BAYESIAN DATA ANALYSIS

Two examples

Bayesian stat workshop 24th Aug 2020

DAVID BOCK,
BIOSTATISTICS, SCHOOL OF PUBLIC HEALTH AND COMMUNITY MEDICINE, INSTITUTE OF MEDICINE
SCANDINAVIAN SURGICAL OUTCOMES RESEARCH GROUP (SSORG)

Aim

- Give a reminder of basic concepts of bayesian thinking
- Illustrate two simple situations where bayesian strategies are reasonable
  - Sequential learning
  - Clustered data

- Prerequisites: Basic knowledge of statistics

Bayesian vs frequentist approaches (simplified view)

Bayesian

Probability

- Reflecting degree of belief

Inference

- Use Bayes formula

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

Frequentist

- Explaining behaviour in hypothetical repetitions under the same conditions

- Use methods that works well “in the long run”

Inference based on bayes formula

- Population parameter of interest: θ
- Prior probability reflecting degree of belief about θ : $P(\theta)$
- Data (likelihood): $P(X|\theta)$
- Posterior distribution: $P(\theta|X) = \frac{P(\theta)P(X|\theta)}{P(X)}$
- Posterior ∝ Prior * data
- Inference takes the form of updating priors to yield posteriors

Belief + data ⟶ New belief

Learning from data

- Inference takes the form of updating priors to yield posteriors
- Bayesian updating
- Old posterior becomes our new prior $P(\theta|X_1)$
- New data X_2
- New posterior $P(\theta|X_1, X_2) = \dfrac{P(\theta|X_1)P(X_2|\theta)}{P(X_2)}$

