IE 300: Analysis of data

Chrysafis Vogiatzis

Written during Fall 2020 to accompany video lectures, worksheets, in-class and at home activities, quizzes, and exams. If you would like this material, too, please **email me**!

Dedicated to my wife, Eleftheria Kontou, who supported me throughout the COVID-19 quarantine period, and our whole lives.

Dedicated of course to all of the students in the Fall 2020 semester of IE 300. Thank you for being so kind and flexible.

Dedicated to our dog, Ralphie, even though he is responsible for some of the typos here 😊

Analysis of data

Chrysafis Vogiatzis

Contents

Part 1: Lectures 1–9

Random experiments, sample spaces, and events

Learning objectives

After this lecture, we will be able to:

- Give examples of experiments, sample spaces, and events.
- Explain sets and why they are used to describe events.
- Use Venn diagrams to represent events.
- Describe events using set operations.
- Give examples and recognize mutually exclusive events.
- Calculate the cardinality of an event.

Motivation: Monopoly

Is Monopoly a game of luck or strategy? It is your turn and your friends have built hotels *everywhere*. You need to roll two dies and get a 6 or a 7 to avoid paying your friends and declaring bankruptcy. *Everyone* expects you to lose: what are the "chances" you will roll a 6 or a 7, after considering all the scenarios?

In this first lecture, we will introduce and discuss all the necessary definitions in order to be able to quantify risks and chances.

Motivation: A card game

You are playing a card game on a deck with 52 cards of 4 different suits: ♡, ♣, ♢, ♠. You also know that there are 13 cards of each suit. The game is simple: *pick a card, any card*. If that card is red, you win; otherwise, you lose. For the intents of this game, we assume that the order of numbers is 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 followed by three face cards noted as J, Q, K.

Should we play? How should we play? Are there ways to mathematically quantify our risks and our gains?

Definitions

Random experiments

A **random experiment**[1] is defined as an activity or situation where the outcome obtained may be different, even when executed the same way. This randomness in the outcome could be due to the inherent

[1] Examples include the flip of a coin (could be Heads or Tails) or rolling a six-sided dice (could get any integer number between 1 and 6.

nature of the experiment, due to different levels of skill required to get a better result, or due to differences in instrumentation.

What are some other random experiments you can think of? Is measuring the width of a coffee table using a ruler a random experiment? How about cooking? Is a football game an experiment?

Back to the card game

Is our card game a random experiment? In essence, if you always follow the same strategy (e.g., pick the card at the top, or pick the card at the bottom), are you guaranteed the same result?

Sample spaces

With the term **sample space**[2], we refer to the set of all possible outcomes that can be obtained for a random experiment.

A sample space can have a finite or countably infinite number of possible outcomes (e.g., "1, 2, 3, 4, 5, or 6" or any integer number) or it can be an interval of real numbers (e.g., any number between 0 and 1, $[0, 1]$). We call the first type of sample space **discrete**. We will focus on that type of sample spaces in the beginning of the semester. The second type of sample space (where the outcome is a real number belonging to some interval) is called **continuous**. The rest of this lecture is devoted to discrete sample spaces.

[2] In a game of tic-tac-toe, the possible outcomes are win, lose, and tie, whereas in a graded course, the possible outcomes are $A, A-, B+, B, B-, \ldots, F$.

Is food poisoning a possible outcome of cooking? Is snow a possible outcome for tomorrow's weather?

Define the sample space for rolling a die and for rolling two dies. Define the sample space for the distance of any person at any point from the closest McDonald's.

Give an example of a sample space with a finite number of possible outcomes, and an example of a sample space defined over an interval of real numbers.

Back to the card game

Let us think about our card game. The number of outcomes is finite, that is for sure, so our sample space is discrete. *But what is our sample space?* There are multiple ways to describe the sample space here: $S = \{1\heartsuit, 2\heartsuit, \ldots, K\heartsuit, 1\clubsuit, 2\clubsuit, \ldots, K\spadesuit\}$ or $S = \{red, black\}$ or even $S = \{\heartsuit, \clubsuit, \diamondsuit, \spadesuit\}$.

The selection of the proper sample space is guided by what we are trying to achieve. In our motivation, we spoke about the color of the suit, so a sample space of $S = \{red, black\}$ seems the better choice.

Events

The term **event**[3] is used to define a subset of outcomes from the sample space. It can be just one or it can include many of the outcomes. An event can be a combination of outcomes ("get a 4 or more in a six-sided die") or the negation of an outcome ("no rain"). An event is *simple* if it has one outcome ("get dealt a Queen of ♣ in a deck of cards") or *compound* if it includes multiple outcomes ("don't lose" implies a win or a tie).

[3] For a student taking a graded class, an event can be to *pass* or to *get a grade higher than or equal to a B*.

Define some events for tomorrow's weather. Is "less than 10 minutes" a possible event for the experiment of counting the time until the next bus arrives? Is "farther than 10 miles" an event for the closest gas station?

In a board game where players roll two six-sided dies, is "getting a 10" a simple or a compound event?

Back to the card game

Let us return to the card game from our motivation. Assume that

$$S = \{1\heartsuit, 2\heartsuit, \ldots, K\heartsuit, 1\clubsuit, 2\clubsuit, \ldots, K\spadesuit\},$$

then the event E "picking a red card" is a **compound event** as there are 26 outcomes that satisfy it and

$$E = \{1\heartsuit, 2\heartsuit, \ldots, K\heartsuit, 1\diamondsuit, \ldots, K\diamondsuit\}.$$

Had we picked that $S = \{red, black\}$, then the event E "picking a red card" is a **simple event** as there is only one outcome that satisfies it (note that $E = red$ in this case).

Sets and set operations

Set operations

Set operations are a very useful way to describe events based on several outcomes. The most common set operations (and the ones we will predominantly use in this class) are:

- The union of two events E_1, E_2 as $\boxed{E_1 \cup E_2}$ [4].
- The intersubsection of two events E_1, E_2 as $\boxed{E_1 \cap E_2}$ [5].
- The complement of an event E as $\boxed{\overline{E}}$ (sometimes is also written as E^c or E') [6].
- The relative complement (sometimes termed as the *difference*) of an event E_2 from event E_1 as $\boxed{E_1 \setminus E_2}$ [7].

[4] *Either* event E_1 *or* E_2 (or *both*!) should happen.

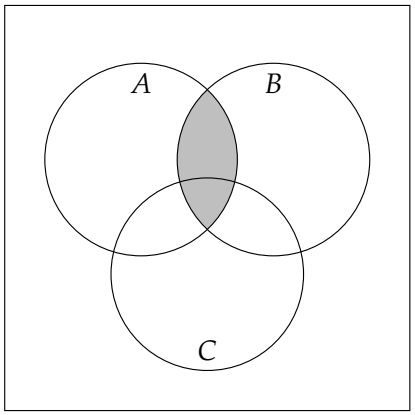
[5] *Both* events E_1 *and* E_2 should happen.

[6] *Any* other event *but* E.

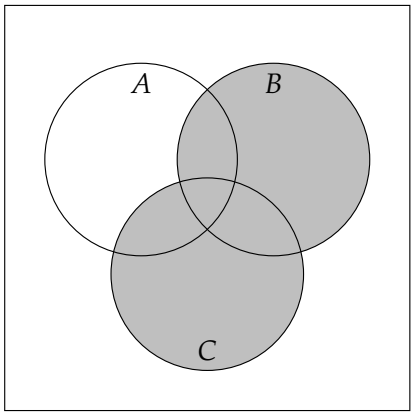
[7] *All outcomes in* E_1 *that are not also in* E_2.

Set operations are nicely described through the use of Venn diagrams. Consider the following examples, where the whole sample space is A, B, and C.

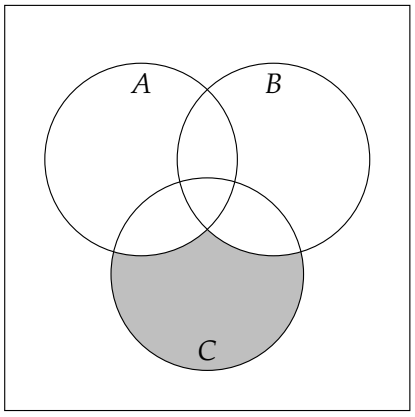
1. A and B should both happen $\rightarrow A \cap B$:



2. B or C should happen $\rightarrow B \cup C$:



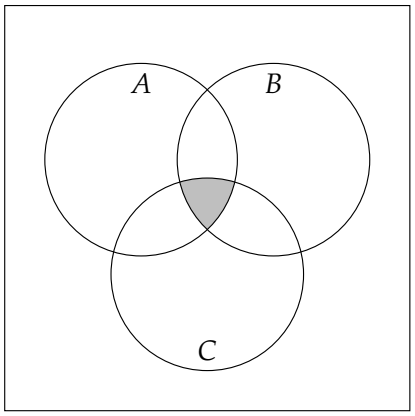
3. Neither A nor B should happen $\rightarrow \overline{A} \cap \overline{B}$: [8]



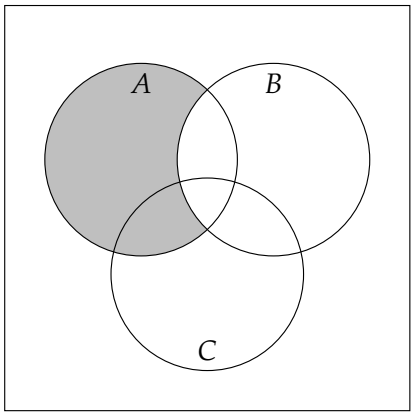
4. A, B, and C should all happen $\rightarrow A \cap B \cap C$:

[8] This can also be expressed as:

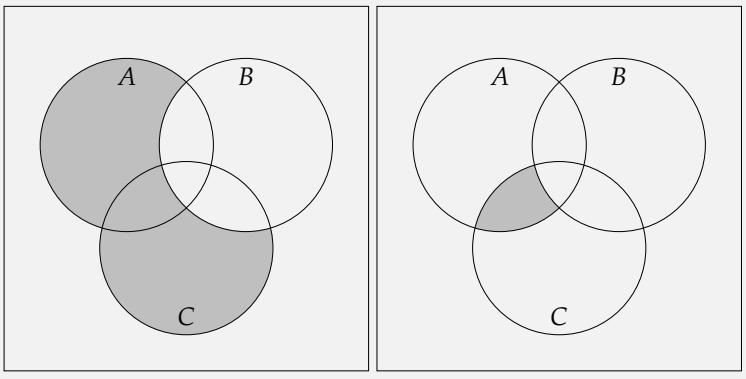
- Only C should happen but not A nor $B \rightarrow C \setminus (A \cup B)$.
- A or B should not happen $\rightarrow \overline{(A \cup B)}$.



5. A but not B should happen $\rightarrow A \setminus B$:



Describe mathematically and in English these two diagrams.



Some more definitions about sets:

- A set that contains no elements (outcomes) is called an empty or a null set, and is denoted by $\emptyset$.

- The sample space is a set containing all outcomes and is typically denoted by S.
- We say that event E_1 is a subset of event E_2 if all outcomes of event E_1 are included in event E_2 [9].
  - We denote this as $E_1 \subseteq E_2$.
  - By definition, $\emptyset \subseteq S$.

[9] In English, this also implies that event E_1 happening immediately signals that event E_2 is happening, too.

In the Venn diagrams earlier, we had $S = A \cup B \cup C$, as these were the only three possible outcomes.

Give an example of a two events where one is a subset of the other.

Finally, we say that two events are **mutually exclusive** [10] if they contain no common outcomes. Mathematically, two events E_1, E_2 are mutually exclusive if

$$E_1 \cap E_2 = \emptyset.$$

[10] You cannot both get a B in a class and *fail* the class at the same time.

Give an example of a pair of mutually exclusive events.

Back to the card game

Assume that

$$S = \{1\heartsuit, 2\heartsuit, \ldots, K\heartsuit, 1\clubsuit, 2\clubsuit, \ldots, K\spadesuit\},$$

and consider three events:

- A =“get a card with the value 3 or less”
- B =“get a red card”
- C =“get a face card”

What is:

- the union of A and B?

  $A \cup B$: “get a card with the value of 3 or less or a red card.”
  - The event happens if we get $7\heartsuit$.
  - The event happens if we get $2\diamondsuit$.
  - The event happens if we get $1\spadesuit$.
  - The event does not happen if we get $10\clubsuit$.

Back to the card game

What is:

- the intersubsection of B and C?

  $B \cap C$: "get a red and face card."

  - The event happens if we get $Q\heartsuit$.
  - The event happens if we get $J\diamondsuit$.
  - The event does not happen if we get $1\spadesuit$.
  - The event does not happen if we get $K\clubsuit$.

- the intersubsection of A and C?

  $B \cap C$: "get face card that is less than or equal to 3 in value."

  - The event never happens.
  - In set notation, we have $B \cap C = \emptyset$.
  - B and C are mutually exclusive events.

- the complement of C?

  $\overline{C}$: "not get a face card."

  - The event does not happen if we get $Q\heartsuit$.
  - The event does not happen if we get $J\diamondsuit$.
  - The event happens if we get $1\spadesuit$.
  - The event happens if we get $6\clubsuit$.

Set operation laws

Assume S is the sample space, and A, B, C are some events. Then:

1. $A \cup \overline{A} = S$, $A \cap \overline{A} = \emptyset$, $\overline{\overline{A}} = A$.
2. $A \cup B = B \cup A$ and $A \cap B = B \cap A$.
3. De Morgan's laws:
   - $\overline{(A \cup B)} = \overline{A} \cap \overline{B}$.
   - $\overline{(A \cap B)} = \overline{A} \cup \overline{B}$.
4. $A \cup (B \cup C) = (A \cup B) \cup C$, $A \cap (B \cap C) = (A \cap B) \cap C$.
5. $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$, $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.

Cardinality

The **cardinality** of an event E [11] is the number of outcomes that it contains and it is denoted by $|E|$. Some important cardinality properties are:

- $E = \emptyset \implies |E| = 0$.
- If E_1 is a subset of E_2, then $|E_1| \leq |E_2|$.
- If events E_1, E_2 are mutually exclusive, then
  - $|E_1 \cap E_2| = 0$.
  - $|E_1 \cup E_2| = |E_1| + |E_2|$.
- For *any* two events E_1, E_2, then
  - $|E_1 \cup E_2| = |E_1| + |E_2| - |E_1 \cap E_2|$ [12].

[11] In a graded course (where $S = \{A, A-, B+, B, B-, \ldots F\}$, the cardinality of $E = grade \geq B$ is 4 ($B, B+, A-$, and A), that is $|E| = 4$.

[12] Proving this is part of your worksheet for the day.

Back to the card game

Let us finally discuss the actual problem of our motivation. Assume, once again, that our sample space is defined as:

$$S = \{1\heartsuit, 2\heartsuit, \ldots, K\heartsuit, 1\clubsuit, 2\clubsuit, \ldots, K\spadesuit\}.$$

Our game states that we win if we pick a red card. There are 13 $\heartsuit$ and 13 $\diamondsuit$ cards in the game. This gives us a cardinality of 26 outcomes. Recall that in total, our sample space consists of 52 outcomes, that is $|S| = 52$. One might want to reason then that we have 26 favorable outcomes in a total of 52 outcomes...

A class at the University of Illinois at Urbana-Champaign is taught by three different professors. The number of students that took the class and the grades they received are shown in the following table.

Letter Grade	Professor 1	Professor 2	Professor 3	Total
A	108	20	30	158
B	44	49	46	139
C	11	15	15	41
D	0	1	8	9
Total	163	85	99	347

- How many students received an A in Professor 1's class?
- How many students were in Professor 1's class or got an A?
- Are the students who got an A and the students who got a B in Professor 1's class mutually exclusive events?
- How many students got an A but were not in Professor 1's class?

Back to Monopoly

In the beginning of this lecture, we only needed to roll a 6 or a 7 to "survive" another round (so to speak). Let us finish this lecture with the following thought process:

1. The sample space of rolling two dies is:
$$S = \{(1,1), (1,2), (1,3), \ldots, (1,6), (2,1), \ldots, (6,6)\}.$$
2. Counting all possible outcomes, we have that $S = |36|$.
3. The event "roll a 6" contains 5 outcomes:
$$\{(1,5), (2,4), (3,3), (4,2), (5,1)\}.$$
4. The event "roll a 7" contains 6 outcomes:
$$\{(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)\}.$$
5. "Roll a 6" and "Roll a 7" are mutually exclusive, hence we have a total of
$$5 + 6 = 11 \text{ favorable outcomes.}$$
6. One could again argue that our "chances" albeit small are not *that small* after all..

Counting

Learning objectives

After this lecture, we will be able to:

- Count how many outcomes satisfy an event.
- Recall the multiplication rule to count.
- Use the multiplication rule to count.
- Differentiate between permutations and combinations.
- Use permutations and combinations to count.
- Differentiate between different types of permutations.
- Interpret probabilities and recall fundamental probability properties.

Motivation: quantifying probabilities

When we discuss **probability**, there are two worldviews:

1. the frequentist view: which states that the probability of an event happening represents a relative frequency of the times the event happens versus all the times the random experiment is conducted ("in the long run").
2. the Bayesian view: which states that probability is a subjective measure of quantifying the likelihood of an event happening (as a "degree of belief").

Definition 1 (Probability) *With every event, we associate a real number called probability to represent the likelihood of that event happening. Probabilities satisfy three main rules* [13]*:*

1. $P(E) \geq 0$*, for any event* E*.*
2. *If an event* E *comprises the whole sample space (in which case, we write that* $E = S$*), then* $P(E) = 1$*.*
3. *If* $E_1, E_2, \ldots, E_m$ *are* m *mutually exclusive events* [14]*, then*

$$P\left(E_1 \cup E_2 \cup \ldots \cup E_m\right) = P\left(E_1\right) + P\left(E_2\right) + \ldots + P\left(E_m\right),$$

or even more concisely:

$$P\left(\bigcup_{i=1}^{m} E_i\right) = \sum_{i=1}^{m} P(E_i).$$

[13] Also known as the Kolmogorov axioms of probability.

[14] See the previous lecture.

From the definition, we can deduce that all probabilities are in $[0, 1]$, where a probability of 0 implies that an event can never happen, and a probability of 1 implies that an event is certain (will always happen).

Motivation: equally likely outcomes

When the outcomes in a discrete random experiment with sample space S are **equally probable**, we assume that the probability of each outcome happening is $\frac{1}{|S|}$. Hence, our question becomes:

> "how can we count all favorable outcomes and contrast them to all possible outcomes to derive a measure of probability?"

Why would that be useful?

Counting

The multiplication rule

In the previous lecture and worksheet, we fully enumerated all possible outcomes. For example, rolling two dies results in a total of 36 outcomes:

$$S = \{(1,1), (1,2), (1,3), \ldots, (1,6), (2,1), \ldots, (6,6)\}.$$

What happens if I need to find the cardinality of the sample space of rolling 10 dies?

A new burrito restaurant

In a new burrito place, you are allowed to choose *only one* of two types of tortillas (flour and wheat), *only one* of four types of "meats" (chicken, pork, steak, no meat), and *only one* of two types of beans (refried and black beans). A food critic needs to try one burrito every day until they have tried all possible burritos. How many days will they be visiting the restaurant to do that?

When our outcomes arise from a sequence of k steps, each of them with n_i, $i = 1, \ldots k$ options (i.e., n_1 options in step 1, n_2 options in step 2, and so on), then

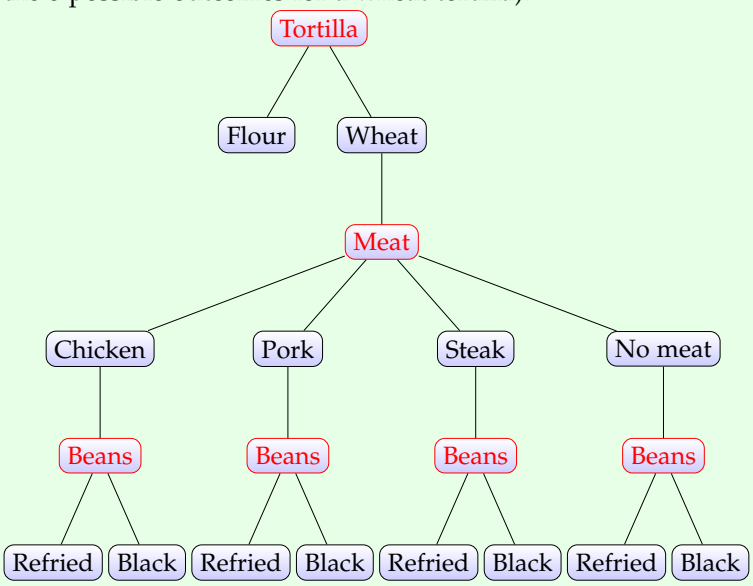
$$\text{the number of possible outcomes is } n_1 \cdot n_2 \cdot \ldots \cdot n_k.$$

Two key observations:

1. at each step i, we can have to pick exactly one of the n_i options.
2. the order does not matter.

A new burrito restaurant

Hence, in our burrito place example, we have 3 options (tortilla type, meat type, bean type), leading to a total of $2 \cdot 4 \cdot 2 = 16$ combinations (in the figure below, we show the 8 possible outcomes for a wheat tortilla).



In Greece, a vehicle is required to have a license plate with 3 letters (from the Greek alphabet!) and 4 numbers (integer numbers between 0 and 9). How many plates can there be, seeing as the Greek alphabet has 24 letters?

Permutations

A permutation is an **ordered** sequence of elements selected from some set. For example, consider the sample space $S = \{1, 2, 3\}$. All permutations are:

- $\{1, 2, 3\}$
- $\{1, 3, 2\}$
- $\{2, 1, 3\}$
- $\{2, 3, 1\}$
- $\{3, 1, 2\}$
- $\{3, 2, 1\}$

The number of permutations for a sample space with n possible outcomes is [15]

$$P_n = n!$$

[15] $n!$ is defined for any integer number as $n \cdot (n-1) \cdot (n-2) \cdot 1$. $n!$ is read as "n factorial". By definition, we say that $0! = 1$.

A random draw

5 people have been named the winners of a competition. There are 5 different books that will be given to them. How many different outcomes (assignments of winners to books) do we have?

There are $P_5 = 5! = 120$ possible outcomes (assignments of winners to books).

We can also opt to select only $r < n$ from the available elements in the set. For example, consider the set $S = \{1, 2, 3, 4\}$. The permutations of $r = 2$ elements from that set are:

- $\{1, 2\}$
- $\{2, 1\}$
- $\{3, 1\}$
- $\{4, 1\}$
- $\{1, 3\}$
- $\{2, 3\}$
- $\{3, 2\}$
- $\{4, 2\}$
- $\{1, 4\}$
- $\{2, 4\}$
- $\{3, 4\}$
- $\{4, 3\}$

The number of permutations of r outcomes from a total of n outcomes is:

$$P_{n,r} = n \cdot (n-1) \cdot \ldots \cdot (n-r) = \frac{n!}{(n-r)!}$$

A random draw (cont'd)

A total of 100 people are participating in a draw. 5 of the participants will be named winners and get one of 5 different books. How many different outcomes (assignments of winners to books) do we have now?
There are $P_{100,5} = \frac{100!}{(95)!} = 9034502400$ possible outcomes (assignments of winners to books).

Another type of permutation arises when some of the outcomes are the same (for example, two entries in a competition belonging to the same person). In that case, there are fewer **distinguishable permutations**. Formally, assume that:

- we have k different types of outcomes;
- n_1 outcomes of type 1, n_2 outcomes of type 2, $\ldots$, n_k ourcomes of type k;
- such that $n_1 + n_2 + \ldots + n_k = n$.

How many different permutations can we obtain? As an example, assume we are given the following 5 letters in Scrabble: $2 \times E$, $2 \times S$, $1 \times T$. Some of the possible possible 5-letter "words" we can create are:

- EESST
- EESTS
- EETSS
- ETESS
- TEESS
- SEEST
- SEETS
- SETES
- STEES
- TSEES
- ESEST
- ...

Why is this setup different than the previous permutations we discussed?

If we have k types of elements with n_i objects of type i ($i = 1, \ldots, k$) such that $\sum_{i=1}^{k} n_i = n$, then the number of distinguishable permutations is:

$$\frac{n!}{n_1! \cdot n_2! \cdot \ldots n_k!}.$$

A game of Scrabble

How many different 7-letter words (maybe nonsensical) can we create in a game of Scrabble, where we have $2 \times A, 1 \times B, 2 \times S, 1 \times T, 1 \times X$?
There are 5 different letters with $n_1 = 2, n_2 = 1, n_3 = 2, n_4 = 1, n_5 = 1$. Hence, the total number of distinguishable, 7-letter words we can create is:

$$\frac{7!}{2! \cdot 1! \cdot 2! \cdot 1! \cdot 1!} = 1260 \text{ words.}$$

Combinations

In all of our discussion so far, *order matters*. Often, though, we do not care about it. For example consider the cases of:

- creating a group of 4 people for a class project.
- checking the numbers on ten dies.
- picking the winning numbers in a lottery.

We define a **combination** as an unordered subset of $r < n$ outcomes selected from a sample spance with n outcomes. The number of all possible combinations is calculated by [16]:

$$C_{n,r} = \binom{n}{r} = \frac{n!}{r! \cdot (n-r)!}$$

[16] $\binom{n}{r}$ is also read as "n choose r".

Permutation or combination?

- Choosing 5 students out of 80 candidates to participate in a group?
- Choosing 5 students out of 80 for 5 specific and different positions in a group?
- Locating 10 different facilities in 10 cities in the USA?
- Locating 2 different headquarter facilities from 50 candidate cities in the USA?

Distinguishing between permutations and combinations

You have to select between 10 students for 3 positions. You are allowed pick the same student for all three positions, if you'd like.

- How many possible outcomes are there if the 3 positions are different?

  We need to use the multiplication rule $10 \cdot 10 \cdot 10 = 1000$ possible outcomes.

- What if you are not allowed to select the same student more than once? Assume again the 3 positions are different.

  We need to use a permutation (10 students for 3 different positions): $P_{10,3} = 720$ potential outcomes.

- What if all positions are actually for the same type of work?

  We now have a combination (10 students for 3 positions): $C_{10,3} = 120$ outcomes.

From counting to calculating probability

As mentioned in the Motivation subsection, **probability** is a real number between 0 and 1 that quantifies how likely an outcome (event) is. Adopting the frequentist view of probabilities [17] we could possibly count the number of outcomes that are favorable and divide by the total number of possible outcomes and thus estimate probability.

[17] The frequentist view states that the probability of an outcome is the relative frequency with which that outcome appears over all possible outcomes (see the Motivation subsection).

Quality control

A package is set to leave a factory and be sent to a retailer. The package contains 100 items. We already know that exactly 3 of the 100 items are defective. The quality control team over at the retailer works as follows: they select a sample of 6 items from the 100, and check them. If there are 0 defective items in the selected sample of 6, they accept the package and sell its contents; otherwise, they send the package back. What is the probability that the quality control rejects the package and sends it back?

To answer this question, we decompose the problem into its components. We need to know:

1. how many ways are there to select 6 items from the 100?
2. how many ways are there to have 1, 2, or 3 defective items in the 6 selected?

Let us begin by addressing the first question.

Quality control: How many ways are there to select 6 items from the 100 in the package?

Step 1: How many ways are there to select 6 items from the 100 in the package? This is a *combination* and we get:

$$C_{100,6} = \binom{100}{6} = \frac{100!}{6! \cdot 94!} = 1192052400 \text{ ways.}$$

For the second question, we need to think slightly differently. Let x be the number of defective items and $6 - x$ the number of non-defective items in the sample of 6. Then:

Quality control: How many ways are there to have 1, 2, or 3 items in the sample of 6 selected?

Step 2: How many ways are there to have 1, 2, or 3 defective items from the 3 available in the selected sample of 6? This is another *combination*, albeit requiring more calculations.

- **Step 2a:** How to select $x = 1$ defective items in the sample?

  We would need to pick 1 out of the 3 defective and 5 out of the 97 non-defective!

  $$C_{3,1} = \binom{3}{1} = 3.$$
  $$C_{97,5} = \binom{97}{5} = 64446024.$$

  We should now use the multiplication rule between the two, as we have to pick one option from the 3 possible options from the first selection and one option from the 64446024 possible ones in the second selection, for a total of

  $$3 \cdot 64446024 = 193338072 \text{ ways.}$$

- **Step 2b:** How to select $x = 2$ defective items in the sample?

  Similarly:

  $$C_{3,2} = \binom{3}{2} = 3.$$
  $$C_{97,4} = \binom{97}{4} = 3464840.$$

  The total is 10394520 ways.

- **Step 2c:** How to select $x = 3$ defective items in the sample?

  Similarly:

  $$C_{3,3} = \binom{3}{3} = 1.$$
  $$C_{97,3} = \binom{97}{3} = 147440.$$

  The total is 147440 ways.

To finish this example, we need to divide the number of desired outcomes (obtained in Step 2) to the total number of outcomes (obtained in Step 1). Note that we may add the three numbers from Step

2 to calculate the total number of desired outcomes [18].

Quality control: What is the probability?

Step 3: Let $E =$ fail inspection. Then:

$$P(E) = \frac{193338072 + 10394520 + 147440}{1192052400} = \frac{203880032}{1192052400} = 0.171.$$

You pick 3 cards at random from a deck with 52 cards. What is the probability that all 3 are face cards? What is the probability that 2 are face cards?

[18] We observe that the three events (selecting $x = 1$, $x = 2$, or $x = 3$ defective in the sample) are **mutually exclusive** and hence the cardinality of the union of the three events is equal to the summation of the individual cardinalities

Basic probability theory

Learning objectives

After this lecture, we will be able to:

- Recall and explain the basic properties that probability has.
- Calculate the probability of an event.
- Apply set operations in probability calculations.
- Define and provide examples of conditional probabilities.
- Apply the conditional probability formula.
- Recognize independence.

Motivation: Will I miss my flight?

Flying from Urbana-Champaign almost always requires a layover in another airport. For example, flying to New York City usually is done through Chicago with two legs: Urbana-Champaign to Chicago, and Chicago to New York City. My layover is only 45 minutes in Chicago, so I am naturally worried about making my connection. I would feel much better if I knew whether my first flight leaves on time or not. What is the probability that I make my second flight given that my first flight is delayed by 15 minutes or more?

Motivation: Data collection

A company has undertaken the large effort of contact tracing and testing for COVID-19 in the Urbana-Champaign area. It is expected that from the people that leave in the area, 1% has been in close contact with an already known case of COVID-19, 15% has been working in close contact with multiple people as an essential worker, and 6% has traveled to a location (in and outside the USA) with high risk of contagion. A random person is selected for a test, but will only be administered the test if they fall within one of the three categories above. What is the probability that the person will get the test?

Probabilities

Definition

As a continuation from the motivation in the previous lecture, there are two interpretations of probabilities:

1. relative frequency of favorable outcomes versus all outcomes.

2. subjective "degree of belief".

In both cases, we can use *basic probability theory* to calculate the likelihood of an event happening or not. Once again, recall the definition of probability:

Definition 2 (Probability) *With every event, we associate a real number called probability to represent the likelihood of that event happening. Probabilities satisfy three main rules* [19]*:*

[19] Also known as the Kolmogorov axioms of probability.

1. $P(E) \geq 0$, *for any event* E.
2. *If an event* E *comprises the whole sample space (in which case, we write that* $E = S$*), then* $P(E) = 1$.
3. *If* $E_1, E_2, \ldots, E_m$ *are* m *mutually exclusive events, then*

$$P\left(E_1 \cup E_2 \cup \ldots \cup E_m\right) = P\left(E_1\right) + P\left(E_2\right) + \ldots + P\left(E_m\right),$$

or even more concisely:

$$P\left(\bigcup_{i=1}^{m} E_i\right) = \sum_{i=1}^{m} P(E_i).$$

The first two axioms imply that probability is a real number in $[0, 1]$ [20], where 0 is an *impossible event* (one that can never happen) and 1 signals a *certain event* (one that will always happen) [21].

[20] We sometimes present probability as a percentage (%): for example, a probability of 0.4 can be also written as a probability of 40%.

[21] Consider a sample space S that consists of three events: A, B, C. Then the event that neither A nor B nor C happen is *impossible*; the event that A or B or C happens is *certain*.

Probabilities of mutually exclusive events

Assume that two events E_1, E_2 are mutually exclusive: for example, let E_1 be the event that you get an A in IE 300, and E_2 the probability that you get an $A-$. Your *personal belief* is that you have a 30% "chance" at an A and a 20% "chance" at an $A-$. Then, the probability that you get *at least an* $A-$ *in the class* is

$$P(\text{“at least an } A- \text{ in IE 300"}) = P(E_1) + P(E_2) = 50\%.$$

From the three laws of probability, we also deduce that:

- $P(\overline{E}) = 1 - P(E)$.
- $P(\emptyset) = 0$.
- If $E_1 \subseteq E_2$, then $P(E_1) \leq P(E_2)$.

Recall that we say that one event E_1 is contained in another event E_2 and write that $E_1 \subseteq E_2$ if all outcomes that satisfy E_1 are included in E_2.

When $E_1 \subseteq E_2$

A pizza store advertises delivery in 30 minutes or less. Assume that the probability of an order being delivered in 30 minutes or less is 0.9. Then:

- the probability of an order being delivered in 15 minutes or less is at most 0.9.
- the probability of an order being delivered in 1 hour or less is at least 0.9.

An email is categorized as one of the following 2 (mutually exclusive) categories: spam or not-spam. Emails that are not-spam are also categorized as one of 3 (mutually exclusive again) categories: urgent, normal priority, and advertisements. Answer the following questions.

- If the probability of a message being spam is 0.45, then the probability of a message being not-spam is 0.55. True or False?
- What is the probability that an urgent email is spam?
- An email is urgent with probability 0.1 and an email is of normal priority with probability 0.2. Which of the following cases is true?
  1. $P(\textit{not-spam}) < 0.3$.
  2. $P(\textit{not-spam}) = 0.3$.
  3. $P(\textit{not-spam}) \geq 0.3$.

Unions and intersubsections of events

For any two events E_1, E_2, define $E = E_1 \cup E_2$. We then have that $E_1 \subseteq E$ and $E_2 \subseteq E$, which leads to

$$\boxed{P(E_1) \leq P(E),\ \ P(E_2) \leq P(E)}$$

and

$$\boxed{P(E) \leq P(E_1) + P(E_2).}$$ [22]

[22] Recall that we already saw when the equality holds.

We can use a similar deduction for two events E_1, E_2 and $E = E_1 \cap E_2$. We have that $E \subseteq E_1$ and $E \subseteq E_2$, and get

$$\boxed{P(E) \leq P(E_1),\ \ P(E) \leq P(E_2).}$$

But how can we calculate $P(E_1 \cup E_2)$ exactly in the general case [23]? Let us turn back to sets and cardinalities.

[23] Recall again that if E_1 and E_2 are mutually exclusive, then $P(E_1 \cup E_2) = P(E_1) + P(E_2)$ by the third Kolmogorov axiom.

Deriving that $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$

We have already seen how $E_1 \cup E_2$ can be written as the union of three mutually exclusive events: $E_1 \setminus E_2$, $E_1 \cap E_2$, and $E_2 \setminus E_1$. From the third Kolmogorov axiom, we have that:

$$P(E_1 \cup E_2) = P(E_1 \setminus E_2) + P(E_1 \cap E_2) + P(E_2 \setminus E_1). \quad (1)$$

Now, we note that E_1 (and E_2, respectively) can also be written as the union of two mutually exclusive events: $E_1 = (E_1 \setminus E_2) \cup (E_1 \cap E_2)$ (and $E_2 = (E_2 \setminus E_1) \cup (E_1 \cap E_2)$, respectively). This gives that

$$P(E_1) = P(E_1 \setminus E_2) + P(E_1 \cap E_2) \implies$$
$$P(E_1 \setminus E_2) = P(E_1) - P(E_1 \cap E_2). \quad (2)$$

Combining (1) and (2) gives us that:

$$\boxed{P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2).}$$

Recall the grades from 3 different professors for the same class shown in a previous lecture.

Letter Grade	Professor 1	Professor 2	Professor 3	Total
A	108	20	30	158
B	44	49	46	139
C	11	15	15	41
D	0	1	8	9
Total	163	85	99	347

Assuming you call on one student out of the total 347 students, what is the probability:

1. E_1: you pick a student from Professor 1's class?
$$P(E_1) = 163/347 = 0.4697.$$
2. E_2: you pick a student who received an A in the class?
$$P(E_2) = 158/347 = 0.4553.$$
3. $E_1 \cap E_2$: you pick a student who was both in Professor 1's class and received an A in the class?
$$P(E_1 \cap E_2) = 108/347 = 0.3112.$$

How about the probability that you pick either a student from Professor 1's class or a student who received an A in the class?

Recall that there is a:

- 1% probability for a person to have been in close contact with a known COVID-19 case;
- 15% probability for a person to work as an essential worker;
- 6% probability that a person has traveled to a location with high contagion risk.

However, when estimating the probability that a person qualifies for the test, we have found that only 17% of the population does that. Based on your knowledge so far, does that make sense? How could that happen?

This last question should get us thinking about union of more than 2 events. For the next part see also the Worksheet for Lecture 3.

Deriving the probability of the union of more than 2 events

This will be filled after the actual lecture.

Deriving the probability of the union of more than 2 events (cont'd)

This will be filled after the actual lecture.

In general, for m events $E_1, E_2, \ldots, E_m$, we have:

1. Add the probabilities of the individual events.
2. Subtract the probabilities of the intersubsections of any two events.
3. Add the probabilities of the intersubsections of any three events.
4. Continue subtracting the probabilities of the intersubsections of any $4, 6, \ldots$ events and adding the probabilities of the intersubsections of any $5, 7, \ldots$, events.

Conditional probabilities

Motivation

It is common to want to recalculate our chances as more information become available or under certain conditions. For example, the probability that I miss the second leg of my flights is immediately affected by any delays I might experience the first leg of my flight. In such cases, we turn to *conditional probability*.

Definitions

Definition 3 (Conditional probability) *Conditional probability is defined as the probability that an event E_1 happens given that event E_2 has already happened: this is written as $P(E_1|E_2)$* [24].

[24] This is read as "the probability of E_1 given E_2" or "the probability of E_1 such that E_2 has happened".

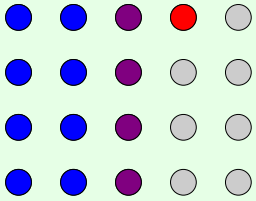
Let us think what we might need to calculate such a probability. Assume that events E_1 and E_2 (not necessarily mutually exclusive) are set to happen. We have already calculated that:

- $P(E_1) = 0.5$.
- $P(E_2) = 0.25$.
- $P(E_1 \cap E_2) = 0.2$.

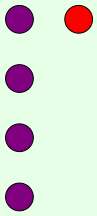
Based on the given probabilities, we can also deduce that $P(E_1 \cup E_2) = 0.5 + 0.25 - 0.2 = 0.55$, even though we do not need this result.

Changing our perception

The probability that E_1 happens is 0.5 (50%). Does this perception change if we are told that E_2 has already happened? Let us try to come up with a visual parallel to the provided probabilities. Here, we have 20 dots, out of which 12 (60%) are red and 5 (25%) are blue. 20% of them (4 in number) are both red and blue for a "purplish" color. Note that the four purple dots are both red and blue.



Had you known that E_2 has already happened, this leaves you with much fewer cases to consider!



Should our perception for the probability of E_1 change then?

Definition 4 (Conditional probability formula) *The conditional probability of one event E_1 conditional to event E_2 is calculated by*[25]

$$P(E_1|E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)}. \tag{3}$$

[25] Note that conditional probabilities are **only** for $P(E_2) > 0$.

What is $P(E_1|E_2)$ in the previous example? How about $P(E_2|E_1)$?

Finally, note that for two mutually exclusive events E_1, E_2, the definition of conditional probabilities certainly implies that

$$P(E_2|E_1) = 0 \text{ and } P(E_1|E_2) = 0.$$

The multiplication rule for probabilities

A very straightforward rewriting of the conditional probability formula gives us a very important result. Solving for the numerator of

the right hand side in (3) gives us that for any two events A, B:

$$P(A \cap B) = P(A|B) \cdot P(B). \tag{4}$$

This will come quite in handy in the next lecture.

Independence

We typically say that two entities are independent if actions of one are completely unaffected (and do not themselves affect) the actions of the other. In probability theory, we say that two events are **independent events** if knowledge that one has happened (or not) does not affect our perception for the probability of the other.

In mathematical terms, we say the following.

Definition 5 (Independent events) *Two events E_1, E_2 are independent if we have that:*

$$P(E_2|E_1) = P(E_2) \;\; \textit{and} \;\; P(E_1|E_2) = P(E_1).$$

Equivalently, we may write that two events E_1, E_2 are independent if

$$P(E_1 \cap E_2) = P(E_1) \cdot P(E_2).$$

Think of two events from real life that are independent. Also think of two events that are clearly dependent.

Two events are mutually exclusive. Are they independent?

Bayes' theorem

Learning objectives

After this lecture, we will be able to:

- Recall and explain the law of total probability.
- Use the law of total probability to calculate probabilities.
- Formulate Bayes' theorem.
- Describe Bayes' theorem.
- Explain what Bayes' theorem implies for probabilities.
- Apply Bayes' theorem in calculating probabilities.

Motivation: The Mantoux test

The Mantoux (sometimes called the Mendel–Mantoux) test is a diagnostic tool for tuberculosis (TB). In the test, a dosage of tuberculin units is injected: some time later, the reaction on the skin is measured and a positive or negative reaction is given. It is assumed that about 0.05% of the children in the world have TB. The test is pretty accurate, with 99% success rate – that is a person with TB receives a positive result 99% of the time, and a person without TB receives a negative result 99% of the time.

The Mantoux test is mandatory in most European countries schools. A random kid did the test, which came up positive. Are you 99% certain the kid has TB?

Motivation: Pilot season

Studios typically make decisions on shows based on a single episode made early on, called a "pilot". This pilot episode is viewed by a carefully selected audience who then reports either favorable or unfavorable reviews. A show is considered highly successful, moderately successful, or unsuccessful depending on its performance while on air.

Historically, 95% of highly successful shows received favorable reviews, 50% of moderately successful shows received favorable reviews, and 10% of unsuccessful shows received favorable reviews.

You are one of the producers of a new TV show, and are showing the pilot episode to a major studio. The audience loved it and gave generally favorable reviews. Will your show definitely be a huge success?

The law of total probability

Motivation

The Spring 2020 semester saw a rapid change of plans for most university courses due to the pandemic. Students were left needing to make a decision about selecting credit or no credit for their classes. Let's focus on one particular case.

Credit/no credit or graded?

A UIUC course requires students to end up with an average of 70 or above to qualify for credit, whereas an average of 60 is enough to qualify for a passing grade. A student believes they will end up with a score between 65 and 80 – so they are definitely passing the class – but they are thinking of opting for the credit/no credit option. Unfortunately, the class is missing two important grades: the final project and a final (non-cumulative) exam.

We are commonly facing problems like this in every day life. Decision-making under uncertainty revolves around us making decisions where the outcomes are not guaranteed. In such cases, the decision-maker *weighs the different futures* and aims to quantify the probability of a favorable outcome. Let's revisit the student from the example.

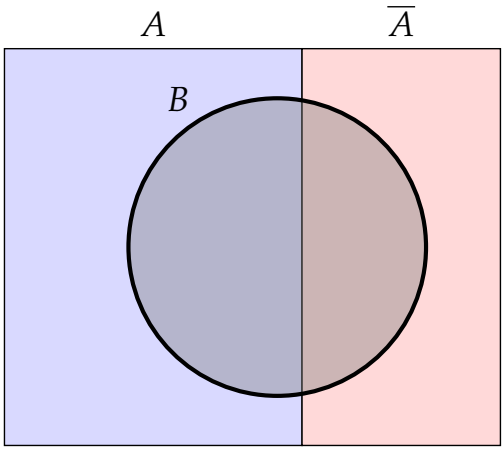
Credit/no credit or graded?

The student has been quite enjoying the material of the final exam and they are optimistic that they will score very highly. They believe that they will end up with a score of 90/100 with probability 50% or a score of 80/100 with probability 50%. They are not as confident for the final project, where they believe they received either a 50/100, a 60/100, or a 70/100 (with probability 30%, 60%, and 10%, respectively). So, say, they go ahead and put all these eventualities in a table.

A table can be used to keep track of all of the events whose outcomes are uncertain. In the case of the student, they would have to be enumerate a total of 6 cases (why? [26]), which are presented next.

[26] Remember counting!

Figure 1: Two events A and B. A is represented by the blue area, whereas B contains all outcomes in the circle.



Credit/no credit or graded?

	Final project	Final exam	Final grade
Scenario 1	50	80	65
Scenario 2	50	90	68
Scenario 3	60	80	69
Scenario 4	60	90	72
Scenario 5	70	80	73
Scenario 6	70	90	76

If the student picks credit/no credit, what is the probability they do not receive credit?

Derivation

Consider two events A and B, marked below as the blue area of the rectangle and the circle in the middle, respectively. We also mark the complement of A in the figure.

Say, we are interested in the probability of B happening. We can present this as a function of A as follows:

From the second figure, we observe that B can be written as a union of two mutually exclusive events as in

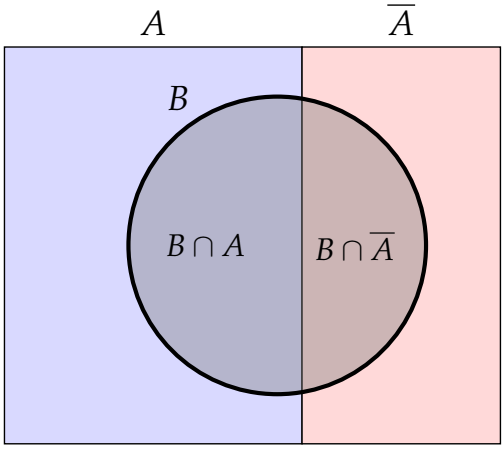
$$\begin{aligned} B &= (B \cap A) \cup \left(B \cup \overline{A}\right) \implies \\ P(B) &= P\left(B \cap A\right) + P\left(B \cap \overline{A}\right). \end{aligned} \tag{5}$$

Finally, we recall here that $P(A \cap B) = P(A) \cdot P(B|A)$ [27], which we can replace in (5) to get:

$$P(B) = P(A) \cdot P(B|A) + P(\overline{A}) \cdot P(B|\overline{A}). \tag{6}$$

[27] The multiplication rule we saw during our previous lecture.

Figure 2: Marking the two (mutually exclusive) parts of B. It is true that $B = (B \cap A) \cup (B \cap \overline{A})$.



This is the law of total probability for two events.

> Two urns contain red and blue balls. The first urn contains 3 red and 3 blue balls, while the second urn contains 5 red and 2 blue balls. We pick one ball at random from the first urn and (without seeing its color) place it in the second urn. What is the probability we pick a blue ball from the second urn?

Use and interpretation

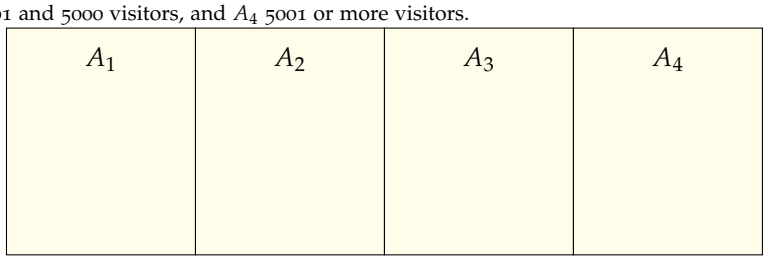
The law of totally probability can be generalized to more than 2 states. Assume we have m mutually exclusive and *collectively exhaustive* events. With the term *collectively exhaustive* we mean events whose union is the whole sample space. Formally:

Definition 6 (Collectively exhaustive events) *Let S be the sample space, and let $A_i, i = 1, \ldots, m$ be some events. Then, events A_i are collectively exhaustive if $\cup_{i=1}^{m} A_i = S$.*

Definition 7 (Mutually exclusive and collectively exhaustive events) *Let S be the sample space, and let $A_i, i = 1, \ldots, m$ be some events. Then, events A_i are mutually exclusive and collectively exhaustive if $\cup_{i=1}^{m} A_i = S$ and $A_i \cap A_j = \emptyset$ for any two sets $A_i, A_j, i \neq j$.*

An example of a series of mutually exclusive and collectively exhaustive events is given in Figure 3, where $S = A_1 \cup A_2 \cup A_3 \cup A_4$ and, as is shown, $A_1 \cap A_2 = A_1 \cap A_3 = A_1 \cap A_4 = A_2 \cap A_3 = A_2 \cap A_4 = A_3 \cap A_4 = \emptyset$.

Figure 3: Four collectively exhaustive and mutually exclusive events. For example, they could represent numbers of unique website visitors in a given day. A_1 could then be up to 1000 visitors, A_2 could represent between 1001 and 3000 visitors, A_3 between 3001 and 5000 visitors, and A_4 5001 or more visitors.



Credit or no credit?

In the student example, the final project can be viewed as three collectively exhaustive and mutually exclusive events, since the student can only have received a score of 50, 60, or 70. On the other hand, the final exam score has two collectively exhaustive and mutually exclusive events (a score of 80 or 90).

In the case of $m > 2$ mutually exclusive and collectively exhaustive events, the law of total probability becomes:

$$P(B) = P(A_1) \cdot P(B|A_1) + P(A_2) \cdot P(B|A_2) + \ldots + P(A_m) \cdot P(B|A_m) = \\ = \sum_{i=1}^{m} P(A_i) \cdot P(B|A_i). \tag{7}$$

Credit or no credit?

Let's go back to the table of scenarios the student had prepared. Let us rewrite the eventualities:

- If they receive a 50 in the final project, then they definitely do not get credit.
- If they receive a 60 in the final project, then they have a 50% of getting credit.
- If they receive a 70 in the final project, then they definitely get credit.

Let A_1, A_2, A_3 be the events of getting a 50, 60, or 70 in the final project and let C be the event of receiving credit in the class. Then, in probability terms, we have:

$$\begin{aligned} P(C) &= P(A_1) \cdot P(C|A_1) + P(A_2) \cdot P(C|A_2) + P(A_3) \cdot P(C|A_3) = \\ &= 0.3 \cdot 0 + 0.6 \cdot 0.5 + 0.1 \cdot 1 = \\ &= 0.4. \end{aligned}$$

You have just booked a two-leg (two-flight) trip. The flights are very close to one another and you are worried you will miss your second flight in the case of a delay. If the first flight is not delayed (which happens 60% of the time), you will be certainly fine (won't miss the flight); if the first flight is delayed up to 30 minutes (which happens 20% of the time), you might miss the second flight with probability 50%; finally, if the first flight is delayed by more than 30 minutes (which happens 20% of the time), you will definitely miss the second flight. What is the probability you miss the second flight?

Bayes' theorem

States of the world

Consider the following paradigm. You wake up in the morning, and unbeknownst to you the worlds is at a certain state. Let's call this state "good" or "bad". In a "good" world, 90% of everyone you talk to is happy and smiling and welcoming. In a "bad" world, only 5% of the people you talk to are happy and smiling and welcoming. Unfortunately, you have no idea which state the world is in today. What could you do to find out?

The above paradigm, as far-fetched as it sounds, applies in multi-

ple aspects of our life. A student could have studied and can answer a multiple choice question correctly, or could have gotten lucky and could give the correct answer by chance. A diagnostic test could come back positive, and this could mean that the patient is indeed positive, or it could be a mistake (referred to as a false positive). Even worse, a diagnostic test could come back negative, when the patient is unfortunately positive (this is called a false negative). In all the above cases, there is a "state of the world" that we are querying through tests, whose outcomes we read.

States versus outcomes

We contrast states to outcomes as follows. We consider that a state is fleeting and unknowable; hence, we perform a test and make an observation of its outcome. However, the outcome of the test does not necessarily reveal the state, as no test is perfect.

The Mantoux test

In the Motivation subsection, we saw the Mantoux test. The states of the world (unknowable for certain) are whether a kid has TB or not. The test here is the Mantoux test. Its outcomes are positive or negative.

We state two key observations:

- **The test outcome is not equivalent to the state**. A positive Mantoux test does not always mean a person with TB. A low score in a test does not always mean a student who did not study. A good review does not necessarily mean you will like a movie.
- **Looking for something rate, we will encounter many false positives**. Think of what happens when searching in a vast desert for an oasis. Most times, the oasis is a mirage.

A two-state example

We present a two-state example, adapted by Daniel Kahneman's "Thinking, Fast and Slow" book.

Farmer or librarian?

You sit next to someone in a flight and you start talking. The person tells you that they are from the USA, they discuss with you how much they enjoy reading books in their free time, and that they enjoy learning about other cultures. They then ask you: "We've been talking for a while. Guess what my occupation is. Do you think I work as a librarian or as a farmer?"

Our mind can create some connections based on what we know and what we *think we know*. We know that there are more farmers than librarians in the USA (roughly 3 million farmers compared to 300,000 librarians). We also *think we know* that librarians probably enjoy books and learning about other cultures. We may jump to a conclusion, but if we do the math, we will see that our mind can rely too much on prior beliefs rather than context.

In summary: we formulate a hypothesis (for the state of the world) and, then, given evidence (outcomes of a test) that we leverage, we check to see if we are right or wrong.

Definitions

Let's collect here some definitions and notations that will be useful throughout the derivation of *Bayes' theorem*:

- $S_i, i = 1, \ldots, n$: n states of the world, which are mutually exclusive and collectively exhaustive.
- $O_j, j = 1, \ldots, m$: m outcomes of a test we administer trying to understand the true state of the world.

We also need to define our beliefs for what the state of the world is. These are called *prior probabilities*, as they reflect prior beliefs and biases [28] (before we see the outcomes of the test):

- $P(S_i)$: prior probability of state S_i.

Additionally, we define *likelihood probabilities* that represent the probability we see a certain outcome of the test when the world is in a certain state [29]:

- $P(O_j|S_i)$: likelihood probability of seeing outcome O_j given that we are in state S_i.

Moreover, we have already established that $P(O_j \cap S_i)$ is the probability that we both experience outcome O_j and we are in state S_i. This is called a *joint probability*:

[28] Examples include the probability that a random student has studied or not, or the probability that a random person is a librarian or a farmer.

[29] Examples include the probability that a student does well in an exam given that they have studied or that a person works as a librarian given that they enjoy reading books

- $P(O_j \cap S_i)$: joint probability of both seeing outcome O_j and we are in state S_i.

Finally, we may calculate the probability that a random test returns a certain outcome O_j. This is referred to as a *marginal probability* [30]:

- $P(O_j)$: marginal probability of seeing outcome O_j.

Don't lose sight of what we are searching for! This would be $P(S_i|O_j)$: the *posterior probability* of being in state S_i given that we observed outcome O_j in the test.

We may now proceed to the derivation of Bayes' theorem.

[30] An example would be the probability of an "A" in an exam, or the probability of a positive Mantoux test.

Derivation and use

We break down the derivation in steps.

Derivation

Step 1: conditional probabilities. What can you say about $P(S_i|O_j)$?

$$P(S_i|O_j) = \frac{P(S_i \cap O_j)}{P(O_j)}. \tag{8}$$

Step 2: joint probability. Which other conditional probability employs $P(S_i \cap O_j)$?

$$P(O_j|S_i) = \frac{P(S_i \cap O_j)}{P(S_i)} \implies P(S_i \cap O_j) = P(S_i) \cdot P(O_j|S_i). \tag{9}$$

Replacing (9) into the numerator of the right hand side in (8), we get:

$$P(S_i|O_j) = \frac{P(S_i) \cdot P(O_j|S_i)}{P(O_j)}. \tag{10}$$

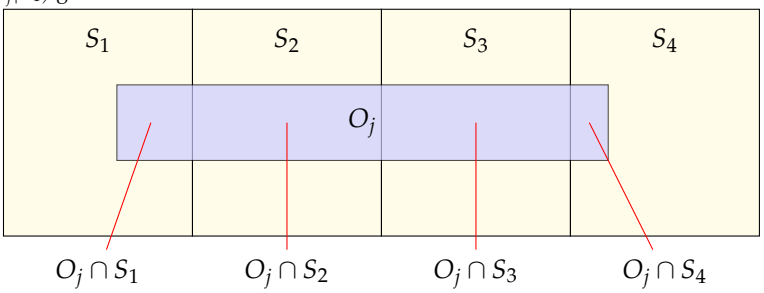
Step 3: marginal probability. What can we say for $P(O_j)$? Let's go back to the multiplication rule:

$$P(O_j) = \sum_{i=1}^{n} P(S_i) \cdot P(O_j|S_i) \tag{11}$$

Replacing (11) in the denominator of the right hand side in (10) gives us Bayes' theorem:

$$\boxed{P(S_i|O_j) = \frac{P(S_i) \cdot P(O_j|S_i)}{\sum_{i=1}^{n} P(S_i) \cdot P(O_j|S_i)}.}$$

Figure 4: Consider 4 mutually exclusive and collectively exhaustive states S_1, S_2, S_3, S_4. When outcome O_j happens, note how the probabilities of each state change. For example, given O_j, we have $P(S_1|O_j) = \frac{P(O_j \cap S_1)}{P(O_j)}$. Replacing $P(O_j \cap S_1)$ by $P(S_1) \cdot P(O_j|S_1)$ and $P(O_j)$ by $P(S_1) \cdot P(O_j|S_1) + P(S_2) \cdot P(O_j|S_2) + P(S_3) \cdot P(O_j|S_3) + P(S_4) \cdot P(O_j|S_4)$ gives the result.



Bayes' theorem states that the posterior probability $P(S_i|O_j)$ depends on our prior probabilities $P(S_i)$ and our joint probabilities $P(O_j|S_i)$.

$$\boxed{P(S_i|O_j) = \frac{P(S_i) \cdot P(O_j|S_i)}{\sum_{i=1}^{n} P(S_i) \cdot P(O_j|S_i)}.}$$

For a visual representation of Bayes' theorem, check Figure 4.

The Mantoux test

Going back to the Mantoux test, let's fill in the information we need to answer the question: "what is the probability a kid has TB given that the test came back positive?"

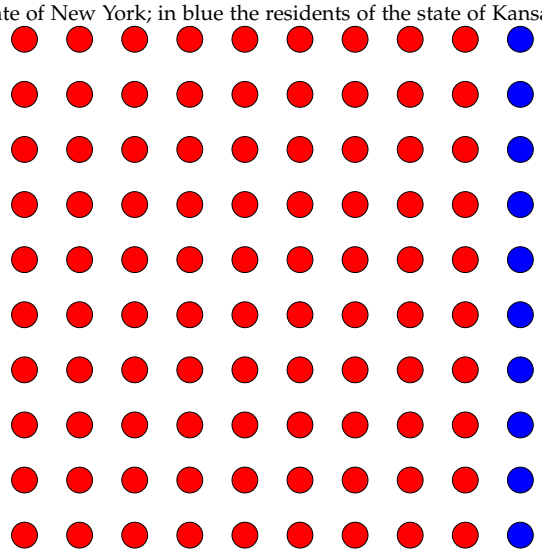
- S_1: kid has TB; S_2: kid does not have TB.
- O_1: positive Mantoux test; O_2: negative Mantoux test.
- $P(S_1) = 0.0005$; $P(S_2) = 0.9995$.
- $P(O_1|S_1) = 0.99$; $P(O_1|S_2) = 0.01$; $P(O_2|S_1) = 0.01$; $P(O_2|S_2) = 0.99$.

Using the Bayes' theorem, we have:

$$P(S_1|O_1) = \frac{P(S_1) \cdot P(O_1|S_1)}{P(S_1) \cdot P(O_1|S_1) + P(S_2) \cdot P(O_1|S_2)} = \\ = \frac{0.0005 \cdot 0.99}{0.0005 \cdot 0.99 + 0.9995 \cdot 0.01} = 0.0472.$$

We deduce that a positive Mantoux test implies a 4.72% chance of actually having TB.

Figure 5: A visual representation of the population picked. In red, we have the residents of the state of New York; in blue the residents of the state of Kansas.



Answer the probability question in the pilot episode example of the motivation.

Another visual representation of Bayes' theorem

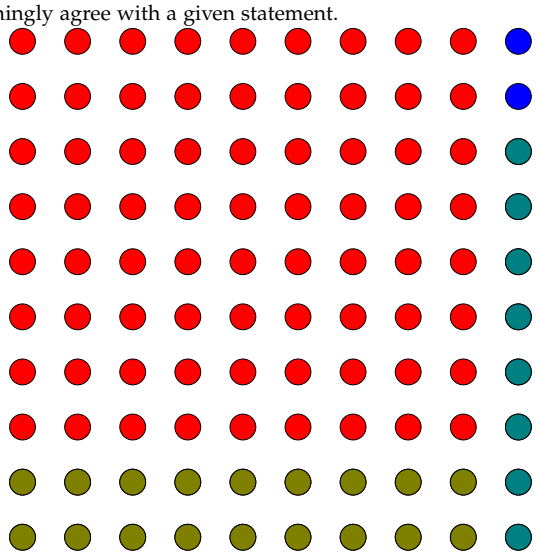
Assume that we get a representative population from two states: Kansas and New York. Seeing as Kansas is about 9 times smaller than New York, we pick 90 people from New York and 10 from Kansas, represented pictorially in Figure 5.

Statistically 20% of the population of New York works in an agriculture-related job. The same percentage is 80% for Kansas. Without loss of generality, we show that with the following Figure where farmers are shaded in green (light green for New York, dark green for Kansas).

Finally, assume the person next to you is flying from Kansas to New York (and has made it clear that they are either from Kansas or New York) and works in a farm. While your original bias may be that the person has to be in Kansas (look at the percentage of agriculture-related jobs for the Kansas population!), Bayes' theorem states that the probability is only $8/26 = 0.31$. Formally:

$$P(Kansas|farmer) = \frac{P(Kansas \cap farmer)}{P(farmer)} = \frac{0.1 \cdot 0.8}{0.1 \cdot 0.8 + 0.9 \cdot 0.2} = 0.31.$$

Figure 6: A visual representation of 100 registered voters. The voters in shades of green (darker green for red, lighter green for blue) are voters from Party A and Party B that overwhelmingly agree with a given statement.



Discrete random variables

Learning objectives

After these lectures, we will be able to:

- Define discrete and continuous random variables.
- Differentiate between discrete and continuous random variables.
- Differentiate between when to use cumulative distribution functions and probability mass functions.
- Give examples of at least four different discrete distributions.
- Recall when to and how to use:
  - binomially distributed random variables.
  - geometric distributed random variables.
  - hypergeometric distributed random variables.
  - Poisson distributed random variables.

Motivation: 2-engine vs. 4-engine aircraft

Suppose that for a flight to be completed successfully (which we would really *love*) we need at least half of the engines to be operational at the end of the trip. In the case of a 2-engine aircraft, this means at least one; in a 4-engine aircraft, we'd need at least two.

We are ordering engines from a production company and we would like to see whether buying two (for a 2-engine plane) or four (for a 4-engine plane). Which one would be safer?

Motivation: Big in Japan

Over the last 135 years, there have been 5 earthquakes of seismic intensity over 7.0 in the Kanto region of Japan. However, especially for those of us living in seismogenic zones [31], we probably grew up hearing statements such as "the probability of an earthquake in X within the next Y years is Z%". Hey, this is what we do in this class!

[31] An area with high seismic/earthquake activity.

We make the following assumptions for big earthquakes:

1. Big earthquakes are independent events – that is the fact that a big earthquake happened does not increase or decrease the probability of another big earthquake soon.
2. The probability of an earthquake occurring is the same throughout the year [32].

[32] This is a property also called homogeneity. More on that later.

What is the probability that there will be one big earthquake in the Kanto region in the next year? What is the probability that there will be one big earthquake in the Kanto region in the next decade?

Random variables

The world around us is a series of random processes, whose outcomes affect the way we perceive things. Mathematically, we need to somehow define these outcomes – using numerical representations.

Definition 8 (Random variables) *With the term random variable* [33] *we mean a* ***real-valued function*** *defined over the sample space.*

[33] Also termed random quantities or stochastic variables.

Definition 9 (Random variables) *A random variable is a function that associates a number with each element of the sample space.*

Classification

In the next few lectures, we will separate our discussion between discrete and continuous random variables.

- Discrete random variables take countable, discrete values. [34]
- Continuous random variables can take any real-value. [35]

[34] For example, the side of a die, the number of customers.

[35] For example, the time until the next bus arrives, the lifetime of a light bulb, the pressure of a gas.

Recall that we had made that distinction in the past for random experiments and their sample spaces! [36]

[36] See Lecture 1.

Classify these random variables as discrete or continuous:

1. the time it takes for a biker to go from one side of campus to the other.
2. the number of red lights the biker has to stop at when going from one side of the campus to the other.
3. the distance a biker traverses to go from one side of campus to the other.
4. the number of times the biker changed speed gear while going from one side of campus to the other.

Functions

When a random variable behaves a specific way, we say that it follows a **probability distribution**. A probability distribution is typically described by two distribution functions:

- The **probability mass function** for discrete random variables or **probability density function** for continuous random variables.
- The **cumulative distribution function**.

We formally define those where needed for discrete and continuous random variables.

The remainder for our lecture notes is devoted to discrete random variables. See Lectures 7 and 8 for a discussion on continuous random variables.

Discrete random variables

Let X be a **discrete** random variable.

Definition 10 *We define the* **probability mass function** *(pmf)* $p(x)$ [37] *of a discrete random variable* X *as the probability that it takes a specific value* x*:*

$$p(x) = P(X = x).$$

[37] Some textbooks may use $f(x)$ for the probability mass function. In our notes, we will use $p(x)$ for discrete random variables and their probability mass functions.

One thing we need to be very careful with:

- Distinguish between X (upper case X) and x (lower case x)! X is the random variables; x is a given value. [38]

[38] Example: we can write $P(X = 7)$ and read "what is the probability that *random variable* X is equal to the *value* 7"?

For example, the question "what is the probability that a store has **exactly** 20 customers enter in the next hour?" can be addressed using

the **probability mass function** as follows. First, let X be a random variable that represents the number of customers that enter the store in the next hour. Then, express the probability as $P(X = 20)$.

Definition 11 *We define the* **cumulative distribution function** *(cdf) $F(x)$ of a discrete random variable as the probability that it takes* up to *a value x, i.e.,*

$$F(x) = P(X \leq x) = \sum_{y:y \leq x} P(X = y) = \sum_{y:y \leq x} p(y).$$

For example, the question "what is the probability that a store has **up to** 20 customers enter in the next hour?" can be addressed using the **cumulative distribution function** as follows. First, let X be a random variable that represents the number of customers that enter the store in the next hour. Then, express the probability as

$$\begin{aligned} F(20) &= P(X \leq 20) = \\ &= P(X = 0) + P(X = 1) + \ldots + P(X = 20) = \\ &= \sum_{y \leq 20} P(X = y). \end{aligned}$$

An immediate result from the definition of the cdf is that if we are interested in the probability of seeing more than a certain value x we may write:

$$P(X > x) = 1 - P(X \leq x) = 1 - F(x).$$

Combining the two definitions (of $P(X \leq x)$ and $P(X > x)$) we get that the probability of X taking values between a and b is [39]:

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

[39] This is easy to work out. It is left as an exercise to the reader.

Defining these two functions helps us classify random variables based on their properties, as we will spend the rest of the lectures finding out.

2-engine vs. 4-engine

We saw that a plane performs a trip safely if at least half of its engines are operational. Let X be the number of engines that have failed during a trip. Then, we should be looking for:

- $P(X \leq 1)$: for the probability of a successful trip with a 2-engine aircraft.
- $P(X \leq 2)$: for the probability of a successful trip with a 4-engine aircraft.

Earthquake probabilities

In a similar fashion, let X be the number of earthquakes in the Kanto region in the next year and Y be the same number in the next decade. Then, we should be looking for:

- $P(X = 1)$: for the probability of one big earthquake in the next year.
- $P(Y = 1)$: for the probability of one big earthquake in the next decade.

Should you use the pmf or the cdf?

- To avoid paying your friends in a game of Monopoly you need to get a 6 or less when throwing two dies.
- An exam has 10 multiple choice questions. What is the probability you answer all of them correctly?
- An exam has 10 multiple choice questions. What is the probability you answer more than or equal to 8 questions correctly?
- Two people are playing a game that is best out of three. What is the probability the first player wins with a score of 2 to 1?

Assume a discrete random variable X with n outcomes x_i, $i = 1, \ldots, n$. Then, the probability mass function $p(x)$ of random variable X has to satisfy the following three rules:

1. $p(x_i) = P(X = x_i)$, for every outcome $x_i, i = 1, \ldots, n$.
2. $p(x_i) \geq 0$.
3. $\sum_{i=1}^{n} p(x_i) = 1$

Urns and balls

Two balls are drawn from an urn containing 5 red and 4 black balls. Define a random variable X as the number of red balls drawn. What is its probability mass function?

- X has three outcomes: $0, 1, 2$.
- To select 0 red balls: $p(0) = P(X = 0) = \frac{C_{4,2}}{C_{9,2}} = \frac{6}{36}$
- To select 1 red ball: $p(1) = P(X = 1) = \frac{C_{4,1} \cdot C_{5,1}}{C_{9,2}} = \frac{20}{36}$
- Finally, for 2 red balls: $p(2) = P(X = 2) = \frac{C_{5,2}}{C_{9,2}} = \frac{10}{36}$

We may verify that these probabilities satisfy all three rules of a valid probability mass function.

Urns and balls

Two balls are drawn from an urn containing 5 red and 4 black balls. Define a random variable X as the number of red balls drawn. What is its probability mass function?

- X has three outcomes: $0, 1, 2$.
- To select 0 red balls: $p(0) = P(X = 0) = \frac{C_{4,2}}{C_{9,2}} = \frac{6}{36}$
- To select 1 red ball: $p(1) = P(X = 1) = \frac{C_{4,1} \cdot C_{5,1}}{C_{9,2}} = \frac{20}{36}$
- Finally, for 2 red balls: $p(2) = P(X = 2) = \frac{C_{5,2}}{C_{9,2}} = \frac{10}{36}$

We may verify that these probabilities satisfy all three rules of a valid probability mass function.

Calculating probabilities

A sample space is described by two mutually exclusive outcomes A and B. We have observed that for some real number x, the pmf is $P(A) = 3 \cdot x$ and $P(B) = 10 \cdot x^2$. What is x?
This type of question needs us to use the pmf rules. We need to verify what x should be in order to satisfy $P(A), P(B) \geq 0$ and $P(A) + P(B) = 1$. Replacing the pmf in the second equality we have:

$$\begin{aligned} P(A) + P(B) = 1 &\implies 3 \cdot x + 10 \cdot x^2 = 1 \\ &\implies 10 \cdot x^2 + 3 \cdot x - 1 = 0 \implies \\ &\implies x = \begin{cases} \text{-0.5} \\ \\ 0.2 \end{cases} \end{aligned}$$

Replacing $x = -0.5$, we get that $P(B) = 2.5$, but $P(A) = -1.5 < 0$. Replacing $x = 0.2$, we get that $P(A) = 0.6$ and $P(A) = 0.4$, and is the correct answer.

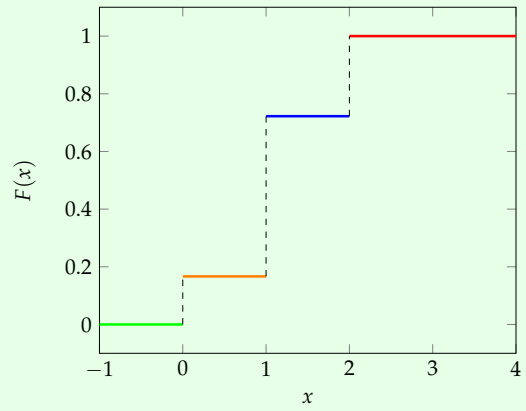
The cumulative distribution function (cdf) of a discrete random variable X needs to satisfy in turn two rules:

1. $0 \leq F(x) \leq 1$.
2. If $x \leq y$, then $F(x) \leq F(y)$.

Urns and balls

Consider the previous sample space $0, 1, 2$ with $p(0) = \frac{6}{36}, p(1) = \frac{20}{36}, p(2) = \frac{10}{36}$. Then:

$$F(x) = \begin{cases} 0, & x < 0 \\ \frac{6}{36}, & 0 \le x < 1 \\ \frac{26}{36}, & 1 \le x < 2 \\ 1, & x \ge 2. \end{cases}$$



Let the sample space be $S = \{1, 2, 3\}$ with $p(1) = \frac{1}{2}, p(2) = \frac{1}{3}, p(3) = \frac{1}{6}$.

- Verify this is a valid pmf.
- Write the cdf $F(x)$.
- Draw the cdf (like we showed in the previous example).

The binomial distribution

Before we get to the **binomial distribution**, we need to introduce **Bernoulli random variables**. This first random variable we will introduce is also (probably) the simplest!

Consider a single experiment that has only two probable outcomes:

1. **success** which happens with probability p; and
2. **failure** which happens with probability $q = 1 - p$.

Now, define a random variable X based on that single experiment:

$$X = \begin{cases} 0, & \text{if the experiment failed;} \\ 1, & \text{if the experiment succeeded.} \end{cases}$$

The key is that we consider only **one** experiment [40]. For the Bernoulli distribution, we have:

$$\text{pmf:} \quad \begin{aligned} P(X=0) &= q = 1-p \\ P(X=1) &= p \end{aligned}$$

$$\text{cdf: } F(x) = \begin{cases} 0, & x < 0 \\ 1-p, & x < 1 \\ 1, & x \geq 1 \end{cases}$$

[40] Will the next coin toss be a heads (success) or a tail (failure)? Will it rain (success) or not (failure)? Will my favorite NBA team win (success) its next game or not (failure)? Will the next patient be cured (success) or not (failure)?

Urns and balls

An urn contains 40 black and 10 red balls. You pick at random one ball from the urn. Let X be the number of black balls you pick from the urn. What is the pmf of X?

X is a Bernoulli distributed random variable with pmf:

- $P(X=0) = \frac{10}{50} = 0.2$.
- $P(X=1) = \frac{40}{50} = 0.8$.

What if we consider more than one experiments? What if, say, we picked 5 balls and wanted to get 3 black ones [41] This is where **binomially distributed random variables** come in play! The setup is simple:

[41] Multiple experiments could mean multiple coin tosses, or a control group of 100 patients, or a best-of-five game series!

- n independent experiments/trials.
- each experiments ends up in a success with probability p and a failure with probability $1-p$;
  - that is, each trial is a Bernoulli random variable.
- Let X be the number of successes.

Then X is a binomial random variable. We may also write that $X = binom(n, p)$ as n (number of experiments) and p (probability of success in each individual experiment) are the only necessary parameters to fully define this random variable.

Coin tosses

The most common example to explain binomial random variables comes from coin tosses. Assume we possess a "fair" coin with probability of Heads $p = 0.5$, and probability of Tails $q = 1-p = 0.5$? What is the probability that there will be exactly 2 Heads in $n = 3$ tosses of the coin? This would be a binomial distribution with $n = 3$, $p = 0.5, q = 0.5$, and $x = 2$ Heads.

The formula for calculating the probability for binomially distributed random variables is: [42]

$$\boxed{p(x) = P(X = x) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x}, \quad \text{for } x = 0, 1, \ldots, n.}$$

Recall that $\binom{n}{x} = \frac{n!}{x! \cdot (n-x)!}$ as we had seen in a previous lecture. [43]

[42] The derivation of the formula is part of Lecture 5's worksheet.

[43] See Lecture 2.

