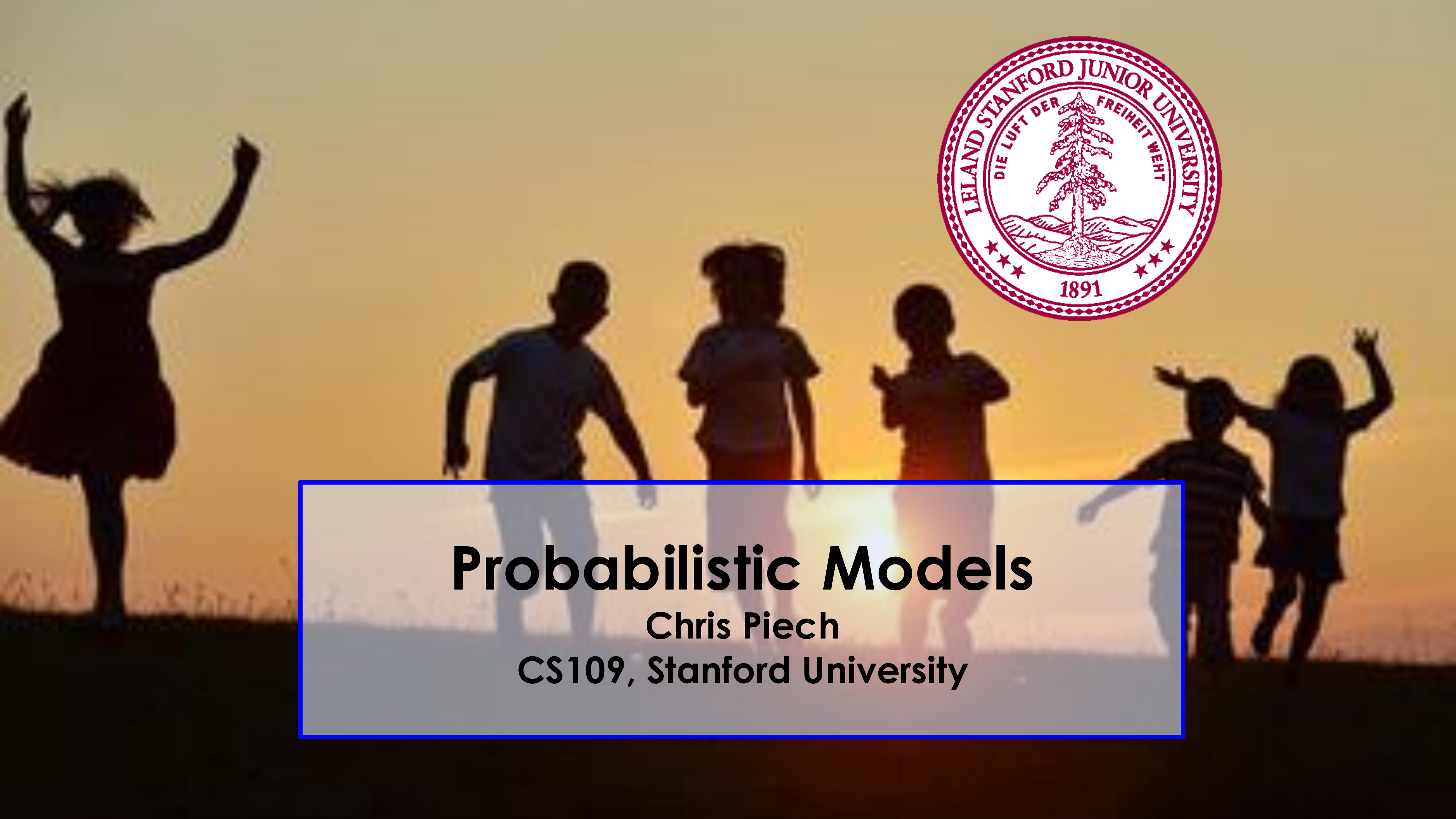
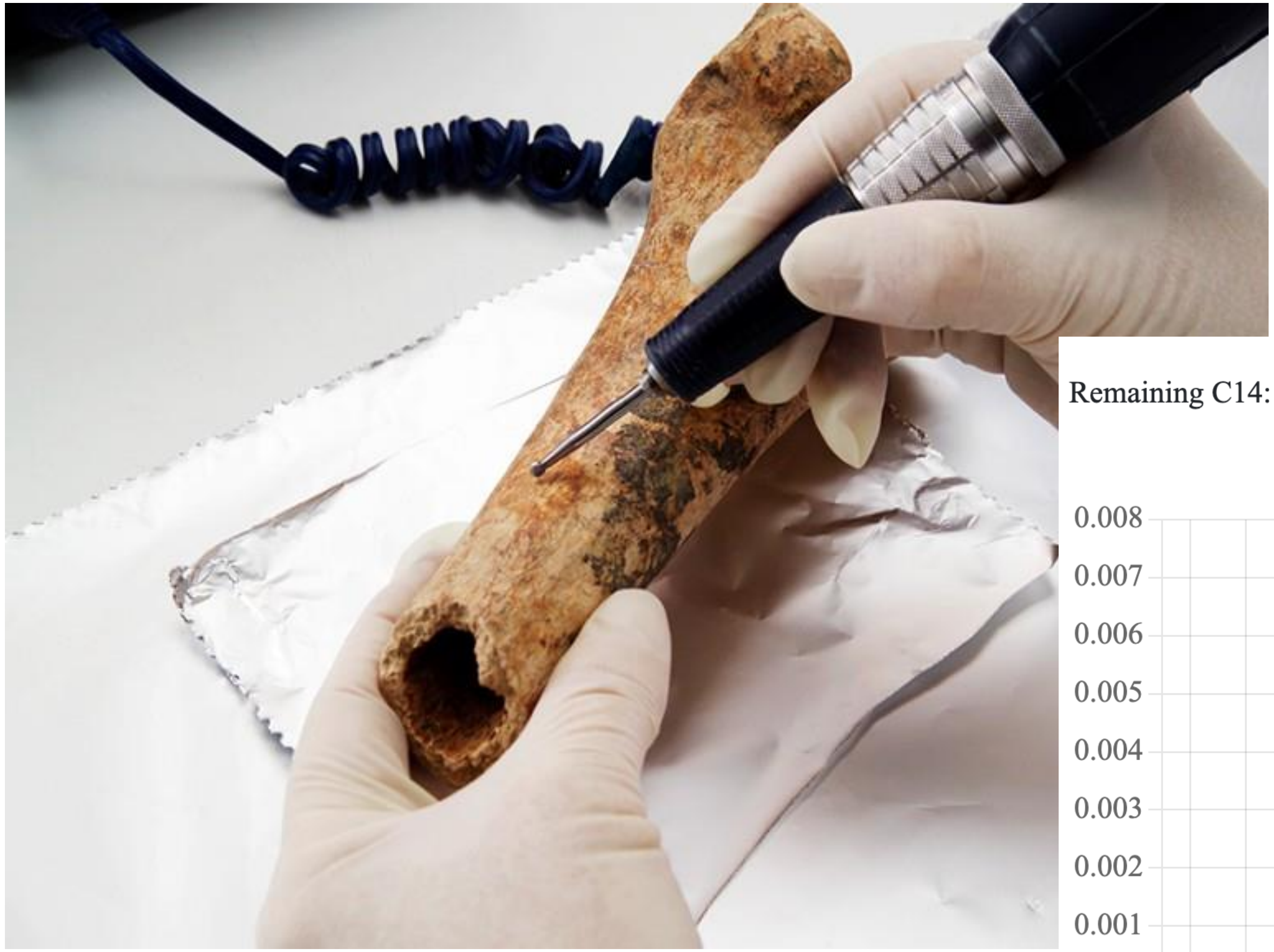
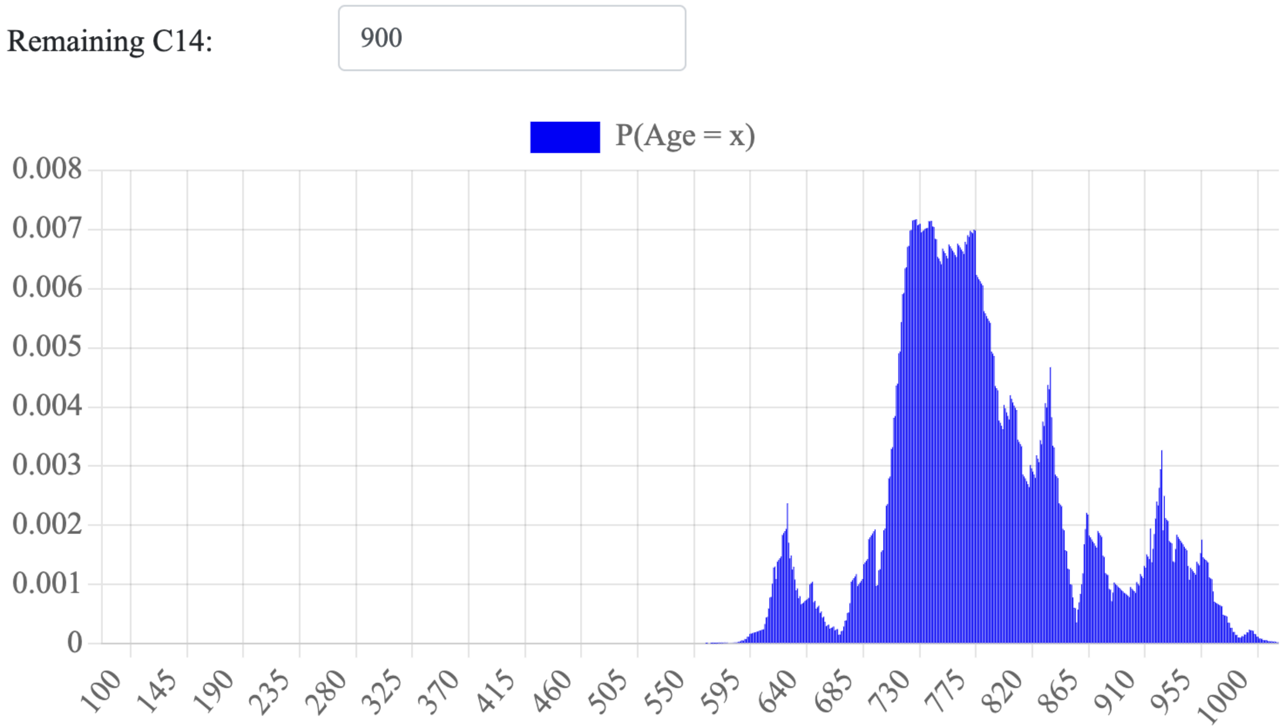


