GÖTEBORGS UNIVERSITET

BAYESIAN STATISTICS WITH R

DAVID BOCK, BIOSTATISTICS, SCHOOL OF PUBLIC HEALTH, GÖTEBORG UNIVERSITY

Course aim

- The aim of the course is to give an introduction to Bayesian methods for statistical analysis with a focus on:
  - the understanding of basic concepts, methods and simpler problems
  - comparative discussion on basic bayesian and frequentist concepts
  - practical applications with the program R

Course content

- Introduction to Bayesian methods for statistical analysis focusing on:
  - Introduction to bayesian and frequentist concepts such as data and parameters, probability and likelihood as well as a prior and posterior distribution. *
  - analysis of various statistical questions using estimation, prediction, hypothesis testing and interval estimates
  - practical applications with the R program
- Less focus on mathematical concepts

* Bayesian statistics is easy to understand. Frequentist statistics is hard to understand.
Need to understand it (at least to some degree) to assess pros and cons of the two schools
Frequentist statistics still major scientific method

Course goals

- Knowledge and understanding
  - Explain key concepts in Bayesian statistics
  - Explain similarities and differences between basic bayesian and frequentist concepts
- Skills and Abilities
  - Suggest a Bayesian analysis method for a simpler statistical problem
  - Carry out a simpler Bayesian statistical analysis of data with the program R
- Evaluation ability and approach
  - Evaluate the quality of and interpret results of simple Bayesian analyzes

Course information

Course schedule spring semester 2021

Activity	Date	Time
Lecture 1	Mon 15/3	9-11
Lecture 2	Tues 16/3	9-11
Computer lab 1	Tues 16/3	14-16
Lecture 3	Fri 19/3	9-11
Lecture 4	Mon 22/3	9-11
Computer lab 2	Mon 22/3	14-16
Lecture 5	Wed 24/3	9-11
Computer lab 3	Wed 24/3	14-16

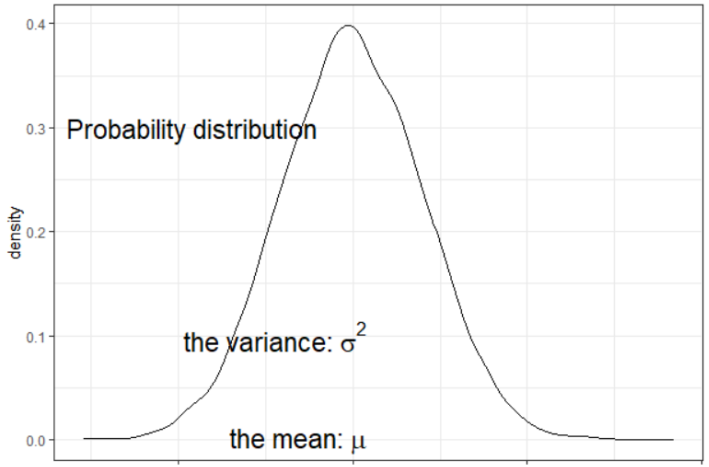
Scientific questions and uncertainty

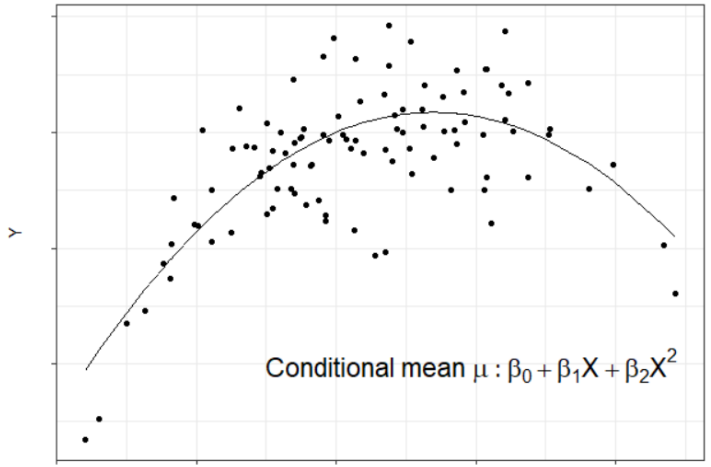
- Answer scientific questions or engage in decision-making where uncertainty is present
- Understanding uncertainty
  - Uncertainty due to unexplained random* behaviour
- Limited information available to support scientist/decision-maker
- Re-formulate objective as a task of gaining information on unknown quantities (”parameters”) that characterize a *probability/statistical model*

* synonym: ”Stochastic”

Probability model

- Mathematical representation of a random phenomenon.
- Components of a probability model
  - Characterize the randomness by probability distribution
  - The distributional parameters
  - The conditional dependence between parameters and other information

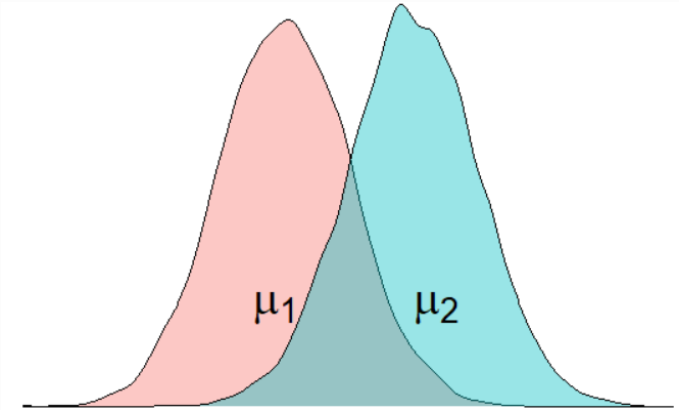




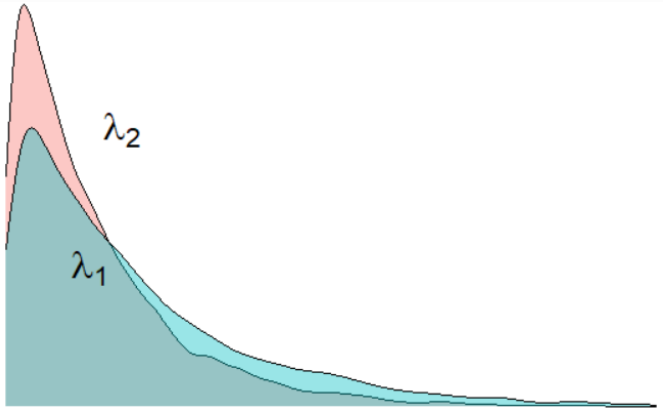
Parameters and probability distribution

- Probability distribution
  - Characterize by parameters
  - Parameter: θ (or μ, λ, β, etc)

Normal distribution



Exponential distribution



Probability and statistical models

- Probability model
  - All components are known
    - Calculate probabilities of observing specific outcomes
- Statistical model
  - Probability model suggested for explaining observed data
  - Components of the model are unknown
    - Use data to construct a probability model
    - Use model to answer questions about data

Inference about statistical models

- Re-formulate scientific objective as a task of gaining information on unknown quantities that characterize a *statistical model*
- Aim of statistical inference is to use data (realizations from the probability distribution) to infer about the parameters
  - Give numerical estimate of the the parameter (”a good guess”)
  - Quantify the uncertainty of the estimate
  - Statements of parameters relevant for scientific hypotheses (hypothesis testing)
- Y = outcome of interest
  - y = observations of Y (data)
- Statistical model
  - Y has a probability distribution with unknown parameter θ

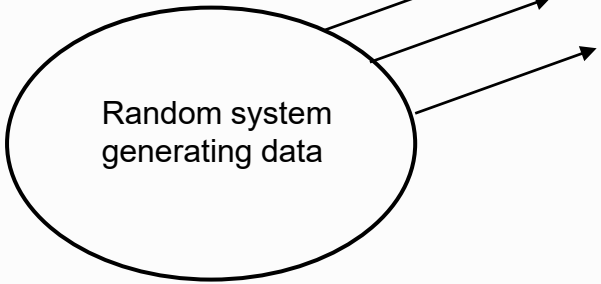
Inference about statistical models

- Aim is to estimate θ
  - Use data y to caclulate an *estimator* of θ, $\hat{\theta}$ (”point estimate”)
- What is the **uncertainty** by which we estimate θ ?
- Hypothesis testing regarding θ
  - H0: θ = 0
  - H1: θ ≠ 0
- Flip a coin, sometimes head, sometimes tails
  - Characterized as random behaviour
  - θ = probability of head (1- θ = probability of tails)

Probability

- Uncertainty due to unexplained random behaviour
- Probability is a characterization of uncertainty
- ”Frequentist”:
  - Random behaviour: physical phenomena
    - Characterize behaviour in hypothetical repetitions under the same conditions
    - Random behaviour in the long run
    - The relative ***frequency*** of different outcomes
- ”Bayesian”:
  - Probability reflects degree of belief
    - Quantify subjective belief as a probability distribution (”*prior distribution*”)
    - Your personal thoughts about uncertainty are fully valid

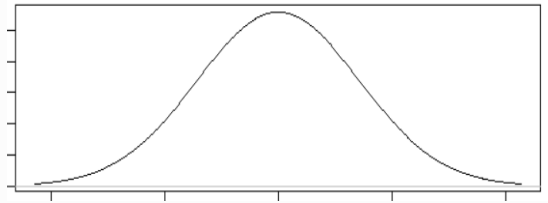
Frequentist probability distribution



Start system. Observe what comes out (y)
Start system. Observe what comes out
Start system. Observe what comes out

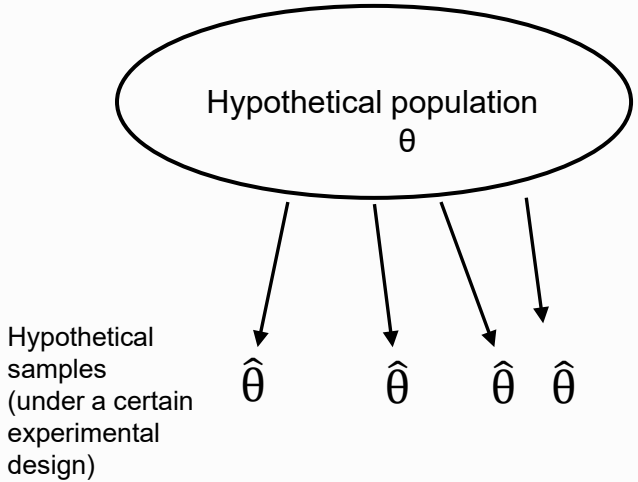
<do this inifinitely many times>

Note the value of Y each time
The number of times you observe a **certain value y**
(divided by by total number of looks = INFINITELY MANY)
= **Probability of that value for y**



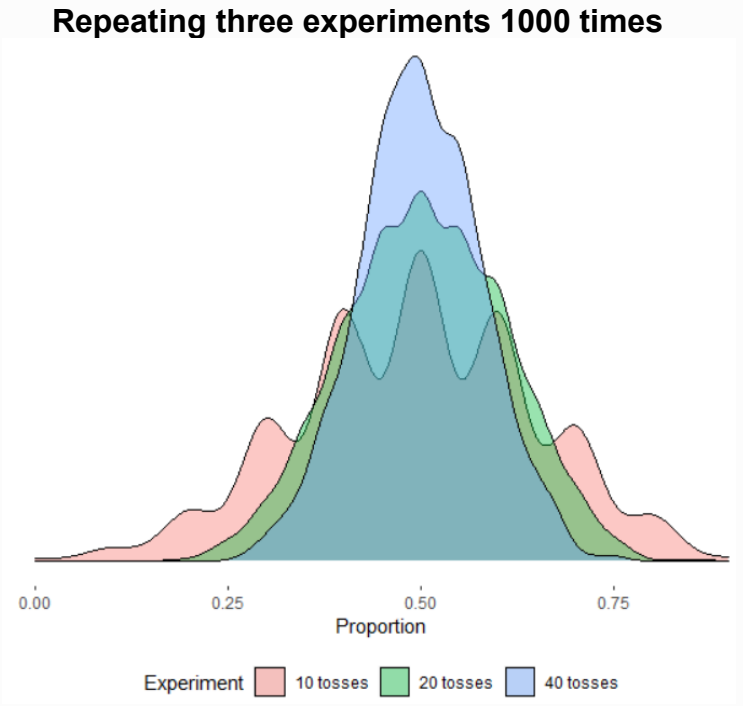
Frequentist Sampling distribution

- Relative proportion of heads is an estimate of θ
- Uncertainty of θ characterized by hypothetical repetitions under the same conditions (same experiment)
  - ”sampling distribution”
- Data determines our estimate of $\hat{\theta} = \hat{\theta}(y)$
- Sampling distribution: The distribution of estimates $\hat{\theta}$



Frequentist Sampling distribution

- Sampling distribution assess the plausibility of different outcomes (such as an estimate $\hat{\theta}$) when repeating the same experiment under identical conditions, e.g. under H0: $\theta = 0$
- Samping distribution is the only link between $\hat{\theta}$ and θ
- Sampling distribution given the same experiment
- Sampling distribution is used for constructing confidence intervals and calculating p-values



Example

- Toss a coin 10 times. Calculate an estimate of θ

$\hat{\theta} = \frac{\sum y_i}{10} = 0.6$

95% Confidence interval:

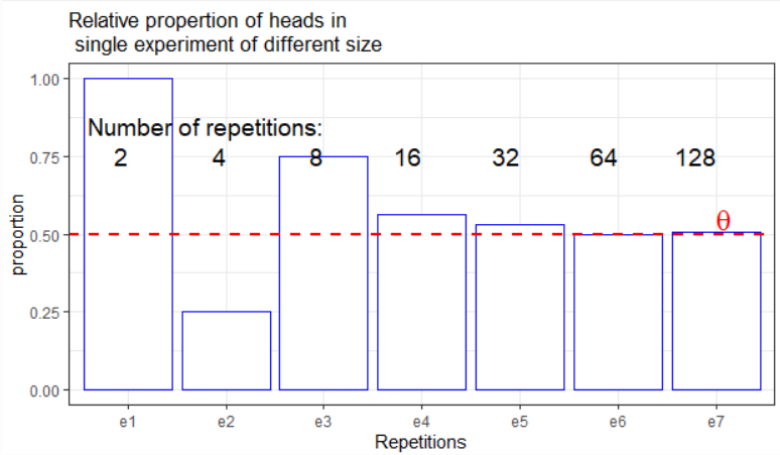
$\hat{\theta} \pm 1.96 * SE(\hat{\theta}) = (0.30; 0.85)$

H0: θ=0.5

H1: θ≠0.5

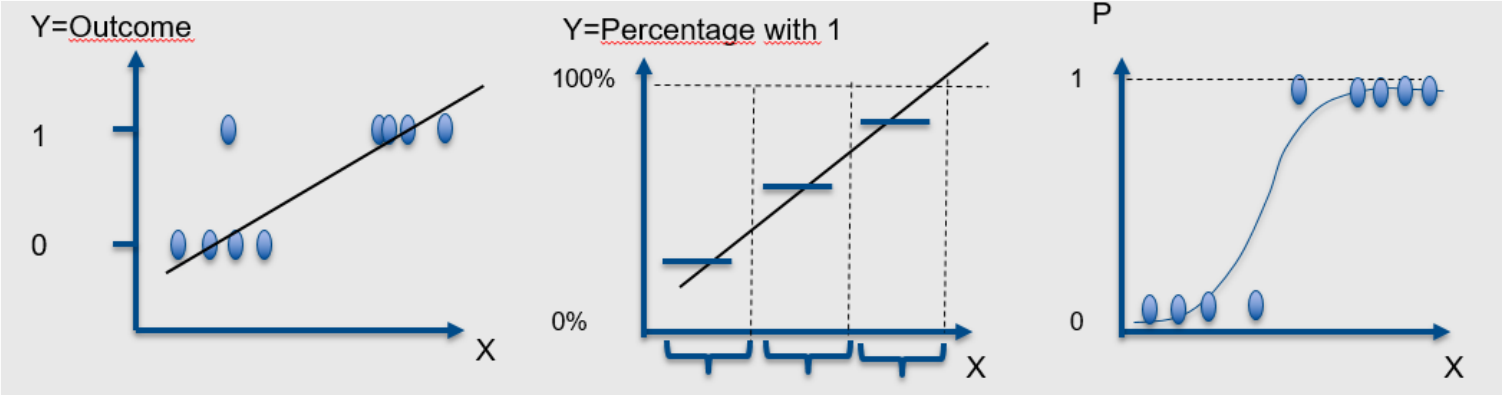
P-value = 0.53

Uncertainty about θ becomes smaller when sample size increase.



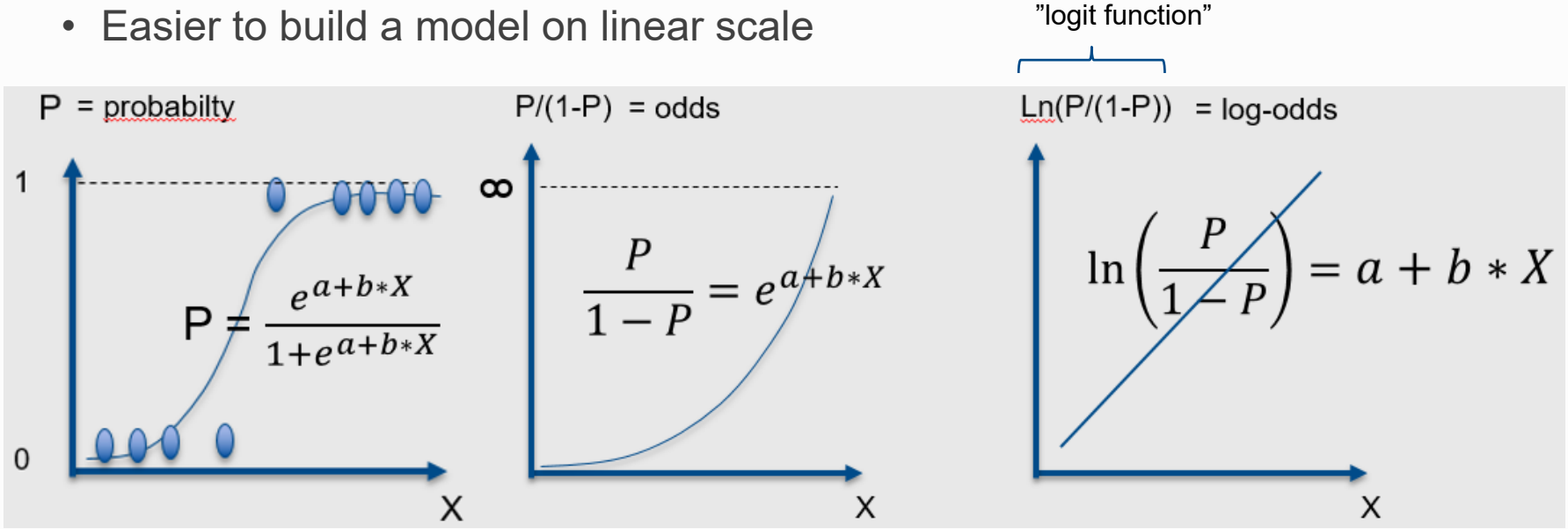
Logistic regression

- Y is binary outcome (can take two values)
  - Bernoulli distribution with probability parameter p
    - (or *Number of outcomes of certain among Total number: Binomial distribution*)
  - How to describe relationship between Y and indendent variable X?



Logistic regression

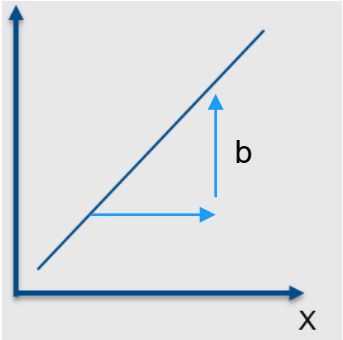
- Model the probability p as a function of X
- Logistic function ensures p is within 0 and 1
- Easier to build a model on linear scale



Logistic regression

Logit function: $\ln\left(\frac{P}{1-P}\right) = a + b * X$

$$\frac{P}{1-P} = e^{a+b*X}$$



- If X increased by 1 unit,
  - log-odds absolute increase by *b* units
  - Odds relative increase by $e^{b} - 1$
- Odds ratio = e^{b}

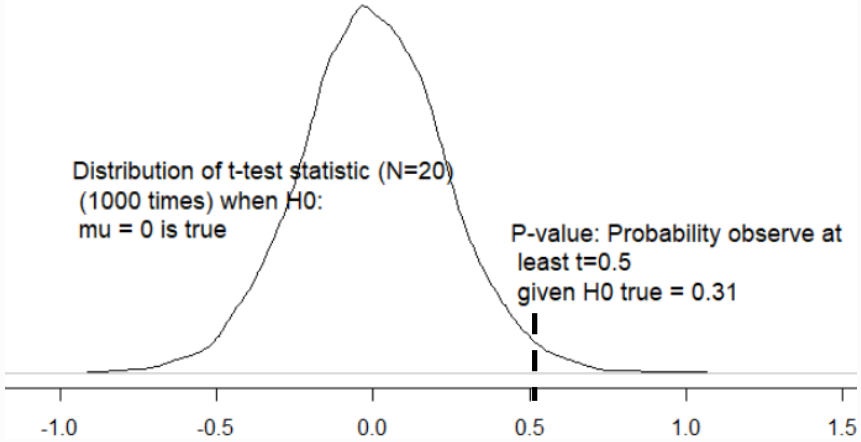
P-value

- The probability under a specified statistical model that a statistical summary of the data would be equal to or more extreme than its observed value*
- Probability of obtaining results *at least as extreme as* those observed given that the null hypothesis is true
  - Based on hypothetical outcomes more extreme than the ones observed
- Gives probability of data given null hypothesis *P*(*data*|*hypothesis*) rather than probability of alternative hypothesis given data *P*(*hypothesis*|*data*)

* Wasserstein & Lazar, 2016, *American Statistician*)

Sampling distribution and p-values

- Sampling distribution is the only link between $\hat{\theta}$ and θ
- Instead of $\hat{\theta}$ look at t(y) = t-test statistic $\hat{\theta}/SE(\hat{\theta})$
- Hypothesis testing regarding θ
  - H0: θ = 0
  - H1: θ ≠ 0
- P-value:
  - Depends on the sampling distribution
    - Depends on the experiment
- The only way to test H0 is through the p-value



Statistical inference and nature of probability

- Aim is to estimate θ ($\hat{\theta}$), quantify uncertainty of θ and make statements on θ relevant for your scientific objectives by means of new information
- Frequentist:
  - **Can´t** express beliefs about θ as probability distribution
  - θ assumed unknown and fixed
    - Aim is to use information to estimate a single value $\hat{\theta}$ (”point estimate”)
  - Bayesian:
  - **Can** express beliefs about θ as *prior probability distribution*
    - Aim is to characterize the distribution in light on new information (”*posterior distribution*”)

Frequentist and bayesian statistical models. Example

- Frequentist

  $Y \sim N(\mu , \sigma)$

  can be expressed as:

