# Gatsby Bridging Program

## Probability: Discrete Distributions

Cameron Stewart

Gatsby Computational Neuroscience Unit

# Table of contents

# Random Variables, Probability Mass Functions, and Cumulative Distribution Functions

# What is a Random Variable?

Random variables are functions which map the possible outcomes of an experiment to numerical values. In general, for random variable $X$ and sample space $\Omega$, we have that $X : \Omega \to \mathbb{R}$.

# What is a Random Variable?

Random variables are functions which map the possible outcomes of an experiment to numerical values. In general, for random variable $X$ and sample space $\Omega$, we have that $X : \Omega \to \mathbb{R}$.

## Example 1

Consider a bag containing 3 red balls (R) and 5 green balls (G). In our experiment, we are going to draw 2 balls from the bag, with replacement. The sample space is $\Omega = \{\text{RR, RG, GR, GG}\}$.

We are interested in determining the probabilities of drawing various numbers of red balls. To do this, we could start by defining a random variable $X : \Omega \to \{0, 1, 2\}$ such that

$$X(\omega) = \begin{cases} 0 & \text{if } \omega = \text{GG} \\ 1 & \text{if } \omega \in \{\text{RG, GR}\} \\ 2 & \text{if } \omega = \text{RR} \end{cases}.$$

# What is a Random Variable?

Random variables are often utilised without any explicit reference to the sample space. Instead of writing $X(\omega)$, we will typically just write $X$. Instead of writing $\mathbb{P}(E)$ for some event $E$, we typically write $\mathbb{P}(X \in A)$ for some set of numerical values $A$.

# What is a Random Variable?

Random variables are often utilised without any explicit reference to the sample space. Instead of writing $X(\omega)$, we will typically just write $X$. Instead of writing $\mathbb{P}(E)$ for some event $E$, we typically write $\mathbb{P}(X \in A)$ for some set of numerical values $A$.

**Example 1 Cont.**

What is the probability of drawing one red ball? We could express this as $\mathbb{P}(\{\text{RG}, \text{GR}\})$, but it is common and accepted notation to instead write $\mathbb{P}(X \in \{1\})$, or $\mathbb{P}(0 < X < 2)$, or most preferably in this case $\mathbb{P}(X = 1)$.

$$\begin{aligned}\mathbb{P}(X = 1) &= \frac{3}{8}\frac{5}{8} + \frac{5}{8}\frac{3}{8} \\ &= \frac{15}{32}.\end{aligned}$$

# The Chain Rule and Independence

In a previous lecture, we covered the following relationship for events $E$ and $F$:

$$\begin{aligned}\mathbb{P}(E \cap F) &= \mathbb{P}(E \mid F)\,\mathbb{P}(F) \\ &= \mathbb{P}(F \mid E)\,\mathbb{P}(E)\ .\end{aligned}$$

# The Chain Rule and Independence

In a previous lecture, we covered the following relationship for events $E$ and $F$:

$$\begin{aligned}\mathbb{P}(E \cap F) &= \mathbb{P}(E \mid F)\,\mathbb{P}(F) \\ &= \mathbb{P}(F \mid E)\,\mathbb{P}(E)\ .\end{aligned}$$

This also applies to random variables. Here we are looking at the probability of random variables $X$ and $Y$ taking values in sets $A$ and $B$ respectively:

## Chain Rule for Two Random Variables

$$\begin{aligned}\overbrace{\mathbb{P}(X \in A, Y \in B)}^{\text{Joint Prob.}} &= \overbrace{\mathbb{P}(X \in A \mid Y \in B)}^{\text{Conditional Prob.}}\,\overbrace{\mathbb{P}(Y \in B)}^{\text{Marginal Prob.}} \\ &= \mathbb{P}(Y \in B \mid X \in A)\,\mathbb{P}(X \in A)\end{aligned}$$

