A Gentle Introduction to Bayesian Estimation

Day 1: Introduction

Sara van Erp (s.j.vanerp@uu.nl)

These slides are in part based on the original slides from Rens van de Schoot, available here.

Tea experiment

Please have some tea and write down what was poured first: the (oat) milk or the tea?

Overview

- **Day 1**: Conceptual introduction
- **Day 2**: WAMBS-checklist (When to worry and how to Avoid the Misuse of Bayesian Statistics)
- **Day 3**: Algorithms and checks
- **Day 4**: Priors: Cautionary tails and possibilities
- **Day 5**: Informative priors

Daily schedule

09:00-12:00 Lecture
12:00-13:00 Lunch
13:00-16:00 Computer lab

Note: During the computer labs, you will work on the exercises yourself but we will check in regularly.

Feel free to ask questions throughout the lectures and labs, also on your own applications (see also lab Friday).

IOPS participants have an additional hand-in assignment about analyzing your own data

Teachers

Sara van Erp

Duco Veen

Florian Metwaly

Course website

https://utrechtuniversity.github.io/BayesianEstimation/

Software

We use brms in R and online applications

nature reviews methods primers

View all journals Search My Account

Explore content Journal information Publish with us Sign up for alerts RSS feed

nature › nature reviews methods primers › primers › article › table

Table 2 A non-exhaustive summary of commonly used and open Bayesian software programs

From: Bayesian statistics and modelling

Software package	Summary
General-purpose Bayesian inference software	
BUGS[231,232]	The original general-purpose Bayesian inference engine, in different incarnations. These use Gibbs and Metropolis sampling. Windows-based software (WinBUGS[233]) with a user-specified model and a black-box MCMC algorithm. Developments include an open-source version (OpenBUGS[234]) also available on Linux and Mac
JAGS[235]	An open-source variation of BUGS that can run cross-platform and can run from R via rjags[236]
PyMC3[237]	An open-source framework for Bayesian modelling and inference entirely within Python; includes Gibbs sampling and Hamiltonian Monte Carlo
Stan[98]	An open-source, general-purpose Bayesian inference engine using Hamiltonian Monte Carlo; can be run from R, Python, Julia, MATLAB and Stata
NIMBLE[238]	Generalization of the BUGS language in R; includes sequential Monte Carlo as well as MCMC. Open-source R package using BUGS/JAGS-model language to develop a model; different algorithms for model fitting including MCMC and sequential Monte Carlo approaches. Includes the ability to write novel algorithms
Programming languages that can be used for Bayesian inference	
TensorFlow Probability[239,240]	A Python library for probabilistic modelling built on Tensorflow[203] from Google
Pyro[241]	A probabilistic programming language built on Python and PyTorch[204]
Julia[242]	A general-purpose language for mathematical computation. In addition to Stan, numerous other probabilistic programming libraries are available for the Julia programming language, including Turing.jl[243] and Mamba.jl[244]
Specialized software doing Bayesian inference for particular classes of models	
JASP[245]	A user-friendly. higher-level interface offering Bayesian analysis. Open source and relies on a collection of open-source R packages
R-INLA[230]	An open-source R package for implementing INLA[246]. Fast inference in R for a certain set of hierarchical models using nested Laplace approximations
GPstuff[247]	Fast approximate Bayesian inference for Gaussian processes using expectation propagation; runs in MATLAB, Octave and R

Why this course?

"... It is clear that it is not possible to think about learning from experience and acting on it without coming to terms with Bayes' theorem."

- Jerome Cornfield (in de Finetti, 1974a)

"…whereas the 20th century was dominated by NHST [null hypothesis significance testing], the 21st century is becoming Bayesian…"

- Kruschke (2011, p.272) in a special 'Bayesian' issue of Perspectives on Psychological Science

"... over the last few decades, it has become the major approach in the field of statistics, and has come to be accepted in many or most of the physical, biological and human sciences ..."

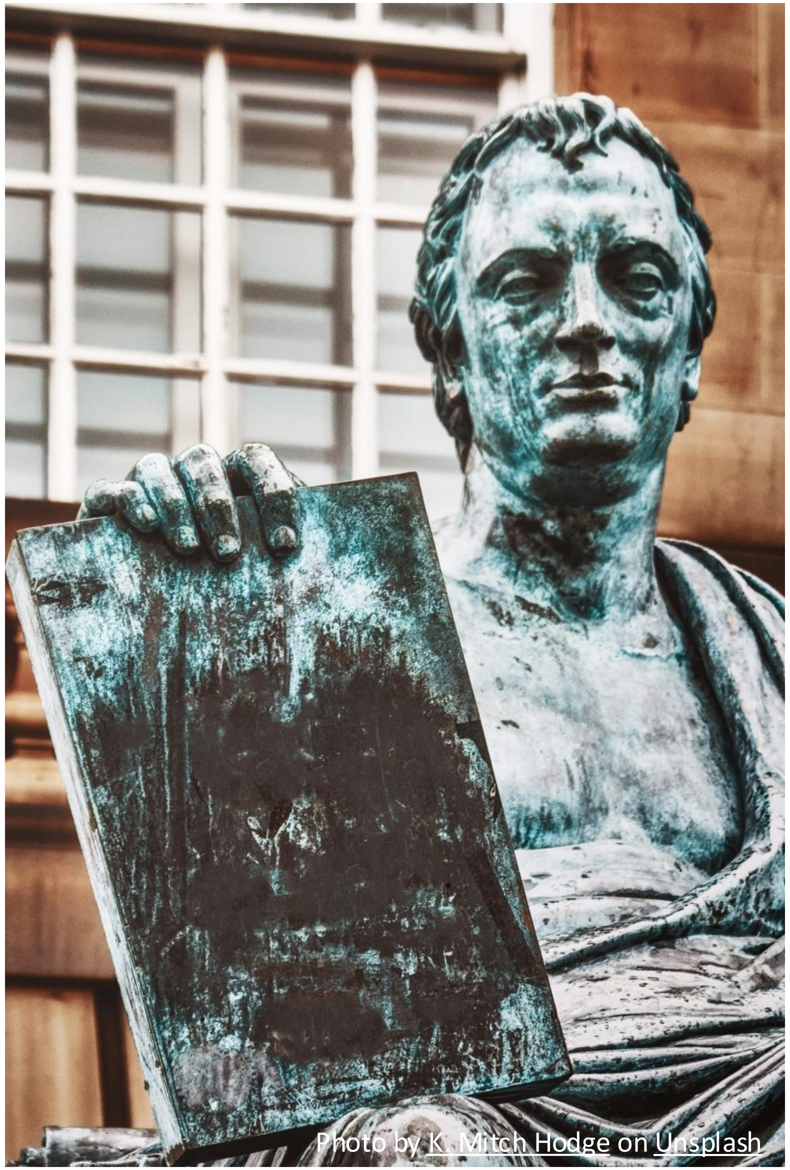
- Lee (2011, p1)

Why are you taking this course?

A brief history of Bayes

It all started...

In 1748 when Hume published an essay about uncertainty



A brief history of Bayes

This essay inspired Thomas Bayes (1701-1761) who was enrolled at the University of Edinburgh to study logic and theology.

He worked on the question whether God exists using Inverse Probability, but he never published any work on this topic

A brief history of Bayes

After T. Bayes passed, his relatives asked Richard Price (1723-1791) to go through his unfinished work and it was Price who discovered the essay on inverse probability.

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

Dear Sir,

Read Dec. 23, 1763.

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

He had, you know, the honour of being a member of that illuſtrious Society, and was much eſteemed by many in it as a very able mathematician. In an introduction which he has writ to this Eſſay, he ſays, that his deſign at firſt in thinking on the ſubject of it was, to find out a method by which we might judge concerning the probability that an event has to happen, in given circumſtances, upon ſuppoſition that we know nothing concerning it but that, under the ſame

circum-

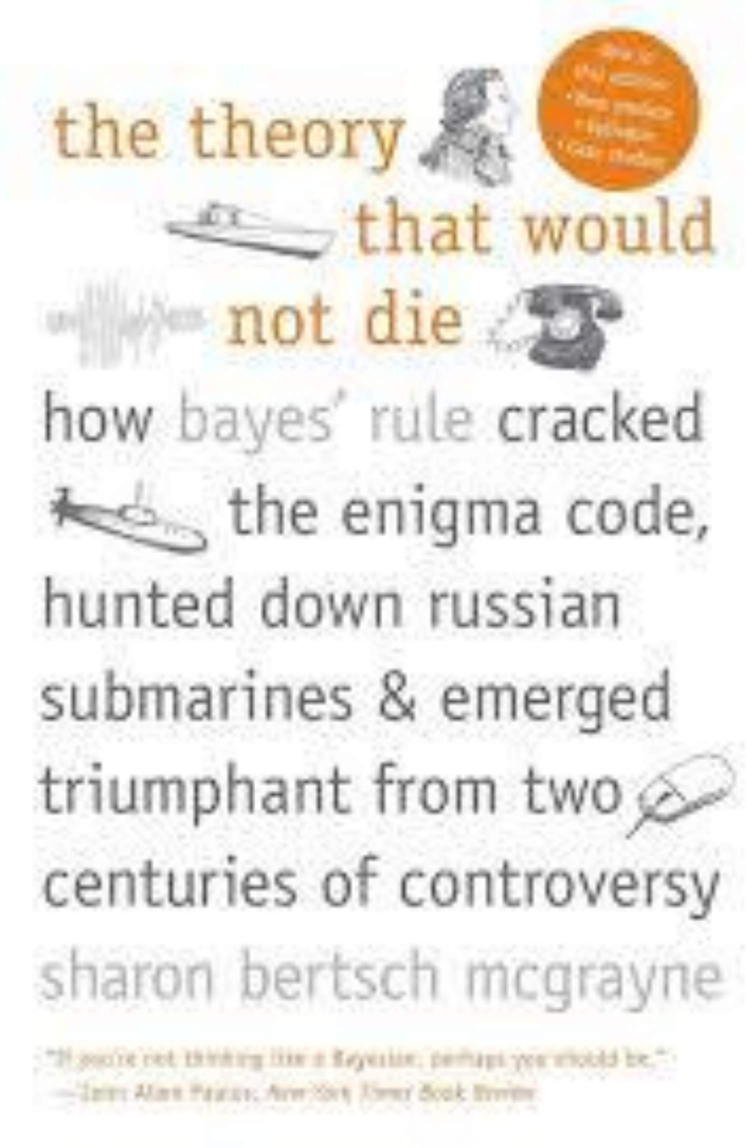
circumſtances, it has happened a certain number of times, and failed a certain other number of times. He adds, that he ſoon perceived that it would not be very difficult to do this, provided ſome rule could be found according to which we ought to eſtimate the chance that the probability for the happening of an event perfectly unknown, ſhould lie between any two named degrees of probability, antecedently to any experiments made about it; and that it appeared to him that the rule muſt be to ſuppoſe the chance the ſame that it ſhould lie between any two equidifferent degrees; which, if it were allowed, all the reſt might be eaſily calculated in the common method of proceeding in the doctrine of chances. Accordingly, I find among his papers a very ingenious ſolution of this problem in this way. But he afterwards conſidered, that the *poſtulate* on which he had argued might not perhaps be looked upon by all as reaſonable; and therefore he choſe to lay down in another form the propoſition in which he thought the ſolution of the problem is contained, and in a *ſcholium* to ſubjoin the reaſons why he thought ſo, rather than to take into his mathematical reaſoning any thing that might admit diſpute. This, you will obſerve, is the method which he has purſued in this eſſay.

