Foundations of Statistical Inference

Julien Berestycki

Department of Statistics
University of Oxford

MT 2016

Lecture 6 : Bayesian Inference

Ideas of probability

The majority of statistics you have learned so are are called classical or Frequentist. The probability for an event is defined as the proportion of successes in an infinite number of repeatable trials.

Ideas of probability

The majority of statistics you have learned so are are called classical or Frequentist. The probability for an event is defined as the proportion of successes in an infinite number of repeatable trials.

By contrast, in Subjective Bayesian inference, probability is a measure of the strength of belief.

Ideas of probability

The majority of statistics you have learned so are are called classical or Frequentist. The probability for an event is defined as the proportion of successes in an infinite number of repeatable trials.

By contrast, in Subjective Bayesian inference, probability is a measure of the strength of belief.

We treat parameters as random variables. Before collecting any data we assume that there is uncertainty about the value of a parameter. This uncertainty can be formalised by specifying a pdf (or pmf) for the parameter. We then conduct an experiment to collect some data that will give us information about the parameter. We then use Bayes Theorem to combine our prior beliefs with the data to derive an updated estimate of our uncertainty about the parameter.

The history of Bayesian Statistics

- Bayesian methods originated with Bayes and Laplace (late 1700s to mid 1800s).

The history of Bayesian Statistics

- Bayesian methods originated with Bayes and Laplace (late 1700s to mid 1800s).
- In the early 1920's, Fisher put forward an opposing viewpoint, that statistical inference must be based entirely on probabilities with direct experimental interpretation i.e. the repeated sampling principle.

The history of Bayesian Statistics

- Bayesian methods originated with Bayes and Laplace (late 1700s to mid 1800s).
- In the early 1920's, Fisher put forward an opposing viewpoint, that statistical inference must be based entirely on probabilities with direct experimental interpretation i.e. the repeated sampling principle.
- In 1939 Jeffrey's book 'The theory of probability' started a resurgence of interest in Bayesian inference.

The history of Bayesian Statistics

- Bayesian methods originated with Bayes and Laplace (late 1700s to mid 1800s).
- In the early 1920's, Fisher put forward an opposing viewpoint, that statistical inference must be based entirely on probabilities with direct experimental interpretation i.e. the repeated sampling principle.
- In 1939 Jeffrey's book 'The theory of probability' started a resurgence of interest in Bayesian inference.
- This continued throughout the 1950-60s, especially as problems with the Frequentist approach started to emerge.

The history of Bayesian Statistics

- Bayesian methods originated with Bayes and Laplace (late 1700s to mid 1800s).
- In the early 1920's, Fisher put forward an opposing viewpoint, that statistical inference must be based entirely on probabilities with direct experimental interpretation i.e. the repeated sampling principle.
- In 1939 Jeffrey's book 'The theory of probability' started a resurgence of interest in Bayesian inference.
- This continued throughout the 1950-60s, especially as problems with the Frequentist approach started to emerge.
- The development of simulation based inference has transformed Bayesian statistics in the last 20-30 years and it now plays a prominent part in modern statistics.

quodque ſolum, certa nitri ſigna præbere, ſed plura concurrere debere, ut de vero nitro producto dubium non relinquatur.

LII. *An Eſſay towards ſolving a Problem in the Doctrine of Chances. By the late Rev. Mr.* Bayes, *F. R. S. communicated by Mr.* Price, *in a Letter to* John Canton, *A. M. F. R. S.*

Dear Sir,

Read Dec. 23, 1763.

I Now ſend you an eſſay which I have found among the papers of our deceaſed friend Mr. Bayes, and which, in my opinion, has great merit, and well deſerves to be preſerved. Experimental philoſophy, you will find, is nearly intereſted in the ſubject of it; and on this account there ſeems to be particular reaſon for thinking that a communication of it to the Royal Society cannot be improper.

Problems with Frequentist Inference - I

Frequentist Inference generally does not condition on the observed data

Problems with Frequentist Inference - I

Frequentist Inference generally does not condition on the observed data

A confidence interval is a set-valued function $C(X) \subseteq \Theta$ of the data X which covers the parameter $\theta \in C(X)$ a fraction $1 - \alpha$ of repeated draws of X taken under the null H_0.