# The Chain Rule and Independence

In a previous lecture, we covered the following relationship for events $E$ and $F$:

$$\begin{aligned} \mathbb{P}(E \cap F) &= \mathbb{P}(E \mid F)\,\mathbb{P}(F) \\ &= \mathbb{P}(F \mid E)\,\mathbb{P}(E) \,. \end{aligned}$$

This also applies to random variables. Here we are looking at the probability of random variables $X$ and $Y$ taking values in sets $A$ and $B$ respectively:

## Chain Rule for Two Random Variables

$$\begin{aligned} \overbrace{\mathbb{P}(X \in A, Y \in B)}^{\text{Joint Prob.}} &= \overbrace{\mathbb{P}(X \in A \mid Y \in B)}^{\text{Conditional Prob.}}\,\overbrace{\mathbb{P}(Y \in B)}^{\text{Marginal Prob.}} \\ &= \mathbb{P}(Y \in B \mid X \in A)\,\mathbb{P}(X \in A) \end{aligned}$$

A future lecture will cover joint, conditional, and marginal distributions and the chain rule, independence, and marginalisation in more detail.

# The Chain Rule and Independence

In this lecture, we will only consider independent random variables. For two random variables to be independent, the realisation of one must have no effect on the distribution of the other. E.g. a coin flip is independent of the outcome of the previous coin flip.

# The Chain Rule and Independence

In this lecture, we will only consider independent random variables. For two random variables to be independent, the realisation of one must have no effect on the distribution of the other. E.g. a coin flip is independent of the outcome of the previous coin flip. Mathematically, we write this as

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\,\mathbb{P}(Y \in B) \ ,$$

taking note that independence implies

$$\mathbb{P}(X \in A \mid Y \in B) = \mathbb{P}(X \in A) \ .$$

# The Chain Rule and Independence

In this lecture, we will only consider independent random variables. For two random variables to be independent, the realisation of one must have no effect on the distribution of the other. E.g. a coin flip is independent of the outcome of the previous coin flip. Mathematically, we write this as

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\, \mathbb{P}(Y \in B) \ ,$$

taking note that independence implies

$$\mathbb{P}(X \in A \mid Y \in B) = \mathbb{P}(X \in A) \ .$$

For $N$ independent random variables, this generalises to:

### Independence of $N$ Random Variables

$$\mathbb{P}(X_1 \in A_1, \ldots, X_N \in A_N) = \prod_{n=1}^{N} \mathbb{P}(X_n \in A_n)$$

# The Chain Rule and Independence

**Example 1 Cont.**

In our red ball example, we could also use separate random variables for the outcomes of each draw from the bag. Let $X_1 \in \{0, 1\}$ be the random variable representing the first draw and $X_2 \in \{0, 1\}$ be the random variable representing the second draw (0 for a green ball and 1 for red ball). Then it is straightforward to see that $X = X_1 + X_2$.

# The Chain Rule and Independence

## Example 1 Cont.

In our red ball example, we could also use separate random variables for the outcomes of each draw from the bag. Let $X_1 \in \{0, 1\}$ be the random variable representing the first draw and $X_2 \in \{0, 1\}$ be the random variable representing the second draw (0 for a green ball and 1 for red ball). Then it is straightforward to see that $X = X_1 + X_2$.

Are $X_1$ and $X_2$ independent?

**Example 1 Cont.**

In our red ball example, we could also use separate random variables for the outcomes of each draw from the bag. Let $X_1 \in \{0, 1\}$ be the random variable representing the first draw and $X_2 \in \{0, 1\}$ be the random variable representing the second draw (0 for a green ball and 1 for red ball). Then it is straightforward to see that $X = X_1 + X_2$.

Are $X_1$ and $X_2$ independent? Yes, they are independent random variables, as the first draw has no effect on the second draw.

**Example 1 Cont.**

In our red ball example, we could also use separate random variables for the outcomes of each draw from the bag. Let $X_1 \in \{0, 1\}$ be the random variable representing the first draw and $X_2 \in \{0, 1\}$ be the random variable representing the second draw (0 for a green ball and 1 for red ball). Then it is straightforward to see that $X = X_1 + X_2$.

