Computer Science

CSC696H: Probabilistic Methods in ML

Probability and Statistics : Review

Prof. Jason Pacheco

Outline

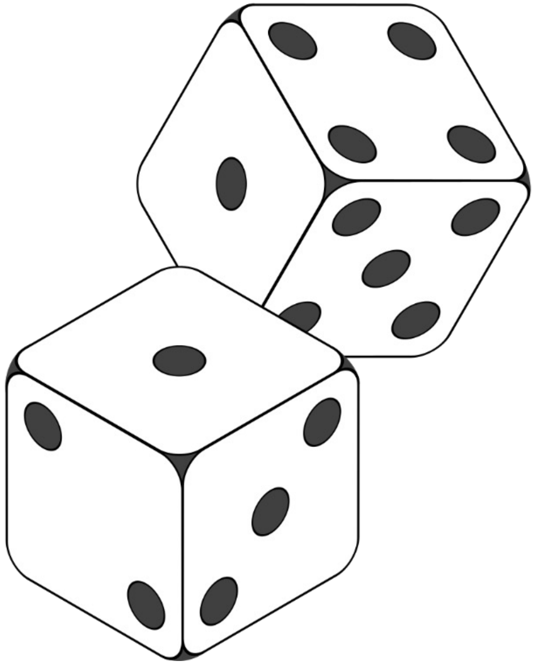
- Random Variables and Discrete Probability
- Fundamental Rules of Probability
- Expected Value and Moments
- Continuous Probability
- Bayesian Inference

Outline

- Random Variables and Discrete Probability
- Fundamental Rules of Probability
- Expected Value and Moments
- Continuous Probability
- Bayesian Inference

Random Variables

(Informally) A random variable is an unknown quantity whose value depends on the outcome of a random process



Example Roll 2 dice and let random variable X represent their sum. It takes values,

$$X \in \{2, 3, 4, \ldots, 12\}$$

Example Flip a coin and let random variable Y represent the outcome,

$$Y \in \{\text{Heads}, \text{Tails}\}$$

Discrete vs. Continuous Probability

Discrete RVs take on a finite or countably infinite set of values

Continuous RVs take an uncountably infinite set of values

- Representing / interpreting / computing probabilities becomes more complicated in the continuous setting
- We will focus on discrete RVs for now…

Random Variables and Probability

Capitol letters represent random variables

Lowercase letters are realized *values*

$$X = x$$

$X = x$ is the **event** that X takes the value x

Example Let X be the random variable (RV) representing the sum of two dice with values,

$$X \in \{2, 3, 4, \ldots, 12\}$$

X=5 is the *event* that the dice sum to 5.

Probability Mass Function

A function $p(X)$ is a **probability mass function (PMF)** of a discrete random variable if the following conditions hold:

(a) It is nonnegative for all values in the support,

$$p(X = x) \geq 0$$

(b) The sum over all values in the support is 1,

$$\sum_x p(X = x) = 1$$

Intuition Probability mass is conserved, just as in physical mass. Reducing probability mass of one event must increase probability mass of other events so that the definition holds...

Probability Mass Function

Example Let X be the outcome of a single fair die. It has the PMF,

$$p(X = x) = \frac{1}{6} \qquad \text{for } x = 1, \ldots, 6$$

Uniform Distribution

Example We can often represent the PMF as a vector. Let S be an RV that is the *sum of two fair dice*. The PMF is then,

Observe that S does not follow a uniform distribution

$$p(S) = \begin{pmatrix} p(S=2) \\ p(S=3) \\ p(S=4) \\ \vdots \\ p(S=12) \end{pmatrix} = \begin{pmatrix} 1/36 \\ 1/18 \\ 1/2 \\ \vdots \\ 1/36 \end{pmatrix}$$

Functions of Random Variables

Any function $f(X)$ of a random variable X is also a random variable and it has a probability distribution

Example Let X_1 be an RV that represents the result of a fair die, and let X_2 be the result of another fair die. Then,

$$S = X_1 + X_2$$

Is an RV that is the *sum of two fair dice* with PMF $p(S)$.

NOTE Even if we know the PMF $p(X)$ and we know that the PMF $p(f(X))$ exists, it is not always easy to calculate!

PMF Notation

- We use $p(X)$ to refer to the probability mass *function* (i.e. a function of the RV X)

- We use $p(X=x)$ to refer to the probability of the *outcome* $X=x$ (also called an "event")

- We will often use $p(x)$ as shorthand for $p(X=x)$

Outline

- Random Variables and Discrete Probability
- Fundamental Rules of Probability
- Expected Value and Moments
- Continuous Probability
- Bayesian Inference

Joint Probability

Definition Two (discrete) RVs X and Y have a *joint PMF* denoted by $p(X, Y)$ and the probability of the event X=x and Y=y denoted by $p(X = x, Y = y)$ where,

(a) It is nonnegative for all values in the support,

$$p(X = x, Y = y) \geq 0$$

(b) The sum over all values in the support is 1,

$$\sum_x \sum_y p(X = x, Y = y) = 1$$

Joint Probability

Let X and Y be *binary RVs.* We can represent the joint PMF p(X,Y) as a 2x2 array (table):

		Y	
		0	1
X	0	0.04	0.36
	1	0.30	0.30

All values are nonnegative

Joint Probability

Let X and Y be *binary RVs.* We can represent the joint PMF p(X,Y) as a 2x2 array (table):

		Y = 0	Y = 1
X	0	0.04	0.36
	1	0.30	0.30

The sum over all values is 1:
0.04 + 0.36 + 0.30 + 0.30 = 1

Joint Probability

Let X and Y be *binary RVs.* We can represent the joint PMF p(X,Y) as a 2x2 array (table):

		Y = 0	Y = 1
X	0	0.04	0.36
	1	0.30	0.30

P(X=1, Y=0) = 0.30

Fundamental Rules of Probability

Given two RVs X and Y the **conditional distribution** is:

$$p(X \mid Y) = \frac{p(X,Y)}{p(Y)} = \frac{p(X,Y)}{\sum_x p(X=x,Y)}$$

Multiply both sides by $p(Y)$ to obtain the **probability chain rule**:

$$p(X, Y) = p(Y)p(X \mid Y)$$

For N RVs $X_1, X_2, \ldots, X_N$:

$$p(X_1, X_2, \ldots, X_N) = p(X_1)p(X_2 \mid X_1) \ldots p(X_N \mid X_{N-1}, \ldots, X_1)$$

$$= p(X_1) \prod_{i=2}^{N} p(X_i \mid X_{i-1}, \ldots, X_1)$$

Chain rule valid for any ordering

Fundamental Rules of Probability

Law of total probability

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

- P(y) is a **marginal** distribution
- This is called **marginalization**

Proof

$$\sum_x p(Y, X = x) = \sum_x p(Y) p(X = x \mid Y) \quad \text{(chain rule)}$$

$$= p(Y) \sum_x p(X = x \mid Y) \quad \text{(distributive property)}$$

$$= p(Y) \quad \text{(PMF sums to 1)}$$

Generalization for conditionals:

$$p(Y \mid Z) = \sum_x p(Y, X = x \mid Z)$$

Tabular Method

Let X, Y be binary RVs with the joint probability table

For Binomial use K-by-K probability table.

		Y: y_1	Y: y_2	$P(x)$
X	x_1	0.04	0.36	0.4 ($P(x_1)$)
X	x_2	0.30	0.30	0.6 ($P(x_2)$)
$P(y)$		0.34 ($P(y_1)$)	0.66 ($P(y_2)$)	

$P(y_1)=P(x_1,y_1)+P(x_2,y_1)$
$P(y_2)=P(x_1,y_2)+P(x_2,y_2)$
[i.e., sum down columns]

$P(x_1)=P(x_1,y_1)+P(x_1,y_2)$
$P(x_2)=P(x_2,y_1)+P(x_2,y_2)$
[i.e., sum across rows]

Tabular Method

We don’t care about event $Y=y_2$

		Y	
		y_1	y_2
X	x_1	0.04	Censored!
	x_2	0.30	
		0.34	

0.34: $P(y_1)$

$P(x|y_1)=?$

Tabular Method

		$Y=y_1$
X	x_1	0.04
	x_2	0.30
		0.34

$P(y_1)$

→

0.04 / 0.34

0.30 / 0.34

$P(x|y_1)$

These sum to one:
A conditional probability distribution is still a probability distribution

Summary

- A **random variable** is an unknown quantity whose value depends on the outcome a random process (informal definition)
- $X = x$ Is an event with probability mass $p(X = x)$
- p(X) is a **probability mass function** (PMF) satisfying

$$p(X = x) \geq 0 \qquad \sum_x p(X = x) = 1$$

- Some fundamental rules of probability:
  - Conditional: $p(X \mid Y) = \frac{p(X,Y)}{p(Y)} = \frac{p(X,Y)}{\sum_x p(X=x,Y)}$
  - Law of total probability: $p(Y) = \sum_x p(Y, X = x)$
  - Probability chain rule: $p(X, Y) = p(Y)p(X \mid Y)$

Outline

- Random Variables and Discrete Probability
- Fundamental Rules of Probability
- Expected Value and Moments
- Continuous Probability
- Bayesian Inference

Moments of RVs

Definition *The* *<u>expectation</u>* *of a discrete RV* X*, denoted by* $\mathbf{E}[X]$*, is:*

$$\mathbf{E}[X] = \sum_x x\, p(X = x)$$

Summation over all values in domain of X

Example Let X be the sum of two fair dice, then:

$$\mathbf{E}[X] = \frac{1}{36} \cdot 2 + \frac{1}{18} \cdot 3 + \ldots + \frac{1}{36} \cdot 12 = 7$$

Theorem (Linearity of Expectations) *For any finite collection of discrete RVs* $X_1, X_2, \ldots, X_N$ *with finite expectations,*

$$\mathbf{E}\left[\sum_{i=1}^N X_i\right] = \sum_{i=1}^N \mathbf{E}[X_i]$$

Corollary *For any constant c*
$\mathbf{E}[cX] = c\mathbf{E}[X]$

E.g. for two RVs X and Y
$\mathbf{E}[X+Y] = \mathbf{E}[X] + \mathbf{E}[Y]$

Moments of RVs

Law of Total Expectation *Let X and Y be discrete RVs with finite expectations, then:*

$$\mathbf{E}[X] = \mathbf{E}_Y[\mathbf{E}_X[X \mid Y]]$$

Proof

$$\mathbf{E}_Y[\mathbf{E}_X[X \mid Y]] = \mathbf{E}_Y\left[\sum_x x \cdot p(x \mid Y)\right]$$

$$= \sum_y \left[\sum_x x \cdot p(x \mid y)\right] \cdot p(y)$$ **(Definition of expectation)**

$$= \sum_y \sum_x x \cdot p(x, y)$$ **(Probability chain rule)**

$$= \sum_x x \sum_y \cdot p(x, y)$$ **(Linearity of expectations)**

$$= \sum_x x \cdot p(x) = \mathbf{E}[X]$$ **(Law of total probability)**

Moments of RVs

Definition *The <u>conditional expectation</u> of a discrete RV X, given Y is:*

$$\mathbf{E}[X \mid Y = y] = \sum_x x\, p(X = x \mid Y = y)$$

Example Roll two standard six-sided dice and let X be the result of the first die and let Y be the sum of both dice, then:

$$\mathbf{E}[X_1 \mid Y = 5] = \sum_{x=1}^{4} x\, p(X_1 = x \mid Y = 5)$$

$$= \sum_{x=1}^{4} x \frac{p(X_1 = x, Y = 5)}{p(Y = 5)} = \sum_{x=1}^{4} x \frac{1/36}{4/36} = \frac{5}{2}$$

Conditional expectation follows properties of expectation (linearity, etc.)

