CSC535: Probabilistic Graphical Models

Probability and Statistics

Prof. Jason Pacheco

Administrative Items

Office Hours
- Fridays @ 3-5pm
- **Need to shift hours earlier this Friday - will announce on Piazza**
- Held via Zoom
- Zoom meetings on D2L and link in Piazza

HW1
- Assignment on D2L
- Due by 11:59pm Monday January 30
- 10 points, 4 questions, discrete and continuous probability

Outline

➢ Random Variables and Discrete Probability

➢ Fundamental Rules of Probability

➢ Expected Value and Moments

➢ Useful Discrete Distributions

➢ Continuous Probability

Outline

➢ Random Variables and Discrete Probability

➢ Fundamental Rules of Probability

➢ Expected Value and Moments

➢ Useful Discrete Distributions

➢ Continuous Probability
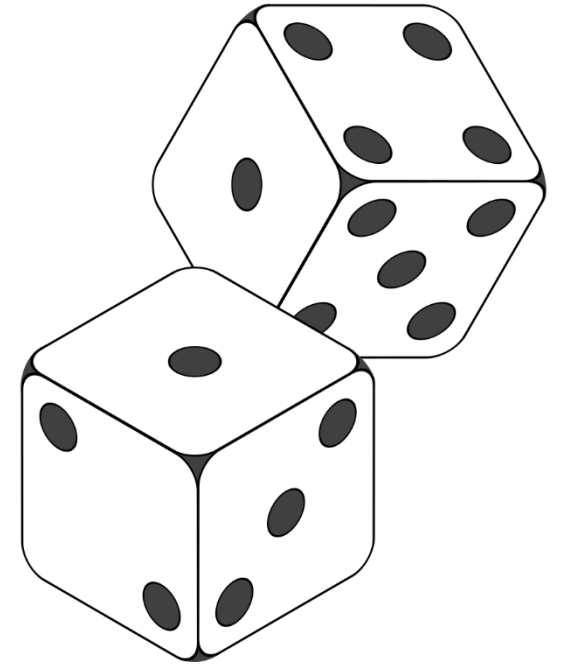
Suppose we roll <u>two fair dice</u>…

> ➢ What are the possible outcomes?
> ➢ What is the *probability* of rolling **even** numbers?
> ➢ What is the *probability* of rolling **odd** numbers?

…probability theory *gives a mathematical formalism to addressing such questions…*

Definition An **experiment** or **trial** is any process that can be repeated with well-defined outcomes. It is *random* if more than one outcome is possible.

Random Events and Probability

Definition An **outcome** is a possible result of an experiment or trial, and the collection of all possible outcomes is the **sample space** of the experiment,

Outcome

Example (1,1), (1,2), …, (6,1), (6,2), …, (6,6)

Sample Space
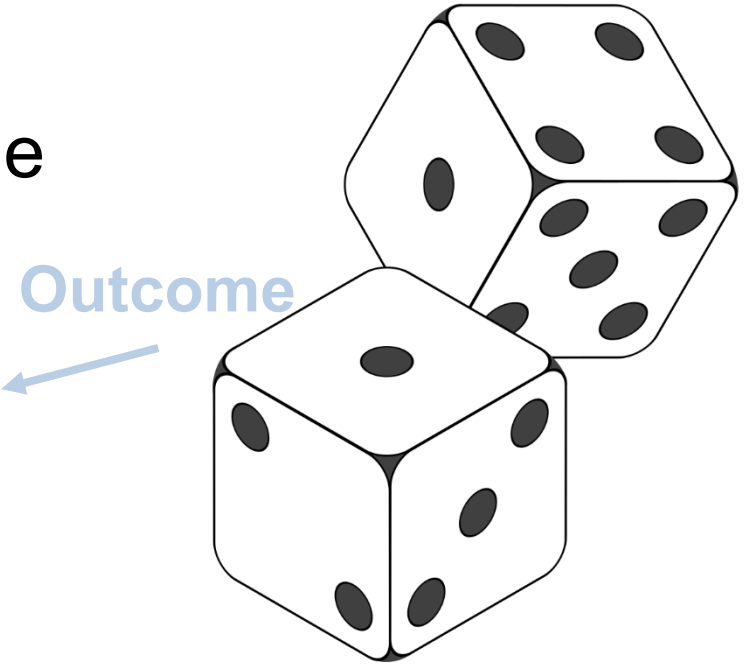
Definition An **event** is a *set* of outcomes (a subset of the sample space),

Example Event Roll at least a single 1
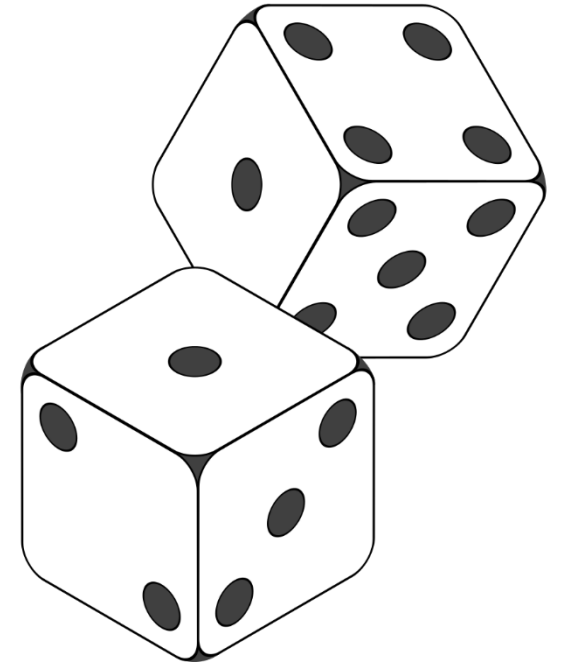
{(1,1), (1,2), (1,3), …, (1,6), …, (6,1)}

Assume each outcome is equally likely, and sample space is <u>finite</u>, then the probability of event is:

Number of outcomes in event set

$$P(E) = \frac{|E|}{|\Omega|}$$

Number of possible outcomes in sample space

This is the **uniform probability distribution**

Example Probability that we roll *only* even numbers,

$$E^{\mathrm{even}} = \{(2,2), (2,4), \ldots, (6,4), (6,6)\}$$

$$P(E^{\mathrm{even}}) = \frac{|E^{\mathrm{even}}|}{|\Omega|} = \frac{9}{36}$$

Example Probability that the *sum of both dice* is even,

$$E^{\text{sum even}} = \{(1,1),(1,3),(1,5),\ldots,(2,2),(2,4),\ldots\}$$

$$P(E^{\text{sum even}}) = \frac{|E^{\text{sum even}}|}{|\Omega|} = \frac{18}{36} = \frac{1}{2}$$

