The Generalization Jigsaw

BAYESIANISM

- (methods)
- Parameter confidence
- Model choice: model weighting

FREQUENTISM

- (methods)
- Parameter confidence
- Model choice: hypothesis testing

EMPIRICISM

- Evaluating model fit
- Model choice: holdout evaluation

PUTTING THE JIGSAW TOGETHER

Bayes's rule, done right

For two **discrete** random variables X and Y,

$$\mathbb{P}(X = x|Y = y) = \frac{\mathbb{P}(X = x)\mathbb{P}(Y = y|X = x)}{\mathbb{P}(Y = y)} \quad \text{when } \mathbb{P}(Y = y) > 0$$

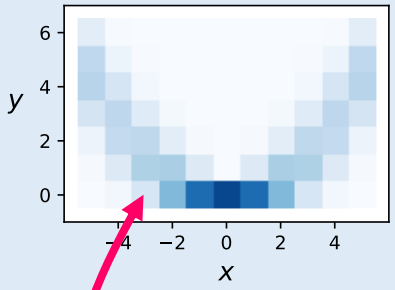
For two **discrete or continuous** random variables X and Y,

$$\Pr_X(x|Y = y) = \frac{\Pr_X(x)\,\Pr_Y(y|X = x)}{\Pr_Y(y)} \quad \text{when } \Pr_Y(Y) > 0$$

What do these "conditional likelihoods" even mean?

Joint distribution


```
def rxy():
    x = np.random.randint(low=-5, high=6)  # from -5 to +5 inclusive
    y = np.random.binomial(n=6, p=(x/6)**2)
    return (x,y)
```



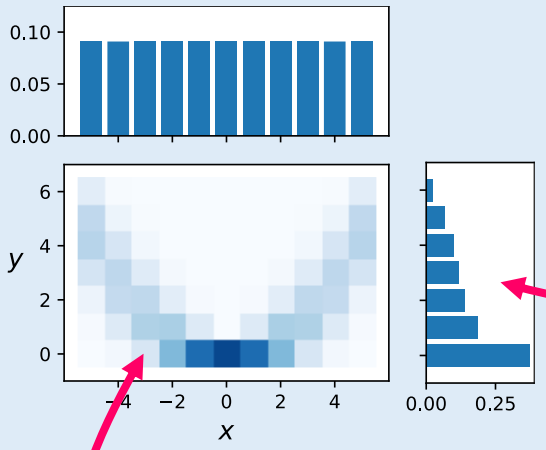
The joint pmf of (X, Y)

$\Pr_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y)$

by defn. of conditional probability

$= \mathbb{P}(X=x)\,\mathbb{P}(Y=y \mid X=x) = \frac{1}{11}\binom{6}{y}\left[\left(\frac{x}{6}\right)^2\right]^y\left[1-\left(\frac{x}{6}\right)^2\right]^{6-y}$

Marginal random variables




```
def rxy():
    x = np.random.randint(low=-5, high=6)  # from -5 to +5 inclusive
    y = np.random.binomial(n=6, p=(x/6)**2)
    return (x,y)
```

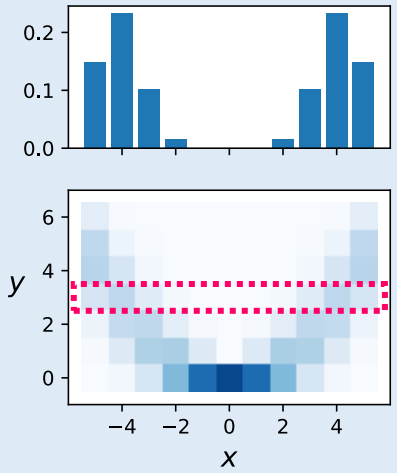
The joint pmf of (X, Y)

$$\Pr_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y)$$

The marginal of Y

$$\begin{aligned}\Pr_Y(y) &= \mathbb{P}(Y = y) \\ &= \sum_x \mathbb{P}(X = x, Y = y) \quad \text{by the Sum Rule} \\ &= \sum_x \Pr_{X,Y}(x, y)\end{aligned}$$

