SOCI 620: Quantitative methods 2

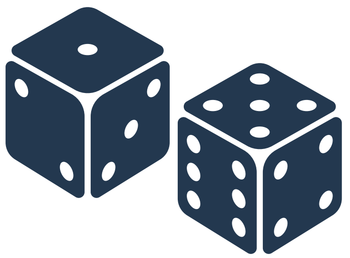
Jan. 14

Probability distributions and random samples

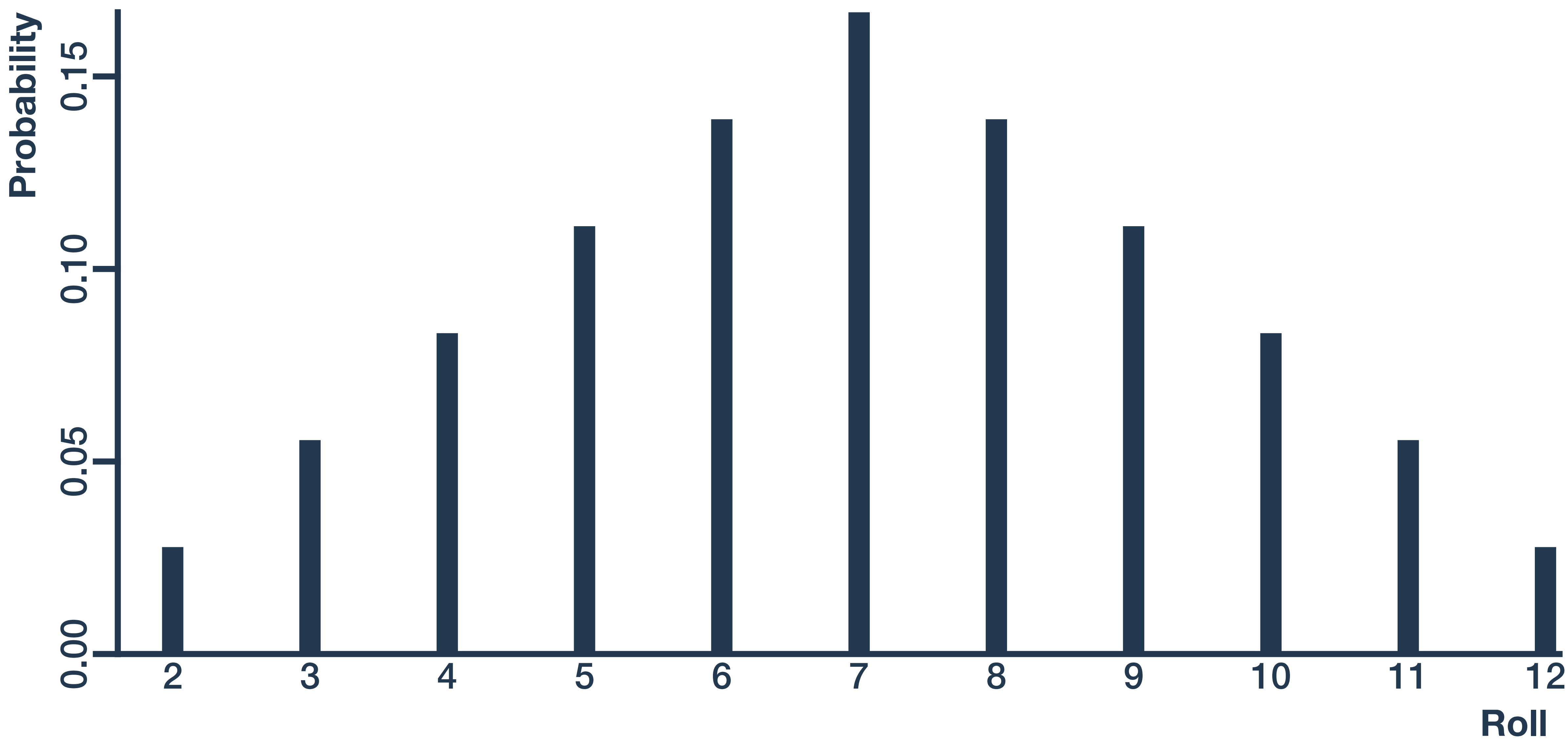
1. Probability distributions
2. Probability models
3. Summarizing random variables
4. Sampling from distributions

