HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY
SCHOOL OF APPLIED MATHEMATICS AND INFORMATICS
DEPARTMENT OF APPLIED MATHEMATICS

MI2020E

ALSO MI2026, MI2036, ETC.

STATISTICS AND PROBABILITY

THEORY SUMMARY AND ANSWER OF EXERCISES

SUMMARIZED AND EDITED BY TA NGOC MINH
BSc. student in Data Science & AI @ HUST-SOICT

Hanoi - 2022

Contents

Preface

To all of my friends who are using this document,

This document is written to summarize all the theories and give a recommended solution for the exercise of the subject Statistics and Probability, a course in the curriculum of students of ELiTECH Program at Hanoi University of Science and Technology. Although I studied the course with code MI2020E, this document is also applicable to students who study MI2026, MI2036, and so on. I hope you all use it as a reference and support while studying, and do not use it as a tool for not doing your homework.

To write this document, I would like to give many thanks to professor Nguyen Van Hanh for his instructions, professor Nguyen Thi Thu Thuy, for her course materials, and many other professors at the School of Applied Mathematics and Informatics, Hanoi University of Science and Technology about their public documents about Statistics and Probability. Thanks to the authors of *Probability & Statistics for Engineers and Scientists*, Walpole R.E, Myers R.H, Myers S.L, Ye K. and *The course of Probability and Statistics*, professor Tong Dinh Quy for their books as reference. Moreover, I also want to say thanks to several universities in the world such as Indiana University–Purdue University Indianapolis, Massachusetts Institute of Technology, the University of Toronto, etc. because of their helpful public documents.

Because this document is written in such a short time and all the solutions to mathematical problems are made by me, it is so hard to make sure it is completely correct. I would love to receive all of your contributions to my document via **minh.tn214918@sis.hust.edu.vn**.

Thank you for using this document.

Ngoc-Minh Ta

Chapter 1

Probability.

1.1 Theory.

1.1.1 Basic Notion.

- An **experiment** is the process by which an observation (or measurement) is obtained.
- Sample space
  - An **outcome** of an experiment is any possible observation of that experiment.
  - The **sample space** of an experiment is the set of all possible outcomes for an experiment, denoted by S.
- Events
  - An **event** is a set of outcomes of an experiment (or a subset of a sample space).
  - A **simple event** is an event that consists of exactly one outcome.
- Event Relations
  - The **union** of events A and B, denoted by $A \cup B$ (or $A + B$) is the event that either A or B or both occur.
  - The **intersection** of events A and B, denoted by $A \cap B$ (or AB), is the event that both A and B occur.
  - Two events, A and B, are **mutually exclusive/disjoint** if, when one event occurs, the others cannot, and vice versa. That is, $A \cap B = AB = \emptyset$.
  - A collection of events $A_1, A_2, ..., A_n$ is **mutually exclusive** if and only if $A_i \cap A_j = \emptyset,\ i \neq j$.
  - A collection of events $A_1, A_2, ..., A_n$ is **collectively exhaustive** if and only if $A_1 \cup A_2 \cup ... \cup A_n = S$.
  - The **complement** of event A is the set of all outcomes in the sample space that are not included in event A. Denoted by A^c or A'.
  - An event space is a collectively exhaustive, mutually exclusive set of events.
- Counting sample points
  - **Multiplication Rule:** If an operation can be performed in n_1 ways, and if for each of these a second operation can be performed in n_2 ways, and for each of the first two a third operation can be performed in n_3 ways, and so forth, then the sequence of k operations can be performed in $n_1 n_2 ... n_k$ ways.
  - **Permutation:** number of permutations of n distinct objects taken k at a time is

$$A_n^k = n \times (n-1) \times (n-2) \times ... \times (n-k+1) = \frac{n!}{(n-k)!}$$

> **Notice**
> Permutation of n objects taken k at a time: **ORDER, NO REPEAT.**

- The number of n-permutations of n distinguishable objects is $P_n = A_n^n = n!$.
- The number of distinct **combinations** of n distinct objects that can be formed, taking them k at a time, is

$$C_n^k = \frac{n!}{k! \times (n-k)!}$$

> **Notice**
> Combination of n objects taken k at a time: **NO REPEAT, NO ORDER.**

- **Sampling with Replacement**: Given n distinguishable objects, there are $\overline{A}_n^k = n^k$ ways to choose with replacement an ordered sample of k objects.

1.1.2 Probability of an event.

- The probability $P[A]$ of event A is a measure of our belief that A will occur.
- **Theoretical probability (Classical approach):** If an experiment has n possible equally likely outcomes, this method would assign a probability of $\frac{1}{n}$ to each outcome. Then if an event A contains exactly m outcomes, the probability of event A is

$$P[A] = \frac{m}{n}$$

- In general, the probability P(A) of event A is the sum of the probabilities assigned to the outcomes (simple events) contained in A:

$$P[A] = \sum_{O_i \in A} P[O_i]$$

- **Empirical Probability (Relative frequency):** assigning probabilities based on experimentation or historical data. If an experiment is performed n times, then the relative frequency of a particular occurrence say, A is

$$\text{relative frequency} = \frac{\text{frequency}}{n}$$

where the frequency is the number of times the event A occurred. Then the relative frequency of the event A is defined as the probability of event A, that is

$$P[A] = \lim_{n \in +\infty} \frac{\text{frequency}}{n}$$

1.1.3 Probability rule.

- Complement rule:

$$P[\overline{A}] = 1 - P[A]$$

- Addition rule:
  - $P[A \cup B] = P[A] + P[B] - P[A \cap B]$.

- If A and B are mutually exclusive (i.e, $A \cap B = \emptyset$, $P[A \cup B] = P[A] + P[B]$.
- In general,

$$P\left[\bigcup_{i=1}^{n} A_i\right] = \sum_{i=1}^{n} P[A_i] - \sum_{i<j} P[A_iA_j] + \sum_{i<j<k} P[A_iA_jA_k] - ... + (-1)^n P[A_iA_j...A_n]$$

- **Conditional probability rule:** Conditional probability of event A given event B, denoted by $P[A|B]$, is the probability of event A given that the event B has occurred. The conditional probability formula is:

$$P[A|B] = \frac{P[AB]}{P[B]}$$

- The multiplication rule:
  - $P[AB] = P[A] \times P[B|A] = P[B] \times P[A|B]$.
  - In general, $P[A_1A_2...A_n] = P[A_1]P[A_2|A_1]P[A_3|A_1A_2]....$
  - A and B are independent if $P[A|B] = P[A]$ or $P[B|A] = P[B]$ or $P[AB] = P[A]P[B]$.

1.1.4 Bayes' rule.

- **The total probability:** For an event space $\{A_1, A_2, ..., A_n\}$ for all i and an event A, the probability of the event A can be expressed as:

$$P[A] = P[AA_1] + P[AA_2] + ... + P[AA_n] = P[A_1]P[A|A_1] + ... + P[A_n]P[A|A_n]$$

- **Bayes' rule:** For an event space $\{A_1, A_2, ..., A_n\}$ for all i and an event A, posterior probability of the event A_i given A can be expressed as

$$P[A_i|A] = \frac{P[A_i] \times P[A|A_i]}{P[A]} = \frac{P[A_i] \times P[A|A_i]}{P[A_1]P[A|A_1] + ... + P[A_n]P[A|A_n]}$$

- A **Bernoulli trial** (or **binomial trial**) is a random experiment with exactly two possible outcomes, "success" and "failure", in which the probability of success is the same (equals p) every time. The probability of k successes and $n-k$ failures in n Bernoulli trials is: $P_n(k) = C_n^k \times p^k \times (1-p)^{n-k}$.

1.2 Answer of Exercises.

1.2.1 Experiments.

Problem 1.1.
A fax transmission can take place at any of three speeds depending on the condition of the phone connection between the two fax machines. The speeds are high (h) at 14400b/s, medium (m) at 9600b/s, and low (l) at 4800b/s. In response to requests for information, a company sends either short faxes of two (t) pages, or long faxes of four (f) pages. Consider the exp. of monitoring a fax trans and observing the trans speed and length. An observation is a two-letter word, e.g., a high-speed, two-page fax is ht.

(a) What is the sample space of the experiment?
(b) Let A_1 be the event "medium-speed fax." What are the outcomes in A_1?
(c) Let A_2 be the event "short fax." What are the outcomes in A_2?
(d) Let A_3 be the event "high-speed fax or low-speed fax." What are the outcomes in A_3?
(e) Are A_1, A_2, and A_3 mutually exclusive?
(f) Are A_1, A_2, and A_3 collectively exhaustive?

Solution

(a) $S = \{ht, mt, lt, hf, mf, lf\}$.
(b) $A_1 = \{mt, mf\}$.
(c) $A_2 = \{ht, mt, lt\}$.
(d) $A_3 = \{ht, hf, lt, lf\}$.
(e) No, because $A_1A_2 = \{mt\}, ...$
(f) Yes, because $A_1 \cup A_2 \cup A_3 = S$.

Problem 1.2.
An integrated circuit factory has 3 machines X, Y, and Z. Test 1 integrated circuit produced by each machine. Either a circuit is acceptable (a) or fails (f). An observation is a sequence of 3 test results corr. to the circuits from machines X, Y, and Z, respectively. E.g., aaf is the observation that the circuits from X and Y pass and the circuit from Z fails.

(a) What are the elements of the sample space of this experiment?
(b) What are the elements of the sets Z_F = {circuit from Z fails}, X_A = {circuit from X is acceptable}.
(c) Are Z_F and X_A mutually exclusive?
(d) Are Z_F and X_A collectively exhaustive?
(e) What are the elements of the sets C = {more than one circuit acceptable}, D = {at least two circuits fail}.
(f) Are C and D mutually exclusive / collectively exhaustive?

Solution

(a) $S = \{aaa, aaf, afa, faa, aff, faf, ffa, fff\}$.
(b) $Z_F = \{aaf, aff, faf, fff\}$, $X_A = \{aaa, aaf, afa, aff\}$.

(c,d) No.
(e) $C = \{aaf, afa, faa, aaa\}$, $D = \{aff, faf, ffa, fff\}$
(f) C and D are both mutually exclusive and collectively exhaustive.

Problem 1.3.
Find out the birthday (month and day but not year) of a randomly chosen person. What is the sample space of the experiment. How many outcomes are in the event that the person is born in July?

Solution

- $S = \{1/1, 1/2, ..., 31/12\}$.
- Since July has 31 days so there are 31 outcomes that the person is born in July.

Problem 1.4.
Let the sample space of the experiment consist of the measured resistances of two resistors. Give four examples of event spaces.

1.2.2 Counting Methods.

Problem 1.5.
Consider a binary code with 5 bits (0 or 1) in each code word. An example of a code word is 01010. How many different code words are there? How many code words have exactly three 0's?

Solution

- There are 2 choices of each bit. So, there are $2^5 = 32$ different code words.
- Take three 0s, we can choose 3 out of 5 positions to put them in, so there are $C_5^3 = 10$ choices. The last 2 positions are left is for 1, so there is only 1 choice for putting two 1s in these position.
  So, there are 10 code words that satisfy the question.

Problem 1.6.
Consider a language containing four letters: A, B, C, D. How many three-letter words can you form in this language? How many four-letter words can you form if each letter appears only once in each word?

Solution

- There are 4 choices of each letter, so, there are $4^3 = 64$ three-letter words can be formed.
- There are 4 choices to choose first letter, 3 choices to choose second letter (except for the first one),... So, there are $4! = 24$ four-letter words can be formed.

Problem 1.7.

On an American League baseball team with 15 field players and 10 pitchers, the manager must select for the starting lineup, 8 field players, 1 pitcher, and 1 designated hitter. A starting lineup specifies the players for these positions and the positions in a batting order for the 8 field players and designated hitter. If the designated hitter must be chosen among all the field players, how many possible starting lineups are there?

Solution

- Ways of choosing 9 players from 15 field players (with orders): $A_1^9 5$.
- Way of choosing 1 players from 10 pitchers is $C_1^1 0 = 10$.
- So, the number of possible starting lineups is: $N = A_1^9 5 \times 10 = 18, 162, 144, 000$.

Problem 1.8.

A basketball team has three pure centers, four pure forwards, four pure guards, and one swingman who can play either guard or forward. A pure position player can play only the designated position. If the coach must start a lineup with one center, two forwards, and two guards, how many possible lineups can the coach choose?

Solution

- If the swingman plays guard: $N_1 = C_3^1 \times C_4^2 \times C_4^1 = 72$.
- If the swingman plays forward: $N_1 = C_3^1 \times C_4^1 \times C_4^2 = 72$.
- If the swingman doesn't play: $N_1 = C_3^1 \times C_4^2 \times C_4^2 = 108$.
- So, there are $72 + 72 + 108 = 252$ possible lineups can be chosen.

1.2.3 Probability.

Problem 1.9.

In a certain city, three newspapers A, B, and C are published. Suppose that 60 percent of the families in the city subscribe to newspaper A, 40 percent of the families subscribe to newspaper B, and 30 percent of the families subscribe to newspaper C. Suppose also that 20 percent of the families subscribe to both A and B, 10 percent subscribe to both A and C, 20 percent subscribe to both B and C, and 5 percent subscribe to all three newspaper A, B, and C. What percentage of the families in the city subscribe to at least one of the three newspapers?

Solution

$$\begin{aligned} P[\text{at least one}] &= P[A \cup B \cup C] = P[A] + P[B] + P[C] - P[AB] - P[BC] - P[CA] + P[ABC] \\ &= 0.6 + 0.4 + 0.3 - 0.2 - 0.1 - 0.2 + 0.05 = 0.85 = 85\% \end{aligned}$$

Problem 1.10.

From a group of 3 freshmen, 4 sophomores, 4 juniors and 3 seniors a committee of size 4 is randomly selected. Find the probability that the committee will consist of

(a) 1 from each class;
(b) 2 sophomores and 2 juniors;
(c) Only sophomores and juniors.

Solution

(a) $P[A] = \dfrac{C_3^1 \times C_4^1 \times C_4^1 \times C_3^1}{C_{14}^4} = \dfrac{144}{1001}$.

(b) $P[B] = \dfrac{C_4^2 \times C_4^2}{C_{14}^4} = \dfrac{36}{1001}$.

(c) $P[C] = \dfrac{C_8^4 - C_4^4}{C_{14}^4} = \dfrac{69}{1001}$.

Problem 1.11.

A box contains 24 light bulbs of which four are defective. If one person selects 10 bulbs from the box in a random manner, and a second person then takes the remaining 14 bulbs, what is the probability that all 4 defective bulbs will be obtained by the same person?

Solution

$$P[X=4] = \frac{C_{20}^6 + C^1 0_{20}}{C_{24}^{10}} = \frac{173}{1518}.$$

Problem 1.12.

Suppose that three runners from team A and three runners from team B participate in a race. If all six runners have equal ability and there are no ties, what is the probability that three runners from team A will finish first, second, and third, and three runners from team B will finish fourth, fifth, and sixth?

Solution

- The number of order can be happened is $6! = 720$.
- The number of order of team A is 1st, 2nd, 3rd and team B is 4th, 5th, 6th is $3! \times 3! = 36$.

  Therefore, the probability $P[X] = \dfrac{36}{720} = \dfrac{1}{20}$.

Problem 1.13.

Suppose that a school band contains 10 students from the freshman class, 20 students from the sophomore class, 30 students from the junior class, and 40 students from the senior class. If 15 students are selected at random from the band, what is the probability that at least one students from each of the four classes?

Solution

- Denote A_i be the probability of no student in ith year student is chosen.

$$\begin{aligned}P[\overline{X}] = &P[A_1]+P[A_2]+P[A_3]+P[A_4]-P[A_1A_2]-P[A_1A_3]-P[A_1A_4]-P[A_2A_3]\\ &-P[A_2A_4]-P[A_3A_4]+P[A_1A_2A_3]+P[A_1A_2A_4]+P[A_1A_3A_4]+P[A_2A_3A_4]\\ &-P[A_1A_2A_3A_4]\\ =&\frac{C_{90}^{15}+C_{80}^{15}+C_{70}^{15}+C_{60}^{15}}{C_{100}^{15}}-\frac{C_{70}^{15}+C_{60}^{15}+C_{50}^{15}+C_{50}^{15}+C_{40}^{15}+C_{30}^{15}}{C_{100}^{15}}+\frac{C_{40}^{15}+C_{30}^{15}+C_{20}^{15}}{C_{100}^{15}}\\ &(\text{Since } P[A_1A_2A_3A_4]=0)\end{aligned}$$

- So, $P[X] = 1 - P[\overline{X}] =$

Problem 1.14.
Suppose that 10 cards, of which 5 are red and 5 are green, are placed at random in 10 envelopes, of which 5 are red and 5 are green. Determine the probability that exactly x envelopes will contain a card with a matching color ($x = 0, 1, 2, ..., 10$).

Solution

- Considering choose 5 envelopes at random into which the 5 red letters will be placed. If there are exactly r red envelopes among the five selected envelopes ($r = 0, 1, ..., 5$), then exactly $x = 2r$ envelopes will contain a card with a matching color.
- Hence, the only possible value of x are $0, 2, 4, ..., 10$. Thus, for $x = 0, 2, ..., 10$ and $r = \frac{x}{2}$, the desired probability is the probability that there are exactly r red envelopes among the five selected envelopes, which is $\frac{C_5^r \times C_5^{5-r}}{C_{10}^5}$.

Problem 1.15.
Consider two events A and B with $P(A) = 0.4$ and $P(B) = 0.7$. Determine the maximum and minimum possible values of $P(A \cap B)$ and the conditions under which each of these values is attained.

Solution

- Since $P(AB) \leq P(A) and P(AB) \leq P(B) \Rightarrow \max P(AB) = P(A) = 0.4$.
- Since $1 \geq P(A \cup B) = P(A) + P(B) - P(AB) = 1.1 - P(AB) \Rightarrow \min P(AB) = 0.1$.

Problem 1.16.
Suppose that four guests check their hats when they arrive at a restaurant, and that these hats are returned to them in a random order when they leave. Determine the probability that no guest will receive the proper hat.

Solution

- The random experiment "4 hats are returned to 4 guests" $\Rightarrow |S| = 4!$.

- Let A be the event "No guest ...", so A^c is "At least one ..."; A_i is "the i^th guest receive his proper hat".
- We have $A^c = A_1 \cup A_2 \cup A_3 \cup A_4$

$$\begin{aligned} \Rightarrow P[A^c] &= \sum_{i=1}^{4} P[A_i] - \sum_{i<j} P[A_i A_j] + \sum_{i<j<k} P[A_i A_j A_k] - P[A_1 A_2 A_3 A_4] \\ &= 4P[A_1] - 6P[A_1 A_2] + 4P[A_1 A_2 A_3] - P[A_1 A_2 A_3 A_4] \end{aligned}$$

- Since $P[A_1 A_2 A_3 A_4] = \dfrac{1}{4!} \Rightarrow P[A_1] = \dfrac{3!}{4!}, P[A_1 A_2] = \dfrac{2!}{4!}, P[A_1 A_2 A_3 A_4] = \dfrac{1}{4!}$
  $\Rightarrow P[A^c] = \dfrac{5}{8} \Rightarrow P[A] = 1 - P[A^c] = \dfrac{3}{8}$.

Problem 1.17.

Suppose that four guests check their hats when they arrive at a restaurant, and that these hats are returned to them in a random order when they leave. Determine the probability that at least 2 guests will receive the proper hat.

Solution

- The random experiment "4 hats are returned to 4 guests" $\Rightarrow |S| = 4!$.
- There is 1 ways for 4 hats are returned properly, and the event "Only 3 hats are returned properly" cannot be happened.
- Consider returning exact 2 proper hats. There are $C_4^2 = 6$ ways of choosing two hats to return properly. There is only 1 way of returning the last 2 hats have to be returned improperly to 2 guests.
- So, there are total $6 + 1 = 7$ ways of returning the hats that satisfy the demand.
- The probability need calculating is: $P = \dfrac{7}{24}$.

Problem 1.18.

Suppose that A, B and C are three independent events such that $P(A) = \frac{1}{4}, P(B) = \frac{1}{3}$ and $P(C) = \frac{1}{2}$.

(a) What is the probability that none of these three events will occur?
(b) Determine the probability that exactly one of these three events will occur.

Solution

(a) $P[\overline{A}\overline{B}\overline{C}] = P[\overline{A}].P[\overline{B}].P[\overline{C}] = \dfrac{1}{4}$.
(b) $P[\text{one}] = P[A\overline{B}\overline{C}] + P[\overline{A}B\overline{C}] + P[\overline{A}\overline{B}C] = \dfrac{1}{12} + \dfrac{1}{8} + \dfrac{1}{4} = \dfrac{11}{24}$.

Problem 1.19.

Three players A, B and C take turns tossing a fair coin. Suppose that A tosses the coin first, B tosses the second and C tosses third and cycle is repeated indefinitely until someone wins by being the first player to obtain a head. Determine the probability that each of the three players will win.

Solution

- If all loss in one cycle, the game will continue, and the probability is $\left(\frac{1}{2}\right)^3 = \frac{1}{8}$.

$$P[A] = \sum_{n=0}^{+\infty} \frac{1}{2} \times \left(\frac{1}{8}\right)^n = \frac{4}{7}$$

$$P[B] = \sum_{n=0}^{+\infty} \frac{1}{2} \times \frac{1}{2} \times \left(\frac{1}{8}\right)^n = \frac{2}{7}$$

$$P[C] = 1 - P[A] - P[B] = \frac{1}{8}$$

Problem 1.20.

Computer programs are classified by the length of the source code and by the execution time. Programs with ≥ 150 lines in the source code are big (B). Programs with ≤ 150 lines are little (L). Fast programs (F) run in less than 0.1s. Slow programs (W) require at least 0.1s. Monitor a program executed by a computer. Observe the length of the source code and the run time. The probability model for this experiment contains the following information: $P[LF] = 0.5, P[BF] = 0.2, P[BW] = 0.2$. What is the sample space of the experiment? Calculate the following probabilities:

(a) $P[W]$
(b) $P[B]$;
(c) $P[W \cup B]$.

Solution

(a) The sample space of the experiment is $S = LF, BF, LW, BW$. From the problem statement, we know that $P[LF] = 0.5, P[BF] = 0.2$ and $P[BW] = 0.2$.
This implies $P[LW] = 1 - 0.5 - 0.2 - 0.2 = 0.1$ (because BF, BW, and LF are mutually exclusive). Since $LW \cap BW = \emptyset \Rightarrow P[W] = P[LW] + P[BW] = 0.1 + 0.2 = 0.3$.
(b) $P[B] = P[BF] + P[BW] = 0.2 + 0.2 = 0.4$.
(c) $P[W \cup B] = P[W] + P[B] - P[BW] = 0.3 + 0.4 - 0.2 = 0.5$.

Problem 1.21.

You have a six-sided die that you roll once and observe the number of dots facing upwards. What is the sample space? What is the probability of each sample outcome?What is the probability of E, the event that the roll is even?

Solution

- Let s_i denote the outcome that the down face has i dots. The sample space is $S = s_1, ..., s_6$. The probability of each sample outcome is $P[s_i] = \frac{1}{6}$.
- The probability of the event E that the roll is even is

$$P[E] = P[s_2] + P[s_4] + P[s_6] = \frac{3}{6}.$$

Problem 1.22.
A student's score on a 10-point quiz is equally likely to be any integer between 0 and 10. What is the probability of an A, which requires the student to get a score of 9 or more? What is the probability the student gets an F by getting less than 4?

Solution

- Let s_i equal the outcome of the student's quiz. The sample space is then composed of all the possible grades that she can receive.

$$S = 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10$$

- Since each of the 11 possible outcomes is equally likely, the probability of receiving a grade of i, for each $i = 0, 1, ..., 10$ is $P[s_i] = \frac{1}{11}$. The probability that the student gets an A is the probability that she gets a score of 9 or higher. That is

$$P[\text{Grade of A}] = P[9] + P[10] = \frac{1}{11} + \frac{1}{11} = \frac{2}{11}$$

- The probability of failing requires the student to get a grade less than 4.

$$P[\text{Failing}] = P[3] + P[2] + P[1] + P[0] = \frac{1}{11} + \frac{1}{11} + \frac{1}{11} + \frac{1}{11} = \frac{4}{11}$$

Problem 1.23.
Mobile telephones perform hand-offs as they move from cell to cell. During a call, a telephone either performs zero hand-offs (H0), one hand-off (H1), or more than one hand-off (H2). In addition, each call is either long (L), if it lasts more than three minutes, or brief (B). The following table describes the probabilities of the possible types of calls.

	H_0	H_1	H_2
L	0.1	0.1	0.2
B	0.4	0.1	0.1

What is the probability $P[H_0]$ that a phone makes no hand-offs? What is the probability a call is brief? What is the probability a call is long or there are at least two hand-offs?

Solution

- From the table we look to add all the disjoint events that contain H0 to express the probability that a caller makes no hand-offs as

$$P[H_0] = P[LH_0] + P[BH_0] = 0.1 + 0.4 = 0.5.$$

- In a similar fashion we can express the probability that a call is brief by

$$P[B] = P[BH_0] + P[BH_1] + P[BH_2] = 0.4 + 0.1 + 0.1 = 0.6$$

- The probability that a call is long or makes at least two hand-offs is

$$P[L \cup H_2] = P[LH_0] + P[LH_1] + P[LH_2] + P[BH_2] = 0.1 + 0.1 + 0.2 + 0.1 = 0.5.$$

Problem 1.24.

Proving the following facts: (a) $P[A \cup B] \geq P[A]$; (b) $P[A \cup B] \geq P[B]$; (c) $P[A \cap B] \leq P[A]$; (d) $P[A \cap B] \leq P[B]$.

Solution

(a) Since $P[A \cup B] = P[A] + P[B] - P[AB] = P[A] + P[B] - P[B].P[A|B] \geq P[A] + P[B] - P[B] = P[A]$.
(b) Similar to (a).
(c) Since $P[A \cap B] = P[A].P[B|A]$. Since $P[B|A] \leq 1$, we have $P[A \cap B] \leq P[A]$.
(d) Similar to (c).

Problem 1.25.

Proving by induction the union bound: For any collection of events $A_1, ..., A_n$,

$$P[A_1 \cup A_2 \cup ... \cup A_n] \leq \sum_{i=1}^{n} P[A_i]$$

Solution

- We have the function with $n = 2$ is true: $P[A \cup B] \leq P[A] + P[B]$
- Also the $n = 3$: $P[A \cup B \cup C] \leq P[A \cup B] + P[C] \leq P[A] + P[B + P[C]]$
- We assume that the function is true with $n = k$. Let prove it is true with $n = k + 1$. Indeed,

$$P[A_1 \cap A_2 \cap ... \cap A_{k+1}] \leq P[A_1 \cap A_2 \cap ... \cap A_k] + P[A_{k+1}] \leq \sum_{i=1}^{k+1} P[A_i]$$

- By the induction, we have the function is true for all n.

Problem 1.26.

Proving $P[\varnothing] = 0$.

Solution

- Choose $A = S$ and $B_i = \varnothing$. We have $A \cup B_i = S \ \forall i$ and A, B_i are mutually exclusive.
- So, $P[A \cup B_i] = 1 = P[S] + P[\varnothing] + ... + P[\varnothing] \Rightarrow 1 + P[\varnothing] + ... + P[\varnothing] = 1 \Rightarrow P[\varnothing] = 0$.

1.2.4 Law of Total Probability.

Problem 1.27.

Given the model of hand-offs and call lengths in Problem 1.23,
(a) What is the probability that a brief call will have no hand-offs?
(b) What is the probability that a call with one hand-off will be long?
(c) What is the probability that a long call will have one or more hand-offs?

Solution

(a) $P[H_0|B] = \dfrac{P[H_0B]}{P[B]} = \dfrac{0.4}{0.4+0.1+0.1} = \dfrac{2}{3}$

(b) $P[L|H_1] = \dfrac{P[H_1L]}{P[H_1]} = \dfrac{0.1}{0.1+0.1} = \dfrac{1}{2}$

(c) $P[H_1 \cup H_2|L] = \dfrac{P[H_1L \cup H_2L]}{P[L]} = \dfrac{P[H_1L] + P[H_2L]}{P[L]} = \dfrac{3}{4}$

Problem 1.28.

You have a six-sided die that you roll once. Let R_i denote the event that the roll is i. Let G_j denote the event that the roll is greater than j. Let E denote the event that the roll of the die is even-numbered.
(a) What is $P[R_3|G_1]$, the conditional probability that 3 is rolled given that the roll is greater than 1?
(b) What is the conditional probability that 6 is rolled given that the roll is greater than 3?
(c) What is $P[G_3|E]$, the conditional probability that the roll is greater than 3 given that the roll is even?
(d) Given that the roll is greater than 3, what is the conditional probability that the roll is even?

Solution

Let s_i denote the outcome that the roll is i and all outcomes have probability $\frac{1}{6}$. So, for $1 \le i \le 6, R_i = s_i$. Similarly, $G_j = \{s_{j+1}, ..., s_6\}$.

(a) Since $G_1 = \{s_2, ..., s_6\}$, $P[G_1] = \dfrac{5}{6}$, so that $P[R_3|G_1] = \frac{P[R_3G_1]}{P[G_1]} = \frac{1}{5}$.

(b) The conditional probability that 6 is rolled given that the roll is greater than 3 is $P[R_6|G_3] = \frac{P[R_6G_3]}{P[G_3]} = \frac{P[s_6]}{P[s_4,s_5,s_6]} = \frac{1}{3}$.

(c) The event E that the roll is even is $E = s_2, s_4, s_6$ and has probability $\frac{3}{6}$. The joint probability of G_3 and E is $P[G_3E] = P[s_4, s_6] = \frac{1}{3}$. So, the conditional probabilities of G3 given E is $P[G_3|E] = \frac{P[G_3E]}{P[E]} = \frac{2}{3}$.

(d) The conditional probability that the roll is even given that it's greater than 3 is $P[E|G_3] = \frac{P[EG_3]}{P[G_3]} = \frac{2}{3}$.

Problem 1.29.

You have a shuffled deck of three cards: 2, 3, and 4. You draw one card. Let C_i denote the event that card i is picked. Let E denote the event that card chosen is a even-numbered card.
(a) What is $P[C_2|E]$, the probability that the 2 is picked given that an even-numbered card is chosen?
(b) What is the conditional probability that an evennumbered card is picked given that the 2 is picked?

Solution

(a) Since the 2 of clubs is an even numbered card, $C_2 \in E$ so that $P[C_2E] = P[C_2] = \dfrac{1}{3}$.
Since $P[E] = \dfrac{2}{3} \Rightarrow P[C_2|E] = \dfrac{P[C_2E]}{P[E]} = \dfrac{1}{2}$.

(b) The probability that an even numbered card is picked given that the 2 is picked is $P[E|C_2] = \dfrac{P[C_2E]}{P[C_2]} = 1$

Problem 1.30.
Two different suppliers, A and B, provide a manufacturer with the same part. All suppliers of this part are kept in a large bin. In the past, 5 percent of the parts supplied by A and 9 percent of the parts supplied by B have been defective. A supplies four times as many parts as B. Suppose you reach into the bin and select a part and find it is non-defective. What is the probability that it was supplied by A?

Solution

- Denote D :"The selected part is non-defective". So, $P[D|A] = 0.95$ and $P[D|B] = 0.91$.
- A supplies four times as many as B $\Rightarrow P[A] = 0.8$, $P[B] = 0.2$.
- The probability that we need to find is $P[A|D]$. We have $P[D] = P[A].P[D|A] + P[B].P[D|B] = 0.942$.
  Therefore: $P[A|D] = \dfrac{P[D|A].P[A]}{P[D]} = 0.807$.

Problem 1.31.
Suppose that 30 percent of the bottles produced in a certain plant are defective. If a bottle is defective, the probability is 0.9 that an inspector will notice it and remove it from the filling line. If a bottle is not defective, the probability is 0.2 that the inspector will think that it is defective and remove it from the filling line.
(a) If a bottle is removed from the filling line, what is the probability that it is defective?
(b) If a customer buys a bottle that has not been removed from the filling line, what is the probability that it is defective?