Conditional random variables




```
def rxy():
    x = np.random.randint(low=-5, high=6)  # from -5 to +5 inclusive
    y = np.random.binomial(n=6, p=(x/6)**2)
    return (x,y)
```

QUESTION. **What is X conditional on $Y = 3$?**

$$\mathbb{P}(X = x|Y = 3) = \frac{\mathbb{P}(X = x, Y = 3)}{\mathbb{P}(Y = 3)} = \frac{\Pr_{X,Y}(x, 3)}{\Pr_Y(3)}$$

$\text{pmf}_3(x)$

i.e. take the Y=3 row, then rescale it to sum to 1

We can think of "X conditional on $Y = 3$" as a random variable …

We've provided a valid probability mass function:

$\text{pmf}_3(\cdot) \geq 0$ $\quad \sum_x \text{pmf}_3(x) = 1$

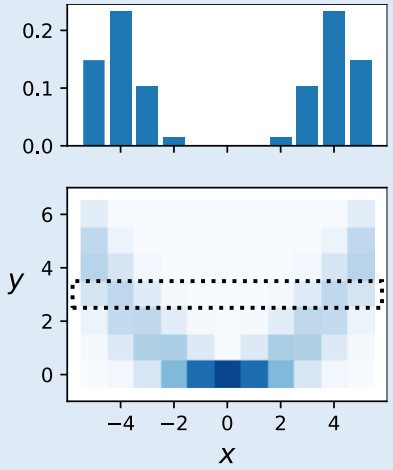
Sample space: $\Omega = \{-5, -4, \ldots, 4, 5\}$

Code to generate values from it:

```
def rx_given_y():
    while True:
        x,y = rxy()
        if y == 3: break
    return x
```

```
def rx_given_y():
    Ω = {-5,...,5}
    p = [pmf_3(x) for x in Ω]
    return np.random.choice(Ω, p=p)
```

Conditional random variables




```
def rxy():
    x = np.random.randint(low=-5, high=6)  # from -5 to +5 inclusive
    y = np.random.binomial(n=6, p=(x/6)**2)
    return (x,y)
```

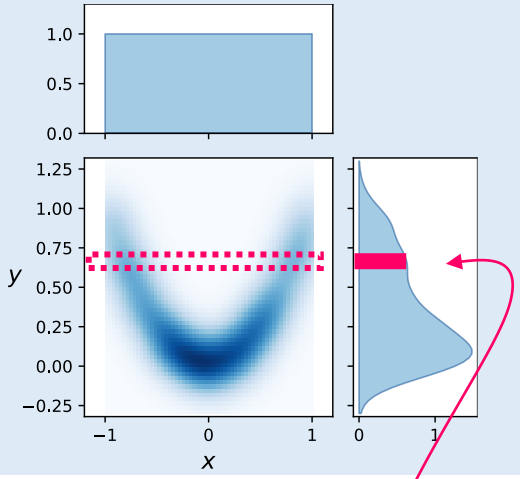
We *define* the **conditional random variable**, written $(X|Y=y)$, by specifying its likelihood:

$$\Pr_{(X|Y=y)}(x) = \frac{\Pr_{X,Y}(x,y)}{\Pr_Y(y)}$$

This likelihood is also written $\Pr_X(x|Y=y)$.

i.e. take the Y=y row from the joint pmf, then rescale it

Conditional random variables (continuous case)




```
def rxy():
    x = np.random.uniform(-1,1)
    y = np.random.normal(loc=x**2, scale=0.1)
    return (x,y)
```

QUESTION. **What is X conditional on $Y = 0.6$?**

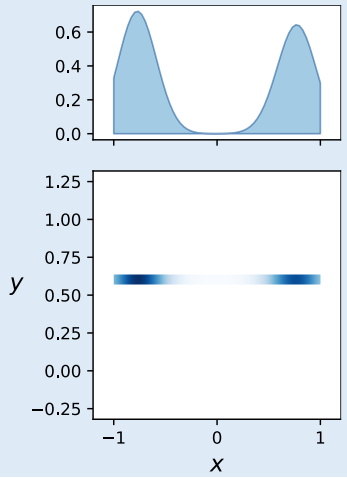
Take the Y=0.6 slice of the joint pdf,
then rescale it so it integrates to 1
i.e. so we get a legitimate pdf.