Probabilistic Models

Chris Piech
CS109, Stanford University

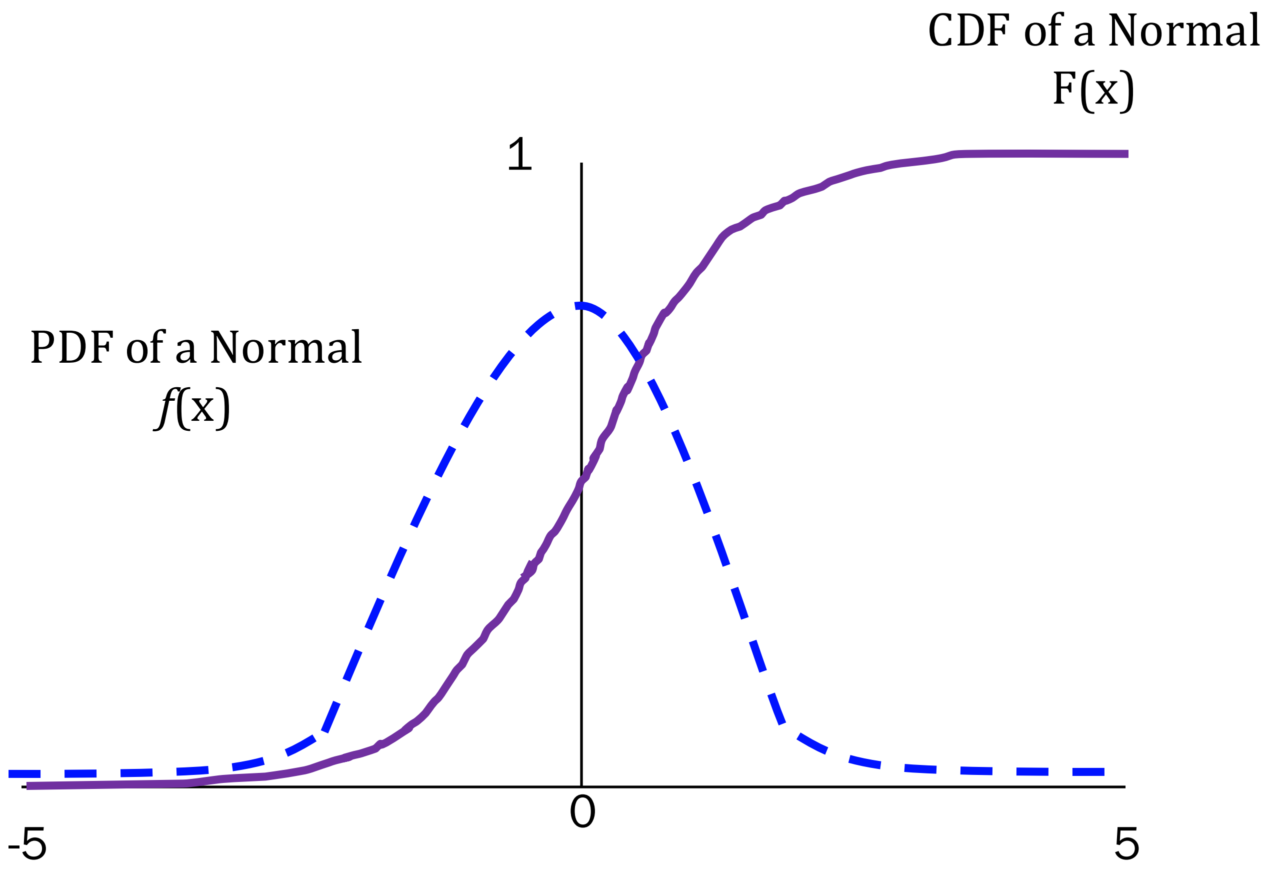
Bayesian Carbon Dating





First, some review

Normal Distribution



f(x) = derivative of probability

F(x) = P(X < x)

Cumulative Distribution Function (CDF)

$$\mathcal{N}(\mu, \sigma^2)$$

CDF of Standard Normal: A function that has been solved for numerically

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

The cumulative density function (CDF) of any normal

Table of $\Phi(z)$ values are precomputed

Probability Density Function (PDF)

$$\mathcal{N}(\mu, \sigma^2)$$

the distance to the mean

"exponential"

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

probability density at *x*

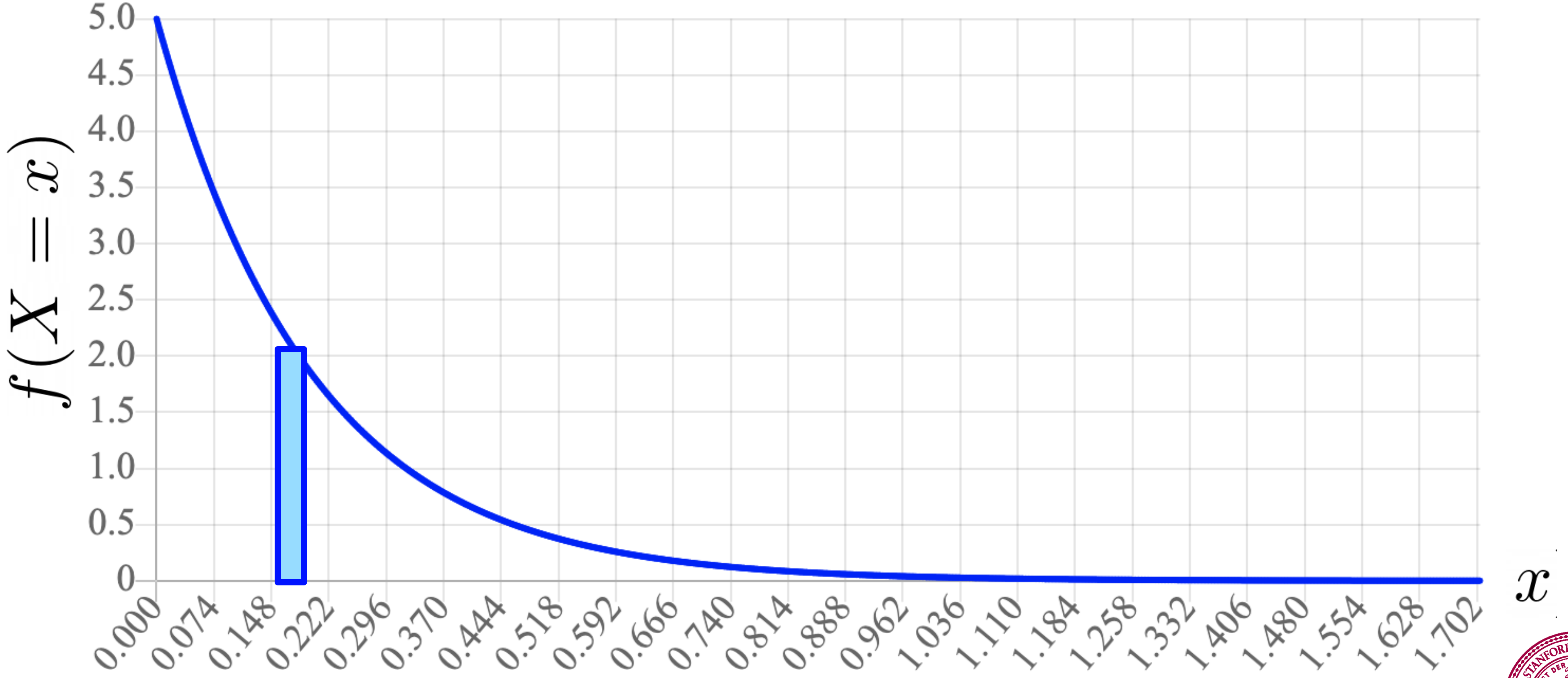
a constant

sigma shows up twice

Great Question

Epsilon: Useful perspective

$$P(X = x) = \boxed{f(X = x) \cdot \epsilon_x}$$

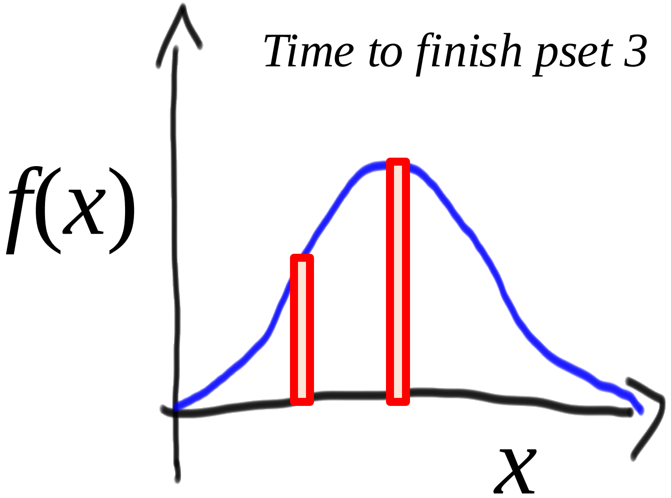


* In the limit as ϵ_x goes to zero

Relative Probability of Continuous Variables

X = time to finish pset 3

X ~ N(μ = 10, σ^2 = 2)



How much more likely are you to complete pset 3 in 10 hours than in 5?

$$\frac{P(X=10)}{P(X=5)} = \frac{\varepsilon f(X=10)}{\varepsilon f(X=5)}$$

$$= \frac{f(X=10)}{f(X=5)}$$

$$= \frac{\frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(10-\mu)^2}{2\sigma^2}}}{\frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(5-\mu)^2}{2\sigma^2}}}$$

$$= \frac{\frac{1}{\sqrt{4\pi}} e^{-\frac{(10-10)^2}{4}}}{\frac{1}{\sqrt{4\pi}} e^{-\frac{(5-10)^2}{4}}}$$

$$= \frac{e^0}{e^{-\frac{25}{4}}} = 518$$

End of review

Where are we in CS109?

Overview of Topics

Counting Theory | Core Probability | Random Variables | Probabilistic Models | Uncertainty Theory | Machine Learning

CS109

Machine Learning

Uncertainty Theory

Single Random Variables

Probabilistic Models

Counting

Probability Fundamentals

[suspense]

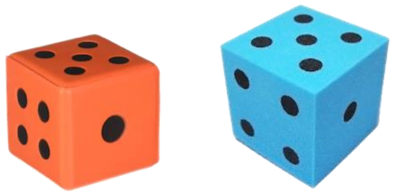
Discrete Probabilistic Models

The world is full of interesting probability problems

Have multiple random variables interacting with one another

Joint probability mass functions

Roll two 6-sided dice, yielding values X and Y.



$$X$$

random variable

$$P(X = 1)$$

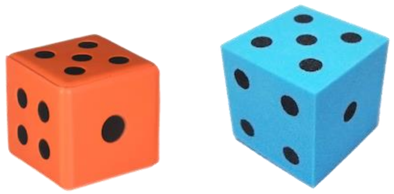
probability of an event

$$P(X = k)$$

probability mass function

Joint probability mass functions

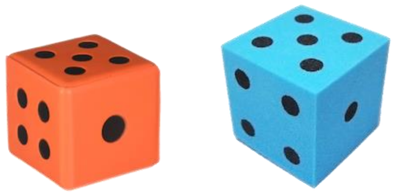
Roll two 6-sided dice, yielding values X and Y.



X random variable	$P(X = 1)$ probability of an event	$P(X = k)$ probability mass function
X, Y random variables	$P(X = 1 \text{ and } Y = 6)$ $P(X = 1, Y = 6)$ recall: the comma probability of the intersection of two events	$P(X = a, Y = b)$ joint probability mass function

Two dice

Roll two 6-sided dice, yielding values X and Y.



1. What is the joint PMF of X and Y?

$$P(X = a, Y = b) = 1/36 \qquad (a, b) \in \{(1,1), \dots, (6,6)\}$$

		X					
		1	2	3	4	5	6
Y	1	1/36	…	…	…	…	1/36
	2	…	…	…	…	…	…
	3	…	…	…	…	…	…
	4	…	…	…	…	…	…
	5	…	…	…	…	…	…
	6	1/36	…	…	…	…	1/36

$P(X = 4, Y = 3)$

Probability table
- All possible outcomes for several discrete RVs
- Not parametric (e.g., parameter p in Ber(p))

Dating at Stanford. Data from a few years ago

	Single	In a relationship	It's complicated
Freshman	0.13	0.08	0.02
Sophomore	0.17	0.11	0.02
Junior	0.09	0.10	0.02
Senior	0.02	0.07	0.01
5+	0.06	0.09	0.04

Joint is Complete Information!

