A Mini Introduction to Bayesian Statistics

Yingbo Li

March 3, 2018

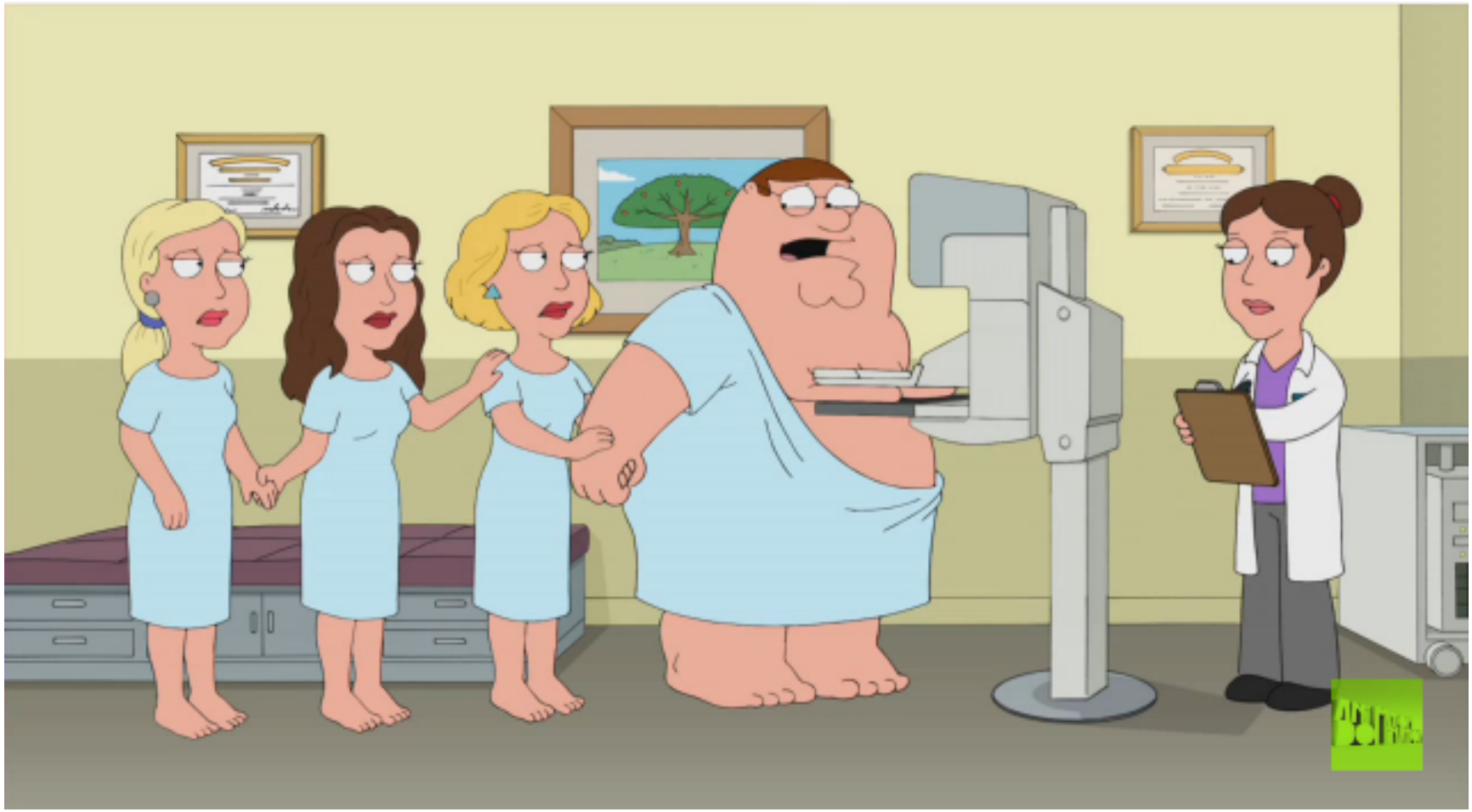
Outline

Bayes rule: the conditional probability

Parameter estimation: dynamic belief updating

Hypothesis testing: no more p-values

If the mammogram is positive, how likely s/he has breast cancer?



Source: http://familyguy.wikia.com/wiki/File:Mammogram.png

If the mammogram is positive, how likely s/he has breast cancer?

- A mammogram is right 75% of the time, if a women has breast cancer

$$P(\textbf{+} \mid 😱) = 0.75$$

Source: Silver, N., (2012) The Signal and the Noise, Chapter 8, page 245

If the mammogram is positive, how likely s/he has breast cancer?

- A mammogram is right 75% of the time, if a women has breast cancer

$$P(\textbf{+} \mid 😱) = 0.75$$

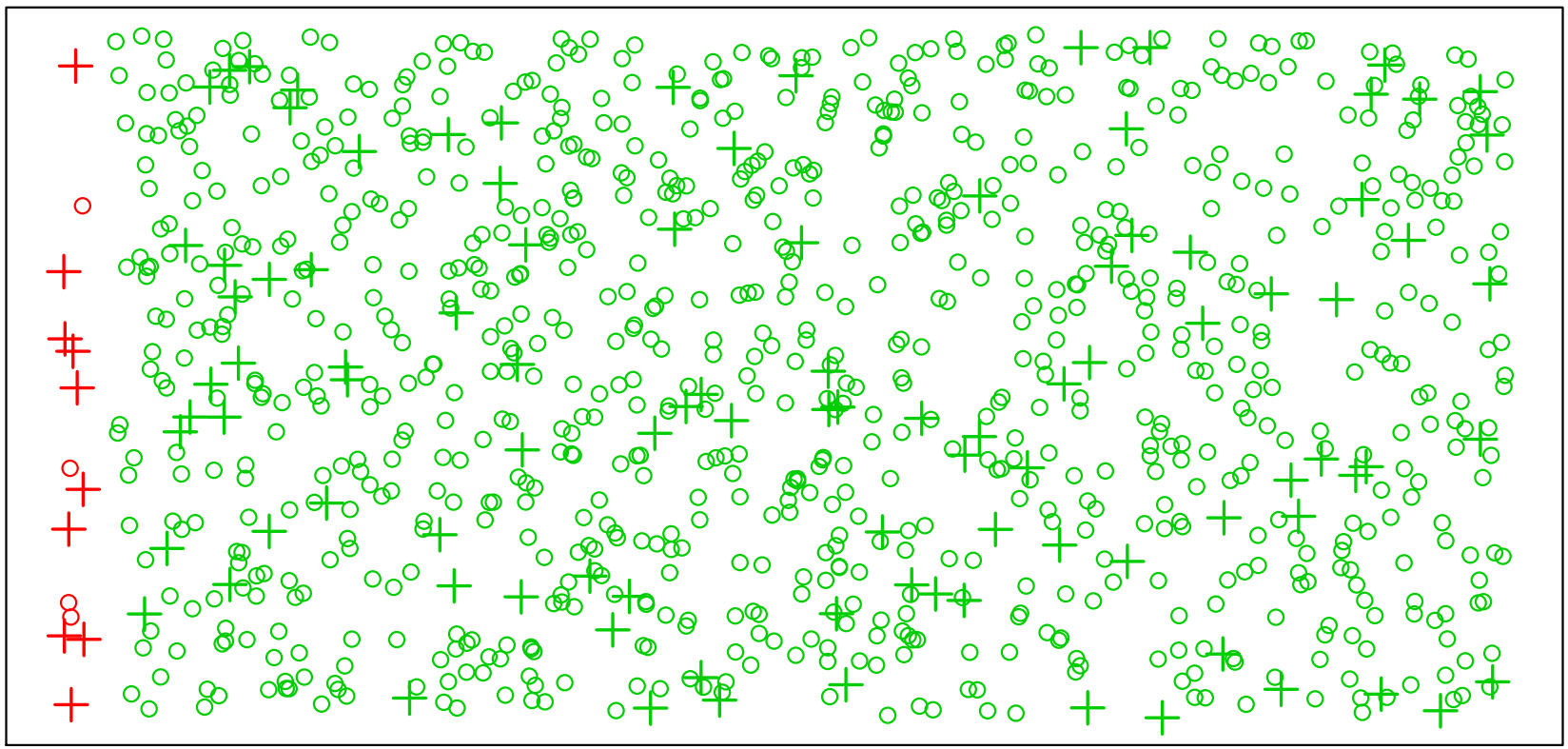
- A mammogram is right 90% of the time, if a women doesn't have breast cancer

$$P(\textbf{+} \mid 🙂) = 0.1$$

Source: Silver, N., (2012) The Signal and the Noise, Chapter 8, page 245

If the mammogram is positive, how likely s/he has breast cancer?

- For a women in her 40's, the chance of having breast cancer is 1.4%.



Source: Silver, N., (2012) The Signal and the Noise, Chapter 8, page 245

If the mammogram is positive, how likely s/he has breast cancer?

$$P(😱 \mid \textbf{+}) = \frac{P(\textbf{+} \mid 😱) \cdot P(😱)}{P(\textbf{+} \mid 😱) \cdot P(😱) + P(\textbf{+} \mid 🙂) \cdot P(🙂)}$$

If the mammogram is positive, how likely s/he has breast cancer?

$$P(😱 \mid \textbf{+}) = \frac{P(\textbf{+} \mid 😱) \cdot P(😱)}{P(\textbf{+} \mid 😱) \cdot P(😱) + P(\textbf{+} \mid 🙂) \cdot P(🙂)}$$

$$= \frac{0.75 \times 0.014}{0.75 \times 0.014 \quad + \quad 0.1 \times 0.986}$$

If the mammogram is positive, how likely s/he has breast cancer?

$$P(😱 \mid \textbf{+}) = \frac{P(\textbf{+} \mid 😱) \cdot P(😱)}{P(\textbf{+} \mid 😱) \cdot P(😱) + P(\textbf{+} \mid 🙂) \cdot P(🙂)}$$

$$= \frac{0.75 \times 0.014}{0.75 \times 0.014 \quad + \quad 0.1 \times 0.986}$$

$$= 0.096$$

Prior vs posterior probabilities

- Event of interest: 😱, having breast cancer

Prior vs posterior probabilities

- Event of interest: 😱, having breast cancer
- Prior probability: opinion on how likely the event will occur before we collecting data

$$P(😱) = 1.4\%$$

Prior vs posterior probabilities

- Event of interest: 😱, having breast cancer
- Prior probability: opinion on how likely the event will occur before we collecting data

$$P(😱) = 1.4\%$$

- Posterior probability: opinion on how likely the event will occur after observing the data

$$P(😱 \mid \textbf{+}) = 9.6\%$$

Bayes rule

$$P(A \mid D) = \frac{P(D \mid A) \cdot P(A)}{P(D \mid A) \cdot P(A) + P(D \mid A^c) \cdot P(A^c)}$$

A occurs

A doesn't occur

What if a 2nd mammogram is again +?

- Assuming the two results are independent

What if a 2nd mammogram is again +?

- Assuming the two results are independent

$$P(😱 \mid \textbf{+}) = \frac{P(\textbf{+} \mid 😱) \cdot P(😱)}{P(\textbf{+} \mid 😱) \cdot P(😱) + P(\textbf{+} \mid 🙂) \cdot P(🙂)}$$

$$= \frac{0.75 \times 0.096}{0.75 \times 0.096 + 0.1 \times (1 - 0.096)}$$

$$= 0.443$$

What if a 2nd mammogram is again +?

- Dynamic update of information: she's no longer an average woman in her 40's, but who already has one positive mammogram.

