Lecture Notes for Math 372: Elementary Probability and Statistics

Last updated: May 6, 2026

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

0 Tentative course outline **6**

1 2025-01-12 | Week 01 | Lecture 01 **7**

2 2025-01-14 | Week 01 | Lecture 02 **8**
2.1 A general framework for probability . . . 8
2.2 The enumeration principle . . . 9
2.3 Events . . . 10

3 2025-01-16 | Week 01 | Lecture 03 **11**
3.1 Powerball . . . 11
3.2 Recap of terminology . . . 11
3.3 Understanding the "events" as sets . . . 12

4 2026-01-21 | Week 02 | Lecture 04 **14**
4.1 Repeated coin flipping . . . 14
4.2 The three probability axioms . . . 15
4.3 Inclusion-exclusion principle . . . 16

5 2026-01-23 | Week 02 | Lecture 05 **17**
5.1 St. Petersburg Paradox . . . 17
5.2 What is probability, continued . . . 17

6 2026-01-26 | Week 03 | Lecture 06 **19**
6.1 Multiplication principle . . . 19
6.2 Permutations and combinations . . . 21

7 2026-01-28 | Week 03 | Lecture 07 **22**
7.1 Review of multiplication principle . . . 22
7.2 The permutation principle and the combination principle . . . 22
7.3 Examples of the three combinatorial principles . . . 23

8 2025-01-30 | Week 03 | Lecture 08 **25**
8.1 Applications of the 3 combinatorial principles . . . 25
8.2 Introduction to craps . . . 26
8.3 Conditional probability . . . 26

9 2026-02-02 | Week 04 | Lecture 09 **27**
9.1 Conditional probability . . . 27

10 2025-02-04 | Week 04 | Lecture 10 **29**
10.1 Conditional Probability: Examples . . . 29
10.2 The law of total probability . . . 30

11 2026-02-06 | Week 04 | Lecture 11 **32**
11.1 Application of the law of total probability: diagnostic tests . . . 32
11.2 Independence . . . 33

12 2026-02-11 | Week 05 | Lecture 12 **34**
12.1 Examples of using independence . . . 34
12.2 Dicrete Random Variables . . . 34

13 2026-02-13 | Week 05 | Lecture 13 **37**
13.1 Exam review problems . . . 37

14 2026-02-23 | Week 07 | Lecture 14 **38**
14.1 Expectation . . . 38
14.2 Variance . . . 39

15 2026-02-25 | Week 07 | Lecture 15 **41**
15.1 Properties of Variance . . . 41
15.2 Binomial random variable . . . 42

16 2026-02-27 | Week 07 | Lecture 16 **44**
16.1 Application of the Law of Total Probability: The probability of winning at craps . . . 44
16.2 Binomial random variable (continued) . . . 46

17 2026-03-02 | Week 08 | Lecture 17 **47**
17.1 Poisson random variable . . . 47

18 2026-03-04 | Week 08 | Lecture 18 **50**
18.1 Applications of the Poisson random variable . . . 50
18.2 Continuous random variables . . . 51

19 2026-03-06 | Week 08 | Lecture 19 **53**
19.1 Continuous random variables . . . 53

20 2026-03-09 | Week 09 | Lecture 20 **55**
20.1 Continuous random variables . . . 55

21 2026-03-11 | Week 09 | Lecture 21 **56**
21.1 Normal random variables . . . 56

22 2026-03-23 | Week 10 | Lecture 22 **58**
22.1 Joint distributions - discrete case . . . 58
22.2 Joint distributions - continuous case . . . 59

23 2026-03-25 | Week 10 | Lecture 23 **61**
23.1 An example of using a joint distribution to compute probabilities . . . 61
23.2 LOTUS Theorem - multivariate case . . . 61

24 2025-03-27 | Week 10 | Lecture 24 **64**
24.1 Statistics: a general framework . . . 64
24.2 What is a statistic? . . . 64

25 2026-04-06 | Week 12 | Lecture 25 **66**
25.1 The central limit theorem . . . 66
25.2 Point estimation . . . 68

26 2026-04-08 | Week 12 | Lecture 26 **69**
26.1 Point estimation . . . 69

27 2026-04-13 | Week 13 | Lecture 27 **72**
27.1 Confidence intervals . . . 72

28 2026-04-15 | Week 13 | Lecture 28 **73**
28.1 The Coupon Collector Problem . . . 73
28.2 Hypothesis Testing . . . 74

29 2026-04-17 | Week 13 | Lecture 29 **75**
29.0.1 What is a hypothesis test? . . . 75
29.0.2 Type I and Type II errors . . . 76

30 2026-04-20 | Week 14 | Lecture 30 **77**
30.1 Decision errors . . . 77
30.1.1 Type I and Type II errors . . . 77

31 2026-04-22 | Week 14 | Lecture 31 **79**
31.1 Tests of independence . . . 79
31.1.1 Ingedient 1: The chi-squared distribution . . . 79
31.1.2 Ingedient 2: The chi-squared statistic . . . 79
31.1.3 Example of a test of independence . . . 80

32 2025-04-21 | Week 14 | Lecture 32 **81**
32.1 Interpreting tests of independence . . . 82

33 2025-05-01 | Week 15 | Lecture 33 **84**
33.1 Simple linear regression (i.e., finding a best-fit line) . . . 84

34 2026-05-04 | Week 16 | Lecture 34 **88**

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)
- **Midterm 1 (Feb 18)**
- **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.
- **Week 9:** Introduction to statistical estimation
  - Point and confidence interval estimation. Maximum likelihood, optimal, and unbiased estimators. Examples. (Sections 6.1, 6.2)
- **Midterm 2 (March 27)**
- **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)
- **Last day of instruction: May 6**
- **Final Exam: May 15**

1 2025-01-12 | Week 01 | Lecture 01

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

2 2025-01-14 | Week 01 | Lecture 02

The nexus question of this lecture: *What is the general framework for probability?*
Reading assignment: *chapters 2.1 - 2.5*

2.1 A general framework for probability

We begin with a general framework and some terminology to formalize the notions of probability.

- An ***experiment*** is an activity or process whose outcome is subject to uncertainty, and about which an observation is made.

  Examples include flipping a coin, rolling a dice, measuring the size of a wave, or the amount of rainfall, conducting a poll, performing a diagnostic test, opening a pack of Pokemon cards, etc.

- The ***sample space*** S of an experiment is the set of all possible outcomes. The elements of the sample space are called ***sample points***.

  We think of each sample point as representing a unique outcome of the experiment. In the case of rolling a dice, the sample points are $1, 2, 3, 4, 5$ and 6, and the sample space is $S = \{1, 2, 3, 4, 5, 6\}$.

- We use the term ***event*** to refer to a collection of outcomes. 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: if our experiment is rolling a 6-sided dice, here are some events

  $$\begin{aligned} A &= [\text{observe an odd number}] & E_2 &= [\text{observe a 2}] \\ B &= [\text{observe an even number}] & E_3 &= [\text{observe a 3}] \\ C &= [\text{observe a number less than 5}] & E_4 &= [\text{observe a 4}] \\ D &= [\text{observe a 2 or a 3}] & E_5 &= [\text{observe a 5}] \\ E_1 &= [\text{observe a 1}] & E_6 &= [\text{observe a 6}] \end{aligned}$$

- There are two types of events: ***compound events***, which can be decomposed into other events, and ***simple events***, which cannot.

  In the above example, the events A, B, C and D are compound events. $E_1, \ldots, E_6$ are simple events.

  $$\begin{aligned} A &= [\text{observe an odd number}] = \{1, 3, 5\} & E_2 &= [\text{observe a 2}] = \{2\} \\ B &= [\text{observe an even number}] = \{2, 4, 6\} & E_3 &= [\text{observe a 3}] = \{3\} \\ C &= [\text{observe a number less than 5}] = \{1, 2, 3, 4\} & E_4 &= [\text{observe a 4}] = \{4\} \\ D &= [\text{observe a 2 or a 3}] = \{2, 3\} & E_5 &= [\text{observe a 5}] = \{5\} \\ E_1 &= [\text{observe a 1}] = \{1\} & E_6 &= [\text{observe a 6}] = \{6\} \end{aligned}$$

Example 1 (Examples of sample spaces)**.**

- 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 rainfall in a day is $S = \{x \in \mathbb{R} : x \geq 0\}$

  An example of an event for this sample space is:

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

- 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 1. □

2.2 The enumeration principle

Example 2 (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 2. □

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

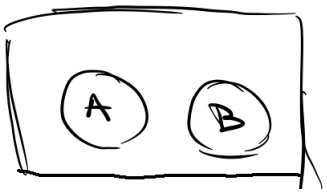
> If your sample space is finite and consists of *equally likely outcomes*, then you can compute lots of probabilities easily by listing outcomes and counting them. To be prcecise, for any event A,
>
> $$\mathbb{P}[A] = \frac{|A|}{|S|} = \frac{\#\text{ of outcomes in } A}{\#\text{ of outcomes in } S}$$

I already had you do a bunch of these problems on Monday.

2.3 Events

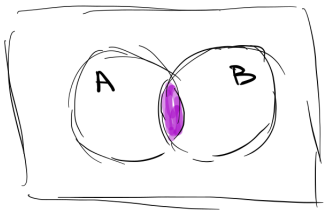
Some observations about events:

- Two events E and F are ***mutually exclusive*** if $E \cap F = \emptyset$. This means that E and F cannot both happen at the same time.



  In the dice example, the events A and B are mutually exclusive, since the dice roll cannot be both even and odd. But A and C are not mutually exclusive because $A \cap C = \{1, 3\} \neq \emptyset$. If a 1 or a 3 is rolled, then both A and C occur.

- The sample points are *elements* of S. The simple events are *singleton subsets* of S. In the dice example, we have:
  - Sample points: 1,2,3,4,5,6.
  - Simple events: $\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}$.
- The empty set $\emptyset$ and the whole sample space S are always both events: $\emptyset$ is the event "nothing happens" and S is the event "something happens".
- If E and F are events, then $E \cap F$ is the event that both E **and** F occur:



  Here, $E \cap F$ is the ***intersection*** of E and F; that is, $E \cap F$ the set of sample points contained in both E and F.

3 2025-01-16 | Week 01 | Lecture 03

> ***The nexus question of this lecture:*** *What is the general framework for probability (continued)?*

- Go over problem 4, parts (b) and (c), and problem 6 on the worksheet.

3.1 Powerball

Example 3 (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 3. □

3.2 Recap of terminology

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

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

Example 4 (Sum 2d4)**.** Roll a 4-sided dice twice (this is the **experiment**). There are 16 possible **outcomes**. The **sample space** is

$$S = \{(x, y) : x, y \in \{1, 2, 3, 4\}\},$$

which consists of 16 ordered pairs (x, y), the **sample points**. An **event** is any subset of these.

We'll give two examples of events.

Let X be the sum of the two rolls. We can represent X by the following table:

		Dice 2			
		1	2	3	4
Dice 1	1	2	3	4	5
	2	3	4	5	6
	3	4	5	6	7
	4	5	6	7	8

The event that $X = 6$ is:

$$[X = 6] = \{(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)\}.$$

Here's another example of an event:

$$[X \text{ is even}] = \{(1,1),(1,3),(2,2),(2,4),(3,1),(3,3),(4,2),(4,4)\}.$$

The **enumeration principle** says that, when outcomes are equally likely, we can compute the probability of an event E by

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

So

$$\mathbb{P}[X = 6] = \frac{3}{16}$$

and

$$\mathbb{P}[X \text{ is even}] = \frac{8}{16} = \frac{1}{2}.$$

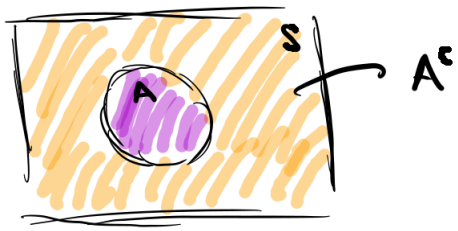
End of Example 4. □

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

3.3 Understanding the "events" as sets

Let A and B be events. That is, A and B are subsets of S.

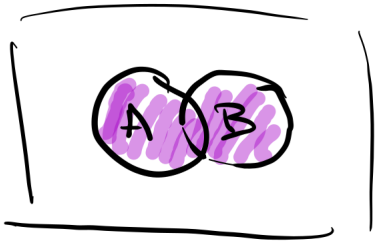
- Suppose A is an event. It's a subset of S, like this:



- The ***complement*** of A, denoted A^c is the set of points in S that are **not** in A.

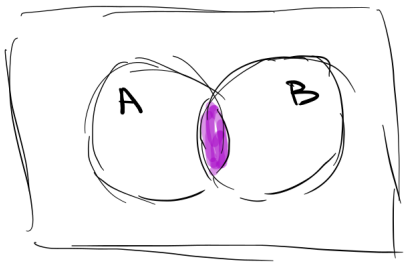
  We interpet A^c as the event that A didn't happen.

- The ***union*** of A and B, denoted $A \cup B$, to denote the set of all outcomes that are in A or B this includes outcomes that are in both):



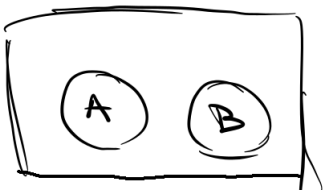
  We interpret $A \cup B$ as the event that either A or B (or both) occur.

- The ***intersection*** of A and B, denoted $A \cap B$ is the set of points that are contained in both A and B:



We interpret $A \cap B$ as the event that A and B both occur.

- 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 ***disjoint***. Disjoint events are ***mutually exclusive***, in the sense that they cannot happen simultaneously.

For example, when rolling a dice, the following two events are mutually exclusive:

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

- An 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, the experiment from last lecture involving repeatedly drawing a ball from an urn with replacement until we get the gold ball. Let X be the number of balls we had to draw until we got the gold ball. 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.

4 2026-01-21 | Week 02 | Lecture 04

4.1 Repeated coin flipping

Example 5 (Repeated coin flipping)**.** Consider an experiment in which we flip a coin repeatedly until we get a heads. The **output** of this experiment is *the number of times we flipped the coin.*

The sample space of the experiment is

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

Let X = (number of coin flips). Here, X is a random quantity which will vary between experiments. 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

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

Question 1: How many times do we expect to flip the coin? In other words, 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

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

and applying the geometric series formula again, we see that

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

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

End of Example 5. □

4.2 The three probability axioms

Sections 2.1-2.2 in the textbook.

Definition 6 (Probability Measure)**.** 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) If S is the whole sample space, then $\mathbb{P}[S] = 1$.

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

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

Notice that Definition 9 doesn't say how to assign probabilities. It merely tells us the conditions that probabilities must satisfy. The first two, **A.1** and **A.2** seem almost self-evident. The last one, **A.3**, is more technical and harder to deciper. The next example gives involves an application of axiom **A.3**.

Example 7 (Continuation of Example 5.)**.** Assume the same experiment and notation as Example 5.

