Introduction to Bayesian Statistics with R

Day 1

Oliver Lindemann

Erasmus University Rotterdam, NL

- Dr. Oliver Lindemann
  - Email: lindemann@essb.eur.nl

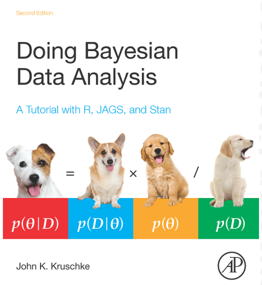
- Website: `http://bit.ly/bayes2018`

Key readings

Kruschke, John K.(2015) *Doing Bayesian Data Analysis* (2nd Edition), Academic Press, Amsterdam.

see `http://bit.ly/2D6vGPR`

Further literature will be provide during the course



Overview

Day 1 Basics
- Motivation, Probabilites, Distributions, Bayes Theorem

Day 2 Bayesian Hypothesis Test, the Easy Way
- Bayes factor, JASP, BayesFactor R package

Day 3 Under the Hood
- Estimating posterior distribution, MCMC Sampling, JAGS

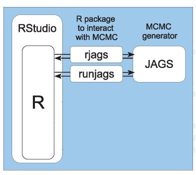
Day 4 Bayesian Modelling
- Modeling with JAGS, Multidimensional space, Hierarchical models

Day 5 Applications
- Hypothesis testing, Parameter Estimation, Linear Regressions, Model Comparisons

---

Technical Requirements

- preferable: own laptop
- Only free software
- Bayesian Modelling
  - **R 3.1**
  - **RStudio OpenSource Edition**
  - **JAGS 3.4**
- "SPSS"-like statistical software package
  - Jasp



Links

R	`http://mirrors.softliste.de/cran/`
RStudio	`http://www.rstudio.com/products/rstudio/download/`
JAGS	`http://sourceforge.net/projects/mcmc-jags/files/JAGS/3.x`
Jasp	`https://jasp-stats.org/`

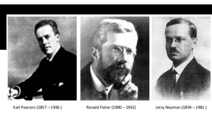
Introduction

Why Inferential Statistics?

Procedure

Theory → **Hypothesis** → Experiment → **Data** → Statistics

Classical statistics



- Ronald Fischer:
  - terms "null-hypothesis" & "significant"
  - urged the distinction between sample and population
  - degrees of freedom
  - suggested $p < .05$
  - random assignment of conditions, random sampling
- Neyman and Pearson:
  - Power and Type II error
  - effect size
  - **Formal decision rule**
    - following this rule, in the long run, we will not be often wrong
    - error rate (α) of the decision process

Interpreting Frequentist Inferential Statistics

What does the p-value tell us?

$$p(\text{Data}|H_0)$$

And what do we what to know from the data?

$$p(H_1|\text{Data})$$

But

$$p(H_1|\text{Data}) \neq p(\text{Data}|H_0)$$

Thus, p does NOT tell us anything about the likelihood of the hypothesis, neither H_1 nor H_0!

Misconceptions for Frequentist Statistics

- $p < .05$ means that H_0 is unlikely to be true, and can be rejected.
- $p > .10$ means that H_0 is likely to be true.
- For a given parameter μ, a 95% confidence interval from, say, a to b means that there is a 95% chance that μ lies in between a and b.

p-values and strength of evidence

Neyman-Pearson approach: p-values as only interpretable as binary decision rule (effect or not)

- Why can't we use p-values as measure of evidence? Why is a small p-value not more evidence for H_1?

If H_0 then NOT R
R
$\Rightarrow$ Not H_0

(valid)

If H_0 then probably NOT R
R
$\Rightarrow$ Probably not H_0

(invalid!)

Example: Probabilistic logic

If H_0 then probably NOT R
R
$\Rightarrow$ Probably not H_0

If a person is German, she is probably not member of the Bundestag
The person a member of the Bundestag
$\Rightarrow$ Person is probably not German

Bayesian Inference in a Nutshell

In Bayesian inference

- Uncertainty or degree of belief (in a hypothesis) is quantified by probability.
- **Prior** beliefs are updated by means of the data to yield **posterior** beliefs.