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

Prior $P(p)$

Likelihood $P(X_1|p)$

Posterior $P(p|X_1)$

- Round 2

Prior $P(p|X_1)$

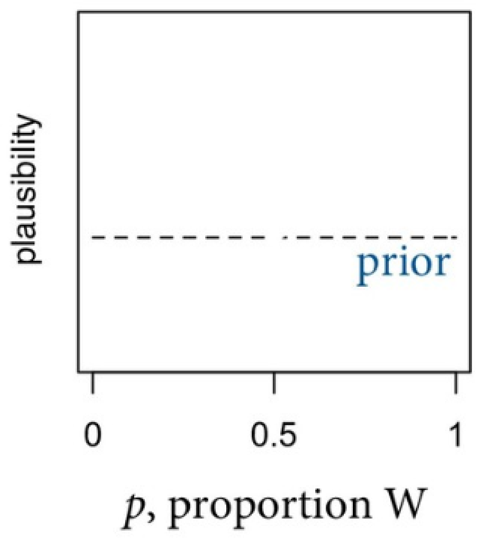
Likelihood $P(X_2|p)$

Posterior $P(p|X_1, X_2)$

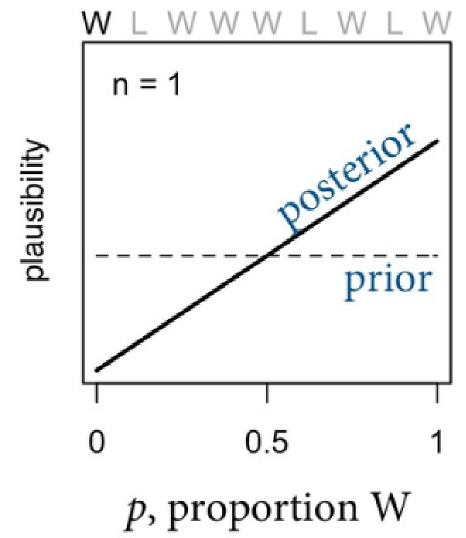
- Etc, etc …

Round 1

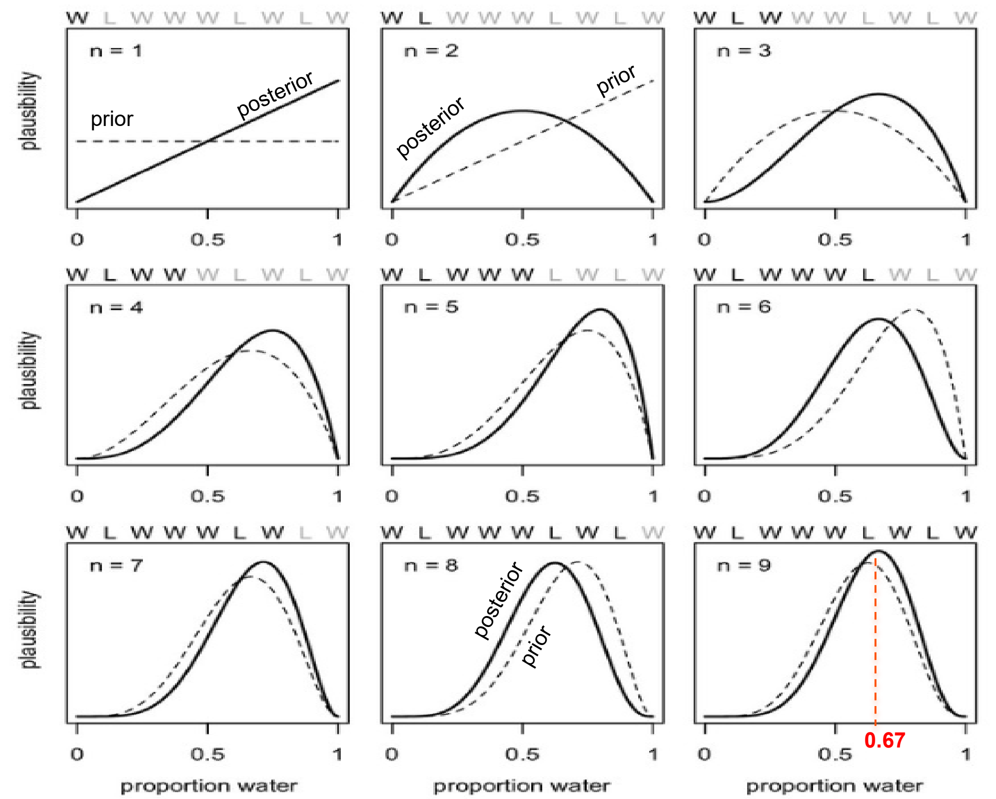
$P(p)$ → $P(X_1|p)$ → $P(p|X_1)$



$X_1 = W$



Round 2, 3, …



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

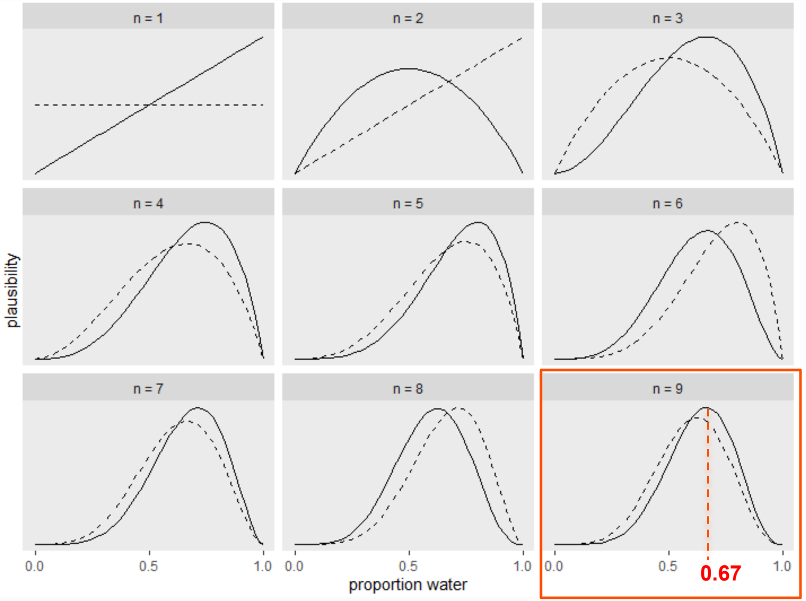
Truth:

Approx 71% of earth is water

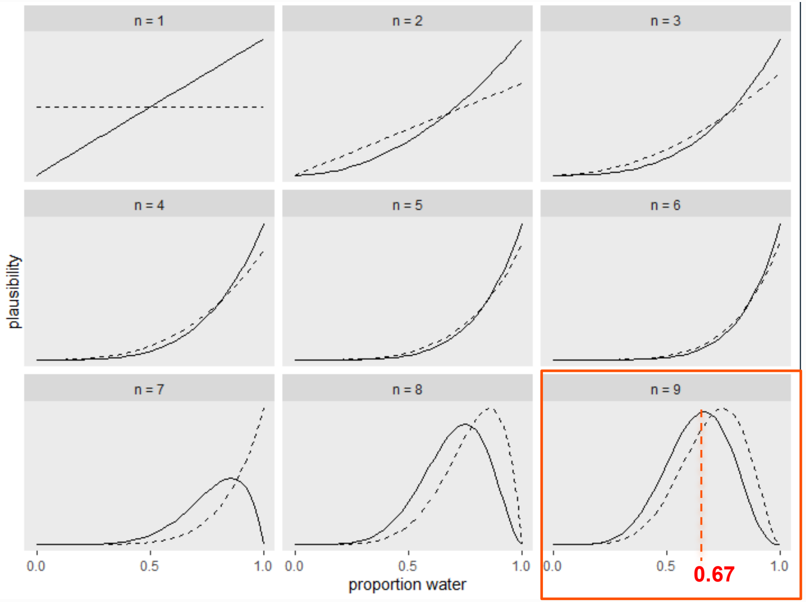
Does the order of tossing matter?

