Lecture Notes for Math 372: Elementary Probability and Statistics

Last updated: November 4, 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 **4**

1 2026-08-24 | Week 01 | Lecture 01 **5**

2 2026-08-26 | Week 01 | Lecture 02 **6**
2.1 Set theory . . . 6
2.1.1 Definitions and set-builder notation . . . 6
2.1.2 Set operations . . . 7

3 2026-08-28 | Week 01 | Lecture 03 **9**
3.1 A general framework for probability . . . 9
3.2 Random variable . . . 10
3.3 The enumeration principle . . . 11

4 2026-08-31 | Week 02 | Lecture 04 **12**
4.1 Recap and coin flipping example . . . 12
4.2 What is probability? The three axioms . . . 12
4.3 Basic properties of probability functions . . . 13

5 2026-09-02 | Week 02 | Lecture 05 **15**
5.1 The multiplication principle . . . 15
5.2 Permutations . . . 16
5.3 Combinations . . . 17

6 2026-09-04 | Week 02 | Lecture 06 **18**
6.1 Examples of counting principles . . . 18

7 2026-09-09 | Week 03 | Lecture 07 **20**
7.1 Introduction to craps . . . 20
7.2 Conditional probability . . . 20

8 2026-09-11 | Week 03 | Lecture 08 **22**
8.1 Conditional probability examples . . . 22

9 2026-09-14 | Week 04 | Lecture 09 **24**
9.1 The law of total probability . . . 24

10 2026-09-16 | Week 04 | Lecture 10 **26**
10.1 Conditional probability example: diagnostic tests . . . 26
10.2 Independence . . . 27
10.3 Examples of using independence . . . 27

11 2026-09-18 | Week 04 | Lecture 11 **28**
11.1 The "at least one" trick . . . 28
11.2 Dicrete Random Variables . . . 28

12 2026-09-21 | Week 05 | Lecture 12 **31**

13 2026-09-23 | Week 05 | Lecture 13 **33**
13.1 Expected value examples . . . 33
13.2 Law of the unconscious statistician . . . 34

14 2026-09-25 | Week 05 | Lecture 14 **36**
14.1 The "spread" of a distribution . . . 36
14.2 Variance . . . 36
14.3 Variance of the binomial distribution . . . 37

15 2026-09-28 | Week 06 | Lecture 15 **39**
15.1 Motivation . . . 39
15.2 The Poisson distribution . . . 39

16 2026-09-30 | Week 06 | Lecture 16 **41**

17 2026-10-02 | Week 06 | Lecture 17 **42**

18 2026-10-04 | Week 07 | Lecture 18 **43**
18.1 The probability of winning at craps . . . 43
18.2 Continuous random variables . . . 45

19 2026-10-07 | Week 07 | Lecture 19 **47**

20 2026-10-09 | Week 07 | Lecture 20 **48**

21 2026-10-12 | Week 08 | Lecture 21 **50**
21.1 Normal random variables . . . 50

22 2026-10-14 | Week 08 | Lecture 22 **52**
22.1 Joint distributions - discrete case . . . 52

23 2026-10-16 | Week 08 | Lecture 23 **55**
23.1 Joint distributions - continuous case . . . 55

24 2026-10-19 | Week 09 | Lecture 24 **56**

25 2026-10-21 | Week 09 | Lecture 25 **57**
25.1 LOTUS Theorem - multivariate case . . . 57
25.2 Covariance . . . 57

26 2026-10-26 | Week 10 | Lecture 26 **60**
26.1 Statistics: a general framework . . . 60
26.2 What is a statistic? . . . 60

27 2026-10-28 | Week 10 | Lecture 27 **62**
27.1 The central limit theorem . . . 62

28 2026-10-30 | Week 10 | Lecture 28 **65**
28.1 Central Limit Theorem Example 65
28.2 Introduction to point estimation 65

29 2026-11-02 | Week 11 | Lecture 29 **67**
29.1 Examples of Point Estimators 67
29.2 The Law of Large Numbers 68

30 2026-11-04 | Week 11 | Lecture 30 **70**
30.1 Confidence intervals 70

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 1-3:** 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 4-5:** Discrete random variables (1.5 weeks)
  - Expected values, mean, variance, binomial distribution, Poisson distribution. Moment generating functions. (Sections: 3.1-3.6)
- **End of week 6: Midterm 1 (Friday, October 2)**
- **Weeks 6-7:** Continuous Random variables (1.5 weeks)
  - Uniform, exponential, gamma, and normal distributions. Intuitive treatment of the Poisson process and development of the relationship with gamma distributions. (Sections: 4.1-4.4)
- **Weeks 8-9:** 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)
- **End of week 9: Midterm 2 (Tentatively: Friday, October 23)**
- **Week 10:** Introduction to statistical estimation
  - Point and confidence interval estimation. Maximum likelihood, optimal, and unbiased estimators. Examples. (Sections 6.1, 6.2)
- **Weeks 11-12:** 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: Wednesday, December 9**
- **Final Exam: December 18**

1 2026-08-24 | Week 01 | Lecture 01

- give syllabus
- did the pirates worksheet

2 2026-08-26 | Week 01 | Lecture 02

> ***The nexus question of this lecture:*** *What is a set and and what are the basic set operations?*
> ***Reading assignment:*** *chapter 2.1*

2.1 Set theory

The mathematical formalization of modern probability theory rests on the foundation of set theory.

2.1.1 Definitions and set-builder notation

- A ***set*** is an unordered collection of distinct objects. The objects of a set are called ***elements***. We write the elements of a set inside curly braces, $\{\ldots\}$.

  Examples: $\{1, 4, 2, 3\}$, $\{\text{cat}, \text{fish}, 5\}$, $\mathbf{Z} := \{\ldots, -2, -1, 0, 1, 2, \ldots\}$, $\mathbf{R} = \{\text{the real numbers}\}$

- Two sets A and B are ***equal***, denoted $A = B$, if they share precisely the same elements. The order that the elements are written is irrelevant.

  $\{1, 2, 3\} = \{3, 1, 2\}$.

  Repeated elements are ignored:

  $\{1, 2, 3\} = \{1, 1, 2, 3\}$.

- A set is ***finite*** if it contains a finite number of distinct elements. This number is called the ***size*** of the set, and is denoted $|A|$.

  Example: $|\{\text{cat}, 5\}| = 2$.

- The ***empty set*** is the set which contains no elements. This is denoted as

  $$\emptyset \quad \text{or} \quad \{\}$$

- ***Set-builder notation*** describes a set by specifying the properties that its elements must satisfy.

  Examples:

  The set of integers between 1 and 36:

  $$\{n \in \mathbb{Z} : 1 \leq n \leq 36\}$$

  The set of ordered pairs (x, y) such that x and y are both positive real numbers:

  $$\{(x, y) \in \mathbb{R}^2 : x, y > 0\}$$

  The set of even integers:

  $$\{n \in \mathbb{Z} : n \text{ is even}\}$$

  The set of ordered pairs x, y such that both x and y are integers beween 1 and 4:

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

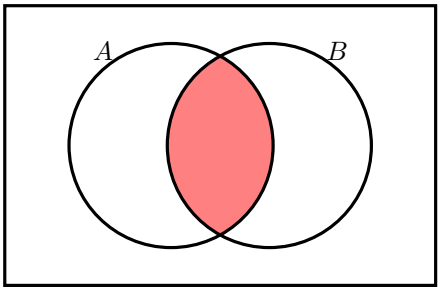
  In set builder notation, the colon ':' is read as "such that" and the symbol '$\in$' (which looks like an E) is read as "is an element of"

2.1.2 Set operations

Let A, B be sets.

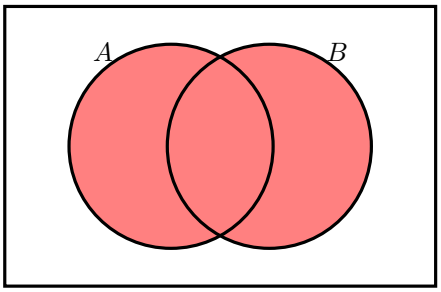
- The ***intersection*** of A and B, denoted $A \cap B$, is defined by

$$A \cap B = \{x : x \in A \text{ and } x \in B\}.$$



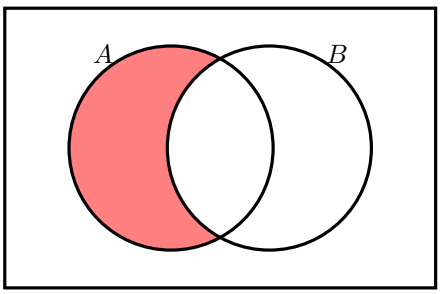
- The ***union*** of A and B, denoted $A \cup B$, is defined by

$$A \cup B = \{x : x \in A \text{ or } x \in B\}.$$



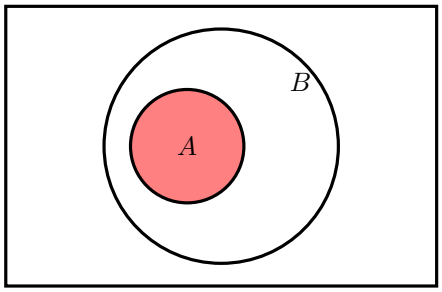
- The ***set difference*** of A and B, denoted $A \setminus B$, is

$$A \setminus B = \{x : x \in A \text{ and } x \notin B\}.$$



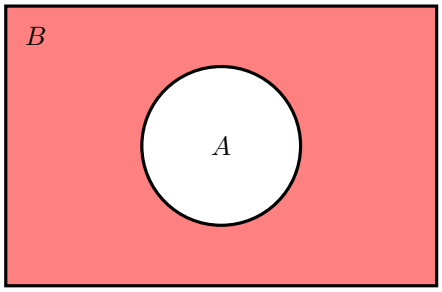
- We say that A is a ***subset*** of B, and write $A \subseteq B$, if

$$x \in A \implies x \in B.$$

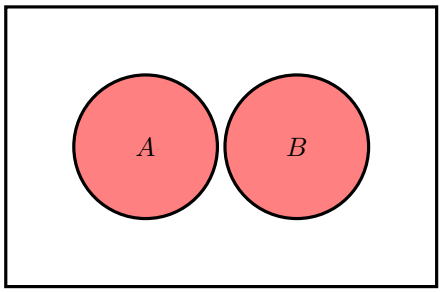


- If $A \subseteq B$, the ***complement of A in B***, denoted A^c, is

$$A^c = B \setminus A = \{x : x \in B \text{ and } x \notin A\}.$$



- We say that A and B are ***disjoint*** if $A \cap B = \emptyset$.



3 2026-08-28 | Week 01 | Lecture 03

> ***The nexus question of this lecture:*** *What does set theory have to do with probability?*
> ***Reading assignment:*** *finish chapter 2.1, start 2.2*

3.1 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, asking your crush on a date, 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***. Each sample point represents a **unique** outcome of the experiment.

  Roll a dice. There are six outcomes:

  $$1, 2, 3, 4, 5, 6$$

  These are the sample points. The sample space is

  $$S = \{1, 2, 3, 4, 5, 6\}.$$

- In probability, we refer to subsets of the sample space as ***events***.

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

  We represent events using brackets $[\ldots]$, and usually denote them with capital letters.

