2.2 Probability Mass Functions

Definition

The **probability mass function (pmf)** of a discrete random variable X is a function $p_X : \mathbb{R} \to [0,1]$ defined by:
$p_X(x) = P(X = x), \quad x \in R_X,$
where $R_X = \mathrm{range}(X)$ is the set of values the random variable can take.
The pmf of a discrete random variable defines the **probability distribution** of a random variable.

- We frequently assign names to common probability distributions.

Range of a random variable

The range of a discrete random variable X is defined as
$R_X = \{x \in \mathbb{R} : P(X = x) > 0\}$.
When X is discrete, then R_X contains a countable number of elements.

Range versus support

The **support** of a discrete random variable is sometimes used synonymously with the range of X (**Casella and Berger 2002**; **Blitzstein and Hwang 2019**; **Reich and Ghosh 2019**).
Other **resources** define the support of X, S_X, as the smallest closed set such that
$\sum_{x \in S_X} p_X(x) = 1$.
Under this second definition, the range and support will be the same for discrete random variables but can differ for continuous random variables.

Notation

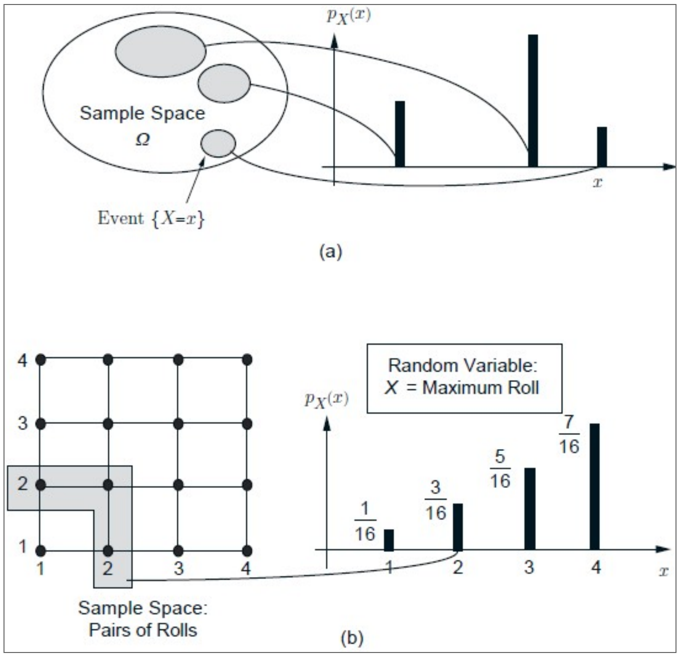
- Uppercase X will denote a random variable.
- Lowercase x will denote a specific numeric value.

Calculation of a discrete pmf

For each possible value x of X:

1. Collect all the possible outcomes that give rise to the event $\{X = x\}$.
2. Add their probabilities to obtain $p_X(x)$.

Visualization of pmf calculation



Calculating a pmf

Simple example

Suppose we toss a fair coin twice in a row and count the number of heads.

$$p_X(x) = \begin{cases} & \text{if } x = 0 \\ & \text{if } x = 1 \\ & \text{if } x = 2. \end{cases}$$

Properties of pmfs

The pmf of a random variable X, $p_X(x)$, satisfies the following properties:

- $p_X(x) \geq 0$ for all x
- $\sum_{x \in R_X} p_X(x) = 1$, where R_X is the range of X.
- If $A \subset R_X$, then $P(A) = \sum_{x \in A} p_X(x)$.

E.g., continuing our simple coin flipping example,
$P(X > 0) = p_X(1) + p_X(2) = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}$.

The Bernoulli random variable

A Bernoulli random variable models a single trial that has only two possible outcomes: success or failure.

- A success occurs with probability $p \in [0, 1]$ and is assigned the value 1.
- A failure occurs with probability $1 - p$ and is assigned the value 0.

The Bernoulli random variable defined

Let X be a Bernoulli random variable such that the probability of a success is $p \in [0, 1]$.

- Then $X \sim \text{Bernoulli}(p)$ (read as X is distributed as Bernoulli p).

The pmf of X is:

$$p_X(x) = \begin{cases} & \text{if } x = 1 \\ & \text{if } x = 0 \\ & \text{otherwise.} \end{cases}$$

R_X :

Bernoulli examples

- Website clicks: A user either clicks on an ad (success) or doesn't (failure).
- Medical diagnosis: A test either detects a disease (positive result) or not (negative result).
- Email spam filter: An income email is marked as spam (success) or not (failure).
- Sensor activation: A motion sensor either detects movement (success) or doesn't (failure) in a given time frame.