  $Y = \mu + \varepsilon, \varepsilon \sim N(0, \sigma)$

- Bayesian

  $Y \sim N(\mu , \sigma)$

  $\mu \sim$ <*choice of prior*>

  $\sigma \sim$ <*choice of prior*>

Read as:
Y has a Normal distribution with unknown parameters mu and sigma

Read as:
Y has a Normal distribution with <choice of prior> for mu and <choice of prior> sigma

Summary

- Scientific research as matter of gaining information on unknown quantities (”parameters”) that characterize a *probability/statistical model*
  - Probability model describes random phenomena
  - Statistical model: Model suggested for explaining observed data
- Frequentist sampling distribution characterizes probability as behavior under hypothetical repetitions
- Bayesian view on probability: Reflect degree of belief
- Logistic regression: Model for describing random behavior for binary outcomes

Components we are interested in

- θ = Quantity of interest
  - Parameter of a probability distribution
  - Future outcome of observation (prediction)
- y = data
- P(θ) = prior probability distribution

- f(y|θ) = ”Likelihood”
- P(θ|y) = ”Posterior distribution”

Re-expression of our scientific question

New information

or no prior if probability does not reflect degree of belief (frequentist)

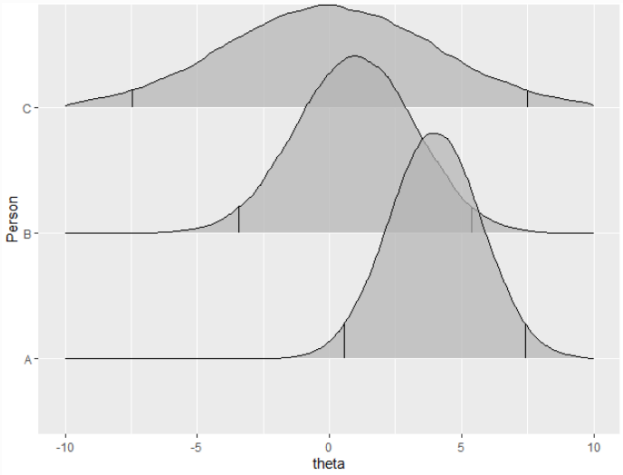
Our probability model (connection between X and θ)

New belief of θ in light of data

Prior probability distribution

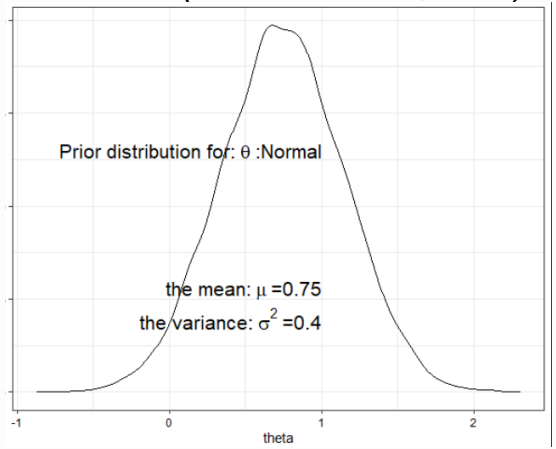
- Reflect degree of prior belief
  - I chose to characterize my belief about the treatment efficacy (Drug vs Placebo) θ as:

Person A, B, C´s belief



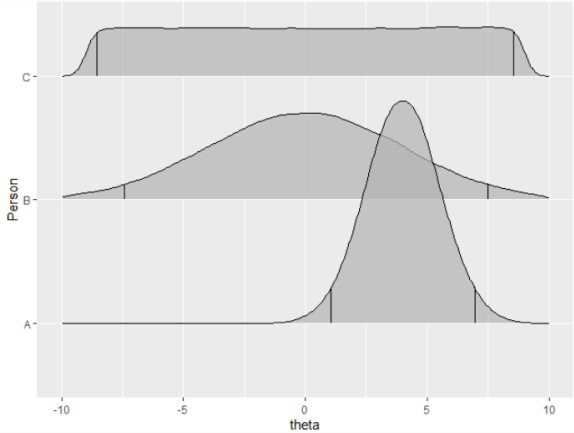
Reflect past information

”*Based on a meta-analysis the Average effect is belief to be around 0.75 (95% CI: -0.10; 1.90)*”



Prior probability distribution

- Very vaguely reflect past information (weakly informative)
- Reflect ”abscence” of prior knowledge (noninformative prior)



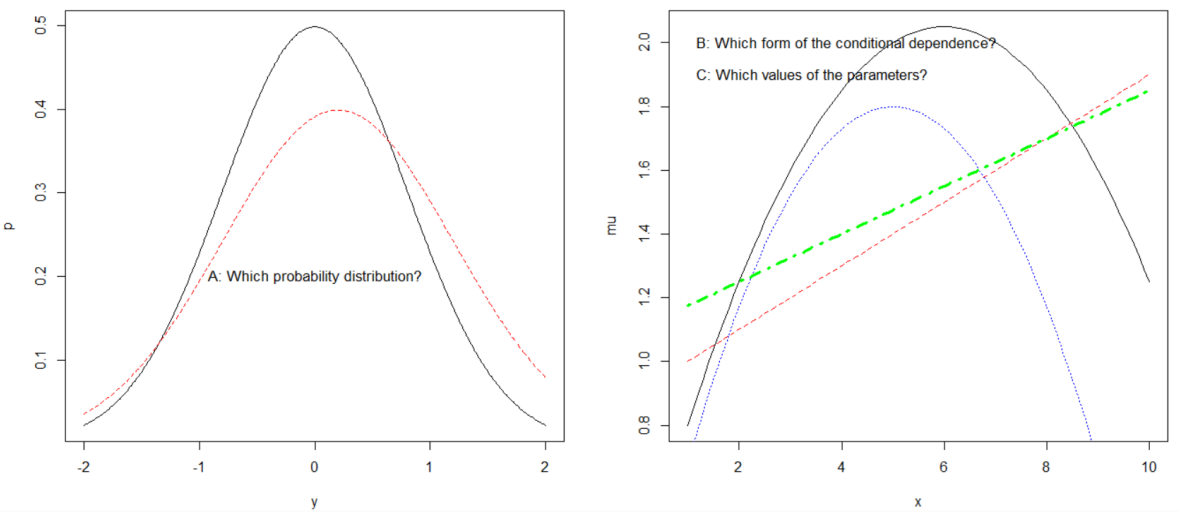
- Chose prior to make computations and interpretations easy (”conjugate”)
- Choise prior to fulfill different goals with analysis (e.g. ”regularization”)

Likelihood

- Links probability model to data
- Given a probability model, to what extent does it fit to data?
- It tells us the probability of observing our data given the value(s) of some parameter(s).
- f(y|model) = Mathematical function of Y and the model
  - Concordance between data and the model
  - Simple case f(y|θ) = single parameter model
- Each probability model yields a unique form of the likelihood function
  - Matematical function reflecting
    - The models functional form
    - The probability distribution

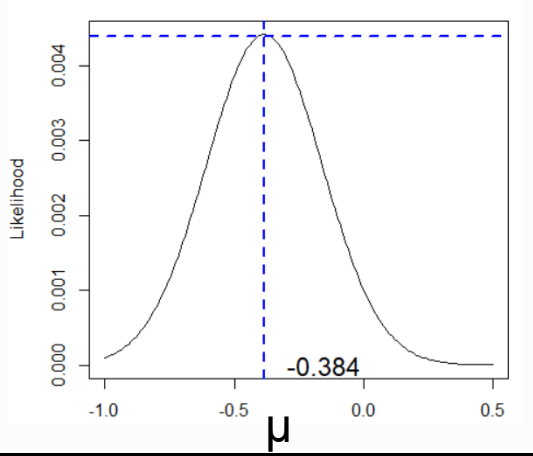
Likelihood

- What is the evidence in data (y) for the model?
  - probability distribution
  - functional form
  - parameter values



Likelihood

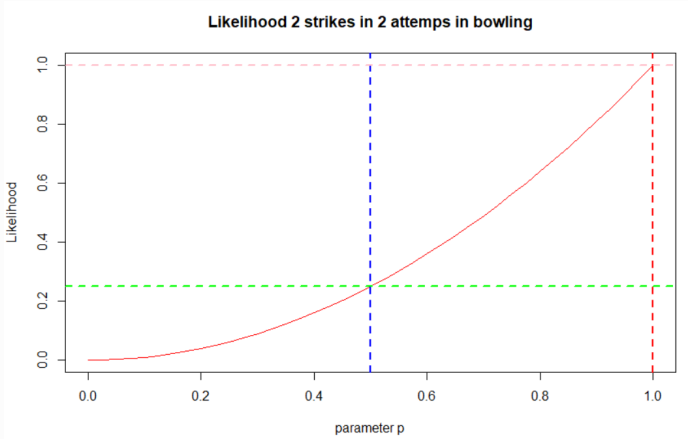
- Model: Y~ iid N(μ, σ), iid = ”indepdent and identically distributed”
- Data: $y_1, y_2, \ldots, y_n$,
- Likelihood: $f(y|\mu,\sigma) = (2\pi\sigma^2)^{-n/2}\exp\left(\frac{-\sum_{i=1}^{n}(y_i-\mu)^2}{2\sigma^2}\right)$, θ=(μ, σ)
- Given observations, likelihood, and value of θ, it becomes a single value (often expressed as log-likelihood)
- Evidence in favor of certain θ
  - The larger the likelihood (given data) the more plausible is θ



Likelihood

- Example
  - Number of strikes among attempts when playing bowling
  - Simple single parameter model (f(y|θ)
    - Model: Binomial distribution with parameter p

$$f(y|\theta) = \binom{n}{r} \theta^{y} * (1-\theta)^{N-y}$$



Likelihood for p=0.5 given 2 strikes in 2 attemps = 0.25

Likelihood for p=1.0 given 2 strikes in 2 attemps = 1.0

Likelihood

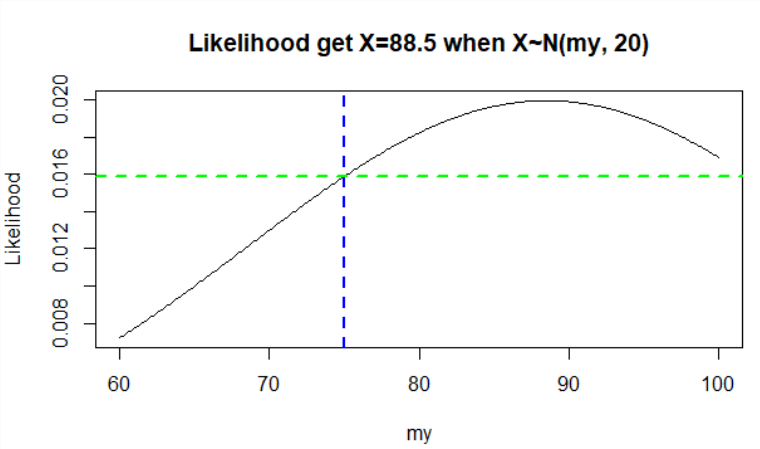
Example:

Body weight = 88.5 is observed.

Model: Body weight ~N(μ, 20)

Likelihood of μ=75 given x= 88.5 is 0.016

$$f(y|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(y-\mu)^2}{2\sigma}\right)$$



Likelihood vs sampling distribution

- Sampling distribution: θ constant, y is varying
  - $Y \sim p(Y|\theta)$

- Likelihood: y constant, θ is varying
  - $\theta \sim p(\theta|y)$

Bayes formula

$$P(\theta|y) = \frac{f(y|\theta) * P(\theta)}{f(y)}$$

BAYES FORMULA

- Where $f(y) = \int f(y|\theta)P(\theta)d\theta$ is used for normalizing P(θ|y) to be within (0, 1)
  - *f(y)* is a marginal likelihood
- Since f(y) just normalizing constant, we can say

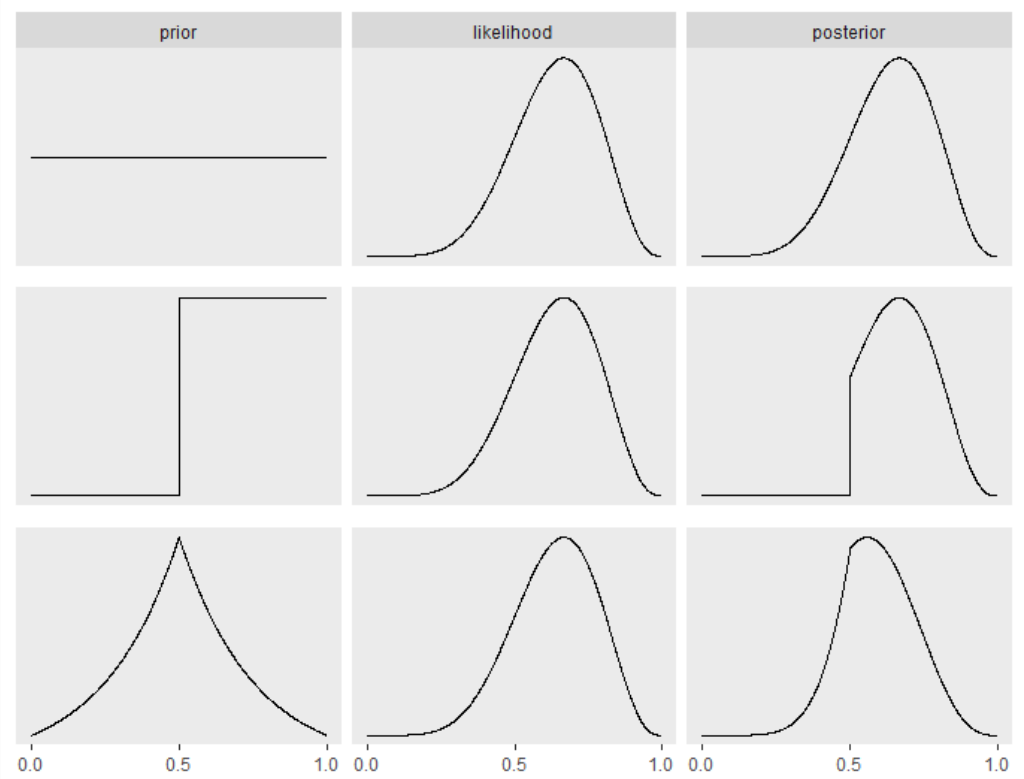
$$P(\theta|y) \propto f(y|\theta) * P(\theta)$$

∝: “proportional to”

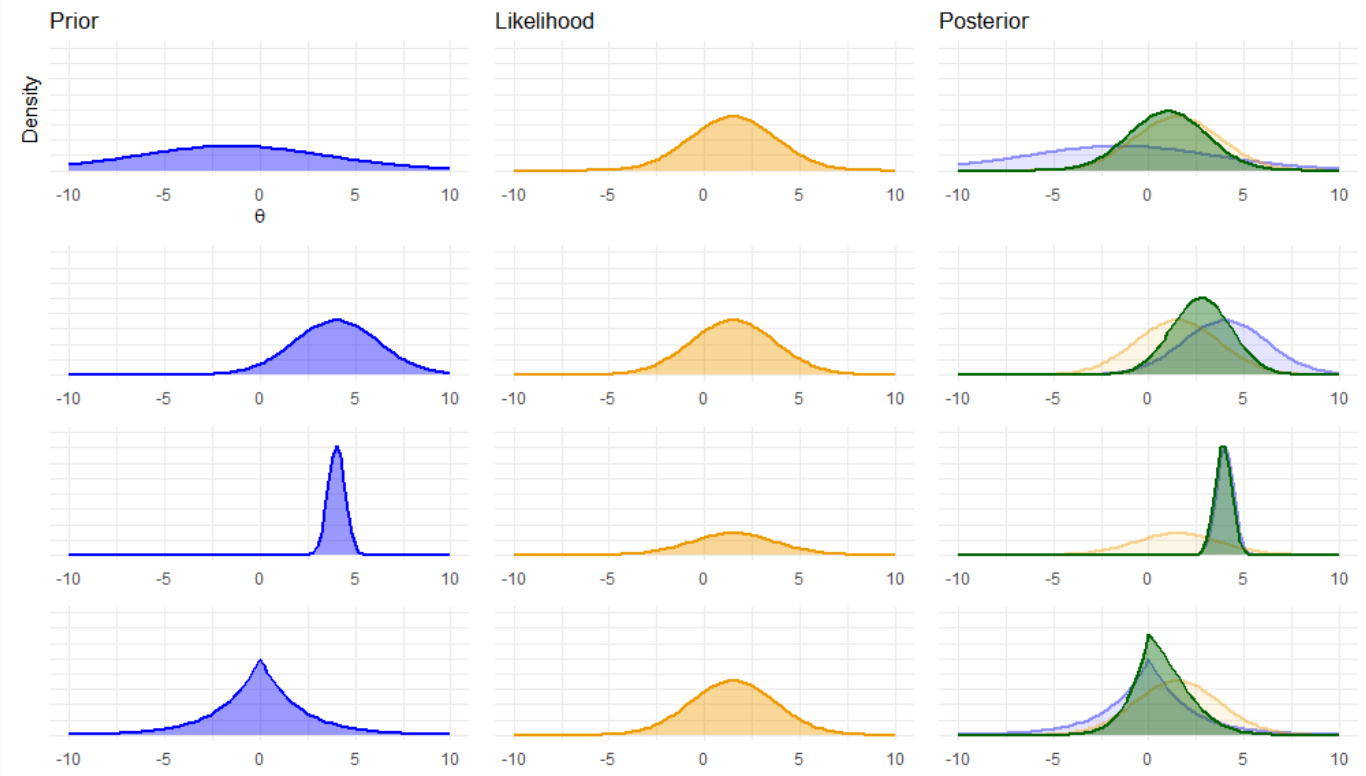
- Posterior ∝ Prior * likelihood

Posterior ∝ Prior * likelihood

$$P(\theta|y) \propto f(y|\theta) * P(\theta)$$

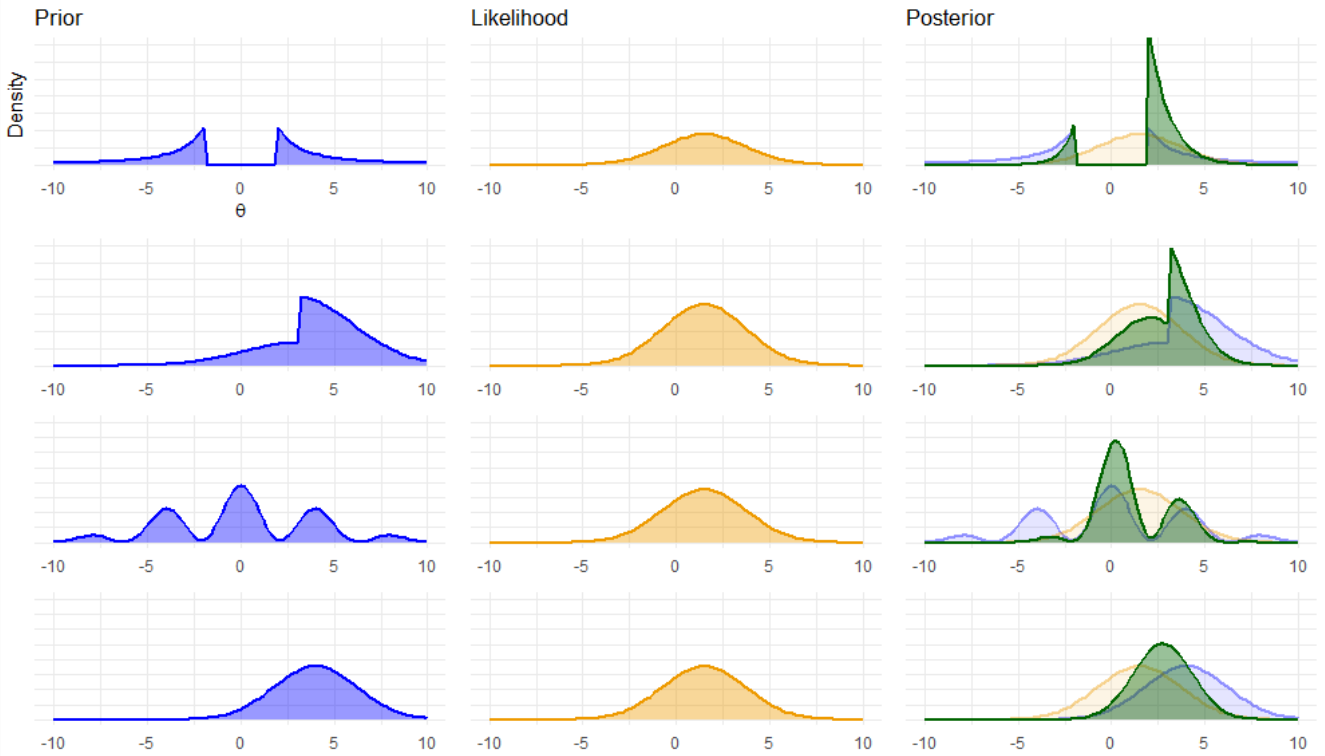


Posterior ∝ Prior * likelihood



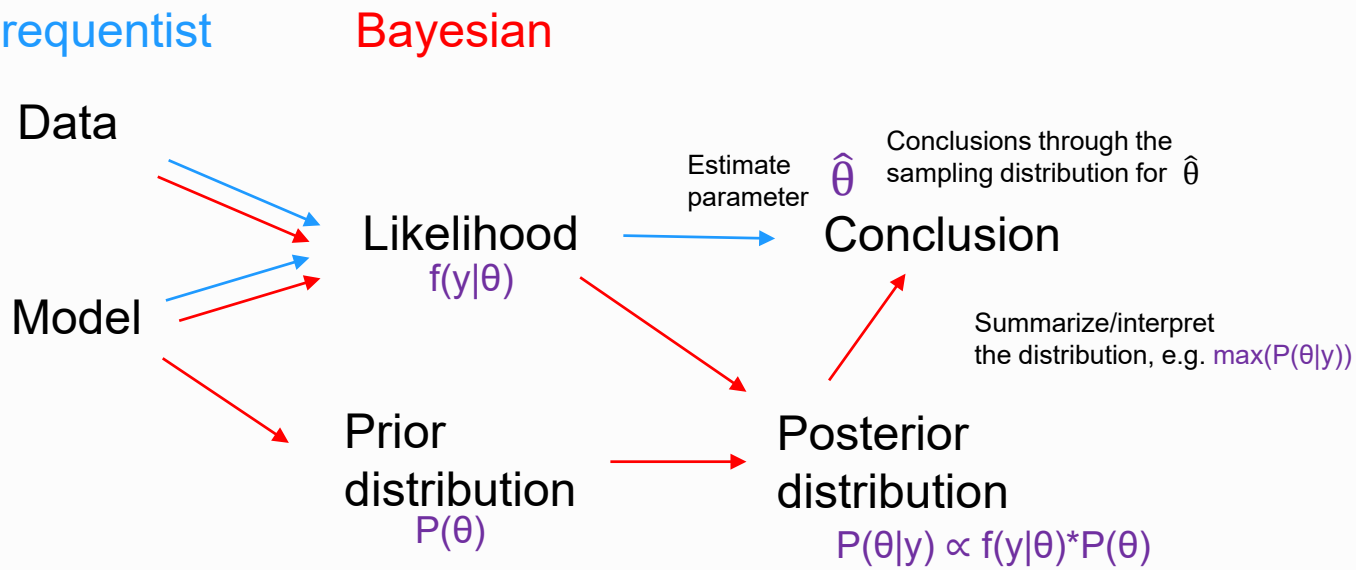
https://github.com/mattansb/bayesian-evidence-iscop-2021/blob/main/bayesian-evidence-iscop-2021.Rmd

Posterior ∝ Prior * likelihood



https://github.com/mattansb/bayesian-evidence-iscop-2021/blob/main/bayesian-evidence-iscop-2021.Rmd

Approaches to inference



Likelihood and sampling distribution

- Likelihood: how plausible model is given data
- Sampling distribution: how plausible is data given model and experimental design

Frequentist

Likelihood: data → model (to make statements about θ)

Sampling distribution: model → data (to adress uncertainty of your statement)

Bayesian

Likelihood: data → model (to make statements about θ)

Sampling distribution: model → data (to adress uncertainty of your statement) (not needed)

- Frequentists must go through the sampling distribution to make statements on θ (because you need the sampling properties of $\hat{\theta}$)
- Bayesians must use a prior distribution to make statements on θ

Summary

- Prior probability distribution
  - Reflect degree of belief about parameters of a statistical model
- Likelihood
  - Concordance between observed data and different values of parameters or statistical model
- Posterior probability distribution
  - Prior + likelihood
- Frequentists must go through the sampling distribution to make statements on θ (because you need the sampling properties of $\hat{\theta}$)
- Bayesians must use a prior distribution to make statements on θ

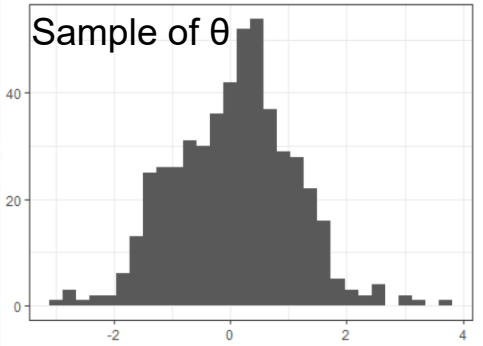
Thomas Bayes

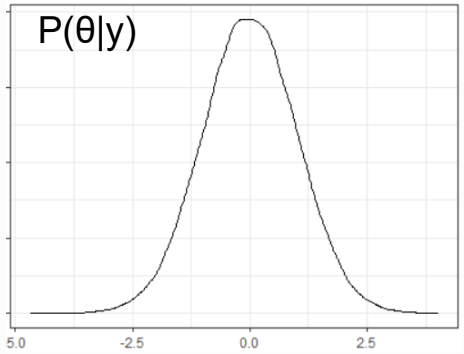
- Bayesian statistics named after Thomas Bayes (1702-1761)
- English statistician, philosopher and Presbyterian minister.

$$P(\theta|y) \propto f(y|\theta) * P(\theta)$$

How to calculate the posterior distribution?

