Gatsby Bridging Program

Probability: Discrete Distributions

Cameron Stewart

Gatsby Computational Neuroscience Unit

Today's Lecture Topics

1. Random Variables, Probability Mass Functions, and Cumulative Distribution Functions

2. Expectation and Variance

3. Important Discrete Distributions

4. Introduction to Stochastic Processes

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 Ω, 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 Ω, 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 $\Omega = \{\mathrm{RR}, \mathrm{RG}, \mathrm{GR}, \mathrm{GG}\}$.

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

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

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 $\mathbb{P}(\{RG, GR\})$, 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 $\mathbb{P}(X = 1)$.

$$\begin{aligned}
\mathbb{P}(X = 1) &= \mathbb{P}(\{RG\} \cup \{GR\}) \\
&= \mathbb{P}(\{RG\}) + \mathbb{P}(\{GR\}) - \mathbb{P}(\{RG\} \cap \{GR\}) \\
&= \frac{3}{8}\frac{5}{8} + \frac{5}{8}\frac{3}{8} - 0 \\
&= \frac{15}{32}.
\end{aligned}$$

The Chain Rule and Independence

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

$$\mathbb{P}(E \cap F) = \mathbb{P}(E \mid F) \mathbb{P}(F)$$
$$= \mathbb{P}(F \mid E) \mathbb{P}(E) .$$

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

$$\mathbb{P}\left(E \cap F\right) = \mathbb{P}\left(E \mid F\right)\mathbb{P}\left(F\right)$$
$$= \mathbb{P}\left(F \mid E\right)\mathbb{P}\left(E\right).$$

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

$$\overbrace{\mathbb{P}\left(X \in A, Y \in B\right)}^{\substack{\text{Joint} \\ \text{Probability}}} = \overbrace{\mathbb{P}\left(X \in A \mid Y \in B\right)}^{\substack{\text{Conditional} \\ \text{Probability}}}\overbrace{\mathbb{P}\left(Y \in B\right)}^{\substack{\text{Marginal} \\ \text{Probability}}}$$
$$= \mathbb{P}\left(Y \in B \mid X \in A\right)\mathbb{P}\left(X \in A\right)$$

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

$$\mathbb{P}(E \cap F) = \mathbb{P}(E \mid F)\mathbb{P}(F)$$
$$= \mathbb{P}(F \mid E)\mathbb{P}(E).$$

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

$$\overbrace{\mathbb{P}(X \in A, Y \in B)}^{\substack{\text{Joint} \\ \text{Probability}}} = \overbrace{\mathbb{P}(X \in A \mid Y \in B)}^{\substack{\text{Conditional} \\ \text{Probability}}} \overbrace{\mathbb{P}(Y \in B)}^{\substack{\text{Marginal} \\ \text{Probability}}}$$

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

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

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

taking note that independence implies

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

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

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

taking note that independence implies

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

For n independent random variables, this generalises to:

Independence of n Random Variables

$$\mathbb{P}\left(X_1 \in A_1, \ldots, X_n \in A_n\right) = \prod_{i=1}^{n} \mathbb{P}\left(X_i \in A_i\right)$$

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 $X = X_1 + X_2$.

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.

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? 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.

A hint for the problem set questions:

Whilst we won't be dealing with dependent random variables today, some problem set questions will still require you to work with conditional probabilities. Take note that

$$\mathbb{P}\left(X \in A \mid X \in B\right) = \frac{\mathbb{P}\left(X \in A, X \in B\right)}{\mathbb{P}\left(X \in B\right)}.$$

There is nothing special about this. We are simply looking at the probability of one event conditioned on the occurrence of another.

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

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

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

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

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 $p_X(x)$ and $p_Y(y)$ than $p(x)$ and $p(y)$ 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

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

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:

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

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

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

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

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:

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

Cumulative Distribution Functions

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

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{25}{64} & \text{if } 0 \le x < 1 \\ \frac{55}{64} & \text{if } 1 \le x < 2 \\ 1 & \text{if } x \ge 2 \end{cases}.$$
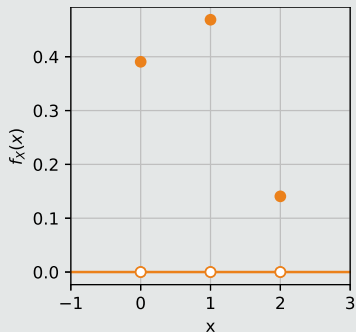
PDF and CDF Plots

Example 1 Cont.