Solution

(a) Denote D: "The bottle is defective" and R: "The bottle is removed". So, $P[D] = 0.3, P[R|D] = 0.9, P[R|\overline{D}] = 0.2$.
We have $P[R] = P[D].P[R|D] + P[\overline{D}].P[R|\overline{D}] = 0.41$
$\Rightarrow P[D|R] = \frac{P[R|D].P[D]}{P[R]} = \frac{27}{41}$.

(b) From $P[\overline{R}] = 1 - P[R] = 0.59$, $P[D] = P[D|\overline{R}].P[\overline{R}] + P[D|R].P[R]$
$\Rightarrow P[D|\overline{R}] = \frac{P[D]-P[D|R].P[R]}{P[\overline{R}]} = \frac{3}{59}$.

Problem 1.32.
Suppose that traffic engineers have coordinated the timing of two traffic lights to encourage a run of green lights. In particular, the timing was designed so that with probability 0.75 a driver will find the second light to have the same color as the first. Assuming the first light is equally likely to be red or green.
(a) What is the probability that the second light is green?
(b) What is the probability that you wait for at least one light?

Solution

We denote that R_i and G_i are the red and green light at the time i.
(a) $P[G_2] = 0.5$.
(b) $P[W] = P[R_1] + P[R_2 G_1] = 0.5 + P[R_2|G_1].P[G_1]$
$= 0.5 + 0.25 \times 0.5 = 0.625$.

Problem 1.33.
A factory has three machines A, B, and C. Past records show that the machine A produced 40% of the items of output, the machine B produced 35% of the items of output, and machine C produced 25% of the items. Further 2% of the items produced by machine A were defective, 1.5% produced by machine B were defective, and 1% produced by machine C were defective.
(a) If an item is drawn at random, what is the probability that it is defective?
(b) An item is acceptable if it is not defective. What is the probability that an acceptable item comes from machine A?

Solution

(a) We have $P[A] = 0.4, P[B] = 0.35, P[C] = 0.25$. Denote D: "The product is defective",
$\Rightarrow P[D|A] = 0.02, P[D|B] = 0.015, P[D|C] = 0.01$.
$\Rightarrow P[D] = P[A].P[D|A] + P[B].P[D|B] + P[C].P[D|C] = \frac{63}{4000}$
(b) We have $P[\overline{D}] = 1 - P[D] = \dfrac{3937}{4000}$ and $P[A|D] = \dfrac{P[A].P[D|A]}{P[D]} = \dfrac{32}{63}$
$\Rightarrow P[A|\overline{D}] = \dfrac{P[A] - P[D].P[A|D]}{P[\overline{D}]} = \dfrac{49}{492125}$.

1.2.5 Independent Trials.

Problem 1.34.
Is it possible for A and B to be independent events yet satisfy $A = B$?

Solution

- We denote that R_i and G_i are the red and green light at the time i.
- By definition, events A and B are independent if and only if $P[AB] = P[A]P[B]$.
- We can see that if A = B, that is they are the same set, then $P[AB] = P[AA] = P[A] = P[B]$.
- Thus, for A and B to be the same set and also independent, $P[A] = P[AB] = P[A].P[B] = (P[A])^2$.
- There are two ways that this requirement can be satisfied:
  ◦ $P[A] = 1$ implying $A = B = S$.
  ◦ $P[A] = 0$ implying $A = B = \varnothing$.

Problem 1.35.
Use a Venn diagram in which the event areas are proportional to their probabilities to illustrate two events A and B that are independent.

Solution Disclaimer: I do not have enough time to draw an image here.

Problem 1.36.

In an experiment, A, B, C, and D are events with probabilities $P[A] = \frac{1}{4}$, $P[B] = \frac{1}{8}$, $P[C] = \frac{5}{8}$, and $P[D] = \frac{3}{8}$. Furthermore, A and B are disjoint, while C and D are independent.

(a) Find $P[A \cap B], P[A \cup B], P[A \cap B^c]$, and $P[A \cup B^c]$.
(b) Are A and B independent?
(c) Find $P[C \cap D], P[C \cap D^c]$, and $P[C^c \cap D^c]$.
(d) Are C^c and D^c independent?

Solution

(a) $P[A \cap B] = 0, P[A \cup B] = \dfrac{3}{8}, P[A \cap B^c] = \dfrac{1}{4}, P[A \cup B^c] = \dfrac{7}{8}$

(b) Since A and B are disjoint, so $P[A \cap B] = 0 \neq P[A].P[B]$. This implies that A and B are not independent.

(c) $P[C \cap D] = \dfrac{15}{64}, \quad P[C \cap D^c] = P[C] - P[C \cap D] = \dfrac{25}{64}$,
$P[C^c \cap D^c] = P\left[(C \cup D)^c\right] = 1 - P[C \cup D] = 1 - (P[C] + P[D] - P[C \cap D]) = \dfrac{15}{64}$.

(d) Since C and D are independent, we have $P[CD] = P[C].P[D]$. Thus:
$P[C^c D^c] = P\left[(C \cup D)^c\right] = 1 - P[C \cup D]$
$= 1 - (P[C] + P[D] - P[C \cap D]) = 1 - P[C] - P[D] + P[C]P[D]$
$= (1 - P[C])(1 - P[D]) = P[C^c]P[D^c]$.
This implies that C^c and D^c are independent.

Problem 1.37.

In an experiment, A, B, C, and D are events with probabilities $P[A \cup B] = \dfrac{5}{8}, P[A] = \dfrac{3}{8}, P[C \cap D] = \dfrac{1}{3}, P[C] = \dfrac{1}{2}$. Furthermore, A and B are disjoint, while C and D are independent.

(a) Find $P[A \cap B], P[B], P[A \cap B^c]$, and $P[A \cup B^c]$.
(b) Are A and B independent?
(c) Find $P[D], P[C \cap D^c], P[C^c \cap D^c], P[C|D]$.
(d) Find $P[C \cup D]$ and $P[C \cup D^c]$.
(e) Are C and D^c independent?

Solution

(a) $P[A \cap B] = 0, P[B] = \dfrac{1}{4}, P[A \cap B^c] = \dfrac{3}{8}, P[A \cup B^c]\dfrac{3}{4}$.

(b) Since A and B are disjoint, so $P[A \cap B] = 0 \neq P[A].P[B]$. This implies that A and B are not independent.

(c) $P[D] = \dfrac{P[CD]}{P[C]} = \dfrac{2}{3}, P\left[C \cap D^c\right] = P[C] - P[CD] = \dfrac{1}{6}$,
$P\left[C^c \cap D^c\right] = P\left[(C \cup D)^c\right] = 1 - (P[C] + P[D] - P[CD]) = \dfrac{1}{6}, P[C|D] = P[C] = \dfrac{1}{2}$.

(d) $P[C \cup D] = P[C] + P[D] - P[CD] = \dfrac{5}{6}, P[C \cup D^c] = 1 - P[D] + P[CD] = \dfrac{2}{3}$

(e) Since $P\left[C \cap D^c\right] = P[C] - P[CD] = P[C] - P[C]P[D] = P[C](1 - P[D]) = P[C].P\left[D^c\right]$. This implies that C and D^c are independent.

Problem 1.38.
Consider a binary code with 5 bits (0 or 1) in each code word. An example of a code word is 01010. In each code word, a bit is a zero with probability 0.8, independent of any other bit.

Solution Insufficient data

Problem 1.39.
Suppose each day that you drive to work a traffic light that you encounter is either green with probability $\frac{7}{16}$, red with probability $\frac{7}{16}$, or yellow with probability $\frac{1}{8}$, independent of the status of the light on any other day. If over the course of five days, G, Y, and R denote the number of times the light is found to be green, yellow, or red, respectively, what is the probability that $P[G = 2, Y = 1, R = 2]$? Also, what is the probability $P[G = R]$?

Solution

- We know that the probability of a green and red light is $\dfrac{7}{16}$, and that of a yellow light is $\dfrac{1}{8}$. Since there are always 5 lights, G, Y , and R obey the Bernoulli Trial Calculator:

$$P[G = 2, Y = 1, R = 2] = \frac{5!}{2!1!2!}.\left(\frac{7}{16}\right)^2 \frac{1}{8}\left(\frac{7}{16}\right)^2 \approx 0.137$$

- The probability that the number of green lights equals the number of red lights: $P[G = R] = P[G = 1, Y = 3, R = 1] + P[G = 2, Y = 1, R = 2] + P[G = 0, Y = 5, R = 0]$ ≈ 0.145.

Problem 1.40.
We wish to modify the cellular telephone coding system in example below to reduce the number of errors. In particular, if there are 2 or 3 zeroes in the received sequence of 5 bits, we will say that a deletion (event D) occurs. Otherwise, if at least 4 zeroes are received, then the receiver decides a zero was sent. Similarly, if at least 4 ones are received, then the receiver decides a one was sent. We say that an error occurs if either a one was sent and the receiver decides zero was sent or if a zero was sent and the receiver decides a one was sent. For this modified protocol, what is the prob. $P[E]$ of an error? What is the prob. $P[D]$ of a deletion?

Example

We wish to modify the cellular telephone coding system in example below to reduce the number of errors. In particular, if there are 2 or 3 zeroes in the received sequence of 5 bits, we will say that a deletion (event D) occurs. Otherwise, if at least 4 zeroes are received, then the receiver decides a zero was sent. Similarly, if at least 4 ones are received, then the receiver decides a one was sent. We say that an error occurs if either a one was sent and the receiver decides zero was sent or if a zero was sent and the receiver decides a one was sent. For this modified protocol, what is the prob. $P[E]$ of an error? What is the prob. $P[D]$ of a deletion?

Solution of the Example.

- We denote s_k is the case that there are k successes in 5 trials.
- We have 5 trials corresponding to the 5 times the binary symbol is sent. On each trial, a success occurs when a binary symbol is received correctly. The probability of a success is $p = 1 - q = 0.9$. The error event E occurs when the number of successes is less than three:

$$P[E] = P[s_0] + P[s_1] + P[s_2] = q^5 + C_5^1.pq^4 + C_5^2.p^2q^3 = \frac{107}{12500} = 8.56.10^{-3}$$

Solution of Problem 1.40.

- The probability of a success is $P[S] = P[s_4] + P[s_5] = C_5^4.p^4.q + p^5 = 0.91854$.
- The probability of a deletion is $P[D] = P[s_2] + P[s_3] = C_5^2.p^2q^3 + C_5^3.p^3q^2 = 0.081$.
- So, the probability of an error is $P[E] = 1 - P[S] - P[D] = 4.6.10^{-4}$

Problem 1.41.

An airline sells 200 tickets for a certain flight on an airplane that has only 198 seats because, on the average, 1 percent of purchasers of airline tickets do not appear for the departure of their flight. Determine the probability that everyone who appears for the departure of this flight will have a seat.

Solution

- Let X = "number of people who do not appear for their flight". We can assume that the passengers independently decide to show up or not, and we'll consider not appearing for the flight a success (for the airline!).
- Then X is the number of successes in a sequence of n = 200 independent Bernoulli trials with probability of success $p = 1\% = 0.01$. So, $X\ Bin(n = 200, p = 0.01)$.
- Because $n = 200$ is large, $p = 0.01$ is small and $np = 2$ is moderate, we can approximate the $Bin(n = 200, p = 0.01)$ distribution by Poisson $(= np = 2)$ distribution.
- Now, everyone will get a seat if and only if at least 2 passengers do not appear, i.e., $X \geq 2$.
- Therefore, required probability:

$$P[X \geq 2] = 1 - P[X \leq 1] = 1 - P[0] - P[1] \approx 1 - e^{-2}\frac{2^0}{0!} - e^{-2}\frac{2^1}{1!} = 1 - 3e^{-2} = 0.59.$$

1.2.6 In-class exercise

Exercise 1.
Let A and B be 2 events such that $P[A] = 0.6, P[B|A] = 0.7$ and $P[B|\overline{A}] = 0.8$. What is the value of $P[A|B]$?

Solution

- $P[AB] = P[B|A]P[A] = 0.42 \quad ; \quad P[\overline{A}B] = P[B|\overline{A}].P[\overline{A}] = 0.32$
  $P[B] = P[AB] + P[\overline{A}B] = 0.74$.
- So, $P[A|B] = \dfrac{P[AB]}{P[B]} \approx 0.568$

Exercise 2.
Let A and B be 2 events such that $P[AB] = 0.4; P[\overline{AB}] = 0.3$. Furthermore $A\overline{B}$ and $\overline{A}B$ are equally likely. What is the probability of $\overline{A}$?

Solution

- $P[AB] + P[A\overline{B}] + P[\overline{A}B] + P[\overline{AB}] = 1 \Rightarrow 2P[A\overline{B}] = 1 - 0.3 - 0.4 = 0.3 \Rightarrow P[A\overline{B}] = P[\overline{A}B] = 0.15$
- $P[\overline{A}] = P[\overline{A}B] + P[\overline{AB}] = 0.15 + 0.3 = 0.45$

Exercise 3.
There are 3 boxes of marbles: the first box contains 3 red marbles, 2 white marbles; the second box contains 2 red marbles, 2 white marbles; the third box has no marbles. Draw randomly 1 marble from the first box and 1 marble from the second box and put them in the third box. Then, from the third box, 1 marble is drawn at random. Given that the marble drawn from the third box is red, what is the probability that the marble drawn from the first box is red?

Solution

- We denote that R and W are the event that the drawn marble is red and white respectively. Since the drawn marble in the third box is red, there are two cases for two drawn marbles which satisfy the requirement: RR and RW.
- Hence, the probability of getting a red one in the first and second box is $\frac{3}{5}; \frac{2}{5}$; getting a white one in the first and second box is $\frac{2}{4}; \frac{2}{4}$ respectively.
- The probability that the drawn marble is red is: $P[R] = P[R|RR].P[RR] + P[R|RW].P[RW] + P[R|WR].P[WR] + P[R|WW].P[WW] = 0.55$
- The probability for the red drawn marble in the first box is: $P[RR + RW|R] = P[RR|R] + P[RW|R] = \dfrac{P[RR].P[R|RR] + P[RW].P[R|RW]}{P[R]}$
- In conclusion, $P[RR + RW|R] = \dfrac{0.2 \times 1 + 0.2 \times 0.5}{0.55} = \dfrac{9}{11} \approx 0.818$.

Chapter 2

Random Variables and Probability Distributions.

2.1 Theory.

2.1.1 Concept of random variables.

- A variable X is a random variable if the value that it assumes, corresponding to the outcome of an experiment, is random.
- X is a discrete random variable if the range of X is a countable set: $S_X = \{x_1, x_2, ..., x_n\}$.
- X is a continuous random variable if the range of X is a uncountably infinite set (a continuous interval $[a, b]$).

2.1.2 Discrete Probability Distributions.

- The probability distribution (PD) of a discrete random variable is given by the following table:

X	x_1	x_2	x_3	...
$P[X = x_i]$	p_1	p_2	p_3	...

- X takes the values: $x_1, x_2, ...$
- Probability $p_i = P[X = x_i] \in [0, 1]$ such that $\sum p_i = 1$.
- The probability mass function (PMF) of a discrete random variable X is the following:

$$P_X(x) = \begin{cases} p_i, & \text{if} x = x_i,\ i = 1, 2, 3, ... \\ 0, & \text{otherwise.} \end{cases}$$

- The cumulative distribution function (CDF) $F_X(x)$ of a discrete random variable X with probability distribution $P_X(x)$ is

$$F_X(x) = P[X < x] = \sum_{t<x} P_X(t) = \begin{cases} 0, & x \leq x_1, \\ p_1, & x_1 < x \leq x_2, \\ p_1 + p_2 & x_2 < x \leq x_3, \\ ... \\ 1, & x > x_n, \end{cases}$$

2.1.3 Continuous Probability Distributions.

- The cumulative distribution function (CDF) of random variable X is $F_X(x) = P[X < x]$, $x \in \mathbb{R}$. X is a continuous random variable if the CDF $F_X(x)$ is a continuous function. The following properties of the CDF of X:
  - $0 \leq F_X(x) \leq 1$, $\forall x \in \mathbb{R}$.
  - $F_X(-\infty) = 0$ and $F_X(+\infty) = 1$.
  - $F_X(x)$ is a non-decreasing function on $\mathbb{R}$.
  - $P[a \leq x < b] = F_X(b) - F_X(a)$.
  - $F_X(x) = \int\limits_{-\infty}^{x} f_X(u)\mathrm{d}u$.
- The probability distribution of a continuous random variable is given by a probability density function (PDF) $f_X(x)$ that satisfies the following conditions:
  - $f_X(x) = 0$, $\forall x \in \mathbb{R}$.
  - $\int\limits_{-\infty}^{+\infty} f_X(x)\mathrm{d}x = 1$ (the area under the curve $y = f_X(x)$ and Ox is 1).
  - $P[X = a] = 0 \forall x \in \mathbb{R}$
  - $P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b) = \int\limits_{a}^{b} f_X(x)\mathrm{d}x$ ((the area under the curve $y = f_X(x)$ and between $x = a$ and $x = b$).
  - $f_X(x) = \dfrac{\mathrm{d}F_X(x)}{\mathrm{d}x}$.

2.1.4 Characteristic measures of a random variable.

- Central location of a random variable:
  - The expected value (the expectation, the averaged value, the mean) of X is:
    $\mu = E(X) = \sum\limits_i x_i P[X = x_i]$ if X is a discrete random variable.
    $\mu = E(X) = \int\limits_{-\infty}^{+\infty} x f_X(x)$ if X is a continuous random variable.
    Properties: $E(aX + bY) = aE(X) + bE(Y); E(XY) = E(X)E(Y)$ if X, Y are independent.
  - A mode of random variable X is a number $x$mod satisfying:

$$P_X[\text{mod}] \geq P_X[x] \ \forall x$$

  - A median, $x$med, of random variable X is a number that satisfies

$$P_X[\text{mod} \geq x] = P_X[x < \text{mod}]$$

- Functions of a Random Variable:
  - For a discrete random variable X, the PMF of $Y = g(X)$ is $\boxed{P_Y(y) = \sum\limits_{x:g(x)=y} P_X(x)}$
  - Let X be a random variable with probability distribution $P_X(x)$, or $f_X(x)$. The expected value of the random variable $Y = g(X)$ is
    $\mu = E(g(X)) = \sum\limits_{x \in S(x)} g(x) P_X[x]$ if X is a discrete random variable.
    $\mu = E(g(X)) = \int\limits_{-\infty}^{+\infty} g(x) f_X(x)$ if X is a continuous random variable.

- Variability of a random variable:
  - The variance of X is $Var[X] = E[(X-\mu)^2] = E[X^2] - (E[X])^2$.
  - Properties of variance:
    $Var(X) = \sum_i x_i^2 P[X = x_i] - \mu^2$ if X is a discrete random variable.
    $Var(X) = \int_{-\infty}^{+\infty} x^2 f_X(x) - \mu^2$ if X is a continuous random variable.
  - $Var(aX+b) = a^2 Var(X); Var(X+Y) = Var(X-Y) = Var(X) + V(Y)$ if X, Y are independent.
  - The standard deviation of X is $\sigma[X] = \sqrt{Var[X]}$

2.2 Answer of Exercises.

2.2.1 Discrete Random Variables.

Problem 2.1.
A civil engineer is studying a left-turn lane that is long enough to hold 7 cars. Let X be the number of cars in the lane at the end of a randomly chosen red light. The engineer believes that the probability that $X = x$ is proportional to $(x+1)(8-x)$ for $x = 0, 1, ..., 7$.

(a) Find the probability mass function of X.
(b) Find the probability that X will be at least 5.

Solution

(a) We denote PMF of X is $P_X(x) = c(x+1)(8-x)$. To find c, we have $\sum P_X(x) = 1$, that means $c \sum_{x=0}^{7}(x+1)(8-x) = 1 \Rightarrow c = \frac{1}{120}$. This implies that PMF of X is $P_X(x) = \frac{1}{120}(x+1)(8-x)$.
(b) $P[X \geq 5] = P_X(5) + P_X(6) + P_X(7) = \frac{3}{20} + \frac{7}{60} + \frac{1}{15} = \frac{1}{3}$.

Problem 2.2.
A midterm test has 4 multiple choice questions with four choices with one correct answer each. If you just randomly guess on each of the 4 questions, what is the probability that you get exactly 2 questions correct? Assume that you answer all and you will get (+5) points for 1 question correct, (-2) points for 1 question wrong. Let X is number of points that you get. Find the probability mass function of X and the expected value of X.

Solution

- Since the outcomes of choosing one answer out of 4 are equally likely $\Rightarrow$ The probability of choosing one answer is 0.25. There is one correct answer and the others are wrong.
  $\Rightarrow$ The probability that choosing correct answer is 0.25 and choosing wrong answer is 0.75. Apply Bernoulli formula: $P_4(2) = C_4^2 \times 0.75^2 = \frac{27}{128}$.

- The range of X is $S_X = 20, 13, 6, -1, -8$. Apply Bernoulli formula:

$$\begin{cases} P[X=20] = C_4^4 \times 0.25^4 \times 0.75^0 = \dfrac{1}{256}, \\ P[X=13] = C_4^3 \times 0.25^3 \times 0.75^1 = \dfrac{3}{64}, \\ P[X=6] = C_4^2 \times 0.25^2 \times 0.75^2 = \dfrac{27}{128}, \\ P[X=-1] = C_4^1 \times 0.25^1 \times 0.75^3 = \dfrac{27}{64}, \\ P[X=-8] = C_4^0 \times 0.25^0 \times 0.75^4 = \dfrac{81}{256}. \end{cases}$$

Problem 2.3.

The random variable N has PMF $P_N(n) \begin{cases} c\left(\frac{1}{2}\right)^2, & n = 0,1,2 \\ 0, & \text{otherwise} \end{cases}$

(a) What is the value of the constant c?
(b) What is $P[N \leq 1]$?
(c) What is the PD of X?

Solution

(a) We have $\sum_0^2 c\left(\dfrac{1}{2}\right)^n = 1 \Leftrightarrow c\left(1 + \dfrac{1}{2} + \dfrac{1}{4}\right) \Rightarrow c = \dfrac{4}{7}$.
(b) $P[N \leq 1] = P_N[0] + P_N[1] = \dfrac{6}{7}$?
(c) PD of X:

$X = x_i$	0	1	2
$P[X = x_i]$	$\frac{4}{7}$	$\frac{2}{7}$	$\frac{1}{7}$

Problem 2.4.

The random variable V has PMF $P_N(n) \begin{cases} cv^2, & v = 1,2,3,4 \\ 0, & \text{otherwise} \end{cases}$

(a) Find the value of the constant c.
(b) Find $P[V \in u^2 \mid u = 1,2,3,...]$.
(c) Find the probability that V is an even number.
(d) Find $P[V > 2]$.

Solution

(a) $c = \frac{1}{30}$.
(b) $P[V \in u^2 \mid u = 1,2,3,...] = P[V \in 1,4,9,16,25] = P_V(1) + P_V(4) = \dfrac{1^2 + 4^2}{30} = \dfrac{17}{30}$.
(c) $P[V \mid V \bmod 2 = 0] = P_V(2) + P_V(4) = \dfrac{2}{3}$.
(d) $P[V > 2] = P_V(3) + P_V(4) = \dfrac{3^2 + 4^2}{30} = \dfrac{5}{6}$.

Problem 2.5.

Suppose when a baseball player gets a hit, a single is twice as likely as a double which is twice as likely as a triple which is twice as likely as a home run. Also, the player's batting average, i.e., the probability the player gets a hit, is 0.300. Let B denote the number of bases touched safely during an at-bat. For example, B = 0 when the player makes an out, B = 1 on a single, and so on. What is the PMF of B?

Solution

- We denote: $\begin{cases} P(B=0) = 1 - 0.3 = 0.7 \\ P(B=1) = x \end{cases} \Rightarrow \begin{cases} P(B=2) = x/2 \\ P(B=3) = x/4 \\ P(B=4) = x/8 \end{cases}$

  $\Rightarrow 0.7 + x + \dfrac{x}{2} + \dfrac{x}{4} + \dfrac{x}{8} = 1 \Leftrightarrow \dfrac{15x}{8} = 0.3 \Leftrightarrow x = 0.16.$

- As a result, the PMF of B is given by $P_B(b) = \begin{cases} 0.7, & b = 0 \\ 0.16, & b = 1 \\ 0.08, & b = 2 \\ 0.04, & b = 3 \\ 0.02, & b = 4 \end{cases}$

Problem 2.6.

In a package of M&Ms, Y, the number of yellow M&Ms, is uniformly distributed between 5 and 15.

(a) What is the PMF of Y?
(b) What is $P[Y < 10]$?
(c) What is $P[Y > 12]$?
(d) What is $P[8 \leq Y \leq 12]$?

Solution

(a) PMF of Y: $\begin{cases} P_Y(y) = \frac{1}{11}, & \text{if } y = \overline{5, 15}, \\ 0, & \text{otherwise.} \end{cases}$

(b) $P[Y < 10] = \sum\limits_{i=5}^{i=9} P_Y(i) = \dfrac{5}{11}.$

(c) $P[Y > 12] = \dfrac{3}{11}.$

(d) $P[8 \leq Y \leq 12] = \dfrac{5}{11}.$

Problem 2.7.

When a conventional paging system transmits a message, the probability that the message will be received by the pager it is sent to is p. To be confident that a message is received at least once, a system transmits the message n times.

(a) Assuming all transmissions are independent, what is the PMF of K, the number of times the pager receives the same message?
(b) Assume $p = 0.8$. What is the minimum value of n that produces a probability of 0.95 of receiving the message at least once?

Solution

(a) $P_K(k) = C_n^k p^k (1-p)^{n-k}$ with $k = 0..n$.
(b) $P[\text{at least one}] = 1 - P_K[0] = 1 - 0.2^n$. From the assumption, $P[\text{at least one}] = 0.95$, that implies $0.2^n = 0.05 \Leftrightarrow n = \dfrac{\ln 0.05}{\ln 0.2} = 1.86 \approx 2$.

Problem 2.8.

When a two-way paging system transmits a message, the probability that the message will be received by the pager it is sent to is p. When the pager receives the message, it transmits an acknowledgment signal (ACK) to the paging system. If the paging system does not receive the ACK, it sends the message again.

(a) What is the PMF of N, the number of times the system sends the same message?
(b) The paging company wants to limit the number of times it has to send the same message. It has a goal of $P[N \leq 3] \geq 0.95$. What is the minimum value of p necessary to achieve the goal?

Solution

(a) The n same messages are sent when the pager doesn't receive the message n times. So, the PMF of N is: $P_N(n) = \begin{cases} (1-p)^{n-1}.p, & \text{if n} = 1, 2, ... \\ 0, & \text{otherwise.} \end{cases}$
(b) We have

$$P[N \leq 3] = 1 - P[N > 3] = 1 - \sum_{n=4}^{+\infty} P_N(n) = 1 - p\left[(1-p)^3 + (1-p)^4 + ... + (1-p)^n\right]$$

$$= 1 - p.\lim_{n\to+\infty}(1-p)^3.\frac{1-(1-p)^n}{1-(1-p)} = 1 - (1-p)^3.$$

Since $P[N \leq 3] \geq 0.95 \Rightarrow (1-p)^3 \leq 0.05 \Leftrightarrow p \geq 1 - 0.05^{1/3} \approx 0.63$.

Problem 2.9.

The number of bits B in a fax transmission is a geometric $(p = 2.5 \times 10^{-5})$ random variable. What is the probability $P[B > 500,000]$ that a fax has over 500,000 bits?

Solution

- The PMF of B is $P_B(b) = \begin{cases} (1-p)^{b-1}.p, & \text{if b} = 1, 2, ... \\ 0, & \text{otherwise.} \end{cases}$. We denote $q = 1 - p$, so we have $P_B(b) = \begin{cases} (1-p).p^{b-1}, & \text{if b} = 1, 2, ... \\ 0, & \text{otherwise.} \end{cases}$.

  This implies $\sum_{i=1}^{n} P_B(b) = 1 - q + q - q^2 + ... = 1 - q^n$.

- So, $P[B > 500,000] = 1 - P[B \le 500,000]$
  $= 1 - (1 - q^{500,000}) = (1-p)^{500,000} = (1 - 2.5 \times 10^{-5})^{500,000}$.

Problem 2.10.

The random variable X has CDF: $F_X(x) = \begin{cases} 0, & x < -1, \\ 0.2, & -1 \le x < 0, \\ 0.7, & 0 \le x < 1, \\ 1, & x \ge 1. \end{cases}$

(a) Draw a graph of the CDF.
(b) Write $P_X(x)$, the PMF of X. Be sure to write the value of $P_X(x)$ for all x from $-\infty$ to $+\infty$.
(c) Write the PD of X.

Solution

(a) I'm too lazy to draw.

(b) PMF of X: $P_X(x) = \begin{cases} 0.2, & x = -1, \\ 0.5, & x = 0, \\ 0.3, & x = 1, \\ 0, & \text{otherwise.} \end{cases}$

(c) PD of X is

$X = x_i$	-1	0	1
$P[X = x_i]$	0.2	0.5	0.3

Problem 2.11.

The random variable X has CDF: $F_X(x) = \begin{cases} 0, & x < -3, \\ 0.4, & -3 \le x < 5, \\ 0.8, & 5 \le x < 7, \\ 1, & x \ge 7. \end{cases}$

(a) Draw a graph of the CDF.
(b) Write $P_X(x)$, the PMF of X.
(c) Write the PDT of X.

Solution

(a) I'm too lazy to draw.

(b) PMF of X: $P_X(x) = \begin{cases} 0.4, & x = -3, \\ 0.4, & x = 5, \\ 0.2, & x = 7, \\ 0, & \text{otherwise.} \end{cases}$

(c) Similar to Problem 2.10.

Problem 2.12.

In **Problem 2.5**, find and sketch the CDF of B, the number of bases touched safely during an at-bat.

Solution

CDF of B: $F_B(b) = \begin{cases} 0.7, & b = 0 \\ 0.86, & b = 1 \\ 0.94, & b = 2 \\ 0.98, & b = 3 \\ 1, & b = 4 \end{cases}$. Here, we can sketch it, but I'm too lazy to do.

Problem 2.13.

Let X have the uniform PMF $P_X(x) = \begin{cases} 0.01, & x = 1, 2, ..., 100, \\ 0, & \text{otherwise.} \end{cases}$

(a) Find a mode x_{mod} of X. If the mode is not unique, find the set X_{mod} of all modes of X.
(b) Find a median x_{med} of X. If the median is not unique, find the set X_{med} of all numbers x that are medians of X.

Solution

(a) Since the mode of the discrete random variable is the value x_{mod} where $P[X = x_{\text{mod}}]$ is maximum. Here, we have $x_{\text{mod}} \in \{1, 2, ..., 100\}$ as the values for which $P[X = x]$ is maximum. Therefore, the set $X_{\text{mod}} = \{1, 2, ..., 100\}$.
(b) x_{med} is the value such that

$$P[X \leq x_{\text{med}}] = P[X \geq x_{\text{med}}] \text{ and } P[X \leq x_{\text{med}}] + P[X \geq x_{\text{med}}] = 1$$

Then, $P[X \leq x_{\text{med}}] = F_X[x_{\text{med}}] = 0.5$

Since CDF of X is $\begin{cases} 0, & x < 1, \\ 0.01, & 1 \leq x < 2, \\ ... \\ 0.5, & 50 \leq x < 51, \\ ... \\ 1, & x \geq 100. \end{cases}$ then $x_{\text{med}} \in [50, 51)$

Problem 2.14.

Voice calls cost 20 cents each and data calls cost 30 cents each. C is the cost of one telephone call. The probability that a call is a voice call is $P[V] = 0.6$. The probability of a data call is $P[D] = 0.4$.

(a) Find $P_C(c)$, the PMF of C.
(b) Find the PD of C.
(c) What is $E[C]$, the expected value of C?

Solution

(a) The PMF of C: $P_C(c) = \begin{cases} 0.6, & c = 20, \\ 0.4, & c = 30, \\ 0, & \text{otherwise.} \end{cases}$

(b) The PD of C:

$X = x_i$	20	30
$P[X = x_i]$	0.6	0.4

(c) Expected value: $E[C] = 0.6 \times 20 + 0.4 \times 30 = 24$.

Problem 2.15.

Find the expected value of the random variable X in **Problem 2.10**.

Solution

- $E[X] = 0.2 \times (-1) + 0.5 \times 0 + 0.3 \times 1 = 0.1$.

Problem 2.16.

Find the expected value of the random variable X in **Problem 2.11**.

