Lecture Notes for Math 372: Elementary Probability and Statistics

Last updated: May 6, 2025

Probability theory has a right and a left hand. On the right is the rigorous foundational work using the tools of measure theory. The left hand "thinks probabilistically," reduces problems to gambling situations, coin-tossing, models of a physical particle. – Leo Breiman

Contents

About these notes

These lecture notes were prepared by Max Hill for a 16-week course on probability and statistics (MATH 372) at University of Hawaii at Manoa in Spring 2025. The textbook used is *Probability and Statistics for Engineering and the Sciences* (9th edition) by Jay L. Devore. In addition to this text, the material in these lecture notes is also based on a number of other sources as well, including:

- Milton and Arndold's *Introduction to Probability and Statistics* (2nd edition)
- D. Zeilberger's lecture notes `https://sites.math.rutgers.edu/~zeilberg/math477/`
- Unpublished lecture notes from E. Gross, R. Willett, J. Kim, and B. Xie.
- My notes from when I took probability with W. Peterson in 2012.
- L. Breiman's *Probability* (1992).
- K.L. Chung's wonderful textbook *Elementary Probability Theory with Stochastic Processes*, (1975).
- H. Pishro-Nik's online textbook "Introduction to Probability, Statistics, and Random Processes"
- "Introduction to Probability" by D. Anderson, T. Seppalainen, and B. Valko.

0 Tentative course outline

This course is a problem-oriented introduction to the basic concepts of probability and statistics, providing a foundation for applications and further study.

- **Weeks 1-2:** Introduction to probability theory
  - Experiments, events, sets, probabilities, random variables. Equally likely outcomes, counting techniques. Conditional probability. Independence. Bayes' theorem. (Sections: 2.1-2.5)
- **Weeks 3-5:** Random variables
  - Discrete random variables (1.5 weeks): Expected values, mean, variance, binomial distribution, Poisson distribution. Moment generating functions. (Sections: 3.1-3.6)
  - Continuous Random variables (1.5 weeks): Uniform, exponential, gamma, and normal distributions. Intuitive treatment of the Poisson process and development of the relationship with gamma distributions. (Sections: 4.1-4.4)
- **Weeks 6-7:** Multivariate distributions
  - Calculation of probability, covariance, correlation, marginals, conditions. Distributions of sums of random variables and sampling distributions. Central limit theorem. (Sections: 1.1, 1.3, 1.4, 5.1-5.7)
- **Week 8:** Catch-up, review, and midterm at the end of the week (Friday March 7).
- **Week 9:** Introduction to statistical estimation
  - Point and confidence interval estimation. Maximum likelihood, optimal, and unbiased estimators. Examples. (Sections 6.1, 6.2)
- **Weeks 10-12:** Large sample inference
  - Estimation (1.5 weeks): Types and comparison of estimators; sampling distributions for means/proportions, and their use in large sample estimation; sample size. (Sections 7.1, 7.2)
  - Hypothesis testing (1.5 weeks): Components of a test; signficance and power; p-values; large-sample tests for means and proportions (Sections: 8.1-8.4)
- **Week 13:** Small sample inference
  - t-distribution, with applications to small sample estimation and testing; χ^2 and F distributions, with applications to inference about variances (Sections: 7.3, 7.4, 8.3)
- **Weeks 14-16:** Regression and χ^2 tests
  - Regression (1.5 weeks): Least squares, correlation coefficient, inference (Sections: 12.1-12.5)
  - χ^2 tests: multinomial dsitributions, contingency tables, goodness-of-fit (Sections: 14.1-14.3)

1 2025-01-12 | Week 1 | Class 1

- give syllabus
- do activity with why you're in this course
- do the pirates worksheet

2 2025-01-15 | Week 1 | Class 2

Tell the students to read chapters 2.1 - 2.5

Example 1 (Powerball Lottery)**.** The a ticket for the powerball lottery costs $2. There are two outcomes:

- You win $300,000,000.
- You don't win any money.

The probability of winning is approximately $\frac{1}{300,000,000}$. Let X be the net payoff from the game, in dollars:

- If you lose, then $X = -2$, since you had to pay $2 to play the game.
- If you win, then your net payoff is $X = 299,999,998$.

What is the expected value of X?

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{P}[\text{win}] \cdot (-2) + \mathbb{P}[\text{lose}] \cdot (299,999,998) \\ &= \frac{299,999,999}{300,000,000}(-2) + \frac{1}{300,000,000}(299,999,999) \\ &= -1.\end{aligned}$$

Conclusion: you "expect" to lose $1 every time you play. Similarly, if you play twice, you expet to lose $2. If you play 10 times, you expect to lose $10. Etc. This property is called *linearity* of expectation.

End of Example 1. □

- Go over problem 4, parts (b) and (c). This is the problem where you look ath te expected value of 2 or 120 dice.
- Go over problem 6.
- I'll put problems 7 and 8 on the first homework.
- "Elementary" doesn't mean easy.

2.1 Sample space

An *experiment* is an activity or process whose outcome is subject to uncertainty. Examples include flipping a coin, rolling a dice, measuring the size of a wave, or the amount of rainfall. WE USUALLY DENOTE RANDOM QUANTITIES WITH CAPITAL LETTERS, LIKE 'X'. (We also tend to denote *sets* with capital letters, so ask if you get confused).

The *sample space* of an experiment is the *set* of all possible outcomes.

Example 2 (Sample space)**.**

- If I roll dice, the sample space is
$$S = \{1, 2, 3, 4, 5, 6\}$$
- If I flip a coin, the sample space is
$$S = \{T, H\}$$
- The amount of daily rainfall is
$$S = \{x \in \mathbb{R} : x \geq 0\}$$

- You've got an urn filled with 300,000,000 balls. Exactly one ball is made of gold. Draw a ball out at random. If it's not the gold ball, put it back and keep repeating until you get the gold ball. Once you get the gold ball, you are done. The output of this experiment is *the number of times you drew a ball from the urn*. The sample space is
$$S = \{1, 2, 3, 4, \ldots\}.$$

The last two examples show that the sample space need not be finite.

End of Example 2. □

The *elements* of S are called outcomes. A specified collection of outcomes is called an event. So we think of an event as being a *subset* of the sample space:

$$\text{an event} = \text{a set of outcomes} = \text{a subset of } S.$$

Example 3 (Rolling a dice)**.** To illustrate, suppose we roll a dice. Let X be the value of the dice roll. We know the sample space of the experiment is $S = \{1, 2, 3, 4, 5, 6\}$. Some possible events, along with the sets we identify them with, are shown below:

- "the dice roll is even" ..... $\{2, 4, 6\}$
- "the dice roll is at least 3" ..... $\{3, 4, 5, 6\}$
- "the dice roll equals 1" ..... $\{1\}$

It's common to represent events using square brackets like this: [event description]. For example, the first event above could also be written as:

$$[X \text{ is even}] \quad \text{or} \quad [X \in \{2, 4, 6\}].$$

And we might write the other two events as

$$[X \geq 3] \quad \text{and} \quad [X = 1].$$

End of Example 3. □

Example 4 (Rolling two dice)**.** When rolling a red and a blue dice, the sample space consists of 36 possible outcomes:

1 1	1 2	1 3	1 4	1 5	1 6
2 1	2 2	2 3	2 4	2 5	2 6
3 1	3 2	3 3	3 4	3 5	3 6
4 1	4 2	4 3	4 4	4 5	4 6
5 1	5 2	5 3	5 4	5 5	5 6
6 1	6 2	6 3	6 4	6 5	6 6

so we can write the sample space as:

$$\begin{aligned} S &= \{(x, y) : x, y \in \{1, 2, 3, 4, 5, 6\}\} \\ &= \{(1, 1), (1, 2), \ldots, (6, 6)\} \end{aligned}$$

The event that the dice are equal is

$$[\text{dice are equal}] = \{(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6)\}$$

We could have other events too, like that the dice sum to 3:

$$[\text{dice sum to } 4] = \{(1,3),(2,2),(3,1)\}$$

All of the outcome pairs are equally likely, so we can compute the probabilities of by counting entires in the table. Let

$$Z = \text{the sum of the red dice and the blue dice}$$

By counting entries in our table, we see that

$$\mathbb{P}[Z=4] = \frac{3}{36}.$$

Similarly,

$$\mathbb{P}[Z=7] = \frac{6}{36} = \frac{1}{6}$$

and

$$\mathbb{P}[Z \leq 5] = \frac{10}{36}.$$

End of Example 4. □

The previous example illustrates the critically imporant idea of **enumeration:**

> If your sample space is finite and consists of *equally likely outcomes*, then you can compute lots of probabilities easily by listing outcomes and counting them. To be prcecise, for any event A,
>
> $$\mathbb{P}[A] = \frac{|A|}{|S|} = \frac{\#\text{ of outcomes in } A}{\#\text{ of outcomes in } S}$$

I already had you do a bunch of these problems on Monday.
For the rainfall example,

$$[\text{between 1 and 2 inches of rain}] = \{x \in \mathbb{R} : 1 \leq x \leq 2\}.$$

3 2025-01-17 | Week 1 | Class 3

3.1 Cool coin flipping experiment

Example 5 (Cool coin flipping experiment)**.** Consider the following experiment:

1. Flip a coin.
2. If the coin is tails, go back to step 1. Otherwise, stop.

In other words, we flip a coin repeatedly until we get a heads.
Output: the number of times we flipped the coin.

This experiment is very similar to the example from last lecture in which we had an urn with 300,000,000 balls, exactly one of which was gold.

The sample space is

$$S = \{1, 2, 3, 4, \ldots\}$$

since there is no limit to the number of times we might have to flip the coin before getting a heads!

Let X be the number of times we flipped the coin. Then

$$\mathbb{P}[X = 1] = \mathbb{P}[\text{first coin flip was heads}] = \frac{1}{2}$$

and

$$\mathbb{P}[X = 2] = \mathbb{P}[\text{first flip was tails and second was heads}] = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}.$$

Similarly, for any $n = 1, 2, 3, 4, \ldots$, we have

$$\mathbb{P}[X = n] = \left(\frac{1}{2}\right)^n \tag{1}$$

How many times do we expect to flip the coin? I mean, what is $\mathbb{E}[X]$?

We use the formula

$$\mathbb{E}[X] = \sum_{n=1}^{\infty} n\mathbb{P}[X = n]$$

which gives

$$\begin{aligned}
\mathbb{E}[X] =& \sum_{n=1}^{\infty} n\left(\frac{1}{2}\right)^n \\
=& \frac{1}{2} + 2\left(\frac{1}{2}\right)^2 + 3\left(\frac{1}{2}\right)^3 + 4\left(\frac{1}{2}\right)^4 + \ldots \\
=& \frac{1}{2} + 2 \cdot \frac{1}{4} + 3 \cdot \frac{1}{8} + 4 \cdot \frac{1}{16} + \ldots \\
=& \frac{1}{2} + \left(\frac{1}{4} + \frac{1}{4}\right) + \left(\frac{1}{8} + \frac{1}{8} + \frac{1}{8}\right) + \left(\frac{1}{16} + \frac{1}{16} + \frac{1}{16} + \frac{1}{16}\right) + \ldots \\
=& \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32} + \ldots \\
&\quad + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32} + \ldots \\
&\qquad + \frac{1}{8} + \frac{1}{16} + \frac{1}{32} + \ldots \\
&\qquad\quad + \frac{1}{16} + \frac{1}{32} + \ldots \\
&\qquad\qquad + \frac{1}{32} + \ldots
\end{aligned}$$

The first row adds up to 1 by the geometric series formula (which says that $\sum_{n=1}^{\infty} r^n = \frac{r}{1-r}$). The second row adds up to 1/2 (because it's 1/2 less than the first row). The third row adds up to 1/4. The fourth adds up to 1/8. etc. So

$$\mathbb{E}[X] = 1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \dots$$

and applying the geometric series formula again, we see that

$$\mathbb{E}[X] = 2.$$

In other words, we expect to flip the coin twice.

This problem was easy to state (and some people guessed the correct answer ahead of time!), but it took a lot of technical work to compute the answer. (This sort of interplay between our intuition about games of chance, and the sometimes difficult technical work needed to answer questions conclusively, is a large part of why I think this topic is so interesting.)

End of Example 5. □

3.2 Set theory

Recall: The *sample space* of an experiment is the *set* of all possible *outcomes*. An *event* is a collection of outcomes.

$$\text{an event} = \text{a set of outcomes} = \text{a subset of the sample space } S$$

Roll a red and a blue dice. The **sample space** is the set

$$S = \{(1,1), (1,2), \dots (6,6)\}$$

with 36 ordered pairs. Each ordered pair is an **outcome**. The **event** that the dice are equal is the set

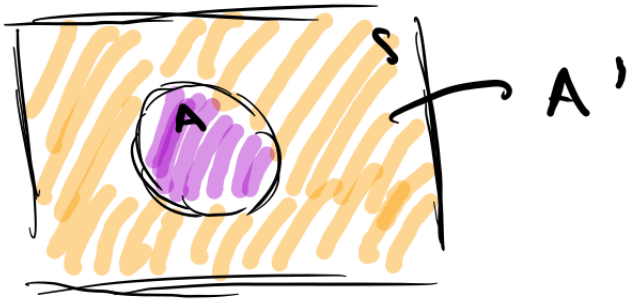
$$E = \{(1,1), (2,2), (3,3), (4,4), (5,5), (6,6)\}$$

And the enumeration principle says that

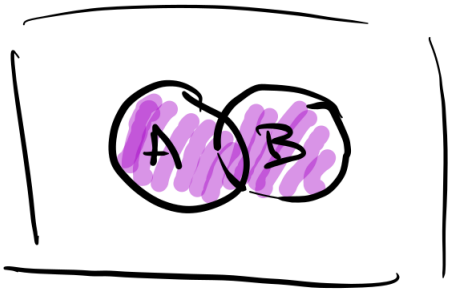
$$\mathbb{P}[E] = \frac{|E|}{|S|} = \frac{6}{36} = \frac{1}{6}.$$

Since probability theory is formalized in terms of sets, we need to have some intuition about set theory.

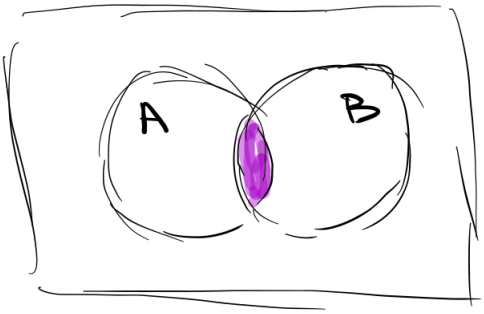
- Suppose A is an event. It's a subset of S, like this:



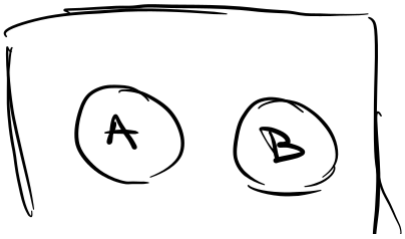
- Define the complmenent of A: A' or A^c. [see diagram above]
- Suppose B is also an event. We write $A \cup B$ to denote the set of all outcomes that are in A or B this includes outcomes that are in both):



- Do the same with $A \cap B$



- What if A and B don't overlap at all? In that case, their intersection has nothing in it!



  In this case, we say that $A \cap B$ is the "empty set", which is the set containing no elements. We use the symbol $\emptyset$ to represent the empty set; that is,

$$\emptyset := \{\}$$

  This is sometimes called the *null event*. If $A \cap B = \emptyset$, then we say that A and B are *mutually exclusive*, or *disjoint*. For example, if I roll a dice, the events

$$[\text{dice is even}] \quad \text{and} \quad [\text{dice is odd}]$$

  are mutually exclusive: they cannot happen simultaneously.

- A sequence of events $\{E_1, E_2, \ldots\}$ is said to be *pairwise disjoint* if and only if

$$E_i \cap E_j = \emptyset \quad \text{whenever } i \neq j.$$

  In other words, none of the events in E "overlap" with any other events. For an example of this, consider the experiment of Example 5, where X is the number of times we flip the coin in the experiment. Let

$$E_n = [X = n] \text{ for } n = 1, 2, 3, \ldots$$

  Then $\{E_1, E_2, \ldots\}$ is a pairwise disjoint sequence of events.

3.3 Probability Axioms

A *probability measure* $\mathbb{P}$ is a function which assigns to each event a probability. We denote the probability of an event E by

$$\mathbb{P}[E] \quad \text{or} \quad \mathbb{P}(E).$$

To be a *probability measure*, $\mathbb{P}$ must satisfy the following three axioms:

A.1 (Nonnegativity) For every event E, we have

$$\mathbb{P}[E] \geq 0.$$

A.2 (Sum-to-one) $\mathbb{P}[S] = 1$

A.3 (Countable additivity) Let $E_1, E_2, \ldots$ be an infinite sequence of events. If the sequence is pairwise disjoint, then

$$\mathbb{P}[E_1 \cup E_2 \cup \ldots] = \mathbb{P}[E_1] + \mathbb{P}[E_2] + \ldots.$$

Example 6 (Example 5 continued)**.** As an example of **A.3**, consider the experiment from Example 5, where we had

$$X = (\text{number of coin flips}).$$

For each $n = 1, 2, \ldots$, define the event E_n as

$$E_n = [X = n].$$

We already know that $\mathbb{P}[E_n] = \left(\frac{1}{2}\right)^n$ from Eq. (1). Also, you should verify for yourself that the sequence

$$\{E_1, E_2, E_3, \ldots\}$$

is pairwise disjoint.

Now, let's say we want to know the probability that X is even. Observe that

$$[X \text{ is even}] = E_2 \cup E_4 \cup E_6 \cup E_8 \cup \ldots$$

Next we will use **A.3** to compute the probability of this event:

$$\begin{aligned}
\mathbb{P}[X \text{ is even}] &= \mathbb{P}[E_2 \cup E_4 \cup E_6 \cup E_8 \cup \ldots] \\
&= \mathbb{P}[E_2] + \mathbb{P}[E_4] + \mathbb{P}[E_6] + \mathbb{P}[E_8] + \ldots \qquad (\text{by } \mathbf{A.3}) \\
&= \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^4 + \left(\frac{1}{2}\right)^6 + \left(\frac{1}{2}\right)^8 + \ldots \qquad (\text{by Eq. (1)}) \\
&= \frac{1}{4} + \left(\frac{1}{4}\right)^2 + \left(\frac{1}{4}\right)^3 + \left(\frac{1}{4}\right)^4 + \ldots \\
&= \frac{\frac{1}{4}}{1 - \frac{1}{4}} \qquad \text{by the geometric series formula } \sum_{n=1}^{\infty} r^n = \frac{r}{1-r} \\
&= \frac{1}{3}.
\end{aligned}$$

So the probability that you flip the coin an *even* number of times is only 1/3.

End of Example 6. □

4 2025-01-22 | Week 2 | Class 4

Recall, a *probability measure* $\mathbb{P}$ is a function which assigns to each possible event a probability, so that the probability of an event E is denoted $\mathbb{P}[E]$ or $\mathbb{P}(E)$. The following axioms hold:

A.1 (Nonnegativity) For every event E, we have

$$\mathbb{P}[E] \geq 0.$$

A.2 (Sum-to-one) $\mathbb{P}[S] = 1$

A.3 (Countable additivity) Let $E_1, E_2, \ldots$ be an infinite sequence of events. If the sequence is pairwise disjoint, then

$$\mathbb{P}[E_1 \cup E_2 \cup \ldots] = \mathbb{P}[E_1] + \mathbb{P}[E_2] + \ldots.$$

recall that paiwise dijoint means that there is no overlap.

Proposition 7 (Basic properties of probability measure)**.**

(i.) *(The null event has probability zero)* $\mathbb{P}[\emptyset] = 0$

(ii.) *(Finite additivity) Let $\{E_1, \ldots, E_n\}$ be a finite sequence of events. If the sequence is pairwise disjoint, then*

$$\mathbb{P}[E_1 \cup E_2 \cup \ldots \cup E_n] = \mathbb{P}[E_1] + \mathbb{P}[E_2] + \ldots + \mathbb{P}[E_n]$$

(iii.) *("With probability one, an event E either does occur or doesn't")* $\mathbb{P}[E^c] = 1 - \mathbb{P}[E]$

(iv.) *(Excision Property) If A, B are events and $A \subseteq B$, then*

$$\mathbb{P}[B \backslash A] = \mathbb{P}[B] - \mathbb{P}[A].$$

(v.) *("The particular is less likely than the general") If A, B are events and $A \subseteq B$, then $\mathbb{P}[A] \leq \mathbb{P}[B]$*

(vi.) *("Probabilities are between 0 and 1") For any event E, $\mathbb{P}[E] \in [0, 1]$*

Proof of Proposition 7. First we will prove (i.). Let $E_i = \emptyset$ for all $i = 1, 2, 3, \ldots$. Then

- $\{E_1, E_2, \ldots\}$ is a pairwise disjoint sequence.
- If $E_1 \cup E_2 \cup \ldots = \emptyset$.

Therefore,

$$\begin{aligned} \mathbb{P}[\emptyset] &= \mathbb{P}[E_1 \cup E_2 \cup \ldots] \\ &= \mathbb{P}[E_1] + \mathbb{P}[E_2] + \ldots && \text{by } \mathbf{A.3} \\ &= \mathbb{P}[\emptyset] + \mathbb{P}[\emptyset] + \ldots. \end{aligned}$$

This equation implies $\mathbb{P}[\emptyset] = 0$. We have now proved (i.).

Next we prove (ii.). Let $E_1, \ldots, E_n$ be a finite sequence of events which are pairwise disjoint. Then expand the sequence by taking $E_i = \emptyset$ for all $i \in \{n+1, n+2, \ldots\}$. Then by **A.3**,

$$\begin{aligned} \mathbb{P}\left[\bigcup_{i=1}^{n} E_i\right] & \\ &= \mathbb{P}\left[\bigcup_{i=1}^{\infty} E_i\right] \\ &= \sum_{i=1}^{\infty} \mathbb{P}[E_i] \\ &= \mathbb{P}[E_1] + \ldots \mathbb{P}[E_n] + \underbrace{\mathbb{P}[\emptyset] + \mathbb{P}[\emptyset] + \ldots}_{=0 \text{ by Proposition 7 (i.)}} \\ &= \mathbb{P}[E_1] + \ldots \mathbb{P}[E_n]. \end{aligned}$$

Next we prove (iii.). Let E be any event. Then

$$\begin{aligned} 1 &= \mathbb{P}[S] && \text{by } \mathbf{A.2} \\ &= \mathbb{P}[E \cup E^c] && \text{since } S = E \cup E^c \\ &= \mathbb{P}[E] + \mathbb{P}[E^c] && \text{by Proposition 7 (ii.), since } E \cap E^c = \emptyset. \end{aligned}$$

This proves (iii.).

