STATS 225: Bayesian Analysis
Bayesian Inference for Simple Models

Babak Shahbaba

Department of Statistics, UCI

Winter, 2015

Outline

- Bayes' theorem
- Exchangeability and deFinetti's theorem
- Bayesian inference
- Prior, likelihood, and posterior
- Posterior predictive probability
- Conjugate priors
- Binomial; Normal; Poisson
- Multinomial; Multivariate normal

Bayes' theorem

- For two events, A and B, *Bayes' theorem* can be simply presented as follows:

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

- Also, recall that if $B = (B_1, B_2, ..., B_n)$ are a set of events that partition the sample space, Ω, using the law of total probability, we have:

$$\begin{aligned} P(A) &= P(A|B_1)P(B_1) + ... + P(A|B_n)P(B_n) \\ P(B_i|A) &= \frac{P(B_i)P(A|B_i)}{\sum_{i'}^{n} P(B_{i'})P(A|B_{i'})} \end{aligned}$$

- This simple formula which is known as *Bayes' theorem* is the basis of Bayesian analysis. (However, using this theorem does not automatically make you a Bayesian!)

Monty Hall problem revisited

- The problem is based on a TV game show hosted by Monty Hall. In this show, contestants had the chance to win cars. Before each show, a car is put behind one of three closed doors. The other two doors have goats behind them. The contestant is asked to choose one of the three doors. Then, Monty Hall then opens one of the other two doors, which he knows does not have a car behind it. After that, the contestant can choose to switch the the remaining unopened door, or stay with his original selection. The question is: should he switch?

Monty Hall problem revisited

- At the beginning, the car can be behind any of the three doors with equal probability. That is,

$$P(D_1) = P(D_2) = P(D_3) = 1/3$$

- Let's say we choose door number 1, and Monty open door number 2.
- Now let's write down the conditional probability of opening, OD_2, given the three possibilities (i.e., D_1, D_2, and D_3):

$$\begin{aligned} P(OD_2|D_1) &= 1/2 \\ P(OD_2|D_2) &= 0 \\ P(OD_2|D_3) &= 1 \end{aligned}$$

- Now using the law of total probability we can find the marginal probability for opening door number 2:

$$P(OD_2) = 1/2 \times 1/3 + 1/2 \times 0 + 1/2 \times 1/3 = 1/2$$

Monty Hall problem revisited

- Using Bayes' theorem, we have:

$$
\begin{aligned}
P(D_3|OD_2) &= \frac{p(D_3)P(OD_2|D_3)}{P(OD_2)} \\
&= \frac{1/3 \times 1}{1/2} = 2/3 \\
P(D_1|OD_2) &= \frac{P(D_1)P(openD_2|D_1)}{P(OD_2)} \\
&= \frac{1/3 \times 1/2}{1/2} = 1/3
\end{aligned}
$$

- Therefore, probability of winning doubles if we switch.

- You can try this using a penny and three cups.

Exchangeability and deFinetti's representation theorem

- Consider a set of observations $y = (y_1, ..., y_n)$. In constructing the joint distribution of these observations, we might believe that the indices are uninformative.
- For example, we toss an old-fashoined thumbtack on a soft surface and keep track of whether the sharp point is up or down. After n tosses, we believe the joint distribution remains the same regardless of which order we consider.
- In this experiment, we do not expect those tosses close together in time to be more similar to each other compared to other tosses.
- we also believe that the above comments are true for any subsets of tosses. That is, if $n = 100$, (y_4, y_{17}) has the same joint distribution as (y_{91}, y_{12}), and (y_{33}, y_{12}, y_{95}) has the same joint distribution of (y_{18}, y_{10}, y_9) and so forth.

Exchangeability and deFinetti's representation theorem

- Such *symmetry* or *similarity* could be expressed as $P(y_1, ..., y_n) = P(y_{\pi_1}, ..., y_{\pi_n})$, where π represent all permutations on $\{1, 2, ..., n\}$. That is, the joint distribution, $P(y_1, ..., y_n)$ is *invariant* to permutations of indices.

- If this property holds for any finite subset of observations, the sequence is called to be *exchangeable*.

- For a sequence of coin tossing, let's denote head as 1 and tail as 0. Then exchangeability means the joint probability of any fixed set of 0's and 1's does not change when we permute them.

- For example, if $n = 3$, then $P(100) = P(010) = p(001)$, i.e., there is nothing special about the location of 1. Also $P(110) = P(101) = P(011)$.

Exchangeability and deFinetti's representation theorem

- We should be careful about our judgement of exchangeability.
- Consider the age of students in this class. Assume that all we know about students is their names.
- We might regard their age as exchangeable, which means students' names are not informative in defining the joint distribution.
- However, what if we also know whether a student is in a master's program or a PhD program.
- In general, master's students tend to be younger.
- Therefore, it would be more appropriate to assume exchangeability only within each group (i.e., master's and PhD).

Exchangeability and deFinetti's representation theorem

- Why do we care about exchangeability?
- deFinetti's representation theorem implies that:

  > *If we can judge an infinite sequence of observations, $y_1, y_2, ...,$ to be exchangeable, we can model any subset of them, $y = (y_1, y_2, ..., y_n)$, as independent and identically distributed (iid) samples from a parametric distribution $P(y|\theta)$.*

- Note that, P is the density function for continuous random variables and probability mass function for discrete variables.
- The actual theorem is more elaborate and its proof is complicated. (See Chapter 1 of Schervish's book for more details.)

Exchangeability and deFinetti's representation theorem

- We derive the following from deFinetti's representation theorem:
  - The conditional distribution of y given θ is

$$P(y|\theta) = P(y_1, y_2, ..., y_n|\theta) = \prod_{i=1}^{n} p(y_i|\theta)$$

  - There exists a *prior* probability distribution $P(\theta)$ over the parameters of the model such that we can find the unconditional (or marginal) joint distribution of observations as follows:

$$P(y) = P(y_1, y_2, ..., y_n) = \int_{\Omega} \prod_{i=1}^{n} P(y_i|\theta)P(\theta)d\theta$$

Bayesian inference

- deFinetti's theorem is an *existence* theorem. We still need to specify the form of such distributions.

- Therefore, we first need to specify the model $P(y|\theta)$ for the observed data, and the prior $P(\theta)$ for the parameter of the model.

- The next step in Bayesian inference is to make probabilistic conclusions regarding the unobserved quantity θ conditional on the observed data y.

- That is, we are interested in $P(\theta|y)$, which is called *posterior distribution*.

Bayesian inference

- Bayes' theorem provides a mathematical formula for obtaining $P(\theta|y)$ based on $P(\theta)$ and $P(y|\theta)$:

$$P(\theta|y) \quad = \quad \frac{P(\theta)P(y|\theta)}{P(y)}$$

- Since $P(y)$ does not depend on θ, we can use the following unnormalized form of the posterior distribution:

$$P(\theta|y) \quad \propto \quad P(\theta)P(y|\theta)$$

- This simple formula is the essential part of Bayesian analysis.
- It is used not only for expressing our updated belief about model parameters (i.e., θ), but also for making decisions (e.g., accepting or rejecting a hypothesis) and predicting unknown observable (e.g., for future cases).

Parametric models

- Next, we will discuss some simple models commonly used for typical random variables.
- These models are based on our assumption for the underlying mechanism that generates the observed data.
- The focus in this model is on one single parameter, which represents the population mean.
- If there are other parameters in the model, we would regard them as nuisance parameters.
- Later, we will discuss multi-parameter models.

Binomial model

- Consider a sequence of independent binary random variables, $x_1, x_2, x_3, \ldots$, such as "heads/tails", "cancer/non-cancer", or "win/lose", such that $x_i \in \{0, 1\}$.

- Denote the probability of observing 1 (e.g., win) as θ.

- We assume x_i has a Bernoulli distribution:

$$P(x_i|\theta) = \theta^{x_i}(1-\theta)^{1-x_i}$$

- If $x_1, x_2, \ldots, x_n$ are n exchangeable binary random variables with Bernoulli distribution, the $y = \sum_i x_i$ (i.e., number of 1's in the sequence) has a Binomial(n, θ) distribution:

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

Binomial model

- Assuming the prior $P(\theta)$, the marginal distribution of y can be obtained as

$$P(y) = \int_0^1 \binom{n}{y} \theta^y (1-\theta)^{n-y} P(\theta) d\theta$$

- We therefore obtain the posterior distribution as follows:

$$P(\theta|y) \quad = \quad \frac{\binom{n}{y}\theta^y(1-\theta)^{n-y}P(\theta)}{\int_0^1 \binom{n}{y}\theta^y(1-\theta)^{n-y}P(\theta)d\theta}$$

- Let's say we are quite ignorant about the possible value of θ. That is to say, we think θ is uniformly distributed in [0, 1], i.e., $P(\theta) = 1, 0 \leq \theta \leq 1$.

Binomial model

- Then we have

$$P(y) \quad = \quad \int_0^1 \binom{n}{y} \theta^y (1-\theta)^{n-y} d\theta = \frac{1}{n+1}$$

- Exercise: Using the pdf of Beta distribution, show that the marginal distribution of y is in fact $1/(n+1)$.

- The posterior distribution simplifies to

$$P(\theta|y) \quad = \quad \frac{(n+1)!}{y!(n-y)!} \theta^y (1-\theta)^{n-y}$$

- This is a Beta$(y+1, n-y+1)$ distribution with expectation

$$E(\theta|y) = \frac{y+1}{n+2}$$

Posterior predictive distribution

- Sometimes our objective is to use the posterior distribution to predict future observations.
- That is, after observing some data, $y = (y_1, y_2, ..., y_n)$, we want to predict the next observation, y_{n+1}, which we denote as $\tilde{y}$.
- Note that we still have uncertainty regarding our prediction, and therefore we express such prediction in the form of probability distribution, called *posterior predictive distribution*, which is obtained by summing (or integrating) over posterior distribution of θ, i.e., $P(\theta|y)$:

$$P(\tilde{y}|y) \quad = \quad \int_\theta P(\tilde{y}|\theta, y)P(\theta|y)d\theta$$

Since $\tilde{y}$ is independent of y given θ, we have

$$P(\tilde{y}|y) \quad = \quad \int_\theta P(\tilde{y}|\theta)P(\theta|y)d\theta$$

Binomial model

- For the above binomial model, since $P(y = 1|\theta) = \theta$, the posterior predictive distribution can be obtained as

$$P(\tilde{y} = 1|y) \quad = \quad \int_0^1 \theta P(\theta|y) d\theta$$

- Recall that the posterior distribution of $\theta|y$ is a $\text{Beta}(y + 1, n - y + 1)$ distribution.

- Therefore, $\int_0^1 \theta P(\theta|y) d\theta$ is the posterior expectation of θ:

$$P(\tilde{y} = 1|y) = \frac{y + 1}{n + 2}$$

- Exercise: Use the above results and find the probability that sun rises tomorrow.

Predicting the election result

- We want to predict which one of two candidates, A or B, will win the election.

- Let's denote the probability that A wins as θ, and we assume *a priori* the probability of winning for candidate A has a uniform distribution.

- We ask 10 people which candidate they would choose in this election. Of 10 people surveyed, 3 people said they are going to vote for A.

- Our updated belief in A's winning has now a Beta(4, 8) distribution.

- The posterior expectation of A's winning is $\frac{4}{12} = 0.33$, which is also the probability that the next person we survey votes for A.

- Note that this is almost the same as the maximum likelihood estimation $\frac{3}{10} = 0.3$ (the similarity is superficial however since the underlying philosophies are different).

Conjugate priors

- In the above example, the derivation of posterior distribution was quite simple since it had a closed form.
- This was due to our choice of prior, i.e., uniform distribution.
- Note that uniform prior on $[0,1]$ is in fact Beta(1, 1) distribution.
- Therefore, for the above binomial model, both prior and posterior are Beta distributions.
- This is called "conjugacy" and the prior is called a "conjugate" prior.
- Conjugacy is informally defined as a situation where the prior distribution $P(\theta)$ and the corresponding posterior distribution, $P(\theta|y)$ belong to the same distributional family.
- Using conjugate priors makes sampling and Bayesian inference much easier compared to using non-conjugate priors.

Exponential family

- A large class of distributions, called *exponential family*, have the following form:

$$P(y_i|\theta) = h(y_i)g(\theta)\exp(\phi(\theta)^T s(y_i))$$

- Many widely used distributions such as normal, bernoulli, and Poisson belong to the exponential family.

- $\phi(\theta)$ is called the "natural parameter" of the family.

- The joint distribution for a set of conditionally (given θ) independent observations, $y = (y_1, y_2, ..., y_n)$ is

$$P(y|\theta) = \left[\prod_i h(y_i)\right] g(\theta)^n \exp(\phi(\theta)^T \sum_i s(y_i))$$

- $t(y) = \sum_i s(y_i)$ is a *sufficient statistic* for θ.

Sufficient statistic

- In the classical framework, $t(y)$ is said to be sufficient since given $t(y)$, the distribution of the data becomes independent of θ.

- In the Bayesian framework, $t(y)$ is said to be a sufficient statistic since θ depends on the data y only through t, i.e., $P(\theta|y) = P(\theta|t)$ for every prior $P(\theta)$.

Conjugate priors (formal definition)

- If for an exponential family we define our prior as

$$P(\theta) \propto g(\theta)^{\eta} \exp(\phi(\theta)^T \nu)$$

  then the posterior would have a similar form:

$$P(\theta|y) \propto g(\theta)^{\eta+n} \exp(\phi(\theta)^T (\nu + t(y)))$$

- In this case, $P(\theta)$ is a conjugate prior.

Binomial model

- Let's look at the binomial model again:

$$\begin{aligned} P(y|\theta, n) &= \binom{n}{y}\theta^y(1-\theta)^{(n-y)} \\ &= \binom{n}{y}\exp\left[y\log(\frac{\theta}{1-\theta}) + n\log(1-\theta)\right] \\ &= \binom{n}{y}(1-\theta)^n\exp[y\log(\frac{\theta}{1-\theta})] \end{aligned}$$

Therefore,

$$g(\theta) = (1-\theta)$$
$$\phi(\theta) = \log(\frac{\theta}{1-\theta})$$

Binomial model

- Recall that a conjugate prior is proportional to

$$P(\theta) \propto g(\theta)^{\eta} \exp(\phi(\theta)^{T} \nu)$$

- Therefore, the conjugate prior for the Binomial model has the following form:

$$P(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- This is a Beta(α, β) distribution.

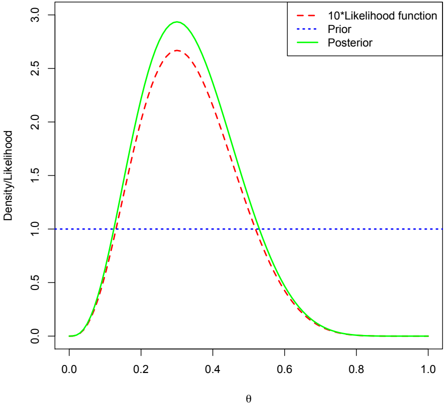
- We can interpret this prior as observing $\alpha - 1$ prior success and $\beta - 1$ prior failure. That is, the prior acts as additional data.

- The posterior distribution is also Beta, with parameters $\alpha + y$ and $\beta + n - y$, i.e., Beta$(\alpha + y, \beta + n - y)$

- Note that the uniform distribution we previously used is in fact a special cases of Beta distribution where $\alpha = \beta = 1$.

The election example

The likelihood function, the uniform prior, and the posterior distribution for the election example.



The election example with informative prior

- As before, assume that we have surveyed 10 people and 3 of them are going to vote for candidate A.

- This time, however, we know that candidate A belongs to the party that in the previous elections won about 55% of votes.

- Instead of a uniform prior, we could use a more informative Beta prior which reflects such prior information.

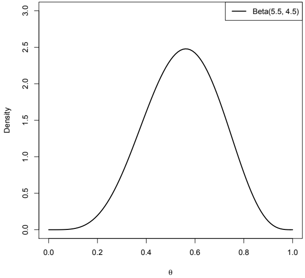
- For example, we could choose a Beta prior whose mean is $\frac{\alpha}{\alpha+\beta} = 0.55$, and it is broad enough to reflect the extent of our uncertainty.

- We should always use a reasonably broad prior. As Savage (1954) said: "Keep the mind open, or at least ajar".

The election example with informative prior

- We choose a $P(\theta) = \mathrm{Beta}(5.5, 4.5)$ as our prior

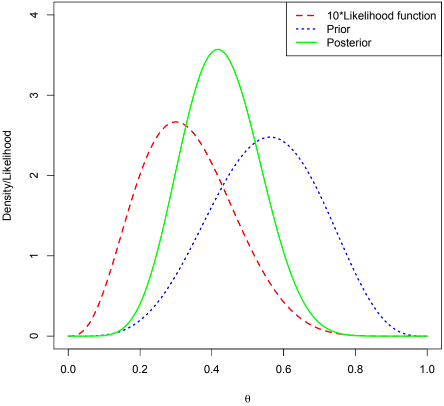
- *Either plot your prior distribution or generate samples from it to make sure it is a good representation of your opinion.*



- The posterior distribution of $\theta|y$ is $\mathrm{Beta}(8.5, 11.5)$. So while the MLE is 0.3, the posterior expectation is 0.425, which is a compromise between the observed data and the prior.

The election example with informative prior

The likelihood function, the Beta prior (which is informative), and the posterior distribution for the election example.



The election example when more data are observed

- Now assume that we have obtained additional budget to survey 20 more people. The result shows that 12 out 20 are going to vote for candidate A.

- It makes sense to update our opinion based on this new information. It is also reasonable not to ignore the previous data.

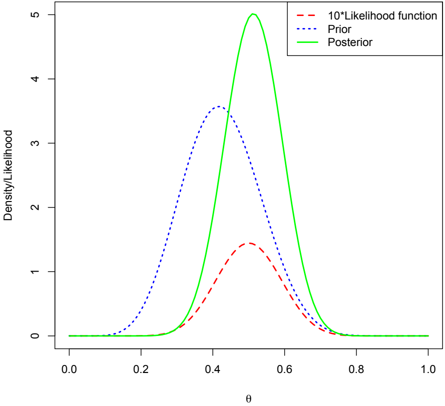
- However, we do not need to start our analysis from the beginning. We can use the previous posterior distribution, $P(\theta) = \mathrm{Beta}(8.5, 11.5)$, as our new prior and obtain a new posterior based on the more recent data.

The election example when more data are observed

- Exercise: Show that the following two approaches provide the same results:
  - We start with our original prior Beta(5.5, 4.5), update our prior based on the first set of data, use the resulting posterior as our new prior and update it based on the second set of data
  - We combine the two data sets and use them to update our original Beta(5.5, 4.5) prior.
- Our new posterior is therefore $P(\theta|y) = \mathrm{Beta}(20.5, 19.5)$.
- The posterior expectation ($\frac{20.5}{20.5+19.5} = 0.51$) and the MLE ($15/30 = 0.5$) are now getting closer as the amount of data increases.

The election example with informative prior

The likelihood function, the Beta prior (based on our previous posterior distribution), and the new posterior distribution for the election example with more data.



Poisson model

- Poisson model is another member of exponential family and is commonly used for count data.
- Assume we have observed $y = (y_1, y_2, ..., y_n)$:

$$\begin{aligned} P(y|\theta) &= \prod_i \frac{\theta^{y_i} \exp(-\theta)}{y_i!} \\ &\propto \exp(-n\theta) \exp(\log(\theta) \sum y_i) \end{aligned}$$

- The conjugate prior would have the following form:

$$\begin{aligned} P(\theta) &\propto (\exp(-\theta))^\eta \exp(\nu \log(\theta)) \\ &\propto \exp(-\eta\theta)\theta^\nu \end{aligned}$$

- Using $P(\theta) \propto \exp(-\beta\theta)\theta^{\alpha-1}$, which is a Gamma$(\alpha, \beta)$ distribution, as our prior, we obtain the following posterior:

$$\theta|y \sim \mathrm{Gamma}(\alpha + \sum_{i=1}^{n} y_i, \beta + n)$$

Poisson model

- When David Beckham joined LA Galaxy, he scored one goal in his first two MLS games.
- Assume that after the manager of LA Galaxy wanted to predict the number of goals Beckham would score in the remaining games.
- We model the number of goals, y_i, he scores in a game using a Poisson model with parameter θ.
- The maximum likelihood is $\hat{\theta} = 0.5$.
- Now let's use a Gamma(α, β) prior for θ.
- What should we choose for α and β?
- If we don't have a clue, we should use a noninformative prior (discussed later) that reflects our lack of information.
- Alternatively, we might want to use Beckham's history in Real Madrid to build a prior opinion.

Poisson model

- When in Madrid, Beckham scored 3 goals in 22 games (i.e., $3/22 = 0.14$ on average) during 06-07 season.

- We could choose a Gamma prior with mean around 0.14, for example, making sure it is broad enough to reflect our uncertainty.

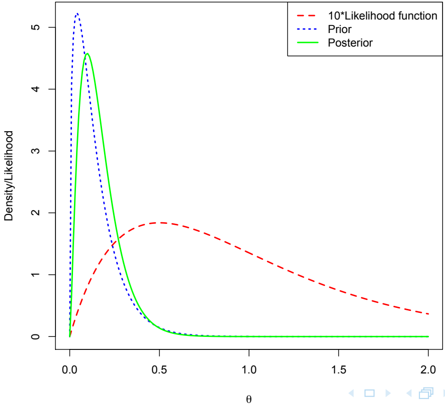
- You should be careful when working with Gamma distribution since there are two different ways of parameterizing it: $f(x|a,b) = [b^a\Gamma(a)]^{-1}x^{a-1}e^{-x/b}$, and $f(x|a,b) = [b^a\Gamma(a)]^{-1}x^{a-1}e^{-bx}$. Here, we use the latter form.

- The mean of Gamma(α, β) is α/β

- For our example, we could use the conjugate Gamma(1.4, 10) prior with mean $1.4/10 = 0.14$.

Poisson model

- Since Gamma is a conjugate prior for the parameter of poisson model, the posterior also has a Gamma distribution, which in this case is a Gamma(1.4+1, 10+2) distribution.
- The expected number of goals is therefore $2.4/12 = 0.2$



Poisson model

- Posterior is again a compromise between the prior and the data (likelihood).
- In this example, as shown in the graph, the posterior is more similar to the prior than the likelihood.
- This is due to the fact that the amount of data is small.
- As the amount of data increases the influence of prior on posterior decreases while the effect of likelihood increases.
- In 2008-2009, Beckham played 25 games and scored 5 goals. This is a 0.2 average, which is much closer to our estimate 0.19 compared to the MLE, which is 0.5 (when I first wrote this example, Beckham had played only two games for Galaxy).

Univariate normal model

- The normal distribution is also a member of exponential families.
- We first consider a situation where there is only one observation and the variance is known.

$$\begin{aligned} y &\sim N(\theta, \sigma^2) \\ P(y|\theta,\sigma) &= \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{\theta^2}{2\sigma^2})\exp(-\frac{y^2}{2\sigma^2}+\frac{\theta y}{\sigma^2}) \end{aligned}$$

- So the general form of a conjugate prior is

$$P(\theta) \propto \exp(a\theta^2 + b\theta)$$

which can be parameterized as

$$P(\theta) \propto \exp(-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2)$$

which is a $N(\mu_0, \tau_0^2)$ distribution.

Univariate normal model

- As the result, the posterior distribution would be

$$P(\theta|\sigma, y) \propto \exp(-\frac{1}{2\sigma^2}(y-\theta)^2 - \frac{1}{2\tau_0^2}(\theta-\mu_0)^2)$$

- When you complete the square, the posterior would also become a normal distribution:

$$P(\theta|\sigma, y) \propto \exp(-\frac{1}{2\tau_1^2}(\theta-\mu_1)^2)$$

which is a $N(\mu_1, \tau_1^2)$ distribution with

$$\mu_1 = \frac{\frac{\mu_0}{\tau_0^2} + \frac{y}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}} \qquad \frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{1}{\sigma^2}$$

- For n observations, we write the model for $\overline{y}$; therefore,

$$\mu_n = \frac{\frac{\mu_0}{\tau_0^2} + \frac{n\overline{y}}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \qquad \frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}$$

Derivations for a single observation

$$
\begin{aligned}
P(\theta|\sigma, y) &\propto \exp(-\frac{1}{2\sigma^2}(y-\theta)^2 - \frac{1}{2\tau_0^2}(\theta-\mu_0)^2) \\
&= \exp[-\frac{1}{2}(\theta^2(\frac{1}{\sigma^2}+\frac{1}{\tau_0^2}) - 2\theta(\frac{y}{\sigma^2}+\frac{\mu_0}{\tau_0^2}) + (\frac{y^2}{\sigma^2}+\frac{\mu_0^2}{\tau_0^2}))] \\
&= \exp[-\frac{a}{2}(\theta^2 - 2\frac{b}{a}\theta + \frac{c}{a})] \\
&= \exp[-\frac{a}{2}(\theta^2 - 2\frac{b}{a}\theta + \frac{b^2}{a^2} - \frac{b^2}{a^2} + \frac{c}{a})] \\
&= \exp[-\frac{a}{2}(\theta - \frac{b}{a})^2] \exp[-\frac{a}{2}(\frac{c}{a} - \frac{b^2}{a^2})] \\
&\propto \exp[-\frac{a}{2}(\theta - \frac{b}{a})^2]
\end{aligned}
$$

This is a normal distribution with mean b/a and variance $1/a$.

Univariate normal model

- Let's assume that the height (in inch) of students in this class follows a normal distribution $N(\theta, 16)$.
- We use a $\theta \sim N(65, 9)$ prior.
- We measure the hight of three students: $y_1 = 72$, $y_2 = 75$, and $y_3 = 70$.
- The posterior distribution of θ is also a normal distribution $N(\mu_n, \tau_n^2)$, where

$$\mu_n = \frac{\frac{65}{9} + \frac{217}{16}}{\frac{1}{9} + \frac{3}{16}} \qquad \frac{1}{\tau_n^2} = \frac{1}{9} + \frac{3}{16}$$

- Therefore, $\theta | y \sim N(69.6, 3.4)$
- The role of prior is substantial here due to the small sample size. The prior modifies the likelihood based estimate (i.e., $\overline{y} = 72.3$), which could have been misleading since all the observed data points happened to be from tall people.

Univariate normal model

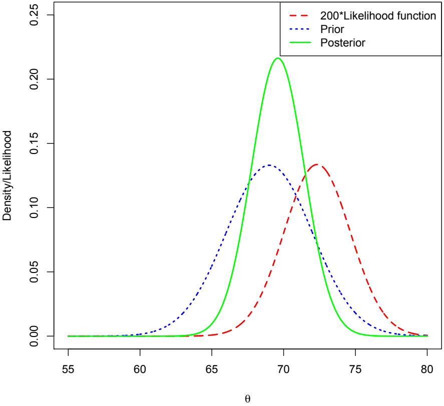
- Let's assume that the height (in inch) of students in this class follows a normal distribution $N(\theta, 16)$.
- We use a $\theta \sim N(65, 9)$ prior.
- We measure the hight of three students: $y_1 = 72$, $y_2 = 75$, and $y_3 = 70$.
- The posterior distribution of θ is also a normal distribution $N(\mu_n, \tau_n^2)$, where

$$\mu_n = \frac{\frac{65}{9} + \frac{217}{16}}{\frac{1}{9} + \frac{3}{16}} \qquad \frac{1}{\tau_n^2} = \frac{1}{9} + \frac{3}{16}$$

- Therefore, $\theta | y \sim N(69.6, 3.4)$
- The role of prior is substantial here due to the small sample size. The prior modifies the likelihood based estimate (i.e., $\overline{y} = 72.3$), which could have been misleading since all the observed data points happened to be from tall people.

Univariate normal model

Again, the posterior distribution could be interpreted as a compromise between the prior and the likelihood.



Univariate normal model

- In the above example, what would be our prediction for the height of next person we observe?

- Denote our prediction as $\tilde{y}$, and the corresponding distribution as $P(\tilde{y}|y)$, i.e., the posterior predictive probability. As before,

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

- By integrating out θ, the conditional distribution of $\tilde{y}$ given y is normal with the following mean and variance:

$$\begin{aligned} E(\tilde{y}|y) &= E(E(\tilde{y}|\theta, y)|y) = E(\theta|y) = \mu_n \\ Var(\tilde{y}|y) &= E(var(\tilde{y}|\theta, y)|y) + Var(E(\tilde{y}|\theta, y)|y) \\ &= E(\sigma^2|y) + Var(\theta|y) \\ &= \sigma^2 + {\tau_n}^2 \end{aligned}$$

Univariate normal model

- We could use μ_n, the posterior expectation of θ, as our single point estimate for $\tilde{y}$.

- The variation around this estimate (i.e., our uncertainty) comes from two difference sources: σ^2, the sampling variation (which is assumed fixed here) of data according to the model, and τ_n^2, the posterior variation of the model parameter, θ, given the observed data.

- In the height example, our guess for the height of the forth student can be expressed by a $N(69.6, 19.4)$ distribution.

Univariate normal model

- For situations where the mean is fixed and the variance, σ^2, is the parameter of interest, we use the following scaled inverse-χ^2 prior:

$$\sigma^2 \sim \mathrm{Inv}\text{-}\chi^2(\nu_0, \sigma_0^2)$$

  where ν_0 is the degrees of freedom and σ_0^2 is the scale parameter.

- The posterior would also be scaled inverse-χ^2 with $\nu_0 + n$ degrees of freedom scale equal to $\frac{\nu_0\sigma_0^2 + \sum_{i=1}^{n}(y_i - \mu)^2}{\nu_0 + n}$.

- Recall that $\chi^2(\nu)$ is a special case of Gamma(α, β) distribution with $\alpha = \nu/2$ and $\beta = 1/2$.

- Therefore, we can also use an inv-Gamma prior for σ^2 or Gamma prior for precision $\gamma^2 = 1/\sigma^2$ as conjugate priors.

Univariate normal model

- When both μ and σ^2 are unknown, the only way to make the priors conjugate is to make the prior for μ dependent on σ^2 as follows:

$$\begin{aligned} \mu|\sigma^2 &\sim N(\mu_0, \sigma^2/k_0) \\ \sigma^2 &\sim \mathrm{Inv}-\chi^2(\nu_0, \sigma_0^2) \end{aligned}$$

- In general, I do not recommend this prior.

- As we will see later, we don't have to specify our prior this way.

Multinomial model

- This is a multiparameter generalization of binomial distribution.
- For example, in the election problem, we might have more than two candidates.
- If y has a multinomial distribution with J groups, the sampling distribution would have the following form

$$P(y|\theta) \propto \prod_{j=1}^{J} \theta_j^{y_j}$$

Multinomial model

- The conjugate prior for this model is the Dirichlet distribution

$$P(\theta|\alpha) \propto \prod_{j=1}^{J} \theta_j^{\alpha_j - 1} \qquad \theta_j \geq 0, \sum_{j=1}^{J} \theta_j = 1, \alpha_j > 0.$$

  which is a multivariate generalization of the Beta distribution.

- For this distribution, $E(\theta_j) = \alpha_j / \sum_{j'} \alpha_{j'}$

Multinomial model

- The posterior distribution is also a Dirichlet$(y_1 + \alpha_1, ..., y_J + \alpha_J)$ distribution

$$P(\theta|\alpha, y) \propto \prod_{j=1}^{J} \theta_j^{y_j+\alpha_j-1}$$

- If we do not want to use informative priors, similar to the binomial model, we could use a flat prior by setting $\alpha_j = 1$.

- Alternatively, we could use an improper prior by setting $\alpha_j = 0$. As long as we have at least one observation, the posterior would still be proper.

Multinomial model

- Consider the election example. Let's assume another candidate, C, enters the race and a new poll shows that out of 100 people surveyed 24 people vote for A, 45 for B, and 31 for C.
- Let's denote the probability of winning by θ_j, where $j \in \{A, B, C\} \equiv \{1, 2, 3\}$.
- Assume a flat Dirichlet prior with $\alpha_j = 1$.
- The posterior distribution of θ has a Dirichlet (25, 46, 32) distribution.
- The probability of winning (i.e., $E(\theta_j)$) for candidates A, B, C becomes 25/103, 46/103, and 32/103 respectively

Multivariate normal model

- For multivariate normal distribution with known covariance, Σ, we assume

$$\begin{aligned} \mathbf{x} &\sim N_p(\boldsymbol{\mu}, \Sigma) \\ \boldsymbol{\mu} &\sim N_p(\boldsymbol{\mu}_0, \Sigma_0) \end{aligned}$$

- The posterior distribution of $\boldsymbol{\mu}$ given n observations is also a multivariate normal distribution,

$$\boldsymbol{\mu}|\bar{\mathbf{x}} \sim N_p(\boldsymbol{\mu}_n, \Sigma_n)$$

  where

$$\begin{aligned} \boldsymbol{\mu}_n &= (\Sigma_0^{-1} + n\Sigma^{-1})^{-1}(\Sigma_0^{-1}\boldsymbol{\mu}_0 + n\Sigma^{-1}\bar{\mathbf{x}}) \\ \Sigma_n &= (\Sigma_0^{-1} + n\Sigma^{-1})^{-1} \end{aligned}$$

Multivariate normal model

- For multivariate normal distribution with known mean, $\boldsymbol{\mu}$, we assume

$$\begin{aligned} \mathbf{x} &\sim N_p(\boldsymbol{\mu}, \Sigma) \\ \Sigma &\sim \text{Inv-Wishart}(\nu_0, \Lambda_0) \end{aligned}$$

- The posterior distribution of Σ given n observations is also a multivariate normal distribution,

$$\Sigma | \bar{\mathbf{x}} \sim \text{Inv-Wishart}(\nu_n, \Lambda_n)$$

where

$$\begin{aligned} \nu_n &= \nu_0 + n \\ \Lambda_n &= \Lambda_0 + \sum_i (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\top \end{aligned}$$