Beta: The Random Variable
for Probabilities

Chris Piech
CS109, Stanford University
Which video are you more likely to like?

Davie504

👍 10,000  👎 50

Not Davie504

👍 10  👎 0
Philosophical Ponderings:
You ask about the probability of rain tomorrow.

**Person A:** My leg itches when it rains and it's kind of itchy.... Uh, $p = .80$

**Person B:** I have done complex calculations and have seen 10,451 days like tomorrow... $p = 0.80$

What is the difference between the two estimates?
“Those who are able to represent what they do not know make better decisions”
- CS109
Today we are going to learn something unintuitive, beautiful and useful
Pset 4 is out!

Here is the structure of the probabilistic model:

- **Risk Factors**
  - Tick Bite
  - Stress

- **Diseases**
  - Lyme
  - Flu

- **Symptoms**
  - Fever
  - Rash
  - Tired
  - Cough
  - Runny Nose
  - Headache

The input to your program is given to you via a constant, OBSERVATION. Your code should print out the probability that a person has Lyme disease given the observation. It should also print out the probability that a person has the Flu given the observation. The numeric answer to this problem validates the probability: $P(\text{Lyme} = 1 | \text{Fever} = 1, \text{Tick} = 1, \text{Cough} = 0)$, however your code should run rejection sampling for any input observation.

Specifically, your input is given to you via this constant:

1

Stanford University
Pset 4 is out!
Biometric Keystrokes

Did you know that computers can know who you are not, just by what you write, but also by how you write it? Coursera uses Biometric Keystroke signatures for plagiarism detection. If you can’t write a sentence with the same statistical distribution of key press timings as in your previous work, they assume that it is not you who is sitting behind the computer. In this problem we provide you with three files: 

- `ps4.txt`
- `personKeyTimingA.txt`
- `personKeyTimingB.txt`

`ps4.txt` has keystroke timing information for a user A writing a passage. The first column is the time in milliseconds (since the start of writing) when the user hit each key. The second column is the key that the user hit.

`personKeyTimingA.txt` has keystroke timing information for a second user (user B) writing the same passage as the user A. Even though the content of the passage is the same the timing of how the second user wrote the passage is different.

`email.txt` has keystroke timing information for an unknown user. We would like to know if the author of the email was user A or user B.

Let $X$ and $Y$ be random variables for the duration of time, in milliseconds, for users A and B (respectively) to type a key. Assume that each keystroke from a user has a duration that is an independent random variable with the same distribution.

Your Task: Calculate the ratio of the probability that user A wrote the email over the probability that user B wrote the email. To do so, first approximate $X$ and $Y$ as Normals with mean and variance that match their biometric data.

Explain your work and justify your answer.
Pset 4 is out!

Learning While Helping

You are designing a randomized algorithm that delivers one of two new drugs (which we call drug A and drug B) to patients who come to your clinic. Each patient can only receive one of the drugs. Initially you know nothing about the effectiveness of the two drugs. You are simultaneously trying to learn which drug is the best and, at the same time, cure the maximum number of people. To do so we will use the Thompson Sampling Algorithm.

Your job is to implement the \texttt{thompson\_sampling} function which will decide whether to give drug A or drug B, based on a limited history of observations.

Thompson Sampling Algorithm:

For each drug we maintain a Beta distribution to represent the drug’s probability of being successful. Our initial belief in the probability of success is uniform for both drug A and drug B: \( \theta \sim \text{Beta}(1, 1) \).

When choosing which drug to give to the next patient we sample a value from the Beta representing drug A, and we sample a value from the Beta representing drug B. We select the drug with the largest sampled value. We administer the drug, observe if the patient was cured, and update the Beta that represents our belief about the probability of the drug being successful.

Running a single game with 10 trials
Your choice: A, Success? = False
Your choice: A, Success? = False
Your choice: A, Success? = False
Your choice: A, Success? = True
Your choice: A, Success? = True
Your choice: A, Success? = False
Your choice: A, Success? = False
Your choice: A, Success? = True
True probabilities: ‘A’ = 0.500, ‘B’ = 0.087
total successes for your algorithm: 4
total successes for the oracle implementation: 1
New Features

Go into “Focus Mode”  
Class wide resource list

Public Resources:

No resources added yet

Resource Title
URL
Add Resource
Coverage. You are ready!

