Agenda

Probability distributions & random samples

1. Administrative
2. Probability distributions
3. From distributions to models
4. Summarizing random variables
5. *Hands on: random samples in R*

Worksheet 1 posted

- Available in MyCourses
  Content > Worksheets > Worksheet 1
- Download the linked `.Rmd` file to your computer and open in your preferred editor (e.g. RStudio)
- When you're done, upload to the same spot in MyCourses

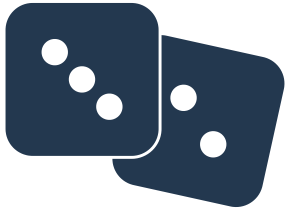
Lab

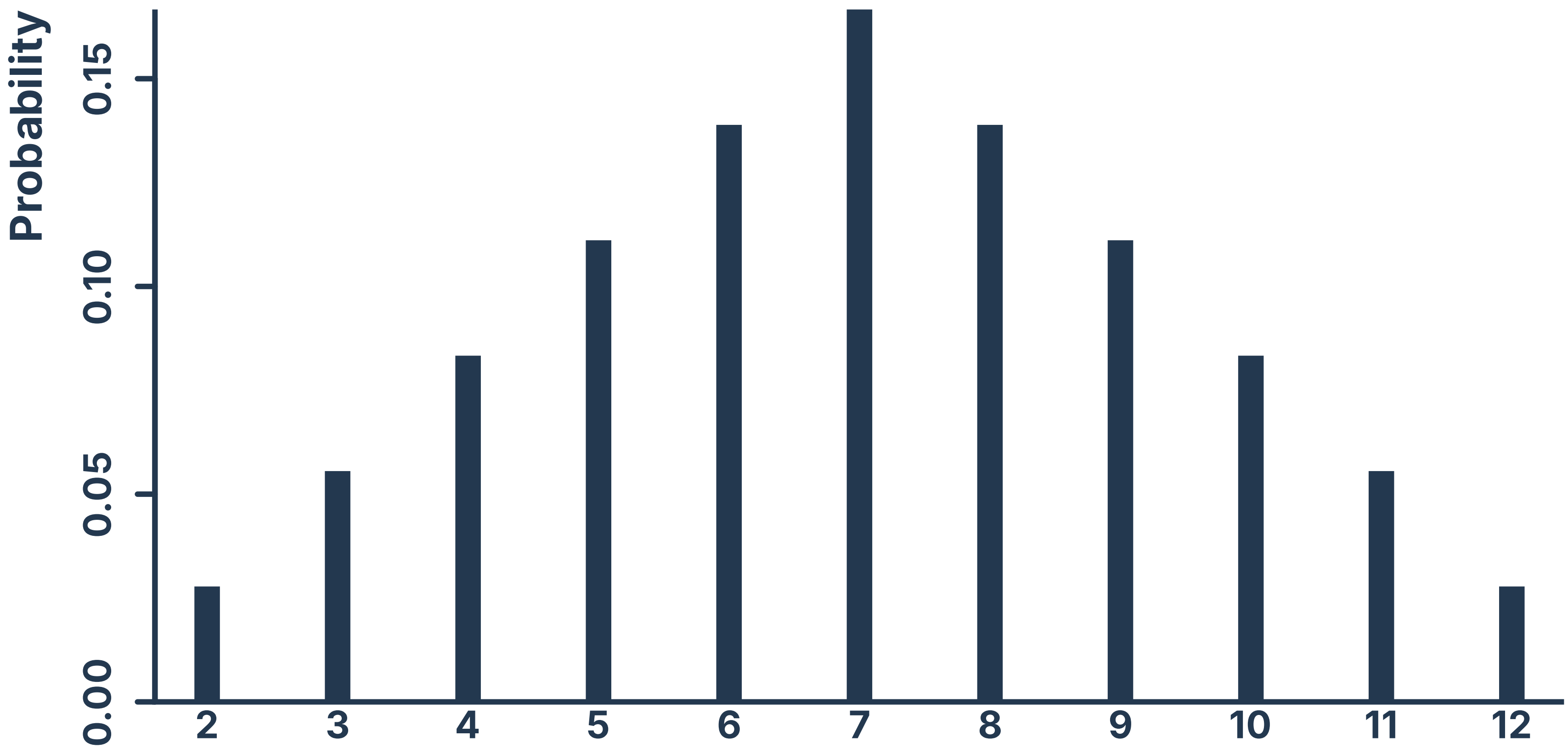
- First lab today, directly after class in Leacock 808 (one floor down)