Question 2: What is the probability that we flip the coin an even number of times?
For each $n = 1, 2, \ldots$, define the event E_n as

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

We want to know the probability that X is even. Observe that

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

We note that the sequence $E_2, E_4, E_6 \ldots$ is pairwise disjoint. So we can use **A.3** to compute the probability of this event:

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

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

End of Example 7. □

4.3 Inclusion-exclusion principle

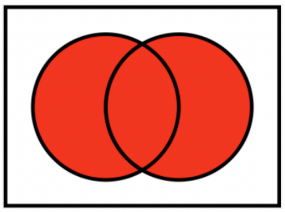
Theorem 8 (Inclusion-Exclusion Prinicple)**.** *If A and B are events, then*

$$\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B]$$

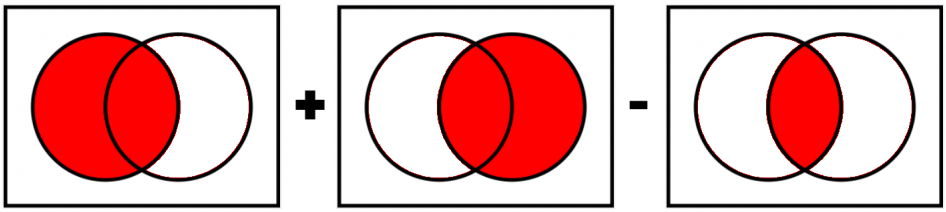
In particular, if A and B are distjoint, then

$$\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B].$$

Idea of proof. $\mathbb{P}[A \cup B]$ is the area of the red-shaded region:



Compare with $\mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B]$:



□

5 2026-01-23 | Week 02 | Lecture 05

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?

$$\begin{aligned}\mathbb{E}[X] &= \sum_x x\mathbb{P}[X = x] \\ &= 2\mathbb{P}[X = 2] + 4\mathbb{P}[X = 4] + 8\mathbb{P}[X = 8] + 16\mathbb{P}[X = 16] + \ldots\end{aligned}$$

Next, we observe that $\mathbb{P}\left[X = 2^k\right] = \frac{1}{2^k}$ for all $k = 1, 2, \ldots$. Plugging these probabilities in, we get

$$\mathbb{E}[X] = 1 + 1 + 1 + 1 + 1 + \ldots = +\infty.$$

The expected value of playing this game is positive infinity!!!!! N matter how much you pay to play this game, you will eventually make money if you play enough times.

5.2 What is probability, continued

Definition 9 (Probability Measure)**.** 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) If S is the whole sample space, then $\mathbb{P}[S] = 1$.

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

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

Proposition 10 (Basic properties of probability measure)**.**

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

(ii.) *(Finite additivity) If $E_1, \ldots, E_n$ are 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, $\mathbb{P}[E] \in [0,1]$*

Proposition 11 (De Morgan's Laws)**.** *The following equalities hold:*

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

Proof. Draw a picture. □

6 2026-01-26 | Week 03 | Lecture 06

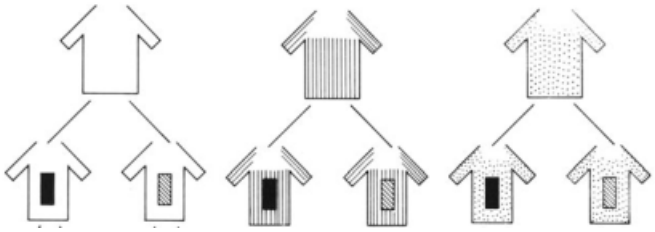
This lecture is based on section 2.3 in the textbook.

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

- multiplication principle
- permutation principle
- combination principle

6.1 Multiplication principle

Example 12. A man has 3 shirts and two ties. How many ways can he dress himself?



We can schematize this as three shirts s_1, s_2, s_3 and two ties t_1, t_2, in which case the possible outfits are:

$$(s_1, t_1)\ (s_2, t_1)\ (s_3, t_1)$$
$$(s_1, t_2)\ (s_2, t_2)\ (s_3, t_2)$$

Of course, we don't need to write s and t; it is enough to write

$$(1, 1)\ (2, 1)\ (3, 1)$$
$$(1, 2)\ (2, 2)\ (3, 2)$$

where the first slot is "shirt" and the second is "tie". Thus the methematical way to name the collection of outfits is the set of all ordered couples (a, b) with $a = 1, 2$ and $b = 1, 2, 3$. So you can see the total number of outfits is $2 \times 3 = 6$.

In general we can talk about ordered k-tuples $(a_1, \ldots, a_k)$, where for each j from 1 to k, the symbol a_j is the assignment (choice) for the j^{th} slot and it may be denoted by a numeral between 1 and n_j. In our example, $k = 2$ and $n_1 = 3$ and $n_2 = 2$.

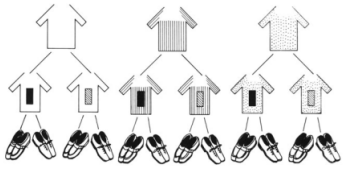
If the man also have two pairs of shoes, we can add a third slot for "shoes", so the question is asking to count the number of elements in the set

$$\{(a_1, a_2, a_3) : a_1 \in \{1, 2, 3\}, a_2, a_3 \in \{1, 2\}\}$$

In this case, $n_1 = 3$, $n_2 = 2$ and $n_3 = 2$, so the number of outfits is

$$n_1 n_2 n_3 = 3 \cdot 2 \cdot 2 = 12,$$

which is shown in the following picture:



End of Example 12. □

In this example, we've used the "multiplication principle". Let's state this principle formally now. To do so, we need the following definition:

Definition 13 (k-tuple)**.** A ***k*-tuple** is an ordered list of k objects.

For example, $(1, 2, 5, 4)$ and $(3, 3, 5, 4)$ are both examples of 4-tuples. Note that repeats are allowed.

Theorem 14 (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*

$$n_1 \times n_2 \times \cdots \times n_k$$

possible k-tuples.

Example 15 (Application of the multiplication principle to probability)**.** Consider an experiment in which we roll a dice 5 times in a row. One possible outcome is

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

This is an ordered list with five entries, i.e., a 5-tuple.

Problem 1: What is the sample space? How many elements does it have?

- there are 6 choices for the first entry
- 6 choices for the second entry

  ⋮
- 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 2: Let $E = [\text{All 5 dice are less than or equal to 3}]$. What is $\mathbb{P}[E]$?

By the enumeration principle,

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

now we use again the multiplication principle 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 15. □

6.2 Permutations and combinations

Example 16. How many ways are there to order the 4 letters A, B, C, D?

Use the multiplication principle

$$4 \times 3 \times 2 \times 1$$

End of Example 16. □

Proposition 17 (Counting permutations)**.** *The total number of ways to order n distinct objects is*

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

(Note that we define $0! = 1$.)

Definition 18 (permutation and combination)**.** 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

A ***permutation*** is an ordered subset. An unordered subset is called a ***combination***.

7 2026-01-28 | Week 03 | Lecture 07

This lecture is based on section 2.3 in the textbook.

7.1 Review of multiplication principle

Definition 19 (k-tuple)**.** A ***k*-tuple** is an ordered list of k objects.

Theorem 20 (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*

$$n_1 \times n_2 \times \cdots \times n_k$$

possible k-tuples.

7.2 The permutation principle and the combination principle

Proposition 21 (Counting permutations)**.** *The total number of ways to order n distinct objects is*

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

(Note that we define $0! = 1$.)

Definition 22. A ***set*** is an *unordered* collection of elements, all of which are *distinct.* A ***permutation*** is an ordered subset. An unordered subset is called a ***combination***.

Theorem 23 (The permutation principle)**.** *The number of permutations of size k that can be formed from n objects is*

$$(n)_k := n \cdot (n-1) \cdot (n-2) \cdots (n-k+1) = \frac{n!}{(n-k)!}.$$

There are k terms in this product. (The book uses the notation $P_{k,n}$ instead of $(n)_k$.)

Proof. Suppose our n objects are n balls in an urn, labeled $1, \ldots, n$. We draw n balls sequentially, each ball drawn is left out of the urn. The number permutations is the number of ways that k balls can be drawn out of the urn. We are dealing with ordered k-tuples

$$(a_1, \ldots, a_k)$$

where $a_1, \ldots, a_k \in \{1, \ldots, n\}$ are integers between 1 and n that must all be different. (Obviously, $k \leq n$ since there are n balls in the urn and we seek to draw out k of them).

For the first ball, we have n choices. For the second ball we have $n-1$ choice (since there remain only $n-1$ balls in the urn after our first draw). For the third ball we have $n-2$ choices, and so forth. By the k^{th} ball (our final draw), there are $n-k+1$ balls to choose from in the urn.

Therefore, by the multiplication principle,

$$\text{the number of permutations} = n \cdot (n-1) \cdot (n-2) \cdots (n-k+1) = \frac{n!}{(n-k)!}.$$

□

Theorem 24 (The combination principle)**.** *Let S be a set with n elements. The number of (unordered) subsets of size k is*

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

Proof. To understand why the formula of Theorem 24 is correct, we begin by making two observations:

- **Observation 1.** For every unordered subset of size k, there are $k!$ ways to order it by Proposition 17. Therefore the following equation holds:

$$k! \times (\text{number of unordered subsets of size } k) = (\text{number of ordered subsets of size } k)$$

- **Observation 2.** The permutation principle (Theorem 23) tells us that

$$(\text{number of ordered subsets of size } k) = \frac{n!}{(n-k)!}.$$

We now put these two observations together. Let X and Y be the following quantities:

$$X = (\text{number of ordered subsets of size } k)$$
$$Y = (\text{number of unordered subsets of size } k).$$

Then

$$\begin{aligned} k! \cdot Y &= X && \text{by Observation 1} \\ &= \frac{n!}{(n-k)!} && \text{by Observation 2} \end{aligned}$$

Dividing both sides by $k!$ gives

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

and this proves Theorem 24. □

7.3 Examples of the three combinatorial principles

Example 25. A dessert company has 32 distinct desserts on its menu: 6 types of pie, 9 types of cake, and 17 flavors of ice cream.

(a) **Question:** After purchasing 4 different desserts, a customer decides he will eat them all in a row, and he cares about the order in which he eats them. How many different ways can he eat his 4 desserts?

Solution: Here we are counting the number of ways to order set of four objects. This number is 4! by Proposition 17.

(b) **Question:** How many ways are there for the customer to choose four desserts and then eat them in a particular order of his choice?

Solution 1: By the combination principle, there are $\binom{32}{4}$ combinations of desserts that the customer could choose. Moreover, by part (a), for each combination of 4 desserts, he can order them 4! ways. Therefore by the multiplication principle, there are

$$\binom{32}{4} \times 4! = 32 \cdot 31 \cdot 30 \cdot 29$$

ways that the customer could choose and then eat four desserts.

Solution 2: Another solution would be to note that we want to count the number of permutations of size $k = 4$ from a set of $n = 32$ distinct objects. (Think: the customer is selecting 4 desserts from an urn with 32 distinct desserts in it). By the permutation principle, this number is

$$(32)_4 = 32 \cdot 31 \cdot 30 \cdot 29$$

which agrees with our answer from solution 1.

(c) **Question:** If 9 desserts are randomly selected, how many ways are there to obtain 3 deserts of each type (pie, cake, and ice cream)?

Solution: For pie, there are $\binom{6}{3}$ combinations possible. For cake, there are $\binom{9}{3}$ combinations possible, and for ice cream, there are $\binom{17}{3}$.

By the multiplication principle, there are

$$\binom{6}{3} \times \binom{9}{3} \times \binom{17}{3}$$

ways to simultaneously select 3 pies, 3 cakes, and 3 ice cream flavors.

End of Example 25. □

8 2025-01-30 | Week 03 | Lecture 08

This lecture is based on sections 2.3 and 2.4 in the textbook.

8.1 Applications of the 3 combinatorial principles

Example 26. Let $S = \{1, 2, 3, 4, 5, 6\}$. How many unordered subsets (i.e., combinations) of size 2 are there?

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

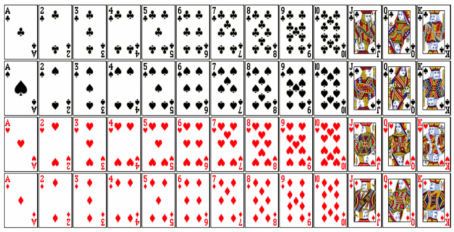
The fiften combinations are

$$\begin{aligned} \{1,2\}, \{1,3\}, \{1,4\}, \{1,5\}, \{1,6\} \\ \{2,3\}, \{2,4\}, \{2,5\}, \{2,6\} \\ \{3,4\}, \{3,5\}, \{3,6\} \\ \{4,5\}, \{4,6\} \\ \{5,6\} \end{aligned}$$

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

End of Example 26. □

Example 27 (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. A **flush** occurs when all five cards are the same suit. How many ways are there to get a flush?

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

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

3. What's the probability of getting a flush?

   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 27. □

8.2 Introduction to craps

Craps is a dice-rolling game in which a player, known as the shooter, rolls a pair of 6-sided dice. The shooter wins, loses or re-rolls depending on what happens.

Here's a set of rules for the shooter:

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.

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:

8.3 Conditional probability

This section is based on section 2.4 in the text.

Definition 28 (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

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

9 2026-02-02 | Week 04 | Lecture 09

9.1 Conditional probability

This section is based on section 2.4 in the text.

Definition 29 (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

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

Interpretation: $\mathbb{P}[A \mid B]$ is the probability of A when we know that event B happened.

Multiplication rule: $\mathbb{P}[A \cap B] = \mathbb{P}[B \mid A]\,\mathbb{P}[A]$

Definition 30. We say that there exists a ***positive relationship*** between events A and B if

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

and a ***negative relationship*** if

$$\mathbb{P}[B \mid A] < \mathbb{P}[B].$$

- Smoking and getting lung canger have a positive relationship: if you know someone is a smoker (A), then the probability that they get lung cancer (B) is higher than if you didn't know they were a smoker.

Example 31 (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 29,

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

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 31. □

Example 32 (Similar to examples 2.27 and 2.28 in the textbook)**.** A vampire goes to the blood bank looking to find some type O+ blood (the most delicious type). He finds 4 unlabeled bags of blood. Only one of the bags is O+, but the vampire doesn't know which one, so he 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

$$A = [\text{first bag isn't O+}] \quad \text{and} \quad B = [\text{second bag isn't O+}]$$

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

(b) What's the probability that he must test exactly 3 bags?

Solution: We want $\mathbb{P}[X = 3] = \mathbb{P}[\text{first bag isn't O+} \cap \text{second bag isn't O+} \cap \text{third bag is O+}]$.

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

Here we have used the following multiplication rule:

$$\begin{aligned}\mathbb{P}[A \cap B \cap C] &= \mathbb{P}[C \mid A \cap B]\,\mathbb{P}[A \cap B] \\ &= \mathbb{P}[C \mid A \cap B]\,\mathbb{P}[B \mid A]\,\mathbb{P}[A]\end{aligned}$$

with $A = [\text{first bag isn't O+}]$, $B = [\text{second bag isn't O+}]$, and $C = [\text{third bag is O+}]$.

End of Example 32. □

10 2025-02-04 | Week 04 | Lecture 10

This section is based on sections 2.4 in the textbook.

10.1 Conditional Probability: Examples

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

Definition 33 (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 34 (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 possibilities. 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 34. □

Example 35 (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 A 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

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

Now,

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

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

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

Therefore by Eq. (2),

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

End of Example 35. □

10.2 The law of total probability

Definition 36. We say that a collection of events $A_1, \ldots, A_n$ is ***exhaustive*** if at least one must occur; that is,

$$A_1 \cup A_2 \cup \ldots \cup A_n = S$$

where S is the sample space.

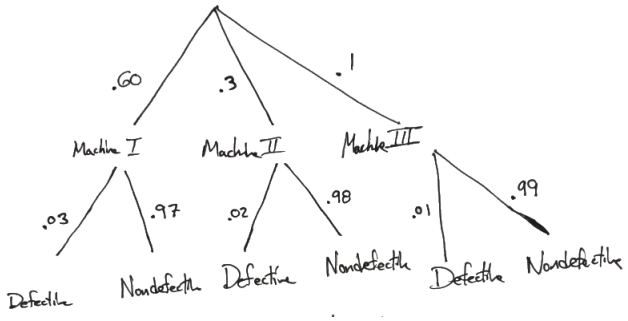
Theorem 37 (Law of Total Probability)**.** *Let $A_1, \ldots, A_n$ be events which are (1) mutually exclusive and (2) exhaustive. Then for any event B,*

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

Example 38. A munitions factory produces 155mm artillery rounds using three machines: 60% of rounds are produced by machine I, 30% by machine II, and 10% by machine III. Engineers discover that

- 3% of rounds produced by machine I are defective
- 2% of rounds produced by machine II are defective
- 1% of rounds produced by machine III are defective.