Example Probability that the *sum of both dice* is <u>greater than 12,</u>

$$E^{>12} = \emptyset$$

$$P(E^{>12}) = \frac{|E^{\text{sum even}}|}{|\Omega|} = 0$$

i.e. we can reason about the probability of <u>impossible</u> outcomes

Random Variables

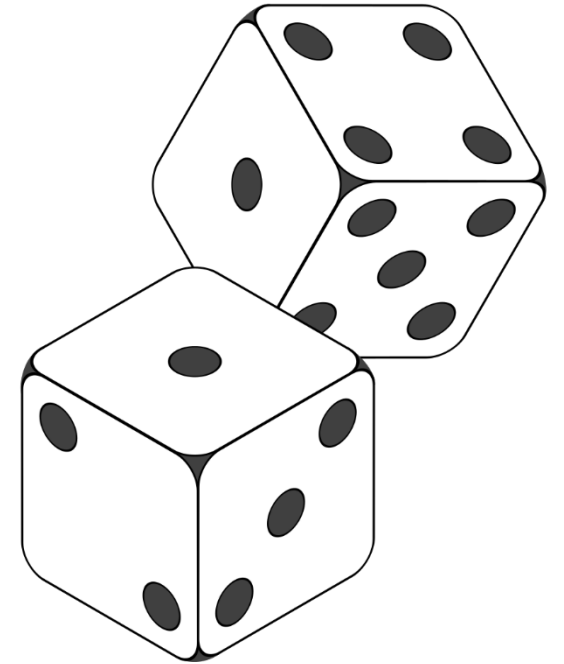*Suppose we are interested in a distribution over the <u>sum of dice</u>…*

Option 1 Let E_i be event that the sum equals *i*

Two dice example:

$$E_2 = \{(1,1)\} \quad E_3 = \{(1,2),(2,1)\} \quad E_4 = \{(1,3),(2,2),(3,1)\}$$

$$E_5 = \{(1,4),(2,3),(3,2),(4,1)\} \quad E_6 = \{(1,5),(2,4),(3,3),(4,2),(5,1)\}$$

Enumerate all possible means of obtaining desired sum. Gets cumbersome for N>2 dice…

Random Variables

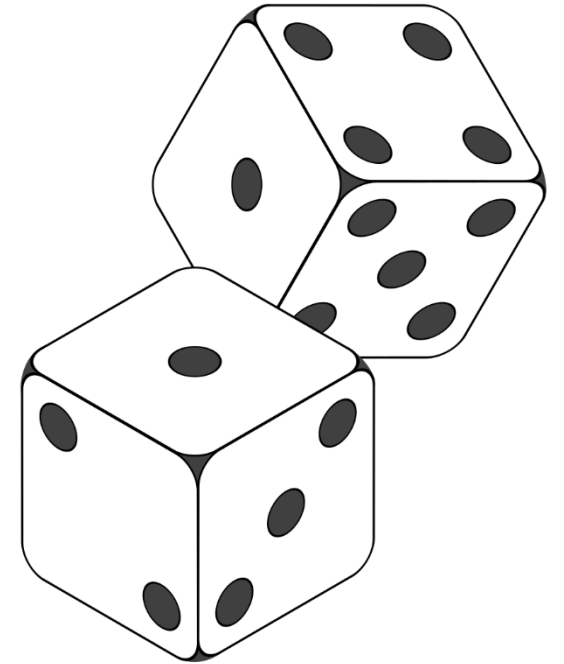**Option 2** Use a function of sample space…

(Informally) A random variable is an unknown quantity that maps events to numeric values.

Example X is the *sum of two dice* with 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

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

<u>Any</u> 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 $p(X=x)$

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,

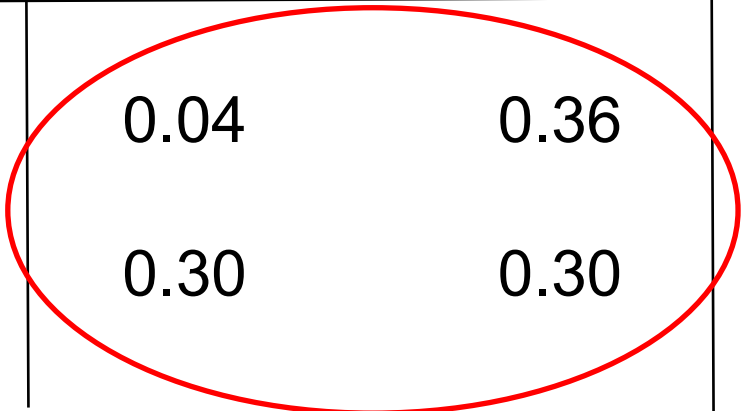(a) It is nonnegative for all values in the support,

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

(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	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	1	
	0	0.04	0.36	
X	1	0.30	0.30	

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

Outline

- ➢ Random Variables and Discrete Probability

- ➢ **Fundamental Rules of Probability**

- ➢ Expected Value and Moments

- ➢ Useful Discrete Distributions

- ➢ Continuous Probability

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**:

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

The probability chain rule extends to N RVs $X_1, X_2, \ldots, X_N$:

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

Chain rule valid for any ordering

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

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_2
x_1	0.04	0.36
x_2	0.30	0.30

X

0.4 ← $P(x_1)$

0.6 ← $P(x_2)$

P(x)

P(y)   0.34   0.66

$P(y_1)$ $P(y_2)$

$P(y_1)=P(x_1,y_1)+P(x_2,y_1)$
$P(y_2)=P(x_1,y_2)+P(x_2,y_2)$
 [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