  Examples of events when rolling a dice are:

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

- The first three events shown are ***singleton*** sets, containing only one sample point. The book calls these ***simple events***. But note

  $$1 \neq \{1\}$$

- The empty set $\emptyset$ and the whole sample space S are always both events:
  - S is the event that an outcome occurs ("something happens").
  - $\emptyset$ is the event that no outcome occurs. This is called the ***null event.***

- Two events are E, F ***mutually exclusive*** if $E \cap F = \emptyset$ (i.e., the sets E, F are disjoint). This means that the events E and F cannot both happen at the same time.

  The events A, B defined above are mutually exclusive, since the dice roll cannot be both even and odd.

- Translating set operations into the language of events,

- $E \cap F$ is the event that both E **and** F occur
- $E \cup F$ is the event that at least one of the events E **or** F occurs
- $E^c = S \backslash E$ is the event that E does **not** occur.

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

- 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. □

3.2 Random variable

We often associate numbers with outcomes of an experiment.

Definition 2 (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 the S. Formally, it is a function whose domain is S and whose range is $\mathbb{R}$.

Most interesting events are defined using random variables.

Example 3 (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**. We can enumerate the sample space as a table:

		Dice 2			
		1	2	3	4
Dice 1	1	(1, 1)	(1, 2)	(1, 3)	(1, 4)
	2	(2, 1)	(2, 2)	(2, 3)	(2, 4)
	3	(3, 1)	(3, 2)	(3, 3)	(3, 4)
	4	(4, 1)	(4, 2)	(3, 4)	(4, 4)

Let X be the sum of the two rolls. This is an example of a **random variable**. 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

We can define events using X. For example

$$\begin{aligned}
[X \leq 3] &= \{(x,y) \in S : x+y \leq 3\} = \{(1,1),(1,2),(2,1)\} \\
[X = 4] &= \{(x,y) \in S : x+y = 4\} = \{(1,3),(2,2),(3,1)\} \\
[X \leq 4] &= [X \leq 3] \cup [X = 4] = \{(1,1),(1,2),(2,1),(1,3),(2,2),(3,1)\} \\
[X \text{ is even}] &= \{(1,1),(1,3),(2,2),(2,4),(3,1),(3,3),(4,2),(4,4)\}. \\
&\text{etc.}
\end{aligned}$$

End of Example 3. □

3.3 The enumeration principle

On Monday's worksheet, I had you do a bunch of problems where you computed probabilities by counting outcomes. This is a smiple but critically important idea, called the ***enumeration principle***.

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

So using X from Example 44,

$$\mathbb{P}[X = 4] = \frac{3}{16}$$

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

4 2026-08-31 | Week 02 | Lecture 04

The nexus question of this lecture: *What is probability?*
Reading assignment: *finish chapter 2.2, start 2.3*

4.1 Recap and coin flipping example

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 (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*. There is no upper bound on how many times we might have to flip the coin, so the sample space is

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

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

$$\mathbb{P}[X = 1] = \mathbb{P}\left[1^{\text{st}} \text{ flip heads}\right] = \frac{1}{2}$$

and

$$\mathbb{P}[X = 2] = \mathbb{P}\left[1^{\text{st}} \text{ flip tails, } 2^{\text{nd}} \text{ flip heads}\right] = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}.$$

Similarly,

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

End of Example 4. □

We will need one more definition.

Definition 5 (Pairwise disjoint)**.** A sequence of events $(A_1, A_2, \ldots)$ is ***pairwise disjoint*** if

$$A_i \cap A_j = \emptyset \quad \text{whenever } i \neq j.$$

In other words, no two events have any "overlap". Equivalently, no two events can simultaneously occur.

4.2 What is probability? The three axioms

We denote the probability of an event E by $\mathbb{P}[E]$ or $\mathbb{P}(E)$. But what is "probability"? And what is $\mathbb{P}$?!

Definition 6 (Probability Axioms)**.** Let S be a sample space associated with an experiment. A function $\mathbb{P}$ is called a ***probability function*** on S if it satisfies the following three axioms:

A.1 (Nonnegativity) $\mathbb{P}[E] \geq 0$ for every event $E \subseteq S$.

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

A.3 (Countable additivity) If $A_1, A_2, \ldots$ is an infinite sequence of pairwise disjoint events then

$$\mathbb{P}[A_1 \cup A_2 \cup \ldots] = \mathbb{P}[A_1] + \mathbb{P}[A_2] + \ldots.$$

If $\mathbb{P}$ is a probabilty function, then for every event $E \subseteq S$, the number $\mathbb{P}[E]$ is called the ***probability*** of E.

The above definition may feel unsatisfying becuase it only tells us the conditions that probabilities must satisfy; it doesn't tell us how to assign specific probabilities to events. But before judging this definition, we should first expore the implications of these axioms.

The first two axioms, **A.1** and **A.2** are self-evident. The last one, **A.3**, is more technical and harder to deciper. The next example gives an application of axiom **A.3**.

Example 7 (Continuation of Example 4)**.** What is the probability that we flip the coin an even number of times? That is, what is $\mathbb{P}[X \text{ is even}]$?

Define

$$E_n := [X = n], \qquad n \in \{1, 2, \ldots\}$$

Then

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

The sequence $E_2, E_4, E_6 \ldots$ is pairwise disjoint. Therefore

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

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

End of Example 7. □

4.3 Basic properties of probability functions

The next theorem enumerates some basic properties that can be derived using only the three axioms.

Proposition 8 (Basic properties of probabilities)**.**

(i.) *(The probability of no outcome occurring is 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*

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

"The probability that you get hit by a non-red car is the probability that you get hit by a car minus the probability that you get hit by a red car."

(v.) *("The particular is less likely than the general") If A, B are events and $A \subseteq B$, then $\mathbb{P}[A] \leq \mathbb{P}[B]$*
"The probability that you get hit by a car is higher than the probability that you get hit by a red car."

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

Proof of Proposition 8.

- First we will prove (i.). Let $E_i = \emptyset$ for all $i = 1, 2, 3, ...$ Then $(E_i)_{i=1}^{\infty}$ is a pairwise disjoint sequence, and $E_1 \cup E_2 \cup \ldots = \emptyset$. Therefore,

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

  The equation $\mathbb{P}[\emptyset] = (\mathbb{P}[\emptyset] + \mathbb{P}[\emptyset] + \ldots)$ forces $\mathbb{P}[\emptyset] = 0$. This proves (i.).

- Next we prove (ii.). Let $E_1, \ldots, E_n$ be a finite sequence of events which are pairwise disjoint. Then expand the sequence by taking $E_i = \emptyset$ for every positive integer i greater than n. Then

$$\begin{aligned} \mathbb{P}\left[\bigcup_{i=1}^{n} E_i\right] &= \mathbb{P}\left[\bigcup_{i=1}^{\infty} E_i\right] \\ &= \sum_{i=1}^{\infty} \mathbb{P}[E_i] && \text{by } \mathbf{A.3} \\ &= \mathbb{P}[E_1] + \ldots + \mathbb{P}[E_n] + \underbrace{\mathbb{P}[\emptyset] + \mathbb{P}[\emptyset] + \ldots}_{=0 \text{ by (i.)}} \\ &= \mathbb{P}[E_1] + \ldots \mathbb{P}[E_n]. \end{aligned}$$

  This proves (ii.).

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

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

  This proves (iii.).

- Next we prove (iv.). Since $A \subseteq B$,

$$B = A \cup (B \cap A^c).$$

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

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

  Rearranging terms proves (iv.).

- Next we prove (v.). Assume $A \subseteq B$. Then by (iv.),

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

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

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

  and this implies (v.).

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

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

□

5 2026-09-02 | Week 02 | Lecture 05

The topic for this lecture: *Combinatorial principles*
Reading assignment: *chapter 2.3*

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

- multiplication principle
- permutation principle
- combination principle

5.1 The multiplication principle

Notation 9. Denote the set consisting of the first n positive integers by

$$[n] = \{1, \ldots, n\}.$$

Definition 10 (Tuple)**.** Let k be a positive integer. A ***k-tuple*** is an ordered list

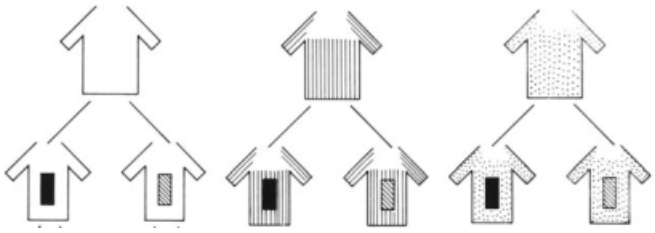
$$(a_1, a_2, \ldots, a_k)$$

of k objects. We call a_i the i^{th} ***coordinate*** of the tuple.

Remark 11. The word *ordered* is key. $(1, 2)$ and $(2, 1)$ are different 2-tuples, and $(1, 1, 2)$ is a perfectly fine 3-tuple even thought it repeats an entry.

Theorem 12 (The multiplication principle)**.** *Suppose an object is built by making k choices in order, where the number of options for the 1^{st} choice is n_1, the number of options for the 2^{nd} is n_2, and so forth. Then the number of objects that can be built is $n_1 n_2 \cdots n_k$.*

Example 13. A man has 3 shirts and 2 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". Equivalently,

$$\{(a_1, a_2) : a_1 \in [3], a_2 \in [2]\}.$$

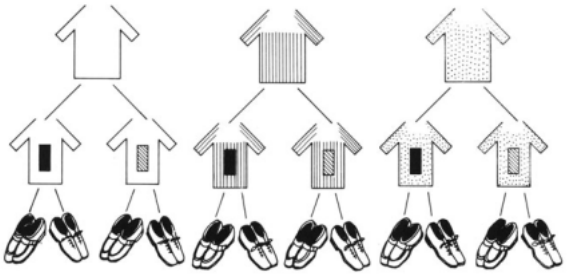
The total number of outfits is $3 \times 2 = 6$.

If the man also has 2 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

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

By Theorem 12, the number of elements in this set is $3 \times 2 \times 2 = 12$. Thus the number of ways he can dress himself 12.

This is shown in the following tree diagram:



End of Example 13. □

Corollary 14. $|\{(a_1, a_2, \ldots, a_k) : a_1 \in [n_1],\ a_2 \in [n_2], \ldots, a_k \in [n_k]\}| = n_1 n_2 \cdots n_k$.

Proof. Follows immediately by Theorem 12 □

5.2 Permutations

Example 15. Suppose I have a set of 6 objects:

$$\{\text{coin}, \text{card}, \text{pen}, \text{clip}, \text{craine}, \text{dice}\}$$

Question 1. How many ways can I order these six objects?
Answer: $6! = 6 \times 5 \times 4 \times 3 \times 2 \times 1$.

Question 2. How many ways can I arrange just four elements of S?
Answer: $6! = 6 \times 5 \times 4 \times 3$.

End of Example 15. □

In the remainder, we will formalize these ideas.

Definition 16 (Permutation)**.** Given a set S, a ***permutation*** of S is an ordering of all the elements of S.

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 $0! = 1$.)

Definition 18 (Permutation of size k)**.** Given a set S, a ***permutation of size k*** is an ordered subset of S containing exactly k elements.

Theorem 19 (The permutation principle)**.** *Let $P_{k,n}$ be the number of permutations of size k that can be formed from a set of n objects. Then*

$$P_{k,n} = n \cdot (n-1) \cdot (n-2) \cdots (n-k+1) = \frac{n!}{(n-k)!}$$

(Note: the right-hand side is a product of k terms.)