	Single	Relationship	Complicated
Frosh	0.13	0.08	0.02
Soph	0.17	0.11	0.02
Junior	0.09	0.10	0.02
Senior	0.02	0.07	0.01
5+	0.06	0.09	0.04

A joint distribution is complete information. It can be used to answer any probability question.

Joint table: mutually exclusive and covers sample space.

	Single	Relationship	Complicated
Frosh	0.13	0.08	0.02
Soph	0.17	0.11	0.02
Junior	0.09	0.10	0.02
Senior	0.02	0.07	0.01
5+	0.06	0.09	0.04

Each combination is mutually exclusive, and they span the sample space

$$\sum_{x \in X} \sum_{y \in Y} P(x, y) = 1$$

X is dating status.
Y is year.

Joint table: mutually exclusive and covers sample space.

	Single	Relationship	Complicated
Frosh	0.13	0.08	0.02
Soph	0.17	0.11	0.02
Junior	0.09	?	0.02
Senior	0.02	0.07	0.01
5+	0.06	0.09	0.04

Each combination is mutually exclusive, and they span the sample space

$$\sum_{x \in X} \sum_{y \in Y} P(x, y) = 1$$

X is dating status.
Y is year.

Joint table: mutually exclusive and covers sample space.

	Single	Relationship	Complicated
Frosh	0.13	0.08	0.02
Soph	0.17	0.11	0.02
Junior	0.09	0.10	0.02
Senior	0.02	0.07	0.01
5+	0.06	0.09	0.04

Each combination is mutually exclusive, and they span the sample space

$$\sum_{x \in X} \sum_{y \in Y} P(x, y) = 1$$

X is dating status.
Y is year.

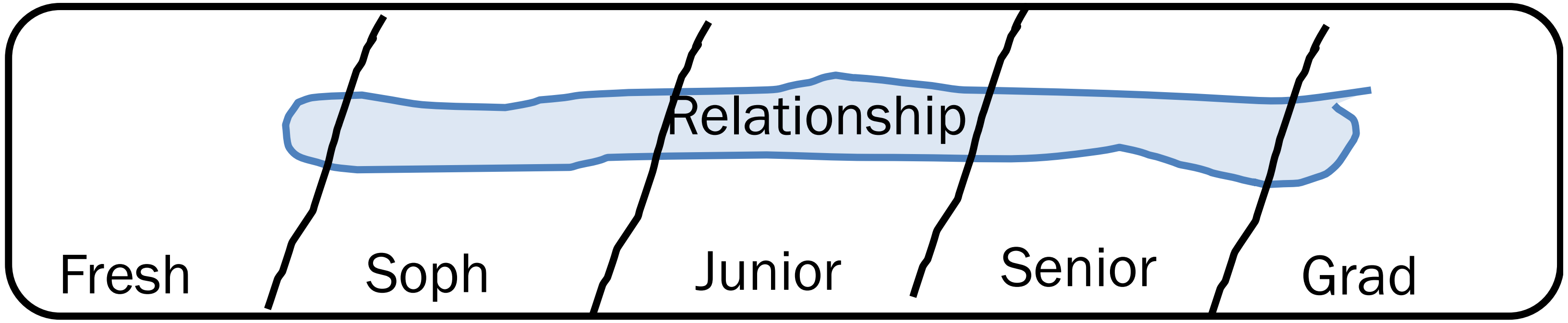
What is the probability someone is in a relationship?

	Single	Relationship	Complicated
Frosh	0.13	0.08	0.02
Soph	0.17	0.11	0.02
Junior	0.09	0.10	0.02
Senior	0.02	0.07	0.01
5+	0.06	0.09	0.04

We can use the law of total probability!

X is dating status. Y is year.

$$P(X = \text{relation}) = \sum_{y \in Y} P(X = \text{relation}, Y = y)$$



What is the probability someone is in a relationship?

	Single	Relationship	Complicated
Frosh	0.13	0.08	0.02
Soph	0.17	0.11	0.02
Junior	0.09	0.10	0.02
Senior	0.02	0.07	0.01
5+	0.06	0.09	0.04

We can use the law of total probability!

X is dating status. Y is year.

$$P(X = \text{relation}) = \sum_{y \in Y} P(X = \text{relation}, Y = y)$$

$$P(X = \text{single}) = \sum_{y \in Y} P(X = \text{single}, Y = y)$$

$$P(Y = \text{frosh}) = \sum_{x \in X} P(X = x, Y = \text{frosh})$$

$$P(Y = \text{soph}) = \sum_{x \in X} P(X = x, Y = \text{soph})$$

Welcome the marginal

Marginal Distribution

For two discrete joint random variables X and Y,
the **joint probability mass function** is defined as:

$$P(X = a, Y = b)$$

The **marginal distributions** of the joint PMF are defined as:

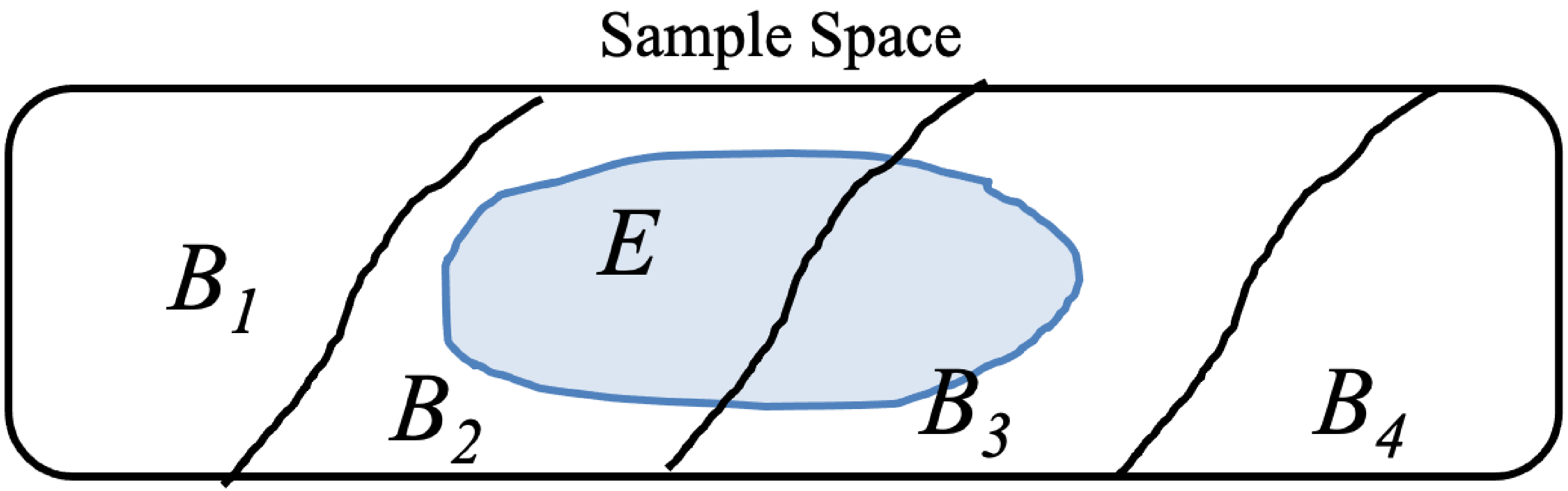
$$P(X = a) = \sum_y P(X = a, Y = y)$$

$$P(Y = b) = \sum_x P(X = x, Y = b)$$

Use marginal distributions to get a 1-D RV from a joint PMF.

Marginal Distribution. Law of Total Probability for RVs

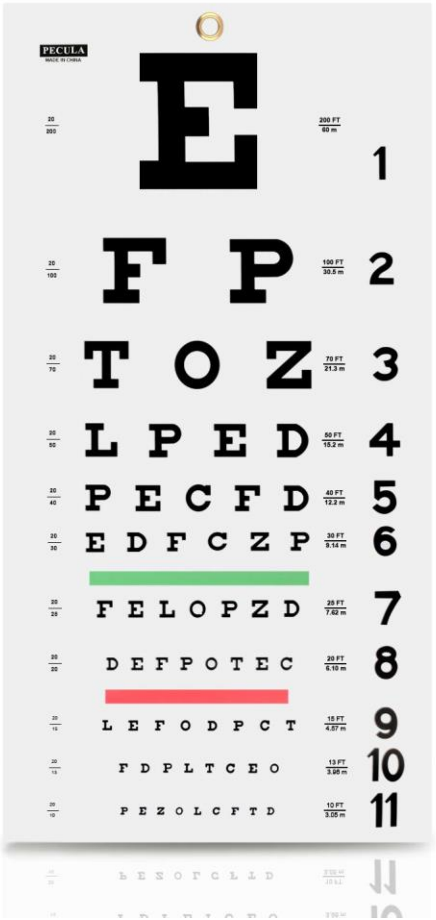
$$P(\underline{X = a}) = \sum_y P(X = a, Y = y)$$



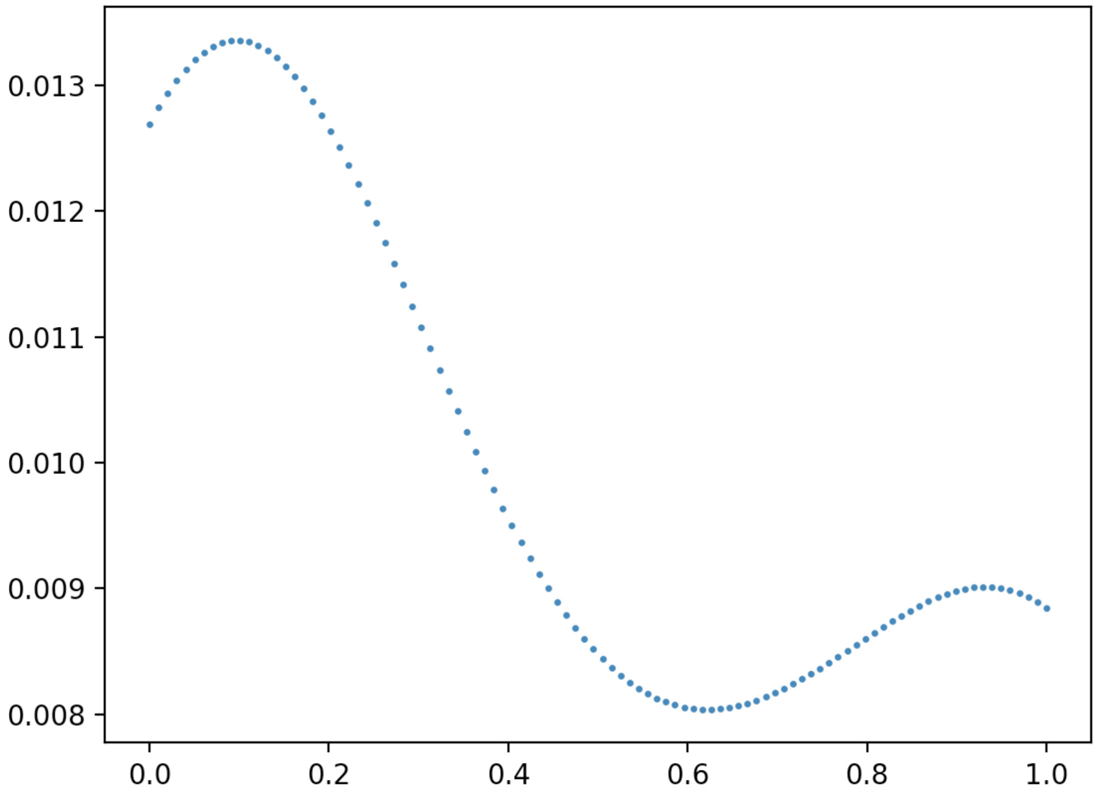
Why is that called the marginal?

