Come to the dark side: we got cookies! An introduction to Bayesian statistics

Gianluca Baio

University College London
Department of Statistical Science

gianluca@stats.ucl.ac.uk

(Thanks to Tom Fearn, University College London)

28th Annual Chemometrics Symposium

Utrecht, Thursday 7 November 2013

OPEN ACCESS Freely available online

PLOS ONE

When Math Hurts: Math Anxiety Predicts Pain Network Activation in Anticipation of Doing Math

Ian M. Lyons[1,2]*, **Sian L. Beilock**[1]*

1 Department of Psychology, University of Chicago, Chicago, Illinois, United States of America, **2** Department of Psychology, Western University, London, Ontario, Canada

Abstract

Math can be difficult, and for those with high levels of mathematics-anxiety (HMAs), math is associated with tension, apprehension, and fear. But what underlies the feelings of dread effected by math anxiety? Are HMAs' feelings about math merely psychological epiphenomena, or is their anxiety grounded in simulation of a concrete, visceral sensation – such as pain – about which they have every right to feel anxious? We show that, when anticipating an upcoming math-task, the higher one's math anxiety, the more one increases activity in regions associated with visceral threat detection, and often the experience of pain itself (bilateral dorso-posterior insula). Interestingly, this relation was not seen during math performance, suggesting that it is not that math itself hurts; rather, the anticipation of math is painful. Our data suggest that pain network activation underlies the intuition that simply anticipating a dreaded event can feel painful. These results may also provide a potential neural mechanism to explain why HMAs tend to avoid math and math-related situations, which in turn can bias HMAs away from taking math classes or even entire math-related career paths.

Citation: Lyons IM, Beilock SL (2012) When Math Hurts: Math Anxiety Predicts Pain Network Activation in Anticipation of Doing Math. PLoS ONE 7(10): e48076. doi:10.1371/journal.pone.0048076

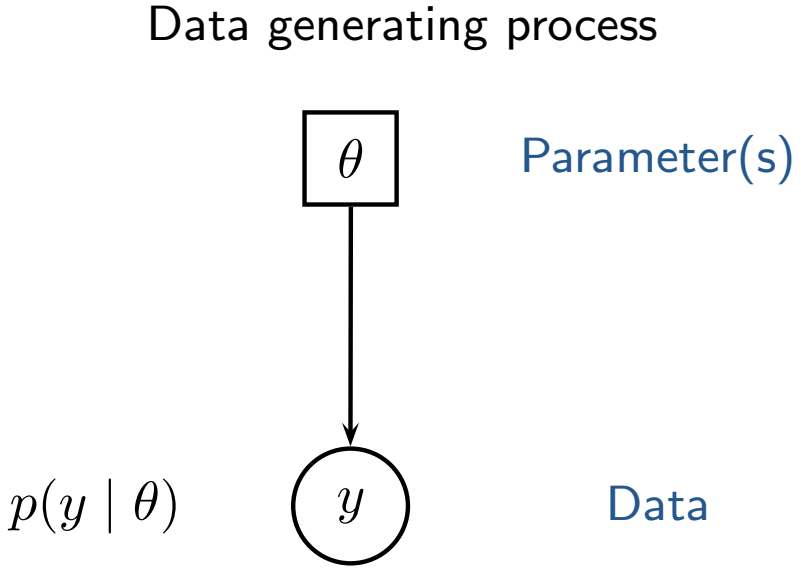
Editor: Georges Chapouthier, Université Pierre et Marie Curie, France

Received February 23, 2012; **Accepted** September 20, 2012; **Published** October 31, 2012

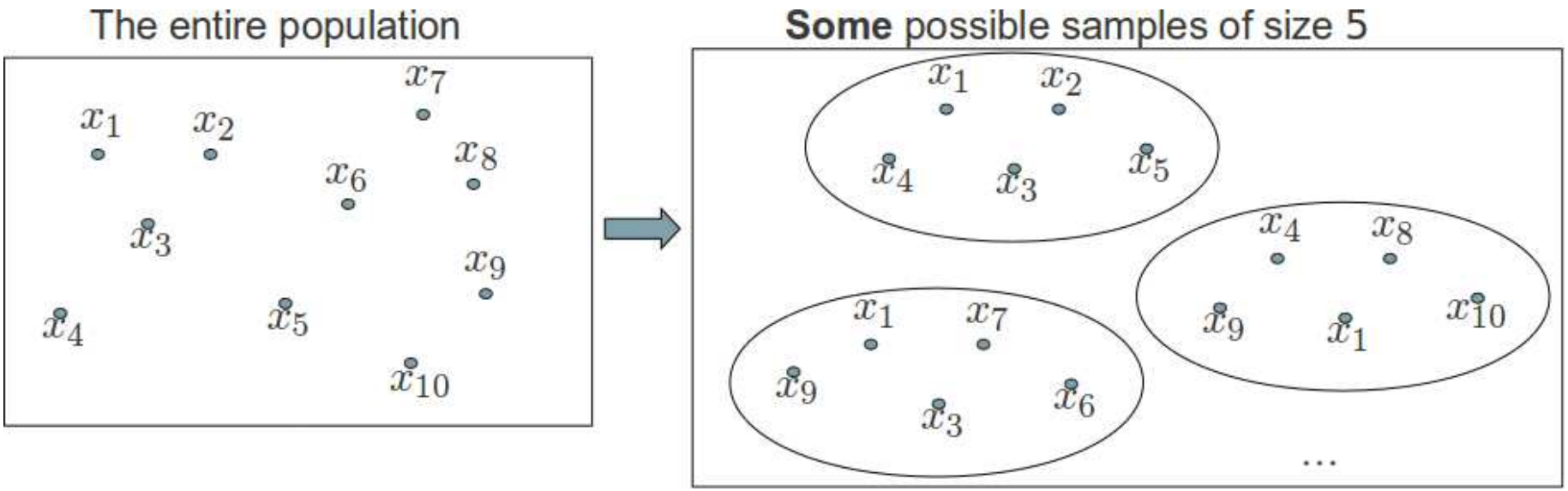
Funding: Research supported by National Science Foundation CAREER DRL-0746970 and the National Science Foundation Spatial Intelligence Learning Center to SLB. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Statistical inference

- Typically, we observe some data and we want to use them to learn about some unobservable feature of the general population that we are interested in
- To do this, we use statistical models to describe the probabilistic mechanism by which (**we assume!**) that the data have arisen



Sampling variability



- Size $N = 10$
- Mean μ
- Standard deviation σ

- Size $n = 5$
- Mean $\bar{x}$
- Standard deviation s_x

In reality we observe **only one** such sample (out of the many possible — in fact there are 252 different ways of picking **at random** 5 units out of the population!) and we want to use the information contained in **that** sample to **infer** about the population parameters (e.g. the true mean and standard deviation)

Running example: Binomial model

- Suppose in a study we observe that, in a single postcode sector, n kits for cancer screening are sent out in a certain period, of which only y ($\leq n$) are returned by patients. A reasonable model in this case is

$$y \mid \theta, n \sim \text{Binomial}(\theta, n)$$

as a function of a parameter θ, which in this case represents the **screening uptake rate** (annual uptake probability) in the overall population

Running example: Binomial model

- Suppose in a study we observe that, in a single postcode sector, n kits for cancer screening are sent out in a certain period, of which only y ($\leq n$) are returned by patients. A reasonable model in this case is

$$y \mid \theta, n \sim \text{Binomial}(\theta, n)$$

  as a function of a parameter θ, which in this case represents the **screening uptake rate** (annual uptake probability) in the overall population

- This is equivalent to assuming

$$p(y \mid \theta, n) = \binom{n}{y} \theta^y (1-\theta)^{(n-y)} \propto \theta^y (1-\theta)^{(n-y)}$$

Running example: Binomial model

- Suppose in a study we observe that, in a single postcode sector, n kits for cancer screening are sent out in a certain period, of which only y ($\leq n$) are returned by patients. A reasonable model in this case is

$$y \mid \theta, n \sim \text{Binomial}(\theta, n)$$

  as a function of a parameter θ, which in this case represents the **screening uptake rate** (annual uptake probability) in the overall population

- This is equivalent to assuming

$$p(y \mid \theta, n) = \binom{n}{y} \theta^y (1-\theta)^{(n-y)} \propto \theta^y (1-\theta)^{(n-y)}$$

- The objective of statistical inference is to find a way of "learning" about θ, using
  - The **evidence** (observed data, possibly also on some covariates $\boldsymbol{X}$)
  - The **assumptions** that we are making about the random phenomenon under study
- Some times we are interested in **prediction** (e.g. for a yet unobserved unit)

Schools of inference

There are 3 major schools of inference; those taught in basic stats courses are:

Schools of inference

There are 3 major schools of inference; those taught in basic stats courses are:

- **Frequentist** (Neyman-Pearson)
  - The frequentist school does not attempt to make inference for a specific set of data, but rather it considers and evaluates *inference procedures* (like the way in which an estimator is defined)
  - Inference consists in the probabilistic assessment of the properties of the procedure (ie unbiasedness, robustness, etc)

Schools of inference

There are 3 major schools of inference; those taught in basic stats courses are:

- **Frequentist** (Neyman-Pearson)
  - The frequentist school does not attempt to make inference for a specific set of data, but rather it considers and evaluates *inference procedures* (like the way in which an estimator is defined)
  - Inference consists in the probabilistic assessment of the properties of the procedure (ie unbiasedness, robustness, etc)
- **Likelihood** (Fisher)
  - The likelihood school maintains that inference from the data at hand is completely determined by the *likelihood function*, that is the statistical model that we use to describe the problem, but as a mathematical function of the parameters
  - For example, the *Maximum Likelihood Estimator* (MLE) is the value of θ that maximises $\mathcal{L}(\theta \mid y) = p(y \mid \theta)$

Schools of inference

There are 3 major schools of inference; those taught in basic stats courses are:

- **Frequentist** (Neyman-Pearson)
  - The frequentist school does not attempt to make inference for a specific set of data, but rather it considers and evaluates *inference procedures* (like the way in which an estimator is defined)
  - Inference consists in the probabilistic assessment of the properties of the procedure (ie unbiasedness, robustness, etc)
- **Likelihood** (Fisher)
  - The likelihood school maintains that inference from the data at hand is completely determined by the *likelihood function*, that is the statistical model that we use to describe the problem, but as a mathematical function of the parameters
  - For example, the *Maximum Likelihood Estimator* (MLE) is the value of θ that maximises $\mathcal{L}(\theta \mid y) = p(y \mid \theta)$

Often, these two schools are presented as a combined and unified theory, although they are actually separated, and, to some scholars, irreconcilable!

Example

- Consider again the cancer screening example and focus on y, the total number of kits returned out of the n sent out, and suppose in our data we observe $n = 32$ and $y = 18$

Example

- Consider again the cancer screening example and focus on y, the total number of kits returned out of the n sent out, and suppose in our data we observe $n = 32$ and $y = 18$

- If we use a standard analysis, inference is generally performed using the MLE, which in this case is

$$\hat{\theta} = \frac{y}{n} = \frac{18}{32} = 0.5625$$

with standard error

$$\text{se}(\hat{\theta}) = \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} = \sqrt{\frac{0.5625 \times 0.4375}{32}} = 0.0877$$

and 95% CI

$$\hat{\theta} \pm 1.96 \times \text{se}(\hat{\theta}) = 0.5625 \pm 1.96 \times 0.0877 = [0.3096; 0.7344]$$

Example

- Consider again the cancer screening example and focus on y, the total number of kits returned out of the n sent out, and suppose in our data we observe $n = 32$ and $y = 18$

- If we use a standard analysis, inference is generally performed using the MLE, which in this case is

$$\hat{\theta} = \frac{y}{n} = \frac{18}{32} = 0.5625$$

with standard error

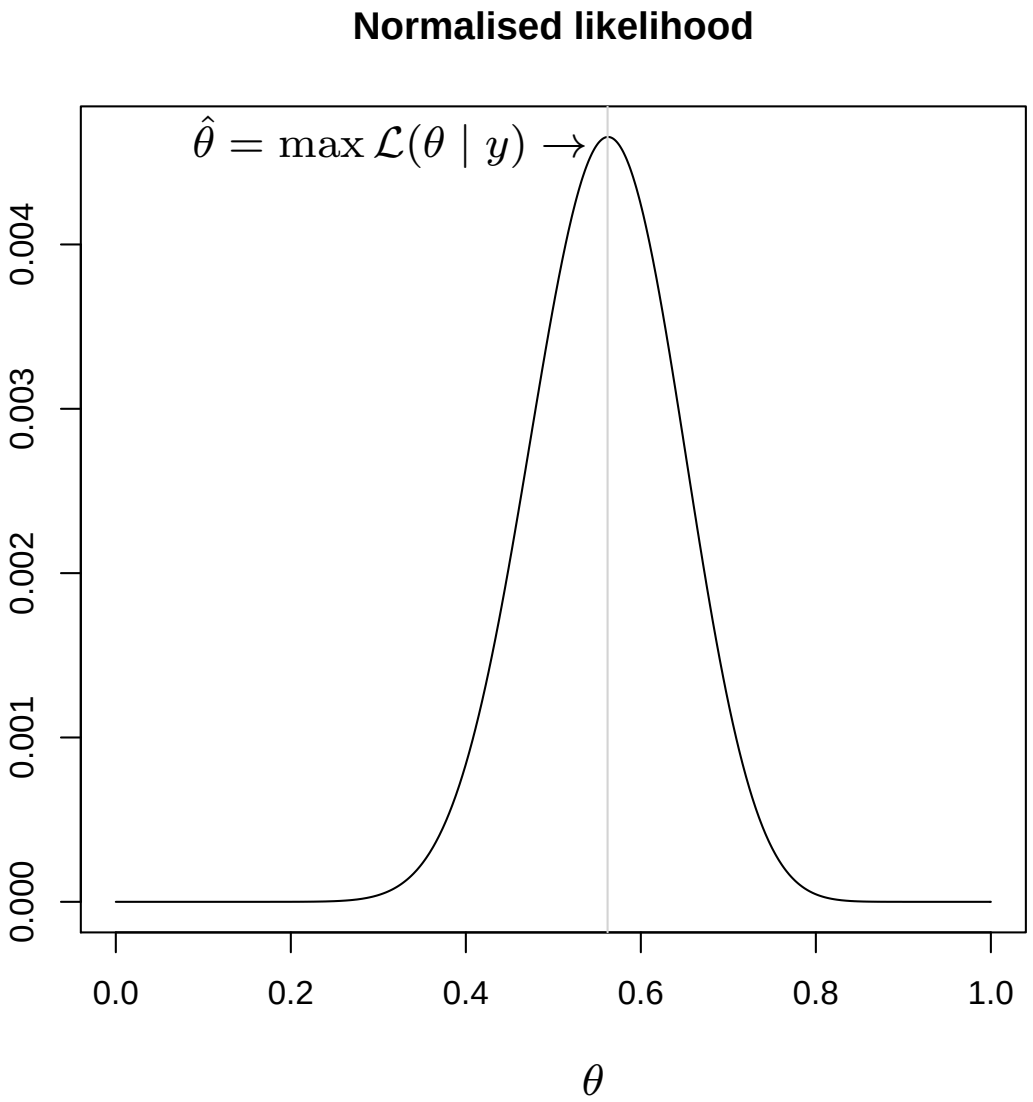
$$\text{se}(\hat{\theta}) = \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} = \sqrt{\frac{0.5625 \times 0.4375}{32}} = 0.0877$$

and 95% CI

$$\hat{\theta} \pm 1.96 \times \text{se}(\hat{\theta}) = 0.5625 \pm 1.96 \times 0.0877 = [0.3096; 0.7344]$$

- **NB**: The MLE has all good frequentist properties!