Proof. Build the list one entry at a time. There are n choices for the first entry, then $n-1$ for the second (since one element was used up). Then $n-2$, and so on, with $n-k+1$ choices for the k^{th} entry. By Theorem 12, the number of choices multiply, giving the formula. □

5.3 Combinations

Definition 20 (Combination of size k)**.** Given a set S, a ***combination of size*** k is a subset of S containing exactly k elements.

Theorem 21 (The combination principle)**.** *Let S be a set with n elements. The number of combinations of size k is*

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

Proof. Let $C_{k,n}$ be the number of subsets of S of size k. Observe that for every unordered subset of size k, there are $k!$ ways to order it by Proposition 17. Hence,

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

Changing notation,

$$k!\, C_{k,n} = P_{k,n}.$$

Therefore

$$C_{k,n} = \frac{P_{k,n}}{k!}$$

By the formula from Theorem 19,

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

□

6 2026-09-04 | Week 02 | Lecture 06

6.1 Examples of counting principles

Example 22. 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. By the permutation principle, this number is

$$P_{4,32} = 32 \cdot 31 \cdot 30 \cdot 29.$$

(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 $\binom{17}{3}$.

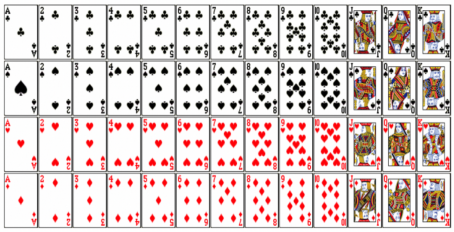
By the multiplication principle, there are

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

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

End of Example 22. □

Example 23 (Poker)**.** A poker hand consist of 5 randomly chosen cards from a standard 52-card deck.



(a) **Question:** How many distinct poker hands are there?

Solution:

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

(b) **Question:** A **flush** occurs when all five cards are the same suit. How many ways are there to get a flush?

Solution:

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

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

(c) **Question:** What's the probability of getting a flush?

Solution: 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 23. □

7 2026-09-09 | Week 03 | Lecture 07

The topic for this lecture: *Conditional probability*
Reading assignment: *Chapter 2.4*

7.1 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 from the in-class activity, we saw that shooters won about 56% of the time. Is this a fair game? Does the shooter have an advantage? 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:

7.2 Conditional probability

This section is based on section 2.4 in the text.

Definition 24 (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: $\mathbb{P}[B \mid A]$ is the probability of B when we know that event A happened.

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

Definition 25. 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 26 (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 24,

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

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 26. □

Example 27 (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 27. □

8 2026-09-11 | Week 03 | Lecture 08

The topic for this lecture: *Conditional probability*
Reading assignment: *Chapter 2.4*

8.1 Conditional probability examples

Example 28 (Urn - example of multiplication rule)**.** An urn contains 6 white balls and 9 black balls. If 4 balls are drawn at random, what is the probability that the first 2 are white and the last 2 are black?
Solution: 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. (23),

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

End of Example 28. □

Example 29 (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 29. □

Example 30. Suppose two cards are drawn at random from a standard 52 card playing deck. Let

$$A = [\text{at least one ace was drawn}]$$
$$B = [\text{both cards drawn are aces}]$$
$$C = [\text{the ace of clubs was drawn}]$$

Are $\mathbb{P}(B \mid A)$ and $\mathbb{P}(B \mid C)$ the same?

End of Example 30. □

9 2026-09-14 | Week 04 | Lecture 09

The topic for this lecture: *The Law of Total Probability*
Reading assignment: *Chapter 2.4, start chapter 2.5*

9.1 The law of total probability

Definition 31. Let S be the sample space of some experiment, and let $A_1, \ldots, A_n \subseteq S$ be events. We say that $A_1, \ldots, A_n$ ***partition*** S if

(i) $A_1 \cup A_2 \cup \cdots \cup A_n = S$; and

(ii) $A_i \cap A_j = \varnothing$ whenever $i \neq j$.

Remark 32. Condition (i) says that at least one of the events $A_1, \ldots, A_n$ occurs. Condition (ii) says that at most one of them occurs.

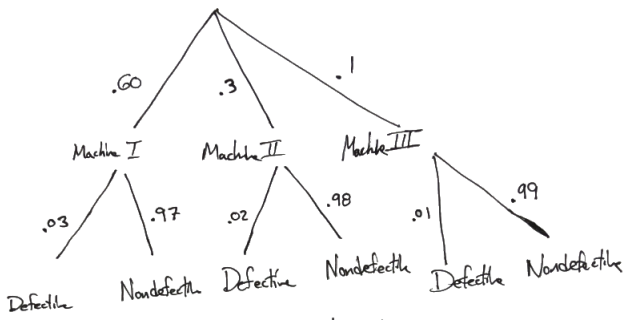
Theorem 33 (Law of Total Probability)**.** *Let $A_1, \ldots, A_n$ be events which partition their sample space. Then for any event B,*

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

Example 34. 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 $\mathbb{P}[M_3 \mid D] = \mathbb{P}[\text{machine III} \mid \text{defective}]$.

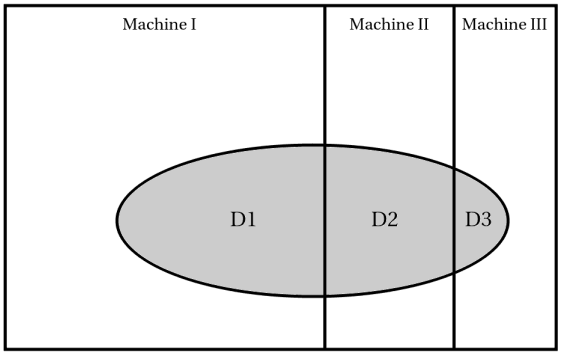
$$\mathbb{P}[M_3 \mid D] = \frac{\mathbb{P}[M_3 \cap D]}{\mathbb{P}[D]} \tag{3}$$

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. (3),

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

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

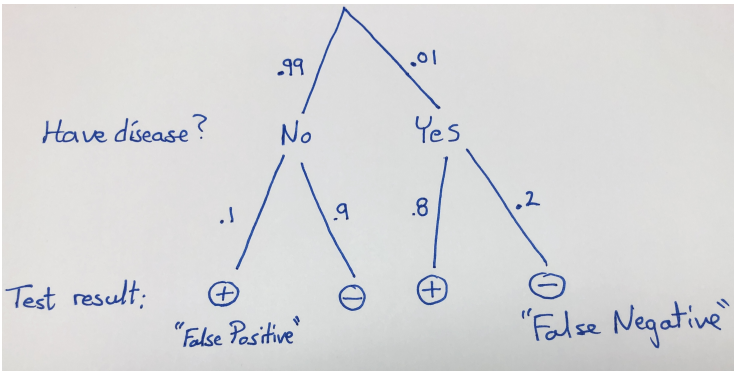
End of Example 34. □

10 2026-09-16 | Week 04 | Lecture 10

The topic for this lecture: *diagnostic tests and independence*
Reading assignment: *Chapter 2.4, 2.5*

10.1 Conditional probability example: diagnostic tests

Example 35 (Diagnostic tests)**.** 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. (24).
- (Denominator) By the law of total probability,
$$\begin{aligned}\mathbb{P}\left[T^+\right] &= \underbrace{\mathbb{P}\left[T^+|D\right]\mathbb{P}\left[D\right]}_{=0.008 \text{ by Eq. (21)}} + \underbrace{\mathbb{P}\left[T^+ \mid D^c\right]\mathbb{P}\left[D^c\right]}_{=(.1)(.99)=0.099} \\ &= (.8)(.01) + (.1)(.99) \\ &= 0.008 + 0.099\end{aligned}$$

Finally, having done the necessary work, we can now plug our results into Eq. (24):

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

We conclude that if you test positive in a random screening, your probability of actually having the disease is about 7.5%. *This result is perhaps suprising: even though you tested positive, you still probably don't have the disease.*

End of Example 35. □

10.2 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 36 (Independence)**.** Events A, B are ***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 37 (Mutual independence)**.** Events $A_1, \ldots, A_n$ are ***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 38. *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 39. (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 38 simply formalizes this idea using conditional probabilities.

Remark 40. 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!

10.3 Examples of using independence

Example 41. 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 41. □

11 2026-09-18 | Week 04 | Lecture 11

The topic for this lecture: *random variables*
Reading assignment: *Chapter 3.1, 3.2*

11.1 The "at least one" trick

Example 42 (Independence, and the "at least one" trick)**.** 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 42. □

11.2 Dicrete Random Variables

Based on sections 3.1 and 3.2 of the textbook

We often associate numbers with outcomes of an experiment.

Definition 43 (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 44. 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 44. □

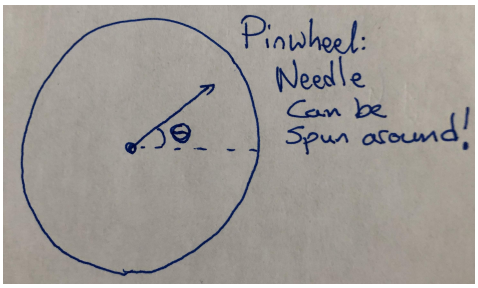
Definition 45 (Discrete random variable)**.** We say that a random variable X is a ***discrete random variable*** if it can assume only a finite or countably infinite number of distinct values.

Definition 46 (Continuous random variable)**.** A random variable X is said to be ***continuous*** if

(i) it takes values in an interval $I \subseteq \mathbb{R}$ or a union of such intervals; and,

(ii) $\mathbb{P}[X = x] = 0$ for all x.

Example 47 (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

$$0 \leq X < 360$$

but X could take *any* value between 0 and 360, not just integer values. We have no reason to believe any one angle is more likely than any others – for our idealized pinwheel, all values between 0 and 360 should all be equally likely.

I claim that it must be the case that

$$\Pr[X = x] = 0, \qquad \text{for all } x \in [0, 360).$$

We'll prove this by contradiction:

Suppose on the contrary that $\Pr[X = x] = c > 0$ for all x. Let N be any positive integer greater than $1/c$. Let A be a set of any N numbers between 0 and 360. Then using **A.3**

$$\begin{aligned}
\Pr[X \in A] &= \sum_{a \in A} \Pr[X = a] \\
&= \sum_{a \in A} c \\
&= c|A| \\
&= cN \\
&> 1.
\end{aligned}$$

But this contradicts (vi.). Thus we must have $c = 0$.

End of Example 47. □

Definition 48 (Continuous random variable)**.** We

Example 49 (Bernoulli random variable)**.** Let $p \in (0, 1)$. We say that X is a ***Bernoulli random variable*** with success parameter p and write $X \sim \text{Bern}(p)$, if

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

Conceptually, a Bernoulli random variable represents the flip of a biased coin that has probability p of landing on heads.