Coin tosses

We may now address the earlier question:

$$p(2) = \binom{3}{2} \cdot 0.5^2 \cdot (1-0.5)^{3-2} = \frac{3!}{2! \cdot 1!} \cdot 0.25 \cdot 0.5 = 0.375.$$

2-engine vs. 4-engine

In our motivational example, let p be the probability that an engine fails during a trip, and hence $q = 1 - p$ is the probability it does not fail. For the success of each plane, we have:

2-engine:

$$\begin{aligned} P(X \leq 1) &= P(X = 0) + P(X = 1) = \\ &= \binom{2}{0} \cdot p^0 \cdot (1-p)^2 + \binom{2}{1} \cdot p^1 \cdot (1-p)^1 = \\ &= 1 - 2p + p^2 + 2 \cdot p - 2 \cdot p^2 = 1 - p^2 \end{aligned}$$

4-engine:

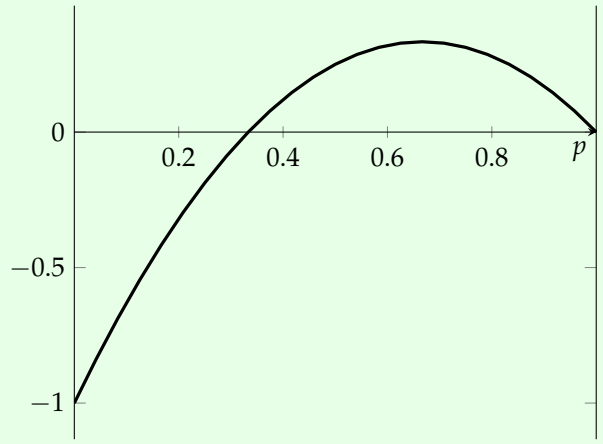
$$\begin{aligned} &P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) = \\ &= \binom{4}{0} \cdot p^0 \cdot (1-p)^4 + \binom{4}{1} \cdot p^1 \cdot (1-p)^3 + \binom{4}{2} \cdot p^2 \cdot (1-p)^2 \\ &= 1 - 4 \cdot p + 6 \cdot p^2 - 4 \cdot p^3 + p^4 + 4 \cdot p - 12 \cdot p^2 + 12 \cdot p^3 - 4 \cdot p^4 \\ &+ 6 \cdot p^2 - 12 \cdot p^3 + 6 \cdot p^4 = 1 + 3 \cdot p^4 - 4 \cdot p^3 \end{aligned}$$

Can we compare the two? To prefer a 2-engine plane we need its probability of success to be higher, that is we need:

$$\begin{aligned} 1 - p^2 \geq 1 + 3 \cdot p^4 - 4 \cdot p^3 \implies -3 \cdot p^4 + 4 \cdot p^3 - p^2 \geq 0 \implies \\ \implies -3 \cdot p^2 + 4 \cdot p - 1 \geq 0. \end{aligned}$$

Let's plot $y = -3 \cdot p^2 + 4 \cdot p - 1$ and see what we get! Since for $y \geq 0$ we prefer a 2-engine aircraft, it suffices to see when y is nonnegative in the plot!

2-engine vs. 4-engine (cont'd)



We observe that for probability of engine failure $p \geq \frac{1}{3}$, then a 2-engine plane is favored!

An urn contains 40 black and 10 red balls. You pick at random one ball from the urn, check its color, and after checking its color, you put it back in the urn. Let X be the number of black balls you pick from the urn in $n = 10$ tries. What is the probability that $X = 6$? What is the probability that $X \geq 9$?

Food for thought: why was it important in the previous example to put the ball back in the urn? What changes if I remove it from the urn?

The geometric distribution

Let's look at another extension of Bernoulli random variables. Earlier, during our discussion for binomially distributed random variables, we cared about the number of successes in a series of trials. How about the **first success** though? When did it occur? Since we are talking about a series of experiments, this first success can occur at the first, second, third, and so on, try.

Coin tosses

Assume again we are in possession of a fair coin. What is the probability the first Heads appears after three tries?

In general, we have that the probability that the first success is seen after exactly x trials is: [44]

$$P(X = x) = (1 - p)^{x-1} \cdot p.$$

[44] The derivation of this formula is also, albeit easier, part of Lecture 5's worksheet.

Learning basketball

A kid learning basketball is shooting free throws with a probability of scoring equal to 25%. What is the probability the kid has to shoot four free throws until scoring the first one?

The number of free throws until the first one is scored X is a geometric random variable with $p = 0.25$ and $x = 4$, hence we have

$$P(X = 4) = (1 - 0.25)^3 \cdot 0.25 = 0.75^3 \cdot 0.25 = 0.1055.$$

Assume we have a fair coin. What is the probability..

- the first Heads appears in the 2nd toss?
- the first Heads appears in the 5th toss?
- the first Heads appears in the 10th toss?

The hypergeometric distribution

What if..

- we had N items;
- $K \leq N$ of them are successes (the remaining $N - K$ are failures);
- we drew n of them;
- what is the probability we get k successes in the sample of n?

We have actually dealt with this problem before. Recall the example from Lecture 2:

Quality control

A package is set to leave a factory and be sent to a retailer. The package contains 100 items. We already know that exactly 3 of the 100 items are defective. The quality control team over at the retailer works as follows: they select a sample of 6 items from the 100, and check them. If there are 0 defective items in the selected sample of 6, they accept the package and sell its contents; otherwise, they send the package back. What is the probability that the quality control rejects the package and sends it back?

The answer to this probability was $0.171 = 17.1\%$. Now, check how this problem matches the setup of the hypergeometric distribution:

$N = 100$ total population size; $K = 3$ defective ones; $n = 6$ sample size. If X is the number of defective items picked in the sample, then the pmf for the hypergeometric is:

$$\text{pmf: } P(X = x) = \frac{\binom{K}{x} \cdot \binom{N-K}{n-x}}{\binom{N}{n}}$$

> An urn contains 40 black and 10 red balls. You pick at random a sample of five balls from the urn. Let X be the number of black balls in the sample. What is the probability that $X = 3$?

The big difference between the binomial and the hypergeometric distribution is in the **sampling with replacement** [45] and the **sampling without replacement** [46] The main difference is that with replacement, the probability of picking an item stays the same throughout the experiment, no matter how many times it is repeated; without replacement, the probability changes with every selection.

[45] For example, taking 5 balls from the urn one-by-one, looking at each one's color, and putting it back in, before picking the next; or selecting 5 items from a box one-by-one, checking it, and placing it back in again before picking the next.

[46] For example, taking 5 balls from the urn at the same time and looking at their colors together; or selecting 5 items from a box at the same time, checking them and seeing if they are defective.

The Poisson distribution

We do the opposite of what we normally do: we will motivate the Poisson distribution from a mathematical perspective instead of through an example. Recall the binomial distribution and its probability mass function:

$$p(x) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x}.$$

Assume that we try way too many experiments (in essence let $n \to \infty$) and define λ as the number of successes we get: that is, $p = \frac{\lambda}{n}$. Let's replace this in the probability mass function itself:

$$\begin{aligned} p(x) &= \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x} \\ &= \frac{n!}{x! \cdot (n-x)!} \cdot \left(\frac{\lambda}{n}\right)^x \cdot \left(1 - \frac{\lambda}{n}\right)^{n-x}. \end{aligned}$$

Let us now employ some of the cool limit properties that we know as $n \to \infty$!

$$\begin{aligned} \lim_{n\to\infty} P(X = x) &= \lim_{n\to\infty} \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x} = \\ &= \lim_{n\to\infty} \frac{n!}{x! \cdot (n-x)!} \cdot \left(\frac{\lambda}{n}\right)^x \cdot \left(1 - \frac{\lambda}{n}\right)^{n-x} = \\ &= \lim_{n\to\infty} \frac{\lambda^x}{x!} \cdot \underbrace{\frac{n!}{(n-x)! \cdot n^x}}_{1} \cdot \underbrace{\left(1 - \frac{\lambda}{n}\right)^n}_{e^{-\lambda}} \cdot \underbrace{\left(1 - \frac{\lambda}{n}\right)^{-x}}_{1} = \\ &= e^{-\lambda} \frac{\lambda^x}{x!} \end{aligned}$$

Definition 12 *A discrete random variable X taking values $0, 1, 2, \ldots$ is a Poisson random variable with parameter (rate) $\lambda > 0$ if:*

$$p(x) = P(X = x) = e^{-\lambda}\frac{\lambda^x}{x!}.$$

Poisson random variables have a wide, *wide* array of applications. They have been used to model:

- the number of phone calls that a call center gets every day.
- the number of shark attacks in California every year.
- the number of home runs in a baseball series.
- the number of patients arriving in an emergency department every night.
- the number of website requests per second.
- the number of earthquakes expected to hit a seismogenic area every decade.

As can be seen from the examples, Poisson distributed random variables are commonly used to model the number of events that happen in a given interval. Poisson distributed random variables need to satisfy three main conditions:

1. independence: an event happening should not affect the rate with which more events happen.
2. homogeneity: the rate with which events happen is constant.
3. no two events can occur at *exactly* the same time. Instead there is a small interval of time that separates two consecutive events.

Earthquake probabilities

We will model our motivating example with predicting the probability of an earthquake in the Kanto region of Japan using the Poisson distribution. We need to estimate λ, the rate of events. From the data, we are told that there have been 5 big earthquakes over the last 135 years, and hence:

$$\lambda = \frac{5}{135} = 0.037 \text{ earthquakes per year.}$$

We are interested in:

- $P(X = 1)$: for the probability of one big earthquake in the next year.

$$P(X = 1) = e^{-0.037} \cdot \frac{0.037^1}{1!} = 0.0357 = 3.57\%.$$

When interested in the probability of one big earthquake over the next decade:

- $P(Y = 1)$: for the probability of one big earthquake in the next decade.

Here we need to adapt the rate to accommodate periods of 10 years. Hence, $\lambda = 0.37$ earthquakes per year. Finally:

$$P(Y = 1) = e^{-0.37} \cdot \frac{0.37^1}{1!} = 0.2556 = 25.56\%.$$

Finally, let's address the probability that there is *at least one* big earthquake in the next decade:

$$P(Y \geq 1) = 1 - P(Y = 0) = 1 - e^{-0.37} \cdot \frac{0.37^0}{0!} =$$
$$= 1 - 0.6907 = 0.3093 = 30.93\%.$$

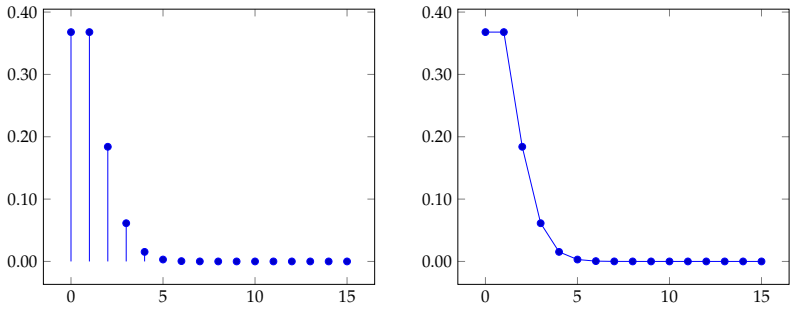
Is it fair to assume that typos appearing in notes follow a Poisson distribution? Why/Why not?

Assume typos appear in my notes following a Poisson distribution with a rate of $\lambda = 0.5/page$. What is the probability that no typos exist in the first page? What is the probability that there exist more than 1 typo in the first 10 pages?

Plotting the Poisson distribution also proves an interesting endeavor. Let's remember that all distributions we are discussing are

discrete: hence, we will simply plot each point and then connect the points with a line. For example, in Figure 7, we show the case for $\lambda = 1$ and how we would connect the different data points.

Figure 7: The Poisson distribution for $\lambda = 1$.



We do the same for $\lambda = 2, 5, 10$ in Figures 8, 9, 10.

Figure 8: The Poisson distribution for $\lambda = 2$.

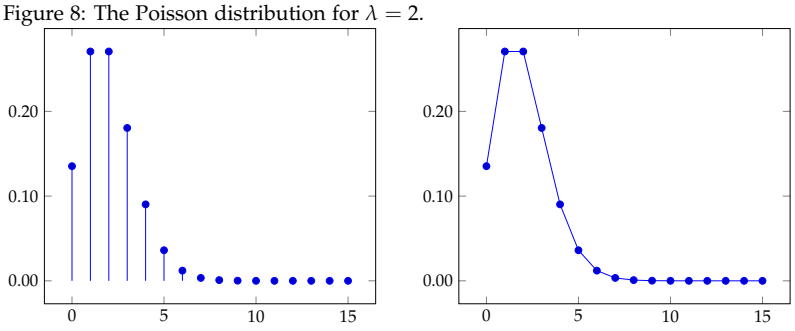
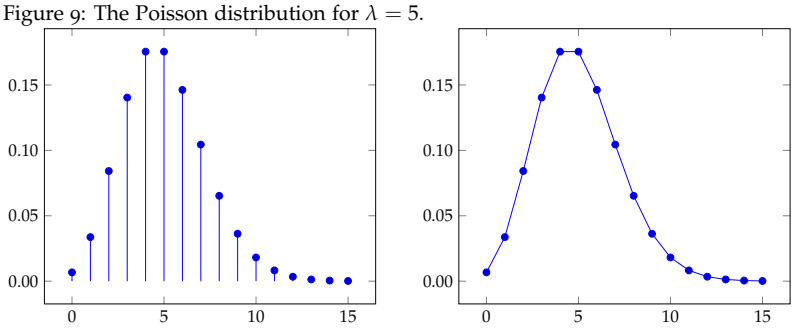


Figure 9: The Poisson distribution for $\lambda = 5$.



Finally, take a look at Figure 11. See what happens when λ takes on very big values...

The uniform distribution

Finally, we see the simplest discrete distribution, the uniform distribution. Think of a discrete random variable with n different outcomes $x_i, i = 1, \ldots, n$. Now assume that:

Figure 10: The Poisson distribution for $\lambda = 10$.

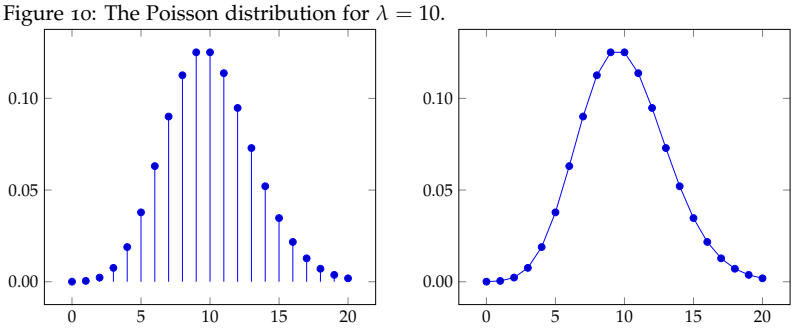
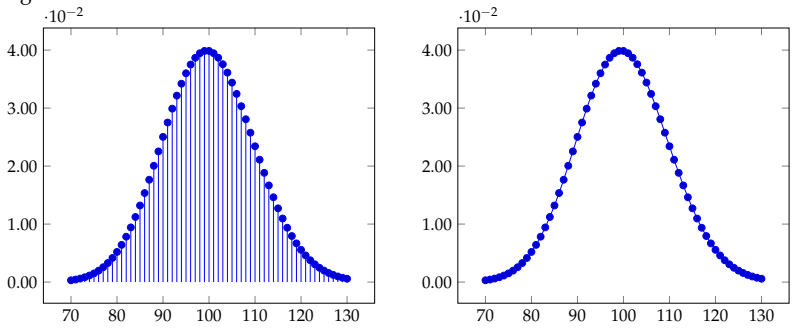


Figure 11: The Poisson distribution for $\lambda = 100$.



- all n outcomes are equally probable, then we have a uniform random variable.
- each of the outcomes is equally probable, i.e., $p_i = P(X = x_i) = \frac{1}{n}$.

In a special case, the discrete random variable take *integer* values in $[a, b]$. In that case, the pmf is

$$p_i = \frac{1}{b - a + 1}, \text{ for all } i = a, a + 1, \ldots, b.$$

A diamond cutting facility

A demanding customer has shown up in a diamond cutting facility and has asked for 2 *custom-made fine-cut diamond castings*. They are willing to buy 2 of those, as long as they are of high quality.. Diamond cutting is an expensive process, but you can make a lot of money out of it, and hence decide to take on the order. You plan to buy enough material for $Q = 4$ castings, just to be safe. Assuming that diamond cutting is a purely random process and all outcomes (producing $x = 0, 1, \ldots, Q$ high quality diamonds) are equally probable. What is the probability you satisfy your customer?

We need $x \geq 2$ high-quality fine-cut castings. The number of high-quality castings produced follows a uniform distribution, so:

$$P(X = x) = \frac{1}{Q+1} = \frac{1}{5}.$$

Hence, to satisfy the customer we have a probability of:

$$P(X \geq 2) = P(X = 2) + P(X = 3) + P(X = 4) = \frac{3}{5}.$$

Summary

In Table 4, we provide all of our results from Lectures 5 and 6. One could simply refer to these (and the keyword at the end of the page) for all information about discrete probability distributions.

Table 1: A summary of all results from Lectures 5 and 6.

Name	**Parameters**	**Values**	**pmf**
Bernoulli	$0 < p < 1$	$\{0, 1\}$	$p(0) = 1 - p$ $p(1) = p$
Binomial	$0 < p < 1$, $n \geq 0$	$\{0, 1, \ldots, n\}$	$p(x) = \binom{n}{x} p^x \cdot (1-p)^{n-x}$
Geometric	$0 < p < 1$	$\{1, 2, \ldots\}$	$p(x) = (1-p)^{x-1} \cdot p$
Hypergeometric	$N, K, n \geq 0$	$\{0, 1, \ldots, n\}$	$p(x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$
Poisson	$\lambda > 0$	$\{0, 1, \ldots\}$	$p(x) = e^{-\lambda}\frac{\lambda^x}{x!}$
Uniform	-	$[a, b]$	$p(x) = \frac{1}{b-a+1}$

Some keywords that might help you narrow down your search.

Bernoulli: "one single experiment/trial"; "success/failure"; "p and $q = 1 - p$".

Binomial: "multiple experiments/trials"; "success/failure"; "probabilities stay the same from experiment to experiment"; "how many successes in n tries?"; "with replacement".

Geometric: "number of experiments/trials until first success"; "success/failure"; "probabilities stay the same from experiment to experiment".

Hypergeometric: "sample"; "success/failure"; "without replacement"; "how many successes in a sample of size n?".

Poisson: "rate of events"; "number of events in an interval".

Uniform: "equally probable"; "outcomes are integer in $[a, b]$".

Continuous random variables: part 1

Learning objectives

After these lectures, we will be able to:

- Calculate probabilities of continuous random variables using their probability distribution and cumulative distribution functions.
- Give examples of uniform and exponentially distributed random variables.
- Recall when to and how to use:
  - uniformly distributed random variables.
  - exponentially distributed random variables.
- Define the memorylessness property and apply it exponentially distributed random variables.
- Use Poisson random variables and exponential random variables and provide examples of their relationship.

Motivation: continuous vs. discrete random variables

Guess which number I am thinking between 0 and 10 is a tricky proposition. If asked to do so in integer numbers (that is, 0 or 1 or 2...) then it is difficult to guess correctly, but not nearly impossible: we'd get a probability of 1 over 11 or a little more than 9%. On the other hand, when asked to do so with *any* number...

Motivation: Big in Japan

In an earlier example, we discussed the probability of seeing a certain number of earthquakes in the Kanto region of Tokyo in Japan. What if though we are interested in the timing of the next earthquake? How would we go about modeling this using continuous random variables?

Continuous random variables

Let X be a **continuous** random variable. Recall here that a continuous random variable is allowed to take any **real value** within some interval, say in $[a, b]$. Hence, there is an *infinite* number of possible outcomes associated with each continuous random variable!

Definition 13 *We define the* **probability distribution function** *(pdf)* $f(x)$ [47] *of a continuous random variable X as the "relative likelihood" that*

[47] Contrast with the definition of a probability mass function (pmf) $p(x)$ of a discrete random variable here...

X will be equal to a specific value x. This definition is a little open-ended, so we will address it more carefully shortly.

We again need to be careful with one item here:

- The actual probability that a continuous random variable X is exactly equal to some value x is 0! [48]

[48] Surprised?

This last note probably changes the way we need to discuss continuous probabilities. What if, instead of asking for the probability that continuous random variable X is exactly equal to some value x, we focus on the probability that continuous random variable X *belongs to some interval* of values?

Continuous random variables

Instead of the probability that:

- the average temperature tomorrow is exactly 78.3 Fahrenheit;
- the next bus passes in exactly 3 minutes and 25 seconds;
- the error of an ammeter (used to measure the current in a circuit) is exactly 0.1 A;

we may ask for the probability that:

- the average temperature tomorrow is between 78 and 79 Fahrenheit;
- the next bus passes between 3 and 4 minutes from now;
- the error of an ammeter is within 0.1 A.

This gives rise to the need for defining and using the **cumulative distribution function**. First, though, let us provide a different definition for the probability density function of a continuous random variable X.

Definition 14 *A random variable is* **continuous** *if it can take uncountably many values such that there exists some function $f(x)$ called a* **probability density function** *defined over real values $(-\infty, +\infty)$ such that:*

- $f(x) \geq 0$;
- $\int_{-\infty}^{+\infty} f(x)dx = 1$;
- $P(X \in B) = \int_{B} f(x)dx$.

The last property essentially states that to find the probability that a random variable X belongs to some interval B, then we need to take the integral of the probability density function of X, $f(x)$, over the interval B.

Continuous random variables

What is the probability that random variable X with pdf $f(x)$ is between 0 and 10?

$$P(0 \leq X \leq 10) = \int_0^{10} f(x)dx.$$

Due to the continuous nature of random variable X, and due to the fact that $P(X = x) = 0$, for any value x, we also get that:

$$P(0 \leq X \leq 10) = P(0 < X < 10) = P(0 < X \leq 10) = P(0 \leq X < 10).$$

Assume that continuous random variable X is distributed with probability density function $f(x)$ in $[0, \infty)$. What is:

a) the probability that X is between 2 and 5?

b) the probability that X is below 5 or above 10?

c) the probability that X is exactly equal to 5?

Definition 15 *We define the* **cumulative distribution function** *of a continuous random variable as the probability that it takes* up to *a value a, i.e.,*

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

By definition, $F'(x) = f(x)$: the derivative of the cdf gives us the pdf. Moreover, what we observed for discrete random variables is also true here and $P(a \leq X \leq b) = F(b) - F(a)$.

Timing chemical reactions

The time until a chemical reaction is over is measured in milliseconds (ms). The probability that the reaction is over by time x is given by the following cdf:

$$F(x) = \begin{cases} 0, & x < 0 \\ 1 - e^{-0.01x}, & x \geq 0. \end{cases}$$

Answer the following questions.

1. What is the pdf?
2. What proportion of chemical reactions are performed in less than or equal to 200 ms?
3. What proportion of chemical reactions take more than or equal to 100 ms and less than or equal to 200 ms?

1. For the pdf, by definition we have that

$$f(x) = F'(x) = \begin{cases} 0 \\ \left(1 - e^{-0.01x}\right)' \end{cases} = \begin{cases} 0, & x < 0 \\ 0.01 \cdot e^{-0.01x}, & x \geq 0 \end{cases}$$

2. We need to calculate $F(200)$:

$$P(x \leq 200) = F(200) = 1 - e^{-2} = 0.8647.$$

3. We now need $P(100 \leq x \leq 200)$:

$$P(100 \leq x \leq 200) = \int_{100}^{200} f(x)dx = \int_{100}^{200} 0.01 \cdot e^{-0.01x}dx = \\ = e^{-1} - e^{-2} = 0.2325.$$

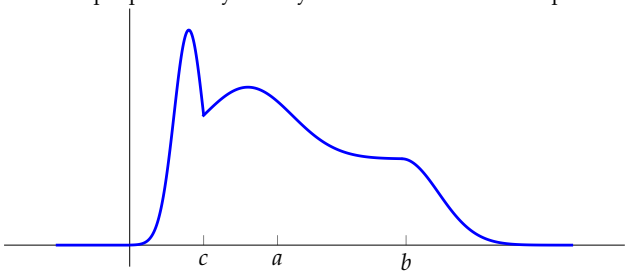
To answer this last part, we could have also used the fact that $P(a \leq X \leq b) = F(b) - F(a)$:

$$P(100 \leq x \leq 200) = F(200) - F(100) = 0.2325.$$

We may also represent the cdf visually. If we plot $f(x)$ (the pdf), then the cdf is the area under the curve. For example, consider the $f(x)$ plotted in Figure 12. It could be a valid pdf as it satisfies $f(x) \geq 0$, and it also can be shown to satisfy $\int_{-\infty}^{+\infty} f(x)dx = 1$, even though it would be impossible to do so without knowing the exact function.

That said, we may observe that it is equal to 0 for small enough and large enough numbers, which indicates that the integral from minus to plus infinity is equal to a finite number (i.e., the area under the curve is a finite number).

Figure 12: The example probability density function we have come up with here.



Then, in Figures 13 and 14, we show what the cdf appears to be visually.

Figure 13: How to visually represent the cumulative probability distribution $F(x)$ (here, we specifically present $F(c)$).

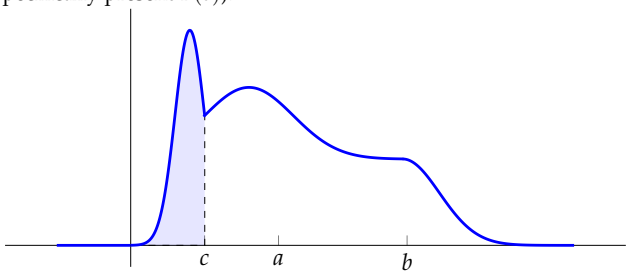
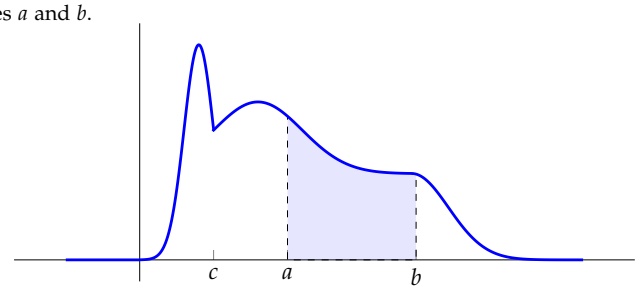


Figure 14: How to visually represent the probability that a random variable is between two values a and b.



Since we discussed the validity of a pdf, let's see an example of how we could use that.

Valid pdf?

Assume a continuous random variable taking values between 0 and 10 with a pdf of $f(x) = c \cdot x$. What is c?

First of all, c has to be nonnegative ($c \geq 0$), otherwise $f(x)$ may become negative, which is not allowed. We then employ the fact that $\int_{-\infty}^{+\infty} f(x)dx = 1$:

$$\int_{-\infty}^{+\infty} f(x)dx = 1 \implies \int_{0}^{10} cxdx = 1 \implies c \cdot \frac{x^2}{2}\Big|_0^{10} = 1 \implies c = 0.02.$$

Are the following valid pdfs?

- $f(x) = 0.01, 0 \leq x \leq 100$?
- $f(x) = \lambda \cdot e^{-\lambda \cdot x}, x \geq 0$?
- $f(x) = \lambda \cdot e^{-\mu \cdot x}, x \geq 0$ if we are told that $\lambda \neq \mu$?

The uniform distribution

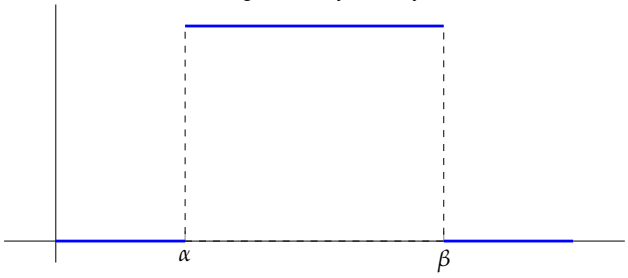
We begin this time from the simplest continuous distribution, the uniform distribution. In essence, it mimics its discrete counterpart, where everything is equally likely. However, since we are discussing continuous random variables, this implies that all values of $f(x)$ are equal, having equal relative likelihood. Its pdf and cdf are shown next.

$$f(x) = \begin{cases} \frac{1}{\beta - \alpha}, & \text{if } \alpha \leq x \leq \beta \\ 0, & \text{otherwise.} \end{cases}$$

$$F(x) = \int_{-\infty}^{x} f(y)dy = \begin{cases} 0, & \text{if } x < \alpha \\ \frac{x-\alpha}{\beta-\alpha}, & \text{if } \alpha \leq x \leq \beta \\ 1, & \text{if } x > \beta \end{cases}$$

Visually, the uniform distribution is presented in Figure 15.

Figure 15: The uniform distribution probability density function.



Totally random buses

Assume you are told that the next bus will arrive at any point in the next 5 to 15 minutes. Hence, in this case, the time until the next bus shows up is uniformly distributed. Then, what is the probability the bus arrives before:

a) 2′? b) 7′? c) 10′? d) 18′?

Note that here we have that $\alpha = 5$, $\beta = 15$. Thus:

a) 2′: $x = 2 < \alpha \implies F(2) = 0$.

b) 7′: $x = 7 \implies F(7) = \frac{7-5}{15-5} = 0.2$.

c) 10′: $x = 10 \implies F(10) = \frac{10-5}{15-5} = 0.5$.

d) 18′: $x = 18 > \beta \implies F(18) = 1$.

We visually present the probability for the bus to arrive in the next 10 minutes as an area under the curve in Figure 16.

Figure 16: The area under the curve for the probability of the bus arriving in 10′ from the example. The area is marked in green. We can tell that it is half the total area under the curve of the pdf, and hence corresponds to a probability of 50%.

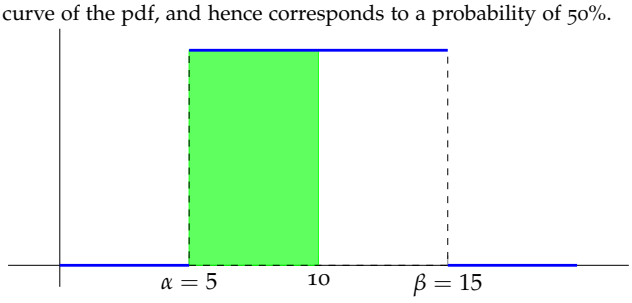
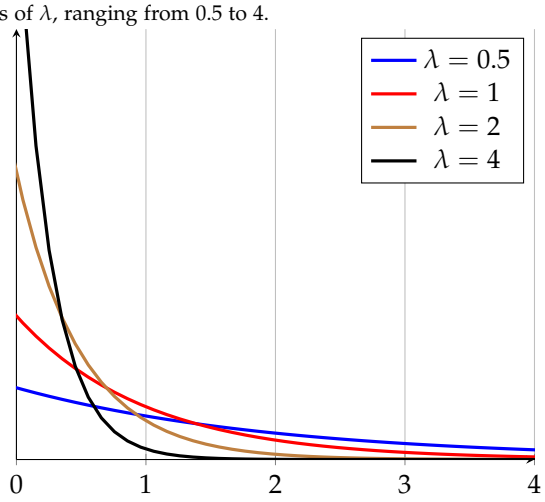


Figure 17: The exponential distribution probability density function visualized for different values of λ, ranging from 0.5 to 4.



The exponential distribution

It is time to move to one of the most important and consequential probability distributions. The exponential distribution takes its name from the fact that it is based on the exponential function. This is shown when considering its pdf and cdf [49]:

[49] In this lecture's worksheet, you are asked to derive $F(x)$, so brush up on your integration skills!

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

$$F(x) = \begin{cases} 1 - e^{-\lambda x}, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0. \end{cases}$$

We also present the two functions visually in Figures 21 and 18. Formally, the exponential distribution is defined as in Definition 16.

Definition 16 (The exponential distribution) *A continuous random variable X defined over the interval of $[0, \infty)$ is exponentially distributed if it has probability density function given by*

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

where $\lambda > 0$ is a parameter. We sometimes write that $X \sim Exp(\lambda)$ if it follows the exponential distribution with rate λ. [50]

[50] Where have we seen *rates* before?

One of the many applications that the exponential distribution sees in practice has to do with quantifying the probability of **the time until the next event**. When events happen with some rate λ, we can quantify the risk or chance that the next event will happen

Figure 18: The exponential distribution cumulative density function visualized for different values of λ, ranging from 0.5 to 4.

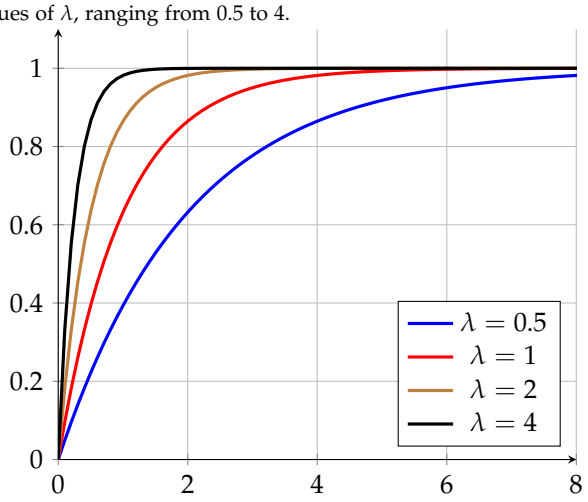


Figure 19: Taken from *The Simpsons*.

Figure 20: Taken from *The Office*.

within some time interval using the exponential distribution. Let us motivate this better with an example.

No accident in x days

How many days until the next accident? In many real-life cases, we assume that the **time to the next event** follows an exponential distribution. Assume that you work in a facility that has typically a rate of $\lambda = 2$ major accidents per year.

What is the probability the next accident happens in the next year? What is the probability that the next accident happens in the next 1 month?

Before we answer the question, we need to address the déjà vue feeling we may be experiencing. **We have seen** this family of questions before! When dealing with the **Poisson random variables**, we were talking about rates and about the probability of having a certain number of events within some time interval. The relationships do not stop here: both distributions make use of the exponential function, and both distributions rely on rates $\lambda > 0$. This brings us to the next subsubsection: how are Poisson random variables and exponentially

distributed random variables related?

The exponential distribution and the Poisson distribution

We have already motivated the fact that these two appear to be "sibling" distributions. Let us go back to a question we addressed in a previous worksheet. [51] We repeat this here for convenience.

[51] Recall Lecture 6 Worksheet and, specifically, Problem 10.

Lecture 6 worksheet: Problem 10 repeat

We saw in class the probability mass function for a Poisson distributed random variable with rate λ. Assume that $\lambda = 3$ per year. What is the probability that there will be no events in the next year? Can you say that this means that the next event will happen more than a year from now? Let T be the time of the next event: what is $P(T > 1 \text{ year})$?

Let X be the number of events during the next year. Then, we have that:

$$P(T > 1 \text{ year}) = P(X = 0) = e^{-\lambda} \cdot \lambda^0/0! = e^{-3} = 0.05.$$

Note that we could also find the probability that the next event does happen during the next year:

$$P(T \leq 1 \text{ year}) = 1 - P(X = 0) = 1 - e^{-\lambda} \cdot \lambda^0/0! = 1 - e^{-3} = 0.95.$$

Let us put the previous result in perspective. The time to the next event is exponentially distributed if the number of events is distributed as a Poisson random variable! The full relationship between the exponential and the Poisson distributions is presented in tabular form in Table 2.

Table 2: The relationship between an exponentially distributed and a Poisson distributed random variable.

Exponential distribution	Poisson distribution
Rate λ	Rate λ
Time to next event	Number of events within some time
Continuous, $[0, \infty)$	Discrete $\{0, 1, \ldots\}$

No accident in x days (cont'd)

How many days until the next accident? In many real-life cases, we assume that the **time to the next event** follows an exponential distribution. Assume that you work in a facility that has typically a rate of $\lambda = 2$ major accidents per year.

What is the probability the next accident happens in the next year? What is the probability that the next accident happens in the next 1 month?

Let X be the time until the next accident. Recall that $\lambda = 2$ per year, or equivalently $\lambda = 2$ per 12 months. We then have:

1. $P(X \leq 1 \text{ year}) = F(1) = 1 - e^{-2 \cdot 1} = 0.8647$.
2. $P(X \leq 1 \text{ month}) = F(1) = 1 - e^{-\frac{2}{12} \cdot 1} = 0.1535$.

Note how we used $\lambda = 2/12$ for the second question.

Historically, an emergency room after hours (10pm–6am) sees 48 patient requests every 8 hours. The time until the next patient arrives is exponentially distributed with that rate.

a) What is the probability that the next patient arrives in the next 10 minutes?

b) What is the probability there are 5 patients during the next hour (60 minutes)?

Memorylessness

Definition 17 (Memoryless random variables) *A random variable X is said to be* memoryless *(without memory) if:*

$$P(X > s + t | X > s) = P(X > t).$$

In English, the memorylessness property states that information available to us for what has happened so far does not alter our perception for the future. Let us see that with an example.

Memorylessness and the exponential distribution

A car transmission fails in time that is exponentially distributed with a rate of 1 every 80,000 miles. What is the probability that the transmission does not fail within its first 40,000 miles?

We need $P(T > \text{40,000 miles})$, when knowing that T is exponentially distributed with $\lambda = 1/80000$. We have:

$$P(T > \text{40,000 miles}) = 1 - P(T \leq \text{40,000 miles}) = 1 - F(40000) =$$
$$= e^{-\frac{1}{80000} \cdot 40000} = e^{-0.5} = 0.6065 = 60.65\%.$$

The next part is left as an exercise to the reader. [52]

[52] See also the Worksheet of Lecture 7!

Memorylessness and the exponential distribution

For the car from the previous example, assume we know that its transmission has been working for 80,000 miles already. What is the probability that the transmission does not fail in the next 40,000 miles?

From our answer to the previous question, we get to the point we wanted to make:

Exponentially distributed random variables are memoryless. [53]

[53] Again, the proof is something we will do in the Worksheet of Lecture 7.

Memorylessness and the uniform distribution

Assume that the time (in minutes) until the next customer shows up is uniformly distributed in $[0, 60]$. What is the probability the next customer shows up after the first minute? What is the probability the next customer shows up between the 59th and 60th minute, given that no customer has shown up until the 58th minute?

Let T be the time to the next customer arrival. In the first part, we are looking for $P(T > 1)$:

$$P(T > 1) = \frac{59}{60}.$$

Now, note that this answer is significantly different than the answer we would get by calculating $P(T > 59 | T > 58)$, which is:

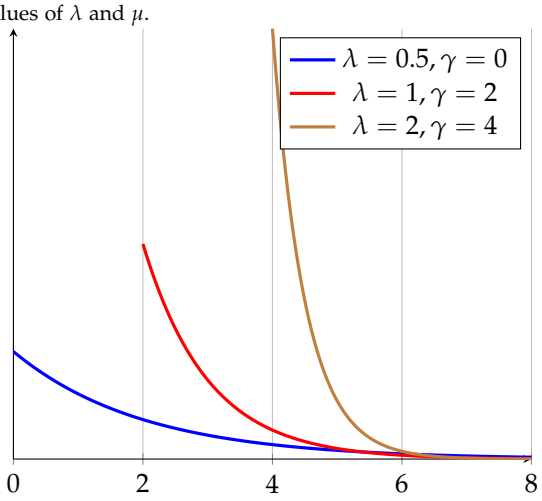
$$P(T > 59 | T > 58) = \frac{P(T > 59)}{P(T > 58)} = \frac{1/60}{2/60} = \frac{1}{2}.$$

From the example, we may deduce that the uniform distribution is not memoryless. Memorylessness is a rare property (actually, the only **continuous distribution** to possess the property is the exponential).

General exponential distribution

The exponential distribution we have discussed so far only requires a single parameter: the rate $\lambda > 0$. In some cases, we may be interested in a "shifted version" of the exponential distribution (as in Figure **??**).

Figure 21: The general exponential distribution probability density function visualized for different values of λ and μ.



We call the "shift" the location parameter and we represent it with $\gamma > 0$ [54]. Hence, the general exponential distribution is a two-parameter distribution requiring the presence of a rate $\lambda > 0$ and a location parameter $\gamma > 0$, rendering the pdf and the cdf of the general exponential distribution as follows:

[54] Some textbooks employ μ instead of γ. In this class, I will try to reserve μ for something different.

$$f(x) = \begin{cases} \lambda \cdot e^{-\lambda(x-\gamma)}, & \text{if } x \geq \gamma \\ 0, & \text{if } x < \gamma. \end{cases}$$

$$F(x) = \begin{cases} 1 - e^{-\lambda(x-\gamma)}, & \text{if } x \geq \gamma \\ 0, & \text{if } x < \gamma. \end{cases}$$

Doctor FaceTime

A doctor sees patients in time that is exponentially distributed with rate 1 patient every 40 minutes. However, every patient will spend at least 10 minutes logged in the appointment while they are answering survey questions. In essence, this means that no patient will leave before these 10 minutes are up. What is the probability the next patient is seen for:

a) more than 30 minutes?

b) more than 1 hour?

c) more than 2 hours?

Let T be the time the next patient will require: T is exponentially distributed with rate $\lambda = 1/40$ minutes and $\gamma = 10$ minutes. Then, we have:

a) $P(T > 30 \text{ minutes}) = 1 - P(T \leq 30 \text{ minutes}) = 1 - F(30) = e^{-\frac{1}{40} \cdot (30-10)} = e^{-0.5} = 0.607$.

b) $P(T > 1 \text{ hour}) = 1 - F(60) = e^{-\frac{1}{40} \cdot (60-10)} = e^{-5/4} = 0.287$.

c) $P(T > 2 \text{ hours}) = 1 - F(120) = e^{-\frac{1}{40} \cdot (120-10)} = e^{-11/4} = 0.064$.

Summary

We have seen a lot of material in this lecture. To help place everything together, we provide in Table 4 a summary of all results from Lecture 7. You could again refer to these (and the keyword that follow) for all information about the uniform and the exponential probability distributions.

Table 3: A summary of all results from Lecture 7.

Name	**Parameters**	**Values**	**pmf**
Uniform	$-$	$[a, b]$	$f(x) = \begin{cases} \frac{1}{\beta-\alpha}, & \text{if } \alpha \leq x \leq \beta \\ 0, & \text{otherwise.} \end{cases}$
Exponential	$\lambda > 0$	$[0, +\infty)$	$f(x) = \begin{cases} \lambda \cdot e^{-\lambda x}, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0. \end{cases}$
General exponential	$\lambda, \gamma > 0$	$[\gamma, +\infty)$	$f(x) = \begin{cases} \lambda \cdot e^{-\lambda(x-\gamma)}, & \text{if } x \geq \gamma \\ 0, & \text{if } x < \gamma. \end{cases}$

Some keywords that might help you narrow down your search. For convenience we also include the Poisson distribution, seeing as it is related to the exponential distribution.

Uniform: "equally probable"; "$f(x) = c$, where c is a constant".

Exponential: "time to next event"; "rate of events"; "memoryless distribution/memorylessness property".

General exponential: "time to next event"; "rate of events"; "location parameter"; "no event before a certain point".

Poisson: "number of events in an interval"; "rate of events".

Continuous random variables: part 2

Learning objectives

After these lectures, we will be able to:

- Give examples of Gamma, Erlang, and normally distributed random variables.
- Recall when to and how to use:
  - Gamma and Erlang distributed random variables.
  - normally distributed random variables.
- Recognize when to use the exponential, the Poisson, and the Erlang distribution.
- Use the standard normal distribution table to calculate probabilities of normally distributed random variables.

Motivation: Congratulations, you are our 100,000th customer!

Last time, we discussed about the probability of the next customer arriving in the next hour, next day, next year. What about the probability of the 10th customer arriving at a certain time? Or, consider a printer that starts to fail and needs maintenance after the 1000th job: what is the probability these failures start happening a month from now?

Motivation: Food poisoning and how to avoid it

A chef is using a new thermometer to tell whether certain foods have been adequately cooked. For example, chicken has to be at an internal temperature of 165 Fahrenheit or more to be adequately cooked; otherwise, we run the risk of salmonella. The restaurant wants to take *no chances*! The chef, then, takes a look at the temperature reading at the thermometer and sees 166. What is the probability that the chicken is adequately cooked, if we assume that the thermometer is right within a margin of error?

The Gamma and the Erlang distribution

Assume again that you are given a rate λ with which some events happen. So far, we have addressed two related questions:

1. What is the probability that the next event happens during some time interval? This is addressed through defining and using an exponentially distributed random variable (continuous).

2. What is the probability that we see a number of events during some time interval? This is addressed through defining and using a Poisson distributed random variable (discrete).

It is time to address a third question: what is the probability that the k-th event happens during some time interval? Like the first question, this is addressed using a continuous random variable; like the second question, we need a number of events to happen first.

Definition 18 (The Gamma distribution) *A continuous random variable X defined over the interval of $[0, \infty)$ is Gamma distributed if it can be written it has probability density function given by*

$$f(x) = \begin{cases} \frac{\lambda^k \cdot x^{k-1} \cdot e^{-\lambda x}}{\Gamma(k)}, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0, \end{cases}$$

where $\lambda > 0$ and $k > 0$ are given parameters and $\Gamma(k)$ is the Gamma function. [55] *We sometimes write that $X \sim Gamma(k, \lambda)$ if it follows the Gamma distribution with rate λ and shape parameter k.*

[55] In this class, we will only deal with integer values of k, and hence $\Gamma(k) = (k-1)!$.

When k is a positive integer number, the Gamma distribution is referred to as the **Erlang distribution**. In essence, the definition follows.

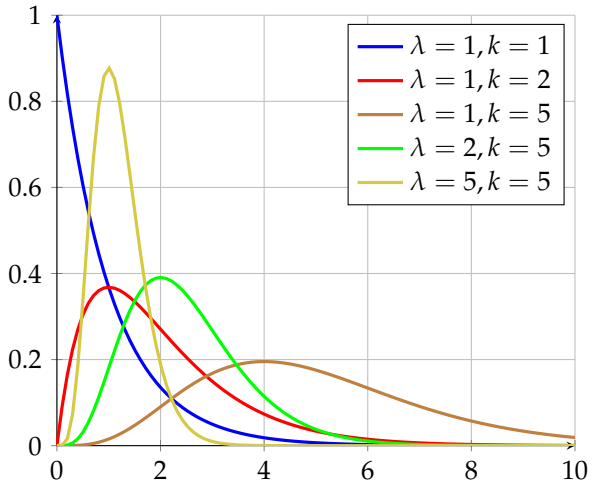
Definition 19 (The Erlang distribution) *A continuous random variable X defined over the interval of $[0, \infty)$ is Erlang distributed if it can be written as* ***the summation of exponentially distributed random variables*** *$X = \sum_{i=1}^{k} X_i$, where X_i is an exponentially distributed random variable. When X is Erlang distributed it has probability density function given by*

$$f(x) = \begin{cases} \frac{\lambda^k \cdot x^{k-1} \cdot e^{-\lambda x}}{(k-1)!}, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0, \end{cases}$$

where real $\lambda > 0$ and integer $k > 0$ are given parameters.

Due to the nature of the Γ function, it is typically easier to integrate the pdf after we plug in the value for k that we are interested in. For example, consider the problem from our motivation.

Figure 22: The Erlang distribution probability density function visualized for different values of λ and k.



Congratulations, you are our 10th customer

A store, which is open for 8 hours every day, gives a gift card to the (exactly) 10th customer of every day. The store has observed that customers show up at a rate of 1 every 20 minutes (exponentially distributed). What is the probability the 10th customer of the day shows up in the second half of the day?

Let T be the time the $k = 10$-th customer arrives. T can then be written as the summation of the arrival times of the first plus the second plus the third, all the way to the 10-th customer: since it is a summation of exponentially distributed random variables, T is Erlang distributed with parameters λ (the rate) and $k = 10$.
We are interested in $P(T > 4 \text{ hours})$. For convenience, we translate the rate to hours, getting that $\lambda = 3$ per hour:

$$P(T > 4 \text{ hours}) = \int_4^8 f(x)dx = \int_4^8 \frac{\lambda^k \cdot x^{k-1} \cdot e^{-\lambda x}}{\Gamma(k)} dx =$$
$$= \int_4^8 \frac{3^{10} \cdot x^9 \cdot e^{-3 \cdot x}}{9!} dx = \frac{59049}{362880} \int_4^8 x^9 \cdot e^{-3 \cdot x} =$$
$$= \frac{59049}{362880} \cdot 1.487 = 0.242.$$

Replacing parts

A machine requires a component to work. The component is replaced every two times the machine is doing a job. The machine works at a rate of 3 jobs per 8 hours. What is the probability the component is not replaced in the first 8 hours?

Once again, we use $\lambda = 3/8$ per hour, and then define T as the Erlang distributed random variable of the time the 2nd job appears ($k = 2$, integer and hence Erlang). We then have:

$$P(T > 8 \text{ hours}) = \int_8^\infty f(x)dx = \int_8^\infty \frac{\lambda^k \cdot x^{k-1} \cdot e^{-\lambda x}}{\Gamma(k)} dx = \\ = \int_8^\infty \frac{\left(\frac{3}{8}\right)^2 \cdot x \cdot e^{-\frac{3}{8}x}}{\Gamma(2)} dx = \frac{9}{64} \int_8^\infty x \cdot e^{-\frac{3}{8}x}. \quad (12)$$

Recall that we can integrate by parts to get:

$$\int_8^\infty x \cdot e^{-\frac{3}{8}x} = -\frac{8}{3} \int_8^\infty x \cdot \left(e^{-\frac{3}{8}x}\right)' = -\frac{8}{3} \cdot \left(x \cdot e^{-\frac{3}{8}x} \Big|_8^\infty - \int_8^\infty e^{-\frac{3}{8}x} \right) = \\ = -\frac{8}{3} \cdot \left(8e^{-3} + \frac{8}{3} \cdot e^{-3}\right) = \frac{256}{9} \cdot e^{-3} = 1.4162. \quad (13)$$

Plugging the result from (13) into (12), we get 0.19915.

The lifetime of a printer toner is exponentially distributed: it needs to be replaced once every 9 months. What probability distribution would you use for each of the following cases?

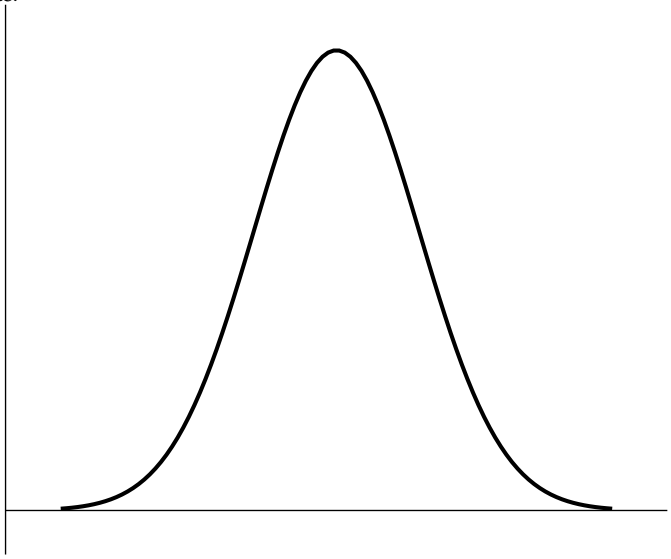
a) The number of toners you need to buy in the next 3 years.

b) The time until the toner is replaced.

c) The time until you run out of toners, if you have bought a package with 3 toners.

d) The time until the toner is replaced, given that the toner currently in use has not been replaced for 6 months already.

Figure 23: An example of how the normal distribution probability density function looks like.



The normal distribution

We have come to a big one. This is arguably the most well-studied, used, and applied distribution among the ones we have studied so far. It is defined through two parameters referred to as the mean (μ) and the variance (σ). We then say that a normally distributed random variable X is $\mathcal{N}(\mu, \sigma^2)$. [56]

[56] Note that we replace the standard deviation σ, with its square σ^2.

Definition 20 (Normal distribution) *A random variable X is said to follow a normal distribution with mean μ and standard deviation σ, if it has a probability density function of:*

$$f(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}.$$

We then commonly write that $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$.

We present an example for how the normal distribution pdf looks like in Figure 23. We observe that it is **symmetric** and **bell-shaped**. From the definition of the normal distribution, we may also get the cumulative distribution function as:

$$F(x) = \int_{-\infty}^{x} f(t)dt.$$

This is clearly an integral that we would rather not have to deal with!

Parameters

The two parameters that describe a normal distribution affect its *location* (μ) and its *spread* (σ). Visually, we show this relationship in

Figure 24: Some examples of how the normal distribution is affected by its mean and standard deviation.

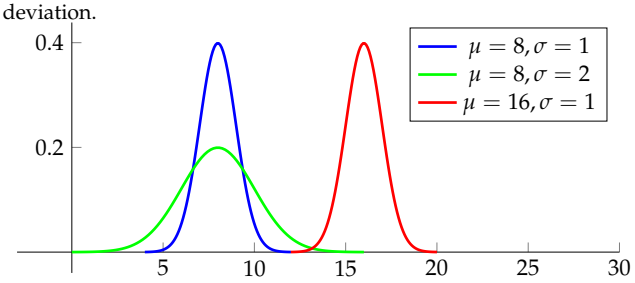


Figure 24. Note how the mean affects the location of the normal distribution, whereas the standard deviation affects how far it spreads.

The standard normal distribution

When $\mu = 0$ and $\sigma = 1$, we call the resulting normal distribution, the **standard normal distribution** and denote it as $\mathcal{N}(0, 1)$. Due to the applicability of the normal distribution in many real-life instances, the standard normal distribution has been extensively studied and we have in our possession tables containing the values of the cumulative density function. An example of such a table is provided to you in the next page.

For convenience, we refer to the pdf and the cdf of the standard normal distribution $\mathcal{N}(0, 1)$ as $\phi(z)$ and $\Phi(z)$, respectively. Note the use of z rather than the typically used x!

NORMAL CUMULATIVE DISTRIBUTION FUNCTION ($\Phi(z)$)

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

A normally distributed random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ can be converted to the standard normal distribution $\mathcal{N}(0,1)$ through one small, simple transformation, called the z-transform:

$$\text{If } X \text{ is } \mathcal{N}(\mu, \sigma^2), \text{ then } Z = \tfrac{X-\mu}{\sigma} \text{ is } \mathcal{N}(0,1).$$

This implies that for any normally distributed random variable X, $P(X = x)$ can be written as $P(Z = \frac{x-\mu}{\sigma})$, where Z is distributed following the standard normal distribution!

Doing transformations

Let X be $\mathcal{N}(400, 400)$ (i.e., $\sigma^2 = 400 \implies \sigma = 20$). What is:

a) $P(X \leq 400)$? b) $P(X \leq 451)$? c) $P(X \leq 375)$?

a) $x = 400 \implies z = \frac{400-\mu}{\sigma} = 0.$

b) $x = 451 \implies z = \frac{51}{20} = 2.55.$

c) $x = 375 \implies z = \frac{-25}{20} = -1.25.$

With z at hand, calculating a probability becomes merely a look-up operation! Indeed, all you need to do is find the z value in the cdf table. The rows reveal the two most important digits and the columns the third most important digit. For example, finding $z = 1.37$, we'd go to the 1.3 row and the 0.07 column.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	**0.9147**	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319

Using the z-table

Let X be $\mathcal{N}(400, 400)$ (i.e., $\sigma^2 = 400 \implies \sigma = 20$). What is:

a) $P(X \leq 400)$? b) $P(X \leq 451)$? c) $P(X \leq 375)$?

We already have found that:

a) $z = 0$. b) $z = 2.55$. c) $z = -1.25$.

Now is the time to find these values in the table. For the first one:

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	**0.5000**	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359

For the second one:

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	**0.9946**	0.9948	0.9949	0.9951	0.9952

For the third one, we run into a problem. The table provided does not give any negative values for z!

Recall of two facts:

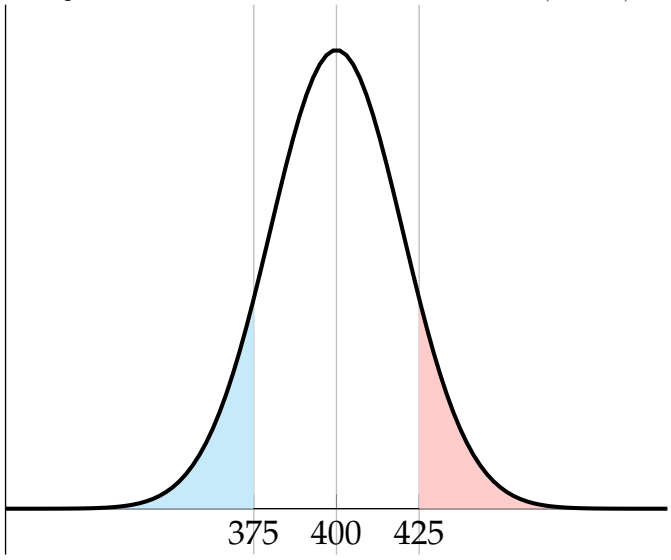
1. The normal distribution is symmetric.
2. For any continuous random variable, the probability can be found by looking at the area under the curve of the pdf.

Let us combine these two facts in an image. Consider the random variable $X \sim \mathcal{N}(400, 400)$ in Figure 25. As a reminder, we are interested in $P(X \leq 375)$.

Due to symmetry, the two shaded areas (in blue and red) have to be equal, as they are symmetric from the mean (400). Hence, we have that $P(X \leq 375) = P(X \geq 425)$. That said, we do know that $P(X \geq 425) = 1 - P(X \leq 425)$. Finally, recall that $P(X \leq 425)$ can be found in the z-table, as it corresponds to a positive value! Hence, to recap, when dealing with negative values of z, we can follow the next steps:

1. Instead of $z < 0$, search for $-z$.
2. Find the value in the z-table, $\Phi(-z)$.
3. Then, $\Phi(z) = 1 - \Phi(-z)$.

Figure 25: The pdf of the distribution of random variable $X \sim \mathcal{N}(400, 400)$.



Negative values of z

Let X be $\mathcal{N}(400, 400)$ (i.e., $\sigma^2 = 400 \implies \sigma = 20$). What is:

a) $P(X \leq 400)$? b) $P(X \leq 451)$? c) $P(X \leq 375)$?

We have solved the first two:

a) $P(X \leq 400) = 0.$ b) $P(X \leq 451) = 0.9946.$ c) $P(X \leq 375)$?

Finally, for $P(X \leq 375)$, with corresponding $z = -1.25$, we find $-z$ on the table.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	**0.8944**	0.8962	0.8980	0.8997	0.9015

Then, we report that

$$P(X \leq 375) = \Phi(-1.25) = 1 - \Phi(1.25) = 0.1056.$$

Finally, like in all distributions, if we are interested in the probability of a quantity being within a range of values (say, $P(a \leq X \leq b)$), then we may calculate $F(b) - F(a)$, or using the corresponding z-values (z_a, z_b), we may calculate that probability as $P(a \leq X \leq b) = F(b) - F(a) = \Phi(z_b) - \Phi(z_a)$.

A newsvendor is deciding how many newspapers to order for the following day. The demand for newspapers follows a normal distribution with a mean of 100 and a standard deviation of 10.

- What is the probability of selling all the newspapers they order if they place an order for:
  a) 120 newspapers?
  b) 80 newspapers?
- How many newspapers should the newsvendor order if:
  a) the newsvendor is risk-averse and would like at least a 90% chance of selling all of them?
  b) the newsvendor is risk-seeking and would like at least a 90% chance of satisfying all demand?

Summary

In Table 4, we provide all of the results from Lectures 7 and 8. We are bundling them together (even though we already provided a summary of results for Lecture 7 alone) so that we have everything for continuous distributions in one place.

Table 4: A summary of all results from Lectures 7 and 8.

Name	**Parameters**	**Values**	**pmf**
Uniform	$-$	$[a, b]$	$f(x) = \begin{cases} \frac{1}{\beta-\alpha}, & \text{if } \alpha \leq x \leq \beta \\ 0, & \text{otherwise.} \end{cases}$
Exponential	$\lambda > 0$	$[0, +\infty)$	$f(x) = \begin{cases} \lambda \cdot e^{-\lambda x}, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0. \end{cases}$
General exponential	$\lambda, \gamma > 0$	$[\gamma, +\infty)$	$f(x) = \begin{cases} \lambda \cdot e^{-\lambda(x-\gamma)}, & \text{if } x \geq \gamma \\ 0, & \text{if } x < \gamma. \end{cases}$
Gamma	$\lambda > 0, k > 0$	$[0, +\infty)$	$f(x) = \begin{cases} \frac{\lambda^k \cdot x^{k-1} \cdot e^{-\lambda x}}{\Gamma(k)}, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0, \end{cases}$
Erlang	$\lambda > 0$, integer $k > 0$	$[0, +\infty)$	$f(x) = \begin{cases} \frac{\lambda^k \cdot x^{k-1} \cdot e^{-\lambda x}}{(k-1)!}, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0, \end{cases}$
Normal	μ, σ^2	$(-\infty, +\infty)$	$f(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$

Some keywords that might help you narrow down your search. For convenience we also include the Poisson distribution, seeing as it is related to the exponential distribution.

Uniform: "equally probable"; "$f(x) = c$, where c is a constant".

Exponential: "time to next event"; "rate of events"; "memoryless distribution/memorylessness property".

General exponential: "time to next event"; "rate of events"; "location parameter"; "no event before a certain point".

Poisson: "number of events in an interval"; "rate of events".

Erlang: "time to k-th event"; "rate of events".

Normal: "normally distributed"; "average/summation of multiple identical random variables"; "central limit theorem". [57]

[57] These keywords are provided here, but are explained in Lecture 10.

Expectations and variances

Learning objectives

After these lectures, we will be able to:

- Define and explain with examples what expectations and variances are.
- Calculate the expectation and variance of discrete and continuous random variables.
- Calculate the expectation and variance of functions of discrete and continuous random variables.
- Recall and use basic properties of expectations and variances.

Motivation: Printer lifetime

A printer has a working lifetime that is exponentially distributed with rate $\lambda = 1$ broken printer every 3 years. In English, we typically replace the printer every 3 years. Assume the company decides to change the printer every 2 years (if it hasn't broken down) or when it breaks down (whichever happens first). What is the expected time the company keeps a printer?

Expectation

When describing a random variable and its probability distribution, we sometimes are interested in answering a simple question "what should I expect?" Seeing as a random variable is inherently, well, random, expectations are important and they reveal a "center" of the probability distribution.

Definition 21 (Expectation) *With the term expectation (sometimes we use the term mean or expected value), we imply a measure of the center of the probability distribution. Intuitively, we may think of the expected value of a random variable X as an "average" of the values that X is allowed to take weighted by their respective probabilities.*

This definition is very open-ended, so we provide more specific definitions for discrete and continuous random variables in the next subsubsections.

Discrete random variables

Definition 22 (Expectation of a discrete random variable) *Let X be a numerically-valued discrete random variable with sample space S and*

probability mass function $p(x)$. Then, the expected value of X is written as $E[X]$ and is calculated as:

$$E[X] = \sum_{x \in S} x \cdot p(x).$$

The expected value is commonly referred to as the mean *and is also written as μ.*

An unfair die

Consider an "unfair" die with sample space $S = \{1,2,3,4,5,6\}$ and $p(1) = 1/3, p(2) = 1/6, p(3) = 1/6, p(4) = 1/6, p(5) = 1/12, p(6) = 1/12$. What is the expected value?

- $E[X] = 1 \cdot \frac{1}{3} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{12} + 6 \cdot \frac{1}{12} = \frac{33}{12} = 2.75.$

Note how the value we expect can never actually happen, as the die can only take the values of 1, 2, 3, 4, 5, or 6!

What about for a fair die, where each side (1, 2, 3, 4, 5, or 6) are equally probable? What is the expectation for this die?

Continuous random variables

Definition 23 (Expectation of a continuous random variable) *Let X be a numerically-valued real random variable defined over $(-\infty, +\infty)$ and probability density function $f(x)$. Then, the expected value of X is written as $E[X]$ and is calculated as:*

$$E[X] = \int_{-\infty}^{+\infty} x \cdot f(x) dx.$$

The expected value is commonly referred to as the mean *and is also written as μ.*

A rating system

A company is rating their employees with a system that assigns a score between 1 and 5. We assume the score is continuous (that is, a score of 4, 4.2, and 4.31478 are all valid) and the probability with which score x appears is given by pdf $f(x) = \frac{3}{124} \cdot x^2$, for $1 \leq x \leq 5$. What is the expected rating score of a random employee?

- $E[X] = \int_1^5 x \cdot f(x) dx = \frac{3}{124} \int_1^5 x^3 dx = \frac{117}{31} = 3.77.$

Good employees are the ones that receive a rating score of 4 or above. Their scores are distributed with a slightly different distribution: they follow pdf $f(x) = \frac{2}{9} \cdot x$ for $4 \leq x \leq 5$. What is the expected rating score of a good employee?

Properties of the expectation

The expectation satisfies the following properties

1. Let α be a real number and X be a random variable. Then:

$$E[\alpha \cdot X] = \alpha \cdot E[X].$$

2. Let X, Y be two random variables. Then:

$$E[X + Y] = E[X] + E[Y].$$

   - This generalizes to as many random variables as you would want to. In essence, we have for n random variables $X_1, X_2, \ldots, X_n$:

$$E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i].$$

3. Combining 1 and 2, we have the following. Let α, β be two real numbers and X, Y be two random variables. Then:

$$E[\alpha \cdot X + \beta \cdot Y] = \alpha \cdot E[X] + \beta \cdot E[Y].$$

   - Once again, this can be generalized. For n random variables $X_1, X_2, \ldots, X_n$ and n real numbers $\alpha_1, \alpha_2, \ldots, \alpha_n$:

$$E\left[\sum_{i=1}^{n} \alpha_i \cdot X_i\right] = \sum_{i=1}^{n} \alpha_i \cdot E[X_i].$$

4. Let $g(X)$ be a function of the random variable. Then, the expectation of $g(X)$ is denoted by $E[g(X)]$ and is equal to:

   - for discrete random variable X with sample space S:

$$E[g(X)] = \sum_{x \in S} g(x) \cdot p(x).$$

   - for continuous random variable X:

$$E[g(X)] = \int_{-\infty}^{+\infty} g(x) \cdot f(x) dx.$$

Let us see a couple of examples.

Profit expectation

A company makes \$2,000 if they sell 4 units, \$1,800 if they sell 3 units, \$1,200 if they sell 2 units, lose \$1,000 if they sell 1 unit, and lose \$3,000 if they sell no units. Each event from 0 to 4 customers is equally probable. How much should they expect to make?

$$\begin{aligned} E[g(X)] &= \sum_{x=0}^{4} g(x) \cdot p(x) = \\ &= 2000 \cdot \frac{1}{5} + 1800 \cdot \frac{1}{5} + 1200 \cdot \frac{1}{5} - 1000 \cdot \frac{1}{5} - 3000 \cdot \frac{1}{5} = \\ &= \$1000. \end{aligned}$$

Circuit heat

Let X be a continuous random variable measuring the current (in milliamperes, mA) in a wire with pdf $f(x) = 0.05$, for $0 \leq x \leq 20$. The heat produced from the current is given by the function $g(x) = 10 \cdot x$ (with x in milliamperes). What is the mean heat produced by the current?

$$\begin{aligned} E[g(X)] &= \int_{x=0}^{20} g(x) \cdot f(x) \cdot dx = \int_{x=0}^{20} g(x) \cdot f(x) dx = \\ &= \int_{x=0}^{20} 10 \cdot x \cdot 0.05 \cdot dx = \int_{x=0}^{20} 0.5 \cdot x \cdot dx = 100. \end{aligned}$$

Variance

Expectations are important; they are also utterly revealing of a single point of interest. Your decision-making process is bound to be very different if I tell you that the expectation is you will make \$1000 in the following two scenarios:

1. You will make \$500 or \$1500 with probability 50% each;
2. You will lose \$3000 or make \$5000 with probability 50% each.

While in both cases the expected value is \$1000, the second one is much more "spread out" than the first one (where all values that random variable X can take are closer to the expectation).

Variance is a quantity that helps answer the question "how spread out is my distribution?" or "what is the variability of a random variable?" Once again, due to the fact that random variables are random, variances are important and they reveal the "spread" of the probability distribution. A variance of 0 implies that the expectation always comes true.

Definition 24 (Variance) *With the term variance, we imply a measure of the spread of the probability distribution. Intuitively, we may think of the variance of a random variable X as the expected squared deviation of the values that X is allowed to take compared to the expected value of X.*

In mathematical terms:

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

The definition implies that the variance is **always nonnegative**! That is, we always have

$$Var\left[X\right] \geq 0.$$

The variance is sometimes replaced by the standard deviation.

Definition 25 (Standard deviation) *Standard deviation is also a measure of the spread of a probability distribution. It is represented by $SD\left[X\right]$ and is related to the variance with the following expression:*

$$SD\left[X\right] = \sqrt{Var\left[X\right]}.$$

Unsurprisingly, it is commonly denoted by σ.

We refer to this measure as the standard deviation because it *standardizes* the units of the deviation. Note the following:

- Assume that X is a random variable measured in *units* (e.g., miles, Kelvin, \$, etc.).
- Then, $E\left[X\right]$ (or μ) is the expectation of random variable X and it is also measured in units.
- On the other hand, $Var\left[X\right]$ (or σ^2) is the variance of random variable X and it is measured in units *squared* (e.g., miles^2, Kelvin^2, \$^{2}).
- Contrary to the variance, $SD\left[X\right]$ (or σ) is the standard deviation of random variable X and it is measured in units (the same as X, miles, Kelvin, \$).

Now, like we did for expectations, we separate the discussion in discrete and continuous random variables. From now on, whenever we are interested in the standard deviation we may simply take the squared root of the variance.

Discrete random variables

Definition 26 (Variance of a discrete random variable) *Let X be a numerically-valued discrete random variable with sample space S and probability mass function $p(x)$. Then, the variance of X is written as $Var[X]$ and is calculated as:*

$$Var[X] = E\left[(X - E[X])^2\right] = \sum_{x \in S} (x - E[X])^2 \cdot p(x).$$

The variance is commonly written as σ^2.

An unfair die

Consider the same "unfair" die as before with sample space $S = \{1, 2, 3, 4, 5, 6\}$ and $p(1) = 1/3, p(2) = 1/6, p(3) = 1/6, p(4) = 1/6, p(5) = 1/12, p(6) = 1/12$. What is the variance?

Remember that $E[X] = 2.75$, as we calculated earlier. Then, applying the formula for discrete random variables, we get:

- $Var[X] = (1 - 2.75)^2 \cdot \frac{1}{3} + (2 - 2.75)^2 \cdot \frac{1}{6} + (3 - 2.75)^2 \cdot \frac{1}{6} + (4 - 2.75)^2 \cdot \frac{1}{6} + (5 - 2.75)^2 \cdot \frac{1}{12} + (6 - 2.75)^2 \cdot \frac{1}{12} = \frac{33}{12} = 2.6875$.

Recall that per Definition 24 $Var[X]$ is also equal to $E\left[X^2\right] - (E[X])^2$. Hence, we could have answered the question, using a slightly different logic:

An unfair die: second take

Remember that $E[X] = 2.75$, as we calculated earlier. Now, applying the other formula, we get:

- $Var[X] = E\left[X^2\right] - (E[X])^2 = E\left[X^2\right] - 2.75^2$.

Let us focus on the first quantity ($E\left[X^2\right]$). We have a function of a random variable, $g(X) = X^2$. Hence, applying the fourth of the expectation properties, we may compute this as:

- $E\left[X^2\right] = \sum_{x=1}^{6} x^2 \cdot p(x) = 1^2 \cdot \frac{1}{3} + 2^2 \cdot \frac{1}{6} + 3^2 \cdot \frac{1}{6} + 4^2 \cdot \frac{1}{6} + 5^2 \cdot \frac{1}{12} + 6^2 \cdot \frac{1}{12} = 10.25$.

Subtracting $2.75^2 = 7.5625$, we get that $Var[X] = 2.6875$, as expected.

What about for a fair die, where each side (1, 2, 3, 4, 5, or 6) are equally probable? What is the variance for this die? Provide a brief explanation why there is a difference in the variance of the two dies.

Continuous random variables

Definition 27 (Variance of a continuous random variable) *Let X be a numerically-valued continuous random variable defined over $(-\infty, +\infty)$ with probability distribution function $f(x)$. Then, the variance of X is written as $Var[X]$ and is calculated as:*

$$Var[X] = E\left[(X - E[X])^2\right] = \int_{-\infty}^{+\infty} (x - E[X])^2 \cdot f(x)dx.$$

The variance is commonly written as σ^2.

Back to the rating system

Earlier, we saw that the expected rating for an employee in the company was 3.77. How about the variance? Remember that the ratings are between 1 and 5 (continuous) and have pdf $f(x) = \frac{3}{124} \cdot x^2$.

We again apply the formula (but for continuous random variables now) and get:

- $Var[X] = \int_{-\infty}^{+\infty} (x - E[X])^2 \cdot f(x)dx = \int_{1}^{5} (x - 3.77)^2 \cdot dx = \int_{1}^{5} (x - 3.77)^2 \cdot dx = \int_{-2.77}^{1.23} y^2 \cdot dy = \frac{y^3}{3}\Big|_{-2.77}^{1.23} = \frac{(1.23)^3}{3} - \frac{(-2.77)^3}{3} = 7.7.$

Like we did earlier, we can again apply the formula that $Var[X] = E\left[X^2\right] - (E[X])^2$ and get the same result. This is left as an exercise to the reader.

Earlier, we saw that good employees receive a rating score of 4 or above and the distribution of their scores has pdf $f(x) = \frac{2}{9} \cdot x$ for $4 \leq x \leq 5$. What is the variance of the rating score of a good employee?

Properties of the variance

The variance satisfies the following properties.

1. Let α be a real number (not a random variable). Then:

$$Var[\alpha] = 0.$$

   - In essence, this states that when you know what is going to happen, there is no variance!

2. Let α be a real number and X be a random variable. Then:

$$Var[\alpha \cdot X] = \alpha^2 \cdot Var[X].$$

3. Let X, Y be two **independent** [58] random variables. Then:

$$Var[X + Y] = Var[X] + Var[Y].$$

[58] We did not need this assumption when we were dealing with expectations!

   - Like with the expectation, this also generalizes to more than two random variables. We then have for n **independent** random variables $X_1, X_2, \ldots, X_n$:

$$Var\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} Var[X_i].$$

   - Let α, β be real numbers and X be a random variable. Combining 1, 2, and 3 leads to:

$$Var[\alpha \cdot X + \beta] = Var[\alpha \cdot X] + Var[\beta] = \alpha^2 \cdot Var[X].$$

4. We combine 2 and 3 to get the following. Let α, β be two real numbers and X, Y be two independent random variables. Then:

$$Var[\alpha \cdot X + \beta \cdot Y] = \alpha^2 \cdot Var[X] + \beta^2 \cdot Var[Y].$$

   - In general, for n independent random variables $X_1, X_2, \ldots, X_n$ and n real numbers $\alpha_1, \alpha_2, \ldots, \alpha_n$:

$$Var\left[\sum_{i=1}^{n} \alpha_i \cdot X_i\right] = \sum_{i=1}^{n} \alpha_i^2 \cdot Var[X_i].$$

Expectation and variance of well-known distributions

In this part, we will turn our focus to the distributions we have already discussed. What is the expected number of successes in n trials? What is the expected number of earthquakes in the next decade? What is the variance of a Gamma distributed random variable? We will both derive and apply these quantities in this coming part.

