

Information Retrieval

Venkatesh Vinayakarao

Term: Aug – Dec, 2018

Indian Institute of Information Technology, Sri City



So much of life, it seems to me, is determined by pure randomness.

– **Sidney Poitier.**

God does not play dice with the universe.

- **Einstein.**



Agenda

Updating Beliefs using Probability Theory (Bayesian Data Analysis)

Role of Beta Distributions

Will Discuss

- ✓ Concepts
- ✓ Illustrations
- ✓ Intuitions
- ✓ Purpose
- ✓ Properties

Will not Discuss

- ⊗ Details
- ⊗ Definitions
- ⊗ Formalism
- ⊗ Derivations
- ⊗ Proofs

The Case of Coin Flips

General Assumption: If a coin is fair! Heads (H) and Tails (T) are equally likely. But, coin need not be fair 😞



Experiment with **Coin - 1**

HHHTTTHTHT

Experiment with **Coin - 2**

HHHHHHHHHT

Coin-1 more likely to be fair when compared to **coin-2**.

Our Beliefs

- Can we find a structured way to determine coin's nature?



Coin-1

Data Observed	None	H	T	H	T
Belief	Fair	Skewed	Fair	Skewed	Fair



Coin-2

Data Observed	None	H	H	H	H
Belief	Fair	Skewed	More Skewed	Even More...	Even Even More...

Prior Belief

Belief Updates

Priors

- Priors can be strong or weak

Weak Prior

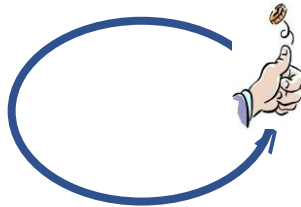
New Coin



A few observations sufficient to change our belief significantly.

Strong Prior

Coin is lab tested for 1 Million Tosses.
50% H, 50% T observed.



One more observation will not change our belief significantly.

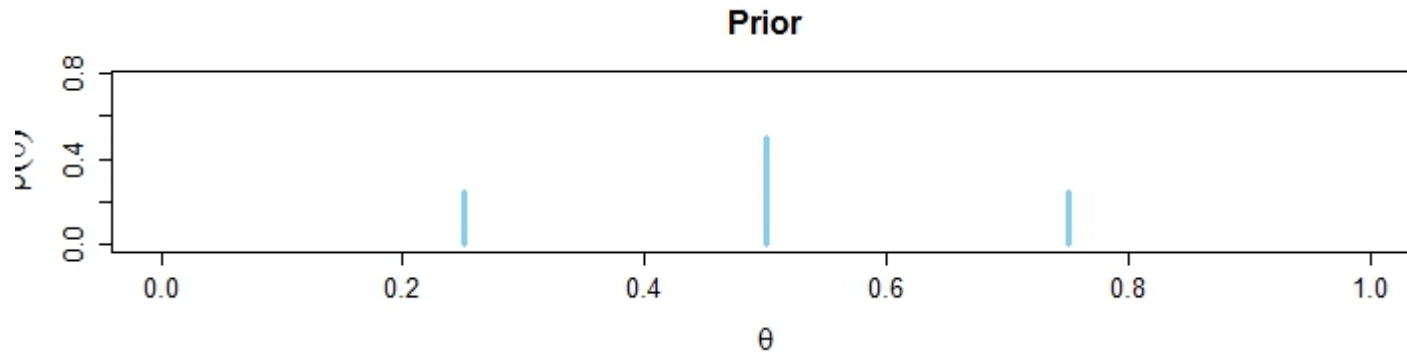
HyperParameter

- Prior probability (of Heads) could be anything:
 - 0.5 → Fair Coin
 - 0.25 → Skewed towards Tails
 - 0.75 → Skewed towards Heads
 - 1 → Head is guaranteed!
 - 0 → Both sides are Tails.

We use θ as a HyperParameter to visualize what happens for different values.

