IE 300: Analysis of data (worksheets and active learning)

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

1. Random experiments, sample spaces, and events

Every worksheet will work as follows.

1. You will be asked to form a group with other students in the class: you can make this as big or as small as you'd like, but groups of 4-5 work best.
2. Read through the worksheet, discussing any questions with the other members of your group.
   - You can call me at any time for help!
   - I will also be interrupting you for general guidance and announcements at random points during the class time.
3. Answer each question (preferably in the order provided) to the best of your knowledge.
4. While collaboration between students is highly encouraged and expected, each student has to submit their own version.
5. You will have 24 hours (see gradescope) to submit your work.

Activity 1: Playing with coins

Problem 1

A coin is tossed 4 times in a row and we mark whether it has come up Heads or Tails each time. Explain (in a sentence) why this is a random experiment. [1]

[1] Consider what happens if we repeat this process. Do we always get the same sequence?

Answer to Problem 1.

Problem 2

In the experiment from Problem 1, what is the sample space? [2] What is the cardinality of the sample space?

[2] In general, there are multiple ways to define a sample space. Here, we may want to focus on each individual coin toss as it happens. For example, $\{Heads, Heads, Tails, Heads\}$ could be a potential outcome, whereas $\{Tails, Tails, Tails, Heads\}$ would be another.

Answer to Problem 2.

$S =$

$|S| =$

Problem 3

Once again, consider the experiment from Problem 1. What is the cardinality of the following events:

1. Get the sequence Heads, Tails, Tails, Heads.
2. The third coin toss comes up Heads.
3. Get at least three Heads.

Answer to Problem 3.

1.

2.

3.

You know what's interesting? When all outcomes in the sample space are equally likely, then the "likelihood" of an event happening (*probability*) is equal to the cardinality of the event over the cardinality of the sample space! For example, "Get at least three Heads" has cardinality 5, and the sample space S has cardinality $|S| = 16$ for a "likelihood" of $5/16 = 31.25\%$. [3]

[3] More on that next time!

Problem 4

Consider the events in Problem 3. Are the first and the second events mutually exclusive? How about the first and the third events? Finally, what can you say about the second and the third events? Justify (in a sentence) your answer.

Answer to Problem 4.

1. "Get the sequence Heads, Tails, Tails, Heads" and "The third coin toss comes up Heads":

2. "Get the sequence Heads, Tails, Tails, Heads" and "Get at least three Heads":

3. "The third coin toss comes up Heads" and "Get at least three Heads":

Activity 2: An experimental design

Problem 5

An experiment happens in an environment where the temperature is always between 20 and 100 Fahrenheit – that is, the temperature belongs to $S = [20, 100]$. Define the events:

- $A = [80, 100]$: temperature is greater than or equal to 80 F.
- $B = [20, 40]$: temperature is between 20 and 40 F.
- $C = [32, 85]$: experiment is successful.

When event C has happened, we say that the experiment is successful. In other cases, the experiment is unsuccessful. We also have that the experiment is hot when the temperature belongs to A, and we say that event A has happened. Similarly, the experiment is cold when the temperature belongs to B, and we say that B has happened. [4]

For this next question, write what range of values satisfy each of the given events. For example, if asked to find what $A \cap C$, you would answer $[80, 85]$, as $A \cap C$ is the range of values that are both in A and in C at the same time.

[4] Be careful with the inclusion and exclusion notation [,] and (,), respectively.

Answer to Problem 5.

1. What is $\overline{C}$?
2. What is $C \setminus A$?
3. What is $B \cap C$?
4. What is $(\overline{A} \cap B) \cup (\overline{A} \cap C)$?
5. Define the set of outcomes where an experiment that is both successful while simultaneously being neither hot (≥ 80) nor cold (≤ 40) in set notation.

Activity 3: Set and cardinality properties

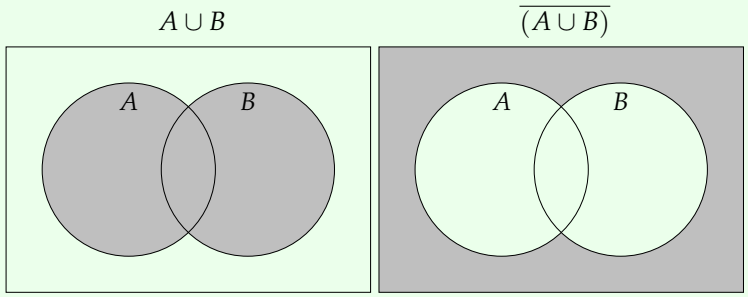
We saw many interesting properties in the first pre-lecture video and the accompanying slides. Now, it is time to see their derivations. To begin with, consider the first of the two DeMorgan's law we saw in the lecture notes:

$$\boxed{\overline{(A \cup B)} = \overline{A} \cap \overline{B}.}$$

Problem 6

In the two Venn diagrams below, we mark in gray the event $(A \cup B)$ and $\overline{(A \cup B)}$ (corresponding to the left hand side). We have solved this for you, so you can move to Problem 7.

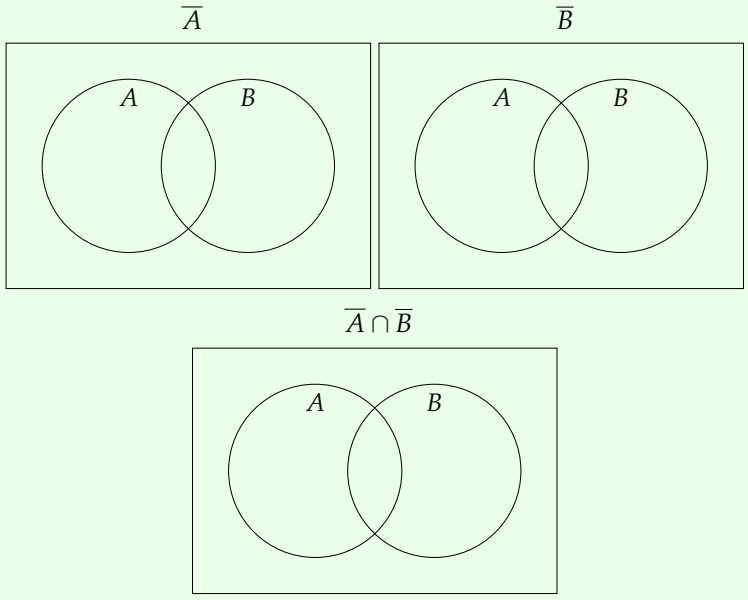
Answer to Problem 6.



Problem 7

In the three Venn diagrams provided, mark the events $\overline{A}$, $\overline{B}$, and $\overline{A} \cap \overline{B}$ (corresponding to the right hand side of the DeMorgan's law).

Answer to Problem 7.



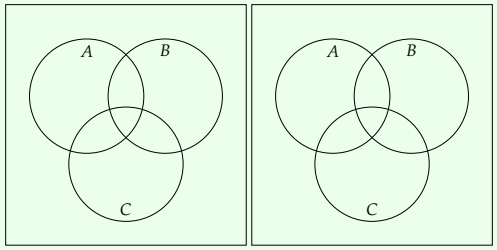
Note how the last Venn diagram in Problem 7 (the one for $\overline{A} \cap \overline{B}$) is the same as the last Venn diagram in Problem 6 (the one corresponding to $\overline{(A \cup B)}$, "showing" DeMorgan's law.

Problem 8

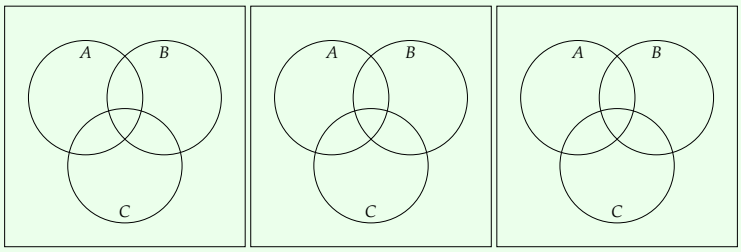
Based on the previous constructive derivation you did, what can you say about the extension of this DeMorgan's law to more than 2 events? Specifically, what is $\overline{(A \cup B \cup C)}$? Use the diagrams below to do something similar to what you did in Problems 6 and 7.

Answer to Problem 8.

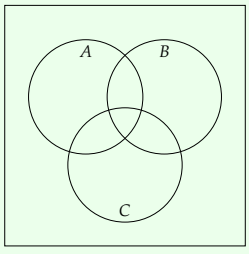
$A \cup B \cup C$ $\qquad$ $\overline{(A \cup B \cup C)}$



$\overline{A}$ $\qquad$ $\overline{B}$ $\qquad$ $\overline{C}$



$\overline{A} \cap \overline{B} \cap \overline{C}$



Finally, we have that:

$$\overline{(A \cup B \cup C)} =$$

Problem 9

Consider two **mutually exclusive** [5] events A and B. What can you say about the cardinality of $A \cap B$ and of $A \cup B$? You may express your answer as a function of the individual cardinalities $|A|$ and $|B|$.

[5] That is, two events that *share no common outcomes*.

Answer to Problem 9.

- $|A \cap B| =$
- $|A \cup B| =$

Problem 10

It is true that for two general sets A, B, we have that $|A \cup B| = |A| + |B| - |A \cap B|$. Let us construct a proof for the statement using your answers in Problem 9.

Consider an event X that is comprised of **three mutually exclusive events** X_1, X_2, X_3. What can you say about the cardinality of X and the cardinalities of X_1, X_2, X_3? Circle the (one) correct answer.

Answer to Problem 10.

$$|X| \quad \begin{matrix} < \\ \leq \\ = \\ \geq \\ > \end{matrix} \quad |X_1| + |X_2| + |X_3|$$

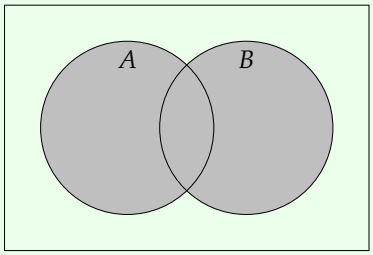
This is true, in general, no matter how many mutually exclusive events we can break an event down to. Say that event X consisted of n mutually exclusive events $X_1, X_2, \ldots, X_n$, we could write that its cardinality $|X|$ is equal to the summation of the individual event cardinalities:

$$|X| = \sum_{i=1}^{n} |X_i|.$$

Problem 11

Based on your observation in Problem 10, can we think of $A \cup B$ as three mutually exclusive events? Check the following Venn diagram and mark the **three mutually exclusive events** that describe the greyed area. How are they described mathematically?

Answer to Problem 11.



Problem 12

Combine your observations from Problems 10 and 11 to derive that $|A \cup B| = |A| + |B| - |A \cap B|$.

Answer to Problem 12.

2. Counting

Every worksheet will work as follows.

1. You will be asked to form a group with other students in the class: you can make this as big or as small as you'd like, but groups of 4-5 work best.
2. Read through the worksheet, discussing any questions with the other members of your group.
   - You can call me at any time for help!
   - I will also be interrupting you for general guidance and announcements at random points during the class time.
3. Answer each question (preferably in the order provided) to the best of your knowledge.
4. While collaboration between students is highly encouraged and expected, each student has to submit their own version.
5. You will have 24 hours (see gradescope) to submit your work.

Activity 1: Texas hold'em

Texas hold'em is a variant of poker (a card game) where each player is given 2 starting cards out of the 52 cards in a deck. There are 13 cards of each suit; and there are four suits: ♡, ♢, ♠, ♣. The 13 cards are from 1 to 10, along with three special cards named J, Q, K.

Problem 1

How many combinations of starting cards are there? For the sake of this problem consider that order does not matter (and hence a starting card setup of ♡3♠7 and ♠7♡3 are one and the same). Also assume that the same numbers of different suits are different setups (and hence ♢5♡K and ♣K♡5 are two different starting setups). [6]

Answer to Problem 1.

[6] Translation: you have 52 cards to pick from, and you need to see how many ways you have of picking 2 of them – without any care for the order.

Problem 2

We define a pair as two same cards (of different suit). For example, ♡K♠K and ♡4♢4 are pairs. Considering again that order does not matter, and that the same numbers of different suits are different setups (and, for example, ♡4♢4 and ♣4♡4 are two different setups), how many starting setups that are pairs are there? [7]

[7] If it helps, consider how many pairs of a single card (say, "7"s) you can create. Then, whatever that number is, you can multiply it by all 13 possible cards.

Answer to Problem 2.

Problem 3

Considering your answers in Problems 1 and 2, what is the probability that we are dealt a pair in our 2 starting cards?

Answer to Problem 3.

Activity 2: Elections

Problem 4

Assume that there is a country with two political parties: Party A and Party B. A small town has 45 registered voters. Typically in an election, you have to count all votes, and the party with the most votes wins. That said, the town does not want to invest the resources to count all 45 votes and are contemplating a new system. In this new system, **only three registered voters** are selected at random and asked to vote.

Assume we have independently polled all voters and we are aware that Party A is set to win with 30 votes compared to 15 votes for Party B. If the town implements this new system, what is the probability that Party B is the one that wins the election? [8]

[8] For a hint, look at the Quality Control example in our notes!

Answer to Problem 4.

Problem 5

How does the calculation change if we were to pick 5 voters among all registered ones (instead of 3)?

Answer to Problem 5.

Problem 6

What do you observe when comparing your answers to Problem 4 and Problem 5? Which of the two parties would prefer a bigger number of voters go to the polls?

Answer to Problem 6.

Activity 3: Deriving the permutations and combinations formulae

Problem 7: From multiplications to permutations

In the lecture, we saw that the formula for a permutation of n items is $P_n = n!$. Use the multiplication rule to show that this is indeed the case. [9]

[9] If it helps, consider the case of assigning n items to n people. How many ways are there to assign the first item? After assigning the first item, how many ways are there to assign the second one? How about the k-th one (for $k < n$)?

Answer to Problem 7.

Problem 8: From multiplications to permutations of $r < n$ items

Similarly, in the lecture we derived the formula for a permutation of $r < n$ items from a total of n items. Again, use the multiplication rule to show that $P_{n,r} = \frac{n!}{(n-r)!}$.

Answer to Problem 8.

Problem 9: From permutations to combinations

Recall that a permutation $P_{n,r}$ is an ordered sequence of r items out of n possible items, whereas a combination $C_{n,r}$ is an unordered sequence of the same r items. Use the multiplication rule to show that $P_{n,r} = C_{n,r} \cdot r!$. Then, use that fact to derive the combinations formula.

Answer to Problem 9.

Activity 4: Summary of results

Problem 10: Committing to memory

We have seen several different formulae for counting in this lecture. In your answer, provide one example and write the formula associated with each setup.

Answer to Problem 10.

- **Multiplication**:

Formula	Small example

- **Permutation of n items**:

Formula	Small example

- **Permutation of r<n items**:

Formula	Small example

- **Distinguishable permutations of groups of items**:

Formula	Small example

- **Combination of r<n items**:

Formula	Small example

3. Basic probability theory

Every worksheet will work as follows.

1. You will be asked to form a group with other students in the class: you can make this as big or as small as you'd like, but groups of 4-5 work best.
2. Read through the worksheet, discussing any questions with the other members of your group.
   - You can call me at any time for help!
   - I will also be interrupting you for general guidance and announcements at random points during the class time.
3. Answer each question (preferably in the order provided) to the best of your knowledge.
4. While collaboration between students is highly encouraged and expected, each student has to submit their own version.
5. You will have 24 hours (see gradescope) to submit your work.

Activity 1: Game of dies

Problem 1

Consider a game where you roll two fair dice (i.e., where each number on the side of the dice from 1 to 6 has an equal probability of appearing). What is the probability that the sum of the numbers on the two dice is 7?

Answer to Problem 1.

Problem 2

What is the probability the sum of the numbers on the two dice is 7 given that the first of the two dice rolled on a 3?

Answer to Problem 2.

Problem 3

Based on your answers on part (a) and (b), what can you claim about the independence of the events "the sum of the two dice is equal to 7" and "the first dice is a 3"? Is that true for all pairs of events "the sum of the two dice is 7" and "the first dice is a i" where $i = 1, 2, 3, 4, 5, 6$?

Answer to Problem 3.

Problem 4

Prove or disprove[10] the following statement.

- When throwing two dies, the two events "the sum of the two dies is $j = 2, 3, \ldots 12$" and "the first die is a $i = 1, 2, \ldots, 6$" for $i < j$ are independent events.

[10] To disprove a statement, you may simply find an example where the statement is **not** true.

Answer to Problem 4.

Activity 2: Quality control revisited

Problem 5

A manufacturing facility is making 2 different products. Every product can be classified as defective (D) or non-defective (ND). In addition to that, some products appear to have cosmetic damage (C) or not (NC). The company has collected data for both products over the last 400 items for each. [11]

[11] You may treat these numbers as "probabilities": for example a product 1 is defective and has cosmetic damage with probability 5/400, whereas a product 2 that is known to be defective has cosmetic damage with probability 2/20.

Product 1:

Def.	Cosm. dam. Yes (C)	No (NC)	Total
Yes (D)	5	23	28
No (ND)	24	348	372
Total	29	371	400

Product 2:

Def.	Cosm. dam. C	NC	Total
D	2	18	20
ND	38	342	380
Total	40	360	400

You pick up an item from the recent production of Product 1. If you see it has cosmetic damage, does this alter your perception that the product is defective? [12]

Answer to Problem 5.

[12] Could we compare the probability that an item is defective (let it be $P(D)$) with the probability that it is defective *given* that it has cosmetic damage (let it be $P(D|C)$)? What if these probabilities are equal to one another? What if they are not?

Problem 6

Using the data provided from the previous problem, what can you deduce about Product 2? Does knowing that it has cosmetic damage alter your perception that the product is defective?

Answer to Problem 6.

Activity 3: The birthday problem

Our class has 84 students. If our class had 365 students (assume for now with me that February 29th does *not* exist and, hence, every year has 365 days), then we would be guaranteed that at least two of you share the same birthday.

The question though becomes: in a class of 84, like ours, what is the probability that two of you have the same birthday?

Problem 7

Let's start simple. If there are only two of you, what is the probability that you share the same birthday?

Answer to Problem 7.

Problem 8

With your answer in Problem 7 in mind, add a third person in the mix. What is the probability that the third person has the same birthday with either the first or the second person? [13]

Answer to Problem 8.

In many instances when interested in finding the probability of an event E, it is easier to calculate $P(\overline{E})$ and use that to find $P(E) = 1 - P(\overline{E})$, rather than trying to calculate $P(E)$ immediately!

[13] Hint: consider the event that you do not have the same birthday as E and then calculate $P(\overline{E})$. Also: how many possible triplets of birthday dates can you create such that no two birthdays are the same? And how many possible triplets of birthday dates can you create in total? Probability can de calculated as the first number over the second on, if only we knew how to count the number of events...

Problem 9

Based on your reasoning in Problems 7 and 8, you must have reached a probability of

$$\frac{365}{365} \cdot \frac{364}{365} \cdot \frac{363}{365} \cdot \ldots \cdot \frac{365 - n + 1}{365}$$

for the event that no two people share a birthday in a group of n people. Using the fact that $P(E) = 1 - P(\overline{E})$, we can deduce that the probability that two people share a birthday is:

$$P(\text{share birthday}) = 1 - \frac{365}{365} \cdot \frac{364}{365} \cdot \frac{363}{365} \cdot \ldots \cdot \frac{365 - n + 1}{365} \qquad (1)$$

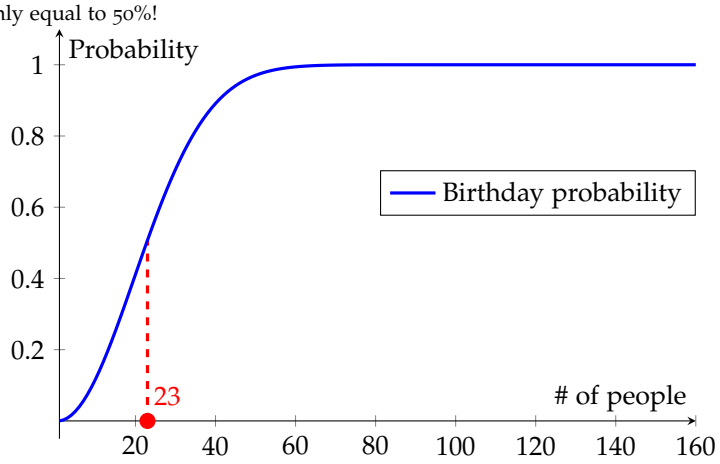
What does expression (1) evaluate to in a class of 84 students? What does it evaluate to in a class of 25 students? [14]

[14] Check our Jupyter notebook for the birthday paradox, too.

Answer to Problem 9.

Next time a person in any class of a significant size shares the same birthday with you, remember that this is decidedly **not** the biggest coincidence in the world, but instead a rather common observance.

Figure 1: The birthday problem probabilities, visualized. It is at 23 people that this is roughly equal to 50%!



Say we had run this for a many values of n, starting from $n = 1$ (one person alone has a 0% chance of sharing the birthday with someone else), $n = 2$ (two people sharing a birthday with a $1/365$ chance), $n = 3$ (triplet of people sharing a birthday), and so on, until $n = 160$ people. We would have then obtained a figure like the one in Figure 1. What do you observe?

See how few people are needed for the probability to get *almost* equal to 100%! This is a very important realization and helps us understand that sometimes rather common observations are viewed as rare.

4. Bayes' theorem

Every worksheet will work as follows.

1. You will be asked to form a group with other students in the class: you can make this as big or as small as you'd like, but groups of 4-5 work best.
2. Read through the worksheet, discussing any questions with the other members of your group.
   - You can call me at any time for help!
   - I will also be interrupting you for general guidance and announcements at random points during the class time.
3. Answer each question (preferably in the order provided) to the best of your knowledge.
4. While collaboration between students is highly encouraged and expected, each student has to submit their own version.
5. You will have 24 hours (see gradescope) to submit your work.

Activity 1: The law of total probability

Problem 1: U.S. Open

In a game of tennis [15] two players compete for the next point. The person serving has (historically) an advantage and wins the next point with probability 70%. This probability changes to 50% (so both players are equally probable to win the next point), if the person serving loses the point [16]. When the person serving wins the point again, the probability returns to 70%.

What is the probability the player serving wins two points in a row? What is the probability the player serving wins the second point (no matter what happens in the first point)? What is the probability the other player wins the second point (no matter what happens in the first point)?

[15] The U.S. Open is taking place these days, hence the motivation for this question! The early rounds of the tournament have just started.

[16] You may assume that this is because tennis is also a sport of momentum and morale.

Answer to Problem 1.

Server wins both points:

Server wins second point:

Opponent wins second point:

Problem 2

A company needs to make a decision on whether to open a new facility or not. Their chief economic strategist has recommended one of three options:

1. Invest in a new facility.
2. Play it safe and invest in renovations in existing facilities.
3. Cut down on all expenses.

The market next year can be bullish, bearish, or stagnant [17]. The economic strategist has identified the following probabilities of success for the company depending on the economy:

	Market		
Option	Bullish	Bearish	Stagnant
Option 1	80%	5%	30%
Option 2	50%	50%	80%
Option 3	10%	95%	40%

What should the company choose to do assuming that next year's market is expected to be bullish with probability 30%, bearish with probability 30%, and stagnant with probability 40%? To answer the question, get the probability of success for each of three options and pick the one with the highest.

[17] A bullish market is recognized by an increase in prices, due to the market participants having optimistic views of the economy; a bearish market on the other hand sees a decline in prices, as market participants become pessimistic of the economy. Finally, a stagnant economy sees neither increases or declines in prices and prices stay in a constant level.

Answer to Problem 2.

Option 1: $P(success) =$

Option 2: $P(success) =$

Option 3: $P(success) =$

Activity 2: Using Bayes' theorem

In this activity, we will bring smaller and bigger applications of the Bayes' theorem. We already saw an interesting one that has to do with the Mantoux test (any diagnostic tool in general). Here, we will see examples from fair grading to detecting spam messages and blood supply logistics.

Problem 3: Fair grader

A quiz in IE 300 is scheduled to have just one multiple choice question with 5 different answers. A student decides to only study half the material for the quiz: hence with probability 50% they will know the answer to the question, and with probability 50% they will guess an answer at random [18]. A student receives a perfect score in the quiz! What is the probability they guessed the answer?

[18] Recall: this implies that all 5 answers are equally probable events!

Answer to Problem 3.

Problem 4: SPAM?

Gmail has observed that the word "inheritance" appears in 20% of all spam emails. It appears at only 0.1% of all known non-spam email communications. Roughly, the estimate right now is that 45% of all email communications are spam. You received an email entitled **Inheritance**: what is the probability it is spam?

Answer to Problem 4.

Problem 5: A need for blood

In Greece, blood type distribution is as follows: 44.4% type O, 37.9% type A, 13% type B, and 4.7% type AB [19]. However for people born in the 1960s or earlier, typing was not correctly done. We now know that:

- for a person with blood type A, they would be (correctly) found as A with probability 85%;
- for a person with blood type O, they would be (incorrectly) found as A with probability 5%;
- for a person with blood type B, they would be (incorrectly) found as A with probability 15%;
- for a person with blood type AB, they would be (incorrectly) found as A with probability 25%;

A person born in the period before the 1960s is brought to a hospital and needs blood. They are listed as having type A blood: what is the probability they actually need type A blood?

[19] To be clear: blood types are mutually exclusive and collectively exhaustive.

Answer to Problem 5.

Activity 3: Deriving Bayes' theorem

In Bayes' theorem, we assume the existence of n states or hypotheses $S_i, i = 1, \ldots, n$ that are either reinforced or weakened through the appearance of m test outcomes or evidence $O_j, j = 1, \ldots, m$.

Bayes' theorem also assumes the availability of prior information in the form of probabilities for each state ($P(S_i), i = 1, \ldots, n$) and the availability of historical information in the form of probabilities for each outcome depending on the state ($P(O_j|S_i), j = 1, \ldots, m, i = 1, \ldots, n$).

Finally, we are concerned with deriving what $P(S_i|O_j)$ or what is the probability that state/hypothesis S_i is true given the existence of a test outcome or evidence O_j.

With these in mind, answer the following questions.

Problem 6

Write $P(S_i|O_j)$ using the conditional probability formula. [20]

[20] How can we calculate a conditional probability? Check the previous lecture notes..

Answer to Problem 6.

Problem 7

You should have that $P(S_i|O_j)$ can be written as a fraction of two probabilities. Focus on the probability in the numerator, $P(S_i \cap O_j)$. In a sentence, explain why $P(S_i \cap O_j) = P(O_j \cap S_i)$.

Answer to Problem 7.

Problem 8

Use the multiplication rule to write $P(O_j \cap S_i)$. [21]

Answer to Problem 8.

Problem 9

Go back to your answer in Problem 6, however now focus on the denominator, $P(O_j)$. Use the total probability law to write $P(O_j)$ as a function of $P(S_i), i = 1, \ldots, n$ and $P(O_j|S_i), i = 1, \ldots, n$.

Answer to Problem 9.

Problem 10

Combine your answer to Problem 8 and Problem 9 to show what $P(S_i|O_j)$ is according to Bayes' theorem.

Answer to Problem 10.

[21] Recall that the multiplication rule for probabilities states that

$$P(A \cap B) = P(B) \cdot P(A|B).$$

As a reminder, Bayes' theorem states the following: given n mutually exclusive and collectively exhaustive events S_i with probabilities $P(S_i)$, and m outcomes O_j with probabilities $P(O_j|S_i)$, then:

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

Activity 4: Flipping a flipped classroom

Problem 11

Create one exercise using the Bayes' theorem. To help you:

1. Find an interesting application where you need to know something and you may run a test to figure it out.
2. Look online to see if you can find probabilities that appear realistic.
3. Pose the question of searching for a probability.
4. Solve the question. [22]

[22] Bonus :) if the result is counterintuitive!

Answer to Problem 11.

5. Discrete random variables

Every worksheet will work as follows.

1. You will be asked to form a group with other students in the class: you can make this as big or as small as you'd like, but groups of 4-5 work best.
2. Read through the worksheet, discussing any questions with the other members of your group.
   - You can call me at any time for help!
   - I will also be interrupting you for general guidance and announcements at random points during the class time.
3. Answer each question (preferably in the order provided) to the best of your knowledge.
4. While collaboration between students is highly encouraged and expected, each student has to submit their own version.
5. You will have 24 hours (see gradescope) to submit your work.

Activity 1: Basic discrete random variable questions

A movie magazine gives each movie they review a score from 0 to 4 stars – no movie is allowed to have a fractional number of stars though. Over the years, the magazine has observed that the movies get a score that is distributed as a discrete random variable X with probability mass function

$$p(x) = \begin{cases} \frac{2x+2}{c}, & x = 0, 1, 2, 3, 4 \\ 0, & \text{otherwise.} \end{cases}$$

Problem 1: Probability mass functions

What is the value of c for which $p(x)$ is a valid probability mass function?