Bayesian statistics: 250+ yrs history

Thomas Bayes
(1701 - 1761)

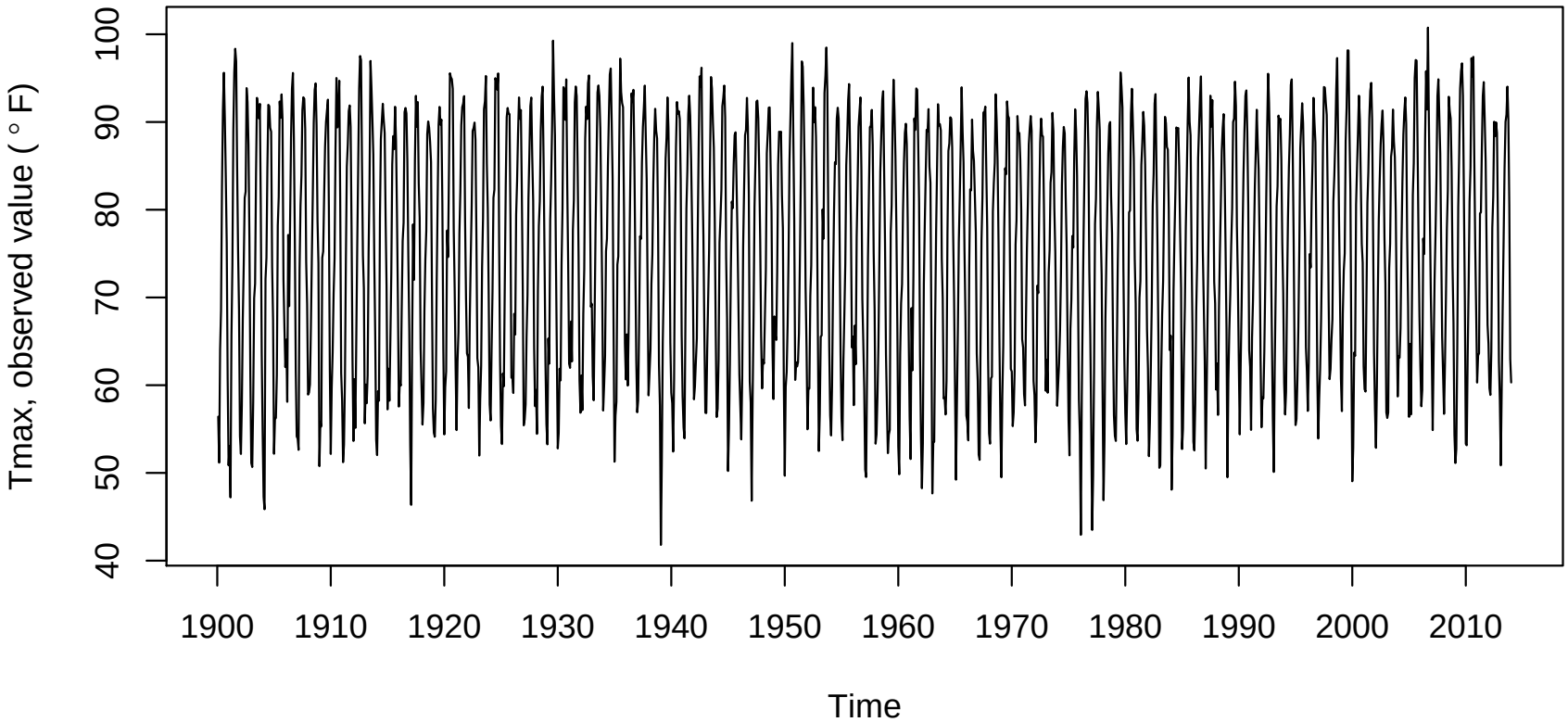
Ronald Fisher
(1890-1962)

Why Bayesian statistics?

- Can incorporate **subjective prior belief** or domain experts knowledge
- No large sample assumption needed
- Markov chain Monte Carlo (MCMC) enables easy inference for sophisticated models
- Revives with revolutionary computation power

