Bayesian Inference

Loukia Meligkotsidou,
National and Kapodistrian University of Athens

MSc in Statistics and Operational Research,
Department of Mathematics

Outline of the course

This course provides theory and practice of the Bayesian approach to statistical inference. Applications are performed with the statistical package R.

Topics:

- Bayesian Updating through Bayes' Theorem
- Prior Distributions
- Multi-parameter Problems
- Summarizing Posterior Information
- Prediction
- Asymptotics
- Markov chain Monte Carlo Methods
- The Gibbs Sampler
- Data Augmentation
- The Metropolis-Hastings Algorithm
- Applications

Unit 1: Introduction

'What is statistical inference?'

Many definitions are possible, but most boil down to the principle that statistical inference is the science of making conclusions about a 'population' from 'sample', items drawn from that population. (This itself begs many questions about what is meant by a population, how the sample relates to the population,etc).

Unit 1: Introduction

'What is statistical inference?'
Many definitions are possible, but most boil down to the principle that statistical inference is the science of making conclusions about a 'population' from 'sample', items drawn from that population. (This itself begs many questions about what is meant by a population, how the sample relates to the population,etc).

Parametric Inference

Statistical Modelling: Build a stochastic model, containing a few unknown parameters, to describe the dynamics of a random process. (Distributional assumptions, linear models, GLMs, etc).
Statistical Inference: Develop techniques to infer the model's parameters from data, observations on the random process. (Estimation, Confidence Intervals, Hypothesis Tests, Predictions).

Unit 1: Introduction

Uncertainty holds a central role in statistics!

Unit 1: Introduction

Uncertainty holds a central role in statistics!

Bayesian Inference is based on a simple idea:

Unit 1: Introduction

Uncertainty holds a central role in statistics!

Bayesian Inference is based on a simple idea:
The only satisfactory description of uncertainty is achieved through probability.

Unit 1: Introduction

Uncertainty holds a central role in statistics!

Bayesian Inference is based on a simple idea:
The only satisfactory description of uncertainty is achieved through probability.

The rule:

Unit 1: Introduction

Uncertainty holds a central role in statistics!

Bayesian Inference is based on a simple idea:
The only satisfactory description of uncertainty is achieved through probability.

The rule:
All the unknown quantities should be described through probabilities.

Unit 1: Introduction

Uncertainty holds a central role in statistics!

Bayesian Inference is based on a simple idea:
The only satisfactory description of uncertainty is achieved through probability.

The rule:
All the unknown quantities should be described through probabilities.

This means that the parameters of a statistical model should be treated as random variables.

Introduction: Example

Suppose the Forestry Commission wish to estimate the proportion of trees in a large forest which suffer from a particular disease. It's impractical to check every tree, so they select a sample of n trees.

Random sampling: if θ is the proportion of trees having the disease in the forest, then each tree in the sample will have the disease, independently of all others in the sample, with probability θ.

Introduction: Example

Suppose the Forestry Commission wish to estimate the proportion of trees in a large forest which suffer from a particular disease. It's impractical to check every tree, so they select a sample of n trees.

Random sampling: if θ is the proportion of trees having the disease in the forest, then each tree in the sample will have the disease, independently of all others in the sample, with probability θ.

X: the number of diseased trees in the sample
$X = x$: the observed value of the random variable X

Introduction: Example

Suppose the Forestry Commission wish to estimate the proportion of trees in a large forest which suffer from a particular disease. It's impractical to check every tree, so they select a sample of n trees.

Random sampling: if θ is the proportion of trees having the disease in the forest, then each tree in the sample will have the disease, independently of all others in the sample, with probability θ.

X: the number of diseased trees in the sample
$X = x$: the observed value of the random variable X

Inference: *point estimate* ($\hat{\theta} = 0.1$);
confidence interval (95 % confident that θ lies in [0.08,0.12]);
hypothesis test (reject the hypothesis that $\theta = 0.07$ at sig. 5%);
prediction (predict that 15% of trees will be affected by next year).

Introduction: Example

Statistical inferences are made by specifying a probability model, also called the likelihood model, $f(x|\theta)$, which determines how, for a given value of θ, the probabilities of the different values of X are distributed. Here, $X|\theta \sim Binomial(n, \theta)$, therefore

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

Inference about the population parameter θ are made on the basis of observing $X = x$!

Introduction: Example

Statistical inferences are made by specifying a probability model, also called the likelihood model, $f(x|\theta)$, which determines how, for a given value of θ, the probabilities of the different values of X are distributed. Here, $X|\theta \sim Binomial(n, \theta)$, therefore

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

Inference about the population parameter θ are made on the basis of observing $X = x$!

The principle of maximum likelihood: values of θ which give high probability to the observed value x are 'more likely' than those which assign x low probability.

The MLE: choose, as the best point estimate of θ the value that maximizes the likelihood function!

The Classical or Frequentist Approach to Inference

The most fundamental point in classical inference is that the parameter θ, whilst not known, is being treated as constant rather than random. This is the cornerstone of classical theory, but leads to some problems of interpretation.

The Classical or Frequentist Approach to Inference

The most fundamental point in classical inference is that the parameter θ, whilst not known, is being treated as constant rather than random. This is the cornerstone of classical theory, but leads to some problems of interpretation.

For example, we'd like a 95% CI of $[0.08, 0.12]$ to mean there's a 95% probability that θ lies between 0.08 and 0.12. It *cannot* mean this, since θ is not random: it either *is* in the interval, or it *isn't*.

The Classical or Frequentist Approach to Inference

The most fundamental point in classical inference is that the parameter θ, whilst not known, is being treated as constant rather than random. This is the cornerstone of classical theory, but leads to some problems of interpretation.

For example, we'd like a 95% CI of $[0.08, 0.12]$ to mean there's a 95% probability that θ lies between 0.08 and 0.12. It *cannot* mean this, since θ is not random: it either *is* in the interval, or it *isn't*.

The only random element in this probability model is the data, so the correct interpretation of the CI is that if we applied our procedure 'many times', then 'in the long run', the intervals we construct will contain θ on 95% of occasions.

All inferences based on classical theory are forced to have this type of long–run–frequency interpretation. This leads to the, so called, frequentist approach to inference.

The Bayesian Approach to Inference

The overall framework which Bayesian inference works within is identical to that above: there is a population parameter θ which we wish to make inferences about, and a probability mechanism $f(x \mid \theta)$ which determines the probability of observing different data x, under different parameter values θ.

The Bayesian Approach to Inference

The overall framework which Bayesian inference works within is identical to that above: there is a population parameter θ which we wish to make inferences about, and a probability mechanism $f(x \mid \theta)$ which determines the probability of observing different data x, under different parameter values θ.

So, Bayesian inference is still likelihood-based inference. The fundamental difference is that θ is treated as a random quantity.

The Bayesian Approach to Inference

The overall framework which Bayesian inference works within is identical to that above: there is a population parameter θ which we wish to make inferences about, and a probability mechanism $f(x \mid \theta)$ which determines the probability of observing different data x, under different parameter values θ.

So, Bayesian inference is still likelihood-based inference. The fundamental difference is that θ is treated as a random quantity.

In essence, inferences are based on $f(\theta \mid x)$ rather than $f(x \mid \theta)$; that is the probability distribution of the parameter given the data, rather than the data given the parameter. This leads to a probabilistic approach to statistical inference.

The Bayesian Approach to Inference

The overall framework which Bayesian inference works within is identical to that above: there is a population parameter θ which we wish to make inferences about, and a probability mechanism $f(x \mid \theta)$ which determines the probability of observing different data x, under different parameter values θ.

So, Bayesian inference is still likelihood-based inference. The fundamental difference is that θ is treated as a random quantity.

In essence, inferences are based on $f(\theta \mid x)$ rather than $f(x \mid \theta)$; that is the probability distribution of the parameter given the data, rather than the data given the parameter. This leads to a probabilistic approach to statistical inference.

To achieve this, it is necessary to specify a prior distribution, $f(\theta)$, which represents beliefs about θ *prior* to observing data.

The Coin Sampling Example

Five coins have been tossed. We are required to estimate the proportion θ of these coins which are tails, by looking at **a sample of just two coins**.

The Coin Sampling Example

Five coins have been tossed. We are required to estimate the proportion θ of these coins which are tails, by looking at **a sample of just two coins**.

There are only 6 possible values of θ: $0, 1/5, 2/5, 3/5, 4/5, 1$.

The Coin Sampling Example

Five coins have been tossed. We are required to estimate the proportion θ of these coins which are tails, by looking at **a sample of just two coins**.

There are only 6 possible values of θ: $0, 1/5, 2/5, 3/5, 4/5, 1$.

Let X be **the number of tails in the sample** of two coins, and suppose we observe $X = 1$.

The **likelihood** of a given value of θ is the probability of observing $X = 1$ depending on this value of θ.

The Coin Sampling Example

Five coins have been tossed. We are required to estimate the proportion θ of these coins which are tails, by looking at **a sample of just two coins**.

There are only 6 possible values of θ: $0, 1/5, 2/5, 3/5, 4/5, 1$.

Let X be **the number of tails in the sample** of two coins, and suppose we observe $X = 1$.

The **likelihood** of a given value of θ is the probability of observing $X = 1$ depending on this value of θ.

For example, $f(X = 1|\theta = \frac{3}{5})$ is the probability of observing $X = 1$ (we find **1 tail and 1 head** in the sample of 2 coins), if $\theta = \frac{3}{5}$ (we have **3 tails and 2 heads** in the set of 5 coins).

