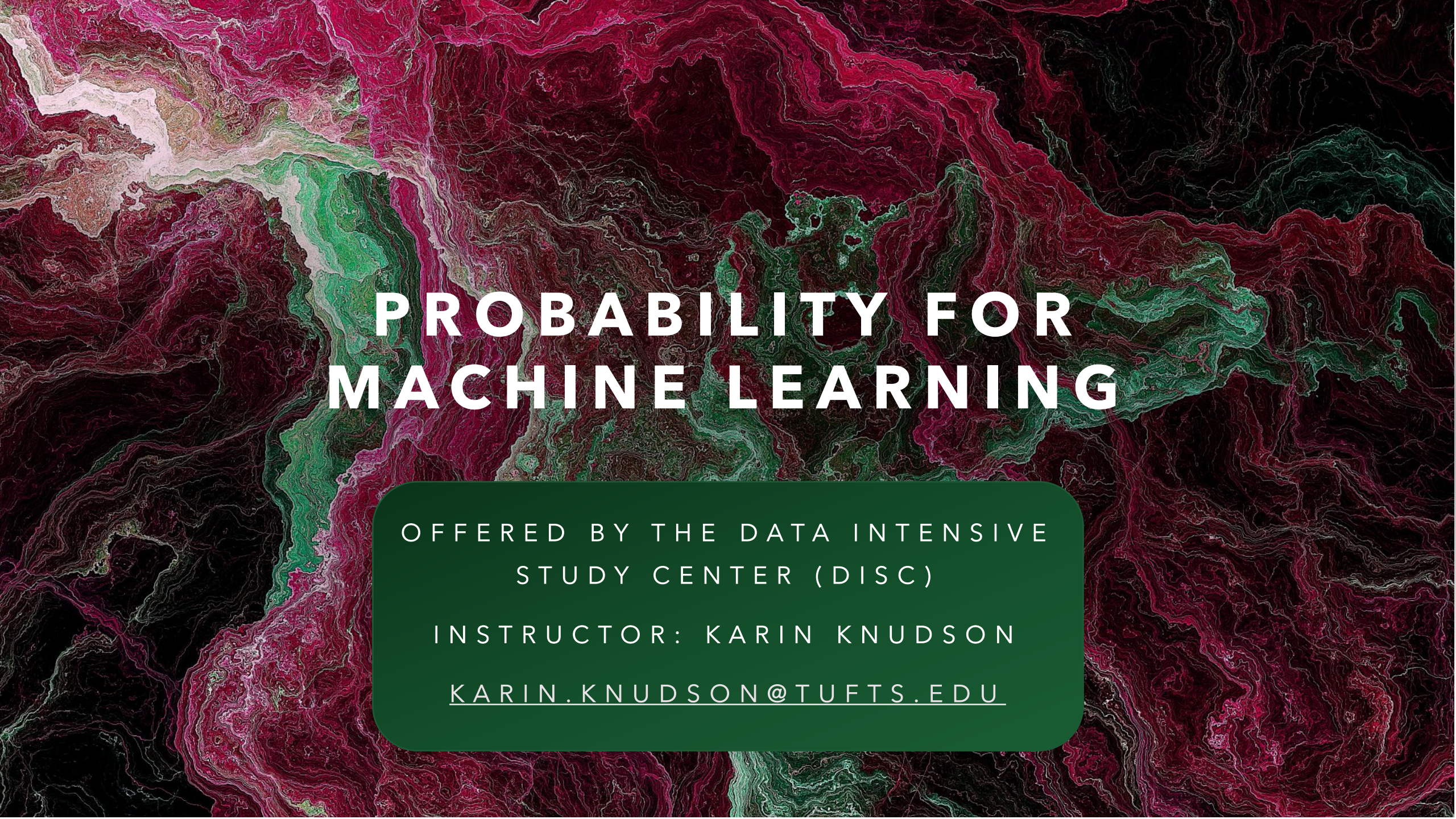


PROBABILITY FOR MACHINE LEARNING

OFFERED BY THE DATA INTENSIVE STUDY CENTER (DISC)