Bernoulli, binomial, geometric, hypergeometric

Bernoulli distribution Recall that we say X is Bernoulli distributed if it can take two values 0 or 1 (failure or success) with probabilities $q = 1 - p$ and p, respectively. Then, based on the definition of expectation, we have:

$$E[X] = p \cdot 1 + (1-p) \cdot 0 = p.$$

Similarly, based on the definition of variance, we get:

$$Var[X] = E\left[X^2\right] - (E[X])^2 = p \cdot 1^2 + (1-p) \cdot 0^2 - p^2 = p \cdot (1-p).$$

Binomial distribution We again apply the formula from the definition of the binomial distribution with parameters n and p:

$$\begin{aligned}
E[X] &= \sum_{x=0}^{n} x \cdot p(x) = \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x} = \\
&= \sum_{x=0}^{n} x \frac{n!}{x! \cdot (n-x)!} p^x (1-p)^{n-x} = \\
&= \sum_{x=0}^{n} \frac{n \cdot (n-1)!}{(x-1)! \cdot (n-x)!} p \cdot p^{x-1} \cdot (1-p)^{n-x} = \\
&= np \sum_{x=1}^{n} \binom{n-1}{x-1} p^{x-1} (1-p)^{n-x} = \\
&= np \sum_{k=0}^{n-1} \binom{n-1}{k} p^k (1-p)^{n-k-1} = \qquad (k = x-1) \\
&= np \sum_{k=0}^{m} \binom{m}{k} p^k (1-p)^{m-k} = \qquad (m = n-1) \\
&= np
\end{aligned}$$

Why is $\sum_{k=0}^{m} \binom{m}{k} p^k (1-p)^{m-k} = 1$?

Finally, we omit the derivation for the variance, but the end result is that:

$$Var[X] = n \cdot p \cdot (1-p).$$

A certificate program

Students accepted in a certificate program graduate with probability $p = 0.75$. This year, the certificate program has accepted 300 students. How many are expected to successfully finish the program?

Let X be the random variable of the number of students that successfully finish the program. Then:

$$E[X] = n \cdot p = 300 \cdot 0.75 = 225 \text{ students.}$$

Geometric distribution We now have:

$$E[X] = \frac{1}{p}$$

and

$$Var[X] = \frac{1-p}{p^2}.$$

Shooting free throws

A kid learning basketball is shooting free throws with a probability of scoring equal to 25%. What are the expected free throws the kid has to attempt until scoring for the first time?

Let X be the number of free throws shot until the first one is made. X is a geometric random variable with $p = 0.25$, hence:

$$E[X] = \frac{1}{p} = 4 \text{ free throws.}$$

Hypergeometric distribution As a reminder, we have a population of N elements, K of which are successes (and $N - K$ are failures). We pick a sample of size n from the big population of N elements. We, then, have for a random variable X that is following a hypergeometric distribution:

$$E[X] = n \cdot \frac{K}{N}$$

and

$$Var[X] = n\frac{K}{N}\frac{(N-K)}{N}\frac{N-n}{N-1}.$$

Poisson, exponential, and Gamma

In all three distributions, we assume the availability of a rate parameter $\lambda > 0$.

Poisson distribution As a reminder, $p(x) = e^{-\lambda} \cdot \frac{\lambda^x}{x!}$. Let us apply this in the general expectation formula:

$$\begin{aligned}
E[X] &= \sum_{x=0}^{\infty} x \cdot p(x) = \sum_{x=0}^{\infty} x e^{-\lambda} \cdot \frac{\lambda^x}{x!} = \\
&= \sum_{x=1}^{\infty} x e^{-\lambda} \cdot \frac{\lambda^x}{x!} = \\
&= \lambda \cdot e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} = \\
&= \lambda \cdot e^{-\lambda} \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} = \\
&= \lambda \cdot e^{-\lambda} \cdot e^{\lambda} = \\
&= \lambda.
\end{aligned}$$

For the variance, we have:

$$\begin{aligned}
Var[X] &= E\left[X^2\right] - (E[X])^2 = \sum_{x=0}^{\infty} x^2 \cdot p(x) - \lambda^2 = \\
&= \sum_{x=0}^{\infty} x^2 e^{-\lambda} \cdot \frac{\lambda^x}{x!} - \lambda^2 = \\
&= \sum_{x=1}^{\infty} x e^{-\lambda} \cdot \frac{\lambda^x}{x!} - \lambda^2 = \\
&= \lambda \cdot e^{-\lambda} \sum_{x=1}^{\infty} x \cdot \frac{\lambda^{x-1}}{(x-1)!} - \lambda^2 = \\
&= \lambda \cdot e^{-\lambda} \left(\sum_{x=1}^{\infty} \lambda \cdot \frac{\lambda^{x-2}}{(x-2)!} + \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} \right) - \lambda^2 = \\
&= \lambda \cdot e^{-\lambda} \left(\sum_{y=0}^{\infty} \lambda \cdot \frac{\lambda^y}{y!} + \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} \right) - \lambda^2 = \\
&= \lambda \cdot e^{-\lambda} \left(\lambda e^{-\lambda} + e^{-\lambda} \right) - \lambda^2 = \\
&= \lambda^2 + \lambda - \lambda^2 = \lambda.
\end{aligned}$$

Hence, both the expectation and the variance of the Poisson is distribution is λ.

Exponential distribution This is a continuous distribution, hence the derivation of the expectation and the variance slightly change.

$$E[X] = \int_0^\infty x \cdot f(x) \cdot dx = \int_0^\infty \lambda x e^{-\lambda x} \cdot dx =$$
$$= \frac{1}{\lambda} \cdot \int_0^\infty y e^{-y} dy \qquad (y = \lambda \cdot x \implies dx = dy/\lambda)$$
$$= \frac{1}{\lambda} \left(-e^{-y} - y e^{-y}\right)\Big|_0^\infty = \frac{1}{\lambda}.$$

The variance is

$$Var[X] = \frac{1}{\lambda^2}.$$

Gamma distribution We finish this part with the Gamma distribution.

$$E[X] = \frac{n}{\lambda},$$
$$Var[X] = \frac{n}{\lambda^2}.$$

Chasing cars

A transportation engineer is counting vehicles that are passing through an intersubsection. They have observed that vehicles pass following a Poisson distribution with rate 1 vehicle every 30 seconds.

- The expected number of vehicles in the next 30 seconds is
$$\lambda = 1.$$
- The expected time until the next vehicle is
$$\frac{1}{\lambda} = \frac{1}{1/30 \text{ seconds}} = 30 \text{ seconds}.$$
- The expected time until the 10th vehicle passes is
$$\frac{10}{\lambda} = \frac{10}{1/30 \text{ seconds}} = 300 \text{ seconds} = 5 \text{ minutes}.$$

Be careful with the rate you are using! In general, given a rate λ in some time unit, then if we are asked to find an expectation in time t, we need to replace λ with $\lambda \cdot t$.

Chasing cars: part 2

A transportation engineer is counting vehicles that are passing through an intersubsection. They have observed that vehicles pass following a Poisson distribution with rate 1 vehicle every 30 seconds. How many vehicles should they expect to see in 3 hours?

- This is still a Poisson distribution with a rate of 1 every 30 seconds.
- That said, it would be easier to transform the rate into the period asked – 3 hours.
- 3 hours $=$ 10800 seconds $\implies \lambda = 360$ vehicles per 3 hours.

Uniform

Recall that there is a discrete and a continuous uniform distribution. We have:

- Discrete between a and b, that is X takes values in $a, a+1, \ldots, b$:
  1. $E[X] = \frac{a+b}{2}$.
  2. $Var[X] = \frac{(b-a+1)^2-1}{12}$.
- Continuous in (a, b):
  1. $E[X] = \frac{a+b}{2}$.
  2. $Var[X] = \frac{(b-a)^2}{12}$.

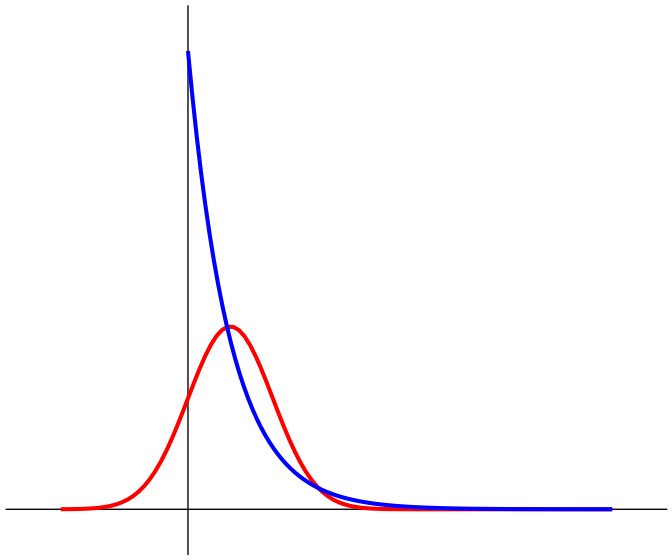
Normal

Last, but not least, we see the normal. The good news is that μ and σ^2 are both in the definition of the distribution!

Food for thought

Are two distributions with the same expectation and variance the same distributions? The answer is no: consider for a counterexample an exponential distribution with $\lambda = 0.5$ and a normal distribution $\mathcal{N}(2, 4)$. Their means are both equal to 2 and their variances are both equal to 4, but they are categorically not the same distribution (see Figure 26).

Figure 26: An example of how two distributions can have the same mean and variance but yet not be similar at all.



Review

Discrete random variables

Name	Parameters	Values	pmf	E [X]	Var [X]
Bernoulli	$0 < p < 1$	$\{0,1\}$	$p(0) = 1-p$ $p(1) = p$	p	$p(1-p)$
Binomial	$0 < p < 1, n \geq 0$	$\{0,1,\ldots,n\}$	$p(x) = \binom{n}{x} p^x \cdot (1-p)^{n-x}$	np	$np(1-p)$
Geometric	$0 < p < 1$	$\{1,2,\ldots\}$	$p(x) = (1-p)^{x-1} \cdot p$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Hypergeometric	$N, K, n \geq 0$	$\{1,2,\ldots\}$	$p(x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$	$n\frac{K}{N}$	$n \cdot \frac{K}{N} \cdot \frac{N-K}{N} \cdot \frac{N-n}{N-1}$
Poisson	$\lambda > 0$	$\{0,1,\ldots\}$	$p(x) = e^{-\lambda}\frac{\lambda^x}{x!}$	λ	λ
Uniform	-	$[a,b]$	$p(x) = \frac{1}{b-a+1}$	$\frac{a+b}{2}$	$\frac{(b-a+1)^2-1}{12}$

Continuous random variables

Name	Parameters	Values	pdf	E [X]	Var [X]
Uniform	-	$[a, b]$	$f(x) = \frac{1}{b-a}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponential	$\lambda > 0$	$[0, +\infty)$	$f(x) = \lambda \cdot e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gamma	$\lambda > 0, k > 0$	$[0, +\infty)$	$f(x) = \frac{\lambda^k \cdot x^{k-1} \cdot e^{-\lambda x}}{\Gamma(k)}$	$\frac{n}{\lambda}$	$\frac{n}{\lambda^2}$
Erlang	$\lambda > 0$, integer $k > 0$	$[0, +\infty)$	$f(x) = \frac{\lambda^k \cdot x^{k-1} \cdot e^{-\lambda x}}{(k-1)!}$	$\frac{n}{\lambda}$	$\frac{n}{\lambda^2}$
Normal	μ, σ^2	$(-\infty, +\infty)$	$f(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2

Part 2: Lectures 10–19

The central limit theorem

Learning objectives

After these lectures, we will be able to:

- Explain why the normal distribution appears often in real life.
- Recall and use the central limit theorem.

Motivation: The normal distribution

Why is the normal distribution so ubiquitous? Why is it that in many instances we see normally distributed quantities around us?

Motivation: Testing a hypothesis

Consider the case of trying to predict the outcome of an election. A good way to do so would be to pick a sample n of potential voters, and ask them what they would be voting for. Say x say they are voting for Candidate 1: this could lead you to deduce that x/n is the proportion of the vote that Candidate 1 would get in the general election. But, what can you say for the distribution of the proportion? How probable is it that your prediction is off?

Introduction

These lecture notes are organized as follows. First, we motivate why the central limit theorem applies; later in the notes, we state the theorem in its entirety. We finish this lecture with an example.

Motivating the central limit theorem

Say we throw a "fair" die (uniform distribution of getting any of the six numbers) 100,000 times and we collect back the appearances of each number. We then plot our results (see Figure 27) and observe that, as expected, every number appears equally probably. Now, let's consider a game of Monopoly. In this board game, you throw 2 dies and the summation of the two numbers is the number of steps you are expected to take: note that this number goes from 2 (both dies land on the side of 1) to 12 (both dies land on the side of 6). We already saw earlier how the probability of getting a 7 is higher than

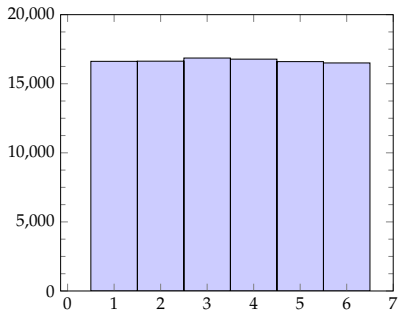


Figure 27: The number of occurrences for each number from 1 to 6 for one fair die.

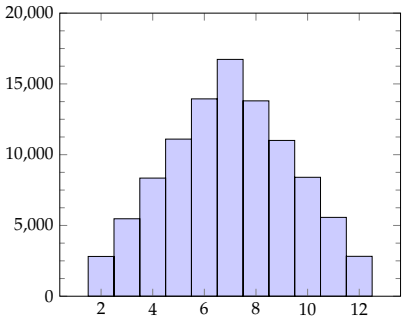


Figure 28: The number of occurrences for numbers from 2 to 12 for two fair dies.

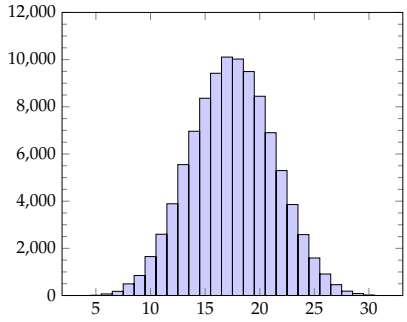


Figure 29: The number of occurrences for numbers from 5 to 30 for five fair dies.

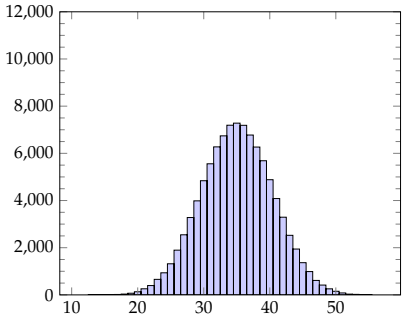


Figure 30: The number of occurrences for numbers from 10 to 60 for ten fair dies.

the rest, which is revealed also in Figure 28. There seems to be an interesting pattern emerging... Let's investigate this more!

We now proceed to show what happens when we toss 5 and 10 dies (see Figures 29 and 30). The pattern is even clearer now: it seems like **the summation of many random variables with the same distribution follows a normal distribution**! Let's see whether we can formally state what we observe.

Theorem 1 (The central limit theorem) *Let X_i, $i = 1, \ldots, n$ be a series of independent, identically distributed random variables.* [59] *Also, define $Z = \sum_{i=1}^{n} X_i$ (i.e., as the summation of all random variables X_i) or define $Y = \sum_{i=1}^{n} X_i / n$ (i.e., as the average of all X_i).*

Then both Z and Y follow a normal distribution when n is large enough. [60]

What is the implication of this result? Say we are measuring some random variable that is an average of independent random variables coming from the same distribution; then this average is expected to be normally distributed! This is why the normal distribution appears so often in real life. And this is pretty interesting since we live in a world full of data that does not seem to follow any "clean", nice

[59] Continuous or discrete, Bernoulli, binomial, geometric, Poisson, exponential, uniform, normal – any distribution. Note though that all random variables need to follow the same distribution.

[60] What constitutes "large enough"? We will investigate this later in the semester.

distributions: yet, selecting samples from this data and analyzing them provides us with a nice normal distribution to work with.

Waiting for a bus

Assume that the time you have to wait for a bus every day is uniformly distributed between 0 and 4 minutes.

a) What is the probability you have to wait for more than 3 minutes for the bus today?

b) What is the probability you have to wait on average for more than 3 minutes for the bus during 5 days of waiting for the bus every day?

c) What is the probability you have to wait on average for more than 3 minutes for the bus during your stay in Urbana-Champaign?

a) The first one is pretty straightforward: uniform distribution, continuous between 0 and 4: hence, the probability is $\frac{1}{4} = 0.25$.

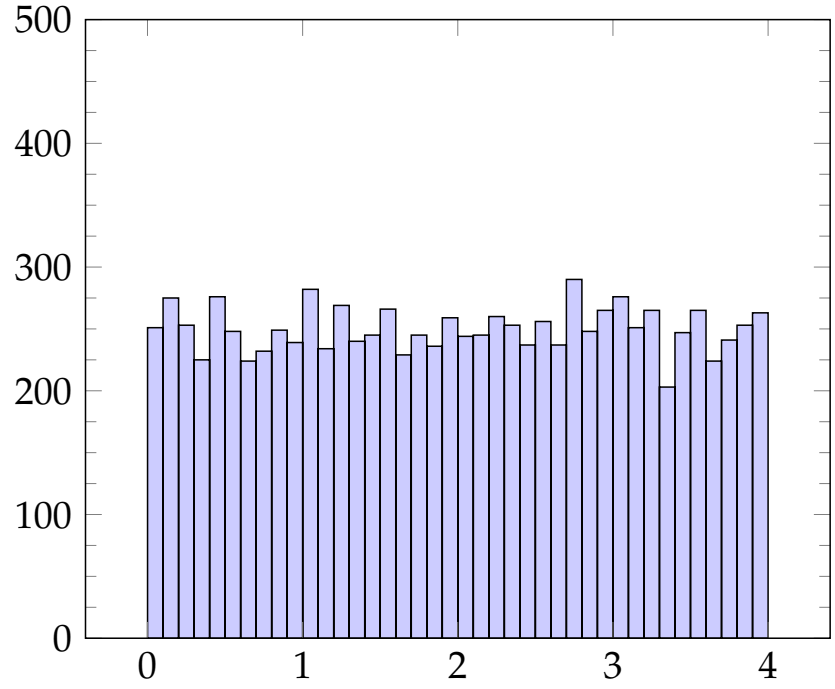
b) The second one is tougher. Is $n = 5$ (for 5 days of waiting for the bus every day) big enough for the central limit theorem to apply? And does it help to apply it?

c) The third one is similar to our previous reasoning–**but** keeping in mind that your stay in Urbana-Champaign is for 3-4 years, we may assume that n (number of days waiting for a bus) is pretty big. Does the central limit theorem help?

Based on the central limit theorem, the average time we wait for the bus is normally distributed. Let us visualize what this means (much like what we did for the dies earlier). We will generate 10000 random variables and present the results depending on the number of times each interval of numbers appears.

Finally, for the last question (where the central limit theorem would clearly hold as n is very big), we can immediately find the probability using a normal distribution. The details of the normal distribution will be discussed later in the semester.

Figure 31: The amount of time we spend waiting for one bus.



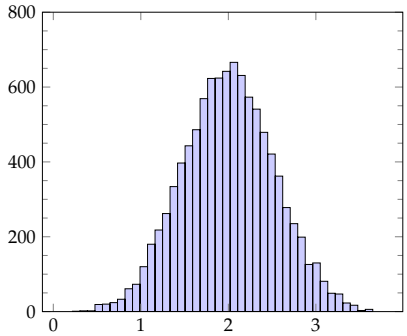


Figure 32: The average amount of time we spend waiting for a bus for a total of $n = 5$ days.

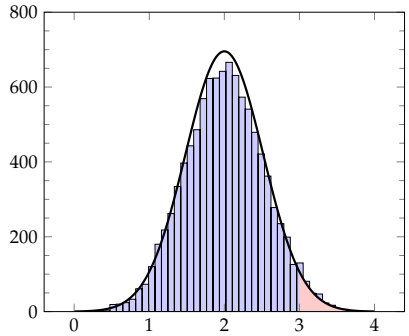


Figure 33: Visualizing the "probability" of waiting (on average) for more than 3 minutes in 5 days. It is shown in red.

The central limit theorem

Expectation and variance review

Let us recall a few important properties from calculating expectations and variances. Given a series of independent random variables $X_i, i = 1, \ldots, n$, we have that:

$$E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E\left[X_i\right].$$
$$Var\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} Var\left[X_i\right].$$

Moreover, recall that:

$$E\left[\alpha \cdot X\right] = \alpha \cdot E\left[X\right].$$
$$Var\left[\alpha \cdot X\right] = \alpha^2 \cdot Var\left[X\right].$$

Combining, we have that for quantities $\frac{\sum_{i=1}^{n} X_i}{n}$, we get:

$$E\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] = \frac{\sum_{i=1}^{n} E\left[X_i\right]}{n}.$$
$$Var\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] = \frac{\sum_{i=1}^{n} Var\left[X_i\right]}{n^2}.$$

Assuming that we have $\mu = E\left[X_1\right] = E\left[X_2\right] = \ldots = E\left[X_n\right]$ and $\sigma^2 = Var\left[X_1\right] = Var\left[X_2\right] = \ldots = Var\left[X_n\right]$, we may write that:

$$E\left[\sum_{i=1}^{n} X_i\right] = n \cdot \mu \qquad E\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] = \frac{n \cdot \mu}{n} = \mu.$$
$$Var\left[\sum_{i=1}^{n} X_i\right] = n \cdot \sigma^2 \qquad Var\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] = \frac{n \cdot \sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

The full theorem

We are now ready to state the central limit theorem in its entirety.

Theorem 2 (The central limit theorem) *Let X_i, $i = 1, \ldots, n$ be a series of independent, identically distributed random variables with expected value $E[X_i] = \mu$ and variance $Var[X_i] = \sigma^2$. Define $Z = \sum_{i=1}^{n} X_i$ (i.e., as the summation of all random variables X_i) and $Y = \sum_{i=1}^{n} X_i / n$ (i.e., as the average of all X_i).*

Then:

- *Z follows a normal distribution when n is large enough with parameters $\mu_Z = \sum_{i=1}^{n} E[X_i] = n \cdot \mu$ and $\sigma_Z^2 = \sum_{i=1}^{n} Var[X_i] = n \cdot \sigma^2$.*

$$\boxed{Z \sim \mathcal{N}\left(n \cdot \mu, n \cdot \sigma^2\right).}$$

- *Y follows a normal distribution when n is large enough with parameters $\mu_Y = \frac{1}{n} \sum_{i=1}^{n} E[X_i] = \mu$ and $\sigma_Z^2 = \frac{1}{n} \sum_{i=1}^{n} Var[X_i] = \frac{\sigma^2}{n}$.*

$$\boxed{Y \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).}$$

Random buses

The time you have to wait for a bus every day is uniformly distributed between 0 and 4 minutes. What is the probability you have to wait for more than or equal to 2.2 minutes for the bus on average in the next 300 days?

We may now fully use the central limit theorem.

- As $n = 300$ (big enough), we know that the average time you will wait for the bus in the next 300 days is normally distributed with mean μ and variance $\sigma^2/300$.
- The time you wait for a single bus is uniformly distributed with mean $\mu = \frac{0+4}{2} = 2$ minutes and variance $\sigma^2 = \frac{(4-0)^2}{12} = 4/3$ minutes^2.
- Combining, the average time T you wait for the bus follows $\mathcal{N}(2, \frac{4}{900})$.

Random buses (cont'd)

We are interested in $P(T > 2.2) = 1 - P(T \leq 2.2)$. First let us convert to the proper z value:

$$Z = \frac{X - \mu}{\sigma} = \frac{2.2 - 2}{2/30} = 3.$$

Looking at the z-table:

$$P(T \leq 2.2) = 0.9987 \implies P(T \geq 2.2) = 1 - 0.9987 = 0.0013.$$

Jointly distributed random variables

Learning objectives

After these lectures, we will be able to:

- Describe and recognize jointly distributed random variables.
- Define joint, marginal, and conditional probability mass functions for discrete random variables.
- Define joint, marginal, and conditional probability distribution functions for continuous random variables.
- Use joint, marginal, and conditional probability mass and distribution functions to calculate probabilities.

Motivation: "Can you hear me now"?

Not all of us pay attention all the time in a Zoom call; sometimes it is our fault (we are distracted or busy), but others it is not (technical difficulties, bad reception). So the question becomes: how many times does something need to be repeated before you hear it? Note that it does not only depend on whether you are paying attention (which is a random variable), but also on whether you are having a clear connection (another random variable).

Jointly distributed random variables

Real life and its outcomes can be viewed as a combination of random events, rather than a single random event. Succeeding in an exam has many factors that do not rely on only your preparation: you need to be healthy and well-rested, you need to be focused during the exam, you need to have luck at your side, you need to have a calculator whose batteries are still working. And even when all of these things align, you also need to be there on time, which means that you need to catch a bus, that there is no construction causing traffic jams, etc. We can go on like this *a lot*.

The truth is that in this class we have focused on single random variables that are distributed their own way. What about the case where two random variables are distributed alongside each other?

But, wait? Did we not discuss the probability of two events happening hand-in-hand during the first lectures of the course? And did we not discuss specifically what happens if those two events are independent[61] *or not?*

You are correct. We have discussed what happens when two events are happening at the same time. We also discussed what the

[61] Recall independence: it implies that knowledge of one event happening does not affect our probabilities of the other event happening.

probability is that an event happens given another event happening. However, there are two caveats in our discussion earlier:

1. We only focused on discrete (countable) events: it is time we see what this implies in the continuous space too.
2. We saw this in terms of events and sets. We are now going to have that discussion in terms of distributions, probability mass/density functions.

Definition

We begin with the definition of jointly distributed random variables.

Definition 28 (Jointly distributed random variables) *Let X and Y be two random variables. The probability distribution that defines their simultaneous behavior is referred to as a* ***joint probability distribution****. The two random variables X and Y are then called* ***jointly distributed random variables****.*

Examples of jointly distributed random variables

Here are some examples of jointly distributed random variables.

- The times you have to repeat yourself on the phone and your signal reception.
- The grade you receive in an exam and the amount of sleep you've had the night before.
- The performance of two or more stocks in your portfolio.
- The box office of a movie and the critical reception.

We observe here that jointly distributed does not imply immediate effect. For example, a student could get a very high grade in an exam, even if they slept very little the night before; or a movie could make a lot of money in the box office, despite being universally hated by reviewers. However, jointly distributed random variables imply that what we see is a combination of random variables rather than outcome of a single random variable.

Are the following better modeled as a single random variable or as jointly distributed random variables?

- Getting a higher grade in an exam than the person sitting next to you?
- Throwing a die?
- Throwing two dies and having the first die land on a higher number than the second one?

An example

Securing a position after college might require some effort.. If the economy is doing well ("is good"), then a student could get more job interview invitations, and consequently there are more chances for a job opportunity. If the economy is average, or if it is outright bad, then a student may struggle to get interviews and/or a job..

Based on our definitions, the state of the economy is a random variable. The same can be said about the number of job interviews that a student gets invited to. In the end of the day, the number of interviews that a student needs to go on before they secure a position after graduation is a **jointly distributed random variable**. Let's assume that the probabilities are as given in Table 5.

Table 5: Number of job interviews required to get a job depending on the state of the economy.

	Y=state of the economy		
X=job interviews to get a job	Bad	Average	Good
1	0.01	0.05	0.20
2	0.03	0.05	0.18
3	0.03	0.12	0.08
≥ 4	0.08	0.12	0.05

Jointly distributed discrete random variables

If X and Y are discrete random variables, then (X, Y) is called a jointly discrete bivariate random variable.

Definition 29 (Joint probability mass function) *The **joint probability mass function** is defined as:*

$$f_{XY}(x, y) = P(X = x, Y = y).$$

It follows the next three properties:

1. $f_{XY}(x,y) \geq 0, \forall x, y.$
2. $\sum_x \sum_y f_{XY}(x,y) = 1.$
3. $P((X,Y) \in A) = \sum\sum_{(x,y)\in A} f_{XY}(x,y).$

A couple of quick notes about the notation here. You will observe that the joint probability mass function is given by $f_{XY}(x,y)$. We had previously reserved $f(\cdot)$ for continuous random variables, keeping $p(\cdot)$ for discrete ones. For convenience, we only use $f(\cdot)$ for joint probability distributions.

Additionally, we notice that there is a subscript in the function. The subscript is supposed to reveal which random variables the function is including. For example $f_{XY}(3,4)$ would imply that the function is considering random variables X and Y and is asking for them to be equal to 3 and 4, respectively.

Note that this definition is easily generalized for more than two variables: if X_i are discrete random variables for $i = 1, \ldots, n$, then $(X_1, \ldots, X_n)$ is called a **jointly distributed discrete multivariate random variable** with joint pmf:

$$f_{X_1 X_2 \ldots X_n}(x_1, x_2, \ldots, x_n) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n).$$

Following the same notation as before, we see from the subscript of the function that this distribution contains random variables $X_1, X_2, \ldots, X_n$.

Getting a job after college

Let's see: do the probabilities provided in the example earlier satisfy the first two properties?

1. $f_{XY}(x,y) \geq 0, \forall x, y$. This is true, as all entries for the 12 cases are all positive.
2. $\sum_x \sum_y f_{XY}(x,y) = 0.01 + 0.05 + 0.20 + \ldots + 0.05 = 1$. This is also true.

Let's dwell a little on the third property now.

Getting a job after college

What is the probability that:

1. a student gets a job in 1 interview and that the economy is good?

	Y		
X	Bad	Average	Good
1	0.01	0.05	**0.20**
2	0.03	0.05	0.18
3	0.03	0.12	0.08
≥ 4	0.08	0.12	0.05

The probability is 20%.

2. a student gets a job in less than or equal to 3 interviews and the economy is average?

	Y		
X	Bad	Average	Good
1	0.01	**0.05**	0.20
2	0.03	**0.05**	0.18
3	0.03	**0.12**	0.08
≥ 4	0.08	0.12	0.05

The probability is 22%.

3. a student gets a job in more than 3 interviews?
4. the economy is good?
5. a student gets a job in 1 interview if we know that the economy is good?

Questions 3, 4, and 5 seem to require a little different logic. Could we add all the outcomes that include the specific clause we are after? For example, could we simply add all the probabilities of a good economy and say that this is the probability that the economy is good? But for the last one, we know that the economy is good. How can we use this fact to calculate the required probability? Could we use conditional probabilities?

Marginal probability mass function

Definition 30 (Marginal probability mass function) *The **marginal probability mass function** (marginal pmf) of a discrete random variable is computed by summing over all possible values of the other random variable. For two random variables, X and Y:*

1. *The marginal distribution of X:*

$$f_X(x) = P(X = x) = \sum_y f_{XY}(x, y)$$

2. *The marginal distribution of Y:*

$$f_Y(y) = P(Y = y) = \sum_x f_{XY}(x, y)$$

The marginal distribution of a random variable answers the question: "what is the probability that X takes a certain value, regardless of Y?" [62]

[62] And vice versa for the marginal distribution of Y.

Going back to the motivation from earlier:

Getting a job after college

What is the probability that:

1. a student gets a job in 1 interview and that the economy is good? The probability is 20%.
2. a student gets a job in less than or equal to 3 interviews and the economy is average. The probability is 22%.
3. a student gets a job in more than 3 interviews?

   We are after the probability of $P(X > 3)$. Based on the definition of marginal distributions, we have that:

$$P(X = x) = \sum_y f_{XY}(x, y) \implies$$

$$\implies P(X > 3) = P(X \geq 4) = 0.08 + 0.12 + 0.05 = 0.25.$$

	Y		
X	Bad	Average	Good
1	0.01	0.05	0.20
2	0.03	0.05	0.18
3	0.03	0.12	0.08
≥ 4	**0.08**	**0.12**	**0.05**

The probability is 25%.

Getting a job after college (cont'd)

4. the economy is good?

   We are after the probability of $P(Y = \text{Good})$. Following a similar logic:

$$P(Y = y) = \sum_x f_{XY}(x, y) \implies$$

$$\implies P(Y = \text{Good}) = 0.20 + 0.18 + 0.08 + 0.05 = 0.51.$$

	Y		
X	Bad	Average	Good
1	0.01	0.05	**0.20**
2	0.03	0.05	**0.18**
3	0.03	0.12	**0.08**
≥ 4	0.08	0.12	**0.05**

The probability is 51%.

5. a student gets a job in 1 interview if we know that the economy is good?

For calculating marginal distributions in discrete random events given in tabular format, we may also add up the probabilities in the columns and rows and obtain:

	Y=state of the economy			
X=job interviews to get a job	Bad	Average	Good	$f_X(x)$
1	0.01	0.05	0.20	0.26
2	0.03	0.05	0.18	0.26
3	0.03	0.12	0.08	0.23
≥ 4	0.08	0.12	0.05	0.25
$f_Y(y)$	0.15	0.34	0.51	1

Here the columns are showing the probability of the state of the economy (alone) which are bad with 15%, average with 34%, and good with 51%, whereas the rows are showing the number of interview (26% for 1, 26% for 2, 23% for 3, 25% for more than 3).

Conditional probability mass function

Definition 31 (Conditional probability mass function) *The **conditional probability mass function** (conditional pmf) of a discrete random variable given values for the other ones is computed by dividing the joint pmf of all over the marginal pmf of the others. For two random variables, X and Y:*

1. *The conditional distribution of X given $Y = y$:*

$$f_{X|Y=y}(x) = f_{X|y} = P(X = x|Y = y) = \frac{f_{XY}(x, y)}{f_Y(y)}$$

2. *The conditional distribution of Y given $X = x$:*

$$f_{Y|X=x}(y) = f_{Y|x} = P(Y = y|X = x) = \frac{f_{XY}(x,y)}{f_X(x)}$$

Of course, for the conditional pmf to make sense, we need that $f_X(x) > 0$ and $f_Y(y) > 0$.

One more note of notation. Observe how the subscript has changed to reflect the fact that we know what Y or X is. We write:

$$f_{X|Y=y} = f_{X|y},$$

which is read as the "conditional pmf of random variable X given that random variable Y is equal to y" or, simply the "conditional pmf of random variable X given y."

Let us revisit our motivation.

Getting a job after college

What is the probability that:

1. a student gets a job in 1 interview and that the economy is good? The probability is 20%.
2. a student gets a job in less than or equal to 3 interviews and the economy is average. The probability is 22%.
3. a student gets a job in more than 3 interviews? The probability is 25%.
4. the economy is good? The probability is 51%.
5. a student gets a job in 1 interview if we know that the economy is good?

   This is the definition of a conditional probability. Specifically, we want to calculate $P(X = 1|Y = \text{Good})$.

$$P(X = 1|Y = \text{Good}) = \frac{f_{XY}(1,\text{Good})}{f_Y(\text{Good})} = \frac{0.2}{0.51} = 0.3922.$$

	Y			
X	Bad	Average	Good	$f_X(x)$
1	0.01	0.05	**0.20**	0.26
2	0.03	0.05	0.18	0.26
3	0.03	0.12	0.08	0.23
≥ 4	0.08	0.12	0.05	0.25
$f_Y(y)$	0.15	0.34	**0.51**	1

One full example

As interesting as this example has been, the truth is that in many cases we cannot enumerate easily all cases. In those instances, we turn to calculus. Let us see a similar case:

Jointly distributed discrete random variables

Two discrete random variables X and Y have a joint distribution of $f_{XY}(x,y) = \frac{x+y+1}{c}$, for x and y equal to 0, 1, or 2.

1. What should c be?
2. What is $P(X \leq 1, Y = 1)$?
3. What is $P(Y = 1)$?
4. What is $P(X \leq 1 | Y = 1)$?

For calculating c, we need to use the second property.

Getting the joint pmf

$$\sum_{x=0}^{2}\sum_{y=0}^{2} f_{XY}(x,y) = \sum_{x=0}^{2}\sum_{y=0}^{2} \frac{x+y+1}{c} = 1 \implies$$
$$\implies \frac{1}{c} + \frac{2}{c} + \frac{3}{c} + \frac{2}{c} + \frac{3}{c} + \frac{4}{c} + \frac{3}{c} + \frac{4}{c} + \frac{5}{c} = 1$$
$$\implies c = 27.$$

With the full joint pmf, we can answer the remaining questions:

Using the joint pmf

$$P(X \leq 1, Y = 1) = \sum_{x=0}^{1} f_{XY}(x,1) = \sum_{x=0}^{2} \frac{x+2}{27} = \frac{2}{27} + \frac{3}{27} = \frac{5}{27}.$$

We now move our focus to the marginal distribution, which can be found as:

Computing and using the marginal pmf

$$f_Y(y) = P(Y = y) = \sum_{x=0}^{2} f_{XY}(x,y) =$$
$$= \frac{0+y+1}{27} + \frac{1+y+1}{27} + \frac{2+y+1}{27} \implies f_Y(y) = \frac{3y+6}{27}.$$

Hence $P(Y = 1) = f_Y(1) = \frac{9}{27} = \frac{1}{3}$.

To conclude this, we may combine the joint and marginal pmf to get the conditional pmf:

Computing and using the conditional pmf

$$f_{X|Y=y}(x) = P(X = x|Y = y) = \frac{f_{XY}(x,y)}{f_Y(y)} = \frac{\frac{x+y+1}{27}}{\frac{3y+6}{27}} = \frac{x+y+1}{3y+6}$$

Finally: $P(X \leq 1|Y = 1) = f_{X|Y=1}(0) + f_{X|Y=1}(1) = \frac{2}{9} + \frac{3}{9} = \frac{5}{9}$.

Jointly distributed continuous random variables

If X and Y are continuous random variables, then (X, Y) is called a jointly continuous bivariate random variable.

Definition 32 (Joint probability distribution function) *The **joint probability distribution function** is defined as:*

$$f_{XY}(x,y).$$

Like in the simple continuous random variables, f_{XY} reveals a relative likelihood rather than a probability value. It also follows three properties:

1. $f_{XY}(x,y) \geq 0, \forall x, y.$
2. $\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{XY}(x,y)dxdy = 1.$
3. $P((X,Y) \subset R) = \iint_R f_{XY}(x,y)dxdy.$

Once again, the definitions is easy to generalize to more than two variables: if X_i are continuous random variables for $i = 1, \ldots, n$, then $(X_1, \ldots, X_n)$ is called a **jointly distributed continuous multivariate random variable** with joint pdf:

$$f_{X_1X_2\ldots X_n}(x_1, x_2, \ldots, x_n).$$

In the end of the subsection, we given an example with more than 2 random variables so that you can practice with it – it will come in hand during Lecture 13.

A chemical mixture

A product is a mixture of two materials: let the volume of material 1 used be represented as X, and the volume of material 2 used be represented as Y. The joint probability density function of the two random variables is

$$f_{XY}(x,y) = c\,(2x+3y)\,, \quad 0 \le x \le 1, 0 \le y \le 1.$$

1. What is c?
2. What is the probability the first material has volume less than or equal to 0.5, and the second material has volume between 0.25 and 0.5?

Similarly to what we did for discrete random variables (with the main difference that we now need to integrate over the values that X and Y can take), we get:

Computing the joint pdf

1. From the second property of joint pdfs for continuous random variables, we have:

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{XY}(x,y)dxdy = 1 \implies \int_0^1\int_0^1 c\,(2x+3y)\,dxdy = 1 \implies$$

$$\implies c\int_0^1 \left(x^2+3xy\right)\Big|_0^1 dy = 1 \implies c\int_0^1 (3y+1)dy = 1 \implies$$

$$\implies c\left(3\frac{y^2}{2}+y\right)\Big|_0^1 = 1 \implies c\frac{5}{2} = 1 \implies c = \frac{2}{5}.$$

Using the joint pdf

2. Knowing that $f_{XY}(x,y) = \frac{2}{5}(2x+3y)$, we may calculate the probability $P(X \leq 0.5, 0.25 \leq Y \leq 0.5)$ as follows. Recall that we are talking about continuous random variables, so we will always integrate!

$$P(X \leq 0.5, 0.25 \leq Y \leq 0.5) = \int_0^{0.5}\int_{0.25}^{0.5} f_{XY}(x,y)dydx =$$
$$= \int_0^{0.5}\int_{0.25}^{0.5} \frac{2}{5}(2x+3y)\,dydx = \frac{2}{5}\int_0^{0.5}\left(2xy+3\frac{y^2}{2}\right)\Big|_{0.25}^{0.5} dx =$$
$$= \frac{2}{5}\int_0^{0.5}\left(0.5x+\frac{9}{32}\right)dx = \frac{2}{5}\left(0.5\frac{x^2}{2}+\frac{9}{32}x\right)\Big|_0^{0.5} = \frac{13}{160}$$

Marginal probability distribution function

Definition 33 (Marginal probability distribution function) *The **marginal probability distribution function** (marginal pdf) of a continuous random variable is computed by integrating over all possible values of the other random variable. For two random variables, X and Y:*

1. *The marginal distribution of X:*

$$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x,y)dy$$

2. *The marginal distribution of Y:*

$$f_Y(y) = \int_{-\infty}^{+\infty} f_{XY}(x,y)dx$$

Let us return to our chemical mixture:

Computing and using the marginal pdf

As a reminder, we have $f_{XY}(x,y) = \frac{2}{5}(2x+3y)$ for the joint pdf of two continuous random variables X and Y (volume of material 1 and 2, respectively) taking values between 0 and 1. What is the probability that:

1. the volume of material 1 is less than 0.5?
2. the volume of material 2 is between 0.25 and 0.5?

Computing and using the marginal pdf

1. First, we calculate $f_X(x)$. It is:

$$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x,y)dy = \int_0^1 \frac{2}{5}(2x+3y)\,dy = \frac{1}{5}(4x+3).$$

We can now use $f_X(x)$:

$$P(X \leq 0.5) = \int_0^{0.5} f_X(x)dx = \int_0^{0.5} \frac{1}{5}(4x+3)\,dx = 0.4.$$

2. Then, we do the same for $f_Y(y)$:

$$f_Y(y) = \int_{-\infty}^{+\infty} f_{XY}(x,y)dx = \int_0^1 \frac{2}{5}(2x+3y)\,dx = \frac{1}{5}(6y+2).$$

And we finish the question by calculating the proper integral:

$$P(0.25 \leq Y \leq 0.5) = \int_{0.25}^{0.5} f_Y(y)dy = \int_{0.25}^{0.5} \frac{1}{5}(6y+2)\,dy = \frac{17}{80}.$$

Conditional probability distribution function

Definition 34 (Conditional probability distribution function) *The **conditional probability distribution function** (conditional pdf) of a continuous random variable given values for the other ones is computed by dividing the joint pdf of all over the marginal pdf of the others. For two random variables, X and Y:*

1. *The conditional distribution of X given $Y = y$:*

$$f_{X|Y=y}(x) = f_{X|y} = \frac{f_{XY}(x,y)}{f_Y(y)}$$

2. *The conditional distribution of Y given $X = x$:*

$$f_{Y|X=x}(y) = f_{Y|x} = \frac{f_{XY}(x,y)}{f_X(x)}$$

Once again, the conditional pdf is only defined when $f_X(x) > 0$ and $f_Y(y) > 0$.

Let's go back to our chemical mixture example.

Computing and using the conditional pdf

What is the probability the first material has a proportion less than or equal to 50%, given that the second material has a proportion equal to 30%?

To contrast this with the following question, we shall name this the conditional distribution route.

The conditional distribution route. First, we calculate $f_{X|Y=y}(x)$ as:

$$f_{X|Y=y}(x) = \frac{f_{XY}(x,y)}{f_Y(y)} = \frac{\frac{2}{5}(2x+3y)}{\frac{1}{5}(6y+2)} = \frac{4x+6y}{6y+2}.$$

Replacing $Y = 0.3$ as is known, we get:

$$f_{X|Y=0.3}(x) = \frac{4x+1.8}{3.8}.$$

Finally, we may calculate $P(X \leq 0.5|Y = 0.3)$ as follows:

$$P(X \leq 0.5|Y = 0.3) = \int_0^{0.5} \frac{4x+1.8}{3.8} dx = 0.3684.$$

And here is one more conditional to practice basic probability theory:

Calculating conditional probabilities

What is the probability the first material has a proportion less than or equal to 50%, given that the second material has a proportion between 25% and 50%?

The basic probability theory route. Remember that $P(A|B) = P(A \cap B)/P(B)$. In our case, we have already calculated $P(A \cap B) = P(X \leq 0.5, 0.25 \leq Y \leq 0.5) = \frac{13}{160}$ and $P(B) = P(0.15 \leq Y \leq 0.5) = \frac{17}{80}$. Combining:

$$P(X \leq 0.5|0.25 \leq Y \leq 0.5) = \frac{13/160}{17/80} = \frac{13}{34}.$$

A multivariate example

Here, we provide a small example for a joint distribution with 4 random variables. More specifically:

A 4-component machine

Suppose that a machine consists of four components, whose lifetimes (in years) are jointly distributed with the following pdf:

$$f_{X_1X_2X_3X_4}(x_1,x_2,x_3,x_4) = c \cdot e^{-2x_1}e^{-x_2}e^{-3x_3}e^{-0.5x_4}.$$

- What should c be for this to be a valid pdf?
- What is the probability the first component survives for more than one year?

For the first question, we want

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{X_1X_2X_3X_4}(x_1,x_2,x_3,x_4)dx_4dx_3dx_2dx_1 = 1 \implies$$

$$\implies \int_{0}^{+\infty}\int_{0}^{+\infty}\int_{0}^{+\infty}\int_{0}^{+\infty} f_{X_1X_2X_3X_4}(x_1,x_2,x_3,x_4)dx_4dx_3dx_2dx_1 = 1 \implies$$

$$\implies c\int_{0}^{+\infty}\int_{0}^{+\infty}\int_{0}^{+\infty}\int_{0}^{+\infty} e^{-2x_1}e^{-x_2}e^{-3x_3}e^{-0.5x_4}dx_4dx_3dx_2dx_1 = 1 \implies$$

$$\implies \frac{c}{0.5}\int_{0}^{+\infty}\int_{0}^{+\infty}\int_{0}^{+\infty} e^{-2x_1}e^{-x_2}e^{-3x_3}dx_3dx_2dx_1 = 1 \implies$$

$$\implies \frac{c}{0.5\cdot 3}\int_{0}^{+\infty}\int_{0}^{+\infty} e^{-2x_1}e^{-x_2}dx_2dx_1 = 1 \implies$$

$$\implies \frac{c}{0.5\cdot 3\cdot 1}\int_{0}^{+\infty} e^{-2x_1}dx_1 = 1 \implies \frac{c}{0.5\cdot 3\cdot 1\cdot 2} = 1 \implies c = 3.$$

For the second question, first calculate the marginal pdf $f_{X_1}(x_1)$:

$$f_{X_1}(x_1) = \int_{0}^{+\infty}\int_{0}^{+\infty}\int_{0}^{+\infty} f_{X_1X_2X_3X_4}(x_1,x_2,x_3,x_4)dx_4dx_3dx_2 = 2e^{-2x_1}.$$

Then, the probability it survives for more than one year is

$$P(X_1 > 1) = 1 - P(X_1 \leq 1) = 1 - \int_{0}^{1} 2e^{-2x_1}dx_1 = 0.1353.$$

Review

A very brief summary of today's lecture follows. For two random variables X and Y that are *jointly* distributed, we have the following:

Joint pmf/pdf

- TL;DR: How are both variables distributed as *simultaneously*?
  - **Discrete**: what is $P(X = x \text{ and } Y = y)$?
  - **Continuous**: what is the relative likelihood of X having the value of x and Y getting the value of y?
- Denoted by $f_{XY}(x, y)$.
- Follows three properties:

  - **Discrete**:
    1. $f_{XY}(x, y) \geq 0$.
    2. $\sum_x \sum_y f_{XY}(x, y) = 1$.
    3. $P((X, Y) \in A) = \sum\sum_{(x,y)\in A} f_{XY}(x, y)$.

  - **Continuous**:
    1. $f_{XY}(x, y) \geq 0, \forall x, y$.
    2. $\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{XY}(x, y)dxdy = 1$.
    3. $P((X, Y) \subset R) = \int\int_R f_{XY}(x, y)dxdy$.

Marginal pmf/pdf

- TL;DR: I am only interested in one of the random variables.
  - **Discrete**: Let's lose Y. What $P(X = x)$?
  - **Continuous**: Let's lose Y. What the relative likelihood of X getting the value of x?
- Denoted by $f_X(x)$ or $f_Y(y)$.

Conditional pmf/pdf

- TL;DR: I am given information on one of the random variables.
  - **Discrete**: I know what Y is! It is equal to y. What is $P(X = x|Y = y)$?
  - **Continuous**: I know what Y is! It is equal to y. What is the relative likelihood of X taking value x now?
- Denoted by $f_{X|y}(x)$ or $f_{Y|x}(y)$.

Recall that all definitions shown here for both discrete and continuous random variables may be extended to more than 2 random variables $X_1, X_2, \ldots$

Joint distributions: extensions

Learning objectives

After these lectures, we will be able to:

- Calculate and use the expectation and variance of jointly distributed random variables.
- Define, calculate, and use the conditional expectation.
- Recognize independence in joint distributions.
- Use independence to calculate probabilities.
- Quantify the level of dependence using covariance and correlation.
- Calculate the expectation and variance of multiple random variables, independent or not.

Motivation: What should I expect?

Consider two jointly distributed random variables (X, Y). What should I expect X to be? What should I expect Y to be? What should I expect X to be if I already know what Y is? Does that change or does it stay the same?

Motivation: Dependence

Knowing whether the value of X affects Y or not is important. Consider, for example, a movie studio planning a series of superhero movies. If the first movie is unsuccessful, and is panned by critics and the audience, then the studio may want to rethink the sequel and subsequent movies. We would want to know the level of dependence between two random variables to help us make decisions better.

Expectations and variances

Once again, we have discussed expectations and variances before; however, those were in the setting of single random variables. Here, we generalize in two or more random variables. As a motivating example, consider a student taking two classes: the student may be interested in the expected grade in one of the two classes alone.

This may ring a bell. Recall that during our last lecture, we discussed the **marginal pmf/pdf** of jointly distributed random variables (X, Y). We specifically said that they come in play when we want to answer the question "what is the probability that X takes a certain

value, regardless of Y?" Well, we will use this to calculate expectations and variances. Specifically, we want to answer the questions:

1. "what is the expected value that X takes, regardless of Y?"
2. "what is the variance of X, regardless of Y?"

Of course, both are easily adapted for Y (regardless of X).

Let (X, Y) be two jointly distributed random variables with marginal pmf/pdf $f_X(x)$ and $f_Y(y)$. Then:

	Discrete	Continuous	
$E[X] =$	$\sum_x x f_X(x)$	$= \int_{-\infty}^{+\infty} x f_X(x) dx$	$= \mu_X$
$Var[X] =$	$\sum_x x^2 f_X(x) - \mu_X^2$	$= \int_{-\infty}^{+\infty} x^2 f_X(x) dx - \mu_X^2$	$= \sigma_X^2$
$E[Y] =$	$\sum_y y f_Y(y)$	$= \int_{-\infty}^{+\infty} y f_Y(y) dy$	$= \mu_Y$
$Var[Y] =$	$\sum_y y^2 f_Y(y) - \mu_Y^2$	$= \int_{-\infty}^{+\infty} y^2 f_Y(y) dy - \mu_Y^2$	$= \sigma_Y^2$

Applying the formulae: a chemical mixture

Last time we saw a chemical mixture problem with the volumes of two materials. Here, X and Y are continuous random variables between 0 and 1 representing the material volumes with joint pdf $f_{XY}(x, y) = \frac{2}{5}(2x + 3y)$. Recall that last time we calculated the marginal pdf for X and Y as $f_X(x) = \frac{1}{5}(4x + 3)$ and $f_Y(y) = \frac{6y+2}{5}$.

1. What is the expectation and the variance of the volume of the first material?
2. What is the expectation and the variance of the volume of the second material?

Applying the formulae: a chemical mixture

1. What is the expectation and the variance of the volume of the first material?

   Recall that $f_X(x) = \frac{1}{5}(4x+3)$:

   $$E[X] = \int_0^1 x \cdot \frac{1}{5}(4x+3)\,dx = \frac{17}{30} = 0.566\ldots$$

   $$Var[X] = \int_0^1 x^2 \cdot \frac{1}{5}(4x+3)\,dx - (E[X])^2 = \frac{71}{900} = 0.0788\ldots$$

2. What is the expectation and the variance of the volume of the second material?

   Recall that $f_Y(y) = \frac{6y+2}{5}$:

   $$E[Y] = \int_0^1 y \cdot \frac{6y+2}{5} dy = 0.6$$

   $$Var[Y] = \int_0^1 y^2 \cdot \frac{6y+2}{5} dy - (E[Y])^2 = \frac{11}{150} = 0.073\ldots$$

Practice with the following joint distributions:

- Discrete: $f_{XY}(x,y) = \frac{x \cdot y}{18}$, $x = 1,2$, $y = 1,2,3$.
- Continuous: $f_{XY}(x,y) = x + y$, $0 \le x \le 1, 0 \le y \le 1$.

Conditional expectations and variances

Conditional means and variances can also be defined in the case of two jointly distributed random variables X, Y. They would answer the question: "what should I expect X to be if I know that Y is equal to y?" Clearly, the same question can be asked for Y.

The conditional mean and variance of random variable X given a value for random variable $Y = y$ and of random variable Y given $X = x$ are:

	Discrete	Continuous	
$E[X\|y]$	$=\sum_x x f_{X\|y}(x)$	$=\int_{-\infty}^{+\infty} x f_{X\|y}(x)dx$	$=\mu_{X\|y}$
$Var[X\|y]$	$=\sum_x x^2 f_{X\|y}(x) - \mu^2_{X\|y}$	$=\int_{-\infty}^{+\infty} x^2 f_{X\|y}(x)dx - \mu^2_{X\|y}$	$=\sigma^2_{X\|y}$
$E[Y\|x]$	$=\sum_y y f_{Y\|x}(y)$	$=\int_{-\infty}^{+\infty} y f_{Y\|x}(y)dy$	$=\mu_{Y\|x}$
$Var[Y\|x]$	$=\sum_y y^2 f_{Y\|x}(y) - \mu^2_{Y\|x}$	$=\int_{-\infty}^{+\infty} y^2 f_{Y\|x}(y)dy - \mu^2_{Y\|x}$	$=\sigma^2_{Y\|x}$

Back to the chemical mixture

Recall that we had calculated during the previous lecture that $f_{X|y}(x) = \frac{4x+6y}{6y+2}$.

What is the expectation and the variance of the volume of the first material, given that the second material's volume is equal to 0.6?

We have:

$$E[X|y] = \int_0^1 x \frac{4x + 6 \cdot 0.6}{6 \cdot 0.6 + 2} dx = 0.5595$$

$$Var[X|y] = \int_0^1 x^2 \frac{4x + 6 \cdot 0.6}{6 \cdot 0.6 + 2} dx - 0.5595^2 = 0.0799.$$

Find the conditional distributions of the two joint distributions below.

- Discrete: $f_{XY}(x,y) = \frac{x \cdot y}{18}$, $x = 1,2$, $y = 1,2,3$.
- Continuous: $f_{XY}(x,y) = x + y$, $0 \le x \le 1, 0 \le y \le 1$.

Then, find:

- For the first one (the discrete distribution): $E[X|Y=2]$,$Var[X|Y=2]$.
- For the second one (the continuous distribution): $E[Y|X=0.5]$,$Var[Y|X=0.5]$.

Expectations of functions

Finally, as far as expectations are concerned, we take a look at the expectation of a function. Let $h(X,Y)$ be a function of two jointly distributed random variables X, Y. Very similarly to what we did for expectations of functions of single random variables, we get: [63]

$$\text{discrete :} \qquad E\left[h(X,Y)\right] = \sum_x \sum_y h(x,y) f_{XY}(x,y)$$

$$\text{continuous :} \qquad E\left[h(X,Y)\right] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} h(x,y) f_{XY}(x,y) dxdy$$

[63] Recall that for random variable X and function $g(X)$, we have:

$$\text{discrete :} \qquad E\left[g(X)\right] = \sum_x g(x)p(x)$$

$$\text{continuous :} \qquad E\left[g(X)\right] = \int_{-\infty}^{+\infty} g(x)f(x)dx$$

Again with the chemical mixture

The chemical mixture's quality is evaluated by function $h(X,Y) = 3X + 7Y$. We note here that material 2 is preferable to material 1, and hence the quality is severely favored when material 2 has higher volume. The maximum possible quality is equal to 10 (when both material volumes are equal to 1): in general, the quality ranges from 0 to 10.

What is the expected mixture quality?

We can calculate this as:

$$\begin{aligned} E\left[h(X,Y)\right] &= \int_0^1 \int_0^1 (3x+7y)\,\frac{2}{5}\,(2x+3y)\,dxdy = \\ &= \int_0^1 \frac{1}{5}\left(42y^2+23y+4\right)dy = \\ &= 5.9. \end{aligned}$$

Independence

Independence is a fundamental property of events and random variables. In the past [64] we have discussed independence for events: events A and B are independent if

$$P(A|B) = P(A) \;\text{ or }\; P(A \cap B) = P(A) \cdot P(B).$$

This property proved very useful for calculating basic probabilities. Now, we extend its definition to jointly distributed random variables.

[64] See Lecture 3.

Definition 35 (Independence for random variables) *Two random variables X, Y are independent if any of the following statements hold:*

1. $f_{XY}(x,y) = f_X(x)f_Y(y), \forall x, y$
2. $f_{X|y}(x) = f_X(x), \forall x, y$
3. $f_{Y|x}(y) = f_Y(y), \forall x, y$
4. $P(X \in A, Y \in B) = P(X \in A) \cdot P(Y \in B), \forall A, B$

In English, the four statements say the same thing; but they do it in different ways. The first one claims that two independent random variables will see their joint pmf/pdf be equal to the product of the individual marginal pmfs/pdfs. **This is typically the easiest way to show (or not) independence of two random variables.** If we find the marginal pmf/pdf of X and X and their product does not always equate to their joint pmf/pdf, then the two variables are not independent.

The second and the third statements are similar to the first definition of independence of events. In essence, they claim that the conditional pmf/pdf of one random variable given the other is equal to the marginal pmf/pdf.

The last statement is interesting, but it needs to be shown for any two sets of values A and B. It states that the probability of both X belonging to set A and Y belonging to set B can be found through the product of the individual probabilities. The last statement is very similar to the first, but instead focuses on sets of values rather than the pmf/pdf.

Discrete random variable independence

Consider two discrete random variables with joint pmf $f_{XY}(x,y) = e^{-3}\frac{2^x}{x!\cdot y!}$ for $x, y \geq 0$. Are random variables X and Y independent?

This looks very intimidating, but we can use some well-known facts from calculus to obtain the answer. Recall that $\sum_{i=0}^{\infty} \frac{\alpha^i}{i!} = e^{\alpha}$. This will be useful.
Now, to find the marginal pmfs:

1. $f_X(x) = \sum_{y=0}^{\infty} e^{-3}\frac{2^x}{x!\cdot y!} = e^{-3}\frac{2^x}{x!} \cdot e = e^{-2}\frac{2^x}{x!}$.
2. $f_Y(y) = \sum_{x=0}^{\infty} e^{-3}\frac{2^x}{x!\cdot y!} = e^{-3}\frac{1}{y!} \cdot \sum_{x=0}^{\infty} \frac{2^x}{x!} = e^{-3}\frac{1}{y!} \cdot e^2 = e^{-1}\frac{1}{y!}$.

We then observe that $f_{XY}(x,y) = f_X(x) \cdot f_Y(y)$, showing independence.

We could have used statements 2 and 3 to show the same thing!

Discrete random variable independence

Consider two discrete random variables with joint pmf $f_{XY}(x,y) = e^{-3}\frac{2^x}{x!\cdot y!}$ for $x, y \geq 0$. Are random variables X and Y independent?

We could first find the conditional pmfs:

1. $f_{X|y}(x) = \frac{f_{XY}(x,y)}{f_Y(y)} = \frac{e^{-3}\frac{2^x}{x!\cdot y!}}{e^{-1}\frac{1}{y!}} = e^{-2}\frac{2^x}{x!} = f_X(x)$.

2. $f_{Y|x}(y) = \frac{f_{XY}(x,y)}{f_X(x)} = \frac{e^{-3}\frac{2^x}{x!\cdot y!}}{e^{-2}\frac{2^x}{x!}} = e^{-1}\frac{1}{y!} = f_Y(y)$.

We then observe that $f_{X|y}(x) = f_X(x)$ and $f_{Y|x} = f_Y(y)$, showing independence again.

The last statement would prove a little tougher, as we would need to show that it is true for any pair of sets of values. Let's see an example where independence does not hold.

Continuous random variable independence

Consider two continuous random variables with joint pdf $f_{XY}(x,y) = x + y, 0 \leq x \leq 1, 0 \leq y \leq 1$. Are the random variables independent?

Similarly to the previous example, let's calculate the marginal pdfs:

1. $f_X(x) = \int_{y=0}^{1} (x+y)\,dy = xy + \frac{y^2}{2}\Big|_0^1 = x + \frac{1}{2}$.

2. Similarly, $f_Y(y) = y + \frac{1}{2}$.

We now note that $f_X(x) \cdot f_Y(y) = x \cdot y + \frac{1}{2}x + \frac{1}{2}y + \frac{1}{4}$, which does not reveal independence, as the product is not always equal to $x + y$. At this point we may claim that the two events are not independent. The same can be said using statements 2 and 3:

1. $f_{X|y}(x) = \frac{x+y}{x+\frac{1}{2}}$.

2. Similarly, $f_{Y|x}(y) = \frac{x+y}{y+\frac{1}{2}}$.

Again, these two are not necessarily equal to $f_X(x)$ or $f_Y(y)$, respectively.

Last, let us consider statement 4: if we are able to find at least one pair of values A, B such that $P(X \in A, Y \in B) \neq P(X \in A) \cdot P(Y \in B)$ should be enough to disprove independence.

Continuous random variable independence

Consider two continuous random variables with joint pdf $f_{XY}(x,y) = x + y, 0 \leq x \leq 1, 0 \leq y \leq 1$. Are the random variables independent?

Consider $A = [0, 0.25]$ and $B = [0.5, 1]$. We have:

$$P(X \in A, Y \in B) = \int_0^{0.25} \int_{0.5}^{1} (x+y)\, dydx = \\ = \int_0^{0.25} \left(0.5x + \frac{3}{8}\right) dx = 0.109375.$$

On the other hand:

1. $P(X \in A) = \int_0^{0.25} f_X(x)dx = \int_0^{0.25} \left(x + \frac{1}{2}\right) dx = 0.15625.$
2. $P(Y \in B) = \int_{0.5}^{1} f_Y(y)dy = \int_{0.5}^{1} \left(y + \frac{1}{2}\right) dy = 0.625.$

We observe that $P(X \in A) \cdot P(Y \in B) = 0.15625 \cdot 0.625 = 0.09765625$ which is not equal to $P(X \in A, Y \in B) = 0.109375$. Hence, the two random variables are not independent.

Alright, so we are able to tell if two random variables are independent or not. Another useful metric though, would be to be able to tell **how dependent** two random variables are.

Covariance

Definition 36 (Covariance) *Covariance is a measure of the* association *between two random variables. For two random variables X and Y, we define* covariance *as:*

$$\sigma_{XY} = Cov[X, Y] = E[(X - E[X]) \cdot (Y - E[Y])] = E[XY] - E[X] \cdot E[Y].$$

Remember the definition of variance for a single random variable? It was:

$$\sigma_X^2 = Var[X] = E\left[(X - E[X])^2\right] = E\left[X^2\right] - (E[X])^2.$$

This is very similar to the definition of covariance, extended to include jointly distributed random variables.

A few observations that we may make based on the definition follow:

1. If $X \geq E[X]$ whenever $Y \geq E[Y]$ and if $X \leq E[X]$ whenever $Y \leq E[Y]$, then the covariance will be positive.
2. If $X \geq E[X]$ whenever $Y \leq E[Y]$ and if $X \leq E[X]$ whenever $Y \geq E[Y]$, then the covariance will be negative.
3. Finally, and very importantly: **two independent random variables X, Y will have $\sigma_{XY} = Cov[X, Y] = 0$. The inverse is not necessarily true.**

A small Florida example

In Gainesville, FL, summer days are classified as either sunny or rainy. Whenever it is sunny, Floridians go to watch a local baseball team; whenever it is rainy, they tend to forget their umbrellas and they need to buy one. If it is rainy, profits skyrocket for an umbrella selling grocery store and they make \$4,500; at the same time, the local team only makes \$1,000. If it is sunny, the grocery store only makes \$500; the team though makes \$2,500 from tickets. Summers are sunny in Florida 65% of the time, and rainy the remaining 35%. What is the covariance of the two company profits?

Let U be the umbrella profits, and T the team profits. To help us collect all data we may construct a small table as follows:

	Sunny	Rainy
Probability	0.65	0.35
Team (T)	\$2500	\$1000
Umbrellas (U)	\$500	\$4500

- First, calculate the expected profits:

$$E[U] = 0.65 \cdot 500 + 0.35 \cdot 4500 = \$1900$$
$$E[T] = 0.65 \cdot 2500 + 0.35 \cdot 1000 = \$1975.$$

A small Florida example

- Now, on to calculating $(X - E[X]) \cdot (Y - E[Y])$:
  - when it is sunny:

$$(U - E[U]) \cdot (T - E[T]) =$$
$$= (500 - 1900) \cdot (2500 - 1975) = -805000.$$

  - when it is rainy:

$$(U - E[U]) \cdot (T - E[T]) =$$
$$= (4500 - 1900) \cdot (1000 - 1975) = -2340000.$$

- Last, calculate the covariance:

$$Cov(U, T) = E[(U - E[U]) \cdot (T - E[T]) = (4500 - 1900)] =$$
$$= 0.65 \cdot (-805000) + 0.35 \cdot (-2340000) =$$
$$= -1342250.$$

Covariance is easier calculate as $E[X \cdot Y] - E[X] \cdot E[Y]$ when the probabilities are given in joint pmf/pdf format.

Covariance for continuous random variables

Earlier, we saw an example of two jointly distributed continuous random variables X and Y with pdf $f_{XY}(x, y) = x + y, 0 \leq x \leq 1, 0 \leq y \leq 1$. We actually calculated their marginal pdfs:

- $f_X(x) = x + \frac{1}{2}$.
- $f_Y(y) = y + \frac{1}{2}$.

We also found that these two are not independent. What is their covariance?

Covariance for continuous random variables

Recall that we can calculate the covariance of X and Y as $E[XY] - E[X] \cdot E[Y]$.

- $E[X] = \int_0^1 x\left(x + \frac{1}{2}\right) dx = \frac{7}{12}$.
- $E[Y] = \int_0^1 y\left(y + \frac{1}{2}\right) dy = \frac{7}{12}$.
- Also: $E[XY] = \int_0^1 \int_0^1 xy\,(x+y)\, dydx = \frac{1}{3}$.

Finally:

$$\sigma_{XY} = \frac{1}{3} - \frac{7}{12} \cdot \frac{7}{12} = -\frac{1}{144}.$$

Correlation

The problem with covariance is that it is not normalized. A very big covariance or a very small covariance do not necessarily imply the actual level of dependence. This is why we introduce **correlation**, a measure that directly relates its value to the magnitude of dependence.

Definition 37 (Correlation) *Correlation is a measure of the linear relationship between two random variables X and Y. It is calculated as:*

$$\rho_{XY} = \frac{Cov[X,Y]}{\sqrt{Var[X]}\sqrt{Var[Y]}} = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y}.$$

By definition, $-1 \leq \rho_{XY} \leq 1$.

Notice that the numerator of correlation is the covariance itself, normalized by the product of the individual standard deviations (square roots of the variances). A few observations we can make based on this definition.

1. When X and Y are independent then $\sigma_{XY} = \rho_{XY} = 0$.
2. $\rho_{XY} = 1$ implies that X and Y are fully positively correlated.
3. $\rho_{XY} = -1$ implies that X and Y are fully negatively correlated.

Back to Florida

What is the correlation in the Florida example?

First, we calculate individual variances:

- $Var[U] = 0.65 \cdot (500 - 1900)^2 + 0.35 \cdot (4500 - 1900)^2 = 3640000$.
- $Var[T] = 0.65 \cdot (2500 - 1975)^2 + 0.35 \cdot (1000 - 1975)^2 = 508471.25$.

Since we already calculated $Cov(U, T) = -1342250$, we can compute the correlation as:

$$\rho_{UT} = \frac{-1342250}{\sqrt{3640000} \cdot \sqrt{508471.25}} = -0.9866.$$

Hence, as expected, the two profits are almost totally negatively correlated!

Correlation for continuous random variables

Another example we saw earlier: $f_{XY}(x, y) = x + y, 0 \leq x, y \leq 1$ with:

- $f_X(x) = x + \frac{1}{2}$.
- $f_Y(y) = y + \frac{1}{2}$.
- $E[X] = \frac{7}{12}$.
- $E[Y] = \frac{7}{12}$.
- $\sigma_{XY} = -\frac{1}{144}$.

What is the correlation?
First, we find the variances:

$$Var[X] = \int_0^1 x^2 f_X(x) dx - (E[X])^2 = \int_0^1 x^2 \left(x + \frac{1}{2}\right) dx - \left(\frac{7}{12}\right)^2 =$$
$$= \frac{5}{12} - \left(\frac{7}{12}\right)^2 = \frac{11}{144}.$$

We may similarly calculate $Var[Y] = \frac{11}{144}$, too. Finally:

$$\rho_{XY} = \frac{-\frac{1}{144}}{\sqrt{\frac{11}{144}} \cdot \sqrt{\frac{11}{144}}} = -\frac{1}{11}.$$

When x and y restrict each other

This serves as more of a reminder from calculus. When taking the integral of more than one variable at the same time, we need to be very careful with the bounds we are using.

Let us see this with an example. Assume (X, Y) are two jointly distributed continuous random variables with joint pdf equal to $f_{XY}(x, y) = 3 \cdot (x + y)$. Moreover, assume that $X, Y \geq 0$ and (and this is important!) $X + Y \leq 1$.

Note how the value of random variable X affects the range of values that Y is allowed to take; and vice versa. We need to be very careful with how we proceed in this case. There are three things we need to be able to do.

1. Calculate probability for **both random variables at the same time** and expectations for a function of both random variables.
2. Calculate marginal distributions **for one random variable at a time**.
3. Calculate probabilities, expectations, and variances **for one random variable (forgetting the other exists)** and calculate conditional probabilities, expectations, and variances **for one random variable setting the other equal to a value**.

Let us specifically focus on these three items for the pdf provided: $f_{XY}(x, y) = 3 \cdot (x + y)$ for $x, y \geq 0$, such that $x + y \leq 1$.

Both at the same time

Typical questions:

- Verify that $f_{XY}(x, y)$ is a valid pdf.
- What is the probability that $X \leq 0.3$ and $Y > 0.5$?

To answer these questions, we need to allow x and y to consider all values they can get. For example, if x is allowed to go from 0 to 1, then y is allowed to go from 0 to $1 - x$. On the other hand, if y is allowed to go from 0 to 1, then x is only allowed to go from 0 to $1 - y$.

What we could do wrong: we could possibly allow both x and y to go from 0 to 1, allowing $x + y$ to potentially go higher than 1, breaking the requirement.

Both at the same time

Finally:

- Verify that $f_{XY}(x,y)$ is a valid pdf.

$$\int_0^1 \int_0^{1-x} 3\cdot(x+y)\,dydx = \int_0^1 3\cdot\left(xy+\frac{y^2}{2}\right)\Big|_0^{1-x} dx =$$
$$= \int_0^1 3\cdot\left(\frac{1}{2}-\frac{x^2}{2}\right)dx = \left(\frac{3}{2}x-\frac{1}{2}x^3\right)\Big|_0^1 = 1,$$

or

$$\int_0^1 \int_0^{1-y} 3\cdot(x+y)\,dxdy = \int_0^1 3\cdot\left(\frac{x^2}{2}+yx\right)\Big|_0^{1-y} dy =$$
$$= \int_0^1 3\cdot\left(-\frac{y^2}{2}+\frac{1}{2}\right)dy = \left(-\frac{1}{2}y^3+\frac{3}{2}y\right)\Big|_0^1 = 1.$$

- What is the probability that $X \le 0.3$ and $Y > 0.5$?

$$\int_0^{0.3} \int_{0.5}^{1-x} 3\cdot(x+y)\,dydx = \int_0^{0.3} 3\cdot\left(xy+\frac{y^2}{2}\right)\Big|_{0.5}^{1-x} dx =$$
$$= \int_0^{0.3} \left(1.125-1.5x-1.5x^2\right)dx = \left(\frac{9x}{8}-\frac{3x^2}{4}-\frac{x^3}{2}\right)\Big|_0^{0.3} =$$
$$= 0.2565.$$

Let us now check the second case: one at a time.

One at a time

Typical questions:

- Find the marginal distribution of X.
- Find the marginal distribution of Y.

To answer the questions, we need to either express y as a function of x or express x as a function of y.

What we could do wrong: we could possibly allow both x and y to go from 0 to 1, allowing $x + y$ to potentially go higher than 1, breaking the requirement.
Let's see how we could go about solving this:

- Find the marginal distribution of X.

$$f_X(x) = \int_0^{1-x} 3 \cdot (x + y)\, dy = \frac{3}{2} - \frac{3x^2}{2}.$$

- Find the marginal distribution of Y.

$$f_Y(y) = \int_0^{1-y} 3 \cdot (x + y)\, dy = \frac{3}{2} - \frac{3y^2}{2}.$$

The third case involves forgetting that one of the variables exist.

One alone

Typical questions:

- Find the probability that $X \leq 0.3$.
- What is the expectation of Y?

To answer the questions, we simply use the bounds as given!

What we could do wrong: we could possibly try to restrict x or y as a function of the other, when the other no longer exists – as we do not care about it in the setup.

One alone

- Find the probability that $X \leq 0.3$.

$$\int_0^{0.3} f_X(x)dx = \int_0^{0.3} \left(\frac{3}{2} - \frac{3x^2}{2}\right) dx = 0.4365.$$

- What is the expectation of Y?

$$\int_0^1 y f_Y(y)dy = \int_0^1 y \left(\frac{3}{2} - \frac{3y^2}{2}\right) dx = 0.375.$$

Extension to more than 2 random variables

Everything we have discussed today can be generalized to more than 2 random variables. [65] More specifically, for a multivariate jointly distributed random variable $(X_1, X_2, \ldots, X_n)$ (hence n random variables), with joint pmf/pdf $f(x_1, x_2, \ldots, x_n)$, we have expectations and variances:

[65] This will prove useful for Lecture 13.

Discrete:

$$E[X_i] = \sum_x x_i f_{X_i}(x_i) = \mu_{X_i}$$
$$Var[X_i] = \sum_x (x_i - \mu_{X_i})^2 f_{X_i}(x_i) = \sigma^2_{X_i}$$

Continuous:

$$E[X_i] = \int_{-\infty}^{+\infty} x_i f_{X_i}(x_i)\, dx_i = \mu_{X_i}$$
$$Var[X_i] = \int_{-\infty}^{+\infty} (x_i - \mu_{X_i})^2 f_{X_i}(x_i)\, dx_i = \sigma^2_{X_i}$$

Similarly, for independence:

Definition 38 (Independence) *Random variables $X_1, X_2, \ldots, X_n$ are independent if and only if for all values of $x_1, x_2, \ldots, x_n$, we have that*

$$f_{X_1 X_2 \ldots X_n}(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1) \cdot f_{X_2}(x_2) \cdots f_{X_n}(x_n).$$

A 4-component machine

Recall the machine (from Lecture 11) that consists of four components, whose lifetimes (in years) are jointly distributed with the following pdf:

$$f_{X_1X_2X_3X_4}(x_1, x_2, x_3, x_4) = 3 \cdot e^{-2x_1}e^{-x_2}e^{-3x_3}e^{-0.5x_4}.$$

Are the random variables from the pdf in this example independent?