Answer to Problem 1.

This can prove very useful. When we have an idea for the probability mass function of a *discrete* random variable, but we are missing part of it (e.g., missing a coefficient's value), then we can sum over all possible values of the pmf and it should equate to 1. This summation only holds for *discrete random variables*. More on what we anticipate for *continuous random variables* that are allowed to take any real value within a range in Lecture 7.

Problem 2: Constructing cumulative distribution functions

Using the same probability mass function $p(x)$ that you were given in Problem 1 (replace the value for c that you calculated), what is the cumulative distribution function (cdf) of discrete random variable X? We could write it as a summation! That is,

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

Use the cdf to calculate the probability that a movie gets up to 2 stars (including 2 stars) out of 4.

Answer to Problem 2.

$P(X \leq 2) =$

Problem 3: Calculating probabilities

For the same distribution, calculate the following probabilities. [23]

Answer to Problem 3.

$P(1 \leq X \leq 3) =$

$P(1 < X \leq 3) =$

[23] For discrete random variables, we can calculate the probability of X being between a and b in many different ways:

- $P(a \leq X \leq b) = \sum_{x=a}^{b} p(x)$.
- $P(a \leq X \leq b) = P(a-1 < X \leq b) = F(b) - F(a-1)$.

Be careful with $\leq$ vs. $<$.

Activity 2: Binomial, geometric, or hypergeometric?

Problem 4

A foundry has received an order for **5 castings**, made from precious metals. Each casting produced is of high quality (and hence can be sold to the customer) with probability 0.97. All castings are produced independently. You decide to schedule 6 castings for production, knowing full well that the customer only wants 5. What is the probability you get more than or equal to 5 high quality castings? [24]

Answer to Problem 4.

[24] Let X be the number of high quality castings produced out of the 6 you tried to produce.

1. What is X distributed as?
2. What is $P(X \geq 5)$? Can we make the claim that

$$P(X \geq 5) = P(X = 5) + P(X = 6)?$$

Problem 5

The foundry from earlier has received the same order for **5 castings**. However, instead of producing new ones, they use a batch of older castings already produced. The batch contains 100 castings, 97 of whom are of high quality. They decide to pick 6 of them at random and give them to the customer. What is the probability that the customer gets more than or equal to 5 high quality castings in the sample of 6 they receive? [25]

Answer to Problem 5.

[25] Once again, let X be the number of high quality castings you get in the group of 6 castings you pick from the batch.

1. What is X distributed as now?
2. Which formula should we use to calculate $P(X = 5)$ and $P(X = 6)$?

Problem 6

5% of all bits (a signal of 0 or 1) transmitted are sent in error (a 0 is sent instead of a 1, or vice versa). The message stops transmitting when the first bit is transmitted in error. Let X be the length of the message. What is the probability that $X = 5$?

Answer to Problem 6.

Problem 7

For the length of the message from Problem 6, what is the probability that $X \leq 5$?

Answer to Problem 7.

Activity 3: Overbooking

An airline has found that 10% of people buying a first class ticket do not show up to travel on the day of their flight (and independent of one another). This is why the company has decided to **sell 32 first class tickets for a flight that contains 30 first class seats**. If more than 30 first class ticket holders show up, then the company has to pay a penalty.

Problem 8

What is the probability the company does not pay a penalty for a flight? Equivalently, what is the probability that at most 30 first class passengers show up?

Answer to Problem 8.

Problem 9

A flight just departed without having to pay a penalty. What is the probability the flight departed with at least one empty first class seat?

Answer to Problem 9.

By the way, congratulations for answering this! This was a question in **last year's midterm exam**!

6. Discrete random variables

Every worksheet will work as follows.

1. You will be asked to form a group with other students in the class: you can make this as big or as small as you'd like, but groups of 4-5 work best.
2. Read through the worksheet, discussing any questions with the other members of your group.
   - You can call me at any time for help!
   - I will also be interrupting you for general guidance and announcements at random points during the class time.
3. Answer each question (preferably in the order provided) to the best of your knowledge.
4. While collaboration between students is highly encouraged and expected, each student has to submit their own version.
5. You will have 24 hours (see gradescope) to submit your work.

Activity 1: Selling on craigslist

You have decided to sell an item on craigslist. You have given no price in your ad: instead, you ask the interested people reading the ad to submit an offer. Their offers will be uniformly distributed between 0 and 99 dollars. [26] You would like to make at least \$75 from the item, so any offer that is greater than or equal to \$75 dollars would be acceptable. In addition, you receive *exactly* one offer every day since posting the ad.

[26] You may assume that only whole dollar amount offers can be given, such as \$37 or \$73

Problem 1

What is the probability a random offer you receive is higher than or equal to \$75?

Answer to Problem 1.

$P(\text{offer} \geq \$75) =$

Problem 2

What is the probability you sell the item on the 3rd day? What distribution does this follow? [27]

[27] Hint: consider each offer/day a success or a failure... Does that sound *geometric*?

Answer to Problem 2.

Problem 3

What is the probability you sell the item in one of the first 3 days? [28]

Answer to Problem 3.

[28] Again, assuming that the day you sell the item follows a gemoetric distribution, then you'd want to calculate $P(X = 1) + P(X = 2) + P(X = 3)$.

Problem 4

You decide not to look at your email for 10 days. When opening your email again to check on the (10) offers you have received, what is the probability that at least one of them offers you $75 or more? What distribution does the number of "successful" offers follow? [29]

Answer to Problem 4.

[29] Hint: consider as if you get 10 "tries" and need at least one of them be a "success"...

Problem 5

You decide not to look at your email for 3 days. What is the probability that you sell the item on the 6th day, given that in the first 3 days you receive no satisfying offers? Contrast this answer to your answer in Problem 2. [30]

Answer to Problem 5.

[30] Hint: use conditional probabilities and the geometric distribution probability mass function.

Will you look at this. *Knowing that the first three days are failed* does not change (either increase or decrease) the probability of getting a success in the next three days! This is observed because our answers in Problems 2 and 5 are the same. We say that the geometric distribution is **memoryless**.

Activity 2: Practice with the Poisson distribution

Problem 6

A transportation engineer is collecting data during rush hour on the intersection of University and Neil in Champaign. They have noticed that the number of vehicles during rush hour follows a Poisson distribution with a rate of 7.5 per minute. What is the probability that exactly 10 vehicles pass through the intersection in the next minute?

Answer to Problem 6.

Problem 7

What is the probability that exactly 10 vehicles pass through the intersection in the **next 2 minutes**?

Answer to Problem 7.

Note how our first order of business with a Poisson distributed random variable is to convert the rate λ in the necessary time units. In Problem 6, we needed $\lambda = 7.5/\text{minute}$, whereas in Problem 7, we had to change this to $\lambda = 15$ every 2 minutes.

Problem 8

During a pandemic, the number of patients in need of a special treatment is Poisson distributed with a rate λ of 1 patient every 4 hours. Assuming a hospital can offer 4 of those treatments per day (24 hours), what is the probability the hospital runs out of available treatments during a day (24 hours) of operations? [31]

[31] Assume this means that at least 5 patients show up in a day.

Answer to Problem 8.

Problem 9

Scientists in the midwestern states have observed an increase in the frequency of floods and river overflows. The most recent estimate is that a devastating flood in some location in the midwest may happen with a rate of 1 every 50 years, whereas earlier estimates had placed that number in a rate of 1 every 200 years. Quantify the change in probability of a devastating flood in the midwest between the two estimates. What is the probability of seeing at least one devastating flood in the next year with the earlier and the more recent estimates?

Answer to Problem 9.

Activity 3: Interesting Poisson distribution properties

Problem 10

We saw in class the probability mass function for a Poisson distributed random variable with rate λ. [32] 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 $P(T > 1 \text{ year})$?

[32] It is $P(X = x) = e^{-\lambda} \cdot \frac{\lambda^x}{x!}$.

Answer to Problem 10.

$P(T > 1 \text{ year}) =$

So, the **time to the next event** is related to the number of events.. But, time is of a continuous nature whereas the number of events is discrete (has to be integer). Does that mean there is a relationship between a Poisson distributed random variable and a continuous (real-valued) random variable? *Interesting*: let's keep that in mind for our next lecture on continuous random variables.

7. Continuous random variables: the uniform and the exponential distribution

Every worksheet will work as follows.

1. You will be asked to form a group with other students in the class: you can make this as big or as small as you'd like, but groups of 4-5 work best.
2. Read through the worksheet, discussing any questions with the other members of your group.
   - You can call me at any time for help!
   - I will also be interrupting you for general guidance and announcements at random points during the class time.
3. Answer each question (preferably in the order provided) to the best of your knowledge.
4. While collaboration between students is highly encouraged and expected, each student has to submit their own version.
5. You will have 24 hours (see gradescope) to submit your work.

Activity 1: Basic continuous probability distribution properties

Let X be a continuous random variable measuring the current (in milliamperes, mA) in a wire with probability density function (pdf) given by $f(x) = 0.05$, for $0 \leq x \leq \alpha$. Answer the following questions.

Problem 1: Valid pdf?

What is α in order for $f(x)$ to be a valid pdf? [33]

Answer to Problem 1.

[33] Recall that this means that $f(x) \geq 0$ (which is clearly true here) **and** that $\int f(x)dx = 1$ over *all values* that random variable X is allowed to take...

Problem 2: Constructing cumulative distribution functions

What is the cumulative distribution function? [34]

Answer to Problem 2.

[34] How is a cdf defined for *continuous* random variables?

Problem 3: Calculating probabilities

The wire is said to be overheating if the current is more than 10mA. It is also said to be working within the normal range of operations if it is between 5 and 10mA. Answer the following questions:

- What is the probability the wire is overheating?
- What is the probability the wire is within its normal range?

Answer to Problem 3.

Activity 2: The exponential distribution

The time until the next customer arrives is exponentially distributed with a rate of 1 customer every 10 minutes. Answer the following questions.

Problem 4: Calculating probabilities

Let T be the time until the next customer arrives. What is the probability the next customer shows up in the next:

a) 1 minute? b) 5 minutes? c) 10 minutes? d) 20 minutes?

Answer to Problem 4.

a) 1 minute: $P(T \leq 1) =$

b) 5 minutes: $P(T \leq 5) =$

c) 10 minutes: $P(T \leq 10) =$

d) 20 minutes: $P(T \leq 20) =$

This is the big difference between using a rate for a Poisson or for an exponential distribution. For the Poisson distribution, we first convert the rate into the time units of the question. For the exponential distribution, we simply multiply the rate by the necessary time interval length!

For example, for a rate $\lambda = 3$ per minute, the probability we see 5 customers in 3 minutes would be written as

$$P(X = x) = e^{-\lambda} \cdot \frac{\lambda^x}{x!} \implies P(X = 10) = e^{-9} \cdot \frac{9^5}{5!},$$

because λ would be set equal to 9 (customers per 3 minutes).

On the other hand, for the same rate, the probability the next customer shows up within the first 3 minutes would be written as

$$P(T \leq t) = 1 - e^{-\lambda t} \implies P(T \leq 3) = 1 - e^{-3 \cdot 3} = 1 - e^{-9}.$$

Problem 5: Inverting the question

The employee of the store wants to take a break, but they do not want to miss the next customer arrival. How long should the break be for in order to have a 50% chance of not missing the next customer? [35]

[35] In essence, we want the time $\tilde{t}$ such that the probability $P(T \leq \tilde{t}) = 0.5$.

Answer to Problem 5.

Problem 6: Taking a break

In the same store, assume that the employee takes a 2 minute break during which no customer arrived. What is the probability the next customer does not arrive in the next 5 minutes now? Contrast it to your answer for $P(T \leq 5)$ in Problem 4.

Answer to Problem 6.

Based on your answer in Problem 6, we must be deducing that the exponential distribution is indeed **memoryless**. In the next activity, we prove that property for all exponentially distributed random variables.

As a reminder, we say that a distribution is memoryless if for random variable X following that distribution, we have that

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

For example, in the case of the time of arrival of the next customer (random variable T) in more than 5 minutes from now, we would write $P(T > 5)$. After we take a break for 2 minutes, then that same probability would be $P(T > 7 | T > 2)$.

Activity 3: Memorylessness

Let's check the proof!

Problem 7: Memorylessness proof

In the first step of the proof, we will calculate the right hand side of what we want to show: $P(X > s+t|X > s)$. Using conditional probabilities [36], what is this equal to?

[36] Remember that $P(A|B) = \frac{P(A \cap B)}{P(B)}$.

Answer to Problem 7.

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

Problem 8: Memorylessness proof (cont'd)

In the second step of the proof, we need to make one observation. If $A \subseteq B$ then $A \cap B = A$. Is this the case here for the denominator of your answer in Problem 7? Specifically what can you write about $P(X > s+t \ \cap \ X > s)$? [37]

[37] Recall what you can say for $P(A \cap B)$ if $B \subseteq A$...

Answer to Problem 8.

$P(X > s+t \ \cap \ X > s) =$

Problem 9: Memorylessness proof (cont'd)

In the third step of the proof, consider the original identity we are trying to show: $P(X > s+t|X > s) = P(X > t)$. Assume that X is exponentially distributed with rate λ: that is, $P(X > t) = e^{-\lambda t}$. Fill in the blanks below to get the result. Hints are given along the way (look at the final equation!).

Answer to Problem 9.

$$P(X > s+t|X > s) = \frac{P(X > s+t \ \cap \ X > s)}{P(X > s)} = \quad \text{(from Prob. 7)}$$

$$= \frac{P(X > s+t)}{P(X > s)} = \quad \text{(from Prob. 8)}$$

$$= \frac{\qquad\qquad\qquad}{\qquad\qquad\qquad} =$$

$$= \_\_\_\_\_ = P(X > t).$$

The next worksheet problem comes pre-filled. However, going through this will help with understanding why the exponential distribution is unique.

Problem 10: Only the exponential distribution!

So, is the exponential distribution one of many distributions to be memoryless? The answer is a resounding no: it is the **only** continuous distribution to be memoryless. [38] To show that, we need to use all of the tools in our toolbox.

[38] Again: the **only continuous** distribution. We saw a discrete distribution that is memoryless not too long ago!

Answer to Problem 10.

Like we did in Problem 9:

$$P(X > s+t | X > s) = \frac{P(X > s+t \cap X > s)}{P(X > s)} = \qquad \text{(from Prob. 7)}$$
$$= \frac{P(X > s+t)}{P(X > s)}. \qquad \text{(from Prob. 8)}$$

We now note that if X is memoryless, then $P(X > s+t | X > s) = P(X > t)$:

$$P(X > s+t | X > s) = P(X > t) \implies \frac{P(X > s+t)}{P(X > s)} = P(X > t)$$
$$\implies P(X > s+t) = P(X > s) \cdot P(X > t).$$

For simplicity, define $\overline{F}(x) = P(X > x)$ and hence our previous equality becomes:

$$\overline{F}(s+t) = \overline{F}(s) \cdot \overline{F}(t).$$

Now, take the logarithms on both sides!

$$\ln \overline{F}(s+t) = \ln \overline{F}(s) \cdot \overline{F}(t) = \ln \overline{F}(s) + \ln \overline{F}(t).$$

Once again, for simplicity, define $g(x) = \ln \overline{F}(x)$, rendering our final equality as:

$$g(s+t) = g(s) + g(t).$$

The **only continuous function** that satisfies $g(s+t) = g(s) + g(t)$ is the linear one (see the citation for the proof in the end), hence we deduce that $g(x) = \beta \cdot x$, for some β.
Combine everything:

1. We need $P(X > s+t) = P(X > s) \cdot P(X > t)$ for memorylessness to hold.
2. Equivalently, after some definitions, we saw this is equivalent to $g(s+t) = g(s) + g(t)$, where $g(x) = \ln \overline{F}(x)$.
3. We saw though that this is only true if $g(s)$ is linear, that is $g(x) = \beta \cdot x$.

We replace backwards: $g(s) = \beta \cdot s \implies \ln \overline{F}(s) = \beta \cdot s \implies \overline{F}(s) = e^{\beta \cdot s} \implies F(s) = 1 - e^{-\beta \cdot s}$. Recall that the exponential distribution has $F(x) = 1 - e^{-\lambda \cdot x}$. Pretty close! Letting $\beta = -\lambda$ finishes the proof!

Problem 11: Just checking

Circle the correct choice in each of the three questions.

Answer to Problem 11.

The exponential distribution is the only *continuous* probability distribution that is memoryless.

a) True b) False

The exponential distribution is the only probability distribution that is memoryless.

a) True b) False

Assume that the time until the next event is exponentially distributed. At 1pm, we calculated the probability that the next event happens before or on 1.30pm is 30%. It is 1.30pm now, and no event has happened yet. What is the probability that the next event happens before or on 2pm? Pick the correct answer in the multiple choice question.

a) It is higher than 30%: it was 30% at 1pm and it never happened, so now it should happen with higher certainty.

b) It is lower than 30%: it was 30% at 1pm and it never happened, so now we should consider that maybe our probability calculation was off to begin with.

c) It is equal to 30%: it was 30% at 1pm and while it never happened, memorylessness states that the probability is still 30%.

Activity 4: Exponential and Poisson

We will play with this relationship (and introduce Gamma/Erlang distributed random variables, too) even more in the next lecture. That said, let us practice the fact that **the exponential distribution and the Poisson distribution go hand-in-hand**. [39]

[39] See the notes for more details!

If the time between events is **exponentially distributed** with rate λ then the number of events in some time t is **Poisson distributed** with rate $\lambda \cdot t$, and vice versa.

8. Continuous random variables: the Gamma/Erlang distribution and the normal distribution

Every worksheet will work as follows.

1. You will be asked to form a group with other students in the class: you can make this as big or as small as you'd like, but groups of 4-5 work best.
2. Read through the worksheet, discussing any questions with the other members of your group.
   - You can call me at any time for help!
   - I will also be interrupting you for general guidance and announcements at random points during the class time.
3. Answer each question (preferably in the order provided) to the best of your knowledge.
4. While collaboration between students is highly encouraged and expected, each student has to submit their own version.
5. You will have 24 hours (see gradescope) to submit your work.

Activity 1: Exponential, Poisson, and Erlang

A manufacturing process requires the completion of 4 small steps: pre-processing, processing, inspection, packaging. Each step of the process requires time that is that is exponentially distributed with a rate of 1 completion every 3 minutes. That is, all steps are exponentially distributed with $\lambda = 1/3$ minutes. The steps have to be performed sequentially. Answer the following questions.

Problem 1

What is the probability that the "pre-processing" step alone (that is, **the first step alone**, from its start to its end) is completed within 4 minutes? [40]

[40] Is the time in which a step, any step, is completed Poisson, exponential, or Erlang? Based on your answer, does it matter if you were asked about any other step or would your answer stay the same?

Answer to Problem 1.

Problem 2

An inspector shows up to watch the operations take place. They only have time to be there for 10 minutes. What is the probability that there are exactly 3 steps that are completed in the 10 minutes after the inspector is there? [41]

[41] First, calculate the rate of completed steps in 10 minutes; then decide if you are using Poisson, exponential, or Erlang..

Answer to Problem 2.

Problem 3

What is the probability that a manufacturing order is completed within 10 minutes (all 4 steps, one after the other)? [42]

[42] Feel free to use an online calculator for your integral if you need to.

Answer to Problem 3.

We can answer Problem 3 using random variable T for the time until 4 steps are completed: T is Erlang with $k = 4$ steps and $\lambda = 1/3$ minutes. We could also calculated this using random variable X for the number of steps that are completed in 10 minutes: X is a Poisson random variable with $\lambda = 10/3$. Then, we can show that:

$$P(T \leq 10 \text{ minutes}) = P(X \geq 4 \text{ steps}).$$

Activity 2: The normal distribution

In this part of the worksheet, we turn our focus to the normal distribution. For this next part, we assume that $\Phi(z)$ is the standard normal distribution cumulative function. We can use the z-table provided in the last page to find $\Phi(z)$!

Problem 4: Converting to z values

Let's practice with converting to the proper z values. Let X be a normally distributed random variable with $\mu = 10, \sigma^2 = 4$.

Answer to Problem 4.

- $X = 12 \implies z =$
- $X = 8 \implies z =$
- $X = 4 \implies z =$

Problem 5: Simple normal distribution probabilities

For the previous random variable $X \sim \mathcal{N}(10, 4)$, find the probabilities. Use the z values you calculated earlier. [43]

[43] A z-table as described in the lecture notes is provided in the last page of the worksheet. Also recall that $\Phi(-z) = 1 - \Phi(z)$ due to symmetry.

Answer to Problem 5.

- $P(X \leq 12) =$
- $P(X \geq 4) =$
- $P(4 \leq X \leq 12) =$

Problem 6: Interesting probabilities

As we saw in class, the normal distribution is centered at μ. [44] A follow-up question would be to find the range of values centered at μ that satisfy a certain probability. Let's see an example here: what is the probability that X is within 1 unit from its center (μ), that is what is the probability that X is between 9 and 11? How about 2 units from its mean?

[44] So, following the previous random variable X, it would be centered at 10.

Answer to Problem 6.

- $P(9 \leq X \leq 11) =$
- $P(8 \leq X \leq 12) =$

Problem 7

Let's now focus on the opposite problem. What should the range be (centered at μ) so that the probability of the range is 50%? In essence, what should a be in order for $P(\mu - a \leq X \leq \mu + a) = 0.5$? Remember that earlier we calculated two range probabilities:

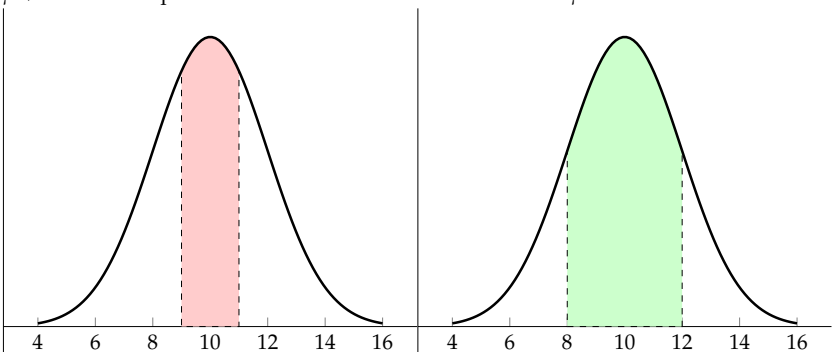
1. $P(9 \leq X \leq 11) = 0.383$.
2. $P(8 \leq X \leq 12) = 0.6826$.

Based on that, we should anticipate a to fall somewhere above 1 unit but below 2 units. But how big should it be exactly? Recall that we assume that $X \sim \mathcal{N}(10, 4)$. [45]

[45] $\mu = 10, \sigma^2 = 4 \implies \sigma = 2$.

Answer to Problem 7.

Figure 2: What should a be for $P(\mu - a \leq X \leq \mu + a) = 0.95$? Here we show in red the area for $P(\mu - 1 \leq X \leq \mu + 1) = 0.383$ and in green the area for $P(\mu - 2 \leq X \leq \mu + 2) = 0.6826$. It should make sense that $P(\mu - a \leq X \leq \mu + a) = 0.95$, $\mu - a$ and $\mu + a$ have to be points located even farther from the center μ.

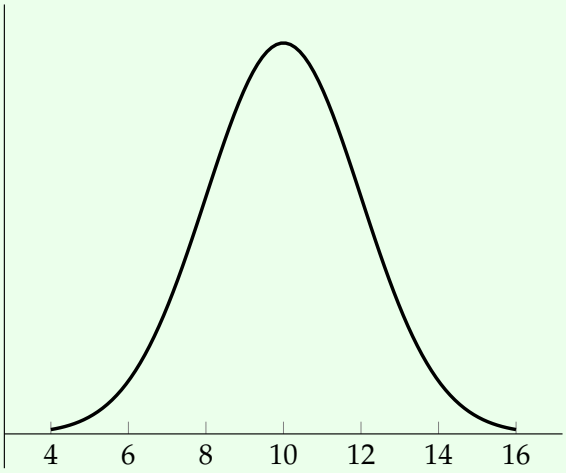


Problem 8

For $X \sim \mathcal{N}(10, 4)$, we want to see how big the range should be in order to have a probability equal to 95% that X falls in that range. In essence, we are now interested in $P(\mu - a \leq X \leq \mu + a) = 0.95$. Graphically:

Answer to Problem 8.

After you have found a, try to draw the resultant area!



Problem 9

We may have started observing that a does not depend on μ all that much. Instead it depends on σ. For example, seeing as we may write $P(\mu - a \leq X \leq \mu + a) = 0.5$ as a probability of z as follows:

1. $z_1 = \frac{\mu - \alpha - \mu}{\sigma} = -\frac{\alpha}{\sigma}$
2. $z_2 = \frac{\mu + \alpha - \mu}{\sigma} = \frac{\alpha}{\sigma}$.
3. Note that $z_1 = -z_2$.

Based on that: $P(\mu - a \leq X \leq \mu + a) = P(z_1 \leq Z \leq z_2)$. Now recall that $P(z_1 \leq Z \leq z_2) = \Phi(z_2) - \Phi(z_1) = \Phi(z_2) - \Phi(-z_2) = \Phi(z_2) - (1 - \Phi(z_2)) = 2\Phi(z_2) = 2\Phi(\frac{\alpha}{\sigma}) - 1$.

With that in mind, answer the following three questions. Try to answer them generally, not only for $X \sim \mathcal{N}(10, 4)$.

Answer to Problem 9.

- $P(\mu - \sigma \leq X \leq \mu + \sigma) =$
- $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) =$
- $P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) =$

Based on our answers, we have the following realization: σ is very important. Any normally distributed random variable probability can be expressed as a "distance" in terms of "σ". In essence:

$$\boxed{x = \mu + z\sigma \implies F(x) = \Phi(z).}$$

See this link for an interesting mnemonic, called the **68-95-99.7 rule**:

https://en.wikipedia.org/wiki/68–95–99.7_rule

Activity 3: Contrasting exponentials

Consider two exponentially distributed random variables X_1, X_2 with rates λ_1, λ_2.

Problem 10

What is the probability of $X_1 > X_2$, given that $X_2 = x$? [46]

[46] Can't we say that $P(X_1 > X_2 | X_2 = x)$ is simply $P(X_1 > x)$?

Answer to Problem 10.

$$P(X_1 > X_2 | X_2 = x) = P(X_1 > x) =$$

Problem 11

What is the probability of $X_1 > X_2$, in general? Recall the total probability law? [47] We can apply this to continuous distributions, too! We cannot sum here, but we may integrate. Let X_1 be random variable distributed with pdf $f(\cdot)$ and X_2 be a random variable distributed with pdf $g(\cdot)$, then:

[47] For an event B, and m mutually exclusive and collectively exhaustive events A_i, $i = 1, \ldots, m$, then we have $P(B) = \sum_{i=1}^{m} P(B|A_i) \cdot P(A_i)$.

$$\boxed{P(X_1 > X_2) = \int_{-\infty}^{+\infty} P(X_1 > X_2 | X_2 = x) g(x) dx.}$$

Use this to answer the following question:

Answer to Problem 11.

$$P(X_1 > X_2) =$$

From this last part, we see that for two exponentially distributed random variables X_1, X_2 with rates λ_1, λ_2, respectively, we have:

$$\boxed{P(X_1 \leq X_2) = \frac{\lambda_1}{\lambda_1 + \lambda_2}}$$

Problem 12

Two employees are using a website to place an order with a supplier at exactly the same time. The first person is more tech savvy and completes an order with rate 1 order every 3 minutes. The second person is just starting the job and learning, so they are a little slower and complete an order with rate 1 order every 5 minutes. Both times are exponentially distributed. What is the probability that the second person completes the order faster than or equal to the time the first person takes to complete an order?

Answer to Problem 12.

As a summary of the exponential distribution:

1. If the time between events is exponentially distributed, then the time until the k-th event ($k > 1$) is Erlang distributed, and the number of events within some time is Poisson distributed.

   Customers arrive with time that is exponentially distributed with rate $\lambda = 3$ customers every hour. Then:

   - "What is the probability the next customer shows up in 10 minutes?" is an **exponential distribution** question.
   - "What is the probability there are more than one customer in the next 20 minutes" is a **Poisson distribution** type of question. Recall that we will update λ to be in 20 minute intervals $\implies \lambda = 1$ customer every 20 minutes.
   - "What is the probability the second customer shows up within 20 minutes" is an **Erlang distribution** question.

2. The exponential distribution is **memoryless**! Hence:

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

3. If $X_1, X_2, \ldots, X_k$ are independent exponentially distributed random variables with rates $\lambda_1, \lambda_2, \ldots, \lambda_k$ respectively, then the probability X_i is the smallest one (i.e., $P(X_i < X_1, X_i < X_2, \ldots)$) is:

$$\frac{\lambda_i}{\lambda_1 + \lambda_2 + \ldots + \lambda_k}.$$

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

9. Expectations and variances

Every worksheet will work as follows.