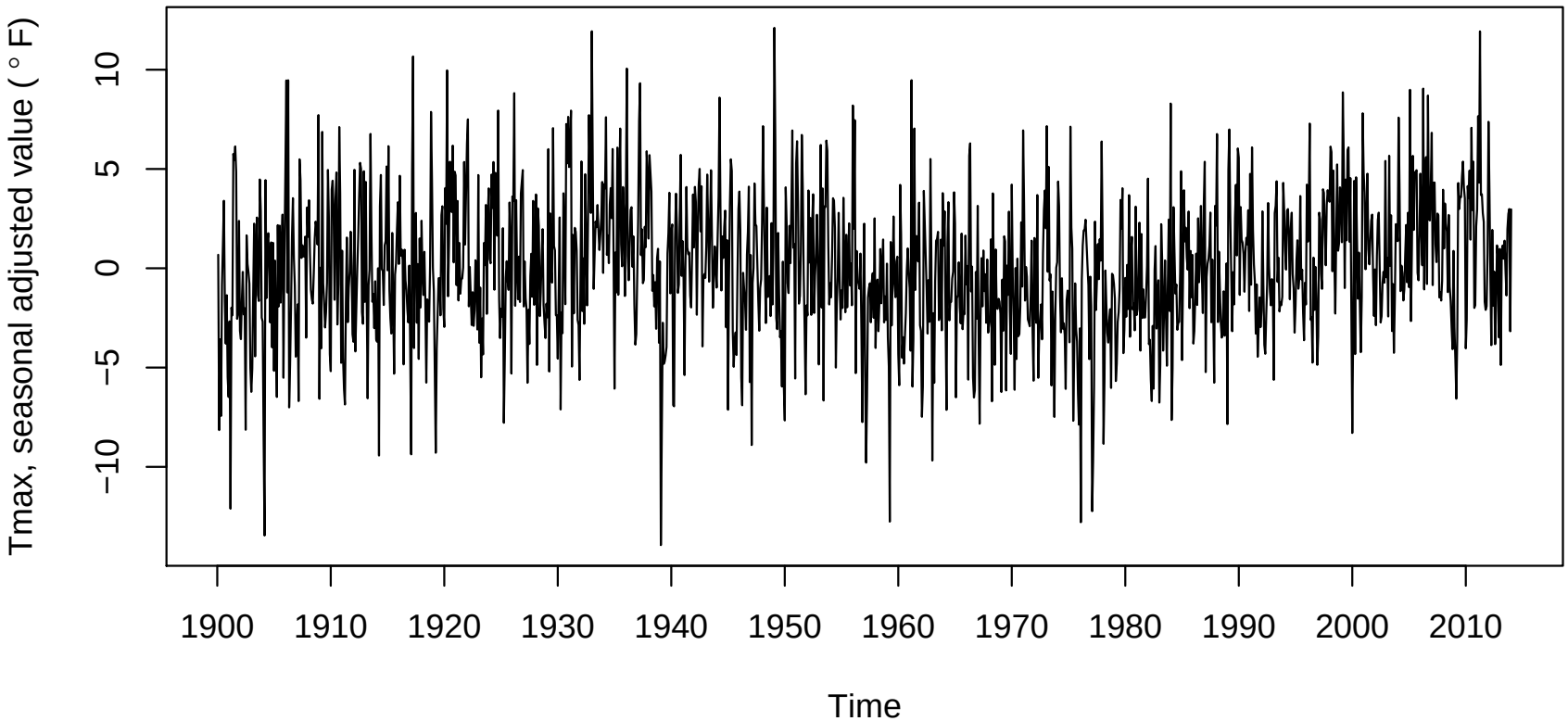
Monthly temperature, Tuscaloosa AL

Observed data



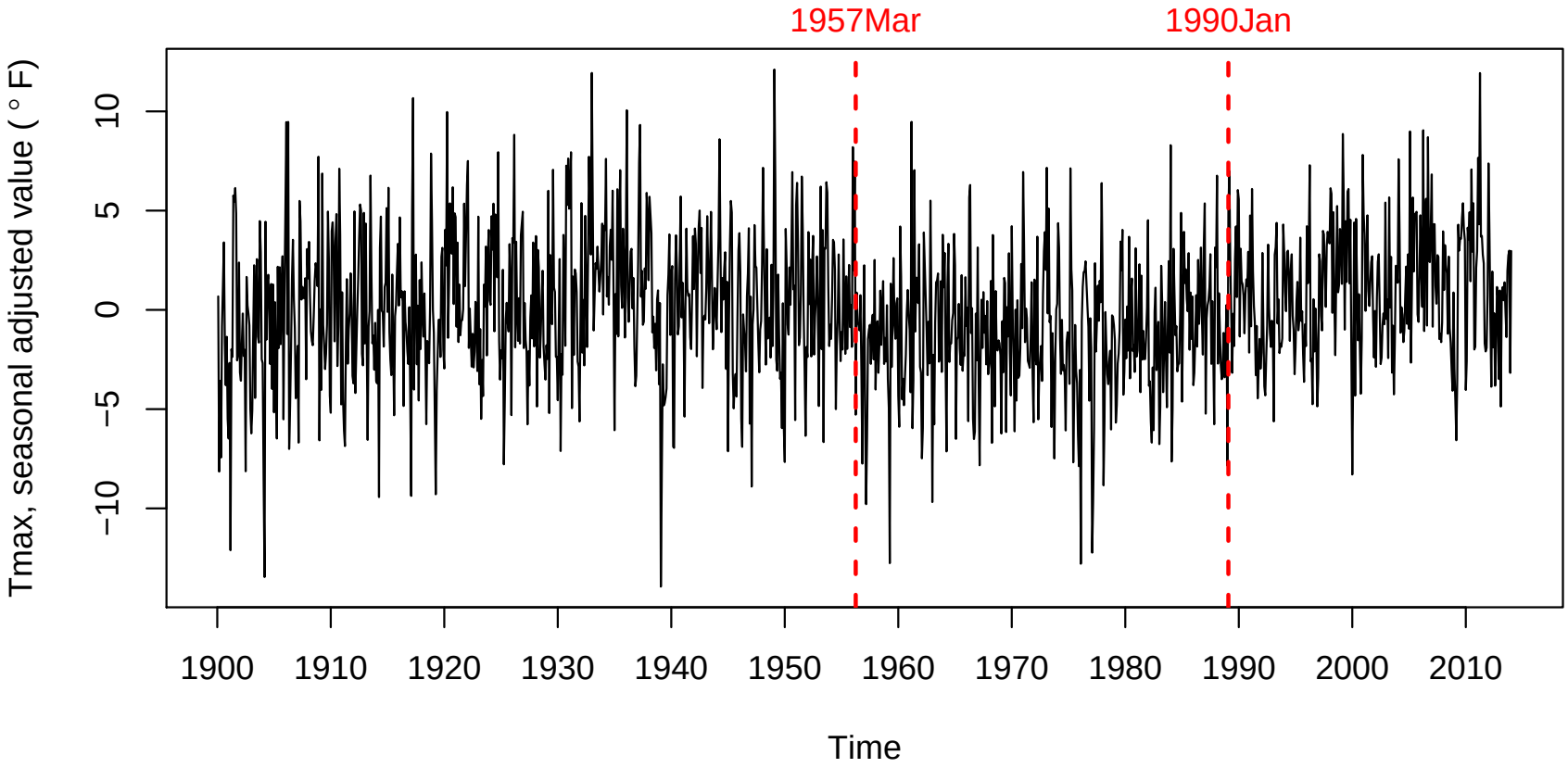
Monthly temperature, Tuscaloosa AL

Observed data − sample seasonal mean



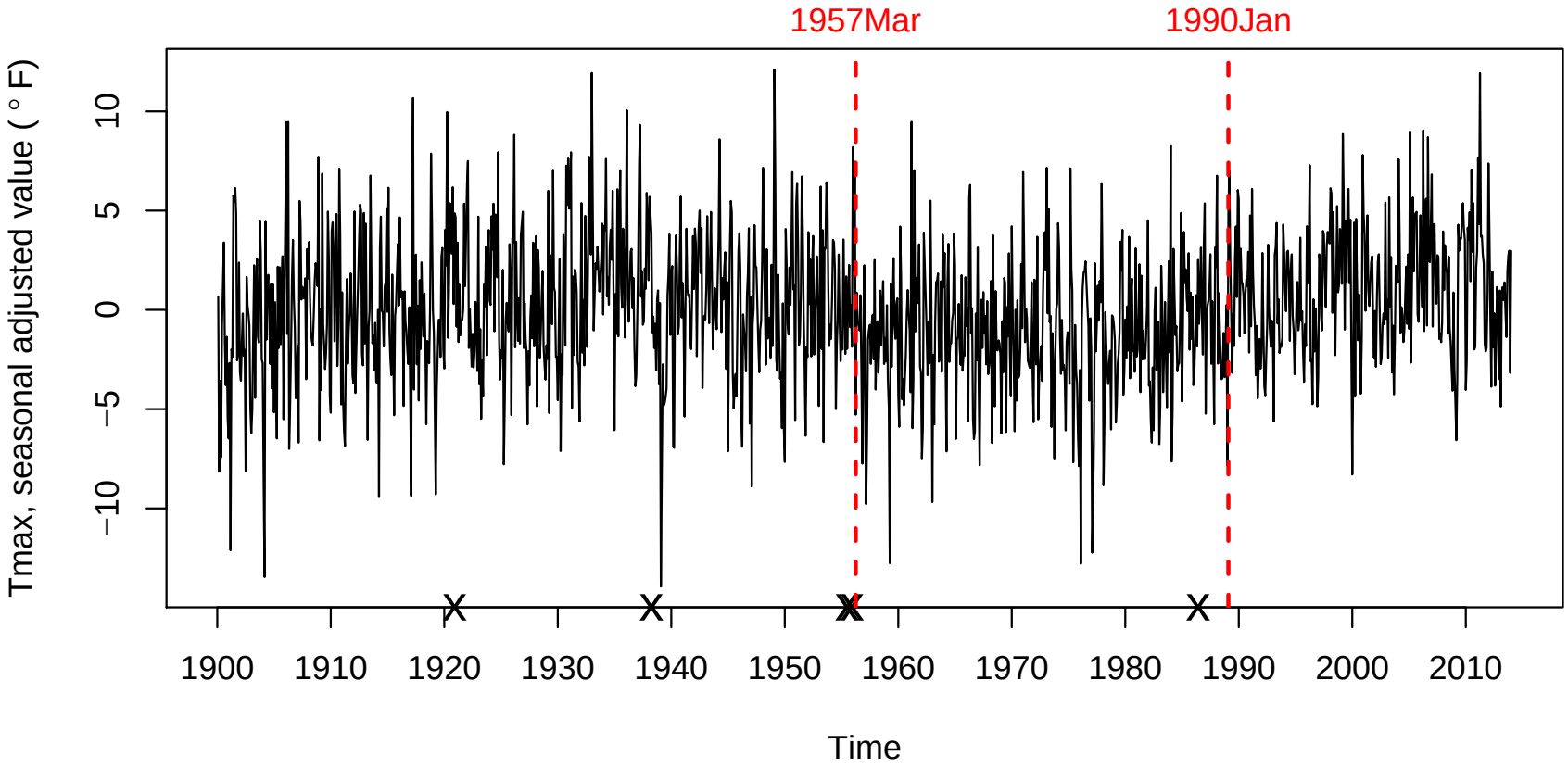
Monthly temperature, Tuscaloosa AL

Observed data − sample seasonal mean



Monthly temperature, Tuscaloosa AL

Observed data – sample seasonal mean

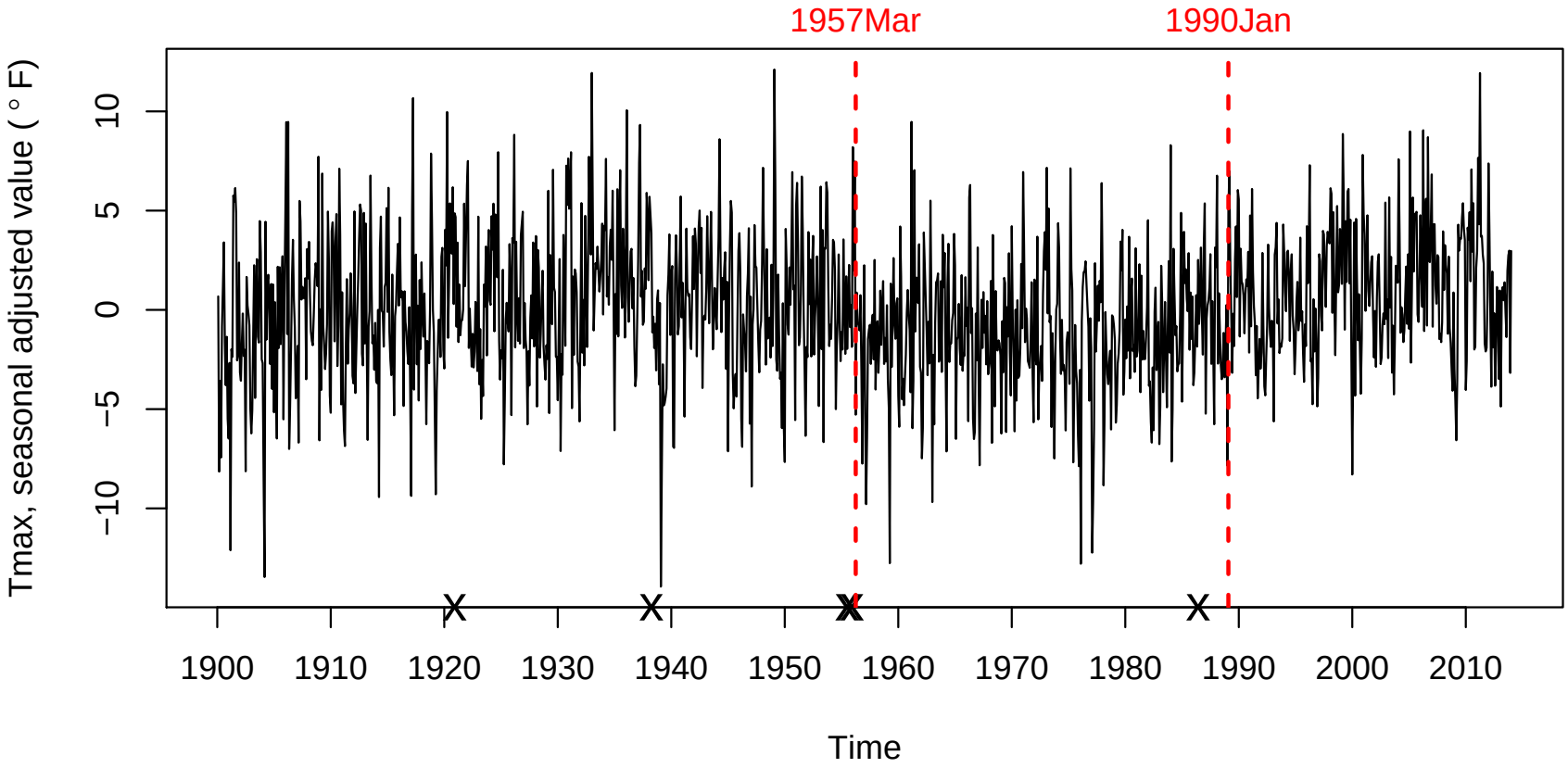


Metadata (station history logs)

- Station relocations: 1921/11, 1939/03, 1956/06
- Instrumentation changes: 1956/11, 1987/05

Monthly temperature, Tuscaloosa AL

Observed data – sample seasonal mean

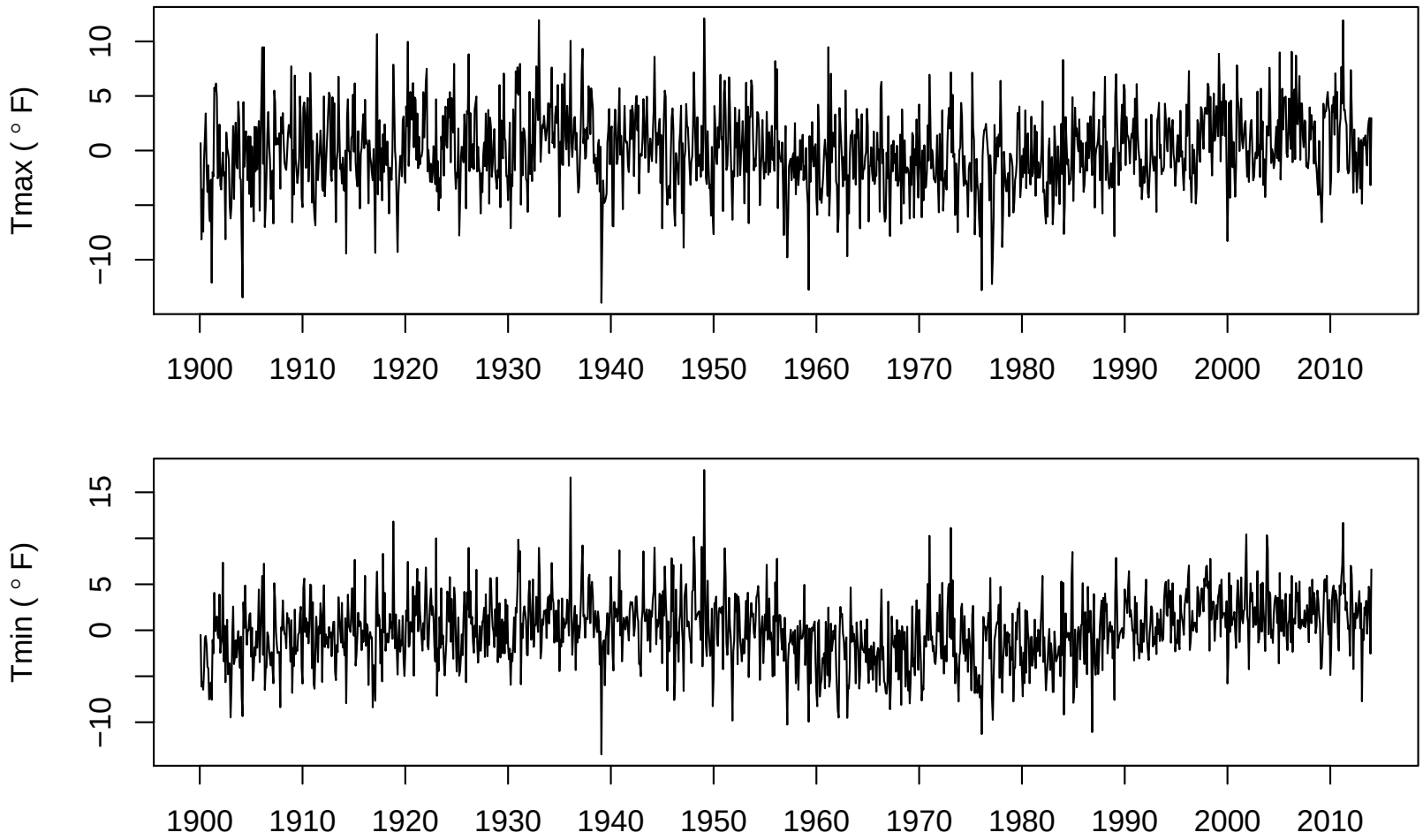


Metadata: more likely to induce mean shifts

- Station relocations: 1921/11, 1939/03, 1956/06
- Instrumentation changes: 1956/11, 1987/05