First, we calculate all 4 marginal pdfs:

1. $f_{X_1}(x_1) = 2e^{-2x_1}$
2. $f_{X_2}(x_2) = e^{-x_2}$
3. $f_{X_3}(x_3) = 3e^{-3x_3}$
4. $f_{X_4}(x_4) = 0.5e^{-0.5x_4}$

Finally, we have that

$$f_{X_1X_2X_3X_4}(x_1, x_2, x_3, x_4) = f_{X_1}(x_1) f_{X_2}(x_2) f_{X_3}(x_3) f_{X_4}(x_4),$$

so they are independent.

We finish today's notes with the following **very important result**: For a series of random variables, we have:

$$\begin{aligned} E\left[\sum_{i=1}^{n} a_i X_i\right] &= \sum_{i=1}^{n} a_i E[X_i] \\ Var\left[\sum_{i=1}^{n} a_i X_i\right] &= \sum_{i=1}^{n} a_i^2 Var[X_i] + \sum_{i=1}^{n}\sum_{j=1:i\neq j}^{n} Cov[X_i, X_j] \\ &= \sum_{i=1}^{n} a_i^2 Var[X_i] + 2a_i a_j \cdot \sum_{i=1}^{n}\sum_{i<j} Cov[X_i, X_j] \end{aligned}$$

Since when we have independence, $Cov[X_i, X_j] = 0$ for all X_i, X_j, then, for multiple independent random variables, we have: [66]

[66] We had already derived this result! Check Lecture 9.

$$\begin{aligned} E\left[\sum_{i=1}^{n} a_i X_i\right] &= \sum_{i=1}^{n} a_i E[X_i] \\ Var\left[\sum_{i=1}^{n} a_i X_i\right] &= \sum_{i=1}^{n} a_i^2 Var[X_i] \end{aligned}$$

Joint distributions: common distributions

Learning objectives

After these lectures, we will be able to:

- Find the pmf/pdf of a function of a random variable.
- Recognize multinomial distributions.
- Calculate probabilities (including marginal and conditional ones) for multinomially distributed random variables.
- Recognize bivariate normal distributions.
- Describe and explain bivariate normal distributions and their correlations.
- Calculate probabilities (including marginal and conditional ones) for bivariate normally distributed random variables.

Motivation: Success or failure? More like full success, or somewhat success, or ...

In Lectures 5-6, we introduced a lot of discrete distributions. One of the most fundamental ones is the binomial distribution. Its premise is simple: perform an experiment n times and count the number of successes, assuming the remainders are failures. This works pretty well when we have two outcomes: for example, a patient may have an infection or not, a student may pass a class or not, etc.

What happens when the number of outcomes is higher than 2? What if a patient may have a severe infection, or a moderate infection, or no infection? What if a student can get an A, a B, a C, a D, or fail a class?

Motivation: Normally distributed random variables with correlation

We sometimes are aware that a specific random variable is normally distributed. However, its exact parameters may depend (and in turn may also affect) another normally distributed random variable. For example, consider a Sunday night at HBO. A TV series starting at 10pm may expect a normally distributed share of viewers with a known mean and standard deviation. However, if the TV series showing at 9pm has its grand finale, we may anticipate a higher viewership for the 10pm show, too! This relationship needs to be modeled somehow...

Distribution of a function

We have already discussed what we expect will happen for a function of a random variable. [67] We repeat the definitions here for convenience:

[67] Recall Lecture 9 and the properties of expectation subsection.

Definition 39 (Expectation of a function of a random variable) *Let $g(X)$ be a function of a random variable X. Then, the expectation of $g(X)$ is denoted by $E[g(X)]$ and is equal to:*

- *for discrete random variable X with sample space S:*

$$E[g(X)] = \sum_{x \in S} g(x) \cdot p(x).$$

- *for continuous random variable X:*

$$E[g(X)] = \int_{-\infty}^{+\infty} g(x) \cdot f(x) dx.$$

It is time we discuss how the function is **distributed**: rather than addressing questions of expectation ("what should I expect the function value to be?"), we will be addressing questions of probability ("what is the probability the function value is...").

Some examples of why this would be useful:

- What is the probability my profits are higher than $2000 today?
  - My profits depend on the number of customers, discrete random variable X.
- What is the probability the circuit overheats?
  - The heat of the circuit is a function of its current, continuous random variable X.
- What is the probability the crop has high yield?
  - The yield of a crop is a function of the location temperature, continuous random variable X.

Formally, let $Y = h(X)$ be a **one-to-one** transformation of a random variable X to a random variable Y. The one-to-one transformation is important: it implies that solving $y = h(x)$ provides us with a unique solution. Assume that the solution is [68]

[68] Recall the definition of inverse functions.

$$x = h^{-1}(y) = u(y).$$

Examples of inverses

- $Y = X^2 \implies x = u(y) = \sqrt{y}$.
- $Y = 2\ln x \implies x = e^{y/2}$.

Definition 40 (Distribution of a function) *Let $Y = h(X)$ be a one-to-one function of random variable X to Y. X is distributed with pmf/pdf $f_X(x)$. Then, the pmf/pdf of random variable $Y = h(X)$ can be found using the **chain rule**:*

1. *Discrete X:* $f_Y(y) = f_X(u(y))$.
2. *Continuous X:* $f_Y(y) = f_X(u(y)) \cdot |u'(y)|$,
   where $u'(y)$ is the derivative of function $u(y)$.

Printer speed

A printer has speed that is equal to $h(x) = \frac{\sqrt{x+1}}{x+1}$, where x is the condition of the printer. The condition of the printer is a continuous random variable distributed exponentially with rate $\lambda = 1$. What is the probability the printer is faster than 0.5?

First of all, let's see what we have:

- X is the condition of the printer, a random variable with $f_X(x) = \lambda \cdot e^{-\lambda x}$ and $x \geq 0$.
- Y is the speed of the printer, a random variable which is a function of X and has $Y = h(x) = \frac{\sqrt{x+1}}{x+1}$. By definition $0 \leq y \leq 1$.
- We may solve for $u(y)$:

$$y = \frac{\sqrt{x+1}}{x+1} \implies x = 1 - \frac{1}{y^2} \implies u(y) = \frac{y^2-1}{y^2}.$$

Based on the chain rule: $f_Y(y) = e^{-\frac{y^2-1}{y^2}} \cdot \frac{2}{y^3}$. Finally, we have:

$$P(Y > 0.5) = \int_{0.5}^{1} e^{-\frac{y^2-1}{y^2}} \cdot \frac{2}{y^3} dy = 0.9502.$$

The multinomial distribution

Flashback! Let's review together the **binomial distribution**, one of the first discrete probability distributions we studied together. We had the following setup.

What if we perform n independent trials of the same experiment? Each trial may result in a success (with probability p) or a failure (with probability $q = 1 - p$). Let X be the number of successes we observe: then X is said to be binomially distributed with parameters n and p. Some interesting things about the binomial distribution:

- pmf: $P(X = x) = p(x) = \binom{n}{x} p^x (1-p)^{n-x}$, for $0 \leq x \leq n$.
- expectation: $E[X] = np$.
- variance: $Var[X] = np(1-p)$.

What if we tried to generalize this? We will still perform n independent trials; however now each trial will result in one of k outcomes (instead of just two). Each outcome appears with each own probability p_i. Clearly we need $\sum_{i=1}^{k} p_i = 1$. This is called the **multinomial distribution**. Formally:

Definition 41 (The multinomial distribution) *Let X_i be the number of times that outcome i appears in n independent trials. Each outcome i appears with probability p_i such that $\sum_{i=1}^{k} p_i = 1$. Then, $(X_1, X_2, \ldots, X_k)$ is distributed following a multinomial distribution with parameters n and $p_i, i = 1, \ldots, k$. The joint probability mass function is given by:*

$$P(X_1 = x_1, X_2 = x_2, X_3 = x_3, \ldots, X_k = x_k) = \frac{n!}{x_1! x_2! \ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}.$$

Note that $\sum_{i=1}^{k} x_i = n$.

Grades

According to the grade disparity website at UIUC, a student registering for CS 101 gets an A 73% of the time, a B 17% of the time, a C 6% of the time, a D 2% of the time, and an F the remaining 2% of the time. In a subsection of the class, there are 20 students. What is the probability that:

a) 10 students get an A and 10 students get a B?

b) everyone gets an A?

c) 12 students get an A, 5 students get a B, 2 students get a C, and 1 student gets a D?

Grades

This is a multinomial distribution with $n = 20$, $p_1 = 0.73$, $p_2 = 0.17$, $p_3 = 0.06$, $p_4 = 0.02$, $p_5 = 0.02$ for A, B, C, D, and F, respectively. Let X_1, X_2, X_3, X_4, X_5 be the number of students getting an A, B, C, D, F. Then, we have:

a) 10 students get an A and 10 students get a B?

$$P(X_1 = 10, X_2 = 10, X_3 = 0, X_4 = 0, X_5 = 0) = \\ = \frac{20!}{10!10!0!0!0!} 0.73^{10} 0.17^{10} 0.06^0 0.02^0 0.02^0 = \\ = 0.00016.$$

b) everyone gets an A?

$$P(X_1 = 20, X_2 = 0, X_3 = 0, X_4 = 0, X_5 = 0) = \\ = \frac{20!}{20!0!0!0!0!} 0.73^{20} 0.17^0 0.06^0 0.02^0 0.02^0 = \\ = 0.00185.$$

c) 12 students get an A, 5 students get a B, 2 students get a C, and 1 student gets a D?

$$P(X_1 = 12, X_2 = 5, X_3 = 2, X_4 = 1, X_5 = 0) = \\ = \frac{20!}{12!5!2!1!0!} 0.73^{12} 0.17^5 0.06^2 0.02^1 0.02^0 = \\ = 0.00495.$$

The marginal distribution

Let us derive the marginal distribution of $f_{X_1 X_2 \cdots X_k}(x_1, x_2, \ldots, x_k) = P(X_1 = x_1, X_2 = x_2, X_3 = x_3, \ldots, X_k = x_k)$ for the multinomial distribution.

First of all, a little notation. Like we did in Lectures 11 and 12, the marginal distribution of, say, X_i can be written as $f_{X_i}(x_i)$. Remember

that X_i is the random variable capturing the number of outcomes i we see in n tries.

Now, let outcome i be a "success"; otherwise, a "failure". Does this ring a bell? Also, remember that outcome i happens with probability p_i; everything else with probability $1 - p_i$. Hence, X_i is the number of successes we see in n independent tries. What is X_i distributed like when we view it like this?

The marginal distribution of X_i is the binomial distribution: i.e., every single one of the X_i is **binomially distributed** with parameters n, p_i.

Grades

For the example discussed earlier, what is the probability that:

a) 10 students get an A?

b) at most 1 students fails?

Also, how many students are expected to get each grade?

The first one is binomially distributed with $n = 20, p = 0.73$. The second one is binomially distributed with $n = 20, p = 0.02$. Overall, we have:

a) $P(X_1 = 10) = \binom{20}{10} 0.73^{10} 0.27^{10} = 0.01635$.

b) $P(X_5 \leq 1) = \binom{20}{0} 0.02^0 0.98^{20} + \binom{20}{1} 0.02^1 0.98^{19} = 0.6676 + 0.2725 = 0.9401$.

To answer the expectation question, if each outcome is binomially distributed with $n = 20$ and p_i, the expectations are:

a) A: 14.6 b) B: 3.4 c) C: 1.2 d) D: 0.4 e) F: 0.4

The conditional distribution

Now, let us consider the conditional distribution of the multinomial distribution. Say that among all outcomes we already know that X_j has happened x_j times. This implies that there is no uncertainty about x_j of the n tries. Let's keep that in mind.

Furthermore, if we were to remove these x_j outcomes, what we are left with is $n - x_j$ tries; however these tries do not have all k outcomes happening, but instead only $k - 1$.

Let us consider this with an example: if we have 20 students tak-

ing a class, and we know that 15 students ended up with an A, then the stochastic nature of this distribution only affects the remaining 5 students – after we remove the students whose grade is known to be an A.

Finally, in the remaining outcomes, we know that x_j is missing. However, we also know that if we sum the probabilities of all outcomes we should be getting 1. In the case of the grades from before, after we remove the As, we get a summation of probabilities equal to $p_2 + p_3 + p_4 + p_5 = 0.27 \neq 1$. To fix this issue, we renormalize the remainder of the probabilities. Instead of p_i, we now use $q_i = \frac{p_i}{\sum_{\ell \neq j} p_\ell}$.

Summing up:

> The conditional distribution of $X_1, X_2, \ldots, X_{j-1}, X_{j+1}, \ldots, X_k$ given $X_j = x_j$ is the multinomial distribution again but with parameters $n - x_j$, $q_i = \frac{p_i}{\sum_{\ell \neq j} p_\ell}$.

Grades

Back to the example we have been using. We have just been informed that 3 students failed. What is the probability that:

a) 10 students get an A and 7 students get a B?

b) 10 students get an A, 5 students get a B, and 2 students get a C?

Both are multinomial distributions with parameters $n = 20 - 3 = 17$ and $q_1 = \frac{p_1}{p_1+p_2+p_3+p_4} = \frac{0.73}{0.98} = 0.7449$; $q_2 = \frac{p_2}{p_1+p_2+p_3+p_4} = \frac{0.17}{0.98} = 0.1735$; $q_3 = 0.0612$; and $q_4 = 0.0204$. Then, we have:

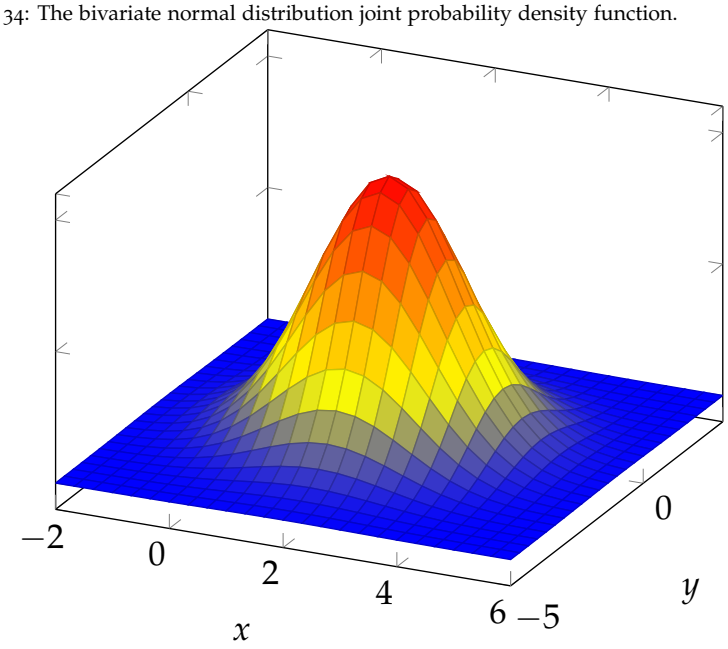
a) $P(X_1 = 10, X_2 = 7, X_3 = 0, X_4 = 0) = \frac{17!}{10!7!0!0!} 0.7449^{10} 0.1735^7 0.0612^0 0.0204^0 = 0.0048$.

b) $P(X_1 = 10, X_2 = 5, X_3 = 2, X_4 = 0) = \frac{17!}{10!5!2!0!} 0.7449^{10} 0.1735^5 0.0612^2 0.0204^0 = 0.01265$.

The bivariate normal distribution

Similarly to what we did for the binomial and its extension to the multinomial, we will also extend the normal distribution. What if, we have two jointly distributed variables that are *individually* normally distributed with their own means and variances? In essence, what if we have the three-dimensional pdf portrayed in Figure 34?

Figure 34: The bivariate normal distribution joint probability density function.



Formally, we provide the definition that follows:

Definition 42 (Bivariate normal distribution) *Consider two normally distributed random variables X, Y with means μ_X, μ_Y and variances σ_X^2, σ_Y^2. That is, $X \sim \mathcal{N}\left(\mu_X, \sigma_X^2\right)$ and $Y \sim \mathcal{N}\left(\mu_Y, \sigma_Y^2\right)$. We also assume that the two random variables are correlated with correlation ρ_{XY}.*

*Then, two random variables X and Y with the above parameters are **jointly distributed with a bivariate random distribution** if:*

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{XY}^2}} \cdot e^{\frac{-z}{2\left(1-\rho_{XY}^2\right)}},$$

where $z = \frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{2\rho_{XY}(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}$

We observe that ρ_{XY} plays an important role. When $\rho_{XY} = 0$, we have the simplified version of the bivariate random distribution as:

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y} \cdot e^{-\left(\frac{(x-\mu_X)^2}{2\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right)}.$$

However, we know that $\rho_{XY} = 0$ implies that X and Y are independent. And, for two independent random variables, we know that their joint pdf is equal to the product of the individual pdfs. Let's see if that is the case here:

Figure 35: The bivariate normal distribution and its contour plot. Here, we have that $\rho_{XY} = 0$.

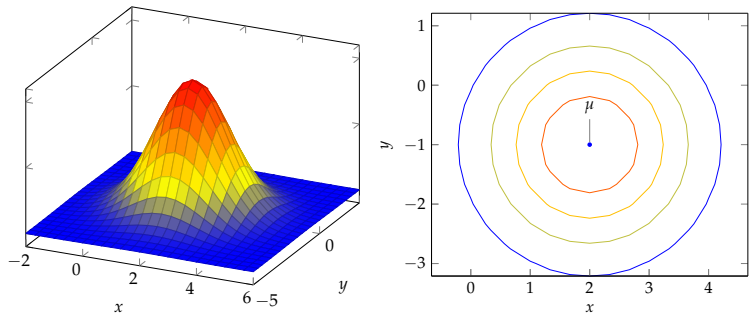
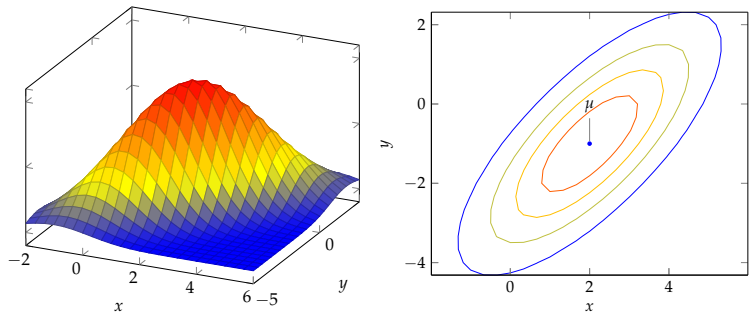


Figure 36: The bivariate normal distribution and its contour plot. Here, we have that $\rho_{XY} > 0$.



$$f_X(x) = \frac{1}{\sqrt{2\pi}\cdot\sigma_X} e^{-\frac{(x-\mu_X)^2}{2\sigma_X^2}}$$

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\cdot\sigma_Y} e^{-\frac{(y-\mu_Y)^2}{2\sigma_Y^2}}$$

$$f_X(x)\cdot f_Y(y) = \frac{1}{\sqrt{2\pi}\cdot\sigma_X} e^{-\frac{(x-\mu_X)^2}{2\sigma_X^2}} \cdot \frac{1}{\sqrt{2\pi}\cdot\sigma_Y} e^{-\frac{(y-\mu_Y)^2}{2\sigma_Y^2}} =$$

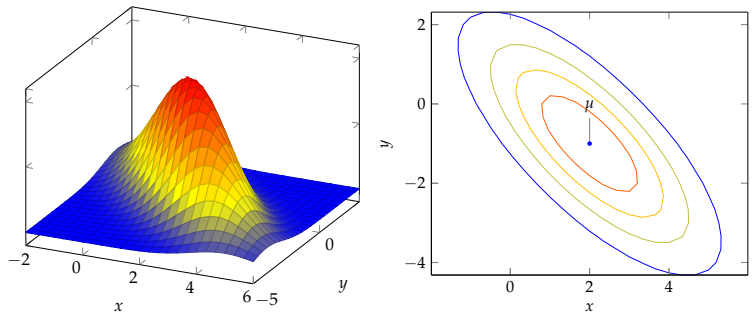
$$= \frac{1}{2\pi\sigma_X\sigma_Y}\cdot e^{-\left(\frac{(x-\mu_X)^2}{2\sigma_X^2}+\frac{(y-\mu_Y)^2}{\sigma_Y^2}\right)} = f_{XY}(x,y).$$

This independence is shown in Figure 35. What happens when $\rho > 0$ or $\rho < 0$? What happens when $\rho = 1$ or $\rho = -1$?

When we have positive correlation, this implies that higher/lower values of X will imply higher/lower values of Y and vice-versa (for Y and X). This is showcased with the pdf and the contour in Figure 36, which appears to be "positively" skewed.

On the other hand, if $\rho_{XY} < 0$, this means that higher/lower values of X will lead to lower/higher values of Y and vice-versa (for Y and X). This is exactly the opposite. This is again shown visually with the pdf and the contour in Figure 37, which appears to be "negatively" skewed.

Figure 37: The bivariate normal distribution and its contour plot. Here, we have that $\rho_{XY} = 0$.



What happens when $\rho = 1$ or $\rho = -1$? Let's leave this as food for thought.

The marginal and the conditional distribution

Much like what we did for the multinomial distribution, we may also derive the marginal and conditional distributions for the bivariate normal distribution. More specifically, both the *marginal* and the *conditional* distributions for the bivariate normal distribution are normal distributions themselves!

$$
\begin{aligned}
\text{Marginal pdf:} \quad & X \sim \mathcal{N}\left(\mu_X, \sigma_X^2\right) \\
& Y \sim \mathcal{N}\left(\mu_Y, \sigma_Y^2\right) \\
\text{Conditional pdf:} \quad & X|Y = y \sim \mathcal{N}\left(\mu_{X|Y=y}, \sigma_{X|Y=y}^2\right) \\
& \mu_{X|Y=y} = \mu_X + \rho_{XY}\left(\frac{\sigma_X}{\sigma_Y}\right)(y - \mu_Y) \\
& \sigma_{X|Y=y}^2 = \sigma_X^2\left(1 - \rho_{XY}^2\right) \\
& Y|X = x \sim \mathcal{N}\left(\mu_{Y|X=x}, \sigma_{Y|X=x}^2\right) \\
& \mu_{Y|X=x} = \mu_Y + \rho_{XY}\left(\frac{\sigma_Y}{\sigma_X}\right)(x - \mu_X) \\
& \sigma_{Y|X=x}^2 = \sigma_Y^2\left(1 - \rho_{XY}^2\right)
\end{aligned}
$$

For the conditional pdf, the question from earlier comes back. What if X and Y are independent? What if they are perfectly correlated ($\rho_{XY} = 1$ or $\rho_{XY} = -1$)?

We finish this lecture with a big, comprehensive example that combines information from Lectures 12 and 13, as well as Lecture 7. Pay close attention to the derivations and calculations that follow!

Bivariate normal distribution example

A class has two exams, both of which have grades that are normally distributed with $\mu_1 = 80, \mu_2 = 82.5$ and $\sigma_1^2 = 100, \sigma_2^2 = 225$. Finally the two exams are positively correlated with $\rho = 0.6$. What is the probability that:

a) a random student scores over 75 in Exam 2?

b) a random student scores over 75 in Exam 2 given that they scored an 85 in the first exam?

c) the sum of the two exams of a random student is less than or equal to 175?

d) a random student did better on the second exam than the first exam?

Let's get to it. Let X_1 be the grade of the first exam, and X_2 the grade of the second exam. Then:

a) We know that $X_2 \sim \mathcal{N}(82.5, 225)$. Hence:

- $z = \frac{75-82.5}{15} = -0.5$.
- $P(X_2 > 75) = 1 - P(X_2 \leq 75) = 1 - \Phi(z) = \Phi(-z) = \Phi(0.5) = 0.6915$.

b) We also know that $X_2|X_1 \sim \mathcal{N}\left(\mu_{X_2|X_1}, \sigma^2_{X_2|X_1}\right)$. We calculate:

- $\mu_{X_2|X_1=85} = \mu_{X_2} + \rho_{X_1X_2}\left(\frac{\sigma_{X_2}}{\sigma_{X_1}}\right)(85 - \mu_{X_1}) = 82.5 + 0.6 \cdot \frac{15}{10} \cdot 5 = 87$.
- $\sigma^2_{X_2|X_1=85} = \sigma^2_{X_2}\left(1 - \rho^2_{X_1X_2}\right) = 225 \cdot 0.64 = 144$.
- $z = \frac{75-87}{12} = -1$.
- $P(X_2 > 75|X_1 = 85) = 1 - P(X_2 \leq 75|X_1 = 85) = 1 - \Phi(z) = \Phi(-z) = \Phi(1) = 0.8413$.

Hence, knowing that the student did better than average in the first exam changes our perspective for their probability to do well in the second exam, too.

Bivariate normal distribution example

c) The sum of two normally distributed random variables is also normally distributed! Additionally, we have:

$$
\begin{aligned}
E\left[\sum_{i=1}^{n} a_i X_i\right] &= \sum_{i=1}^{n} a_i E\left[X_i\right] \\
Var\left[\sum_{i=1}^{n} a_i X_i\right] &= \sum_{i=1}^{n} a_i^2 Var\left[X_i\right] + \sum_{i=1}^{n} \sum_{j=1:i\neq j}^{n} a_i a_j Cov\left[X_i, X_j\right] \\
&= \sum_{i=1}^{n} a_i^2 Var\left[X_i\right] + 2a_i a_j \cdot \sum_{i=1}^{n} \sum_{i<j} Cov\left[X_i, X_j\right]
\end{aligned}
$$

In our case, we have two random variables, so:

$$
\begin{aligned}
E\left[X_1 + X_2\right] &= E\left[X_1\right] + E\left[X_2\right] = 162.5 \\
Var\left[X_1 + X_2\right] &= Var\left[X_1\right] + Var\left[X_2\right] + 2Cov\left[X_1, X_2\right] = \\
&= 325 + 2\sigma^2_{X_1 X_2}.
\end{aligned}
$$

To calculate $\sigma^2_{X_1 X_2}$ we use the definition of correlation (see Lecture 12):

$$
\rho_{X_1 X_2} = \frac{\sigma^2_{X_1 X_2}}{\sigma_{X_1}\sigma_{X_2}} \implies 0.6 = \frac{\sigma^2_{X_1 X_2}}{10 \cdot 15} = \sigma^2_{X_1 X_2} = 90.
$$

This leads to a final variance of $Var\left[X_1 + X_2\right] = 505$. Finally:

- $X_1 + X_2 \sim \mathcal{N}\left(162.5, 505\right)$.
- $z = \frac{175-162.5}{\sqrt{505}} = 0.56$.
- $P(X_1 + X_2 \leq 175) = \Phi(z) = \Phi(0.56) = 0.7123$.

d) For this question, we want $X_2 > X_1 \implies X_2 - X_1 > 0$. The difference of two normally distributed random variables is—again—normally distributed! Its details:

$$
\begin{aligned}
E\left[X_2 - X_1\right] &= E\left[X_2\right] - E\left[X_1\right] = 2.5 \\
Var\left[X_2 - X_1\right] &= Var\left[X_1\right] + Var\left[X_2\right] - 2Cov\left[X_1, X_2\right] = 125.
\end{aligned}
$$

- $X_2 + X_1 \sim \mathcal{N}\left(2.5, 125\right)$.
- $z = \frac{0-2.5}{\sqrt{125}} = -0.22$.
- $P(X_2 > X_1) = P(X_2 - X_1 > 0) = 1 - P(X_2 - X_1 \leq 0) = 1 - \Phi(z) = \Phi(-z) = \Phi(0.22) = 0.5871$.

Descriptive statistics

Learning objectives

After these lectures, we will be able to:

- Differentiate between populations and samples.
- Given a sample, calculate the sample mean, variance, range, and quartiles.
- Use graphical devices to present data, and more specifically:
  - histograms;
  - box plots;
  - scatter plots;
  - time series plots;
  - Q-Q plots.

Motivation: Summarizing information

We live in the era of big data. The size of the data we collect is doubling every 2 years (and this is a conservative estimate). When confronted with so much information, one way to make sense of it is to distill it in smaller, more manageable chunks. This is what we will be doing in this lecture.

Probabilities and statistics

In Lecture 3, we defined **probability** using the words "with every event, we associate a real number called probability to represent the likelihood of that event happening." We may use probability theory to help us address questions such as:

- How likely is it that we get a 6 and a 1 if we roll two dice?
- What is the probability that a patient survives a disease?

On the other hand, we define **statistics** as the field including all methods involved with collecting, describing, analyzing, interpreting data. We use statistics to answer questions such as:

- Are two dice fair?
- What is a good estimate for the mortality rate of a disease?

We show visually an example of the first question. Say we rolled dice multiple times and reported the average number we obtained

Figure 38: The average number obtained by rolling the blue dice.

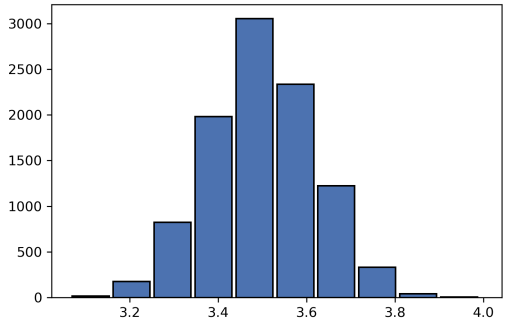
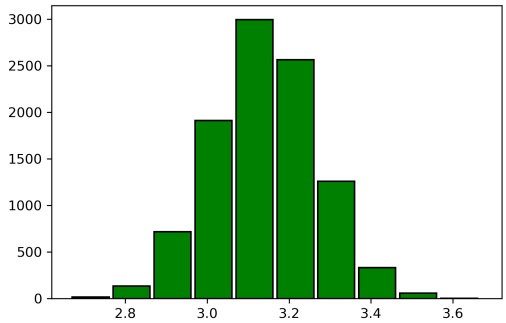


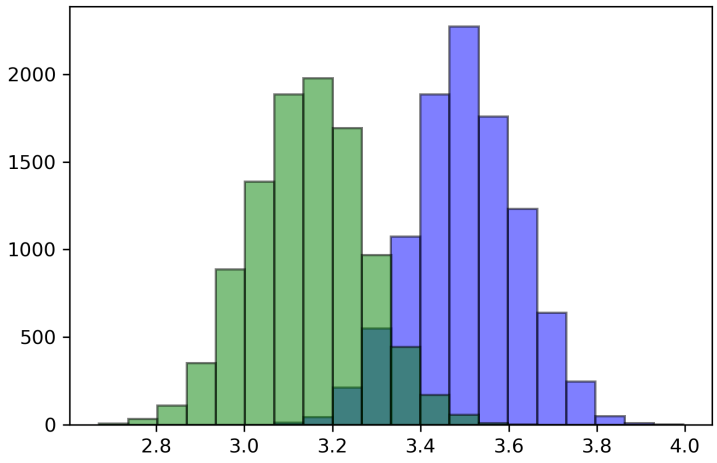
Figure 39: The average number obtained by rolling the green dice.



for each series of rolls. We have two dice: a green and a blue one and their results are show in Figures 39 and 38.

We want to make a decision about whether the two dice are fair. Let us plot both numbers together (see Figure 40). This makes it easy to compare them and deduce that the two dice do not look very similar. It appears the blue one is fair, with an average peak at 3.5 as expected, but the green one seems to favor smaller numbers, and thus unfair.

Figure 40: Plotting both dice average numbers at the same time to make comparisons easier.



Statistical methods

In the remainder of the semester, we will be dealing with statistical methods. We differentiate methods in three very important categories:

1. **Descriptive statistics**: methods to *describe* and *present* data.
2. **Inferential statistics**: methods to use observations in a smaller **sample** to *draw conclusions* for the larger **population**.
3. **Model building**: methods to build models to *predict* future data based on past observations.

The airline industry

The airline industry uses all three categories of statistical methods to help them guide decision-making based on available data. More specifically, example questions they use statistical methods include:

1. **Descriptive statistics**. Present the average delay for each route: this could be used to identify routes that are on average very late to depart from their origin or to arrive at their destination.
2. **Inferential statistics**. Select a subset of the routes to perform some prescriptive action: if the routes do indeed decrease their delays, can we claim that the action will work for all routes?
3. **Model building**. Build a model to predict delays: this is useful for identifying routes that are prone to be delayed and reroute passengers with connections that would be missed.

All three will be seen in subsequent lectures. However, for now, we will focus on **descriptive statistics** alone.

Descriptive statistics

This is the main part of today's lecture. We will specifically see two types of descriptive statistics: numerical and graphical.

What we will focus on in this lecture and in the worksheet is **descriptive statistics**. More specifically:

1. Numerical summaries of data.
   - sample mean, mode, median.
   - sample variance, standard deviation.
   - percentiles, quartiles, ranges.
2. Graphical displays of data.
   - Dot diagrams.
   - Histograms.
   - Stem-and-leaf diagrams.
   - Box plots.
   - Scatter diagrams.

- Time series plots.
- Q-Q plots.

Populations and samples

With the term **population** we refer to all possible observations we can collect. For example, a population could be the list of heights of every person in the world; or the SAT scores of every student in Illinois; or the time delays in all flights of a specific airline. The number of observations can grow to be very, very big and impractical to work with.

With the term **random sample** we refer to a subset of the observations selected from a population. For example, a sample could be the list of heights of 12 randomly selected people from our class; or the SAT scores of every student from a specific high school in Illinois; or the time delays in flights leaving ORD of a specific airline. This number of observations in the sample is expected to be significantly smaller than the population size, and hence, manageable to work with.

Formal definitions

More formally, assume a population X where each of its element is distributed with the same distribution (assume mean μ and variance σ^2). Then, a random sample is a set of randomly selected elements from X referred to as $X_1, X_2, \ldots, X_n$. Each X_i is independently selected, and comes from the same population X with mean μ and variance σ^2. Hence, we have:

- $E[X_i] = E[X] = \mu$.
- $Var[X_i] = Var[X] = \sigma^2$.

Numerical summaries of data

Sample mode

Definition 43 (Sample mode) *Given n observations $x_1, x_2, \ldots, x_n$ in a random sample, the **sample mode** is the value(s) x_i that appears most times.*

Small example

Assume that the heights of the 5 people in the leadership team of a student chapter are: 60, 67, 72, 63, 60. Then, the sample mode is 60 as it appears twice.

Sample mean

Definition 44 (Sample mean/average) *Given n observations $x_1, x_2, \ldots, x_n$ in a random sample, the **sample mean** or **average** is calculated as*

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Small example

Assume that the heights of the 5 people in the leadership team of a student chapter are: 60, 67, 72, 63, 60. Then, the sample mean is

$$\frac{1}{5}(60 + 67 + 72 + 63 + 60) = 64.4.$$

Sample variance

Definition 45 (Sample variance) *Given n observations $x_1, x_2, \ldots, x_n$ in a random sample, the **sample variance** is calculated as*

$$s^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{n-1} =$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}{n-1}.$$

*The sample standard deviation is denoted by $s = \sqrt{s^2}$. Furthermore, $n-1$ is also called the **degrees of freedom** of the sample.*

Small example

Assume that the heights of the 5 people in the leadership team of a student chapter are: 60, 67, 72, 63, 60 with $\overline{x} = 64.4$. Then, the sample variance is

$$\frac{1}{4}\left(4.4^2 + 2.7^2 + 7.8^2 + 1.4^2 + 4.4^2\right) = \frac{108.81}{4} = 27.2025.$$

Population mean and variance

Percentiles and quartiles

Definition 46 (Percentile) *The number below which we can approximately find p% of the data in the sample is called the p-percentile.*

Based on the definition, we may calculate any p-percentile as follows:

1. Sort the data in increasing order.
2. Calculate $k = (n+1) \cdot \frac{p}{100}$.
3. The element at the k-th position in the sorted data is the p percentile.

Note that the calculation of $(n+1)p/100$ may well be fractional (i.e., the number has us search between two values). When this is the case, then we interpolate. [69]

[69] For example, say we calculate $k = 7.4$. Then the percentile is between the 7th and the 8th value. However, due to the .4 decimal part we would interpolate as: $0.6 \cdot x_7 + 0.4 \cdot x_8$.

Bigger example

Assume the heights of 9 people are 62, 64, 67, 58, 70, 61, 67, 65, 64. What is the 30% and the 67% percentile?

The ordered heights are 58, 61, 62, 64, 64, 65, 67, 67, 70.

30% percentile: Plugging in the formula $\frac{(n+1)p}{100} = \frac{10 \cdot 30}{100} = 3$. The 3rd value is 62.

67% percentile: Plugging in the formula $\frac{(n+1)p}{100} = \frac{10 \cdot 67}{100} = 6.7$. The 6th value is 65 and the 7th is 67: interpolating, we get: $0.3 \cdot 65 + 0.7 \cdot 67 = 66.4$.

A special type of percentiles are the quartiles. They separate the data in four parts, each of which contains 25% of the data. Specifically, we have three quartiles, typically denoted as $Q1, Q2, Q3$:

- Q1: Splits the lower 25% from the rest of the data.
- Q2: Splits the lower 50% from the rest of the data.
- Q3: Splits the lower 75% from the rest of the data.

Q2 is also called the **median**.

Definition 47 (Sample median) *Given n observations $x_1, x_2, \ldots, x_n$ in a random sample, the **sample median** is the value below which (and above which) we find 50% of the observations. It is denoted by $\tilde{x}$ or $Q2$ (the second quartile).*

Bigger example

Earlier, we got the ordered 9 heights to be 58, 61, 62, 64, 64, 65, 67, 67, 70.

Q1: $\frac{(n+1)p}{100} = \frac{10 \cdot 25}{100} = 2.5$. So $Q1 = 61.5$.

Q2: $\frac{(n+1)p}{100} = \frac{10 \cdot 50}{100} = 5 \implies Q2 = \tilde{x} = 64$.

Q3: $\frac{(n+1)p}{100} = \frac{10 \cdot 75}{100} = 7.5 \implies Q3 = 67$.

Ranges and outliers

Definition 48 (Range) *The range of values in a sample or population is calculated as the difference of the maximum and the minimum value in the sample or population:* $R = \max\{x_i\} - \min\{x_i\}$.

By definition, the range of a population will always be greater than or equal to the range of a sample.

Definition 49 (Interquartile range) *The interquartile range is calculated as the difference of the third to the first quartile:* $IQR = Q3 - Q1$.

The IQR is in essence a measure of range but focusing on the middle part of the data considered.

Definition 50 (Outliers) *An outlier is a value that affects the range of our data but leaves the IQR unaffected. Specifically, we say that a data point is an outlier if it lies outside* $[Q1 - 1.5IQR, Q3 + 1.5IQR]$.

Describing aluminum-lithium specimens

A company has collected the following data for compressive strength (psi) of aluminum-lithium specimens: 105, 221, 183, 186, 121, 181, 180, 143, 97, 154, 153, 174, 120, 168, 167, 141, 245, 228, 174, 199, 181, 158, 176, 110, 163, 131, 154, 115, 160, 208, 158, 133, 207, 180, 190, 193, 194, 133, 156, 123, 134, 178, 76, 167, 184, 135, 229, 146, 218, 157, 101, 171, 165, 172, 158, 169, 199, 151, 142, 163, 145, 171, 148, 158, 160, 175, 149, 87, 160, 237, 150, 135, 196, 201, 200, 176, 150, 170, 118, 149.

What are the outliers?

We would first have to sort the data in increasing order, and then calculate $Q1, Q3$. Doing so gives us $Q3 = 181$, $Q1 = 144.5$ and $IQR = 36.5$. We may also calculate the minimum and maximum values as 76 and 245, respectively, leading to a range of 169.

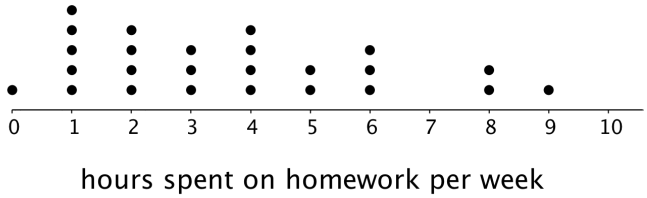
Potential outliers would lie outside the range of $[Q1 - 1.5IQR, Q3 + 1.5IQR] = [89.75, 235.75]$. The only values satisfying this are: 245, 76, 87, 237.

Graphical devices of data

In this subsection, we discuss some visual tools to represent data.

Dot diagrams Dot diagrams (as the name suggests) asks to place a dot on top of each data point. The mode and median are revealed pretty easily in a dot diagram: simply find the tallest set of dots for the mode, and the value below which 50% of the dots lie for the median. See Figure 41 for an example.

Figure 41: An example of a dot diagram representing the amount of time each student spent on Homework assignment 1 (self-reported) during Fall 2019, rounded to the closest integer.



In the example, the mode was 1 hour, and the median is at 13 "dots" (for a total of 25 dots)[70] and is found at 3 hours.

It becomes clear from the way this is constructed that the dot diagram is only useful for smaller sized datasets.

[70] Recall the median calculation is $(n+1)\,50/100 = 13$

Stem-and-leaf plots A stem-and-leaf diagram only makes sense when all of the data consists of at least two digits. It is a striking visual tool to showcase frequency. The way it is constructed is simple: we pick a series of stems (the first, more important digits) and leaves (the least important digit). For example, the number 311, could be represented as a stem of 31 and a leaf of 1. The leaves are sorted in increasing order. An example is presented in Figure 42.

Alongside the diagram, we typically present frequency (the cumulative number of observations up to and including a stem). In the example, we see that up to and including the stem of 10 we have five observations; on the other hand up to and including the stem of 18 we have 64 observations. This can be used to measure individual frequency: for example the stem 17 has 10 observations – we can tell because up to and including 17 we have 57 observations, whereas up to and including 16 we have 47 observations.

Scatter plots Scatter diagrams are particularly useful when we suspect that the data has some hidden relationship, either positive or negative. For example, what can you say about the following scatter plot of Figure 43 showing data points of activity and obesity in the US?

Scatter plots may reveal a positive or negative relationship. They may also show that there seems to be no relationship between two variables. We reveal small, simple examples of each of the three cases

Figure 42: An example of a stem-and-leaf diagram representing the data from the aluminum-lithium specimens.

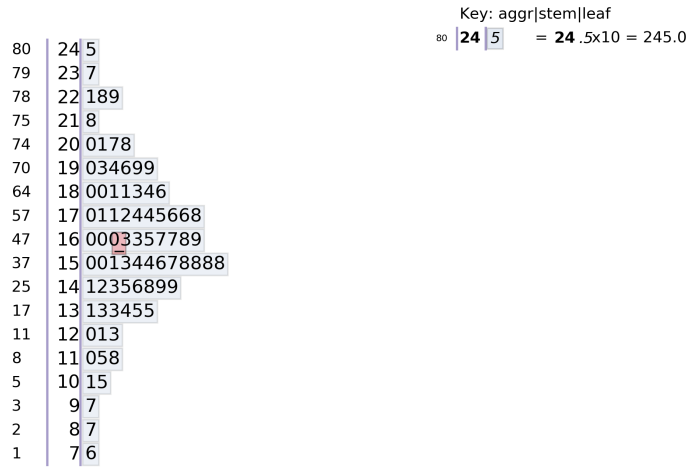
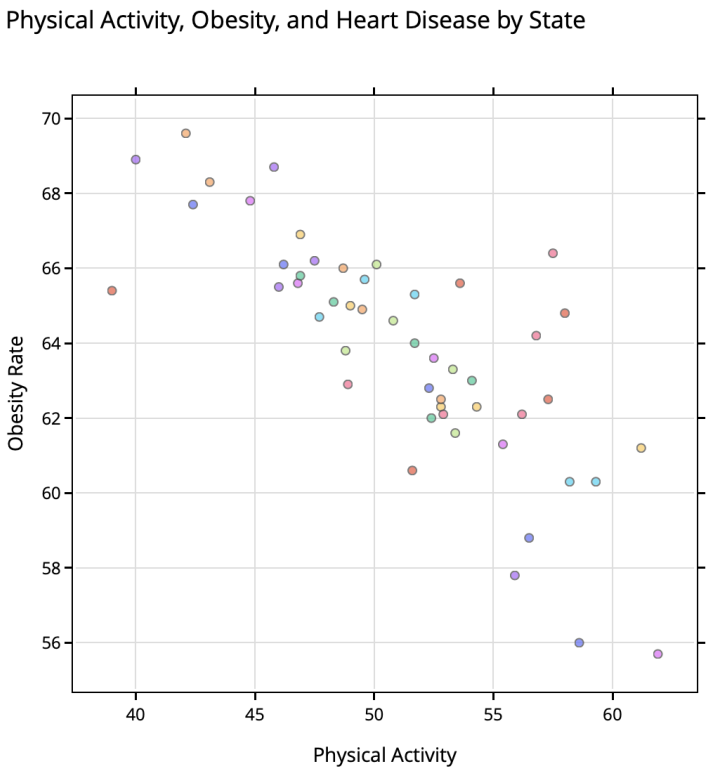
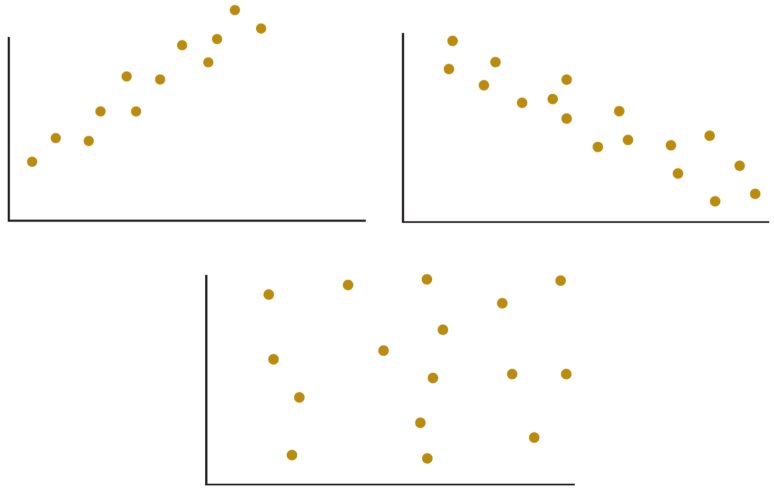


Figure 43: A scatter plot of the relationship between the rate of obesity cases and the average physical activity levels at each state.



in Figure 44.

Figure 44: Positive relationship (left), negative relationship (right), and no discernible relationship (below).



Time series plots A time series plot is useful when the data are recorded in the order of time. For example, if we are given data that presents some number that changes every month, then it may be suitable to present in a plot where the x axis represents time, and the y axis the number of interest. Below we present two examples: from the city of Chicago for the number of reported crimes per month in Figure 45 (notice the huge drop every February, due to the fact that February has fewer days!) and from Champaign county on the number of COVID-19 cases every week in Figure 46.

Figure 45: Reported crimes in Chicago by date. Data obtained by `https://data.cityofchicago.org` on October 8, 2019.

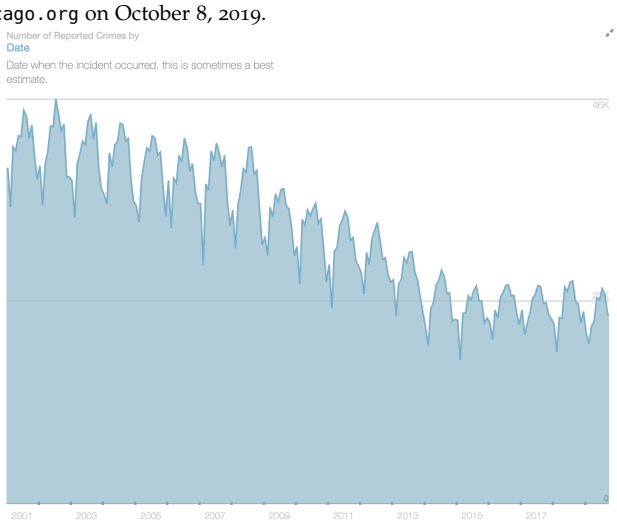
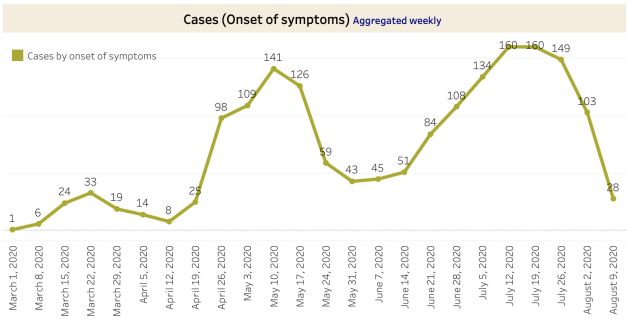


Figure 46: Number of cases (onset of symptoms) per week in Champaign county. Data obtained by `http:c-uphd.org` on August 14, 2020.



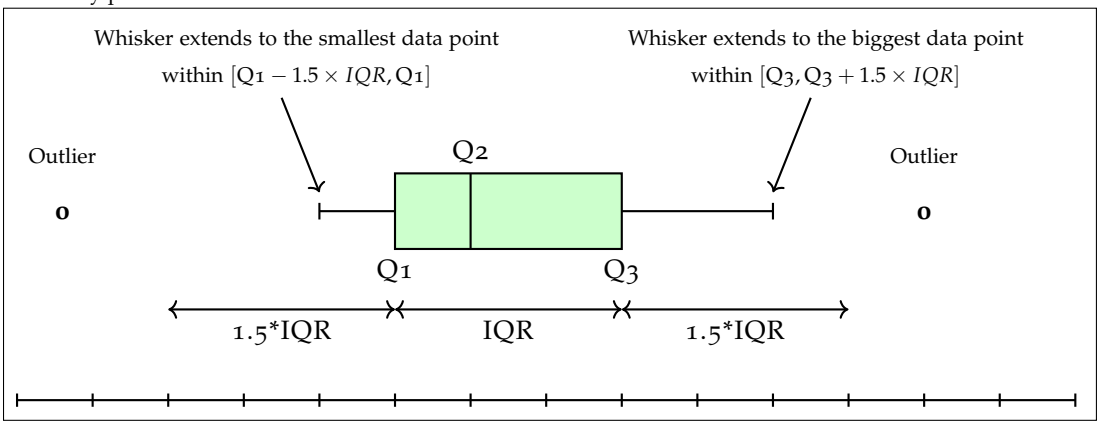
Box plots Box plots, sometimes also called box-and-whisker plots, are graphical devices built to reveal multiple interesting properties at once. Seeing a box plot reveals:

1. the center of the data;
2. the spread of the data;
3. the shape of the data;
4. and the outliers in the data.

Seeing a box plot immediately shows the *min* value, the first quartile $Q1$, the median $Q2$, the third quartile $Q3$, the interquartile range IQR, and the *max* value of the data.

Figure 47 shows all the inner workings of constructing a box plot.

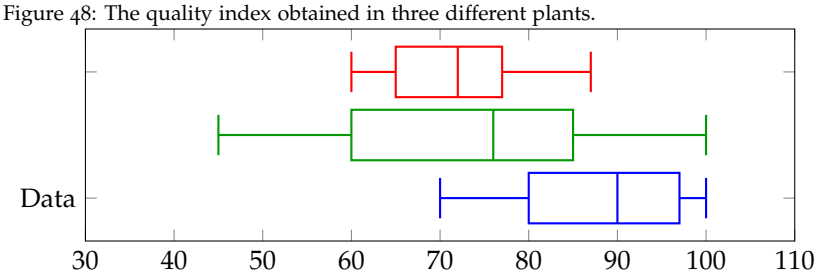
Figure 47: A box plot. To construct it, we create a rectangle ranging from Q1 to Q3. We separate it into two parts drawing a line where the median Q2 is. Then, we extend two whiskers on the two sides all the way to the smallest and biggest value respectively so long as that value is less than $1.5 \times IQR$ away from the quartile. Finally, we note every point outside the whiskers as outliers with a "o".



It is useful to compare box plots one next to the other. For example, see the box plot of Figure 48 containing information about the

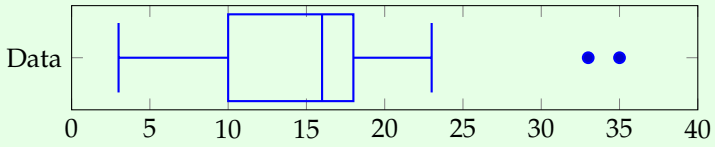
quality obtained in three different plants. We observe that the blue plant provides us with the highest quality index. The green one has better median quality index than the red one, but the red one has a narrower range of possible quality indices, making it more consistent.

Figure 48: The quality index obtained in three different plants.



A small example

We are given a set of data points, and we have calculated that $Q1 = 10, Q2 = 16, Q3 = 18$. The points outside the $[Q1, Q3]$ range are 3, 7, 8, 8, 9 from below and 19, 23, 33, and 35 from above. Draw the boxplot.



Histograms A histogram is a graphical construct that presents data by placing them in *bins*.

The bins could be numbers (in the case of Figure 49, the number of friends on fb.

The bins could represent age ranges (in the case of Figure 50, the age of Florida residents).

The bins could even represent letter grades (as you are probably used to seeing letter grade distributions after exams as in Figure 51!).

Histograms possess three important characteristics:

1. modality.
2. heavy/light tailedness.
3. skewness.

The **modality of a histogram** is concerned with the number of "noticeable peaks" in the data. Recall that a single peak would imply a single mode (most frequent value, or in a histogram's case most frequent range of values). A histogram can then be:

Figure 49: The number of friends that a person has on Facebook.

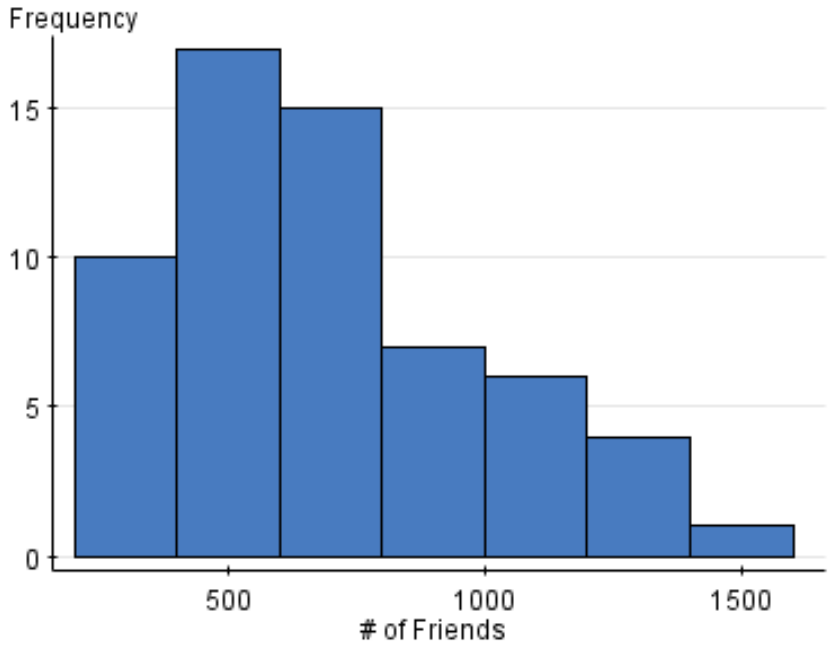


Figure 50: The age of Florida residents in 2018.

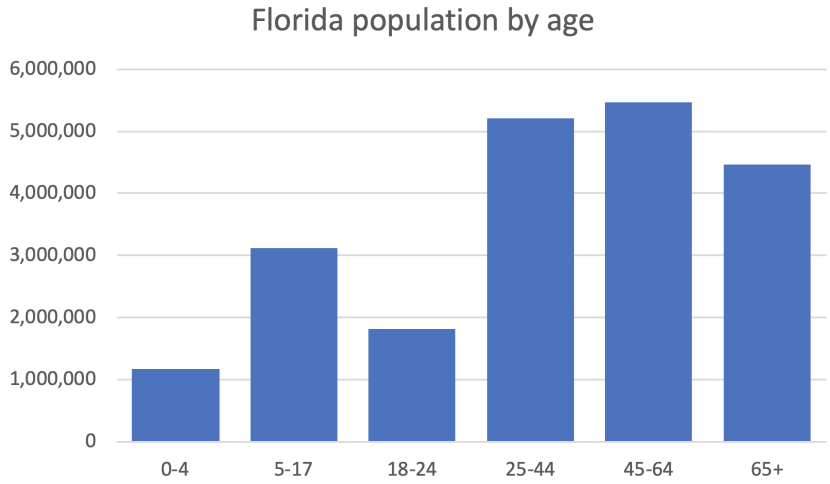
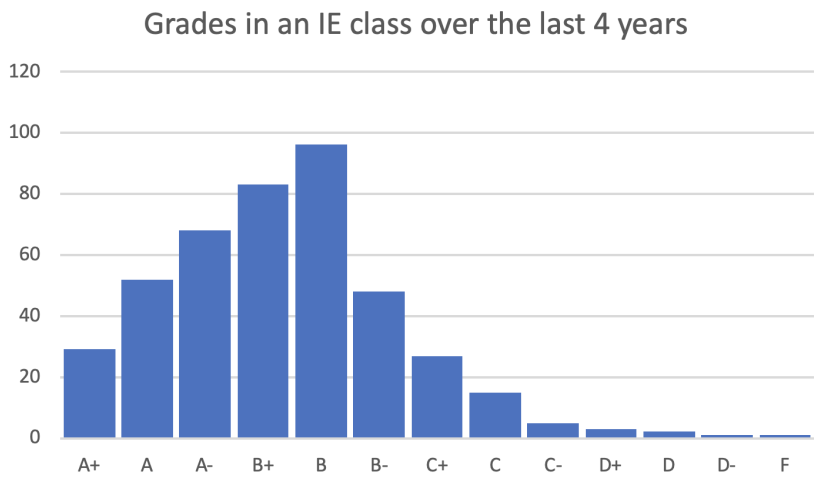


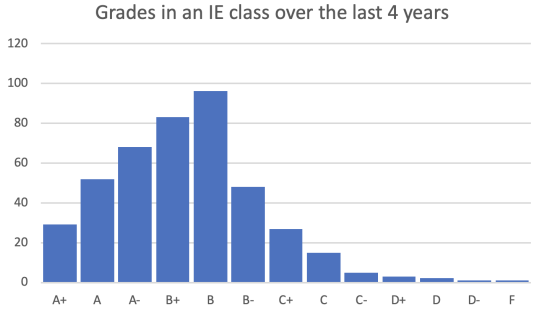
Figure 51: The final grade distribution in an IE course over the last four years.



- unimodal (single mode).
- bimodal (two modes).
- multimodal (multiple modes).
- uniform (no mode).

We present four examples to showcase each of the four types in Figures 52–55.

Figure 52: **Unimodal**: we note one observable peak at the "B" letter grade.



A second histogram characteristic is whether it possesses a **heavy or light tail**. We say that a tail is "heavy" if it is "heavier" than the exponential distribution.

Let's turn our focus back to histograms. Here are two examples of how a heavy-tailed a light-tailed histogram would look like:

Figure 53: **Bimodal**: we note one observable peak at the "5-17" and the "45-65" age ranges.

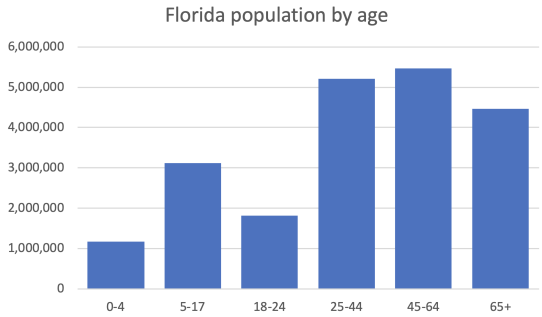


Figure 54: **Multimodal**: we can find four noticeable peaks here when observing the height of NBA players (in inches, all heights from the 2013 league).

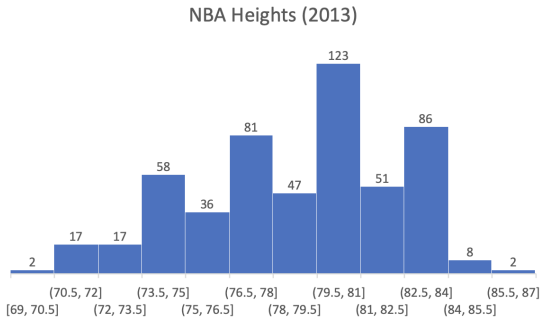


Figure 55: **Uniform**: when we roll multiple dice and report the outcomes.

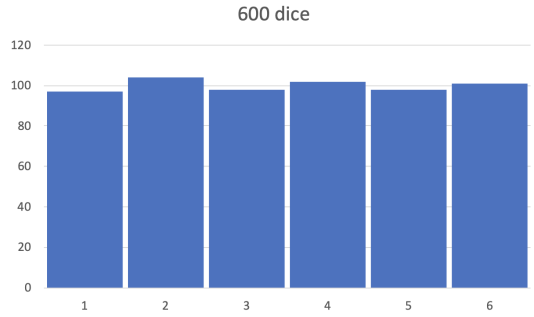
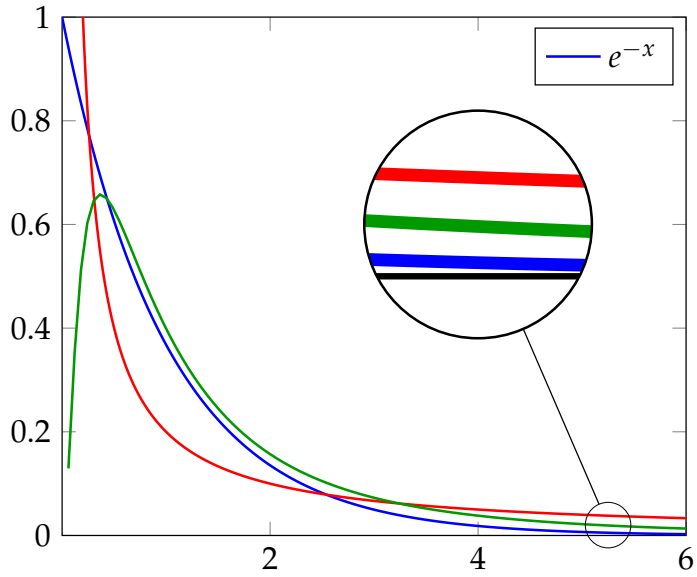
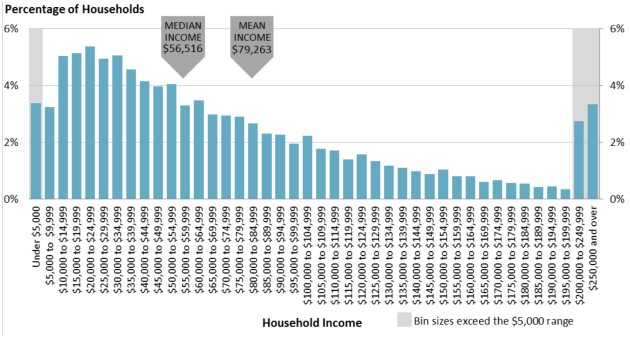


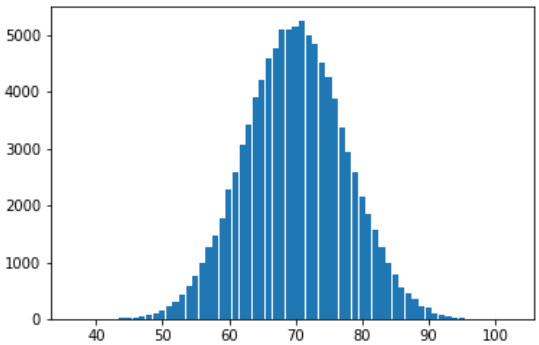
Figure 56: What heavier than an exponential distribution means. Here both the green and the red functions are located higher than the exponential for bigger values of x, so they would both be characterized as heavy-tailed.



a) **Heavy-tailed**: the household income. Note how there is a heavy tail for some very high incomes.



b) **Light-tailed**: heights of a sample of the population in Denmark (restricted to people who self-identify as male).



Why should we care about this characteristic? Well, say we are devising a policy and we need to figure out whether the same policy should apply to all. When the feature we are looking at has a heavy tail, this might have us thinking twice before having the same policy, because many observations would lie far from the average or the bulk of our observations.

Finally, we discuss **skewness**, the third histogram characteristic. In essence, we want to answer the question of whether the histogram is symmetric or not. And, if not, is it right-skewed? Or left-skewed? How do we figure this out?

- If the tail is to the left, then it is left-skewed or has negative skewness.
- If the tail is to the right, then it is right-skewed or has positive skewness.
- Otherwise, it is symmetric.

Figure 57: An example of a left-skewed histogram, as the tail is to the left.

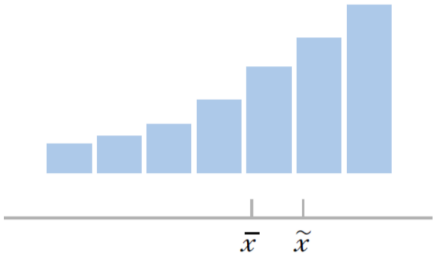
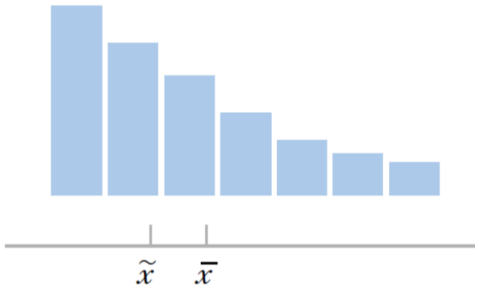


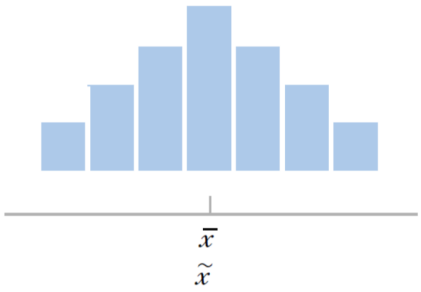
Figure 58: An example of a right-skewed histogram, as the tail is to the right.



In all figures, $\bar{x}$ represents the average, and $\tilde{x}$ the median. Note that this helps us provide another definition for skewness. If the median is:

- to the right of the average, then the histogram is left-skewed.

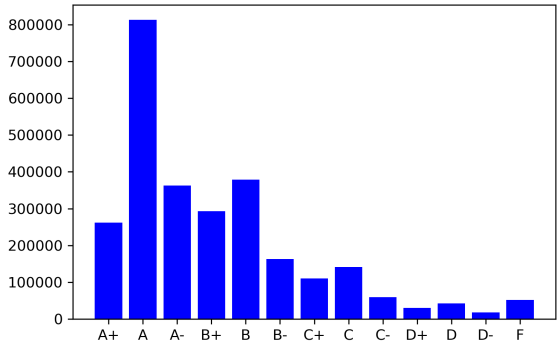
Figure 59: An example of a symmetric histogram with no discernible tail.



- to the left of the average, then the histogram is right-skewed.
- in a similar location to the average, the the histogram is symmetric.

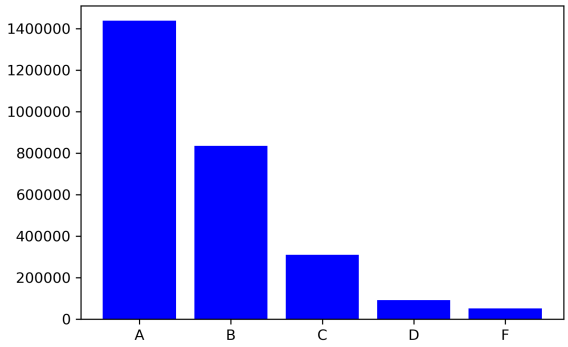
Finally, we address another important aspect. It is true that the same data can be presented in many, many different ways using histograms:

- Here we show every distinct letter grade that a student may



receive:

- Whereas here we show only the main letter grades (for example, no “B+” or “B-”, instead we only have a “B” grade):



The more bins we introduce, the less width each bin has, and the more the shape resembles the actual distribution of the data (for larger amounts of data).

Q-Q plots Q stands for **quantile**. A Q-Q plot is useful when *comparing* two probability distributions or two samples. It is more "powerful" (as in easier to interpret) than comparing two histograms. It may also be used for "goodness of fit" to check whether our data follows a specific distribution. The most well-known Q-Q plot is the normal Q-Q plot that helps verify whether our data follows a normal distribution or not.

Before we introduce how Q-Q plots are built and read, we need to discuss quantile functions. What are quantile functions? As we saw earlier in the lecture, for any sample we have:

- p percentile: p% of the observations are below that value.
- Q1: first quartile, $p = 25$.
- Q2: second quartile, $p = 50$, also known as the median.
- Q3: third quartile, $p = 75$.

Now, for any random variable X with CDF $F(x)$, we define the **quantile function** as:

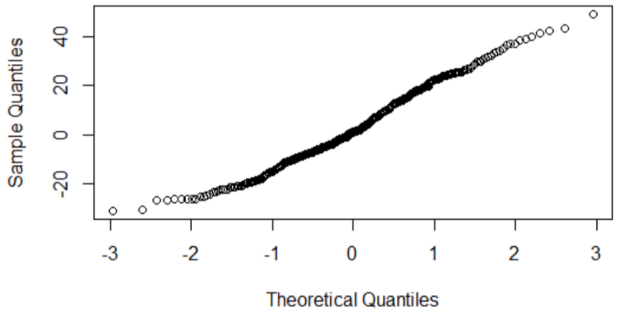
$$Q(p) = \inf\{x : F(x) \leq p\}, \quad \text{for } 0 \leq p \leq 1.$$

In English: look for the smallest value of x such that $F(x)$ is smaller than or equal to the given probability p. An interesting note: $Q(p) = F^{-1}(x)$.

So how do we build a Q-Q plot? We have a sample of n observations, and we have a theoretical distribution we believe our sample follows (e.g., exponential, normal, etc.). With this information at hand, we follow the procedure:

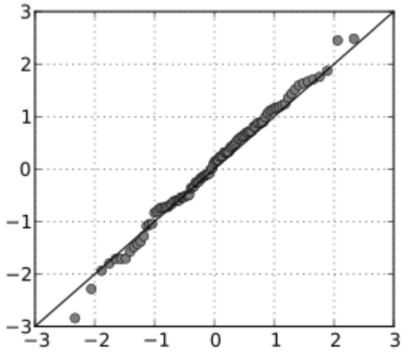
1. First, identify some quantile levels of interest: $0 < p_1 < p_2 < \ldots < p_n$.
   - Typically, we choose $p_i = \frac{i}{n+1}$, for n observations.
   - We could also choose $p_i = \frac{i-0.5}{n}$.
2. Then, we compute the *sample*'s quantiles. Let them be $X_1, X_2, \ldots, X_n$.
3. Now, we compute the *theoretical* quantiles, based on the $F(x)$ selected. Let them be $F^{-1}(p_1)$, $F^{-1}(p_2)$, ..., $F^{-1}(p_n)$.
4. Finally, plot the sample quantiles against the theoretical quantiles in the same plot. See Figure 60 for an example.

Figure 60: An example of a Q-Q plot. Here the x axis shows the theoretical quantiles and the y axis the sample quantiles.



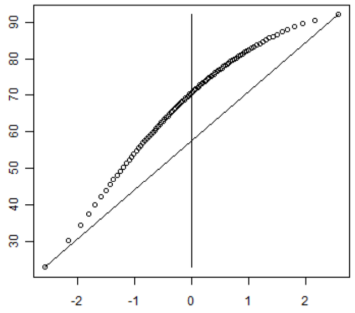
After we have a Q-Q plot, how do we use it? How do we read it? Well, the nice thing is that if indeed the sample follows that (theorized) distribution, then the Q-Q plot will look like a straight (45°) line, as in Figure 63!

Figure 61: Here we assumed the sample follows a normal distribution so we plotted the sample's quantiles to the normal quantiles. We get a straight line, meaning our data comes indeed from the normal distribution!

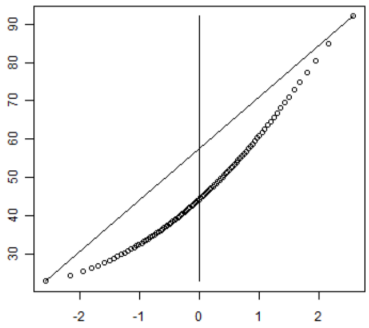


We can also tell whether the distribution is left or right skewed.

a) Left-skewed:

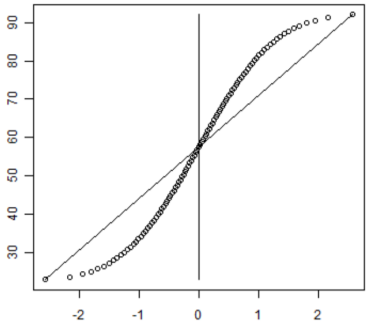


b) Right-skewed:

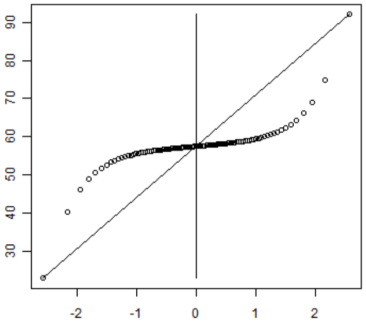


Finally, we can tell if it light- or heavy-tailed.

a) Light-tailed:



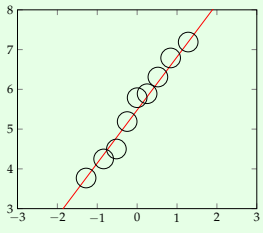
b) Heavy-tailed:



Constructing a Q-Q plot

Assume we collected the following observations from some population: $3.77, 4.25, 4.50, 5.19, 5.89, 5.79, 6.31, 6.79, 7.19$. Do the observations seem to come from a normal distribution? Let us construct a Q-Q plot to prove or disprove this.

We have $n = 9$ observations, so we can get $p_1 = 10\%, p_2 = 20\%, \ldots, p_9 = 90\%$. Find the z-values corresponding to the 9 percentage: -1.28, -0.84, -0.52, -0.25, 0, 0.25, 0.52, 0.84, 1.28. Finally, we plot them and see if they appear to form a 45° line or not.



Point estimators

Learning objectives

After these lectures, we will be able to:

- Describe point estimation.
- Explain the difference between bias and variance of a point estimator.
- Evaluate the bias, variance, mean square error of a point estimator.
- Compare two point estimators and pick the better one.

Motivation: Inferring parameters

In most applications, we have a good enough idea on the distribution that we need to follow. When a company makes vehicles, we could pick some of them to check for their quality and count how many are of high quality: this can be modeled as a binomial or a hypergeometric distribution. When a student takes an exam, they will expect to do similarly to their previous exams plus or minus some points if they prepare in a similar manner: their score can be modeled as a normal distribution.

However, one of the questions we need to answer is: what are the parameters of the distributions? What is the mean and variance of that normal distribution? What is p in a binomial distribution? So far, we have been given this as part of our data. What happens when we are given data and need to infer their values, based on real-life observations?

Motivation: Predicting an election

Before an election takes place, we see many polls. Some of them appear to better resemble the final result (after the election); others fail to capture reality. Given this data based on a *sample* of the whole *population*, what can we say about the election? What can we say about the probabilities of one candidate versus another?

Statistical inference

Statistical inference takes us from the sample to the whole. For example, consider any of the next scenarios:

- We interviewed 50 people about the next election. What do the results imply for the general election?

