https://vvtesh.sarahah.com/

Information Retrieval

Venkatesh Vinayakarao

Term: Aug – Sep, 2019
Chennai Mathematical Institute

So much of life, it seems to me, is determined by pure randomness.
– Sidney Poitier.
God does not play dice with the universe.
- Einstein.

The Law – Robert M. Coates

From the book, “The World of Mathematics – Volume IV”.

Triborough Bridge, NY, USA.
(aka Robert F. Kennedy Bridge)

Late 1940s, NY: *No other bridge or main highway was affected, and though the two preceding nights had been equally balmy and moonlit, on both of these the bridge traffic had run close to* ***normal****.*

And then, one day…

It just looked as if ***everybody*** *in Manhattan who owned a car had decided to drive out to Long Island that evening.*

No Reason!

Sergeant: “I kept askin’ them” he said, “Is there night football somewhere that we don’t know about? Is it the races you’re goin’ to?”

But the funny thing was half the time they’d be askin’ me. “What’s the crowd for, Mac?” they would say. And I’d just look at them.

If normal things stop happening, if we lose regularities in life, our planet could become unlivable!

Time for Action

- At this juncture, it was inevitable that Congress should be called on for action.

- Senator said, "*You can control it*". Re-education and reforms were decided upon. He said, (we need to lead people back to) "*the basic regularities, the homely* ***averageness*** *of the American way of life*".

The Law of Large Numbers

Known as the Fundamental theorem of Probability

The average of the results obtained from a large number of trials should be close to the expected value, and will tend to become closer as more trials are performed.

Expectation

- Roll a dice. Assume you may see 1 to 6 with equal probability.
- Expected Value = ?

$$\frac{1+2+3+4+5+6}{6} = 3.5$$

According to the law of large numbers, if a large number of six-sided dice are rolled, the average of their values (sometimes called the sample mean) is likely to be close to 3.5, with the precision increasing as more dice are rolled. – Wikipedia.

A Minor Digression

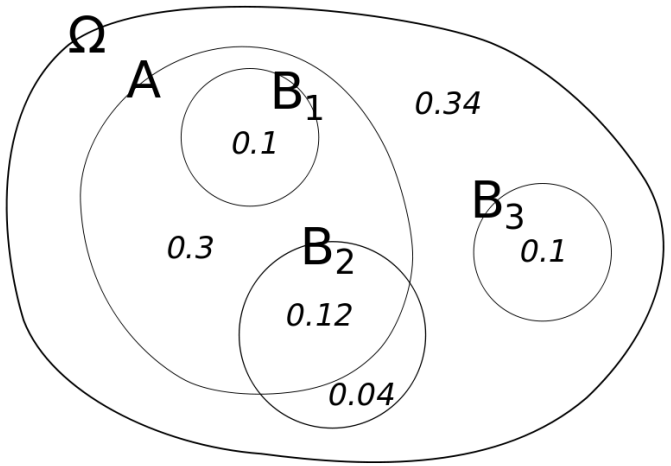
- What is the fundamental theorem of algebra?

Quiz

- What is the fundamental theorem of algebra?
  - Loosely, “Every polynomial has root(s)”
  - More Precisely, “*every non-constant single-variable polynomial with complex coefficients has at least one complex root*.” [Source: Wikipedia].

Conditional Probability

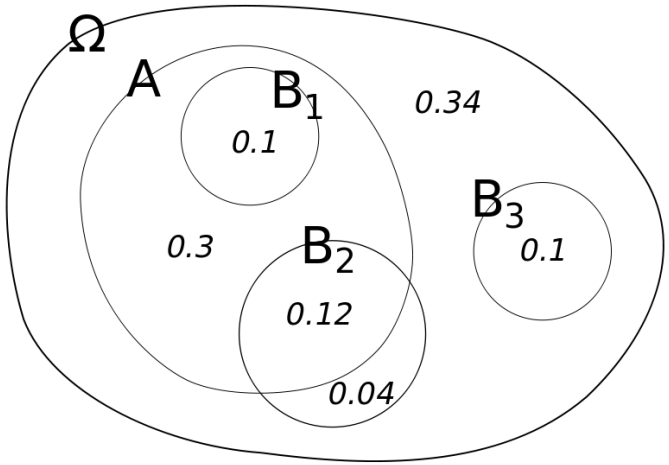
- $P(A) = 0.52$,
- $P(B_1) = 0.1$, and so on as shown below
- What is $P(A|B_1)$?
  - $P(A|B_1) = 1$.



Euler Diagram

Quiz: Conditional Probability

- What is $P(A|B_2)$?



Euler Diagram

Revisiting Probability

- Developers in two companies are distributed as follows. Compute Joint Probabilities.

	Java	C	Total
Company-X	1	17	18
Company-Y	37	20	57
Total	38	37	75

	Java	C	Total
Company-X	0.013	0.227	0.24
Company-Y	??	??	??
Total	0.506	??	??