Biggest Game Changer this half of CS109

Belief in **Vision** Given **User Responses**

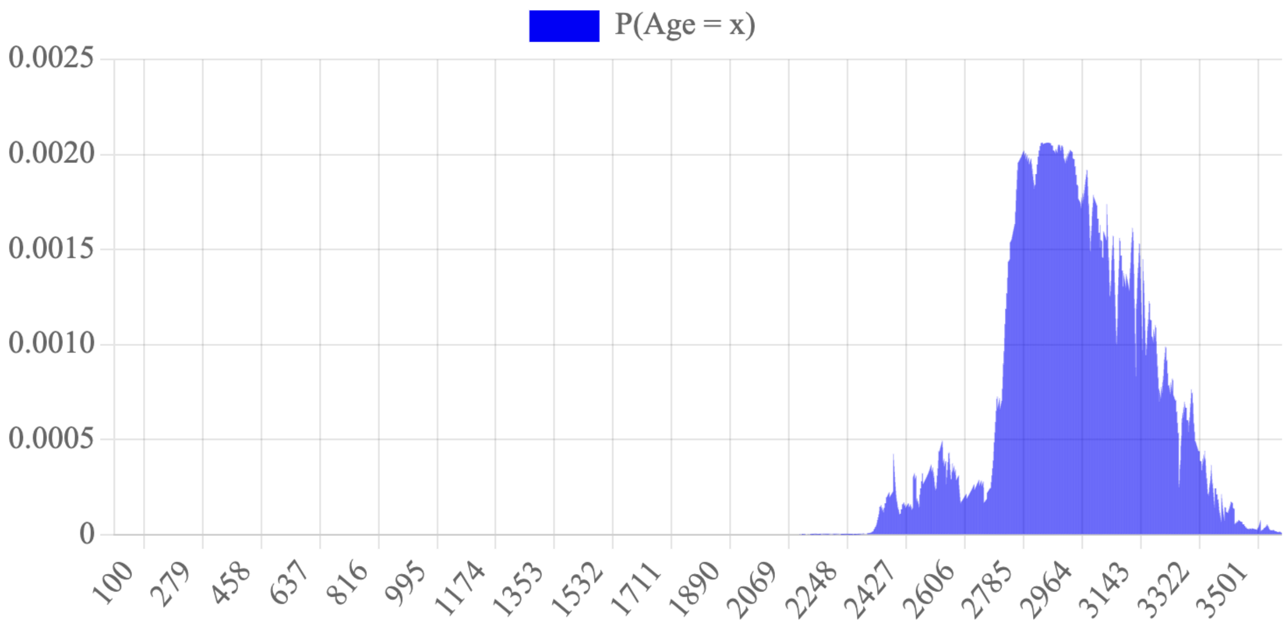


P(Ability to See | Observed Responses)



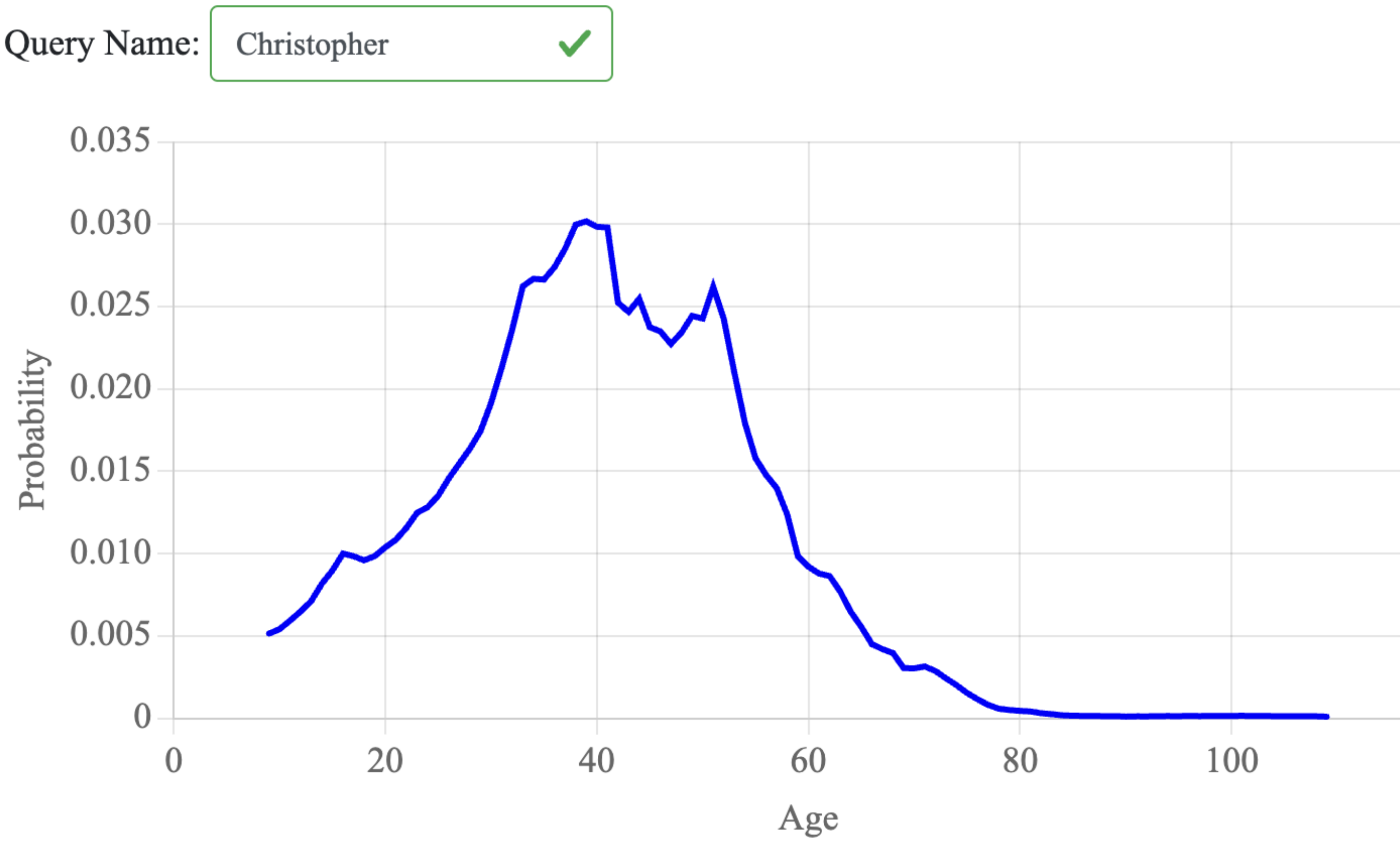
Belief in **Age** Given **Observed C14**

P(Age | Observed C14)

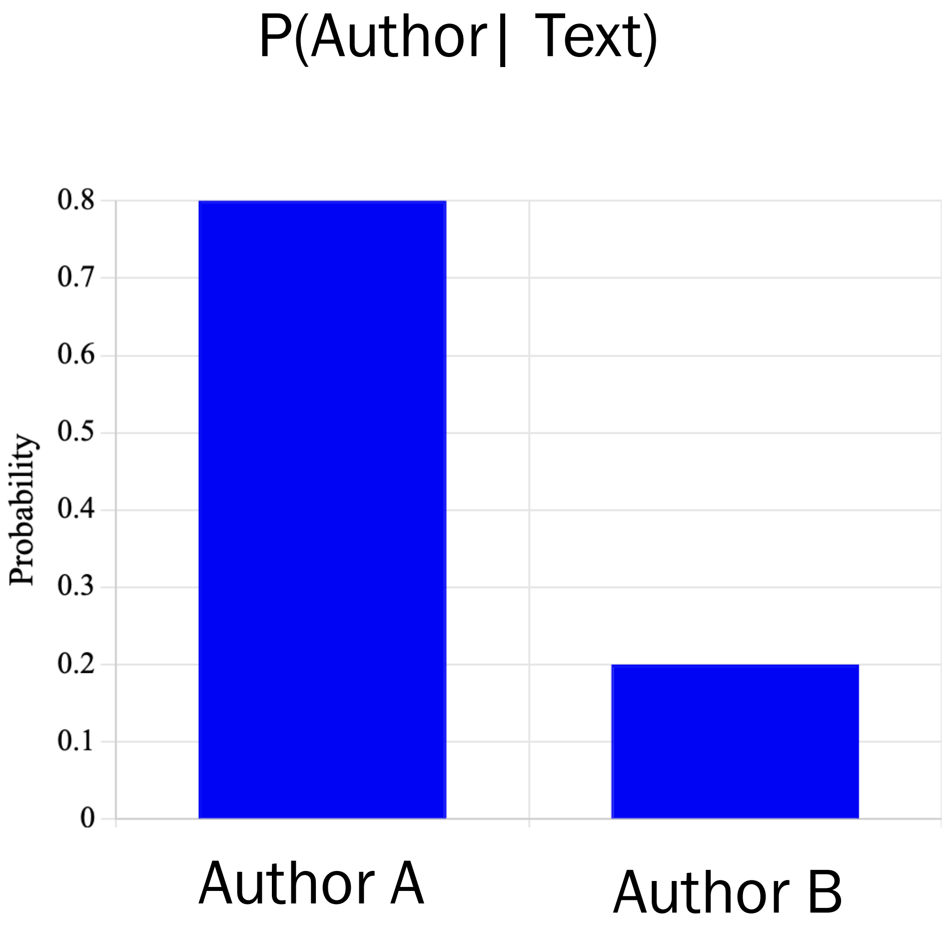


Belief in **Age** Given **Name**

P(Age | Name)



Belief in **Author** Given **Text**



Today: Inference

Inference *noun*

Updating one's belief about a random variable (or multiple) based on conditional knowledge regarding another random variable (or multiple) in a probabilistic model.

TLDR: conditional probability with random variables.