Problems with Frequentist Inference - I

Frequentist Inference generally does not condition on the observed data

A confidence interval is a set-valued function $C(X) \subseteq \Theta$ of the data X which covers the parameter $\theta \in C(X)$ a fraction $1-\alpha$ of repeated draws of X taken under the null H_0.

This is not the same as the statement that, given data $X = x$, the interval $C(x)$ covers θ with probability $1-\alpha$. But this is the type of statement we might wish to make. (observe that this statement makes sense iff θ is a r.v.)

Problems with Frequentist Inference - I

Frequentist Inference generally does not condition on the observed data

A confidence interval is a set-valued function $C(X) \subseteq \Theta$ of the data X which covers the parameter $\theta \in C(X)$ a fraction $1 - \alpha$ of repeated draws of X taken under the null H_0.

This is not the same as the statement that, given data $X = x$, the interval $C(x)$ covers θ with probability $1 - \alpha$. But this is the type of statement we might wish to make. (observe that this statement makes sense iff θ is a r.v.)

Example 1 Suppose $X_1, X_2 \sim U(\theta - 1/2, \theta + 1/2)$ so that $X_{(1)}$ and $X_{(2)}$ are the order statistics. Then $C(X) = [X_{(1)}, X_{(2)}]$ is a $\alpha = 1/2$ level CI for θ. Suppose in your data $X = x$, $x_{(2)} - x_{(1)} > 1/2$ (this happens in an eighth of data sets). Then $\theta \in [x_{(1)}, x_{(2)}]$ with probability one.

Problems with Frequentist Inference - II

Frequentist Inference depends on data that were never observed

Problems with Frequentist Inference - II

Frequentist Inference depends on data that were never observed

The likelihood principle Suppose that two experiments relating to θ, E_1, E_2, give rise to data y_1, y_2, such that the corresponding likelihoods are proportional, that is, for all θ

$$L(\theta; y_1, E_1) = cL(\theta; y_2, E_2).$$

then the two experiments lead to identical conclusions about θ.

Problems with Frequentist Inference - II

Frequentist Inference depends on data that were never observed

The likelihood principle Suppose that two experiments relating to θ, E_1, E_2, give rise to data y_1, y_2, such that the corresponding likelihoods are proportional, that is, for all θ

$$L(\theta; y_1, E_1) = cL(\theta; y_2, E_2).$$

then the two experiments lead to identical conclusions about θ.

Key point MLE's respect the likelihood principle i.e. the MLEs for θ are identical in both experiments. But significance tests do not respect the likelihood principle.

Consider a pure test for significance where we specify just H_0. We must choose a test statistic $T(x)$, and define the p-value for data $T(x) = t$ as

$$p\text{-value} = P(T(X) \text{ at least as extreme as } t | H_0).$$

The choice of $T(X)$ amounts to a statement about the direction of likely departures from the null, which requires some consideration of alternative models.

Consider a pure test for significance where we specify just H_0. We must choose a test statistic $T(x)$, and define the p-value for data $T(x) = t$ as

$$p\text{-value} = P(T(X) \text{ at least as extreme as } t | H_0).$$

The choice of $T(X)$ amounts to a statement about the direction of likely departures from the null, which requires some consideration of alternative models.

Note 1 The calculation of the p-value involves a sum (or integral) over data that was not observed, and this can depend upon the form of the experiment.

Consider a pure test for significance where we specify just H_0. We must choose a test statistic $T(x)$, and define the p-value for data $T(x) = t$ as

$$p\text{-value} = P(T(X) \text{ at least as extreme as } t | H_0).$$

The choice of $T(X)$ amounts to a statement about the direction of likely departures from the null, which requires some consideration of alternative models.

Note 1 The calculation of the p-value involves a sum (or integral) over data that was not observed, and this can depend upon the form of the experiment.

Note 2 A p-value is not $P(H_0 | T(X) = t)$.

Example

A Bernoulli trial succeeds with probability p.

Example

A Bernoulli trial succeeds with probability p.

E_1 fix n_1 Bernoulli trials, count number y_1 of successes

Example

A Bernoulli trial succeeds with probability p.

E_1 fix n_1 Bernoulli trials, count number y_1 of successes

E_2 count number n_2 Bernoulli trials to get fixed number y_2 successes

Example

A Bernoulli trial succeeds with probability p.

E_1 fix n_1 Bernoulli trials, count number y_1 of successes

E_2 count number n_2 Bernoulli trials to get fixed number y_2 successes

$$L(p; y_1, E_1) = \binom{n_1}{y_1} p^{y_1} (1-p)^{n_1 - y_1} \quad \textit{binomial}$$

Example

A Bernoulli trial succeeds with probability p.