Thomas Bayes (1701 1761)

Bayes Theorem

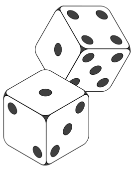
$$p(H|D) = \frac{p(D|H)\,p(H)}{p(D)}$$

Aims of this Course

- Recent developments in behavioral statistics
  - critical view on frequentist interference (i.e., *classical* statistics)
- Basic principles of Bayes statistics
  - understanding the some mathematical principles
  - Markov-chain-Monte-Carlo (MCMC) sampling
- **Hands on experiences in conducting Bayesian analyses**
  - Using R and the sampling software JAGS (later more)

Probabilities

Random Events

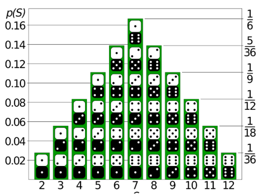


Single dice

$$S = \{1, 2, 3, 4, 5, 6\}$$
$$p(S) = 1/6$$

Two dices

$$S = \{2, \ldots, 12\}$$



Probabilities

Probabilities assign numbers to possibilities

A probability $p(x_i)$ have to satisfy three properties:

$$0 \leq p(x_i) \leq 1 \tag{1}$$
$$\sum_i p(x_i) = 1 \tag{2}$$
$$p(A \text{ or } B) = p(A) + p(B); \quad A \neq B \tag{3}$$

Example dice:

$$p(x \in \{5, 6\}) = p(x = 5) + p(x = 6) = 1/6 + 1/6 = 1/3$$

Types of Probabilities

1. Fairness, **objective property** of the coin
   - $\theta = p(H)$: probability of head
   - e.g., fair coin: $\theta = 0.5$
2. Degree of **subjective belief** about the fairness of a particular coin
   - $p(\theta)$: Probability about a particular θ
   - e.g. $p(\theta = 0.5) = 0.95$

Note

The probability of an event, θ, is a state of the world.
The probability of a coin bias, $p(\theta)$, is merely "inside our heads".

Probability in Statistics

Frequentist approach
Objective Probability

- fact about the property in the world
- independently of our beliefs
- can be observed
- reference class or collective
- e.g. mean frequency of head in 10 coin flips

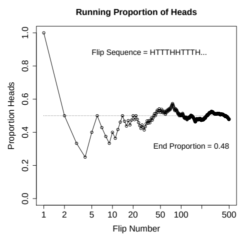
Bayesian approach
Subjective Probability

- degree of conviction in a belief
- state of knowledge
- "in the mind"
- single event
- e.g. Probability that it snows today?

States of the World

How to determine the probability of an event?

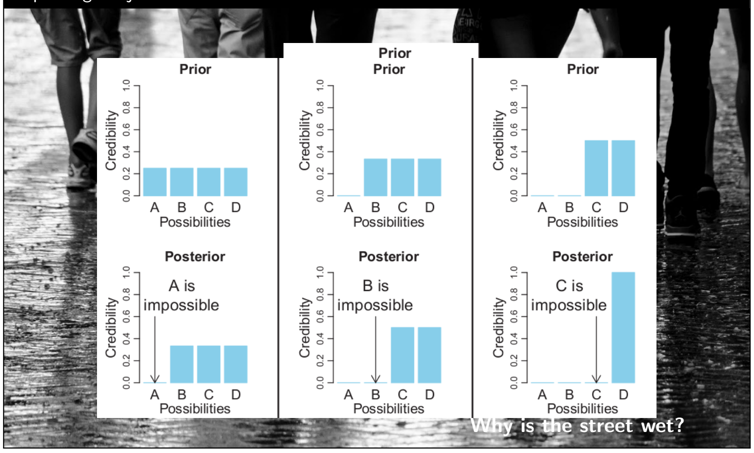
1. deriving mathematically
   - Example of the dice
   - not always possible
2. Simulation long-run relative frequencies
   - Example N coin flips
   - count head H
   - $\hat{\theta} = \theta = H/N$