Probability distributions

Probability mass function

E.g. the sum of two fair dice:
Categorical distribution

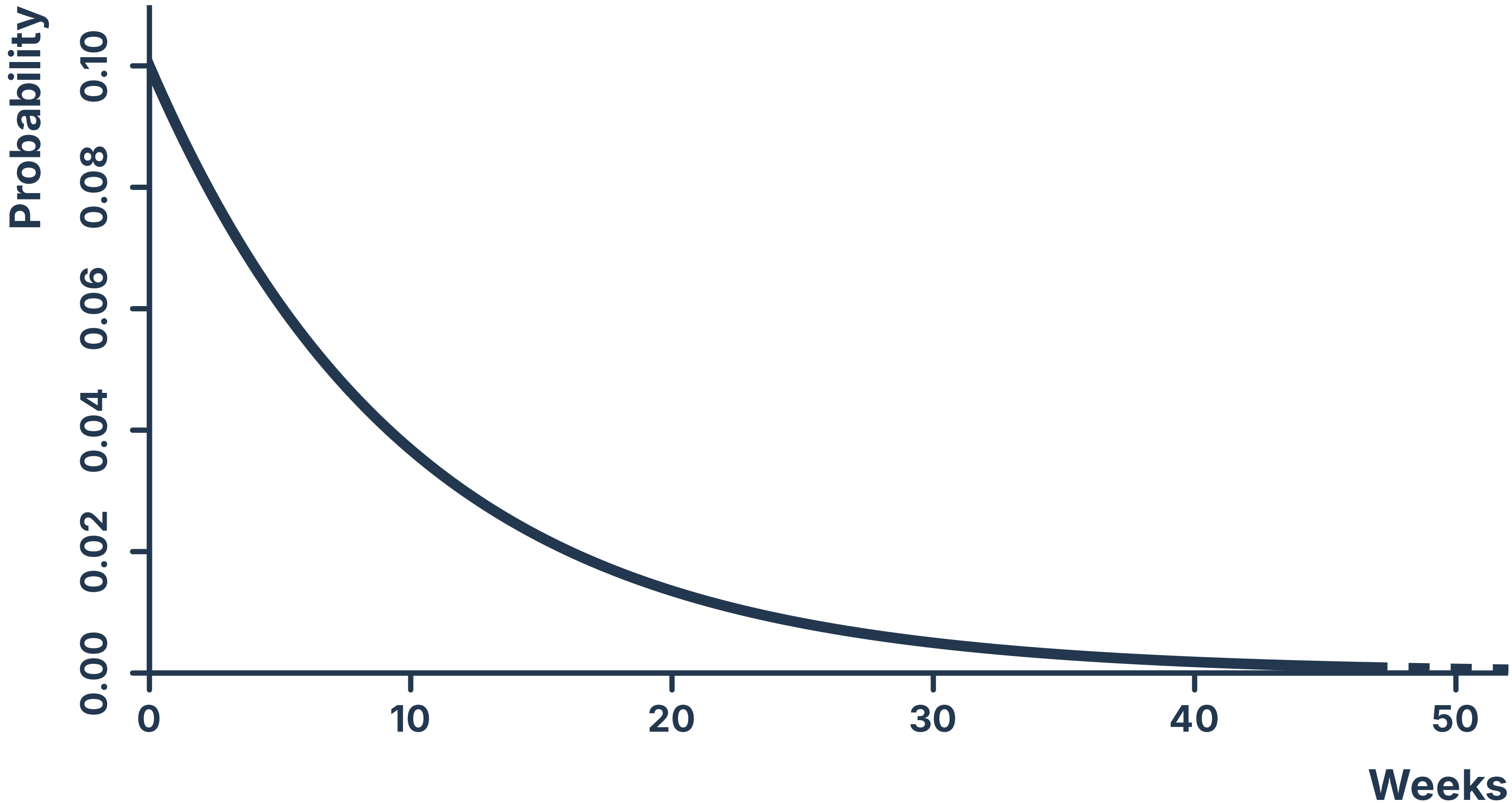




Support: Integers from 2 to 12 (discrete)

Probability density function

E.g. the time between shark attacks in Austrailia:
Exponential distribution ($\lambda = 0.1$)