The Coin Sampling Example: The Likelihood

The number of ways of picking 1 tail and 1 head, out of the 3 tails and 2 heads, is $\binom{3}{1} \times \binom{2}{1} = 3 \times 2 = 6$.

The total number of ways of picking 2 coins out of 5 is $\binom{5}{2} = 10$.

The Coin Sampling Example: The Likelihood

The number of ways of picking 1 tail and 1 head, out of the 3 tails and 2 heads, is $\binom{3}{1} \times \binom{2}{1} = 3 \times 2 = 6.$

The total number of ways of picking 2 coins out of 5 is $\binom{5}{2} = 10$.

Then, $f(X = 1|\theta = 3/5) = \frac{\binom{3}{1} \times \binom{2}{1}}{\binom{5}{2}} = \frac{6}{10} = 0.6.$

The table of likelihoods

θ	0	1/5	2/5	3/5	4/5	1
$f(X = 1\|\theta)$	0.0	0.4	0.6	0.6	0.4	0.0

The Coin Sampling Example: The Prior

The Bayesian approach uses the likelihood function, but **combines it with prior knowledge**. This is described by the prior distribution of θ.

The Coin Sampling Example: The Prior

The Bayesian approach uses the likelihood function, but **combines it with prior knowledge**. This is described by the prior distribution of θ.

Suppose that, prior to observing data, we believe that **the coins are fair**: 1/2 probability of each coin being tail. Then, for example,

$$f(\theta = 0) = \binom{5}{0}\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^5 = \frac{1}{32} = f(\theta = 1)$$

The Coin Sampling Example: The Prior

The Bayesian approach uses the likelihood function, but **combines it with prior knowledge**. This is described by the prior distribution of θ.

Suppose that, prior to observing data, we believe that **the coins are fair**: 1/2 probability of each coin being tail. Then, for example,

$$f(\theta = 0) = \binom{5}{0} \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^5 = \frac{1}{32} = f(\theta = 1)$$

θ	0	1/5	2/5	3/5	4/5	1
$f(X = 1 \mid \theta)$	0.0	0.4	0.6	0.6	0.4	0.0
$f(\theta)$	1/32	5/32	10/32	10/32	5/32	1/32

The Coin Sampling Example: The Posterior

The *joint* distribution of X and θ is obtained as
$f(X = 1, \theta) = f(X = 1|\theta) \times f(\theta)$.

The Coin Sampling Example: The Posterior

The *joint* distribution of X and θ is obtained as $f(X=1,\theta)=f(X=1|\theta)\times f(\theta)$.

The important step is to get the **conditional distribution** of θ given $X=1$, i.e. the posterior distribution of θ, by dividing through by the sum $f(X=1)=\sum f(X=1|\theta)\times f(\theta)$:

$$f(\theta|X=1)=\frac{f(X=1|\theta)\times f(\theta)}{f(X=1)}$$

θ	0	1/5	2/5	3/5	4/5	1
$f(X=1\|\theta)$	0.0	0.4	0.6	0.6	0.4	0.0
$f(\theta)$	1/32	5/32	10/32	10/32	5/32	1/32
$f(X=1,\theta)$	0	2/32	6/32	6/32	2/32	0
$f(\theta\|X=1)$	0	4/32	12/32	12/32	4/32	0

Prior Beliefs

In almost all situations, when we are trying to estimate a parameter θ, we do have some knowledge, or some belief, about the value of θ before we take account of the data.

Prior Beliefs

In almost all situations, when we are trying to estimate a parameter θ, we do have some knowledge, or some belief, about the value of θ before we take account of the data.

An Example

You look out of a window and see *a large wooden thing with branches covered by small green things*. You entertain two hypotheses: the thing **is a tree** or **it's the postman**.

Prior Beliefs

In almost all situations, when we are trying to estimate a parameter θ, we do have some knowledge, or some belief, about the value of θ before we take account of the data.

An Example

You look out of a window and see *a large wooden thing with branches covered by small green things*. You entertain two hypotheses: the thing **is a tree** or **it's the postman**.

Define:

A: the event that you see a wooden thing with green bits

B_1: the event it's a tree

B_2: the event it's the postman

You **reject** B_2 in favour of B_1 because $f(A|B_1) > f(A|B_2)$ (the principle of **maximum likelihood**)

Prior Beliefs

But, you might also entertain a third possibility
B_3: the thing is a replica of a tree

Prior Beliefs

But, you might also entertain a third possibility
B_3: the thing is a replica of a tree

In this case it may well be that $f(A|B_1) = f(A|B_3)$, and yet you would still **reject** this hypothesis in favour of B_1.

Prior Beliefs

But, you might also entertain a third possibility
B_3: the thing is a replica of a tree

In this case it may well be that $f(A|B_1) = f(A|B_3)$, and yet you would still **reject** this hypothesis in favour of B_1.

That is, even though the probability of seeing what you observed is the same whether it is a tree or a replica, your **prior belief** is that *it's more likely* to be a tree than a replica and so you include this information when making your *decision*.

More Examples

Consider another example, where in each of the following cases our data model is $X|\theta \sim Bin(10, \theta)$ and we observe $x = 10$ so that the hypothesis $H_0 : \theta \leq 0.5$ is rejected in favour of $H_1 : \theta > 0.5$:

1. A woman tea–drinker claims she can detect from a cup of tea whether the milk was added before or after the tea. She does so correctly for ten cups.
2. A music expert claims he can distinguish between a page of Hayden's work and a page of Mozart. She correctly categorizes 10 pieces.
3. A drunk friend claims he can predict the outcome of tossing a fair coin, and does so correctly for 10 tosses.

More Examples

Consider another example, where in each of the following cases our data model is $X|\theta \sim Bin(10, \theta)$ and we observe $x = 10$ so that the hypothesis $H_0 : \theta \leq 0.5$ is rejected in favour of $H_1 : \theta > 0.5$:

1. A woman tea–drinker claims she can detect from a cup of tea whether the milk was added before or after the tea. She does so correctly for ten cups.
2. A music expert claims he can distinguish between a page of Hayden's work and a page of Mozart. She correctly categorizes 10 pieces.
3. A drunk friend claims he can predict the outcome of tossing a fair coin, and does so correctly for 10 tosses.

Just in terms of the data, we would draw **the same inferences** in each case. But our prior beliefs suggest that we are likely to remain sceptical about the drunk friend, impressed about the tea–drinker, and not surprised at all about the music expert.

The Prior Distribution

The essential point is this: experiments are not abstract devices. Invariably, we have some knowledge about the process being investigated before obtaining the data. It is sensible (many would say essential) that inferences should be based on the combined information that this prior knowledge *and* the data represent. **Bayesian inference is the mechanism for drawing inference from this combined knowledge**.

The Prior Distribution

The essential point is this: experiments are not abstract devices. Invariably, we have some knowledge about the process being investigated before obtaining the data. It is sensible (many would say essential) that inferences should be based on the combined information that this prior knowledge *and* the data represent. **Bayesian inference is the mechanism for drawing inference from this combined knowledge**.

Just to put the alternative point of view, it's this very **reliance on prior beliefs which opponents of the Bayesian viewpoint object to**. Different prior beliefs will lead to different inferences in the Bayesian view of things, and it's whether you see this as a good or a bad thing which determines your acceptability of the Bayesian framework.

Characteristics of the Bayesian Approach

- **Prior Information.** All problems are unique and have their own context, which derives prior information. This is taken into account in Bayesian analysis.

Characteristics of the Bayesian Approach

- **Prior Information.** All problems are unique and have their own context, which derives prior information. This is taken into account in Bayesian analysis.
- **Subjective Probability.** Classical statistics hinges on an objective 'long–run–frequency' definition of probabilities. Bayesian statistics formalizes explicitly the notion that all probabilities are subjective, depending on knowledge to hand. Inference is based on the *posterior* distribution $f(\theta|x)$, whose form depends (through Bayes' theorem) on the prior $f(\theta)$.

Characteristics of the Bayesian Approach

- **Prior Information.** All problems are unique and have their own context, which derives prior information. This is taken into account in Bayesian analysis.
- **Subjective Probability.** Classical statistics hinges on an objective 'long–run–frequency' definition of probabilities. Bayesian statistics formalizes explicitly the notion that all probabilities are subjective, depending on knowledge to hand. Inference is based on the *posterior* distribution $f(\theta|x)$, whose form depends (through Bayes' theorem) on the prior $f(\theta)$.
- **No 'adhockery'.** Because classical inference cannot make probability statements about θ, various criteria are developed to judge whether a particular estimator is in some sense 'good'. Bayesian statistics treats the parameter θ as random and, hence its whole development stems from probability theory and all inferences are probabilistic.

Review of Bayes Theorem

In its basic form, Bayes' Theorem is a simple result concerning conditional probabilities:

If A and B are two events with $\Pr(A) > 0$. Then

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

The use of Bayes' Theorem, in probability applications, is to reverse the conditioning of events. That is, it shows how the probability of $B|A$ is related to $A|B$.