Next we prove (iv.). Observe that

$$\begin{aligned} B &= (B \cap A) \cup (B \cap A^c) \\ &= A \cup B \cap A^c && \text{since } A \subseteq B. \end{aligned}$$

This is a disjoint union. Therefore by (ii.),

$$\begin{aligned} \mathbb{P}[B] &= \mathbb{P}[A] + \mathbb{P}[B \cap A^c] \\ &= \mathbb{P}[A] + \mathbb{P}[B \backslash A]. \end{aligned}$$

Rearranging terms proves (iii.). Next we prove (v.). Assume $A \subseteq B$. Then by (iv.),

$$\mathbb{P}[B \backslash A] = \mathbb{P}[B] - \mathbb{P}[A]$$

Moreover, by **A.1**, $\mathbb{P}[B \backslash A] \geq 0$. Hence

$$0 \leq \mathbb{P}[B] - \mathbb{P}[A]$$

and this implies (v.).

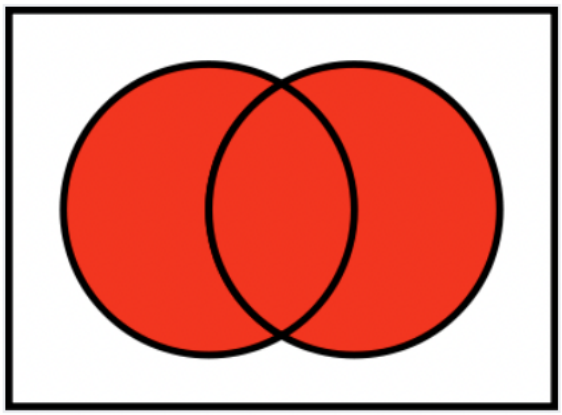
Finally, we will show (vi.). Let E be an event. Then

$$\begin{aligned} 0 \leq \mathbb{P}[E] \leq \mathbb{P}[S] \leq && && \text{by } \mathbf{A.1} \\ \leq \mathbb{P}[S] && && \text{by (v.)} \\ = 1 && && \text{by } \mathbf{A.2}. \end{aligned}$$

□

Proposition 8 (Inclusion-Exclusion Prinicple)**.** $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B]$

Idea:



Proposition 9 (De Morgan's Laws)**.** *The following equalities hold:*

$$(A \cup B)^c = A^c \cap B^c \quad \textit{and} \quad (A \cap B)^c = A^c \cap B^c.$$

Proposition 10 (Distributive Laws)**.** *The following equalities hold:*

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

and

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

4.1 Introduction to counting techniques

A *k-tuple* is an ordered list of k numbers.

Example 11 (tuples)**.** Consider the experiment in which you roll a dice 5 times in a row:

$$(1, 4, 6, 4, 4)$$

is a 5-tuple.
Problem: What is the sample space? How many elements does it have?
Idea: we will use the *the product rule for k-tuples*:

- there are 6 choices for the first entry
- 6 choices for the second entry
- $\vdots$
- 6 choices for the fifth entry

so the number of possible outcomes is

$$6 \times 6 \times 6 \times 6 \times 6 = 6^5 = 7776.$$

Problem: Let $E = [\text{All 5 dice are less than or equal to 3}]$. What is $\mathbb{P}[E]$?
By the counting principle,

$$\mathbb{P}[E] = \frac{|E|}{|S|} = \frac{\#\text{ ways that all dice are} \leq 3}{7776}$$

now we use again the product rule to count the k-tuples with all entries less than 3:

$$3 \times 3 \times 3 \times 3 \times 3 = 3^5 = 243$$

So

$$\mathbb{P}[E] = \frac{243}{7776} = \frac{1}{32}$$

End of Example 11. □

5 2025-01-24 | Week 2 | Lecture 5

5.1 St. Petersburg Paradox

The game is the following: repeatedly flip a coin until you get heads. You win 2^n dollars, where n is the number of coin flips. How much would you be willing to pay to play this game????

Let X be your (random) payoff. The sample space of X is

$$S = \{2, 4, 8, 16, 32, \ldots\}.$$

What is the expected value of X?

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{n=1}^{\infty} n\mathbb{P}[X = n] \\
&= \mathbb{P}[X = 1] + 2\mathbb{P}[X = 2] + 3\mathbb{P}[X = 3] + 4\mathbb{P}[X = 4] + 5\mathbb{P}[X = 5] + 6\mathbb{P}[X = 6] + 7\mathbb{P}[X = 7] \\
&\quad + 8\mathbb{P}[X = 8] + 9\mathbb{P}[X = 9] + \ldots \\
&= 2\mathbb{P}[X = 2] + 4\mathbb{P}[X = 4] + 8\mathbb{P}[X = 8] + 16\mathbb{P}[X = 16] + \\
&= \sum_{k=1}^{\infty} 2^k \mathbb{P}\left[X = 2^k\right]
\end{aligned}$$

where the third equality follow because $\mathbb{P}[X = i] = 0$ whenever $i \notin S$.

Next, we observe that $\mathbb{P}\left[X = 2^k\right] = \frac{1}{2^k}$ (check this for, say, $k = 1, 2, 3$).

Plugging these probabilities in, we get

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} 2^k \cdot \frac{1}{2^k} = \sum_{k=1}^{\infty} 1 = 1 + 1 + 1 + 1 + 1 + \ldots = +\infty.$$

The expected value of playing this game is positive infinity!!!!!

5.2 Counting Techniques

There are three main counting principles we'll look at today:

- multiplication principle
- permutation principle
- combination principle

Recall that a k-tuple is an ordered list of things (usually but not always numbers). A k-tuple can have repeats!

Proposition 12 (Multiplication Principle)**.** *Suppose S consist of k-tuples and there are n_1 choices for the first element, n_2 choices for the second, etc. Then there are*

$$n_1 \times n_2 \times n_k$$

possible k-tuples.

Example 13. I roll my lucky dice, flip a gold coin, then flip a silver coin. How many possible outcomes are there?

[draw a tree to illustrate the product rule]

$$6 \times 2 \times 2 = 24$$

End of Example 13. □

Example 14. How many ways are there to order the 3 letters A, B, C?
Use the multiplication principle

$$3 \times 2 \times 1$$

End of Example 14. □

Proposition 15 (Counting permutations)**.** *The total number of ways to order n distinct objects is*

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

Each ordering is called a permutation. (Note that we define $0! = 1$.)
(Permutations don't have repeats).

Definition 16 (set)**.** A *set* is an *unordered* collection of elements, all of which are *distinct.*

- $\{1, 2, 4, 5, 3\}$ is a set
- $\{1, 1, 2, 4, 5, 3\}$ is not a set

Proposition 17 (Counting subsets – this is the "combination principle")**.** *Let S be a set with n elements. The number of (unordered) subsets of size k is*

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

Example 18. Let $S = \{1, 2, 3, 4, 5, 6\}$. How many subsets of size 2 are there?

$$\binom{6}{2} = \frac{6!}{2!\cdot 4!} = \frac{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{(2 \cdot 1) \cdot (4 \cdot 3 \cdot 2 \cdot 1)} = \frac{6 \cdot 5}{2} = 15.$$

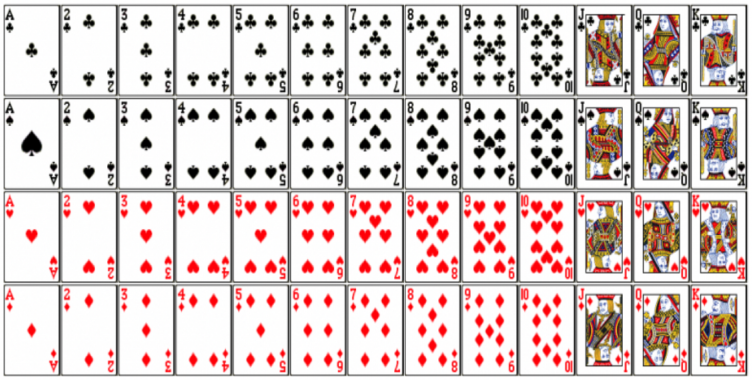
These are

$$\{1,2\}, \{1,3\}, \{1,4\}, \{1,5\}, \{1,6\}$$
$$\{2,3\}, \{2,4\}, \{2,5\}, \{2,6\}$$
$$\{3,4\}, \{3,5\}, \{3,6\}$$
$$\{4,5\}, \{4,6\}$$
$$\{5,6\}$$

note that the order doesn't matter, i.e., $\{2,1\} = \{1,2\}$ so this just counts as one, not two.

End of Example 18. □

Example 19 (Poker)**.** A poker hand consist of 5 randomly chosen cards from a standard 52-card deck.



Some questions about poker hands:

1. How many distinct poker hands are there?

$$\binom{52}{5} = \frac{52!}{5! \cdot 47!} = 2,598,960$$

2. How many ways are there to get a flush (i.e., all the same suit)?

$$4 \times \binom{13}{5} = 4 \cdot \frac{13!}{5!\,8!} = 4 \cdot \frac{13 \cdot 12 \cdot 11 \cdot 10 \cdot 9}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 5148$$

   The 4 is to choose a suit. Then we choose 5 out of 13 cards in that suit.

3. What's the probability that we are dealt a hand in which this occurs? (i.e., in which all cards are of the same suit?)

   To answer this, we use the enumeration principle:

$$\frac{\binom{4}{1} \times \binom{13}{5}}{\binom{52}{5}} = \frac{5,148}{2,598,960} \approx 0.002$$

   In other words, the probability of a flush is pretty low: about 0.2%.

End of Example 19. □

6 2025-01-27 | Week 3 | Lecture 6

6.1 Conditional probability

This lecture is based on section 2.4 in the text.

First, play craps.

Tallying the results, we saw that shooters won about 50% of the time. Is this a fair game? We will set out to answer this question. To do so, we'll need to introduce some new ideas. A key idea is that of conditional probability:

Definition 20 (Conditional Probability)**.** Let A, B be events, and assume that $\mathbb{P}[A] > 0$. Then the *conditional probability of B, given A*, denoted $\mathbb{P}[B \mid A]$, is given by the formula

$$\mathbb{P}[B \mid A] := \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[A]}.$$

[draw picture]

- Intuitively, $\mathbb{P}[B \mid A]$ is the probability of B when we know that the event H as occurred.
- The idea is that if we know that event A has occured, then the sample space becomes A, and the new event is $A \cap B$.

For example,

Example 21 (Conditional probability)**.** I roll a dice behind a screen. I tell you that I rolled an even number. What's the probability that the dice roll is a 4 or a 6?

Solution: We have $S = \{1, 2, 3, 4, 5, 6\}$, $A = \{2, 4, 6\}$, $B = \{4, 6\}$. We want to compute $\mathbb{P}[B \mid A]$. By Definition 20,

$$\mathbb{P}[B \mid A] = \frac{\mathbb{P}[AB]}{\mathbb{P}[A]} = \frac{2/6}{1/2} = \frac{2}{3}.$$

Put differently, we have three outcomes in A, which are all equally likely. Two of them (4 and 6) mean that event B occurs. So by the enumeration principle, the probability is 2/3.

End of Example 21. □

Then we worked on `worksheet-02.pdf`.

7 2025-01-29 | Week 3 | Lecture 7

7.1 Review of Set Operations

Students can contact the grader (name: Hunter) directly `hvt7hawaii.edu` with any questions.

Example 22 (Set Operations)**.** Let's review set operations. Suppose

$$A = \text{[I brought a black marker to class]} \quad \text{and} \quad B = \text{[I brought a blue marker to class]}$$

Then the following sets are defined as follows:

$$A^c = \text{[I didn't bring a black marker to class]}$$

and

$$B^c = \text{[I didn't bring a blue marker to class]}$$

and

$$A \cup B = \text{[I brought a black marker OR a blue marker, possibly both]}$$

and

$$A \cap B = \text{[I brought both a blue and a black marker]}$$

and we also have *De Morgan's Laws:*

$$(A \cup B)^c = A^c \cap B^c = \text{[I brought neither a blue nor a black marker]}$$

$$(A \cap B)^c = A^c \cup B^c = \text{[I didn't bring } \textit{both} \text{ a blue marker and a black marker (but maybe I brought one)]}$$

End of Example 22. □

7.2 Conditional Probability

For any two events A and B, conditional probability describes the probability that A happens given that we know B happened.

Definition 23 (Conditional Probability)**.** Let A, B be events, and assume that $\mathbb{P}[A] > 0$. Then the *conditional probability of B, given A*, denoted $\mathbb{P}[B \mid A]$, is given by the formula

$$\mathbb{P}[B \mid A] := \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[A]}.$$

Example 24 (Conditional Probability)**.** I roll two dice and add them up. Call this number X. Then $\mathbb{P}[X = 7] = \frac{1}{6}$, which we can see by counting up the possibilies. Now, what about if I tell you that I rolled the dice and got a sum greater than 4? Now that you have a little additional information, you can exclude the possibilities that $X = 1, 2, 3$ or $X = 4$. So with this new information, you should evaluate the probability that $X = 7$ as greater than before. That is, now we want

$$\begin{aligned}
\mathbb{P}[X = 7 | X > 4] &= \frac{\mathbb{P}[X = 7 \text{ and } X > 4]}{\mathbb{P}[X > 4]} \\
&= \frac{\mathbb{P}[X = 7]}{\mathbb{P}[X > 4]} \\
&= \frac{1/6}{30/36} \\
&= \frac{1}{5}.
\end{aligned}$$

This answer makes sense because $\frac{1}{5} > \frac{1}{6}$, consistent with our intution.

End of Example 24. □

Example 25 (Conditional probability)**.** Suppose you flip a coin 10 times and get more than 6 heads. What's the probability that you get less than 9 heads?
Solution: Let X be the number of heads. Let $A = [X < 9]$, and let $B = [X > 6]$. Then

$$\mathbb{P}[B] = \frac{\binom{10}{7} + \binom{10}{8} + \binom{10}{9} + \binom{10}{10}}{2^{10}}$$

Observe that $A \cap B = [X \in \{7, 8\}]$. Therefore

$$\mathbb{P}[A \cap B] = \frac{\binom{10}{7} + \binom{10}{8}}{2^{10}}$$

Therefore

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[AB]}{\mathbb{P}[B]} = \frac{\binom{10}{7} + \binom{10}{8}}{\binom{10}{7} + \binom{10}{8} + \binom{10}{9} + \binom{10}{10}} = \frac{15}{16}.$$

End of Example 25. □

Notation 26 (Set intersections)**.** If A, B are events, then we will sometimes write AB to denote the set $A \cap B$.

Proposition 27 (Multiplication Rule)**.** *For two events A, B,*

$$\mathbb{P}[AB] = \mathbb{P}[A]\,\mathbb{P}[B \mid A]$$

For three events A, B, C,

$$\mathbb{P}[ABC] = \mathbb{P}[A]\,\mathbb{P}[B \mid A]\,\mathbb{P}[C \mid AB]$$

This rule is useful for analyzing experiments which proceed in stages, like the following problem:

Example 28 (Urn - example of multiplication rule)**.** An urn contains 6 white balls and 9 black balls. If 4 balls are drawn at random, what is the probability that the first 2 are white and the last 2 are black?
Solution: Next time.

End of Example 28. □

8 2025-01-31 | Week 3 | Lecture 8

8.1 The Multiplication Rule for Conditional Probability

Recall, the multiplication rule for conditional probability is

$$\mathbb{P}[AB] = \mathbb{P}[A \mid B]\,\mathbb{P}[B]$$

where A and B are any events.

Example 29 (Urn - example of multiplication rule)**.** An urn contains 6 white balls and 9 black balls. If 4 balls are drawn at random, what is the probability that the first 2 are white and the last 2 are black?
Solution: Let W be the event that the first two balls are white, and B the event that the last two balls are black. Then the desired probability is

$$\mathbb{P}[W \cap B] = \mathbb{P}[B \mid W]\,\mathbb{P}[W] \tag{2}$$

Now,

$$\mathbb{P}[W] = \frac{\binom{6}{2}}{\binom{15}{2}} = \frac{1}{7}.$$

Moreover, given that W occurs, then 4 white balls and 9 black balls remain. So

$$\mathbb{P}[B \mid W] = \frac{\binom{9}{2}}{\binom{13}{2}} = \frac{6}{13}.$$

Therefore by Eq. (2),

$$\mathbb{P}[W \cap B] = \frac{1}{7} \cdot \frac{6}{13} = \frac{6}{91} \approx 0.07$$

End of Example 29. □

Example 30 (Application of multiplication rule (c.f. Example 2.27 in textbook))**.** A vampire goes to the blood bank looking to find some type O+ blood (the most delicious type). He finds four unlabeled bags of blood. Only one of the bags is O+, but he doesn't know which one. The vampire resorts to taste-testing to find the O+ bag.

(a) What is the probability that he must test at least 3 bags to find the desired type?
Solution: Let X be the number of tested bags. We want to find $\mathbb{P}[X \geq 3]$.
Let $A = [\text{first bag isn't O+}]$. Let $B = [\text{second bag isn't O+}]$

$$\begin{aligned}\mathbb{P}[X \geq 3] &= \mathbb{P}[A \cap B] \\ &= \mathbb{P}[A]\,\mathbb{P}[B \mid A] \\ &= \frac{3}{4} \cdot \frac{2}{3} \\ &= \frac{1}{2}.\end{aligned}$$

(b) What's the probability that the third bag he tests is the one containing the O+ blood?
Solution: We want $\mathbb{P}[X = 3] = \mathbb{P}[\text{third bag is O+}]$.

$$\begin{aligned}\mathbb{P}[\text{third bag is O+}] &= \mathbb{P}[\text{third bag is} \mid \text{first isn't} \cap \text{second isn't}] \times \mathbb{P}[\text{second isn't} \mid \text{first isn't}] \cdot \mathbb{P}[\text{first isn't}] \\ &= \frac{1}{2} \cdot \frac{2}{3} \cdot \frac{3}{4} \\ &= \frac{1}{4}\end{aligned}$$

End of Example 30. □

9 2025-02-03 | Week 4 | Lecture 9

If you haven't yet, please finish reading sections 2.4 and 2.5.

9.1 Law of Total Probability + Bayes' Theorem

Theorem 31 (The Law of Total Probability)**.** *Let $A_1, \ldots, A_n$ be events which partition the sample space (i.e., they are mutually exclusive and $A_1 \cup \cdots \cup A_n = S$). Then*

$$\mathbb{P}[B] = \sum_{i=1}^{n} \mathbb{P}[B \mid A_i]\,\mathbb{P}[A_i]$$

Proof. Proof by picture. □

Example 32 (Questing adventurers)**.** *[Similar to example 2.29 from text]* A large group of adventurers are given the choice of one of three quests: slay a dragon (Q_1), defeat the lich king (Q_2), or retrieve the long-lost scepter of fire in the underground kingdom of Avernus (Q_3). Out of this group, 50% of the adventurers undertake Q_1, 30% undertake Q_2, and the remaining 20% choose Q_3.

It is known that within the first year, 25% of adventurers who set out to slay a dragon become dragon food instead. Moreover, 20% of adventurers who attempt to stop the lich king end up joining the ranks of his growing undead army, and 10% of adventurers who descend into Avernus to retrieve the scepter of fire get slain by underground lizardmen.

(a) What is the probability that a randomly-selected adventurer undertakes quest (1) and gets eaten by a dragon?

Solution: In other words, we want to compute $\mathbb{P}\left[☠ \cap Q_1\right]$. Using the multiplication rule, we have

$$\mathbb{P}\left[☠ \cap Q_1\right] = \mathbb{P}\left[☠ \mid Q_1\right] \cdot \mathbb{P}[Q_1] = \frac{1}{4} \cdot \frac{1}{2} = \frac{1}{8} = .125$$

(b) What is the probability that a randomly-selected adventurer meets an untimely demise?

Solution:

$$\begin{aligned}
\mathbb{P}\left[☠\right] &= \mathbb{P}\left[(Q_1 \cap ☠) \cup (Q_2 \cap ☠) \cup (Q_3 \cap ☠)\right] \\
&= \mathbb{P}\left[Q_1 \cap ☠\right] + \mathbb{P}\left[Q_2 \cap ☠\right] + \mathbb{P}\left[Q_3 \cap ☠\right] \\
&= .125 + .06 + .02 \\
&= .205,
\end{aligned}$$

so we conclude that 20.5% of adventurers meet an untimely demise ☠ within the first year.

(c) If an adventurer meets an untimely demise, what is the probability that they undertook quest (1)? What about quests (2) and (3)?

Solution: In other words, we want to compute $\mathbb{P}\left[Q_i \mid ☠\right]$ for $i = 1, 2, 3$. For this we use the definition of conditional probabiblity. For $i = 1$, we have:

$$\begin{aligned}
\mathbb{P}\left[Q_1 \mid ☠\right] &= \frac{\mathbb{P}\left[Q_1 \cap ☠\right]}{\mathbb{P}\left[☠\right]} \\
&= \frac{.125}{.205} && \text{by parts (a) and (b)} \\
&= \frac{25}{41} \\
&\approx 61\%.
\end{aligned}$$

We conclude that 61% of adventurers who met an untimely end were done in by a dragon.

End of Example 32. □

In part (b), we used the *law of total probability*:

$$\mathbb{P}[A] = \sum_i \mathbb{P}[A \cap B_i] = \sum_i \mathbb{P}[A \mid B_i]\,\mathbb{P}[B_i],$$

(which holds as long as the sequence $B_1, B_2, \ldots$ are mutually exclusive and exhaustive). In particular, we found $\mathbb{P}\left[\text{☠}\right]$ by taking $A = \text{☠}$, along with $B_1 = Q_1$, $B_2 = Q_2$, and $B_3 = Q_3$.

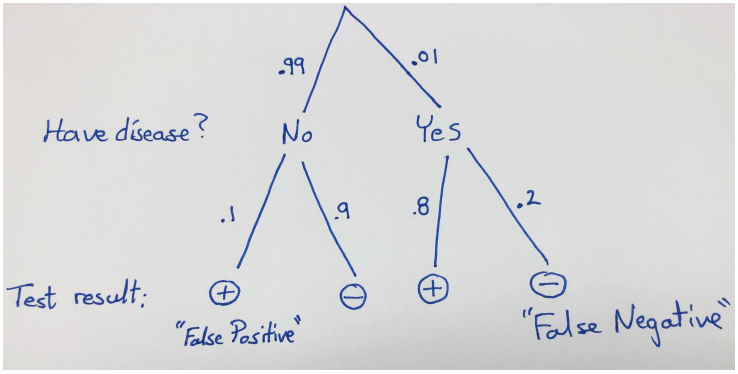
In part (c), we have used *Bayes' Theorem*, which is usually stated as

$$\mathbb{P}[B_i \mid A] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[A]} = \frac{\mathbb{P}[A \mid B_i]\,\mathbb{P}[B_i]}{\sum_k \mathbb{P}[A \mid B_k]\,\mathbb{P}[B_k]}.$$

In part (c) of Example 32, we had $A = \text{☠}$ and $B_i = Q_i$ for $i = 1, 2, 3$. We were able to interpet $\mathbb{P}\left[Q_1 \mid \text{☠}\right]$ as "the proportion of dead aventurers who were done in by a dragon". More generally, the right-hand side of Bayes' theorem suggests that we can think of $\mathbb{P}[B_i \mid A]$ as "the contribution of B_i to the total probability of A".

