2758672 Bayes Prin App

Week 1: Introduction

อ.ดร.สิวะโชติ ศรีสุทธิยากร
ภาควิชาวิจัยและจิตวิทยาการศึกษา
คณะครุศาสตร์ จุฬาลงกรณมหาวิทยาลัย

Statistics Inference

- การอนุมานเชิงสถิติ (inferential statistics) เป็นวิธีการทางสถิติที่ช่วยผู้วิเคราะห์ข้อมูล/นักวิจัย สร้างข้อสรุปเกี่ยวกับปรากฏการณ์ที่สนใจ
    - Estimation
    - Prediction
    - Model Comparison/Hypothesis Testing

กิจกรรม 1 : เหรียญเที่ยงตรงหรือไม่ ?

1. หาเหรียญอะไรก็ได้ขึ้นมา 1 เหรียญ
2. นิสิตคิดว่าเหรียญดังกล่าวมีความเที่ยงตรงในการออกหน้าหัวและก้อยหรือไม่? มั่นใจกับคำตอบนี้มากแค่ไหน?
3. นิสิตคิดว่าจะใช้วิธีการใดเพื่อพิสูจน์ความเชื่อเกี่ยวกับความเที่ยงตรงของเหรียญในข้อ 2.

The confidence interval for a population proportion, therefore, becomes:

$$p = p' \pm \left[Z_{\left(\frac{a}{2}\right)} \sqrt{\frac{p'(1-p')}{n}} \right]$$

$Z_{\left(\frac{a}{2}\right)}$ is set according to our desired degree of confidence and $\sqrt{\frac{p'(1-p')}{n}}$ is the standard deviation of the sampling distribution.

S.6 Test of Proportion

Let us consider the parameter p of population proportion. For instance, we might want to know the proportion of males within a total population of adults when we conduct a survey. A test of proportion will assess whether or not a sample from a population represents the true proportion from the entire population.

Critical Value Approach

The steps to perform a test of proportion using the critical value approval are as follows:

1. State the null hypothesis H_0 and the alternative hypothesis H_A.
2. Calculate the test statistic:

   $$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

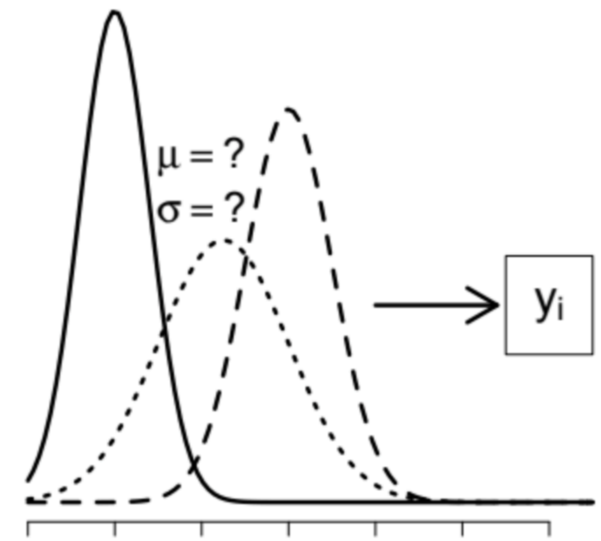
   where p_0 is the null hypothesized proportion i.e., when $H_0 : p = p_0$
3. Determine the critical region.
4. Make a decision. Determine if the test statistic falls in the critical region. If it does, reject the null hypothesis. If it does not, do not reject the null hypothesis.

https://openstax.org/books/introductory-business-statistics/pages/8-3-a-confidence-interval-for-a-population-proportion

https://online.stat.psu.edu/statprogram/reviews/statistical-concepts/proportions

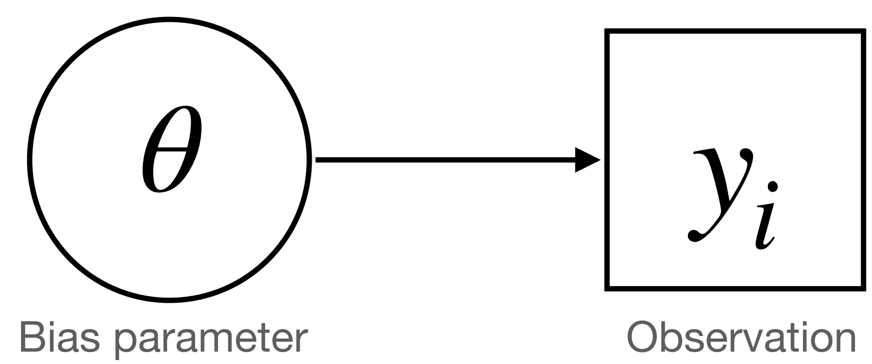
Mathematical Models

นักวิจัยใช้ math models เพื่อเป็นแบบจำลองของปรากฏการณ์ที่สนใจ โมเดลดังกล่าวใช้เลียนแบบกระบวนการสร้างค่าสังเกตของตัวแปรตามในประชากรที่ทำการศึกษา



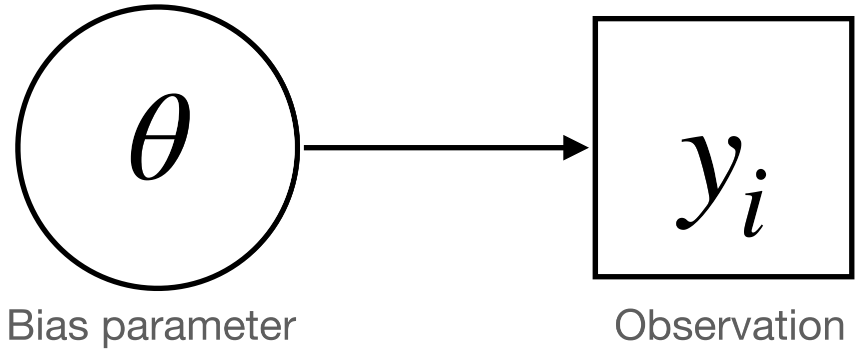
- Deterministic models
- **Statistical models** —-> probability models

  - สมมุติให้เหรียญมีคุณลักษณะที่เรียกว่า**ความลำเอียง (bias)** ที่ไม่สามารถสังเกตได้โดยตรง โดยเป็นคุณลักษณะที่อยู่เบื้องหลังและเป็นตัวกำหนดโอกาสของการออกหน้าหัวและก้อย
  - ความลำเอียงดังกล่าวคงที่ในแต่ละการโยน
  - การโยนแต่ละครั้งเป็นอิสระซึ่งกันและกัน
  - การทราบค่าความลำเอียงของเหรียญจะช่วยให้สร้างข้อสรุปได้ว่าเหรียญดังกล่าวมีความเที่ยงตรงหรือไม่ อย่างไร?



โดยที่ $y_i = 0,\ 1$

Mathematical Models



โดยที่ $y_i = 0,\ 1$

หากกำหนดให้ความลำเอียง มีค่าเท่ากับความน่าจะเป็นที่เหรียญจะออกหัว

กล่าวคือ $P(y_i = 1) = \theta$ และในทำนองเดียวกันจะได้ว่า $P(y_i = 0) = 1 - \theta$ โดยที่ $\theta \in [0,1]$

จะได้**โมเดลของค่าสังเกต (model of observation)** ในการโยนเหรียญเป็น

$$P(y_i | \theta) = \theta^{y_i}(1-\theta)^{1-y_i}$$

$\implies p(\mathbf{y} | \theta) = \theta^{\sum_{i=1}^{n} y_i}(1-\theta)^{1-\sum_{i=1}^{n} y_i}$ **<— เรียกฟังก์ชันนี้ว่า likelihood function**

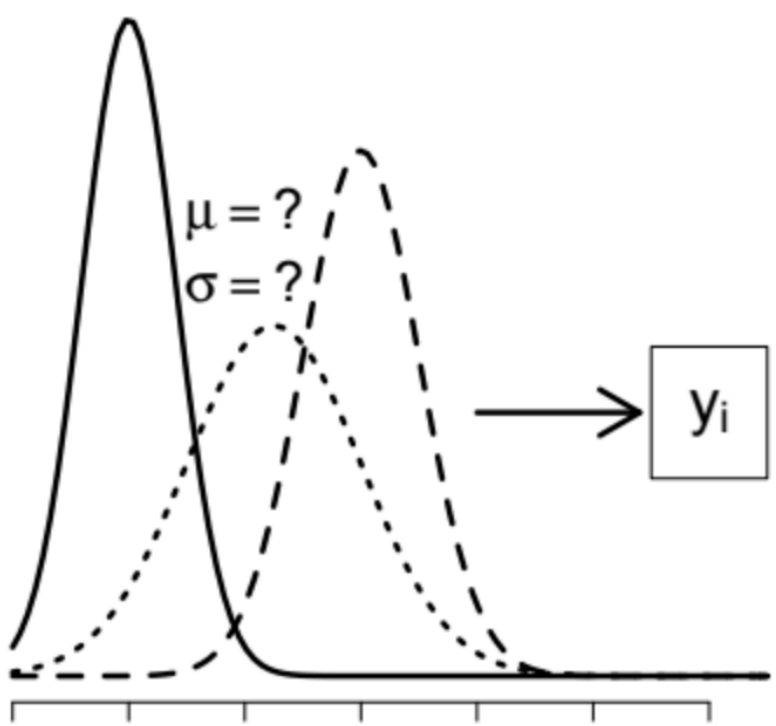
ในทางสถิติจะเรียกคุณลักษณะความลำเอียงภายในโมเดลว่า **พารามิเตอร์ (parameter)**

- นิสิตจะเห็นว่าหากทราบค่าของพารามิเตอร์ดังกล่าว ก็จะสามารถอธิบายแนวโน้มการเกิดหน้าหัวหรือก้อยของเหรียญนี้ได้ และสามารถอนุมานเกี่ยวกับความเที่ยงตรงของเหรียญได้
- หากโมเดลมีความเหมาะสม พารามิเตอร์จึงเป็นส่วนสำคัญในการหาคำตอบหรือสร้างข้อสรุปเกี่ยวกับปรากฏการณ์ที่ต้องการศึกษา

- ปัจจุบันคะแนนเฉลี่ยรวม 3 วิชา (Math, Sci, Eng) ของนักเรียนเป็นอย่างไร?
  - **โมเดลของค่าสังเกต (observation model)**

$$y_i = \mu + \epsilon_i \quad ; \; \epsilon_i \sim N(0, \sigma^2) \implies y_i \sim N(\mu, \sigma^2)$$

$$\implies p(y_i | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\{-\frac{1}{2\sigma^2}(y_i - \mu)^2\}$$



- **ฟังก์ชันภาวะความควรจะเป็น (likelihood function)**

ฟังก์ชันภาวะความควรจะเป็น เป็นเครื่องมือในทางสถิติที่สามารถใช้หาค่าพารามิเตอร์ที่เหมาะสมของโมเดลนั้นได้ โดยอิงจากระดับความสอดคล้องเชิงประจักษ์ระหว่างโมเดลกับข้อมูล นอกจากนี้ยังสามารถใช้ฟังก์ชันภาวะความควรจะเป็นในการอนุมานเชิงสถิติอื่น ๆ ได้แก่ การทดสอบสมมุติฐาน และการเปรียบเทียบโมเดลได้อีกด้วย