End of Example 49. □

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

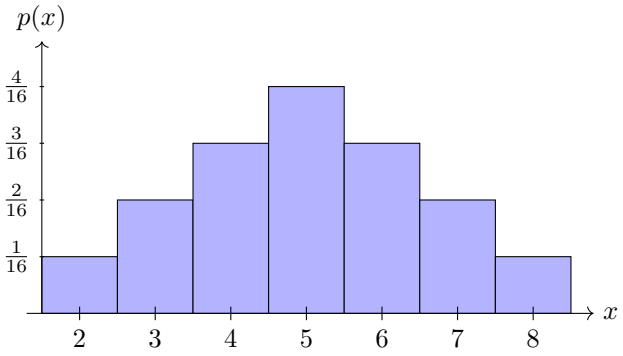
$$p(x) = \mathbb{P}[X = x], \quad x \in \mathbb{R}.$$

Example 51. The pmf of X in Example 44 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, and visualize the distribution of X as a histogram:

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

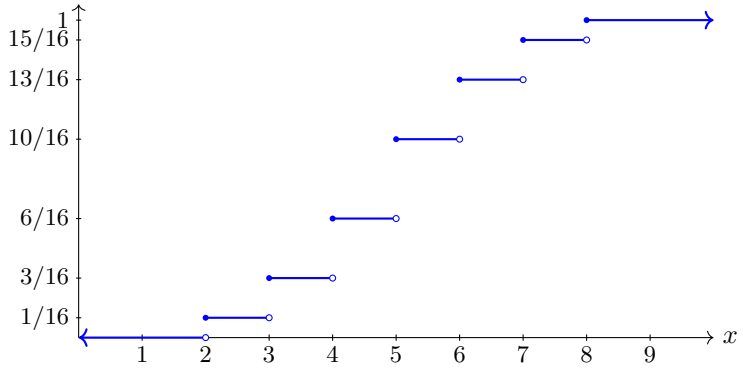
$$F(x) = \mathbb{P}[X \leq x],$$

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

Example 53. For the random variable in Example 44, 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 53. □

12 2026-09-21 | Week 05 | Lecture 12

The topic for this lecture: *expected value*
Reading assignment: *Chapter 3.3*

Definition 54 (Expected value - discrete random-variable)**.** For a discrete random variable X, the ***expected value*** of X, denoted $\mathbb{E}[X]$, is

$$\mathbb{E}[X] = \sum_x x\mathbb{P}[X = x] \tag{5}$$

where the sum ranges over all distinct values that X can take. (Technical note: the sum on the right hand side of may be an infinite series, which may or may not converge. If the sum on the right-hand side of Eq. (5) does not converge absolutely, then we say that the expected value of X ***does not exist***.)

Remark 55. $\mathbb{E}[X]$ is the long-run average of X, if we were to repeat the experiment many times.

The following is a simple but important fact:

Theorem 56 (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

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]. \tag{6}$$

Corollary 57 (Linearity of expectation)**.** *If X and Y are any random variables, then*

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y].$$

Proof. Take $a = b = 1$ in Eq. (6). □

Example 58 (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.

Assume that

$$\mathbb{P}[\text{win}] = \frac{1}{300,000,000}, \qquad \mathbb{P}[\text{lose}] = \frac{299,999,999}{300,000,000}.$$

Let X be the net payoff from the game, in dollars. Then, since you had to pay \$2 to buy the ticket,

$$X = \begin{cases} 299,999,998 & : \text{ if you win} \\ -2 & : \text{ if you lose} \end{cases}$$

What is the expected value of X? By Definition 54,

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

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 is linearity of expectation.

End of Example 58. □

Example 59 (Binomial random variable)**.** Suppose we flip a coin n times. Assume that for each flip, the coin has probability of heads p and probability of tails $1-p$, where $p \in (0,1)$.

Let

$$X = (\#\text{ of heads in } ntextcoinflips)$$

Constructed in this way, the random variable X is a ***binomial*** random variable with n trials and probability of success p, and we write $X \sim \text{Bin}(n,p)$.

What is $\mathbb{E}[X]$?

Solution: For each $k \in \{1,2,\ldots,n\}$ define

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

Then the key observation is that

$$X = Y_1 + \ldots + Y_n.$$

Therefore by linearity of expectation,

$$\mathbb{E}[X] = \mathbb{E}[Y_1] + \ldots + \mathbb{E}[Y_n]. \tag{7}$$

Now, what is $\mathbb{E}[Y_k]$? Observe that for each $k = 1,\ldots,n$ we have

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

Therefore by Definition 54,

$$\mathbb{E}[Y_k] = (1)(p) + (0)(1-p) = p$$

for every $k \in \{1,\ldots,n\}$. Plugging this into Eq. (7) implies

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

End of Example 59. □

13 2026-09-23 | Week 05 | Lecture 13

The topic for this lecture: *expected value*
Reading assignment: *Chapter 3.3*

13.1 Expected value examples

Example 60 (Repeated coin flipping)**.** Consider an experiment in which we flip a coin repeatedly until we get a heads. The sample space of the experiment is

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

Let

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

As discussed in Example 4, we have

$$\mathbb{P}[X = n] = \left(\frac{1}{2}\right)^n, \qquad n \in \{1, 2, \ldots\}. \tag{8}$$

Question: What is the expected number of coin flips? In other words, what is $\mathbb{E}[X]$?

We use the expected value 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 60. □

Example 61 (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!!!!! No matter how much you pay to play this game, you will eventually make money if you play enough times.

End of Example 61. □

13.2 Law of the unconscious statistician

Theorem 62 (LOTUS - disrete case)**.** *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 63. 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):

$$\begin{aligned}\mathbb{E}[3Y+7] &= 3\mathbb{E}[Y] + 7 && \text{by Theorem 56} \\ &= 3(0.3) + 7 && \text{by our answer to part (a).} \\ &= 7.9.\end{aligned}$$

Solution to (c): Here we will apply Theorem 62 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 62 with $h(x) = e^x$:

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

End of Example 63. □

14 2026-09-25 | Week 05 | Lecture 14

The topic for this lecture: *variance*
Reading assignment: *Chapter 3.3, 3.4*

14.1 The "spread" of a distribution

We begin by illustrating the distributions of some random variables we've seen so far. In the plots below, for each x, the height of the bar at x is the probability that the rv takes vales x (ie it is the value of $p(x)$). Note that the shaded areas sum to one.

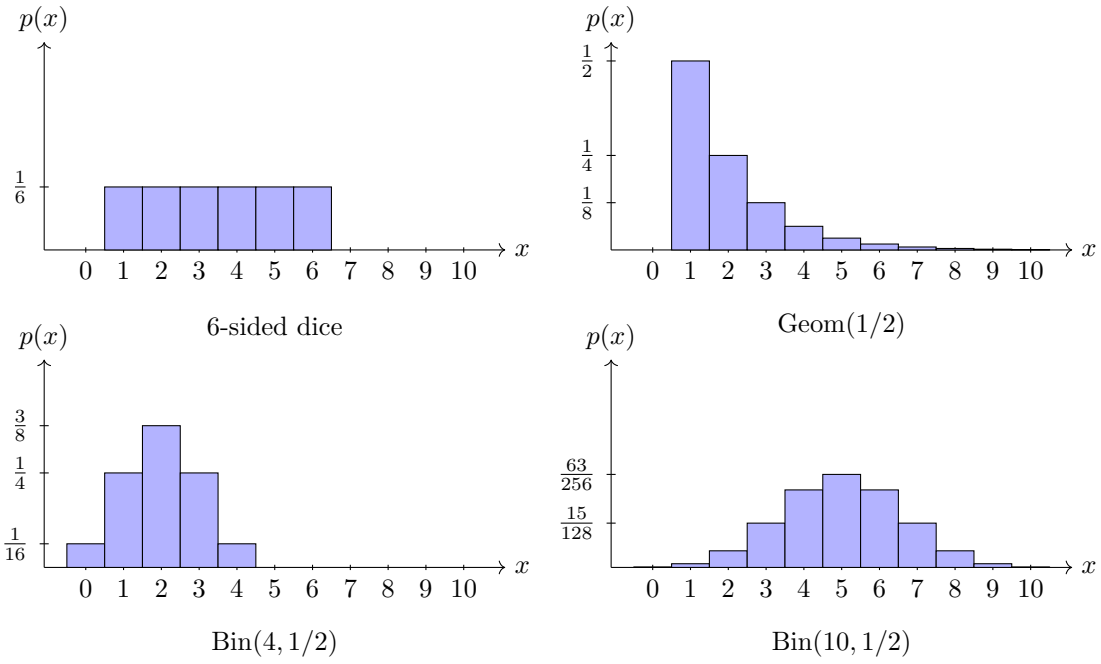


Figure 1: The distributions fora 6-sided dice (top left), the coin-fliping random variable from Example 4 (top right), and binomial random variables with parameters $(n, p) = (4, .5)$ (bottom left) and $(n, p) = (10, .5)$ (bottom right).

Question: How can we measure how much variation a random variable exibits? That is, how spread out or concentrated is its distribution?

One useful measure is called the *variance*.

14.2 Variance

Definition 64 (Variance)**.** Let X be a random variable. The ***variance*** of X is

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

where $\mu = \mathbb{E}[X]$. The positive square root of the variance is called the ***standard deviation*** of X.

Remark 65. Some notes about the above formula:

- The random quantity $X - \mu$ is the deviation of X from its mean. Since this gets squared, large deviations are amplified.

- Since $(X-\mu)^2 \geq 0$, the variance is always nonnegative.
- Notation: standard deviation is usually denoted by σ, and variance by σ^2.
- **Interpretation of variance:** Variance is a measure of how spread out the distribution of a random variable is, e.g., how likely the value of a random variable is to be far from its expected value.

Taking $h(x) = (x-\mu)^2$ in Theorem 62, we can write the variance as

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

where p is the pmf of X.

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

Theorem 66 (Variance shortcut formula)**.** *Let X be a random variable with $\mathbb{E}\left[X^2\right] < \infty$. Then*

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

The condition $\mathbb{E}\left[X^2\right] < \infty$ in Theorem 66 is there to exclude the possibility that the right hand side of the formula be the undefined form $\infty - \infty$.

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

Theorem 67 (Properties of Variance)**.** *Let X be a random variable, and $a \in \mathbb{R}$ a scalar. Then*

$$\mathrm{Var}(X+a) = \mathrm{Var}(X). \tag{9}$$

and

$$\mathrm{Var}(aX) = a^2 \mathrm{Var}(X) \tag{10}$$

In words, Eq. (9) says that shifting the distribution of an r.v. to the left or right does not change its variance. Eq. (10) say that scaling a random variable by a results in a quadratic scaling of its variance.

By linearity of expectation, $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$. The next theorem shows that a similar property holds for variance, provided the random variables are indpendent.

Theorem 68 (Additivity of variance)**.** *If X and Y are independent random variables, then*

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y). \tag{11}$$

This formula can be though of as analogous to the Pythagorean theorem.

14.3 Variance of the binomial distribution

Definition 69 (Binomial random variable)**.** A random variable X is said to be a ***Binomial random variable*** with n trials and success probability p if it has pmf

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x \in \{0, 1, \ldots, n\}.$$

In this case, we write $X \sim \mathrm{Bin}(n, p)$.

Remark 70 (Useful characterization of binomial random variables)**.** If X is a Binomial random variable with n trials and success probability p, then we can write it as a sum of independent random variables $Y_1, \ldots, Y_n$:

$$X = Y_1 + \ldots + Y_n$$

where

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

In words Y_i is either one or zero depending on whether the i^{th} coin flip was 'heads' or not. Hence, the sum $Y_1 + \ldots + Y_n$ counts the number of heads in n coin flips.

Example 71 (Variance of binomial rv)**.** In Example 59, we showed that $\mathbb{E}[X] = np$. Now we compute the variance. Using the shortcut formula (Theorem 66),

$$\mathrm{Var}(Y_i) = p - p^2 = p(1-p).$$

Since $Y_1, \ldots, Y_n$ are independent, Theorem 68 gives

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{Var}(Y_1) + \ldots + \mathrm{Var}(Y_n) \\ &= p(1-p) + \ldots + p(1-p) \\ &= np(1-p).\end{aligned}$$

Note that the variance increases with n. As a function of p it is maximized when $p = 1/2$.

