Introduction to Bayesian Statistics

Dimitris Fouskakis
Dept. of Applied Mathematics
National Technical University of Athens
Greece
fouskakis@math.ntua.gr

Thomas Bayes

Thomas Bayes (Encyclopedia Britannica)

Born 1702, London, England.
Died April 17, 1761, Tunbridge Wells, Kent.

English Nonconformist theologian and mathematician who was the first to use probability inductively and who established a mathematical basis for probability inference (a means of calculating, from the frequency with which an event has occurred in prior trials, the probability that it will occur in future trials).

Bayes set down his findings on probability in “Essay Towards Solving a Problem in the Doctrine of Chances” (1763), published posthumously in the Philosophical Transactions of the Royal Society.

Fundamental Ideas

Bayesian statistical analysis is based on the premise that all uncertainty should be modeled with probability and that statistical inferences should be logical conclusions based on the laws of probability.

Fundamental Ideas

Bayesian statistical analysis is based on the premise that all uncertainty should be modeled with probability and that statistical inferences should be logical conclusions based on the laws of probability.

This typically involves the explicit use of subjective information provided by the scientist, since initial uncertainty about unknown parameters must be modeled from a priori expert opinions. Bayesian methodology is consistent with the goals of science.

Fundamental Ideas (cont.)

For large amounts of data, scientists with different subjective prior beliefs will ultimately agree after (separately) incorporating the data with their “prior" information.

Fundamental Ideas (cont.)

For large amounts of data, scientists with different subjective prior beliefs will ultimately agree after (separately) incorporating the data with their “prior" information.

On the other hand, “insufficient" data can result in (continued) discrepancies of opinion about the relevant scientific questions.

Fundamental Ideas (cont.)

For large amounts of data, scientists with different subjective prior beliefs will ultimately agree after (separately) incorporating the data with their “prior" information.

On the other hand, “insufficient" data can result in (continued) discrepancies of opinion about the relevant scientific questions.

We believe that the best statistical analysis of data involves a collaborative effort between subject matter scientists and statisticians, and that it is both appropriate and necessary to incorporate the scientist’s expertise into making decisions related to the data.

Simple Probability Calculations

For two events A and B the conditional probability of A given B is defined as

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

Simple Probability Calculations

For two events A and B the conditional probability of A given B is defined as

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

The simplest version of Bayes Theorem is that

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

Example: Drug Screening

Let D indicate a drug user and C indicate someone who is clean of drugs. Let $+$ indicate that someone tests positive on a drug test, and $-$ indicates testing negative.

Example: Drug Screening

Let D indicate a drug user and C indicate someone who is clean of drugs. Let $+$ indicate that someone tests positive on a drug test, and $-$ indicates testing negative.

The overall *prevalence* of a drug use in the population is, say, $P(D) = 0.01$. Therefore $P(C) = 0.99$.

Example: Drug Screening

Let D indicate a drug user and C indicate someone who is clean of drugs. Let $+$ indicate that someone tests positive on a drug test, and $-$ indicates testing negative.

The overall *prevalence* of a drug use in the population is, say, $P(D) = 0.01$. Therefore $P(C) = 0.99$.

The *sensitivity* of the drug test is $P(+|D) = 0.98$.

Example: Drug Screening

Let D indicate a drug user and C indicate someone who is clean of drugs. Let $+$ indicate that someone tests positive on a drug test, and $-$ indicates testing negative.

The overall *prevalence* of a drug use in the population is, say, $P(D) = 0.01$. Therefore $P(C) = 0.99$.

The *sensitivity* of the drug test is $P(+|D) = 0.98$.

The *specificity* of the drug test is $P(-|C) = 0.95$.

Example: Drug Screening

Let D indicate a drug user and C indicate someone who is clean of drugs. Let $+$ indicate that someone tests positive on a drug test, and $-$ indicates testing negative.

The overall *prevalence* of a drug use in the population is, say, $P(D) = 0.01$. Therefore $P(C) = 0.99$.

The *sensitivity* of the drug test is $P(+|D) = 0.98$.

The *specificity* of the drug test is $P(-|C) = 0.95$.

$$P(D|+) = \frac{P(+|D)P(D)}{P(+|D)P(D) + P(+|C)P(C)} = 0.165$$

Bayesian Statistics and Probabilities

Fundamentally, the field of statistics is about using probability models to analyze data.

Bayesian Statistics and Probabilities

Fundamentally, the field of statistics is about using probability models to analyze data.

There are two major philosophical positions about the use of probability models.

Bayesian Statistics and Probabilities

Fundamentally, the field of statistics is about using probability models to analyze data.

There are two major philosophical positions about the use of probability models.

One is that probabilities are determined by the outside world.

Bayesian Statistics and Probabilities

Fundamentally, the field of statistics is about using probability models to analyze data.

There are two major philosophical positions about the use of probability models.

One is that probabilities are determined by the outside world.

The other is that probabilities exist in peoples' heads.

Bayesian Statistics and Probabilities

Fundamentally, the field of statistics is about using probability models to analyze data.

There are two major philosophical positions about the use of probability models.

One is that probabilities are determined by the outside world.

The other is that probabilities exist in peoples' heads.

Historically, probability theory was developed to explain games of chance.

Bayesian Statistics and Probabilities (cont.)

The notion of probability as a belief is more subtle. For example, suppose I flip a coin.

Bayesian Statistics and Probabilities (cont.)

The notion of probability as a belief is more subtle. For example, suppose I flip a coin.

Prior to flipping the coin, the physical mechanism involved suggests probabilities of 0.5 for each of the outcomes heads and tails.

Bayesian Statistics and Probabilities (cont.)

The notion of probability as a belief is more subtle. For example, suppose I flip a coin.

Prior to flipping the coin, the physical mechanism involved suggests probabilities of 0.5 for each of the outcomes heads and tails.

But now I have flipped the coin, looked at the result, but not told you the outcome.

Bayesian Statistics and Probabilities (cont.)

The notion of probability as a belief is more subtle. For example, suppose I flip a coin.

Prior to flipping the coin, the physical mechanism involved suggests probabilities of 0.5 for each of the outcomes heads and tails.

But now I have flipped the coin, looked at the result, but not told you the outcome.

As long as you believe I am not cheating, you would naturally continue to describe the probabilities for heads and tails as 0.5.

Bayesian Statistics and Probabilities (cont.)

The notion of probability as a belief is more subtle. For example, suppose I flip a coin.

Prior to flipping the coin, the physical mechanism involved suggests probabilities of 0.5 for each of the outcomes heads and tails.

But now I have flipped the coin, looked at the result, but not told you the outcome.

As long as you believe I am not cheating, you would naturally continue to describe the probabilities for heads and tails as 0.5.

But this probability is no longer the probability associated with the physical mechanism involved, because you and I have different probabilities.

Bayesian Statistics and Probabilities (cont.)

The notion of probability as a belief is more subtle. For example, suppose I flip a coin.

Prior to flipping the coin, the physical mechanism involved suggests probabilities of 0.5 for each of the outcomes heads and tails.

But now I have flipped the coin, looked at the result, but not told you the outcome.

As long as you believe I am not cheating, you would naturally continue to describe the probabilities for heads and tails as 0.5.

But this probability is no longer the probability associated with the physical mechanism involved, because you and I have different probabilities.

I know whether the coin is heads or tails, and your probability is simply describing your personal state of knowledge.

Bayesian Statistics and Probabilities (cont.)

Bayesian statistics starts by using (prior) probabilities to describe your current state of knowledge.

Bayesian Statistics and Probabilities (cont.)

Bayesian statistics starts by using (prior) probabilities to describe your current state of knowledge.

It then incorporates information through the collection of data, and

Bayesian Statistics and Probabilities (cont.)

- Bayesian statistics starts by using (prior) probabilities to describe your current state of knowledge.
- It then incorporates information through the collection of data, and
- Results in new (posterior) probabilities to describe your state of knowledge after combining the prior probabilities with the data.

Bayesian Statistics and Probabilities (cont.)

Bayesian statistics starts by using (prior) probabilities to describe your current state of knowledge.

It then incorporates information through the collection of data, and

Results in new (posterior) probabilities to describe your state of knowledge after combining the prior probabilities with the data.

In Bayesian statistics, all uncertainty and all information are incorporated through the use of probability distributions, and

Bayesian Statistics and Probabilities (cont.)

Bayesian statistics starts by using (prior) probabilities to describe your current state of knowledge.

It then incorporates information through the collection of data, and

Results in new (posterior) probabilities to describe your state of knowledge after combining the prior probabilities with the data.

In Bayesian statistics, all uncertainty and all information are incorporated through the use of probability distributions, and all conclusions obey the laws of probability theory.

Data and Parameter(s)

In statistics a data set is becoming available via a random mechanism.

Data and Parameter(s)

In statistics a data set is becoming available via a random mechanism.

A model (law) $f(x|\theta)$ is used to describe the data generation procedure. The model is either available with the design (e.g. a Binomial experiment with known number of trials, where $x|\theta \sim Bin(n, \theta)$), or we need to elicit it from the data (e.g. strength required to brake a steel cord) and thus we need some assurance (testing) of whether we made the appropriate choice.

Data and Parameter(s)

The model comes along with a (univariate or multivariate) set of parameters that fully describe the random mechanism which produces the data. For example:

Data and Parameter(s)

The model comes along with a (univariate or multivariate) set of parameters that fully describe the random mechanism which produces the data. For example:

$x|\theta \sim Bin(n, \theta)$

Data and Parameter(s)

The model comes along with a (univariate or multivariate) set of parameters that fully describe the random mechanism which produces the data. For example:

$x|\theta \sim Bin(n, \theta)$

$x|\boldsymbol{\theta} \sim N(\theta_1, \theta_2)$

Data and Parameter(s)

The model comes along with a (univariate or multivariate) set of parameters that fully describe the random mechanism which produces the data. For example:

$x|\theta \sim Bin(n, \theta)$

$x|\boldsymbol{\theta} \sim N(\theta_1, \theta_2)$

$\mathbf{x}|\boldsymbol{\theta} \sim N_p(\boldsymbol{\theta}, \Sigma)$

Data and Parameter(s)

The model comes along with a (univariate or multivariate) set of parameters that fully describe the random mechanism which produces the data. For example:

$x|\theta \sim Bin(n, \theta)$

$x|\boldsymbol{\theta} \sim N(\theta_1, \theta_2)$

$\mathbf{x}|\boldsymbol{\theta} \sim N_p(\boldsymbol{\theta}, \Sigma)$

Usually we are interested in either drawing inference (point/interval estimates, hypothesis testing) for the unknown parameter θ ($\boldsymbol{\theta}$) and/or provide predictions for future observable(s).

Likelihood

Unless otherwise specified we assume that the data constitute a random sample, i.e. they are independent and identically distributed (iid) observations (given the parameter).

Likelihood

Unless otherwise specified we assume that the data constitute a random sample, i.e. they are independent and identically distributed (iid) observations (given the parameter).

Then the joint distribution of the data $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is given by:

$$f(\mathbf{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta) = L(\theta)$$

which is known as likelihood.

Likelihood

Unless otherwise specified we assume that the data constitute a random sample, i.e. they are independent and identically distributed (iid) observations (given the parameter).

Then the joint distribution of the data $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is given by:

$$f(\mathbf{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta) = L(\theta)$$

which is known as likelihood.

The likelihood is a function of the parameter θ and is considered to capture all the information that is available in the data.

Sampling Density vs. Likelihood

In statistics, we eventually get to see the data, say $d = d_{obs}$, and want to draw inferences (conclusions) about θ.

Sampling Density vs. Likelihood

In statistics, we eventually get to see the data, say $d = d_{obs}$, and want to draw inferences (conclusions) about θ.

Thus, we are interested in the values of θ that are most likely to have generated d_{obs}. Such information comes from $f(d_{obs}|\theta)$ but with d_{obs} fixed and θ allowed to vary. This new way of thinking about d and θ determines the likelihood function.

Sampling Density vs. Likelihood

In statistics, we eventually get to see the data, say $d = d_{obs}$, and want to draw inferences (conclusions) about θ.

Thus, we are interested in the values of θ that are most likely to have generated d_{obs}. Such information comes from $f(d_{obs}|\theta)$ but with d_{obs} fixed and θ allowed to vary. This new way of thinking about d and θ determines the likelihood function.

On the other hand in the sampling density $f(d|\theta)$, θ is fixed and d is the variable.

Sampling Density vs. Likelihood

In statistics, we eventually get to see the data, say $d = d_{obs}$, and want to draw inferences (conclusions) about θ.

Thus, we are interested in the values of θ that are most likely to have generated d_{obs}. Such information comes from $f(d_{obs}|\theta)$ but with d_{obs} fixed and θ allowed to vary. This new way of thinking about d and θ determines the likelihood function.

On the other hand in the sampling density $f(d|\theta)$, θ is fixed and d is the variable.

The likelihood function and the sampling density are different concepts based on the same quantity.

Example: Drugs on the job

Suppose we are interested in assessing the proportion of U.S. transportation industry workers who use drugs on the job.

Example: Drugs on the job

Suppose we are interested in assessing the proportion of U.S. transportation industry workers who use drugs on the job.

Let θ denote this proportion and assume that a random sample of n workers is to be taken while they are actually on the job.

Example: Drugs on the job

Suppose we are interested in assessing the proportion of U.S. transportation industry workers who use drugs on the job.

Let θ denote this proportion and assume that a random sample of n workers is to be taken while they are actually on the job.

Each individual will be tested for drugs.

Example: Drugs on the job

Suppose we are interested in assessing the proportion of U.S. transportation industry workers who use drugs on the job.

Let θ denote this proportion and assume that a random sample of n workers is to be taken while they are actually on the job.

Each individual will be tested for drugs.

Let y_i be a one if the ith individual tests positive and zero otherwise.

Example: Drugs on the job

Suppose we are interested in assessing the proportion of U.S. transportation industry workers who use drugs on the job.

Let θ denote this proportion and assume that a random sample of n workers is to be taken while they are actually on the job.

Each individual will be tested for drugs.

Let y_i be a one if the ith individual tests positive and zero otherwise.

θ is the probability that someone in the population would have tested positive for drugs

Example: Drugs on the job (cont.)

We have (independently) $y_1, \ldots, y_n | \theta \sim Bernoulli(\theta)$.

Example: Drugs on the job (cont.)

We have (independently) $y_1, \ldots, y_n|\theta \sim Bernoulli(\theta)$.

Assume that a random sample of n workers is to be taken while they are actually on the job.

Example: Drugs on the job (cont.)

We have (independently) $y_1, \ldots, y_n|\theta \sim Bernoulli(\theta)$.

Assume that a random sample of n workers is to be taken while they are actually on the job.

Each individual will be tested for drugs.

Example: Drugs on the job (cont.)

We have (independently) $y_1, \ldots, y_n|\theta \sim Bernoulli(\theta)$.

Assume that a random sample of n workers is to be taken while they are actually on the job.

Each individual will be tested for drugs.

Let y_i be a one if the ith individual tests positive and a zero otherwise.

Example: Drugs on the job (cont.)

We have (independently) $y_1, \ldots, y_n | \theta \sim Bernoulli(\theta)$.

Assume that a random sample of n workers is to be taken while they are actually on the job.

Each individual will be tested for drugs.

Let y_i be a one if the ith individual tests positive and a zero otherwise.

θ is the probability that someone in the population would have tested positive for drugs

Example: Drugs on the job (cont.)

We have (independently) $y_1, \ldots, y_n | \theta \sim Bernoulli(\theta)$.

Assume that a random sample of n workers is to be taken while they are actually on the job.

Each individual will be tested for drugs.

Let y_i be a one if the ith individual tests positive and a zero otherwise.

θ is the probability that someone in the population would have tested positive for drugs

Because the y_is are iid, the (sampling) density of $y = (y_1, \ldots, y_n)^T$ is

$$f(y|\theta) = \prod_{i=1}^{n} \theta^{y_i} (1-\theta)^{1-y_i}$$

Example: Drugs on the job (cont.)

Suppose that 10 workers were sampled and that two of them tested positive for drug use. The likelihood is then

$$L(\theta|y) \propto \theta^2(1-\theta)^8.$$

Example: Drugs on the job (cont.)

Suppose that 10 workers were sampled and that two of them tested positive for drug use. The likelihood is then

$$L(\theta|y) \propto \theta^2(1-\theta)^8.$$

Both $\theta = 0$ or 1 are impossible, since they exclude the possibility of seeing drug tests that are both positive and negative.

Example: Drugs on the job (cont.)

Suppose that 10 workers were sampled and that two of them tested positive for drug use. The likelihood is then

$$L(\theta|y) \propto \theta^2(1-\theta)^8.$$

Both $\theta = 0$ or 1 are impossible, since they exclude the possibility of seeing drug tests that are both positive and negative.

Values of θ above 0.5 are particularly unlikely to have generated these data.

Example: Drugs on the job (cont.)

Suppose that 10 workers were sampled and that two of them tested positive for drug use. The likelihood is then

$$L(\theta|y) \propto \theta^2(1-\theta)^8.$$

Both $\theta = 0$ or 1 are impossible, since they exclude the possibility of seeing drug tests that are both positive and negative.

Values of θ above 0.5 are particularly unlikely to have generated these data.

In fact, the most likely value is the sample proportion, $0.20 = 2/10$.

Example: Drugs on the job (cont.)

Suppose that 10 workers were sampled and that two of them tested positive for drug use. The likelihood is then

$$L(\theta|y) \propto \theta^2(1-\theta)^8.$$

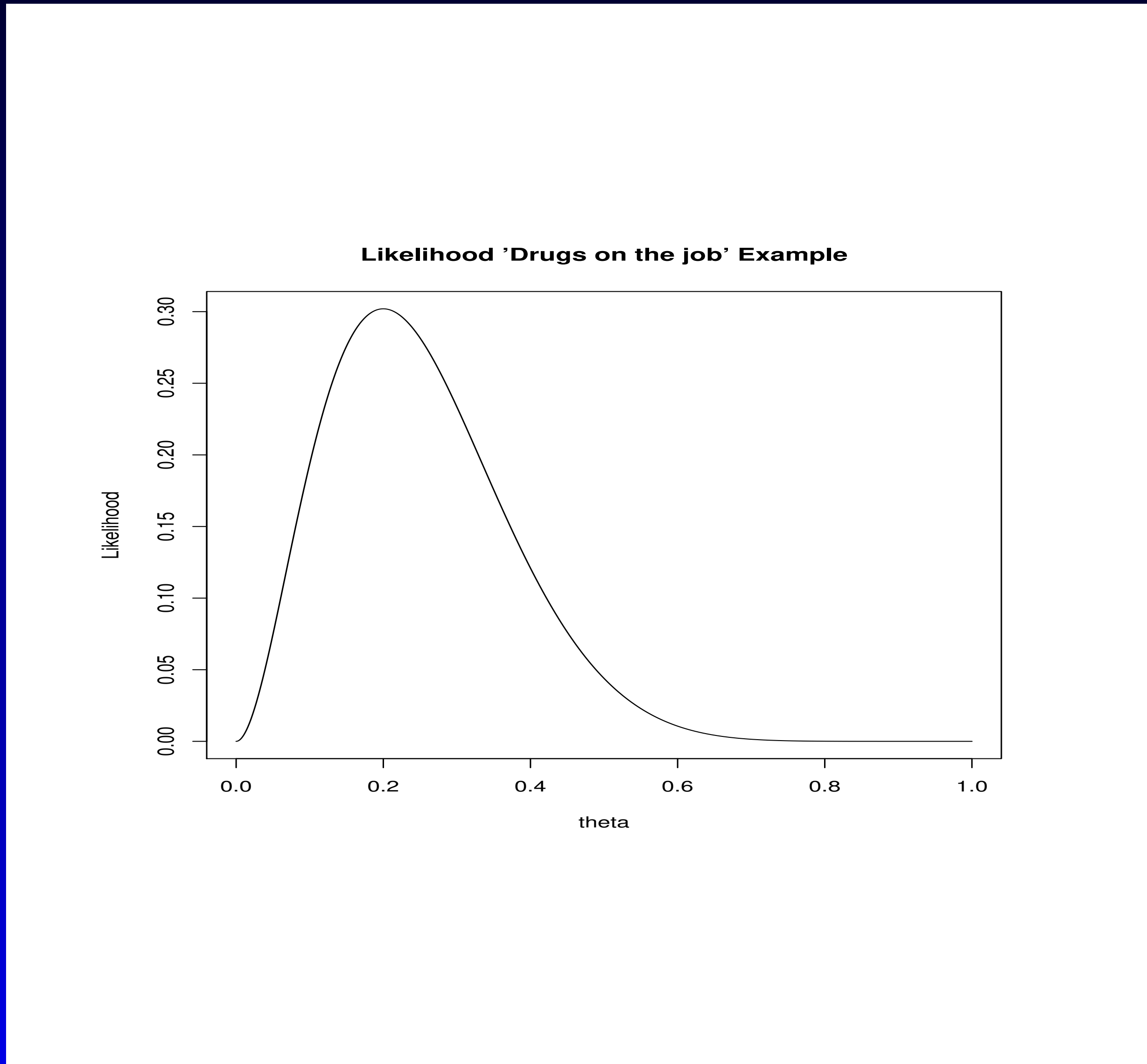
Both $\theta = 0$ or 1 are impossible, since they exclude the possibility of seeing drug tests that are both positive and negative.