1. You will be asked to form a group with other students in the class: you can make this as big or as small as you'd like, but groups of 4-5 work best.
2. Read through the worksheet, discussing any questions with the other members of your group.
   - You can call me at any time for help!
   - I will also be interrupting you for general guidance and announcements at random points during the class time.
3. Answer each question (preferably in the order provided) to the best of your knowledge.
4. While collaboration between students is highly encouraged and expected, each student has to submit their own version.
5. You will have 24 hours (see gradescope) to submit your work.

Activity 1: Basic expectation and variance properties

First, let's practice some basic expectation and variance properties. Answer the following questions. If False, correct the statement. [48]

[48] For these statements, consult with pages 3-4 and 8-9 of the notes.

Problem 1: True or False

Answer to Problem 1.

$$E[3X+2] = 3E[X]$$

a) True b) False

$$E\left[3X^2\right] = 3E\left[X^2\right]$$

a) True b) False

$$Var[3X+2] = 9Var[X] + 4$$

a) True b) False

For the next one, you may assume that X and Y are independent random variables.

$$Var[3X+Y] = 9Var[X] + Var[Y]$$

a) True b) False

Activity 2: Rating the latest Ant-Man and the Wasp movie

A movie magazine decides to allow its reviewers to provide a rating ranging from 1 to 4 stars. Here, we are specifically focusing on Ant-Man and the Wasp: Quantumania, the latest blockbuster in the Marvel Cinematic Universe. In the next two problems, we will see what the expected stars the movie will get if reviewers are allowed to provide an integer number of stars, or when they are allowed to providen *any real number* of stars.

Problem 2: Integer stars

Assume that the number of stars for any movie is represented as a **discrete random variable** X with pmf equal to $p(x) = \frac{x^2}{30}$ for $x = 1, 2, 3, 4$. What is the expected number of stars for any movie (that is, what is the expectation of X)? What is the variance of X? [49]

[49] For the variance, recall that you may use the formula

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

Answer to Problem 2.

$E[X] =$

$Var[X] =$

Problem 3: Real stars

Now, assume that the number of stars (again, from 1 to 4) a reviewer can give to a movie can be represented as a **continuous random variable** X with pdf equal to $f(x) = \frac{x^2}{21}$ for $1 \leq x \leq 4$. [50] What is the expected value of X in this case? How about the variance of X? [51]

[50] This means that a reviewer may opt to give a movie 3.5 stars, while another may decide to give out 2.78554 stars.

[51] The same variance formula $Var[X] = E\left[X^2\right] - (E[X])^2$ holds for both continuous and discrete random variables.

Answer to Problem 3.

$E[X] =$

$Var[X] =$

Activity 3: A printer replacement policy

An office has bought a new printer which is supposed to have lifetime that is **exponentially distributed** with an expected lifetime at 2 years. [52] Answer the following questions.

[52] Usually you were provided rates for exponential distributions: however, we may now equivalently provide the expectation, since we know that for an exponentially distributed random variable X, its expectation $E[X]$ is $\frac{1}{\lambda}$, where λ is the rate.

Problem 4

What is the probability that the printer still works after 2 years?

Answer to Problem 4.

Problem 5

Based on your answer in Problem 4, if the company buys $n = 30$ printers for their 30 offices, how many of them should we expect to still work after 2 years? [53]

[53] To answer this question, consider what type of distribution fits the question.. It is as if we have $n = 30$ "tries" for printers to "survive for two years"..

Answer to Problem 5.

Problem 6

The company is starting a new policy. They will replace the printer either when it breaks down (recall that it breaks down in time that is exponentially distributed with like in Problem 4), or when it becomes 2 years old, whichever comes first. What is the expected lifetime of every printer the company buys? [54]

Note that if we allowed a printer to work until it breaks down, then we would (on average) expect to replace a printer every two years – as this is the given expectation of the exponential distribution. But if we replace any printer the moment they are 2 years old, then the expectation should go down, shouldn't it?

[54] An equivalent question: every how often does the company buy a new printer?

Answer to Problem 6.

*Activity 4: The law of **total expectation***

In this activity, we derive the law of total expectation for both discrete random variables (in the form of a summation) and continuous random variables (in the form of an integral).

Problem 7: A simple case

Let us begin with something simple. An experiment is successful 90% of the time (and failed the remaining time). We perform 10 experiments. How many should we expect to be successful? [55]

[55] Think about what distribution this could be modeled as. Then, you may use the expectation formula for that specific distribution!

Answer to Problem 7.

Problem 8: External conditions

Let us complicate this slightly. Once again we perform 10 experiments, where an experiment can be successful or failed. However, the success probability depends on some external conditions. If the conditions are good, the probability of success is 95%; in average conditions, the probability becomes 90%; in bad conditions, the probability is lower at 75%. [56] All experiments will take place at the same conditions; so either all experiments will have good conditions, or all experiments will have average conditions, or all experiments will have bad conditions.

[56] Think of it like that: if the conditions are good, then the expected number of successes would be $0.95 \cdot 10 = 9.5$; if the conditions are average, then the expected number of successes would be $0.90 \cdot 10 = 9$; finally, if the conditions are bad, then the expectation becomes $0.75 \cdot 10 = 7.5$. Could we multiply each expectation with its respective probability? Are we allowed to do that?

Assuming conditions are equally probable (that is, good/average/bad conditions appear $\frac{1}{3}$ of the time), what is the expected number of successful experiments now?

Answer to Problem 8.

Problem 9: Generalizing the result

Can we generalize the previous result? What if we had m different possible conditions, each appearing with probability $\pi_i, i = 1, \ldots, m$ and each leading to probability of success p_i? How many experiments should we expect to be successful if we perform $n = 10$ experiments?

Answer to Problem 9.

Based on your answers so far, we observe that if we can partition the space in m mutually exclusive and collectively exhaustive events A_i each with probability of appearing equal to $P(A_i)$ [57], then the expected value of random variable X can be found by:

$$\boxed{E[X] = \sum_{i=1}^{m} E[X|A_i] \cdot P(A_i)}$$

How do you think this should look like for continuous random variables? [58]

[57] Look at this! This is also the setup for the law of total probability (see Lecture 4).

[58] Recall during the previous lecture we saw that summations become integrations, and probabilities become probability distribution functions..

Problem 10

Consider a continuous random variable with pdf $f(x) = \frac{1}{2}(1 + \theta \cdot x)$ for $-1 \leq x \leq 1$, where θ is uniformly distributed between 0 and 1. What is the expected value of X?

Answer to Problem 10.

The law of total expectation applies to continuous random variables, too. Consider X, Y as continuous random variables, such that we know $E[X|Y = y]$. Also assume that Y has pdf $g(y)$. Then, we have:

$$\boxed{E[X] = \int_{-\infty}^{+\infty} E[X|Y = y] \cdot g(y) dy}$$

Activity 5: Expectations of functions

This is a slightly bigger question, that I am anticipating you can work on after lecture. Recall that for any function $g(X)$ of a random variable X, we can calculate its expectation as:

$$E\left[g(X)\right] = \begin{cases} \sum_S g(x) \cdot p(x), & \text{if } X \text{ is discrete,} \\ \int_S g(x) \cdot f(x)dx, & \text{if } X \text{ is continuous.} \end{cases}$$

Problem 11

Consider again the printer from Activity 3. The speed with which the printer works (and prints documents) is a function of its age. When the printer is x years old, its speed is given by $g(x) = \frac{\sqrt{x+1}}{x+1}$ velocity units. For example, when it is just bought, and the printer is 0 years old, its speed is equal to 1 velocity unit; on the other hand, after 2 years its speed drops to $\sqrt{3}/3$ velocity units. [59]

[59] If it helps ground the question better, you may think of 1 velocity unit as the biggest speed there is, and 0 velocity units as the lowest speed there is.

What is the expected speed of a printer if the company implements the policy of replacing the printer when it breaks down or when it becomes 2 years old, whichever comes first?

Answer to Problem 11.

10. The central limit theorem

Every worksheet will work as follows.

1. You will be asked to form a group with other students in the class: you can make this as big or as small as you'd like, but groups of 4-5 work best.
2. Read through the worksheet, discussing any questions with the other members of your group.
   - You can call me at any time for help!
   - I will also be interrupting you for general guidance and announcements at random points during the class time.
3. Answer each question (preferably in the order provided) to the best of your knowledge.
4. While collaboration between students is highly encouraged and expected, each student has to submit their own version.
5. You will have 24 hours (see gradescope) to submit your work.

Activity 1: Normally distributed random variables

We begin this section by mentioning something very important:

Adding two (or more) independent normal distributed random variables together results in a normal distribution.

In mathematical terms: if $X_i, i = 1, \ldots, n$ are independent random variables distributed normally, then $Z = \sum_{i=1}^{n} X_i$ is also normally distributed. The same is true for any linear combination, i.e., $Z = \sum_{i=1}^{n} a_i \cdot X_i$ is normally distributed. Recall that to fully describe a normally distributed random variable, we simply need its expectation and variance.

Problem 1: Expectation and variance of the sum of normally distributed random variables

Consider two normally distributed random variables $X \sim \mathcal{N}\left(\mu_X, \sigma_X^2\right)$ and $Y \sim \mathcal{N}\left(\mu_Y, \sigma_Y^2\right)$, which are independent. What is the expectation and variance of $X + Y$?

Answer to Problem 1.

In general, $Z = \sum_{i=1}^{n} a_i \cdot X_i$ is normally distributed with $\mu = E[Z] = \sum_{i=1}^{n} a_i \cdot E[X_i]$ and $\sigma^2 = Var[Z] = \sum_{i=1}^{n} a_i^2 \cdot Var[X_i]$.

Problem 2: Application

You own a portfolio of stocks that consists of 5 stocks S_1 and 10 stocks S_2. Let X_1 be the price of stock S_1 and X_2 the price of stock S_2 one year from now. X_1 is normally distributed with $\mathcal{N}(50, 100)$[60] and X_2 is normally distributed with $\mathcal{N}(60, 100)$. Last, assume that the two stocks are totally unrelated (i.e., they are independent of one another).

What is the probability that your portfolio is worth more than \$1000 one year from now? What is the probability that your portfolio is worth less than \$800 one year from now?

[60] That is, its mean is 50 and its variance is 100.

Answer to Problem 2.

Problem 3: Comparing random variables

Consider two normally distributed random variables X and Y, which are independent. Can you make the claim that $X - Y$ is normally distributed? What is the expectation and variance of $X - Y$? [61]

[61] Sure we can! Remember that if $X_1, X_2, \ldots$ are normally distributed then any linear combination of them is also normally distributed!

Answer to Problem 3.

Problem 4: Comparing stock prices

Again, consider that X_1 is the price of stock S_1 and X_2 is the price of stock S_2 one year from now. Both are normally distributed: $X_1 \sim \mathcal{N}(50, 100)$ and $X_2 \sim \mathcal{N}(60, 100)$. Finally, assume that X_1, X_2 are independent.

What is the probability that S_1 (a single stock of that) is worth more than S_2 (again, consider only a single stock of that one, too) a year from now? In mathematical terms, what is $P(X_1 > X_2)$? [62]

[62] Think of $Z = X_1 - X_2$... If Z is a normally distributed random variable, then we can calculate the probability that $P(Z > 0)$, no?

Answer to Problem 4.

What about the case where two random variables (distributed normally) are **not independent**? One could make the argument that two stocks affect each other, and knowing that one has gone up may affect our perspective of the other one going up, too. In this case, where independence is hard to assume and use, can we still compare two stocks (or calculate the probability of our portfolio being above a desired value a year from now)?

The answer is yes! However, we need to introduce ideas like **dependence** and measure it with **covariance** and **correlation**. More on that in Lecture 12...

Activity 2: Using the central limit theorem

In the online course world, certain classes have thousands of students: these classes are sometimes referred to as *Massive Open Online Courses* (MOOC). On average there are 8000 students in a course. For the purposes of this exercise, we assume that a student successfully finishes a MOOC 7.5% of the time and that all students behave independently. Assume that probability is the same regardless of the course. Finally, assume that every class has **exactly** 8000 students.

Answer the following questions.

Problem 5: Setting up the distribution

What is the best distribution to model the number of students to successfully finish a single class? What is the mean and the variance of the number of students who successfully finish a single class? [63]

[63] Recall: you have 8000 students, each of whom may succeed or fail.. Which distribution is this? And what is its mean?

Answer to Problem 5.

Problem 6: Setting up the central limit theorem

Now, consider the case of an online course provider, such as Coursera. Assume the total number of classes the provider offers is equal to 1000. For each individual class, the number of students who successfully finish that specific class follows the distribution you found in Problem 5.

What distribution does the average number of students graduating from all class of the provider follow? The average number of students graduating in all 1000 classes can be found by summing the number of graduates in each class and dividing by 1000. [64]

[64] Does the central limit theorem apply? If so, then the distribution is...

Answer to Problem 6.

Activity 3: Setting up hypothesis testing

Recall that we have made the hypothesis that a student successfully finishes a class 7.5% of the time. We have also found that the number of students successfully finishing one class follows a binomial distribution (see Problem 5) and the average number of students across 1000 courses follows a normal distribution (based on Problem 6).

The online course provider has decided to check whether this "7.5%" success rate is true or not. They have decided to survey 15 of the 1000 courses they offer and they found the following: [65]

Course 1	588	Course 2	645	Course 3	632
Course 4	623	Course 5	635	Course 6	641
Course 7	644	Course 8	611	Course 9	630
Course 10	628	Course 11	569	Course 12	637
Course 13	635	Course 14	677	Course 15	610

[65] You will not need the numbers until Problem 8.

Problem 7: Checking the numbers

What should the average number of graduating students in these 15 classes be distributed as? Once again, to find the average number of graduates, simply add up the numbers for the 15 classes and divide by $n = 15$. [66] What is the mean and the variance of the distribution?

[66] Treat 15 as a large enough number for the central limit theorem to apply.

Answer to Problem 7.

Problem 8: Using the central limit theorem

From the numbers they found (see the tabulated data from Activity 3) that on average a pretty big 627 students in each class successfully finishes it [67].

[67] Recall they were expecting 600 on average, or $8000 \cdot 0.075$.

Based on the distribution you have identified in Problem 7, what is the probability that there are more than 627 students on average finishing each class? [68]

[68] Remember that z values that do not appear in the table (i.e., are larger than 3.9) can be treated as corresponding to 1 (100%)!

Answer to Problem 8.

Problem 9: Huh?

Hopefully, you have gotten a very, very small [69] probability for Problem 8. This implies that the numbers they got appear to be *highly* improbable. While we are at it, let's calculate one more probability. What is the probability that the number of students that (on average) finish successfully each course is more than 610?

[69] Even zero!

Answer to Problem 9.

Problem 10: Rejecting a hypothesis

Considering the small probability you got for the average to be as high as 610 or more, what can you deduce for the success rate of 7.5%? Should they believe it? Do *you* think it is valid? Or is the true success rate higher/lower? [70]

[70] Think: if the success rate is higher, and the mean of the normal distribution (for the average of 15 classes) is also higher, then does that mean the probabilities in Problem 8 and 9 go up or down?

Answer to Problem 10.

Congratulations, you just performed a fully-fledged data analysis experiment! We will focus a lot on this part after our second midterm.

NORMAL CUMULATIVE DISTRIBUTION FUNCTION

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

This is an in-class activity.

1. I will not ask you to work in groups (until we finish it).
2. Instead, we will read through and solve it together.
   - Feel free to ask **any questions** and participate with me!
3. You do not need to submit this. This is only for your better understanding of the **Central Limit Theorem**.

Part 1: Setting up the problem

A streaming service wants to see how many "units of entertainment" (episodes, movies, documentaries, etc.) a person consumes at one sitting. For convenience, the service assumes that a person starting something implies they have consumed it: hence, a person starting 3 episodes and 1 movie will have consumed 4 units during their session.

Problem 1: Distribution for a single person

A single person is assumed to watch a number of "units" of entertainment that follows a Poisson distribution. What is the expectation and variance, given rate λ? [71]

[71] Assume that λ is given in number of "units" per "sitting"/session.

Answer to Problem 1.

Problem 2: Probabilities

The service assumes that a specific type of person watches "units" of entertainment with a rate of 2.7 units per streaming session. What is the probability this type of person watches more than or equal to 2 items during their next streaming venture?

Answer to Problem 2.

Part 2: Setting up the central limit theorem

The service is having second thoughts about that type of person and they believe that their streaming routine has significantly changed after March 2020. They are now proposing that the person watches "units" of entertainment with a rate of $\lambda_2 = 3.1$ units per session, compared to $\lambda_1 = 2.7$ that was used earlier.

Problem 3: Collecting data

The streaming service has collected 20 sessions of that person since March 2020, and have found the following:

$$\{3, 2, 3, 2, 1, 4, 2, 4, 3, 3, 4, 3, 5, 5, 0, 4, 2, 3, 5, 3\}\,.$$

What is the average number of units consumed? [72]

[72] While we have not *officially* defined the average yet, use the traditional idea of summing the numbers and dividing by 20.

Answer to Problem 3.

Problem 4: Setting up the central limit theorem

Assume that the 20 sessions are a large enough sample for the central limit theorem to apply. What is the average number of units *distributed as*? What are the mean and the variance?

Answer to Problem 4.

Part 3: Analysis and aftermath

Recall that we have made the *hypothesis* that a person watches "units" of entertainment with rate λ_1 (before March 2020) or with rate λ_2 (after March 2020). Let's do both.

Problem 5: For λ_1

What is the probability a person following $\lambda_1 = 2.7$ units per session sees more than or equal to an average of 3.05 in 20 sessions?

Answer to Problem 5.

Problem 6: For λ_2

What is the probability a person following $\lambda_2 = 3.1$ units per session sees more than or equal to an average of 3.05 in 20 sessions?

Answer to Problem 5.

Problem 7: Aftermath

Can we reject? Are we *really sure* that the rate has changed from 2.7 to 3.1?

Answer to Problem 7.

Problem 8: Answering a few questions

I claim not really. The probabilities are not too small for either rate. So, what should we do to address this question once and for all?

Answer to Problem 8.

- It is easier to prove or disprove a change if the sample size is higher.

  a) True b) False

- It is easier to prove or disprove a change if the suspected change is bigger.

  a) True b) False

NORMAL CUMULATIVE DISTRIBUTION FUNCTION

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

11. Jointly distributed random variables

Every worksheet will work as follows.

1. You will be asked to form a group with other students in the class: you can make this as big or as small as you'd like, but groups of 4-5 work best.
2. Read through the worksheet, discussing any questions with the other members of your group.
   - You can call me at any time for help!
   - I will also be interrupting you for general guidance and announcements at random points during the class time.
3. Answer each question (preferably in the order provided) to the best of your knowledge.
4. While collaboration between students is highly encouraged and expected, each student has to submit their own version.
5. You will have 24 hours (see gradescope) to submit your work.

Before we get started, please remember (it is crucial to differentiate these two notions): we refer to X and Y (upper case, by convention) as two random variables, whereas x and y (lower case) are values they take.

Activity 1: Jointly distributed discrete random variables

In the next four questions, we focus on a pair of *discrete* random variables X and Y, distributed with joint probability mass function

$$f_{XY}(x,y) = c \cdot \frac{x+1}{y}.$$

At this point, it is useful to remind ourselves that the probability mass function values *are actual probabilities*, since we are discussing about discrete random variables.

Problem 1: Joint probability mass functions

Assume that $X = \{1,2,3\}$ and $Y = \{10,20\}$, that is X is allowed to take values 1, 2, or 3, and Y is allowed to be equal to either 10 or 20. What should c be in order for this to be a valid joint pmf? [73]

Answer to Problem 1.

[73] Remember the first axiom of probability mass functions for discrete random variables: summing over all possible values should give us 1. In math terms:

$$\sum_x \sum_y f_{XY}(x,y) = 1 \implies \ldots \implies c = \ldots$$

Problem 2: Table form

We may construct a table! Based on your answer in Problem 1, we may collect the different probability mass function values in tabular form. Fill the table below with the actual probabilities for each pair of values.

Answer to Problem 2.

		Y	
		10	20
X	1		
	2		
	3		

Verify once again that indeed the summation all of them is equal to 1, just to be sure.

Problem 3: Marginal probabilities

Where may we find the *marginal probabilities* for X or Y alone? In the table form we saw earlier, they are found by summing over a row or a column, depending on which one we are looking for! In this case, what is:

a) $P(X = 1)$?

b) $P(Y = 20)$?

Answer to Problem 3.

$P(X = 1) =$

$P(Y = 20) =$

Problem 4: Conditional probabilities

Where may we find the *conditional probabilities* for X given $Y = y$ or for Y given $X = x$? In the table form we saw earlier, they are found by focusing on one column or row, and then dividing the probability we are looking for over the summation of all elements in that column or row. [74] In our example, what is:

- $P(X = 1|Y = 20)$?
- $P(Y = 20|X = 1)$?

[74] Without a table, we would calculate a conditional probability by dividing appropriately. For example, the probability of getting $P(X = x|Y = y)$ could be found by $\frac{f_{XY}(x,y)}{f_Y(y)}$.

Answer to Problem 4.

$P(X = 1|Y = 20) =$

$P(Y = 20|X = 1) =$

Of course, this table form is valuable; but it is also limited to smaller sample spaces. What happens when we are dealing with a huge number of cases? In that case, we need to resort to the actual formulations for each and every one of our probability calculations. We'll see how that algebraic way of dealing with probabilities works in Activity 3 (in Page 6).

First, though, let's take a walk in the realm of continuous random variables in the next activity.

Activity 2: Jointly distributed continuous random variables

Let X and Y be two continuous random variables that are allowed to take any value between 0 and 1: that is, $0 \leq X \leq 1$ and $0 \leq Y \leq 1$. We further assume that they are jointly distributed with probability density function:

$$f_{XY}(x,y) = \frac{12}{11}\left(x^2 + y^2 + xy\right).$$

Problem 5: Calculating a probability

Recall that with continuous random variables, we need to integrate properly to calculate a probability. With that in mind, what is the probability that $0.3 \leq X \leq 0.7$ and $Y \geq 0.75$? [75]

Answer to Problem 5.

[75] You'll need to do a double integration. I'll get you started:

$$\int_{0.3}^{0.7}\int_{0.75}^{1} \ldots\ldots$$

Problem 6: Deriving a marginal distribution

Again, like you did earlier, derive the two marginal distributions. However, remember, that we are no longer summing! In the continuous space, we integrate. With that in mind, what is the marginal distribution of X? What is the marginal distribution of Y? [76]

Answer to Problem 6.

[76] In our case, because X and Y are only allowed to be between 0 and 1, we have (for Y, but it is very similar for X):

$$f_Y(y) = \int_0^1 f_{XY}(x,y)dx.$$

Problem 7: Deriving the conditional distribution

What is $P(X \leq 0.5|Y = 1)$? [77]

Answer to Problem 7.

Think about your approach. Could you have calculated the probability $P(X \leq 0.5|Y = y)$ for any value y?

Problem 8: Final touch

This is a little tougher. What is $P(X \leq Y)$? To help you get started, we have begun the solution approach, but do take a look at the hint for another explanation.. [78]

Answer to Problem 8.

From the law of total probability for continuous random variables, we have:

$$P(X \leq Y) = \int_0^1 P(X \leq Y|Y = y) f_Y(y) dy = \int_0^1 P(X \leq y) f_Y(y) dy = \dots$$

[77] For the conditional distribution, apart from the fact we are integrating instead of summing, we follow exactly the same logic as for discrete random variables. Remember to divide appropriately!

[78] Well, if we knew that $Y = y$, we could then calculate $P(X \leq Y|Y = y) = P(X \leq y)$, right? And, if we need to do that for any value that Y is allowed to take, can we use the **law of total probability for continuous random variables**? How does the law of total probability for continuous random variables look like again?

Activity 3: Discrete, but infinite

Consider two discrete random variables X and Y that take on integer values $X \geq 0$ and $1 \leq Y \leq 5$. That is, X could be 3, 107, or 0, and Y could be equal to 1, 2, 3, 4, or 5. Their joint probability mass function is given by:

$$f_{XY}(x,y) = e^{-y} \cdot \frac{y^x}{5 \cdot x!}.$$

As a side note, remember that

$$\boxed{\sum_{i=0}^{\infty} e^{-\alpha} \cdot \frac{\alpha^i}{i!} = 1 \text{ for any } \alpha > 0.}$$

Problem 9: Using the joint pmf

Let's start easy. What is the probability that both X and Y are equal to 1? That is, what is $P(X = 1 \cap Y = 1)$? [79]

Answer to Problem 9.

[79] This can also be written as $P(X = 1, Y = 1)$. Recall that (because this is a pmf for a discrete random variable) this can be found as simply the value of the pmf for the given X and Y.

Problem 10: Deriving a marginal distribution

What is the probability that $Y = 1$, regardless of what X is? That is, what is $P(Y = 1)$? [80]

Answer to Problem 10.

[80] Remember! To find the marginal distribution of one discrete random variable, sum over the other variable! In our case:

$$f_Y(y) = P(Y = y) = \sum_{x=0}^{\infty} f_{XY}(x,y).$$

After answering this, what is the probability that $Y = y$, for any value y? Is it always $\frac{1}{5} = 20\%$?

Problem 11: Deriving a conditional distribution

What is the probability that $X \geq 1$ given that $Y = 1$? That is, what is $P(X \geq 1 | Y = 1)$? [81]

Answer to Problem 11.

[81] You will probably get something similar to what we had seen earlier in the semester...

12. Jointly distributed random variables: extensions

Every worksheet will work as follows.

1. You will be asked to form a group with other students in the class: you can make this as big or as small as you'd like, but groups of 4-5 work best.
2. Read through the worksheet, discussing any questions with the other members of your group.
   - You can call me at any time for help!
   - I will also be interrupting you for general guidance and announcements at random points during the class time.
3. Answer each question (preferably in the order provided) to the best of your knowledge.
4. While collaboration between students is highly encouraged and expected, each student has to submit their own version.
5. You will have 24 hours (see gradescope) to submit your work.

Activity 1: Jointly distributed discrete random variables

During our last worksheet, we saw two discrete random variables $X = \{1, 2, 3\}$ and $Y = \{10, 20\}$ that were jointly distributed with probability mass function:

$$f_{XY}(x, y) = \frac{20}{27} \cdot \frac{x+1}{y}.$$

Additionally, we can calculate the marginal distribution of X and Y:

$$f_X(x) = \sum_y f_{XY}(x, y) = \frac{20}{27}\frac{x+1}{10} + \frac{20}{27}\frac{x+1}{20} = \frac{x+1}{9},$$
$$f_Y(y) = \sum_x f_{XY}(x, y) = \frac{20}{27}\frac{2}{y} + \frac{20}{27}\frac{3}{y} + \frac{20}{27}\frac{4}{y} = \frac{180}{27y} = \frac{60}{9y}.$$

Finally, recall that the conditional distribution for X given $Y = y$ can be calculated as:

$$f_{X|Y=y}(x) = \frac{f_{XY}(x, y)}{f_Y(y)} = \frac{\frac{20}{27} \cdot \frac{x+1}{y}}{\frac{60}{9y}} = \frac{x+1}{9}.$$

It seems that the conditional and the marginal distribution are the same – this is not always the case! This only happens when X and Y are independent. More on that later today.

Problem 1: Expectations and variances

What is the expectation of X and what is the variance of X? [82]

Answer to Problem 1.

[82] Recall that the expectation of one of two jointly distributed random variables can be found by properly summing (if discrete, as is the case here) or integrating (when continuous) its **marginal distribution**. The marginal distribution is given in Page 1.

Problem 2: Conditional expectations and variances

What is the expectation of X and what is the variance of X *given that* $Y = 10$? [83]

Answer to Problem 2.

[83] The previous hint still applies! However, we now replace the marginal with the **conditional distribution**. Again, the conditional distribution is given in Page 1.

Problem 3: Independent? Covariance? Correlation?

As hinted at when we calculated the marginal and conditional distributions, it appears that $f_X(x) = f_{X|Y=y}(x)$. Equivalently, we could make the observation that $f_{XY}(x,y) = f_X(x) \cdot f_Y(y)$. Hence, the two random variables X and Y are independent!

Based on your this observation of independence, what is the covariance? What is the correlation? [84]

[84] Recall that two independent random variables have **zero** covariance and, consequently, **no** correlation.

Answer to Problem 3.

$\sigma_{XY} = Cov[X,Y] =$

$\rho_{XY} = Corr[X,Y] =$

Activity 2: Jointly distributed continuous random variables

Consider two jointly distributed *continuous* random variables X, Y with domains $0 \leq X \leq 2$ and $0 \leq Y \leq 1$ and with joint probability density function equal to:

$$f_{XY}(x,y) = \frac{3}{4}x^3y^2.$$

Problem 4: Warm-up with marginal distributions

Let's repeat what we had done during our previous lecture. What are the marginal distributions of X and Y? [85]

[85] As a reminder, the marginal distribution of X will be a function of x and the marginal distribution of Y will be a function of y.

