Probability Theory

Matt Arthur

UCR GradQuant

April 15, 2021

Building Blocks

To apply probability theory, we need to define three objects:

- Sample Space: the set of all possible outcomes of an experiment. Usually denoted Ω.
- A collection of subsets of Ω for which we can calculate probabilities. Usually denoted $\mathcal{S}$
- A probability function, $\mathbb{P}$, which maps objects in $\mathcal{S}$ to the interval $[0, 1]$.

Building Blocks

Example: Flip a fair coin twice.

- Sample Space: $\Omega = \{\text{HH}, \text{HT}, \text{TH}, \text{TT}\}$
- A collection of subsets:

$$\mathcal{S} = \left\{ \begin{array}{llll} \{\emptyset\} & \{\text{HH}\} & \{\text{HT}\} & \{\text{TH}\} \\ \{\text{TT}\} & \{\text{HH}, \text{HT}\} & \{\text{HH}, \text{TH}\} & \{\text{HH}, \text{TT}\} \\ \{\text{HT}, \text{TH}\} & \{\text{HT}, \text{TT}\} & \{\text{TH}, \text{TT}\} & \{\text{HH}, \text{HT}, \text{TH}\} \\ \{\text{HH}, \text{HT}, \text{TT}\} & \{\text{HH}, \text{TH}, \text{TT}\} & \{\text{HT}, \text{TH}, \text{TT}\} & \{\text{HH}, \text{HT}, \text{TH}, \text{TT}\} \end{array} \right\}$$

- A probability function, $\mathbb{P} : \mathcal{S} \to [0, 1]$:

  Used to calculate expressions such as:

  - $\mathbb{P}(\{\text{HH}\}) =$ the probability that the result of the experiment is HH
  - $\mathbb{P}(\{\text{HH}, \text{HT}, \text{TT}\}) =$ the probability that the result of the experiment is HH or HT or TT.

Events

Elements of $\mathcal{S}$ are called *events*.

Events

Elements of $\mathcal{S}$ are called *events*.

- Two events A_1 and A_2 are said to be *mutually exclusive* if they cannot happen simultaneously. That is, if A_1 and A_2 are disjoint.

Example: $\{\text{HH}\}$ and $\{\text{HT}\}$ are mutually exclusive because $\{\text{HH}\} \cap \{\text{HT}\} = \emptyset$

Example: $\{\text{HH}, \text{HT}\}$ and $\{\text{HH}, \text{TT}\}$ are not mutually exclusive because

$$\{\text{HH}, \text{HT}\} \cap \{\text{HH}, \text{TT}\} = \{\text{HH}\} \neq \emptyset$$

Choosing $\mathbb{P}$

Selection of the probability function $\mathbb{P}$ is up to us. There are a few restrictions on $\mathbb{P}$ which constitute the main axioms of probability:

(1) $\mathbb{P}(\emptyset) = 0$. The probability of obtaining nothing from the experiment is zero.

(2) $\mathbb{P}(\Omega) = 1$. We are certain to obtain something from the experiment.

(3) If $A_1, A_2, A_3, \ldots$ is a sequence of disjoint events, then

$$\mathbb{P}\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} \mathbb{P}(A_k)$$

Unpacking (3)

Condition (3) on $\mathbb{P}$ is sometimes called a condition of *countable additivity*:

If $A_1, A_2, A_3, \ldots$ is a sequence of disjoint events, then

$$\mathbb{P}\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} \mathbb{P}(A_k)$$

Unpacking (3)

Condition (3) on $\mathbb{P}$ is sometimes called a condition of *countable additivity*:

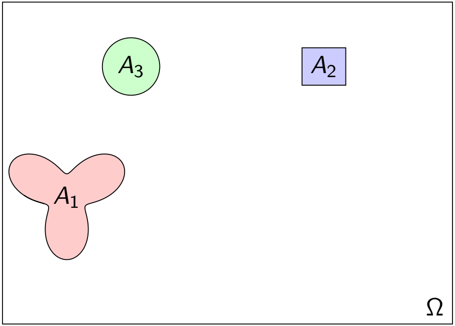
If $A_1, A_2, A_3, \ldots$ is a sequence of disjoint events, then

$$\mathbb{P}\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} \mathbb{P}(A_k)$$

Motivation for (3): If events have no outcomes in common, the likelihood that any of the events occurs is the sum of the likelihoods for each individual event.

Visualization of (3)

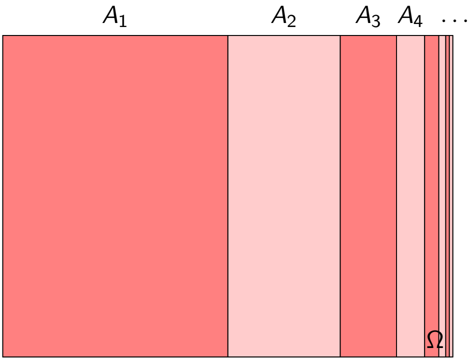
Figure: Disjoint Events in Ω



Since A_1, A_2, and A_3 have no outcomes in common, we expect the likelihood of observing A_1 or A_2 or A_3 to be: $\mathbb{P}(A_1) + \mathbb{P}(A_2) + \mathbb{P}(A_3)$.

