Intro to Bayesian Statistics: Likelihoods, Priors & Posteriors

CSDE Workshop

Jessica Godwin

April 9, 2024

Resources and Support

Probability Basics

Likelihoods

Likelihoods, Priors & Posteriors

Resources and Support

Resources and Support

Texts

- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. (2013). Bayesian Data Analysis, 3rd ed. Chapman and Hall/CRC.
- McElreath, R. (2020). Statistical Rethinking: A Bayesian Course with Examples in R and Stan, 2nd ed. Chapman and Hall/CRC.
- Casella, G., & Berger, R. L. (2002). Statistical Inference, 2nd ed. Cengage Learning.

Probability Basics

Set Notation

A **set** is a collection of elements from a population.

Examples

- Positive Integers ≤ 5 : $A = \{1, 2, 3, 4, 5\}$
- Primary Colors: $B = \{\text{blue, red, yellow }\}$
- Odd Numbers: $C = \{1, 3, 5, 7, 9...\}$

Set Notation

A **set** is a collection of elements from a population.

Examples

- Positive Integers $\leq 5: \quad A = \{1, 2, 3, 4, 5\}$
- Primary Colors: $B = \{\text{blue, red, yellow }\}$
- Odd Numbers: $C = \{1, 3, 5, 7, 9...\}$

A set is an **empty set** if it contains no elements: written $\emptyset$ or $D = \{\emptyset\}$, e.g. integers that are greater than 4 and less than 1.

Set Notation

A **set** is a collection of elements from a population.

Examples

- Positive Integers ≤ 5 : $\quad A = \{1, 2, 3, 4, 5\}$
- Primary Colors: $B = \{$blue, red, yellow $\}$
- Odd Numbers: $C = \{1, 3, 5, 7, 9...\}$

A set is an **empty set** if it contains no elements: written $\emptyset$ or $D = \{\emptyset\}$, e.g. integers that are greater than 4 and less than 1.

A set is called the **universal set** if it contains all the elements in the population, denoted Ω or $E = \{\Omega\}$.

Intersections & Unions

- The **intersection** of two sets A, B is the set of all elements that are in A **AND** B.
  - The intersection is denoted $A \cap B$.
- The **union** of two sets A, B is the set of all elements that are in A **OR** B.
  - The intersection is denoted $A \cup B$.

Examples

- $A = \{1, 2, 3, 4, 5\}$, $B = \{2, 4, 6, 8, 10\} \Rightarrow A \cap B = \{2, 4\}$
  $A \cup B = \{1, 2, 3, 4, 5, 6, 8, 10\}$
- $A = \{$ Odd numbers $\}$, $B = \{$ Even numbers $\} \Rightarrow A \cap B = \{\emptyset\}$ $A \cup B = \{$All integers $\}$

Sample Spaces & Experiments

- An **experiment** is an action or process of observation.
- Once performed an experiment has only one **outcome**, but we do not know what it will be with certainty until the experiment is carried out.
  - e.g. rolling a dice or flipping a coin.
- The **sample space** is the set all the possible outcomes of the experiment and usually denoted by S.
  - If the experiment is rolling a die, $S = \{1, 2, 3, 4, 5, 6\}$.
  - If the experiment is flipping a coin, $S = \{$ Heads, Tails $\}$.

Events

An **event** is a subset of the sample space, and is a collection of one or more outcomes.

Suppose we flip one coin three times.

- **Sample Space:** $S = \{HHH, HHT, HTH, THH, TTH, THT, HTT, TTT\}$
  - Getting 2 heads: $\{HHT, HTH, THH\}$
  - Getting an odd number of tails: $\{HHT, HTH, THH, TTT\}$
  - Getting more than 1 head: $\{HHH, HHT, HTH, THH\}$

Events

An **event** is a subset of the sample space, and is a collection of one or more outcomes.

Suppose we flip one coin three times.

- **Sample Space:** $S = \{HHH, HHT, HTH, THH, TTH, THT, HTT, TTT\}$
  - Getting 2 heads: $\{HHT, HTH, THH\}$
  - Getting an odd number of tails: $\{HHT, HTH, THH, TTT\}$
  - Getting more than 1 head: $\{HHH, HHT, HTH, THH\}$

Suppose we flip one coin and roll one die.

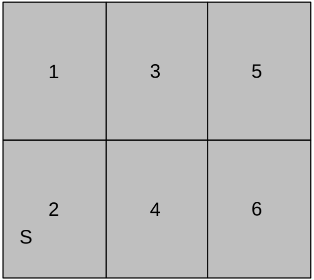
- **Sample Space:** $S = \{1H, 2H, 3H, 4H, 5H, 6H, 1T, 2T3T, 4T, 5T, 6T\}$
  - Rolling higher than a 4: $\{5H, 6H, 5T, 6T\}$
  - Getting a head: $\{1H, 2H, 3H, 4H, 5H, 6H\}$
  - Rolling a 3 or a 2: $\{2H, 3H, 2T, 3T\}$

Probability: Finite Sample Spaces & Equal Probability

When the sample space is finite and each outcome has equal probability, we can find the **probability** of an event by dividing the number of outcomes in the event by the size of the sample space.

Example - Rolling a fair die:

- $S = \{1, 2, 3, 4, 5, 6\}$
- $A = \{\text{roll} \leq 4\} = \{1, 2, 3, 4\} \Rightarrow P(A) = \frac{4}{6} = \frac{2}{3}$
- $B = \{\text{roll odd}\} = \{1, 3, 5\} \Rightarrow P(B) = \frac{3}{6} = \frac{1}{2}$



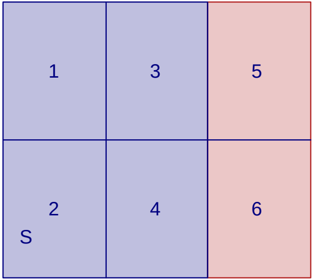
Probability: Unions & Mutually Exclusive Events

When the intersection of two events is the empty set, the two events are called **mutually exclusive** and

$$P(A \cup B) = P(A) + P(B).$$

Example - Rolling a fair die:

- $S = \{1, 2, 3, 4, 5, 6\}$
- $A = \{\text{roll} \leq 4\} = \{1, 2, 3, 4\} \Rightarrow P(A) = \frac{4}{6} = \frac{2}{3}$
- $B = \{\text{roll} \geq 5\} = \{5, 6\} \Rightarrow P(B) = \frac{2}{6} = \frac{1}{3}$
- $P(A \cup B) = P(A) + P(B) = \frac{2}{3} + \frac{1}{3} = 1 = \frac{6}{6}$



Probability: Unions

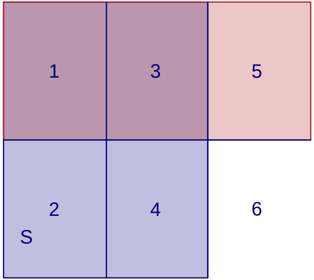
In general, the probability of the **union** of two events is

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Example - Rolling a fair die:

- $S = \{1, 2, 3, 4, 5, 6\}$
- $A = \{\text{roll} \leq 4\} \Rightarrow P(A) = \frac{4}{6}$
- $B = \{1, 3, 5\} \Rightarrow P(B) = \frac{3}{6}$
- $\Rightarrow A \cap B = \{1, 3\} \Rightarrow$ $P(A \cap B) = \frac{2}{6}$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{4}{6} + \frac{3}{6} - \frac{2}{6} = \frac{5}{6}$$



Probability: Conditional Probability

Sometimes knowing that one event has occurred changes what you know about the probability of another event. For example if the sidewalk is wet in the morning you might think it is more likely that it rained last night than if you didn't know anything about the sidewalk.