Solution

- $E[X] = 0.4 \times (-3) + 0.4 \times 5 + 0.2 \times 7 = 2.2$.

Problem 2.17.

Find the expected value of a binomial ($n = 4,\ p = \frac{1}{2}$) random variable X.

Solution

- Since the binomial distribution with $n = 4,\ p = \dfrac{1}{2}$, so PMF of X is

$$P_X(x) = \begin{cases} C_n^x \left(\dfrac{1}{2}\right)^x \cdot \left(\dfrac{1}{2}\right)^{4-x}, & x = 0, 1, ..., 4, \\ 0, & \text{otherwise.} \end{cases} \Leftrightarrow P_X(x) = \begin{cases} \dfrac{C_n^x}{16}, & x = 0, 1, ..., 4, \\ 0, & \text{otherwise.} \end{cases}$$

- The expected value of X is $E[X] = np = 2$.

Problem 2.18.

Give examples of practical applications of probability theory that can be modeled by the following PMFs. In each case, state an experiment, the sample space, the range of the random variable, the PMF of the random variable, and the expected value: (a) Bernoulli;(b) Binomial; (c) Poisson. Make up your own examples.

Solution

No idea for this problem.

Problem 2.19.

Given the random variable X in Problem 2.10, let $V = g(X) = |X|$.

(a) Find $P_V(v)$.
(b) Find $F_V(v)$.
(c) Find $E[V]$.

Solution

(a) $P_V[1] = P_X[-1] + P_X[1] = 0.5$, and $P_V[0] = P_X[0] = 0.5$, so: $P_V(v) = \begin{cases} 0.5, & v = 0, 1, \\ 0, & \text{otherwise.} \end{cases}$.

(b) $F_V(v) = \begin{cases} 0, & v < 0, \\ 0.5, & 0 \le v < 1, \\ 1, & v \ge 1. \end{cases}$.

(c) $E[V] = 0 \times 0.5 + 1 \times 0.5 = 0.5$.

Problem 2.20.

In a certain lottery game, the chance of getting a winning ticket is exactly one in a thousand. Suppose a person buys one ticket each day (except on the leap year day February 29) over a period of fifty years. What is the expected number $E[T]$ of winning tickets in fifty years? If each winning ticket is worth \$1000, what is the expected amount $E[R]$ collected on these winning tickets? Lastly, if each ticket costs \$2, what is your expected net profit $E[Q]$?

Solution

- Whether a lottery ticket is a winner is a Bernoulli trial with a success probability of 0.001. If we buy one every day for 50 years for a total of $50 \times 365 = 18250$ tickets, then the number of winning tickets T is a binomial random variable with mean

$$E[T] = 18250 \times 0.001 = 18.25$$

- Since each winning ticket grosses \$1000, the revenue we collect over 50 years is $R = 1000T$ dollars. The expected revenue is

$$E[R] = 1000E[T] = 18250$$

- But buying a lottery ticket everyday for 50 years, at \$2.00 a pop isn't cheap and will cost us a total of $\$18250 \times 2 = \$36500$. Our net profit is then $Q = R - 36500$ and the result of our loyal 50 year patronage of the lottery system, is disappointing expected loss of

$$E[Q] = E[R] - 36500 = -18250$$

Problem 2.21.

In an experiment to monitor two calls, the PMF of N, the number of voice calls, is

$$F_N(n) \begin{cases} 0.2, & n = 0, \\ 0.7, & n = 1, \\ 0.1, & n = 2, \\ 0, & \text{otherwise.} \end{cases}$$

(a) Find $E[N]$, the expected number of voice calls.
(b) Find $E[N^2]$, the second moment of N.
(c) Find $Var[N]$, the variance of N.
(d) Find σ_N, the standard deviation of N.

Solution

(a) $E[N] = 0.2 \times 0 + 0.7 \times 1 + 0.1 \times 2 = 0.9$.
(b) $E[N^2] = 0.2 \times 0^2 + 0.7 \times 1^2 + 0.1 \times 2^2 = 1.1$.
(c) $Var[N] = E[(N - 0.9)^2] = (0 - 0.9)^2 \times 0.2 + (1 - 0.9)^2 \times 0.7 + (2 - 0.9)^2 \times 0.1 = 0.29$.
(d) $\sigma_N = \sqrt{Var[N]} = \dfrac{\sqrt{29}}{10}$.

Problem 2.22.

Find the variance of the random variable X in **Problem 2.10**.

Solution

- $Var[X] = E[(X - 0.1)^2] = (-1 - 0.1)^2 \times 0.2 + (0 - 0.1)^2 \times 0.5 + (1 - 0.1)^2 \times 0.3 = 0.49$.

Problem 2.23.

Let X have the binomial PMF $P_X(x) = C_4^x \left(\frac{1}{2}\right)^4$.

(a) Find the standard deviation of the random variable X.
(b) What is $P[\mu_X - \sigma_X \leq X \leq \mu_X + \sigma_X]$, the probability that X is within one standard deviation of the expected value?

Solution

(a) The mean of X is $E(X) = 0 \times \dfrac{1}{16} + ... + 4 \times \dfrac{1}{16} = 2$

The variance of X is $Var[X] = E(X^2) - [E(X)]^2 = 0^2 \times \dfrac{1}{16} + 4^2 \times \dfrac{1}{16} - 2^2 = 1$

So the standard deviation of X is $\sigma_X = \sqrt{Var[X]} = 1$

(b) $P[\mu_X - \sigma_X \leq X \leq \mu_X + \sigma_X] = P[2-1 \leq X \leq 2+1] = P[1 \leq X \leq 3] = P[X = 1] + P[X=2] + P[X=3] = \frac{14}{16} = \frac{7}{8}$

Problem 2.24.

Show that the variance of $Y = aX + b$ is $Var[Y] = a^2 Var[X]$

Solution

$$\begin{aligned}
Var[Y] &= E[Y^2] - (E[Y])^2 = E[a^2X^2 + 2abX + b^2] - (aE[X]+b)(aE[X]+b) \\
&= a^2E[X^2] + 2abE[X] + b^2 - a^2E[X]^2 - 2abE[X] - b^2 \\
&= a^2E[X^2] - a^2E[X]^2 = a^2Var[X]
\end{aligned}$$

Problem 2.25.

Given a random variable X with mean μ_X and variance σ_X^2 , find the mean and variance of the standardized random variable $Y = \frac{X-\mu_X}{\sigma_X}$.

Solution

- Since $Y = \frac{X-\mu_X}{\sigma_X}$
  $\Rightarrow E(Y) = E\left(\frac{X-\mu_X}{\sigma_X}\right) = \frac{1}{\sigma_X}E(X-\mu_X) = \frac{1}{\sigma_X}\left(E(X) - E(\mu_X)\right) = \frac{1}{\sigma_X}(\mu_X - \mu_X) = 0$
- $Var[Y] = Var\left[\frac{X-\mu_X}{\sigma_X}\right] = \frac{1}{\sigma_X^2}Var[X-\mu_X] = \frac{1}{\sigma_X^2}Var[X] = \frac{1}{\sigma_X^2}.\sigma_X^2 = 1$

Problem 2.26.

In **Problem 2.10**, find $P_{X|B}(x)$, where the condition $B = |X| > 0$. What are $E[X|B]$ and $Var[X|B]$?

Solution

- PMF of X: $P_X(x) = \begin{cases} 0.2, & x = -1, \\ 0.5, & x = 0, \\ 0.3, & x = 1, \\ 0, & \text{otherwise.} \end{cases}$.
- Since $B = |X| > 0 \Rightarrow P[B] = 1 - P[X=0] = \frac{1}{2}$. So, the PMF of $X|B$ is:

$$P_{X|B}(x) = \frac{P_X(x)}{P_B(x)} \begin{cases} 0.4, & x = -1, \\ 0.6, & x = 1, \\ 0, & \text{otherwise.} \end{cases}$$

  Since B is non-zero variables of X, so $P_{X|B}$ only takes the non-zero values where $X \neq 0$.
- $E[X|B] = 0.4 \times (-1) + 0.6 \times 1 = 0.2$.

- $Var[X|B] = E[X^2|B] - (E[X|B])^2 = 0.4 \times (-1)^2 + 0.6 \times 1^2 - 0.2^2 = 0.96.$

Problem 2.27.
In **Problem 2.23**, find $P_{X|B}(x)$, where the condition $B = X \neq 0$. What are $E[X|B]$ and $Var[X|B]$?

Solution

- PMF of X: $P_X(x) = C_4^x \left(\frac{1}{4}\right)$.
- Since $B = X \neq 0 \Rightarrow P[B] = 1 - P[X = 0] = \frac{15}{16}$. Here, the PMF of $X|B$ is $P_{X|B} = \frac{P_X(x)}{P_B(x)} = \frac{C_4^x}{15}$ with $x = 1..4$.
- $E[X|B] = 1 \times \frac{C_4^1}{15} + ... + 4 \times \frac{C_4^4}{15} = \frac{32}{15}$.
- $Var[X|B] = E[X^2|B] - (E[X|B])^2 = \frac{80}{15} - \frac{32}{15} = \frac{16}{5}$.

2.2.2 Continuous Random Variables.

Problem 2.28.
The cumulative distribution function of random variable X is

$$F_X(x) = \begin{cases} 0, & x \leq -1, \\ \frac{x+1}{2}, & -1 < x \leq 1, \\ 1, & x > 1 \end{cases}$$

(a) What is $P[X \leq 1/2]$?
(b) What is $P[-1/2 \leq X < 3/4]$?
(c) What is $P[|X| > 1/2]$?
(d) What is the value of a such that $P[X < a] = 0.8$?

Solution

(a) We have PDF of X: $f_X(x) = \begin{cases} \frac{1}{2}, & -1 < x \leq 1, \\ 0, & \text{otherwise}. \end{cases}$.

So, $P[X \leq 1/2] = \int\limits_{-\infty}^{1/2} f_X(x)dx = \int\limits_{-\infty}^{-1} 0dx + \int\limits_{-1}^{1/2} \frac{1}{2} = 0.75.$

(b) $P[-1/2 \leq X < 3/4] = \int\limits_{-1/2}^{3/4} \frac{1}{2} = \frac{5}{8}$.

(c) $P[|X| > 1/2] = \int\limits_{-1}^{-1/2} \frac{1}{2} + \int\limits_{1/2}^{1} \frac{1}{2} = \frac{1}{2}$.

(d) $P[X < a] = 0.8 \Leftrightarrow \int\limits_{-\infty}^{a} f_X(x) = 0.8 \Leftrightarrow \int\limits_{-1}^{a} \frac{1}{2} = 0.8 \Leftrightarrow \frac{a}{2} + \frac{1}{2} = 0.8 \Leftrightarrow a = 0.6.$

Problem 2.29.

The cumulative distribution function of the continuous random variable V is

$$F_V(v) = \begin{cases} 0, & v \le -5, \\ c(v+5)^2, & -5 < x \le 7, \\ 1, & x > 7 \end{cases}$$

(a) What is c?
(b) What is $P[V \ge 4]$?
(c) What is $P[-3 \le v < 0]$?
(d) What is the value of a such that $P[V \ge a] = 2/3$?

Solution

(a) For V to be a continuous variable, we should choose the point that should not have any discontinuity. Take $x = 7$, so $F_V(7) = 1 = c(7+5)^2 \Rightarrow c = \frac{1}{144}$.
(b) $P[V \ge 4] = 1 - P[V < 4] = 1 - F_V(4) = \frac{9}{16}$.
(c) $P[-3 \le v < 0] = F_V[0] - F_V[-3] = \frac{7}{48}$.
(d) $P[V \ge a] = 1 - F_V(a) = 1 - \frac{(a+5)^2}{144} = \frac{2}{3} \Leftrightarrow a = -5 + 4\sqrt{3}$.

Problem 2.30.

The random variable X has probability density function $f_X(x) = \begin{cases} cx, & 0 \le x \le 2, \\ 0, & \text{otherwise}. \end{cases}$.

Use the PDF to find

(a) the constant c.
(b) $P[0 \le X \le 1]$.
(c) $P\left[-\frac{1}{2} \le X \le \frac{1}{2}\right]$.
(d) the CDF $F_X(x)$.

Solution

(a) We have $1 = \int\limits_{-\infty}^{+\infty} f_X(x)\mathrm{d}x = \int\limits_0^2 cx\mathrm{d}x = 2c \Rightarrow c = \frac{1}{2}$.
(b) $P[0 \le X \le 1] = \int\limits_0^1 \frac{x}{2}\mathrm{d}x = \frac{1}{4}$.
(c) $P\left[-\frac{1}{2} \le X \le \frac{1}{2}\right] = \int\limits_0^{\frac{1}{2}} \frac{x}{2}\mathrm{d}x = \frac{1}{16}$.
(d) CDF of X: $F_X(x) = \begin{cases} 0, & x < 0, \\ \frac{cx^2}{2}, & 0 \le x \le 2, \\ 1, & x > 2. \end{cases}$.

Problem 2.31.

The cumulative distribution function of random variable X is

$$F_X(x) = \begin{cases} 0, & x \leq -1, \\ \dfrac{x+1}{2}, & -1 < x \leq 1, \\ 1, & x > 1 \end{cases}$$

Find the PDF $f_X(x)$ of X.

Solution

$$\text{The PDF of X} : f_X(x) = \begin{cases} \dfrac{1}{2}, & -1 < x \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$

Problem 2.32.

Continuous random variable X has PDF $f_X(x) = \begin{cases} 1/4, & -1 \leq x \leq 3, \\ 0, & \text{otherwise.} \end{cases}$. Define the random variable Y by $Y = h(X) = X^2$.

(a) Find $E[X]$ and $Var[X]$.
(b) Find $h(E[X])$ and $E[h(X)]$.
(c) Find $E[Y]$ and $Var[Y]$.

Solution

(a) $E[X] = \int\limits_{-\infty}^{+\infty} x f_X(x)\mathrm{d}x = \int\limits_{-1}^{3} \dfrac{x}{4}\mathrm{d}x = 1.$; $Var[X] = E[X^2] - E[X]^2 = \dfrac{7}{3} - 1^2 = \dfrac{4}{3}.$

(b) $h(E[X]) = E[X]^2 = 1.$; $E[h(X)] = E[X^2] = \int\limits_{-\infty}^{+\infty} x^2 f_X(x)\mathrm{d}x = \int\limits_{-1}^{3} \dfrac{x^2}{4}\mathrm{d}x = \dfrac{7}{3}.$

(c) $E[Y] = E[h(X)] = \dfrac{7}{3}.$

$Var[Y] = E[Y^2] - E[Y]^2 = E\left[X^4\right] - E\left[X^2\right]^2 = \dfrac{61}{5} - \dfrac{49}{9} = \dfrac{304}{45}.$

Problem 2.33.

Random variable X has CDF $F_X(x) = \begin{cases} 0, & x \leq 0, \\ \dfrac{x}{2}, & 0 < x \leq 2, \\ 1, & x > 2 \end{cases}$.

(a) What is $E[X]$?
(b) What is $Var[X]$?

Solution

(a) The PDF of X is $f_X(x) = \begin{cases} \dfrac{1}{2}, & 0 < x \leq 2, \\ 0, & \text{otherwise.} \end{cases}$. So, $E[X] = \int\limits_{-\infty}^{+\infty} x f_X(x)\mathrm{d}x = \int\limits_{0}^{2} \dfrac{x}{2}\mathrm{d}x = 1.$

(b) $Var[X] = E[X^2] - E[X]^2 = \dfrac{1}{3}$.

Problem 2.34.

Y is an exponential random variable with variance $Var[Y] = 25$.

(a) What is the PDF of Y?
(b) What is $E[Y^2]$?
(c) What is $P[Y > 5]$?

Solution

We have Y is an exponential random variable with $\lambda > 0$ has PDF:

$$f_Y(y) = \begin{cases} \lambda e^{-\lambda y}, & y \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

In addition, the mean and variance of Y is $E[Y] = \dfrac{1}{\lambda}$ and $Var[Y] = \dfrac{1}{\lambda^2}$.

(a) Since $Var[Y] = 25 \Rightarrow \lambda = \dfrac{1}{5} \Rightarrow f_Y(y) = \begin{cases} \dfrac{1}{5}e^{-\frac{y}{5}}, & y \geq 0, \\ 0, & \text{otherwise.} \end{cases}$.

(b) $E[Y^2] = Var[Y] + E[Y]^2 = 50$.

(c) $P[Y > 5] = 1 - F_Y[5] = 1 - (1 - e^{-1}) = \dfrac{1}{e}$.

Problem 2.35.

X is a continuous uniform $(-5, 5)$ random variable.

(a) What is the PDF $f_X(x)$?
(b) What is the CDF $F_X(x)$?
(c) What is $E[X]$?
(d) What is $E[X^5]$?
(e) What is $E[e^X]$?

Solution

(a) PDF: $f_X(x) = \begin{cases} \dfrac{1}{10}, & -5 \leq x \leq 5, \\ 0, & \text{otherwise.} \end{cases}$.

(b) CDF: $F_X(x) = \begin{cases} 0, & x \leq a, \\ \dfrac{x+5}{10}, & -5 < x < 5, \\ 1, & x \geq b. \end{cases}$.

(c) $E[X] = 0$.

(d) $E[X^5] = \int\limits_{-\infty}^{+\infty} x^5 f_X(x)\mathrm{d}x = \int\limits_{-5}^{5} \dfrac{x^5}{10}\mathrm{d}x = 0$.

(e) $E[e^X] = \int\limits_{-\infty}^{+\infty} e^x f_X(x)\mathrm{d}x = \int\limits_{-5}^{5} \dfrac{e^x}{10}\mathrm{d}x = \dfrac{e^5 - e^{-5}}{10}$.

Problem 2.36.
X is a uniform random variable with expected value $\mu_X = 7$ and variance $Var[X] = 3$. What is the PDF of X?

Solution

- Since $\mu_X = 7 \Rightarrow a + b = 14$ and $Var[X] = 3 \Rightarrow b - a = 6$. So, $a = 10,\ b = 4$.
- Therefore, the PDF of X: $f_X(x) = \begin{cases} \frac{1}{6}, & 4 \le x \le 10, \\ 0, & \text{otherwise.} \end{cases}$.

Problem 2.37.
The peak temperature T, as measured in degrees Fahrenheit, on a July day in New Jersey is the Gaussian $(85, 10)$ random variable. What is $P[T > 100]$, $P[T < 60]$, and $P[70 \le T \le 100]$?

Solution

- We denote $y = \dfrac{x - \mu}{\sigma} = \dfrac{x - 85}{10}$ to get

$$F_T(t) = \int_{-\infty}^{\frac{x-82}{10}} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \mathrm{d}y = \Phi\left(\frac{x - 85}{10}\right)$$

- Here, we get

$$P[T > 100] = 1 - P[T \le 100] = 1 - F_T(100) = 1 - \Phi(1.5)$$
$$P[T < 60] = P[T \le 60] = \Phi(-2.5) = 1 - \Phi(2.5)$$
$$P[70 \le T \le 100] = F_T(100) - F_T(70) = \Phi(1.5) - \Phi(-1.5) = 2\Phi(1.5) - 1$$

Problem 2.38.
What is the PDF of Z, the standard normal random variable?

Solution

$$\varphi_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

Problem 2.39.
X is a Gaussian random variable with $E[X] = 0$ and $P[|X| \le 10] = 0.1$. What is the standard deviation σ_X?

Solution

- Since $E[X] = 0 \Rightarrow \mu = 0$.
- Besides, $P[|X| \le 10] = P[-10 \le X \le 10] = F_X(10) - F_X(-10) = 2\Phi\left(\frac{10}{\sigma}\right) - 1 = 0.1$

$\Leftrightarrow \Phi\left(\frac{10}{\sigma}\right) = 0.55 \Rightarrow \frac{10}{\sigma} \approx 0.13 \Leftrightarrow \sigma \approx 77$

Problem 2.40.
The peak temperature T, in degrees Fahrenheit, on a July day in Antarctica is a Gaussian random variable with a variance of 225. With probability $\frac{1}{2}$, the temperature T exceeds 10 degrees. What is $P[T > 32]$, the probability the temperature is above freezing? What is $P[T < 0]$? What is $P[T > 60]$?

Solution

- T is a Gaussian random variable with mean μ and variance $Var[T] = 225$. Here, we have $P[T > 10] = 0.5 \Leftrightarrow$ the median is 10 degrees, so $\mu = 10$.
- $P[T > 32] = P\left[\frac{T-10}{15} > \frac{32-10}{15}\right] = P\left[Z > \frac{22}{15}\right] = P[Z > 1.47] = 0.071.$

Problem 2.41.
The voltage X across a 1Ω resistor is a uniform random variable with parameters 0 and 1. The instantaneous power is $Y = X^2$. Find the CDF $F_Y(y)$ and the PDF $f_Y(y)$ of Y.

Solution

- CDF of X $F_X(x) = \begin{cases} 0, & x < 0, \\ x, & 0 \le x \le 1, \\ 1, & x > 1. \end{cases}$.
- Since $Y = X^2$, so CDF of Y is $F_Y(y) = \begin{cases} 0, & y < 0, \\ \sqrt{y}, & 0 \le y \le 1, \\ 1, & y > 1. \end{cases}$.
- The PDF of Y is derivative of CDF, so $f_Y(y) = \begin{cases} \frac{1}{2\sqrt{y}}, & 0 \le y \le 1, \\ 0, & \text{otherwise}. \end{cases}$.

Problem 2.42.
X is uniform random variable with parameters 0 and 1. Find a function $g(x)$ such that the PDF of $Y = g(X)$ is $f_Y(y) = \begin{cases} 3y^2, & 0 \le y \le 1, \\ 0, & \text{otherwise}. \end{cases}$

Solution

- CDF of X and Y are : $F_X(x) = \begin{cases} 0, & x < 0, \\ x, & 0 \le x \le 1, \\ 1, & x > 1. \end{cases}$ and $F_Y(y) = \begin{cases} 0, & y < 0, \\ y^3, & 0 \le y \le 1, \\ 1, & y > 1. \end{cases}$
- This implies $X = Y^3 \Leftrightarrow Y = \sqrt[3]{X}$. That means $g(x) = \sqrt[3]{x}$.

Problem 2.43.

X is a uniform random variable with parameters -5 and 5. Given the event $B = |X| \leq 3$,

(a) Find the conditional PDF, $f_{X|B}(x)$.
(b) Find the conditional expected value, $E[X|B]$.
(c) What is the conditional variance, $Var[X|B]$?

Solution

(a) We have PDF of X: $f_X(x) = \begin{cases} \frac{1}{10}, & -5 \leq x \leq 5, \\ 0, & \text{otherwise.} \end{cases}$.

Therefore, $P[B] = P[-3 \leq X \leq 3] = F_X[3] - F_X[-3] = 0.6$.

So, the conditional PDF: $f_{X|B}(x) = \begin{cases} \frac{f_X(x)}{P[B]}, & -3 \leq x \leq 3, \\ 0, & \text{otherwise.} \end{cases} = \begin{cases} \frac{1}{6}, & -3 \leq x \leq 3, \\ 0, & \text{otherwise.} \end{cases}$.

(b) $E[X|B] = \frac{1}{P[B]} \int\limits_{-\infty}^{+\infty} x f_X(x) \mathrm{d}x = \frac{1}{P[B]} \int\limits_{-3}^{3} \frac{x}{6} \mathrm{d}x = 0.$

(c) $E[X^2|B] = \frac{1}{P[B]} \int\limits_{-\infty}^{+\infty} x^2 f_X(x) \mathrm{d}x = \frac{1}{P[B]} \int\limits_{-3}^{3} \frac{x^2}{6} \mathrm{d}x = 5.$

So, $Var[X|B] = E[X^2|B] - E[X|B]^2 = 5$.

Problem 2.44.

Y is an exponential random variable with parameter $\lambda = 0.2$. Given the event $A = \{Y < 2\}$,

(a) Find the conditional PDF, $f_{Y|A}(x)$.
(b) Find the conditional expected value, $E[Y|A]$.

Solution

(a) We have PDF of Y: $f_Y(y) = \begin{cases} 0.2 \times e^{-0.2x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$.

Therefore, $P[A] = P[Y < 2] = F_Y[2] = 1 - e^{-0.2 \times 2} = 1 - e^{0.4}$.

$$\Rightarrow f_{Y|A}(y) = \begin{cases} \frac{f_Y(y)}{P[A]}, & x \geq 0, \\ 0, & x < 0. \end{cases} = \begin{cases} \frac{0.2 \times e^{-0.2x}}{1 - e^{0.4}}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

(b) $E[Y|A] = \frac{1}{P[A]} \int\limits_{-\infty}^{+\infty} y f_Y(y) \mathrm{d}y = \frac{1}{P[A]} \int\limits_{0}^{+\infty} y \times 0.2 \times e^{-0.2x} \mathrm{d}y = \frac{E[Y]}{P[A]} = \frac{5}{1 - e^{0.4}}.$

Problem 2.45.

The cumulative distribution function of the continuous random variable X is

$$F_X(x) = \begin{cases} 0, & x \leq 0, \\ \frac{1}{2} - k\cos x, & 0 < x \leq \pi, \\ 1, & x > \pi. \end{cases}$$

(a) What is k?

(b) What is $\mathrm{P}\left[0 < x < \frac{\pi}{2}\right]$?

(c) What is $E[X]$?

Solution

(a) For X to be a continuous random variable, there should not be any discontinuity in all values of X. That implies $F_X(0) = 0 = \frac{1}{2} - k$ and $F_X(\pi) = 1 = \frac{1}{2} + k \Rightarrow k = 0.5$.

(b)

$$P\left[0 < x < \frac{\pi}{2}\right] = F_X\left(\frac{\pi}{2}\right) - F_X(0) = \frac{1}{2}$$

(c)

$$E[X] = \int_{-\infty}^{+\infty} x f_X(x)\mathrm{d}x = \int_{0}^{\pi} \frac{1}{2} x \sin x \mathrm{d}x = \pi$$

Problem 2.46.

The cumulative distribution function of the continuous random variable X is

$$F(x) = \begin{cases} 0, & x \leq -a, \\ A + B\arcsin\frac{x}{a}, & -a < x < a, \\ 1, & x \geq a. \end{cases}$$

(a) What are A and B?

(b) What is PDF $f_X(x)$?

Solution

(a) For X to be a continuous random variable, there should not be any discontinuity in all values of X. That implies

$$F_X(-a) = 0 = A - \frac{B\pi}{2} \quad ; \quad F_X(a) = 1 = A + \frac{B\pi}{2} \Rightarrow A = \frac{1}{2}, B = \frac{1}{\pi}$$

(b) $F(x) = \begin{cases} 0, & x \leq -a, \\ \frac{1}{2} + \frac{1}{\pi}\arcsin 2a, & -\frac{1}{2} \leq x \leq \frac{1}{2}, \\ 1, & x \geq a. \end{cases} \Rightarrow f_X(x) = \begin{cases} \frac{2}{\sqrt{1-4x^2}}, & -\frac{1}{2} \leq x \leq \frac{1}{2}, \\ 0, & \text{otherwise.} \end{cases}$

Problem 2.47.

The cumulative distribution function of the continuous random variable X is

$$F(x) = a + b \arctan x, \; (-\infty < x < +\infty)$$

(a) What are a and b?
(b) What is PDF $f_X(x)$?
(c) What is $P[-1 < X < 1]$?

Solution

(a) We have $\lim\limits_{x\to-\infty} F(x) = a - \dfrac{b\pi}{2} = 0$ and $\lim\limits_{x\to-\infty} F(x) = a + \dfrac{b\pi}{2} = 1 \Rightarrow a = \dfrac{1}{2}, \; b = \dfrac{1}{\pi}$.

(b) From (a), we have CDF of X: $F(x) = \dfrac{1}{2} + \dfrac{1}{\pi} \arctan x, \; (-\infty < x < +\infty)$

$\Rightarrow f_X(x) = \dfrac{1}{\pi(x^2+1)}, \quad (-\infty < x < +\infty)$

(c) $P[-1 < X < 1] = F[1] - F[-1] = \dfrac{3}{4} - \dfrac{1}{4} = \dfrac{1}{2}$.

Problem 2.48.

The cumulative distribution function of the continuous random variable X is

$$F(x) = \frac{1}{2} + \frac{1}{\pi} \arctan \frac{x}{2}$$

What is the value of x_1 such that $P(X > x_1) = \dfrac{1}{4}$?

Solution

Since $P(X > x_i) = 1 - F(x_i) = \dfrac{1}{4} \Leftrightarrow F(x_i) = \dfrac{3}{4} \Leftrightarrow x = 2$.

Problem 2.49.

The continuous random variable X has probability density function

$$f(x) = \begin{cases} k \sin 3x, & x \in \left(0; \dfrac{\pi}{3}\right), \\ 0, & x \notin \left(0; \dfrac{\pi}{3}\right). \end{cases}$$

Use the PDF to find

(a) the constant k.
(b) CDF $F_X(x)$.
(c) $P\left[\dfrac{\pi}{6} \le X < \dfrac{\pi}{3}\right]$.

Solution

(a) We have $\int\limits_{-\infty}^{+\infty} f(x)\mathrm{d}x = \int\limits_{0}^{\frac{\pi}{3}} k \sin 3x \mathrm{d}x = 2k$. Since the result should be 1, so $k = \dfrac{1}{2}$.

(b) CDF of X: $F_X(x) = \begin{cases} 0, & x \leq 0, \\ -\dfrac{1}{2}\cos 3x, & x \in \left(0; \dfrac{\pi}{3}\right), \\ 1, & x \geq \dfrac{\pi}{3}. \end{cases}$.

(c) $P\left[\dfrac{\pi}{6} \leq X < \dfrac{\pi}{3}\right] = F_X\left(\dfrac{\pi}{3}\right) - F_X\left(\dfrac{\pi}{6}\right) = \dfrac{1}{2}.$

Problem 2.50.

The continuous random variable X has PDF $f(x) = \dfrac{c}{e^x + e^{-x}}$. What is $E[X]$?

Solution

Since $\int\limits_{-\infty}^{+\infty} f(x)\mathrm{d}x = 1$, so

$$\int\limits_{-\infty}^{+\infty} \frac{c}{e^x + e^{-x}}\mathrm{d}x = c\int\limits_{-\infty}^{+\infty} \frac{e^x}{e^{2x}+1}\mathrm{d}x = c\int\limits_{-\infty}^{+\infty} \frac{\mathrm{d}e^x}{e^{2x}+1} = c\arctan e^x|_{-\infty}^{+\infty} = c\pi \Rightarrow c = \frac{1}{\pi}.$$

Here,

$$E[X] = \int\limits_{-\infty}^{+\infty} xf(x)\mathrm{d}x = \frac{1}{\pi}\int\limits_{-\infty}^{+\infty} \frac{x}{e^x + e^{-x}}\mathrm{d}x = ...$$

Problem 2.51.

The continuous random variable X has PDF $f(x) = ae^{-|x|}$, $(-\infty < x < +\infty)$. Define the random variable Y by $Y = X^2$.

(a) What is a?
(b) What is the CDF $F_Y(x)$?
(c) What is $E[X]$? What is $Var[X]$?

Solution

(a) Since $f(x)$ is a PDF of x, then

- $f(x) \geq 0\ \forall x \Rightarrow ae^{-|x|} \geq 0\ \forall x \in \mathbb{R} \Leftrightarrow a \geq 0$
- $\int\limits_{-\infty}^{+\infty} f(x)\mathrm{d}x = 1 \Leftrightarrow \int\limits_{-\infty}^{+\infty} ae^{-x}\mathrm{d}x = 1 \Leftrightarrow 2\int\limits_{0}^{+\infty} ae^{-x}\mathrm{d}x = 1 \Leftrightarrow 2a = 1 \Leftrightarrow a = \dfrac{1}{2}$

So $a = \dfrac{1}{2}$ and $f(x) = \dfrac{1}{2}e^{-|x|}$

(b) The CDF of Y is $F_Y(x) = P(Y < x) = P(X^2 < x)$
If $x \leq 0$ then $F_Y(x) = P(\varnothing) = 0$
If $x > 0$ then $F_Y(x) = P(-\sqrt{x} \leq x \leq \sqrt{x})$...

