $Y_1, ..., Y_n|\theta \overset{i.i.d.}{\sim} \text{Bernoulli}(\theta)$.
- Suppose that we observe $y_1, ..., y_n$ with $\sum_{i=1}^n y_i = y$, then the posterior distribution for θ is $p(\theta|y) = \text{Beta}(\tilde{a}, \tilde{b})$ where $\tilde{a} = a + y$ and $\tilde{b} = b + n - y$.

  The posterior predictive distribution for m new observations $y_{n+1}, ..., y_{n+m}$ with $\tilde{y} = \sum_{j=1}^m y_{n+j}$ is:

$$p(y_{n+1}, ..., y_{n+m}|y_1, ..., y_n) = \frac{\Gamma(a+b+n)}{\Gamma(a+y)\Gamma(b+n-y)} \frac{\Gamma(a+y+\tilde{y})\Gamma(b+n-y+m-\tilde{y})}{\Gamma(a+b+n+m)}$$

Binomial model: an example

- Problem: A study reported the long-term effects of exposure to low levels of lead in childhood.

  Researchers analyzed children's shed primary teeth for lead content. Of the children whose teeth had a lead content of more than 22.22 parts per million (ppm), 22 eventually graduated from high school and 7 did not. Suppose your prior density for the proportion of all such children who will graduate from high school is Beta$(1{,}1)$, and so your posterior density is ....

  Based on this information, if 10 more children are found to have a lead content of more than 22.22 ppm, what is your probability that 9 or 10 of them will graduate from high school?

assume binomial: y_i=1, graduate; y_i=0 not graduate

Beta(1,1): uniformed

P(\Sigma_{i=1}^10 ^\sim y_i l y_1, y_2…y_29) \geq 9 = P(^\sim y=9 l y_1…y_29) + P(^\sim y = 10 l y_1,…y_29)

parately: \frac{\binom{10}{9} P(^\sim y_1, … ^\sim y_10l y_1, …y_29) }{\hat{y_1=1, y_2=1…y_9=1, y_10=0}}

Binomial model: an example

- Answer: The data collected are $n = 29$ with $y = 22$ and $n - y = 7$. Since the prior is $\text{Beta}(1, 1)$, the posterior distribution for θ is $\text{Beta}(1 + 22, 1 + 7) = \text{Beta}(23, 8)$.

  We want to find the posterior predictive probability that $m = 10$ new observations $y_{n+1}, ..., y_{n+m}$ are such that $\sum_{j=1}^{10} y_{n+j} = \tilde{y}$ is 9 or 10. Using the formulas above, we have:

$$P(\tilde{y} = 9|y = 22) = \binom{10}{9} \frac{\Gamma(1+1+29)}{\Gamma(1+22)\Gamma(1+7)} \frac{\Gamma(1+22+9)\Gamma(1+7+10-9)}{\Gamma(1+1+29+10)}$$

$$= 10 \cdot \frac{\Gamma(31)}{\Gamma(23)\Gamma(8)} \cdot \frac{\Gamma(32)\Gamma(9)}{\Gamma(41)} = 10 \times 0.0190 = 0.190$$

and

$$P(\tilde{y} = 10|y = 22) = \binom{10}{10} \frac{\Gamma(1+1+29)}{\Gamma(1+22)\Gamma(1+7)} \frac{\Gamma(1+22+10)\Gamma(1+7+10-10)}{\Gamma(1+1+29+10)}$$

$$= 1 \cdot \frac{\Gamma(31)}{\Gamma(23)\Gamma(8)} \cdot \frac{\Gamma(33)\Gamma(8)}{\Gamma(41)} = 0.076$$

Binomial model: an example

- Therefore, the probability that among the 10 children found to have elevated lead content, 9 or more will graduate from high school is:

$$P(\tilde{y} \geq 9 | y = 22) = 0.190 + 0.076 = 0.266$$

Binomial model: prior choice

- So far, we have conducted inference on our data using a prior that is:
  - non-informative: $p(\theta) = \text{Uniform}(0, 1)$
  - conjugate: $p(\theta) = \text{Beta}(a, b)$
- Can we use other priors?

Binomial model: prior distributions

- How to choose a prior distribution $p(\theta)$? There are various types of prior distributions:
  - Jeffreys' prior: Jeffreys' prior distribution is a prior that satisfies the local uniformity prior property for non-informative priors. It is a prior distribution that is based on Fisher information.
  - Definition: Let $Y_1, ..., Y_n$ be random variables for which we assume the following sampling model $p(y|\theta)$. Fisher information is:

$$I(\theta) = -E_{y|\theta}\left[\frac{\partial^2 \log p(y|\theta)}{\partial \theta^2}\right]$$

If θ is a q-dimensional vector of parameters, then:

$$I(\theta) = -E_{y|\theta}\left[\frac{\partial^2 \log p(y|\theta)}{\partial \theta_i \partial \theta_j}\right]_{q \times q}$$

and $I(\theta)$ is a $q \times q$ matrix.