Are $X_1$ and $X_2$ independent? Yes, they are independent random variables, as the first draw has no effect on the second draw.

What if we had the same experimental setup, but without replacing the ball after the first draw?

**Example 1 Cont.**

In our red ball example, we could also use separate random variables for the outcomes of each draw from the bag. Let $X_1 \in \{0, 1\}$ be the random variable representing the first draw and $X_2 \in \{0, 1\}$ be the random variable representing the second draw (0 for a green ball and 1 for red ball). Then it is straightforward to see that $X = X_1 + X_2$.

Are $X_1$ and $X_2$ independent? Yes, they are independent random variables, as the first draw has no effect on the second draw.

What if we had the same experimental setup, but without replacing the ball after the first draw? In this case they are not independent, as the colour of the first drawn ball will dictate the probabilities of the second drawn ball.

# Discrete and Continuous Random Variables

Discrete random variables can take a countable number of values. For example:

- $X \in \{0, 1\}$
- $X \in \{0, 1, 2, \dots\}$
- $X$ representing the number of buses arriving within an hour.

# Discrete and Continuous Random Variables

Discrete random variables can take a countable number of values. For example:

- $X \in \{0, 1\}$
- $X \in \{0, 1, 2, \dots\}$
- $X$ representing the number of buses arriving within an hour.

Continuous random variables can take values in continuous ranges. For example:

- $X \in [0, 1]$
- $X \in \mathbb{R}$
- $X$ representing the waiting time until the next bus.

# Probability Mass Functions

Discrete distributions can be defined by their probability mass function (PMF). The PMF of random variable $X$ is often denoted by $f_X$ or $p_X$, and is defined as:

## Probability Mass Functions

$$f_X(x) = \mathbb{P}(X = x)$$

# Probability Mass Functions

Discrete distributions can be defined by their probability mass function (PMF). The PMF of random variable $X$ is often denoted by $f_X$ or $p_X$, and is defined as:

**Probability Mass Functions**

$$f_X\left(x\right) = \mathbb{P}\left(X = x\right)$$

$f_X$ is simply a function name. It is fine to use a different name, as long as it is clear how the function is defined. Occasionally, the same name is used for PMFs if it is clear from context how these are defined, but I'd advise against this practice for the sake of clarity. E.g. it is clearer to write $p_X\left(x\right)$ and $p_Y\left(y\right)$ than $p\left(x\right)$ and $p\left(y\right)$ if these correspond to 2 different PMFs.

# Probability Mass Functions

Probabilities can't be negative and must sum to 1 over the set of all possible values $\mathcal{X}$, so we have the following constraints:

$$f_X(x) \geq 0 \text{ for all } x$$

and

$$\sum_{x \in \mathcal{X}} f_X(x) = 1\,.$$

$\mathcal{X}$ is referred to as the support of $X$. $f_X(x) = 0$ for $x \notin \mathcal{X}$.

**Example 1 Cont.**

Let's continue with our red ball example. What is the PMF of $X$? First, we recognise that $\mathcal{X} = \{0, 1, 2\}$. You can verify for yourself that

$$f_X(x) = \begin{cases} \frac{25}{64} & \text{if } x = 0 \\ \frac{15}{32} & \text{if } x = 1 \\ \frac{9}{64} & \text{if } x = 2 \\ 0 & \text{otherwise} \end{cases},$$

and that this function satisfies the conditions for a PMF on the previous slide.

Often, we use a $\sim$ to mean ”is distributed as”. It is very common to see the notation

$$X \sim f_x \,,$$

which, in the discrete case, means that $X$ has the PMF $f_x$. Variations of this notation exist, but there should never be any ambiguity about the distribution of a random variable when using a $\sim$.