Core of Bayesian Statistics

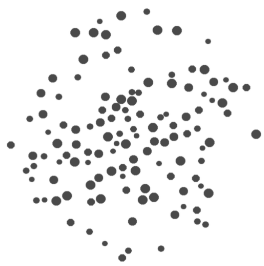
Updating Beliefs

Updating Subjective Prior Beliefs



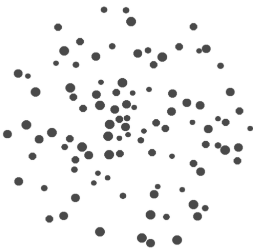
Subjective Beliefs about Continuous Parameter

How many dots?



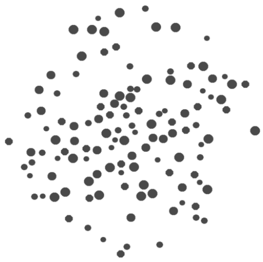
Subjective Beliefs about Continuous Parameter

100 dots

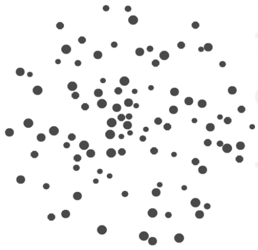


Subjective Beliefs about Continuous Parameter

How many dots?



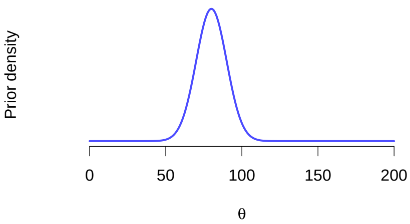
100 dots



Subjective Beliefs about Continuous Parameter

Mathematical functions describe our beliefs:

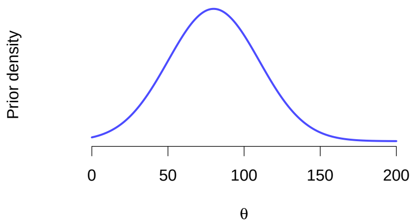
$\sim$ 80 dots



Subjective Beliefs about Continuous Parameter

Mathematical functions describe our beliefs:

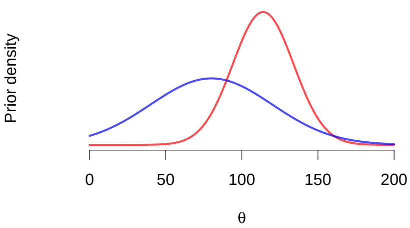
$\sim$ 80 dots



Subjective Beliefs about Continuous Parameter

Mathematical functions describe our beliefs:

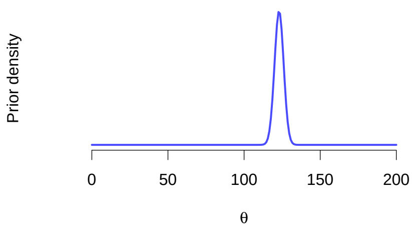
updated belief after getting baseline: $\sim$ 114 dots



Subjective Beliefs about Continuous Parameter

Mathematical functions describe our beliefs:

123 dots

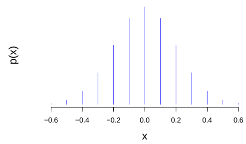


Distributions

Probability Distributions

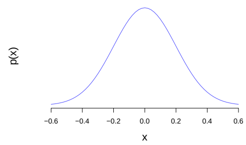
Discrete Distributions

- Probability mass function (PMF)
- *Wahrscheinlichkeitsfunktion*
- $\sum_i p(x_i) = 1$



Continuous Distributions

- Probability density functions (PDF)
- *Wahrscheinlichkeitsdichtefunktion*
- $\int_i p(x_i) = 1$



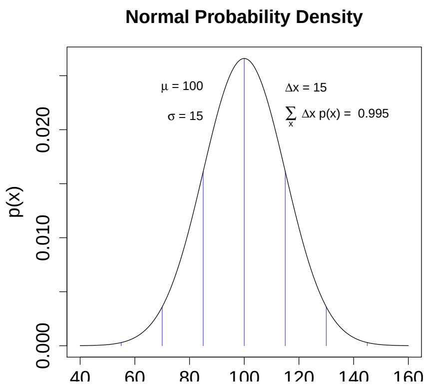
Normal Probability Density

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left[\frac{x-\mu}{\sigma}\right]^2 \right)$$