Binomial model: prior distributions: Jeffreys' prior

- Jeffreys' prior is defined as:

  p(\psi (\theta)) = I(\psi)|^{1/2} = |\frac{d \psi}{d \theta}|^{-1} |I(\theta)|^{1/2}

  $$p(\theta) \propto |I(\theta)|^{\frac{1}{2}}$$

  where $|\cdot|$ denotes the determinant.

- It can be proved that Jeffreys' prior is locally uniform and therefore non-informative.

  This property of Jeffreys' prior is quite useful since it provides an automated scheme for finding a non-informative prior for any parametric sampling model $p(y|\theta)$.

- Jeffreys' prior might be an improper prior.

- An appealing property of Jeffreys' prior is that it is invariant with respect to one-to-one transformation.

  This follows from the fact that if $\psi = \psi(\theta)$ is a one-to-one transformation of θ, then:

  $$I(\theta) = I(\psi(\theta)) \cdot \overset{\text{sqrt of first derivative}}{\left(\frac{d\psi(\theta)}{d\theta}\right)^2} \implies |I(\theta)|^{\frac{1}{2}} = |I(\psi(\theta))|^{\frac{1}{2}} \cdot \left|\frac{d\psi(\theta)}{d\theta}\right|$$

Binomial model: Jeffreys' invariance principle

- The invariance property of Jeffreys' prior for θ means that if we consider a new parameter $\psi = \psi(\theta)$ and derive Jeffreys' prior for ψ, we obtain the same answer as if we'd have computed Jeffreys' prior for θ and then performed the usual Jacobian transformation to the ψ scale.

- The invariance property of Jeffreys' prior is a property that we would like to hold in general.

- Jeffreys' invariance principle: Any rule for determining a prior distribution for θ should yield an equivalent result if applied to a transformed parameter $\psi = \psi(\theta)$.

Binomial model: Jeffreys' prior

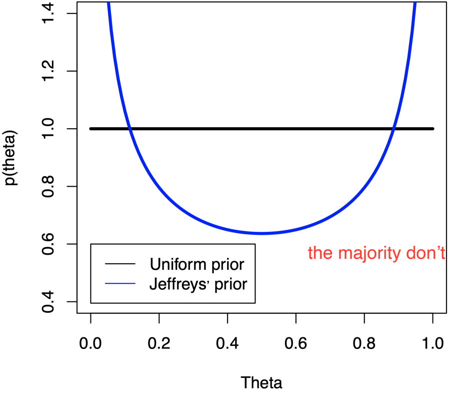
- What is Jeffreys' prior for the sampling model: $p(y|\theta) = \text{Binomial}(n; \theta)$?
- We have:

$$\begin{aligned} I(\theta) &= -E_{y|\theta}\left[\frac{\partial^2 \log p(y|\theta)}{\partial\theta^2}\right] = -E_{y|\theta}\left[-\frac{y}{\theta^2} - \frac{n-y}{(1-\theta)^2}\right] \\ &= \frac{E(y)}{\theta^2} + \frac{n-E(y)}{(1-\theta)^2} = \frac{n}{\theta(1-\theta)} \end{aligned}$$

- This implies that Jeffreys' prior for θ is: $p(\theta) \propto \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}}$.
- Can you recognize what density this is?

Binomial model: Jeffreys' prior

Uniform and Jeffreys' prior for θ



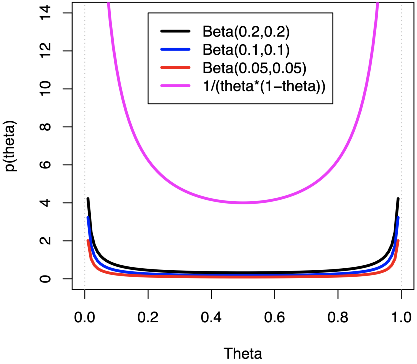
the majority don't vary a lot

- Why is Jeffreys' prior for θ not equal to the uniform prior?

Binomial model: improper priors

eg. sum of P != 1 -> but it’s fine because the posterior will be proper

- We could use an improper prior for θ. How do we construct an improper prior for θ?
- We could obtain an improper prior for θ by taking $p(\theta) = \text{Beta}(a, b)$ with a and b both going to 0.
  More precisely, let $p(\theta) \propto \frac{1}{\theta(1-\theta)}$.