- Posterior distribution $P(\theta|y) = \frac{f(y|\theta) * P(\theta)}{f(y)}$
  - Realistic/complex problems impossible compute analytically (closed form expression)
    - $P(\theta|y)$ unkown
  - $f(y) = \int f(y|\theta)P(\theta)d\theta$ can´t be computed
  - A formula: θ →f(θ) =$P(\theta|y)$ not available
- Create a sample of values of θ and try to approximate $P(\theta|y)$ numerically





How to calculate the posterior distribution?

- How can we sample θ if $P(\theta|y)$ is unknown?
  - What do we know is the unnormalized $P(\theta|y)$, $f(y|\theta) * P(\theta)$
- Take a ***sample*** of values of $\theta = \theta_i$ i, = 1, 2…,
- We want the relative number of times we pick certain values of θ to have a histogram that approximates $P(\theta|y)$
  - Let sampling depend on: $f(y|\theta) * P(\theta)$
  - We want to chose values of θ around the peak(s) of the unnormalized $P(\theta_i|y)$ more often than in the tails

$f(y|\theta) * P(\theta) \rightarrow$ sample values θ to achieve a histogram of the same shape $\rightarrow$ approximate $P(\theta|y)$

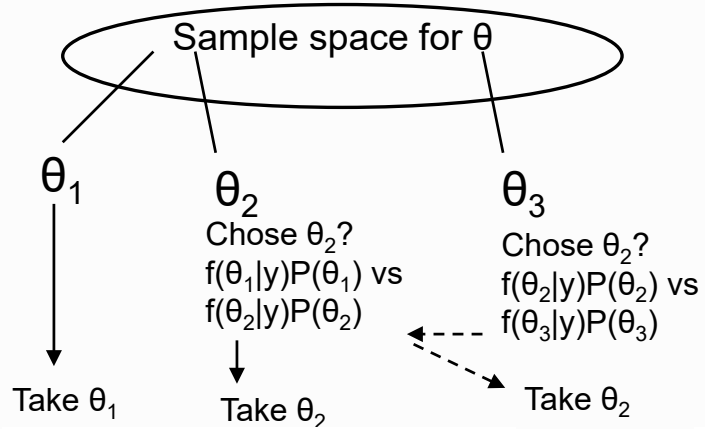
How to calculate the posterior distribution?

- What value of θ is chosen for θ_{i+1} is made dependent on the previous θ_i
  - $f(y|\theta_{i+1}) * P(\theta_{i+1})$ vs $f(y|\theta_i) * P(\theta_i)$
- How often certain values are “chosen” according to a *Markov Chain*
  - Markov Chain: Random process. Future state depends on current state
  - Stationary distribution: The probability of θ_{i+1} = θ regardless of θ_i
- The marginal distribution of the values of θ chosen approximate $P(\theta|y)$

$f(y|\theta) * P(\theta) \rightarrow$ sample values θ to achieve a histogram of the same shape $\rightarrow$ approximate $P(\theta|y)$

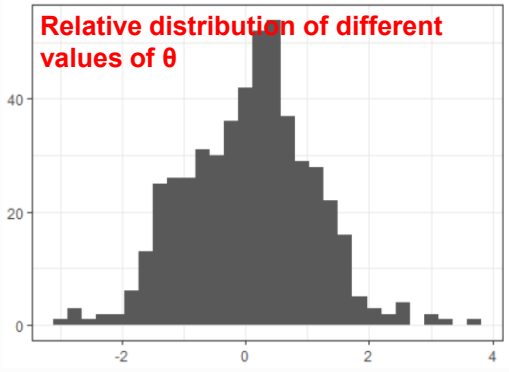
The ”choice” of which values to sample and how often is based on ”Markov Chain Monte Carlo” (MCMC)

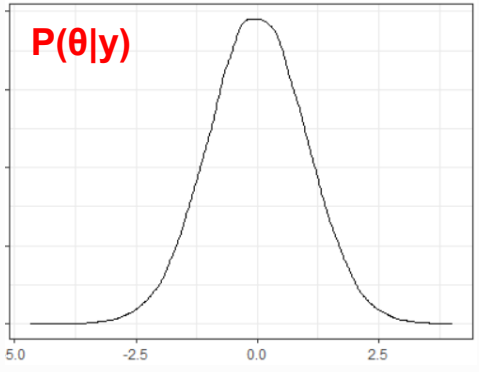
How to calculate the posterior distribution?



Do this several selection of θ´s thousand times

Some times the same θ is chosen, sometimes not





Example: Predict rain using bayesian inference

- I believe the probability it will rain today is 0.3 (that is will not rain=0.7)
  - $P(R^+)=0.3$, $P(R^-)=0.7$

PRIOR

- You observe dark clouds (C^+)
  - Likelihood observe dark clouds prior rain? $P(C^+|R^+) = 0.9$.
  - Likelihood observe dark clouds prior to no rain? $P(C^+|R^-) = 0.4$.

DATA

LIKELIHOOD

- Probability of R^+ given observed C^+ :

POSTERIOR PROBABILITY

$$P(R^+|C^+) = \frac{P(C^+|R^+) * P(R^+)}{P(C^+)} = \frac{P(C^+|R^+) * P(R^+)}{P(C^+|R^+) * P(R^+) + P(C^+|R^-) * P(R^-)} =$$

$$\frac{0.9 * 0.3}{0.9 * 0.3 + 0.4 * 0.7} = 0.49$$

Example: Predict rain using bayesian inference

- First $P(R^+) = 0.3$
- Observing dark clouds strengthened belief of rain to 0.49
- What if observe no clouds (C^-) instead.
  - Likelihood observe dark clouds prior rain? $P(C^- | R^+) = 0.1$.
  - Likelihood observe dark clouds prior to no rain? $P(C^- | R^-) = 0.8$.
- Probability of R^+ given observed C^-

PRIOR

DATA

LIKELIHOOD

POSTERIOR PROBABILITY

$$P(R^+|C^-) = \frac{P(C^-|R^+) * P(R^+)}{P(C^-)} = \frac{P(C^-|R^+) * P(R^+)}{P(C^-|R^+) * P(R^+) + P(C^-|R^-) * P(R^-)} =$$

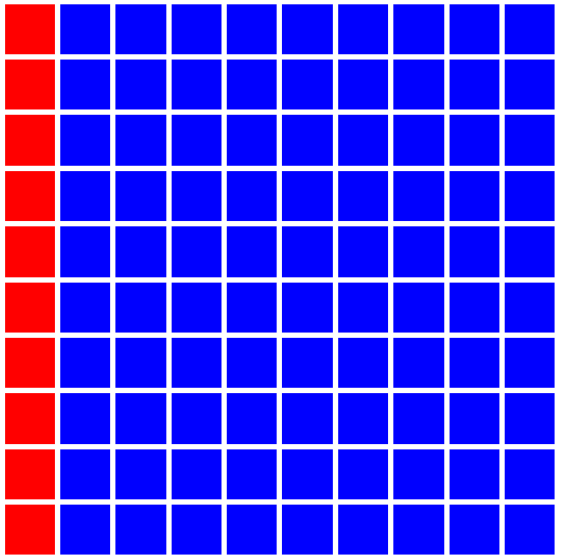
$$= \frac{0.1*0.3}{0.1*0.3+0.8*0.7} = 0.051$$

- Abscence of clouds weakened your belief

Medical test paradox example

- Say you want to test of you have a disease by diagnostic test
  - Prevalence of disease is 10% in the population (100)
  - TP = 9, FP = 10
  - Sensitivitet ($P(+|D^{+})$) = 9 out of 10 (90%)
  - Specificity ($P(-|D^{-})$) = 80 out of 90 (89%)

  How probable is it that you actually have the disease given you have tested + ?



See https://www.youtube.com/watch?v=lG4VkPoG3ko&feature=youtu.be&ab_channel=3Blue1Brown

Medical test paradox example

- Probability you have the disease given tested positive
  - Positive predictive value (PPV) = $P(D^+|+)$ = Posterior probability

Sensitivity

Prevalence of disease

$$P(D^+|+) = \frac{P(+|D^+) * P(D^+)}{P(+)}$$

Probability testing positive

- $P(+|D^+)=0.9$
- $P(D^+)=0.1$
- $P(+)=P(+|D^+)*P(D^+) + P(+|D^-)*P(D^-)$
- $P(+)=(TP+FP)/100=(9+10)/100=0.19$

Medical test paradox example

$$P(D^+|+) = \frac{P(+|D^+) * P(D^+)}{P(+)} = \frac{0.9 * 0.1}{0.19} = 0.47$$

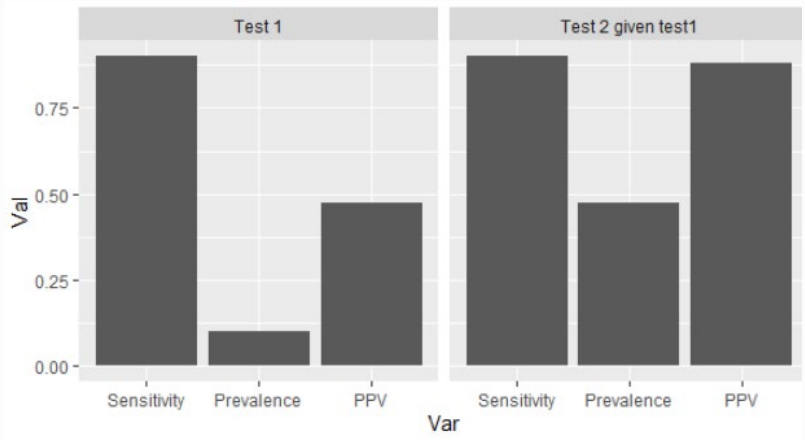
- Probability having disease given + = 0.47
- Sensitivity was 0.90
- Prior probability of disease $P(D^+)$ low (0.10)
  - Despite positive test it is not likely you have the disease

Medical test paradox example

- Lets say you do a second test
  - New prior: Current PPV = 0.47
  - P(+) = P(+|D^+)*PPV + P(+|D^-)*(1-0.47) = 0.485

$$P(D^+|++) = \frac{P(+|D^+)*PPV}{P(+)} = \frac{0.9*0.47}{0.485} = 0.88$$

  - Overall prevalence of 0.10 replaced with update P(D^+)|+) = 0.47



Estimation

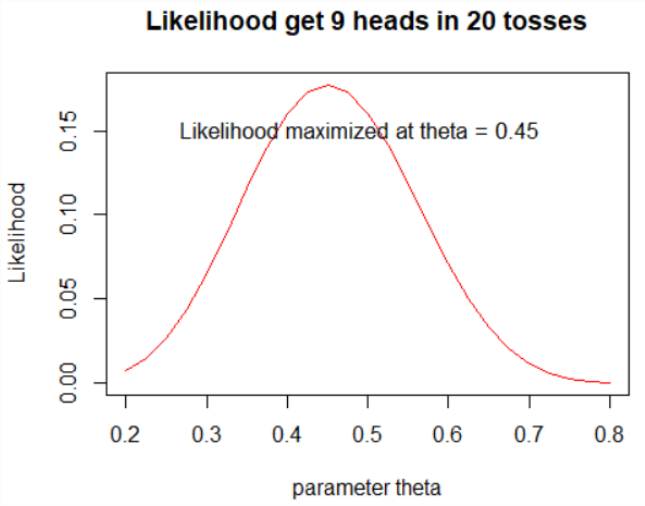
- Aim is to estimate θ
- Frequentist: $\hat{\theta} = \hat{\theta}(y)$
  - Chose a single value $\hat{\theta}$ based on y ($\hat{\theta}(y)$) where likelihood is maximized ("ML estimator")
  - "ML estimator" has good statistical properties, easy to work with, etc
- Bayesian: P(θ|y) *
  - Summarize the posterior distribution by for example
    - Median
    - Average
    - Peak ("Maximum A Posteriori" (MAP))

Estimation

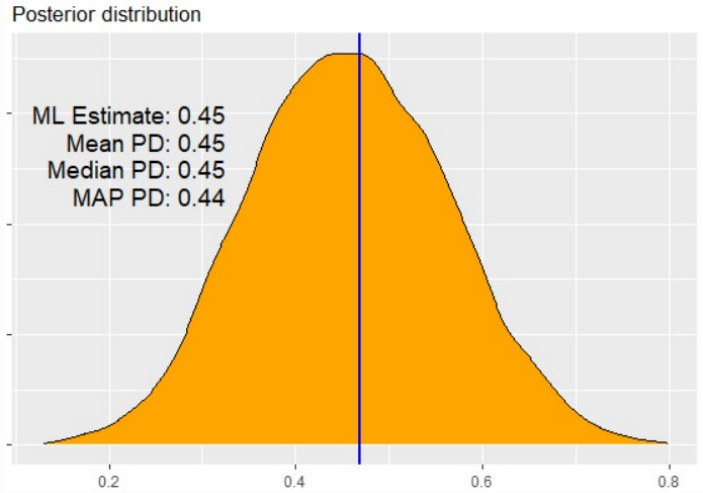
- Example: 20 coin tosses. Aim is to estimate probability of heads

Statistical model

Frequentist:
Y~Binomial(N, θ)



Bayesian:
Y~Binomial(N, θ)
Logit(θ) ~ N(μ=0, σ=2.5)



Likelihood
Prior

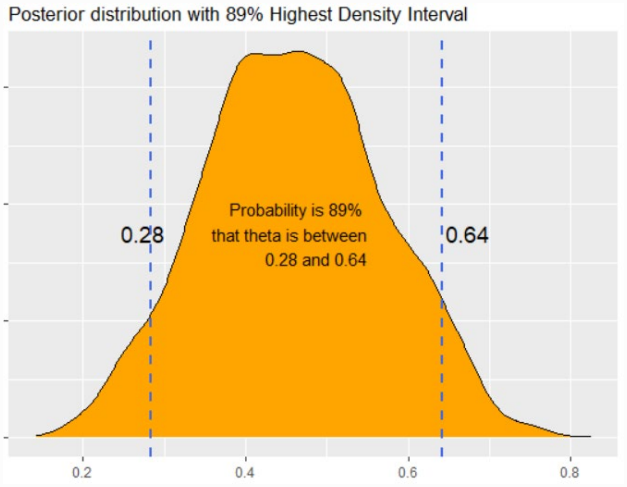
Logit(θ) = ln(θ/(1- θ))

Frequentist interval estimation: ”Confidence interval”

- $\hat{\theta} = \hat{\theta}(y)$ is a single guess from a single sample
- Aim is to create interval with plausible values of θ
- Confidence interval $\hat{\theta} \pm 1.96 * SE(\hat{\theta})$
- Sampling distribution ensures we can make statements on θ
- Bounds are random but θ is a fixed unknown value
- Interpretation
  - A set of values of θ that are compatible with the observed data and the statistical model we chose to describe data with
  - Of all possible samples I could have obtained, 95% of those samples would generate an interval which actually contain the true population value
- Previous example: 20 coin tosses. $\hat{\theta}$ = 0.45, 95%CI: 0.25; 0.66

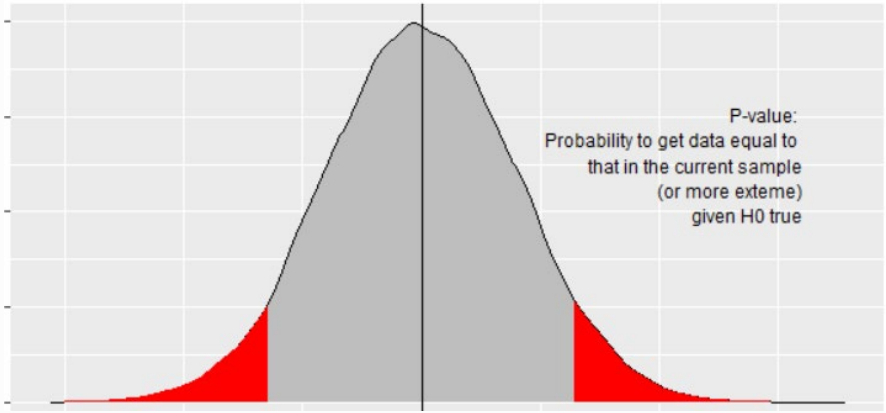
Bayesian interval estimation: ”Credible interval”

- Interval within which an unobserved parameter value falls with a particular probability
  - The range of values of θ that θ will be in with 95% probability
  - Range containing a particular percentage of probable values, e.g. 95%
  - Often CIs are computed with 89% intervals
- Bounds are fixed but θ is a random variable
- Highest Density Interval (HDI) is a credible interval
  - All points within this interval have a higher probability density than points outside the interval.



Frequentist hypothesis testing

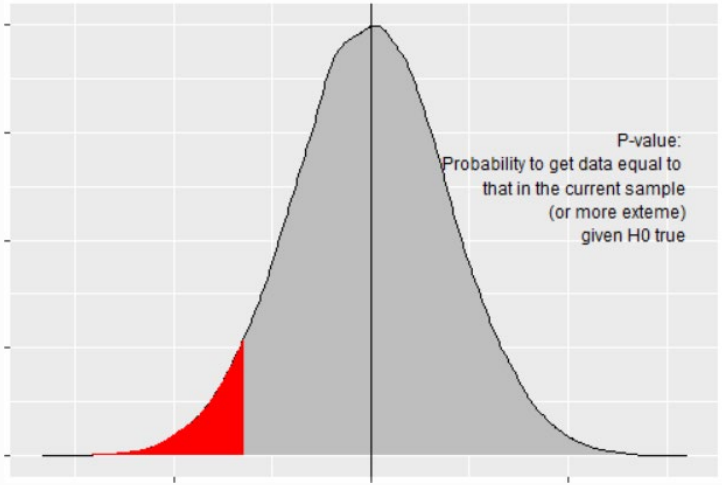
- H0: $\theta=0.5$ vs H1: $\theta \neq 0.5$
- Previous example: 20 coin tosses:
- $\hat{\theta} = 0.45$, yield p-value = 0.655.
- Not reject H0
- Note:
  - H0 is never accepted (θ can never be *exactly* 0.5)
  - P-value depends on study design (sample size)
    - P-value is a probability about random data given fix θ



$\theta=0.5 \rightarrow \text{logit}(\theta)=0$

Frequentist hypothesis testing

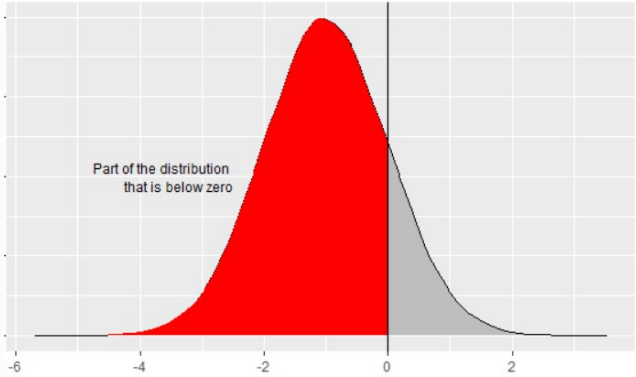
- One-sided test
- H0: $\theta \geq 0.5$ vs H1: $\theta < 0.5$
- $\hat{\theta} = 0.45$, yield p-value = 0.328 (two-sided p-value/2).
- Rarely used in practice
  - Non-inferiority studies



$\theta=0.5 \rightarrow \text{logit}(\theta)=0$

Bayesian approaches related to hypothesis testing

- Probability of Direction (pd)
  - Probability θ strictly positive (or negative)
  - pd = 0.66
  - H0: θ ≥ 0.5 vs H1: θ < 0.5



- pd = P(H1|y) = 0.66
- 1-pd = P(H0|y) = 0.34 ≈ P-value one-sided frequentist test
- 2*(1-pd) = 0.68 ≈ P-value two-sided frequentist test

P-value: Degree of evidence in data in favor of H0
Probability of direction: Probability of H0 given data

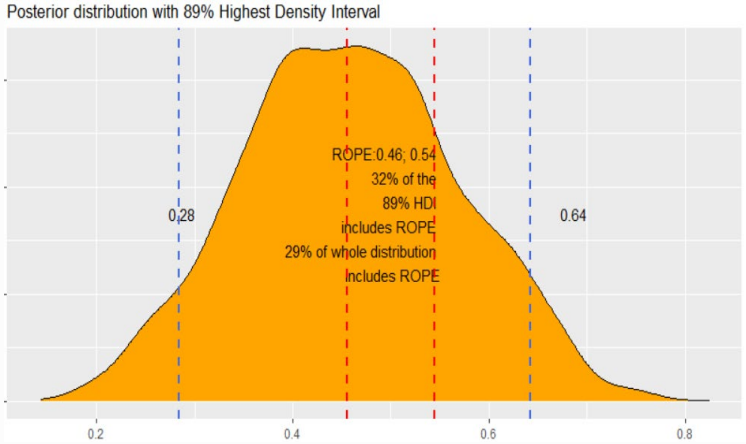
θ=0.5 → logit(θ)=0

Bayesian approaches related to hypothesis testing

- Again, P(θ=0.5|y) = 0
- What is the probability θ is in an *area around* 0.5?
- If θ is in area around 0.5 then *θ is practically 0.5.*
- Region of Practical Equivalence (ROPE)
  - How to define the ROPE range?
    - *Suggestion:* -0.1 to 0.1 of a standardized parameter (negligible effect size according to Cohen, 1988)
  - Proportion of the 89% most probable values (the 89% CI) which are not null, *i.e.*, which are outside this range
  - How big part of the distribution is in area around 0.5?

