# Mathematical Statistics I

## Chapter 1: Probability

Jesse Wheeler

# Outline

# Introduction

# Introduction

- Formally, a random variable is a function from a sample space $\Omega$ to the real numbers[1].
- That is, for any element $\omega \in \Omega$, a random variable $X$ will map $\omega$ to a real number: $X(\omega) \in \mathbb{R}$.
- Most often people think of random variables as random numbers rather than functions; in most instances in this class, this treatment will be sufficient.

[1] In this class, will assume real-valued spaces, though more generally a random variable can map to any measureable space

# Example of a random variable

Consider the experiment of flipping three coins. The sample space is

$$\Omega = \{hhh, hht, hth, thh, htt, tht, tth, ttt\}.$$

- Some possible random variables include (1) the number of heads, (2) the number of tails, (3) the number of heads minus the number of tails.
- Importantly, a random variable must assign a value to all possible outcomes $\omega \in \Omega$.

**Number of Heads**

Let $X$ be the random variable representing the number of heads. If the result of the outcome is the event $hth$, the $X(\{hth\}) = 2$.

# A few comments on random variables

- Sometimes in this course I will use the abbreviation RV to mean "random variable", and you can do so as well.
- It is conventional to use uppercase letters (math text or italics) to denote random variables.
- While a random variable is a function, the outcome of an experiment $\omega \in \Omega$ is random (that's the point), and we only ever see a single outcome. Thus, the fact that $X$ is a function is often dropped, and we just write $X$. The realized value of $X$ is random, because the input is random.

# Discrete Random Variables

# Discrete Random Variables

**Definition: Discrete random variable**

A discrete random variable is a random variable that can take on only a finite or at most a countably infinite number of values.

- Example: The number of heads in three coin flips can only be in the set $\{0, 1, 2, 3\}$. Alternatively, consider flipping a coin indefinitely until you achieve a heads. The possible outcomes are in the set $\{1, 2, 3, \ldots\}$, which is countably infinite.

# Probabilities

- The probability measure on the sample space determines the probability of the values of $X$.
- In our example, if a coin is fair, then we can assign a uniform probability measure on the sample set of flipping a coin three times.
- That is, all outcomes are equally likely, each with probability $1/8$.
- The probability that $X$ takes on it's potential values is easily computed, by counting the number of outcomes that result in the particular value of $X$:

# Probabilities II

$$P(X = 0) = \frac{1}{8}$$
$$P(X = 1) = \frac{3}{8}$$
$$P(X = 2) = \frac{3}{8}$$
$$P(X = 3) = \frac{1}{8}.$$

# Probabilities III

- More generally, let's assume that $X$ is a discrete RV, and denote the possible values as $x_1, x_2, \ldots$. There exists a function $p$ such that $p(x_i) = P(X = x_i)$ that satisfies $\sum_i p(x_i) = 1$. This function $p$ is called the probability mass function (pmf) of the random variable $X$.
- We may also be interested in calculating for all values $x \in \mathbb{R}$, the probability $F(x) = P(X \leq x)$; the function $F$ is called the cumulative distribution function (cdf). The cdf plays a number of important roles in probability and statistics that we will see later on.

# Probabilities IV

Some notes:

- The cdf is non-decreasing (see Theorem 1.2), and

$$\lim_{x \to -\infty} F(x) = 0 \quad \text{and} \quad \lim_{x \to \infty} F(x) = 1.$$

- The pmf and cdf are connected: the cdf "jumps" at all values that the pdf $p(x) > 0$.
- Conventionally, the pmf is usually denoted with lower-case letters (e.g., $p$, $f$), whereas the cdf is usually denoted with upper-case letters (e.g., $F$).

# Independence

- Jumping ahead a little bit, we will define what it means for random variables to be independent (a chapter 3 topic).

**Definition: Independent random variables**

Let $X$ and $Y$ be discrete random variables defined on the same probability space, taking values $x_1, x_2, \ldots$ and $y_1, y_2, \ldots$, respectively. $X$ and $Y$ are said to be independent if, for all $i, j$,

$$P(X = x_i, Y = y_i) = P(X = x_i)P(Y = y_i).$$

- This definition follows very similarly to that of independent events. We can also extend this definition to mutual independence of many variables if the probabilities of all combinations of variables can be factored.