Answer to Problem 4.

$f_X(x) = \int_0^1 f_{XY}(x,y)dy =$

$f_Y(y) = \int_0^2 f_{XY}(x,y)dx =$

Problem 5: Getting the expectations

What are the expectations of X and Y? Don't forget that they are defined over *different domains*! [86]

[86] X is defined over $[0, 2]$, whereas Y is defined over the range $[0, 1]$.

Answer to Problem 5.

Problem 6: Independent?

Are X and Y independent? Why/Why not? [87]

[87] Check whether $f_{XY}(x, y) = f_X(x) \cdot f_Y(y)$: if so, then they are independent!

Answer to Problem 6.

Activity 3: When X and Y restrict each other

Assume that random variables X and Y are jointly distributed with probability density function $f_{XY}(x,y) = \frac{1}{4}(x+y)$ defined over $0 \leq X \leq Y \leq 2$. Note how random variable X always takes values that are at most as big as the value of random variable Y. [88] X and Y are not independent; this is clear from their definition, as knowing the one restricts the values the other one may take. What is the covariance of X and Y then? Well, to answer that we will need a lot of things. Namely, we need:

1. the marginal distributions $f_X(x), f_Y(y)$;
2. the expectation $E[X], E[Y]$;
3. the expectation of function $X \cdot Y$: $E[X \cdot Y]$;
4. finally, you'll get

$$Cov[X,Y] = E[X \cdot Y] - E[X] \cdot E[Y].$$

Let's get to it!

[88] If you are wondering how this is a valid pdf, we may show that the double integration is equal to 1. Be **very careful** with how you are integrating this. Following are the two correct ways (for an incorrect way, look at the notes!):

$$\int_0^2 \int_0^y \frac{1}{4}(x+y)\,dxdy = 1$$

$$\int_0^2 \int_x^2 \frac{1}{4}(x+y)\,dydx = 1$$

Problem 7: Marginal distributions

What are the marginal distributions of X and Y? Verify that X and Y are not independent by checking whether $f_X(x) \cdot f_Y(y)$ is different than $f_{XY}(x,y)$. Here, we help you get $f_X(x)$ started as a hint.

Answer to Problem 7.

$$f_X(x) = \int_x^2 f_{XY}(x,y)dy =$$

Problem 8: Expectation of X and Y

Using the marginal distributions from earlier, what is the expectation of X and what is the expectation of Y? [89]

Answer to Problem 8.

[89] Ok.. So expectations are typically found by integrating the marginal over the domain of the random variable. What is the domain here? Well, if Y is "gone", the domain for random variable X is $[0, 2]$, and if X is gone the domain of Y is also $[0, 2]$...

Problem 9: Expectation of $X \cdot Y$

And.. what is the expectation of $X \cdot Y$? [90]

Answer to Problem 9.

[90] Assume we have two random variables X and Y that are jointly distributed with joint pdf $f_{XY}(x, y)$. Then, for a function of those random variables, $g(X, Y)$, its expectation is $\int\int g(x, y) f_{XY}(x, y) dxdy$. But what about the integral limits? See Hint 7 in Page 5 :)

Problem 10: Covariance and correlation

Almost there. So, what is the covariance and what is the correlation of X and Y? For the correlation calculations as well as for any other remaining parts in this specific activity, use that $Var[X] = \frac{43}{180}$ and $Var[Y] = \frac{3}{20}$.

Answer to Problem 10.

Problem 11: Applying the variance identity

In the lecture notes, we discuss how the variance identity becomes [91]:

$$Var[aX + bY] = a^2 Var[X] + b^2 Var[Y] + 2abCov[X, Y].$$

[91] Recall that if X and Y are independent, then $Cov[X, Y] = 0$, and $Var[aX + bY] = a^2 Var[X] + b^2 Var[Y]$.

So, assume we are interested in the variance of random variable $Z = 3 \cdot X + 2 \cdot Y$: what can it be evaluated as?

Answer to Problem 11.

13. Jointly distributed random variables: some common distributions

Every worksheet will work as follows.

1. You will be asked to form a group with other students in the class: you can make this as big or as small as you'd like, but groups of 4-5 work best.
2. Read through the worksheet, discussing any questions with the other members of your group.
   - You can call me at any time for help!
   - I will also be interrupting you for general guidance and announcements at random points during the class time.
3. Answer each question (preferably in the order provided) to the best of your knowledge.
4. While collaboration between students is highly encouraged and expected, each student has to submit their own version.
5. You will have 24 hours (see gradescope) to submit your work.

Activity 1: The multinomial distribution

A product is manufactured in one of three factories located in Fargo, ND, Atlanta, GA, and Portland, OR. The product can be of high or low quality (assume these are the only two possibilities). The three factories have slightly different quality outputs, that are provided in Table 1.

Table 1: The high and low quality output rates of each factory. For example, Fargo, ND will have 95% of their products be of high quality, whereas Atlanta, GA will only have 92% of their products be of high quality.

	Fargo, ND	Atlanta, GA	Portland, OR
High quality	95%	92%	97.5%
Low quality	5%	8%	2.5%

Last, assume that Fargo, ND is responsible for 40% of the production in the United States, with Atlanta, GA and Portland, OR sharing the remaining production.

Problem 1: Law of total probability

Remember the law of total probability? Let's apply it here (in its discrete form). What is the probability a random product you pick up is of low quality?

Answer to Problem 1.

Problem 2: An application for the multinomial

You pick 10 products at random. What is the probability that exactly 4 of them were produced in Fargo, ND, exactly 3 of them were produced in Atlanta, GA, and the other 3 of them were produced in Portland, OR?

Answer to Problem 2.

Problem 3: Marginal and conditional distributions

You pick 10 products at random again. Answer the two probability questions here:

1. What is the probability that at most one of them was produced in Atlanta, GA? [92]
2. What is the probability 4 of them were produced in Fargo and the other 4 of them were produced in Portland given that 2 of them were manufactured in Atlanta? [93]

[92] Use the **marginal** distribution for this.

[93] Use the **conditional** distribution for this.

Answer to Problem 3.

In general:

1. The marginal distribution of a multinomial distribution is the binomial.
2. The conditional distribution of a multinomial distribution is another multinomial.

Activity 2: The bivariate normal distribution

We have asked a set of people about their views on climate change and associated sustainability policies proposed to curb its effects. Specifically, we want to measure the net favorability of each question (measured in %) defined as the % of respondents agreeing minus the % of respondents disagreeing. The survey in several different states has shown that the net favorability in the climate change question is $X \sim \mathcal{N}(4,4)$, while the net favorability in the policies question is $Y \sim \mathcal{N}(2,9)$. In English, we assume the net favorability in agreeing that climate change is a real threat is (continuous) random variable X normally distributed with mean $\mu_X = 4$ and variance $\sigma_X^2 = 4$; while the net favorability in agreeing with policies proposed is another normally distributed random variable called Y with mean $\mu_Y = 2$ and $\sigma_Y^2 = 9$.

Problem 4: Simple normal

Let's jog our memories with the normal distribution. Answer the two next probability questions. [94]

- What is the probability that there is more than a 6% net favorability difference for climate change? That is, what is $P(X > 6)$?

[94] A z-table is provided in the last page of the worksheet.

- What is the probability that there is less than or equal to a 1% net favorability difference adopting sustainability policies? That is, what is $P(Y \leq 1)$?

Answer to Problem 4.

$P(X > 6) =$

$P(Y \leq 1) =$

Problem 5: Independent?

Let's now look at the problem from a "data analyst" perspective. Do you think it is fair to assume that these two answers are independent? Or would they be correlated? And is that correlation positive or negative? [95] Simply a sentence of explaining why yes/no would suffice here.

[95] Think of the following setup: would a person agreeing that climate change is a threat also be more prone to agreeing with policies meant to curb it? Would the opposite be true?

Answer to Problem 5.

Hopefully you have agreed that the two are indeed correlated. From now on, please assume that we have a positive correlation equal to $\rho_{XY} = 0.5$.

Problem 6: The case of NC

We are analyzing the results from North Carolina. It appears that in NC respondents agree that climate changes is an important threat. Specifically, we observe that the net favorability of this question is $X = 3\%$. What is the probability that respondents disagree with policies to promote sustainability given that $X = 3\%$? That is, what is $P(Y \leq 0 | X = 3)$? [96] Recall that we have $\rho_{XY} = 0.5$.

[96] What is the **conditional** distribution of a bivariate normal distribution? Is it.. another normal distribution?

Answer to Problem 6.

Problem 7: Overwhelming support for one; so-and-so on the other

What is the probability of getting a state that both has a higher than 6% net favorability for the first question but is very close on the second question? Let's define "very close" as being between -1% and +1% of net favorability. In mathematical terms, what is $P\left((X > 6) \cap -1 \leq (Y \leq 1)\right)$? [97]

Answer to Problem 7.

[97] Also written as $P(X > 6, -1 \leq Y \leq 1)$ or $P(X > 6 \text{ and } -1 \leq Y \leq 1)$. Recall that to answer questions like these we need to integrate properly the pdf of the bivariate normal distribution. Thankfully there are online calculators!

Activity 3: The velocity of a particle

A particle's velocity (measured in m/s) in a gas is a continuous random variable (let it be V) with pdf $f_V(v) = av^2e^{-bv}, v > 0$, where b is a constant that depends on the temperature of the gas and the mass of the particle. While we could make the case that the speed of anything can be no larger than the speed of light, we do allow speed to go up to infinity here, as it makes the math a little easier :) so please consider that $v \in (0, +\infty)$.

Problem 8: Back to basics

Once again, what should a be in order for $f_V(v)$ to be a valid pdf? Note that your result will be a function fo b. [98]

Answer to Problem 8.

[98] As a reminder, we want $\int_{-\infty}^{+\infty} f(x)dx$ to be equal to 1. In our case, random variable velocity V cannot be negative, and we assume it goes to $+\infty$, so we would like $\int_{0}^{+\infty} f_V(v)dv = 1$.

Problem 9: The expectation of a function

Back to expectation properties now. The kinetic energy of a particle is $W = \frac{1}{2}mV^2$, where V is the velocity (the random variable from earlier). What is the expected kinetic energy of the particle? [99] Of course, as in Problem 1, your final answer will be a function of b.

Answer to Problem 9.

[99] Check Lecture 9: the expectation of a function of a *continuous* random variable X, $g(X)$, is calculated as $E[g(X)] = \int_{-\infty}^{+\infty} g(x)f(x)dx$, where $f(x)$ is the pdf of X itself. As a reminder, that would be slightly different in the discrete case (the integration becomes a summation).

Problem 10: The probability density function of a function

Assuming again that the kinetic energy of a particle is $W = \frac{1}{2}mV^2$, where V is the velocity random variable: what is the pdf of W?

Answer to Problem 10.

This is a very useful derivation. Let X be a random variable distributed with pdf $f_X(X)$. Also consider Y a random variable that is a function of X, as in $Y = g(X)$. Let $u(y)$ be the inverse function, i.e., $u(y) = g^{-1}(y)$. Then, Y is distributed with pdf:

$$\boxed{f_Y(y) = f_X(u(y)) \cdot |u'(y)|}$$

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

14. Descriptive statistics

Every worksheet will work as follows.

1. You will be asked to form a group with other students in the class: you can make this as big or as small as you'd like, but groups of 4-5 work best.
2. Read through the worksheet, discussing any questions with the other members of your group.
   - You can call me at any time for help!
   - I will also be interrupting you for general guidance and announcements at random points during the class time.
3. Answer each question (preferably in the order provided) to the best of your knowledge.
4. While collaboration between students is highly encouraged and expected, each student has to submit their own version.
5. You will have 24 hours (see gradescope) to submit your work.

Activity 1: Descriptive statistics

The average high September temperature in Chicago over the last 15 years (from 2009 to 2023) has been as in Table 2 (in Fahrenheit).

Table 2: The average high September temperatures over the last 15 years.

Year	Temperature
2023	78
2022	76
2021	80
2020	74
2019	75
2018	70
2017	70
2016	72
2015	71
2014	66
2013	69
2012	66
2011	63
2010	65
2009	66

Let's use this small dataset today to describe the information presented and visually showcase it. In the remaining exercises you will only use this small part of the data. Of course, in real life, we would be given a bigger volume of data. In such instances, we would resort to using software such as Excel or Pandas on Python.

Problem 1: Sample average and variance

What is the average temperature in the sample? What is the sample variance? [100]

Answer to Problem 1.

[100] Recall how the calculation of variance is different for a sample compared to a population...

Problem 2: Sample mode

What is the mode of the data? [101]

Answer to Problem 2.

[101] The mode is the most frequently observed data point.

Problem 3: Quartiles

What are the first, second, and third quartile? [102]

Answer to Problem 3.

[102] The second quartile is also called the median.

Problem 4: Range and interquartile range

What is the range and the interquartile range? Based on your answers: are there any outliers in the data?

Answer to Problem 4.

Activity 2: Graphical representations

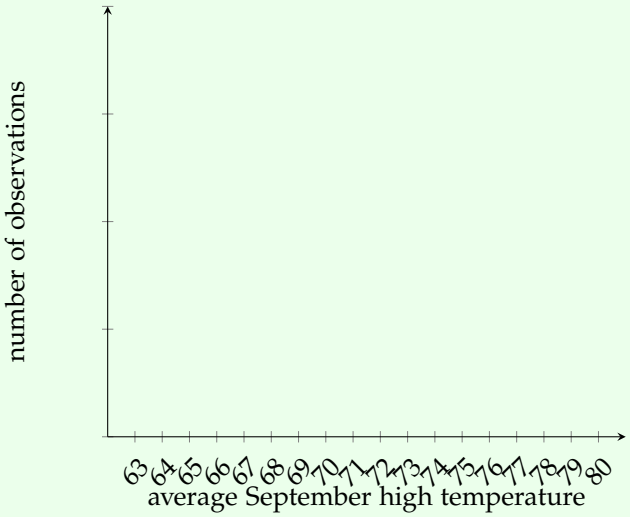
In this activity, we will present the data given in a series of graphical tools. We omit some of them due to time constraints (for a full description, please read the Lecture 14 notes).

Problem 5: Dot diagrams

Dot diagrams [103] (as the name suggests) asks to place a dot on top of each data point. For example, say there are 5 observations for a specific data point, we would put 5 dots one on top of the other! Create a dot diagram for the data of Table 2.

[103] See Page 8-9 in the notes!

Answer to Problem 5.



Note how easy it is to find the mode now! Simply look for the tallest set of dots.

Problem 6: Stem-and-leaf diagrams

A stem-and-leaf diagram [104] only makes sense when all of the data consists of at least two digits. How to construct one? We need to separate a numerical observation into a stem (the first, more important digits) and leaves (the least important digit). For example, the temperature of 32 Fahrenheit can be decomposed into a stem of "3" and a leaf of "2", or the number 538 can be decomposed into a stem of "53" and a leaf of "8".

Just for this example, consider the average low temperature in two different states:

[104] See Page 9 in the notes!

State	J	F	M	A	M	J	J	A	S	O	N	D
State 1	15	19	33	41	51	61	65	63	54	43	33	20
State 2	36	40	47	53	62	68	71	71	65	60	51	40

Create a stem-and-leaf diagrams using the data provided for State 2. Observe my stem-and-leaf for State 1 as an example!

Answer to Problem 6.

State 1

6	1 3 5
5	1 4
4	1 3
3	3 3
2	0
1	5 9

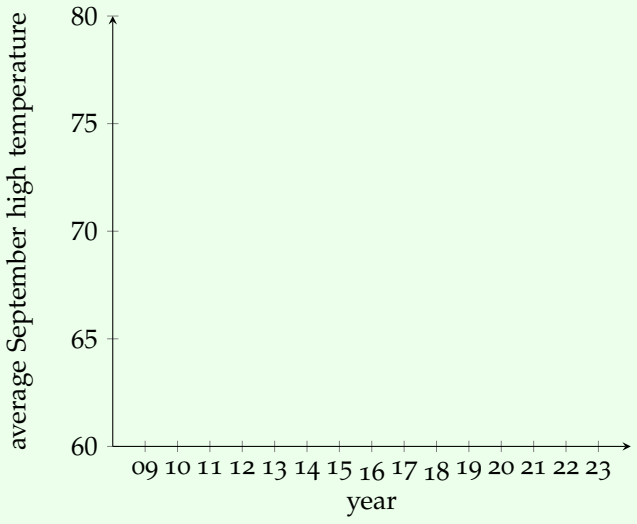
It is a *striking* visual tool to showcase frequency.

Problem 7: Time series plots

A time series plot [105] is especially useful when the data are recorded in the order of time. For example, in our original dataset of Table 2, all temperatures are given in order of time from 2006 to 2020. Create a time series plot by adding each observation (y coordinate) in the appropriate time (x coordinate), and then connect the observations using straight lines.

[105] See Pages 11-12 in the notes.

Answer to Problem 7.



Activity 3: Box plots

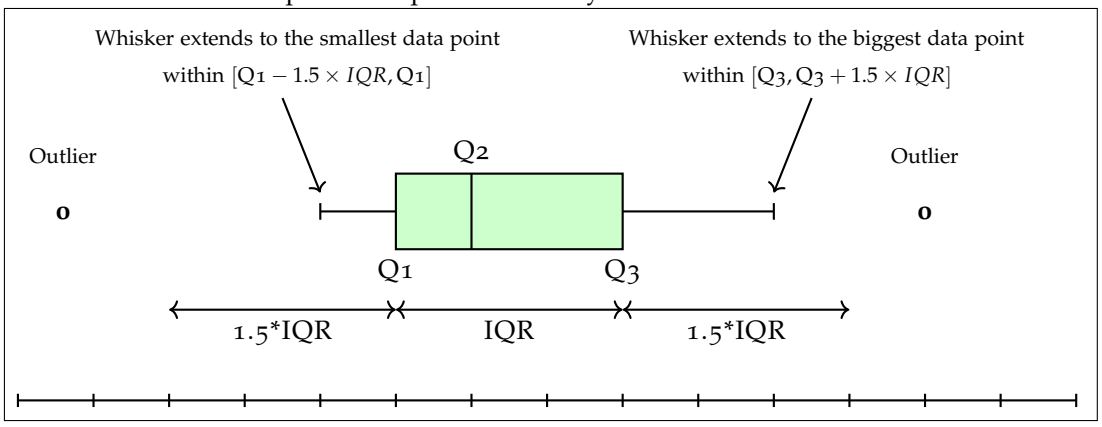
Box plots [106], sometimes also called box-and-whisker plots, are graphical devices built to reveal multiple interesting properties at once. Seeing a box plot reveals the center, the spread, the shape, and the outliers in our data!

To build one, follow the next few steps:

1. Identify Q_1, Q_2, Q_3 and the IQR. Create a small box with three lines: the left most line is at Q_1, the middle one is at Q_2, and the right most one is at Q_3.
2. Calculate $Q_1 - 1.5 \cdot IQR$ and $Q_3 + 1.5 \cdot IQR$. Points that are outside these limits are *outliers*. Mark every outlier with a "o".
3. Extend the one whisker all the way to the smallest observation within $[Q_1 - 1.5 \cdot IQR, Q_1]$ and the other whisker all the way to the largest observation within $[Q_3, Q_3 + 1.5 \cdot IQR]$.

[106] See Pages 12-13 in the notes!

For convenience, we present here the same figure from the notes, where the details of a box plot are explained visually.



Problem 8: Designing a box plot

Design a box plot based on the data of Table 2.

Answer to Problem 8.

Problem 9: Missing whiskers?

The compressive strength of concrete is the subject of a test by civil engineers. Nine different specimens were tested and the civil engineers obtained (in psi): 2210, 2230, 2200, 2240, 2250, 2240, 2330, 2250, 2210. Describe the data as a **box plot**.

Answer to Problem 9.

Note then how when there are no points in $[Q_1 - 1.5IQR, Q1]$ or $[Q_3, Q_3 + 1.5IQR]$ then the whisker simply disappears!

Activity 4: Histograms

A histogram [107] is a graphical construct that presents data by placing them in *bins*. Histograms possess three important characteristics:

1. modality.
2. heavy/light tailedness.
3. skewness.

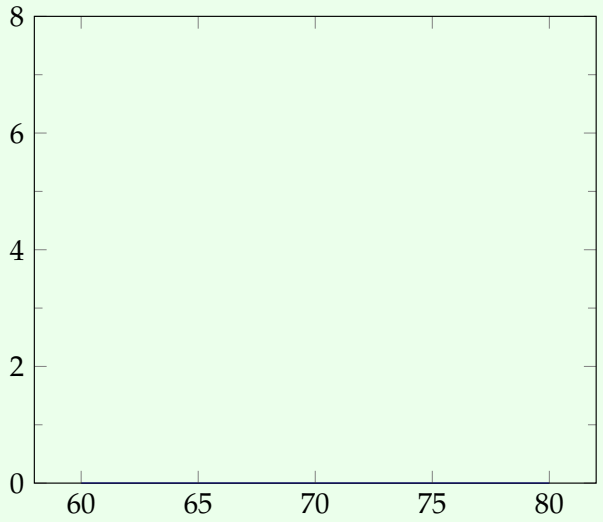
We will investigate all three of them in the next few activities.

[107] See Pages 13-19 in the notes!

Problem 10: Designing a histogram

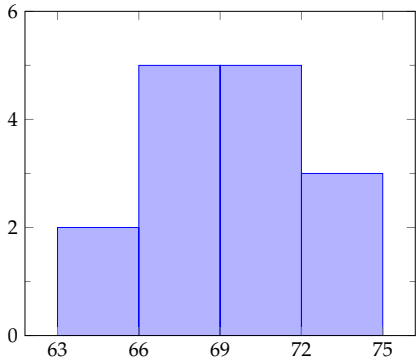
First of all, we may create different histograms depending on the number of bins and the type of bins we create. Assume we create 4 bins corresponding to temperatures $[60,65)$,$[65,70)$,$[70,75)$,$[75,80)$. Design this histogram.

Answer to Problem 10.



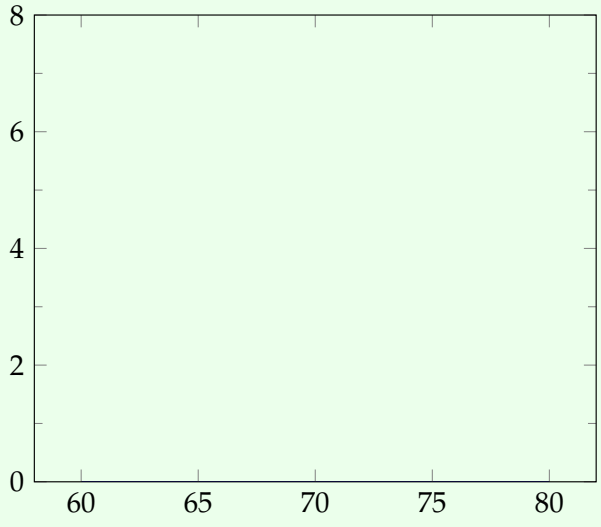
Problem 11: Different histograms

What if we change the number of bins? Assume we want to generate n bins. What we could do, is find the range of values and divide this by n: call this (fractional, probably) number q. Then, create bins as follows: $[min, min + q), [min + q, min + 2q), \ldots, [max - q, max]$. For example, if we created $n = 4$ bins with our data, we would have $q = 12/4 = 3$ and the following bins: $[63, 66), [66, 69), [69, 72), [72, 75]$. The corresponding histogram would be:



Draw a histogram with $n = 6$ bins.

Answer to Problem 11.



Problem 12: Histogram details

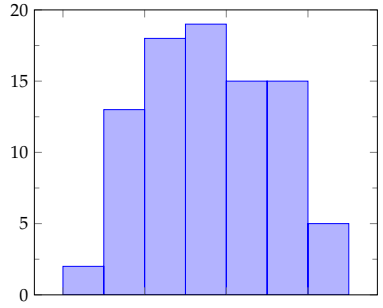
What do you observe about your two histograms (in Problems 10 and 11)? What are their modalities (i.e., number of discernible peaks)? Are they heavy-tailed? Are they left or right skewed? [108] After you answer these for your histograms, then match the following histograms (a, b, c, and d) with the histogram below.

[108] For example, the histogram with $n = 4$ bins is **unimodal** (one discernible peaks, one around 66–72), heavy-tailed, and it appears to be right-skewed (i.e., with a tail to the right).

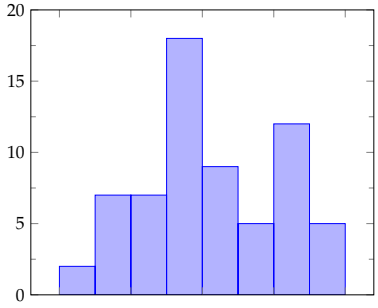
Which of the following four histograms:

1. are unimodal?
2. are bimodal?
3. are uniform?
4. are right-skewed?
5. are left-skewed?
6. are symmetric?

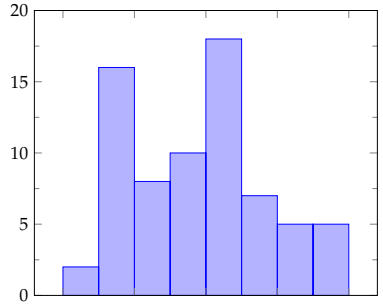
(a)



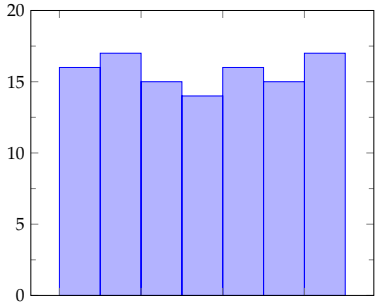
(b)



(c)



(d)



Answer to Problem 12.

- **unimodal**:
- **bimodal**:
- **uniform**:
- **right-skewed**:
- **left-skewed**:
- **symmetric**:

15. Point estimators

Every worksheet will work as follows.

1. You will be asked to form a group with other students in the class: you can make this as big or as small as you'd like, but groups of 4-5 work best.
2. Read through the worksheet, discussing any questions with the other members of your group.
   - You can call me at any time for help!
   - I will also be interrupting you for general guidance and announcements at random points during the class time.
3. Answer each question (preferably in the order provided) to the best of your knowledge.
4. While collaboration between students is highly encouraged and expected, each student has to submit their own version.
5. You will have 24 hours (see gradescope) to submit your work.

Activity 1: Anchoring activity

Let us begin with a small game activity. Consider a (discrete) **uniformly distributed population** X that can take any integer value between 1 and α, where $\alpha > 1$ is sadly unknown. To estimate it, we will try three point estimators. Given a sample of $n = 3$ numbers (call them X_1, X_2, X_3) from the population, we will estimate α as:

a) $\hat{\Theta}_1 = \frac{X_1+X_2+X_3}{3}$.

b) $\hat{\Theta}_2 = \frac{X_1+2X_2+X_3}{3}$.

c) $\hat{\Theta}_3 = \frac{2}{3}(X_1 + X_2 + X_3) - 1$.

Problem 1: Using estimators

First, navigate yourself to the website

`http://vogiatzis.web.illinois.edu/random.html`

and generate three numbers. Call them X_1, X_2, X_3. What is the point estimate [109] you'd get for α with each of the three estimators?

[109] Recall that we call point estimate the actual value you get by using a point estimator.

Answer to Problem 1.

$X_1 =$, $X_2 =$, $X_3 =$

a) $\hat{\theta}_1 =$

b) $\hat{\theta}_2 =$

c) $\hat{\theta}_3 =$

Problem 2: Estimator bias

If we are trying to estimate α, recall that the bias of some estimator $\hat{\Theta}$ can be calculated as

$$bias\left[\hat{\Theta}\right] = E\left[\hat{\Theta}\right] - \alpha.$$

Moreover, recall that because X_1, X_2, X_3 were obtained from the same population X we know that the following are true:

$$E[X_1] = E[X_2] = E[X_3] = E[X] = \mu$$
$$Var[X_1] = Var[X_2] = Var[X] = \sigma^2.$$

You do not need the variance in this question, but you may need it later today!

Finally, remember that the mean μ depends on the distribution we have, so in our case it would be $\mu = \frac{1+\alpha}{2}$ (seeing as the numbers obtained come from a discrete uniform distribution between 1 and α).

Putting all of the above together, what is the bias of each of the three estimators? Your bias could possibly be a function of α!

Answer to Problem 2.

a) $bias\left[\hat{\Theta}_1\right] =$

b) $bias\left[\hat{\Theta}_2\right] =$

c) $bias\left[\hat{\Theta}_3\right] =$

You must have gotten that the bias of estimator $\hat{\Theta}_3$ is equal to zero! This is great news. Let us check what its variance is.