E_1 fix n_1 Bernoulli trials, count number y_1 of successes

E_2 count number n_2 Bernoulli trials to get fixed number y_2 successes

$$\begin{aligned} L(p; y_1, E_1) &= \binom{n_1}{y_1} p^{y_1}(1-p)^{n_1-y_1} \quad \textit{binomial} \\ L(p; n_2, E_2) &= \binom{n_2-1}{y_2-1} p^{y_2}(1-p)^{n_2-y_2} \quad \textit{negative binomial} \end{aligned}$$

Example

A Bernoulli trial succeeds with probability p.

E_1 fix n_1 Bernoulli trials, count number y_1 of successes

E_2 count number n_2 Bernoulli trials to get fixed number y_2 successes

$$
\begin{aligned}
L(p; y_1, E_1) &= \binom{n_1}{y_1} p^{y_1} (1-p)^{n_1 - y_1} \quad \textit{binomial} \\
L(p; n_2, E_2) &= \binom{n_2 - 1}{y_2 - 1} p^{y_2} (1-p)^{n_2 - y_2} \quad \textit{negative binomial}
\end{aligned}
$$

If $n_1 = n_2 = n$, $y_1 = y_2 = y$ then $L(p; y_1, E_1) \propto L(p; n_2, E_2)$.

Example

A Bernoulli trial succeeds with probability p.

E_1 fix n_1 Bernoulli trials, count number y_1 of successes

E_2 count number n_2 Bernoulli trials to get fixed number y_2 successes

$$
\begin{aligned}
L(p; y_1, E_1) &= \binom{n_1}{y_1} p^{y_1}(1-p)^{n_1-y_1} \quad \textit{binomial} \\
L(p; n_2, E_2) &= \binom{n_2-1}{y_2-1} p^{y_2}(1-p)^{n_2-y_2} \quad \textit{negative binomial}
\end{aligned}
$$

If $n_1 = n_2 = n$, $y_1 = y_2 = y$ then $L(p; y_1, E_1) \propto L(p; n_2, E_2)$.
So MLEs for p will be the same under E_1 and E_2.

Example

But significance tests contradict : eg, $H_0 : p = 1/2$ against $H_1 : p < 1/2$ and suppose $n = 12$ and $y = 3$.

Example

But significance tests contradict : eg, $H_0 : p = 1/2$ against $H_1 : p < 1/2$ and suppose $n = 12$ and $y = 3$.

The p-value based on E_1 is

$$P\left(Y \leq y|\theta = \frac{1}{2}\right) = \sum_{k=0}^{y} \binom{n}{k} (1/2)^k (1 - 1/2)^{n-k} (= 0.073)$$

Example

But significance tests contradict : eg, $H_0 : p = 1/2$ against $H_1 : p < 1/2$ and suppose $n = 12$ and $y = 3$.

The p-value based on E_1 is

$$P\left(Y \leq y|\theta = \frac{1}{2}\right) = \sum_{k=0}^{y} \binom{n}{k} (1/2)^k (1-1/2)^{n-k} (= 0.073)$$

while the p-value based on E_2 is

$$P\left(N \geq n|\theta = \frac{1}{2}\right) = \sum_{k=n}^{\infty} \binom{k-1}{y-1} (1/2)^k (1-1/2)^{n-k} (= 0.033)$$

so different conclusions at significance level 0.05.

Example

But significance tests contradict : eg, $H_0 : p = 1/2$ against $H_1 : p < 1/2$ and suppose $n = 12$ and $y = 3$.

The p-value based on E_1 is

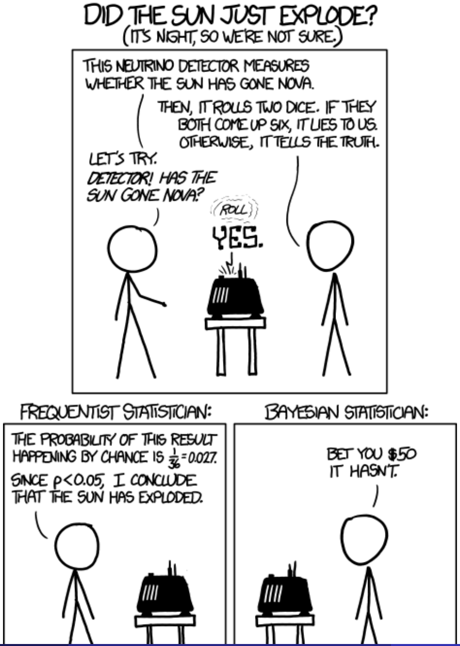
$$P\left(Y \leq y|\theta = \frac{1}{2}\right) = \sum_{k=0}^{y} \binom{n}{k} (1/2)^k (1-1/2)^{n-k} (= 0.073)$$