(Rescale it by dividing by $\Pr_Y(0.6)$
where $\Pr_Y(\cdot)$ is the marginal for Y.)

The marginal for Y

$$\Pr_Y(y) = \int_x \Pr_{X,Y}(x,y)\ dx$$

Conditional random variables (continuous case)




```
def rxy():
    x = np.random.uniform(-1,1)
    y = np.random.normal(loc=x**2, scale=0.1)
    return (x,y)
```

QUESTION. **What is X conditional on $Y = 0.6$?**

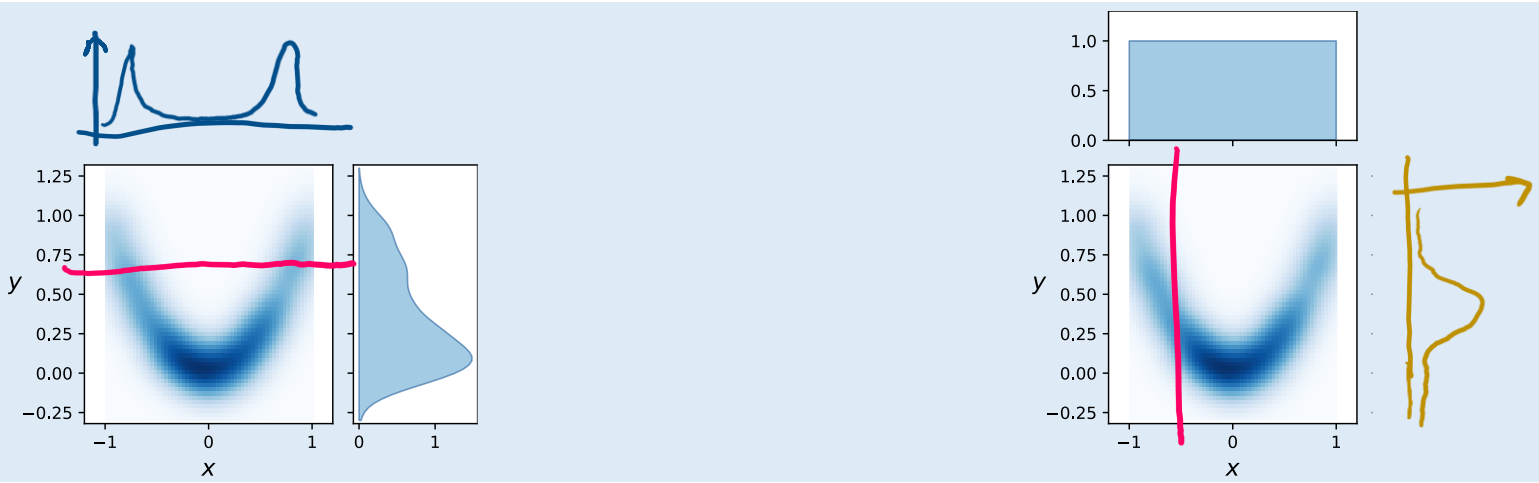
Take the Y=0.6 slice of the joint pdf,
then rescale it so it integrates to 1
i.e. so we get a legitimate pdf.

(Rescale it by dividing by $\Pr_Y(0.6)$
where $\Pr_Y(\cdot)$ is the marginal for Y.)

We *define* the conditional random variable $(\boldsymbol{X}|\boldsymbol{Y} = \boldsymbol{y})$ by specifying its likelihood:

$$\Pr_X(x|Y = y) = \frac{\Pr_{X,Y}(x, y)}{\Pr_Y(y)}$$

Bayes's rule



$$\Pr_X(x \mid Y=y) = \frac{\Pr_{X,Y}(x,y)}{\Pr_Y(y)}$$

$$\Pr_Y(y \mid X=x) = \frac{\Pr_{X,Y}(x,y)}{\Pr_X(x)}$$

$$\Pr_X(x \mid Y=y) = \frac{\Pr_{X,Y}(x,y)}{\Pr_Y(y)} = \frac{\Pr_X(x)\,\Pr_Y(y \mid X=x)}{\Pr_Y(y)}$$

Bayes's rule is true for any pair of random variables X, Y.
It's only useful for "sequential models" i.e. when the question tells us $\Pr_X(x)$ and $\Pr_Y(y|X=x)$.