Independent and identically distributed

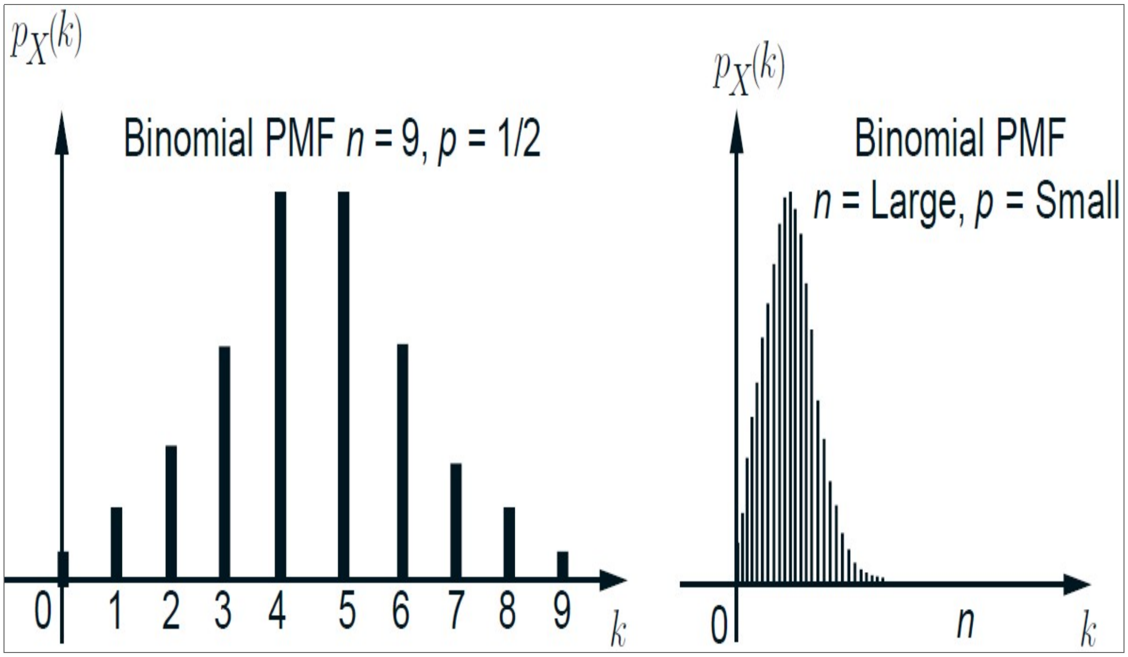
We often assume that random variables are **independent and identically distributed** (i.i.d.).
i.i.d. random variables are defined by two characteristics:

1. All of the random variables are independent.
   - Conceptually, the value of one random variable has no impact on the others (and so on).
2. Each of the random variables has the exact same probability distribution.

The Binomial random variable

A Binomial random variable models the number of successes in a sequence of n i.i.d. $\mathrm{Bernoulli}(p)$ trials.
If $X \sim \mathrm{Binomial}(n, p)$, then the pmf is
$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad \text{for } k = 0, 1, 2, \ldots, n.$
R_X :

Visualizing a Binomial pmf



The pmf of a Binomial random variable

Binomial examples

Product defects: counting the number of defective items in a batch of manufactured goods.
Email campaigns: counting how many recipients open an email out of a fixed number sent.
Vaccination cffectiveness: out of 100 vaccinated individuals, how many avoid infection?
Memory tests: how many words a subject recalls correctly out of a list.