P(Company-X,Java) = 1/75 = 0.013

Revisiting Probability

- Developers in two companies

	Java	C	Total
Company-X	1	17	18
Company-Y	37	20	57
Total	38	37	75

	Java	C	Total
Company-X	0.013	0.227	0.24
Company-Y	0.493	0.267	0.76
Total	0.506	0.494	1

- Joint Probability P(Company-X,Java) = 0.013.
- P(Company-Y, Java) = 0.493
- Sometimes written as P(AB) or P(A ∩ B)

Revisiting Probability

- Developers in two companies

	Java	C	Total
Company-X	1	17	18
Company-Y	37	20	57
Total	38	37	75

	Java	C	Total
Company-X	0.013	0.227	0.24
Company-Y	0.493	0.267	0.76
Total	0.506	0.494	1

- P(Company-Y|Java) = ??
- Is P(Company-Y|Java) == P(Java|Company-Y) ?

Revisiting Probability

- Developers in two companies

	Java	C	Total
Company-X	1	17	18
Company-Y	37	20	57
Total	38	37	75

	Java	C	Total
Company-X	0.013	0.227	0.24
Company-Y	0.493	0.267	0.76
Total	0.506	0.494	1

- P(Company-Y|Java) = 37/38 = 0.974
- P(Java|Company-Y) = 37/57 = 0.649

Odds

- Odds, $O(A) = P(A)/P(A') = PA/(1-P(A))$

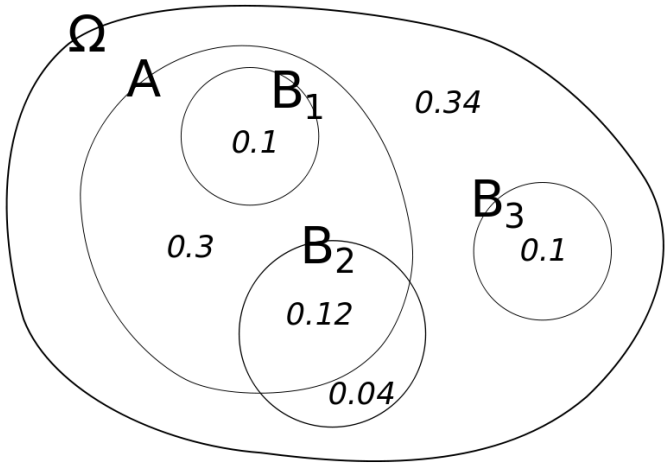
Quiz

- What is the probability of getting a 5 when rolling a six sided die? Assume a fair die.
- What is the odds of the same event?

Quiz: Conditional Probability

- What is $P(A|B_2)$?

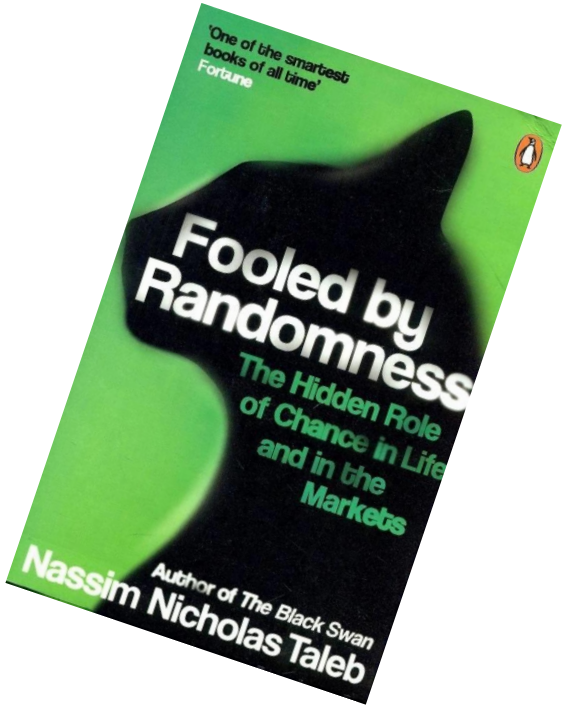
$P(A|B_2) = 0.12/(0.12 + 0.04) = 0.75.$



Euler Diagram

Reading

Probability and Computing - Eli Upfal and Michael Mitzenmacher



Fooled by Randomness - Nassim Nicholas Taleb

Thomas Bayes, 1701 to 1761

Bayesian Data Analysis and Beta Distribution

Venkatesh Vinayakarao

Agenda

Updating Beliefs using Probability Theory

Role of Beta Distributions

Will Discuss	Will not Discuss
✓ Concepts	⦸ Details
✓ Illustrations	⦸ Definitions
✓ Intuitions	⦸ Formalism
✓ Purpose	⦸ Derivations
✓ Properties	⦸ 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

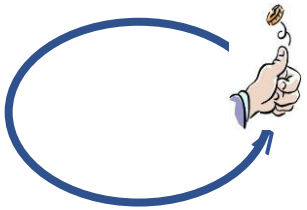
- Priors can be strong or weak

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.