Given a defective round, what is the probability that it was produced by machine III?
Solution: Let's make a diagram to illustrate this:



Let

$$\begin{aligned} D &= [\text{shell is defective}] \\ M_1 &= [\text{shell was made by machine I}] \\ M_2 &= [\text{shell was made by machine II}] \\ M_3 &= [\text{shell was made by machine III}]. \end{aligned}$$

We want to find $\mathbb{P}[M_3 \mid D] = \mathbb{P}[\text{machine III} \mid \text{defective}]$.

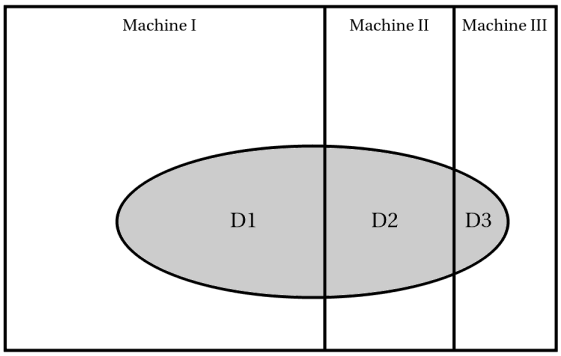
$$\mathbb{P}[M_3 \mid D] = \frac{\mathbb{P}[M_3 \cap D]}{\mathbb{P}[D]} \tag{3}$$

Now we compute the numerator and denominator separately:

By the multiplication rule,

$$\mathbb{P}[M_3 \cap D] = \mathbb{P}[D \mid M_3]\,\mathbb{P}[M_3] = (.01)(.1) = .001.$$

Next, we compute the denominator. The following figure illustrates the sample space, which consists of shells produces by machines I, II and III. The grey region represents the event that the randomly selected shell is defective.



$$\begin{aligned}
\mathbb{P}[D] &= \mathbb{P}[D1] + \mathbb{P}[D2] + \mathbb{P}[D3] \\
&= \mathbb{P}[D \cap M_1] + \mathbb{P}[D \cap M_1] + \mathbb{P}[D \cap M_1] \\
&= \mathbb{P}[D \mid M_1]\,\mathbb{P}[M_1] + \mathbb{P}[D \mid M_2]\,\mathbb{P}[M_2] + \mathbb{P}[D \mid M_3]\,\mathbb{P}[M_3] \\
&= (.03)(.6) + (.02)(.3) + (.01)(.1) \\
&= .025
\end{aligned}$$

Hence by Eq. (26),

$$\mathbb{P}[M_3 \mid D] = \frac{.001}{.025} = .04.$$

In other words, 4% of defective rounds are produced by machine III.

End of Example 38. □

Example 39 (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

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

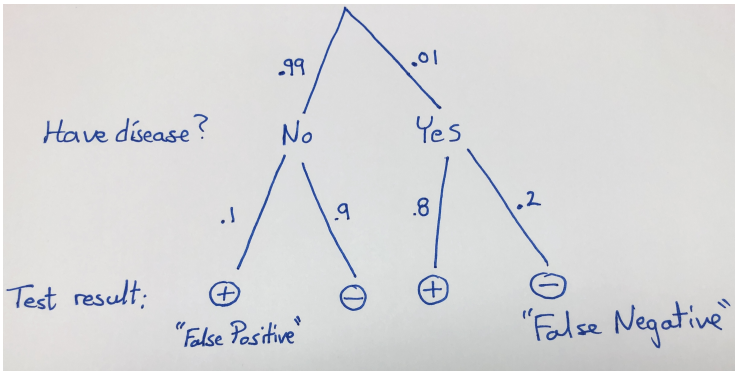
End of Example 39. □

11 2026-02-06 | Week 04 | Lecture 11

11.1 Application of the law of total probability: diagnostic tests

This based on section 2.4

Example 40 (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 $\mathbb{P}\left[\text{have disease} \mid \text{test positive}\right]$. 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,

$$\mathbb{P}\left[D \mid T^+\right] = \frac{\mathbb{P}\left[T^+ \cap D\right]}{\mathbb{P}\left[T^+\right]} \tag{4}$$

We will compute the numerator and denominator separately:

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

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

  We've now computed the numerator of Eq. (4).

- (Denominator) By the law of total probability,

  $$\begin{aligned}\mathbb{P}\left[T^+\right] &= \underbrace{\mathbb{P}\left[T^+|D\right]\mathbb{P}\left[D\right]}_{=0.008 \text{ by Eq. (16)}} + \underbrace{\mathbb{P}\left[T^+ \mid D^c\right]\mathbb{P}\left[D^c\right]}_{=(.1)(.99)=0.099} \\ &= (.8)(.01) + (.1)(.99) \\ &= 0.008 + 0.099\end{aligned}$$

  Finally, having done the necessary work, we can now plug our results into Eq. (4):

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

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 40. □

11.2 Independence

This is based on section 2.5 in the textbook.

Previously, we saw that A and B have a **positive relationship** if

$$\mathbb{P}[B \mid A] > \mathbb{P}[B]$$

and a **negative relationship** if

$$\mathbb{P}[B \mid A] < \mathbb{P}[B]$$

What about when $\mathbb{P}[B \mid A] = \mathbb{P}[B]$?

Definition 41 (Independence)**.** Two events A and B are said to be ***independent*** if $\mathbb{P}[A \cap B] = \mathbb{P}[A]\,\mathbb{P}[B]$. Otherwise, the events are said to be ***dependent***.

This definition generalizes in the natural way:

Definition 42 (Mutual independence)**.** For any $n \geq 2$, we say that the events $A_1, \ldots, A_n$ are ***mutually independent*** if

$$\mathbb{P}[A_1 \cap A_2 \cap \cdots \cap A_n] = \mathbb{P}[A_1]\,\mathbb{P}[A_2] \cdots \mathbb{P}[A_n].$$

Proposition 43. *If A, B are events with positive probabilities then the following are equivalent:*

(i.) *A and B are independent.*

(ii.) $\mathbb{P}[A \mid B] = \mathbb{P}[A]$ *and* $\mathbb{P}[B \mid A] = \mathbb{P}[B]$

Remark 44. (ii.) says that A and B have neither a positive nor negative relationship. In words, independence means that the probabilities of each event are unaffected by whether or not the other event occurs. Proposition 43 simply formalizes this idea using conditional probabilities.

Remark 45. A common mistake is to mix up "independence" with "mutually exclusive". Mutually exclusive events are typically **not** independent. That's because if you know that one happened, then you know that the other didn't. That's about as far from independence as you can get!

12 2026-02-11 | Week 05 | Lecture 12

12.1 Examples of using independence

This is based on chapter 2.5 in the textbook.

Recall that two events A and B are **independent** if $\mathbb{P}[A \cap B] = \mathbb{P}[A]\,\mathbb{P}[B]$. This means that knowledge of one event gives no information about the other.

Example 46. The proportions of blood types in the US are

A: 40% B: 11% AB: 4% O: 45%

(a) What is the probability that two randomly-selected people are both O.

Solution: Assuming that the two blood types are independent, we have

$$\begin{aligned}\mathbb{P}[\text{person 1 has O blood} \cap \text{person 2 has O blood}] &= \mathbb{P}[\text{person 1 has O blood}]\,\mathbb{P}[\text{person 2 has O blood}] \\ &= (.45)(.45) \\ &= 0.2025.\end{aligned}$$

(b) What is the probability that two randomly-selected people have matching blood types?

Solution: Let M be the event that the people have matching blood type, and let M_X be the event that both people have type X blood, for $X \in \{A, B, AB, O\}$.

$$\begin{aligned}\mathbb{P}[M] &= \mathbb{P}[M_A \cup M_B \cup M_{AB} \cup M_O] \\ &= \mathbb{P}[M_A] + \mathbb{P}[M_B] + \mathbb{P}[M_{AB}] + \mathbb{P}[M_O],\end{aligned}$$

where the last equality followed from the fact that the events M_A, M_B, M_{AB}, M_O are pairwise disjoint. We already computed $\mathbb{P}[M_O] = (.45)(.45) = .2025$. By similar calculations, we get

$$\mathbb{P}[M] = (.40)(.40) + (.11)(.11) + (.04)(.04) + (.45)(.45) = .3762.$$

End of Example 46. □

12.2 Dicrete Random Variables

Based on sections 3.1 and 3.2 of the textbook

We often associate numbers with outcomes of an experiment.

Definition 47 (Random variables)**.** Let S be a sample space of an experiment. A ***random variable*** is any rule that associates a number with each outcome in S. In other words, it is a function whose domain is S and whose range is $\mathbb{R}$.

The value of a random variable is thought of as varying depending on the outcome of the experiment. Random variables are usually denoted with capital letters, like X, Y, Z.

Example 48. Roll a 4-sided dice twice (this is the **experiment**). There are 16 possible **outcomes**. The **sample space** is

$$S = \{(x, y) : x, y \in \{1, 2, 3, 4\}\}.$$

Let X be the sum of the two rolls. We can represent X by the following table:

		Dice 2			
		1	2	3	4
Dice 1	1	2	3	4	5
	2	3	4	5	6
	3	4	5	6	7
	4	5	6	7	8

End of Example 48. □

Definition 49 (Discrete random variable)**.** We say that a random variable X is a ***discrete random variable*** if it can assume only a finite or countably infinite number of distinct values.

Definition 50 (Probability mass function (pmf))**.** Let X be a discrete random variable. The ***probability mass function*** (or ***pmf***) of X is the function

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

defined for every $x \in \mathbb{R}$.

Example 51. The pmf of X in Example 48 is

$$p(2) = 1/16 \quad p(3) = 2/16, \quad p(4) = 3/16, \quad p(5) = 4/16, \quad p(6) = 3/16, \quad p(7) = 2/16, \quad p(8) = 1/16$$

and $p(x) = 0$ for all other $x \in \mathbb{R}$. It is often useful to write these as a table:

x	$\mathbb{P}[X = x]$
2	1/16
3	2/16
4	3/16
5	4/16
6	3/16
7	2/16
8	1/16

End of Example 51. □

Definition 52. For a discrete random variable X, the ***expected value*** of X, denoted $\mathbb{E}[X]$, is

$$\mathbb{E}[X] = \sum_x x\mathbb{P}[X = x]$$

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

Remark 53. $\mathbb{E}[X]$ is the long-run average of X, if we were to repeat the experiment many times.

Example 54. Let X be the random variable in Example 48. Then

$$\begin{aligned}
\mathbb{E}[X] &= \sum_x x\mathbb{P}[X = x] \\
&= 2\left(\frac{1}{16}\right) + 3\left(\frac{2}{16}\right) + 4\left(\frac{3}{16}\right) + 5\left(\frac{4}{16}\right) + 6\left(\frac{3}{16}\right) + 7\left(\frac{2}{16}\right) + 8\left(\frac{1}{16}\right) \\
&= \frac{80}{16} \\
&= 5.
\end{aligned}$$

Therefore, the sum of two 4-sided dice rolls is, on average, 5.

End of Example 54. □

Definition 55 (Cumulative distribution function (cdf))**.** Let X be any random variable. The ***cumulative distribution function*** (or ***cdf***) of X is the function

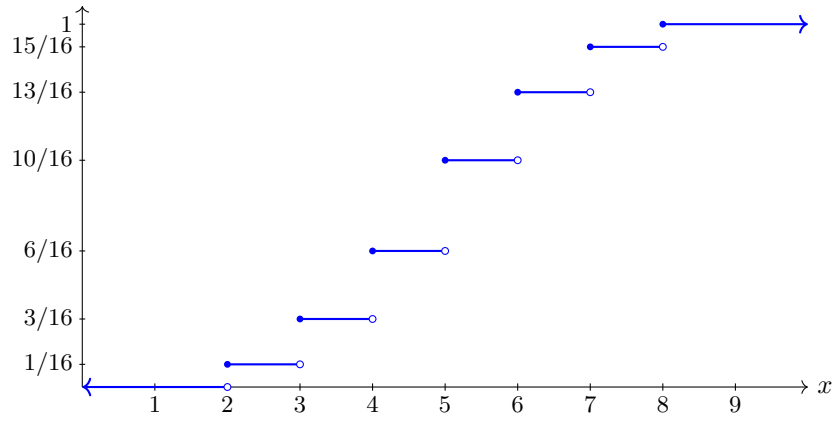
$$F(x) = \mathbb{P}[X \leq x],$$

defined for all $x \in \mathbb{R}$.

Example 56. For the random variable in Example 48, the cdf is given by the following table

x	$F(x) = \mathbb{P}[X \leq x]$
2	1/16
3	3/16
4	6/16
5	10/16
6	13/16
7	15/16
8	1

Here is a plot of $F(x)$. It is the step function shown in blue:



For dicrete random variables, the graph of a cdf will always look similar to the above: it will be an increasing step function with open circles on the right endpoints and filled-in circles on the left endpoints.

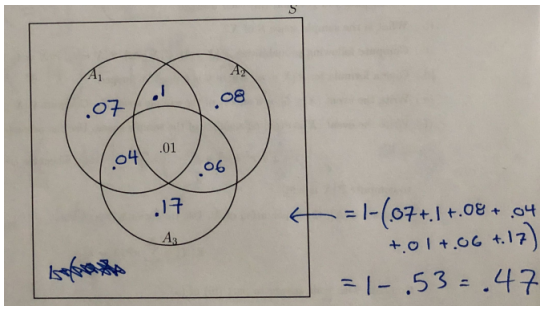
End of Example 56. □

13 2026-02-13 | Week 05 | Lecture 13

13.1 Exam review problems

We went over some example problems:

- The solution to problem 1 on homework 2. The key idea involves correctly filling out the venn diagram as shown:



- The solution to problem 14 on homework 4.
- An example about computing probabilities invovling the phrase "at least" (shown below).

Example 57. Suppose we roll a 6-sided dice three times. What's the probability that we get at least one 6?

Solution: The key idea is the following

$$\mathbb{P}\left[\text{at least one six is rolled}\right] = 1 - \mathbb{P}\left[\text{no sixes are rolled}\right]$$

To put this idea into practice, let $A_1 = [\text{the first dice roll is a 6}]$, $A_2 = [\text{the second dice roll is a 6}]$, and $A_3 = [\text{the third dice roll is a 6}]$. Clearly, A_1, A_2, A_3 are independent because the dice rolls don't influence each other. Then

$$\begin{aligned}
\mathbb{P}\left[\text{at least one six is rolled}\right] &= \mathbb{P}\left[A_1 \cup A_2 \cup A_3\right] \\
&= 1 - \mathbb{P}\left[A_1^c \cap A_2^c \cap A_3^c\right] && \text{by de Morgan's law} \\
&= 1 - \mathbb{P}\left[A_1^c\right]\mathbb{P}\left[A_2^c\right]\mathbb{P}\left[A_3^c\right] && \text{by independence} \\
&= 1 - \left(\frac{5}{6}\right)\left(\frac{5}{6}\right)\left(\frac{5}{6}\right) \\
&= 1 - \left(\frac{5}{6}\right)^3.
\end{aligned}$$

End of Example 57. □

14 2026-02-23 | Week 07 | Lecture 14

This lecture is based on chapter 3.3 in the textbook.

14.1 Expectation

Recall that 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 $\mathbb{E}[X]$.

Remark 58 (Notation)**.** We sometimes use the notation μ or μ_X for $\mathbb{E}[X]$.

The following is a simple but important fact:

Proposition 59 (Linearity of Expectation)**.** *Let a, b be numbers and X and Y be a random variables. Then*

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

and

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

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

Theorem 60 (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 61. 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):

$$\begin{aligned}\mathbb{E}[3Y+7] &= 3\mathbb{E}[Y]+7 && \text{by Proposition 59}\\ &= 3(0.3)+7 && \text{by our answer to part (a).}\\ &= 7.9.\end{aligned}$$

Solution to (c): Here we will apply Theorem 60 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 60 with $h(x) = e^x$:

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

End of Example 61. □

14.2 Variance

Definition 62 (Variance)**.** The ***variance*** of a random variable X is

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

where $\mu = \mathbb{E}[X]$. (Note: variance is often denoted by σ^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.