(c) We have $E(X) = \int\limits_{-\infty}^{+\infty} xf(x)\mathrm{d}x = \int\limits_{-\infty}^{+\infty} \frac{1}{2}xe^{-|x|}\mathrm{d}x = 0$

$$Var[X] = \int\limits_{-\infty}^{+\infty} x^2 f(x)\mathrm{d}x - \mu_x^2 = \int\limits_{-\infty}^{+\infty} \frac{1}{2}x^2 e^{-|x|}\mathrm{d}x - 0^2$$
$$= -\int\limits_{0}^{+\infty} x^2 \mathrm{d}e^x = -x^2e^{-x} \mid_0^{+\infty} + \int\limits_{0}^{+\infty} e^{-x}2x\mathrm{d} - x^2e^{-x} \mid_0^{+\infty} + \int\limits_{0}^{+\infty} e^{-x}2x\mathrm{d}x$$
$$= -2\int\limits_{0}^{+\infty} x\mathrm{d}e^{-x} = -2xe^{-x} \mid_0^{+\infty} + 2\int\limits_{0}^{+\infty} e^{-x}\mathrm{d}x = 0 - 2e^{-x} \mid_0^{+\infty} = 2$$

Problem 2.52.

X is an exponential random variable with the PDF $f_X(x)$ is $f(x) = \begin{cases} 5e^{-5x}, & x > 0, \\ 0, & x \leq 0. \end{cases}$.

(a) What is $E[X]$?
(b) What is $P[0.4 < X < 1]$?

Solution

(a) $E[X] = \frac{1}{5} = 0.2$.

(b) The CDF of X is: $F_X(x) = \begin{cases} 0, & x < 0, \\ 1 - e^{-5x}, & x \geq 0. \end{cases}$.

So, $P[0.4 < X < 1] = F_X(1) - F_X(0.4) = e^{-2} - e^{-5} \approx 0.129$.

Problem 2.53.

X is a Gaussian random variable with $E[X] = 0$ and $\sigma_X = 0.4$.

(a) What is $P[X > 3]$?
(b) What is the value of c such that $P[3 - c < X < 3 + c] = 0.9$?

Solution

Since $E[X] = 0 \Rightarrow \mu = 0$ and $\sigma_X = 0.4$, we denote $Z = \frac{X}{0.4}$. Hence, we have

$$F_X(x) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\frac{x}{0.4}} e^{-\frac{z^2}{2}}\mathrm{d}z = \Phi\left(\frac{x}{0.4}\right)$$

(a) $P[X > 3] = 1 - F_X(3) = 1 - \Phi(7.5) \approx$

(b) $P[3 - c < X < 3 + c] = F_X(3 - c) - F_X(3 + c) = \Phi\left(\frac{3 + c}{0.4}\right) - \Phi\left(\frac{3 - c}{0.4}\right) = 0.9$

Chapter 3

Important Probability Distributions.

3.1 Theory.

3.1.1 Some Discrete Probability Distributions.

- Binomial distribution $\mathcal{B}(n, p)$.
  - X is a Bernoulli random variable if the PMF of X has the form
  $$\begin{cases} 1-p, & x=0, \\ p, & x=1, \\ 0, & \text{otherwise.} \end{cases}$$
  where the parameter p is in the range $0 < p < 1$.
  - The mean and variance of the Bernoulli random variable X are $\mu = E[X] = p$ and $\sigma^2 = Var[X] = p(1-p)$.
  - X is a binomial random variable if the PMF of X has the form
  $$\begin{cases} C_n^x p^x (1-p)^{n-x}, & x = 0, 1, 2, ..., n, \\ 0, & \text{otherwise.} \end{cases}$$
  where $0 < p < 1$ and n is an integer such that $n \geq 1$.
  - The probability distribution of this discrete random variable is called the binomial distribution, and is denoted by $\mathcal{B}(n, p)$ (or $X \sim \mathcal{B}(n, p)$).
  - The mean and variance of the binomial random variable X are $\mu = E[X] = np$ and $\sigma^2 = Var[X] = npq,\ q = 1-p$.
- Discrete uniform distribution
  - X is a discrete uniform random variable if the probability distribution of X has the form

X	1	2	...	n
p	$\frac{1}{n}$	$\frac{1}{n}$	...	$\frac{1}{n}$

  - For the discrete uniform random variable X: $E[X] = \dfrac{n+1}{2},\ Var[X] = \dfrac{n^2-1}{12}$.
- Poisson distribution $\mathcal{P}(\lambda)$.
  - X is a Poisson random variable if the PMF of X has the form
  $$\begin{cases} e^{-\lambda} \times \dfrac{\lambda^x}{x!}, & x = 0, 1, 2, ... \\ 0, & \text{otherwise.} \end{cases}$$

where the parameter λ is in the range $\lambda > 0$.
- Both the mean and the variance of the Poisson distribution $\mathcal{P}(\lambda)$ are λ.

Remark

Here are some examples of experiments for which the random variable X can be modeled by the Poisson random variable:

1. The number of calls received by a switchboard during a given period of time.
2. The number of customer arrivals at a checkout counter during a given minute.
3. The number of machine breakdowns during a given day.
4. The number of traffic accidents at a given intersection during a given time period.

In each example, X represents the number of events that occur in a period of time or space during which an average of λ such events can be expected to occur. The only assumptions needed when one uses the Poisson distribution to model experiments such as these are that the counts or events occur randomly and independently of one another.

- Approximation of Binomial Distribution by a Poisson Distribution.
  - Let X be a binomial random variable with probability distribution $\mathcal{B}(n, p)$.
  - When $n \to \infty, p \to 0$, and $np \to \mu$ as $n \to \infty$ remains constant, $\mathcal{B}(n, p) \to \mathcal{P}(\lambda)$ as $n \to \infty$.

Remark

The Poisson distribution provides a simple, easy-to-compute, and accurate approximation to binomial probabilities when n is large and $\lambda = np$ is small, preferably with $np < 7$.

3.1.2 Some Continuous Probability Distributions.

- Continuous uniform distribution.
  - X is a uniform $\mathcal{U}[a, b]$ random variable if the PDF of X is

$$f_X(x) = \begin{cases} \dfrac{1}{b-a}, & x \in [a, b], \\ 0, & x \notin [a, b]. \end{cases}$$

  - The CDF of X is

$$F_X(x) = \begin{cases} 0, & x \le a, \\ \dfrac{x-a}{b-a}, & a < x \le b, \\ 1, & x > b. \end{cases}$$

  - The expected value of X is $\mu = E[X] = \dfrac{a+b}{2}$.
  - The variance of X is $\sigma^2 = Var[X] = \dfrac{(b-a)^2}{12}$.

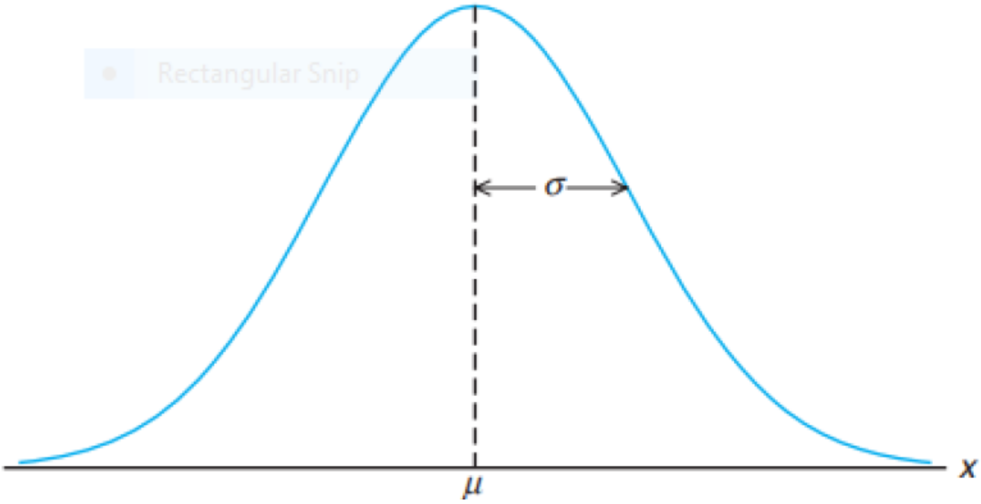
- Exponential distribution.
  - X is an exponential $exp(\lambda)$ random variable if the PDF of X is

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

    where the parameter $\lambda < 0$.
  - CDF of X is

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

  - The expected value of X is $\mu = E[X] = \dfrac{1}{\lambda}$.
  - The variance of X is $\sigma^2 = Var[X] = \dfrac{1}{\lambda^2}$.
- Normal distribution $\mathcal{N}(\mu, \sigma)$.
  - The most important continuous probability distribution in the entire field of statistics is the normal distribution. Its graph, called the normal curve, is the bell-shaped curve of the graph below, which approximately describes many phenomena that occur in nature, industry, and research.



  - **Introduction:** The normal distribution is often referred to as the Gaussian distribution, in honor of Karl Friedrich Gauss (1777–1855), who also derived its equation from a study of errors in repeated measurements of the same quantity. The mathematical equation for the probability distribution of the normal variable depends on the two parameters μ and σ, its mean and standard deviation, respectively. Hence, we denote the normal distribution by $\mathcal{N}(\mu, \sigma)$.
  - The PDF of the normal random variable X, with mean μ and variance σ^2, is

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

  - Once μ and are specified, the normal curve is completely determined. For example, if $\mu = 50$ and $\sigma = 5$, then the ordinates $\mathcal{N}(50, 5)$ can be computed for various values of x and the curve drawn.
  - The mean and variance of the normal random variable are μ and σ^2, respectively. Hence, the standard deviation is σ.

- Standard normal distribution.
  - Put $Z = \dfrac{X = \mu}{\sigma}$. If X is a normal random variable with mean μ and variance σ^2 then Z is seen to be a normal random variable with mean 0 and variance 1.
  - The distribution of a normal random variable with mean 0 and variance 1 is called a standard normal distribution $\mathcal{N}(0,1)$.
  - The PDF of the standard normal random variable Z is

$$\varphi_Z(z) = f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

  - The CDF of the standard normal random variable Z is

$$\Phi_Z(z) = \frac{1}{\sigma\sqrt{2\pi}} \int\limits_{-\infty}^{z} e^{-\frac{t^2}{2}} \mathrm{d}t$$

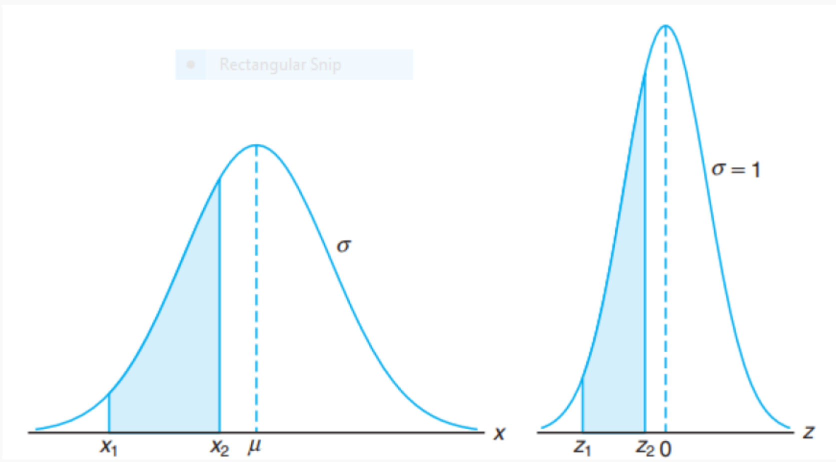
  - If X is a normal random variable $\mathcal{N}(\mu, \sigma)$, the CDF of X is $F_X(x) = \Phi\left(\dfrac{x-\mu}{\sigma}\right)$.
  - The probability that X is in the interval (x_1, x_2) is

$$P[x_1 < x < x_2] = \Phi\left(\frac{x_2 - \mu}{\sigma}\right) - \Phi\left(\frac{x_1 - \mu}{\sigma}\right)$$

Note that: $\Phi(-z) = 1 - \Phi(z)$

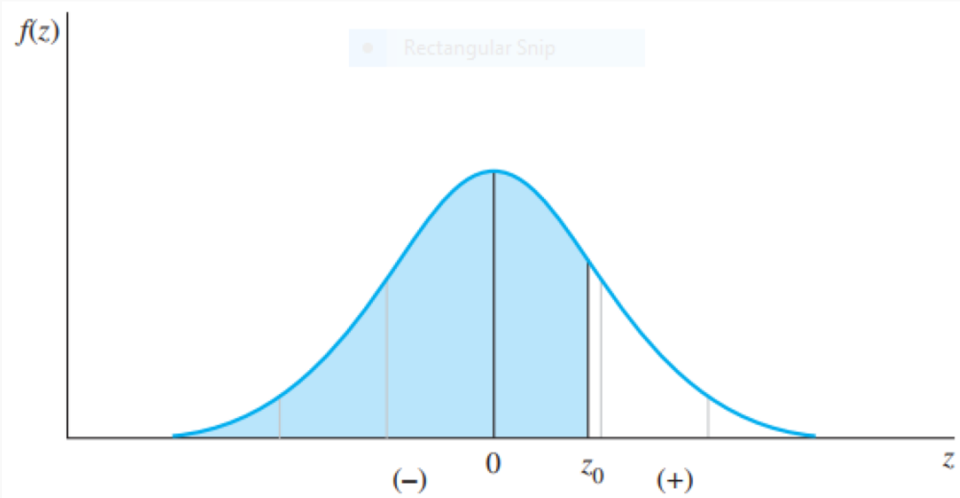
Remark

(a) In using the above formulae, we transform values of a norm random variable, X, to equivalent values of the standard normal random variable, Z. For a sample value x of the random variable X, the corresponding sample value of Z is $z = \dfrac{x-\mu}{\sigma}$ or equivalently $x = \mu + z\sigma$.
The original and transformed distributions are illustrated in the figure below. Since all the values of X falling between x_1 and x_2 have corresponding z values between z_1 and z_2, the area under the X-curve between the ordinates x_1 and x_2 in the figure equals the area under the Z-curve between the transformed ordinates z_1 and z_2.



Remark

(b) The probability distribution for Z, shown below, is called the standardized normal distribution because its mean is 0 and its standard deviation is 1. Values of Z on the left side of the curve are negative, while values on the right side are positive. The area under the standard normal curve to the left of a specified value of Z say, z_0 is the probability $P[Z \leq z_0]$. This cumulative area is shown as the shaded area in the figure below.



- Binomial approximation to Normal distribution
  - If X is a binomial random variable with mean $\mu = np$ and variance $\sigma^2 = npq$, then the limiting form of the distribution of $Z = \dfrac{x - np}{\sqrt{npq}}$ as $n \to +\infty$, is the standard normal distribution $\mathcal{N}(0, 1)$.
  - Let X be a binomial random variable with n trials and probability p of success. The probability distribution of X is approximated using a normal curve with $\mu = np$ and $\sigma = \sqrt{npq}$ and

$$P[x_1 < x < x_2] = \sum_{k=x_1}^{x_2} C_n^k p^k (1-p)^{n-k} \approx \Phi\left(\frac{x_2 + 0.5 - \mu}{\sigma}\right) - \Phi\left(\frac{x_1 - 0.5 - \mu}{\sigma}\right)$$

and the approximation will be good if np and $n(1-p)$ are greater than or equal to 5.

3.2 Answer of Exercises.

3.2.1 Some Discrete Probability Distributions.

Problem 3.1.

see Problem 2.6. In a package of M&Ms, Y, the number of yellow M&Ms, is uniformly distributed between 5 and 15.

(a) What is the PMF of Y?
(b) What is $P[Y < 10]$?
(c) What is $P[Y > 12]$?
(d) What is $P[8 \leq Y \leq 12]$?

Solution

(a) PMF of Y: $\begin{cases} P_Y(y) = \frac{1}{11}, & \text{if } y = \overline{5, 15}, \\ 0, & \text{otherwise.} \end{cases}$
(b) $P[Y < 10] = \sum_{i=5}^{i=9} P_Y(i) = \dfrac{5}{11}$.
(c) $P[Y > 12] = \dfrac{3}{11}$.
(d) $P[8 \leq Y \leq 12] = \dfrac{5}{11}$.

Problem 3.2.

see Problem 2.7. When a conventional paging system transmits a message, the probability that the message will be received by the pager it is sent to is p. To be confident that a message is received at least once, a system transmits the message n times.

(a) Assuming all transmissions are independent, what is the PMF of K, the number of times the pager receives the same message?
(b) Assume $p = 0.8$. What is the minimum value of n that produces a probability of 0.95 of receiving the message at least once?

Solution

(a) $P_K(k) = C_n^k p^k (1-p)^{n-k}$ with $k = 0..n$.
(b) $P[\text{at least one}] = 1 - P_K[0] = 1 - 0.2^n$. From the assumption, $P[\text{at least one}] = 0.95$, that implies $0.2^n = 0.05 \Leftrightarrow n = \dfrac{\ln 0.05}{\ln 0.2} = 1.86 \approx 2$.

Problem 3.3.

see Problem 2.8. When a two-way paging system transmits a message, the probability that the message will be received by the pager it is sent to is p. When the pager receives the message, it transmits an acknowledgment signal (ACK) to the paging system. If the paging system does not receive the ACK, it sends the message again.

(a) What is the PMF of N, the number of times the system sends the same message?
(b) The paging company wants to limit the number of times it has to send the same message. It has a goal of $P[N \leq 3] \geq 0.95$. What is the minimum value of p necessary to achieve the goal?

Solution

(a) The n same messages are sent when the pager doesn't receive the message n times. So, the PMF of N is: $P_N(n) = \begin{cases} (1-p)^{n-1}.p, & \text{if n} = 1, 2, ... \\ 0, & \text{otherwise.} \end{cases}$

(b) We have

$$P[N \leq 3] = 1 - P[N > 3] = 1 - \sum_{n=4}^{+\infty} P_N(n) = 1 - p\left[(1-p)^3 + (1-p)^4 + ... + (1-p)^n\right]$$

$$= 1 - p. \lim_{n \to +\infty} (1-p)^3 . \frac{1-(1-p)^n}{1-(1-p)} = 1 - (1-p)^3.$$

Since $P[N \leq 3] \geq 0.95 \Rightarrow (1-p)^3 \leq 0.05 \Leftrightarrow p \geq 1 - 0.05^{1/3} \approx 0.63$.

Problem 3.4.

Four microchips are to be placed in a computer. Two of the four chips are randomly selected for inspection before assembly of the computer. Let X denote the number of defective chips found among the two chips inspected. Find the probability mass and distribution function of X if

(a) Two of the microchips were defective.
(b) One of the microchips was defective.
(c) None of the microchips was defective.

Solution

(a) The values of X are $0, 1, 2$ and $P_X(x) = \dfrac{C_2^x \times C_2^{2-x}}{C_4^2}$. So, PMF: $P_X(x) = \begin{cases} \frac{1}{6}, & x = 0, \\ \frac{2}{3}, & x = 1, \\ \frac{1}{6}, & x = 2. \\ 0, & \text{otherwise.} \end{cases}$

(b) The values of X are $0, 1$ and $P_X(x) = \dfrac{C_1^x \times C_3^{3-x}}{C_4^2}$. So, PMF: $P_X(x) = \begin{cases} 0.5, & x = 0, \\ 0.5, & x = 1, \\ 0, & \text{otherwise.} \end{cases}$

(c) The values of X is 0 and $P_X(0) = \dfrac{C_0^0 \times C_4^2}{C_4^2} = 1$. So, PMF: $P_X(x) = \begin{cases} 1, & x = 0, \\ 0, & \text{otherwise.} \end{cases}$

Problem 3.5.

A four engine plane can fly if at least two engines work.

(a) If the engines operate independently and each malfunctions with probability q, what is the probability that the plane will fly safely?
(b) A two engine plane can fly if at least one engine works and if an engine malfunctions with probability q, what is the probability that plane will fly safely?
(c) Which plane is the safest?

Solution

(a) Let X is a random variable shows the number of engine function correctly. Here X has binomial distribution with $n = 4, p = 1 - q$.

So, the probability that the plane will fly safely is $P[X \geq 2] = \sum\limits_{k=2}^{4} C_4^k (1-q)^k q^{4-k}$.

(b) Let X is a random variable shows the number of engine function correctly. Here X has binomial distribution with $n = 2, p = 1 - q$.

So, the probability that the plane will fly safely is $P[X \geq 1] = \sum\limits_{k=1}^{2} C_2^k (1-q)^k q^{2-k}$.

(c) Which plane is the safest?

Problem 3.6.

A rat maze consists of a straight corridor, at the end of which is a branch; at the branching point the rat must either turn right or left. Assume 10 rats are placed in the maze, one at a time.

(a) If each is choosing one of the two branches at random, what is the distribution of the number that turn right?
(b) What is the probability at least 9 will turn the same way?

Solution

(a) Consider x be the number of right turns taken by the rats.

Therefore, $n = 10, p = q = \dfrac{1}{2}$.

(b) $P_X[\text{at least 9 rats takes the same turn}] = P[X \geq 9] + P[X \leq 1]$
$= P_X(0) + P_X(1) + P_X(9) + P_X(10) = 0.0214$.

Problem 3.7.

A student who is trying to write a paper for a course has a choice of two topics, A and B. If topic A is chosen, the student will order 2 books through inter-library loan, while if topic B is chosen, the student will order 4 books. The student feels that a good paper necessitates receiving and using at least half the books ordered for either topic chosen.

(a) If the probability that a book ordered through inter-library loan actually arrives on time is 0.9 and books arrive independently of one another, which 2 topics should the student choose to maximize the probability of writing a good paper?
(b) What if, the arrival probability is only 0.5 instead of 0.9?

Solution

(a) For topic A $(n = 2, p = 0.9, q = 0.1)$, so the probability of topic A is

$$P[A] = P_A[a \geq 1] = P_A(1) + P_A(2) = 0.99.$$

For topic B $(n = 4, p = 0.9, q = 0.1)$, so the probability of topic B is

$$P[B] = P_B[b \geq 1] = P_B(2) + P_B(3) + P_B(4) = 0.9963.$$

From that, we can implies that choosing topic B is better.

(b) Similarly, we can compute $P[A] = 0.75$ and $P[B] = 0.6875$. From that, we can implies that choosing topic A is better.

Problem 3.8.

The number of phone calls at a post office in any time interval is a Poisson random variable. A particular post office has on average 2 calls per minute.

(a) What is the probability that there are 5 calls in an interval of 2 minutes?
(b) What is the probability that there are no calls in an interval of 30 seconds?
(c) What is the probability that there are no less than one call in an interval of 10 seconds?

Solution

Problem 3.9.

see Problem 1.41. An airline sells 200 tickets for a certain flight on an airplane that has only 198 seats because, on the average, 1 percent of purchasers of airline tickets do not appear for the departure of their flight. Determine the probability that everyone who appears for the departure of this flight will have a seat.

Solution

- Let X = "number of people who do not appear for their flight". We can assume that the passengers independently decide to show up or not, and we'll consider not appearing for the flight a success (for the airline!).
- Then X is the number of successes in a sequence of n = 200 independent Bernoulli trials with probability of success $p = 1\% = 0.01$. So, $X\ Bin(n = 200, p = 0.01)$.
- Because $n = 200$ is large, $p = 0.01$ is small and $np = 2$ is moderate, we can approximate the $Bin(n = 200, p = 0.01)$ distribution by Poisson $(= np = 2)$ distribution.
- Now, everyone will get a seat if and only if at least 2 passengers do not appear, i.e., $X \geq 2$.
- Therefore, required probability:

$$P[X \geq 2] = 1 - P[X \leq 1] = 1 - P[0] - P[1] \approx 1 - e^{-2}\frac{2^0}{0!} - e^{-2}\frac{2^1}{1!} = 1 - 3e^{-2} = 0.59.$$

3.2.2 Some Continuous Probability Distributions.

Problem 3.10.

see Problem 2.52. X is an exponential random variable with the PDF $f_X(x)$ is

$$f(x) = \begin{cases} 5e^{-5x}, & x > 0, \\ 0, & x \le 0. \end{cases}$$

(a) What is $E[X]$?
(b) What is $P[0.4 < X < 1]$?

Solution

(a) $E[X] = \dfrac{1}{5} = 0.2$.

(b) The CDF of X is: $F_X(x) = \begin{cases} 0, & x < 0, \\ 1 - e^{-5x}, & x \ge 0. \end{cases}$.

So, $P[0.4 < X < 1] = F_X(1) - F_X(0.4) = e^{-2} - e^{-5} \approx 0.129$.

Problem 3.11.

see Problem 2.53. X is a Gaussian random variable with $E[X] = 0$ and $\sigma_X = 0.4$.

(a) What is P[X > 3]?
(b) What is the value of c such that $P[3 - c < X < 3 + c] = 0.9$?

Solution

Problem 3.12.

Let X be an exponential random variable with parameter and define $Y = [X]$, the largest integer in X, (ie. $[x] = 0$ for $0 \le x < 1$, $[x] = 1$ for $1 \le x < 2$ etc.)

(a) Find the probability function for Y.
(b) Find $E(Y)$.
(c) Find the distribution function of Y.
(d) Let Y represent the number of periods that a machine is in use before failure. What is the probability that the machine is still working at the end of 10th period given that it does not fail before 6th period?

Solution

(a) Since X is an exponential random variable, so CDF of X: $F_X(x) = \begin{cases} 0, & x \le 0, \\ 1 - e^{-\lambda x}, & x > 0. \end{cases}$.

$\Rightarrow P_Y(y) = P[y \le X < y+1] = F_X(y+1) - F_X(y) = e^{-\lambda y} - e^{-\lambda(y+1)} = e^{-\lambda y}(1 - e^{-\lambda})$.
Since Y is a discrete random variable, so PMF of Y is $P_Y(y) = ce^{-\lambda y}(1 - e^{-\lambda})$.
Because $\sum P_X(x) = 1$, we have $c = ...$

(b) According to expected value of geometric distribution function: $E(Y) = \dfrac{1}{1 - e^{-1}}$.

(c) Geometric distribution.

(d)

$$P[Y \geq 10 \mid Y \geq 6] = \frac{P[Y \geq 10]}{P[Y \geq 6]} = \frac{1 - P[Y \leq 9]}{1 - P[Y \leq 5]} = \frac{(1 - 1 + e^{-1})^9}{(1 - 1 + e^{-1})^5} = e^{-4}.$$

Problem 3.13.

Starting at 5:00 am, every half hour there is a flight from San Francisco airport to Los Angeles International Airport. Suppose that none of these planes sold out and that they always have room for passengers. A person who wants to fly LA arrives at the airport at a random time between 8:45 - 9:45 am. Find the probability that she waits at most 10 minutes and at least 15 minutes.

Solution

- Let the passenger arrive at the airport X minutes pass 8:45. Then X is a uniform random variable over the interval (0, 60). Hence the PDF of X is given by

$$f_X(x) = \begin{cases} \dfrac{1}{60}, & 0 \leq x \leq 60. \\ 0, & \text{otherwise.} \end{cases}$$

- Now the passenger waits at most 10 minutes if she arrives between 8:50 and 9:00 or 9:20 and 9:30; that is, if $5 < X < 15$ or $35 < X < 45$. Hence,

$$P[A] = \int_5^{15} \frac{1}{60} \mathrm{d}x + \int_{35}^{45} \frac{1}{60} \mathrm{d}x = \frac{1}{3}.$$

- The passenger waits at least 15 minutes if she arrives between 9:00 and 9:15 or 9:30 and 9:45; that is, if $15 < X < 30$ or $45 < X < 60$. Hence,

$$P[B] = \int_{15}^{30} \frac{1}{60} \mathrm{d}x + \int_{45}^{60} \frac{1}{60} \mathrm{d}x = \frac{1}{2}.$$

Problem 3.14.

X is a Gaussian random variable with $E[X] = 0$ and $P[|X| \leq 10] = 0.1$. What is the standard deviation σ_X?

Solution

- X is a Gaussian random variable with zero mean but unknown variance. We do know, however, that $P[|X| \leq 10] = 0.1$
- We can find the variance Var[X] by expanding the above probability in terms of the $\Phi(\cdot)$ function:

$$P[-10 \leq X \leq 10] F_X(10) - F_X(-10) = \Phi\left(\frac{10}{\sigma_X}\right) - \left(1 - \Phi\Phi\left(\frac{10}{\sigma_X}\right)\right) = 2\Phi\left(\frac{10}{\sigma_X}\right) - 1$$

- This implies $\Phi\left(\dfrac{10}{\sigma_X}\right) = 0.55 \Rightarrow \dfrac{10}{\sigma_X} \approx 0.125 \Rightarrow \sigma_X \approx 80.$

An extra problem

The probability of having a new born boy is 0.51. Take the survey of 1000 new born babies in a hospital. Calculate the probability that the number of boys is smaller than the number of girls.

Solution

- We denote A is the number of new born boys. Apply the binomial distribution formula with $n = 1000$; $p = 0.51$; $q = 0.49$. For the number of boy is smaller than that of girls, we have $0 \leq A \leq 499$.
- Apply the normal approximation to binomial distribution, we have $\mu = np = 510$; $\sigma = \sqrt{npq} = 0.7\sqrt{510}$. Therefore,

$$
\begin{aligned}
P[0 \leq A \leq 499] &= \sum_{k=0}^{499} C_{1000}^{k} \times p^{k} \times q^{1000-k} \\
&\simeq \Phi\left(\frac{499 + 0.5 - \mu}{\sigma}\right) - \Phi\left(\frac{499 - 0.5 - \mu}{\sigma}\right) = \Phi(-0.664) - \Phi(-0.727) \\
&= \Phi(0.727) - \Phi(0.664) \approx 0.76730 - 0.74537 = 0.02193
\end{aligned}
$$

Mid-term exam questions.

MI2020E - 20201

Question 1 (3.0 point) Suppose that 15 cards are selected at random from a regular deck of 52 playing cards.

(a) If it is known that one ace has been selected, what is the probability that at least two aces have been selected?
(b) If it is known that the ace of hearts has been selected, what is the probability that at least two aces have been selected?

Question 2 (2.0 points) A student who is trying to write a paper for a course has a choice of two topics, A and B. If the topic A is chosen, the student will order 2 books through inter-library loan, while if the topic B is chosen, the student will order 4 books. The student feels that a good paper necessitates receiving and using at least half the books ordered for either topic chosen. If the probability that a book ordered through inter-library loan actually arrives on time is 0.85 and books arrive independently of one another, which topic should the student choose to maximize the probability of writing a good paper?

Question 3 (2.0 points) A box contains 10 red balls, 15 white balls, and 5 blue balls. Suppose that 3 balls are selected at random one at a time and let X be the number of colors missing from the 3 selected balls. Determine the probability mass function of X.

Question 4 (3.0 points) A light bulb has a lifetime Y which is exponential distribution with parameter $\lambda = 0.5/year$. Define $Z = [Y]$, the largest integer in Y , (i.e. $[y] = 0$ for $0 \leq y < 1$, $[y] = 1$ for $1 \leq y < 2$ etc.)

(a) What is $P(Z = 1)$?
(b) If the bulb is still working after 2 years then it is replaced by a new one. Let W be the time that the bulb is in use. Find the cumulative distribution function of W.

Solution

Question 1.

(a)
- We denote A is a random variable of the number of aces selected. Here, the probability that need calculating is

$$P[A \geq 2 \mid A \geq 1] = \frac{P[A \geq 2]P[A \geq 1 \mid A \geq 2]}{P[A \geq 1]} = \frac{P[A \geq 2]}{P[A \geq 1]}$$

(Since $P[A \geq 1 \mid A \geq 2] = 1$).

- Since the number of aces in a deck is 4, so the probability

$$P[A \geq 1] = 1 - P[A = 0] = 1 - \frac{C_{48}^{15}}{C_{52}^{15}} = \frac{344}{455}$$

$$P[A \geq 2] = P[A \geq 1] - P[A = 1] = \frac{344}{455} - \frac{C_4^1 \times C_{48}^{14}}{C_{52}^{15}} = \frac{2518}{7735}$$

- Hence, $P[A \geq 2 \mid A \geq 1] = \dfrac{P[A \geq 2]}{P[A \geq 1]} \approx 0.443.$

(b)
- We denote A is a random variable of the number of aces selected. Here, the probability that need calculating is $P[A \geq 2 \mid \text{Ace of heart is selected}]$
- We have

$$P[\text{Ace of heart is selected}] = \frac{C_{51}^{14}}{C_{52}^{15}} = \frac{15}{52}$$

$$P[A \geq 2 \cap \text{Ace of heart is selected}] = \frac{C_3^1 \times C_{48}^{13} + C_3^2 \times C_{48}^{12} + C_3^3 \times C_{48}^{11}}{C_{52}^{15}} = \frac{1119}{5188}$$

- Hence,

$$P[A \geq 2 \mid \text{Ace of heart is selected}] = \frac{P[A \geq 2 \cap \text{Ace of heart is selected}]}{P[\text{Ace of heart is selected}]} \approx 0.627$$

Question 2.