# Bernoulli Random Variables

- A Bernoulli RV only takes on two values[2], $0$ and $1$, with probabilities $1-p$ and $p$, respectively. The pmf is therefore

$$\begin{aligned} p(1) &= p \\ p(0) &= 1-p \\ p(x) &= 0, \quad \text{if } x \neq 0 \text{ and } x \neq 1. \end{aligned}$$

- By using the output of $0$ and $1$, the pmf is usually written in a more compact form:

$$p(x) = \begin{cases} p^x(1-p)^x, & \text{if } x = 0 \text{ or } x = 1, \\ 0 & \text{otherwise} \end{cases}$$

[2]Sometimes you'll see the random variable take values $-1$ and $1$.

# Indicator functions

- A common instance of a Bernoulli RV is an indicator random variable. Let $I_A$ be the random variable that takes on the value of $1$ if the event $A \subset \Omega$ occurs, and $0$ otherwise:

$$I_A(\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \text{otherwise} \end{cases}$$

- Here, we see that $P(I_A = 1) = P(A)$.

# Binomial Distribution

- Using what we know about independent RVs and Bernoulli RVs, we can derive the pmf for a Binomial distribution.
- Suppose that we have $n$ independent experiments, where $n$ is a fixed (positive) integer. Let each experiment have two outcomes with probabilities $p$ and $1-p$, respectively, which we call "success" or "failure". We are interested now in the random variable $X$, the number of "successes" in $n$ independent trials.
- *Question:* What is the probability that $X = k$, for some $k \in \{0, 1, 2, \ldots\}$?

# Binomial Distribution II

*Solution Sketch*:

# Binomial Distribution III

## Flipping Coins

Suppose that a coin is flipped $10$ times. What is the probability that the coin lands heads heads exactly $6$ times?

Here, $n = 10$, and success= Heads. Assuming the coin is fair, we have

$$P(\text{Num Heads} = 6) = \binom{10}{6}(0.5)^6(0.5)^4 \approx 210 \times 0.00098 \approx .205$$

# Binomial Distribution IV

- Suppose a five 6-sided (fair) dice are rolled simultaneously. What is the probability that at least two of the dice show the value 6?
- Let $X$ denote the number of 6s in this experiment, which takes values in the set $\{0, 1, \ldots 5\}$. We want the probability that $X \geq 2$.
- Because the different values of $X$ are mutually exclusive events (i.e., $X = 2$ implies $X \neq 3$), we can calculate this as:

$$P(X \geq 2) = \sum_{i \in \{2,3,4,5\}} p(i) \approx 0.1962449,$$

where $p(i)$ is the pmf of the binomial$(5, 1/6)$ distribution.

# Binomial Distribution V

- Alternatively, we can use the complement set, which is smaller:

$$P(X \geq 2) = 1 - P(X < 2) = 1 - \big(p(0) + p(1)\big) \approx 0.1962449$$

- *Note:* A binomial RV can be expressed as the sum of independent Bernoulli RVs. That is, let $X_1, X_2, \ldots, X_n$ be independent Bernoulli RVs, each with $P(X_i = 1) = p$. Then, $Y = X_1 + X_2 + \ldots + X_n$ is a Binomial RV, with parameters $(n, p)$.

# Geometric Distribution

- We can construct a geometric RV in a similar way that we did with the binomial distribution.
- Suppose instead of having a fixed number of trials, we continue having a trial until our first success. That means that if $X = k$, we will have $k - 1$ failures, one success, and then stop.
- Thus, the pmf can easily be constructed to be:

$$p(k) = P(X = k) = (1 - p)^{k-1} p, \quad k = 1, 2, 3, \ldots$$

# Geometric Distribution II

## Geometric Series

Recall from calculus the geometric series:

$$\sum_{i=0}^{\infty} r^i = \frac{1}{1-r}, \quad \text{if } 0 < r < 1.$$

This identity occurs in the pmf of the geometric series. Let $0 < p < 1$, then

$$\sum_{k=1}^{\infty} (1-p)^{k-1} p = p \sum_{j=0}^{\infty} (1-p)^j = p \frac{1}{1-(1-p)} = 1.$$