Example 63. Let X and Y be random variables defined as follows:

$$X = \begin{cases} +1 & : \text{ with probability } \frac{1}{2}\\ -1 & : \text{ with probability } \frac{1}{2}\end{cases} \quad \text{and} \quad Y = \begin{cases} +1000 & : \text{ with probability } \frac{1}{2}\\ -1000 & : \text{ with probability } \frac{1}{2}\end{cases}$$

In both cases, $\mathbb{E}[X] = 0$ and $\mathbb{E}[Y] = 0$. But despite having the same expected value, these are very different random variables: the value of Y deviates from its mean much more than X does. Indeed, direct computation (shown below) shows the variance of Y to be much larger than the variance of X:

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}\left[(X-\mu)^2\right]\\ &= \mathbb{E}\left[X^2\right] && \text{since } \mu = \mathbb{E}[X] = 0\\ &= (1)^2(.5)+(-1)^2(.5) && \text{by Theorem 60}\\ &= 1.\end{aligned}$$

On the other hand,

$$\begin{aligned}
\mathrm{Var}(Y) &= \mathbb{E}\left[(Y-\mu)^2\right] \\
&= \mathbb{E}\left[Y^2\right] && \text{since } \mu = \mathbb{E}[Y] = 0 \\
&= (1000)^2(.5) + (-1000)^2(.5) && \text{by Theorem 60} \\
&= 1000000.
\end{aligned}$$

End of Example 63. □

15 2026-02-25 | Week 07 | Lecture 15

15.1 Properties of Variance

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

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

where $\mu = \mathbb{E}[X]$.

For a discrete random variable X, this expectation can be written as

$$\mathrm{Var}(X) = \sum_x (x-\mu)^2 \mathbb{P}[X=x]$$

by Theorem 60.

The positive square root of the variance is called the ***standard deviation***.

There is a shortcut for computing variance, given in the next theorem.

Theorem 64 (Variance shortcut formula)**.** *Let X be a random variable with $\mathbb{E}\left[X^2\right] < \infty$. Then*

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

The condition $\mathbb{E}\left[X^2\right] < \infty$ in Theorem 64 is there to exclude the possibility that the right hand side of the formula be the undefined form $\infty - \infty$.

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 65. 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

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

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 $\mathbb{E}[Y] = -10\frac{1}{2} + 10\frac{1}{2} = 0$, 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 65. □

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.

Variance has several properties that are often useful, described in the next theorem.

Theorem 66 (Properties of Variance)**.** *Let X be a random variable. Then:*

- *For any scalars a, b,*
$$\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X). \tag{5}$$
- *If X and Y are independent random variables then*
$$\mathrm{Var}(aX + bY + c) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y). \tag{6}$$

Implications of Theorem 66:

- Taking $b = 0$ in Eq. (5) gives
$$\mathrm{Var}(aX) = a^2\mathrm{Var}(X). \tag{7}$$
Here is a direct proof of Eq. (7):
$$\begin{aligned}
\mathrm{Var}(aX) &= \mathbb{E}\left[(aX - \mathbb{E}[aX])^2\right] && \text{by Definition 62}\\
&= E\left[(aX - a\mathbb{E}[X])^2\right] && \text{by Proposition 59}\\
&= \mathbb{E}\left[a^2(X - \mathbb{E}[X])\right]\\
&= a^2\mathbb{E}[X - \mathbb{E}[X]] && \text{by Proposition 59}\\
&= a^2\mathrm{Var}(X) && \text{by Definition 62}
\end{aligned}$$
□
- Taking $a = 1$ in Eq. (5) gives
$$\mathrm{Var}(X + b) = \mathrm{Var}(X)$$
which says that adding a nonrandom quantity to a random variable doesn't change its variance.
- Taking $a = b = 1$ and $c = 0$ in Eq. (17) implies that
$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y). \tag{8}$$
This says that for independent random variables, variance is additive.

  (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.)

15.2 Binomial random variable

Next we introduce an important type of random variable, called the "binomial" random variable. The motivation is the following idea: if you flip a coin n times, and let $p \in (0, 1)$ be the probability of 'heads' on any given coin flip. We can define a random variable $X = (\#$ of heads in n flips$)$.

The probability that you get exactly k heads in n flips is given by the formula:

$$\binom{n}{k} p^k (1 - p)^{n-k}. \tag{9}$$

This is the probability that $X = k$. To see where Eq. (9) comes from, work out an example, say $n = 5$ and $k = 3$, and the general idea will become clear.

The precise definition of a binomial random variable is as follows:

Definition 67 (Binomial random variable)**.** A random variable is said to be a ***Binomial random variable*** with n trials and success probability p if

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

for every $k \in \{0, 1, \ldots, n\}$. In this case, we write $X \sim \text{Bin}(n, p)$.

Remark 68 (Useful characterization of binomial random variables)**.** If X is a Binomial random variable with n trials and success probability p, then we can write

$$X = \xi_1 + \ldots + \xi_n$$

where

$$\xi_i = \begin{cases} 1 & : \text{ with probability } p \\ 0 & : \text{ with probability } 1-p \end{cases}.$$

In words ξ_i is either one or zero depending on whether the i^{th} coin flip was 'heads' or not. Hence, the sum $\xi_1 + \ldots + \xi_n$ counts the number of heads in n coin flips.

Example 69. Suppose we wish to compute $\mathbb{E}[X]$, when X is binomial with n trials and success parameter p.

To do this computation, we'll use the idea of Remark 68:

$$\begin{aligned} \mathbb{E}[X] &= \mathbb{E}[\xi_1 + \ldots + \xi_n] \\ &= \mathbb{E}[\xi_1] + \ldots + \mathbb{E}[\xi_n] && \text{by Proposition 59.} \end{aligned}$$

Observe that $\mathbb{E}[\xi_i] = 1 \cdot + 0 \cdot (1-p) = p$ for all $i \in \{1, \ldots, n\}$, and hence

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

End of Example 69. □

16 2026-02-27 | Week 07 | Lecture 16

16.1 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

$$\mathbb{P}[W] = \sum_{k=2}^{12} \mathbb{P}[W \mid F_k] \times \mathbb{P}[F_k] \tag{10}$$

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 short 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,

$$\begin{aligned}
\mathbb{P}[W \mid F_4] &= \mathbb{P}[\text{roll 4} \mid \text{roll 4 or 7}] \\
&= \frac{\mathbb{P}[(\text{roll 4}) \text{ and } (\text{roll 4 or 7})]}{\mathbb{P}[\text{roll 4 or 7}]} && \text{by def of conditional prob.} \\
&= \frac{\mathbb{P}[\text{roll 4}]}{\mathbb{P}[\text{roll 4 or 7}]} \\
&= \frac{3/36}{3/36 + 6/36} \\
&= \frac{3}{9} && = \frac{1}{3}.
\end{aligned}$$

Continue in this manner to get all the values of the table. Then plugging the values from the table into Eq. (10) gives

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

Your chance of winning is about 49.3%.

16.2 Binomial random variable (continued)

(*Based on chapter 3.4 in the textbook.*)

Recall that a random variable is called a ***binomial random variable*** with number of trials n and success probability $p \in (0,1)$ if

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

for every $k \in \{0, \ldots, n\}$.

Interpretation: If p is the probability of a coin landing on heads, then

$$X = (\#\text{ of heads in } n \text{ coin tosses}).$$

Example 70 (Variance of a binomial random variable)**.** Let X be a binomial random variable. In the last lecture we showed that

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

Question: What is the variance of X?

Solution: Recall that we can write

$$X = \xi_1 + \ldots + \xi_n,$$

where

$$\xi_i = \begin{cases} 1 & : \text{ with probability } p \\ 0 & : \text{ with probability } 1-p \end{cases}.$$

Therefore, using the additivity of variance (Theorem 66), we have

$$\mathrm{Var}(X) = \mathrm{Var}(\xi_1 + \ldots + \xi_n) = \mathrm{Var}(\xi_1) + \ldots + \mathrm{Var}(\xi_n). \tag{11}$$

We need to compute $\mathrm{Var}(\xi_i)$, for $i = 1, \ldots, n$. Observe that by the shortcut formula for variance (Theorem 64, for every i, we have

$$\begin{aligned} \mathrm{Var}(\xi_i) &= \mathbb{E}\left[\xi_i^2\right] - \left(\mathbb{E}[\xi_i]\right)^2 \\ &= p - p^2 \\ &= p(1-p). \end{aligned}$$

Plugging this into Eq. (11) implies that

$$\mathrm{Var}(X) = np(1-p). \tag{12}$$

End of Example 70. □

17 2026-03-02 | Week 08 | Lecture 17

The next example presents another important discrete random variable: the geometric random variable.

17.1 Poisson random variable

Fix $\lambda \in (0,1)$. Suppose have the following coins:

- 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.

Let us focus now on the n^{th} experiment. Let $\xi_1, \ldots, \xi_n$ represent the n coin flips, where

$$\xi_k = \begin{cases} 1 & : \quad k^{\text{th}} \text{ coin flip is heads} \\ 0 & : \quad \text{otherwise} \end{cases}$$

for each $k \in \{1, \ldots, n\}$.

Let

$$X_n = \xi_1 + \xi_2 + \ldots + \xi_n = (\#\text{ of heads in } n \text{ flips})$$

Question: How many heads do we expect to get? We can use the linearity of expectation:

$$\begin{aligned} \mathbb{E}[X_n] &= \mathbb{E}[\xi_1 + \xi_2 + \ldots + \xi_n] \\ &= \mathbb{E}[\xi_1] + \mathbb{E}[\xi_2] + \ldots + \mathbb{E}[\xi_n] \\ &= \underbrace{\frac{\lambda}{n} + \frac{\lambda}{n} + \ldots + \frac{\lambda}{n}}_{n \text{ terms}} \\ &= \lambda. \end{aligned}$$

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 Y, one that we haven't seen yet? What properties would that new random variable have?

- Its expectation should be λ, since $\mathbb{E}[Y] = \lim_{n\to\infty} \mathbb{E}[X_n] = \lambda$
- It should be able to take any positive integer value, and not any other values.

Question: What do we expect the variance of Y to be?

To answer this, observe that, as the sum of n coin flips (each with $\mathbb{P}[\text{Heads}] = \frac{\lambda}{n}$), it follows by Remark 68 that $X_n \sim \text{Bin}(n, \lambda/n)$.

Applying Eq. (12) from Example 70, we find that

$$\text{Var}(X_n) = \lambda\left(1 - \frac{\lambda}{n}\right).$$

Therefore we expect the following to be true of Y:

$$\text{Var}(Y) = \lim_{n\to\infty} \text{Var}(X_n) = \lim_{n\to\infty} \lambda\left(1 - \frac{\lambda}{n}\right) = \lambda.$$

If this limit were 0, then we would expect our final result to not vary at all—in other words, we'd expect the limit to be a fixed number, rather than a random quantity. But since the limiting variance λ is positive, we expect the limit to be a random variable.

Qustion: What should the probability mass function of Y look like?

By Definition 67, the probability mass function of X_n is

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

It turns out that when n is large and p is small, the following approximation holds:

$$\binom{n}{k} p^k (1-p)^{n-k} \approx e^{-\lambda}\frac{\lambda^k}{k!}$$

So we might expect that in the limit, if our our coin-flipping process converges to a random variable Y, that random variable should have probability mass function

$$\mathbb{P}[Y = k] = e^{-\lambda}\frac{\lambda^k}{k!}.$$

Theorem 71 (The exponential function)**.** *For any real number x, the* ***exponential function****, denoted e^x, is the function*

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

Definition 72 (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

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

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

$$\begin{aligned}
\sum_{k=0}^{\infty} p(k) &= \sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!} \\
&= e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} \\
&= e^{-\lambda}e^{\lambda} && \text{by Theorem 71} \\
&= 1.
\end{aligned}$$

Proposition 73. *Let $X \sim Pois(\lambda)$. Then*

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

The Poisson distribution is a common model since it models "accidents".
We didn't finish the following example, so I'll start next time with it.

Example 74. 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 74. □

18 2026-03-04 | Week 08 | Lecture 18

18.1 Applications of the Poisson random variable

(Based on section 3.6 in the textbook.)

Recall that a random variable X is said to have a ***Poisson distribution*** with mean $\lambda > 0$ if

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

The Poisson distribution is a common model for modeling the number of "accidents". Recall that the expected value and variance of a Poisson random variable are both equal to λ.

Example 75. 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 75. □

In the previous lecture, we introduce the Poisson distribution as a mystery limit of coin flips, where (1) the number of coin flips is increasing and (2) the probability of heads is decreasing at a proportional rate. Let us now formalized that idea in the following theorem:

Theorem 76 (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:

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

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

Let's try an application of this:

Example 77 (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 77. □

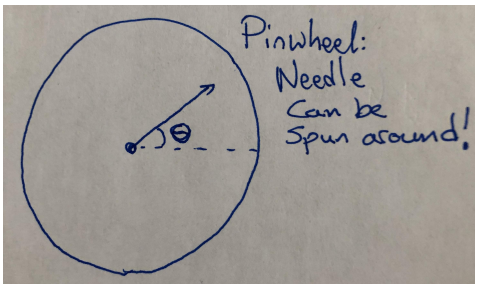
18.2 Continuous random variables

(This section is based on Chapter 4.1)

Recall that a random variable is ***discrete*** if it can assume only a finite number of values, or if its values can be listed in an infinite sequence. Chapter 4 introduces another class of random variables, which are called continuous random variables.

Definition 78 (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 $x \in I$.

Example 79 (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. So X is a random variable. If X is measured in degrees, then

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

but A 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.

End of Example 79. □

Definition 80 (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*** of X.

19 2026-03-06 | Week 08 | Lecture 19

19.1 Continuous random variables

(This lecture is based on sections 4.1-4.3 in the textbook)

Definition 81 (pdf)**.** Let X be a continuous random variable. The ***probability density function*** (aka "pdf" or "density") of X is a function f such that for any real numbers a, b with $a \leq b$, we have

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

The graph of $y = f(x)$ is called the ***density curve*** of X.

Observatioins about pdfs:

- $f(x) \geq 0$ for all x
- $\int_{-\infty}^{\infty} f(x)dx = 1$
- $f(x)$ is not a probability because it could be greater than 1. Insteady, you can think of it vaguely as a "relative likeliood".

Proposition 82. *Let $F(x) = \mathbb{P}[X \leq x]$ be the cdf of X, and let f be the pdf of X. Then*

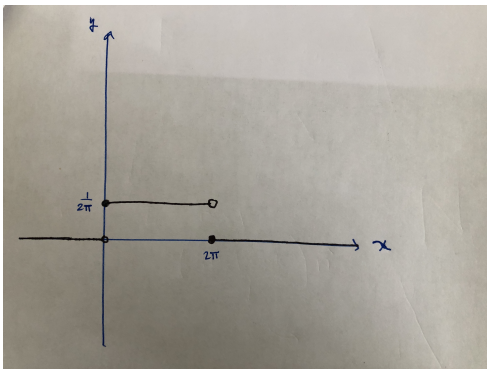
$$F(x) = \int_{-\infty}^{x} f(t)dt \quad \textit{and} \quad F'(x) = f(x)$$

Proposition 83 (Properties of cdfs of uniform random variables)**.**

- $\mathbb{P}[a \leq X \leq b] = F(b) - F(a)$
- *F is a nondecreasing function*
- *$F(-\infty) = 0$ and $F(+\infty) = 1$.*

Example 84 (Pinwheel example continued)**.** Recall the example of the pinwheel Example 79, 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

$$\begin{aligned} F_X(x) &= \int_{-\infty}^{x} \frac{1}{2\pi} \cdot \mathbf{1}_{[0,2\pi)}(t)dt \\ &= \frac{1}{2\pi} \int_{-\infty}^{x} \mathbf{1}_{[0,2\pi)}(t)dt. \end{aligned} \tag{13}$$

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. (13) is zero.
- When $x \geq 2\pi$, the integral equals

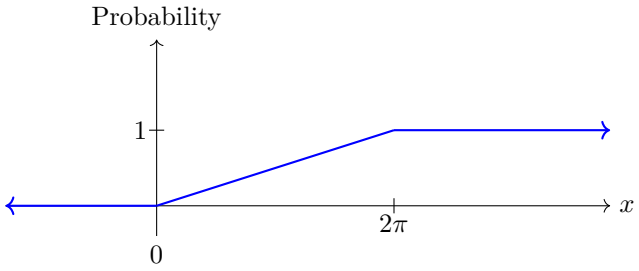
$$\int_{-\infty}^{0} 0dt + \frac{1}{2\pi} \int_{0}^{2\pi} dt + \int_{2\pi}^{x} 0dt = 0 + 1 + 0 = 1$$

- 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. (13), we can conclude that

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



End of Example 84. □

Definition 85 (Uniform random variable)**.** 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 it has pdf

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

In this case, X is called a ***uniform*** random variable.