Probabilistic Models

Today!
Review
Let M be a **discrete** random variable

Let N be a **discrete** random variable

\[ P(M = 2|N = 3) = \frac{P(N = 3|M = 2)P(M = 2)}{P(N = 3)} \]

\[ P(M = m|N = n) = \frac{P(N = n|M = m)P(M = m)}{P(N = n)} \]

More generally

Shorthand notation

\[ P(m|n) = \frac{P(n|m)P(m)}{P(n)} \]
Inference on a non-bernoulli random variable

We can perform inference when there are two random variables using Bayes!

\[ P(A = a) \quad \text{Observation } Y = 0 \quad P(A = a | Y = 0) \]
Number or Dictionary?

\[ P(A = a | Y = 0) = \frac{P(Y = 0 | A = a)P(A = a)}{P(Y = 0)} \]

\[
\text{belief} = \frac{0.02 \times 0.001}{0.02} = 0.001
\]

Value of a \( \uparrow \)

\( P(A=a) \)
Inference on a non-bernoulli random variable

In plain English: run bayes for each value of a

\[
P(A = a|Y = 0) = \frac{P(Y = 0|A = a)P(A = a)}{P(Y = 0)}
\]

# RV bayes as code

def update(belief, obs):
    for a in support:
        prior_a = belief[a]
        likelihood = calc_likelihood(a, obs)
        belief[a] = prior_a * likelihood
    normalize(belief)
Normalize???

def update(belief, obs):
    for a in support:
        prior_a = belief[a]
        likelihood = calc_likelihood(a, obs)
        belief[a] = prior_a * likelihood
    normalize(belief)

\[
P(A = a | Y = 0) = \frac{P(Y = 0 | A = a)P(A = a)}{P(Y = 0)} \\
= \frac{P(Y = 0 | A = a)P(A = a)}{\sum_a P(Y = 0, A = a)} \\
= \frac{P(Y = 0 | A = a)P(A = a)}{\sum_a P(Y = 0 | A = a)P(A = a)}
\]

In plain English: this is the sum of all the things in belief
End Review
Where are we in CS109?

Overview of Topics

- Counting Theory
- Core Probability
- Random Variables
- Probabilistic Models
- Uncertainty Theory
- Machine Learning
Let's play a game!

Roll a dice three times. If I roll a six twice (or more) I win $1 million. Otherwise you win $1 million. What should we charge to play?

\[ P(W) = \left( \frac{5}{6} \right)^2 \approx 0.69 \]
What if you don't know a probability?
What if you don't know a probability?
Pirate Supply Store, San Francisco
What is your belief that you successfully roll a 6 on my die?
The parameter $p$ to a binomial can be a random variable
9 Heads out of 10 Flips. What is your Belief in $p$?

$$p = \frac{9}{10}$$
9 Heads out of 10 Flips. What is your Belief in $p$?

Let $X$ be our belief about the probability of heads:

$$P(X = x | H = 9, T = 1) = \frac{P(H = 9, T = 1 | X = x) f(X = x)}{P(H = 9, T = 1)}$$

Binomial

Uniform?
9 Heads out of 10 Flips. What is your Belief in $p$?

Let $X$ be our belief about the probability of heads:

$$P(X = x|H = 9, T = 1) = \frac{P(H = 9, T = 1|X = x)f(X = x)}{P(H = 9, T = 1)} = \frac{\binom{10}{9}x^9(1 - x)^1}{P(H = 9, T = 1)}$$
9 Heads out of 10 Flips. What is your Belief in p?