- Given the same framework and data:

Sequence *wlwwwlwlw* gave rise to:



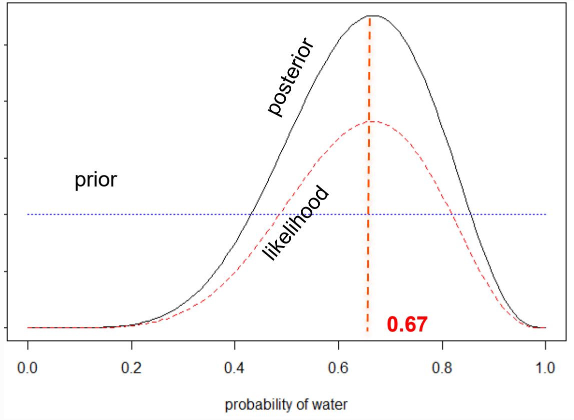
Another sequence of data, wwwwwwlll give rise to:



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

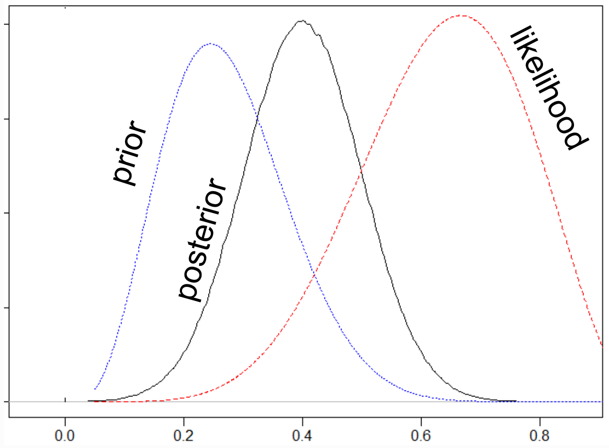
Prior	$P(p)$
Likelihood	$P(6\ W\ and\ 3\ L \mid p)$
Posterior	$P(p \mid 6\ W\ and\ 3\ L)$



The value of p that maximize posterior distribution $P(p|X)$ will be the same as the value that maximize the likelihood function $P(6\ W\ and\ 3\ L|p)$

My belief about proportion of water on earth*

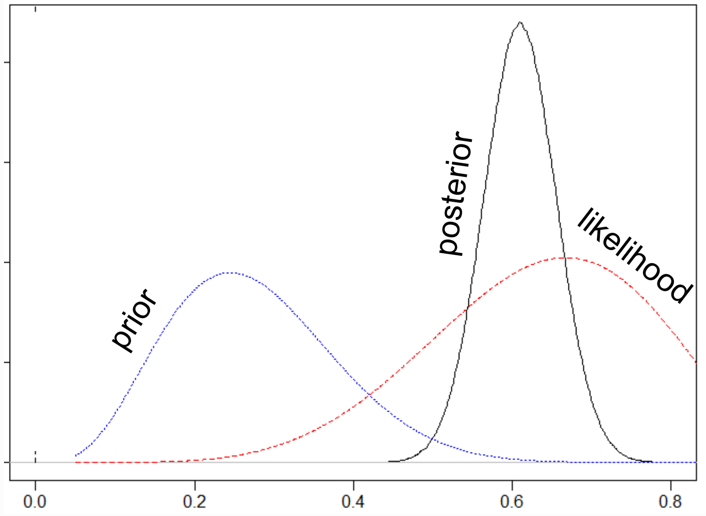
- Lets say I´ve read too many science fiction books about other planets where 30% of surface is water (*p = 0.3*).
  - Hence, I believe earth has 10-50% of surface under water
  - I observe 6 W and 3 L



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

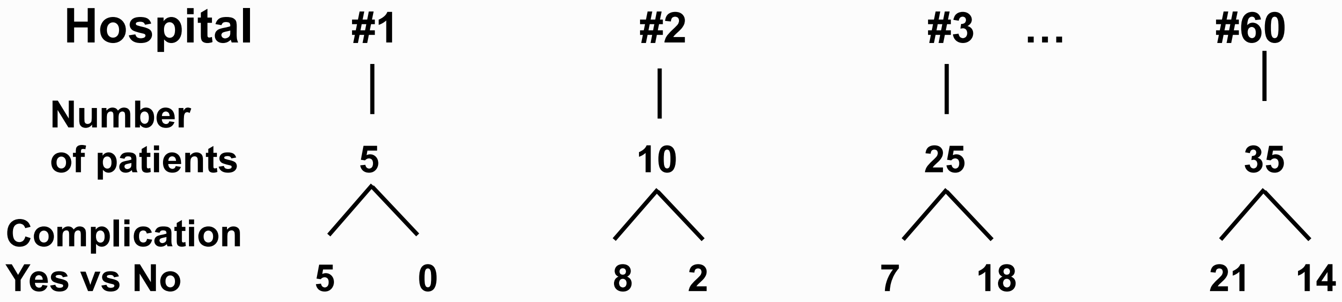
My belief about proportion of water on earth

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



Hospital example*

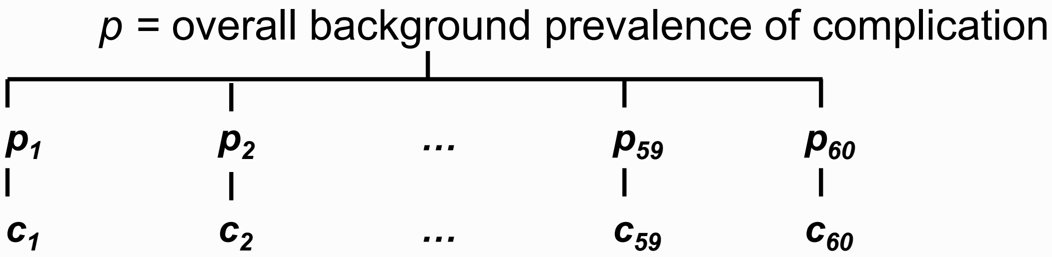
- **Simulated data**: 60 hospitals. Each hospital has a varying number of patients undergoing a certain type of surgery. Some of these patients experience a serious post-operative complication
- Estimate the prevalence of post-operative complication
  - Per hospital and overall