Every judicious perſon will be ſenſible that the

A brief history of Bayesian statistics

Pierre Simon Laplace (1749-1827) independently discovered the same theorem and published the formula we now know as Bayes’ rule.

More history on Bayesian statistics



Bayes' rule

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

where H = Hypothesis and E = Evidence (i.e., the data)

Bayes’ rule

Ignoring the denominator $P(e)$ for now, we get:

$$posterior \propto likelihood \times prior$$

Classical frequentist statistics relies only on the likelihood.

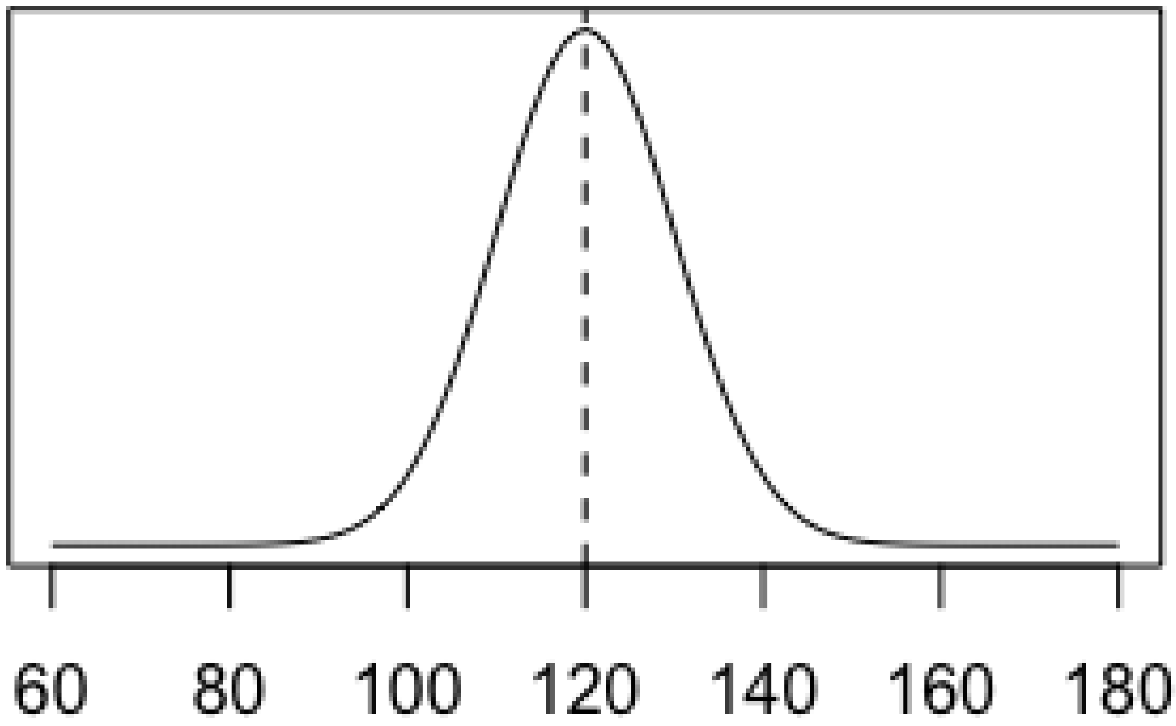
The likelihood

Suppose we want to know the average IQ in the general population.

We use a convenience sample of university students and measure their IQ.

This information can be found in the likelihood function:

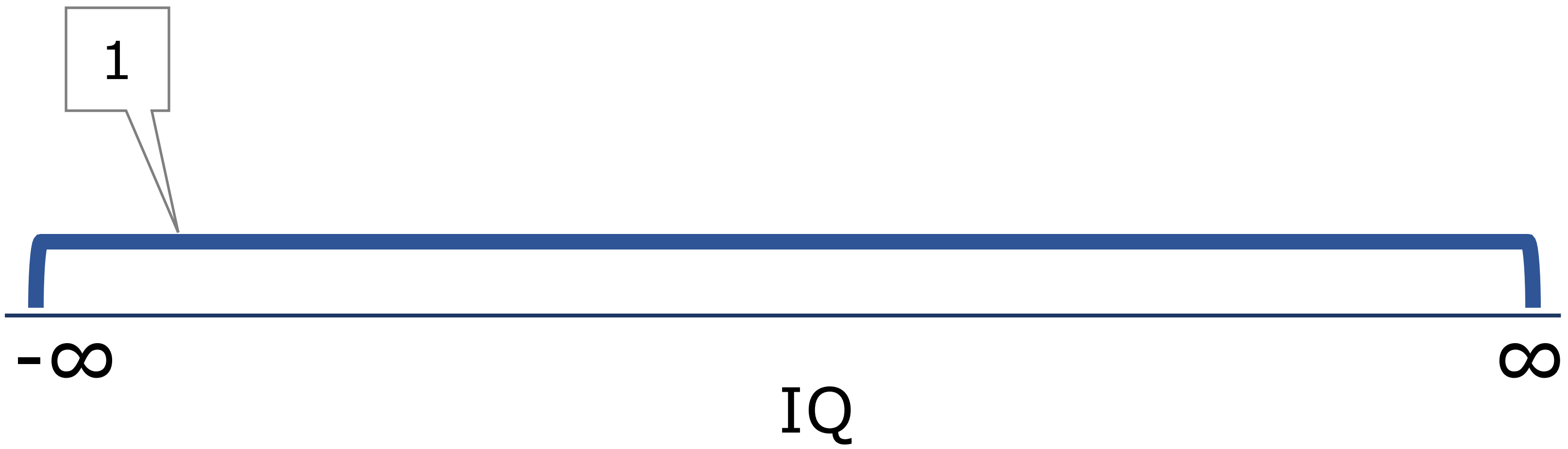
- y-axis = the likelihood or: how likely are the observed data given specific IQ values?



Bayesian statistics combines likelihood & prior

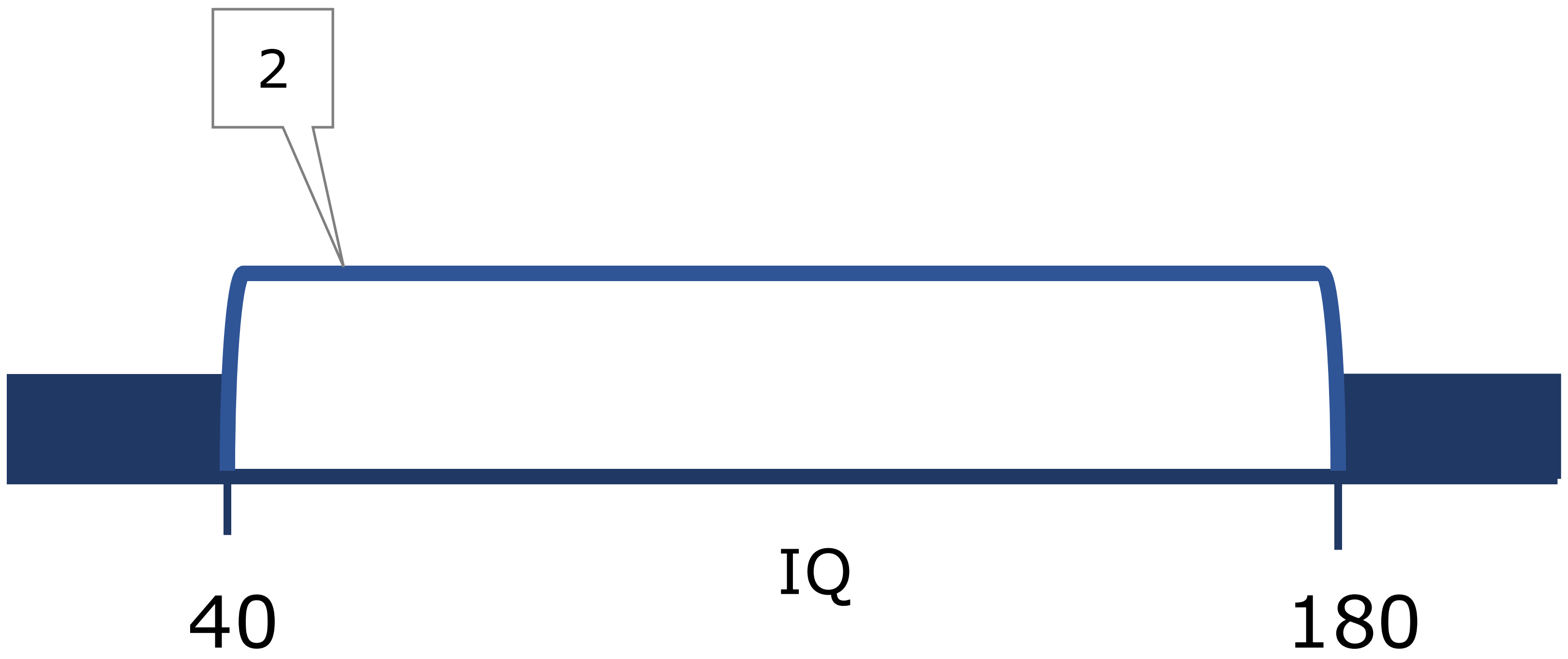
- A prior is a probability distribution containing information about your parameters *before* you collect the data.
- Prior information can come from different sources, such as previous research, expert knowledge, knowledge about the parameters (see day 5)
- Priors vary in their informativeness
- Some software programs rely on “default” prior distributions (see day 4)

Estimating the average IQ: Prior knowledge

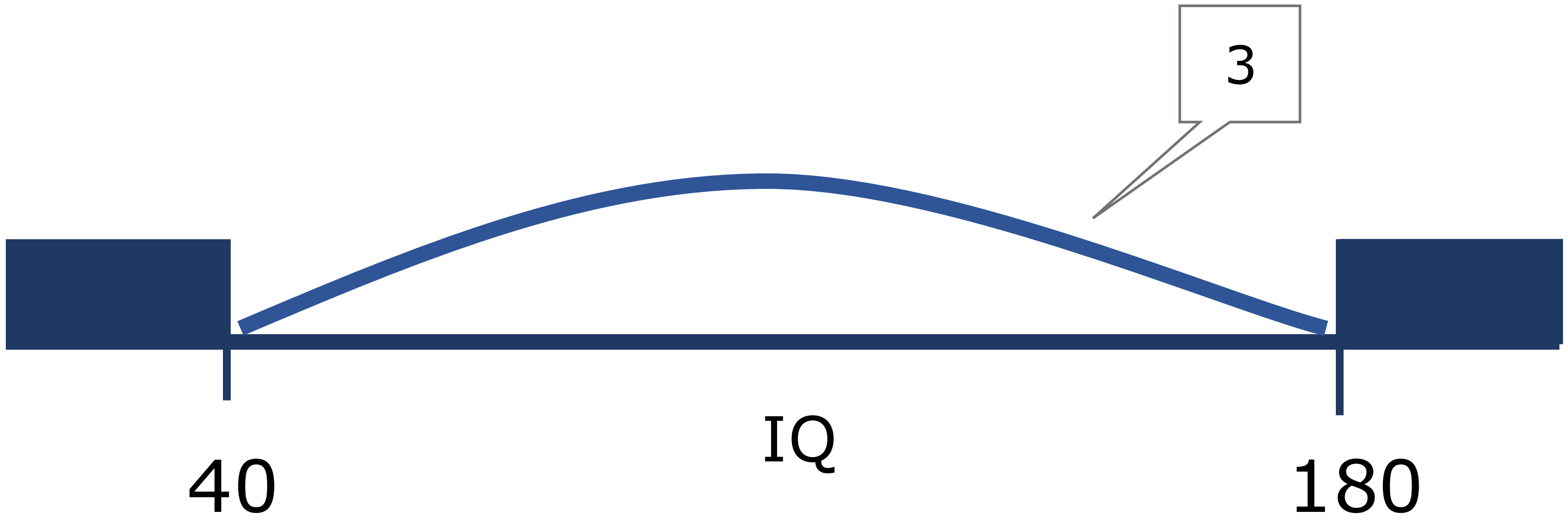


From: Van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & Van Aken, M. A. (2014). A gentle introduction to Bayesian analysis: Applications to developmental research. *Child development*, *85*(3), 842-860.