Moments of RVs

Definition *The* *variance* *of a RV* X *is defined as,*

$$\mathbf{Var}[X] = \mathbf{E}[(X - \mathbf{E}[X])^2]$$ (X-units)^2

The *standard deviation* *is* $\sigma[X] = \sqrt{\mathbf{Var}[X]}$. (X-units)

Lemma An equivalent form of variance is:

$$\mathbf{Var}[X] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2$$

Proof Keep in mind that $E[X]$ is a constant,

$$\mathbf{E}[(X - \mathbf{E}[X])^2] = \mathbf{E}[X^2 - 2X\mathbf{E}[X] + \mathbf{E}[X]^2]$$ **(Distributive property)**

$$= \mathbf{E}[X^2] - 2\mathbf{E}[X]\mathbf{E}[X] + \mathbf{E}[X]^2$$ **(Linearity of expectations)**

$$= \mathbf{E}[X^2] - \mathbf{E}[X]^2$$ **(Algebra)**

Moments of RVs

Definition *The covariance of two RVs X and Y is defined as,*

$$\mathbf{Cov}(X,Y) = \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])]$$

Lemma *For any two RVs X and Y,*

$$\mathbf{Var}[X+Y] = \mathbf{Var}[X] + \mathbf{Var}[Y] + 2\mathbf{Cov}(X,Y)$$

e.g. variance is not a linear operator.

Proof $\mathbf{Var}[X+Y] = \mathbf{E}[(X+Y-\mathbf{E}[X+Y])^2]$

(Linearity of expectation) $= \mathbf{E}[(X+Y-\mathbf{E}[X]-\mathbf{E}[Y])^2]$

(Distributive property) $= \mathbf{E}[(X-\mathbf{E}[X])^2 + (Y-\mathbf{E}[Y])^2 + 2(X-\mathbf{E}[X])(Y-\mathbf{E}[Y])]$

(Linearity of expectation) $= \mathbf{E}[(X-\mathbf{E}[X])^2] + \mathbf{E}[(Y-\mathbf{E}[Y])^2] + 2\mathbf{E}[(X-\mathbf{E}[X])(Y-\mathbf{E}[Y])]$

(Definition of Var / Cov) $= \mathbf{Var}[X] + \mathbf{Var}[Y] + 2\mathbf{Cov}(X,Y)$

Summary

Moments and Expected Value

- Expected value of a discrete RV:

$$\mathbf{E}[X] = \sum_x x\, p(X = x)$$

- Expectation is a linear operator

$$\mathbf{E}\left[\sum_{i=1}^{N} X_i\right] = \sum_{i=1}^{N} \mathbf{E}[X_i]$$

- Variance of a RV:

$$\mathbf{Var}[X] = \mathbf{E}[(X - \mathbf{E}[X])^2]$$

- Variance is **not** a linear operator

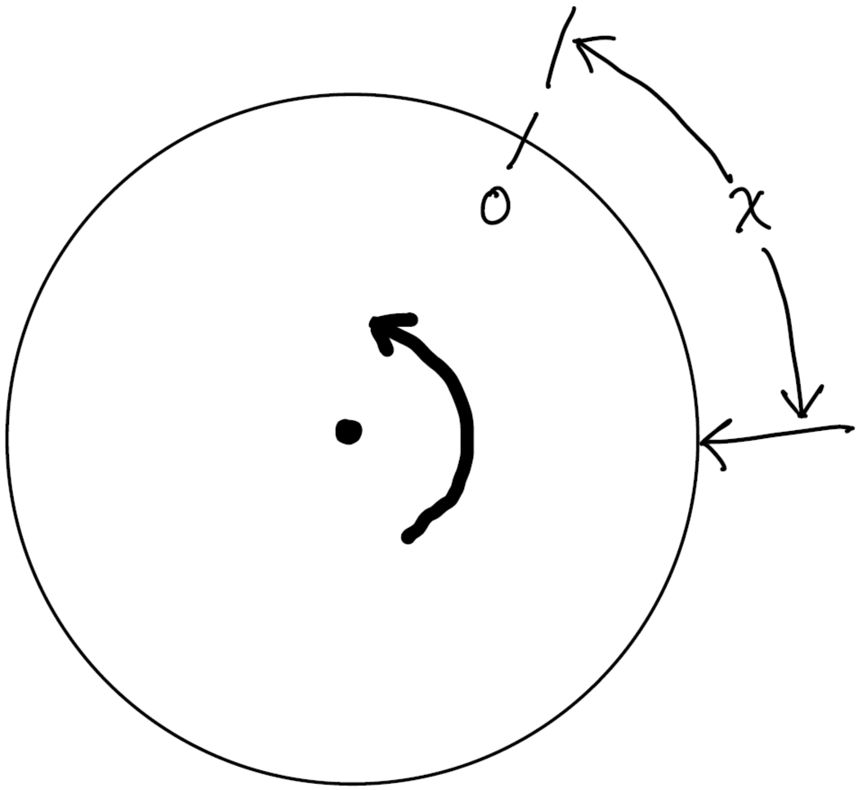
$$\mathbf{Var}[X + Y] = \mathbf{Var}[X] + \mathbf{Var}[Y] + 2\mathbf{Cov}(X, Y)$$

Outline

- Random Variables and Discrete Probability
- Fundamental Rules of Probability
- Expected Value and Moments
- Continuous Probability
- Bayesian Inference

Continuous Probability

Experiment Spin continuous wheel and measure X displacement from 0



Question Assuming uniform probability, what is $p(X = x)$?

Continuous Probability

- Let $p(X = x) = \pi$ be the probability of any single outcome

- Let $S(k)$ be set of any k *distinct* points in $[0, 1)$ then,

$$P(x \in S(k)) = k\pi$$

- Since $0 < P(x \in S(k)) < 1$ by axioms of probability, $k\pi < 1$ for any k

- Therefore: $\pi = 0$ and $P(x \in S(k)) = p(X = x) = 0$

Continuous Probability

- We have a well-defined event that x takes a value in set $x \in S(k)$
- Clearly this event can happen… i.e. **it is possible**
- But we have shown it has zero probability of occurring,

$$P(x \in S(k)) = 0$$

- By the axioms of probability, the probability that it **doesn't happen** is,

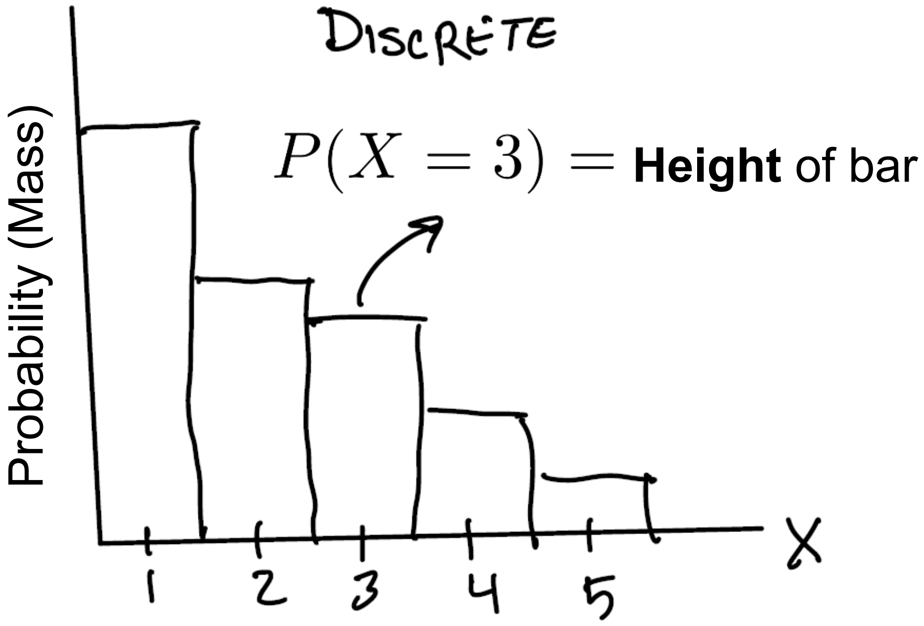
$$P(x \notin S(k)) = 1 - P(x \in S(k)) = 1$$

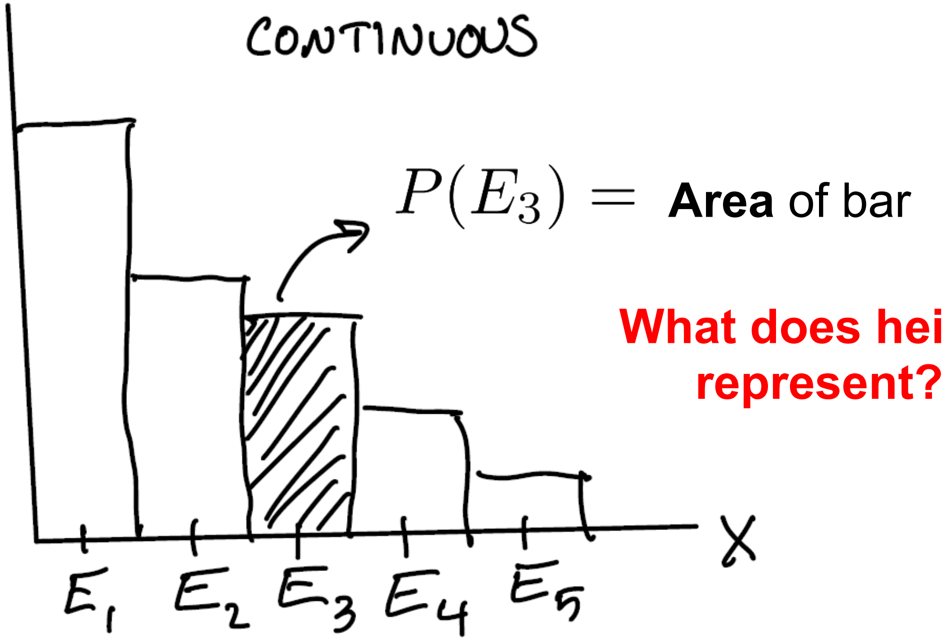
We seem to have a paradox!