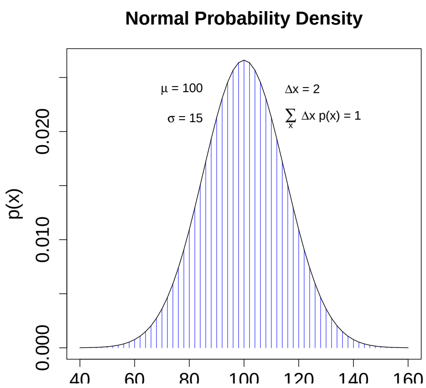
- mean: μ
- standard deviation: σ

```
# R code
px = dnorm(x, mean = 0, sd = 0.2)
```


Normal Probability Density



Normal Probability Density



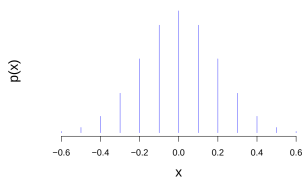
Properties of Probability Mass Functions

Mean

$$E[x] = \sum_x p(x)x$$

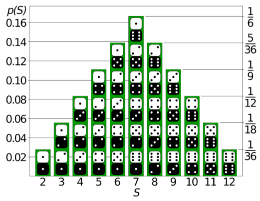
Standard deviation

$$\mathrm{var}_x = \sum_x p(x)(x - E[x])$$



Properties of Probability Mass Functions

Example two dices



Mean

$$E[x] = \sum_x p(x)x$$

$$E = 2 \cdot \frac{1}{36} \quad + 3 \cdot \frac{1}{18} \ldots + 11 \cdot \frac{1}{18} + 12 \cdot \frac{1}{36}$$
$$= 7$$

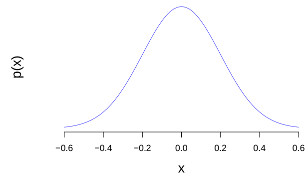
Properties of Probability Density Functions

Mean

$$E[x] = \int_x p(x)x$$

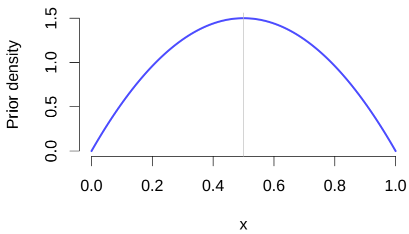
Standard deviation

$$\mathrm{var}_x = \int_x p(x)(x - E[x])$$



Mean of Continuous Distribution

example PDF: $p(x) = 6x(1 - x)$



Mean?

Example Mean of Continuous Distribution

$$\begin{aligned}
E(x) &= \int_0^1 dx\ p(x)x \\
&= \int_0^1 dx\ 6x(1-x)x \\
&= 6\int_0^1 dx\ (x^2 - x^3) \\
&= 6\left[\frac{1}{3}x^3 - \frac{1}{4}x^4\right]_0^1 \\
&= 6\left[\left(\frac{1}{3}1^3 - \frac{1}{4}1^4\right) - \left(\frac{1}{3}0^3 - \frac{1}{4}0^4\right)\right] \\
&= 0.5
\end{aligned}$$

Conditional Probabilities

Conjoint, Marginal & Conditional Probability

	Hair Color				
Eye Color	Black	Brunette	Blond	Red	
Blue	.03	.14	.16	.03	*.36*
Brown	.12	.20	.01	.04	*.37*
Green	.03	.14	.04	.05	*.27*
	.18	*.48*	*.21*	*.12*	

Proportion of sample of University of Delaware students 1974, N=592. Data adapted from Snee (1974).