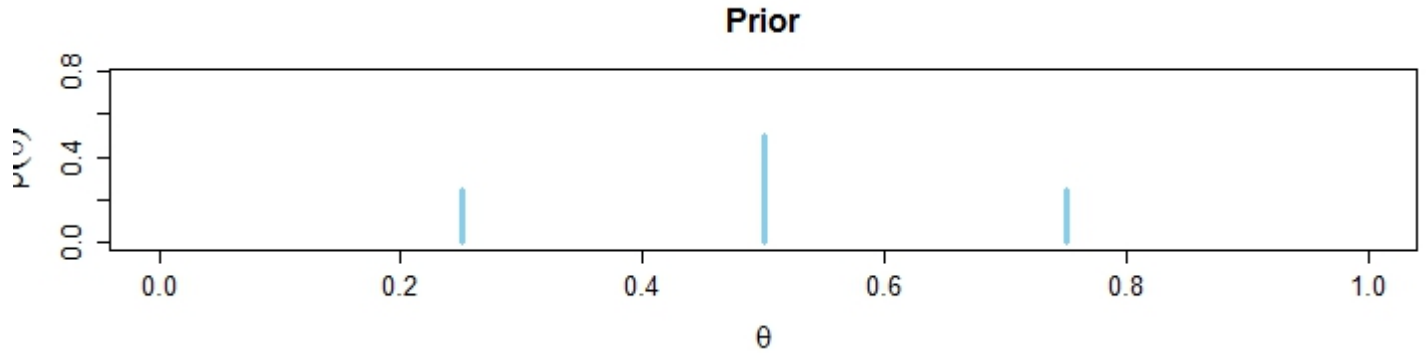
HyperParameter

- Prior probability (of Heads) could be anything:
  - O.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.