Visualization of (3)

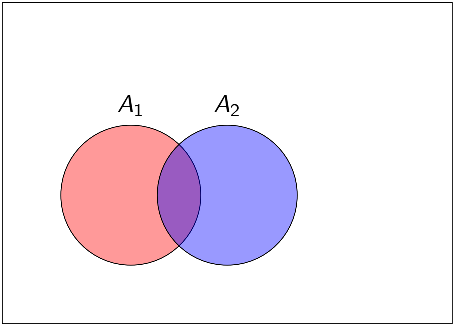
Figure: Disjoint Events in Ω



Since $A_1, A_2, \dots$ have no outcomes in common, we expect the likelihood of observing any one of these events to be $\sum_{k=1}^{\infty} \mathbb{P}(A_k)$.

Non-Disjoint Events

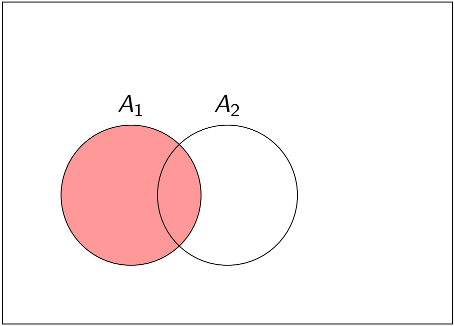
Figure: Non-Disjoint Events in Ω



$\mathbb{P}(A_1 \cup A_2) =$

Non-Disjoint Events

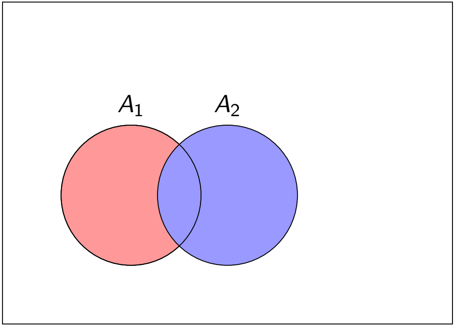
Figure: Non-Disjoint Events in Ω



$\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1)$

Non-Disjoint Events

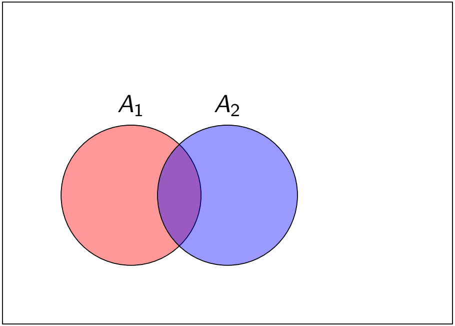
Figure: Non-Disjoint Events in Ω



$$\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2)$$

Non-Disjoint Events

Figure: Non-Disjoint Events in Ω



$$\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2) - \mathbb{P}(A_1 \cap A_2)$$

Non-Disjoint Events

This gives the general formula:

$$\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2) - \mathbb{P}(A_1 \cap A_2)$$

This works whether or not the events are mutually exclusive. If A_1 and A_2 are disjoint, then $A_1 \cap A_2 = \emptyset$ and so $\mathbb{P}(A_1 \cap A_2) = 0$.

Conditional Probability

Suppose we have two events $A_1, A_2 \subset \Omega$.

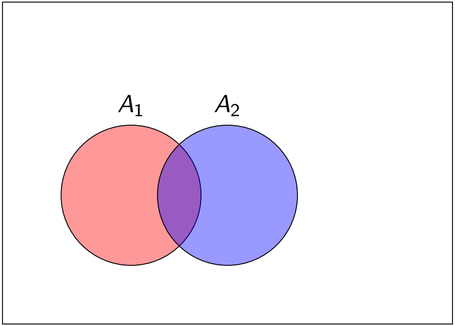
We know that A_1 has occurred. Should this change our thoughts about the likelihood that A_2 has occurred?

Suppose we have two events $A_1, A_2 \subset \Omega$.

We know that A_1 has occurred. Should this change our thoughts about the likelihood that A_2 has occurred? Sometimes!

Conditional Probability

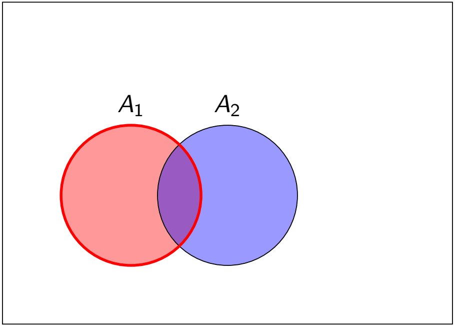
Figure: Non-Disjoint Events in Ω



If we have no information about whether A_1 has occurred, then the likelihood that A_2 has occurred is simply $\mathbb{P}(A_2)$.