Review of Bayes Theorem

A slight extension of **Bayes' Theorem** is obtained by considering events $C_1, \ldots, C_k$ which partition the sample space Ω, so that $C_i \cap C_j = \phi$ if $i \neq j$ and $C_1 \cup \ldots \cup C_k = \Omega$. Then

$$\Pr(C_i|A) = \frac{\Pr(A|C_i)\Pr(C_i)}{\sum_{j=1}^{k}\Pr(A|C_j)\Pr(C_j)} \quad \text{for } i = 1, \ldots, k.$$

Review of Bayes Theorem

A slight extension of **Bayes' Theorem** is obtained by considering events $C_1, \ldots, C_k$ which partition the sample space Ω, so that $C_i \cap C_j = \phi$ if $i \neq j$ and $C_1 \cup \ldots \cup C_k = \Omega$. Then

$$\Pr(C_i|A) = \frac{\Pr(A|C_i)\Pr(C_i)}{\sum_{j=1}^{k}\Pr(A|C_j)\Pr(C_j)} \quad \text{for } i = 1, \ldots, k.$$

A further extension is to **random variables**:

$$f(\theta \mid x) = \frac{f(x \mid \theta) f(\theta)}{f(x)}.$$

Example 1

A screening procedure for HIV is applied to a population which is **at high risk for HIV**; 10% of this population are believed to be HIV positive.

The screening test is **positive** for 90% of people who are **genuinely HIV positive**, and **negative** for 85% of people who are **not HIV positive**.

What are the probabilities of false positive and false negative results?

Example 1

Denote by
A: person is HIV positive, and **B: test result is positive**

Example 1

Denote by

A: person is HIV positive, and **B: test result is positive**

10% of the population are HIV positive: $\Pr(A) = 0.1$, $\Pr(A^c) = 0.9$
(**prior knowledge - before observing data)**

Example 1

Denote by

A: person is HIV positive, and **B: test result is positive**

10% of the population are HIV positive: $\Pr(A) = 0.1$, $\Pr(A^c) = 0.9$
(**prior knowledge - before observing data**)

The test is positive for 90% of people who are genuinely HIV positive: $\Pr(B|A) = 0.9$**,** $\Pr(B^c \mid A) = 0.1$
and negative for 85% of people who are not HIV positive:
$\Pr(B^c|A^c) = 0.85$**,** $\Pr(B \mid A^c) = 0.15$
(information in the data - likelihood)

Example 1

Denote by

A: person is HIV positive, and **B: test result is positive**

10% of the population are HIV positive: $\Pr(A) = 0.1$, $\Pr(A^c) = 0.9$
(**prior knowledge - before observing data)**

The test is positive for 90% of people who are genuinely HIV positive: $\Pr(B|A) = 0.9$, $\Pr(B^c \mid A) = 0.1$
and negative for 85% of people who are not HIV positive:
$\Pr(B^c|A^c) = 0.85$, $\Pr(B \mid A^c) = 0.15$
(information in the data - likelihood)

Probability of false positive: $\Pr(A^c|B)$**=?**
Probability of false negative: $\Pr(A|B^c)$**=?**
(posterior knowledge - after observing data)

Example 1

Compute $\Pr(B)$ through the **law of total probability**:
$\Pr(B) = \Pr(B \mid A)\Pr(A) + \Pr(B \mid A^c)\Pr(A^c) =$
$0.9 \times 0.1 + 0.15 \times 0.9 = 0.09 + 0.135 = 0.225$
and $\Pr(B^c) = 1 - \Pr(B) = 0.775$

Apply **Bayes' Theorem**:

$$\Pr(A^c|B) = \frac{\Pr(B|A^c)\Pr(A^c)}{\Pr(B)} = \frac{0.15 \times 0.9}{0.225} = 0.6$$

and

$$\Pr(A|B^c) = \frac{\Pr(B^c|A)\Pr(A)}{\Pr(B^c)} = \frac{0.1 \times 0.1}{0.775} = 0.0129$$

Example 2

In a bag there are 6 balls of unknown colours. **Three balls are drawn without replacement and are found to be black**. Find the probability that no black ball is left in the bag.

Example 2

In a bag there are 6 balls of unknown colours. **Three balls are drawn without replacement and are found to be black**. Find the probability that no black ball is left in the bag.

Let A**: 3 black balls are drawn**, and C_i**: there were** i **black balls in the bag**. Then, we need to calculate $\Pr(C_3|A)$.

Example 2

In a bag there are 6 balls of unknown colours. **Three balls are drawn without replacement and are found to be black**. Find the probability that no black ball is left in the bag.

Let A**: 3 black balls are drawn**, and C_i**: there were i black balls in the bag**. Then, we need to calculate $\Pr(C_3|A)$.

By Bayes' Theorem:

$$\Pr(C_i|A) = \frac{\Pr(A|C_i)\Pr(C_i)}{\sum_{j=0}^{6}\Pr(A|C_j)\Pr(C_j)}, \quad i = 0, \ldots, 6$$

Example 2

In a bag there are 6 balls of unknown colours. **Three balls are drawn without replacement and are found to be black**. Find the probability that no black ball is left in the bag.

Let A**: 3 black balls are drawn**, and C_i**: there were i black balls in the bag**. Then, we need to calculate $\Pr(C_3|A)$.

By Bayes' Theorem:

$$\Pr(C_i|A) = \frac{\Pr(A|C_i)\Pr(C_i)}{\sum_{j=0}^{6}\Pr(A|C_j)\Pr(C_j)}, \quad i = 0, \dots, 6$$

But here's the key issue: what values do we give $\Pr(C_0), \dots, \Pr(C_6)$? These are the probabilities of the different numbers of black balls in the bag, *prior* to having seen the data.

Example 2

Without any information to the contrary, we might well assume that **all possible numbers are equally likely**, i.e. $\Pr(C_0) = \Pr(C_1) = \ldots = \Pr(C_6) = \frac{1}{7}$.

Example 2

Without any information to the contrary, we might well assume that **all possible numbers are equally likely**, i.e. $\Pr(C_0) = \Pr(C_1) = \ldots = \Pr(C_6) = \frac{1}{7}$.

In fact, we will use this prior specification for the problem.

Example 2

Without any information to the contrary, we might well assume that **all possible numbers are equally likely**, i.e. $\Pr(C_0) = \Pr(C_1) = \ldots = \Pr(C_6) = \frac{1}{7}$.

In fact, we will use this prior specification for the problem.
But, is it the most sensible?

Example 2

Without any information to the contrary, we might well assume that **all possible numbers are equally likely**, i.e.
$\Pr(C_0) = \Pr(C_1) = \ldots = \Pr(C_6) = \frac{1}{7}$.

In fact, we will use this prior specification for the problem.
But, is it the most sensible?

You could take the view that it's quite likely that **all balls in the bag are likely to be of the same colour**, and consequently give higher prior probabilities to $\Pr(C_0)$ and $\Pr(C_7)$.

Example 2

Without any information to the contrary, we might well assume that **all possible numbers are equally likely**, i.e. $\Pr(C_0) = \Pr(C_1) = \ldots = \Pr(C_6) = \frac{1}{7}$.

In fact, we will use this prior specification for the problem.
But, is it the most sensible?

You could take the view that it's quite likely that **all balls in the bag are likely to be of the same colour**, and consequently give higher prior probabilities to $\Pr(C_0)$ and $\Pr(C_7)$.

Or you could find out from the ball manufacturers that **they produce balls of 10 different colours**. You might then take the prior view that each ball is black with probability $\frac{1}{10}$.

Example 2

Without any information to the contrary, we might well assume that **all possible numbers are equally likely**, i.e. $\Pr(C_0) = \Pr(C_1) = \ldots = \Pr(C_6) = \frac{1}{7}$.

In fact, we will use this prior specification for the problem.
But, is it the most sensible?

You could take the view that it's quite likely that **all balls in the bag are likely to be of the same colour**, and consequently give higher prior probabilities to $\Pr(C_0)$ and $\Pr(C_7)$.

Or you could find out from the ball manufacturers that **they produce balls of 10 different colours**. You might then take the prior view that each ball is black with probability $\frac{1}{10}$.

The point is we have to *think hard* about **how to express our prior beliefs**, since the answer will depend on that.

Example 2

Apply **Bayes' Theorem**:

$$\begin{aligned}
\Pr(C_3|A) &= \frac{\Pr(C_3)\Pr(A|C_3)}{\sum_{j=0}^{6}\Pr(A|C_j)\Pr(C_j)} \\
&= \frac{\frac{1}{7}\times(\frac{3}{6}\times\frac{2}{5}\times\frac{1}{4})}{\frac{1}{7}\{3\times 0+(\frac{3}{6}\times\frac{2}{5}\times\frac{1}{4})+(\frac{4}{6}\times\frac{3}{5}\times\frac{2}{4})+(\frac{5}{6}\times\frac{4}{5}\times\frac{3}{4})+1} \\
&= \frac{1}{35}.
\end{aligned}$$

Thus, the data has updated our prior belief of $\Pr(C_3) = \frac{1}{7}$ to the posterior probability $\Pr(C_3|A) = \frac{1}{35}$. That is, the event is much less likely having seen the data than it was previously.