while the p-value based on E_2 is

$$P\left(N \geq n|\theta = \frac{1}{2}\right) = \sum_{k=n}^{\infty} \binom{k-1}{y-1} (1/2)^k (1-1/2)^{n-k} (= 0.033)$$

so different conclusions at significance level 0.05.

Note The p-values disagree because they sum over portions of two different sample spaces.


DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY.
DETECTOR! HAS THE SUN GONE NOVA?
ROLL
YES.
FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS 1/36 = 0.027.
SINCE p<0.05, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.

Bayesian Inference - Revison

Likelihood $f(x \mid \theta)$ and prior distribution $\pi(\theta)$ for ϑ

Bayesian Inference - Revison

Likelihood $f(x \mid \theta)$ and prior distribution $\pi(\theta)$ for ϑ. The posterior distribution of ϑ at $\vartheta = \theta$, given x, is

$$\pi(\theta \mid x) = \frac{f(x \mid \theta)\pi(\theta)}{\int f(x \mid \theta)\pi(\theta)d\theta} \Rightarrow \quad \pi(\theta \mid x) \quad \propto \quad f(x \mid \theta)\pi(\theta)$$

Bayesian Inference - Revison

Likelihood $f(x \mid \theta)$ and prior distribution $\pi(\theta)$ for ϑ. The posterior distribution of ϑ at $\vartheta = \theta$, given x, is

$$\pi(\theta \mid x) = \frac{f(x \mid \theta)\pi(\theta)}{\int f(x \mid \theta)\pi(\theta)d\theta} \Rightarrow \quad \begin{array}{rcl} \pi(\theta \mid x) & \propto & f(x \mid \theta)\pi(\theta) \\ \text{posterior} & \propto & \text{likelihood} \times \text{prior} \end{array}$$

The same form for θ continuous ($\pi(\theta \mid x)$ a pdf) or discrete ($\pi(\theta \mid x)$ a pmf). We call $\int f(x \mid \theta)\pi(\theta)d\theta$ the marginal likelihood.

Bayesian Inference - Revison

Likelihood $f(x \mid \theta)$ and prior distribution $\pi(\theta)$ for ϑ. The posterior distribution of ϑ at $\vartheta = \theta$, given x, is

$$\pi(\theta \mid x) = \frac{f(x \mid \theta)\pi(\theta)}{\int f(x \mid \theta)\pi(\theta) d\theta} \Rightarrow \quad \pi(\theta \mid x) \propto f(x \mid \theta)\pi(\theta)$$

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

The same form for θ continuous ($\pi(\theta \mid x)$ a pdf) or discrete ($\pi(\theta \mid x)$ a pmf). We call $\int f(x \mid \theta)\pi(\theta) d\theta$ the marginal likelihood.

Likelihood principle Notice that, if we base all inference on the posterior distribution, then we respect the likelihood principle. If two likelihood functions are proportional, then any constant cancels top and bottom in Bayes rule, and the two posterior distributions are the same.

Example 1

$X \sim \text{Bin}(n, \vartheta)$ for known n and unknown ϑ

Example 1

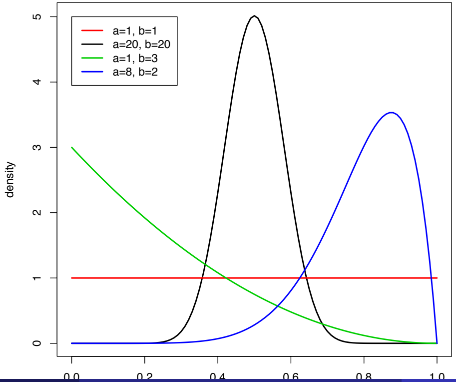
$X \sim \text{Bin}(n, \vartheta)$ for known n and unknown ϑ. Suppose our prior knowledge about ϑ is represented by a Beta distribution on $(0, 1)$, and θ is a trial value for ϑ.