- We picked a sample of 10 cars and performed a crash test. What do the observations imply for the whole production line?
- We collected exit interview data from 100 alumni. What do their answers imply for the starting salary of our alumni?

Right away, we can make some intuitive observations:

1. Checking a sample, rather than the whole, saves us time and effort.
2. Checking a sample, rather than the whole, comes with a loss of information.
3. Checking a sample, rather than the whole, we want to recreate the whole.

Statistical inference theme

The general theme for this part of the class can be summarized pretty well in the following Figure 62.

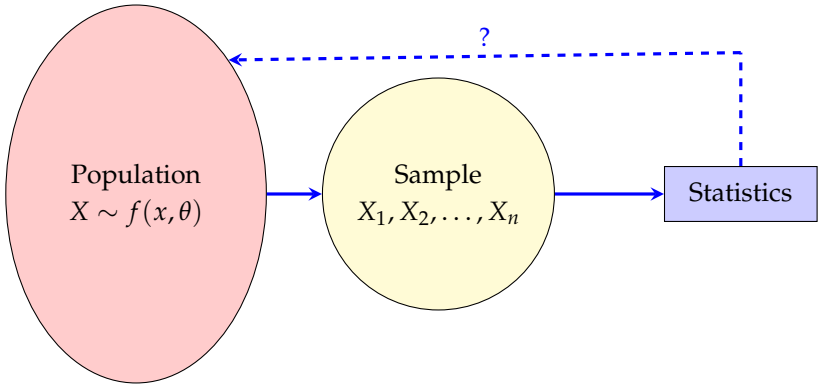


Figure 62: The theme of statistical inference. We collect a **sample** of the whole **population** that we analyze to get some **statistic**, which we then use to infer information about the population.

Sample averages and variances

Assume a large population X: you decide to collect only 5 random variables X_1, X_2, X_3, X_4, X_5. Then, we may calculate the sample average $\frac{X_1+X_2+X_3+X_4+X_5}{5}$ and the sample variance and use those in lieu of the population mean (unknown) and the population variance (unknown).

For example, say I want to figure out the average height of every Chicago resident, I could (i) travel to Chicago, (ii) ask 10 people about their height, and (iii) calculate the average of these 10 people. Is this the true average?

Statistics

We proceed with some preliminaries that will be used throughout the next few classes.

Definition 51 (Statistic) *A statistic is any value obtained by random data.*

Well, this is not very useful. The only recurring theme here is *random* data. In essence, the definition claims that any value that is different for differently obtained data can be considered a statistic!

Height statistics

Assume that the heights of the 5 people in the leadership team of a student chapter are: 60, 67, 72, 63, 60. Then, the height of the second person picked (67) is a statistic. The average height is another statistic. Finally, getting the height of the first person and multiplying it by 3 and adding to it the height of the last person is *also* a statistic.

Clearly, some statistics are more useful than others. For example, in our earlier discussion the average is more useful than the last statistic. To verify, answer the following questions.

Are the following statistics? True or False.

- The sample variance.
  a) True b) False
- The value of the first element of a sample.
  a) True b) False
- The value of the first element of a sample minus 3.
  a) True b) False

Some basic properties of statistics:

1. **Statistics depend on the sample selected**: the value a statistic gets will be different depending on the sample selected. If I select 5 students in the class and report their average exam score (a statistic), it will be different depending on the 5 students selected.
2. **Statistics are functions of the sample selected**: we can use the same "formula" or follow the same "approach" to estimate the value of a statistic given a sample, no matter the sample.

3. **Statistics are random variables**: statistics are distributed as random variables. Certain values may appear more often than others, we can define expectations for statistics, etc.

Polling with few people

For a poll, we are asked to find 10 people and ask them a question on a Likert scale (that is from 1 to 5). Then, we take their answers and add them up: we say that if the score is ≥ 30 then the people agree with that statement on average.

Unfortunately you were only able to find 5 people, who provided the following answers X_i: 2, 1, 3, 3, 3. You then decide to use $Y = 2 \cdot \sum X_i$ as the statistic you report back. In this case, you'd report $Y = 2 \cdot (2+1+3+3+3) = 24$.

- **Is this a statistic?** Yes.
- **Does it depend on the sample selected?** Yes. Change the sample asked to obtain a different number.
- **Is it a function of the sample selected?** Yes. We always add up the answers and multiply by 2.
- **Is this a random variable?** Yes. We can calculate an expectation and a variance, and we can estimate probabilities!

Sampling distribution

So, if a statistic is a random variable, what is its distribution? The distribution of a statistic, called the **sampling distribution** depends on three things:

1. The distribution of the whole population. Of course we should expect that the distribution of the population will be reflected when looking at a sample!
2. The size of the sample. Once again, it should make sense that the bigger the sample we pick the more accurately we will reflect the population distribution.
3. The way the sample was selected. We will not devote a lot of time in this: but, picking a sample in a non-random way will affect the distribution we see.

Confused? Don't be! We have done that already..

Back to the normal distribution

Assume you have a population where each individual is distributed following a normal distribution with mean μ and variance σ^2. You pick, at random, a set of n individuals X_i. What is the average distributed as?

We have seen that the average $Y = \frac{\sum X_i}{n}$ is also normally distributed with the same mean μ and variance $\frac{\sigma^2}{n}$. This normal distribution $\mathcal{N}(\mu, \sigma^2/n)$ is the sampling distribution.

Point estimators

Let's put everything formally. Let X be a population distributed with some pdf $f(x, \theta)$, where θ is some unknown parameter. By the way, you may treat θ as a vector of multiple parameters. [71]

Furthermore, let $X_1, X_2, \ldots, X_n$ be a series of random elements picked from the population. They form the sample of size n that was selected. By definition, seeing as $X_1, X_2, \ldots, X_n$ all come from the same place, they are identically distributed and independent random variables.

[71] Recall that every distribution we have seen had some parameters that were required to define it. For example, the binomial distribution needed $n > 0$ and $p \geq 0$, whereas the exponential distribution or the Poisson distribution needed $\lambda > 0$.

Definition 52 (Point estimators) *We define point estimator(s) $\hat{\Theta}$ as a statistic that is used to approximate the unknown parameter(s) θ.*

By definition, $\hat{\Theta}$ is a function of the sample selected $(X_1, X_2, \ldots, X_n)$, hence we may say that $\hat{\Theta}$ is a random variable depending on the sample ($\hat{\Theta} = h(X_1, X_2, \ldots, X_n)$, where $h(\cdot)$ is some function).

Of course, once we have picked a sample then $\hat{\Theta}$ can be calculated and assigned a value. This value is called the **point estimate** $\hat{\theta}$. To summarize, $\hat{\Theta}$ is the general statistic used (e.g., the point estimator can be found if we take the average and add 2) whereas $\hat{\theta}$ is the value the estimator receives for a specific sample (e.g., for our sample, the average is 7 so the point estimate is 9). To summarize, $\hat{\Theta}$ is typically a "formula" or an "expression", whereas $\hat{\theta}$ is an actual number.

Common point estimators

Such a topic (statistical inference) is so broad and useful that we definitely already have some estimators that are typically used. Are you looking for the (unknown) mean of a population? Collect a sample and report its average. Are you looking for the (unknown) population variance? Collect a sample and report its sample variance. We differentiate between single and two populations.

Single population. For a single population:

Parameters	Point estimators
Population mean μ	Sample average $\hat{\Theta} = \overline{x}$
Population variance σ^2	Sample variance $\hat{\Theta} = s^2$
Population proportion p	Sample proportion $\frac{\hat{n}}{n}$

Single population proportions

Assume a population that we want to ask whether they agree or disagree with a new policy. Should we enact it? If it is difficult or impossible to collect feedback from all, we may pick a sample and ask them if they agree or not. Let n be the sample size and $\hat{n}$ be the number of people who agree.

Finally, we may report that $\frac{\hat{n}}{n}$ is the point estimator for the unknown proportion.

Two populations. For two populations:

Parameters	Point estimators
Difference in population means $\mu_1 - \mu_2$	Difference in sample averages $\hat{\Theta} = \overline{x}_1 - \overline{x}_2$
Ratio in population variances $\frac{\sigma_1^2}{\sigma_2^2}$	Ratio in sample variance $\frac{s_1^2}{s_2^2}$
Difference in population proportions $p_1 - p_2$	Difference in sample proportions $\hat{\Theta} = \frac{\hat{n}_1}{n_1} - \frac{\hat{n}_2}{n_2}$

Two population proportions

Assume a population that we want to ask whether they agree or disagree with a new policy. However, we are also aware of the existence of two populations: say, for example, people who make more that \$100,000 and people who make less. Is the policy more preferred to people of one category versus the other? Let n_1 be the sample size from the first population and $\hat{n}_1$ be the number of people who agree from that population. Similarly, define n_2 and $\hat{n}_2$.

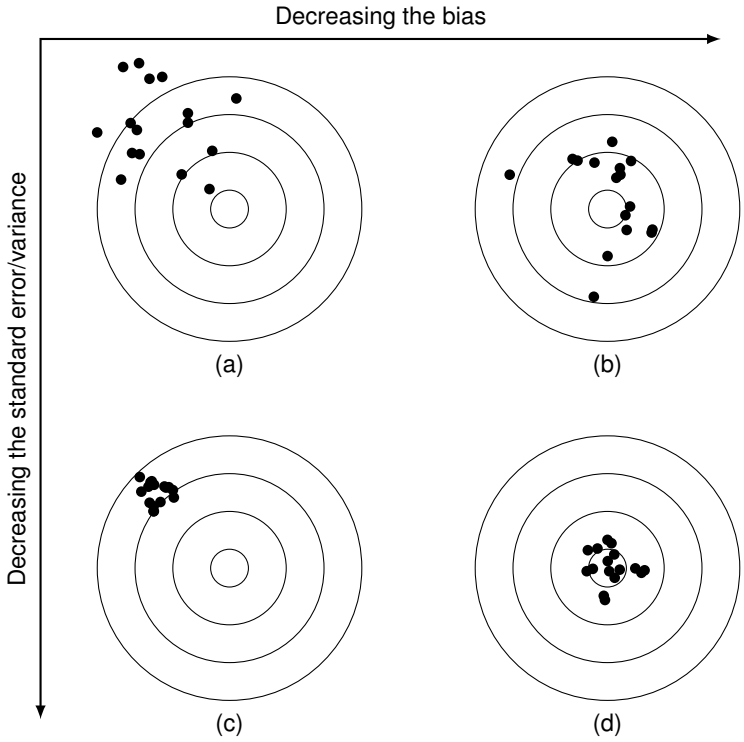
Finally, we may report that the difference between the two proportions is $\frac{\hat{n}_1}{n_1} - \frac{\hat{n}_2}{n_2}$ as the point estimator for the unknown proportion difference.

Before moving to the next subsection, take a moment to summarize what we have seen. We have a population distributed with some

probability density function ($f(x)$) but with unknown parameters θ. We then come up with a plan: select a sample, and calculate a point estimator $\hat{\Theta}$. For a given sample, you obtain a point estimate (actual value) $\hat{\theta}$. You use that to estimate the unknown parameter. Additionally, recall that $\hat{\Theta}$ is a random variable, so it should come as no surprise that we can analyze it as such!

What makes a good estimator?

Every estimator has two main items we want to evaluate it by: **accuracy** and **precision**. In statistics terms, we refer to them as **bias** and **standard error** or **variance**. We present the effect that each of the two would have to our estimation process: decreasing the bias would lead to better results *on average*; decreasing the standard error/variance would lead to smaller dispersion.



A "good" estimator should have zero bias and zero variance! However, this is practically impossible: hence, we settle for zero bias and minimum variance. Let us proceed with the definitions.

Definition 53 (Bias) *We define the **bias** of a point estimator as the difference between its expectation and the parameter itself.*

$$bias\left[\hat{\Theta}\right] = E\left[\hat{\Theta}\right] - \theta.$$

An estimator with zero bias is referred to as unbiased.

Bias example

Assume a population with mean μ and variance σ^2. As the mean is unknown you decide to use the following three approaches to estimate it:

1. Get the average from a sample of 3 randomly picked observations.
2. Get a sample of 3 randomly picked observations and calculate $\frac{2 \cdot X_1 + X_2 - X_3}{2}$.
3. Get a sample of 3 randomly picked observations and calculate $2X_1 + X_2 - X_3$.

What are the biases of each of the three point estimators?

1. $\hat{\Theta}_1 = \frac{X_1 + X_2 + X_3}{3}$:

$$E\left[\hat{\Theta}_1\right] = E\left[\frac{X_1 + X_2 + X_3}{3}\right] = \frac{1}{3}\left(\underbrace{E\left[X_1\right]}_{\mu} + \underbrace{E\left[X_2\right]}_{\mu} + \underbrace{E\left[X_3\right]}_{\mu}\right) =$$
$$= \frac{1}{3}\left(\mu + \mu + \mu\right) = \mu \implies$$
$$\implies bias(\hat{\Theta}_1) = 0.$$

2. $\hat{\Theta}_2 = \frac{2 \cdot X_1 + X_2 - X_3}{2}$:

$$E\left[\hat{\Theta}_2\right] = E\left[\frac{2 \cdot X_1 + X_2 - X_3}{2}\right] = \frac{2\mu + \mu - \mu}{2} = \mu \implies$$
$$\implies bias\left(\hat{\Theta}_2\right) = 0.$$

3. $\hat{\Theta}_3 = 2 \cdot X_1 + X_2 - X_3$:

$$E\left[\hat{\Theta}\right] = E\left[2 \cdot X_1 + X_2 - X_3\right] = 2\mu + \mu - \mu = 2\mu \implies$$
$$\implies bias\left(\hat{\Theta}_3\right) = \mu.$$

So, the first two estimators will be unbiased (zero bias)! The last one is biased and its bias is as big as the unknown mean.

Definition 54 (Standard error and variance) *We define the **standard error** of a point estimator as the square root of its **variance**.*

$$SE\left[\hat{\Theta}\right] = \sqrt{Var\left[\hat{\Theta}\right]}.$$

We want this to be minimum. A point estimator with minimum variance and zero bias is called a minimum variance unbiased estimator.

Variances example

Assume the same population with unknown mean μ and variance σ^2. We use again the three estimators from before (referred to as $\hat{\Theta}_1, \hat{\Theta}_2, \hat{\Theta}_3$). What are the variances of each of the three point estimators?

1. $\hat{\Theta}_1 = \frac{X_1+X_2+X_3}{3}$:

$$\begin{aligned} Var\left[\hat{\Theta}_1\right] &= Var\left[\frac{X_1+X_2+X_3}{3}\right] = \\ &= \frac{1}{9}\left(\underbrace{Var\left[X_1\right]}_{\sigma^2} + \underbrace{Var\left[X_2\right]}_{\sigma^2} + \underbrace{Var\left[X_3\right]}_{\sigma^2}\right) = \\ &= \frac{1}{9}3\sigma^2 = \frac{\sigma^2}{3}. \end{aligned}$$

2. $\hat{\Theta}_2 = \frac{2\cdot X_1+X_2-X_3}{2}$:

$$\begin{aligned} Var\left[\hat{\Theta}_2\right] &= Var\left[\frac{2\cdot X_1+X_2-X_3}{2}\right] = \\ &= Var\left[X_1\right] + \frac{1}{4}Var\left[X_2\right] + \frac{1}{4}Var\left[X_3\right] = \\ &= \sigma^2 + \frac{1}{4}\sigma^2 + \frac{1}{4}\sigma^2 \implies Var\left[\hat{\Theta}_2\right] = \frac{3}{2}\sigma^2. \end{aligned}$$

3. $\hat{\Theta}_3 = 2\cdot X_1 + X_2 - X_3$:

$$\begin{aligned} Var\left[\hat{\Theta}_3\right] &= Var\left[2\cdot X_1 + X_2 - X_3\right] = \\ &= 4\sigma^2 + \sigma^2 + \sigma^2 \implies Var\left[\hat{\Theta}_3\right] = 6\sigma^2. \end{aligned}$$

Comparing, the first estimator has a significantly smaller variance than the other two. Among the three options, $\hat{\Theta}_1$ is the minimum variance unbiased estimator.

Definition 55 (Mean square error) *We define the **mean square error** of a point estimator as the expected value of the square error $\left(\hat{\Theta}-\theta\right)^2$:*

$$MSE = E\left[\left(\hat{\Theta}-\theta\right)^2\right].$$

We can use this to derive the fact that the mean square error is equal to

the summation of the variance plus the square of the bias:

$$\begin{aligned} MSE(\hat{\Theta}) &= E\left[\left(\hat{\Theta}-\theta\right)^2\right] = \\ &= E\left[\hat{\Theta}-E\left[\hat{\Theta}\right]\right]^2 + \left(\theta - E\left[\hat{\Theta}\right]\right)^2 = \\ &= Var\left[\hat{\Theta}\right] + bias(\hat{\Theta})^2. \end{aligned}$$

By definition, the MSE tries to capture both bias and variance at the same time. Hence, we typically say that one estimator is better than another if its MSE is smaller. We may also define the **relative efficiency** as the ratio of two estimator mean square errors:

$$\text{Relative efficiency} = \frac{MSE\left(\hat{\Theta}_1\right)}{MSE\left(\hat{\Theta}_2\right)}.$$

If the relative efficiency is less than 1, then we say that point estimator $\hat{\Theta}_1$ is preferred to point estimator $\hat{\Theta}_2$.

Mean square errors example

Assume the same population with unknown mean μ and variance σ^2. We use for one last time the three estimators $\hat{\Theta}_1, \hat{\Theta}_2$, and $\hat{\Theta}_3$. What are the mean square errors of each of the three point estimators? Which one would we prefer? What are the relative efficiencies of $\hat{\Theta}_1, \hat{\Theta}_2$, $\hat{\Theta}_1, \hat{\Theta}_3$, $\hat{\Theta}_2, \hat{\Theta}_3$?

1. $\hat{\Theta}_1 = \frac{X_1+X_2+X_3}{3}$: $MSE(\hat{\Theta}_1) = \frac{\sigma^2}{3} + 0 = \frac{\sigma^2}{3}$.
2. $\hat{\Theta}_2 = \frac{2\cdot X_1+X_2-X_3}{2}$: $MSE(\hat{\Theta}_2) = \frac{3\sigma^2}{2} + 0 = \frac{3\sigma^2}{2}$.
3. $\hat{\Theta}_3 = 2\cdot X_1 + X_2 - X_3$: $MSE(\hat{\Theta}_3) = 6\sigma^2 + \mu^2$.

$\hat{\Theta}_1$ has the smallest MSE (as expected), followed by $\hat{\Theta}_2$. The relative efficiencies can be found as:

1. $\hat{\Theta}_1, \hat{\Theta}_2$: $\frac{MSE\left(\hat{\Theta}_1\right)}{MSE\left(\hat{\Theta}_2\right)} = \frac{\frac{\sigma^2}{3}}{\frac{3\sigma^2}{2}} = \frac{2}{9} < 1$, so $\hat{\Theta}_1$ is preferred.
2. $\hat{\Theta}_1, \hat{\Theta}_3$: $\frac{MSE\left(\hat{\Theta}_1\right)}{MSE\left(\hat{\Theta}_3\right)} = \frac{\frac{\sigma^2}{3}}{6\sigma^2+\mu^2} < 1$, so $\hat{\Theta}_1$ is preferred.
3. $\hat{\Theta}_2, \hat{\Theta}_3$: $\frac{MSE\left(\hat{\Theta}_2\right)}{MSE\left(\hat{\Theta}_3\right)} = \frac{\frac{3\sigma^2}{2}}{6\sigma^2+\mu^2} < 1$, so $\hat{\Theta}_2$ is preferred.

Review

Let us review very quickly the notions we have seen in this lecture:

- **Population**: X, where each element in the population is distributed with the same distribution (assume pdf $f(x)$) and with potentially unknown parameter(s) θ.
- **Random sample**: $X_1, X_2, \ldots, X_n$ each independent and from the same population with mean μ and variance σ^2.
  - $E[X_i] = E[X] = \mu$.
  - $Var[X_i] = Var[X] = \sigma^2$.
- **Statistic**: any function of a random variable.
- **Sampling distribution**: the distribution of a statistic.
- **Parameter**: (potentially unknown) information necessary to fully define the distribution of the population.
- **Point estimator** $\hat{\Theta}$: a statistic to estimate or approximate an unknown parameter θ.
- **Bias**: $E\left[\hat{\Theta}\right] - \theta$. we want this to be zero.
- **Standard error**: $\sqrt{Var\left[\hat{\Theta}\right]}$. we want this to be small.
- **Minimum variance unbiased estimator**: an estimator $\hat{\Theta}$ with zero bias and minimum variance.

Methods of point estimation

Learning objectives

After these lectures, we will be able to:

- Find point estimators for unknown parameters.
- Use the method of moments to find point estimators for unknown parameters.
- Use maximum likelihood estimation to find point estimators for unknown parameters.
  - Compare and identify when it is easiest to use likelihood and when log-likelihood.
- Propose new point estimators for unknown parameters based on these three methods.
- Calculate the unknown rate of an exponential distribution, or the unknown success probability of a Bernoulli distribution using the three methods.

Motivation: "I guess it is exponentially distributed. But what is λ?"

Motivation: Estimating the mortality risk

We call **mortality risk** of a hospital the probability of death occurring for any patient admitted to the hospital. The question we need to answer is "what is the mortality risk" of a given hospital? It depends on many factors, such as the type of conditions the patients admitted in this hospital have; the equipment of the hospital; the personnel of the hospital; among many, many others. Our intuition says the following, though: could we not *observe* the hospital for a period of time and then deduce what the risk is based on the obtained data? Is this fair/unfair/correct/misleading? What is the number of deaths in a hospital truly distributed as?

Estimation

During Lectures 15 and 16, we saw what makes a good point estimator $\hat{\Theta}$. We would like to have:

- small **bias** (zero, if possible). $bias = E\left[\hat{\Theta}\right] - \theta.$
- small **variance** (minimum among all estimators). $Var\left[\hat{\Theta}\right].$
- small **mean square error**. $MSE = bias^2 + Var\left[\hat{\Theta}\right].$

- We also defined the **relative efficiency** of two estimators $\hat{\Theta}_1, \hat{\Theta}_2$ as $\frac{MSE(\hat{\Theta}_1)}{MSE(\hat{\Theta}_2)}$.

So, *given* two or more estimators, you may calculate these items and infer which one to use/which one is better. However, where do these estimators come from? When faced with the problem of recognizing a parameter based on data, what can we do? In this series of lecture, we will work on deriving, using, and comparing three methods of point estimation:

1. **Method of moments** estimators.
2. **Maximum likelihood** estimators.
3. **Bayesian** estimators.

In this set of notes, we only deal with the first two. The third one is addressed in Lecture 19. Before we get to their details, we provide a definition and a motivating example.

Definition 56 (Method of estimation) *Assume we are provided a population X distributed with unknown parameter(s) θ. We want to estimate θ. Given a series of observations (sample) $X_1, X_2, \ldots, X$, how to come up with a "good" point estimator $\hat{\Theta}$?*

Mortality risk

Let us go back to our original motivating example with calculating/estimating the mortality risk of a hospital based on observations. Say, we have been observing the hospital over the last 2 months, and we have observed 18 deaths in the first 150 patient admissions. What would we estimate the mortality rate as?

Some more examples we may consider?

- How to estimate the rate of an exponential distribution?

  "We know the time between accidents in a factory is exponentially distributed. How do we find out what the rate is?"

- How to estimate the probability (proportion) of a binomial distribution?

  "We know the number of students graduating from the College of Engineering is binomially distributed. How do we find out what the probability of success (graduation) is?"

- How to estimate the mean and variance of a binomial distribution?

"We know exam grades in IE 300 are normally distributed. But, what is μ and σ^2?"

Method of moments

Methods

We begin the subsection with the definition of **sample** (empirical) **moments** and **population moments**. Assume we have a population X distributed with pdf $f(x)$. We have managed to collect a set of samples from the population $X_1, X_2, \ldots, X_n$. Then:

Definition 57 (Population moments) *The k-th population moment of a continuous population X (also referred to as the k-th moment of $f(x)$) is calculated as*

$$E\left[X^k\right] = \int_{-\infty}^{+\infty} x^k f(x) dx.$$

The same logic applies to the k-th population moment of a discrete population X, with a summation rather than an integration:

$$E\left[X^k\right] = \sum_{x \in X} x^k p(x) dx.$$

Definition 58 (Sample (empirical) moments) *The k-th sample moment of X (also referred to as the k-th empirical moment of X) is calculated as*

$$\frac{1}{n} \sum_{i=1}^{n} X_i^k,$$

where $X_1, X_2, \ldots, X_n$ are samples from the population X.

By definition, the first population moment of X is the population mean, and the first sample moment of X is the sample average. On the other hand, the second population moment of X is **not** the population variance; instead, $E\left[X^2\right]$ is only part of the calculation of the variance:

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

Similarly, the second sample moment of X is **not** the sample variance!

Calculating population moments

Assume $f(x) = \frac{1}{2}(1 - \alpha \cdot x)$ where α is some parameter. What are the first three population moments?

- first moment:

$$E[X] = \int_{-1}^{+1} x \cdot f(x) dx = \int_{-1}^{+1} x \cdot \frac{1}{2}(1 - \alpha \cdot x)\, dx = -\frac{\alpha}{3}.$$

- second moment:

$$E\left[X^2\right] = \int_{-1}^{+1} x^2 \cdot f(x) dx = \int_{-1}^{+1} x^2 \cdot \frac{1}{2}(1 - \alpha \cdot x)\, dx = \frac{1}{3}$$

- third moment:

$$E\left[X^3\right] = \int_{-1}^{+1} x^3 \cdot f(x) dx = \int_{-1}^{+1} x^3 \cdot \frac{1}{2}(1 - \alpha \cdot x)\, dx = -\frac{\alpha}{5}$$

Calculating sample (empirical) moments

Assume we have collected $n = 5$ samples from the population distributed with $X_1 = 0.7, X_2 = 0.77, X_3 = 0.65, X_4 = 0.5, X_5 = 0.83$. What are the first three sample moments?

- first moment:

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{5}(0.7 + 0.77 + 0.65 + 0.5 + 0.83) = 0.69.$$

- second moment:

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = \frac{1}{5}\left(0.7^2 + 0.77^2 + 0.65^2 + 0.5^2 + 0.83^2\right) = 0.48886.$$

- third moment:

$$\frac{1}{n}\sum_{i=1}^{n} X_i^3 = \frac{1}{5}\left(0.7^3 + 0.77^3 + 0.65^3 + 0.5^3 + 0.83^3\right) = 0.354189.$$

The method

The main idea behind the method is the following: **we want to match empirical (sample) moments of a distribution to the population moments**. Before we apply the method, we make a couple of observations.

Observation 1 The k-th moment of $f(x)$, $E\left[X^k\right]$ depends only on the unknown parameters $\theta_1, \theta_2, \ldots, \theta_m$.

Observation 2 The k-th moment of the sample, $\frac{1}{n}\sum_{i=1}^{n} X_i^k$ depends only on the data (the sample itself)!

So, if the 1st population moment is expected to *match* the 1st sample moment, and the 2nd population moment is expected to *match* the 2nd sample moment, and so on, then.. how many moments do we need to be able to solve a system of equations?

Based on the above discussion, we are now ready to formally state the method of moments. Assume we have m unknown parameters $\theta_1, \theta_2, \ldots, \theta_m$. The *method of moment estimators* $\hat{\Theta}_1, \hat{\Theta}_1, \ldots, \hat{\Theta}_m$ can be obtained by:

1. Get the first m [72] moments of $f(x)$ and of the sample.
2. Equate them.
3. Solve a system of equations with m unknowns (parameters θ_i)!

The solution obtained are the **method of moment estimators** $\hat{\Theta}_i$ for each parameter θ_i.

[72] Need to take more than m if some moments are zero or produce equations on the same variables as the previous ones.

Our first method of moments estimator

Recall earlier the population X distributed with $f(x) = \frac{1}{2}(1 - \alpha \cdot x)$ where α is some (unknown) parameter. We have collected a sample of $n = 5$ observations from X and we found the observations to be $X_1 = 0.7, X_2 = 0.77, X_3 = 0.65, X_4 = 0.5, X_5 = 0.83$. What is the method of moments estimator for α?

We have already found both the first population and the first sample moments. Looking at the earlier solutions, we have $E[X] = -\frac{\alpha}{3}$ and $\frac{1}{n}\sum_{i=1}^{n} X_i = 0.69$. Equating we get:

$$-\frac{\alpha}{3} = 0.69 \implies \hat{\alpha} = -2.07.$$

Observe how we put a "hat" ($\hat{}$) on top of α when we assign a value to it in the end. This is done to signal that this is merely an estimator and is not necessarily its true value.

The general case

In general, letting $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, given any sample of n observations, we may calculate the method of moments estimator for α as:

$$\hat{\alpha} = -3 \cdot \overline{X}.$$

The method of moments estimator for an exponential distribution

Assume we suspect X is a population that is exponentially distributed, but with unknown rate λ. Thankfully, we have collected a sample from that population: $X_1, X_2, \ldots, X_n$. We have one unknown parameter (λ) so we will need one equation.

Let us try the first population moment [73]:

$$E[X] = \frac{1}{\lambda}.$$

[73] Easy to find as it is the expected value of an exponential distribution!

Similarly, we may obtain the first sample moment as:

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}.$$

Recall that it is typical to denote a sample average as $\overline{X}$.

Equating the two (per the method of moments), we get:

$$\boxed{\hat{\lambda} = 1/\overline{X} = \frac{n}{\sum_{i=1}^{n} X_i}}$$

The method of moments estimator for a normal distribution

Assume we have some normally distributed population with mean μ and variance σ^2. Alas, they are both unknown. However, we have collected n observations (a sample) from the population: $X_1, X_2, \ldots, X_n$. What is the method of moments estimators for μ and σ^2.

We divide this proof in two parts:

1 The population moments.

For the population moments, we need (at least) the first two: $E[X]$ and $E\left[X^2\right]$. The first one is easy, as it is equal to μ. The second one on the other hand is **not the variance**: it is used in the variance calculation! Recall that $\sigma^2 = E\left[X^2\right] - (E[X])^2 \implies E\left[X^2\right] = \sigma^2 + (E[X])^2 = \sigma^2 + \mu^2$. In summary, we have:

$$E[X] = \mu$$
$$E\left[X^2\right] = \sigma^2 + \mu^2.$$

Estimating the rate of earthquakes

We assume that the time between two earthquakes of magnitude greater than or equal to 7 in Japan is exponentially distributed. Here is a list of earthquakes that satisfy these criteria from the last decade and when they have happened:

1	April 16, 2016
2	May 30, 2015
3	October 26, 2013
4	December 7, 2012
5	July 10, 2011
6	April 11, 2011
7	April 7, 2011
8	March 11, 2011
9	March 11, 2011
10	March 9, 2011
11	December 21, 2010
12	February 26, 2010

What is the method of moments estimator for the rate λ?

We first consider the time between the earthquakes. We have 11 such observations (between the first and the second, between the second and the third, etc.). Let us count this in days (for consistency): 322 days, 581 days, 323 days, 516 days, 90 days, 4 days, 27 days, 0 days, 2 days, 78 days, 318 days.
From the method of moments, we want the first population and sample method (as we only have one unknown parameter), so:

$$\begin{cases} E[X] = \frac{1}{\lambda} \\ \frac{1}{n}\sum_{i=1}^{n} X_i = 205.55 \end{cases} \implies \hat{\lambda} = 1 \text{ earthquake per } 205.55 \text{ days.}$$

2 The sample moments. These are easier to calculate as:

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$$
$$\frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

Here, again, we use $\overline{X}$ to represent the sample average.

Equating the two, we get the following system of equations:

$$\mu = \overline{X} \implies \hat{\mu} = \overline{X}.$$

$$\mu^2 + \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \implies \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} X_i^2 - n\mu^2}{n} = \frac{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}{n}.$$

Estimating the grade distribution of an exam

Assume we want to estimate the grade distribution of an exam *before* we grade all of the exams! If we expect the distribution to be normally distributed, we could grade the first five exams (at random) and get their grades $X_1 = 80, X_2 = 97, X_3 = 50, X_4 = 67, X_5 = 84$. Then, we calculate $\frac{1}{n}\sum_{i=1}^{n} X_i = 75.6$ and $\frac{1}{n}\sum_{i=1}^{n} X_i^2 = 5970.8$. Finally, from the method of moments, we have:

$$\begin{cases} E[X] = \frac{1}{n}\sum_{i=1}^{n} X_i = 75.6 \\ E[X^2] = \frac{1}{n}\sum_{i=1}^{n} X_i^2 = 5970.8 \end{cases} \implies \begin{cases} \hat{\mu} = 75.6. \\ \hat{\sigma}^2 = 255.44. \end{cases}$$

The method of moments estimator for a Bernoulli distribution

Finally, let X be a Bernoulli random variable with probability of success p, that is unknown. How to estimate it using the method of moments? Well, we resort to the following setup. Let us run n experiments of that Bernoulli random variable and let's mark each of them as X_i with a 1 (when successful) or a 0 (when failed). Then:

$$E[X] = p$$
$$\frac{1}{n}\sum_{i=1}^{n} X_i$$

Equating the two, we get that

$$p = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Fair or unfair?

Assume you have an unfair coin, but you have no idea how unfair it is – that is, p is not known. Say you toss the coin $n = 10$ times and get 7 Heads, 3 Tails. What is the estimator you get for the probability of getting Heads from the method of moments?

Let Heads be equal to 1 and Tails equal to 0. Then: $\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{7}{10} = 0.7$. From the method of moments $\hat{p} = 0.7$.

A few extra examples

This is an example from the slides. In the slides, we mention that the distribution is normal; but this is not necessary!

A delivery problem

We believe the times it takes to deliver a package are identically distributed with the same unknown mean μ and variance σ^2. We have collected information on 10 packages and the time to delivery (in hours) are: 49.1, 47.9, 48.6, 50.4, 49.5, 49.8, 48.2, 50.3, 45.2, 46.2. What are good mean and variance estimators for the normal distribution using the method of moments?

We have two unknown parameters (mean and variance), so we will need at least two population and sample moments. Let us take the first two:

- Population 1st moment:

$$E\left[X^1\right] = E\left[X\right] = \mu$$

- Sample 1st moment:

$$\frac{1}{10}\sum_{i=1}^{10} X_i^1 = 48.52$$

- Population 2nd moment:

$$E\left[X^2\right] = Var\left[X\right] + \left(E\left[X\right]\right)^2 = \\ = \sigma^2 + \mu^2$$

- Sample 2nd moment:

$$\frac{1}{10}\sum_{i=1}^{10} X_i^2 = 2356.844$$

Equating the two and solving the system of equations, we get $\hat{\mu} = 48.52$ and $\hat{\sigma}^2 = 2.6536$.

A discrete distribution

Assume we have a discrete random variable X defined over $0, 1, 2, 3, 4$ and distributed with probabilities $p(0) = \frac{\theta_1}{3}, p(1) = \frac{\theta_1}{6}, p(2) = \frac{\theta_1}{6}, p(3) = \frac{\theta_2}{2}, p(4) = \frac{\theta_2}{2}$.

Now, assume we have collected a sample of $n = 10$ observations: $0, 1, 1, 3, 4, 2, 2, 3, 4, 1$. Based on this, what is the method of moments estimators for θ_1 and for θ_2?

Now, let's see. At first glance we have two estimators.. But we know better than that. We probably remember that $\sum_{x=0}^{4} p(x) = 1$, which implies that:

$$\sum_{x=0}^{4} p(x) = 1 \implies \frac{\theta_1}{3} + \frac{\theta_1}{6} + \frac{\theta_1}{6} + \frac{\theta_2}{2} + \frac{\theta_2}{2} = 1 \implies \frac{2\theta_1}{3} + \theta_2 = 1.$$

Based on that, if we knew, say θ_1 we could obtain θ_2 right away. Let us get the first moments and equate them:

$$E[X] = 0 \cdot \frac{\theta_1}{3} + 1 \cdot \frac{\theta_1}{6} + 2 \cdot \frac{\theta_1}{6} + 3 \cdot \frac{\theta_2}{2} + 4 \cdot \frac{\theta_2}{2} = \frac{\theta_1 + 7\theta_2}{2}.$$

$$\frac{1}{n} \sum_{i=1}^{10} X_i = \frac{1}{10} (0 + 1 + 1 + 3 + 4 + 2 + 2 + 3 + 4 + 1) = 2.1.$$

Finally, we have a system of equations at our hands!

$$\begin{cases} \frac{2\theta_1}{3} + \theta_2 = 1 \\ \frac{\theta_1 + 7\theta_2}{2} = 2.1 \end{cases} \implies \begin{cases} \hat{\theta}_1 = \frac{42}{55} \\ \hat{\theta}_2 = \frac{27}{55} \end{cases}$$

Maximum likelihood estimation

Basics

Recall that we already have a population X distributed with pdf $f(x)$. Also recall that the pdf has one or more unknown parameters θ. We may then write that the pdf is actually a function of x and θ as $f(x, \theta)$. That is, we need inputs for both the value x and the parameter(s) θ before evaluating $f(x)$. Finally, we have already collected a sample of n observations from the population, let them be $X_1, X_2, \ldots, X_n$.

This brings us to the definition of the **likelihood function**.

Definition 59 (Likelihood function) *The **likelihood function** of a sample of n observations $X_1, X_2, \ldots, X_n$ is defined as*

$$L(\theta) = f(X_1, \theta) \cdot f(X_2, \theta) \cdot \ldots \cdot f(X_n, \theta) = \prod_{i=1}^{n} f(X_i, \theta).$$

Observe how the likelihood function is only a function of θ as $X_i, i = 1, \ldots, n$ are known quantities.

The method

The main idea is pretty simple: for the sample to have been obtained the way it has, then the observations must have been **likely**. Hence, they must be values that maximize the likelihood function! This is summarized in the following statement:

The **maximum likelihood estimators** $\hat{\Theta}$ are
the values that *maximize* the likelihood function.

The maximum likelihood estimators are also referred to as MLE. To find this maximizer, we take the first derivative of the likelihood function and equate it to 0:

$$\frac{\partial L}{\partial \theta} = 0$$

and solve for θ to obtain the estimator.

Our first MLE estimator

Go back again to the population X distributed with $f(x) = \frac{1}{2}(1 - \alpha \cdot x)$ where α is the unknown parameter we would like to estimate. We have a sample from X as $X_1 = 0.7, X_2 = 0.77, X_3 = 0.65, X_4 = 0.5, X_5 = 0.83$. What is the MLE estimator for α?

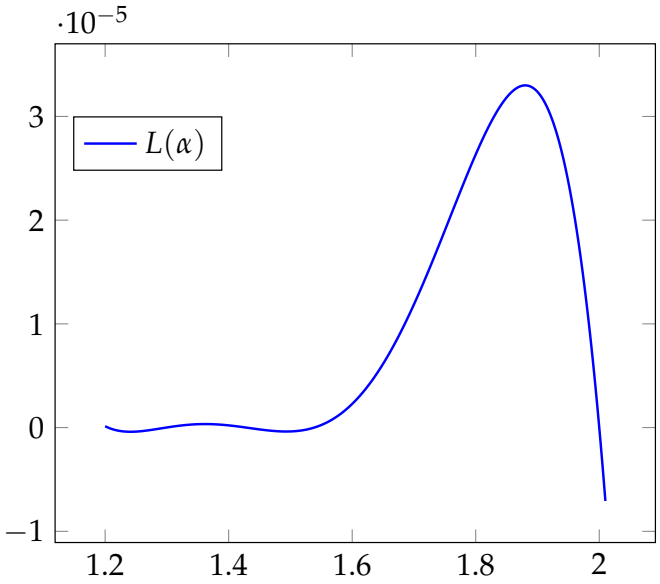
First to build the likelihood function:

$$\begin{aligned} L(\alpha) &= f(X_1) \cdot f(X_2) \cdot f(X_3) \cdot f(X_4) \cdot f(X_5) = \\ &= \frac{1}{32}(1 - 0.7\alpha)(1 - 0.77\alpha)(1 - 0.65\alpha)(1 - 0.5\alpha)(1 - 0.83\alpha) = \\ &= \frac{1}{32} - 0.107813\alpha + 0.147784\alpha^2 - 0.100557\alpha^3 + 0.0339432\alpha^4 - 0.0045436\alpha^5 \end{aligned}$$

Then, we get the first derivative and set it equal to o to find the maximizer. We get:

$$\frac{\partial L}{\partial \alpha} = 0 \implies \alpha = 1.88.$$

This solution could also be found visually! Here is a plot of the likelihood function and the point where it is maximized is easier to find.



Finally, observe how we got a different estimator here compared to the method of moments!

Extension to log-likelihood

Since the likelihood involves a product of n pdf values, it comes as no surprise that our end result may be a little difficult to control and use. This is why we may also introduce the **log-likelihood**:

Definition 60 (Log-likelihood function) *The **log-likelihood function** of a sample of n observations $X_1, X_2, \ldots, X_n$ is defined as*

$$\ln L(\theta) = \ln f(X_1, \theta) + \ln f(X_2, \theta) + \ldots + \ln f(X_n, \theta) = \sum_{i=1}^{n} \ln f(X_i, \theta).$$

Observe how also the log-likelihood function is only a function of θ as $X_i, i = 1, \ldots, n$ are known quantities. Contrary to the simple likelihood function, the log-likelihood is a summation which makes it easier to differentiate.

Our first MLE estimator using log-likelihood

We have:

- pdf $f(x) = \frac{1}{2}(1 - \alpha \cdot x)$;
- sample $X_1 = 0.7, X_2 = 0.77, X_3 = 0.65, X_4 = 0.5, X_5 = 0.83$.

We build the log-likelihood function as:

$$\begin{aligned}\ln L(\alpha) &= \ln f(X_1) + \ln f(X_2) + \ln f(X_3) + \ln f(X_4) + \ln f(X_5) = \\ &= \ln \frac{1}{2}(1 - 0.7\alpha) + \ln \frac{1}{2}(1 - 0.77\alpha) + \ln \frac{1}{2}(1 - 0.65\alpha) + \\ &+ \ln \frac{1}{2}(1 - 0.5\alpha) + \ln \frac{1}{2}(1 - 0.83\alpha).\end{aligned}$$

Here, we note that $\left(\frac{1}{2}(1 - X_i\alpha)\right)' = \frac{X_i}{\alpha X_i - 1}$. Hence, in our case we have:

$$\begin{aligned}&\frac{\partial \ln L}{\partial \alpha} = 0 \implies \\ &\frac{0.7}{1 - 0.7\alpha} + \frac{0.77}{1 - 0.77\alpha} + \frac{0.65}{1 - 0.65\alpha} + \frac{0.5}{1 - 0.5\alpha} + \frac{0.83}{1 - 0.83\alpha} = 0 \implies \\ &\implies \alpha = 1.88.\end{aligned}$$

The result obtained using the likelihood or the log-likelihood function will be the same.

The MLE estimator for an exponential distribution

Assume we have obtained a sample of n observations $X_1, X_2, \ldots, X_n$ with average $\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$. We also assume that the population

is exponentially distributed with rate λ. What is the MLE estimator for λ?

First, we build the log-likelihood function as:

$$\begin{aligned}\ln L(\lambda) &= \ln \lambda e^{-\lambda X_1} + \ln \lambda e^{-\lambda X_2} + \ldots + \ln \lambda e^{-\lambda X_n} = \\ &= \ln \lambda - \lambda X_1 + \ln \lambda - \lambda X_2 + \ldots + \ln \lambda - \lambda X_n = \\ &= n \ln \lambda - \lambda \left(X_1 + X_2 + \ldots + X_n\right)\end{aligned}$$

Again, find the maximizer:

$$\frac{\partial \ln L(\lambda)}{\partial \lambda} = 0 \implies \left(n \ln \lambda - \lambda \left(X_1 + X_2 + \ldots + X_n\right)\right)' = 0 \implies$$
$$\frac{n}{\lambda} - \left(X_1 + X_2 + \ldots + X_n\right) = 0 \implies n - \lambda \sum_{i=1}^{n} X_i = 0 \implies \lambda = \frac{n}{\sum\limits_{i=1}^{n} X_i} = \frac{1}{\overline{X}}$$

Observe how we have reached the same result as when using the method of moments. Recall that this is **not necessarily** always the case.

Say we had not wanted to use the log-likelihood and instead used the simple likelihood function $L(\lambda)$:

$$\begin{aligned}L(\lambda) &= \lambda e^{-\lambda X_1} \cdot \lambda e^{-\lambda X_2} \cdot \ldots \cdot \lambda e^{-\lambda X_n} = \lambda^n \cdot e^{-\lambda (X_1 + X_2 + \ldots + X_n)} = \\ &= \lambda^n \cdot e^{-\lambda \sum\limits_{i=1}^{n} X_i}\end{aligned}$$

Take the derivative:

$$\begin{aligned}&\frac{\partial L(\lambda)}{\partial \lambda} = 0 \implies \left(\lambda^n \cdot e^{-\lambda \sum\limits_{i=1}^{n} X_i}\right)' = 0 \implies \\ &\implies n \cdot \lambda^{n-1} \cdot e^{-\lambda \sum\limits_{i=1}^{n} X_i} - \lambda^n \cdot \sum_{i=1}^{n} X_i \cdot e^{-\lambda \sum\limits_{i=1}^{n} X_i} = 0.\end{aligned}$$

Observe how we can simplify quite a bit: we may divide by λ^{n-1} (because we know that $\lambda > 0$). This gives us:

$$n \cdot e^{-\lambda \sum\limits_{i=1}^{n} X_i} - \lambda \cdot \sum_{i=1}^{n} X_i \cdot e^{-\lambda \sum\limits_{i=1}^{n} X_i} = 0$$

We may also divide by $e^{-\lambda \cdot \sum\limits_{i=1}^{n} X_i}$ because it is also definitely positive. This leads to the much more manageable:

$$n - \lambda \cdot \sum_{i=1}^{n} X_i = 0 \implies \lambda = \frac{n}{\sum\limits_{i=1}^{n} X_i} = \frac{1}{\overline{X}}$$

Note how getting the same result with the log-likelihood was significantly easier due to the nature of this probability density function.

The MLE estimator for a Bernoulli distribution

Once again, consider that we have a population producing random variables distributed as Bernoulli with probability of success p. We have obtained a sample of $n = 10$ observations with 7 successes (let them be $X_i = 1$) and 3 failures ($X_i = 0$). What is the MLE estimator for the unknown p?

Based on the MLE method we first need to calculate the likelihood function. Recall that for a Bernoulli random variable its probability mass function (as it is a discrete random variable) is $P(0) = 1 - p$ and $P(1) = p$. Without loss of generality, assume we arrange the observations with the successes first (the first, say, 7 observations) and the failures next (the remaining $10 - 7 = 3$ observations).

We are now ready to build the likelihood function:

$$L(p) = \left(\prod_{i=1}^{7} p\right) \cdot \left(\prod_{i=8}^{10} (1-p)\right) = p^7 \cdot (1-p)^{10-7} = p^7 \cdot (1-p)^3 .$$

The derivative of the likelihood function can be found as:

$$\frac{\partial L}{\partial p} = \left(p^7 \cdot (1-p)^3\right)' = 7p^6(1-p)^3 - 3(1-p)^2 p^7.$$

Now, equate this to 0 to get the maximizer:

$$\begin{aligned} 7p^6(1-p)^3 - 3(1-p)^2 p^7 = 0 \implies \\ \implies 7(1-p) - 3p = 0 \implies \hat{p} = 0.7. \end{aligned}$$

In the above, we make the assumption that $p \in (0, 1)$: that is, it cannot be 0 or 1. If we allowed this to be the case, then solving would give three solutions $p = 0, p = 1, p = 0.7$. However, the first two solutions are *minima* rather than *maxima*, and we would still pick $\hat{p} = 0.7$ as our estimator.

A few extra examples

This is an example from the slides. In the slides, we mention that the distribution is normal; but this is not necessary!

A delivery problem

We believe the times it takes to deliver a package are identically distributed with the same unknown mean μ and variance σ^2. We have collected information on 10 packages and the time to delivery (in hours) are: 49.1, 47.9, 48.6, 50.4, 49.5, 49.8, 48.2, 50.3, 45.2, 46.2. What are good mean and variance estimators for the normal distribution using the method of moments?

We have two unknown parameters (mean and variance), so we will need at least two population and sample moments. Let us take the first two:

- Population 1st moment:

$$E\left[X^1\right] = E[X] = \mu$$

- Sample 1st moment:

$$\frac{1}{10}\sum_{i=1}^{10} X_i^1 = 48.52$$

- Population 2nd moment:

$$E\left[X^2\right] = Var[X] + (E[X])^2 = \\ = \sigma^2 + \mu^2$$

- Sample 2nd moment:

$$\frac{1}{10}\sum_{i=1}^{10} X_i^2 = 2356.844$$

Equating the two and solving the system of equations, we get $\hat{\mu} = 48.52$ and $\hat{\sigma}^2 = 2.6536$.

A discrete distribution

Assume we have a discrete random variable X defined over $0, 1, 2, 3, 4$ and distributed with probabilities $p(0) = \frac{\theta_1}{3}, p(1) = \frac{\theta_1}{6}, p(2) = \frac{\theta_1}{6}, p(3) = \frac{\theta_2}{2}, p(4) = \frac{\theta_2}{2}$.

Now, assume we have collected a sample of $n = 10$ observations: $0, 1, 1, 3, 4, 2, 2, 3, 4, 1$. Based on this, what is the method of moments estimators for θ_1 and for θ_2?

Now, let's see. At first glance we have two estimators.. But we know better than that. We probably remember that $\sum_{x=0}^{4} p(x) = 1$, which implies that:

$$\sum_{x=0}^{4} p(x) = 1 \implies \frac{\theta_1}{3} + \frac{\theta_1}{6} + \frac{\theta_1}{6} + \frac{\theta_2}{2} + \frac{\theta_2}{2} = 1 \implies \frac{2\theta_1}{3} + \theta_2 = 1.$$

Based on that, if we knew, say θ_1 we could obtain θ_2 right away. Let us get the first moments and equate them:

$$E[X] = 0 \cdot \frac{\theta_1}{3} + 1 \cdot \frac{\theta_1}{6} + 2 \cdot \frac{\theta_1}{6} + 3 \cdot \frac{\theta_2}{2} + 4 \cdot \frac{\theta_2}{2} = \frac{\theta_1 + 7\theta_2}{2}.$$

$$\frac{1}{n} \sum_{i=1}^{10} X_i = \frac{1}{10} (0 + 1 + 1 + 3 + 4 + 2 + 2 + 3 + 4 + 1) = 2.1.$$

Finally, we have a system of equations at our hands!

$$\begin{cases} \frac{2\theta_1}{3} + \theta_2 = 1 \\ \frac{\theta_1 + 7\theta_2}{2} = 2.1 \end{cases} \implies \begin{cases} \hat{\theta}_1 = \frac{42}{55} \\ \hat{\theta}_2 = \frac{27}{55} \end{cases}$$

Bayesian estimation

Learning objectives

After this lecture, we will be able to:

- Use Bayesian estimation to find point estimators for unknown parameters.
- Propose new point estimators for unknown parameters based on prior information.

Motivation: Heads or Tails?

We flip a coin 10 times and we get 6 Heads and 4 Tails. Do you believe it is a fair coin? What does the method of moments and the maximum likelihood estimation method say about this situation?

Quick review

During these past two lectures, we discussed two methods to identify "good" estimators $\hat{\Theta}$ for a series of unknown parameters:

- **the method of moments**.
  1. Compute the moments of the population, calculated as $E\left[X^k\right]$.
  2. Compute the moments of the sample (empirical moments), calculated as $\frac{1}{n}\sum_{i=1}^{n} X_i^k$.
  3. Equate the two and solve a system of equations for the unknown parameters.
- **maximum likelihood estimation**.
  1. Calculate the likelihood function as
  $$L(\theta) = f(X_1, \theta) \cdot f(X_2, \theta) \cdot \ldots \cdot f(X_n, \theta)$$
  2. Or the log-likelihood function as
  $$\ln(L(\theta)) = \ln(f(X_1, \theta)) + \ln(f(X_2, \theta)) + \ldots + \ln(f(X_n, \theta))$$
  3. Find the maximizer (usually found by setting the derivative per each unknown parameter equal to 0).

Both these methods have one thing in common: they require no prior information to work, but instead they base all of their observations on the obtained sample. What if I already know some more information about what is going on?

Bayesian estimation through an example

We begin in a slightly different way than usually. We begin with an example to help us build intuition! Assume I carry 3 coins with me:

1. One with both sides showing Heads.
2. One with both sides showing Tails.
3. One that is fair and has a side of Heads and a side of Tails.

Assume I randomly pick one coin and start flipping it. I report to you the number of tries (n) and the number of Heads (x). For example, I may tell you $n = 8$, $x = 5$ or $n = 2$, $x = 0$, and so on.

Flipping the coin: first take

I let you know that I flipped the coin three times and got Heads both times: $n = 3, x = 2$. What are the method of moments and the maximum likelihood estimators for p?

We will have $E[X] = p$ and $\frac{1}{3} \cdot (1+1+0) = \frac{2}{3}$, and equating will give $\hat{p} = \frac{2}{3}$. The likelihood function is $L(p) = p^2 \cdot (1-p)$, and maximizing will also give $\hat{p} = \frac{2}{3}$.

But... I carry three coins with me. Shouldn't I use this information somehow?

1. Can it be my "2-Heads" coin?
2. Can it be my "2-Tails" coin?
3. Does it have to be my "50-50" coin?

This is the key to realizing what Bayesian estimation brings to the table: extra information in the form of prior probabilities for the parameters that are unknown.

Bayesian estimation

We separate the discussion between discrete sets for the values the parameter can take (like in the previous example where I carried 3 distinct coins with me) and between continuous sets, where the parameter can be any real number in a range of values.

For discrete parameter values

Before describing the method, we provide some notation:

- **prior probabilities** ("priors"): the probability of seeing a certain parameter $P(\theta)$.

- **likelihood probabilities** ("likelihoods"): the likelihood of seeing an outcome *given* a certain parameter $P(X|\theta)$.
- **posterior probabilities** ("posteriors"): the multiplication of the two $P(\theta) \cdot P(X|\theta)$.

A quick note about the likelihoods: those are calculated in identical manner as the likelihood function in the maximum likelihood estimation method!
The Bayesian estimation method then states that:

"The higher the posterior probability,
the better the chance of having that parameter."

This is it! This is the whole method!

Flipping the coin: second take

Let us go back to the example where I carried three coins ("2-Heads"; "2-Tails"; and "50-50") and I picked one at random. After 3 tries, we got 2 Heads: $n = 3, x = 2$. Let's see what we have for these three distinct cases of $p = 1, p = 0, p = 0.5$:

- priors $P(p)$: probability of picking a certain coin, that is $P(p = 1) = P(p = 0) = P(p = 0.5) = \frac{1}{3}$.
- likelihoods $P(X = 2|p)$: likelihood function of seeing two Heads for each coin. For example, the likelihood function for the $p = 0.5$ coin with $x = 2$ Heads in $n = 3$ tries would be: $p^2 \cdot (1 - p) = 0.5^2 \cdot 0.5 = 0.125$.
- posteriors $P(p) \cdot P(X = 2|p)$: we will need to calculate this for each coin.

Let us put this in table format.

parameter p	prior $P(p)$	likelihood $P(X = 2\|p)$	posterior $P(p) \cdot P(X = 2\|p)$
0	$\frac{1}{3}$	$0^2 \cdot 1^1 = 0$	$\frac{1}{3} \cdot 0 = 0$
1	$\frac{1}{3}$	$1^2 \cdot 0^1 = 0$	$\frac{1}{3} \cdot 0 = 0$
0.5	$\frac{1}{3}$	$0.5^2 \cdot 0.5^1 = 0.125$	$\frac{1}{3} \cdot 0.125 = 0.0416\overline{6}$

The maximum value (and only non-zero probability!) is achieved for the "50-50" coin so it must be this!

See? It is pretty intuitive. Of course, we may complicate things by making the probability of picking a coin a little more general.

Flipping the coin: third take

I still have three types of coins on me. But, given that I am an adult that carries money wherever I go, I carry more actual coins ("50-50") than novelty coins ("2-Heads", "2-Tails"). More specifically, I carry 8 real coins and 1 of each novelty coin. I take a coin out and toss it twice and get two Heads! Which coin is it?

Let us try the table format again:

parameter p	prior $P(p)$	likelihood $P(X=2\|p)$	posterior $P(p)\cdot P(X=2\|p)$
0	$\frac{1}{8}$	$0^2=0$	$\frac{1}{8}\cdot 0=0$
1	$\frac{1}{8}$	$1^2=1$	$\frac{1}{8}\cdot 1=0.125$
0.5	$\frac{3}{4}$	$0.5^2=0.25$	$\frac{3}{4}\cdot 0.25=0.1875$

The maximum value is still achieved for a "50-50" coin, so we are inclined to think we picked one. Note how much closer the posteriors are, though..

It would actually take one more Heads to change our parameter estimation towards the "2-Heads" novelty coin! Why is that?

Normalizing may also be useful. Instead of looking at the posterior values as they are in the end, we may turn them into actual "%" values to help compare them. To normalize simply take each posterior and divide it by the summation of all posterior probabilities. For example, in our third take (see above) we would end up with probabilities:

- $P(p=0)=\frac{0}{0+0.125+0.1875}=0.$
- $P(p=1)=\frac{0.125}{0+0.125+0.1875}=0.4.$
- $P(p=0.5)=\frac{0.1875}{0+0.125+0.1875}=0.6.$

This helps us quantify our parameter estimation even more. There is a 40% chance we picked the "2-Heads" coin and a 60% chance we picked one of the "50-50" coins.
Let's work one more example before we move to the continuous case.

A computer vision system: third take

A machine learning algorithm for computer vision is trained to observe the first vehicle that passes from an intersubsection at or after 8am every day. Then, it reports the time from that vehicle to the next one again and again. We assume this time is exponentially distributed but with unknown λ.

- If the first vehicle of the day was a personal car, then $\lambda_1 = 1$ per minute.
- If the first vehicle of the day was a motorcycle, then $\lambda_2 = 1$ per 5 minutes.
- If the first vehicle was a truck, then $\lambda_3 = 1$ per 10 minutes.
- If the first vehicle was a bike, then $\lambda_4 = 1$ per 12 minutes.

We observe a sample of 5 times: $X_1 = 9$ minutes, $X_2 = 8.5$ minutes, $X_3 = 8$ minutes, $X_4 = 10.5$ minutes. What is the probability of each parameter λ?

First, we need to calculate the prior probabilities $P(\lambda)$. The first vehicle of the day is:

- a personal car with probability $\frac{\lambda_1}{\lambda_1+\lambda_2+\lambda_3+\lambda_4} = 0.732$ (why?),
- a motorcycle with probability $\frac{\lambda_2}{\lambda_1+\lambda_2+\lambda_3+\lambda_4} = 0.146$,
- a truck with probability $\frac{\lambda_3}{\lambda_1+\lambda_2+\lambda_3+\lambda_4} = 0.073$,
- or a bike with probability $\frac{\lambda_4}{\lambda_1+\lambda_2+\lambda_3+\lambda_4} = 0.049$.

With these in hand, we calculate the likelihood functions as being $\lambda \cdot e^{-\lambda \cdot X_1} \cdot \lambda \cdot e^{-\lambda \cdot X_2} \cdot \ldots \cdot \lambda \cdot e^{-\lambda \cdot X_n}$ since we have an exponentially distributed random variable.
For example, if $\lambda = \lambda_1 = 1$, then we would have $1 \cdot e^{-1 \cdot 9} \cdot 1 \cdot e^{-1 \cdot 8.5} \cdot 1 \cdot e^{-1 \cdot 8} \cdot 1 \cdot e^{-1 \cdot 10.5} = e^{-36} = 2.32 \cdot 10^{-16}$.
Finally:

parameter λ	prior $P(\lambda)$	likelihood $P(X_1, X_2, X_3, X_4 \mid \lambda)$	posterior $P(\lambda) \cdot P(X_1, X_2, X_3, X_4 \mid \lambda)$
$\lambda_1 = 1$	0.732	$2.32 \cdot 10^{-16}$	$1.70 \cdot 10^{-16}$
$\lambda_2 = 0.2$	0.146	$1.19 \cdot 10^{-6}$	$1.74 \cdot 10^{-7}$
$\lambda_3 = 0.1$	0.073	$2.73 \cdot 10^{-6}$	$1.99 \cdot 10^{-7}$
$\lambda_4 = 0.06\overline{6}$	0.049	$1.79 \cdot 10^{-6}$	$8.77 \cdot 10^{-8}$

From the results it seems that the vehicle that first passed today is more likely a truck!

If we wanted to assign probability values to each type of vehicle, we would report:

- personal car: $\frac{6.24 \cdot 10^{-17}}{6.24 \cdot 10^{-17} + 1.43 \cdot 10^{-7} + 1.81 \cdot 10^{-7} + 8.18 \cdot 10^{-8}} = 1.54 \cdot 10^{-10} \approx 0.$
- motorcycle: 0.3527.
- truck: 0.4458.
- bike: 0.2015.

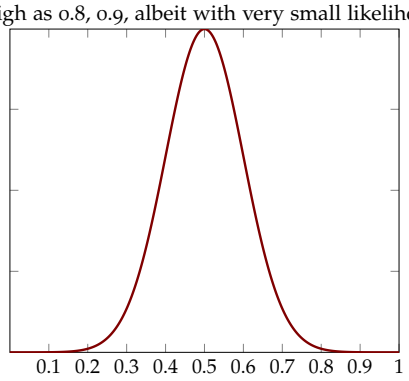
We can now move to the continuous case.

For continuous parameter values

Let us again begin with an example. The method is largely still the same; but the definitions of some of the items change slightly to accommodate the continuous nature of the unknown parameter(s).

Say we have a coin that is made with the goal of being fair; that is, "50-50". But, materials fail and get deposited more on one side than the other resulting in different compositions for the probability of Heads and Tails. Say, in the end, the probability of Heads is normally distributed with $\mathcal{N}(0.5, 0.01)$, that is a mean of $\mu = 0.5$ and a variance $\sigma^2 = 0.01 \implies \sigma = 0.1$. Visually, we would get the distribution of Figure 63

Figure 63: The distribution of the probability of getting Heads in the continuous version of the problem. We see how $p = 0.5$ is more likely, but we can get values as low as 0.1, 0.2, or as high as 0.8, 0.9, albeit with very small likelihood.

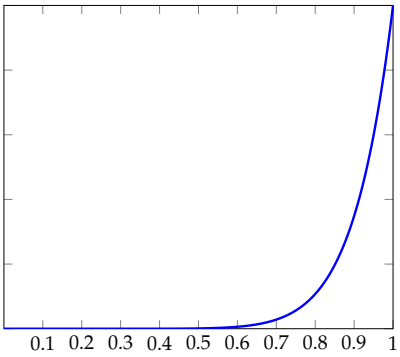


Now that we know this, say we tossed a coin 10 times, and got 10 straight times Heads! Recall that both the method of moments and the maximum likelihood estimation method would simply assume that the coin has $p = 1$ and proceed.

Getting 10 Heads in 10 tosses would be highly improbable for a coin that is "50-50", but it could mean that I have a biased coin towards Heads. So, what should our estimate be?

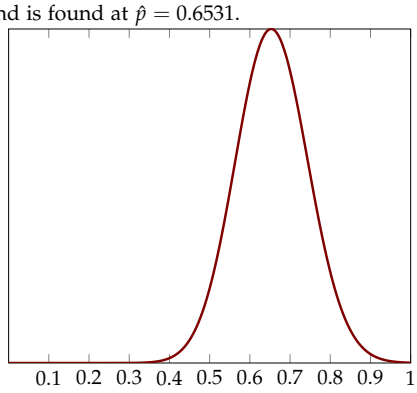
First, calculate the likelihood function, the way we did during the maximum likelihood estimation calculations. In this case, it would be $L(p) = p^{10}$. Let's plot that (see Figure 64).

Figure 64: The likelihood function of getting 10 Heads after tossing a coin 10 times. It is maximized at $p = 1$, which would then be our maximum likelihood estimator.



In Bayesian estimation for discrete-valued parameters earlier, we calculated $P(\theta)$ (priors) with $P(X|\theta)$ (likelihoods) to obtain a series of posteriors that we would compare. In the continuous version, we calculate $f(\theta)$ (prior distribution) with $L(\theta)$ (likelihood function) to obtain a posterior distribution that we would then find the maximizer at! Confused? Let's look at this visually again in Figure 65.

Figure 65: The posterior distribution, found by multiplying $f(\theta)$ (the pdf of the normal distribution $\mathcal{N}(0.5, 0.01)$) with the likelihood function $L(\theta)$. The maximizer here is the Bayesian estimator and is found at $\hat{p} = 0.6531$.



Let us define the notation for the method then:

- **prior distribution**: the distribution of the real-valued and continuous parameter θ, $f(\theta)$.
- **likelihood function**: the likelihood function, built just as in the maximum likelihood estimation metho, $L(\theta)$.

- **posterior distribution**: the multiplication of the two functions $f(\theta) \cdot L(\theta)$.

The Bayesian estimation method for continuous parameters states that:

> "The Bayesian estimator is found by maximizing the posterior distribution."

And, yes! This sums it up. Let us view one example from beginning to end using the method.

Mortality risk

We call mortality risk of a hospital the probability of death occurring for any patient admitted to the hospital. The mortality risk in US hospitals is in general **exponentially distributed** with a mean at 1.5% (that is, $\lambda = \frac{1}{1.5\%}$). You have been observing a hospital and have seen 25 deaths in the first 150 patient admissions. What is the Bayesian estimator for the true mortality rate of the hospital?

Right away, distinguish between two items:

1. the prior distribution that we believe the mortality risk to be distributed as (exponential)
2. the mortality rate itself is a Bernoulli random variable (p and $1 - p$); in our case, we have a sample we have collected (25 deaths in 150 admissions) to help us estimate p.

So, let us start collecting what we need one-by-one.

Prior distribution:

$$f(p) = \frac{1}{1.5} e^{-\frac{1}{1.5}p}.$$

Likelihood function:

$$L(p) = p^{25} \cdot (1-p)^{125}.$$

Posterior distribution:

$$f(p) \cdot L(p) = \frac{1}{1.5} e^{-\frac{1}{1.5}p} \cdot p^{25} \cdot (1-p)^{125}.$$

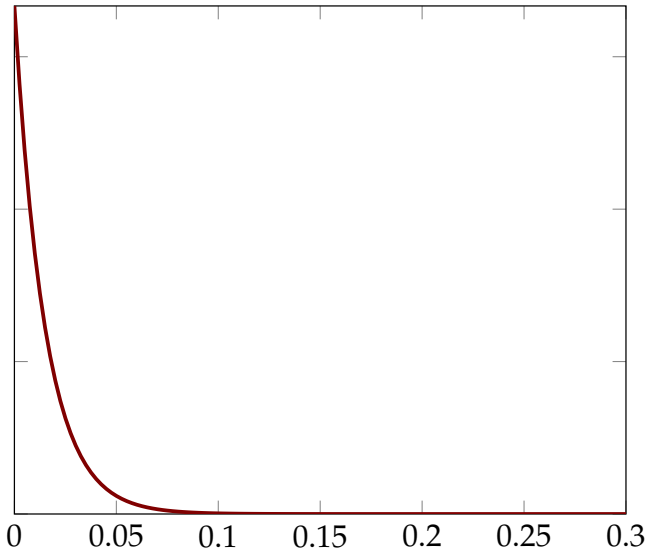
Mortality risk (cont'd)

To maximize, get the derivative and equate to 0 to get:

$$\frac{\partial f(p) \cdot L(p)}{\partial p} = \frac{4}{9} e^{-\frac{2}{3}p} (p-1)^{124} p^{24} ((p-226)\, p + 37.5) = 0 \implies$$
$$\implies ((p-226)\, p + 37.5) = 0 \implies p = 0.16605.$$

The maximizer is, then, at $\hat{p} = 0.16605$.

If we want to, we can see the same result visually. First, plot our prior beliefs/distribution:



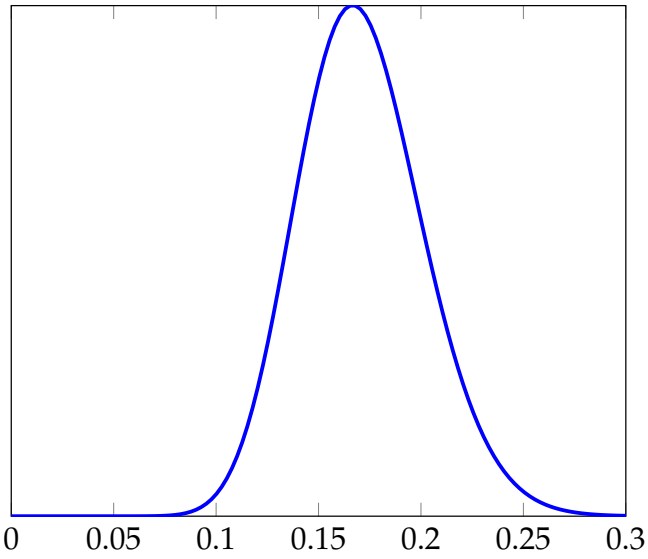
Then, plot our likelihood function based on the sample collected:

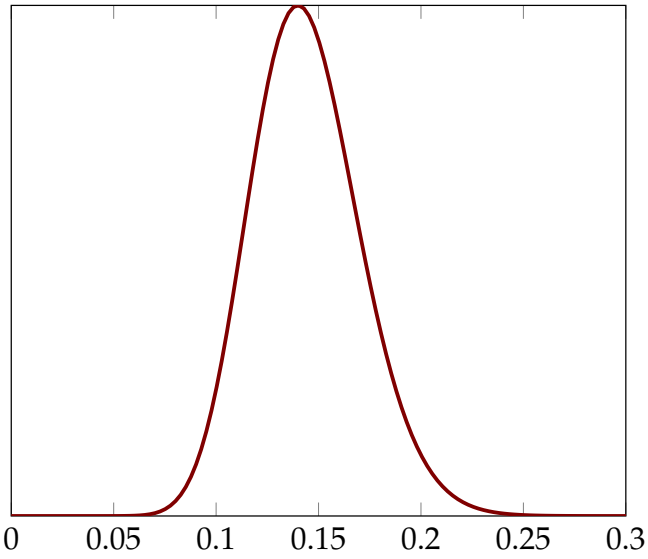
And finally plot the posterior distribution, and check that the maximizer is indeed at $\hat{p} = 0.16605$:

One last example

Let us work on one more example for continuously distributed parameters. Assume we have a population distribution with pdf $f(x) = (\theta + 1)\, x^{\theta}$, for $0 \leq x \leq 1$. Moreover, assume that θ is not totally random, but is instead distributed with pdf $f(\theta) = \frac{1}{12}(3 - \theta)$, defined over $-2 \leq \theta \leq 2$. Assume we have collected a sample of $X_1 = 0.9, X_2 = 0.89, X_3 = 0.76, X_4 = 0.96$. What is the Bayesian estimator for θ?

You may inspect the solution visually as a homework assignment. Algebraically, though, we would multiply the prior distribution





($f(\theta)$) with the likelihood function ($L(\theta)$) to obtain the posterior distribution. In mathematical terms:

$$
\begin{aligned}
f(\theta) &= \frac{1}{12}(3-\theta) \\
L(\theta) &= (\theta+1)\,X_1^\theta \cdot (\theta+1)\,X_2^\theta \cdot (\theta+1)\,X_3^\theta \cdot (\theta+1)\,X_4^\theta = \\
&= (\theta+1)^4\,(X_1 \cdot X_2 \cdot X_3 \cdot X_4)^\theta = (\theta+1)^4\,0.5844096^\theta \\
f(\theta) \cdot L(\theta) &= \frac{1}{12}(3-\theta) \cdot (\theta+1)^4\,0.5844096^\theta
\end{aligned}
$$

Getting the derivative of the posterior, and equating it to 0, we get:

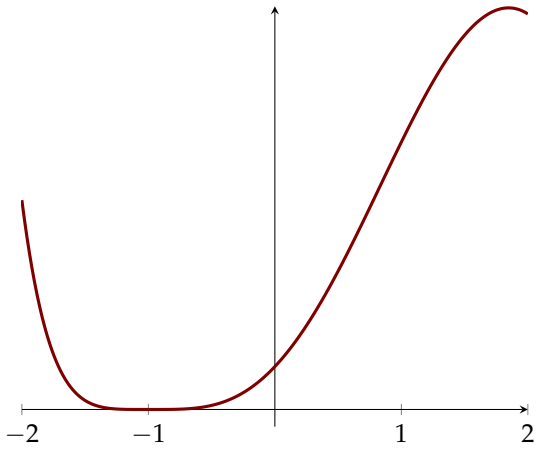
$$\frac{\partial f(\theta)L(\theta)}{\partial \theta} = 0 \implies 0.0447628 \cdot 0.58441^{\theta}(1+\theta)^3(17.4783 + \theta(-11.3083 + \theta)) = 0.$$

We get three possible solution: $\theta = -1$, $\theta = 1.85$, or $\theta = 9.46$. We note that the last one cannot happen as θ is between -2 and 2. Between the two remaining possible solutions, we compare their posterior distribution values:

- $f(-1) \cdot L(-1) = \frac{1}{12}(3 - (-1)) \cdot ((-1) + 1)^4 \, 0.5844096^{-1} = 0.$
- $f(1.85) \cdot L(1.85) = \frac{1}{12}(3 - 1.85) \cdot (1.85 + 1)^4 \, 0.5844096^{1.85} = 2.34.$

Hence, $\hat{\theta} = 1.85$ is the maximizer and the Bayesian estimator.

I lied.. Here is the visual version of the posterior also. It is clear that 1.85 is indeed the maximizer!



Part 3: Lectures 20–29

Confidence intervals for single population means

Learning objectives

After lectures 20–23, we will be able to:

- Build confidence intervals for:
  - unknown means;
  - unknown variances;
  - unknown proportions.
- Build confidence intervals for:
  - the difference between two unknown means;
  - the ratio between two unknown variances;
  - the difference between two unknown proportions.
- Understand the effect of Type I error, or probability α.
- Calculate errors and interval margins.
- Select appropriate sample sizes to keep errors below a limit.

Motivation: Point estimates lie

Assume we are told that a new smartphone has a battery of 24 hours, compared to a battery of 22 hours of the previous iteration. A person upgrades to the new phone to see that their new phone also has the same battery as the older one! Should that surprise them?

Motivation: Elections

During an election, many (*many*) polls are released to capture the momentum of the different political parties and candidates. However, surprises and upsets still happen: does that really mean that polling is off? Or should we start caring about the set of plausible outcomes rather than fixating on a single point estimate?

Quick review

In Part 2 of the class, we discussed parameter estimation. Specifically, we saw that:

- Given a population X...
- distributed with some probability density function $f(x)$...
- but with unknown parameter θ...
- we may estimate θ using an estimator $\hat{\Theta}$...
- and use a sample $X_1, X_2, \ldots, X_n$ from the original population...
- to arrive at a single conclusion: a point estimate $\hat{\theta}$!

We also discussed how "wrong" the estimator is, by calculating its bias, variance, mean square error. **But what about the probability our true parameter θ is smaller than $\hat{\theta}$** (or bigger than, or equal to)?

Motivating question 1

We are interested in estimating the unknown mean battery of a new smartphone (population X includes all new smartphones in circulation). We bought a new phone (a random sample from population X, let us call it X_1) and got that the new battery is equal to 21 hours, smaller than what our previous phone had! We are naturally disappointed, so we start asking questions about this new estimate $\hat{\mu} = 21$. What is the probability that the true battery life of smartphones in X is above 21 hours?