Conditional Probability

Figure: Non-Disjoint Events in Ω



If we know that A_1 has occurred, then the likelihood of A_2 should be the "proportion" of the A_2's likelihood that also lies in A_1: $\dfrac{\mathbb{P}(A_2 \cap A_1)}{\mathbb{P}(A_1)}$

Conditional Probability

We write "the probability of A_2 given that A_1 has occurred" as

$$\mathbb{P}\left(A_2 \mid A_1\right) = \frac{\mathbb{P}\left(A_2 \cap A_1\right)}{\mathbb{P}\left(A_1\right)}$$

Conditional Probability

We write "the probability of A_2 given that A_1 has occurred" as

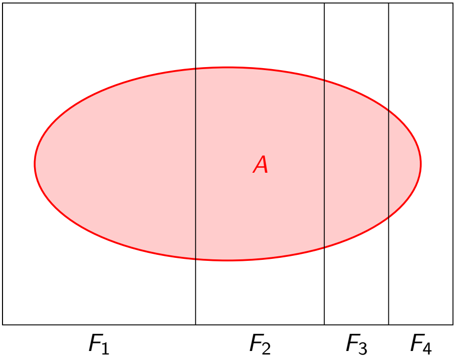
$$\mathbb{P}(A_2 \mid A_1) = \frac{\mathbb{P}(A_2 \cap A_1)}{\mathbb{P}(A_1)}$$

From which we obtain the important relationship:

$$\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_2 \mid A_1)\,\mathbb{P}(A_1)$$

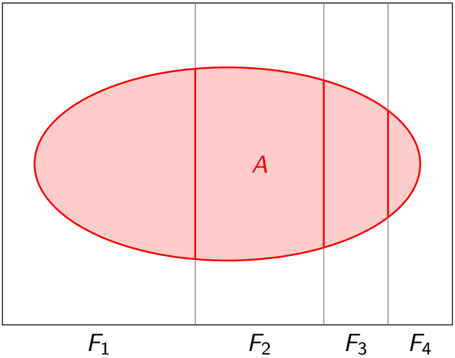
Partitioning the Sample Space

Figure: Partitioning Ω



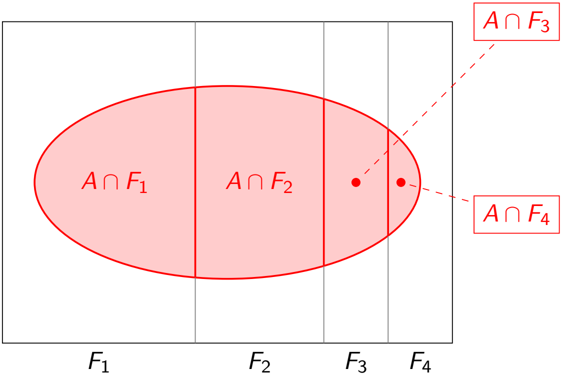
Partitioning the Sample Space

Figure: Partitioning Ω



Partitioning the Sample Space

Figure: Partitioning Ω



Partitioning the Sample Space

F_1, F_2, F_3, F_4 are said to form a *partition* of the sample space Ω:

- $F_1 \cup F_2 \cup F_3 \cup F_4 = \Omega$
- F_1, F_2, F_3, F_4 are disjoint.

Partitioning the Sample Space

F_1, F_2, F_3, F_4 are said to form a *partition* of the sample space Ω:

- $F_1 \cup F_2 \cup F_3 \cup F_4 = \Omega$
- F_1, F_2, F_3, F_4 are disjoint.

Note that the following events are also disjoint:

$$A \cap F_1, A \cap F_2, A \cap F_3, A \cap F_4$$

Partitioning the Sample Space

Note that the following events are also disjoint:

$$A \cap F_1, A \cap F_2, A \cap F_3, A \cap F_4$$

Therefore, we can write $\mathbb{P}(A)$ as

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(A \cap F_1) + \mathbb{P}(A \cap F_2) + \mathbb{P}(A \cap F_3) + \mathbb{P}(A \cap F_4) \\ &= \mathbb{P}(A \mid F_1)\,\mathbb{P}(F_1) + \mathbb{P}(A \mid F_2)\,\mathbb{P}(F_2) + \mathbb{P}(A \mid F_3)\,\mathbb{P}(F_3) + \mathbb{P}(A \mid F_4)\,\mathbb{P}(F_4)\end{aligned}$$

Partitioning the Sample Space

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(A \cap F_1) + \mathbb{P}(A \cap F_2) + \mathbb{P}(A \cap F_3) + \mathbb{P}(A \cap F_4) \\ &= \mathbb{P}(A \mid F_1)\,\mathbb{P}(F_1) + \mathbb{P}(A \mid F_2)\,\mathbb{P}(F_2) + \mathbb{P}(A \mid F_3)\,\mathbb{P}(F_3) + \mathbb{P}(A \mid F_4)\,\mathbb{P}(F_4)\end{aligned}$$

Therefore, we can determine $\mathbb{P}(A)$ if we know:

1. The unconditional probability of each F_i
2. The conditional probability of A, given each F_i.

The expressed for $\mathbb{P}(A)$ above is called the *law of total probability*

Law of Total Probability

Suppose $F_1, F_2, \ldots, F_n$ are disjoint and $\bigcup_{k=1}^{n} F_k = \Omega$. Then,

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A \mid F_i)\,\mathbb{P}(F_i)$$

Bayes Rule

Suppose we have a partition of Ω: $F_1, F_2, \ldots, F_n$.