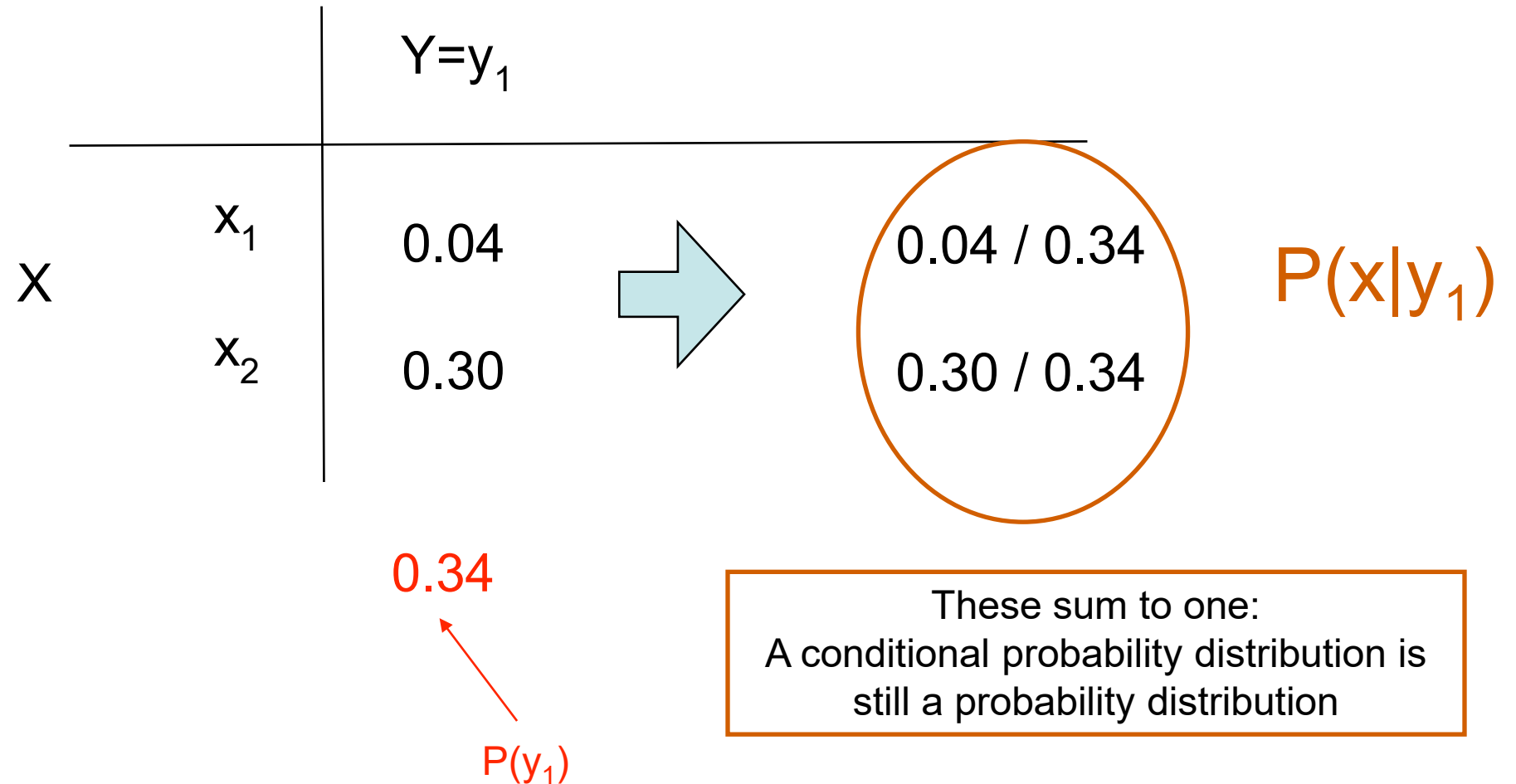|	y_1	y_2
x_1	0.04	
x_2	0.30	Censored!

X

0.34

P(x|y_1)=?

P(y_1)

Tabular Method

	Y=y_1
x_1	0.04
X	
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

Question: Roll two dice and let their outcomes be $X_1, X_2 \in \{1, \dots, 6\}$ *for die 1 and die 2, respectively. Recall the definition of conditional probability,*

$$p(X_1 \mid X_2) = \frac{p(X_1, X_2)}{p(X_2)}$$

Which of the following are true?

a) $p(X_1 = 1 | X_2 = 1) > p(X_1 = 1)$

b) $p(X_1 = 1 | X_2 = 1) = p(X_1 = 1)$ Outcome of die 2 doesn't *affect* die 1

c) $p(X_1 = 1 | X_2 = 1) < p(X_1 = 1)$

Question: Let $X_1 \in \{1, \ldots, 6\}$ *be outcome of die 1, as before. Now let* $X_3 \in \{2, 3, \ldots, 12\}$ *be the* sum of both dice. Which of the following are true?

a) $p(X_1 = 1 | X_3 = 3) > p(X_1 = 1)$

b) $p(X_1 = 1 | X_3 = 3) = p(X_1 = 1)$

c) $p(X_1 = 1 | X_3 = 3) < p(X_1 = 1)$

Only 2 ways to get $X_3 = 3$ *, each with equal probability:*

$(X_1 = 1, X_2 = 2)$ *or* $(X_1 = 2, X_2 = 1)$

so

$p(X_1 = 1 \mid X_3 = 3) = \dfrac{1}{2} > \dfrac{1}{6} = p(X_1 = 1)$

Intuition…

Consider $P(B|A)$ where you want to bet on B

Should you pay to know A?

In general you would pay something for A if it changed your belief about B. In other words if,

$$P(B|A) \neq P(B)$$

Independence of RVs

Definition *Two random variables X and Y are* <u>*independent*</u> *if and only if,*

$$p(X = x, Y = y) = p(X = x)p(Y = y)$$

for all values x and y, and we say $X \perp Y$.

Definition RVs $X_1, X_2, \ldots, X_N$ are <u>mutually independent</u> if and only if,

$$p(X_1 = x_1, \ldots, X_N = x_N) = \prod_{i=1}^{N} p(X_i = x_i)$$

➢ Independence is *symmetric:* $X \perp Y \Leftrightarrow Y \perp X$
➢ Equivalent definition of independence: $p(X \mid Y) = p(X)$

Definition *Two random variables X and Y are <u>conditionally independent</u> given Z if and only if,*

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

for all values x, y, and z, and we say that $X \perp Y \mid Z$.

➤ N RVs conditionally independent, given Z, if and only if:

$$p(X_1, \ldots, X_N \mid Z) = \prod_{i=1}^{N} p(X_i \mid Z)$$

> Shorthand notation
> Implies for all *x, y, z*

➤ Equivalent def'n of conditional independence: $p(X \mid Y, Z) = p(X \mid Z)$

➤ Symmetric: $X \perp Y \mid Z \Leftrightarrow Y \perp X \mid Z$

Outline

➢ Random Variables and Discrete Probability

➢ Fundamental Rules of Probability

➢ Expected Value and Moments

➢ Useful Discrete Distributions

➢ Continuous Probability

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,*

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

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

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

Theorem: *If $X \perp Y$ then $\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$.*

Proof:

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

$$= \sum_x \sum_y (x \cdot y) p(X = x) p(Y = y) \qquad \text{(Independence)}$$

$$= \left(\sum_x x \cdot p(X = x) \right) \left(\sum_y y \cdot p(Y = y) \right) = \mathbf{E}[X]\mathbf{E}[Y] \quad \textbf{(Linearity of Expectation)}$$

Example *Let $X_1, X_2 \in \{1, \ldots, 6\}$ be RVs representing the result of rolling two fair standard die.* *What is the mean of their product?*

$$\mathbf{E}[X_1 X_2] = \mathbf{E}[X_1]\mathbf{E}[X_2] = 3.5^2$$

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.)

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

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

$$= \sum_y \left[\sum_x x \cdot p(x \mid y)\right] \cdot p(y) \qquad \text{(Definition of expectation)}$$

$$= \sum_y \sum_x x \cdot p(x, y) \qquad \text{(Probability chain rule)}$$

$$= \sum_x x \sum_y \cdot p(x, y) \qquad \text{(Linearity of expectations)}$$

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

Definition *The <u>variance</u> of a RV X is defined as,*

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

The <u>standard deviation</u> 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)**

Definition *The <u>covariance</u> 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 <u>not a linear operator</u>.