Motivating question 2

We are interested in estimating the unknown proportion of people in support of candidate A. We have interviewed 10 voters (randomly selected) and have found that 7 of them support candidate A. The point estimate then, based on our sample, is that $\hat{p} = 70\%$. What is:

- the probability that the true proportion of voters in support of candidate A is above 70%?
- the probability that the true proportion of voters in support of candidate A is below 70%?
- the probability that the true proportion of voters in support of candidate A is exactly equal to 70%?

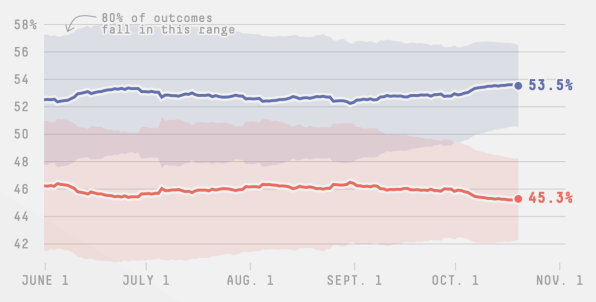
An introduction to confidence intervals

At least one of the previous questions is easy: the probability that the estimate is exactly equal to the true unknown parameter is... zero! This is because for any continuously distributed random variable, $P(X = x) = 0$... So, one thing is for certain: your estimate is not the true parameter value. This is why we introduce **confidence intervals**.

Confidence intervals appear very often in every day life. Some examples include the following.

Election time!

We may want to estimate the proportion of the population preferring one candidate over another. However simply averaging out all the polls is not a solution: we want to reveal all possible scenarios. To do show, we may produce an **interval** of all plausible scenarios and shade them. See the following image taken from 538 (fivethirtyeight.com) on October 20, 2020.



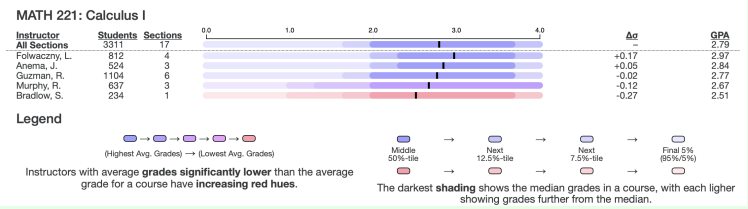
Note the small legend on the upper right corner that "warns" us: 80% of the outcomes fall in the shaded range! So there is still a 20% chance we get something *outside that range*.

Grade disparity

Many students are familiar with a tool, built by by Devin Oliver, Johnny Guo, Joe Tan, Jerry Li, Tina Abraham, Andy (Tianyue) Mao, Kara Landolt, Nathan Cho and Wade Fagen-Ulmschneider (see `https://waf.cs.illinois.edu/discovery/grade_disparity_between_subsections_at_uiuc/`) which shows the historical grade distribution for different classes at UIUC. When presenting this information, we do not only want to look at the average GPA that a class has.

Grade disparity (cont'd)

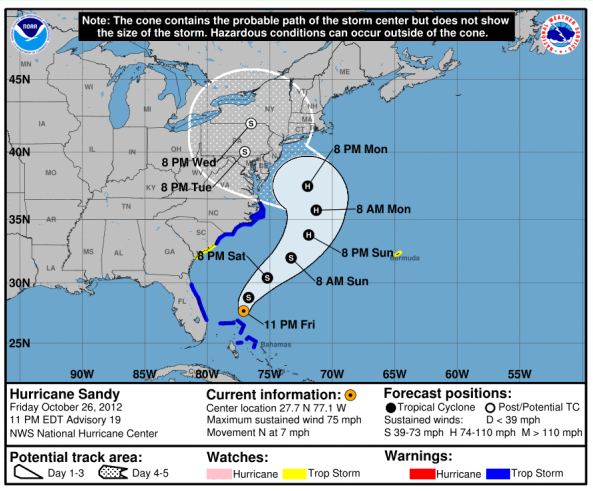
On the contrary, we are also interested in addressing the question: what is the probability a random student ends up with a grade within some range of the expected grade?



And this is clearly presented here, too! Looking at the screenshot, some classes have a very narrow range of values (meaning less deviations from the expected grade), whereas others (look at this last one above!) have a wide array of possible grades.

Cone of uncertainty

For any of us that have lived in a state that gets hit by hurricanes, we have grown used to seeing a map like this one:



This is actually from Hurricane/Superstorm Sandy (end of October 2012). When someone sees this, they may think that this reveals the areas that may be hit, or even that this is the size of the storm. But this is not the case! This "cone of uncertainty" reveals an interval of 60-70% of expected paths based on historical information!

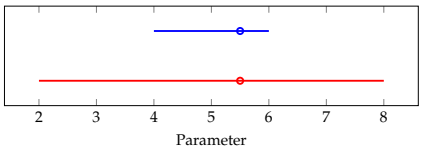
What do all of the above have in common? They are based on this understanding that the true, unknown parameters may be difficult to capture with certainty; so they resort to presenting a series of outcomes (in range form) that reveal a certain percentage of scenarios that can happen. In the election, that was 80% of the scenarios; in the grade disparity case, the first shading includes 50% of the grades historically; in the cone of uncertainty about 60-70% of historical information.

So, let us summarize really quickly before moving to the definition of confidence intervals.

- **Point** estimation: a *single* estimate with our best guess at what the unknown parameter is.
- **Interval** estimation: an interval of values where the unknown parameter is believed to belong in.

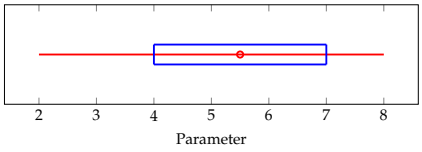
What are the advantages of interval estimation?

1. A point estimation reveals a single point of information about the whereabouts of the unknown parameter, leaving us with no idea of how close the actual parameter is expected to be:



For example, in the above figure the red parameter can be any value between 2 and 8, but our estimate places it at 5.5. On the other hand, the blue one is also estimated at 5.5; however it can only take values within 4 to 6. An interval estimate would reveal more information about these ranges.

2. An interval estimation reveals a margin of error as a measure of accuracy for our parameter.



In this example, we still estimate the red parameter to be 5.5; but now we are also told that we would not be surprised to see it be in any point between 4 and 7.

Now, a **confidence interval**, usually presented in the form of $[L, U]$, contains the most "believable" values for the estimated parameter. Every confidence interval is associated with a **confidence level**, which represents the probability that the true parameter value falls in that interval. In mathematical terms:

$$P(L \leq \theta \leq U) = 1 - \alpha.$$

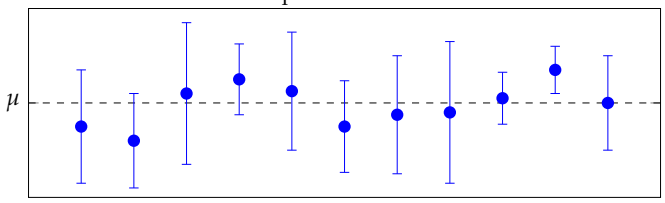
Combining, we write that $[L, U]$ is a $100 \cdot (1 - \alpha)\,\%$ confidence interval for parameter θ.

A couple of notes about confidence intervals and how they came to be. First of all, both L and U are obtained from the random sample we selected. That is, they depend on the sample selected. Secondly, α is an external parameter and we can make it as small or as big as we want to. Smaller α values lead to higher confidence for our interval estimates and vice versa. Typical value for α is 5%, which leads to the creation of 95% confidence intervals. Finally, we may also represent confidence intervals as

$$\boxed{\text{Point estimate} \pm \text{Margin}}$$

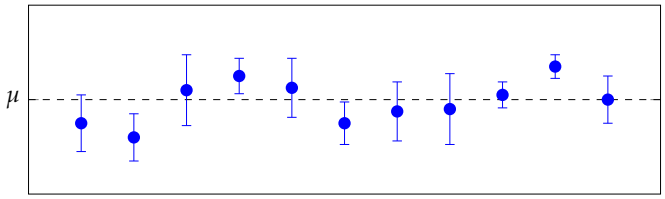
Visually, this is naturally showcased in a plot with the point estimate in the center and the margins on each side as whiskers. As an example we provide Figures 66 and 67.

Figure 66: Here we run 10 experiments and provide the estimate obtained from each of them with a blue dot. We then build a **95% confidence interval** around the estimate; this is shown with the whiskers of the plot.



In both figures we sometimes underestimate and sometimes overestimate the parameter. When we extend our estimate to include a range of "believable" values, we see that the number of experiments that include the true parameter changes sharply. Observe how almost all intervals contain the true mean in the first example; and about half of them do in the second example. As an example take the second experiment: we are underestimating the true value of the parameter through our estimation process. However, in the first figure the 95% confidence interval includes the true parameter μ. In the second figure, it does not.

Figure 67: Here we run 10 experiments and provide the estimate obtained from each of them with a blue dot (note that the estimates are the same as in Figure 66). We then build a **50% confidence interval** around the estimate; this is shown with the whiskers of the plot.



The confidence interval then reveals interesting properties. If we build a 95% confidence interval around an unknown parameter, then this means that:

1. we are 95% certain the parameter is in that range.
2. if we obtain 100 samples, 95 of them will have a parameter in that range.
3. there is a 5% chance we are wrong and the parameter is outside that range (either higher or lower).

Sampling distributions

As a reminder, when picking a sample out of a population, then we say that the sample is distributed with some sampling distribution. For convenience, let us focus on the case of trying to estimate the unknown mean μ of a population X: the population is distributed with some distribution and mean μ (unknown) and variance σ^2 (possibly known).

To estimate μ, we resort to collecting a sample $X_1, X_2, \ldots, X_n$ and calculate the sample average $\overline{X} = \frac{X_1+X_2+\ldots+X_n}{n}$. Recall that the sample average **is a random variable**, distributed with some sampling distribution with:

- expectation $E\left[\overline{X}\right] = \mu$.
- variance $Var\left[\overline{X}\right] = \frac{\sigma^2}{n}$.

Let's distinguish between two cases:

1. X is normally distributed.
2. X is not normally distributed.

If X is normally distributed, the $\overline{X}$ is **also normally distributed** with mean μ and variance σ^2/n. On the other hand, if X is not normally distributed, then $\overline{X}$ is **normally distributed only if the sample size is large enough** (due to the central limit theorem).

Let us assume that one of the above two conditions hold. Then:

- $P(\overline{X} = \mu) = 0$ – $\overline{X}$ is a random variable and the probability it is exactly equal to some other value is zero.
- $P(\overline{X} \geq \mu) = 0.5$ – due to the symmetry of the normal distribution.
- $P(\overline{X} \leq \mu) = 0.5$ – due to the symmetry of the normal distribution.

Recall that in some of our previous worksheets, we had already identified that for a normally distributed random variable, we can calculate the probability of a range of values around the mean as: [74]

[74] See Worksheet 8, Questions 7-8-9.

$$P(\mu - r \leq \overline{X} \leq \mu + r) = 2\Phi\left(\frac{r}{\sigma}\right) - 1$$

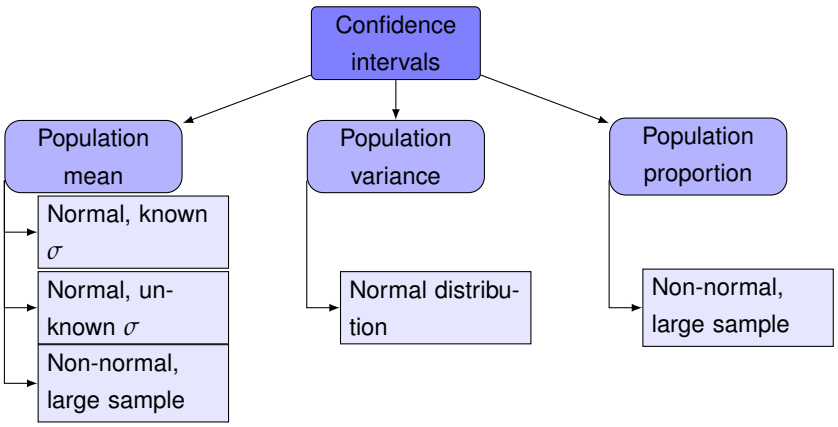
Say, we were looking to build a 95% confidence interval, that would translate to:

$$P(\mu - r \leq \overline{X} \leq \mu + r) = 2\Phi\left(\frac{r}{\sigma}\right) - 1 = 0.95 \implies$$
$$\implies 2\Phi\left(\frac{r}{\sigma}\right) = 1.95 \implies \Phi\left(\frac{r}{\sigma}\right) = 0.975.$$

Hence, $\frac{r}{\sigma}$ has to be the value that leads to 0.975... Let's keep that in the back of our minds for now.

Single population confidence intervals

Before we proceed to the next calculations, we provide an overview of where we are headed at:

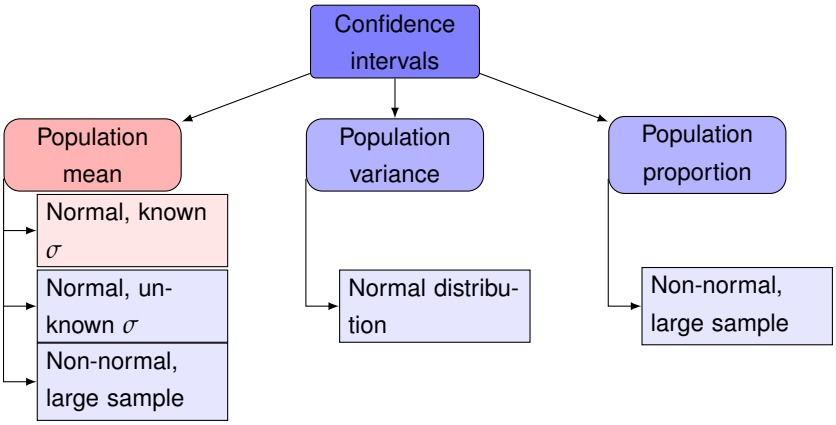


In all of them we assume the existence of one population with some unknown parameter (the mean, the variance, a proportion). To

estimate the unknown parameter, we collect a sample, estimate the unknown parameter based on it and we create an interval around it. We begin with the simplest case: the mean.

Population mean confidence intervals

Assume X is a population with unknown mean. We have collected a sample $X_1, X_2, \ldots, X_n$ to estimate the mean. As we have discussed in previous classes, the sample average $\overline{X}$ is an unbiased estimator for the unknown mean. But what should be the interval around it?



Let us assume that we know X to be normally distributed. And while we are missing the true population mean μ we know its variance σ^2 (or its standard deviation σ).

Recall that we are looking for L, U such that $P(L \leq \overline{X} \leq U) = 1 - \alpha$. As $\overline{X} \sim \mathcal{N}\left(\mu, \sigma^2/n\right)$, we have that $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$. Then, we may write $P(L \leq \overline{X} \leq U) = 1 - \alpha$ as $P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2} = 1 - \alpha$, where $z_{\alpha/2}$ is called the **critical z value** and is found as $P(Z > z_{\alpha/2}) = \frac{\alpha}{2}$. We show what these critical values represent in visual format in Figures 68–71.

Finding critical values

Some common critical values:

- $\alpha = 10\% \implies z_{0.05} = 1.645$ as $\Phi(1.645) = 95\% = 1 - \alpha/2$.
- $\alpha = 5\% \implies z_{0.025} = 1.96$ as $\Phi(1.96) = 97.5\% = 1 - \alpha/2$.
- $\alpha = 1\% \implies z_{0.005} = 2.576$ as $\Phi(2.576) = 99.5\% = 1 - \alpha/2$.

Try it yourselves! What is $z_{\alpha/2}$ for:

- $\alpha = 20\% \implies z_{0.1} =$
- $\alpha = 2\% \implies z_{0.01} =$
- $\alpha = 0.1\% \implies z_{0.0005} =$

Figure 68: $\alpha = 1\%$.

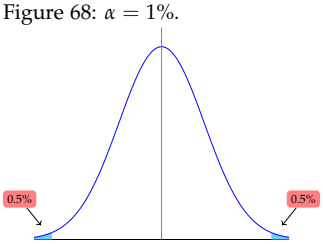


Figure 69: $\alpha = 5\%$.

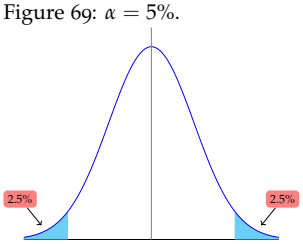


Figure 70: $\alpha = 10\%$.

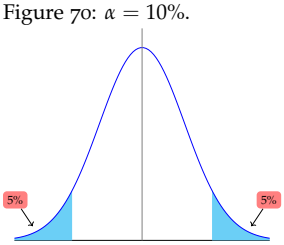
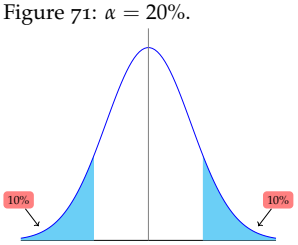


Figure 71: $\alpha = 20\%$.



Based on this discussion, and based on the symmetry of the normal distribution, we have our first confidence interval:

$$\boxed{P(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha}$$

And consequently, we have a lower bound for our interval at $L = \overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ and an upper bound at $U = \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

Our first confidence interval

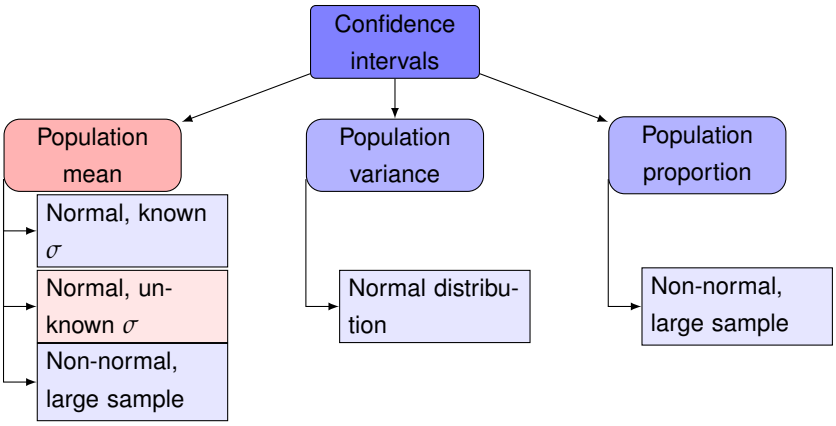
A class at UIUC gives out grades that are normally distributed with known variance equal to 100 (i.e., $\sigma = 10$). Build a 95% confidence interval for the mean of the class grades, assuming that in the previous 8 semesters, the average has been a 77.

First, a 95% confidence interval implies that $\alpha = 0.05$. Hence, we are looking at $z_{\alpha/2} = z_{0.025} = 1.96$. Then our interval will be

$$[L, U] = \left[77 - 1.96 \cdot 10/\sqrt{8}, 77 + 1.96 \cdot 10/\sqrt{8}\right] = [70.07, 83.93].$$

Let us move on to the second part of our discussion about means. What if we know that X is normally distributed but we have no idea

what the variance or standard deviation is?

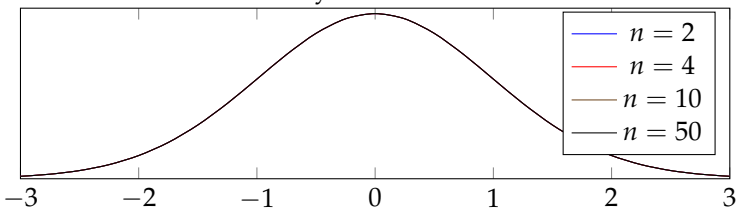


Since we do not know σ, we need to estimate it. And what better way to estimate σ other than using the sample standard deviation s! What is the big deal? Can't we just do everything we did earlier, and simply use s instead of σ?

The short answer is **no**. Unfortunately the statistic $\frac{\overline{X}-\mu}{s/\sqrt{n}}$ which was earlier normally distributed because we knew σ is not any more. Replacing σ with s leads to the statistic to be distributed with the so-called **Student's T distribution**. More specifically, we write that $T = \frac{\overline{X}-\mu}{s/\sqrt{n}}$ is distributed following a Student's T distribution with $n-1$ degrees of freedom.

What kind of name is that? The distribution was introduced by W.S. Gosset, who published his findings under the fake name "Student". This happened because the Guinness brewery (where he was employed at that time) did not allow its employees to publish their findings.

The distribution looks eerily similar to the normal distribution:



It is symmetric, but it has thicker tails. As the degrees of freedom increase, then it starts looking more and more like the actual normal distribution. Observe how for large values of n (say, $n \to \infty$) the z and the t values are identical!

Finally, just like the normal distribution, we may calculate any value we are interested in by looking up a table of values. The table is offered in the last page of the notes for convenience.

Similarly to before then, after replacing z (normal distribution) with t (Student's T distribution) values and replacing σ (known standard deviation) with s (sample standard deviation), we get:

$$\boxed{P(\overline{X} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}) = 1 - \alpha}$$

And consequently, we have a lower bound for our interval at $L = \overline{X} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$ and an upper bound at $U = \overline{X} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$.

Finding critical values for the T distribution

Some critical values for the T distribution:

- $t_{0.025,15} = 2.131$
- $t_{0.05,10} = 1.812$
- $t_{0.05,25} = 1.708$
- $t_{0.10,5} = 1.476$

STUDENT'S t CRITICAL VALUES

ν	0.4	0.33	0.25	0.2	0.125	0.1	0.05	0.025	0.01	0.005	0.001
1	0.325	0.577	1.000	1.376	2.414	3.078	6.314	12.706	31.821	63.657	318.31
5	0.267	0.457	0.727	0.920	1.301	1.476	2.015	2.571	3.365	4.032	5.893
10	0.260	0.444	0.700	0.879	1.221	1.372	1.812	2.228	2.764	3.169	4.144
15	0.258	0.439	0.691	0.866	1.197	1.341	1.753	2.131	2.602	2.947	3.733
25	0.256	0.436	0.684	0.856	1.178	1.316	1.708	2.060	2.485	2.787	3.450
∞	0.253	0.431	0.674	0.842	1.150	1.282	1.645	1.960	2.326	2.576	3.090

Another confidence interval

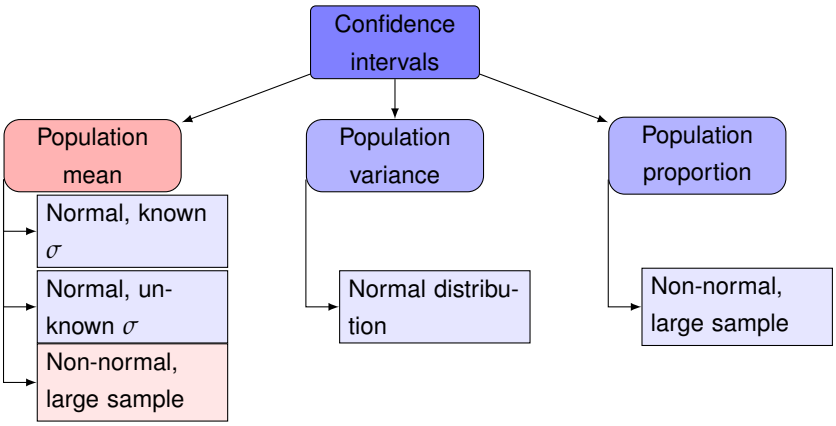
The same class at UIUC gives out grades that are normally distributed with unknown variance; we have observe that over the last 8 semesters the sample variance was equal to 100 (i.e., $s = 10$). Build a 95% confidence interval for the mean of the class grades, assuming that in the previous 8 semesters, the average has been a 77.

We still use $\alpha = 0.05$. Now, though, we are looking at $t_{\alpha/2,7} = t_{0.025,7} = 2.365$. Finally, our interval will be

$$[L, U] = \left[77 - 2.365 \cdot 10/\sqrt{8}, 77 + 2.365 \cdot 10/\sqrt{8}\right] = [68.64, 85.36].$$

Note how the confidence interval has been extended, due to the fact that we do not know what σ is.

For the last case, we will be assuming a general distribution (not necessarily normal) with known or unknown σ: however we also assume the existence of a large enough sample (say, $n \geq 30$).



This case is very similar to the first one. If the sample is big enough, then the **central limit theorem** applies and the average $\overline{X}$ is still normally distributed. If we know σ, we may use it in our calculations; if we do not, then we replace it with the sample standard deviation s. All in all:

$$\boxed{P(\overline{X} - z_{\alpha/2}\frac{s}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\alpha/2}\frac{s}{\sqrt{n}}) = 1 - \alpha}$$

Consequently, the confidence interval is given as

$$\left[\overline{X} - z_{\alpha/2}\frac{s}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{s}{\sqrt{n}}\right].$$

Extensions In all of our previous discussion, we assumed **two-sided** confidence intervals. However, in some instance we only care about the one side. Consider the following examples:

- contamination levels (only interested if they are too high);
- grades (only interested if they are too low);
- cholesterol levels (only interested if they are too high);
- and others.

In these cases, we want to build **one-sided** confidence intervals that look like

$$[L, +\infty) \quad \text{or} \quad (-\infty, U].$$

What is the repercussion of having one side rather than two sides? Recall that we choose a level of confidence $1 - \alpha$. With two sides,

Figure 72: $\alpha = 5\%$ (upper bound only).

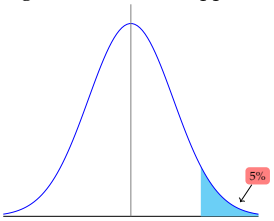


Figure 73: $\alpha = 5\%$ (lower bound only).

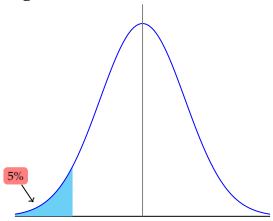


Figure 74: $\alpha = 10\%$ (upper bound only).

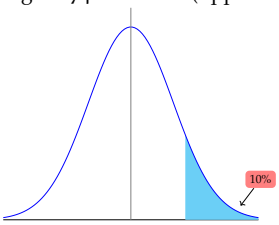
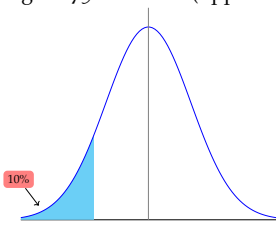


Figure 75: $\alpha = 10\%$ (upper bound only).



this was divided evenly on both sides! Now, all of α gets on one side. Visually, we have the situation of Figures 72–75.

This, in essence, is all that changes: instead of $z_{\alpha/2}$ or $t_{\alpha/2,n-1}$ use z_α or $t_{\alpha,n-1}$ and only calculate a lower or an upper bound as needed.

Cholesterol testing

When testing for cholesterol the sample that a patient has given is divided into 5 parts, each of which is tested individually: assume that each individual test has known $\sigma = 8$. The average measurement was 194; the upper limit out of which the patient may need to start being careful is 200. What is the two-sided 95% confidence interval? What is the one-sided upper 95% confidence interval?
We know the standard deviation, so we are in the first of the three cases we discussed. Hence, the average measurement $\overline{X}$ is normally distributed.

- Two-sided: $z_{\alpha/2} = z_{0.025} = 1.96$.

$$[L, U] = \left[194 - 1.96 \cdot 8/\sqrt{5}, 194 + 1.96 \cdot 8/\sqrt{5}\right] = [186.99, 201.01].$$

- One-sided: $z_\alpha = z_{0.05} = 1.645$.

$$(-\infty, U] = \left(-\infty, 194 + 1.645 \cdot 8/\sqrt{5}\right] = (-\infty, 199.89].$$

Let's see the implication of this. The doctor cannot be 95% certain that your cholesterol level is below 200 units if they want to give you a two-sided interval. They can be 95% certain though if they do a one-sided interval!

Another very interesting topic has to do with the question: **how big a sample guarantees me a small error**? The question has two components to address. What do we define as error? And what do we define as small error?

The **estimation error** is the absolute difference between the measured and the true value:

$$E = \left|\overline{X} - \mu\right| \leq z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

The **precision error** is the width of the confidence interval:

$$2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Increasing n will have a positive effect on both errors: they go down when we collect more samples. But, of course, the natural question is *how many samples are enough*? Enough for what? This leads us to the following result.

For sample size $n = \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2$, the estimation error is at most E.

For a one-sided interval, the estimation error is $z_\alpha \frac{\sigma}{\sqrt{n}}$ and we need sample size $n = \left(\frac{z_\alpha \sigma}{E}\right)^2$ for the error to be at most E.

As we are discussing the number of samples to obtain, we always round up the number we get if it is fractional.

Cholesterol testing

The doctor from earlier would like to get as many samples as necessary to be sure that your true cholesterol levels are within a estimation error of 7 units. How many samples should they take for a 95% two-sided and a 95% upper-sided (one-sided) confidence interval? Recall that we know $\sigma = 8$.

- two-sided: $z_{\alpha/2} = z_{0.025} = 1.96$:

$$n = \left(\frac{1.96 \cdot 8}{7}\right)^2 = 5.0176 \to 6.$$

- one-sided: $z_\alpha = z_{0.05} = 1.645$:

$$n = \left(\frac{1.645 \cdot 8}{7}\right)^2 = 3.5344 \to 4.$$

NORMAL CUMULATIVE DISTRIBUTION FUNCTION ($\Phi(z)$)

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

STUDENT'S t CRITICAL VALUES

ν	0.4	0.33	0.25	0.2	0.125	0.1	0.05	0.025	0.01	0.005	0.001
1	0.325	0.577	1.000	1.376	2.414	3.078	6.314	12.706	31.821	63.657	318.31
2	0.289	0.500	0.816	1.061	1.604	1.886	2.920	4.303	6.965	9.925	22.327
3	0.277	0.476	0.765	0.978	1.423	1.638	2.353	3.182	4.541	5.841	10.215
4	0.271	0.464	0.741	0.941	1.344	1.533	2.132	2.776	3.747	4.604	7.173
5	0.267	0.457	0.727	0.920	1.301	1.476	2.015	2.571	3.365	4.032	5.893
6	0.265	0.453	0.718	0.906	1.273	1.440	1.943	2.447	3.143	3.707	5.208
7	0.263	0.449	0.711	0.896	1.254	1.415	1.895	2.365	2.998	3.499	4.785
8	0.262	0.447	0.706	0.889	1.240	1.397	1.860	2.306	2.896	3.355	4.501
9	0.261	0.445	0.703	0.883	1.230	1.383	1.833	2.262	2.821	3.250	4.297
10	0.260	0.444	0.700	0.879	1.221	1.372	1.812	2.228	2.764	3.169	4.144
11	0.260	0.443	0.697	0.876	1.214	1.363	1.796	2.201	2.718	3.106	4.025
12	0.259	0.442	0.695	0.873	1.209	1.356	1.782	2.179	2.681	3.055	3.930
13	0.259	0.441	0.694	0.870	1.204	1.350	1.771	2.160	2.650	3.012	3.852
14	0.258	0.440	0.692	0.868	1.200	1.345	1.761	2.145	2.624	2.977	3.787
15	0.258	0.439	0.691	0.866	1.197	1.341	1.753	2.131	2.602	2.947	3.733
16	0.258	0.439	0.690	0.865	1.194	1.337	1.746	2.120	2.583	2.921	3.686
17	0.257	0.438	0.689	0.863	1.191	1.333	1.740	2.110	2.567	2.898	3.646
18	0.257	0.438	0.688	0.862	1.189	1.330	1.734	2.101	2.552	2.878	3.610
19	0.257	0.438	0.688	0.861	1.187	1.328	1.729	2.093	2.539	2.861	3.579
20	0.257	0.437	0.687	0.860	1.185	1.325	1.725	2.086	2.528	2.845	3.552
21	0.257	0.437	0.686	0.859	1.183	1.323	1.721	2.080	2.518	2.831	3.527
22	0.256	0.437	0.686	0.858	1.182	1.321	1.717	2.074	2.508	2.819	3.505
23	0.256	0.436	0.685	0.858	1.180	1.319	1.714	2.069	2.500	2.807	3.485
24	0.256	0.436	0.685	0.857	1.179	1.318	1.711	2.064	2.492	2.797	3.467
25	0.256	0.436	0.684	0.856	1.178	1.316	1.708	2.060	2.485	2.787	3.450
26	0.256	0.436	0.684	0.856	1.177	1.315	1.706	2.056	2.479	2.779	3.435
27	0.256	0.435	0.684	0.855	1.176	1.314	1.703	2.052	2.473	2.771	3.421
28	0.256	0.435	0.683	0.855	1.175	1.313	1.701	2.048	2.467	2.763	3.408
29	0.256	0.435	0.683	0.854	1.174	1.311	1.699	2.045	2.462	2.756	3.396
30	0.256	0.435	0.683	0.854	1.173	1.310	1.697	2.042	2.457	2.750	3.385
40	0.255	0.434	0.681	0.851	1.167	1.303	1.684	2.021	2.423	2.704	3.307
50	0.255	0.433	0.679	0.849	1.164	1.299	1.676	2.009	2.403	2.678	3.261
60	0.254	0.433	0.679	0.848	1.162	1.296	1.671	2.000	2.390	2.660	3.232
∞	0.253	0.431	0.674	0.842	1.150	1.282	1.645	1.960	2.326	2.576	3.090

Confidence intervals for variances and proportions

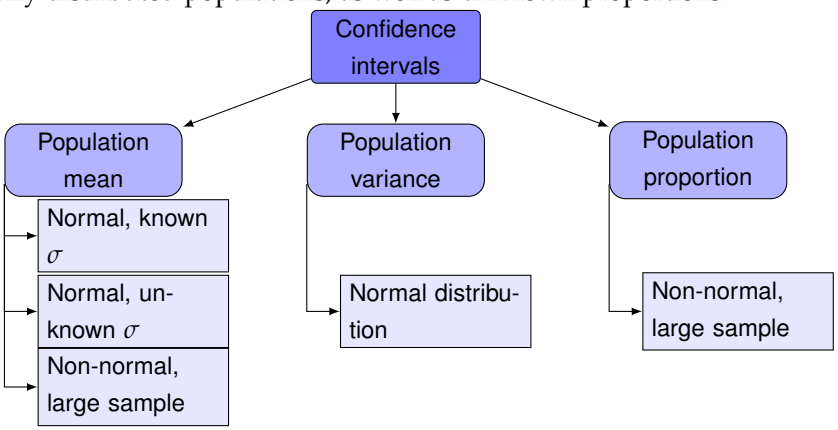
Learning objectives

After lectures 20–23, we will be able to:

- Build confidence intervals for:
  - unknown means;
  - unknown variances;
  - unknown proportions.
- Build confidence intervals for:
  - the difference between two unknown means;
  - the ratio between two unknown variances;
  - the difference between two unknown proportions.
- Understand the effect of Type I error, or probability α.
- Calculate errors and interval margins.
- Select appropriate sample sizes to keep errors below a limit.

Single population confidence intervals

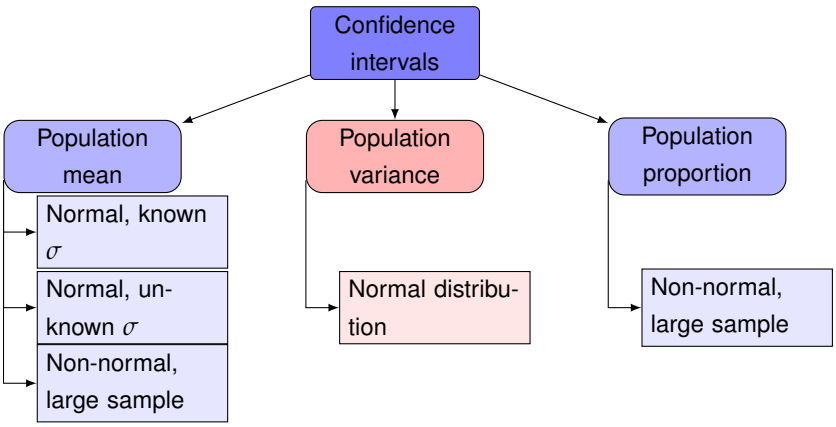
Continuing from last lecture, we are still building confidence intervals for a single population. In this lecture, though, we will talk about creating confidence intervals for unknown variances of normally distributed populations, as well as unknown proportions.



Population variance confidence intervals

Assume X is a normally distributed population with unknown variance. We have collected a sample $X_1, X_2, \ldots, X_n$ to estimate the variance. As we have discussed in previous classes, the sample variance

s^2 is an unbiased estimator for the unknown variance. But what should be the interval around it? This is our focus:



Once again, we are looking for L, U such that $P(L \leq s^2 \leq U) = 1 - \alpha$. However, we first need to discuss what s^2 is distributed as.

Sampling distribution for σ^2

Recall that we have a good estimator for the population variance σ^2:

- pick a sample $X_1, X_2, \ldots, X_n$.
- estimate the variance by the sample variance: s^2.

We have already proven that $E\left[s^2\right] = \sigma^2$. The question now is: what is the sampling distribution of s^2? It turns out it follows the χ^2 **distribution**. [75] The distribution is formally defined as follows:

[75] Pronounced "Chi-Squared".

> Let $X_1, X_2, \ldots, X_n$ be a sample from a normally distributed population with $\mathcal{N}\left(\mu, \sigma^2\right)$. Then, the random variable
>
> $$X^2 = \frac{(n-1)\, s^2}{\sigma^2}$$
>
> is distributed with a χ^2-distribution with $n - 1$ degrees of freedom.

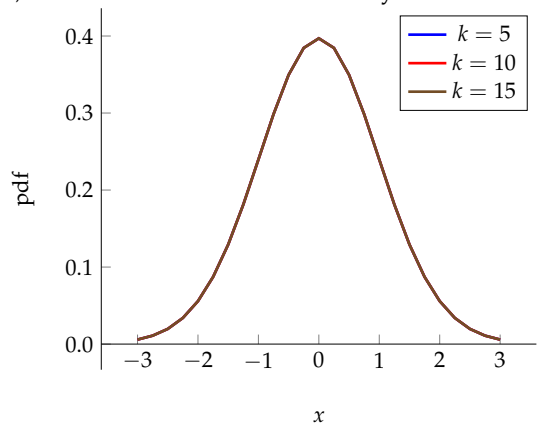
It is the sum of the squares of $n - 1$ normally distributed random variables. For a visual representation see Figure 76.

Very similarly to our previous operations for other confidence intervals, we again focus on identifying critical values for the χ^2-distribution, that is values such that:

$$P(X^2 \geq \chi^2_{\alpha,k}) = \alpha.$$

Luckily, we again may use a table containing these values, referred to as (you guessed it) a χ^2-table.

Figure 76: Here we present the χ^2 distribution for three different degrees of freedom equal to $k = 5, 10, 15$. Note how the distribution is **not symmetric**.



Practice with the χ^2 distribution

For example, let us practice with some values:

- $\chi^2_{0.05,5} = 11.07$
- $\chi^2_{0.1,5} = 9.236$
- $\chi^2_{0.9,20} = 12.443$
- $\chi^2_{0.95,55} = 38.958$

Here are some values taken from the tables in the last two pages. These should help with finding the above critical values. Again, we look at the rows for the degrees of freedom, and at the columns for the percentages.

ν	99%	97.5%	95%	90%	10%	5%	2.5%	1%
1	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635
5	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
55	33.570	36.398	38.958	42.060	68.796	73.311	77.380	82.292

Once more, assume we have a sample $X_1, X_2, \ldots, X_n$. Then:

$$X^2 = \frac{(n-1)\,s^2}{\sigma^2} \sim \chi^2_{n-1}$$

and hence:

$$P\left(\chi^2_{1-\alpha/2,n-1} \leq X^2 \leq \chi^2_{\alpha/2,n-1}\right) = 1 - \alpha.$$

By converting back to the σ^2 space, we get:

$$P\left(\frac{(n-1)\,s^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)\,s^2}{\chi^2_{1-\alpha/2,n-1}}\right),$$

where the two bounds are (in $[L, U]$ form):

$$L = \frac{(n-1)\,s^2}{\chi^2_{\alpha/2,n-1}}$$
$$U = \frac{(n-1)\,s^2}{\chi^2_{1-\alpha/2,n-1}}$$

A couple of notes of caution for when building a variance confidence interval:

1. There are no *actual* squares involved! You do not "square" the value: this is simply the name of the distribution!
2. Notice that the critical values are not symmetric: in the normal and the t distribution, the values are symmetric.
   - For the lower bound, use $\chi^2_{\alpha/2,n-1}$;
   - For the upper bound, use $\chi^2_{1-\alpha/2,n-1}$.
3. Because of the lack of symmetry in the critical values, there is no symmetry in the bounds.
   - On top of that, you are dividing the estimator by a value (rather than adding it and subtracting it to the estimator, which was the case earlier).

Our first variance confidence interval

An engineer is concerned about soil contamination, which is assumed to be normally distributed. They pick 15 soil samples and measure the contaminant levels finding that $\overline{X} = 13.7$ ppm and $s = 3.15$ ppm. You may assume that the soil contamination level has unknown mean and variance. What is:

1. a 95% confidence interval for μ?
2. a 95% confidence interval for σ^2?

A mean confidence interval first

Wait! The first part is for a mean confidence interval. Let us do a quick activity then to find it. We have:

1. normally distributed population;
2. unknown variance.

Hence, we need values from the t-table. More specifically, we need:

- $t_{0.025,14} = 2.145$ to build the mean confidence interval.

This leads to an interval that:

$$\mu \in \left[13.7 - 2.145 \cdot \frac{3.15}{\sqrt{15}}, 13.7 + 2.145 \cdot \frac{3.15}{\sqrt{15}}\right] = [11.96, 15.44]\,.$$

And a variance confidence interval next

For the variance confidence interval, we look at the χ^2 table (look at the last two pages of this set of nodes) to find the two values we need:

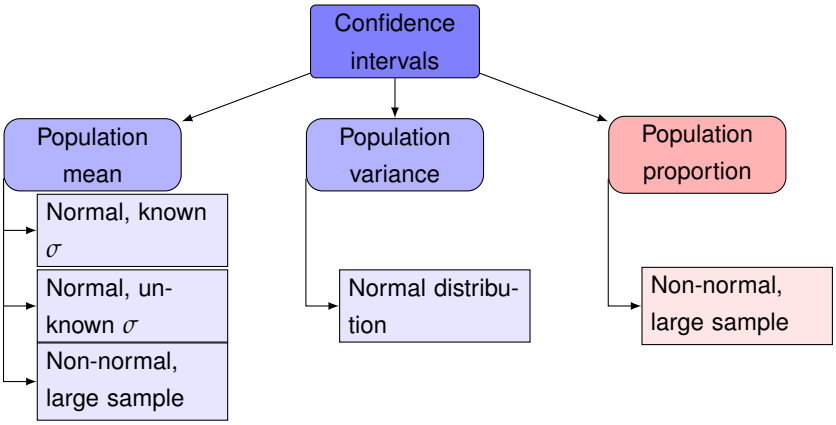
- $\chi^2_{0.025,14} = 26.119, \quad \chi^2_{0.975,14} = 5.629.$

The interval then is found as:

$$\sigma^2 \in \left[\frac{14 \cdot 3.15^2}{26.119}, \frac{14 \cdot 3.15^2}{5.629}\right] = [5.32, 24.68]$$

Note how it is not at all symmetric!

Population proportion confidence intervals



Let us see the last case now. We begin with a motivational example.

Policy making

Assume we are deciding for a new law, and want to make sure that the population of a city (estimated at 100,000) supports it. Moreover, assume that support means 50% or more people like the law.

What can we do?

- Ask a random set of n people whether they support the law.
- Count how many support the law. Let them be X.
- Estimate $\hat{p} = \frac{X}{n}$.

Suppose $\hat{p} = 0.6$ after asking $n = 30$ people.
Should we enact the law? *Are we 95% sure the majority supports it?*

In the previous example, we have that $X \sim \text{binomial}(n, p)$. When n is big enough, then X is approximated by a normal distribution with mean np and variance $np(1-p)$. [76] Let us state this more formally.

[76] Why is that?

Definition 61 (Normal approximation to the binomial distribution) *Assume that X is binomially distributed with parameters n, p. Further assume that $np > 5$ and $n(1-p) > 5$. Then, X can be written as a normally distributed random variable $\mathcal{N}(np, np(1-p))$.*

Because of that, the statistic $Z = \frac{X-np}{\sqrt{np(1-p)}}$ follows the standard normal distribution (i.e., $\mathcal{N}(0,1)$). Note how we can rewrite Z as follows:

$$Z = \frac{X - np}{\sqrt{np(1-p)}} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim \mathcal{N}(0,1).$$

Now, let us derive the confidence intervals. Let $\hat{p}$ be the proportion of observations that are of interest (for example, the number of people who agree with a statement versus the total number n of people asked). Then:

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Policy making

We asked 30 people and 18 said they support the law. What is the 95%-confidence interval for the true proportion supporting the law in the city?

$$0.6 - 1.96 \cdot \sqrt{\frac{0.6 \cdot 0.4}{30}} \leq p \leq 0.6 + 1.96 \cdot \sqrt{\frac{0.6 \cdot 0.4}{30}} \implies$$
$$0.4247 \leq p \leq 0.7753.$$

Bounding the error

The **estimation error** for our point estimate $\hat{p}$ is

$$E = |\hat{p} - p|.$$

Assume we are asked to calculate a $100 \cdot (1-\alpha)\%$ confidence interval. Then, its error is bounded above by:

$$E \leq z_{\alpha/2} \sqrt{p(1-p)/n}.$$

Expectedly, as n increases, the error bound goes down. But the real question is: **how big should n be for the error to be at a pre-specified level?** We may calculate this as:

$$n \geq \left(\frac{z_{\alpha/2}}{E}\right)^2 p(1-p).$$

However... the true proportion p is unknown – but we can show that $p(1-p) \leq 0.25$ [77]. Hence, we use just that to finally get that:

$$\boxed{n \geq 0.25 \left(\frac{z_{\alpha/2}}{E}\right)^2.}$$

[77] This is the maximum value for $p \cdot (1-p)$ for any value of p $[0,1]$.

Policy making

In the previous example, we want to have a 95%-confidence interval with an error of at most $E = 5\%$. How many people should we ask?

95%-confidence level $\implies z_{0.025} = 1.96$. Hence, we get:

$$n \geq 0.25 \cdot \left(\frac{1.96}{0.05}\right)^2 = 384.16 \implies n = 385.$$

We should ask at least 385 people.

Observe that the number does **not** depend on the specific population, but *only* on the confidence level and the pre-specified error.

ν	99.9%	99.5%	99.0%	97.5%	95.0%	90.0%	87.5%	80.0%	75.0%	66.7%	50.0%
1	0.000	0.000	0.000	0.001	0.004	0.016	0.025	0.064	0.102	0.186	0.455
2	0.002	0.010	0.020	0.051	0.103	0.211	0.267	0.446	0.575	0.811	1.386
3	0.024	0.072	0.115	0.216	0.352	0.584	0.692	1.005	1.213	1.568	2.366
4	0.091	0.207	0.297	0.484	0.711	1.064	1.219	1.649	1.923	2.378	3.357
5	0.210	0.412	0.554	0.831	1.145	1.610	1.808	2.343	2.675	3.216	4.351
6	0.381	0.676	0.872	1.237	1.635	2.204	2.441	3.070	3.455	4.074	5.348
7	0.598	0.989	1.239	1.690	2.167	2.833	3.106	3.822	4.255	4.945	6.346
8	0.857	1.344	1.646	2.180	2.733	3.490	3.797	4.594	5.071	5.826	7.344
9	1.152	1.735	2.088	2.700	3.325	4.168	4.507	5.380	5.899	6.716	8.343
10	1.479	2.156	2.558	3.247	3.940	4.865	5.234	6.179	6.737	7.612	9.342
11	1.834	2.603	3.053	3.816	4.575	5.578	5.975	6.989	7.584	8.514	10.341
12	2.214	3.074	3.571	4.404	5.226	6.304	6.729	7.807	8.438	9.420	11.340
13	2.617	3.565	4.107	5.009	5.892	7.042	7.493	8.634	9.299	10.331	12.340
14	3.041	4.075	4.660	5.629	6.571	7.790	8.266	9.467	10.165	11.245	13.339
15	3.483	4.601	5.229	6.262	7.261	8.547	9.048	10.307	11.037	12.163	14.339
16	3.942	5.142	5.812	6.908	7.962	9.312	9.837	11.152	11.912	13.083	15.338
17	4.416	5.697	6.408	7.564	8.672	10.085	10.633	12.002	12.792	14.006	16.338
18	4.905	6.265	7.015	8.231	9.390	10.865	11.435	12.857	13.675	14.931	17.338
19	5.407	6.844	7.633	8.907	10.117	11.651	12.242	13.716	14.562	15.859	18.338
20	5.921	7.434	8.260	9.591	10.851	12.443	13.055	14.578	15.452	16.788	19.337
21	6.447	8.034	8.897	10.283	11.591	13.240	13.873	15.445	16.344	17.720	20.337
22	6.983	8.643	9.542	10.982	12.338	14.041	14.695	16.314	17.240	18.653	21.337
23	7.529	9.260	10.196	11.689	13.091	14.848	15.521	17.187	18.137	19.587	22.337
24	8.085	9.886	10.856	12.401	13.848	15.659	16.351	18.062	19.037	20.523	23.337
25	8.649	10.520	11.524	13.120	14.611	16.473	17.184	18.940	19.939	21.461	24.337
26	9.222	11.160	12.198	13.844	15.379	17.292	18.021	19.820	20.843	22.399	25.336
27	9.803	11.808	12.879	14.573	16.151	18.114	18.861	20.703	21.749	23.339	26.336
28	10.391	12.461	13.565	15.308	16.928	18.939	19.704	21.588	22.657	24.280	27.336
29	10.986	13.121	14.256	16.047	17.708	19.768	20.550	22.475	23.567	25.222	28.336
30	11.588	13.787	14.953	16.791	18.493	20.599	21.399	23.364	24.478	26.165	29.336
35	14.688	17.192	18.509	20.569	22.465	24.797	25.678	27.836	29.054	30.894	34.336
40	17.916	20.707	22.164	24.433	26.509	29.051	30.008	32.345	33.660	35.643	39.335
45	21.251	24.311	25.901	28.366	30.612	33.350	34.379	36.884	38.291	40.407	44.335
50	24.674	27.991	29.707	32.357	34.764	37.689	38.785	41.449	42.942	45.184	49.335
55	28.173	31.735	33.570	36.398	38.958	42.060	43.220	46.036	47.610	49.972	54.335
60	31.738	35.534	37.485	40.482	43.188	46.459	47.680	50.641	52.294	54.770	59.335

ν	40.0%	33.3%	25.0%	20.0%	12.5%	10.0%	5.0%	2.5%	1.0%	0.5%	0.1%
1	0.708	0.936	1.323	1.642	2.354	2.706	3.841	5.024	6.635	7.879	10.828
2	1.833	2.197	2.773	3.219	4.159	4.605	5.991	7.378	9.210	10.597	13.816
3	2.946	3.405	4.108	4.642	5.739	6.251	7.815	9.348	11.345	12.838	16.266
4	4.045	4.579	5.385	5.989	7.214	7.779	9.488	11.143	13.277	14.860	18.467
5	5.132	5.730	6.626	7.289	8.625	9.236	11.070	12.833	15.086	16.750	20.515
6	6.211	6.867	7.841	8.558	9.992	10.645	12.592	14.449	16.812	18.548	22.458
7	7.283	7.992	9.037	9.803	11.326	12.017	14.067	16.013	18.475	20.278	24.322
8	8.351	9.107	10.219	11.030	12.636	13.362	15.507	17.535	20.090	21.955	26.125
9	9.414	10.215	11.389	12.242	13.926	14.684	16.919	19.023	21.666	23.589	27.877
10	10.473	11.317	12.549	13.442	15.198	15.987	18.307	20.483	23.209	25.188	29.588
11	11.530	12.414	13.701	14.631	16.457	17.275	19.675	21.920	24.725	26.757	31.264
12	12.584	13.506	14.845	15.812	17.703	18.549	21.026	23.337	26.217	28.300	32.910
13	13.636	14.595	15.984	16.985	18.939	19.812	22.362	24.736	27.688	29.819	34.528
14	14.685	15.680	17.117	18.151	20.166	21.064	23.685	26.119	29.141	31.319	36.123
15	15.733	16.761	18.245	19.311	21.384	22.307	24.996	27.488	30.578	32.801	37.697
16	16.780	17.840	19.369	20.465	22.595	23.542	26.296	28.845	32.000	34.267	39.252
17	17.824	18.917	20.489	21.615	23.799	24.769	27.587	30.191	33.409	35.718	40.790
18	18.868	19.991	21.605	22.760	24.997	25.989	28.869	31.526	34.805	37.156	42.312
19	19.910	21.063	22.718	23.900	26.189	27.204	30.144	32.852	36.191	38.582	43.820
20	20.951	22.133	23.828	25.038	27.376	28.412	31.410	34.170	37.566	39.997	45.315
21	21.991	23.201	24.935	26.171	28.559	29.615	32.671	35.479	38.932	41.401	46.797
22	23.031	24.268	26.039	27.301	29.737	30.813	33.924	36.781	40.289	42.796	48.268
23	24.069	25.333	27.141	28.429	30.911	32.007	35.172	38.076	41.638	44.181	49.728
24	25.106	26.397	28.241	29.553	32.081	33.196	36.415	39.364	42.980	45.559	51.179
25	26.143	27.459	29.339	30.675	33.247	34.382	37.652	40.646	44.314	46.928	52.620
26	27.179	28.520	30.435	31.795	34.410	35.563	38.885	41.923	45.642	48.290	54.052
27	28.214	29.580	31.528	32.912	35.570	36.741	40.113	43.195	46.963	49.645	55.476
28	29.249	30.639	32.620	34.027	36.727	37.916	41.337	44.461	48.278	50.993	56.892
29	30.283	31.697	33.711	35.139	37.881	39.087	42.557	45.722	49.588	52.336	58.301
30	31.316	32.754	34.800	36.250	39.033	40.256	43.773	46.979	50.892	53.672	59.703
35	36.475	38.024	40.223	41.778	44.753	46.059	49.802	53.203	57.342	60.275	66.619
40	41.622	43.275	45.616	47.269	50.424	51.805	55.758	59.342	63.691	66.766	73.402
45	46.761	48.510	50.985	52.729	56.052	57.505	61.656	65.410	69.957	73.166	80.077
50	51.892	53.733	56.334	58.164	61.647	63.167	67.505	71.420	76.154	79.490	86.661
55	57.016	58.945	61.665	63.577	67.211	68.796	73.311	77.380	82.292	85.749	93.168
60	62.135	64.147	66.981	68.972	72.751	74.397	79.082	83.298	88.379	91.952	99.607

Confidence intervals for two populations

Learning objectives

After lectures 20–23, we will be able to:

- Build confidence intervals for:
  - unknown means;
  - unknown variances;
  - unknown proportions.
- Build confidence intervals for:
  - the difference between two unknown means;
  - the ratio between two unknown variances;
  - the difference between two unknown proportions.
- Understand the effect of Type I error, or probability α.
- Calculate errors and interval margins.
- Select appropriate sample sizes to keep errors below a limit.

Motivation: Do masks work?

There has been an ongoing discussion about whether mandating universal mask wearing curbs COVID-19. All politics aside, there was a very interesting study coming from Kansas: apparently, counties that mandated masks saw smaller increases (or decreases) in the onset of new COVID-19 cases, than counties that did not mandate masks. Could we prove that (within a given specified level of confidence)?

Motivation: Does IE 300 have more variable grades than IE 310?

When deciding a technical elective, we also look for how *variable* the grading is; not only what the average is! A class that has an A- average is not necessarily "easier" or "more straightforward" than a class that has a B average. Instead, we also want to see what the variances are. The question we would like to answer then: how much more variable is class A compared to class B? Or, to put it in confidence interval terms, what is the ratio of variances between two populations with 95% confidence?

Two population confidence intervals

In this set of notes, we turn our focus to two different populations and how they compare. More specifically, in this lecture we see confi-

dence intervals on:

1. the difference of two means, $\mu_1 - \mu_2$.
2. the difference between two proportions, $p_1 - p_2$.
3. the ratio of two variances, $\frac{\sigma_1^2}{\sigma_2^2}$.

Why would we look at two populations? Well, in many practical applications, we are given more than one populations to compare. For example, we may want to compare the performance of a drug in two groups of patients. At a similar vein, we may want to check the differences in driving on ice between more (> 10 years) and less experienced ($0 - 10$ years) drivers. Finally, at a problem that we can relate to in 2020, we may want to see how people living in two different states vote?

One thing is for sure: in all these cases, it is imperative to create confidence intervals for more than just one population.

Difference in means

Consider two **normally distributed** populations with unknown means μ_1, μ_2. We are interested in quantifying the difference in their means:

$$\mu_1 - \mu_2.$$

How about we try again what we did before? That is:

- Take a sample of size n_1 from the first population and calculate the sample average $\overline{X}_1$.
- Take a sample of size n_2 from the second population and calculate the sample average $\overline{X}_2$.
- Estimate $\mu_1 - \mu_2$ by $\overline{X}_1 - \overline{X}_2$.

This will be a good point estimate, for sure [78]. But what about the confidence interval?

[78] Why? Can you prove that?

Difference in means of two populations with known variances

Before we get to the confidence intervals, let us define a new "kind" of standard deviation.

Definition 62 (Pooled standard deviation) *Given two samples of sizes n_1, n_2, with known respective standard deviations σ_1, σ^2, we define their pooled standard deviation as:*

$$\sigma_P = \sqrt{\frac{(n_1 - 1)\,\sigma_1^2 + (n_2 - 1)\,\sigma_2^2}{n_1 + n_2 - 2}}.$$

With this definition, and keeping the same logic as in Lecture 20, we get our first confidence interval:

$$\boxed{\overline{X}_1 - \overline{X}_2 - z_{\alpha/2}\sigma_P\sqrt{\tfrac{1}{n_1}+\tfrac{1}{n_2}} \leq \mu_1 - \mu_2 \leq \overline{X}_1 - \overline{X}_2 + z_{\alpha/2}\sigma_P\sqrt{\tfrac{1}{n_1}+\tfrac{1}{n_2}}}$$

Note how similar the setup is as in the case of single population means with known standard deviation! The only differences are in the point estimate used ($\overline{X}$ vs. $\overline{X}_1 - \overline{X}_2$), in the standard deviation used (σ vs. the pooled standard deviation σ_P), and in the population size (we multiplied by $1/\sqrt{n}$ vs. multiplying by $\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$).

BMI

The body mass index of 1545 people was found to be on average 28.8. The same index for another population of 1781 people was calculated as (on average) 27.6. The body mass index has variance of 9 (in both populations). Build a 99%-confidence interval for the difference in the BMI between the two populations.

First, we will need $z_{0.005} = 2.576$. Then, we have:

- $\overline{X}_1 - \overline{X}_2 = 1.2$.
- $\sigma_P = \sqrt{\frac{1544\cdot 9 + 1780\cdot 9}{3324}} = 3$. (unsurprising as both populations had the same σ to begin with.)
- $L = 1.2 - 2.576 \cdot 3 \cdot \sqrt{\frac{1}{1545}+\frac{1}{1781}} = 0.931$.
- $U = 1.2 + 2.576 \cdot 3 \cdot \sqrt{\frac{1}{1545}+\frac{1}{1781}} = 1.469$.

Hence, the difference in the mean BMI between these two populations is

$$[0.931, 1.469]\,.$$

Difference in means of two populations with unknown variances

Now, assume we do not know the population variances! If we do not know them, then we can take a page from the single population confidence intervals book: we can estimate these unknown variances using the sample variances, s_1^2 and s_2^2. As soon as we estimate the variances though (rather than having the true ones), we get two changes:

Change 1: we no longer have a z-value (from a normal distribution), but we have a t-value (from a Student's T distribution) with $n_1 + n_2 - 2$ degrees of freedom.

Change 2: as we do not know σ_1, σ_2, we **estimate** the pooled standard deviation as

$$s_P = \sqrt{\frac{(n_1-1)\,s_1^2 + (n_2-1)\,s_2^2}{n_1+n_2-2}}.$$

$$\overline{X}_1 - \overline{X}_2 - t_{\alpha/2,n_1+n_2-2} s_P \sqrt{\tfrac{1}{n_1}+\tfrac{1}{n_2}} \le \mu_1 - \mu_2 \le \overline{X}_1 - \overline{X}_2 + t_{\alpha/2,n_1+n_2-2} s_P \sqrt{\tfrac{1}{n_1}+\tfrac{1}{n_2}}$$

BMI: take 2

Assume that we run a smaller experiment on the body mass index of two populations. We now have collected a sample of 4 from some population with $\overline{X}_1 = 26.2$ and $s_1 = 2$, and a sample of 6 from a different population with $\overline{X}_2 = 28$ and $s_2 = 3.6$. Build a 99% confidence interval for $\overline{X}_1 - \overline{X}_2$.

Now, instead of a z-value, we will need $t_{0.005,9} = 4.297$. We have:

- $\overline{X}_1 - \overline{X}_2 = -1.8$.
- $s_P = \sqrt{\frac{3\cdot 4 + 5\cdot 12.96}{8}} = \sqrt{9.6} = 3.1$.
- $L = -1.8 - 4.297 \cdot 3.1 \cdot \sqrt{\frac{1}{4}+\frac{1}{6}} = -10.4$.
- $U = -1.8 + 4.297 \cdot 3.1 \cdot \sqrt{\frac{1}{4}+\frac{1}{6}} = 6.8$.

Hence, the difference in the mean BMI between these two populations has now become

$$[-10.4, 6.8]\,.$$

A bit of critical analysis in these two results. Take a look at the first confidence interval from the larger experiment with the known standard deviations. We got (with 99% confidence) that the first population has bigger BMI values by at least 0.931 points and up to 1.469 points. Hence, we could say that the first population has **more BMI** with 99% confidence! On the other hand, looking at the second smaller experiment with unknown standard deviations, we got a much bigger and much less clear confidence interval: specifically, we got that the BMI difference is between -10.4 and 6.8. This translates to possibly the first population having smaller BMI by a whole 10.4 points or bigger BMI by 6.8 points! Hence, we could **not** claim that the first population has **more nor less BMI** with 99% confidence!

This type of critical thinking will be invaluable when we move to hypothesis testing.

Confidence intervals for the ratio of the variances of two normally distributed populations

Before we calculate confidence intervals on the variances of two normally distributed populations, we define *the ratio of two sample variances* as:

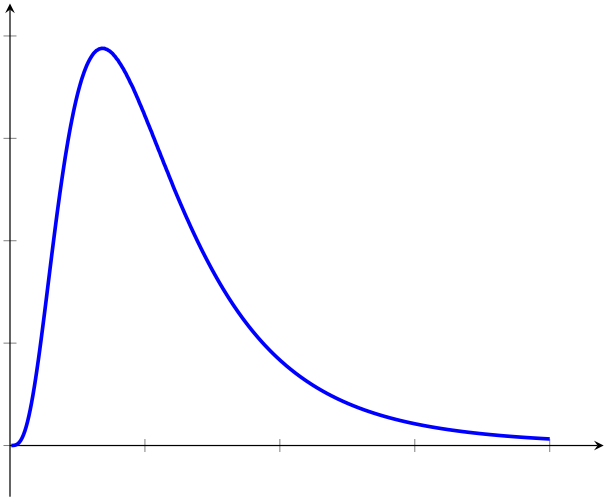
$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}.$$

Recall that variances have degrees of freedom: hence, assuming we have a sample of n_1 observations from the first and n_2 observations from the second population, then we say that F is distributed as an F distribution with $n_1 - 1$ degrees of freedom in the numerator and $n_2 - 1$ degrees of freedom in the denominator, or, mathematically, we write that:

$$F \sim F_{n_1-1,n_2-1}.$$

We show this visually in Figure 77. Note that much like the χ^2 distribution, the F distribution is also not symmetric.

Figure 77: The F distribution visually.



Much like with the other distributions we have seen, we also have a table for the F distribution (see the last two pages here)! It is a little different to read: here the columns and rows represent the degrees of freedom of the numerator (v_1) and the denominator (v_2). Then, we find the value of α we are interested in to get the value.

Finding F distribution values

- $n_1 = 9$, $n_2 = 5$, $\alpha = 10\%$: we then look for $\nu_1 = n_1 - 1 = 8$, $\nu_2 = n_2 - 1 = 4$, $\alpha = 0.1$ and find $f_{8,4,0.1} = 3.95$.
- $n_1 = 5$, $n_2 = 16$, $\alpha = 5\%$: we then look for $\nu_1 = n_1 - 1 = 4$, $\nu_2 = n_2 - 1 = 15$, $\alpha = 0.05$ and find $f_{4,15,0.05} = 3.06$.

Wait! What do I do if I am looking at other values, such as the ones used for $\alpha = 90\%$? Those are clearly not available in the table, right? Well, in that case we have:

$$\boxed{f_{u,v,\alpha} = \frac{1}{f_{v,u,1-\alpha}}}$$

In English: the f value for u degrees of freedom in the numerator, v degrees of freedom in the denominator and α is equal to 1 over the f value for v degrees of freedom in the numerator, u degrees of freedom in the denominator and $1 - \alpha$. Nifty, no? Let's put it to the use.

Finding F distribution values

In general, for two-sided confidence intervals we need values for $\alpha/2$ and $1 - \alpha/2$. So, assume we are building 80% confidence intervals and 95% confidence intervals for:

- $n_1 = 9$, $n_2 = 5$, $\alpha = 20\%$: we then look for $\nu_1 = n_1 - 1 = 8$, $\nu_2 = n_2 - 1 = 4$, $\alpha/2 = 0.10$ and find $f_{8,4,0.1} = 3.95$. We also need the same value but for $1 - \alpha/2$, $f_{8,4,0.9}$. We know it can be found as:

$$f_{8,4,0.9} = \frac{1}{f_{4,8,0.1}} = \frac{1}{2.81} = 0.356.$$

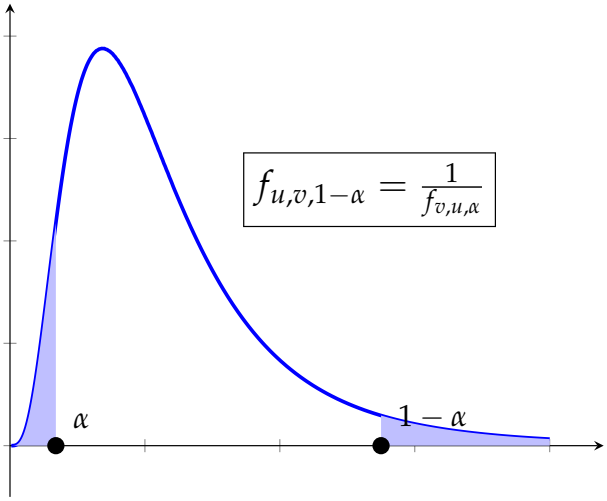
- $n_1 = 5$, $n_2 = 15$, $\alpha = 10\%$: we then look for $\nu_1 = n_1 - 1 = 4$, $\nu_2 = n_2 - 1 = 15$, $p = 1 - \alpha = 95\%$ and find $f_{4,15,0.05} = 3.06$. We also need $f_{4,15,0.95}$. We know it can be found as:

$$f_{4,15,0.95} = \frac{1}{f_{15,4,0.05}} = \frac{1}{5.86} = 0.171.$$

Please check this awesome online tool to help you do these calculations: `https://stattrek.com/online-calculator/f-distribution.aspx`. Additionally, see Figure 78 for an example of how α and $1 - \alpha$ are located in the F distribution.

Following a similar logic to the single population variance case, we finally have:

Figure 78: The F distribution and the marks for $1-\alpha$ and α.



$$f_{n_2-1,n_1-1,1-\alpha/2}\frac{s_1^2}{s_2^2} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq f_{n_2-1,n_1-1,\alpha/2}\frac{s_1^2}{s_2^2}$$

Before we put this to the test, let us remember what a critical value is! In essence, for a critical value $f_{u,v,\alpha}$ we need:

$$P(F \geq f_{u,v,\alpha}) = \alpha.$$

This implies the following: for a given value for α, we look at $p = 1-\alpha$ in the F-table!
Let's put this to use right away!

Semiconductor wafers and their oxide layers

The variability in the thickness of oxide layers in semiconductor wafers is a critical characteristic, where low variability is desirable. A company is investigating two different ways to mix gases so as to reduce the variability of the oxide thickness. We produce 16 wafers with each gas mixture and our results indicate that the standard deviation is $s_1 = 1.96$Å and $s_2 = 2.13$Å for the two mixtures. What is the 95% confidence intervals for the ratio between the two variances?

Semiconductor wafers and their oxide layers

We have been given a series of information:

- size of population 1: $n_1 = 16$;
- sample standard deviation for sample from population 1: $s_1 = 1.96$;
- size of population 2: $n_2 = 16$;
- sample standard deviation for sample from population 2: $s_2 = 2.13$.

Since we are looking for a 95% confidence interval we need two f values:

- $f_{n_2-1,n_1-1,\alpha/2} = f_{15,15,0.025} = 2.86$.
- $f_{n_2-1,n_1-1,1-\alpha/2} = \frac{1}{f_{n_1-1,n_2-1,\alpha/2}} = \frac{1}{2.86} = 0.35$.

Finally, the confidence interval for σ_1^2/σ_2^2 is found as:

$$\left[f_{n_2-1,n_1-1,1-\alpha/2}\frac{s_1^2}{s_2^2}, f_{n_2-1,n_1-1,\alpha/2}\frac{s_1^2}{s_2^2}\right] =$$
$$= [0.35 \cdot 0.847, 2.86 \cdot 0.847] = [0.296, 2.422]\,.$$

Confidence intervals for the difference of the proportions of two populations

As a reminder, if we had a single population with proportion p, then our confidence intervals, are given by:

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\cdot(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\cdot(1-\hat{p})}{n}}.$$

Recall that $\hat{p}$ is the observed (estimated) proportion based on the sample collected. Similarly, $\sqrt{\frac{\hat{p}\cdot(1-\hat{p})}{n}}$ is the estimated square error (standard deviation).

Now, assume we have two populations: one with true proportion p_1 and the other with true proportion p_2. If we do not know what p_1 and p_2 are, how can we estimate $p_1 - p_2$?

Well, let us follow a similar process:

1. Collect a sample from the first population of size n_1 and calculate the observed proportion $\hat{p}_1$.
2. Collect a sample from the second population of size n_2 and calculate the observed proportion $\hat{p}_2$.

3. Estimate $p_1 - p_2$ as $\hat{p}_1 - \hat{p}_2$.

Great! That will do! But, what about the confidence interval around it? Following the theory from the single population proportions, we get...

If $n_1 p_1, n_2 p_2, n_1(1-p_1), n_2(1-p_2)$ are all greater than or equal to 5, then $\hat{p}_1 - \hat{p}_2$ is **normally distributed** with mean $p_1 - p_2$ and variance $\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}$.

Using that, we finally obtain our confidence interval as:

$$\hat{p}_1 - \hat{p}_2 - z_{\alpha/2}\sqrt{\frac{\hat{p}_1\,(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2\,(1-\hat{p}_2)}{n_2}} \leq p_1 - p_2 \leq$$
$$\leq \hat{p}_1 - \hat{p}_2 + z_{\alpha/2}\sqrt{\frac{\hat{p}_1\,(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2\,(1-\hat{p}_2)}{n_2}}.$$

Environmentally conscious

Residents of major metropolitan areas in the US were asked whether they agree with the following statement:

"I consider my self environmentally conscious."

The answers they could give were either a "Yes" or a "No". Specifically, we focus on two cities: Portland and Philadelphia.

- Out of $n = 91$ respondents in Portland, 61 answered Yes.
- Out of $n = 100$ respondents in Philadelphia, 45 said Yes.

Build a 95% confidence interval for the true proportion difference $p_1 - p_2$, where p_1 is the proportion of people agreeing with the statement in Portland and p_2 the proportion of the same people in Philadelphia.

We have:

- $n_1 = 91, \hat{p}_1 = \frac{61}{91} = 0.67$.
- $n_2 = 100, \hat{p}_2 = \frac{45}{100} = 0.45$.

We also have $z_{\alpha/2} = z_{0.025} = 1.96$. Plugging everything together:

$$p_1 - p_2 \in [0.22 - 1.96 \cdot 0.07, 0.22 + 1.96 \cdot 0.07] = [0.08, 0.36]\,.$$

Let us dwell on this last result for one minute. What do we learn from this confidence interval? Well, we learn that residents of Portland are (with 95% confidence) **more environmentally conscious** than residents of Philadelphia! We could never make the same assertion simply by looking at the individual observed proportions: that is, we cannot make the claim simply through the argument that $\hat{p}_1 > \hat{p}_2$. But now? We most definitely can make it! Always with the sidenote that "with 95% confidence".

More on that, in the coming lectures.