Problem 3: Estimator variance

What is the variance of $\hat{\Theta}_3$? [110] Recall that the variance too can be a function of the unknown parameter α. The variance of a uniformly distributed discrete random variable between a and b is $\frac{(b-a+1)^2-1}{12}$. In our case, our population X is between 1 and α so the formula gives us a variance of

$$Var[X] = \frac{\alpha^2 - 1}{12}.$$

[110] You **will** need this variance property. If X and Y are independent random variables, then:

$$Var[aX + bY + c] = a^2 Var[X] + b^2 Var[Y].$$

Answer to Problem 3.

$Var\left[\hat{\Theta}_3\right] =$

Problem 4: Estimator variance for different sample sizes

How would the variance change if we picked a sample of $n = 5$ observations $(X_1, X_2, X_3, X_4, X_5)$ and then calculated the estimator as $\hat{\Theta} = \frac{2}{5}(X_1 + X_2 + X_3 + X_4 + X_5) - 1$? Would it increase, decrease, or stay the same? What happens as n increases more and more?

Answer to Problem 4.

$Var\left[\hat{\Theta}\right] =$

This brings us to our first realization. With an unbiased estimator, larger samples will lead to smaller variances!

Let's make it interesting. Use this estimator to produce some values from the website and note here your best estimate for what α is. What is the biggest number you can possibly obtain from this website? The biggest number that can be produced from the website is...

My best estimate is...

Activity 2: Weird point estimators

Assume that a population is distributed with pdf $f(x) = c(1 + \theta x)$, $-1 \leq x \leq 1$, where θ is an unknown parameter, and c a constant. [111]

[111] That means, in English, that c has to be one value and one value alone, whereas θ can be *anything*.

Problem 5: Back to basics

Let's return to the basics for a second! What should c be equal to in order for $f(x)$ to be a valid continuous pdf? [112]

[112] Integrate $f(x)$ over its domain of all values of x allowed and equate to 1.

Answer to Problem 5.

If we are doing things right, θ should disappear after taking the integral. This implies that θ can take any from a series of values, hence being a **parameter** rather than a **constant**. A parameter can be any thing; a constant has to take on a specific value.

Problem 6: Where did you come up with this?

Assume you obtain a sample of n observations. Consider the sample average $\overline{X} = (X_1 + X_2 + \ldots + X_n)/n$. Show that $\hat{\Theta} = 3\overline{X}$ is an **unbiased estimator** for θ. [113]

[113] To do so first you need to calculate $E\left[\hat{\Theta}\right] = E\left[3\overline{X}\right] = 3E\left[\overline{X}\right]$. But, isn't $E\left[\overline{X}\right] = E\left[\frac{\sum X_i}{n}\right] = E[X]$? We could calculate this as $E[X] = \int_{-1}^{+1} xf(x)dx$, which may be a function of θ...

Answer to Problem 6.

Problem 7: Variance and standard error

What is the standard error of the point estimator $\hat{\Theta} = 3\overline{X}$? [114] The standard error can be calculated as

$$SE\left[\hat{\Theta}\right] = \sqrt{Var\left[\hat{\Theta}\right]}.$$

[114] To calculate this you will first need to calculate the expectation and the variance of population X. They could very well be a function of θ as you do not know what the parameter is equal to...

Answer to Problem 7.

This is quite common: when dealing with an unknown parameter (in our case, θ), the bias and the variance can depend on the value that the parameter actually has. So, how do we compare estimators for unknown parameters?

Activity 3: Comparing point estimators

Assume we have collected a sample of $n = 3$ observations X_1, X_2, X_3 coming from a population X distributed with *some pdf* with unknown μ and known $\sigma^2 = 16$. We have devised three point estimators for the unknown population mean:

- Get the average from the first two observations omitting the third, i.e.,
$$\hat{\Theta}_1 = \frac{X_1 + X_2}{2}.$$
- Add the "odd" observations once and the "even" observations doubled and divide everything by 4, i.e.,
$$\hat{\Theta}_2 = \frac{X_1 + 2X_2 + X_3}{4}.$$
- Once again omit the third observation and simply add the first two and divide by 4, i.e.,
$$\hat{\Theta}_3 = \frac{X_1 + X_2}{4}.$$

Now, for one last time in this worksheet, we go ahead and calculate the bias and variance of each of the estimators.

- We have for the first estimator, $\hat{\boldsymbol{\Theta}}_{\mathbf{1}}$:
$$bias\left[\hat{\Theta}_1\right] = E\left[\hat{\Theta}_1\right] - \mu = E\left[\frac{X_1 + X_2}{2}\right] - \mu = \frac{1}{2}E[X_1] + \frac{1}{2}E[X_2] - \mu =$$
$$= \frac{1}{2}E[X] + \frac{1}{2}E[X] - \mu = E[X] - \mu = \mu - \mu = 0.$$
$$Var\left[\hat{\Theta}_1\right] = Var\left[\frac{X_1 + X_2}{2}\right] = \frac{1}{4}Var[X_1 + X_2] = Var[X_1] + \frac{1}{4}Var[X_2] =$$
$$= \frac{1}{4}Var[X] + \frac{1}{4}Var[X] = \frac{1}{2}Var[X] = \frac{1}{2}\sigma^2 = 8.$$
- Similarly, for the second estimator, $\hat{\boldsymbol{\Theta}}_{\mathbf{2}}$, we would end up with: [115]
$$bias\left[\hat{\Theta}_2\right] = 0.$$
$$Var\left[\hat{\Theta}_2\right] = 6.$$
- Finally, for the third estimator, $\hat{\boldsymbol{\Theta}}_{\mathbf{3}}$:
$$bias\left[\hat{\Theta}_3\right] = -\frac{\mu}{2}.$$
$$Var\left[\hat{\Theta}_3\right] = 2.$$

[115] These calculations are left as an exercise.

Or, in tabular form:

	$\hat{\Theta}_1$	$\hat{\Theta}_2$	$\hat{\Theta}_3$
bias	0	0	$-\mu/2$
variance	8	6	2

Problem 8: Comparison I

What is the MSE of each of the estimators? Which estimator is the best according to its MSE, if we have been told that $\mu > 4$? [116]

Answer to Problem 8.

Note how the result would have changed if $\mu < 4$. On the other hand, there is no way that the first estimator has the smallest variance among the three.

Problem 9: Observation

Earlier in this activity, we observe that the first two estimators are unbiased (i.e., zero bias). In general, assume you are collecting a sample of n observations $(X_1, X_2, \ldots, X_n)$ and are using $\hat{\Theta} = a_1 X_1 + a_2 X_2 + \ldots + a_n X_n$ to estimate the unknown mean. What condition should $a_1 + a_2 + \ldots + a_n$ satisfy in order for $\hat{\Theta}$ to have bias equal to zero? [117]

Answer to Problem 9.

[116] As a reminder:

$$MSE = bias^2 + variance.$$

[117] Hmmm.. What can you tell about $E\left[\sum_{i=1}^{n} a_i X_i\right]$? Additionally, never forget that $E[X_1] = E[X_2] = \ldots = E[X_n] = E[X]$ because all observations come from the same population X!

16. Point estimators

We will solve this in-class activity before breaking out into groups (as usual) and working on our worksheets. Just as a reminder: the Lecture 15 worksheet is due **tomorrow** by noon on gradescope (instead of the original deadline).

Activity 1: Uniform distribution

Last time, we worked with estimating the upper bound of a discrete uniformly distributed population X; today, we begin with the continuous version! Assume we are trying to estimate the upper bound of a continuous uniformly distributed population X. In English: if we have a random number generator producing (real) numbers between 0 and α, how can we estimate what α is after collecting some data from the generator?

Problem 1: Intuition

For example, say we have obtained $X_1 = 7.7, X_2 = 15.21, X_3 = 9.1$, then what is a "good" estimate for the true value of α? Is saying that all numbers that are produced by this generator are between 0 and 10 true? How about between 0 and 20?

Answer to Problem 1.

A good estimate we can come up with for the upper bound of the generator is that it is ...

Problem 2: Bias for $n = 3$

Assume that we use the following estimator:

$$\hat{\Theta} = \max\{X_1, X_2, \ldots, X_n\},$$

where n is the number of data points we have collected. For example, if $X_1 = 7.7, X_2 = 15.21, X_3 = 9.1$ are the data obtained then $n = 3$, and $\hat{\theta}$ (the estimate obtained) is 15.21. **What is the bias of the estimator?** Assume that we have already obtained [118] that

$$E\left[\hat{\Theta}\right] = \frac{n}{n+1} \cdot \alpha.$$

[118] Ask me why, if you are interested!

Answer to Problem 2.

$bias\left[\hat{\Theta}\right] =$

Problem 3: Bias for $n \to \infty$

What happens to the bias as we increase the number of observations from the generator? Specifically, what is $\lim_{n\to\infty} bias\left[\hat{\Theta}\right]$?

Answer to Problem 3.

We see that the bias goes to 0 as we obtain more and more observations. This is great news! We define **a consistent estimator** (sometimes also called an *asymptotically* consistent estimator) as an estimator that has bias converging to 0 as the number of data points used in its calculation increases. In practice, this means that the estimates obtained get closer and closer to the true value of the parameter we are estimating as we use more and more data.

Problem 4: Variance for $n = 3$

Let's go back to our estimator $\hat{\Theta} = \max\{X_1, X_2, \ldots, X_n\}$ for $n = 3$ (i.e., only 3 data points from the generator). What is the variance of the estimator? Assume we have already calculated that [119]

$$E\left[\hat{\Theta}^2\right] = \frac{n}{n+2} \cdot \alpha^2.$$

[119] Again: ask me for the derivation if you are interested!

Answer to Problem 4.

$Var\left[\hat{\Theta}\right] =$

Problem 5: The mean square error

What is the MSE for the estimator $\hat{\Theta}$ for $n = 3$? How about for $n = 10$? When $n \to \infty$?

Answer to Problem 5.

For $n = 3$: $MSE = bias^2\left[\hat{\Theta}\right] + Var\left[\hat{\Theta}\right] =$

For $n = 10$:

$$MSE =$$

As $n \to \infty$:

$$\lim_{n\to\infty} MSE =$$

Activity 2: Comparing point estimators

Assume that a population is distributed with pdf $f(x) = \theta\left(x - \frac{1}{2}\right) + 1$, $0 \leq x \leq 1$, where θ is an unknown parameter. We have also been able to collect a sample of n observations and have calculated their sample average as $\overline{X}$. We have been trying estimators for θ and we want to compare three of them and pick the best:

1. $\hat{\Theta}_1 = 2\overline{X} - 1$
2. $\hat{\Theta}_2 = 12\overline{X} - 6$

Problem 6: MSE

Which one of the two has the smallest mean square error? [120]

Answer to Problem 6.

[120] As a reminder, for an estimator $\hat{\Theta}$, the mean square error is:

$$MSE = bias\left[\Theta\right]^2 + Var\left[\hat{\Theta}\right].$$

Problem 7: Application

For the previous population from Problem 6, we have collected a sample of 6 items and found them equal to $X_1 = 0.8, X_2 = 0.83, X_3 = 0.95, X_4 = 0.72, X_5 = 0.85, X_6 = 0.65$. What is a good estimate for θ? Use the point estimator that provided the best MSE from Problem 8.

Answer to Problem 7.

Activity 3: Point estimators for the exponential distribution

Assume that a population X is exponentially distributed; however, we have no idea what λ is. You recalled one thing though from IE 300: **the expected time between events is equal to** $\frac{1}{\lambda}$. This gives you an idea. You will wait to observe the times between some events and use them to estimate λ. More specifically, you will estimate λ as 1 over the average time between events.

For example, if the sampled times between two consecutive events are equal to 3 minutes, 2 minutes, 5 minutes, 4 minutes, 2 minutes, 2 minutes, you will estimate λ as 1 over 3 minutes!

Problem 8: Biased or not?

Is this estimate for λ biased or not? In mathematical terms, is $\hat{\lambda} = \frac{1}{\overline{X}}$, where $\overline{X}$ is the average of n observations, biased? [121]

[121] Can you make the claim that $E\left[\frac{1}{X}\right] = \frac{1}{E[X]}$? Yes or no?

Answer to Problem 8.

Problem 9: MSE

What is the mean square error for the estimator $\hat{\lambda} = \frac{1}{\overline{X}}$ for the unknown rate of an exponential distribution after observing n data points?

Answer to Problem 9.

17. Methods of estimation: the method of moments

Every worksheet will work as follows.

1. You will be asked to form a group with other students in the class: you can make this as big or as small as you'd like, but groups of 4-5 work best.
2. Read through the worksheet, discussing any questions with the other members of your group.
   - You can call me at any time for help!
   - I will also be interrupting you for general guidance and announcements at random points during the class time.
3. Answer each question (preferably in the order provided) to the best of your knowledge.
4. While collaboration between students is highly encouraged and expected, each student has to submit their own version.
5. You will have 24 hours (see gradescope) to submit your work.

Activity 1: Streaming services and their data

A TV streaming service has collected a lot of information from several customers and is interested in analyzing it for patterns about their streaming habits. More specifically, they have collected data on two items: how much time they have spent using the service (from login to closing the tab/exiting the app) and how many episodes they have watched. Part of the data follows.

	Session #									
	1	2	3	4	5	6	7	8	9	10
Time (in hours)	0.75	1.20	1.33	0.97	0.80	1.43	0.87	1.41	1.09	1.05
# of episodes	1	3	3	3	2	5	1	5	3	1

You will use the time data (first row, provided in hours) in Problems 1 and 2; you will also need the count data (second row, provided in number of episodes) in Problem 3.

Let us help them find good estimators using the method of moments! As a reminder, for the method of moments, we:

1. Calculate (generally, there are exceptions [122]) as many population and sample moments as the unknown parameters.
2. Equate each population moment with the corresponding sample moment. Typically the population moment(s) will be a function of the unknown parameter(s), whereas the sample moment(s) will be a value based on the observations.
3. Solve a system of equations for the unknown parameters.

[122] For example, a population moment may not exist.

Problem 1: The method of moments for a uniform distribution

The streaming service assumes that people spend time that is **uniformly distributed** in $[\theta, 2\theta]$ using their service. Use the method of moments to provide an estimator for θ, $\hat{\Theta}$. What is the point estimate $\hat{\theta}$ you get when using the data of the previous table? [1][2][3]

You'll only need the first row of the previous table. Here it is again for convenience:

	Session #									
	1	2	3	4	5	6	7	8	9	10
Time (in hours)	0.75	1.20	1.33	0.97	0.80	1.43	0.87	1.41	1.09	1.05

[1][2][3] Recall that for a population X that follows a uniform distribution (continuous) in $[a, b]$, the first moment (i.e., the expectation) is easily found as

$$E[X] = \frac{a+b}{2}.$$

Answer to Problem 1.

Problem 2: The general uniform distribution

What if the streaming time is not distributed in $[\theta, 2\theta]$ but is instead uniformly distributed in $[a, b]$? [124] How would we go about estimating a and b? What are the point estimates $\hat{a}$ and $\hat{b}$ when you use the data from the earlier table? [125]

Again, you will simply need the first row from the earlier table:

	Session #									
	1	2	3	4	5	6	7	8	9	10
Time (in hours)	0.75	1.20	1.33	0.97	0.80	1.43	0.87	1.41	1.09	1.05

You *may* also need that $Var[X] = \frac{(b-a)^2}{12}$ for a uniformly distributed random variable X.

Answer to Problem 2.

[124] You will notice how $E[X] = \overline{X}$ is not enough! Hence, we will need a second equation to solve that system..

[125] Additionally, you'll need $E\left[X^2\right]$. Recall that

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

Problem 3: Number of episodes as a Poisson distributed random variable

Let us turn our focus to the number of episodes users watch. We assume the number of episodes watched during a session is a Poisson distributed random variable with rate λ; alas, λ is not known.

Using the method of moments, provide an estimator for λ. [126] Use that estimator on the data from the table to obtain a point estimate $\hat{\lambda}$. As a reminder, here are the data on the number of episodes per session:

[126] What is $E[X]$ for a Poisson distributed random variable? What is $\overline{X}$ for the data given? Equating should provide the answer for the estimate.

	Session #									
	1	2	3	4	5	6	7	8	9	10
# of episodes	1	3	3	3	2	5	1	5	3	1

Answer to Problem 3.

Activity 2: Coming up with an estimator

The outcome of an experiment is a number between 0 and 1 (where 0 is considered an utter failure and 1 is considered a big success). The outcome is distributed with pdf $f(x) = \theta\left(x - \frac{1}{2}\right) + 1$, defined over $0 \leq x \leq 1$, where θ is an unknown parameter.

Problem 4: Applying the method of moments

Using the method of moments, obtain an estimator for θ. You may assume that you have been given a sample $X_1, X_2, \ldots, X_n$ and have obtained the sample average (sample first moment) as $\overline{X}$.

Answer to Problem 4.

Hopefully, you have gotten that $\hat{\Theta} = 12\overline{X} - 6$, where $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the average of n observations.
If you did not get that result, go back, find the mistake, and correct it. Did you take the first population and the first sample moments and equate them? Did you remember to equate the first sample moment (the sample average $\overline{X}$) to the first population moment (expected value $E[X] = \int_0^1 xf(x)dx = \frac{\theta+6}{12}$)?

Problem 5: MSE

What is the MSE of your estimator? [127] If you get stuck remember you will need that $E\left[\overline{X}\right] = E\left[X\right]$ and $Var\left[\overline{X}\right] = \frac{1}{n} Var\left[X\right]$. [128] Finally, keep in mind that both the bias and the variance may end up being a function of the unknown parameter or of the sample size.

Answer to Problem 5.

[127] For an estimator $\hat{\Theta}$, the mean square error is:

$$MSE = bias\left[\Theta\right]^2 + Var\left[\hat{\Theta}\right].$$

[128] Additionally, remember that

$$Var\left[X\right] = E\left[X^2\right] - \left(E\left[X\right]\right)^2 = \\ = \int_0^1 x^2 f(x) dx - \left(E\left[X\right]\right)^2.$$

Problem 6: Application

For this experiment, we have collected a sample of 6 items and found them equal to $X_1 = 0.8, X_2 = 0.83, X_3 = 0.95, X_4 = 0.72, X_5 = 0.85, X_6 = 0.65$. What is a good point estimate for θ? Use the point estimator you came up with in Problem 5. Using that estimate, what is the probability the experiment is at least a moderate success? Assume that moderate success is any value above 0.75.

Answer to Problem 6.

Activity 3: The Pareto distribution

Problem 7: A 2 × 2 system

The Pareto distribution (named after Italian mathematician and engineer, Vlifredo Pareto) is described by $f(x) = \frac{\alpha\beta^\alpha}{x^{\alpha+1}}$ for $x \geq \beta$. Originally it was designed to apply to the distribution of wealth where a small portion of the population accounts for a large part of the wealth; but it has been successfully applied in other settings, too. In this exercise, you are asked to estimate α and β using the method of moments. You may assume that $\alpha > 2$ and $\beta \geq 1$.

No need to actually solve the system of equations! Simply set it up as if you had a sample of n observations $X_1, X_2, \ldots, X_n$ and leave it as a system of equations.

Answer to Problem 7.

Activity 4: Special case

In this last activity, we have you run into a weird case. What if the first moment does not work? [129] Then, we need to take the second moments and equate them. Here is such a case.

[129] Maybe the first moments do not exist, or are not a function of the parameter.

Problem 8: A weird situation

Let X be a continuous population distributed with probability density function $f(x) = \frac{1}{2\beta} e^{-\frac{|x|}{\beta}}, x \in (-\infty, +\infty)$. Using the method of moments provide an estimator for β.

Answer to Problem 8.

So, in this case, we observe that we do not always take the first k moments when there are k parameters that we are estimating. Instead, we need to only consider moments that are a function of the parameters. In the above case, the first moment of X ended up being a constant (not a function of β) so we had to take the second moments and equate them.

18. Methods of estimation: maximum likelihood estimation

Every worksheet will work as follows.

1. You will be asked to form a group with other students in the class: you can make this as big or as small as you'd like, but groups of 4-5 work best.
2. Read through the worksheet, discussing any questions with the other members of your group.
   - You can call me at any time for help!
   - I will also be interrupting you for general guidance and announcements at random points during the class time.
3. Answer each question (preferably in the order provided) to the best of your knowledge.
4. While collaboration between students is highly encouraged and expected, each student has to submit their own version.
5. You will have 24 hours (see gradescope) to submit your work.

Activity 1: Our first MLE

As a reminder, to get the maximum likelihood estimators, we:

1. build the likelihood function $L(\theta)$ by multiplying the probability mass function (for discrete) or the probability density function (for continuous) for each of the sample values.
2. either plot $L(\theta)$, or (for larger sample sizes) get the derivative(s) $\frac{\partial L(\theta)}{\partial \theta}$ for the unknown parameter(s) and equate to zero.
3. if plotting, check to find the maxima it attains; if using the derivative(s), solve the system and obtain the maxima.
   - When we have multiple maxima, then pick the absolute biggest amongst them.

Maximum likelihood estimators can get pretty tough to calculate as we have samples of larger and larger size, so let us start easy.

Problem 1: Building a likelihood function

Assume we have been told that a continuous population X is distributed with pdf $f(x) = \frac{1}{2}(1+\theta x), -1 \leq x \leq 1$, where θ is a parameter with $-1 \leq \theta \leq 1$. We have collected a sample of size $n = 3$: $X_1 = 0.25, X_2 = 0.85, X_3 = -0.50$. What is the likelihood function $L(\theta)$?

Answer to Problem 1.

Problem 2: MLE for a sample of $n = 3$ observations

Based on your likelihood function from Problem 1, what is the maximum likelihood estimator for θ? You can solve this problem in two ways:

1. Plot the (degree 3) polynomial you obtained for $L(\theta)$. Then find the value for θ that leads to the maximum $L(\theta)$. If there are multiple maxima, pick the "highest" one.
2. Get the derivative $\frac{\partial L(\theta)}{\partial \theta}$ and set it equal to 0. Then solve the equation to obtain the value for θ. If there are multiple solutions, pick the one that leads to maximum $L(\theta)$. [130]

[130] If you are solving a quadratic equation, you are bound to get more than one solutions. That is because setting the derivative equal to 0 gives both maxima and minima: we only want the absolute maximum here!

Answer to Problem 2.

Problem 3: When the derivatives (and the plotting!) lie

We cannot *always* visualize and pick the maximum or get the derivatives and trust that the maximum is the estimator we are looking for. Let's see a case where this "issue" happens.

Recall again how continuous population X is distributed with pdf $f(x) = \frac{1}{2}(1+\theta x), -1 \leq x \leq 1$. Also recall that we have $-1 \leq \theta \leq 1$. We again pick a random sample of size $n = 3$. This time, though, it is: $X_1 = 0.75, X_2 = 0.65, X_3 = 0.80$. What is the maximum likelihood estimator now?

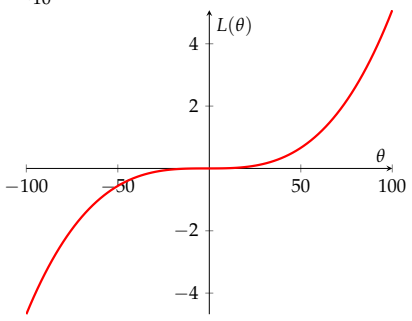
Answer to Problem 3.

So, this was weird, especially if we decided to take derivatives. Let's recap what we ran into in the next page.

1. **If we plot the likelihood function** $L(\theta) = f(X_1) \cdot f(X_2) \cdot f(X_3)$, or

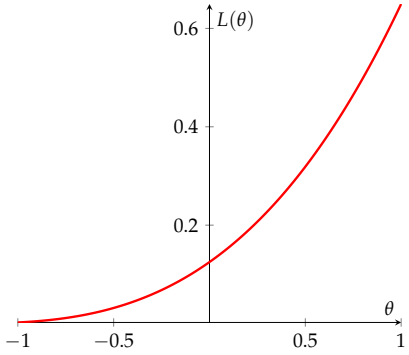
$$L(\theta) = \left(\frac{1}{2}\right)^3 \cdot (1 + 0.75 \cdot \theta) \cdot (1 + 0.65 \cdot \theta) \cdot (1 + 0.80 \cdot \theta),$$

we obtain:



It seems like the maximum is achieved for very very large values of θ.

Recall though that $-1 \leq \theta \leq +1$. Then, we have:



The maximum is clearly attained at $\boxed{\hat{\boldsymbol{\theta}} = \mathbf{1}}$.

2. **If we take the derivative and set to 0**, we get:

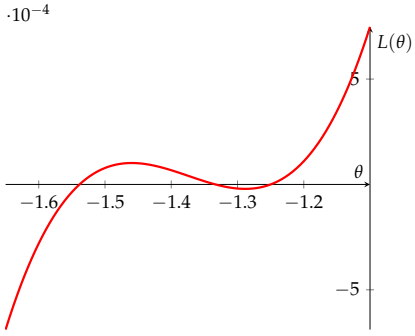
$$\frac{\partial L(\theta)}{\partial \theta} = 0 \implies 0.275 + 0.401875 \cdot \theta + 0.14625 \cdot \theta^2 = 0 \implies \theta = \begin{cases} -1.45964 \\ -1.28822 \end{cases}.$$

Neither is between -1 and $+1$, hence, before deciding the maximum, we'd need to also check the values at -1 and $+1$. Comparing, we get:

$$L(-1) = \left(\frac{1}{2}\right)^3 \cdot (1 + 0.75 \cdot (-1)) \cdot (1 + 0.65 \cdot (-1)) \cdot (1 + 0.80 \cdot (-1)) = 0.0021875.$$
$$L(+1) = \left(\frac{1}{2}\right)^3 \cdot (1 + 0.75 \cdot 1) \cdot (1 + 0.65 \cdot 1) \cdot (1 + 0.80 \cdot 1) = 0.6496875.$$

Contrasting, we would pick $\boxed{\hat{\boldsymbol{\theta}} = \mathbf{1}}$.

As a parenthesis, here are the two "zero derivative" points: they are not both maxima! And, as discussed earlier, even the local maximum here is not as big as the "global" maximum achieved at $\hat{\theta} = 1$.



Activity 2: Streaming services and their data (reloaded)

Let us revisit the example from last time. Remember that TV streaming giant with the data they had collected? Well, they are back and they'd like to see what else they can do to come up with estimators.

As a reminder, they have provided you with data on the number of episodes people watch during a session (an integer number).

	Session #									
	1	2	3	4	5	6	7	8	9	10
# of episodes	1	3	3	3	2	5	1	5	3	1

Let us help them find a good estimator using the maximum likelihood method this time around.

Problem 4: Poisson rates in the general case

You have been provided with the actual observations in the sample. What if we had not collected a sample yet, though? Can you still calculate the maximum likelihood estimator as a function of the sample, whatever it may be?

Let's put this to practice for the rate of a Poisson distributed population. What is λ as a function of the sample collected $X_1, X_2, \ldots, X_n$? Don't use the provided data just yet. See if you can derive the MLE as a general function of $X_1, X_2, \ldots, X_n$.

Answer to Problem 4.

Problem 5: Poisson rates in the general case (reloaded)

While the previous result is not terribly difficult to derive, it is still confusing sometimes to take the derivative of a product of functions, as is the case with general likelihood.

Following our logic from earlier, the likelihood function for a general sample of size n (let it be $X_1, X_2, \ldots, X_n$) would be:

$$L(\lambda) = e^{-\lambda} \cdot \frac{\lambda^{X_1}}{X_1!} \cdot e^{-\lambda} \cdot \frac{\lambda^{X_2}}{X_2!} \cdot \ldots e^{-\lambda} \cdot \frac{\lambda^{X_n}}{X_n!}.$$

We would need the derivative of this [131], and this is not always easy. In the notes, we mentioned something called the log-likelihood. For the log-likelihood, you should still calculate the likelihood function but then take its logarithm to obtain $\ln(L(\theta))$. Then, you can take its derivative and equate it to 0. [132] Using the logarithm properties we may calculate

$$\ln(L(\lambda)) = \ln\left(e^{-\lambda} \cdot \frac{\lambda^{X_1}}{X_1!}\right) + \ln\left(e^{-\lambda} \cdot \frac{\lambda^{X_2}}{X_2!}\right) + \ldots + \ln\left(e^{-\lambda} \cdot \frac{\lambda^{X_n}}{X_n!}\right),$$

which is **so much easier to differentiate**! What would be the maximum likelihood estimator?

Answer to Problem 5.

It is of course the same either way! Using the likelihood or the log-likelihood to differentiate and equate to 0 will always give the same result.

[131] In terms of λ, recall that the observations X_i in the sample are supposed to be known values!