Definition 86 (Exponential random variable)**.** We say that X is an ***exponential*** random variable with rate $\lambda > 0$ if it has pdf

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

Remark 87. In Definition 86, the term $\mathbf{1}_{[x \geq 0]}$ is an "indicator function" which takes value 1 when $x \geq 0$ and 0 if $x < 0$.

20 2026-03-09 | Week 09 | Lecture 20

20.1 Continuous random variables

Example 88 (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

$$f(x) = \lambda e^{-\lambda x} \mathbf{1}_{[x \geq 0]} \tag{14}$$

for a certain (fixed) value of $\lambda > 0$.

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 88. □

Definition 89 (Expected value – continuous random variables)**.** The ***expected value*** of a continuous random variable X is

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

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

We also have an analogue of the law of the unconscious statistician (Theorem 60)

Theorem 90 (LOTUS - continuous case)**.** *Let X be a continous random variable with pdf f. Then for any function $h : \mathbb{R} \to \mathbb{R}$,*

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

provided that the integral exists and is finite.

Suppose $\mu = \mathbb{E}[X]$. Applying Theorem 90 to the function $h(x) = (x - \mu)^2$ gives the following formula for the variance of a continuous random variable:

$$\mathrm{Var}(X) := \mathbb{E}\left[(X - \mu)^2\right] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx.$$

And again, we have the shortcut formula

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

21 2026-03-11 | Week 09 | Lecture 21

21.1 Normal random variables

This section is based on chapter 4.3.

Definition 91 (Normal / Gaussian Distribution)**.** A random variable X is said to be ***normally distributed*** with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$ if it has pdf

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

defined for all $x \in \mathbb{R}$. We abbreviate this by writing $X \sim \mathcal{N}(\mu, \sigma^2)$. If $X \sim \mathcal{N}(0,1)$, we say Y has ***standard normal distribution***. The parameter σ is called the ***standard deviation*** of X, and is the positive square root of the variance.

Remark 92 (Notation)**.**

- It is a common convention to denote a standard normal random variable with the letter Z.
- random variables/distributions are also called ***Gaussian*** random variables/distributions.

Example 93. Normal distributions are the classic "bell curve" distributions. Whenever a random quantity is the result of adding up many small, independent random effects, we expect it to be approximately normally distributed.

- A classic example is human height, which is influenced by the combined effects of thousands of genes together with environmental factors.
- Another example is repeated measurements of the same physical quantity (e.g., if we all went out and measured the temperature outside tomorrow at midday), this is because the total error is the sum of many small independent sources of noise (e.g., instrument precision, time, location, shade, wind, etc).

End of Example 93. □

Remark 94 (Standardization)**.** If $Y \sim \mathcal{N}(\mu, \sigma^2)$ then the random variable

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

is a standard normal random variable (i.e., $Z \sim \mathcal{N}(0,1)$. The transformation

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

in which we subtract off the mean of X and then divide by its standard deviation is called ***standardization***.

Remark 95 (The Empirical Rule)**.** The **Empirical Rule** (also called the **68-95-99.7 rule**; see textbook section 1.3) is summarized by the following figure, for normal distributions:

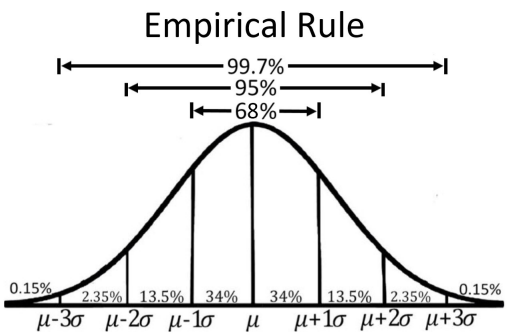


Image source: https://andymath.com/wp-content/uploads/2019/12/empirical-rule-normdist2.jpg

Remark 96. Note that μ determines where the peak of the bell curve is located, and σ determines how wide or narrow it is.

The next example shows how we can use standardization to do calculations:

Example 97. 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?

$$\begin{aligned}
\mathbb{P}\left[\mu - \sigma \leq X \leq \mu + \sigma\right] &= \mathbb{P}\left[\frac{(\mu-\sigma)-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{(\mu+\sigma)-\mu}{\sigma}\right] \\
&= \mathbb{P}\left[\frac{(\mu-\sigma)-\mu}{\sigma} \leq Z \leq \frac{(\mu+\sigma)-\mu}{\sigma}\right] && \text{by Remark 94} \\
&= \mathbb{P}\left[-1 \leq Z \leq 1\right] \\
&= \int_{-1}^{1} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx \\
&\approx .68 && \text{using the figure in Remark 95}
\end{aligned}$$

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

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

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

End of Example 97. □

22 2026-03-23 | Week 10 | Lecture 22

22.1 Joint distributions - discrete case

Please read chapter 5.1

We are interested in understanding the properties of ***random vectors***, which are vectors whose entries are random variables. We'll start with the case where the random variables are discrete.

Definition 98 (Joint pmf)**.** Let X and Y be two discrete random variables. The ***joint pmf*** of X and Y is the function

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

defined for all $x, y \in \mathbb{R}$.

Remark 99.

- It must be the case that $p(x,y) \geq 0$ and $\sum_{x,y} p(x,y) = 1$.
- Let A be any set consisting of pairs of (x,y) values. Then the probability

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

Example 100. 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
	0	.840	.030	.020	.010
X	1	.060	.010	.008	.002
	2	.010	.005	.004	.001

The probability that there are no defects in the car is

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

The probability that there will be exactly one defect is

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

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:

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

End of Example 100. □

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, and in this context we call them "marginal pmfs". The manner in which this is done is suggested by the previous example.

Definition 101. If $p(x, y)$ is the pmf of (X, Y), then the ***marginal pmf*** of X is

$$p_X(x) = \mathbb{P}[X = x] = \sum_{\text{all } y} p(x, y).$$

Similary, the ***marginal pmf*** of Y

$$p_Y(y) = \mathbb{P}[Y = y] = \sum_{\text{all } x} p(x, y).$$

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 102 (Independence of rvs - discrete case)**.** Let X, Y be discrete random variables with joint pmf $p(x, y)$. We say that X and Y are ***independent*** if for every pair of (x, y)-values,

$$p(x, y) = p_X(x) \cdot p_Y(y)$$

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

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

Example 103. In the car factory example, X and Y have with joint pmf is specified by the table

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

where, for example $p(1, 3) = .002$.

The marginal pmfs are given by the following tables:

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

and

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

Are X and Y independent? 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 103. □

22.2 Joint distributions - continuous case

Based on chapter 5.1 in the textbook – in particular, please read examples 5.5 and 5.14 in the textbook.

The theory for continuous random vectors parallels closely the theory for discrete random variables presented in Section 22.1.

Definition 104. 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,

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

The ***marginal densities*** of X and Y are

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

We say that X and Y are ***independent*** if

$$f(x,y) = f_X(x) \cdot f_Y(y).$$

for all $x, y \in \mathbb{R}$.

23 2026-03-25 | Week 10 | Lecture 23

23.1 An example of using a joint distribution to compute probabilities

Example 105. Let X and Y be the lifetimes of two lightbulbs, in months. Assume that the expected lifetimes are 10 months and 12 months, respectively. Also assume that the lifetimes are independent, exponentially-distributed random variables.

Then

$$f_X(x) = \frac{1}{10}e^{-\frac{1}{10}x}\mathbf{1}_{[x>0]} \quad \text{and} \quad f_Y(y) = \frac{1}{12}e^{-\frac{1}{12}y}\mathbf{1}_{[y>0]}.$$

Since X and Y are independent, the joint pdf is

$$\begin{aligned} f(x,y) &= f_X(x)f_Y(y) \\ &= \left(\frac{1}{10}e^{-\frac{1}{10}x}\mathbf{1}_{[x>0]}\right)\left(\frac{1}{12}e^{-\frac{1}{12}y}\mathbf{1}_{[y>0]}\right) \\ &= \frac{1}{120}e^{-\frac{1}{10}x-\frac{1}{12}y}\mathbf{1}_{[x,y>0]} \\ &= \begin{cases} \frac{1}{120}e^{-\frac{1}{10}x-\frac{1}{12}y} & : \quad x>0 \text{ and } y>0 \\ 0 & : \quad \text{else} \end{cases} \end{aligned}$$

Question: What is the probability that both lightbulbs last at least 9 months?

Want:

$$\begin{aligned} \mathbb{P}\left[X \geq 9, Y \geq 9\right] &= \int_9^\infty \int_9^\infty f(x,y)dxdy \\ &= \int_9^\infty \int_9^\infty f_X(x)f_Y(y)dxdy \\ &= \int_9^\infty \left[\int_9^\infty f_X(x)f_Y(y)dx\right]dy \\ &= \int_9^\infty \left[\int_9^\infty f_X(x)dx\right]f_Y(y)dy \\ &= \left(\int_9^\infty f_X(x)dx\right)\left(\int_9^\infty f_Y(y)dy\right) \\ &= \left(\int_9^\infty \frac{1}{10}e^{-\frac{1}{10}x}dx\right)\left(\int_9^\infty \frac{1}{12}e^{-\frac{1}{12}y}dy\right) \\ &= \left[-e^{\frac{1}{10}x}\right]_{x=9}^{x=\infty} \cdot \left[-e^{\frac{1}{12}y}\right]_{y=9}^{y=\infty} \\ &= e^{-9\left(\frac{1}{10}+\frac{1}{12}\right)} \\ &= e^{-1.65} \\ &\approx .2. \end{aligned}$$

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

End of Example 105. □

23.2 LOTUS Theorem - multivariate case

Chapter 5.2 in the textbook.

Theorem 106 (LOTUS - multivariate case)**.** *Let X, Y be random variables with joint pmf $p(x,y)$ if X, Y are discrete, or joint pdf $f(x,y)$ if X, Y are continuous. Then for any function $h : \mathbb{R}^2 \to \mathbb{R}$,*

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

An important aplication of Theorem 106:

Theorem 107. *If X, Y are indepenedent random variables then*

$$\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$$

Proof. Let's prove the case with X, Y discrete. Let $p(x, y)$ be the joint pmf of X and Y. Taking $h(x, y) = xy$, Theorem 106 implies

$$\begin{aligned}
\mathbb{E}[XY] &= \sum_{\text{all } x,y} xyp(x, y) \\
&= \sum_{\text{all } x,y} xyp_X(x) \cdot p_Y(y) && \text{by independence} \\
&= \left(\sum_{\text{all } x} xp_X(x)\right)\left(\sum_{\text{all } y} yp_Y(y)\right) \\
&= \mathbb{E}[X]\,\mathbb{E}[Y] && \text{by Theorem 60}
\end{aligned}$$

The continuous case is similar. □

One important function for Theorem 106 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***:

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

By foiling this out, we get

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

Interpetation: covariance is a rough measure of the relationship between the two random variables. By Theorem 107, if X, Y are independent, then

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mu_X\mu_Y = \mathbb{E}[X]\,\mathbb{E}[Y] - \mu_X\mu_Y = 0.$$

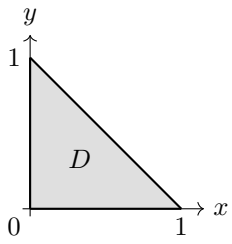
Example 108 (Covariance calculation)**.** Let X, Y be random variables with joint distribution

$$f(x, y) = 24xy$$

whose domain is

$$D := \left\{(x, y) \in \mathbb{R}^2 : x \geq 0, y \geq 0, x + y \leq 1\right\}$$

which is shown below:



What is the covariance 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

$$\begin{aligned}
\mathbb{E}[XY] &= \int_D h(x,y) f(x,y) dxdy \\
&= \int_D (xy) 24xy dxdy \\
&= 24 \int_D x^2 y^2 dxdy \\
&= 24 \int_0^1 \int_0^{1-y} x^2 y^2 dxdy \\
&= 24 \int_0^1 y^2 \left[\int_0^{1-y} x^2 dx \right] dy \\
&= 8 \int_0^1 y^2 (1-y)^3 dy \\
&= \dots \\
&= -8.
\end{aligned}$$

End of Example 108. □

24 2025-03-27 | Week 10 | Lecture 24

24.1 Statistics: a general framework

Chapters 5.3 and 5.4

The objective of statistics to is to make an inference about a population based on information contained in a sample from that population and to provide an associated measure of goodness for the inference. Here we introduce a conceptual framework for mathematical statistics.

A ***population*** is a large body of data that is the target of our interest. The subset collected from it is our ***sample***.

Example 109 (Populations)**.** A population can be real or theoretical. Here are some examples

- The set of people in Hawai'i (real, finite)
- The set of voters in the 2026 Midterm elections (hypothetical, finite)
- The decimal expansion of π (countably infinite)
- The infinitely many observations that could be made during a laboratory experiment if the experiment were repeated over and over again (hypothetical)
- The lifetimes of light bulbs produced by a factory

Importantly, a population can also be a *probability distribution*, specified by a pdf, pmf, or cdf

- Observations made from an exponential distribution with rate $\lambda > 0$ (i.e., the distribution with pdf $f(x) = \lambda e^{-\lambda x}\mathbf{1}_{[x>0]}$.)

End of Example 109. □

More commonly in mathematical statistics, we think of the population as a distribution.

Definition 110 (Random sample from a distribution)**.** Consider a given probability distribution on $\mathbb{R}$ that can be represented by a pdf or pmf f. We say that the random variables $X_1, \ldots, X_n$ form a ***random sample*** from this distribution if these random variables are independent and the distribution of each is given by f. Such random variables are also said to be ***independent and identically distributed*** (iid). The number of random variables n is the ***sample size***.

24.2 What is a statistic?

Definition 111 (Statistic)**.** A ***statistic*** is any quantity whose value can be calculated from sample data.

Here are two important examples of statistics:

Example 112. Let $X_1, \ldots, X_n$ be random sample.

- The ***sample mean***, denoted $\overline{X}$, is
$$\overline{X} = \frac{X_1 + \ldots + X_n}{n}.$$
- The ***sample variance***, denoted S^2, is
$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

End of Example 112. □

Because a statistic is a *function* of random variables, it is itself a random variable. The probability distribution of a statistic is called its ***sampling distribution***.