Likelihood function & MLE



- Since

$$p(y \mid \theta) = \binom{n}{y} \theta^y (1-\theta)^{(n-y)}$$

the likelihood function is

$$\mathcal{L}(\theta \mid y) = \theta^{18} (1-\theta)^{(32-18)}$$

- As is easy to see, the point $\hat{\theta}$ is the one associated with the maximum value of the likelihood
- Therefore, we deem it the "most likely", or the "most supported" by the observed data

Comments

- In both the frequentist and the likelihood approaches to inference, the parameters are considered as **fixed and unknown** quantities
- In other words, the only form of uncertainty (and the very reason why we need statistics) is the *individual* (sampling) variability

Comments

- In both the frequentist and the likelihood approaches to inference, the parameters are considered as **fixed and unknown** quantities
- In other words, the only form of uncertainty (and the very reason why we need statistics) is the *individual* (sampling) variability

- We are **not** entitled to make probabilistic statements on the value of the parameters (as they do not possess a probability distribution!)
- Accordingly, the 95% interval is interpreted as the procedure such that **if** applied to many (identical) replications of the same study/experiment would include the "true" value of θ in 95% of the cases

Subjective probability as the unique measure of uncertainty

- Every single uncertain events is associated with a *probability*, which represents the experimenter's **degree of belief** in its realisation — this does not necessarily coincide with the *frequency* with which the event is observed
- Each individual is entitled to their own, subjective evaluation. According to the evidence that becomes sequentially available, individuals tend to update their belief
- The probability of a given event also **depends on the individual whose uncertainty is expressed and on the information background behind the evaluation**. Upon varying these quantities, so does the measure of probability
- Consequently, there is no need for the assumption of the existence of a unique, "true" (yet unknown) value for the probability of an event

Subjective probability as the unique measure of uncertainty

- Every single uncertain events is associated with a *probability*, which represents the experimenter's **degree of belief** in its realisation — this does not necessarily coincide with the *frequency* with which the event is observed
- Each individual is entitled to their own, subjective evaluation. According to the evidence that becomes sequentially available, individuals tend to update their belief
- The probability of a given event also **depends on the individual whose uncertainty is expressed and on the information background behind the evaluation**. Upon varying these quantities, so does the measure of probability
- Consequently, there is no need for the assumption of the existence of a unique, "true" (yet unknown) value for the probability of an event

- The Bayesian philosophy does not deny the usefulness of frequencies and the fact that parameters may be "fixed and unknown", physical quantities. **But these concepts are just not essential!**

Where did it all start?

Reverend Thomas Bayes (1702 - 1761)

PROBLEM.

Given the number of times in which an unknown event has happened and failed: *Required* the chance that the probability of its happening in a ſingle trial lies ſomewhere between any two degrees of probability that can be named.

In modern language: given $r \sim \text{Binomial}(\theta, n)$, find $\Pr(\theta_1 \leq \theta \leq \theta_2 \mid r, n)$

Some historical references:

http://www.bayesian.org/resources/bayes.html

S. Bertsch McGrayne (2011). The Theory That Would Not Die

S. Fienberg (206). When did Bayesian inference become Bayesian?

Bayesian inference (cont'd)

$p(\theta)$

Prior (subjective knowledge)

Bayesian inference (cont'd)

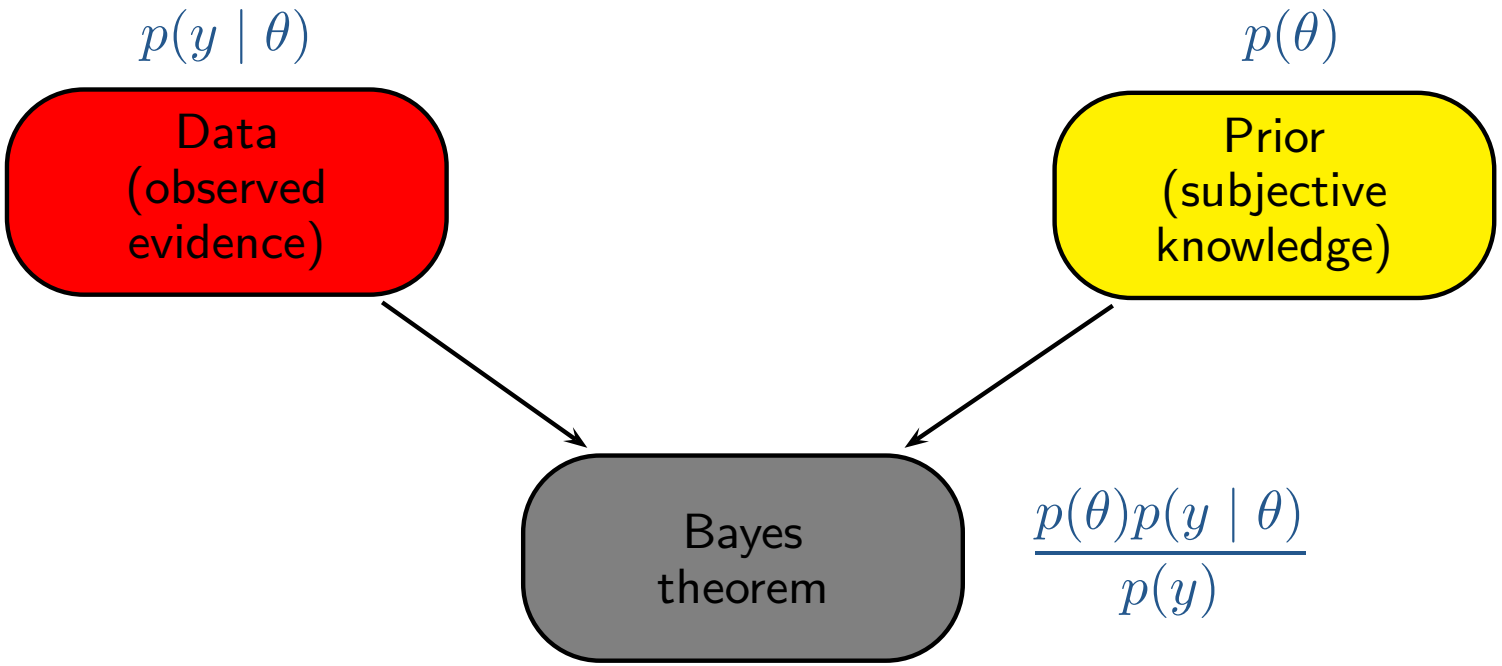
$p(y \mid \theta)$

Data (observed evidence)

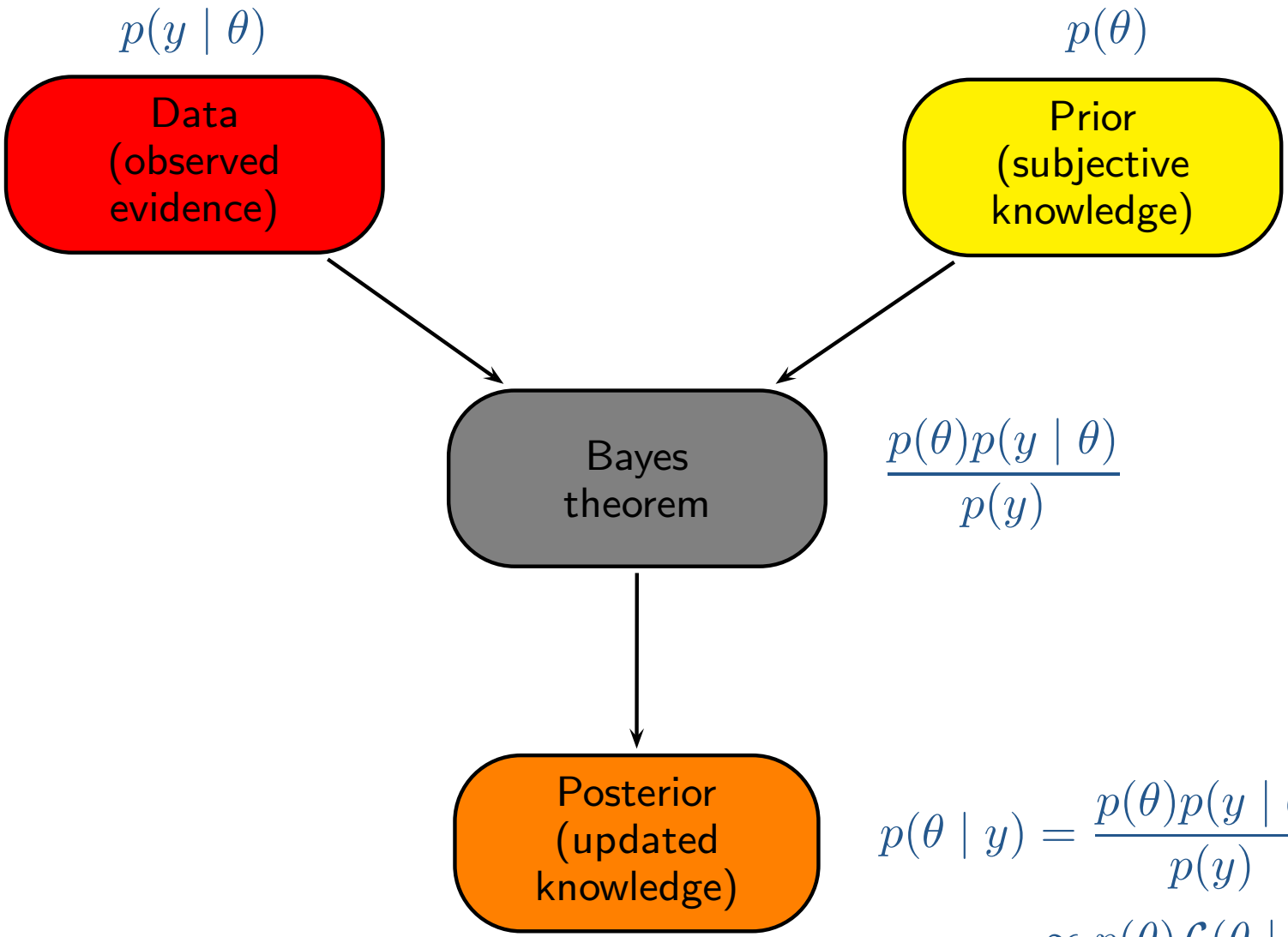
$p(\theta)$

Prior (subjective knowledge)

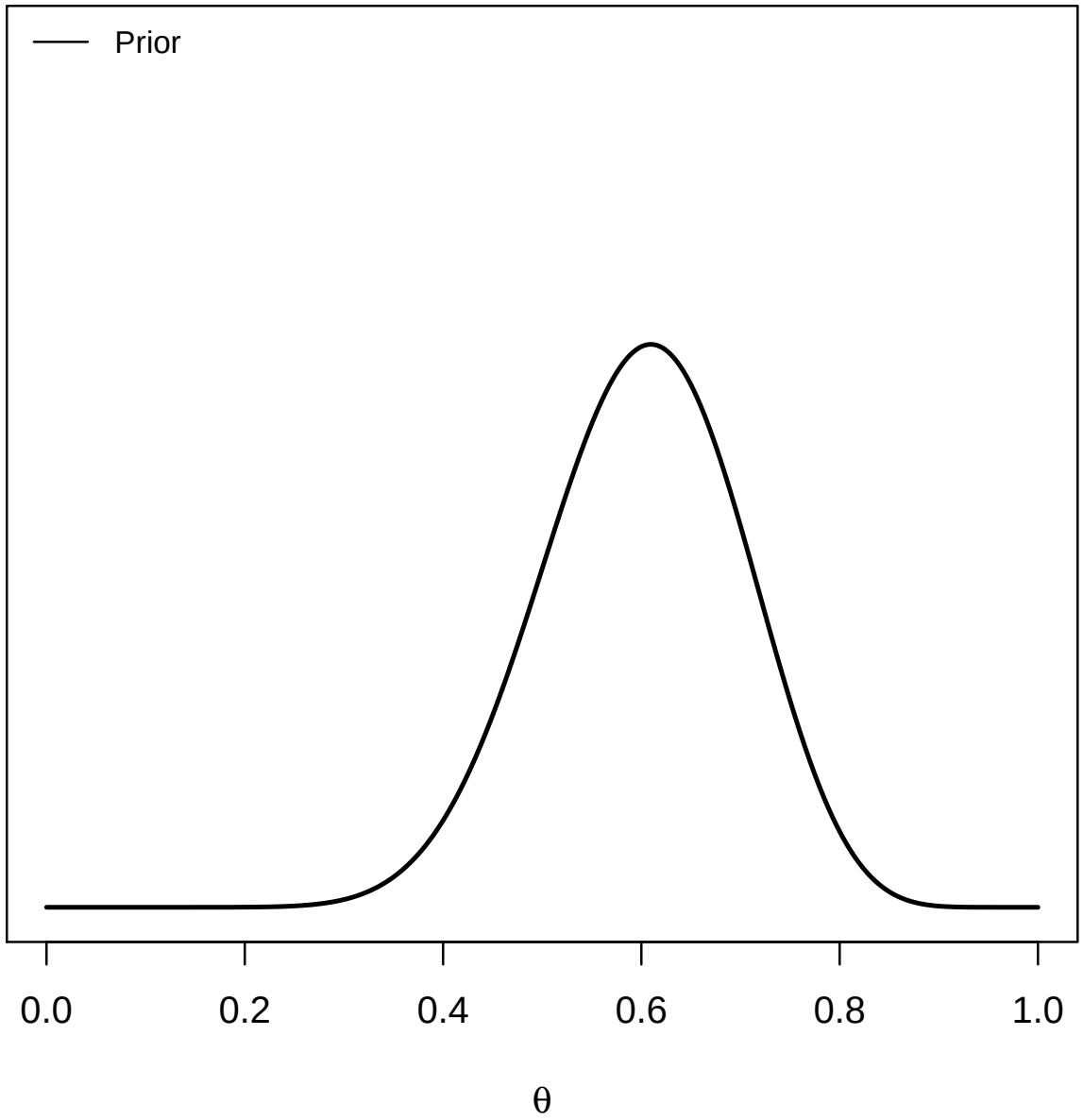
Bayesian inference (cont'd)