You can memorize the RHS of the above formula, but it should suffice to know how to use tree digrams like we did in this problem, which will work as long as you know the law of total probability and the definition of conditional probability.

Example 33 (Bayes' Theorem)**.** Prevalance of a disease in a population is 1%. A new diagnostic test is advertised as having a false positive rate of .1 and a false negative rate of .2. (In other words, 10% of positive tests are wrong, and 20% of negative tests are wrong.) Suppose you are selected for a random screening, and you test posititve. What's the chance that you have the disease?



We want $\mathbb{P}[\text{no disease} \mid \text{test positive}]$. For brevity, let's let

$$D = [\text{have disease}] \quad \text{and} \quad T^+ = [\text{test positive}].$$

In other words, we want

$$\mathbb{P}\left[D \mid T^+\right]$$

By the definition of conditional probability,

$$\mathbb{P}\left[D \mid T^+\right] = \frac{\mathbb{P}[T^+ \cap D]}{\mathbb{P}[T^+]} \tag{3}$$

We will compute the numerator and denominator separately:

- (Numerator) By the multiplication rule for conditional probability ($\mathbb{P}[A \cap B] = \mathbb{P}[A|B]\,\mathbb{P}[B]$), we have

$$\mathbb{P}\left[T^+ \cap D\right] = \mathbb{P}\left[T^+ \mid D\right]\mathbb{P}[D] = (.8)(.01) = 0.008. \tag{4}$$

  We've now computed the numerator of Eq. (3).

  Similarly, recalling that $D^c = [\text{don't have disease}]$, we could compute

$$\mathbb{P}\left[T^+ \cap D^c\right] = \mathbb{P}\left[T^+ \mid D^c\right]\mathbb{P}[D^c] = (.1)(.99) = 0.099. \tag{5}$$

- (Denominator) By the law of total probability,

$$\begin{aligned}\mathbb{P}\left[T^+\right] &= \mathbb{P}\left[T^+ \cap D\right] + \mathbb{P}\left[T^+ \cap D^c\right] \\ &= 0.008 + 0.099 \qquad \text{by Eqs. (4) and (5)}\end{aligned}$$

- Finally, having done the necessary work, we can now plug our results into Eq. (3):

$$\mathbb{P}\left[H \mid T^+\right] = \frac{0.008}{0.099 + 0.008} \approx .075$$

We conclude that if test positive in a random screening, your probability of actually having the disease is about 7.5%. Is this result surprising?

End of Example 33. □

10 2025-02-05 | Week 4 | Lecture 10

10.1 Application of the Law of Total Probability: The probability of winning at craps

Recall the rules of craps:

1. You roll two dice and add them up.
   - If the sum is $2, 3$, or 12, you lose ("craps")
   - If you roll 7 or 11, then you win ("natural")
   - Otherwise you establish "point", which is whatever number you got.
2. If you established point, then you must keep rolling until one of two events occurs:
   - You roll a 7: you lose
   - You roll your "point" number: you win.

Let's introduce some useful notation: Let

$$W = [\text{You win}]\,,$$

and for each $k = 2, 3, 4, \ldots, 12$, define

$$F_k = [\text{Your first roll is } k]\,.$$

Question: What is $\mathbb{P}[W]$?
Solution: By the *Law of Total Probability*, we have

$$\mathbb{P}[W] = \sum_{k=2}^{12} \mathbb{P}[W \mid F_k] \times \mathbb{P}[F_k] \tag{6}$$

In words,

$$\mathbb{P}[W \mid F_k] = \text{ probability of winning, given you rolled } n \text{ on the first roll}$$

k	$P[W \mid F_i]$	$\mathbb{P}[F_k]$
2	0	1/36
3	0	2/36
4	1/3	3/36
5	4/10	4/36
6	5/11	5/36
7	1	6/36
8	5/11	5/36
9	4/10	4/36
10	1/3	3/36
11	1	2/36
12	0	1/36

How do we get, say, $\mathbb{P}[W \mid F_4]$?
(Method 1: The long way) Let's enumerate the ways we can win:

4	win on the first roll after establishing point
$*\,4$	win on second roll
$**4$	win on third roll
$***\,4$	etc
$****4$	
$\ldots$	

where $*$ denotes any roll other than a 4 or a 7.

These have probabitilies

$$\begin{aligned}
\mathbb{P}[4] &= \frac{3}{36} \\
\mathbb{P}[*4] &= \frac{27}{36}\frac{3}{36} \\
\mathbb{P}[**4] &= \left(\frac{27}{36}\right)^2 \frac{3}{36} \\
\mathbb{P}[***4] &= \left(\frac{27}{36}\right)^3 \frac{3}{36} \\
&\\
\mathbb{P}[****4] &= \left(\frac{27}{36}\right)^4 \frac{3}{36} \\
&\\
&\dots
\end{aligned}$$

So

$$\begin{aligned}
\mathbb{P}[W \mid F_4] &= \frac{3}{36} + \frac{27}{36}\frac{3}{36} + \left(\frac{27}{36}\right)^2 \frac{3}{36} \\
&= \frac{3}{36}\left(1 + \frac{27}{36} + \left(\frac{27}{36}\right)^2 + \dots\right) \\
&= \frac{3}{36} \cdot \frac{1}{1 - \frac{27}{36}} \\
&= \frac{3}{9} \\
&= \frac{1}{3}.
\end{aligned}$$

(Method 2: The sort way). Given F_4, you know you're gonna keep rolling until you get a 4 or a 7, which is eventually going to happen. And whether you win or lose is determined by what you roll on that last roll. Your probability of winning is the probability of rolling a 4 on your last roll, given that your last roll is a 4 or a 7. Thus,

$$\begin{aligned}
\mathbb{P}[W \mid F_4] &= \mathbb{P}[\text{roll 4} \mid \text{roll 4 or 7}] \\
&= \frac{\mathbb{P}[(\text{roll 4}) \text{ and } (\text{roll 4 or 7})]}{\mathbb{P}[\text{roll 4 or 7}]} && \text{by def of conditional prob.} \\
&= \frac{\mathbb{P}[\text{roll 4}]}{\mathbb{P}[\text{roll 4 or 7}]} \\
&= \frac{3/36}{\frac{3}{36} + 6/36} \\
&= \frac{1}{3}.
\end{aligned}$$

Continue in this manner to get all the values of the table. Then plugging the values from the table into Eq. (6) gives

$$\mathbb{P}[W] = 0 \cdot \frac{1}{36} + 0 \cdot \frac{2}{36} + \frac{1}{3} \cdot \frac{3}{36} + \frac{4}{10} \cdot \frac{4}{36} + \frac{5}{11} \cdot \frac{5}{36} + 1 \cdot \frac{6}{36} + \ldots + 0 \cdot \frac{1}{36} = 0.492999...$$

Your chance of winning is about 49.3%.

10.2 Independence

Two events A and B are said to be *indpendent* if and only if

$$\mathbb{P}[A \cap B] = \mathbb{P}[A] \times \mathbb{P}[B].$$

We denote this by writing $A \perp\!\!\!\perp B$.

Otherwise, we say that A and B are *dependent*.

Remarks:

- *Intuition:* Independence means that if I know that A occurred or didn't occur, that gives me no information about whether B occurred or not. And vis-versa. Then

$$\begin{aligned}\mathbb{P}[A \mid B] &= \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} \\ &= \frac{\mathbb{P}[A\cap]\,\mathbb{P}[B]}{\mathbb{P}[B]} && \text{by independence} \\ &= \mathbb{P}[A]\end{aligned}$$

- Physical indpendence always implies mathematical independence. For example, when you roll two dice, the results are independent because they are physically independent.
- Mutually exclusive events are never independent.[1]. That's because if you know that one happened, then you know that the other didn't!

10.3 Random Variables

Based on sections 3.1 and 3.2 of the textbook

A *random variable* is a variable whose value depends on the outcome of a random process or phenomenon.

Technically, a random variable X is a function which assigns to each outcome $\omega \in S$ in the sample space a real number $X(\omega)$:

$$X : S \to \mathbb{R}.$$

But in this class, we usually won't think of a random variable in this way.

It is customary to denote random variables by capital letters like X, Y etc., and the values they take by x, y, etc.

The *state space* of a random variable X is the set of values it can take. We say that a random variable is *discrete* if the state space is "countable" (either finitely many values, or, for example integers or rational numbers).

Given a discrete random variable X, the *probability mass function (pmf)* is the function

$$p(x) := \mathbb{P}[X = x]$$

To avoid ambiguity, sometimes we name the pmf of X $p_X(x)$ instead of just $p(x)$.

Here are some standard random variables to be familiar with:

Example 34 (Bernoulli random variable)**.** Any random varible whose only values are 0 or 1. For any fixed value $0 < \alpha < 1$, we say that X *is a Bernoulli random variable with success parameter* σ if

$$X = \begin{cases} 1 & \text{with probability } \alpha \\ 0 & \text{with probability } 1-\alpha \end{cases}$$

In this case, we write $X \sim \text{Bern}(\alpha)$. The pmf of X is the function taking the following values:

$$p(1) = \alpha, \quad p(0) = 1-\alpha, \quad \text{and} \quad p(x) = 0 \text{ for all other values of } x$$

A Bernoulli random variable is nothing other than a coin flip (with a coin that lands 'head' with probability α.)

End of Example 34. □

[1] As long as the events are of positive probability

11 2025-02-07 | Week 4 | Lecture 11

- please read sections 3.1-3.5 in the textbook
- there is a webassign homework due next Friday
- handwritten assignment too

11.1 Discrete random variables (continued)

We often associate numbers with outcomes of experiments. This is what we mean by random variables.

Example 35 (Random variables and pmfs)**.** Say we flip a fair coin three times

	HHH	HHT	HTH	THH	TTH	THT	HTT	TTT
$X =$ # heads	3	2	2	2	1	1	1	0
$Y =$ # tails prior to first heads	0	0	0	1	2	1	0	3
$Z =$ (more tails than heads?)	1	1	1	1	0	0	0	0

Here $Z = \begin{cases} 1 & : \text{ more tails than heads} \\ 0 & : \text{ otherwise} \end{cases}$ X, Y, Z are examples of random variables. This is an example of what's called an "indicator" random variable. To be precise random variables are functions $X : S \to \mathbb{R}$ and $Y : S \to \mathbb{R}$.

In this example, the sample space is

$$S = \{\text{HHH, HHT, HTH, THH, TTH, THT, HTT, TTT}\}$$

$X = 2$ is an event:

$$[X = 2] = \{s \in S : X(s) = 2\} = \{\text{HHT, HTH, THH}\}$$

$X \leq 2$ is also an event:

$$[X \leq 2] = \{s \in S : X(s) \leq 2\} = \{\text{HHT, HTH, THH, TTH, THT, HTT, TTT}\}$$

We can specify X, Y, Z by the following tables:

x	$\mathbb{P}[X = x]$
0	1/8
1	3/8
2	3/8
3	1/8

y	$\mathbb{P}[Y = y]$
0	4/8
1	2/8
2	1/8
3	1/8

z	$\mathbb{P}[Z = z]$
0	4/8
1	4/8

Recall that given a discrete random variable X, the *probability mass function (pmf)* is the function

$$p(x) := \mathbb{P}[X = x]$$

For example, the pmf of X is the function p which take the following values

$$p(0) = 1/8 \quad p(1) = 3/8 \quad p(2) = 3/8, \quad \text{and} \quad p(3) = 1/8$$

It's just a different way of writing the above table!

General fact: For any discrete r.v., we have $p(x) \geq 0$ for all x, and that

$$\sum_x p(x) = 1,$$

where the sum ranges over all values that X can take.

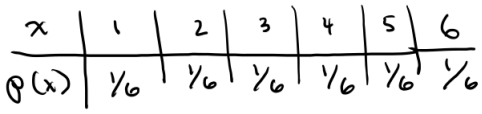
End of Example 35. □

Let X be a discrete random varible. The *cumulative distribution function (cdf)* of X is the function

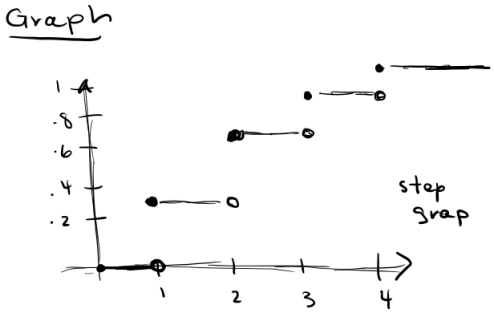
$$F_X(x) := \mathbb{P}[X \leq x]$$

Example 36 (cdf of a dice roll)**.** What is the pmf and cdf a dice roll?

The pmf is:



The cdf is



End of Example 36. □

Proposition 37 (cdf properties)**.** *We note that* $F_X : \mathbb{R} \to [0,1]$ *and that*

$$F_X(x) = \sum_{y \leq x} p_X(y)$$

From this equation, we can deduce that

$$F_X(x_1) \leq F_X(x_2) \text{ whenever } x_1 \leq x_2.$$

In other words, F_X *is an increasing function (or maybe more appropriately, "nondecreasing"). Also,*

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

Example 38 (Geometric distribution)**.** Throw a basketbball. Then

$$1 = \text{success} = \text{made basket} \tag{7}$$

$$0 = \text{failed} = \text{missed basket} \tag{8}$$

Fix a parameter $\alpha \in (0,1)$. Say that success has probability α and failure has probability $1-\alpha$.

Let X be the number of throws until I make a basket. Let $p(x)$ the pmf of X. Then

x	$p(x)$
1	α
2	$(1-\alpha)\alpha$
3	$(1-\alpha)^2\alpha$
4	$(1-\alpha)^3\alpha$
$\vdots$	$\vdots$

In other words, the pmf is

$$p(x) = \begin{cases} (1-\alpha)^{x-1}\alpha & \text{if } x = 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases}$$

A random variable with this pmf is called a *geometric random variable with success parameter* α.

End of Example 38. □

12 2025-02-10 | Week 5 | Lecture 12

Today's lecture is on sections 3.3 and 3.4 in the textbook

12.1 Expectation

The *expected value* of a discrete random variable is

$$\mu = \mathbb{E}[X] = \sum_x x \cdot p(x)$$

where the x in the summation runs over all possible values of X.

In words, the expected value is the *long-run average*, meaning that if you repeated an experiment many times (independently), the long-run average would converge to $\mathbb{E}[X]$.

Textbook uses the notation μ or μ_X for $\mathbb{E}[X]$.

The following is a simple but important fact:

Proposition 39 (Linearity of Expectation)**.** *Let a, b be numbers and X be a random variable. Then*

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$$

We saw this earlier, when we looked at the expected value of playing the lotter once ($-\$1$) and of playing 10 times ($-\$10$), etc.

Theorem 40 (Expectation of a function of X)**.** *Let X be a discrete random variable with pmf $p(x)$. Let $h : \mathbb{R} \to \mathbb{R}$ be any function. Then*

$$\mathbb{E}[h(X)] = \sum_x h(x)p(x)$$

provided that the sum on the right hand side is absolutely convergent. As usual, the summation runs over all possible values of X.

Example 41. Suppose Y is a discrete random variable with pmf

$$\begin{aligned} p(-2) &= \mathbb{P}[Y = -2] = 0.1 \\ p(-1) &= \mathbb{P}[Y = -1] = 0.3 \\ p(1) &= \mathbb{P}[Y = 1] = 0.4 \\ p(2) &= \mathbb{P}[Y = 2] = 0.2 \end{aligned}$$

and such that $\mathbb{P}[Y = y] = 0$ if $y \notin \{-2, -1, 1, 2\}$ Find the following quantities:

(a) $\mathbb{E}[Y]$

(b) $\mathbb{E}[3Y + 7]$

(c) $\mathbb{E}\left[Y^3 + 2Y\right]$

(d) $\mathbb{E}\left[e^X\right]$

Solution to (a):

$$\begin{aligned} \mathbb{E}[Y] &= -2(0.1) + (-1)(0.3) + (1)(0.4) + (2)(0.2) \\ &= -.2 - .3 + .4 + .4 \\ &= .3 \end{aligned}$$

Solution to (b):

$$\begin{aligned} \mathbb{E}[3Y + 7] &= 3\mathbb{E}[Y] + 7 && \text{by Proposition 39} \\ &= 3(0.3) + 7 && \text{by our answer to part (a).} \\ &= 7.9. \end{aligned}$$

Solution to (c): Here we will apply Theorem 40 with $h(x) = x^3 + 2x$:

$$\begin{aligned}\mathbb{E}\left[Y^3 + 2Y\right] &= \sum_{y\in\{-2,-1,1,2\}} h(x)p(x) \\ &= h(-2)p(-2) + h(-1)p(-1) + h(1)p(1) + h(2)p(2) \\ &= (-12)(0.1) + (-3)(0.3) + (3)(0.4) + (12)(0.2) \\ &= 1.5.\end{aligned}$$

Solution to (d): Here we will apply Theorem 40 with $h(x) = e^x$:

$$\begin{aligned}\mathbb{E}\left[e^Y\right] &= \sum_{y\in\{-2,-1,1,2\}} h(y)p(y) \\ &= \sum_{y\in\{-2,-1,1,2\}} e^y p(y) \\ &= e^{-2}p(-2) + e^{-1}p(-1) + e^1 p(1) + e^2 p(2) \\ &= e^{-2}\cdot 0.1 + e^{-1}\cdot 0.3 + e^1\cdot 0.4 + e^2\cdot 0.2 \\ &\approx 2.689.\end{aligned}$$

End of Example 41. □

12.2 Variance

The *variance* of a random variable X is

$$\mathrm{Var}(X) := \sum_x (x-\mu)^2 p(x)$$

where $\mu = \mathbb{E}[X]$ and x ranges over all possible values that X can take. The variance of a random variable is often denoted σ^2.

Important formula: There is a shortcut for computing variance

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

Variance is a measure of how likely the value of a random variable is to be far from its expected value. In an ideal world, we would measure this by

$$\mathbb{E}[|X-\mu|] = (\text{expected distance of } X \text{ from its mean})$$

But unfortunately, the absolute value function is mathematically difficult to work with, so instead we use

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

which is mathematically easer to work with because it doesn't have an absolute value.

We'll do some examples next time

13 2025-02-12 | Week 5 | Lecture 13

Please read 3.4, 3.6. Skip 3.5.

In the homework, we say the following problem:

Problem 42. Two cards are randomly drawn from a deck of cards. Let A be the event that at least one ace is drawn. Let A_s be the event that the ace of spades is chosen. And let B be the event that both cards are aces. Compute the following conditional probabilities:

(a) $\mathbb{P}[B \mid A_s]$

(b) $\mathbb{P}[B \mid A]$

We computed that $\mathbb{P}[B \mid A_s] = \frac{1}{17} \approx .06$ and that $\mathbb{P}[B \mid A] = \frac{1}{33} \approx .03$. The first of these is twice as big as the second. Why on earth does knowing A_s increase the probability of B, compared to knowing A?!

13.1 Variance

Recall, the *variance* of a random variable X is

$$\mathrm{Var}(X) := \sum_x (x-\mu)^2 p(x)$$

where $\mu = \mathbb{E}[X]$ and x ranges over all possible values that X can take. The variance of a random variable is often denoted σ^2.

Important formula: There is a shortcut for computing variance

$$\mathrm{Var}(X) = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 \tag{9}$$

Also, we have the following useful properties. Let a, b be fixed, nonrandom numbers. Then

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

Moreover, if X, Y are independent random variables, then

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$

(Of course, we always have $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$, even if X and Y are not independent. This is called *linearity of expectation*. But for variance, we need X and Y to be independent to do something similar.)

Example 43. Suppose

$$U = \begin{cases} 2 & : \text{ with probability } \alpha \\ 3 & : \text{ with probability } 1-\alpha \end{cases}$$

What is $\mathbb{E}[U]$ and $\mathrm{Var}(U)$?

Solution: First, we will compute the expectation:

$$\begin{aligned} \mathbb{E}[U] &= 2\alpha + 3(1-\alpha) \\ &= 3 - \alpha \end{aligned} \tag{10}$$

Next, we will compute $\mathrm{Var}(U)$. We will use the formula Eq. (9). In particular we wil need $\mathbb{E}\left[U^2\right]$, so let's compute that:

$$\begin{aligned} \mathbb{E}\left[U^2\right] &= 2^2\alpha + 3^2(1-\alpha) && \text{by Theorem 40} \\ &= 4\alpha + 9(1-\alpha) \\ &= 9 - 5\alpha \end{aligned} \tag{11}$$

We can now plug the values from Eqs. (10) and (11) into Eq. (9) to get

$$\begin{aligned}\mathrm{Var}(U) &= \mathbb{E}\left[U^2\right] - \left(\mathbb{E}\left[U\right]\right)^2 \\ &= 9 - 5\alpha - (3-\alpha)^2\end{aligned}$$

For example, if $\alpha = \frac{1}{2}$, then we get $\mathbb{E}\left[U\right] = 3 - \frac{1}{2} = 2.5$ and

$$\mathrm{Var}(U) = 9 - \frac{5}{2} - \frac{25}{4} = \frac{1}{4}$$

Now, someone asked "how do we interpret this 1/4?" The short answer is, it's complicated. There isn't a nice interpretation that works for every random variable, so often the best we can do is understand this as some measure of how much our random variable tends to deviate from its mean. In this case, the 1/4 is pretty small, so the random variable doesn't seem to deviate all that much from its mean.

End of Example 43. □

While variance does give us some information, it's only a single number and so the amount of information it gives us about the random variable is actually very limited, and it is impossible to interpret it precisely without additional information about the random variable. This is illustrated in the following example.

Example 44. Let $M = 1,000,000$. And let

$$X = \begin{cases} 0 & : \text{ with probability } .99 \\ 10M & : \text{ with probability } .01 \end{cases}$$

The random variable X is like winning the lottery.

On the other hand, let

$$Y = \begin{cases} -10 & : \text{ with probability } \frac{1}{2} \\ 10 & : \text{ with probability } \frac{1}{2} \end{cases}$$

In the case of X, we have

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}\left[X^2\right] - \left(\mathbb{E}\left[X\right]\right)^2 \\ &= \left[(10M)^2(0.01) + 0^2(0.99)\right] - \left[10M(0.01) + 0(0.99)\right]^2 \\ &= M^2 - \frac{M^2}{100} \\ &= \frac{99}{100}M^2\end{aligned}$$

This is a HUGE variance, because $M = 10,000,000$. But for the most part, intuitvely, few people would say that this r.v. "varies" all that much. After all, it's almost always 0.