Theorem 113. *Let $X_1, \ldots, X_n$ be a random sample from a distribution with mean μ and standard deviation $\sigma > 0$. Then*

- $\mu_{\overline{X}} := \mathbb{E}\left[\overline{X}\right] = \mu$
- $\sigma^2_{\overline{X}} := \mathrm{Var}(\overline{X}) = \sigma^2/n$ *(and $\sigma_{\overline{X}} = \sigma/\sqrt{n}$)*

Proof. First we compute $\mathbb{E}\left[\overline{X}\right]$:

$$\begin{aligned}
\mathbb{E}\left[\overline{X}\right] &= \mathbb{E}\left[\frac{X_1 + \ldots + X_n}{n}\right] && \text{by definition of } \overline{X} \\
&= \frac{1}{n}\left(\mathbb{E}\left[X_1\right] + \ldots + \mathbb{E}\left[X_n\right]\right) && \text{by linearity of expectation} \\
&= \frac{1}{n}\left(n\mu\right) && \text{since } \mathbb{E}\left[X_i\right] = \mu \text{ for all } i \\
&= \mu.
\end{aligned}$$

Next we compute $\mathrm{Var}(\overline{X})$:

$$\begin{aligned}
\mathrm{Var}(\overline{X}) &= \mathrm{Var}\left(\frac{X_1 + \ldots + X_n}{n}\right) && \text{by defintion of } \overline{X} \\
&= \frac{1}{n^2}\mathrm{Var}\left(X_1 + \ldots + X_n\right) && \text{by Eq. (7)} \\
&= \frac{1}{n^2}\left(\mathrm{Var}(X_1) + \ldots + \mathrm{Var}(X_n)\right) && \text{by additivity (see Eq. (8))} \\
&= \frac{1}{n^2}\left(n\sigma^2\right) && \text{since } \mathrm{Var}(X_i) = \sigma^2 \text{ for all } i \\
&= \frac{\sigma^2}{n}.
\end{aligned}$$

□

Theorem 114. *Let $X_1, \ldots, X_n$ be a random sample from a normal distribution with mean μ and standard deviation σ. Then for any n, $\overline{X} \sim \mathcal{N}(\mu, \sigma^2/n)$.*

25 2026-04-06 | Week 12 | Lecture 25

Based on chapter 5.4

25.1 The central limit theorem

Theorem 115 (Central limit theorem)**.** *Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean μ and variance $\sigma^2 > 0$. If n is sufficiently large, then $\overline{X}$ has approximately a normal distribution with mean $\mu_{\overline{X}} = \mu$ and variance $\sigma^2_{\overline{X}} = \frac{\sigma^2}{n}$. The larger n, the better the approximation.*

In symbols,

$$\overline{X} \approx \underbrace{\mu + \frac{\sigma}{\sqrt{n}} Z}_{\mathcal{N}(\mu, \frac{\sigma^2}{n})} \tag{15}$$

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

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

Corollary 116. *The distribution of the standardardized sample mean is approximately a standard normal; that is,*

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \approx Z, \tag{16}$$

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

Example 117 (Witch's brew - similar to example 5.27 in textbook)**.** 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 $\sigma = 1.5g$. Assume 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 $3.5 \leq \overline{X} \leq 3.8$?
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

$$\begin{aligned}
\mathbb{P}\left[3.5 \leq \overline{X} \leq 3.8\right] &\approx \mathbb{P}\left[3.5 \leq 4 + \frac{3}{2\sqrt{50}} Z \leq 3.8\right] \\
&= \mathbb{P}\left[-\frac{1}{2} \leq \frac{3}{2\sqrt{50}} Z \leq -\frac{1}{5}\right] \\
&= \mathbb{P}\left[-\frac{\sqrt{50}}{3} \leq Z \leq -\frac{2\sqrt{50}}{15}\right] \\
&= \mathbb{P}\left[-2.36 \leq Z \leq -.94\right] \\
&= \int_{-2.36}^{-.94} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy \\
&\approx .16
\end{aligned}$$

End of Example 117. □

Example 118 (CLT Coin Flipping)**.** Flip a fair coin 100 times. What is the probability that you get at least 55 heads?

Solution: Let

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

We want to compute $\mathbb{P}[X_1 + \ldots + X_n \geq 55]$, or equivalently,

$$\mathbb{P}\left[\overline{X} \geq .55\right].$$

We will use Eq. (16). First observe that

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

and that

$$\sigma = \sqrt{\text{Var}(X_i)} = \frac{1}{2}$$

(you should check these equations).

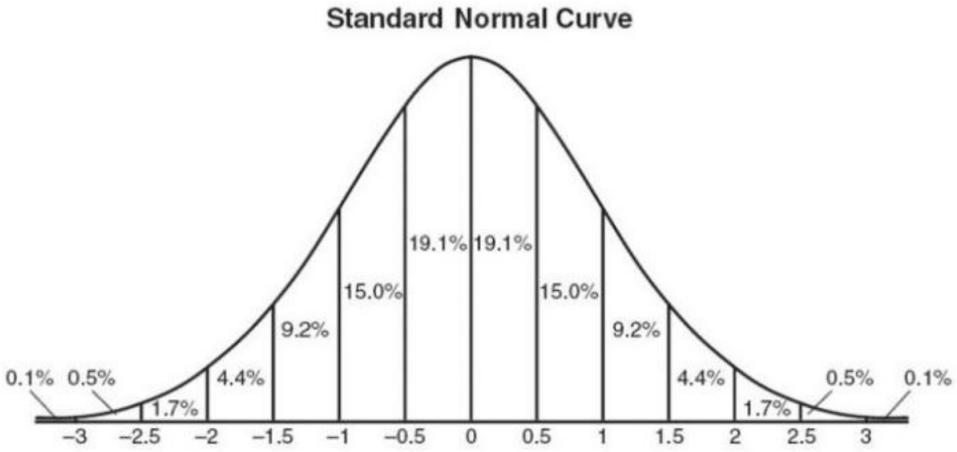
Therefore

$$\begin{aligned} \mathbb{P}\left[\overline{X} \geq .55\right] &= \mathbb{P}\left[\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \geq \frac{.55 - \mu}{\sigma/\sqrt{n}}\right] \\ &\approx \mathbb{P}\left[Z \geq \frac{.55 - \mu}{\sigma/\sqrt{n}}\right] \qquad \text{by CLT (Eq. (16))} . \end{aligned}$$

The above is the key step. Plugging in $n = 100$, $\mu = \frac{1}{2}$ and $\sigma = \frac{1}{2}$ gives

$$\mathbb{P}\left[\overline{X} \geq .55\right] = \mathbb{P}[Z \geq 1] \approx .16,$$

where we get .16 by consulting the following figure:



End of Example 118. □

Remark 119. When reviewing the key step in Example 118, it is helpful to recall the standardization transformation defined in Remark 94. The central limit theorem says that, approximately, $\overline{X} \sim \mathcal{N}(\mu, \sigma^2/n)$. Hence, by Remark 94,

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \approx Z$$

where $Z \sim \mathcal{N}(0, 1)$. This is what justifies the key step in Example 118.

25.2 Point estimation

Section 6.1

Recall a **population** is a large body of data that is the target of our interest, and a **statistic** is any quantity whose value can be computed from a sample.

Often, we have some quantity of interest, called a ***target parameter***, and we want a single "best guess" of some quantity of interest. When this is the case, the we call a statistic an estimator:

Definition 120 (Estimator)**.** An ***estimator*** is a statistic, that is a function $T(X_1, \ldots, X_n)$ of a sample, that is used to approximate a target parameter. An ***estimate*** is the realized value of an estimator (e.g., a number) that is obtained when the sample is actually taken.

In words, an estimator is a rule, often expressed as a formula, that tells us how to calculate the value of an estimate based on the measurements contained in a sample.

26 2026-04-08 | Week 12 | Lecture 26

26.1 Point estimation

Chapter 6.1

Recall a **population** is a large body of data that is the target of our interest, and a **statistic** is any quantity whose value can be computed from a sample.

Often, we have some quantity of interest, called a ***target parameter***, and we want a single "best guess" of some quantity of interest. When this is the case, the we call a statistic an estimator:

Definition 121 (Estimator)**.** An ***estimator*** is a statistic, that is a function $T(X_1, \ldots, X_n)$ of a sample, that is used to approximate a target parameter. An ***estimate*** is the realized value of an estimator (e.g., a number) that is obtained when the sample is actually taken.

In words, an estimator is a rule, often expressed as a formula, that tells us how to calculate the value of an estimate based on the measurements contained in a sample.

Example 122 (Broccoli Poll)**.** Suppose the fraction of people in Hawaii who like broccoli is p. This is an unknown **target 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

$$\widehat{p} = \frac{20}{100} = .2$$

The **sample** in this example consists of n random variables $X_1, \ldots, X_n$, where

$$X_i = \begin{cases} 1 & : \text{ i-th sampled person likes broccoli} \\ 0 & : \text{ they don't} \end{cases}$$

Our **point estimate** is $\widehat{p} = 0.2$ and the **point estimator** is the function $\widehat{p} = \frac{X_1 + \ldots + X_n}{n}$, which is a **sample mean**.

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

$$\widehat{p} \to p \text{ as } n \to \infty.$$

End of Example 122. □

Definition 123 (Bias and standard error)**.** Let $\widehat{\theta}$ be a point estimator of a target parameter θ. The ***bias*** of $\widehat{\theta}$ is the number $\text{Bias}(\widehat{\theta}) = \mathbb{E}[\widehat{\theta}] - \theta$. If $\text{Bias}(\widehat{\theta}) = 0$, we say that $\widehat{\theta}$ is ***unbiased***, otherwise we say it is ***biased***.

The ***standard error*** of $\widehat{\theta}$, denoted $\sigma_{\widehat{\theta}}$, is its standard deviation:

$$\sigma_{\widehat{\theta}} = \sqrt{\text{Var}(\widehat{\theta})}$$

Example 124 (Broccoli poll continued)**.** In the broccoli example, the estimator we used was the ***sample mean:***

$$\widehat{p} = \frac{X_1 + \ldots + X_n}{n}.$$

Claim 1: $\widehat{p}$ is unbiased.

Proof of Claim 1.

$$\begin{aligned}\mathbb{E}\left[\widehat{p}\right] &= \mathbb{E}\left[\frac{X_1 + \ldots + X_n}{n}\right] \\ &= \frac{1}{n}\left(\mathbb{E}\left[X_1\right] + \ldots + \mathbb{E}\left[X_n\right]\right) \\ &= \frac{1}{n}(np) \\ &= p.\end{aligned}$$

□ Claim

Claim 2: $\sigma_{\widehat{p}} = \sqrt{\frac{p(1-p)}{n}}$.

Proof of Claim 2. First observe that for every $i \in \{1, \ldots, n\}$, we have

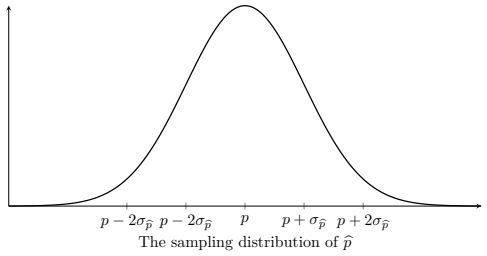
$$\mathrm{Var}(X_i) = \mathbb{E}\left[X_i^2\right] - \left(\mathbb{E}\left[X_i\right]\right)^2 = p - p^2 = p(1-p).$$

Using Theorem 66,

$$\begin{aligned}\mathrm{Var}(\widehat{p}) &= \mathrm{Var}\left(\frac{X_1 + \ldots + X_n}{n}\right) \\ &= \frac{1}{n^2}\left(\mathrm{Var}(X_1 + \ldots + X_n)\right) \\ &= \frac{1}{n^2}\left(\mathrm{Var}(X_1) + \ldots + \mathrm{Var}(X_n)\right) && \text{by independence} \\ &= \frac{1}{n^2}\left(np(1-p)\right) \\ &= \frac{p(1-p)}{n}\end{aligned}$$

Taking square roots implies the statement of the claim. □ Claim

Finally, since $\widehat{p}$ is a sample mean, the central limit theorem implies that its distribution is approximately normal:



The sampling distribution of $\widehat{p}$

As n grows, the value of $\sigma_{\widehat{p}} = \sqrt{\frac{p(1-p)}{n}}$ gets smaller, so the above curve (which is centered around the true parameter value p) gets narrower.

End of Example 124. □

Example 125 (Estimating the sides of a dice)**.** 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.

Given n rolls, the sample consists of the n random variables

$$X_1, \ldots, X_n$$

where X_i is the i^{th} dice roll. We'll give two examples of estimators:

- Our first example is called the *running maximum*, and it is defined as

$$\widehat{\theta} = \max(X_1, \ldots, X_n).$$

For example, if I roll the dice four times and obtain $3, 11, 7$, and 4, then your estimate would be that the dice has 11 sides.

This sounds silly, but with enough data this estimator will give you the correct answer with high probability. This is because, if I roll the dice enough times, then I will eventually roll the highest number; hence,

$$\lim_{n\to\infty} \widehat{\theta}(X_1, \ldots, X_n) = \theta.$$

One drawback is that this estimator is biased because $\mathbb{E}[\widehat{\theta}] < \theta$. This is because the estimator will "usually" be less than the true number of sides, and will never be more—so on average, $\widehat{\theta}$ will be smaller than θ.

- Our second example is what is called the *method of moments estimator*. In this setting, it's defined as

$$\widehat{\theta} = 2\left(\frac{X_1 + \ldots + X_n}{n}\right) - 1.$$

Let's demystify this formula. Suppose X a random variable representing one roll of the dice. We know that since the dice has θ sides, X has expected value

$$\mathbb{E}[X] = \frac{1}{\theta}(1 + 2 + \ldots + \theta) = \frac{\theta + 1}{2}.$$

Rearranging, we get

$$\theta = 2\mathbb{E}[X] - 1. \tag{17}$$

We also know that by the law of large numbers, $\mathbb{E}[X] \approx \frac{X_1 + \ldots + X_n}{n}$. Combining this with Eq. (17) implies

$$\theta \approx 2\left(\frac{X_1 + \ldots + X_n}{n}\right) - 1.$$

The right hand side is what we've chosen to be our $\widehat{\theta}$, and hopefuly this calculation makes it seem like a reasonable choice.

This estimator is unbiased. To see why, observe that

$$\begin{aligned}
\mathbb{E}[\widehat{\theta}] &= \mathbb{E}\left[2\left(\frac{X_1 + \ldots + X_n}{n}\right) - 1\right] \\
&= 2\mathbb{E}\left[\frac{X_1 + \ldots + X_n}{n}\right] - 1 \\
&= 2\mathbb{E}[X] - 1 \\
&= \theta && \text{by Eq. (17).}
\end{aligned}$$

In our example, with dice rolls $3, 11, 7$, and 4, our estimator $\widehat{\theta}$ returns the estimate

$$2\left(\frac{3 + 11 + 7 + 4}{4}\right) - 1 = 2\left(\frac{25}{4}\right) - 1 = 11.5.$$

So your estimate would be that the dice has 11.5 sides.

End of Example 125. □

Given two unbiased estimators, then *ceteris paribus*, we would typically prefer the estimator which has lower variance.

27 2026-04-13 | Week 13 | Lecture 27

27.1 Confidence intervals

Section 7.1

Definition 126 (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

$$\mathbb{P}\left[|\widehat{p}-p| < \epsilon\right] \geq .95 \tag{18}$$

In words, the interval contains the true parameter with probability at least .95.

Example 127. 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

$$X_i = \begin{cases} 1 & : \text{ i-th sampled person likes broccoli} \\ 0 & : \text{ they don't} \end{cases}$$

By the central limit theorem,

$$\widehat{p} \approx p + \frac{\sigma}{\sqrt{n}} Z \tag{19}$$

where $Z \sim \mathcal{N}(0,1)$ and $\sigma = \sqrt{\mathrm{Var}(X_i)} = \sqrt{p(1-p)}$. We also don't know what σ is, since it depends on the unknown parameter p, but we can approximate it:

$$\sigma = \sqrt{\mathrm{Var}(X_i)} = \sqrt{p(1-p)} \approx \sqrt{\widehat{p}(1-\widehat{p}) = (.2)(.8)} = \sqrt{.16} = .4 \tag{20}$$

By Eq. (19),

$$\begin{aligned} |\widehat{p}-p| &= \frac{\sigma}{\sqrt{n}}|Z| \\ &= \frac{.4}{10}|Z| \\ &= .04|Z| \end{aligned}$$

Therefore

$$\begin{aligned} \mathbb{P}\left[|\widehat{p}-p| \leq \epsilon\right] &= \mathbb{P}\left[.04|Z| \leq \epsilon\right] \\ &= \mathbb{P}\left[|Z| \leq 25\epsilon\right] \\ &= \mathbb{P}\left[-25\epsilon \leq Z \leq 25\epsilon\right] \end{aligned}$$

By the empirical rule, $\mathbb{P}\left[-2 \leq Z \leq 2\right] = .95$, so we need $25\epsilon = 2$, or $\epsilon = .08$.