Probability distributions

A discrete distribution



Probability mass function (PMF)
Sum of two fair dice

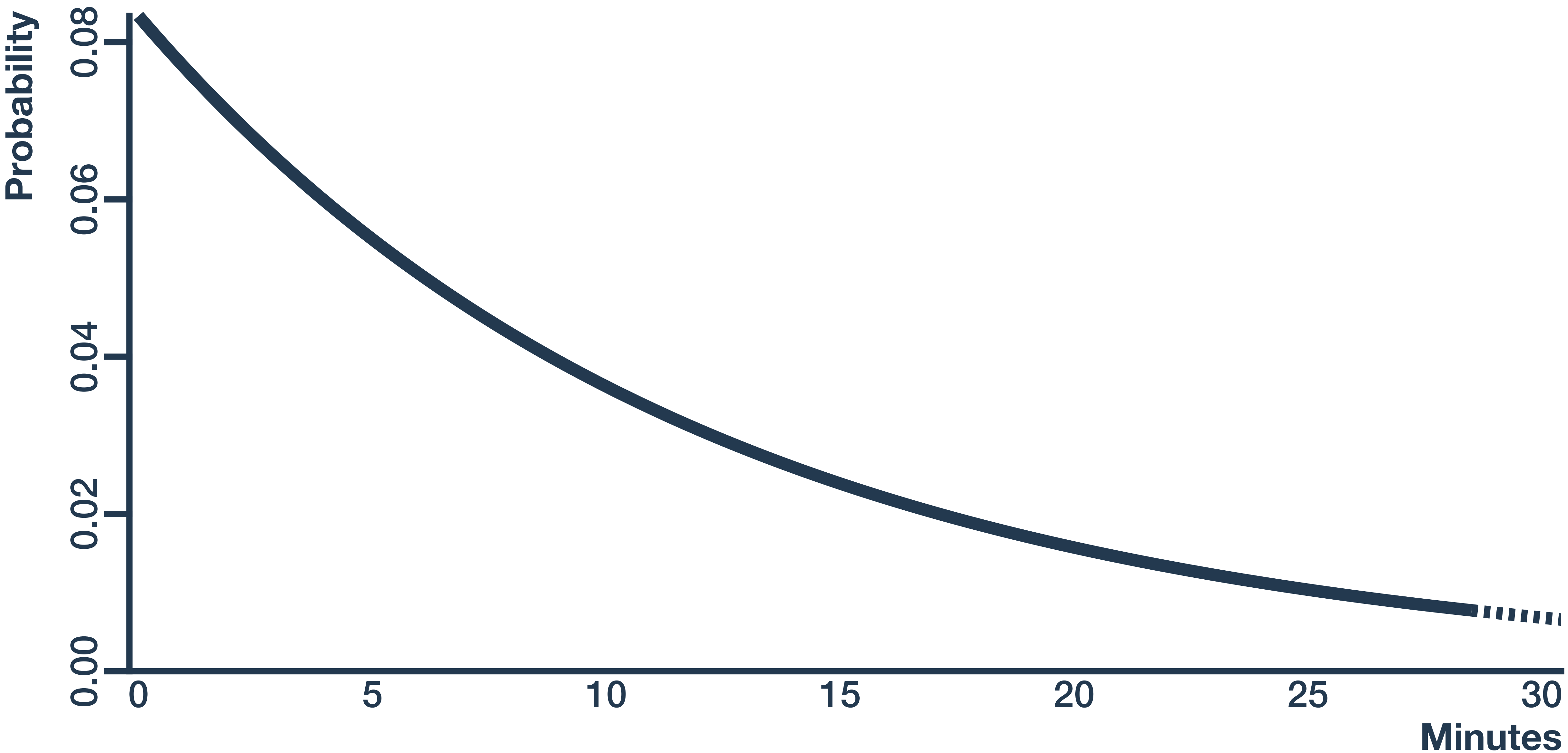


Support: integers from 2 to 12 (discrete)

A continuous distribution

Probability density function (PDF)

Time between Metro arrivals, (λ=1/12)



Support: non-negative, real [0,∞)

A discrete bivariate distribution

Contingency table

Questions measuring authoritarian attitudes

		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.