We would like to calculate $\mathbb{P}(F_i \mid A)$ for all $i = 1, \ldots, n$. But we only have the following information:

1. $\mathbb{P}(A \mid F_i)$ for all $i = 1, \ldots, n$
2. The unconditional probabilities, $\mathbb{P}(F_i)$ for $i = 1, \ldots, n$.

How can we do this?

Bayes Rule

How can we do this?

$$\begin{aligned}\mathbb{P}(F_i \mid A) &= \frac{\mathbb{P}(F_i \cap A)}{\mathbb{P}(A)} \\ &= \frac{\mathbb{P}(A \cap F_i)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A \mid F_i)\,\mathbb{P}(F_i)}{\mathbb{P}(A)} \\ &= \frac{\mathbb{P}(A \mid F_i)\,\mathbb{P}(F_i)}{\sum_{i=1}^{n} \mathbb{P}(A \mid F_i)\,\mathbb{P}(F_i)} \qquad (*)\end{aligned}$$

$(*)$ is called *Bayes Rule*

Bayes Rule

Bayes Rule: Suppose $F_1, F_2, \ldots, F_n$ form a partition of Ω. Then for each $i = 1, \ldots, n$:

$$\mathbb{P}(F_i \mid A) = \frac{\mathbb{P}(A \mid F_i)\,\mathbb{P}(F_i)}{\sum_{i=1}^{n} \mathbb{P}(A \mid F_i)\,\mathbb{P}(F_i)}$$

Bayes Rule: Example

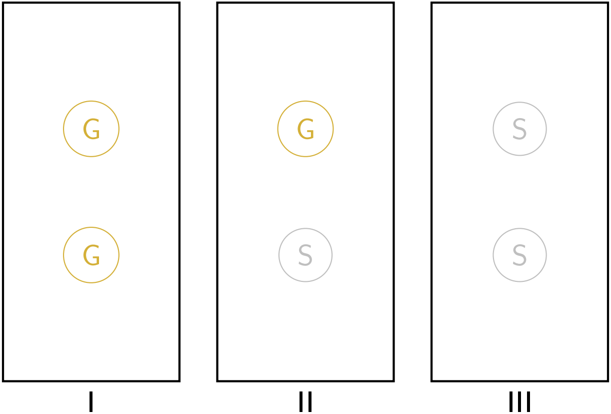
Suppose we have three drawers.

- Drawer I contains 2 Gold Coins.
- Drawer II contains 1 Gold Coin and 1 Silver Coin.
- Drawer III contains 2 Silver Coins.

I draw a drawer at random. From that drawer, I randomly draw a coin. I give you the coin, and you find that it is a Gold Coin.

What is the probability that I drew this coin from Drawer I?

Bayes Rule: Example



Bayes Rule: Example

Let D_1, D_2, and D_3 denote the events that I chose drawers I, II, and III, respectively. Let G and S denote the events that I chose a Gold and Silver coin, respectively. We wish to calculate:

$$\mathbb{P}(D_1 \mid G)$$

Since the drawer was chosen randomly, we assume that $\mathbb{P}(D_1) = \mathbb{P}(D_2) = \mathbb{P}(D_3) = 1/3$. We also know the conditional probabilities:

- $\mathbb{P}(G \mid D_1) = 1$
- $\mathbb{P}(G \mid D_2) = 1/2$
- $\mathbb{P}(D \mid D_3) = 0$

Bayes Rule: Example

Calculate using Bayes Rule:

$$\mathbb{P}(D_1 \mid G) = \frac{\mathbb{P}(G \mid D_1)\,\mathbb{P}(D_1)}{\mathbb{P}(G \mid D_1)\,\mathbb{P}(D_1) + \mathbb{P}(G \mid D_2)\,\mathbb{P}(D_2) + \mathbb{P}(G \mid D_3)\,\mathbb{P}(D_3)}$$

$$= \frac{(1)\frac{1}{3}}{(1)\frac{1}{3} + \frac{1}{2}\left(\frac{1}{3}\right) + (0)\frac{1}{3}} = \frac{\frac{1}{3}}{\frac{1}{3} + \frac{1}{6}} = \frac{2}{3}$$

Thus, it is more likely that I picked the coin from Drawer I, if we know that the coin is Gold.

Independence

Suppose A_1 and A_2 are two events in our sample space. Sometimes, our feelings about the likelihood of A_1 will **not** change if we know that A_2 has occurred. Mathematically speaking:

$$\mathbb{P}(A_1 \mid A_2) = \mathbb{P}(A_1)$$

In this case, the events A_1 and A_2 are said to be *independent*

Independence

Suppose A_1 and A_2 are two events in our sample space. Sometimes, our feelings about the likelihood of A_1 will **not** change if we know that A_2 has occurred. Mathematically speaking:

$$\mathbb{P}(A_1 \mid A_2) = \mathbb{P}(A_1)$$

In this case, the events A_1 and A_2 are said to be *independent*

Note that

$$\mathbb{P}(A_1 \mid A_2) = \mathbb{P}(A_1) \Leftrightarrow \frac{\mathbb{P}(A_1 \cap A_2)}{\mathbb{P}(A_2)} = \mathbb{P}(A_1) \Leftrightarrow \mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1)\,\mathbb{P}(A_2)$$

this is traditionally given as the definition of independence for events A_1 and A_2/

Independence

Definition: Two events A_1 and A_2 are said to be *independent* if

$$\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1)\,\mathbb{P}(A_2)$$

Often, we want to associate elements in our sample space with numerical outcomes.