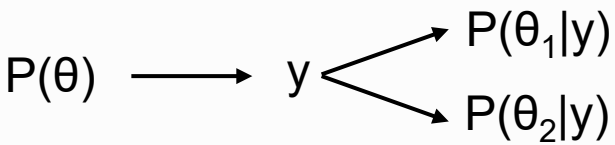
Bayesian approaches related to hypothesis testing

- Previous example: 20 coin tosses:
  - Proportion of the 89% most probable values (the 89% CI) which are not null, *i.e.*, which are outside this range
    - ROPE around θ= 0.5 is here chosen to 0.46; 0.55 (*default*)
  - A majority (89%) of PD is in 0.28; 0.64.
  - 32% of this majority is within ROPE. 29% of the whole distribution consists of ROPE



Bayes factor

- Relative evidence of one “model”/ parameter value over another in light of data



- H0: $\theta=\theta_1$ vs H1: $\theta = \theta_2$

$$P(H_0|y) \propto f(y|H_0) * P(H_0)$$
$$P(H_1|y) \propto f(y|H_1) * P(H_1)$$

$$\underbrace{\frac{P(H_0|y)}{P(H_1|y)}}_{\text{Posterior odds}} = \underbrace{\frac{P(y|H_0)}{P(y|H_1)}}_{\text{Likelihood ratio}} * \underbrace{\frac{P(H_0)}{P(H_1)}}_{\text{Prior odds}}$$

Bayes factor

- BF = $\frac{P(y|H_0)}{P(y|H_1)} = \frac{P(H_0|y)}{P(H_1|y)} * \frac{P(H_1)}{P(H_0)}$
- The likelihood of Y=y given H0 relative to H1 (likelihood ratio $\frac{P(y|H_0)}{P(y|H_1)}$)
- The update in favor of H0 by Y=y ($\frac{P(H_0|y)}{P(H_0)}$) relative to the update in favor of H1 by Y=y ($\frac{P(H_1|y)}{P(H_1)}$)
- Bayes factor indicates the degree by which the mass of the posterior distribution has shifted further away from or closer to the null value(s) (relative to the prior distribution)
- Indicating if the null value has become less or more likely given the observed data

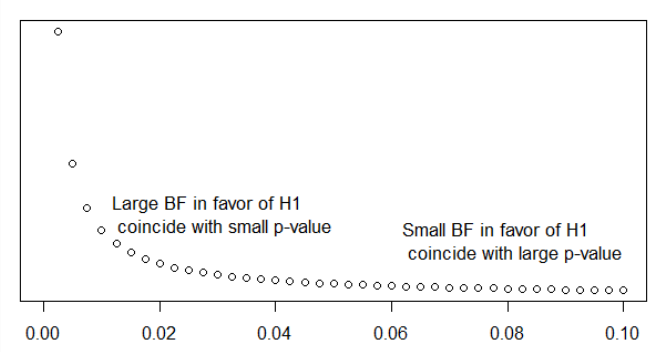
$$\frac{P(H_0|y)}{P(H_0)} \text{ vs } \frac{P(H_1|y)}{P(H_1)}$$

Bayes factor

- $BF_{10} = \frac{P(y|H_1)}{P(y|H_0)} = 8$ means that there is 8 times more evidence for H_1 than H_0 (and is equivalent to a $BF_{01} = \frac{P(y|H_0)}{P(y|H_1)} = 1/8 = 0.125$)

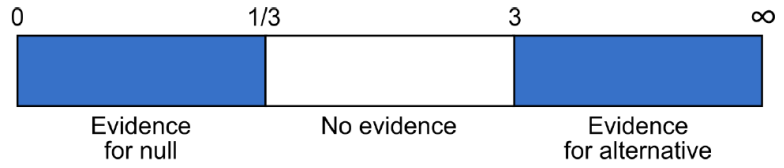
BF_{10}	Interpretation
> 100	Extreme evidence for H_1
30 - 100	Very strong evidence for H_1
10 - 30	Strong evidence for H_1
3 - 10	Moderate evidence for H_1
1 - 3	Anecdotal evidence for H_1
1	Equal evidence for H_1 and H_0
1/3 - 1	Anecdotal evidence for H_0
1/10 - 1/3	Moderate evidence for H_0
1/30 - 1/10	Strong evidence for H_0
1/100 - 1/30	Very strong evidence for H_0
$< 1/100$	Extreme evidence for H_0

P-value:
- Function of the sampling distribution $P(y|H_0)$



Bayes factor vs p-value

- P-value:
  - Data that are unlikely under H0 may lead to its rejection, even though these data are just as unlikely under H1.
  - Can´t accept H0
- Bayes factor
  - Gives relative evidence for H1.
  - Can accept H0
- Drawbacks
  - Rely on same assumptions as frequentist statistics
  - Results can be sensitive to priors



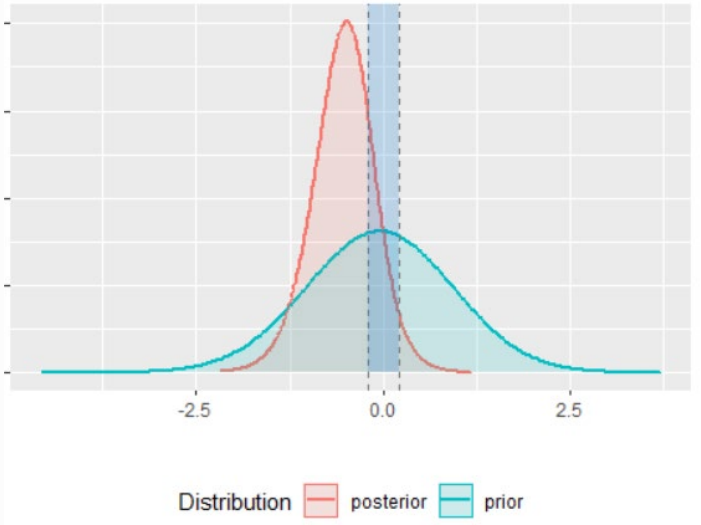
Bayes factor

- Y~Binomial(N, θ)
- Logit(θ) ~ N(μ=0, σ=2.5)
- H0: θ is in area around 0.5 where *θ is practically 0.5 (ROPE)*
- H1: θ is outside this area

H0: θ ∈ (0.45; 0.55)
H1: θ ∉ (0.45; 0.55)

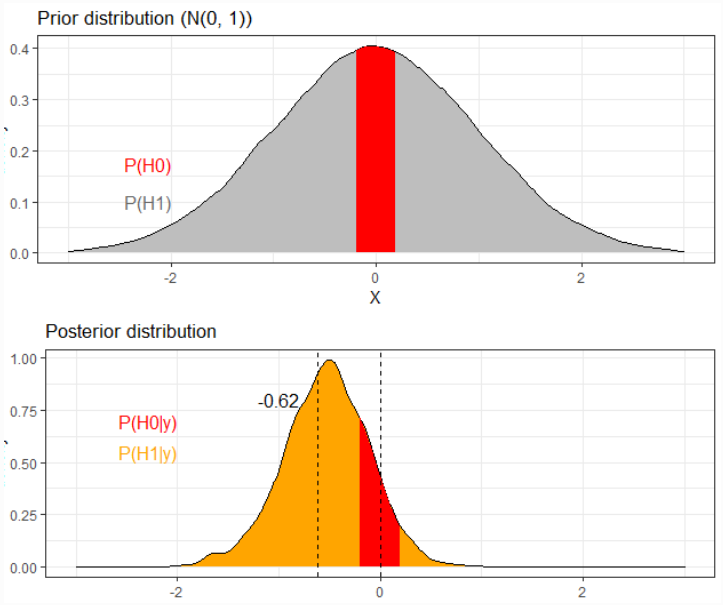
- Data: 20 coin tosses results in 7 heads

θ=0.5 → logit(θ)=0



Bayes factor

- Prior odds = $\frac{P(H_0)}{P(H_1)} = \frac{P(\theta \in (0.45; 0.55))}{P(\theta \notin (0.45; 0.55))}$
- Posterior odds = $\frac{P(H_0|y)}{P(H_1|y)} = \frac{P(\theta \in (0.45; 0.55)|y)}{P(\theta \notin (0.45; 0.55)|y)}$
- Comparing the prior and posterior odds of the parameter falling within or outside the null interval
  - Evidence against H0:
  - BF = $\frac{P(\theta \in (0.45; 0.55)|y)}{P(\theta \notin (0.45; 0.55)|y)} * \frac{P(\theta \notin (0.45; 0.55))}{P(\theta \in (0.45; 0.55))}$
  - BF = 0.947
  - Evidence in favor of H0: 1/0.947 = 1.06
  - Anecdotical evidence in favor of H0



Bayes factor

Y~N(μ, 1)
μ~N(0, 1)

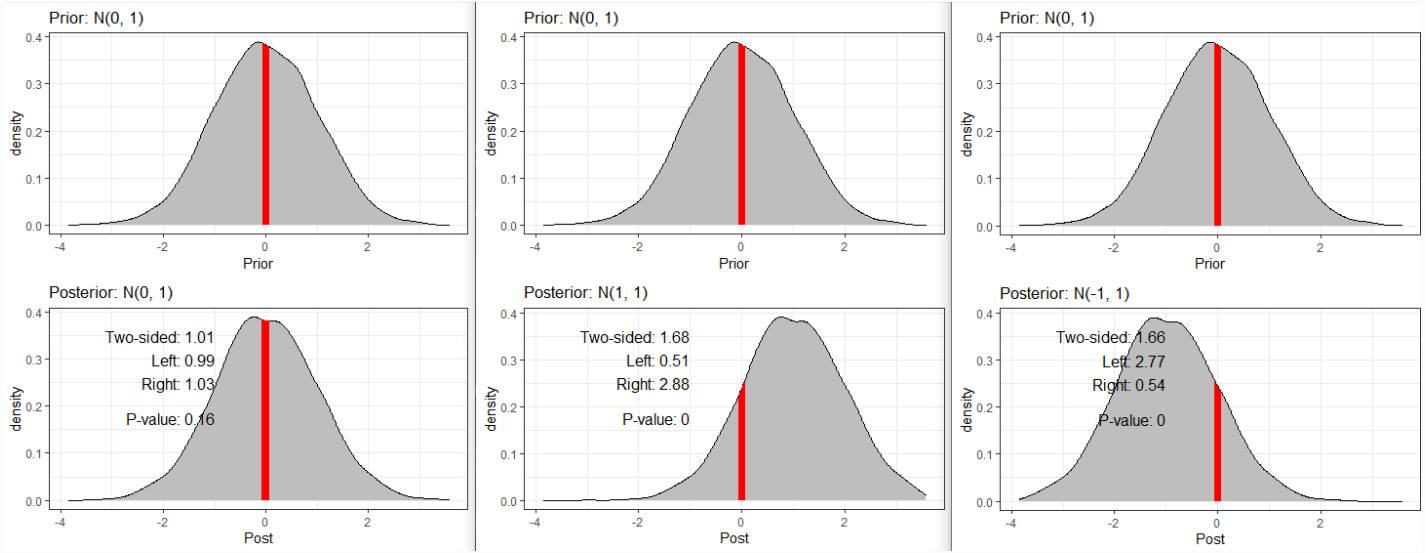
$$BF = \frac{P(y|H_0)}{P(y|H_1)} = \frac{P(H_0|y)}{P(H_1|y)} * \frac{P(H_1)}{P(H_0)}$$

Two-sided H0: μ=0
H1: μ≠0

Left H0: μ≥0
H1: μ<0

Right H0: μ≤0
H1: μ>0

P-value One-sample t-test
H0: μ=0
H1: μ≠0



Bayes factor for model comparison

- Which model does data give most support for?
- M_1: θ = logit(p) = constant (θ = {a})
- M_2: θ = logit(p) = a + b*X (θ = {a, b})

$$\underbrace{\frac{P(M_1|y)}{P(M_2|y)}}_{\text{Posterior odds}} = \underbrace{\frac{P(y|M_1)}{P(y|M_2)}}_{\text{Likelihood ratio}} * \underbrace{\frac{P(M_1)}{P(M_2)}}_{\text{Prior odds}}$$

$P(y|M_1)$ = Marginal likelihood for model M_1

$P(y|M_2)$ = Marginal likelihood for model M_2

Bayes factor for model comparison

$$P(y|M_1) = \int P(y|M_1,\theta)P(\theta|M_1)d\theta = \int P(y|M_1,a)P(a|M_1)da$$
$$P(y|M_2) = \int P(y|M_2,\theta)P(\theta|M_2)d\theta = \int\int P(y|M_2,a,b)P(a|M_2)P(b|M_2)dadb$$

- The marginalized likelihood is the probability of the data given the model type, not assuming any particular model parameter θ

$$\underbrace{\frac{P(M_1|y)}{P(M_2|y)}}_{\text{Posterior odds}} = \underbrace{\frac{P(y|M_1)}{P(y|M_2)}}_{\text{Likelihood ratio}} * \underbrace{\frac{P(M_1)}{P(M_2)}}_{\text{Prior odds}}$$

Frequentist model comparison: Likelihood ratio test (LRT)

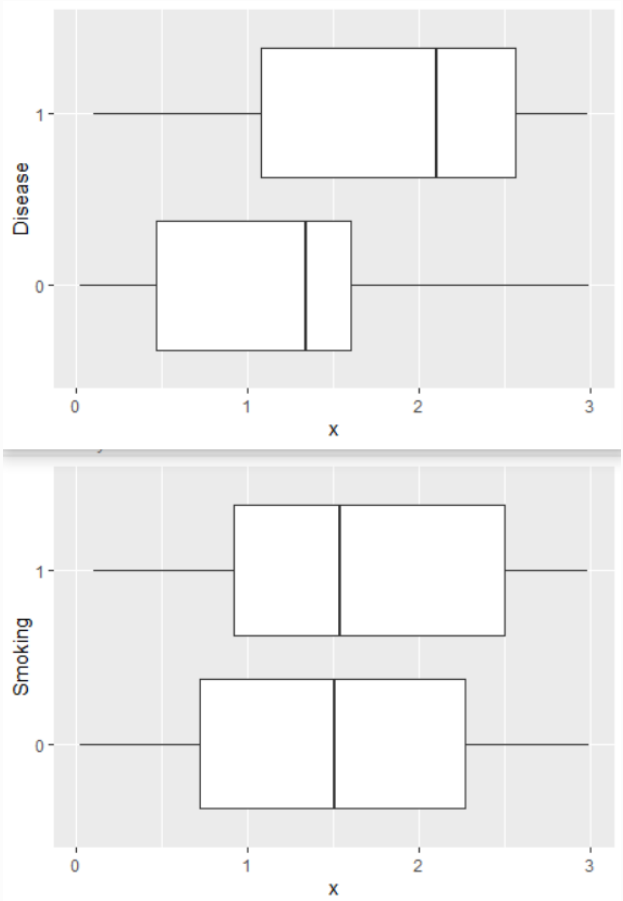
- Compare the goodness of fit of two statistical models
- The LRT compares two hierarchically nested models to determine whether or not adding complexity to your model (i.e., adding more parameters) makes the model significantly more accurate.
- The “hierarchically nested models”
  - the complex model differs only from the simpler (or “nested”) model by the addition of one or more parameters.
- LRT tells us if it is beneficial to add parameters to the model, or if we should stick with our simpler model.

Frequentist model comparison: Likelihood ratio test (LRT)

- H0: Use Nested model. ($\theta = \{a\}$)
- H1: Use Complex model. ($\theta = \{a, b\}$)
  - If you reject the H0: Conclude that the complex model is significantly more accurate than the nested model. Chose the complex model.
  - If you fail to reject the H0: Conclude that the complex model is NOT significantly more accurate than the nested model. Choose to use the nested model instead.
- Test statistic = -2 * [loglikelihood(nested)-loglikelihood(complex)]
  - Sampling distribution: chi-squared distribution.
    - degrees of freedom equal to the difference in dimensionality of the models

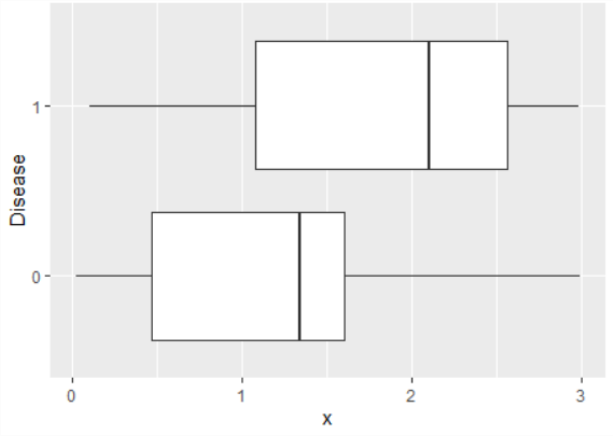
Example model comparison

- Example 70 subjects with information on
  - diagnosis Sick/Healthy
  - Smoking status (smoker Yes/No)
  - measurement of biomarker X
- Aim is to explain risk of being diseased (p)
  - Models
    - D_1: θ = logit(p) = constant (θ = {a})
    - D_2: θ = logit(p) = a + b*X (θ = {a, b})
- Aim is to explain probability of being a smoker (p)
  - Models
    - S_1: θ = logit(p) = constant (θ = {a})
    - S_2: θ = logit(p) = a + b*X (θ = {a, b})



Example model comparison

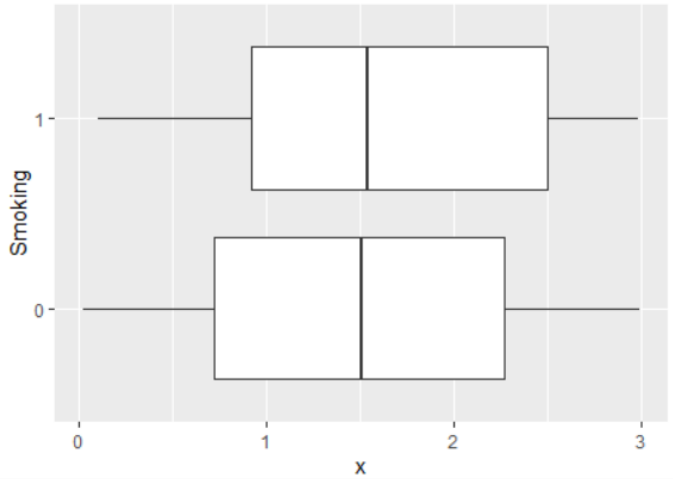
- Model for being diseased:
  - D_1: θ = logit(p) = constant (θ = {a})
  - D_2: θ = logit(p) = a + b*X (θ = {a, b})
- Frequentist analysis
  - logLik(D_1) = -47.8, (nested)
  - logLik(D_2) = -44.8 (complex)
  - -2 * [loglikelihood(D_1)-loglikelihood(D_2)] = 5.87
  - P-value = P(Teststatistic ≥5.87|H0 true) = 0.015.
  - Reject H0
- Risk of being dieased can be explain by biomarker



Has a Chi2 distribution with df = 1

Example model comparison

- Model for being smoker:
  - S_1: θ = logit(p) = constant (θ = {a})
  - S_2: θ = logit(p) = a + b*X (θ = {a, b})
- Frequentist analysis
  - logLik(S_1) = -48.5, (nested)
  - logLik(S2) = -48.4 (complex)
  - -2 * [loglikelihood(S_1)-loglikelihood(S_2)] = 0.33
  - P-value = P(Teststatistic ≥0.33|H0 true) = 0.56.
  - Do NOT Reject H0
- Risk of being smoking not further explained by biomarker



Has a Chi2 distribution with df = 1

Example model comparison

- Model for being diseased:
  - D_1: θ = logit(p) = constant (θ = {a})
  - D_2: θ = logit(p) = a + b*X (θ = {a, b})
  - a ~ N(0, 2.5)
  - b ~ N(0, 2.5)
- B_{21} = 2.028
  - anecdotal evidence in favour of D_2

- Model for being smoker:
  - D_1: θ = logit(p) = constant (θ = {a})
  - D_2: θ = logit(p) = a + b*X (θ = {a, b})
  - a ~ N(0, 2.5)
  - b ~ N(0, 2.5)
- B_{21} = 0.115
  - moderate evidence against D_2

Summary

- Parameter estimation
  - Frequentist: Maximize the likelihood
  - Bayesian: Maximize the posterior distribution
- Interval estimation
  - Frequentist:
    - A set of values of θ that are compatible with the observed data and the statistical model we chose to describe data with
    - Of all possible samples I could have obtained, 95% of those samples would generate an interval which actually contain the true population value
  - Bayesian: Range of values of θ with high probability
    - The range of values of θ that θ will be in with 95% probability
- Bayes factor
  - Relative evidence of one “model”/ parameter value over another in light of data

Posterior Predictive Distribution

- Consider a new data sample $\tilde{Y}$.
- Want to find p($\tilde{Y}$|y); the probability of the new data sample given our current data y.
  - Used to forecast
  - To check model.
- p($\tilde{Y}$|y) = Posterior predictive distribution
- p($\tilde{Y}$|y) = p($\tilde{Y}$|θ)* p(θ|y)

Sampling Distribution **Posterior Distribution**

Get p($\tilde{Y}$|y) in two steps:

1. Sample θ´s from p(θ|y)
2. Sample $\tilde{Y}$´s from p($\tilde{Y}$|θ)

Posterior Predictive Distribution

- The probability distribution for a new data sample $\tilde{Y}$ given our current data y.
- p($\tilde{Y}$|y) for prediction
- p($\tilde{Y}$|y) for model checking.
  - Does predictions mimic observed data, i.e. is model OK?