* 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

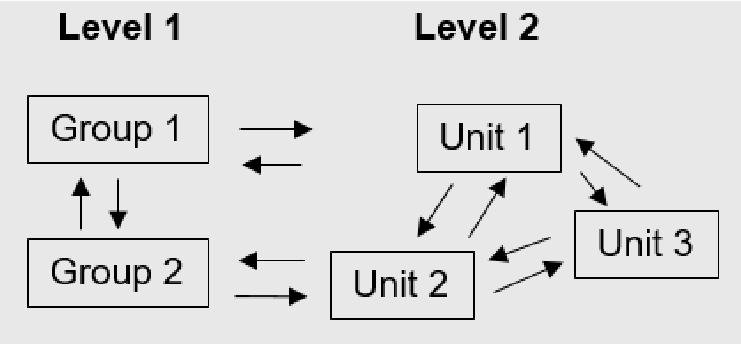
- Number of complications (*C*) in each hospital has a binomial distribution
  - $C_i \sim Bin(p_i, N_i), i = 1, 2 \dots, 60$



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

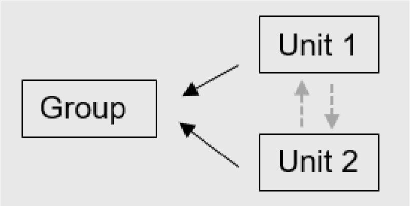
Multilevel models

- 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



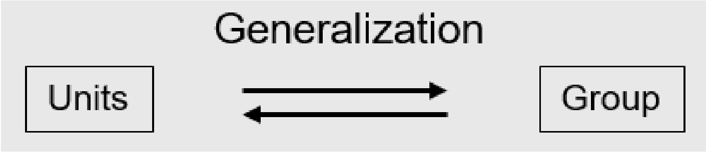
Multilevel models

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

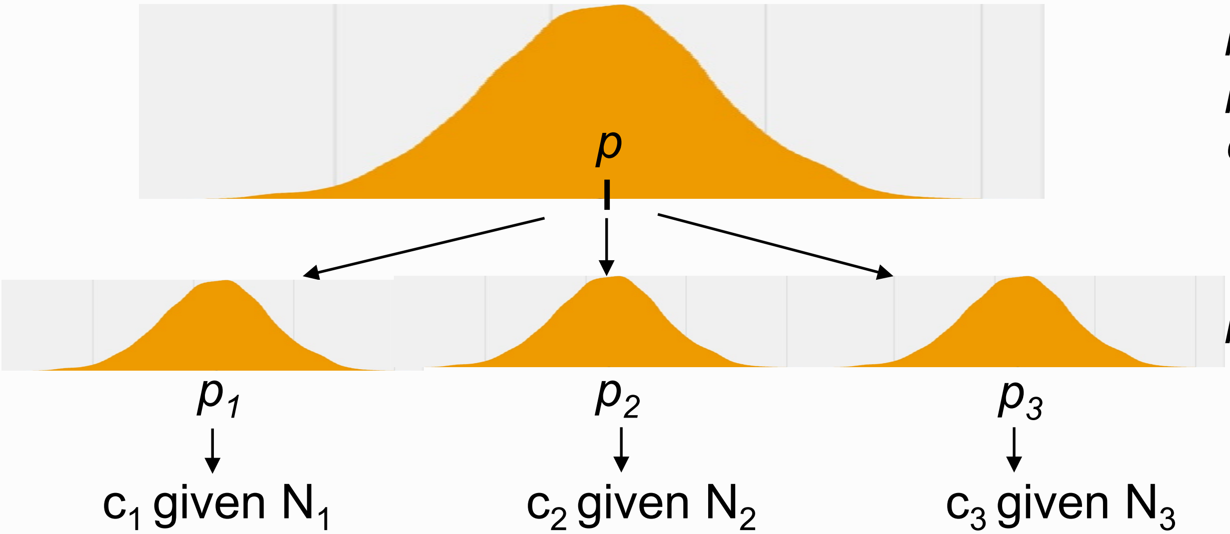


If units are very similar (correlated) each additional unit will add less extra information compared to when uncorrelated

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



Multilevel models



p, σ = ”hyperparameter”