# Probability Mass Functions

Often, we use a $\sim$ to mean ”is distributed as”. It is very common to see the notation

$$X \sim f_x \, ,$$

which, in the discrete case, means that $X$ has the PMF $f_x$. Variations of this notation exist, but there should never be any ambiguity about the distribution of a random variable when using a $\sim$.

Sometimes you will see

$$X_1, \ldots, X_N \overset{\text{i.i.d.}}{\sim} f_x \, ,$$

which implies that all $N$ random variables are independent and identically distributed (i.i.d.) with PMF $f_x$.

A probability distribution can also be defined in terms of its cumulative distribution function (CDF). The CDF of a random variable $X$ is a monotonically increasing function defined as:

### Cumulative Distribution Functions

$$F_X(x) = \mathbb{P}(X \leq x)$$

# Cumulative Distribution Functions

A probability distribution can also be defined in terms of its cumulative distribution function (CDF). The CDF of a random variable $X$ is a monotonically increasing function defined as:

## Cumulative Distribution Functions

$$F_X\left(x\right) = \mathbb{P}\left(X \leq x\right)$$

For discrete random variables, we can relate this definition to the PMF as follows:

$$F_X\left(x\right) = \sum_{y \leq x} f_X\left(y\right) \; .$$

# Cumulative Distribution Functions

As probabilities must sum to 1 over the support, we have the following emergent properties for CDFs:

$$\lim_{x \to -\infty} F_X\left(x\right) = 0$$
$$\lim_{x \to \infty} F_X\left(x\right) = 1$$

# Cumulative Distribution Functions

As probabilities must sum to 1 over the support, we have the following emergent properties for CDFs:

$$\lim_{x \to -\infty} F_X(x) = 0$$
$$\lim_{x \to \infty} F_X(x) = 1$$

We can also see that the following is true:

$$\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a) \; .$$

## Example 1 Cont.

Back to the red ball example. What is the CDF of $X$? Previously, we found that

$$f_X(x) = \begin{cases} \frac{25}{64} & \text{if } x = 0 \\ \frac{15}{32} & \text{if } x = 1 \\ \frac{9}{64} & \text{if } x = 2 \\ 0 & \text{otherwise} \end{cases}.$$

Hence, the CDF is

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{25}{64} & \text{if } 0 \leq x < 1 \\ \frac{55}{64} & \text{if } 1 \leq x < 2 \\ 1 & \text{if } x \geq 2 \end{cases}.$$

# PDF and CDF Plots

## Example 1 Cont.

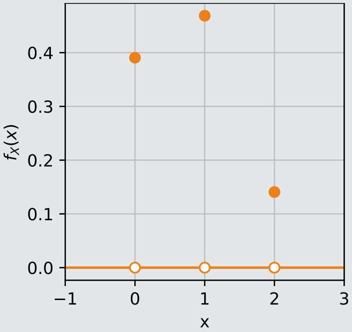


**Figure 1:** Probability Mass Function

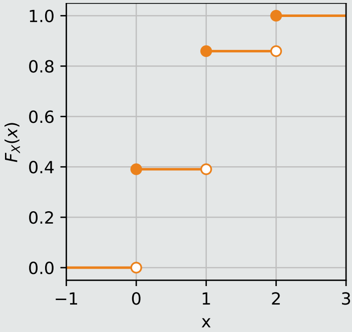


**Figure 2:** Cumulative Distribution Function

# Sampling

Similarly to how we refer to realisations of random variables, we can also talk about samples of distributions. Sampling from a discrete distribution with PMF $f_x$ simply implies that we randomly generate a number $x$ with probability $f_x(x)$.

# Sampling

Similarly to how we refer to realisations of random variables, we can also talk about samples of distributions. Sampling from a discrete distribution with PMF $f_x$ simply implies that we randomly generate a number $x$ with probability $f_x(x)$.