Posterior Predictive Distribution

Example:

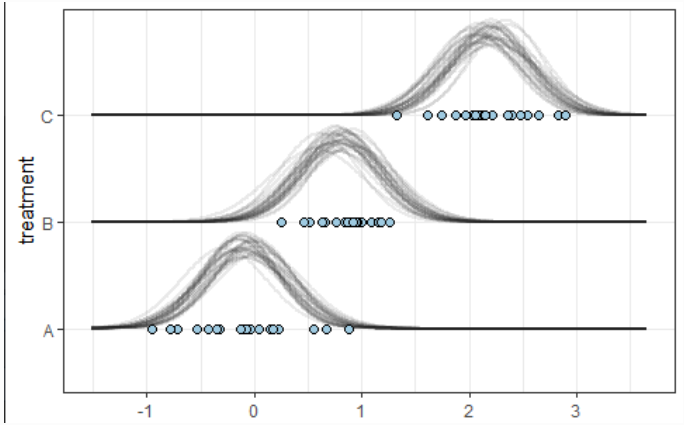
3 treatment, n=20/group

$Y \sim N(\mu, \sigma)$

$\mu_i \sim t(3, 0.9, 2.5), i = 1, 2, 3$

$\sigma \sim$ half-t(3, 0, 2.5)

Posterior predictive distribution



Posterior Predictive Distribution

- Example: Linear regression

  $Y \sim N(\mu, \sigma)$

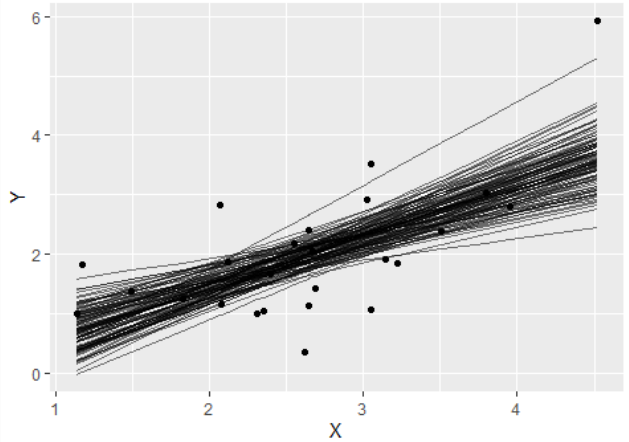
  $\mu = \beta_0 + \beta_1 * X$

  $\beta_0 \sim t(3, 1.9, 2.5)$

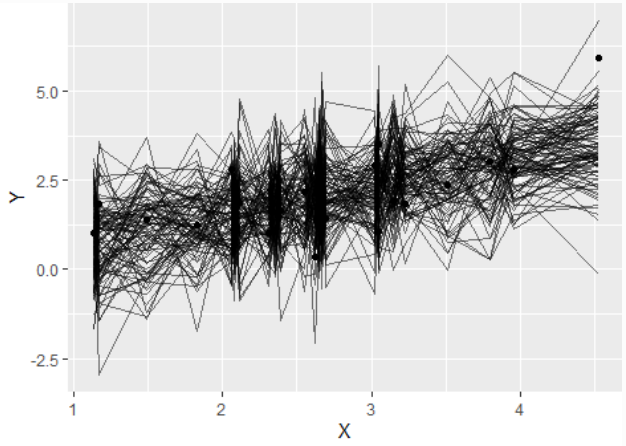
  $\beta_1 \sim t(10, 0, 1)$

  $\sigma \sim$ half-t(3, 0, 2.5)

Predicted mean curves based on the posterior distribution

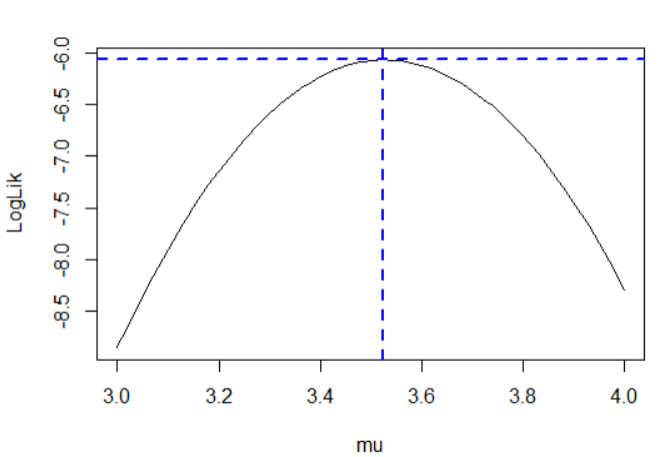


Individual predictions based on the posterior distribution



One sample normal distribution, single parameter

- Frequentist model:
  - Y ~ N(μ, σ = 1)
  - Data: Random sample (10 observations), estimate μ
  - Likelihood maximized for μ = 3.53 (same as $\bar{y} = {\Sigma y}/{10}$)



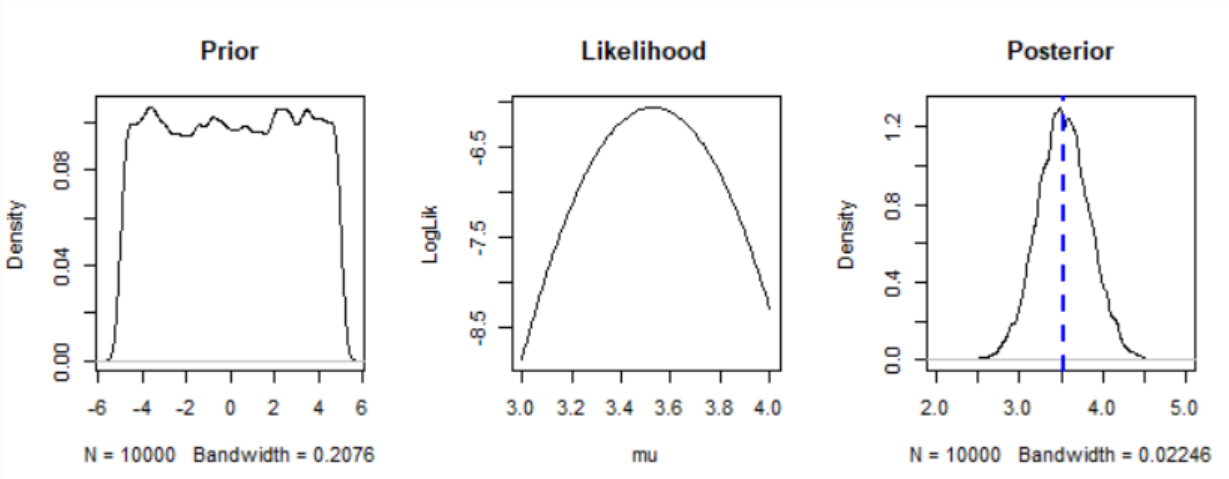
95% confidence intervals

$\bar{y}$ =3.54
Lower: $\bar{y}$ - 1.96*SE = 2.94
Upper: $\bar{y}$ + 1.96*SE = 4.12

One sample normal distribution, single parameter

- Bayesian model:
  - $Y \sim N(\mu, \sigma = 1)$
    - $\mu \sim U(-5, 5)$
  - Data: Random sample (10 observations),

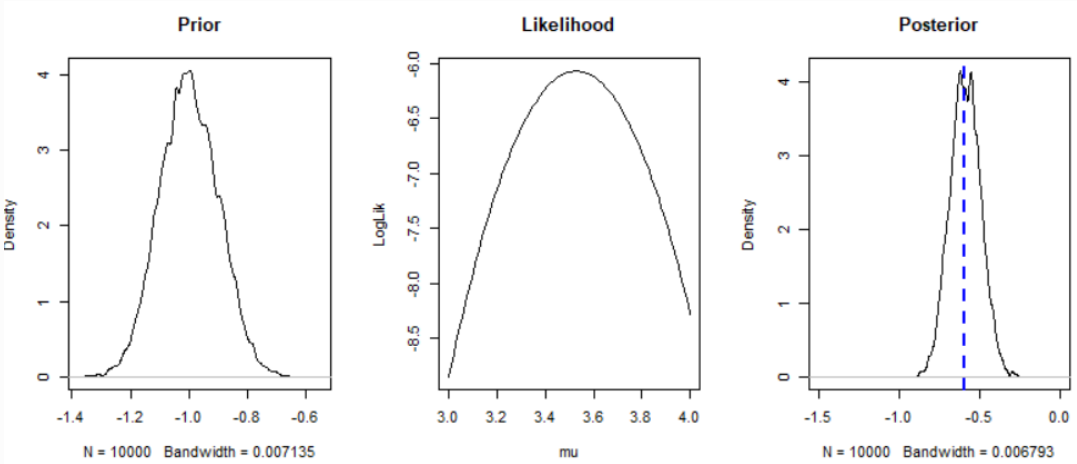
The mean value of posterior μ is 3.53 (same as frequentist estimate 3.53)



*Uniform prior not actually a good idea

One sample normal distribution, single parameter

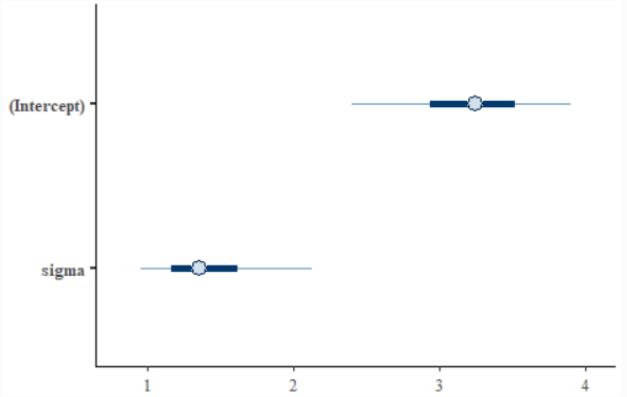
- Bayesian model:
  - Y ~ N(μ, σ = 1)
    - μ ~ N(-1, 0.1) (strong prior belief of μ = -1)
  - Data: Random sample (10 observations),

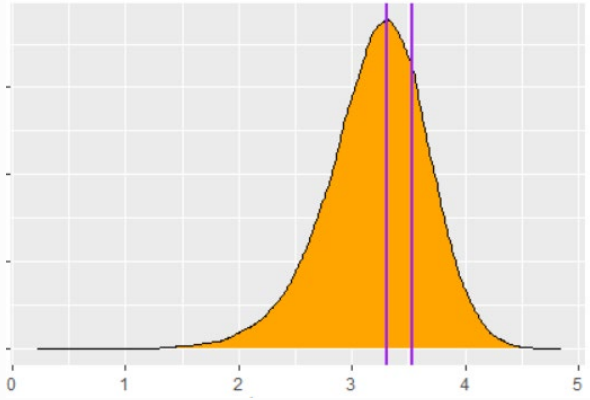


The mean value of posterior μ is 0.59
A compromise between strong prior belief of -1 and data indicating 3.53

One sample normal distribution, single parameter

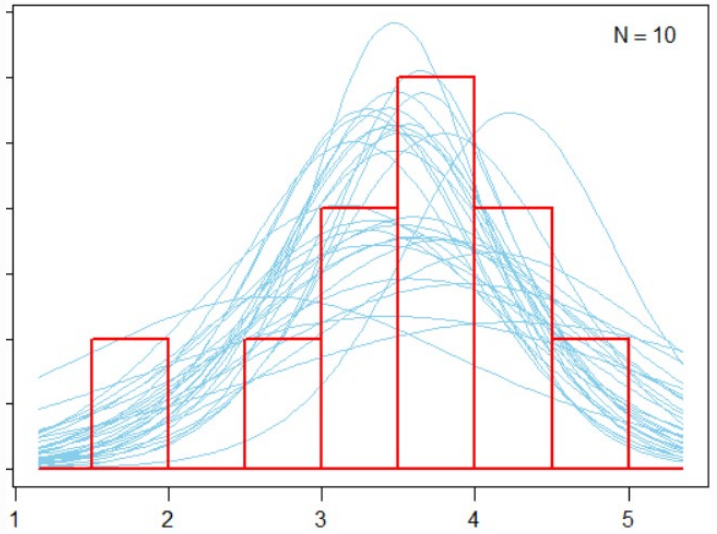
- Bayesian model:
  - Y ~ N(μ, σ = 1)
    - μ ~ N(0, 1)
    - σ ~Exponentialt(rate = 1)
  - Maximum posterior estimate (MAP): 3.31
  - 95% Credible interval: 2.27; 4.07
- Evidence again H0: μ = 3
  - BF = 0.076
- Evidence again H0: μ ≤ 3
  - BF = 2.32





One sample normal distribution, single parameter

- Posterior predictive distribution



https://cran.r-project.org/web/packages/BEST/vignettes/BEST.pdf

Two independent samples. Normal distribution

- Frequentist two-sample t-test
- $Y \sim N(\mu_i, \sigma)$, $\mu_i = \begin{cases} \mu_1, \text{Group} = 1 \\ \mu_2, \text{Group} = 2 \end{cases}$
- H0: $\mu_1 - \mu_2 = 0$
- H1: $\mu_1 - \mu_2 \neq 0$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

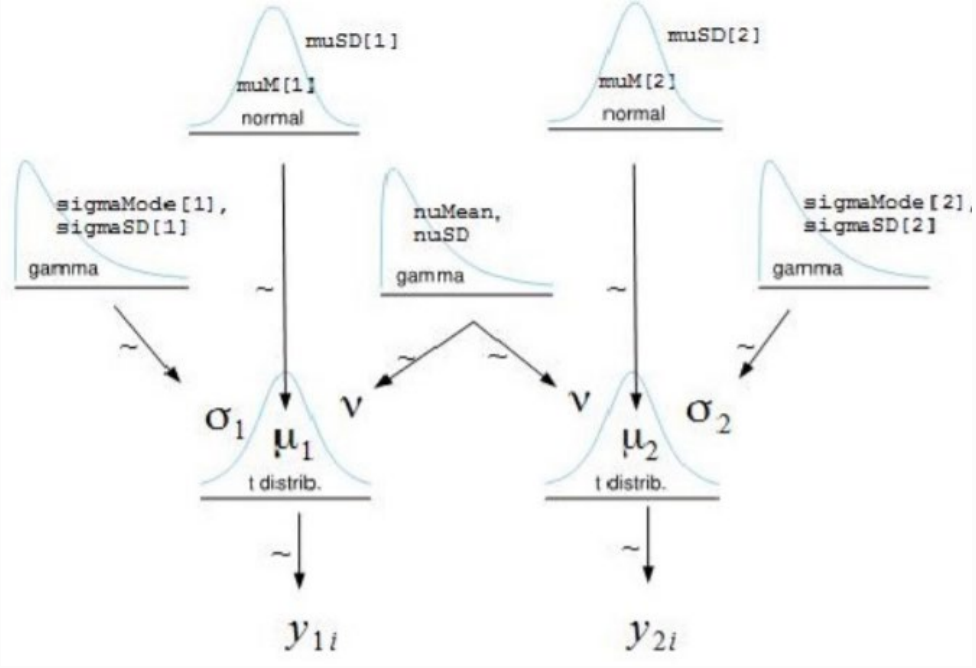
$$s_p = \sqrt{\frac{(n_1 - 1)\, s_{X_1}^2 + (n_2 - 1)\, s_{X_2}^2}{n_1 + n_2 - 2}}$$

- t | H0 ~ t-dist ($n_1 + n_2 - 2$)

(sampling distribution)

Two independent samples. Normal distribution

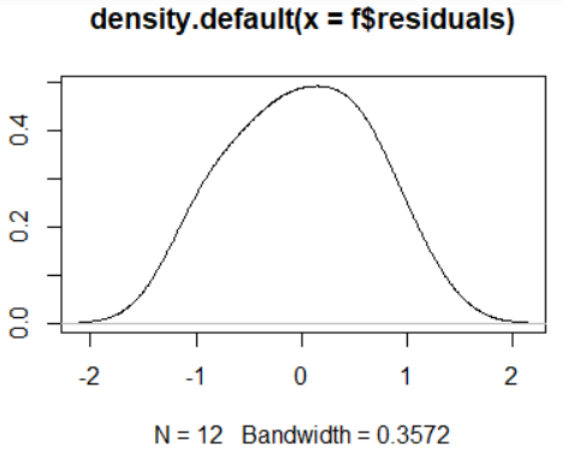
- Meredith & Kruschke´s Bayesian t-test
  - Y ~t-distribution(μ_i, σ_i, v) ,
  - $\mu_i = \begin{cases} \mu_1 \text{ , Group} = 1 \\ \mu_2 \text{ , Group} = 2 \end{cases}$
- H0: $\mu_1 - \mu_2 = 0$
- H1: $\mu_1 - \mu_2 \neq 0$



https://cran.r-project.org/web/packages/BEST/vignettes/BEST.pdf

Two independent samples. Normal distribution

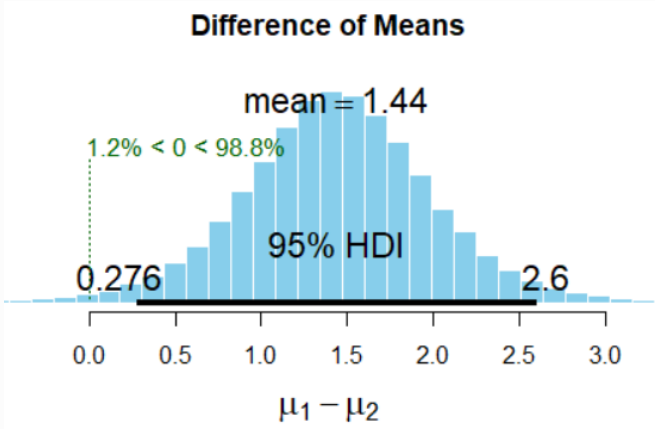
- Example:
  - Group 1 (n=6) consumes a drug which may increase reaction times while Group 2 (n=6) is a control group that consumes a placebo.
- Frequentist analysis:
  - $\bar{y}_1 = 4.69$, $\bar{y}_2 = 3.21$, SD_1= 0.75, SD_2=0.61
  - $\bar{y}_1$ - $\bar{y}_2$ = 1.49 (95% CI: 0.60; 2.37), p-value = 0.004
  - Reject H0: μ_1 - $\mu_2 = 0$
  - P-value (LR test): 0.001
  - R2 = 0.59
    - Two group characterization explains variability



https://cran.r-project.org/web/packages/BEST/vignettes/BEST.pdf

Two independent samples. Normal distribution

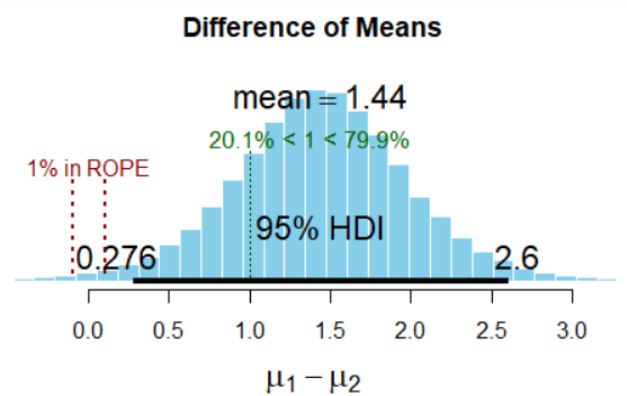
- Bayesian analysis:
  - Based on previous experience, we expect reaction times to be approximately 6 secs, but they vary a lot, so we'll set μ~N(6, 2).
  - Probability that the true value is greater than zero is shown as 98.8%



https://cran.r-project.org/web/packages/BEST/vignettes/BEST.pdf

Two independent samples. Normal distribution

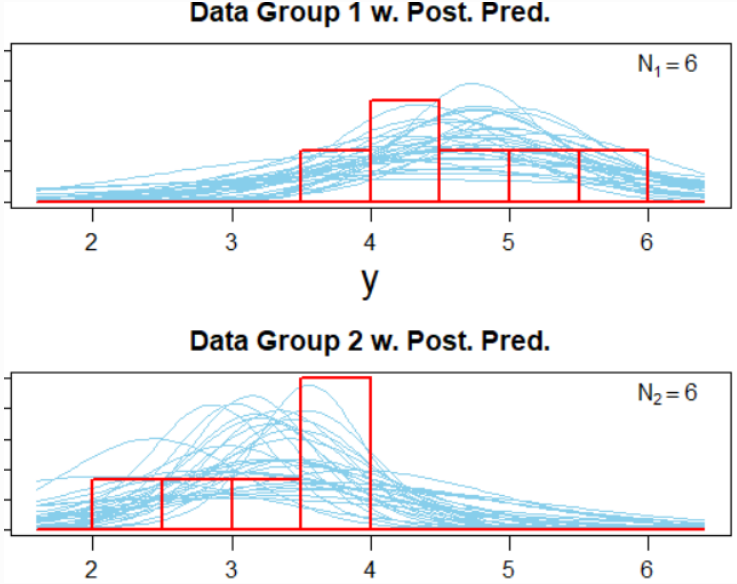
- Bayesian analysis:
  - increase reaction time of 1 unit may indicate that users of the drug should not drive or operate equipment
  - high probability that the reaction time increase is > 1
  - Probability that the true value is greater than one is shown as 79.9%



https://cran.r-project.org/web/packages/BEST/vignettes/BEST.pdf

Two independent samples. Normal distribution

- Bayesian analysis:
  - Posterior predictive distribution



https://cran.r-project.org/web/packages/BEST/vignettes/BEST.pdf

Two-sample t-test

- https://rpsychologist.com/d3/bayes/

The role of the prior: Shrinkage (”regularization”)

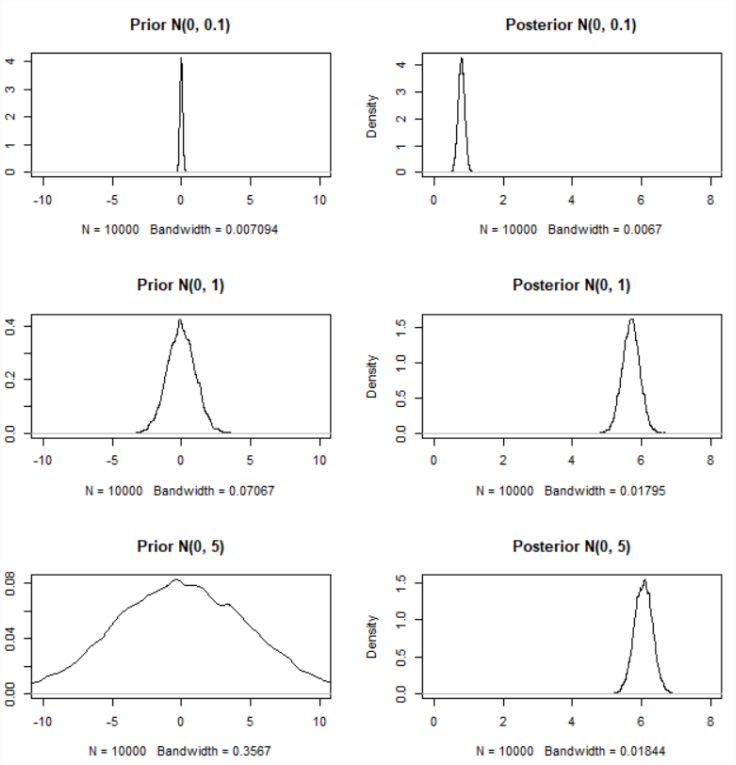
- The posterior is a weighted average of the prior and likelihood (data).
- Changes in position of prior or likelihood are reflected in posterior.
- The weighting towards the likelihood increases as more data is collected
  - models with a lot of data are less dependent on priors.
- Exception to this is “zero” priors.