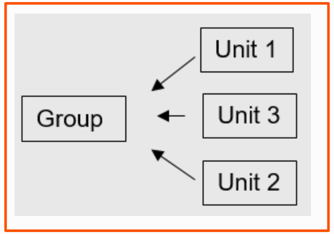
$p \sim N(0, 1)$

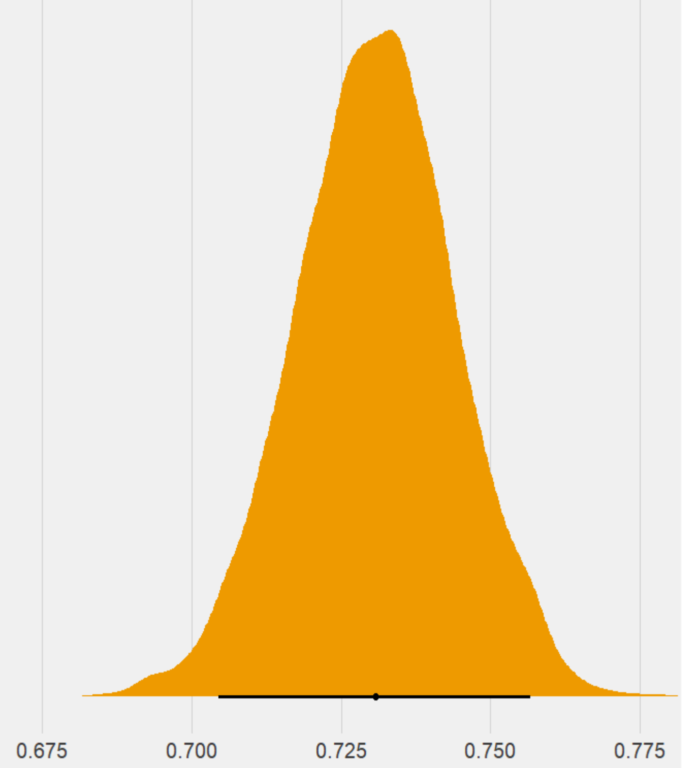
$\sigma \sim Exp(1)$

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

Ignore multi-level structure (complete pooling)

- $C_i \sim Bin(p_i, N_i), i = 1, 2 \dots, 60$
- 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)

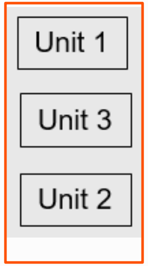




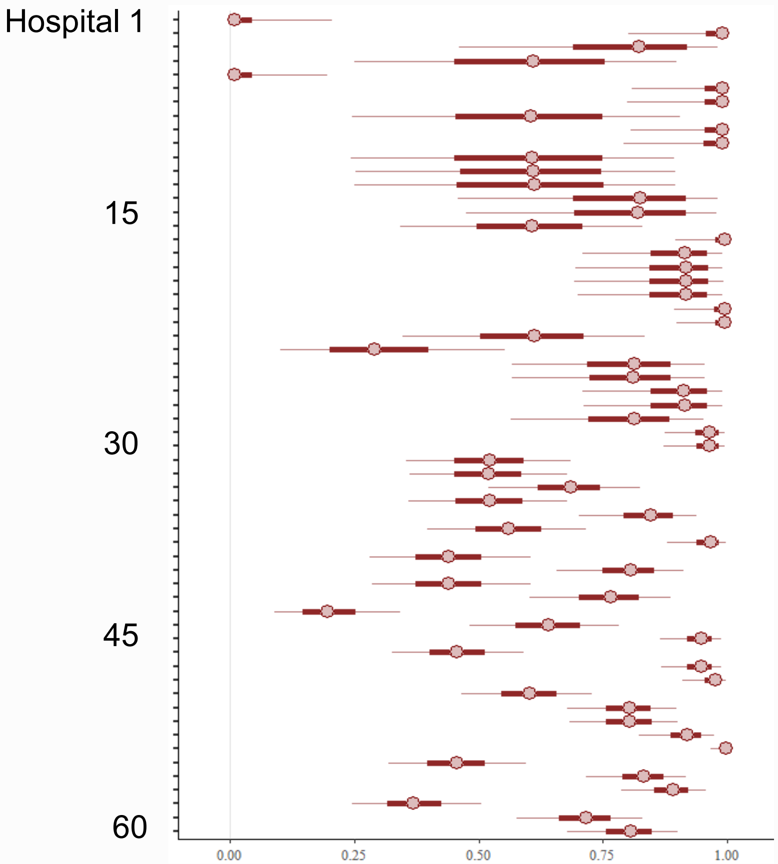
* Same results given by:
- Frequentist logistic regression with single intercept
- $\frac{1}{60}\sum_{i=1}^{60} \frac{C_i}{N_i}$

Ignore multi-level structure (no pooling at all)