In this case: $\int_{\theta\in\Theta} p(\theta)d\theta \propto \int_{\theta\in[0,1]} \frac{1}{\theta(1-\theta)} d\theta = \infty$

Binomial model: improper priors

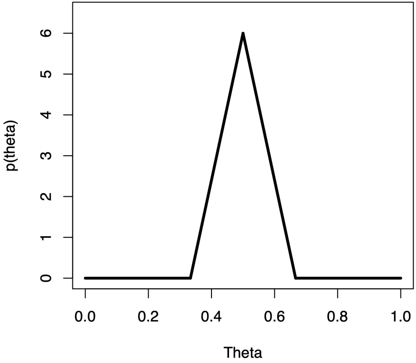
- Note that even if we place as prior on θ, $p(\theta) \propto \frac{1}{\theta(1-\theta)}$, the posterior distribution is still proper!
- Example: Using again the data on the ischemic stroke drug: $n = 129$ patients were given the drug, $\sum_{i=1}^{n} y_i = 118$ did not have any unfavorable outcome and $n - \sum_{i=1}^{n} y_i = 11$ did. Suppose we take as prior on θ, $p(\theta) \propto \frac{1}{\theta(1-\theta)}$, then:

$$\begin{aligned} p(\theta|y_1, ..., y_{129}) &\propto p(y_1, ..., y_{129}|\theta)p(\theta) \\ &= \theta^{118}(1-\theta)^{11}\theta^{-1}(1-\theta)^{-1} \\ &= \theta^{117}(1-\theta)^{10} \propto \text{Beta}(118, 11) \end{aligned}$$

$\Longrightarrow$ the posterior distribution is proper!

Binomial model: other prior choices

- Suppose now that we place the following prior on θ:



  then, the posterior distribution cannot be computed in closed form.

- When that happens, posterior inference on θ can still be conducted, however, we need to use Monte Carlo methods (we will see this later in the class).

Poisson model: Topics

- Sufficient statistic
- Conjugate prior and posterior distribution
- Prediction

Poisson model: Sufficient statistic

- Let's consider now the case where $Y_1, ..., Y_n$ are exchangeable random variables that can take only non-negative integer values.
- Example: suppose that Y_i is the number of children of individual i.
- Suppose that we use the following sampling model for Y_i: $p(y|\theta) = \text{Poisson}(\theta)$ where $\theta > 0$:

$$p(y|\theta) = \text{Poisson}(y|\theta) = \frac{1}{y!}\theta^y \exp(-\theta), \qquad y = 0, 1, 2, ...$$

  Then, $\text{E}(Y|\theta) = \theta$ and $\text{Var}(Y|\theta) = \theta$.

- If we assume that $Y_1, ..., Y_n$ are conditionally i.i.d. given θ:

$$\begin{aligned} p(y_1, ..., y_n|\theta) &= \textstyle\prod_{i=1}^n p(y_i|\theta) = \prod_{i=1}^n \frac{1}{y!}\theta^y \exp(-\theta) \\ &= \textstyle\frac{1}{y_1!y_2!...y_n!}\theta^{\sum_{i=1}^n y_i} \exp(-n\theta) \end{aligned}$$

- This implies that the posterior distribution of θ is:

$$p(\theta|y_1, ..., y_n) \propto p(y_1, ..., y_n|\theta)p(\theta) \propto \theta^{\sum_{i=1}^n y_i} \exp(-n\theta)p(\theta)$$

Poisson model: Sufficient statistic

- Let's evaluate the posterior odds of θ_a relative to θ_b. We have

$$\frac{p(\theta_a|y_1,...,y_n)}{p(\theta_b|y_1,...,y_n)} = \frac{p(y_1,...,y_n|\theta_a)p(\theta_a)}{p(y_1,...,y_n|\theta_b)p(\theta_b)}$$
$$= \frac{\theta_a^{\sum_{i=1}^n y_i} \exp(-n\theta_a)p(\theta_a)}{\theta_b^{\sum_{i=1}^n y_i} \exp(-n\theta_b)p(\theta_b)}$$

- The posterior odds depend on the data only through $\sum_{i=1}^n y_i$, therefore this statistic contains all the information about θ that is contained in the sample $\Longrightarrow \sum_{i=1}^n y_i$ is a sufficient statistic.