[132] Some properties logarithms have:

$$\ln(a \cdot b) = \ln a + \ln b$$

$$\ln\left(a^b\right) = b \cdot \ln a$$

Problem 6: Solving the example

Based on your answer in Problem 4 or 5, use the data provided for the number of episodes watched per session and calculate the maximum likelihood estimator for the rate λ. Is it the same as the rate you got when you used the method of moments during the previous lecture?

Answer to Problem 6.

This rings a bell. This is the same as the estimator we obtained last time using the method of moments! Are these two methods (MLE and method of moments) always giving us the same estimators?

Activity 3: Extra details

In this activity, we see a couple of special cases and answer the following questions:

1. Are the method of moments and MLE always giving us the same estimators?

   See Problem 6 (where they are the same) and Problem 7 (where they are not).

2. Is there a time when the actual estimator is different than the one obtained when setting the derivative(s) equal to 0? Or maybe where the parameter can be calculated without derivatives?

   See Problem 3 (earlier) and Problems 8-9.

3. How can we estimate two or more parameters using MLE?

   See Problem 10 (comes pre-filled).

Problem 7: Not always the same

Consider the following probability density function [133] $f(x) = \theta x^{\theta-1}$ defined over $0 \leq x \leq 1$. Assume we have been provided a sample of size n: $X_1, X_2, \ldots, X_n$. We can apply the method of moments to get:

$$E[X] = \frac{1}{n}\sum_{i=1}^{n} X_i \implies \int_0^1 x f(x)dx = \overline{X} \implies \int_0^1 \theta x^{\theta} dx = \overline{X} \implies$$

$$\implies \theta \frac{x^{\theta+1}}{\theta+1}\Big|_0^1 = \overline{X} \implies \frac{\theta}{\theta+1} = \overline{X} \implies \hat{\theta} = \frac{\overline{X}}{1-\overline{X}}.$$

How about the maximum likelihood estimator? What would it be? [134]

[133] Taken from last semester's exam 2! So... good practice for sure!

[134] You will need to take the derivative of a^x as far as x is concerned. We have:

$$\frac{\partial a^x}{\partial x} = a^x \cdot \ln a.$$

Answer to Problem 7.

So, as we just saw the method of moments and the maximum likelihood estimation method match often; but not always.

Problem 8: When the derivatives "lie" (revisited)

We go back to the streaming service. For convenience, here is some data they obtained for the streaming time (continuous random variable) for 10 sessions.

	Session #									
	1	2	3	4	5	6	7	8	9	10
Time (in hours)	0.75	1.20	1.33	0.97	0.80	1.43	0.87	1.41	1.09	1.05

Let us assume that they believe people to be watching episodes in time that is uniformly distributed and continuous in $[\theta, 2\theta]$. What is the likelihood function based on the sample they have obtained? [135]

[135] Recall that $f(x) = \frac{1}{\theta}$ for $\theta \leq x \leq 2\theta$. So, let us create $L(\theta) = f(X_1) \cdot f(X_2) \dots f(X_n)$. Wait, is this always the same? No matter the sample?

Answer to Problem 8.

$L(\theta) =$

Problem 9: When the derivatives "lie" (revisited)

Hopefully, we have gotten that $L(\theta) = \left(\frac{1}{\theta}\right)^n$, where $n = 10$ in our case. This is an always decreasing function! [136] Because of that, for $L(\theta)$ to be maximum, we want θ to be smallest..

[136] We can tell: taking the derivative we get $-n\left(\frac{1}{\theta}\right)^{n-1}$ which is negative.

What is the smallest value that θ is allowed to take? Recall that we need $\theta \leq X_i$ and $2\theta \geq X_i$ for **all our observations** X_i! This means that $\theta \leq 0.75$ and $2\theta \geq 0.75$, $\theta \leq 1.20$ and $2\theta \geq 1.20$, and so on. With that in mind, what is the maximum likelihood estimator for θ?

Answer to Problem 9.

$\hat{\theta} =$

To easily solve this? Only pick the smallest and the largest observation and then only write these two constraints: $\theta \leq 0.75$ and $2\theta \geq 1.43$, which implies that $\theta \leq 0.75$ and $\theta \geq 0.715$, which finally gives that $\hat{\theta} = 0.715$. Cool, no? No derivatives at all!

This last problem comes pre-solved. **Please study the derivation** for an example of using MLE for two parameters!

Problem 10: The normal distribution

Assume that you have n observations from a normally distributed population with unknown mean and variance. What are the maximum likelihood estimators for μ and σ^2? [137]

[137] We need the pdf of the normal distribution to answer this question. You may use that $f(x) = \frac{1}{\sqrt{2\pi}\cdot\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.

Answer to Problem 10.

Let's build the likelihood function:

$$L(\mu,\sigma) = f(X_1)\cdot f(X_2)\cdot f(X_3)\cdot f(X_4)\cdot f(X_5) = \\ = \frac{1}{\sqrt{2\pi}\cdot\sigma} e^{-\frac{(X_1-\mu)^2}{2\sigma^2}} \cdot \frac{1}{\sqrt{2\pi}\cdot\sigma} e^{-\frac{(X_2-\mu)^2}{2\sigma^2}} \ldots \frac{1}{\sqrt{2\pi}\cdot\sigma} e^{-\frac{(X_5-\mu)^2}{2\sigma^2}}.$$

Focus on the exponent calculation for a second. First of all, we know that $e^a \cdot e^b = e^{a+b}$. Hence, we need to sum the exponents. We would have:

$$-\frac{(X_1-\mu)^2}{2\sigma^2} - \frac{(X_2-\mu)^2}{2\sigma^2} - \frac{(X_3-\mu)^2}{2\sigma^2} - \frac{(X_4-\mu)^2}{2\sigma^2} - \frac{(X_5-\mu)^2}{2\sigma^2} = \\ = \frac{-X_1^2 + 2\mu X_1 - \mu^2 - X_2^2 + 2\mu X_2 - \mu^2 - \ldots - X_5^2 + 2\mu X_5 - \mu^2}{2\sigma^2} = \\ = -\frac{\sum X_i^2 - 2\mu\sum X_i + n\mu^2}{2\sigma^2}.$$

Finally, this gives us:

$$L(\mu,\sigma) = \left(\frac{1}{\sqrt{2\pi}\cdot\sigma}\right)^n \cdot e^{-\frac{\sum X_i^2 - 2\mu\sum X_i + n\mu^2}{2\sigma^2}}.$$

Now, on to the derivatives... Why plural? Well, we have two unknown parameters! First, for μ:

$$\frac{\partial L(\mu,\sigma)}{\partial\mu} = 0 \implies \frac{\partial\left[\left(\frac{1}{\sqrt{2\pi}\cdot\sigma}\right)^n \cdot e^{-\frac{\sum X_i^2 - 2\mu\sum X_i + n\mu^2}{2\sigma^2}}\right]}{\partial\mu} = 0.$$

Let's use some derivative properties. First of all, the left part of the function is a constant as far as μ is concerned. Hence, we can just take it out:

$$\frac{\partial\left[\left(\frac{1}{\sqrt{2\pi}\cdot\sigma}\right)^n \cdot e^{-\frac{\sum X_i^2 - 2\mu\sum X_i + n\mu^2}{2\sigma^2}}\right]}{\partial\mu} = \left(\frac{1}{\sqrt{2\pi}\cdot\sigma}\right)^n \cdot \frac{\partial\left[e^{-\frac{\sum X_i^2 - 2\mu\sum X_i + n\mu^2}{2\sigma^2}}\right]}{\partial\mu}.$$

Answer to Problem 10 (continued).

Then, for the right part of the function, we have that

$$\left(e^{-g(x)}\right)' = -\left(g(x)\right)' e^{-g(x)}.$$

Specifically, in this case we have

$$g(\mu) = \frac{\sum X_i^2 - 2\mu \sum X_i + n\mu^2}{2\sigma^2} \text{ and } \left(g(\mu)\right)' = 2\mu n - 2\sum X_i.$$

Combining:

$$\begin{aligned}
\frac{\partial L(\mu,\sigma)}{\partial \mu} = 0 \implies & \frac{\partial\left[\left(\frac{1}{\sqrt{2\pi}\cdot\sigma}\right)^n \cdot e^{-\frac{\sum X_i^2 - 2\mu\sum X_i + n\mu^2}{2\sigma^2}}\right]}{\partial\mu} = 0 \implies \\
\implies & \left(\frac{1}{\sqrt{2\pi}\cdot\sigma}\right)^n \cdot \left(2\sum X_i - 2\mu n\right) e^{-\frac{\sum X_i^2 - 2\mu\sum X_i + n\mu^2}{2\sigma^2}} = 0 \\
\implies & \left(2\sum X_i - 2\mu n\right) = 0 \implies \hat{\mu} = \frac{\sum X_i}{n} = \overline{X}.
\end{aligned}$$

This is true as both $\left(\frac{1}{\sqrt{2\pi}\cdot\sigma}\right)^n > 0$ and $e^{-\frac{\sum X_i^2 - 2\mu\sum X_i + n\mu^2}{2\sigma^2}} > 0$.
Now, for the derivative as far as σ is concerned:

$$\frac{\partial L(\mu,\sigma)}{\partial \sigma} = 0 \implies \frac{\partial\left[\left(\frac{1}{\sqrt{2\pi}\cdot\sigma}\right)^n \cdot e^{-\frac{\sum X_i^2 - 2\mu\sum X_i + n\mu^2}{2\sigma^2}}\right]}{\partial\sigma} = 0.$$

Recall that $(f(x)g(x))' = f'(x)g(x) + f(x)g'(x)$. Using for simplicity that $g(\mu) = \sum X_i^2 - 2\mu\sum X_i + n\mu^2$, leads to:

$$\begin{aligned}
\frac{\partial L(\mu,\sigma)}{\partial \sigma} = 0 \implies & \frac{\partial\left[\left(\frac{1}{\sqrt{2\pi}\cdot\sigma}\right)^n\right]}{\partial\sigma} \cdot e^{-\frac{g(\mu)}{2\sigma^2}} + \left(\frac{1}{\sqrt{2\pi}\cdot\sigma}\right)^n \cdot \frac{\partial\left[e^{-\frac{g(\mu)}{2\sigma^2}}\right]}{\partial\sigma} = 0 \\
\implies & -\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \frac{n}{\sigma} \cdot e^{-\frac{g(\mu)}{2\sigma^2}} + \left(\frac{1}{\sqrt{2\pi}\cdot\sigma}\right)^n \cdot e^{-\frac{g(\mu)}{2\sigma^2}} \cdot \frac{g(\mu) e^{-\frac{g(\mu)}{2\sigma^2}}}{\sigma^3} = 0 \\
\implies & \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{g(\mu)}{2\sigma^2}} \left(-\frac{n}{\sigma} + \frac{g(\mu)}{\sigma^3}\right) = 0 \implies \\
\implies & -\frac{n}{\sigma} + \frac{g(\mu)}{\sigma^3} = 0 \implies -n\sigma^2 + g(\mu) = 0 \implies \\
\implies & \hat{\sigma}^2 = \frac{g(\mu)}{n} = \frac{\sum X_i^2 - 2\mu\sum X_i + n\mu^2}{n}
\end{aligned}$$

Replacing that $\mu = \overline{X}$, we get:

$$\hat{\sigma}^2 = \frac{\sum X_i^2}{n} - \overline{X}^2.$$

19. Methods of estimation: Bayesian estimation

Every worksheet will work as follows.

1. You will be asked to form a group with other students in the class: you can make this as big or as small as you'd like, but groups of 4-5 work best.
2. Read through the worksheet, discussing any questions with the other members of your group.
   - You can call me at any time for help!
   - I will also be interrupting you for general guidance and announcements at random points during the class time.
3. Answer each question (preferably in the order provided) to the best of your knowledge.
4. While collaboration between students is highly encouraged and expected, each student has to submit their own version.
5. You will have 24 hours (see gradescope) to submit your work.

Activity 1: A simple Bayesian estimation

As a reminder, to get the Bayesian estimators, we:

1. identify our prior probabilities/distribution: this quantifies what we think the unknown parameters are *before* observing any sample.
2. build the likelihood function $L(\theta)$ by multiplying the probability mass function (for discrete) or the probability density function (for continuous) for each of the sample values (**exactly what we did last time for MLE**).
3. compute the posterior probabilities/distribution by multiplying the priors with the likelihood.
4. find the maximum posterior probability or find the point where the posterior distribution attains its maximum.
   - This latter part can be done by either comparing all probabilities and picking the biggest one or
   - by getting the derivative(s) of the posterior distribution for each of the unknown parameter(s) and setting it to 0

Problem 1: Creating a table

For smaller problems, it may be useful to create a table. In the table, we store the prior probabilities (based on past information or prior beliefs), the likelihood functions (based on the sample), and computer the posterior probabilities. Let us see this in an example.

A school can be in one of three categories: $R1, R2, R3$. $R1$ schools place 90% of their graduates in good jobs, $R2$ schools place 75% of their graduates in good jobs, and $R3$ schools place 50% of their graduates in good jobs. You may assume that each school has a $\frac{1}{3}$ chance of appearing.

You have been observing a school and have found that 82 of their last 100 graduates have been placed in good jobs. Is the school an $R1$, $R2$, or $R3$ school? Equivalently, if p is the probability of placing a graduate in a good position, what is $\hat{p}$ for the specific school begin observed?

Answer to Problem 1.

	priors $\pi(p)$	likelihood $L(X\|p)$	posterior $\pi(p) \cdot L(X\|p)$
$p = 0.9$			
$p = 0.75$			
$p = 0.5$			

Problem 2: Normalizing the result

Based on your calculations, you probably noticed that the posterior probabilities are very small. Can you normalize them so as to answer the question: "how certain are you that the school is an R1/R2/R3 institution?" [138]

[138] Normalizing entails getting each end result and dividing by the summation of all of your results.

Answer to Problem 2.

Activity 2: Streaming services and their data (Bayesian remix)

We turn one last time to the data from the TV streaming giant. As a reminder, they have given you data on the time people spend during a session (continuous); as well as the number of episodes people watch during a session (discrete number). In this activity, we focus solely on the number of episodes streamers watch per session.

	Session #									
	1	2	3	4	5	6	7	8	9	10
# of episodes	1	3	3	3	2	5	1	5	3	1

Problem 3: Poisson with prior beliefs

The streaming giant believes that there are three types of customers. For all three types the number of episodes they watch is Poisson distributed; they do have different rates λ, though. The three types of customers are:

- 75% of their clientele watches a number of episodes that is Poisson distributed with $\lambda = 1$.
- 10% of their clientele watches a number of episodes that is Poisson distributed with $\lambda = 3$.
- 15% of their clientele watches a number of episodes that is Poisson distributed with $\lambda = 4$.

Using this new information, and assuming that all data has been collected from one type of customers alone, what is the Bayesian estimator for λ? [139]

[139] In other words, has the data been obtained from a customer with $\lambda = 1$, $\lambda = 3$, or $\lambda = 4$?

Answer to Problem 3.

Problem 4: Poisson with more general prior beliefs

Here we go in the general, continuous case. What if the streaming giant was wrong and customers are not discrete (that is $\lambda = 1, 3,$ or 4), but are instead continuous (λ is *anything* in the 1 to 4 range)? For this problem, you may assume that λ is uniformly distributed between 1 to 4. What is the Bayesian estimator for λ in this case? [140]

[140] You will need again the fact that the pdf of the uniform distribution is $\frac{1}{b-a}$, where a and b are the lower and upper bounds of the uniform.

Answer to Problem 4.

We may summarize this result as follows. When presented with discrete cases, we will calculate posterior *probabilities* (normalize them, if we prefer) and report the maximum among them. In the continuous case, we will calculate the posterior *distribution* and find the maximizer (potentially by plotting it and observing it, or by setting the derivative equal to 0, when possible).

Activity 3: Coins

Problem 5: Step-by-step

Assume you carry with you three coins with probability of Heads $p = 0.25, 0.5, 0.75$. You pick a coin and you flip it. Which coin do you (believe you) have picked to flip if: [141]

[141] Assume the events presented later come in sequence.

Answer to Problem 5.

the first coin comes up Heads?

the second coin comes up Heads?

the third coin comes up Tails?

the fourth coin comes up Tails?

the fifth coin comes up Tails?

Now, let us run into one of the interesting aspects of Bayesian estimation...

Problem 6: Coins again

You still have an unfair coin at your disposal that brings Heads with unknown probability p. This time though, your coin is not one of three like earlier. Instead, the probability p is uniformly distributed between 0.4 and 0.6.

You decide to run an experiment to estimate p. You will toss the coin as many times as necessary to get Heads and record the number of times until Heads are observed. [142] You have obtained $N_1 = 5, N_2 = 4, N_3 = 5$.

With this in mind, what is the Bayesian estimator for p?

[142] This is... geometric, no?

Answer to Problem 6.

This is interesting! If our maximizer is above the largest allowable value, or below the smallest allowable value, then we need to pick the largest/smallest one! You may also show this graphically, if you prefer, to see what happens.

Activity 4: Continuous case only

What if both the distribution we are estimating is continuous **and** the distribution of the parameter is continuous at the same time?

Problem 7: Normal and exponential

The time between two consecutive vehicles passing through an intersection is exponentially distributed with unknown rate λ; however, we do know that λ is normally distributed with $\mu = 2$ per minute and variance $\sigma^2 = 0.25$. In essence, this states that typically 2 vehicles would pass every minute, but this can go lower to 1 every 2 minutes (when it is less busy) and higher to 7 every 2 minutes (when it is much busier).

A new day has just begun and we would like to see how many vehicles to expect in the next 2 hours. During the first 5 minutes, we saw the first vehicle appear in $T_1 = 0.5$ minutes, $T_2 = 1$ minutes, $T_3 = 2$ minutes, $T_4 = 1.5$ minutes. Recall that these are times between two consecutive vehicles.

With this in mind, what is the Bayesian estimator for λ?

Answer to Problem 7.

20. Confidence intervals for single population means

Every worksheet will work as follows.

1. You will be asked to form a group with other students in the class: you can make this as big or as small as you'd like, but groups of 4-5 work best.
2. Read through the worksheet, discussing any questions with the other members of your group.
   - You can call me at any time for help!
   - I will also be interrupting you for general guidance and announcements at random points during the class time.
3. Answer each question (preferably in the order provided) to the best of your knowledge.
4. While collaboration between students is highly encouraged and expected, each student has to submit their own version.
5. You will have 24 hours (see gradescope) to submit your work.

Activity 1: Our first confidence intervals

Today we build our first confidence intervals!

Radon (a naturally occurring radioactive gas) is considered safe if found in quantities less than or equal to 4.0 picocuries per liter of air, or 4.0 pCi/L. Radon quantity is assumed to be **normally distributed**. To measure radon, house inspectors collect samples over the course of a few days. Throughout this first activity, we will assume that every day leads to exactly one measurement.

Problem 1: Estimating the unknown mean of Radon

A house inspector has taken measurements using equipment that measures with a standard deviation of 1 pCi/L over a period of 3 days (and hence has 3 measurements) and has found that the level of Radon was $X_1 = 3.6\text{pCi/L}, X_2 = 4.1\text{pCi/L}, X_3 = 3.4\text{pCi/L}$. What is the method of moments point estimate based on the sample obtained? [143]

Answer to Problem 1.

[143] Don't forget! The method of moments estimator for the unknown mean of a normal distribution is just the sample average! Or, in math terms:

$$\hat{\mu} = \overline{X}.$$

Problem 2: Building a confidence interval

So, the average is below 4pCi/L. But, are you sure it is safe to live in the house without doing any treatment for Radon? What is the 90% confidence interval for the mean Radon quantity in the house? Recall that σ is known and is equal to 1pCi/L.

Answer to Problem 2.

Problem 3: Increasing the number of observations

Assume we are collecting 7 days worth of observations, and the average is still 3.7pCi/L. What is the 95% confidence interval now? Would you say the house is safe with probability 95% based on this new confidence interval? Again, don't forget that σ is given to be equal to 1pCi/L.

Answer to Problem 3.

Activity 2: One-sided confidence intervals

One could claim that doing a two-sided confidence interval is too conservative. Wouldn't we still want to live in the house if the radon quantity is below the lower bound of the confidence interval built?

Assume we have collected $n = 7$ days worth of observations (so a sample of size $n = 7$) with a reported sample average of $\overline{X} =$ 3.7pCi/L. Also recall that we already know $\sigma = 1 pCi/L$.

Problem 4: One-sided 90% confidence interval

What is the **one-sided** (upper) 90% confidence interval? [144]

Answer to Problem 4.

[144] In mathematical terms, what is the $(-\infty, U]$ confidence interval? Of course, Radon cannot be negative: but this shouldn't change the operations we do.

Problem 5: One-sided 95% confidence interval

What is the **one-sided** (upper, again) 95% confidence interval?

Answer to Problem 5.

Problem 6: Standard deviations

What if σ was different? The standard deviation depends on our measurement tools, no? So, an interesting follow-up question would be: what is the maximum value of σ such that the one-sided (upper) 95% confidence bound is below 4 pCi/L?

Answer to Problem 6.

Activity 3: Errors

Problem 7: Limiting the error for two-sided intervals

Before buying a house, interested buyers want to be sure (within some limit) about the Radon quantities. So, they ask the house inspector to verify that. How many days worth of observations should the inspector collect to address their concerns with an **estimation error** that is at most equal to 0.5pCi/L for the **two-sided** 95% confidence interval? [145]

[145] Look at Page 15 of your notes. :)

Answer to Problem 7.

Problem 8: Limiting the error for one-sided intervals

Recall though that the customers do not care about an error on the lower side of the spectrum! They only care for the upper bound; so they are more interested in a one-sided, upper 95% confidence interval. Still, though, the inspector wants to address their questions with a number of measurements enough to limit the estimation error less than or equal to 0.5pCi/L. How many days should they collect data for when interested in limiting the error for the one-sided (upper) confidence interval? [146]

[146] Think about the difference between two- and one-sided confidence intervals. Then, think about where that difference appears in the formula you used earlier.

Answer to Problem 8.

This leads us to recognize that for one-sided confidence intervals, to limit the error to below E we would need the proper number of samples n equal to:

$$\boxed{n = \left(\frac{z_\alpha \sigma}{E}\right)^2}$$

Activity 4: Extensions

Problem 9: Estimating the unknown mean of Radon

Let us assume the quantity of Radon is still normally distributed. But, **what if the standard deviation is unknown?** [147] Assume that the sample collected (over the course of $n = 7$ days, so $n = 7$ observations) has lead us to $\overline{X} = 3.7\text{pCi/L}$ with sample standard deviation $s = 0.9\text{pCi/L}$. What are the 95% two-sided and one-sided (upper) confidence intervals now? [148]

[147] Then, we need the sample standard deviation. But using it leads to a non-normal sampling distribution...

[148] Use the t-table provided in the notes.

Answer to Problem 9.

Contrast the results you got with the results when σ was known. What do you observe about the width of the intervals?

Problem 10: Estimating the unknown mean of Radon again

What if Radon quantities are not normally distributed? You may again assume that $\overline{X} = 3.7\text{pCi/L}$ and $s = 0.9\text{pCi/L}$, but now we have this after analyzing $n = 40$ observations (more than a month of measurements) [149]. What are the 95% two-sided and one-sided (upper) confidence intervals now?

Recall that in this case, you have a **large enough sample size**, but the population is **not normally distributed**.

[149] Hint: you may assume that $n = 40$ is a large enough sample.

Answer to Problem 10.

21. Confidence intervals for single population variances and proportions

Every worksheet will work as follows.

1. You will be asked to form a group with other students in the class: you can make this as big or as small as you'd like, but groups of 4-5 work best.
2. Read through the worksheet, discussing any questions with the other members of your group.
   - You can call me at any time for help!
   - I will also be interrupting you for general guidance and announcements at random points during the class time.
3. Answer each question (preferably in the order provided) to the best of your knowledge.
4. While collaboration between students is highly encouraged and expected, each student has to submit their own version.
5. You will have 24 hours (see gradescope) to submit your work.

Activity 1: Finding the proper χ^2 critical values

For variance confidence intervals, we need to use the χ^2 distribution. Due to its lack of symmetry, it can be confusing and painful to find the correct values to use. We begin with such an activity to help us get used to it.

Problem 1: Retrieving values from the χ^2 table

Answer to Problem 1.

- $\chi^2_{0.025,14} =$
- $\chi^2_{0.05,23} =$
- $\chi^2_{0.995,10} =$

Problem 2: Retrieving values to build variance confidence intervals

What are the χ^2 values you'd need to build the following two-sided confidence intervals? [150]

[150] Remember! Since χ^2 is not symmetric, you need to find **two values** for each two-sided confidence interval. You'd need one for a one-sided confidence interval.

Answer to Problem 2.

- $\alpha = 0.05$, $n = 10$:
- $\alpha = 0.05$, $n = 15$:
- $\alpha = 0.10$, $n = 20$:

Activity 2: A soil contamination problem

An engineer is concerned about soil contamination. They pick 15 soil samples and measure the contaminant levels finding that the sample average is $\overline{X} = 13.7$ ppm and the sample standard deviation is $s = 3.15$. You may assume that the soil contamination level is normally distributed with unknown mean and variance.

Problem 3: Back to basics

Construct a two-sided 95%-confidence interval for the unknown mean soil contamination, μ. [151]

[151] Check Lecture 20! What critical values should we use for a mean of a normally distributed population with unknown variance?

Answer to Problem 3.

Problem 4: A variance confidence interval

Construct a two-sided 95%-confidence interval for the unknown variance of the soil contamination, σ^2.

Answer to Problem 4.

Activity 3: One-sided confidence intervals

Like we did last time, we again check how to derive one-sided confidence intervals for the unknown variance.

Problem 5: One-sided 95% confidence interval

Using the data from the previous exercises (that is, $n = 15, s = 3.15$), construct a one-sided upper 95%-confidence interval for the unknown variance of the soil contamination, σ^2. [152]

Answer to Problem 5.

[152] That is, your variance would be in $[0, U]$, since the variance can never be negative!

Activity 4: Proportions

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".

Problem 6: Estimating the proportion in Portland, OR

Out of $n = 91$ respondents in Portland, 61 answered Yes. Create a 95% two-sided confidence interval on the proportion of Portland residents considering themselves environmentally conscious.

Answer to Problem 6.

Problem 7: Estimating the proportion in Philadelphia, PA

The same survey in Philadelphia [153] for $n = 100$ respondents gave that 45 of them agreed with the statement. Create a 95% two-sided confidence interval on the proportion of Philadelphia residents now considering themselves environmentally conscious.

[153] Where it's always sunny!

Answer to Problem 7.

What do you think? Based on their answers, it appears that Portland is more environmentally conscious than Philadelphia: however, would you believe it if we said that we are 95% certain that Portland is indeed more environmentally conscious? Yes or No?

Problem 8: Designing an experiment

Assume we want to perform the same survey in other cities in the United States. What is the smallest sample size we would need in order to estimate the proportion of residents that consider themselves environmentally conscious with a margin of error smaller than 3% (with 95% confidence again)?

Answer to Problem 8.

Observe that the number only depends on α and the desired E. For the same risk we are willing to take (α) and for the same estimation error (E), the number of samples to collect is identical regardless the application/context!

23. Confidence intervals for two populations

Every worksheet will work as follows.

1. You will be asked to form a group with other students in the class: you can make this as big or as small as you'd like, but groups of 4-5 work best.
2. Read through the worksheet, discussing any questions with the other members of your group.
   - You can call me at any time for help!
   - I will also be interrupting you for general guidance and announcements at random points during the class time.
3. Answer each question (preferably in the order provided) to the best of your knowledge.
4. While collaboration between students is highly encouraged and expected, each student has to submit their own version.
5. You will have 24 hours (see gradescope) to submit your work.

Activity 1: Comparing battery lives

When a new smartphone, tablet, laptop is released, one of the things that everyone wants to know is how good their batteries are. Assume a new phone has just come out and we want to see whether the new battery is better than the old one.

For the rest of Activity 1, we assume that the standard deviation of the previous model battery life is equal to $\sigma_1 = 1$ hour, and the standard deviation of the new model battery life is $\sigma_2 = 0.6$ hours.

Problem 1: Pooled standard deviation

What is the **pooled standard deviation** σ_P, assuming that we have a sample of $n_1 = 12$ previous model phones, and $n_2 = 20$ new model phones?

Answer to Problem 1.

$\sigma_P =$

Problem 2: Computing a 95% confidence interval

What is the (two-sided) 95% confidence interval on the difference between the two means $\mu_2 - \mu_1$? You may assume that you calculated the averages $\overline{X}_1 = 22.5$ hours and $\overline{X}_2 = 23$ hours. [154]

Answer to Problem 2.

[154] Do we need a z or a t critical value here? No matter which one you need, recall there are z and t tables in Lecture 20!

Problem 3: Comparing battery lives

When comparing the two battery lives, can you make the claim (with 95% confidence) that the new model has better battery life than the older model? Why/Why not? Briefly explain in a sentence or two. [155]

Answer to Problem 3.