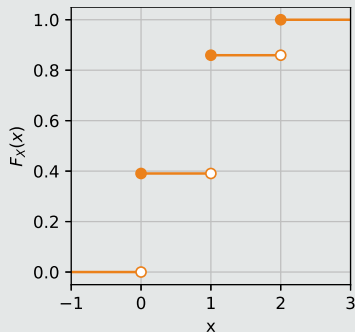Figure 1: Probability Mass Function

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

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

Intermission

Expectation and Variance

Expectations

Suppose we want to know the average value of a function $g\left(x\right)$ 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 $\{g\left(x_1\right),\ldots,g\left(x_n\right)\}$. 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.

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

$$\underset{X \sim f_X}{\mathbb{E}}[g(X)] \approx \frac{1}{n} \sum_{i=1}^{n} g(x_i) \qquad \text{for large } n$$

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

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

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

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

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

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

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

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

$$\mathop{\mathbb{E}}_{X \sim f_X} [X] \ .$$

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.

Expectations

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?

Expectations

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

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

Expectations

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.

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.

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!

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

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

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

Variance

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:

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

Example 2 Cont.

Method 2:

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

It is typical to see the notations μ, σ^2, and σ 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:

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

Intermission

Important Discrete Distributions

Why Study Specific Distributions?

- Many real-world scenarios can be modelled with a few simple distributions. E.g. modelling the chance of success, the number of successes in a finite number of trials, the number of failures until a success, the waiting time until an event happens, and the uncertainty in an estimate after multiple measurements.

Why Study Specific Distributions?

- Many real-world scenarios can be modelled with a few simple distributions. E.g. modelling the chance of success, the number of successes in a finite number of trials, the number of failures until a success, the waiting time until an event happens, and the uncertainty in an estimate after multiple measurements.

- A large number of these simple distributions have very nice mathematical properties, including important relationships between them.

- Many real-world scenarios can be modelled with a few simple distributions. E.g. modelling the chance of success, the number of successes in a finite number of trials, the number of failures until a success, the waiting time until an event happens, and the uncertainty in an estimate after multiple measurements.

- A large number of these simple distributions have very nice mathematical properties, including important relationships between them.

After the lecture on continuous distributions, I'd recommend doing some exploring of these distributions and their relationships.

Functions of Random Variables

The next few slides will deal with simple functions of random variables. We have an entire lecture later on which will cover this topic in much more depth. For now, all you need to know are the following:

- Functions of random variables are also random variables. E.g. if g is a function and X, Y are random variables, then $Z = g(X, Y)$ is also a random variable. The distribution of Z can often be pretty complicated!

Functions of Random Variables

The next few slides will deal with simple functions of random variables. We have an entire lecture later on which will cover this topic in much more depth. For now, all you need to know are the following:

- Functions of random variables are also random variables. E.g. if g is a function and X, Y are random variables, then $Z = g(X, Y)$ is also a random variable. The distribution of Z can often be pretty complicated!

- The realisation of Z is given by the realisations of X and Y passed through g. I.e. $z = g(x, y)$. So sampling can often be pretty easy!

Functions of Random Variables

The next few slides will deal with simple functions of random variables. We have an entire lecture later on which will cover this topic in much more depth. For now, all you need to know are the following:

- Functions of random variables are also random variables. E.g. if g is a function and X, Y are random variables, then $Z = g(X, Y)$ is also a random variable. The distribution of Z can often be pretty complicated!
- The realisation of Z is given by the realisations of X and Y passed through g. I.e. $z = g(x, y)$. So sampling can often be pretty easy!

A hint for the problem set questions:

If X and Y are discrete and independent, then $Z = X + Y$ has the PMF

$$p_Z(z) = \sum_{x \in \mathcal{X}} p_X(x) \, p_Y(z - x) \ .$$

You may use this without proof. A summation of this form is called a convolution.

Discrete Uniform Distribution

If a random variable takes any integer between a and b inclusive with equal probability, then it follows a discrete uniform distribution.