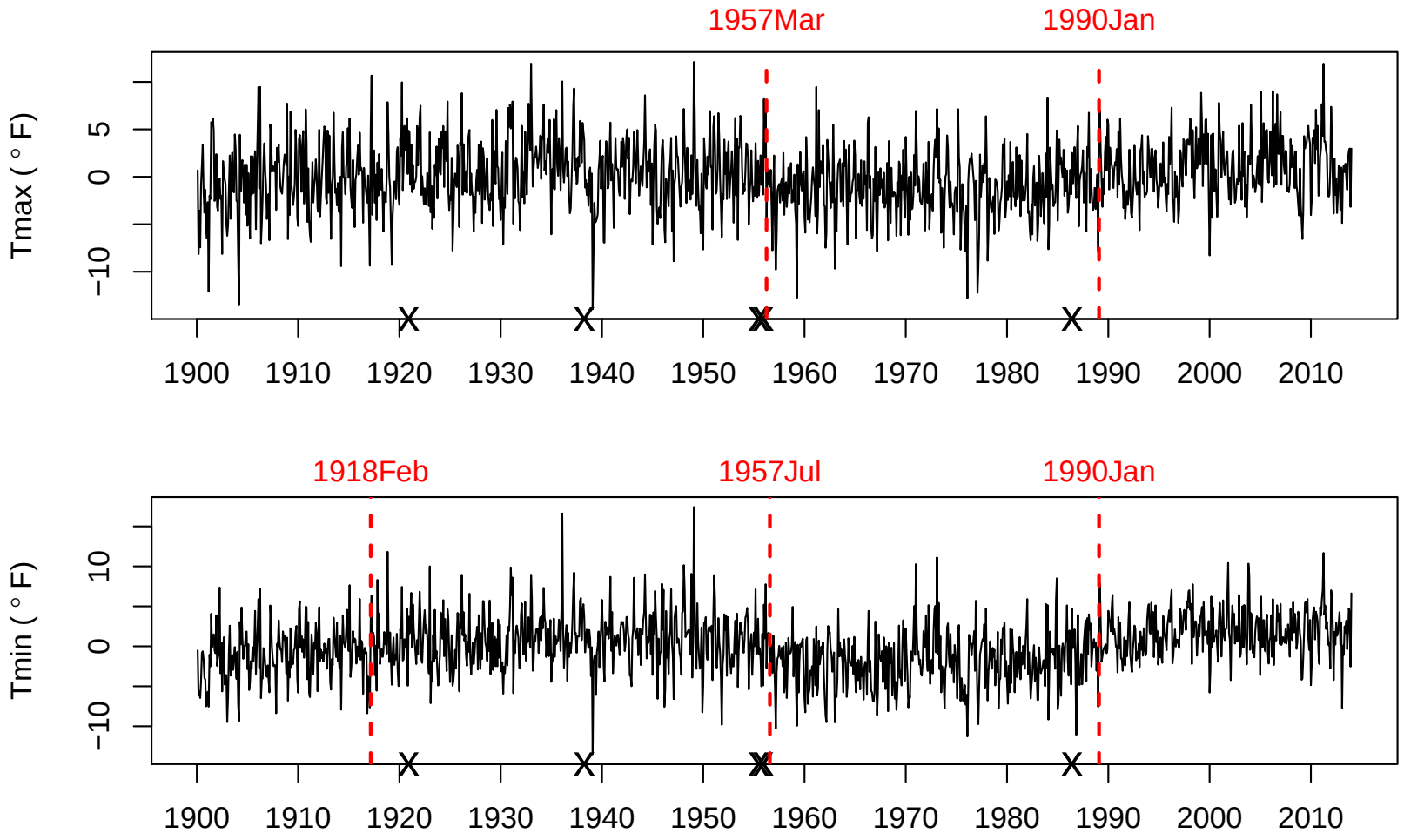
Monthly max and min temperatures

Observed data − sample seasonal mean

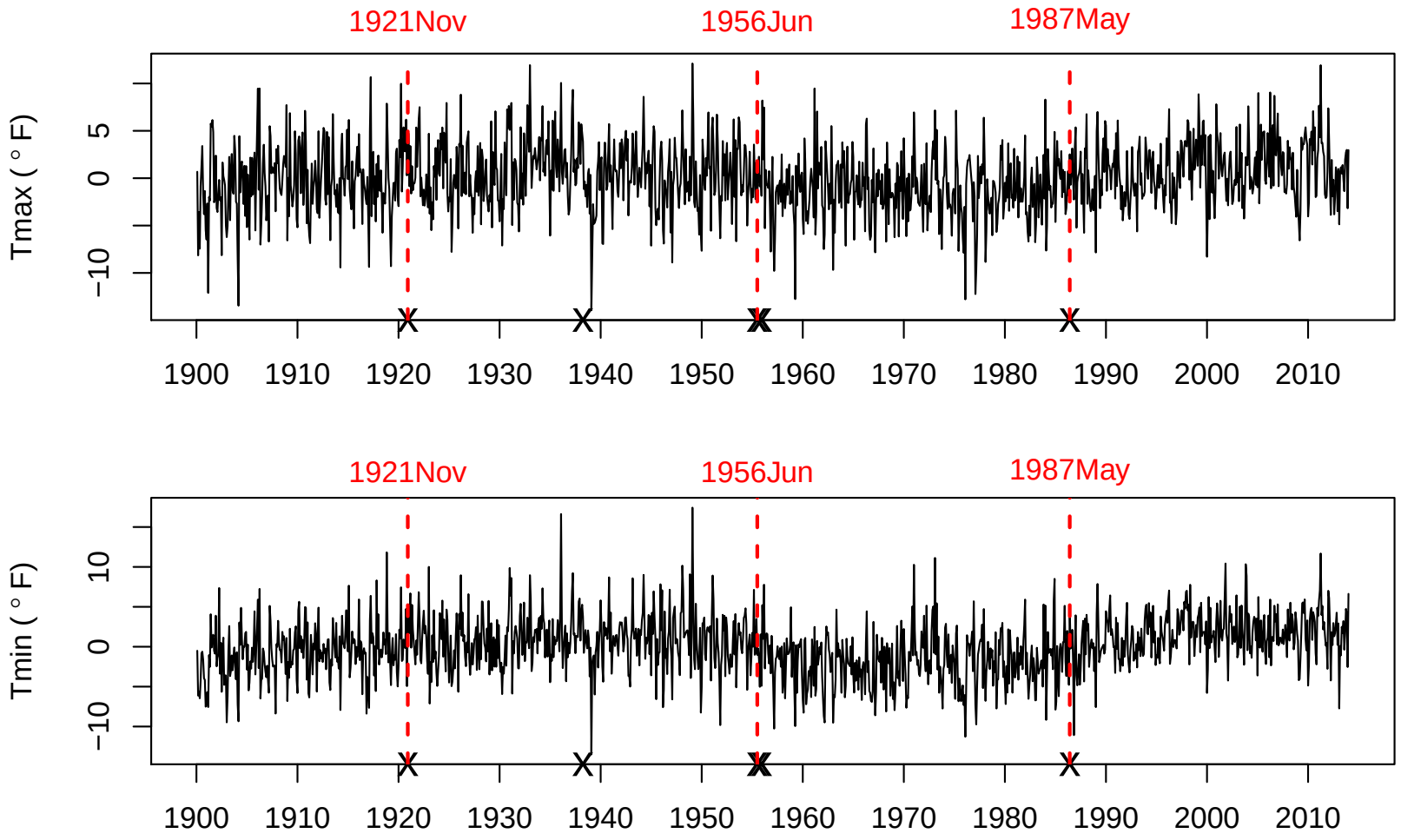


Tmax and Tmin: likely to shift at the same time

Existing changepoint detection



Bayesian changepoint detection



The denominator is not important

$$P(A \mid D) = \frac{P(D \mid A) \cdot P(A)}{P(D \mid A) \cdot P(A) + P(D \mid A^c) \cdot P(A^c)}$$

$$P(A^c \mid D) = \frac{P(D \mid A^c) \cdot P(A^c)}{P(D \mid A) \cdot P(A) + P(D \mid A^c) \cdot P(A^c)}$$

The denominator is not important

$$P(A \mid D) = \frac{P(D \mid A) \cdot P(A)}{P(D \mid A) \cdot P(A) + P(D \mid A^c) \cdot P(A^c)}$$

$$P(A^c \mid D) = \frac{P(D \mid A^c) \cdot P(A^c)}{P(D \mid A) \cdot P(A) + P(D \mid A^c) \cdot P(A^c)}$$

The denominator is not important

$$P(A \mid D) \propto P(D \mid A) \cdot P(A)$$

$$P(A^c \mid D) \propto P(D \mid A^c) \cdot P(A^c)$$

Bayes rule: distribution version

- Data D is randomly generated from a model
- With unknown parameter θ

$$p(\theta \mid D) \quad \propto \quad p(D \mid \theta) \times \quad p(\theta)$$

posterior distribution | likelihood | prior distribution

Bayes rule: distribution version

- Data D is randomly generated from a model
- With unknown parameter θ

$$p(\theta \mid D) \quad \propto \quad p(D \mid \theta) \times \quad p(\theta)$$

posterior distribution | likelihood | prior distribution

- Estimate θ using the posterior mean $E(\theta \mid D)$

Is it a fair coin?

Source: https://www.li-lacchocolates.com/assets/images/Chanukah/Gold-Chocolate-Coins.jpg

Is it a fair coin?