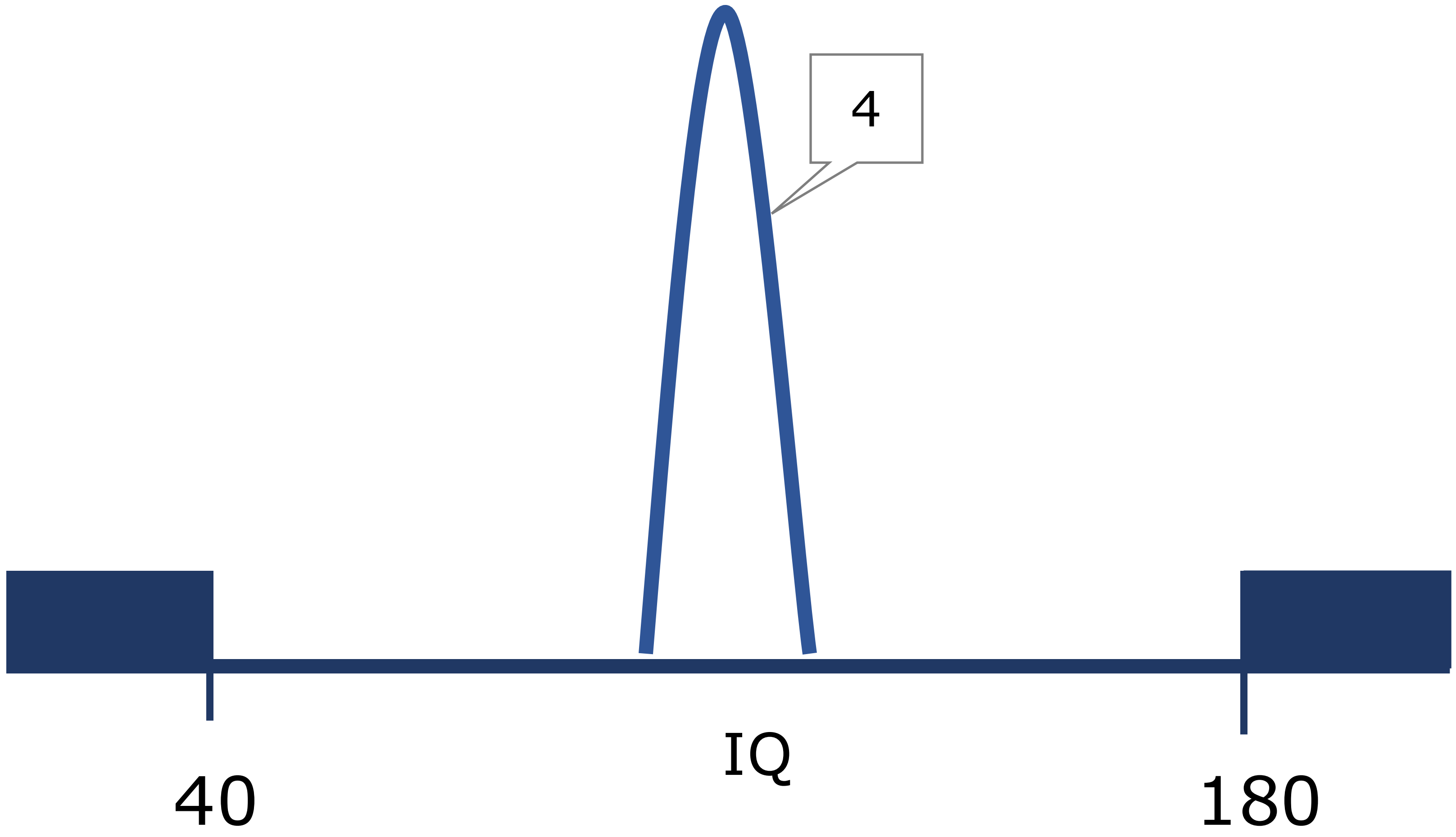
Estimating the average IQ: Prior knowledge



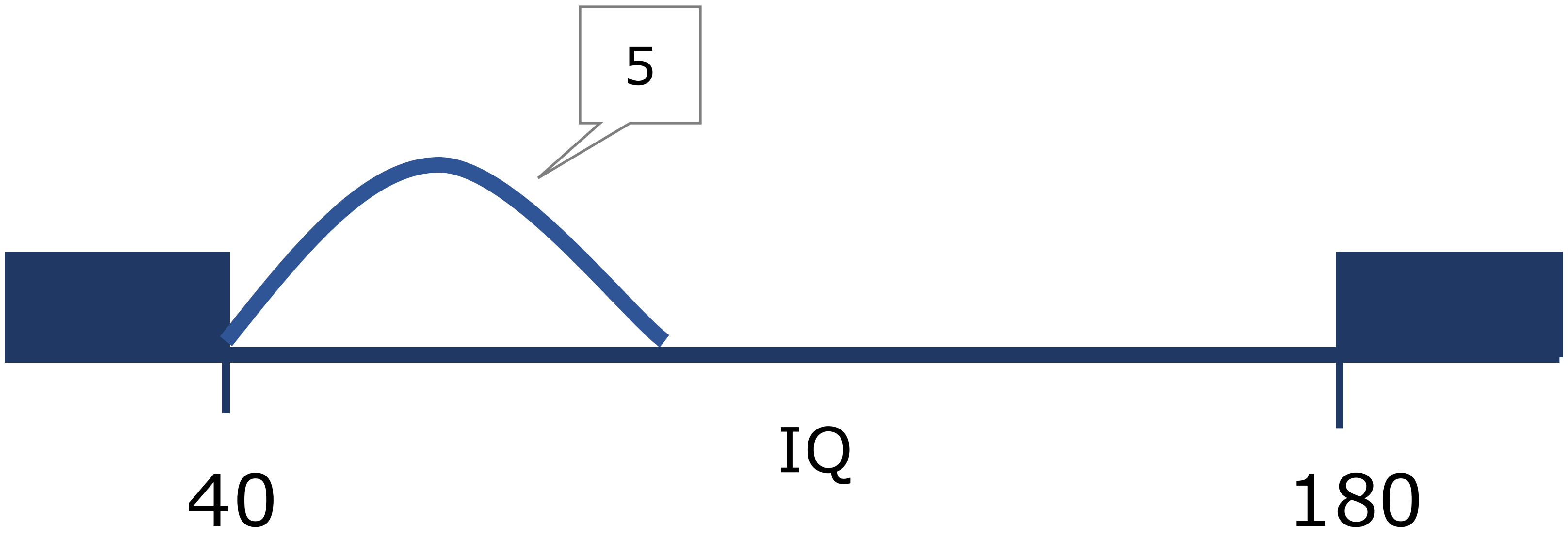
Estimating the average IQ: Prior knowledge