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

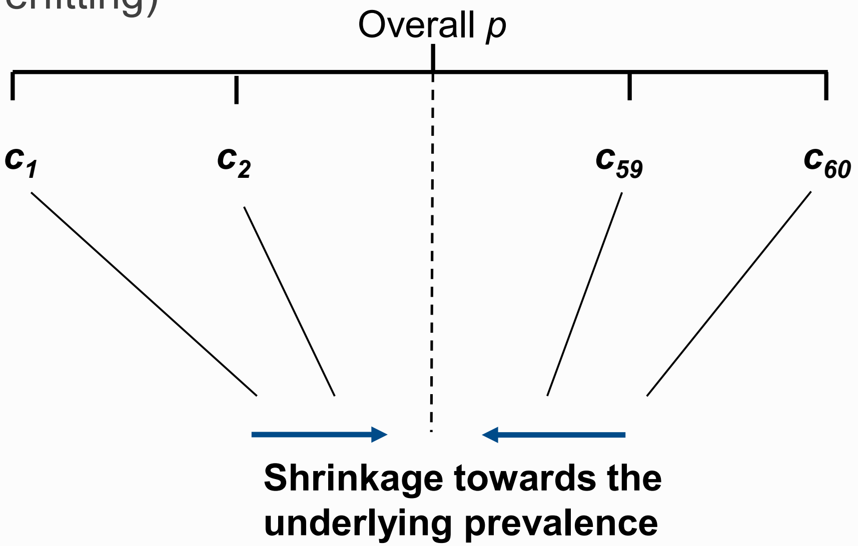


Median posterior estimates & credible intervals

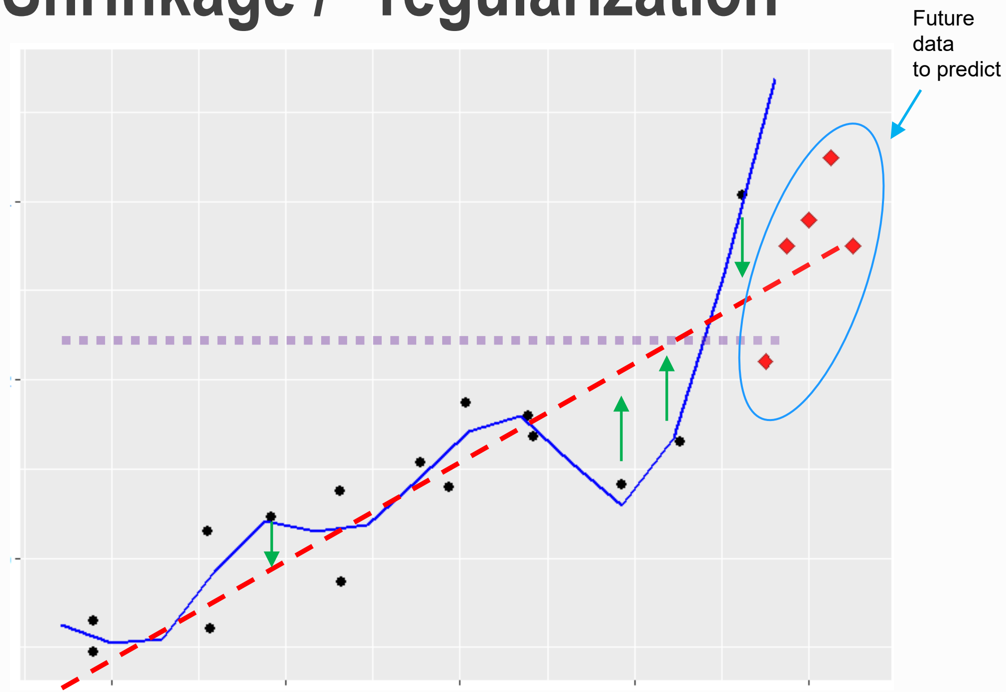


Account for multi-level structure (”shrinkage”)

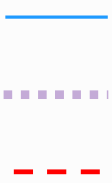
- Shrinkage
  - Dampen the effect of between-hospital variability
    - extreme values less influence
  - Improve prediction (avoid overfitting)
  - ”Empirical bayes”



Shrinkage / ”regularization”

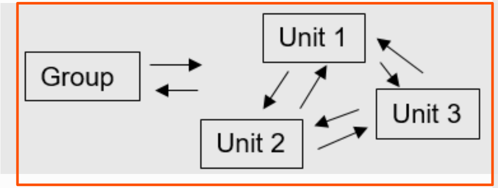


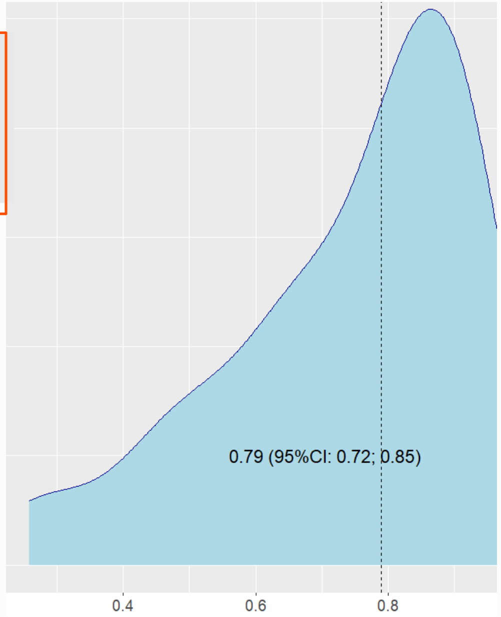
- Shrinkage
  - ”regularization”
- Parsimonous model
- Enhanced external validity



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

- Random intercept logistic regression
  - $C_i \sim Bin(p_i, N_i), i = 1, 2 \dots, 60$
  - Logit(p_i) $= a + b_i$
  - $b_i \sim$ N(0, σ)