- Parameter θ: chance of Head
- Data D: coin flipping Head/Tail outcomes
- Likelihood function:

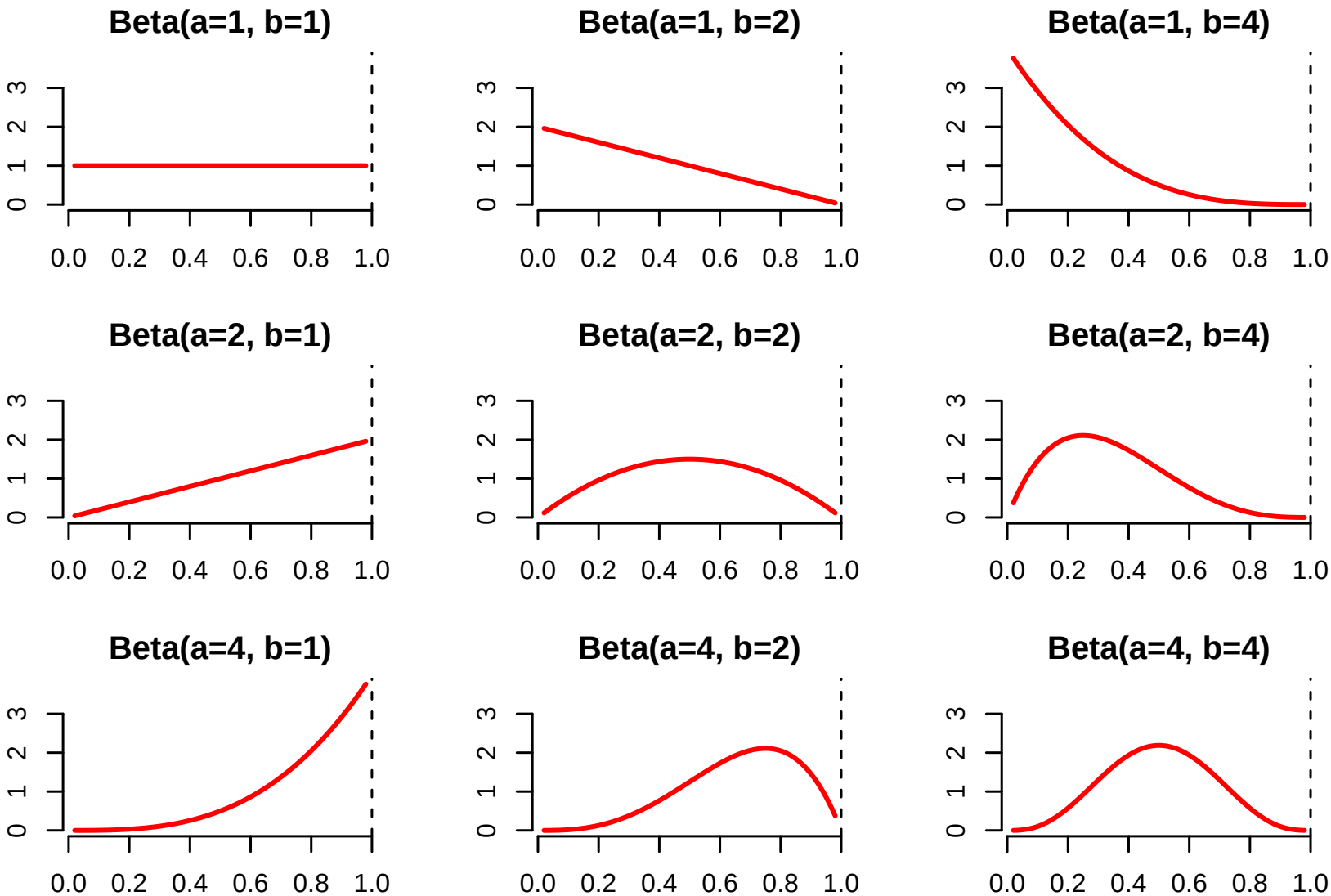
$$p(D \mid \theta) = \theta^{\#\text{Head}}(1-\theta)^{\#\text{Tail}}$$

Beta distribution: between 0 and 1

$$p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$$

Beta distribution: between 0 and 1

$$p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$$



Posterior distribution is again Beta

$$p(\theta \mid D) \propto p(D \mid \theta) \times p(\theta)$$

posterior | likelihood | prior

Posterior distribution is again Beta

$$p(\theta \mid D) \propto \left[\theta^{\#\text{Head}}(1-\theta)^{\#\text{Tail}}\right] \times p(\theta)$$

posterior — likelihood — prior

Posterior distribution is again Beta

$$p(\theta \mid D) \propto \left[\theta^{\#\text{Head}}(1-\theta)^{\#\text{Tail}}\right] \times \left[\theta^{a-1}(1-\theta)^{b-1}\right]$$

posterior likelihood prior

Posterior distribution is again Beta

$$p(\theta \mid D) \propto \left[\theta^{\#\text{Head}}(1-\theta)^{\#\text{Tail}}\right] \times \left[\theta^{a-1}(1-\theta)^{b-1}\right]$$

posterior likelihood prior

$$\theta \mid D \sim \text{Beta}\left(\#\text{Head} + a, \#\text{Tail} + b\right)$$

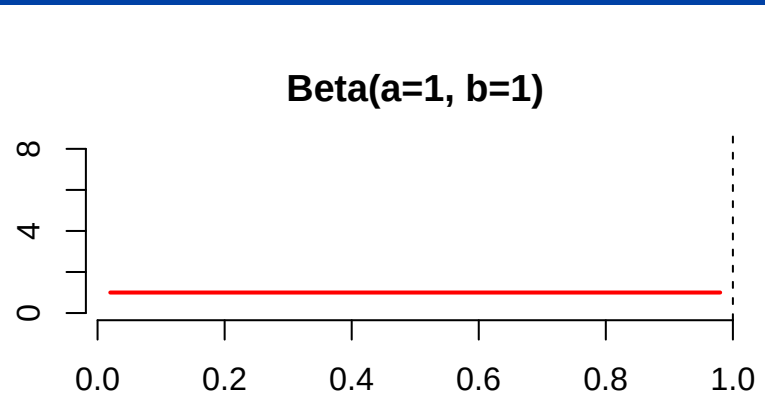
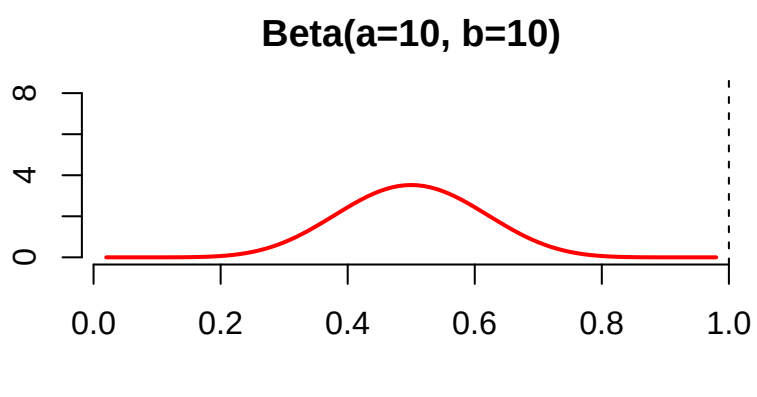
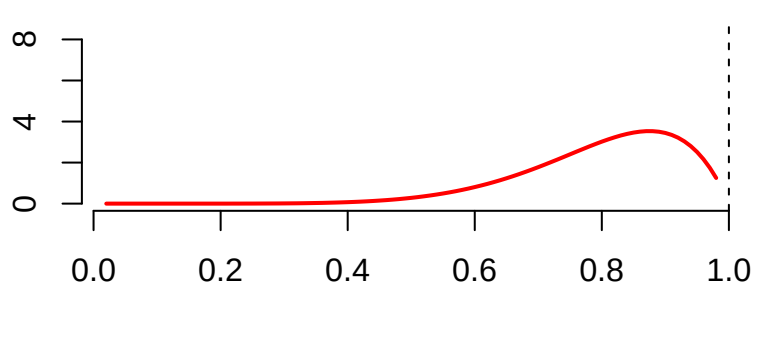
Posterior distribution is again Beta

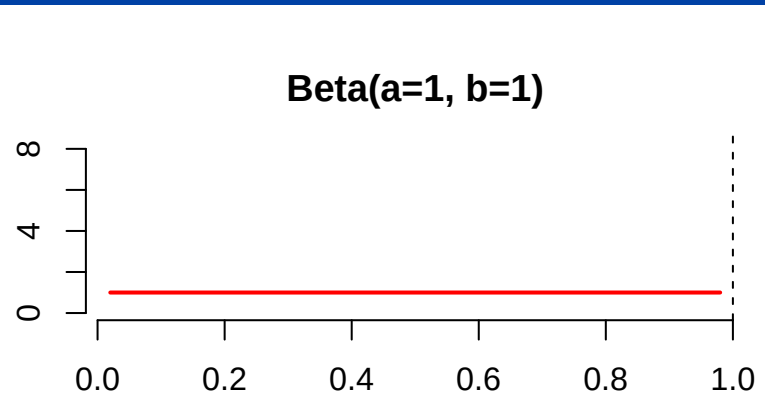
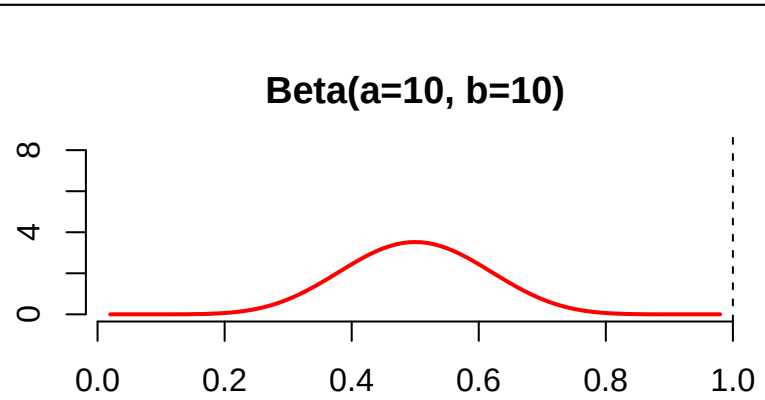
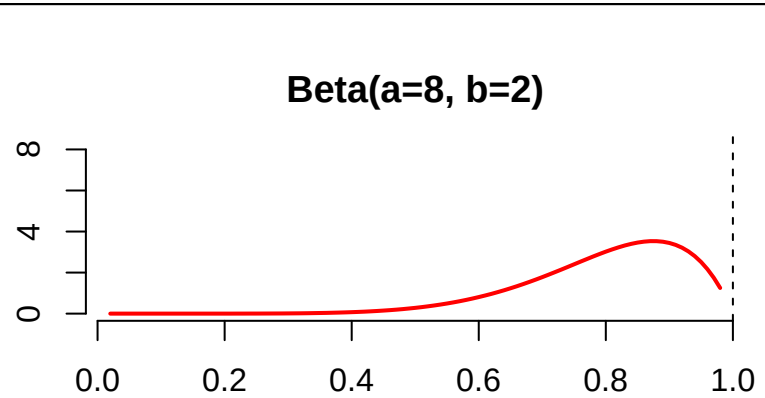
$$p(\theta \mid D) \propto \left[\theta^{\#\text{Head}}(1-\theta)^{\#\text{Tail}}\right] \times \left[\theta^{a-1}(1-\theta)^{b-1}\right]$$