End of Example 71. □

15 2026-09-28 | Week 06 | Lecture 15

Topic of this lecture: *Poisson distribution*
Reading assignment: *chapters 3.4, 3.6*

15.1 Motivation

Fix a small positive number $\lambda > 0$. For each $i \in \{1, \ldots, n\}$ let

$$Y_i = \begin{cases} 1 & : \text{ with probability } \frac{\lambda}{n} \\ 0 & : \text{ with probability } 1 - \frac{\lambda}{n} \end{cases}$$

Assume that $Y_1, \ldots, Y_n$ are independent. Informally, we think of them as coin flips, where $\mathbb{P}[\texttt{HEADS}] = \frac{\lambda}{n}$. Now consider their sum:

$$\begin{aligned} S_n &= Y_1 + \ldots + Y_n \\ &= \left(\#\text{ of heads in } n \text{ flips, assuming } \mathbb{P}[\texttt{HEADS}] = \frac{\lambda}{n} \right). \end{aligned}$$

Since $S_n \sim \text{Binom}(n, \frac{\lambda}{n})$, our earlier computations with binomial rvs in Examples 59 and 71 gives

$$\mathbb{E}[S_n] = n\left(\frac{\lambda}{n}\right) = \lambda$$

and

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

Question: What happens as $n \to \infty$? Think! As n grows, the probability of heads is getting smaller, but the number of coins is increasing. These two forces balance out, so that the expected number of heads is always λ, which doesn't change as n grows. Could this procedure converge to a new random variable that we haven't seen yet? What would it look like?

If there exists a random variable X such that $X = \lim_{n\to\infty} S_n$, then we anticipate that it should satisfy

$$\mathbb{E}[X] = \lim_{n\to\infty} \mathbb{E}[S_n] = \lambda$$

and

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

Such a random variable exists. It is called a Poisson random variable, and we introduce it next.

15.2 The Poisson distribution

Theorem 72 (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 73 (Poisson ("Pwason"))**.** Let $\lambda > 0$. A random variable X is a ***Poisson random variable*** with mean $\lambda > 0$ if it has probability mass function

$$p(k) = e^{-\lambda}\frac{\lambda^k}{k!} \quad \text{for } k \in \{0, 1, 2, \ldots\}.$$

Equivalently, we say that X has a ***Poisson distribution*** with mean λ, and write $X \sim \text{Pois}(\lambda)$.

Remark 74. Before continuing, we should check that $p(k)$ is a valid probability mass function. This means we need to check two things:

- Check that $p(k) \geq 0$ for all k.
- Check that $\sum_k p(k) = 1$.

It is clear that $p(k) = \frac{e^{-\lambda}\lambda^k}{k} \geq 0$, so the first condition is satisfied. To check the second condition, we have:

$$\begin{aligned}\sum_{k=0}^{\infty} p(k) &= \sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!} \\ &= e^{-\lambda}\left(\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\right) \\ &= e^{-\lambda}e^{\lambda} && \text{by Theorem 72} \\ &= 1.\end{aligned}$$

Therefore p is indeed a valid pmf.

Proposition 75. *Let $X \sim Pois(\lambda)$. Then*

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

The Poisson distribution is commonly-used to model *accidents*.

Example 76. 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

$$\mathbb{P}[X \leq 2]\,?$$

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 $\mathbb{P}[X \leq 2] = p(0) + p(1) + p(2) \approx 0.002 + 0.015 + 0.045 = 0.062$. Hence there is about 6.2% chance that a given day has 2 or fewer accidents. are

End of Example 76. □

Theorem 77 (Poisson approximation, AKA: The Law of Rare Events)**.** *Assume that $X \sim \mathrm{Binom}(n, p)$. This means that X has pmf*

$$\mathbb{P}[X = k] = \binom{n}{k}p^k(1-p)^{n-k}, \qquad k \in \{0, 1, \ldots, n\}$$

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

$$\mathbb{P}[X = k] \approx e^{-\lambda}\frac{\lambda^k}{k!}, \qquad k \in \{0, 1, \ldots\}$$

Remark 78. In other words, Theorem 79 says that when n is large and p is small, the probability distribution of X is well-approximated by a Poisson distribution with parameter $\lambda = np$. That is,

$$\mathrm{Binom}(n, p) \approx \mathrm{Pois}(\lambda),$$

where $\lambda = np$.

16 2026-09-30 | Week 06 | Lecture 16

Theorem 79 (The Law of Rare Events)**.** *Assume that $X \sim \text{Binom}(n, p)$. When n is large and p is small (e.g., $n > 50$ and $\lambda := np < 5$), then*

$$\mathbb{P}[X = k] \approx e^{-\lambda} \frac{\lambda^k}{k!}, \qquad k \in \{0, 1, \ldots\}. \tag{12}$$

Proof. Recall that

$$\mathbb{P}[X = k] = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k \in \{0, 1, \ldots, n\}.$$

Fix $\lambda > 0$ and let $p = \frac{\lambda}{n}$. To prove Eq. (12), it will be enough to show that

$$\lim_{n\to\infty} \mathbb{P}[X = k] = e^{-\lambda} \frac{\lambda^k}{k!}. \tag{13}$$

Using $p = \lambda/n$, write

$$\mathbb{P}[X = k] = \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}.$$

Since $\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!}$, we may rewrite the above as

$$\mathbb{P}[X = k] = \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}.$$

Rearranging terms on the right-hand side gives

$$\mathbb{P}[X = k] = \underbrace{\frac{n(n-1)(n-2)\cdots(n-k+1)}{n^k}}_{A_n} \underbrace{\left(1 - \frac{\lambda}{n}\right)^{-k}}_{B_n} \underbrace{\frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{n}\right)^{n}}_{C_n}.$$

Let's consider A_n, B_n and C_n separately:

- Observe that

$$\begin{aligned} A_n &= \frac{n}{n} \cdot \frac{n-1}{n} \cdot \frac{n-2}{n} \cdots \frac{n-k+1}{n} \\ &= 1 \cdot \left(1 - \frac{1}{n}\right) \cdot \left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{k-1}{n}\right) \end{aligned}$$

  Therefore $\lim_{n\to\infty} A_n = 1$.

- Since k is fixed,

$$\lim_{n\to\infty} B_n = \lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^{-k} = 1^{-k} = 1.$$

- Finally, using the well-known limit $\lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n = e^x$, we have

$$\lim_{n\to\infty} C_n = \frac{\lambda^k}{k!} e^{-\lambda}.$$

Therefore

$$\lim_{n\to\infty} \mathbb{P}[X = k] = \left(\lim_{n\to\infty} A_n\right) \left(\lim_{n\to\infty} B_n\right) \left(\lim_{n\to\infty} C_n\right) = (1)(1) \left(\frac{\lambda^k}{k!} e^{-\lambda}\right) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

This proves Eq. (13), as desired. □

How accurate is the Poisson approximation of Theorem 79?

Example 80 (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

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

The expected value of X is five. The pmf of X is

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

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 80. □

17 2026-10-02 | Week 06 | Lecture 17

Midterm day.

18 2026-10-04 | Week 07 | Lecture 18

18.1 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

$$\mathbb{P}[W] = \sum_{k=2}^{12} \mathbb{P}[W \mid F_k]\,\mathbb{P}[F_k] \tag{14}$$

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

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

$$\mathbb{P}[****4] = \left(\frac{27}{36}\right)^4 \frac{3}{36}$$

$$\ldots$$

So

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

(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. (14) gives

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

Hence, the shooter's chance of winning is about 49.3%. The house has a slight advantage.

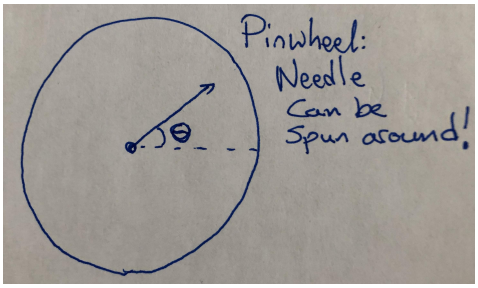
18.2 Continuous random variables

(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 81 (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 82 (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. 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 360 should all be equally likely.

End of Example 82. □

Definition 83 (pdf)**.** Let X be a continuous random variable. The ***probability density function (pdf)*** of X is a function f such that

$$\mathbb{P}[a \leq X \leq b] = \int_a^b f(x)dx, \qquad a \leq b. \tag{15}$$

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

Remark 84. A pdf is similar to a pmf, with one big exception: the value of $f(x)$ is **not** a probability because it could be greater than 1. Instead it is useful to think of it as a "relative likeliood".

Recall the ***cdf*** of a rv X is the function

$$F(x) := \Pr(X \leq x).$$

For a continuous random variable, we have the following theorem:

Theorem 85. *If X is a continuous random variable with cdf F and pdf f, then*

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

Proof. The first formula follows by setting $x = b$ sending $a \to -\infty$ in Eq. (15). The second formula is due to the fundamental theorem of calculus. □

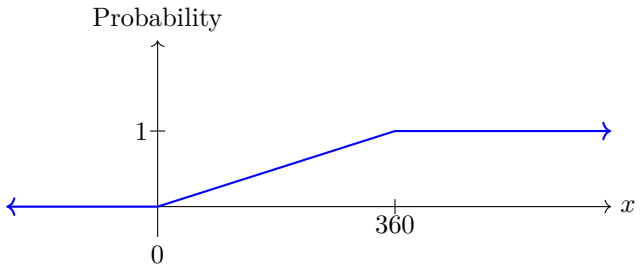
For Example 82, the pdf of X is

$$f(x) = \begin{cases} \frac{1}{360} & : \quad 0 \leq x < 360 \\ 0 & : \quad \text{otherwise} \end{cases}$$

and by Theorem 85 the cdf is

$$F(x) = \begin{cases} 0 & : \quad x < 0 \\ \frac{x}{360} & : \quad 0 \leq x < 360 \\ 1 & : \quad x \geq 360 \end{cases}$$

The cdf has plot



Remark 86. To be a valid pmf, a function f must satisfy two properties:

$$f(x) \geq 0 \text{ for all } x \in \mathbb{R}, \quad \text{and} \quad \int_{-\infty}^{\infty} f(x)dx = 1.$$

Proposition 87 (Properties of cdfs)**.** *If F is the cdf of X then*

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

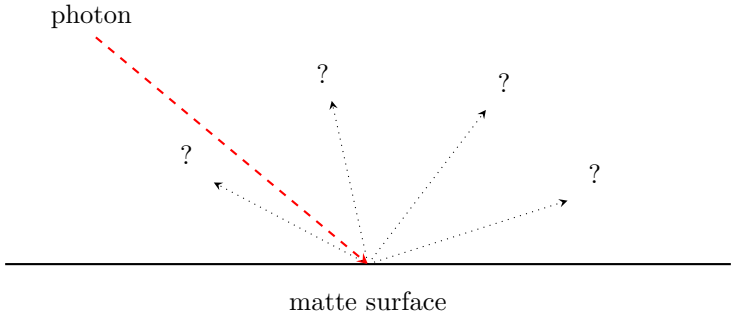
19 2026-10-07 | Week 07 | Lecture 19

worked on in-class worksheet (due as homework)

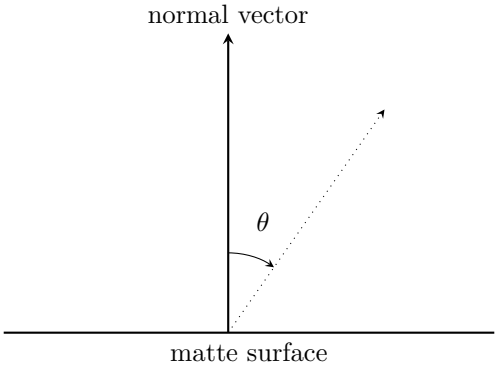
20 2026-10-09 | Week 07 | Lecture 20

Topic of this lecture: *Continuous random variables*
Reading assignment: *Chapters 4.1, 4.2*

Example 88 (Diffuse reflection of light)**.** When light hits a matte surface, such as a sheet of paper, it scatters in many directions. For a single photon, the angle of reflection is random:



Let X be the random angle θ between the reflected ray and the normal vector to the surface, as shown:



Measuring the angle in radians, we have

$$-\frac{\pi}{2} \leq \theta \leq \frac{\pi}{2},$$

where

- A positive angle means the ray exits to the right of the normal vector;
- A negative angle means it exits to the left.