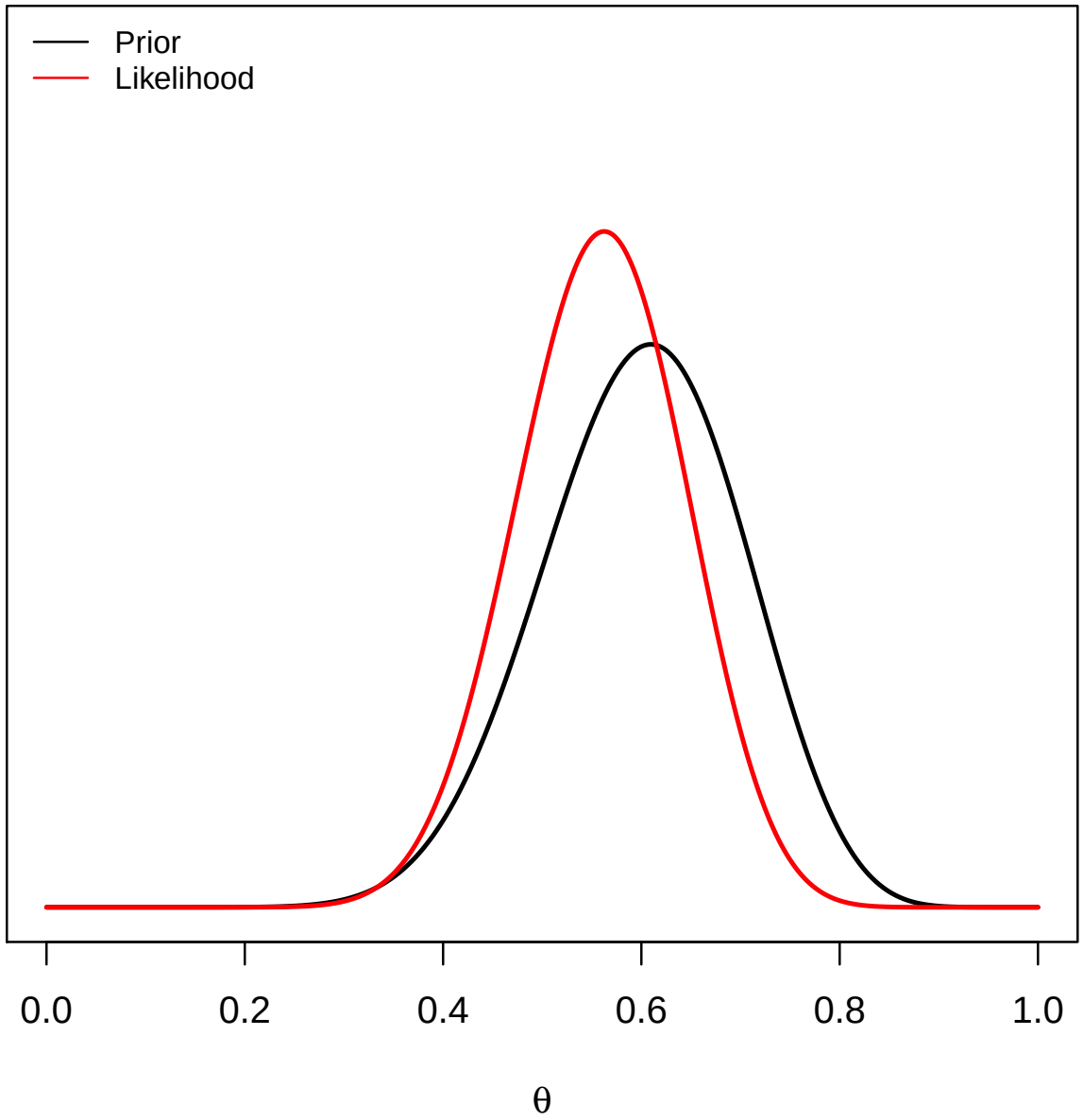
Bayesian inference (cont'd)



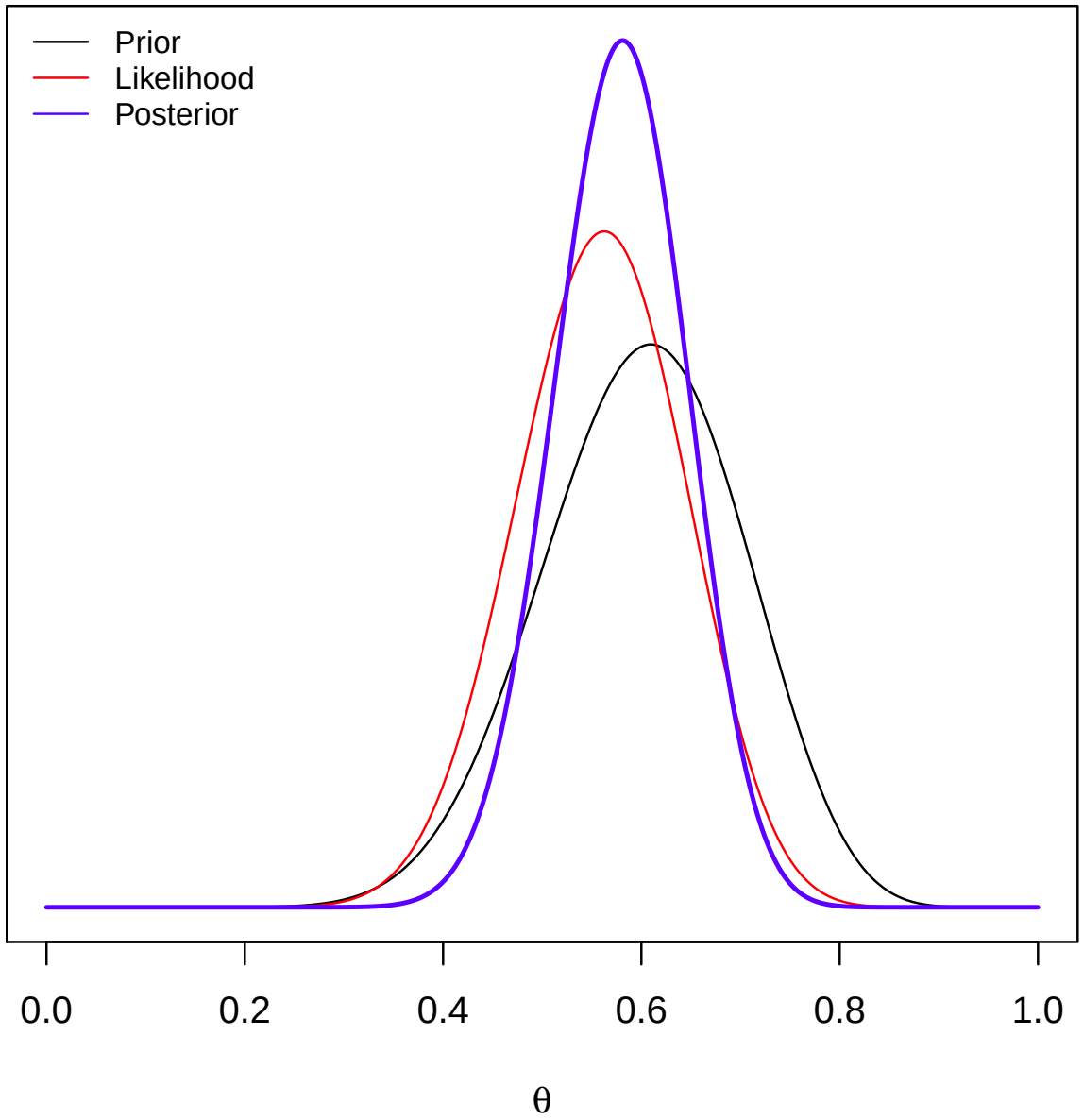
Bayesian inference — updating knowledge



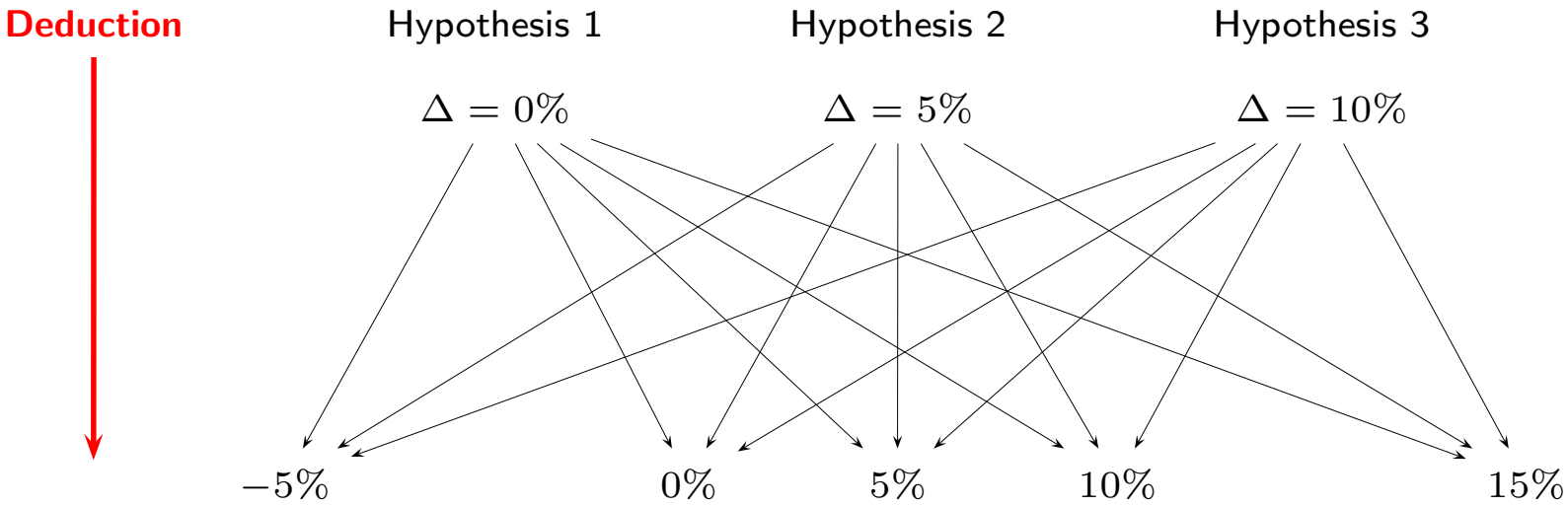
Bayesian inference — updating knowledge



Bayesian inference — updating knowledge

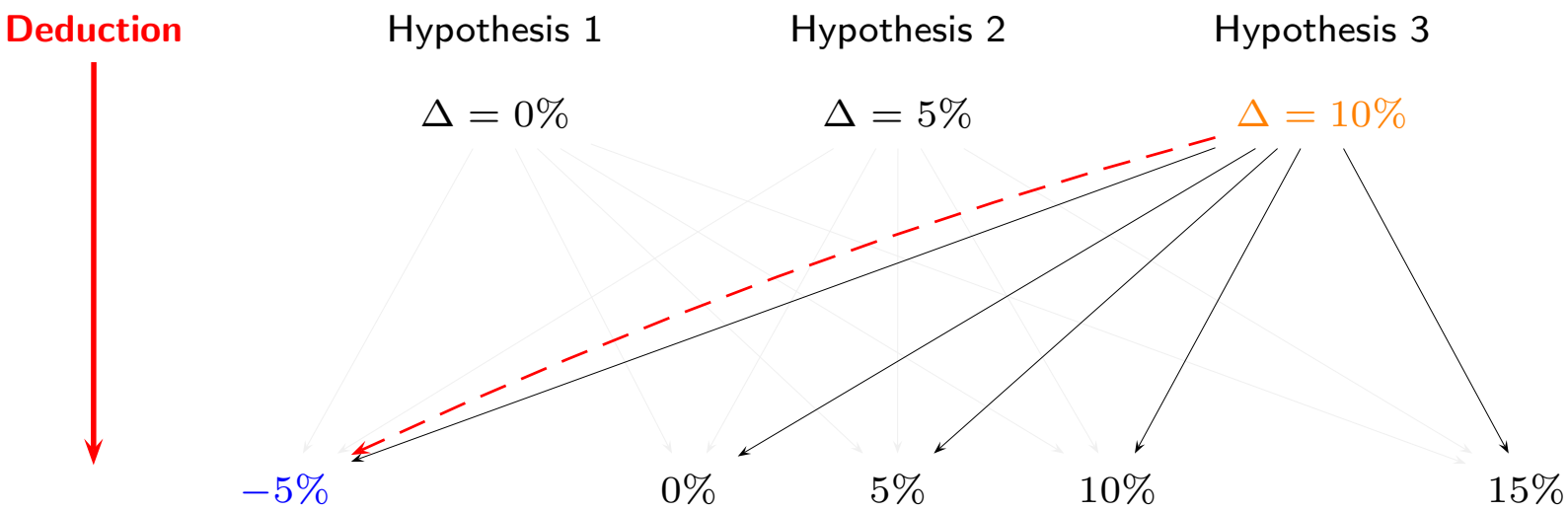


Deductive vs inductive inference



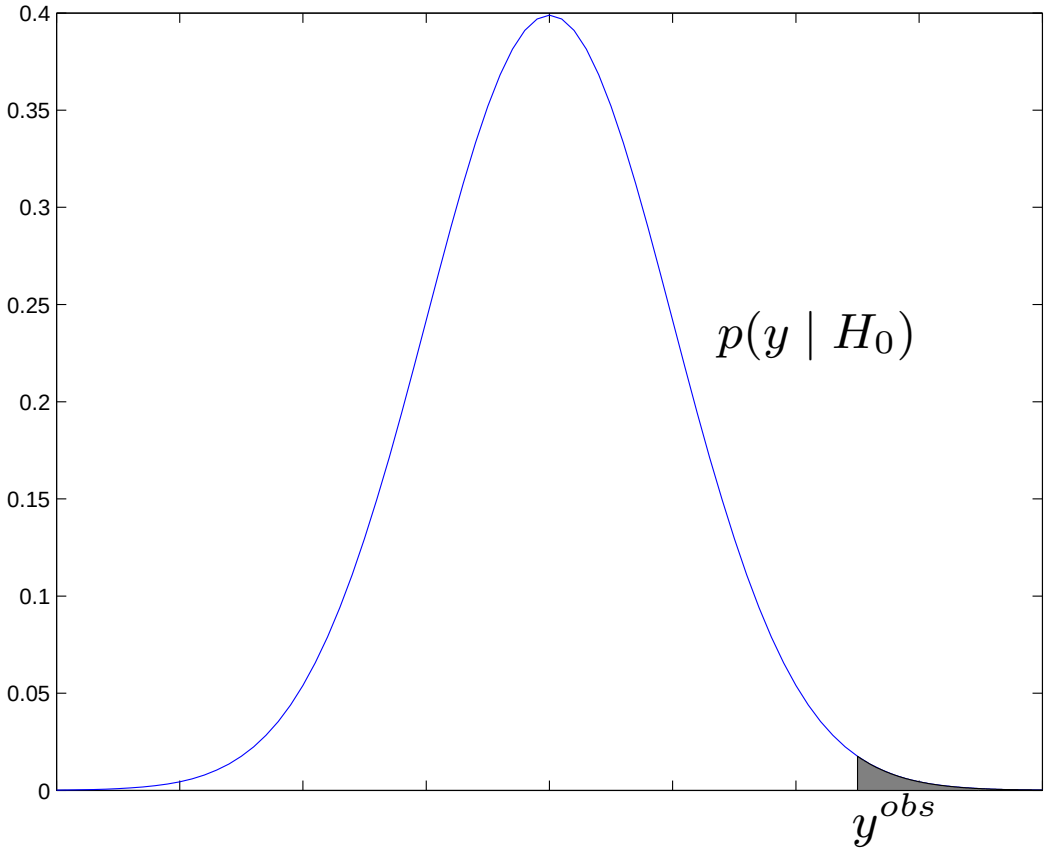
- Standard (frequentist) procedures fix the working hypotheses and, **by deduction**, make inference on the observed data:
  - If my hypothesis is true, what is the probability of randomly selecting the data that I actually observed? If small, then *deduce* weak support of the evidence to the hypothesis