Values of θ above 0.5 are particularly unlikely to have generated these data.

In fact, the most likely value is the sample proportion, $0.20 = 2/10$.

The value that maximizes the likelihood is called the maximum likelihood estimate (MLE).

Example: Drugs on the job (cont.)



Treating the unknown parameter θ

There are three main schools in statistics on how one should deal with the parameter θ:

Treating the unknown parameter θ

There are three main schools in statistics on how one should deal with the parameter θ:

(1) Likelihood

Treating the unknown parameter θ

There are three main schools in statistics on how one should deal with the parameter θ:

(1) Likelihood

(2) Frequentist

Treating the unknown parameter θ

There are three main schools in statistics on how one should deal with the parameter θ:

(1) Likelihood

(2) Frequentist

(3) Bayesian

Treating the unknown parameter θ

There are three main schools in statistics on how one should deal with the parameter θ:

(1) Likelihood

(2) Frequentist

(3) Bayesian

All the above share the idea of the likelihood function, $f(\mathbf{x}|\theta)$, that is available from the data, but they differ drastically on the way they handle the unknown parameter θ.

Likelihood School

All the information regarding the parameter should come exclusively from the likelihood function.

Likelihood School

All the information regarding the parameter should come exclusively from the likelihood function.

The philosophy of this school is based on the likelihood principle, where if two experiments produce analogous likelihoods then the inference regarding the unknown parameter should be identical.

Likelihood School

Likelihood Principle:
If the data from two experiments are $\mathbf{x}$, $\mathbf{y}$ and for the respective likelihoods $f(\mathbf{x}|\theta)$, $f(\mathbf{y}|\theta)$ it holds:

$$f(\mathbf{x}|\theta) \propto k(\mathbf{x},\mathbf{y}) f(\mathbf{y}|\theta)$$

then the inference regarding θ should be identical in both experiments.

Likelihood School

Likelihood Principle:
If the data from two experiments are $\mathbf{x}$, $\mathbf{y}$ and for the respective likelihoods $f(\mathbf{x}|\theta)$, $f(\mathbf{y}|\theta)$ it holds:

$$f(\mathbf{x}|\theta) \propto k(\mathbf{x},\mathbf{y}) f(\mathbf{y}|\theta)$$

then the inference regarding θ should be identical in both experiments.

Fiducial Inference:
Within this school R. A. Fisher developed the idea of transforming the likelihood to a distribution function (naively, think of $f(\mathbf{x}|\theta)/\int f(\mathbf{x}|\theta)d\theta = L(\theta)/\int L(\theta)d\theta$).

Frequentist School

Within this school the parameter θ is considered to be a **fixed** unknown constant.

Frequentist School

Within this school the parameter θ is considered to be a **fixed** unknown constant.

Inference regarding θ becomes available thanks to long term frequency properties. Precisely, we consider (infinite) repeated sampling, for **fixed** value of θ.

Frequentist School

Within this school the parameter θ is considered to be a **fixed** unknown constant.

Inference regarding θ becomes available thanks to long term frequency properties. Precisely, we consider (infinite) repeated sampling, for **fixed** value of θ.

While point estimation seems to be well aligned in this school, the assumption of a **fixed** parameter value can cause great difficulty in the interpretation of interval estimates (confidence intervals) and/or hypotheses testing.

Frequentist School

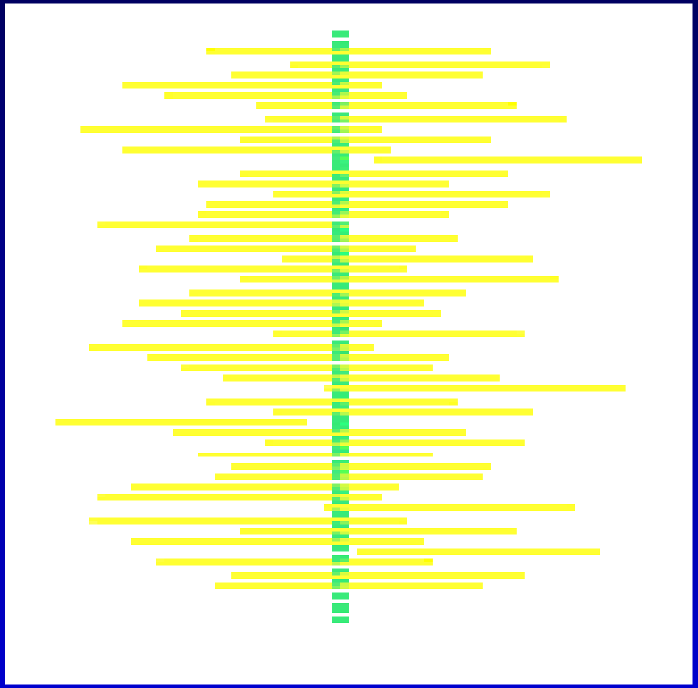
Typical example is the confidence interval, where the confidence level is quite often misinterpreted as the probability that the parameter belongs to the interval.

Frequentist School

Typical example is the confidence interval, where the confidence level is quite often misinterpreted as the probability that the parameter belongs to the interval.

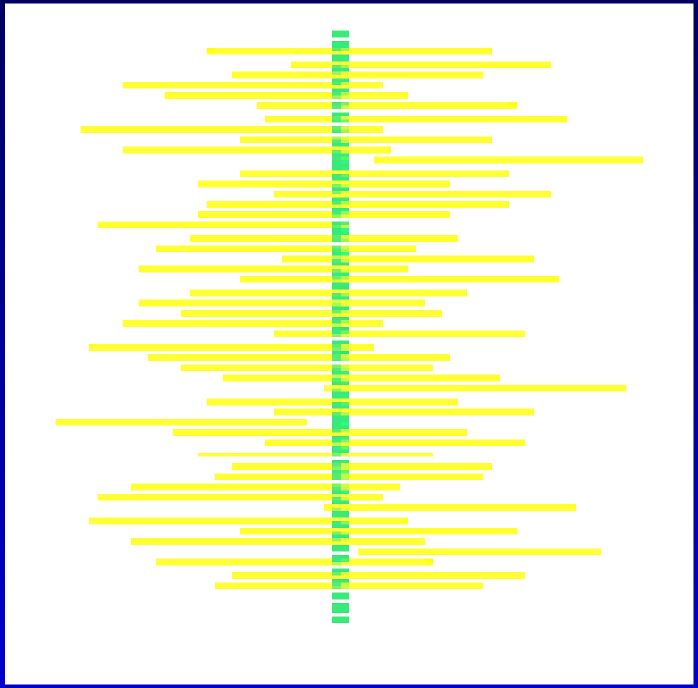
Frequentist School

Typical example is the confidence interval, where the confidence level is quite often misinterpreted as the probability that the parameter belongs to the interval.



Frequentist School

Typical example is the confidence interval, where the confidence level is quite often misinterpreted as the probability that the parameter belongs to the interval.



The parameter is constant, the interval is the random quantity.

Frequentist School

The frequentist's approach can violate the likelihood principle.

Frequentist School

The frequentist's approach can violate the likelihood principle.

Example (Lindley and Phillips (1976)):
Suppose we are interested in testing θ, the unknown probability of heads for possibly biased coin. Suppose, $H_0 : \theta = 1/2$ versus $H_1 : \theta > 1/2$. An experiment is conducted and 9 heads and 3 tails are observed. This information is not sufficient to fully specify the model $f(x|\theta)$. Specifically:

Frequentist School

Scenario 1: Number of flips, $n = 12$ is predetermined. Then number of heads x is $B(n, \theta)$, with likelihood:
$L_1(\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} = \binom{12}{9}\theta^9(1-\theta)^3$

Frequentist School

Scenario 1: Number of flips, $n = 12$ is predetermined. Then number of heads x is $B(n, \theta)$, with likelihood:
$L_1(\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} = \binom{12}{9}\theta^9(1-\theta)^3$

Scenario 2: Number of tails (successes) $r = 3$ is predetermined, i.e, the flipping is continued until 3 tails are observed. Then, x=number of heads (failures) until 3 tails appear is $NB(3, 1-\theta)$ with likelihood:
$L_2(\theta) = \binom{r+x-1}{r-1}(1-\theta)^r\theta^x = \binom{11}{2}\theta^9(1-\theta)^3$

Frequentist School

Scenario 1: Number of flips, $n = 12$ is predetermined. Then number of heads x is $B(n, \theta)$, with likelihood:
$L_1(\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} = \binom{12}{9}\theta^9(1-\theta)^3$

Scenario 2: Number of tails (successes) $r = 3$ is predetermined, i.e, the flipping is continued until 3 tails are observed. Then, x=number of heads (failures) until 3 tails appear is $NB(3, 1-\theta)$ with likelihood:
$L_2(\theta) = \binom{r+x-1}{r-1}(1-\theta)^r\theta^x = \binom{11}{2}\theta^9(1-\theta)^3$

Since $L_1(\theta) \propto L_2(\theta)$, based on the likelihood principle the two scenarios ought to give identical inference regarding θ.

Frequentist School

However, for a frequentist, the p-value of the test is:

Frequentist School

However, for a frequentist, the p-value of the test is:

Scenario 1:

$P(X \geq 9|H_0) = \sum_{x=9}^{12} \binom{12}{x} (0.5)^x (1-0.5)^{12-x} = 0.073$

Frequentist School

However, for a frequentist, the p-value of the test is:

Scenario 1:

$P(X \geq 9|H_0) = \sum_{x=9}^{12} \binom{12}{x} (0.5)^x (1-0.5)^{12-x} = 0.073$

Scenario 2:

$P(X \geq 9|H_0) = \sum_{x=9}^{\infty} \binom{3+x-1}{2} (1-0.5)^x (0.5)^3 = 0.0327$

Frequentist School

However, for a frequentist, the p-value of the test is:

Scenario 1:

$P(X \geq 9|H_0) = \sum_{x=9}^{12} \binom{12}{x}(0.5)^x(1-0.5)^{12-x} = 0.073$

Scenario 2:

$P(X \geq 9|H_0) = \sum_{x=9}^{\infty} \binom{3+x-1}{2}(1-0.5)^x(0.5)^3 = 0.0327$

and if we consider $\alpha = 0.05$ under the first scenario we fail to reject, while in the second we reject the H_0.

Bayesian School

In this school the parameter θ is considered to be a random variable. Given that θ is unknown, the most natural thing to do is to consider probability theory in quantifying what is unknown to us.

Bayesian School

In this school the parameter θ is considered to be a random variable. Given that θ is unknown, the most natural thing to do is to consider probability theory in quantifying what is unknown to us.

We will quantify our (subjective) opinion regarding θ (before looking the data) with a prior distribution: $p(\theta)$.

Bayesian School

In this school the parameter θ is considered to be a random variable. Given that θ is unknown, the most natural thing to do is to consider probability theory in quantifying what is unknown to us.

We will quantify our (subjective) opinion regarding θ (before looking the data) with a prior distribution: $p(\theta)$.

Then Bayes theorem will do the magic updating the prior distribution to posterior, under the light of the data.

Bayesian School

The Bayesian approach consists of the following steps:

Bayesian School

The Bayesian approach consists of the following steps:

(a) Define the likelihood: $f(\mathbf{x}|\theta)$

Bayesian School

The Bayesian approach consists of the following steps:

(a) Define the likelihood: $f(\mathbf{x}|\theta)$

(b) Define the prior distribution: $p(\theta)$

Bayesian School

The Bayesian approach consists of the following steps:

(a) Define the likelihood: $f(\mathbf{x}|\theta)$

(b) Define the prior distribution: $p(\theta)$

(c) Compute the posterior distribution: $p(\theta|\mathbf{x})$

Bayesian School

The Bayesian approach consists of the following steps:

(a) Define the likelihood: $f(\mathbf{x}|\theta)$

(b) Define the prior distribution: $p(\theta)$

(c) Compute the posterior distribution: $p(\theta|\mathbf{x})$

(d) Decision Making: Draw inference regarding θ – do predictions

Bayesian School

The Bayesian approach consists of the following steps:

(a) Define the likelihood: $f(\mathbf{x}|\theta)$

(b) Define the prior distribution: $p(\theta)$

(c) Compute the posterior distribution: $p(\theta|\mathbf{x})$

(d) Decision Making: Draw inference regarding θ – do predictions

We have already discussed **(a)** and we will proceed with **(c)**, **(b)** and conclude with **(d)**.

Computing the posterior

The Bayes theorem for events is given by:

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

Computing the posterior

The Bayes theorem for events is given by:

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

while for density functions it becomes:

$$p(\theta|\mathbf{x}) = \frac{f(\mathbf{x}, \theta)}{f(\mathbf{x})} = \frac{f(\mathbf{x}|\theta)p(\theta)}{\int f(\mathbf{x}|\theta)p(\theta)d\theta} \propto f(\mathbf{x}|\theta)p(\theta)$$

Computing the posterior

The Bayes theorem for events is given by:

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

while for density functions it becomes:

$$p(\theta|\mathbf{x}) = \frac{f(\mathbf{x}, \theta)}{f(\mathbf{x})} = \frac{f(\mathbf{x}|\theta)p(\theta)}{\int f(\mathbf{x}|\theta)p(\theta)d\theta} \propto f(\mathbf{x}|\theta)p(\theta)$$

The denominator $f(\mathbf{x})$ is the marginal distribution of the observed data, i.e. it is a single number (known as normalizing constant) that is responsible for making $p(\theta|\mathbf{x})$ to become a density.

Computing the (multivariate) posterior

Moving from univariate to multivariate we obtain:

Computing the (multivariate) posterior

Moving from univariate to multivariate we obtain:

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{f(\mathbf{x}, \boldsymbol{\theta})}{f(\mathbf{x})} = \frac{f(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int \cdots \int f(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}$$

Computing the (multivariate) posterior

Moving from univariate to multivariate we obtain:

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{f(\mathbf{x}, \boldsymbol{\theta})}{f(\mathbf{x})} = \frac{f(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int \cdots \int f(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}$$

The normalizing constant was the main reason for the underdevelopment of the Bayesian approach and its limited use in science for decades (if not centuries). However, the MCMC revolution, started in mid 90's, overcame this technical issue (providing a sample from the posterior) making widely available the Bayesian school of statistical analysis in all fields of science.

Bayesian Inference

it is often convenient to summarize the posterior information into objects like the posterior median, say $m(d)$, where $m(d)$ satisfies

$$\frac{1}{2} = \int_{-\infty}^{m(d)} p(\theta|d)d\theta$$

Bayesian Inference

it is often convenient to summarize the posterior information into objects like the posterior median, say $m(d)$, where $m(d)$ satisfies

$$\frac{1}{2} = \int_{-\infty}^{m(d)} p(\theta|d)d\theta$$

or the posterior mean

$$E(\theta|d) = \int \theta p(\theta|d)d\theta$$

Bayesian Inference (cont.)

Other quantities of potential interest are the posterior variance

$$V(\theta|d) = \int [\theta - E(\theta|d)]^2 \, p(\theta|d) d\theta$$

Bayesian Inference (cont.)

Other quantities of potential interest are the posterior variance

$$V(\theta|d) = \int [\theta - E(\theta|d)]^2 \, p(\theta|d) d\theta$$

the posterior standard deviation $\sqrt{V(\theta|d)}$

Bayesian Inference (cont.)

Other quantities of potential interest are the posterior variance

$$V(\theta|d) = \int [\theta - E(\theta|d)]^2 \, p(\theta|d) d\theta$$

the posterior standard deviation $\sqrt{V(\theta|d)}$

and, say, the 95% probability intervals $[a(d), b(d)]$ where $a(d)$ and $b(d)$ satisfy

$$0.95 = \int_{a(d)}^{b(d)} p(\theta|d) d\theta$$

Prior distribution

This is the key element of the Bayesian approach.

Prior distribution

This is the key element of the Bayesian approach.

Subjective Bayesian approach: The parameter of interest takes eventually a single number, which is used in the likelihood to provide the data. Since we do not know this value, we use a random mechanism (the prior $p(\theta)$) to describe the uncertainty about this parameter value. Thus, we simply use probability theory to model the uncertainty.

Prior distribution

This is the key element of the Bayesian approach.

Subjective Bayesian approach: The parameter of interest takes eventually a single number, which is used in the likelihood to provide the data. Since we do not know this value, we use a random mechanism (the prior $p(\theta)$) to describe the uncertainty about this parameter value. Thus, we simply use probability theory to model the uncertainty.

The prior should reflect our personal (subjective) opinion regarding the parameter, before we look at the data. The only think we need to be careful about, is to be coherent, which will happen if we will obey the probability laws (see de Finetti, DeGroot, Hartigan etc.)

Prior distribution

Main issues regarding prior distributions:

Prior distribution

Main issues regarding prior distributions:

- Posterior lives in the range defined by the prior.

Prior distribution

Main issues regarding prior distributions:

- Posterior lives in the range defined by the prior.
- The more data we get the less the effect of the prior in determining the posterior distribution (unless extreme choices, like point mass priors are made.)

Prior distribution

Main issues regarding prior distributions:

- Posterior lives in the range defined by the prior.
- The more data we get the less the effect of the prior in determining the posterior distribution (unless extreme choices, like point mass priors are made.)
- Different priors applied on the same data will lead to different posteriors.

Prior distribution

Main issues regarding prior distributions:

- Posterior lives in the range defined by the prior.
- The more data we get the less the effect of the prior in determining the posterior distribution (unless extreme choices, like point mass priors are made.)
- Different priors applied on the same data will lead to different posteriors.

The last bullet, raised (and keeps raising) the major criticism from non-Bayesians (see for example Efron (1986), “Why isn’t everyone a Bayesian”). However, Bayesians love the opportunity to be subjective. Lets see an example:

Prior distribution - Example 1

We have two different binomial experiments.

Prior distribution - Example 1

We have two different binomial experiments.

Setup 1: We ask from a sommelier (wine expert) to taste 10 glasses of wine and decide whether each glass is Merlot or Cabernet Sauvignon.

Prior distribution - Example 1

We have two different binomial experiments.

Setup 1: We ask from a sommelier (wine expert) to taste 10 glasses of wine and decide whether each glass is Merlot or Cabernet Sauvignon.

Setup 2: We ask from a drank man to guess the sequence of H and T in 10 tosses of a fair coin.

Prior distribution - Example 1

We have two different binomial experiments.

Setup 1: We ask from a sommelier (wine expert) to taste 10 glasses of wine and decide whether each glass is Merlot or Cabernet Sauvignon.

Setup 2: We ask from a drank man to guess the sequence of H and T in 10 tosses of a fair coin.

In both cases we have a $B(10, \theta)$ with unknown the probability of success (θ).

Prior distribution - Example 1

We have two different binomial experiments.

Setup 1: We ask from a sommelier (wine expert) to taste 10 glasses of wine and decide whether each glass is Merlot or Cabernet Sauvignon.

Setup 2: We ask from a drank man to guess the sequence of H and T in 10 tosses of a fair coin.

In both cases we have a $B(10, \theta)$ with unknown the probability of success (θ).

The data become available and we have 10 successes in both setups, i.e. based on the frequentist MLE $\hat{\theta} = 1$ in both cases.

Prior distribution - Example 1

But is this really what we believe?

Prior distribution - Example 1

But is this really what we believe?

Before looking in the data, if you were to bet money to the higher probability of success, would you put your money to setup 1 or 2? or did you think that the probabilities were equal?

Prior distribution - Example 1

But is this really what we believe?

Before looking in the data, if you were to bet money to the higher probability of success, would you put your money to setup 1 or 2? or did you think that the probabilities were equal?

For the sommelier we expect to have the probability of success close to 1, while for the drunk man we would expect his success rate to be close to 1/2.

Prior distribution - Example 1

But is this really what we believe?

Before looking in the data, if you were to bet money to the higher probability of success, would you put your money to setup 1 or 2? or did you think that the probabilities were equal?

For the sommelier we expect to have the probability of success close to 1, while for the drunk man we would expect his success rate to be close to 1/2.

Adopting the appropriate prior distribution for each setup would lead to different posteriors, in contrast to the frequentist based methods that yield identical results.

Prior distribution – Example 2

At the end of the semester you will have a final exam on this course.

Prior distribution – Example 2

At the end of the semester you will have a final exam on this course.

Please write down, what is the probability that you will pass the exam.

Prior distribution – Example 2

At the end of the semester you will have a final exam on this course.

Please write down, what is the probability that you will pass the exam.

Lets look in the future now: you will either pass or fail the exam. Thus the frequentist MLE point estimate of the probability of success will be either 1 (if you pass) or 0 (if you fail).

Prior distribution – Example 2

At the end of the semester you will have a final exam on this course.

Please write down, what is the probability that you will pass the exam.

Lets look in the future now: you will either pass or fail the exam. Thus the frequentist MLE point estimate of the probability of success will be either 1 (if you pass) or 0 (if you fail).

If you wrote down any number in (0,1) then you are a Bayesian! (consciously or unconsciously).

Prior distribution – Elicitation

The prior distribution should reflect our personal beliefs for the unknown parameter, before the data becomes available.

Prior distribution – Elicitation

The prior distribution should reflect our personal beliefs for the unknown parameter, before the data becomes available.

If we do not know anything about θ, expert's opinion or historic data can be used, but *not* the current data.

Prior distribution − Elicitation

The prior distribution should reflect our personal beliefs for the unknown parameter, before the data becomes available.

If we do not know anything about θ, expert's opinion or historic data can be used, but *not* the current data.

The elicitation of a prior consists of the following two steps:

Prior distribution – Elicitation

The prior distribution should reflect our personal beliefs for the unknown parameter, before the data becomes available.

If we do not know anything about θ, expert's opinion or historic data can be used, but *not* the current data.

The elicitation of a prior consists of the following two steps:

- Recognize the function form which best expresses our uncertainty regarding θ (i.e. modes, symmetry etc.)

Prior distribution – Elicitation

The prior distribution should reflect our personal beliefs for the unknown parameter, before the data becomes available.

If we do not know anything about θ, expert's opinion or historic data can be used, but *not* the current data.

The elicitation of a prior consists of the following two steps:

- Recognize the function form which best expresses our uncertainty regarding θ (i.e. modes, symmetry etc.)
- Decide on the parameters of the prior distribution, that most closely match our beliefs.

Prior distribution – Subjective vs Objective

There exist setups where we have good knowledge about θ (like an industrial statistician that supervises a production line). In such cases the subjective Bayesian approach is highly preferable since it offers a well defined framework to incorporate this (subjective) prior opinion.

Prior distribution – Subjective vs Objective

There exist setups where we have good knowledge about θ (like an industrial statistician that supervises a production line). In such cases the subjective Bayesian approach is highly preferable since it offers a well defined framework to incorporate this (subjective) prior opinion.

But what about cases where no information whatsoever about θ is available?

Prior distribution – Subjective vs Objective

There exist setups where we have good knowledge about θ (like an industrial statistician that supervises a production line). In such cases the subjective Bayesian approach is highly preferable since it offers a well defined framework to incorporate this (subjective) prior opinion.

But what about cases where no information whatsoever about θ is available?

Then one could follow an objective Bayesian approach.

Prior distribution – Conjugate analysis

A family of priors is called conjugate when the posterior is a member of the same family as the prior.

Prior distribution – Conjugate analysis

A family of priors is called conjugate when the posterior is a member of the same family as the prior.

Example:
$f(x|\theta) \sim B(n, \theta)$ and for the parameter θ we assume:
$p(\theta) \sim Beta(\alpha, \beta)$

Prior distribution − Conjugate analysis

A family of priors is called conjugate when the posterior is a member of the same family as the prior.

Example:
$f(x|\theta) \sim B(n, \theta)$ and for the parameter θ we assume:
$p(\theta) \sim Beta(\alpha, \beta)$

Then:

$$p(\theta|x) \propto f(x|\theta)p(\theta) \propto \left[\theta^x(1-\theta)^{n-x}\right]\left[\theta^{\alpha-1}(1-\theta)^{\beta-1}\right]$$
$$= \theta^{\alpha+x-1}(1-\theta)^{n+\beta-x-1}$$

Prior distribution – Conjugate analysis

A family of priors is called conjugate when the posterior is a member of the same family as the prior.

Example:

$f(x|\theta) \sim B(n, \theta)$ and for the parameter θ we assume:

$p(\theta) \sim Beta(\alpha, \beta)$

Then:

$$
\begin{aligned}
p(\theta|x) \propto f(x|\theta)p(\theta) &\propto \left[\theta^{x}(1-\theta)^{n-x}\right]\left[\theta^{\alpha-1}(1-\theta)^{\beta-1}\right] \\
&= \theta^{\alpha+x-1}(1-\theta)^{n+\beta-x-1}
\end{aligned}
$$

Thus, $p(\theta|x) \sim Beta(\alpha + x, \beta + n - x)$

Prior distribution − Conjugate analysis

With a conjugate prior there is no need for the evaluation of the normalizing constant (i.e. no need to calculate the integral in the denominator).

Prior distribution − Conjugate analysis

With a conjugate prior there is no need for the evaluation of the normalizing constant (i.e. no need to calculate the integral in the denominator).

To guess for a conjugate prior it is helpful to look at the likelihood as a function of θ.

Prior distribution − Conjugate analysis

With a conjugate prior there is no need for the evaluation of the normalizing constant (i.e. no need to calculate the integral in the denominator).

To guess for a conjugate prior it is helpful to look at the likelihood as a function of θ.

Existence theorem:
When the likelihood is a member of the exponential family a conjugate prior exists.

Prior distribution – Non-informative (Objective)

A prior that does not favor one value of θ over another.

Prior distribution – Non-informative (Objective)

A prior that does not favor one value of θ over another.

For compact parameter spaces the above is achieved by a "flat" prior, i.e. uniform over the parameter space.

Prior distribution – Non-informative (Objective)

A prior that does not favor one value of θ over another.

For compact parameter spaces the above is achieved by a "flat" prior, i.e. uniform over the parameter space.

For non-compact parameter spaces (like $\boldsymbol{\theta} \in (-\infty, +\infty)$) then the flat prior ($p(\theta) \propto c$) is not a distribution. However, it is still legitimate to be used iff: $\int f(\mathbf{x}|\theta) d\theta = K < \infty$. These priors are called "improper" priors and they lead to proper posteriors since:

$$p(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)p(\theta)}{\int f(\mathbf{x}|\theta)p(\theta)d\theta} = \frac{f(\mathbf{x}|\theta)c}{\int f(\mathbf{x}|\theta)c d\theta} = \frac{f(\mathbf{x}|\theta)}{\int f(\mathbf{x}|\theta)d\theta}$$

(remember the Fiducial inference).

Prior distribution – Non-informative (Objective)

Example:

$f(x|\theta) \sim B(n, \theta)$ and for the parameter θ we assume:

$p(\theta) \sim U(0, 1)$

Prior distribution − Non-informative (Objective)

Example:

$f(x|\theta) \sim B(n,\theta)$ and for the parameter θ we assume:

$p(\theta) \sim U(0,1)$

Then:

$$\begin{aligned} p(\theta|x) &\propto f(x|\theta)p(\theta) \propto \left[\theta^x(1-\theta)^{n-x}\right]1 \\ &= \theta^{(x+1)-1}(1-\theta)^{(2-x)-1} \end{aligned}$$

Prior distribution – Non-informative (Objective)

Example:

$f(x|\theta) \sim B(n,\theta)$ and for the parameter θ we assume:

$p(\theta) \sim U(0,1)$

Then:

$$
\begin{aligned}
p(\theta|x) &\propto f(x|\theta)p(\theta) \propto \left[\theta^{x}(1-\theta)^{n-x}\right]1 \\
&= \theta^{(x+1)-1}(1-\theta)^{(2-x)-1}
\end{aligned}
$$

Thus, $\quad p(\theta|x) \sim Beta(x+1, 2-x)$

Prior distribution – Non-informative (Objective)

Example:

$f(x|\theta) \sim B(n,\theta)$ and for the parameter θ we assume:

$p(\theta) \sim U(0,1)$

Then:

$$\begin{aligned} p(\theta|x) \propto f(x|\theta)p(\theta) &\propto \left[\theta^x(1-\theta)^{n-x}\right]1 \\ &= \theta^{(x+1)-1}(1-\theta)^{(2-x)-1} \end{aligned}$$

Thus, $\quad p(\theta|x) \sim Beta(x+1, 2-x)$

Remember that $U(0,1) \equiv Beta(1,1)$ which we showed earlier to be conjugate for the Binomial likelihood.

Prior distribution – Non-informative (Objective)

Example:

$f(x|\theta) \sim B(n,\theta)$ and for the parameter θ we assume:

$p(\theta) \sim U(0,1)$

Then:

$$
\begin{aligned}
p(\theta|x) \propto f(x|\theta)p(\theta) &\propto \left[\theta^x (1-\theta)^{n-x}\right] 1 \\
&= \theta^{(x+1)-1}(1-\theta)^{(2-x)-1}
\end{aligned}
$$

Thus, $\quad p(\theta|x) \sim Beta(x+1, 2-x)$

Remember that $U(0,1) \equiv Beta(1,1)$ which we showed earlier to be conjugate for the Binomial likelihood.

In general with flat priors we do not get posteriors in closed forms and use of MCMC techniques is inevitable.

Prior distribution – Jeffreys prior

It is the prior, which is invariant under 1-1 transformations.
It is given as:

$$p_0(\theta) \propto [I(\theta)]^{1/2}$$

where $I(\theta)$ is the expected Fisher information i.e.:

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

Jeffreys prior is not necessarily a flat prior.

Prior distribution – Jeffreys prior

It is the prior, which is invariant under 1-1 transformations. It is given as:

$$p_0(\theta) \propto [I(\theta)]^{1/2}$$

where $I(\theta)$ is the expected Fisher information i.e.:

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

Jeffreys prior is not necessarily a flat prior.

As we mentioned earlier we should not take into account the data in determining the prior. Jeffreys prior is consistent with this principle, since it makes use of the form of the likelihood and *not* of the actual data.

Prior distribution – Jeffreys prior

Example: Jeffreys prior when $f(x|\theta) \sim B(n, \theta)$.

Prior distribution − Jeffreys prior

Example: Jeffreys prior when $f(x|\theta) \sim B(n,\theta)$.

$$
\begin{aligned}
logL(\theta) &= log\binom{n}{x} + xlog\theta + (n-x)log(1-\theta) \\
\frac{\partial logL(\theta)}{\partial\theta} &= \frac{x}{\theta} - \frac{n-x}{1-\theta} \\
\frac{\partial^2 logL(\theta)}{\partial\theta^2} &= -\frac{x}{\theta^2} - \frac{n-x}{(1-\theta)^2} \\
E_{X|\theta}\left[\frac{\partial^2 logL(\theta)}{\partial\theta^2}\right] &= -\frac{n\theta}{\theta^2} - \frac{n-n\theta}{(1-\theta)^2} = -\frac{n}{\theta(1-\theta)} \\
p_0(\theta) &\propto \theta^{-1/2}(1-\theta)^{-1/2} \equiv Beta(1/2,1/2)
\end{aligned}
$$

Prior distribution – Vague (low information)

In some cases we try to make the support of the prior distribution to be vague by “flatten” it out. This can be done by “exploding” the variance, which will make the prior almost flat (from a practical perspective) for the range of values we are concerned with.

Prior distribution – Mixture

When we need to model different a-priori opinions, we might end up with a multimodal prior distribution. In such cases we can use a mixture of prior distributions:

Prior distribution – Mixture

When we need to model different a-priori opinions, we might end up with a multimodal prior distribution. In such cases we can use a mixture of prior distributions:

$$p(\theta) = \sum_{i=1}^{k} p_i(\theta)$$

Prior distribution – Mixture

When we need to model different a-priori opinions, we might end up with a multimodal prior distribution. In such cases we can use a mixture of prior distributions:

$$p(\theta) = \sum_{i=1}^{k} p_i(\theta)$$

Then the posterior distribution will be a mixture with the same number of components as the prior.

Hyperpriors − Hierarchical Modeling

The prior distribution will have its own parameter values: $\boldsymbol{\eta}$, i.e. $p(\theta|\boldsymbol{\eta})$. Thus far we assumed that $\boldsymbol{\eta}$ were known exactly.

Hyperpriors – Hierarchical Modeling

The prior distribution will have its own parameter values: $\boldsymbol{\eta}$, i.e. $p(\theta|\boldsymbol{\eta})$. Thus far we assumed that $\boldsymbol{\eta}$ were known exactly.

If $\boldsymbol{\eta}$ are unknown, then the natural thing to do, within the Bayesian framework, is to assign a prior on them $h(\boldsymbol{\eta})$, i.e. a second level prior or hyperprior. Then:

$$
\begin{aligned}
p(\theta|\mathbf{x}) &= \frac{f(\mathbf{x},\theta)}{\int f(\mathbf{x},\theta)d\theta} = \frac{\int f(\mathbf{x},\theta,\boldsymbol{\eta})d\boldsymbol{\eta}}{\int\int f(\mathbf{x},\theta,\boldsymbol{\eta})d\theta d\boldsymbol{\eta}} \\
&= \frac{\int f(\mathbf{x}|\theta)p(\theta|\boldsymbol{\eta})h(\boldsymbol{\eta})d\boldsymbol{\eta}}{\int\int f(\mathbf{x}|\theta)p(\theta|\boldsymbol{\eta})h(\boldsymbol{\eta})d\boldsymbol{\eta}d\theta}
\end{aligned}
$$

Hyperpriors – Hierarchical Modeling

The prior distribution will have its own parameter values: $\boldsymbol{\eta}$, i.e. $p(\theta|\boldsymbol{\eta})$. Thus far we assumed that $\boldsymbol{\eta}$ were known exactly.

If $\boldsymbol{\eta}$ are unknown, then the natural thing to do, within the Bayesian framework, is to assign a prior on them $h(\boldsymbol{\eta})$, i.e. a second level prior or hyperprior. Then:

$$
\begin{aligned}
p(\theta|\mathbf{x}) &= \frac{f(\mathbf{x},\theta)}{\int f(\mathbf{x},\theta)d\theta} = \frac{\int f(\mathbf{x},\theta,\boldsymbol{\eta})d\boldsymbol{\eta}}{\int\int f(\mathbf{x},\theta,\boldsymbol{\eta})d\theta d\boldsymbol{\eta}} \\
&= \frac{\int f(\mathbf{x}|\theta)p(\theta|\boldsymbol{\eta})h(\boldsymbol{\eta})d\boldsymbol{\eta}}{\int\int f(\mathbf{x}|\theta)p(\theta|\boldsymbol{\eta})h(\boldsymbol{\eta})d\boldsymbol{\eta}d\theta}
\end{aligned}
$$

This build up hierarchy can continue to a 3^{rd}, 4^{th}, etc level, leading to hierarchical models.

Sequential updating

In the Bayesian analysis we can work sequentially (i.e. update from prior to posterior as each data becomes available) or not (i.e. first collect all the data and the obtain the posterior).

Sequential updating

In the Bayesian analysis we can work sequentially (i.e. update from prior to posterior as each data becomes available) or not (i.e. first collect all the data and the obtain the posterior).

The posterior distributions obtained working either sequentially or not will be identical as long as the data are conditionally independent, i.e.:

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

Sequential updating

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

Sequential updating

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

In some settings the sequential analysis is very helpful since it can provide inference for θ in an online fashion and not once the data collection is completed.

Sensitivity Analysis

At the end of our analysis it is wise to check how robust (sensitive) our results are to the particular choice of the prior we made.

Sensitivity Analysis

At the end of our analysis it is wise to check how robust (sensitive) our results are to the particular choice of the prior we made.

So it is proposed to repeat the analysis with a vague, noninformative, etc, priors and observe the effect these changes have to the obtained results.

Example: Drugs on the job (cont.)

Suppose that (i) a researcher has estimated that 10% of transportation workers use drugs on the job, and (ii) the researcher is 95% sure that the actual proportion was no larger than 25%. Therefore our best guess is $\theta \approx 0.1$ and $P(\theta < 0.25) = 0.95$.

Example: Drugs on the job (cont.)

Suppose that (i) a researcher has estimated that 10% of transportation workers use drugs on the job, and (ii) the researcher is 95% sure that the actual proportion was no larger than 25%. Therefore our best guess is $\theta \approx 0.1$ and $P(\theta < 0.25) = 0.95$.

We assume the prior is a member of some parametric family of distributions and to use the information to identify an appropriate member of the family.

Example: Drugs on the job (cont.)

Suppose that (i) a researcher has estimated that 10% of transportation workers use drugs on the job, and (ii) the researcher is 95% sure that the actual proportion was no larger than 25%. Therefore our best guess is $\theta \approx 0.1$ and $P(\theta < 0.25) = 0.95$.

We assume the prior is a member of some parametric family of distributions and to use the information to identify an appropriate member of the family.

For example, suppose we consider the family of $Beta(a, b)$ distributions for θ

Example: Drugs on the job (cont.)

Suppose that (i) a researcher has estimated that 10% of transportation workers use drugs on the job, and (ii) the researcher is 95% sure that the actual proportion was no larger than 25%. Therefore our best guess is $\theta \approx 0.1$ and $P(\theta < 0.25) = 0.95$.

We assume the prior is a member of some parametric family of distributions and to use the information to identify an appropriate member of the family.

For example, suppose we consider the family of $Beta(a, b)$ distributions for θ

We identify the estimate of 10% with the mode

$$m = \frac{a-1}{a+b-2}$$

Example: Drugs on the job (cont.)

So we set

$$0.10 = \frac{a-1}{a+b-2} \Rightarrow a = \frac{1+0.1b}{0.9}$$

Example: Drugs on the job (cont.)

So we set

$$0.10 = \frac{a-1}{a+b-2} \Rightarrow a = \frac{1+0.1b}{0.9}$$

Using Chun-lung Su's Betabuster, we can search through possible b values until we find a distribution Beta(a, b) for which $P(\theta < 0.25) = 0.95$

Example: Drugs on the job (cont.)

So we set

$$0.10 = \frac{a-1}{a+b-2} \Rightarrow a = \frac{1+0.1b}{0.9}$$

Using Chun-lung Su's Betabuster, we can search through possible b values until we find a distribution Beta(a, b) for which $P(\theta < 0.25) = 0.95$

The $Beta(a = 3.4, b = 23)$ distribution actually satisfies the constraints given above for the transportation industry problem

Example: Drugs on the job (cont.)

Suppose $n = 100$ workers were tested and that 15 tested positive for drug use. Let y be the number who tested positive. Therefore we have $y|\theta \sim Bin(n, \theta)$.

Example: Drugs on the job (cont.)

Suppose $n = 100$ workers were tested and that 15 tested positive for drug use. Let y be the number who tested positive. Therefore we have $y|\theta \sim Bin(n, \theta)$.

The posterior is $\theta|y \sim Beta(y + a = 18.4, n - y + b = 108)$

Example: Drugs on the job (cont.)

Suppose $n = 100$ workers were tested and that 15 tested positive for drug use. Let y be the number who tested positive. Therefore we have $y|\theta \sim Bin(n, \theta)$.

The posterior is $\theta|y \sim Beta(y + a = 18.4, n - y + b = 108)$

The prior mode is

$$0.098 \approx \frac{a-1}{a+b-2}$$

Example: Drugs on the job (cont.)

Suppose $n = 100$ workers were tested and that 15 tested positive for drug use. Let y be the number who tested positive. Therefore we have $y|\theta \sim Bin(n, \theta)$.

The posterior is $\theta|y \sim Beta(y + a = 18.4, n - y + b = 108)$

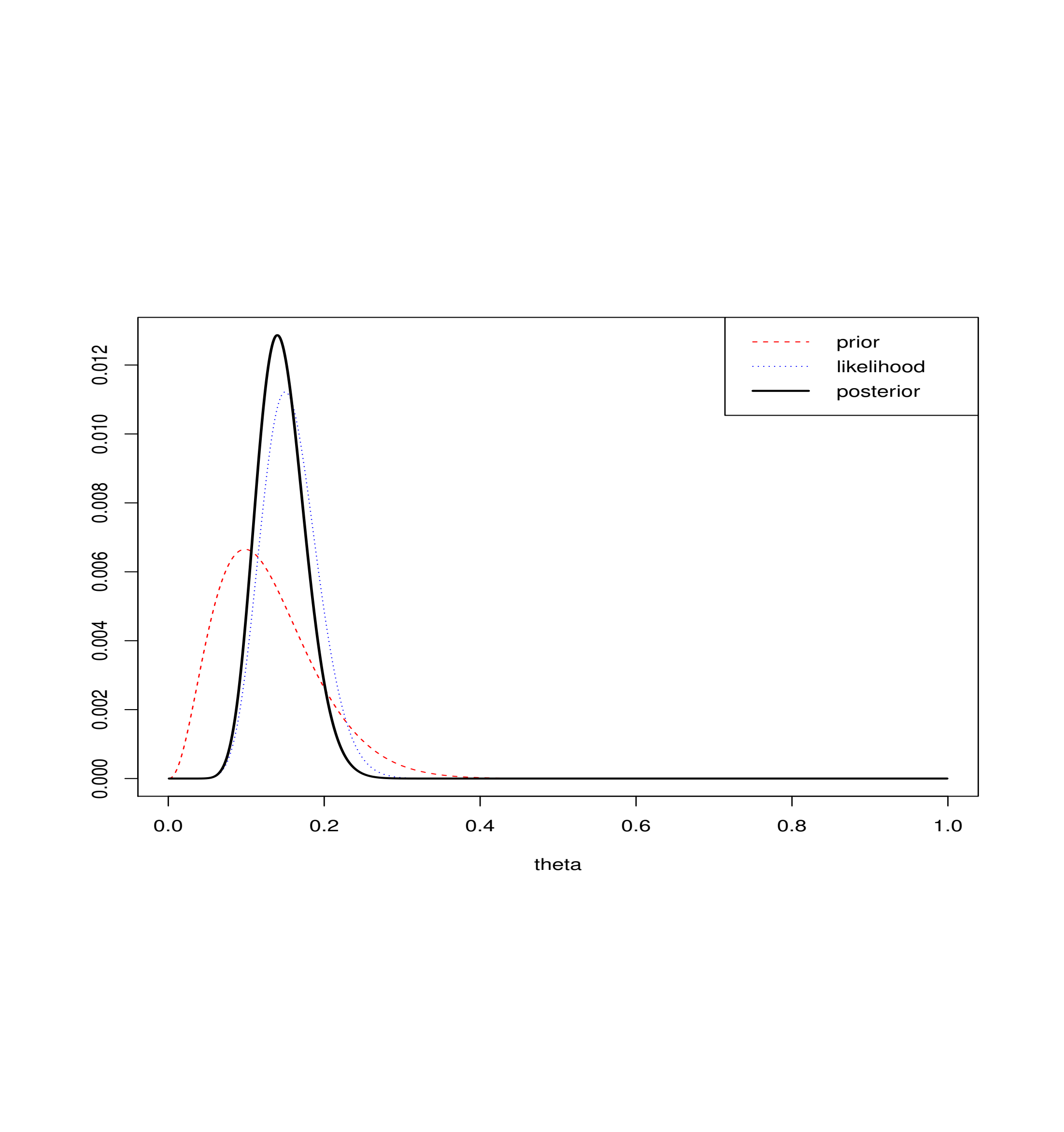
The prior mode is

$$0.098 \approx \frac{a-1}{a+b-2}$$

The posterior mode is

$$0.14 \approx \frac{y+a-1}{n+a+b-2}$$

Example: Drugs on the job (cont.)



Example: Drugs on the job (cont.)

We also consider the situation with $n = 500$ and $y = 75$

Example: Drugs on the job (cont.)

We also consider the situation with $n = 500$ and $y = 75$

The posterior is now
$\theta|y \sim Beta(y + a = 78.4, n - y + b = 448)$

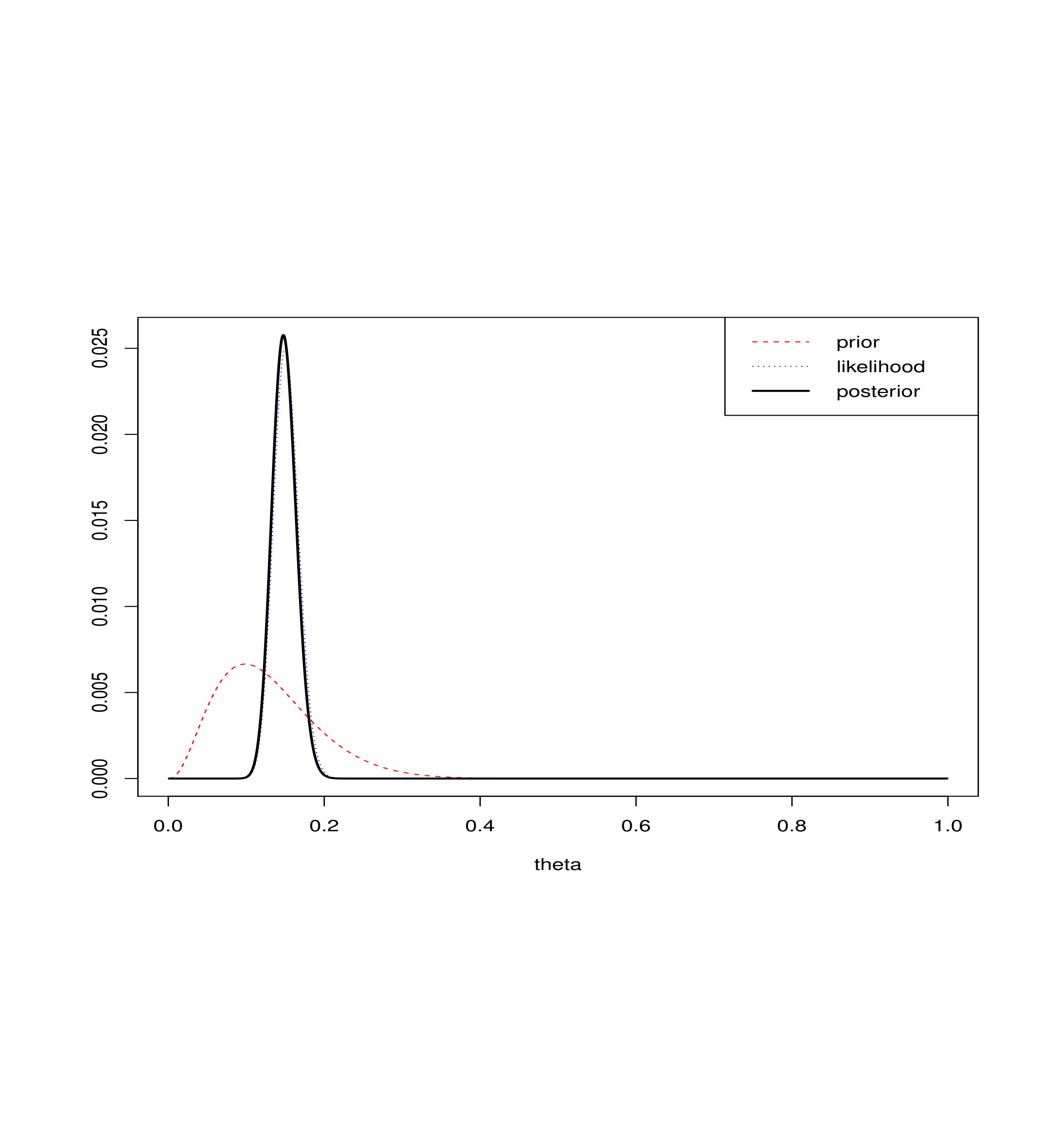
Example: Drugs on the job (cont.)

We also consider the situation with $n = 500$ and $y = 75$

The posterior is now
$\theta|y \sim Beta(y + a = 78.4, n - y + b = 448)$

Notice how the posterior is getting more concentrated

Example: Drugs on the job (cont.)



Example: Drugs on the job (cont.)

These data could have arisen as the original sample of size 100, which resulted in then $Beta(18.4, 108)$ posterior.

Example: Drugs on the job (cont.)

These data could have arisen as the original sample of size 100, which resulted in then $Beta(18.4, 108)$ posterior.

Then, if an additional 400 observations were taken with 60 positive outcomes, we could have used the $Beta(18.4, 108)$ as our prior, which would have been combined with the current data to obtain the $Beta(78.4, 448)$ posterior.

Example: Drugs on the job (cont.)

These data could have arisen as the original sample of size 100, which resulted in then $Beta(18.4, 108)$ posterior.

Then, if an additional 400 observations were taken with 60 positive outcomes, we could have used the $Beta(18.4, 108)$ as our prior, which would have been combined with the current data to obtain the $Beta(78.4, 448)$ posterior.

Bayesian methods thus handle sequential sampling in a straightforward way.

Example 1 (Carlin and Louis)

We give to 16 customers of a fast food chain to taste two patties (one is expensive and the other is cheap) in a random order. The experiment is double blind, i.e. neither the customer nor the chef/server knows which is the expensive patty. We had 13 out of the 16 customers to be able to tell the difference (i.e. they preferred the more expensive patty).

Example 1 (Carlin and Louis)

We give to 16 customers of a fast food chain to taste two patties (one is expensive and the other is cheap) in a random order. The experiment is double blind, i.e. neither the customer nor the chef/server knows which is the expensive patty. We had 13 out of the 16 customers to be able to tell the difference (i.e. they preferred the more expensive patty).

Assuming that the probability (θ) of being able to discriminate the expensive patty is constant, then we had X=13, where:

$$X|\theta \sim B(16, \theta)$$

Example 1 (Carlin and Louis)

Our goal is to determine whether $\theta = 1/2$ or not, i.e. whether the customers guess or they can actually tell the difference.

Example 1 (Carlin and Louis)

Our goal is to determine whether $\theta = 1/2$ or not, i.e. whether the customers guess or they can actually tell the difference.

We will make use of three different prior distributions:

Example 1 (Carlin and Louis)

Our goal is to determine whether $\theta = 1/2$ or not, i.e. whether the customers guess or they can actually tell the difference.

We will make use of three different prior distributions:

- $\theta \sim Beta(1/2, 1/2)$, which is the Jeffreys prior

Example 1 (Carlin and Louis)

Our goal is to determine whether $\theta = 1/2$ or not, i.e. whether the customers guess or they can actually tell the difference.

We will make use of three different prior distributions:

- $\theta \sim Beta(1/2, 1/2)$, which is the Jeffreys prior
- $\theta \sim Beta(1, 1) \equiv U(0, 1)$, which is the noninformative prior

Example 1 (Carlin and Louis)

Our goal is to determine whether $\theta = 1/2$ or not, i.e. whether the customers guess or they can actually tell the difference.

We will make use of three different prior distributions:

- $\theta \sim Beta(1/2, 1/2)$, which is the Jeffreys prior
- $\theta \sim Beta(1, 1) \equiv U(0, 1)$, which is the noninformative prior
- $\theta \sim Beta(2, 2)$, which is a skeptical prior, putting the prior mass around 1/2

Example 1 (Carlin and Louis)

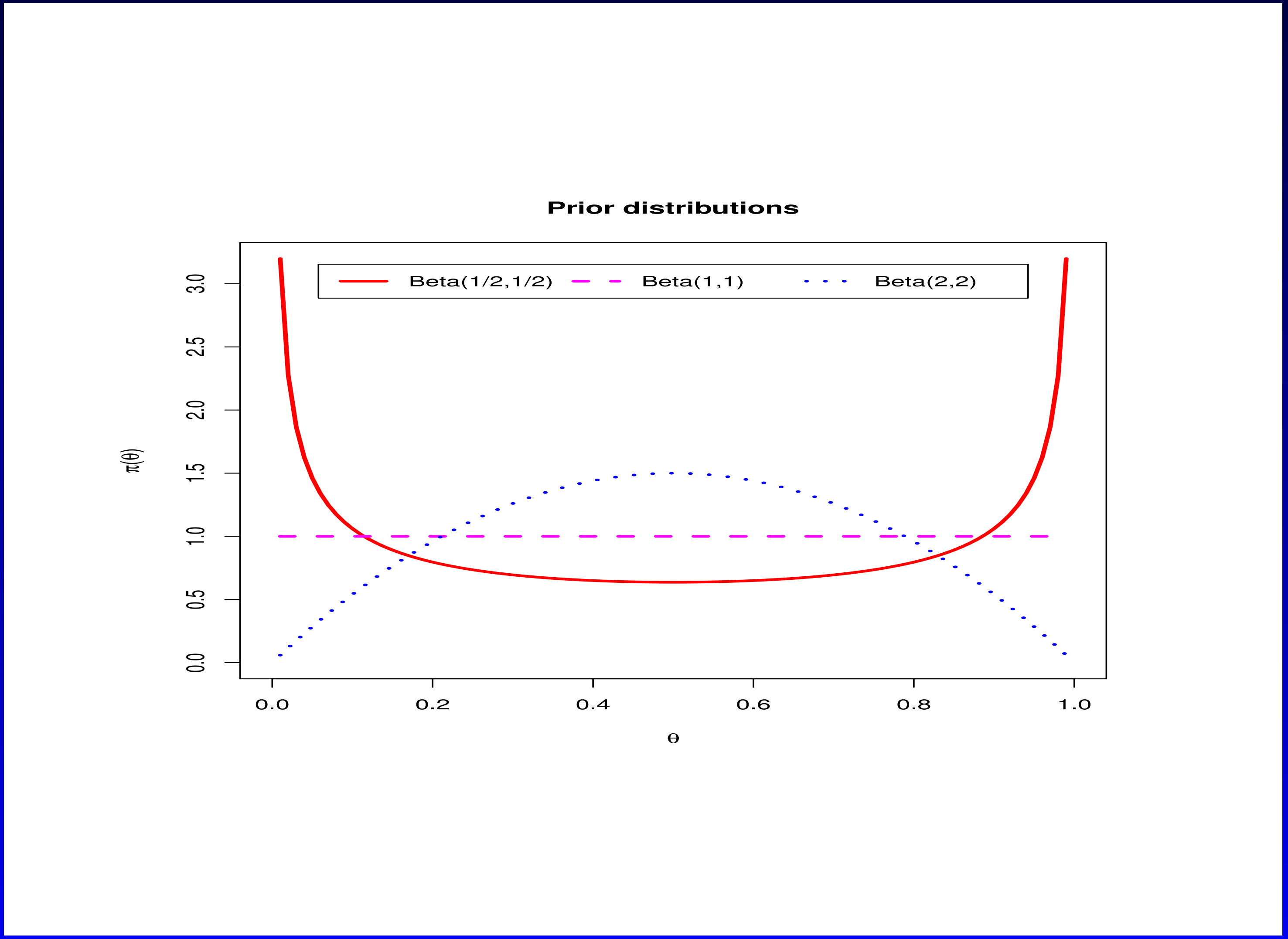
Plot of the three prior distributions:

Example 1 (Carlin and Louis)

Plot of the three prior distributions:

Example 1 (Carlin and Louis)

Plot of the three prior distributions:



Example 1 (Carlin and Louis)

As we showed earlier the posterior distribution under this conjugate setup will be given as:

$$p(\theta|x) \sim Beta(\alpha + x, \beta + n - x)$$

Example 1 (Carlin and Louis)

As we showed earlier the posterior distribution under this conjugate setup will be given as:

$$p(\theta|x) \sim Beta(\alpha + x, \beta + n - x)$$

Thus the respective posteriors of the three prior choices will be:

Example 1 (Carlin and Louis)

As we showed earlier the posterior distribution under this conjugate setup will be given as:

$$p(\theta|x) \sim Beta(\alpha + x, \beta + n - x)$$

Thus the respective posteriors of the three prior choices will be:

- $p(\theta|x) \sim Beta(13.5, 3.5)$, for the Jeffreys prior

Example 1 (Carlin and Louis)

As we showed earlier the posterior distribution under this conjugate setup will be given as:

$$p(\theta|x) \sim Beta(\alpha + x, \beta + n - x)$$

Thus the respective posteriors of the three prior choices will be:

- $p(\theta|x) \sim Beta(13.5, 3.5)$, for the Jeffreys prior
- $p(\theta|x) \sim Beta(14, 4)$, for the noninformative prior

Example 1 (Carlin and Louis)

As we showed earlier the posterior distribution under this conjugate setup will be given as:

$$p(\theta|x) \sim Beta(\alpha + x, \beta + n - x)$$

Thus the respective posteriors of the three prior choices will be:

- $p(\theta|x) \sim Beta(13.5, 3.5)$, for the Jeffreys prior
- $p(\theta|x) \sim Beta(14, 4)$, for the noninformative prior
- $p(\theta|x) \sim Beta(15, 5)$, for the skeptical prior

Example 1 (Carlin and Louis)

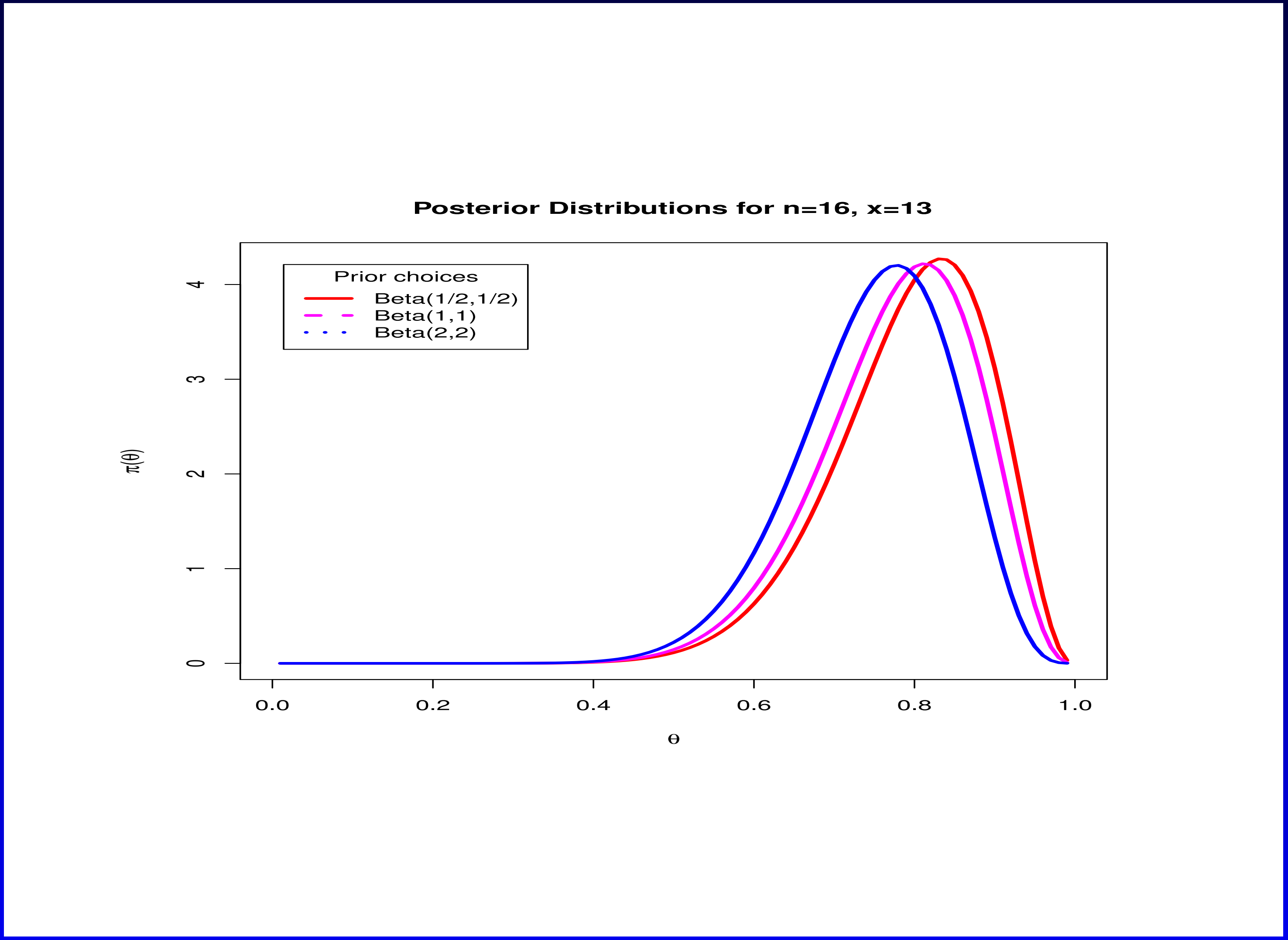
Plot of the three posterior distributions:

Example 1 (Carlin and Louis)

Plot of the three posterior distributions:

Example 1 (Carlin and Louis)

Plot of the three posterior distributions:



Example 2: Normal/Normal model

Assume that $x_i|\theta \overset{iid}{\sim} N(\theta, \sigma^2)$ for $i = 1, 2, \ldots, n$ with σ^2 being known. Then we have: $\overline{x}|\theta \sim N\left(\theta, \sigma^2/n\right)$

Example 2: Normal/Normal model

Assume that $x_i|\theta \overset{iid}{\sim} N(\theta, \sigma^2)$ for $i = 1, 2, \ldots, n$ with σ^2 being known. Then we have: $\overline{x}|\theta \sim N\left(\theta, \sigma^2/n\right)$

The conjugate prior is: $p(\theta) \sim N\left(\mu, \tau^2\right)$

Example 2: Normal/Normal model

Assume that $x_i|\theta \overset{iid}{\sim} N(\theta, \sigma^2)$ for $i = 1, 2, \ldots, n$ with σ^2 being known. Then we have: $\overline{x}|\theta \sim N\left(\theta, \sigma^2/n\right)$

The conjugate prior is: $p(\theta) \sim N\left(\mu, \tau^2\right)$

Then the posterior distribution is given by:

$$p(\theta|\overline{x}) \sim N\left(\frac{\frac{\sigma^2}{n}\mu + \tau^2\overline{x}}{\frac{\sigma^2}{n} + \tau^2}, \frac{\frac{\sigma^2}{n}\tau^2}{\frac{\sigma^2}{n} + \tau^2}\right)$$

Example 2: Normal/Normal model

Assume that $x_i|\theta \overset{iid}{\sim} N(\theta, \sigma^2)$ for $i = 1, 2, \ldots, n$ with σ^2 being known. Then we have: $\overline{x}|\theta \sim N\left(\theta, \sigma^2/n\right)$

The conjugate prior is: $p(\theta) \sim N\left(\mu, \tau^2\right)$

Then the posterior distribution is given by:

$$p(\theta|\overline{x}) \sim N\left(\frac{\frac{\sigma^2}{n}\mu + \tau^2\overline{x}}{\frac{\sigma^2}{n} + \tau^2}, \frac{\frac{\sigma^2}{n}\tau^2}{\frac{\sigma^2}{n} + \tau^2}\right)$$

If we will define:

$$K_n = \frac{\frac{\sigma^2}{n}}{\frac{\sigma^2}{n} + \tau^2}$$

where $0 \leq K_n \leq 1$ we have:

Example 2: Normal/Normal model

$$E[\theta|\overline{x}] = K_n\mu + (1 - K_n)\overline{x}$$

Example 2: Normal/Normal model

$$E[\theta|\overline{x}] = K_n\mu + (1 - K_n)\overline{x}$$

$$V[\theta|\overline{x}] = K_n\tau^2 = (1 - K_n)\sigma^2/n$$

Example 2: Normal/Normal model

$$E[\theta|\overline{x}] = K_n\mu + (1 - K_n)\overline{x}$$

$$V[\theta|\overline{x}] = K_n\tau^2 = (1 - K_n)\sigma^2/n$$

- $E[\theta|\overline{x}]$ is a convex combination of the prior mean and the current data with the weight depending on the variance terms

Example 2: Normal/Normal model

$$E[\theta|\overline{x}] = K_n\mu + (1 - K_n)\overline{x}$$

$$V[\theta|\overline{x}] = K_n\tau^2 = (1 - K_n)\sigma^2/n$$

- $E[\theta|\overline{x}]$ is a convex combination of the prior mean and the current data with the weight depending on the variance terms
- $V[\theta|\overline{x}] \leq min\{\tau^2, \sigma^2/n\}$

Example 2: Normal/Normal model

$$E[\theta|\overline{x}] = K_n\mu + (1 - K_n)\overline{x}$$

$$V[\theta|\overline{x}] = K_n\tau^2 = (1 - K_n)\sigma^2/n$$

- $E[\theta|\overline{x}]$ is a convex combination of the prior mean and the current data with the weight depending on the variance terms
- $V[\theta|\overline{x}] \leq min\{\tau^2, \sigma^2/n\}$
- As $n \uparrow$ the posterior converges to a point mass at $\overline{x}$ (the MLE)

Example 2: Normal/Normal model

$$E[\theta|\overline{x}] = K_n\mu + (1 - K_n)\overline{x}$$

$$V[\theta|\overline{x}] = K_n\tau^2 = (1 - K_n)\sigma^2/n$$

- $E[\theta|\overline{x}]$ is a convex combination of the prior mean and the current data with the weight depending on the variance terms
- $V[\theta|\overline{x}] \leq min\{\tau^2, \sigma^2/n\}$
- As $n \uparrow$ the posterior converges to a point mass at $\overline{x}$ (the MLE)
- As $\tau^2 \uparrow$ then the posterior converges to the $N(\overline{x}, \sigma^2/n)$

Example 2: Normal/Normal model

$$E[\theta|\overline{x}] = K_n\mu + (1 - K_n)\overline{x}$$

$$V[\theta|\overline{x}] = K_n\tau^2 = (1 - K_n)\sigma^2/n$$

- $E[\theta|\overline{x}]$ is a convex combination of the prior mean and the current data with the weight depending on the variance terms
- $V[\theta|\overline{x}] \leq min\{\tau^2, \sigma^2/n\}$
- As $n \uparrow$ the posterior converges to a point mass at $\overline{x}$ (the MLE)
- As $\tau^2 \uparrow$ then the posterior converges to the $N(\overline{x}, \sigma^2/n)$
- As $\tau^2 \downarrow$ then the posterior converges a point mass at μ (the prior mean)

Example 2: Normal/Normal model

Lets look on some graphical illustrations regarding the effect of the sample size n and the variance of the prior distribution, τ^2. Specifically, lets assume that $\overline{x} = 4$ and:

Example 2: Normal/Normal model

Lets look on some graphical illustrations regarding the effect of the sample size n and the variance of the prior distribution, τ^2. Specifically, lets assume that $\overline{x} = 4$ and:

- $n = 1, 10, 100$ with $p(\theta) \sim N(0, 1)$

Example 2: Normal/Normal model

Lets look on some graphical illustrations regarding the effect of the sample size n and the variance of the prior distribution, τ^2. Specifically, lets assume that $\overline{x} = 4$ and:

- $n = 1, 10, 100$ with $p(\theta) \sim N(0, 1)$
- $n = 1$ with $p(\theta) \sim N\left(0, 10^2\right)$

Example 2: Normal/Normal model

Lets look on some graphical illustrations regarding the effect of the sample size n and the variance of the prior distribution, τ^2. Specifically, lets assume that $\overline{x} = 4$ and:

- $n = 1, 10, 100$ with $p(\theta) \sim N(0, 1)$
- $n = 1$ with $p(\theta) \sim N\left(0, 10^2\right)$
- $n = 1, 10, 100$ with $p(\theta) \sim N\left(0, 0.1^2\right)$

Example 2: Normal/Normal model

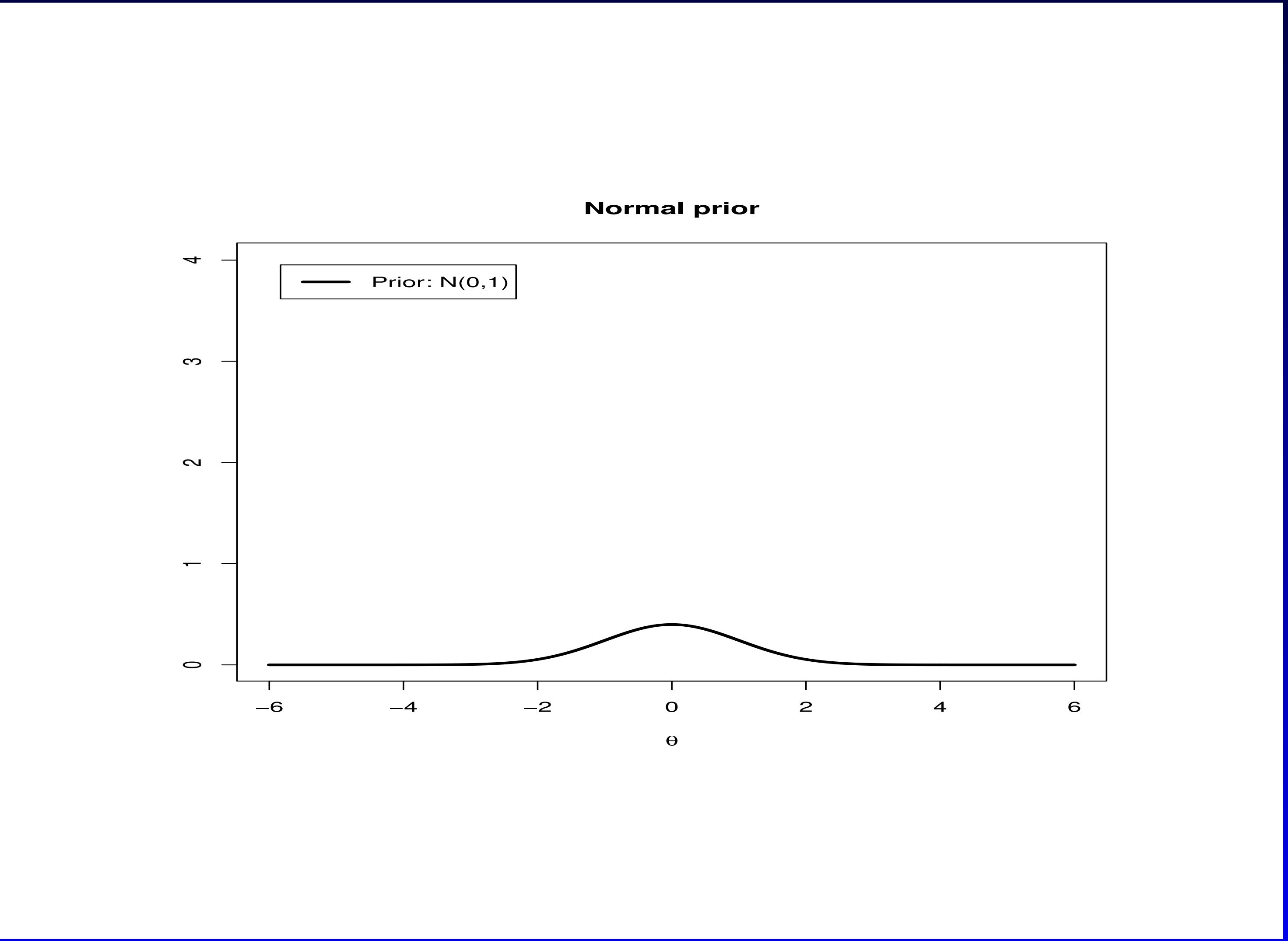
Plot of the $N(0, 1)$ prior distribution:

Example 2: Normal/Normal model

Plot of the $N(0,1)$ prior distribution:

Example 2: Normal/Normal model

Plot of the $N(0, 1)$ prior distribution:



Example 2: Normal/Normal model

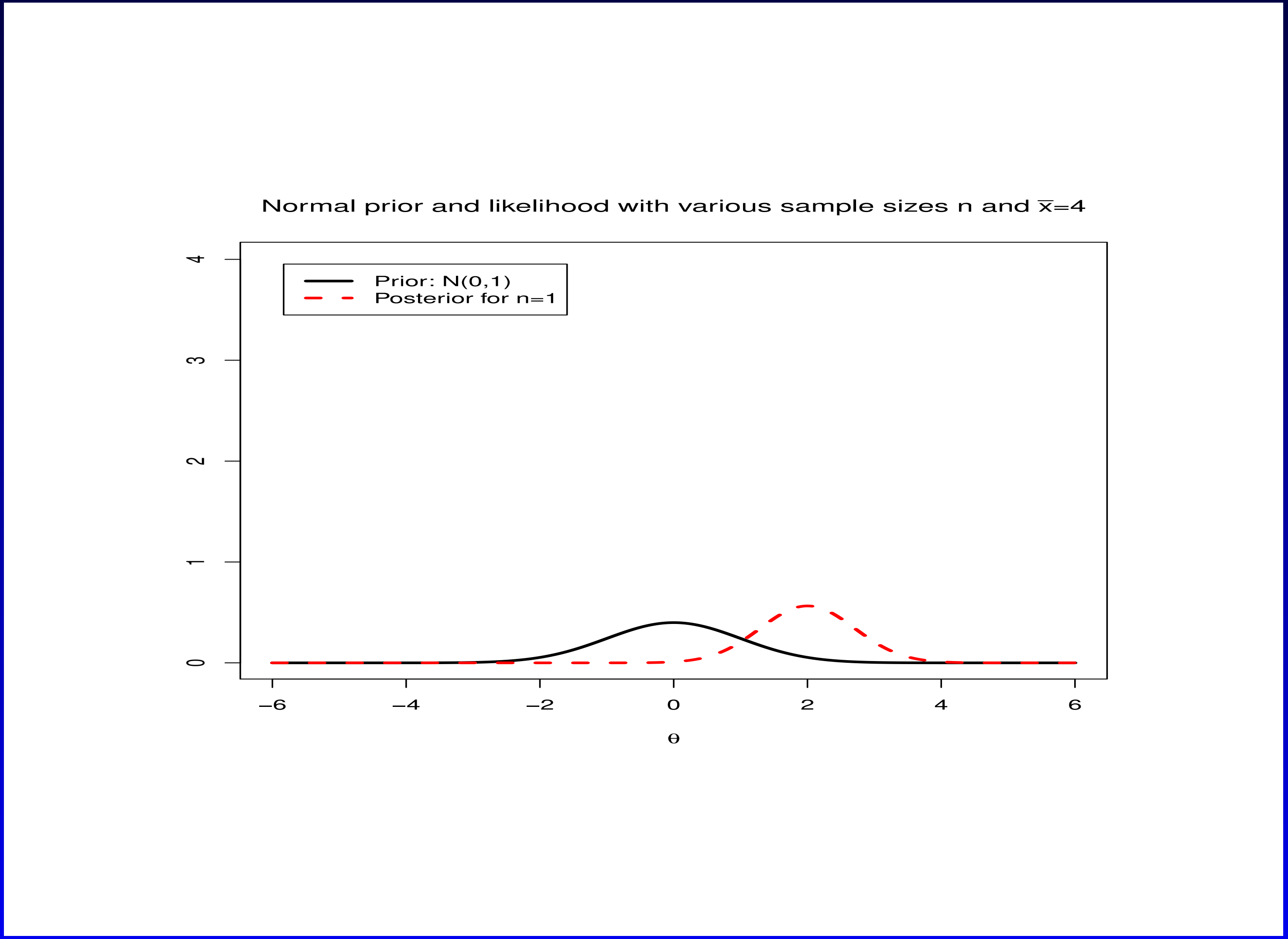
Plot of $p(\theta|x)$, when $n = 1$ with $p(\theta) \sim N(0,1)$:

Example 2: Normal/Normal model

Plot of $p(\theta|x)$, when $n = 1$ with $p(\theta) \sim N(0, 1)$:

Example 2: Normal/Normal model

Plot of $p(\theta|x)$, when $n = 1$ with $p(\theta) \sim N(0, 1)$:



Example 2: Normal/Normal model

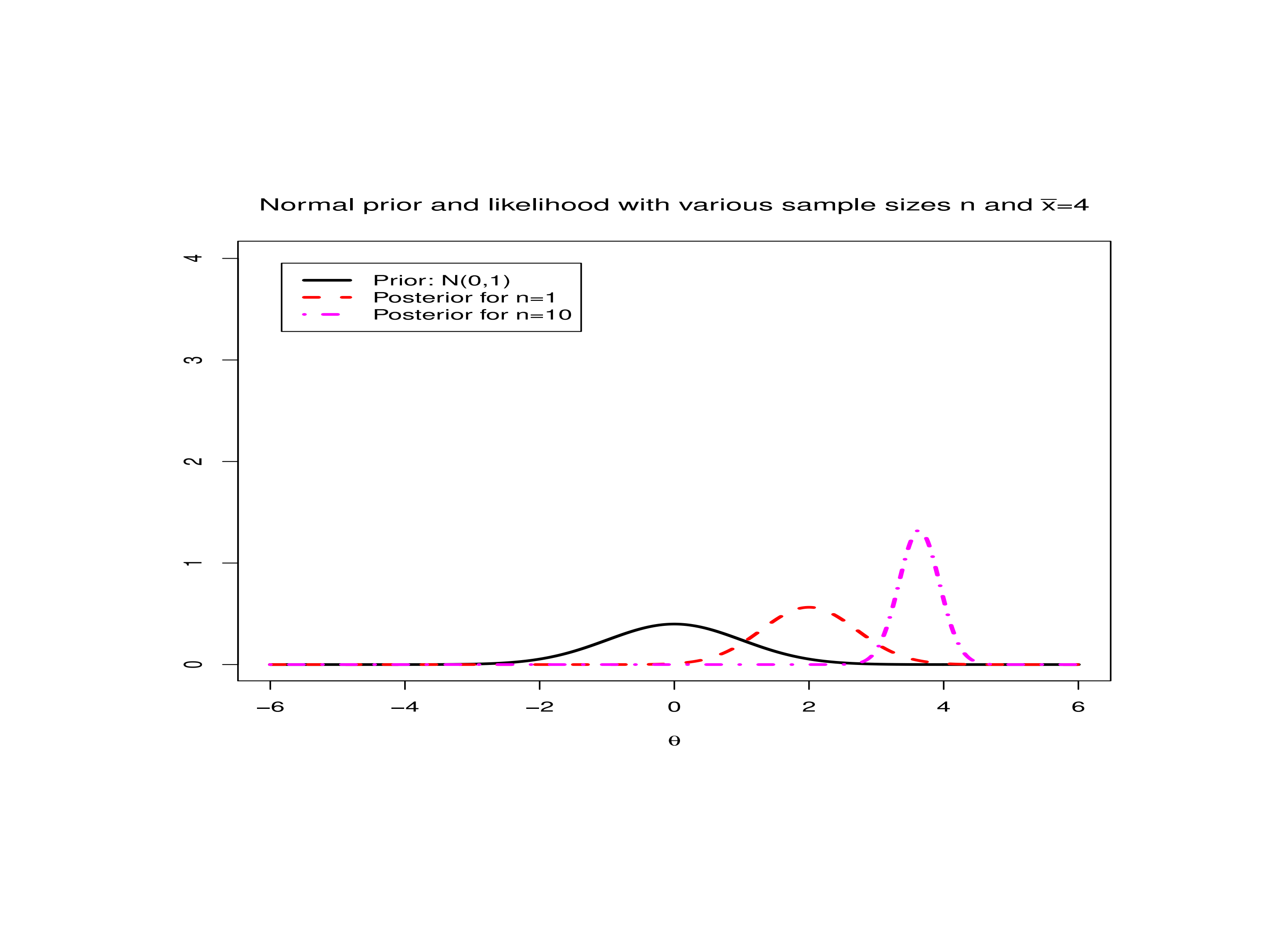
Plot of $p(\theta|x)$, when $n = 1, 10$ with $p(\theta) \sim N(0, 1)$:

Example 2: Normal/Normal model

Plot of $p(\theta|x)$, when $n = 1, 10$ with $p(\theta) \sim N(0, 1)$:

Example 2: Normal/Normal model

Plot of $p(\theta|x)$, when $n = 1, 10$ with $p(\theta) \sim N(0, 1)$:



Example 2: Normal/Normal model

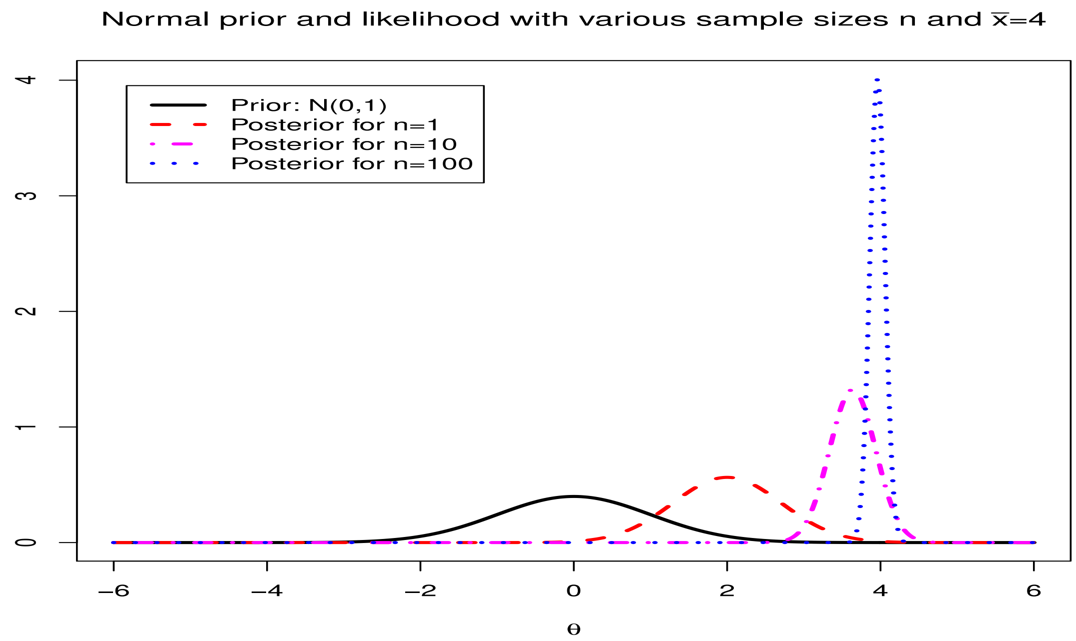
Plot of $p(\theta|x)$, when $n = 1, 10, 100$ with $p(\theta) \sim N(0, 1)$:

Example 2: Normal/Normal model

Plot of $p(\theta|x)$, when $n = 1, 10, 100$ with $p(\theta) \sim N(0, 1)$:

Example 2: Normal/Normal model

Plot of $p(\theta|x)$, when $n = 1, 10, 100$ with $p(\theta) \sim N(0, 1)$:



Example 2: Normal/Normal model

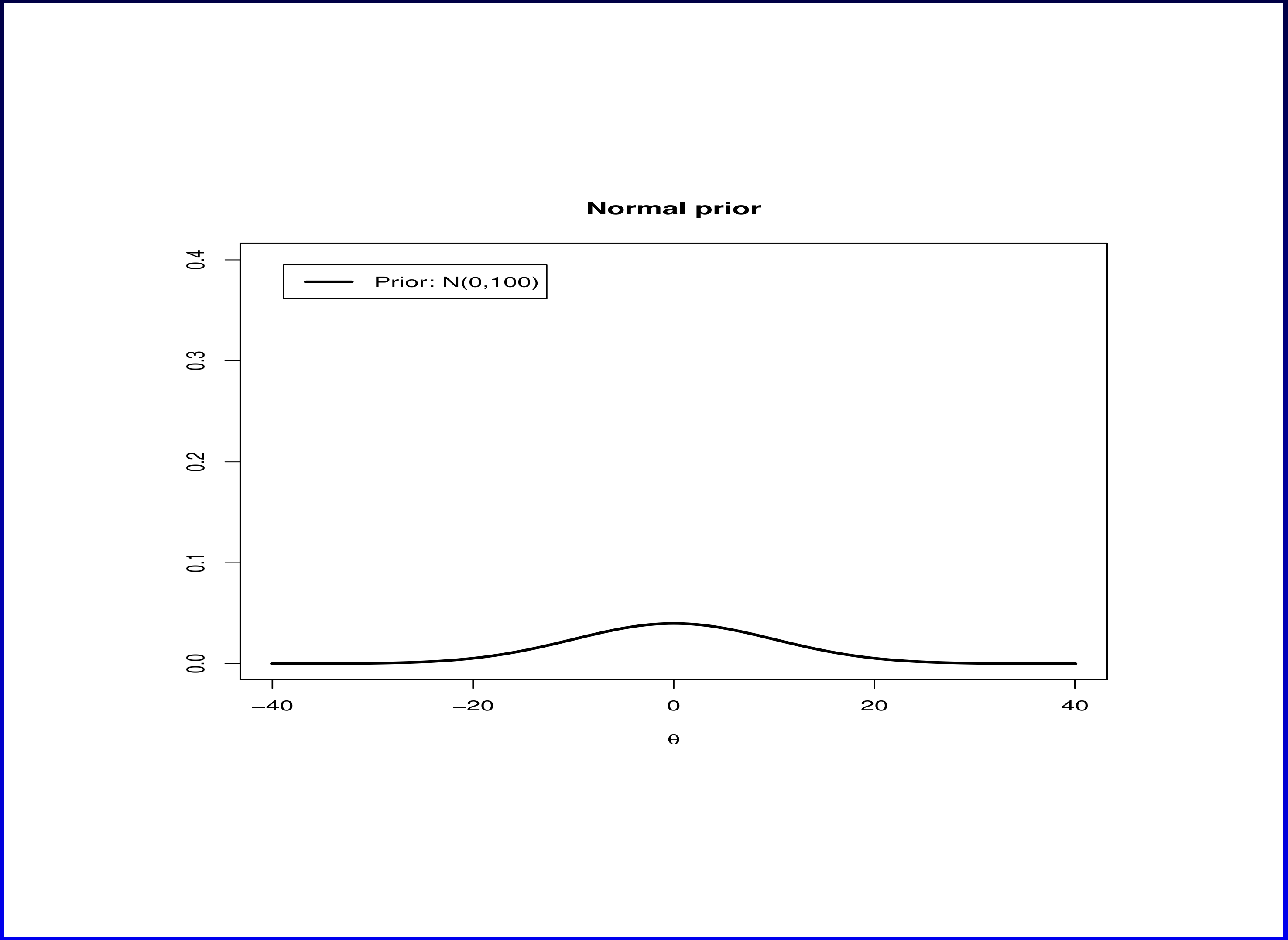
Plot of the $N\left(0, 10^2\right)$ prior distribution:

Example 2: Normal/Normal model

Plot of the $N\left(0, 10^2\right)$ prior distribution:

Example 2: Normal/Normal model

Plot of the $N\left(0, 10^2\right)$ prior distribution:



Example 2: Normal/Normal model

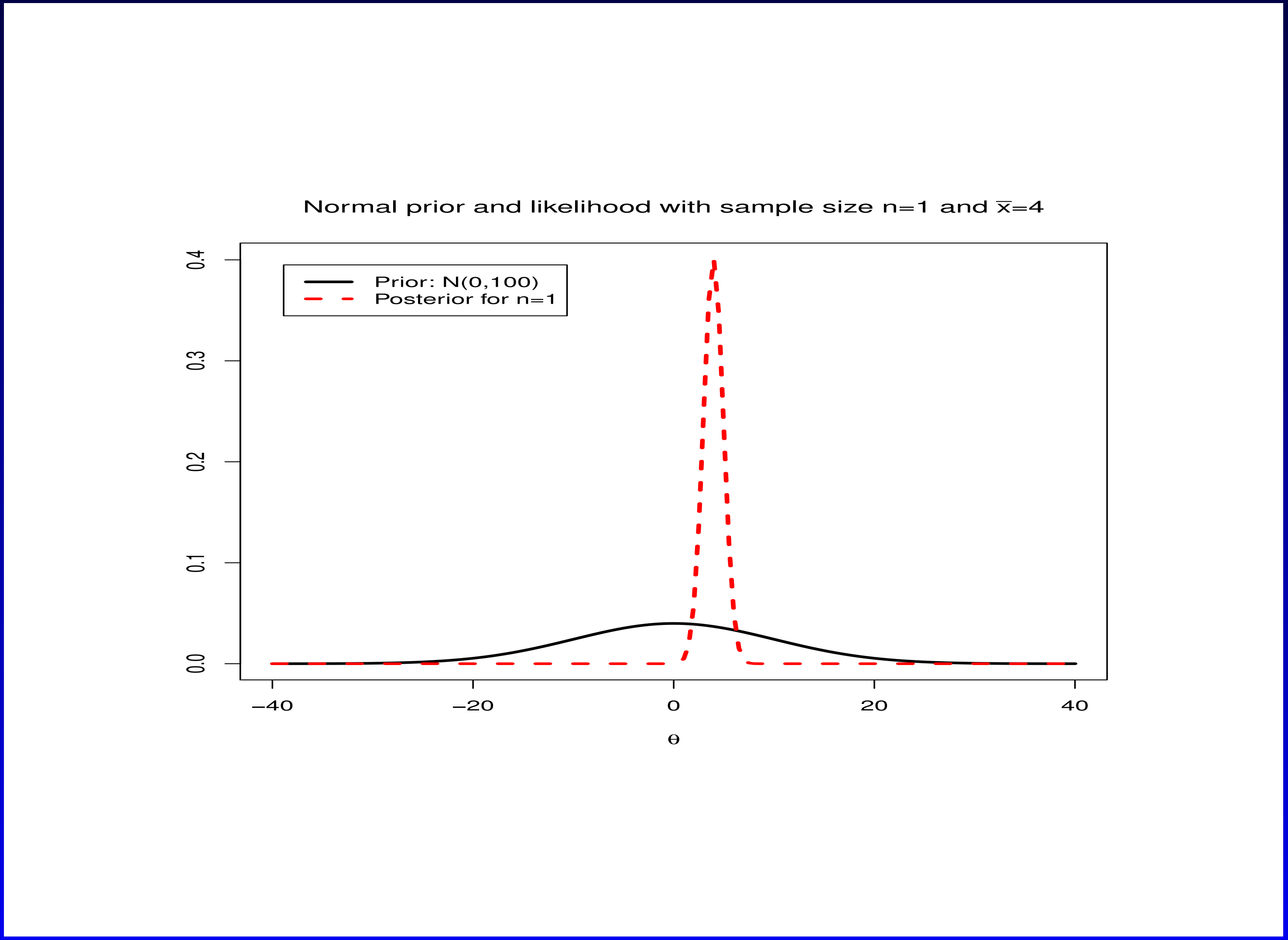
Plot of $p(\theta|x)$, when $n = 1$ with $p(\theta) \sim N\left(0, 10^2\right)$:

Example 2: Normal/Normal model

Plot of $p(\theta|x)$, when $n = 1$ with $p(\theta) \sim N\left(0, 10^2\right)$:

Example 2: Normal/Normal model

Plot of $p(\theta|x)$, when $n = 1$ with $p(\theta) \sim N\left(0, 10^2\right)$:



Example 2: Normal/Normal model

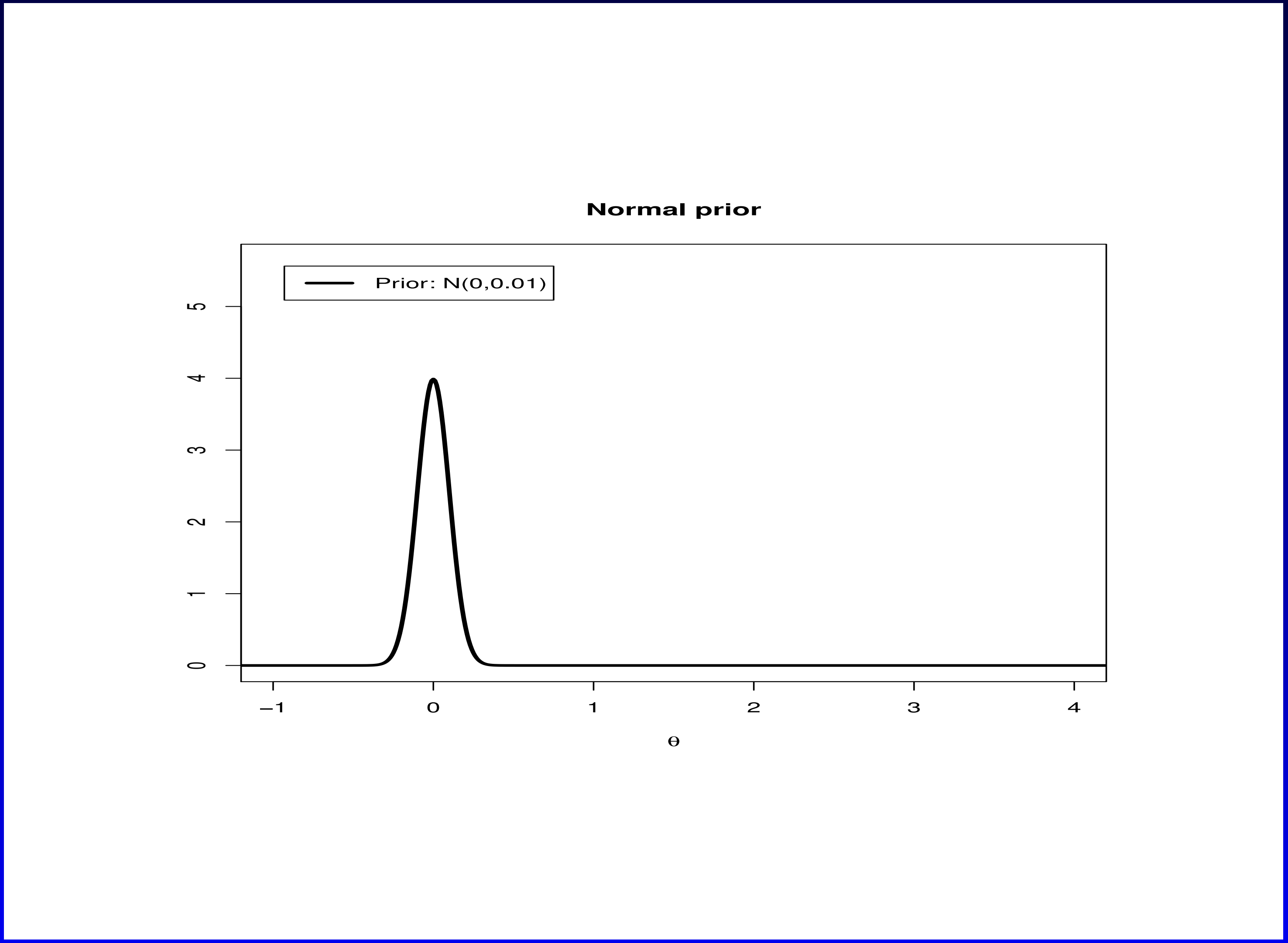
Plot of the $N\left(0, 0.1^2\right)$ prior distribution:

Example 2: Normal/Normal model

Plot of the $N\left(0, 0.1^2\right)$ prior distribution:

Example 2: Normal/Normal model

Plot of the $N\left(0, 0.1^2\right)$ prior distribution:



Example 2: Normal/Normal model

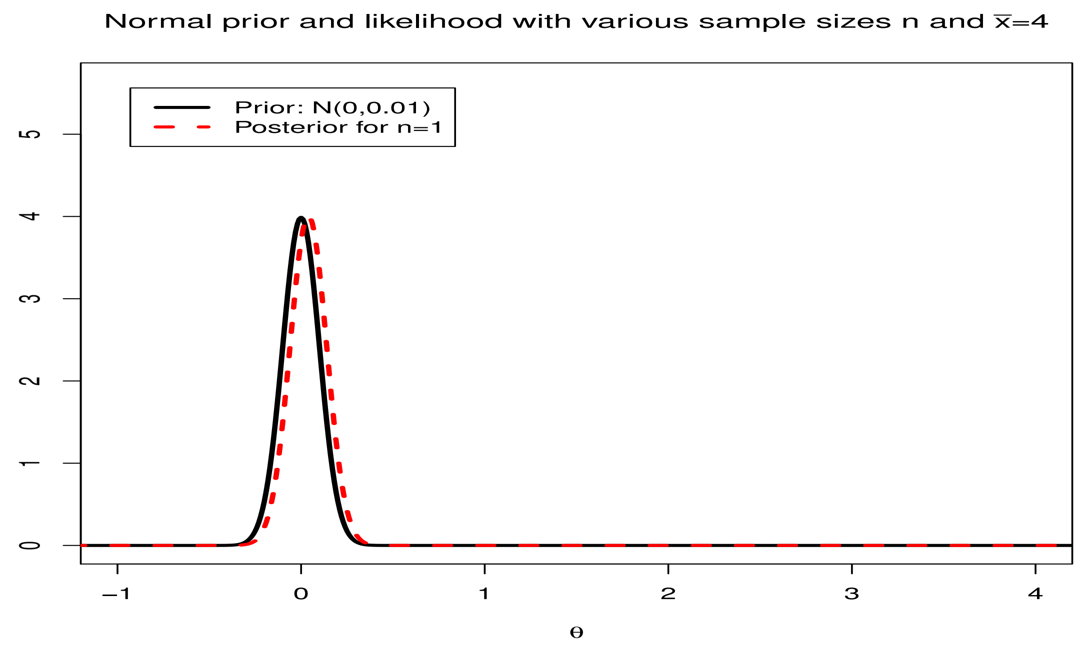
Plot of $p(\theta|x)$, when $n = 1$ with $p(\theta) \sim N\left(0, 0.1^2\right)$:

Example 2: Normal/Normal model

Plot of $p(\theta|x)$, when $n = 1$ with $p(\theta) \sim N\left(0, 0.1^2\right)$:

Example 2: Normal/Normal model

Plot of $p(\theta|x)$, when $n = 1$ with $p(\theta) \sim N\left(0, 0.1^2\right)$:



Example 2: Normal/Normal model

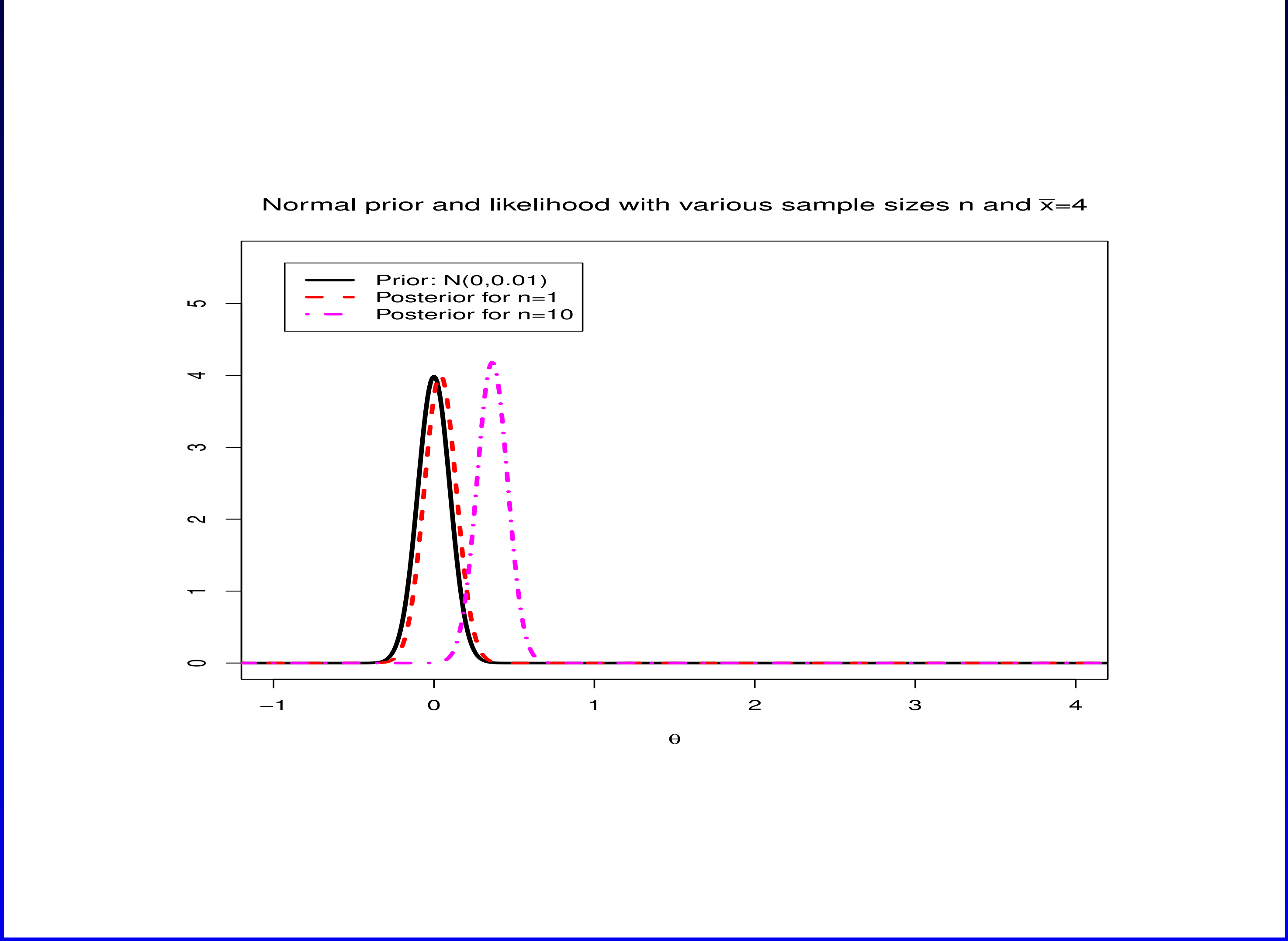
Plot of $p(\theta|x)$, when $n = 1, 10$ with $p(\theta) \sim N\left(0, 0.1^2\right)$:

Example 2: Normal/Normal model

Plot of $p(\theta|x)$, when $n = 1, 10$ with $p(\theta) \sim N(0, 0.1^2)$:

Example 2: Normal/Normal model

Plot of $p(\theta|x)$, when $n = 1, 10$ with $p(\theta) \sim N\left(0, 0.1^2\right)$:



Example 2: Normal/Normal model

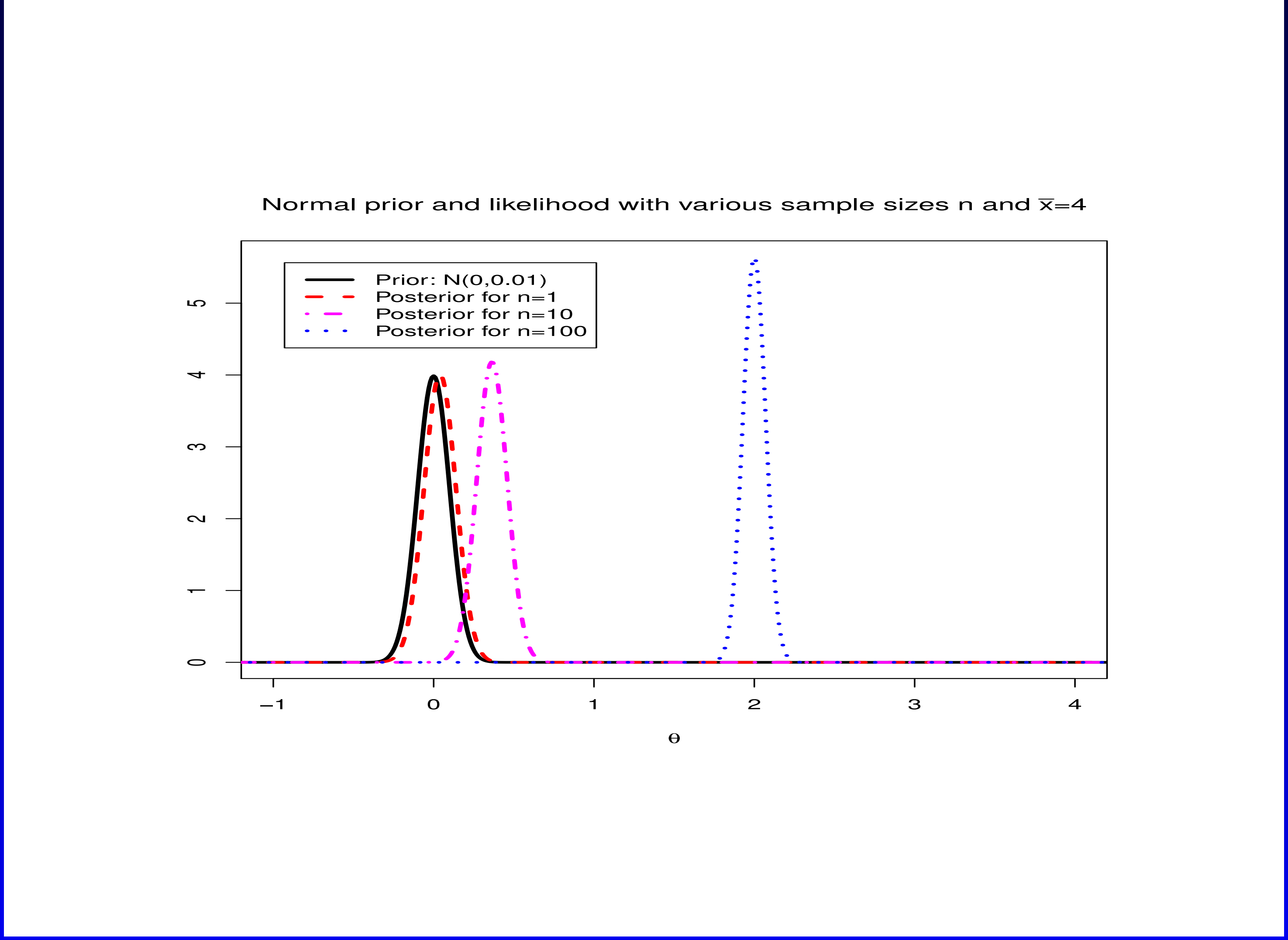
Plot of $p(\theta|x)$, when $n = 1, 10, 100$ with $p(\theta) N\left(0, 0.1^2\right)$:

Example 2: Normal/Normal model

Plot of $p(\theta|x)$, when $n = 1, 10, 100$ with $p(\theta)N\left(0, 0.1^2\right)$:

Example 2: Normal/Normal model

Plot of $p(\theta|x)$, when $n = 1, 10, 100$ with $p(\theta) N\left(0, 0.1^2\right)$:



Inference regarding θ

For a Bayesian, the posterior distribution is a complete description of the unknown parameter θ. Thus for a Bayesian the posterior distribution is the inference.

Inference regarding θ

For a Bayesian, the posterior distribution is a complete description of the unknown parameter θ. Thus for a Bayesian the posterior distribution is the inference.

However, most people (especially non statisticians) are accustomed to the usual form of frequentist inference procedures, like point/interval estimates and hypothesis testing for θ.

Inference regarding θ

For a Bayesian, the posterior distribution is a complete description of the unknown parameter θ. Thus for a Bayesian the posterior distribution is the inference.

However, most people (especially non statisticians) are accustomed to the usual form of frequentist inference procedures, like point/interval estimates and hypothesis testing for θ.

In what follows we will provide, with the help of decision theory, the most representative ways of summarizing the posterior distribution to the well known frequentist's forms of inference.

Decision Theory: Basic definitions

- $\Theta =$ parameter space, all possible values of θ

Decision Theory: Basic definitions

- $\Theta =$ parameter space, all possible values of θ
- $\mathcal{A} =$ action space, all possible values a for estimating θ

Decision Theory: Basic definitions

- $\Theta =$ parameter space, all possible values of θ
- $\mathcal{A} =$ action space, all possible values a for estimating θ
- $L(\theta, a) : \Theta \times \mathcal{A} \longrightarrow \Re$, loss occurred (profit if negative) when we take action $a \in \mathcal{A}$ and the the true state is $\theta \in \Theta$.

Decision Theory: Basic definitions

- $\Theta =$ parameter space, all possible values of θ
- $\mathcal{A} =$ action space, all possible values a for estimating θ
- $L(\theta, a) : \Theta \times \mathcal{A} \to \Re$, loss occurred (profit if negative) when we take action $a \in \mathcal{A}$ and the the true state is $\theta \in \Theta$.
- The triplet $(\Theta, \mathcal{A}, L(\theta, a))$ along with the data $\mathbf{x}$ from the likelihood $f(\mathbf{x}|\theta)$ constitute a statistical decision problem.

Decision Theory: Basic definitions

- $\Theta =$ parameter space, all possible values of θ
- $\mathcal{A} =$ action space, all possible values a for estimating θ
- $L(\theta, a) : \Theta \times \mathcal{A} \to \Re$, loss occurred (profit if negative) when we take action $a \in \mathcal{A}$ and the the true state is $\theta \in \Theta$.
- The triplet $(\Theta, \mathcal{A}, L(\theta, a))$ along with the data $\mathbf{x}$ from the likelihood $f(\mathbf{x}|\theta)$ constitute a statistical decision problem.
- $\mathcal{X} =$ all possible data of the experiment.

Decision Theory: Basic definitions

- $\Theta =$ parameter space, all possible values of θ
- $\mathcal{A} =$ action space, all possible values a for estimating θ
- $L(\theta, a) : \Theta \times \mathcal{A} \longrightarrow \Re$, loss occurred (profit if negative) when we take action $a \in \mathcal{A}$ and the the true state is $\theta \in \Theta$.
- The triplet $(\Theta, \mathcal{A}, L(\theta, a))$ along with the data $\mathbf{x}$ from the likelihood $f(\mathbf{x}|\theta)$ constitute a statistical decision problem.
- $\mathcal{X} =$ all possible data of the experiment.
- $\delta(\mathbf{x}) : \mathcal{X} \rightarrow \mathcal{A}$, decision rule (strategy), which indicates which action $a \in \mathcal{A}$ we will pick, when $\mathbf{x} \in \mathcal{X}$ is observed.

Decision Theory: Basic definitions

- $\Theta =$ parameter space, all possible values of θ
- $\mathcal{A} =$ action space, all possible values a for estimating θ
- $L(\theta, a) : \Theta \times \mathcal{A} \longrightarrow \Re$, loss occurred (profit if negative) when we take action $a \in \mathcal{A}$ and the the true state is $\theta \in \Theta$.
- The triplet $(\Theta, \mathcal{A}, L(\theta, a))$ along with the data $\mathbf{x}$ from the likelihood $f(\mathbf{x}|\theta)$ constitute a statistical decision problem.
- $\mathcal{X} =$ all possible data of the experiment.
- $\delta(\mathbf{x}) : \mathcal{X} \longrightarrow \mathcal{A}$, decision rule (strategy), which indicates which action $a \in \mathcal{A}$ we will pick, when $\mathbf{x} \in \mathcal{X}$ is observed.
- $\mathcal{D} =$ set of all available decision rules.

Decision Theory: Evaluating decision rules

Our goal is to obtain the decision rule (strategy), from the set $\mathcal{D}$, for which we have the minimum loss.

Decision Theory: Evaluating decision rules

Our goal is to obtain the decision rule (strategy), from the set $\mathcal{D}$, for which we have the minimum loss.

But the loss function, $L(\theta, a)$, is a random quantity.

Decision Theory: Evaluating decision rules

Our goal is to obtain the decision rule (strategy), from the set $\mathcal{D}$, for which we have the minimum loss.

But the loss function, $L(\theta, a)$, is a random quantity.

From a Frequentist perspective it is random in $\mathbf{x}$ (since we fixed θ).

Decision Theory: Evaluating decision rules

Our goal is to obtain the decision rule (strategy), from the set $\mathcal{D}$, for which we have the minimum loss.

But the loss function, $L(\theta, a)$, is a random quantity.

From a Frequentist perspective it is random in $\mathbf{x}$ (since we fixed θ).

From a Bayesian perspective it is random in θ (since we fixed the data $\mathbf{x}$).

Decision Theory: Evaluating decision rules

Our goal is to obtain the decision rule (strategy), from the set $\mathcal{D}$, for which we have the minimum loss.

But the loss function, $L(\theta, a)$, is a random quantity.

From a Frequentist perspective it is random in $\mathbf{x}$ (since we fixed θ).

From a Bayesian perspective it is random in θ (since we fixed the data $\mathbf{x}$).

Thus, each school will evaluate a decision rule differently, by finding the average loss, with respect to what is random each time.

Decision Theory: Frequentist & Posterior Risk

- **Frequentist Risk:** $FR(\,.\,,\delta(\mathbf{x})) : \Theta \rightarrow \Re$, where:

$$FR(\theta,\delta(\mathbf{x})) = E_{X|\theta}\left[L(\theta,\delta(\mathbf{x}))\right] = \int L(\theta,\delta(\mathbf{x}))f(\mathbf{x}|\theta)d\mathbf{x}$$

Decision Theory: Frequentist & Posterior Risk

- **Frequentist Risk:** $FR(\ .\ , \delta(\mathbf{x})) : \Theta \to \Re$, where:

$$FR(\theta, \delta(\mathbf{x})) = E_{X|\theta}\left[L(\theta, \delta(\mathbf{x}))\right] = \int L(\theta, \delta(\mathbf{x})) f(\mathbf{x}|\theta) d\mathbf{x}$$

- **Posterior Risk:** $PR(\theta\ , \delta(.)) : \mathcal{X} \to \Re$, where:

$$PR(\theta, \delta(\mathbf{x})) = E_{\theta|\mathbf{x}}\left[L(\theta, \delta(\mathbf{x}))\right] = \int L(\theta, \delta(\mathbf{x})) p(\theta|\mathbf{x}) d\theta$$

Decision Theory: Frequentist & Posterior Risk

- **Frequentist Risk:** $FR(\ .\ , \delta(\mathbf{x})) : \Theta \rightarrow \Re$, where:

$$FR(\theta, \delta(\mathbf{x})) = E_{X|\theta}\left[L(\theta, \delta(\mathbf{x}))\right] = \int L(\theta, \delta(\mathbf{x})) f(\mathbf{x}|\theta) d\mathbf{x}$$

- **Posterior Risk:** $PR(\theta\ , \delta(.)) : \mathcal{X} \rightarrow \Re$, where:

$$PR(\theta, \delta(\mathbf{x})) = E_{\theta|\mathbf{x}}\left[L(\theta, \delta(\mathbf{x}))\right] = \int L(\theta, \delta(\mathbf{x})) p(\theta|\mathbf{x}) d\theta$$

FR assumes θ to be fixed and $\mathbf{x}$ random, while PR treats θ as random and $\mathbf{x}$ as fixed. Thus each approach takes out (averages) the uncertainty from one source only.

Decision Theory: Bayes risk

For the decision rules to become comparable, it is necessary to integrate out the remaining source of uncertainty to each of the FR and PR. This is achieved with the Bayes Risk:

Decision Theory: Bayes risk

For the decision rules to become comparable, it is necessary to integrate out the remaining source of uncertainty to each of the FR and PR. This is achieved with the Bayes Risk:

$$\begin{aligned} BR(p(\theta), \delta(\mathbf{x})) &= E_{\theta}\left[FR(\theta, \delta(\mathbf{x}))\right] = \int FR(\theta, \delta(\mathbf{x}))p(\theta)d\theta \\ &= E_{X}\left[PR(\theta, \delta(\mathbf{x}))\right] = \int PR(\theta, \delta(\mathbf{x}))f(\mathbf{x})d\mathbf{x} \end{aligned}$$

Decision Theory: Bayes risk

For the decision rules to become comparable, it is necessary to integrate out the remaining source of uncertainty to each of the FR and PR. This is achieved with the Bayes Risk:

$$\begin{aligned} BR(p(\theta), \delta(\mathbf{x})) &= E_\theta\left[FR(\theta, \delta(\mathbf{x}))\right] = \int FR(\theta, \delta(\mathbf{x}))p(\theta)d\theta \\ &= E_X\left[PR(\theta, \delta(\mathbf{x}))\right] = \int PR(\theta, \delta(\mathbf{x}))f(\mathbf{x})d\mathbf{x} \end{aligned}$$

Thus the BR summarizes each decision rule with a single number: the average loss, with respect to random θ and random $\mathbf{x}$ (being irrelevant to which quantity we integrate out first).

Decision Theory: Bayes rule

The decision rule which minimizes the Bayes Risk is called Bayes Rule and is denoted as $\delta_p(.)$. Thus:

Decision Theory: Bayes rule

The decision rule which minimizes the Bayes Risk is called Bayes Rule and is denoted as $\delta_p(.)$. Thus:

$$\delta_p(.) = \inf_{\delta \in \mathcal{D}} \{BR(p(\theta), \delta(\mathbf{x}))\}$$

Decision Theory: Bayes rule

The decision rule which minimizes the Bayes Risk is called Bayes Rule and is denoted as $\delta_p(.)$. Thus:

$$\delta_p(.) = \inf_{\delta \in \mathcal{D}} \{BR(p(\theta), \delta(\mathbf{x}))\}$$

The Bayes rule minimizes the expected (under both uncertainties) loss. It is known as the "rational" player's criterion in picking up a decision rule from $\mathcal{D}$.

Decision Theory: Bayes rule

The decision rule which minimizes the Bayes Risk is called Bayes Rule and is denoted as $\delta_p(.)$. Thus:

$$\delta_p(.) = \inf_{\delta \in \mathcal{D}} \{BR(p(\theta), \delta(\mathbf{x}))\}$$

The Bayes rule minimizes the expected (under both uncertainties) loss. It is known as the “rational” player’s criterion in picking up a decision rule from $\mathcal{D}$.

Bayes rule might not exist for a problem (just as the minimum of function does not always exists).

Decision Theory: Minimax rule

A more conservative player does not wish to minimize the expected loss. He/She is interested in putting a bound to the worst that can happen.

Decision Theory: Minimax rule

A more conservative player does not wish to minimize the expected loss. He/She is interested in putting a bound to the worst that can happen.

This leads to the minimax decision rule $\delta^*(.)$ which is defined as the decision rule for which:

$$\sup_{\theta\in\Theta}\{FR(\theta,\delta^*(.))\} = \inf_{\delta\in\mathcal{D}}\left[\sup_{\theta\in\Theta}\{FR(\theta,\delta(.))\}\right]$$

Decision Theory: Minimax rule

A more conservative player does not wish to minimize the expected loss. He/She is interested in putting a bound to the worst that can happen.

This leads to the minimax decision rule $\delta^*(.)$ which is defined as the decision rule for which:

$$\sup_{\theta\in\Theta}\{FR(\theta,\delta^*(.))\} = \inf_{\delta\in\mathcal{D}}\left[\sup_{\theta\in\Theta}\{FR(\theta,\delta(.))\}\right]$$

The minimax rules takes into account the worst that can happen, ignoring the performance anywhere else. This can lead in some cases to very poor choices.

Inference for θ: Point estimation

The goal is to summarize the posterior distribution to a single summary number.

Inference for θ: Point estimation

The goal is to summarize the posterior distribution to a single summary number.

From a decision theory perspective we assume that $\mathcal{A} = \Theta$ and under the appropriate loss function $L(\theta, a)$ we search for the Bayes rule.

Inference for θ: Point estimation

The goal is to summarize the posterior distribution to a single summary number.

From a decision theory perspective we assume that $\mathcal{A} = \Theta$ and under the appropriate loss function $L(\theta, a)$ we search for the Bayes rule.

E.g.1

If $L(\theta, a) = (\theta - a)^2$ then $\delta_p(\mathbf{x}) = E[\theta|\mathbf{x}]$

Inference for θ: Point estimation

The goal is to summarize the posterior distribution to a single summary number.

From a decision theory perspective we assume that $\mathcal{A} = \Theta$ and under the appropriate loss function $L(\theta, a)$ we search for the Bayes rule.

E.g.1

If $L(\theta, a) = (\theta - a)^2$ then $\delta_p(\mathbf{x}) = E[\theta|\mathbf{x}]$

E.g.2

If $L(\theta, a) = |\theta - a|$ then $\delta_p(\mathbf{x}) = median\{p(\theta|\mathbf{x})\}$

Inference for θ: Interval estimation

In contrast to the frequentist's Confidence Interval (CI), where the parameter θ belongs to the CI with probability 0 or 1, within the Bayesian framework we can have probability statements regarding the parameter θ. Specifically:

Inference for θ: Interval estimation

In contrast to the frequentist's Confidence Interval (CI), where the parameter θ belongs to the CI with probability 0 or 1, within the Bayesian framework we can have probability statements regarding the parameter θ. Specifically:

Any subset $C_\alpha(\mathbf{x})$ of Θ is called a $(1-\alpha)100\%$ credible set if:

$$\int_{C_\alpha(\mathbf{x})} p(\theta|\mathbf{x})d\theta = 1-\alpha$$

Inference for θ: Interval estimation

In contrast to the frequentist's Confidence Interval (CI), where the parameter θ belongs to the CI with probability 0 or 1, within the Bayesian framework we can have probability statements regarding the parameter θ. Specifically:

Any subset $C_\alpha(\mathbf{x})$ of Θ is called a $(1-\alpha)100\%$ credible set if:

$$\int_{C_\alpha(\mathbf{x})} p(\theta|\mathbf{x})d\theta = 1-\alpha$$

In simple words the $(1-\alpha)100\%$ credible set is any subset of the parameter space Θ that has posterior coverage probability equal to $(1-\alpha)100\%$.

Inference for θ: Interval estimation

In contrast to the frequentist's Confidence Interval (CI), where the parameter θ belongs to the CI with probability 0 or 1, within the Bayesian framework we can have probability statements regarding the parameter θ. Specifically:

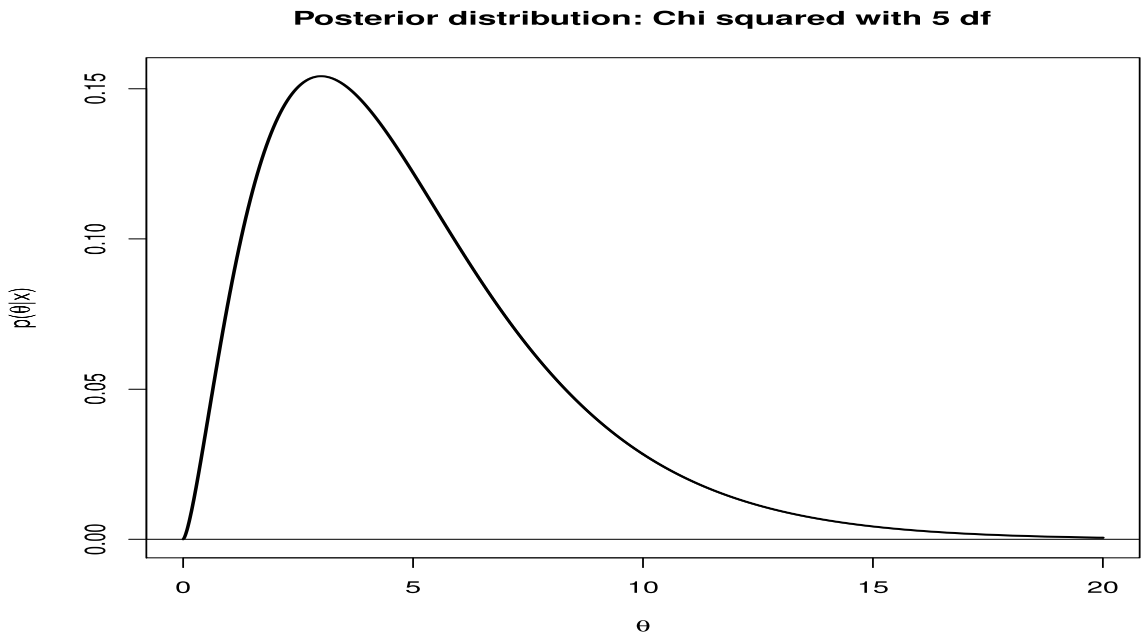
Any subset $C_\alpha(\mathbf{x})$ of Θ is called a $(1-\alpha)100\%$ credible set if:

$$\int_{C_\alpha(\mathbf{x})} p(\theta|\mathbf{x})d\theta = 1-\alpha$$

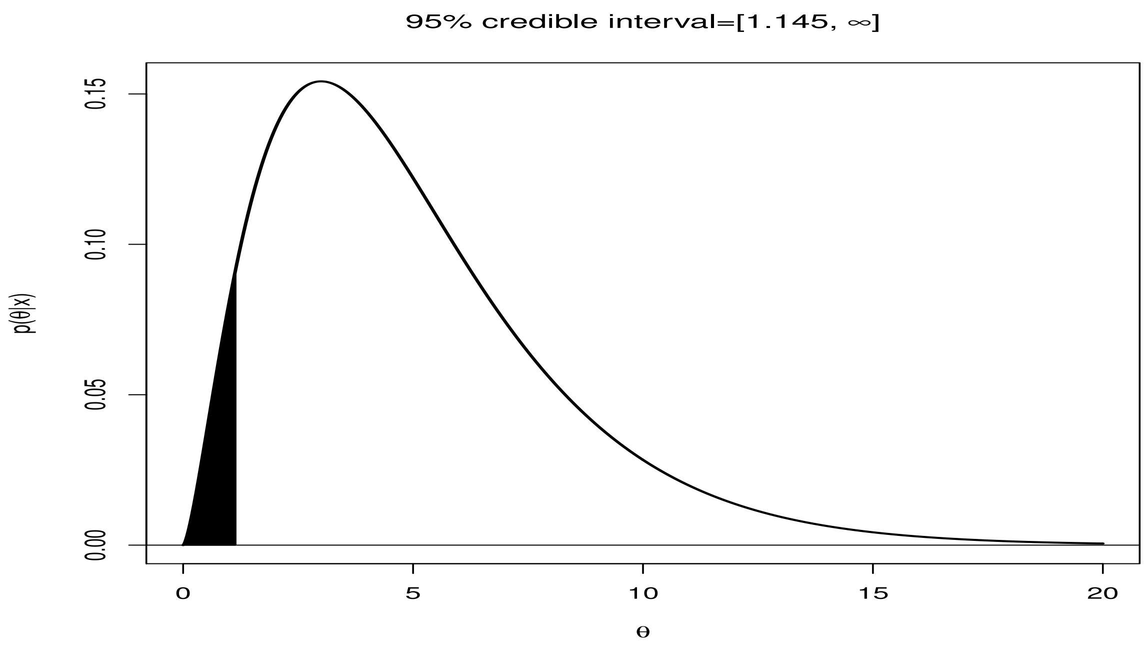
In simple words the $(1-\alpha)100\%$ credible set is any subset of the parameter space Θ that has posterior coverage probability equal to $(1-\alpha)100\%$.