$\Pr(X_1{=}\text{A}, X_2{=}\text{A}) = 0.05$ $\Pr(X_1{=}\text{A}, X_2{=}\text{D}) = 0.53$

$\Pr(X_1{=}\text{D}, X_2{=}\text{A}) = 0.33$ $\Pr(X_1{=}\text{D}, X_2{=}\text{D}) = 0.09$

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
	Disagree	0.33	0.09

Conditional probability: measures probability of one variable in a joint distribution, holding the other constant

$\Pr(X_2 \mid X_1{=}\mathrm{D})$

	Agree	Disagree
X_1 = Disagree	$\frac{0.33}{0.33+0.09} = 0.79$	$\frac{0.09}{0.33+0.09} = 0.21$

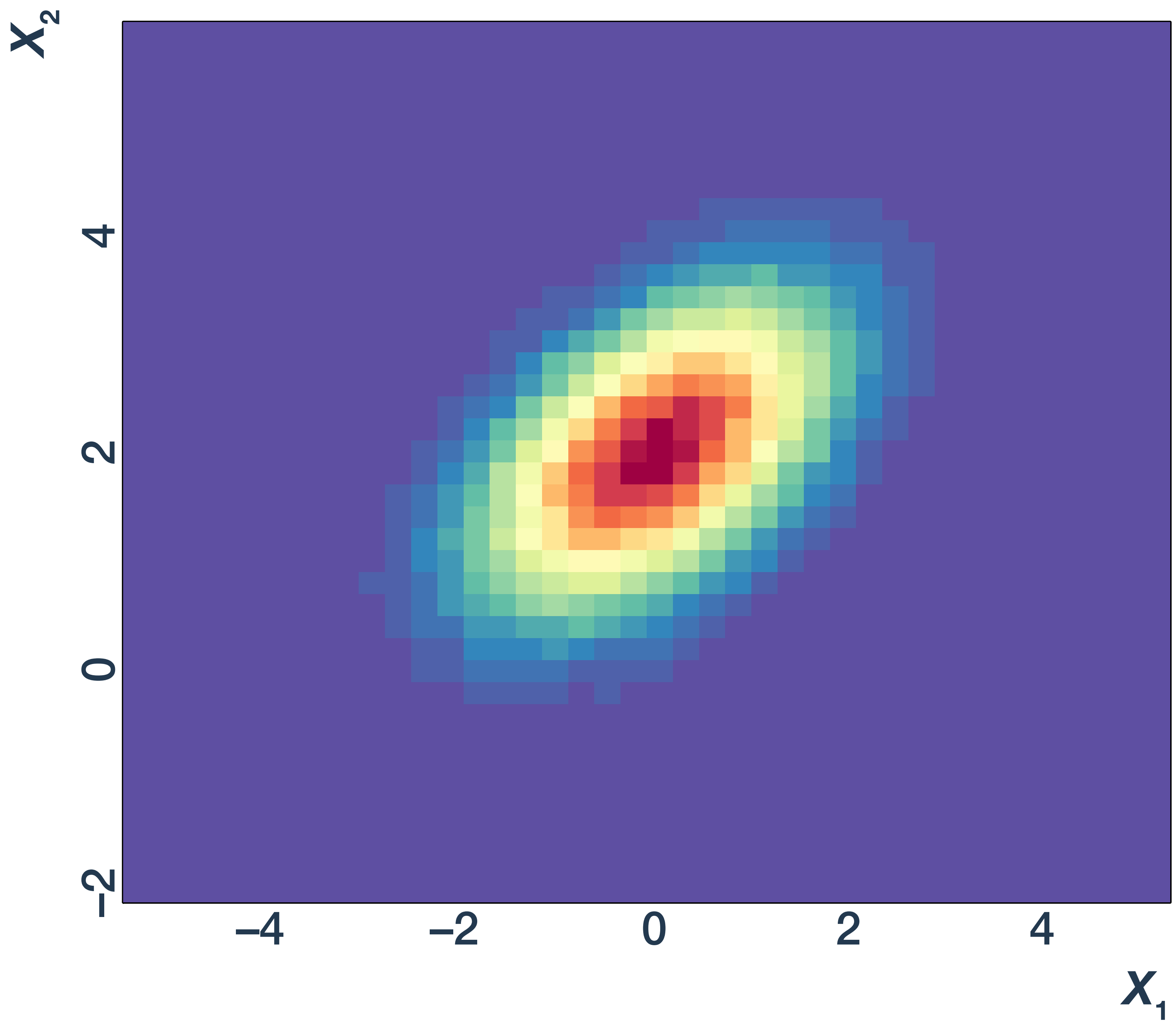
A discrete bivariate distribution

		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

	Agree	Disagree
$\Pr(X_2)$	0.5 + 0.33 = 0.38	0.53 + 0.09 = 0.62

A continuous bivariate distribution



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

Some common distributions

	Type	Parameters	Support
Binomial	*Discrete*	n, p	$\{0,\ldots,n\}$
Poisson	*Discrete*	λ	$\{0,1,2,\ldots\}$
Geometric	*Discrete*	p	$\{0,1,2,\ldots\}$
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 *many* 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 our (known) data with our (unknown) parameters allows succinct communication