Example: Perhaps we are interested in the number of heads obtained in two flips of a fair coin. Our sample space for this experiment was given earlier as

$$\Omega = \{\text{HH}, \text{HT}, \text{TH}, \text{TT}\}$$

Each element of Ω is associated with some number of heads:

- HH $\to 2$
- HT $\to 1$
- TH $\to 1$
- TT $\to 0$

The number of heads obtained is said to be a *random variable* because it maps elements of the sample space to the real number line.

Random Variables

In general, a random variable X is a measurable function $X : \Omega \to \mathbb{R}$.

Measurable:

- For any subset of the form $(-\infty, x]$. Define:

$$C_x = X^{-1}\Big((-\infty, x]\Big) = \{\omega \in \Omega \mid X(\omega) \in (-\infty, x]\} \subset \Omega.$$

- We must be able to calculate $\mathbb{P}\left(C_x\right)$ for any $x \in \mathbb{R}$.

Random Variables

Notation: For any random variable, X, and any $A \subset \mathbb{R}$, we typically define

$$\mathbb{P}(X \in A) = \mathbb{P}\left(X^{-1}(A)\right) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) \in A\})$$

If A is a singleton set, $A = \{a\}$, it is common to write:

$$\mathbb{P}(X = a) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) = A\})$$

We can interpret this as the "the probability that the variable X assumes a value in A"

Distribution Function of a Random Variable

Based on our assumptions about X, we must be able to define $\mathbb{P}(X \leq x)$ for all $x \in \mathbb{R}$. We can think of this probability as a function of x. Define

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

F_X is called the *Distribution Function* of X. Every random variable has a distribution function.

Distribution Function of a Random Variable

Define

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

F_X is called the *Distribution Function* of X. Every random variable has a distribution function.

It can be shown that F_X has the following properties:

1. $\lim_{x\to\infty} F_X(x) = 1$
2. $\lim_{x\to-\infty} F_X(x) = 0$
3. $\lim_{x\to a^+} F_X(x) = F_X(a)$ (right-continuity)
4. F_X is nondecreasing.

Distribution Function of a Random Variable

The distribution function plays a very important role in probability theory:

- Random variables are uniquely identified by their distribution function.
- Distribution functions assign probabilities to events of interest—They define the function $\mathbb{P}$ at all subsets of Ω that are of interest to us.

Some families distribution functions are used so often they are given special names.

Distribution Functions: Example

The following is a valid distribution function:

$$F_X(x) = \begin{cases} 1 - e^{-x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

1. $\lim_{x \to \infty} F_X(x) = 1$
2. $\lim_{x \to -\infty} F_X(x) = 0$
3. $\lim_{x \to a^+} F_X(x) = F_X(a)$ (right-continuity)
4. F_X is nondecreasing.

Discrete Random Variables

A random variable that assumes a positive probability on a countable number of points is called a *discrete* random variable.

Example: $X =$ number of heads in two tosses of a fair coin.

- Since the coin is fair, all 4 outcomes in the sample space are equally likely: $\{\mathrm{HH}, \mathrm{HT}, \mathrm{TH}, \mathrm{HH}\}$
- Then, we have

$$\mathbb{P}(X = 2) = \tfrac{1}{4}$$
$$\mathbb{P}(X = 1) = \tfrac{1}{2}$$
$$\mathbb{P}(X = 0) = \tfrac{1}{4}$$

Discrete Random Variables: Exercise

Since X is a random variable, it must have a distribution function. How would we sketch the distribution function of X?

Discrete Random Variables: Exercise

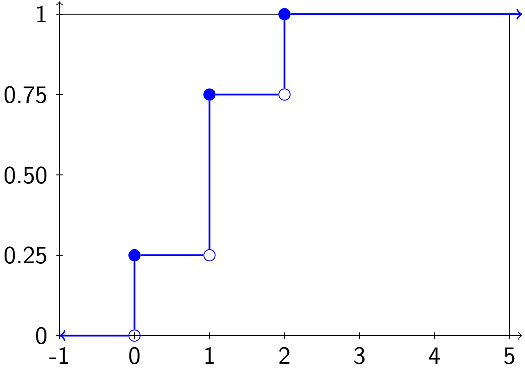
Since X is a random variable, it must have a distribution function. How would we sketch the distribution function of X?

We need to calculate $\mathbb{P}(X \leq x)$ for all values of x.

$$\mathbb{P}(X = 2) = \frac{1}{4}$$
$$\mathbb{P}(X = 1) = \frac{1}{2}$$
$$\mathbb{P}(X = 0) = \frac{1}{4}$$

Discrete Random Variables: Exercise

Figure: Sketch of Distribution Function for X



Note: We can recover the probabilities: $\mathbb{P}(X = k)$, $k = 0, 1, 2$ by calculating the size of the "jump" at k.