On the other hand, what if we compute the variance of Y? In this, we have $\mathbb{E}\left[Y\right] = -10\frac{1}{2} + 10\frac{1}{2} = 0$, so

$$\begin{aligned}\mathrm{Var}(Y) &= \mathbb{E}\left[Y^2\right] - 0 \\ &= (-10)^2\frac{1}{2} + (10)^2\frac{1}{2} \\ &= 100\end{aligned}$$

In this case, the variance is pretty big—not huge, but pretty big. Yet the random variable Y doesn't really seem to "vary" all that much: after all, we know it is always either 10 more than the mean (which is zero), or 10 less than the mean.

End of Example 44. □

Hopefully this example doesn't give you the impression that variance is useless. It's actually a really important measure. The issue is just that random variables can "vary" in lots of different ways that can be hard to summarize as a single value.

Definition 45 (Binomial Distribution)**.** Flip a coin n times, and let $\alpha \in (0,1)$ be the probability of 'heads' on any given coin flip. The probability that you get exactly k heads is

$$\binom{n}{k}\alpha^k(1-\alpha)^{n-k}$$

A random variable with state space $\{0,1,2,\ldots,n\}$ is said to have *Binomial distribution with n trials and success probability α* if its pmf is

$$p(k) = \binom{n}{k}\alpha^k(1-\alpha)^{n-k}$$

for $k = 0,1,\ldots,n$. We write $X \sim \text{Bin}(n,\alpha)$.

You've already seen binomial random variables before. For an example, the random variable X from Example 35 is an example of a random variable with $X \sim \text{Bin}(3,\frac{1}{2})$.

13.2 Poisson convergence

Fix $\lambda \in (0,1)$. Suppose have a sequence of magic coins $1,2,\ldots$.

- Coin 1 has probability of heads λ
- Coin 2 has probability of heads $\frac{\lambda}{2}$
- Coin 3 has probability of heads $\frac{\lambda}{3}$
- And so forth.

Consider the following sequence of experiments:

- Experiment 1: Flip coin 1 once.
- Experiment 2: Flip coin 2 twice.
- Experiment 3: Flip coin 3 three times.
- Experiment n: Flip coin n, n times.

Now let us consider the nth experiment. Let $F_1,\ldots,F_n$ represent the n coin flips, with

$$F_1 = \begin{cases} 1 & : \text{ first coin flip is heads} \\ 0 & : \text{ first coin flip is tails} \end{cases}$$

$$F_2 = \begin{cases} 1 & : \text{ second coin flip is heads} \\ 0 & : \text{ second coin flip is tails} \end{cases}$$

and so forth for all n coin flips. So

$$X_n = F_1 + F_2 + \ldots + F_n$$

is the number of heads that you get.

Question: What is the expectation of X_n? We can use the linearity of expectation:

$$\begin{aligned} \mathbb{E}[X_n] &= \mathbb{E}[F_1 + F_2 + \ldots + F_n] \\ &= \mathbb{E}[F_1] + \mathbb{E}[F_2] + \ldots + \mathbb{E}[F_n] \\ &= \underbrace{\frac{\lambda}{n} + \frac{\lambda}{n} + \ldots + \frac{\lambda}{n}}_{n \text{ terms}} \\ &= \lambda. \end{aligned}$$

So the expected number of heads that we get doesn't change from experiment to experiment, even if we send $n \to \infty$. Think!—for each experiment, the probability of heads is $\frac{\lambda}{n}$ and that is getting smaller. However the total number of coin flips is getting larger and larger.

Question Is it possible that as $n \to \infty$, the random variable X_n converges to some new random variable, one that we haven't seen yet?

Yes.

14 2025-02-14 | Week 5 | Lecture 14

Read 3.1-3.4 and 3.6

14.1 Poisson convergence

Fix $\lambda \in (0,1)$. Suppose have a sequence of magic coins $1, 2, \ldots$.

- Coin 1 has probability of heads λ
- Coin 2 has probability of heads $\frac{\lambda}{2}$
- Coin 3 has probability of heads $\frac{\lambda}{3}$
- And so forth.

Consider the following sequence of experiments:

- Experiment 1: Flip coin 1 once.
- Experiment 2: Flip coin 2 twice.
- Experiment 3: Flip coin 3 three times.
- Experiment n: Flip coin n, n times.

Now let us consider the nth experiment. Let $F_1, \ldots, F_n$ represent the n coin flips, with

$$F_1 = \begin{cases} 1 & : \text{ first coin flip is heads} \\ 0 & : \text{ first coin flip is tails} \end{cases}$$

$$F_2 = \begin{cases} 1 & : \text{ second coin flip is heads} \\ 0 & : \text{ second coin flip is tails} \end{cases}$$

and so forth for all n coin flips. So

$$X_n = F_1 + F_2 + \ldots + F_n$$

is the number of heads that you get.

Question: What is the expectation of X_n? We can use the linearity of expectation:

$$\begin{aligned} \mathbb{E}[X_n] &= \mathbb{E}[F_1 + F_2 + \ldots + F_n] \\ &= \mathbb{E}[F_1] + \mathbb{E}[F_2] + \ldots + \mathbb{E}[F_n] \\ &= \underbrace{\frac{\lambda}{n} + \frac{\lambda}{n} + \ldots + \frac{\lambda}{n}}_{n \text{ terms}} \\ &= \lambda. \end{aligned}$$

So the expected number of heads that we get doesn't change from experiment to experiment, even if we send $n \to \infty$. Think!—for each experiment, the probability of heads is $\frac{\lambda}{n}$ and that is getting smaller. However the total number of coin flips is getting larger and larger.

Question Is it possible that as $n \to \infty$, the random variable X_n converges to some new random variable, one that we haven't seen yet?

Yes.

Theorem 46 (The exponential function)**.** *For any real number x, it holds that*

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}.$$

Definition 47 (Poisson ("Pwason"))**.** Let $\lambda > 0$. A random variable X that takes on the values $0, 1, 2, \ldots$ is a *Poisson* random variable with parameter λ if its probability mass function is given by

$$p(k) := \mathbb{P}[X = k] = e^{-\lambda}\frac{\lambda^k}{k!} \quad \text{for } k = 0, 1, 2, \ldots.$$

Before explaining this, we should check that it is actually a probability mass function: remember, it needs to sum to 1:

$$\begin{aligned}\sum_{k=0}^{\infty} p(k) &= \sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!} \\ &= e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} \\ &= e^{-\lambda}e^{\lambda} && \text{by Theorem 46} \\ &= 1.\end{aligned}$$

The Poisson distribution is a common model since it models "accidents".

Example 48. Oahu has about a million cars and averages about 6 car accidents per day. The probability that any one car will have an accident is very small, and they are all (nearly) independent from each other, bus since there are so many cars, you expect 6 accidents, on average. But actual number of accidents each day is $X = 0$ or 1 or 2 or .... The number of car accidents is thus poisson distributed with mean $\lambda = 6$ What is

$$F(2) = \mathbb{P}[X \leq 2]\,?$$

The probabilities of no accidents, one accident, and two accidents are

$$p(0) = e^{-6}\frac{6^0}{0!} = e^{-6} \approx 0.002$$

$$p(1) = e^{-6}\frac{6^1}{1!} = 6e^{-6} \approx 0.015$$

$$p(2) = e^{-6}\frac{6^2}{2!} = 18e^{-6} \approx 0.045$$

Therefore $F(2) = p(0) + p(1) + p(2) \approx 0.002 + 0.015 + 0.045 = 0.062$

End of Example 48. □

Proposition 49. *Let $X \sim Pois(\lambda)$. Then*

$$\mathbb{E}[X] = \lambda \quad \textit{and} \quad \mathrm{Var}(X) = \lambda$$

That the variance and expected values are the same for the Poisson random variable is a neat coincidence.

Theorem 50 (Poisson approximation)**.** *Recall that if $X \sim Binom(n, p)$, that means*

$$\mathbb{P}[X = k] = \binom{n}{k}p^k(1-p)^{n-k}, \quad \textit{for } k = 0, 1, 2, \ldots.$$

Let $\lambda = np$. When n is large and p is small (e.g., $n > 50$ and $\lambda < 5$), then we have the following approximation:

$$\mathbb{P}[X = k] \approx \frac{e^{-\lambda}\lambda^k}{k!} \quad \textit{for } k = 0, 1, 2, \ldots.$$

In other words, when n is large and p is small, X is well-approximated by a Poisson distribution with parameter $\lambda = np$.

This convergence/approximation idea is actually how I introduced the Poisson random variable in the first place, by considering the sequence of experiments involving magic coin flips.

Let's try an application of this:

Example 51 (Poisson approximation)**.** In *Dungeons and Dragons*, a 20-sided dice is used (aka the 'd20'). When you roll a 1 out of 20, that's called a *critical failure*, which usually results in something terrible happening.

Suppose that in the course of a game, a twenty side dice is rolled $n = 100$ times. Let X be the number of critical failures. The probability of a critical failure is $p = 1/20 = .05$. So

$$X \sim \text{Bin}\left(100, \frac{1}{20}\right).$$

So, for example, we have $\mathbb{E}[X] = 100 \times \frac{1}{20} = 5$.

The pmf of X is

$$\mathbb{P}[X = k] = \binom{100}{k}\left(\frac{1}{20}\right)^k\left(1 - \frac{1}{20}\right)^{100-k}, \text{ for } k = 0, 1, 2, \ldots, 100$$

Since $p = 0.05$ is small and $n = 100$ is large, we can approximate X with a Poisson random variable Y with parameter $\lambda = np = 5$. That is, $Y \sim \text{Pois}(5)$. The pmf of Y is

$$\mathbb{P}[Y = k] = e^{-5} \cdot \frac{5^k}{k!}, \text{ for } k = 0, 1, 2, \ldots$$

Let's compute some probabilities, rounding to three decimal places

x	$\mathbb{P}[X = k]$	$\mathbb{P}[Y = k]$
0	.006	.007
1	.031	.034
2	.081	.084
3	.140	.140
4	.178	.175

End of Example 51. □

15 2025-02-19 | Week 6 | Lecture 15

Definition 52 (Continuous random variable)**.** A random variable X is said to be *continuous* if it takes values in an interval I (or disjoint union of intervals) AND $\mathbb{P}[X = x] = 0$ for all $x \in I$.

To be precise, a random variable x is continuous if its cdf

$$F(x) = \mathbb{P}[X \leq x]$$

is a continuous function.

Definition 53 (pdf)**.** Let X be a continuous random variable. The *probability density function* (aka "pdf" or "density") of X is a function f_X such that for any real numbers a, b with $a \leq b$, we have

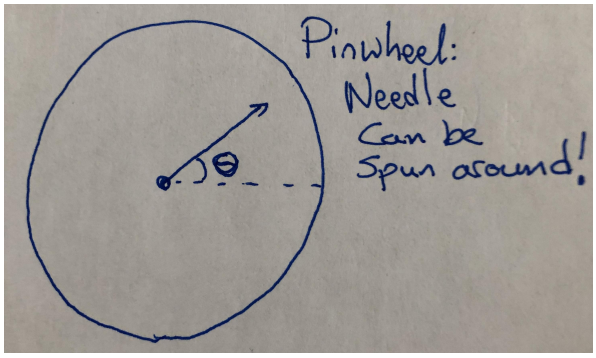
$$\mathbb{P}[X \in (a, b)] = \int_a^b f_X(x)dx$$

The graph of f_X is called the *density curve*.

Observations

- $f_X(x) \geq 0$ for all x
- $\int_{-\infty}^{\infty} f_X(x)dx = 1$
- $f_X(x)$ is not a probability!!! But you can think of it as "relative likeliood"

Example 54 (Pinwheel)**.** Suppose you have a pinwheel as follows.



Suppose we spin the needle. When the needle stops spinning, we measure its angle from the dotted line. Let X be the angle of the needle. If X is measured in degrees, then

$$X \in [0, 360)$$

but X could take *any* value, not just integer values. On the other hand, if we measure X in radians, then

$$X \in [0, 2\pi).$$

Moreover we have no reason to believe any one angle is more likely than any others – for our idealized pinwheel, all values between 0 and 2π should all be equally likely. This is an example of a uniform random variable, which we define next.

End of Example 54. □

Definition 55 (Uniform distribution)**.** Let I be an interval with endpoints a and b, such that $a < b$. A continuous random variable X has *uniform* distribution on I if

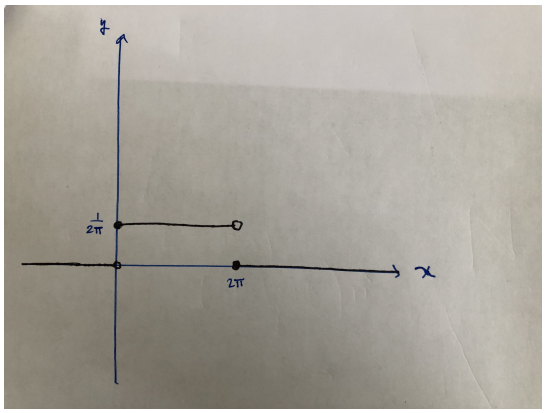
$$\begin{cases} \frac{1}{b-a} & : \quad x \in I \\ 0 & : \quad x \notin I \end{cases}$$

We note that $b - a$ is the length of the interval.

In the case of the pinwheel example, the pdf is

$$f_X(x) = \begin{cases} \frac{1}{2\pi} & : \quad 0 \leq x < 2\pi \\ 0 & : \quad \text{else} \end{cases}$$

so the density curve is



Proposition 56. *The cumulative density function or cdf is the function*

$$F_X(x) = \mathbb{P}[X \leq x] = \int_{-\infty}^{x} f_X(t)dt$$

Observations

- $\mathbb{P}[a \leq X \leq b] = F_X(b) - F_X(a)$

16 2025-02-21 | Week 6 | Lecture 16

Recall the example of the pinwheel Example 54, with the angle X measured in radians, so that the state space of X is $[0, 2\pi)$, with all values equally likely. Recall that we computed the pdf as

$$f_X(x) = \frac{1}{2\pi} \cdot \mathbf{1}_{[0,2\pi)}(x) = \begin{cases} \frac{1}{2\pi} & : \quad x \in [0, 2\pi) \\ 0 & : \quad x \notin [0, 2\pi) \end{cases}$$

Recall also that the cdf is defined as

$$F_X(x) := \int_{-\infty}^{x} f_X(t)dt.$$

So that for the pinwheel angle X, we have

$$\begin{aligned} F_X(x) &= \int_{-\infty}^{x} \frac{1}{2\pi} \cdot \mathbf{1}_{[0,2\pi)}(t)dt \\ &= \frac{1}{2\pi} \int_{-\infty}^{x} \mathbf{1}_{[0,2\pi)}(t)dt. \end{aligned} \tag{12}$$

To evaluate an integral of this form, we have to consider three cases:

- $x < 0$ (i.e. when x is to the left of the interval)
- $x \geq 2\pi$ (i.e. when x is to the right of the interval)
- $0 \leq x < 2\pi$ (i.e. when x is in the interval)

Let's think through these cases:

- When $x < 0$, the integral on the right-hand side of Eq. (12) is zero.
- When $x \geq 2\pi$, the integral equals

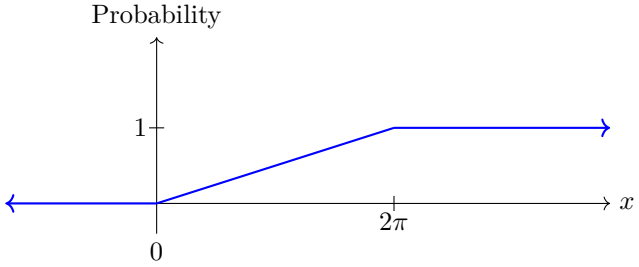
$$\int_{-\infty}^{0} 0dt + \frac{1}{2\pi} \int_{0}^{2\pi} dt + \int_{2\pi}^{x} 0dt = 0 + 1 + 0 = 1$$

- When $0 \leq x < 2\pi$, the integral equals

$$\int_{-\infty}^{0} 0dt + \frac{1}{2\pi} \int_{0}^{x} dt = 0 + \frac{x}{2\pi} = \frac{x}{2\pi}$$

From this analysis and Eq. (12), we can conclude that

$$F_X(x) = \begin{cases} 0 & : \quad x < 0 \\ \frac{x}{2\pi} & : \quad 0 \leq x < 2\pi \\ 1 & : \quad x \geq 2\pi \end{cases}$$



The *median* of a continuous random variable is the point x^* where $F(x^*) = \frac{1}{2}$. Interpretation: A random variable is equally likely to be greater or less than than its median.

The *expected value* of a continuous random variable X is

$$\mathbb{E}[X] := \int_{-\infty}^{\infty} x f_X(x) dx$$

where $f_X(x)$ is the pdf of X. (Provided that the integral above converges absolutely).

Similarly, for any function $h : \mathbb{R} \to \mathbb{R}$ we have

$$\mathbb{E}[h(X)] = \int_{-\infty}^{\infty} h(x) f_X(x) dx$$

This is the continuous analogue of the very important theorem Theorem 40 which was for discrete random variables.

We also have variance defined like it was for discrete random variables:

$$\mathrm{Var}(X) := \mathbb{E}\left[(X-\mu)^2\right],$$

where $\mu = \mathbb{E}[X]$, and as before we have the formula

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

Example 57 (Waiting times)**.** Suppose you station yourself at spot on a road and watch as cars pass by. Let T be the time in minutes until the first car passes by. This random variable is called the *waiting time*.

Under certain circumstances, it is reasonable for T to have pdf

$$f(x) = \lambda e^{-\lambda x} \mathbf{1}_{[x \geq 0]} \tag{13}$$

for a certain (fixed) value of $\lambda > 0$. In this expression, the term $\mathbf{1}_{[x \geq 0]}$ is an "indicator function" which takes value 1 when $x \geq 0$ and 0 if $x < 0$. A random variable with pdf like in Eq. (13) is called an *exponential random variable with rate* λ, and these are some of the most important random variables.

For concreteness, let's assume that $\lambda = 3$, so

$$f(x) = 3e^{-3x} \mathbf{1}_{[x \geq 0]}$$

Then the cdf is

$$\begin{aligned} F(t) &= \int_{-\infty}^{t} f_T(x) dx \\ &= \int_0^t 3e^{-3x} dx \\ &= 3\left[\frac{-e^{-4x}}{3}\right]_0^t \\ &= 1 - e^{-3t}. \end{aligned}$$

Equivalently,

$$\mathbb{P}[T > t] = e^{-3t}.$$

For example the probability that you have to wait more than $t = 30$ seconds for a car to pass by is

$$\mathbb{P}[T > 5] = e^{-3 \cdot 0.5} \approx .22$$

What about

$$\mathbb{E}[T]\,?$$

If the cars come at a rate of $\lambda = 3$ cars per minute, then on average how long would you have to wait for a car to arrive? This should be 20 seconds.

End of Example 57. □

17 2025-02-24 | Week 7 | Lecture 17

Topic: moment generating functions – this isn't in the textbook

Definition 58 (Equal in distribution)**.** Let X and Y be two random variables. We say that X and Y *have the same distribution*, denoted $X \overset{d}{=} Y$ if they have the same cdf (i.e., if $F_X = F_Y$).

- If X and Y are discrete and have pmfs p_X and p_Y, then this is equivalent to $p_X = p_Y$.
- If X and Y are continuous with pdfs f_X and f_Y, then this is equivalent to $f_X = f_Y$.

Example 59. Warning: equality in distribution does not imply that $X = Y$. For example, if I roll a red dice X and a blue dice Y, both with 6 sides, the chances are pretty good that $X \neq Y$. But of course these are equal in distribution, since

$$p_X(k) = p_Y(k) = \frac{1}{6} \quad \text{for all } k = 1, 2, \ldots, 6.$$

End of Example 59. □

Definition 60 (Standard deviation)**.** The *standard deviation* of X is the square root of $\mathrm{Var}(X)$:

$$\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)}.$$

Standard deviation is often denoted σ.

Definition 61 (Moment generating function)**.** The *moment generating function* or *m.g.f.* of a random variable X is the function

$$M(t) := \mathbb{E}\left[e^{tX}\right].$$

- For a discrete random variable with pmf $p(x)$, we have

$$M(t) = \sum_x e^{tx} p(x) dx.$$

- For a continuous r.v., with pdf $f(x)$, we have

$$M(t) = \int_{-\infty}^{\infty} e^{tx} f(x) dx.$$

The moment generating function is important because if two random variables X and Y have the same mgf, then they have the same distribution.

Question: But what are "moments" and why is this called a "moment-generating" function?!

Definition 62. Let X be a r.v., and let k be a nonnegative integer. The *k-th moment of X* is the quantity

$$\mathbb{E}\left[X^k\right].$$

So, the 0-th moment is always 1. The first moment is just the expectation of the r.v. The second moment shows up in the variance fomrula:

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

To the best of my knowledge, moments don't have a straightforward interpretation that holds generally. But they are important for two reasons (1) they often show up in things we care about, like the variance formula, as well as in the assumptions and proofs of important theorems, like the central limit theorem, and (2) when taken together, all the moments of a random variable $\left\{\mathbb{E}\left[X^k\right] : k = 0, 1, 2, \ldots\right\}$ will usually completely determine the distribution of the random variable.[2]

[2] It is necessary that the sequence $\mathbb{E}\left[X\right], \mathbb{E}\left[X^2\right], \mathbb{E}\left[X^3\right], \ldots$ does not grow too quickly, but we can ignore this technical issue for now.

As the name suggests, the mgf of X has a close connection to the moments of X. To see this, observe that

$$e^{tX} = \sum_{k=0}^{\infty} \frac{(tX)^k}{k!} = \sum_{k=0}^{\infty} \frac{t^k X^k}{k!}.$$

So

$$\mathbb{E}\left[e^{tX}\right] = \mathbb{E}\left[\sum_{k=0}^{\infty} \frac{t^k X^k}{k!}\right] = \sum_{k=0}^{\infty} \frac{t^k}{k!} \mathbb{E}\left[X^k\right] \tag{14}$$

Example 63 (Using a mgf to compute moments)**.** Let $X \sim \exp(\lambda)$. That is, X is an exponential random variable with rate λ.

Question #1: What is the moment generating function of X?

First, recall that

- The pdf of X is $f_X(x) = \lambda e^{-\lambda x}$, for all $x > 0$.
- For any function $h : \mathbb{R} \to \mathbb{R}$, we have

$$\mathbb{E}\left[h(X)\right] = \int_{-\infty}^{\infty} h(x) f_X(x) dx$$

Using these two facts, we have:

$$\begin{aligned} M(t) &= \mathbb{E}\left[e^{tX}\right] \\ &= \int_0^{\infty} e^{tx} \lambda e^{-\lambda x} dx \\ &= \lambda \int_0^{\infty} e^{(t-\lambda)x} dx \\ &= \frac{\lambda}{t-\lambda} \left[e^{(t-\lambda)x}\right]_{x=0}^{x=\infty}. \\ &= \frac{\lambda}{t-\lambda} \left[\left(\lim_{x\to\infty} e^{(t-\lambda)x}\right) - 1\right] \end{aligned}$$

When we try to evaluate the limit as $x \to \infty$, something goes wrong if $t > \lambda$. Indeed, we have:

$$\lim_{x\to\infty} e^{(t-\lambda)x} = \begin{cases} +\infty & : \quad t > \lambda \\ 0 & : \quad t < \lambda \end{cases}$$

Therefore we have

$$M(t) = \frac{\lambda}{t-\lambda}(-1) = \frac{\lambda}{\lambda - t}$$

but this function is defined only for $t < \lambda$.