Example 2

Calculation of posterior probabilities:

C_i	C_0	C_1	C_2	C_3	C_4	C_5	C_6
$\Pr(C_i)$	$1/7$	$1/7$	$1/7$	$1/7$	$1/7$	$1/7$	$1/7$

Example 2

Calculation of posterior probabilities:

C_i	C_0	C_1	C_2	C_3	C_4	C_5	C_6
$\Pr(C_i)$	$1/7$	$1/7$	$1/7$	$1/7$	$1/7$	$1/7$	$1/7$
$\Pr(A \mid C_i)$	0	0	0	$1/20$	$1/5$	$1/2$	1

Example 2

Calculation of posterior probabilities:

C_i	C_0	C_1	C_2	C_3	C_4	C_5	C_6
$\Pr(C_i)$	$1/7$	$1/7$	$1/7$	$1/7$	$1/7$	$1/7$	$1/7$
$\Pr(A \mid C_i)$	0	0	0	$1/20$	$1/5$	$1/2$	1
$\Pr(A, C_i)$	0	0	0	$1/140$	$1/35$	$1/14$	$1/7$

Example 2

Calculation of posterior probabilities:

C_i	C_0	C_1	C_2	C_3	C_4	C_5	C_6
$\Pr(C_i)$	1/7	1/7	1/7	1/7	1/7	1/7	1/7
$\Pr(A \mid C_i)$	0	0	0	1/20	1/5	1/2	1
$\Pr(A, C_i)$	0	0	0	1/140	1/35	1/14	1/7
$\Pr(C_i \mid A)$	0	0	0	1/35	4/35	10/35	20/35

Note: $\Pr(A) = \frac{1}{140} + \frac{1}{35} + \frac{1}{14} + \frac{1}{7} = \frac{35}{140}$

Example 2

First alternative prior: all the balls in the bag are of the same colour.

C_i	C_0	C_1	C_2	C_3	C_4	C_5	C_6
$\Pr(C_i)$	0.5	0	0	0	0	0	0.5

Example 2

First alternative prior: all the balls in the bag are of the same colour.

C_i	C_0	C_1	C_2	C_3	C_4	C_5	C_6
$\Pr(C_i)$	0.5	0	0	0	0	0	0.5
$\Pr(A \mid C_i)$	0	0	0	1/20	1/5	1/2	1

Example 2

First alternative prior: all the balls in the bag are of the same colour.

C_i	C_0	C_1	C_2	C_3	C_4	C_5	C_6
$\Pr(C_i)$	0.5	0	0	0	0	0	0.5
$\Pr(A \mid C_i)$	0	0	0	1/20	1/5	1/2	1
$\Pr(A, C_i)$	0	0	0	0	0	0	0.5

Example 2

First alternative prior: all the balls in the bag are of the same colour.

C_i	C_0	C_1	C_2	C_3	C_4	C_5	C_6
$\Pr(C_i)$	0.5	0	0	0	0	0	0.5
$\Pr(A \mid C_i)$	0	0	0	$1/20$	$1/5$	$1/2$	1
$\Pr(A, C_i)$	0	0	0	0	0	0	0.5
$\Pr(C_i \mid A)$	0	0	0	0	0	0	1

Example 2

Second alternative prior: With probability $\frac{1}{10}$ each ball is black and with probability $\frac{9}{10}$ it isn't.

Example 2

Second alternative prior: With probability $\frac{1}{10}$ each ball is black and with probability $\frac{9}{10}$ it isn't.

$$\Pr(C_i) = \binom{6}{i}\left(\frac{1}{10}\right)^i\left(\frac{9}{10}\right)^{6-i}$$

Example 2

Second alternative prior: With probability $\frac{1}{10}$ each ball is black and with probability $\frac{9}{10}$ it isn't.

$$\Pr(C_i) = \binom{6}{i} \left(\frac{1}{10}\right)^i \left(\frac{9}{10}\right)^{6-i}$$

C_i	C_0	C_1	C_2	C_3	C_4	C_5	C_6
$\Pr(C_i)$	0.53	0.35	0.098	0.015	0.0012	0.0001	¡0.0000

Example 2

Second alternative prior: With probability $\frac{1}{10}$ each ball is black and with probability $\frac{9}{10}$ it isn't.

$$\Pr(C_i) = \binom{6}{i} \left(\frac{1}{10}\right)^i \left(\frac{9}{10}\right)^{6-i}$$

C_i	C_0	C_1	C_2	C_3	C_4	C_5	C_6
$\Pr(C_i)$	0.53	0.35	0.098	0.015	0.0012	0.0001	¡0.0000
$\Pr(A \mid C_i)$	0	0	0	1/20	1/5	1/2	1

Example 2

Second alternative prior: With probability $\frac{1}{10}$ each ball is black and with probability $\frac{9}{10}$ it isn't.

$$\Pr(C_i) = \binom{6}{i}\left(\frac{1}{10}\right)^i\left(\frac{9}{10}\right)^{6-i}$$

C_i	C_0	C_1	C_2	C_3	C_4	C_5	C_6
$\Pr(C_i)$	0.53	0.35	0.098	0.015	0.0012	0.0001	¡0.0000
$\Pr(A \mid C_i)$	0	0	0	1/20	1/5	1/2	1
$\Pr(C_i \mid A)$	0	0	0	0.7290	0.2430	0.0270	0.0010

Example 3

A seed collector, who has acquired a small number of seeds from a plant, has a prior belief that the probability θ of germination of each seed is uniform over the range $0 \leq \theta \leq 1$. She experiments by sowing two seeds and finds that they both germinate.
(i) Write down the likelihood function for θ deriving from this observation, and obtain the collector's posterior distribution of θ.
(ii) Compute the posterior probability that θ is less than one half and compare it with the prior probability that θ is less than a half.

Example 3

A seed collector, who has acquired a small number of seeds from a plant, has a prior belief that the probability θ of germination of each seed is uniform over the range $0 \leq \theta \leq 1$. She experiments by sowing two seeds and finds that they both germinate.
(i) Write down the likelihood function for θ deriving from this observation, and obtain the collector's posterior distribution of θ.
(ii) Compute the posterior probability that θ is less than one half and compare it with the prior probability that θ is less than a half.

X: the number of seeds that germinate in the sample of 2 seeds

$$X \sim Binomial(2, \theta)$$

θ: the probability of germination ($0 \leq \theta \leq 1$)

Example 3

Binomial model:

$$f(x \mid \theta) = \binom{2}{x} \theta^x (1-\theta)^{2-x}$$

likelihood ($X = 2$): $f(x = 2 \mid \theta) = \theta^2$

Example 3

Binomial model:

$$f(x \mid \theta) = \binom{2}{x} \theta^x (1-\theta)^{2-x}$$

likelihood ($X = 2$): $f(x = 2 \mid \theta) = \theta^2$

prior of θ : $f(\theta) = 1,\ 0 \leq \theta \leq 1$

Example 3

Binomial model:

$$f(x \mid \theta) = \binom{2}{x} \theta^x (1-\theta)^{2-x}$$

likelihood ($X = 2$): $f(x = 2 \mid \theta) = \theta^2$

prior of θ : $f(\theta) = 1,\ 0 \leq \theta \leq 1$

likelihood x prior: $f(x \mid \theta) f(\theta) = \theta^2$

Example 3

Binomial model:

$$f(x \mid \theta) = \binom{2}{x} \theta^x (1-\theta)^{2-x}$$

likelihood ($X = 2$): $f(x = 2 \mid \theta) = \theta^2$

prior of θ : $f(\theta) = 1,\ 0 \leq \theta \leq 1$

likelihood x prior: $f(x \mid \theta) f(\theta) = \theta^2$

posterior: $f(\theta \mid x) = \frac{f(x|\theta)f(\theta)}{f(x)} = 3\theta^2$, since
$f(x) = \int_0^1 f(x \mid \theta) f(\theta) d\theta = \int_0^1 \theta^2 d\theta = \frac{1}{3}$

Example 3

Binomial model:

$$f(x \mid \theta) = \binom{2}{x} \theta^x (1-\theta)^{2-x}$$

likelihood ($X = 2$): $f(x = 2 \mid \theta) = \theta^2$

prior of θ : $f(\theta) = 1,\ 0 \leq \theta \leq 1$

likelihood x prior: $f(x \mid \theta) f(\theta) = \theta^2$

posterior: $f(\theta \mid x) = \frac{f(x|\theta) f(\theta)}{f(x)} = 3\theta^2$, since
$f(x) = \int_0^1 f(x \mid \theta) f(\theta) d\theta = \int_0^1 \theta^2 d\theta = \frac{1}{3}$

Then, $\Pr(\theta < 1/2) = \int_0^{1/2} f(\theta) d\theta = \int_0^{1/2} d\theta = 1/2$

Example 3

Binomial model:

$$f(x \mid \theta) = \binom{2}{x} \theta^x (1-\theta)^{2-x}$$

likelihood ($X = 2$): $f(x = 2 \mid \theta) = \theta^2$

prior of θ : $f(\theta) = 1,\ 0 \leq \theta \leq 1$

likelihood x prior: $f(x \mid \theta) f(\theta) = \theta^2$

posterior: $f(\theta \mid x) = \frac{f(x|\theta)f(\theta)}{f(x)} = 3\theta^2$, since
$f(x) = \int_0^1 f(x \mid \theta) f(\theta) d\theta = \int_0^1 \theta^2 d\theta = \frac{1}{3}$

Then, $\Pr(\theta < 1/2) = \int_0^{1/2} f(\theta) d\theta = \int_0^{1/2} d\theta = 1/2$
$\Pr(\theta < 1/2 \mid x) = \int_0^{1/2} f(\theta \mid x) d\theta = \int_0^{1/2} 3\theta^2 d\theta = 1/8$

Unit 2: Bayesian updating

Key steps of the Bayesian approach:

1. Specification of a likelihood model $f(x \mid \theta)$;
2. Determination of a prior $f(\theta)$;
3. Calculation of posterior distribution, $f(\theta \mid x)$ from Bayes' Theorem;
4. Drawing inferences from this posterior information.