The role of the prior: Shrinkage (”regularization”)

- Shrinkage:
  - Downplay the role of data
  - Avoid overfitting
- Overfitting:
  - Estimate a model that closely or exactly to a particular set of data, and may therefore fail to fit additional data or predict future observations reliably (lack external validity)
  - An overfitted model is a statistical model that contains more parameters than needed
  - An overfitted models tend to over-emphasize effects

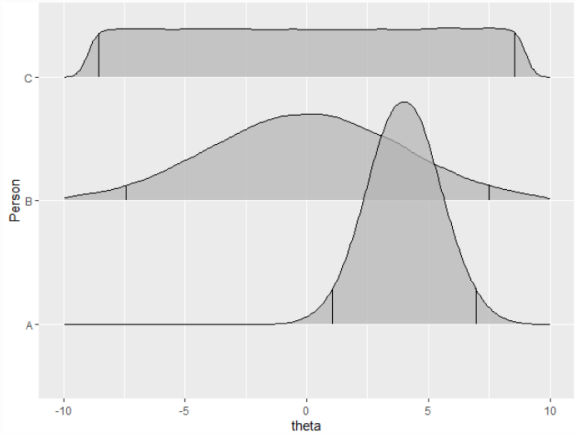
The role of the prior: Shrinkage (”regularization”)

- Posterior distribution = combination of data and subjective belief
- Shrink down posterior down to the prior
  - Data infleunce less
  - Avoid over-optimistic parameter estimates
- Chose prior to reflect desired trade-off between prior belief and new data



How to chose prior?

- Reflect ”abscence” of prior knowledge (noninformative prior)
  - Do **not** use uniform distribution
    - Puts unrealistic restrictions on parameters
    - No regulatization
  - Better use regularizing prior with large variability



How to chose prior?: Conjugate prior

- Conjugate prior: Chose prior to make things easy
  - If the posterior distributions P(θ|y) are in the same probability distribution family as the prior probability distribution P(θ), the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior for the likelihood function P(y|θ).
  - Possible with a closed-form expression of P(θ|y)
    - No numerical integration nedded
  - Exemple
    - Y ~Binomial(N, p)
    - p ~ Beta(α, β) yields a posterior:
    - p|y ~ Beta(α+y, β+(n-y))

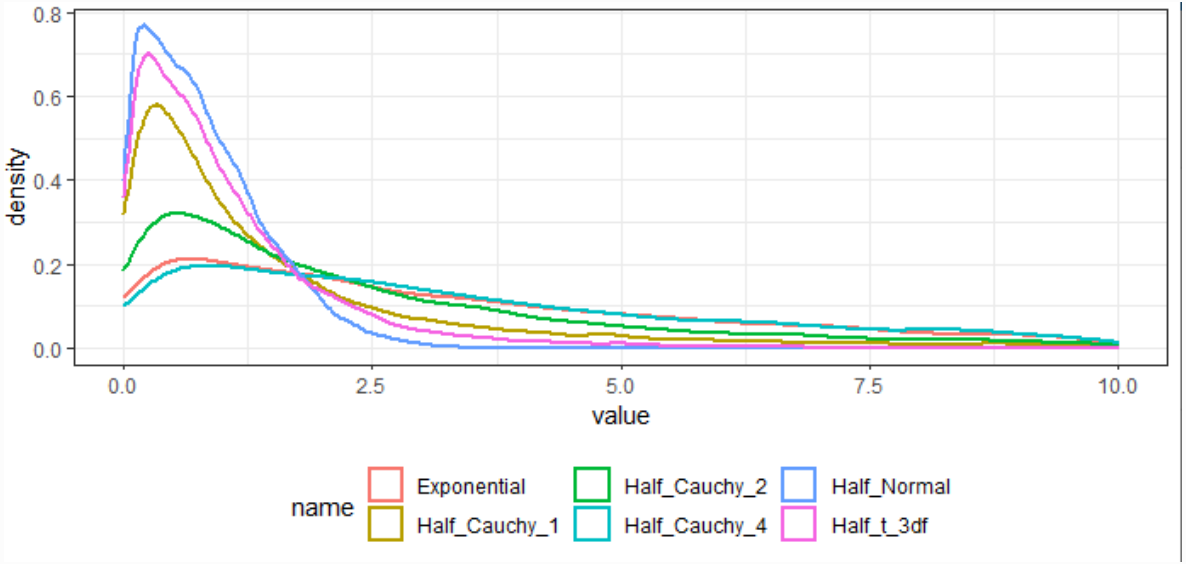
α, β = ”hyperparameters”
(parameters for parameters)

How to chose prior?: Conjugate prior

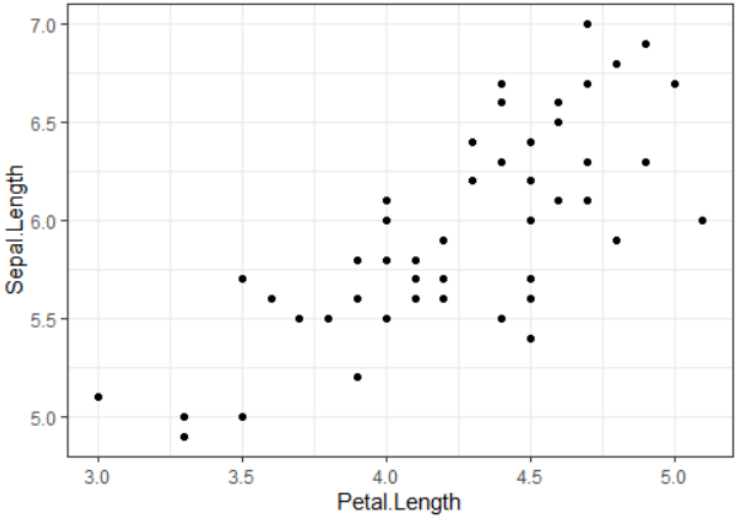
Likelihood	Prior	Posterior
Bernoulli	$\text{Beta}(\alpha, \beta)$	$\text{Beta}(\alpha + \sum_{i=1}^{n} X_i, \beta + n - \sum_{i=1}^{n} X_i)$
Binomial	$\text{Beta}(\alpha, \beta)$	$\text{Beta}(\alpha + \sum_{i=1}^{n} X_i, \beta + \sum_{i=1}^{n} N_i - \sum_{i=1}^{n} X_i)$
Poisson	$\text{Gamma}(\alpha, \beta)$	$\text{Gamma}(\alpha + \sum_{i=1}^{n} X_i, \beta + n)$
Multinomial	$\text{Dirichlet}(\boldsymbol{\alpha})$	$\text{Dirichlet}(\boldsymbol{\alpha} + \sum_{i=1}^{n} \boldsymbol{X}_i)$
Normal	Normal-inv-Γ	Normal-inv-Γ

Prior for standard deviation?

- $Y \sim N(\mu, \sigma)$
- $\sigma > 0$



Linear regression

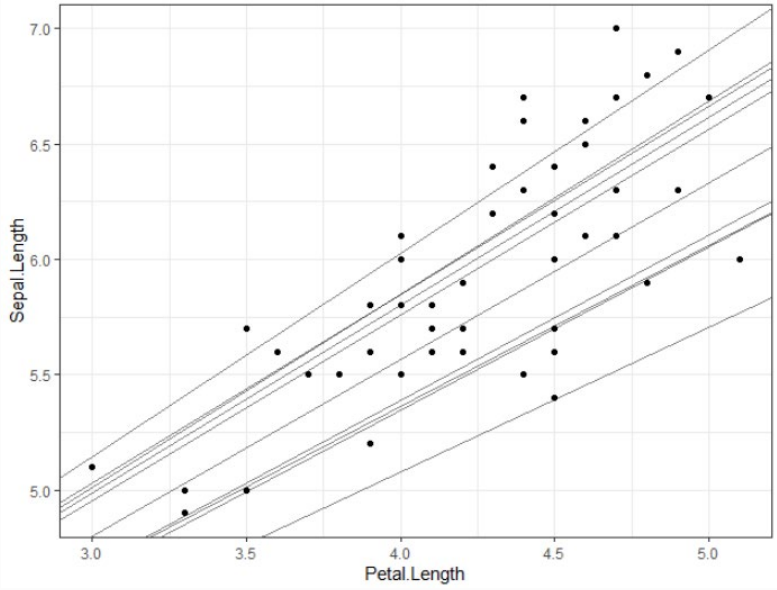


- Aim is to explain how Sepal length depend on Petal length

Linear regression

- Probability model:
  - Sepal length ~ N(μ, σ)
  - $\mu = \beta_0 + \beta_1 *$ Petal length
  - β_0 , β_1 and σ unknown

- Which line describe relationship best?
  - Which values of β_0 , β_1 and σ are reasonable?



Linear regression

Frequentist:
Sepal length ~ N(μ, σ)
$\mu = \beta_0 + \beta_1 *$ Petal length

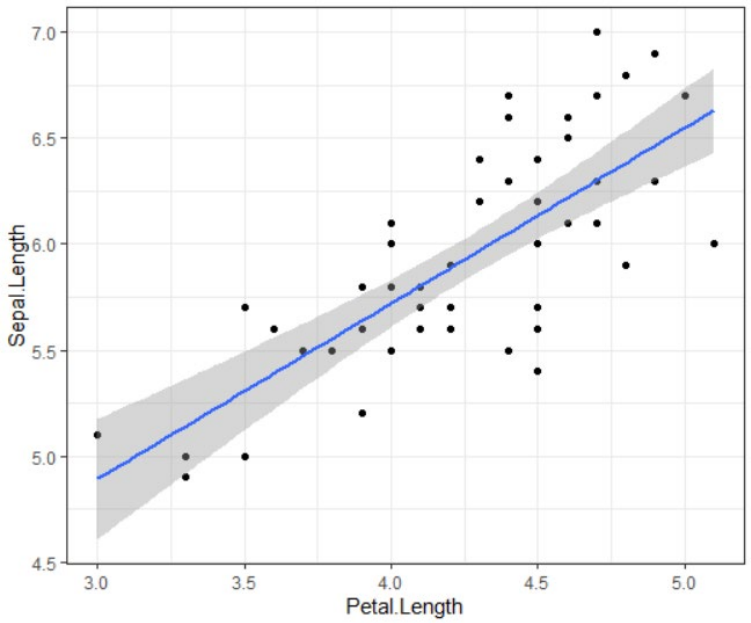
Maximum likelihood estimates of β_0 and β_1

$\hat{\beta}_0 = 2.41$

$\hat{\beta}_1 = 0.83 \quad (95\%\ \text{CI}: 0.62; 1.04)$

P-value < 0.0001 (H0: $\beta_1 = 0$)

Likelihood ratio test
H0: $\mu = \beta_0$
H1: $\beta_0 + \beta_1 *$ Petal length
P-value < 0.001. Reject H0



Linear regression

Bayesian:

Sepal length ~ N(μ, σ)

$\mu = \beta_0 + \beta_1 *$ Petal length

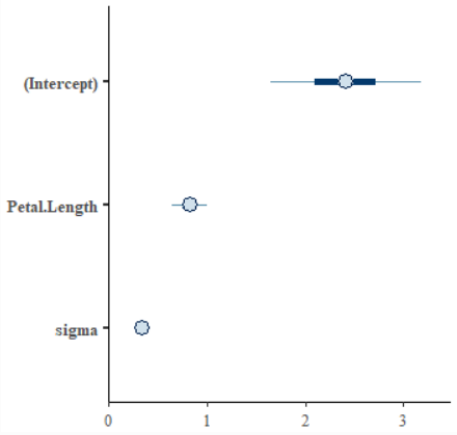
β_0 ~ N(5.93, 2.5)

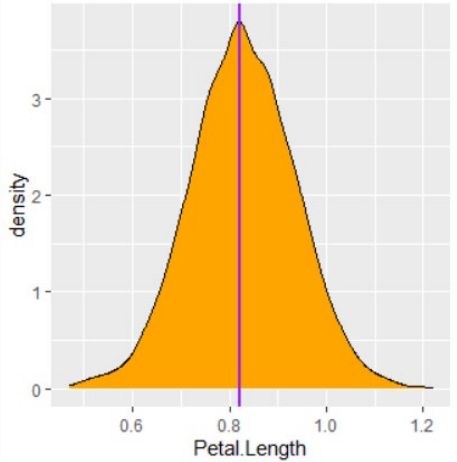
β_1 ~ N(0, 2.5)

σ ~ Exponential(rate = 1)

Maximum posterior estimate (MAP):

$\hat{\beta}_1 = 0.83 \quad (95\%\ \mathrm{HDI}: 0.62; 1.04)$

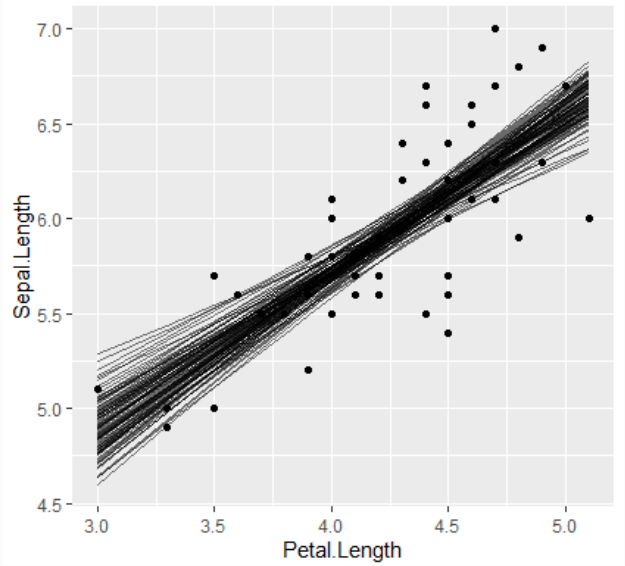




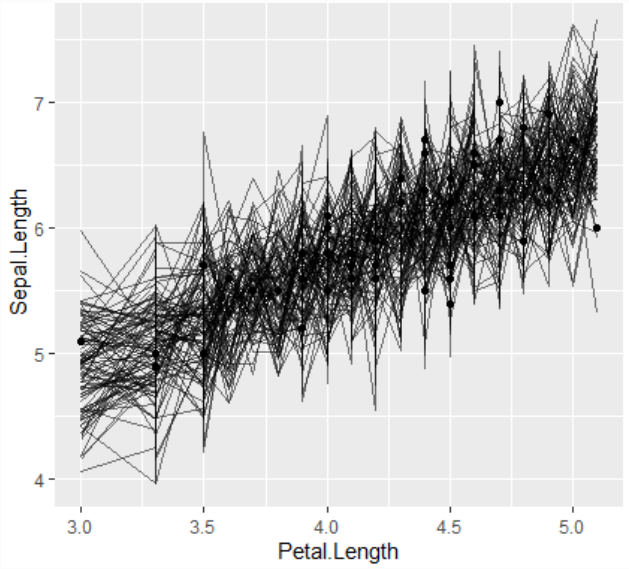
Linear regression

- Posterior predictions

Predicted mean curves based on the posterior distribution



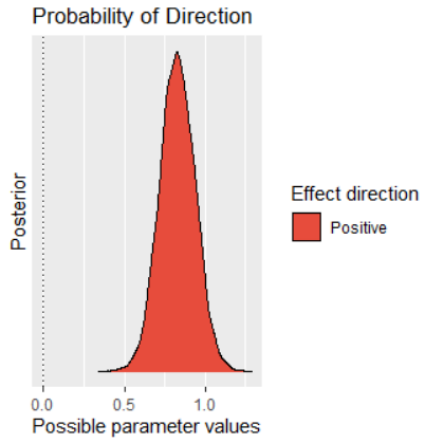
Individual predictions based on the posterior distribution

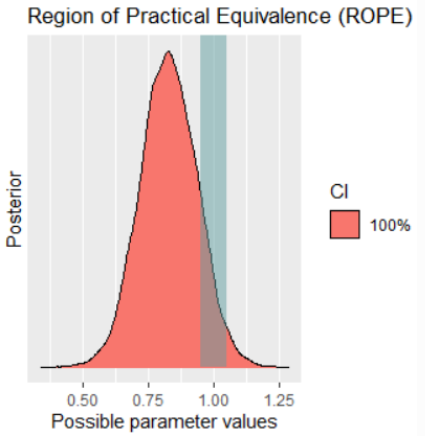


Linear regression

- Probability of Direction (pd)
- Probability β_1 strictly positive (or negative)
- H0: $\beta_1 \leq 0$ vs H1: $\beta_1 \geq 0$
- pd = 0.

- Region of Practical Equivalence (ROPE):
- What is the probability β_1 is in an *area around* 1?
- ROPE: (0.95; 1.05)
  - How big part of the distribution is in area around 1?
- ROPE = 11%





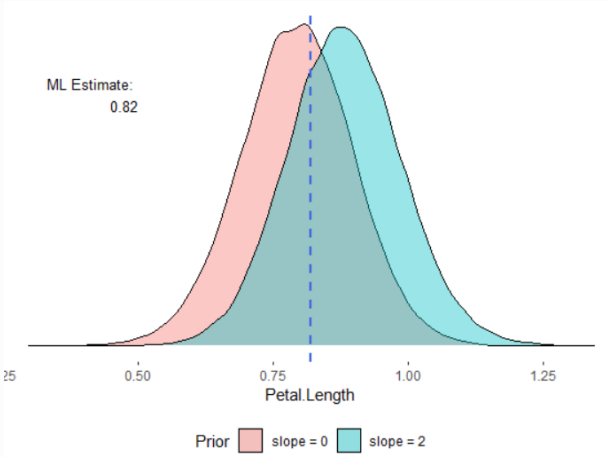
Linear regression

- Model 1:
  - $\mu = \beta_0$
  - $\beta_0 \sim N(5.93, 2.5)$
  - $\sigma \sim$ Exponential(rate = 1)
- Model 2:
  - $\mu = \beta_0 + \beta_1 *$ Petal length
  - $\beta_0 \sim N(5.93, 2.5)$
  - $\beta_1 \sim N(0, 2.5)$
  - $\sigma \sim$ Exponential(rate = 1)
- $BF_{21} > 100$: Extreme evidence for Model 2

Linear regression

- Very strong prior belief that $\beta_1 = 0$
  - $\mu = \beta_0 + \beta_1 *$ Petal length
  - $\beta_0 \sim N(6.9, 2.5)$
  - $\beta_1 \sim N(\mathbf{0, 0.5})$
  - $\sigma \sim$ Exponential(rate = 1)

- Very strong prior belief that $\beta_1 = 2$
  - $\mu = \beta_0 + \beta_1 *$ Petal length
  - $\beta_0 \sim N(6.9, 2.5)$
  - $\beta_1 \sim N(\mathbf{2, 0.5})$
  - $\sigma \sim$ Exponential(rate = 1)



Summary

- Posterior predictive distribution
  - The probability distribution for a new data sample $\tilde{Y}$ given our current data y.
- Posterior = Prior + likelihood
  - Damped the influence of data (avoid overfitting = shrinkage)
- Conjugate prior:
  - Posterior & prior on the same form

Learning from data

- Inference takes the form of updating priors to yield posteriors
- Bayesian updating
- Old posterior becomes our new prior $P(\theta|y_1)$
- New data y_2
- New posterior

$$P(\theta|y_1, y_2) \propto P(\theta|y_1) f(y_2|\theta)$$

PRIOR

DATA **LIKELIHOOD**

POSTERIOR PROBABILITY

Belief + data ⟶ New belief ⟶ Newer data ⟶ Even newer belief

Globe tossing example*

- You have a globe representing our planet, the Earth.
- How much of the surface is covered in water?
- Toss the globe up in the air.
  - When you catch it, you will record whether or not the surface under your right index finger is water or land.
- Repeat the procedure

* Example taken from McElreath, R: Statistical Rethinking (2020), see https://xcelab.net/rm/statistical-rethinking/
See also https://bookdown.org/content/3890/

Globe tossing example

- How to learn how much of the surface is covered in water by tossing the globe?
- Specify framework for learning
  - (1) The true proportion of water covering the globe is *p*.
    - Any p between 0 and 1 initially equally plausible
  - (2) A single toss of the globe has a probability *p* of producing a water (W) observation and 1 − *p* of producing a land (L) observation.
  - (3) Each toss of the globe is independent of the others.