Deductive vs inductive inference



- Standard (frequentist) procedures fix the working hypotheses and, **by deduction**, make inference on the observed data:
  - If my hypothesis is true, what is the probability of randomly selecting the data that I actually observed? If small, then *deduce* weak support of the evidence to the hypothesis
  - Assess $\Pr(\text{Observed data} \mid \text{Hypothesis})$

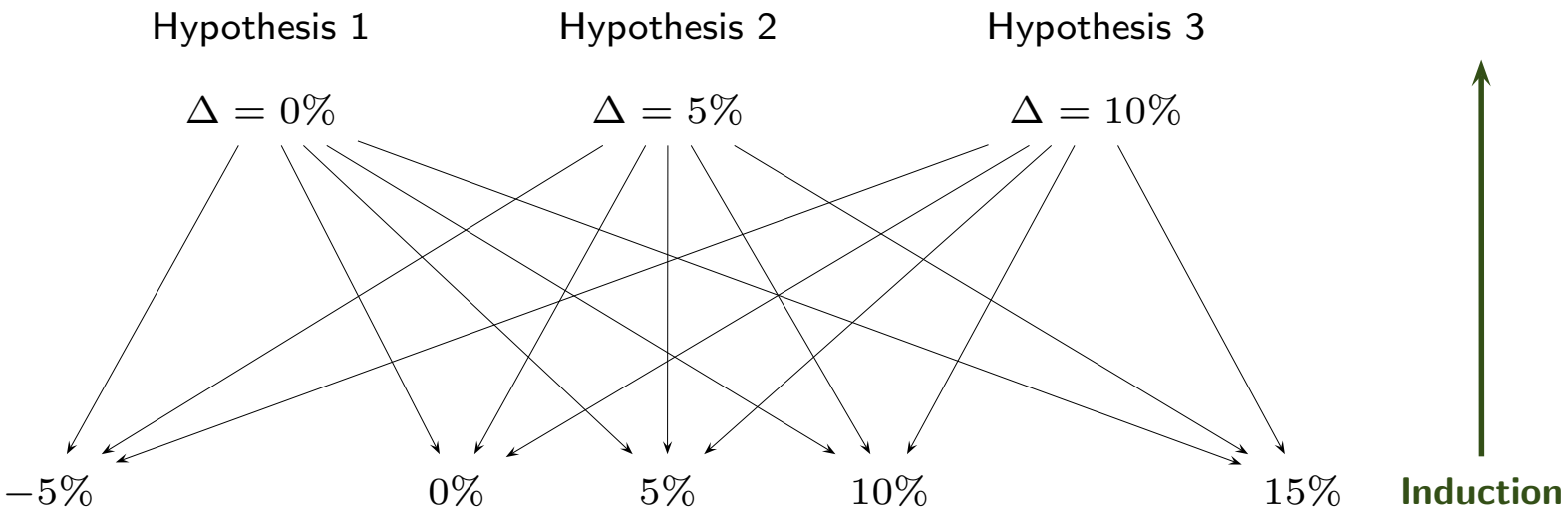
Deductive vs inductive inference



Fisher's interpretation of p-value P (grey area):

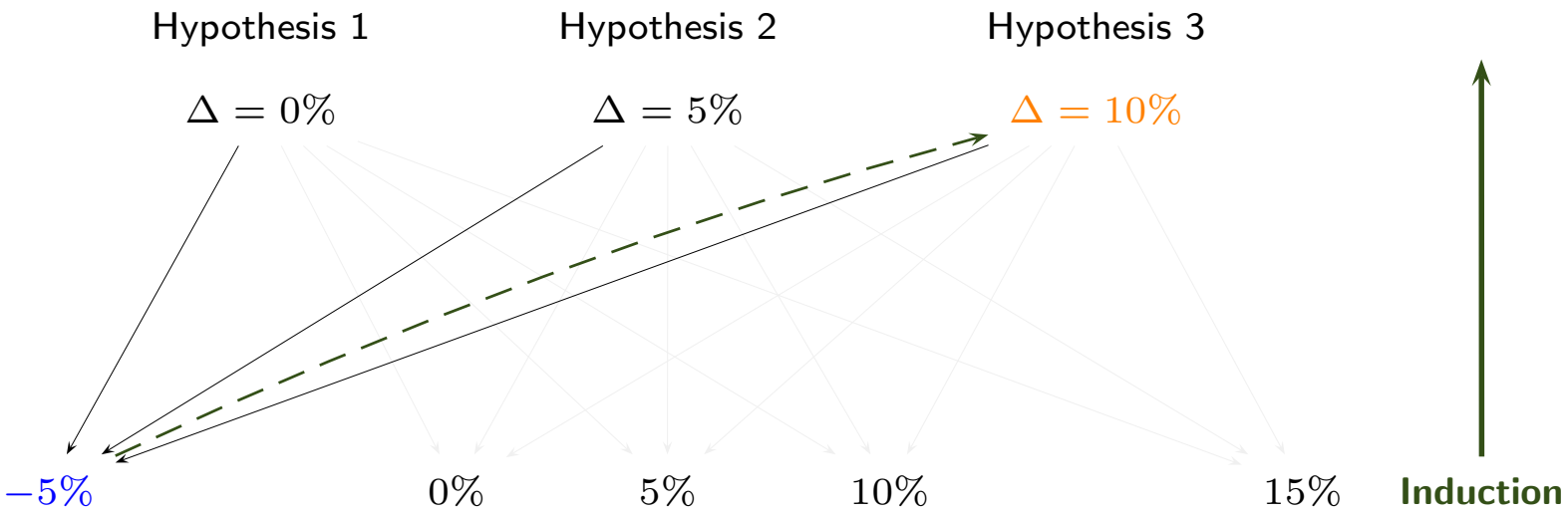
- If $P < 0.01 \Rightarrow$ **strong** evidence against H_0
- If $0.01 < P < 0.05 \Rightarrow$ **fairly strong** evidence against H_0
- If $P > 0.05 \Rightarrow$ **little or no** evidence against H_0

Deductive vs inductive inference



- On the contrary, the Bayesian philosophy proceeds fixing the value of the observed data and, **by induction**, makes inference on unobservable hypotheses
  - What is the probability of my hypothesis, given the data I observed? If less than the probability of other competing hypotheses, then weak support of the evidence to the hypothesis

Deductive vs inductive inference



- On the contrary, the Bayesian philosophy proceeds fixing the value of the observed data and, **by induction**, makes inference on unobservable hypotheses
  - What is the probability of my hypothesis, given the data I observed? If less than the probability of other competing hypotheses, then weak support of the evidence to the hypothesis
  - Assess $\Pr(\text{Hypothesis} \mid \text{Observed data})$

Bayesian inference (cont'd)

The Bayesian procedure allows a straightforward **sequential update** of the evidence

Bayesian inference (cont'd)

The Bayesian procedure allows a straightforward **sequential update** of the evidence

- Define a prior distribution on θ: $p(\theta)$

Bayesian inference (cont'd)

The Bayesian procedure allows a straightforward **sequential update** of the evidence

- Define a prior distribution on θ: $p(\theta)$
- Observe the available data y and update the uncertainty about the parameter into the posterior distribution: $p(\theta \mid y)$

Bayesian inference (cont'd)

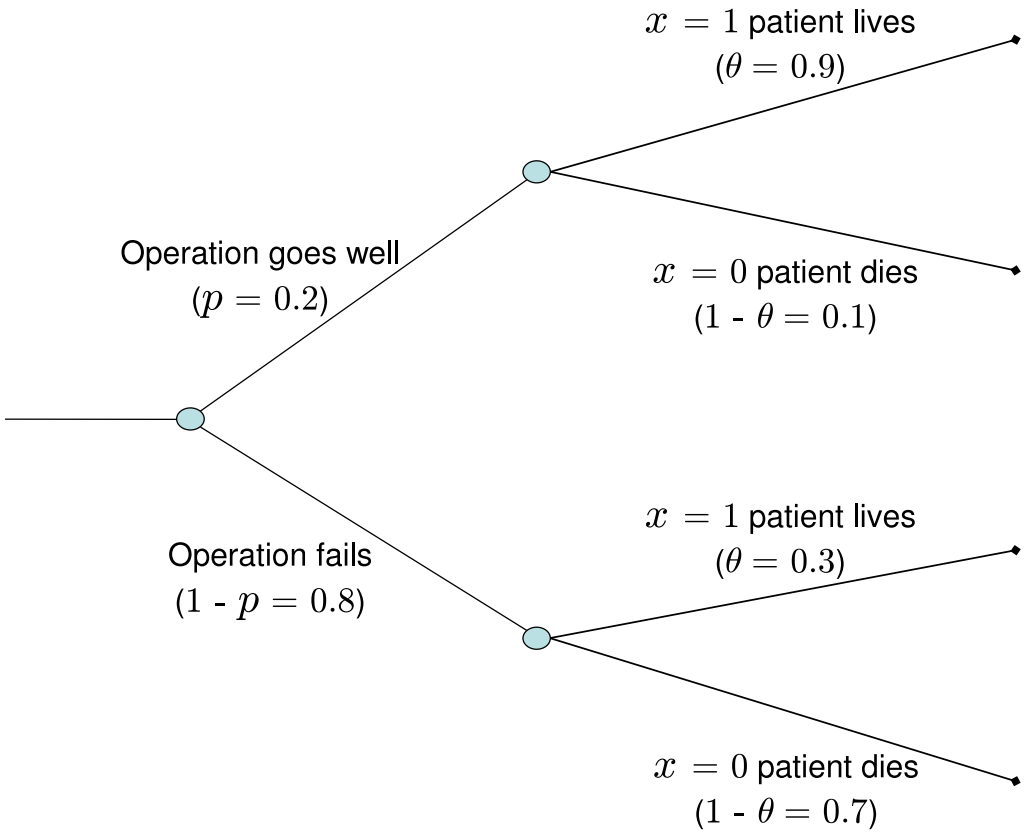
The Bayesian procedure allows a straightforward **sequential update** of the evidence

- Define a prior distribution on θ: $p(\theta)$
- Observe the available data y and update the uncertainty about the parameter into the posterior distribution: $p(\theta \mid y)$
- If further ("similar") evidence z is made available, it is possible to integrate it in the updating process, using the posterior distribution given y as the new prior:

$$p(\theta \mid y, z) = \frac{p(z \mid \theta, y)p(\theta \mid y)}{p(z \mid y)} \propto p(z \mid \theta, y)p(\theta \mid y)$$

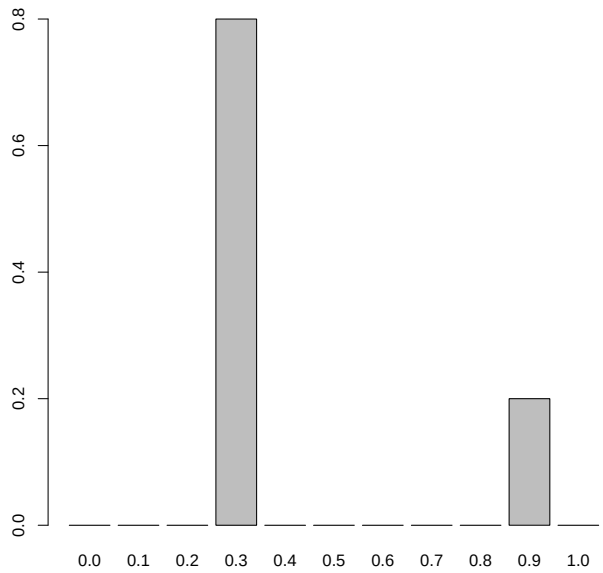
"Today's posterior is tomorrow's prior"

Bayesian analysis — Example



- In this situation, our prior knowledge on θ can be encoded in the probability distribution

$$\theta = \begin{cases} 0.9 & \text{with probability } 0.2 \\ 0.3 & \text{with probability } 0.8 \end{cases}$$



Example (cont'd)

- Suppose we see $n = 10$ patients and a total of $y = 7$ of them are alive at 1 week
- Assuming again a Binomial model for the number of "successes", we then have:
  Model for the observed data: $p(y \mid \theta) \propto \theta^y (1-\theta)^{(n-y)}$
  Prior distribution: $p(\theta = 0.9) = 0.2, \quad p(\theta = 0.3) = 0.8$

Example (cont'd)

- Suppose we see $n = 10$ patients and a total of $y = 7$ of them are alive at 1 week
- Assuming again a Binomial model for the number of "successes", we then have:
  Model for the observed data: $p(y \mid \theta) \propto \theta^y (1-\theta)^{(n-y)}$
  Prior distribution: $p(\theta = 0.9) = 0.2, \quad p(\theta = 0.3) = 0.8$

- The likelihood for the two possible values of θ is then
  $\mathcal{L}(\theta = 0.9 \mid y = 7, n = 10) = 0.9^7 \times 0.1^3 = 0.00047$;
  $\mathcal{L}(\theta = 0.3 \mid y = 7, n = 10) = 0.3^7 \times 0.7^3 = 0.00007$

Example (cont'd)

- Suppose we see $n = 10$ patients and a total of $y = 7$ of them are alive at 1 week
- Assuming again a Binomial model for the number of "successes", we then have:
  Model for the observed data: $p(y \mid \theta) \propto \theta^y (1-\theta)^{(n-y)}$
  Prior distribution: $p(\theta = 0.9) = 0.2, \quad p(\theta = 0.3) = 0.8$

- The likelihood for the two possible values of θ is then
  $\mathcal{L}(\theta = 0.9 \mid y = 7, n = 10) = 0.9^7 \times 0.1^3 = 0.00047$;
  $\mathcal{L}(\theta = 0.3 \mid y = 7, n = 10) = 0.3^7 \times 0.7^3 = 0.00007$