Support: Non-negative real numbers $[0, \infty)$

Contingency table

E.g. questions measuring authoritarian attitudes:
Bivariate categorical distribution

		X_2: *Women should have to promise to obey their husbands when they get married.*	
		Agree	**Disagree**
X_1: *Gays and lesbians are just as healthy and moral as anybody else.*	**Agree**	0.05	0.53
	Disagree	0.33	0.09

Joint probability distributions measure probability across multiple variables *and* the association between those variables.

$$\Pr(X_1 = A, X_2 = A) = 0.05 \qquad \Pr(X_1 = A, X_2 = D) = 0.53$$
$$\Pr(X_1 = D, X_2 = A) = 0.33 \qquad \Pr(X_1 = D, X_2 = D) = 0.09$$

		X_2	
		Women should have to promise to obey their husbands when they get married.	
		Agree	**Disagree**
X_1 *Gays and lesbians are just as healthy and moral as anybody else.*	**Agree**	0.05	0.53
	Disagree	0.33	0.09

Conditional probability measures probability of one variable in a joint distribution, holding the other constant at a specific value. The probabilities must be *normalized*.

$$\Pr(X_2 = A | X_1 = D) = \frac{0.33}{0.33 + 0.09} = 0.79$$

$$\Pr(X_2 = D | X_1 = D) = \frac{0.09}{0.33 + 0.09} = 0.21$$

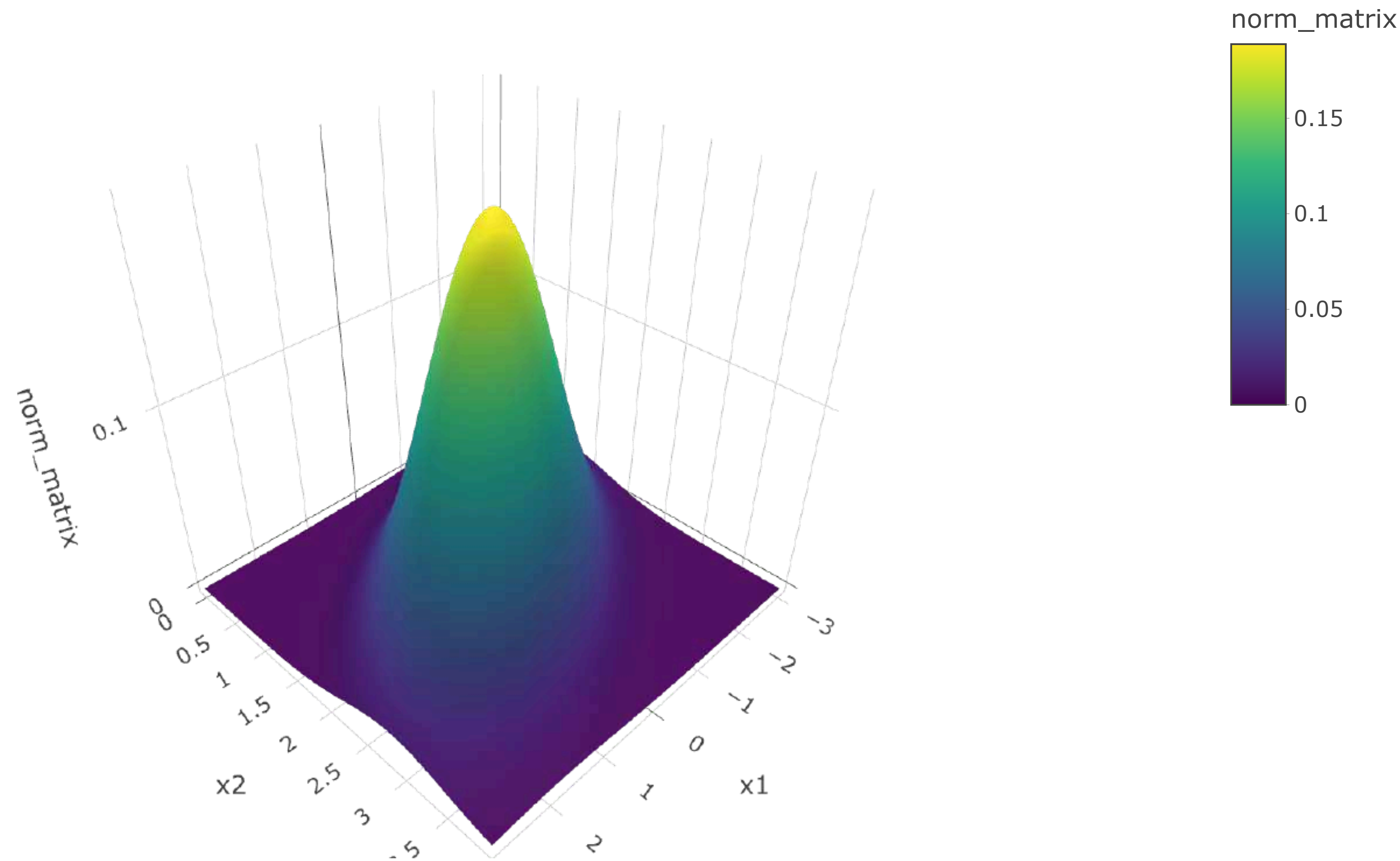
A DISCRETE BIVARIATE DISTRIBUTION

			X_2: *Women should have to promise to obey their husbands when they get married.*		
			Agree	**Disagree**	
X_1	*Gays and lesbians are just as healthy and moral as anybody else.*	**Agree**	0.05	0.53	**0.58**
		Disagree	0.33	0.09	**0.42**
			0.38	**0.62**	

Marginal probability measures probability of one variable in a joint distribution, across all possible values of the other.

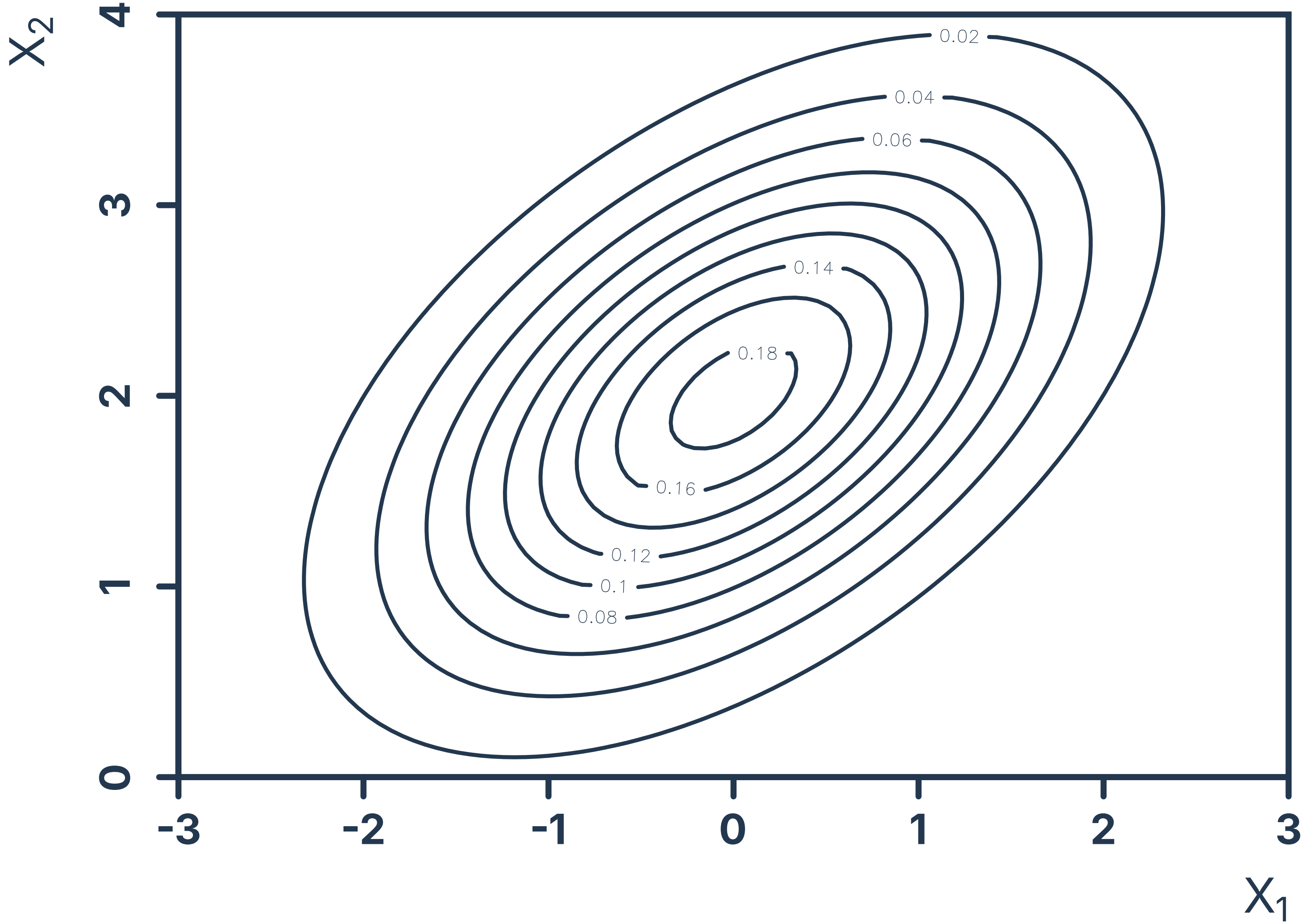
$$\Pr(X_2 = A) = 0.05 + 0.33 = 0.38$$
$$\Pr(X_2 = D) = 0.53 + 0.09 = 0.62$$