$$\pi(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}, \; 0 < \theta < 1.$$

Example 1

$X \sim \text{Bin}(n, \vartheta)$ for known n and unknown ϑ. Suppose our prior knowledge about ϑ is represented by a Beta distribution on $(0, 1)$, and θ is a trial value for ϑ.

$$\pi(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}, \; 0 < \theta < 1.$$



Example 1

Prior probability density

$$\pi(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)},\ 0 < \theta < 1.$$

Example 1

Prior probability density

$$\pi(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)},\ 0 < \theta < 1.$$

Likelihood

$$f(x \mid \theta) = \binom{n}{x}\theta^{x}(1-\theta)^{n-x},\ x = 0, \dots, n$$

Example 1

Prior probability density

$$\pi(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)},\ 0 < \theta < 1.$$

Likelihood

$$f(x \mid \theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x},\ x = 0, \dots, n$$

Posterior probability density

$$\pi(\theta \mid x) \quad \propto \quad \text{likelihood} \times \text{prior}$$

Example 1

Prior probability density

$$\pi(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)},\ 0 < \theta < 1.$$

Likelihood

$$f(x \mid \theta) = \binom{n}{x}\theta^{x}(1-\theta)^{n-x},\ x = 0,\ldots,n$$

Posterior probability density

$$\begin{aligned}\pi(\theta \mid x) &\propto \text{likelihood} \times \text{prior}\\ &\propto \theta^{a-1}(1-\theta)^{b-1}\theta^{x}(1-\theta)^{n-x}\end{aligned}$$

Example 1

Prior probability density

$$\pi(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}, \ 0 < \theta < 1.$$

Likelihood

$$f(x \mid \theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, \ x = 0, \ldots, n$$

Posterior probability density

$$\begin{aligned}
\pi(\theta \mid x) &\propto \text{likelihood} \times \text{prior} \\
&\propto \theta^{a-1}(1-\theta)^{b-1}\theta^x(1-\theta)^{n-x} \\
&= \theta^{a+x-1}(1-\theta)^{n-x+b-1}
\end{aligned}$$

Example 1

Prior probability density

$$\pi(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)},\ 0 < \theta < 1.$$

Likelihood

$$f(x \mid \theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x},\ x = 0, \dots, n$$

Posterior probability density

$$\begin{aligned}\pi(\theta \mid x) &\propto \text{likelihood} \times \text{prior} \\ &\propto \theta^{a-1}(1-\theta)^{b-1}\theta^x(1-\theta)^{n-x} \\ &= \theta^{a+x-1}(1-\theta)^{n-x+b-1}\end{aligned}$$

Here posterior has the same form as the prior (conjugacy) with updated parameters a, b replaced by $a + x, b + n - x$

Example 1

Prior probability density

$$\pi(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}, \ 0 < \theta < 1.$$

Likelihood

$$f(x \mid \theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, \ x = 0, \dots, n$$

Posterior probability density

$$\begin{aligned}\pi(\theta \mid x) &\propto \text{likelihood} \times \text{prior} \\ &\propto \theta^{a-1}(1-\theta)^{b-1}\theta^x(1-\theta)^{n-x} \\ &= \theta^{a+x-1}(1-\theta)^{n-x+b-1}\end{aligned}$$

Here posterior has the same form as the prior (conjugacy) with updated parameters a, b replaced by $a+x, b+n-x$, so

$$\pi(\theta \mid x) = \frac{\theta^{a+x-1}(1-\theta)^{n-x+b-1}}{B(a+x, b+n-x)}$$

Example 1

For a Beta distribution with parameters a, b

$$\mu = \frac{a}{a+b}, \ \sigma^2 = \frac{ab}{(a+b)^2(a+b+1)}$$

Example 1

For a Beta distribution with parameters a, b

$$\mu = \frac{a}{a+b}, \ \sigma^2 = \frac{ab}{(a+b)^2(a+b+1)}$$

The posterior mean and variance are

$$\frac{a+X}{a+b+n}, \ \frac{(a+X)(b+n-X)}{(a+b+n)^2(a+b+n+1)}$$

Example 1

For a Beta distribution with parameters a, b

$$\mu = \frac{a}{a+b}, \; \sigma^2 = \frac{ab}{(a+b)^2(a+b+1)}$$

The posterior mean and variance are

$$\frac{a+X}{a+b+n}, \; \frac{(a+X)(b+n-X)}{(a+b+n)^2(a+b+n+1)}$$

Suppose $X = n$ and we set $a = b = 1$ for our prior. Then posterior mean is

$$\frac{n+1}{n+2}$$

i.e. when we observe events of just one type then our point estimate is not 0 or 1 (which is sensible especially in small sample sizes).