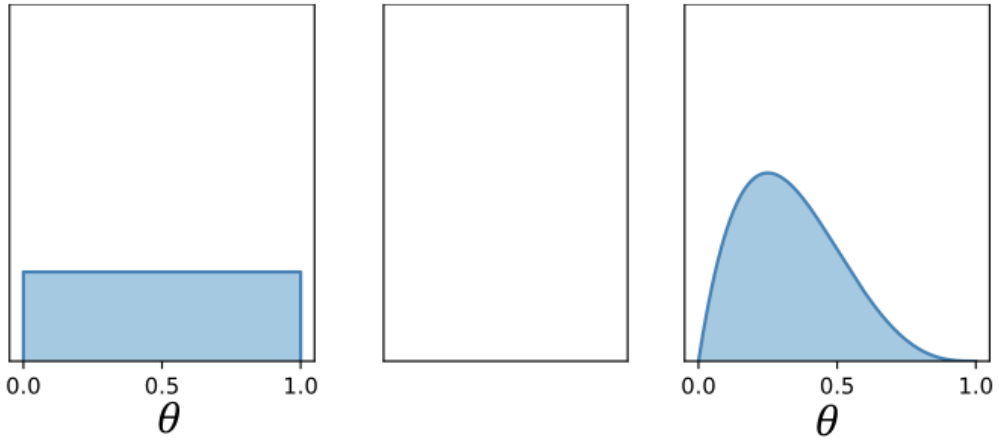
Now we have the tech to apply Bayes's rule to problems with continuous random variables.

$\Theta \sim U[0,1]$

to express complete prior ignorance about θ

$X \sim \mathrm{Bin}(n, \Theta)$

prior belief $\Pr_\Theta(\theta)$ + data x → posterior belief $\Pr_\Theta(\theta | X = x)$



Bayes's rule for discrete or continuous random variables

For two random variables X and Y,

$$\Pr_X(x|Y=y) = \frac{\Pr_X(x)\,\Pr_Y(y|X=x)}{\Pr_Y(y)} \quad \text{when} \quad \Pr_Y(y) > 0$$

In practice, we write it as

constant that doesn't involve x

$$\Pr_X(x|Y=y) = \kappa\,\Pr_X(x)\,\Pr_Y(y|X=x)$$

$\Pr_{(X|Y=y)}(x)$

then figure out κ so that $\Pr_X(\cdot\,|Y=y)$ is a legitimate likelihood function

Choose κ so that

$$\int_x \kappa\,\Pr_X(x)\,\Pr_Y(y|X=x)\,dx = 1$$

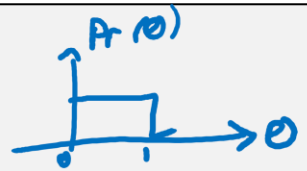
Exercise.

Consider the pair of random variables (Θ, X) where

$$\Theta \sim U[0,1], \qquad X \sim \mathrm{Bin}(4, \Theta)$$

Find the distribution of $(\Theta | X = 1)$.

$\Pr_\Theta(\theta) =$ 1



$\Pr_X(x|\Theta = \theta) =$ $\binom{4}{x}\theta^x(1-\theta)^{4-x} = 4\theta(1-\theta)^3$ for $n=4, x=1$

$\Pr_\Theta(\theta|X=1) = \kappa \Pr_\Theta(\theta) \Pr_X(1|\Theta=\theta)$ $= k \times 1 \times 4\theta(1-\theta)^3$

$= k' \theta(1-\theta)^3$ k' amalgamates non-θ terms.

Normalize: $\int_{\theta=0}^{1} k'\theta(1-\theta)^3 d\theta = 1 \Rightarrow k' = \dfrac{1}{\int_{\theta=0}^{1} \theta(1-\theta)^3 d\theta}$.

Exercise.

Consider the pair of random variables (Θ, X) where

$$\Theta \sim U[0,1], \qquad X \sim \text{Bin}(4, \Theta)$$

Find the distribution of $(\Theta | X = 1)$.

Beta

Probability density function	
Notation	$\text{Beta}(\alpha, \beta)$
PDF	$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$ where $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ and Γ is the Gamma function.

$$\Pr_\Theta(\theta | X = 1) = \kappa \Pr_\Theta(\theta) \Pr_X(1 | \Theta = \theta)$$

$$= k\, \theta\, (1-\theta)^3 = k\, B(\alpha, \beta) \cdot \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}$$

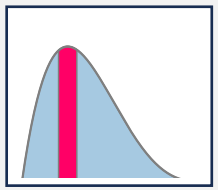
this is a standard pdf

where $\alpha = 2, \quad \beta = 4$

so this constant must be 1 (otherwise this pdf wouldn't integrate to 1 wr.t. θ)

Thus $(\Theta | X = 1) \sim \text{Beta}(\alpha = 2, \beta = 4)$

It's easier to communicate the posterior distribution by reporting a confidence interval.
E.g. what is $\mathbb{P}(\Theta \in [.2, .3] \mid X = 1)$?



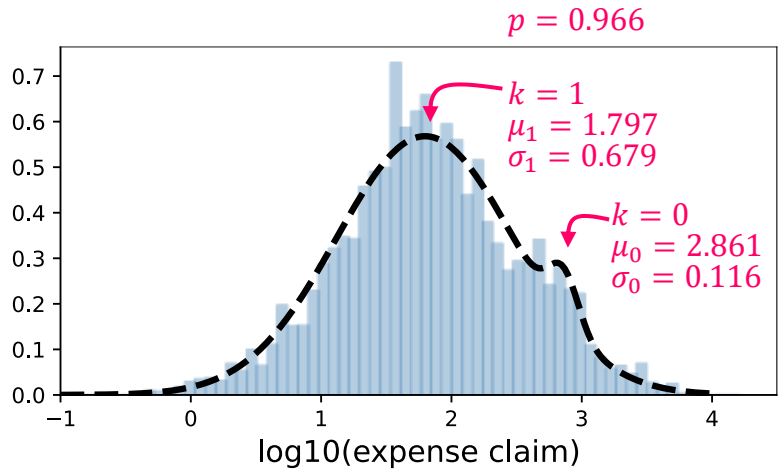
```
D = scipy.stats.beta(a=2,b=4)
D.cdf(.3) - D.cdf(.2)
```

Exercise 5.2.3 (Bayesian classification)

In a dataset of MP expense claims, let y_i be $\log_{10}$ of the claim amount in record i. A histogram of the y_i suggests we use a Gaussian mixture model with two components,

$$K = \begin{cases} 1 & \text{with prob } p \\ 0 & \text{with prob } 1-p \end{cases}$$
$$Y \sim \text{Normal}(\mu_K, \sigma_K^2)$$

Find the probability that a claim amount £5000 belongs to the component $k = 0$.



In other words, find $\mathbb{P}(K = 0 | Y = y)$, where $y = \log_{10} 5000 = 3.70$.

$\Pr_K(k \mid Y=y)$

$\Pr_K(k) =$

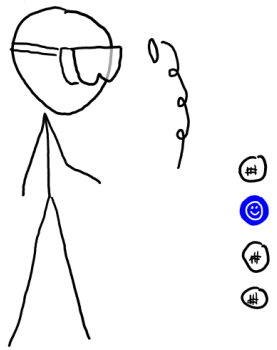
$\Pr_Y(y|K = k) =$

$\Pr_K(k|Y = y) = \kappa \Pr_K(k) \Pr_Y(y|K = k)$

choose κ s.t. $\sum_k \Pr_K(k) \Pr_Y(y|k=k) = 1$

EXERCISE

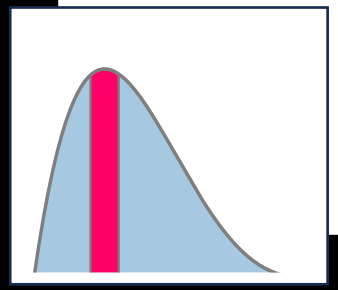
We've used Bayes's rule to find the posterior distribution of random variable.
We've had tricky integrals / sums, to find the normalizing constant.



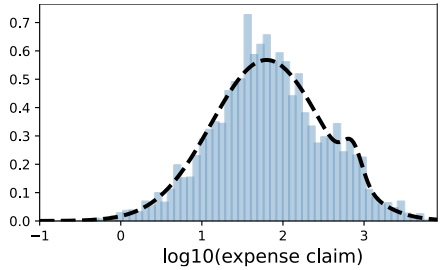
$$\Theta \sim U[0,1]$$

$$\downarrow$$

$$X \sim \mathrm{Bin}(n, \Theta)$$



Want $\Pr_\Theta(\theta | X = x)$
Or $\mathbb{P}(\Theta \in [.2, .3] | X = x)$



$$K \sim \mathrm{Bin}(1, p)$$

$$\downarrow$$

$$Y \sim N(\mu_K, \sigma_K^2)$$

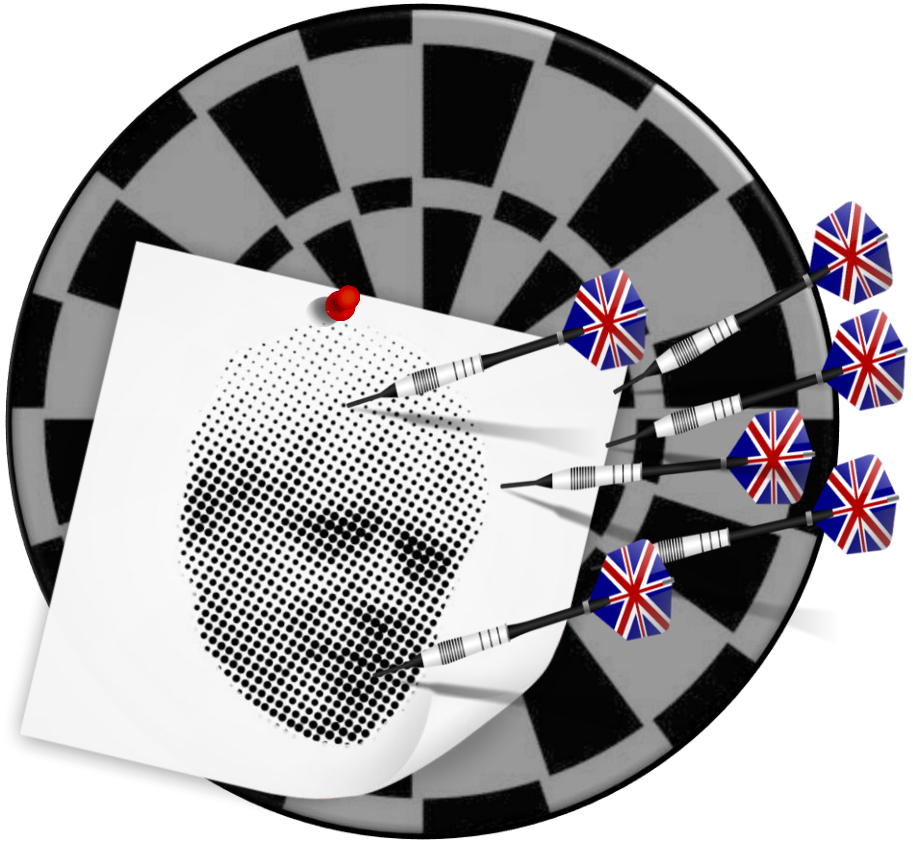
Want $\Pr_K(k | Y = y)$

If we want to give our answer in a more practical form, as a confidence interval, there's another tricky integral.

Next: how we can use sampling and computation instead of tricky integrals

§6. Computational methods

What's the chance that a randomly thrown dart will hit the mystery object A?



Let X be the location of a randomly thrown dart, and let $x_1, \ldots, x_n$ be some throws.