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

Question: *What is the variance of the sum of <u>independent RVs</u>*

$$\mathbf{Var}[X_1 + X_2] = \mathbf{Var}[X_1] + \mathbf{Var}[X_2] + 2\mathbf{Cov}(X_1, X_2)$$

$$= \mathbf{Var}[X_1] + \mathbf{Var}[X_2] + 2\mathbf{E}[(X_1 - \mathbf{E}[X_1])(X_2 - \mathbf{E}[X_2])]$$

$$= \mathbf{Var}[X_1] + \mathbf{Var}[X_2] + 2\mathbf{E}[(X_1 - \mathbf{E}[X_1])]\mathbf{E}[(X_2 - \mathbf{E}[X_2])]$$

$$= \mathbf{Var}[X_1] + \mathbf{Var}[X_2] + 2\left(\mathbf{E}[X_1] - \mathbf{E}[X_1]\right)\left(\mathbf{E}[X_2] - \mathbf{E}[X_2]\right)$$

$$= \mathbf{Var}[X_1] + \mathbf{Var}[X_2]$$

E.g. variance is a *linear operator* for <u>independent RVs</u>

Theorem: *If* $X \perp Y$ *then* $\mathbf{Var}[X + Y] = \mathbf{Var}[X] + \mathbf{Var}[Y]$

Corollary: *If* $X \perp Y$ *then* $\mathbf{Cov}(X, Y) = 0$

Outline

➢ Random Variables and Discrete Probability

➢ Fundamental Rules of Probability

➢ Expected Value and Moments

➢ Useful Discrete Distributions

➢ Continuous Probability

Useful Discrete Distributions

Bernoulli *A.k.a. the **coinflip** distribution on <u>binary</u> RVs* $X \in \{0, 1\}$

$$p(X) = \pi^X (1 - \pi)^{(1-X)}$$

*Where π is the probability of **success** (e.g. heads), and also the mean*

$$\mathbf{E}[X] = \pi \cdot 1 + (1 - \pi) \cdot 0 = \pi$$

Suppose we flip N <u>independent</u> coins $X_1, X_2, \ldots, X_N$, what is the distribution over their sum $Y = \sum_{i=1}^{N} X_i$

Num. "successes" out of N trials

Num. ways to obtain k successes out of N

Binomial Dist. $\qquad p(Y = k) = \binom{N}{k} \pi^k (1 - \pi)^{N-k}$

Binomial Mean: $\mathbf{E}[Y] = N \cdot \pi$ **Sum of means for N indep. Bernoulli RVs**

Question: How many flips until we observe a success?

Geometric *Distribution on number of independent draws of* $X \sim \text{Bernoulli}(\pi)$ *until success:*

$$p(Y = n) = (1 - \pi)^{n-1}\pi \qquad \mathbf{E}[Y] = \frac{1}{\pi}$$

E.g. for fair coin $\pi = 1/2$ takes two flips on avg.

e.g. there must be n-1 failures (tails) before a success (heads).

Question: How many more flips of we have already seen k failures?

$$p(Y = n + k \mid Y > k) = \frac{p(Y = n+k, Y > k)}{p(Y > k)} = \frac{p(Y = n+k)}{p(Y > k)}$$

$$= \frac{(1-\pi)^{n+k-1}\pi}{\sum_{i=k}^{\infty}(1-\pi)^i\pi} = \frac{(1-\pi)^{n+k-1}\pi}{(1-\pi)^k} = (1-\pi)^{n-1}\pi = p(Y = n)$$

For $0 < x < 1, \sum_{i=k}^{\infty} x^i = x^k/(1-x)$

Corollary: $p(Y > k) = (1 - \pi)^{k-1}$

Categorical *Distribution on integer-valued RV* $X \in \{1, \ldots, K\}$

$$p(X) = \prod_{k=1}^{K} \pi_k^{\mathbf{I}(X=k)} \qquad \text{or} \qquad p(X) = \sum_{k=1}^{K} \mathbf{I}(X=k) \cdot \pi_k$$

with parameter $p(X = k) = \pi_k$ *and Kroenecker delta:*

$$\mathbf{I}(X = k) = \begin{cases} 1, & \text{If } X = k \\ 0, & \text{Otherwise} \end{cases}$$
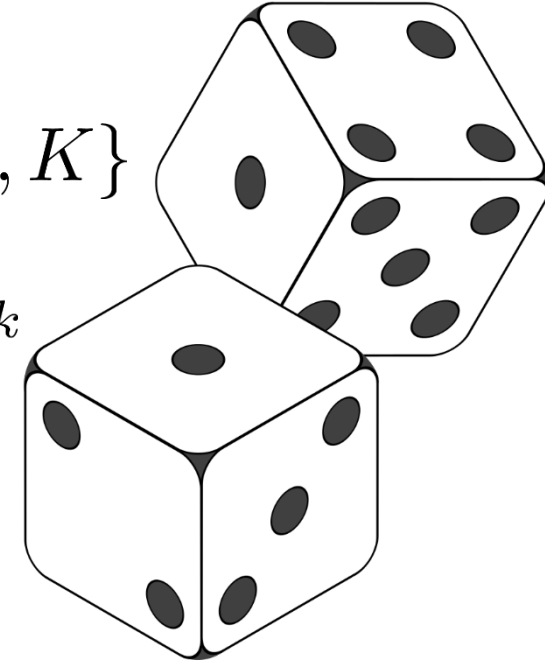
Can also represent X as *one-hot* binary vector,

$$X \in \{0, 1\}^K \quad \text{where} \quad \sum_{k=1}^{K} X_k = 1 \quad \text{then} \quad p(X) = \prod_{k=1}^{K} \pi_k^{X_k}$$

This representation is special case of the **multinomial distribution**

Useful Discrete Distributions

What if we count outcomes of *N* independent categorical RVs?

Multinomial *Distribution on K-vector $X \in \{0, N\}^K$ of counts of N repeated trials $\sum_{k=1}^{K} X_k = N$ with PMF:*

$$p(x_1, \ldots, x_K) = \binom{n}{x_1 x_2 \ldots x_K} \prod_{k=1}^{K} \pi_k^{x_k}$$