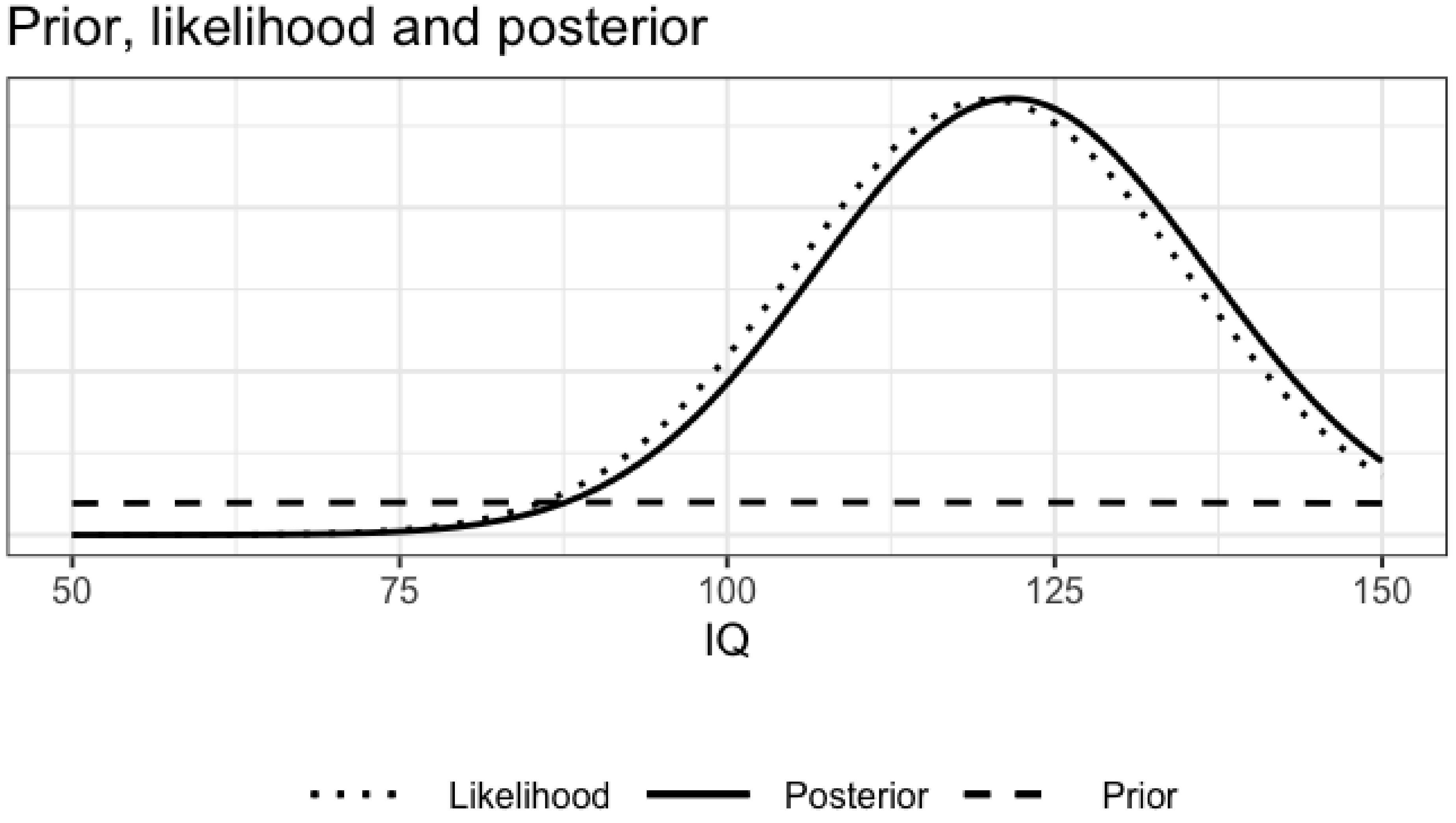


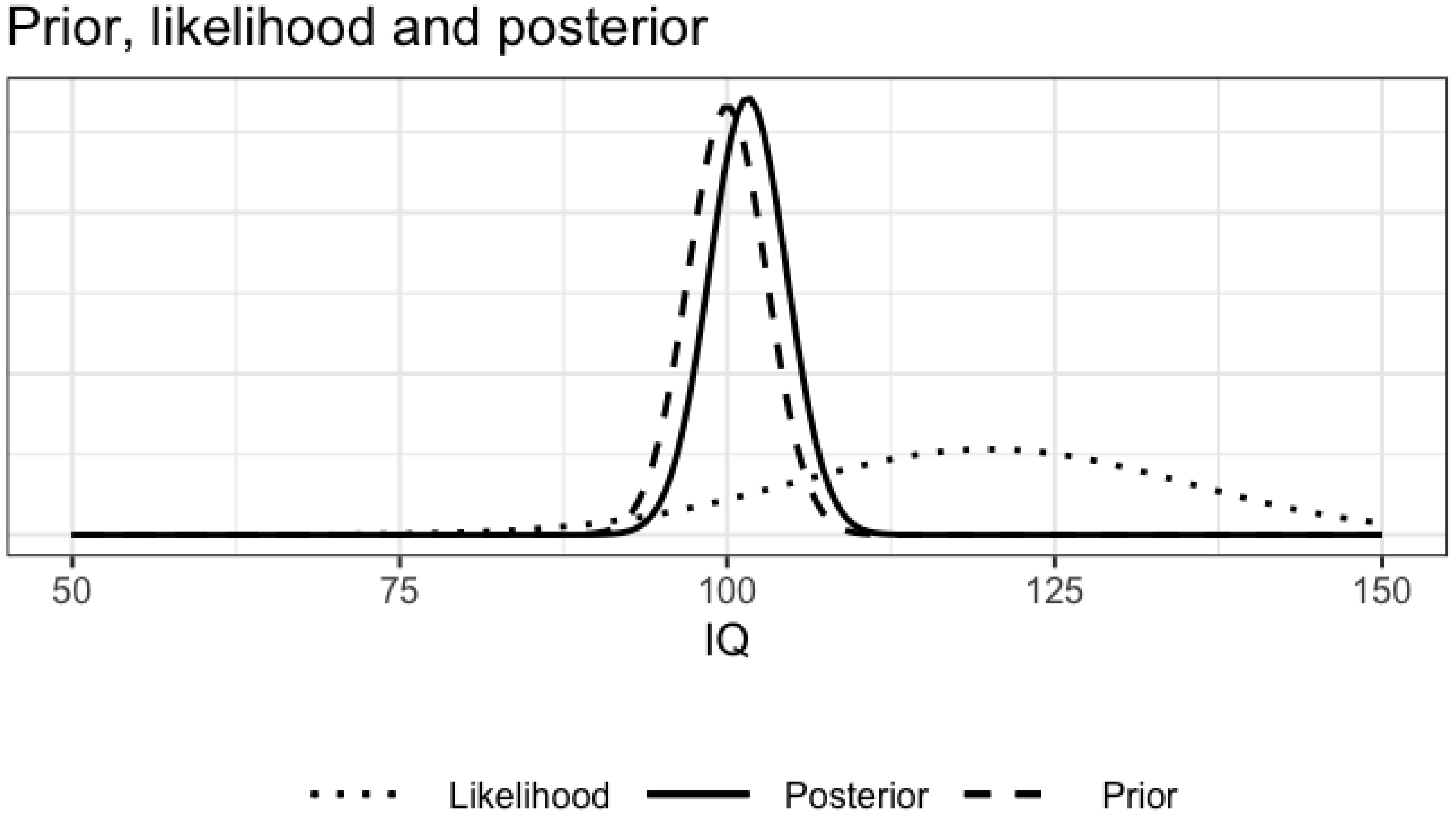
Estimating the average IQ: Prior knowledge

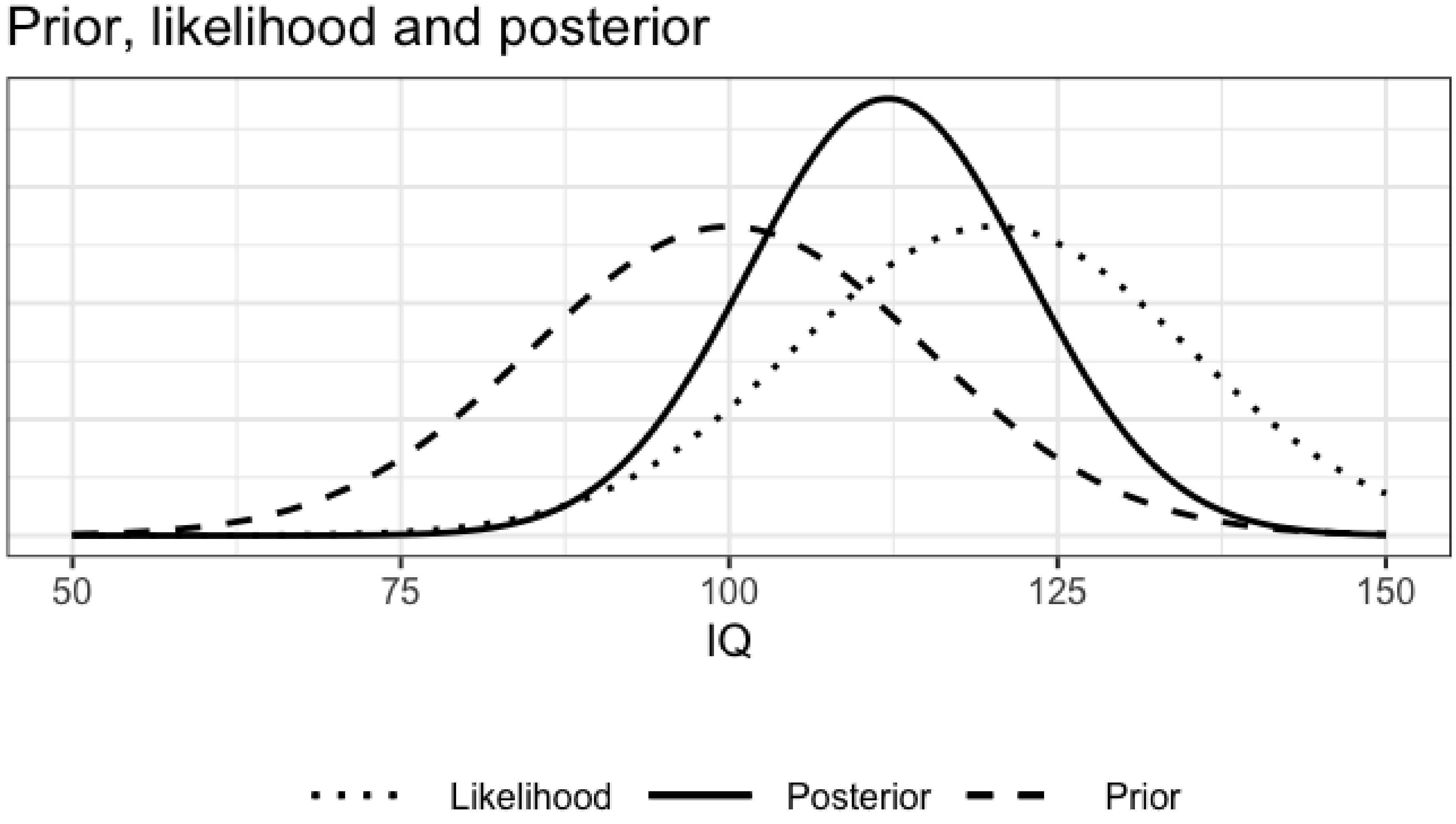


Estimating the average IQ: Prior knowledge




Prior, likelihood and posterior
50
75
100
125
150
IQ
Likelihood
Posterior
Prior


Prior, likelihood and posterior
50
75
100
125
150
IQ
Likelihood
Posterior
Prior


Prior, likelihood and posterior
50
75
100
125
150
IQ
Likelihood
Posterior
Prior

Some notes about the prior distribution

- A distributional form is needed (e.g., normal, gamma, Wishart, binomial, uniform, beta, etc...)
- Hyperparameters need to be specified (e.g., the mean of the normal prior and its variance)
- These choices should result in a prior that accurately reflects the current state of knowledge about the problem
    - Is this even possible?
- The resulting prior can greatly influence the results of the analysis

How to obtain the posterior?

- In complex models, the posterior is often intractable (impossible to compute exactly)
- Solution: approximate posterior by simulation – generate many draws from posterior distribution
- Compute mode, median, mean, 95% interval, etc. from the simulated draws

Regression example: Model

Suppose we have a regression model with 3 predictors.

-> 3 unknown regression coefficients ($\beta_1, \beta_2, \beta_3$) and one common but unknown σ^2.

Statistical model assuming centered data:

$$Y_i = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e_i$$

With $e_i \sim N(0,\ \sigma^2)$

Regression example: Priors

Specify prior: $P(\beta_1, \beta_2, \beta_3, \sigma^2)$

Conjugate: when the posterior is in the same distributional family as the prior

For illustration, we use ***conjugate*** priors here. Note that this is no longer needed in many software programs, including brms (and sometimes it might be better not to use the “default” conjugate priors, see day 4).