Bayes' Theorem and Bayesian Inference

Stated in terms of random variables with densities denoted generically by f, Bayes Theorem takes the form:

$$f(\theta \mid x) = \frac{f(\theta)f(x \mid \theta)}{\int f(\theta)f(x \mid \theta)d\theta}$$

Note: We will use this notation to cover the case where x is either continuous or discrete, where in the continuous case f is the **p.d.f.** as usual, but in the discrete case, f is the **p.m.f.** of x. Similarly, θ can be discrete or continuous, but in the discrete case $\int f(\theta)f(x \mid \theta)d\theta$ is to be interpreted as $\sum_j f(\theta_j)f(x \mid \theta_j)$.

Notice that the denominator in Bayes' Theorem is a function of x only — θ having been 'integrated out'.

Bayes' Theorem and Bayesian Inference

Thus another way of writing Bayes' Theorem is

$$\begin{aligned} f(\theta \mid x) &= cf(\theta)f(x \mid \theta) \\ &\propto f(\theta)f(x \mid \theta) = h(\theta) \end{aligned}$$

'the posterior is proportional to the prior times the likelihood'.

Bayes' Theorem and Bayesian Inference

Thus another way of writing Bayes' Theorem is

$$\begin{aligned} f(\theta \mid x) &= cf(\theta)f(x \mid \theta) \\ &\propto f(\theta)f(x \mid \theta) = h(\theta) \end{aligned}$$

'the posterior is proportional to the prior times the likelihood'.

The constant of proportionality c, which may depend on x but not θ, is a *normalising constant* (makes the posterior integrate to one).

Note: There is a unique pdf, say $g(\theta)$ which is proportional to any given function $h(\theta)$, because $g(\theta)$ can be determined uniquely as $g(\theta) = ch(\theta)$ where $c = 1/\int h(\theta)d\theta$.

Bayes' Theorem and Bayesian Inference

Thus another way of writing Bayes' Theorem is

$$\begin{aligned} f(\theta \mid x) &= cf(\theta)f(x \mid \theta) \\ &\propto f(\theta)f(x \mid \theta) = h(\theta) \end{aligned}$$

'the posterior is proportional to the prior times the likelihood'.

The constant of proportionality c, which may depend on x but not θ, is a *normalising constant* (makes the posterior integrate to one).

Note: There is a unique pdf, say $g(\theta)$ which is proportional to any given function $h(\theta)$, because $g(\theta)$ can be determined uniquely as $g(\theta) = ch(\theta)$ where $c = 1/\int h(\theta)d\theta$.

This allows us to remove any factors of $h(\theta) = f(\theta)f(x|\theta)$, which do NOT depend upon θ, before carrying out the normalisation.

Choice of Likelihood Model

Statistical Modelling: Assume a parametric model which is suitable to describe the dynamics of the observed process. This leads to a parametric form of the likelihood function associated with the model.

Choice of Likelihood Model

Statistical Modelling: Assume a parametric model which is suitable to describe the dynamics of the observed process. This leads to a parametric form of the likelihood function associated with the model.

Therefore, the likelihood model depends on the mechanics of the problem to hand and its formulation is the same problem faced using classical inference — **what is the most suitable model for our data?**

Choice of Likelihood Model

Statistical Modelling: Assume a parametric model which is suitable to describe the dynamics of the observed process. This leads to a parametric form of the likelihood function associated with the model.

Therefore, the likelihood model depends on the mechanics of the problem to hand and its formulation is the same problem faced using classical inference — **what is the most suitable model for our data?**

Often, knowledge of the structure by which the data is obtained may suggest appropriate models (Binomial sampling, or Poisson counts, for example), but often a model will be 'hypothesised' (Y is linearly related to X with independent Normal errors, for example) and its plausibility assessed later in the context of the data.

Choice of Prior

- Because the prior represents our beliefs about θ before observing the data, it follows that the subsequent analysis is unique to us. Different priors lead to different posteriors.
- As long as the prior is not 'completely unreasonable', then the effect of the prior becomes *less influential as more data become available*.
- Often we might have a 'rough idea' what the prior should look like (perhaps we could give its mean and variance), but cannot be more precise than that. In such situations we could use a 'convenient' form for the prior which is *consistent with our beliefs, but which also makes the mathematics easy*.
- Sometimes we might feel that we have no prior information about a parameter. In such situations we might wish to use a prior which reflects our *ignorance about the parameter*.

Bayesian Computation

Though straightforward enough in principle, the implementation of Bayes' Theorem in practice can be computationally difficult, mainly as a result of the **normalizing integral** in the denominator.

Bayesian Computation

Though straightforward enough in principle, the implementation of Bayes' Theorem in practice can be computationally difficult, mainly as a result of the **normalizing integral** in the denominator.

For some choices of prior-likelihood combination, this integral can be **avoided**, but in general, specialised techniques are required to simplify this calculation (for example, **numerical or Monte Carlo integration**).

Bayesian Computation

Though straightforward enough in principle, the implementation of Bayes' Theorem in practice can be computationally difficult, mainly as a result of the **normalizing integral** in the denominator.

For some choices of prior-likelihood combination, this integral can be **avoided**, but in general, specialised techniques are required to simplify this calculation (for example, **numerical or Monte Carlo integration**).

In complex, multi-parameter problems, the multi-dimensional integral in the denominator of Bayes' theorem can be **impossible** to compute. For such problems, simulation based techniques have been developed, known as **Markov chain Monte Carlo (MCMC) methods**.

Bayesian Inference

Bayesian analysis gives a more complete inference in the sense that all knowledge about θ available from the prior and the data is represented in the posterior distribution. That is, $f(\theta|x)$ *is* the inference.

Bayesian Inference

Bayesian analysis gives a more complete inference in the sense that all knowledge about θ available from the prior and the data is represented in the posterior distribution. That is, $f(\theta|x)$ *is* the inference.

Still, it is often desirable to summarize that inference in the form of a **point estimate, or an interval estimate**.

Bayesian Inference

Bayesian analysis gives a more complete inference in the sense that all knowledge about θ available from the prior and the data is represented in the posterior distribution. That is, $f(\theta|x)$ *is* the inference.

Still, it is often desirable to summarize that inference in the form of a **point estimate, or an interval estimate**.

Moreover, desirable properties or concepts of **statistics**, functions of the data that are used for inferential purposes, are also present in Bayesian analysis. For example, the concept of sufficiency has analogous role in Bayesian inference, but is more intuitively appealing. It can be characterised by saying that if we partition our data by $x = (x_1, x_2)$, then x_1 is sufficient for θ if $f(\theta|x)$ depends only on x_1 and does not depend on x_2.

Example 1. Binomial Sample

Suppose our likelihood model is $X \sim Binomial(n, \theta)$, and we wish to make inferences about θ, from a single observation x. So,

$$f(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}; \qquad x = 0, \ldots, n.$$

Example 1. Binomial Sample

Suppose our likelihood model is $X \sim Binomial(n, \theta)$, and we wish to make inferences about θ, from a single observation x. So,

$$f(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}; \qquad x = 0, \ldots, n.$$

As prior distribution for θ we will consider the Beta distribution:

$$\theta \sim Beta(p, q), \quad p > 0, \quad q > 0.$$

so that

$$\begin{aligned} f(\theta) &= \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \theta^{p-1} (1-\theta)^{q-1} \qquad (0 \leq \theta \leq 1) \\ &\propto \theta^{p-1}(1-\theta)^{q-1}. \end{aligned}$$

The Beta Distribution

The Beta distribution is also written

$$f(\theta) = \frac{\theta^{p-1}(1-\theta)^{q-1}}{B(p,q)}, \quad \text{where}$$

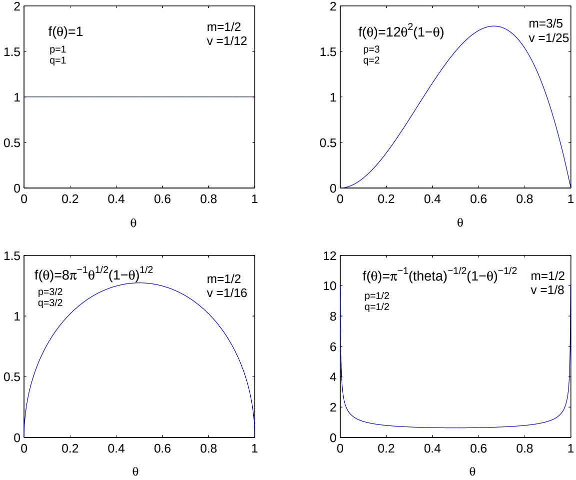
$$B(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)} = \int_0^1 \theta^{p-1}(1-\theta)^{q-1}\mathrm{d}\theta.$$

We call $B(p,q)$ the beta function.

The mean and variance of this distribution are

$$E(\theta) = m = \frac{p}{p+q} \quad \text{and} \quad Var(\theta) = v = \frac{pq}{(p+q)^2(p+q+1)}.$$

Cases of the Beta Distribution



The Posterior Distribution

$$
\begin{aligned}
f(\theta|x) &\propto f(\theta)f(x|\theta) \\
&\propto \theta^{p-1}(1-\theta)^{q-1} \times \theta^{x}(1-\theta)^{n-x} \\
&= \theta^{p+x-1}(1-\theta)^{q+n-x-1} \\
&= \theta^{P-1}(1-\theta)^{Q-1}
\end{aligned}
$$

where $P = p + x$ and $Q = q + n - x$.

The Posterior Distribution

$$
\begin{aligned}
f(\theta|x) &\propto f(\theta)f(x|\theta) \\
&\propto \theta^{p-1}(1-\theta)^{q-1} \times \theta^{x}(1-\theta)^{n-x} \\
&= \theta^{p+x-1}(1-\theta)^{q+n-x-1} \\
&= \theta^{P-1}(1-\theta)^{Q-1}
\end{aligned}
$$

where $P = p + x$ and $Q = q + n - x$.