# Negative Binomial Distribution

- The negative binomial (NB) distribution can be thought of as a generalization of the geometric distribution; rather than stopping when we have exactly one success, we now will stop when we have $r$ successes.
- For any particular sequence of trials of length $k$ that satisfy this condition, the probability is $p^r(1-p)^{k-r}$.
- The last trial must be a success (because we stopped), so we need to choose the location of the remaining $r-1$ successes.
- Thus, if $X$ has a negative binomial distribution, the pmf is:

$$p(k) = P(X = k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}.$$

# Negative Binomial Distribution II

- Another way that can be helpful for thinking about the NB-distribution is considering it as the sum of $r$ independent geometric random variables:
- We want to represent the total number of trials until the $r$th success, which is the sum of the number of trials until (and including) the first success, plus the number of trials from the first success until (and including) the second success, and continued until we get $r$ successes.

# Negative Binomial Distribution III

## Negative Binomial Lottery

Suppose that there is a type of lottery where each purchased ticket has equal probability of winning $(p = 1/100)$, and there are 3 total prizes to be won. What is the probability that exactly $k$ tickets will be sold until all prizes have been won?

$$P(X = k) = \binom{k-1}{3-1}(0.01)^3(0.99)^{k-3}.$$

# The Hypergeometric Distribution

- Suppose that there is a total population of size $n$, and $r$ have some trait of interest ("success"), and $n - r$ do not ("failure").
- If we sample $m$ items from the population, then the total number of "successes" in our sample of size $m$ follows a hypergeometric distribution:

$$P(X = k) = \frac{\binom{r}{k}\binom{n-r}{m-k}}{\binom{n}{m}}.$$

# The Hypergeometric Distribution II

- Combinatorially, for $X = k$, we must select $k$ successes out of the total possible $r$ successes in the entire population; there are $\binom{r}{k}$ ways to do this.
- Since we selected $m$ objects in our sample, and we want $m - k$ of them to be failures, we must pick $m - k$ failures from the $n - r$ failures in the population; there are $\binom{n-r}{m-k}$ ways to do this.
- Together, the multiplication principle implies there are $\binom{r}{k}\binom{n-r}{m-k}$ ways that a sample of size $m$ contains $k$ successes from described population.
- Finally, there are a total of $\binom{n}{m}$ ways we can pick our sample.

# The Hypergeometric Distribution III

## Balls in a basket

Suppose that there are $n$ balls in a basket, and $r$ balls are black, $n-r$ balls are some other color. If we select $1 \leq m < n$ balls randomly (without replacement), let $X$ denote the number of black balls in our sample of size $m$. Then, for all $0 \leq k \leq r$,

$$P(X = k) = \frac{\binom{r}{k}\binom{n-r}{m-k}}{\binom{n}{m}}.$$

# The Poisson Distribution

- The Poisson distribution is used very frequently in both theory and practice, though the derivation is less intuitive than other distributions, so we will first just provide the pmf:

**Definition: Poisson Distribution**

The pmf of a random variable $X$ that follows a Poisson distribution with parameter $\lambda > 0$ is

$$p(k) = P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \ldots$$

# The Poisson Distribution II

- Recall from calculus that $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$. Thus, like all pmf's, the pmf of a Poisson distributed RV sums to one:

$$\begin{aligned}\sum_{k=0}^{\infty} p(k) &= \sum_{k=0} \frac{\lambda^k}{k!} e^{-\lambda} \\ &= e^{-\lambda} \sum_{k=0} \frac{\lambda^k}{k!} \\ &= e^{-\lambda} e^{\lambda} = 1.\end{aligned}$$

# The Poisson Distribution III

- The value of $\lambda$ controls the *shape* of the distribution:

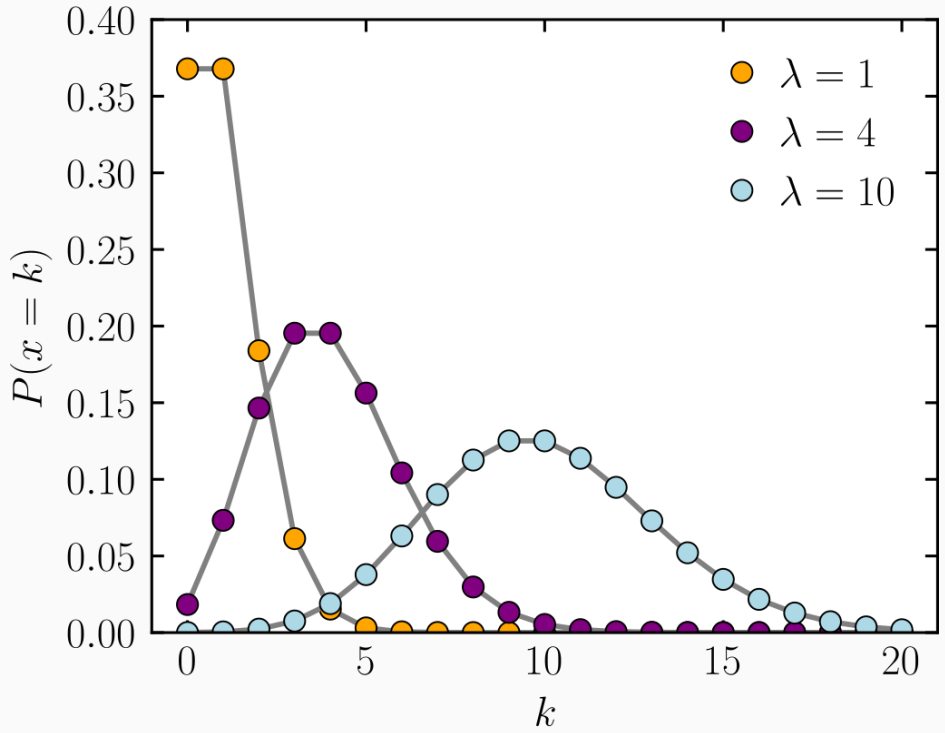


**Figure 1:** Shape of the Poisson distribution for various values of $\lambda$ (Wikipedia contributors, 2025b).

# The Poisson Distribution IV

- The Poisson distribution can be derived as the limit of a binomial distribution as the number of trials $n \to \infty$, and $p \to 0$, such that $np = \lambda$.

See next slide.

# The Poisson Distribution V

*Derivation:*

# The Poisson Distribution VI

*Derivation (continued):*

# The Poisson Distribution VII

This derivation suggests how a Poisson distribution can arise in practice.

- Let $X$ denote the random variable representing the number of times some event occurs in a fixed time interval.
- Think of dividing the interval into very large number of small sub-intervals of equal length.
- Assume that the sub-intervals are so small that the probability of more than one event in a sub-interval is negligible relative to the probability of one event (which itself is small).
- Finally, assume that the probability of an event in a given sub-interval is identical and independent of that of other sub-intervals.

# The Poisson Distribution VIII

- Following this, X is nearly binomially distributed, with $n$ being the number of sub-intervals, and $p = \lambda/n$ the probability of the event in each sub-interval.
- Taking the limit, we get something that is nearly Poisson distributed.

# The Poisson Distribution IX

- This idea can actually be formalized and made rigorous; you would probably see something like this in a course on stochastic processes.
- The Poisson distribution is often used to model the number of events that occur in a fixed interval.

# The Poisson Distribution X

The Poisson distribution is often good model for the number of events in a fixed time interval if the following conditions are met:

- The occurrence of one event does not affect the occurrence of another.
- The rate at which events occur is fixed.
- Two events cannot occur at the exact same instant.

In this scenario, the random (stochastic) process that generates the data is called a Poisson process, which gives rise to the name rate for the parameter $\lambda$.

# The Poisson Distribution XI

**Example: Telephone calls**

Suppose that an office receives telephone calls as a Poisson process with $\lambda = 0.5$ calls per minute. The number of calls in a 5-min. interval follows a Poisson distribution with parameter $5\lambda = 2.5$. Thus, the probability of no calls in a 5-min. interval is $p(0) = e^{-2.5} \approx .082$; the probability one one call is $p(1) = 2.5e^{-2.5} \approx .205$

# Continuous Random Variables

# Introduction

- Because discrete RVs take only a finite number of possibilities, they are relatively simple to define.
- In many situations, however, we are interested in random variables that can take on a continuum of values rather than a finite (or countably infinite) number.