Let $X$ be our belief about the probability of heads:

$$P(X = x|H = 9, T = 1) = \frac{P(H = 9, T = 1|X = x)f(X = x)}{P(H = 9, T = 1)}$$

$$= \binom{10}{9} x^9 (1-x)^1$$

$$= K \cdot x^9 (1-x)^1$$
9 Heads out of 10 Flips. What is your Belief in $p$?

$$P(X = x | H = 9, T = 1)$$
Flip a coin with unknown probability

Flip a coin \((n + m)\) times, comes up with \(n\) heads

- We don’t know probability \(X\) that coin comes up heads

Frequentist (never prior)

\[
X = \lim_{n+m \to \infty} \frac{n}{n + m} \\
\approx \frac{n}{n + m}
\]

Bayesian (prior is great)

\[
f_{X|N}(x|n) = \\
P(N = n|X = x) f_X(x) \\
P(N = n)
\]

\(X\) is (often) a single value

\(X\) is a random variable. Leads to a belief distribution which captures confidence
Flip a coin with unknown probability!

Flip a coin \((n + m)\) times, comes up with \(n\) heads
- We don’t know probability \(X\) that coin comes up heads
- Our belief before flipping coins is that: \(X \sim \text{Uni}(0, 1)\)
- Let \(N = \text{number of heads}\)
- Given \(X = x\), coin flips independent: \((N | X) \sim \text{Bin}(n+m, x)\)

\[
f_{X|N}(x|n) = \frac{P(N = n|X = x)f_X(x)}{P(N = n)}
\]

Bayesian "posterior" probability distribution

Bayesian "prior" probability distribution
Flip a coin with unknown probability!

Flip a coin \((n + m)\) times, comes up with \(n\) heads

- We don’t know probability \(X\) that coin comes up heads heads
- Our belief before flipping coins is that: \(X \sim \text{Uni}(0, 1)\)
- Let \(N = \) number of heads
- Given \(X = x\), coin flips independent: \((N \mid X) \sim \text{Bin}(n + m, x)\)

\[
\begin{align*}
  f_{X \mid N}(x \mid n) &= \frac{P(N = n \mid X = x) f_X(x)}{P(N = n)} \\
  &= \frac{(n+m)_n x^n (1-x)^m}{P(N = n)} \\
  &= \frac{(n+m)_n}{P(N = n)} x^n (1-x)^m \\
  &= \frac{1}{c} \cdot x^n (1-x)^m \quad \text{where } c = \int_0^1 x^n (1-x)^m \, dx
\end{align*}
\]
If you start with a $X \sim \text{Uni}(0, 1)$ prior over probability, and observe:

$n$ “successes” and
$m$ “failures”…

Your new belief about the probability is:

$$f_X(x) = \frac{1}{c} \cdot x^n(1 - x)^m$$

where

$$c = \int_0^1 x^n(1 - x)^m$$
Belief after 7 success and 1 fail

\[ f_X(x) = \frac{1}{c} \cdot x^n (1 - x)^m \]

\( n = 7 \)

\( m = 1 \)
Equivalently!

If you start with a $X \sim \text{Uni}(0, 1)$ prior over probability, and observe:

let $a = \text{num "successes"} + 1$

let $b = \text{num "failures"} + 1$

Your new belief about the probability is:

$$f_X(x) = \frac{1}{c} \cdot x^{a-1}(1 - x)^{b-1}$$

where

$$c = \int_{0}^{1} x^{a-1}(1 - x)^{b-1}$$
X is a **Beta Random Variable**: \( X \sim \text{Beta}(a, b) \)

- Probability Density Function (PDF): (where \( a, b > 0 \))
  \[
  f(x) = \begin{cases} 
  \frac{1}{B(a,b)} x^{a-1} (1-x)^{b-1} & 0 < x < 1 \\
  0 & \text{otherwise}
  \end{cases} 
  \]

\[
B(a,b) = \int_0^1 x^{a-1} (1-x)^{b-1} \, dx
\]

- Symmetric when \( a = b \)
  \[
  E[X] = \frac{a}{a+b} \\
  Var(X) = \frac{ab}{(a+b)^2(a+b+1)}
  \]
Beta is the Random Variable for Probabilities

Used to represent a distributed belief of a probability
Beta Parameters can come from experiments:

\[ a = \text{“successes”} + 1 \]
\[ b = \text{“failures”} + 1 \]
Think about the difference between a **point estimate** and a **distribution**

\[ p = 0.75 \]
Beta is a distribution for probabilities. Its range is values between 0 and 1.
Beta Parameters *can* come from experiments:

\[ a = \text{"successes"} + 1 \]
\[ b = \text{"failures"} + 1 \]
If the Prior was Beta?