Discrete Uniform Distribution

Discrete uniform random variable $X \sim \mathcal{U}\{a, b\}$ has the following properties:

Parameters	$a, b \in \mathbb{Z} \mid b \geq a$
Support	$\mathcal{X} = \{a, \ldots, b\}$
PMF	$f_X(x) = \frac{1}{b-a+1}$ $\quad$ for $x \in \mathcal{X}$
Mean	$\mathbb{E}[X] = \frac{a+b}{2}$
Variance	$\mathrm{Var}(X) = \frac{(b-a+1)^2-1}{12}$

Example 2 Cont.

Back to the die roll example. Is X uniformly distributed?

Example 2 Cont.

Back to the die roll example. Is X uniformly distributed?

Yes, it is uniformly distributed with $a = 1$ and $b = 6$.

Discrete Uniform Distribution

Example 2 Cont.

Back to the die roll example. Is X uniformly distributed?

Yes, it is uniformly distributed with $a = 1$ and $b = 6$. We can easily confirm our previous calculations with the formulas on the last slide:

$$\mathbb{E}\left[X\right] = \frac{1 + 6}{2}$$
$$= 3.5$$

and

$$\text{Var}\left(X\right) = \frac{(6 - 1 + 1)^2 - 1}{12}$$
$$= 2.91\dot{6}\,.$$

If a random variable takes the value 1 with probability p, and 0 otherwise, then it follows a Bernoulli distribution. It models the probability of "success".

Bernoulli Distribution

Bernoulli random variable $X \sim \text{Bern}(p)$ has the following properties:

Parameters	$p \in [0, 1]$
Support	$\mathcal{X} = \{0, 1\}$
PMF	$f_X(x) = p^x (1-p)^{1-x}$ for $x \in \mathcal{X}$
Mean	$\mathbb{E}[X] = p$
Variance	$\text{Var}(X) = p(1-p)$

If $X_1, \ldots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Bern}(p)$ and $X = X_1 + \cdots + X_n$, then X is a binomial random variable. It models the number of successes in n Bernoulli trials.

Binomial Distribution

Binomial random variable $X \sim \text{Bin}(n, p)$ has the following properties:

Parameters	$n \in \{0, 1, 2 \ldots\}$ and $p \in [0, 1]$
Support	$\mathcal{X} = \{0, \ldots, n\}$
PMF	$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x}$ for $x \in \mathcal{X}$
Mean	$\mathbb{E}[X] = np$
Variance	$\text{Var}(X) = np(1-p)$

Clearly, if $X \sim \text{Bin}(1, p)$ then X is also Bernoulli distributed.

Why is the binomial PMF given by $\binom{n}{x} p^x (1-p)^{n-x}$?

Why is the binomial PMF given by $\binom{n}{x}p^x\left(1-p\right)^{n-x}$?

In n trials we have x successes and $n-x$ failures. The $p^x\left(1-p\right)^{n-x}$ term gives the probability of observing the successes at specific times. E.g. the probability of running 3 trials and having only trials 1 and 3 produce successes.

Binomial Distribution

Why is the binomial PMF given by $\binom{n}{x} p^x (1-p)^{n-x}$?

In n trials we have x successes and $n-x$ failures. The $p^x (1-p)^{n-x}$ term gives the probability of observing the successes at specific times. E.g. the probability of running 3 trials and having only trials 1 and 3 produce successes.

More precisely,

$$
\begin{aligned}
\mathbb{P}\left(X_1 = x_1, \ldots, X_n = x_n\right) &= \prod_{i=1}^{n} \mathbb{P}\left(X_i = x_i\right) \\
&= \prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i} \\
&= p^{x_1 + \cdots + x_n} (1-p)^{n-(x_1 + \cdots + x_n)} .
\end{aligned}
$$

However, we have to consider all possible ways in which x successes can occur. If we run 3 trials and get 2 successes, we have three different possibilities:

- Trials 1 and 2 were successes.

Binomial Distribution

However, we have to consider all possible ways in which x successes can occur. If we run 3 trials and get 2 successes, we have three different possibilities:

- Trials 1 and 2 were successes.
- Trials 1 and 3 were successes.