To be entirely correct, drawing $N$ samples from $f_x$ involves running $N$ random experiments with $N$ associated random variables $X_1, \ldots, X_N \overset{\text{i.i.d.}}{\sim} f_x$ and generating $N$ realisations $x_1, \ldots, x_N$.

# Sampling

Similarly to how we refer to realisations of random variables, we can also talk about samples of distributions. Sampling from a discrete distribution with PMF $f_x$ simply implies that we randomly generate a number $x$ with probability $f_x(x)$.

To be entirely correct, drawing $N$ samples from $f_x$ involves running $N$ random experiments with $N$ associated random variables $X_1, \ldots, X_N \overset{\text{i.i.d.}}{\sim} f_x$ and generating $N$ realisations $x_1, \ldots, x_N$.

As $N \to \infty$, we will observe that

$$\frac{\sum_{n=1}^{N} [x_n = x]}{N} \to f_x(x)$$

Intermission

# Expectation and Variance

# Expectations

Suppose we want to know the average value of a function $g(x)$ evaluated at samples from a distribution. More precisely, we wish to draw realisations of $X_1, \ldots, X_N \overset{\text{i.i.d.}}{\sim} f_x$ and then compute the mean of $\{g(x_1), \ldots, g(x_N)\}$. As $N$ increases, this will converge to a value which we call the expectation (or expected value). This theorem is referred to as the law of large numbers.

# Expectations

Suppose we want to know the average value of a function $g(x)$ evaluated at samples from a distribution. More precisely, we wish to draw realisations of $X_1, \ldots, X_N \overset{\text{i.i.d.}}{\sim} f_x$ and then compute the mean of $\{g(x_1), \ldots, g(x_N)\}$. As $N$ increases, this will converge to a value which we call the expectation (or expected value). This theorem is referred to as the law of large numbers.

We can estimate the expectation for finite $N$ as follows:

## Empirical Estimates of Expectations

$$\mathop{\mathbb{E}}_{X \sim f_X}[g(X)] \approx \frac{1}{N} \sum_{n=1}^{N} g(x_n) \qquad \text{for large } N$$

# Expectations

If $X$ is discrete with support $\mathcal{X}$ and PMF $f_X$, we can also define this expectation as a weighted average:

## Expectations on Discrete Distributions

$$\mathop{\mathbb{E}}_{X \sim f_X} \left[ g\left(X\right) \right] = \sum_{x \in \mathcal{X}} f_X\left(x\right) g\left(x\right)$$

# Expectations

If $X$ is discrete with support $\mathcal{X}$ and PMF $f_X$, we can also define this expectation as a weighted average:

## Expectations on Discrete Distributions

$$\underset{X \sim f_X}{\mathbb{E}} \left[ g \left( X \right) \right] = \sum_{x \in \mathcal{X}} f_X \left( x \right) g \left( x \right)$$

If the distribution on which the expectation is taken is clear from context, the text under the $\mathbb{E}$ is often omitted.

# Expectations

If $X$ is discrete with support $\mathcal{X}$ and PMF $f_X$, we can also define this expectation as a weighted average:

## Expectations on Discrete Distributions

$$\underset{X \sim f_X}{\mathbb{E}} \left[ g \left( X \right) \right] = \sum_{x \in \mathcal{X}} f_X \left( x \right) g \left( x \right)$$

If the distribution on which the expectation is taken is clear from context, the text under the $\mathbb{E}$ is often omitted.

When talking about the mean of a distribution defined by $f_X$, this specifically refers to the value given by

$$\underset{X \sim f_X}{\mathbb{E}} \left[ X \right] .$$

# Properties of Expectations

- Linearity:

$$\mathbb{E}\left[ag\left(X\right)+bh\left(X\right)\right]=a\mathbb{E}\left[g\left(X\right)\right]+b\mathbb{E}\left[h\left(X\right)\right]\ ,$$

  for constants $a, b$ and functions $g, h$.