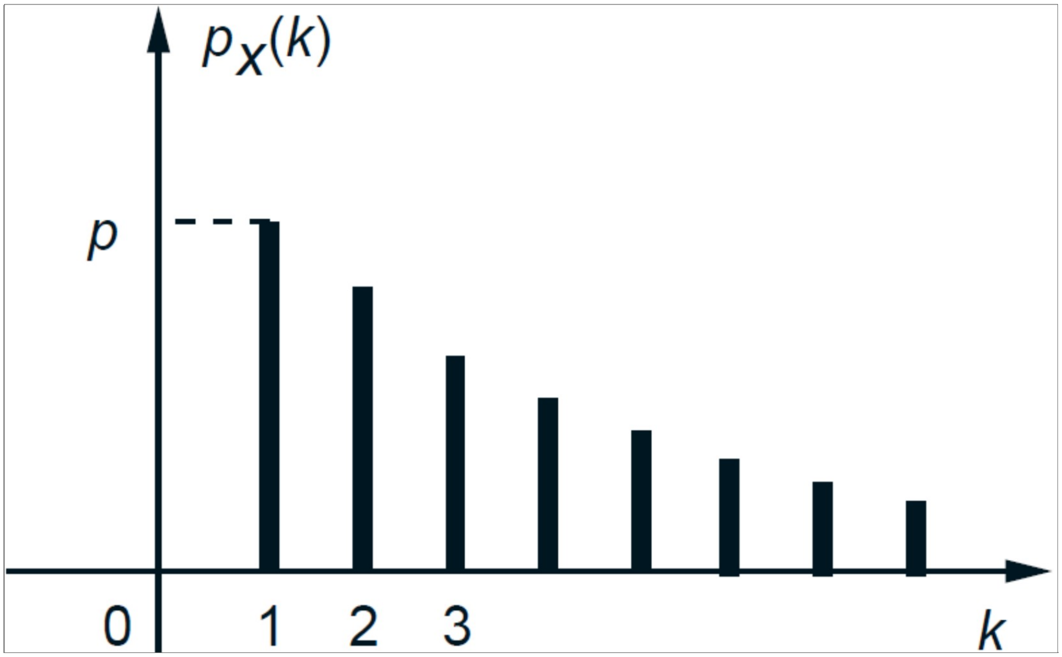
Computing a binomial probability

Only 5% of a lot of 20 electrical fuses are supposed to be defective. What is the probability that at least four defective fuses are observed?

The Geometric random variable

A Geometric random variable counts the number of i.i.d. $\mathrm{Bernoulli}(p)$ trials until the first success.
If $X \sim \mathrm{Geometric}(p)$, then the pmf is
$p_X(x) = (1-p)^{x-1}p, \quad x = 1, 2, 3, \ldots$
$R_X :$

Visualizing a Geometric pmf



The pmf of a Geometric random variable

Geometric examples

How many calls must a call center agent make until they reach a customer who answers?
How many items must we inspect until we find a defective one?
How many free throws must a person take until they make one?

Geometric computation

A call center agent must keep making calls until they reach a customer who answers. The probability of a customer answering is 0.10. What is the probability that the agent needs to make at least 20 calls?