ฟังก์ชันภาวะความควรจะเป็นหาได้จากผลคูณระหว่างฟังก์ชันความน่าจะเป็นของค่าสังเกต เขียนแทนด้วย $p(\mathbf{y}|\theta) = \Pi_{i=1}^{n} p(y_i|\theta)$

$$p(\mathbf{y}|\theta) = \Pi_{i=1}^{n} p(y_i|\mu, \sigma^2) = (\frac{1}{\sqrt{2\pi\sigma^2}})^n exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\}$$

ในเชิงปฏิบัติเราสุ่มตัวอย่าง $\mathbf{y}$ มาจากประชากร แปลว่าข้อมูลตัวอย่างเป็นส่วนที่คงที่แล้ว ในขณะที่พารามิเตอร์ที่เหมาะสมของโมเดลเป็นส่วนที่ต้องพิจารณา จึงเขียนสัญลักษณ์ของฟังก์ชันภาวะความควรจะเป็นใหม่เป็น $L(\theta|\mathbf{y})$

- **Maximum Likelihood Estimation**

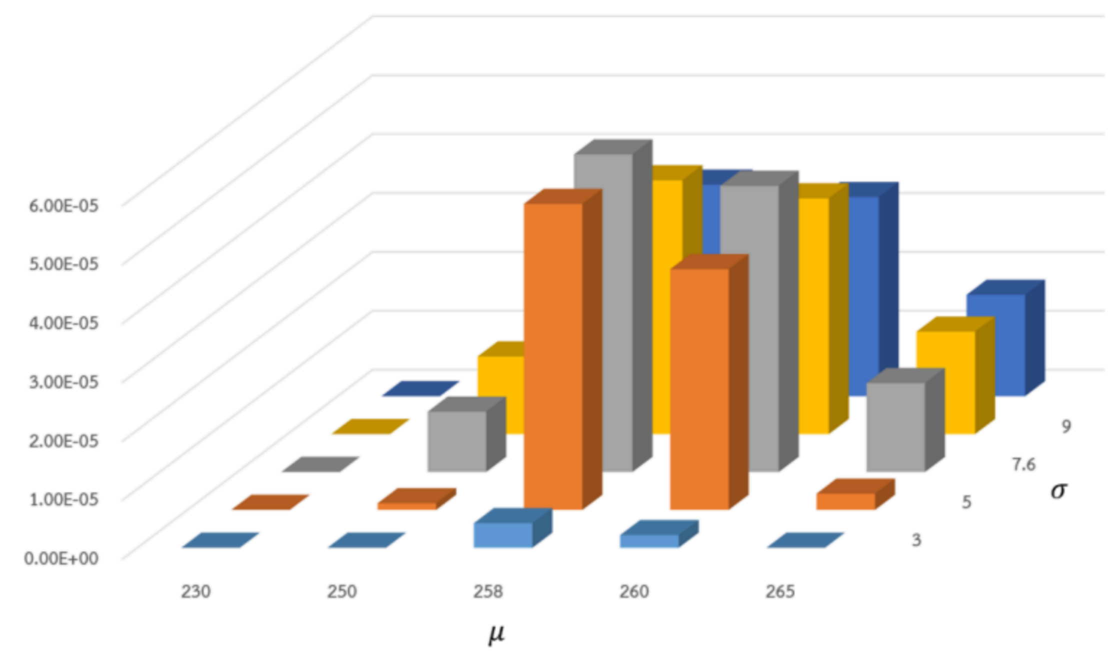
การพิจารณาเลือกค่าพารามิเตอร์ของโมเดลที่เหมาะสมโดยอิงกับฟังก์ชันภาวะความควรจะเป็น จะเลือกค่าพารามิเตอร์ที่ทำให้ค่าของฟังก์ชันภาวะความควรจะเป็นมีค่าสูงที่สุด สมมุติว่านักวิจัยสุ่มตัวอย่างขนาด 3 หน่วยจากประชากร

$$\mathbf{y} = \{y_1 = 250,\ y_2 = 265,\ y_3 = 259\}$$

$$L(\theta|\mathbf{y}) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\}$$

- ค่าประมาณของพารามิเตอร์ค่าเฉลี่ยควรมีค่าเท่ากับเท่าไหร่?
- ค่าประมาณของพารามิเตอร์ความแปรปรวนควรมีค่าเท่ากับเท่าไหร่?

σ \ μ	230	250	258	260	265
3	7.46E-63	9.74E-11	4.18E-06	2.14E-06	1.19E-09
5	1.93E-25	1.12E-06	5.20E-05	4.09E-05	2.75E-06
7.6	7.75E-14	1.02E-05	5.39E-05	4.86E-05	1.51E-05
9	2.13E-11	1.32E-05	4.31E-05	4.00E-05	1.74E-05
10	2.80E-10	1.37E-05	3.59E-05	3.38E-05	1.72E-05



กิจกรรม 2 : Prior distribution

ในกิจกรรมที่ 1 ข้อ 2 นิสิตได้คาดการณ์ความเที่ยงตรงของเหรียญผ่านการสังเกตและประสบการณ์ที่สะสมมาแล้ว ผลจากการคาดการณ์ดังกล่าวสามารถนำมาให้น้ำหนักของแต่ละความเป็นไปได้ของพารามิเตอร์ความลำเอียงนี้ได้อย่างไร

Theta	Weight —> p(Theta)
0.2	
0.5	
0.8	
รวม	1.00

การแจกแจงของความลำเอียงในตารางข้างต้น เป็นความเชื่อที่ได้จากการสังเกต/ประสบการณ์ หรือความรู้ที่ผ่านมาของนิสิต และไม่ได้เกี่ยวข้องกับข้อมูลเชิงประจักษ์ใด ๆ เราเรียกความเชื่อนี้ว่า **ความเชื่อก่อนหน้า (prior beliefs)** และการแจกแจงความน่าจะเป็นของพารามิเตอร์ที่สอดคล้องกับความเชื่อดังกล่าว จะเรียกว่า **การแจกแจงความน่าจะเป็นก่อนหน้า (prior distribution)**

กิจกรรม 3 : Posterior Distribution

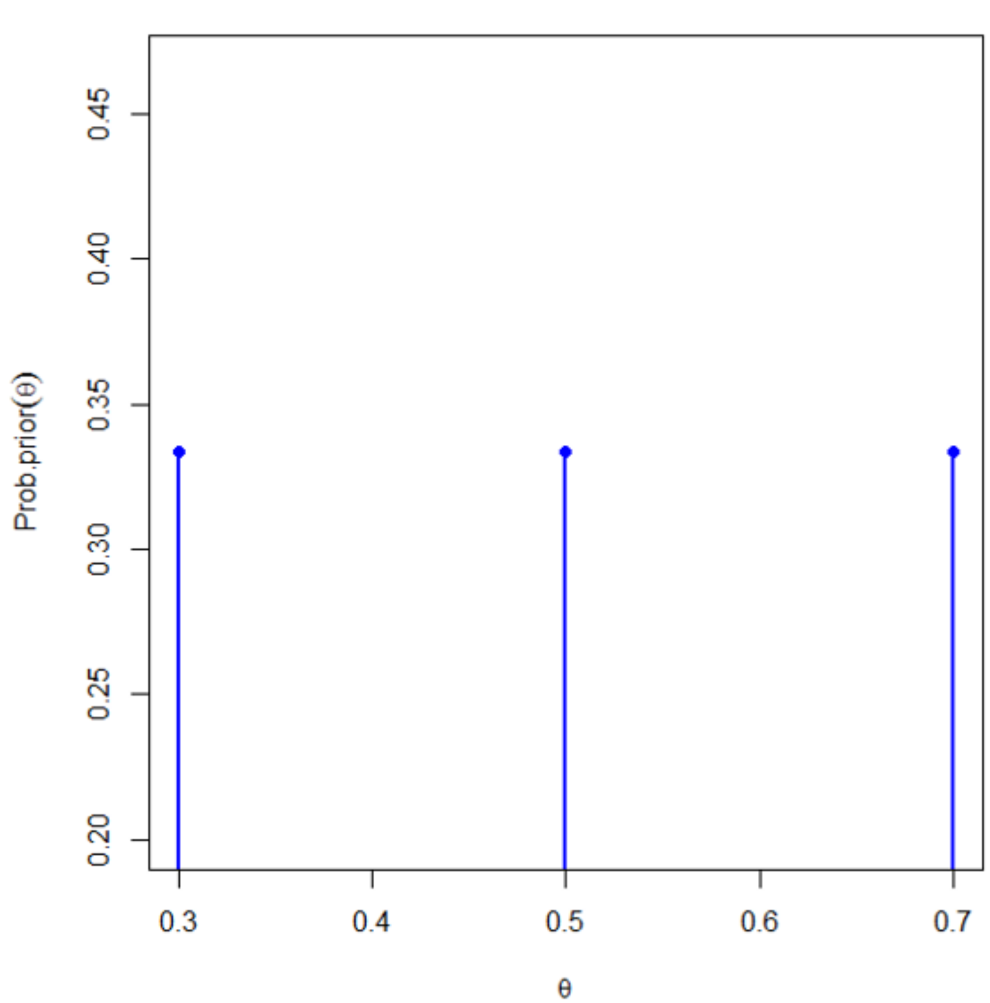
1. ทดลองโยนเหรียญดังกล่าว 10 ครั้ง บันทึกผลลัพธ์ที่ได้