Conjoint Probabilities: $p(E, H)$

	Hair Color				
Eye Color	Black	Brunette	Blond	Red	
Blue	.03	.14	.16	.03	*.36*
Brown	.12	.20	.01	.04	*.37*
Green	.03	.14	.04	.05	*.27*
	.18	*.48*	*.21*	*.12*	

Conjoint probabilities: $p(E, H)$

For example,
$p(E = \text{blue}, H = \text{black}) = .03$

Marginal Probabilities: $p(E)$

	Hair Color				
Eye Color	Black	Brunette	Blond	Red	
Blue	.03	.14	.16	.03	*.36*
Brown	.12	.20	.01	.04	*.37*
Green	.03	.14	.04	.05	*.27*
	.18	*.48*	*.21*	*.12*	

For example
$p(E = \text{blue}) = \sum_H p(E = \text{blue}, H) = .36$

Marginal Probabilities: $p(H)$

	Hair Color				
Eye Color	Black	Brunette	Blond	Red	
Blue	.03	.14	.16	.03	*.36*
Brown	.12	.20	.01	.04	*.37*
Green	.03	.14	.04	.05	*.27*
	.18	*.48*	*.21*	*.12*	

For example
$p(H = \text{black}) = \sum_H p(E, H = \text{black}) = .18$ $\quad p(H)$ **without** information about E

Conjoint, Marginal & Conditional Probability

Marginal probability

$$p(A) = \sum_i p(A, B_i)$$

Conditional probability

$$p(A|B) := \frac{p(A, B)}{p(B)}$$

Conditional Probabilities: $p(H|E = \text{blue})$

	Hair Color				
Eye Color	Black	Brunette	Blond	Red	
Blue	.03	.14	.16	.03	*.36*
Brown	.12	.20	.01	.04	*.37*
Green	.03	.14	.04	.05	*.27*
	.18	*.48*	*.21*	*.12*	

$p(H|E = \text{blue})$ is
$p(H)$ **with** information that $E = \text{blue}$

Conditional Probabilities: $p(H|E = \text{blue})$

	Hair Color				
Eye Color	Black	Brunette	Blond	Red	
Blue	.03	.14	.16	.03	*.36*
Brown	.12	.20	.01	.04	*.37*
Green	.03	.14	.04	.05	*.27*
	.18	*.48*	*.21*	*.12*	

$$p(H|E = \text{blue}) = \frac{p(H, E = \text{blue})}{p(E = \text{blue})}$$

	Black	Brunette	Blond	Red	
Blue	$.03/.36 =$ .08	$.14/.36 =$ .39	$.16/.36 =$ .45	$.03/.36 =$ .08	$.36/.36 =$ *1*

Conditional Probabilities: $p(H|E = \text{blue})$

	Hair Color				
Eye Color	Black	Brunette	Blond	Red	
Blue	.03	.14	.16	.03	*.36*
Brown	.12	.20	.01	.04	*.37*
Green	.03	.14	.04	.05	*.27*
	.18	*.48*	*.21*	*.12*	

$$p(H|E = \text{brown}) = \frac{p(H, E = \text{brown})}{p(E = \text{brown})}$$

	Black	Brunette	Blond	Red	
Brown	$.12/.37 =$ .32	$.20/.37 =$ .54	$.01/.37 =$ .03	$.04/.37 =$ .11	$.37/.37 =$ *1*

Bayes Theorem

Interrelation of conditional probabilities

Bayes Theorem Derivation

$$p(A|B) = \frac{p(A, B)}{p(B)}$$

$$p(A, B) = p(A|B) \cdot p(B) = p(B|A) \cdot p(A)$$

$$p(B|A) = \frac{p(A|B) \cdot p(B)}{p(A)}$$

Bayes Rule: Card Example

$$p(B|A) = \frac{p(A|B)\, p(B)}{p(A)}$$

Let's check:

- $p(\diamondsuit) = 1/8$, $p(A) = 4/32$
- $p(\diamondsuit|\text{A}) = 1/4$
- $p(\text{A}|\diamondsuit) = 1/8$?

$$p(\text{A}|\diamondsuit) = \frac{p(\diamondsuit|\text{A})\, p(\text{A})}{p(\diamondsuit)} = \frac{1/4 \cdot 4/32}{1/4} = \frac{1}{8}$$

Applying Bayes Rule

Example

The probability of breast cancer is **1%** for a woman at age 40 who participates in routine screening. If a woman has breast cancer, the probability is **80%** that she will have a positive test. If a woman does not have breast cancer, the probability is **9.6%** that she will also have a positive test.