Lambert's cosine law implies that the probability density at angle θ is proportional to $\cos(\theta)$; that is, the pdf of X is

$$f(\theta) = C\cos(\theta), \qquad -\frac{\pi}{2} \leq \theta \leq \frac{\pi}{2}$$

for some constant C.
Question 1. Find C.
Solution: The total area under a pdf is always 1. Since $f(\theta) = 0$ for all $\theta \notin \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$, we have

$$\int_{-\theta/2}^{\pi/2} f(\theta)d\theta = 1.$$

Therefore

$$C\int_{-\pi/2}^{\pi/2} \cos(\theta)d\theta = 1$$

The integral evaluates to 2, and hence

$$C = \frac{1}{2}.$$

Therefore, the pdf of X is

$$f(\theta) = \begin{cases} \frac{1}{2}\cos(\theta) & : \quad -\frac{\pi}{2} \leq \theta \leq \frac{\pi}{2} \\ 0 & : \quad \text{otherwise/} \end{cases}$$

Question 2. What is the probability that the photon bounces off the surface at an angle of less than 30° from the normal vector?
Solution: Since 30° is $\pi/6$ radians,

$$\begin{aligned} \mathbb{P}\left[-\frac{\pi}{6} \leq X \leq \frac{\pi}{6}\right] &= \int_{-\pi/6}^{\frac{\pi}{6}} f(\theta) d\theta \\ &= \frac{1}{2}\int_{-\pi/6}^{\frac{\pi}{6}} \cos(\theta) d\theta \\ &= \frac{1}{2}\left[\sin(\theta)\right]_{-\pi/6}^{\pi/6} \\ &= \frac{1}{2}\left[\frac{1}{2} - \left(-\frac{1}{2}\right)\right] \\ &= \frac{1}{2}. \end{aligned}$$

Question 3. What is the expected angle of the reflectced photon ray?
Solution: We want to compute $\mathbb{E}[X]$.

$$\begin{aligned} \mathbb{E}[X] &= \int_{-\infty}^{\infty} \theta f(\theta)\, d\theta \\ &= \frac{1}{2}\int_{-\pi/2}^{\pi/2} \theta\cos(\theta)\, d\theta \\ &= 0, \end{aligned}$$

where we used the fact that (1) the function $\theta\cos(\theta)$ is odd, and (2) the domain of integration $[-\pi/2, \pi/2]$ is symmetric around the origin. (You can check that the integral equals zero using integration by parts).

End of Example 88. □

21 2026-10-12 | Week 08 | Lecture 21

21.1 Normal random variables

This section is based on chapter 4.3.

Definition 89 (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 90 (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 91. 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 91. □

Remark 92 (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 93 (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:

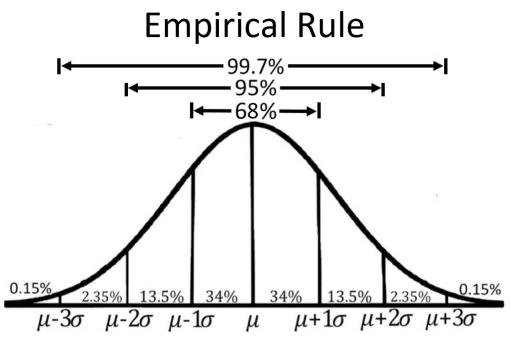


Image source: https://andymath.com/wp-content/uploads/2019/12/empirical-rule-normdist2.jpg

Remark 94. Note that μ determines where the peak of the bell curve is located, and σ determines how wide or narrow it is.

Definition 95. The ***standard normal curve*** is the graph of the standard normal density

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$$

If Z is a standard normal random variable, we denote its cdf by

$$\Phi(z) := \mathbb{P}[Z \leq z] = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz.$$

Further, we denote by z_α the value on the x-axis for which α of the area under the curve lies to the right of z_α. We call z_α the $100(1-\alpha)^{\text{th}}$ percentile of the standard normal distribution.

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

Example 96. Due to variation in the manufacturing process, the peak power consumption X of the latest Nvidia GPU varies slightly from chip to chip. Assume the power consumption is normally distributed $X \sim \mathcal{N}(\mu, \sigma^2)$. What is the probability that X is within one standard deviation of its mean?

$$\begin{aligned}
\mathbb{P}[\mu - \sigma \leq X \leq \mu + \sigma] &= \mathbb{P}\left[\frac{(\mu-\sigma)-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{(\mu+\sigma)-\mu}{\sigma}\right] \\
&= \mathbb{P}\left[\frac{(\mu-\sigma)-\mu}{\sigma} \leq Z \leq \frac{(\mu+\sigma)-\mu}{\sigma}\right] && \text{by Remark 92} \\
&= \mathbb{P}[-1 \leq Z \leq 1] \\
&= \int_{-1}^{1} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \\
&\approx .68 && \text{using the figure in Remark 93}
\end{aligned}$$

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

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

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

End of Example 96. □

22 2026-10-14 | Week 08 | Lecture 22

Topic of this lecture: *joint distributions of discrete random variables*
Reading assignment: *chapter 5.1, chapter 4.4 (focus on exponential rv), and chapter 4.2 (focus on definition of expectation and variance for continuous random variables, and the LOTUS theorem for continuous random variables. These are important, but I don't have time to present them in lecture)*

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

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

Remark 98. To be a valid joint pmf, two conditions must hold:

1. $p(x,y) \geq 0$
2. $\sum_{x,y} p(x,y) = 1$

Theorem 99. *If A is 2-dimensional set (e.g., consisting of pairs of (x,y) values), then*

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

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 pmf $p(x,y)$ of (X,Y) is given by the following table:

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

For example, the probability that a car has exactly one defective weld and three improperly tightened bolts is

$$p(1,2) = 0.008.$$

- Question 0. Is $p(x,y)$ a valid pmf?

  Answer: Yes. The entries of the table are (1) nonnegative and (2) sum to one, so $p(x,y)$ is a valid pmf.

- Question 1. What is the probability that a car has *no defects*?

  Answer: The probability that there are no defects in the car is

$$p(0,0) = \mathbb{P}\left[X = 0, Y = 0\right] = 0.84.$$

- Question 2. What is the probability that a car has *exactly one* defect?

  Answer: The probability that there will be exactly one defect is

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

- Question 3. What is the probability that there will be no improperly tightened bolts?

  Answer: This question concerns only the variable Y. The probability can be obtained by summing over all values of x:

$$\begin{aligned}\mathbb{P}[Y = 0] &= \sum_{x=0}^{2} \mathbb{P}[X = x, Y = 0] \\ &= p(0,0) + p(1,0) + p(2,0) \\ &= .84 + .06 + .01 \\ &= 0.91.\end{aligned}$$

  Therefore 91% of cars produced by the factory have no improperly tightened bolts.

  To find this value, observe that we summed over all values in the 0-column of the table above. This is called a ***column sum***.

- Question 4. What is the probability that there is exactly one defective weld?

  Answer: The probability is

$$\begin{aligned}\mathbb{P}[X = 1] &= \sum_{y=0}^{3} p(1,y) \\ &= p(1,0) + p(1,1) + p(1,2) + p(1,3) \\ &= .060 + .010 + .008 + .002 \\ &= .08.\end{aligned}$$

  So 8% of cars produced by the factory have exactly one defective weld.

  To find this value, we summed over all values in the 1-row of the table. This is called a ***row sum***.

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

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

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

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

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

$$p(x,y) = p_X(x) \cdot p_Y(y), \qquad \text{for all } x, y \in \mathbb{R}. \tag{16}$$

If this condition isn't satisfied for all pairs (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

- Question 5. What are the marginal pmfs p_X and p_Y?

  Answer: 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

- Question 6. Are X and Y independent?

  Answer: No. Observe that

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

  Therefore Eq. (16) is not satisfied. Hence X and Y are dependent.

End of Example 103. □

23 2026-10-16 | Week 08 | Lecture 23

Topic of this lecture: *joint distributions - continuous case*
Reading assignment: *textbook section 5.1*

23.1 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,

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

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

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

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

$$f(x,y) = f_X(x) \cdot f_Y(y), \qquad \text{for all } x, y \in \mathbb{R}.$$

Then worksheet for rest of class.

24 2026-10-19 | Week 09 | Lecture 24

Topic of this lecture: *joint distributions*
Reading assignment: *textbook section 5.1, 5.2*

- wed office hours pushed to thursday 2-3pm
- problem 1 is example 5.3 in the textbook
- quiz on wednesday (one problem from hw 6 and one from hw 7)

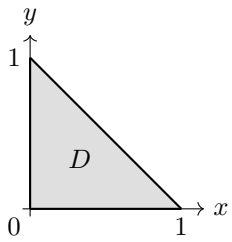
Example 105. Let X, Y be the random variables with joint pdf

$$f(x,y) = \begin{cases} 2 & : \quad x > 0, y > 0, x + y < 1 \\ 0 & : \quad \text{otherwise.} \end{cases}$$

The support of this function is the set

$$D = \big\{(x,y) \in \mathbb{R}^2 : x > 0, y > 0, x + y < 1\big\}$$

which is the region in $\mathbb{R}^2$ shown below:



(a) What is $\mathbb{P}[X > 1/2]$?

Solution: Recall that if f_X is the marginal density of X, then

$$\int_a^b f_X(x)dx = \mathbb{P}[a \le X \le b], \qquad -\infty \le a \le b \le \infty.$$

We start by computing the marginal density f_X. If $0 \le x \le 1$, then

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy = \int_0^{1-x} f(x,y)dy = \int_0^{1-x} 2dy = 2(1-x). \tag{17}$$

On the other hand, if $x \notin [0,1]$, then $f(x,y) = 0$ for all y, so

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy = 0. \tag{18}$$

Putting Eqs. (17) and (18) together,

$$f_X(x) = \begin{cases} 2(1-x) & : \quad 0 \le x \le 1 \\ 0 & : \quad \text{otherwise.} \end{cases}$$

Hence

$$\mathbb{P}[X > 1/2] = \int_{1/2}^{\infty} f_X(x)dx = \int_{1/2}^{1} 2(1-x)dx = \big[2x - x^2\big]_{1/2}^{1} = 1 - \frac{3}{4} = \frac{1}{4}.$$

End of Example 105. □

25 2026-10-21 | Week 09 | Lecture 25

Topic of this lecture: *Multivariate LOTUS Theorem and Covariance*
Reading assignment: *5.1, 5.2*

25.1 LOTUS Theorem - multivariate case

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 $h : \mathbb{R}^2 \to \mathbb{R}$,*

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

An important aplication of Theorem 106:

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

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

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

$$\begin{aligned} \mathbb{E}\left[XY\right] &= \sum_{\text{all } x,y} xyp(x,y) \\ &= \sum_{\text{all } x,y} xyp_X(x) \cdot p_Y(y) && \text{by independence} \\ &= \left(\sum_{\text{all } x} xp_X(x)\right)\left(\sum_{\text{all } y} yp_Y(y)\right) \\ &= \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right] && \text{by Theorem 62} \end{aligned}$$

The continuous case is similar, but with integrals. □

25.2 Covariance

The ***covariance*** of two random variables X, Y is

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

where $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$.