Solution Rethink how we interpret probability in continuous setting

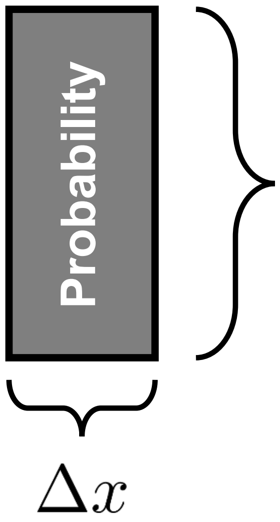
- Define events as *intervals* instead of discrete values
- Assign probability to those intervals

Continuous Probability





What does height represent?



Height $= \dfrac{\text{Probability}}{\Delta x}$

Height represents *probability per unit* in the x-direction

We call this a **probability density** (as opposed to probability mass)

Continuous Probability

- We denote the **probability density function** (PDF) as, $p(X)$
- An event E corresponds to an *interval* $a \leq X < b$
- The probability of an interval is given by the *area under the PDF,*

$$P(a \leq X < b) = \int_a^b p(X = x)\, dx$$

- Specific outcomes have zero probability $P(X = 0) = P(x \leq X < x) = 0$
- But may have nonzero *probability density* $p(X = x)$

Continuous Probability Measures

Definition *The* *<u>cumulative distribution function</u>* *(CDF) of a real-valued continuous RV X is the function given by,*

$$P(x) = P(X \leq x)$$

Different ways to represent probability of interval, CDF is just a convention.

- Can easily measure probability of closed intervals,

$$P(a \leq X < b) = P(b) - P(a)$$

- If *X* is *absolutely continuous* (i.e. differentiable) then,

Fundamental Theorem of Calculus

$$p(x) = \frac{dP(x)}{dx} \quad \text{and} \quad P(t) = \int_{-\infty}^{t} p(x)\, dx$$

Where $p(x)$ is the *probability density function* (PDF)

Continuous Probability

Most definitions for discrete RVs hold, replacing PMF with PDF/CDF…

Two RVs X & Y are **independent** if and only if,

$$p(x,y) = p(x)p(y) \qquad \text{or} \qquad P(X \le x, Y \le y) = P(X \le x)P(Y \le y)$$

Conditionally independent given Z iff,

Shorthand: $P(x) = P(X \le x)$

$$p(x,y \mid z) = p(x \mid z)p(y \mid z) \qquad \text{or} \qquad P(x,y \mid z) = P(x \mid z)P(y \mid z)$$

Probability chain rule,

$$p(x,y) = p(x)p(y \mid x) \qquad \text{and} \qquad P(x,y) = P(x)P(y \mid x)$$

Continuous Probability

…and by replacing summation with integration…

Law of Total Probability for continuous distributions,

$$p(x) = \int_{\mathcal{Y}} p(x, y) \, dy$$

Expectation of a continuous random variable,

$$\mathbf{E}[X] = \int_{\mathcal{X}} x \cdot p(x) \, dx$$

Covariance of two continuous random variables X & Y,

$$\mathbf{Cov}(X, Y) = \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])] = \int_{\mathcal{X}} \int_{\mathcal{Y}} (x - \mathbf{E}[X])(y - \mathbf{E}[Y]) p(x, y) \, dx dy$$

Continuous Probability

Caution *Some technical subtleties arise in continuous spaces…*

For **discrete** RVs X & Y, the conditional

P(Y=y)=0 means impossible

$$P(X = x \mid Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$$

is **undefined** when P(Y=y) = 0 … no problem.

For **continuous** RVs we have,

$$P(X \leq x \mid Y = y) = \frac{P(X \leq x, Y = y)}{P(Y = y)}$$

but numerator and denominator are 0/0.

P(Y=y)=0 means improbable, but not impossible

Continuous Probability

Defining the conditional distribution as a limit fixes this…

$$P(X \le x \mid Y = y) = \lim_{\delta \to 0} P(X \le x \mid y \le Y \le y + \delta)$$

$$= \lim_{\delta \to 0} \frac{P(X \le x, y \le Y \le y + \delta)}{P(y \le Y \le y + \delta)}$$

$$= \lim_{\delta \to 0} \frac{P(X \le x, Y \le y + \delta) - P(X \le x, Y \le y)}{P(Y \le y + \delta) - P(Y \le y)}$$

$$= \int_{-\infty}^{x} \lim_{\delta \to 0} \frac{\frac{\partial}{\partial x} P(u, y + \delta) - \frac{\partial}{\partial x} P(u, y)}{P(y + \delta) - P(y)} \, du$$

(Fundamental theorem of calculus)

(Assume interchange limit / integral)

$$= \int_{-\infty}^{x} \lim_{\delta \to 0} \frac{\left(\frac{\partial}{\partial x} P(u, y + \delta) - \frac{\partial}{\partial x} P(u, y)\right) / \delta}{\left(P(y + \delta) - P(y)\right) / \delta} \, du$$

(Multiply by $\frac{\delta}{\delta} = 1$)

$$= \int_{-\infty}^{x} \frac{\frac{\partial^2}{\partial x \partial y} P(u, y)}{\frac{\partial}{\partial y} P(y)} \, du \quad = \int_{-\infty}^{x} \frac{p(u, y)}{p(y)} \, du$$

(Definition of partial derivative)

(Definition PDF)

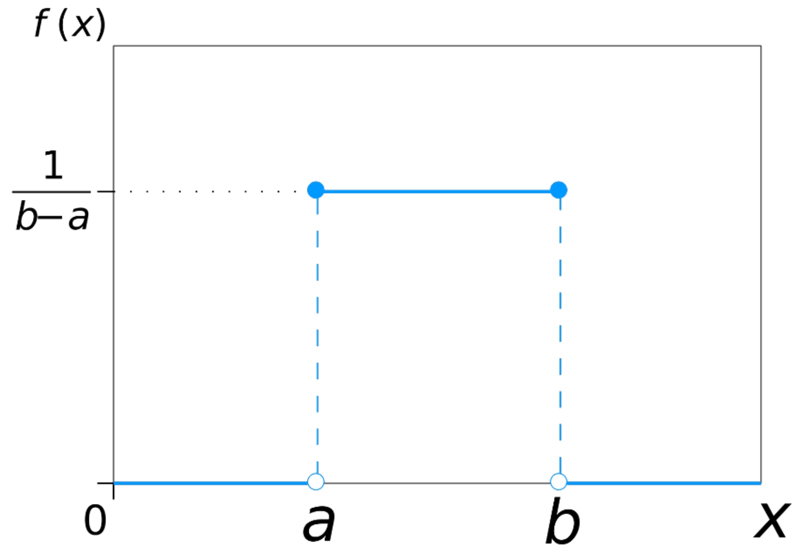
Definition The conditional PDF is given by,

$$p(x \mid y) = \frac{p(x, y)}{p(y)}$$

Useful Continuous Distributions

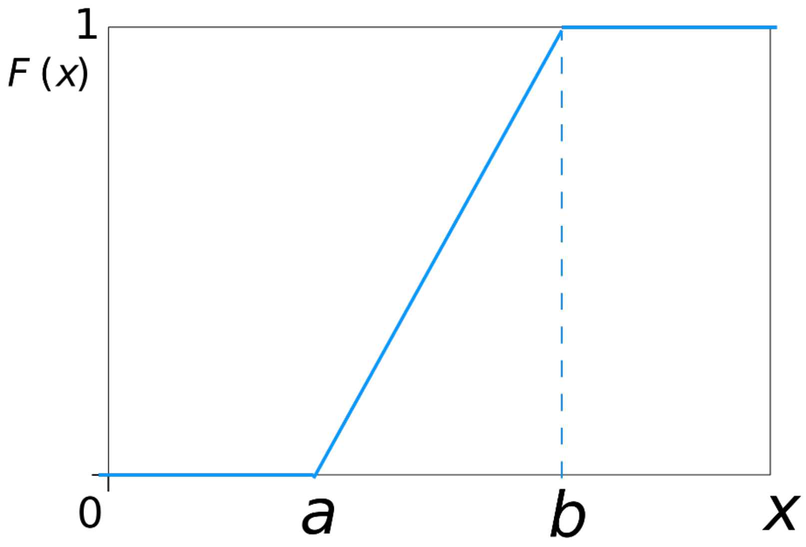
Uniform distribution on interval $[a, b]$,

$$p(x) = \begin{cases} 0 & \text{if } x \le a, \\ \frac{1}{b-a} & \text{if } a \le x \le b, \\ 0 & \text{if } b \le x \end{cases} \qquad P(X \le x) = \begin{cases} 0 & \text{if } x \le a, \\ \frac{x-a}{b-a} & \text{if } a \le x \le b, \\ 1 & \text{if } b \le x \end{cases}$$



Say that $X \sim U(a, b)$ whose moments are,

$$\mathbf{E}[X] = \frac{b+a}{2} \qquad \mathbf{Var}[X] = \frac{(b-a)^2}{12}$$



Suppose $X \sim U(0,1)$ and we are told $X \le \frac{1}{2}$ what is the conditional distribution?