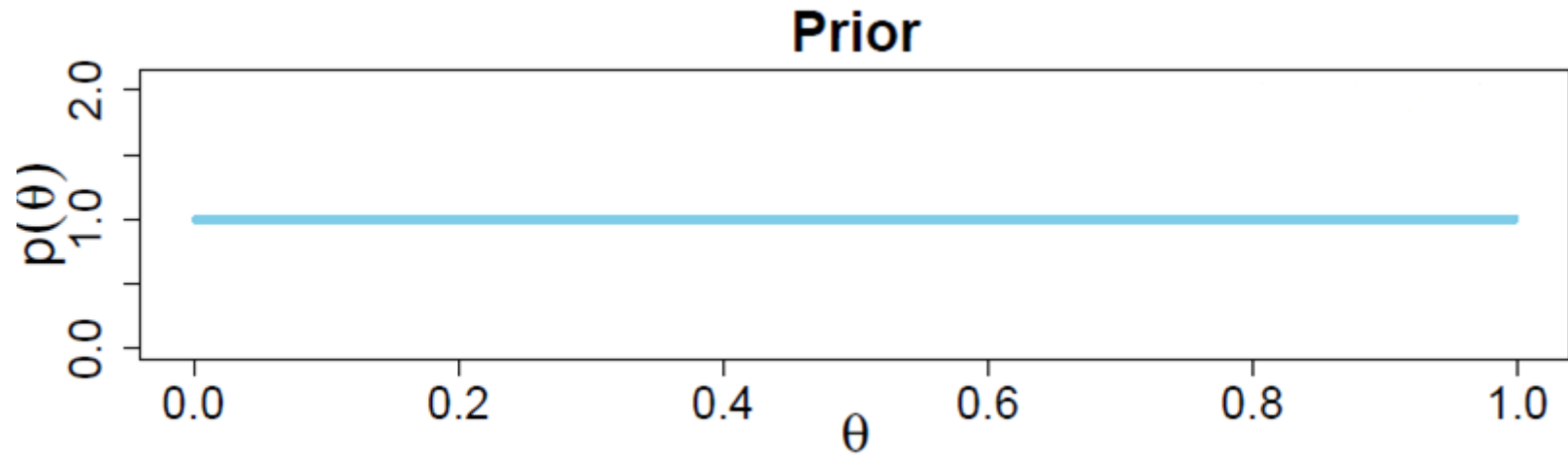
World of Distributions

Discrete Distribution of Prior. Since I typically perceive coins as fair, Prior belief peaks at 0.5.



Another Possibility

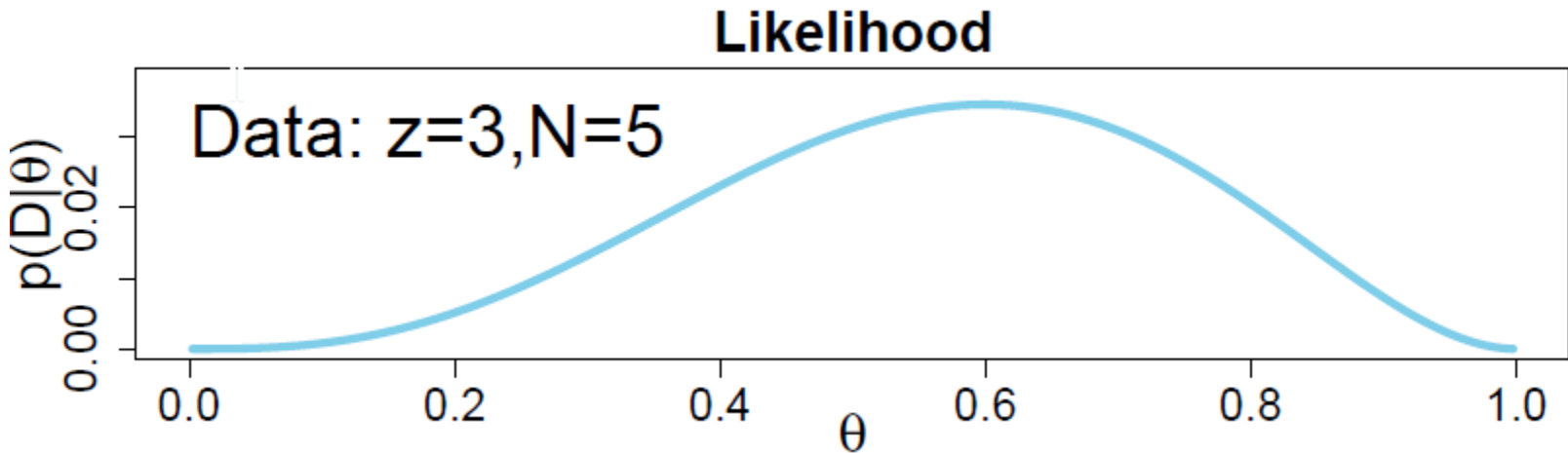
I may also choose to be unbiased! i.e., θ may take any value equally likely.



A Continuous Uniform Distribution!

Observations

Let's flip the coin (N) 5 times. We observe (z) 3 Heads.



Impact of Data

Belief is influenced by Observations. But, note that:

Belief \neq observation



Bayes' Rule

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

Eeks... what's in the denominator?

Numerator is easy

- $p(\theta)$ was uniform. So, nothing to calculate.
- How to calculate $p(D|\theta)$?



Jacob Bernoulli
1655 – 1705.

$$\theta^z (1 - \theta)^{n-z}$$

If D observed is HHHTT and θ is 0.5,
We have:

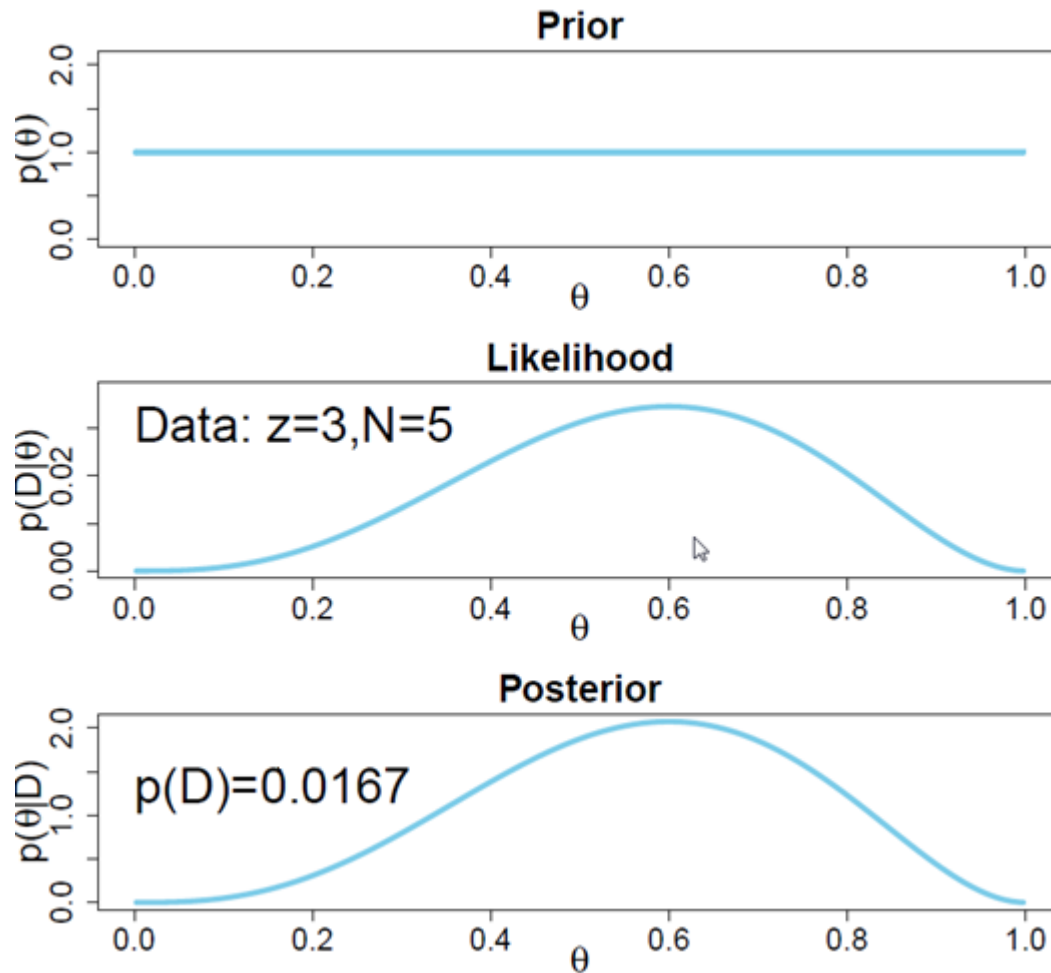
$$p(D|\theta) = (0.5)^3 (1 - 0.5)^{5-3}$$

Remember, two things: 1) we are interested in the distribution
2) Order of H,T does not matter.

Quiz

- Calculate $p(D | \theta)$ for the observation TTHHH and $\theta = 0.3$.
 - $(0.3)^3(0.7)^2 = 0.013$

Bayesian Update



Painful Denominator

- Recall, for discrete distributions:

$$P(\theta|D) = \frac{P(D|\theta) P(\theta)}{\sum_i P(D|\theta_i) P(\theta_i)}$$

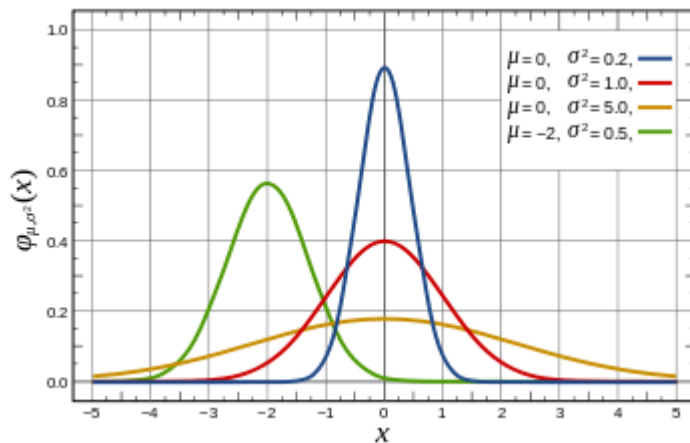
- And, for continuous distributions:

$$P(\theta|D) = \frac{P(D|\theta) P(\theta)}{\int P(D|\theta) P(\theta) d\theta}$$

A Simpler Way

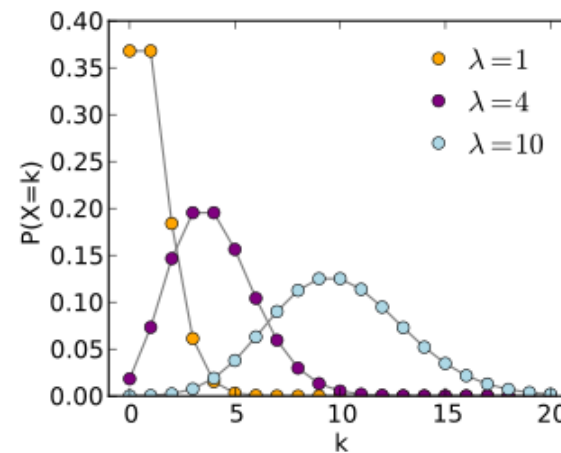
Form, Functions and Distributions

Normal (or Gaussian)



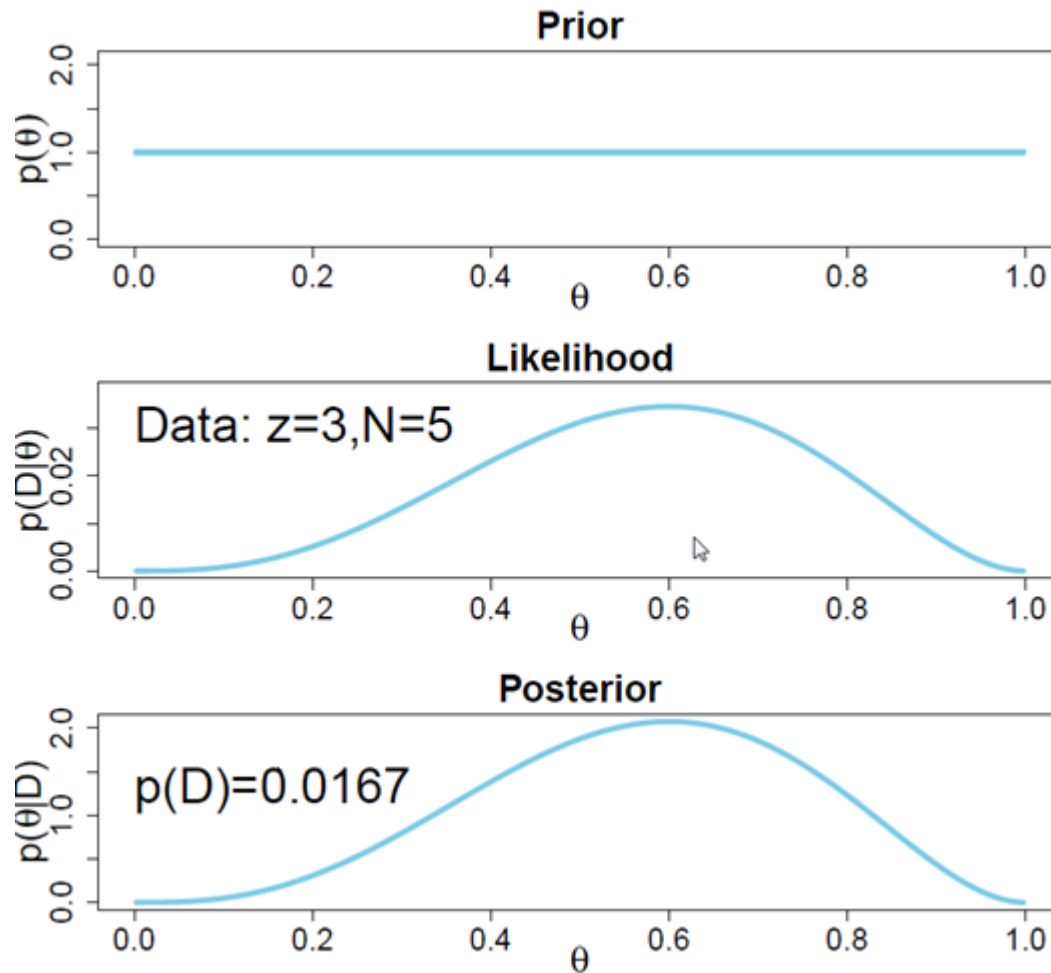
$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Poisson

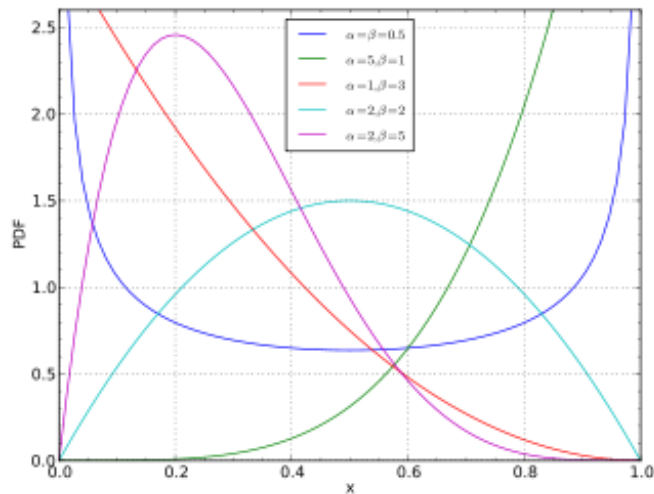


$$f(k; \lambda) = \Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!},$$

What form will suit us?



Beta Distribution



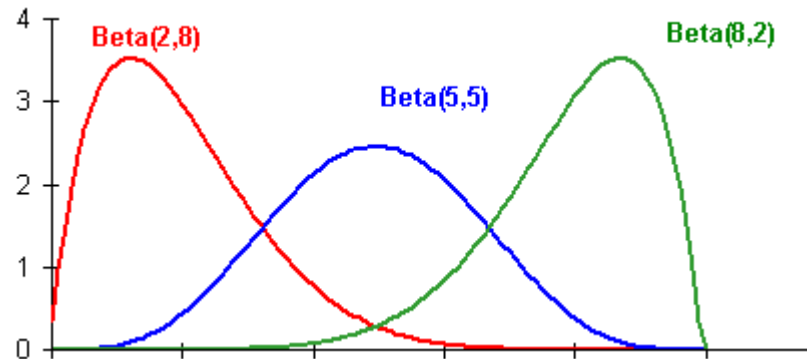
$$f(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1 - x)^{\beta-1}$$



Vadivelu, a famous Tamil comedian. This is one of his great expressions - terrified and confused.

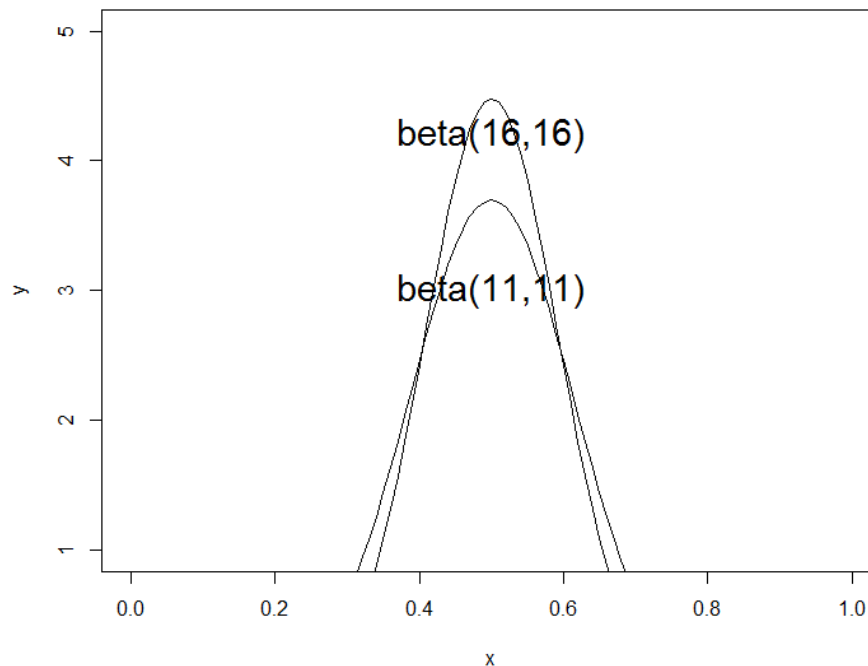
Beta Distribution

- Takes two parameters Beta(a, b)
- For now, assume $a = \#H + 1$ and $b = \#T + 1$.
- After 10 flips, we may end up in one of these three:



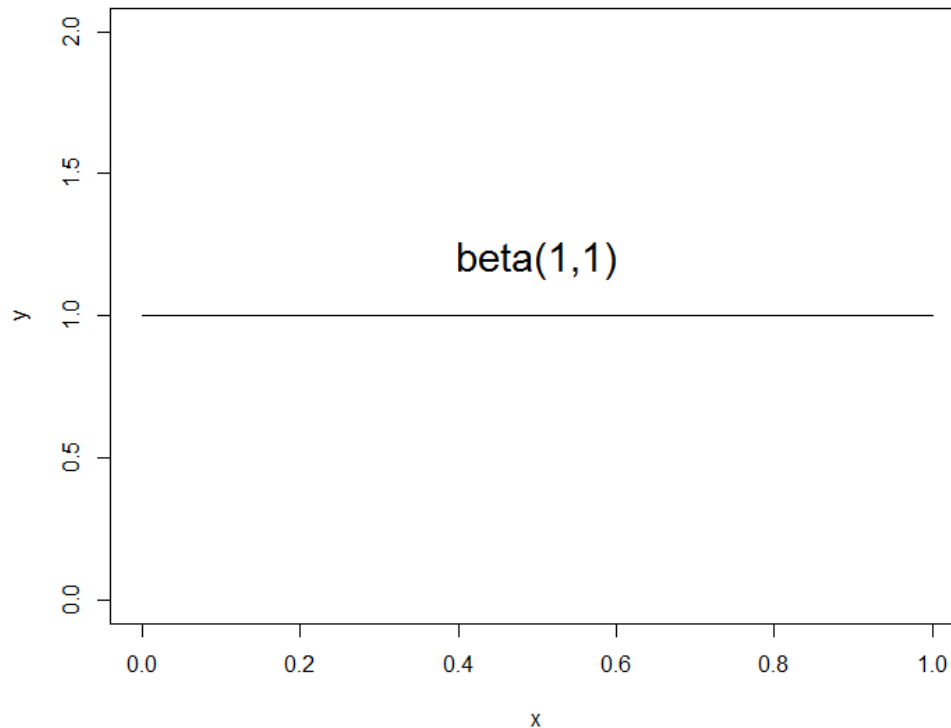
Prior and Posterior

Let's say we have a Strong Prior – Beta (11,11). What should happen if we see 10 more observations with 5H and 5T?



Prior and Posterior...

What if we have not seen any data?



Conjugate Prior

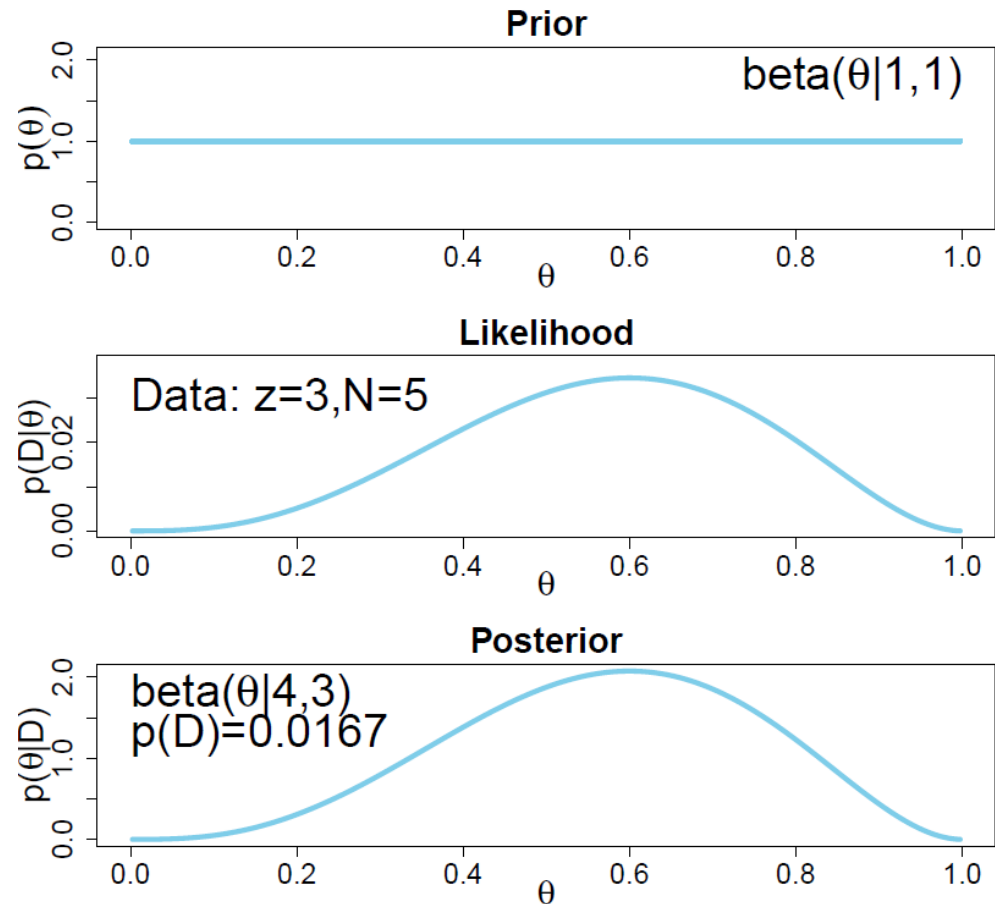
So, we see that:



Such Priors that have same form are called Conjugate Priors

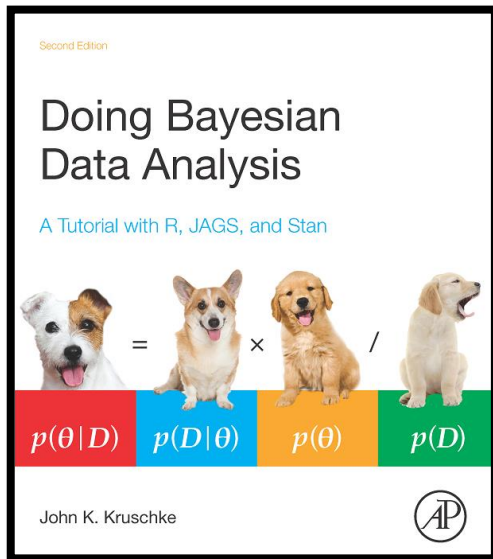
Summary

- ✓ Prior
- ✓ Likelihood
- ✓ Posterior
- ✓ Bayes' Rule
- ✓ Bernoulli Distribution
- ✓ Beta Distribution

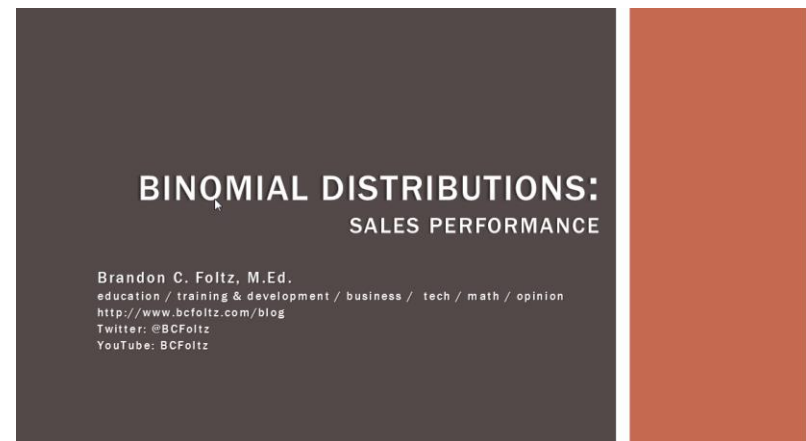


References

[Book on Doing Bayesian Data Analysis](#) – John K. Kruschke.



[YouTube Video on Statistics 101: The Binomial Distribution](#) – Brandon Foltz.



Maximum Likelihood Estimation

- An Observation
 - HHTT TTHH TTTT TTTT
- What values of $P(H)$ and $P(T)$ will maximize the probability of the above observation?
- $P(\dots\text{Observation}\dots) = x^4(1 - x)^{16}$
- For what value of x will this function maximize?

Probabilistic Retrieval

- Information Need: Taj Mahal
- Let a query q be “Taj”
- Let the results be:
 - d_1 : Taj
 - d_2 : Taj Mahal
 - d_3 : Taj Tea
- Two judges were asked to provide relevance judgments:

Document	Judge 1	Judge 2
Taj	R	N
Taj Mahal	R	R
Taj Tea	N	N

Probability of Relevance

- Documents can have probability of being relevant and of being non-relevant at the same time.
- Example:
 - Documents in our collection :

Document	$P(R=0 d, q)$	$P(R=1 d, q)$
Taj	0.5	0.5
Taj Mahal	?	?
Taj Tea	?	?

$R = 0 \rightarrow$ Non-Relevant

$R = 1 \rightarrow$ Relevant

Probability of Relevance

- Documents can have probability of being relevant and of being non-relevant at the same time.
- Example:
 - Documents in our collection :

Document	$P(R=0 d, q)$	$P(R=1 d, q)$
Taj	0.5	0.5
Taj Mahal	0	1
Taj Tea	1	0

$R = 0 \rightarrow$ Non-Relevant

$R = 1 \rightarrow$ Relevant

Probability Ranking Principle

**Rank documents by the
probability of relevance,
 $P(R=1 | q, d)$ $R \in \{0, 1\}$**

Probability Ranking Principle

**Rank documents by the probability of relevance,
 $P(R=1 | q, d)$ $R \in \{0, 1\}$**

Document	$P(R=0 d, q)$	$P(R=1 d, q)$
Taj	0.5	0.5
Taj Mahal	0	1
Taj Tea	1	0

$R = 0 \rightarrow$ Non-Relevant

$R = 1 \rightarrow$ Relevant

Search Result:

1. Taj Mahal
2. Taj
3. Taj Tea

Bayes Optimal Decision Rule

d is relevant if
 $P(R=1 | d,q) > P(R=0 | d,q)$

Document	$P(R=0 d,q)$	$P(R=1 d,q)$
Taj	0.5	0.5
Taj Mahal	0	1
Taj Tea	1	0

Search Result:
1. Taj Mahal

Predicting Relevance

Document	$P(R=0 d, q)$	$P(R=1 d, q)$
Taj	0.5	0.5
Taj Mahal	0	1
Taj Tea	1	0



This is user given relevance.

Can we estimate/predict relevance based on term occurrence ?

Predicting Relevance

- You may use labeled set from judges (or mined from clicklogs)
- You may assume query and document as set of words.

Query	Document	Relevance
q1 = (x1,x2,...)	d1 = (..xi, xj,...)	1
q1	d2	1
q1	d3	0
q2	d1	0
q2	d2	0
q2	d3	1



This is user given relevance.

Can we estimate/predict relevance ?

Binary Independence Model (BIM)

- Each document is a binary vector of terms.
- Occurrence of terms is mutually independent.

$$P(R = 1|d, q) = \frac{P(d|R = 1, q)P(R = 1|q)}{P(d|q)}$$

Bayes Rule

Quiz

$$P(R = 1|d, q) = \frac{P(d|R = 1, q)P(R = 1|q)}{P(d|q)}$$

Bayes Rule

- $P(d=\text{Taj} | R=1, q) = ?$
- $P(R=1 | q) = ?$
- $P(d | q) = ?$

Document	$P(R=0 d, q)$	$P(R=1 d, q)$
Taj	0.5	0.5
Taj Mahal	0	1
Taj Tea	1	0

Quiz

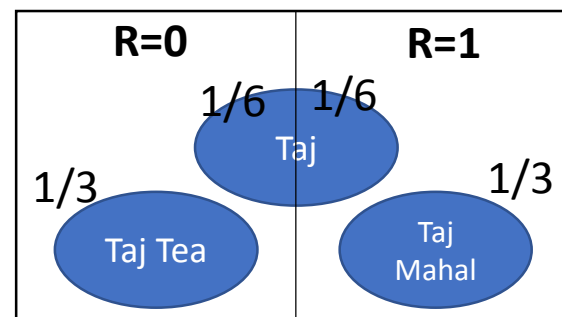
$$P(R = 1|d, q) = \frac{P(d|R = 1, q)P(R = 1|q)}{P(d|q)}$$

Bayes Rule

- $P(d=\text{Taj} | R=1, q) = 1/3$
- $P(R=1 | q) = 1/2$
- $P(d=\text{Taj} | q) = 1/3$

$$P(R=1 | d=\text{Taj}, q) = (1/3)(1/2)/(1/3) = 1/2$$

Document	$P(R=0 d, q)$	$P(R=1 d, q)$
Taj	0.5	0.5
Taj Mahal	0	1
Taj Tea	1	0



Predicting Relevance

- Odds of Relevance is easier to calculate

$$O(R|\vec{x}, \vec{q}) = \frac{P(R = 1|\vec{x}, \vec{q})}{P(R = 0|\vec{x}, \vec{q})} = \frac{\frac{P(R=1|\vec{q})P(\vec{x}|R=1,\vec{q})}{P(\vec{x}|\vec{q})}}{\frac{P(R=0|\vec{q})P(\vec{x}|R=0,\vec{q})}{P(\vec{x}|\vec{q})}} = \frac{P(R = 1|\vec{q})}{P(R = 0|\vec{q})} \cdot \frac{P(\vec{x}|R = 1, \vec{q})}{P(\vec{x}|R = 0, \vec{q})}$$

Constant term.

- In BIM, we assume that the term occurrence is mutually independent

$$\begin{aligned} \frac{P(\vec{x}|R = 1, \vec{q})}{P(\vec{x}|R = 0, \vec{q})} &= \prod_{t=1}^M \frac{P(x_t|R = 1, \vec{q})}{P(x_t|R = 0, \vec{q})} = \prod_{t:x_t=1} \frac{P(x_t = 1|R = 1, \vec{q})}{P(x_t = 1|R = 0, \vec{q})} \cdot \prod_{t:x_t=0} \frac{P(x_t = 0|R = 1, \vec{q})}{P(x_t = 0|R = 0, \vec{q})} \\ &= \prod_{t:x_t=q_t=1} \frac{p_t}{u_t} \cdot \prod_{t:x_t=0, q_t=1} \frac{1 - p_t}{1 - u_t} \end{aligned}$$

Thank You