A CONTINUOUS BIVARIATE DISTRIBUTION



$$Y \sim \text{Norm}\left(\mu = (0, 2), \Sigma = \begin{bmatrix} 1.2 & 0.5 \\ 0.5 & 0.8 \end{bmatrix}\right)$$

A CONTINUOUS BIVARIATE DISTRIBUTION



$$Y \sim \mathrm{Norm}\left(\mu = (0, 2), \Sigma = \begin{bmatrix} 1.2 & 0.5 \\ 0.5 & 0.8 \end{bmatrix}\right)$$

SOME COMMON DISTRIBUTIONS

	Type	Parameters	Support
Binomial	Discrete	n, p	$\{0, 1, \ldots, n\}$
Poisson	Discrete	λ	$\{0, 1, 2, \ldots\}$
Normal (Gaussian)	Continuous	μ, σ	$(-\infty, \infty)$
Cauchy	Continuous	x_0, γ	$(-\infty, \infty)$
Beta	Continuous	α, β	$[0, 1]$
Exponential	Continuous	λ	$[0, \infty)$

(Statisticians have devised and named innumerable distributions over time.
See https://en.wikipedia.org/wiki/List_of_probability_distributions for an incomplete list)

Probability models

DESCRIBING MODELS

A language for describing probabilistic models

Using probability distributions to link (known) data with (unknown) parameters succinctly and clearly communicates a model.

Example from last week:
Estimating the unemployment rate p from count of unemployed (Y) in our sample of n individuals

Data

$$Y \sim \mathrm{Binom}(n, p)$$

$$p \sim \mathrm{Unif}(0, 1)$$

Parameter Hyperparameters

A language for describing probabilistic models

Using probability distributions to link (known) data with (unknown) parameters succinctly and clearly communicates a model.

Changes to the model are clear:

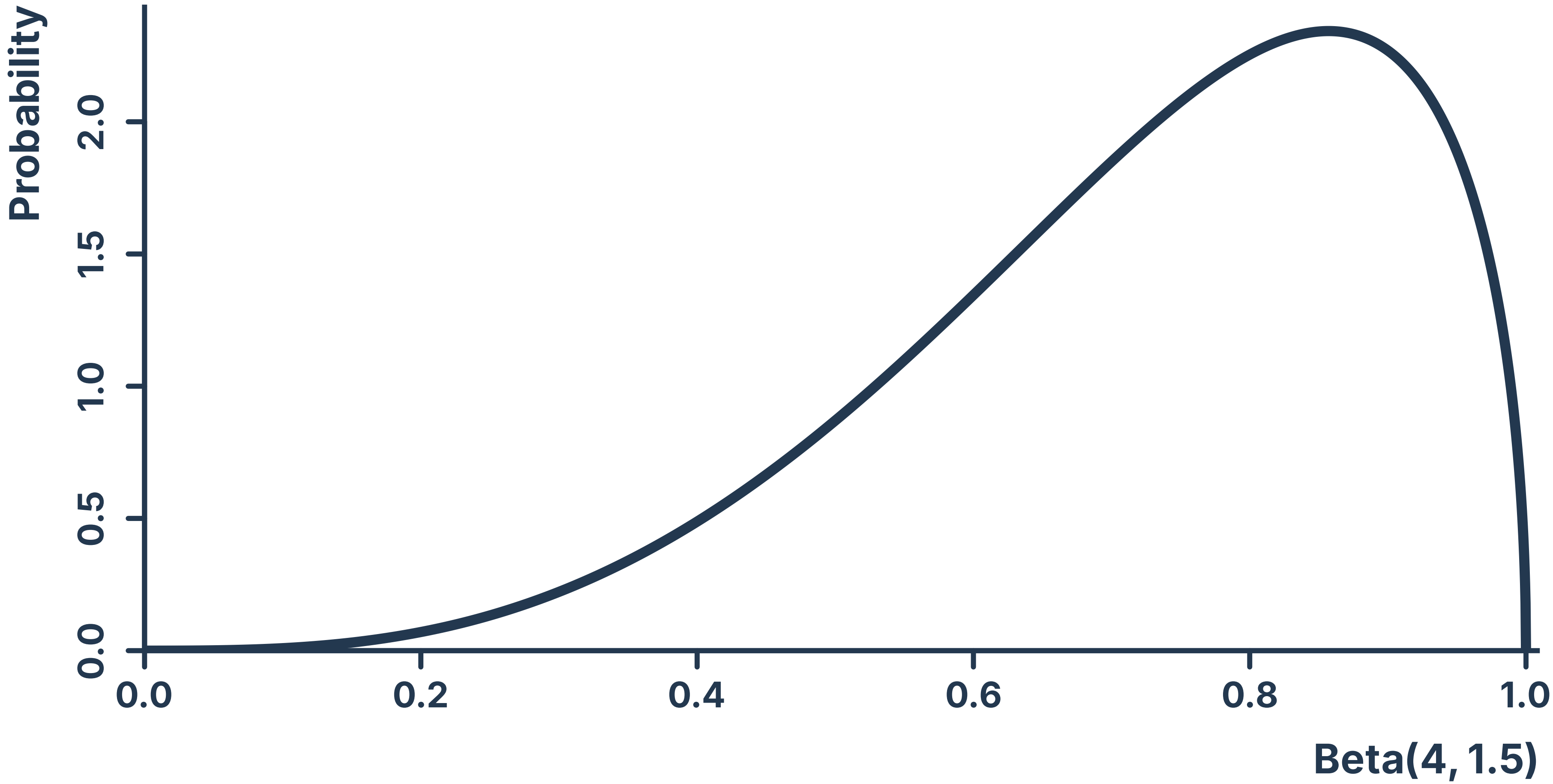
$$\begin{aligned} Y &\sim \text{Binom}(n, p) \\ p &\sim \text{Unif}(0, 1) \end{aligned}$$

$$\begin{aligned} Y &\sim \text{Binom}(n, p) \\ p &\sim \text{Beta}(1.01, 1.01) \end{aligned}$$