There is only one density function proportional to this, so it must be the case that

$$\theta|x \sim Beta(P, Q).$$

Some Notes

Thus, by careful choice, we have obtained a posterior distribution which is **in the same family** as the prior distribution, and in doing so have avoided the need to calculate explicitly any integrals for the normalising constant.

Some Notes

Thus, by careful choice, we have obtained a posterior distribution which is **in the same family** as the prior distribution, and in doing so have avoided the need to calculate explicitly any integrals for the normalising constant.

The effect of the data is to modify the parameters of the beta distribution from their prior values of (p, q), to the posterior values of $(p + x, q + n - x)$.

Some Notes

Thus, by careful choice, we have obtained a posterior distribution which is **in the same family** as the prior distribution, and in doing so have avoided the need to calculate explicitly any integrals for the normalising constant.

The effect of the data is to modify the parameters of the beta distribution from their prior values of (p, q), to the posterior values of $(p + x, q + n - x)$.

The posterior values $P = p + x, \quad Q = q + n - x$ involve both the data, through x and n, and the prior values p, q.

Binomial Sample: A Numerical Example

Of 70 patients given a new treatment protocol for a particular form of cancer, 34 are found to survive beyond a specified period. Denote by θ the probability of a patient's survival.

Binomial Sample: A Numerical Example

Of 70 patients given a new treatment protocol for a particular form of cancer, 34 are found to survive beyond a specified period. Denote by θ the probability of a patient's survival.

Medical experts, who are familiar with similar trials, express the prior belief that $E(\theta) = 0.4$ and $Var(\theta) = 0.02$.

Binomial Sample: A Numerical Example

Of 70 patients given a new treatment protocol for a particular form of cancer, 34 are found to survive beyond a specified period. Denote by θ the probability of a patient's survival.

Medical experts, who are familiar with similar trials, express the prior belief that $E(\theta) = 0.4$ and $Var(\theta) = 0.02$.

Now, if a beta distribution is reasonable for their prior beliefs, then we should choose $\theta \sim Beta(p, q)$ such that

$$E(\theta) = m = \frac{p}{p+q} = 0.4 \text{ and } Var(\theta) = v = \frac{pq}{(p+q)^2(p+q+1)} = 0.02$$

These equations are solved by

$$p = \frac{(1-m)m^2}{v} - m = 4.4 \quad \text{and} \quad q = \frac{(1-m)^2 m}{v} - (1-m) = 6.6,$$

Binomial Sample: A Numerical Example

Then, the posterior is Beta(P, Q) with updated parameters $P = 4.4 + 34 = 38.4$ and $Q = 6.6 + 70 - 34 = 42.6$.

This posterior distribution summarizes all available information about θ and represents the complete inference about θ.

Binomial Sample: A Numerical Example

Then, the posterior is Beta(P, Q) with updated parameters $P = 4.4 + 34 = 38.4$ and $Q = 6.6 + 70 - 34 = 42.6$.

This posterior distribution summarizes all available information about θ and represents the complete inference about θ.

By comparing prior and posterior expectations we can see:

$$E(\theta|x) = \frac{P}{P+Q} = 0.474 > E(\theta) = \frac{p}{p+q} = 0.4.$$

Binomial Sample: A Numerical Example

Then, the posterior is Beta(P, Q) with updated parameters $P = 4.4 + 34 = 38.4$ and $Q = 6.6 + 70 - 34 = 42.6$.

This posterior distribution summarizes all available information about θ and represents the complete inference about θ.

By comparing prior and posterior expectations we can see:

$$E(\theta|x) = \frac{P}{P+Q} = 0.474 > E(\theta) = \frac{p}{p+q} = 0.4.$$

The effect of the observed data has been to increase the prior estimate of θ from 0.4 to 0.474. On the other hand, a natural estimate for θ on the basis of the data only is $x/n = 0.486$, which is the **M.L.E** $\hat{\theta}$.

Some Notes

Actually,

$$E(\theta|x) = \frac{P}{P+Q} = \frac{p+x}{p+q+n}.$$

Thus, the posterior estimate is a **balance between our prior beliefs and the information provided by the data**.

Some Notes

Actually,

$$E(\theta|x) = \frac{P}{P+Q} = \frac{p+x}{p+q+n}.$$

Thus, the posterior estimate is a **balance between our prior beliefs and the information provided by the data**.

More generally, if x and n are large relative to p and q then the posterior expectation is approximately x/n, the M.L.E.

Some Notes

Actually,

$$E(\theta|x) = \frac{P}{P+Q} = \frac{p+x}{p+q+n}.$$

Thus, the posterior estimate is a **balance between our prior beliefs and the information provided by the data**.

More generally, if x and n are large relative to p and q then the posterior expectation is approximately x/n, the M.L.E.

On the other hand, if p and q are moderately large then they will have reasonable influence on the posterior mean.

Example 2. Poisson Sample

Suppose we have a random sample (i.e. independent observations) of size n, $x = (x_1, x_2, \ldots, x_n)$ of a random variable X whose distribution is Poisson(θ) , so that

$$f(x \mid \theta) = \frac{\theta^x e^{-\theta}}{x!}, \quad \theta \geq 0.$$

The mean and the variance of this distribution are:

$$E(X) = Var(X) = \theta$$

Example 2. Poisson Sample

Suppose we have a random sample (i.e. independent observations) of size n, $x = (x_1, x_2, \ldots, x_n)$ of a random variable X whose distribution is Poisson(θ) , so that

$$f(x \mid \theta) = \frac{\theta^x e^{-\theta}}{x!}, \quad \theta \geq 0.$$

The mean and the variance of this distribution are:

$$E(X) = Var(X) = \theta$$

The likelihood is

$$f(x|\theta) = \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!} \propto e^{-n\theta}\theta^{\Sigma x_i}$$

The Prior Distribution

Prior beliefs about θ will vary from problem to problem, but we'll look for a form which gives a range of different possibilities, but is also mathematically tractable.

The Prior Distribution

Prior beliefs about θ will vary from problem to problem, but we'll look for a form which gives a range of different possibilities, but is also mathematically tractable.

We consider a gamma prior distribution:

$$\theta \sim Gamma(p, q),$$

so

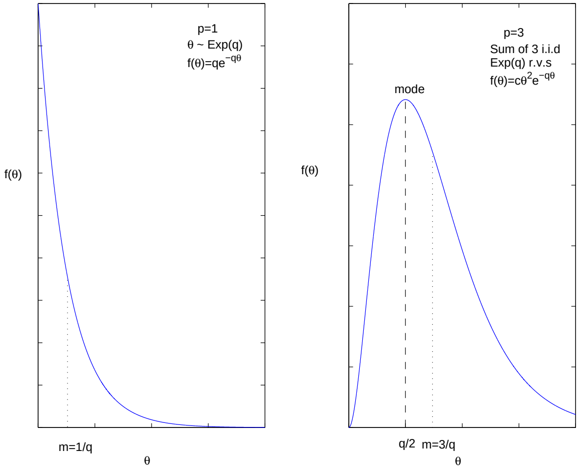
$$f(\theta) = \frac{q^p}{\Gamma(p)} \theta^{p-1} \exp\{-q\theta\}, \quad \theta > 0.$$

The parameter $p > 0$ is a shape parameter, and $q > 0$ is a scale parameter.

The mean and variance of this distribution are

$$E(\theta) = m = \frac{p}{q} \text{ and } Var(\theta) = v = \frac{p}{q^2}.$$

Examples of the Gamma Distribution



The Posterior Distribution

Applying Bayes' Theorem with the gamma prior distribution,

$$\begin{aligned} f(\theta|x) &\propto \theta^{p-1}\exp\{-q\theta\} \times \exp\{-n\theta\}\theta^{\Sigma x_i} \\ &= \theta^{(p+\Sigma x_i-1)}\exp\{-(q+n)\theta\} \\ &= \theta^{P-1}\exp(-Q\theta) \end{aligned}$$

where $P = p + \sum x_i$ and $Q = q + n$.

The Posterior Distribution

Applying Bayes' Theorem with the gamma prior distribution,

$$\begin{aligned} f(\theta|x) &\propto \theta^{p-1}\exp\{-q\theta\} \times \exp\{-n\theta\}\theta^{\Sigma x_i} \\ &= \theta^{(p+\Sigma x_i - 1)}\exp\{-(q+n)\theta\} \\ &= \theta^{P-1}\exp(-Q\theta) \end{aligned}$$

where $P = p + \sum x_i$ and $Q = q + n$.

Again, there is only one p.d.f. proportional to this:

$$\theta|x \sim Gamma(P, Q),$$

a gamma distribution whose parameters are modified by the sum of the data, $\sum_{i=1}^{n} x_i$, and the sample size n. (Note that $\sum x_i$ is sufficient for θ).

Example 3. Normal Mean

Let $x = (x_1, x_2, \ldots, x_n)$ be a random sample of size n of a random variable X with the $N(\theta, \sigma^2)$ distribution, where σ^2 is known:

$$f(x \mid \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\theta)^2}{2\sigma^2}\right\}.$$

The mean and the variance of this distribution are:
$E(X) = \theta$ and $Var(X) = \sigma^2$.