Binomial Distribution

However, we have to consider all possible ways in which x successes can occur. If we run 3 trials and get 2 successes, we have three different possibilities:

- Trials 1 and 2 were successes.
- Trials 1 and 3 were successes.
- Trials 2 and 3 were successes.

Binomial Distribution

However, we have to consider all possible ways in which x successes can occur. If we run 3 trials and get 2 successes, we have three different possibilities:

- Trials 1 and 2 were successes.
- Trials 1 and 3 were successes.
- Trials 2 and 3 were successes.

In general, we have $\binom{n}{x}$ ways in which n trials can produce x successes. Adding the probabilities from these cases together gives the binomial PMF.

Binomial Distribution

Example 3

An avid card collector wishes to purchase 10 packs of trading cards in hopes of finding a rare card. A rare card is known to exist in 1% of card packs. No more than 1 rare card is ever in a pack. What is the probability the collector receives at least 1 rare card?

Binomial Distribution

Example 3

An avid card collector wishes to purchase 10 packs of trading cards in hopes of finding a rare card. A rare card is known to exist in 1% of card packs. No more than 1 rare card is ever in a pack. What is the probability the collector receives at least 1 rare card?

We can model the distribution of received rare cards with a binomial random variable. Let $X \sim \mathrm{Bin}\left(10, 0.01\right)$ with PMF f_X. Then the probability of receiving at least 1 rare card is given by:

$$\begin{aligned}
\mathbb{P}\left(X \geq 1\right) &= 1 - \mathbb{P}\left(X = 0\right) \\
&= 1 - f_X\left(0\right) \\
&= 1 - \binom{10}{0} 0.01^0 \left(1 - 0.01\right)^{10-0} \\
&\approx 0.0956\,.
\end{aligned}$$

If $X_1, X_2, \ldots \overset{\text{i.i.d.}}{\sim} \text{Bern}(p)$ and $X = \min\{i \in \mathbb{N} \mid X_i = 1\}$, then X is a geometric random variable. It models the number of sequential Bernoulli trials performed until getting a success.

Geometric Distribution

Geometric random variable $X \sim \text{Geo}(p)$ has the following properties:

Parameters	$p \in [0, 1]$
Support	$\mathcal{X} = \mathbb{N}$
PMF	$f_X(x) = (1-p)^{x-1} p \qquad \text{for } x \in \mathcal{X}$
Mean	$\mathbb{E}[X] = \frac{1}{p}$
Variance	$\text{Var}(X) = \frac{1-p}{p^2}$

Example 3 Cont.

The card collector was disappointed to find no rare cards in the 10 opened packs. They decide to continue purchasing packs until they find a rare card. What is the expected number of packs they will open from this point onward?

Geometric Distribution

Example 3 Cont.

The card collector was disappointed to find no rare cards in the 10 opened packs. They decide to continue purchasing packs until they find a rare card. What is the expected number of packs they will open from this point onward?

The geometric distribution is perfect for modelling this problem. Let $Y \sim \mathrm{Geo}\,(0.01)$. Then the expected number of packs opened is simply

$$\mathbb{E}\,[Y] = \frac{1}{0.01}$$
$$= 100\,,$$

which makes sense intuitively.

Geometric Distribution

Example 3 Cont.

The card collector was disappointed to find no rare cards in the 10 opened packs. They decide to continue purchasing packs until they find a rare card. What is the expected number of packs they will open from this point onward?

The geometric distribution is perfect for modelling this problem. Let $Y \sim \mathrm{Geo}\,(0.01)$. Then the expected number of packs opened is simply

$$\mathbb{E}\,[Y] = \frac{1}{0.01}$$
$$= 100\,,$$

which makes sense intuitively.

The collector opens 100 packs, but still doesn't find a rare card. What is the expected number of packs they will open from this point onward? Still 100; the geometric distribution is memoryless.

If $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bern}\left(\frac{\lambda}{n}\right)$, $X = X_1 + \cdots + X_n$, and $n \to \infty$, then X is a Poisson random variable. It models the number of independently occurring events in a fixed interval of time or space, where λ is the mean rate of occurrence.