[155] Does the confidence interval include cases where the new phone is better? How about cases where the old phone is better?

This is a nice consequence of confidence intervals. When comparing two means, if the confidence interval contains both positive and negative values, this means that the believable range of values for the mean difference $(\mu_1 - \mu_2)$ contains values where population 1 or population 2 are bigger. In our case, then, since the confidence interval contains negative values, this means that we cannot be sure (with 95% confidence) that the claim that the new phone has better battery life is true.

Activity 2: Unknown variances

In the previous activity, we assumed that the true standard deviations were known. What if this is not true?

Problem 4: Using the sample standard deviations

We use the same samples as before: we got $n_1 = 12$ previous model phones, and $n_2 = 20$ new model phones. These two samples led to averages $\overline{X}_1 = 22.5$ hours and $\overline{X}_2 = 23$ hours as well as sample standard deviations $s_1 = 1.6$ hours and $s_2 = 1.3$ hours. What is the **estimated pooled standard deviation** s_P now?

Answer to Problem 4.

$s_P =$

Problem 5: A 95% confidence interval (again)

Construct a two-sided 95%-confidence interval on the difference between the two means $\mu_2 - \mu_1$. [156]

[156] Do we need a z or a t critical value here? And, if we need a t value, what are the degrees of freedom?

Answer to Problem 5.

Activity 3: The F distribution and ratios of variances

During Lecture 23, we saw the "weird" F distribution. As a reminder, it is useful for constructing confidence intervals for the ratio of two variances. Formally the F distribution describes the ratio of two χ^2 random variables. Before we put it to the use (to construct a confidence interval), let's find some values.

Problem 6: Values straight from the F table

Let us experiment reading the table. What are the f values for: [157]

[157] We give a small hint by setting the first one up :)

Answer to Problem 6.

- 12 degrees of freedom in the numerator, 15 degrees of freedom in the denominator, and error α equal to 0.05?

  $f_{0.05,12,15} =$

- 5 degrees of freedom in the numerator, 5 degrees of freedom in the denominator, and error α equal to 0.10?

- 15 degrees of freedom in the numerator, 12 degrees of freedom in the denominator, and error α equal to 0.05?

Problem 7: Values that do not exist in the F table

What about the following values that do **not** exist in the F table? Recall that you can use the property that $f_{u,v,1-\alpha} = \frac{1}{f_{v,u,\alpha}}$. [158]

[158] Observe how $1-\alpha$ is changed to α and the degrees of freedom u, v are switched around to v, u.

Answer to Problem 7.

- 12 degrees of freedom in the numerator, 15 degrees of freedom in the denominator, and error α equal to 0.95?

- 5 degrees of freedom in the numerator, 5 degrees of freedom in the denominator, and error α equal to 0.90?

- 15 degrees of freedom in the numerator, 12 degrees of freedom in the denominator, and error α equal to 0.95?

Problem 8: Confidence intervals on ratios of variances

Let us go back to the new cell phone vs. the old cell phone batteries. Another important characteristic is how variable the battery life itself is. We have now collected a sample of $n_1 = 16$ new phones and $n_2 = 13$ old ones, which resulted in sample variances equal to $s_1^2 = 2$ hours^2 and $s_2^2 = 3.81$ hours^2 respectively. What is the 90% confidence interval on the ratio of the true variances? You may assume that **both battery lives** (for new and older version models) **are normally distributed**.

Answer to Problem 8.

Note how the interval includes values that are both greater and smaller than one. Hence, we cannot believe with certainty 90% that the one variance is bigger than the other. If the 90% confidence interval was all below 1, then we could believe (with 90% confidence) that $\sigma_1^2 < \sigma_2^2$ and vice versa.

Activity 4: Election Day 2020

A few days before an election, there is always quite the influx of new polls. Specifically, during the Presidential Election of 2020, there were a lot of polls that tried to analyze whether different populations vote differently in a significant manner. Let us try to analyze one of them (from our neighboring Iowa) with the help of a confidence interval.

Problem 9: Comparing proportions

The poll in question asked $n_1 = 444$ female likely voters and $n_2 = 409$ male likely voters [159]. Among the first sample, the observed population voting for a candidate was 188 (so, $\hat{p}_1 = \frac{188}{444} = 0.4$). In the second sample, the same population was found to be 233 (and hence, $\hat{p}_2 = \frac{233}{409} = 0.57$).

Can we make the claim (with 95% confidence) that the two populations view this upcoming election similarly? Construct the 95% confidence interval on $p_1 - p_2$ to help you answer the question.

[159] The poll separated voters only for these two identities. Had the poll included more options, the results would have definitely been more characteristic of the true population.

Answer to Problem 9.

If our $p_1 - p_2$ confidence interval includes 0, then we cannot make the claim that the two population proportions are different. For example, had we found that the $1 - \alpha$ confidence interval was $-0.005 \leq p_1 - p_2 \leq 0.225$, we may feel inclined to believe that p_1 is larger than p_2 on average... but we are not $1 - \alpha$ certain about it!

24-25. Introduction to hypothesis testing: hypothesis testing for proportions

Every worksheet will work as follows.

1. You will be asked to form a group with other students in the class: you can make this as big or as small as you'd like, but groups of 4-5 work best.
2. Read through the worksheet, discussing any questions with the other members of your group.
   - You can call me at any time for help!
   - I will also be interrupting you for general guidance and announcements at random points during the class time.
3. Answer each question (preferably in the order provided) to the best of your knowledge.
4. While collaboration between students is highly encouraged and expected, each student has to submit their own version.
5. You will have 24 hours (see gradescope) to submit your work.

Activity 1: Formulating hypotheses

Before we get started formulating statistical hypotheses, we need to clarify a couple of points. Formulating statistical hypotheses can be pretty difficult at first. That said, there are a few rules we may follow:

1. **How to state the null and alternative hypotheses?**
   - State the null hypothesis as an equality.
   - State the alternative hypothesis as either a two-sided inequality ($\neq$) or a one-sided inequality ($\leq, \geq$) depending on the hypothesis being tested.
2. **What are we proving or disproving?**
   - Rejecting a hypothesis in favor of the alternative hypothesis is a **strong conclusion**.
   - Failing to reject the null hypothesis is a **weak conclusion**.

   Due to that, we formulate our hypothesis in the following way. We set what we are interested in proving as the **alternative hypothesis**. Let us practice that in the next few problems.

Problem 1: Proportion of houses in the market

The house market has been interesting over the last few months with a lot of houses being in the market. Formulate the hypothesis that the true proportion of houses in the market right now is **not 10%**.

Answer to Problem 1.

H_0 :

H_1 :

Problem 2: Proportion of "A"s in a class

A class has recently changed instructors and students believe that the new instructor is "easier". Formulate the hypothesis that the true proportion of students that got an "A" in a class is **more than 30%**.

Answer to Problem 2.

H_0 :

H_1 :

Problem 3: Proportion of "F"s in a class

The same class (with the new instructor) is also assumed to have now fewer students failing. Formulate the hypothesis that the true proportion of students that got an "F" in a class is **less than 10%**.

Answer to Problem 3.

H_0 :

H_1 :

Activity 2: Internships

Problem 4: The proportion of students with an internship

The University of Illinois at Urbana-Champaign (UIUC) is interested in finding how many students have internships lined up for next summer. UIUC believes that more than 50% of them have secured internships. Formulate the statistical hypothesis that the true percentage is higher than 50%.

Answer to Problem 4.

H_0 :

H_1 :

Problem 5: The α and the β errors

Assume you ask a sample of UIUC students of size n through an online survey. In the survey, x of them will reply that they have indeed already secured an internship. Define (in a sentence or two) what α (Type I error) and what β (Type II error) are in this setup for the hypothesis you formulated in Problem 4.

Answer to Problem 5.

Activity 3: World Tourism Organization

The World Tourism Organization (WTO) needs your help! Due to COVID-19 and the effects it's had in tourism and travel, the WTO is preparing a new advertising campaign to encourage travel within the United States for US residents, after COVID-19 has been placed under control.

The WTO has certain ideas about United States residents and their travel patterns, but no idea whether they are right or not! Help them by performing a full scale hypothesis test.

Problem 6: Formulating a hypothesis about a proportion

The WTO claims that a fraction $p = 0.55$ of the population has visited 7 or more states. Is that the case or maybe it is a different fraction of the population has visited 7 or more states? Formulate this proportion statistical hypothesis test so as to try to prove the WTO wrong.

Answer to Problem 6.

Problem 7: Collecting a sample

Our plan is as follows. We will ask the $n = 90$ students in the class and help WTO reject or fail to reject their null hypothesis of $p = 0.55$. Let x be the number of students who have indeed been to 7 or more states. Assume that the true proportion of the population is indeed $p = 0.55$. What is the observed proportion $\hat{p} = \frac{x}{n}$ of $n = 90$ distributed as? [160]

Answer to Problem 7.

$\hat{p} \sim$

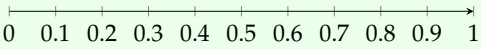
[160] First of all, make the argument that n is large enough for the central limit theorem to hold. Then, recall that $\hat{p}$ has to be normally distributed with the same mean as the original population and a variance that is the original variance over n.

Problem 8: Drawing

Get your artistic selves out! Based on your answer in Problem 5, draw the distribution here. [161]

[161] To help you: draw a normal distribution with a "peak" at 0.55 and a tapering off towards 0 at $0.55 + 3\sigma$ and $0.55 - 3\sigma$ (based on the σ you got from Problem 5).

Answer to Problem 8.



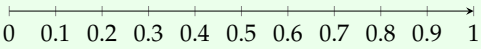
Problem 9: Adding a significance level

Let us select a significance level of 95% $\implies$ $\alpha = 5\%$. Mark the acceptance region for $\hat{p}$ by redrawing your plot with marks at

$$p_0 \pm z_{\alpha/2} \cdot \sqrt{\frac{p_0\,(1-p_0)}{n}},$$

where $p_0 = 0.55$.

Answer to Problem 9.



Why is that true? Well, as we remember from confidence intervals, if the true distribution is centered at $p_0 = 0.55$, then any sample we obtain would fall within the selection region with confidence 95%!

Problem 10: Bringing this together

Visit `http://vogiatzis.web.illinois.edu/random_lec24.html` to get your personalized number of respondents who have been to 7 or more states. If it is not working, call me in your breakout rooms, and I will give you the number.

Based on the sample you were given (or that you got online), should you reject or fail to reject the null hypothesis that the true proportion is $p = 0.55$? For example, say you got a number of respondents equal to 67 – should you reject or fail to reject that the null hypothesis of $p = 0.55$ based on the observed proportion $\hat{p} = 67/90 = 0.7444$?

Answer to Problem 10.

Activity 4: Extensions

Problem 11: The β error for overestimating

The WTO is worried that they may be overestimating the true proportion. What is the power of the test $(1 - \beta)$ should the true proportion be 0.3?

Answer to Problem 11.

Problem 12: The β error for underestimating

What if they are underestimating? What is the power of the test $(1 - \beta)$ should the true proportion be 0.8?

Answer to Problem 12.

Problem 13: P-values

Before we head out, let us check something interesting. Given a specific observed proportion $\hat{p}$, what is the maximum value of α such that $\hat{p}$ leads to a rejection? First, try it for a specific value. What is the maximum value of α such that $\hat{p} = 0.67$ leads to rejection? You may assume that we are still dealing with the original hypothesis test of

$$H_0 : p = 0.55.$$
$$H_1 : p \neq 0.55.$$

Answer to Problem 13.

Problem 14: P-values

How would you calculate this in the general case? We call the maximum value of α that leads to rejection the P-value. How can a P-value be calculated in general?

Answer to Problem 14.

26-27. Hypothesis testing for means and variances

Every worksheet will work as follows.

1. You will be asked to form a group with other students in the class: you can make this as big or as small as you'd like, but groups of 4-5 work best.
2. Read through the worksheet, discussing any questions with the other members of your group.
   - You can call me at any time for help!
   - I will also be interrupting you for general guidance and announcements at random points during the class time.
3. Answer each question (preferably in the order provided) to the best of your knowledge.
4. While collaboration between students is highly encouraged and expected, each student has to submit their own version.
5. You will have 24 hours (see gradescope) to submit your work.

Activity 1: Hypotheses for means of normally distributed populations

A tablet advertises that it has an all-day battery. The company claims that their battery will expectedly last 24 hours. Batteries of that capacity are assumed to have lives that are normally distributed with (known) standard deviation equal to 1.25 hours.

Problem 1: Formulating the hypothesis

Formulate a suitably hypothesis to check whether the battery life is smaller than 24 hours. [162]

[162] What do you think: is this a two-sided or a one-sided hypothesis test?

Answer to Problem 1.

H_0 :

H_1 :

Never forget that the rejecting our null hypothesis is the strong conclusion. Hence, if we are trying to check whether the battery life is less than 24 hours, we should have an alternate hypothesis of $\mu < 24$.

Problem 2: The hypothesis testing procedure

A random sample of $n = 10$ tablets revealed a sample average life of $\overline{X} = 23.2$ hours. Using $\alpha = 0.05$, can you support the claim that the battery life is 24 hours; or should you reject it in favor of the alternate? [163]

[163] Recall that σ is known and equal to 1.25 hours.

Answer to Problem 2.

Problem 3: Practicing P-values

What is the P-value for the observed average from Problem 2? Is this a "surprising" value or is it expected based on your decision in Problem 2?

Answer to Problem 3.

This shouldn't surprise us and we should indeed be expecting that P-value $< \alpha$, seeing as we ended up rejecting the null hypothesis. Had we failed to reject the null hypothesis, we would have anticipated that P-value $\geq \alpha$.

Problem 4: Unknown variance

Earlier in this activity, we made the assumption that the variance was known. What if the variance is unknown? If we collect a sample of $n = 10$ tablets that result in sample average $\overline{X} = 23.2$ hours and sample standard deviation $s = 1.8$ hours explain whether you can now reject or fail to reject the null hypothesis. [164]

[164] Use the same (one-sided!) hypothesis as formulated in Problem 1.

Answer to Problem 4.

Activity 2: The Type II error

It is high time we discuss β like we did during the the Lecture 24-25 worksheet. What changes when dealing with means? Assume that we are again using a sample of $n = 10$ tablets; further assume that the lifetime is normally distributed with known standard deviation $\sigma = 1.25$ hours. Finally, still use the original hypothesis as formulated in Problem 1:

$H_0 : \ \mu = 24$

$H_1 : \ \mu < 24$

Problem 5: What if the true mean is 23 hours?

What is the β error associated with the hypothesis test assuming the true mean of the battery is actually 23 hours? [165]

[165] If you are stuck, look at the bottom of the next page for a summary of the operations you need to do.

Answer to Problem 5.

Problem 6: Improving β

Assume that the β error you found in Problem 5 is unacceptable. The company is asking you for a way to improve this to which you recommend "get a bigger sample!" While you are right, and a bigger sample should decrease β, we would still like to keep the sample small enough. What is the smallest sample you should use in order for β to be equal to at most 10%? Use the same standard deviation $\sigma = 1.25$ and the same alternate hypothesis of $\mu_1 = 23$ hours (as in Problem 5).

Answer to Problem 6.

This is an interesting approach. In the calculation of β we perform three operations:

1. Find the distribution of $\overline{X}$ assuming that the true mean is equal to the alternate.
2. Use the bounds (L, U) from the original null hypothesis.
3. Calculate $P(L \leq \overline{X} \leq U)$.

In this exercise, we were told what $P(L \leq \overline{X} \leq U)$ is! Knowing L and U, as well as σ and μ_1 allows us to know everything we need, except for n. Hence, we solve for that (using the z-table).

Activity 3: Hypotheses for means of not normally distributed populations

Let us switch gears and move away from the tablet and battery lifetime world. The emergency department in a local hospital has observed that their average waiting time has historically been 45 minutes. The hospital hired new personnel and trained the previous hires in early 2020 in an effort to improve waiting times. We make the assumption that these times are **not normally distributed**.

Problem 7: Stating the hypotheses

State the null and alternate hypotheses for whether the waiting times *have improved* after the effort. [166]

[166] Recall that rejecting the null in favor of the alternate is the strong conclusion!

Answer to Problem 7.

H_0 :

H_1 :

Problem 8: Drawing a conclusion

A sample of $n = 94$ patients were audited after the improvement and the sample average waiting time was 42.1 minutes with a sample standard deviation of 10 minutes. Should you reject or fail to reject the hypothesis based on the proper test statistic (under $\alpha = 0.05$)? What is the corresponding P-value?

Answer to Problem 8.

Problem 9: Back to the normal distribution

Would your approach and answer in Problem 8 be different had you known that the times were normally distributed? Do not re-solve the problem; simply explain the differences (if any).

Answer to Problem 9.

Activity 4: Hypotheses for variances of normally distributed populations

The time to get served at a bank with three tellers [167] is normally distributed. The bank had already studied the standard deviation of the time to serve a customer and had found it to be equal to 5.6 minutes. The bank went ahead and hired an industrial engineer who recommended a new queuing setup. Instead of people waiting in three different lines (one for each teller), there is one super-queue where people wait for the first available teller.

[167] This is unimportant information for the purposes of this exercise.

Problem 10: Variance testing

After testing this new system, the sample standard deviation for $n = 30$ customers was found to be $s = 3.2$ minutes. Under $\alpha = 0.05$, is there enough evidence to support the statement that the new queue results in different time variance? [168]

[168] "Different" time variance should imply a two-sided hypothesis test!

Answer to Problem 10.

Problem 11: Variance testing

Once more, with the new super-queue, the sample standard deviation for $n = 30$ customers was found to be $s = 3.2$ minutes. Using $\alpha = 0.05$, is there enough evidence to support the statement that the new queue results in smaller time variance? [169]

Answer to Problem 11.

[169] The question now becomes: is it one- or two-sided?

Activity 5: Type I and Type II errors

Consider a hypothesis testing for the unknown mean of a UIUC student GPA (assumed to be a normally distributed population with known $\sigma = 0.5$):

$H_0: \ \mu = 3.33$

$H_1: \ \mu \neq 3.33$

To test the hypothesis, we have decided to ask $n = 20$ students what their GPA is and calculate their average.

Problem 12: Calculating α

If the $n = 20$ sample's GPA is either above 3.48 or below 3.18, we will reject the null hypothesis; else, we will fail to reject. What is α? [170]

Answer to Problem 12.

[170] This is new! We are typically given α and find the bounds of rejection. How do we deal with the opposite setup (given bounds, find α)?

Problem 13: Calculating β

For the bounds of Problem 12 ($L = 3.18$, $U = 3.48$), what is the power of test $1 - \beta$ assuming the true mean is 3.6 (with the same standard deviation of $\sigma = 0.5$)?

Answer to Problem 13.

28. Hypothesis testing for two populations

Every worksheet will work as follows.

1. You will be asked to form a group with other students in the class: you can make this as big or as small as you'd like, but groups of 4-5 work best.
2. Read through the worksheet, discussing any questions with the other members of your group.
   - You can call me at any time for help!
   - I will also be interrupting you for general guidance and announcements at random points during the class time.
3. Answer each question (preferably in the order provided) to the best of your knowledge.
4. While collaboration between students is highly encouraged and expected, each student has to submit their own version.
5. You will have 24 hours (see gradescope) to submit your work.

Activity 1: Comparing means

A pharmaceutical company is researching a new drug that has been cleared for human testing. This new drug (if cleared after testing) will replace a previously used drug that had a side effect: it impaired driving.

Two samples are selected and given the old drug (sample 1 of size $n_1 = 15$) and the new drug (sample 2 of size $n_2 = 10$). The reaction times while driving of people in the first sample ended up begin an average of $\overline{X}_1 = 4.65$ seconds with a standard deviation of $s_1 = 0.5$ seconds. For the second sample, the same numbers were $\overline{X}_2 = 4.36$ seconds and $s_2 = 0.3$ seconds, respectively.

You may assume that reaction times while driving are always **normally distributed**.

Problem 1: Formulating the hypothesis

Formulate a suitable hypothesis to check whether the second drug leads to different reaction times in driving.

Answer to Problem 1.

H_0 :

H_1 :

Problem 2: Formulating the hypothesis (correctly)

What if we want to check whether the new drug *improves* the side effect? That is, we want to check whether the reaction times are better than the ones of the old drug.

Answer to Problem 2.

H_0 :

H_1 :

This is as good a time as any to remind ourselves of one item of importance. When formulating a hypothesis, if we are interested in proving a claim, then we typically set it as the alternative hypothesis! The reason is that rejecting the null hypothesis is a stronger conclusion; rejecting the null in favor of the alternative essentially implies that we prove the alternative.

Hence, before you proceed, please verify that you have the following hypothesis setup:

$$\boxed{H_0 : \mu_1 = \mu_2 \text{ vs. } H_1 : \mu_1 > \mu_2.}$$

Problem 3: Known variances

Assume we know the variances of the reaction times when driving are known and equal to $\sigma_1^2 = 0.09$ seconds^2 for the first drug and $\sigma_2^2 = 0.16$ seconds^2 for the second drug. Using $\alpha = 0.05$, do you have enough evidence to reject the null hypothesis? That is, do you have enough evidence to deduce that the second drug lead to better side effects (=faster reaction times)? What is the corresponding P-value?

Answer to Problem 3.

Problem 4: β errors

It is not an easy feat to use the Student's T distribution (which appears alongside unknown variances) to calculate β errors and P-values. On the other hand, when the variances are known, and we have a standard normal distribution, calculations become easier.

With that in mind, assume you are formulating a hypothesis where the null hypothesis is that the two means are the same (i.e., $H_0 : \mu_1 - \mu_2 = 0$) vs. an alternative hypothesis that $\mu_1 < \mu_2$ (i.e., $H_1 : \mu_1 - \mu_2 < 0$). What is the β error of accepting the null hypothesis assuming that the true difference is $\mu_1 - \mu_2 = -0.3$? [171]

Answer to Problem 4.

[171] This means that the second drug is actually worsening reaction times while driving by 0.3 seconds.

Activity 2: Unknown variances

In this set of exercises, we will use the same data as in Activity 1. However, we will no longer assume that the variances are known.

Problem 5: Unknown (but equal!) variances

What if we have no idea what the true variances are, but we know they are supposed to be equal to one another? Then, using $\alpha = 0.05$, should you reject the null hypothesis? [172]

[172] Recall that we have been told that the sample averages and standard deviations for the two samples were $\overline{X}_1 = 4.65$ seconds, $s_1 = 0.5$ seconds, and $\overline{X}_2 = 4.36$ seconds, $s_2 = 0.3$ seconds, respectively.

Answer to Problem 5.

Problem 6: Unknown (and not necessarily equal) variances

Continuing on the same line of logic, what if we have the most general case? Assume now that not only you do not know the true variances, but you also do not know whether or not they are equal to one another.

What would you deduce in this case? Under $\alpha = 0.05$, should you reject the null hypothesis? [173]

[173] Do **not forget** the approximate degrees of freedom! See Canvas for a small "gift"!

Answer to Problem 6.

Activity 3: The "paired" t-test

The difference of two independent normally distributed random variables is also normally distributed. We have used this fact in many of our derivations.

Now, consider two independent and normally distributed populations with unknown variances σ_1^2 and σ_2^2. We get a random sample $X_1, X_2, \ldots, X_n$ from the first population and a random sample $Y_1, Y_2, \ldots, Y_n$ from the second population. Note how both samples are of equal size n. Now, consider $W_i = X_i - Y_i$. Clearly, if X and Y have the same mean, then we should expect each of the W_i to be small, no?

Problem 7: Devising a hypothesis testing procedure

Based on that, can you recommend a hypothesis test to check whether the two populations have the same mean? Explain what the setup of the hypothesis test is, which statistic you would use, and when you would reject or fail to reject the null hypothesis. [174]

Answer to Problem 7.

[174] $H_0 : \mu_1 - \mu_2 = 0$ vs. $H_1 : \mu_1 - \mu_2 \neq 0$.

How would you change this to accommodate one-sided hypotheses?

Problem 8: The paired t-test to the practice

A group of researchers at Illinois is investigating different modalities for teaching. To observe whether the different mode offered helps students learn better, they have conducted the following experiment. They will match students (based on performance) in pairs: one of the students in the pair will not have access to different modalities, whereas the other student will. Then, they will compare the performance of one group versus another, one pair of students at a time. The results are presented in Table 3.

Table 3: The observed performances per pair. For example, the first pair both got 20 points; on the other hand, the second pair saw the student without access get 14 points, whereas the student who had access performed better at 18 points.

Pair	No access	With access
1	20	20
2	14	18
3	13	18
4	14	15
5	15	14
6	12	15
7	10	15
8	8	13

Using this (limited) data on $n = 8$ pairs of students, can we claim that students with access perform better? Use $\alpha = 5\%$.

Answer to Problem 8.

29. Hypothesis testing for two populations

Every worksheet will work as follows.

1. You will be asked to form a group with other students in the class: you can make this as big or as small as you'd like, but groups of 4-5 work best.
2. Read through the worksheet, discussing any questions with the other members of your group.
   - You can call me at any time for help!
   - I will also be interrupting you for general guidance and announcements at random points during the class time.
3. Answer each question (preferably in the order provided) to the best of your knowledge.
4. While collaboration between students is highly encouraged and expected, each student has to submit their own version.
5. You will have 24 hours (see gradescope) to submit your work.

Activity 1: Comparing variances

There is a number of applications where **consistency** is key. Hence, we want to devise a procedure that helps us identify if the variance of our (normally distributed) populations has changed. In this activity, you are asked to do exactly that.

In a Senior Design project at a manufacturing facility in Northern Kentucky, a team of engineering students has been asked to compare two different facility layouts. Raw material is picked up from the storage rooms; they are transported to the main floor for processing; they are then transported to packaging; after that a finalized product is sent out for shipping. Both facility layouts have led to *similar* average times from beginning of the process to the end. Additionally, both layouts appear to lead to normally distributed times. That said, the company is particularly interested in what the variances are; smaller variances are definitely prepared.

To investigate these two layouts the company has allowed the students to collect data on $n_1 = 9$ replications from layout 1 and $n_2 = 16$ replications from layout 2. The students calculated that the sample standard deviation was $s_1 = 14$ minutes and $s_2 = 8.5$ minutes. Answer the following questions.

Problem 1: Formulating the hypothesis

Formulate a suitable hypothesis to test whether the two layouts lead to **different** variances.

Answer to Problem 1.

H_0 :

H_1 :

Problem 2: Comparing variances

Using $\alpha = 5\%$, do you have enough evidence to deduce that the two layouts lead to **different** variances?

Answer to Problem 2.

Problem 3: Comparing variances (one-sided)

What if we are interested in a one-sided hypothesis test? Using $\alpha = 5\%$, do you have enough evidence to deduce that the second layout leads to a **smaller variance**? Remember that you will need to formulate a different hypothesis now than the one in Problem 1.

Answer to Problem 3.

The F table can be confusing at times, so please practice with retrieving the values you need!

Activity 2: Comparing proportions

In a previous worksheet [175], we had discussed how environmentally conscious residents of different cities viewed themselves to be. Specifically, we had focused on the responses of two cities in the United States: one in the East and one in the West coast. Let us start from the one in the West: out of $n_1 = 91$ respondents in Portland, 61 answered that they viewed themselves to be environmentally conscious. In Philadelphia, on the other hand, $n_2 = 100$ respondents gave us 45 of them that stated they were environmentally conscious.

[175] See Worksheet 21.

Problem 4: Comparing proportions

Using $\alpha = 5\%$, can we claim that the two city resident populations differ in their views; or do they both consider themselves equally environmentally conscious? [176]

[176] Here, then, we ask you to consider $H_0 : p_1 - p_2 = 0$ vs. $H_1 : p_1 - p_2 \neq 0$.

Answer to Problem 4.

Problem 5: Comparing proportions (one-sided)

Using $\alpha = 10\%$, can we claim that the residents of Portland view themselves more environmentally conscious by 10 percentage points or more?

Answer to Problem 5.

Problem 6: Comparing proportions revisited

Can you make the same claim for $\alpha = 1\%$?

Answer to Problem 6.

Problem 7: P-values

What are the P-values for the hypothesis tests in Problems 4 and 5?

Answer to Problem 7.

Activity 3: A "full" test

Let us go back to the hypothesis tests of **means and variances** for two normally distributed populations. One class is trying a new tool for educational purposes. The students with access to the new tool are $n_1 = 8$ and have received final scores of (in decreasing order):

$$95, 93, 89, 89, 85, 80, 75, 67.$$

The students without access to the new tool are $n_2 = 6$ and have received final scores of:

$$83, 80, 80, 79, 77, 77.$$

Using $\alpha = 1\%, 5\%, 10\%$ perform suitable hypothesis tests to address the following two statements:

1. The new tool helps students do better in class.
2. The new tool amplifies discrepancies between students' access to technologies. This is shown in the variance of the students who got access to the tool (i.e., the variance seems to be higher).