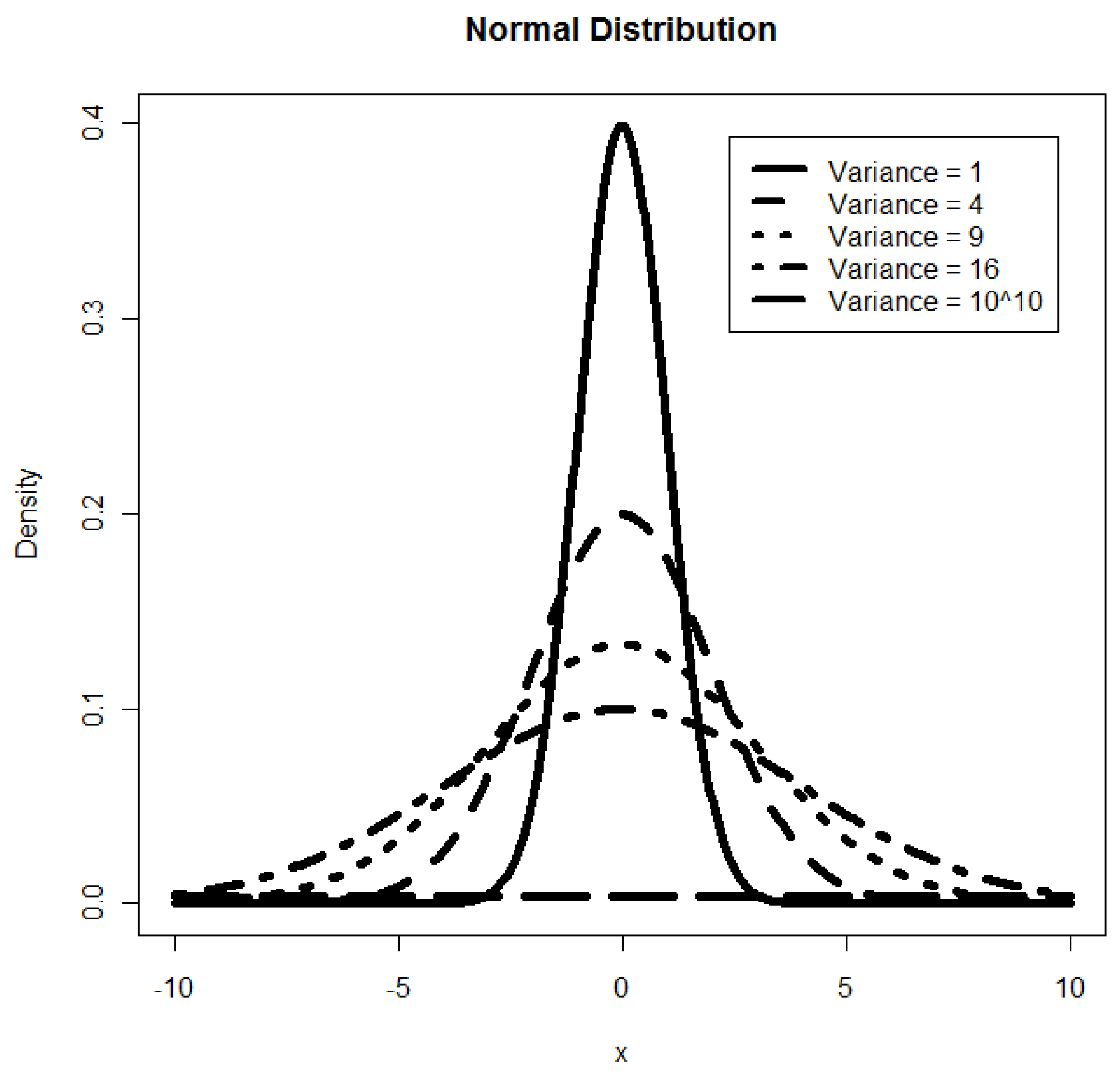
Regression example: Priors

Specify prior: $P(\beta_1, \beta_2, \beta_3, \sigma^2)$

- Prior (β_j) ~ Normal(μ_0, var_0)
- Prior (β_j) ~ Normal(0, 10000)

Normal priors

Hyperparameters:
μ (mean)
σ^2 (variance) or σ (SD)



Regression example: Priors

Specify prior: $P(\beta_1, \beta_2, \beta_3, \sigma^2)$

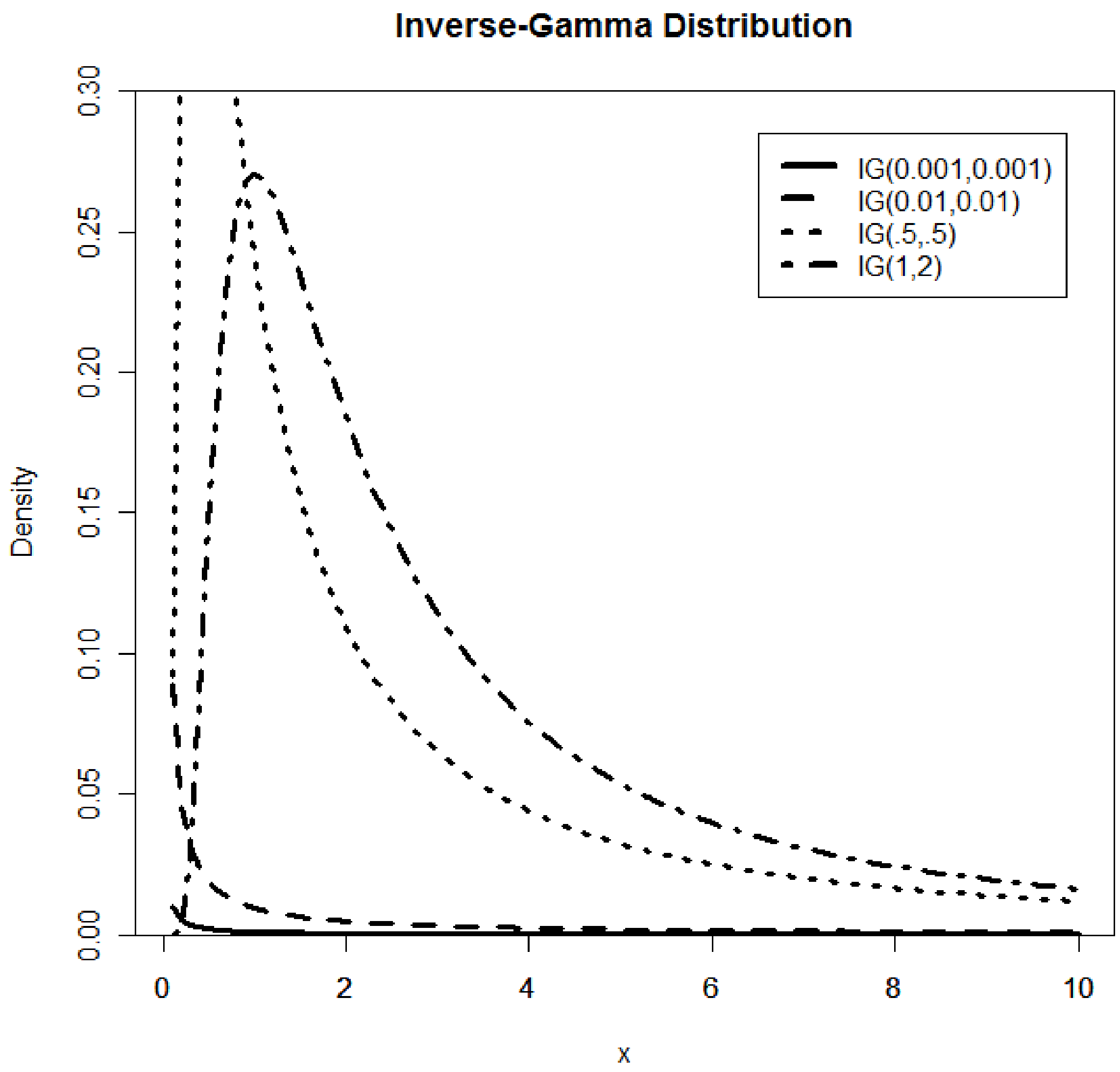
- Prior (β_j) ~ Normal(μ_0, var_0)
- Prior (β_j) ~ Normal(0, 10000)

- Prior (σ^2) ~ Inverse-gamma(0.001, 0.001)

Inverse gamma priors

Hyperparameters:
α (shape), β (scale)

More on this prior on day 4!



Regression example: Posterior

Combining the prior with the likelihood gives the posterior:

P($\beta_1, \beta_2, \beta_3, \sigma^2$ | data) -> this is a 4-dimensional distribution

Regression example: Gibbs sampling

Iterative evaluation via conditional distributions:

$$Post(\beta_1|\beta_2,\beta_3,\sigma^2,data) \sim Prior(\beta_1) \times\ likelihood$$

$$Post(\beta_2|\beta_1,\beta_3,\sigma^2,data) \sim Prior(\beta_2) \times\ likelihood$$

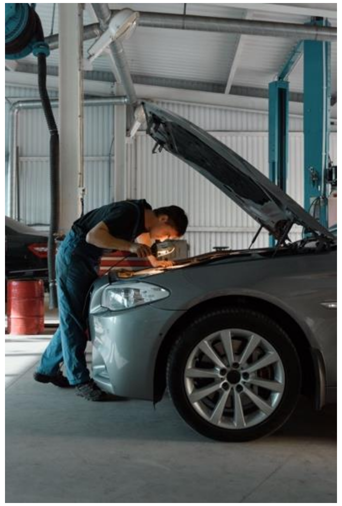
$$Post(\beta_3|\beta_1,\beta_2,\sigma^2,data) \sim Prior(\beta_3) \times\ likelihood$$

$$Post(\sigma^2|\beta_1,\beta_2,\beta_3,data) \sim Prior(\sigma^2) \times\ likelihood$$

Regression example: Gibbs sampling

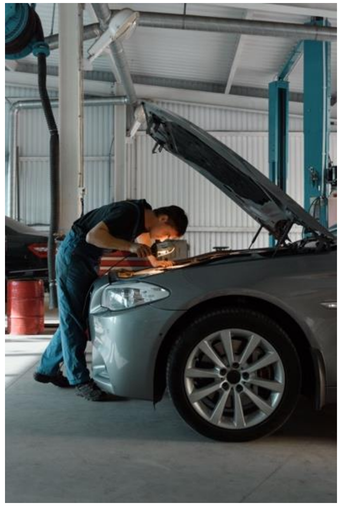
1. Assign starting values
2. Sample β_1 from conditional distribution
3. Sample β_2 from conditional distribution
4. Sample β_3 from conditional distribution
5. Sample σ^2 from conditional distribution
6. Go to step 2 and repeat

Gibbs sampling



$$Y_i = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e_i$$

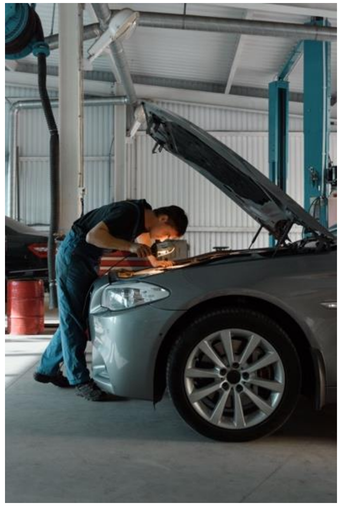
Gibbs sampling: Step 1



$$Y_i = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e_i$$

Step 1: $3 * X_1 + 5 * X_2 + 8 * X_3 + 10$

Gibbs sampling: Step 2

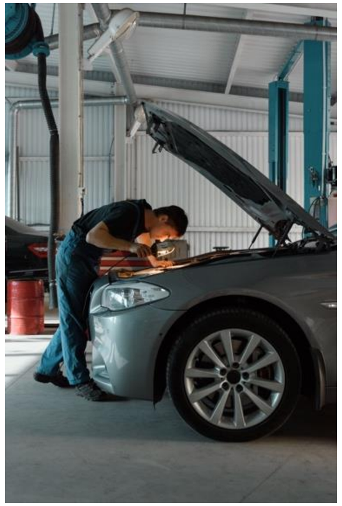


$$Y_i = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e_i$$

Step 1: $3 * X_1 + 5 * X_2 + 8 * X_3 + 10$

Step 2: $\beta_1 X_1 + 5 * X_2 + 8 * X_3 + 10$

Gibbs sampling: Step 3



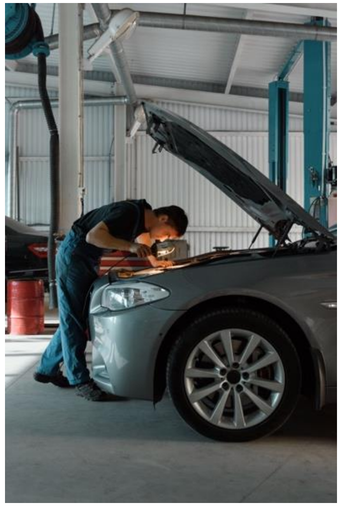
$$Y_i = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e_i$$