Your observation is:

Update variable is:

	Discrete	Continuous
Binary	Classic Bayes	
Multi-Valued		

Today

Bayes Theorem

Posterior belief

Likelihood of evidence

Prior belief

$$P(B|E) = \frac{P(E|B) \quad P(B)}{P(E)}$$

Normalization constant

Bayes with Discrete Random Variables

Let M be a **discrete** random variable

Let N be a **discrete** random variable

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

More generally

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

Shorthand notation

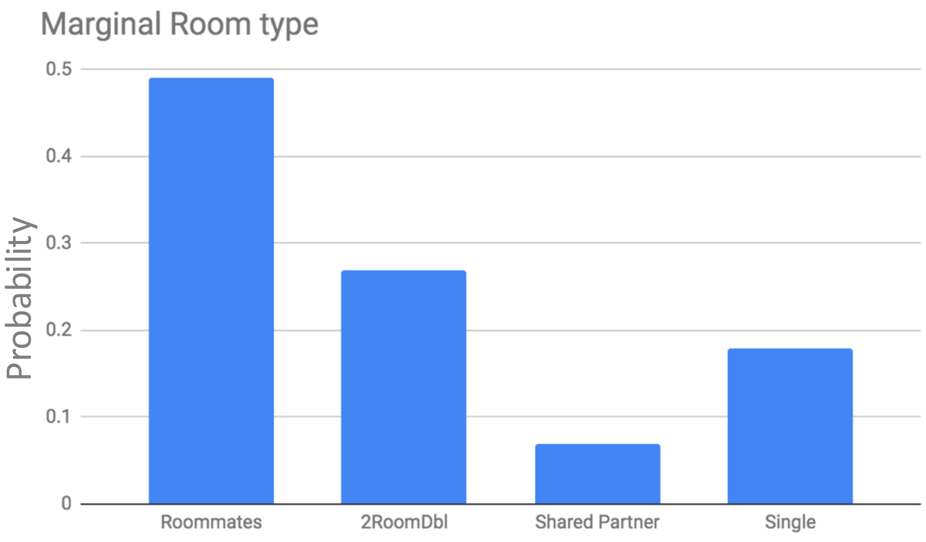
$$\boxed{P(m|n) = \frac{P(n|m) P(m)}{P(n)}}$$

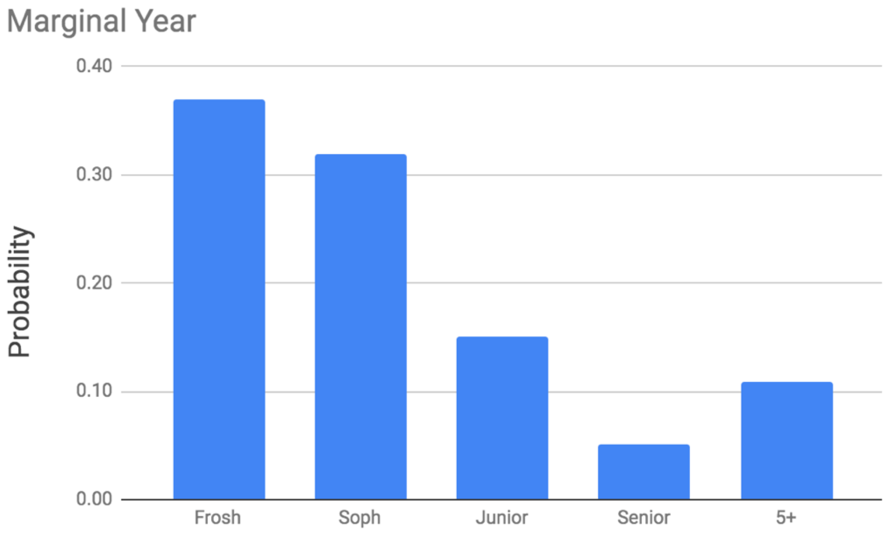
What is the probability distribution of rooms | student is a senior?

Joint

	Roommates	2RoomDbl	Shared Partner	Single	
Frosh	0.30	0.07	0.00	0.00	0.37
Soph	0.12	0.18	0.00	0.03	0.32
Junior	0.04	0.01	0.00	0.10	0.15
Senior	0.01	0.02	0.02	0.01	0.05
5+	0.02	0.00	0.05	0.04	0.11
	0.49	0.27	0.07	0.18	1.00

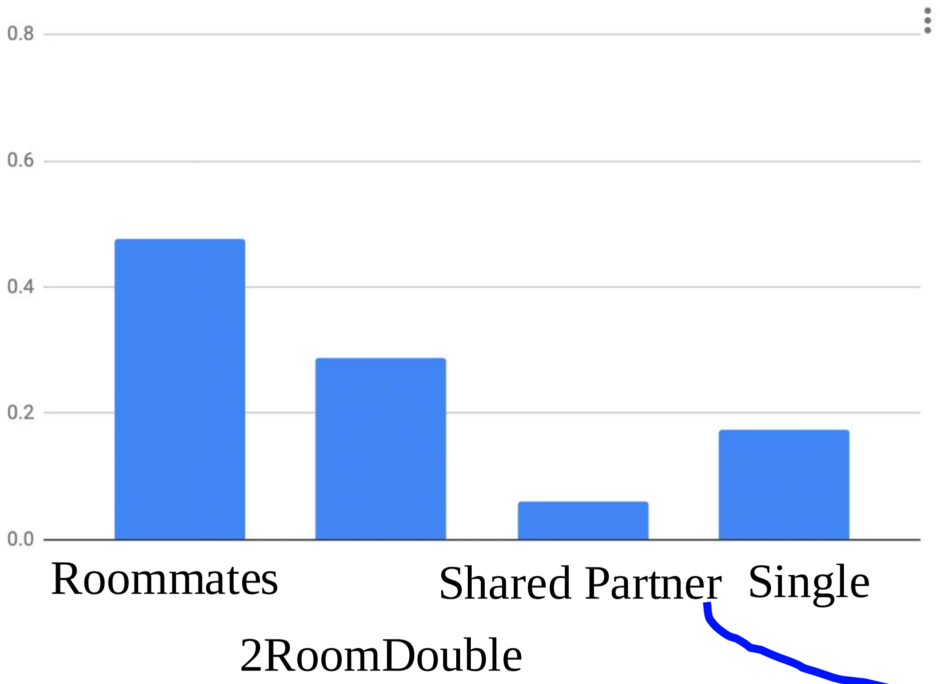
Marginals



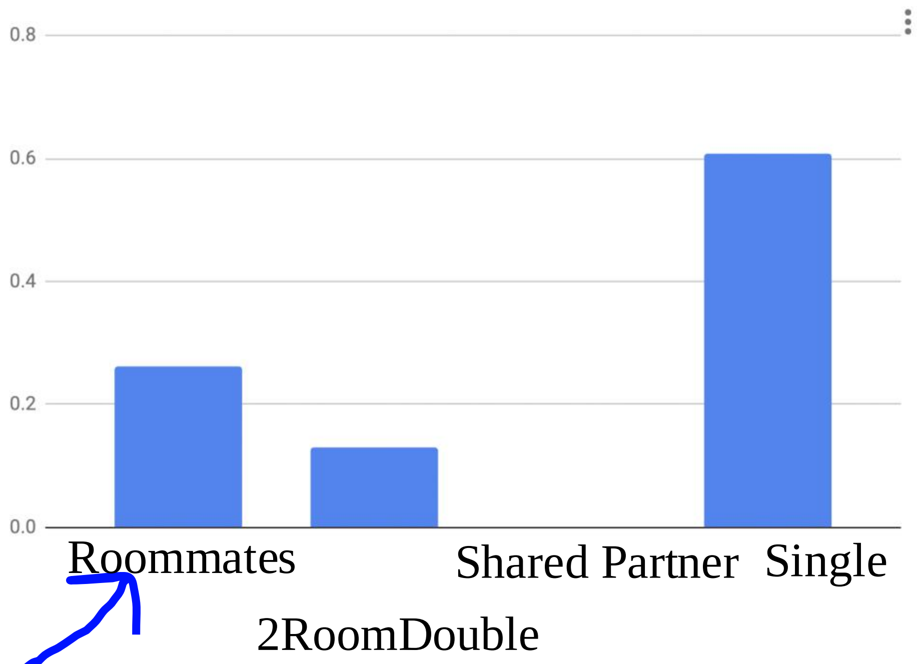


What is the probability distribution of rooms | student is a senior?

$$P(R = r)$$



$$P(R = r | Y = \text{senior})$$



Inference

I Heard That

Let X be the **change in gaze** (measured in degrees) over 3 seconds after a sound is played

Value of X	PMF of X given Baby can hear the sound	PMF of X given Baby can **not** hear the sound
0 to 5	0.08	0.40
5 to 10	0.15	0.30
10 to 15	0.35	0.12
15 to 20	0.20	0.08
20 to 25	0.12	0.05
Above 25	0.10	0.05

$$\text{P(can hear the sound)} = \frac{3}{4}$$

You observe $\boxed{X = 0}$. What is the probability the baby **can** hear the sound?

Question: Have I Been Given the Joint?

Let X be the **change in gaze** (measured in degrees) over 3 seconds after a sound is played

Value of X	PMF of X given Baby can hear the sound	PMF of X given Baby can **not** hear the sound
0 to 5	0.08	0.40
5 to 10	0.15	0.30
10 to 15	0.35	0.12
15 to 20	0.20	0.08
20 to 25	0.12	0.05
Above 25	0.10	0.05

$$\text{P(can hear the sound)} = \frac{3}{4}$$

You observe $X = 0$. What is the probability the baby **can** hear the sound?

I Heard That

Value of X	PMF of X given Baby can hear the sound	PMF of X given Baby can **not** hear the sound
0 to 5	0.08	0.40
5 to 10	0.15	0.30
10 to 15	0.35	0.12
15 to 20	0.20	0.08
20 to 25	0.12	0.05
Above 25	0.10	0.05

$$\text{P(can hear the sound)} = \frac{3}{4}$$

You observe $X = 0$. What is the probability the baby **can** hear the sound?

Y = 1 means the child can hear the sound

$$P(Y=1|X=0) = \frac{(X=0|Y=1)P(Y=1)}{P(X=0|Y=1)P(Y=1)+P(X=0|Y=0)P(Y=0)}$$

$$P(Y=1|X=0) = \frac{0.08 * 0.75}{0.08 * 0.75 + 0.40 * 0.25} = \frac{3}{8}$$

I Heard That with Continuous

Normal Assumption:

For babies who can hear sounds, we approximate their gaze movement after the sound is played as: $N(\mu = 15, \sigma^2 = 25)$.

For babies who can **not** hear sounds, we approximate gaze movement as $N(\mu = 8, \sigma^2 = 25)$.

For a new baby we observe a 14 degree movement after the sound is played. What is your belief that a baby can hear, under The Normal Assumption?

How do you handle observing a continuous value?

Aside

Mixing Discrete and Continuous

Let X be a **continuous** random variable

Let N be a **discrete** random variable

$$P(N = n|X = x) = \frac{P(X = x|N = n)P(N = n)}{P(X = x)}$$

$$P(N = n|X = x) = \frac{f(X = x|N = n) \cdot \epsilon \cdot P(N = n)}{f(X = x) \cdot \epsilon}$$

$$\boxed{P(N = n|X = x) = \frac{f(X = x|N = n) \cdot P(N = n)}{f(X = x)}}$$

Mixing Discrete and Continuous

Let X be a **continuous** random variable

Let N be a **discrete** random variable

$$P(X = x|N = n) = \frac{P(N = n|X = x)P(X = x)}{P(N = n)}$$

Change notation

$$P(x|n) = \frac{P(n|x)P(x)}{P(n)}$$

$$f(x|n) \cdot \epsilon_x = \frac{P(n|x)f(x) \cdot \epsilon_x}{P(n)}$$

$$\boxed{f(x|n) = \frac{P(n|x)f(x)}{P(n)}}$$

End Aside

I Heard That with Continuous

Normal Assumption:

For babies who can hear sounds, we approximate their gaze movement after the sound is played as: $N(\mu = 15, \sigma^2 = 25)$.

For babies who can **not** hear sounds, we approximate gaze movement as $N(\mu = 8, \sigma^2 = 25)$.

For a new baby we observe a 14 degree movement after the sound is played. What is your belief that a baby can hear, under The Normal Assumption?

Equivalently

Normal Assumption: For babies who can hear sounds, we approximate their gaze movement after the sound is played as: $N(\mu = 15, \sigma^2 = 25)$.

For babies who can **not** hear sounds, we approximate gaze movement as $N(\mu = 8, \sigma^2 = 25)$.

For a new baby we observe a 14 degree movement after the sound is played. What is your belief that a baby can hear, under The Normal Assumption?

```
def sample ():
  # bernoulli sample
  can_hear = rand_bern(0.75)
  if can_hear == 1:
     # gaussian sample
     return rand_gauss (mu = 15 , std = 5)
  else:
     # gaussian sample
     return rand_gauss (mu = 8, std = 5)
```

The function sample returned the value 14. What is the probability that can_hear was 1?