Discrete Random Variables: Formally Defined

Definition: A random variable is said to be *discrete* if there is some countable set $A \subset \mathbb{R}$ such that $\mathbb{P}(X \in A) = 1$.

A is defined to be the set of all points at which x has positive probability:

$$A = \{k \in \mathbb{R} \mid \mathbb{P}(X = k) > 0\}$$

The function $\mathbb{P}(X = k)$ is called the *Probability Mass Function* (PMF) of the discrete random variable. From the definition above, must have:

$$\sum_{k \in A} \mathbb{P}(X = k) = 1$$

PMFs allow us to calculate probabilities for subsets of our sample space. In this way,

Sidebar: Combinations

The number of ways to select k objects from a group of n is

$$\binom{n}{k} = \frac{n!}{(n-k)!k!}, \quad 0 \leq k \leq n$$

where $m! = m(m-1)(m-2) \times \cdots \times (2)(1)$.

Binomial Distribution

Generalizing our two-flip coin example:

- Suppose we have a biased coin, for which the probability of heads on a particular flip is p.
- Suppose we flip the coin n times, and each flip is independent of all other flips.
- Let $X =$ number of heads in n flips.

How could we calculate $\mathbb{P}(X = k)$, for $k = 0, 1, 2, \ldots, n$?

Binomial Distribution

How could we calculate $\mathbb{P}(X = k)$, for $k = 0, 1, 2, \ldots, n$?

- If the probability of heads on a particular flip is p, then the probability of tails is $1 - p$.
- There are $\binom{n}{k}$ possible arrangements of k heads in n flips.
- The probability of observing any one of these arrangements is $p^k(1-p)^{n-k}$—since there are k heads, $n - k$ tails, and all flips are independent.

Binomial Distribution

How could we calculate $\mathbb{P}(X = k)$, for $k = 0, 1, 2, \ldots, n$?

- If the probability of heads on a particular flip is p, then the probability of tails is $1 - p$.
- There are $\binom{n}{k}$ possible arrangements of k heads in n flips.
- The probability of observing any one of these arrangements is $p^k(1-p)^{n-k}$—since there are k heads, $n - k$ tails, and all flips are independent.

Therefore,

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Binomial Distribution

Let X be a random variable representing the number of successes in n independent trials, each having probability of success p. Then X is said to be a *Binomial* Random Variable, and

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

for $k = 0, 1, 2, \ldots, n$

Geometric Distribution

Suppose we flip a coin repeatedly. The probability of heads on any particular flip is p. The flips are independent.

- Let X be the number of trials until the first head appears.

Geometric Distribution

Suppose we flip a coin repeatedly. The probability of heads on any particular flip is p. The flips are independent.

Let X be the number of trials until the first head appears.

- In this case, X could take infinitely many values.
- Let's calculate $\mathbb{P}(X = k)$ for $k = 1, 2, 3, \dots$

Geometric Distribution

Suppose we flip a coin repeatedly. The probability of heads on any particular flip is p. The flips are independent.

Let X be the number of trials until the first head appears.

To obtain $X = k$, we need $k - 1$ tails followed by 1 head. Since the flips are independent, this probability is given by

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

Geometric Distribution

Definition: Consider a sequence of independent trials each having probability p of success. Let X denote the number of trials until the first success. X is called a *geometric* random variable, and

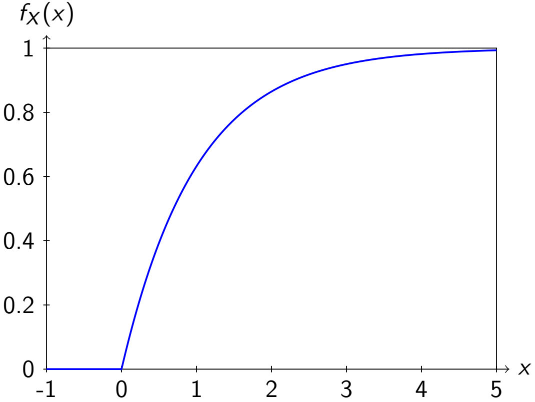
$$\mathbb{P}(X = k) = p(1-p)^{k-1}, \quad k = 1, 2, 3, \ldots$$

Continuous Distributions

Earlier, we claimed the following was a valid distribution function:

$$F_X(x) = \begin{cases} 1 - e^{-x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

Figure: Plot of $F_X(x) = \begin{cases} 1 - e^{-x} & x > 0 \\ 0 & x \leq 0 \end{cases}$



Continuous Distributions

This looks much different from the distribution function we sketched for the two-coin flip experiment!

- In the discrete case, we were table to recover $\mathbb{P}(X = x)$ by measuring the "jump size" of the distribution function at the point x.
- But this distribution function has no jumps...

Continuous Distributions

This looks much different from the distribution function we sketched for the two-coin flip experiment!