Bayesian updating

- Round 1
  - Any p between 0 and 1 initially equally plausible
  - Toss the globe and register the outcome
  - Update your belief about *p* by means of bayes formula
- Round 2

  Prior $P(p|X_1)$

  Likelihood $P(X_2|p)$

  Posterior $P(p|X_1, X_2)$
- Etc, etc …

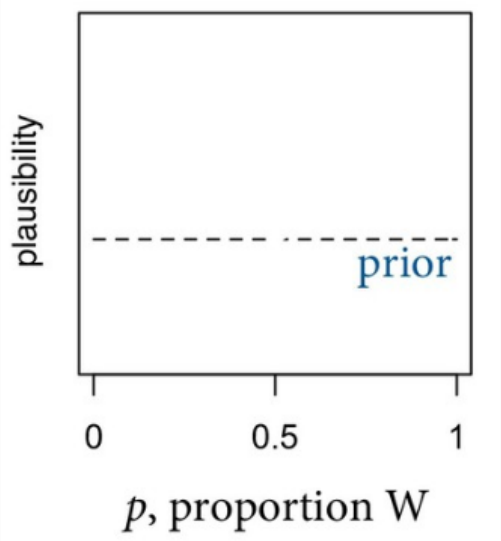
Prior $P(p)$

Likelihood $P(X_1|p)$

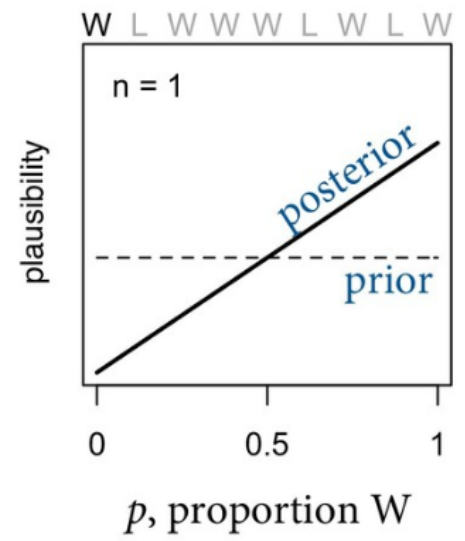
Posterior $P(p|X_1)$

Round 1

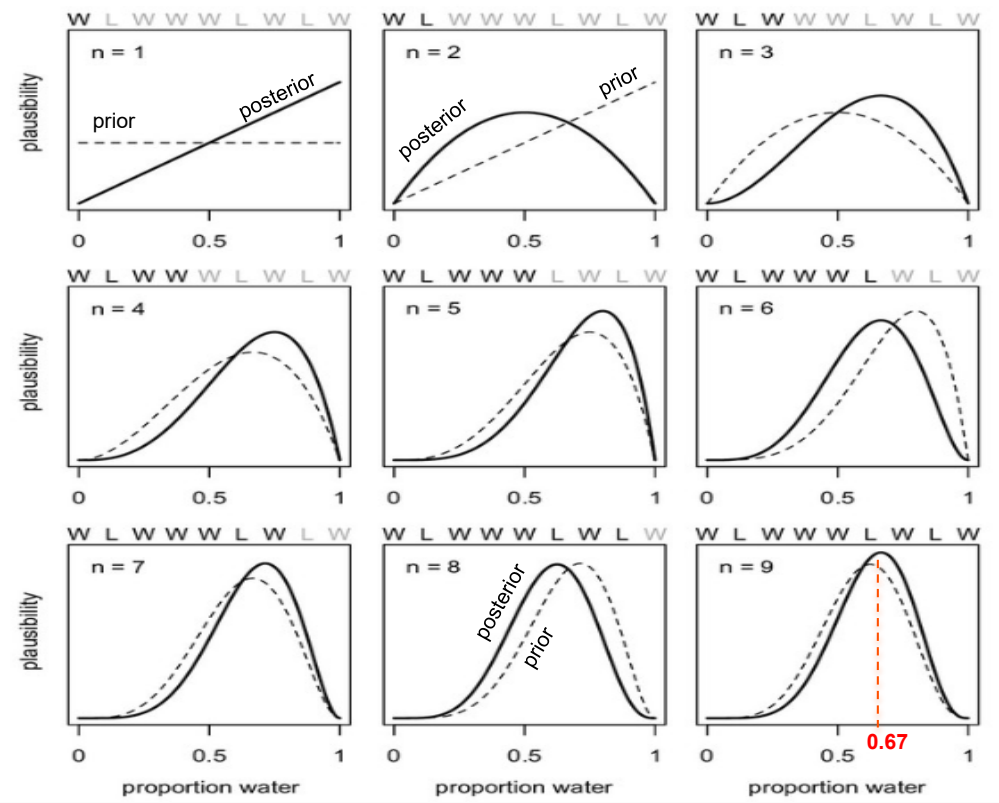
$P(p)$ → $P(X_1|p)$ → $P(p|X_1)$



X_1 = *W*



Round 2, 3, …



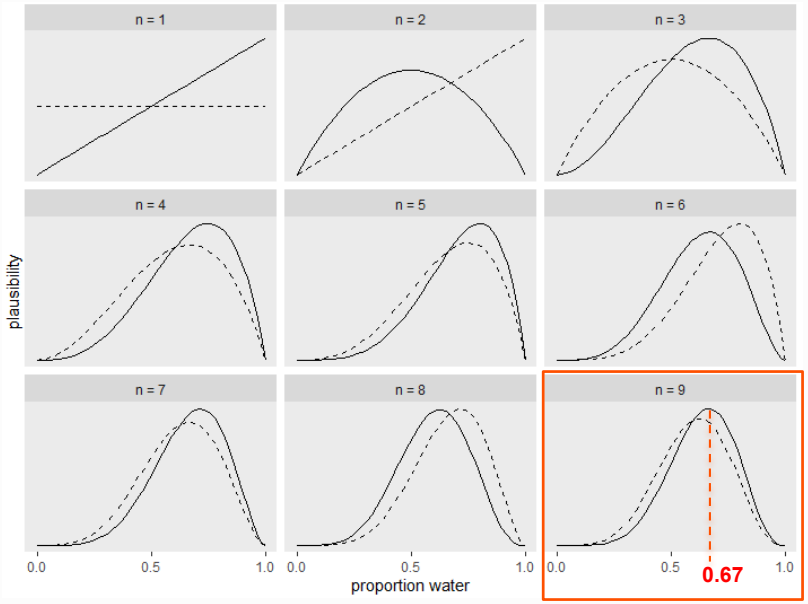
N	Posterior mode
1:	$p = 1$
2:	$p = 0.49$
3:	$p = 0.67$
..	
9:	$p = 0.67$

Truth:
Approx 71% of earth is water

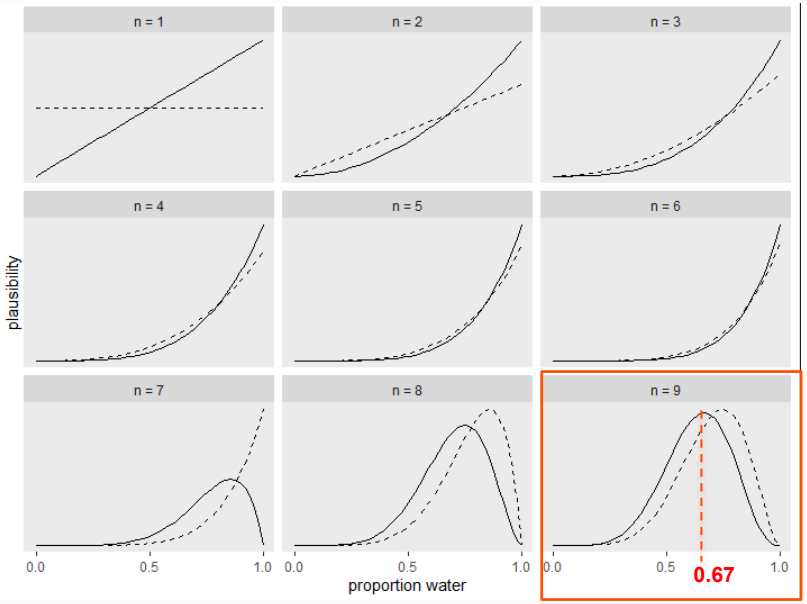
Does the order of tossing matter?

- Given the same framework and data:

Sequence *wlwwwlwlw* gave rise to:



Another sequence of data, wwwwwwlll give rise to:



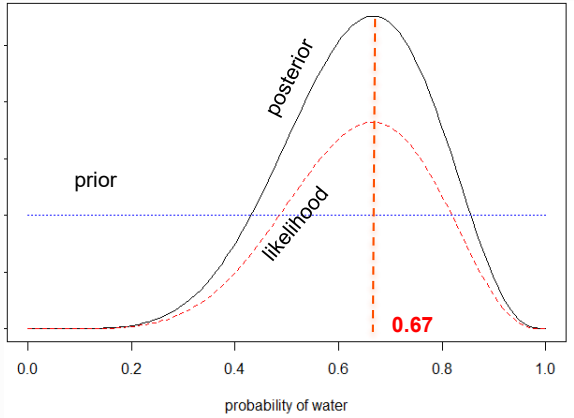
Can we do everything once when n=9?

- Do we need to update after every toss?
- Given the same framework and data:
  - Any value of p equally plausible
  - Observe 6 W and 3 L
  - Some values of p are more plausible

Prior $P(p)$

Likelihood $P(6\ W\ and\ 3\ L|p)$

Posterior $P(p|6\ W\ and\ 3\ L)$



Why does the posterior become the same in the end?

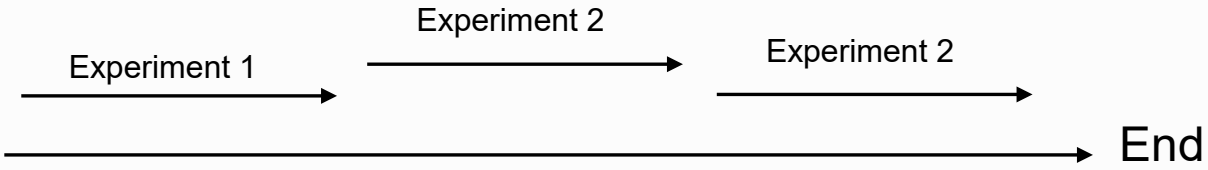
- Different order of tossing or observed 6W and 3 L at once gave same posterior distribution – Why?
- Posterior distribution depends only on data through $f(y|\theta)$.
- Likelihood principle
  - Identical likelihoods should give identical inference
  - *How* likelihood was constructed (what order of sequence) does not matter
    - Data 1: Do two experiments, get: (4, 3) and (5, 6, 1)
    - Data 2: Do one experiment, get: (4, 3, 5, 6, 1)
    - Data 3: Do one experiment, get: (4, **6**, 5, **3**, 1)

Frequentist statistics violates the likelihood principle

- Recall 1:
  - Sampling distribution assess the plausibility of different outcomes when repeating the same experiment under identical conditions, e.g. under H0: $\theta = 0$
- Recall 2:
  - statements about θ (e.g.hypothesis testing) depends on
    - sampling distribution
      - In turn depends on the experimental design
        - E.g. p-value depends on the sample size.
- Frequentist methods violates the likelihood principle
  - *How* likelihood was constructed (what order of sequence) does matter

Sampling distribution and p-values

- Each hypothesis test has it´s own sampling distribution under H0
  - Fix type I error (incorrectly reject a true H0) for each test
  - Can´t combine p-values (in an intuitive way)
  - Repeated hypothesis testing increase familywise type I error (incorrectly reject at least one true H0)
  - Handling repeated hypothesis testing is complex and unintuitive
  - Frequentist can not flexibly learn from accumulated data
    - Must handle each experiment separately



Stopping rules and p-values

- The p-value:
  - Behaviour under repeated sampling under H0 where the sampling is stopped at size N.
- Another rule:
  - Behaviour under repeated sampling under H0 where the sampling is stopped at time point T (different samples can have different samples).
- ACTUAL OUTCOME (data) is the same but we compared with different things because we have a different stopping rule.
- Bayesians can flexibly learn from data as it accumulates
  - The way the likelihood (and prior) are constructed doesn´t matter

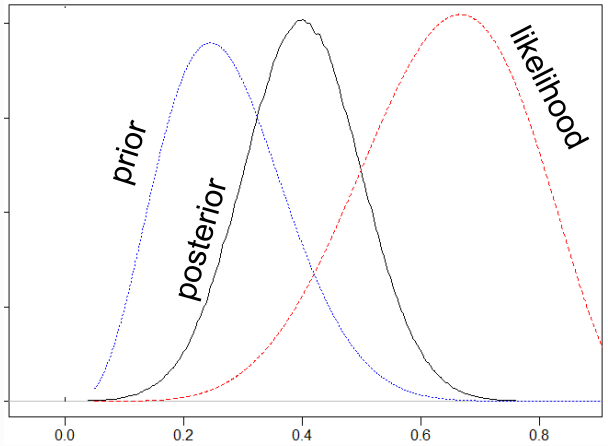
Example of violation of likelihood principle

- Test of coin is fair: H0: θ = 0 vs H1: θ ≠ 0

- Experiment 1: Flip coin until *Heads* turn up.
  - Result: 6 flips
  - P-value = $0.5 * (1 - 0.5)^5 + 0.5 * (1 - 0.5)^6 + \ ... \ = 0.03125$
- Experiment 2: Flip coin 6 times.
  - Result: One *Heads* turn up
  - P-value = $\binom{6}{1} * 0.5 * (1 - 0.5)^5 + \binom{6}{0} * (1 - 0.5)^6 = 0.1093$

My belief about proportion of water on earth*

- Lets say I´ve read too many science fiction books about other planets where 30% of surface is water (*p = 0.3).*
  - Hence, I believe earth has 10-50% of surface under water
  - I observe 6 W and 3 L



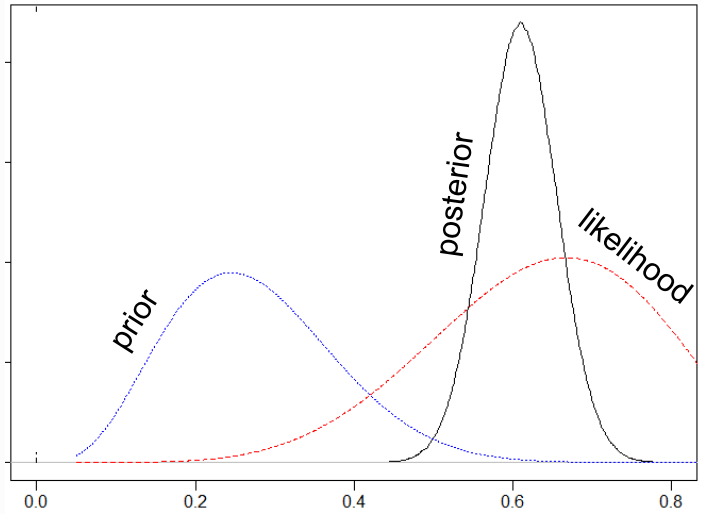
Likelihood maximum at p=0.67
Posterior distribution max at p=0.40

Frequentist and bayesian results differ due to strong prior belief

* Inspired by https://staff.math.su.se/hoehle/blog/2017/06/22/interpretcis.html

My belief about proportion of water on earth

- Let´s say I continue to toss the globe a total of 100 times.
  - I observe 66 W and 34 L



Likelihood maximum at p=0.67
Posterior distribution max at p=0.60

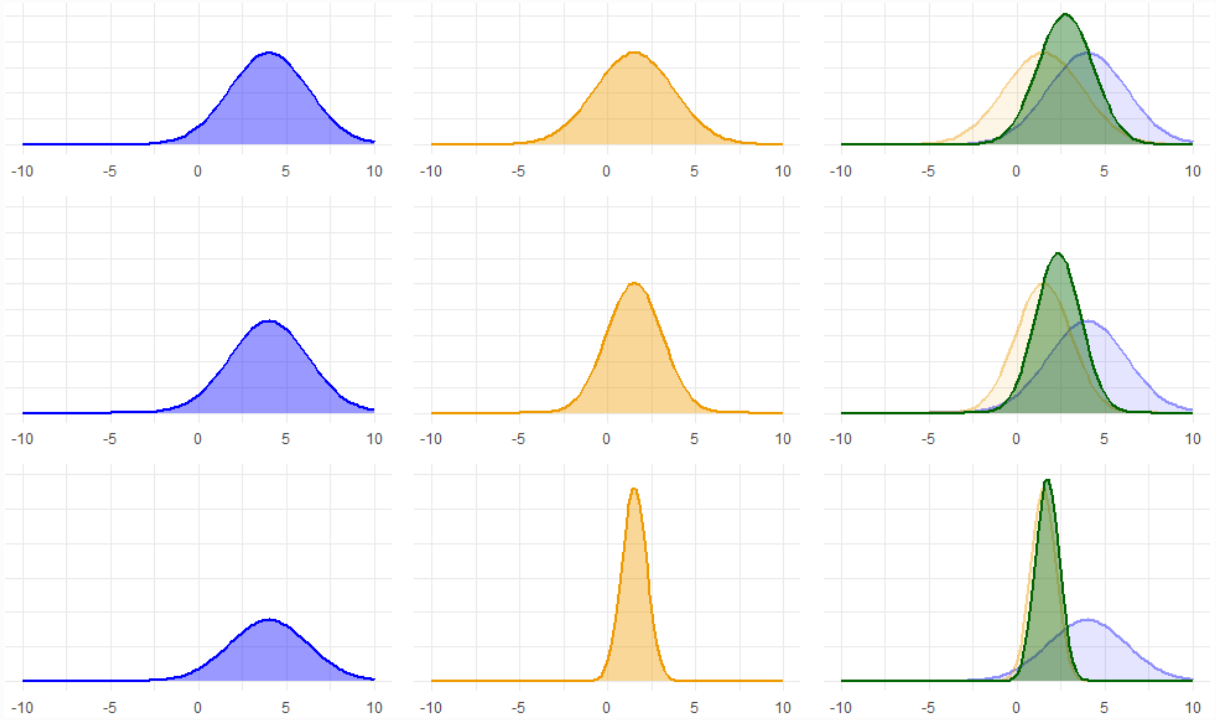
Frequentist and bayesian results differ due to strong prior belief but much data (=much evidence) makes strong prior belief influence less

Number of observations will influence the posterior

N=30

N=60

N=300



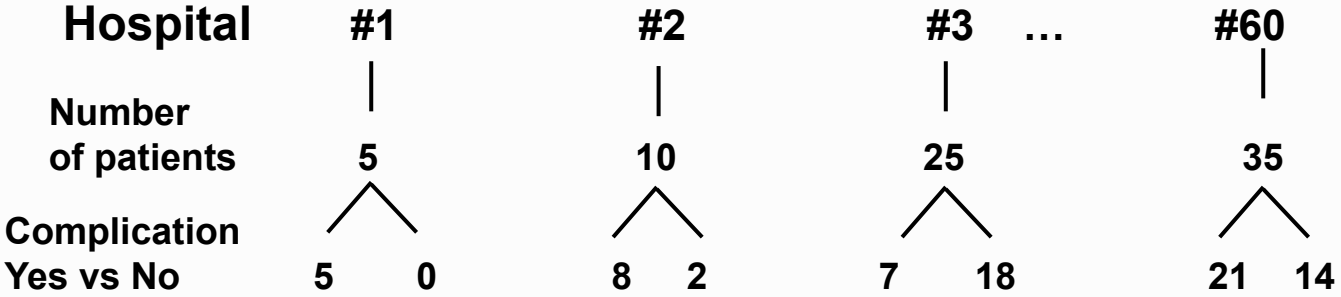
https://github.com/mattansb/bayesian-evidence-iscop-2021/blob/main/bayesian-evidence-iscop-2021.Rmd

Summary

- Likelihood principle
  - Bayesian updating (learning from data)
- Frequentist inference violates the likelihood principle
  - Depends on the experiment and how/why data was collected (sampling intention)

Hospital example*

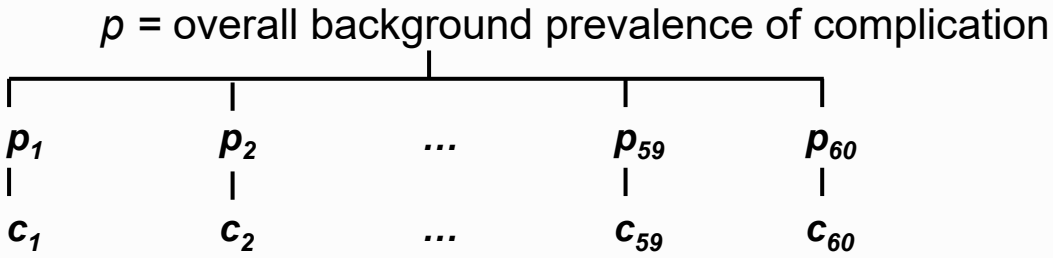
- **Simulated data**: 60 hospitals. Each hospital has a varying number of patients undergoing a certain type of surgery. Some of these patients experience a serious post-operative complication
- Estimate the prevalence of post-operative complication
  - Per hospital and overall



* Example taken from (but converted to a clinical setting) McElreath, R: Statistical Rethinking (2020), see https://xcelab.net/rm/statistical-rethinking/
See also https://bookdown.org/content/3890/

Hospital example

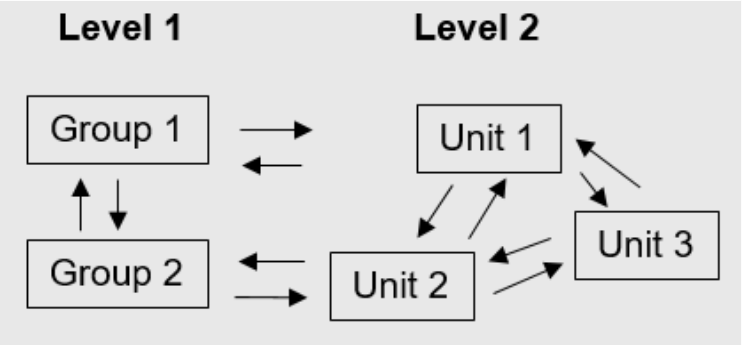
- Number of complications (*C*) in each hospital has a binomial distribution
  - $C_i \sim Bin(p_i, N_i), i = 1, 2 \ldots, 60$



- Differences in *p* across hospitals (clustering)
- Arise from a common underlying prevalence *p*

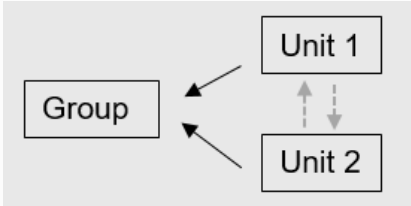
Multilevel models

- Accounts for relationships between units and groups
- Accounts for how much information available in the units and groups
- Dynamically combine information across and within levels
  - Effects within/between levels are random and correlated



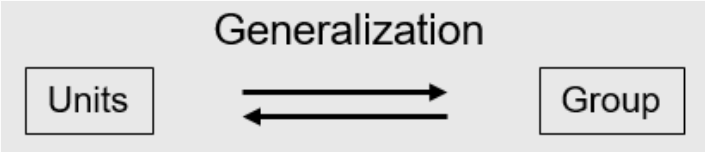
Multilevel models

- Simplification of models may cause misspecifications
  - By ignoring the dependency structure in the data we incorrectly believe data contain more information than it does

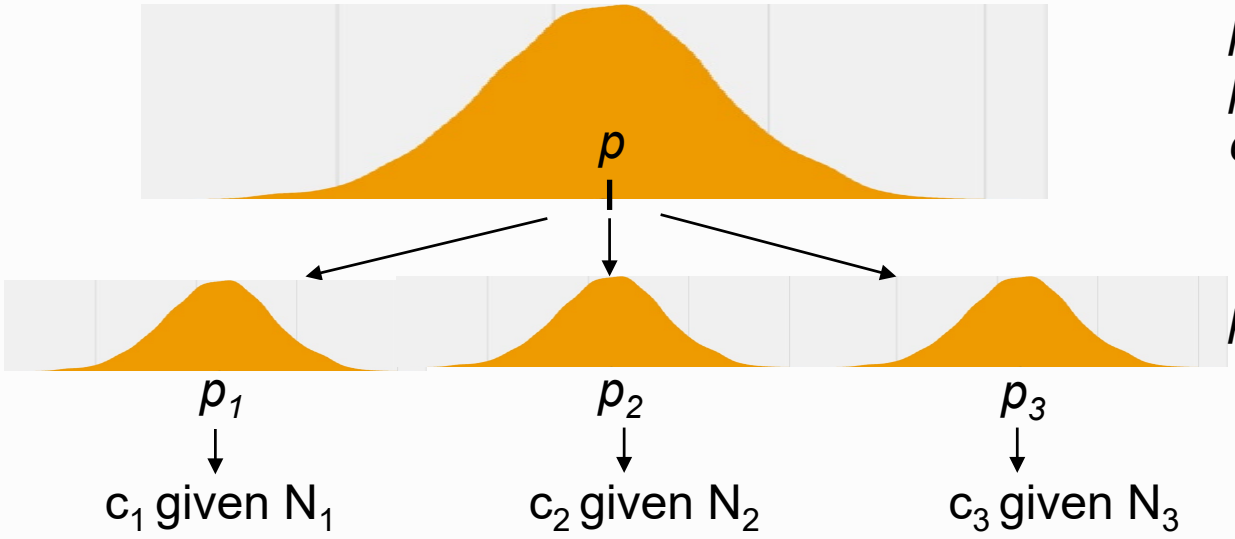


If units are very similar (correlated) each additional unit will add less extra information compared to when uncorrelated