ผลลัพธ์	**จำนวน**
H	
T	
รวม	**10**

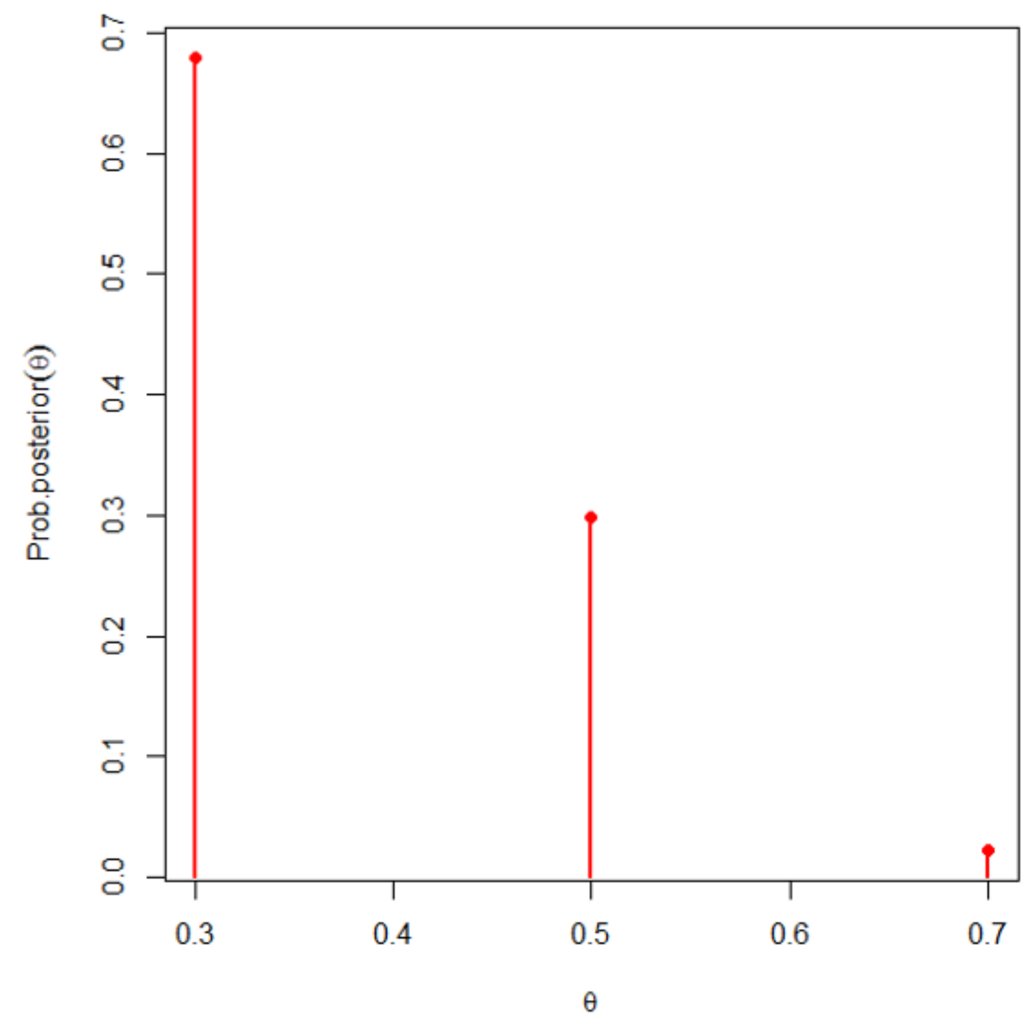
2. จากผลลัพธ์ดังกล่าว ความเชื่อเกี่ยวกับความลำเอียงของเหรียญยังคงเดิมหรือไม่ หรือควรเปลี่ยนแปลงไปอย่างไร?

Theta	**Weight —> p(Theta|Data)**
0.2	
0.5	
0.8	
รวม	1.00

ความเชื่อที่ถูกปรับปรุงใหม่เมื่อมีข้อมูลเชิงประจักษ์เพิ่มเติมจะเรียกว่า **ความเชื่อภายหลัง (posterior belief)** การแจกแจงความน่าจะเป็นที่สอดคล้องกับความเชื่อดังกล่าว จะเรียกว่า **การแจกแจงความน่าจะเป็นภายหลัง (posterior distribution)**



Prior distribution



Posterior distribution

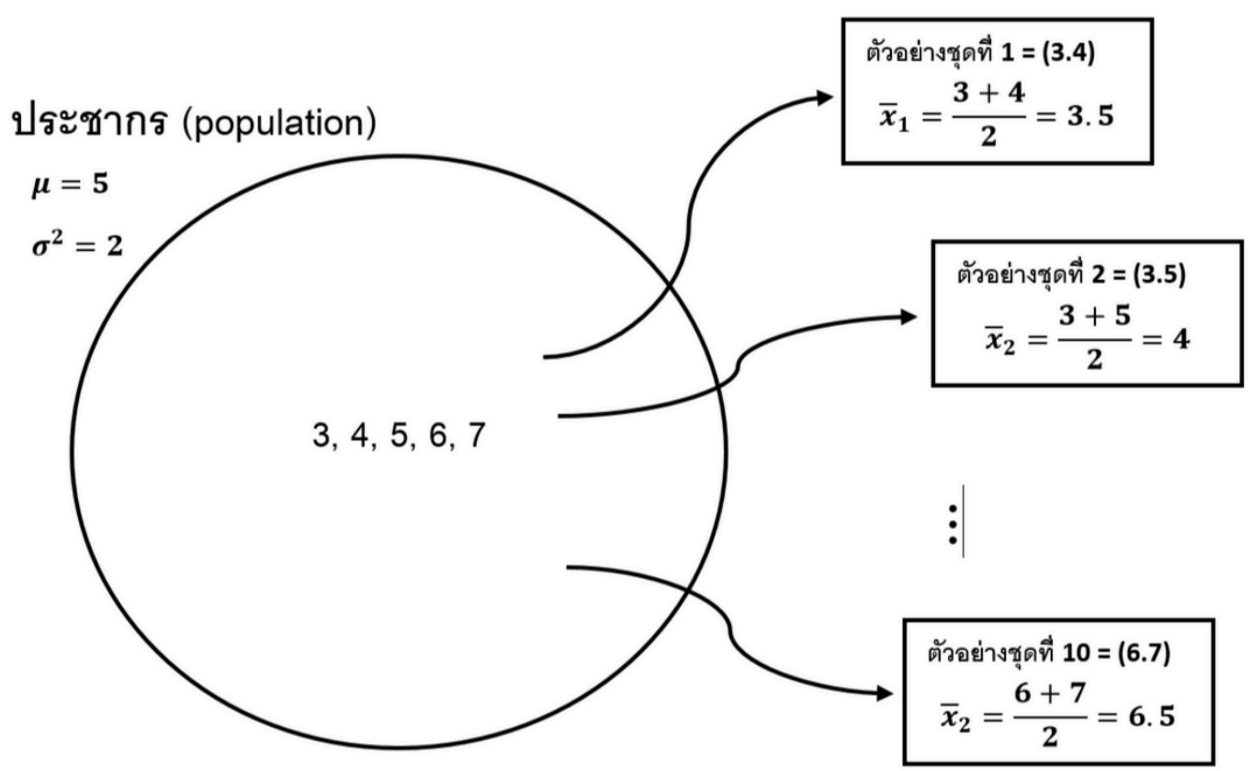
(given n=10 and H=3)

- การจัดสรรน้ำหนักให้กับแต่ละความเป็นไปได้ของความลำเอียงเมื่อมีข้อมูลจริงจากการโยนเหรียญเพิ่มเติม ควรทำอย่างไร?
- การทราบการแจกแจงความน่าจะเป็นภายหลัง ทำให้ผู้วิเคราะห์ทราบ (1) ขอบเขตความเป็นไปได้ของพารามิเตอร์ (2) แนวโน้ม/น้ำหนักความน่าเชื่อถือในแต่ละความเป็นไปได้ สารสนเทศนี้ทำให้เราสามารถอนุมานเกี่ยวกับพารามิเตอร์ได้อย่างมีประสิทธิภาพ

Why Bayes?

Frequentist Statistics

- พารามิเตอร์เป็นค่าคงที่ ที่มีอยู่จริงในประชากร แต่ไม่ทราบค่า
- ข้อมูลตัวอย่าง (samples) เป็นหนึ่งในผลลัพธ์ที่สุ่มได้จากประชากร โดยผลลัพธ์ที่เป็นไปได้นี้สมมุติให้มีจำนวนเป็นอนันต์
- จากการที่ตัวอย่างเป็นค่าสุ่มในข้างต้น ค่าสถิติที่คำนวณจากข้อมูลตัวอย่างจึงมีคุณสมบัติเป็นค่าสุ่มด้วย การแจกแจงของค่าสถิติดังกล่าว เรียกว่า **การแจกแจงของฟังก์ชันที่ได้จากตัวอย่างสุ่ม (sampling distribution)** การแจกแจงนี้เป็นหัวใจของการอนุมานเชิงสถิติแบบดั้งเดิม



ข้อสังเกต การอนุมานเชิงสถิติแบบดั้งเดิม ไม่ได้อธิบายความไม่แน่นอนของค่าพารามิเตอร์ที่ต้องการศึกษาโดยตรง แต่เป็นการอธิบายความไม่แน่นอนของผลลัพธ์ที่ได้จากตัวอย่างสุ่ม นิสิตลองนึกย้อนถึงการแปลผลการวิเคราะห์ที่ได้เรียนมาแล้ว ได้แก่

- Standard Error
- ช่วงความเชื่อมั่น
- P-value

```
##
##  One Sample t-test
##
## data:  dat$Ach
## t = 5.078, df = 149, p-value = 1.123e-06
## alternative hypothesis: true mean is not equal to 5.5
## 95 percent confidence interval:
##  5.709732 5.976934
## sample estimates:
## mean of x
##  5.843333
```

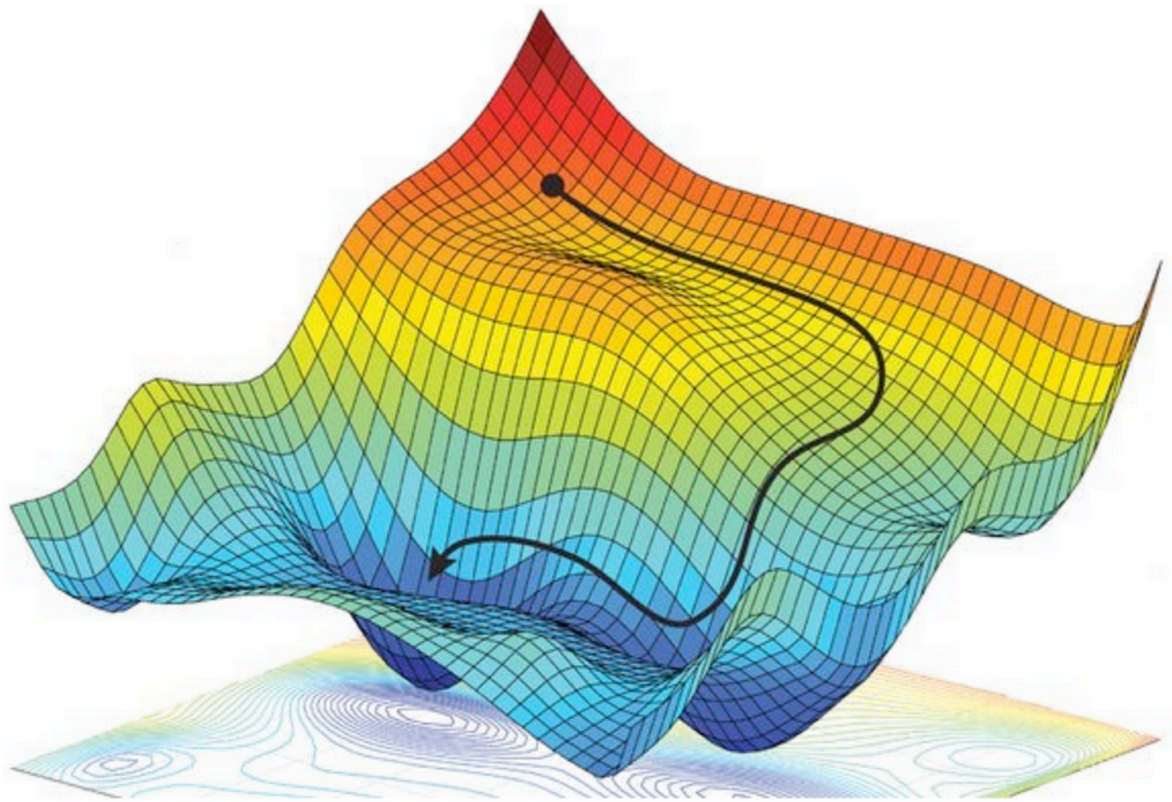
คำถาม

1. output ข้างต้นแสดงผลการประมาณช่วงความเชื่อมั่น และทดสอบสมมุติฐานทางสถิติเกี่ยวกับผลสัมฤทธิ์ทางการเรียนของนักเรียน (คะแนนเต็ม 10) โดยจากการประมาณด้วยช่วงความเชื่อมั่น 95% ในข้างต้นพบว่ามีค่าเท่ากับ $[5.71, 5.98]$ ซึ่งหมายความว่าอะไร?