Proposition 108 (Covariance shortcut formula)**.** *This can be computed with Theorem 106, but by foiling it out, we get the shortcut formula*

$$\text{Cov}(X,Y) = \mathbb{E}\left[XY\right] - \mu_X\mu_Y \tag{19}$$

Interpetation: covariance measures the strength of the relationship between X and Y.

Proposition 109 (Covariance of independent rvs)**.** *By Theorem 107, if X, Y are independent, then*

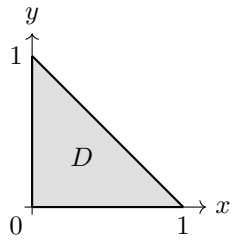
$$\text{Cov}(X,Y) = \mathbb{E}\left[XY\right] - \mu_X\mu_Y = \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right] - \mu_X\mu_Y = 0.$$

Note: We didn't finish the next example, but it is quite informative..

Example 110 (Covariance calculation)**.** Let

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

as shown:



Let X, Y be random variables with joint distribution

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

(a) **What are the marginal densities?**

Let $x \in [0,1]$, then

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f(x,y)dy \\ &= \int_0^{1-x} 24xydy \\ &= \left[12xy^2\right]_0^{1-x} \\ &= 12x(1-x)^2. \end{aligned}$$

If $x \notin [0,1]$, then (you should check) $f_X(x) = 0$. Hence,

$$f_X(x) = \begin{cases} 12x(1-x)^2 & : \quad 0 \leq x \leq 1 \\ 0 & : \quad \text{otherwise.} \end{cases}$$

A similar argument shows that

$$f_Y(y) = \begin{cases} 12y(1-y)^2 & : \quad 0 \leq y \leq 1 \\ 0 & : \quad \text{otherwise.} \end{cases}$$

(b) **What are $\mathbb{E}[X]$ and $\mathbb{E}[Y]$?**

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)dx = \int_0^1 12x^2(1-x)^2dx = 0.4.$$

By the same calculation with x replaced by y, we also have

$$\mathbb{E}[Y] = 0.4.$$

(c) **What is $\mathbb{E}[XY]$?**

Apply Theorem 106 with $h(x,y) = xy$:

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

(d) **What is** $\mathrm{Cov}(X,Y)$**?**

Using the shortcut formula Eq. (19),

$$\begin{aligned}
\mathrm{Cov}(X,Y) &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \\
&= (-8) - (0.4)^2 \\
&= -8.16.
\end{aligned}$$

End of Example 110. □

26 2026-10-26 | Week 10 | Lecture 26

Topic of this lecture: *What is statistics?*
Reading assignment: *Chapters 5.3, 5.4*

26.1 Statistics: a general framework

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 111 (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 111. □

The simpleset sampling procedure is called simple random sampling.

Definition 112 (Simple random sampling)**.** Let N and n denote the numbers of elements in the population and sample, respectively. If the sampling is conducted in such a way that each of the $\binom{N}{n}$ samples has an equal probability of being selected, the sampling is called ***simple random sampling***, and the result is a ***simple random sample***.

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

Definition 113 (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***.

26.2 What is a statistic?

Definition 114 (Statistic)**.** A ***statistic*** is any quantity whose value can be calculated from sample data.

In other words, if $X_1, \ldots, X_n$ is a random sample and $T : \mathbb{R}^n \to \mathbb{R}$ is a function, then the random variable

$$Y = T(X_1, \ldots, X_n)$$

is a statistic. 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***.

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 115 (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. Here are two important examples of estimators.

Example 116. 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 116. □

The following example illustrates some of the terminology we've introduced.

Example 117 (Estimating average lightbulb lifetime)**.** Suppose the lifetime of each light bulb produced in a certain factory is distributed according to the following pdf

$$f(y) = \begin{cases} \lambda e^{-\lambda y} & : \quad y > 0 \\ 0 & : \quad y \leq 0 \end{cases}$$

for some parameter $\lambda > 0$, where y is measured in years. Here, the **population** is either the set of lightbulbs produced by the factory, which we idealize as the probability distribution given by f. Suppose we do not know λ, and we wish to estimate it. Hence λ is the **target parameter**.

In this case f is an exponential distribution with rate $\lambda > 0$. The exponential distribution is used for modeling lifetimes of systems that are *not subject to degredation* (e.g., lifetimes of light bulbs, not diesel engines) as well as *waiting times* (e.g., the length of time intervals between cars passing by a tree on an idyllic country road). In the road example, if the rate of cars passing is 2 cars per minute, then on average you have to wait only 1/2 of a minute for a car to pass. That is, the *rate* and *expected values* are reciprocals of each other. Let's prove this fact in the following claim:

Claim 1: If Y has pdf f, then $\mathbb{E}[Y] = \frac{1}{\lambda}$.

Proof of Claim 1. By definition of expected value for a continuous random variable,

$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} y f(y) dy = \int_{0}^{\infty} \lambda y e^{-\lambda y} dy \overset{\text{IBP}}{=} \frac{1}{\lambda}.$$

□ Claim

Suppose a **random sample** $Y_1, \ldots, Y_n$ is taken from f. This means that $Y_1, \ldots, Y_n$ are independent random variables, each with pdf f.

Consider the **statistic**

$$\overline{Y} = \frac{Y_1 + \ldots + Y_n}{n}.$$

It is reasonable to believe that when n is large,

$$\overline{Y} \approx \mathbb{E}[Y] = \frac{1}{\lambda}.$$

Therefore a reasonable **estimator** for λ is

$$\frac{1}{\overline{Y}} = \frac{n}{Y_1 + \ldots + Y_n}.$$

If $n = 10$ and the observed values of $Y_1, \ldots, Y_{10}$ are

$$2, \ 0.8, \ 0.1, \ 2.3, \ 3.2, \ 5.4, \ 2, \ 0.4, \ 2.8, \quad \text{and } 3.4$$

then $\overline{Y} = 2.24$, and hence our **estimate** for λ would be $\frac{1}{2.24} \approx 0.45$.

End of Example 117. □

27 2026-10-28 | Week 10 | Lecture 27

Topic of this lecture: *Distribution of a Sample mean, Central Limit Theorem*
Reading assignment: *5.4*

Recall that if $X_1, \ldots, X_n$ is a random sample, then the ***sample mean*** is

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

Note that $\overline{X}$ is a random variable.

Definition 118. The mean and standard deviation of $\overline{X}$ are, respectively,

$$\mu_{\overline{X}} := \mathbb{E}\left[\overline{X}\right], \qquad \sigma_{\overline{X}} := \sqrt{\mathrm{Var}(\overline{X})}.$$

Theorem 119 (Key sample mean properties)**.** *Let $X_1, \ldots, X_n$ be a random sample from a distribution with mean μ and standard deviation $\sigma > 0$. Then*

$$\mu_{\overline{X}} = \mu, \qquad \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}.$$

Proof. First we compute $\mu_{\overline{X}}$

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

Next we compute $\sigma^2_{\overline{X}}$:

$$\begin{aligned}
\sigma^2_{\overline{X}} &= \mathrm{Var}(\overline{X}) \\
&= \mathrm{Var}\left(\frac{X_1 + \ldots + X_n}{n}\right) && \text{by defintion of } \overline{X} \\
&= \frac{1}{n^2}\mathrm{Var}\left(X_1 + \ldots + X_n\right) && \text{by Eq. (10)} \\
&= \frac{1}{n^2}\left(\mathrm{Var}(X_1) + \ldots + \mathrm{Var}(X_n)\right) && \text{by additivity (see Eq. (11))} \\
&= \frac{1}{n^2}\left(n\sigma^2\right) && \text{since } \mathrm{Var}(X_i) = \sigma^2 \text{ for all } i \\
&= \frac{\sigma^2}{n}.
\end{aligned}$$

Taking square roots gives the formula for $\sigma_{\overline{X}}$. □

27.1 The central limit theorem

Main idea of the central limit theorem: If $X_1, \ldots, X_n$ is a random sample from (almost) any distribution, then $\overline{X}$ will be approximately normal when n is large.

Theorem 120 (Central limit theorem)**.** *Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with common mean μ and variance $\sigma^2 > 0$. Define*

$$U_n := \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}.$$

Then as $n \to \infty$, we have $U_n \to Z$, where $Z \sim \mathcal{N}(0,1)$. In particular, this means that

$$\mathbb{P}[a \le U_n \le b] \approx \mathbb{P}[a \le Z \le b]$$

for any $-\infty \le a \le b \le \infty$.

Remark 121. When n is large, we have $U_n \approx Z$. Rearranging terms gives

$$\overline{X} \approx \mu + \frac{\sigma}{\sqrt{n}} Z. \tag{20}$$

The right-hand side is a normal random variable with mean μ and standard devaiation $\sigma/\sqrt{n}$. Thus, in words, the central limit theorem says that

$\overline{X}$ is asymptotically normally distributed with mean $\mu_{\overline{X}} = \mu$ and standard deviation $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$.

Remark 122. Usually $n \ge 30$ is big enough for the approximation to be valid.

Example 123 (CLT Coin Flipping)**.** Flip a fair coin 100 times. What is the probability that you get at least 55 heads?

Solution: Let

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

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

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

We will use Eq. (21). First observe that

$$\mu = \mathbb{E}[X_i] = \frac{1}{2}, \qquad \sigma = \sqrt{\mathrm{Var}(X_i)} = \frac{1}{2}, \qquad \sqrt{n} = \sqrt{100} = 10$$

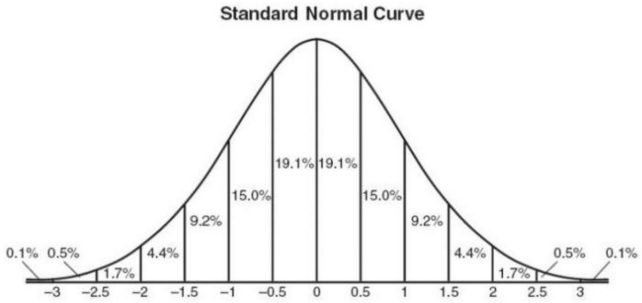
(you should check these equations). Therefore by the Central Limit Theorem,

$$\overline{X} \approx \mu + \frac{\sigma}{\sqrt{n}} Z = \frac{1}{2} + \frac{1}{20} Z.$$

Hencxe

$$\begin{aligned} \mathbb{P}\left[\overline{X} \ge .55\right] &\approx \mathbb{P}\left[\frac{1}{2} + \frac{1}{20} Z \ge .55\right] \\ &= \mathbb{P}[Z \ge 1] \\ &\approx .16. \end{aligned}$$

where we get .16 by consulting the following figure:



End of Example 123. □

Example 124. The average production time for an artillery shell at a factory is 1.5 minutes, with a variance of $\sigma^2 = 1$. Find the probability that the factory will produce at least 100 shells in the next two hours.