  - Give rise to atomistic and ecological fallacies
    - Conclusions on units and group levels are confused and distorted



Multilevel models

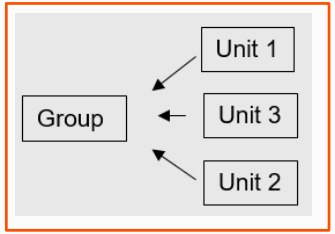


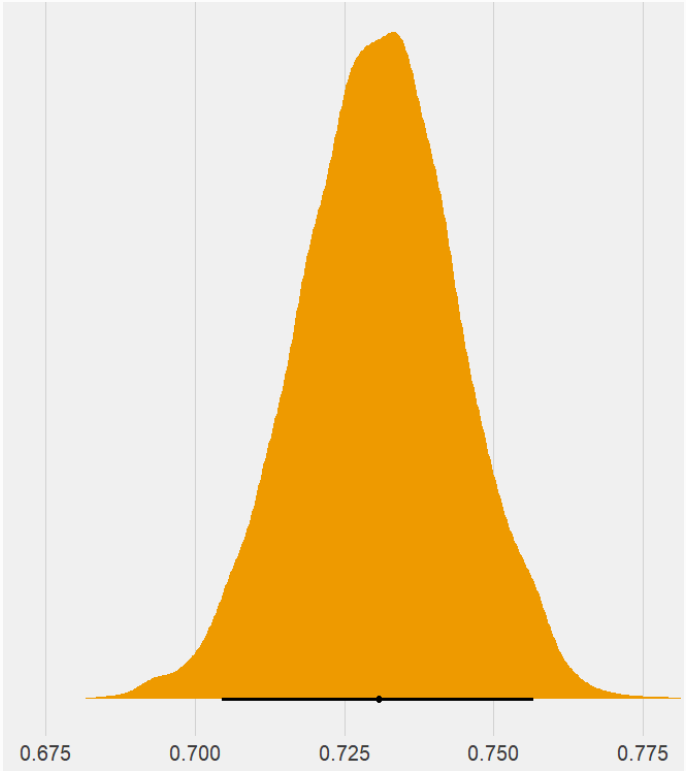
p, σ = ”hyperparameter”
$p \sim N(0, 1)$
$\sigma \sim Exp(1)$

$p_i \sim N(p, \sigma)$

Ignore multi-level structure (complete pooling)

- $C_i \sim Bin(p_i, N_i), i = 1, 2 \ldots, 60$
- Logit(p_i) = a
- A single prior for all hospitals:
- $a \sim t(3, 0, 2.5)$
- Overall prevalence *
  - p = 0.73
  - (95% Credible interval: 0.70; 0.76)

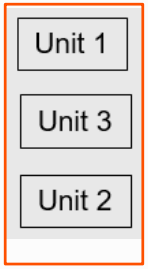




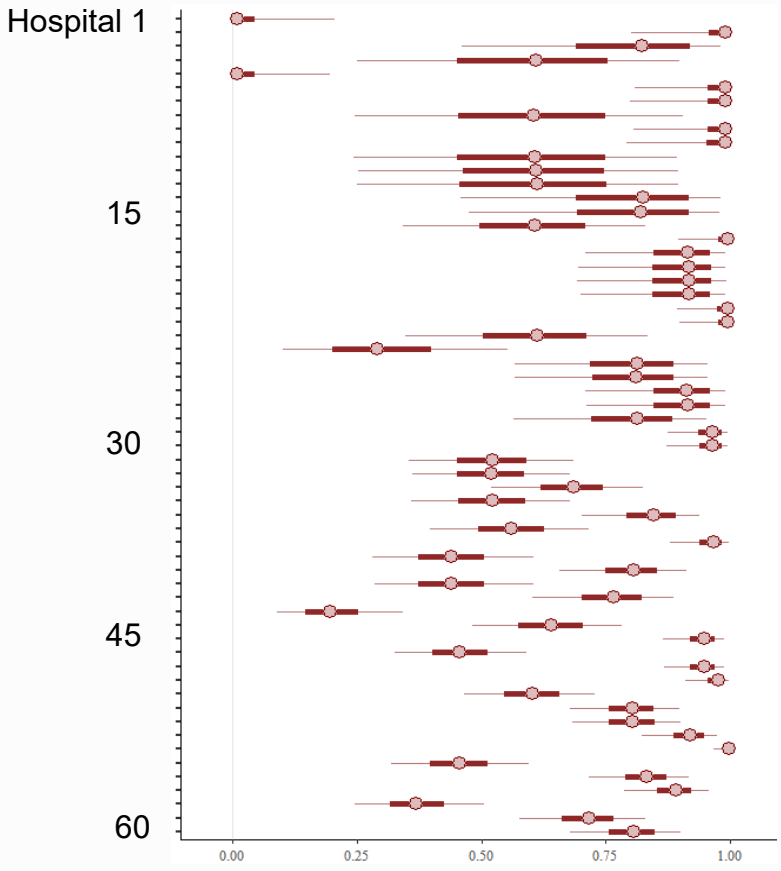
* Same results given by:
- Frequentist logistic regression with single intercept
- $\frac{1}{60}\sum_{i=1}^{60}\frac{C_i}{N_i}$

Ignore multi-level structure (no pooling at all)

- Estimate prevalence for each hospital
- $C_i \sim Bin(p_i, N_i), i = 1, 2 \ldots, 60$
- Logit(p_i) = b_i
- One prior for each hospital
- $b_i \sim N(0, 5), i = 1, 2 \ldots, 60$

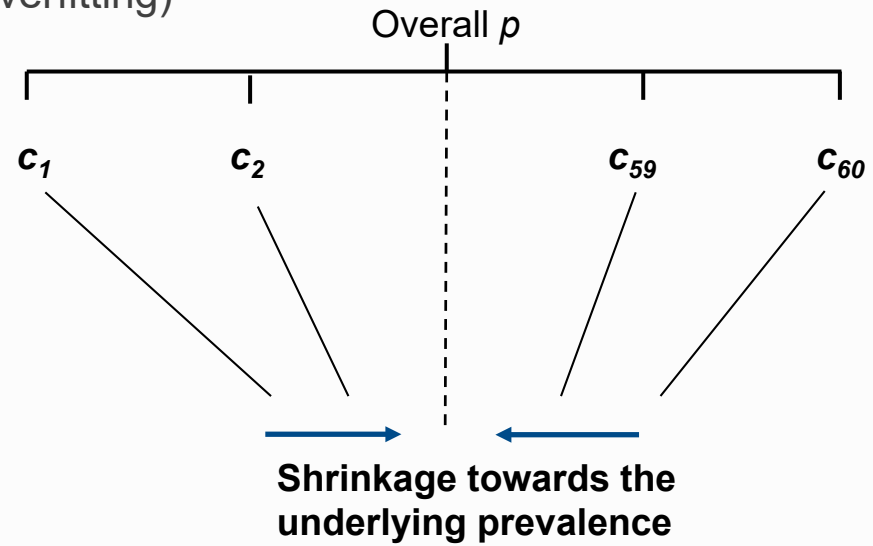


Median posterior estimates & credible intervals

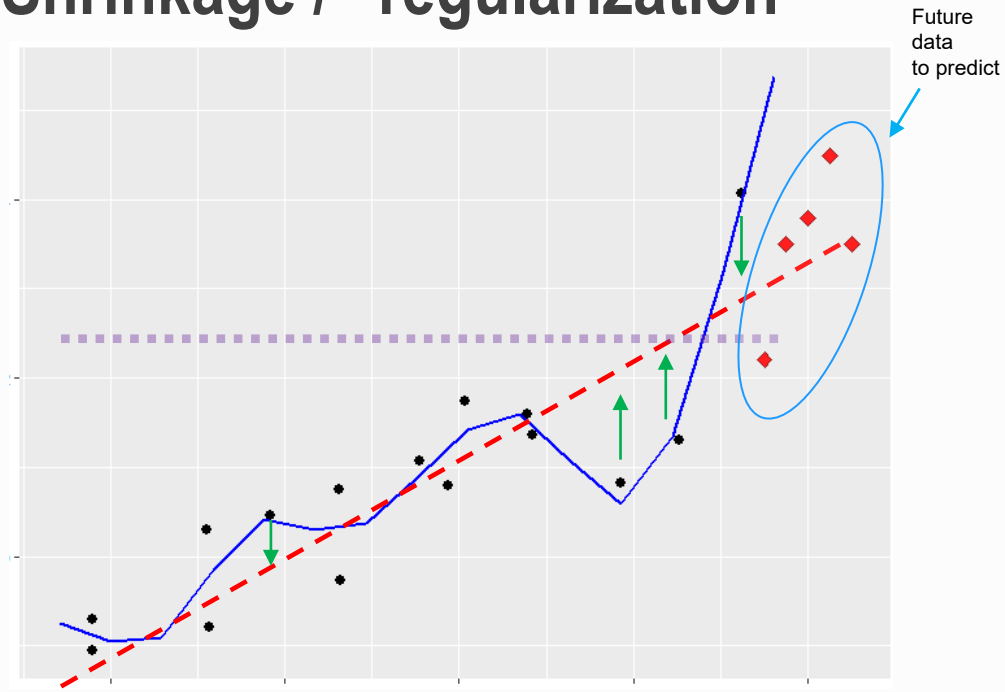


Account for multi-level structure (”shrinkage”)

- Shrinkage
  - Dampen the effect of between-hospital variability
    - extreme values less influence
  - Improve prediction (avoid overfitting)
  - ”Empirical bayes”



Shrinkage / ”regularization”



- Shrinkage
  - ”regularization”
- Parsimonous model
- Enhanced external validity

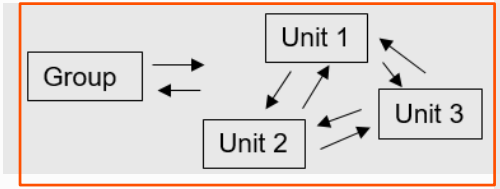
Too much flexibility

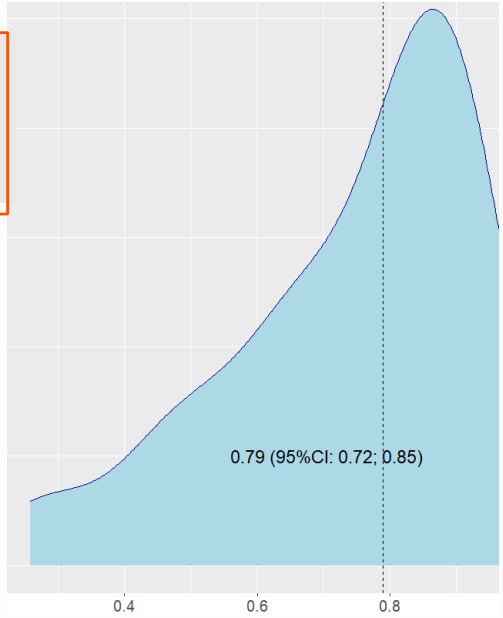
Too little flexibility

Reasonable flexibility

Account for multi-level structure (partial pooling): Frequentist

- Random intercept logistic regression
  - $C_i \sim Bin(p_i, N_i), i = 1, 2 \ldots, 60$
  - Logit(p_i) $= a + b_i$
  - $b_i \sim$ N$(0, \sigma)$
- Overall prevalence (fixed effect estimate)
  - p = 0.79
  - (95% Confidence interval: 0.72; 0.85)





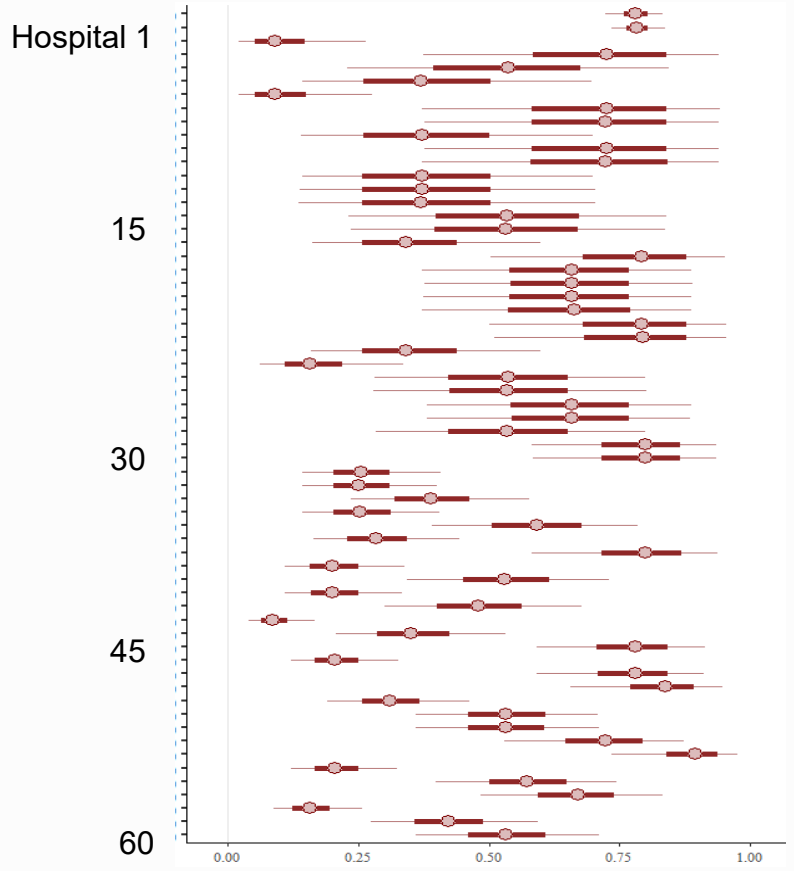
Account for multi-level structure (partial pooling): Bayesian

- $C_i \sim Bin(p_i, N_i), i = 1, 2 \ldots, 60$
- Logit(p_i) = b_i
- $b_i \sim$ N(b, σ)
- $b \sim$ N(0, 1)
- $\sigma \sim$ HalfCauchy(0, 1)

b is a hyperparameter. Parameter for the average hospital prevalence

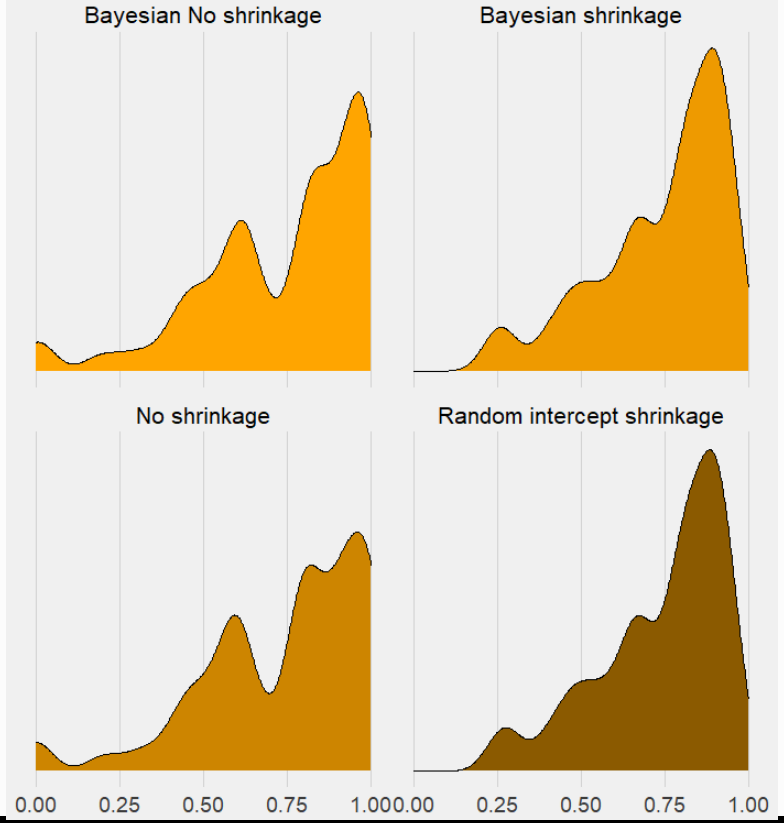
- Overall prevalence
  - p = 0.78
  - (95% Credible interval: 0.71; 0.84)

Median posterior estimates & credibility intervals



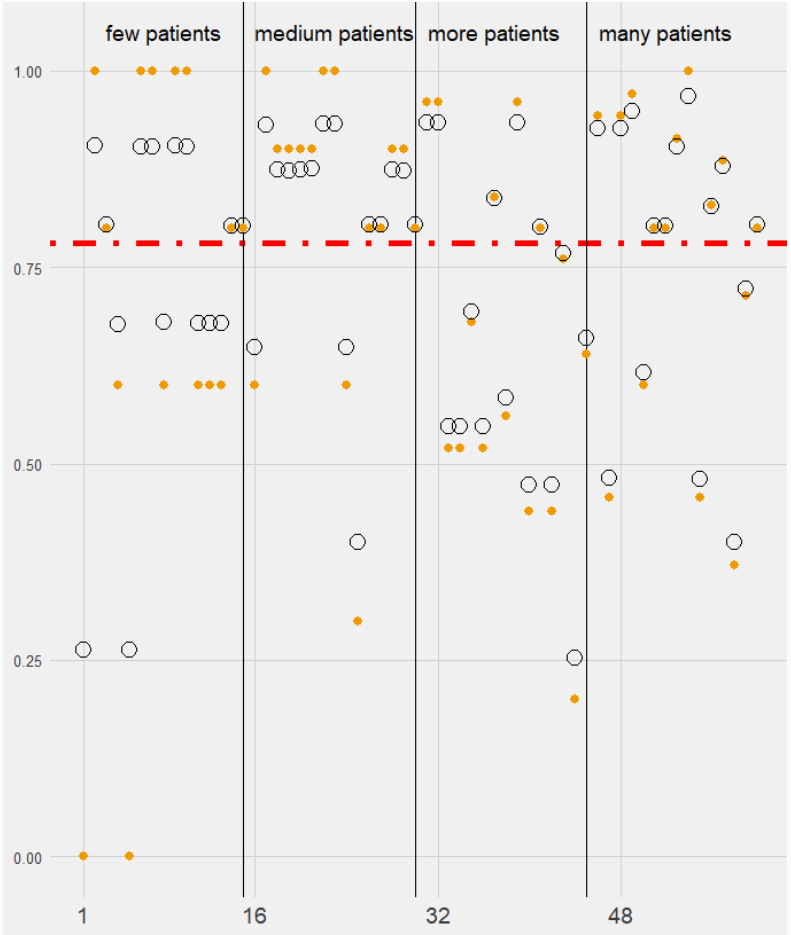
Shrinkage vs No shrinkage

Distribution of estimated prevalence across hospitals



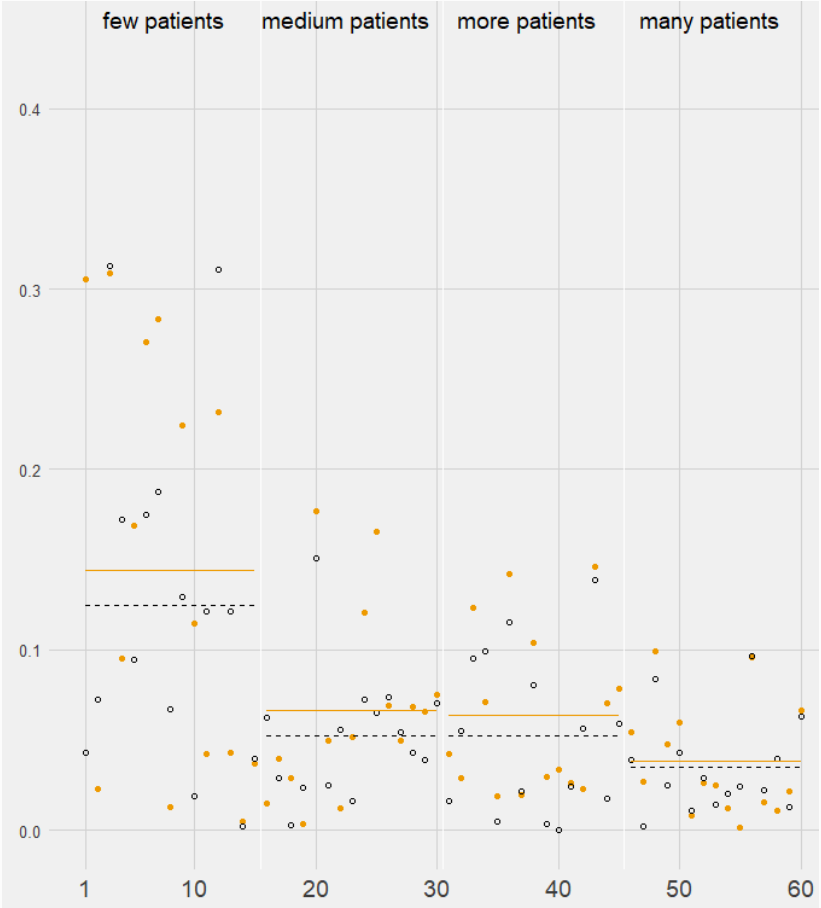
Shrinkage vs No shrinkage

- The empirical proportions are in orange
- Modelled proportions are the black circles.
- Dashed red line is the model-implied average proportion (0.78).
- The fewer patients per hospital the more pronounced is the shrinkage



Shrinkage vs No shrinkage

- Absolute error (estimated *p* – true *p)*
- No-pooling shown in orange.
- Partial pooling shown in black
- Lines show the average error



Consequences of ignoring multi-level structure

- By ignoring the dependency structure in the data we incorrectly believe data contain more information than it does
- Standard errors underestimated
  - Credible intervals and confidence intervals too narrow
  - P-values too small (Type I errors)

	Estimated overall prevalence			
Model	**Estimate**	**95% Credible interval**	**Width of interval**	**bias (true rate=0.80)**
Complete pooling	0.73	0.70; 0.76	0.06	-0.07
Frequentist multilevel	0.79	0.72; 0.85 *	0.13	-0.01
Bayesian multilevel	0.78	0.71; 0.84	0.13	-0.02

* Confidence interval

- What about bias?

Summary

- Multi-level/hierarchical modelling reflect correct amount on information available in hierarchical correlated data
- Ignoring multi-level structure will falsely claim we have more information than we actually have
- Shrinkage to the population favor more realistic effect estimates