2. จากผลการทดสอบสมมุติฐาน $H_0 : \mu_{Ach} = 5.5$ พบว่ามีค่า p-value < .000 หมายความว่าอะไร?

3. เราสามารถตัดสินใจยอมรับสมมุติฐาน H_0 ได้หรือไม่?

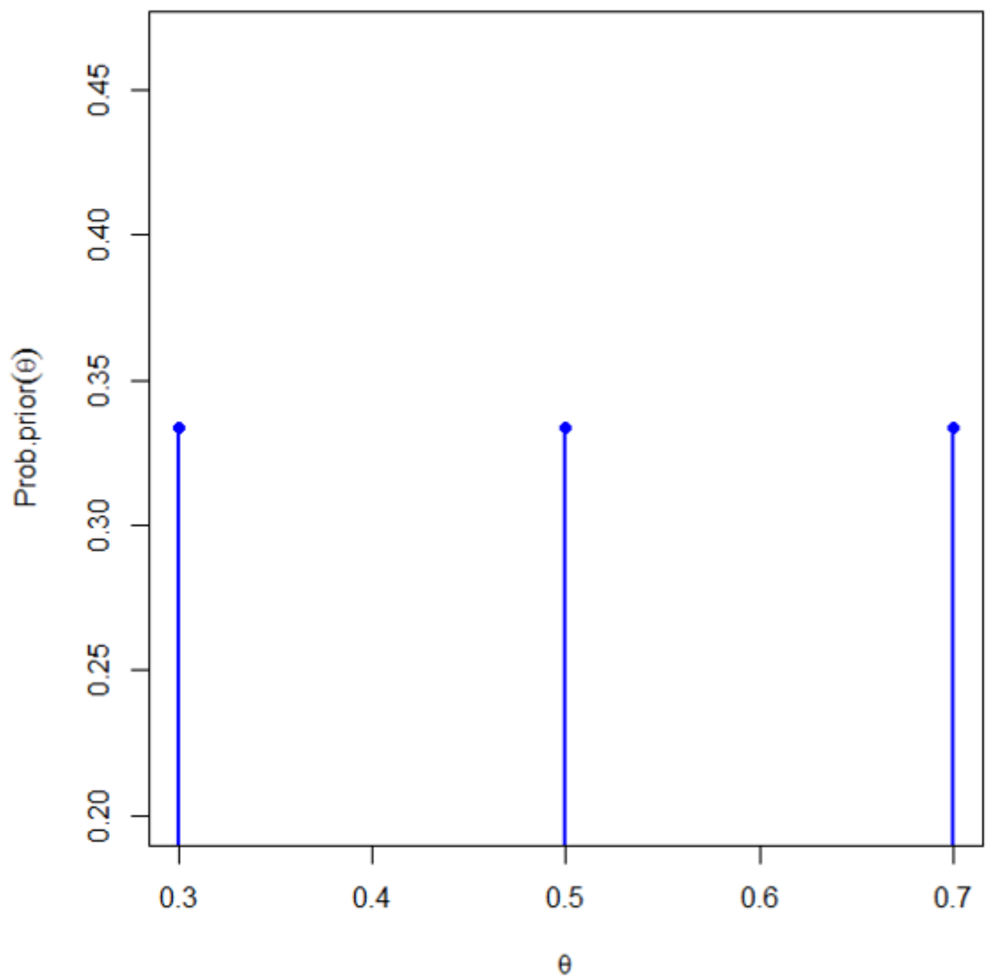
- การแปลผลการวิเคราะห์ที่ไม่ตรงไปตรงมา
- การประมาณค่าพารามิเตอร์ภายใต้โมเดลซับซ้อนที่อาจไม่มีประสิทธิภาพในบางกรณี
- การเปรียบเทียบ/คัดเลือกโมเดล
- การวิเคราะห์ภายใต้สถานการณ์ที่ขนาดตัวอย่างเล็ก



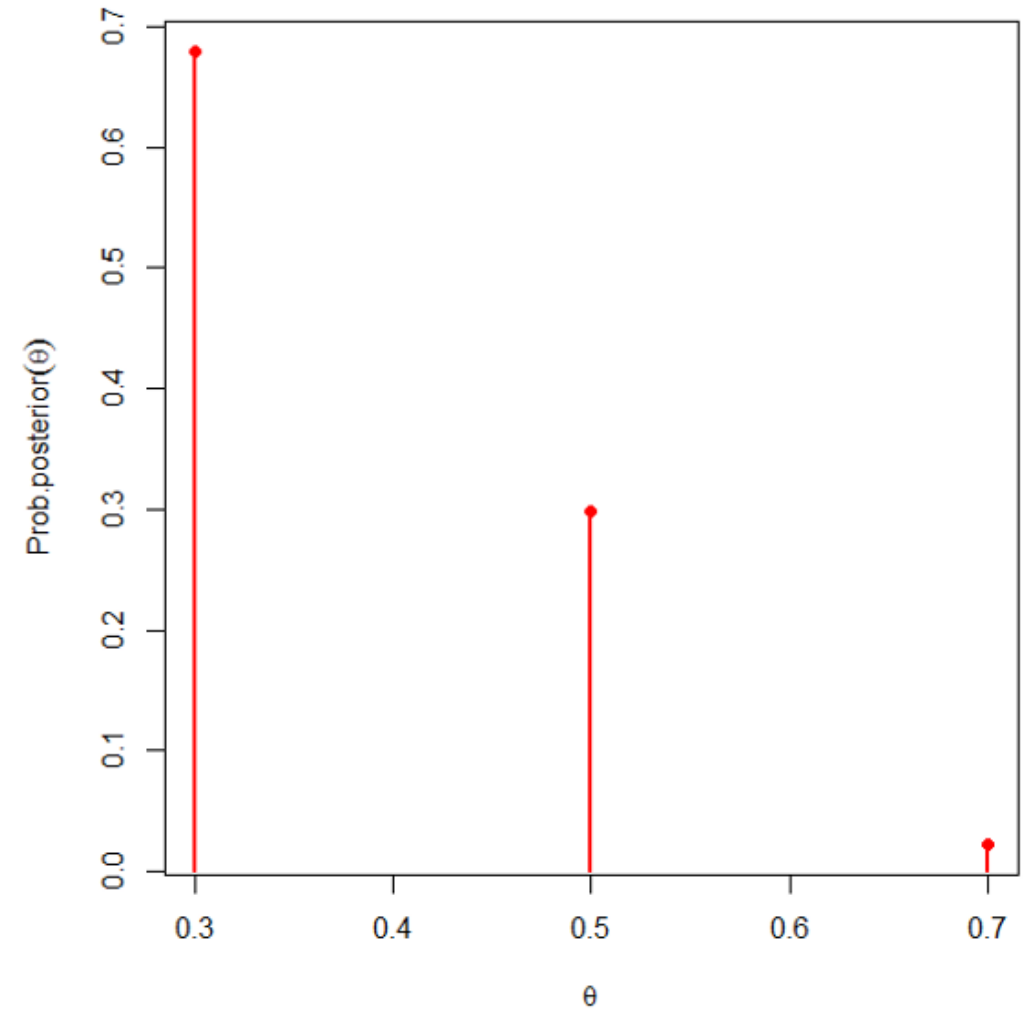
https://towardsdatascience.com/coding-deep-learning-for-beginners-linear-regression-gradient-descent-fcd5e0fc077d

Bayesian Statistics

- พารามิเตอร์เป็นค่าสุ่ม มีความไม่แน่นอน สามารถกำหนดการแจกแจงความน่าจะเป็นเพื่ออธิบายความไม่แน่นอนดังกล่าวได้
- ข้อมูลตัวอย่างเป็นค่าคงที่
- ไม่มีค่าสถิติ เพราะในการอนุมานจะทำผ่านการแจกแจงของพารามิเตอร์โดยตรง
- การแจกแจงความน่าจะเป็นภายหลัง (posterior distribution) เป็นเครื่องมือหลักสำหรับการอนุมานเชิงสถิติแบบเบส์ และการอนุมานก็ไม่ยุ่งยากซับซ้อนเหมือนกับสถิติแบบดั้งเดิม เพราะใช้เพียงสถิติพื้นฐานในการสรุปสารสนเทศเกี่ยวกับพารามิเตอร์ที่ต้องการจาก posterior distribution ได้เลย แตกต่างจากสถิติแบบดั้งเดิมที่มีวิธีการวิเคราะห์และตัวสถิติที่แตกต่างกันในแต่ละปัญหา



Prior distribution



Posterior distribution

(given n=10 and H=3)

- การจัดสรรน้ำหนักให้กับแต่ละความเป็นไปได้ของความลำเอียงเมื่อมีข้อมูลจริงจากการโยนเหรียญเพิ่มเติม ควรทำอย่างไร?

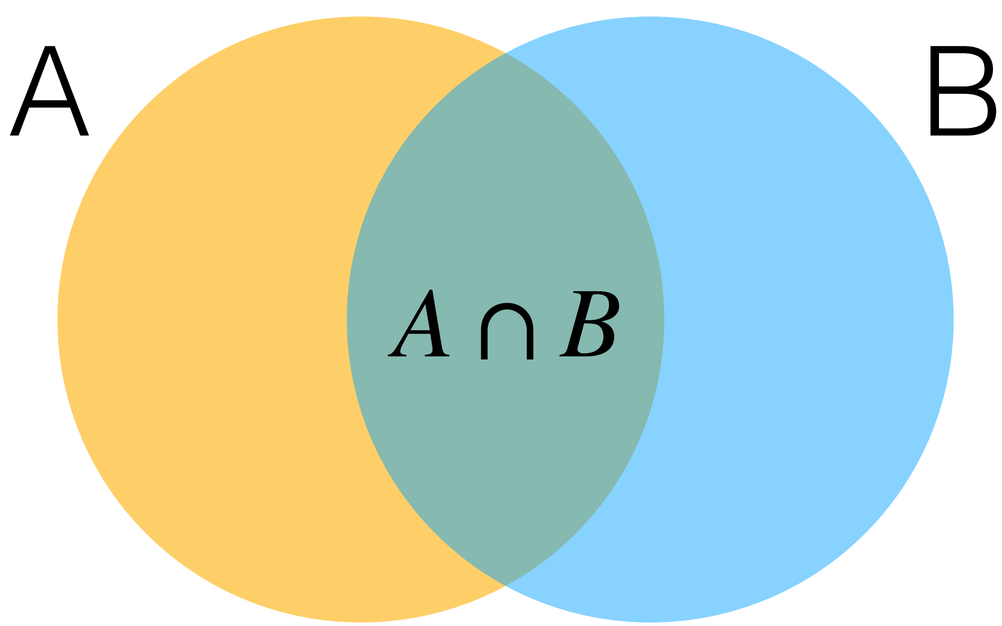
Thomas Bayes

Portrait used of Bayes in a 1936 book,[1] but it is doubtful whether the portrait is actually of him.[2] No earlier portrait or claimed portrait survives.