The **conditional probability** of A given B is the probability that A occurs given that B has been observed. It is denoted $P(A|B)$.

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

Note, this implies another important and often used relationship of probabilities:

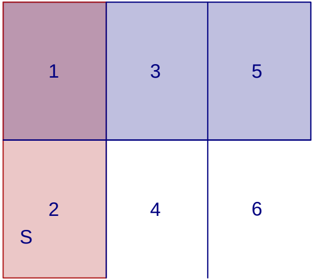
$$P(A \cap B) = P(A|B) \times P(B).$$

Probability: Conditional Probability - Example

What is the probability that your roll is odd if you know that it is less than 3?

- $A = \{\text{odd number}\} \Rightarrow P(A) = \dfrac{3}{6}$
- $B = \{\text{roll} < 3\} \Rightarrow P(B) = \dfrac{2}{6}$
- $A \cap B = \{1\} \Rightarrow (A \cap B) = \dfrac{1}{6}$

$$P(\text{Odd}| < 3) = \frac{P(\text{Odd} \cap < 3)}{P(< 3)} = \frac{1/6}{2/6} = \frac{1}{6} \times \frac{6}{2} = \frac{1}{2}.$$



Probability: Independence

What if knowing B does not give us any information about A? That is, if $P(A|B) = P(A)$, then we say that A and B are **independent**.

Independence also means:

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

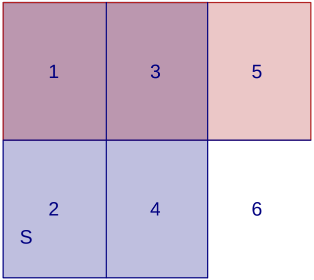
Thus, $P(A) \cdot P(B) = P(A \cap B)$ allows us to check for independence.

Probability: Independence - Example

Rolling a fair die:

- $A = \{1, 2, 3, 4\} \Rightarrow P(A) = \frac{4}{6}$
- $B = \{\text{odd number}\} \Rightarrow P(B) = \frac{3}{6}$
- $A \cap B = \{1, 3\} \Rightarrow P(A \cap B) = \frac{2}{6}$

$$P(A) \cdot P(B) = \frac{4}{6} \cdot \frac{3}{6} = \frac{12}{36} = \frac{2}{6}$$



Knowing that you have rolled a 1, 2, 3 or 4 doesn't give you any information about whether or not you rolled an odd number (because there are 2 even and 2 odd) and vice versa.

Probability: Bayes' Rule

Sometimes you may know one conditional probability, but not the other. How can you use the first conditional probability to find the other one?

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

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

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

Probability: Bayes' Rule - Testing Example

A disease has a prevalence of 1% in the population. A blood test for the disease has high sensitivity (the probability of a positive test if someone is sick) and specificity (the probability of a negative test if someone is not sick).

- If someone has the disease, there is a 98% chance they will test positive.
- If someone does not have the disease, there is a 95% chance they will test negative

Suppose you test positive for the disease and you want to figure out the probability that you have the disease. That is, given someone has tested positive for the disease what is the chance that they have the disease?

Probability: Bayes' Rule - Testing Example

A disease has a prevalence of 1% in the population. A blood test for the disease has high sensitivity (the probability of a positive test if someone is sick) and specificity (the probability of a negative test if someone is not sick).

- If someone has the disease, there is a 98% chance they will test positive.
- If someone does not have the disease, there is a 95% chance they will test negative

Suppose you test positive for the disease and you want to figure out the probability that you have the disease. That is, given someone has tested positive for the disease what is the chance that they have the disease?

What do we know?

Probability: Bayes' Rule - Testing Example

A disease has a prevalence of 1% in the population. A blood test for the disease has high sensitivity (the probability of a positive test if someone is sick) and specificity (the probability of a negative test if someone is not sick).

- If someone has the disease, there is a 98% chance they will test positive.
- If someone does not have the disease, there is a 95% chance they will test negative

Suppose you test positive for the disease and you want to figure out the probability that you have the disease. That is, given someone has tested positive for the disease what is the chance that they have the disease?

What do we know?

- $P(+|D^+) = 0.98, \ P(-|D^+) = 0.02$

Probability: Bayes' Rule - Testing Example

A disease has a prevalence of 1% in the population. A blood test for the disease has high sensitivity (the probability of a positive test if someone is sick) and specificity (the probability of a negative test if someone is not sick).

- If someone has the disease, there is a 98% chance they will test positive.
- If someone does not have the disease, there is a 95% chance they will test negative

Suppose you test positive for the disease and you want to figure out the probability that you have the disease. That is, given someone has tested positive for the disease what is the chance that they have the disease?

What do we know?

- $P(+|D^{+}) = 0.98,\ P(-|D^{+}) = 0.02$
- $P(-|D^{-}) = 0.95,\ P(+|D^{-}) = 0.05$

Probability: Bayes' Rule - Testing Example

A disease has a prevalence of 1% in the population. A blood test for the disease has high sensitivity (the probability of a positive test if someone is sick) and specificity (the probability of a negative test if someone is not sick).

- If someone has the disease, there is a 98% chance they will test positive.
- If someone does not have the disease, there is a 95% chance they will test negative

Suppose you test positive for the disease and you want to figure out the probability that you have the disease. That is, given someone has tested positive for the disease what is the chance that they have the disease?

What do we know?

- $P(+|D^+) = 0.98$, $P(-|D^+) = 0.02$
- $P(-|D^-) = 0.95$, $P(+|D^-) = 0.05$
- $P(D^+) = 0.01$, $P(D^-) = 0.99$

Probability: Bayes' Rule - Testing Example

A disease has a prevalence of 1% in the population. A blood test for the disease has high sensitivity (the probability of a positive test if someone is sick) and specificity (the probability of a negative test if someone is not sick).

- If someone has the disease, there is a 98% chance they will test positive.
- If someone does not have the disease, there is a 95% chance they will test negative

Suppose you test positive for the disease and you want to figure out the probability that you have the disease. That is, given someone has tested positive for the disease what is the chance that they have the disease?

What do we know?