Summarizing random variables

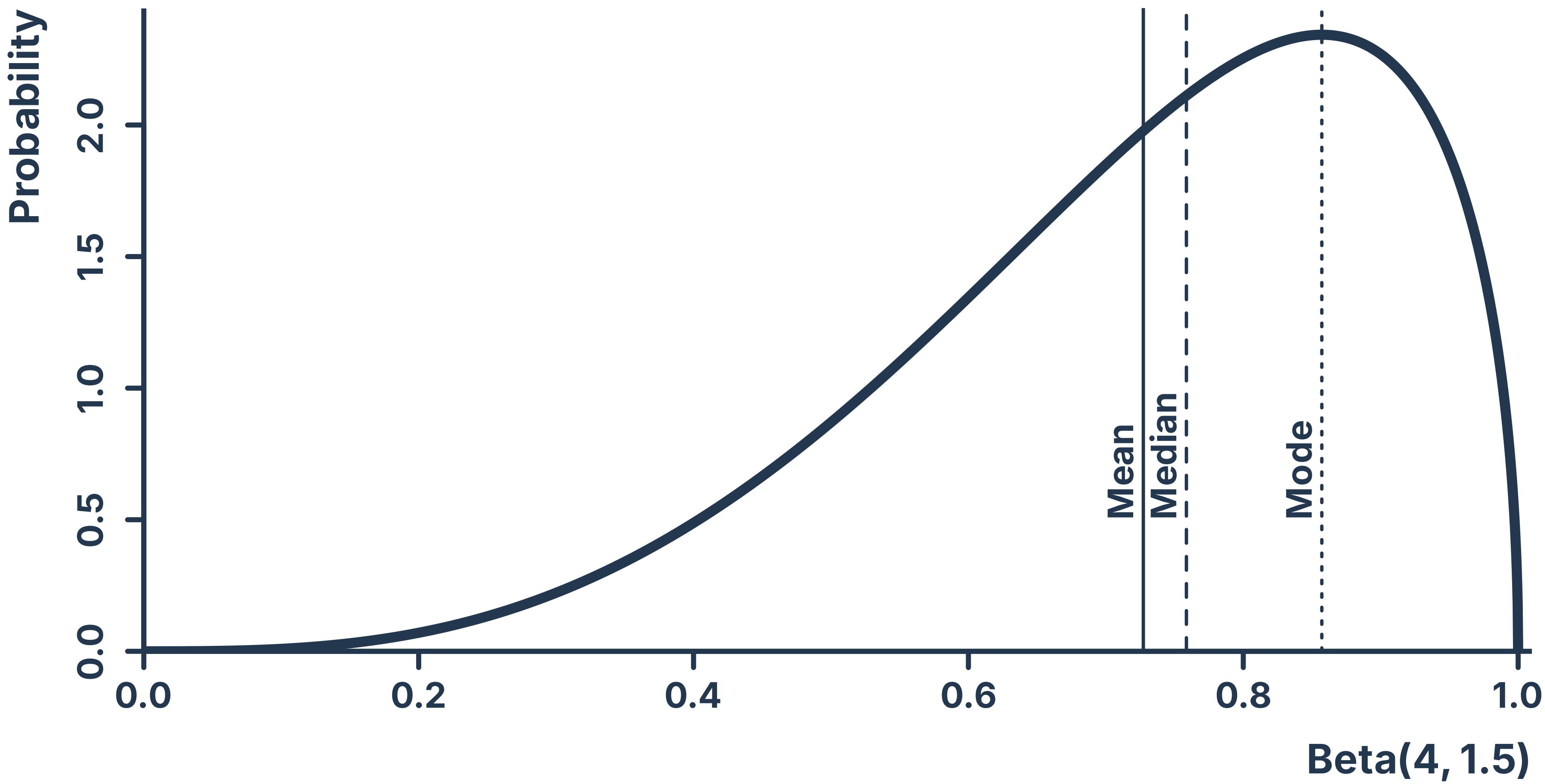
Describing the shape of a distribution

Posterior distributions contain a lot of information



Point summaries

Describe the “center” of the distribution
Mean, median, and mode all have different meanings



SUMMARIZING DISTRIBUTIONS

Mean

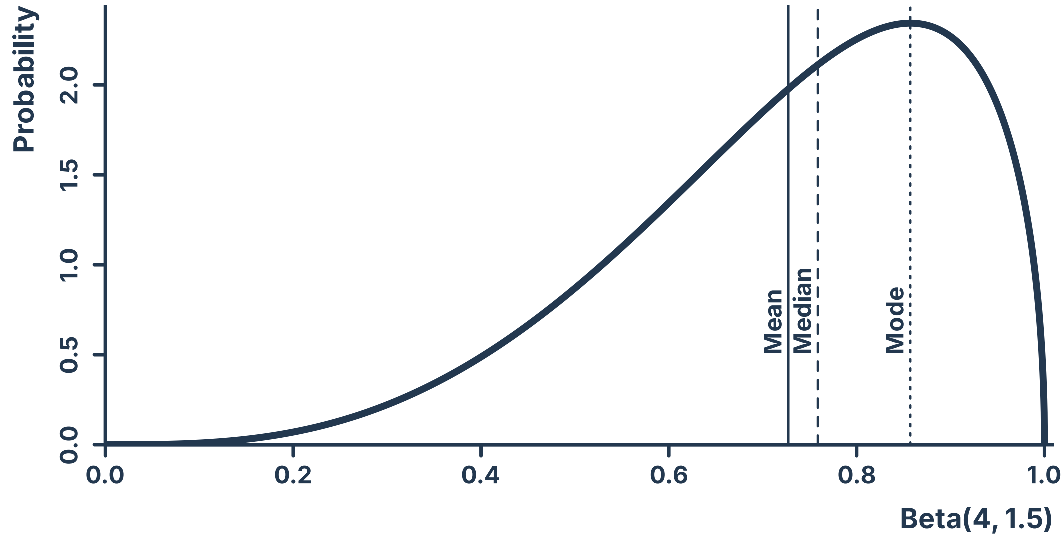
- Colloquially "average"
- Accounts for the magnitude of all data
- Sensitive to extreme values