A woman in this age group had a positive test in a routine screening. What is the probability that she actually has breast cancer?

$p(H^+) = .01$	probability of BC (Hypothesis)
$p(D|H^+) = .80$	probability of positive test (data) if BC
$p(D|H^-) = .096$	probability of positive test if no BC (False Positive)

Applying Bayes Rule

Example: $p(H^+) = .01$, $p(D|H^+) = .80$, $p(D|H^-) = .096$

$$p(H^+|D) = \frac{p(D|H^+)\, p(H^+)}{p(D)}$$

$$\begin{aligned} p(D) &= \sum_i p(D|H^i)\, p(H^i) = p(D|H^+)\, p(H^+) + p(D|H^-)\, p(H^-) \\ &= 0.80 \cdot 0.01 + 0.96 \cdot (1 - 0.01) = 0.008 + 0.095 \\ &= 0.103 \end{aligned}$$

$$p(H^+|D) = \frac{0.01 \cdot 0.80}{0.103} = 0.078$$

Correct: 7.8%

Excursus: Psychology of the Communication of Bayesian Information

Gigerenzer & Hoffrage (1999) presented this problem to subjects:

- Only 18% of the subjects sovled it correctly
- Subjects are poor in dealing with probability information
- **and ignore prior probability estimate for a hypothesis**, $p(H^+)$

How to communicate Bayesian information to laypersons?

- 46% subjects solved the task when presenting natural frequencies

Natural frequencies version

Out of every 1000 women at age 40 who participate in routine screening 10 have breast cancer. Of these 10 women with breast cancer, 8 will have a positive test. Of the remaining 990 women without breast cancer, 95 will still have a positive test.

Bayes Rule: Hypothesis/Beliefs & Data

$$p(H|D) = \frac{p(D|H) \cdot p(H)}{p(D)}$$

$$\underbrace{p(H|D)}_{\text{Posterior}} = \frac{\overbrace{p(D|H)}^{\text{Likelihood}} \cdot \overbrace{p(H)}^{\text{Prior}}}{\underbrace{p(D)}_{\text{Evidence}}}$$

Posterior $\propto$ Likelihood $\cdot$ Prior

Bayes Rule and Parameter Estimation

Usual scenario

We have collected data, D, in order to get information about our (psychological) model or theory, $\mathcal{M}$, with the not directly observable parameter, θ.

known: $p(D|\theta)$ unknow: $p(\theta|D)$

$$\underbrace{p(\theta|D)}_{\text{Posterior}} = \underbrace{p(D|\theta)}_{\text{Likelihood}} \underbrace{p(\theta)}_{\text{Prior}} \Big/ \underbrace{p(D)}_{\text{Evidence}}$$

Likelihood $\neq$ probability function

Why is $p(D|\theta)$ called *likelihood* function and not *probability* function?

1. $p(D|\theta)$ as **"probability function of** D**"** with fixed θ
   - $\sum_D p(D|\theta) = 1$ (discrete values)
   - $\int dD\, p(D|\theta) = 1$ (continuous values)
2. $p(D|\theta)$ as **"likelihood function of** θ**"** with fixed y
   - $\sum_\theta p(D|\theta) \neq 1$
   - $\int d\theta\, p(D|\theta) \neq 1$

In Bayesian statistics, we usually have fixed (observed) data, D, and the variable θ. We therefore call $p(D|\theta)$ *likelihood function*.