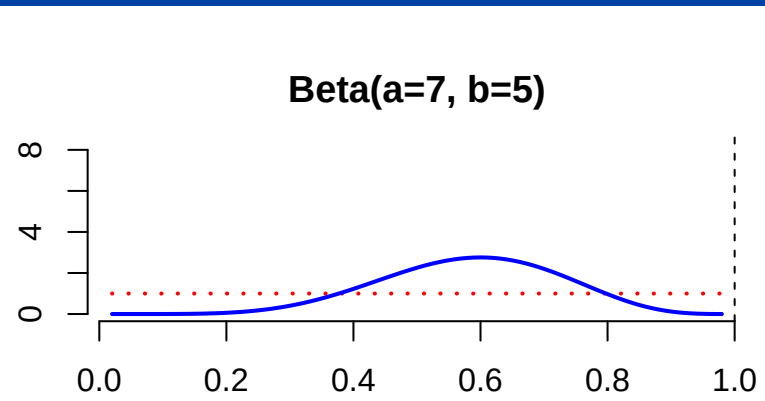
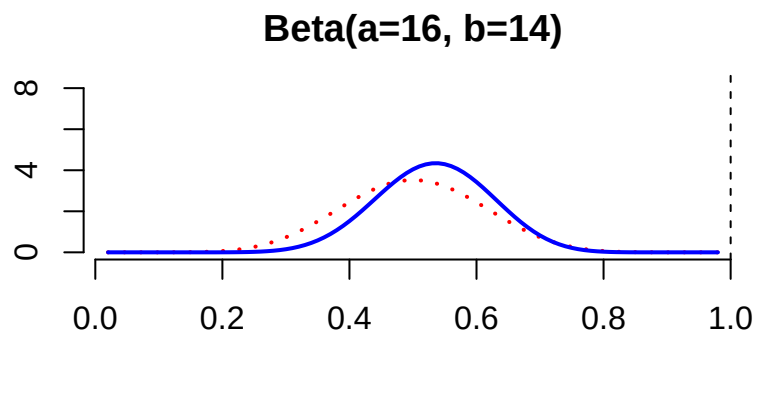
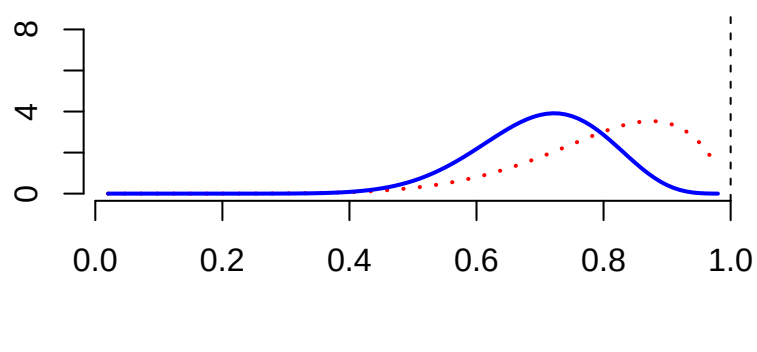
posterior likelihood prior

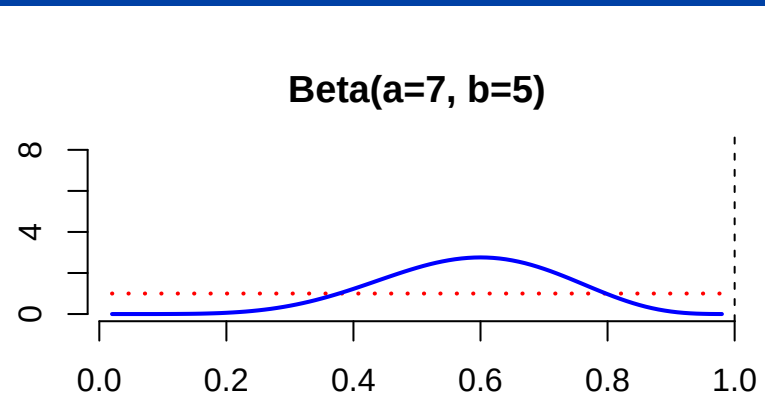
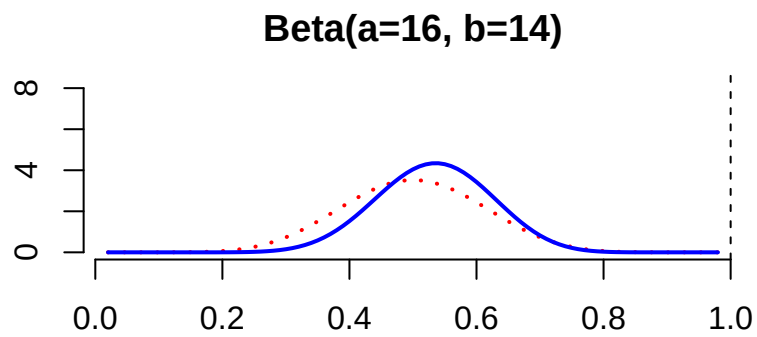
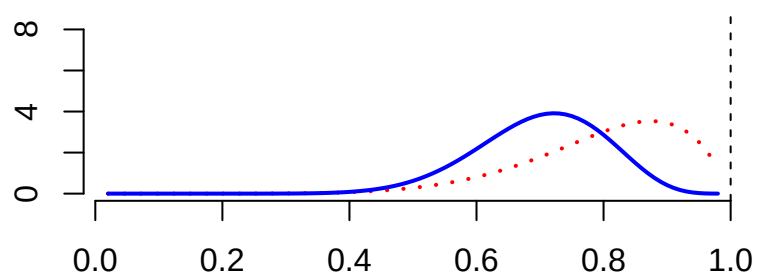
$$\theta \mid D \sim \text{Beta}\left(\#\text{Head} + a, \#\text{Tail} + b\right)$$

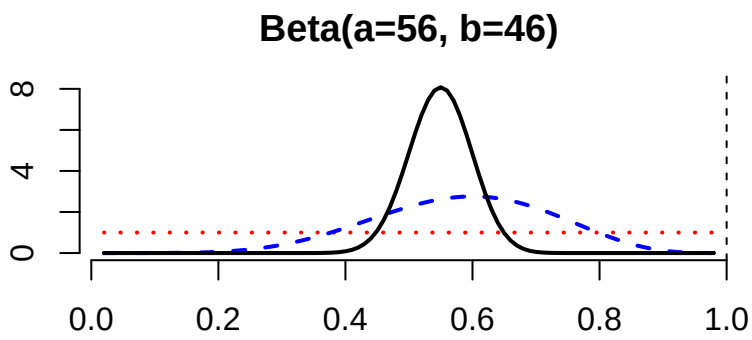
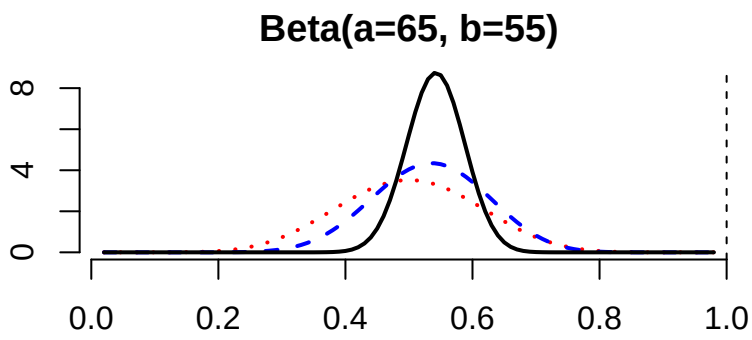
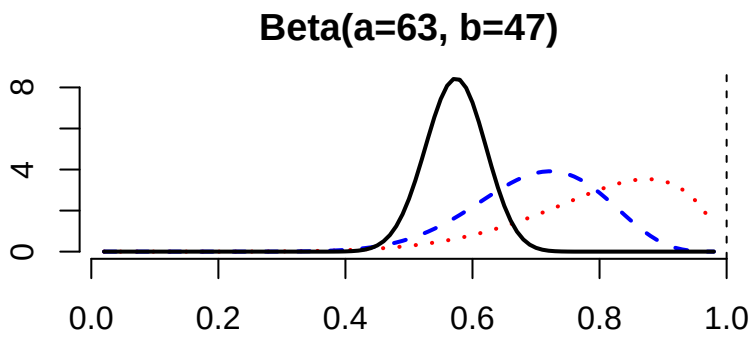
- Conjugacy: prior and posterior in the same distribution family

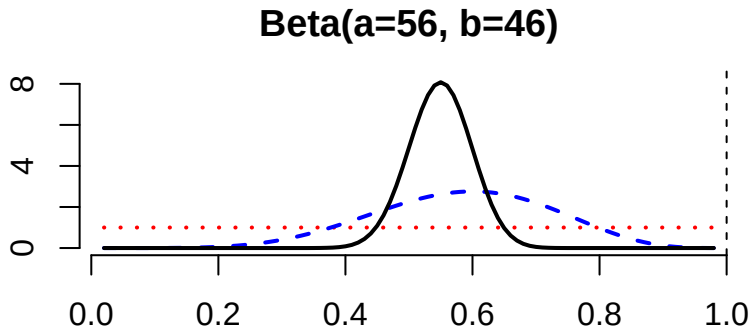
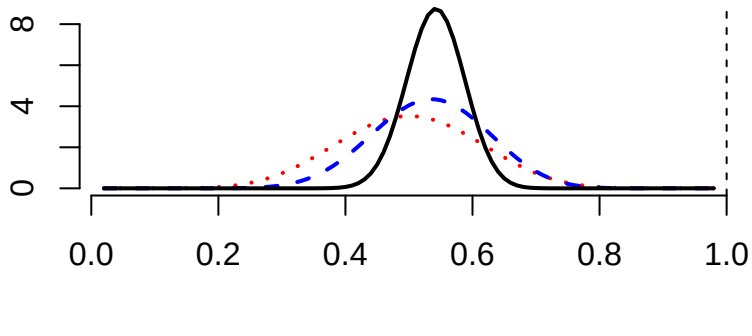
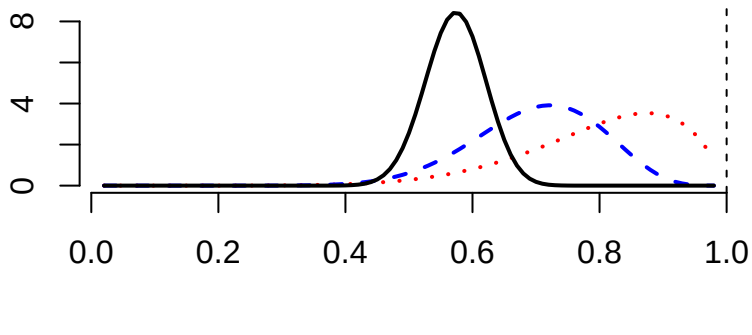
Person	Prior	Mean	SD
1			
2			
3			

Person	Prior	Mean	SD
1		0.50	0.29
2		0.50	0.11
3		0.80	0.12

Person	Posterior, 6 head / 4 tails	Mean	SD
1		0.50	0.29
2		0.50	0.11
3		0.80	0.12

Person	Posterior, 6 head / 4 tails	Mean	SD
1		0.50 0.58	0.29 0.14
2		0.50 0.53	0.11 0.09
3		0.80 0.70	0.12 0.10

Person	Posterior, 55 head / 45 tails	Mean	SD
1		0.50 0.58	0.29 0.14
2		0.50 0.53	0.11 0.09
3		0.80 0.70	0.12 0.10

Person	Posterior, 55 head / 45 tails	Mean	SD
1		0.50 0.58 0.55	0.29 0.14 0.05
2		0.50 0.53 0.54	0.11 0.09 0.05
3		0.80 0.70 0.57	0.12 0.10 0.05

Movie ratings on Rotten Tomatoes

Rotten Tomatoes staffs

- collect reviews from certified movie critics, and
- determine whether each review is