The likelihood of θ from a single observation x_i is given by

$$\begin{aligned} f(x_i \mid \theta) &= \tfrac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\tfrac{(x_i-\theta)^2}{2\sigma^2}\right\} \\ &\propto \exp\left\{-\tfrac{1}{2\sigma^2}(x_i^2 - 2x_i\theta + \theta^2)\right\} \end{aligned}$$

Example 3. Normal Mean

Let $x = (x_1, x_2, \ldots, x_n)$ be a random sample of size n of a random variable X with the $N(\theta, \sigma^2)$ distribution, where σ^2 is known:

$$f(x \mid \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\theta)^2}{2\sigma^2}\right\}.$$

The mean and the variance of this distribution are:
$E(X) = \theta$ and $Var(X) = \sigma^2$.

The likelihood of θ from a single observation x_i is given by

$$\begin{aligned} f(x_i \mid \theta) &= \tfrac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\tfrac{(x_i-\theta)^2}{2\sigma^2}\right\} \\ &\propto \exp\left\{-\tfrac{1}{2\sigma^2}(x_i^2 - 2x_i\theta + \theta^2)\right\} \\ &= \exp\left(-\tfrac{1}{2\sigma^2}x_i^2\right) \exp\left(\tfrac{1}{\sigma^2}x_i\theta - \tfrac{1}{2\sigma^2}\theta^2\right) \\ &\propto \exp\left(-\tfrac{1}{2\sigma^2}\theta^2 + \tfrac{1}{\sigma^2}x_i\theta\right). \end{aligned}$$

The Likelihood and the Prior

The likelihood of the whole sample is then

$$\begin{aligned} f(x \mid \theta) &= \prod_i f(x_i \mid \theta) \\ &\propto \prod_i \exp\left(-\frac{1}{2\sigma^2}\theta^2 + \frac{1}{\sigma^2}x_i\theta\right) \end{aligned}$$

The Likelihood and the Prior

The likelihood of the whole sample is then

$$\begin{aligned} f(x \mid \theta) &= \prod_i f(x_i \mid \theta) \\ &\propto \prod_i \exp\left(-\frac{1}{2\sigma^2}\theta^2 + \frac{1}{\sigma^2}x_i\theta\right) \\ &= \exp\left[\sum_i \left(-\frac{1}{2\sigma^2}\theta^2 + \frac{1}{\sigma^2}x_i\theta\right)\right] \\ &= \exp\left[-\left(\frac{n}{2\sigma^2}\right)\theta^2 + \left(\frac{1}{\sigma^2}\sum_i x_i\right)\theta\right]. \end{aligned}$$

The Likelihood and the Prior

The likelihood of the whole sample is then

$$\begin{aligned} f(x \mid \theta) &= \prod_i f(x_i \mid \theta) \\ &\propto \prod_i \exp\left(-\tfrac{1}{2\sigma^2}\theta^2 + \tfrac{1}{\sigma^2}x_i\theta\right) \\ &= \exp\left[\sum_i \left(-\tfrac{1}{2\sigma^2}\theta^2 + \tfrac{1}{\sigma^2}x_i\theta\right)\right] \\ &= \exp\left[-\left(\tfrac{n}{2\sigma^2}\right)\theta^2 + \left(\tfrac{1}{\sigma^2}\sum_i x_i\right)\theta\right]. \end{aligned}$$

Suppose our prior beliefs about θ can be represented as a normal distribution: $\theta \sim N(b, d^2)$. Then,

$$f(\theta) = \frac{1}{\sqrt{2\pi}d}\exp\left\{-\frac{(\theta-b)^2}{2d^2}\right\}$$

The Likelihood and the Prior

The likelihood of the whole sample is then

$$
\begin{aligned}
f(x \mid \theta) &= \prod_i f(x_i \mid \theta) \\
&\propto \prod_i \exp\left(-\frac{1}{2\sigma^2}\theta^2 + \frac{1}{\sigma^2}x_i\theta\right) \\
&= \exp\left[\sum_i \left(-\frac{1}{2\sigma^2}\theta^2 + \frac{1}{\sigma^2}x_i\theta\right)\right] \\
&= \exp\left[-\left(\frac{n}{2\sigma^2}\right)\theta^2 + \left(\frac{1}{\sigma^2}\sum_i x_i\right)\theta\right].
\end{aligned}
$$

Suppose our prior beliefs about θ can be represented as a normal distribution: $\theta \sim N(b, d^2)$. Then,

$$
\begin{aligned}
f(\theta) &= \frac{1}{\sqrt{2\pi}d}\exp\left\{-\frac{(\theta-b)^2}{2d^2}\right\} \\
&\propto \exp\left(-\frac{1}{2d^2}\theta^2 + \frac{1}{d^2}\,b\,\theta - \frac{1}{2d^2}b^2\right) \\
&\propto \exp\left(-\frac{1}{2d^2}\theta^2 + \frac{1}{d^2}\,b\,\theta\right).
\end{aligned}
$$

The Posterior

We now derive the posterior distribution of θ as follows

$$\begin{aligned} f(\theta \mid x) &\propto f(\theta) f(x \mid \theta) \\ &\propto \exp\left(-\frac{1}{2d^2}\theta^2 + \frac{1}{d^2} b\theta\right) \exp\left[-\left(\frac{n}{2\sigma^2}\right)\theta^2 + \left(\frac{1}{\sigma^2}\sum_i x_i\right)\theta\right] \end{aligned}$$

The Posterior

We now derive the posterior distribution of θ as follows

$$
\begin{aligned}
f(\theta \mid x) &\propto f(\theta) f(x \mid \theta) \\
&\propto \exp\left(-\frac{1}{2d^2}\theta^2 + \frac{1}{d^2} b\theta\right) \exp\left[-\left(\frac{n}{2\sigma^2}\right)\theta^2 + \left(\frac{1}{\sigma^2}\sum_i x_i\right)\theta\right] \\
&= \exp\left[-\frac{1}{2}\left(\frac{1}{d^2} + \frac{n}{\sigma^2}\right)\theta^2 + \left(\frac{1}{d^2} b + \frac{1}{\sigma^2}\sum_i x_i\right)\theta\right]
\end{aligned}
$$

The Posterior

We now derive the posterior distribution of θ as follows

$$\begin{aligned}
f(\theta \mid x) &\propto f(\theta) f(x \mid \theta) \\
&\propto \exp\left(-\frac{1}{2d^2}\theta^2 + \frac{1}{d^2} b\theta\right) \exp\left[-\left(\frac{n}{2\sigma^2}\right)\theta^2 + \left(\frac{1}{\sigma^2}\sum_i x_i\right)\theta\right] \\
&= \exp\left[-\frac{1}{2}\left(\frac{1}{d^2} + \frac{n}{\sigma^2}\right)\theta^2 + \left(\frac{1}{d^2} b + \frac{1}{\sigma^2}\sum_i x_i\right)\theta\right] \\
&= \exp\left(-\frac{1}{2D^2}\theta^2 + \frac{1}{D^2} B\theta\right).
\end{aligned}$$

The Posterior

We now derive the posterior distribution of θ as follows

$$\begin{aligned}
f(\theta \mid x) &\propto f(\theta)f(x \mid \theta) \\
&\propto \exp\left(-\tfrac{1}{2d^2}\,\theta^2 + \tfrac{1}{d^2}\,b\theta\right)\exp\left[-\left(\tfrac{n}{2\sigma^2}\right)\theta^2 + \left(\tfrac{1}{\sigma^2}\sum_i x_i\right)\theta\right] \\
&= \exp\left[-\tfrac{1}{2}\left(\tfrac{1}{d^2} + \tfrac{n}{\sigma^2}\right)\theta^2 + \left(\tfrac{1}{d^2}\,b + \tfrac{1}{\sigma^2}\sum_i x_i\right)\theta\right] \\
&= \exp\left(-\tfrac{1}{2D^2}\theta^2 + \tfrac{1}{D^2}B\theta\right).
\end{aligned}$$

Therefore, we can conclude that the posterior distribution of θ is

$$\theta|x \;\sim\; N(B, D^2)$$

where

$$B = E(\theta|x) = \frac{\frac{1}{d^2}b + \frac{n}{\sigma^2}\bar{x}}{\frac{1}{d^2} + \frac{n}{\sigma^2}} \quad \text{and} \quad D^2 = V(\theta|x) = \left(\frac{1}{d^2} + \frac{n}{\sigma^2}\right)^{-1},$$

and we have replaced $\sum x_i$ by $n\bar{x}$.

The Precision

This result is expressed more concisely if we define **'precision'** to be the reciprocal of variance. Let $\tau = 1/\sigma^2$ and $c = 1/d^2$, then

$X \sim N(\theta, \tau^{-1})$ and $f(x_i \mid \theta) \propto \exp\left\{-\frac{\tau\theta^2}{2} + \tau x_i \theta\right\}$

$\theta \sim N(b, c^{-1})$ and $f(\theta) \propto \exp\left\{-\frac{c\theta^2}{2} + cb\theta\right\}$

The Precision

This result is expressed more concisely if we define **'precision'** to be the reciprocal of variance. Let $\tau = 1/\sigma^2$ and $c = 1/d^2$, then

$X \sim N(\theta, \tau^{-1})$ and $f(x_i \mid \theta) \propto \exp\left\{-\frac{\tau\theta^2}{2} + \tau x_i \theta\right\}$

$\theta \sim N(b, c^{-1})$ and $f(\theta) \propto \exp\left\{-\frac{c\theta^2}{2} + cb\theta\right\}$

The posterior is obtained as

$f(\theta \mid x) \propto \exp\left\{-\frac{(n\tau+c)\theta^2}{2} + (n\tau\bar{x} + cb)\theta\right\}$, that is

$$\theta | x \sim N\left(\frac{cb + n\tau\bar{x}}{c + n\tau}, \frac{1}{c + n\tau}\right)$$

Some Notes

1. $E(\theta|x) = \frac{c}{c+n\tau}b + (1 - \frac{c}{c+n\tau})\bar{x} = \gamma_n b + (1 - \gamma_n)\bar{x}$.
   The posterior mean is a weighted average of the prior mean and $\bar{x}$. If $n\tau$ is large relative to c, then $\gamma_n \approx 0$ and the posterior mean is close to $\bar{x}$.