CRITICAL VALUES OF THE F DISTRIBUTION

$\nu_2 \backslash \nu_1$	α	2	3	4	5	6	7	8	10	12	15	20	30	50	∞
1	0.100	49.5	53.6	55.8	57.2	58.2	59.1	59.7	60.5	61.0	61.5	62.0	62.6	63.0	63.3
	0.050	199	216	225	230	234	237	239	242	244	246	248	250	252	254
	0.025	800	864	900	922	937	948	957	969	977	985	993			
2	0.100	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.39	9.41	9.43	9.44	9.46	9.47	9.49
	0.050	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5
	0.025	39.0	39.2	39.2	39.3	39.3	39.4	39.4	39.4	39.4	39.4	39.4	39.5	39.5	39.5
3	0.100	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.23	5.22	5.20	5.18	5.17	5.15	5.13
	0.050	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.79	8.74	8.70	8.66	8.62	8.58	8.53
	0.025	16.0	15.4	15.1	14.9	14.7	14.6	14.5	14.4	14.3	14.3	14.2	14.1	14.0	13.9
4	0.100	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.92	3.90	3.87	3.84	3.82	3.79	3.76
	0.050	6.94	6.59	6.39	6.26	6.16	6.09	6.04	5.96	5.91	5.86	5.80	5.75	5.70	5.63
	0.025	10.6	9.98	9.60	9.36	9.20	9.07	8.98	8.84	8.75	8.66	8.56	8.46	8.38	8.26
5	0.100	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.30	3.27	3.24	3.21	3.17	3.15	3.10
	0.050	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.74	4.68	4.62	4.56	4.50	4.44	4.36
	0.025	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.62	6.52	6.43	6.33	6.23	6.14	6.02
6	0.100	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.94	2.90	2.87	2.84	2.80	2.77	2.72
	0.050	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.06	4.00	3.94	3.87	3.81	3.75	3.67
	0.025	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.46	5.37	5.27	5.17	5.07	4.98	4.85
7	0.100	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.70	2.67	2.63	2.59	2.56	2.52	2.47
	0.050	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.64	3.57	3.51	3.44	3.38	3.32	3.23
	0.025	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.76	4.67	4.57	4.47	4.36	4.28	4.14
8	0.100	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.54	2.50	2.46	2.42	2.38	2.35	2.29
	0.050	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.35	3.28	3.22	3.15	3.08	3.02	2.93
	0.025	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.29	4.20	4.10	4.00	3.89	3.81	3.67
9	0.100	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.42	2.38	2.34	2.30	2.25	2.22	2.16
	0.050	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.14	3.07	3.01	2.94	2.86	2.80	2.71
	0.025	5.71	5.08	4.72	4.48	4.32	4.20	4.10	3.96	3.87	3.77	3.67	3.56	3.47	3.33
10	0.100	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.32	2.28	2.24	2.20	2.16	2.12	2.06
	0.050	4.10	3.71	3.48	3.33	3.22	3.14	3.07	2.98	2.91	2.84	2.77	2.70	2.64	2.54
	0.025	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.72	3.62	3.52	3.42	3.31	3.22	3.08
11	0.100	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.25	2.21	2.17	2.12	2.08	2.04	1.97
	0.050	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.85	2.79	2.72	2.65	2.57	2.51	2.40
	0.025	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.53	3.43	3.33	3.23	3.12	3.03	2.88
12	0.100	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.19	2.15	2.10	2.06	2.01	1.97	1.90
	0.050	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.75	2.69	2.62	2.54	2.47	2.40	2.30
	0.025	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.37	3.28	3.18	3.07	2.96	2.87	2.72
13	0.100	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.14	2.10	2.05	2.01	1.96	1.92	1.85
	0.050	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.67	2.60	2.53	2.46	2.38	2.31	2.21
	0.025	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.25	3.15	3.05	2.95	2.84	2.74	2.60

CRITICAL VALUES OF THE F DISTRIBUTION

$\nu_2 \backslash \nu_1$	α	2	3	4	5	6	7	8	10	12	15	20	30	50	∞
14	0.100	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.10	2.05	2.01	1.96	1.91	1.87	1.80
	0.050	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.60	2.53	2.46	2.39	2.31	2.24	2.13
	0.025	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.15	3.05	2.95	2.84	2.73	2.64	2.49
15	0.100	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.06	2.02	1.97	1.92	1.87	1.83	1.76
	0.050	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.54	2.48	2.40	2.33	2.25	2.18	2.07
	0.025	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.06	2.96	2.86	2.76	2.64	2.55	2.40
16	0.100	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.03	1.99	1.94	1.89	1.84	1.79	1.72
	0.050	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.49	2.42	2.35	2.28	2.19	2.12	2.01
	0.025	4.69	4.08	3.73	3.50	3.34	3.22	3.12	2.99	2.89	2.79	2.68	2.57	2.47	2.32
17	0.100	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.00	1.96	1.91	1.86	1.81	1.76	1.69
	0.050	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.45	2.38	2.31	2.23	2.15	2.08	1.96
	0.025	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.92	2.82	2.72	2.62	2.50	2.41	2.25
18	0.100	2.62	2.42	2.29	2.20	2.13	2.08	2.04	1.98	1.93	1.89	1.84	1.78	1.74	1.66
	0.050	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.41	2.34	2.27	2.19	2.11	2.04	1.92
	0.025	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.87	2.77	2.67	2.56	2.44	2.35	2.19
19	0.100	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.96	1.91	1.86	1.81	1.76	1.71	1.63
	0.050	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.38	2.31	2.23	2.16	2.07	2.00	1.88
	0.025	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.82	2.72	2.62	2.51	2.39	2.30	2.13
20	0.100	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.94	1.89	1.84	1.79	1.74	1.69	1.61
	0.050	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.35	2.28	2.20	2.12	2.04	1.97	1.84
	0.025	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.77	2.68	2.57	2.46	2.35	2.25	2.09
25	0.100	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.87	1.82	1.77	1.72	1.66	1.61	1.52
	0.050	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.24	2.16	2.09	2.01	1.92	1.84	1.71
	0.025	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.61	2.51	2.41	2.30	2.18	2.08	1.91
30	0.100	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.82	1.77	1.72	1.67	1.61	1.55	1.46
	0.050	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.16	2.09	2.01	1.93	1.84	1.76	1.62
	0.025	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.51	2.41	2.31	2.20	2.07	1.97	1.79
60	0.100	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.71	1.66	1.60	1.54	1.48	1.41	1.29
	0.050	3.15	2.76	2.53	2.37	2.25	2.17	2.10	1.99	1.92	1.84	1.75	1.65	1.56	1.39
	0.025	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.27	2.17	2.06	1.94	1.82	1.70	1.48
80	0.100	2.37	2.15	2.02	1.92	1.85	1.79	1.75	1.68	1.63	1.57	1.51	1.44	1.38	1.24
	0.050	3.11	2.72	2.49	2.33	2.21	2.13	2.06	1.95	1.88	1.79	1.70	1.60	1.51	1.32
	0.025	3.86	3.28	2.95	2.73	2.57	2.45	2.35	2.21	2.11	2.00	1.88	1.75	1.63	1.40
100	0.100	2.36	2.14	2.00	1.91	1.83	1.78	1.73	1.66	1.61	1.56	1.49	1.42	1.35	1.21
	0.050	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.93	1.85	1.77	1.68	1.57	1.48	1.28
	0.025	3.83	3.25	2.92	2.70	2.54	2.42	2.32	2.18	2.08	1.97	1.85	1.71	1.59	1.35
∞	0.100	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.60	1.55	1.49	1.42	1.34	1.26	1.00
	0.050	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.83	1.75	1.67	1.57	1.46	1.35	1.00
	0.025	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.05	1.94	1.83	1.71	1.57	1.43	1.00

Hypothesis testing for proportions

Learning objectives

After lectures 24–25, we will be able to:

- Formulate statistical hypotheses for testing.
  - Carefully define null and alternative hypotheses.
  - Define what are the errors of Type I and Type II in hypothesis testing.
- Accept or reject hypotheses for proportions.
- Define the *P*-value and use it to accept or reject a hypothesis.

Motivation: True or False?

1. 94% of UIUC's College of Engineering graduates secure employment or go to graduate school within a year of graduation.
2. The average starting salary for these Engineering graduates is $78,159.
3. Electrical Engineering or Construction Management? Electrical engineers earn more in the start of their careers.
4. Electrical Engineering or Construction Management? The top 10% construction management professionals earn more than the top 10% electrical engineering professionals.
5. The majority of customers prefers Coke to Pepsi.
6. People with a dog in the house live longer.

What do all the above have in common and what are their differences? How can we *test these claims*? This is what hypothesis testing is all about!

Motivation: Grainger College of Engineering internships

The University of Illinois is interested in finding how many of their Engineering students already have internships lined up for next summer. The University believes that the proportion is 50%: that is, roughly half the students have secured internships.

The Universite sent out a survey that 140 students filled out with 84 of them stating they have an internship offer at their hands. Is the true percentage 50%? Or is it different than that?

Hypothesis testing

Once more, let us go back to the last weeks of lectures. We have seen **point estimation**, **confidence intervals**, and we are now moving to **hypothesis testing**. A quick review:

How do we estimate an unknown parameter/quantity?

1. **Point estimation**: provides us with a *single estimate* for some unknown parameter of a population.
   - Example: 63% prefer Coke to Pepsi.
2. **Interval estimation**: provides us with a *range/interval containing believable values* for some unknown parameter of a population.
   - Example: The percentage of people preferring Coke to Pepsi lies somewhere between 55% and 71%.
3. **Hypothesis testing**. We form a *hypothesis* or a *claim* for some unknown parameter of a population.
   - Example: Our claim is that more than half of the population prefers Coke to Pepsi.
   - We now need to somehow accept that claim; or reject it, based on **observations**.

Before we formally define hypothesis testing, we ask ourselves a series of motivating questions. Namely, we want to address the following:

1. How do we **formally state a hypothesis**? How do we put it in the proper mathematical terms?
2. When do we **accept** and when do we **reject a hypothesis**?
   - What does "accepting" mean in this mathematical context?
   - What does "rejecting" mean in this mathematical context?
3. What is the **likelihood of reaching the wrong conclusion**? That could mean that..
   - either we accept something that is false.
   - or we reject something that is true.

We are now ready to formally define hypothesis testing. We have the following definitions.

Definition 63 (Statistical hypothesis) *With the term **statistical hypothesis** we mean a claim about some unknown parameters or the unknown distributions of a population. Some examples include:*

- *The mean grade of a student in a class is a B+.*
- *The proportion of students that end up with an A in a class is 25%.*
- *The grade of a student in a class is normally distributed.*

A statistical hypothesis is divided in two parts. The first one is referred to as a **null hypothesis**, H_0, which is the hypothesis/claim that is being tested. As an example, our null hypothesis could be that the mean grade is a B+, or that the true proportion of students with an A is 25%.

The second one is the **alternative hypothesis**, H_1, which is either the opposite of or simply an alternative to the null hypothesis/claim. For example, the alternative hypothesis could be that the mean grade is **not** a B+, or that the true proportion of students with an A is smaller than 25%. We proceed with some examples of formulating statistical hypotheses.

Average grades

Let's assume our claim is the following:

"The average grade of a student in a class is 84&."

Define μ as the average score of a student in a class. Based on that we formulate the statistical hypothesis as:

$$H_0 : \mu = 84\%.$$
$$H_1 : \mu \neq 84\%.$$

Coke vs. Pepsi

Say, our claim is now that:

"More than half of the population prefers Coke to Pepsi."

Let p be the proportion of people preferring Coke to Pepsi. Then, we can formulate this hypothesis as

$$H_0 : p = 0.5$$
$$H_1 : p < 0.5$$

Note that the null hypothesis is always an equality. The alternate hypothesis though changes depending on our original claim.

Eating greens

What about the following claim?

"There is no life expectancy change by eating vegetables."

First, we assume we have two populations: one that eats vegetables and one that does not. Let μ_i be the true mean life expectancy of each group. Then, we formulate our hypothesis as:

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_1 : \mu_1 - \mu_2 \neq 0.$$

Eating greens: reformulated

This will look very similar. Pay attention to the detail that changes!

"There is no life expectancy increase by eating vegetables."

Again, we assume the existence of two (eating vs. non-eating vegetables) populations. However, our hypothesis changes slightly now to:

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_1 : \mu_1 - \mu_2 > 0.$$

Before we head to the next definitions, we summarize some finer details of formulating a hypothesis.

- The null hypothesis is always an equality.
- The alternative hypothesis can be one- or two-sided, depending on the claim we are trying to prove/disprove.
- The hypothesis can deal with a single population; or with the comparison between two populations.

Let us get to the fundamental part of this lecture. **How do we perform hypothesis tests?** How do we decide whether we have enough information to accept or reject a hypothesis?

Definition 64 (Hypothesis test) *A **hypothesis test** is a statistical procedure to collect information based on a random sample, which can lead to making a decision about the null hypothesis.*

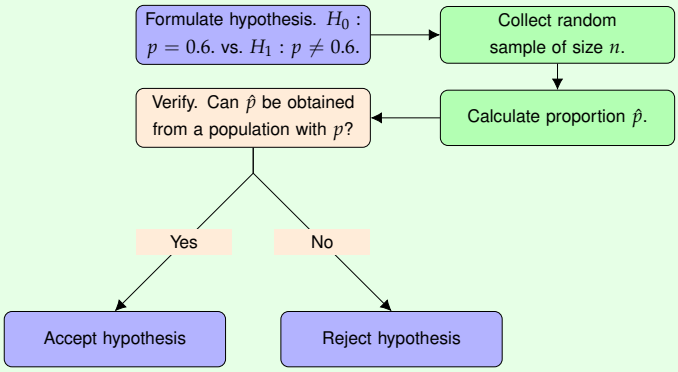
Let's see this with an example.

Coke vs. Pepsi

Assume you want to check whether 60% of the people prefer Coke to Pepsi. We can do the following operations.

1. First, formulate the statistical hypothesis as $H_0 : p = 0.6$ vs. $H_1 : p \neq 0.6$. We could have formulated a one-sided alternative hypothesis as $H_1 : p > 0.6$ or $H_1 : p < 0.6$ if we had more information about the original claim, but now $H_1 : p \neq 0.6$ will do.
2. Secondly, collect a random sample. Use it to estimate the proportion of people observed that prefer Coke to Pepsi.
3. Thirdly, try to verify. If your original hypothesis/claim is true, could you have gotten the observed proportion in the sample? If so, accept; if not, reject.

Visually:



We proceed to discuss how we may accept or reject a hypothesis. First of all, let us get one thing out of the way. While "accepting" and "rejecting" are universally used for hypothesis testing, it is more correct to think of them as "failing to reject" and "rejecting".

Think of the following parallel: say you are the jury at a trial. The hypothesis is that the defendant is innocent, no? The attorneys present data (observations) and it is up to you to decide whether it is enough to "reject innocence" or "fail to reject innocence". Note that failing to reject innocence is not the same as being innocent! It merely implies that there was not enough evidence to persuade you.

So, how does that translate to hypothesis testing? Let us study this using proportions.

Hypothesis testing for proportions

Assume we have a population X that has some unknown proportion p. We collect a sample of size n from that population. Recall that if we have $np \geq 5, n(1-p) \geq 5$, then we may make the claim that the observed proportion out of a sample of size n (defined as $\hat{p} = \frac{x}{n}$) is distributed as $\hat{p} \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$. [79]

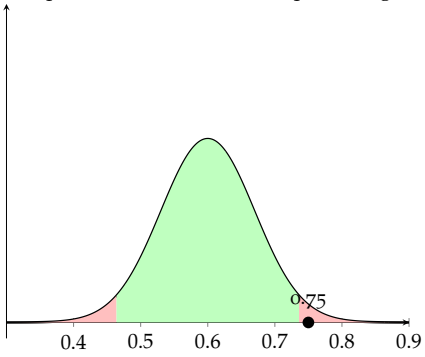
[79] This comes straight from our discussion about confidence intervals on proportions. See Lectures 20-23.

For the sake of the example, assume that we have

$$H_0 : p = 0.6.$$
$$H_1 : p \neq 0.6.$$

Say we have collected a sample of size $n = 50$. Then, **if the null hypothesis is true**, we'd expect a distribution of $\mathcal{N}(0.6, 0.0048)$. Visually, we get a normal distribution as the one presented in Figure 79. Now, say we select a confidence level of 95%. That means, visually, we'd expect 95% of the potential sample averages to fall in the green area; not the red. Finally, let's say that our sample average (for this $n = 50$) amounts to $\hat{p} = 0.75$. We also mark that in the figure.

Figure 79: A figure showing the distribution of the population **if the null hypothesis is true**, green and red areas marking the critical acceptance and rejection regions, and a point at $\hat{p} = 0.75$ that represents the obtained sample average.



We claim that the above figure essentially captures hypothesis testing. The question becomes: does the observed proportion $\hat{p} = 0.75$ fall in the range of believable values (the critical regions of **accepting**)? Or does it fall outside them (in the critical regions of **rejecting** the hypothesis)?

Based on this, let us revisit the terms we use. **Accepting** and **rejecting** a hypothesis are not the most appropriate terms for the outcomes of a hypothesis test. Instead, from now on, we will write that we:

- **Reject** the hypothesis, when we have sufficient observations to claim that the null hypothesis is not true.

  - This is a s **strong conclusion**.
  - It implies the existence of sufficient evidence agains the hypothesis.
  - In the end of this, we are quite certain that H_0 is wrong.

- **Fail to reject** the hypothesis, when we are not sure about the validity of the null hypothesis.

  - Consequently, it is a **weak conclusion**.
  - It merely implies the lack of sufficient evidence agains the hypothesis.
  - It does not mean that H_0 is true! It only implies that we are uncertain about either H_0 or H_1 being true.

Reaching the wrong conclusions

How many types of errors do you foresee appearing with this way of testing a hypothesis? Let's see this in tabular form:

Decision	H_0 is true	H_0 is false
Reject H_0	incorrect decision	correct decision
Fail to reject H_0	correct decision	incorrect decision

We will then need to formally define these two types of errors. Their names are "uninspired". These two are defined as **Type I** or α error [80] and **Type II** or β error.

[80] α? Is it.. the same as α in confidence intervals? Oh, yes. Yes it is.

Definition 65 (Type I errors) *The Type I or α error happens when we reject H_0 even though H_0 is valid. It is quantified as*

$$\alpha = P(\textit{reject } H_0 \textit{ when } H_0 \textit{ is true}) = P(\textit{reject } H_0 | H_0).$$

Definition 66 (Type II errors) *The Type II or β error happens when we fail to reject H_0 even though H_0 is not true. It is quantified as*

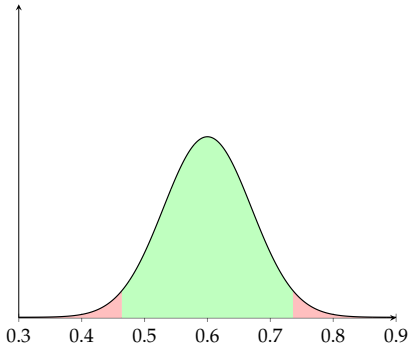
$$\beta = P(\textit{fail to reject } H_0 \textit{ when } H_0 \textit{ is false}) = P(\textit{fail to reject } H_0 | \overline{H}_0).$$

The courthouse parallel

Take a minute and think of the parallels to the jury trial example from before. What is α and β in a trial setting?

Let us focus on α. We have been using it all along!

$$\boxed{\alpha = P(\text{Type I error}) = P(\text{reject } H_0 \text{ when } H_0 \text{ is true}).}$$



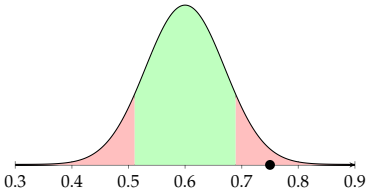
- $1 - \alpha$ is called the *significance* or the *size* of the test.
- It is equivalent to the red shaded areas.
- It can be improved by selecting stricter confidence levels.

The courthouse parallel (cont'd)

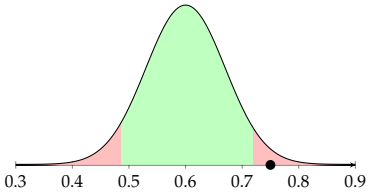
You can improve α in a trial by asking for more and more evidence. For example, "I will not find anyone guilty unless you present video evidence that they have done it" increases α signficantly, doesn't it? I wonder what happens to β, though...

Let us see another example for the effect of α. In our motivating example, we asked $n = 50$ people to check the hypothesis that $p = 0.6$. Assume out of that sample we get $\hat{p} = 0.75$. Then, the following would be the visual results for different significance levels (values for α):

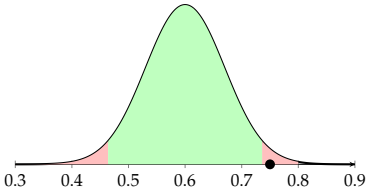
$\alpha = 20\%$



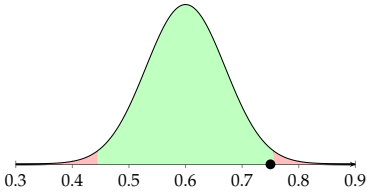
$\alpha = 10\%$



$\alpha = 5\%$



$\alpha = 2.5\%$



Hence, the bigger the α we are willing to accept, the tougher it becomes to reject a hypothesis. When $\alpha = 0$ (no error accepted!), then we no longer can reject a hypothesis.

We now move to β.

$$\boxed{\beta = P(\text{Type II error}) = P(\text{fail to reject } H_0 \text{ when } H_0 \text{ is false}).}$$

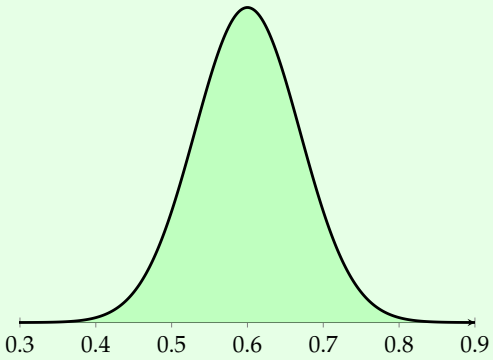
- It is related to $1 - \beta$, the *power* of the test.
- To formulate, it requires a **specific alternative hypothesis**.
- It decreases as the difference between the hypothesized and the true value of the hypothesis increases.

What this tells us is that β is not universal, given a hypothesis test. Instead, it depends on what we are comparing H_0 to. We show this in practice in the next pages.

Type II errors

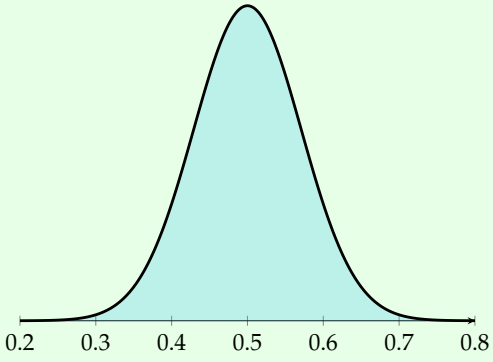
We have a company that offers a service that needs to be at 60% or above. The company is in trouble when the service quality lowers at 50%. To avoid this, they have an inspection mechanism in place. From time to time, they collect a sample of $n = 50$ services and make sure that average lies in the acceptance region! How often are they wrong and they *believe* they are good when they are not? What is the probability they accept the hypothesis that $p = 0.6$ when in fact the true p has lowered to 0.5?

The sampling distribution for $n = 50$ **if** the null hypothesis that $p = 0.6$ is true. Recall this is $\mathcal{N}(0.6, 0.0048)$.



In a similar manner, we can represent the sampling distribution for $n = 50$ when $p = 0.5$ instead! It would be $\mathcal{N}(0.5, 0.05)$.

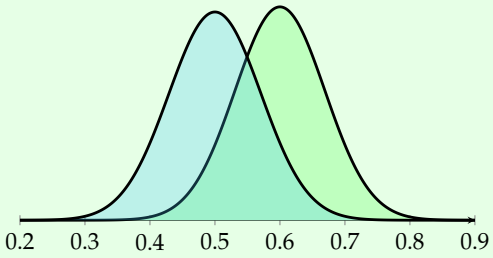
The sampling distribution for $n = 50$ **if** $p = 0.5$ is true.



Let us try to plot these two together!

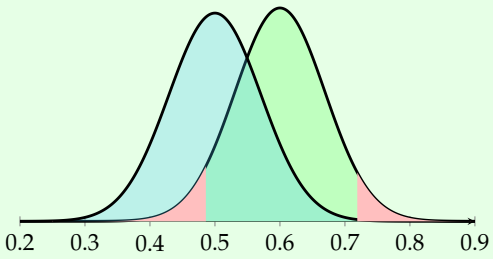
Type II errors

Plotting the two together reveals quite the overlap.



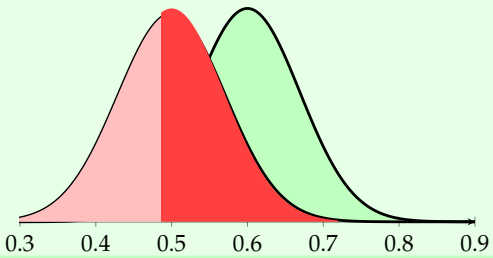
This means that it is quite possible that a value in the overlap may correspond to a "reality" of $p = 0.6$ or one of $p = 0.5$. But, remember! We only accept part of the first curve, depending on our α! Let's add this to the plot!

When we add the regions where we'd reject the original null hypothesis $H_0 : p = 0.6$. Here we use $\alpha = 10\%$.



Still, though. Observe the area in dark red below.

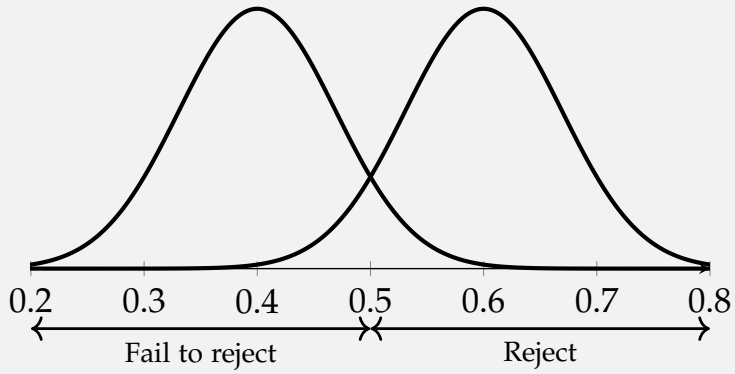
The dark area represents the β error! The lighter red area shows the power of the test $(1 - \beta)$.



As practice, paint the following.

Practice with the visuals

Here, we assume that $H_0 : p = 0.4$ (the null hypothesis) and say the alternative is $p = 0.6$. We have already provided where you would reject the null hypothesis and where you would not. Mark the following areas: (i) the region where you reject the null hypothesis, (ii) the region where you have the β error, (iii) the region of the power of the test $(1 - \beta)$.



How could we mathematically calculate the β error? Let us see the red area we need to be covering. That would be between $0.6 - z_{\alpha/2} \cdot 0.0693$ (recall that we have $\mathcal{N}(0.6, 0.048)$, so $\sqrt{0.048} = 0.0693$) and $0.6 + z_{\alpha/2} \cdot 0.0693$. For the sake of the example let us use $\alpha = 10\%$, which leads to $z_{0.05} = 1.645 \implies 0.6 - z_{\alpha/2} \cdot 0.0693 = 0.486$ and $0.6 + z_{\alpha/2} \cdot 0.0693 = 0.714$. Hence, we have, assuming that $p = 0.5$ is right and hence distributed with $\mathcal{N}(0.5, 0.005)$:

$$\beta = P(0.486 \leq p \leq 0.714) = P(p \leq 0.714) - P(p \leq 0.486) =$$
$$= \Phi\left(\frac{0.714 - 0.5}{\sqrt{0.005}}\right) - \Phi\left(\frac{0.486 - 0.5}{\sqrt{0.005}}\right) = \Phi(3.03) - \Phi(-0.20) = \Phi(3.03) - 1 + \Phi(0.20) =$$
$$= 0.9988 - 1 + 0.5793 = 57.81\%.$$

Before we finish this discussion, we provide a couple of observations about the Type I and Type II errors:

- Observation #1: assuming a fixed sample size, then decreasing one error will result in an increase of the other error.
  - Decreasing α will imply an increase in β.

  - Decreasing β will imply an increase in α.
- Observation #2: both errors can be reduced by increasing the sample size.

Finishing the proportion hypothesis testing procedure

We are *finally* ready to finish the discussion on hypothesis testing for proportions! We separate our discussion in three cases, depending on the hypothesis testing format (two-sided or one-sided).

Two-sided hypothesis testing

1. **Preliminaries.**
   - Select the desired α (significance $1-\alpha$).
   - Set up your hypothesis test as:

$$H_0 : p = p_0$$
$$H_1 : p \neq p_0.$$

2. **Compute test statistic** based on sample of size n.
   - $\hat{p}$

     or
   - $Z_0 = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$.
3. **Check.**
   - Is $\hat{p}$ below $p_0 - z_{\alpha/2}\sqrt{\frac{p_0(1-p_0)}{n}}$ or above $p_0 + z_{\alpha/2}\sqrt{\frac{p_0(1-p_0)}{n}}$?
   - Equivalently, is Z_0 below $-z_{\alpha/2}$ or above $z_{\alpha/2}$?
4. **Decide.**
   - If the check is true, reject the hypothesis.
   - Otherwise, fail to reject it.

To calculate the power of the test (or β), first identify the alternative you are investigating, say $p = p_1$. Then, assume that your sample is distributed such that $p \sim \mathcal{N}\left(p_1, \frac{p_1(1-p_1)}{n}\right)$. Finally, calculate:

$$\beta = P\left(p_0 - z_{\alpha/2}\sqrt{\frac{p_0\left(1-p_0\right)}{n}} \leq p \leq p_0 + z_{\alpha/2}\sqrt{\frac{p_0\left(1-p_0\right)}{n}}\right).$$

One-sided hypothesis testing Assume now that we are looking for the upper alternative hypothesis ($p > p_0$).

1. **Preliminaries.**
   - Select the desired α (significance $1 - \alpha$).
   - Set up your hypothesis test as:

$$H_0 : p = p_0$$
$$H_1 : p > p_0.$$

2. **Compute test statistic** based on sample of size n.
   - $\hat{p}$

     or
   - $Z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$.
3. **Check.**
   - Is $\hat{p}$ above $p_0 + z_\alpha \sqrt{\frac{p_0(1-p_0)}{n}}$?
   - Equivalently, is Z_0 above z_α?
4. **Decide.**
   - If the check is true, reject the hypothesis.
   - Otherwise, fail to reject it.

To calculate the power of the test (or β), again identify the alternative you are investigating, say $p = p_1$. Then, assume that your sample is distributed such that $p \sim \mathcal{N}\left(p_1, \frac{p_1(1-p_1)}{n}\right)$. Finally, calculate:

$$\beta = P\left(p \leq p_0 + z_\alpha \sqrt{\frac{p_0\,(1-p_0)}{n}}\right).$$

For the lower alternative hypothesis ($p < p_0$), we take very similar steps.

1. **Preliminaries.**
   - Select the desired α (significance $1 - \alpha$).
   - Set up your hypothesis test as:

$$H_0 : p = p_0$$
$$H_1 : p < p_0.$$

2. **Compute test statistic** based on sample of size n.
   - $\hat{p}$
     or
   - $Z_0 = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$.
3. **Check.**
   - Is $\hat{p}$ below $p_0 - z_\alpha\sqrt{\frac{p_0(1-p_0)}{n}}$?
   - Equivalently, is Z_0 below $-z_\alpha$?
4. **Decide.**
   - If the check is true, reject the hypothesis.
   - Otherwise, fail to reject it.

To calculate the power of the test (or β), first identify the alternative you are investigating, say $p = p_1$. Then, assume that your sample is distributed such that $p \sim \mathcal{N}\left(p_1, \frac{p_1(1-p_1)}{n}\right)$. Finally, calculate:

$$\beta = P\left(p_0 - z_\alpha\sqrt{\frac{p_0\left(1-p_0\right)}{n}} \leq p\right).$$

A comprehensive example

We claim that the percentage of people in favor of a law is 0.5. A sample of 50 people gave $\hat{p} = 0.62$. Our hypothesis then is that $H_0 : p = 0.5$.

1. We would like the limits of our hypothesis test to be between 0.45 and 0.55. What is α?
2. What is the acceptance region for $\alpha = 0.05$ and a two-sided test? Can we reject the null hypothesis in favor of the alternative $p \neq 0.5$?
3. What is the acceptance region for $\alpha = 0.05$ and a one-sided test (alternative is $H_1 : p > 0.5$)? Can we reject the null hypothesis in favor of the alternative?
4. What is β if the true percentage in favor of the law is 0.70? Assume we are interested in a one-sided (upper) hypothesis test.

A comprehensive example

Recall that $\hat{p} \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$ and if $p = 0.5$ then $\hat{p} \sim \mathcal{N}(0.5, 0.005)$. We have

$$\begin{aligned} 1-\alpha &= P\left(0.45 \leq \hat{p} \leq 0.55\right) \implies \\ \implies \alpha &= P(\hat{p} < 0.45) + P(\hat{p} > 0.55) = \\ &= 2 - 2\Phi(0.71) = 2 - 1.5222 = 0.4778. \end{aligned}$$

Hence $\alpha = 0.5222 = 52.22\%$.

For the second part, this is easier: for $\alpha = 0.05$, we have $z_{\alpha/2} = z_{0.025} = 1.96$. Hence, the acceptance region would be between $0.5 - 1.96\sqrt{0.005} = 0.361$ and $0.5 + 1.96\sqrt{0.005} = 0.639$. We fail to reject the hypothesis, and hence we do not have enough evidence to disagree with $p = 50\%$.

For the third part, the only difference is that we are only focused on the alternative hypothesis of $H_1 : p > p_0$. Hence, we could only reject on that side. For $\alpha = 0.05$, we now use $z_\alpha = z_{0.05} = 1.645$ and we get: $0.5 + 1.645\sqrt{0.005} = 0.616$. The acceptance region is between 0 and 0.616. This means that we do have enough evidence to reject the null hypothesis now! We have enough evidence to disagree with $p = 50\%$ in favor of $p > 50\%$.

Finally, for the power of the test against $p = 0.7$: we already have that the upper limit is equal to 0.616. Hence, we would reject the hypothesis for any $\hat{p}$ above this. We are then looking at

$$\begin{aligned} 1-\beta &= P(\hat{p} > 0.616) = 1 - \Phi\left(\frac{0.616 - 0.7}{\sqrt{0.7 \cdot 0.3/50}}\right) = \\ &= 1 - \Phi(-1.30) = \Phi(1.30) = 0.9032. \end{aligned}$$

This is a pretty powerful test, even with a small sample size (comparatively) at $n = 50$.

Hypothesis testing for means and variances

Learning objectives

After lectures 26–27, we will be able to:

- Accept or reject hypotheses for means:
  - normally distributed population with known variance.
  - normally distributed population with unknown variance.
  - not normally distributed population, but with a large enough sample.
- Accept or reject hypotheses for normally distributed population variances.

P-values

Before we get to the means and variances, we begin from our previous worksheet. The last two exercises asked you to compute a so called P-value. Let us see the reason behind investigating this quantity.

Definition 67 (P-values) *In hypothesis testing, the P-value is the largest probability that still leads to the null hypothesis being correct (failing to reject).*

We begin by comparing α to the P-value. The probability α is a **rigid, pre-specified limit** to the risk we are willing to take. The risk, of course, translates to $P(\text{reject } H_0 | H_0 \text{ is true})$. No matter how useful α is fails to reveal the whole picture of statistical hypothesis testing.

On the other hand, the P-value is an **observed significance level**, that depends on the observed (obtained) sample. As it is the largest probability that would still allow us to fail to reject H_0, we immediately get the following as a consequence:

- if we accept the null hypothesis, then $\alpha \leq P$-value;
- if we reject the null hypothesis, then $\alpha > P$-value.

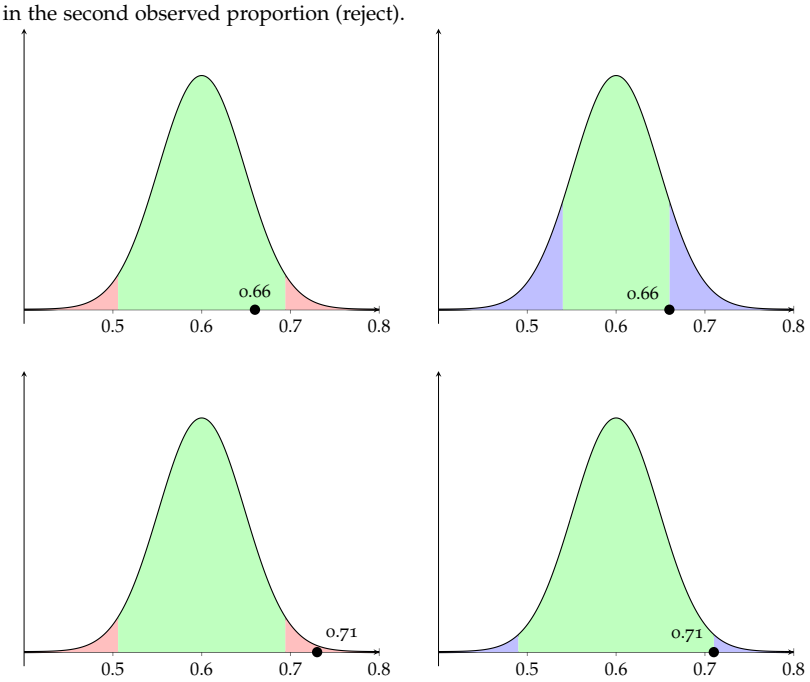
We show this consequence visually in Figure 80.

So, how to calculate a P-value? It depends on whether we are testing two sides or one side. Recall that for a given observed proportion $\hat{p}$, we may compute the test statistic $Z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$. Then, we have for the P-values: [81]

- Two-sided hypothesis: $P = 2\left(1 - \Phi\left(|Z_0|\right)\right)$

[81] See Worksheet 24-25 for the details.

Figure 80: In the first pair, we have that the observed proportion is $\hat{p} = 0.66$. We then show α (as in the rejection areas, in red) and the observed P-value (in blue). In the second pair, we show the same values but now the observed proportion is $\hat{p} = 0.71$. Note how P-value $\geq \alpha$ in the first observed proportion (fail to reject) and P-value $< \alpha$ in the second observed proportion (reject).



- One-sided (upper) hypothesis: $P = 1 - \Phi(Z_0)$
- One-sided (lower) hypothesis: $P = \Phi(Z_0)$

Based on our discussion here, we have **two ways to recommend rejection of a null hypothesis**:

1. Check whether the observed proportion $\hat{p}$ or the Z_0 statistic fall in the rejection region.
2. Calculate the P-value and compare to α.

In summary, we have for proportion hypothesis testing:

Proportion hypothesis testing

Null hypothesis:	Test statistic:	Distribution:
$H_0 : p = p_0.$	$Z_0 = \dfrac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}.$	$Z_0 \sim \mathcal{N}(0,1).$

H_1	Rejection region	P-value
$p \neq p_0$	$\lvert Z_0 \rvert > z_{\alpha/2}$	$2 \cdot (1 - \Phi(\lvert Z_0 \rvert))$
$p > p_0$	$Z_0 > z_\alpha$	$1 - \Phi(Z_0)$
$p < p_0$	$Z_0 < -z_\alpha$	$\Phi(Z_0)$

Reject if Z_0 falls in the rejection region or if P-value $< \alpha$.

Polling for a law

We surveyed 100 people on whether they support a new proposed law that will be on the ballot. 58% said that they do. Our hypothesis is that our county is evenly divided and hence 50% actually do support the law. What is the observed P-value? Should we reject the hypothesis that there are 50% in support of the law when $\alpha = 0.05$?

We have that

$$Z_0 = \frac{0.58 - 0.5}{0.05} = 1.6.$$

Now, we calculate $\Phi(Z_0) = \Phi(1.6) = 0.9452$. Hence, we get that P-value $= 2 \cdot (1 - 0.9452) = 0.1096$. Because P-value $\geq \alpha$, we fail to reject the hypothesis.

Polling for a law

Assume that in a different county, we surveyed 100 people on whether they support the law: we now got that 38% said that they do. Our hypothesis is that our city is again that the county supports it by 50%; but now our alternative hypothesis is the lower side only (i.e., $H_1 : p < 0.5$). What is the observed P-value in this case? Should we reject the hypothesis that there are 50% in support of the law when $\alpha = 0.05$?

We have that

$$Z_0 = \frac{0.38 - 0.5}{0.05} = -2.4.$$

Now, we calculate $\Phi(Z_0) = \Phi(-2.4) = 1 - \Phi(2.4) = 0.0082$. This is also the P-value. Note that P-value $< \alpha$, and thus we should reject the hypothesis in favor of the alternative (that is, less than 50% support the law).

Hypothesis testing for means

We now do the same for means. Recall the three cases we are interested in!

1. normally distributed population with known variance σ^2.
2. normally distributed population with unknown variance.
3. not normally distributed population, but we have a large enough sample.

Their derivation is again based on their confidence intervals, so we simply provide a summary of their results.

Normally distributed population with known variance σ^2

Mean with known variance hypothesis testing

Null hypothesis:	Test statistic:	Distribution:
$H_0 : \mu = \mu_0.$	$Z_0 = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}.$	$Z_0 \sim \mathcal{N}(0,1)\,.$

H_1	Rejection region	P-value
$\mu \neq \mu_0$	$\lvert Z_0 \rvert > z_{\alpha/2}$	$2 \cdot (1 - \Phi(\lvert Z_0 \rvert))$
$\mu > \mu_0$	$Z_0 > z_\alpha$	$1 - \Phi(Z_0)$
$\mu < \mu_0$	$Z_0 < -z_\alpha$	$\Phi(Z_0)$

Like earlier, we may reject if:

1. Check whether the observed sample average $\overline{X}$ or the Z_0 statistic fall in the rejection region.
2. Calculate the P-value and compare to α.

A life expectancy example

We select a random sample of 100 recorded deaths in the city of Urbana. The sample average is 71.8 years old. Assuming that life expectancy is normally distributed with a (known) standard deviation of 9 years, can we claim that life expectancy in Urbana is 70 years or is it higher? Use $\alpha = 5\%$.

First, state the null and alternate hypotheses. In our case:

$$H_0 : \mu = 70 \qquad H_1 : \mu > 70$$

Now, calculate $Z_0 = \frac{71.8-70}{9/\sqrt{100}} = 2$. Since it is one-sided, check $z_\alpha = z_{0.05} = 1.645$. We finally reject the hypothesis, as $Z_0 > z_\alpha$.

In this example, we could have built the upper confidence interval (for $\alpha = 5\%$) around the hypothesized mean as:

$$[0, U] = \left[0, \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}\right] = [0, 71.48]\,.$$

We can see that even doing it this way, we still note that the observed sample average ($\overline{X} = 71.8$) is outside the confidence interval, and hence we should reject the null hypothesis that life expectancy is at 70 years in favor of the alternative that it is higher than 70 years.

Normally distributed population with unknown variance σ^2

When the variance is unknown, we recall from the confidence interval discussion that the sampling distribution is no longer the normal one; instead we used the so-called Student's T distribution.

Mean with unknown variance hypothesis testing

Null hypothesis:	Test statistic:	Distribution:
$H_0 : \mu = \mu_0.$	$T_0 = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}}.$	$T_0 \sim T_{n-1}.$

H_1	Rejection region	P-value
$\mu \neq \mu_0$	$\lvert T_0 \rvert > t_{\alpha/2,n-1}$	$2 \cdot (1 - T_{n-1}(\lvert T_0 \rvert))$
$\mu > \mu_0$	$T_0 > t_{\alpha,n-1}$	$1 - T_{n-1}(T_0)$
$\mu < \mu_0$	$T_0 < -t_{\alpha,n-1}$	$T_{n-1}(T_0)$

A quick note about the notation. With lower case t we typically refer to the T distribution critical values (e.g., $t_{\alpha,n-1}$). On the other hand, with upper case T we typically refer to the cumulative distribution of the T distribution (e.g., $T_{n-1}(t) = P(T \leq t)$): for these values we would typically consult a cumulative distribution function for the T distribution table.

A life expectancy example

We select a random sample of 16 recorded deaths in the city of Urbana. The sample average is 71.8 years old and the sample standard deviation is 9 years. Assuming that life expectancy is normally distributed but with no known standard deviation, can we claim that life expectancy in Urbana is 70 years or is it higher? Use $\alpha = 5\%$.

We have the same null and alternate hypotheses as in the previous case, because it is again one-sided. However, now, we have a different test statistic:

$$H_0 : \mu = 70 \qquad H_1 : \mu > 70$$
$$T_0 = \frac{71.8 - 70}{9/\sqrt{16}} = 0.8.$$

The corresponding critical value we want to find is $t_{\alpha,n-1} = t_{0.05,15} = 1.753$. Due to that, we cannot reject the hypothesis, as $T_0 \leq t_{\alpha,n-1}$.

Much like what we did earlier, we will again build a confidence interval around the unknown mean. We'd have:

$$[0, U] = \left[0, \mu_0 + t_{\alpha,n-1}\frac{s}{\sqrt{n}}\right] = [0, 73.94]\,.$$

Note how the observed sample average is $\overline{X} = 71.8$ years and it

totally is part of the confidence interval. This is another way we could have deduced that we cannot reject the hypothesis.

Not normally distributed population

Not normally distributed population mean hypothesis testing

Null hypothesis:	Test statistic:	Distribution:
$H_0 : \mu = \mu_0.$	$Z_0 = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}}.$	$Z_0 \sim \mathcal{N}(0,1).$

H_1	Rejection region	P-value
$\mu \neq \mu_0$	$\lvert Z_0 \rvert > z_{\alpha/2}$	$2 \cdot (1 - \Phi(\lvert Z_0 \rvert))$
$\mu > \mu_0$	$Z_0 > z_\alpha$	$1 - \Phi(Z_0)$
$\mu < \mu_0$	$Z_0 < -z_\alpha$	$\Phi(Z_0)$

The only difference from the first case? We need a bigger sample size (say, $n \geq 30$) and we do not necessarily need the variance. Instead, we may estimate it (if unknown) as the sample variance s^2 and use it instead.

Hypothesis testing for normally distributed population variances

We are ready to show the last hypothesis testing procedure for a single population! For the variance of a normally distributed population we may reject or fail to reject a hypothesis on its true value following (again) the same logic as for its confidence interval, which was based on the χ^2 distribution.

Normally distributed population variance hypothesis testing

Null hypothesis:	Test statistic:	Distribution:
$H_0 : \sigma^2 = \sigma_0^2.$	$\chi_0^2 = \frac{(n-1)\,s^2}{\sigma_0^2}.$	$\chi_0^2 \sim \chi_{n-1}^2.$

H_1	Rejection region	CI region
$\sigma^2 \neq \sigma_0$	$\chi_0^2 > \chi_{\alpha/2,n-1}^2$ $\chi_0^2 < \chi_{1-\alpha/2,n-1}^2$	$\left[\frac{(n-1)\sigma_0^2}{\chi_{\alpha/2,n-1}^2}, \frac{(n-1)\sigma_0^2}{\chi_{1-\alpha/2,n-1}^2}\right]$
$\sigma^2 > \sigma_0$	$\chi_0^2 > \chi_{\alpha,n-1}^2$	$\left[\frac{(n-1)\sigma_0^2}{\chi_{\alpha,n-1}^2}, +\infty\right)$
$\sigma^2 < \sigma_0$	$\chi_0^2 < \chi_{1-\alpha,n-1}^2$	$\left(-\infty, \frac{(n-1)\sigma_0^2}{\chi_{1-\alpha,n-1}^2}\right]$

Hence, we would reject the null hypothesis whenever the χ_0^2 statistic is inside the rejection region, or whenever the hypothesized σ_0^2 is outside the confidence interval region.

A life expectancy example

We select a random sample of 16 recorded deaths in the city of Urbana. The sample average is 71.8 years old and the sample standard deviation is 9 years. Assuming that life expectancy is normally distributed but with no known standard deviation, can we claim that the standard deviation is equal to 7 years? Or is it different than that? Use $\alpha = 5\%$.

We have:

$$H_0 : \sigma^2 = 49 \qquad H_1 : \sigma^2 \neq 49$$

The test statistic is:

$$\chi_0^2 = \frac{15 \cdot 81}{49} = 24.796.$$

The corresponding critical value is $\chi^2_{\alpha/2,n-1} = \chi^2_{0.025,15} = 27.488$ and $\chi^2_{1-\alpha/2,n-1} = \chi^2_{0.975,15} = 6.262$. Hence, we cannot reject the hypothesis, as $\chi^2_{1-\alpha/2,n-1} \leq \chi_0^2 \leq \chi^2_{\alpha/2,n-1}$.

Like we did earlier, note that building the confidence interval we would have gotten:

$$[L, U] = \left[\frac{15 \cdot 49}{27.488}, \frac{15 \cdot 49}{6.262}\right] = [26.74, 117.37]\,,$$

which includes the sample variance $s^2 = 81$.

Hypothesis testing for two populations

Learning objectives

After lectures 28–29, we will be able to:

- Accept or reject hypotheses for two populations:
  - for the difference between their means.
  - for the difference between their proportions.
  - for the ratio of their variances.
- Use this statistical tool to compare two populations and make decisions about them.

Motivation: weather differences

You may have heard people say something along the lines "The weather is so different nowadays!" or "It used to snow during Halloween when I was a kid!" or even something like "Last year, it was much warmer/colder!". How can we employ statistics and probability theory to **reject** or **fail to reject** such claims? Could we somehow compare the mean temperature/snow/humidity/etc. from one year to the next?

Motivation: electoral considerations

During an election, political parties and candidates would like to know how specific populations behave. Do farmers overwhelmingly care about one item versus another? How about first-generation college students? We then would like to check and compare different populations, hopefully to find common ground that can help us address as many issues as possible, without alienating one group or another.

Motivation: online education and audiovisual tools

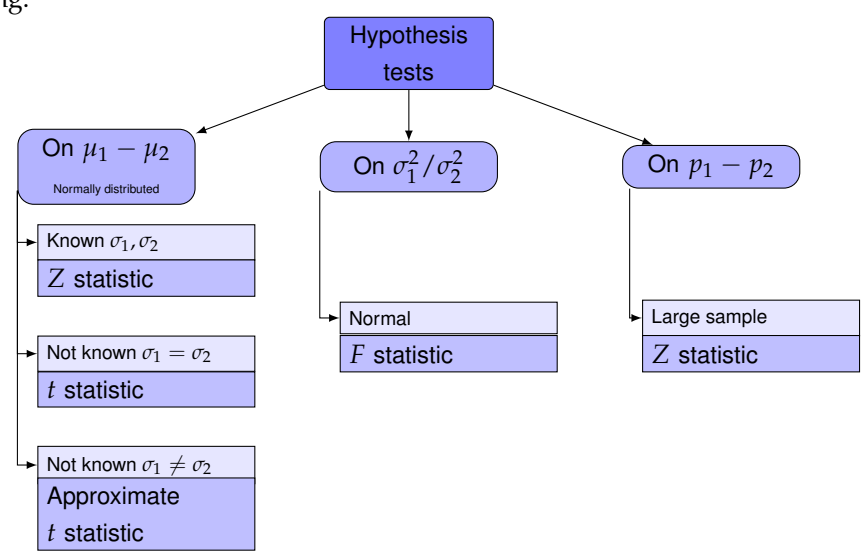
It is easy to completely demonize or completely agree with online education and its tools. What is more difficult is to quantify what happens with the variability of the performance of students in an online setting with the audiovisual tools at our disposal. Can we test the claim whether properly designed online courses lead to *lower variability* in the learning of students?

Hypothesis testing for two populations

I know we just discussed some motivating examples. Let me state them here in a more specific setting:

- Is the weather in Chicago significantly different than it was 10 years ago?
- Do students who have access to audiovisual aids for a class perform better than students who do not?
- Does the variability in the duration of a call decrease when the signal reception is improved?
- Do voters from one group overwhelmingly prefer one candidate over another in a local election compared to another group of voters?

All of the above examples have one thing in common: they deal with two populations and how they **compare** and **contrast**. Like we did in the past, we will again deal here with hypothesis testing for means, variances, and proportions. Visually, we discuss the following:



Hypothesis testing for means of two normally distributed populations

We have three cases (consult the earlier figure). They are:

1. normally distributed populations with known variances σ_1^2, σ_2^2.
2. normally distributed populations with unknown variances that are known to be equal, that is unknown $\sigma_1^2 = \sigma_2^2$.

3. normally distributed populations with unknown variances that are not known to be equal, that is unknown $\sigma_1^2 \neq \sigma_2^2$.

Their derivation is again based on their confidence intervals, so we simply provide a summary of their results.

Normally distributed populations with known variances σ_1^2, σ_2^2

A quick review before getting started.

- Assume two normally distributed populations X, Y with mean μ_1, μ_2 and standard deviations σ_1, σ_2. Then:
  1. Pick a sample of n_1 elements from X: $\overline{X} \sim \mathcal{N}\left(\mu_1, \sigma_1^2/n_1\right)$.
  2. Pick a sample of n_2 elements from Y: $\overline{Y} \sim \mathcal{N}\left(\mu_2, \sigma_2^2/n_2\right)$.

- Additionally, for combinations of the two populations, we have:
  1. $X + Y \sim \mathcal{N}\left(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2\right)$.
  2. $X - Y \sim \mathcal{N}\left(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2\right)$.
  3. $aX + bY \sim \mathcal{N}\left(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2\right)$.

- Finally, consider we pick a sample n_1 from X and a sample n_2 from Y:
  1. Pick a sample of n_1 elements from X: $\overline{X} \sim \mathcal{N}\left(\mu_1, \sigma_1^2/n_1\right)$.
  2. Pick a sample of n_2 elements from Y: $\overline{Y} \sim \mathcal{N}\left(\mu_2, \sigma_2^2/n_2\right)$.
  3. Combine to get that

$$\boxed{\overline{X} - \overline{Y} \sim \mathcal{N}\left(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2\right).}$$

Means with known σ_1^2, σ_2^2

Null hypothesis: $H_0 : \mu_1 - \mu_2 = \Delta_0$.

Test statistic: $Z_0 = \dfrac{\left(\overline{X}_1 - \overline{X}_2\right) - \Delta_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$.

Distribution: $Z_0 \sim \mathcal{N}(0,1)$.

H_1	Rejection region	P-value
$\mu_1 - \mu_2 \neq \Delta_0$	$\lvert Z_0\rvert > z_{\alpha/2}$	$2 \cdot (1 - \Phi(\lvert Z_0\rvert))$
$\mu_1 - \mu_2 > \Delta_0$	$Z_0 > z_\alpha$	$1 - \Phi(Z_0)$
$\mu_1 - \mu_2 < \Delta_0$	$Z_0 < -z_\alpha$	$\Phi(Z_0)$

Like in the single population cases, we should reject the null hypothesis under the following conditions:

1. Check whether the observed sample average $\overline{X}$ or the Z_0 statistic fall in the rejection region.
2. Calculate the P-value and compare to α.

Vaping

Two vaping products are being tested for their relationship with the outbreak of lung injury (see this CDC link). The first product has been responsible for more illnesses, so we are interested in seeing whether the nicotine content is at least 0.2 milligrams higher than in the second product. We have found that $n_1 = 50$ products of the first kind had an average nicotine content of $\overline{X}_1 = 2.61$ milligrams and $n_2 = 40$ products of the second kind had $\overline{X}_2 = 2.38$ milligrams. Using $\alpha = 0.05$, can we claim that the first product has 0.2 milligrams of difference or is it higher? Assume that standard deviations per product are known and equal to $\sigma_1 = 0.8$ and $\sigma_2 = 1.1$ milligrams, respectively.

We want to compare two population means: more specifically we want to see whether the difference is $\Delta_0 = 0.2$. We then have:

$$H_0 : \mu_1 - \mu_2 = 0.2 \quad H_1 : \mu_1 - \mu_2 > 0.2.$$

We pick:

- from the first population: $n_1 = 50, \overline{X}_1 = 2.61, \sigma_1 = 0.8$
- from the second population: $n_2 = 40, \overline{X}_2 = 2.38, \sigma_2 = 1.1$

Vaping

Now, calculate the test statistic as:

$$Z_0 = \frac{2.61 - 2.38 - 0.2}{\sqrt{\frac{0.8^2}{50} + \frac{1.1^2}{40}}} = \frac{0.03}{0.21} = 1/7 = 0.14.$$

It is one-sided, so find critical value $z_\alpha = z_{0.05} = 1.645$. Seeing as $Z_0 \leq z_\alpha$, we *fail to reject*.

We did not end up needing this, but we could have used the corresponding confidence intervals to decide whether we want to reject a hypothesis or not. How? First, calculate $\overline{X}_1 - \overline{X}_2$. Then, check the CI region: if the difference of the sample averages falls within or the confidence interval, then we fail to reject the null hypothesis.

H_1	CI region
$\mu_1 - \mu_2 \neq \Delta_0$	$(\mu_1 - \mu_2) \pm z_{\alpha/2}\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$
$\mu_1 - \mu_2 > \Delta_0$	$\left(-\infty, (\mu_1 - \mu_2) + z_{\alpha}\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}\right]$
$\mu_1 - \mu_2 < \Delta_0$	$\left[(\mu_1 - \mu_2) - z_{\alpha}\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}, +\infty\right)$

Normally distributed populations with unknown, but equal, variances $\sigma_1^2 = \sigma_2^2$

Let's try to derive the procedure now! First, assume that $\sigma_1 = \sigma_2 = \sigma$. Then the test statistic can be written as:

$$Z_0 = \frac{(\overline{X}_1 - \overline{X}_2) - \Delta_0}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

You've guessed the next step! If σ is unknown, I need to somehow estimate it.. Can we use a sample standard deviation? Recall that the sample standard deviation s can be a good estimator for the unknown population standard deviation σ. However, that was for a single population. What can we do here?

Since we have two samples from two populations, each with their own (possibly different) sample standard deviations s_1, s_2, we use the so-called **pooled estimator**, where we treat both as if they are one population. We then get:

$$s_p^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}.$$

A couple of notes:

- s_p^2 is a weighted average of the two variances s_1, s_2.
- Letting $n_1 = n_2$ leads to $s_p^2 = \left(s_1^2 + s_2^2\right)/2$.

Finally, as we are moving from *known* variances to *unknown* ones, we also need to account for it by moving from a normal distribution (and its z values) to a Student's T distribution (and the corresponding t values). Note that the distribution has $n_1 + n_2 - 2$ degrees of freedom. Overall we have:

Means with unknown but equal $\sigma_1^2 = \sigma_2^2$

Null hypothesis: Test statistic: Distribution:

$$H_0 : \mu_1 - \mu_2 = \Delta_0. \quad T_0 = \frac{(\overline{X}_1 - \overline{X}_2) - \Delta_0}{s_p\sqrt{1/n_1 + 1/n_2}}. \quad T_0 \sim T_{n_1+n_2-2}.$$

H_1	Rejection region	P-value
$\mu_1 - \mu_2 \neq \Delta_0$	$\lvert T_0\rvert > t_{\alpha/2,n_1+n_2-2}$	$2 \cdot (1 - T_{n_1+n_2-2}(\lvert T_0\rvert))$
$\mu_1 - \mu_2 > \Delta_0$	$T_0 > t_{\alpha,n_1+n_2-2}$	$1 - T_{n_1+n_2-2}(T_0)$
$\mu_1 - \mu_2 < \Delta_0$	$T_0 < -t_{\alpha,n_1+n_2-2}$	$T_{n_1+n_2-2}(T_0)$

Let us not forget that we may also decide to reject or not based on whether $\overline{X}_1 - \overline{X}_2$ falls within or outside the corresponding confidence interval!

H_1	CI region
$\mu_1 - \mu_2 \neq \Delta_0$	$(\mu_1 - \mu_2) \pm t_{\alpha/2,n_1+n_2-2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
$\mu_1 - \mu_2 > \Delta_0$	$\left(-\infty, (\mu_1 - \mu_2) + t_{\alpha,n_1+n_2-2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right]$
$\mu_1 - \mu_2 < \Delta_0$	$\left[(\mu_1 - \mu_2) - t_{\alpha,n_1+n_2-2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, +\infty\right)$

Catalyst comparison

Two catalysts, A and B, are being compared to see how they affect the mean yield of a chemical process. We have devised a pilot operation and results using the two catalysts are shown below for 8 runs. Using $\alpha = 0.05$ and assuming unknown but equal standard deviations, can we deduce that the two catalysts affect the yield differently?

Run	A	B	Run	A	B
1	91.5	89.19	5	91.79	97.19
2	94.18	90.95	6	89.07	97.04
3	92.18	90.46	7	94.72	91.07
4	95.39	93.21	8	89.21	92.75

We have:

- Population 1 for Catalyst A with: $n_1 = 8, \overline{X}_1 = 92.255, s_1 = 2.39$
- Population 2 for Catalyst B with: $n_2 = 8, \overline{X}_2 = 92.7325, s_2 = 2.98$

Recall that we know that $\sigma_1 = \sigma_2$, but this will not imply that the sample standard deviations will be equal too!
Now, on to formulating the hypothesis. We have:

$$H_0 : \mu_1 - \mu_2 = 0 \quad H_1 : \mu_1 - \mu_2 \neq 0.$$

The pooled standard deviation is:

$$s_P = \sqrt{\frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{7 \cdot 2.39^2 + 7 \cdot 2.98^2}{14}} = 2.7.$$

With all that, we get the corresponding test statistic as:

$$T_0 = \frac{92.255 - 92.7325 - 0}{2.7\sqrt{\frac{1}{8} + \frac{1}{8}}} = \frac{-0.4775}{1.35} = -0.35.$$

Since $\alpha = 0.05$ and we have a two-sided hypothesis, we need to identify the proper critical value as $t_{\alpha/2,14} = t_{0.025,14} = 2.145$. As $-t_{\alpha/2,14} \leq t_0 \leq t_{\alpha/2,14}$, we *fail to reject*.

Normally distributed populations with unknown, and not necessarily equal, variances $\sigma_1^2 \neq \sigma_2^2$

In this case, things get a little more complicated. Had we known what σ_1, σ_2 were:

$$Z_0 = \frac{(\overline{X}_1 - \overline{X}_2) - \Delta_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim \mathcal{N}(0,1).$$

Replacing σ_1, σ_2 with their sample counterparts s_1, s_2, we get:

$$T_0 = \frac{(\overline{X}_1 - \overline{X}_2) - \Delta_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \sim T_v.$$

We then say that T_0 is distributed *approximately* as the T distribution, but with degrees of freedom equal to v:

$$v = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}}.$$

This number will usually be fractional – we typically round down when needing to consult a t-table.

Means with unknown and not necessarily equal $\sigma_1^2 \neq \sigma_2^2$

Null hypothesis:	Test statistic:	Distribution:
$H_0 : \mu_1 - \mu_2 = \Delta_0.$	$T_0 = \frac{(\overline{X}_1 - \overline{X}_2) - \Delta_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}.$	$T_0 \sim T_v.$

H_1	Rejection region	P-value
$\mu_1 - \mu_2 \neq \Delta_0$	$\lvert T_0 \rvert > t_{\alpha/2,v}$	$2 \cdot (1 - T_v(\lvert T_0 \rvert))$
$\mu_1 - \mu_2 > \Delta_0$	$T_0 > t_{\alpha,v}$	$1 - T_v(T_0)$
$\mu_1 - \mu_2 < \Delta_0$	$T_0 < -t_{\alpha,v}$	$T_v(T_0)$

In the above, we calculate the approximate degrees of freedom v as:

$$v = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}}.$$

Let's put that to the use. We will follow the same example as before, however now we will drop the assumption that the two variances are equal.

Catalyst comparison

Two catalysts, A and B, are being compared to see how they affect the mean yield of a chemical process. We have devised a pilot operation and results using the two catalysts are shown below for 8 runs. Using $\alpha = 0.05$ and assuming unknown standard deviations, can we deduce that the two catalysts affect the yield differently?

Run	A	B	Run	A	B
1	91.5	89.19	5	91.79	97.19
2	94.18	90.95	6	89.07	97.04
3	92.18	90.46	7	94.72	91.07
4	95.39	93.21	8	89.21	92.75

We follow a very similar logic to earlier. However, we now will need the approximate degrees of freedom before proceeding (plus the T_0 statistic calculation changes slightly). We have the same hypothesis $H_0 : \mu_1 - \mu_2 = 0 \quad H_1 : \mu_1 - \mu_2 \neq 0$ and the same $n_1 = 8, \overline{X}_1 = 92.255, s_1 = 2.39, n_2 = 8, \overline{X}_2 = 92.7325, s_2 = 2.98$. Here is where things change now:

1. Calculate test statistic:

$$T_0 = \frac{92.255 - 92.7325 - 0}{\sqrt{\frac{2.39^2}{8} + \frac{2.98^2}{8}}} = \frac{-0.4775}{1.35} = -0.35.$$

2. Calculate approximate degrees of freedom:

$$v = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}} = \frac{(0.714 + 1.11)^2}{\frac{0.714^2}{7} + \frac{1.11^2}{7}} =$$
$$= \frac{1.824^2}{0.073 + 0.176} = 13.361 \to 13.$$

Finally, we find $t_{\alpha/2,v} = t_{0.025,13} = 2.16$. Becase $-t_{\alpha/2,13} \leq t_0 \leq t_{\alpha/2,13}$, we *fail to reject*.

Hypothesis testing for the ratio of the variances of two normally distributed populations

As must be obvious by now, we are taking each two population confidence interval and adapting it to the hypothesis testing procedure. Next up is the ratio of the two unknown variances of two **normally distributed** populations.

Ratio of variances

Null hypothesis:	Test statistic:	Distribution:
$H_0 : \sigma_1^2 = \sigma_2^2.$	$F_0 = \frac{s_1^2}{s_2^2}.$	$F_0 \sim F_{n_1-1,n_2-1}.$

H_1	Rejection region
$\sigma_1^2 \neq \sigma_2^2$	$F_0 > f_{\alpha/2,n_1-1,n_2-1}$ or
	$F_0 < f_{1-\alpha/2,n_1-1,n_2-1}$
$\sigma_1^2 > \sigma_2^2$	$F_0 > f_{\alpha,n_1-1,n_2-1}$
$\sigma_1^2 < \sigma_2^20$	$F_0 < f_{1-\alpha,n_1-1,n_2-1}$

Variability in thickness

The variability in the thickness of oxide layers in semiconductor wafers is a critical characteristic, where low variability is desirable. A company is investigating two different ways to mix gases so as to reduce the variability of the oxide thickness. We produce 16 wafers with each gas mixture and our results indicate that the standard deviation is $s_1 = 1.96$Å and $s_2 = 2.13$Å for the two mixtures. Using $\alpha = 0.05$, is there evidence to indicate that either gas is preferable for better wafers?

We have two populations: i) population 1 with: $n_1 = 16, s_1 = 1.96$ and ii) population 2 with: $n_2 = 16, s_2 = 2.13$. As always, we begin by formulating our hypothesis as

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad H_1 : \sigma_1^2 \neq \sigma_2^2.$$

Proceed to calculate our test statistic, based on the sample variances as:

$$F_0 = \frac{s_1^2}{s_2^2} = \frac{1.96^2}{2.13^2} = 0.8467.$$

Since this is a two-sided hypothesis test, we need two critical values (recall that the F distribution is not symmetric!): $f_{\alpha/2,n_1-1,n_2-1} = f_{0.025,15,15} = 2.86$ and $f_{1-\alpha/2,n_1-1,n_2-1} = \frac{1}{f_{\alpha/2,n_2-1,n_1-1}} = \frac{1}{2.86} = 0.35$. Seeing as $f_{1-\alpha/2,n_1-1,n_2-1} \leq F_0 \leq f_{\alpha/2,n_1-1,n_2-1}$, we *fail to reject*: that is, we do not have enough evidence to claim that the two variances are not equal.

Hypothesis testing for the difference in the proportions of two populations

We finish our venture in two population hypothesis testing with proportions. It should come as no surprise that this also emulates the discussion of the two population proportions confidence interval we had seen earlier in the class! A few definitions and assumptions before we start:

1. *Large* samples from the two populations ($n_i p_i \geq 30$ and $n_1 (1 - p_i) \geq 30$ for both populations $i = 1, 2$).
2. sample size and observed proportion from population 1: $n_1, \hat{p}_1$, and sample size and observed proportion from population 2: $n_2, \hat{p}_2$.
3. a (hypothesized) difference $p_1 - p_2 = \Delta_0$.
4. a **pooled proportion estimator** in the form of

$$\hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}.$$

   - Much like a *weighted average* of the two observed proportions.

With these available, we may derive the hypothesis testing procedure as follows:

Proportions of two populations, p_1, p_2

Null hypothesis:	Test statistic:	Distribution:
$H_0 : p_1 - p_2 = \Delta_0.$	$Z_0 = \dfrac{(\hat{p}_1 - \hat{p}_2) - \Delta_0}{\sqrt{\hat{p}\,(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$	$\sim \mathcal{N}(0,1)\,.$

H_1	Rejection region	P-value
$p_1 - p_2 \neq \Delta_0$	$\lvert Z_0 \rvert > z_{\alpha/2}$	$2 \cdot (1 - \Phi(\lvert Z_0 \rvert))$
$p_1 - p_2 > \Delta_0$	$Z_0 > z_\alpha$	$1 - \Phi(Z_0)$
$p_1 - p_2 < \Delta_0$	$Z_0 < -z_\alpha$	$\Phi(Z_0)$

Politicians favorability ratings

A recent survey asked people in Urbana and Champaign whether they like their elected officials. Out of 118 Urbana residents, 37 said yes; for Champaign citizens there were 135 residents, among whom 61 said yes. Is there significant evidence (using $\alpha = 0.05$) to assume that Champaign's citizens showcase higher approval rates for their elected officials?

First collect our information:

- Urbana: $n_1 = 118, \hat{p}_1 = \frac{37}{118} = 0.314$.
- Champaign: $n_2 = 135, \hat{p}_2 = \frac{61}{135} = 0.452$.

We formulate the hypothesis as

$$H_0 : p_1 = p_2 \quad H_1 : p_1 < p_2.$$

Then, calculate the pooled proportion estimator as $\hat{p} = \frac{37+61}{118+135} = \frac{98}{253} = 0.387$. We are now ready to calculate the test statistic:

$$Z_0 = \frac{(\hat{p}_1 - \hat{p}_2) - \Delta_0}{\sqrt{\hat{p}\,(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \\ = \frac{0.138}{\sqrt{0.387 \cdot 0.613 \cdot (1/118 + 1/135)}} = 2.25.$$

To reject or fail to reject, we need the critical value for $\alpha = 0.05$:$z_\alpha = z_{0.05} = 1.64$. Because $Z_0 > z_\alpha$, we have to *reject the hypothesis*: hence, we deduce that Champaign does indeed have higher approval rates for elected officials (under $\alpha = 0.05$)!

Activity: Practicing hypothesis testing with real-life data

In this activity, you will use Python (specifically, numpy and pandas, as discussed in the lab sections) to reject or fail to reject a hypothesis on **whether masks work or not**. More specifically, this extra credit opportunity aims to give you the following tools:

1. Practice on formulating a hypothesis.
2. Use real data to prove or disprove claims.
3. Use pandas for real data wrangling and cleaning.
4. Use scientific communication to address real-life problems.

The case of Kansas

The state of Kansas adopted a state-wide mask mandate in early July. For the intents of this activity, we will set the time at July 15 (07/14 is the last day then before the mandate). However, not all counties accepted this mandate; on the contrary, many counties rescinded the mask mandate or did not enforce it. This includes 80 counties, given to you in a csv file. The other 25 counties (also presented in the csv file) adopted the mandate. In this extra credit activity I ask you to formulate and reject/fail to reject the following hypothesis:

> Counties that adopted the mask mandate ended up seeing a smaller (average) increase in cases than the counties that did not adopt it.

To put everything in perspective, we need to add the populations to the mix. Let p_i be the population of county i. Also assume that $\overline{X}_i^{before}$ and $\overline{X}_i^{after}$ are the average number of cases before and after 07/15. Now define $Y_i = \frac{\overline{X}_i^{after} - \overline{X}_i^{before}}{p_i}$.

Then, this activity asks you to compare the mean Y_i values for counties that rescinded the mask mandate versus counties that adopted the mask mandate. **Do masks work?** That is, using $\alpha = 0.05$, do you have enough evidence to reject the hypothesis that masks do not help in favor of the alternative hypothesis that masks help (in curbing the rate of increase of the number of cases)? All data is given in two files:

1. covid_KS.csv: a file that includes *all the cases* in Kansas per county from March 15 to November 14. Use the first part of the data (from March 15 to July 14) for the "before" period and the second part of the data (from July 15 to November 14) for the "after" period.
2. population_KS.csv: a file that includes all counties and their populations.

Enjoy! I will be there to help in any way I can.

Part 4: Lectures 30–34

Linear regression

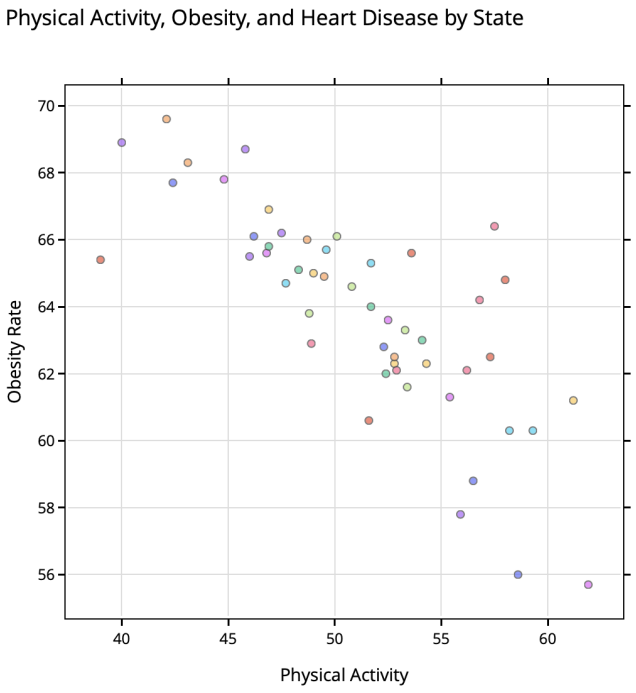
Learning objectives

After lectures 30-31, we will be able to:

- Explain the difference between regression and classification.
- Describe regression and linear regression.
- Derive, use, and interpret the results of the least squares line.
- Check whether a simple linear regression is significant or not.

Motivation: Physical activity and obesity

See below a figure representing the different levels of physical activity in each of the 50 states (x axis) and the resulting obesity rates (y axis). Do we see a relationship between activity and obesity? Is it linear? And, very importantly, is it significant?



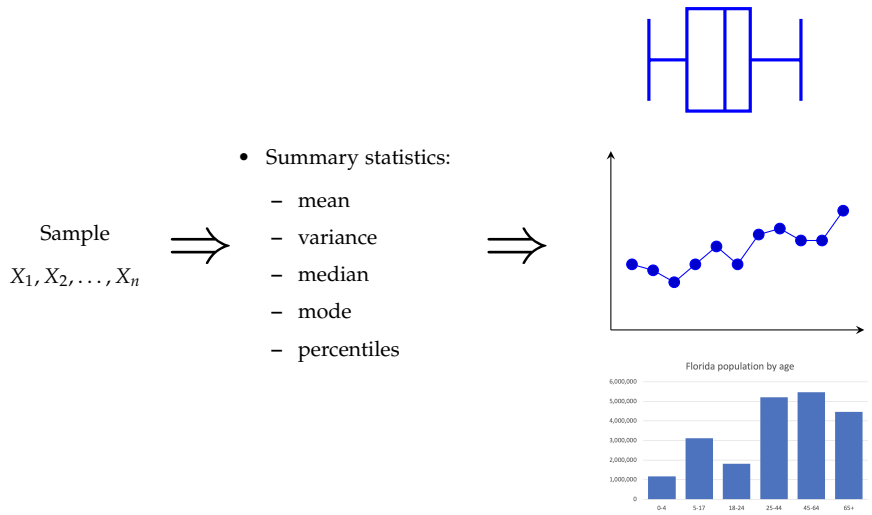
Motivation: Education level and income

Is there a relationship between the annual income of a person and their education level? And, if so, can we predict the income of a person before and after they have obtained a Master's degree?

Model building

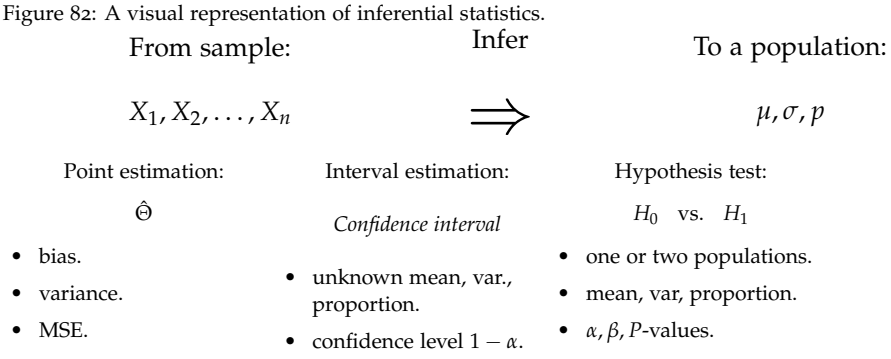
In the second part of the class, we saw **descriptive statistics**. Visually, see Figure 81, where from a sample we obtain a series of descriptive information (referred to as summary statistics), that we then present in a pictorial form (for ease of use and understanding).

Figure 81: A visual representation of descriptive statistics.



Then, in the third part of the class, we moved towards **inferential statistics**. Again, we represent this part visually in Figure 82.

Figure 82: A visual representation of inferential statistics.



There are three classifications of modern statistical methods:

1. **Descriptive statistics**: techniques to describe and visualize data.

2. **Inferential statistics**: techniques to draw conclusions for a large, unknown population based on observations of a smaller group (sample).

3. **Model building**: techniques to find relationships between data points, measure how strong these relationships are, and build models that can make predictions about the future.

In this last part of the class, we will focus on **model building**. Model building has three goals then:

- Goal #1: **investigate whether a relationship exists** between variables of our model.

  Does a relationship exist?

  - Do students perform better in tests that are in the morning or in the evening? Does time of day affect performance?
  - Does cold weather increase the number of accidents? Does the temperature affect driving patterns? Or do weather conditions, regardless of temperature, affect driving?
  - Does physical activity affect obesity rates? Does income affect obesity rates?