- Overall prevalence (fixed effect estimate)
  - p = 0.79
  - (95% Confidence interval: 0.72; 0.85)

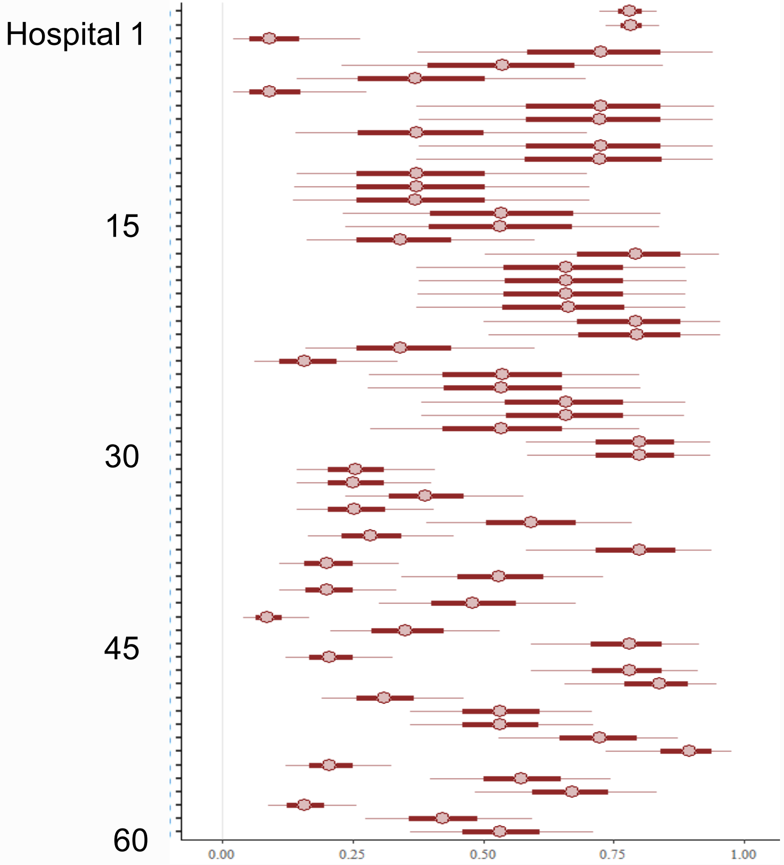
Account for multi-level structure (partial pooling): Bayesian

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

b is a hyperparameter. Parameter for the average hospital prevalence

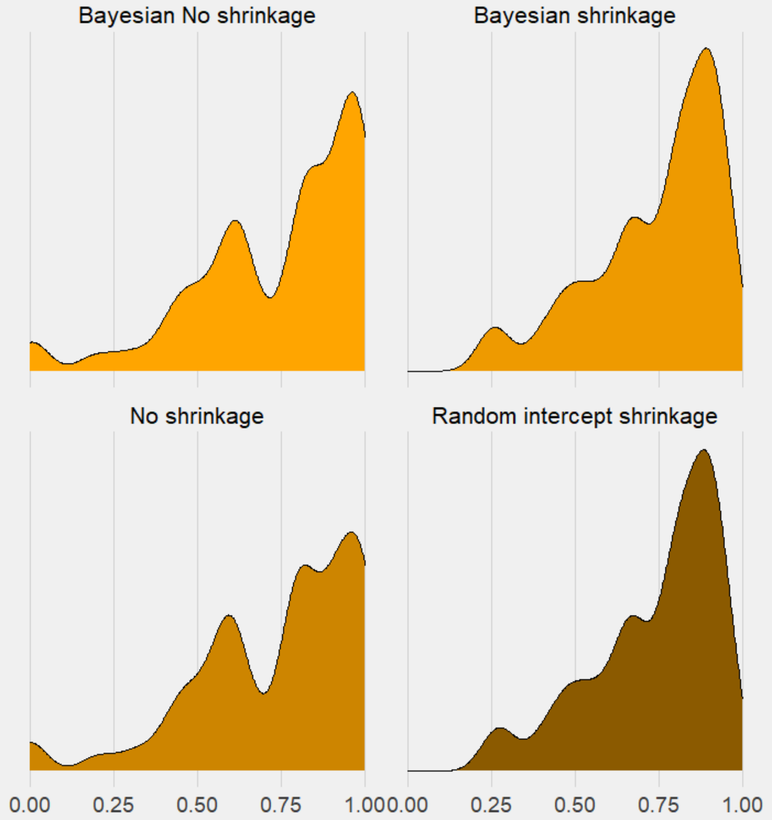
- Overall prevalence
  - p = 0.78
  - (95% Credible interval: 0.71; 0.84)