- $P(+|D^+) = 0.98$, $P(-|D^+) = 0.02$
- $P(-|D^-) = 0.95$, $P(+|D^-) = 0.05$
- $P(D^+) = 0.01$, $P(D^-) = 0.99$

Probabilty: Bayes' Rule - Testing Example

- $P(+|D^+) = 0.98,\ P(-|D^+) = 0.02$
- $P(-|D^-) = 0.95,\ P(+|D^-) = 0.05$
- $P(D^+) = 0.01,\ P(D^-) = 0.99$

$$
\begin{aligned}
P(D^+|+) = \frac{P(+|D^+)P(D^+)}{P(+)} &= \frac{P(+|D^+)P(D^+)}{P(+\cap D^+) + P(+\cap D^-)} \\
&= \frac{P(+|D^+)P(D^+)}{P(+|D^+)P(D^+) + P(+|D^-)P(D^-)} \\
&= \frac{0.98\cdot 0.01}{0.98\cdot 0.01 + 0.05\cdot 0.99} \\
&= 0.165
\end{aligned}
$$

Likelihoods

Random Variables

A **random variable** is a function which assigns a number to each element in the sample space. (Think of it as the answer to a question you are asking about each element in the space).

- **Variable**, because the answer will be different for each element.
- **Random** because we can't predict the answer with any great certainty.

Random variables are usually denoted with capital letters, X, Y, Z. Examples:

- Roll a die $\Rightarrow S = \{1, 2, \ldots, 6\}$
  - $X =$ the number rolled
  - Possible values: $1, 2, \ldots, 6$
- Roll two dice: $\Rightarrow S = \{(1,1), (1,2), \ldots, (6,5), (6,6)\}$
  - $X =$ sum of the dice
  - Possible values: $2, 3, 4, \ldots, 12$

Probability Distribution

The **probability distribution** of a random variable is a function that assigns a probability to each possible value of X.

The probability distribution can be written as:

X-Value	x_1	x_2	...	x_n
$P(X = x_i)$	p_1	p_2	...	p_n

where each possible value x_i for X is listed with its probability p_i. Find each p_i by summing the probabilities of the elements such that $X = x_i$.

Example: If we roll a die and let X be the number on rolled, then

X-Value	1	2	3	4	5	6
$P(X = x_i)$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$

Probability Mass Functions

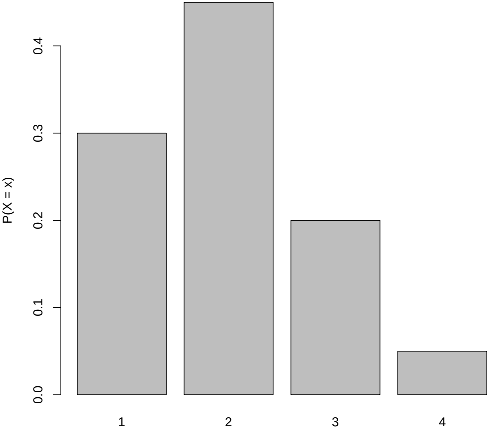
The **probability distribution function** is defined differently for discrete and continuous random variables.

Discrete Random Variables

- **probability mass function** (pmf)
- Countable number of outcomes n, can write down

$$P(X = x_i) \ \forall \, x_i, \quad i = 1, \ldots, n$$

.

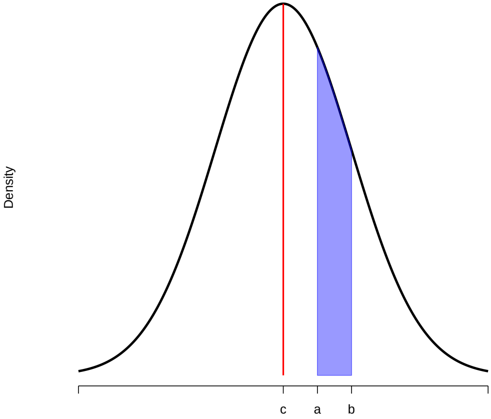


Probability Density Functions

The **probability distribution function** is defined differently for discrete and continuous random variables.

Continuous Random Variables

- **probability density function** (pdf)
- Too many (∞) possible values to write down.
- $P(X = c) = 0$ for any value of x.
- $P(a < X < b) > 0$, if a and b are values X can take on.



Likelihood or pmf?

Let X be a Binomial random variable, with n independent trials and probability of success in each trial p. The pmf for X is

$$P(X|n,p) = \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}.$$

Now, suppose we flip a coin $n = 100$ times and get 57 heads, so $X = 57$. What is different?

Likelihood or pmf?

Let X be a Binomial random variable, with n independent trials and probability of success in each trial p. The pmf for X is

$$P(X|n,p) = \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}.$$

Now, suppose we flip a coin $n = 100$ times and get 57 heads, so $X = 57$. What is different?

$$L(p|X,n) = \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x} = \frac{100!}{57!(100-57)!}p^{57}(1-p)^{100-57}$$

Likelihood or pmf?

Let X be a Binomial random variable, with n independent trials and probability of success in each trial p. The pmf for X is

$$P(X|n,p) = \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}.$$

Now, suppose we flip a coin $n = 100$ times and get 57 heads, so $X = 57$. What is different?

$$L(p|X,n) = \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x} = \frac{100!}{57!(100-57)!}p^{57}(1-p)^{100-57}$$

In a pmf, the **parameters** of the distribution are known and the value of the random variable or outcome of the experiment is unknown. In a **likelihood**, the value of the random variable is known, but the parameter is not.

Likelihood or pdf?

Let X_i, $i = 1, \ldots, n$ be independent, identically distribution normal random variables, with mean μ and variance σ^2. The pdf for each X_i is

$$P(X_i|\mu, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{(x_i - \mu)^2}{2\sigma^2}\right\}.$$

However, when we have n, X_i we observed, then

$$\begin{aligned} L(\mu, \sigma^2|X_i, i = 1, \ldots, n) &= \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{(x_i - \mu)^2}{2\sigma^2}\right\} \\ &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2}\right\} \end{aligned}$$

Likelihoods, Priors & Posteriors

Bayes's Rule and a Bayesian model

So what does Bayes' rule have to do with Bayesian statistics? Recall,

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

Suppose we have observations y_i, $i = 1, \ldots, n$ and we assume they come from a probability distribution with parameters θ. Our inference goal is to learn something about θ from our observed data.

$$\underbrace{P(\theta|y)}_{\text{posterior}} = \frac{\overbrace{P(y|\theta)}^{\text{likelihood}}\ \overbrace{P(\theta)}^{\text{prior}}}{P(y)}.$$

Bayesian modeling - Testing Example

Returning to our testing example, we can consider each test a Bernoulli random variable where $X_i = 1$ with probability p. If we use the population prevalence of the disease, $P(D^+)$ as a **prior** on our probability of testing positive, then

$$\underbrace{P(D^+|X_i = 1)}_{\text{posterior}} = \frac{\overbrace{P(X_i = 1|D^+)}^{\text{likelihood}}\,\overbrace{P(D^+)}^{\text{prior}}}{\underbrace{P(X_i) = P(X_i = 1|D^+)P(D^+) + P(X_i = 1|D^-)P(D^-)}_{\text{marginal distribution of the data}}}.$$

Bayesian modeling

- In general, $P(\theta)$ is a probability distribution over possible values of θ. It can be uninformative or informed by prior knowledge.
- The hardest part of Bayesian statistics is $P(y)$, often referred to as the **normalizing constant** or the **marginal distribution** of the data or, sometimes, the **prior predictive distribution**. $P(y)$ is the probability of seeing your data accounting for all possible values of θ in the prior.