$$P(X \le x \mid X \le \tfrac{1}{2}) = U(0, \tfrac{1}{2})$$

Holds generally: Uniform closed under conditioning

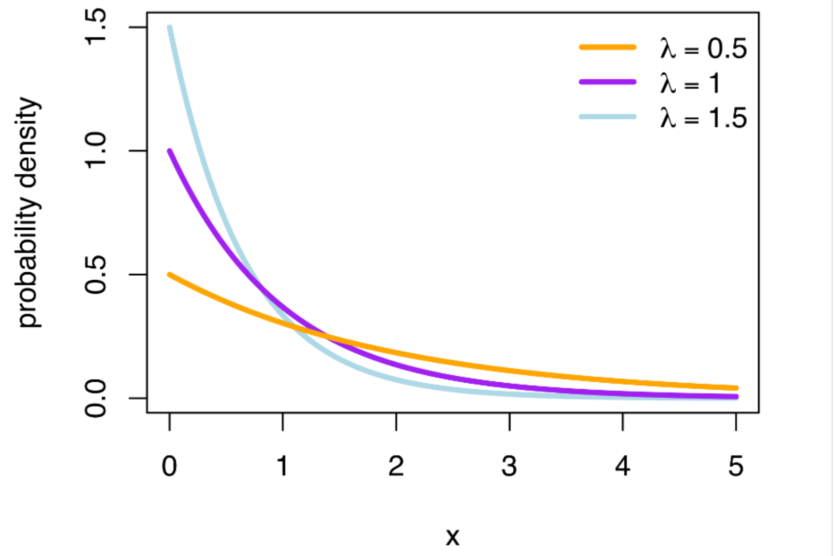
Useful Continuous Distributions

Exponential distribution with scale λ,

$$p(x) = \lambda e^{-\lambda x} \qquad P(x) = 1 - e^{-\lambda x}$$

for X>0. Moments given by,

$$\mathbf{E}[X] = \frac{1}{\lambda} \qquad \mathbf{Var}[X] = \frac{2}{\lambda^2}$$



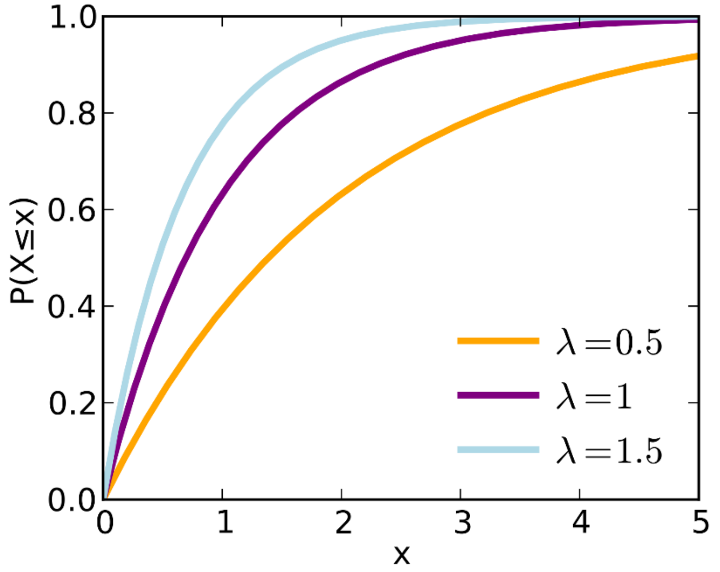
Useful properties

- **Closed under conditioning** If $X \sim \text{Exponential}(\lambda)$ then,

$$P(X \geq s + t \mid X \geq s) = P(X \geq s) = e^{-\lambda s}$$

- **Minimum** Let $X_1, X_2, \ldots, X_N$ be i.i.d. exponentially distributed with scale parameters $\lambda_1, \lambda_2, \ldots, \lambda_N$ then,

$$P(\min(X_1, X_2, \ldots, X_N)) = \text{Exponential}(\textstyle\sum_i \lambda_i)$$

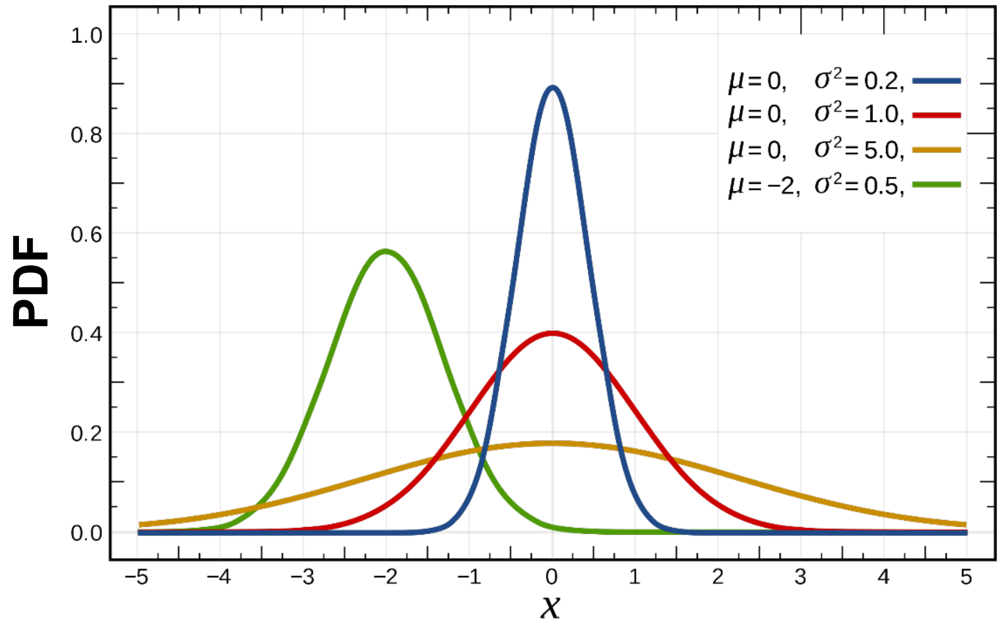


Useful Continuous Distributions

Gaussian (a.k.a. Normal) distribution with mean (location) μ and variance (scale) σ^2 parameters,

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{1}{2}(x-\mu)^2/\sigma^2$$

We say $X \sim \mathcal{N}(\mu, \sigma^2)$.



Useful Properties

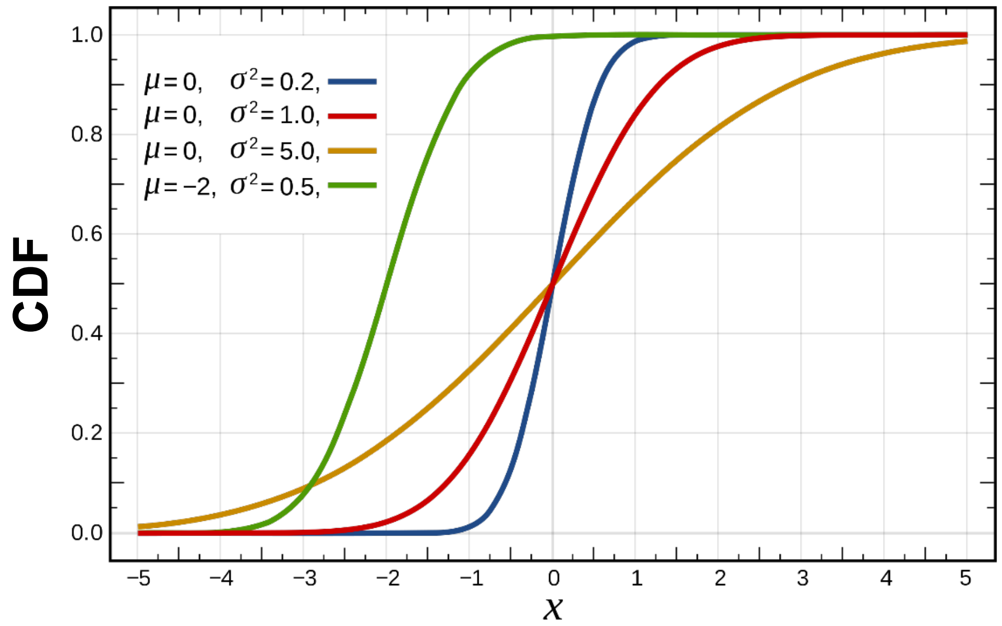
- Closed under additivity:

$$X \sim \mathcal{N}(\mu_x, \sigma_x^2) \qquad Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$$

$$X + Y \sim \mathcal{N}(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$$

- Closed under linear functions (a and b constant):

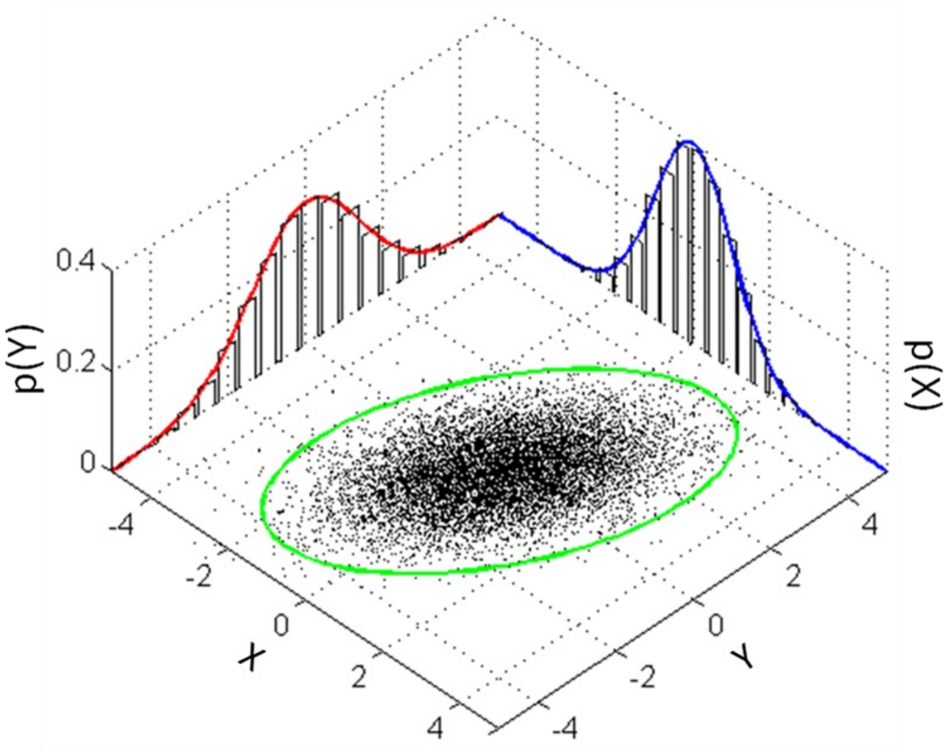
$$aX + b \sim \mathcal{N}(a\mu_x + b, a^2\sigma_x^2)$$



Useful Continuous Distributions

Multivariate Gaussian On RV $X \in \mathcal{R}^d$ with mean $\mu \in \mathcal{R}^d$ and positive semidefinite covariance matrix $\Sigma \in \mathcal{R}^{d \times d}$,

$$p(x) = |2\pi\Sigma|^{-1/2} \exp -\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)$$



Moments given by parameters directly.