- Finally, combining the likelihood with the prior we get the posterior distribution

$$p(\theta = 0.9 \mid y = 7, n = 10) = \frac{\overset{\mathcal{L}(\theta=0.9|y)}{0.00047} \times \overset{p(\theta=0.9)}{0.2}}{(\underset{\mathcal{L}(\theta=0.9|y)}{0.00047} \times \underset{p(\theta=0.9)}{0.2}) + (\underset{\mathcal{L}(\theta=0.3|y)}{0.00007} \times \underset{p(\theta=0.3)}{0.8})} = 0.626$$

$$p(\theta = 0.3 \mid y = 7, n = 10) = \frac{\overset{\mathcal{L}(\theta=0.3|y)}{0.00007} \times \overset{p(\theta=0.3)}{0.8}}{(\underset{\mathcal{L}(\theta=0.9|y)}{0.00047} \times \underset{p(\theta=0.9)}{0.2}) + (\underset{\mathcal{L}(\theta=0.3|y)}{0.00007} \times \underset{p(\theta=0.3)}{0.8})} = 0.374$$

(where the denominator is $p(y)$, the product of likelihood and prior, summed over all possible values of θ)

Choice of the prior distribution

- **Non-informative prior**
  - Attempts to include minimal information in the prior to "let the data speak for themselves" (sometimes known as "minimally informative")
  - Need to be careful in defining the scale in which non-informativeness is selected
  - Sometimes helpful as preliminary approximation — often leads to essentially the same inference as using maximum likelihood

- **Conjugate prior**
  - Convenient mathematical formulation
  - Prior and posterior in the same family
  - E.g. Prior $=$ Normal(m_0, s_0) + Data $=$ Normal$(\mu, \sigma^2) \Rightarrow$
  - Posterior $=$ Normal(m_1, s_1)
  - Can formally express (subjective) knowledge and include prior information

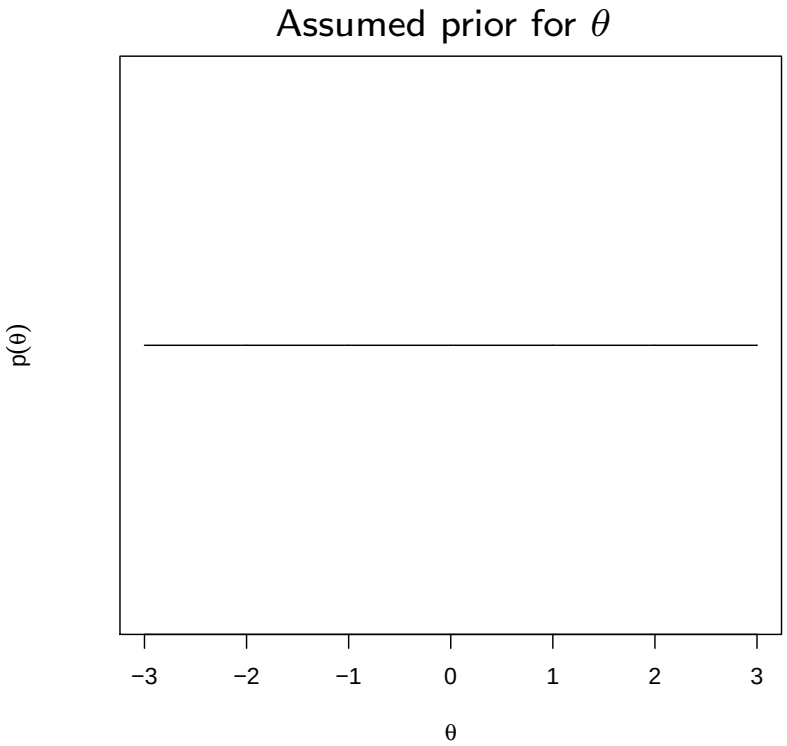
- **Non-conjugated prior**
  - Overcome limitations of conjugate priors, i.e.
    - Too restrictive
    - Not available for widely used models (e.g. logistic regression)
  - More difficult to handle computationally, so needs to resort to simulation-based methods (e.g. MCMC) or clever approximations (e.g. INLA/ABC)

"Non informative" priors

- "Ignorance" on θ should imply ignorance on any function of θ. Unfortunately, non informative prior distributions are sensitive to changes of scale
- For example, suppose we consider θ, the log-odds ratio in a logistic regression model, and assume $p(\theta) = k$. Typically, we are interested in the transformation $\phi = \exp(\theta)$, that represents the odds ratio on the natural scale

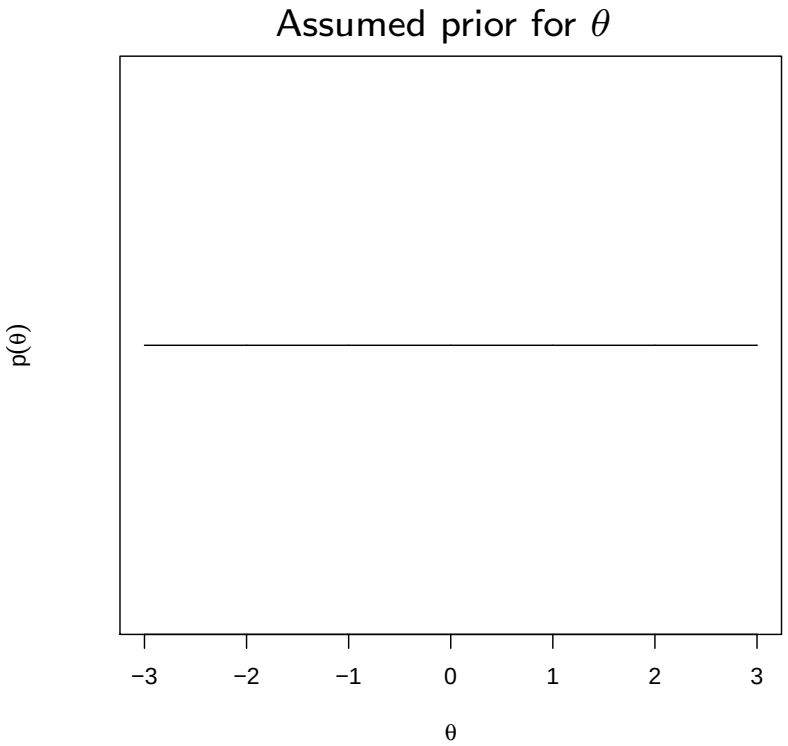
"Non informative" priors

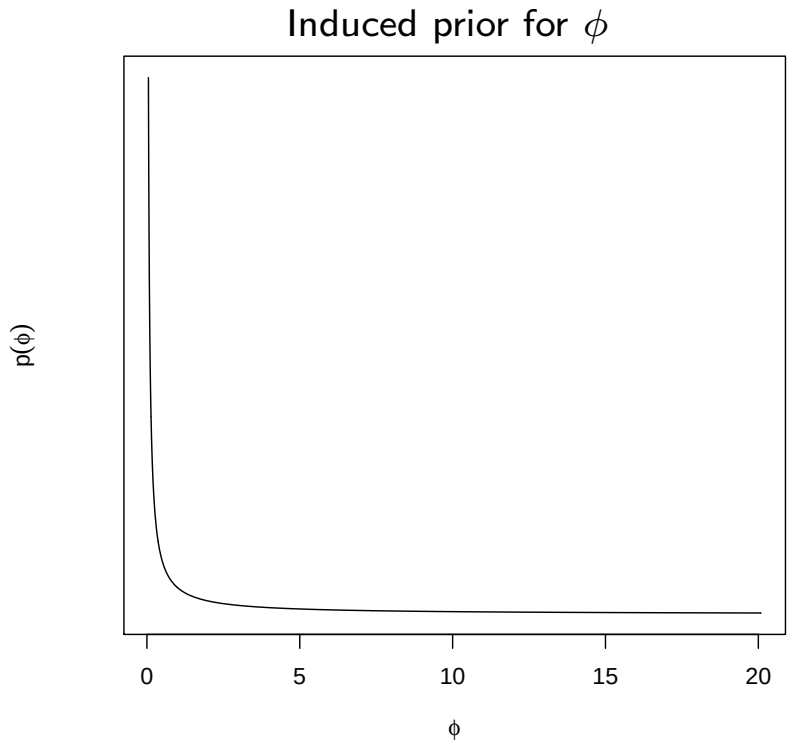
- "Ignorance" on θ should imply ignorance on any function of θ. Unfortunately, non informative prior distributions are sensitive to changes of scale
- For example, suppose we consider θ, the log-odds ratio in a logistic regression model, and assume $p(\theta) = k$. Typically, we are interested in the transformation $\phi = \exp(\theta)$, that represents the odds ratio on the natural scale



"Non informative" priors

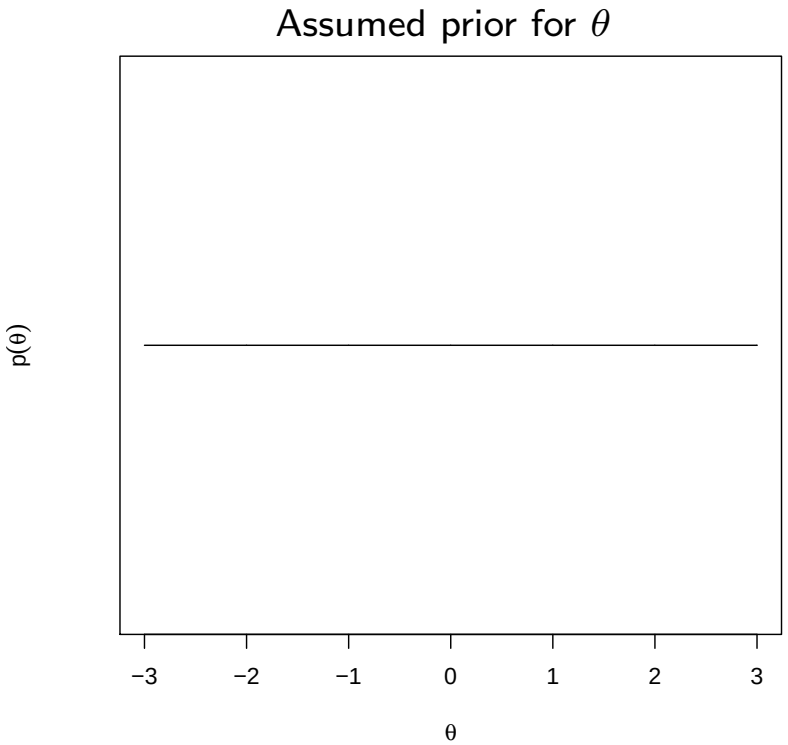
- "Ignorance" on θ should imply ignorance on any function of θ. Unfortunately, non informative prior distributions are sensitive to changes of scale
- For example, suppose we consider θ, the log-odds ratio in a logistic regression model, and assume $p(\theta) = k$. Typically, we are interested in the transformation $\phi = \exp(\theta)$, that represents the odds ratio on the natural scale

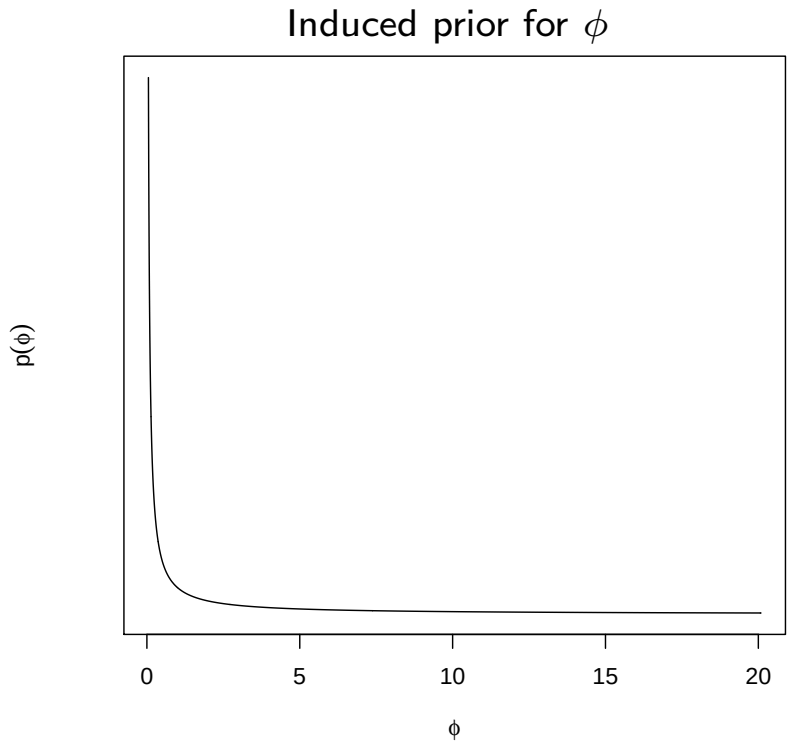




"Non informative" priors

- "Ignorance" on θ should imply ignorance on any function of θ. Unfortunately, non informative prior distributions are sensitive to changes of scale
- For example, suppose we consider θ, the log-odds ratio in a logistic regression model, and assume $p(\theta) = k$. Typically, we are interested in the transformation $\phi = \exp(\theta)$, that represents the odds ratio on the natural scale





- The assumed prior ignorance on θ turns out to be extremely informative on ϕ. So what formulation should one use?