Number of ways to partition N objects into K groups:

$$\binom{n}{x_1 x_2 \ldots x_K} = \frac{n!}{x_1! x_2! \ldots x_K!}$$

Leading term ensures PMF is properly normalized:

$$\sum_{x_1} \sum_{x_2} \ldots \sum_{x_K} p(x_1, x_2, \ldots, x_K) = 1$$

Useful Discrete Distributions

*A **Poisson** RV X with <u>rate</u> parameter λ has the following distribution:*

$$p(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

$$\mathbf{E}[X] = \mathbf{Var}[X] = \lambda$$

> Mean and variance both scale with parameter

Represents number of times an *event* occurs in an interval of time or space.
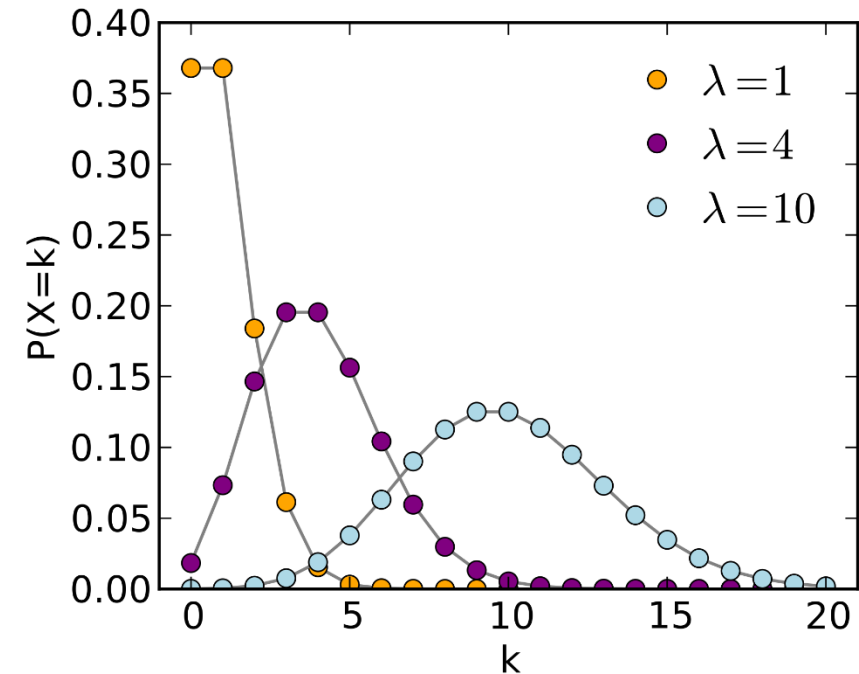
***Ex.** Probability of overflow floods in 100 years,*

$$p(k \text{overflow floods in 100 yrs}) = \frac{e^{-1}1^k}{k!}$$

> Avg. 1 overflow flood every 100 years, makes setting rate parameter easy.

***Lemma (additive closure)** The sum of a finite number of Poisson RVs is a Poisson RV.*

$$X \sim \text{Poisson}(\lambda_1), \quad Y \sim \text{Poisson}(\lambda_2), \quad X + Y \sim \text{Poisson}(\lambda_1 + \lambda_2)$$

Outline

➢ Random Variables and Discrete Probability

➢ Fundamental Rules of Probability

➢ Expected Value and Moments

➢ Useful Discrete Distributions

➢ Continuous Probability

Experiment Spin continuous wheel and measure X displacement from 0
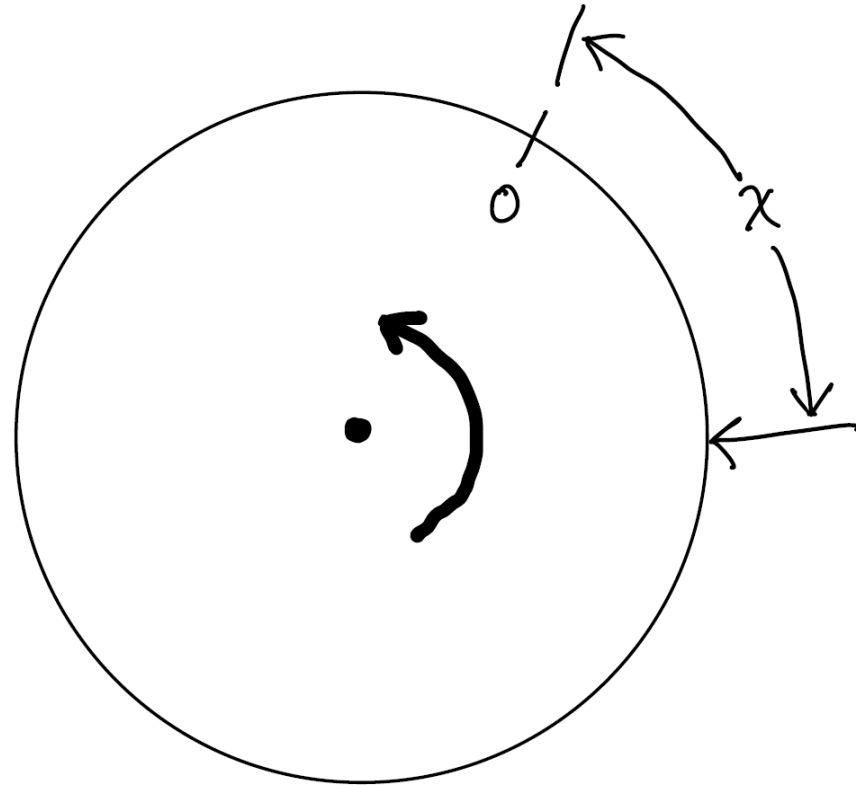
Question Assuming uniform probability, what is $p(X = x)$?

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 $0 < P(x \in S(k)) < 1$ we have that $k\pi < 1$ 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,

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

➢ 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

DISCRETE

$P(X = 3) = $ **Height** of bar

Probability (Mass)

1 2 3 4 5 X

CONTINUOUS

$P(E_3) = $ **Area** of bar

What does height represent?