Some Notes

1. $E(\theta|x) = \frac{c}{c+n\tau}b + (1 - \frac{c}{c+n\tau})\bar{x} = \gamma_n b + (1 - \gamma_n)\bar{x}$.
   The posterior mean is a weighted average of the prior mean and $\overline{x}$. If $n\tau$ is large relative to c, then $\gamma_n \approx 0$ and the posterior mean is close to $\overline{x}$.
2. 'posterior precision': $[Var(\theta \mid x)]^{-1} = c + n\tau$
   If $n\tau$ is large relative to c, then $Var(\theta \mid x) \approx \frac{\sigma^2}{n}$.

Some Notes

1. $E(\theta|x) = \frac{c}{c+n\tau}b + (1 - \frac{c}{c+n\tau})\bar{x} = \gamma_n b + (1 - \gamma_n)\bar{x}$.
   The posterior mean is a weighted average of the prior mean and $\overline{x}$. If $n\tau$ is large relative to c, then $\gamma_n \approx 0$ and the posterior mean is close to $\overline{x}$.
2. 'posterior precision': $[Var(\theta \mid x)]^{-1} = c + n\tau$
   If $n\tau$ is large relative to c, then $Var(\theta \mid x) \approx \frac{\sigma^2}{n}$.
3. As $n \to \infty$, then (loosely) $\theta|x \sim N(\overline{x}, \frac{\sigma^2}{n})$, so that the prior has no effect in the limit.

Some Notes

1. $E(\theta|x) = \frac{c}{c+n\tau}b + (1 - \frac{c}{c+n\tau})\bar{x} = \gamma_n b + (1-\gamma_n)\overline{x}$.
   The posterior mean is a weighted average of the prior mean and $\overline{x}$. If $n\tau$ is large relative to c, then $\gamma_n \approx 0$ and the posterior mean is close to $\overline{x}$.
2. 'posterior precision': $[Var(\theta \mid x)]^{-1} = c + n\tau$
   If $n\tau$ is large relative to c, then $Var(\theta \mid x) \approx \frac{\sigma^2}{n}$.
3. As $n \to \infty$, then (loosely) $\theta|x \sim N(\overline{x}, \frac{\sigma^2}{n})$, so that the prior has no effect in the limit.
4. As $d \to \infty$ $(c \to 0)$, we again obtain $\theta|x \sim N(\overline{x}, \frac{\sigma^2}{n})$

Some Notes

1. $E(\theta|x) = \frac{c}{c+n\tau} b + (1 - \frac{c}{c+n\tau})\bar{x} = \gamma_n b + (1 - \gamma_n)\bar{x}$.
   The posterior mean is a weighted average of the prior mean and $\bar{x}$. If $n\tau$ is large relative to c, then $\gamma_n \approx 0$ and the posterior mean is close to $\bar{x}$.
2. 'posterior precision': $[Var(\theta \mid x)]^{-1} = c + n\tau$
   If $n\tau$ is large relative to c, then $Var(\theta \mid x) \approx \frac{\sigma^2}{n}$.
3. As $n \to \infty$, then (loosely) $\theta|x \sim N(\bar{x}, \frac{\sigma^2}{n})$, so that the prior has no effect in the limit.
4. As $d \to \infty$ $(c \to 0)$, we again obtain $\theta|x \sim N(\bar{x}, \frac{\sigma^2}{n})$
5. The posterior distribution depends on the data only through $\bar{x}$ and not through the individual values of the x_i themselves. We say that $\bar{x}$ is sufficient for θ.

Sequential Updating

We have seen that Bayes' Theorem provides the machine by which your **prior information is updated by data to give your posterior information**. This then can serve as your 'new' prior information before more data become available.

Sequential Updating

We have seen that Bayes' Theorem provides the machine by which your **prior information is updated by data to give your posterior information**. This then can serve as your 'new' prior information before more data become available.

Consider two independent variables X_1 and X_2, each having density $f(x|\theta)$. Suppose we observe x_1, and update our prior through

$$f(\theta|x_1) \propto f(\theta)f(x_1|\theta).$$

This becomes our new **prior** before observing x_2. Thus,

$$f(\theta|x_1, x_2) \quad \propto \quad f(\theta|x_1)f(x_2|\theta) = f(\theta)f(x_1|\theta)f(x_2|\theta)$$

Sequential Updating

We have seen that Bayes' Theorem provides the machine by which your **prior information is updated by data to give your posterior information**. This then can serve as your 'new' prior information before more data become available.

Consider two independent variables X_1 and X_2, each having density $f(x|\theta)$. Suppose we observe x_1, and update our prior through

$$f(\theta|x_1) \propto f(\theta)f(x_1|\theta).$$

This becomes our new **prior** before observing x_2. Thus,

$$\begin{aligned} f(\theta|x_1, x_2) &\propto f(\theta|x_1)f(x_2|\theta) = f(\theta)f(x_1|\theta)f(x_2|\theta) \\ &= f(\theta)f(x_1, x_2|\theta) \end{aligned}$$

which is the same result we would have obtained by updating on the basis of the entire information (x_1, x_2) directly.

Sufficiency

The classical result by which we recognise that a function $s(x)$, of the data alone, is a sufficient statistic for a parameter θ, is that

$$f(x|\theta) = g(x)h(s, \theta)$$

where $g(x)$ does not involve θ, only the data.

Sufficiency

The classical result by which we recognise that a function $s(x)$, of the data alone, is a sufficient statistic for a parameter θ, is that

$$f(x|\theta) = g(x)h(s, \theta)$$

where $g(x)$ does not involve θ, only the data.

If this is the case, in the Bayesian analysis

$$f(x|\theta) \propto h(s, \theta)$$

so the likelihood depends on the data only through the sufficient statistic $s(x)$.

Sufficiency

The classical result by which we recognise that a function $s(x)$, of the data alone, is a sufficient statistic for a parameter θ, is that

$$f(x|\theta) = g(x)h(s, \theta)$$

where $g(x)$ does not involve θ, only the data.

If this is the case, in the Bayesian analysis

$$f(x|\theta) \propto h(s, \theta)$$

so the likelihood depends on the data only through the sufficient statistic $s(x)$.

In that case the posterior distribution $f(\theta|x)$ also depends on the data only through the sufficient statistic $s(x)$.

$$f(\theta \mid x) \propto f(\theta)f(x \mid \theta) \propto f(\theta)h(s, \theta)$$

The Likelihood Principle

The likelihood principle states that **if two experiments yield the same likelihood (up to proportionality), then all inferences we draw about θ should be the same in each case**.

The Likelihood Principle

The likelihood principle states that **if two experiments yield the same likelihood (up to proportionality), then all inferences we draw about θ should be the same in each case**.

A major virtue of the Bayesian framework is that **Bayesian techniques are inherently consistent with the likelihood principle**, whereas many simple procedures from classical statistics violate it.

An Example

Consider two experiments concerned with estimating the probability of success θ in independent trials. In the first experiment, the number x of successes in n trials is recorded. In the second, the number y of trials required to obtain m successes is recorded.

An Example

Consider two experiments concerned with estimating the probability of success θ in independent trials. In the first experiment, the number x of successes in n trials is recorded. In the second, the number y of trials required to obtain m successes is recorded.

The distributions of the random variables X, Y describing the outcomes of these experiments differ. They are the **Binomial** and **Negative Binomial** distributions, respectively.

$$f(x|\theta) = P(X = x) = \binom{n}{x} \theta^x (1-\theta)^{n-x}, \quad x = 0, 1, \ldots, n$$

and

$$f(y|\theta) = P(Y = y) = \binom{y-1}{m-1} \theta^m (1-\theta)^{y-m}, \quad y = m, m+1, \ldots, \infty.$$

An Example

The corresponding **M.L.E.s** are $\hat{\theta}_x = x/n$ and $\hat{\theta}_y = m/y$. However, **their sampling distributions are quite different**.

An Example

The corresponding **M.L.E.s** are $\hat{\theta}_x = x/n$ and $\hat{\theta}_y = m/y$.
However, **their sampling distributions are quite different**.

If $n = 2$, then x/n can take the values 0, 1/2 and 1.

An Example

The corresponding **M.L.E.s** are $\hat{\theta}_x = x/n$ and $\hat{\theta}_y = m/y$. However, **their sampling distributions are quite different**.

If $n = 2$, then x/n can take the values 0, 1/2 and 1.

If $m = 1$, then m/y can take the values $1, 1/2, 1/3, \ldots$.

An Example

The corresponding **M.L.E.s** are $\hat{\theta}_x = x/n$ and $\hat{\theta}_y = m/y$. However, **their sampling distributions are quite different**.

If $n = 2$, then x/n can take the values 0, 1/2 and 1.

If $m = 1$, then m/y can take the values 1, 1/2, 1/3, ...

But if it happened that also $x = 1$ and $y = 2$,

$$f(x|\theta) = 2\theta(1-\theta) \text{ and } f(y|\theta) = \theta(1-\theta)$$

so that the likelihoods are **proportional**, and the Bayesian inference would be the same in both cases.