Choice of the prior distribution

- **Non-informative prior**
  - Attempts to include minimal information in the prior to "let the data speak for themselves" (sometimes known as "minimally informative")
  - Need to be careful in defining the scale in which non-informativeness is selected
  - Sometimes helpful as preliminary approximation — often leads to essentially the same inference as using maximum likelihood

- **Conjugate prior**
  - Convenient mathematical formulation
  - Prior and posterior in the same family
    E.g. Prior $=$ Normal(m_0, s_0) + Data $=$ Normal$(\mu, \sigma^2) \Rightarrow$
    Posterior $=$ Normal(m_1, s_1)
  - Can formally include prior information in the definition

- **Non-conjugated prior**
  - Overcome limitations of conjugate priors, i.e.
    - Too restrictive
    - Not available for widely used models (e.g. logistic regression)
  - More difficult to handle computationally, so needs to resort to simulation-based methods (e.g. MCMC) or clever approximations (e.g. INLA/ABC)

"Informative" conjugate prior

- Consider again the cancer screening example

$$y \mid \theta, n \sim \text{Binomial}(\theta, n)$$

and suppose further that (for instance from previous similar studies) we know that the probability that a patient returns their kit has been estimated between 20 and 60%, a condition that we can represent by assuming

  - $\text{mean}(\theta) = 0.4$
  - $\text{sd}(\theta) = 0.1$

"Informative" conjugate prior

- Consider again the cancer screening example

$$y \mid \theta, n \sim \text{Binomial}(\theta, n)$$

and suppose further that (for instance from previous similar studies) we know that the probability that a patient returns their kit has been estimated between 20 and 60%, a condition that we can represent by assuming

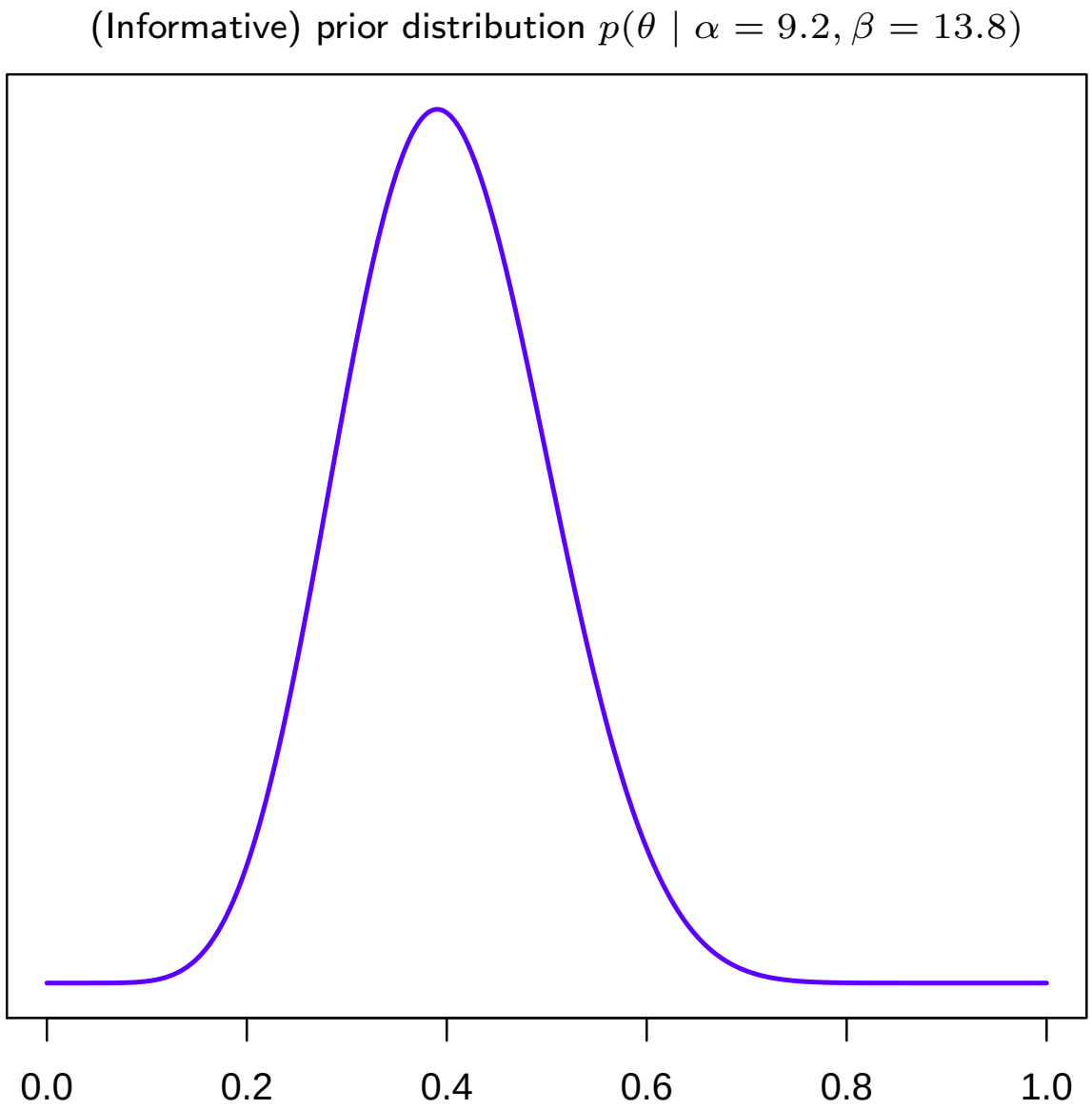
  - $\text{mean}(\theta) = 0.4$
  - $\text{sd}(\theta) = 0.1$

- We can **encode** this information into a suitable prior distribution. One possibility is to model the prior using a **Beta** distribution

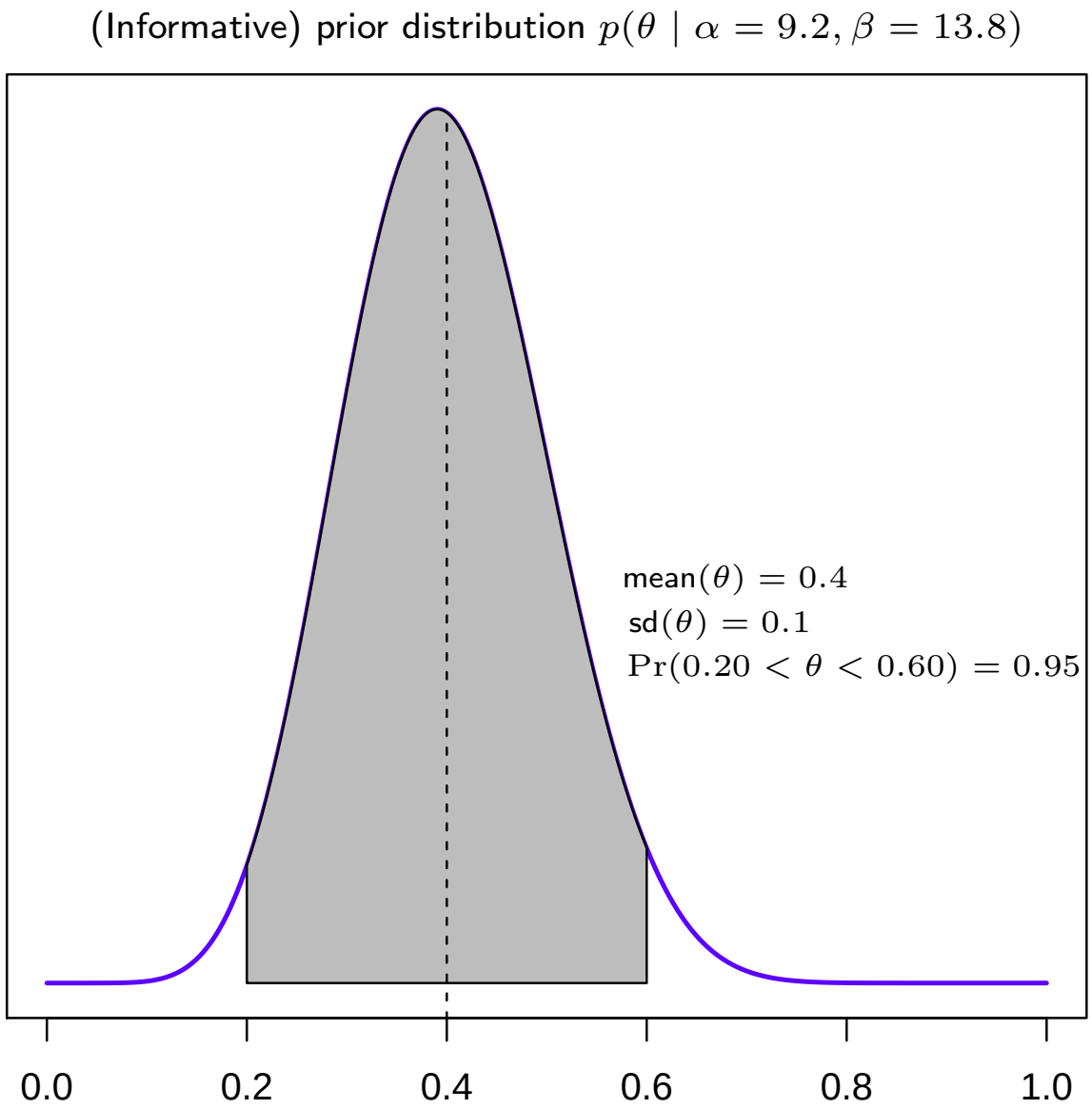
$$\theta \mid \alpha, \beta \sim \text{Beta}(\alpha, \beta)$$

(the values α and β are called *hyper-parameters*. Upon varying them, we obtain different forms for the prior)

"Informative" conjugate prior (cont'd)

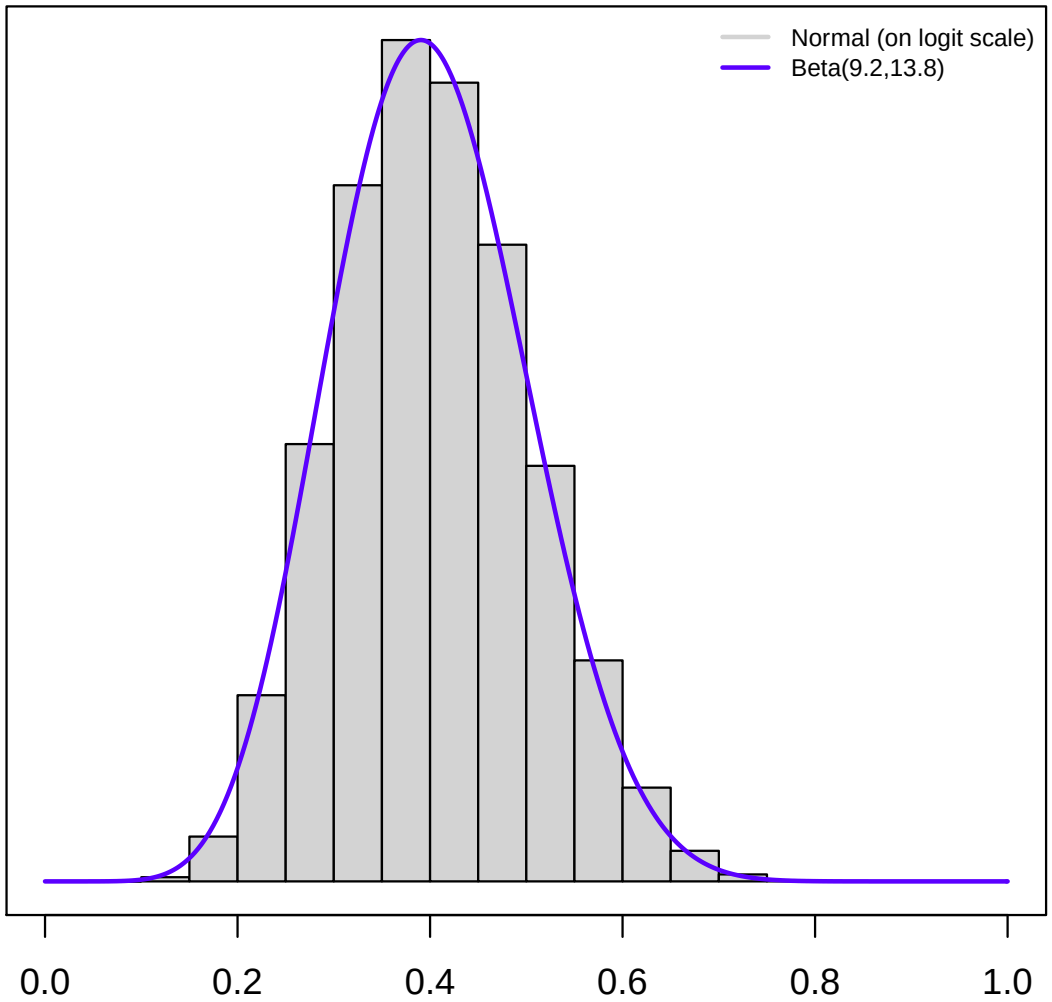


"Informative" conjugate prior (cont'd)



Prior **information** vs prior **distribution**

- **NB**: Using a Beta distribution is only **one** possibility! There are different ways of encoding the prior knowledge
  - For example, we could model $\psi = \text{logit}(\theta) \sim \text{Normal}(-0.41, 0.43)$, which effectively implies the same prior information!



"Informative" conjugate prior (cont'd)

- Apart from being extremely versatile, the Beta distribution is also conjugated for the Binomial model
- Consequently,

$$\theta \mid \alpha, \beta \sim \textbf{Beta}(\alpha, \beta) = \text{Beta}(9.2, 13.8)$$
$$y \mid \theta \sim \text{Binomial}(\theta, n)$$

then $\theta \mid y \sim \textbf{Beta}(\alpha^*, \beta^*)$

where

- $\alpha^* = (\alpha + y)$,
- $\beta^* = (n + \beta - y)$

"Informative" conjugate prior (cont'd)

- Apart from being extremely versatile, the Beta distribution is also conjugated for the Binomial model
- Consequently,

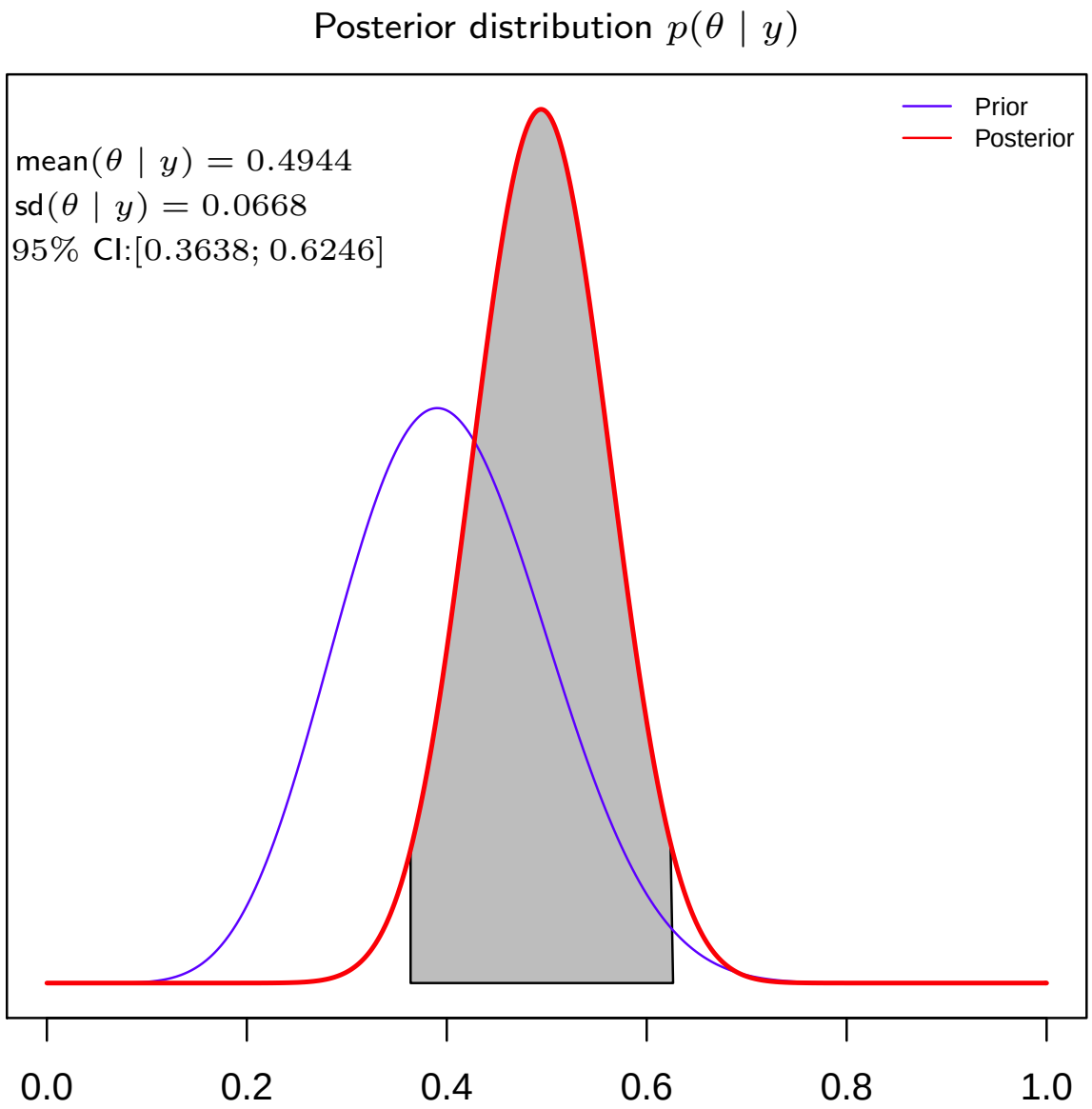
$$\theta \mid \alpha, \beta \sim \textbf{Beta}(\alpha, \beta) = \text{Beta}(9.2, 13.8)$$
$$y \mid \theta \sim \text{Binomial}(\theta, n)$$

then $\theta \mid y \sim \textbf{Beta}(\alpha^*, \beta^*)$

where
  - $\alpha^* = (\alpha + y)$,
  - $\beta^* = (n + \beta - y)$

- If we observe that $n = 32$ kits are sent out and $y = 18$ are returned, the posterior distribution then becomes $\textbf{Beta}(\alpha^*, \beta^*) = \textbf{Beta}(27.2, 27.8)$
- **NB**: Since the distributional form is known, it is easy to characterise the posterior (i.e. compute mean, sd, ...)