Parameter Estimation

Three Goals of Interference

1. Estimating parameter values

$$p(\theta|D)$$

2. Prediction of data values

$$\text{e.g.:} \quad p(y=1) = \sum_\theta p(y=1|\theta)\ p(\theta)$$

3. Model Comparison

$$\frac{p(D|\mathcal{M}_1)}{p(D|\mathcal{M}_2)} = \text{Bayes factor}$$

Bayes Rule and Parameter Estimation

We actually talk about the probability of a particular model, $\mathcal{M}$, and should write:

$$p(\theta, \mathcal{M}|D) = p(D|\theta, \mathcal{M})\, p(\theta, \mathcal{M}) \,/\, p(D)$$

Reformulating Denominator $p(D)$ in Bayes' rule

1. Marginal probability
$$p(A) = \sum_i p(A, B_i)$$
2. Conditional probability
$$p(A|B) = \frac{p(A,B)}{p(B)}$$
$$p(A,B) = p(A|B) \cdot p(B)$$

Marginal probability as function of conditional probabilities

$$p(D) = \sum_\theta p(D|\theta)\, p(\theta) \quad (2.4)$$

Bayes theorem for Discrete & Continuous Parameter

Bayes Rule (discrete values)

$$p(\theta|D) = \frac{p(D|\theta) \cdot p(\theta)}{\sum_\theta p(D|\theta)\, p(\theta)} \quad (2.5)$$

Bayes Rule (continuous values)

$$p(\theta|D) = \frac{p(D|\theta) \cdot p(\theta)}{\int d\theta\ p(D|\theta)\ p(\theta)} \quad (2.6)$$

Example

Coin Flip

$\underbrace{p(\theta|D)}_{\text{Posterior}} = \underbrace{p(D|\theta)}_{\text{Likelihood}}\ \underbrace{p(\theta)}_{\text{Prior}} \,/\, \underbrace{p(D)}_{\text{Evidence}}$

"Coin flip model" $\mathcal{M}$:

- $p(y=1|\theta) = \theta$
- $p(y=0|\theta) = 1-\theta$

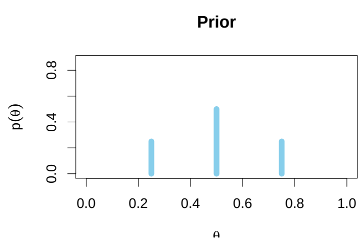
Example: Prior $p(\theta)$

$\underbrace{p(\theta|D)}_{\text{Posterior}} = \underbrace{p(D|\theta)}_{\text{Likelihood}}\ \underbrace{p(\theta)}_{\text{Prior}} \,/\, \underbrace{p(D)}_{\text{Evidence}}$

Specifying the Prior distribution $p(\theta)$

Let's assume three possibilities:

- $\theta = 0.5$: coin is fair
- $\theta = 0.25$: biased toward head
- $\theta = 0.75$: biased toward tail



Example: Data

We flip 12 times and observe 3 heads and 9 tails

- $y_i = [1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0]$
  - 1 for head
  - 0 for tail
- $z = 3, \ N = 12$

What is the likelihood to observe this data under our three priors?

Bernoulli Distribution

$$p(y|\theta) = \begin{cases} \theta, & \text{if } y = 1 \text{ (head)} \\ 1-\theta, & \text{if } y = 0 \text{ (tail)} \end{cases}$$

Bernoulli likelihood function (single coin flip)

$$p(y|\theta) = \theta^y \, (1-\theta)^{(1-y)}$$

Bernoulli Distribution (multiple coin flips)

$$\begin{aligned} p(\{y_1, \ldots, y_N\}|\theta) &= \prod_i p(y_i|\theta) \\ &= \prod_i \theta^{y_i} \, (1-\theta)^{(1-y_i)} \end{aligned}$$

Bernoulli likelihood for multiple coin flips

- number of 'heads': $z = \sum_i^N y_i$

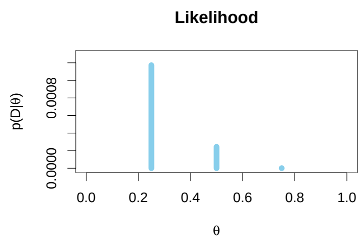
$$p(z, N|\theta) = \theta^z \, (1-\theta)^{N-z}$$

Example: Likelihood $p(D|\theta)$

What is the **likelihood**, $p(D|\theta)$, of the data?