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

Data

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

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

Parameter

Hyperparameters

Describing models

A language for describing probabilistic models
Using probability distributions to link our (known) data with our (unknown) parameters allows succinct communication

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

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

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

Changes to model are clear:

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

$$p \sim \text{Beta}(1.01, 1.01)$$

A note on likelihood

Posterior probability | Likelihood | Prior probability

$$\Pr(p \mid n, Y) = \frac{\Pr(Y \mid n, p)\,\Pr(p)}{\Pr(Y)}$$

Posterior and _prior_ are distributions over _p_:

Posterior tells us “probability of any p, given the data”
Prior tells us “probability of any p, *a priori*”

When we plot posterior and prior for values of p, we see a valid probability distribution

Likelihood is a distribution over the _data_:

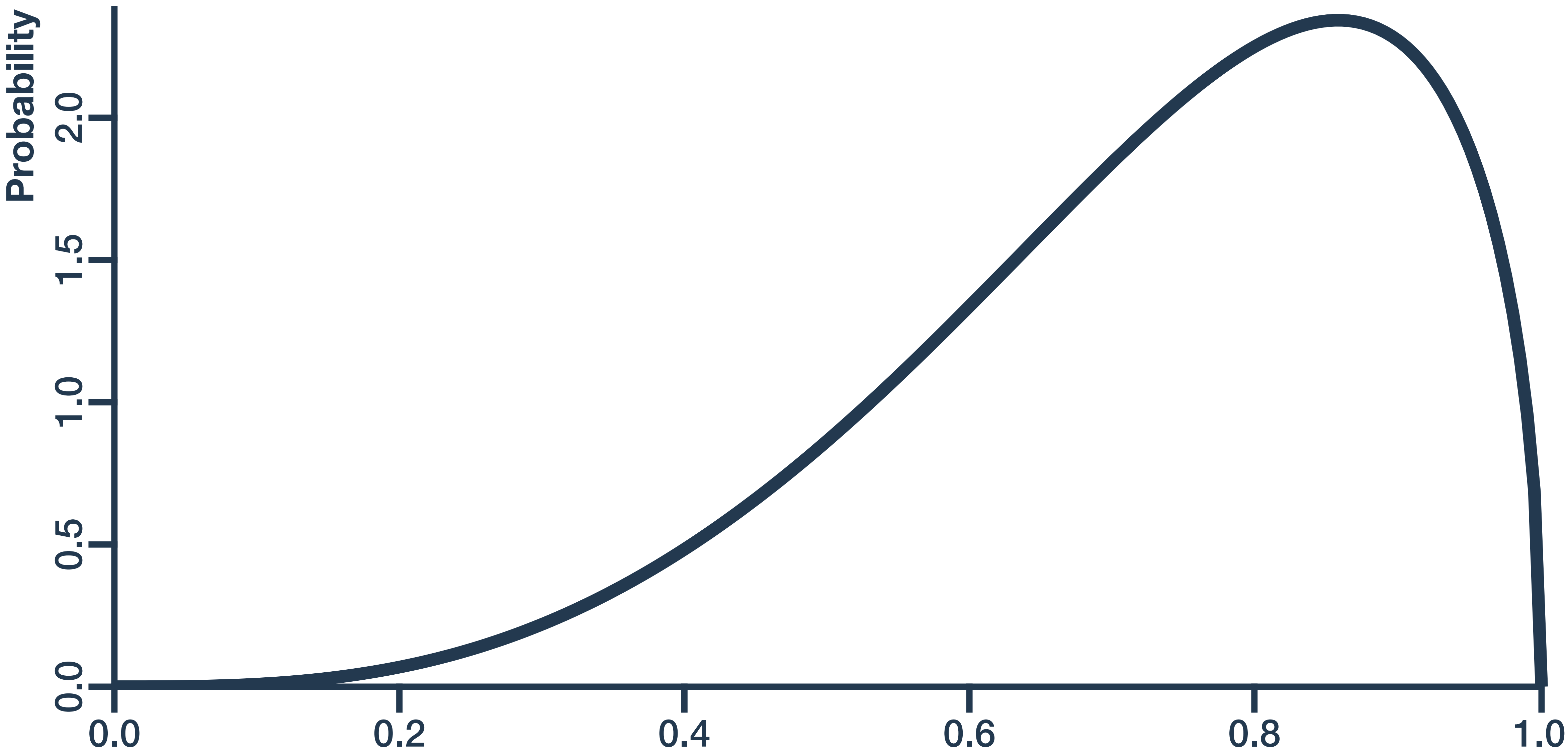
Likelihood tells us “probability of the data, given any p”