E_1 E_2 E_3 E_4 E_5 X
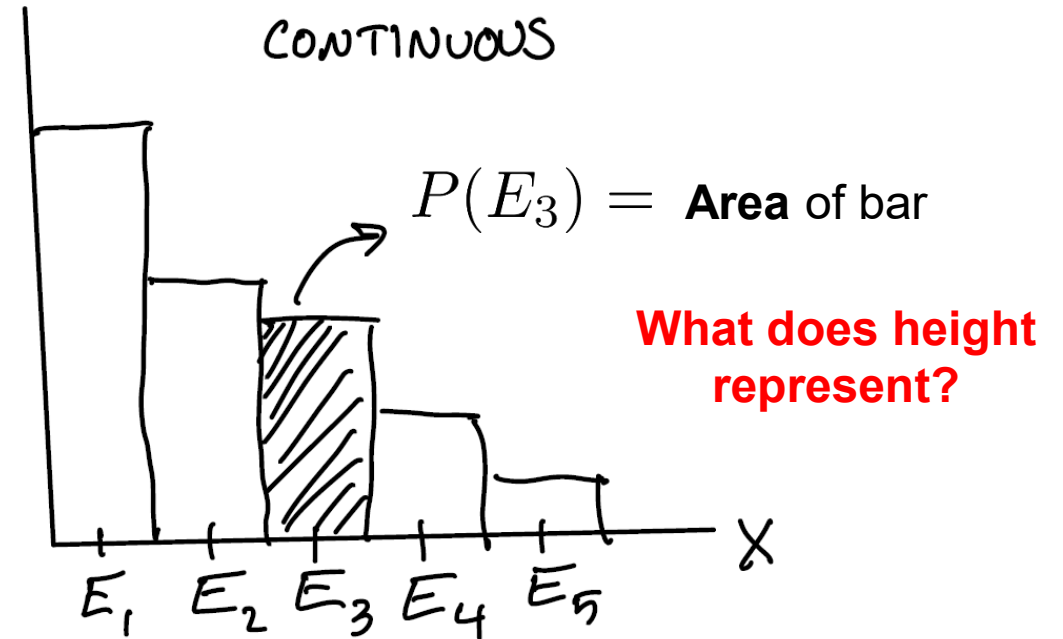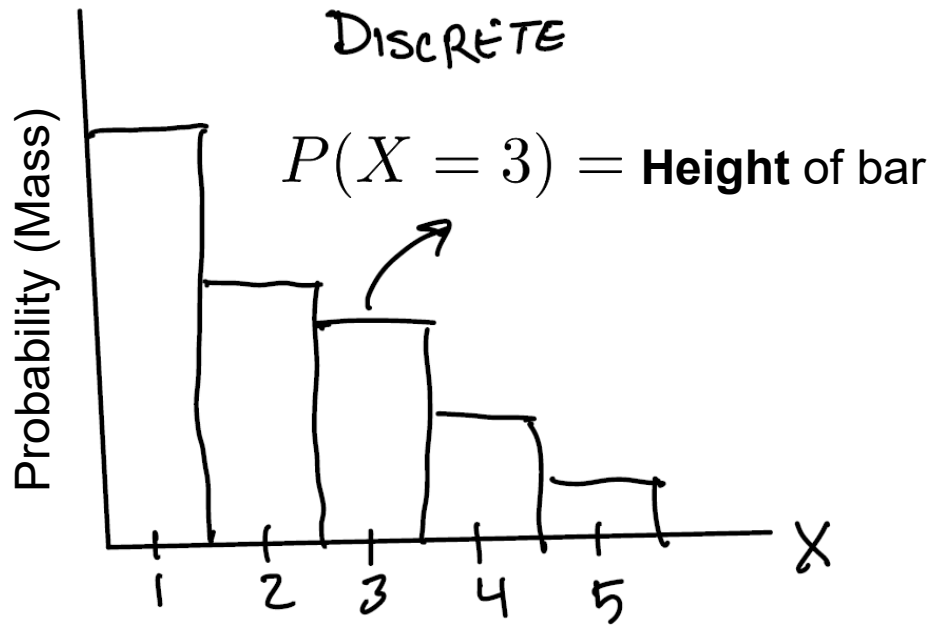
Probability

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

Δx

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,*

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

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

➢ Can easily measure probability of closed intervals,

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

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

Fundamental Theorem of Calculus

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

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,

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

Conditionally independent given Z iff,

$$\boxed{\textbf{Shorthand: } P(x) = P(X \leq x)}$$

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

Probability chain rule,

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

…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

Defining the conditional distribution as a limit fixes this…

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

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

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

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

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

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

> **Definition** The <u>conditional PDF</u> is given by,
> $$p(x \mid y) = \frac{p(x, y)}{p(y)}$$

(Fundamental theorem of calculus)

(Assume interchange limit / integral)

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

(Definition of partial derivative)

(Definition PDF)

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

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

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 $X \leq \frac{1}{2}$ what is the conditional distribution?

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

Holds generally: Uniform closed under conditioning

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 $X \sim \mathrm{Exponential}(\lambda)$ 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,

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

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):

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

Multivariate Gaussian On RV $X \in \mathcal{R}^d$
with mean $\mu \in \mathcal{R}^d$ and <u>positive semidefinite</u>
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
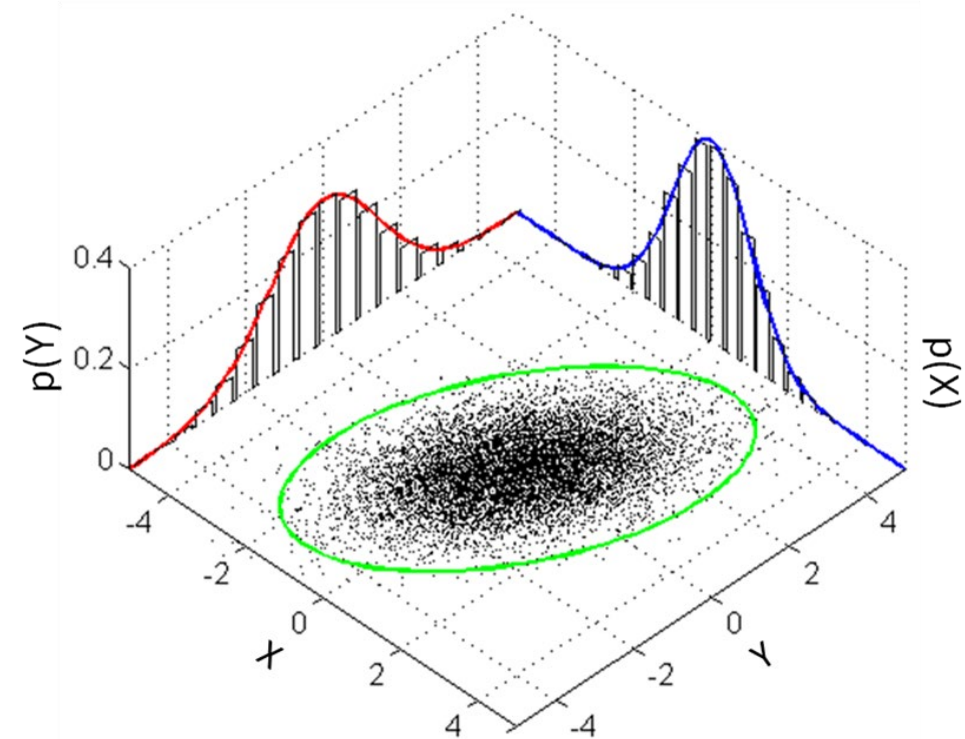
Covariance