# Properties of Expectations

- Linearity:

$$\mathbb{E}\left[ag\left(X\right) + bh\left(X\right)\right] = a\mathbb{E}\left[g\left(X\right)\right] + b\mathbb{E}\left[h\left(X\right)\right] \ ,$$

  for constants $a, b$ and functions $g, h$.

- Non-multiplicativity: In general,

$$\mathbb{E}\left[g\left(X\right)h\left(X\right)\right] \neq \mathbb{E}\left[g\left(X\right)\right]\mathbb{E}\left[h\left(X\right)\right] \ .$$

# Properties of Expectations

- Linearity:

$$\mathbb{E}\left[ag\left(X\right)+bh\left(X\right)\right]=a\mathbb{E}\left[g\left(X\right)\right]+b\mathbb{E}\left[h\left(X\right)\right]\ ,$$

  for constants $a, b$ and functions $g, h$.
- Non-multiplicativity: In general,

$$\mathbb{E}\left[g\left(X\right)h\left(X\right)\right]\neq\mathbb{E}\left[g\left(X\right)\right]\mathbb{E}\left[h\left(X\right)\right]\ .$$

- $\mathbb{E}\left[a\right]=a$ for constant $a$. Hence, $\mathbb{E}\left[\mathbb{E}\left[X\right]\right]=\mathbb{E}\left[X\right]$.

## Properties of Expectations

- Linearity:

$$\mathbb{E}\left[ag\left(X\right)+bh\left(X\right)\right]=a\mathbb{E}\left[g\left(X\right)\right]+b\mathbb{E}\left[h\left(X\right)\right] ,$$

  for constants $a, b$ and functions $g, h$.

- Non-multiplicativity: In general,

$$\mathbb{E}\left[g\left(X\right)h\left(X\right)\right]\neq\mathbb{E}\left[g\left(X\right)\right]\mathbb{E}\left[h\left(X\right)\right] .$$

- $\mathbb{E}\left[a\right]=a$ for constant $a$. Hence, $\mathbb{E}\left[\mathbb{E}\left[X\right]\right]=\mathbb{E}\left[X\right]$.

Additional properties exist for multivariate distributions.

## Example 2

Suppose we roll a six-sided die. Let $X \in \{1, 2, 3, 4, 5, 6\}$ represent this roll. What is the expected value of $X$?

## Example 2

Suppose we roll a six-sided die. Let $X \in \{1, 2, 3, 4, 5, 6\}$ represent this roll. What is the expected value of $X$?

Firstly, we define the support as $\mathcal{X} \in \{1, 2, 3, 4, 5, 6\}$ and the PMF $f_X$ as

$$f_X(x) = \begin{cases} \frac{1}{6} & \text{if } x \in \mathcal{X} \\ 0 & \text{otherwise} \end{cases}.$$

## Example 2

Suppose we roll a six-sided die. Let $X \in \{1, 2, 3, 4, 5, 6\}$ represent this roll. What is the expected value of $X$?

Firstly, we define the support as $\mathcal{X} \in \{1, 2, 3, 4, 5, 6\}$ and the PMF $f_X$ as

$$f_X(x) = \begin{cases} \frac{1}{6} & \text{if } x \in \mathcal{X} \\ 0 & \text{otherwise} \end{cases}.$$

Then the expected value of $X$ is given by

$$\begin{aligned} \mathbb{E}[X] &= \sum_{x \in \mathcal{X}} f_X(x)\, x \\ &= \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 3 + \frac{1}{6} \cdot 4 + \frac{1}{6} \cdot 5 + \frac{1}{6} \cdot 6 \\ &= 3.5 \end{aligned}$$

# Expectations

Why care about expectations?

- A huge number of statistical estimation problems, where we try to estimate model parameters given some data, can be written in terms of expectations.

# Expectations

Why care about expectations?

- A huge number of statistical estimation problems, where we try to estimate model parameters given some data, can be written in terms of expectations.
- This includes machine learning problems. Training a machine learning model involves minimising a loss function. The loss function and its gradient is very often written in terms of expectations.