Median

- 50th percentile
- Not sensitive to extreme values

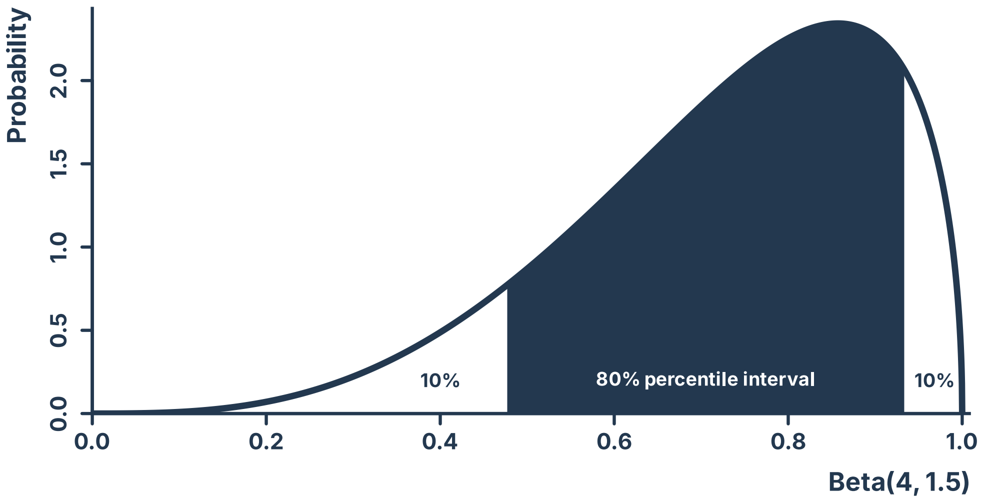
Mode

- Value of X with highest probability density
- Maximum likelihood methods find the mode

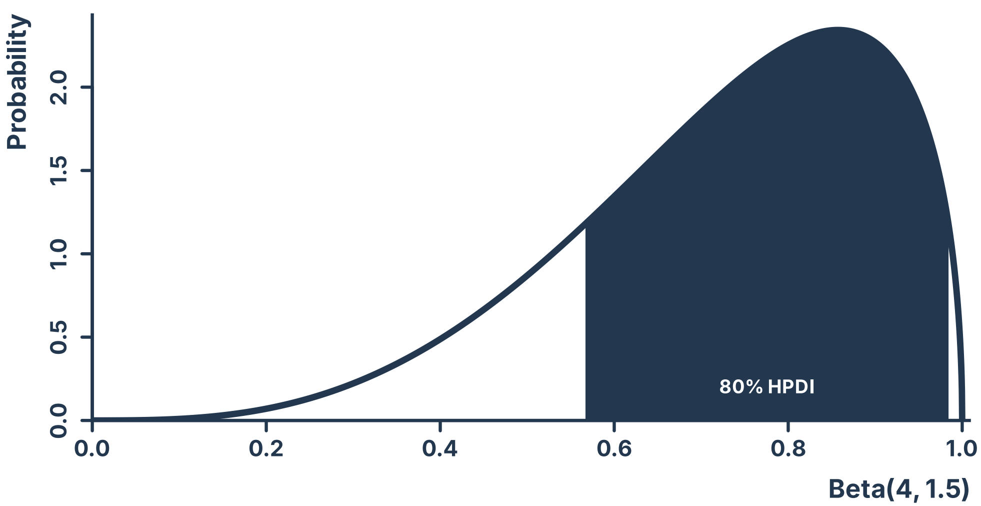


Credible intervals

Describe the "spread" of the distribution

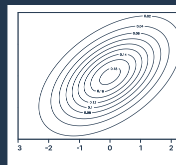


Percentile (aka quantile) intervals leave the same amount of density on either end of the distribution.



Highest posterior density intervals find the narrowest possible interval containin the target density.

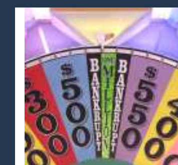
Image credit



Figures by Peter McMahan (source code)

Poster detail for Jaws (1975)

Photo by Leo Reynolds on Flickr



Still from Wheel of Fortune