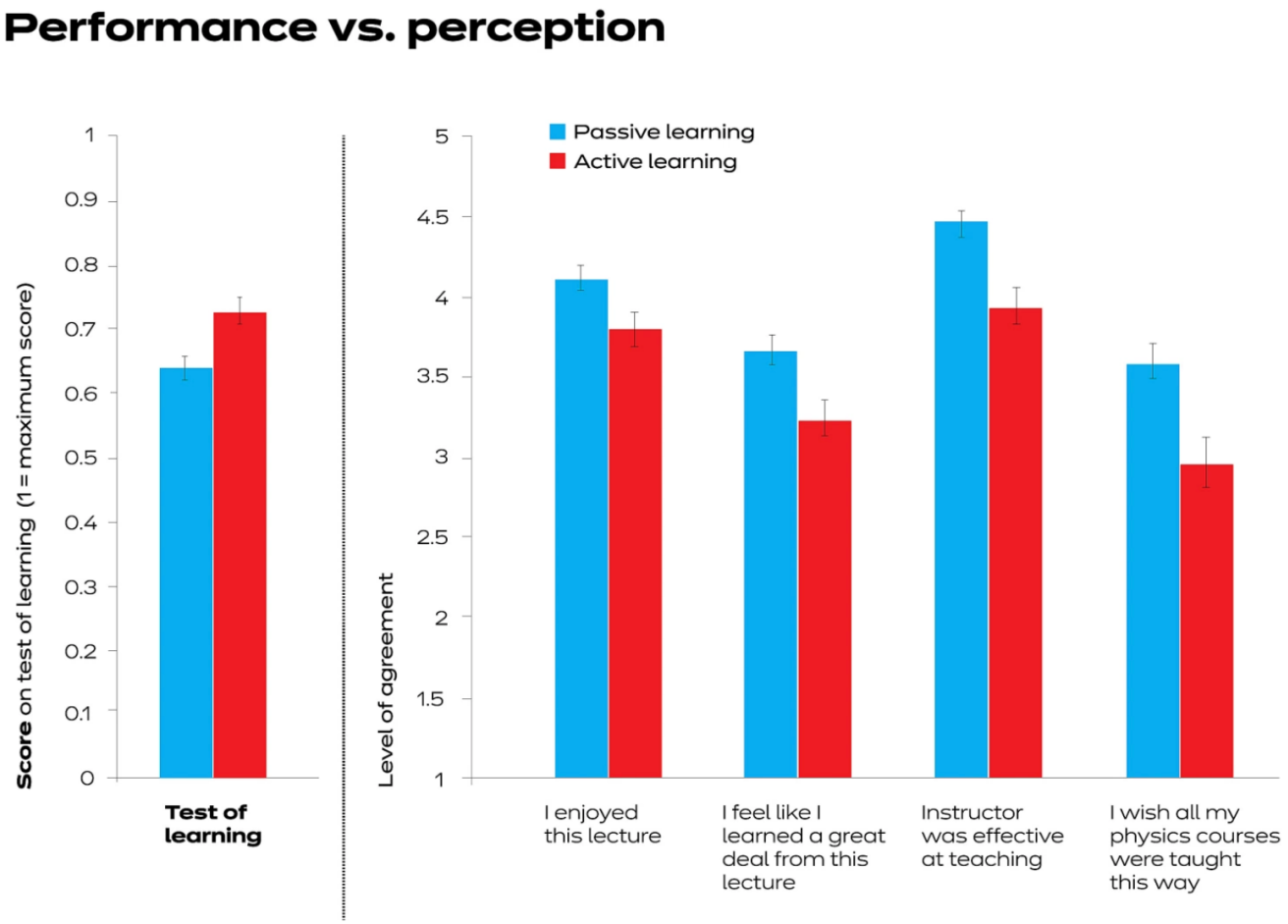
INSTRUCTOR: KARIN KNUDSON

KARIN.KNUDSON@TUFTS.EDU

DAILY FORMAT

- Whole group instruction
- Hands-on practice in breakout groups
- Breaks

Source: "Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom," Louis Deslauriers, Logan S. McCarty, Kelly Miller, Kristina Callaghan, and Greg Kestin



https://news.harvard.edu/gazette/story/2019/09/study-shows-that-students-learn-more-when-taking-part-in-classrooms-that-employ-active-learning-strategies/

WEEK SCHEDULE

- Monday: Probability foundations
- Tuesday: Regression and classification*
- Wednesday: Regression and classification*
- Thursday: Unsupervised learning*
- Friday: Misc: (time series, deep learning, other topics)*

* Will be both exploring specific techniques and using these techniques to illustrate broader probabilistic principles

"The probability of flipping heads is 50%"

– what does this mean?

"The probability of flipping heads is 50%" – what does this mean?

- If we flip the coin many times, we expect about half of the flips to be heads
- 50% quantifies our degree of belief that the coin will be heads the next time we flip
- We might pay 50 cents to play a game where the coin is flipped once and we get 1 dollar if it is heads

RANDOM VARIABLES

- Discrete random variable – countable number of possible values:

  Coin flip: {heads, tails}

  Die roll: {1, 2, 3, 4, 5, 6}

  Number of times a neuron fires in a time interval: {0, 1, 2, …}

- Continuous random variable – uncountable number of possible values:

  Temperature outside

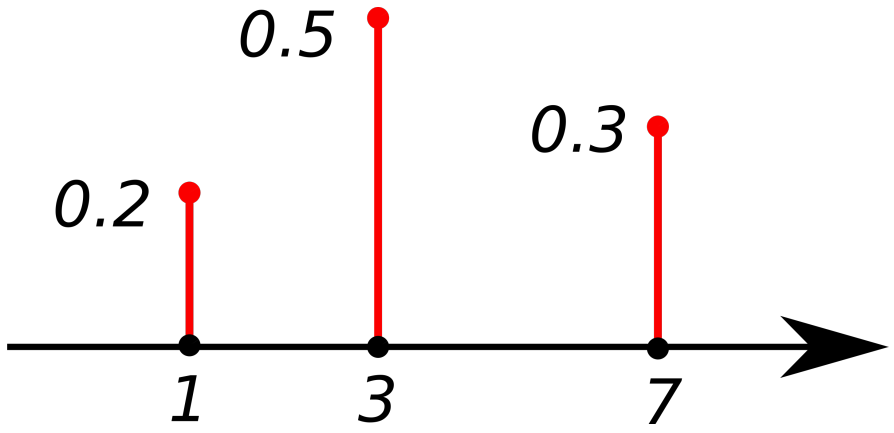
  Bias of a coin: [0, 1]

- Distinction can blur– e.g. approximating the count of vehicles on a road in a day with a continuous distribution

- Set of possible values a random variables: support

DISCRETE

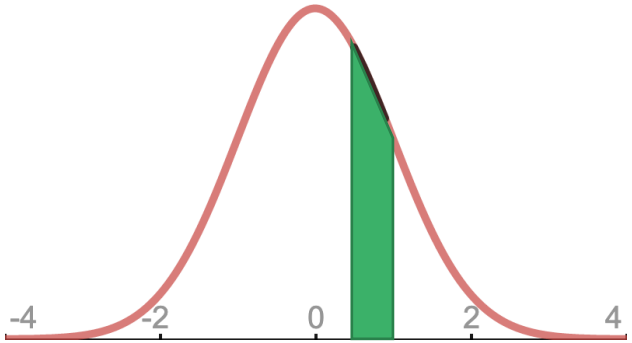
probability mass function (pmf)



$$\sum p(x_i) = 1$$
$$p(x_i) = P(X = x_i) \geq 0$$

CONTINUOUS

probability density function pdf



$$P(X \in (a, b)) = \int_a^b p(x)\, dx$$
$$\int p(x)\, dx = 1$$
$$p(x) \geq 0$$

EXPECTATION

$$\mathbb{E}(X) = \sum xp(x)$$
$$\mathbb{E}(f(x)) = \sum f(x)p(x)$$

$$\mathbb{E}(X) = \int xp(x)\,dx$$
$$\mathbb{E}(f(x)) = \int f(x)p(x)\,dx$$

Weighted average (weighted by probability)

Linear: $$\begin{aligned}\mathbb{E}(aX+b) &= \sum (ax+b)p(x) \\ &= \sum axp(x) + \sum bp(x) \\ &= a\sum xp(x) + b\sum p(x) \\ &= a\mathbb{E}(X) + b\end{aligned}$$

VARIANCE

$$\mathrm{var}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2]$$

Expected squared difference from the mean

$$\mathrm{var}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2$$

Proof: (exercise)

$$\mathrm{var}(aX + b) = \ ??\ \ (\mathrm{exercise})$$

COVARIANCE

$$\operatorname{cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]$$

	x = 1	x=2	
y=1	.1	.3	
y=2	.2	.1	
y=3	.2	0	
y=4	0	.1	

Joint probability
P(X=2, Y= 1)

	x = 1	x=2	
y=1	.1	.3	.4
y=2	.2	.1	.3
y=3	.2	0	.2
y=4	0	.1	.1
	.5	.5	

Joint probability
P(X=2, Y= 1)

Marginal probability
P(Y=2)

$$= \sum_x p(Y = 2, X = x)$$

	x = 1	x=2	
y=1	.1	.3	.4
y=2	.2	.1	.3
y=3	.2	0	.2
y=4	0	.1	.1
	.5	.5	

Marginal distribution of Y

Joint probability
P(X=2, Y= 1)

Marginal probability
P(Y=2)

$$= \sum_x p(Y = 2, X = x)$$

	x = 1	x=2	
y=1	.1	.3	.4
y=2	.2	.1	.3
y=3	.2	0	.2
y=4	0	.1	.1
	.5	.5	

Joint probability
P(X=2, Y= 1)