Inference with Continuous

Q: At birth, girl elephant weights are distributed as a Gaussian with mean = 160kg, std = 7kg. At birth, boy elephant weights are distributed as a Gaussian with mean = 165kg, std = 3kg. All you know about a newborn elephant is that it is $\boxed{163\text{kg}}$. What is the probability that it is a girl?

Inference with Continuous

Q: At birth, girl elephant weights are distributed as a Gaussian with mean = 160kg, std = 7kg. At birth, boy elephant weights are distributed as a Gaussian with mean = 165kg, std = 3kg. All you know about a newborn elephant is that it is 163kg. What is the probability that it is a girl?

Model:

Let G be an indicator that the elephant is a girl. G is Bern(p = 0.5)

Let X be the distribution of weight of the elephant.

$X \mid G = 1$ is $N(\mu = 160, \sigma^2 = 7^2)$

$X \mid G = 0$ is $N(\mu = 165, \sigma^2 = 3^2)$

Inference with Continuous

Q: What is $P(G = 1 \mid X = 163)$

Let G be an indicator that the elephant is a girl. G is $\text{Bern}(p = 0.5)$

Let X be the distribution of weight of the elephant.

$X \mid G = 1$ is $N(\mu = 160, \sigma^2 = 7^2)$

$X \mid G = 0$ is $N(\mu = 165, \sigma^2 = 3^2)$

All the Bayes Belong to Us

M,N are discrete. X, Y are continuous

OG Bayes

$$P(m|n) = \frac{P(n|m)P(m)}{P(n)}$$

Mix Bayes #1

$$f(x|n) = \frac{P(n|x)f(x)}{P(n)}$$

Mix Bayes #2

$$P(n|x) = \frac{f(x|n)P(n)}{f(x)}$$

$$f(x|y) = \frac{f(y|x)f(x)}{f(y)}$$

TLDR:

If random variable is discrete: use PMF

If random variable is continuous: use PDF

LOTP? Chain Rule? You can play too!

N is discrete. **X is continuous**

Chain Rule

$$f(N = n, X = x) = f(X = x | N = n)P(N = n)$$

Law of total probability

$$f(X = x) = \sum_{n} f(X = x | N = n)P(N = n)$$

Pedagogical Pause

Your observation is:

Update variable is:		Discrete	Continuous
	Binary	Classic Bayes	
	Multi-Valued		

Today

Goal: Inference

Change your belief
distribution
(Joint, PMF, or PDF)
of random variables,
based on
observations

*Note in the earlier examples, we were updating Bernoulli Random Variables

Lets Play Number of Function!

Number or Function?

$$P(X = 2|Y = 5)$$

Number

Number or Function?

$$P(X = x|Y = 2)$$

Function

(a probability mass function if discrete)

100 Binomial Problems
Winning Series
Jury Selection
Grading Eye Inflamation
Grades are Not Normal
Curse of Dimensionality
Probability of Baby Delivery

Part 3: Probabilistic Models
Joint Probability
Multinomial
Continuous Joint
Inference
Bayesian Networks
Independence in Variables
Correlation
General Inference
Applications
Fairness in AI
Federalist Paper Authorship
Name to Age
Bayesian Carbon Dating
Digital Vision Test
Bridge Distribution
CS109 Logo
Tracking in 2D

Part 4: Uncertainty Theory
Beta Distribution

Popularized in 2009:

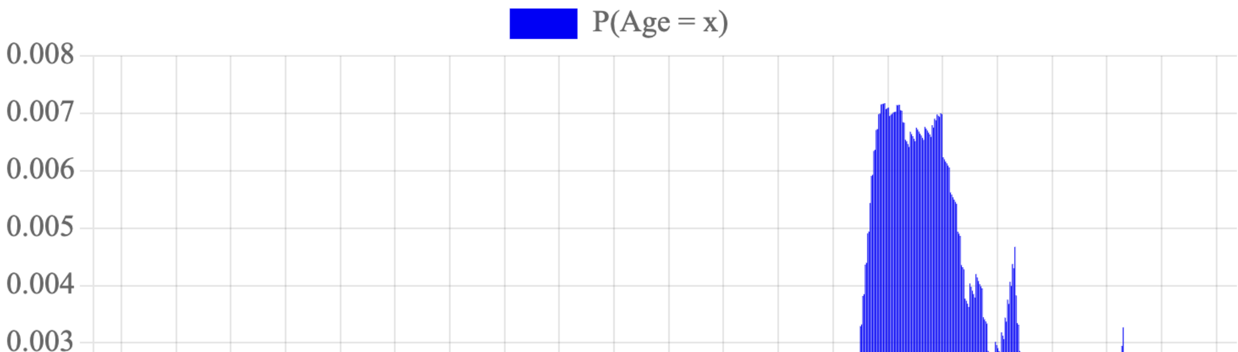
Bayesian Carbon Dating

We are able to know the age of ancient artefacts using a process called carbon dating. This process involves a lot of uncertainty! You observe a measurement of 90% of natural C14 molecules in a sample. What is your belief distribution over the age of the sample? This task requires probabilistic models because we have to think about two random variables together: the age A of the sample, and M the remaining C14 molecules.

Carbon Dating Demo

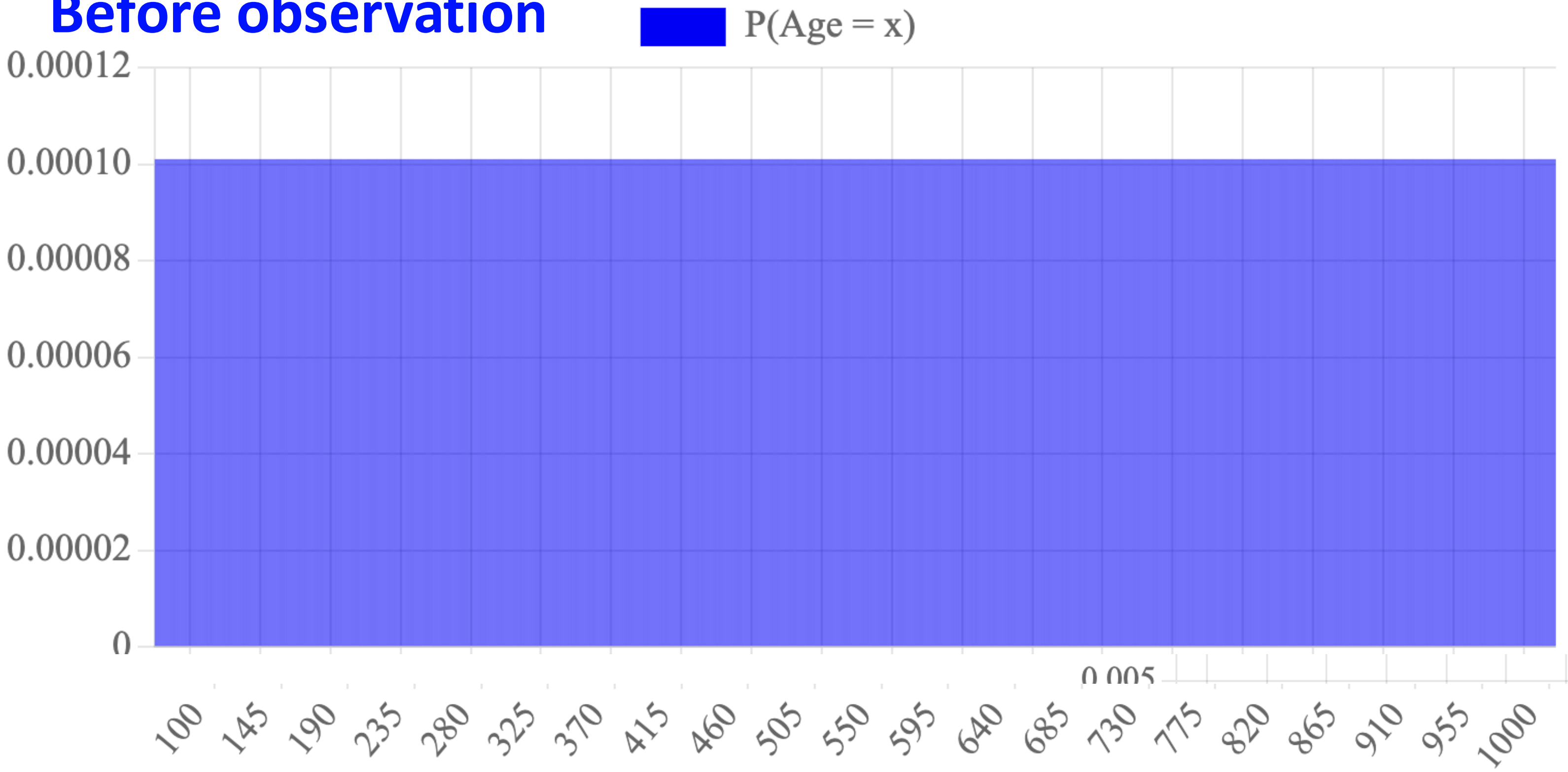
Imagine you have just taken a sample from your artifact. For the sample size you took, a living organism would have had 1000 molecules of C14. Use this demo to explore the relationship between how much C14 is left and your belief ditribution for how old your artifact is.

Remaining C14: 900



Update Belief PMF

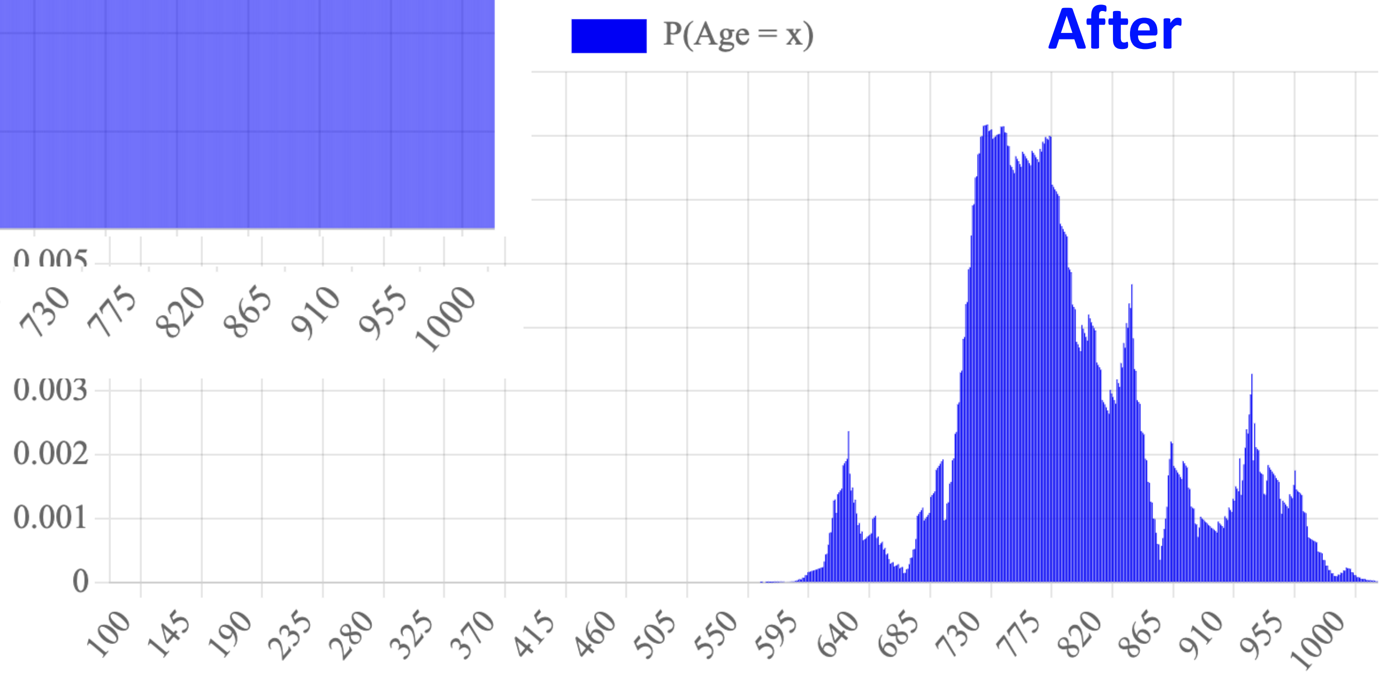
Before observation



Observation

Remaining C14: 900

After



Warmup with Code

```
def update_belief(m = 900):
  """
  Returns a dictionary A, where A[i] contains the
  corresponding probability, P(A = i| M = 900).
  m is the number of C14 molecules remaining and i
  is age in years. i is in the range 100 to 10000
  """
  pr_A = {}
  n_years = 9901
  for i in range(100,10000+1):
    prior = 1 / n_years # P(A = i)
    likelihood = calc_likelihood(m, i) # P(M=m | A=i)
    pr_A[i] = prior * likelihood
  # implicitly computes the normalization constant
  normalize(pr_A)
  return pr_A
```