- Since $Y_1, ..., Y_n$ are conditionally i.i.d. given θ as Poisson(θ), $p(\sum_{i=1}^n Y_i|\theta) = \text{Poisson}(n\theta)$.

the summation will also be poisson

Poisson model: Conjugate prior and posterior distribution

- We want to determine (if it exists) a conjugate prior for the sampling model $p(y|\theta) = \text{Poisson}(\theta)$.
- Given a prior $p(\theta)$ for θ, the posterior distribution of θ given the data $y_1, ..., y_n$ is given by:

$$p(\theta|y_1, ..., y_n) \propto \theta^{\sum_{i=1}^{n} y_i} \exp(-n\theta)p(\theta)$$

- This implies that in order for a prior distribution $p(\theta)$ to be a conjugate prior it needs to contain terms of the form $\theta^{c_1} \exp(c_2\theta)$.
- A positive random variable X is said to have a Gamma(a, b) distribution if

$$p(x; a, b) = \text{Gamma}(x; a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx) \quad \text{for} \quad a, b > 0.$$

Then, E$(X) = \frac{a}{b}$, Var$(X) =$ 文字 $\frac{E(X)}{b} = \frac{a}{b^2}$, and

$$\text{Mode}(X) = \begin{cases} \frac{a-1}{b} & \text{if } a > 1 \\ 0 & \text{if } a \leq 1 \end{cases}$$

if prior assum poisson -> posterior is poisson

Poisson model: Conjugate prior and posterior distribution

- Let $p(\theta) = \text{Gamma}(a, b) = \frac{b^a}{\Gamma(a)}\theta^{a-1}\exp(-b\theta)$, then:

$$p(\theta|y_1, ..., y_n) \propto \theta^{\sum_{i=1}^n y_i}\exp(-n\theta)\frac{b^a}{\Gamma(a)}\theta^{a-1}\exp(-b\theta)$$
$$\propto \theta^{a+\sum_{i=1}^n y_i - 1}\exp(-(b+n)\theta)$$

  that is, $p(\theta|y_1, ..., y_n) = \text{Gamma}(\tilde{a}, \tilde{b})$ where:
  - $\tilde{a} = a + \sum_{i=1}^n y_i$
  - $\tilde{b} = b + n$
  - Thus, b is interpreted as the prior sample size and a as the prior sum of counts in b observations.

- The posterior mean $E(\theta|y_1, ..., y_n)$ is given by:

$$E(\theta|y_1, ..., y_n) = \frac{\tilde{a}}{\tilde{b}} = \frac{a + \sum_{i=1}^n y_i}{b+n} = \frac{b}{b+n}\cdot\frac{a}{b} + \frac{n}{b+n}\cdot\frac{\sum_{i=1}^n y_i}{n}$$

Poisson model: Conjugate prior and posterior distribution

- $E(\theta|y_1, ..., y_n) = \frac{b}{b+n} \cdot \frac{a}{b} + \frac{n}{b+n} \cdot \frac{\sum_{i=1}^n y_i}{n}$

- Since $p(\theta) =$ Gamma(a, b) implies that $E(\theta) = \frac{a}{b}$, the posterior mean is a weighted average of the prior mean and the sample average.

E(\bar{y}|\thet
a) = \theta

- It is clear that if $n \gg b$, then, the weight of the prior mean $\rightarrow 0$ and the posterior mean $\approx$ the sample average, i.e., the information in the data dominates the prior information.

P(y_i l \theta) = poisson
E(y_i l\theta) = \theta
Var(y_i l \theta) = \theta

- For large n:

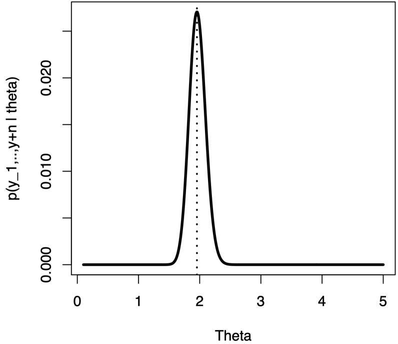
$$E(\theta|y_1, ..., y_n) \approx \frac{\sum_{i=1}^n y_i}{n} = \bar{y}$$

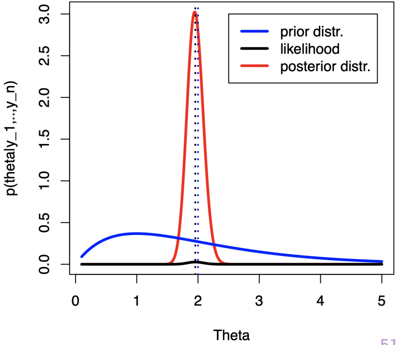
$$\text{Var}(\theta|y_1, ..., y_n) = \frac{E(\theta|y_1, ..., y_n)}{\tilde{b}} = \frac{E(\theta|y_1, ..., y_n)}{b+n} \approx \frac{\bar{y}}{n}$$

Var(\bar{y} | \theta) =
\frac{\sigma^2}{n} = \theta / n =

Poisson model: Conjugate prior and posterior distribution

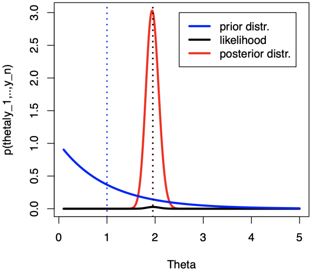
- Example: During the course of the 90's, the General Society Survey collected data on the educational attainment and number of children of 155 women who were in their 20s during the 1970s. Of these 155 women, $n = 111$ had less than a bachelor degree. For these women, let Y_i be the number of children of woman i. The data collected is: $\sum_{i=1}^{111} Y_i = 217$ with $\bar{Y} = 1.95$.
  If we place a $p(\theta) = \text{Gamma}(2, 1)$ prior on θ, then:
  $p(\theta|Y_1, ..., Y_n) = \text{Gamma}(2 + 217, 1 + 111)$.