# Expectations

Why care about expectations?

- A huge number of statistical estimation problems, where we try to estimate model parameters given some data, can be written in terms of expectations.
- This includes machine learning problems. Training a machine learning model involves minimising a loss function. The loss function and its gradient is very often written in terms of expectations.
- We can estimate these expectations simply by drawing samples!

# Variance

How about if we want to know how "spread out" a distribution is? A way of measuring this is with the variance. It measures the expected deviation from the mean.

## Variance of a Random Variable

$$\begin{aligned}
\text{Var}\left(X\right) &= \mathbb{E}\left[\left(X - \mathbb{E}\left[X\right]\right)^2\right] \\
&= \mathbb{E}\left[X^2 - 2X\mathbb{E}\left[X\right] + \mathbb{E}\left[X\right]^2\right] \\
&= \mathbb{E}\left[X^2\right] - 2\mathbb{E}\left[X\right]\mathbb{E}\left[X\right] + \mathbb{E}\left[X\right]^2 \\
&= \mathbb{E}\left[X^2\right] - \mathbb{E}\left[X\right]^2
\end{aligned}$$

# Properties of Variance

- $\text{Var}\left(aX + b\right) = a^2\text{Var}\left(X\right)$ for constants $a, b$.

# Properties of Variance

- $\text{Var}(aX + b) = a^2 \text{Var}(X)$ for constants $a, b$.
- $\text{Var}(a) = 0$ for constant $a$.

# Properties of Variance

- $\text{Var}(aX+b) = a^2\text{Var}(X)$ for constants $a, b$.
- $\text{Var}(a) = 0$ for constant $a$.
- $\text{Var}(X) \geq 0$.

# Properties of Variance

- $\text{Var}(aX + b) = a^2 \text{Var}(X)$ for constants $a, b$.
- $\text{Var}(a) = 0$ for constant $a$.
- $\text{Var}(X) \geq 0$.

Additional properties exist for multivariate distributions.

### Example 2 Cont.

Continuing with the die roll example. What is the variance of $X$?

## Example 2 Cont.

Continuing with the die roll example. What is the variance of $X$?

We will compute the variance of $X$ in two separate ways. Method 1:

$$\begin{aligned}
\text{Var}(X) &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\
&= \mathbb{E}\left[(X - 3.5)^2\right] \\
&= \sum_{x \in \mathcal{X}} f_X(x)(x - 3.5)^2 \\
&= \frac{1}{6}(1 - 3.5)^2 + \frac{1}{6}(2 - 3.5)^2 + \frac{1}{6}(3 - 3.5)^2 + \frac{1}{6}(4 - 3.5)^2 \\
&\quad + \frac{1}{6}(5 - 3.5)^2 + \frac{1}{6}(6 - 3.5)^2 \\
&= 2.91\dot{6}
\end{aligned}$$

# Variance

## Example 2 Cont.

Method 2:

$$\begin{aligned}\mathsf{Var}(X) &= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 \\ &= \sum_{x\in\mathcal{X}} f_X(x)\, x^2 - 3.5^2 \\ &= \frac{1}{6}\cdot 1^2 + \frac{1}{6}\cdot 2^2 + \frac{1}{6}\cdot 3^2 + \frac{1}{6}\cdot 4^2 + \frac{1}{6}\cdot 5^2 + \frac{1}{6}\cdot 6^2 - 3.5^2 \\ &= 2.91\dot{6}\end{aligned}$$

# Notation

It is typical to see the notations $\mu$, $\sigma^2$, and $\sigma$ used for the mean, variance, and standard deviation of a distribution respectively. Clearly, the standard deviation is just the square root of the variance. In other words, we have:

$$\mu = \mathbb{E}[X]$$
$$\sigma^2 = \mathsf{Var}(X) \ .$$

Intermission

# Common Discrete Distributions

Intermission

# Introduction to Stochastic Processes