Question #2: For concreteness, let's suppose that $\lambda = 9$. What is the 2-nd moment of X?

Solution 1. We can compute the second moment using the formula

$$\mathbb{E}\left[X^2\right] = \int_0^{\infty} x^2 f_X(x) dx = \int_0^{\infty} x^2 9 e^{-9x} dx$$

This would work. But this would require integrating by parts twice, which we may wish to avoid. So instead, I'll present a different approach which utilizes the moment generating function.

Solution 2. From our earlier calculations, we have

$$M(t) = \frac{9}{9-t}, \quad \text{defined for } t < 9$$

Differentiating M with respect to t gives

$$M'(t) = \frac{9}{(9-t)^2}.$$

Differentiating again, we have:

$$M''(t) = \frac{9}{2(9-t)^3} \tag{15}$$

But recall by Eq. (14), we have

$$M(t) = \sum_{k=0}^{\infty} \frac{t^k}{k!}\mathbb{E}\left[X^k\right] = 1 + t\mathbb{E}\left[X\right] + \frac{t^2}{2!}\mathbb{E}\left[X^2\right] + \frac{t^3}{3!}\mathbb{E}\left[X^3\right] + \frac{t^4}{4!}\mathbb{E}\left[X^4\right] + \ldots,$$

and differentiating *this* equation twice gives

$$M''(t) = \mathbb{E}\left[X^2\right] + t\mathbb{E}\left[X^3\right] + \frac{t^2}{2!}\mathbb{E}\left[X^4\right] + \ldots. \tag{16}$$

Combining Eqs. (15) and (16), it follows that

$$\mathbb{E}\left[X^2 e^{tX}\right] = \frac{9}{2(9-t)^3}.$$

The above equation holds for all t such that both sides are defined. In particular, it holds when $t = 0$. Setting $t = 0$, something almost magical happens:

$$\mathbb{E}\left[X^2 e^{0}\right] = \frac{9}{2(9)^3}$$

which simplifies to

$$\mathbb{E}\left[X^2\right] = \frac{2}{81}.$$

We've computed the second moment (and all we had to do was differentiate the mgf twice and then plug in zero!).

End of Example 63. □

This example illustrates the following general property of a moment generating function:

$$\mathbb{E}\left[X^k\right] = \left[\frac{d^k}{dx^k}\left[M(t)\right]\right]_{t=0} \tag{17}$$

And this is how the moment generating function got its name.

18 2025-02-26 | Week 7 | Lecture 18

Please read chapter 4.3

Announcment: I would like to pushing back the midterm until March 14.

Example 64. Let $X \sim \exp(3/2)$. What is the probability is that X is greater than twice its standard deviation σ?

Solution: Recall that for an exponential r.v. Y with rate λ, we have

$$\mathbb{E}[Y] = \frac{1}{\lambda}. \tag{18}$$

and

$$F_Y(t) := \mathbb{P}[Y \leq t] = 1 - e^{-\lambda t}.$$

This implies that

$$\mathbb{P}[Y > t] = e^{-\lambda t}. \tag{19}$$

In our case, we have $\lambda = 3/2$. Let σ be the standard deviation of X. We want to compute

$$\mathbb{P}[X > 2\sigma].$$

By Eq. (19), with $\lambda = \frac{3}{2}$ and $t = 2\sigma$, we have

$$\mathbb{P}[X > 2\sigma] = e^{-\frac{3}{2}(2\sigma)} = e^{-3\sigma}. \tag{20}$$

So it remains to find σ. We have

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

From Eq. (18), have $\mathbb{E}[X] = \frac{2}{3}$, so

$$\mathrm{Var}(X) = \mathbb{E}\left[X^2\right] - \frac{4}{9} \tag{21}$$

Next, let's compute $\mathbb{E}\left[X^2\right]$. Recall from Example 63 that

$$M(t) = \begin{cases} \frac{\lambda}{\lambda - t} & : \quad t < \lambda \\ 0 & : \quad t \geq \lambda \end{cases}$$

which for our setting is

$$M(t) = \frac{\frac{3}{2}}{\frac{3}{2} - t} \quad \text{for } t < \frac{3}{2}.$$

From Eq. (17), we have $M''(0) = \mathbb{E}\left[X^2\right]$.

Indeed, $M''(t) = 3\left(\frac{3}{2} - t\right)^{-3}$. Then

$$\mathbb{E}\left[X^2\right] = M''(0) = 3\left(\frac{3}{2}\right)^{-3} = \frac{8}{9}$$

Plugging this into Eq. (21), we get

$$\mathrm{Var}(X) = \frac{8}{9} - \frac{4}{9} = \frac{4}{9}.$$

And hence

$$\sigma = \sqrt{\mathrm{Var}(X)} = \sqrt{\frac{4}{9}} = \frac{2}{3}.$$

Therefore by Eq. (20), we have

$$\mathbb{P}[X > 2\sigma] = e^{-3\sigma} = e^{-2} \approx 0.14.$$

End of Example 64. □

18.1 Normal random variables

The most important continuous r.v. is the following,

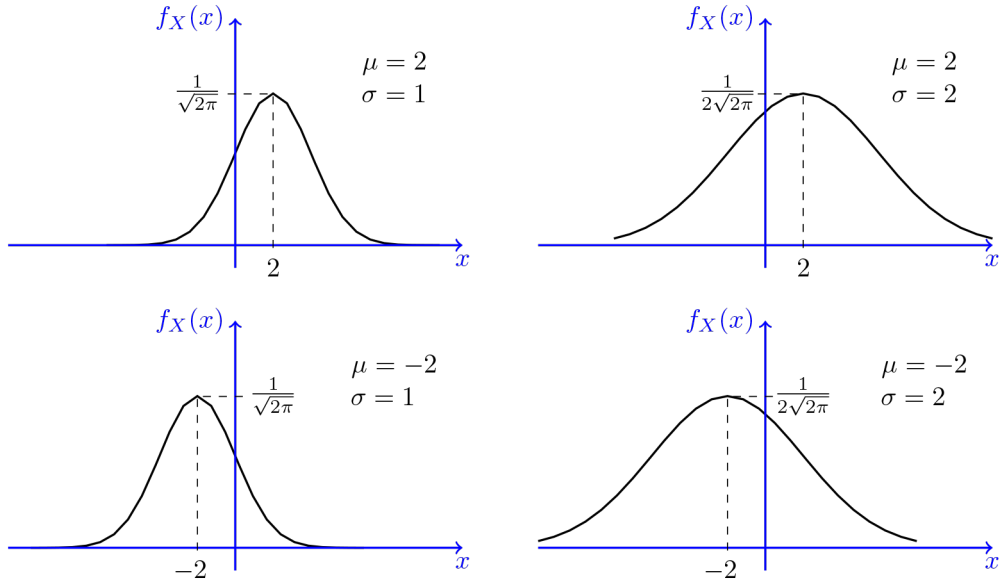
Example 65 (Gaussian random variable)**.** We say that X is a *Gaussian random variable with mean μ and variance σ^2* if X is a continuous r.v. with density given by the formula

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$

In this case, we write $X \sim \mathcal{N}(\mu, \sigma^2)$. Gaussian r.v.'s are also known as *Normal* random variables.

End of Example 65. □

Normal random variables have a "bell-curve" distribution. Here are some example density curves for various values of μ and σ. Note that μ determines where the peak of the bell curve is located, and σ determines how wide or narrow it is:



Normal random variables arise naturally when we have a random quantity that is the result of adding up many small random quantities. A classic example of this is the height of a random person, as height is influenced by the effects of thousands of genes and environmental factors.

Most important special case: The most important special case is where $\mu = 0$ and $\sigma^2 = 1$, in which case the density is

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad -\infty < x < \infty$$

A random variable with this density is called a *standard normal* and is usually denoted with the letter Z.

Fact: if $X \sim \mathcal{N}(\mu, \sigma^2)$, then the random variable

$$Z = \frac{X - \mu}{\sigma} \tag{22}$$

is a standard normal random variable.

This transformation

$$X \mapsto \frac{X - \mu}{\sigma}$$

is called *standardization*. Because we can standardize any normal random variable, that means we can focus our attention on standard normal random variables.

But first, we should really check that ϕ is really a density. Namely, we need to check that

$$\int_{-\infty}^{\infty} \phi(x)dx = 1.$$

We showed how to do this by using Fubini's theorem and converting to polar coordinates.

The CDF of a standard normal: The cdf of a standard normal is usually denoted with the symbol Φ:

$$\Phi(x) = \int_{-\infty}^{x} \phi(u)du = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{u^2}{2}} du.$$

There is no elementary form for the antiderivative Φ, so generally we compute the values with a computer.

19 2025-02-28 | Week 7 | Lecture 19

An example of normal random variable and standardization:

Example 66. Due to variation in the manufacturing process, the peak power consumption X of the latest Nvidia GPU varies slightly from chip to chip. Assume the power consumption is normally distributed $X \sim \mathcal{N}(\mu, \sigma^2)$. What is the probability that X is within one standard deviation of its mean?

$$\begin{aligned}\mathbb{P}[\mu - \sigma \leq X \leq \mu + \sigma] &= \mathbb{P}\left[\frac{(\mu-\sigma)-\mu}{\sigma} \leq Z \leq \frac{(\mu+\sigma)-\mu}{\sigma}\right] \\ &= \mathbb{P}[-1 \leq Z \leq 1] \\ &\Phi(1) - \Phi(-1) \\ &\approx .68\end{aligned}$$

So if $\mu = 700$ watts and $\sigma = 35$ watts (realistic values for the Nvidia's new \$30,000 H100 GPU), we have

$$\mathbb{P}[665 \leq X \leq 735] \approx 0.68.$$

In other words, about 68% of chips will have peak power consumption between 665 and 735 watts.

End of Example 66. □

19.1 Joint distributions

Please read chapter 5.1

We are interested in understanding the properties of *random vectors*, which are vectors whose entries are random variables.

Let X and Y be two discrete r.v.s defined on the sample space $\mathcal{S}$. The *joint pmf* $p(x, y)$ is defined as

$$p(x, y) = \mathbb{P}[X = x, Y = y]$$

for all $x, y \in \mathcal{S}$. It must be the case that $p(x, y) \geq 0$ and $\sum_{x,y \in \mathcal{S}} p(x, y) = 1$.

Let A be any set consisting of pairs of (x, y) values. For example,

$$A = \{(x, y) : x + y = 5\}$$

or

$$A = \{(1, 2), (1, 4), (1, 6), (2, 5)\}$$

Then the probability

$$\mathbb{P}[(X, Y) \in A] = \sum_{\substack{x,y \in S \\ (x,y) \in A}} p(x, y)$$

Example 67. Cars made in a factory experience two kinds of defects: defective joint welds, and and improperly tightened bolts. Let

- X = number of defective welds in a new car
- Y = number of improperly tightened bolts

Past data suggests that the joint pdf of (X, Y) is given by the following table:

		Y			
		0	1	2	3
X	0	.840	.030	.020	.010
	1	.060	.010	.008	.002
	2	.010	.005	.004	.001

The probability that there are no defects in the car is

$$p(0,0) = \mathbb{P}[X = 0, Y = 0] = .84$$

The probability that there will be exactly one defect is

$$p(0,1) + p(1,0) = \mathbb{P}[X = 0, Y = 1] + \mathbb{P}[X = 1, Y = 0] = .06 + .03 = .09.$$

What is the probability that there will be no improperly tightened bolts? This concerns only the variable Y. This can be obtained by summing over all values of x:

$$\begin{aligned}\mathbb{P}[Y = 0] &= \sum_{x=0}^{2} \mathbb{P}[X = x, Y = 0] \\ &= p(0,0) + p(1,0) + p(2,0) \\ &= .84 + .06 + .01 \\ &= .91\end{aligned}$$

End of Example 67. □

Given the joint pmf for a two-dimensional discrete random vector (X, Y), it is easy to derive the individual pmfs for X and Y. The manner in which this is done is suggested by the previous example:

If p is the pdf of (X, Y), then the pdf of X alone is obtained by the formula

$$p_X(x) = \sum_{\text{all } y} p(x, y)$$

and the pdf of Y alone is given by

$$p_Y(y) = \sum_{\text{all } x} p(x, y)$$

In this setting, the pdfs p_X and p_Y are called the *marginal pmfs* of X and Y (respectively).

Similarly, if X, Y are continuous r.v.s, then the *joint pdf* $f(x, y)$ is a function satisfying $f(x, y) \geq 0$ and $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)dxdy = 1$ such that for any 2-dimensional set A,

$$\mathbb{P}[(X, Y) \in A] = \int_A \int f(x, y)dxdy$$

In addition, we can compute the *marginal density* of X and Y as

$$f_X(x) = \int_{\text{all } y} f(x, y)dy$$

and

$$f_Y(y) = \int_{\text{all } x} f(x, y)dx$$

20 2025-03-03 | Week 8 | Lecture 20

Sections 5.1, 5.2 Recall that two events are independent if knowledge that one has occurred gives no information about whether the other has occurred. We can extend this idea to random variables:

Definition 68 (Independence of rvs)**.** Two random variables X and Y are *independent* if for every pair of (x, y)-values,

$$p(x, y) = p_X(x) \cdot p_Y(y) \quad \text{when } X \text{ and } Y \text{ are discrete}$$

or

$$f(x, y) = f_X(x) \cdot f_Y(y) \quad \text{when } X \text{ and } Y \text{ are continuous}$$

If this condition isn't satisfied for all (x, y), then X and Y are said to be *dependent*.

[For clarity in the above: $p(x, y)$ is the joint pmf of (X, Y). And $f(x, y)$ is the joint pdf of (X, Y).]

The intuition is this: two random variables are independent if observing one of them doesn't give you any information about the other.

Example 69. In the car factory example from last lecture, we had two random variables, X and Y, with joint pmf given by

		Y			
		0	1	2	3
X	0	.840	.030	.020	.010
	1	.060	.010	.008	.002
	2	.010	.005	.004	.001

Are X and Y independent? To determine this, we need to first obtain the pmfs:

x	$p_X(x)$
0	.9
1	.08
2	.02

and

y	$p_Y(y)$
0	.91
1	.045
2	.032
3	.013

Having computed the pmfs, we can verify that X and Y are **not independent**. This is because

$$p(0, 0) = .84 \neq (.9)(.91) = .819 = p_X(0)p_Y(0).$$

End of Example 69. □

Example 70. Let X and Y be the lifetimes of two lightbulbs. Assume that $X \sim \exp(\alpha)$ and $Y \sim \exp(\beta)$, where α and β are positive numbers. It's reasonable to assume of these as independent. Then the joint pdf is

$$\begin{aligned} f(x, y) &= f_X(x) f_Y(y) \\ &= \left(\alpha e^{-\alpha x} \mathbf{1}_{[x>0]}\right) \left(\beta e^{-\beta y} \mathbf{1}_{[y>0]}\right) \\ &= \alpha\beta e^{-\alpha x - \beta y} \mathbf{1}_{[x,y>0]} \\ &= \begin{cases} \alpha\beta e^{-\alpha x - \beta y} & : \quad x, y > 0 \\ 0 & : \quad \text{else} \end{cases} \end{aligned}$$

Suppose $\alpha = 1/1000$ and $\beta = 1/1200$. Then the expected lifetimes of the bulbs are 1000 hours and 1200 hours, respectively.

Question: What is the probability that both lightbulbs last at least 900 hours?

Want:

$$\begin{aligned}
\mathbb{P}\left[X \geq 900, Y \geq 900\right] &= \int_{900}^{\infty}\int_{900}^{\infty}\left(\alpha e^{-\alpha x}\right)\left(\beta e^{-\beta y}\right)dxdy \\
&= \int_{900}^{\infty}\left[\int_{900}^{\infty}\alpha e^{-\alpha x}dx\right]\left(\beta e^{-\beta y}\right)dy \\
&= \left(\int_{900}^{\infty}\alpha e^{-\alpha x}dx\right)\left(\int_{900}^{\infty}\beta e^{-\beta y}dy\right) \\
&= \left(\left[e^{-\alpha x}\right]_{x\to\infty}^{x=900}\right)\left(\left[e^{-\beta y}\right]_{y\to\infty}^{y=900}\right) \\
&= \left(e^{-900\alpha}-0\right)\left(e^{-900\beta}-0\right) \\
&= e^{-900(\alpha+\beta)} \\
&= e^{-900(\frac{1}{1000}+\frac{1}{1200})} \\
&= e^{-1.65} \\
&\approx .2
\end{aligned}$$

So with probability about one fifth, both ligthbulbs will last at least 900 hours.

A faster way to do this. Observe that the events $[X \geq 900]$ and $[Y \geq 900]$ are independent, because the random variables are. Then

$$\begin{aligned}
\mathbb{P}\left[X \geq 900, Y \geq 900\right] &= \mathbb{P}\left[X \geq 900\right]\mathbb{P}\left[Y \geq 900\right] \\
&= e^{-900\alpha}e^{-900\beta} \\
&= e^{-1.65}
\end{aligned}$$

End of Example 70. □

21 2025-03-05 | Week 8 | Lecture 21

Please read 5.1-5.4

Example 71. Let X and Y be the lifetimes of two lightbulbs. Assume that $X \sim \exp(\alpha)$ and $Y \sim \exp(\beta)$, where α and β are positive numbers. It's reasonable to assume of these as independent. Then the joint pdf is

$$\begin{aligned} f(x,y) &= f_X(x) f_Y(y) \\ &= \left(\alpha e^{-\alpha x} \mathbf{1}_{[x>0]}\right)\left(\beta e^{-\beta y} \mathbf{1}_{[y>0]}\right) \\ &= \alpha\beta e^{-\alpha x - \beta y} \mathbf{1}_{[x,y>0]} \\ &= \begin{cases} \alpha\beta e^{-\alpha x - \beta y} & : \quad x, y > 0 \\ 0 & : \quad \text{else} \end{cases} \end{aligned}$$

Suppose $\alpha = 1/1000$ and $\beta = 1/1200$. Then the expected lifetimes of the bulbs are 1000 hours and 1200 hours, respectively.

Question: What is the probability that both lightbulbs last at least 900 hours?

Want:

$$\begin{aligned} \mathbb{P}\left[X \geq 900, Y \geq 900\right] &= \int_{900}^{\infty}\int_{900}^{\infty} \left(\alpha e^{-\alpha x}\right)\left(\beta e^{-\beta y}\right) dxdy \\ &= \int_{900}^{\infty} \left[\int_{900}^{\infty} \alpha e^{-\alpha x} dx\right] \left(\beta e^{-\beta y}\right) dy \\ &= \left(\int_{900}^{\infty} \alpha e^{-\alpha x} dx\right)\left(\int_{900}^{\infty} \beta e^{-\beta y} dy\right) \\ &= \left(\left[e^{-\alpha x}\right]_{x\to\infty}^{x=900}\right)\left(\left[e^{-\beta y}\right]_{y\to\infty}^{y=900}\right) \\ &= \left(e^{-900\alpha} - 0\right)\left(e^{-900\beta} - 0\right) \\ &= e^{-900(\alpha+\beta)} \\ &= e^{-900\left(\frac{1}{1000}+\frac{1}{1200}\right)} \\ &= e^{-1.65} \\ &\approx .2 \end{aligned}$$

So with probability about one fifth, both ligthbulbs will last at least 900 hours.

A faster way to do this. Observe that the events $[X \geq 900]$ and $[Y \geq 900]$ are independent, because the random variables are. Then

$$\begin{aligned} \mathbb{P}\left[X \geq 900, Y \geq 900\right] &= \mathbb{P}\left[X \geq 900\right] \mathbb{P}\left[Y \geq 900\right] \\ &= e^{-900\alpha} e^{-900\beta} \\ &= e^{-1.65} \end{aligned}$$

End of Example 71. □

Example 72 (Treasure chest)**.** A randomly looted pirate chest contains treasure, consisting of a mix of gemstones, gold pieces, and valuable navigational charts. The weight of the treasure chest is 10lbs, but the weight contribution of each type of treasure is random. Let X be the proportion of the treasure consisting of gemstones (by weight). Let Y be the proportion which is gold pieces. And let Z the proportion which consisting of charts.

Because the proportions sum to 1, it is enough to consider only two of them, X and Y.

Because X and Y are *proportions*, they take values in the following set:

$$D = \{(x,y) : 0 \leq x \leq 1, 0 \leq y \leq 1, x + y \leq 1\}$$

For this problem, we will assume the joint pdf of (X, Y) is

$$f(x,y) = \begin{cases} 24xy & : \quad (x,y) \in D \\ 0 & : \quad (x,y) \notin D \end{cases}$$

Observe that the density increases as x and y increase. So points near the diagonal boundary are *relatively more likely* than points in the bottom left corner. This is appropriate: since gold and gems are heavier than paper, we expect that most of the weight of the treasure will consist of these things, rather than navigational charts.

Question: What is the probability that more than half of the weight of the treasure comes from navigational charts, rather than gems and gold?

We want to compute

$$\mathbb{P}\left[Z \geq 1/2\right]$$

We know $X + Y + Z = 1$, so $Z = 1 - X - Y$. Therefore we want to compute

$$\mathbb{P}\left[1 - X - Y \geq 1/2\right]$$

or equivalently,

$$\mathbb{P}\left[X + Y \leq 1/2\right]$$

Let $A = \{(x,y) : 0 \leq x \leq 1, 0 \leq y \leq 1, x + y \leq .5\}$

$$\mathbb{P}\left[X + Y \leq 1/2\right] = \mathbb{P}\left[(X,Y) \in A\right] = \int_0^{.5} \int_0^{.5-x} f(x,y) dy dx$$

Some other things we can do:

- check of f is a pdf
- compute the marginal densities f_X and f_Y

End of Example 72. □

22 2025-03-07 | Week 8 | Lecture 22

Please read sections 5.1-5.4

Theorem 73 (Expectation of functions of random vectors)**.** *Let (X, Y) be a random vector with pmf $p(x, y)$ if X, Y are discrete or pdf $f(x, y)$ if X, Y are continuous. Let $h : \mathbb{R}^2 \to \mathbb{R}$ be a function. [Note: then $h(X, Y)$ is a random variable.] Then*

$$\mathbb{E}\left[h(X,Y)\right] \begin{cases} \sum_{all\ x,y} h(x,y)p(x,y) & : \ \textit{discrete case} \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y)f(x,y)dxdy & : \ \textit{continuous case} \end{cases}$$

Example 74. Gemstones are worth $\$22,500$ per pound, gold is worth $\$15,000$ per pound, and navigational charts are worth \$7500 per pound. The value of a random pound of treasure is

$$15,000X + 22,500Y + 7,500(1 - X - Y)$$

Recalling that the chest is 10lbs, the total value of the contents of the treasure chest is

$$\begin{aligned} h(X,Y) &= 150000X + 225000Y + (75000)(1 - X - Y) \\ &= 75000 + 75000X + 150000Y \end{aligned}$$

The expected total value is therefore

$$\mathbb{E}\left[h(X,Y)\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y)f(x,y)dxdy = ...\$1,650,000$$

End of Example 74. □

One important function for Theorem 73 is when $h(X, Y) = (X - \mu_X)(Y - \mu_Y)$, where $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$. That gives us a formula for what is called the *covariance*:

$$\mathrm{Cov}(X,Y) = \mathbb{E}\left[(X - \mu_X)(Y - \mu_Y)\right]$$

We also have

$$\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mu_X\mu_Y$$

For example, we can compute the covariance of X and Y from the treasure example.

$$\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mu_X\mu_Y$$

To compute $\mathbb{E}[XY]$, let's take $h(x, y) = xy$. Recalling that $f(x, y) = 24xy$ for $(x, y) \in D$, we have

$$\begin{aligned} \mathbb{E}[XY] &= \int_D h(x,y)f(x,y)dxdy \\ &= \int_D (xy)24xydxdy \\ &= 24\int_D x^2y^2dxdy \\ &= 24\int_0^1\int_0^{1-y} x^2y^2dxdy \\ &= 24\int_0^1 y^2\left[\int_0^{1-y} x^2dx\right]dy \end{aligned}$$

and so forth.