- Goal #2: **measure how strong the relationship is**.

  Strong relationship?

  - Obesity rates have been shown to depend on physical activity, income, age, education, built environment, etc.
  - Physical activity and age have been found to be more important.

- Goal #3: **predict an outcome given data**.

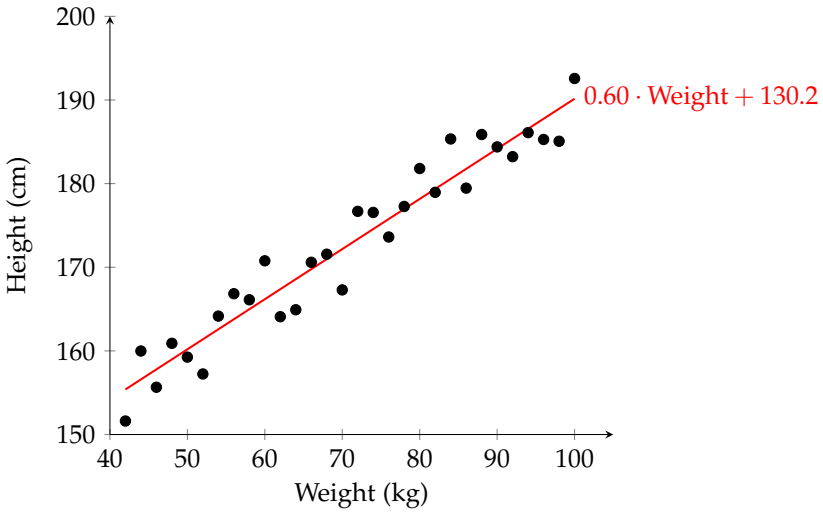
  Predicting

  - Given physical activity levels, predict the obesity rates for a specific state.
  - Given the weather conditions, predict the number of accidents at an interstate.

We define two types of models: **regression** and **classification**. They are visually contrasted in Figures 83 and 84. In the remainder of the semester, we will focus solely on regression.

1. **Regression**: for given values of *independent variables* x_i, predict the value of *dependent variable* y. Typically, regression applies to **continuous** y variables.

Figure 83: A possible regression line helping us predict the height of someone given their weight.



2. **Classification**: for given values of *independent variables* x_i, predict the class where *dependent variable* y belongs to. Typically, classification applies to **discrete** y variables.

In the remainder of today's lecture, we shall focus on regression models.

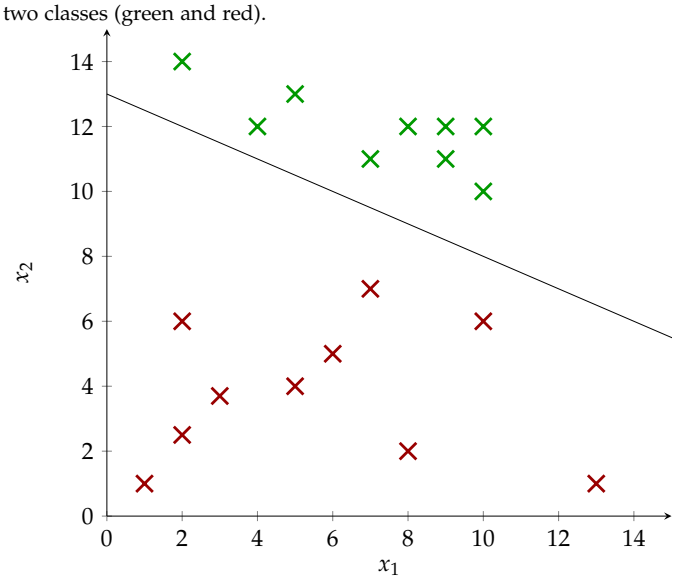
Linear regression

Before we begin with linear regression, a really quick overview of some necessary notation. We will assume the existence of two types of variables:

1. independent variables x: these may also be called predictor variables or regressors.
2. dependent variables y: sometimes also referred as response variables, outcome variables, or regressands.

Typically, independent variables are given to us in an attempt to predict the value of a dependent variable. Of course, this depends on the specifics of the problem we are tackling at each time!

Figure 84: An example of a classification problem. The line here separates our observations in two classes (green and red).



Independent vs. dependent variables

- Does the duration of a call (y) depend on the reception signal (x)?
- Does income (y) depend on years of education (x)?
- Does obesity rate (y) depend on income (x_1), days of physical activity per week (x_2), and age (x_3)?

As we note with the earlier example, it is not necessary to only have one independent variable x! Formally, we define regression as follows:

Definition 68 (Regression) *Regression is a statistical technique that is used to model the relationships between the response variable (also called the **dependent** variable) y and the predictor variables (also called the **independent** variables) x.*

We may define multiple types of regression:

1. Simple linear regression: one independent and one dependent variables tied together through a linear relationship.
2. Multiple linear regression: multiple independent and one dependent variables tied together using a linear relationship.
3. Polynomial regression: one or more independent and one dependent variables tied together using a polynomial relationship.

4. Logistic regression: one or more independent and one dependent variable tied together using any relationship. However now, the dependent variable takes on two discrete values (true or false, healthy or unhealthy, etc.). This is also called a dichotomous regression.

Simple linear regression

In simple linear regression, we want to express the dependent variable y as a linear function of the independent variable x. In mathematical terms, we are looking for coefficients β_0, β_1 such that:

$$y = \beta_0 + \beta_1 x.$$

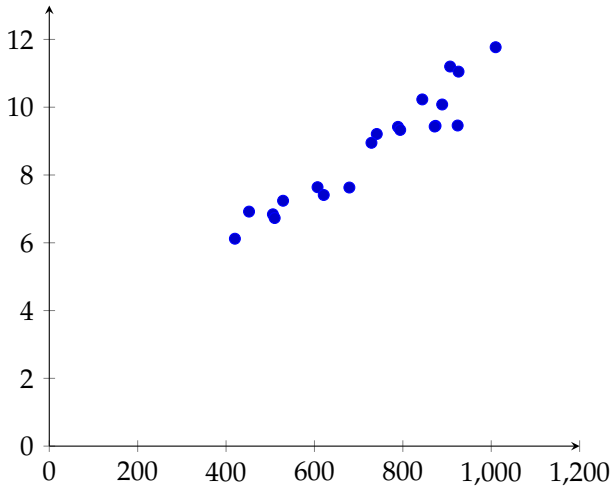
A webstore example

A webstore has collected the following data on the weekly visitors of the website and the profits from the past 20 weeks. They want to investigate that relationship and see whether they can direct more clicks towards their store. The data they have collected is as follows:

n	Visitors	Profit	n	Visitors	Profit
1	907	11.2	2	926	11.05
3	506	6.84	4	741	9.21
5	789	9.42	6	889	10.08
7	874	9.45	8	510	6.73
9	529	7.24	10	420	6.12
11	679	7.63	12	872	9.43
13	924	9.46	14	607	7.64
15	452	6.92	16	729	8.95
17	794	9.33	18	844	10.23
19	1010	11.77	20	621	7.41

What is the relationship between the profit and the number of visitors in their website?

Let's try this again, from a visual perspective...

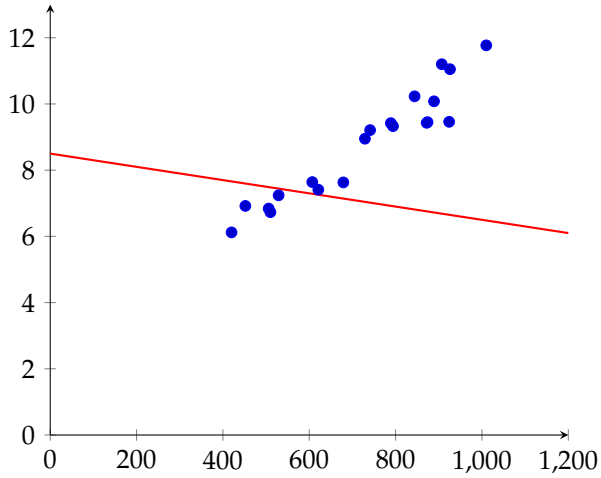


Some of the questions you may have already:

1. Do we see a relationship between profits and visits?
2. Does the relationship appear to be linear?
3. Does the relationship appear to be strong?
4. Can we predict profits based on the number of visitors?

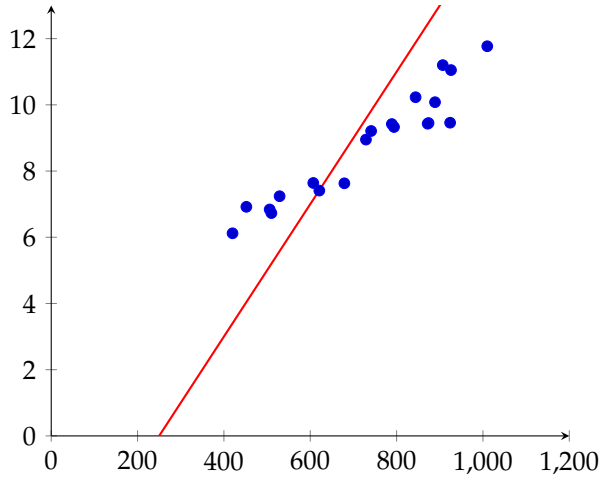
Our answers must have been Yes, Yes, Yes, and We sure hope so. Since there appears to be a linear relationship, what is the **best line** we can come up with to connect the dots? Let us try some and discuss why they work and why they do not work.

Line 1:



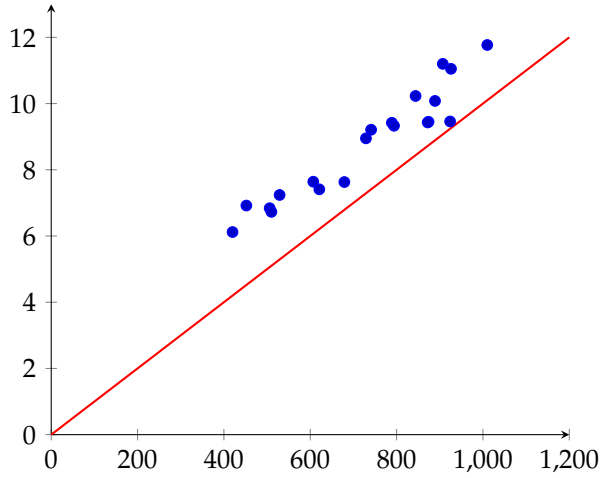
Bad line as it does not seem to capture the data provided.

Line 2:



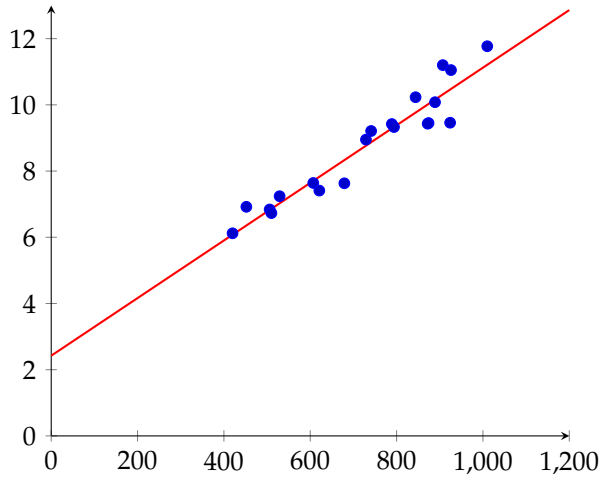
Better than before, but it still seems to **miss the "trend" of the data**, doesn't it?

Line 3:



This one seems to follow the trend, but is **underestimating the outcome** at each point...

Line 4:



The best fit line is the one that **minimizes the deviations of the data from the estimated regression line**.

Let's see what that means from a mathematical point of view. Based on our available data, we have n pairs of independent variables (x_i, y_i), for $i = 1, \ldots, n$. **If our line is correct**, then we should expect $y_i = \beta_0 + \beta_1 x_i$, no?

However, we recall that real life is not modeled exactly and neatly by a model, so maybe we can **incorporate some noise**? In that case, we now should get $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. In this last equation,$\beta_0$ is the **intercept**, β_1 is the **slope**; and ϵ_i is the noise related to data point (x_i, y_i).

In order for the quantity referred to as noise to make sense, we need to make some assumptions. Namely, we have for all noises ϵ_i that:

- they are independent normally distributed random variables;
- with zero mean;
- and with the same variance;
- $\epsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$.

Let us consider the total "error". What could this mean? We could potentially define it as:

- the sum of all errors (positive or negative): $L = \sum_{i=1}^{n} \epsilon_i$.
- the sum of all absolute errors: $L = \sum_{i=1}^{n} |\epsilon_i|$.
- the sum of all squared errors: $L = \sum_{i=1}^{n} \epsilon_i^2$.

The last one is called the **least squares** error. Recall that

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \implies \epsilon_i = y_i - \beta_0 - \beta_1 x_i.$$

Hence, we may derive for the least squares error:

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

A quadratic term! And one that we need to minimize in order to identify the least squares line. What are our unknowns? Those would be the slope and the intercept, β_0 and β_1. And what are our known parameters? Of course all the pairs (x_i, y_i) for all $i = 1, \ldots, n$ known data points.

Finally, how can we minimize $L = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$? We could take the derivative for each of the unknowns and equate to zero, leading to:

$$\frac{\partial L}{\partial \beta_1} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) x_i = 0 \implies$$

$$\implies \boxed{\hat{\beta}_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}}$$

$$\frac{\partial L}{\partial \beta_0} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0 \implies$$

$$\implies \boxed{\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}}$$

Practice with least squares

Earlier, we saw a webstore and part of the data they had collected about the number of visitors and their profits. As a reminder, here is the table with the data again:

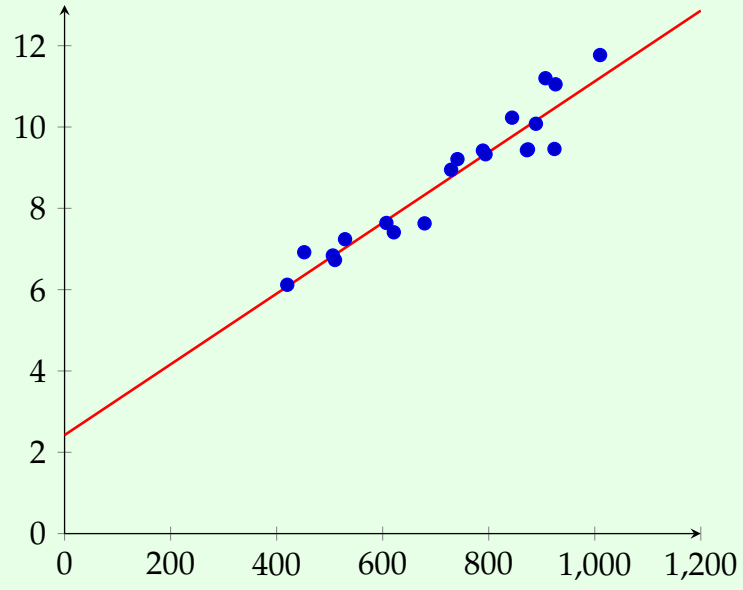
n	Visitors	Profit	n	Visitors	Profit
1	907	11.2	2	926	11.05
3	506	6.84	4	741	9.21
5	789	9.42	6	889	10.08
7	874	9.45	8	510	6.73
9	529	7.24	10	420	6.12
11	679	7.63	12	872	9.43
13	924	9.46	14	607	7.64
15	452	6.92	16	729	8.95
17	794	9.33	18	844	10.23
19	1010	11.77	20	621	7.41

What is the least squares line?

First, calculate $\sum x_i = 907 + 506 + \ldots = 14623, \sum y_i = 11.2 + 6.84 + \ldots = 176.11, \sum x_i y_i = 907 \cdot 11.2 + 506 \cdot 6.84 + \ldots = 134127.9, \sum x_i^2 = 907^2 + 506^2 + \ldots = 11306209$.

- $\hat{\beta}_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = 0.0087.$
- $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = 8.8055 - 0.0087 \cdot 731.15 = 2.423.$

Or, visually:



How can we use the regression line to help predict outcomes? Well, for a given value x, we may now predict y by plugging x in the regression line formula..

Using the regression line

For the previous webstore, how many profits should they anticipate on a very good day with 1200 visitors?

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 2.423 + 0.0087 \cdot 1200 = 12.863.$$

From now on, we will use the following terminology:

1. *observed values*:
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$
where (x_i, y_i) are the pairs of independent and dependent variables and ϵ_i the noise for $i = 1, \ldots, n$.

2. *fitted values*:
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i,$$
where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the intercept and slope.

3. *residuals/errors*:
$$e_i = y_i - \hat{y}_i,$$
the difference between the observed and the fitted dependent values.

4. *sum of squares of errors*:
$$SS_E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Significance of simple linear regression

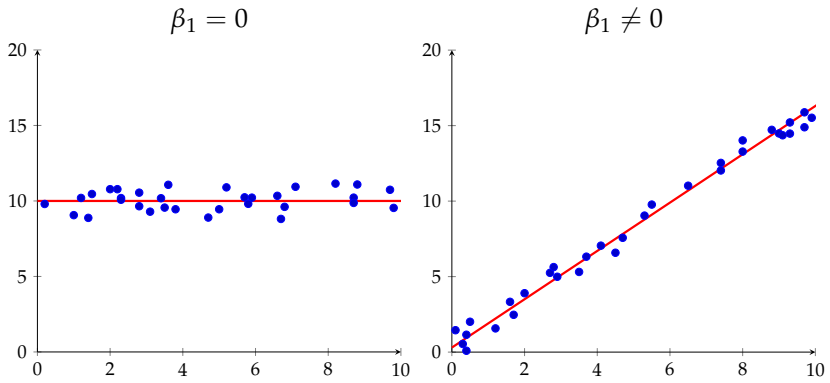
We got our intercept and slope; but is the regression line we got **significant**. What does that mean? What we want to ask is: "is there enough evidence to suggest that x and y are related?" Or does it appear to be just a random phenomenon, a coincidence?

Well, every time we want to check if we have enough evidence to "reject" something, we need *hypothesis testing*. When are x and y unrelated? When $\beta_1 = 0$! So, this is what we will formulate a hypothesis for.

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0.$$

An example of what this looks like is presented below in Figure 85.

Figure 85: An example of an insignificant regression (left), where the slope is 0, and an example of a significant regression (right), where the slope is non-zero.



Before we proceed with this, let us redefine the slope calculations. This will come in handy later. We have:

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}.$$

If we define:

- $S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$
- $S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2$

then we may get that:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}.$$

So, how is $\hat{\beta}_1$ distributed as? Recall that $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. We then have that:

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}\right) \to \mathcal{N}\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

Unfortunately, σ^2 is not known – we will need some way to estimate it. Luckily, there is an easy to calculate estimator. We will need to keep track of the following notions:

- Recall that a sample variance can be calculated as $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}$, where $n-1$ are the degrees of freedom as we needed to estimate one parameter in the calculation.

- In our case, we want to compare y_i to the average y value. $SS_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$. However, it comes with $n-2$ degrees of freedom as we needed to estimate two parameters in its calculation ($\hat{\beta}_0, \hat{\beta}_1$).
- Hence, we may use $\frac{SS_E}{n-2}$ as an estimator for σ^2!

This last quantity is called the **mean square error**:

$$MS_E = \frac{SS_E}{n-2}$$

and we can show that

$$E[MSE] = \sigma^2,$$

which serves to show that it is an unbiased estimator for our unknown variance:

$$\hat{\sigma}^2 = MS_E.$$

Finally, we are ready to pose the hypothesis test for the significance of our regression.

Simple linear regression significance

Null hypothesis:	Test statistic:	Distribution:
$H_0 : \beta_1 = 0.$	$T_0 = \frac{\hat{\beta}_1}{\sqrt{MS_E/S_{xx}}}.$	$T_0 \sim T_{n-2}.$

H_1	Rejection region	CI region
$\beta_1 \neq 0$	$\lvert T_0 \rvert > t_{\alpha/2,n-2}$	$\left[\hat{\beta}_1 - t_{\alpha/2,n-2}\sqrt{\frac{MS_E}{S_{xx}}}, \hat{\beta}_1 + t_{\alpha/2,n-2}\sqrt{\frac{MS_E}{S_{xx}}}\right]$
$\beta_1 > 0$	$T_0 > t_{\alpha,n-2}$	$\left(\infty, \hat{\beta}_1 + t_{\alpha,n-2}\sqrt{\frac{MS_E}{S_{xx}}}\right]$
$\beta_1 < 0$	$T_0 < -t_{\alpha,n-2}$	$\left[\hat{\beta}_1 - t_{\alpha,n-2}\sqrt{\frac{MS_E}{S_{xx}}}, +\infty\right)$

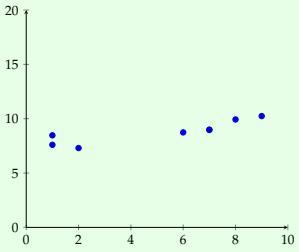
Note that this hypothesis test can be easily adapted to test for any value (not just zero!). How?

Let us put this to the test.

Is the regression significant?

Consider the following points:

x	y
1	7.6
9	10.24
2	7.3
7	8.97
6	8.74
7	8.99
8	9.93
1	8.47



1. Calculate $\hat{\beta}_0, \hat{\beta}_1$.
2. Using $\alpha = 0.10$, is there significant evidence that $\beta_1 \neq 0$?
3. Build a 90% confidence interval around $\hat{\beta}_1$.

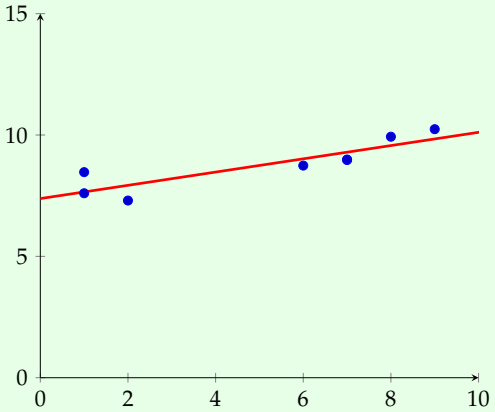
We'll again need to calculate: $n = 8, \sum x_i = 41, \sum y_i = 70.24, \sum x_i y_i = 380.43, \sum x_i^2 = 285$.
First, to calculate $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = 0.273.$$

Now, we can calculate $\hat{\beta}_0$:

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = \frac{70.24}{8} - 0.273 \cdot \frac{41}{8} = 7.381.$$

Overall: $\hat{y} = 7.381 + 0.273 \cdot \hat{x}$.



Is the regression significant?

Recall that for our hypothesis test, we will need an estimator of the variance of the error σ^2..

- $\hat{\sigma}^2 = \frac{SS_E}{n-2}$.

To calculate SS_E, consider the original data, and append a new column (called $\hat{y}$). Populate it with the result $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i$:

x	y	$\hat{y}$
1	7.6	7.654
9	10.24	9.838
2	7.3	7.927
7	8.97	9.292
6	8.74	9.019
7	8.99	9.292
8	9.93	9.565
1	8.47	7.654

Finally,

$$SS_E = \sum (y_i - \hat{y}_i)^2 = 1.629$$

and hence $\hat{\sigma}^2 = \frac{1.629}{6} = 0.272$.
We finally move to the hypothesis testing part.

$$H_0 : \beta_1 = 0 \qquad H_1 : \beta_1 \neq 0.$$

- $T_0 = \frac{\hat{\beta}_1 - 0}{\sqrt{MS_E / S_{xx}}} = \frac{0.273}{\sqrt{0.272/74.875}} = 4.529$, where $S_{xx} = \sum (x_i - \overline{x})^2 = 74.875$.
- Compare to $t_{0.05,6} = 1.943$.
- Because $|T_0| > 1.943$, we reject the null hypothesis and deduce that with 90% confidence $\beta_1 \neq 0$.

Also note that

$$\beta_1 \in [0.273 - 1.943 \cdot 0.06, 0.273 + 1.943 \cdot 0.06] = [0.156, 0.390] .$$

Wait.. So does that mean that we can also use hypothesis testing to check whether $\hat{\beta}_1$ (the slope) has a certain value or not? The answer is a resounding yes!

Simple linear regression slope testing

Null hypothesis:	Test statistic:	Distribution:
$H_0 : \beta_1 = \beta_{10}.$	$T_0 = \dfrac{\hat{\beta}_1 - \beta_{10}}{\sqrt{MS_E / S_{xx}}}.$	$T_0 \sim T_{n-2}.$

H_1	Rejection region	CI region
$\beta_1 \neq \beta_{10}$	$\lvert T_0 \rvert > t_{\alpha/2,n-2}$	$\left[\hat{\beta}_1 - t_{\alpha/2,n-2}\sqrt{\frac{MS_E}{S_{xx}}}, \hat{\beta}_1 + t_{\alpha/2,n-2}\sqrt{\frac{MS_E}{S_{xx}}}\right]$
$\beta_1 > \beta_{10}$	$T_0 > t_{\alpha,n-2}$	$\left(\infty, \hat{\beta}_1 + t_{\alpha,n-2}\sqrt{\frac{MS_E}{S_{xx}}}\right]$
$\beta_1 < \beta_{10}$	$T_0 < -t_{\alpha,n-2}$	$\left[\hat{\beta}_1 - t_{\alpha,n-2}\sqrt{\frac{MS_E}{S_{xx}}}, +\infty\right)$

A different perspective

For the previous example we have hypothesized that the line is $7.381 + 0.273 \cdot \hat{x}$. New data come in and give us the following four points: $(7, 9.97), (2, 7.95), (5, 8.91), (5, 8.14)$.

Using $\alpha = 0.05$, is there enough evidence in the new data to suggest that the slope has changed and we now have $\beta_1 > 0.273$?

Again we may calculate (for the new set of points) that: $n = 4, \sum x_i = 19, \sum y_i = 34.97, (\sum x_i) \cdot (\sum y_i) = 664.43, \sum x_i y_i = 170.94, \sum x_i^2 = 103, (\sum x_i)^2 = 361$. This leads to:

- $\hat{\beta}_1 = \frac{4 \cdot 170.94 - 664.43}{4 \cdot 103 - 361} = \frac{35.61}{51} = 0.379.$
- $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = 6.942.$

We then have $(y_i - \hat{y}_i)^2 = (y_i - 6.942 - 0.379 \cdot x_i)^2$, which leads to:

- $(y_1 - \hat{y}_1)^2 = 0.140$
- $(y_2 - \hat{y}_2)^2 = 0.062$
- $(y_3 - \hat{y}_3)^2 = 0.005$
- $(y_4 - \hat{y}_4)^2 = 0.486$

This finally gives $SS_E = 0.694$ and a $\hat{\sigma} = \sqrt{MS_E} = \sqrt{\frac{SS_E}{n-2}} = \sqrt{\frac{0.694}{2}} = 0.589$. We are ready to formulate our hypothesis:

$$\boxed{H_0 : \beta_1 = 0.273 \quad H_1 : \beta_1 > 0.273}$$

A different perspective

On to our hypothesis testing calculations:

- **the test statistic**: $T_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\hat{\sigma}/\sqrt{S_{xx}}} = \frac{0.106}{0.589/\sqrt{12.75}} = 0.643$, where $S_{xx} = \sum (x_i - \overline{x})^2 = 12.75$.
- **the critical value**: $t_{0.05,2} = 2.92$. Recall that the hypothesis is one-sided here.
- **the comparison**: we have that $T_0 \quad < \quad t_{\alpha,n-2}$, which means that we accept the null hypothesis.

Hence, we deduce that with 95% confidence β_1 is still equal to 0.273 (even with the new data suggesting otherwise). Also note that

$$\beta_1 \in (-\infty, 0.273 + 2.92 \cdot 0.165] = (-\infty, 0.755],$$

which further reinforces that the new data should be even more indicative of a change (result in $\hat{\beta}_1 > 0.755$) to accept the change.

So.. this is how it works in simple linear regression with one dependent and one independent variable. How about we generalize this to more than just one independent variable? More on that, next time!

Multiple linear regression

Learning objectives

After lecture 32, we will be able to:

- Recall the ANOVA identity.
- Recall and use the R^2 and R^2_{adj} parameters to evaluate how good a regression is.
- Understand when to use and how to apply regression with multiple independent variables.
- Derive, use, and interpret the results of the last squares line for multiple independent variables.
- Perform hypothesis testing on multiple parameters of the least squares line.

Motivation: Maintenance fees

What happens when we are trying to derive a (linear) relationship between one dependent variable y and multiple $k > 1$ different independent variables x_j? Well, in that case, we need multiple different parameters (slopes), one for each independent variable!

For example, what is the linear relationship between the maintenance fees (costs y) of a bank as a function of the number of the new applications (x_1) and the number of outstanding loans (x_2)?

Motivation: realtor.com

Taken from , here are 8 recently (August 2019) sold homes in Urbana:

	Sq. ft.	Year built	Garages	#bedrooms	#bathrooms	Price
1	1547	1950	1	3	3	158500
2	1834	1957	0	4	2	183000
3	2520	1980	3	5	2.5	233000
4	985	1911	1	2	1	69000
5	1275	1968	0	3	1.5	118000
6	2337	1977	2	5	2	249900
7	1880	1967	2	3	2	175000
8	1943	1965	1	4	2.5	169900

Which one of the five predictor variables (sq. ft., year built, garages, #bedrooms, #bathrooms) is the least important for predicting price?

The ANOVA identity

Let us begin with an example of the calculations we will see in this subsection. During the previous lecture, we saw an example that led us to a regression line of $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 7.381 + 0.273 \cdot x$

For the given data (again check the previous lecture for all the details of this example), we finally got:

x	y	$\hat{y}$
1	7.6	7.654
9	10.24	9.838
2	7.3	7.927
7	8.97	9.292
6	8.74	9.019
7	8.99	9.292
8	9.93	9.565
1	8.47	7.654

We used that table to calculate SS_E (**the sum of squares of the error**) as:

$$SS_E = \sum (y_i - \hat{y}_i)^2 = (7.6 - 7.654)^2 + (10.24 - 9.838)^2 + \ldots = 1.629.$$

This would eventually be divided by $8 - 2 = 6$ degrees of freedom to estimate the **mean square error** (MS_E).

In a similar manner, we may define **total sum of squares** as the sum of squares of the differences between each *observed* value y_i versus the expectation:

$$SS_T = \sum (y_i - \overline{y})^2 .$$

We may also define the **regression sum of squares** as the sum of squares of the differences between each *fitted* value $\hat{y}_i$ versus the expectation:

$$SS_R = \sum (\hat{y}_i - \overline{y})^2 .$$

We then claim that:

$$\boxed{SS_T = SS_E + SS_R}$$

This is called the **Analysis of Variance** (ANOVA) identity and it is immensely useful when analyzing how good our regression is.

Using the ANOVA identity

In this example, we have already calculated the sum of squares of errors SS_E to be equal to 1.629. How about the total sum of squares SS_T and the regression sum of squares SS_R?

First, begin by calculate the average y value as

$$\overline{y} = \frac{\sum y_i}{n} = \frac{7.6 + 10.24 + \ldots + 8.47}{8} = \frac{70.24}{8} = 8.78.$$

Then:

- $SS_T = \sum (y_i - \overline{y})^2 = (7.6 - 8.78)^2 + (10.24 - 8.78)^2 + \ldots + (8.47 - 8.78)^2 = 7.2148$.
- Using the ANOVA identity: $SS_T = SS_R + SS_E \implies SS_R = SS_T - SS_E = 7.2148 - 1.629 = 5.5858$.

Note how we could have derived SS_R by applying the formula and getting that $SS_R = \sum (\hat{y}_i - \overline{y})^2 = (7.654 - 8.78)^2 + (9.838 - 8.78)^2 + \ldots + (7.654 - 8.78)^2 = 5.5858$, which is the same result.

We now proceed to define an easy to compute parameter that helps us estimate the quality of our regression line.

The R^2 parameter

We want to somehow quantify how "good" a regression is. We would like to establish some coefficient that tells us how closely our predictions $\hat{y}$ follow the real data (y). We call that parameter R^2 and allow it to be in $[0, 1]$m where a value of 1 implies that all data points fall on the regression line. Of course, we would like high values of R^2 and we hope that they imply a good fit of the regression line. Formally:

Definition 69 *R^2 is a measure of how much of the variability is accounted for by the regression model and is calculated as:*

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}.$$

Recall that:

total: $SS_T = \sum (y_i - \overline{y})^2$.

error: $SS_E = \sum (y_i - \hat{y}_i)^2$.

regression: $SS_R = \sum (\hat{y}_i - \overline{y})^2$.

R^2 calculations

What is the R^2 coefficient for the previous regression?

We have two ways to calculate it!

- $R^2 = \frac{SS_R}{SS_T} = 5.58587.2148 = 0.774.$
- $R^2 = 1 - \frac{SS_E}{SS_T} = 1 - \frac{1.629}{7.2148} = 0.774.$

So, *how high is good enough* for R^2? The answer is that (as so many other things that we have seen) "it depends!" We'll take another look at it (and an adjusted version) shortly.

Multiple linear regression

We now move to more than just one independent variable x. This should make sense, as in most practical cases our "future" depends on more than just one piece of information:

- Success in an exam is not only how much you've studied, but also a function of your physical and mental health, how well rested you are, luck, etc.
- The box office success of a movie is not only how good the movie is, but how much budget they've had for advertising, the recognition of the names starring and directing, etc.
- Any more examples?

Let us begin easy with just two predictor variables x_1, x_2. We need to extend our definitions from the simple case:

- We now have a triple [82] (x_{i1}, x_{i2}, y_i), $i = 1, \ldots, n$, that is a series of n data points with provided values for x_1, x_2, y.
- The main idea is still the same!

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i,$$

  where:

  - β_0 is the intercept intercept;
  - β_1, β_2 are the slopes for x_1, x_2, respectively;
  - ϵ_i is the "noise" associated with point i.

- Hence our goal is to find the "best" $\beta_0, \beta_1, \beta_2$ by optimizing the least squares function:

$$L = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2}\right)^2.$$

[82] Contrast with the pair (x_i, y_i) earlier.

How to derive a solution here? Like earlier, we can take the proper derivatives and set them to zero! How many derivatives, though? Well, in this case, we need to take three derivatives:

$$\frac{\partial L}{\partial \beta_0} = 0 \implies -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}\right) = 0$$
$$\frac{\partial L}{\partial \beta_1} = 0 \implies -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}\right) x_{i1} = 0$$
$$\frac{\partial L}{\partial \beta_2} = 0 \implies -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}\right) x_{i2} = 0$$

Or, simplifying:

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} = \sum_{i=1}^{n} y_i$$
$$\hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1}x_{i2} = \sum_{i=1}^{n} y_i x_{i1}$$
$$\hat{\beta}_0 \sum_{i=1}^{n} x_{i2} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}x_{i2} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2}^2 = \sum_{i=1}^{n} y_i x_{i2}$$

This is a system of equations with three unknowns and three equations; solvable under certain conditions. However, it is much more easily expressed in matrix form, no? Let us go back to the original regression line equation

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i.$$

Written in **matrix form**, we have:

$$y = X\beta + \epsilon$$

- $y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$, $X = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{bmatrix}$, $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$, $\epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$

Once more, we wish to find $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ such that

$$L = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i 1 - \beta_2 x_{i2}\right)^2 = (y - X\beta)^T (y - X\beta)$$

is *minimized*. We may rewrite L as:

$$\begin{aligned} L &= (y - X\beta)^T (y - X\beta) = \\ &= y^T y - \beta^T X^T y - y^T X\beta + \beta^T X^T X\beta = \\ &= y^T y - 2\beta^T X^T y + \beta^T X^T X\beta \end{aligned}$$

We need to take the derivative as far as vector β is concerned:

$$\frac{\partial L}{\partial \beta} = 0 \implies -2X^T y + 2X^T X\beta = 0 \implies X^T X\beta = X^T y.$$

This last equality can be solved by taking the inverse $\left(X^T X\right)^{-1}$ and multiplying on the left [83] to obtain:

$$\boxed{\hat{\beta} = \left(X^T X\right)^{-1} X^T y}$$

[83] Why is that? Well, recall that $Ax = b$ can be solved as $x = A^{-1}b$, when matrix A is invertible!

Overall, we have shown that *in general* (not only for two predictor variables, but for as *many* as we would like to), we have $\hat{\beta} = \left(X^T X\right)^{-1} X^T y$, which can be used

- in matrix form:
$$\hat{y} = X\hat{\beta},$$
- or in scalar form:
$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_{ij}, \quad \text{for all } i = 1, \ldots, n.$$

Like in simple linear regression $e_i = \hat{y}_i - y_i$ is the residual/error for each observation i.

Bank maintenance fee prediction

A small bank is hypothesizing that a lot of the fees they pay have to do with the number of loan applications they process every month as well as the number of outstanding loans they have going on. More specifically, they have collected data over the last 16 months that are presented in the following table.

What is the regression line they should use? How much money should they budget for their maintenance costs if they expect 100 applications and 13 outstanding loans this coming January?

Bank maintenance fee prediction

# Applications	# Outstanding	Cost
80	8	2256
93	9	2340
100	10	2426
82	12	2293
90	11	2330
99	8	2368
81	8	2250
96	10	2409
94	12	2364
93	11	2379
97	13	2440
95	11	2364
100	8	2404
85	12	2317
86	9	2309
87	12	2328

First, build matrix X and calculate $\left(X^T X\right)^{-1}$:

$$X = \begin{bmatrix} 1 & 93 & 9 \\ 1 & 100 & 10 \\ 1 & 82 & 12 \\ 1 & 90 & 11 \\ 1 & 99 & 8 \\ 1 & 81 & 8 \\ 1 & 96 & 10 \\ 1 & 94 & 12 \\ 1 & 93 & 11 \\ 1 & 97 & 13 \\ 1 & 95 & 11 \\ 1 & 100 & 8 \\ 1 & 85 & 12 \\ 1 & 86 & 9 \\ 1 & 87 & 12 \end{bmatrix}, \quad \left(X^T X\right)^{-1} = \begin{bmatrix} 14.176 & -0.130 & -0.223 \\ -0.130 & 1.429 \cdot 10^{-3} & -4.764 \cdot 10^{-5} \\ -0.223 & -4.764 \cdot 10^{-5} & 2.222 \cdot 10^{-2} \end{bmatrix}.$$

Finally, we calculate $X^T y$:

$$X^T y = \begin{bmatrix} 37577 \\ 3429550 \\ 385562 \end{bmatrix}.$$

Bank maintenance fee prediction

Combining all we get:

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T y = \begin{bmatrix} 1566.077 \\ 7.62 \\ 8.58 \end{bmatrix}.$$

This in turn gives us the regression line as:

$$\hat{y} = 1566.077 + 7.62 \cdot \#\text{new loans} + 8.58 \cdot \#\text{loans outstanding}.$$

For January then, we should expect to pay:

$$\hat{y}_{Jan} = 1566.077 + 7.62 \cdot 100 + 8.58 \cdot 13 = 2439.62.$$

The question we should be thinking about at this point: *does the ANOVA identity still hold?* And how can we use that to do hypothesis testing for the regression significance? While we are at it, what does regression significance mean for more than one predictor variables? Let us go ahead and answer all of these questions in the remainder of the lecture.

The ANOVA identity still holds:

$$SS_T = SS_R + SS_E.$$

Each of the three sum of squares is calculated the same way as before. The difference lies with the degrees of freedom:

- SS_T: $n-1$ degrees of freedom [84].
- SS_R: k degrees of freedom.
- SS_E: $n-k-1$ degrees of freedom [85].

Due to that, the mean squares are changed and are now equal to:

- MS_T: $\frac{SS_T}{n-1}$.
- MS_R: $\frac{SS_R}{k}$.
- MS_E: $\frac{SS_E}{n-k-1}$.

Now, back to the derivations from the previous class. We wanted to come up with an estimate for the (unknown!) noise standard deviation σ. We came up with:

$$\hat{\sigma}^2 = MS_E = \frac{SS_E}{n-2}.$$

[84] The same as before.

[85] Different, as we are now estimating $k+1$ parameters. What are those? They are the regression line intercept and slopes: $\beta_0, \beta_1, \ldots, \beta_k$.

Hopefully, you see where we are going with this: our MS_E is different, but other than that the derivation holds. Hence we estimate this standard deviation as:

$$\hat{\sigma}^2 = MS_E = \frac{SS_E}{n-k-1},$$

where the sum of squares of error is calculated as $SS_E = \sum (y_i - \hat{y}_i)^2$ or, in matrix form, as $SS_E = y^T y - \hat{\beta}^T X^T y$.

On to the significance of the regression. Recall that for a single predictor variable our significance testing was easy: either $\beta_1 = 0$ (the slope was zero, and hence insignificant) or not (the slope was nonzero and hence it is significant). When dealing with more than just one predictor variable, though, then all of them need to have zero slopes for the regression to be insignificant! This leads us to:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0 \quad \text{vs.} \quad H_1 : \beta_j \neq 0, \text{ for at least one } j.$$

We now make the observation that **if the null hypothesis is true**, then the mean squares of the regression and the error are distributed following a χ^2 distribution, each with their own degrees of freedom:

- $SS_R/\sigma^2 \sim \chi^2_k$, where $SS_R = \sum (\hat{y}_i - \overline{y})^2$
- $SS_E/\sigma^2 \sim \chi^2_{n-k-1}$, where $SS_E = \sum (y_i - \hat{y}_i)^2$.

We are then comparing two population "variances" (for MS_R and MS_E) and the test statistic for that is:

$$F_0 = \frac{SS_R/k}{SS_E/(n-k-1)} = \frac{MS_R}{MS_E}$$

The rejection area is if $F_0 > f_{\alpha,k,n-k-1}$. Some software will also return a P-value, and the rejection criterion is simply whether P-value $< \alpha$.

Multiple linear regression significance

Null hypothesis:	Test statistic:	Distribution:
$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0.$	$F_0 = \frac{MS_R}{MS_E}.$	$F_0 \sim F_{k,n-k}.$

H_1	Rejection region
At least one $\beta_j \neq 0$	$F_0 > f_{\alpha,k,n-k-1}$

Finally, recall R^2: we have some unfinished business. We already defined $R^2 = 1 - \frac{SS_E}{SS_T}$. We make two observations about it:

- Observation #1: R^2 will always increase or stay the same with the addition of any predictor variable.
- Observation #2: This happens even when that predictor variable is associated with a β_j that is insignificant (i.e., the slope is zero).

We hence define an *adjusted R^2 model*, called R^2_{adj}, that will **penalize more complex regressions** (that is, the use of more predictor variables). Its definition?

$$R^2_{adj} = 1 - \frac{SS_E/(n-k-1)}{SS_T/(n-1)}.$$

Note how adding more predictor variables will lead to a bigger numerator in the fraction which in turn will cause R^2_{adj} to go down.

We claim that this adjusted version is more appropriate than the simple version of R^2. Why? Well, primarily because it does not necessarily increase with the addition of new predictor variables, and thus will not favor more complex models. Indeed, it will many times decrease when an insignificant variable is entered. When R^2 and R^2_{adj} differ by a lot, this is an indication that insignificant terms have been added.

Let us put these things to the test in an example on the regression line we got earlier in the bank example.

Testing significance

In the previous bank example, we already found the line as

$$\hat{y} = 1566.077 + 7.62 \cdot \#\text{new loans} + 8.58 \cdot \#\text{loans outstanding}.$$

Is the regression significant using $\alpha = 0.05$? What is R^2 and how does it compare with R^2_{adj}?
We begin with the calculations of the sum of squares:

- $SS_E = \sum_{i=1}^{16} (y_i - \hat{y}_i)^2 = 3479$
- $SS_R = \sum_{i=1}^{16} (\hat{y}_i - \overline{y})^2 = 44157$
- Using ANOVA, $SS_T = SS_R + SS_E = 47636$.

Now, on to calculate the ratio of the two mean squares:

$$F_0 = \frac{MS_R}{MS_E} = \frac{SS_R/2}{SS_E/13} = 82.5$$

Compared to $f_{\alpha,k,n-k-1} = f_{0.05,2,13} = 3.81$, we overwhelmingly reject. The regression is significant! Let us look at the two R^2 parameter calculations:

- $R^2 = 1 - \frac{SS_E}{SS_T} = 1 - 3479/44157 = 0.921$.
- $R^2_{adj} = 1 - \frac{SS_E/(n-k-1)}{SS_T/(n-1)} = 0.916$.

Note how close the two values are, an indication that no insignificant terms have been added.

What if we were interested in each individual coefficient one-by-one? That is, what if we wanted to check whether the number of new loans is significant; or whether the number of outstanding loans is significant? First of all, let us address why this is not the same question as the one we saw how to address earlier.

Consider a regression with k predictor variables: $k-1$ of them could be insignificant, and one of them could be very significant! Then, the regression as a whole is also significant. Because of that, it is a different question whether the whole regression is significant compared to whether each individual independent variable is significant.

So, if we are interested in whether a single variable is significant or not, this reverts back to checking whether the corresponding slope is zero or not.

$$H_0 : \beta_j = 0, \quad H_1 : \beta_j \neq 0.$$

The test statistic is the same as for simple linear regression:

$$T_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 \cdot C_{jj}}}$$

- where C_{jj} is the j-th [86] diagonal element of $\left(X^T X\right)^{-1}$,
- and $\hat{\sigma}^2 = MS_E = \frac{SS_E}{n-k-1}$.

[86] We assume here that the first row and first column element is C_{00}, i.e., we start counting from zero.

Finally, reject if $|T_0| > t_{\alpha/2,n-k-1}$. Note how the main difference from the simple linear regression to the multiple linear regression comes in the form of C_{jj} which replaces S_{xx}. [87] Let us put this to the test right away.

[87] See Lecture 30-31 for details on S_{xx}.

Multiple linear regression term single significance

Null hypothesis:	Test statistic:	Distribution:
$H_0 : \beta_j = 0.$	$T_0 = \frac{\hat{\beta}_j}{\sqrt{MS_E \cdot C_{jj}}}.$	$T_0 \sim T_{n-k-1}.$

H_1	Rejection region	CI region
$\beta_j \neq 0$	$\lvert T_0 \rvert > t_{\alpha/2,n-k-1}$	$\left[\hat{\beta}_j - t_{\alpha/2,n-k-1}\sqrt{\frac{MS_E}{C_{jj}}}, \hat{\beta}_j + t_{\alpha/2,n-k-1}\sqrt{\frac{MS_E}{C_{jj}}}\right]$

Testing significance one-by-one

Let us go back to the banking example from earlier. We already have that the regression line can be written as

$$\hat{y} = 1566.077 + 7.62 \cdot \text{\#new loans} + 8.58 \cdot \text{\#loans outstanding}.$$

- Is the number of new loans significant?
- Is the number of loans outstanding significant?

Use $\alpha = 0.05$. Recall that we already know that the regression is significant; again, though, this does not necessarily imply that both of them are significant!

Testing significance one-by-one

First of all, recall that:

- $SS_E = \sum_{i=1}^{16} (y_i - \hat{y}_i)^2 = 3479$
- $\hat{\sigma}^2 = MS_E = \frac{SS_E}{13} = 267.62.$

For $\hat{\beta}_1$ (number of new loans):

- We have $\left(X^T X\right)^{-1} = \begin{bmatrix} 14.176 & -0.130 & -0.223 \\ -0.130 & 1.429 \cdot 10^{-3} & -4.764 \cdot 10^{-5} \\ -0.223 & -4.764 \cdot 10^{-5} & 2.222 \cdot 10^{-2} \end{bmatrix}$.
- So..

$$C_{11} = 1.429 \cdot 10^{-3}.$$

Combining, we get

$$T_0 = \frac{7.62}{\sqrt{267.62 \cdot 1.429 \cdot 10^{-3}}} = 12.32.$$

Contrasting to $t_{0.025,13} = 2.16$, we reject. The number of new loans is **significant**.

On the other hand, for $\hat{\beta}_2$ (number of loans outstanding):

- Again, looking at $\left(X^T X\right)^{-1}$:

$$C_{22} = 2.222 \cdot 10^{-2}.$$

And we get that

$$T_0 = \frac{8.58}{\sqrt{267.62 \cdot 2.222 \cdot 10^{-2}}} = 3.52.$$

This leads to rejecting the null hypothesis and hence the number of loans outstanding is **also significant**. That said, there is something to be said about which one of the two predictor variables is more important to the regression, no?

We finish this lecture with one big, comprehensive example, solved over the last few pages.

One big comprehensive example

A real estate problem

Taken from realtor.com, here are 8 of the most recently sold homes in Urbana:

	Sq. ft.	Year built	Garages	#bedrooms	#bathrooms	Price
1	1547	1950	1	3	3	158500
2	1834	1957	0	4	2	183000
3	2520	1980	3	5	2.5	233000
4	985	1911	1	2	1	69000
5	1275	1968	0	3	1.5	118000
6	2337	1977	2	5	2	249900
7	1880	1967	2	3	2	175000
8	1943	1965	1	4	2.5	169900

Which one of the five predictor variables (sq. ft., year built, garages, #bedrooms, #bathrooms) is the least important for predicting price? Use $\alpha = 0.05$.

To solve this problem, we enumerate our steps in a way that makes it easier to memorize, understand, and interpret. Here we go:

1. Build $X = \begin{bmatrix} 1 & 1547 & 1950 & 1 & 3 & 3 \\ 1 & 1834 & 1957 & 0 & 4 & 2 \\ 1 & 2520 & 1980 & 3 & 5 & 2.5 \\ 1 & 985 & 1911 & 1 & 2 & 1 \\ 1 & 1275 & 1968 & 0 & 3 & 1.5 \\ 1 & 2337 & 1977 & 2 & 5 & 2 \\ 1 & 1880 & 1967 & 2 & 3 & 2 \\ 1 & 1943 & 1965 & 1 & 4 & 2.5 \end{bmatrix}$.
2. Calculate $X^T X = \begin{bmatrix} 8 & 14321 & 15675 & 10 & 29 & 16.5 \\ 14321 & 27474233 & 28123127 & 20469 & 55469 & 30798 \\ 15675 & 28123127 & 30716537 & 19654 & 56950 & 32377.5 \\ 10 & 20469 & 19654 & 20 & 40 & 22 \\ 29 & 55469 & 56950 & 40 & 113 & 62 \\ 16.5 & 30798 & 32377.5 & 22 & 62 & 36.75 \end{bmatrix}$
3. Compute $X^T y = \begin{bmatrix} 1356300 \\ 2629528500 \\ 2664759800 \\ 1946200 \\ 5318600 \\ 2944550 \end{bmatrix}$

4. Compute $\left(X^TX\right)^{-1} = \begin{bmatrix} 3654.00381 & 0.09848 & -1.93379 & -15.26988 & -6.28608 & 0.35632 \\ 0.09848 & 0.00002 & -0.00005 & -0.00238 & -0.00486 & -0.00139 \\ -1.93379 & -0.00005 & 0.00102 & 0.00826 & 0.00341 & -0.00038 \\ -15.26988 & -0.00238 & 0.00826 & 0.53030 & 0.63902 & 0.17674 \\ -6.28608 & -0.00486 & 0.00341 & 0.63902 & 1.85652 & 0.37698 \\ 0.35632 & -0.00139 & -0.00038 & 0.17674 & 0.37698 & 0.62852 \end{bmatrix}$

5. Find $\hat{\beta} = \left(X^TX\right)^{-1}X^Ty = \begin{bmatrix} -322042.7 \\ 93.2 \\ 150.2 \\ -4111.9 \\ 7242.7 \\ 4538.8 \end{bmatrix}$

We finally get that the regression line is:

$$\boxed{\hat{y} = -322042.7 + 93.2 \cdot x_1 + 150.2 \cdot x_2 - 4111.9 \cdot x_3 + 7242.7 \cdot x_4 + 4538.8 \cdot x_5}$$

Let us get some of the estimator calculations out of the way now:

- $SS_E = \sum_{i=1}^{8} (y_i - \hat{y}_i)^2 = 1164261866.8$.
- $\hat{\sigma}^2 = MS_E = \frac{SS_E}{8-6} = 582130933.4$.
- $SS_T = \sum (y_i - \overline{y})^2 = 23582558750$.
- Using ANOVA, $SS_R = SS_T - SS_E = 23582558750 - 1164261866.8 = 22418296883.2$.

We now have *everything* we need to do five distinct hypothesis tests for each of the five predictor variables. Specifically, we have:

1. For the square footage:
   - $H_0 : \beta_1 = 0, \quad H_1 : \beta_1 \neq 0$.
   - $T_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 \cdot C_{jj}}} = \frac{93.2}{\sqrt{582130933.4 \cdot 0.00002}} = 0.964$.
2. For the year built:
   - $H_0 : \beta_2 = 0, \quad H_1 : \beta_2 \neq 0$.
   - $T_0 = \frac{150.2}{\sqrt{\hat{\sigma}^2 \cdot 0.00102}} = 0.195$.
3. For the garage spots:
   - $H_0 : \beta_3 = 0, \quad H_1 : \beta_3 \neq 0$.
   - $T_0 = \frac{-4111.9}{\sqrt{\hat{\sigma}^2 \cdot 0.5303}} = -0.234$.

4. For the # bedrooms:
   - $H_0 : \beta_4 = 0, \;\; H_1 : \beta_4 \neq 0.$
   - $T_0 = \frac{7242.7}{\sqrt{\hat{\sigma}^2 \cdot 1.85652}} = 0.22.$
5. For the # bathrooms:
   - $H_0 : \beta_5 = 0, \;\; H_1 : \beta_5 \neq 0.$
   - $T_0 = \frac{4538.8}{\sqrt{\hat{\sigma}^2 \cdot 0.62852}} = 0.237.$

Hm... Apparently all factors are in the "fail to reject" region; in essence, this means that all of them one-by-one can be viewed as insignificant.. Some more (e.g., the year build with a $T_0 = 0.195$) than others (e.g., the square footage with a $T_0 = 0.964$), but still all of them can be declared insignificant when compared to $t_{\alpha/2,n-k-1} = t_{0.025,2} = 4.303$ as for all of them we have that $|T_0| < t_{\alpha/2,n-k-1}$. So, is the regression *significant at all*?

We can answer that through an F test:

$$F_0 = MS_R/MS_E = \frac{\frac{SS_R}{k}}{\frac{SS_E}{n-k-1}} = \frac{4483659376.64}{582130933.4} = 7.7.$$

Checking the critical value we get that $F_0 \leq f_{\alpha,k,n-k-1} = f_{0.05,5,2} = 19.3$, which means that we indeed do not have a good regression in our hands.

Finally, we may calculate the R^2 and adjusted R^2 coefficients:

- $R^2 = 1 - SS_E/SS_T = 0.951.$
- $R^2_{adj} = 1 - \frac{SS_E/(n-k-1)}{SS_T/(n-1)} = 0.827.$

Note the difference between R^2 and the adjusted R^2_{adj} showcasing that some insignificant predictor variables have been added.

Regression extensions and model selection

Learning objectives

After lecture 33, we will be able to:

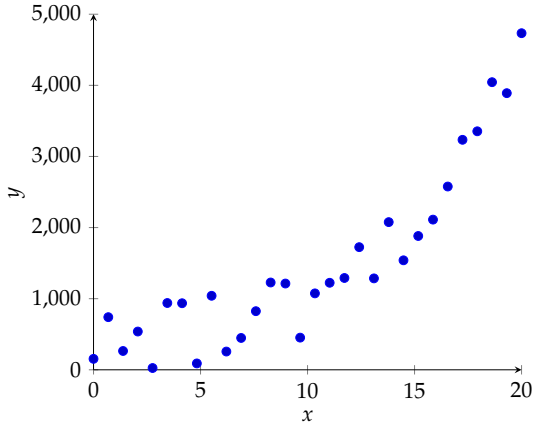
- Perform and interpret polynomial regression.
- Perform and interpret simple nonlinear regression.
- Build regression models with multiple predictors using:
  - all subsets selection.
  - backwards selection.
  - forwards selection.
- Describe and implement an "80-20" validation strategy.
- Describe and implement a K-fold validation strategy.

Motivation: Higher degree terms

What if our relationships is not linear, but is instead a more general **polynomial**? For example, what if I am sure that the yield of a crop is related to the square of the temperature? How could we incorporate this information into our regression models?

Or, what if I plot my data in a scatter plot and get an image like the one in Figure 86? How can I use regression to fit this data to a line?

Figure 86: The scatter plot containing all of our data points.

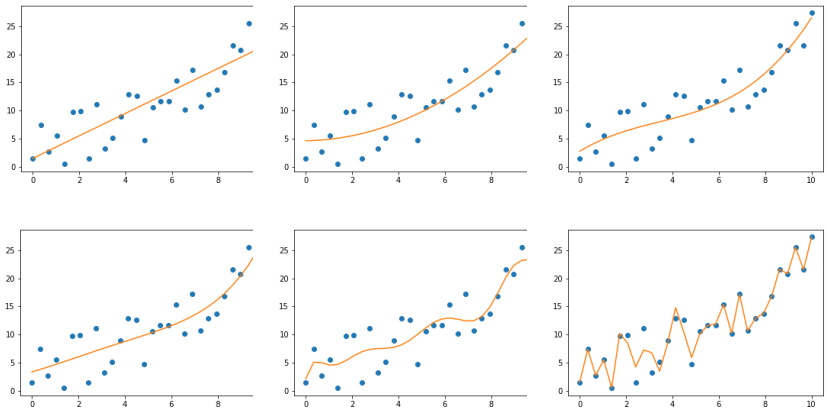


Motivation: Model building

Ok, so we have seen how to build models using 1 or $k > 1$ predictor variables. But given many possible predictor variables, how can we find the combination that works best?

Polynomial regression

Let's start with a question. Which of the following regression models do you believe best captures the data?



The first model (shown on the top left) is your typical simple linear regression. The other five models add some "curvature" by allowing higher degrees in the regression. For example, the second model is a quadratic term, whereas the last two are regressions that includes terms at the power of 10 and 25!

So, assume you have tried simple linear regression and the results have been underwhelming. You would like, instead to try the following line:

$$y = \beta_0 + \beta_1 x + \beta_{11} x_1^2.$$

A couple of notes:

1. We only consider simple linear regression for simplicity: we could very easily extend this to multiple linear regression.
2. There is one predictor variable: but it appears twice in our regression, one with degree 1 and one with degree 2. This is a quadratic regression!

How can we deal with a regression like this? Well, we can follow the next steps:

1. Create a "new" predictor variable (let us call it x_2).
2. Set x_2 equal to x_1^2: $x_2 = x_1^2$.
3. Set up a **multiple linear regression** using matrix X based on *two* predictor variables: x_1 and $x_2 = x_1^2$.

4. Find $\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_{11} \end{bmatrix} = (X^T X)^{-1} X^T y$.

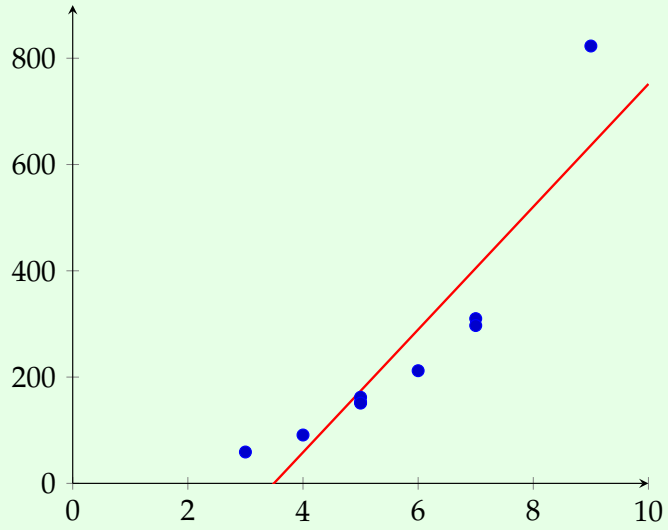
Let us put this to the test right away.

A small quadratic regression model

Consider the following data:

x	y
7	310
3	59
5	153
5	162
4	91
6	212
7	297
5	151
9	823

We tried a linear regression and got the line $y = 7.2404x - 2.2194$.



Since it does not look great, we decide to try a second degree regression polynomial of the form: $y = \beta_0 + \beta_1 x + \beta_{11} x^2$. What are $\hat{\beta}_0, \hat{\beta}_1, \hat{beta}_{11}$?

A small quadratic regression model

1. Add a new column in your data that is equal to x^2.

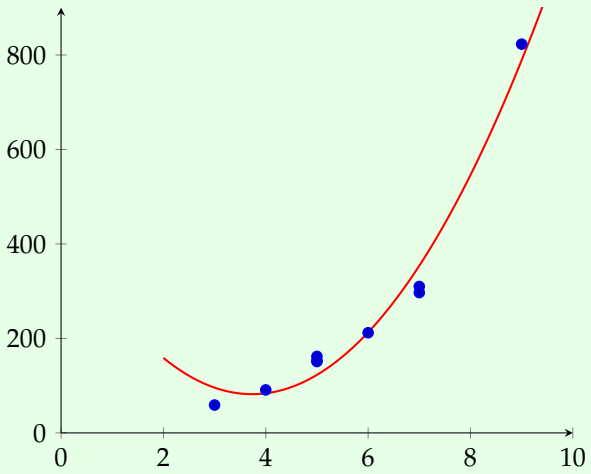
x	x^2	y
7	49	310
3	9	59
5	25	153
5	25	162
4	16	91
6	36	212
7	49	297
5	25	151
9	81	823

2. Construct X:

$$X = \begin{bmatrix} 1 & 7 & 49 \\ 1 & 3 & 9 \\ 1 & 5 & 25 \\ 1 & 5 & 25 \\ 1 & 4 & 16 \\ 1 & 6 & 36 \\ 1 & 7 & 49 \\ 1 & 5 & 25 \\ 1 & 9 & 81 \end{bmatrix}$$

3. Solve for $\hat{\beta}$:

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_{11} \end{bmatrix} = \left(X^T X\right)^{-1} X^T y = \begin{bmatrix} 437.74 \\ -190.47 \\ 25.5 \end{bmatrix}$$



Look at how much nicer this looks like!

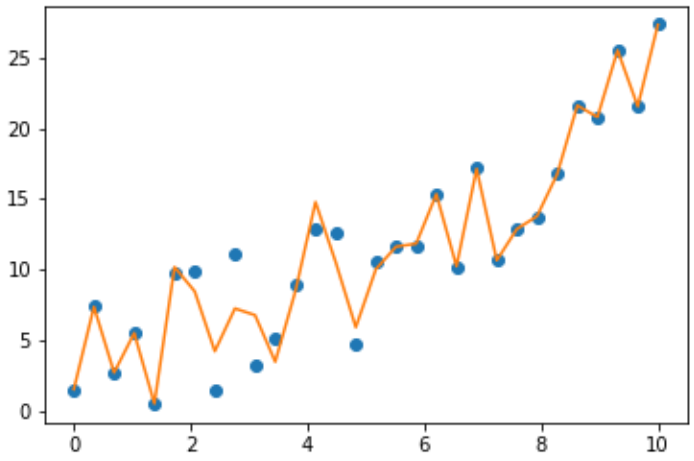
We can follow the same logic with other nonlinear functions!

Some nonlinear transformation examples

- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$
  - Introduce new variable $x_{12} = x_1 x_2$ and solve.
- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{123} x_1 x_2 x_3$
  - Introduce new variable $x_{123} = x_1 x_2 x_3$ and solve.
- We can even do that with other nonlinear functions: for example $y = \beta_0 + \beta_1 x_1 + \beta_2 cos(x_1)$.
  - Introduce new variable $x_2 = cos(x_1)$ and solve.
- Or $y = \beta_0 + \beta_1 x_1 + \beta_2 \log x_1$.
  - Introduce new variable $x_2 = \log x_1$ and solve.

Finally, what is the appropriate model? Now that we can go nonlinear, we could (if we wanted to) make almost all residuals equal to zero! See for example the regression curve in Figure 87.

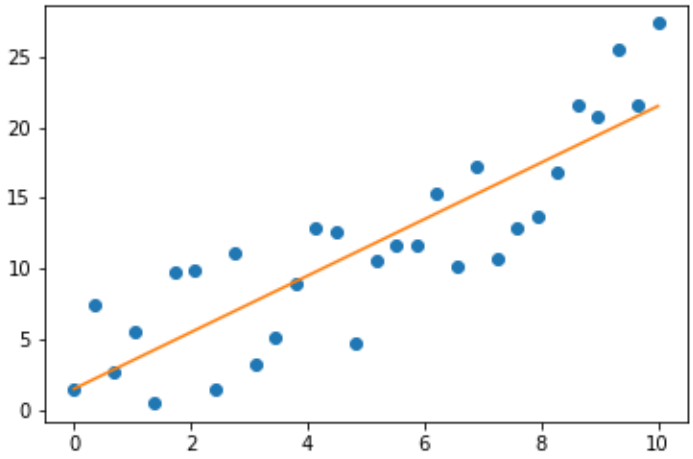
Figure 87: An example of **overfitting**. Here, we end up following the given data too closely, not allowing for any randomness at all.



Of course, the opposite route is still very much possible. We may decide that the simplest, linear regression may be the way to go. The previous two cases are called **overfitting** and **underfitting**.

- Overfitting is an issue because we end up getting too caught up on past information, and hence we lose our edge to predict the future if it doesn't look exactly like the past.

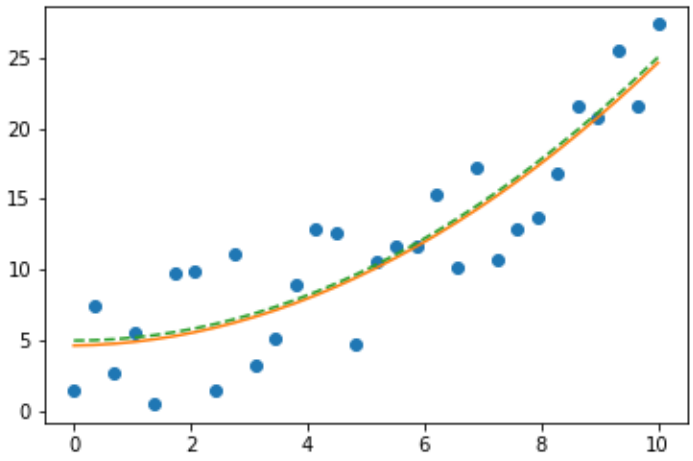
Figure 88: An example of **underfitting**. Here, we end up with a simple linear regression that does not seem to follow the data as well.



- Underfitting, on the other hand, is an issue of oversimplification: our model does not predict well because it is missing information.

We would like to do an **appropriate model selection**. How?

Figure 89: An appropriate model which balances past information and flexibility to new data.



Before we see the how, a couple of quotes that can help us drive the point of model selection home:

1. Paul Valéry (philosopher, 1942)

   "Ce qui est simple est toujours faux. Ce qui ne l'est pas est inutilisable." [88]

2. George Box (statistician, 1978)

   "All models are wrong, but some are useful."

[88] "What is simple is always wrong. What is not simple is impossible to use."

Model selection

In most problems, we have *many* potential variables to consider. To make things worse, we can include different *functions* of the variables themselves! Which ones should be included?

Which to include?

We want to build a regression model using any combination of three factors x_1, x_2, x_3. We can build any of the following models:

1. x_1 alone, or x_2 alone, or x_3 alone. 3 models.
2. x_1 and x_2 but not x_3, or x_1 and x_3 but not x_2, or x_2 and x_3 but not x_2. 3 models.
3. x_1, x_2, and x_3 toghether. 1 model.
4. None of them! 1 model.

The last case happens when all three factors are not significant to the regression.

To make matters worse, assume if we could also add *their squares*: $x_1, x_2, x_3, x_1^2, x_2^2, x_3^2$ for a total of $2^6 = 64$ possible models. Which one should we build?

Hopefully by now you are motivated and you see how this is an important problem that needs to be addressed. Even more so nowadays, with the advent of big data. It is imperative we find out a way to trim the model so that only significant factors are included. In the next few pages, we discuss three model building approaches, called:

1. **all subsets** selection.
2. **backwards** selection.
3. **forwards** selection.

All subsets selection

All subsets selection is a term to signal that we need to consider *all* possible subsets we can create with our factors: these can be quite many. They actually grow exponentially and with k predictor variables, we already have 2^k possible subsets. [89]

[89] Why?

Among all 2^k possible subsets, pick the subset of predictor variables/factors that leads to the largest R^2_{adj}.

Back to the realtor.com example

Consider the realtor.com example from last time. We assumed a house's price depends on area (sq. ft.), the year built, the garage spots, the number of bedrooms, and the number of bathrooms. Which subset of variables gives us the best regression model?

We have 32 combinations to consider (including the empty set, which implies that neither factor is significant). Some are presented here:

- $(x_1, x_2, x_3, x_4, x_5)$: $R^2_{adj} = 0.827$
- (x_1, x_2, x_3, x_4): $R^2_{adj} = 0.882$
- (x_2, x_3, x_4, x_5): $R^2_{adj} = 0.831$
- (x_1, x_3, x_4, x_5): $R^2_{adj} = 0.883$
- ...
- (x_1, x_2, x_3): $R^2_{adj} = 0.910$
- ...
- (x_1, x_3, x_4): $R^2_{adj} = 0.909$
- (x_1, x_3, x_5): $R^2_{adj} = 0.910$
- ...
- (x_1, x_2): $R^2_{adj} = 0.919$
- (x_1, x_3): $R^2_{adj} = 0.926$
- ...
- (x_1): $R^2_{adj} = 0.927$
- $\emptyset$: $R^2_{adj} = 0.857$

Among them, pick the one with the largest R^2_{adj}. In our case, that would be the model with **only** x_1.

Backwards selection

We immediately see the issue with the previous case: too many combinations to consider, even for few predictor variables. To avoid enu-

merating fully all subsets, we investigate two heuristic approaches. With the term heuristic approach, we mean an approach that is not guaranteed to give us the optimal subset; that said, we expect its solution to be obtained faster. We proceed to describe the apporach.

1. First, start by including **all** predictor variables in your regression model.
2. Do a hypothesis test for significance of each individual factor among the predictor variables in your current regression.
3. Check if all P-values are above some threshold (say $p > 0.10$).
4. If not, find the one factor with the *largest* P-value.
   - This is the "least significant" predictor.
5. Remove it from consideration and calculate the new R^2_{adj}. If it is lower than the previously obtained R^2_{adj}, stop. Otherwise, iterate (go back to step 1) after removing the factor.

If P-values are not readily available, we may compare each T-test value $|T_0|$ to $t_{\alpha/2,n-k-1}$ and see if you'd accept/reject the hypothesis. Then, pick the variable which is the farthest from the rejection area and remove it from consideration instead.

Back to the realtor.com example

In the notes from Lecture 32, we did individual hypothesis tests for each of the factors. We had gotten:

1. x_1: $T_0 = 0.964$ P-value $= 0.437$.
2. x_2: $T_0 = 0.195$ P-value $= 0.864$.
3. x_3: $T_0 = -0.234$ P-value $= 0.837$.
4. x_4: $T_0 = 0.220$ P-value $= 0.846$.
5. x_5: $T_0 = 0.237$ P-value $= 0.835$.

The R^2_{adj} for the full model with all five variables is equal to 0.827. Use the backwards selection heuristic approach to find a good regression model.

Back to the realtor.com example

We remove the one with the largest P-value (the one with the T_0 test statistic value that is farthest from rejection): x_2. We then ran the new regression with the remaining four variables to get that :

1. x_1: $T_0 = 1.391$ P-value $= 0.258$.
2. x_3: $T_0 = -0.393$ P-value $= 0.720$.
3. x_4: $T_0 = 0.250$ P-value $= 0.819$.
4. x_5: $T_0 = 0.291$ P-value $= 0.790$.

The new R^2_{adj} is equal to 0.883: since it has improved, continue with the next iteration. From the remaining factors, we now remove x_4 (the highest P-value). The new model (including x_1, x_3, x_5) leads to $R^2_{adj} = 0.910$. This is an improvement, so we continue. Again, the new model includes:

1. x_1: $T_0 = 5.699$ P-value $= 0.005$.
2. x_3: $T_0 = -0.850$ P-value $= 0.443$.
3. x_5: $T_0 = 0.249$ P-value $= 0.816$.

x_5 is set to be removed, leaving us with a model including only x_1, x_3. The new R^2_{adj} is 0.926 – again improving the previous one. Hence, we get:

1. x_1: $T_0 = 7.535$ P-value $= 0.0007$.
2. x_3: $T_0 = -0.993$ P-value $= 0.366$.

Note how x_3 has a P-value above 0.1: let's try removing it and keep a model with **only** x_1. Its R^2_{adj} is equal to 0.927 – another improvement! Removing x_1, though, we obtain the empty model with $R^2_{adj} = 0.857$. Since it worsens, we stick with the model with x_1 alone.

Forwards selection

Forwards selection is – you guessed it – the opposite idea!

1. First, start by including **none** of the predictor variables.
   - That is, we have a line based only on the intercept β_0.

2. Then, run k separate regression models, one for each of the predictor variables.
3. Check whether any of the P-values in each individual test is below some threshold (say $p < 0.10$).
4. Pick the regression variable that leads to the *smallest* P-value. Equivalently, you may check the variable whose addition increases R^2_{adj} the most.
   - This is the "most significant" predictor.
5. Add it to the model and continue to run $k - 1$ separate regression lines: each with the variable from the first part, and one of the remaining variables.
6. Iterate and stop when no more variables have a P-value that is lower than 0.10 (or when no addition leads to an increased R^2_{adj}).

Back to the realtor.com example

Let us solve the same problem, but using forwards selection now.

- First, perform five different regressions, one per variable.
  1. x_1: $T_0 = 9.492$ P-value $= 7.79 \cdot 10^{-5}$.
  2. x_2: $T_0 = -3.354$ P-value $= 0.015$.
  3. x_3: $T_0 = 4.396$ P-value $= 0.005$.
  4. x_4: $T_0 = 6.343$ P-value $= 0.0007$.
  5. x_5: $T_0 = 1.750$ P-value $= 0.131$.
- We add the one with the smallest P-value (the one with the T test statistic value that is the easiest to reject): x_1. The current model has $R^2_{adj} = 0.926$.
- We then ran the new regression with the one variable from earlier (x_1) plus each of the remaining four:
  1. (x_1, x_2): $T_0 = 0.632$ P-value $= 0.555$.
  2. (x_1, x_3): $T_0 = -0.993$ P-value $= 0.366$.
  3. (x_1, x_4): $T_0 = 0.741$ P-value $= 0.492$.
  4. (x_1, x_5): $T_0 = 0.386$ P-value $= 0.716$.

All P-values are above 0.1, so we stop. The model obtained from forwards selection is the one including x_1 alone.

Validation

We have build a regression model based on past data and we now want to put it to the test. But before we do that, we want to check how confident should we be in our model? How can we validate our model selection?

Traditionally, the main idea has been to split our data (the data that we would normally use to build our model!) in two parts: **training** data and **testing** data. The common split between these two is 80%-20% in favor of training. Now, we:

1. Use the training data to build the regression model.
2. Use the testing data to evaluate how well the regression is doing.
   - We quantify the performance through the mean square error:

$$MS_E = \frac{1}{n-2}\sum \left(y_i^{test} - \hat{y}_i^{test}\right)^2.$$

Visually, this is the traditional 80-20 split:

So, what is K-fold validation? K-fold validation involves splitting the data into K parts, typically of equal size. $K-1$ of them are used as training data, with 1 part of them being used as testing data. Then, we:

1. Use the training data to create $K-1$ regression models, one for each of the parts.
2. Use the testing data to test how well **each** of the regression models are performing. Again, you may use the MS_E as defined earlier.
3. Select and return the best model amongst them.

Or, visually: