An introduction to Bayesian Statistics

Çağrı Çöltekin

c.coltekin@rug.nl

Information science/Informatiekunde

2012-04-17

Informally...

- A frequentist is a person whose long-run ambition is to be wrong 5% of the time.
- A Bayesian is one who, vaguely expecting a horse, and catching a glimpse of a donkey, believes he has seen a mule.

Why should you (not) listen to this lecture?

– Probably you will not do any Bayesian analysis: when you want to publish your results you (typically) have to use 'the standard' *frequentist* statistics.

\+ But, maybe you will: Bayesian ideas are becoming more and more popular.

\+ By contrasting with the Bayesian approach you can better understand the frequentist approach.

\+ (Intellectual) curiosity.

\+ It's fun: this is a subject on which lots of disagreement is going on.

Statistical Inference

In most cases, we use statistics for *inference*: given a finite sample from a larger, potentially infinite, population we infer certain properties of the population.

- Height of people in a certain population based on a randomly selected group of individuals from the population.

Statistical Inference

In most cases, we use statistics for *inference*: given a finite sample from a larger, potentially infinite, population we infer certain properties of the population.

- Height of people in a certain population based on a randomly selected group of individuals from the population.
- Ratio of defective memory modules produced on a production line, based on the quality control results on the modules produced so far.

Statistical Inference

In most cases, we use statistics for *inference*: given a finite sample from a larger, potentially infinite, population we infer certain properties of the population.

- Height of people in a certain population based on a randomly selected group of individuals from the population.
- Ratio of defective memory modules produced on a production line, based on the quality control results on the modules produced so far.
- Average length of utterances in child-directed speech, and its relation with children's proficiency of the language

Statistical Inference

In most cases, we use statistics for *inference*: given a finite sample from a larger, potentially infinite, population we infer certain properties of the population.

- Height of people in a certain population based on a randomly selected group of individuals from the population.
- Ratio of defective memory modules produced on a production line, based on the quality control results on the modules produced so far.
- Average length of utterances in child-directed speech, and its relation with children's proficiency of the language

Statistical Inference

In most cases, we use statistics for *inference*: given a finite sample from a larger, potentially infinite, population we infer certain properties of the population.

- Height of people in a certain population based on a randomly selected group of individuals from the population.
- Ratio of defective memory modules produced on a production line, based on the quality control results on the modules produced so far.
- Average length of utterances in child-directed speech, and its relation with children's proficiency of the language

In most cases, we use a point estimate, but we also require a measure of the reliability of this estimate.

A hypothetical example for iference

Throughout this talk we will use a hypothetical example:

- We have a group of extraterrestrials (LGM), visiting Groningen.
- Among other things, we'd like to know the mean height of a LGM. (maybe just curiosity, maybe we have a business idea).
- We managed to measure height of all 10 LGM we know, the data is as follows (in centimeters):
  `122 122 116 134 113 114 113 110 123 130`
- No one knows the population mean in their planet.
- Interestingly, they know that a reliable estimate of the standard deviation of the complete LGM population is 8cm.

Inference for unknown mean: confidence intervals

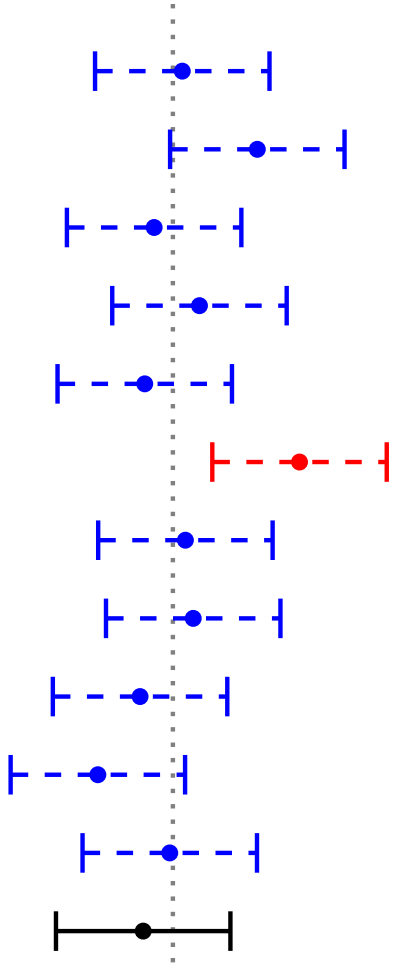
- Sample mean is 118.1 which is our best estimate of the population mean.
- Using our estimate, we calculate a confidence interval

$$\left[\mu - t \times \frac{\sigma}{\sqrt{n}}, \mu + t \times \frac{\sigma}{\sqrt{n}}\right]$$

  where t is the critical value of interest from the t-distribution.
- Here is how to calculate 95% confidence interval in R

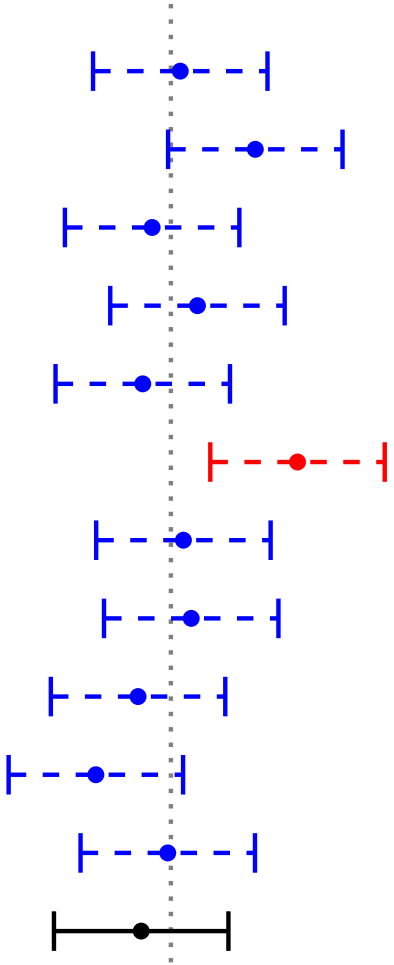
```
> mean(lgm) + qt(.025, df=9) * (8/sqrt(10))
[1] 112.3771
> mean(lgm) + qt(.925, df=9) * (8/sqrt(10))
[1] 122.0813
```

Another look at the confidence intervals



- Population mean is a value that is fixed, unknown, and unknowable.
- We calculate the 95% confidence interval using the only sample mean we have.
- If we had many similar samples from the same population, and calculate the same interval for each, we would have 95% of them including the population mean.
- As a result, we say that we are 95% confident that the interval we calculated contains the sample mean.

Another look at the confidence intervals



- Population mean is a value that is fixed, unknown, and unknowable.
- We calculate the 95% confidence interval using the only sample mean we have.
- If we had many similar samples from the same population, and calculate the same interval for each, we would have 95% of them including the population mean.
- As a result, we say that we are 95% confident that the interval we calculated contains the sample mean.

Does it mean: 'with 0.95 probability the population mean is in the interval we calculated'?

Outline

Probability theory: all you need to know

Three axioms of probability (also called, Kolmogorov axioms).

$P(E) \geq 0$ Probability of any event E is a positive real number.

$P(\Omega) = 1$ Sum of the probabilities of all outcomes is 1.

$P(\cup E_i) = \sum P(E_i)$ For disjoint events E_i, the probability that any of the events happens is the sum of the probabilities individual events.

Probability theory: all you need to know

Three axioms of probability (also called, Kolmogorov axioms).

$P(E) \geq 0$ Probability of any event E is a positive real number.

$P(\Omega) = 1$ Sum of the probabilities of all outcomes is 1.

$P(\cup E_i) = \sum P(E_i)$ For disjoint events E_i, the probability that any of the events happens is the sum of the probabilities individual events.

But probability theory does not tell how to determine probability of an event, say the probability of 'heads' on a coin flip.

Where do probabilities come from?

There are many ways of assigning probabilities events. However, we are interested in two:

Frequentist Probabilities are Long-run frequencies. Probability of an event (e.g., coin-flip resulting in 'heads') is determined by its long-run frequency.

Bayesian Probabilities are degrees of belief (probability you assigned to a coin-flip experiment is your belief that it is a fair coin).

Bayes' theorem

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

- Bayes' theorem gives us a way to calculate conditional probabilities.

Bayes' theorem

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

- Bayes' theorem gives us a way to calculate conditional probabilities.
- Bayes' theorem follows from axioms of probability: there is nothing controversial about it.

Bayes' theorem

$$P(H|Data) = \frac{P(H)P(Data|H)}{P(Data)}$$

- Bayes' theorem gives us a way to calculate conditional probabilities.
- Bayes' theorem follows from axioms of probability: there is nothing controversial about it.
- ... but in Bayesian inference, we use it for calculating probability of a hypothesis, H,

Bayes' theorem

$$P(\theta|Data) = \frac{P(\theta)P(Data|\theta)}{P(Data)}$$

- Bayes' theorem gives us a way to calculate conditional probabilities.
- Bayes' theorem follows from axioms of probability: there is nothing controversial about it.
- ... but in Bayesian inference, we use it for calculating probability of a hypothesis, H,
- ... or a certain parameter of a population, θ, given a sample ($Data$).

Bayes' theorem: some terminology

$$p(\theta|Data) = \frac{p(\theta)p(Data|\theta)}{p(Data)}$$

$p(\theta\|Data)$:	*posterior*
$p(Data\|\theta)$:	*likelihood* ($\mathcal{L}(\theta)$)
$p(\theta)$:	*prior*
$p(Data)$:	Marginal probability of data $\sum P(Data\|\theta)P(\theta)$ or $\int p(Data\|\theta)p(\theta)d\theta$

Bayes' theorem: some terminology

$$p(\theta|Data) = \frac{p(\theta)p(Data|\theta)}{p(Data)}$$

$p(\theta\|Data)$:	*posterior*
$p(Data\|\theta)$:	*likelihood* $(\mathcal{L}(\theta))$
$p(\theta)$:	*prior*
$p(Data)$:	Marginal probability of data $\sum P(Data\|\theta)P(\theta)$ or $\int p(Data\|\theta)p(\theta)d\theta$

$$posterior \propto prior \times likelihood$$

Bayes' theorem: an example

- We have a particular medical test, T, for diagnosing a condition, A. T has the following properties: $P(T^+|A) = 0.99$ and $P(T^+|\textbf{not } A) = 0.02$.

Bayes' theorem: an example

- We have a particular medical test, T, for diagnosing a condition, A. T has the following properties: $P(T^+|A) = 0.99$ and $P(T^+|\textbf{not } A) = 0.02$.
- A patient had a positive T result. Can you tell the probability that he has the condition A, $P(A|T^+)$?

Bayes' theorem: an example

- We have a particular medical test, T, for diagnosing a condition, A. T has the following properties: $P(T^+|A) = 0.99$ and $P(T^+|\textbf{not } A) = 0.02$.
- A patient had a positive T result. Can you tell the probability that he has the condition A, $P(A|T^+)$?
- Not yet, we need one more probability: $P(A)$. Let's assume that $P(A) = 0.0002$.

Bayes' theorem: an example

- We have a particular medical test, T, for diagnosing a condition, A. T has the following properties: $P(T^+|A) = 0.99$ and $P(T^+|\textbf{not } A) = 0.02$.
- A patient had a positive T result. Can you tell the probability that he has the condition A, $P(A|T^+)$?
- Not yet, we need one more probability: $P(A)$. Let's assume that $P(A) = 0.0002$.

Bayes' theorem: an example

- We have a particular medical test, T, for diagnosing a condition, A. T has the following properties: $P(T^+|A) = 0.99$ and $P(T^+|\textbf{not } A) = 0.02$.
- A patient had a positive T result. Can you tell the probability that he has the condition A, $P(A|T^+)$?
- Not yet, we need one more probability: $P(A)$. Let's assume that $P(A) = 0.0002$.

$$P(A|T^+) = \frac{P(A)P(T^+|A)}{P(T^+)}$$

Bayes' theorem: an example

- We have a particular medical test, T, for diagnosing a condition, A. T has the following properties: $P(T^+|A) = 0.99$ and $P(T^+|\textbf{not } A) = 0.02$.
- A patient had a positive T result. Can you tell the probability that he has the condition A, $P(A|T^+)$?
- Not yet, we need one more probability: $P(A)$. Let's assume that $P(A) = 0.0002$.

$$\begin{aligned} P(A|T^+) &= \frac{P(A)P(T^+|A)}{P(T^+)} \\ &= \frac{0.0002 \times 0.99}{0.0002 \times 0.99 + 0.9998 \times 0.02} \end{aligned}$$

Bayes' theorem: an example

- We have a particular medical test, T, for diagnosing a condition, A. T has the following properties: $P(T^+|A) = 0.99$ and $P(T^+|\textbf{not } A) = 0.02$.
- A patient had a positive T result. Can you tell the probability that he has the condition A, $P(A|T^+)$?
- Not yet, we need one more probability: $P(A)$. Let's assume that $P(A) = 0.0002$.

$$\begin{aligned} P(A|T^+) &= \frac{P(A)P(T^+|A)}{P(T^+)} \\ &= \frac{0.0002 \times 0.99}{0.0002 \times 0.99 + 0.9998 \times 0.02} \\ &= 0.0098 \end{aligned}$$

Bayes' theorem: an example

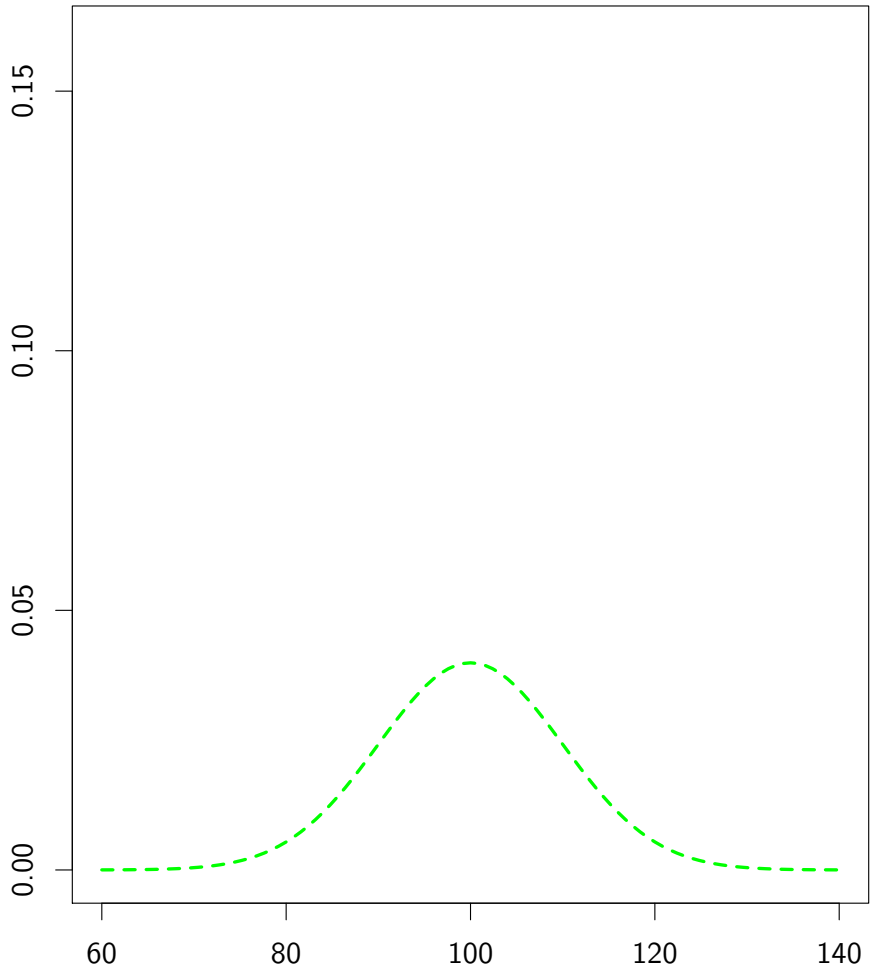
- We have a particular medical test, T, for diagnosing a condition, A. T has the following properties: $P(T^+|A) = 0.99$ and $P(T^+|\textbf{not } A) = 0.02$.
- A patient had a positive T result. Can you tell the probability that he has the condition A, $P(A|T^+)$?
- Not yet, we need one more probability: $P(A)$. Let's assume that $P(A) = 0.0002$.

$$\begin{aligned} P(A|T^+) &= \frac{P(A)P(T^+|A)}{P(T^+)} \\ &= \frac{0.0002 \times 0.99}{0.0002 \times 0.99 + 0.9998 \times 0.02} \\ &= 0.0098 \end{aligned}$$

Prior knowledge is important.

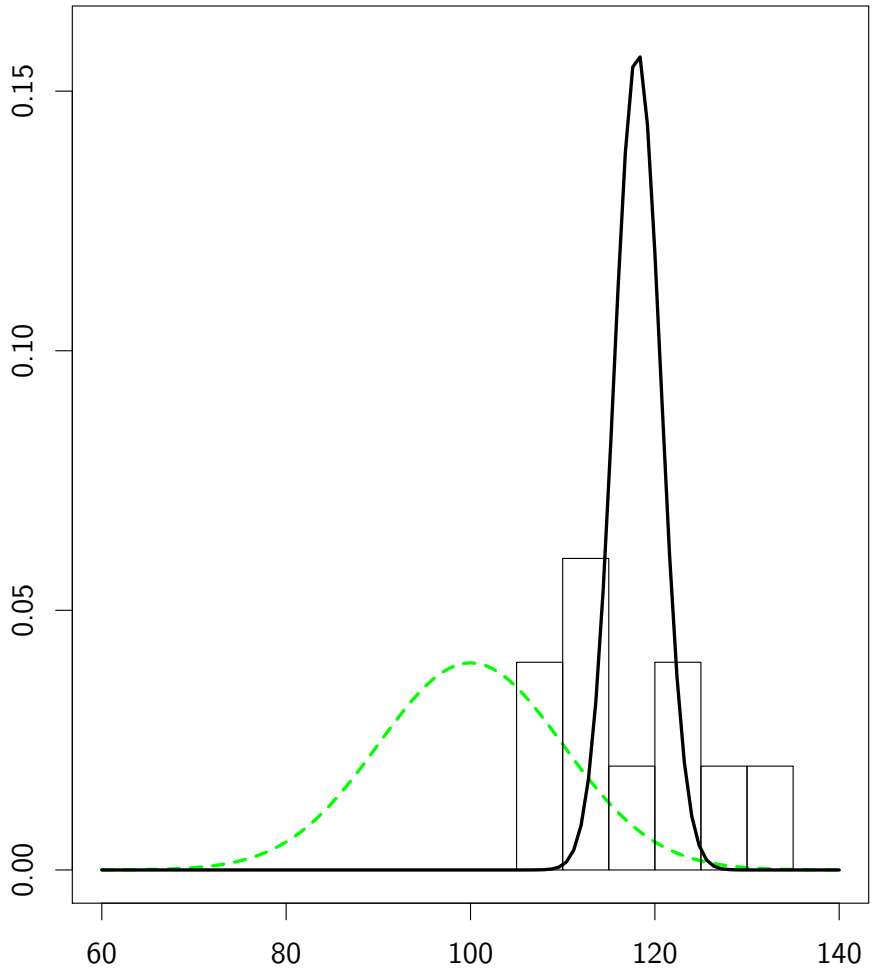
Infering mean height of LGM: the/a Bayesian way

- We start with a Gaussian prior $x \sim N(100, 100)$



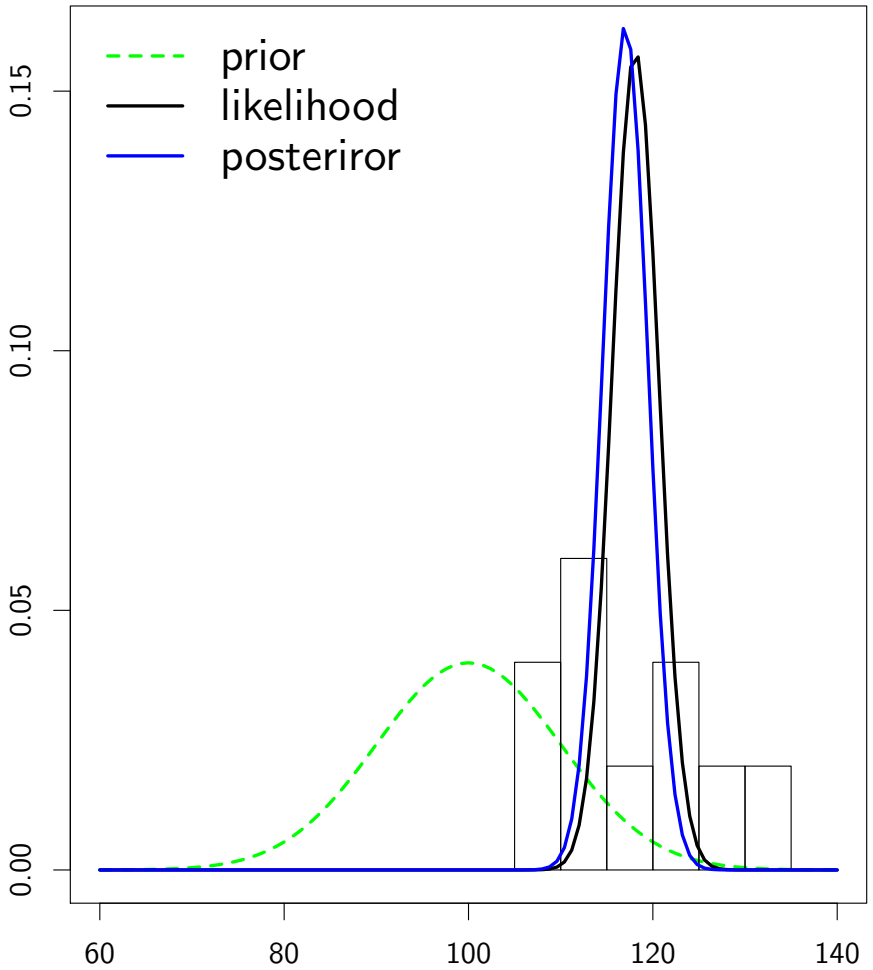
Infering mean height of LGM: the/a Bayesian way

- We start with a Gaussian prior $x \sim N(100, 100)$
- We observe the data. Likelihood is $N(\bar{x}, \sigma^2/n) = N(118.1, 6.4)$.



Infering mean height of LGM: the/a Bayesian way

- We start with a Gaussian prior $x \sim N(100, 100)$
- We observe the data. Likelihood is $N(\bar{x}, \sigma^2/n) = N(118.1, 6.4)$.
- Posterior is again normal $N(117.01, 6.02)$.



Infering mean height of LGM: calculations

Population	$N(\mu, \sigma^2)$
Prior	$N(m, s^2)$
Likelihood	$N(\bar{x}, \sigma^2/n)$
Posterior	$N(m', (s')^2)$

Infering mean height of LGM: calculations

Population	$N(\mu, \sigma^2)$
Prior	$N(m, s^2)$
Likelihood	$N(\bar{x}, \sigma^2/n)$
Posterior	$N(m', (s')^2)$

$$
\begin{aligned}
(s')^2 &= \frac{\sigma^2 \times s^2}{\sigma^2 + ns^2} \\
m' &= \frac{1/s^2}{n/\sigma^2 + 1/s^2} m \\
&+ \frac{n/\sigma^2}{n/\sigma^2 + 1/s^2} \bar{x}
\end{aligned}
$$

Infering mean height of LGM: calculations

Population	$N(\mu, \sigma^2)$
Prior	$N(m, s^2)$
Likelihood	$N(\bar{x}, \sigma^2/n)$
Posterior	$N(m', (s')^2)$

$$
\begin{aligned}
(s')^2 &= \frac{\sigma^2 \times s^2}{\sigma^2 + n s^2} \\
m' &= \frac{1/s^2}{n/\sigma^2 + 1/s^2} m \\
&+ \frac{n/\sigma^2}{n/\sigma^2 + 1/s^2} \bar{x}
\end{aligned}
$$

$$
\begin{aligned}
(s')^2 &= \frac{8^2 \times 10^2}{8^2 + 10 \times 10^2} = 6.02 \\
m' &= \frac{1/8^2}{10/8^2 + 1/10^2} 100 \\
&+ \frac{10/8^2}{10/8^2 + 1/10^2} 118.1 \\
&= 117.01
\end{aligned}
$$

Infering mean height of LGM: calculations

Population	$N(\mu, \sigma^2)$
Prior	$N(m, s^2)$
Likelihood	$N(\bar{x}, \sigma^2/n)$
Posterior	$N(m', (s')^2)$

$$
\begin{aligned}
(s')^2 &= \frac{\sigma^2 \times s^2}{\sigma^2 + n s^2} \\
m' &= \frac{1/s^2}{n/\sigma^2 + 1/s^2} m \\
&+ \frac{n/\sigma^2}{n/\sigma^2 + 1/s^2} \bar{x}
\end{aligned}
$$

$$
\begin{aligned}
(s')^2 &= \frac{8^2 \times 10^2}{8^2 + 10 \times 10^2} = 6.02 \\
m' &= \frac{1/8^2}{10/8^2 + 1/10^2} 100 \\
&+ \frac{10/8^2}{10/8^2 + 1/10^2} 118.1 \\
&= 117.01
\end{aligned}
$$

- Our posterior is N(117.01, 6.02), a normal distribution with mean 117.01cm, and variance 6.02.

... but, where is my confidence inteval or p-value?

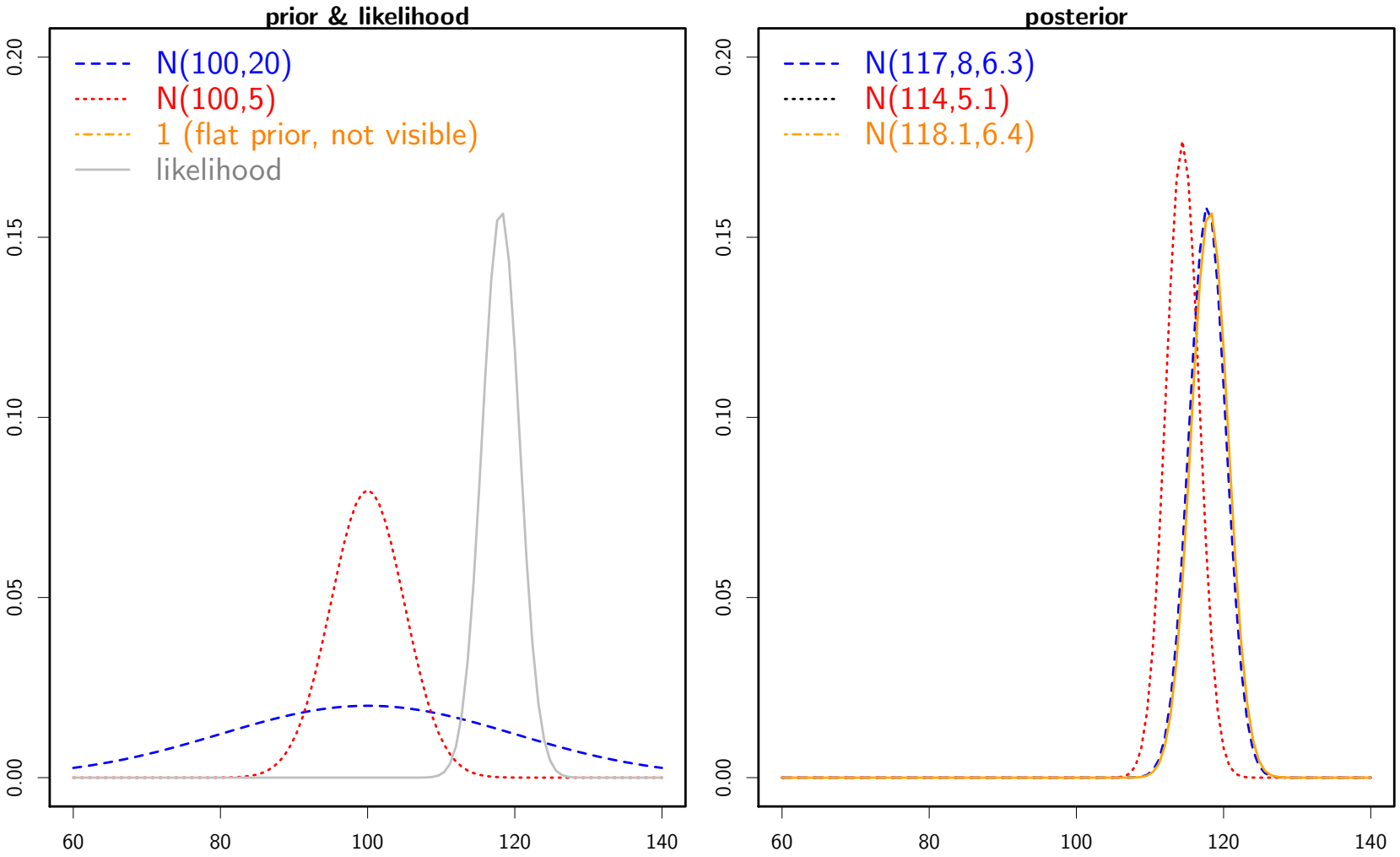
- Bayesian posterior contains all the information you need:
  - m': gives expected value you belief should be centered on.
  - s': gives the variability of your estimate.
- If you like, analogous to confidence intervals, you can calculate the credible intervals.
- 95% credible interval for our LGM example:

```
> qnorm(0.025, 117.01, sqrt(6.02))
[1] 112.20
> qnorm(0.925, 117.01, sqrt(6.02))
[1] 120.54
```

(our frequentist confidence interval was [112.38,122.08])

- Note that we can now say 'with .95 probability population mean is in range [112.20,120.54]'.

Effect of different priors



How do we choose the priors?

- Quantifying expert opinion:
  - Pick the prior mean where expert's belief is centered on.
  - Pick the variance such that the range covers reasonable values of the expected mean, or consider equivalent sample size for your prior. If you want your prior to be equivalent to a sample of size n, your prior variance should be σ^2/n.
- Previous research results. You can use previously reported research results as your prior. Bayesian inference can proceed incrementally.
- Non-informative, flat priors. If there is no reasonable way to form a prior, one can use a flat prior. These often result in similar point estimates as frequentist estimates.

Prior distributions and computation

Bayesian calculations can be difficult. There are two major methods in practice:

- When possible, use conjugate priors, which allow easy calculations. For example, if our likelihood is normal, the conjugate prior is also normal. That's why we could easily compute the posterior distribution for mean LGM height.
- When we cannot compute the solution analytically, we can use methods such as Markov Chain Monte Carlo (MCMC) sampling.

How about hypohesis testing?

- Traditional hypothesis testing is based on specifying a null hypothesis, and rejecting it on the basis of evidence.
- The Bayesian approach is to use posterior odds:

$$\frac{p(H_1|Data)}{P(H_0|Data)}$$

gives you 'which hypothesis to bet for'.

Summary

- Bayesian statistics is another, mathematically more principled, approach to statistical inference.

Summary

- Bayesian statistics is another, mathematically more principled, approach to statistical inference.
- The main difference is about interpretation of probability: probabilities are degreed of belief (as opposed to long-run relative frequencies).

Summary

- Bayesian statistics is another, mathematically more principled, approach to statistical inference.
- The main difference is about interpretation of probability: probabilities are degreed of belief (as opposed to long-run relative frequencies).
- Bayesian inference is based on observed data $P(\theta|Data)$, not based on unobserved data (as in frequentist inference, $P(Data|\theta)$)

Summary

- Bayesian statistics is another, mathematically more principled, approach to statistical inference.
- The main difference is about interpretation of probability: probabilities are degreed of belief (as opposed to long-run relative frequencies).
- Bayesian inference is based on observed data $P(\theta|Data)$, not based on unobserved data (as in frequentist inference, $P(Data|\theta)$)
- Bayesian statistics incorporate prior knowledge.

Summary

- Bayesian statistics is another, mathematically more principled, approach to statistical inference.
- The main difference is about interpretation of probability: probabilities are degreed of belief (as opposed to long-run relative frequencies).
- Bayesian inference is based on observed data $P(\theta|Data)$, not based on unobserved data (as in frequentist inference, $P(Data|\theta)$)
- Bayesian statistics incorporate prior knowledge.
- Posterior probability includes all information you need about the quantity you are interested in after observing the data.

Summary

- Bayesian statistics is another, mathematically more principled, approach to statistical inference.
- The main difference is about interpretation of probability: probabilities are degreed of belief (as opposed to long-run relative frequencies).
- Bayesian inference is based on observed data $P(\theta|Data)$, not based on unobserved data (as in frequentist inference, $P(Data|\theta)$)
- Bayesian statistics incorporate prior knowledge.
- Posterior probability includes all information you need about the quantity you are interested in after observing the data.
- The computation can be difficult, but with new methods and technology, it is far more feasible now.