- For topic A $(n = 2, p = 0.85, q = 0.15)$, so the probability of topic A is

$$P[A] = P_A[a \geq 1] = P_A(1) + P_A(2) = C_2^1 \times 0.85^1 \times 0.15^1 + C_2^2 \times 0.85^2 \times 0.15^0 = 0.9775.$$

- For topic B $(n = 4, p = 0.85, q = 0.15)$, so the probability of topic B is

$$P[B] = P_B[b \geq 1] = P_B(2) + P_B(3) + P_B(4) = 0.9880.$$

- From that, we can implies that choosing topic B is better.

Question 3.

- We denote X is the random variable of the number of missing colors. Hence,

$$P[3] = 0 \quad ; \quad P[0] = \frac{C_{10}^1 \times C_{15}^1 \times C_5^1}{C_{30}^3} = \frac{75}{406} \quad ; \quad P[2] = \frac{C_{10}^3 + C_{15}^3 + C_5^3}{C_{30}^3} = \frac{117}{812}$$

$$\Rightarrow P[1] = 1 - P[0] - P[2] - P[3] = \frac{545}{812}.$$

- Hence, the PMF of X is: $P_X(x) = \begin{cases} \frac{75}{406}, & x = 0, \\ \frac{545}{812}, & x = 1, \\ \frac{117}{812}, & x = 2, \\ 0, & x = 3. \end{cases}$.

Question 4.

(a)
- Since Y is a exponential distribution random variable, the CDF function of Y is $F_X(x) = 1 - e^{-\lambda x}$.

- We have $P[Z = 1] = P[1 \leq Y < 2] = F(2) - F(1) = e^{-0.5} - e^{-1} \approx 0.239$.

(b) CDF of W: $P_W(w) = \begin{cases} 0, & w < 0, \\ 1 - e^{-0.5w}, & 0 \leq w \leq 2, \\ 1, & w > 2. \end{cases}$

MI2026 - 20192

Question 1 (2.5 point) From a group of 3 excellent students, 4 good students, and 5 average students a committee of size 4 is randomly selected. Find the probability that the committee will consist of

(a) One good student.
(b) The good students are more than other types of students.

Question 2 (2.5 points) One box contains 5 products from machine A and 3 from machine B. It is known from past experience that 1% and 2% of the products made by each machine, respectively, are defective. Suppose that two products are randomly selected from the box.

(a) What is the probability that one of them is defective?
(b) If two products were chosen randomly and found to be defective, what is the probability that they were made by machine B?

Question 3 (2.5 points) Continuous random variable X has probability density function

$$f_X(x) = \begin{cases} cx(3-x), & 0 \leq x \leq 3, \\ 0, & \text{otherwise.} \end{cases}$$

(a) Find c?
(b) Find the cumulative distribution function of $Y = 2X - 5$.

Question 4 (2.5 points) BK plays a game, which assigns a number (Z) that is the continuous random variable with the density function

$$f_Z(z) = \begin{cases} \frac{1}{5}, & 0 \leq z \leq 5, \\ 0, & \text{otherwise.} \end{cases}$$

If the game assigns a number less than or equal to k, then he loses 10 dollar, on the other hand, if the game assigns a number larger than k, then he will gain 10 dollar.

(a) Find the expected profit of the game.
(b) Find the variance of the profit.
(c) If you were to play this game 10 times, what is the probability that you gain 20 dollars?

Solution

Question 1.

(a) $P[A] = \dfrac{C_4^1 \times C_8^3}{C_{12}^4} = \dfrac{229}{495}$.

(b) $P[B] = \dfrac{C_4^2 \times C_8^2 + C_4^3 \times C_8^1 + C_4^4 \times C_8^0}{C_{12}^4} = \dfrac{67}{165}$.

Question 2.

(a)
- The probability of choosing one product in the box that produced by machine A or B is $P[A] = \dfrac{5}{8} = 0.625$; $P[B] = \dfrac{3}{8} = 0.375$.
- Because A and B is two disjoint random variable and $A \cup B$ is an event space, so the probability of one selected product is defective is $P[D] = P[A]P[D|A] + P[B]P[D|B] = \dfrac{11}{800}$.
- Hence, the probability need calculating is $C_2^1 \times P[D] \times (1 - P[D]) = 0.0271$.

(b)
- The probability for two selected products are defective is:

$$P[DD] = C_5^2(0.01)^2 + C_3^2(0.02)^2 + C_5^1 \times C_3^1 \times 0.01 \times 0.02 = \frac{13}{2500}$$

- Apply the Bayes' formula, the probability need calculating is

$$P[B|DD] = \frac{P[B] \times P[DD|B]}{P[DD]} = \frac{3}{13}$$

Question 3.

(a) Since X is a random variable whose the given PDF, we have

$$\int_{-\infty}^{+\infty} f_X(x)\mathrm{d}x = 1 \Rightarrow c\int_0^3 x(3-x)\mathrm{d}x = 1 \Leftrightarrow \frac{9}{2}c = 1 \Leftrightarrow c = \frac{2}{9}$$

(b)
- We have

$$F_Y(y) = P[Y \le y] = P\left[X < \frac{Y+5}{2}\right] = \int_0^{\frac{y+5}{2}} f_X(x)\mathrm{d}x = \left(-\frac{2x^3}{27} + \frac{x^2}{3}\right)\Bigg|_0^{\frac{y+5}{2}}$$

$$\Rightarrow F_Y(y) = -\frac{y^3}{108} - \frac{y^2}{18} + \frac{5y}{36} + \frac{25}{27}$$

- Since $0 \le X \le 3 \Leftrightarrow 0 \le \dfrac{Y+5}{2} \le 3 \Leftrightarrow -5 \le Y \le 1$.
- Hence, the CDF of Y is

$$F_Y(y) = \begin{cases} 0, & y < -5, \\ -\dfrac{y^3}{108} - \dfrac{y^2}{18} + \dfrac{5y}{36} + \dfrac{25}{27}, & -5 \le y \le 1, \\ 1, & y > 1. \end{cases}$$

Question 4.

(a) We denote the profit by T. We have

$$E[T] = 10 \times P[T = 10] + (-10)P[T = -10] = 10 \times \frac{5-k}{5} - 10 \times \frac{k}{5} = 10 - 4k.$$

(b) We have

$$E[T^2] = 10^2 \times P[T = 10] + (-10)^2 P[T = -10] = 100 \times \frac{5-k}{5} + 100 \times \frac{k}{5} = 100$$

$$\Rightarrow Var[T] = E[T^2] - E[T]^2 = 100 - (10 - 4k)^2 = 80k - 16k^2.$$

(c) You gain \$20 only if you win exactly 6 times. Apply Bernoulli formula:

$$P(T = 20) = C_{10}^6 \times P[T = 10]^6 \times P[T = -10]^4 = \frac{210 \times (5-k)^6 \times k^4}{5^{10}}.$$

CLB HTHT - 20212

Question 1 (2.5 point) You are mating two purebred Husky dogs, the generation of puppies can have black, gray and white fur with a ratio of 1 : 1 : 2. This Husky couple is about to welcome 7 puppies to be born. Calculate the probability of among those 7:

(a) Having white-fur puppy.
(b) Having gray-fur puppies, known that having white-fur puppies.

Question 2 (2.5 points) An electronics supermarket is famous for its coffee-jelly-making machine. Customers who come here to choose which type of jelly-making machines to buy often pay attention to 3 criteria A, B, C which are the color, the cream maker and the trolley wheel, respectively. Supermarket sales figures show that:

- $P(A) = P(B) = P(C) = 0.6$
- $P(A+B) = 0.7 \quad ; \quad P(C+A) = 0.8 \quad ; \quad P(B+C) = 0.75$
- $P(A+B+C) = 0.85$

Calculate the probability that

(a) Customers choose all 3 criteria.
(b) Customers choose exactly 1 out of 3 criteria.

Question 3 (2.5 points) Gunner A has a 0.6 probability of hitting the target, assuming the gunner shoots 3 times. Let X be the number of hits on the target after 3 shots.

(a) Find the probability distribution table of X.
(b) Calculate mean and variance of random variable $Y = X^2 - 2$. Find the cumulative distribution function of Y.

Question 4 (2.5 points) The length of a machine part (in mm) is a random variable X with probability density function

$$f_X(x) = Ae^{\frac{-32+8x-x^2}{18}}$$

(a) Calculate $E(X)$, $Var(X)$.
(b) The part is qualified when its length deviates from the mean by no more than 1mm. Calculate the probability of having a qualified machine part.
(c) Assume one shipment has 20 independent machine parts. Find the probability that there are at least 15 qualified part.

Appendix

Area under the Normal Curve: $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{x} e^{-\frac{t^2}{2}} \mathrm{d}t.$

$$\Phi\left(\frac{1}{3}\right) = 0.6306 \quad ; \quad \Phi\left(\frac{2}{3}\right) = 0.7475 \quad ; \quad \Phi(1) = 0.8413$$

Solution

Question 1.

(a) $P(W) = 1 - P(\overline{W}) = 1 - 0.5^7 \approx 0.992.$

(b) The probability need calculating: $P[G|W] = \dfrac{P[GW]}{P[W]}$. Besides,

$$P[GW] = 1 - P[\overline{G} + \overline{W}] = 1 - (P[\overline{G}] + P[\overline{W}] + P[\overline{GW}]) = 1 - (0.5^7 + 0.75^7 + 0.25^7) \approx 0.859.$$

Hence, $P[G|W] = \dfrac{P[GW]}{P[W]} = 0.866.$

Question 2.

(a) We denote D: "Customers choose all 3 criteria". We have $P[D] = P[ABC]$ and:

- $\star$ $P[AB] = P[A] + P[B] - P[A + B] = 0.5.$
- $\star$ $P[BC] = P[B] + P[C] - P[B + C] = 0.45.$
- $\star$ $P[CA] = P[C] + P[A] - P[C + A] = 0.4.$

$\Rightarrow P[ABC] = P[A + B + C] - P[A] - P[B] - P[C] + P[AB] + P[BC] + P[CA] = 0.4.$

(b) We denote E: "Customers choose exactly 1 out of 3 criteria". We have

$$P[E] = P[A\overline{B}\overline{C}] + P[\overline{A}B\overline{C}] + P[\overline{A}\overline{B}C].$$

$$\begin{aligned} P[A\overline{B}\overline{C}] &= P[\overline{B}\overline{C}] - P[\overline{A}\overline{B}\overline{C}] = P[\overline{B + C}] - P[\overline{A + B + C}] \\ &= (1 - P[B + C]) - (1 - P[A + B + C]) = P[A + B + C] - P[B + C] = 0.1. \end{aligned}$$

Similarly, $P[\overline{A}B\overline{C}] = 0.05 \quad ; \quad P[\overline{A}\overline{B}C] = 0.15 \quad \Rightarrow \quad P[E] = 0.3.$

Question 3.

(a) PD of X

X	**0**	**1**	**2**	**3**
p	0.064	0.288	0.432	0.216

(b) With $Y = X^2 - 2$, we have PD of Y:

Y	**-2**	**-1**	**2**	**7**
p	0.064	0.288	0.432	0.216

Here, we have $E[Y] = 1.96 \quad ; \quad Var[Y] = 8.4384.$

So, the CDF of Y: $\begin{cases} 0, & y \le 2, \\ 0.064, & -2 < y \le -1, \\ 0.352, & -1 < y \le 2, \\ 0.784, & 2 < y \le 7, \\ 1, & y > 7. \end{cases}$

Question 4.

(a) $f_X(x) = Ae^{\frac{-32+8x-x^2}{18}} = Ae^{-\frac{16}{18}}e^{-\frac{(x-4)^2}{18}} \Rightarrow E[X] = 4;\ Var[X] = \dfrac{18}{2} = 9.$

(b) We have $E[X] - 1 \le X \le E[X] + 1 \Leftrightarrow 3 \le X \le 5$

$\Rightarrow P[3 \le X \le 5] = \Phi\left(\dfrac{5-4}{3}\right) - \Phi\left(\dfrac{3-4}{3}\right) = 2\Phi\left(\dfrac{1}{3}\right) - 1 \approx 0.261.$

(c) $P = \sum_{k=15}^{20} C_{20}^{k} \times 0.261^{k} \times (1-0.261)^{20-k} = 6.8 \times 10^{-6}$.

MI2026 - 20202 - Test 1

Question 1 (2 point) A rare disease exists with which only 5 in 1000 is affected. There is a test for the disease that a correct positive result (patient actually has the disease) occurs 96% of the time while a false positive result (patient does not have the disease) occurs 2% of the time.

(a) What is the prior distribution of the disease and the posterior distribution of the disease given the positive result?
(b) Is the test good? Explain that counter-intuitive result.

Question 2 (2 points) A student has to make a choice between two topic A or B to write his final thesis. There are 6 and 4 essays in the topic A and B, respectively. It is mandatory that students need to pass at least half the number of required essays to reach his final thesis. Previous statistics shows that each essay in topic A has 3% failure and topic B has that figure of 6%. Which topic should that student choose and why?

Question 3 (4 points) Suppose that a bivariate random variable (X, Y) has joint density function

$$f_{X,Y}(x,y) = \begin{cases} \frac{2}{3}(2x+y), & (x,y) \in [0;1] \times [0;1], \\ 0, & \text{otherwise.} \end{cases}$$

(a) Find the probability $P[X > Y]$.
(b) Find marginal distributions of Y and conditional distribution of X, given $Y = \frac{1}{2}$.
(c) Find the expectation of X and Y.
(d) Find the correlation coefficient between X and Y.

Question 4 (2 points) Assume that the lifetime X of an electronic device follows the exponential distribution and it is expected to last 200 hours.

(a) What is probability that the device lasts longer than 150 hours?
(b) If the device has lasted 150 hours then what is the probability that it will last at least 100 hours more? Does it depend on how long the device has lasted?

Solution

Question 1.

(a)
- We denote I: "The subject is infected", P and N: "The test result is Positive/Negative".
- The prior distribution of the disease: $P[I] = 0.005$.
- Given that there is a positive result, the probability that the subject is infected. we have $P[I|P] = \dfrac{P[I] \times P[P|I]}{P[P]} = \dfrac{0.005 \times 0.96}{0.98} \approx 0.0049$.

(b)
- We denote T and F is the test result is true or false.
- We have:
  $P[TP] = P[I] \times P[TP|I] = 0.005 \times 0.96 = 0.0048$;
  $P[FP] = P[\bar{I}] \times P[FP|\bar{I}] = (1 - P[I]) \times P[FP|\bar{I}] = 0.995 \times 0.02 = 0.0199$.

  - When a test is positive, the probability that the test result is true is:

$$P[TP|P] = \frac{P[TP]}{P[TP] + P[FP]} = 0.194$$

  - In conclusion, the test result is **inaccurate**.

Question 2.

- We denote f is the failure of an essay.
- The probability that the student can pass topic A is

$$P[A \geq 3] = \sum_{k=3}^{6} C_6^k \times (1-f)^k \times f^{(6-k)} = \sum_{k=3}^{6} C_6^k \times 0.97^k \times 0.03^{(6-k)} = 0.999988.$$

- The probability that the student can pass topic A is

$$P[B \geq 2] = \sum_{k=2}^{4} C_4^k \times (1-f)^k \times f^{(4-k)} = \sum_{k=2}^{4} C_4^k \times 0.94^k \times 0.06^{(4-k)} = 0.999175.$$

- Since $P[A \geq 3] > P[B \geq 2]$, the student should choose topic A to write final thesis.

Question 3.

(a)
(b)
(c)
(d)

Question 4.

- Since $E[X] = 200 \Rightarrow \lambda = \dfrac{1}{E[X]} = \dfrac{1}{200}$. Hence, CDF of X: $F_X(x) = 1 - e^{-\frac{x}{200}}$ with $x \geq 0$.

(a) The probability that the device lasts longer than 150 hours is

$$P[X > 150] = 1 - F_X[150] = e^{-\frac{150}{200}} = e^{-0.75} \approx 0.4723.$$

(b) The probability that the device lasts longer than 250 hours if it lasts longer than 150 hours is

$$\begin{aligned} P[X > 250|X > 150] &= \frac{P[X > 250] \times P[X > 150|X > 250]}{P[X > 150]} = \frac{P[X > 250]}{P[X > 150]} \\ &= \frac{e^{-\frac{250}{200}}}{e^{-\frac{150}{200}}} = e^{-\frac{100}{200}} \approx 0.6065. \end{aligned}$$

We can easily see that, $P[X > 250|X > 150] = P[X > 100]$, that implies it doesn't depend on how long the device has lasted.

MI2026 - 20202 - Test 2

Question 1 (2.5 point) Suppose traffic engineers have coordinated the timing of three traffic lights to encourage a run of green lights. In particular, the timing was designed so that with probability 0.85 a driver will find the second light to have the same color as the first. The probability 0.7 a driver will find the third light to have the same color as the second. Assuming the first light is equally likely to be red or green.

(a) What is the probability $P[G_3]$ that the third light is green?
(b) What is $P[W]$, the probability that you wait for at least one light?
(c) What is $P[G_1|G_3]$, the conditional probability of a green first light given a green third light?

Question 2 (2.5 points) Suppose that for the general population, 1 in 5500 people carries the human HIV virus. A test for the presence of HIV yields either a positive (+) or negative (-) response. Suppose the test gives the correct answer 99% of the time. What is $P[-|H]$, the conditional probability that person tests negative given that the person does have the HIV virus? What is $P[H|+]$, the conditional probability that a randomly chosen person has the HIV virus given that the person tests positive?

Question 3 (2.5 points) The number of hits at a website in any time interval is a Poisson random variable. A particular site has on average $\lambda = 1$ hits per minute.

(a) What is the probability that there are no hits in an interval of 5 minutes, expected valued of hits in 5 minutes?
(b) What is the probability that there are no more than two hits in an interval of 5 minutes?

Question 4 (2.5 points) Suppose your TOEIC score X on a test is Gaussian $(650, 100)$ random variable.

(a) What is probability that your TOEIC score greater than 814.45.
(b) Given your TOEIC score greater than 650, what is probability your TOEIC score smaller than 814.45.
(c) Find c such that $P(X < c) = 0.05$.

Solution

Question 1.

(a) $P[G_3] = P[G_1G_2G_3] + P[G_1R_2G_3] + P[R_1G_2G_3] + P[R_1R_2G_3]$
$= 0.5 \times 0.85 \times 0.7 + 0.5 \times 0.15 \times 0.3 + 0.5 \times 0.15 \times 0.7 + 0.5 \times 0.85 \times 0.3 = 0.5$
(b) $P[W] = 1 - P[G_1G_2G_3] = 1 - 0.5 \times 0.85 \times 0.7 = 0.7025.$
(c) $P[G_1|G_3] = \dfrac{P[G_1] \times P[G_3|G_1]}{P[G_3]} = \dfrac{0.5 \times (0.5 \times 0.85 \times 0.7 + 0.5 \times 0.15 \times 0.3)}{0.5} = 0.32.$

Question 2.

- Given $P[H] = \dfrac{1}{5500}$; $\quad P[+|H] = 0.99 \Rightarrow P[\overline{H}] = \dfrac{5499}{5500}$; $\quad P[-|\overline{H}] = 0.99$
- We have $P[-|H] = 1 - P[+|H] = 0.01$

- Apply the Bayes' theorem:

$$P[H|+] = \frac{P[H] \times P[+|H]}{P[H] \times P[+|H] + P[\overline{H}] \times P[+|\overline{H}]} = \frac{0.99 \times \frac{1}{5500}}{0.99 \times \frac{1}{5500} + 0.01 \times \frac{5499}{5500}} = 0.0177$$

Question 3.

(a) – Since the number of hits in 5 minutes is a Poisson random variable, we have

$$X \sim \text{Poisson}(5 \times 1) \Rightarrow \lambda = 5$$

– The PMF function is: $P[X = k] = e^{-5}\frac{5^k}{k!}$. Hence, $P[X = 0] = e^{-5}\frac{5^0}{0!} = e^{-5}$.
– The expected value in 5 minutes is $E[X] = \lambda = 5$.

(b) $P[X \leq 2] = P[X = 0] + P[X = 1] + P[X = 2] = e^{-5}\frac{5^0}{0!} + e^{-5}\frac{5^1}{1!} + e^{-5}\frac{5^2}{2!} \approx 0.125.$

Question 4.

(a) We denote $Z = \frac{X - 650}{100}$, so the CDF of Z is

$$F_Z(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{2}} \mathrm{d}t = \Phi(z) = \Phi\left(\frac{x - 650}{100}\right)$$

$$\Rightarrow P[X > 814.45] = 1 - F_Z\left(\frac{814.45 - 650}{100}\right) = 1 - \Phi(1.6445) \approx 0.94950$$

(b) $P[650 < X < 814.45] = \Phi\left(\frac{814.45 - 650}{100}\right) - \Phi\left(\frac{650 - 650}{100}\right) \approx 0.44950.$

(c) Since X is a continuous random variable, so $P[X = c] = 0$, that implies

$$P[X < c] = P[X \leq c] = F_X(c) = F_Z\left(\frac{c - 650}{100}\right) = \Phi\left(\frac{c - 650}{100}\right).$$

Using the given table of standard normal distribution, we have

$$\Phi\left(\frac{c - 650}{100}\right) = \Phi(0) = 0.5 \Rightarrow \frac{c - 650}{100} = 0 \Leftrightarrow c = 650.$$

MI2026 - 20212

Question 1 (2 point) I randomly pick up 6 cards from a regular deck of 52 playing cards.

(a) Find the probability that 3 Jacks have been selected if I tell you that there is a Jack in my hand?
(b) Recalculate the probability if I tell you that there is a Jack of hearts in my hand?

Question 2 (2 points) At first, 1 marble is randomly taken out from the box A containing 4 green marbles and 3 blue marbles, then put in the box B. Next, 1 green marble is put into the box B (so there are 2 marbles in the box B now). Finally, a marble is randomly taken out from the box B.

(a) What is the probability that the marble taken out is green?
(b) Given the fact that the marble taken out is green, how much is the increase of the probability that the marble taken from box A to box B is green?

Question 3 (2 points) Suppose we have 2 coins. Coin C_1 comes up head with probability 0.3 and that of coin C_2 is 0.9. We repeat this process 4 times: choose a coin with equal probability then flip it. Suppose X is the number of heads after 4 flips.

(a) Find the distribution of X.
(b) What is $E[X]$ and $Var[X]$?

Question 4 (2 points) A random variable X has the probability density function $f_X(x) = Cxe^{-3x^2}$ in the interval $[0; +\infty)$ and 0 outside.

(a) Find the normalizing constant C and the probability $P[0 < X < 3]$.
(b) Find the CDF of the random variable $Y = \min(X, 3)$ and the expectation Y.

Question 5 (2 points) A bivariate random variable (X, Y) has joint PDF of the form $f_{X|Y}(x, y) = Kx^2y$, (x, y) in the rectangle $D = 0 < x < 1, x^2 < y < 1$ and 0 outside.

(a) Find the normalizing constant K and the probability $P[0 < X < 0.5; 0 < Y < 0.5]$.
(b) Find the PDF of Y and conditional PDF $X|Y = y$.

Solution

Question 1.

(a)
- We denote J is the number of selected Jacks.
- There is a Jack in your hands means that there will be at least 1 Jack has been selected. So, the probability of choosing 3 Jacks when 1 Jack has been selected is $P[J = 3|J \geq 1]$. Apply the Bayes' formula:

$$P[J = 3|J \geq 1] = \frac{P[J = 3] \times P[J \geq 1|J = 3]}{P[J \geq 1]}$$

- To choose exact 3 Jacks, we need to choose exact 3 Jacks in 4 Jacks, and 3 other cards from 48 remaining cards $\Rightarrow P[J = 3] = \dfrac{C_4^3 \times C_{48}^3}{C_{48}^6} = \dfrac{184}{54145}$.

  - We have $P[J \geq 1] = 1 - P[J = 0] = 1 - \dfrac{C_{48}^6}{C_{52}^6} \approx 0.6028$ and $P[J \geq 1|J = 3] = 1$
  - So, the probability need calculating is

$$P[J = 3|J \geq 1] = \frac{P[J = 3] \times P[J \geq 1|J = 3]}{P[J \geq 1]} = \frac{P[J = 3]}{P[J \geq 1]} = 0.0056.$$

(b)
  - To choose exact 3 Jacks from the deck, given that the Jack of hearts is selected, we have to choose exact 2 Jacks and 3 other cards from 51 remaining cards.
  - The probability of the Jack of hearts is selected is $P[J_{\text{heart}}] = \dfrac{C_{51}^5}{C_{52}^6} = \dfrac{3}{26}$.
  - The probability of choosing a Jack of hearts and exact 2 other Jacks and 3 other cards from 51 remaining cards is $P[J = 3 \cap J_{\text{heart}}] = \dfrac{C_3^2 \times C_{48}^3}{C_{52}^6} = \dfrac{138}{54145}$.
  - Hence, the probability need calculating is

$$P[J = 3|J_{\text{heart}}] = \frac{P[J = 3 \cap J_{\text{heart}}]}{P[J_{\text{heart}}]} = 0.0221.$$

Question 2.

- We denote G_A, B_A is the green and blue marble selected from box A, G_B is the green marble selected from box B, G is the selected green marble.

(a) Apply the total probability formula:

$$P[G] = P[G|G_A G_B] \times P[G_A G_B] + P[G|B_A G_B] \times P[B_A G_B] = 1 \times \frac{C_4^1}{C_7^1} + \frac{1}{2} \times \frac{C_3^1}{C_7^1} = \frac{11}{14}.$$

(b)
  - The probability of the taken marble from A to B is green given that the taken marble is green is $P[G_A G_B|G]$.
  - The probability of the taken marble from A to B is green is $P[G_A G_B] = \dfrac{C_4^1}{C_7^1} = \dfrac{4}{7}$
  - Apply the Bayes' formula: $P[G_A G_B|G] = \dfrac{P[G_A G_B] \times P[G|G_A G_B]}{P[G]} = \dfrac{\frac{4}{7} \times 1}{\frac{11}{14}} = \dfrac{8}{11}$.
  - Hence, the increase of the probability is $\Delta P = P[G_A G_B|G] - P[G_A G_B] = \dfrac{12}{77}$

Question 3.

(a)
  - Since X is the number of heads after 4 flips, we have the range of X is $X \in \{0, 1, 2, 3, 4\}$.
  - We denote H is the random variable that the chosen coin has head up. So, the probability of a flip has a head up is:

$$P[H] = P[HC_1] + P[HC_2] = P[H_1] \times P[C_1] + P[H_2] \times P[C_2] = 0.15 + 0.45 = 0.6.$$

  - To have X heads, we need to have exact $(4 - X)$ tails. That means:

$$P[X = k] = C_4^k \times P[H]^k \times P[\overline{H}]^{(1-k)}.$$

  - Hence, the distribution of X is

X	0	1	2	3	4
$P[X]$	0.0256	0.1536	0.3456	0.3456	0.1296

(b) $E[X] = \sum_{x=0}^{4} x \times P[X = x] = \sum_{x=0}^{4} x \times C_4^k \times P[H]^k \times P[\overline{H}]^{(1-k)} = 2.4.$

$E[X^2] = \sum_{x=0}^{4} x \times P[X = x] = \sum_{x=0}^{4} x \times C_4^k \times P[H]^k \times P[\overline{H}]^{(1-k)} = 6.72$

$\Rightarrow Var[X] = E[X^2] - E[X]^2 = 0.96.$

Question 4.

(a) For the $f_X(x)$ to be a PDF function, we have $\int_{-\infty}^{+\infty} f_X(x)\mathrm{d}x = 1$. Hence,

$$\int_{-\infty}^{+\infty} f_X(x)\mathrm{d}x = c\int_{0}^{+\infty} xe^{-3x^2}\mathrm{d}x = c \times \left.\frac{-e^{-3x^2}}{6}\right|_0^{+\infty} = \frac{c}{6}$$

For $\frac{c}{6} = 1 \Leftrightarrow c = 6$.

(b)
- The CDF of X is $F_X(x) = \begin{cases} 0, & x < 0, \\ 1 - e^{-3x^2}, & 0 \le x. \end{cases}$.
- Since $Y = \min(X, 3)$, then $Y = X$ when $X \le 3$ and $Y = 3$ when $X > 3$.
- Hence, the CDF of Y is $F_Y(k) = P[Y \le k] = P[3 \le k] = 1$ when $k > 3$ and $F_Y(k) = P[Y \le k] = P[X \le k] = F_X(x)$ when $k \le 3$.
- In conclusion, $F_Y(y) = \begin{cases} 0, & y < 0, \\ 1 - e^{-3y^2}, & 0 \le y \le 3, \\ 1, & y > 3. \end{cases}$
- The expected value of Y is

$$\begin{aligned} E[Y] &= E[Y|0 < X \le 3] \times P[X \le 3] + E[Y|X > 3] \times P[X > 3] \\ &= \int_0^3 6x^2e^{-3x^2}\mathrm{d}x + 3 \times (1 - F_X(3)) \approx 0.5117. \end{aligned}$$

Question 5.

(a)

(b)

Chapter 4

Pairs of random variables.

4.1 Theory.

4.1.1 Joint Cumulative Distribution Function.

- **Definition:** The joint cumulative distribution function of random variables X and Y is

$$\boxed{F_{X,Y}(x,y) = P[X < x, Y < y]}$$

- **Properties:**
  - $0 \leq F_{X,Y}(x,y) \leq 1$.
  - $F_X(x) = F_{X,Y}(x,\infty)$.
  - $F_Y(x,y) = F_{X,Y}(\infty,y)$
  - $F_{X,Y}(-\infty,y) = F_{X,Y}(x,-\infty) = 0$.
  - If $x < x_1$ and $y < y_1$, then $F_{X,Y}(x,y) \leq F_{X,Y}(x_1,y_1)$.
  - $F_{X,Y}(\infty,\infty) = 1$.
  - If $x_1 < x_2, y_1 < y_2$, then

$$\begin{aligned} &P[x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2] \\ &= F_{X,Y}(x_2,y_2) - F_{X,Y}(x_2,y_1) - F_{X,Y}(x_1,y_2) + F_{X,Y}(x_1,y_1) \end{aligned}$$

4.1.2 Joint Probability Mass Function.

- **Definition:** The joint probability mass function of discrete random variables X and Y

$$\boxed{P_{X,Y}(x,y) = P[X = x, Y = y]}$$

- Corresponding to S_X, the range of a single discrete random variable, we use the notation $S_{X,Y}$ to denote the set of possible values of the pair (X,Y). That is,

$$\boxed{S_{X,Y} = \{(x,y) | P_{X,Y}(x,y) > 0\}}$$

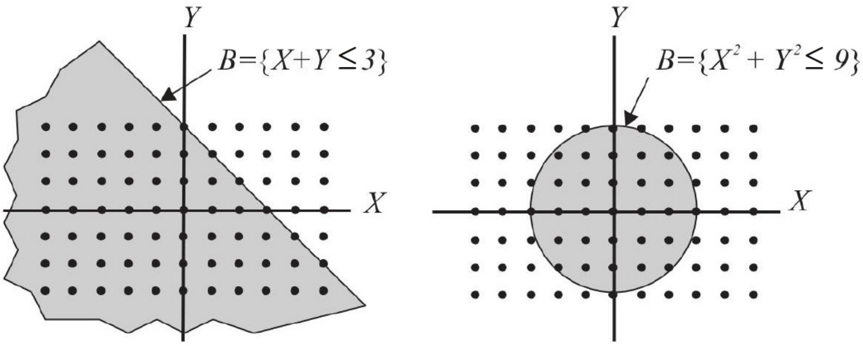
$X \backslash Y$	y_1	$\dots$	y_j	$\dots$	y_n	$\sum_j$
x_1	p_{11}	$\dots$	p_{1j}	$\dots$	p_{1n}	$P[X = x_1]$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	p_{i1}	$\dots$	p_{ij}	$\dots$	p_{in}	$P[X = x_i]$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_m	p_{m1}	$\dots$	p_{mj}	$\dots$	p_{mn}	$P[X = x_m]$
$\sum_i$	$P[Y = y_1]$	$\dots$	$P[Y = y_j]$	$\dots$	$P[Y = y_n]$	$\sum_i \sum_j = 1$

- **Properties:**
  - $0 \leq p_i \leq 1$ for all $i = 1, 2, ..., m$, $j = 1, 2, ..., n$.
  - $\sum\limits_{i=1}^{m} \sum\limits_{j=1}^{n} p_{ij} = 1$.