Marginal probability
P(Y=2)

$$= \sum_x p(Y = 2, X = x)$$

Marginal probability
P(X=2)

$$= \sum_y p(X = 2, Y = y)$$

	x = 1	x=2	
y=1	.1	.3	.4
y=2	.2	.1	.3
y=3	.2	0	.2
y=4	0	.1	.1
	.5	.5	

Joint probability
P(X=2, Y= 1)

Marginal probability
P(Y=2)

$$= \sum_x p(Y = 2, X = x)$$

Marginal probability
P(X=2)

$$= \sum_y p(X = 2, Y = y)$$

	x = 1	x=2	
y=1	.1	.3	.4
y=2	.2	.1	.3
y=3	.2	0	.2
y=4	0	.1	.1
	.5	.5	

Joint probability
P(X=2, Y= 1)

Marginal probability
P(Y=2)

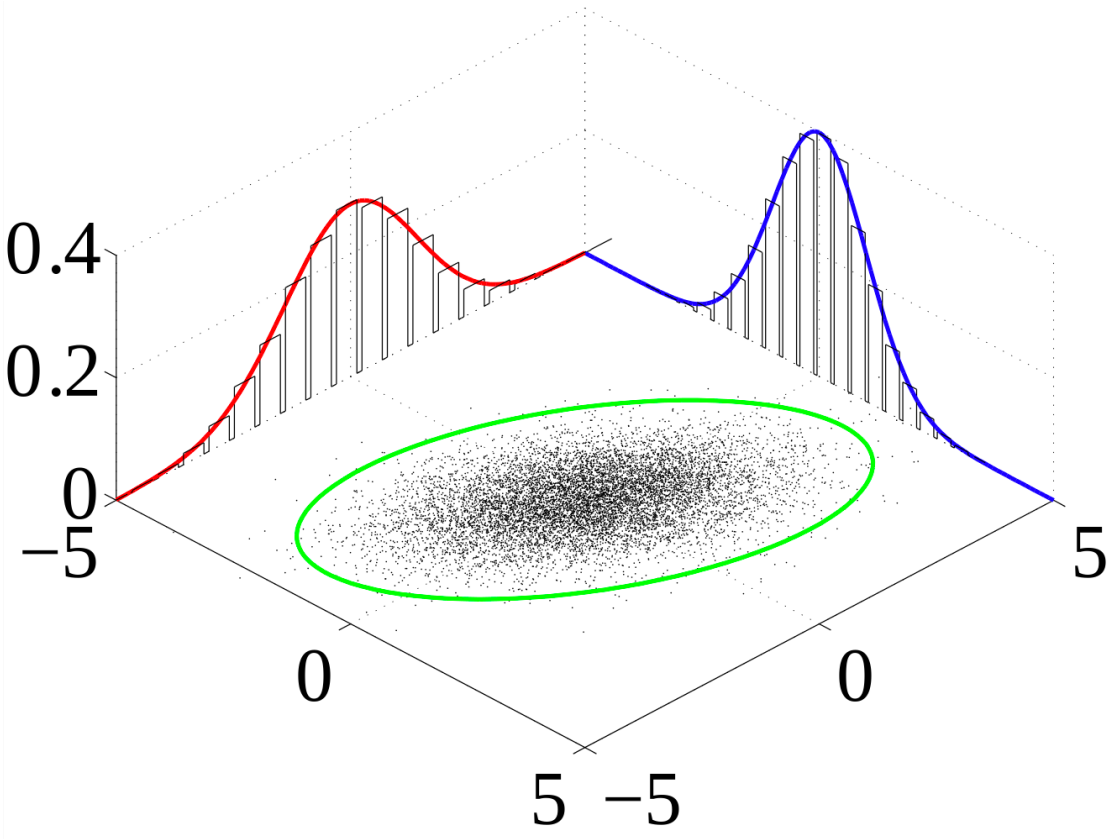
$$= \sum_x p(Y = 2, X = x)$$

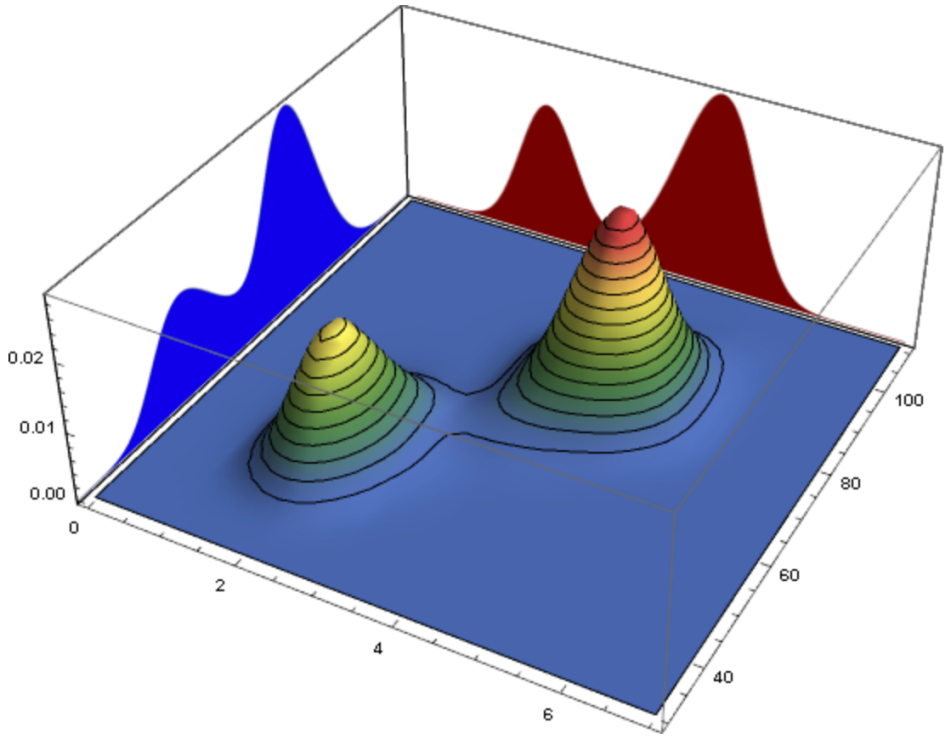
Marginal probability
P(X=2)

$$= \sum_y p(X = 2, Y = y)$$

Conditional probability:

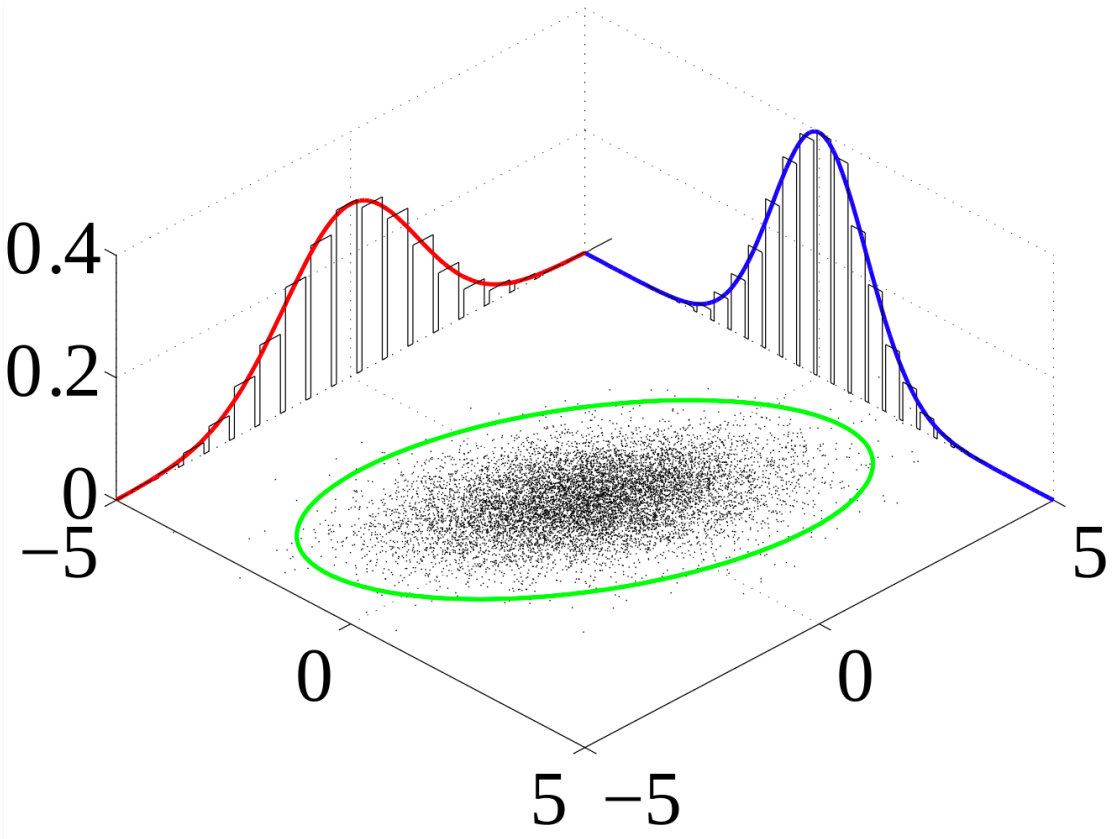
$$P(X = 2|Y = 2) = \frac{P(X=2,Y=2)}{P(Y=2)} = \frac{.1}{.3} = \frac{1}{3}$$

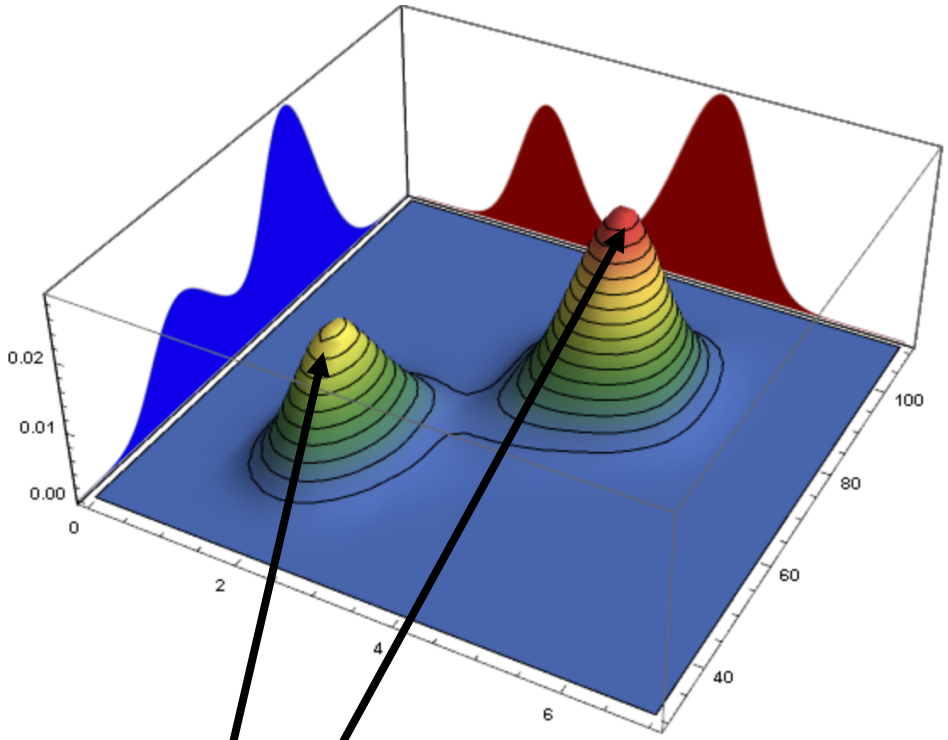




https://www.wolfram.com/mathematica/new-in-8/statistical-visualization/visualize-marginal-distributions-for-estimated-dis.html

CONTINUOUS MULTIVARIATE DISTRIBUTIONS





https://www.wolfram.com/mathematica/new-in-8/statistical-visualization/visualize-marginal-distributions-for-estimated-dis.html

modes

(this is a **multimodal** distribution)

SUM RULE

$$P(X) = \sum_{Y} P(X, Y)$$

$$p(X) = \int p(X, Y)\, dY$$

PRODUCT RULE

$$P(X, Y) = P(X|Y)P(Y)$$

Conditional probability

$$P(X|Y) = \frac{P(X, Y)}{P(Y)}$$

INDEPENDENCE

We say X and Y are **independent** if $P(X, Y) = P(X)\,P(Y)$

BAYES THEOREM

How to Think Like an Epidemiologist

Don't worry, a little Bayesian analysis won't hurt you.

https://www.nytimes.com/2020/08/04/science/coronavirus-bayes-statistics-math.html

BAYES THEOREM

$$P(\theta|Y) = \frac{P(Y|\theta)P(\theta)}{P(Y)}$$

BAYES THEOREM

$$P(\theta|Y) = \frac{P(Y|\theta)P(\theta)}{P(Y)}$$

$$= \frac{P(Y|\theta)P(\theta)}{\sum_\theta P(Y|\theta)P(\theta)}$$

BAYES THEOREM

$$\overbrace{P(\theta|Y)}^{\text{posterior}} = \frac{\overbrace{P(Y|\theta)}^{\text{likelihood}}\overbrace{P(\theta)}^{\text{prior}}}{P(Y)}$$

Y = observed data

θ = parameters of interest (unknown)

$$= \frac{P(Y|\theta)P(\theta)}{\sum_\theta P(Y|\theta)P(\theta)}$$

EX: ESTIMATING BIAS OF A COIN

$$\overbrace{P(\theta|Y)}^{\text{posterior}} = \frac{\overbrace{P(Y|\theta)}^{\text{likelihood}}\overbrace{P(\theta)}^{\text{prior}}}{P(Y)}$$