Solution: We don't need to assume the production times of the shells are normally distributed, but we'll assume they are independent. Let

$$X_i = (\text{time required to produce the } i^{\text{th}} \text{ shell}).$$

Here,

$$\mu = \mathbb{E}[X_i] = 1.5, \qquad \sigma^2 = \text{Var}(X_i) = 1.$$

Observe that

$$[\text{at least 100 shells are produced in two hours}] = [X_1 + \ldots + X_{100} \leq 120] = [\overline{X} \leq 1.2].$$

Here, $n = 100$. By the Central Limit Theorem,

$$\begin{aligned} \overline{X} &\approx \mu + \frac{\sigma}{\sqrt{n}} Z \\ &= 1.5 + \frac{1}{10} Z. \end{aligned}$$

Therefore

$$\begin{aligned} \mathbb{P}\left[\overline{X} \leq 1.2\right] &\approx \mathbb{P}\left[1.5 + \frac{1}{10} Z \leq 1.2\right] \\ &= \mathbb{P}[Z \leq -3] \\ &\approx 0.0013. \end{aligned}$$

Therefore it is very unlikely that the factory will produce more than 100 shells in the next two hours.

End of Example 124. □

28 2026-10-30 | Week 10 | Lecture 28

Topic of this lecture: *Central Limit Theorem, Intro to Point Estimation*
Reading assignment: *textbook section 5.4, 6.1*

28.1 Central Limit Theorem Example

Example 125 (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 $Z \sim \mathcal{N}(0,1)$. Therefore we can substitute $\overline{X}$ by the right-hand side above:

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

End of Example 125. □

28.2 Introduction to point estimation

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*** θ, and we want a single "best guess" $\hat{\theta}$ of some quantity of interest. When this is the case, the we call a statistic an estimator:

Definition 126 (Point estimator)**.** An ***point estimator*** is a statistic, that is a function $\widehat{\theta} = T(X_1, \ldots, X_n)$ of a sample, that is used to approximate a target parameter θ. The ***estimate*** is the realized value of an estimator (e.g., a number) that is obtained when the sample is actually taken, and is also (confusingly) also denoted $\hat{\theta}$.

Definition 127 (Bias)**.** 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***.

Definition 128 (Standard error)**.** The ***standard error*** of an estimator $\widehat{\theta}$, denoted $\sigma_{\widehat{\theta}}$, is its standard deviation:

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

Example 129. In Theorem 119 of the previous lecture, we saw that if $X_1, \ldots, X_n$ is a random sample with common mean $\mu = \mathbb{E}[X_i]$ and standard deviation $\sigma = \sqrt{\text{Var}(X_i)}$, then the estimator $\widehat{\theta} = \overline{X}$ has mean

$$\mu_{\overline{X}} := \mathbb{E}[\overline{X}] = \mu$$

and standard error

$$\sigma_{\overline{X}} := \sqrt{\text{Var}(\overline{X})} = \frac{\sigma}{\sqrt{n}}. \tag{21}$$

End of Example 129. □

Question: why is $\sigma_{\overline{X}} = \frac{\sigma}{n}$?
Answer: Observe that

$$\begin{aligned}
\sigma^2_{\overline{X}} &= \text{Var}(\overline{X}) \\
&= \text{Var}\left(\frac{1}{n}(X_1 + \ldots + X_n)\right) \\
&= \frac{1}{n^2}\left(\text{Var}(X_1) + \ldots + \text{Var}(X_n)\right) && \text{by independence.}
\end{aligned}$$

Since $\sigma^2 = \text{Var}(X_i)$ for every $i = 1, \ldots, n$, the above equation gives

$$\sigma^2_{\overline{X}} = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}.$$

Taking square roots gives

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}},$$

as required.

29 2026-11-02 | Week 11 | Lecture 29

Topic of this lecture: *examples of point estimators, Law of Large numbers*
Reading assignment: *chapter 6.1*

29.1 Examples of Point Estimators

Example 130 (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. 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 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

$$\theta = 2\mathbb{E}[X] - 1. \tag{22}$$

  By the law of large numbers, $\mathbb{E}[X] \approx \frac{X_1 + \ldots + X_n}{n}$. Combining this with Eq. (22) 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 $\widehat{\theta}$. Hopefully this calculation makes it seem like a reasonable choice.

  This estimator is unbiased. To see why, observe that

$$\begin{aligned}
\mathbb{E}[\widehat{\theta}] &= \mathbb{E}\left[2\left(\frac{X_1 + \ldots + X_n}{n}\right) - 1\right] \\
&= 2\mathbb{E}\left[\frac{X_1 + \ldots + X_n}{n}\right] - 1 \\
&= 2\mathbb{E}[X] - 1 \\
&= \theta && \text{by Eq. (22).}
\end{aligned}$$

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 130. □

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

29.2 The Law of Large Numbers

The Central Limit Theorem tells us that if $X_1, \ldots, X_n$ is a random sample with common mean μ and standard deviation $\sigma > 0$, then

$$\overline{X} \approx \mu + \frac{\sigma}{\sqrt{n}} Z.$$

So we expect at $n \to \infty$, $\overline{X} \to \mu$. The fact that

$$\overline{X} = \frac{X_1 + \ldots + X_n}{n} \to \mu \quad \text{as } n \to \infty$$

is called the ***Law of Large Numbers***. A precise statement of this result is shown below (we didn't get to it during lecture).

We will need the following theorem, which says that if the average male height is 5 feet tall, then no more than 10% of men can be more than 50 feet tall.

Theorem 131 (Markov's inequality)**.** *Let X be a nonnegative random variable and let $c > 0$. Then*

$$\mathbb{P}[X \geq c] \leq \frac{\mathbb{E}[X]}{c} \tag{23}$$

Proof. Let $A = [X \geq c]$ and let

$$\mathbf{1}_A = \begin{cases} 1 & : \quad X \geq c \\ 0 & : \quad X < c. \end{cases}$$

Then

$$X \geq X \cdot \mathbf{1}_A \geq c \cdot \mathbf{1}_A.$$

Therefore

$$\begin{aligned} \mathbb{E}[X] &\geq \mathbb{E}[c \cdot \mathbf{1}_A] \\ &\geq c\mathbb{E}[\mathbf{1}_A] \\ &= c\mathbb{P}[A] \\ &= c\mathbb{P}[X \geq a]. \end{aligned}$$

Dividing by c implies Eq. (23). □

Theorem 132 (Weak Law of Large Numbers)**.** *Let $(X_n)_{n=1}^{\infty}$ be a sequence of independent random variables. Assume that for all n, $\mu = \mathbb{E}[X_n]$ and $\mathrm{Var}(X_n) \leq B$ for some fixed bound $B < \infty$. Then for all $\epsilon > 0$,*

$$\mathbb{P}\left[\left|\overline{X} - \mu\right| > \epsilon\right] \to 0 \text{ as } n \to \infty.$$

In words, this says that when n is large, $\overline{X}$ is very unlikely to be far from μ.

Proof. Let $S_n = X_1 + \ldots + X_n$. By independence,

$$\mathrm{Var}(S_n) = \mathrm{Var}(X_1) + \ldots + \mathrm{Var}(X_n) \leq nB. \tag{24}$$

Next, let $\epsilon > 0$ be arbitrary.

$$\begin{aligned}
\mathbb{P}\left[\left|\overline{X} - \mu\right| > \epsilon\right] &= \mathbb{P}\left[|S_n - n\mu| > n\epsilon\right] \\
&= \mathbb{P}\left[(S_n - n\mu)^2 > (n\epsilon)^2\right] \\
&\leq \frac{\mathbb{E}\left[(S_n - n\mu)^2\right]}{n^2\epsilon^2} && \text{by Theorem 131} \\
&= \frac{\mathrm{Var}(S_n)}{n^2\epsilon^2} && \text{by definition of variance} \\
&\leq \frac{B}{\epsilon^2 n} && \text{by Eq. (24).}
\end{aligned}$$

The right hand side tends to zero as $n \to \infty$, proving the theorem. □

30 2026-11-04 | Week 11 | Lecture 30

30.1 Confidence intervals

Topic of this lecture: *Confidence Intervals*
Reading assignment: *Chapter 7.1*

Definition 133 (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

$$\mathbb{P}\left[|\widehat{p}-p| \leq \epsilon\right] \geq .95 \tag{25}$$

The value ϵ is the ***radius*** of the confidence interval.

In words, the interval contains the true parameter with probability at least .95. If we repeated 100 random samples, then we expect the true paramter p to lie in the confidence interval about 95 times.

Example 134 (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

$$X_i := \begin{cases} 1 & : \ i^{\text{th}} \text{ sampled person likes broccoli} \\ 0 & : \ \text{they don't} \end{cases}, \qquad i \in \{1, \ldots, n\}.$$

Our **sample** is $X_1, \ldots, X_n$. Consider the **point estimator**

$$\begin{aligned} \widehat{p} &:= \text{fraction of our sample who like broccoli)} \\ &= \frac{X_1 + \ldots + X_n}{n} \\ &= \overline{X}. \end{aligned}$$

Let

$$\sigma^2 := \text{Var}(X_i) = p(1-p).$$

Then by Theorem 119,

$$\mu_{\widehat{p}} := \mathbb{E}[\widehat{p}] = p, \qquad \sigma_{\widehat{p}} := \text{Var}(\widehat{p}) = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{p(1-p)}{n}}.$$

Recall that $\mu_{\widehat{p}}$ and $\sigma_{\widehat{p}}$ are called the **mean** and **standard error** of the estimator $\widehat{p}$. Since the mean equals the true parameter p, the estimator $\widehat{p}$ is **unbiased**.

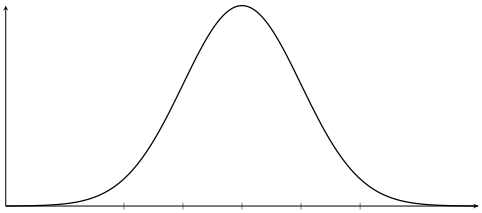
We can think of our estimate as

$$\widehat{p} = p + (\text{error of estimation}),$$

with the error of estimation becoming smaller as n increases. Indeed, since $\widehat{p}$ is a sample mean, the central limit theorem implies that its distribution is approximately normal:

$$\begin{aligned} \widehat{p} &\approx p + \frac{\sigma}{\sqrt{n}} Z \\ &= p + \sigma_{\widehat{p}} Z. \end{aligned}$$

Thus, the distribution of $\widehat{p}$ looks approximately like this:



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.

Suppose we sample $n = 100$ individuals and 20 of them like broccoli. Then

$$\widehat{p} = \frac{20}{100} = 0.2$$

This is our **estimate** of p. Moreover, we have

$$\sigma_{\widehat{p}} \approx \sqrt{\frac{(0.2)(1-0.2)}{100}} = \frac{.16}{10} = .016$$

Finding a confidence interval. 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. To figure out how big ϵ must be, we'll have to use the central limit theorem again.

By the central limit theorem,

$$\widehat{p} \approx p + \frac{\sigma}{\sqrt{n}} Z \tag{26}$$

We don't know what σ is, since it depends on the unknown parameter p, but we can approximate it:

$$\sigma = \sqrt{\mathrm{Var}(X_i)} = \sqrt{p(1-p)} \approx \sqrt{\widehat{p}(1-\widehat{p}) = (.2)(.8)} = \sqrt{.16} = .4 \tag{27}$$

By Eq. (26),

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

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 134. □