$\Rightarrow$ The joint cumulative distribution function of discrete random variables X and Y is

$$\boxed{F_{X,Y}(x, y) = \sum_{x_i < x} \sum_{y_j < y} p_{ij}}$$

- We represent an event B as a region in the xy-plane. Plots below show two examples of events.



- For discrete random variables X and Y and any set B in the xy-plane, the probability of the event $\{(X, Y) \in B\}$ is

$$\boxed{P[B] = \sum_{(x,y) \in B} P_{X,Y}(x, y)}$$

- For discrete random variables X and Y with joint PMF $P_{X,Y}(x, y)$,

$$\boxed{P_X(x) = \sum_{y \in S_Y} P_{X,Y}(x, y) \quad ; \quad P_Y(y) = \sum_{x \in S_X} P_{X,Y}(x, y)}$$

$\Rightarrow$ From the definition of joint probability distribution table, we have

- Marginal CD of X:

X	x_1	x_2	...	x_m
p	$P[X = x_1]$	$P[X = x_2]$	...	$P[X = x_m]$

- Marginal CD of Y:

Y	y_1	y_2	...	y_n
p	$P[Y = y_1]$	$P[Y = y_2]$	...	$P[Y = y_n]$

4.1.3 Joint Probability Density Function.

- **Definition:** The joint PDF of the continuous random variables X and Y is a function fX,Y (x, y) with the property

$$\boxed{F_{X,Y}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(u, v) \mathrm{d}v \mathrm{d}u \quad \Leftrightarrow \quad f_{X,Y}(x, y) = \frac{\partial^2 F_{X,Y}(x, y)}{\partial x \partial y}}$$

- A joint PDF $f_{X,Y}(x,y)$ has the following properties corresponding to first and second axioms of probability:
  - $f_{X,Y}(x,y) \geq 0$ for all x, y.
  - $\int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty} f_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y = 1$.
- The probability that the continuous random variables (X, Y) are in A is

$$P[A] = \iint\limits_{A} f_{X,Y}(x,y)\mathrm{d}x\mathrm{d}yx$$

- If X and Y are random variables with joint PDF $f_{X,Y}(x,y)$,

$$F_X(x) = \int\limits_{-\infty}^{+\infty} f_{X,Y}(x,y)\mathrm{d}y \quad ; \quad F_Y(y) = \int\limits_{-\infty}^{+\infty} f_{X,Y}(x,y)\mathrm{d}x$$

4.1.4 Expected values.

- For random variables X and Y, the expected value of $W = g(X,Y)$ is

$$\begin{aligned} \text{Discrete}: E[W] &= \sum_{x\in S_X}\sum_{y\in S_Y} g(x,y)P_{X,Y}(x,y) \\ E[W] &= \sum_i\sum_j g(x_i,y_j)P_{X,Y}(X=x_i, Y=y_j) \\ \text{Continuous}: E[W] &= \int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty} g(x,y)f_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y \end{aligned}$$

- Hence, we have
  - $E[g_1(X,Y) + ... + g_n(X,Y)] = E[g_1(X,Y)] + ... + E[g_n(X,Y)]$.
  - For any two random variables X and Y, $E[X+Y] = E[X] + E[Y]$.
  - The variance of the sum of two random variables is

$$Var[X+Y] = Var[X] + Var[Y] + 2E[(X-\mu_X)(Y-\mu_Y)]$$

- The covariance of two random variables X and Y is

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

  From the properties of the expected value, $\boxed{Cov(X,Y) = E[XY] - E[X].E[Y]}$, where

$$E[XY] = \begin{cases} \sum\limits_i\sum\limits_j x_iy_jp_{ij}, & \text{discrete} \\ \int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty} xyf_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y, & \text{continuous} \end{cases}$$

- **Properties of covariance:**
  - $Cov(X,Y) = E[XY] - E[X].E[Y] = r_{X,Y} - \mu_X\mu_Y$.

- $Var[X+Y] = Var[X] + Var[Y] + 2Cov[X,Y]$.
- $Var[X] = Cov[X,X], \quad Var[Y] = Cov[Y,Y]$.

- The covariance matrix of two random variables X and Y is

$$\Gamma = \begin{bmatrix} Cov(X,X) & Cov(X,Y) \\ Cov(Y,X) & Cov(Y,Y) \end{bmatrix} = \begin{bmatrix} Var[X] & Cov(X,Y) \\ Cov(Y,X) & Var[Y] \end{bmatrix}$$

- The correlation of X and Y is $r_{X,Y} = E[XY]$, where

$$E[XY] = \begin{cases} \sum\limits_i \sum\limits_j x_i y_j p_{ij}, & \text{discrete} \\ \int\limits_{-\infty}^{+\infty} \int\limits_{-\infty}^{+\infty} xy f_{X,Y}(x,y) \mathrm{d}x\mathrm{d}y, & \text{continuous} \end{cases}$$

From that,

- Random variables X and Y are orthogonal if $r_{X,Y} = 0$.
- Random variables X and Y are uncorrelated if $Cov[X,Y] = 0$.

- The correlation coefficient of two random variables X and Y is

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sqrt{Var[X] \times Var[Y]}} = \frac{Cov(X,Y)}{\sigma_X \times \sigma_Y}$$

Hence, $-1 \le \rho_{X,Y} \le 1$.

If X and Y are random variables such that $Y = aX + b \Rightarrow \rho_{X,Y} = \begin{cases} -1, & a < 0, \\ 0, & a = 0, \\ 1, & a > 0. \end{cases}$

4.1.5 Conditioning by an Event.

- For discrete random variables X and Y and an event, B with $P[B] > 0$, the conditional joint PMF of X and Y given B is

$$P_{X,Y|B}(x,y) = P[X = x, Y = y|B]$$

- For any event B, a region of the xy-plane with $P[B] > 0$,

$$P_{X,Y|B}(x,y) \begin{cases} \dfrac{P_{X,Y}(x,y)}{P[B]}, & (x,y) \in B, \\ 0, & \text{otherwise.} \end{cases}$$

- Given an event B with $P[B] > 0$, the conditional joint probability density function of X and Y is

$$f_{X,Y|B}(x,y) \begin{cases} \dfrac{f_{X,Y}(x,y)}{P[B]}, & (x,y) \in B, \\ 0, & \text{otherwise.} \end{cases}$$

- For random variables X and Y and an event B of nonzero probability, the conditional expected value of $W = g(X,Y)$ given B is

$$\text{Discrete} : E[W|B] = \sum_{x \in S_X} \sum_{y \in S_Y} g(x,y) P_{X,Y|B}(x,y)$$

$$\text{Continuous} : E[W|B] = \int\limits_{-\infty}^{+\infty} \int\limits_{-\infty}^{+\infty} g(x,y) f_{X,Y|B}(x,y) \mathrm{d}x\mathrm{d}y$$

- The conditional variance of the random variable $W = g(X, Y)$ is

$$Var[W|B] = E\left[(W - E[W|B])^2 \mid B\right] = E[W^2|B] - (E[W|B])^2$$

4.1.6 Conditioning by a Random Variable.

- For any event $Y = y$ such that $P_Y(y) > 0$, the conditional PMF of X given $Y = y$ is $P_{X|Y}(x|y) = P[X = x|Y = y]$.
- For random variables X and Y with joint PMF $P_{X,Y}(x, y)$, and x and y such that $P_X(x) > 0$ and $P_Y(y) > 0$,

$$P_{X,Y}(x, y) = P_{X|Y}(x|y)P_Y(y) = P_{Y|X}(y|x)P_X(x)$$

- X and Y are discrete random variables. For any $y \in S_Y$, the conditional expected value of $g(X, Y)$ given $Y = y$ is

$$E[g(X, Y)|Y = y] = \sum_{x \in S_X} g(x, y)P_{X|Y}(x|y)$$

- For y such that $f_Y(y) > 0$, the conditional PDF of X given $\{Y = y\}$ is

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}. \quad \text{This implies : } f_{Y|X}(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)}$$

$$\Rightarrow f_{X,Y}(x, y) = f_{Y|X}(y|x)f_X(x) = f_{X|Y}(x|y)f_Y(y)$$

- For continuous random variables X and Y and any y such that $f_Y(y) > 0$, the conditional expected value of $g(X, Y)$ given $Y = y$ is

$$E[g(X, Y)|Y = y] = \int_{-\infty}^{+\infty} g(x, y)f_{X|Y}(x|y)\mathrm{d}x$$

  The conditional expected value $E[X|Y]$ is a function of random variable Y such that if $Y = y$ then $E[X|Y] = E[X|Y = y]$.

4.1.7 Independent Random Variables.

- **Definitions:** Random variables X and Y are independent if and only if

$$\text{Discrete : } P_{X,Y}(x, y) = P_X(x)P_Y(y)$$
$$\text{Continuous : } f_{X,Y}(x, y) = f_X(x)f_Y(y)$$

- We can implies that:
  - If X and Y are independent discrete random variables:

$$P_{X|Y}(x|y) = P_X(x) \quad ; \quad P_{Y|X}(y|x) = P_Y(y)$$

  - If X and Y are independent continuous random variables:

$$f_{X|Y}(x|y) = f_X(x) \quad ; \quad f_{Y|X}(y|x) = f_Y(y)$$

- **Properties:** For independent random variables X and Y,
  - $E[g(X)h(Y)] = E[g(X)]E[h(Y)]$.
  - $r_{X,Y} = E[XY] = E[X]E[Y]$.
  - $Cov[X,Y] = \rho_{X,Y} = 0$.
  - $Var[X+Y] = Var[X] + Var[Y]$.
  - $E[X|Y=y] = E[X]$ for all $y \in S_Y$.
  - $E[Y|X=x] = E[Y]$ for all $x \in S_X$.

4.2 Answer of Exercises.

Problem 4.1.

Random variables X and Y have the joint CDF

$$F_{X,Y}(x,y) = \begin{cases} (1-e^{-x})(1-e^{-y}), & x \geq 0, y \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

(a) What is $P[X < 2, Y < 3]$?
(b) What is the marginal CDF, $F_X(x)$?
(c) What is the marginal CDF, $F_Y(y)$?

Solution

(a) $P[X < 2, Y < 3] = F_{X,Y}(2,3) = \left(1-e^{-2}\right)\left(1-e^{-3}\right) \approx 0.8216$.

(b) For $x \geq 0$, we have: $F_X(x) = F_{X,Y}(x,+\infty) = (1-e^{-x})\left(1 - \lim\limits_{y\to+\infty} e^{-y}\right) = 1 - e^{-x}$.

So, the marginal CDF $F_X(x) = \begin{cases} 1-e^{-x}, & x \geq 0, \\ 0, & \text{otherwise.} \end{cases}$

(c) For $y \geq 0$, we have: $F_Y(y) = F_{X,Y}(+\infty, y) = \left(1 - \lim\limits_{x\to+\infty} e^{-x}\right)(1-e^{-y}) = 1 - e^{-y}$.

So, the marginal CDF $F_Y(y) = \begin{cases} 1-e^{-y}, & y \geq 0, \\ 0, & \text{otherwise.} \end{cases}$

Problem 4.2.

Express the following extreme values of $F_{X,Y}(x,y)$ in terms of the marginal cumulative distribution functions $F_X(x)$ and $F_Y(y)$.

(a) $F_{X,Y}(x,-\infty)$.
(b) $F_{X,Y}(x,\infty)$.
(c) $F_{X,Y}(-\infty,\infty)$.
(d) $F_{X,Y}(-\infty,y)$.
(e) $F_{X,Y}(\infty,y)$.

Solution

(a) $F_{X,Y}(x,-\infty) = P[X \leq x, Y \leq -\infty] \leq P[Y \leq -\infty] = 0$.
(b) $F_{X,Y}(x,\infty) = F_X(x)$.
(c) $F_{X,Y}(-\infty,\infty) = P[X \leq -\infty, Y \leq \infty] \leq P[X \leq -\infty] = 0$.

(d) $F_{X,Y}(-\infty, y) = P[X \le -\infty, Y \le y] \le P[X \le -\infty] = 0.$
(e) $F_{X,Y}(\infty, y) = F_Y(y).$

Problem 4.3.
Random variables X and Y have CDF $F_X(x)$ and $F_Y(y)$. Is $F_{X,Y}(x, y) = F_X(x)F_Y(y)$ a valid CDF? Explain your answer.

Solution

Problem 4.4.
Random variables X and Y have the joint PMF

$$P_{X,Y}(x,y) = \begin{cases} cxy, & x = 1,2,4;\ y = 1,3, \\ 0, & \text{otherwise.} \end{cases}$$

(a) What is the value of the constant c?
(b) What is $P[Y < X]$?
(c) What is $P[Y > X]$?
(d) What is $P[Y = X]$?
(e) What is $P[Y = 3]$?

Solution

(a) $\sum\limits_{x \in S_X} \sum\limits_{y \in S_Y} cxy = 1 \Leftrightarrow 28c = 1 \Leftrightarrow c = \dfrac{1}{28}.$
(b) $P[Y < X] = P[X = 2, Y = 1] + P[X = 4, Y = 1] + P[X = 4, Y = 3] = \dfrac{9}{14}.$
(c) $P[Y > X] = \dfrac{9}{28}.$
(d) $P[Y = X] = \dfrac{1}{28}.$
(e) $P[Y = 3] = P[X = 1, Y = 3] + P[X = 2, Y = 3] + P[X = 4, Y = 3] = \dfrac{3}{4}.$

Problem 4.5.
Random variables X and Y have the joint PMF

$$P_{X,Y}(x,y) = \begin{cases} c|x+y|, & x = -2,0,2;\ y = -1,0,1, \\ 0, & \text{otherwise.} \end{cases}$$

(a) What is the value of the constant c?
(b) What is $P[Y < X]$?
(c) What is $P[Y > X]$?
(d) What is $P[Y = X]$?
(e) What is $P[X < 1]$?

Solution

(a) $\sum\limits_{x \in S_X} \sum\limits_{y \in S_Y} c|x+y| = 1 \Leftrightarrow 14c = 1 \Leftrightarrow c = \dfrac{1}{14}.$

(b) $P[Y < X] = \dfrac{9}{14}$.
(c) $P[Y > X] = \dfrac{5}{14}$.
(d) $P[Y = X] = 0$
(e) $P[X < 1] = \dfrac{4}{7}$.

Problem 4.6.
Given the random variables X and Y in **Problem 4.4**, find

(a) The marginal PMFs $P_X(x)$ and $P_Y(y)$.
(b) The expected values $E[X]$ and $E[Y]$.
(c) The standard deviations σ_X and σ_Y.

Solution
In **Problem 4.4**, random variables X and Y have the joint PMF

$$P_{X,Y}(x,y) = \begin{cases} \frac{1}{28}xy, & x = 1,2,4;\ y = 1,3, \\ 0, & \text{otherwise.} \end{cases}$$

(a) The marginal PMFs $P_X(x)$ and $P_Y(y)$:

$$P_X(x) = \begin{cases} \frac{4}{28}, & x = 1, \\ \frac{8}{28}, & x = 2, \\ \frac{16}{28}, & x = 4, \\ 0, & \text{otherwise.} \end{cases} \quad ; \quad P_Y(y) = \begin{cases} \frac{7}{28}, & y = 1, \\ \frac{21}{28}, & y = 3, \\ 0, & \text{otherwise.} \end{cases}$$

(b)

$$E[X] = \sum x_i P[X = x_i] = 3 \quad ; \quad E[Y] = \frac{5}{2}.$$

(c)

$$\sigma_X = \sqrt{E[X^2] - E[X]^2} = \frac{\sqrt{70}}{7} \quad ; \quad \sigma_Y = \frac{\sqrt{3}}{2}.$$

Problem 4.7.
Given the random variables X and Y in **Problem 4.5**, find

(a) The marginal PMFs $P_X(x)$ and $P_Y(y)$.
(b) The expected values $E[X]$ and $E[Y]$.
(c) The standard deviations σ_X and σ_Y.

Solution
In **Problem 4.5**, random variables X and Y have the joint PMF

$$P_{X,Y}(x,y) = \begin{cases} \frac{1}{14}|x+y|, & x = -2,0,2;\ y = -1,0,1, \\ 0, & \text{otherwise.} \end{cases}$$

(a) The marginal PMFs $P_X(x)$ and $P_Y(y)$:

$$P_X(x) = \begin{cases} \frac{6}{14}, & x = -2, \\ \frac{2}{14}, & x = 0, \\ \frac{6}{14}, & x = 2, \\ 0, & \text{otherwise.} \end{cases} \quad ; \quad P_Y(y) = \begin{cases} \frac{5}{14}, & y = -1, \\ \frac{4}{14}, & y = 0, \\ \frac{5}{14}, & y = 1, \\ 0, & \text{otherwise.} \end{cases}$$

(b)

$$E[X] = \sum x_i P[X = x_i] = 0 \quad ; \quad E[Y] = 0.$$

(c)

$$\sigma_X = \sqrt{E[X^2] - E[X]^2} = \frac{2\sqrt{42}}{7} \quad ; \quad \sigma_Y = \frac{\sqrt{35}}{7}.$$

Problem 4.8.

Random variables X and Y have the joint PDT

$X \setminus Y$	1	2	3
1	0.12	0.15	0.03
2	0.28	0.35	0.07

(a) Are X and Y independent?
(b) Find the marginal PDTs of X and Y.
(c) Find the PDT of Z, where $Z = XY$.
(d) Find $E(Z)$. Proof $E(Z) = E(X).E(Y)$.

Solution

(a) Since $P[X = 1] \times P[Y = 1] = 0.57 \neq P[X = 1, Y = 1] = 0.12 \Rightarrow X$ and Y are not independent.

(b) The marginal PDTs of X and Y:

x	1	2
$P[X = x]$	0.3	0.7

y	1	2	3
$P[Y = y]$	0.4	0.5	0.1

(c) The PDT of Z, where $Z = XY$.

z	1	2	3	4	6
$P[Z = z = xy]$	0.12	0.43	0.03	0.35	0.07

(d) $E(Z) =$.
We have $Z = XY$, so $E(Z) = \sum z \times P[z] = \sum xy \times P[xy]$

Problem 4.9.

Random variables X and Y have the joint PDT

X | Y	-1	0	1
-1	$\frac{4}{15}$	$\frac{1}{15}$	$\frac{4}{15}$
0	$\frac{1}{15}$	$\frac{2}{15}$	$\frac{1}{15}$
1	0	$\frac{2}{15}$	0

(a) Find $E(X)$, $E(Y)$ and $Cov(X, Y)$.
(b) Are X and Y independent?
(c) Find the marginal PDTs of X and Y.

Solution

(a)
- $E[X] = (-1)\left(\frac{4}{15}+\frac{1}{15}+\frac{4}{15}\right)+0\times\left(\frac{1}{15}+\frac{2}{15}+\frac{1}{15}\right)+1\times\left(0+\frac{2}{15}+0\right) = -\frac{7}{15}$.
- $E[Y] = (-1)\left(\frac{4}{15}+\frac{1}{15}+0\right)+0\times\left(\frac{1}{15}+\frac{2}{15}+\frac{2}{15}\right)+1\times\left(\frac{4}{15}+\frac{1}{15}+0\right) = 0.$
- $E[XY] = (-1)\times(-1)\times\frac{4}{15}+(-1)\times 1\times\frac{4}{15}+1\times(-1)\times 0+1\times 1\times 0 = 0.$
  Hence, $Cov(X, Y) = E[XY] - E[X]E[Y] = 0.$

(b) We have $P(X = -1, Y = -1) \neq P(X = -1) \times P(Y = -1)$, so X, Y are dependent.

(c) PDT of X and$Y$:

X	-1	0	1
$P(X)$	$\frac{9}{15}$	$\frac{4}{15}$	$\frac{5}{15}$

Y	-1	0	1
$P(Y)$	$\frac{5}{15}$	$\frac{5}{15}$	$\frac{5}{15}$

Problem 4.10.

Random variables X and Y have the joint PDT

X | Y	1	2	3
1	0.17	0.13	0.25
2	0.10	0.30	0.05

(a) Find the marginal PDTs of X and Y.
(b) Find the covariance matrix of X and Y.
(c) Find the correlation coefficient of two random variables X and Y.
(d) Are X and Y independent?

Solution

(a) PDT of X and$Y$:

X	1	2
$P(X)$	0.55	0.45

Y	1	2	3
$P(Y)$	0.27	0.43	0.3

(b)
- We have: $E[X] = 1.45, Var[X] = 0.2475, E[Y] = 2.03, Var[Y] = 0.5691$. Moreover, $E[XY] = 2.88 \Rightarrow Cov(X,Y) = -0.0635$.
- Covariance matrix:

$$\Gamma = \begin{bmatrix} Cov(X,X) & Cov(X,Y) \\ Cov(Y,X) & Cov(Y,Y) \end{bmatrix} = \begin{bmatrix} Var[X] & Cov(X,Y) \\ Cov(Y,X) & Var[Y] \end{bmatrix} = \begin{bmatrix} 0.2475 & -0.0635 \\ -0.0635 & 0.5691 \end{bmatrix}$$

(c) Correlation coefficient: $\rho_{X,Y} = \dfrac{Cov(X,Y)}{\sqrt{Var[X] \times Var[Y]}} = -0.1692$.

(d) We have $P(X=1, Y=1) \neq P(X=1)P(Y=1)$, so X, Y are not independent.

Problem 4.11.

Random variables X and Y have the joint PDF

$$f_{X,Y}(x,y) = \begin{cases} c, & x+y \leq 1,\ x \geq 0,\ y \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

(a) What is the value of the constant c?
(b) What is $P[X \leq Y]$?
(c) What is $P\left[X + Y \leq \frac{1}{2}\right]$?

Solution

(a) We have $\begin{cases} c \geq 0, & 0 < y < x < 1, \\ \int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty} c\mathrm{d}x\mathrm{d}y = 1 \end{cases} \Rightarrow \begin{cases} c \geq 0, & 0 < y < x < 1, \\ \int\limits_0^1 \mathrm{d}x \int\limits_0^{1-x} c\mathrm{d}y = 1 \end{cases} \Rightarrow c = 2.$

(b) $P[X \leq Y] = \int\limits_0^{\frac{1}{2}} \mathrm{d}x \int\limits_{1-x}^{1} 2\mathrm{d}y = \dfrac{1}{2}$.

(c) $P\left[X + Y \leq \frac{1}{2}\right] = \int\limits_0^{\frac{1}{2}} \mathrm{d}x \int\limits_0^{\frac{1}{2}-x} 2\mathrm{d}y = \dfrac{1}{4}$.

Problem 4.12.

Random variables X and Y have joint PDF

$$f_{X,Y}(x,y) = \begin{cases} cxy^2, & 0 \leq x \leq 1,\ 0 \leq y \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$

(a) What is the value of the constant c?
(b) Find $P[X > Y]$ and $P[Y < X^2]$.
(c) Find $P\left[\min(X,Y) \leq \frac{1}{2}\right]$.
(d) Find $P\left[\max(X,Y) \leq \frac{3}{4}\right]$.

Solution

(a) $\begin{cases} cxy^2 \geq 0, & 0 < y < x < 1, \\ \int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty} cxy^2\mathrm{d}x\mathrm{d}y = 1 \end{cases} \Rightarrow \begin{cases} cxy^2 \geq 0, & 0 < y < x < 1, \\ \int\limits_0^1 \mathrm{d}x \int\limits_0^1 cxy^2\mathrm{d}y = 1 \end{cases} \Rightarrow c = 6.$

(b)
- $P[X \leq Y] = \int\limits_0^1 \mathrm{d}y \int\limits_0^y 6xy^2 \mathrm{d}x = \dfrac{3}{5} \Rightarrow P[X > Y] = 1 - P[X \leq Y] = \dfrac{2}{5}$.
- $P[Y \leq X^2] = \int\limits_0^1 \mathrm{d}x \int\limits_0^{x^2} 6xy^2 \mathrm{d}y = \dfrac{1}{4}$.

(c) $P\left[\min(X,Y) \leq \frac{1}{2}\right] = 1 - P\left[X > \frac{1}{2}, Y > \frac{1}{2}\right] = 1 - \int\limits_{\frac{1}{2}}^1 \mathrm{d}y \int\limits_{\frac{1}{2}}^1 6xy^2 \mathrm{d}x = 1 - \dfrac{21}{32} = \dfrac{11}{32}$.

(d) $P\left[\max(X,Y) \leq \frac{3}{4}\right] = P\left[X < \frac{3}{4}, Y < \frac{3}{4}\right] = \int\limits_0^{\frac{3}{4}} \mathrm{d}x \int\limits_0^{\frac{3}{4}} 6xy^2 \mathrm{d}y = \dfrac{243}{1024}$.

Problem 4.13.

Random variables X and Y have the joint PDF $f_{X,Y}(x,y) = \begin{cases} \frac{1}{2}, & -1 \leq x \leq y \leq 1, \\ 0, & \text{otherwise.} \end{cases}$

(a) Sketch the region of nonzero probability.
(b) What is $P[X > 0]$?
(c) What is $f_X(x)$?
(d) What is $E[X]$?

Solution

Problem 4.14.

X and Y are random variables with the joint PDF

$$f_{X,Y}(x,y) = \begin{cases} 2, & x + y \leq 1,\ x \geq 0,\ y \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

(a) What is the marginal PDF $f_X(x)$?
(b) What is the marginal PDF $f_Y(y)$?

Solution

Problem 4.15.

Random variables X and Y have the joint PDF $f_{X,Y}(x,y) = \begin{cases} cy, & 0 \leq y \leq x \leq 1, \\ 0, & \text{otherwise.} \end{cases}$

(a) Draw the region of nonzero probability.
(b) What is the value of the constant c?
(c) What is $F_X(x)$?
(d) What is $F_Y(y)$?
(e) What is $P\left[Y \leq \frac{X}{2}\right]$?

Solution

Problem 4.16.

Given random variables X and Y in **Problem 4.5** and the function $W = X + 2Y$, find

(a) The probability mass function $P_W(w)$.
(b) The expected value $E[W]$.
(c) $P[W > 0]$.

Solution

Problem 4.17.

Random variables X and Y have joint PDF

$$f_{X,Y}(x,y) = \begin{cases} x+y, & 0 \le x \le 1,\ 0 \le y \le 1, \\ 0, & \text{otherwise.} \end{cases}.$$

Let $W = \max(X, Y)$.

(a) What is S_W, the range of W?
(b) Find $F_W(w)$ and $f_W(w)$.

Solution

Problem 4.18.

Random variables X and Y have joint PDF $f_{X,Y}(x,y) = \begin{cases} 2, & 0 \le y \le x \le 1, \\ 0, & \text{otherwise.} \end{cases}.$

Let $W = \dfrac{Y}{X}$.

(a) What is S_W, the range of W?
(b) Find $F_W(w)$, $f_W(w)$ and $E[W]$.

Solution

Problem 4.19.

For the random variables X and Y in **Problem 4.4**, find

(a) The expected value of $W = Y/X$,
(b) The correlation, $E[XY]$,
(c) The covariance, $Cov[X,Y]$,
(d) The correlation coefficient, $\rho_{X,Y}$,
(e) The variance of $X+Y$, $Var[X+Y]$.

Solution

Problem 4.20.

Random variables X and Y have joint PMF

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{21}, & x = 0,1,2,3,4,5;\ y = 0,1,...,x, \\ 0, & \text{otherwise.} \end{cases}$$

Find the marginal PMFs $P_X(x)$ and $P_Y(y)$. Also find the covariance $Cov[X,Y]$.

Solution

Problem 4.21.

Random variables X and Y have joint PDF

$$f_{X,Y}(x,y) = \begin{cases} \frac{x+y}{3}, & 0 \le x \le 1,\ 0 \le y \le 2, \\ 0, & \text{otherwise.} \end{cases}$$

(a) What are $E[X]$ and $Var[X]$?
(b) What are $E[Y]$ and $Var[Y]$?
(c) What is $Cov[X,Y]$?
(d) What is $E[X+Y]$?
(e) What is $Var[X+Y]$?

Solution

Problem 4.22.

Random variables X and Y have joint PDF

$$f_{X,Y}(x,y) = \begin{cases} \frac{x+y}{3}, & 0 \le x \le 1,\ 0 \le y \le 2, \\ 0, & \text{otherwise.} \end{cases} .$$

Let $A = Y \le 1$.

(a) What is $P[A]$?
(b) Find $f_{X,Y|A}(x,y)$, $f_{X|A}(x)$, and $f_{Y|A}(y)$.

Solution

Problem 4.23.

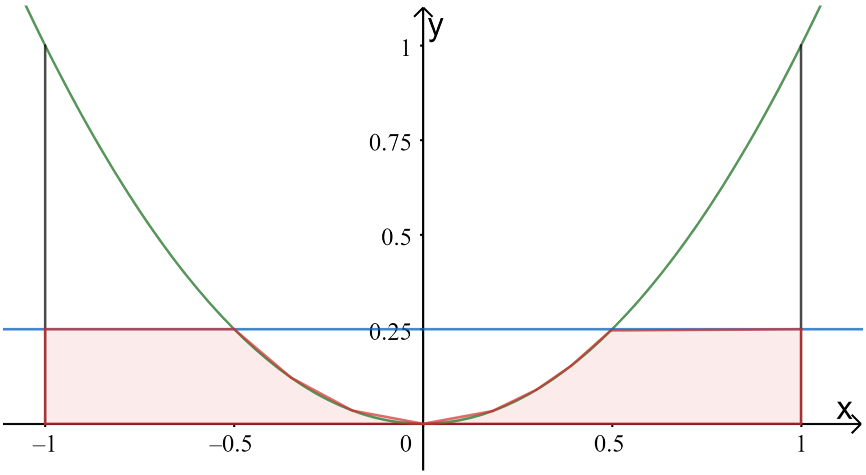
Random variables X and Y have joint PDF

$$f_{X,Y}(x,y) = \begin{cases} \frac{5x^2}{2}, & -1 \le x \le 1,\ 0 \le y \le x^2, \\ 0, & \text{otherwise.} \end{cases}.$$

Let $A = \left\{Y \le \dfrac{1}{4}\right\}$.

(a) What is the conditional PDF $f_{X,Y|A}(x,y)$?
(b) What is $f_{Y|A}(y)$?
(c) What is $E[Y|A]$?
(d) What is $f_{X|A}(x)$?
(e) What is $E[X|A]$?

Solution



Since $A = \left\{Y \le \dfrac{1}{4}\right\} = \begin{cases} 0 \le y \le \frac{1}{4} \\ -1 \le x \le -\sqrt{y} \end{cases} \cup \begin{cases} 0 \le y \le \frac{1}{4} \\ \sqrt{y} \le x \le 1 \end{cases}$. Therefore,

$$P(A) = \int_0^{\frac{1}{4}} \left(\int_{-1}^{-\sqrt{y}} \frac{5x^2}{2} dx \right) dy + \int_0^{\frac{1}{4}} \left(\int_{\sqrt{y}}^{1} \frac{5x^2}{2} dx \right) dy = \frac{48}{19}.$$

(a) $f_{X,Y|A}(x,y) = \begin{cases} \dfrac{120x^2}{19} & \text{if } \begin{cases} 0 \le y \le \frac{1}{4} \\ -1 \le x \le -\sqrt{y} \end{cases} \text{ or } \begin{cases} 0 \le y \le \frac{1}{4} \\ \sqrt{y} \le x \le 1 \end{cases} \\ 0, & \text{otherwise.} \end{cases}$

(b) $f_Y(y) = \begin{cases} \dfrac{5}{3}(1 - y^{1.5}), & y \in [0;1] \\ 0, & \text{otherwise.} \end{cases}$. Hence, $f_{Y|A}(y) = \begin{cases} \dfrac{80}{19}(1 - y^{1.5}), & y \in [0;1] \\ 0, & \text{otherwise.} \end{cases}$

(c) $E[Y|A] = \int_{-\infty}^{+\infty} y f_{Y|A}(y) dy = \dfrac{xx}{xx}$.