Example 1

For large n, the posterior mean and variance are approximately

$$\frac{X}{n}, \; \frac{X(n-X)}{n^3}$$

In classical statistics

$$\widehat{\theta} = \frac{X}{n}, \; \frac{\widehat{\theta}(1-\widehat{\theta})}{n} = \frac{X(n-X)}{n^3}$$

Example 1

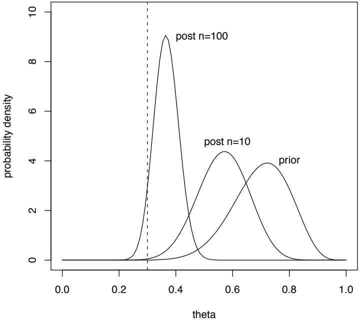
Suppose $\vartheta = 0.3$ but prior mean is 0.7 with std 0.1

Example 1

Suppose $\vartheta = 0.3$ but prior mean is 0.7 with std 0.1. Suppose data $X \sim \text{Bin}(n, \vartheta)$ with $n = 10$ (yielding $X = 3$) and then $n = 100$ (yielding $X = 30$, say).

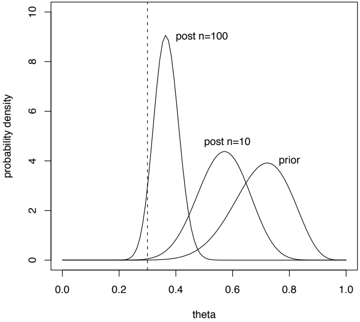
Example 1

Suppose $\vartheta = 0.3$ but prior mean is 0.7 with std 0.1. Suppose data $X \sim \text{Bin}(n, \vartheta)$ with $n = 10$ (yielding $X = 3$) and then $n = 100$ (yielding $X = 30$, say).



Example 1

Suppose $\vartheta = 0.3$ but prior mean is 0.7 with std 0.1. Suppose data $X \sim \text{Bin}(n, \vartheta)$ with $n = 10$ (yielding $X = 3$) and then $n = 100$ (yielding $X = 30$, say).



As n increases, the likelihood overwhelms information in prior.

Example 2 - Conjugate priors

Normal distribution when the mean and variance are unknown.

Example 2 - Conjugate priors

Normal distribution when the mean and variance are unknown.Let $\tau = 1/\sigma^2$, $\theta = (\tau, \mu)$. τ is called the precision.

Example 2 - Conjugate priors

Normal distribution when the mean and variance are unknown.Let $\tau = 1/\sigma^2$, $\theta = (\tau, \mu)$. τ is called the precision. The prior τ has a Gamma distribution with parameters $\alpha, \beta > 0$, and conditional on τ, μ has a distribution $N(\nu, \frac{1}{k\tau})$ for some $k > 0$, $\nu \in \mathbb{R}$.

Example 2 - Conjugate priors

Normal distribution when the mean and variance are unknown.Let $\tau = 1/\sigma^2$, $\theta = (\tau, \mu)$. τ is called the precision. The prior τ has a Gamma distribution with parameters $\alpha, \beta > 0$, and conditional on τ, μ has a distribution $N(\nu, \frac{1}{k\tau})$ for some $k > 0$, $\nu \in \mathbb{R}$.
The prior is

$$\pi(\tau, \mu) = \frac{\beta^\alpha}{\Gamma(\alpha)} \tau^{\alpha-1} e^{-\beta\tau}$$

Example 2 - Conjugate priors

Normal distribution when the mean and variance are unknown.Let $\tau = 1/\sigma^2$, $\theta = (\tau, \mu)$. τ is called the precision. The prior τ has a Gamma distribution with parameters $\alpha, \beta > 0$, and conditional on τ, μ has a distribution $N(\nu, \frac{1}{k\tau})$ for some $k > 0$, $\nu \in \mathbb{R}$.
The prior is

$$\pi(\tau, \mu) = \frac{\beta^\alpha}{\Gamma(\alpha)} \tau^{\alpha-1} e^{-\beta\tau} \cdot (2\pi)^{-1/2} (k\tau)^{1/2} \exp\left\{ -\frac{k\tau}{2} (\mu - \nu)^2 \right\}$$

Example 2 - Conjugate priors

Normal distribution when the mean and variance are unknown.Let $\tau = 1/\sigma^2$, $\theta = (\tau, \mu)$. τ is called the precision. The prior τ has a Gamma distribution with parameters $\alpha, \beta > 0$, and conditional on τ, μ has a distribution $N(\nu, \frac{1}{k\tau})$ for some $k > 0$, $\nu \in \mathbb{R}$.
The prior is

$$\pi(\tau, \mu) = \frac{\beta^\alpha}{\Gamma(\alpha)} \tau^{\alpha - 1} e^{-\beta\tau} \cdot (2\pi)^{-1/2} (k\tau)^{1/2} \exp\left\{ -\frac{k\tau}{2} (\mu - \nu)^2 \right\}$$

or

$$\pi(\tau, \mu) \propto \tau^{\alpha - 1/2} \exp\left[-\tau \left\{ \beta + \frac{k}{2} (\mu - \nu)^2 \right\} \right]$$

Example 2 - Conjugate priors

Normal distribution when the mean and variance are unknown.Let $\tau = 1/\sigma^2$, $\theta = (\tau, \mu)$. τ is called the precision. The prior τ has a Gamma distribution with parameters $\alpha, \beta > 0$, and conditional on τ, μ has a distribution $N(\nu, \frac{1}{k\tau})$ for some $k > 0$, $\nu \in \mathbb{R}$.
The prior is

$$\pi(\tau, \mu) = \frac{\beta^\alpha}{\Gamma(\alpha)} \tau^{\alpha-1} e^{-\beta\tau} \cdot (2\pi)^{-1/2} (k\tau)^{1/2} \exp\left\{-\frac{k\tau}{2}(\mu - \nu)^2\right\}$$

or

$$\pi(\tau, \mu) \propto \tau^{\alpha-1/2} \exp\left[-\tau\left\{\beta + \frac{k}{2}(\mu - \nu)^2\right\}\right]$$

The likelihood is

$$f(x \mid \mu, \tau) = (2\pi)^{-n/2} \tau^{n/2} \exp\left\{-\frac{\tau}{2} \sum_{i=1}^{n} (x_i - \mu)^2\right\}$$

Example 2 - Conjugate priors

Thus

$$\pi(\tau, \mu \mid x) \propto \tau^{\alpha+(n/2)-1/2} \exp\left[-\tau\left\{\beta + \frac{k}{2}(\mu-\nu)^2 + \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}\right]$$

Example 2 - Conjugate priors

Thus

$$\pi(\tau, \mu \mid x) \propto \tau^{\alpha+(n/2)-1/2} \exp\left[-\tau\left\{\beta + \frac{k}{2}(\mu-\nu)^2 + \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}\right]$$

Complete the square to see that

$$\begin{aligned} k(\mu-\nu)^2 &+ \sum(x_i-\mu)^2 \\ &= (k+n)\left(\mu - \frac{k\nu+n\bar{x}}{k+n}\right)^2 + \frac{nk}{n+k}(\bar{x}-\nu)^2 + \sum(x_i-\bar{x})^2 \end{aligned}$$

Example 2 - Completing the square

Your posterior dependence on μ is entirely in the factor

$$\exp\left[-\tau\left\{\beta + \frac{k}{2}(\mu-\nu)^2 + \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}\right]$$

Example 2 - Completing the square

Your posterior dependence on μ is entirely in the factor

$$\exp\left[-\tau\left\{\beta+\frac{k}{2}(\mu-\nu)^2+\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}\right]$$

The idea is to try to write that as

$$c\exp\left[-\tau\left\{\beta'+\frac{k'}{2}(\nu'-\mu)^2\right\}\right]$$

so that (conditional on τ) we recognize a Normal density.