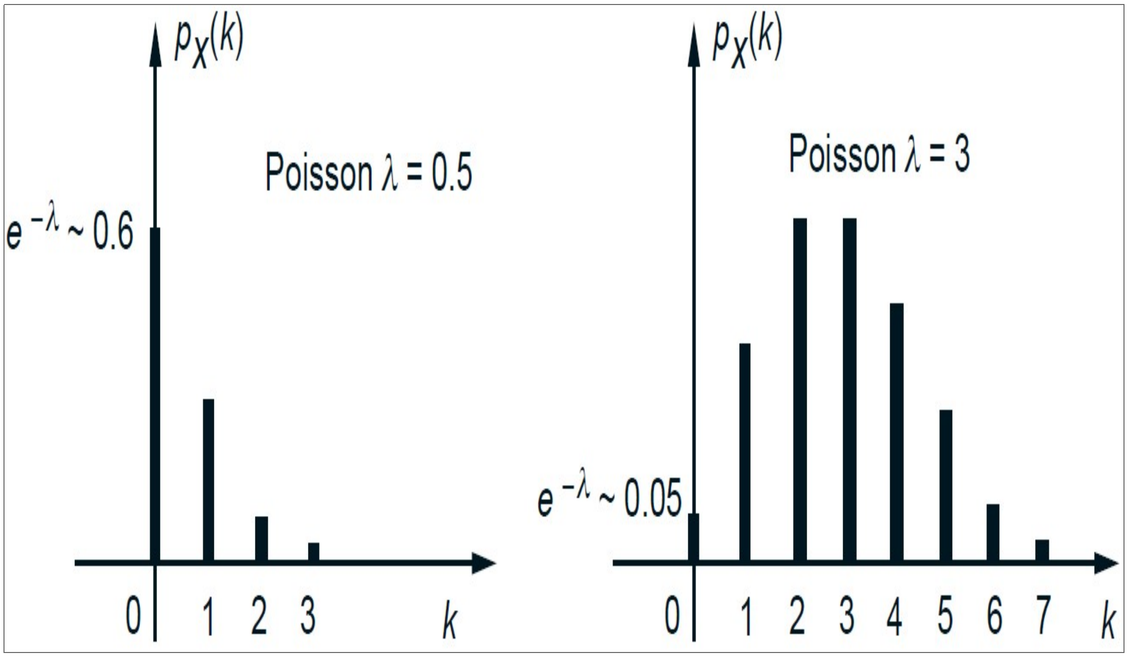
The Poisson distribution

A **Poisson random variable** models the number of events occurring in a fixed interval of time or space, assuming events occur independently and at a constant average rate, $\lambda > 0$.
If $X \sim \mathrm{Poisson}(\lambda)$, then the pmf is

$$p_X(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, \ldots$$

R_X :

Visualizing a Poisson pmf



The pmf of a Poisson random variable

Poisson examples

Emails received: the number of emails received per hour, with an average rate of $\lambda = 5$.
Website hits: the number of website in a minute if the average is 50 hits per minute.
Accidental deaths by horse kick: Ladislaus Bortkiewicz analyzed the number of soldiers in the Prussian army killed by horse kicks. The average was 0.61 deaths per corps per year, and the distribution of deaths followed a Poisson pattern.

Poisson computation

A small bakery receives an average of 4 online orders per hour. They can fulfill 8 online orders in an hour, if necessary. What is the probability that they receive more than 8 orders in an hour?

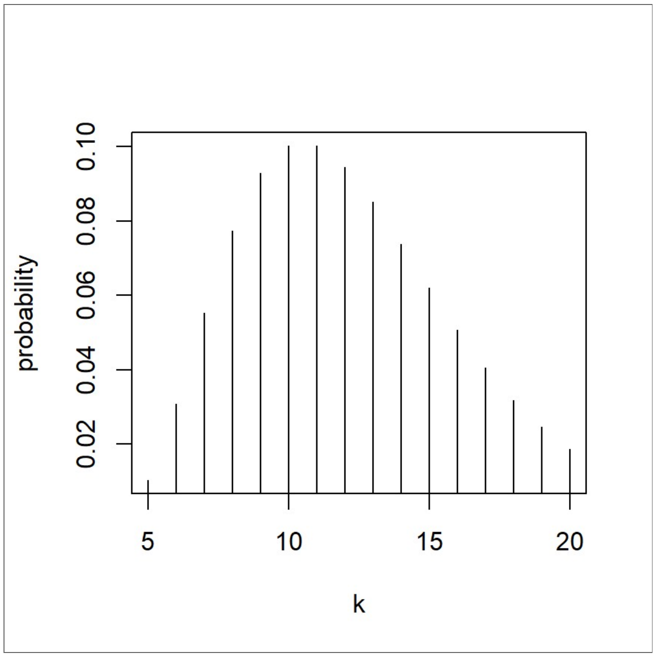
The Negative Binomial distribution

The Negative Binomial distribution models the total number of trials needed to achieve r successes in a sequence of i.i.d. $\text{Bernoulli}(p)$ trials.
If $X \sim \text{NB}(r, p)$, then the pmf is

$$p_X(k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}, \quad x = r, r+1, \ldots.$$

R_X :
Note: there are multiple characterizations, e.g., the number failures until the rth success.

Visualizing a Negative Binomial pmf



The pmf of a Negative Binomial random variable

Negative Binomial examples

Modeling disease outbreaks: model the number of cases until a certain number of recoveries or deaths occur.
Click-through rates: used in online advertising to model the number of ad impressions until a certain number of clicks occur.
Failure analysis: model the number of trials until a component fails a set number of times.

Negative Binomial computation

Suppose a basketball player has a 60% chance of making a free throw. What is the probability that it takes at least 7 shots to make 3 successful free throws?

References

Blitzstein, Joseph K, and Jessica Hwang. 2019. *Introduction to Probability, 2nd Edition*. Chapman; Hall/CRC.

Casella, George, and Roger Berger. 2002. *Statistical Inference, 2nd Edition*. Chapman; Hall/CRC.

Reich, Brian J, and Sujit K Ghosh. 2019. *Bayesian Statistical Methods*. Chapman; Hall/CRC.