When we calculate that probability across values of p, it is not a proper probability distribution
Measure of how surprised we would be by our data for any possible value of p

Summarizing random variables

Summarizing distributions

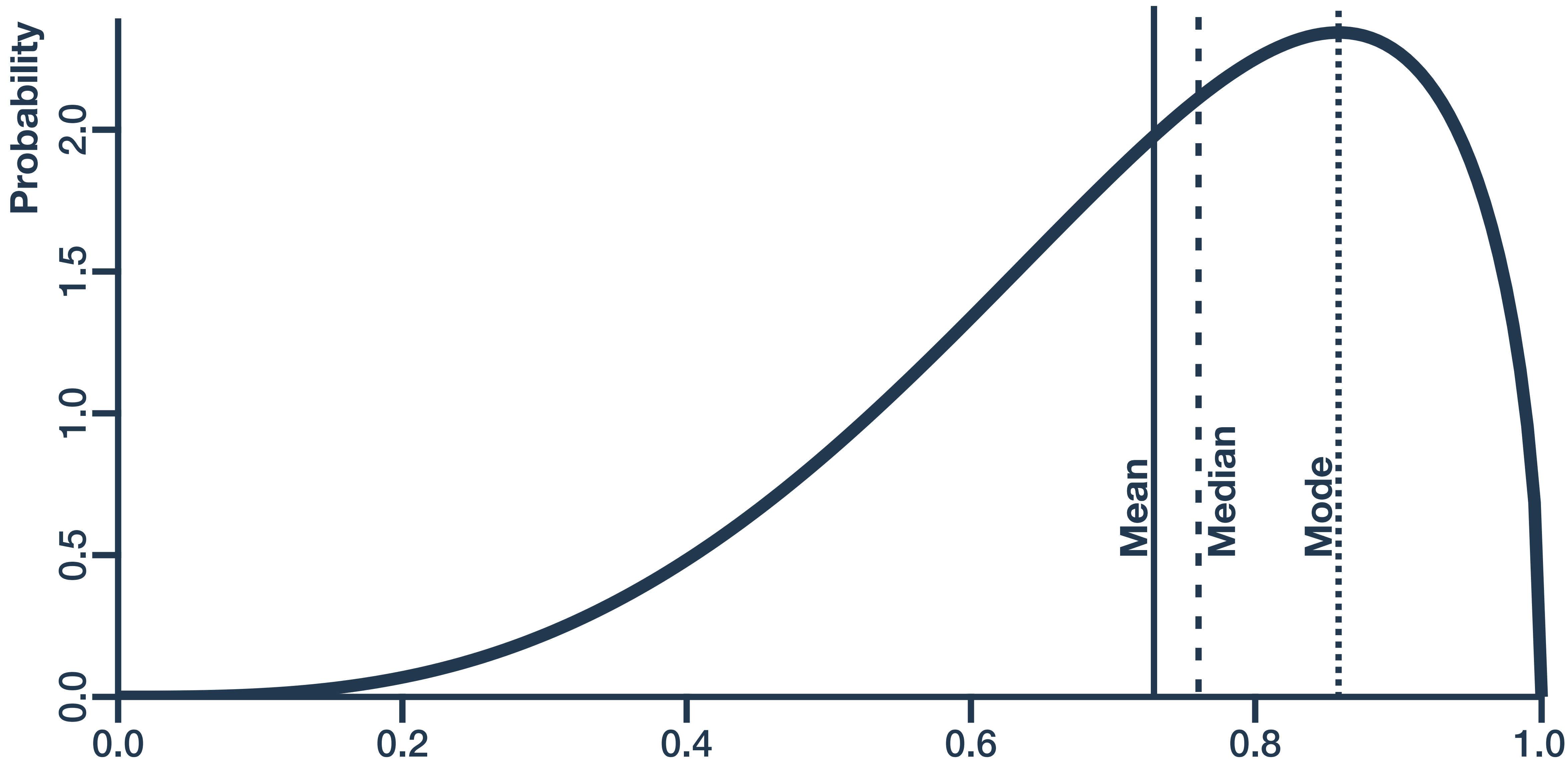
Communicating the shape of a distribution
Probability distributions like those that result from Bayesian analysis are complex