Example 75. Five friends purchased tickets to a concert. The tickets are all in the same row, next to each other, and numberd 1 through 5. What is the expected number of seats separating any two friends?

Let X, Y be the seat number of the first and second individual (chosen randomly). Possible (X, Y) pairs are

$$\{(1,2),(1,3),\ldots,(5,4)\}$$

and the joint pmf is

$$p(x,y) = \begin{cases} \frac{1}{20} & : \quad x,y \in \{1,2,3,4,5\} \text{ and } x \neq y \\ 0 & : \quad else \end{cases}$$

The number of seats separating individuals X and Y is $h(X,Y) = |X-Y|-1$.

(Make a table)

Then

$$\mathbb{E}\left[h(X,Y)\right] = \sum_{(x,y):x\neq y} h(x,y)p(x,y) = \sum_{x=1}^{5}\sum_{\substack{y=1 \\ y\neq x}}^{5} (|x-y|-1)\,\frac{1}{20} = 1.$$

End of Example 75. □

23 2025-03-10 | Week 9 | Lecture 23

Example 76 (Birthday paradox)**.** Let

$$E = [\text{At least one birthday match in a room of } n \text{ people }]$$

Want to compute

$$\mathbb{P}[E].$$

We will use

$$\mathbb{P}[E] = 1 - \mathbb{P}[E^c].$$

In words, $E^c = [\text{no birthday match}]$. Let's assume there are 365 days in a year, and that all days are equally likely. Let

$$D_j = \mathbb{P}[j\text{-th person differs from all predecessors}]$$

Then

$$\begin{aligned}\mathbb{P}[E] &= 1 - \mathbb{P}[E^c] \\ &= 1 - \mathbb{P}[D_1]\,\mathbb{P}[D_2 \mid D_1]\,\mathbb{P}[D_3 \mid D_1 D_2] \ldots \mathbb{P}[D_n \mid D_1 \cdots D_{n-1}] \\ &= 1 - \frac{365}{365} \cdot \frac{364}{365} \cdot \frac{363}{365} \cdots \frac{365-(n-1)}{365}\end{aligned}$$

When $n = 23$, $\mathbb{P}[E] > \frac{1}{2}$. When $n = 35$, $\mathbb{P}[E] \approx .8$.

End of Example 76. □

Definition 77 (iid)**.** We say that a sequence $X_1, \ldots, X_n$ of random variables is *independent and identically distributed (iid)* if

- The X_i's are independent rvs
- Every X_i has the same probability distribution as X_1

In statistics, such a sequence of r.v.s is called a *random sample from the distribution* X_1 The partial sum is

$$T_n = X_1 + \ldots + X_n$$

The *sample mean* is

$$\overline{X} = \frac{X_1 + \ldots + X_n}{n}$$

and the *sample variance* is $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$.

Recall that $\mathcal{N}(\alpha, \beta)$ means "normal distribution with mean α and variance β".

Theorem 78 (Central Limit Theorem)**.** *Let $X_1, \ldots, X_n$ be a random sample from a distribution with mean μ and variance σ^2. If n is sufficiently large, then*

$$T_n \text{ is approximately } \mathcal{N}\left(n\mu, n\sigma^2\right)$$

and

$$\overline{X} \text{ is approximately } \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

The larger n, the better the approximation. Usually $n = 30$ is big enough.

Example 79. Flip a coin 1000 times. What is the probability that you get at least 550 heads?

Let T be the total number of heads. Then T is binomial distribution with $n = 1000$ and $p = \frac{1}{2}$. For each $k \in \{0, 1, \ldots, 1000\}$, we have the formula

$$\mathbb{P}[T = k] = \binom{n}{k} p^k (1-p)^k = \binom{n}{k} \frac{1}{4^k}$$

So we want

$$\mathbb{P}\left[T \geq 550\right] = \sum_{k=550}^{1000} \binom{n}{k} \frac{1}{4^k}$$

Good luck with that. Instead, let's use the central limit theorem. Write

$$T = X_1 + X_2 + \ldots + X_{1000}$$

where

$$X_i = \begin{cases} 0 & : \quad i^{th} \text{ coin flip is tails} \\ 1 & : \quad i^{th} \text{ coin flip is heads} \end{cases}$$

Then

$$\mathbb{E}\left[X_i\right] = \frac{1}{2}$$

and

$$\begin{aligned} \text{Var}(X_i) &= \mathbb{E}\left[X_i^2\right] - \left(\mathbb{E}\left[X\right]\right)^2 \\ &= \frac{1}{2} - \left(\frac{1}{2}\right)^2 \\ &= \frac{1}{2} \end{aligned}$$

So by the central limit theorem, T is approximately $\mathcal{N}(1000\mu, 1000\sigma^2)$ where $\mu = \frac{1}{2}$ and $\sigma^2 = \frac{1}{2}$ or

$$T \sim \mathcal{N}(500, 500)$$

Therefore using the standardization trick,

$$\begin{aligned} \mathbb{P}\left[T \geq 550\right] &\approx \mathbb{P}\left[Z \geq \frac{550 - 500}{\sqrt{500}}\right] \\ &\approx \mathbb{P}\left[Z \geq 2.236\right] \\ &= \frac{1}{\sqrt{2\pi}} \int_{2.236}^{\infty} e^{-x^2/2} dx \\ &\approx 0.013 \end{aligned}$$

End of Example 79. □

24 2025-03-24 | Week 10 | Lecture 24

24.1 Shuffling Cards

I mostly talked about the paper "Trailing the Dovetail Shuffle to its Lair", by Bayer and Diaconis (1992). This is a classical paper that discusses the question, "how many times must one shuffle a deck for it to be sufficiently randomized?" The short answer is "about 7", but the mathematics to get to that answer are quite sophisticated. Some of the key ideas are as follows:

The *symmetric group of order n* is the set

$$S_n = \text{the set of permtuations of } \{1, 2, \ldots, n\} \ ,$$

where we can combine permutations by composing them together.

Formally, a function f is a *permutation* of $\{1, \ldots, n\}$ if

$$f : \{1, \ldots, n\} \to \{1, \ldots, n\}$$

and f is both injective and surjective.

Informally, a *permutation* of $\{1, \ldots, n\}$ is simply any reordering of n distinc objects.

We wrote out all the permutations of $\{1, 2, 3\}$ and showed how to compose them. Composing permutations is like doing one permutation and then doing the other. The result is always a new permutation.

Key idea: shuffling a deck of 52 cards induces a probability distribution on the set S_{52}. The shuffling procedure "fully randomizes" the deck if the resulting probability distribution is close to the uniform distribution on S_{52}.

The first 6 or so riffle shuffles don't do much to randomize, but after 7 shuffles there is an inflection point and, the distribution quickly converges to uniform.

24.2 The Coupon Collector Problem

Problem statement: Suppose a dart board is divided up into n equal-sized sections, labeled 1 through n. Each dart thrown is equally likely to hit any one of the sections. How many times must one throw the dart so that the probability of all sections being hit is at least 99%?

Solution: Let m be any positive integer greater than n. Let

$$V = \text{number of times until each section is hit at least once}$$

and let

$$A_b = \text{section } b \text{ is not hit in the first } m \text{ throws.}$$

To solve the problem, we will need to determine how big m must be so that

$$\mathbb{P}[V > m] < 0.01.$$

Then, if we throw at least that many darts, the probability of all sections being hit will be at least .99.

$$\begin{aligned}
\mathbb{P}[V > m] &= \mathbb{P}[A_1 \cup A_2 \cup \ldots \cup A_n] \\
&\leq \mathbb{P}[A_1] + \mathbb{P}[A_2] + \ldots + \mathbb{P}[A_n] && \text{(the union bound)} \\
&\leq n\left(1 - \frac{1}{n}\right)^m \\
&= n\left[\left(1 - \frac{1}{n}\right)^n\right]^{\frac{m}{n}} \\
&\approx n\left[e^{-1}\right]^{\frac{m}{n}} \\
&= ne^{-m/n}
\end{aligned}$$

Setting

$$ne^{-m/n} < 0.01$$

we can solve for m:

$$e^{-m/n} < \frac{0.01}{n}$$

$$-\frac{m}{n} < \log\left(\frac{0.01}{n}\right)$$

or

$$m > -n\log\left(\frac{0.01}{n}\right)$$

or equivalently,

$$m > n\log(100n)$$

In other words, if we throw at least $n\log(100n)$ darts, then with probability at least 99%, we will hit every section on the dartboard.

25 2025-03-28 | Week 10 | Lecture 25

Start reading 6.1, 6.2

Recall that we say that a sequence $X_1, X_2, \ldots$ of random variables is *independent and identically distributed (IID)* if

- The X_i's are independent rvs
- Every X_i has the same probability distribution as X_1

Theorem 80 (Central limit theorem)**.** *Suppose we have an IID sequence of random variables $X_1, X_2, \ldots$ with finite mean $\mathbb{E}[X_1] = \mu$ and finite variance $\mathrm{Var}(X_1) = \sigma^2 > 0$. Let*

$$S_n = X_1 + \ldots + X_n.$$

Then for any fixed $-\infty \le a \le b \le +\infty$, we have

$$\lim_{n\to\infty} \mathbb{P}\left[a \le \frac{S_n - n\mu}{\sigma\sqrt{n}} \le b\right] = \Phi(b) - \Phi(a) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy \tag{23}$$

The conclusion of the CLT can be stated in many equivalent ways. One useful way involves the sample average $\overline{X} := \frac{S_n}{n}$. The CLT says that

$$\overline{X} \approx \underbrace{\mu + \frac{\sigma}{\sqrt{n}} Z}_{\mathcal{N}(\mu, \frac{\sigma^2}{n})} \tag{24}$$

where $Z \sim \mathcal{N}(0,1)$.

Usually $n \ge 30$ is big enough for the approximation to be valid.

Example 81. A witch decides to make 50 batches of potions. Unfortunately, her hut is not very clean and bugs keep crawling into the brew. Suppose that the amount of bugs contaminating the ith batch is a random variable X_i with mean $\mu = 4g$ and standard deviation $\sigma = 1.5g$. Assuem the witch prepares her 50 batches independently. Let

$$\overline{X} = \frac{X_1 + \ldots + X_{50}}{50}$$

be the average amount of impurity.

Question: What is the probability that $3.5 \le \overline{X} \le 3.8$?
Solution: Here, $n = 50 > 30$ so we can apply the central limit theorem. By the central limit theorem, $\overline{X}$ has distribution approximately

$$\begin{aligned}
\overline{X} &\approx \mu + \frac{\sigma}{\sqrt{n}} Z \\
&= 4 + \frac{3}{2\sqrt{50}} Z
\end{aligned}$$

where $Z \sim \mathcal{N}(0,1)$ is a standard normal random variable. Therefore

$$\begin{aligned}
\mathbb{P}\left[3.5 \le \overline{X} \le 3.8\right] &\approx \mathbb{P}\left[3.5 \le 4 + \frac{3}{2\sqrt{50}} Z \le 3.8\right] \\
&= \mathbb{P}\left[-\frac{1}{2} \le \frac{3}{2\sqrt{50}} Z \le -\frac{1}{5}\right] \\
&= \mathbb{P}\left[-\frac{\sqrt{50}}{3} \le Z \le -\frac{2\sqrt{50}}{15}\right] \\
&= \mathbb{P}\left[-2.36 \le Z \le -.94\right] \\
&= \int_{-2.36}^{-.94} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy \\
&\approx .16
\end{aligned}$$

End of Example 81. □

Example 82 (Application of CLT)**.** The human genome consists of a sequence of nucleotides (A, T, C, G). For this problem we can think of the human genome as a sequence of nucleotides

`CAGGAAGCATTTATATAGAGGTAATAATTAAATTTAACTTTTTATTATTCTCTATGCCCAAGAGAATGTG...`

and the sequence consists of $n = 3.2 \times 10^9 = 3,200,000,000$ nucleotide letters.

The DNA mutation rate in humans is approximately $\mu = 2 \times 10^{-8} = 0.00000002$ mutations per nucleotide per generation. That means each time a person is born, each nucleotide either mutates (with proability μ) or doesn't mutate (with probability $1 - \mu$). Assume that nucleotides mutate *independently*, that is, whether one nucleotide undergoes mutates is indepdenent from any other nucleotides.

Let S_n be the number of new mutations in the genome of a randomly-selected newborn baby. We can write

$$S_n = X_1 + \ldots + X_n$$

where

$$X_i = \begin{cases} 1 & : \text{ mutation at site } i \\ 0 & : \text{ no mutation at site } i \end{cases}$$

Let M be the number of new mutations in the genome of a randomly selected newborn baby. Then $M \sim \text{Bin}(n, \mu)$ where the "number of trials" is $n = 3.2$ billion and the "success probability" is $\mu = 2 \times 10^{-8}$.

Question 1: On average, how many mutations does a newborn baby have?
Solution: We want

$$\begin{aligned} \mathbb{E}[S_n] &= \mathbb{E}[X_1 + \ldots + X_n] \\ &= \mathbb{E}[X_1] + \ldots + \mathbb{E}[X_n] \\ &= n\mu \\ &= (3.2 \times 10^9)(2 \times 10^{-8}) \\ &= 64. \end{aligned} \tag{25}$$

Question 2: What is the probability distribution of S_n?
Solution: S_n is a sume of n "coin flips" each having success probability μ. Hence, S_n is a binomial random variable with number of trials n and success parameter μ.

Question 3: What percentage of people have between 56 and 72 mutations?
Solution: Use the central limit theorem. First observe that

$$\mathbb{E}[X_1] = 1 \cdot \mu + 0 \cdot (1 - \mu) = \mu \tag{26}$$

and

$$\mathbb{E}\left[X_1^2\right] = 1^2 \cdot \mu + 0^2 \cdot (1 - \mu) = \mu$$

Therefore

$$\begin{aligned} \sigma^2 &= \text{Var}(X_1) \\ &= \mathbb{E}\left[X_1^2\right] - (\mathbb{E}[X_1])^2 \\ &= \mu - \mu^2. \end{aligned}$$

Since μ is very small μ^2 is VERY VERY small, and hence we have $\sigma^2 \approx \mu$. Therefore

$$\sigma \approx \sqrt{\mu} = \sqrt{2} \times 10^{-4}. \tag{27}$$

Finally, we have

$$\sqrt{n} = 40000\sqrt{2}. \tag{28}$$

Therefore, using Eq. (23) from the Central limit theorem, we have

$$\mathbb{P}\left[a \leq \frac{S_n - n\mu}{\sigma\sqrt{n}} \leq b\right] \approx \int_a^b \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy$$

and plugging in the values from Eqs. (25), (27) and (28), we get

$$\mathbb{P}\left[a \leq \frac{S_n - 64}{8} \leq b\right] \approx \int_a^b \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy$$

or equivalently

$$\mathbb{P}\left[64 + 8a \leq S_n \leq 64 + 8b\right] \approx \int_a^b \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy$$

Taking $a = -1$ and $b = 1$, we get

$$\mathbb{P}\left[56 \leq S_n \leq 72\right] \approx \int_{-1}^{1} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy \approx .68.$$

Conclusion: about 68% of people have between 56 and 72 mutations.

What percentage of people have between 48 and 80 mutations?

End of Example 82. □

26 2025-03-31 | Week 11 | Lecture 26

26.1 Point estimation

Section 6.1 Suppose the fraction of people in Hawaii who like broccoli is p. This is an unknown parameter. We don't know what p is, but we would like to estimate it. In principle, we could ask everyone in the state, but that would be impractical. Instead, we take a random sample: we choose randomly n individuals and ask each of them whether they like broccoli or not. Let

$$\widehat{p} := \text{ fraction of our sample who like broccoli}$$

Suppose we sample $n = 100$ individuals and 20 of them like broccoli. Then

$$\widehat{p} = \frac{20}{100} = .2$$

Here, $\widehat{p}$ is an example of a *point estimate* of p.

Definition 83 (Point estimate and point estimator)**.** In statistics, we think of "data" as consisting of n IID random variables $X_1, \ldots, X_n$, called a *random sample*.

(i) A *point estimator* is any function $F(X_1, \ldots, X_n)$ of a sample.

(ii) A *point estimate* $\widehat{\theta}$ is any single number, computed from the data, which can be regarded sensible value of some unknown parameter θ.

In general, an *estimator* is a function of the sample, while an *estimate* is the realized numerical value of the estimator that is computed after the sample is actually observed. The standard convention is to refer to point estimates and point estimators by the same letter $\widehat{\theta}$.

For our broccoli example, the data consists of n random variables $X_1, \ldots, X_n$ were

$$X_i = \begin{cases} 1 & : \text{ i-th sampled person likes broccoli} \\ 0 & : \text{ they don't} \end{cases}$$

Our point estimate is $\widehat{p} = 0.2$ and the point estimator is the function $\widehat{p} = \frac{X_1 + \ldots + X_n}{n}$.

Taking a step back, we can think of our estimate as

$$\widehat{p} = p + \text{error of estimation}$$

It is perhaps clear that if we increase the sample size n, then the estimation error should decrease, so that

$$\widehat{p} \to p.$$

But we would like to precisely quantify the accuracy of the estimate. This brings us to the next subject.

26.2 Confidence intervals

Section 7.1

Definition 84 (confidence interval)**.** An interval $(\widehat{p} - \epsilon, \widehat{p} + \epsilon)$ is called *95% confidence interval* for the unknown parameter p if $\epsilon > 0$ is chosen large enough that

$$\mathbb{P}\left[|\widehat{p} - p| < \epsilon\right] \geq .95 \tag{29}$$

In words, the interval contains the true parameter with probability at least .95.

Can we come up with a 95% confidence interval for our broccoli poll? We already have $\widehat{p} = .2$. But we don't yet know how big ϵ needs to be. (Of course we want to pick the smallest ϵ possible, as that would give us a more informative confidence interval). To figure out how big ϵ must be, we'll have to use the central limit theorem again.

Formally, let

$$\widehat{p} = \frac{X_1 + \ldots + X_n}{n}$$

where

$$X_i = \begin{cases} 1 & : \text{ i-th sampled person likes broccoli} \\ 0 & : \text{ they don't} \end{cases}$$

Then

$$\widehat{p} \sim \text{Bin}(n, p)$$

We don't know what p is, but we can still say that by the central limit theorem,

$$\widehat{p} \approx p + \frac{\sigma}{\sqrt{n}} Z \tag{30}$$

where $Z \sim \mathcal{N}(0,1)$ and $\sigma^2 = \text{Var}(X_1) = p(1-p)$. We also don't know what σ^2 is, since it depends on the unknown parameter p. That's okay, we'll cope. It will be enough to observe that for any vale of $p \in [0,1]$, we have

$$\sigma^2 = p(1-p) \leq \frac{1}{4}$$

and therefore that

$$\sigma \leq \frac{1}{2}. \tag{31}$$

We will also use the following fact (and future homework problem):

$$\boxed{\mathbb{P}\left[|Z| < u\right] = 2\Phi(u) - 1} \tag{32}$$

By Eq. (30),

$$|\widehat{p} - p| = \frac{\sigma}{\sqrt{n}} |Z|$$

Therefore

$$\begin{aligned} \mathbb{P}\left[|\widehat{p} - p| < \epsilon\right] &\approx \mathbb{P}\left[\frac{\sigma}{\sqrt{n}}|Z| < \epsilon\right] \\ &= \mathbb{P}\left[|Z| < \frac{\epsilon\sqrt{n}}{\sigma}\right] \\ &\geq \mathbb{P}\left[|Z| < \frac{\epsilon\sqrt{n}}{1/2}\right] \qquad \text{(since } \sigma \leq 1/2\text{).} \\ &= \mathbb{P}\left[|Z| < 2\epsilon\sqrt{n}\right] \\ &= 2\Phi(2\epsilon\sqrt{n}) - 1 \end{aligned}$$

In conclusion we have:

$$\boxed{\mathbb{P}\left[|\widehat{p} - p| < \epsilon\right] \geq 2\Phi(2\epsilon\sqrt{n}) - 1} \tag{33}$$

Thus in order for the left hand side to be greater than .95, it will suffice for

$$2\Phi(2\epsilon\sqrt{n}) - 1 \geq 0.95$$

or equivalently

$$\Phi(2\epsilon\sqrt{n}) \geq 0.975.$$

Using a computer, we can determine that $\Phi(u) \geq 0.975$ whenever $u \geq 1.96$. Thus, we need

$$2\epsilon\sqrt{n} \geq 1.96.$$

For our problem, $n = 100$, so $\sqrt{n} = 10$, so we need ϵ to be at least as big as

$$\epsilon \geq 1.96/20 = 0.098.$$

Thus, our confidence interval is

$$(\widehat{p} - \epsilon, \widehat{p} + \epsilon) = (0.2 - 0.098, 0.2 + 0.098) = (.102, .298).$$

We conclude that with 95% confidence, the true proportion of people in Hawaii who like broccoli is between 10.2% and 29.8%. Our point estimate is 0.2, and our *margin of error* is 0.098.

If we had sampled more people, our margin of error would be smaller. How many people would we have to sample to reduce the margin of error to 0.05? What about to 0.01?

27 2025-04-02 | Week 11 | Lecture 27

27.1 Property of point estimators: Bias

Section 6.1

Let $\theta \in \mathbb{R}$ be an unknown parameter. Suppose that $\widehat{\theta}$ is a point estimator of θ. We say that $\widehat{\theta}$ is *unbiased* if

$$\mathbb{E}\left[\widehat{\theta}\right] = \theta.$$

The *bias* of $\widehat{\theta}$ is the number $\mathbb{E}\left[\widehat{\theta}\right] - \theta$.