- In the discrete case, we were table to recover $\mathbb{P}(X = x)$ by measuring the "jump size" of the distribution function at the point x.
- But this distribution function has no jumps...
- In fact, the function above tells us that $\mathbb{P}(X \in (0, \infty)) = 1$, and for any $a < \infty$, $\mathbb{P}(X \in (0, a)) < 1$. Thus, there is no countable set A for which $\mathbb{P}(X \in A) = 1$.

Therefore, X is not discrete.

Continuous Distributions

Therefore, X is not discrete. In such cases, it makes more sense to consider probability *per-unit length*. Define

$$
\begin{aligned}
f_X(x) &= \lim_{\epsilon \to 0} \frac{\mathbb{P}\left(x - \epsilon < X < x + \epsilon\right)}{2\epsilon} \\
&= \lim_{\epsilon \to 0} \frac{\mathbb{P}\left(X < x + \epsilon\right) - \mathbb{P}\left(X < x - \epsilon\right)}{2\epsilon} \\
&= \lim_{\epsilon \to 0} \frac{F_X(x + \epsilon) - F_X(x - \epsilon)}{2\epsilon} \\
&= \frac{dF_X}{dx}
\end{aligned}
$$

Continuous Distributions

Therefore, X is not discrete. In such cases, it makes more sense to consider probability *per-unit length*. Define

$$\begin{aligned} f_X(x) &= \lim_{\epsilon \to 0} \frac{\mathbb{P}(x - \epsilon < X < x + \epsilon)}{2\epsilon} \\ &= \lim_{\epsilon \to 0} \frac{\mathbb{P}(X < x + \epsilon) - \mathbb{P}(X < x - \epsilon)}{2\epsilon} \\ &= \lim_{\epsilon \to 0} \frac{F_X(x + \epsilon) - F_X(x - \epsilon)}{2\epsilon} \\ &= \frac{dF_X}{dx} \end{aligned}$$

$f_X(x)$ is called the *probability density function* (PDF) of a random variable X.

Continuous Distributions

In our example the distribution function is:

$$F_X(x) = \begin{cases} 1 - e^{-x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

so the density function is:

$$f_X(x) = \begin{cases} e^{-x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

Continuous Distributions

Definition: A random variable X is said to be *absolutely continuous* if X has a density function which is nonzero on some subset of $\mathbb{R}$.

Not all random variables have nonzero density functions. For example, the distribution functions for discrete random variables are step functions. Thus, the PDFs are zero everywhere.

Since distribution functions are nondecreasing, PDFs are always nonnegative.

For any set $A\mathbb{R}$, we can determine $\mathbb{P}(X \in A)$ by integrating the PDF:

$$\mathbb{P}(X \in A) = \int_A f(x)dx$$

Probabilities are interpreted as the area under f_X within A.

Density Functions: Interpretation

We have seen that PDFs and PMFs are related to the distribution function of a random variable.

- For absolutely continuous random variables:

$$f_X(x) = F_X'(x) \quad \text{and} \quad F_X(x) = \int_{-\infty}^{x} f_X(x)dx$$

  PDFs do not uniquely define a random variable's distribution since they could be changed at countable number of points and still result in the same distribution function.

Density Functions: Interpretation

We have seen that PDFs and PMFs are related to the distribution function of a random variable.

- For absolutely continuous random variables:

$$f_X(x) = F_X'(x) \quad \text{and} \quad F_X(x) = \int_{-\infty}^{x} f_X(x)dx$$

  PDFs do not uniquely define a random variable's distribution since they could be changed at countable number of points and still result in the same distribution function.

However, changing f_X at a countable number of points would also not affect how f_X is used to calculate probabilities. Therefore, we often use f_X to identify distributions of random variables.

Exponential Distribution

The distribution function for the exponential distribution is given by

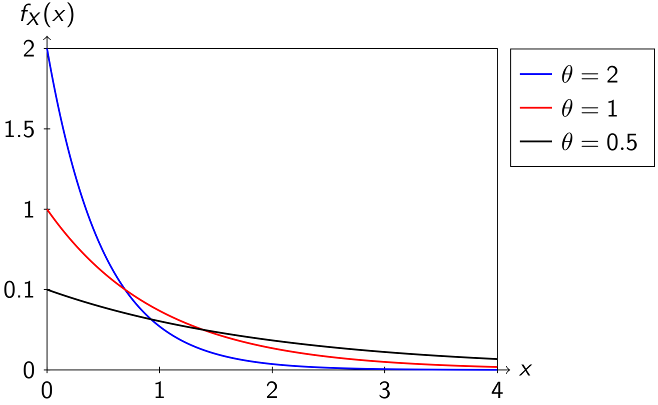
$$F_X(x) = \begin{cases} 1 - e^{-x/\theta} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

This implies a PDF of

$$f_X(x) = \begin{cases} \frac{1}{\theta} e^{-x/\theta} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

Exponential Distributions

Figure: Plot of Exponential PDF for Various θ



Example

Suppose X is exponential with PDF

$$f_X(x) = e^{-x}, \quad x > 0$$

Calculate $\mathbb{P}\left(2 < X < 3\right)$

Example

Suppose X is exponential with PDF

$$f_X(x) = e^{-x}, \quad x > 0$$

Calculate $\mathbb{P}(2 < X < 3)$

$$\mathbb{P}(2 < X < 3) = \int_2^3 e^{-x} dx = -e^{-x}\Big|_2^3$$
$$\approx 0.08555$$

Normal Distributions

The normal distribution is one of the most popular and frequently-used distributions in statistics. We usually write $X \sim \mathrm{N}\left(\mu, \sigma^2\right)$ to indicate that X has a normal distribution with mean μ and variance σ^2

- Mean: Center of the distribution of X.
- Variance: A measure of spread around the mean.