- $z = 3, \ N = 12$

$$\begin{aligned} p(D|\theta) &= \theta^z \, (1-\theta)^{N-z} \\ &= \theta^3 \, (1-\theta)^9 \end{aligned}$$



Example: Evidence $p(D)$

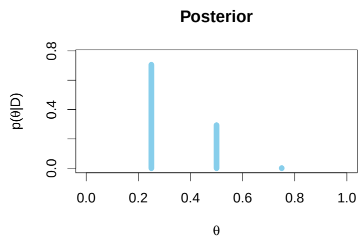
Actually, evidence for the model $\mathcal{M}$: $p(D|\mathcal{M})$

- overall probability
- averaging across all parameter weighted by our belief in them
- normalizer for the posterior distribution

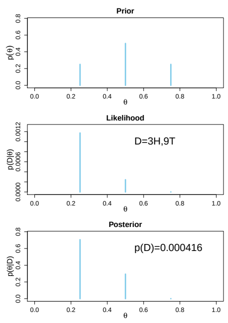
$$\begin{aligned} p(D) &= \sum_\theta p(D|\theta) \, p(\theta) \\ &= 0.000415 \end{aligned}$$

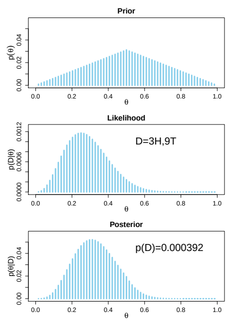
Example: posterior $p(\theta|D)$

just applying Bayes' rule:



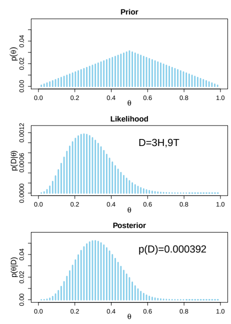
Results: Bayesian updating



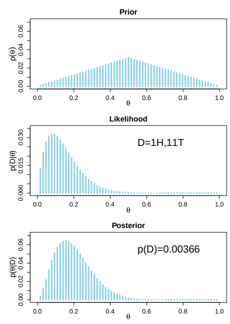


Results: Bayesian updating

Old data



More extreme data



Example: Calculation posterior in R (1)

$$\underbrace{p(\theta|D)}_{\text{Posterior}} = \underbrace{p(D|\theta)}_{\text{Likelihood}}\,\underbrace{p(\theta)}_{\text{Prior}} \,/\, \underbrace{p(D)}_{\text{Evidence}}$$

```r
# Data
Data = c(1,1,1,0,0,0,0,0,0,0,0,0) # order doesn't matter
z = sum( Data )

# Prior
Theta = c(.25, .5, .75)
prior = c(.25, .5, .25)

# Likelihood
N = length( Data )
likelihood = Theta^z * (1-Theta)^(N-z) # Bernoulli likelihood

# Posterior
evidence = sum( likelihood * prior )
posterior = likelihood * prior / evidence # Bayes' rule!

print(posterior)
```


Example: Plotting in R (2)

```r
# plot prior
plot(Theta , prior,
     type="h" , lwd=10 , main="Prior",
     xlim=c(0,1) , xlab=bquote(theta),
     ylim=c(0,1.1*max(posterior)),
              ylab=bquote(p(theta)),
              col="skyblue")

# plot Likelihood
plot( Theta , likelihood ,
     type="h" , lwd=10 , main="Likelihood",
     xlim=c(0,1) , xlab=bquote(theta),
     ylim=c(0,1.1*max(likelihood)),
     ylab=bquote(paste("p(D|",theta,")")),
     col="skyblue" )
```


Example: Plotting in R (2)

```r
# plot posterior
plot( Theta , posterior,
     type="h" , lwd=10 , main="Posterior" ,
     xlim=c(0,1) , xlab=bquote(theta),
     ylim=c(0,1.1*max(posterior)),
     ylab=bquote(paste("p(",theta,"|D)")),
     col="skyblue")
```


Thank you