Expected value

The expected value of a random variable (RV) X is

$$E[X] = \sum_{x} x\, P(X = x)$$

Also, called the average or mean of X.

For example, if X is a 6-sided die, the expected value is:

$$E[X] = 1 * P(X = 1) + 2 * P(X = 1) + \cdots + 6 * P(X = 6)$$
$$= 1 * \frac{1}{6} + 2 * \frac{1}{6} + \cdots + 6 * \frac{1}{6}$$
$$= (1 + 2 + 3 + 4 + 5 + 6) * \frac{1}{6} = \frac{21}{6} = 3.5$$

Expected value

The expected value of a random variable (RV) X is

$$E[X] = \sum_x x\, P(X = x)$$

Also, called the average or mean of X.

Note:

1. Often interested in the expected value of a function $f(X)$. Since RVs are just functions, $E[f(X)] = \sum f(x)P(X = x)$.
2. **Indicator functions,** $I(A)$ (=1 if A is true, else 0), are often used to convert expectations to probabilities.
3. Expected value is linear. For RVs X and Y and constants c and d, $E[cX + dY] = cE[X] + dE[Y]$.

Expected value

The expected value of a random variable (RV) X is

$$E[X] = \sum_{x} x\, P(X = x)$$

For example, lets say you want to know the likely number of heads from 6 coin flips.

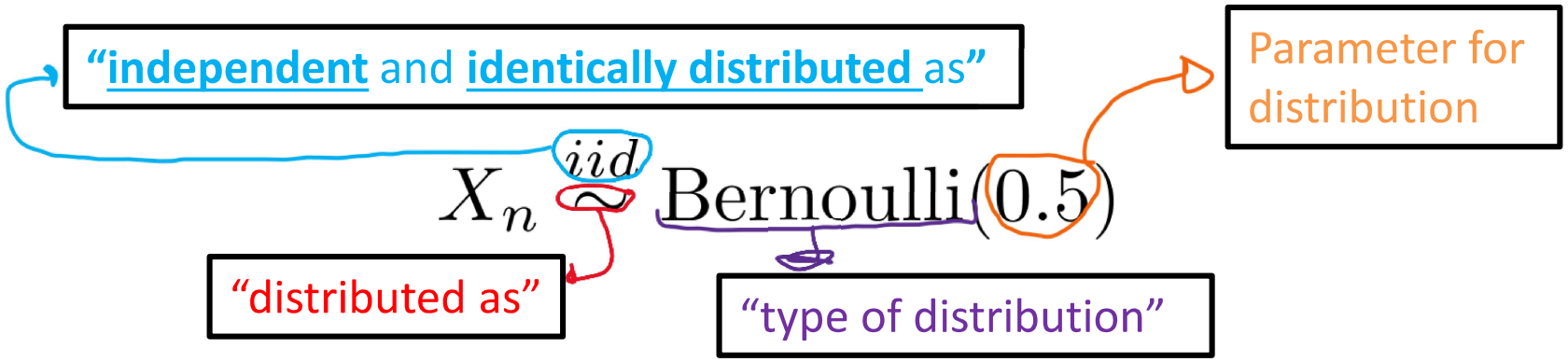
Let X_n be the outcome from the *n*-th coin flip.

Note*:* Heads is usually 1 and Tails is usually 0.

$$E[X_1 + X_2 + X_3 + X_4 + X_5 + X_6]$$
$$= E[X_1] + E[X_2] + E[X_3] + E[X_4] + E[X_5] + E[X_6]$$
$$= 6\, E[X_1] = 6 * (1 * 0.5 + 0 * 0.5) = 3$$

The Bernoulli Distribution

N fair coinflips $X_1, X_2, \ldots, X_N$ can be represented as



The Bernoulli Distribution

N fair coinflips $X_1, X_2, \ldots, X_N$ can be represented as

$$X_n \overset{iid}{\sim} \text{Bernoulli}(0.5) \qquad \theta \in [0, 1]$$

Heads or 1

with probability 0.5

Continuum from 0 to 1 (inclusive)

$$P(X_n) = \begin{cases} H/1 & \text{w. p. } 0.5 \\ T/0/-1 & \text{o. w.} \end{cases}$$

otherwise

Tails or 0 or -1

$$E[X_n] = \sum_{a \in \{1,0\}} aP(X_n = a) = 1 \times P(X_n = 1) + 0 \times P(X_n = 0) = 1 \times 1/2 + 0 \times 1/2 = 1/2$$

What is the expected value for the sum of two coinflips?

$$E[X_1 + X_2] = E[X_1] + E[X_2] = 1/2 + 1/2 = 1$$

N coinflips?

$$E\left[\sum_{n=1}^{N}\right]$$

NA

Sum from 1 to N (with index n)

Probability, Bernoulli and the Expectation Indicator Function "Trick"

The **indicator function** $I(\cdot)$ returns 1 when its argument is true and 0 otherwise.

E.g., $I(10 * 5 < 0) = 0$ or $I(\pi > 3) = 1$

$$E\left[I(X_n = 1)\right]$$

$$E\left[I(X_n = 1)\right] = P(X_n = 1)$$

$\Rightarrow$Convert between expectations and probabilities

$\Rightarrow$Will be critical during the approximation unit

For an event A and random variable Y,

$$P(Y \in A) = E\left[I(Y \in A)\right]$$

- Admin Matters
- Bayes practice and thinking generatively
- Levels of analysis
- **Short rest 1**
- Math to cover:
  - Expected value
  - Continuous probability
  - Gaussian
  - Mixture modeling
  - Beta-Binomial (discrete enters the continuum)
- **Short rest 2**
- Student intros + one paper presentation
  - Go over paper presentation advice here too.

Probabilities in Continuous Land

Probability mass function: $P(\cdot)$

probability mass assigned to a specific value

Discrete random variable (rv) X and continuous rv Y

$$P(X = a) > 0 \qquad \forall a^*, P(Y = a) = 0$$

Probability density function: $p(\cdot)$

Relative likelihood of a value near that precise location

$$p(y = a) \geq 0$$

To get probabilities for continuous distributions, you calculate the probability it takes a value in some **region**.

The probability is the integral of the density for a region

$$P(Y \in A) = \int_A p(y = a)\,da$$

Integral symbol

Region to integrate

Very small change in a

Probabilities in Continuous Land

Probability mass function (PMF): $P(\cdot)$

probability mass assigned to a specific value

Discrete random variable (rv) X and continuous rv Y

$$P(X = a) > 0 \qquad \forall a, P(Y = a) = 0$$

Probability density function (PDF):

Relative likelihood of a value near that precise location

$$p(y = a) \geq 0$$

Note: $$\boxed{P(Y \in \Omega) = \int_{\Omega} p(y = a)da = 1}$$

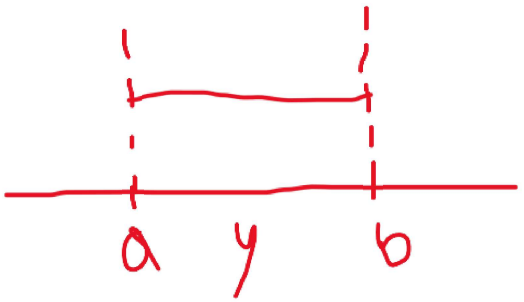
Cumulative distribution function (CDF)

$$F(y) = \int_{-\infty}^{y} p(a)da \qquad P(Y \in [a, b]) = F(b) - F(a)$$

Continuous Uniform Distribution

The continuous uniform distribution over the interval $[a,b]$ is

$$p(Y = y) = \frac{1}{b-a} I(y \in [a, b])$$



$$Y \sim \text{Uniform}([a, b])$$

"distributed as"

E.g., $Y \sim \text{Uniform}([0, 0.5])$ $\quad p(Y = 0.25) = \frac{1}{0.5 - 0} = 2$

What does that mean?

How about prob. that Y is less than 0.1?

$$P(Y \in [0, 0.1]) = \int_0^{0.1} 2dy = 2 \times 0.1 = 0.2$$

Expected value? $E[Y] = \int yp(y)dy$

$$E[Y] = \int_0^{0.5} yp(y)dy =$$

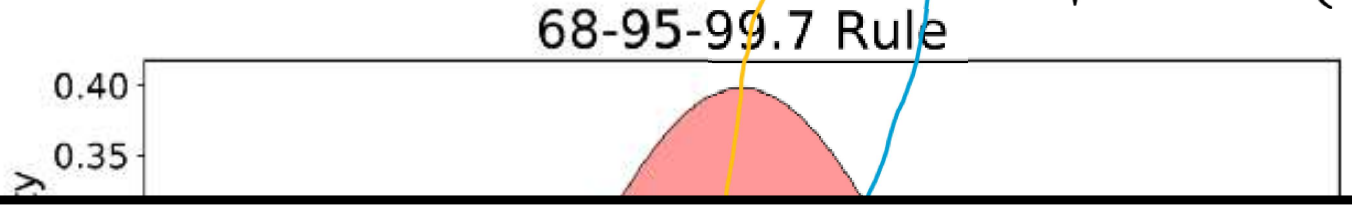
- Admin Matters
- Bayes practice and thinking generatively
- Levels of analysis
- **Short rest 1**
- Math to cover:
  - Expected value
  - Continuous probability
  - Gaussian
  - Mixture modeling
  - Beta-Binomial (discrete enters the continuum)
- **Short rest 2**
- Student intros + one paper presentation
  - Go over paper presentation advice here too.

Gaussian/Normal Distribution

$$Y \sim \text{Gaussian}(\mu, \sigma^2)$$

Alternative parametrization: $\tau = 1/\sigma^2$

$$p(y|\mu, \tau) = \sqrt{\frac{\tau}{2\pi}} \exp\left\{-\frac{\tau^2}{2}(y-\mu)^2\right\}$$



Note: There is no closed form solution for the CDF.

A special case of the CDF is the “Error Function”.

It is often written as Φ .

$$p(y|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}$$

Image from https://towardsdatascience.com/understanding-the-68-95-99-7-rule-for-a-normal-distribution-b7b7cbf760c2

- Admin Matters
- Bayes practice and thinking generatively
- Levels of analysis
- **Short rest 1**
- Math to cover:
  - Expected value
  - Continuous probability
  - Gaussian
  - Mixture modeling
  - Beta-Binomial (discrete enters the continuum)
- **Short rest 2**
- Student intros + one paper presentation
  - Go over paper presentation advice here too.

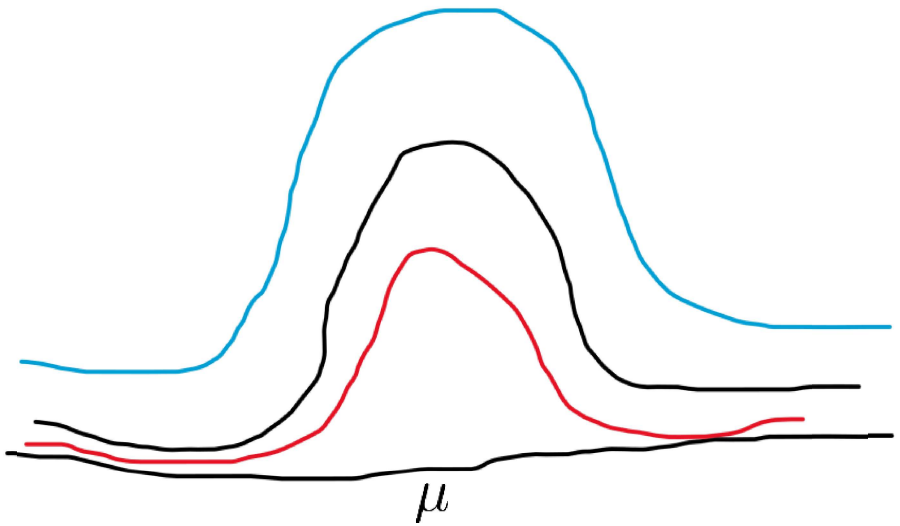
Posterior Mean For Normal Prior Given Known Variances

$$\mu \sim \text{Gaussian}(\mu_0, \sigma_0^2)$$

$$Y|\mu \sim \text{Gaussian}(\mu, \sigma_Y^2)$$

$$p(\mu|Y) = \frac{p(Y|\mu)p(\mu)}{p(Y)}$$

for some $k > 0$ (k does not depend on μ)



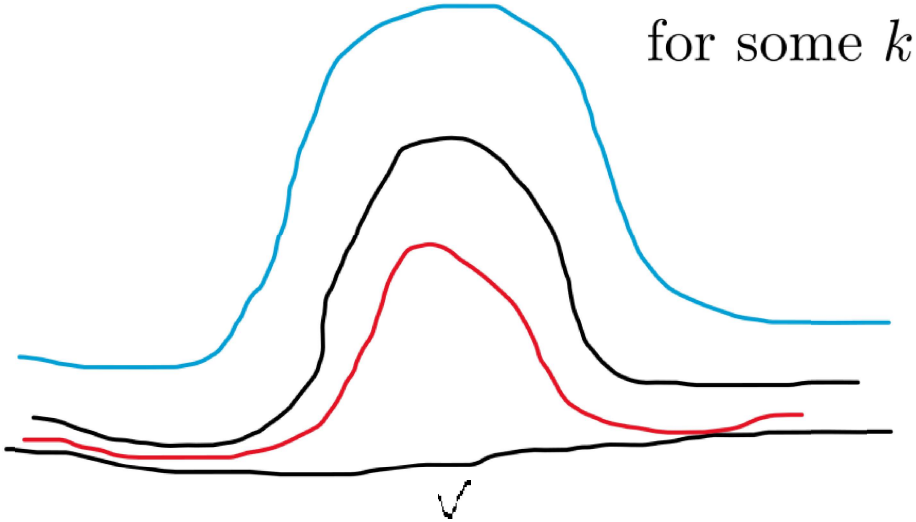
Varying k just stretches the function vertically – neither changes the functional form nor ordering of which points have greater values than others.

k will be the value that makes the integral of the function 1 for the given Y. WHY?

Posterior Mean For Normal Prior Given Known Variances

$$p(\mu|Y) = \frac{p(Y|\mu)p(\mu)}{p(Y)} = kp(Y|\mu)p(\mu)$$

for some $k > 0$ (k does not depend on μ)



Varying k just stretches the function vertically – neither changes the functional form nor ordering of which points have greater values than others.

k will be the value that makes the integral of the function 1 for the given Y. WHY?

Remember $\int kp(y|\mu)p(\mu)d\mu = 1.$

So, $\int kp(y|\mu)p(\mu) = 1$

$1/p(Y)$

"normalizing constant" or "partition function"

Posterior Mean For Normal Prior Given Known Variances

$$p(\mu|Y) = \frac{p(Y|\mu)p(\mu)}{p(Y)} = kp(Y|\mu)p(\mu) \propto p(Y|\mu)p(\mu)$$

$$= \frac{1}{\sigma_Y\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma_Y^2}(Y-\mu)^2\right\}$$

$$= \frac{1}{2\pi\sigma_Y\sigma_0} \exp\left\{-\frac{1}{2\sigma_Y^2}(Y-\mu)^2\right\}$$

$$\propto \exp\left\{-\frac{1}{2\sigma_Y^2}(Y^2 - 2\mu Y + \mu^2)\right\}$$

$$\propto \exp\left\{-\mu^2\left(\frac{1}{2\sigma_Y^2} + \frac{1}{2\sigma_0^2}\right) + 2\mu\left(\frac{Y}{2\sigma_Y^2} + \frac{\mu_0}{2\sigma_0^2}\right)\right\}$$

$$\exp\left\{-\frac{1}{2\sigma_1^2}(\mu-\mu_1)^2\right\}$$

Posterior Mean For Normal Prior Given Known Variances

$$p(\mu|Y) = \frac{p(Y|\mu)p(\mu)}{p(Y)} = kp(Y|\mu)p(\mu) \propto p(Y|\mu)p(\mu)$$

$$\propto \exp\left\{-\mu^2\left(\frac{1}{2\sigma_Y^2} + \frac{1}{2\sigma_0^2}\right) + 2\mu\left(\frac{Y}{2\sigma_Y^2} + \frac{\mu_0}{2\sigma_0^2}\right)\right\}$$

$$\exp\left\{-\frac{1}{2\sigma_1^2}(\mu - \mu_1)^2\right\}$$

$$\sigma_1^2 = \left(\frac{1}{\sigma_x^2} + \frac{1}{\sigma_0^2}\right)^{-1}$$

$$-\frac{1}{2\sigma_1^2}\left(\mu^2 - 2\mu\mu_1 + \mu_1^2\right)$$

$$\mu_1 = \frac{1}{2}\left(\frac{1}{\sigma_x^2} + \frac{1}{\sigma_0^2}\right)^{-1}\left(\frac{x}{\sigma_x^2} + \frac{\mu_0}{\sigma_0^2}\right)$$

Cluster assignment part 1 walkthrough

https://canvas.wisc.edu/courses/446937/assignments/2579703

- Admin Matters
- Bayes practice and thinking generatively
- Levels of analysis
- **Short rest 1**
- Math to cover:
  - Expected value
  - Continuous probability
  - Gaussian
  - Mixture modeling
  - Beta-Binomial (discrete enters the continuum)
- **Short rest 2**
- Student intros + one paper presentation
  - Go over paper presentation advice here too.

Discrete Distribution

Prof. Joe is walking through an animal shelter.

There are three animals there: ferrets, cats, and dogs.

How do we model the probability that he encounters a ferret, cat or dog?

We can model this with $\vec{\theta} = (\theta_1, \theta_2, \theta_3)$, one for each animal for $(\sum_k \theta_k = 1)$.

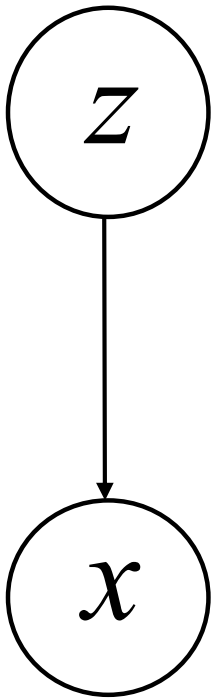
We can define a random variable c to be the next animal encounter. It is **Discrete** distributed:

$$c|\vec{\theta} \sim \text{Discrete}(\vec{\theta})$$

It is sometimes called a *categorical* distribution.

$$P(c = k) = \theta_k$$

Mixture models



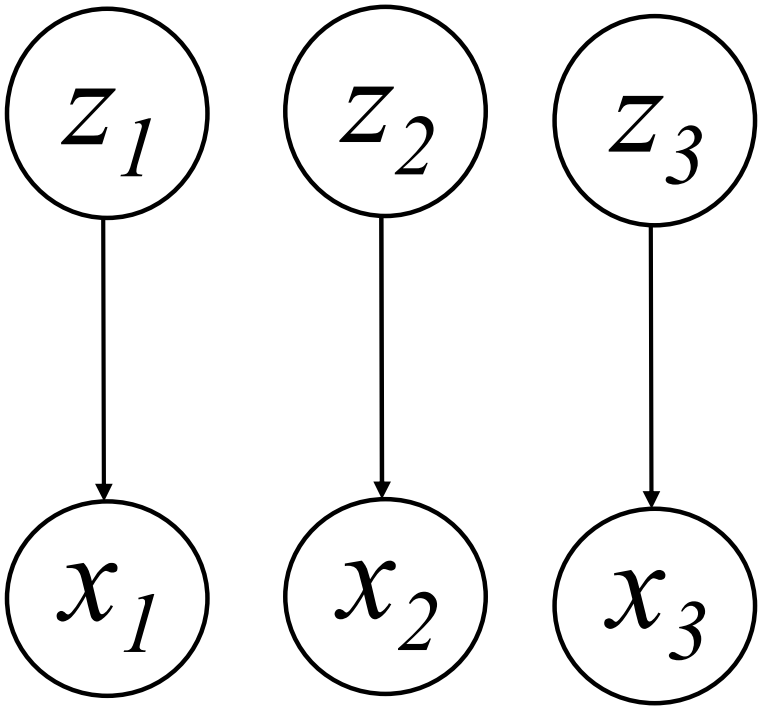
$P(z)$ Latent (unobserved)

$$P(x) = \sum_{z} P(x \mid z) P(z)$$

$P(x|z)$ Observed

Marginalize out (i.e. sum over) latent structure

Plate notation



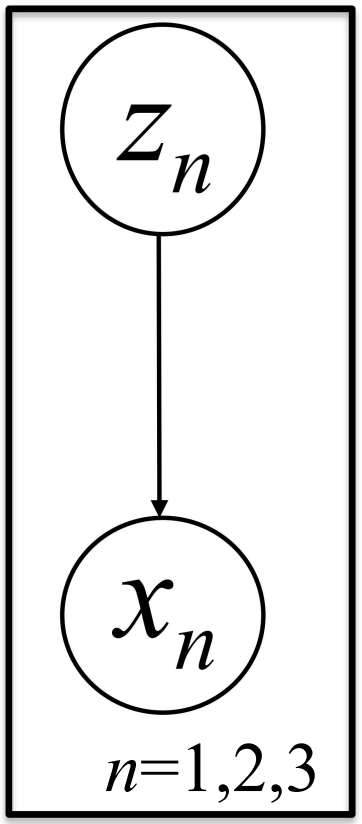
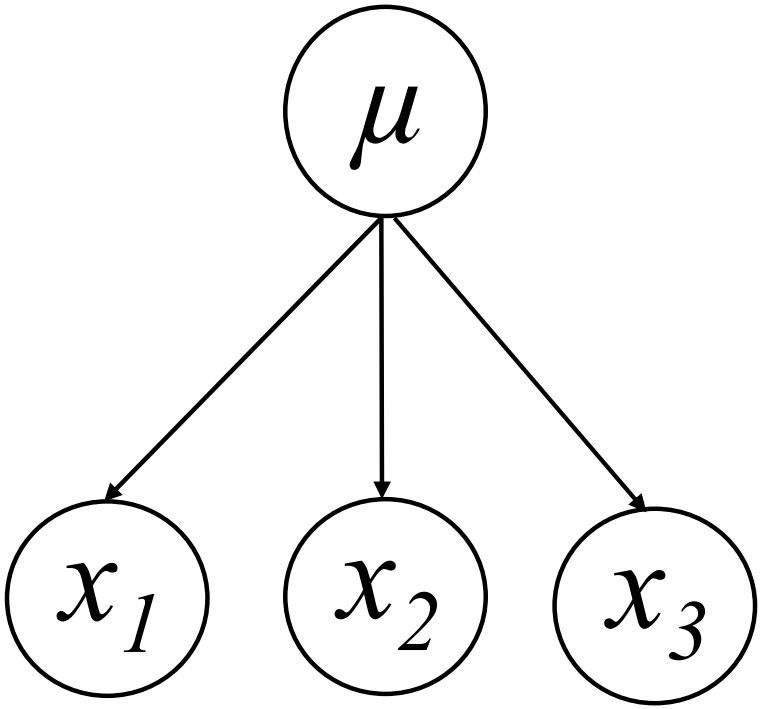
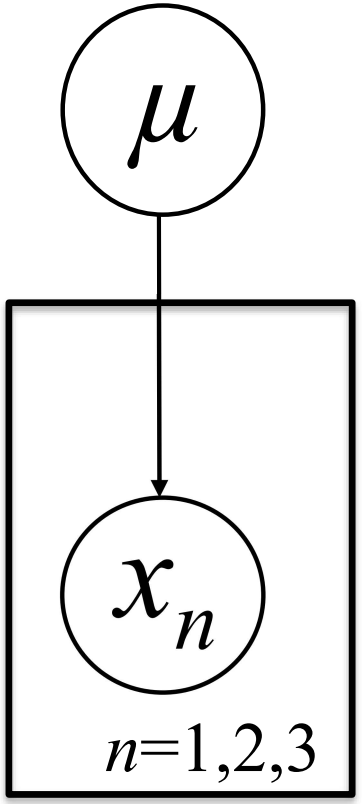


Plate notation





Categorization as Mixture Modeling

To make an animal, first sample which animal we want from $\theta = (.2, \quad .4, \quad .4)$

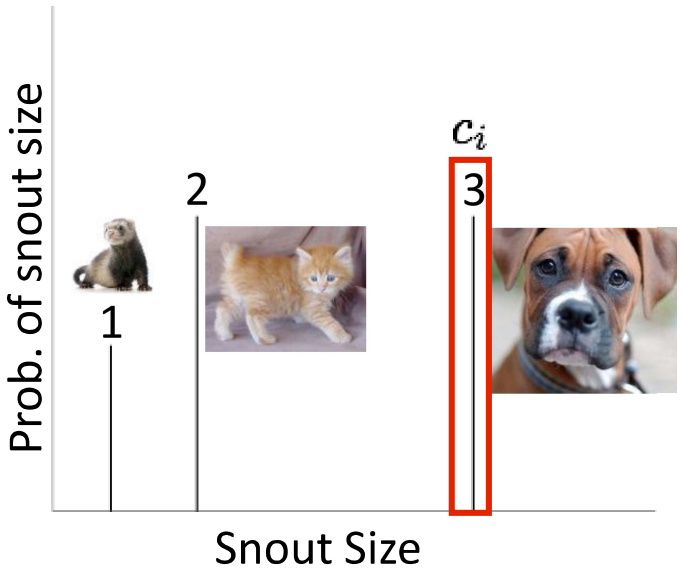
Each animal is created from its associated ideal "snout size" μ_c and variance σ_c^2.

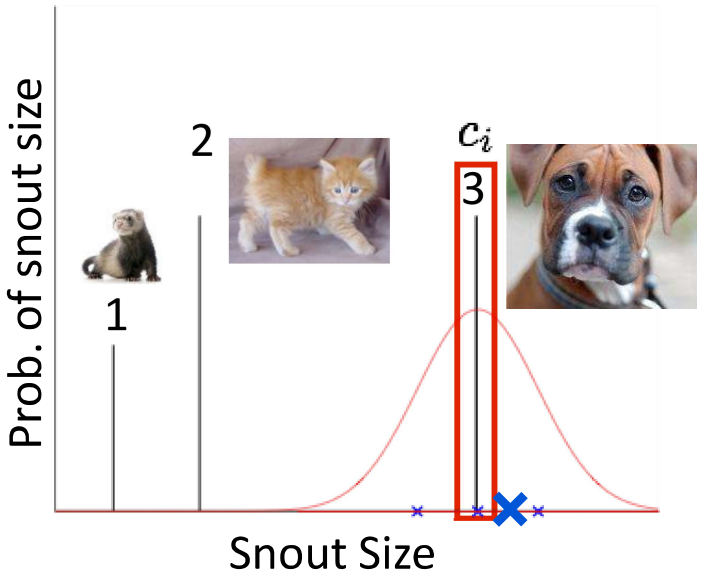
This leads to the following *generative process* for each animal

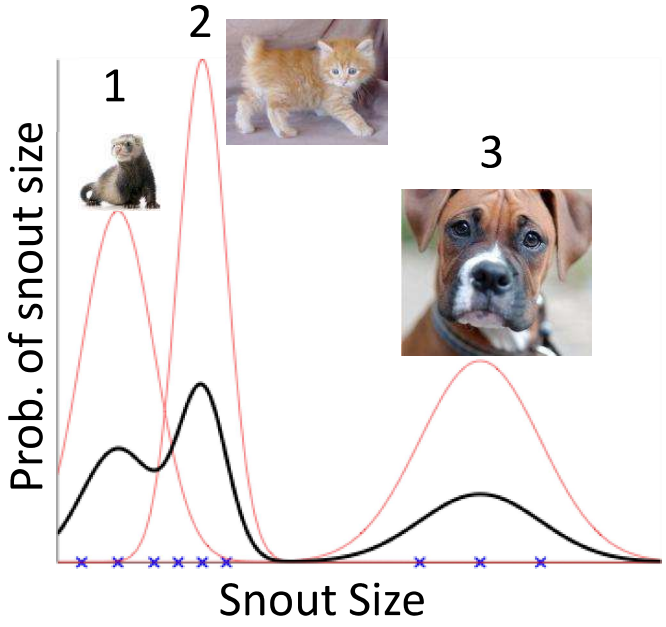
$$c_i|\theta \sim \text{Discrete}(\theta) \qquad x_i|c_i \sim \text{Gaussian}(\mu_{c_i}, \sigma^2_{c_i})$$

Prob. of new snout size: $P(x_i) = \sum_{k=1}^{3} P(x_i, c_i = k) = \sum_{k=1}^{3} P(x_i|\mu_k, \sigma_k^2) P(c_i = k|\theta)$

Categorize a new animal given its snout size: $P(c_i = 1|x_i, \theta) = \frac{p(x_i|\mu_1, \sigma_1^2)P(c = 1|\theta)}{\sum_{k=1}^{3} p(x_i|\mu_k, \sigma_k^2)P(c = k|\theta)}$







Cluster assignment part 2 walkthrough

- https://canvas.wisc.edu/courses/446937/assignments/2579703

- Admin Matters
- Bayes practice and thinking generatively
- Levels of analysis
- **Short rest 1**
- Math to cover:
  - Expected value
  - Continuous probability
  - Gaussian
  - Mixture modeling
  - Beta-Binomial (discrete enters the continuum)
- **Short rest 2**
- Student intros + one paper presentation
  - Go over paper presentation advice here too.

Binomial Distribution

$$X_1, \ldots, X_N | \theta \overset{iid}{\sim} \text{Bernoulli}(\theta) \quad Y = \sum_{n=1}^{N} X_n \quad \Rightarrow Y \sim \text{Binomial}(N, \theta)$$

# of heads from N coinflips.

Create N - k tails

$$P(Y = k|\theta) = \binom{N}{k} \theta^k (1-\theta)^{N-k}$$

Create k heads

Count the number of ways you can get k heads with N coinflips. "N choose k"

Number of ways to order N unique items

$$\binom{N}{k} = \frac{N!}{k!(N-k)!}$$

Number of ways to order N - k tails

Number of ways to order k heads

$N!$ is the "factorial" operator. It counts the number of ways to order N items.

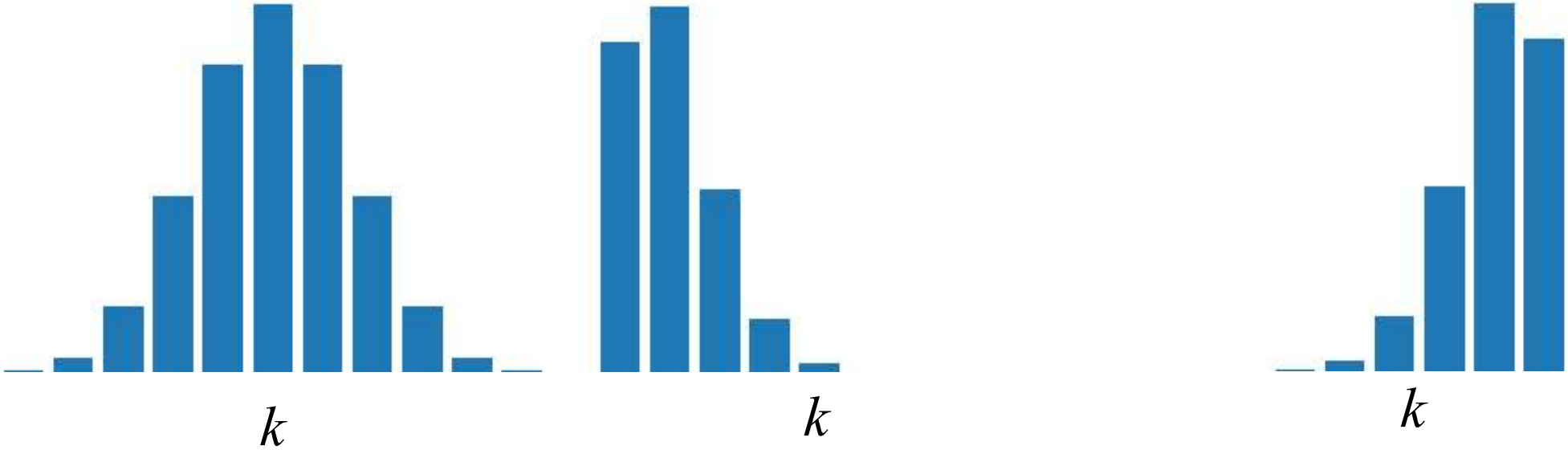
$$N! = 1 \times 2 \times \ldots \times N = \prod_{n=1}^{N} n$$

Binomial Distribution

$$X_1, \ldots, X_N | \theta \overset{iid}{\sim} \text{Bernoulli}(\theta) \quad Y = \sum_{n=1}^{N} X_n \quad \Rightarrow Y \sim \text{Binomial}(N, \theta)$$

$$P(Y = k|\theta) = \binom{N}{k} \theta^k (1-\theta)^{N-k}$$

$P(Y = k|\theta = 0.5)$ $P(Y = k|\theta = 0.1)$ $P(Y = k|\theta = 0.9)$



What if we don't know the coinflip bias? (the probability of heads θ)?

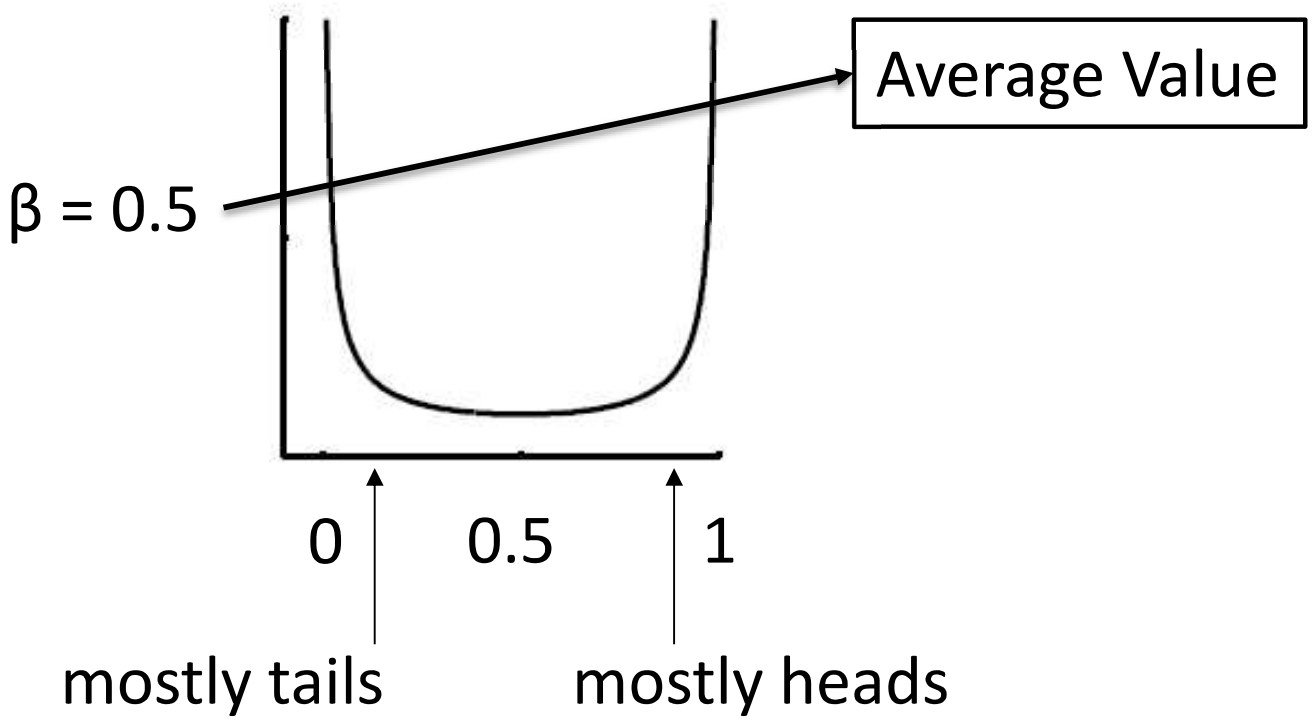
The Beta Distribution: A distribution over the probability of a heads

Beta(a, b)

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

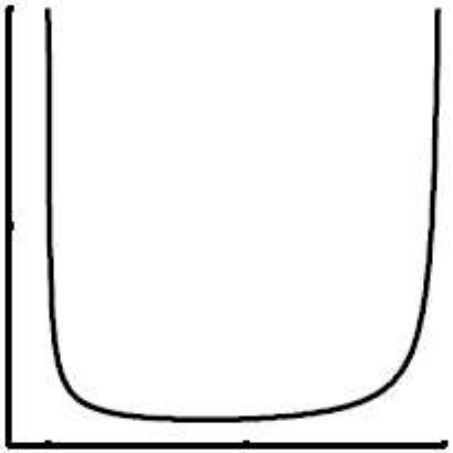
These are Beta(αβ, α(1-β)).

$$p\left(\theta|\alpha\beta,\alpha(1-\beta)\right) = \frac{\Gamma(\alpha)}{\Gamma(\alpha\beta)\Gamma\left(\alpha(1-\beta)\right)}\theta^{\alpha\beta-1}(1-\theta)^{\alpha(1-\beta)-1}$$

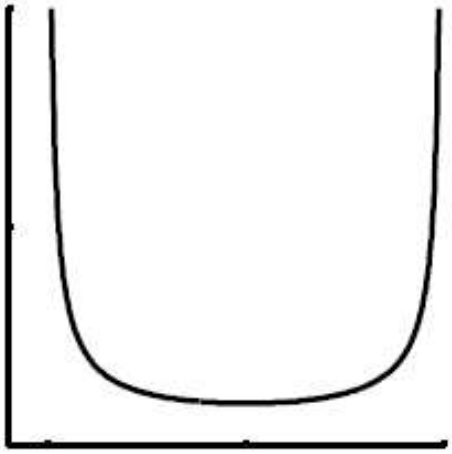


α = 0.5 → Shape

β = 0.8



β = 0.5



0 0.5 1

mostly tails mostly heads

α = 0.5 α = 2 α = 10

Gamma function interlude

In all likelihood, never will need to deal with directly.

It is the generalized factorial function. $\Gamma(x) = \int_0^\infty z^{x-1} e^{-z} dz$

For positive integer x, $\Gamma(x) = (x-1)!$

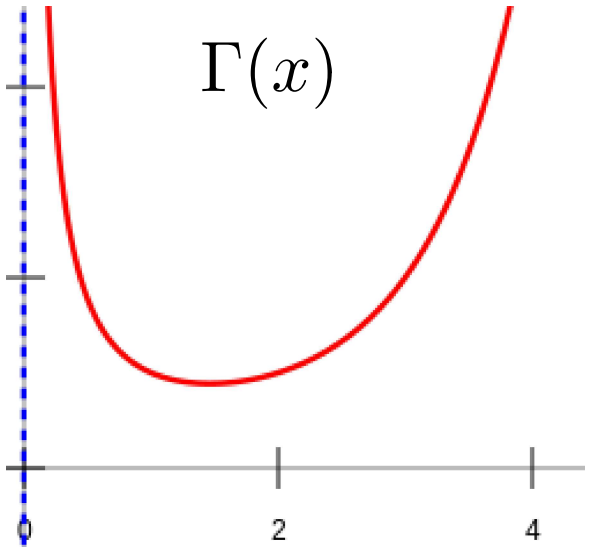
As far as I'm concerned, it is restricted to positive real numbers.

It is the normalization constant for the Gamma distribution.

Some useful facts:

$$\Gamma(x+1) = x\Gamma(x)$$

$$\Gamma(1) = 0! = 1 = \Gamma(2) = 1! = 1$$



Beta-Binomial Posterior Update

$$\theta \sim \text{Beta}(a, b) \qquad Y|\theta \sim \text{Binomial}(N, \theta)$$

$$p(\theta|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \qquad P(Y=k|\theta) = \binom{N}{k}\theta^k(1-\theta)^{N-k}$$

$$p(\theta|Y=N_H) \propto P(Y=N_H|\theta)p(\theta|a,b)$$

$$p(\theta|Y=N_H) \propto \frac{N!}{N_H!(N-N_H)!}\theta^{N_H}(1-\theta)^{N-N_H}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$$

$$p(\theta|Y=N_H) \propto \theta^{N_H}(1-\theta)^{N-N_H}\theta^{a-1}(1-\theta)^{b-1} = \theta^{N_H+a-1}(1-\theta)^{N-N_H+b-1}$$

$$\Rightarrow \theta|Y=N_H \sim \text{Beta}(N_H+a, N-N_H+b)$$

Posterior is of the same functional form as prior (Both are prior).
Parameters are updated to reflect knowledge.
New a adds in the number of heads. New b adds in number of tails.

This is a **Conjugate Prior**. (more on this next week)