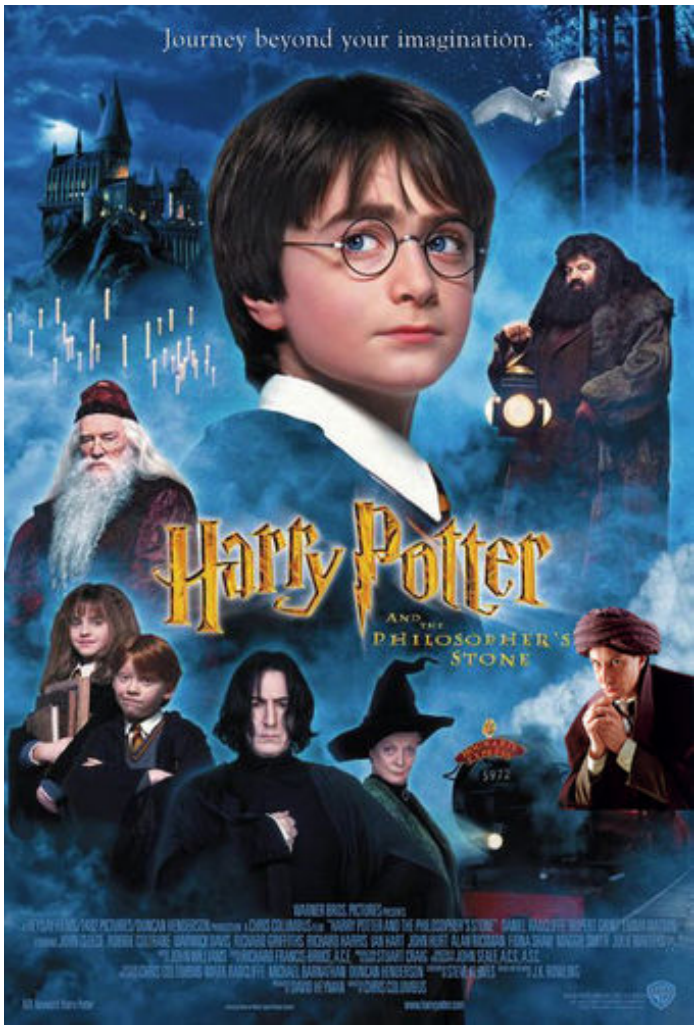
positive “fresh” 

vs

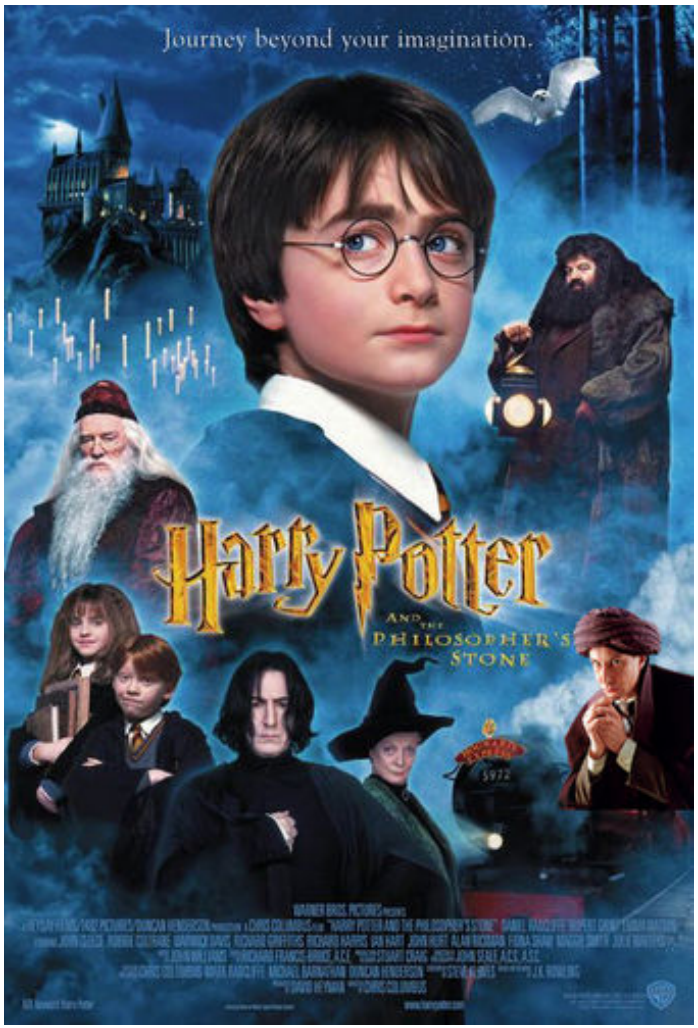
negative “rotten” 

https://en.wikipedia.org/wiki/Rotten_Tomatoes

Harry Potter and the Philosopher's Stone (movie 1)



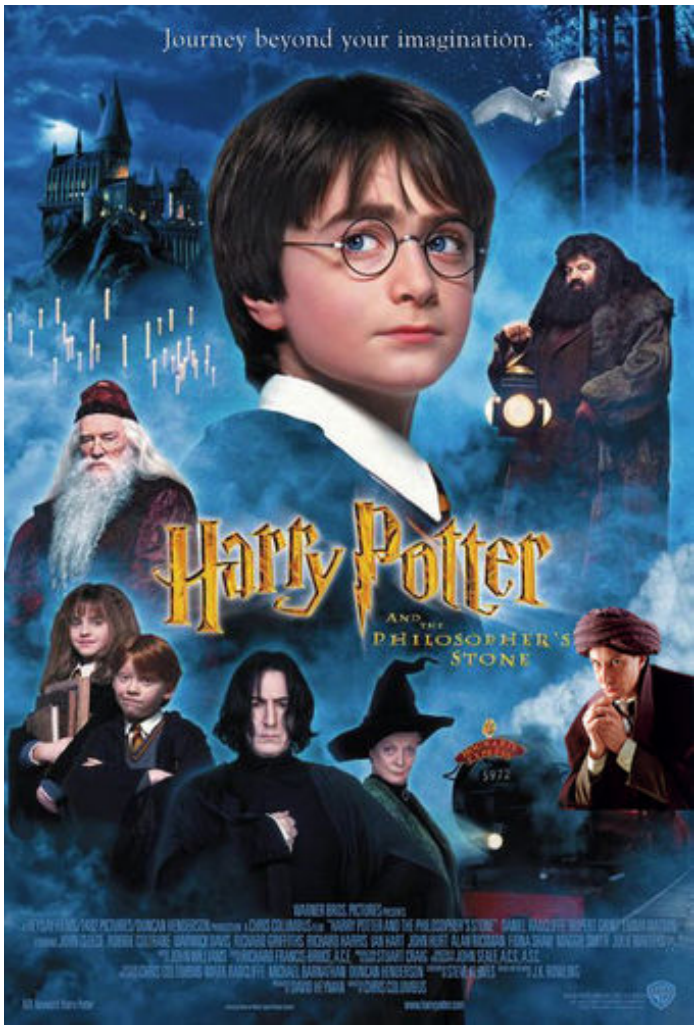
Harry Potter and the Philosopher's Stone (movie 1)



- Prior distribution

$$\text{mean} = 50\%, \ \text{sd} = 28.8\%$$

Harry Potter and the Philosopher's Stone (movie 1)



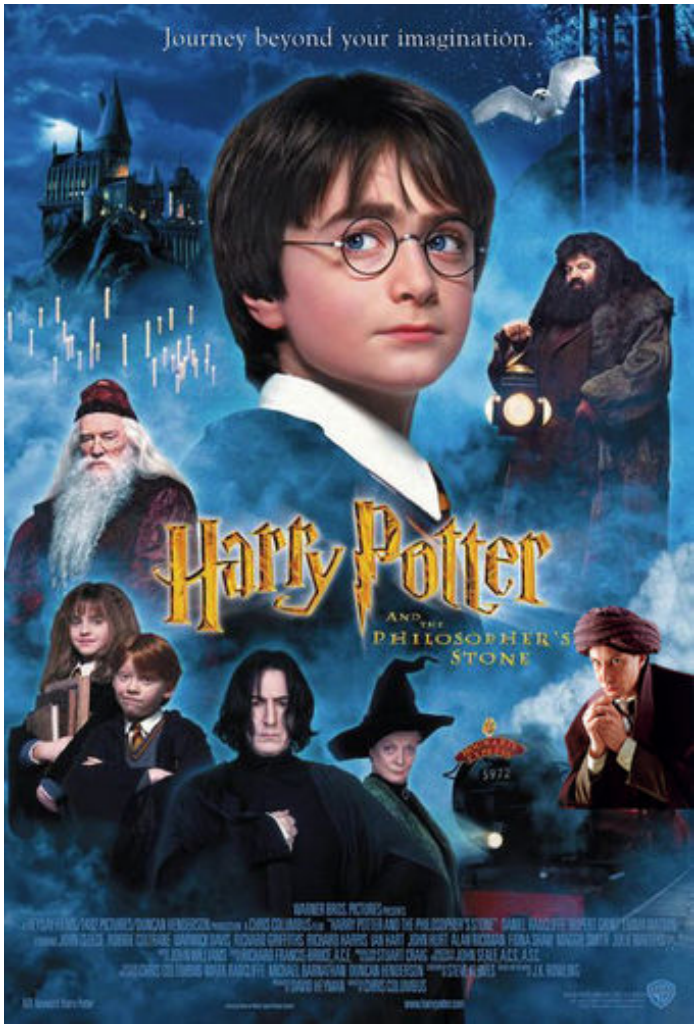
- Prior distribution

$$\text{mean} = 50\%, \ \text{sd} = 28.8\%$$

- Data

🍅 155, 38

Harry Potter and the Philosopher's Stone (movie 1)



- Prior distribution

$$\text{mean} = 50\%, \ \text{sd} = 28.8\%$$

- Data

🍅 155, 38

- Posterior distribution

$$\text{mean} = 80\%, \ \text{sd} = 2.9\%$$

Harry Potter and the Chamber of Secrets (movie 2)

Harry Potter and the Chamber of Secrets (movie 2)

- Prior distribution

$$\text{mean} = 80\%, \ \text{sd} = 2.9\%$$

Harry Potter and the Chamber of Secrets (movie 2)

- Prior distribution

$$\text{mean} = 80\%, \ \text{sd} = 2.9\%$$

- Data

 191,  41

Harry Potter and the Chamber of Secrets (movie 2)

- Prior distribution

$$\text{mean} = 80\%, \ \text{sd} = 2.9\%$$

- Data

 191,  41

- Posterior distribution

$$\text{mean} = 81.3\%, \ \text{sd} = 1.9\%$$

Harry Potter and the Prisoner of Azkaban (movie 3)

Harry Potter and the Prisoner of Azkaban (movie 3)

- Prior distribution

$$\text{mean} = 81.3\%, \ \text{sd} = 1.9\%$$

Harry Potter and the Prisoner of Azkaban (movie 3)

- Prior distribution

$$\text{mean} = 81.3\%, \ \text{sd} = 1.9\%$$

- Data

$$\text{🍅}\ 229, \ \text{🍀}\ 24$$

Harry Potter and the Prisoner of Azkaban (movie 3)

- Prior distribution

  $\text{mean} = 81.3\%, \ \text{sd} = 1.9\%$

- Data

  $229, \ 24$

- Posterior distribution

  $\text{mean} = 84.7\%, \ \text{sd} = 1.4\%$

Harry Potter and the Goblet of Fire (movie 4)

Harry Potter and the Goblet of Fire (movie 4)

- Prior distribution

$$\text{mean} = 84.7\%, \ \text{sd} = 1.4\%$$

Harry Potter and the Goblet of Fire (movie 4)

- Prior distribution

$$\text{mean} = 84.7\%, \ \text{sd} = 1.4\%$$

- Data

🍅 218, 🍀 30

Harry Potter and the Goblet of Fire (movie 4)

- Prior distribution

  mean $= 84.7\%,$ sd $= 1.4\%$

- Data

  🍅 $218,$ 🍀 30

- Posterior distribution

  mean $= 85.6\%,$ sd $= 1.2\%$

Harry Potter and the Order of the Phoenix (movie 5)

Harry Potter and the Order of the Phoenix (movie 5)

- Prior distribution

$$\text{mean} = 85.6\%, \ \text{sd} = 1.2\%$$

Harry Potter and the Order of the Phoenix (movie 5)