Useful Properties

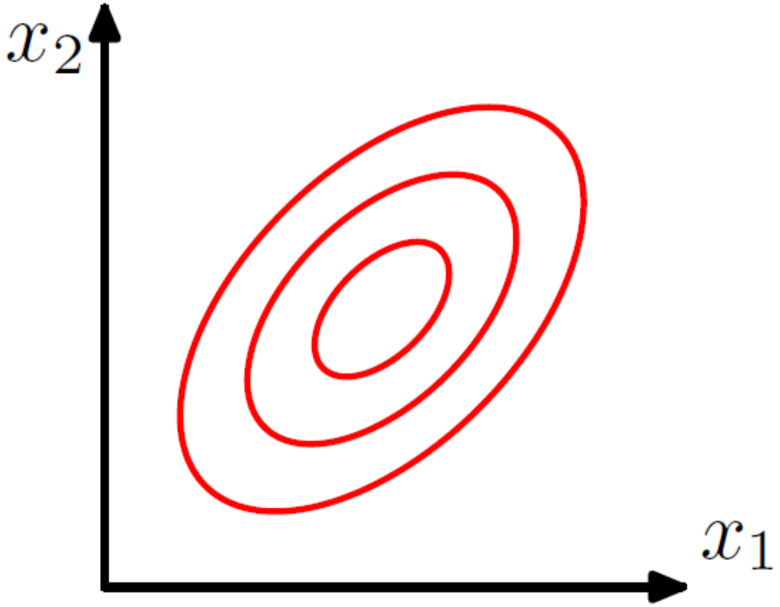
- Closed under additivity (same as univariate case)
- Closed under linear functions,

$$AX + b \sim \mathcal{N}(A\mu_x + b, A\Sigma A^T)$$

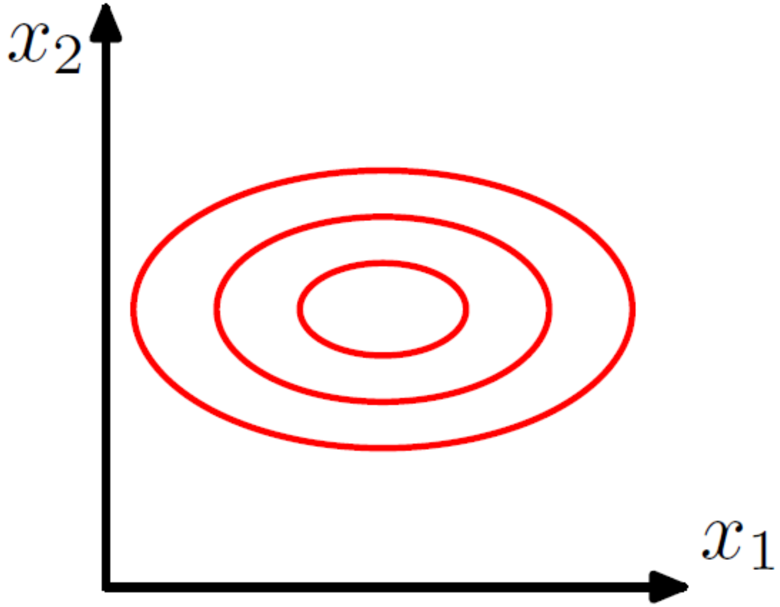
  Where $A \in \mathcal{R}^{m \times d}$ and $b \in \mathcal{R}^m$ (output dimensions may change)
- Closed under conditioning and marginalization

Covariance

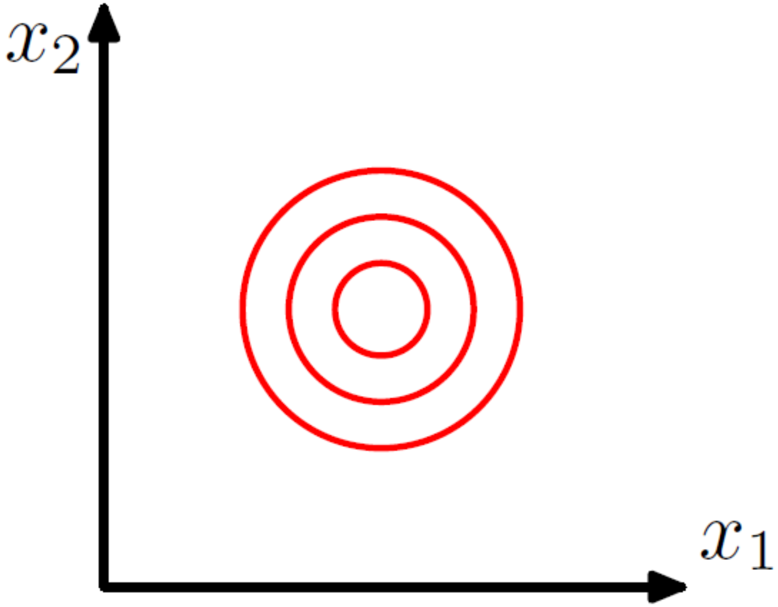
Captures correlation between random variables…can be viewed as set of ellipses…



Positive Correlation



Uncorrelated



Uncorrelated and same variance (isotropic / spherical)

Covariance Matrix

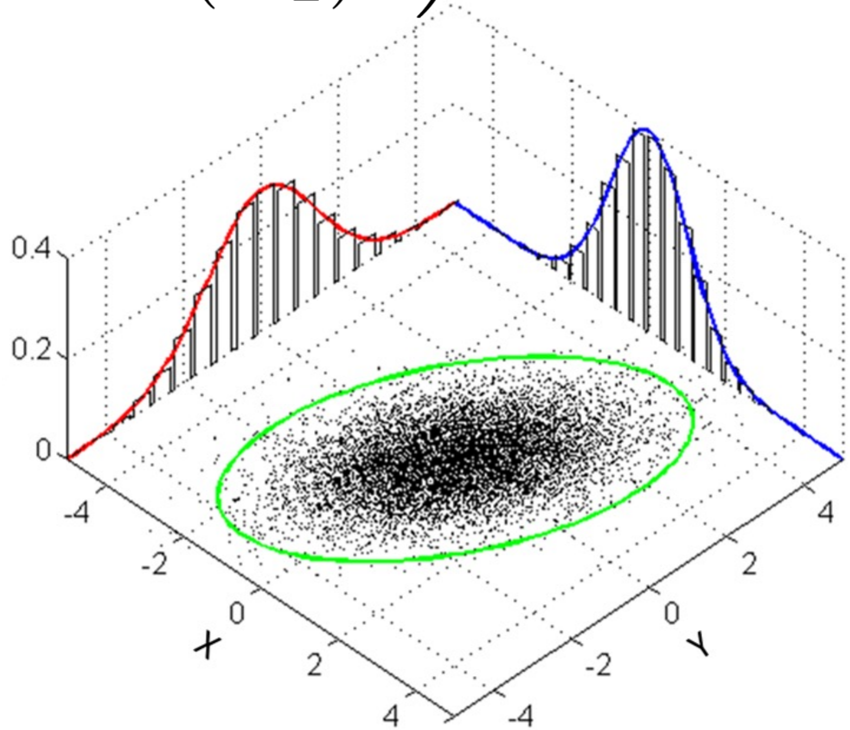
$$\Sigma = \mathrm{Cov}(X) = \begin{pmatrix} \mathrm{Var}(X_1) & \rho\sigma_{X_1}\sigma_{X_2} \\ \rho\sigma_{X_1}\sigma_{X_2} & \mathrm{Var}(X_2) \end{pmatrix}$$

Covariance Matrix

Marginal variance of just the RV X_1

$$\Sigma = \mathrm{Cov}(X) = \begin{pmatrix} \mathrm{Var}(X_1) & \rho\sigma_{X_1}\sigma_{X_2} \\ \rho\sigma_{X_1}\sigma_{X_2} & \mathrm{Var}(X_2) \end{pmatrix}$$

i.e. How "spread out" is the distribution in the X_1 dimension…



Covariance Matrix

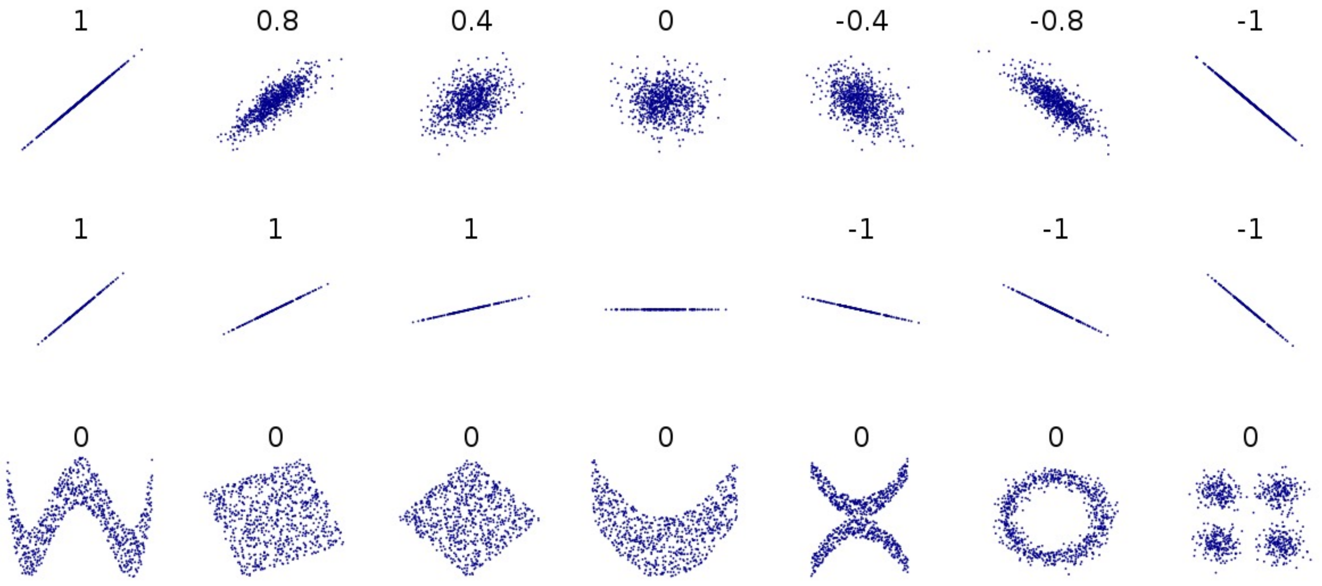
Correlation between X_1 and X_2

$$\Sigma = \mathrm{Cov}(X) = \begin{pmatrix} \mathrm{Var}(X_1) & \rho\sigma_{X_1}\sigma_{X_2} \\ \rho\sigma_{X_1}\sigma_{X_2} & \mathrm{Var}(X_2) \end{pmatrix}$$

Recall, correlation is given by:

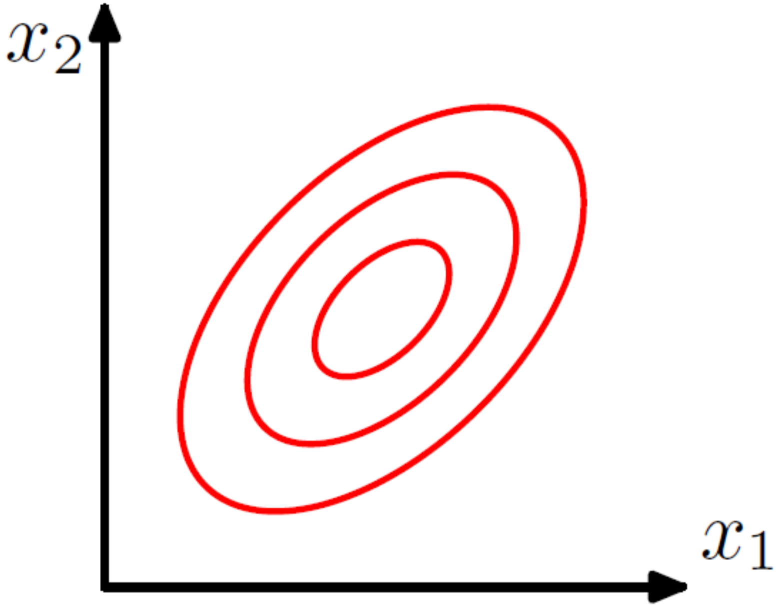
$$\rho = \frac{\mathbf{Cov}(X_1, X_2)}{\sigma_{X_1}\sigma_{X_2}}$$

Captures *linear dependence of RVs*



Covariance

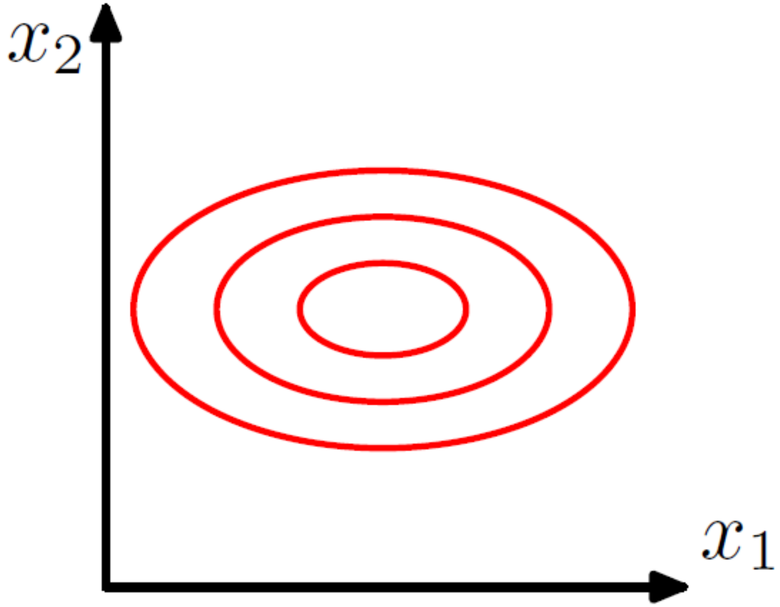
Captures correlation between random variables…can be viewed as set of ellipses…



Positive Correlation

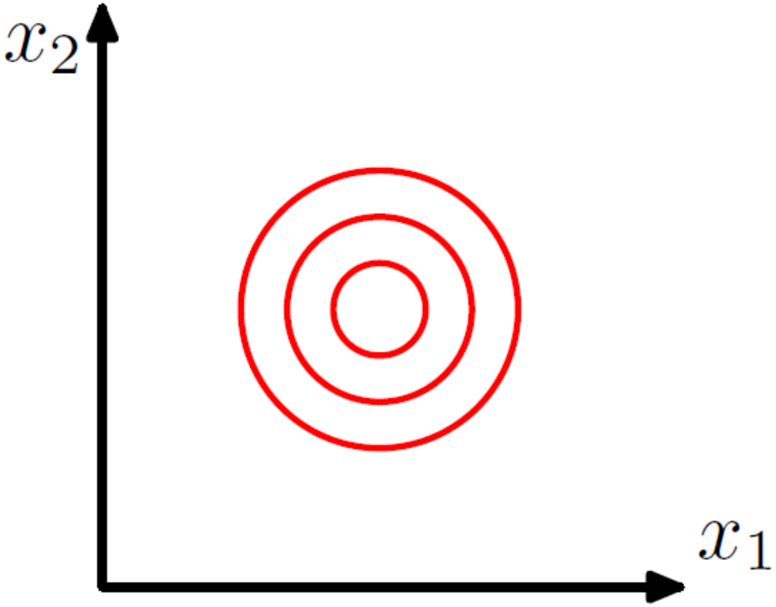
$\rho > 0$

Full matrix Σ



Uncorrelated

$$\Sigma = \begin{pmatrix} \sigma^2_{X_1} & 0 \\ 0 & \sigma^2_{X_2} \end{pmatrix}$$



Isotropic / Spherical

$$\Sigma = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix} = \sigma^2 I$$

Outline

- Random Variables and Discrete Probability
- Fundamental Rules of Probability
- Expected Value and Moments
- Continuous Probability
- Bayesian Inference

What is Probability?

What does it mean that the probability of heads is ½ ?

Two schools of thought…

Frequentist Perspective
Proportion of successes (heads) in repeated trials (coin tosses)

Bayesian Perspective
Belief of outcomes based on assumptions about nature and the physics of coin flips

Neither is better/worse, but we can compare interpretations…

Frequentist & Bayesian Modeling

We will use the following notation throughout:

θ - Unknown (e.g. coin bias) y - Data

Frequentist

(Conditional Model)

$$p(y; \theta)$$

- θ is a non-random unknown parameter
- $p(y; \theta)$ is the *sampling / data generating distribution*

Bayesian

(Generative Model)

Prior Belief → $p(\theta)p(y \mid \theta)$ ← Likelihood

- θ is a random variable (latent)
- Requires specifying $p(\theta)$ the prior belief

Bayes' Rule

Posterior represents all uncertainty after observing data…

prior probability

likelihood function for the parameters

$$p(\theta \mid y) = \frac{p(\theta)p(y \mid \theta)}{p(y)}$$

posterior probability

marginal likelihood
or: evidence
or: partition function
or: normalizer

Bayes' Rule : Marginal Likelihood

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

Often hard to calculate

Often know this (the model)

Marginal likelihood integrates (marginalizes) over unknown θ :

$$p(y) = \int p(\theta)p(y \mid \theta)\, d\theta$$

Marginal likelihood is less problematic in discrete models (not always)

This integral often lacks a closed form and cannot be computed…

Bayesian Inference Example

About 29% of American adults have high blood pressure (BP). Home test has 30% false positive rate and no false negative error.

A recent home test states that you have high BP. Should you start medication?

An Assessment of the Accuracy of Home Blood Pressure Monitors When Used in Device Owners

Jennifer S. Ringrose,[1] Gina Polley,[1] Donna McLean,[2–4] Ann Thompson,[1,5] Fraulein Morales,[1] and Raj Padwal[1,4,6]

Bayesian Inference Example

About 29% of American adults have high blood pressure (BP). Home test has 30% false positive rate and no false negative error.

- Latent quantity of interest is hypertension: $\theta \in \{true, false\}$
- Measurement of hypertension: $y \in \{true, false\}$
- Prior: $p(\theta = true) = 0.29$
- Likelihood: $p(y = true \mid \theta = false) = 0.30$

$$p(y = true \mid \theta = true) = 1.00$$

Bayesian Inference Example

About 29% of American adults have high blood pressure (BP). Home test has 30% false positive rate and no false negative error.

Suppose we get a positive measurement, then posterior is:

$$p(\theta = true \mid y = true) = \frac{p(\theta = true)p(y = true \mid \theta = true)}{p(y = true)}$$

$$= \frac{0.29 * 1.00}{0.29 * 1.00 + 0.71 * 0.30} \approx 0.58$$

What conclusions can be drawn from this calculation?

Aside : Proportionality

Recall PMF / PDF must sum / integrate to 1,

PMF

$$\sum_x p(x) = 1$$

PDF

$$\int p(x)\, dx = 1$$

May only know distribution constant that does not depend on RV x,

$$\int \widetilde{p}(x)\, dx = \mathcal{Z}$$

so

$$p(x) \propto \widetilde{p}(x)$$

Properly normalized distribution by dividing our <u>normalization constant:</u>

$$\int p(x)\, dx = \int \frac{1}{\mathcal{Z}} \widetilde{p}(x)\, dx = \frac{1}{\int \widetilde{p}(x)\, dx} \int \widetilde{p}(x)\, dx = 1$$

Aside : Proportionality

Example Let X be a Bernoulli RV (coinflip) with probabilities *proportional to:*

$$\tilde{p}(X=0) = 0.5 \qquad \tilde{p}(X=1) = 1.5$$

Greater than 1, but
It is an *unnormalized*
probability

Compute normalization constant,

$$\mathcal{Z} = \tilde{p}(X=0) + \tilde{p}(X=1) = 2.0$$

Normalize probability distribution,

$$p(X) = \frac{1}{\mathcal{Z}} \tilde{p}(X) = \begin{pmatrix} 1/4 \\ 3/4 \end{pmatrix}$$

Sums to 1

Frequentist vs. Bayesian Inference

We have data $X_1, \ldots, X_N$ and want to infer unknown parameter θ

Frequentist Inference

The data *uniquely determines* θ*, e.g.* by the likelihood:

Not a distribution on parameter $p(X_1, \ldots, X_N; \theta)$ **How well it explains the data**

Bayesian Inference

The data *updates our belief* about θ, which is random:

$$p(\theta \mid X_1, \ldots, X_N) \propto p(\theta \mid X_1, \ldots, X_{N-1}) p(X_N \mid \theta)$$

Our belief changes with more data

Minimum Mean Squared Error (MMSE)

Posterior mean minimizes squared error,

$$\hat{\theta}^{\mathrm{MMSE}} = \arg\min \mathbb{E}[(\hat{\theta} - \theta)^2 \mid y] = E[\theta \mid y]$$

- Minimizes error conditioned on observed data
- MMSE is an **unbiased estimator**
- MMSE is **asymptotically unbiased** and **asymptotically normal**,

$$\sqrt{N}(\hat{\theta}^{\mathrm{MMSE}} - \theta) \to \mathcal{N}(0, \sigma^2)$$

Bayes Estimators

Minimizes expected loss function,

$$\hat{\theta} = \arg\min_{\hat{\theta}} \mathbf{E}\left[L(\theta, \hat{\theta}) \mid y\right]$$