Example 2 - Completing the square

Your posterior dependence on μ is entirely in the factor

$$\exp\left[-\tau\left\{\beta+\frac{k}{2}(\mu-\nu)^2+\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}\right]$$

The idea is to try to write that as

$$c\exp\left[-\tau\left\{\beta'+\frac{k'}{2}(\nu'-\mu)^2\right\}\right]$$

so that (conditional on τ) we recognize a Normal density.

$$\begin{aligned}
&k(\mu-\nu)^2+\sum(x_i-\mu)^2\\
&\quad=\mu^2(k+n)-\mu(2k\nu+2\sum x_i)+\dots\\
&\quad=(k+n)\left(\mu-\frac{k\nu+n\bar{x}}{k+n}\right)^2+\frac{nk}{n+k}(\bar{x}-\nu)^2+\sum(x_i-\bar{x})^2
\end{aligned}$$

Example 2 - Conjugate priors

Thus the posterior is

$$\pi(\tau, \mu \mid x) \propto \tau^{\alpha' - 1/2} \exp\left[-\tau \left\{\beta' + \frac{k'}{2}(\nu' - \mu)^2\right\}\right]$$

where

$$\begin{aligned} \alpha' &= \alpha + \frac{n}{2} \\ \beta' &= \beta + \frac{1}{2} \cdot \frac{nk}{n+k}(\bar{x} - \nu)^2 + \frac{1}{2}\sum (x_i - \bar{x})^2 \\ k' &= k + n \\ \nu' &= \frac{k\nu + n\bar{x}}{k+n} \end{aligned}$$

This is the same form as the prior, so the class is conjugate prior.

Example 2 - Contd

If we are interested in the posterior distribution of μ alone

$$\pi(\mu|x) = \int \pi(\tau, \mu|x) d\tau$$

Example 2 - Contd

If we are interested in the posterior distribution of μ alone

$$\pi(\mu|x) = \int \pi(\tau, \mu|x)d\tau = \int \pi(\mu|\tau, x)\pi(\tau|x)d\tau$$

Here, we have a simplification if we assume $2\alpha = m \in \mathbb{N}$. Then

$$\tau = W/2\beta, \quad \mu = \nu + Z/\sqrt{k\tau}$$

with $W \sim \chi^2_m$ and $Z \sim N(0, 1)$.

Example 2 - Contd

If we are interested in the posterior distribution of μ alone

$$\pi(\mu|x) = \int \pi(\tau, \mu|x) d\tau = \int \pi(\mu|\tau, x)\pi(\tau|x) d\tau$$

Here, we have a simplification if we assume $2\alpha = m \in \mathbb{N}$. Then

$$\tau = W/2\beta, \quad \mu = \nu + Z/\sqrt{k\tau}$$

with $W \sim \chi^2_m$ and $Z \sim N(0,1)$. Recall that $Z\sqrt{m/W} \sim t_m$ (Studen with m d.f.) we see that the prior of μ is

$$\sqrt{\frac{km}{2\beta}}(\mu - \nu) \sim t_m$$

and the posterior is

$$\sqrt{\frac{k'm'}{2\beta'}}(\mu - \nu') \sim t_{m'}, \quad m' = m + n$$

Example: estimating the probability of female birth given placenta previa

Result of german study: 980 birth, 437 females. In general population the proportion is 0.485.

Example: estimating the probability of female birth given placenta previa

Result of german study: 980 birth, 437 females. In general population the proportion is 0.485.
Using a uniform (Beta(1,1)) prior, posterior is Beta(438,544).

$$\text{post. mean } = 0.446 \quad \text{post. std dev } = 0.016$$
$$\text{central 95\% post. interval } = [0.415, 0.477]$$

Example: estimating the probability of female birth given placenta previa

Result of german study: 980 birth, 437 females. In general population the proportion is 0.485.
Using a uniform (Beta(1,1)) prior, posterior is Beta(438,544).

$$\text{post. mean } = 0.446 \quad \text{post. std dev } = 0.016$$
$$\text{central 95\% post. interval } = [0.415, 0.477]$$

Sensibility to proposed prior. $\alpha/\beta - 2 =$ “prior sample size".

Parameters of the prior distribution		Summaries of the posterior distribution	
		Posterior median of θ	95% posterior interval for θ
$\frac{\alpha}{\alpha+\beta}$	$\alpha+\beta$		
0.500	2	0.446	[0.415, 0.477]
0.485	2	0.446	[0.415, 0.477]
0.485	5	0.446	[0.415, 0.477]
0.485	10	0.446	[0.415, 0.477]
0.485	20	0.447	[0.416, 0.478]
0.485	100	0.450	[0.420, 0.479]
0.485	200	0.453	[0.424, 0.481]