Median posterior estimates & credibility intervals



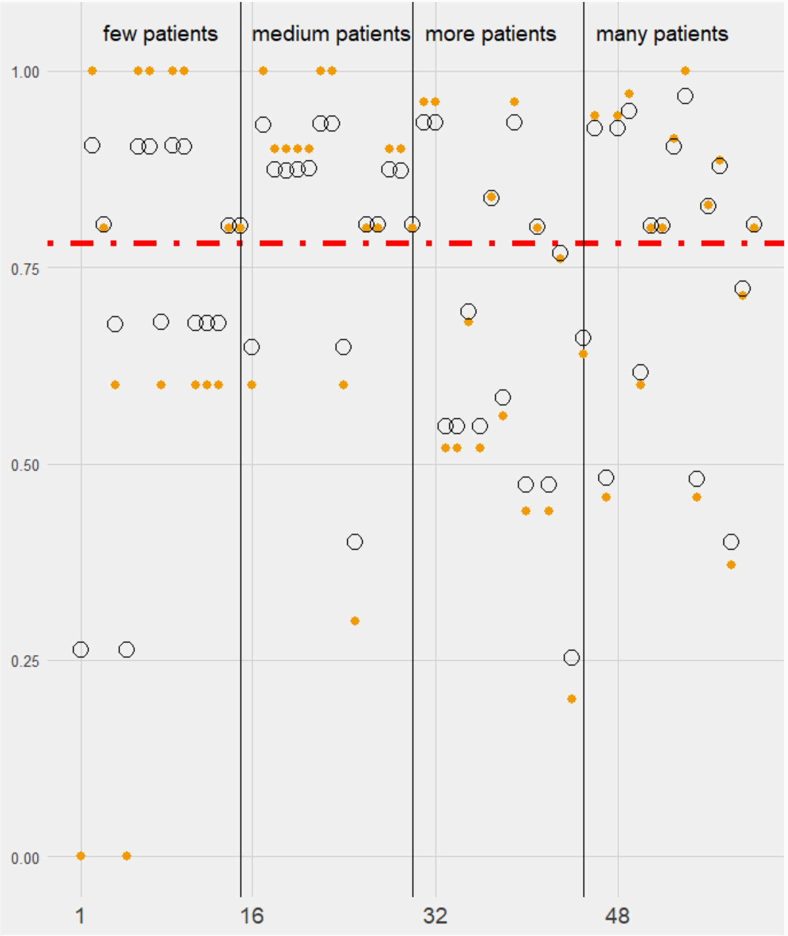
Shrinkage vs No shrinkage

Distribution of estimated prevalence across hospitals



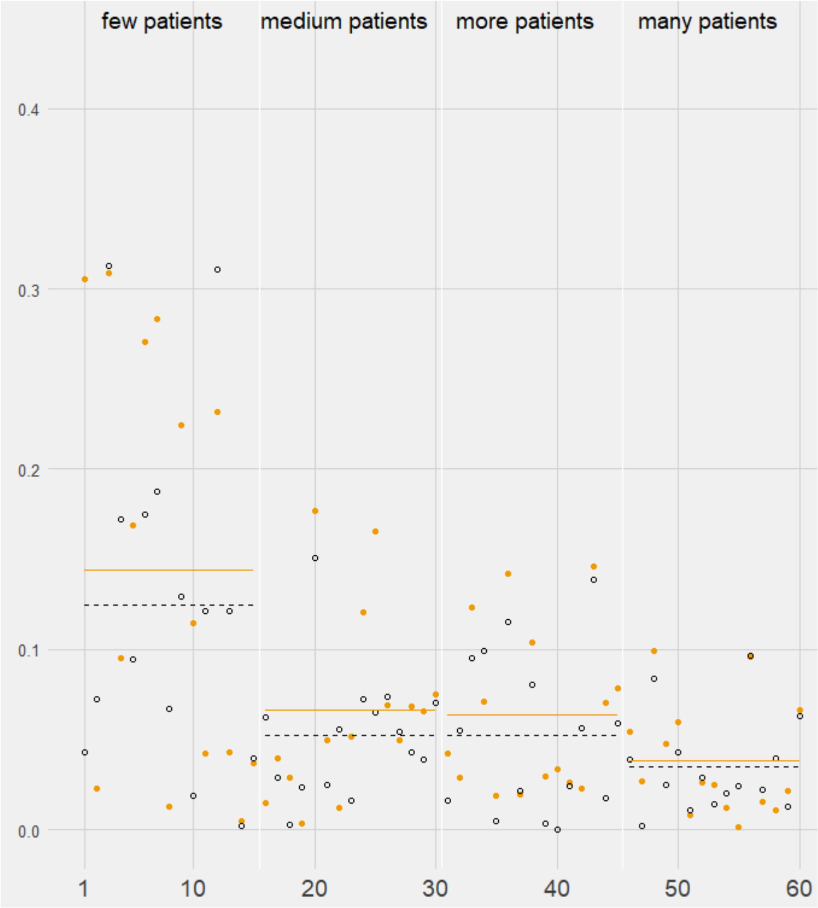
Shrinkage vs No shrinkage

- 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?

Bayesian methods for beginners

- Baysian statistics PhD course Göteborg/Lund (Spring 2021)
- Two good books
  - McElreath, R (2020): Statistical Rethinking. CRC Press
  - Gelman et al (2013): Bayesian data analysis
    - http://www.stat.columbia.edu/~gelman/book/
- Bayesian analysis available in
  - R, SAS, SPSS Stata
- A lot of great online resources