(d)

Since $D = \left\{-1 \le x \le 1; 0 \le y \le x^2\right\}$
$= \{0 \le y \le 1; -1 \le x \le -\sqrt{y}\} \cup \{0 \le y \le 1; \sqrt{y} \le x \le 1\}$

$$\Rightarrow A \cap D = \left\{ y \le \frac{1}{4} \right\} \cap D$$
$$= \left\{ 0 \le y \le \frac{1}{4}; -1 \le x \le -\sqrt{y} \right\} \cup \left\{ 0 \le y \le \frac{1}{4}; \sqrt{y} \le x \le 1 \right\}$$

So, $f_X(x) = \int\limits_{-\infty}^{+\infty} f(x,y)\mathrm{d}y = \dfrac{5x^4}{2}$ if $x \in [-1;1]$. Hence, $f_X(x) = \dfrac{5x^4}{2} \times \dfrac{48}{19}$ if $(x,y) \in D$.

(e)
$$E[X|A] = \iint\limits_{\mathbb{R}^2} x.f_{X,Y|A}(x,y)\mathrm{d}x\mathrm{d}y = \iint\limits_{A \cap D} x \times \frac{120x^2}{9}\mathrm{d}x\mathrm{d}y = 0$$

Problem 4.24.

X and Y have joint PDF $f_{X,Y}(x,y) = \begin{cases} \frac{4x+2y}{2}, & -1 \le x \le 1,\ 0 \le y \le x^2, \\ 0, & \text{otherwise.} \end{cases}$.

(a) For which values of y is $f_{X|Y}(x|y)$ defined? What is $f_{X|Y}(x|y)$?
(b) For which values of x is $f_{Y|X}(y|x)$ defined? What is $f_{Y|X}(y|x)$?

Solution

Problem 4.25.

The joint PDF of two random variables X and Y is

$$f_{X,Y}(x,y) = \begin{cases} kx, & 0 < y < x < 1, \\ 0, & \text{otherwise.} \end{cases}.$$

(a) Find the constant k.
(b) Find the PDFs of X and Y.
(c) Are X and Y independent?

Solution

(a) We have: $\begin{cases} kx \ge 0, & 0 < y < x < 1, \\ \int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty} kx\mathrm{d}x\mathrm{d}y = 1 \end{cases} \Rightarrow \begin{cases} kx \ge 0, & 0 < y < x < 1, \\ \int\limits_{0}^{1} \mathrm{d}x \int\limits_{0}^{x} kx\mathrm{d}y = 1 \end{cases} \Rightarrow k = 3.$

(b)
- Marginal PDF of X:

$$f_X(x) = \int\limits_{-\infty}^{+\infty} f_{X,Y}(x,y)\mathrm{d}y = \begin{cases} \int\limits_{0}^{x} 3x\mathrm{d}y, & 0 < x < 1, \\ 0, & \text{otherwise.} \end{cases} = \begin{cases} 3x^2, & 0 < x < 1, \\ 0, & \text{otherwise.} \end{cases}$$

- Marginal PDF of Y:

$$f_Y(y) = \int\limits_{-\infty}^{+\infty} f_{X,Y}(x,y)\mathrm{d}x = \begin{cases} \int\limits_{y}^{1} 3x\mathrm{d}x, & 0 < y < 1, \\ 0, & \text{otherwise.} \end{cases} = \begin{cases} \frac{3}{2} - \frac{3}{2}y^2, & 0 < y < 1, \\ 0, & \text{otherwise.} \end{cases}$$

(c) Since $f_{X,Y}(x,y) \ne f_X(x)f_Y(x)$ with $0 < y < x < 1$, hence, X, Y are dependent.

Chapter 5

Random Sample.

5.1 Notions.

5.1.1 Populations and Samples.

- A population consists of the totality of the observations with which we are concerned. The number of observations in the population is defined to be the size of the population N.
- A sample is a subset of a population. The number of observations in the sample is defined to be the size of the sample n.

5.1.2 Random sample

Let $X_1, X_2, ..., X_n$ be n independent random variables, each having the same probability distribution $F_X(x)$. Define $(X_1, X_2, ..., X_n)$ to be a random sample of size n from the population $F_X(x)$ and write its joint probability distribution as

$$F_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) = F_{X_1}(x_1) \times F_{X_2}(x_2) \times ... \times F_{X_n}(x_n)$$

5.2 Some important statistics.

Any function of the random variables constituting a random sample is called a statistic.

5.2.1 Sample Mean. Median. Mode.

- Sample mean: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

  Note that the statistic $\overline{X}$ assumes the value $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ when X_1 assumes the value x_1, X_2 assumes the value x_2, and so forth.
- Sample median: $\hat{x} = \begin{cases} x_{\frac{n+1}{2}}, & \text{if } n \text{ is odd,} \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n+1}{2}}\right), & \text{if } n \text{ is even.} \end{cases}$
- The sample mode is the value of the sample that occurs most often.

5.2.2 Sample Variance. Standard Deviation. Range.

- Sample variance: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$. The computed value of S^2 for a given sample is denoted by s^2.

- Sample standard deviation: $S = \sqrt{S^2}$.
- Let $X_{\max}$ denote the largest of the X_i values and $X_{\min}$ the smallest. Sample range: $R = X_{\max} - X_{\min}$.

Theorem

If S^2 is the variance of a random sample of size n, we may write

$$S^2 = \frac{1}{n(n-1)}\left[n \times \sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2\right]$$

Note:

Sample	Population
n: number of measurements in the sample $\overline{x}$: sample mean s^2: sample variance s: sample standard deviation	N: number of measurements in the population μ: population mean σ^2: population variance σ: population standard deviation

5.3 Sampling Distributions.

Definition: The probability distribution of a statistic is called a sampling distribution.

5.3.1 Sampling Distributions of Means and Central Limit Theorem.

- **Introduction:** Suppose that a random sample of n observations is taken from a normal population with mean μ and variance σ^2. Each observation X_i, $i = 1, 2, ..., n$, of the random sample will then have the same normal distribution as the population being sampled.
- **Sampling Distributions of Means:**

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + ... + X_n)$$

  has a normal distribution with mean $\mu_{\overline{X}} = \mu$ and variance $\sigma^2_{\overline{X}} = \frac{\sigma^2}{n}$.

Central Limit Theorem

If $\overline{X}$ is the mean of a random sample of size n taken from a population with mean μ and finite variance σ^2, then the limiting form of the distribution of $Z = \dfrac{\overline{X} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$ as $n \to +\infty$, is the standard normal distribution $\mathcal{N}(0, 1)$.

Note:

- The normal approximation for $\overline{X}$ will generally be good if $n \geq 30$, provided the population distribution is not terribly skewed.
- If $n < 30$, the approximation is good only if the population is not too different from a normal distribution and, as stated above, if the population is known to be normal, the sampling distribution of X will follow a normal distribution exactly, no matter how small the size of the samples.
- Figure 1 illustrates how the theorem works. It shows how the distribution of $\overline{X}$ becomes closer to normal as n grows larger, beginning with the clearly non-symmetric distribution of an individual observation ($n = 1$).

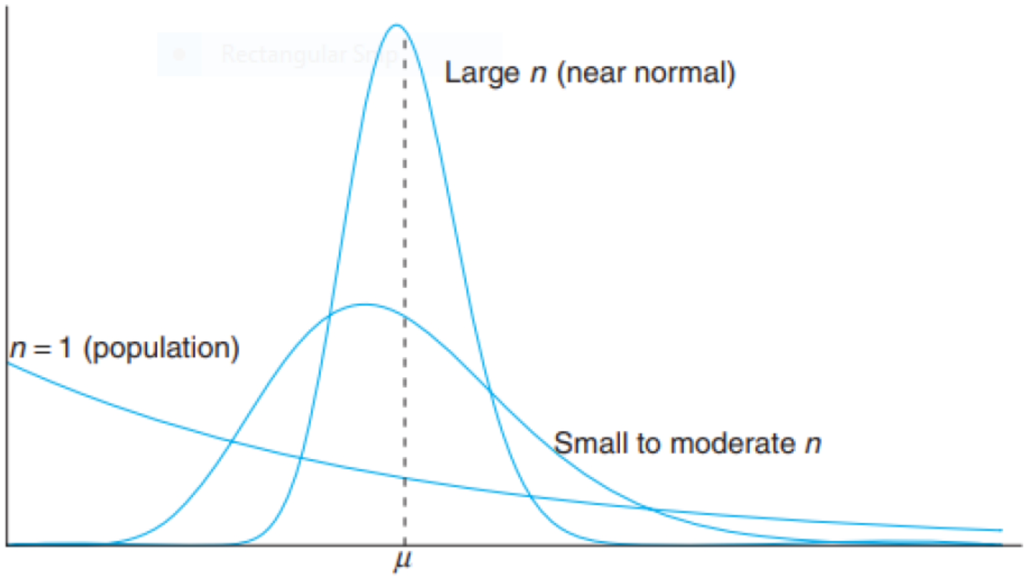


Figure 1: Distribution of $\overline{X}$ for $n = 1$, moderate n, and large n.

5.3.2 Sampling Distribution of the Difference between Two Means.

Theorem

If independent samples of size n_1 and n_2 are drawn at random from two populations, discrete or continuous, with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, respectively, then the sampling distribution of the differences of means, $X_1 - X_2$, is approximately normally distributed with mean and variance given by

$$\mu_{\overline{X}_1 - \overline{X}_2} = \mu_1 - \mu_2 \quad \text{and} \quad \sigma^2_{\overline{X}_1 - \overline{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

$$\Rightarrow Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \text{ is approximately a standard normal variable } \mathcal{N}(0, 1).$$

5.3.3 Sampling Distribution of S^2.

Theorem

If S^2 is the variance of a random sample of size n taken from a normal population having the variance σ^2, then the statistic $\chi^2 = \dfrac{(n-1)S^2}{\sigma^2} = \sum_{i=1} n \dfrac{(X_i - \overline{X})^2}{\sigma^2}$ has a chi-squared distribution with $\nu = n - 1$ degrees freedom.

Note:

- The values of χ^2 are calculated from each sample by the formula $\chi^2 = \dfrac{(n-1)s^2}{\sigma^2}$.
- It is customary to let χ^2_α represent the χ^2 value above which we find an area of α. This is illustrated by the shaded region in Figure 2.

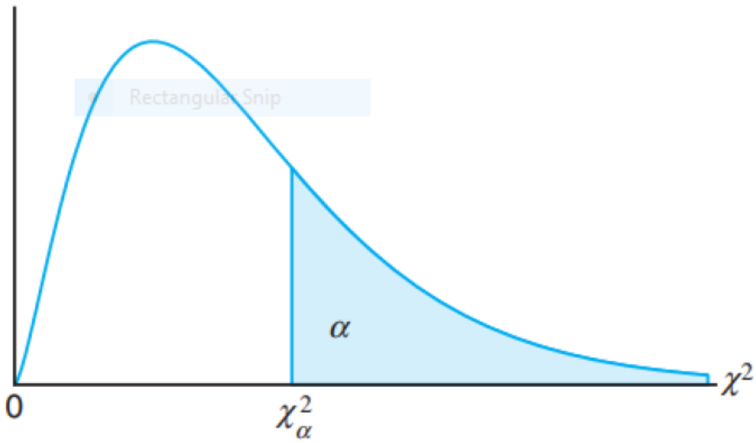


Figure 2: The chi-squared distribution.

5.3.4 Student *t*-distribution.

Theorem

Let Z be a standard normal random variable and V a chi-squared random variable with ν degrees of freedom. If Z and V are independent, then the distribution of the random variable T, where $T = \dfrac{Z}{\sqrt{\dfrac{V}{n-1}}}$ is given by the density function

$$h(t) = \frac{\dfrac{\Gamma(v+1)}{2}}{\Gamma\left(\frac{v}{2}\right)} \left(1 + \frac{t^2}{v}\right)^{-\frac{v+1}{2}}, \quad -\infty < t < +\infty.$$

This is known as the t-distribution with ν degrees of freedom.

- **Corollary:** Let $X_1, X_2, ..., X_n$ be independent random variables that are all normal with mean μ and standard deviation σ. Let

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + ... + X_n) \quad \text{and} \quad S^2 = \frac{1}{n-1} \sum_{i=1} n \left(X_i - \overline{X}\right)^2$$

Then random variable $T = \dfrac{\overline{X} - \mu}{\dfrac{S}{\sqrt{n}}}$ has a t-distribution with $\nu = n - 1$ degrees of freedom.

- Illustration of t-distribution:
  - The distribution of T is similar to the distribution of Z in that they both are symmetric about a mean of zero.
  - Both distributions are bell shaped, but the t-distribution is more variable, owing to the fact that the t-values depend on the fluctuations of two quantities, $\overline{X}$ and S^2, whereas the Z-values depend only on the changes in $\overline{X}$ from sample to sample.
  - The distribution of T differs from that of Z in that the variance of T depends on the sample size n and is always greater than 1.
  - Only when the sample size $n \to +\infty$ will the two distributions become the same.
  - In Figure 3, we show the relationship between a standard normal distribution ($\nu = \infty$) and t-distributions with 2 and 5 degrees of freedom.

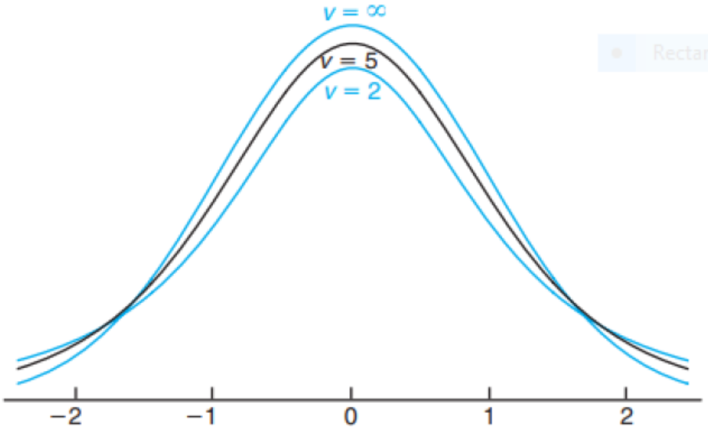
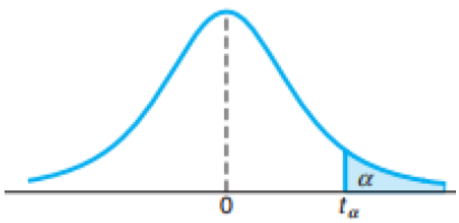


Figure 3: The t-distribution curves for $\nu = 2.5$, and ∞.

- Critical values of the t-distribution:



  - It is customary to let t_α represent the t-value above which we find an area equal to α.
  - The t-value with 10 degrees of freedom leaving an area of 0.025 to the right is $t = 2.228$.
  - Since the t-distribution is symmetric about a mean of zero, we have $t_{1-\alpha} = -t_\alpha$; that is, the t-value leaving an area of $1 - \alpha$ to the right and therefore an area of α to the left is equal to the negative t-value that leaves an area of α in the right tail of the distribution (see Figure 4. That is, $t_{0.95} = -t_{0.05}, t_{0.99} = -t_{0.01}$, and so forth.

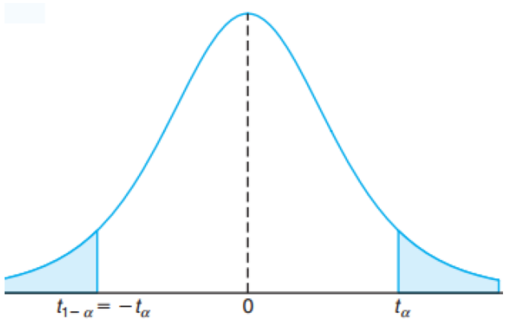


Figure 6: Symmetry property (about 0) of the t-distribution.

- What is the t-distribution used for?
  - The reader should note that use of the t-distribution for the statistic $T = \dfrac{\overline{X} - \mu}{\dfrac{S}{\sqrt{n}}}$ requires that $X_1, X_2, ..., X_n$ be normal.
  - The use of the t-distribution and the sample size consideration do not relate to the Central Limit Theorem. The use of the standard normal distribution rather than T for $n \geq 30$ merely implies that S is a sufficiently good estimator of σ in this case.

5.4 Answer of Exercises.

Problem 5.1.
The lengths of time, in minutes, that 10 patients waited in a doctor's office before receiving treatment were recorded as follows: 5, 11, 9, 5, 10, 15, 6, 10, 5, and 10. Treating the data as a random sample, find (a) the mean; (b) the median; (c) the mode.

Solution

(a) Mean: $\overline{x} = \dfrac{1}{n} \sum\limits_{i=1}^{n} x_i = 8.6.$

(b) Median: $\hat{x} = \dfrac{x_5 + x_6}{2} = 12.5$

(c) Mode: 10.

Problem 5.2.
The numbers of incorrect answers on a true-false competency test for a random sample of 15 students were recorded as follows: 2, 1, 3, 0, 1, 3, 6, 0, 3, 3, 5, 2, 1, 4, and 2. Find (a) the mean; (b) the median; (c) the mode.

Problem 5.3.
The grade-point averages of 20 college seniors selected at random from a graduating class are as follows: 3.2, 1.9, 2.7, 2.4, 2.8, 2.9, 3.8, 3.0, 2.5, 3.3, 1.8, 2.5, 3.7, 2.8, 2.0, 3.2, 2.3, 2.1, 2.5, 1.9. Calculate the standard deviation.

Solution

Standard deviation: $s = \sqrt{\dfrac{1}{n-1} \sum\limits_{i=1}^{n} \left(X_i - \overline{X}\right)^2} = 0.585$

Problem 5.4.

(a) Find $t_{0.025}$ when $\nu = 14$.

(b) Find $-t_{0.10}$ when $\nu = 10$.

(c) Find $t_{0.995}$ when $\nu = 7$.

Solution Calculate by searching in the given table.

Problem 5.5.

(a) Find $P(T < 2.365)$ when $\nu = 7$.
(b) Find $P(T > 1.318)$ when $\nu = 24$.
(c) Find $P(-1.356 < T < 2.179)$ when $\nu = 12$.
(d) Find $P(T > -2.567)$ when $\nu = 17$.

Solution

(a) With 7 degrees of freedom, we have $t_{0.025} = 2.365$, so, $P(T < 2.365) = 0.025$
(b) $P(T > 1.318) = 0.1$ when $\nu = 24$.
(c) Find $P(-1.356 < T < 2.179) = P(T < 2.179) - P(T > 1.356) = 0.875$ when $\nu = 12$.
(d) Find $P(T > -2.567) = 1 - P(T < -2.567) = 0.99$ when $\nu = 17$.

Problem 5.6.

Given a random sample of size 24 from a normal distribution, find k such that:

(a) $P(-2.069 < T < k) = 0.965$
(b) $P(k < T < 2.807) = 0.095$
(c) $P(-k < T < k) = 0.90$

Solution

Problem 5.7.

A manufacturing firm claims that the batteries used in their electronic games will last an average of 30 hours. To maintain this average, 16 batteries are tested each month. If the computed t-value falls between $-t_{0.025}$ and $t_{0.025}$, the firm is satisfied with its claim. What conclusion should the firm draw from a sample that has a mean of $x = 27.5$ hours and a standard deviation of $s = 5$ hours? Assume the distribution of battery lives to be approximately normal.

Solution

- With $\nu = 16 - 1 = 15$ degrees of freedom, we have $t_{0.025} = 2.131$.
- Thus, the interval should be $-2.131 < t < 2.131$.
- We have: $t = \dfrac{\overline{X} - \mu}{\dfrac{S}{\sqrt{n}}} = -2$. Since t satisfies the interval, the firm satisfy with the claim.

Chapter 6

One-sample Estimation Problems.

6.1 Classical methods of Estimation.

6.1.1 Point estimate.

- **Definition:** A statistic $\hat{\Theta}$ is said to be an unbiased estimator of the parameter θ if $\mu_{\hat{\Theta}} = E[\hat{\Theta}] = \theta$.
- **Most efficient estimator:** If we consider all possible unbiased estimators of some parameter θ, the one with the smallest variance is called the most efficient estimator of θ.

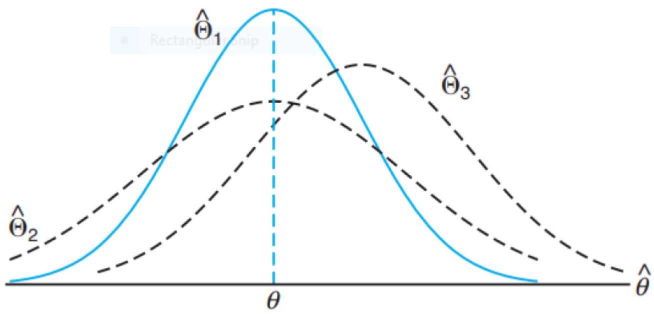


Figure 1: Sampling distributions of different estimators of θ.

Note: Figure 1 illustrates the sampling distributions of three different estimators, $\hat{\Theta}_1, \hat{\Theta}_2$ and $\hat{\Theta}_3$, all estimating θ. It is clear that only $\hat{\Theta}_1$ and $\hat{\Theta}_2$ are unbiased, since their distributions are centered at θ. The estimator $\hat{\Theta}_1$ has a smaller variance than $\hat{\Theta}_2$ and is therefore more efficient. Hence, our choice for an estimator of θ, among the three considered, would be $\hat{\Theta}_1$.

6.1.2 Interval Estimation.

- An interval estimate of a population parameter θ is an interval of the form $\hat{\theta}_L < \theta < \hat{\theta}_U$, where $\hat{\theta}_L$ and $\hat{\theta}_L$ depend on the value of the statistic $\hat{\Theta}$ for a particular sample and also on the sampling distribution of $\hat{\Theta}$.
- **Interpretation of Interval Estimates:**
  - From the sampling distribution of $\hat{\Theta}$, we shall able to determine $\hat{\Theta}_L$ and $\hat{\Theta}_U$ such that $P\left(\hat{\Theta}_L < \theta < \hat{\Theta}_U\right)$ is equal to any positive fractional value we care to specify. If, for instance, we find $\hat{\Theta}_L$ and $\hat{\Theta}_U$ such that $P\left(\hat{\Theta}_L < \theta < \hat{\Theta}_U\right) = 1 - \alpha$ for $0 < \alpha < 1$, then we have a probability of $1 - \alpha$ of selecting a random sample that will produce an interval containing θ.
  - The interval $\hat{\theta}_L < \theta < \hat{\theta}_U$, compute from the selected sample, is called a confidence interval.

- The interval $\hat{\theta}_L < \theta < \hat{\theta}_U$, compute from the selected sample, is called a $100(1-\alpha)\%$ confidence interval.
- The endpoints, $\hat{\theta}_L$ and $\hat{\theta}_U$, are called the lower and upper confidence limits.

6.2 Estimating the Mean.

6.2.1 The case of σ known.

- Two-sided intervals:
  - Writing $z_{\alpha/2}$ for the z-value which we find an area of $\alpha/2$ under the normal curve, we can see from the figure 2 that $P\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right) = 1-\alpha$, where $Z = \dfrac{-\overline{X}-\mu}{\dfrac{\sigma}{\sqrt{n}}}$.
  - Hence, $P\left(-z_{\alpha/2} < \dfrac{-\overline{X}-\mu}{\dfrac{\sigma}{\sqrt{n}}} < z_{\alpha/2}\right) = 1-\alpha$.
  - So, $P\left(\overline{X} - z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}\right) = 1-\alpha$.

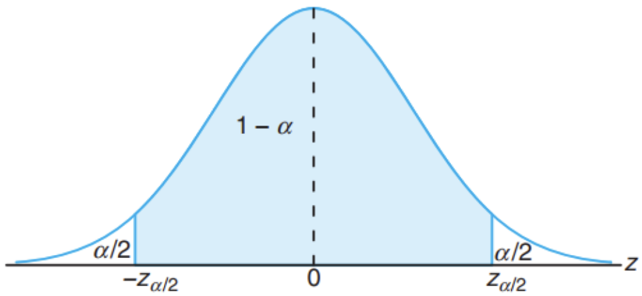


Figure 2: $P\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right) = 1-\alpha$.

Confidence Interval on μ, σ^2 known

If $\overline{x}$ is the mean of a random sample of size n from a population with known variance σ^2, a $100(1-\alpha)\%$ confidence interval for μ is given by $\overline{x} - z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}} < \mu < \overline{x} + z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$ where $z_{\alpha/2}$ is the z-value leaving an area of $\alpha/2$ to the right.

  - **Error:** If $\overline{x}$ is used as an estimate of μ, we can be $100(1-\alpha)\%$ confident that the error will not exceed $e = z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$.
  - **Size:** If $\overline{x}$ is used as an estimate of μ, $100(1-\alpha)\%$ confident that the error will not exceed a specified amount e when the sample size is $n = \left(\dfrac{z_{\alpha/2}\sigma}{e}\right)^2$.
- One-sided confidence bounds:
  - From the Central Limit Theorem: $P\left(\dfrac{\overline{X}-\mu}{\dfrac{\sigma}{\sqrt{n}}} < z_\alpha\right) = 1-\alpha$.

    One can then manipulate the probability statement much as before and obtain $P\left(\mu > \overline{X} - z_\alpha\dfrac{\sigma}{\sqrt{n}}\right) = 1-\alpha$.

– Similarly, $P\left(\dfrac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} > -z_\alpha\right) = 1-\alpha$ gives $P\left(\mu > \overline{X} + z_\alpha\dfrac{\sigma}{\sqrt{n}}\right) = 1-\alpha$.

One-sided confidence bounds on μ, σ^2 known

If $\overline{x}$ is the mean of a random sample of size n from a population with known variance σ^2, then one-sided $100(1-\alpha)\%$ confidence bounds for μ are given by:

- Upper one-sided bound: $\overline{x} + z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$.
- Lower one-sided bound: $\overline{x} - z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$.

6.2.2 The case of σ unknown.

- The random variable $T = \dfrac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}}$ has a t-distribution with $n-1$ degrees of freedom.
- The procedure is the same as that with σ known except that σ is replaced by S and the standard normal distribution is replaced by the t-distribution.

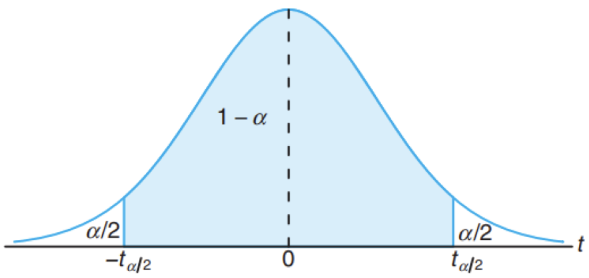


Figure 3: $P\left(-t_{\alpha/2} < T < t_{\alpha/2}\right) = 1-\alpha$ where $t_{\alpha/2}$ is the t-value with $n-1$ degrees of freedom.

- From the above,substituting for T, we write: $P\left(-t_{\alpha/2} < \dfrac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}} < t_{\alpha/2}\right) = 1-\alpha$.
- Hence, $P\left(\overline{X} - t_{\alpha/2}\dfrac{S}{\sqrt{n}} < \mu < \overline{X} + t_{\alpha/2}\dfrac{S}{\sqrt{n}}\right) = 1-\alpha$.

Confidence Interval on μ, σ^2 unknown

If $\overline{x}$ and s are the mean and standard deviation of a random sample from a population with unknown variance σ^2, a $100(1-\alpha)\%$ confidence interval for μ is given by $\overline{x} - t_{\alpha/2}\dfrac{s}{\sqrt{n}} < \mu < \overline{x} + t_{\alpha/2}\dfrac{s}{\sqrt{n}}$ where $t_{\alpha/2}$ is the t-value with $\nu = n-1$ degrees of freedom, leaving an area of $\alpha/2$ to the right.

One-sided confidence bounds on μ, σ^2 unknown

If $\overline{x}$ and s are the mean and standard deviation of a random sample from a population with unknown variance σ^2, then one-sided $100(1-\alpha)\%$ confidence bounds for μ are given by:

- Upper one-sided bound: $\overline{x} + t_{\alpha/2}\dfrac{s}{\sqrt{n}}$.
- Lower one-sided bound: $\overline{x} - t_{\alpha/2}\dfrac{s}{\sqrt{n}}$.

Here, $t_{\alpha/2}$ is the t-value having an area of α to the right.

- **Concept of a large-sample confidence interval:** Often statisticians recommend that even when normality cannot be assumed, σ is unknown, and $n \geq 30$, s can replace σ and the confidence interval $\overline{x} + z_{\alpha/2}\dfrac{s}{\sqrt{n}}$ may be used.

6.3 Estimating proportion.

- A point estimator of the proportion p in a binomial experiment is given by the statistic $\hat{P} = \dfrac{X}{n}$, where X represents the number of successes in n trials.
- Therefore, the sample proportion $\hat{p} = \dfrac{x}{n}$ will be used as the point estimate of the parameter p.
- By the Central Limit Theorem, for n sufficiently large, $\hat{P}$ is approximately normally distributed with mean $\mu_{\hat{P}} = E[\hat{P}] = E\left[\dfrac{X}{n}\right] = \dfrac{np}{n} = p$, and variance $\sigma^2_{\hat{P}} = \sigma^2_{\frac{X}{n}} = \dfrac{\sigma^2_X}{n^2} = \dfrac{npq}{n^2} = \dfrac{pq}{n}$.
- Therefore, we can assert that $P\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right) = 1-\alpha$, with $Z = \dfrac{\hat{P}-p}{\frac{pq}{n}}$ and $z_{\alpha/2}$ is the value above which we find an area of $\alpha/2$ under the standard normal curve.
- Substituting for Z, we write: $P\left(-z_{\alpha/2} < \dfrac{\hat{P}-p}{\sqrt{\frac{pq}{n}}} < z_{\alpha/2}\right) = 1-\alpha$. When n is large, very little error is introduced by substituting the point estimate $\hat{p} = \dfrac{x}{n}$ for the p under the radical sign, where $\hat{p}$ is the proportion of successes in a random sample of size n and $\hat{q} = 1-\hat{p}$, an approximate $100(1-\alpha)\%$ confidence interval, for the binomial parameter p is given by

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

or by

$$\frac{\hat{p} + \frac{z^2_{\alpha/2}}{2n}}{1 + \frac{z^2_{\alpha/2}}{2n}} - \frac{z_{\alpha/2}}{1 + \frac{z^2_{\alpha/2}}{2n}}\sqrt{\frac{\hat{p}\hat{q}}{n} + \frac{z^2_{\alpha/2}}{4n^2}} < p < \frac{\hat{p} + \frac{z^2_{\alpha/2}}{2n}}{1 + \frac{z^2_{\alpha/2}}{2n}} + \frac{z_{\alpha/2}}{1 + \frac{z^2_{\alpha/2}}{2n}}\sqrt{\frac{\hat{p}\hat{q}}{n} + \frac{z^2_{\alpha/2}}{4n^2}}$$

where $z_{\alpha/2}$ is the z-value leaving an area of $\alpha/2$ to the right.

- **Note:**
  - When n is small and the unknown proportion p is believed to be close to 0 or to 1, the confidence-interval procedure established here is unreliable and, therefore, should not be used.
  - To be on the safe side, one should require both $n\hat{p}$ and $n\hat{q}$ to be greater than or equal to 5.
  - Note that although the second formula yields more accurate results, it is more complicated to calculate, and the gain in accuracy that it provides diminishes when the sample size is large enough. Hence, the first one is commonly used in practice.
- **Error:** If $\hat{p}$ is used as an estimate of p, we can be $100(1-\alpha)\%$ confident that the error will not exceed $e = z_{\alpha/2}\sqrt{\dfrac{\hat{p}\hat{q}}{n}}$.
- **Choice of Sample size:**
  - If $\hat{p}$ is used as an estimate of p, we can be $100(1-\alpha)\%$ confident that the error will be less than a specified amount e when the sample size is approximately $n = \dfrac{z^2_{\alpha/2}\hat{p}\hat{q}}{e^2}$.
  - If $\hat{p}$ is used as an estimate of p, we can be at least $100(1-\alpha)\%$ confident that the error will not exceed a specified amount e when the sample size is $n = \dfrac{z^2_{\alpha/2}}{4e^2}$.