โทมัส เบยส์ (Thomas Bayes, 1702-1761) เป็นนักคณิตศาสตร์ผู้พัฒนาทฤษฎีนี้ขึ้น ทฤษฎีนี้เป็นการประยุกต์ใช้ความรู้เกี่ยวกับความน่าจะเป็นแบบมีเงื่อนไข (conditional probability) เพื่อหาค่าความน่าจะเป็นของเหตุการณ์ที่สนใจใหม่โดยใช้ข้อมูลที่มีเพิ่มเติม ดังนั้นก่อนอื่นจะกล่าวถึงความน่าจะเป็นแบบมีเงื่อนไขก่อนแล้วจึงกล่าวถึงทฤษฎีบทของเบย์ต่อไป

Conditional Probability

- P(โพแดง) =
- P(Q) =
- P(โพแดง | Q) =
- P(Q | โพแดง) =



$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

ตัวอย่าง กล่องใบหนึ่งประกอบไปด้วยลูกบอลสีแดง 5 ลูก และลูกบอลสีขาว 3 ลูก หยิบลูกบอล 2 ลูกออกจากกล่อง โดยหยิบทีละลูกและไม่ใส่คืน จงหา

1) ความน่าจะเป็นที่ลูกบอลลูกที่สองจะเป็นสีขาว

2) ความน่าจะเป็นที่ลูกบอลลูกที่สองจะเป็นสีขาว โดยที่ทราบว่าลูกแรกเป็นสีแดง

3) ความน่าจะเป็นที่ลูกบอลลูกที่สองจะเป็นสีขาว โดยที่ทราบว่าลูกแรกเป็นสีขาว

นิยาม กำหนดให้ y และ x เป็นตัวแปรที่สนใจ ซึ่ง $p(x) > 0$ **เราจะเรียก $\boldsymbol{p(y|x)}$ ซึ่งกำหนดค่าโดย**

$$p(y|x) = \frac{p(y,x)}{p(x)} = \begin{cases} \dfrac{p(y,x)}{\sum_y p(x,y)} & \text{กรณีตัวแปร } y \text{ ไม่ต่อเนื่อง} \\ \dfrac{p(y,x)}{\int p(x,y)dy} & \text{กรณีตัวแปร } y \text{ ต่อเนื่อง} \end{cases}$$

ว่า **ความน่าจะเป็นแบบมีเงื่อนไข (conditional probability) ของ $\boldsymbol{y}$ เมื่อกำหนดค่าตัวแปร $\boldsymbol{x}$**

เรียก $p(y,x)$ ว่าฟังก์ชันความน่าจะเป็นร่วมของตัวแปร y และ x

เรียก $p(x)$ ว่าฟังก์ชันความน่าจะเป็นมาร์จินัล (marginal probability) ของตัวแปร x

ตัวอย่าง ในสำนักงานแห่งหนึ่งมีพนักงานจำนวน 100 คน โดยจำแนกตามระดับการศึกษาและเพศได้ดังตาราง

เพศ **ระดับการศึกษา**	**ชาย**	**หญิง**	**รวม**
ป.ตรี	40	10	70
ป.โท	20	30	30
รวม	60	40	100

1) หากสุ่มเลือกพนักงานขึ้นมาคนหนึ่ง จงหาความน่าจะเป็นที่พนักงานคนนี้จะจบการศึกษาระดับปริญญาโท
2) หากผู้จัดการต้องการใช้ให้พนักงานไปซื้อกาแฟจึงสุ่มเลือกพนักงานขึ้นมาคนหนึ่งและพบว่าเป็นเพศชาย จงหาความน่าจะเป็นที่พนักงานคนนี้จะจบการศึกษาระดับปริญญาโท
3) หากประธานบริษัทต้องการแต่งตั้งเลขาส่วนตัวจึงสุ่มเลือกพนักงานขึ้นมาคนหนึ่งและพบว่าเป็นเพศหญิง จงหาความน่าจะเป็นที่พนักงานคนนี้จะจบการศึกษาระดับปริญญาโท

Bayes' Theorem

จากนิยามของความน่าจะเป็นแบบมีเงื่อนไขเราจะได้ว่า

$$p(y|x) = \frac{p(y,x)}{p(x)} \text{ และ } p(x|y) = \frac{p(y,x)}{p(y)}$$

ซึ่งทำให้ได้ว่า

$$p(y,x) = p(x)p(y|x) \text{ หรือ } p(y,x) = p(y)p(x|y)$$

จับสมการทั้งสองเท่ากันดังนี้ $p(x)p(y|x) = p(y)p(x|y)$ เกิดเป็น **ทฤษฎีบทของเบส์ (Bayes' Theorem)** ดังนี้

$$\boxed{p(y|x) = \frac{p(y)p(x|y)}{p(x)}}$$

โดยที่ $p(x) = \sum_y p(x,y) = \sum_y p(y)p(x|y)$ (ในกรณีที่ y เป็นตัวแปรไม่ต่อเนื่อง)

หรือ $p(x) = \int p(x,y)dy = \int p(y)p(x|y)dy$ (ในกรณีที่ y เป็นตัวแปรต่อเนื่อง)

นักวิจัยทางการศึกษาท่านหนึ่งเก็บรวบรวมข้อมูลเกี่ยวกับภาวะโรค Dyslexia จากการสุ่มตัวอย่างเด็กในวัยเรียนของประเทศไทยจำนวน 10000 คน(ภาวะความบกพร่องในการอ่านเขียน เป็นความผิดปรกติเฉพาะด้านการเรียนรู้ประเภทหนึ่งที่เกิดจากการทำงานที่ผิดปรกติของเซลล์สมองซีกซ้าย) ได้ข้อมูลดังนี้ (ตัวเลขสมมติ)

เพศ / **ภาวะโรค**	**ชาย**	**หญิง**	**รวม**
เป็น	2000	1000	3000
ไม่เป็น	4000	3000	7000
รวม	6000	4000	10000

หากทำการเลือกเด็กจากกลุ่มตัวอย่างนี้มา 1 คน แล้วพบว่าเป็นผู้ชาย จงหาความน่าจะเป็นที่จะมีภาวะของโรค Dyslexia

จาก $p(y|x) = \frac{p(y)p(x|y)}{p(x)}$ หากกำหนดให้ $y = \theta$ (ซึ่งแทนพารามิเตอร์ที่ต้องการศึกษา) และ $x = D$ แทนข้อมูลค่าสังเกต แล้วทฤษฎีบทของเบส์ข้างต้นจะเขียนใหม่ได้เป็น

$$p(\theta|D) = \frac{p(\theta)p(D|\theta)}{p(D)}$$

- เรียก $p(\theta)$ ว่าฟังก์ชันความน่าจะเป็นเบื้องต้น (prior probability function) ของพารามิเตอร์ θ ใช้อธิบายความเชื่อที่มีต่อความเป็นไปได้ของค่าพารามิเตอร์ที่สนใจ
- เรียก $p(D|\theta)$ ว่าฟังก์ชันภาวะความควรจะเป็นของข้อมูลค่าสังเกต (likelihood function) มีความหมายเป็นโอกาสในการเกิดข้อมูลค่าสังเกต D ของเมื่อกำหนดค่าพารามิเตอร์ θ ภายใต้โมเดลที่ทำการวิเคราะห์
- เรียก $p(D) = \int p(\theta)p(D|\theta)d\theta$ คือค่าความน่าจะเป็นของการเกิดข้อมูลค่าสังเกต x ภายใต้โมเดลที่ทำการวิเคราะห์ แต่เนื่องจากเรามีข้อมูลค่าสังเกตเพียงชุดเดียว ดังนั้นจึงมองส่วนนี้เป็นส่วนค่าคงที่เพื่อปรับให้ผลคูณ $p(\theta)p(D|\theta)$ มีคุณสมบัติเป็นฟังก์ชันความน่าจะเป็นอยู่เรียกว่าค่า **“normalized constant”**
- **เรียก $p(\theta|D)$ ว่า ฟังก์ชันความน่าจะเป็นภายหลังของพารามิเตอร์ θ**

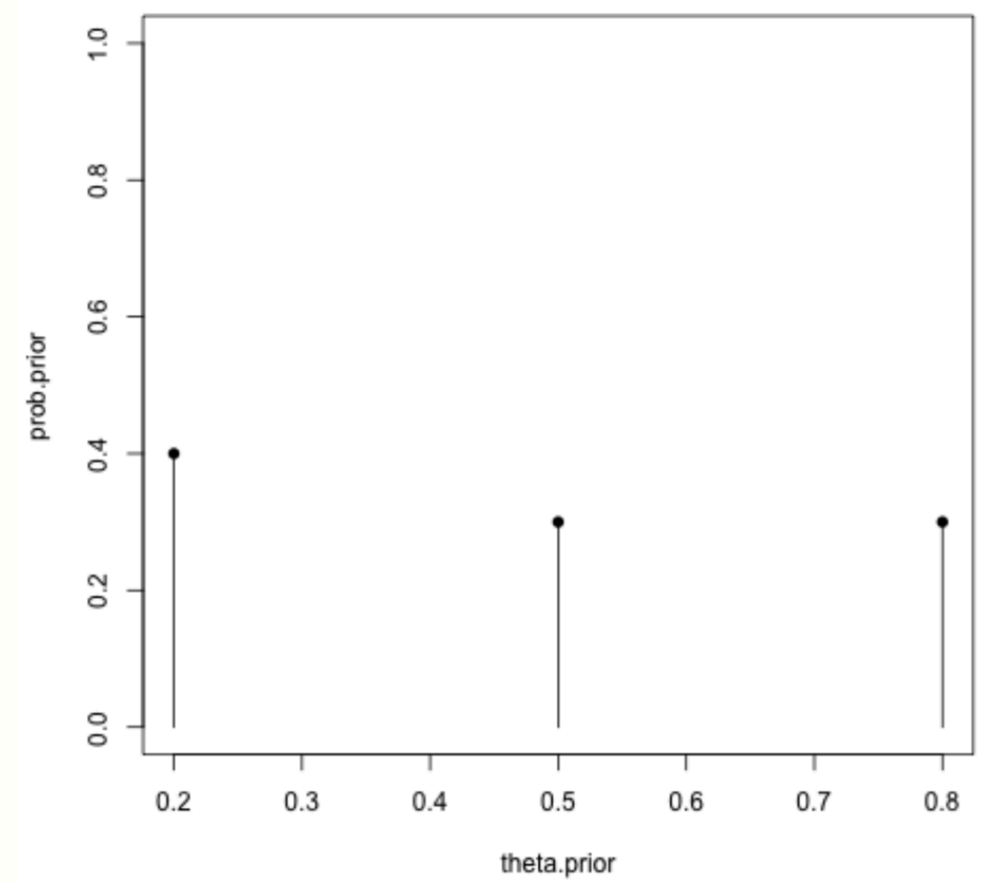
Table 5.5 Applying Bayes' rule to data and parameters

Data	Model parameter			Marginal
	...	θ value	...	
$\vdots$		$\vdots$		$\vdots$
D value	$\cdots$	$p(D,\theta) = p(D\|\theta)\, p(\theta)$	$\cdots$	$p(D) = \sum_{\theta^*} p(D\|\theta^*)\, p(\theta^*)$
$\vdots$		$\vdots$		$\vdots$
Marginal	...	$p(\theta)$	...	

When conditionalizing on row value D, the conditional probability $p(\theta|D)$ is the cell probability $p(D,\theta)$ divided by the marginal probability $p(D)$. When these probabilities are algebraically re-expressed as shown in the table, this is Bayes' rule. This table is merely Table 5.1 with its rows and columns re-named.

สมมุติว่ากำหนดให้ prior distribution ของพารามิเตอร์ความลำเอียง (θ) เป็นดังตาราง

Theta	Weight —> p(Theta)
0.2	0.4
0.5	0.3
0.8	0.2
รวม	1.00



หากไม่มีข้อมูลเพิ่มเติมจะได้ว่า ค่าประมาณแบบจุดของพารามิเตอร์ความลำเอียงมีค่าเท่ากับ

$$E(\theta) = \sum_{\theta} \theta p(\theta) = (0.2, 0.5, 0.8) \times (0.4, 0.3, 0.2)^T$$

$$= (0.2)(0.4) + (0.5)(0.3) + (0.8)(0.2)$$

$$= 0.47$$

$$Var(\theta) = \sum_{\theta} (\theta - 0.47)^2 p(\theta) = 0.0621 \implies SD(\theta) = \sqrt{0.0621} = 0.249$$

กำหนดให้โมเดลของค่าสังเกตเป็น

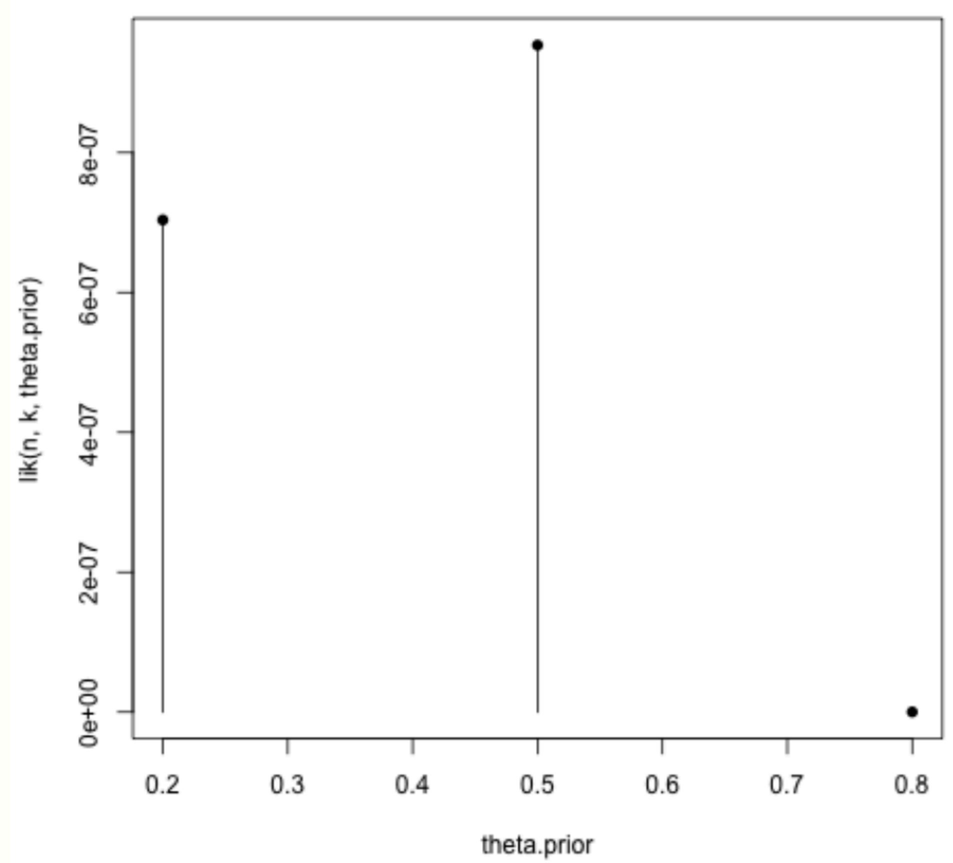
$$p(y_i|\theta) = \theta^{y_i}(1-\theta)^{1-y_i}$$

ดังนั้นฟังก์ชันภาวะความควรจะเป็นจึงมีรูปแบบทั่วไปเป็น

$$p(\mathbf{y}|\theta) = \theta^{\sum y_i}(1-\theta)^{n-\sum y_i}$$

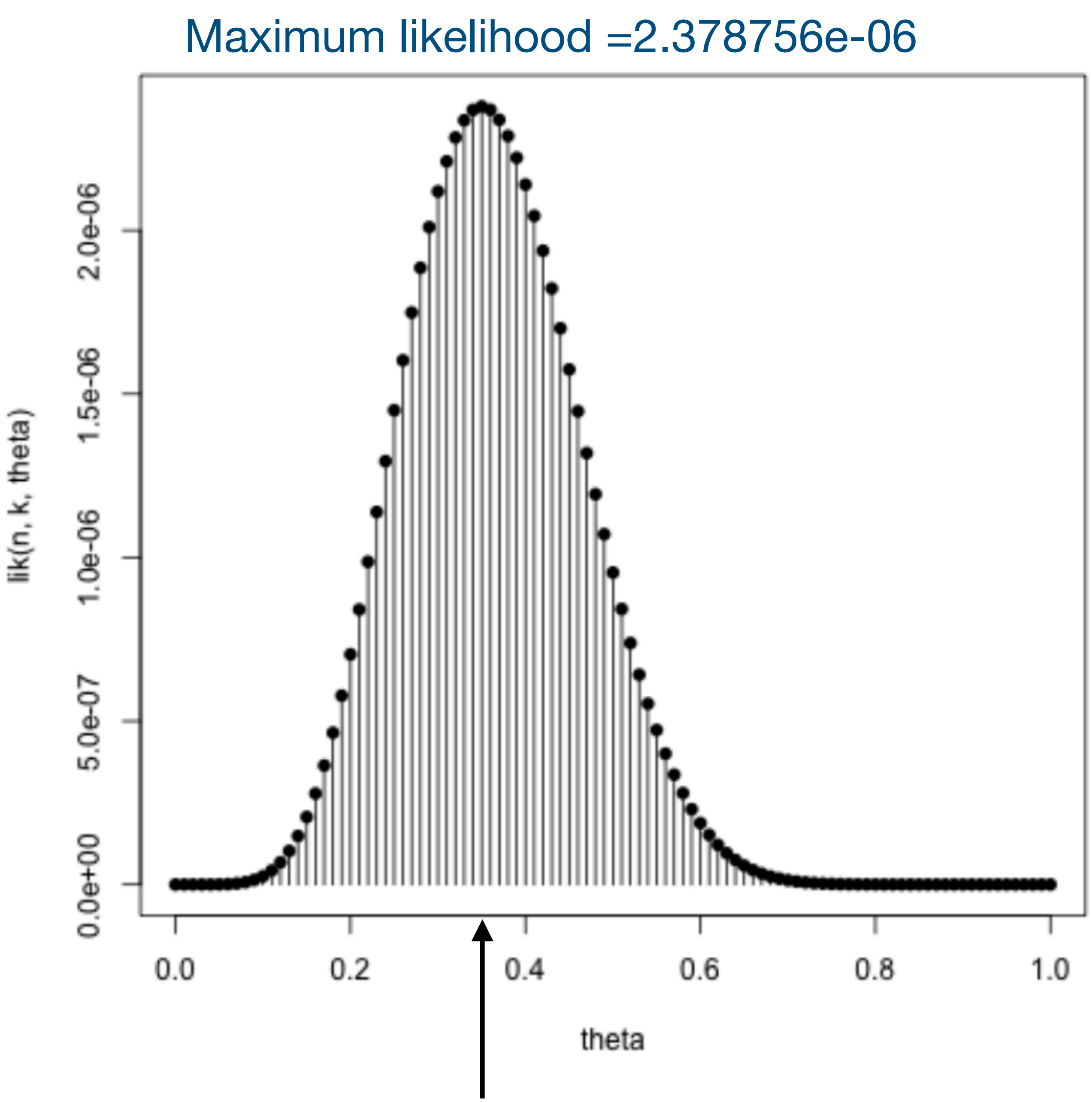
สมมุติว่าทดสอบโยนเหรียญ 20 ครั้ง พบว่าเกิดหน้าหัว 7 ครั้ง ฟังก์ชันภาวะความควรจะเป็นจึงมีค่าเท่ากับ

Theta	p(y|theta)
0.2	7.036874E-07
0.5	9.536743E-07
0.8	1.717987E-10



Maximum likelihood estimate = 0.5

Maximum likelihood estimate



Maximum likelihood estimate = 0.35

Maximum likelihood estimate

ในกรณีทั่วไปสามารถหาค่าประมาณภาวะความควรจะเป็นสูงสุดได้ดังนี้

$$p(\mathbf{y}|\theta) = \theta^{\sum y_i}(1-\theta)^{n-\sum y_i}$$

$$ln(p(\mathbf{y}|\theta)) = \sum y_i \times ln(\theta) + (n - \sum y_i) \times ln(1-\theta)$$

$$= \sum y_i \times ln(\theta) + n \times ln(1-\theta) - \sum y_i \times ln(1-\theta)$$

หาอนุพันธ์อันดับที่ 1 ของ log-likelihood function ข้างต้นเทียบกับพารามิเตอร์ θ

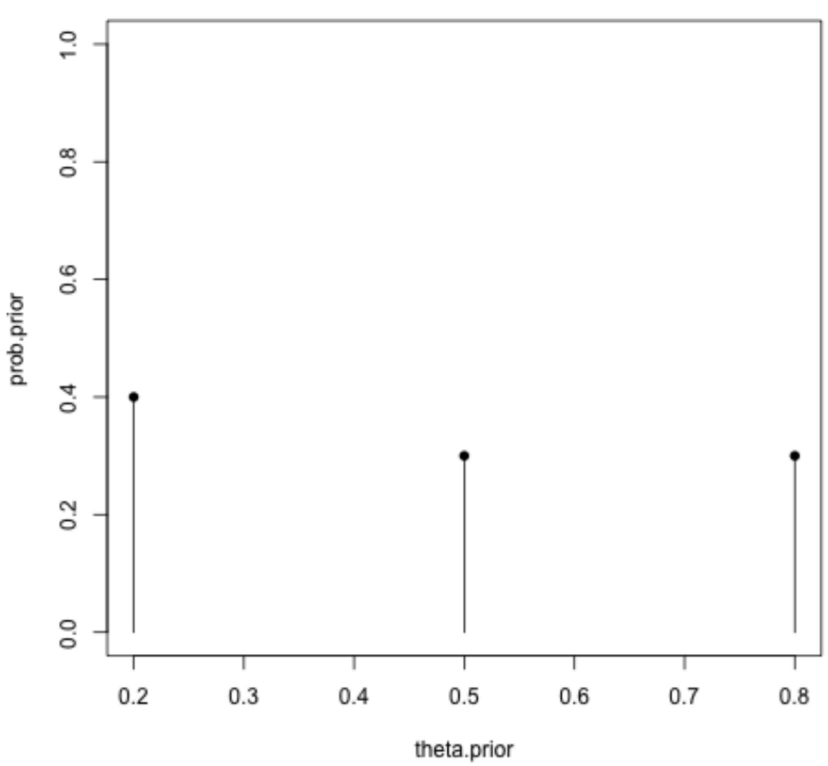
$$\frac{\partial ln(p(\mathbf{y}|\theta))}{\partial\theta} = \frac{\sum y_i}{\theta} + \frac{n}{1-\theta}(-1) - \frac{\sum y_i}{1-\theta}(-1)$$