Expected loss referred to as *Bayes risk.*

MMSE minimizes squared-error loss $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$

Minimum absolute error (MAE) is posterior *median,*

$$\arg\min \mathbb{E}[|\hat{\theta} - \theta| \mid y] = \text{median}(\theta \mid y)$$

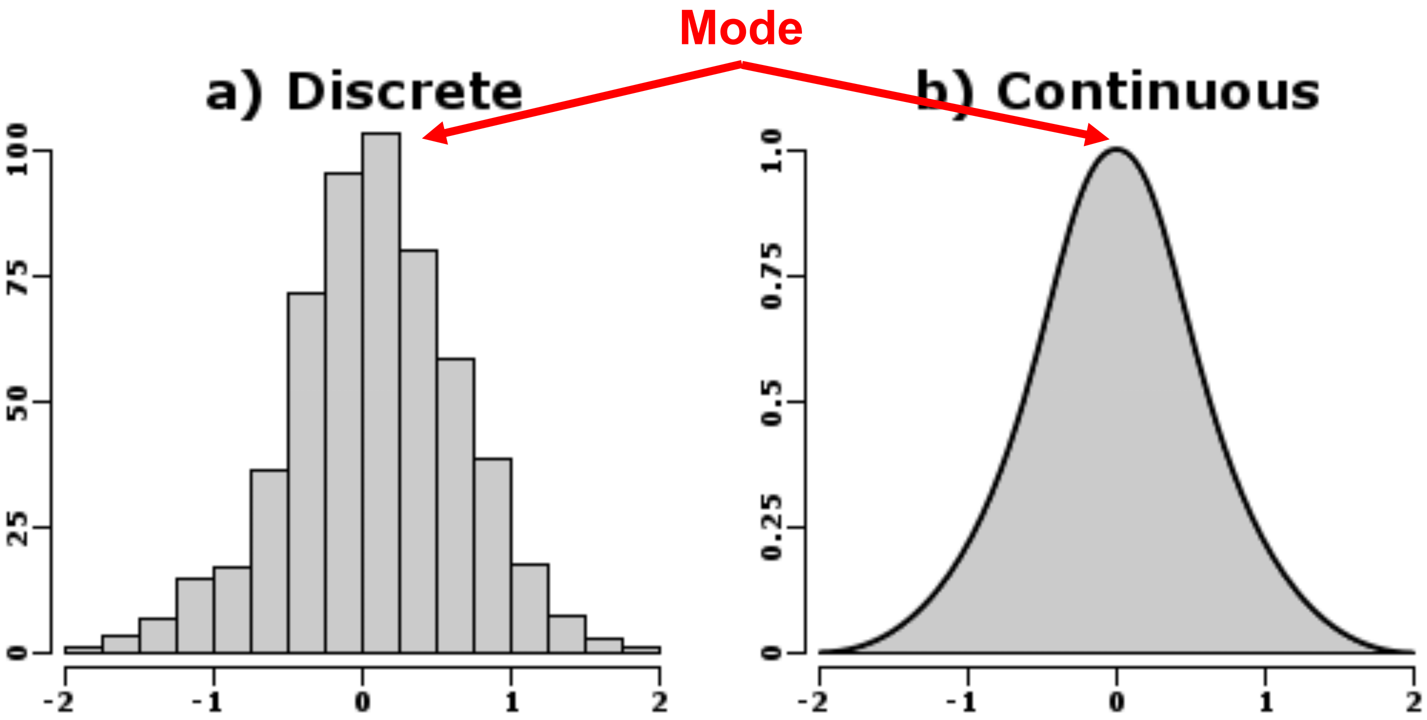
Note: Same answer for linear function: $L(\theta, \hat{\theta}) = c|\hat{\theta} - \theta|$

Maximum a Posteriori (MAP)

Very common to produce maximum probability estimates,

$$\hat{\theta}^{\text{MAP}} = \arg max\; p(\theta \mid y)$$

*MAP is the **mode** (highest probability outcome) of the posterior*



Maximum a Posteriori (MAP)

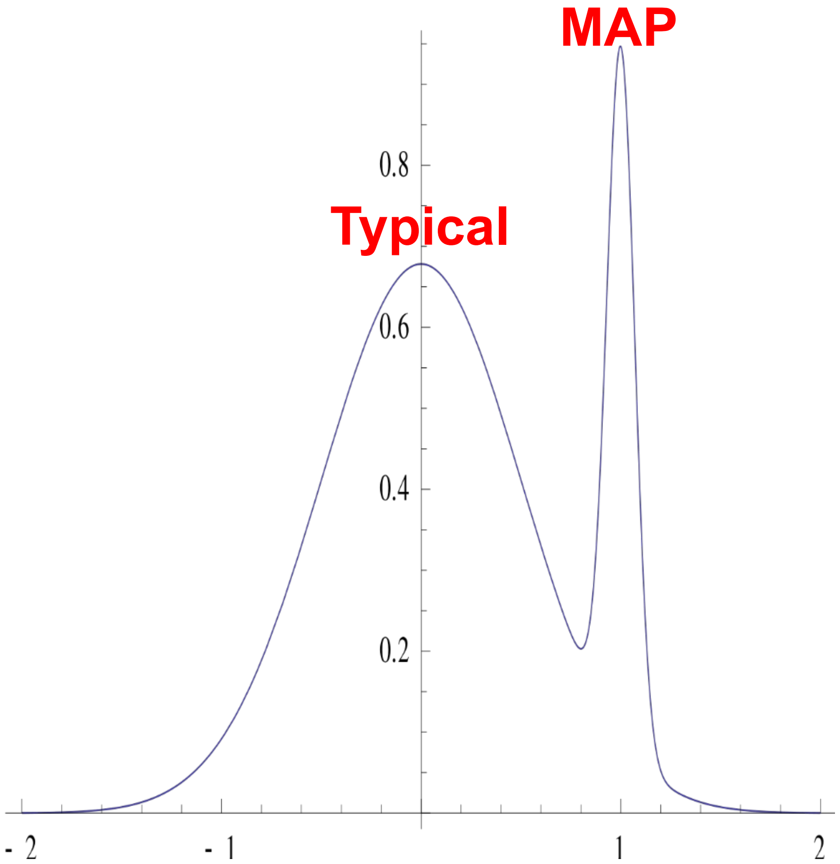
MAP (mode) may not be representative of typical outcomes

Also, not a Bayes estimator (unless discrete),

$$\lim_{c \to 0} L(\theta, \hat{\theta}) = \begin{cases} 0, & \text{if } |\hat{\theta} - \theta| < c \\ 1, & \text{otherwise} \end{cases}$$

Degenerate loss function

Despite its issues, MAP is frequently used in "Bayesian" inference and estimation



Posterior PDF

Example: Beta-Bernoulli MAP

Let $X_1, \ldots, X_N \sim \text{Bernoulli}(\pi)$ and $\pi \sim \text{Beta}(\alpha, \beta)$ then posterior is,

$$p(\pi \mid X_1^N) = \text{Beta}(\alpha + \underbrace{\textbf{number of heads}}_{N_H}, \ \beta + \textbf{number of tails})$$

Highest probability (mode) of Beta given by,

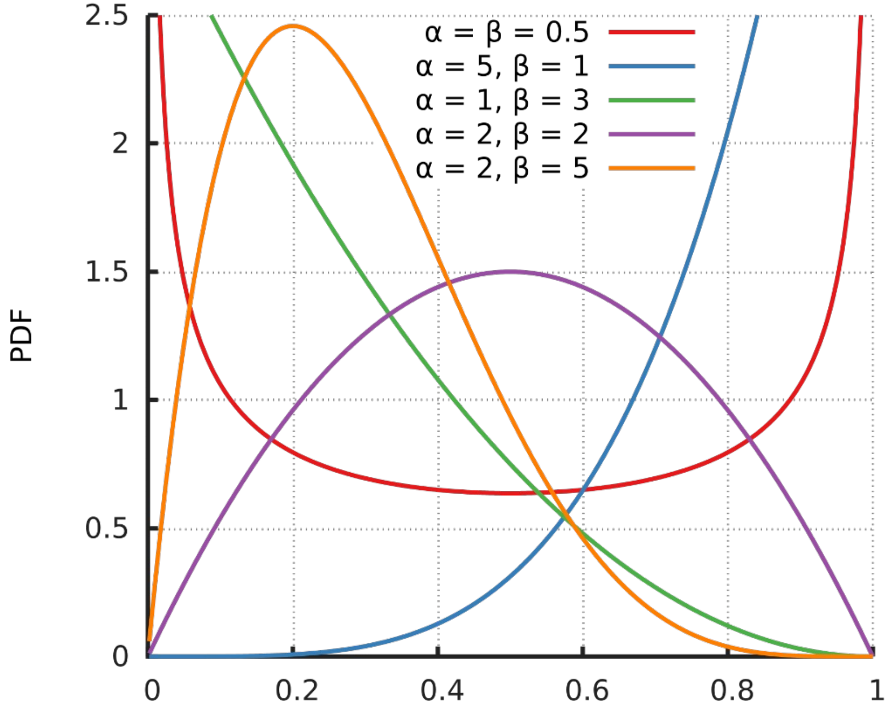
Take derivative, set to zero, solve.

$$\hat{\pi}^{\text{MAP}} = \frac{\alpha + N_H - 1}{\alpha + \beta + N - 2}$$

Beta distribution is not always convex!

- MAP is any value for $\alpha = \beta = 1$
- Two modes (bimodal) for $\alpha, \beta < 1$

Beta Posterior PDF



Maximum a Posteriori (MAP)

Equivalent to maximizing joint probability,

$$\arg\max_{\theta} p(\theta \mid y) = \arg\max_{\theta} \frac{p(\theta, y)}{p(y)} = \arg\max_{\theta} p(\theta, y)$$

Constant ($p(y)$)

For iid $y_1, \ldots, y_N$ solve in log-domain (like *maximum likelihood est.*),

$$\hat{\theta}^{\text{MAP}} = \arg\max_{\theta} \log p(\theta, y_1, \ldots, y_N) = \underbrace{\sum_i \log p(y_i \mid \theta)}_{\text{Log-Likelihood}} + \underbrace{\log p(\theta)}_{\text{Log-Prior}}$$

Log-Likelihood (how well it fits data)

Log-Prior (how well it agrees with prior)

Intuition *MAP is like MLE but with a "penalty" term (log-prior)*

Prediction

Can make predictions of unobserved $\tilde{y}$ before seeing any data,

$$p(\widetilde{y}) = \sum_k p(\theta = k) p(\widetilde{y} \mid \theta = k)$$

Similar calculation to marginal likelihood

*This is the **prior predictive** distribution*

For continuous parameters sum turns into integral,

$$p(\tilde{y}) = \int p(\theta) p(\tilde{y} \mid \theta) \, d\theta$$

*This is a prediction based on **no observed data***

Prediction

When we observe y we can predict future observations $\tilde{y}$,

$$p(\tilde{y} \mid y) = \sum_k \underbrace{p(\theta = k \mid y)}_{\text{This is now the posterior}} p(\tilde{y} \mid \theta = k)$$

*This is the **posterior predictive** distribution*

Again, for continuous parameters sum turns into integral,

$$p(\tilde{y} \mid y) = \int p(\theta \mid y) p(\tilde{y} \mid \theta) \, d\theta$$

Prediction Example

About 29% of American adults have high blood pressure (BP). Home test has 30% false positive rate and no false negative error.