Step 1: $3 * X_1 + 5 * X_2 + 8 * X_3 + 10$

Step 2: $\beta_1 X_1 + 5 * X_2 + 8 * X_3 + 10$

Step 3: $\beta_1 X_1 + \beta_2 X_2 + 8 * X_3 + 10$

Gibbs sampling: Step 4



$$Y_i = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e_i$$

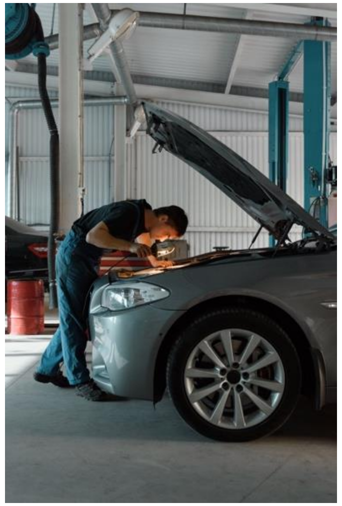
Step 1: $3 * X_1 + 5 * X_2 + 8 * X_3 + 10$

Step 2: $\beta_1 X_1 + 5 * X_2 + 8 * X_3 + 10$

Step 3: $\beta_1 X_1 + \beta_2 X_2 + 8 * X_3 + 10$

Step 4: $\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + 10$

Gibbs sampling: Step 5



$$Y_i = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e_i$$

Step 1: $3 * X_1 + 5 * X_2 + 8 * X_3 + 10$
Step 2: $\beta_1 X_1 + 5 * X_2 + 8 * X_3 + 10$
Step 3: $\beta_1 X_1 + \beta_2 X_2 + 8 * X_3 + 10$
Step 4: $\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + 10$
Step 5: $\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e_i$

This concludes the first iteration.
Replace the initial starting values with the current draws and repeat.

Regression example: Gibbs sampling

Iteration	β_1	β_2	β_3	σ^2
1	3.00	5.00	8.00	10
2	3.75	4.25	7.00	8
3	3.65	4.11	6.78	5
.	.	.	.	.
15	4.45	3.19	5.08	1.1
.	.	.	.	.
.	.	.	.	.
199	4.59	3.75	5.21	1.2
200	4.36	3.45	4.65	1.3

Regression example: Gibbs sampling

This is just one possible algorithm, we will review others on day 3.

Two important consequences:

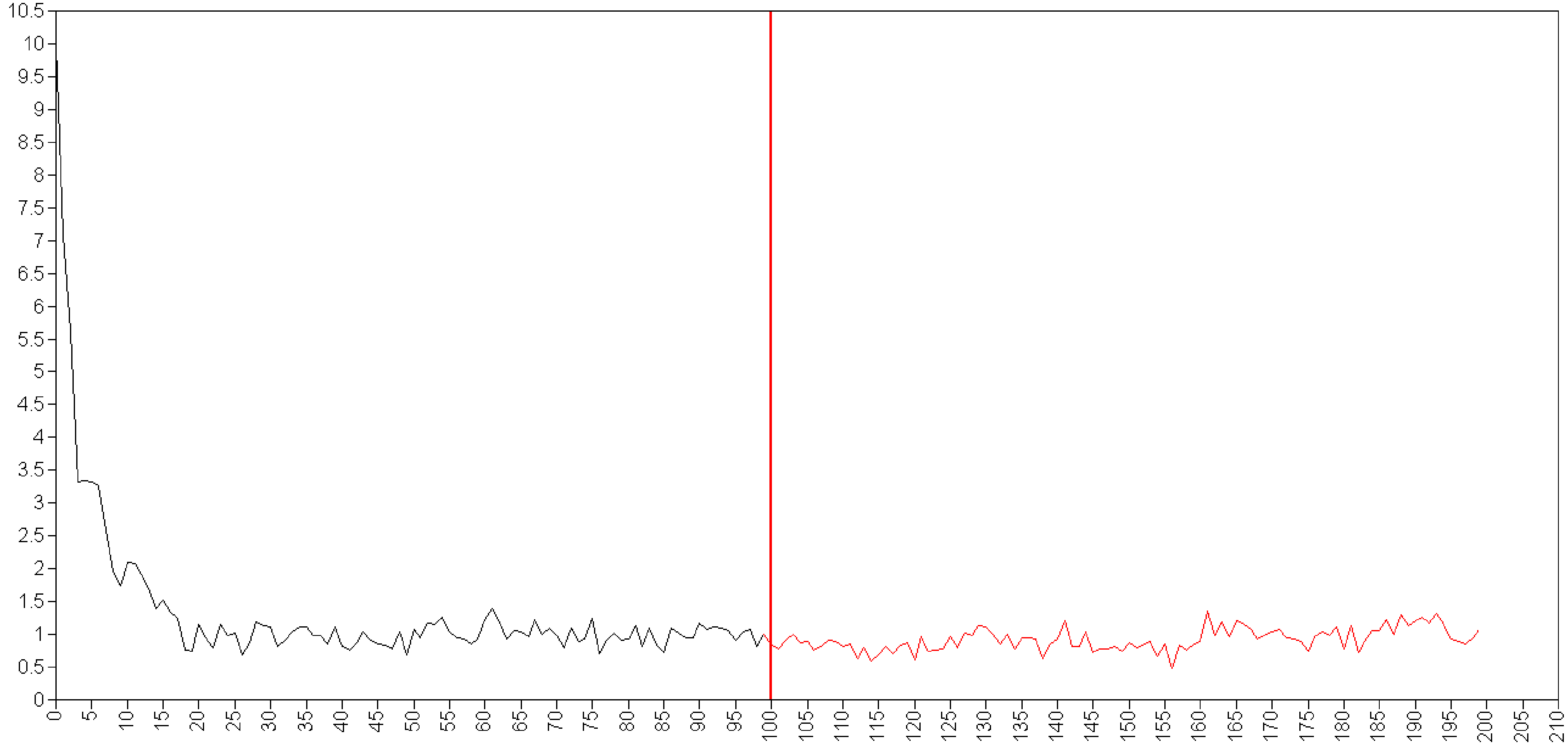
1. We obtain a distribution of samples as our result
2. We need to ensure convergence of the analysis

Interpreting the results of a Bayesian analysis

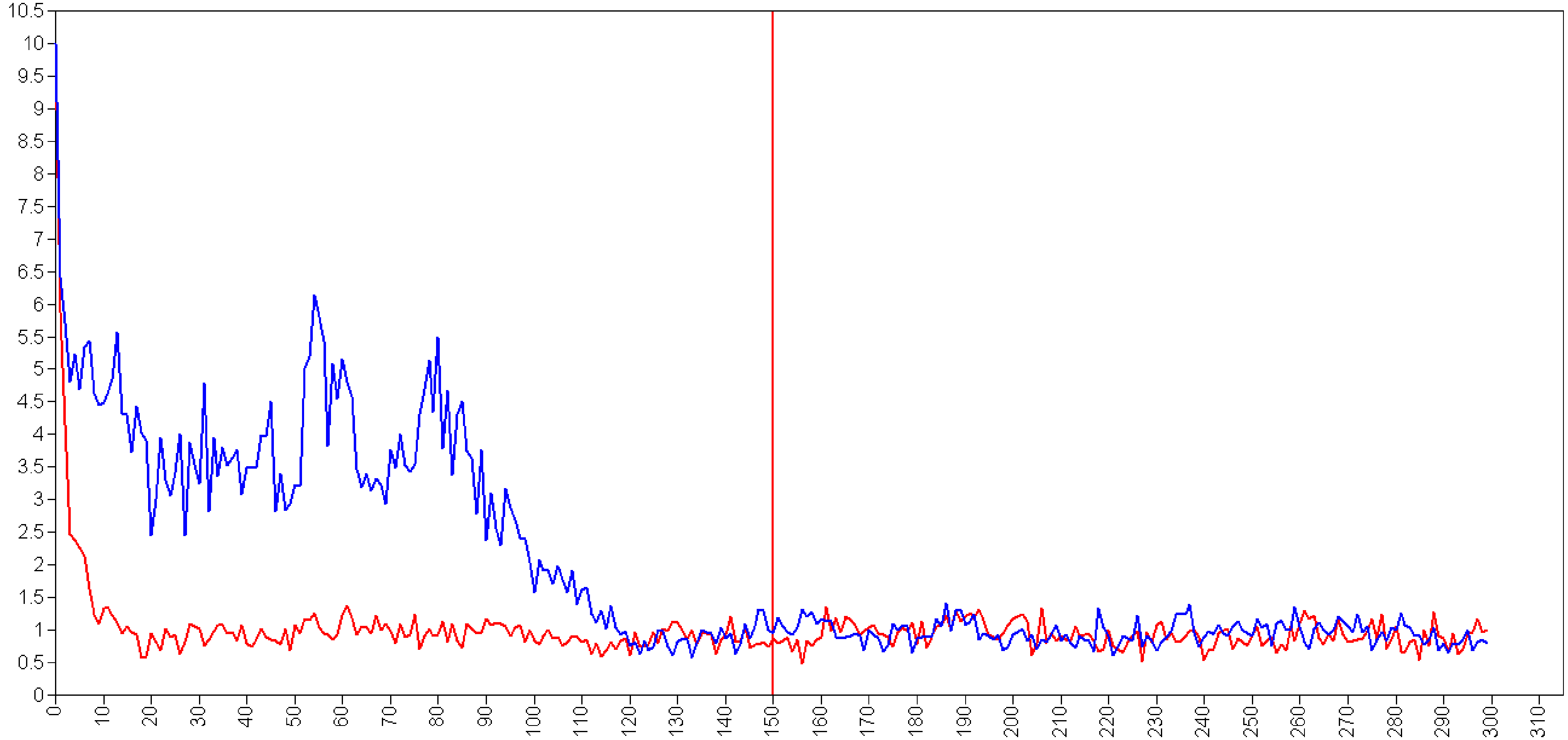
The first step is always to ensure convergence:

1. Visual assessment
2. Numerical diagnostics (and possibly warnings given by the software)

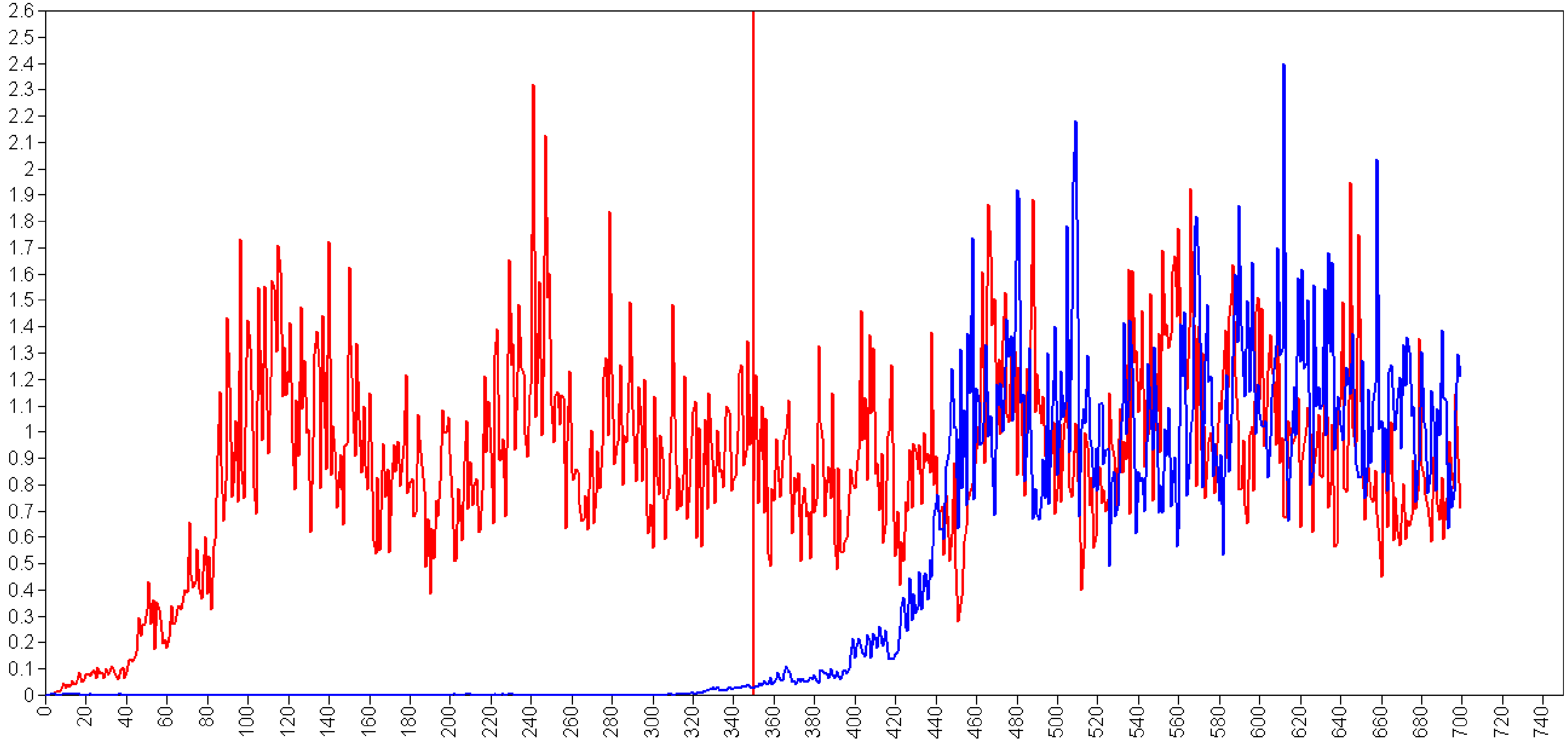
Assessing convergence: Trace plot



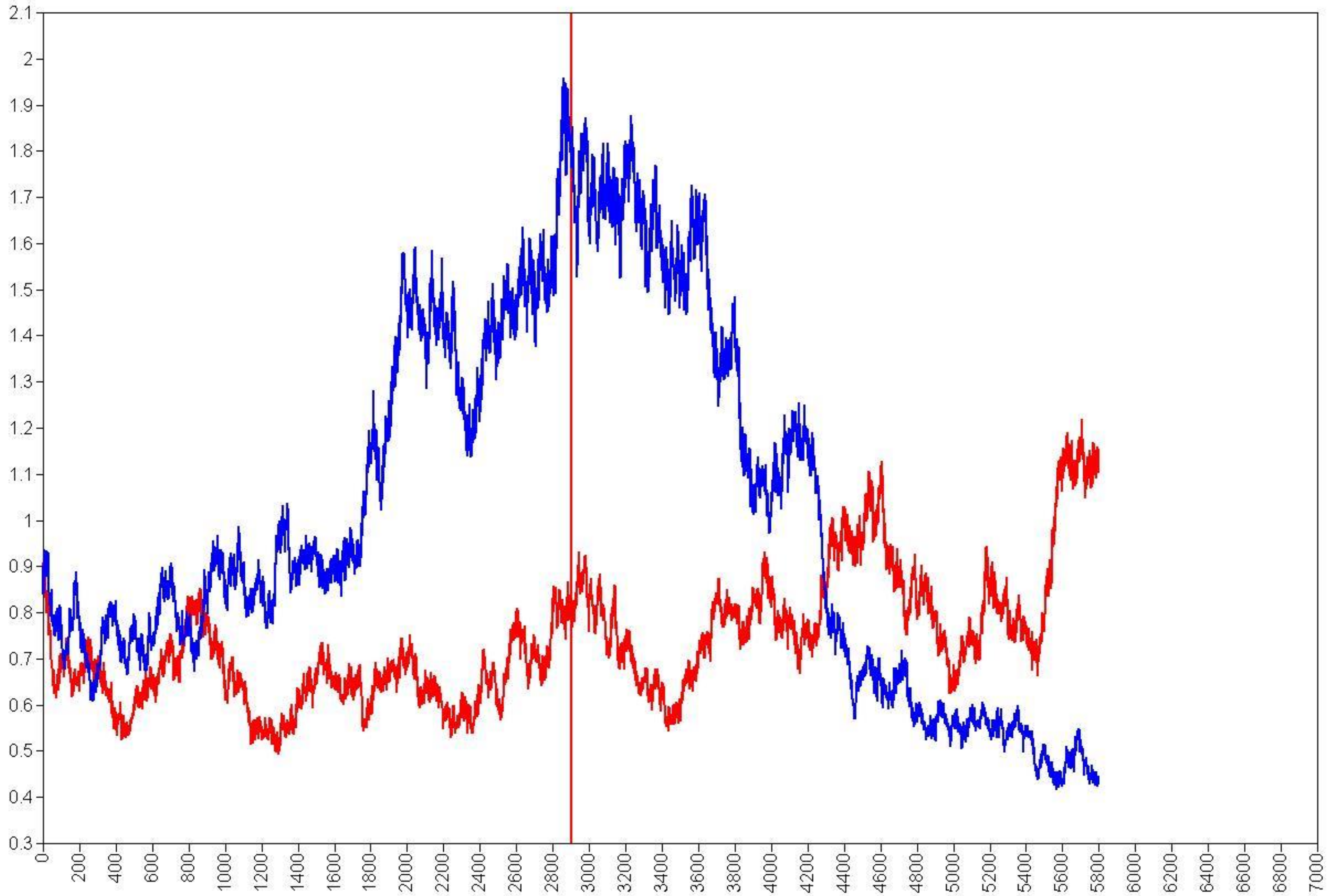
Assessing convergence: Trace plot



Assessing convergence: Trace plot



Assessing convergence: Trace plot



Assessing convergence

Sampler must run *t* iterations ‘burn in’ before we reach the target distribution (our posterior)

How many iterations are needed to converge on the target distribution?

- More iterations = more precision

- Run several chains in parallel
- Trace plot
- Numerical diagnostics

Assessing convergence: Numerical diagnostics

Warning messages:

1: There were 300 divergent transitions after warmup. See https://mc-stan.org/misc/warnings.html#divergent-transitions-after-warmup to find out why this is a problem and how to eliminate them.

2: There were 243 transitions after warmup that exceeded the maximum treedepth. Increase max_treedepth above 10. See https://mc-stan.org/misc/warnings.html#maximum-treedepth-exceeded

3: There were 2 chains where the estimated Bayesian Fraction of Missing Information was low. See https://mc-stan.org/misc/warnings.html#bfmi-low

4: The largest R-hat is 2.62, indicating chains have not mixed. Running the chains for more iterations may help. See https://mc-stan.org/misc/warnings.html#r-hat

5: Bulk Effective Samples Size (ESS) is too low, indicating posterior means and medians may be unreliable. Running the chains for more iterations may help. See https://mc-stan.org/misc/warnings.html#bulk-ess

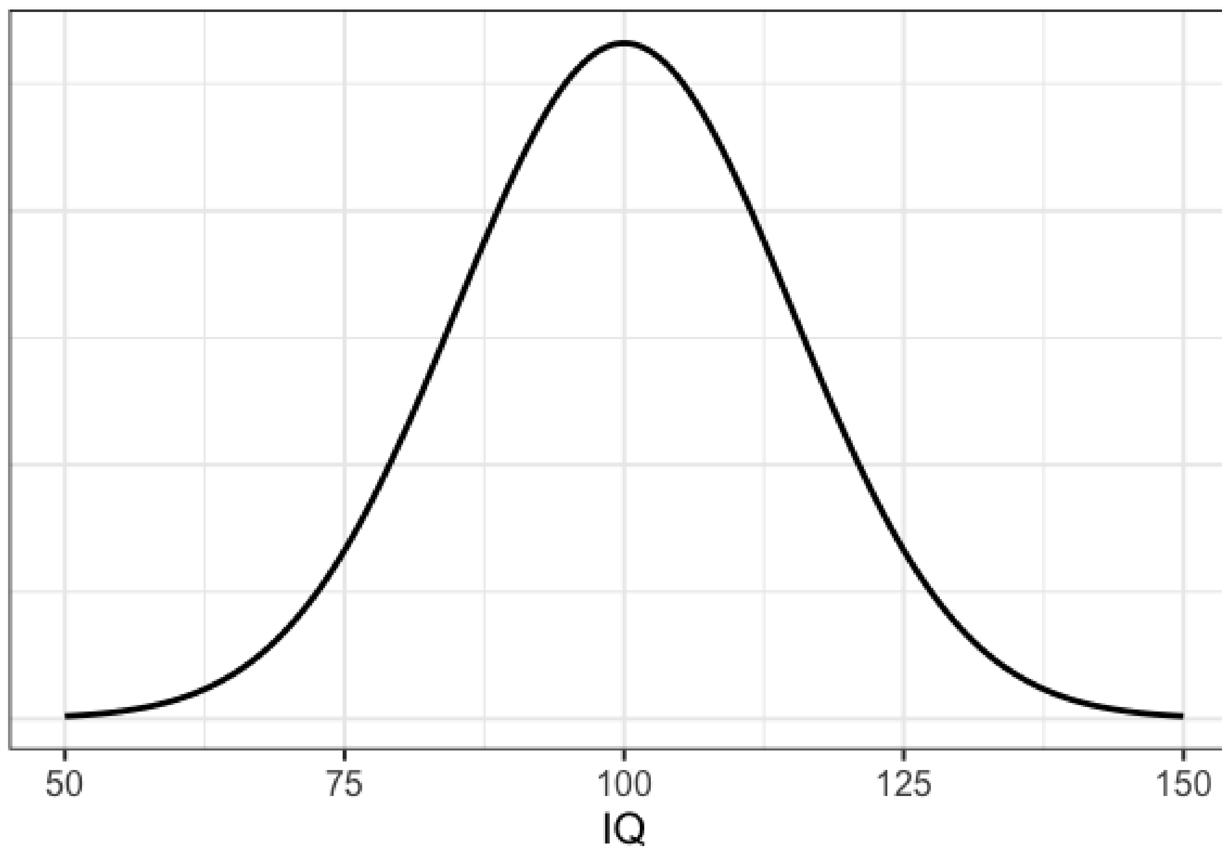
6: Tail Effective Samples Size (ESS) is too low, indicating posterior variances and tail quantiles may be unreliable. Running the chains for more iterations may help. See https://mc-stan.org/misc/warnings.html#tail-ess

Interpreting the results of a Bayesian analysis

The posterior samples provide us with all information.

We can:

- Plot the posterior
- Compute the mean, mode or median
- Compute the SD
- Compute a credible interval



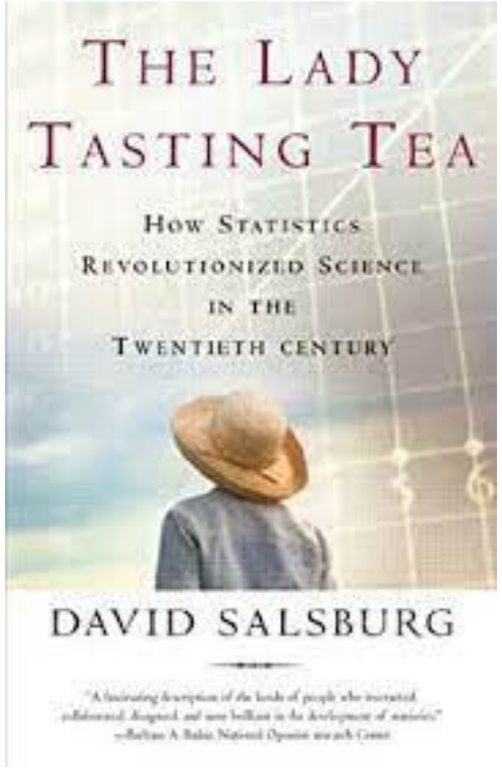
Interpreting the results of a Bayesian analysis

1. Assess convergence (see day 2 & 3)
2. Visualize and summarize the posterior
3. Robustness checks
    - prior sensitivity analysis (day 4)
    - posterior predictive checking (day 3)

The tea experiment

A famous anecdote

Experiment: H0: the lady is guessing



A famous anecdote

Experiment:

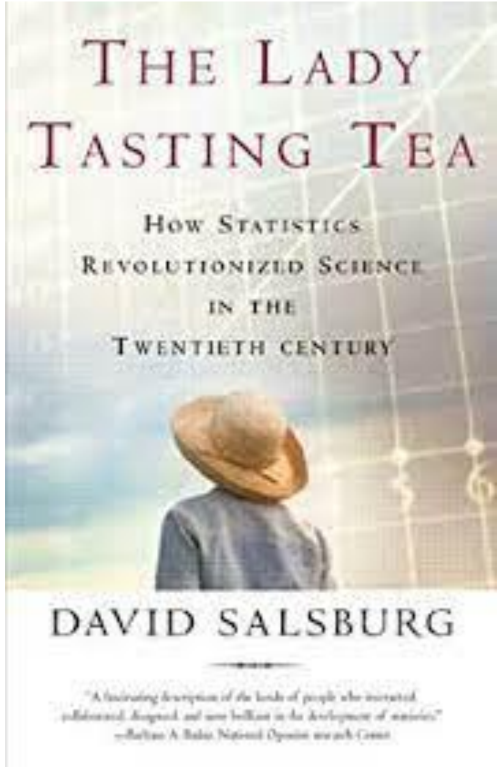
H0: the lady is guessing

Result: 5 out of 6 correct.

Is this a matter of guessing/luck or evidence that the lady can taste the difference?

P-value = Prob(5 or more correct if H0 is true)

Result: p = .109



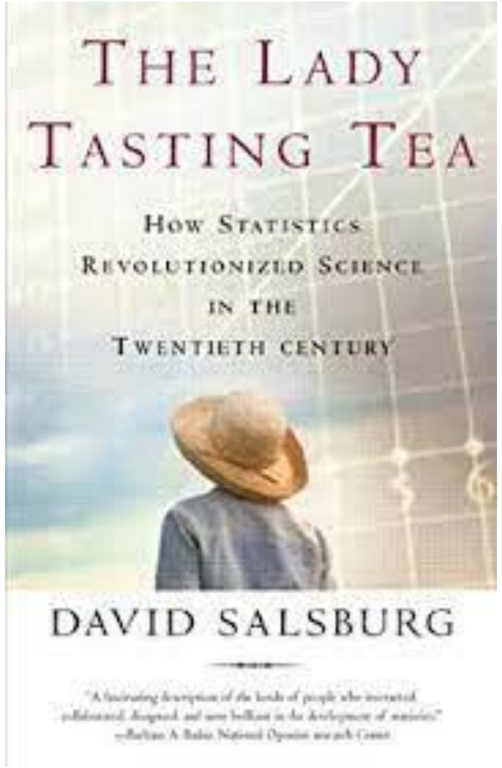
A famous anecdote

Experiment: H0: the lady is guessing

Suppose we use a different sampling plan and continue sampling until we have 5 correct cups

Result: 5 out of 6 correct.

What would we conclude now?



A famous anecdote

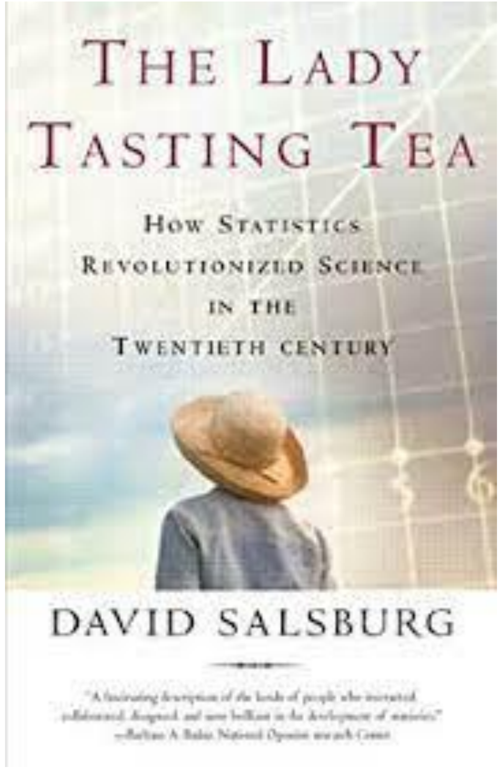
Experiment: H0: the lady is guessing

Result: 5 out of 6 correct.

But we use a different sampling plan and continue until 5 correct.

Result: p=.031

Thus: our results and conclusions depend on the sampling plan.
In the frequentist framework, the same data can offer different conclusions!



Why to use Bayes?

- Frequentist methods based on p-values violate the likelihood principle
- Possibility of incorporating prior information and thus reducing the required sample size (see also day 5)
- Automatic uncertainty quantification (also of functions of parameters)
- Estimating more complex models
- More intuitive interpretation

Interpretation frequentist vs. Bayesian

Frequentist

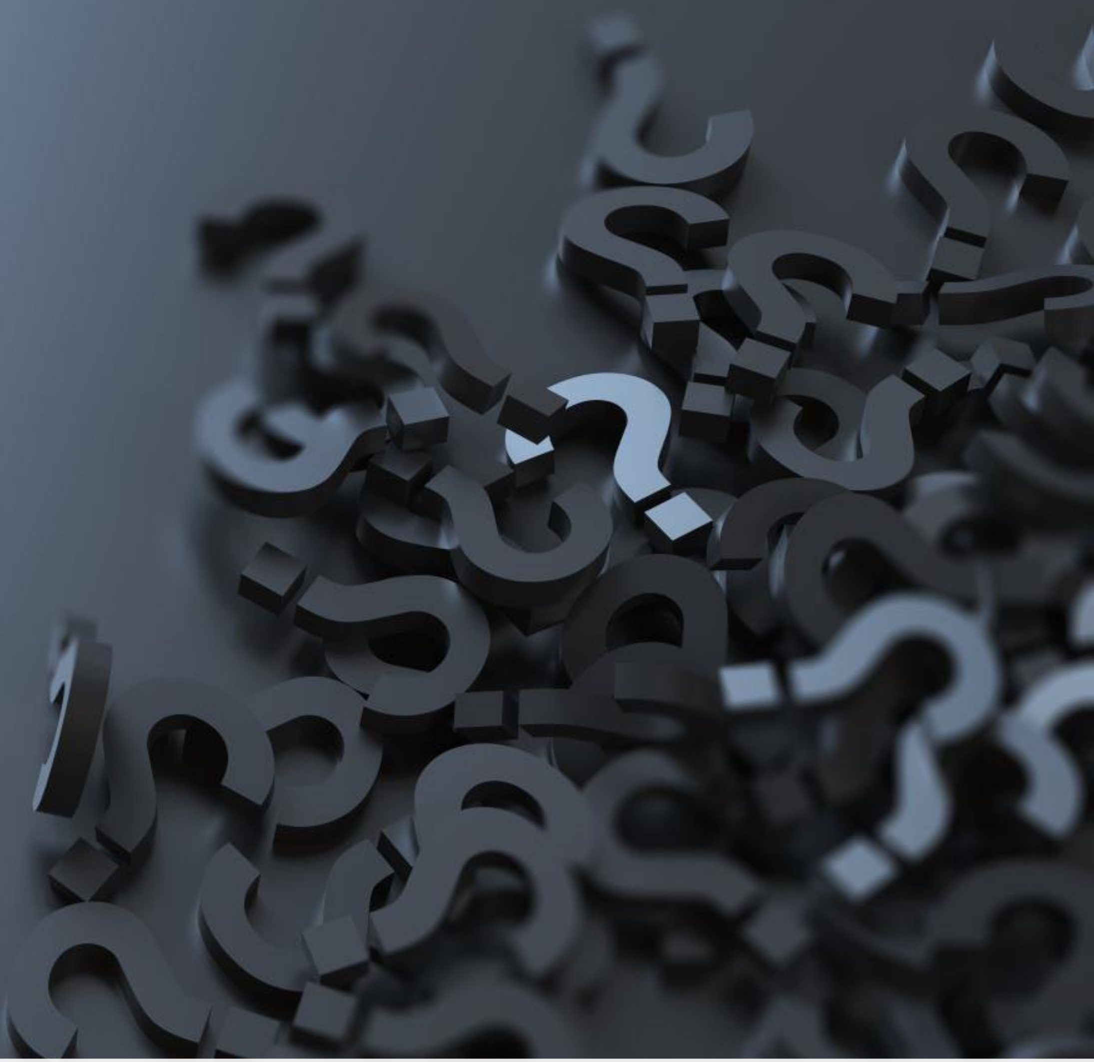
- Parameters are treated as *fixed*: there is only one true parameter value in the population
- Probability as a relative frequency
- If I repeat this experiment infinitely many times, 95% of the computed CIs will contain the true value

Bayesian

- Parameters are treated as *random*: true value is unkown so specify a prior probability to capture our uncertainty or beliefs
- Probability as degree of belief
- There is a 95% probability that the true value will lie in the CI

Recap

- Introduction to the course
- A brief history of Bayesian statistics
- Bayes rule and the idea behind the prior
- How to obtain the posterior and what to do once you have it
- Why to use Bayes?



Questions?