Therefore the 95% confidence interval is the interval

$$(\widehat{p}-\epsilon, \widehat{p}+\epsilon) = (.12, .28).$$

We conclude that with 95% confidence, the true proportion of people in Hawaii who like broccoli is between 12% and 28%. Our point estimate is 0.2, and our ***margin of error*** is 0.08.

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?

End of Example 127. □

28 2026-04-15 | Week 13 | Lecture 28

28.1 The Coupon Collector Problem

Example 128. Joe is collecting coupons, numbered $1, \ldots, k$. There are n different coupons, and when he gets all n of them, he gets a prize. Each time he gets a coupon, it is chosen uniformly at random (i.e., all coupons are equally likely).

Find a value m such that if Joe collects at least m coupons, then his probability of winning the prize is at least 99%.

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

$$V = \text{number of coupons Joe collects until he gets the prize}$$

and let

$$A_k = \text{coupon } k \text{ is not collected in the first } m \text{ coupons.}$$

To solve the problem, want to find a value of m sufficiently large that

$$\mathbb{P}[V \leq m] \geq .99.$$

Equivalently, we want

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

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

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 Joe collects at least $n\log(100n)$ coupons, then with probability at least 99%, he will get the prize.

For example, if $n = 15$, then $15\log(100 \cdot 15) = 109.7$ so $m = 110$ coupons is sufficient to ensure that Joe has 99% probability of getting the prize.

End of Example 128. □

28.2 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_a

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

$$\text{p-value} := \mathbb{P}\left[\text{observing a result at least as extreme as } X \text{ when } H_0 \text{ is true}\right]$$

When the p-value is small, we **reject** H_0.

In particular, we set a ***significance level***, denoted α, as the cutoff such that we reject H_0 is $p < \alpha$ and we fail to reject H_0 if $p \geq \alpha$.

A typical value is $\alpha = .05$ or $\alpha = .01$.

This is the basic story. Let's do an example:

Example 129. 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

$$\begin{aligned} p &= \mathbb{P}\left[\text{at least 8 heads when } \theta = \frac{1}{2}\right] \\ &= \binom{10}{8}\theta^8(1-\theta)^2 + \binom{10}{9}\theta^9(1-\theta)^1 + \binom{10}{10}\theta^{10}(1-\theta)^0 \text{ with } \theta = \frac{1}{2} \\ &= \binom{10}{8}\left(\frac{1}{2}\right)^{10} + \binom{10}{9}\left(\frac{1}{2}\right)^{10} + \binom{10}{10}\left(\frac{1}{2}\right)^{10} \\ &= \frac{7}{128} \\ &\approx 0.0546. \end{aligned}$$

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 129. □

29 2026-04-17 | Week 13 | Lecture 29

29.0.1 What is a hypothesis test?

Definition 130 (Hypothesis Testing)**.** Given a sample from $X_1, \ldots, X_n$ from a distribution depending on parameter θ. We are interested in comparing two competingk hypotheses: a null hypothesis H_0 and an alternative hypothesis H_a. A ***hypothesis test*** is a procedure or decision rule utilizing a statistic T which specifies for which values of T the hypothesis H_0 is rejected, and for which values H_0 is accepted. The statistic T is called the ***test statistic***.

Associated with a hypothesis test is a ***rejection region*** (RR), the set which specifies the values of the test statistic for which the null hypothesis to be rejected in favor of the alternative hypothesis. That is, if $T \in \text{RR}$, then we reject H_0. If $T \notin \text{RR}$, we accept H_0.

Example 131. 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

$$\begin{aligned} p &= \mathbb{P}\left[\text{at least 8 heads when } \theta = \frac{1}{2}\right] \\ &= \binom{10}{8}\theta^8(1-\theta)^2 + \binom{10}{9}\theta^9(1-\theta)^1 + \binom{10}{10}\theta^{10}(1-\theta)^0 \text{ with } \theta = \frac{1}{2} \\ &= \binom{10}{8}\left(\frac{1}{2}\right)^{10} + \binom{10}{9}\left(\frac{1}{2}\right)^{10} + \binom{10}{10}\left(\frac{1}{2}\right)^{10} \\ &= \frac{7}{128} \\ &\approx 0.0546. \end{aligned}$$

In words, if the coin were fair, we would expect it to flip 8 or more heads (out of ten) about 5.46% of the time.

In the previous example, the test statistic was $T =$ (number of heads). A reasonable rejection region might be $[T \geq 8]$ or $[T \geq 9]$.

End of Example 131. □

Example 132. 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 132. □

29.0.2 Type I and Type II errors

There are two types of decision errors in the hypothesis testing framework.

Definition 133 (Decision errors)**.** A ***type I error*** is made if H_0 is rejected when H_0 is true. The probability of a type I error is denoted by α:

$$\alpha = \Pr[\text{Type I error}] = \Pr[\text{rejecting } H_0 \text{ when } H_0 \text{ is true}]$$

The value of α is called the ***significance level*** (or just ***level***) of the test.

Similarly, a ***type II error*** is made if H_0 is accepted when H_a is true. We denote the probability of this happening by β:

$$\beta = \Pr[\text{Type II error}] = \Pr[\text{accepting } H_0 \text{ when } H_a \text{ is true}]$$

The value $1 - \beta$ is called the ***power*** of the test.

These types of errors, and their probabilities, can be summarized in the following table:

	Reject H_0	Accept H_0
H_0 true	Type I error level $= \alpha$	Correctly accept H_0 $1 - \alpha$
H_a true	Correctly reject H_0 power $= 1 - \beta$	Type II error β

30 2026-04-20 | Week 14 | Lecture 30

30.1 Decision errors

Example 134 (Coin flipping hypothesis test)**.** Flip a coin 10 times. Let p be the probability of heads. Consider the hypothesis testing problem

$$H_0 : p = \frac{1}{2} \quad \text{versus} \quad H_a : p \neq \frac{1}{2}.$$

Let our test statistic be

$$T = (\#\text{ of heads})$$

So T takes a value between 0 and 10. If H_0 were true, then (one can check that) there would be about 97.9% chance of getting between 2 and 8 heads. So it might be reasonable to define our rejection region to be the set

$$\text{RR} = \{0, 1, 9, 10\}$$

since it is improbable to get these values with a fair coin. Using this rejection region, if $T = 0, 1, 9$ or 10, we would reject H_0 and conclude the coin is not a fair coin.

This test is not perfect: assuming that H_0 is true, we will nonetheless reject H_0 about 2.1% of the time.

End of Example 134. □

30.1.1 Type I and Type II errors

There are two types of decision errors in the hypothesis testing framework.

	Reject H_0	Accept H_0
H_0 true	Type I error level $= \mathbb{P}[\text{type I error}] = \alpha$	Correctly accept H_0 $1 - \alpha$
H_a true	Correctly reject H_0 power $= 1 - \beta$	Type II error $\mathbb{P}[\text{type II error}] = \beta$

Example 135 (Example 134 continued)**.** In Example 134, the level of the test is $\alpha = 2.1\%$. Note that β is a function of p. For example if $p = 0.4$, then

$$\begin{aligned}
\beta &= \beta(0.4) \\
&= \mathbb{P}[\text{Type II error} \mid p = 0.4] \\
&= \mathbb{P}[T \notin \text{RR} \mid p = 0.4] \\
&= \sum_{k=2}^{8} \binom{10}{k} (0.4)^k (1 - 0.4)^{10-k} \\
&= .952
\end{aligned}$$

Therefore if the true value of p is .4, then the power of the the test is .048.

On the other hand if the true value of p is $p = .1$, then $\beta(.1) = .264$, so the power of the test would be .736. We generally don't know what the true power of the test is, since that would require knowing the value of the parameter (in which case we wouldn't be doing a hypothesis test...).

It is a common mistake to assume that $\alpha + \beta = 1$. In general, $\alpha + \beta \neq 1$.

End of Example 135. □

These terms are clarified by an example.

Remark 136 (Decision Errors)**.** A Ukrainian drone operator spots an unmarked vehicle traveling down a road. She needs to make a decision about whether to strike the truck. The hypothesis test is the following:

H_0 : the target is not a Russian military vehicle

H_a : the target is a Russian military vehicle

Rejecting H_0 means giving an order to destroy the target. Accepting H_0 means 'take no action'.

In this context, a Type-I error occurs when the operator gives the order to destroy the vehicle, but it turns out the vehicle is not Russian military. A Type-II error, on the other hand, has occurred if she doesn't order the strike, but the vehicle really was a Russian military vehicle.

- A Type-I error means she killed a civilian or a friendly unit (an error of commission).
- A Type-II error means she missed an opportunity to kill an enemy (an error of ommission).

Here, the terms "power" and "level" have the following meaning:

- The **power** of the test is the probability that she orders a strike when a Russian military vehicle goes by. (She wants this to be high).
- The **level** of the test is the probability that she orders a strike when a civilian goes by. (She wants this to be low).

Ideally, the operator would like to eliminate both types of errors (accidentally killing civilians vs failing to kill enemy). That is, she would like

$$\alpha = \Pr\left[\text{Type I error}\right] = 0 \quad \text{and} \quad \beta = \Pr\left[\text{Type II error}\right] = 0.$$

Unfortunately, it is impossible to simultaneously eliminate both types of error without perfect information. On one hand, the drone operatore could minimize the probability of a type I error by simply never ordering a strike, in which case $\alpha = \Pr\left[\text{Type I error}\right] = 0$. But this is unsatisfactory, since in that case, $\beta = \Pr\left[\text{Type II error}\right] = 1$. On the other hand, she could maximize the power by simply always ordering a strike; in that case, $\beta = 0$ but $\alpha = 1$. This tradeoff is present in every hypothesis test.

The approach that the hypothesis testing framework takes to address this tradeoff is to (1) set a maximal level α that one is willing to accept, and then (2) find the most powerful test subject to that constraint.

Of course, the appropriate significance level to set depends on the situation: a truck speeding towards the operator's position is very different from a truck traveling away from the front.

31 2026-04-22 | Week 14 | Lecture 31

31.1 Tests of independence

Based on chapters 14.2-14.4

Suppose we want to know if two things A and B are related to each other, or whether they are independent. For example, whether or not a person vapes (A) and whether they get coronary artery disease (B).

Or: whether a student attends class regularly (A) and whether they pass the class (B).

Question: How can we test this (using the hypothesis testing framework)?

Answer: We can utilize what are called "tests of independence", which we'll present in this lecture. For such tests, we require two main ingredients.

31.1.1 Ingedient 1: The chi-squared distribution

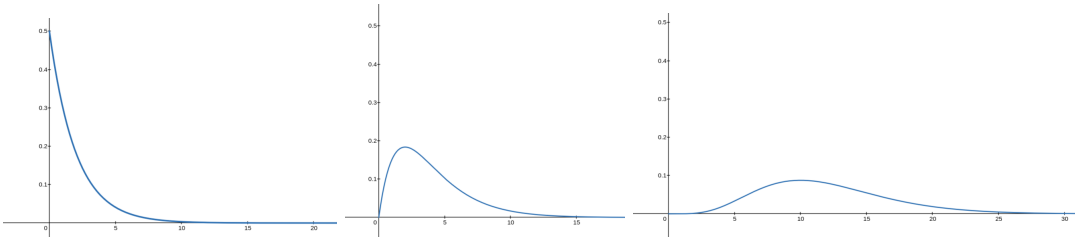
Definition 137 (Chi-Squared Distribution)**.** Let k be a positive integer. A nonnegative random variable X is said to have ***chi-squared distribution*** with d ***degrees of freedom*** if it has pdf

$$f(x) = \begin{cases} \frac{1}{2^{d/2}\Gamma(d/2)} x^{\frac{d}{2}-1} e^{-x/2} & : \quad x \geq 0 \\ 0 & : \quad x < 0 \end{cases}$$

.

Here, $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dx$. Note $\Gamma(k) = (k-1)!$ if k is a positive integer.

From left-to-right, here's what $f(x)$ looks like for $d = 2$, $d = 4$, and $d = 12$:



Proposition 138. *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.

31.1.2 Ingedient 2: The chi-squared statistic

Definition 139 (Chi-squared statistic)**.** Suppose we repeatedly conduct an experiment with k possible outcomes. The ***chi-squared statistic*** is defined as

$$\chi^2 := \sum_{i=1}^{k} \frac{(N_i - E_i)^2}{E_i}, \tag{21}$$

where $E_i = (\text{expected } \# \text{ of times outcome } i \text{ occurs})$, and $N_i = (\# \text{ of times outcome } i \text{ is observed})$.

Remark 140. The right-hand side of Eq. (21) has the form

$$\sum_{i=1}^{k} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

Remark 141. Under general conditions, the chi-squared statistic χ^2 has approximately a chi-squared distribution with degrees of freedom determined by the setting.

31.1.3 Example of a test of independence

Definition 142. 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.

Example 143 (Test of independence)**.** For example, suppose 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

(22)

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

$$H_a = \text{ there is a relationship between attendence and passing.}$$

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?
If attendence and passing are indpendent, then

$$\mathbb{E}\left[\#\text{ of students who attend regularly AND pass}\right] = npq = (54)\left(\frac{31}{54}\right)\left(\frac{33}{54}\right) = 18.9$$

This is what we expect if the null hypothesis is true. Compare this with the actual observed value, which was 25. The idea of the test, which we will finish in the next lecture, is the following:

If the observed values deviate a lot from what we'd expect under H_0, we should reject H_0.

Quantifying this will require computing a chi-squared statistic, which we'll do in the next lecture.

End of Example 143. □

32 2025-04-21 | Week 14 | Lecture 32

Chapter 14.1 through 14.3

We will finish the example of a test of independence from last time.

Example 144 (Test of independence)**.** For example, suppose 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

(23)

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

$$H_a = \text{ there is a relationship between attendence and passing.}$$

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 $n = 54$ 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

(24)

Interpretation: These are the values we would expect if H_0 were true. If the observed values deviate too much from these values, that's evidence against the null hypothesis.

Using the values recorded in Tables 23 and 24, we can compute the chi-squared statistic:

$$\begin{aligned}\chi^2 &= \sum_{\text{all table entries}} \frac{(\text{observed value} - \text{expected value})^2}{\text{expected value}} \\ &= \frac{(25-18.9)^2}{18.9} + \frac{(6-12.1)^2}{12.1} + \frac{(8-14.1)^2}{14.1} + \frac{(15-8.9)^2}{8.9} \\ &= 11.9\end{aligned}$$

By Remark 141, we know that χ^2 has a chi-squared distribution. When computing a chi-squared test statistic from a contingency table for a test of independence, the degrees of freedom is given by the equation

$$\text{df} = (\text{number of rows} - 1) \times (\text{number of columns} - 1)$$

which in our case, for a 2×2 table, is just 1.

Thereore, using Definition 137, the p-value is

$$\begin{aligned}p\text{-value} &= \mathbb{P}\left[\chi^2 \geq 11.9\right] \\ &= \int_{11.9}^{\infty} \frac{1}{\sqrt{2\pi}} x^{-\frac{1}{2}} e^{-x/2} dx \\ &= 0.0006\end{aligned}$$

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.

End of Example 144. □

32.1 Interpreting tests of independence

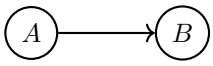
Suppose we want to know if two things A and B are related to each other. We will discuss only the case where A and B are discrete, categorical events. For example

$$A = [\text{a person vapes}]$$

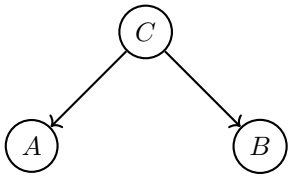
$$B = [\text{person gets coronary artery disease (CAD)}]$$

There are many possible relationships:

- **Causality:** Vaping causes arterial disease.