Warmup with Code: Normalize

```
for i in range(100,10000+1):
  prior = 1 / n_years # P(A = i)
  likelihood = calc_likelihood(m, i) # P(M=m | A=i)
  pr_A[i] = prior * likelihood
# implicitly computes the normalization constant
normalize(pr_A)
return pr_A
```

```
def normalize(prob_dict):
    # first compute the sum of the probability
    sum = 0
    for key, pr in prob_dict.items():
        sum += pr
    # then divide each probability by that sum
    for key, pr in prob_dict.items():
        prob_dict[key] = pr / sum
    # now the probabilities sum to 1 (aka are normalized)
```

Warmup with Code: Normalize

Normalize: scale each value so that they would sum to 1

```
normalize({
  'cat'      : 5,
  'dog'      : 10,
  'axolotyl' : 5
})
```

```
{
  'cat'      : 0.2,
  'dog'      : 0.4,
  'axolotyl' : 0.2
}
```

Bayesian Carbon Dating: Inference Overview

Let A be how many years old the sample is (A = 100 means the sample is 100 years old)
Let M be the observed amount of C14 left in the sample

$$P(A = i | M = 900) = \frac{P(M = 900 | A = i) P(A = i)}{P(M = 900)}$$

$$= P(M = 900 | A = i) \cdot P(A = i) \cdot K$$

Such that

$$K = \frac{1}{\sum_i P(M = 900 | A = i) P(A = i)}$$

Understanding why Denom. is a Constant

A = age
M = measured C14

$$P(A=i|M=900) = P(M=900|A=i) \cdot P(A=i) \cdot \frac{1}{P(M=900)}$$

Notice which term doesn't change as i changes (four example calculations).

$$P(A=100|M=900) = \boxed{P(M=900|A=100) \cdot P(A=100)} \cdot \frac{1}{P(M=900)}$$

$$P(A=200|M=900) = \boxed{P(M=900|A=200) \cdot P(A=200)} \cdot \frac{1}{P(M=900)}$$

$$P(A=300|M=900) = \boxed{P(M=900|A=300) \cdot P(A=300)} \cdot \frac{1}{P(M=900)}$$

$$P(A=400|M=900) = \boxed{P(M=900|A=400) \cdot P(A=400)} \cdot \frac{1}{P(M=900)}$$

Doesn't change

Changes with *i*

Bayesian Carbon Dating: Inference Overview

Let A be how many years old the sample is (A = 100 means the sample is 100 years old)
Let M be the observed amount of C14 left in the sample

$$P(A = i | M = 900) = \frac{P(M = 900 | A = i) P(A = i)}{P(M = 900)}$$

$$= P(M = 900 | A = i) \cdot P(A = i) \cdot K$$

Such that

$$K = \frac{1}{\sum_i P(M = 900 | A = i) P(A = i)}$$

Understanding Through Code

$$P(A = \boxed{i} | M = 900) = P(M = 900 | A = i) \cdot P(A = i) \cdot \boxed{K}$$

```
def update_belief(m = 900):
  """
  Returns a dictionary A, where A[i] contains the
  corresponding probability, P(A = i| M = 900).
  m is the number of C14 molecules remaining and i
  is age in years. i is in the range 100 to 10000
  """
  pr_A = {}
  n_years = 9901
  for i in range(100,10000+1):
    prior = 1 / n_years # P(A = i)
    likelihood = calc_likelihood(m, i) # P(M=m | A=i)
    pr_A[i] = prior * likelihood
  # implicitly computes the normalization constant
  normalize(pr_A)
  return pr_A
```

Carbon Dating Likelihood Math

Probability of Having 900 Remain

$$P(M = 900|A = i)$$

There were originally 1000 C14 molecules.
Each molecule remains independently with equal probability p_i
What is the probability that 900 remain?

$$M \sim \text{Bin}(n = 1000, p = p_i)$$

$$P(M = 900|A = i) = \binom{1000}{900} (p_i)^{900} \cdot (1 - p_i)^{100}$$

Each molecules' time to live is exponential with λ = 1/8267
Let T be the time to decay for any one molecule

$$T \sim \text{Exp}(\lambda = 1/8267) \quad p_i = P(T > i) = 1 - P(T < i) = e^{-\frac{i}{8267}}$$

Probability of Having 900 Remain

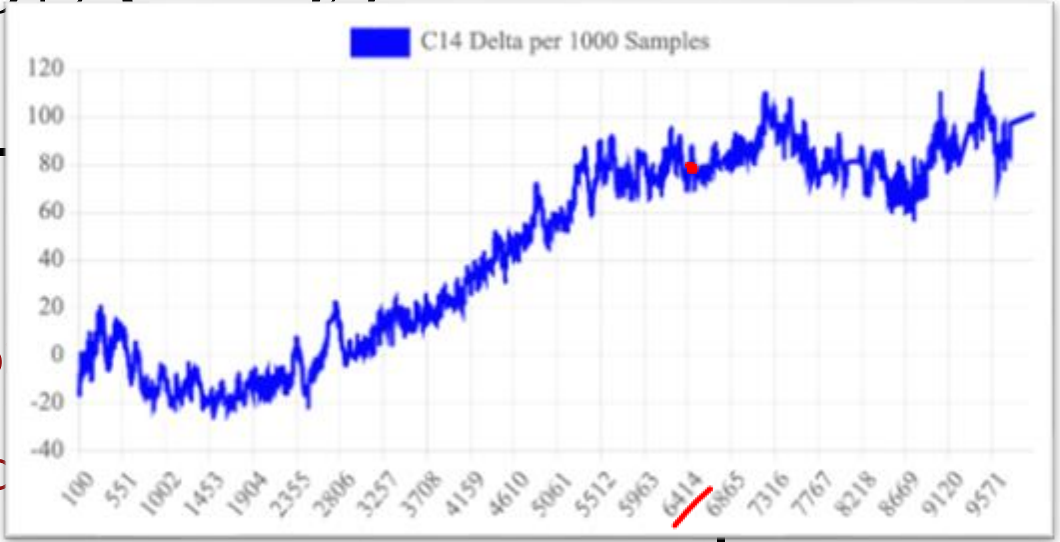
$$P(M = 900 | A = i)$$

```python
def calc_likelihood(m = 900, age):
  """
  Computes P(M = m | A = age), the probability of
  having m molecules left given the sample is age
  years old. Uses the exponential decay of C14
  """
  n_original = 1000
  p_remain = math.exp(-age/C14_MEAN_LIFE)
  return stats.binom.pmf(m, n_original, p_remain)
```

Probability of Having 900 Remain

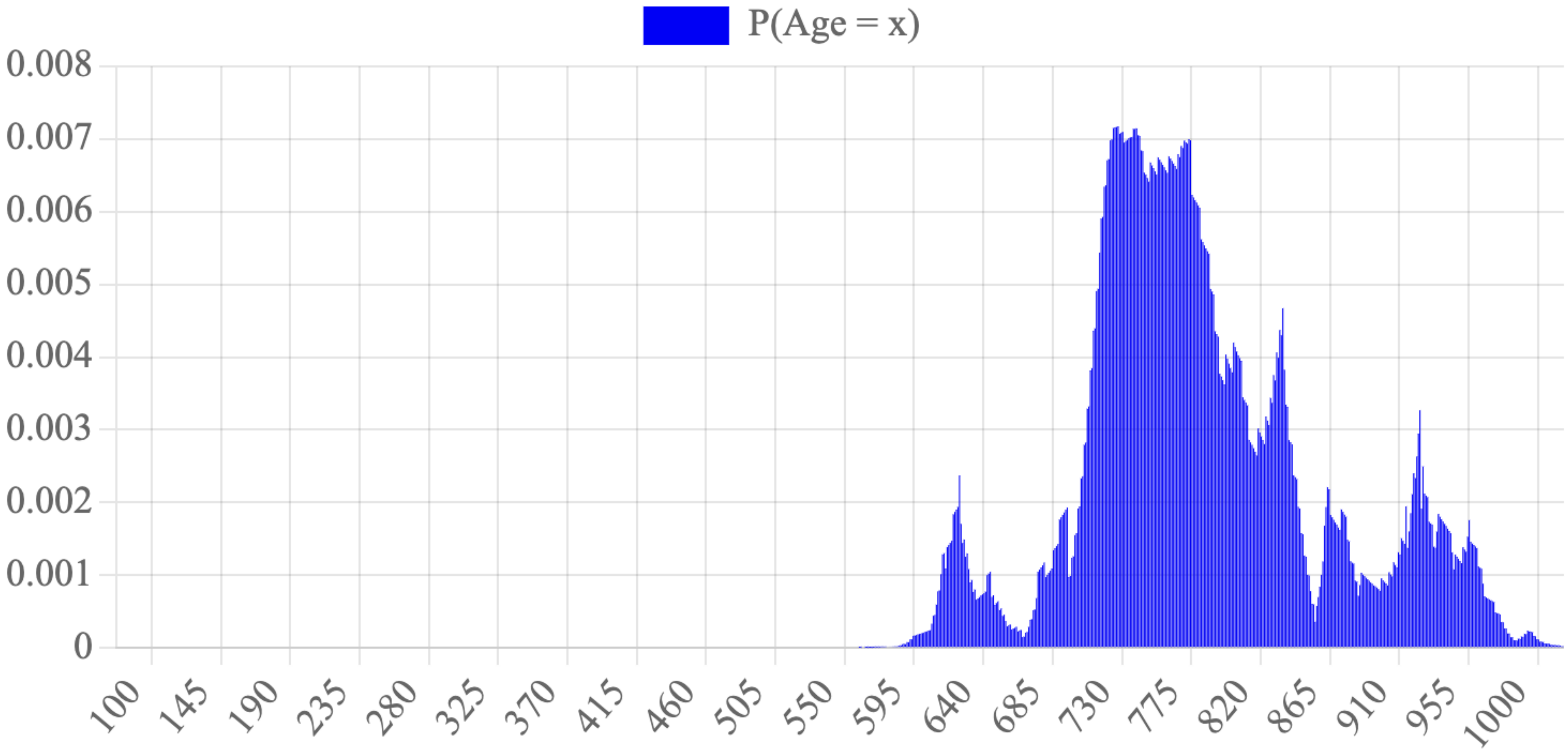
$$P(M = 900 | A = i)$$

```
def calc_likelihood(m = 900, age):
  """
  Computes P(M = m | A = age), the probab
  having m molecules left given the sample
  years old. Uses the exponential decay of C
  """
  n_original = 1000 + delta_start(age)
  p_remain = math.exp(-age/C14_MEAN_LIFE)
  return stats.binom.pmf(m, n_original, p_remain)
```