Poisson Distribution

Poisson random variable $X \sim \text{Pois}(\lambda)$ has the following properties:

Parameters	$\lambda \in (0, \infty)$
Support	$\mathcal{X} = \mathbb{N}_0$
PMF	$f_X(x) = \frac{\lambda^x e^{-\lambda}}{x!} \quad \text{for } x \in \mathcal{X}$
Mean	$\mathbb{E}[X] = \lambda$
Variance	$\text{Var}(X) = \lambda$

Clearly, we can also say if $X \sim \text{Bin}\left(n, \frac{\lambda}{n}\right)$ and $n \to \infty$ then X is Poisson distributed. The proof of this is a problem set question.

Poisson Distribution

Example 4

Consider a single neuron in isolation, which we inject with some current. We find the timings of the spikes to be highly random and almost independent of one another. The timings aren't entirely independent (partly due to a refractory period after a spike), but we consider this a negligible factor. The mean firing rate is 2 Hz. What is the probability of observing at least 1 spike in a 3 second window?

Poisson Distribution

Example 4

Given our assumptions, we can model the number of spikes in this window with a Poisson distribution. Let $X \sim \mathrm{Pois}\,(6)$ with PMF f_X, as the average number of spikes occurring in this window is 3×2. Then the probability of observing at least 1 spike is:

$$\begin{aligned}
\mathbb{P}\,(X \geq 1) &= 1 - \mathbb{P}\,(X = 0) \\
&= 1 - f_X\,(0) \\
&= 1 - \frac{6^0 e^{-6}}{0!} \\
&\approx 0.998\,.
\end{aligned}$$

Intermission

Introduction to Stochastic Processes

What is a Stochastic Process?

A stochastic process describes how a system with some underlying randomness evolves over space or time. Mathematically, we define this by a set of random variables

$$\{X(t) \mid t \in T\}$$

or more precisely

$$\{X(t, \omega) \mid t \in T\},$$

where T is referred to as the index set.

What is a Stochastic Process?

A stochastic process describes how a system with some underlying randomness evolves over space or time. Mathematically, we define this by a set of random variables

$$\{X(t) \mid t \in T\}$$

or more precisely

$$\{X(t, \omega) \mid t \in T\},$$

where T is referred to as the index set.

Very often, T is chosen to represent points in time. In discrete-time, we may have $T = \mathbb{N}$, leading to a countably infinite set of random variables. In continuous-time, usually $T = [0, \infty)$, producing an uncountably infinite set of random variables. Whether these random variables are independent or not is determined by the underlying process.

What is a Stochastic Process?

A stochastic process describes how a system with some underlying randomness evolves over space or time. Mathematically, we define this by a set of random variables

$$\{X(t) \mid t \in T\}$$

or more precisely

$$\{X(t, \omega) \mid t \in T\},$$

where T is referred to as the index set.

Very often, T is chosen to represent points in time. In discrete-time, we may have $T = \mathbb{N}$, leading to a countably infinite set of random variables. In continuous-time, usually $T = [0, \infty)$, producing an uncountably infinite set of random variables. Whether these random variables are independent or not is determined by the underlying process.

Sometimes, the notation X_t is used instead of $X(t)$.

Examples of Stochastic Processes

- Brownian motion: describes the random motion of small particles in fluids.

Examples of Stochastic Processes

- Brownian motion: describes the random motion of small particles in fluids.
- Markov chain Monte Carlo: used to generate samples from complicated distributions.

Examples of Stochastic Processes

- **Brownian motion**: describes the random motion of small particles in fluids.
- **Markov chain Monte Carlo**: used to generate samples from complicated distributions.
- **Diffusion processes**: used in machine learning for image generation (e.g. **Stable Diffusion**).

Examples of Stochastic Processes

- **Brownian motion**: describes the random motion of small particles in fluids.

- **Markov chain Monte Carlo**: used to generate samples from complicated distributions.

- **Diffusion processes**: used in machine learning for image generation (e.g. **Stable Diffusion**).

- **Galves–Löcherbach model**: models the dynamics of networks of biological neurons.

- Brownian motion: describes the random motion of small particles in fluids.
- Markov chain Monte Carlo: used to generate samples from complicated distributions.
- Diffusion processes: used in machine learning for image generation (e.g. Stable Diffusion).
- Galves–Löcherbach model: models the dynamics of networks of biological neurons.