Discrete

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

Continuous

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

This was easy for us to calculate in our testing example when our parameter which determined our data (disease presence) had only 2 values, but often this piece is not possible to compute by hand.

Conjugacy

- The property that the posterior distribution follows the same parametric form as the prior distribution is called **conjugacy**.
- The prior and posterior distributions that have this property with a particular likelihood are called a **conjugate family** to the likelihood.

Examples of conjugate families:

Likelihood	**Conjugate family**
Binomial	Beta
Multinomial	Dirichlet
Poisson	Gamma
Exponential	Gamma
Normal (mean)	Normal
Normal (mean, variance)	Normal (mean), Inverse Gamma (variance)

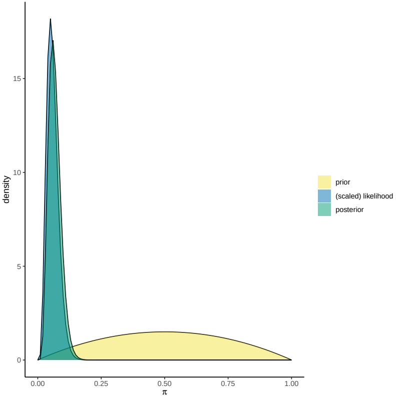
The Beta-Binomial Model

Suppose we test $n = 100$ individuals for our disease we know as a 0.01 prevalence in the population, and $y = 5$ individuals test positive. If we put a $\Gamma(\alpha = 2, \beta = 2)$ prior on p, then

$$P(y|p,n) = \frac{100!}{5!95!} p^5 (1-p)^{100-5}$$

$$P(p) = \frac{\Gamma(2+2)}{\Gamma(2)\Gamma(2)} p^{2-1} (1-p)^{2-1}$$

$$P(p|y,n) = \frac{\Gamma((2+5)(2+100-5))}{\Gamma(2+5)\Gamma(2+100-5)} p^{2+5-1} (1-p)^{2+100-5-1}$$



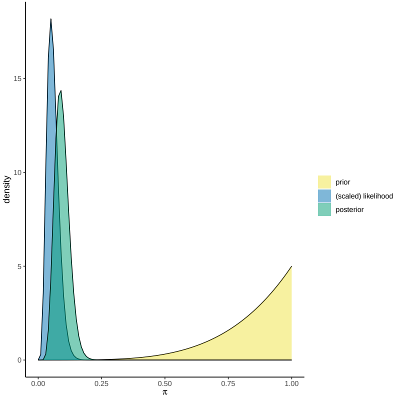
The Beta-Binomial Model

Suppose we test $n = 100$ individuals for our disease we know as a 0.01 prevalence in the population, and $y = 5$ individuals test positive. If we put a $\Gamma(\alpha = 5, \beta = 1)$ prior on p, then

$$P(y|p,n) = \frac{100!}{5!95!} p^{5}(1-p)^{100-5}$$

$$P(p) = \frac{\Gamma(5+1)}{\Gamma(5)\Gamma(1)} p^{5-1}(1-p)^{1-1}$$

$$P(p|y,n) = \frac{\Gamma((5+5)(1+100-5))}{\Gamma(5+5)\Gamma(1+100-5)} p^{5+5-1}(1-p)^{1+100-5-1}$$



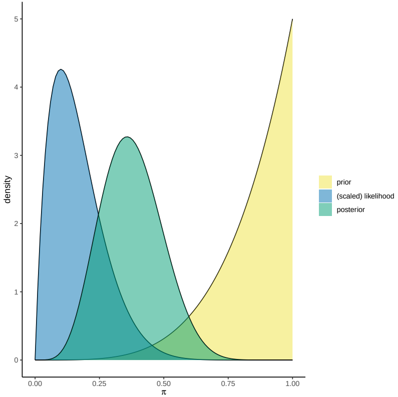
The Beta-Binomial Model

Suppose we test $n = 10$ individuals for our disease we know as a 0.01 prevalence in the population, and $y = 1$ individuals test positive. If we put a $\Gamma(\alpha = 5, \beta = 1)$ prior on p, then

$$P(y|p,n) = \frac{10!}{1!9!} p^{1}(1-p)^{10-1}$$

$$P(p) = \frac{\Gamma(5+1)}{\Gamma(5)\Gamma(1)} p^{5-1}(1-p)^{1-1}$$

$$P(p|y,n) = \frac{\Gamma((5+1)(1+10-1))}{\Gamma(5+1)\Gamma(1+10-1)} p^{5+1-1}(1-p)^{1+10-1-1}$$



The Beta-Binomial Model

Let $X_1, \ldots, X_n$ be iid Bernoulli(p), so that $Y = \sum_{i=1}^n X_i \sim \text{Binomial}(n, p)$. If we assume $p \sim \text{Beta}(\alpha, \beta)$, what is the distribution of $p|y, n$.

$$\begin{aligned}
P(p|y,n) &= \frac{P(y|n,p)P(p)}{P(y)} \\
&\propto P(y|n,p)P(p) \text{ [drop denominator for now]} \\
&= \underbrace{\frac{n!}{y!(n-y)!}p^y(1-p)^{n-y}}_{\text{likelihood}} \times \underbrace{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta-1}}_{\text{prior}} \\
&\propto p^y(1-p)^{n-y} \times p^{\alpha-1}(1-p)^{\beta-1} \text{ [drop terms without } p\text{]}
\end{aligned}$$

The Beta-Binomial Model

Let $X_1, \ldots, X_n$ be iid Bernoulli(p), so that $Y = \sum_{i=1}^{n} X_i \sim$ Binomial(n, p). If we assume $p \sim$ Beta(α, β), what is the distribution of $p|y, n$.

$$p|y, n \sim \text{Beta}(\alpha + y, \beta + n - y)$$

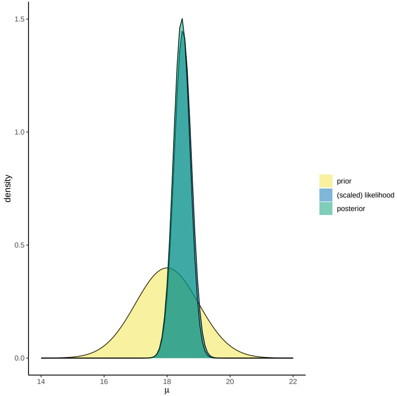
$$E[p|y, n] = \frac{\alpha + y}{\alpha + y + \beta + n - y} = \frac{\alpha + y}{\alpha + \beta + n}$$

$$Var(p|y, n) = \frac{(\alpha + y)(\beta + n - y)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)}$$

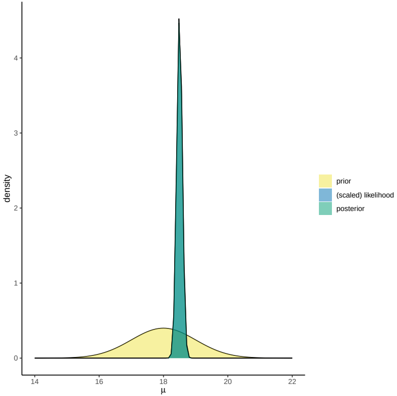
The Normal-Normal Model

Suppose I ask another person what their prior is on the age distribution in an undergraduate course I am teaching at UW, and tell them it is a freshman course. They say Normal($\mu = 18, \sigma^2 = 1$). I ask $n = 10$ or $n = 100$ students in my class their age and $\bar{y} = 18.5$ and $s = 0.86$.

$n = \mathbf{10}$

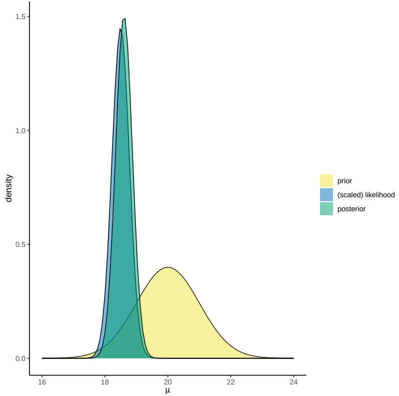


$n = \mathbf{100}$

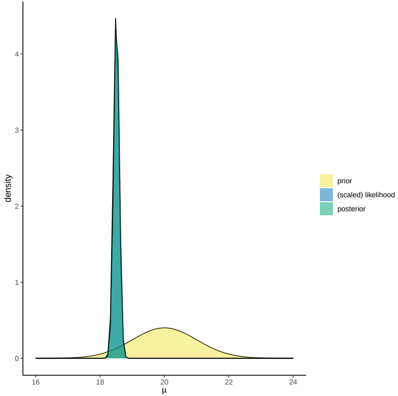


The Normal-Normal Model

Suppose I ask someone what their prior is on the age distribution in an undergraduate course I am teaching at UW. They say Normal$(\mu = 20, \sigma^2 = 1)$. I ask $n = 10$ or $n = 100$ students in my class their age and $\bar{y} = 18.5$ and $s = 0.86$.



$\boldsymbol{n = 10}$

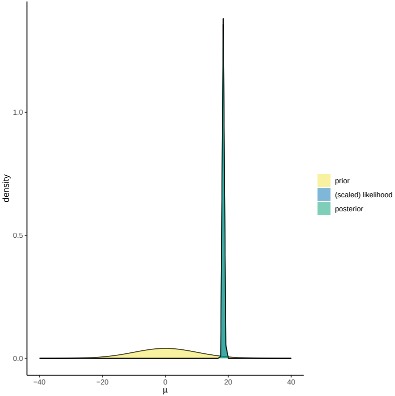


$\boldsymbol{n = 100}$

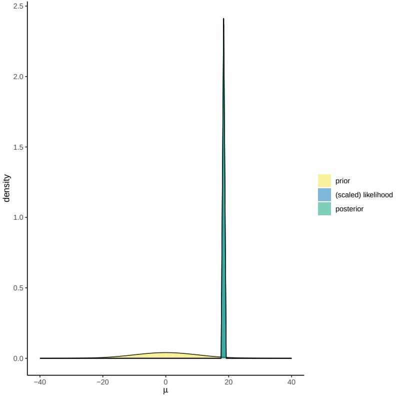
The Normal-Normal Model

Suppose I ask someone what their prior is on the age distribution in an undergraduate course I am teaching at UW. They say Normal$(\mu = 0, \sigma^2 = 10)$. I ask $n = 10$ or $n = 100$ students in my class their age and $\bar{y} = 18.5$ and $s = 0.86$.

$n = 10$



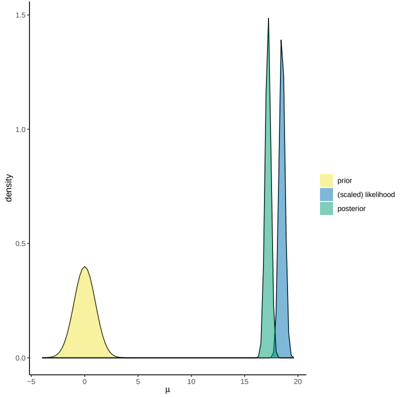
$n = 100$



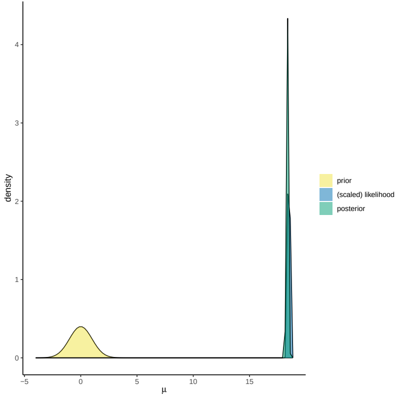
The Normal-Normal Model

Suppose I ask someone what their prior is on the age distribution in an undergraduate course I am teaching at UW. They say Normal$(\mu = 0, \sigma^2 = 1)$. I ask $n = 10$ or $n = 100$ students in my class their age and $\bar{y} = 18.5$ and $s = 0.86$.

$\boldsymbol{n = 10}$



$\boldsymbol{n = 100}$



The Normal-Normal Model

Let $Y \sim \text{Normal}(\mu, \sigma^2)$, where σ^2 is known. If we assume $\mu \sim \text{Normal}(\theta, \tau^2)$, where θ and τ^2 are known values, what is the distribution of $\mu|y, \sigma^2, \theta, \tau^2$?

$$\mu|y, \sigma^2, \theta, \tau^2 \sim \text{Normal}\left(y \times \frac{\tau^2}{\sigma^2 + \tau^2} + \theta \times \frac{\sigma^2}{\sigma^2 + \tau^2}, \frac{\tau^2\sigma^2}{\tau^2 + \sigma^2}\right)$$

$$E[\mu|y, \sigma^2, \theta, \tau^2] = y \times \frac{\tau^2}{\sigma^2 + \tau^2} + \theta \times \frac{\sigma^2}{\sigma^2 + \tau^2}$$

$$Var(\mu|y, \sigma^2, \theta, \tau^2) = \frac{\tau^2\sigma^2}{\tau^2 + \sigma^2}$$

The Normal-Normal Model

Let $Y \sim \text{Normal}(\mu, \sigma^2)$, where σ^2 is known. If we assume $\mu \sim \text{Normal}(\theta, \tau^2)$, where θ and τ^2 are known values, what is the distribution of $\mu | y, \sigma^2, \theta, \tau^2$?

$$\begin{aligned}
P(\mu | y, \sigma^2, \theta, \tau^2) &= \frac{P(y|\mu, \sigma^2) P(\mu|\theta, \tau^2)}{P(y)} \\
&\propto P(y|\mu, \sigma^2) P(\mu|\theta, \tau^2) \text{ [Drop denominator for now]} \\
&= \underbrace{(2\pi\sigma^2)^{-1/2} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)}_{\text{likelihood}} \times \underbrace{(2\pi\tau^2)^{-1/2} \exp\left(-\frac{(\mu-\theta)^2}{2\tau^2}\right)}_{\text{prior}} \\
&\propto \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \times \exp\left(-\frac{(\mu-\theta)^2}{2\tau^2}\right) \text{ [Drop terms without } \mu\text{]}
\end{aligned}$$