Probability of Having 900 Remain

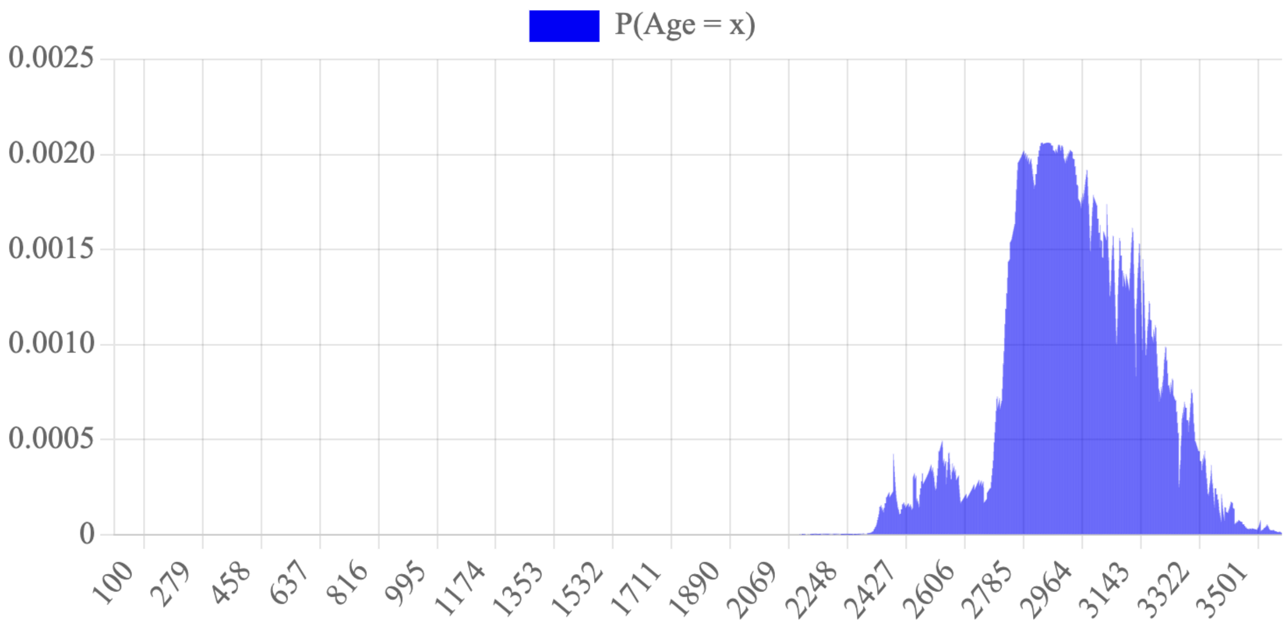
Remaining C14: 900



Come Back for More!

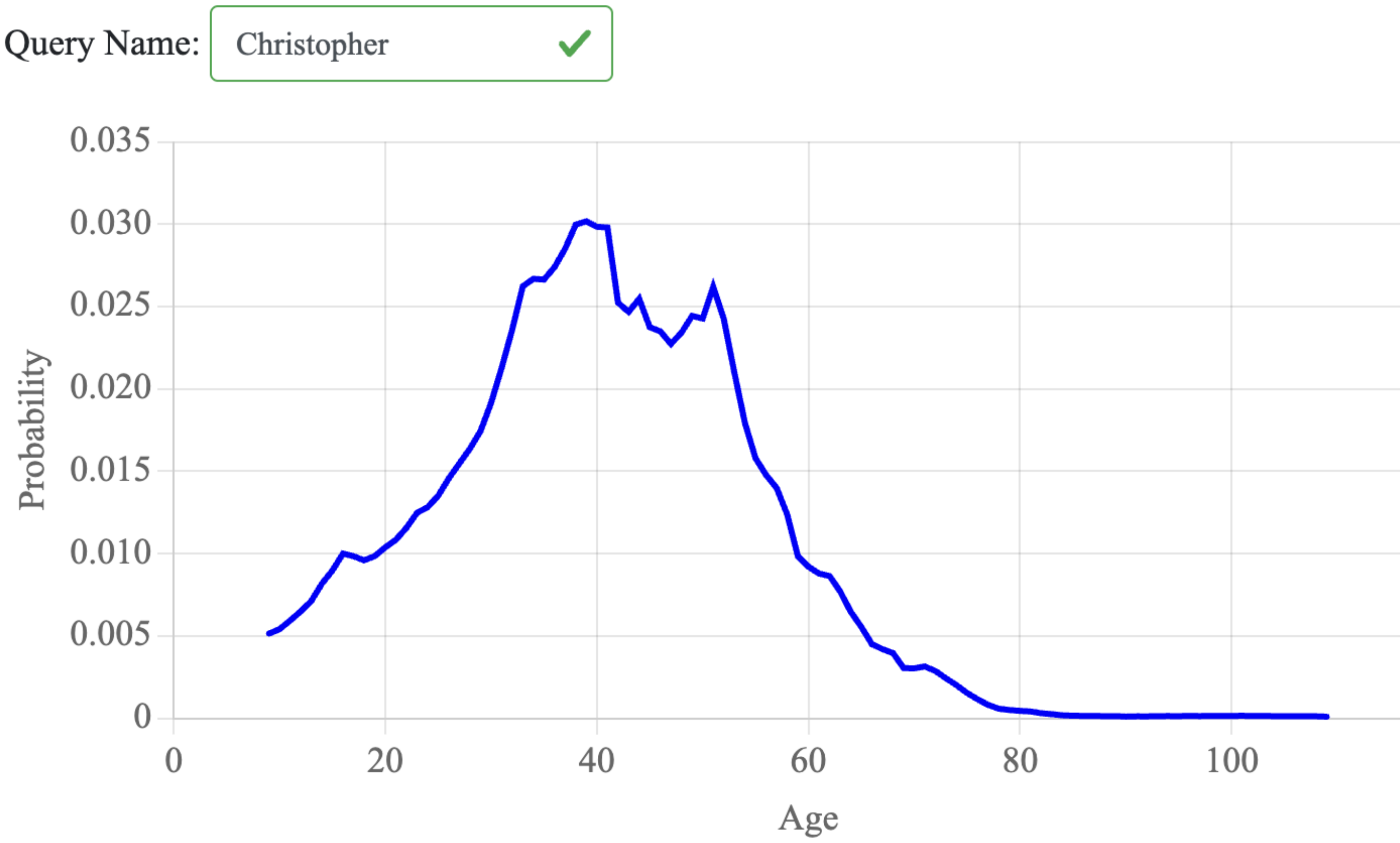
Belief in Age Given Observed C14

P(Age | Observed C14)

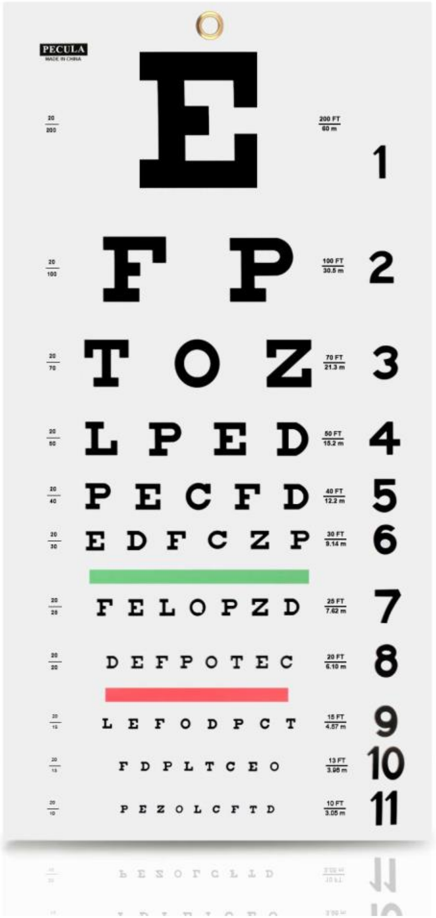


Belief in **Age** Given **Name**

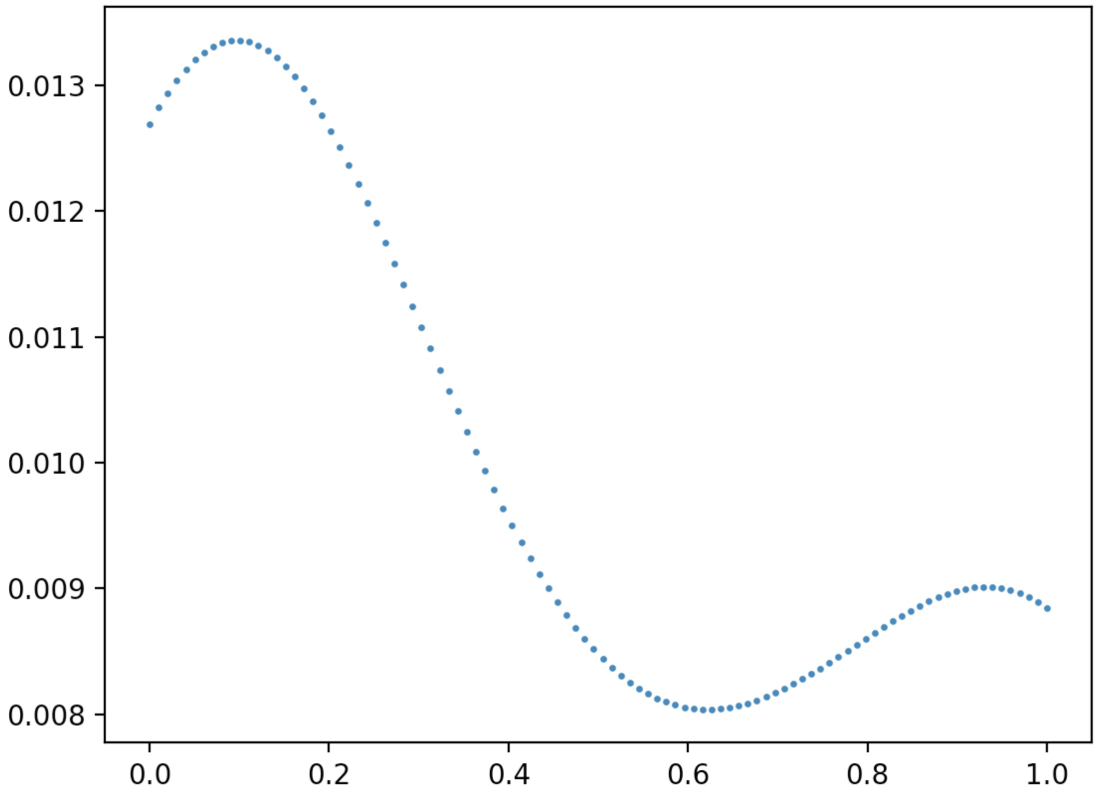
P(Age | Name)



Belief in **Vision** Given **User Responses**



P(Ability to See | Observed Responses)



Belief in **Author** Given **Text**