Captures correlation between random variables…can be viewed as set of ellipses…
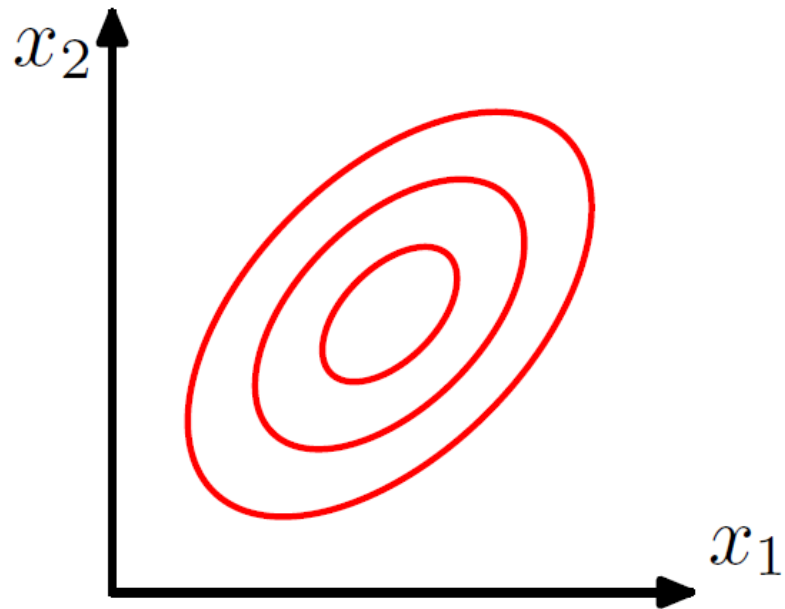
Positive Correlation
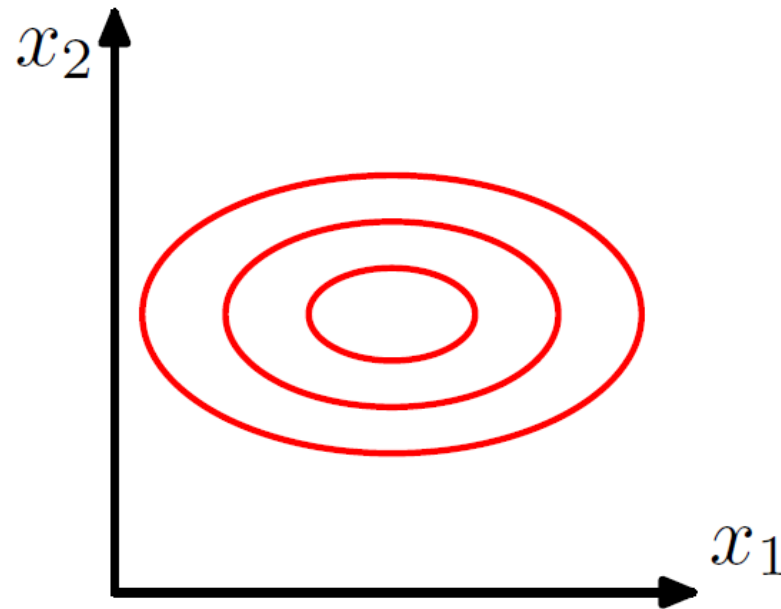
Uncorrelated
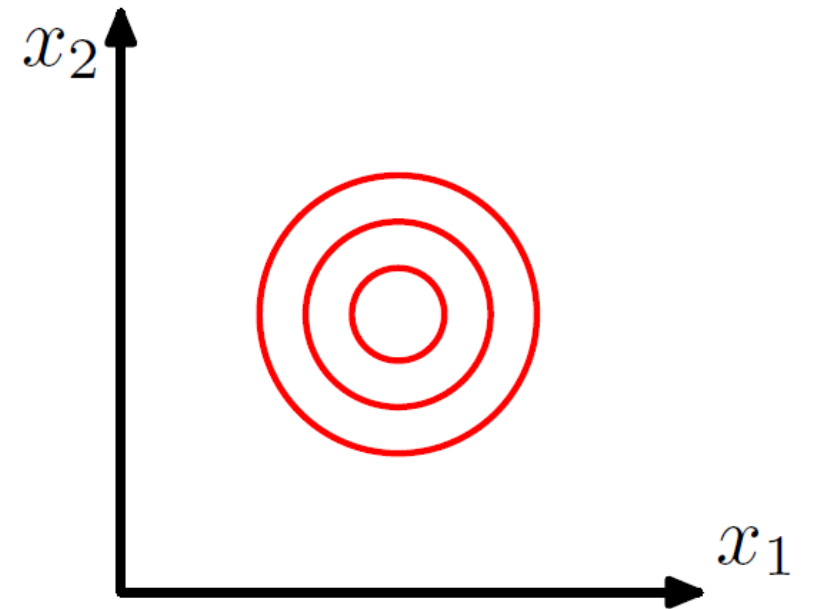
Uncorrelated and same variance (isotropic / spherical)
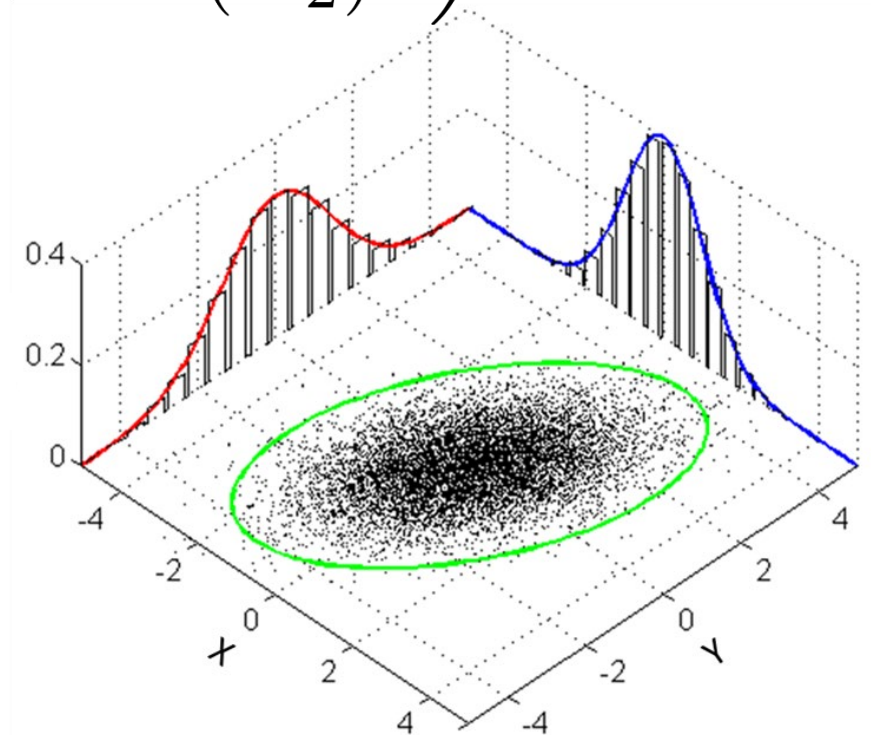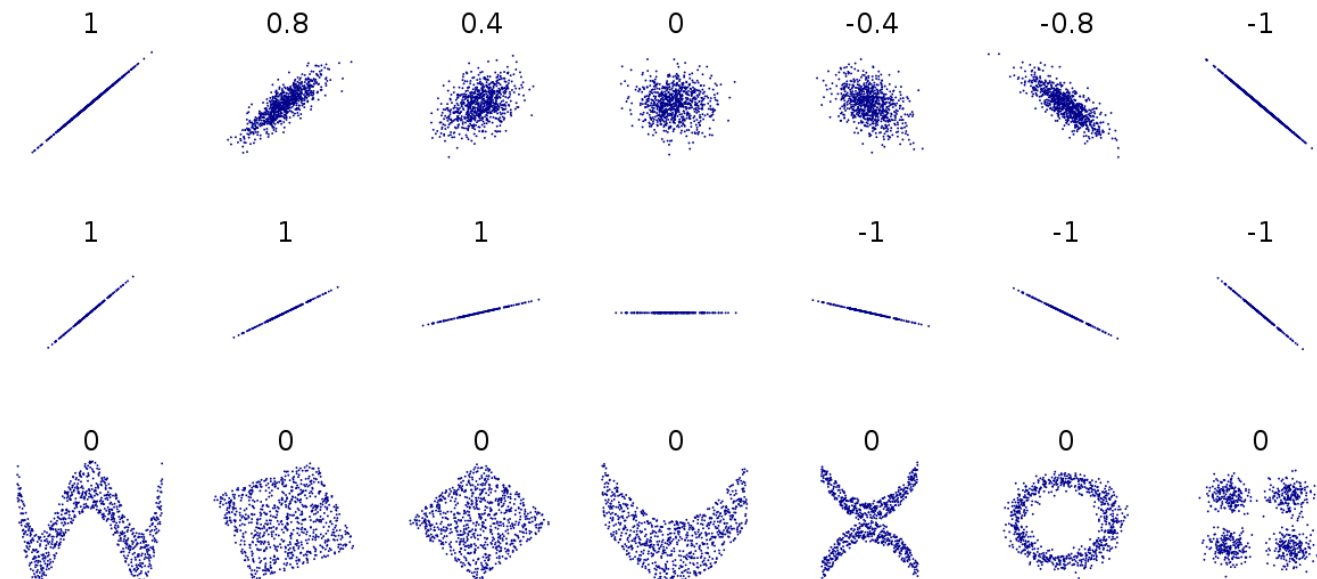
Covariance Matrix