Normal Distributions

The normal distribution is one of the most popular and frequently-used distributions in statistics. We usually write $X \sim \mathrm{N}\left(\mu, \sigma^2\right)$ to indicate that X has a normal distribution with mean μ and variance σ^2

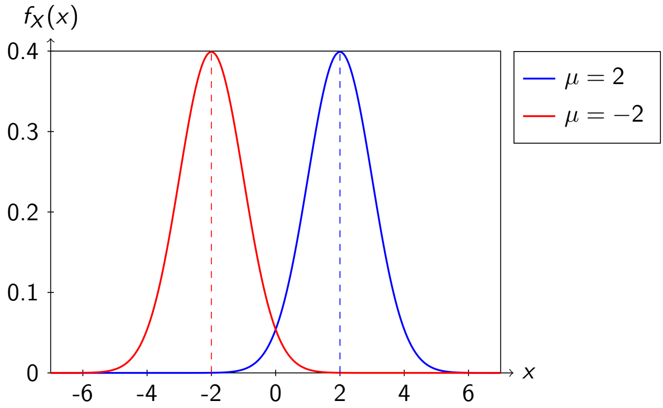
- Mean: Center of the distribution of X.
- Variance: A measure of spread around the mean.

The PDF for X is

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

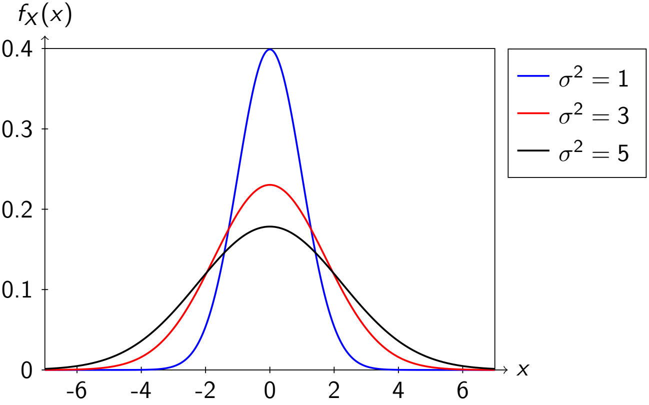
Normal Distributions

Figure: Plot of Normal PDF for Various μ



Normal Distributions

Figure: Plot of Normal PDF for Various σ^2



Connection to Statistics

In statistics, we observe lots of experimental output. For example, suppose $x_1, x_2, \ldots, x_n$ are data points from an experiment.

We assume that $x_1, x_2, \ldots, x_n$ are subject to some sampling error—if we repeat the same experiment again, we'll probably get different data points, for unexplainable reasons.

We model this phenomenon by assuming $x_1, \ldots, x_n$ are observed values from some random variables $X_1, \ldots, X_n$.

Connection to Statistics

In statistics, we observe lots of experimental output. For example, suppose $x_1, x_2, \dots, x_n$ are data points from an experiment.

We assume that $x_1, x_2, \dots, x_n$ are subject to some sampling error—if we repeat the same experiment again, we'll probably get different data points, for unexplainable reasons.

We model this phenomenon by assuming $x_1, \dots, x_n$ are observed values from some random variables $X_1, \dots, X_n$.

We often won't know the exact distribution of $X_1, \dots, X_n$.

- We want to use the information in the observed data $x_1, \dots, x_n$ to make inference about the underlying distribution of $X_1, \dots, X_n$.

Connection to Statistics

We want to use the information in the observed data $x_1, \ldots, x_n$ to make inference about the underlying distribution of $X_1, \ldots, X_n$.

Example: You toss a coin 100 times. Assume the tosses are independent and the coin lands heads with probability p on each toss.

- In $n = 100$ tosses, you observe 86 heads.
- What can you infer about the value of p?

Connection to Statistics

Let X be the number of heads in the n flips. The probability mass function of X is

$$f_X(x \mid p) = \mathbb{P}(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

Connection to Statistics

$$f_X(x \mid p) = \mathbb{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

Maximizing f_X with respect to p gives us the value of p that makes our observed data "most likely".

$$f_X(x \mid p) = \mathbb{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

Maximizing f_X with respect to p gives us the value of p that makes our observed data "most likely". Can show that the value of p that maximizes f_X is:

$$p^\star = \frac{x}{n}$$

So we would estimate $p \approx 86/100$. This is the basis for Maximum Likelihood Estimation.

Further Topics

Additional Topics:

- Linear Regression: $y = \beta_0 + \beta_1 x + \epsilon$
- Hypothesis Testing, p-values
- Confidence Intervals
- Mathematical Expectation

Questions?