**Example: Lifetime of electronic**

We might be interested in the lifetime of an electronic component; the total lifetime may be random, but may take on any positive real number.

# Density function

- For continuous random variables, we no longer have a pmf (which maps all values of the random variable to their corresponding probabilities).
- Instead, the role of the pmf is taken by a probability density function (pdf), which we will denote $f(x)$.

**Basic properties of a pdf**

If $f(x)$ is a pdf, then $f(x) \geq 0$ for all $x$, $f$ is piece-wise continuous, and $\int_{-\infty}^{\infty} f(x)dx = 1$.

# Probabilities

- If $X$ is a random variable with a density function $f$, then for any $a \leq b$, the probability that $X$ falls in the interval $(a, b)$ (with the treatment that if $a = b$, the interval collapses to the set $\{a\}$) is given by:

$$P(a < X < b) = \int_a^b f(x)dx.$$

- An immediate consequence of this definition is that $P(X = a) = 0$ for any $a \in \mathbb{R}$.

# Example: Continuous uniform random variable

- By *uniform* probability, we mean that all outcomes in the given set are equally as likely.
- For example, if $X$ is a RV with uniform distribution on the interval $[0, 1]$, then any real number in this interval is equally likely, and the probability that $X$ is in a sub-interval of length $h$ should be equal to $h$.
- You can verify that the following density satisfies this condition:

$$f(x) = \begin{cases} 1, & 0 \leq x \leq 1 \\ 0 & x < 0 \text{ or } x > 1. \end{cases}$$

# Example: Continuous uniform random variable II

- The previous density can be generalized to any interval $[a, b]$, such that $a < b$.

**Continuous uniform density**

If $X$ is a RV uniformly distributed on an interval $[a, b]$, where $a < b$, then the corresponding density function is:

$$f(x) = \begin{cases} 1/(b-a) & a \leq x \leq b \\ 0 & x < a \text{ or } x > b. \end{cases}$$

# Example: Continuous uniform random variable III

- One important thing to note is that if $X$ is a continuous RV, then

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

- This is because the probability $X$ being any particular value is zero; if this were not the case, then the probability of the entire set would infinite (and probabilities must sum to one).

# Cumulative distribution function

- The cumulative distribution function of a continuous random variable $X$ is defined in the same way as for a discrete random variable:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)dx.$$

- Thus, we can connect the cdf and the pdf of a continuous random variable using the fundamental theorem of calculus.
- Specifically, if $f(x)$ is continuous at $x$, then $F'(x)$, and

$$P(a \leq X \leq b) = \int_{a}^{b} f(x)dx = F(b) - F(a).$$

# Cumulative distribution function II

- This derivation gives some hints at properties of the cdf (both continuous and discrete RVs). Let $F$ be a distribution function. Then the following properties hold:
  - $F$ is right-continuous.
  - $F$ is monotonically increasing (non-decreasing).
  - $F : \mathbb{R} \to [0, 1]$ and satisfies $\lim_{x \to -\infty} F(x) = 0$, and $\lim_{x \to \infty} F(x) = 1$.
- *Note*: every probability distribution supported on the real numbers is uniquely identified by its distribution function $F$ (more to come).

# Cumulative distribution function III

## cdf of continuous uniform density

From the definition, we can calculate the cdf of the continuous uniform density rather easily. Suppose that $X$ is uniformly distributed on $[0, 1]$. Then the cdf $F$ is:

$$\begin{aligned} F(X \leq x) &= \int_{-\infty}^{x} f(x)dx \\ &= \int_{-\infty}^{x} 1[0 \leq x \leq 1]dx \\ &= \begin{cases} 0 & x < 0 \\ x & 0 \leq x \leq 1 \\ 1 & x > 1. \end{cases} \end{aligned}$$

# Percentiles

- You're probably familiar with the term median. For any given sample, the median defines the "mid-point", meaning that half of the values are larger, half are smaller.
- This same concept applies to distribution functions.
- That is, the median of a distribution $F$ is defined to be that value $x_{.5}$ such that $P(X < x_{.5}) = 0.5$.
- Formally, the sample median is the same as the definition above, using the *empirical* distribution function (Wikipedia contributors, 2025a).