The credible sets are not uniquely defined.

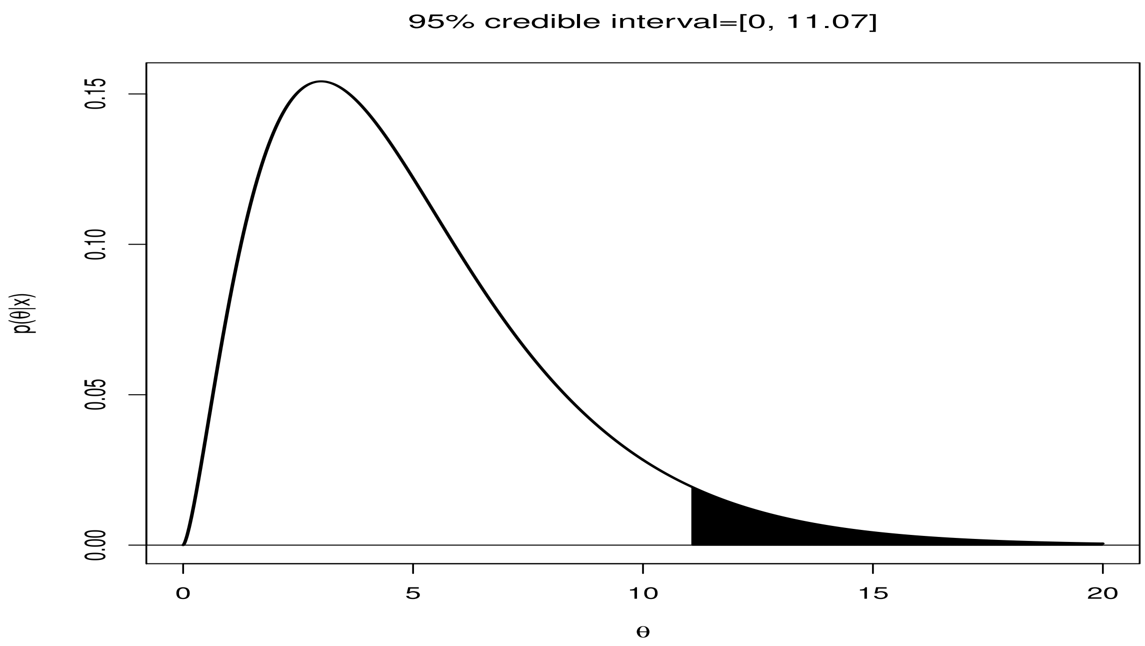
Inference for θ: Interval estimation



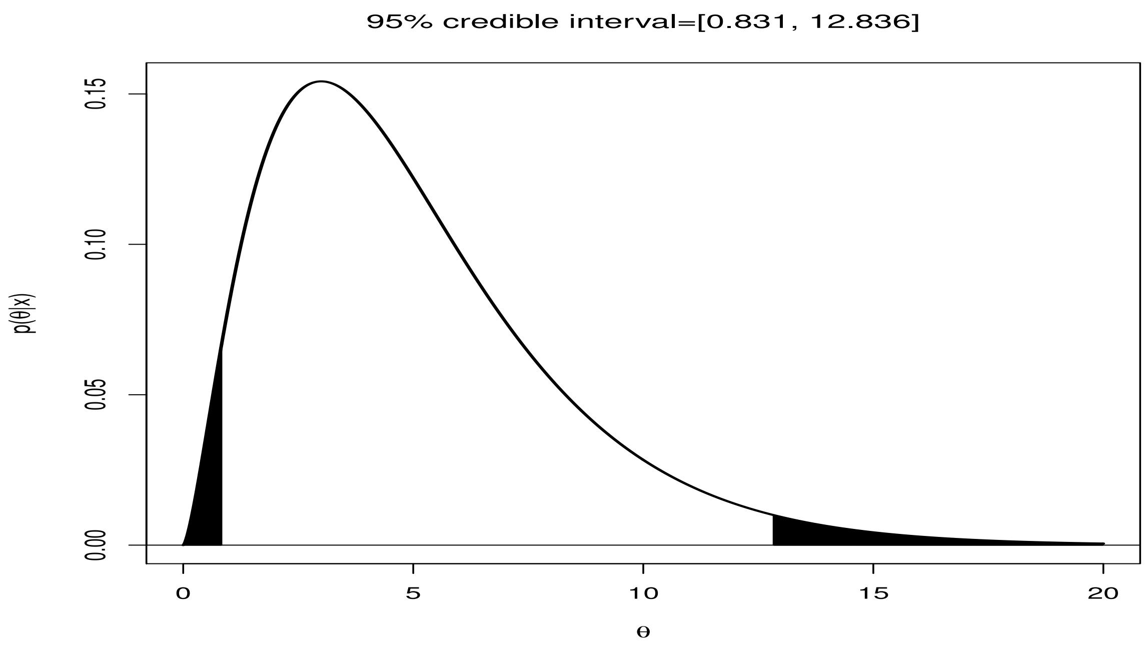
Inference for θ: Interval estimation



Inference for θ: Interval estimation



Inference for θ: Interval estimation



Inference for θ: Interval estimation

For a fixed value of α we would like to obtain the "shortest" credible set. This leads to the credible set that contains the most probable values and is known as Highest Posterior Density (HPD) set. Thus:

Inference for θ: Interval estimation

For a fixed value of α we would like to obtain the "shortest" credible set. This leads to the credible set that contains the most probable values and is known as Highest Posterior Density (HPD) set. Thus:

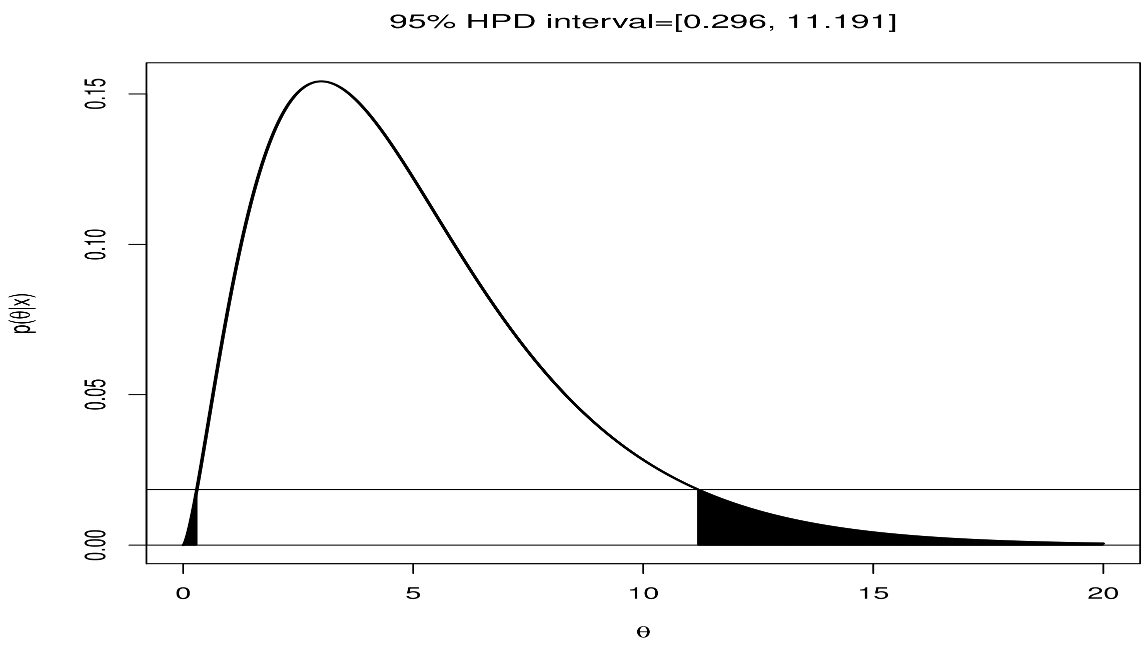
$$HPD_\alpha(\mathbf{x}) = \{\theta : p(\theta|\mathbf{x}) \geq \gamma\}$$

where for the constant γ we have:

$$\int_{HPD_\alpha(\mathbf{x})} p(\theta|\mathbf{x}) d\theta = 1 - \alpha$$

i.e. we keep the most probable region.

Inference for θ: Interval estimation



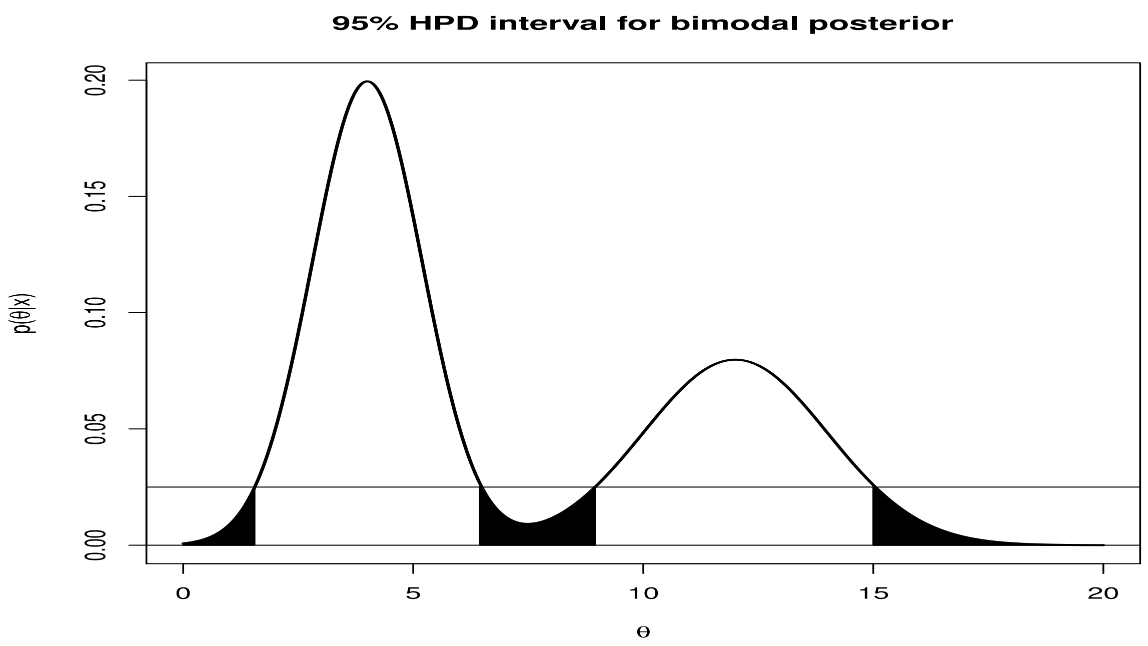
Inference for θ: Interval estimation

- The HPD set is unique and for unimodal, symmetric densities we can obtain it by cutting $\alpha/2$ from each tail.

Inference for θ: Interval estimation

- The HPD set is unique and for unimodal, symmetric densities we can obtain it by cutting $\alpha/2$ from each tail.
- In all other cases we can obtain it numerically. In some cases the HPD might be a union of disjoint sets:

Inference for θ: Interval estimation



Inference for θ: Hypothesis Testing

We are interested in testing $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$

Inference for θ: Hypothesis Testing

We are interested in testing $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$

In frequentist based HT, we assume that H_0 is true and using the test statistics, $T(\mathbf{x})$, we obtain the p-value, which we compare to the level of significance to draw a decision.

Inference for θ: Hypothesis Testing

We are interested in testing $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$

In frequentist based HT, we assume that H_0 is true and using the test statistics, $T(\mathbf{x})$, we obtain the p-value, which we compare to the level of significance to draw a decision.

Several limitations of this approach are known. Like:

Inference for θ: Hypothesis Testing

We are interested in testing $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$

In frequentist based HT, we assume that H_0 is true and using the test statistics, $T(\mathbf{x})$, we obtain the p-value, which we compare to the level of significance to draw a decision.

Several limitations of this approach are known. Like:

- There are cases where the likelihood principle is violated.

Inference for θ: Hypothesis Testing

We are interested in testing $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$

In frequentist based HT, we assume that H_0 is true and using the test statistics, $T(\mathbf{x})$, we obtain the p-value, which we compare to the level of significance to draw a decision.

Several limitations of this approach are known. Like:

- There are cases where the likelihood principle is violated.
- The p-value offers evidence against H_0 (we are not allowed to say “accept H_0” but only “fail to reject”).

Inference for θ: Hypothesis Testing

We are interested in testing $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$

In frequentist based HT, we assume that H_0 is true and using the test statistics, $T(\mathbf{x})$, we obtain the p-value, which we compare to the level of significance to draw a decision.

Several limitations of this approach are known. Like:

- There are cases where the likelihood principle is violated.
- The p-value offers evidence against H_0 (we are not allowed to say “accept H_0” but only “fail to reject”).
- p-values do not have any interpretation as weight of evidence for H_0 (i.e. it is not the probability that H_0 is true).

Inference for θ: Hypothesis Testing

Within the Bayesian framework though, each of the hypotheses are simple subsets of the parameter space Θ and thus we can simply pick the hypothesis with the highest posterior coverage $p(H_i|\mathbf{x})$, where:

$$p(H_i|\mathbf{x}) = \frac{f(\mathbf{x}|H_i)p(H_i)}{f(\mathbf{x})}$$

Inference for θ: Hypothesis Testing

Within the Bayesian framework though, each of the hypotheses are simple subsets of the parameter space Θ and thus we can simply pick the hypothesis with the highest posterior coverage $p(H_i|\mathbf{x})$, where:

$$p(H_i|\mathbf{x}) = \frac{f(\mathbf{x}|H_i)p(H_i)}{f(\mathbf{x})}$$

Jeffreys proposed the use of Bayes Factor, which is the ratio of posterior to prior odds:

$$BF = \frac{p(H_0|\mathbf{x})/p(H_1|\mathbf{x})}{p(H_0)/p(H_1)}$$

where the smaller the BF the more the evidence against H_0

Inference for θ: Hypothesis Testing

From a decision theoretic approach one can derive the Bayes test. Assume that a_i denotes the action of accepting H_i. We make use of the generalized 0-1 loss function:

Inference for θ: Hypothesis Testing

From a decision theoretic approach one can derive the Bayes test. Assume that a_i denotes the action of accepting H_i. We make use of the generalized 0-1 loss function:

$$L(\theta, a_0) = \left\{ \begin{array}{ll} 0, & \theta \in \Theta_0 \\ c_{II}, & \theta \in \Theta_0^c \end{array} \right\}, \; L(\theta, a_1) = \left\{ \begin{array}{ll} c_I, & \theta \in \Theta_0 \\ 0, & \theta \in \Theta_0^c \end{array} \right\}$$

where $c_I(c_{II})$ is the cost of Type I (II) error.

Inference for θ: Hypothesis Testing

From a decision theoretic approach one can derive the Bayes test. Assume that a_i denotes the action of accepting H_i. We make use of the generalized 0-1 loss function:

$$L(\theta, a_0) = \left\{ \begin{array}{ll} 0, & \theta \in \Theta_0 \\ c_{II}, & \theta \in \Theta_0^c \end{array} \right\}, \; L(\theta, a_1) = \left\{ \begin{array}{ll} c_I, & \theta \in \Theta_0 \\ 0, & \theta \in \Theta_0^c \end{array} \right\}$$

where $c_I(c_{II})$ is the cost of Type I (II) error.

Then, the Bayes test (test with minimum Bayes risk) rejects H_0 if:

$$p(H_0|\mathbf{x}) < \frac{c_{II}}{c_I + c_{II}}$$

Predictive Inference

In some cases we are not interested about θ but we are concerned in drawing inference for future observable(s) y.

Predictive Inference

In some cases we are not interested about θ but we are concerned in drawing inference for future observable(s) y.

In the frequentist approach, usually we obtain and estimate of θ ($\hat{\theta}$) which we plug into the likelihood $(f(y|\hat{\theta}))$ and draw inference for the random future observable(s) y.

Predictive Inference

In some cases we are not interested about θ but we are concerned in drawing inference for future observable(s) y.

In the frequentist approach, usually we obtain and estimate of θ ($\hat{\theta}$) which we plug into the likelihood $(f(y|\hat{\theta}))$ and draw inference for the random future observable(s) y.

However, the above does not take into account the uncertainty in estimating θ by $\hat{\theta}$, leading (falsely) to shorter confidence intervals.

Predictive Inference

Within the Bayesian arena though, θ is a random variable and thus its effect can be integrated out leading to the predictive distribution:

Predictive Inference

Within the Bayesian arena though, θ is a random variable and thus its effect can be integrated out leading to the predictive distribution:

$$f(y|\mathbf{x}) = \int f(y|\theta)p(\theta|\mathbf{x})d\theta$$

Predictive Inference

Within the Bayesian arena though, θ is a random variable and thus its effect can be integrated out leading to the predictive distribution:

$$f(y|\mathbf{x}) = \int f(y|\theta)p(\theta|\mathbf{x})d\theta$$

The predictive distribution can be easily summarized to point/interval estimates and/or provide hypothesis testing for future observable(s) y.

Predictive Inference

Example:

We observe the data $f(x|\theta) \sim Binomial(n, \theta)$ and for the parameter θ we assume: $p(\theta) \sim Beta(\alpha, \beta)$. In the future we will obtain N more data points (independently of the first n) with Z referring to the future number of success $(Z = 0, 1, \ldots, N)$. What can be said about Z?

Predictive Inference

Example:

We observe the data $f(x|\theta) \sim Binomial(n, \theta)$ and for the parameter θ we assume: $p(\theta) \sim Beta(\alpha, \beta)$. In the future we will obtain N more data points (independently of the first n) with Z referring to the future number of success $(Z = 0, 1, \ldots, N)$. What can be said about Z?

$$\begin{aligned} p(\theta|x) &\propto f(x|\theta)p(\theta) \\ &\propto \left[\theta^{x}(1-\theta)^{n-x}\right]\left[\theta^{\alpha-1}(1-\theta)^{\beta-1}\right] \\ &= \theta^{\alpha+x-1}(1-\theta)^{n+\beta-x-1} \Rightarrow \\ \Rightarrow p(\theta|x) &\sim Beta(\alpha+x, \beta+n-x) \end{aligned}$$

Predictive Inference

$$
\begin{aligned}
f(z|x) &= \int f(z|\theta)p(\theta|x)d\theta = \\
&= \binom{N}{z}\frac{1}{Be(\alpha+x,\beta+n-x)} \times \\
&\times \int \theta^{\alpha+x-1}(1-\theta)^{n+\beta-x-1}\theta^{z}(1-\theta)^{N-z}d\theta \Rightarrow \\
\Rightarrow f(z|x) &= \binom{N}{z}\frac{Be(\alpha+x+z,\beta+n-x+N-z)}{Be(\alpha+x,\beta+n-x)}
\end{aligned}
$$

with $z = 0, 1, \ldots, N$.

Thus $Z|X$ is Beta-Binomial.

Example: Drugs on the job (cont.)

Recall: We have sampled $n = 100$ individuals and $y = 15$ tested positive for drug use.

Example: Drugs on the job (cont.)

Recall: We have sampled $n = 100$ individuals and $y = 15$ tested positive for drug use.

θ is the probability that someone in the population would have tested positive for drugs

Example: Drugs on the job (cont.)

Recall: We have sampled $n = 100$ individuals and $y = 15$ tested positive for drug use.

θ is the probability that someone in the population would have tested positive for drugs

We use the following prior: $\theta \sim Beta(a = 3.4, b = 23)$

Example: Drugs on the job (cont.)

Recall: We have sampled $n = 100$ individuals and $y = 15$ tested positive for drug use.

θ is the probability that someone in the population would have tested positive for drugs

We use the following prior: $\theta \sim Beta(a = 3.4, b = 23)$

The posterior is then
$\theta | y \sim Beta(y + a = 18.4, n - y + b = 108)$

Example: Drugs on the job (cont.)

Then consider a collection of 50 individuals who have just been selected for testing.

Example: Drugs on the job (cont.)

Then consider a collection of 50 individuals who have just been selected for testing.

We can let y_f be the number of drug users among these $n_f = 50$ and we can consider making inferences about y_f.

Example: Drugs on the job (cont.)

Then consider a collection of 50 individuals who have just been selected for testing.

We can let y_f be the number of drug users among these $n_f = 50$ and we can consider making inferences about y_f.

$y_f = 0, 1, \ldots, 50$

Example: Drugs on the job (cont.)

The predictive density of y_f is

$$\begin{aligned} p(y_f|y) &= \int p(y_f|\theta)p(\theta|y)d\theta = \\ &= \int p(y_f|\theta)Bin(y_f|50,\theta)Beta(\theta|18.4,108)d\theta = \\ &= \binom{50}{y_f}\frac{Be(18.4+y_f,108+50-y_f)}{Be(18.4,108)} \end{aligned}$$

Summary

Summary

Bayes Rocks!!!