The Normal-Normal Model

Let $Y \sim \text{Normal}(\mu, \sigma^2)$, where σ^2 is known. If we assume $\mu \sim \text{Normal}(\theta, \tau^2)$, where θ and τ^2 are known values, what is the distribution of $\mu | y, \sigma^2, \theta, \tau^2$?

$$\begin{aligned}
P(\mu|y, \sigma^2, \theta, \tau^2) &\propto \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \times \exp\left(-\frac{(\mu-\theta)^2}{2\tau^2}\right) \text{ [Drop terms without } \mu] \\
&= \exp\left(-\frac{y^2 - 2y \times \mu + \mu^2}{2\sigma^2} - \frac{\mu^2 - 2\mu \times \theta + \theta^2}{2\tau^2}\right) \text{ [Expand squared terms]} \\
&\propto \exp\left(-\frac{-2y \times \mu + \mu^2}{2\sigma^2} - \frac{\mu^2 - 2\mu \times \theta}{2\tau^2}\right) \text{ [Drop terms without } \mu]
\end{aligned}$$

The Normal-Normal Model

Let $Y \sim \text{Normal}(\mu, \sigma^2)$, where σ^2 is known. If we assume $\mu \sim \text{Normal}(\theta, \tau^2)$, where θ and τ^2 are known values, what is the distribution of $\mu|y, \sigma^2, \theta, \tau^2$?

$$
\begin{aligned}
P(\mu|y,\sigma^2,\theta,\tau^2) &\propto \exp\left(-\frac{-2y \times \mu + \mu^2}{2\sigma^2} - \frac{\mu^2 - 2\mu \times \theta}{2\tau^2}\right) \quad \text{[Drop terms without } \mu\text{]} \\
&= \exp\left(-\frac{-2y \times \mu + \mu^2}{2\sigma^2} \times \frac{\tau^2}{\tau^2} - \frac{\mu^2 - 2\mu \times \theta}{2\tau^2} \times \frac{\sigma^2}{\sigma^2}\right) \quad \text{[Multiply by fancy 1]} \\
&= \exp\left(-\frac{-2y\tau^2 \times \mu - 2\theta\sigma^2 \times \mu}{2\tau^2\sigma^2}\right) \\
&\times \exp\left(-\frac{\sigma^2 \times \mu^2 + \tau^2 \times \mu^2}{2\tau^2\sigma^2}\right) \quad \text{[Combine terms over LCD]}
\end{aligned}
$$

The Normal-Normal Model

Let $Y \sim \text{Normal}(\mu, \sigma^2)$, where σ^2 is known. If we assume $\mu \sim \text{Normal}(\theta, \tau^2)$, where θ and τ^2 are known values, what is the distribution of $\mu | y, \sigma^2, \theta, \tau^2$?

$$
\begin{aligned}
P(\mu | y, \sigma^2, \theta, \tau^2) &\propto \exp\left(-\frac{-2y\tau^2 \times \mu - 2\theta\sigma^2 \times \mu}{2\tau^2\sigma^2}\right) \\
&\quad \times \exp\left(-\frac{\sigma^2 \times \mu^2 + \tau^2 \times \mu^2}{2\tau^2\sigma^2}\right) \quad \text{[Combine terms over LCD]} \\
&= \exp\left(-\frac{(-2y\tau^2 - 2\theta\sigma^2) \times \mu + (\sigma^2 + \tau^2) \times \mu^2}{2\sigma^2\tau^2}\right) \quad \text{[Combine like terms in } \mu\text{]} \\
&= \exp\left(-\frac{B \times \mu + A \times \mu^2}{2\sigma^2\tau^2}\right) \quad \text{[Combine like terms in } \mu\text{]}
\end{aligned}
$$

An aside: Completing the Square

Suppose we have the following quadratic and linear terms in a variable x:

$$0 = B \times x + A \times x^2.$$

We can convert that into a term that looks like $(x + \text{constant})^2$, also known as **completing the square** with the following steps:

$$\begin{aligned} 0 &= B \times x + A \times x^2 \\ &= \frac{B}{A} \times x + x^2 \text{ [Make the coefficient on } x^2 = 1] \\ &= \frac{B}{A} \times x + x^2 + \underbrace{(\frac{1}{2} \times \frac{B}{A})^2 - (\frac{1}{2} \times \frac{B}{A})^2}_{=0=\text{constant}^2-\text{constant}^2} \text{ [Add fancy 0]} \end{aligned}$$

An aside: Completing the Square

We can convert $0 = B \times x + A \times x^2$ into a term that looks like $(x + \text{constant})^2$, also known as **completing the square** with the following steps:

$$
\begin{aligned}
0 &= B \times x + A \times x^2 \\
&= \frac{B}{A} \times x + x^2 \text{ [Make the coefficient on } x^2 = 1\text{]} \\
&= \frac{B}{A} \times x + x^2 + (\frac{1}{2} \times \frac{B}{A})^2 - (\frac{1}{2} \times \frac{B}{A})^2 \text{ [Add fancy 0]} \\
&= (x + \frac{1}{2} \times \frac{B}{A})^2 - (\frac{1}{2} \times \frac{B}{A})^2 \text{ [Rewrite as perfect square]} \\
&= (x + \text{constant})^2 - \text{constant}^2
\end{aligned}
$$

The Normal-Normal Model

Let $Y \sim \text{Normal}(\mu, \sigma^2)$, where σ^2 is known. If we assume $\mu \sim \text{Normal}(\theta, \tau^2)$, where θ and τ^2 are known values, what is the distribution of $\mu | y, \sigma^2, \theta, \tau^2$?

$$
\begin{aligned}
P(\mu|y,\sigma^2,\theta,\tau^2) &\propto \exp\left(-\frac{(-2y\tau^2 - 2\theta\sigma^2) \times \mu + (\sigma^2 + \tau^2) \times \mu^2}{2\sigma^2\tau^2}\right) \quad \text{[Combine like terms in } \mu\text{]} \\
&= \exp\left(-\frac{B \times \mu + A \times \mu^2}{2\sigma^2\tau^2}\right) \quad \text{[Combine like terms in } \mu\text{]} \\
&= \exp\left(-\frac{\frac{B}{A} \times \mu + \mu^2}{2\frac{\sigma^2\tau^2}{A}}\right) \quad \text{[Divide through by A]} \\
&= \exp\left(-\frac{(\frac{1}{2}\frac{B}{A})^2 - (\frac{1}{2}\frac{B}{A})^2 + \frac{B}{A} \times \mu + \mu^2}{2\frac{\sigma^2\tau^2}{A}}\right) \quad \text{[Add fancy 0]}
\end{aligned}
$$

The Normal-Normal Model

Let $Y \sim \text{Normal}(\mu, \sigma^2)$, where σ^2 is known. If we assume $\mu \sim \text{Normal}(\theta, \tau^2)$, where θ and τ^2 are known values, what is the distribution of $\mu | y, \sigma^2, \theta, \tau^2$?

$$P(\mu | y, \sigma^2, \theta, \tau^2) \propto \exp\left(- \frac{(\frac{1}{2}\frac{B}{A})^2 - (\frac{1}{2}\frac{B}{A})^2 + \frac{B}{A} \times \mu + \mu^2}{2\sigma^2\tau^2/A} \right) \text{ [Add fancy 0]}$$

$$= \exp\left(- \frac{1}{2\sigma^2\tau^2/A} \times \left[(\frac{1}{2}\frac{-2y\tau^2 - 2\theta\sigma^2}{\sigma^2 + \tau^2})^2 + \frac{-2y\tau^2 - 2\theta\sigma^2}{\sigma^2 + \tau^2} \times \mu + \mu^2 \right] \right)$$

$$\times \exp(\left(- \frac{1}{2\sigma^2\tau^2/A} \times \left[-(\frac{1}{2}\frac{-2y\tau^2 - 2\theta\sigma^2}{\sigma^2 + \tau^2})^2 \right] \right)$$

The Normal-Normal Model

Let $Y \sim$ Normal(μ, σ^2), where σ^2 is known. If we assume $\mu \sim$ Normal(θ, τ^2), where θ and τ^2 are known values, what is the distribution of $\mu | y, \sigma^2, \theta, \tau^2$?

$$\begin{aligned}
P(\mu|y,\sigma^2,\theta,\tau^2) &\propto \exp\left(-\frac{1}{2\sigma^2\tau^2/A} \times \left[(\frac{1}{2}\frac{-2y\tau^2-2\theta\sigma^2}{\sigma^2+\tau^2})^2 + \frac{-2y\tau^2-2\theta\sigma^2}{\sigma^2+\tau^2} \times \mu + \mu^2\right]\right) \\
&\times \exp(\left(-\frac{1}{2\sigma^2\tau^2/A} \times \left[-(\frac{1}{2}\frac{-2y\tau^2-2\theta\sigma^2}{\sigma^2+\tau^2})^2\right]\right) \\
&= \exp\left(-\frac{\left[\mu + (\frac{1}{2}\frac{-2y\tau^2-2\theta\sigma^2}{\sigma^2+\tau^2})\right]^2 - (\frac{1}{2}\frac{-2y\tau^2-2\theta\sigma^2}{\sigma^2+\tau^2})^2}{2\sigma^2\tau^2/A}\right)
\end{aligned}$$

The Normal-Normal Model

Let $Y \sim \text{Normal}(\mu, \sigma^2)$, where σ^2 is known. If we assume $\mu \sim \text{Normal}(\theta, \tau^2)$, where θ and τ^2 are known values, what is the distribution of $\mu | y, \sigma^2, \theta, \tau^2$?

$$P(\mu|y, \sigma^2, \theta, \tau^2) \propto \exp\left(-\frac{\left[\mu + (\frac{1}{2}\frac{-2y\tau^2 - 2\theta\sigma^2}{\sigma^2 + \tau^2})\right]^2 - (\frac{1}{2}\frac{-2y\tau^2 - 2\theta\sigma^2}{\sigma^2 + \tau^2})^2}{2\sigma^2\tau^2/A} \right)$$

$$= \exp\left(-\frac{\left[\mu - (\frac{y\tau^2 + \theta\sigma^2}{\sigma^2 + \tau^2})\right]^2 - (\frac{1}{2}\frac{-2y\tau^2 - 2\theta\sigma^2}{\sigma^2 + \tau^2})^2}{2\sigma^2\tau^2/A} \right)$$

The Normal-Normal Model

Let $Y \sim \text{Normal}(\mu, \sigma^2)$, where σ^2 is known. If we assume $\mu \sim \text{Normal}(\theta, \tau^2)$, where θ and τ^2 are known values, what is the distribution of $\mu|y, \sigma^2, \theta, \tau^2$?

$$P(\mu|y, \sigma^2, \theta, \tau^2) \propto \exp\left(-\frac{\left[\mu - \left(\frac{y\tau^2 + \theta\sigma^2}{\sigma^2 + \tau^2}\right)\right]^2 - \left(\frac{1}{2}\frac{-2y\tau^2 - 2\theta\sigma^2}{\sigma^2 + \tau^2}\right)^2}{2\sigma^2\tau^2/A} \right)$$

$$\propto \exp\left(-\frac{\left[\mu - \left(\frac{y\tau^2 + \theta\sigma^2}{\sigma^2 + \tau^2}\right)\right]^2}{2\sigma^2\tau^2/(\sigma^2 + \tau^2)} \right) \quad \text{[Drop terms not related to } \mu]$$

The Normal-Normal Model

Let $Y \sim \text{Normal}(\mu, \sigma^2)$, where σ^2 is known. If we assume $\mu \sim \text{Normal}(\theta, \tau^2)$, where θ and τ^2 are known values, what is the distribution of $\mu | y, \sigma^2, \theta, \tau^2$?

$$P(\mu | y, \sigma^2, \theta, \tau^2) \propto \exp\left(- \frac{\left[\mu - \frac{y\tau^2 + \theta\sigma^2}{\sigma^2 + \tau^2}\right]^2}{2\sigma^2\tau^2/(\sigma^2 + \tau^2)} \right)$$

So the posterior distribution of $\mu | y, \sigma^2, \theta, \tau^2$ is

$$\mu \sim \text{Normal}\left(\frac{y\tau^2 + \theta\sigma^2}{\sigma^2 + \tau^2}, \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2} \right)$$

The Normal-Normal Model

Let $Y \sim \text{Normal}(\mu, \sigma^2)$, where σ^2 is known. If we assume $\mu \sim \text{Normal}(\theta, \tau^2)$, where θ and τ^2 are known values, what is the distribution of $\mu | y, \sigma^2, \theta, \tau^2$?

$$\mu | y, \sigma^2, \theta, \tau^2 \sim \text{Normal}\left(y \times \frac{\tau^2}{\sigma^2 + \tau^2} + \theta \times \frac{\sigma^2}{\sigma^2 + \tau^2}, \frac{\tau^2 \sigma^2}{\tau^2 + \sigma^2} \right)$$

$$E[\mu | y, \sigma^2, \theta, \tau^2] = y \times \frac{\tau^2}{\sigma^2 + \tau^2} + \theta \times \frac{\sigma^2}{\sigma^2 + \tau^2}$$

$$Var(\mu | y, \sigma^2, \theta, \tau^2) = \frac{\tau^2 \sigma^2}{\tau^2 + \sigma^2}$$

The Normal-Normal Model

Let $Y_1, ;Y_n \overset{iid}{\sim}$ Normal(μ, σ^2), where σ^2 is known. If we assume $\mu \sim$ Normal(θ, τ^2), where θ and τ^2 are known values, what is the distribution of $\mu | y, \sigma^2, \theta, \tau^2$?

$$\begin{aligned}
P(\mu|\mathbf{y}, \sigma^2, \theta, \tau^2) &= \frac{\prod_{i=1^n} P(y_i|\mu, \sigma^2) P(\mu|\theta, \tau^2)}{P(\mathbf{y})} \\
&\propto \prod_{i=1^n} P(y|\mu, \sigma^2) P(\mu|\theta, \tau^2) \text{ [drop denominator for now]} \\
&= \underbrace{\prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{(y_i - \mu)^2}{2\sigma^2}\right)}_{\text{likelihood}} \times \underbrace{(2\pi\tau^2)^{-1/2} \exp\left(-\frac{(\mu - \theta)^2}{2\tau^2}\right)}_{\text{prior}} \\
&\propto \exp\left(-\frac{\sum_{i=1}^{n} (y_i - \mu)^2}{2\sigma^2}\right) \times \exp\left(-\frac{(\mu - \theta)^2}{2\tau^2}\right) \text{ [drop terms without } \mu \text{]}
\end{aligned}$$

The Normal-Normal Model

Let $Y_1, \ldots, Y_n \overset{iid}{\sim} \text{Normal}(\mu, \sigma^2)$, where σ^2 is known. If we assume $\mu \sim \text{Normal}(\theta, \tau^2)$, where θ and τ^2 are known values, what is the distribution of $\mu | y, \sigma^2, \theta, \tau^2$?

$$\mu | \mathbf{y}, \sigma^2, \theta, \tau^2 \sim \text{Normal}\left(\bar{y} \times \frac{n\tau^2}{\sigma^2 + n\tau^2} + \theta \times \frac{\sigma^2}{\sigma^2 + n\tau^2}, \frac{\tau^2\sigma^2}{n\tau^2 + \sigma^2}\right)$$

$$E[\mu | \mathbf{y}, \sigma^2, \theta, \tau^2] = \bar{y} \times \frac{n\tau^2}{\sigma^2 + n\tau^2} + \theta \times \frac{\sigma^2}{\sigma^2 + n\tau^2}$$

$$Var(\mu | \mathbf{y}, \sigma^2, \theta, \tau^2) = \frac{\tau^2\sigma^2}{n\tau^2 + \sigma^2}$$

The Normal-Normal Model

Let $Y_1, \ldots, Y_n \overset{iid}{\sim} \text{Normal}(\beta_0 + \beta_1 X_i, \sigma^2)$, where σ^2 is known. If we assume $\beta \sim \text{Normal}(\theta, \Sigma_\theta)$, where $\theta = [\ \theta_0 \quad \theta_1\ ]$ and Σ_θ are known values, what is the distribution of $\beta | y, \sigma^2, \theta, \Sigma_\theta$ where $\beta = [\ \beta_0 \quad \beta_1\ ]$?

$$
\begin{aligned}
P(\beta|\mathbf{y}, \mathbf{x}, \sigma^2, \theta, \Sigma_\theta) &= \frac{\prod_{i=1^n} P(y_i|x_i, \beta\sigma^2)P(\beta|\theta, \Sigma_\theta)}{P(\mathbf{y})} \\
&\propto \prod_{i=1^n} P(y|x_i, \beta, \sigma^2)P(\beta|\theta, \Sigma_\theta) \text{ [drop denominator for now]} \\
&= \underbrace{\prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{(y_i - x_i^T\beta)^2}{2\sigma^2}\right)}_{\text{likelihood}} \\
&\times \underbrace{(2\pi|\Sigma_\theta|)^{-1/2} \exp\left(-\frac{1}{2}(\beta - \theta)^T \Sigma_\theta^{-1}(\beta - \theta)\right)}_{\text{prior}}
\end{aligned}
$$

The Normal-Normal Model

Let $Y_1, \ldots, Y_n \overset{iid}{\sim} \text{Normal}(\beta_0 + \beta_1 X_i, \sigma^2)$, where σ^2 is known. If we assume $\beta \sim \text{Normal}(\theta, \Sigma_\theta)$, where $\theta = \begin{bmatrix} \theta_0 & \theta_1 \end{bmatrix}$ and Σ_θ are known values, what is the distribution of $\beta | y, \sigma^2, \theta, \Sigma_\theta$ where $\beta = \begin{bmatrix} \beta_0 & \beta_1 \end{bmatrix}$?

$$\begin{aligned} P(\beta|\mathbf{y}, \mathbf{x}, \sigma^2, \theta, \Sigma_\theta) &\propto \exp\left(-\frac{1}{2\sigma^2}(y - \mathbf{x}^T\beta)^T \mathbf{I}_n (y - \mathbf{x}^T\beta)\right) \\ &\times \exp\left(-\frac{1}{2}(\beta - \theta)^T \Sigma_\theta^{-1} (\beta - \theta)\right) \text{ [drop terms without } \beta] \end{aligned}$$

The Normal-Normal Model

Let $Y_1, \ldots, Y_n \overset{iid}{\sim} \text{Normal}(\beta_0 + \beta_1 X_i, \sigma^2)$, where σ^2 is known. If we assume $\beta \sim \text{Normal}(\theta, \Sigma_\theta)$, where $\theta = \begin{bmatrix} \theta_0 & \theta_1 \end{bmatrix}$ and Σ_θ are known values, what is the distribution of $\beta | y, \sigma^2, \theta, \Sigma_\theta$ where $\beta = \begin{bmatrix} \beta_0 & \beta_1 \end{bmatrix}$?

$$
\begin{aligned}
&P(\beta | \mathbf{y}, \mathbf{x}, \sigma^2, \theta, \Sigma_\theta) \\
&\sim \text{Normal}\left(\left[\Sigma_\theta^{-1} + \frac{\sum_{i=1}^n x_i^2}{\sigma^2} \right]^{-1} \left[\Sigma_\theta^{-1}\theta + \frac{\sum_{i=1}^n x_i y_i}{\sigma^2} \right], \left[\Sigma_\theta^{-1} + \frac{\sum_{i=1}^n x_i^2}{\sigma^2} \right]^{-1} \right) \\
&\sim \text{Normal}\left(\left[\Sigma_\theta^{-1} + \mathbf{x}^T \Sigma^{-1} \mathbf{x} \right]^{-1} \left[\Sigma_\theta^{-1}\theta + \mathbf{x}^T \Sigma^{-1} \mathbf{y} \right], \left[\Sigma_\theta^{-1} + \mathbf{x}^T \Sigma^{-1} \mathbf{x} \right]^{-1} \right) \\
E[\beta | \mathbf{y}, \mathbf{x}, \sigma^2, \theta, \Sigma_\theta] &= \left[\Sigma_\theta^{-1} + \mathbf{x}^T \Sigma^{-1} \mathbf{x} \right]^{-1} \left[\Sigma_\theta^{-1}\theta + \mathbf{x}^T \Sigma^{-1} \mathbf{y} \right] \\
Var(\beta | \mathbf{y}, \mathbf{x}, \sigma^2, \theta, \Sigma_\theta) &= \left[\Sigma_\theta^{-1} + \mathbf{x}^T \Sigma^{-1} \mathbf{x} \right]^{-1}
\end{aligned}
$$