It is important to note that, as defined, the median value may not be unique!

# Percentiles II

**Definition: Percentile**

Let $F$ be the cdf of a continuous random variable. The $p$th quantile of the distribution $F$ is defined to be any value $x_p$ such that $F(x_p) = P(X \leq x_p) = p$. If $F$ is strictly increasing, then $x_p$ is unique and we say that $F^{-1}(p) = x_p$.

If $F$ is not strictly increasing, then $x_p$ may not be unique; in this case, all such values are considered percentiles. If an inverse function is needed in this case, we will define $F^{-1}(p) = \inf\{x \in \mathbb{R} : F(x) \geq p\}$.

- The last bit of the definition is just some important book keeping to ensure the inverse function exists in odd examples, though I don't think it comes up in this course.

# Percentiles III

Some important percentiles have their own names, including:

- Median: $p = 1/2$.
- Quartiles (lower and upper): $p = 1/4$, and $p = 3/4$, resp.
- Min: $p = 0$.
- Max: $p = 1$.

Note that the inverse cdf is sometimes called the quantile function.

# Percentiles IV

## Calculating the inverse cdf

Suppose that

$$F(x) = \begin{cases} 0 & x < 0 \\ x^2 & 0 < x < 1 \\ 1 & x > 1 \end{cases}$$

for $0 \leq x \leq 1$. Find the inverse distribution function $F^{-1}$.

*Solution:*

# Exponential Distribution

- The exponential density function is:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

- Like the Poisson distribution, the exponential density function depends on a single parameter $\lambda$.
- When this is the case, we refer to it as the family of exponential densities that is *indexed* by the parameter $\lambda$.

# Exponential Distribution II

- The cdf is easily found via the fundamental theorem of calculus:

$$F(x) = \int_{-\infty}^{x} f(u)du = \begin{cases} 1 - e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

- From this, we can easily find quantiles of the distribution, such as the median: by solving $F(x_{(.5)}) = 1/2$

$$1 - e^{-\lambda x_{(.5)}} = \frac{1}{2} \quad \Longrightarrow \quad x_{(.5)} = \frac{\log 2}{\lambda}.$$

# Exponential Distribution III

- The exponential distribution is often used to model lifetimes or waiting times (time-to-event).
- In this context, it's conventional to replace the variable $x$ with $t$.
- The exponential distribution has a unique property known as the memoryless property.
- That is, if something follows an exponential distribution and has already lasted a time of $s$, then the probability that it will last another $t$ units of time does not depend on $s$:

# Exponential Distribution IV

*Memoryless property:* Let $T$ be an exponentially distributed RV, and $s, t > 0$. Calculate $P(T > t + s | T > s)$.

# Exponential Distribution V

- It can be shown that any continuous RV with the *memoryless* property must be exponentially distributed.
- Similarly, it can be shown that any discrete RV with the *memoryless* property must be geometrically distributed (maybe a HW question?)

# Exponential Distribution VI

- The exponential distribution is also related to the *Poisson process* that we have discussed.
- Consider a poisson process with rate $\lambda$ over an interval $\mathcal{T} \subset \mathbb{R}$.
- While the number of events in any interval $T_0 \subset \mathcal{T}$ of length $t$ follows a Poisson distribution, the time-to-next-event $T$ follows an exponential distribution:

# References and Acknowledgements

Wikipedia contributors (2025a). "Empirical distribution function — Wikipedia, The Free Encyclopedia." https://en.wikipedia.org/w/index.php?title=Empirical_distribution_function&oldid=1300940691. [Online; accessed 14-August-2025].

Wikipedia contributors (2025b). "Poisson distribution." Accessed 14 August 2025, URL https://en.wikipedia.org/wiki/Poisson_distribution.

- Compiled on August 15, 2025 using R version 4.5.1.

# References and Acknowledgements II



- We acknowledge students and instructors for previous versions of this course / slides.