"Informative" conjugate prior (cont'd)



Bayesian vs "standard" inference

"Standard" analysis

- MLE $\hat{\theta} = \frac{y}{n} = \frac{18}{32} = 0.5625$
- Standard error $= \text{se}(\hat{\theta}) = \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} = \sqrt{\frac{0.5625 \times 0.4375}{32}} = 0.0877$
- 95% confidence interval:
  $\hat{\theta} \pm 1.96 \times \text{se}(\hat{\theta}) = 0.5625 \pm 1.96 \times 0.0877 = [0.3096 - 0.7344]$

Bayesian vs "standard" inference

"Standard" analysis

- MLE $\hat{\theta} = \frac{y}{n} = \frac{18}{32} = 0.5625$
- Standard error $= \text{se}(\hat{\theta}) = \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} = \sqrt{\frac{0.5625 \times 0.4375}{32}} = 0.0877$
- 95% confidence interval:
  $\hat{\theta} \pm 1.96 \times \text{se}(\hat{\theta}) = 0.5625 \pm 1.96 \times 0.0877 = [0.3096 - 0.7344]$

Bayesian analysis

- Prior mean for $\theta = 0.4$;
  prior 95% "credibility" interval $= [0.2 - 0.6]$
- Posterior mean for $\theta = 0.4944$;
  posterior 95% "credibility" interval $= [0.3638 - 0.6246]$

Bayesian vs "standard" inference

"Standard" analysis

- MLE $\hat{\theta} = \frac{y}{n} = \frac{18}{32} = 0.5625$
- Standard error $= \text{se}(\hat{\theta}) = \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} = \sqrt{\frac{0.5625 \times 0.4375}{32}} = 0.0877$
- 95% confidence interval:
  $\hat{\theta} \pm 1.96 \times \text{se}(\hat{\theta}) = 0.5625 \pm 1.96 \times 0.0877 = [0.3096 - 0.7344]$

Bayesian analysis

- Prior mean for $\theta = 0.4$;
  prior 95% "credibility" interval $= [0.2 - 0.6]$
- Posterior mean for $\theta = 0.4944$;
  posterior 95% "credibility" interval $= [0.3638 - 0.6246]$

The standard results are different from the Bayesian estimates, because they do not take into account the existing information about the value of the parameter, coming from previous studies

Choice of the prior distribution

- **Non-informative prior**
  - Attempts to include minimal information in the prior to "let the data speak for themselves" (sometimes known as "minimally informative")
  - Need to be careful in defining the scale in which non-informativeness is selected
  - Sometimes helpful as preliminary approximation — often leads to essentially the same inference as using maximum likelihood

- **Conjugate prior**
  - Convenient mathematical formulation
  - Prior and posterior in the same family
    E.g. Prior $=$ Normal(m_0, s_0) + Data $=$ Normal$(\mu, \sigma^2) \Rightarrow$

- **Non-conjugated prior**
  - Overcome limitations of conjugate priors, i.e.
    - Too restrictive
    - Not available for widely used models (e.g. logistic regression)
  - More difficult to handle computationally, so needs to resort to simulation-based methods (e.g. MCMC) or clever approximations (e.g. INLA/ABC)

Example — multiple regression

- Consider a problem where
  - y_i is a scalar reference measurement, e.g. of a protein
  - $\boldsymbol{x}_i = (x_{1i}, \ldots, x_{pi})$ is a vector with spectral data, e.g. scores on PCs or PLS factors

Example — multiple regression

- Consider a problem where
  - y_i is a scalar reference measurement, e.g. of a protein
  - $\boldsymbol{x}_i = (x_{1i}, \ldots, x_{pi})$ is a vector with spectral data, e.g. scores on PCs or PLS factors
- Typically, we observe a training dataset $\mathsf{TD} = (y, \boldsymbol{x})$ with n cases and we want to use that to predict y_{n+1} for a new case with a given spectrum $\boldsymbol{x}_{n+1}$

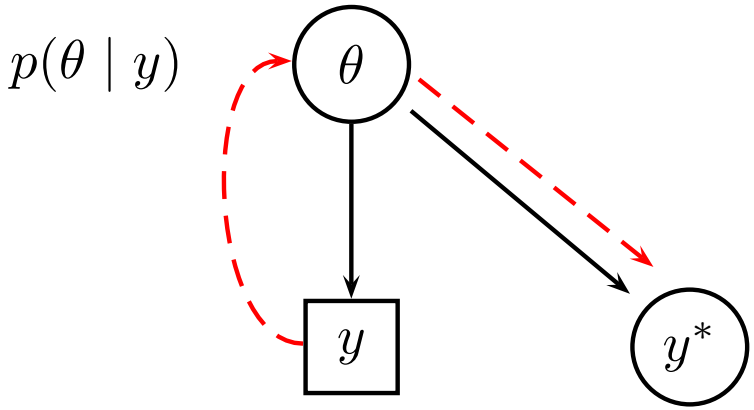
Example — multiple regression

- Consider a problem where
  - y_i is a scalar reference measurement, e.g. of a protein
  - $\boldsymbol{x}_i = (x_{1i}, \ldots, x_{pi})$ is a vector with spectral data, e.g. scores on PCs or PLS factors
- Typically, we observe a training dataset $\mathrm{TD} = (y, \boldsymbol{x})$ with n cases and we want to use that to predict y_{n+1} for a new case with a given spectrum $\boldsymbol{x}_{n+1}$
- The standard approach is to regress y on $\boldsymbol{x}$ for the TD:
  1. Fit the model $y = \boldsymbol{x}\boldsymbol{\beta} + e$, e.g. by least squares
  2. Get an estimate $\boldsymbol{b} = (b_1, \ldots, b_p)$ of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$
  3. predict y_{n+1} by $\boldsymbol{x}_{n+1}\boldsymbol{b}$

Example — multiple regression

- Consider a problem where
  - y_i is a scalar reference measurement, e.g. of a protein
  - $\boldsymbol{x}_i = (x_{1i}, \ldots, x_{pi})$ is a vector with spectral data, e.g. scores on PCs or PLS factors
- Typically, we observe a training dataset $\mathrm{TD} = (y, \boldsymbol{x})$ with n cases and we want to use that to predict y_{n+1} for a new case with a given spectrum $\boldsymbol{x}_{n+1}$
- The standard approach is to regress y on $\boldsymbol{x}$ for the TD:
  1. Fit the model $y = \boldsymbol{x}\boldsymbol{\beta} + e$, e.g. by least squares
  2. Get an estimate $\boldsymbol{b} = (b_1, \ldots, b_p)$ of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$
  3. predict y_{n+1} by $\boldsymbol{x}_{n+1}\boldsymbol{b}$
- **NB**: This model can be considered as a "special case" of a wider class of Bayesian specifications

Exchangeability and prediction



$$p(y^* \mid y) = \int p(y^* \mid \theta) p(\theta \mid y) \mathrm{d}\theta$$

- The assumption of exchangeability implies that $y_{n+1} \mid \boldsymbol{x}_{n+1}$ has the same distribution as $y \mid \boldsymbol{x}$ for the TD
- In other words, we regard the $n+1$-th observation as "similar" to (or, in statistical parlance, **exchangeable** with) the n previously observed

Example — multiple regression (cont'd)

- In the Bayesian version of this model, we need to specify a prior distribution for the parameters
- We can obtain the same results as the frequentist version by assuming very vague information on each regression coefficient, e.g.
  $\beta_j \overset{iid}{\sim} \text{Uniform}(-\infty, \infty)$, which implies $p(\beta_j) = k$ for $j = 1, \ldots, p$

Example — multiple regression (cont'd)

- In the Bayesian version of this model, we need to specify a prior distribution for the parameters
- We can obtain the same results as the frequentist version by assuming very vague information on each regression coefficient, e.g.
  $\beta_j \overset{iid}{\sim} \text{Uniform}(-\infty, \infty)$, which implies $p(\beta_j) = k$ for $j = 1, \ldots, p$
- Applying Bayes theorem we can compute the posterior distribution as

$$\begin{aligned} p(\boldsymbol{\beta} \mid y, \boldsymbol{x}) &= \frac{p(y \mid \boldsymbol{\beta}, \boldsymbol{x}) p(\boldsymbol{\beta})}{p(y)} \\ &\propto \frac{k \times p(y \mid \boldsymbol{\beta}, \boldsymbol{x})}{\int k \times p(y \mid \boldsymbol{\beta}, \boldsymbol{x}) \mathrm{d}\boldsymbol{\beta}} \\ &= \frac{p(y \mid \boldsymbol{\beta}, \boldsymbol{x})}{\int p(y \mid \boldsymbol{\beta}, \boldsymbol{x}) \mathrm{d}\boldsymbol{\beta}} = \frac{\mathcal{L}(\boldsymbol{\beta} \mid \text{TD})}{\int \mathcal{L}(\boldsymbol{\beta} \mid \text{TD}) \mathrm{d}\boldsymbol{\beta}} \end{aligned}$$

  which is just the (**normalised**) likelihood
- Consequently, **the mean of the posterior is identical with the maximum likelihood estimator**

Example — multiple regression (cont'd)

- Assuming exchangeability between the TD and the new observation, we can obtain the **posterior predictive distribution**

$$p(y_{n+1} \mid \boldsymbol{x}_{n+1}, \mathsf{TD}) = \int p(y_{n+1} \mid \boldsymbol{\beta}, \boldsymbol{x}_{n+1}) p(\boldsymbol{\beta} \mid \mathsf{TD}) \mathrm{d}\boldsymbol{\beta}$$

- Under the assumptions specified in this case, **the mean of the predictive distribution is $\boldsymbol{x}_{n+1}\boldsymbol{b}$, which is equivalent to the maximum likelihood estimation**

Example — multiple regression (cont'd)

- Assuming exchangeability between the TD and the new observation, we can obtain the **posterior predictive distribution**

$$p(y_{n+1} \mid \boldsymbol{x}_{n+1}, \text{TD}) = \int p(y_{n+1} \mid \boldsymbol{\beta}, \boldsymbol{x}_{n+1}) p(\boldsymbol{\beta} \mid \text{TD}) \mathrm{d}\boldsymbol{\beta}$$

- Under the assumptions specified in this case, **the mean of the predictive distribution is $\boldsymbol{x}_{n+1}\boldsymbol{b}$, which is equivalent to the maximum likelihood estimation**

- In this sense, the ML analysis is a special case of the general Bayesian procedure
- But of course there is no reason why we have to use the "minimally informative", vague Uniform prior on $\boldsymbol{\beta}$!

Example — multiple regression (cont'd)

- Technically, the Uniform prior on the entire $(-\infty, \infty)$ scale is **improper**
  - This means that it does not integrate to 1
  - Consequently, it is not a "proper" probability distribution, which means we cannot give it a clear qualitative meaning
  - To encode poor information on $\boldsymbol{\beta}$, we can use a proper distribution centered on 0 and with large variance, e.g. $\beta_j \overset{iid}{\sim} \text{Normal}(0, 100000)$

Example — multiple regression (cont'd)

- Technically, the Uniform prior on the entire $(-\infty, \infty)$ scale is **improper**
  - This means that it does not integrate to 1
  - Consequently, it is not a "proper" probability distribution, which means we cannot give it a clear qualitative meaning
  - To encode poor information on $\boldsymbol{\beta}$, we can use a proper distribution centered on 0 and with large variance, e.g. $\beta_j \overset{iid}{\sim} \text{Normal}(0, 100000)$

- In addition, we are artificially discarding any prior information/knowledge we may have on the expected impact of $\boldsymbol{x}$ on y
  - The Uniform prior assumes that any possible value for β is equally likely — but we may have some ideas on the likely magnitude of the effect, or even its sign

Example — multiple regression (cont'd)

- Technically, the Uniform prior on the entire $(-\infty, \infty)$ scale is **improper**
  - This means that it does not integrate to 1
  - Consequently, it is not a "proper" probability distribution, which means we cannot give it a clear qualitative meaning
  - To encode poor information on $\boldsymbol{\beta}$, we can use a proper distribution centered on 0 and with large variance, e.g. $\beta_j \overset{iid}{\sim} \text{Normal}(0, 100000)$