กำหนดให้ $\frac{\partial ln(p(\mathbf{y}|\theta))}{\partial\theta} = 0 \implies \hat{\theta} = \frac{\sum y_i}{n}$

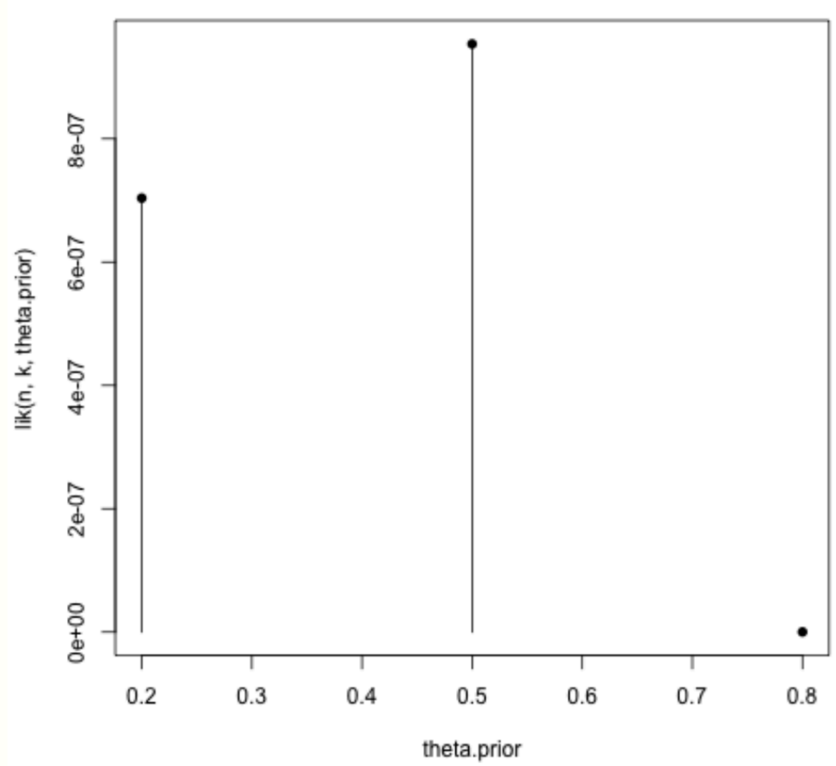
Posterior Distribution

Theta	Weight —> p(Theta)	p(Data | theta)	p(theta) x p(Data | theta)	p(theta|Data)
0.2	0.4	7.036874E-07	0.000000281474960	0.496
0.5	0.3	9.536743E-07	0.000000286102290	0.504
0.8	0.2	1.717987E-10	0.000000000034360	0.000
รวม	1.00		0.000000567611610	1.000

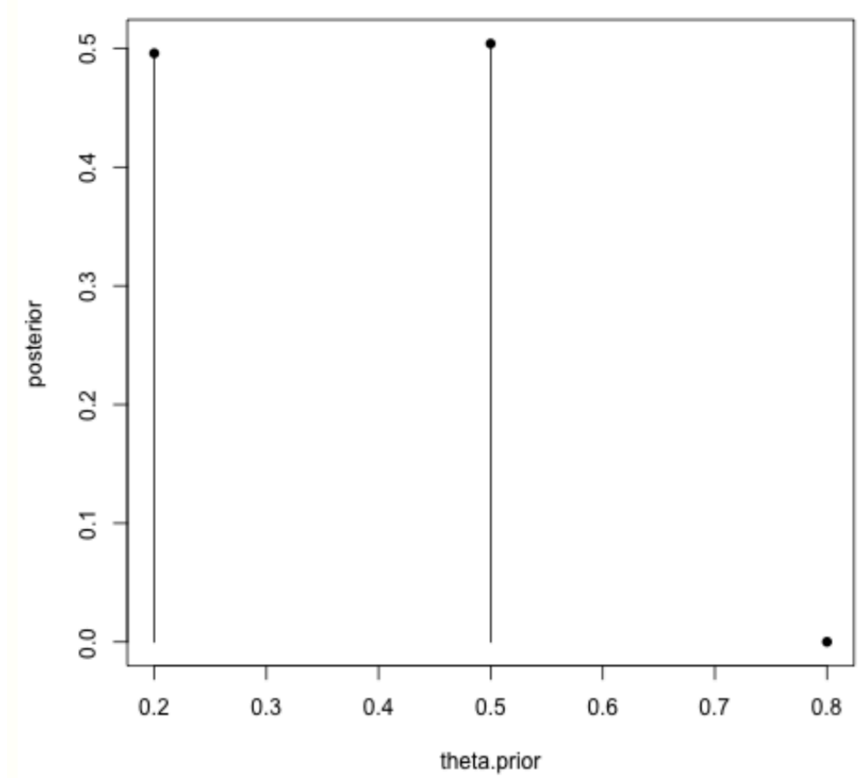
Prior



Likelihood



Posterior



Statistical Inference

- Estimation
- Prediction
- Model Comparison/Hypothesis Testing

Posterior mean

$$E(\theta \mid \mathbf{y}) = \sum_{\theta} \theta \times p(\theta \mid \mathbf{y})$$

$$= (0.2)(0.496) + (0.5)(0.504) + (0.8)(0.000)$$

$$= 0.3512$$

Posterior variance

$$Var(\theta \mid \mathbf{y}) = \sum (\theta - 0.3512)^2 p(\theta \mid \mathbf{y}) = 0.023$$

Prediction

การทำนายค่าสังเกตของโมเดล สามารถทำได้โดยพิจารณาการแจกแจงค่าความน่าจะเป็นของค่าทำนาย (predictive distribution) โดยความน่าจะเป็นของค่าทำนายแต่ละค่าสามารถคำนวณได้จาก

$$p(y_i) = \int_\theta p(y_i | \theta) p(\theta)$$

จากตัวอย่างการโยนเหรียญข้างต้น

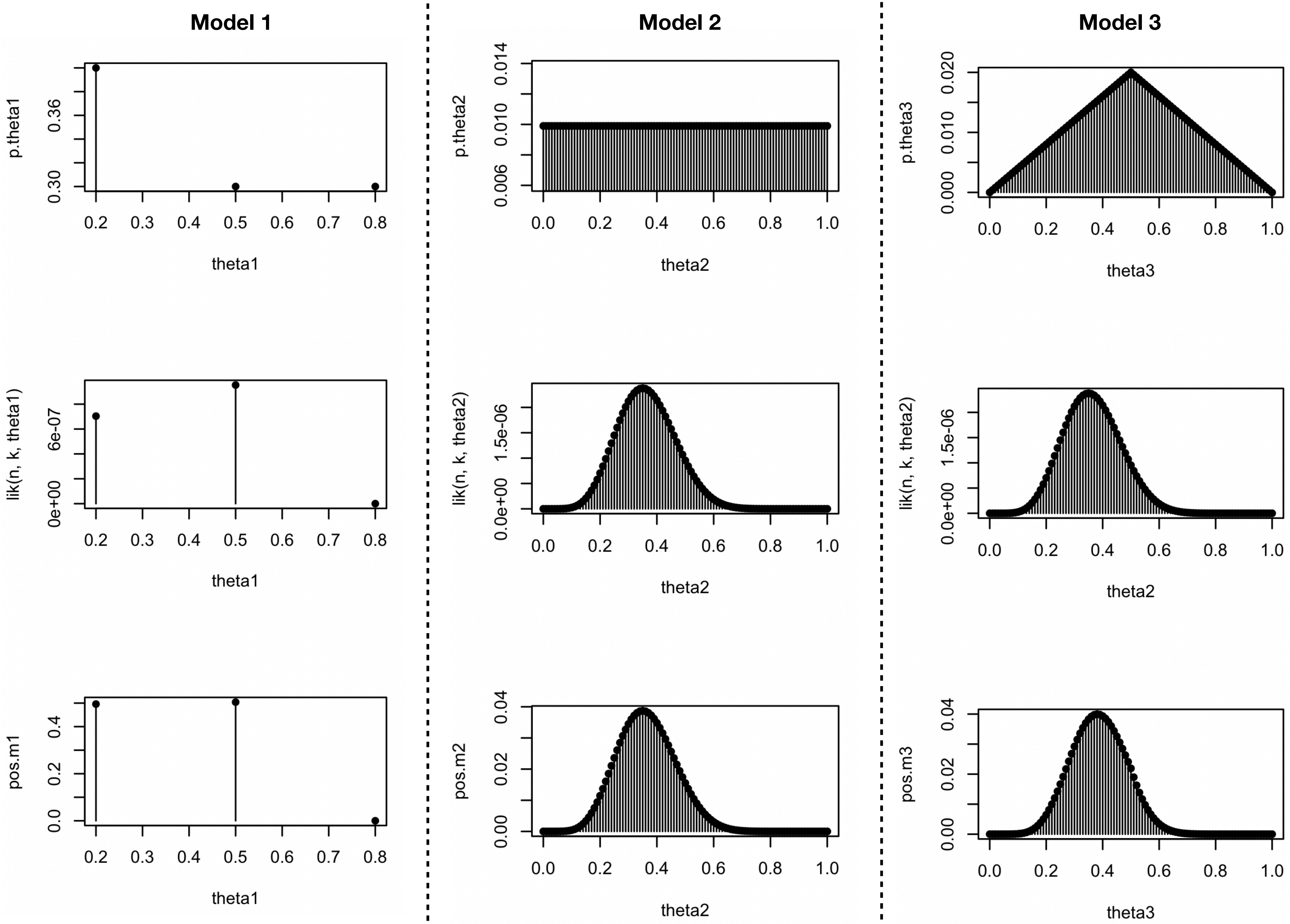
$$\boxed{\begin{aligned} p(y_i = H) &= \sum_\theta p(y_i = H | \theta) p(\theta) \\ &= (0.2)(0.496) + (0.5)(0.504) \\ &= 0.3512 \\ p(y_i = T) &= 1 - 0.3512 = 0.6488 \end{aligned}}$$

<—- Predictive data distribution

Prediction

Predicted value = $\int_y yp(y) = (1)(0.3512) + (0)(0.6488) = 0.3512$

Model Comparison



Bayes Factor

In some applications it can help to make the model explicit in Bayes' rule. Let's call the model *M*. Then, because all the probabilities are defined given that model, we can write:

$$\underbrace{p(\theta|D,M)}_{\text{posterior}} = \underbrace{p(D|\theta,M)}_{\text{likelihood}}\ \underbrace{p(\theta|M)}_{\text{prior}} \Big/ \ \underbrace{p(D|M)}_{\text{evidence}}$$

where the evidence is

$$p(D\,|\,M) = \int_\theta p(D\,|\,\theta, M)p(\theta\,|\,M)$$

Suppose we have two models named M1 and M2

$$\frac{p(M1|D)}{p(M2|D)} = \underbrace{\frac{p(D|M1)}{p(D|M2)}}_{\text{Bayes factor}}\ \frac{p(M1)}{p(M2)}.$$