For clarity, recall that as an estimator, $\widehat{\theta}$ is a function $\widehat{\theta} = f(X_1, \ldots, X_n)$. And so when we write

$$\mathbb{E}\left[\widehat{\theta}\right]$$

we really mean

$$\mathbb{E}\left[f(X_1, \ldots, X_n)\right]$$

In the broccoli example, the estimator we used was the *sample mean:*

$$\widehat{\theta}(X_1, \ldots, X_n) = \frac{X_1 + \ldots + X_n}{n}.$$

This estimator is unbiased because

$$\mathbb{E}\left[\frac{X_1 + \ldots + X_n}{n}\right] = \frac{1}{n}\left(\mathbb{E}\left[X_1\right] + \ldots + \mathbb{E}\left[X_n\right]\right) = \frac{1}{n} np = p.$$

Example 85 (A biased estimator)**.** Suppose I roll a dice behind a screen, shouting out the number rolled each time. Your job is to estimate the number of sides of the dice, which we will call θ. For n rolls, the data consists of n random variables

$$X_1, \ldots, X_n$$

where X_i is the i^{th} dice roll. One natural estimator is

$$\widehat{\theta}(X_1, \ldots, X_n) = \max\left\{X_1, \ldots, X_n\right\}$$

This estimator is called the *running maximum*. For example, if we roll the hidden dice four times and get the following numbers for $X_1, \ldots, X_4$:

$$3, 11, 7, \text{ and } 4,$$

then our estimate would be that the dice has 11 sides.

If I roll the dice enough times, I will eventually roll the highest number; it follows from this that

$$\lim_{n \to \infty} \widehat{\theta} = \theta.$$

But we expect it to be biased (since it will "usually" be less than the true number of sides, and will never be more).

To make this precise, we can use the law of total expectation. Let $p = \mathbb{P}\left[\widehat{\theta} = \theta\right]$, so that $1 - p = \mathbb{P}\left[\hat{\theta} < \theta\right]$.

$$\begin{aligned}
\mathbb{E}\left[\widehat{\theta}\right] &= \mathbb{E}\left[\widehat{\theta} \mid \widehat{\theta} = \theta\right] p + \mathbb{E}\left[\widehat{\theta} \mid \widehat{\theta} < \theta\right] (1 - p) \\
&= \mathbb{E}\left[\theta \mid \widehat{\theta} = \theta\right] p + \mathbb{E}\left[\widehat{\theta} \mid \widehat{\theta} < \theta\right] (1 - p) \\
&< \mathbb{E}\left[\theta \mid \widehat{\theta} = \theta\right] p + \mathbb{E}\left[\theta \mid \widehat{\theta} < \theta\right] (1 - p) \\
&= \theta p + \theta(1 - p) \\
&= \theta.
\end{aligned}$$

End of Example 85. □

Example 86. Here's another method to estimate the number of sides θ of the hidden dice.

Let Y be the dice roll. We know that since the dice has θ sides, it has expected value

$$\mathbb{E}[Y] = \frac{1}{\theta}(1 + 2 + \ldots + \theta) = \frac{\theta + 1}{2}.$$

So maybe let's take the sample average, which in this case is

$$\frac{X_1 + \ldots + X_4}{4} = \frac{22}{4}$$

and see what value of θ would give us the closest average to this number.

That is, let us take as our estimate $\widehat{\theta}$ the value satisfying the equation

$$\frac{\widehat{\theta} + 1}{2} = \frac{22}{4}. \tag{34}$$

Solving for $\widehat{\theta}$, we get $\widehat{\theta} = 10$.

End of Example 86. □

Theorem 87 (Law of Large Numbers)**.** *Let $X_1, X_2 \ldots$ be a sequence of IID random variables. Such that $\mu := \mathbb{E}[X_1]$ is finite, $\sigma^2 := \mathrm{Var}(X_1)$ is finite, and the fourth moment $\mathbb{E}\left[X_i^4\right] < \infty$. Then*

$$\frac{X_1, \ldots, X_n}{n} \to \mathbb{E}[X_1] \qquad \text{as } n \to \infty$$

This theorem says that if you flip a coin repeatedly, the proportion of heads will converge to 1/2.

28 2025-04-04 | Week 11 | Lecture 28

exam average was 30.75 (among those who took the exam)

28.1 The method of moments

Example 88 (Method of Moments)**.** Suppose $X_1, \ldots, X_n$ are the waiting times, in minutes, between customer orders at Raising Cane's at University and King St. Assume that the waiting times are IID exponentially distributed with some unknown rate parameter λ.

Suppose the first 10 waiting times are

$$0.2, 0.1, 1, 0.7, 0.8, 0.2, 0.5, 0.5, 0.4, 2$$

Question: On average, how many orders would you expect per minute?

So the average wait time is

$$\frac{0.2 + 0.1 + 1 + 0.7 + 0.8 + 0.2 + 0.5 + 0.5 + 0.4 + 2}{10} = .64 \text{ minutes/order}$$

So a good estimate for the number of orders per minute is

$$1/.64 \approx 1.56 \text{ orders/minute}$$

We have estimated the rate.

In other words, the pdf for every X_i is

$$f(x) = \lambda e^{-\lambda x}, \quad \lambda \geq 0$$

We have

$$\mathbb{E}[X_1] = \int_0^\infty x f(x) dx = \ldots = \frac{1}{\lambda}$$

' We also know by the Law of Large Numbers that

$$\overline{X} := \frac{X_1 + \ldots + X_n}{n} \to \mathbb{E}[X_1] \quad \text{as } n \to \infty$$

Thus, when n is large, one good way to estimate λ would be to set

$$\overline{X} = \frac{1}{\lambda}$$

and then solve for λ:

$$\lambda = \frac{1}{\overline{X}}.$$

This is precisely what we did. Thus we take as our estimator the function $\widehat{\lambda} = \frac{1}{\overline{X}}$.

This example was brought to you by Raising Cane's.

End of Example 88. □

This example illustrates one general method of estimation, called *the method of moments.* In this method, one has an IID random sample $X_1, \ldots, X_n$ all having distribution X. It is assumed that you know the type of the distribution of X (e.g., exponential, normal, binomial, whatever) but you don't know one or more numerical parameters $\theta_1, \ldots \theta_m$ (e.g., the mean, variance, rates, etc). You then consider one or more of equations of the following form:

$$\begin{aligned}
\frac{X_1 + \ldots + X_n}{n} &= \mathbb{E}[X] \\
\frac{X_1^2 + \ldots + X_n^2}{n} &= \mathbb{E}\left[X^2\right] \\
\frac{X_1^3 + \ldots + X_n^3}{n} &= \mathbb{E}\left[X^3\right] \\
&\vdots
\end{aligned}$$

In these equations, the LHS is what you observed in your data (e.g., $\overline{X} = .64$ orders per minute). The right-hand sides are the moments, which are known functions of the parameters $\theta_1, \ldots, \theta_m$. You then solve the system of equations for $\theta_1, \ldots, \theta_m$. This gives you numerical point estimates for the unknown parameters $\theta_1, \ldots, \theta_m$.

28.2 The Gamma Distribution

The *gamma function* is

$$\Gamma(\alpha) := \int_0^\infty x^{\alpha-1} e^{-x} dx, \quad \alpha > 0.$$

Fact: For any positive integer $n \geq 1$,

$$\Gamma(n) = (n-1)!$$

For any $\alpha > 1$,

$$\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$$

We say that a random variable X has a *gamma distribution with shape parameter $\alpha > 0$ and scale parameter $\beta > 0$* if X has pdf of X is

$$f(x) = \begin{cases} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} & : \quad x \geq 0 \\ 0 & : \quad x < 0 \end{cases}$$

Fact: If X is gamma distributed with shape $\alpha > 0$ and scale $\beta > 0$, then

$$\mathbb{E}[X] = \alpha\beta \quad \text{and} \quad \mathbb{E}\left[X_1^2\right] = \alpha(\alpha+1)\beta^2 \quad \text{and} \quad \text{Var}(X) = \alpha\beta^2$$

This is a very flexible class of distributions, often used in general practice for things like

- amount of degredation or wear
- random rates (e.g. cancer, dna mutation, etc)
- survival and waiting times (e.g. for a given number of events to occur, or how long you survive after being exposed to a high dose of radiation, etc)
- sum of IID exponential waiting times.
- has nice mathematical properties, so it's COMMONLY used in (advanced) Bayesian statistical techniques
- etc

Verification that f is a probability density function: Let $x \geq 0$. THen

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} \geq 0.$$

We need to show that

$$\int_0^\infty f(x) dx = 1$$

This can be shown using a u-subsitution. We will do the u-subsituttion $u = x/\beta$, so that $x = \beta u$ and

$dx = \beta du$

$$\begin{aligned}\int_0^\infty f(x)dx &= \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-x/\beta} dx \\ &= \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^\infty (\beta u)^{\alpha-1} e^{-u} \beta du \\ &= \frac{1}{\beta^\alpha \Gamma(\alpha)} \beta^{\alpha-1} \beta \int_0^\infty u^{\alpha-1} e^{-u} du \\ &= \frac{1}{\Gamma(\alpha)} \underbrace{\int_0^\infty u^{\alpha-1} e^{-u} du}_{=\Gamma(\alpha)} \\ &= 1.\end{aligned}$$

We've shown that f is nonnegative and integrates to 1. Therefore it is a pdf.

Example 89. Suppose $X_1, \ldots, X_n$ are IID gamma distributed random variables with parameters α and β which are unknown to us.
Know:

Can we estimate α and β from data? for example, set

$$\frac{X_1 + \ldots + X_n}{n} = \mathbb{E}[X_1]$$

and

$$\frac{X_1^2 + \ldots + X_n^2}{n} = \mathbb{E}\left[X_1^2\right]$$

We know that for a gamma-distributed random variable X with parameters $\hat{\alpha}$ and $\hat{\beta}$, we have $\mathbb{E}[X] = \hat{\alpha}\hat{\beta}$ and $\mathbb{E}\left[X_1^2\right] = \hat{\alpha}(\hat{\alpha}+1)\hat{\beta}^2$. So these equations become

$$\frac{X_1 + \ldots + X_n}{n} = \hat{\alpha}\hat{\beta}$$

and

$$\frac{X_1^2 + \ldots + X_n^2}{n} = \hat{\alpha}(\hat{\alpha}+1)\hat{\beta}^2$$

We now solve for $\hat{\alpha}$ and $\hat{\beta}$. Suppose that we sample data which is such that

$$\frac{X_1 + \ldots + X_n}{n} = c$$

and

$$\frac{X_1 + \ldots + X_n}{n} = d$$

where c, d are numbers. Then

$$\hat{\alpha}\hat{\beta} = c$$

and

$$d = (\hat{\alpha}\hat{\beta})^2 + (\hat{\alpha}\hat{\beta})\hat{\beta}$$

so that

$$d = c^2 + c\hat{\beta}$$

Solving for $\hat{\alpha}$ and $\hat{\beta}$ gives

$$\hat{\beta} = \frac{d - c^2}{c} \quad \text{and} \quad \hat{\alpha} = \frac{c^2}{d - c^2}.$$

End of Example 89. □

29 2025-04-07 | Week 12 | Lecture 29

29.1 Hypothesis Testing

Textbook section 8.1

The hypothesis testing framework: suppose you observe some random quantity, which is assumed to depend on some unknown numerical parameter θ. You make a *hypothesis*, which can be any claim about the value of some parameter.

We are interested in testing the following two hypotheses:

- The *null hypothesis:* whatever effect you observed was due to random chance. This is denoted H_0.
- The *alternative hypothesis:* the claim you made about the parameter. This is usually denoted H_1

Based on our observation(s) we will make a decision to either reject H_0, or fail to reject H_0.

Suppose the observed data is X. The *p-value* of X is defined as

$$\text{p-value} := \mathbb{P}\left[\text{observing a result at least as extreme as } X \mid H_0\right]$$

When the p-value is small, we **reject** H_0. When the p-value is not small, we **fail to reject** H_0**.** The meaning of "small" here is subjective.

This is the basic story. Let's do an example:

Example 90. Suppose you have a coin but you don't know if it is a fair coin or not. You think maybe the coin is biased in favor of flipping heads. As an experiment, you flip the coin 10 times and you get 8 heads. Is this convincing evidence that the coin is biased?

In this case, the data is an IID sample

$$X_1, \ldots, X_{10}$$

where

$$X_i = \begin{cases} 1 & : \text{ with probability } \theta \\ 0 & : \text{ with probability } 1-\theta \end{cases}$$

The null hypothesis is

$$H_0 : \theta = \frac{1}{2}$$

The alternative hypothesis is that heads is more likely than tails:

$$H_1 : \theta > \frac{1}{2}$$

The p-value is

$$\begin{aligned} p &= \mathbb{P}\left[\text{at least 8 heads} \mid \theta = \frac{1}{2}\right] \\ &= \mathbb{P}\left[S_n = 8 \mid \theta = \frac{1}{2}\right] + \mathbb{P}\left[S_n = 9 \mid \theta = \frac{1}{2}\right] + \mathbb{P}\left[S_n = 10 \mid \theta = \frac{1}{2}\right] \\ &= \binom{10}{8}\theta^8(1-\theta)^2 + \binom{10}{9}\theta^9(1-\theta)^1 + \binom{10}{10}\theta^{10}(1-\theta)^0 \text{ with } \theta = \frac{1}{2} \\ &= \binom{10}{8}\left(\frac{1}{2}\right)^{10} + \binom{10}{9}\left(\frac{1}{2}\right)^{10} + \binom{10}{10}\left(\frac{1}{2}\right)^{10} \\ &= \frac{7}{128} \\ &\approx 0.0546. \end{aligned}$$

In words, if the coin were fair, we would expect it to flip 8 or more heads (out of ten) about 5% of the time. That's not actually super rare, so while this is consistent with the coin being biased, it's also consistent

with the null hypothesis. Hence, I wouldn't find this to be convincing evidence that the coin isn't fair, so I wouldn't reject the null hypothesis.

Maybe you would. The cutoff for what is "convincing" is necessarily subjective.

End of Example 90. □

Example 91. Now, suppose that we flip the coin 100 times and we get 80 of them to be heads. Is this convincing evidence that the coin isn't fair?

To answer this question, we'll compute the p-value. Let $X_1, \ldots, X_{100}$ be our IID random sample, where

$$X_i = \begin{cases} 1 & : \text{ if } i\text{th coin is heads} \\ 0 & : \text{ otherwise} \end{cases}$$

Then the number of heads is the random variable

$$S_n = X_1 + \ldots + X_n.$$

Observe that $\mu = \mathbb{E}[X_i] = \frac{1}{2}$ and Assuming the coin is a fair coin, we have

$$\mu = \mathbb{E}[X_i] = \frac{1}{2}$$

and

$$\sigma^2 = \mathbb{E}\left[X_i^2\right] - \frac{1}{4} = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$$

so that

$$\sigma = \frac{1}{2}.$$

By the central limit theorem, we know that

$$\frac{S_n}{n} \approx \mu + \frac{\sigma}{\sqrt{n}} Z,$$

where Z is a standard normal random variable. For this specific problem, we have:

$$\frac{S_{100}}{100} \approx \frac{1}{2} + \frac{1}{20} Z.$$

We can now compute the p-value:

$$\begin{aligned}
\text{p-value} &= \mathbb{P}[\text{at least 80 heads out of 100 tosses} \mid \text{coin is fair}] \\
&= \mathbb{P}[S_{100} \geq 80 \mid \text{coin is fair}] \\
&= \mathbb{P}\left[\frac{S_{100}}{100} \geq .8 \mid \text{coin is fair}\right] \\
&\approx \mathbb{P}\left[\frac{1}{2} + \frac{Z}{20} \geq .8\right] && \text{Key step. Follows by CLT.} \\
&= \mathbb{P}\left[\frac{Z}{20} \geq .3\right] \\
&= \mathbb{P}[Z \geq 6] \\
&\approx .000000001
\end{aligned}$$

In other words, if the coin were a fair coin, observing heads at least 80 out of 100 times would be extraordinarily unlikely. Therefore we reject H_0.

End of Example 91. □

29.2 The Matching Problem

Suppose there are n people in class. Everyone writes their name on one playing card. The cards are then shuffled, and dealt out again, so that each person gets a new (random) card. What is the probability that at least one person gets their original card?

By enumerating possibilities, we calculated that this has probability $2/3$ when there are $n = 3$ people.

We will return to this problem later.

30 2025-04-09 | Week 12 | Lecture 30

30.1 Simple random walk

Let's track the value of a stock. Suppose that every minute, the stock value goes up \$1 or down \$1. Then the change in stock value from the start of the day is

$$S_n = X_1 + \ldots + X_n$$

where $X_1, \ldots, X_n$ are IID random variables with

$$X_i = \begin{cases} +1 & : \text{ with probability } \theta \\ -1 & : \text{ with probability } 1 - e\theta \end{cases}$$

Suppose we track the value of the stock over the course of 400 minutes, (i.e., about 7 hours). And we observe that the value of the stock is down \$68.

The standard assumption is that the stock price fluctates randomly, with no tendency to go up or down. That is, the null hypothesis is that $\theta = \frac{1}{2}$. Is the observed drop in stock price consistent with this hypothesis? Or, alternatively, is there a reason that the stock is down (the alternative hypothesis)?

Let's compute the p-value. We want to compute

$$\mathbb{P}\left[S_{400} \leq -68 \mid \theta = \frac{1}{2}\right]$$

First observe that

$$\mu = \mathbb{E}[X] = 0 \quad \text{and} \quad \sigma^2 = \mathbb{E}\left[X^2\right] = 1.$$

By the central limit theorem,

$$\frac{S_n}{n} \approx \mu + \frac{\sigma}{\sqrt{n}} Z$$

where Z is a standard normal. Taking $n = 400$, $\mu = 0$, and $\sigma = 1$, we get

$$\frac{S_{400}}{400} \approx \frac{1}{20} Z$$

Therefore

$$\begin{aligned} \mathbb{P}\left[S_{400} \leq -68 \mid \theta = \frac{1}{2}\right] &= \mathbb{P}\left[\frac{S_{400}}{400} \leq -\frac{68}{400} \mid \theta = \frac{1}{2}\right] \\ &\approx \mathbb{P}\left[\frac{Z}{20} \leq -\frac{68}{400}\right] \\ &= \mathbb{P}\left[Z \leq -\frac{68}{20}\right] \\ &= \mathbb{P}[Z \leq -3.4] \\ &= \int_{-\infty}^{-3.4} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy \\ &\approx 0.0003. \end{aligned}$$

In this case, observing such a precipitous drop is very unlikely under the assumption of the null hypothesis. So we might reasonably reject the null hypothesis. It is much more likely that something caused the price drop.

30.2 The Matching Problem

Suppose there are n people in class. Everyone writes their name on one playing card. The cards are then shuffled, and dealt out again, so that each person gets a new (random) card. What is the probability that at least one person gets their original card?

By enumerating possibilities, we calculated that this has probability 2/3 when there are $n = 3$ people.

30.3 Permutations:

- When we shuffle the cards, we permute their order. This induces a reordering of the cards, called a *permutation.*
- Introduce cycle notation
- Define *cycle*
- Define *order* a cycle
- Define *fixed point* of a permutatiion

Question 1: What is the probability that at least one person gets their original card?
Question 2: What is the expected number of fixed points?
Question 3: What is the expected number of cycles of order 2, 3,... ?
Question 4: What is the expected number of cycles?

31 2025-04-11 | Week 12 | Lecture 31

We continued the activity from Wednesday about shuffling cards and permutations. I numbered the cards $1, 2, \ldots$. Then I had each person write their name on the card. Then I collected and shuffled the cards and then dealt them back randomly to all the people.

Then we had everyone stand next to the person whose name is on their new card. (Usually, you don't get your own card back, because the cards were shuffled). This results in several distinct groups of people. These are called *cycles*. The *order* of a cycle is the number of people. A cycle of order 1 is called a *fixed point*. You are a fixed point if you got your original card back. In that case, you have to stand by yourself.

We can think of permutations as functions. Formally, a *permutation on the set* $\{1, 2, \ldots, n\}$ is a function

$$\pi : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}$$

such that π is both injective and surjective (i.e., is a bijection).

By shuffling the deck, we are generating a permutation uniformly at random. We did this experiment several times. The number of fixed points, cycles, and the size of the cycles are all random.

We also introduce *cycle notation*, which is described really good here, (up to about 7 minutes and 40 seconds):

`https://www.youtube.com/watch?v=MpKG6FmcIHk`

Question #1: Suppose we draw a permutation on the set $\{1, 2, \ldots, n\}$ uniformly at random. What's the expected number of fixed points? (That is, on average, how many people get their own card back?)

Let F_i be the event that i is a fixed point. Clearly,

$$\mathbb{P}[F_i] = \frac{1}{n}$$

since each person has a 1-in-n chance of getting their own card back.

Let N be the number of fixed points (this is a random variable). Then

$$N = \mathbf{1}_{F_1} + \mathbf{1}_{F_2} + \ldots + \mathbf{1}_{F_n}$$

Therefore

$$\begin{aligned}
\mathbb{E}[N] &= \mathbb{E}[\mathbf{1}_{F_1} + \mathbf{1}_{F_2} + \ldots + \mathbf{1}_{F_n}] \\
&= \mathbb{E}[\mathbf{1}_{F_1}] + \mathbb{E}[\mathbf{1}_{F_2}] + \ldots + \mathbb{E}[\mathbf{1}_{F_n}] \\
&= \mathbb{P}[F_1] + \mathbb{P}[F_2] + \ldots + \mathbb{P}[F_n] \\
&= \frac{1}{n} + \frac{1}{n} + \ldots + \frac{1}{n} \\
&= 1.
\end{aligned}$$

Therefore, on average, we expect 1 fixed point.

Question #2: What is the probability of having no fixed points?

By enumerating all the permutations, we showed that for $n = 1, 2, 3, 4$, the probability of no fixed points is

$$1, \quad 1/2, \quad 1/3, \quad 3/8$$

which suggests the following conjecture

$$\mathbb{P}[N = 0] = \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \frac{1}{5!} + \ldots + (-1)^n \frac{1}{n!}$$

This would mean that when n is large,

$$\mathbb{P}[N = 0] \approx \frac{1}{e} \approx 0.37$$

32 2025-04-21 | Week 14 | Lecture 32

Chapter 14.1 through 14.3

32.1 Tests of Association

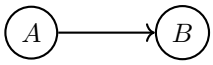
Suppose we want to know if two things A and B are related to each other. We will discuss only the case where A and B are discrete, categorical events. For example

$$A = \text{[a person vapes]}$$

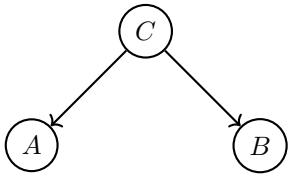
$$B = \text{[person gets coronary artery disease]}$$

There are many possible relationships:

- **Causality:** Vaping causes arterial disease.