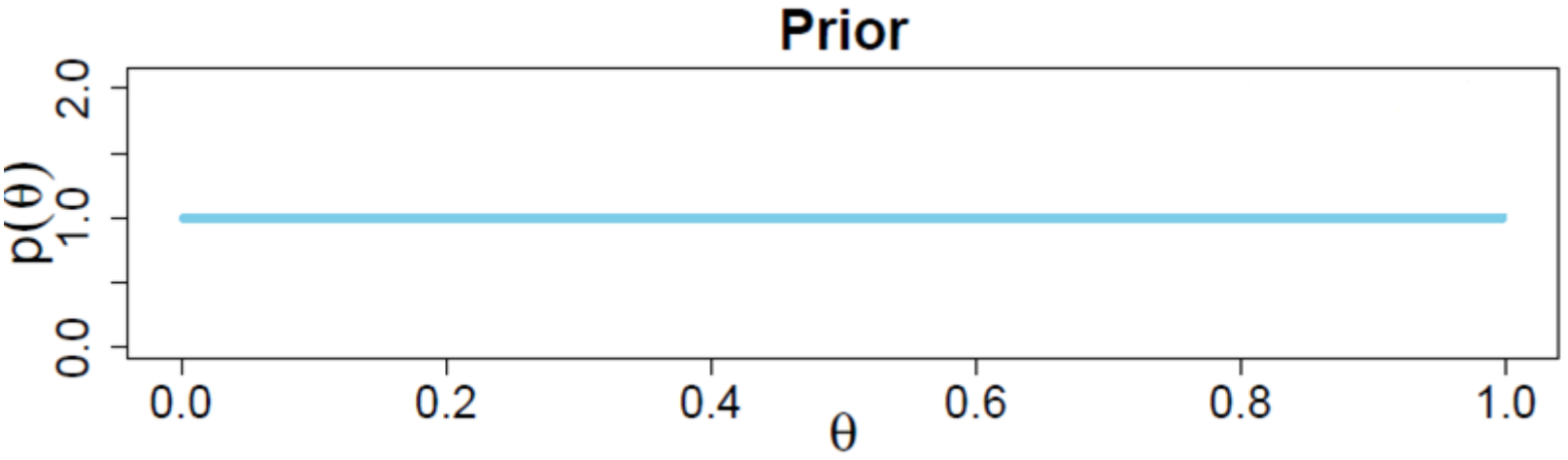
World of Distributions

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



Another Possibility

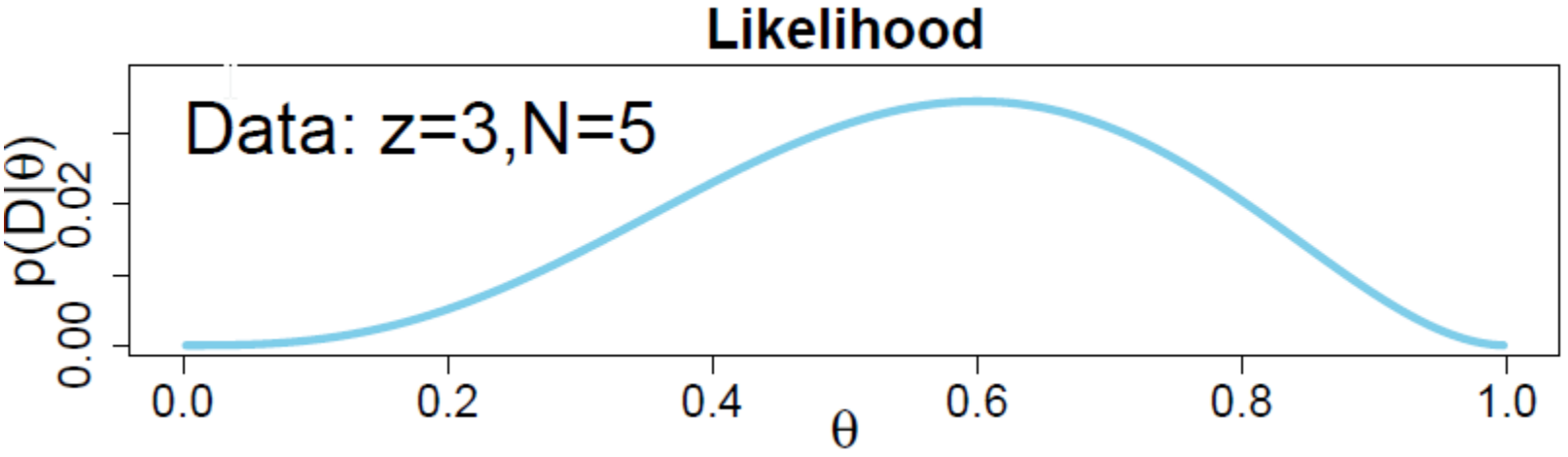
I may also choose to be unbiased! i.e., θ may take any value equally likely.



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 ≠ observation

Bayes' Rule

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

Eeks… what's in the denominator?

Numerator is easy

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

Jacob Bernoulli
1655 – 1705.

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

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

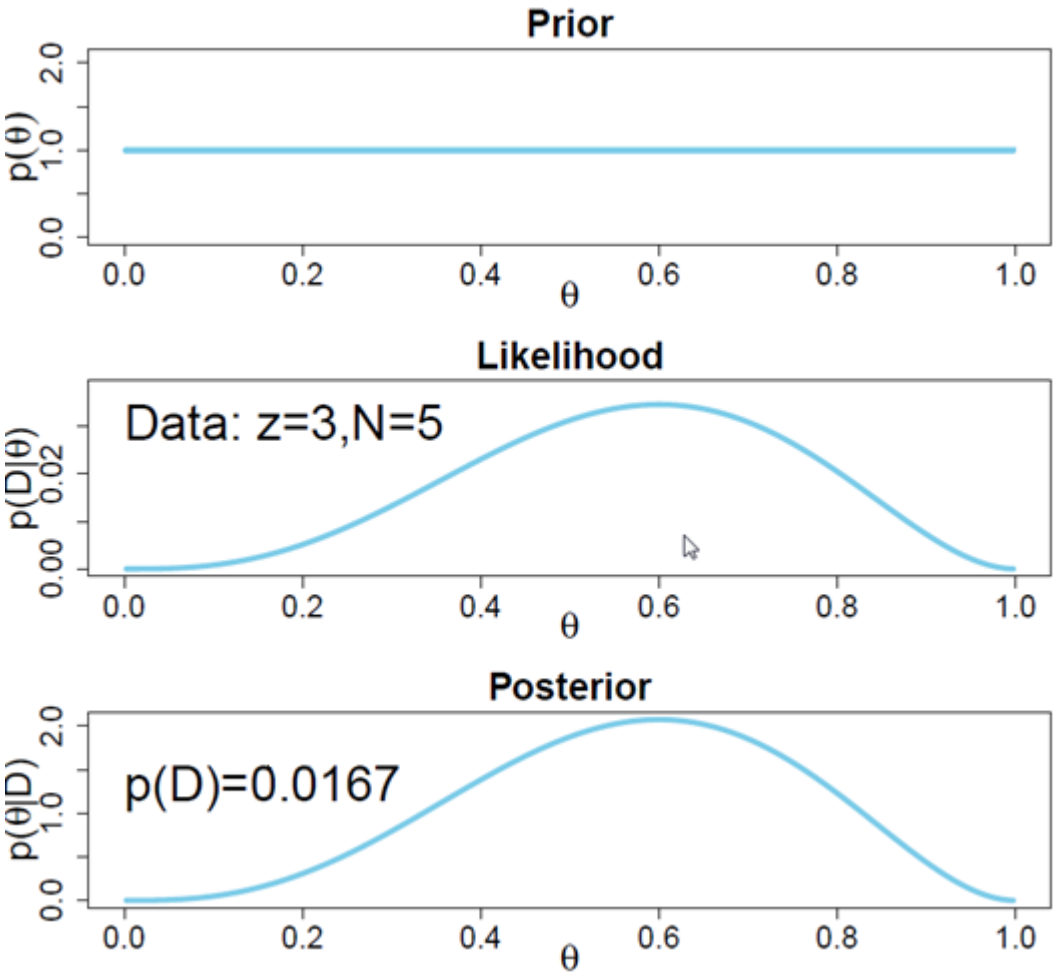
$$p(D|\theta) = (0.5)^{3}(1-0.5)^{5-3}$$

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:

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

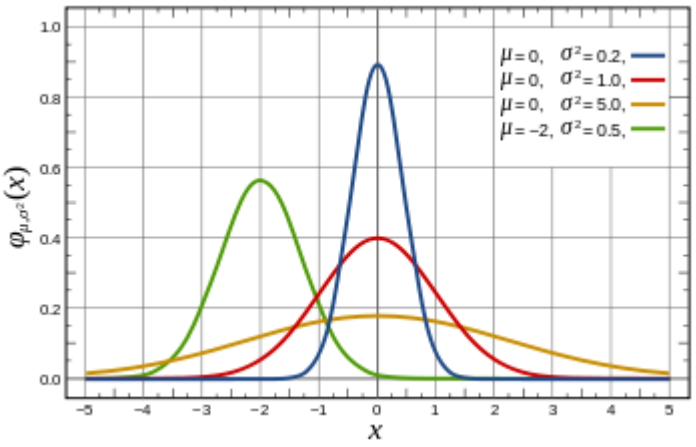
- And, for continuous distributions:

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

A Simpler Way

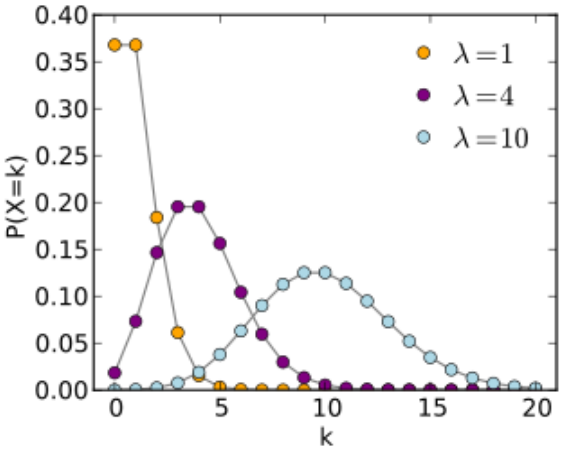
Form, Functions and Distributions

Normal (or Gaussian)



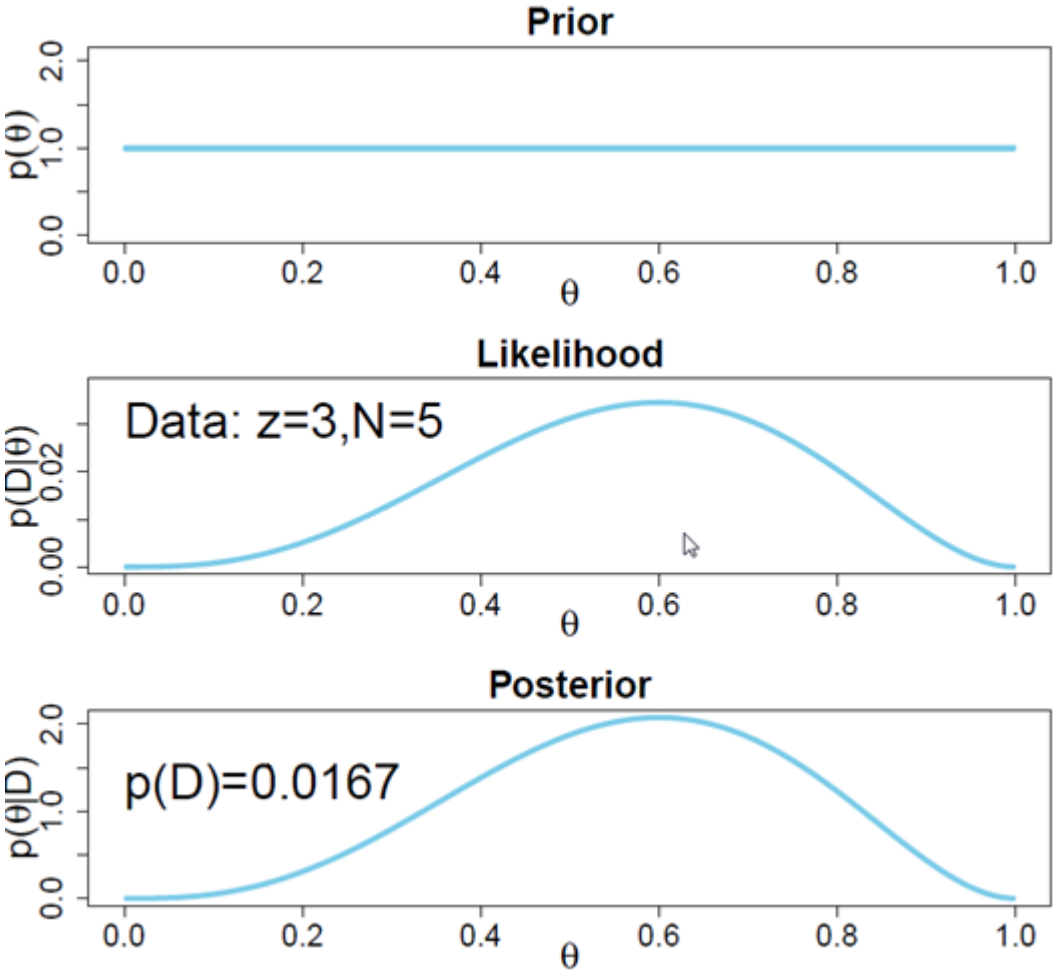
$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Poisson