X is our random variable for probability

If our prior belief about X was beta

\[ f(X = x) = \frac{1}{B(a, b)} x^{a-1} (1 - x)^{b-1} \]

What is our posterior belief about X after observing \( n \) heads (and \( m \) tails)?

\[ f(X = x | N = n) = ??\]
If the Prior was Beta?

\[ f(X = x|N = n) = \frac{P(N = n|X = x)f(X = x)}{P(N = n)} \]

\[ = \frac{(n+m) x^n (1-x)^m f(X = x)}{P(N = n)} \]

\[ = K_1 \cdot \binom{n+m}{n} x^n (1-x)^m \frac{1}{B(a,b)} x^{a-1} (1-x)^{b-1} \]

\[ = K_3 \cdot x^n (1-x)^m x^{a-1} (1-x)^{b-1} \]

\[ = K_3 \cdot x^{n+a-1} (1-x)^{m+b-1} \]

\[ X|N \sim \text{Beta}(n + a, m + b) \]
If “Prior” distribution of X (before seeing flips) is Beta

Then “Posterior” distribution of X (after flips) is Beta

Beta is a **conjugate** distribution for Beta

- Prior and posterior parametric forms are the same!
- Practically, conjugate means easy update:
  - Add number of “heads” and “tails” seen to Beta parameters
A beta understanding

Can set \( X \sim \text{Beta}(a, b) \) as prior to reflect how biased you think coin is apriori

- This is a subjective probability (aka Bayesian)!
- Prior probability for \( X \) based on seeing \((a + b - 2)\) “imaginary” trials, where
  - \((a - 1)\) of them were heads.
  - \((b - 1)\) of them were tails.
- \( \text{Beta}(1, 1) = \text{Uni}(0, 1) \rightarrow \) we haven’t seen any “imaginary trials”, so apriori know nothing about coin

Update to get posterior probability

- \( X \mid (n \text{ heads and } m \text{ tails}) \sim \text{Beta}(a + n, b + m) \)
A beta understanding

\[ X \mid (N = n, M = m) \sim \textbf{Beta}(a = n + 1, b = m + 1) \]

- Prior \( X \sim \textbf{Uni}(0, 1) \)
- Check this out, boss:
  - \( \textbf{Beta}(a = 1, b = 1) =? \)
  
  \[ f(x) = \frac{1}{B(a,b)} x^{a-1} (1-x)^{b-1} = \frac{1}{B(a,b)} x^0 (1-x)^0 = \frac{1}{\int_0^1 1 \, dx} = 1 \text{ where } 0 < x < 1 \]
  
  - \( \textbf{Beta}(a = 1, b = 1) = \textbf{Uni}(0, 1) \)
- So, prior \( X \sim \textbf{Beta}(a = 1, b = 1) \)
Let $X$ be the probability of rolling a “6” on Chris’ die.

Prior: Imagine 5 die rolls where only showed up as a “6”

Observation: Roll it a few times...

What is the updated probability density function of $X$ after our observations?
Check out the Demo!
Damn
A beta example

Before being tested, a medicine is believed to “work” about 80% of the time. The medicine is tried on 20 patients. It “works” for 14 and “doesn’t work” for 6. What is your new belief that the drug works?

Frequentist:

\[
p \approx \frac{14}{20} = 0.7
\]
**A beta example**

Before being tested, a medicine is believed to “work” about 80% of the time. The medicine is tried on 20 patients. It “works” for 14 and “doesn’t work” for 6. What is your new belief that the drug works?

<table>
<thead>
<tr>
<th>Prior: $X \sim \text{Beta}(a = 81, b = 21)$</th>
<th>Interpretation: 80 successes / 100 trials</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X \sim \text{Beta}(a = 9, b = 3)$</td>
<td>8 successes / 10 trials</td>
</tr>
<tr>
<td>$X \sim \text{Beta}(a = 5, b = 2)$</td>
<td>4 successes / 5 trials</td>
</tr>
</tbody>
</table>
Before being tested, a medicine is believed to “work” about 80% of the time. The medicine is tried on 20 patients. It “works” for 14 and “doesn’t work” for 6. What is your new belief that the drug works?