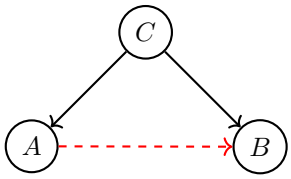


- **Common Response:** Perhaps the effect of vaping on your ateries is perfectly benign, but there is some other variable C that causes both to occur simultaneously:



  For example we might have $C = \text{[person smokes cigarettes]}$. Maybe vapers are disproportionately likely to be cigarette smokers, and it's this subset of vapers that drives the association between vaping and arterial disease. If this is true, then quitting vaping woudn't decrease your chance of arterial disease.

- **Confounding:** A third variable jointly influences or causes both A and B. For example, maybe vapers have lower rates of coronary artery disease than the general population – that doesn't mean vaping necessarily have a protective effect. For example, maybe the vapers are on average younger than the population, and CAD is something that typically hapens to old people. In that case, we might observe a spurious relationship between A and B due to the effects of some other third variable, $C = \text{[youth]}$:



What we can say, however, is that if A and B do not have any relationship, then they are probabilistically independent. In what follows, we will describe how the hypothesis testing framework can be used to test whether two things A and B are independent.

This is what tests of association do. They ask the question **"Are A and B independent?"** The answer to such tests is always either "no" or "maybe". They don't give us more information than that.

We need some theory to describe how to do this.

32.2 Ingredient #1: The Multinomial Distribution

Recall that to construct a *binomial* random variable, you flip coins n times, and count the number of heads. The coin flip only has two outimes. What if, instead of flipping coin, we rolled a dice with k sides, and counted the number of times we get $1, 2, \ldots, k$? In that case, we get an example of what is called *multinomial* random variable. To be precise:

Definition 92 (Multinomial Distribution)**.** Consider an experiment with k possible outcomes, labeled 1 through k.. These outcomes have probabilities $p_1, \ldots, p_k$, respectively. (So $p_1 + \cdots + p_k = 1$). We repeat this experiment n times, and define

$$X_i = \#\text{ of times outcome } i \text{ occurs in the } n \text{ trials}$$

for each $i = 1, \ldots, k$. Then the random vector

$$X = (X_1, \ldots, X_k)$$

is called a *multinomial random variable* with parameters $(p_1, \ldots, p_k)$ and number of trials n.

This is an example of a categorical random variable (k categories).
Let's do an example with $k = 3$.

Example 93 (One armed bandit)**.** A single play of a slot machine has three outcomes:

- outcome 1: win nothing
- outcome 2: win back your bet
- outcome 3: win big

There is a sticker on the machine which says that these outcomes have probabiliites $(p_1, p_2, p_3) = (.5, .3, .2)$. A gambler decides to play the game $n = 100$ times. For each $i = 1, 2, 3$ let N_i be a random variable indicating the number of times that outcome i occurred.

If the sticker is accurate, then the random vector $N = (N_1, N_2, N_3)$ is a multinomial random variable with parameters $(p_1, p_2, p_3) = (.5, .3, .2)$ and $n = 100$. Clearly

$$\mathbb{E}[N_i] = np_i$$

So

$$\mathbb{E}[N_1] = 50 \quad \text{and} \quad \mathbb{E}[N_2] = 30 \quad \text{and} \quad \mathbb{E}[N_3] = 20$$

After playing the game, suppose gambler gets $N = (43, 35, 22)$. It is convenient to arrange this in a table

Category	1	2	3
Observed	43	35	22
Expected	50	30	20

We expect the observed values to be pretty close to the expected values. But if they **differ substantially**, then we conclude that maybe the sticker isn't accurate. In the hypothesis testing framework, sticker represents the null hypothesis

$$H_0 : (p_1, p_2, p_3) = (.5, .3, .2).$$

Question: How do we quantify "differ substantially"?

Answer: One sensible way is to consider the following quantity:

$$\begin{aligned}\chi^2 &:= \frac{(43-50)^2}{50} + \frac{(35-30)^2}{30} + \frac{(22-20)^2}{20} \\ &= \frac{49}{50} + \frac{25}{30} + \frac{4}{20} \\ &\approx 2.01\end{aligned}$$

each term in this sum takes the form

$$\frac{(\text{observed} - \text{expected})^2}{\text{expected}}.$$

The sum in **??** is always nonnegative. If the observed values are close to the expected values, then it is small. Otherwise it is large. So we will look at the size of χ^2, and make the following determination

- if χ^2 is large: we reject the null hypothesis H_0
- if χ^2 is small: we "fail to reject" the null hypothesis

End of Example 93. □

33 2025-04-23 | Week 14 | Lecture 33

33.1 Ingredient #2: The Chi-Squared Distribution

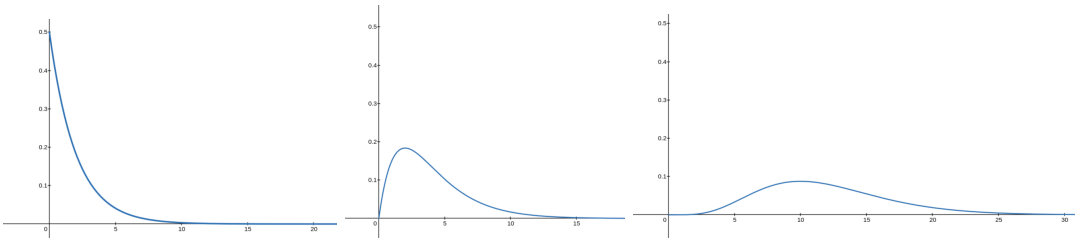
A special kind of gamma distribution:

Definition 94 (Chi-Squared Distribution)**.** Let k be a positive integer. A nonnegative random variable X is said to have *chi-squared distribution with parameter d* if it has pdf

$$f(x) = \begin{cases} \frac{1}{2^{d/2}\Gamma(d/2)} x^{\frac{d}{2}-1} e^{-x/2} & : \quad x \geq 0 \\ 0 & : \quad x < 0 \end{cases}$$

In other words, if X is a Gamma distributed random variable with parameters $\alpha = d/2$ and $\beta = 2$. The parameter d is usually called the *degrees of freedom* of X.

From left-to-right, here's what this looks like for $d = 2$, $d = 4$, and $d = 12$:



Using properties of gamma distribution, we have:

Proposition 95. *A χ^2 random variable X with d degrees of freedom has*

$$\mathbb{E}[X] = \alpha\beta = d \quad \textit{and} \quad \mathrm{Var}(X) = \alpha\beta^2 = 2d$$

The chi-squared distribution plays a central role in statistical inference. It is completely determined by its degrees of freedom.

Theorem 96. *Suppose $X = (X_1, \ldots, X_k)$ is a multinomial random variable with parameters $p_1, \ldots, p_k$. Define a new random variable χ^2, called the chi-squared statistic by*

$$\chi^2 := \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i}.$$

Then χ^2 has approximately chi-squared distribution with $k-1$ degrees of freedom (as long as $np_i \geq 5$ for all $i = 1, \ldots, m$).

33.2 The goodness-of-fit test

With ingredients #1 and #2, we can formulate what is called a "goodness of fit test". This isn't quite what we set out to do (i.e., we want to do *tests of association*), but it's a good first step. Goodness-of-fit tests are described in detail in chatper 14.1 of the textbook.

In Example 93 (one armed bandit), we considered a game with three outcomes, 1,2, and 3 with unknown probabilities p_1, p_2, and p_3. Our null hypothesis was that

$$H_0 : (p_1, p_2, p_3) = (0.5, 0.3, 0.2)$$

and the alternative hypothesis is

$$H_1 : (p_1, p_2, p_3) \neq (0.5, 0.3, 0.2).$$

We played the game 100 times. Our data is summarized in the following table:

Category	1	2	3
Observed	43	35	22
Expected	50	30	20

In this table, the expected values were calculated assuming the null hypothesis that was true. The chi-squred statistic is

$$\begin{aligned}\chi^2 &:= \frac{(43-50)^2}{50} + \frac{(35-30)^2}{30} + \frac{(22-20)^2}{20} \\ &= \frac{49}{50} + \frac{25}{30} + \frac{4}{20} \\ &\approx 2.01\end{aligned}$$

Recall that we reject the null hypothesis if χ^2 is big, and we fail to reject the null hypothesis if χ^2 is small.

Question: So is 2.01 big enough to reject H_0?

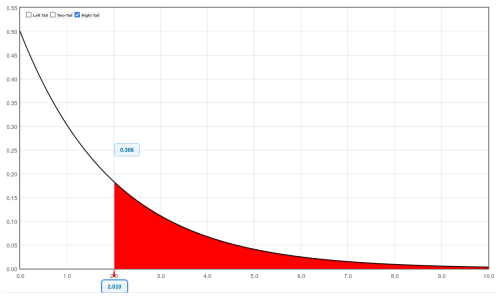
Answer: Let's compute a p-value, which if you recall is the probability of observing results at least as extreme as you observed, assuming the null hypothesis is true.

By Theorem 96, we know that the statistic has a chi-squared distribution with 2 degrees of freedom.

Therefore our p-value is

$$\begin{aligned}p\text{-value} &= \mathbb{P}\left[\chi^2 \geq 2.01\right] \\ &= \int_2^\infty \frac{1}{2} e^{-x/2} && \text{by Definition 94} \\ &= 0.366\end{aligned}$$

In other words, if the null hypothesis were true, we would expect to see our chi-squared statistic be 2.01 or greater about 37% of the time. That's pretty common, so we **fail to reject** the null hypothesis.



The test that we just did is called a goodness of fit test. These are described in chapter 14.1 in the textbook.

34 2025-04-25 | Week 14 | Lecture 34

The example in this lecture is based on a very nice chi-squared tutorial by Caitlin Light, available here: `https://www.ling.upenn.edu/~clight/chisquared.htm`

34.1 Tests of Association, Revisted

We now have the things we need to test whether two things are independent or not.

Definition 97. A *contingency table* is a matrix of with rows labeled as group 1, group 2, and so forth, and columns labeled as outcome 1, outcome 2, etc. The entries of the matrix are COUNTS: the (i, j)-th entry is the number of times that group i experience outcome j.

For example, puppose we have a population of $n = 54$ students, which we divide up according two to factors

(i) whether they attended class regularly or not

(ii) whether they pass the class exam or not

Then our contingency table is

	Pass	**Fail**
Attended	25	6
Skipped	8	15

(35)

This is the *table of observed values*. We usually augment the table with "marginal totals", like so:

	Pass	**Fail**	**Total**
Attended	25	6	31
Skipped	8	15	23
Total	33	21	54

The null hypothesis is

$$H_0 = \text{ attendence and passing are independent}$$

and the alternative hypothesis is

$$H_1 = \text{ there is a relationship between attendence and passing.}$$

Let p denote the probability that a student regularly attends class, and let q denote the probability that a student passes the final exam.
Important: We do not know the true value of p and q. But we can estimate them using the marginal row and column sums from the table:

$$p \approx \frac{31}{54} \quad \text{and} \quad q \approx \frac{33}{54}$$

Question: What are the expected values of the table, assuming the null hypothesis is true?

Under the null hypothesis, we can multiply probabilities, e.g., so that the probability that a student both regularly attends class AND passes the exam is about pq. Therefore, since there are n students,

$$\begin{aligned}
&\mathbb{E}\left[\#\text{ of students who attend regularly AND pass the exam}\right] \\
&= n\mathbb{P}\left[\text{a randomly-selected student attends regularly AND pass the exam}\right] \\
&= n\mathbb{P}\left[\text{attends regularly}\right]\mathbb{P}\left[\text{passes the exam}\right] \\
&= npq
\end{aligned}$$

Similar calculations gives us the following table of expected counts:

	Pass	**Fail**
Attended	npq	$np(1-q)$
Skipped	$n(1-p)q$	$n(1-p)(1-q)$

and plugging in the values $n = 56$ and our approximations $p \approx \frac{31}{54}$ and $q \approx \frac{33}{54}$ gives

	Pass	**Fail**
Attended	18.9	12.1
Skipped	14.1	8.9

(36)

This is the *table of estimated expected values.* Using the values of Equations (35) and (36), we compute a chi-squared statistic:

$$\begin{aligned}\chi^2 &= \sum_{\text{all table entries}} \frac{(\text{observed value} - \text{expected value})^2}{\text{expected value}} \\ &= \frac{(25-18.9)^2}{20.2} + \frac{(6-12.1)^2}{12.1} + \frac{(8-14.1)^2}{14.1} + \frac{(15-8.9)^2}{8.9} \\ &= 11.7\end{aligned}$$

By Theorem 96, we know that χ^2 has a chi-squared distribution. The degrees of freedom thing isn't so clear from that theorem, but when working with a contingency table, the degrees of freedom is given by the equation

$$(\text{number of rows} - 1) \times (\text{number of columns} - 1)$$

which in our case, for a 2×2 table, is just 1.

Thereore, using Definition 94, the p-value is

$$\begin{aligned}p\text{-value} &= \int_{11.7}^{\infty} \frac{1}{\sqrt{2\pi}} x^{-\frac{1}{2}} e^{-x/2} dx \\ &= 0.0006\end{aligned}$$

In other words, if attending and passing were independent, we would see results this extreme only 0.06% of the time. This is very small, so we reject the null hypothesis. Our conclusion is that attendance and passing the final exam are **not independent**, i.e., that there is a relationship between them.

At this point we are tempted to conclude that regularly attending class increases your chance of passing the final exam. This seems obvious and is certainly consistent with the data. But it's not a conclusion that follows from the test that we did. The test only told us that these two things things probably have some relationship; it doesn't give us information about the nature of that relationship.

35 2025-04-30 | Week 15 | Lecture 35

Topic: Finding a "best fit line" using ordinary least squares linear regression. Handwritten notes.

36 2025-05-02 | Week 15 | Lecture 36

Topic: the simple linear regression model with mean-zero gaussian errors. Handwritten notes.

37 2025-05-05 | Week 16 | Lecture 37

37.1 Odds

Suppose A is an event with probability p. The *odds* or *odds ratio* is the fraction

$$\frac{p}{1-p}.$$

The idea of *odds* (which can be any positive number) is older than that of probability (which can only be betwen 0 and 1). When outcomes are equally likely, the *probability* p is understood to be

$$\frac{\text{number of favorable outcomes}}{\text{total number of possible outcomes}},$$

whereas *odds* is

$$\frac{\text{number of favorable outcomes}}{\text{number of unfavorable outcomes}}$$

For example, the probability of rolling a 6 on a dice is 1/6, and the odds is 1/5. In some sense knowing the odds gives you the same amount of information as knowing the probability (since if you know one, you can compute the other), but the interpretation is different. For example, odds of 1/5 tell us that failure is 5 times more likely than success. We would say the odds are "5-to-1 against", or "1-to-5 against".

Odds are commonly used in betting situations. Odds is useful in betting situations because it gives a measure of the player's advantage. The first mathematical text on probability was Girolamo Cardano's 15-page text "The Book on Games of Chance" written in 1564 (though first published only a century later). Much of the work is in terms of "odds", and he appears to be the first to formulate probability in terms of the ratio of favorable outcomes to total outcomes.

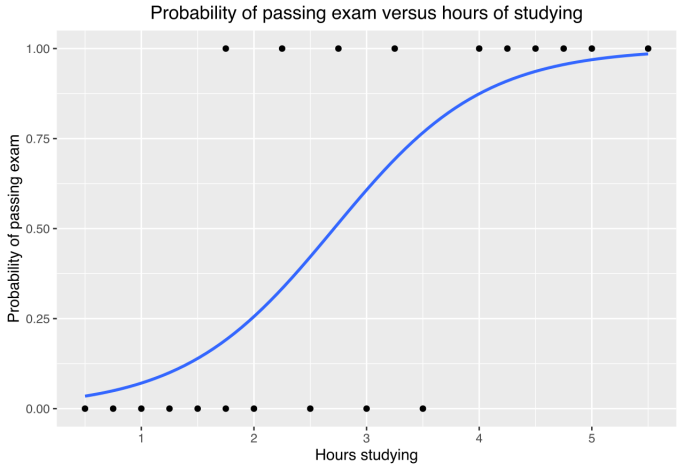
37.2 Logistic Regression

Suppose we have a numerical predictor variable x, which can take a range of values, and we are interested in a binary response variable Y, which can take only values 0 and 1. That is Y provides a binary classification. For example,

$$x = \text{ mileage of car} \quad \text{and} \quad Y = \begin{cases} 1 & : \text{ if the car needs maintenance} \\ 0 & : \text{ if the car doesn't need maintenance} \end{cases}$$

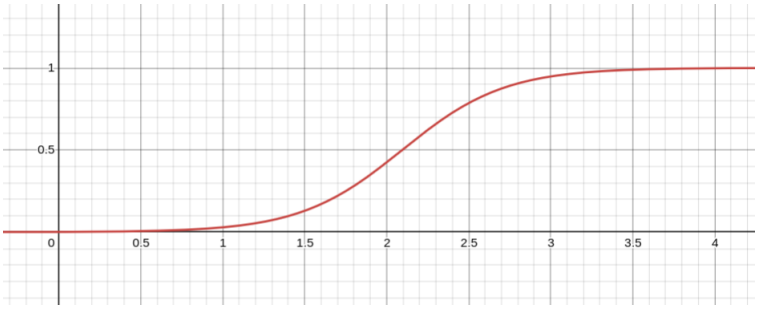
For the wiki problem `https://en.wikipedia.org/wiki/Logistic_regression#Example`, we have data

Hours (x_k)	0.50	0.75	1.00	1.25	1.50	1.75	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50	4.00	4.25	4.50	4.75	5.00	5.50
Pass (y_k)	0	0	0	0	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1



For this setting, it doesn't make sense to try to fit a best line, since there are only 2 possible y-values: 0 and 1. instead, we consider the probability $p(x) = \mathbb{P}[Y = 1]$, which increases from 0 to 1 as the mileage x increases.

The *logistic function* is the function $f(x) = \frac{e^{C+Dx}}{1+e^{C+Dx}}$, where $C, D \in \mathbb{R}$. An example when $D > 0$ is the following increasing function



In *logistic regression*, we make the **assumption** that the logarithm of the odds is approximately linear, i.e, that

$$\log\left(\frac{p(x)}{1-p(x)}\right) \approx C + Dx \tag{37}$$

for some choice of C and D. This is equivalent to assuming that the probability of, say, engine failure is approximated by the logistic function, so we assume that

$$p(x) = \frac{e^{C+Dx}}{1+e^{C+Dx}} \tag{38}$$

for some choice of real numbers C and D.

When we found a "best fit line" for some data, recall that we looked for values of C and D which minimized the magnitude of the error vector

$$e = (e_1, e_2, \ldots, e_n)$$

where $e_i = y_i - (C + Dx_i)$ is the vertical distance between our observed data and what is predicted by the line $y = C + Dx$.

For logistic regression, we do something similar, but instead of minimizing the sum of squares $\|e\|^2 = e_1^2 + e_2^2 + \ldots + e_n^2$, we instead seek to minimize something much more complicated: *the surprisal*.

Consider the single data point (x_i, y_i). Here y_i is either 0 or 1, but $x_i \in \mathbb{R}$.

According to the model given in Eq. (38), for any choice of model parametesr C and D, the predicted probability that $y_i = 1$ is

$$p(x_i) = \frac{e^{C+Dx_i}}{1+e^{C+Dx_i}}$$

Let's call this number p_i and observe that $0 < p_i < 1$. The *surprisal* (or "*log loss*") at the data point (x_i, y_i) is the quantity

$$\ell_i := -\log\left(p_i^{y_i}(1-p_i)^{1-y_i}\right) = \begin{cases} \log(\frac{1}{p_i}) & : \quad y_i = 1 \\ \log\left(\frac{1}{1-p_i}\right) & : \quad y_i = 0 \end{cases}$$

Observations

- $\ell_i > 0$
- ℓ_i really does measure "surprise" in some sense. For example, if $p_i \approx 0$ and $y_i = 0$, then $\ell_i \approx \log(1) = 0$. But if $p_i \approx 1$ and $y_i = 0$, then $\ell_i \approx \log(\text{big number}) = \text{big}$. Bigger values means there is more discrepancy between what is predicted by the model and what is observed.

The *total loss* is the sum of all the surprisals:

$$T(C, D) = \ell_1 + \ell_2 + \ldots + \ell_n$$

Observations

- T is a complicated function of the parameters C and D.
- Each choice of C, D correspond to some model from the family Eq. (38)
- When $T(C, D)$ is large, the observed data $(x_1, y_1), \ldots, (x_n, y_n)$ would be unlikely to occur for that choice of C and D. But if $T(C, D)$ is close to zero, then the model would be more likely to produce data like what was observed.
- In this way the quantity $T(C, D)$ may be regarded as a "distance" between your observed data and the model with parameters C, D. The goal is to find the values of C, D which minimize this distance—these will give you the "best" model for your data.

In logistic regression, the goal is to try to find the values of C, D which minimize the $T(C, D)$. This cannot reasonably be done by hand, as the function T is complicated enough that we don't have closed form solutions. Instead, more typically, one will use software to check many values of C and D and use whichever appears to give you the smallest $T(C, D)$. Actual implementations use sophisticated heuristics to do this.

For the wiki problem, the best C and D we get are $C \approx -4.1$ and $D \approx 1.5$, this gives the "best model" to be

$$p_{\text{best}}(x) = \frac{e^{-4.1+1.5x}}{1 + e^{-4.1+1.5x}}$$

and which gives the curve graphed above.

This model can then be used to make predictions.For example, if you study 3 hours, your probability of passing is predicted to be

$$p_{\text{best}}(3) = \frac{e^{-4.1+1.5(3)}}{1 + e^{-4.1+1.5(3)}} \approx 0.6.$$

The example we have worked has assumed that there is only one explanatory variable x. This restriction isn't essential to the theory. Suppose instead that we have 2 explanatory variables, x and w. Now, instead of Eq. (37) we have

$$\log\left(\frac{p(x)}{1 - p(x)}\right) \approx C + Dx + Ew$$

We have added a new parameter E. So now we must minimize the total loss over all C, D, and E, rather than just over all C and D. Nothing else changes. You can add as many explanatory variables as you like.

If you have a dataset with lots of possible explanatory variables, how do you know which ones to include and which to exclude from a logistic regression analysis? For example: Y indicates whether a surgery patient experiences complications due to surgery (assume same type of surgery). Possible explanatory variables

- patient age, sex
- surgical device type
- year that surgery was performed
- attending physician
- preexisting conditions
- innumerable other risk factors

If you include all of them, you get results that may be overfitted and may be hard to interpret. One option is to repeatedly do many analyes with different subsets of your explanatory variables, with some penalty applied so that models with fewer parameters are preferred (see, e.g. the Bayesian Information Criterion (BIC) score). These are related to goodness of fit tests like the ones we did for the skittle distribution.