Problem 8: Summary

To address the two statements above, we ask you to do the following:

- formulate suitable hypothesis;
- use suitable statistics;
- and deduce we have evidence in favor of the statements or not.

Answer to Problem 8.

Answer to Problem 8.

30. Linear regression

Every worksheet will work as follows.

1. You will be asked to form a group with other students in the class: you can make this as big or as small as you'd like, but groups of 4-5 work best.
2. Read through the worksheet, discussing any questions with the other members of your group.
   - You can call me at any time for help!
   - I will also be interrupting you for general guidance and announcements at random points during the class time.
3. Answer each question (preferably in the order provided) to the best of your knowledge.
4. While collaboration between students is highly encouraged and expected, each student has to submit their own version.
5. You will have 24 hours (see gradescope) to submit your work.

Activity 1: Regression fundamentals

Problem 1: Dependent vs. independent variables

What do we mean by dependent (or response) and independent (or predictor) variables? Provide one example of such a pair.

Answer to Problem 1.

When performing a linear regression, there are some underlying assumptions that we make. They are:

1. **Linearity**: i.e., that the relationship between the independent variable x and the dependent variable y is indeed linear.
2. **Homoscedasticity**: i.e., that the variance of the residuals is the same for any value of x.
3. **Independence**: it implies that the observations (data points) collected are independent from one another.
4. **Normality**: i.e., that the residuals are normally distributed. Equivalently, for any fixed value of the independent variable x, the dependent variable y is normally distributed.

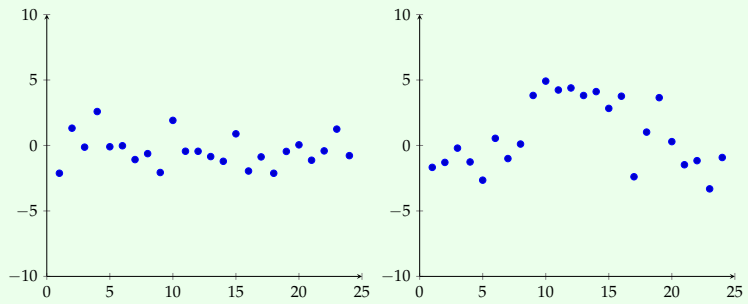
Problem 2: Linearity

Let's recommend a way to check for the linearity assumption. First, remember that we may define the residuals as $y_i - \hat{y}_i$: that is, the difference between the actual value of variable y_i versus the fitted value $\hat{y}_i$. Recall that if our regression model is of high quality, the residuals should have values close to 0.

Now, assume that you create a plot where the y axis is the residuals, and the x axis is the fitted values. This plot is called a **residual by predicted plot** or a **residual vs. fit plot**.

If you were to create such a plot, what should the plot look like in order for the linearity assumption to hold? To help, consider the following two figures showing two residual by predicted plots. Which one of the two appears to portray a linear relationship and which one does not?

Answer to Problem 2.

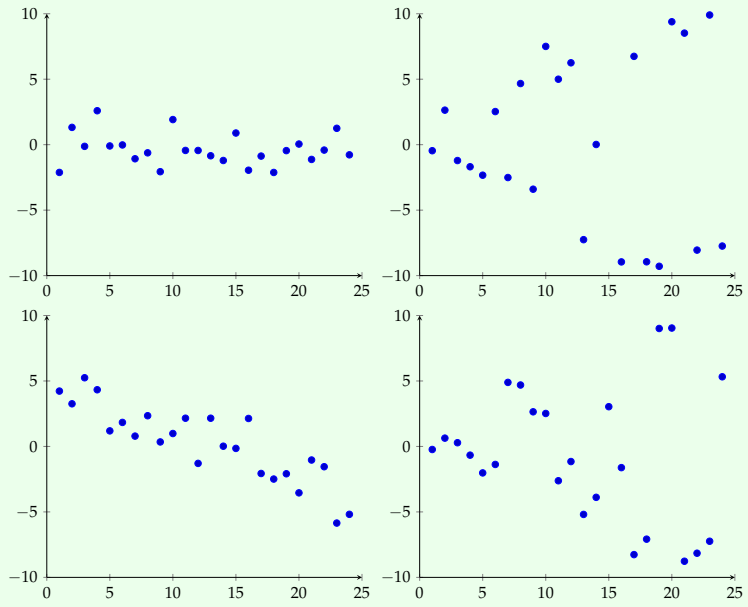


To check the linearity assumption, plotting the residuals should show them be **distributed around a horizontal line centered at 0 without a pattern**. A "bow"-like plot signals nonlinearity!

Problem 3: Homoscedasticity

Recommend a way to check for the homoscedasticity assumption. If we were to construct a residual by predicted plot again, what should the plot look like in order for the homoscedasticity (equal variance everywhere) assumption to hold? To help, consider the following four figures plotting the residuals. Which plots of the four below appear to portray that the variance of the residuals is the same throughout? Which ones appear to have different variances as the predicted values increase?

Answer to Problem 3.



As a general rule of thumb: if there is a "triangle" pattern or, equivalently, the values fan out, then the homoscedasticity assumption is not valid. If the residuals stay close to each other or follow a nondescript pattern, then the assumption can be viewed as valid.

Problem 4: Regression model vs. regression line?

If our data satisfies the linearity assumption, then we assume there exists a slope β_1 and an intercept β_0 such that: $y = \beta_0 + \beta_1 x$. However, when doing our regression, we write $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. Explain in a sentence why we use *observed* or *estimated* values in the regression line, and we do not do that in the original model. What do we have to add to the regression model to be realistic?

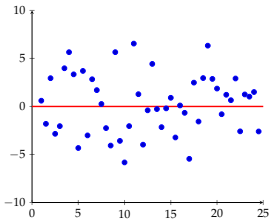
Answer to Problem 4.

The model is missing the **noise**! The "true" values would follow:

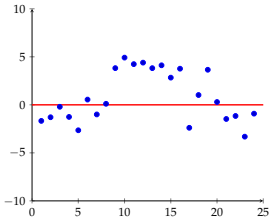
$$\boxed{y = \beta_0 + \beta_1 x + \epsilon}$$

Let us summarize this first activity. When plotting the residual values $y_i - \hat{y}_i$ versus the fitted values $\hat{y}_i$, we should expect them to form a straight line, centered at 0, following no pattern. Here are two such examples:

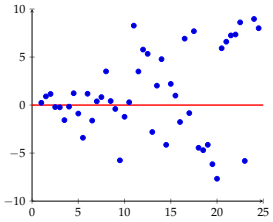
1. Residual vs. fitted values are centered at 0, in a linear manner, with a nondescript pattern.



2. Residual vs. fitted values are centered at 0, however follow a curve pattern (hints at non-linearity).



3. Residual vs. fitted values are centered at 0, but their values "fan out" as the fitted values get higher (hints at different variances – heteroscedasticity).



Activity 2: Fluoride levels

Water fluoridation has been a great way to prevent tooth decay (especially in children). The World Health Organization recently published **a report** where the recommended level of fluoride in a community water supply is ranging between 0.5 to 1.5 mg/L. The following data comes from studying children in 10 cities across the United States. Researchers studied the number of cavities per 100 children versus the level of fluoride in the water supply.

City	Fluoride	Cavities
1	1.9	236
2	2.6	246
3	1.8	252
4	1.2	258
5	1.2	281

City	Fluoride	Cavities
6	1.2	303
7	1.3	323
8	0.9	343
9	0.6	412
10	0.5	444

Problem 5: Finding the line

Find the least squares regression line using the data of the 10 cities given in the table.

Answer to Problem 5.

Problem 6: Using the line

Use the line you calculated to find the expected number of cavities per 100 children for a city with a fluoride level of 1.5.

Answer to Problem 6.

Problem 7: Calculating residuals

The *residuals* or *errors* are a very important measure in regression. They are typically calculated as

$$y_i - \hat{y}_i$$

with y_i being the observed value for city i and $\hat{y}_i$ being the fitted value had we used the regression line for city i.

Calculate the (10) residuals, one for each city. It is useful to do this calculation in table format: so a table is given to you for convenience!

Answer to Problem 7.

City (i)	Fluoride (x_i)	Cavities (y_i)	Fitted value ($\hat{y}_i$)	Residual ($y_i - \hat{y}_i$)
1	1.9	236		
2	2.6	246		
3	1.8	252		
4	1.2	258		
5	1.2	281		
6	1.2	303		
7	1.3	323		
8	0.9	343		
9	0.6	412		
10	0.5	444		

Activity 3: A "qualitative" regression

The Bureau of Labor Statistics has offered the following data on median weekly income per educational level attained (from 2019).

Level	Income	Level	Income
Less than high school	$520	Bachelor's degree	$1173
High school	$712	Master's degree	$1401
Some college	$774	Professional degree	$1836
Associate degree	$836	Doctoral degree	$1743

Problem 8: Translation

Observe how x (our independent variable) is *qualitative*: we need this to become numeric. Think of a way to "translate" the educational level into numbers. Write your idea down.

Answer to Problem 8.

Problem 9: Educational level as level

Let us treat the educational level as $x =$ level. We would then have 8 levels as in: 1 (less than high school), 2 (high school), 3 (some college), 4 (associate degree), 5 (bachelor's degree), 6 (master's degree), 7 (professional degree), 8 (doctoral degree). What is the regression line? What would be the expected salary of a person who quits a year after starting their Master's program (that would be, "some master's education")? [177]

Answer to Problem 9.

[177] You may assume that a Master's degree is typically two years long and takes place after obtaining a Bachelor's degree (so a year of Master's education would lie "between" a Bachelor's and a Master's degree).

Problem 10: Educational level as years of education

Let us treat the educational level as $x =$ years of education. We would then have 9 (less than high school), 12 (high school), 14 (some college), 15 (associate degree), 16 (bachelor's degree), 18 (master's degree), 20 (professional degree), 22 (doctoral degree). What is the regression line? What would be the expected salary of a person who quits a year after starting their Master's program (that would be, "some master's education")?

Answer to Problem 10.

31. Linear regression significance

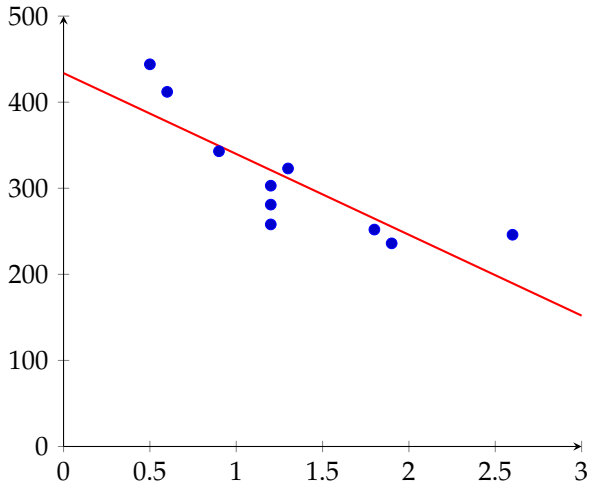
Every worksheet will work as follows.

1. You will be asked to form a group with other students in the class: you can make this as big or as small as you'd like, but groups of 4-5 work best.
2. Read through the worksheet, discussing any questions with the other members of your group.
   - You can call me at any time for help!
   - I will also be interrupting you for general guidance and announcements at random points during the class time.
3. Answer each question (preferably in the order provided) to the best of your knowledge.
4. While collaboration between students is highly encouraged and expected, each student has to submit their own version.
5. You will have 24 hours (see gradescope) to submit your work.

Activity 1: Fluoride levels and cavities

In the previous worksheet (from Lecture 30), you were asked to find the regression line for fluoride levels (x) and the resulting cavities (y). After doing all necessary calculations, we came up with the following line:

$$\hat{y} = 433.75 - 93.9 \cdot x.$$



Using the line, we must have also calculated all residuals ($\hat{y}_i - y_i$) as in the following table:

i	x_i	y_i	$\hat{y}_i$	$\hat{y}_i - y_i$
1	1.9	236	255.34	-19.34
2	2.6	246	189.61	56.39
3	1.8	252	264.73	-12.73
4	1.2	258	321.07	-63.07
5	1.2	281	321.07	-40.07
6	1.2	303	321.07	-18.07
7	1.3	323	311.68	11.32
8	0.9	343	349.24	-6.24
9	0.6	412	377.41	34.59
10	0.5	444	386.8	57.2

Problem 1: Calculating the SS_E

As we noted in this lecture, the sum of squares of error SS_E is an immensely useful quantity. Use the values in the last column of the table to calculate the sum of squares of error for the fluoride level regression.

Answer to Problem 1.

Problem 2: Calculating the noise variance

As has become increasingly clear, the noise plays a fundamental role on how well our regression will behave. The variance of the noise can be estimated as the mean square error, which in turn is based on the sum of squares of the error. What is the noise variance in this case?

Answer to Problem 2.

Problem 3: Significant regression?

Combine your answer in Problem 2 (where you got the estimator for the noise variance) with your calculation of $S_{xx} = \sum (x_i - \overline{x})^2$ [178] to decide whether the regression is significant or not using $\alpha = 5\%$.

[178] Remember that x is the fluoride level.

Answer to Problem 3.

How about for $\alpha = 0.2\%$? We do not need to recalculate everything, right? By the way, this can prove interesting for identifying P-values, even using the T distribution critical values!

Activity 2: Grades and effort

A professor is interested in whether their exams are unfair: they have come up with a plan to check that. They will ask students to note on their exams (for extra credit perhaps?) how many hours the students spent preparing for the exam. Here are the student answers and the grade they received in the exam.

Student	Hours of study	Grade
1	12	80
2	13	86
3	10	94
4	8	77
5	2	45
6	5	87
7	5	60
8	6	98
9	7	95
10	5	50

In any question that looks like this, our answer is always the same: hypothesis testing. In regression specifically, we will always follow the next steps.

1. Calculate the regression line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.
2. Use the regression line to estimate the residuals $y_i - \hat{y}_i$.
3. Calculate the sum of squares of error, SS_E, and the corresponding mean square error, $MS_E/(n-2)$.
4. Calculate S_{xx}.
5. Combine all steps to do a t-test with $n-2$ degrees of freedom. That is, calculate the test statistic of $T_0 = \frac{\hat{\beta}_1}{\sqrt{MS_E/S_{xx}}}$.
   - If we are able to reject, the regression is significant.
   - If we fail to reject, then we lack evidence to claim that the regression is significant.

Let's put this to the test in the data that have been provided to us.

Problem 4: Significant regression?

Using $\alpha = 5\%$, do you have enough evidence that your regression is significant?

Answer to Problem 4.

Problem 5: P-values

When studying the significance of a regression, P-values can help. The smaller the P-value, the more confident we become that the regression is significant. Like in hypothesis testing, if P-value $< \alpha$, then we may reject and claim that the regression is significant. What is the P-value here? How would you go about calculating it? [179]

[179] Maybe.. go through the critical values for the T distribution with $n-2$ degrees of freedom and find a value that is *close*?

Answer to Problem 5.

Activity 3: A "qualitative" regression

Continuing (again) from the last worksheet (Worksheet 30), we saw two ways to quantify the educational level for the purposes of a regression. First, we present the level with integer numbers $1, 2, 3, \ldots, 8$.

Level	Number	Income
Less than high school	1	\$520
High school	2	\$712
Some college	3	\$774
Associate degree	4	\$836
Bachelor's degree	5	\$1173
Master's degree	6	\$1401
Professional degree	7	\$1836
Doctoral degree	8	\$1743

The regression line can be found as

$$\hat{y} = 245.84 + 195.23 \cdot x.$$

Then, we present the level as the "number of years of education" as in $9, 12, 14, \ldots, 22$.

Level	Number	Income
Less than high school	9	\$520
High school	12	\$712
Some college	14	\$774
Associate degree	15	\$836
Bachelor's degree	16	\$1173
Master's degree	18	\$1401
Professional degree	20	\$1836
Doctoral degree	22	\$1743

The line is now becoming:

$$\hat{y} = -630.18 + 111.4 \cdot x.$$

Problem 6: Education level as a numeric level

Is the first regression significant or not?

Answer to Problem 6.

Problem 7: Education level as number of years in education

Is the second regression significant or not?

Answer to Problem 7.

Problem 8: Compare and contrast

Be very careful with this next one. Which one of the two lines appears to provide us with a **more significant** regression? Why? [180]

[180] Would you prefer a bigger or a smaller T statistic? Maybe in absolute value?

Answer to Problem 8.

Interesting! We may be able to use this comparison approach soon.

32. Multiple linear regression

Every worksheet will work as follows.

1. You will be asked to form a group with other students in the class: you can make this as big or as small as you'd like, but groups of 4-5 work best.
2. Read through the worksheet, discussing any questions with the other members of your group.
   - You can call me at any time for help!
   - I will also be interrupting you for general guidance and announcements at random points during the class time.
3. Answer each question (preferably in the order provided) to the best of your knowledge.
4. While collaboration between students is highly encouraged and expected, each student has to submit their own version.
5. You will have 24 hours (see gradescope) to submit your work.

Activity 1: The ANOVA identity and R^2

In this first set of problems, we will use the ANOVA identity and calculate R^2. As a reminder, the **analysis of variance** identity states that the **total variance** can be attributed to either the **regression** or the **error** (noise):

$$SS_T = SS_R + SS_E.$$

This is used in the calculation of R^2, which can be computer as

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}.$$

Now, let us return to the fluoride level dataset that we first saw in Worksheet 30.

i	x_i	y_i	$\hat{y}_i$	$\hat{y}_i - y_i$
1	1.9	236	255.34	-19.34
2	2.6	246	189.61	56.39
3	1.8	252	264.73	-12.73
4	1.2	258	321.07	-63.07
5	1.2	281	321.07	-40.07
6	1.2	303	321.07	-18.07
7	1.3	323	311.68	11.32
8	0.9	343	349.24	-6.24
9	0.6	412	377.41	34.59
10	0.5	444	386.8	57.2

Problem 1: Calculating the SS_T

Recall that the total sum of squares is needed for calculating the variance. It is defined as $SS_T = \sum (y_i - \overline{y})^2$. What is the total sum of squares in our dataset?

Answer to Problem 1.

Problem 2: Calculating the SS_R

During Worksheet 31, you must have calculated the sum of squares of the error as

$$SS_E = \sum (y_i - \hat{y}_i)^2 = 14261.26.$$

Use this fact as well as the ANOVA identity to calculate the sum of squares of the regression.

Answer to Problem 2.

Problem 3: R^2 calculation

Combine Problems 1 and 2 to answer the following: what is R^2?

Answer to Problem 3.

Problem 4: An F test for significance

During Lecture 32, we saw that an F test is a valid test when checking for regression significance. That said, when we only have one predictor variable (x), then we saw that a T test can also do the same thing. Are those two equivalent? Do an F test between the mean square of the regression and the mean square of the error and report whether you'd reject (that is, find the regression to be significant) or not.

Answer to Problem 4.

Interesting! We–again–rejected, like we did in Worksheet 31. This is not unexpected. But what is truly fascinating is that we get *exactly* the same P-values! In Worksheet 31, you must have gotten that $T_0 = -4.23$, which translates to a P-value of 0.003. The F test statistic you got here (equal to 17.89) leads to the same P-value of 0.003 also!

Overall: doing a proper F test or a proper T test on a **simple linear regression model** will lead to exactly the same conclusion. This is not true in the case of multiple linear regression, where the F test provides us information about the regression as a whole ("are all predictors insignificant or is there even one significant among them") and the T test provides us information on each individual predictor ("is this specific predictor insignificant or not").

Activity 2: Predicting the yield

We have collected the following data on $n = 16$ observations for two factors (x_1, x_2) and their effects on the yield (y) of a crop.

Factor 1 (x_1)	Factor 2 (x_2)	Yield (y)
42	29	251
43	29	251
44	30	248
45	30	268
47	30	273
48	30	277
50	31	270
53	31	285
53	31	290
57	32	297
57	32	303
64	32	305
65	32	309
71	32	322
77	33	331
78	33	349

The line obtained from multiple linear regression is

$$\hat{y} = 157.09 + 2.46 \cdot x_1 - 0.18 \cdot x_2.$$

On your way to find the least squares line, you would have built

$$X = \begin{bmatrix} 1 & 42 & 29 \\ 1 & 43 & 29 \\ 1 & 44 & 30 \\ \vdots & \vdots & \vdots \\ 1 & 78 & 33 \end{bmatrix}.$$

You are also given that

$$\left(X^T X\right)^{-1} = \begin{bmatrix} 1.9298 & -0.0203 & -0.0249 \\ -0.0203 & 0.0006 & -0.0004 \\ -0.0249 & -0.0004 & 0.0016 \end{bmatrix}.$$

Problem 5: Multiple linear regression significance

After some calculations, we found that $\sum (y_i - \hat{y}_i)^2 = 708$ and $\sum (\hat{y}_i - \overline{y})^2 = 12447$. Is the regression significant or not? Use $\alpha = 5\%$.

Answer to Problem 5.

So, we got a significant regression in our hands. But are both factors significant?

Problem 6: More significant

Which of the two factors (x_1 or x_2) is more significant? Why?

Answer to Problem 6.

Activity 3: A full case study

Problem 7: Wildfire prediction

Forest fires and wildfires have been a major problem in many parts around the world. Unfortunately, this is an issue that is exacerbated over the last few years. With everything that has happened in the last two years, it is difficult to remember that in the beginning of 2020, Australia experienced some of the worst wildfires (see a BBC article here) or Greece (where I am from) experienced devastating wildfires in 2021 (see a pictorial from The Guardian here).

In this worksheet, we will try to use **multiple linear regression** to predict the level of severity of a fire depending on the underlying conditions. Among all conditions, we picked the following 4:

- The Duff Moisture Code (DMC).
- The Drought Code (DC).
- The Relative Humidity (RH).
- The wind speed (W).

As our dependent variable, we consider y to be the area burned. Out of a huge dataset, we isolated 10 fires to study.

	DMC	DC	RH	W	Area (in hectares)
1	95	670	26	3.1	64.1
2	83	530	43	4	71.3
3	130	720	21	4.5	88.49
4	150	730	27	3.1	95.18
5	130	700	43	2.7	103.39
6	70	670	36	3.1	105.66
7	120	670	25	3.1	154.88
8	140	600	41	5.8	196.48
9	120	650	46	4.5	200.94
10	150	730	40	4.6	212.88

A person you are working with has let you know that DMC and DC are supposed to give similar indications, so not both of them are necessary. **Between the two multiple linear regression models (one with DMC, RH, W and one with DC, RH, W), which one of the two behaves best and why? Compare them based on their F tests, as well as their R^2 and R^2_{adj} coefficients.**

Before you start solving, here is a small roadmap of what you should do. This will hopefully guide you in all multiple linear regression you try to build.

1. Build matrix X.
2. Calculate matrix $\left(X^T X\right)^{-1}$ and vector $X^T y$.
3. Calculate the coefficients as $\hat{\beta} = \left(X^T X\right)^{-1} X^T y$.
4. Use the line to calculate SS_E.
5. Use the data to calculate SS_T.
6. Use ANOVA to calculate SS_R.
7. Estimate the F_0 test statistic as MS_R / MS_E and reject/fail to reject.
8. Estimate the individual T_0 test statistics for each coefficient and reject/fail to reject.
9. Calculate R^2 and R^2_{adj}.

This may take a while [181]. That said, you have probably noticed by now how useful regression can be, so hopefully you will enjoy comparing these two models and helping us solve a real-life, large-scale societal issue.

[181] Do not hesitate to call me in your group if you'd like to see how you could use Excel or Python to do these calculations!

Answer to Problem 7.

Answer to Problem 7 (cont'd).

33. Regression extensions and model building

Every worksheet will work as follows.

1. You will be asked to form a group with other students in the class: you can make this as big or as small as you'd like, but groups of 4-5 work best.
2. Read through the worksheet, discussing any questions with the other members of your group.
   - You can call me at any time for help!
   - I will also be interrupting you for general guidance and announcements at random points during the class time.
3. Answer each question (preferably in the order provided) to the best of your knowledge.
4. While collaboration between students is highly encouraged and expected, each student has to submit their own version.
5. You will have 24 hours (see gradescope) to submit your work.

Activity 1: Respiratory function

In this first set of problems, we tackle linear regression before finally seeing how a quadratic version works.

The following table contains information about respiratory function (as measured by forced expiratory volume) and smoking. The dataset contains information about age (x_1, note that all subjects are between 13 and 19 years old), height (x_2), and whether they smoke (1) or not (0) (x_3). The last column called FEV measures forced expiratory volume and is our dependent variable (y).

ID	Age (x_1)	Height (x_2)	Smoking (x_3)	FEV (y)
1	13	67	1	3.994
2	13	61	0	3.208
3	14	64.5	0	2.997
4	14	72.5	1	4.271
5	16	72	1	4.872
6	16	63	0	2.795
7	19	72	1	5.102
8	19	66	0	3.519
9	18	60	0	2.853
10	17	70.5	1	4.724
11	16	69.5	1	4.070

Problem 1: Simple linear regression

First perform a linear regression between height (x_2) and FEV (y). What is the adjusted R^2 score?

Answer to Problem 1.

Problem 2: Simple quadratic regression

Does the adjusted R^2 score improve if we perform a regression on height *squared* (x_2^2) and FEV (y)? [182]

[182] Recall what we saw in the notes: create a new column with **only** x_2^2 values and use that one!

Answer to Problem 2.

Problem 3: A full quadratic regression

Finally, do a regression between age, height squared, smoking (x_1, x_2^2, x_3) and FEV (y). What is R^2_{adj} now?

Answer to Problem 3.

Activity 2: Blood pressure

In this activity, we check another dataset, one on blood pressure. We have collected the following data:

1. age (in years);
2. time spent in an urban environment (in years).

We want to relate those two factors to the systolic blood pressure. The data is presented in the following table.

ID	Age (x_1)	Years in urban area (x_2)	Systolic pressure (y)
1	22	6	120
2	24	5 5	125
3	28	5	120
4	33	10	114
5	34	15	130
6	35	18	118
7	41	32	128
8	47	1	116
9	50	43	132
10	54	40	152

Problem 4: A ratio regression

Let's not spend time doing a multiple factor regression again. Instead, assume you want to find a regression between **the ratio of years in urban area and age**. That is, you want a regression $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \frac{x_2}{x_1}$. What is the regression line? [183]

[183] Use simple linear regression!

Answer to Problem 4.

Activity 3: Model selection

We have collected $n = 16$ house sales in an urban setting. We are interested in what increases the price of a house per square meter, so we tried to keep track of some sale details. More specifically, we collect:

1. the age of the house (Age, in years);
2. the distance to the closest subway station (Distance, in meters);
3. the number of convenience and grocery stores in walking distance (Stores);
4. the latitude (Latitude);
5. and the longitude (Longitude).

The goal is to use some or all of these factors to predict y, the price per square meter. Here is the table of the data:

#	Age	Distance	Stores	Latitude	Longitude	Price
1	14.7	1717.19	2	24.96	121.52	23
2	12.7	170.13	1	24.97	121.53	37.3
3	26.8	482.76	5	24.97	121.54	35.5
4	7.6	2175.03	3	24.96	121.51	27.7
5	12.7	187.48	1	24.97	121.53	28.5
6	30.9	161.94	9	24.98	121.54	39.7
7	16.4	289.32	5	24.98	121.54	41.2
8	23	130.99	6	24.96	121.54	37.2
9	1.9	372.14	7	24.97	121.54	40.5
10	5.2	2408.99	0	24.96	121.56	22.3
11	18.5	2175.74	3	24.96	121.51	28.1
12	13.7	4082.02	0	24.94	121.5	15.4
13	5.6	90.46	9	24.97	121.54	50
14	18.8	390.97	7	24.98	121.54	40.6
15	8.1	104.81	5	24.97	121.54	52.5
16	6.5	90.46	9	24.97	121.54	63.9

Problem 5: All subsets selection

How many different subsets should you consider before declaring the absolute best combination of factors? [184]

[184] There are 5 factors – how many subsets can we create with 5 factors?

Answer to Problem 5.

Problem 6: Comparing subsets

Clearly enumerating all of the previous subsets is a computationally expensive task. However, comparing two or three of them is much easier. Say we agree that Age, Distance, and Stores are the three most important factors, which one of the three models is a more accurate predictor of price?

1. Age and Distance.
2. Age and Stores.
3. Distance and Stores.

Answer to Problem 6.

Problem 7: Backwards selection

Starting from a full model, do at least one iteration of the backwards selection heuristic. Which one is the least significant factor (that is, which one is the first factor to be removed)?

Answer to Problem 7.

Problem 8: Forwards selection

Starting from an empty model, do at least two iterations of the forwards selection heuristic. Which one is the most significant factor (that is, which one is the first factor to be added to the regression)? Which one is the second one to be added to the regression?

Answer to Problem 8.