**Bayesian:**

**Prior:**

\[ X \sim \text{Beta}(a = 5, b = 2) \]

**Posterior:**

\[ X \sim \text{Beta}(a = 5 + 14, b = 2 + 6) \]

\[ \sim \text{Beta}(a = 19, b = 8) \]

**Evaluating the posterior distribution:**

\[ E[X] = \frac{a}{a + b} = \frac{19}{19 + 8} \approx 0.70 \]

**Mode of the posterior distribution:**

\[ \text{mode}(X) = \frac{a - 1}{a + b - 2} = \frac{19}{18 + 7} \approx 0.72 \]
Laplace Smoothing

Prior: \( X \sim \text{Beta}(a = 2, b = 2) \)

One imagined heads

One imagined tail

Fancy name. Simple prior
Which video are you more likely to like?

👍 10,000 👎 50

👍 10 👎 0
Which video are you more likely to like?

10,000

50

10

0

Beta PDF

Beta PDF
Next level?
Alpha GO mixed deep learning and core reasoning under uncertainty
Multi Armed Bandit
Multi Armed Bandit

Drug A

Drug B

Which one do you give to a patient?
Let's Play!

Drug A

Drug B

Which one do you give to a patient?
import pickle
import random

def main():
    X1, X2 = pickle.load(open('probs.pkl', 'rb'))

    print("Welcome to the drug simulator. There are two drugs")

    while True:
        choice = getChoice()
        prob = X1 if choice == "a" else X2
        success = bernoulli(prob)
        if success:
            print('Success. Patient lives!')
        else:
            print('Failure. Patient dies!')
        print('')
You try drug B, 5 times. It is successful 2 times.
If you had a uniform prior, what is your posterior belief about the likelihood of success?

\[ X \sim \text{Beta}(a = 3, b = 4) \]
You try drug B, 5 times. It is successful 2 times. 

$X$ is the probability of success.

$$X \sim \text{Beta}(a = 3, b = 4)$$

What is expectation of $X$?

$$E[X] = \frac{a}{a + b} = \frac{3}{3 + 4} \approx 0.43$$
You try drug B, 5 times. It is successful 2 times. 

$X$ is the probability of success.

$$X \sim \text{Beta}(a = 3, b = 4)$$

What is the probability that $X > 0.6$

$$P(X > 0.6) = 1 - P(X < 0.6) = 1 - F_X(0.6)$$

Wait what? Chris are you holding out on me?

```
stats.beta.cdf(x, a, b)
```

$$P(X > 0.6) = 1 - F_X(0.6) = 0.1792$$
Explore something new? Or go for what looks good now?
One option: Upper Confidence Bound
Amazing option: Thompson Sampling
Beta:
The probability density for probabilities
Beta is a distribution for probabilities
If you start with a $X \sim \text{Uni}(0, 1)$ prior over probability, and observe:

- let $a = \text{num “successes”} + 1$
- let $b = \text{num “failures”} + 1$

Your new belief about the probability is:

$$f_X(x) = \frac{1}{c} \cdot x^{a-1}(1 - x)^{b-1}$$

where $$c = \int_0^1 x^{a-1}(1 - x)^{b-1}$$
Distributions

- **Binomial**
- **Geometric**
- **Exponential**
- **Poisson**
- **Beta**
- **Neg Binomial**
- **Normal**
- **Uniform**
Think about the difference between a **point estimate** and a **distribution**

\[ p = 0.75 \]

\[ p = \]
Problem with a point estimate:

**Person A:** My leg itches when it rains and it's kind of itchy.... Uh, $p = .80$

**Person B:** I have done complex calculations and have seen 10,451 days like tomorrow... $p = 0.80$

Give me the uncertainty!!!
Any parameter for a “parameterized” random variable can be thought of as a random variable.

Eg:

\[ P(\Lambda = \lambda | N = 5) \]