The probability of hitting A is

$$\mathbb{P}(X \in A) \approx \frac{1}{n}\sum_{i=1}^{n} 1_{x_i \in A}$$

$1_{x_i \in A} = \begin{cases} 1 & \text{if } x_i \in A \\ 0 & \text{else} \end{cases}$

```
1  # Let X ~ N(μ = 1, σ = 3). What is ℙ(X > 5)?
2  x = np.random.normal(loc=1, scale=3, size=10000)
3  i = (x > 5)
4  np.mean(i)
```

i is a vec of 10000 bools

what fraction of them are True?

Expectation

For a real-valued random variable X

$$\mathbb{E}X = \begin{cases} \sum_x x \Pr_X(x), & \text{if } X \text{ is discrete} \\ \int_x x \Pr_X(x)\, dx, & \text{if } X \text{ is continuous} \end{cases}$$

Law of the Unconscious Statistician

For a random variable X and a real-valued function h

$$\mathbb{E}h(X) = \begin{cases} \sum_x h(x)\ \Pr_X(x)\,, & \text{if } X \text{ is discrete} \\ \int_x h(x)\ \Pr_X(x)\,dx\,, & \text{if } X \text{ is continuous} \end{cases}$$

If we want to know the average properties of a rich random variable (e.g. the average length of a random tweet), we have to use real-valued property readout functions h(X) so that we can take averages.

Monte Carlo integration

$$\mathbb{E}h(X) \approx \frac{1}{n}\sum_{i=1}^{n} h(x_i)$$

where $x_1, \ldots, x_n$ is a sample drawn from X

let $h(x) = 1_{x \in A}$

By Monte Carlo.

$$\mathbb{E}\, h(X) \approx \frac{1}{n} \sum_{i=1}^{n} h(x_i)$$

$$\frac{1}{n} \sum_{i} 1_{x_i \in A}$$

let $Y = h(X)$

$$\mathbb{E}Y = 0 \times \mathbb{P}(Y=0) + 1 \times \mathbb{P}(Y=1)$$

$$= \mathbb{P}(Y=1)$$

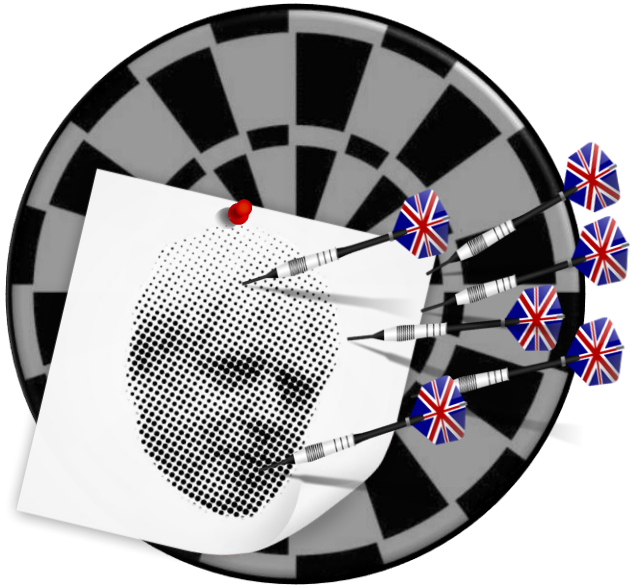
$$= \mathbb{P}(1_{X \in A} = 1)$$

$$= \mathbb{P}(X \in A)$$

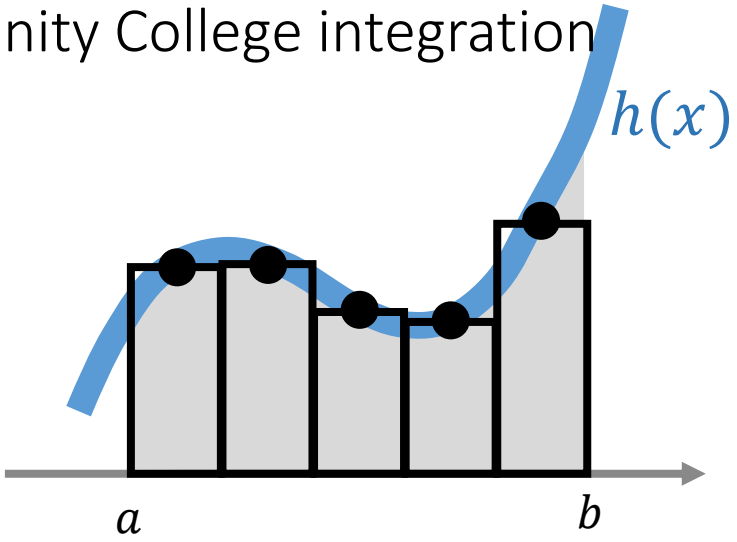
Let X be the location of a randomly thrown dart, and let $x_1, \ldots, x_n$ be some throws.