Bayes factor → **Ratio of evidence (or marginal likelihood)**

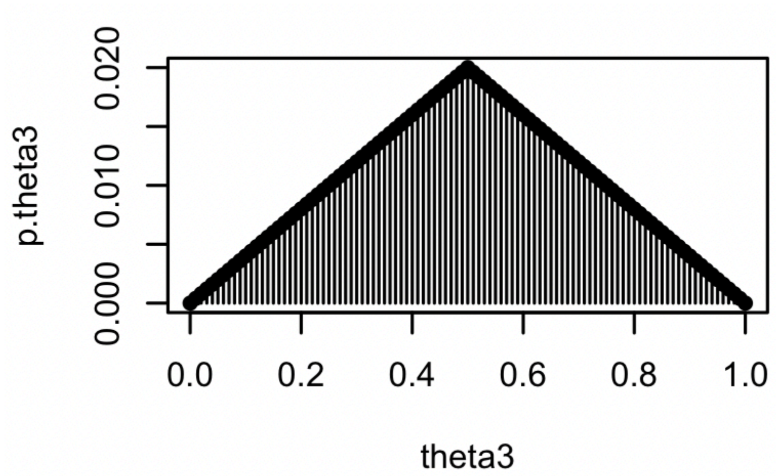
(Kruschke, 2010)

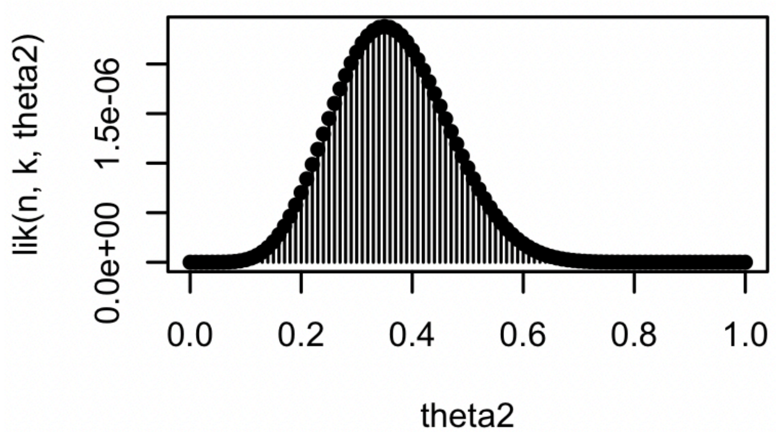
```r
#evidence ---> P(Data | Model)
e1<-sum(lik(n,k,theta1)*p.theta1)
e2<-sum(lik(n,k,theta2)*p.theta2)
e3<-sum(lik(n,k,theta3)*p.theta3)
e1
e2
e3
(B21<-e2/e1)
(B31<-e3/e1)
(B32<-e3/e2)
```

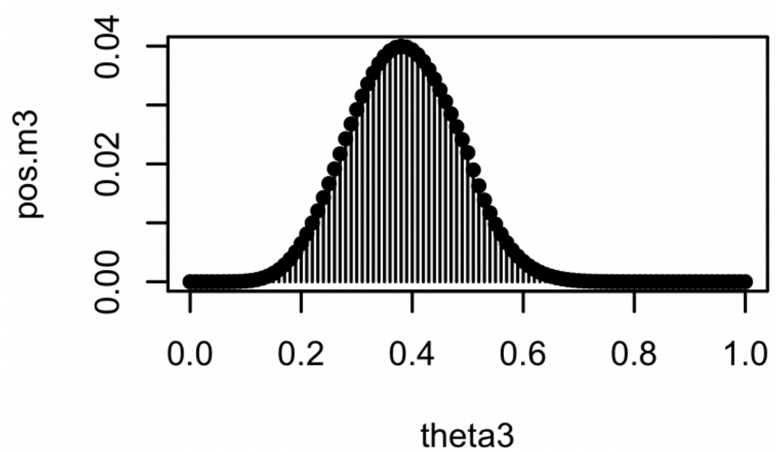
Bayes Factor

Model	1	2	3
1			
2	1.071473		
3	1.533176	1.430905	

Model 3







Posterior mean

$$E(\theta|\mathbf{y}) = \sum_{\theta} \theta p(\theta|\mathbf{y}) = 0.3861$$

Posterior SD

$$SD(\theta|\mathbf{y}) = \sqrt{\sum_{\theta} (\theta - 0.3861)^2 p(\theta|\mathbf{y})} = 0.0953$$

เหรียญเที่ยงตรงหรือไม่?

$$P(\theta = 0.5 \,|\, \mathbf{y}) = 0.0219$$

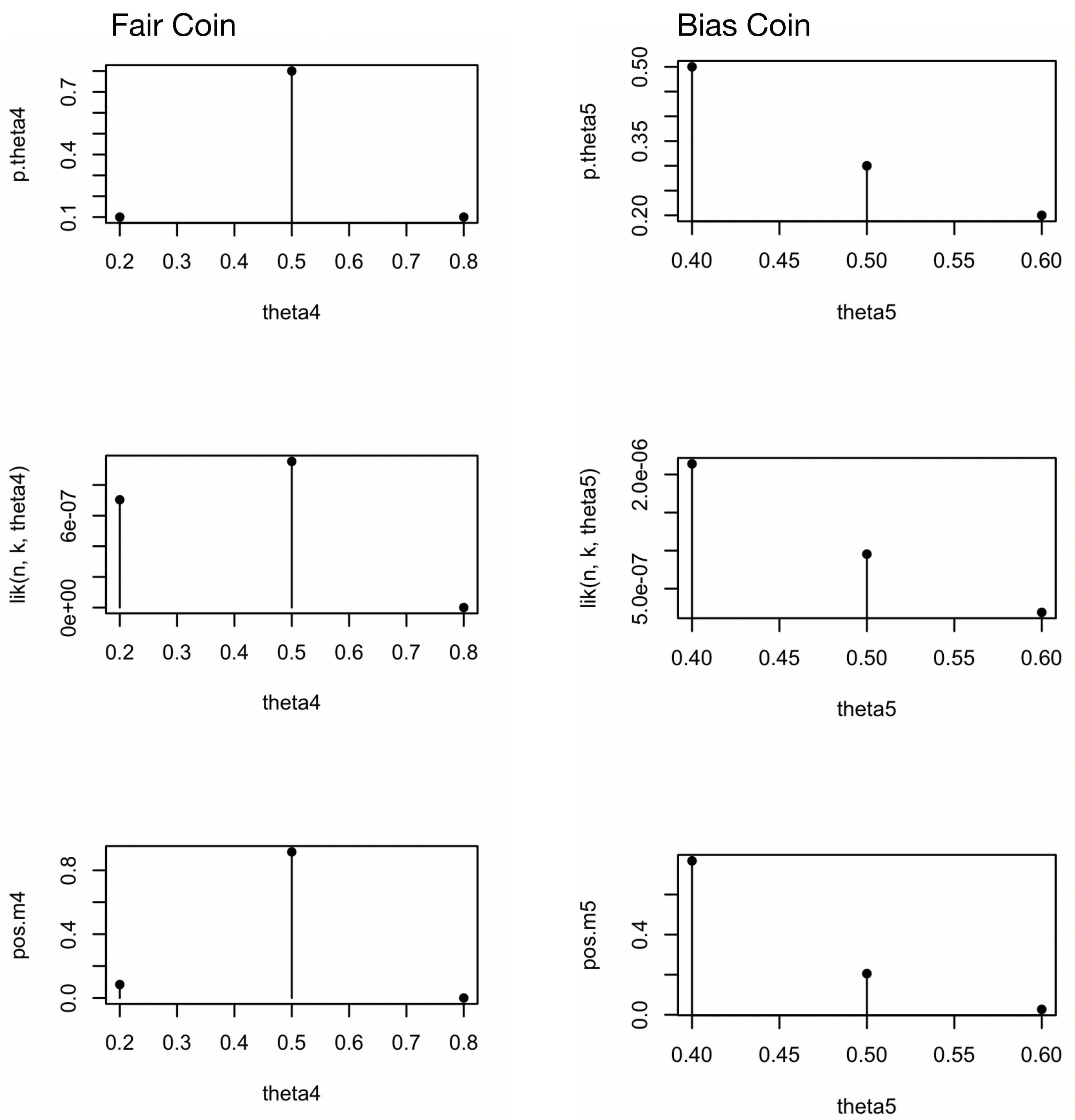
$$P(\theta \neq 0.5 \,|\, \mathbf{y}) = 1 - 0.021 = 0.979$$

เหรียญเที่ยงตรงหรือไม่?

$$P(0.45 < \theta < 0.55 \,|\, \mathbf{y}) = 0.1923$$

เหรียญเที่ยงตรงหรือไม่?

$$\frac{p(\mathbf{y} \mid bias)}{p(\mathbf{y} \mid fair)} = \frac{1.393607e-06}{8.333254e-07} = 1.672344$$



4.3.4 Why Bayesian inference can be difficult

All three goals involve the denominator of Bayes' formula, i.e., the evidence, which usually means computing a difficult integral. There are a few ways out of this difficulty. The traditional way is to use likelihood functions with "conjugate" prior functions. A prior function that is conjugate to the likelihood function simply makes the posterior function come out with the same functional form as the prior. That is, the math works out nicely. If this method doesn't work, an alternative is to approximate the actual functions with other functions that are easier to work with, and then show that the approximation is reasonably good under typical conditions. But this method is still pure, analytical mathematics. Yet another method is to numerically approximate the integral. When the parameter space is small, then it can be covered with a comb or grid of points and the integral can be computed by exhaustively summing across that grid. But when the parameter space gets even moderately large, there are too many grid points, and therefore other methods must be used. A large class of random sampling methods have been developed, which can be referred to as Markov chain Monte Carlo (MCMC) methods, that can numerically approximate probability distributions even for large spaces. It is the development of these MCMC methods that has allowed Bayesian statistical methods to gain practical use. The next major part of this book explains these various methods in some detail. For applications to complex situations, we will ultimately focus on MCMC methods.

Another potential difficulty of Bayesian inference is determining a reasonable prior. What distribution of beliefs should we start with, over all possible parameter values or over competing models? This question may seem daunting, but in practice it is typically addressed in straightforward manner. As will be discussed more in Chapter 11, it is actually advantageous and rational to start with an explicit prior. Prior beliefs *should* influence rational inference from data, because the role of new data is to modify our beliefs from whatever they were without the new data. Prior beliefs are *not* capricious and idiosyncratic and unknowable, but instead are based on publicly agreed facts and theories. Prior beliefs used in data analysis must be admissible by a skeptical scientific audience. When scientists disagree about prior beliefs, the analysis can be conducted with both priors, to assess the robustness of the posterior against changes in the prior. Or, the priors can be mixed together into a joint prior, with the posterior thereby incorporating the uncertainty in the prior. In summary, for most applications, specification of the prior turns out to be technically *un*problematic, although it is conceptually very important to understand the consequences of one's assumptions about the prior. Thus, the main reason that Bayesian analysis can be difficult is the computation of the evidence, and that computation is tractable in many situations via MCMC methods.