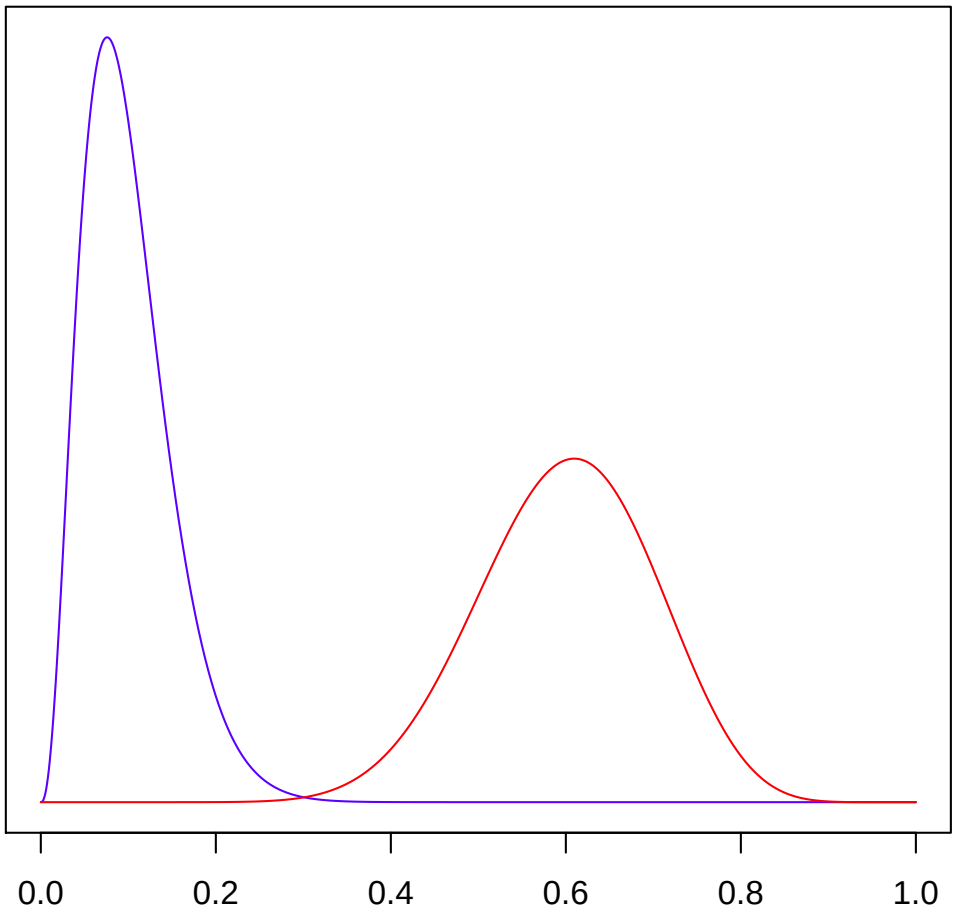
- In addition, we are artificially discarding any prior information/knowledge we may have on the expected impact of $\boldsymbol{x}$ on y
  - The Uniform prior assumes that any possible value for β is equally likely — but we may have some ideas on the likely magnitude of the effect, or even its sign

- Even if in the minimally informative case the numbers may be the same, the qualitative interpretation is quite different
  - Under the Bayesian approach, we are entitled to compute probabilistic assessments on both the posterior and the predictive distributions, e.g. $p(\boldsymbol{\beta} \mid \text{TD}) > 0$, or $p(y_{n+1} \mid \boldsymbol{x}_{n+1}, \text{TD}) > c$ for some specified threshold c

Impact of the prior distribution

Subject 1: $\text{mean}_1(\theta) = 0.05, \text{sd}_1(\theta) = 0.01 \Rightarrow p_1(\theta) \sim \text{Beta}(\alpha_1, \beta_1)$
$= \text{Beta}(3.5, 31.5)$

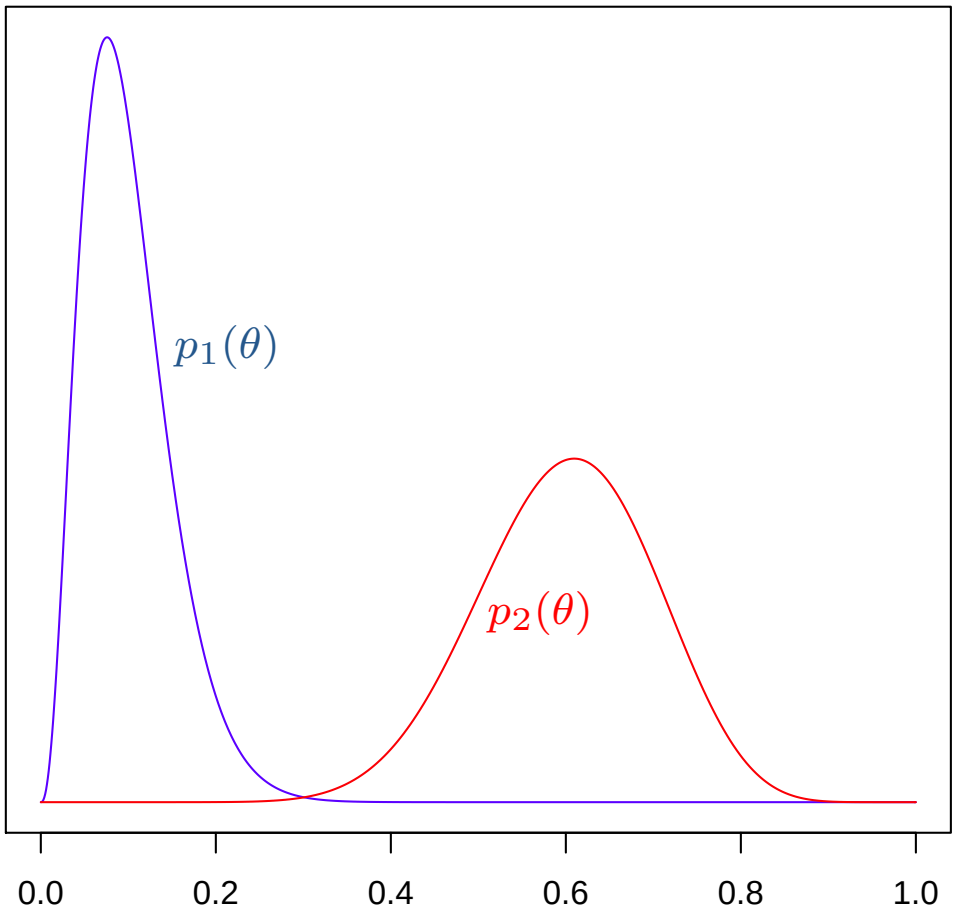
Subject 2: $\text{mean}_2(\theta) = 0.60, \text{sd}_2(\theta) = 0.10 \Rightarrow p_2(\theta) \sim \text{Beta}(\alpha_2, \beta_2)$
$= \text{Beta}(13.8, 9.2)$



Impact of the prior distribution

Subject 1: $\text{mean}_1(\theta) = 0.05, \text{sd}_1(\theta) = 0.01 \Rightarrow p_1(\theta) \sim \text{Beta}(\alpha_1, \beta_1)$
$= \text{Beta}(3.5, 31.5)$

Subject 2: $\text{mean}_2(\theta) = 0.60, \text{sd}_2(\theta) = 0.10 \Rightarrow p_2(\theta) \sim \text{Beta}(\alpha_2, \beta_2)$
$= \text{Beta}(13.8, 9.2)$



Impact of the prior distribution (2)

Suppose we observe $y = 20$ "successes" out of $n = 21$ trials

Subject 1:

$$p_1(\theta \mid y) \sim \text{Beta}(\alpha_1 + y, \beta_1 + n - y)$$
$$= \text{Beta}(23.5, 32.5)$$

Subject 2:

$$p_2(\theta \mid y) \sim \text{Beta}(\alpha_2 + y, \beta_2 + n - y)$$
$$= \text{Beta}(33.8, 10.2)$$

Impact of the prior distribution (2)

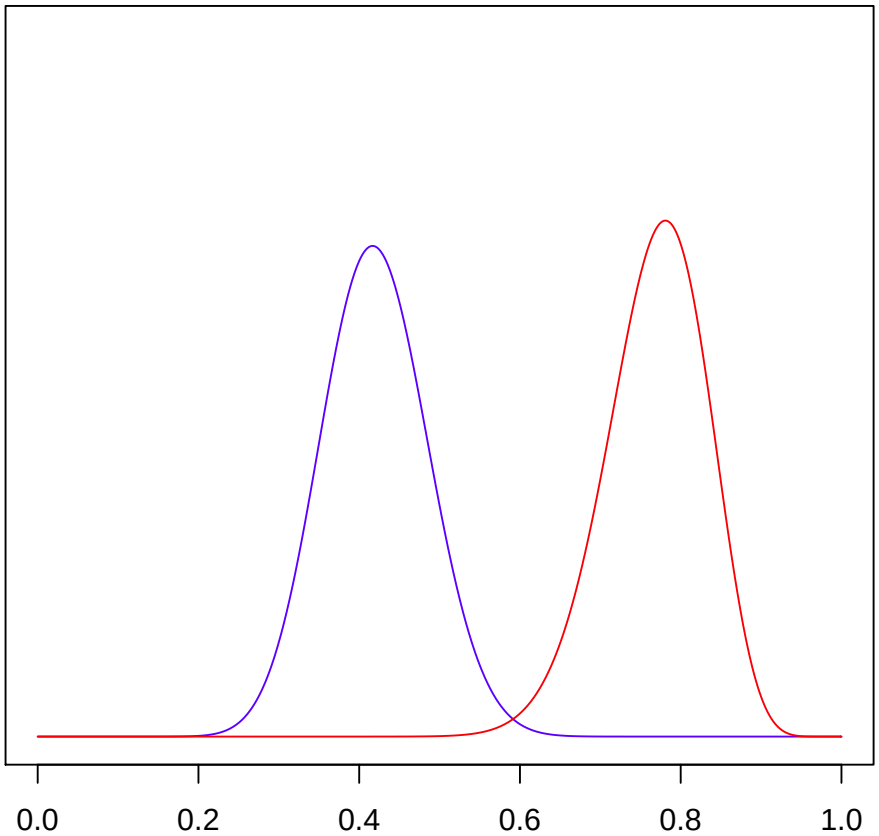
Suppose we observe $y = 20$ "successes" out of $n = 21$ trials

Subject 1:

$$p_1(\theta \mid y) \sim \text{Beta}(\alpha_1 + y, \beta_1 + n - y)$$
$$= \text{Beta}(23.5, 32.5)$$

Subject 2:

$$p_2(\theta \mid y) \sim \text{Beta}(\alpha_2 + y, \beta_2 + n - y)$$
$$= \text{Beta}(33.8, 10.2)$$



$$\text{mean}_1(\theta \mid y) = 0.4196$$
$$\text{sd}_1(\theta \mid y) = 0.0654$$
$$\text{mean}_2(\theta \mid y) = 0.7682$$
$$\text{sd}_2(\theta \mid y) = 0.0629$$

Impact of the prior distribution (3)

Now suppose we observe $y = 200$ "successes" out of $n = 201$ trials

Subject 1:

$$p_1(\theta \mid y) \sim \text{Beta}(\alpha_1 + y, \beta_1 + n - y)$$
$$= \text{Beta}(203.5, 32.5)$$

Subject 2:

$$p_2(\theta \mid y) \sim \text{Beta}(\alpha_2 + y, \beta_2 + n - y)$$
$$= \text{Beta}(213.8, 10.2)$$

Impact of the prior distribution (3)

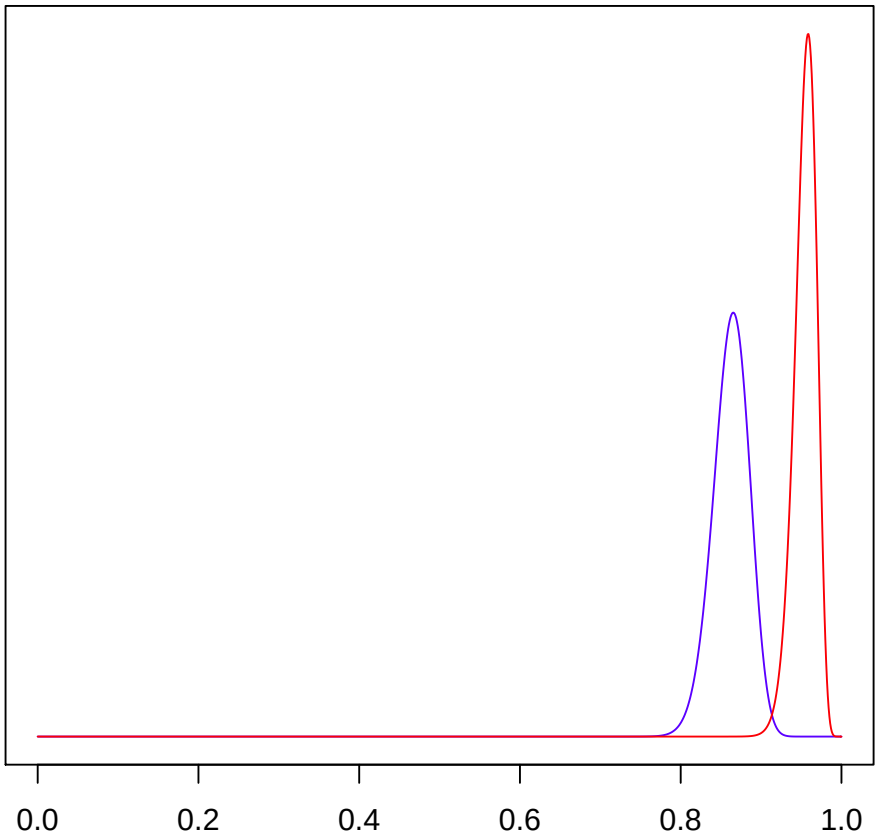
Now suppose we observe $y = 200$ "successes" out of $n = 201$ trials

Subject 1:

$$p_1(\theta \mid y) \sim \text{Beta}(\alpha_1 + y, \beta_1 + n - y)$$
$$= \text{Beta}(203.5, 32.5)$$

Subject 2:

$$p_2(\theta \mid y) \sim \text{Beta}(\alpha_2 + y, \beta_2 + n - y)$$
$$= \text{Beta}(213.8, 10.2)$$



$$\text{mean}_1(\theta \mid y) = 0.8623$$
$$\text{sd}_1(\theta \mid y) = 0.0224$$
$$\text{mean}_2(\theta \mid y) = 0.9545$$
$$\text{sd}_2(\theta \mid y) = 0.0139$$

The two prior opinions tend to converge to a common value "dominated" by the evidence

Conclusions

- Bayesian methods allow the formal and explicit incorporation of knowledge about the specific subject matter
- They are logically sound and directly address the relevant scientific questions of inference
- Particularly good to represent decision problems

Conclusions

- Bayesian methods allow the formal and explicit incorporation of knowledge about the specific subject matter
- They are logically sound and directly address the relevant scientific questions of inference
- Particularly good to represent decision problems

- Increasingly used in many applications
  - Great improvements from the computational point of view
  - Basically can model any problem, allowiung flexibility in the representation of the phenomenon under study
- Defining the prior distribution is indeed a complex matter — but it is doable!
  - Very close collaboration between statisticians and practitioners

Some **selected** references

Baio, G. (2012).
Bayesian Methods in Health Economics.
Chapman Hall CRC, Boca Raton, FL.

Bertsch McGrayne, S. (2011).
The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy.
Yale University Press, New Haven, CT.

Fienberg, S. (2006).
When Did Bayesian Inference Become Bayesian?
Bayesian Analysis 1, 1–40.

Gelman, A., J. Carlin, H. Stern, and D. Rubin (2004).
Bayesian Data Analysis - 2nd edition.
Chapman Hall, New York, NY.

Goodman, S. (1999).
Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy.
Annals of Internal Medicine 130, 995–1004.

Kruschke, J. (2011).
Doing Bayesian Data Analysis.
Academic Press, Burlington, MA.

Lindley, D. (2006).
Understanding Uncertainty.
John Wiley and Sons, New York, NY.

Thank you!