The probability of hitting A is

$$\mathbb{P}(X \in A) \approx \frac{1}{n} \sum_{i=1}^{n} 1_{x_i \in A}$$



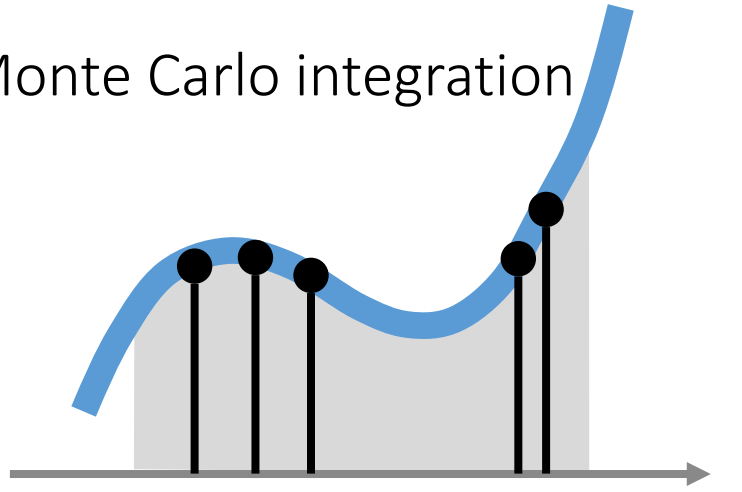
Trinity College integration



$$\int_{x=a}^{b} h(x)\,dx \approx \sum_{i=1}^{n} h(x_i)\frac{b-a}{n}$$

where x_i is the midpoint of interval i

Monte Carlo integration



Let's instead approximate this integral using Monte Carlo. Let $X \sim U[a,b]$.

By Monte Carlo,

$$\underbrace{\mathbb{E}h(X)}_{\int_{x=a}^{b} h(x)\,\Pr_X(x)\,dx} \approx \frac{1}{n}\sum_{i=1}^{n} h(x_i) \quad \text{where } x_1, \ldots, x_n \text{ sampled from } X$$

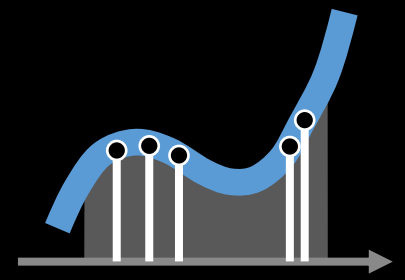
$$\int_{x=a}^{b} h(x)\,\Pr_X(x)\,dx = \int_{x=a}^{b} h(x)\,\frac{1}{b-a}\,dx$$

Thus,

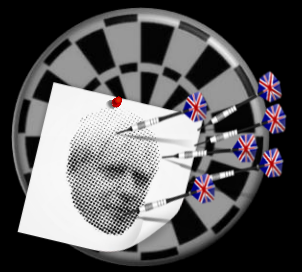
$$\int_{x=a}^{b} h(x)\,dx \approx \frac{b-a}{n}\sum_{i=1}^{n} h(x_i)$$

COMPUTATIONAL METHODS

- If we want $\mathbb{E}h(X)$ but the maths is too complicated, we can approximate
$$\mathbb{E}h(X) \approx n^{-1} \sum_{i=1}^{n} h(x_i)$$
where $x_1, \ldots, x_n$ are sampled from X



- This approximation also tells us how to estimate probabilities, since
$$\mathbb{P}(X \in A) = \mathbb{E}1_{X \in A}$$



- In Bayesian analysis, our aim is to estimate probabilities, e.g. $\mathbb{P}(\Theta \in [.2, .3] | X = 1)$. How can we do this computationally?