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

Covariance Matrix

Marginal variance of just the RV X_1

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

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

Covariance Matrix

Correlation between X₁ and X₂

$$\Sigma = \mathrm{Cov}(X) = \begin{pmatrix} \mathrm{Var}(X_1) & \rho\sigma_{X_1}\sigma_{X_2} \\ \rho\sigma_{X_1}\sigma_{X_2} & \mathrm{Var}(X_2) \end{pmatrix}$$
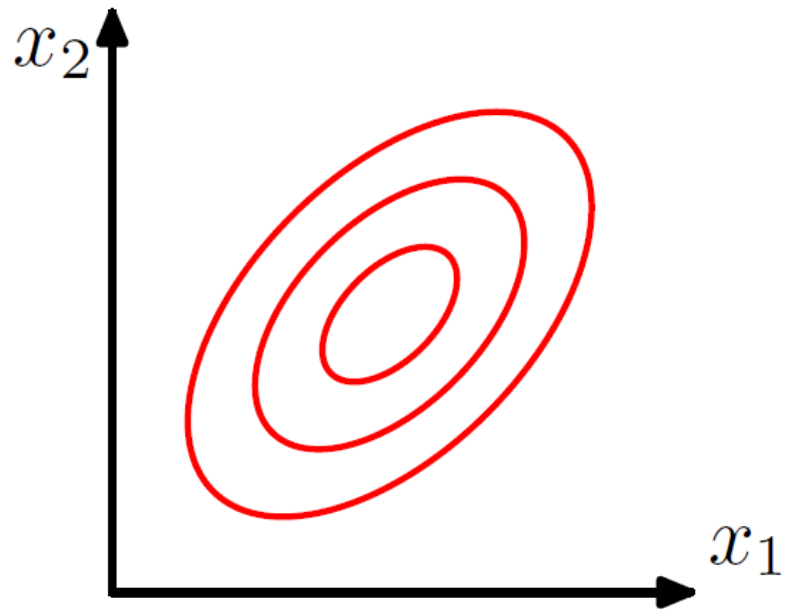
Recall, correlation is given by:

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

Captures *linear dependence of RVs*
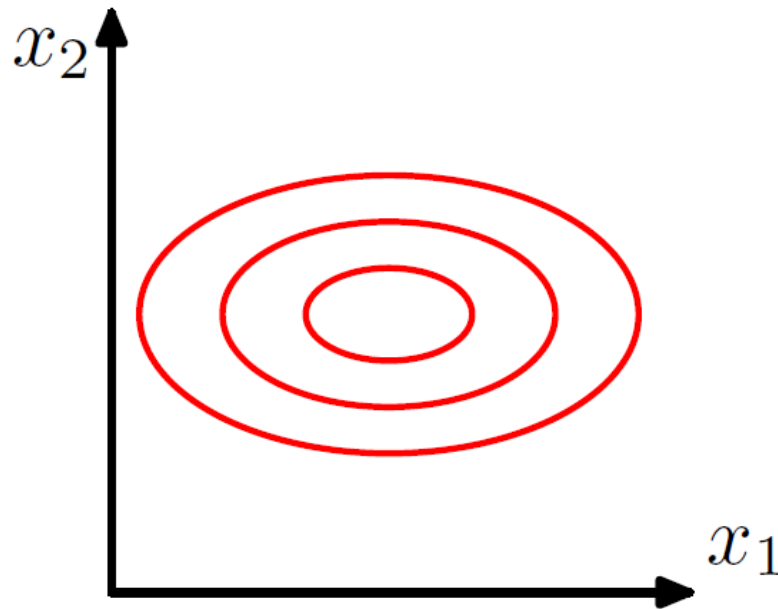
Covariance

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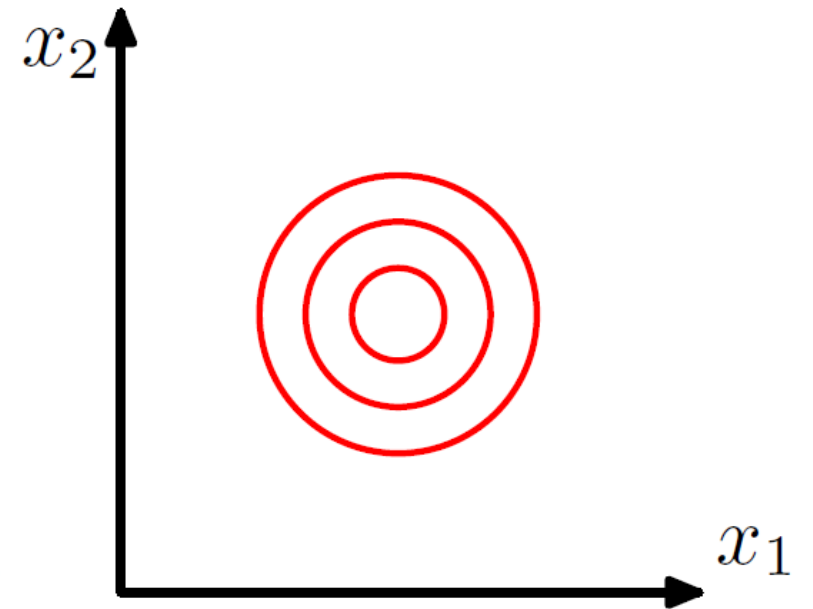

Positive Correlation

$\rho > 0$

Full matrix Σ

Uncorrelated

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

Isotropic / Spherical

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