Y = observed data = # of heads observed in 20 flips

θ = parameter of interest = probability that the coin lands heads

EX: ESTIMATING BIAS OF A COIN

$$\underbrace{P(\theta|Y)}_{\text{posterior}} = \frac{\overbrace{P(Y|\theta)}^{\text{likelihood}}\overbrace{P(\theta)}^{\text{prior}}}{P(Y)}$$

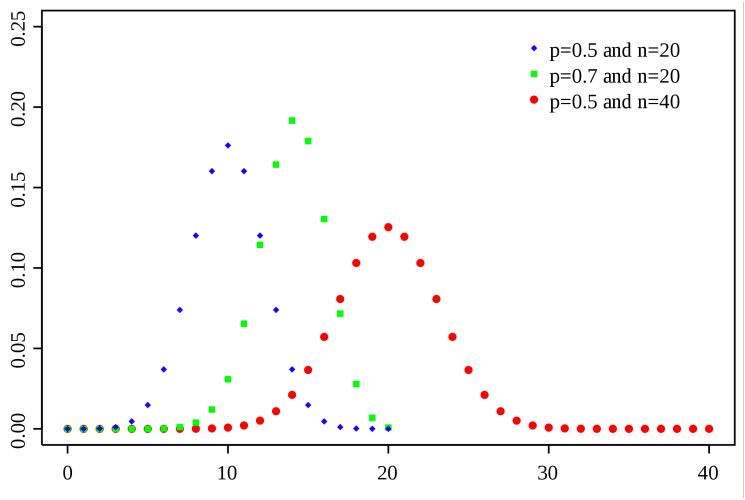
Y = observed data = # of heads observed in 20 flips

θ = parameter of interest = probability that the coin lands heads

likelihood

$$Y|\theta \sim \text{Binomial}(20, \theta)$$

$$p(Y = y|\theta) = \binom{20}{y}\theta^y(1-\theta)^{20-y}$$



EX: ESTIMATING BIAS OF A COIN

$$\underbrace{P(\theta|Y)}_{\text{posterior}} = \frac{\overbrace{P(Y|\theta)}^{\text{likelihood}}\overbrace{P(\theta)}^{\text{prior}}}{P(Y)}$$

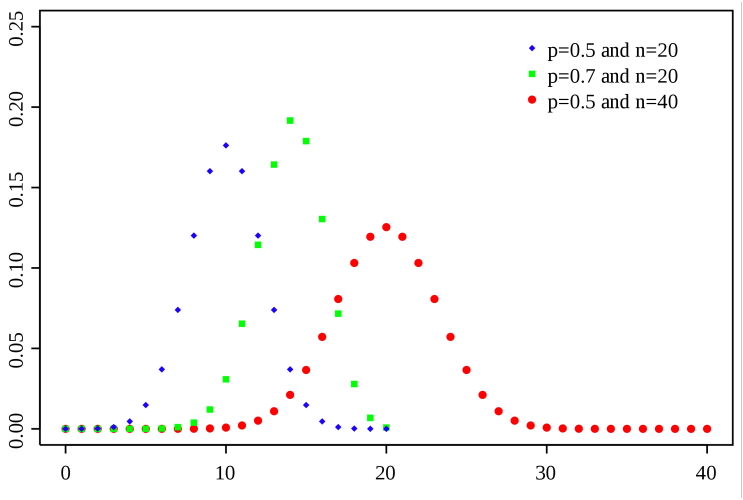
Y = observed data = # of heads observed in 20 flips

θ = parameter of interest = probability that the coin lands heads

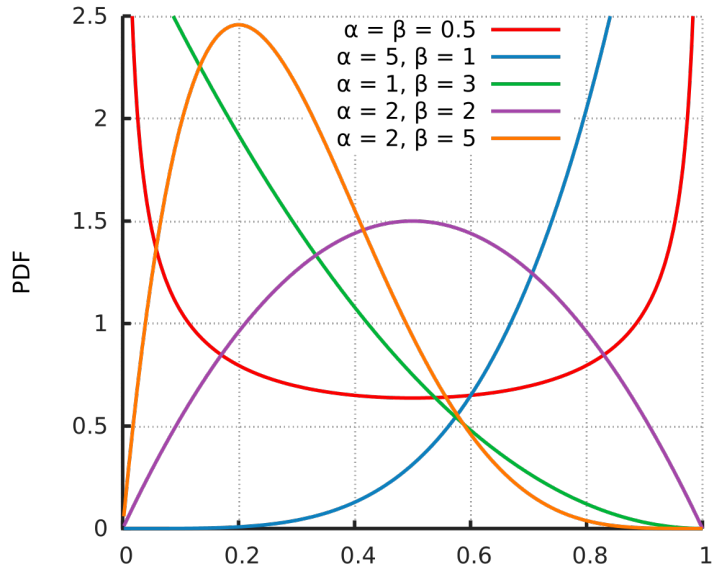
likelihood

$$Y|\theta \sim \text{Binomial}(20, \theta)$$

$$p(Y = y|\theta) = \binom{20}{y}\theta^y(1-\theta)^{20-y}$$



prior



$$\theta \sim \text{Beta}(a, b)$$

$$p(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}$$

EX: ESTIMATING BIAS OF A COIN

$$\underbrace{P(\theta|Y)}_{\text{posterior}} = \frac{\overbrace{P(Y|\theta)}^{\text{likelihood}}\overbrace{P(\theta)}^{\text{prior}}}{P(Y)}$$

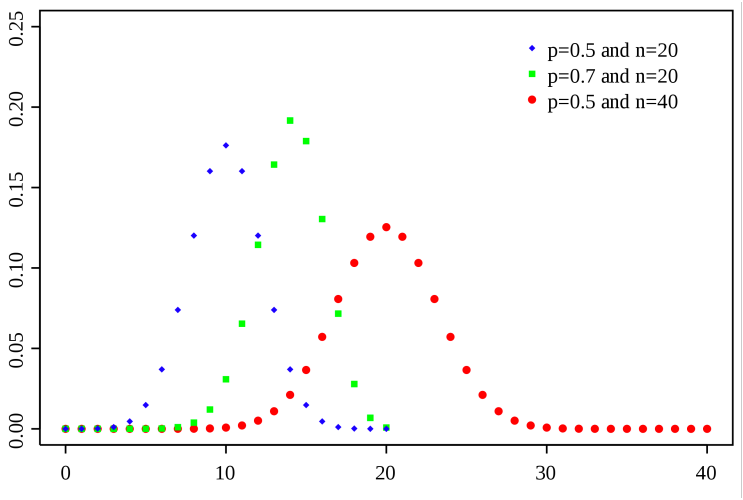
Y = observed data = # of heads observed in 20 flips

θ = parameter of interest = probability that the coin lands heads

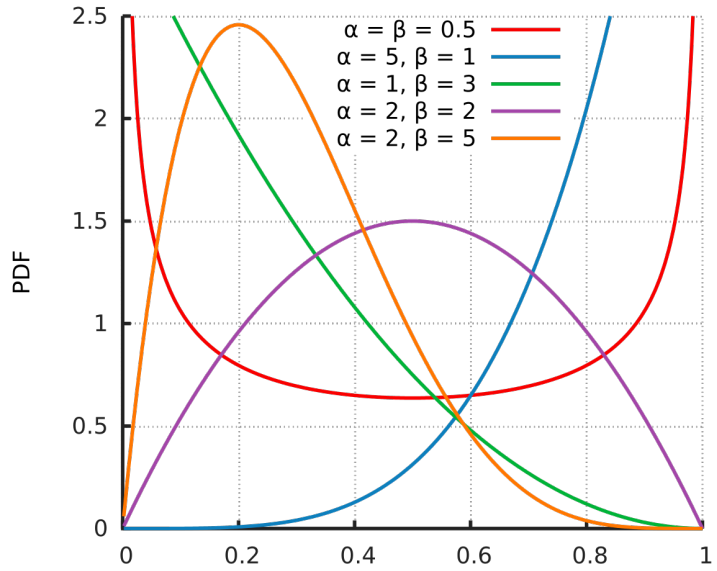
likelihood

$$Y|\theta \sim \text{Binomial}(20, \theta)$$

$$p(Y = y|\theta) = \binom{20}{y}\theta^y(1-\theta)^{20-y}$$



prior



$$\theta \sim \text{Beta}(a, b) \qquad \theta \sim \text{Beta}(2, 2)$$

$$p(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)} \qquad p(\theta) = \frac{\theta(1-\theta)}{B(2,2)}$$

EX: ESTIMATING BIAS OF A COIN

$$\underbrace{P(\theta|Y)}_{\text{posterior}} = \frac{\overbrace{P(Y|\theta)}^{\text{likelihood}}\overbrace{P(\theta)}^{\text{prior}}}{P(Y)}$$

Y = observed data = # of heads observed in 20 flips

θ = parameter of interest = probability that the coin lands heads

likelihood

$$Y \sim \text{Binomial}(20, \theta)$$

$$p(Y = y|\theta) = \binom{20}{y}\theta^y(1-\theta)^{20-y}$$

prior

$$\theta \sim \text{Beta}(a, b) \qquad \theta \sim \text{Beta}(2, 2)$$

$$p(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)} \qquad p(\theta) = \frac{\theta(1-\theta)}{B(2,2)}$$

posterior

EX: ESTIMATING BIAS OF A COIN

$$\underbrace{P(\theta|Y)}_{\text{posterior}} = \frac{\overbrace{P(Y|\theta)}^{\text{likelihood}}\overbrace{P(\theta)}^{\text{prior}}}{P(Y)}$$

Y = observed data = # of heads observed in 20 flips

θ = parameter of interest = probability that the coin lands heads

likelihood

$$Y \sim \text{Binomial}(20, \theta)$$

$$p(Y = y|\theta) = \binom{20}{y}\theta^y(1-\theta)^{20-y}$$

prior

$$\theta \sim \text{Beta}(a, b) \qquad \theta \sim \text{Beta}(2, 2)$$

$$p(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)} \qquad p(\theta) = \frac{\theta(1-\theta)}{B(2,2)}$$

posterior

$$p(\theta|Y = y) = \frac{p(Y=y|\theta)p(\theta)}{p(Y=y)} \propto p(Y = y|\theta)p(\theta)$$

EX: ESTIMATING BIAS OF A COIN

$$\underbrace{P(\theta|Y)}_{\text{posterior}} = \frac{\overbrace{P(Y|\theta)}^{\text{likelihood}}\overbrace{P(\theta)}^{\text{prior}}}{P(Y)}$$

Y = observed data = # of heads observed in 20 flips

θ = parameter of interest = probability that the coin lands heads

likelihood

$$Y \sim \text{Binomial}(20, \theta)$$

$$p(Y = y|\theta) = \binom{20}{y}\theta^y(1-\theta)^{20-y}$$

prior

$$\theta \sim \text{Beta}(a, b) \qquad \theta \sim \text{Beta}(2, 2)$$

$$p(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)} \qquad p(\theta) = \frac{\theta(1-\theta)}{B(2,2)}$$

posterior

$$p(\theta|Y = y) = \frac{p(Y=y|\theta)p(\theta)}{p(Y=y)} \propto p(Y = y|\theta)p(\theta)$$

$$\propto \binom{20}{y}\theta^y(1-\theta)^{20-y}\theta^{a-1}(1-\theta)^{b-1}$$

$$\propto \theta^y(1-\theta)^{20-y}\theta(1-\theta)$$

$$\propto \theta^{y+1}(1-\theta)^{21-y} = \theta^{y+2-1}(1-\theta)^{20-y+2-1}$$

"is proportional to"
- we drop terms in this product that don't depend on theta

EX: ESTIMATING BIAS OF A COIN

$$\underbrace{P(\theta|Y)}_{\text{posterior}} = \frac{\overbrace{P(Y|\theta)}^{\text{likelihood}}\overbrace{P(\theta)}^{\text{prior}}}{P(Y)}$$

Y = observed data = # of heads observed in 20 flips

θ = parameter of interest = probability that the coin lands heads

likelihood

$$Y \sim \text{Binomial}(20, \theta)$$

$$p(Y = y|\theta) = \binom{20}{y}\theta^y(1-\theta)^{20-y}$$

prior

$$\theta \sim \text{Beta}(a, b) \qquad \theta \sim \text{Beta}(2, 2)$$

$$p(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)} \qquad p(\theta) = \frac{\theta(1-\theta)}{B(2,2)}$$

posterior

$$p(\theta|Y = y) = \frac{p(Y=y|\theta)p(\theta)}{p(Y=y)} \propto p(Y = y|\theta)p(\theta)$$

$$\propto \binom{20}{y}\theta^y(1-\theta)^{20-y}\theta^{a-1}(1-\theta)^{b-1}$$

$$\propto \theta^y(1-\theta)^{20-y}\theta(1-\theta)$$

$$\propto \theta^{y+1}(1-\theta)^{21-y} = \theta^{y+2-1}(1-\theta)^{20-y+2-1}$$

"is proportional to"
- we drop terms in this product that don't depend on theta

$$\theta|Y = y \sim \text{Beta}(y+2, (20-y)+2)$$
$$\theta|Y = y \sim \text{Beta}(y+a, (20-y)+b)$$

EX: ESTIMATING BIAS OF A COIN

$$\underbrace{P(\theta|Y)}_{\text{posterior}} = \frac{\overbrace{P(Y|\theta)}^{\text{likelihood}}\overbrace{P(\theta)}^{\text{prior}}}{P(Y)}$$

Y = observed data = # of heads observed in 20 flips

θ = parameter of interest = probability that the coin lands heads

likelihood

$$Y \sim \text{Binomial}(20, \theta)$$

$$p(Y = y|\theta) = \binom{20}{y}\theta^y(1-\theta)^{20-y}$$

prior

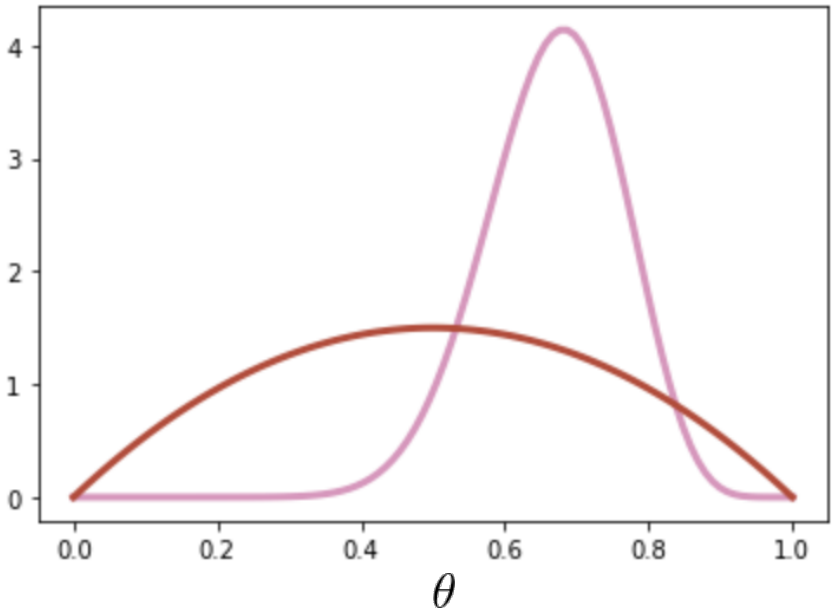
$$\theta \sim \text{Beta}(a, b) \qquad \theta \sim \text{Beta}(2, 2)$$

$$p(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)} \qquad p(\theta) = \frac{\theta(1-\theta)}{B(2,2)}$$

posterior

data: y = 14

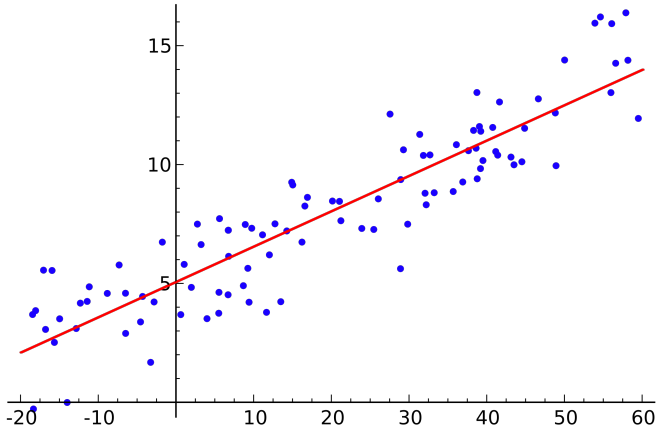
$$\theta|Y = 14 \sim \text{Beta}(16, 8)$$



EX: ESTIMATING BIAS OF A COIN: REFLECTION

- Wrote down our assumptions about the data generating process
- Used existing knowledge of coins to set a prior
- Computed the posterior distribution using the prior and the posterior, updating our beliefs about the parameter of interest
- Posterior distribution for theta had smaller variance than the prior distribution for theta
- Were “lucky” in that the posterior distribution turned out to have a convenient, recognizable form - this won’t always be the case. (When the prior and posterior have the same form, we have what we call **conjugacy**.)

LINEAR REGRESSION



$$y = w_0 + w_1 X_1$$

"independent and identically distributed"

iid

$$y_i | X_i, w_0, w_1 \sim Normal(w_0 + w_1 X_1, \sigma^2)$$

Likelihood

$$p(y_i) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(y_i - (w_0 + w_1 X_i))^2}{2\sigma^2}\right)$$

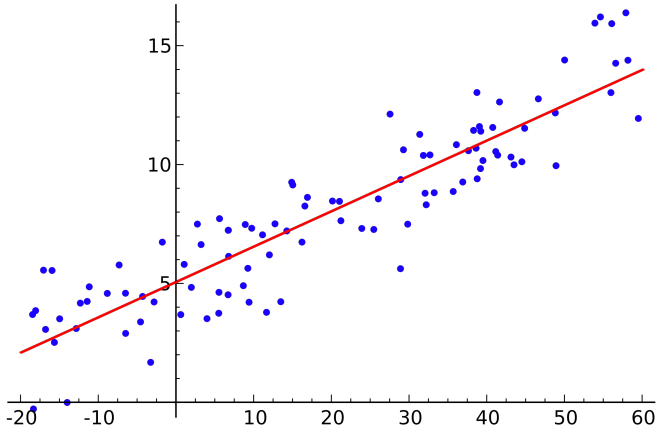
$$p(\mathbf{y}) = \prod_{i=1}^{N} \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(y_i - (w_0 + w_1 X_i))^2}{2\sigma^2}\right)$$

To maximize this with respect to w, it suffices to minimize:

$$\sum_{i=1}^{N} (y_i - (w_0 + w_1 X_i))^2$$

$$\log p(\mathbf{y} | X, w) = -\sum_{i=1}^{N} \left(\frac{(y_i - (w_0 + w_1 X_i))^2}{2\sigma^2}\right) + \sum_{i=1}^{N} \log(1/(2\pi\sigma^2))$$

LINEAR REGRESSION



$$y = w_0 + w_1 X_1$$

"independent and identically distributed"

iid

$$y_i | X_i, w_0, w_1 \sim Normal(w_0 + w_1 X_1, \sigma^2)$$

Likelihood

$$p(y_i) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(y_i - (w_0 + w_1 X_i))^2}{2\sigma^2}\right)$$

$$p(\mathbf{y}) = \prod_{i=1}^{N} \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(y_i - (w_0 + w_1 X_i))^2}{2\sigma^2}\right)$$

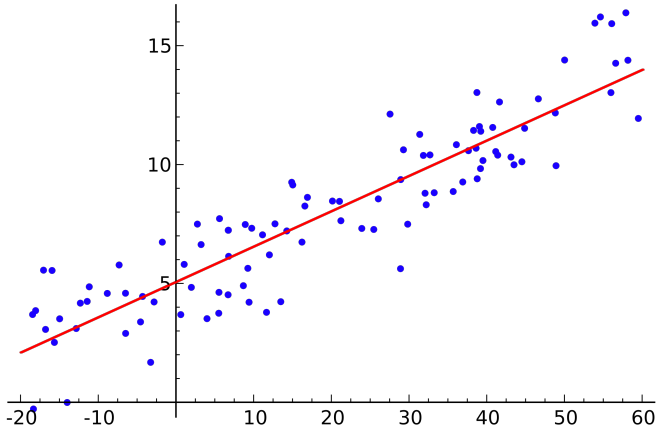
To maximize this with respect to w, it suffices to minimize:

$$\sum_{i=1}^{N} (y_i - (w_0 + w_1 X_i))^2$$

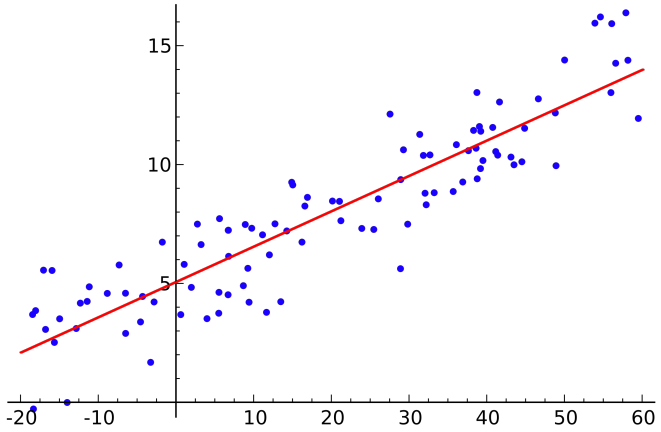
$$\log p(\mathbf{y} | X, w) = -\sum_{i=1}^{N} \left(\frac{(y_i - (w_0 + w_1 X_i))^2}{2\sigma^2}\right) + \sum_{i=1}^{N} \log(1/(2\pi\sigma^2))$$

LINEAR REGRESSION

$$y = w_0 + w_1 X_1$$



LINEAR REGRESSION



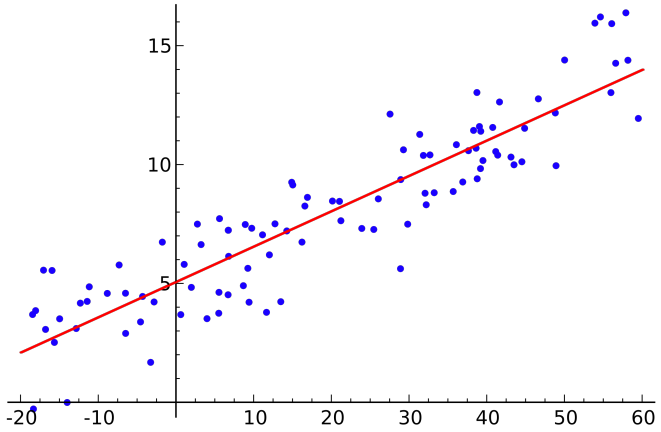
$$y = w_0 + w_1 X_1$$

"independent and identically distributed"

$$y_i | X_i, w_0, w_1 \overset{\text{iid}}{\sim} Normal(w_0 + w_1 X_1, \sigma^2)$$

$$p(y_i) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(y_i - (w_0 + w_1 X_i))^2}{2\sigma^2} \right)$$

LINEAR REGRESSION



$$y = w_0 + w_1 X_1$$

"independent and identically distributed"

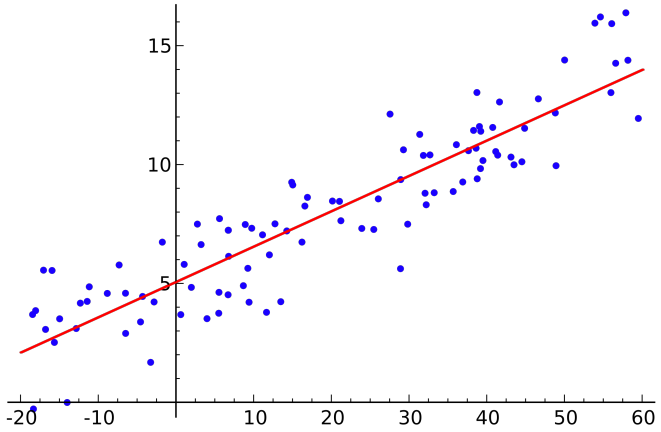
$$y_i | X_i, w_0, w_1 \overset{\text{iid}}{\sim} Normal(w_0 + w_1 X_1, \sigma^2)$$

Likelihood

$$p(y_i) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(y_i - (w_0 + w_1 X_i))^2}{2\sigma^2}\right)$$

$$p(\mathbf{y}) = \prod_{i=1}^{N} \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(y_i - (w_0 + w_1 X_i))^2}{2\sigma^2}\right)$$

LINEAR REGRESSION



$$y = w_0 + w_1 X_1$$

"independent and identically distributed"

iid

$$y_i | X_i, w_0, w_1 \sim Normal(w_0 + w_1 X_1, \sigma^2)$$

Likelihood

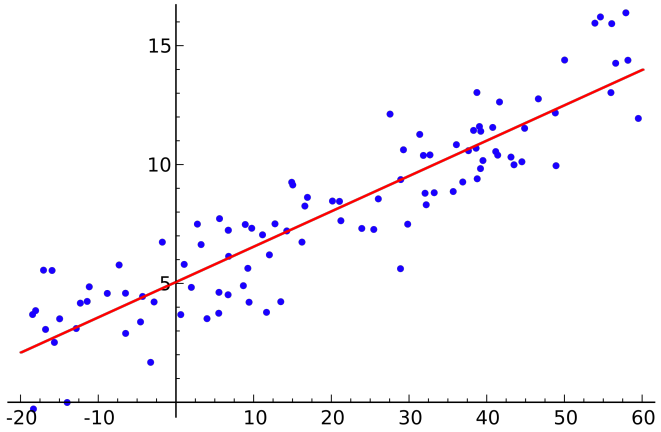
$$p(y_i) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(y_i - (w_0 + w_1 X_i))^2}{2\sigma^2}\right)$$

$$p(\mathbf{y}) = \prod_{i=1}^{N} \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(y_i - (w_0 + w_1 X_i))^2}{2\sigma^2}\right)$$

$$\log p(\mathbf{y} | X, w) = -\sum_{i=1}^{N} \left(\frac{(y_i - (w_0 + w_1 X_i))^2}{2\sigma^2}\right) + \sum_{i=1}^{N} \log(1/(2\pi\sigma^2))$$

Log-likelihood

LINEAR REGRESSION



$$y = w_0 + w_1 X_1$$

"independent and identically distributed"

iid

$$y_i | X_i, w_0, w_1 \sim Normal(w_0 + w_1 X_1, \sigma^2)$$

Likelihood

$$p(y_i) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(y_i - (w_0 + w_1 X_i))^2}{2\sigma^2}\right)$$

$$p(\mathbf{y}) = \prod_{i=1}^{N} \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(y_i - (w_0 + w_1 X_i))^2}{2\sigma^2}\right)$$

To maximize this with respect to w, it suffices to minimize:

$$\sum_{i=1}^{N} (y_i - (w_0 + w_1 X_i))^2$$

$$\log p(\mathbf{y} | X, w) = -\sum_{i=1}^{N} \left(\frac{(y_i - (w_0 + w_1 X_i))^2}{2\sigma^2}\right) + \sum_{i=1}^{N} \log(1/(2\pi\sigma^2))$$

Log-likelihood

REFLECTION

What did you learn?

What's something that made sense to you?

What was your "muddiest point" from today?