Today we will focus on two of the simplest stochastic processes: the Bernoulli process and the Poisson process.

If $T = \mathbb{N}$ and $X(t) \sim \text{Bern}(p)$ for all $t \in T$, then $\{X(t) \mid t \in T\}$ is a Bernoulli process. It has the following important properties:

- The number of successes in any window in time, n time steps long, is binomially distributed:

$$\sum_{i=t}^{t+n-1} X(i) \sim \text{Bin}(n, p) \;\; \text{for all } t \in T$$

If $T = \mathbb{N}$ and $X(t) \sim \text{Bern}(p)$ for all $t \in T$, then $\{X(t) \mid t \in T\}$ is a Bernoulli process. It has the following important properties:

- The number of successes in any window in time, n time steps long, is binomially distributed:

$$\sum_{i=t}^{t+n-1} X(i) \sim \text{Bin}(n, p) \ \text{ for all } t \in T$$

- The waiting time until the next success, at any point in time, is geometrically distributed:

$$\min\{i \in \mathbb{N} \mid X(t+i-1) = 1\} \sim \text{Geo}(p) \ \text{ for all } t \in T$$

The Bernoulli Process

We can also consider an equivalent counting process $\{N(t) \mid t \in T\}$ which is given by

$$N(t) = \sum_{i=1}^{t} X(i) \ .$$

In general, a counting process represents the total number of occurrences of something up to and including time t. It is non-negative, non-decreasing, and takes integer values.

How would you simulate a Bernoulli process? There are a couple of simple ways, one more efficient than the other. This will be a coding exercise at the end of the lecture.

The Poisson Process

The Poisson process is the stochastic process that results from chopping the Bernoulli process into finer and finer intervals. Consider a Bernoulli process with success probability λ and each Bernoulli random variable representing the probability of a success occurring in a 1 second interval.

The Poisson process is the stochastic process that results from chopping the Bernoulli process into finer and finer intervals. Consider a Bernoulli process with success probability λ and each Bernoulli random variable representing the probability of a success occurring in a 1 second interval. Now replace each random variable with n Bernoulli random variables with success probability $\frac{\lambda}{n}$, each representing the probability of a success occurring in sequential $\frac{1}{n}$ second intervals. Letting $n \to \infty$ gives us the Poisson process. We have moved from discrete-time to continuous-time!

Consider a counting process with $T = [0, \infty)$. If:

- $N(0) = 0$,

The Poisson Process

Consider a counting process with $T = [0, \infty)$. If:

- $N(0) = 0$,
- $N(t + s) - N(t) \sim \text{Pois}(s\lambda)$ for all $t \in T$,

The Poisson Process

Consider a counting process with $T = [0, \infty)$. If:

- $N(0) = 0$,
- $N(t + s) - N(t) \sim \text{Pois}(s\lambda)$ for all $t \in T$,
- and $N(t_2) - N(t_1)$ and $N(t_4) - N(t_3)$ are independent for all disjoint intervals $(t_1, t_2]$ and $(t_3, t_4]$,

The Poisson Process

Consider a counting process with $T = [0, \infty)$. If:

- $N(0) = 0$,
- $N(t + s) - N(t) \sim \text{Pois}(s\lambda)$ for all $t \in T$,
- and $N(t_2) - N(t_1)$ and $N(t_4) - N(t_3)$ are independent for all disjoint intervals $(t_1, t_2]$ and $(t_3, t_4]$,

then $\{N(t) \mid t \in T\}$ is a Poisson process with rate λ. Other equivalent definitions exist, but are outside the scope of this lecture.

The Poisson Process

Consider a counting process with $T = [0, \infty)$. If:

- $N(0) = 0$,
- $N(t+s) - N(t) \sim \text{Pois}(s\lambda)$ for all $t \in T$,
- and $N(t_2) - N(t_1)$ and $N(t_4) - N(t_3)$ are independent for all disjoint intervals $(t_1, t_2]$ and $(t_3, t_4]$,

then $\{N(t) \mid t \in T\}$ is a Poisson process with rate λ. Other equivalent definitions exist, but are outside the scope of this lecture.

At any time t, what is the distribution on the waiting time until the next increment in $N(t)$? Find out in the next lecture.

The End