What is the likelihood of *another* positive measurement?

$$p(\tilde{y} = true \mid y = true) = \sum_{\theta \in \{true, false\}} p(\theta \mid y = true) p(\tilde{y} = true \mid \theta)$$

$$= 0.42 * 0.30 + 0.58 * 1.00 \approx 0.71$$

What conclusions can be drawn from this calculation?

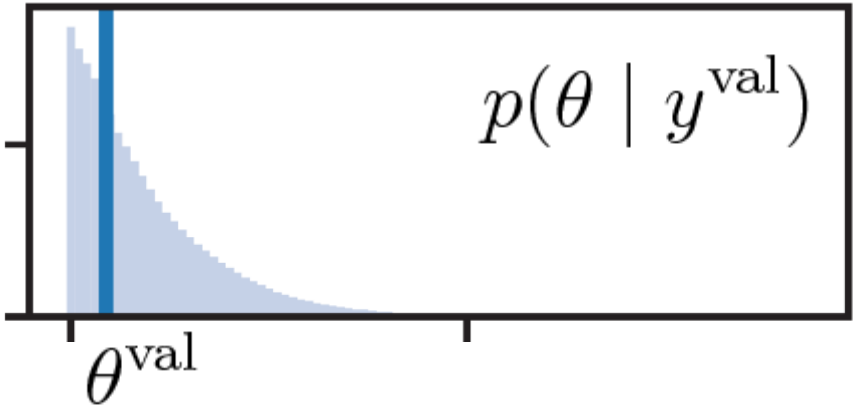
Model Validation

How do we know if the model $p(\theta, y)$ *is* *<u>good</u>*?

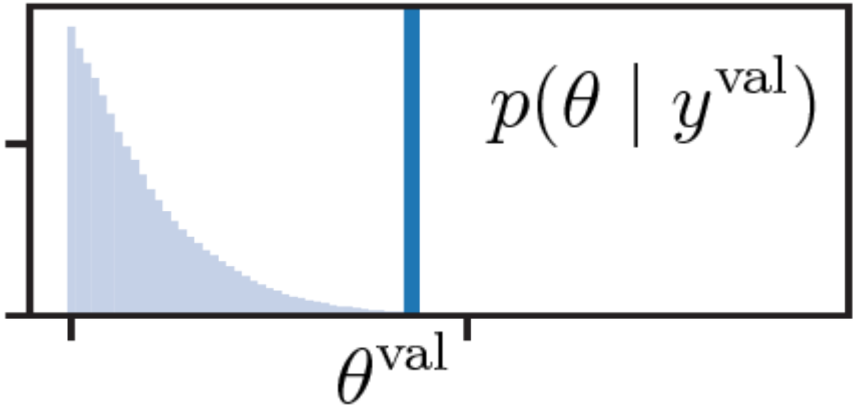
Supervised Learning

Validation set $\{(\theta^{\text{val}}, y^{\text{val}})\}$ consists of known θ^{val}. Are true values typically preferred under the posterior?

Good (maybe lucky)



Not Good (maybe unlucky)



Repeat trials over validation set for more certainty

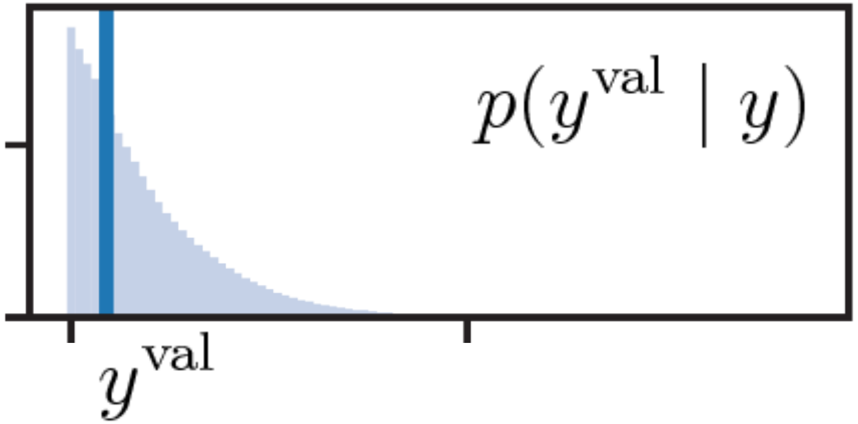
Model Validation

How do we know if the model $p(\theta, y)$ *is* <u>*good*</u>*?*

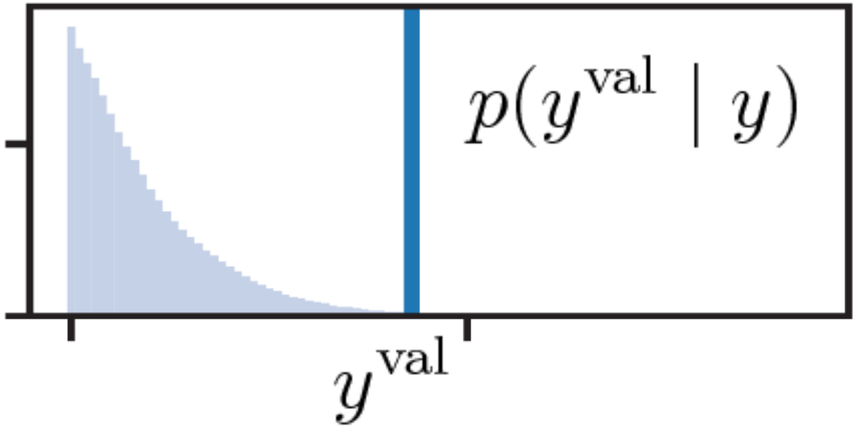
Unsupervised Learning

Validation set $\{y^{\text{val}}\}$ only contains observable data. Check validation data against posterior-predictive distribution.

Good (maybe lucky)



Not Good (maybe unlucky)



Repeat trials over validation set for more certainty

Likelihood and Odds Ratios

Which parameter value θ_1 or θ_2 is more likely to have generated the observed data y?

The **posterior odds ratio** is:

$$\frac{p(\theta_1 \mid y)}{p(\theta_2 \mid y)} = \frac{p(\theta_1)}{p(\theta_2)} \frac{p(y \mid \theta_1)}{p(y \mid \theta_2)} \frac{\cancel{p(y)}}{\cancel{p(y)}}$$

Prior Odds Ratio

Likelihood Ratio

Observe: the marginal likelihood $p(y)$ cancels!

Posterior Summarization

Ideally we would report the <u>full posterior distribution</u> as the result of inference…but this is not always possible

Summary of Posterior Location:

Point estimates: mean (MMSE), mode, median (min. absolute error)

Summary of Posterior Uncertainty:

Credible intervals / regions, posterior entropy, variance

Bayesian analysis should report uncertainty when possible

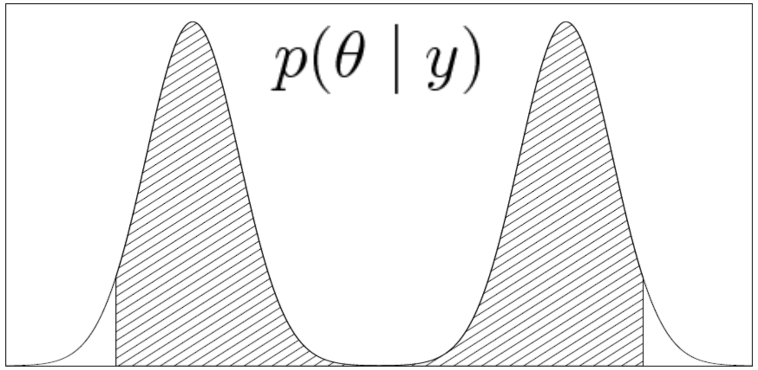
Credible Interval

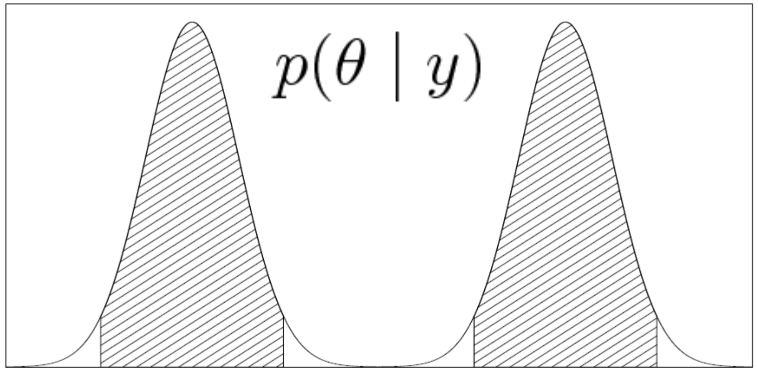
Def. For parameter $0 < \alpha < 1$ the $100(1-\alpha)\%$ credible interval $(L(y), U(y))$ satisfies,

$$p(L(y) < \theta < U(y) \mid y) = \int_{L(y)}^{U(y)} p(\theta \mid y) = 1 - \alpha$$

Interval containing fixed percentage of posterior probability density.

Note: This is not unique -- consider the 95% intervals below:





[Source: Gelman et al., "Bayesian Data Analysis"]

Summary

- Bayesian statistics interprets probability differently than classical stats
  - Frequentist: Probability → Long run odds in repeated trials
  - Bayesian: Probability → Belief of outcome that captures all uncertainty
- Bayesian models treat unknown parameter as random, with a prior
- Bayesian inference via the *posterior distribution* using Bayes' rule

$$p(\theta \mid y) = \frac{p(\theta)p(y \mid \theta)}{p(y)}$$

- Bayesian estimators minimize expected risk (e.g. MMSE)
- Maximum a posteriori (MAP) estimate maximizes posterior probability