Beta(4, 1.5)

Summarizing distributions

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

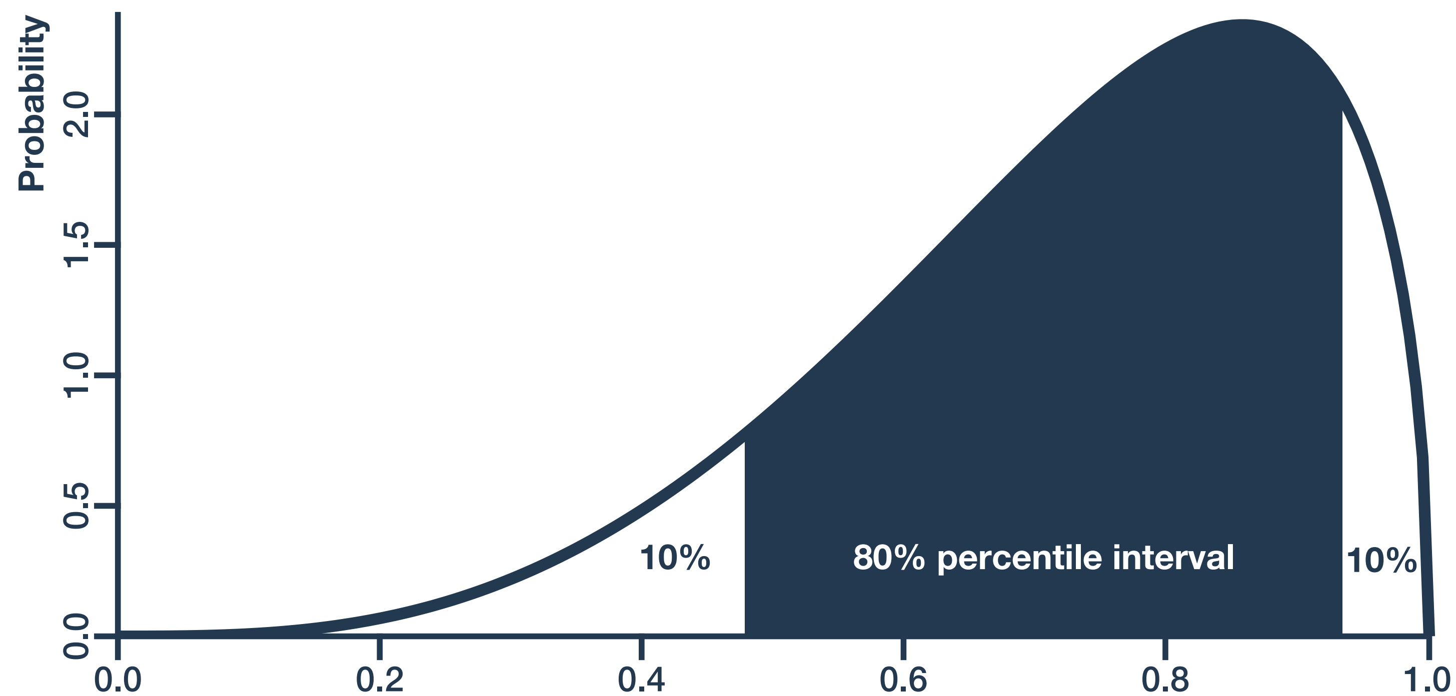


Beta(4, 1.5)

Summarizing distributions

Credible intervals
describe the "spread"
of a 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.