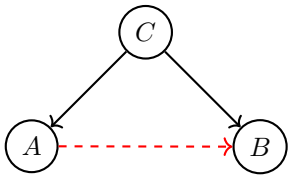


- **Common Response:** Perhaps the effect of vaping on your arteries is perfectly benign (no causal relationship), but CAD and vaping are still positively associated due to the effect of some third variable C that causes both to occur simultaneously:



For example we might have C = [person smokes cigarettes]. Maybe vapers are disproportionately likly to have smoked cigarettes, and perhaps and it's this subset of vapers that drives the association between vaping and CAD, as cigarette smoking is well-established to cause CAD. If this is true, then quitting vaping woudn't decrease your chance of CAD.

- **Confounding:** (A third variable jointly influences or causes both A and B). For example, suppose it were observed that vapers has 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 (CAD usually develops in older people). In that case, we might observe a spurious relationship between A and B due to the effects of some other third variable, C = [youth]:



It could nonetheless be the case that vaping increases your risk of CAD through some biological mechanism (the red dotted line), and due to the effect of age, vapers might *still* have lower incidence of CAD than the general population.

In order to rule this out, we'd need to *control* for the effect of age, perhaps by conducting a study on people from the same age group.

What we *can* say, however, is that if A and B do not have any relationship (causal or otherwise), then they are probabilistically independent. This is what tests of independence 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 much more information than that.

33 2025-05-01 | Week 15 | Lecture 33

33.1 Simple linear regression (i.e., finding a best-fit line)

Based on chapter 12.1 in the textbook

> ***The nexus question of this lecture:*** *How do we find the "best fit" line for data?*

Suppose we have four data points taking the form $(t, b) = (\text{time}, \text{observation})$:

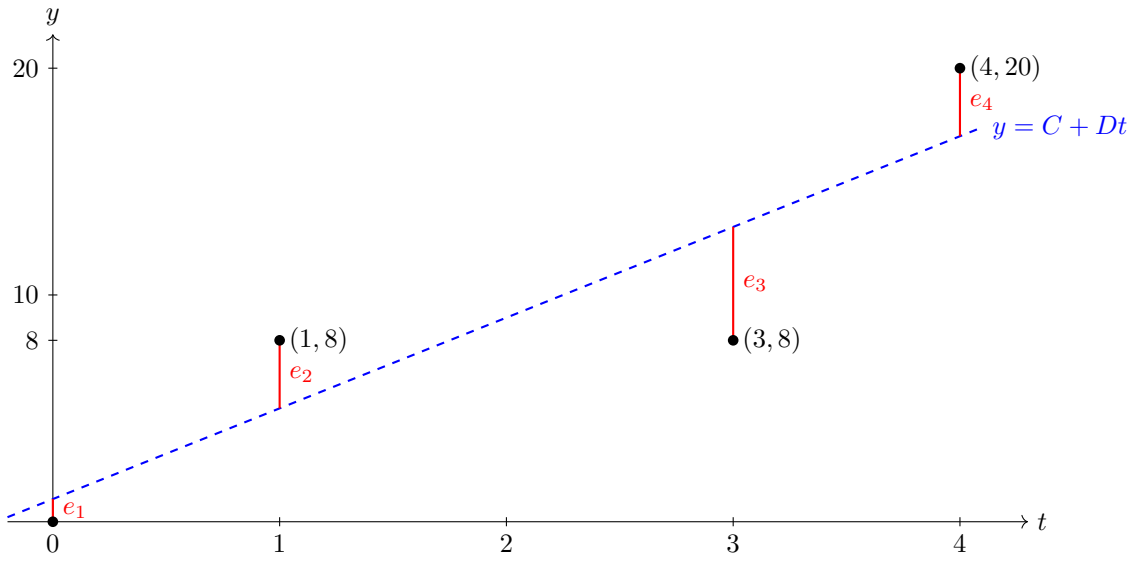
$$(t_1, b_1) = (0, 0)$$
$$(t_2, b_2) = (1, 8)$$
$$(t_3, b_3) = (3, 8)$$
$$(t_4, b_4) = (4, 20)$$

I have plotted these below. The equation of a line is

$$y = C + Dt.$$

Definition 145. The ***simple linear regression problem*** is to find values for C and D which best fit our data.

Here's a picture of the data points and a general line $y = C + Dt$ which aims to fit to the data:



Let $e_i = b_i - (C + Dt_i)$ for each $i = 1, \ldots, 4$. In words, e_i is obtained by measuring the vertical distance between the line $y = C + Dt$ and the i^{th} data point. For this example, we have

$$\mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{bmatrix} = \begin{bmatrix} 0 - C \\ 8 - (C + D) \\ 8 - (C + 3D) \\ 20 - (C + 4D) \end{bmatrix}$$

This vector is called the ***residual vector***, and it consists of the "errors" between

1. the observed data; and,
2. the values predicted by the line $y = C + Dt$.

We want all four errors $e_1, \ldots, e_4$ to be small. One way to measure the total error is to square the four errors and add them up:

$$\text{SSE} = e_1^2 + e_2^2 + e_3^2 + e_4^2 \tag{25}$$

This is called the ***sum of squared errors*** (SSE). Squaring has two advantages: it makes all the errors positive, and it penalizes large errors.

Therefore, to find the best-fitting line $y = C + Dt$, our goal will be to find values for C and D which minimize the sum of squared errors $e_1^2 + e_2^2 + e_3^2 + e_4^2$.

We'll present two approaches to finding the best values for C and D.

Approach #1: Calculus

Consider the function

$$\begin{aligned} f(C, D) &= e_1^2 + e_2^2 + e_3^2 + e_4^2 \\ &= C^2 + (8 - C - D)^2 + (8 - C - 3D)^2 + (20 - C - 4D)^2. \end{aligned}$$

We can use calculus to minimize f. To find the critical point of f, we set $\nabla f = 0$ and solve for C and D:

$$0 \overset{set}{=} \frac{\partial f}{\partial C} = 2C - 2(8 - C - D) - 2(8 - C - 3D) - 2(20 - C - 4D) = -72 + 8C + 16D$$

and

$$0 \overset{set}{=} \frac{\partial f}{\partial D} = -2(8 - C - D) - 6(8 - C - 3D) - 8(20 - C - 4D) = -224 + 16C + 52D.$$

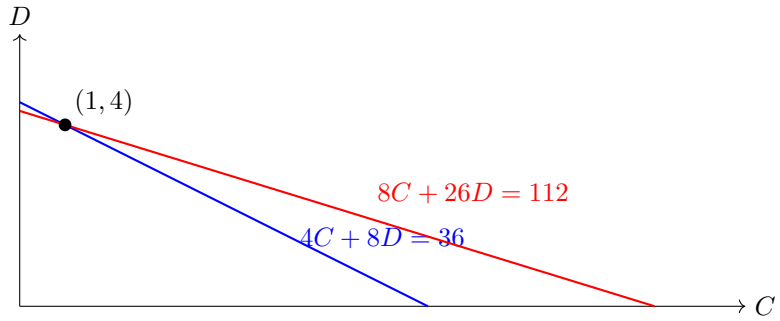
Combining these equations gives

$$\begin{cases} 8C + 16D = 72 \\ 16C + 52D = 224 \end{cases}.$$

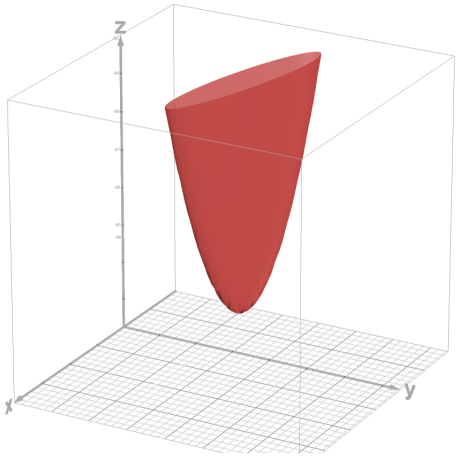
After dividing both equations by 2, we obtain the equivalent system

$$\begin{cases} 4C + 8D = 36 \\ 8C + 26D = 112 \end{cases}. \tag{26}$$

This system corresponds to two lines which intersect at the point $(C, D) = (1, 4)$ (shown below) which is the critical point.



We need to check that $(C, D) = (1, 4)$ is a minimum. We could show this with the second derivative test, but there is also a nice geometric reason why it must be a minimum. In particular, observe that f is a sum of squares, so it is bounded below by zero. By considering cross-sections (i.e., freeze one of either C or D and vary the other), we get an upward-opening parabola in any direction. So the garph is an upward-opening elliptic paraboloid; i.e., a bowl-shape, as shown:



Hence f has exactly one critical point and the critical point is the global minimum.

Therefore f is minimized at $(C, D) = (1, 4)$. Hence, the best fitting line for our data is

$$y = 1 + 4t,$$

as this line, out of all possible lines, minimizes the sum of squared errors $e_1^2 + e_2^2 + e_3^2 + e_4^2$.

Approach #2: Linear Algebra

There is a compact way to organize these computations using matrices, which allows us to compute the best fit line without taking any derivatives.

To motivate this, observe that ideally our best-fit line would pass through all the points, which occurs precisely when C and D satisfy the following system of equations:

$$\begin{cases} C + 0 \cdot D = 0 \\ C + 1 \cdot D = 8 \\ C + 3 \cdot D = 8 \\ C + 4 \cdot D = 20 \end{cases}$$

This is a system of 4 linear equations in 2 unknowns (C, D). We can write it in matrix form as

$$\underbrace{\begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} C \\ D \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} 0 \\ 8 \\ 8 \\ 20 \end{bmatrix}}_{b}$$

or more compactly as

$$Ax = b.$$

The bad news is that there is no choice of C and D for which this equation is true (because the data points **do not lie on a line**).

So instead we want Ax be as close as possible to b; that is

$$Ax \approx b.$$

Since $\mathbf{e} = b - Ax$, we want

$$\mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{bmatrix} \approx 0.$$

So we want the vector $\mathbf{e}$ to be as tiny as possible. By minimizing $e_1^2 + e_2^2 + e_3^2 + e_4^2$, we will also simultaneously minimize $\sqrt{e_1^2 + e_2^2 + e_3^2 + e_4^2}$, which is the length of $\mathbf{e}$. This gives a geometric intepretation of our goal:

finding the values of C and D which give the best fit line $y = C + Dt$ to our data is equivalent to finding the values of C and D which **minimize the the length of the residual vector e**.

To do this, we multiply both sides of $Ax = b$ by

$$A^\top = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 3 & 4 \end{bmatrix}.$$

This yields the following system of equations, called the ***normal equations***:

$$A^\top Ax = A^\top b. \tag{27}$$

Writing out Eq. (27), we get

$$\begin{bmatrix} 4 & 8 \\ 8 & 26 \end{bmatrix} \begin{bmatrix} C \\ D \end{bmatrix} = \begin{bmatrix} 36 \\ 112 \end{bmatrix}$$

or equivalently,

$$\begin{cases} 4C + 8D = 36 \\ 8C + 26D = 112 \end{cases}$$

This system is the normal equations. Observe that it is the same as Eq. (26) and has the same solution: $(C, D) = (1, 4)$, which is the minimizer of $\|\mathbf{e}\|^2 = e_1^2 + \ldots + e_4^2$.

34 2026-05-04 | Week 16 | Lecture 34

Based on chapters 12.1 and 12.2 in the textbook

Suppose we are given data (t_i, b_i), $i = 1, \ldots, n$. The ***simple linear regression problem*** is to find the line $y = C + Dt$ which best fits the data.

In this problem, we understand the "line of best fit" to the be the line $y = C + Dt$ in which C and D are chosen such that the ***sum of squared errors***

$$e_1^2 + \ldots + e_n^2$$

is minimized. Here, the e_i's the the entries of the ***residual vector*** e, defined as:

$$\mathbf{e} = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix} = b - Ax$$

where

$$b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} \quad \text{and} \quad A = \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_n \end{bmatrix}$$

In the last lecture, we saw that minimizing the sum of squared errors is equivalent to minimizing the length of the vector $\mathbf{e}$.

We also saw two appraoches to finding the best values for C and D. The first approach involved calculus. The second approach used linear algebra, which which what we'll focus on today. Recall the following definition:

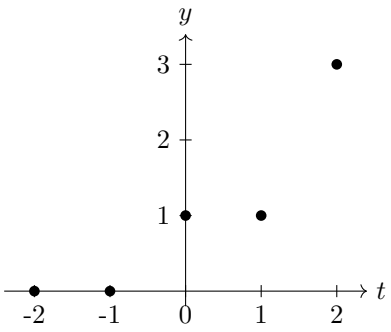
Definition 146 (Normal equations)**.** The ***normal equations*** are

$$A^\top Ax = A^\top b.$$

The linear algebra approach is easy to summarize: to find the best fit line $y = C + Dt$, set up the equation $A^\top Ax = A^\top b$ and solve it for $x = \begin{bmatrix} C \\ D \end{bmatrix}$.

Example 147. Suppose we have the following data:

t	y
-2	0
-1	0
0	1
1	1
2	3



Question: Find the best-fitting line $y = C + Dt$.

Solution: Consider the equation

$$Ax = b$$

where

$$A = \begin{bmatrix} 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}, \quad x = \begin{bmatrix} C \\ D \end{bmatrix}, \quad \text{and} \quad b = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 3 \end{bmatrix}.$$

Since the data points don't lie on a line, there is no solution (i.e., no values of C and C which make this equation true). So instead, we set up the normal equations

$$A^\top Ax = A^\top b$$

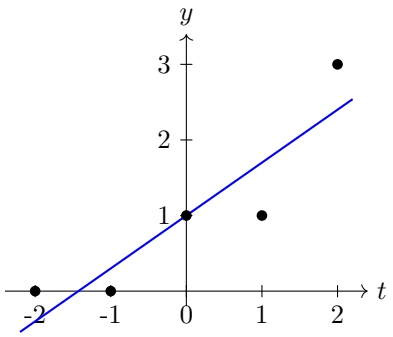
which are

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ -2 & -1 & 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} C \\ D \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ -2 & -1 & 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 3 \end{bmatrix}.$$

This simplifies to

$$\begin{bmatrix} 5 & 0 \\ 0 & 10 \end{bmatrix} \begin{bmatrix} C \\ D \end{bmatrix} = \begin{bmatrix} 5 \\ 7 \end{bmatrix}$$

From this we conclude that $C = 1$ and $D = 7/10$. Hence, our best-fitting line is $y = 1 + .7t$.



End of Example 147. □

Example 148. Question: Find the best fitting line to the data

t	y
0	1
1	2
2	2
3	4
4	5

Solution: We have

$$A = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 2 \\ 2 \\ 4 \\ 5 \end{bmatrix} \quad \text{and} \quad x = \begin{bmatrix} C \\ D \end{bmatrix}.$$

Therefore the normal equations $A^\top Ax = A^\top b$ are

$$\begin{bmatrix} 5 & 10 \\ 10 & 30 \end{bmatrix} \begin{bmatrix} C \\ D \end{bmatrix} = \begin{bmatrix} 14 \\ 38 \end{bmatrix}$$

or equivalently

$$\begin{cases} 5C + 10D = 14 \\ 10C + 30D = 38 \end{cases}$$

Solving this system gives $C = .8$ and $D = 1$. Therefore the best-fit line $y = C + Dt$ is

$$y = .8 + t.$$

End of Example 148. □

Theorem 149. *The values C and D which minimize $e_1^2 + \ldots + e_n^2$, are given by the solution $x = \begin{bmatrix} C \\ D \end{bmatrix}$ to the normal equations*

$$A^\top Ax = A^\top b,$$

where

$$A = \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_n \end{bmatrix} \quad \textit{and} \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

Proof. I tried to explain this geometrically. If x is the solution to the normal equations, then $e = b - Ax$ is orthogonal to $\mathrm{CS}(A)$, which guarantees that the the length of the residual vector is minimized. □