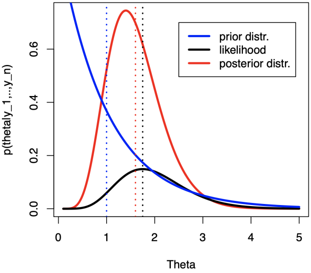


Poisson model: Conjugate prior and posterior distribution

- Here we examine the effect of the sample size on the posterior inference. In both cases, $p(\theta) = \text{Gamma}(1, 1)$ implying that $E(\theta) = 1$ and $\text{Var}(\theta) = 1$.



$n = 111$; $\sum_{i=1}^{n} y_i = 217$; $\bar{y} = 1.95$
$p(\theta|y_1, ..., y_n) = \text{Gamma}(218, 112)$
$E(\theta|y_1, ..., y_n) = \frac{218}{112} = 1.94$
$\text{Var}(\theta|y_1, ..., y_n) = \frac{218}{112 \times 112} = 0.017$



$n = 4$; $\sum_{i=1}^{n} y_i = 7$; $\bar{y} = 1.75$
$p(\theta|y_1, ..., y_n) = \text{Gamma}(8, 5)$
$E(\theta|y_1, ..., y_n) = \frac{8}{5} = 1.6$
$\text{Var}(\theta|y_1, ..., y_n) = \frac{8}{5 \times 5} = 0.32$

Poisson model: Prediction

- Suppose that we are now interested in predicting the number of children for a woman with less than a bachelor degree **given** that we have observed $n = 111$ women and $\sum_{i=1}^{n} y_i = 217$ and we have assumed that $p(\theta) = \text{Gamma}(a, b)$.

- We need to derive the posterior predictive distribution $p(y_{n+1}|y_1, ..., y_n)$ if $p(\theta) = \text{Gamma}(a, b)$.
  We know that in this case
  $p(\theta|y_1, ..., y_n) = \text{Gamma}(a + \sum_{i=1}^{n} y_i, b + n) = \text{Gamma}(\tilde{a}, \tilde{b})$.

- Then:

$$
\begin{aligned}
p(y_{n+1}|y_1, ..., y_n) &= \int_{\theta\in\Theta} p(y_{n+1}|\theta) p(\theta|y_1, ..., y_n) d\theta \\
&= \int_{\theta\in\Theta} \left\{ \tfrac{1}{y_{n+1}!} \theta^{y_{n+1}} e^{-\theta} \right\} \left\{ \tfrac{\tilde{b}^{\tilde{a}}}{\Gamma(\tilde{a})} \theta^{\tilde{a}-1} e^{-\tilde{b}\theta} \right\} d\theta \\
&= \tfrac{\tilde{b}^{\tilde{a}}}{\Gamma(y_{n+1}+1)\Gamma(\tilde{a})} \int_0^\infty \theta^{\tilde{a}+y_{n+1}-1} \exp(-(\tilde{b}+1)\theta) d\theta \\
&= ... = \binom{\tilde{a} + y_{n+1} - 1}{y_{n+1}} (1 - \tfrac{1}{\tilde{b}+1})^{\tilde{a}} (\tfrac{1}{\tilde{b}+1})^{y_{n+1}}
\end{aligned}
$$

Poisson model: Prediction

- The posterior predictive distribution is:

$$p(y_{n+1}|y_1, ..., y_n) = \binom{\tilde{a} + y_{n+1} - 1}{y_{n+1}} (1 - \tfrac{1}{\tilde{b}+1})^{\tilde{a}} (\tfrac{1}{\tilde{b}+1})^{y_{n+1}}$$

  where $y_{n+1} \in \{0, 1, 2, ...\}$.

- A discrete random variable X is said to have a negative binomial distribution with parameters (r, θ) if its density $p(x; r, \theta)$ is given by:

$$p(x; r, \theta) = \text{NegBin}(x; r, \theta) = \binom{r + x - 1}{x} (1-\theta)^r \theta^x, \quad x = 0, 1, 2, ...$$

  where $r = 1, 2, ...$ and $\theta \in [0, 1]$. For such a variable:

$$E(X) = \frac{\theta r}{1 - \theta} \quad \text{and} \quad \text{Var}(X) = \frac{\theta r}{(1 - \theta)^2} = \frac{E(X)}{1 - \theta}$$

- Then, the posterior predictive distribution $p(y_{n+1}|y_1, ..., y_n)$ is a negative binomial distribution with parameters $(\tilde{a}, \frac{1}{\tilde{b}+1}) = (a + \sum_{i=1}^n y_i, \frac{1}{b+n+1})$.

Poisson model: Prediction

- Since $p(y_{n+1}|y_1,...,y_n) = \mathsf{NegBin}(\tilde{a}, \frac{1}{\tilde{b}+1})$, we have

$$E(Y_{n+1}|y_1,...,y_n) = \frac{\tilde{a}\cdot\frac{1}{\tilde{b}+1}}{1-\frac{1}{\tilde{b}+1}} = \frac{\tilde{a}}{\tilde{b}} = E(\theta|y_1,...,y_n) = \frac{a+\sum_{i=1}^n y_i}{b+n}$$

$$\begin{aligned}
\mathsf{Var}(Y_{n+1}|y_1,...,y_n) &= \frac{E(Y_{n+1}|y_1,...,y_n)}{1-\frac{1}{\tilde{b}+1}} = \frac{\tilde{a}}{\tilde{b}}\cdot\frac{\tilde{b}+1}{\tilde{b}} \\
&= \frac{\tilde{a}}{\tilde{b}^2}(\tilde{b}+1) = \mathsf{Var}(\theta|y_1,...,y_n)\cdot(\tilde{b}+1) \\
&= E(\theta|y_1,...,y_n)\cdot\frac{\tilde{b}+1}{\tilde{b}} \\
&= E(\theta|y_1,...,y_n)\cdot\frac{b+n+1}{b+n} \\
&= \frac{a+\sum_{i=1}^n y_i}{b+n}\cdot\frac{b+n+1}{b+n}
\end{aligned}$$

Poisson model: Prediction

- We have obtained:

$$\begin{aligned} E(Y_{n+1}|y_1,...,y_n) &= E(\theta|y_1,...,y_n) \\ \mathsf{Var}(Y_{n+1}|y_1,...,y_n) &= E(\theta|y_1,...,y_n)\cdot \tfrac{b+n+1}{b+n} \\ &= E(\theta|y_1,...,y_n) + \mathsf{Var}(\theta|y_1,...,y_n) \end{aligned}$$

- The posterior predictive variance gives a measure of the posterior uncertainty in Y_{n+1}. This uncertainty comes from uncertainty in the population and sampling variability.

- If n is large, then $\frac{b+n+1}{b+n} \approx 1$ and $\mathsf{Var}(Y_{n+1}|y_1,...,y_n) \approx E(\theta|y_1,...,y_n)$, that is, the uncertainty in Y_{n+1} primarily comes from sampling variability, and this, under a Poisson sampling model, is equal to the expectation, that is θ.

- If n is small, then $\frac{b+n+1}{b+n} > 1$. Note that in this case, the uncertainty in Y_{n+1} has to also account for the uncertainty in θ. Therefore $\mathsf{Var}(Y_{n+1}|y_1,...,y_n)$ is larger than the sampling variability, represented by $E(\theta|y_1,...,y_n)$.

Poisson model: Prediction

- Example: Consider the data from the General Social Survey on the number of children for women with education attainment less than a bachelor's degree. Then, $n = 111$, $\sum_{i=1}^{n} y_i = 217$. Assuming a prior distribution $p(\theta) = \text{Gamma}(2, 1)$, if we want to predict the number of children for a mother with less than a bachelor's degree, we have that:

$$\begin{aligned} p(y_{n+1}|y_1, ..., y_{111}) &= \text{NegBin}(\tilde{a} = 2 + 217, \tfrac{1}{\tilde{b}+1} = \tfrac{1}{(1+111)+1}) \\ &= \text{NegBin}(219, \tfrac{1}{113}) \end{aligned}$$

- If we had to predict the number of children Y_{n+1} for a new woman, we would predict the posterior predictive mean, that is: $E(Y_{n+1}|y_1, ...y_n) = \frac{\tilde{a}}{\tilde{b}} = \frac{219}{112} = 1.96$

- Given the data observed, our level of uncertainty in this prediction is given by: $\text{Var}(Y_{n+1}|y_1, ...y_n) = \frac{\tilde{a}}{\tilde{b}^2} \cdot (\tilde{b} + 1) = \frac{219}{112^2} \times 113 = 1.97$ Note how the results change if: $p(\theta) = \text{Gamma}(2, 1), n = 4$ and $\sum_{i=1}^{n} y_i = 7$.

Poisson model: Jeffreys' prior

- We have seen that if $Y_1, ..., Y_n$ are exchangeable discrete random variables, then we can set up the following model:

$$\theta \sim p(\theta)$$
$$Y_1, ..., Y_n|\theta \overset{i.i.d.}{\sim} \text{Poisson}(\theta)$$

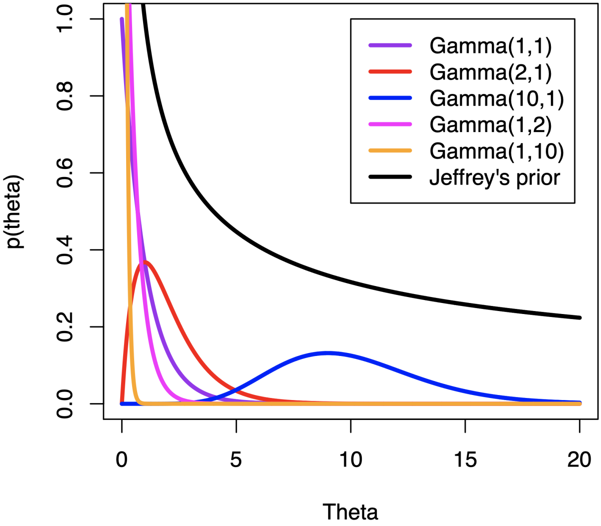
- The prior $p(\theta) = \text{Gamma}(a, b)$ is a conjugate prior for the Poisson sampling model.
- What is Jeffreys' prior for θ in this case?
- We have:

$$I(\theta) = -E_{y|\theta}\left[\frac{\partial^2 \log p(y|\theta)}{\partial\theta^2}\right] = -E_{y|\theta}\left[-\frac{\sum_{i=1}^n y_i}{\theta^2}\right]$$
$$= \frac{E_{y|\theta}(\sum_{i=1}^n y_i)}{\theta^2} = \frac{n\theta}{\theta^2} = \frac{n}{\theta}$$

- This implies that Jeffreys' prior for θ is: $p(\theta) \propto \theta^{-\frac{1}{2}}$.
- This is an improper prior since $\int_0^\infty \theta^{-\frac{1}{2}} d\theta = \infty$.

Poisson model: Jeffreys' prior

- Gamma and Jeffreys' prior for the Poisson sampling model



Exponential family

- The binomial and Poisson sampling models are both instances of one-parameter exponential family model.
- A random variable Y has a density $p(y|\phi)$ that belongs to the one-parameter exponential family model if $p(y|\phi)$ can be written as:

$$p(y|\phi) = h(y)c(\phi)\exp(\phi t(y))$$

  where ϕ is the natural parameter and $t(y)$ is the sufficient statistic.
- If Y has a density $p(y|\phi)$ in the one-parameter exponential family, then:

$$E_{Y|\phi}(t(Y)) = -\frac{c'(\phi)}{c(\phi)}$$

- For densities that belong to the one-parameter exponential family, it is possible to derive conjugate prior distributions $p(\phi)$. These will all be densities of the form:

$$p(\phi; n_0, t_0) = \kappa(n_0, t_0)c(\phi)^{n_0}\exp(n_0 t_0 \phi)$$

Exponential family

- If $Y_1, ..., Y_n$ are random variables such that we can set up the following model:

$$\begin{aligned} Y_1, ..., Y_n|\phi &\overset{i.i.d.}{\sim} p(y|\phi) = h(y)c(\phi)\exp(\phi t(y)) \\ \phi &\sim p(\phi; n_0, t_0) = \kappa(n_0, t_0)c(\phi)^{n_0}\exp(n_0 t_0 \phi) \end{aligned}$$

Then, the posterior distribution is

$$\begin{aligned} p(\phi|y_1, ..., y_n) &\propto p(y_1, ..., y_n|\phi)p(\phi) \\ &= [(\textstyle\prod_{i=1}^n h(y_i))c(\phi)^n \exp(\sum_{i=1}^n \phi t(y_i))] \\ &\quad \cdot \kappa(n_0, t_0)c(\phi)^{n_0}\exp(n_0 t_0 \phi) \\ &\propto c(\phi)^{n_0+n}\exp[\phi(n_0 t_0 + \textstyle\sum_{i=1}^n t(y_i))] \\ &\propto p(\phi; n_0 + n, \tfrac{n_0 t_0 + \sum_{i=1}^n t(y_i)}{n_0+n}) \\ &= p(\phi; n_0 + n, \tfrac{n_0 t_0 + n\bar{t}(\boldsymbol{y})}{n_0+n}) \end{aligned}$$

Exponential family

- since

$$\begin{aligned} Y_1, ..., Y_n|\phi \quad &\overset{i.i.d.}{\sim} \quad p(y|\phi) = h(y)c(\phi)\exp(\phi t(y)) \\ \phi \quad &\sim \quad p(\phi; n_0, t_0) = \kappa(n_0, t_0)c(\phi)^{n_0}\exp(n_0 t_0 \phi) \end{aligned}$$

$$\Longrightarrow p(\phi|y_1, ..., y_n) = p(\phi; n_0 + n, \tfrac{n_0 t_0 + n\bar{t}(\boldsymbol{y})}{n_0+n}),\ \bar{t}(\boldsymbol{y}) = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} t(y_i)$$

  - n_0 is called prior sample size and it measures how informative the prior is.
  - t_0 is called prior guess of $t(Y)$.
    This follows from the fact that:

$$E(t(Y)) = E_\phi(E_{Y|\phi}(t(Y)|\phi)) = E_\phi(-\frac{c'(\phi)}{c(\phi)}) = t_0$$

so, t_0 is the prior expected value of $t(Y)$.

Exponential family

- Consider

  $$p(\phi; n_0, t_0) = \kappa(n_0, t_0) c(\phi)^{n_0} \exp(n_0 t_0 \phi)$$

  and $Y_1, ..., Y_n$ conditionally i.i.d. random variables given ϕ with $p(y|\phi) = h(y)c(\phi)\exp(\phi t(y))$.
  Then:

  $$p(y_1, ..., y_n|\phi) \propto c(\phi)^n \exp(n\bar{t}(\boldsymbol{y})\phi)$$

- The prior $p(\phi; n_0, t_0)$ has the same shape as the likelihood based on n_0 prior observations for which $n\bar{t}(\boldsymbol{y}) = n_0 t_0$, hence the name for n_0 as prior sample size.

Exponential family: binomial model

- The Binomial$(1;\theta)$ = Bernoulli(θ) density belongs to the one-parameter exponential family since:

$$\begin{aligned} p(y|\theta) &= \theta^y(1-\theta)^{1-y} = h(y)(\tfrac{\theta}{1-\theta})^y(1-\theta) \\ &= h(y)\exp(y\log(\tfrac{\theta}{1-\theta}))(1+\exp(\log(\tfrac{\theta}{1-\theta})))^{-1} \\ &= h(y)\exp(y\phi)(1+\exp(\phi))^{-1} = h(y)\exp(y\phi)c(\phi) \end{aligned}$$

  where $\phi = \log(\frac{\theta}{1-\theta})$ is the log-odds, $t(y) = y$ and $c(\phi) = (1+\exp(\phi))^{-1}$.

- The conjugate prior for ϕ is then:

$$p(\phi) \propto (1+\exp(\phi))^{-n_0}\exp(n_0 t_0 \phi)$$

- The prior guess t_0 of $t(Y) = Y$ is the prior expectation of Y , that is our prior guess of the probability that $Y = 1$.

Exponential family: binomial model

- Doing a change of variable from the log-odds $\phi = \log(\frac{\theta}{1-\theta})$ to θ, the conjugate prior for the log-odds

  $$p(\phi) \propto (1 + \exp(\phi))^{-n_0} \exp(n_0 t_0 \phi)$$

  gives a prior for θ of the form

  $$p(\theta; n_0, t_0) \propto \theta^{n_0 t_0 - 1}(1 - \theta)^{n_0 - n_0 t_0 - 1}$$

  which is a Beta$(n_0 t_0, n_0 - n_0 t_0)$ distribution.

- A weakly informative prior can be obtained by setting $n_0 = 1$.

Exponential family: Poisson model

- The Poisson(θ) density belongs to the one-parameter exponential family since:

$$\begin{aligned} p(y|\theta) &= \tfrac{1}{y!}\theta^y \exp(-\theta) = h(y)\exp(y\log(\theta))\exp(-\exp(\log(\theta))) \\ &= h(y)\exp(y\phi)c(\phi) \end{aligned}$$

  where $\phi = \log(\theta), t(y) = y$ and $c(\phi) = \exp(-\exp(\phi))$.

- The conjugate prior for ϕ is then:

$$p(\phi) \propto \exp(-n_0\exp(\phi)) \cdot \exp(n_0 t_0 \phi)$$

- The prior guess t_0 of $t(Y) = Y$ is the prior expectation of Y, that is our prior expectation of the population mean.

Exponential family: Poisson model

- Doing a change of variable from $\phi = \log(\theta)$ to θ, the conjugate prior for ϕ

  $$p(\phi) \propto \exp(-n_0 \exp(\phi)) \cdot \exp(n_0 t_0 \phi)$$

  gives a prior for θ of the form

  $$p(\theta; n_0, t_0) \propto \theta^{n_0 t_0 - 1} \exp(-n_0 \theta)$$

  which is a Gamma$(n_0 t_0, n_0)$ distribution.

- A weakly informative prior can be obtained by setting $n_0 = 1$.