- Prior distribution

$$\text{mean} = 85.6\%, \ \text{sd} = 1.2\%$$

- Data

 192,  53

Harry Potter and the Order of the Phoenix (movie 5)

- Prior distribution

  $\text{mean} = 85.6\%, \ \text{sd} = 1.2\%$

- Data

   192,  53

- Posterior distribution

  $\text{mean} = 84.1\%, \ \text{sd} = 1.1\%$

Harry Potter and the Half-Blood Prince (movie 6)

Harry Potter and the Half-Blood Prince (movie 6)

- Prior distribution

$$\text{mean} = 84.1\%, \ \text{sd} = 1.1\%$$

Harry Potter and the Half-Blood Prince (movie 6)

- Prior distribution

  $\text{mean} = 84.1\%, \ \text{sd} = 1.1\%$

- Data

  🍅 226, 43

Harry Potter and the Half-Blood Prince (movie 6)

- Prior distribution

  mean $= 84.1\%,$ sd $= 1.1\%$

- Data

   $226,$  43

- Posterior distribution

  mean $= 84.0\%,$ sd $= 1.0\%$

Harry Potter and the Deathly Hallows Part 1 (movie 7)

Harry Potter and the Deathly Hallows Part 1 (movie 7)

- Prior distribution

$$\text{mean} = 84.0\%, \ \text{sd} = 1.0\%$$

Harry Potter and the Deathly Hallows Part 1 (movie 7)

- Prior distribution

$$\text{mean} = 84.0\%, \ \text{sd} = 1.0\%$$

- Data

🍅 208, 🧩 57

Harry Potter and the Deathly Hallows Part 1 (movie 7)

- Prior distribution

$$\text{mean} = 84.0\%, \ \text{sd} = 1.0\%$$

- Data

$$208, \ 57$$

- Posterior distribution

$$\text{mean} = 83.2\%, \ \text{sd} = 0.9\%$$

Harry Potter and the Deathly Hallows Part 2 (movie 8)

Harry Potter and the Deathly Hallows Part 2 (movie 8)

- Prior distribution

$$\text{mean} = 83.2\%, \ \text{sd} = 0.9\%$$

Harry Potter and the Deathly Hallows Part 2 (movie 8)

- Prior distribution

  $$\text{mean} = 83.2\%, \ \text{sd} = 0.9\%$$

- Data

  🍅 303, 🧩 13

Harry Potter and the Deathly Hallows Part 2 (movie 8)

- Prior distribution

  $\text{mean} = 83.2\%, \ \text{sd} = 0.9\%$

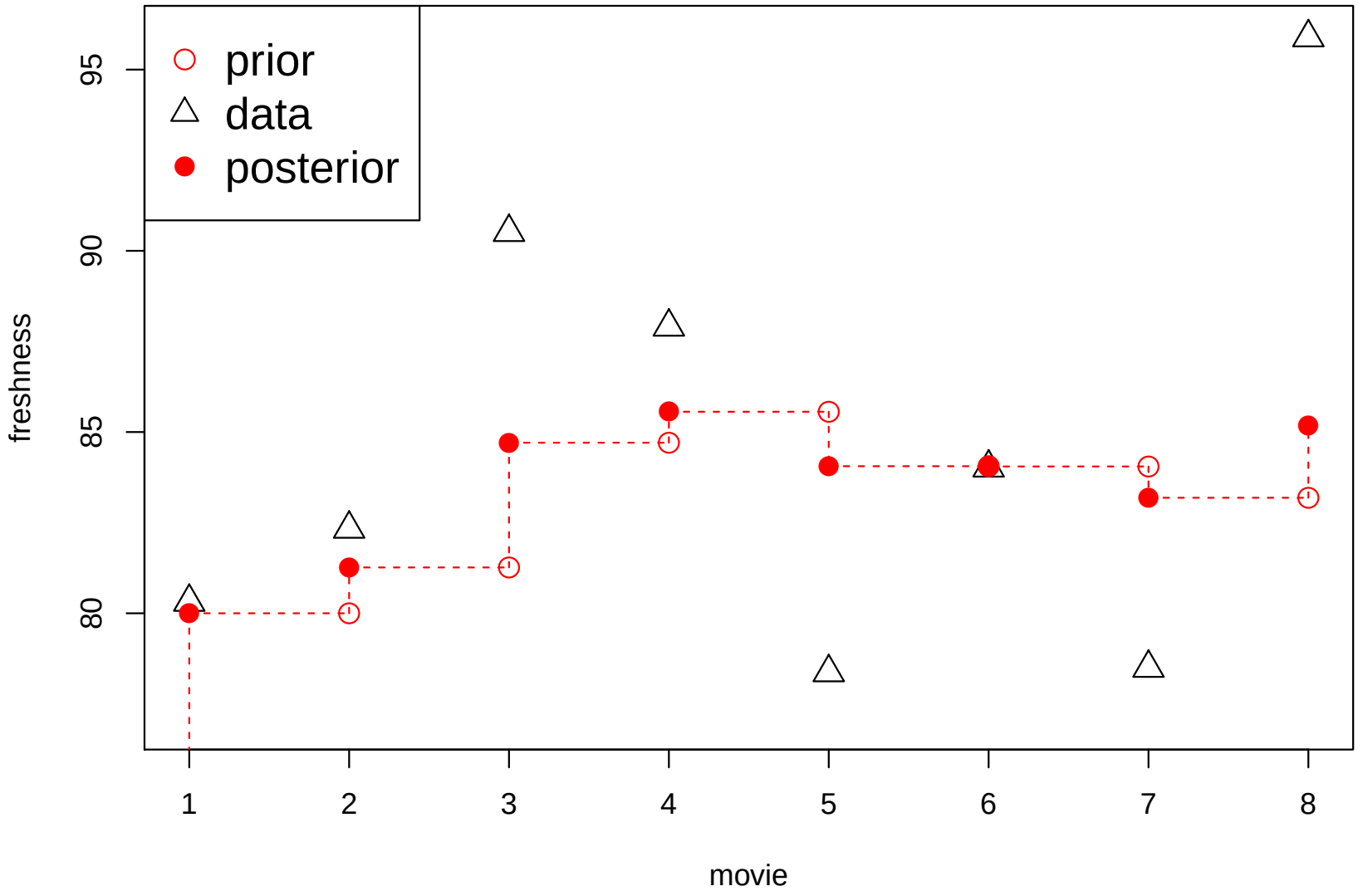
- Data

  🍅 303, 🍀 13

- Posterior distribution

  $\text{mean} = 85.2\%, \ \text{sd} = 0.8\%$

The overall Harry Potter movie series



Monte Carlo approximation

- For complicate problems, posterior $p(\theta \mid D)$ doesn't look like a distribution we know.
- Draw many samples from the posterior

$$\theta^{(1)}, \ldots, \theta^{(N)} \overset{\text{iid}}{\sim} p(\theta \mid D)$$

Monte Carlo approximation

- For complicate problems, posterior $p(\theta \mid D)$ doesn't look like a distribution we know.
- Draw many samples from the posterior

$$\theta^{(1)}, \ldots, \theta^{(N)} \overset{\text{iid}}{\sim} p(\theta \mid D)$$

- Use sample mean to approximate posterior mean

$$E(\theta \mid Y) \approx \frac{1}{N}\left[\theta^{(1)} + \theta^{(2)} + \cdots + \theta^{(N)}\right]$$

Markov chain Monte Carlo (MCMC)

- Drawing independent samples can be hard...
- We can draw samples sequentially:
  next sample's value depends on the current sample's value (Markov property)

$$\theta^{(1)} \longrightarrow \theta^{(2)} \longrightarrow \cdots \longrightarrow \theta^{(N)}$$

Markov chain Monte Carlo (MCMC)

- Drawing independent samples can be hard...
- We can draw samples sequentially:
  next sample's value depends on the current sample's value (Markov property)

$$\theta^{(1)} \longrightarrow \theta^{(2)} \longrightarrow \cdots \longrightarrow \theta^{(N)}$$

- If the number of iteration N large enough,
  - Use MCMC samples' mean to approximate the posterior mean $E(\theta \mid D)$
  - Starting point $\theta^{(1)}$ doesn't matter

Hypothesis testing

- Suppose we want to test whether the coin is fair or not. What are the hypotheses?

Hypothesis testing

- Suppose we want to test whether the coin is fair or not. What are the hypotheses?

- Null hypothesis

$$H_0 : \theta = 0.5$$

  Alternative hypothesis

$$H_1 : \theta \neq 0.5$$

Hypothesis testing

- Suppose we want to test whether the coin is fair or not. What are the hypotheses?
- Null hypothesis

$$H_0 : \theta = 0.5$$

  Alternative hypothesis

$$H_1 : \theta \neq 0.5$$

- Data: in 100 flips, we got 55 heads and 45 tails

P-value: the frequentist (classic) way

$$p\text{-value} = 0.31$$

P-value: the frequentist (classic) way

$$p\text{-value} = 0.31$$

- Reject the null if the p-value is small (< 0.05).

P-value: the frequentist (classic) way

$$p\text{-value} = 0.31$$

- Reject the null if the p-value is small (< 0.05).
- What's the meaning of the p-value?

P-value: the frequentist (classic) way

$$p\text{-value} = 0.31$$

- Reject the null if the p-value is small (< 0.05).
- What's the meaning of the p-value?

$$P(\text{current data or more extreme} \mid H_0) = 0.31$$

Bayesian hypothesis testing

- Equal prior probabilities

$$P(H_0) = P(H_1) = 0.5$$

Bayesian hypothesis testing

- Equal prior probabilities

$$P(H_0) = P(H_1) = 0.5$$

- How likely to get current data

$$P(D \mid H_0) = 0.048, \quad P(D \mid H_1) = 0.010$$

Bayesian hypothesis testing

- Equal prior probabilities

$$P(H_0) = P(H_1) = 0.5$$

- How likely to get current data

$$P(D \mid H_0) = 0.048, \quad P(D \mid H_1) = 0.010$$

- Use the Bayes rule to get posterior probabilities

$$P(H_0 \mid D) = \frac{p(D \mid H_0)P(H_0)}{p(D \mid H_0)P(H_0) + p(D \mid H_1)P(H_1)}$$

Bayesian hypothesis testing

- Equal prior probabilities

$$P(H_0) = P(H_1) = 0.5$$

- How likely to get current data

$$P(D \mid H_0) = 0.048, \quad P(D \mid H_1) = 0.010$$

- Use the Bayes rule to get posterior probabilities

$$P(H_0 \mid D) = 0.83$$
$$P(H_1 \mid D) = 0.17$$

Bayesian hypothesis testing

- Equal prior probabilities

$$P(H_0) = P(H_1) = 0.5$$

- How likely to get current data

$$P(D \mid H_0) = 0.048, \quad P(D \mid H_1) = 0.010$$

- Use the Bayes rule to get posterior probabilities

$$P(H_0 \mid D) = 0.83$$
$$P(H_1 \mid D) = 0.17$$

- Favor H_0 if $P(H_0 \mid D) > P(H_1 \mid D)$

References and resources

- Textbook of my Bayesian intro class:
  Hoff (2010) A First Course in Bayesian Statistical Methods
- Another good reference book:
  Gelman, *et al* (2013) Bayesian Data Analysis, 3rd edition
- A textbook on MCMC:
  Robert and Casella (2004) Monte Carlo Statistical Methods, 2nd edition

- Feel free to check out my teaching slides
  https://github.com/yingboli/MyTeachingSlides