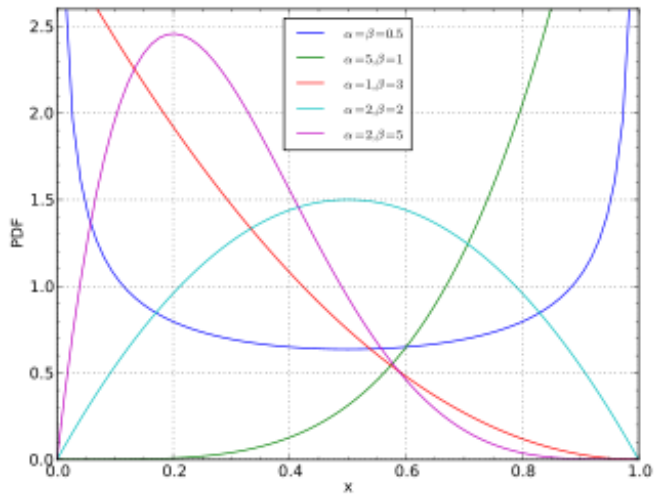


$$f(k; \lambda) = \Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!},$$

What form will suit us?



Beta Distribution

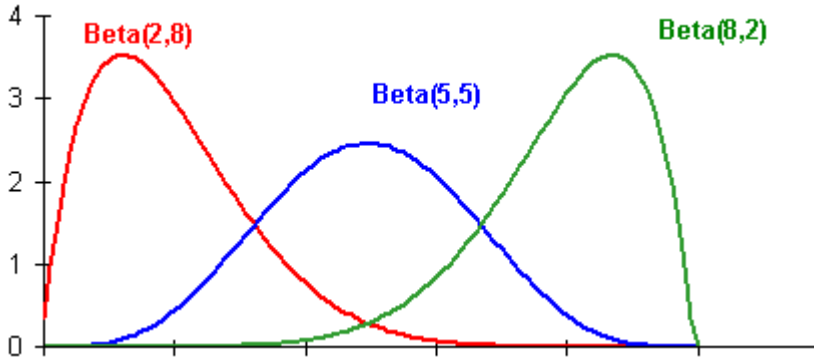


$$f(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, x^{\alpha-1}(1-x)^{\beta-1}$$

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

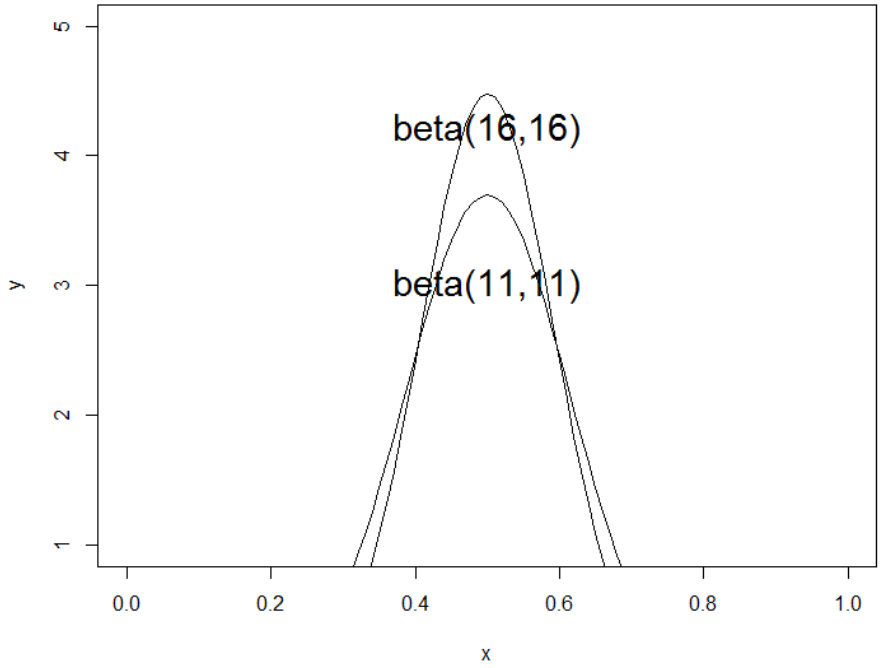
Beta Distribution

- 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:



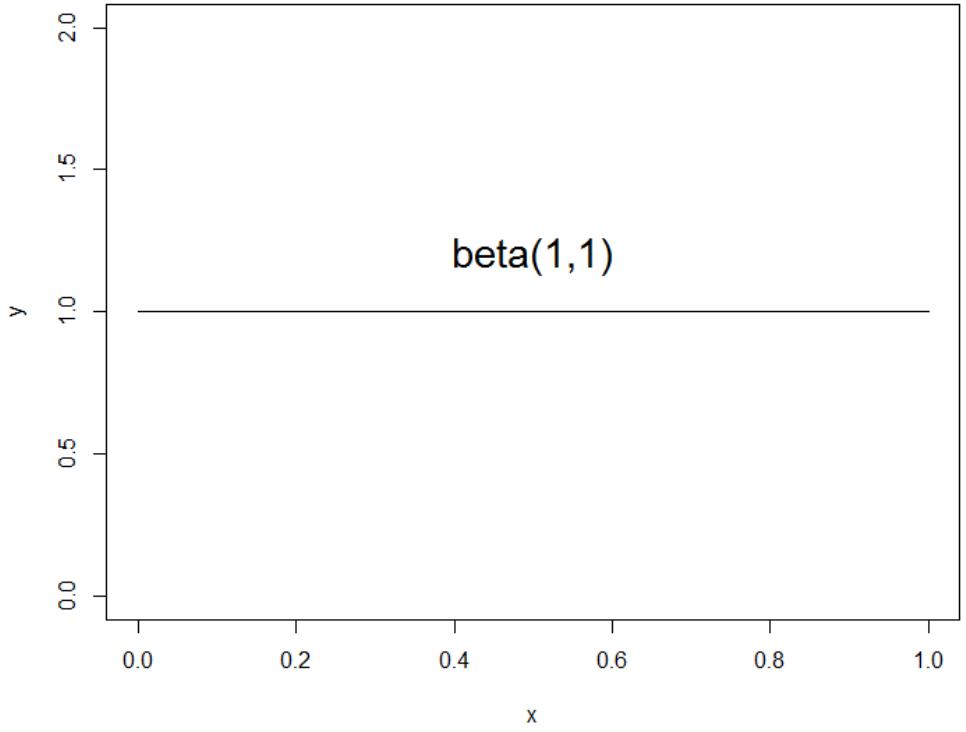
Prior and Posterior

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?



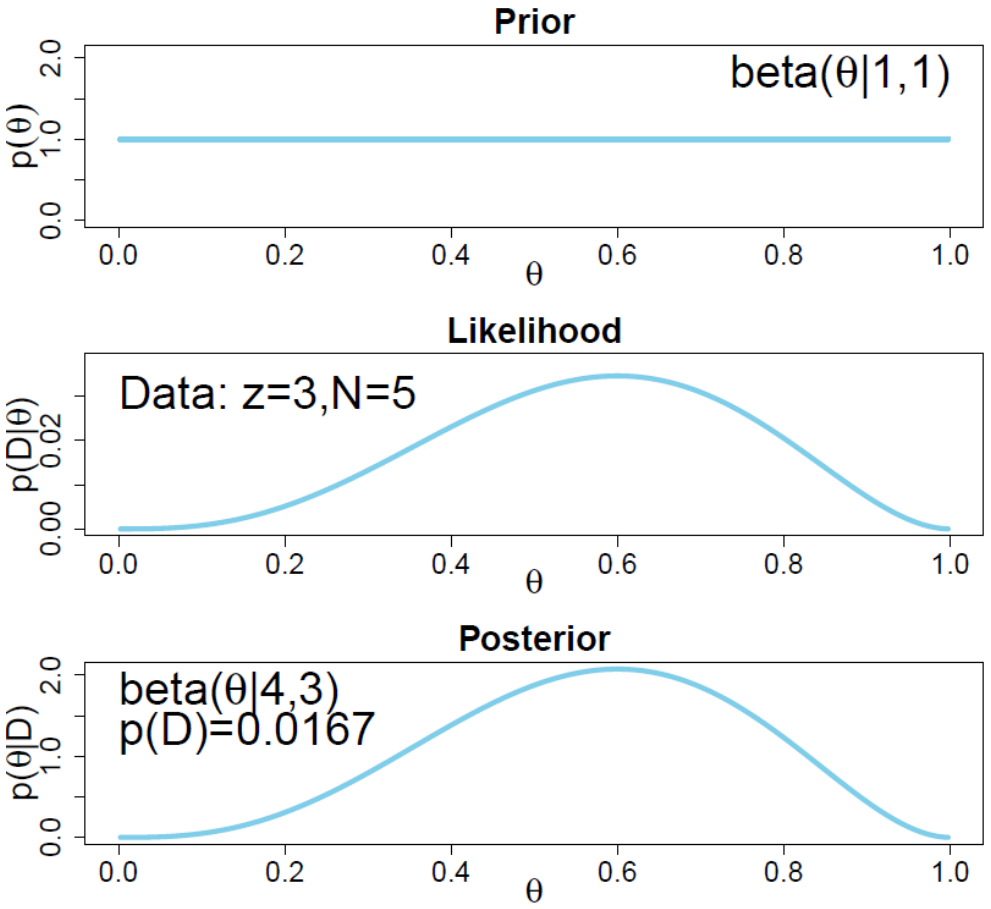
Conjugate Prior

So, we see that:

Such Priors that have same form are called Conjugate Priors

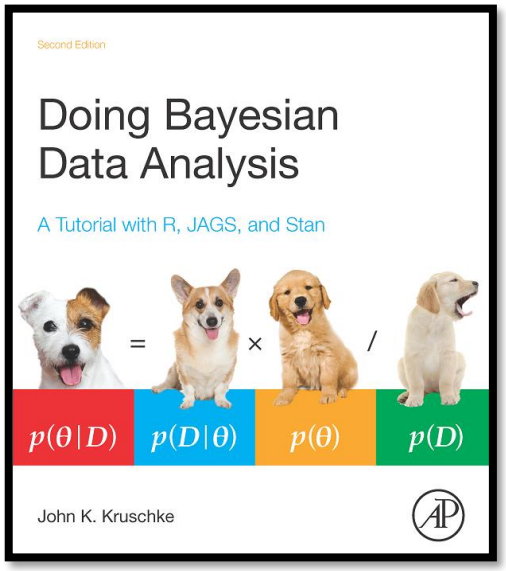
Summary

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

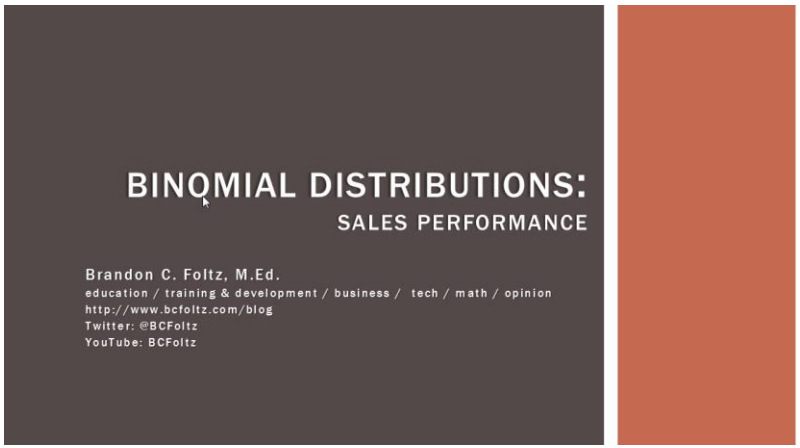


References

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(…Observation…) = $x^4(1-x)^{12}$
- 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:
  - d1: Taj
  - d2: Taj Mahal
  - d3: 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 ➔ Non-Relevant
R = 1 ➔ 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 ➔ Non-Relevant
R = 1 ➔ 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 ➔ Non-Relevant
R = 1 ➔ Relevant

Search Result:

1. Taj Mahal
2. Taj
3. Taj Tea

Bayes Optimal Decision Rule

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

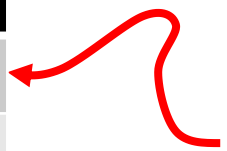
Document	P(R=0|d,q)	P(R=1|d,q)
Taj	0.5	0.5
Taj Mahal	0	1
Taj Tea	1	0

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=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

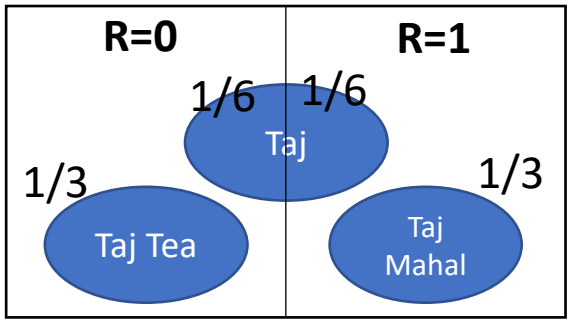
$$P(R = 1|d, q) = \frac{P(d|R = 1, q)P(R = 1|q)}{P(d|q)}$$

Bayes Rule

- P(d=Taj|R=1,q) = 1/3
- P(R=1|q) = 1/2
- P(d=Taj|q) = 1/3

P(R=1|d=Taj,q) = (1/3)(1/2)/(1/3) = ½

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

Constant term.

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

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

$$\frac{P(\vec{x}|R=1,\vec{q})}{P(\vec{x}|R=0,\vec{q})} = \prod_{t=1}^{M} \frac{P(x_t|R=1,\vec{q})}{P(x_t|R=0,\vec{q})} = \prod_{t:x_t=1} \frac{P(x_t=1|R=1,\vec{q})}{P(x_t=1|R=0,\vec{q})} \cdot \prod_{t:x_t=0} \frac{P(x_t=0|R=1,\vec{q})}{P(x_t=0|R=0,\vec{q})}$$

$$= \prod_{t:x_t=q_t=1} \frac{p_t}{u_t} \cdot \prod_{t:x_t=0,q_t=1} \frac{1-p_t}{1-u_t}$$

Retrieval Status Value

Not document specific.
Constant for a query.

$$\prod_{t:x_t=q_t=1} \frac{p_t(1-u_t)}{u_t(1-p_t)} \cdot \prod_{t:q_t=1} \frac{1-p_t}{1-u_t}$$

"We can manipulate this expression by including the query terms found in the document into the right product, but simultaneously dividing through by them in the left product, so the value is unchanged" - CPS.

$$\textbf{RSV} = \prod_{t:x_t=q_t=1} \frac{p_t(1-u_t)}{u_t(1-p_t)}$$

RSV is used for ranking documents.

Read Section 11.3.1 of CPS.

Thank You