6.4 Answer of Exercises.

Problem 6.1.
The average zinc concentration recovered from a sample of measurements taken in 36 different locations in a river is found to be 2.6 grams per milliliter.

(a) Find the 95% and 99% confidence intervals for the mean zinc concentration in the river. Assume that the population standard deviation is 0.3 gram per milliliter.

(b) What are the errors?

(c) How large a sample is required if we want to be 95% confident that our estimate of μ is off by less than 0.05?

Solution

(a)
- The point estimate of μ is $\overline{x} = 2.6$. To find a 95% confidence interval, we find the z-value leaving an area of 0.025 to the right and therefore, an area of 0.975 to the left.
- Hence, the 95% confident interval is $2.6 - 1.96 \times \dfrac{0.3}{\sqrt{36}} < \mu < 2.6 + 1.96 \times \dfrac{0.3}{\sqrt{36}}$ which reduces to $2.50 < \mu < 2.70$.
- Similarly, the 99% confident interval is $2.471 < \mu < 2.729$.

(b)
- 95% confident that the sample mean $\hat{x} = 2.6$ differs from the true mean μ by an amount less than $1.96 \times \dfrac{0.3}{\sqrt{36}} = 0.1$.
- 99% confident that the sample mean $\hat{x} = 2.6$ differs from the true mean μ by an amount less than $1.96 \times \dfrac{0.3}{\sqrt{36}} = 0.1$.

(c) We have $n = \left(\frac{z_{\alpha/2}\sigma}{e}\right)^2 \left(\frac{1.96 \times 0.3}{0.05}\right)^2 = 138.3$. Therefore, we ccan be 95% confident that a random sample of 139 will provide an estimate $\overline{x}$ different from μ by an amount less than 0.05.

Problem 6.2.

An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed with a standard deviation of 40 hours. If a sample of 30 bulbs has an average life of 780 hours, find a 96% confidence interval for the population mean of all bulbs produced by this firm.

Solution

- Since $n = 30, \overline{x} = 780$ and $\sigma = 40$. We have $z_{\alpha/2 = z_{0.02}} = 2.054$.
- So, 96% confidence interval for the population mean of all bulbs produced by this firm is $780 - 2.054 \times \frac{40}{\sqrt{30}} < \mu < 780 + 2.054 \times \frac{40}{\sqrt{30}}$, which reduces to $765 < \mu < 795$.

Problem 6.3.

In a psychological testing experiment, 25 subjects are selected randomly and their reaction time, in seconds, to a particular stimulus is measured. Past experience suggests that the variance in reaction times to these types of stimuli is 4 sec^2 and that the distribution of reaction times is approximately normal. The average time for the subjects is 6.2 seconds. Give an upper 95% bound for the mean reaction time.

Solution

- The upper 95% bound is given by $\overline{x} + \frac{z_\alpha \sigma}{\sqrt{n}} = 6.858$.
- Hence, we are 95% confident that the mean of reaction time is less than 6.858.

Problem 6.4.

The contents of seven similar containers of sulfuric acid are 9.8, 10.2, 10.4, 9.8, 10.0, 10.2, and 9.6 liters. Find a 95% confidence interval for the mean contents of all such containers, assuming an approximately normal distribution.

Solution

- The sample mean for the given data is $\overline{x} = \frac{1}{7}(9.8 + 10.2 + ... + 9.6) = 10$.
- The standard deviation is $s = 0.283$. We have $t_{0.025} = 2.447$ for $\nu = 6$ degrees of freedom. Hence, the 95% confidence interval for μ is $10 - 2.447 \times \frac{0.283}{\sqrt{7}} < \mu < 10 + 2.447 \times \frac{0.283}{\sqrt{7}}$, which reduces to $9.74 < \mu < 10.26$.

Problem 6.5.

A machine produces metal pieces that are cylindrical in shape. A sample of pieces is taken, and the diameters are found to be 1.01, 0.97, 1.03, 1.04, 0.99, 0.98, 0.99, 1.01, and 1.03 centimeters. Find a 99% confidence interval for the mean diameter of pieces from this machine, assuming an approximately normal distribution.

Solution

- The population sample mean:

$$\overline{x} = \frac{1}{9}(1.01 + 0.97 + 1.03 + 1.04 + 0.99 + 0.98 + 0.99 + 1.01 + 1.03) = 1.0056.$$

- The popular standard deviation: s = $\sqrt{\dfrac{\sum\limits_{i=1}^{9}(x_i - \overline{x})^2}{n-1}} = 0.0246.$
- For a 98% confidence interval, we have $\alpha = 0.01$ with $\nu = n - 1 = 8$ degrees of freedom, so $t_{\alpha/2} = t_{0.005} = 3.355$.
- Hence, the interval is $1.0056 - 3.355 \times \dfrac{0.0246}{\sqrt{9}} < \mu < 1.0056 + 3.355 \times \dfrac{0.0246}{\sqrt{9}}$, which reduces to $0.9780 < \mu < 1.0331$.

Problem 6.6.

The following measurements were recorded for the drying time, in hours, of a certain brand of latex paint: 3.4, 2.5, 4.8, 2.9, 3.6, 2.8, 3.3, 5.6, 3.7, 2.8, 4.4, 4.0, 5.2, 3.0, 4.8. Assuming that the measurements represent a random sample from a normal population, find a 95% prediction interval for the drying time for the next trial of the paint.

Solution

- The population sample mean: $\overline{x} = \dfrac{1}{n}\sum\limits_{i=1}^{9} x_i = 3.7867.$
- The popular standard deviation: s = $\sqrt{\dfrac{\sum\limits_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} = 0.9709.$
- For a 95% confidence interval, we have $\alpha = 0.05$ with $\nu = n - 1 = 14$ degrees of freedom, so $t_{\alpha/2} = t_{0.025} = 2.145$.
- Hence, the interval is $3.7867 - 2.145 \times \dfrac{0.9709\sqrt{14}}{\sqrt{15}} < \mu < 3.7867 + 2.145 \times \dfrac{0.9709\sqrt{14}}{\sqrt{15}}$, which reduces to $1.7749 < \mu < 5.7985$.

Problem 6.7.

Scholastic Aptitude Test (SAT) mathematics scores of a random sample of 500 high school seniors in the state of Texas are collected, and the sample mean and standard deviation are found to be 501 and 112, respectively. Find a 99% confidence interval on the mean SAT mathematics score for seniors in the state of Texas.

Solution

- We have 99% confidence interval, so $z_{\alpha/2} = z_{0.005} = 2.575$.
- Therefore, a 99% confidence interval is 501 ± 12.9.

> **Problem 6.8.**
> The heights of a random sample of 50 college students showed a mean of 174.5 centimeters and a standard deviation of 6.9 centimeters.
>
> (a) Construct a 98% confidence interval for the mean height of all college students.
>
> (b) What can we assert with 98% confidence about the possible size of our error if we estimate the mean height of all college students to be 174.5 centimeters?

Solution

(a)
- Since we need to estimate 98% confidence interval, we have $\alpha = 0.02$, hence, $z_{\alpha/2} = z_{0.01} = 2.33$.
- Therefore, the interval is $174.5 - 2.33 \times \dfrac{6.9}{\sqrt{50}} < \mu < 174.5 + 2.33 \times \dfrac{6.9}{\sqrt{50}}$, which can reduces to $172.23 < \mu < 176.77$.

(b) The mean height is $e = 2.33 \times \dfrac{6.9}{\sqrt{50}} = 2.27$.

> **Problem 6.9.**
> In a random sample of $n = 500$ families owning television sets in the city of Hamilton, Canada, it is found that $m = 340$ subscribe to HBO.
>
> (a) Find a 95% confidence interval for the actual proportion of families with television sets in this city that subscribe to HBO.
>
> (b) What is error?
>
> (c) How large a sample is required if we want to be 95% confident that our estimate of p is within 0.02 of the true value?

Solution

(a)
- The point estimate of p is $\hat{p} = \dfrac{340}{500} = 0.68$. We find $z_{0.025} = 1.196$.
- The 95% confidence interval for p is $0.68 - 1.96 \times \sqrt{\dfrac{0.68 \times 0.32}{500}} < p < 0.68 + 1.96 \times \sqrt{\dfrac{0.68 \times 0.32}{500}}$, which reduces to $0.6391 < p < 0.7209$.

(b)
- We are 95% confident that the sample proportion $\hat{p} = 0.68$ differs from the true proportion p by an amount not exceeding: $e = 1.96 \times \sqrt{\dfrac{0.68 \times 0.32}{500}} = 0.04$.

(c) $n = \dfrac{1.96^2 \times 0.68 \times 0.32}{(0.02)^2} = 2089.8 \approx 2090$. Therefore, a random sample of the size 2090 is required it we want to be 95% confident that our estimate of p is within 0.02 of the true value.

Problem 6.10.
In a random sample of 1000 homes in a certain city, it is found that 228 are heated by oil. Find 99% confidence intervals for the proportion of homes in this city that are heated by oil.

Solution

- The sample proportion is $\hat{p} = \dfrac{228}{1000} = 0.228$.
- Since we need to calculate the 99% confidence interval, so $z_{\alpha/2} = z_{0.005} = 2.575$.
- Hence, the interval is $0.194 < p < 0.262$.

Problem 6.11.
(a) A random sample of 200 voters in a town is selected, and 114 are found to support an annexation suit. Find the 96% confidence interval for the fraction of the voting population favoring the suit.
(b) What can we assert with 96% confidence about the possible size of our error if we estimate the fraction of voters favoring the annexation suit to be 0.57?

Solution

(a)
- The proportion of success in the random sample is $\hat{p} = \dfrac{114}{200} = 0.57$, so $\hat{q} = 1 - \hat{p} = 0.43$.
- We have 96% confidence interval so $z_{\alpha/2} = z_{0.02} = 2.05$, so the interval is $\hat{p} - z_{\alpha/2}\sqrt{\dfrac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\dfrac{\hat{p}\hat{q}}{n}}$, which reduces to $0.498 < p < 0.642$.

(b) The size of the error is: $e \leq z_{\alpha/2}\sqrt{\dfrac{\hat{p}\hat{q}}{n}} = 0.072$.

Problem 6.12.
A geneticist is interested in the proportion of African males who have a certain minor blood disorder. In a random sample of 100 African males, 24 are found to be afflicted.
(a) Compute a 99% confidence interval for the proportion of African males who have this blood disorder.
(b) What can we assert with 99% confidence about the possible size of our error if we estimate the proportion of African males with this blood disorder to be 0.24?

Solution

(a)
- The proportion of success in the random sample is $\hat{p} = \dfrac{24}{100} = 0.24$, so $\hat{q} = 1 - \hat{p} = 0.76$.
- We have 96% confidence interval so $z_{\alpha/2} = z_{0.005} = 2.575$, so the interval is $\hat{p} - z_{\alpha/2}\sqrt{\dfrac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\dfrac{\hat{p}\hat{q}}{n}}$, which reduces to $0.130 < p < 0.350$.

(b) The size of the error is: $e \leq z_{\alpha/2}\sqrt{\dfrac{\hat{p}\hat{q}}{n}} = 0.110$.

Chapter 7

Hypothesis Testing.

7.1 Introduction.

7.1.1 Key terms and concepts.

- Null hypothesis. Alternative hypothesis.
  - The two competing hypotheses are the *alternative hypothesis* H_1, generally the hypothesis that the researcher wishes to support, and the *null hypothesis* H_0, a contradiction of the alternative hypothesis.
  - As you will soon see, it is easier to show support for the alternative hypothesis by proving that the null hypothesis is false. Hence, the statistical researcher always begins by *assuming that the null hypothesis* H_0 is true. The researcher then uses the sample data to decide whether the evidence favors H_1 rather than H_0 and draws one of these two conclusions:
    * Reject H_0 and conclude that H_1 is true.
    * Accept (do not reject) H_0 as true.
  - Level of significance: Refers to the degree of significance in which we accept or reject the null-hypothesis. 100% accuracy is not possible for accepting or rejecting a hypothesis, so we therefore select a level of significance that is usually 5%.
- Error:
  - Type I error:
    * When we reject the null hypothesis, although that hypothesis was true.
    * $P[\text{Type I error}] = \alpha$.
    * In hypothesis testing, the normal curve that shows the critical region is called the α region.
  - Type II error:
    * When we accept the null hypothesis but it is false.
    * $P[\text{Type II error}] = \beta$.
    * In hypothesis testing, the normal curve that shows the critical region is called the β region.
  - Power: Usually known as the probability of correctly accepting the null hypothesis. $1 - \beta$ is called power of the analysis.

	H_0 **is true**	H_0 **is false**
Do not reject H_0	Correct decision	Type II error
Reject H_0	Type I error	Correct decision

- Test:

- One-tailed test: A one-tailed test is a statistical test in which the critical area of a distribution is one-sided so that it is either greater than or less than a certain value, but not both.
- Two-tailed test: A two-tailed test is a method in which the critical area of a distribution is two-sided and tests whether a sample is greater than or less than a certain range of values.

7.1.2 Statistical decision for hypothesis testing.

- Statistical decision for hypothesis testing:
  - In statistical analysis, we have to make decisions about the hypothesis. These decisions include deciding:
    - if we should accept the null hypothesis;
    - if we should reject the null hypothesis.
  - The rejection rule is as follows:
    - if the standardized test statistic is not in the rejection region, then we accept the null hypothesis;
    - if the standardized test statistic is in the rejection region, then we should reject the null hypothesis.
- The rejection region is values of test statistic for which the null hypothesis is rejected.
- Approach to Hypothesis Testing with Fixed Probability of Type I Error:
  1. State the null and alternative hypotheses.
  2. Choose a fixed significance level α.
  3. Choose an appropriate test statistic and establish the critical region/rejection region based on α.
  4. Reject H_0 if the computed test statistic is in the critical region/rejection region. Otherwise, do not reject.
  5. Draw scientific or engineering conclusions.

7.2 Single Sample.

7.2.1 Tests on a Single Mean (Variance Known).

- Two-tailed test:
  1. Identify the null and alternative hypotheses: $H_0 : \mu = \mu_0 \quad ; \quad H_1 : \mu \neq \mu_0$.
  2. Find the standardized test statistic $z = \dfrac{\bar{x} - \mu_0}{\sigma}\sqrt{n}$.
  3. Make a decision to reject or fail to reject the null hypothesis:
     - If $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$, then reject H_0.
     - If $-z_{\alpha/2} < z < z_{\alpha/2}$, then fail to reject H_0.
- One-tailed test (right-tailed test):
  1. Identify the null and alternative hypotheses: $H_0 : \mu = \mu_0 \quad ; \quad H_1 : \mu > \mu_0$.
  2. Find the standardized test statistic $z = \dfrac{\bar{x} - \mu_0}{\sigma}\sqrt{n}$.
  3. Make a decision to reject or fail to reject the null hypothesis:
     - If $z > z_\alpha$, then reject H_0.
     - If $z < z_\alpha$, then fail to reject H_0.
- One-tailed test (left-tailed test):

1. Identify the null and alternative hypotheses: $H_0 : \mu = \mu_0 \quad ; \quad H_1 : \mu < \mu_0$.
2. Find the standardized test statistic $z = \frac{\bar{x} - \mu_0}{\sigma}\sqrt{n}$.
3. Make a decision to reject or fail to reject the null hypothesis:
   - If $z < -z_\alpha$, then reject H_0.
   - If $z > -z_\alpha$, then fail to reject H_0.

7.2.2 Tests on a Single Sample (Variance Unknown).

- Two-tailed test:
  1. Identify the null and alternative hypotheses: $H_0 : \mu = \mu_0 \quad ; \quad H_1 : \mu \neq \mu_0$.
  2. Find the standardized test statistic $t = \frac{\bar{x} - \mu_0}{s}\sqrt{n}$.
  3. Make a decision to reject or fail to reject the null hypothesis:
     - If $t < -t_{\alpha/2}^{(n-1)}$ or $t > t_{\alpha/2}^{(n-1)}$, then reject H_0.
     - If $-t_{\alpha/2}^{(n-1)} < t < t_{\alpha/2}^{(n-1)}$, then fail to reject H_0.
- One-tailed test (right-tailed test):
  1. Identify the null and alternative hypotheses: $H_0 : \mu = \mu_0 \quad ; \quad H_1 : \mu > \mu_0$.
  2. Find the standardized test statistic $t = \frac{\bar{x} - \mu_0}{s}\sqrt{n}$.
  3. Make a decision to reject or fail to reject the null hypothesis:
     - If $t > t_{\alpha}^{(n-1)}$, then reject H_0.
     - If $t < t_{\alpha}^{(n-1)}$, then fail to reject H_0.
- One-tailed test (left-tailed test):
  1. Identify the null and alternative hypotheses: $H_0 : \mu = \mu_0 \quad ; \quad H_1 : \mu < \mu_0$.
  2. Find the standardized test statistic $t = \frac{\bar{x} - \mu_0}{s}\sqrt{n}$.
  3. Make a decision to reject or fail to reject the null hypothesis:
     - If $t < -t_{\alpha}^{(n-1)}$, then reject H_0.
     - If $t > -t_{\alpha}^{(n-1)}$, then fail to reject H_0.

7.2.3 Hypothesis testing for proportions.

- Two-tailed test:
  1. Identify the null and alternative hypotheses: $H_0 : \mu = \mu_0 \quad ; \quad H_1 : \mu \neq \mu_0$.
  2. Find the standardized test statistic $z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)}}\sqrt{n}$.
  3. Make a decision to reject or fail to reject the null hypothesis:
     - If $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$, then reject H_0.
     - If $-z_{\alpha/2} < z < z_{\alpha/2}$, then fail to reject H_0.
- One-tailed test (right-tailed test):
  1. Identify the null and alternative hypotheses: $H_0 : \mu = \mu_0 \quad ; \quad H_1 : \mu > \mu_0$.
  2. Find the standardized test statistic $z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)}}\sqrt{n}$.
  3. Make a decision to reject or fail to reject the null hypothesis:
     - If $z > z_\alpha$, then reject H_0.
     - If $z < z_\alpha$, then fail to reject H_0.

- One-tailed test (left-tailed test):
  1. Identify the null and alternative hypotheses: $H_0 : \mu = \mu_0 \quad ; \quad H_1 : \mu < \mu_0$.
  2. Find the standardized test statistic $z = \dfrac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}$.
  3. Make a decision to reject or fail to reject the null hypothesis:
     - If $z < -z_\alpha$, then reject H_0.
     - If $z > -z_\alpha$, then fail to reject H_0.

7.3 Two Sample: Test on Two Means.

- Null and Alternative Hypothesis:
  - In a two-sample hypothesis test, two parameters from two populations arecompared.
  - The null hypothesis H_0 is a statistical hypothesis that usually states there is no difference between the parameters of two populations. The null hypothesis always contains the symbol "=".
  - The alternative hypothesis H_1 is a statistical hypothesis that is true when H_0 is false. The alternative hypothesis always contains the symbol ">, $\neq$, <".
- Null and Alternative Hypothesis:

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases} ; \quad \begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 > \mu_2 \end{cases} ; \quad \begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 < \mu_2 \end{cases}$$

  Regardless of which hypotheses used, $\mu_1 = \mu_2$ is always assumed to be true.

7.3.1 σ_1^2 and σ_2^2 are known.

- Null hypothesis: $H_0 : \mu_1 = \mu_2 = D_0$, where D_0 is some specified difference that you wish to test. For many tests, you will hypothesize that there is no difference between μ_1 and μ_2; that is, $D_0 = 0$.
- Alternative hypothesis:
  - One-tailed test: $H_1 : \mu_1 - \mu_2 > D_0$ or $\mu_1 - \mu_2 < D_0$.
  - Two-tailed test: $H_2 : \mu_1 - \mu_2 \neq D_0$.
- Test statistic: $z = \dfrac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$.
- Rejection region: Reject H_0 when
  - One-tailed test:
    * $z > z_\alpha$ (when the alternative hypothesis is $H_1 : \mu_1 - \mu_2 > D_0$).
    * $z < -z_\alpha$ (when the alternative hypothesis is $H_1 : \mu_1 - \mu_2 < D_0$).
  - Two-tailed test: $z > z_{\alpha/2}$ or $z < -z_{\alpha/2}$.

7.3.2 σ_1^2 and σ_2^2 are unknown, $n_1, n_2 \geq 30$.

- Null hypothesis: $H_0 : \mu_1 = \mu_2 = D_0$, where D_0 is some specified difference that you wish to test. For many tests, you will hypothesize that there is no difference between μ_1 and μ_2; that is, $D_0 = 0$.

- Alternative hypothesis:
  - One-tailed test: $H_1 : \mu_1 - \mu_2 > D_0$ or $\mu_1 - \mu_2 < D_0$.
  - Two-tailed test: $H_2 : \mu_1 - \mu_2 \neq D_0$.
- Test statistic: $z = \dfrac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$.
- Rejection region: Reject H_0 when
  - One-tailed test:
    * $z > z_\alpha$ (when the alternative hypothesis is $H_1 : \mu_1 - \mu_2 > D_0$).
    * $z < -z_\alpha$ (when the alternative hypothesis is $H_1 : \mu_1 - \mu_2 < D_0$).
  - Two-tailed test: $z > z_{\alpha/2}$ or $z < -z_{\alpha/2}$.

7.3.3 σ_1^2 and σ_2^2 are unknown, $n_1, n_2 < 30$.

- If samples of size less than 30 are taken from normally-distributed populations, a t-test may be used to test the difference between the population means μ_1 and μ_2.
- Three conditions are necessary to use a t-test for small independent samples:
  - The samples must be randomly selected.
  - The samples must be independent.
  - Each population must have a normal distribution and population variances are equal.
- Null hypothesis: $H_0 : \mu_1 = \mu_2 = D_0$, where D_0 is some specified difference that you wish to test. For many tests, you will hypothesize that there is no difference between μ_1 and μ_2; that is, $D_0 = 0$.
- Alternative hypothesis:
  - One-tailed test: $H_1 : \mu_1 - \mu_2 > D_0$ or $\mu_1 - \mu_2 < D_0$.
  - Two-tailed test: $H_2 : \mu_1 - \mu_2 \neq D_0$.
- Test statistic: $t = \dfrac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}\left(\dfrac{1}{n_2} + \dfrac{1}{n_2}\right)}}$.
- Rejection region: Reject H_0 when
  - One-tailed test:
    * $t > t_\alpha^{(n_1+n_2-2)}$ (when the alternative hypothesis is $H_1 : \mu_1 - \mu_2 > D_0$).
    * $t < -t_\alpha^{(n_1+n_2-2)}$ (when the alternative hypothesis is $H_1 : \mu_1 - \mu_2 < D_0$).
  - Two-tailed test: $t > t_{\alpha/2}^{(n_1+n_2-2)}$ or $t < -t_{\alpha/2}^{(n_1+n_2-2)}$.

7.4 Testing the Difference Between Proportions.

- Two Sample z-Test for Proportions: A z-test is used to test the difference between two population proportions, p_1 and p_2. Three conditions are required to conduct the test.
  1. The samples must be randomly selected.
  2. The samples must be independent.
  3. The samples must be large enough to use a normal sampling distribution. That is, $n_1 p_1 \geq 5$; $n_1(1-p_1) \geq 5$; $n_2 p_2 \geq 5$; $n_2(1-p_2) \geq 5$.

- Mean and Standard error: If these conditions are met, then the sampling distribution for $\hat{p}_1 - \hat{p}_2$ is a normal distribution with:
  - Mean $\mu_{\hat{p}_1-\hat{p}_2} = p_1 - p_2$.
  - Standard error: $\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\overline{p}(1-\overline{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$.
  - A weighted estimate of p_1 and p_2 can be found by using $\overline{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1+n_2}$.
- Standardized test statistic: A two sample z-test is used to test the difference between two population proportions p_1 and p_2 when a sample is randomly selected from each population. The standardized test statistic is $z = \frac{(\hat{p}_1-\hat{p}_2)-(p_1-p_2)}{\sqrt{\overline{p}(1-\overline{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$.
- Steps for hypothesis testing:
  - **Step 1.** State the claim mathematically and verbally. Identify the null and alternative hypotheses.

Null hypothesis H_0	$p_1 = p_2$	$p_1 = p_2$	$p_1 = p_2$
Alternative hypothesis H_1	$p_1 \neq p_2$	$p_1 \leq p_2$	$p_1 \geq p_2$

  - **Step 2.** Find the standardized test statistic: $z = \frac{(\hat{p}_1-\hat{p}_2)-(p_1-p_2)}{\frac{n_1\hat{p}_1+n_2\hat{p}_2}{n_1+n_2}}$.
  - **Step 3.** Determine the rejection region:

H_0	H_1	Rejection region W_α
$p_1 = p_2$	$p_1 \neq p_2$	$(-\infty; -z_{\alpha/2}) \cup (z_{\alpha/2}; +\infty)$
$p_1 = p_2$	$p_1 \leq p_2$	$(z_\alpha; +\infty)$
$p_1 = p_2$	$p_1 \geq p_2$	$(-\infty; -z_\alpha)$

    where $z_{\alpha/2}$ and $z_{1-\alpha}$ are given.
  - **Step 4.** Make a decision to reject or fail to reject the null hypothesis.
  - **Step 5.** Interpret the decision in the context of the original claim.

7.5 Answer of Exercises.

Problem 7.1.

The average weekly earnings for female social workers is \$670. Do men in the same positions have average weekly earnings that are higher than those for women? A random sample of $n = 40$ male social workers showed $\overline{x} = \$725$. Assuming a population standard deviation of \$102, test the appropriate hypothesis using $\alpha = 0.01$.

Solution

- Null and alternative hypothesis: $\begin{cases} H_0 : \mu = 670, \\ H_1 : \mu > 670 \end{cases}$.
- We have $z = \frac{\overline{x}-\mu_0}{\sigma}\sqrt{n} = 3.41$.
- With $\alpha = 0.01 \Rightarrow z_\alpha = z_{0.01} = 2.33$.

- Since the value of the test statistic is $z = 3.41$, greater than the critical value $z_\alpha = 2.33$. Thus, we can reject H_0. Therefore, the average weekly earnings for male social workers are higher than that of female.

Problem 7.2.

A random sample of 64 bags of white cheddar popcorn weighed, on average, 5.23 ounces with a standard deviation of 0.24 ounce. Test the hypothesis that $\mu = 5.5$ ounces against the alternative hypothesis, $\mu < 5.5$ ounces, at the 0.05 level of significance.

Solution

- Null and alternative hypothesis $\begin{cases} H_0 : \mu = 5.5, \\ H_1 : \mu > 5.5 \end{cases}$.
- We have $z = \dfrac{\overline{x} - \mu_0}{\sigma}\sqrt{n} = -9$.
- With $\alpha = 0.05 \Rightarrow z_\alpha = z_{0.05} = 1.645$.
- Since the value of the test statistic is $z < -z_\alpha$. Thus, we can reject H_0. Therefore, we can conclude that the alternative hypothesis that $\mu < 5.5$ ounces is true.

Problem 7.3.

A local telephone company claims that the average length of a phone call is 8 minutes. In a random sample of 18 phone calls, the sample mean was 7.8 minutes and the standard deviation was 0.5 minutes. Is there enough evidence to support this claim at $\alpha = 0.05$?

Solution

- Null and alternative hypothesis $\begin{cases} H_0 : \mu = 8, \\ H_1 : \mu \neq 8 \end{cases}$. The test is a two-tailed test.
- We compute $t = \dfrac{\overline{x} - \mu_0}{s}\sqrt{n} = -1.70$.
- With $\alpha = 0.05 \Rightarrow$, we have critical region $t < -2.110$ or $t > 2.110$ where $t = \dfrac{\overline{x} - \mu_0}{s}\sqrt{n}$ with 17 degrees freedom.
- Thus, we cannot reject H_0. Therefore, at the 5% level of significance, there is not enough evidence to reject the average length of a phone call in 8 minutes.

Problem 7.4.

Test the hypothesis that the average content of containers of a particular lubricant is 10 liters if the contents of a random sample of 10 containers are 10.2, 9.7, 10.1, 10.3, 10.1, 9.8, 9.9, 10.4, 10.3, and 9.8 liters. Use a 0.01 level of significance and assume that the distribution of contents is normal.

Solution

- Null and alternative hypothesis $\begin{cases} H_0 : \mu = 10, \\ H_1 : \mu \neq 10 \end{cases}$. This is a two-tailed test.
- We compute $z = \dfrac{\overline{x} - \mu_0}{s}\sqrt{n} = 0.775$.

- With $\alpha = 0.01 \Rightarrow$, we have critical region $z < -3.25$ or $z > 3.25$. Thus, we cannot reject H_0. Therefore, at the 1% level of significance, there is not enough evidence to support the claim.

Problem 7.5.
According to a dietary study, high sodium intake may be related to ulcers, stomach cancer, and migraine headaches. The human requirement for salt is only 220 milligrams per day, which is surpassed in most single servings of ready-to-eat cereals. If a random sample of 20 similar servings of a certain cereal has a mean sodium content of 244 milligrams and a sample standard deviation of 24.5 milligrams, does this suggest at the 0.05 level of significance that the average sodium content for a single serving of such cereal is greater than 220 milligrams? Assume the distribution of sodium content to be normal.

Solution

- Null and alternative hypothesis $\begin{cases} H_0 : \mu = 224, \\ H_1 : \mu \neq 224 \end{cases}$. The test is a two-tailed test.
- We compute $t = \dfrac{\overline{x} - \mu_0}{s}\sqrt{n} = 4.381$.
- With $\alpha = 0.05 \Rightarrow$ and $\nu = n - 1 = 19$ degrees of freedom, we have critical region $t < -1.729$ or $t > 1.729$ where $t = \dfrac{\overline{x} - \mu_0}{s}\sqrt{n}$.
- Thus, we can reject H_0. Therefore, at the 5% level of significance, there is enough evidence to claim that the average sodium content for a single serving of such cereal is greater than 220 milligrams.

Problem 7.6.
The daily yield for a local chemical plant has averaged 880 tons for the last several years. The quality control manager would like to know whether this average has changed in recent months. She randomly selects 50 days from the computer database and computes the average and sample standard deviation of the $n = 50$ yields as $\overline{x} = 871$ tons and $s = 21$ tons, respectively. Test the appropriate hypothesis using $\alpha = 0.05$.

Solution

- Null and alternative hypothesis $\begin{cases} H_0 : \mu = 880, \\ H_1 : \mu \neq 880 \end{cases}$. The test is a two-tailed test.
- We have $z = \dfrac{\overline{x} - \mu_0}{\sigma}\sqrt{n} = -3.03$.
- With $\alpha = 0.05 \Rightarrow$ the critical values are $z_{\alpha/2} = z_{0.025} = 1.96$, $-z_{\alpha/2} = -z_{0.025} = -1.96$ and the null hypothesis will be rejected if $z > 1.96$ or $z < -1.96$.
- Since $z = -3.03$, the manager can reject the null hypothesis that $\mu = 880$ tons and conclude that it has changed. The probability of rejecting H_0 when H_0 is true and $\alpha = 0.05$, a fairly small probability.
- Hence, she is reasonably confident that the decision is correct.

Problem 7.7.

A college claims that more than 94% of their graduates find employment within 6 months of graduation. In a sample of 500 randomly selected graduates, 475 of them were employed. Is there enough evidence to support the college's claim at a 1% level of significance?

Solution

- Null and alternative hypothesis $\begin{cases} H_0 : p = 0.94, \\ H_1 : p > 0.94 \end{cases}$. The test is a right-tailed test.
- We have $\hat{p} = \dfrac{475}{500} = 0.95$.
- The standardized test statistic: $z = \dfrac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)}}\sqrt{n} = 0.94$.
- The critical value $z_\alpha = z_{0.01} = 2.33$.
- SInce $z = 0.94 < 2.33 = z_\alpha \Rightarrow H_0$ is not rejected. We can conclude that at the 1% level of significance, there is not enough evidence to support the college's claim.

Problem 7.8.

A cigarette manufacturer claims that 1/8 of the US adult population smokes cigarettes. In a random sample of 100 adults, 5 are cigarette smokers. Test the claim at $\alpha = 0.05$.

Solution

- Null and alternative hypothesis $\begin{cases} H_0 : p = 0.94, \\ H_1 : p \neq 0.94 \end{cases}$. The test is a two-tailed test.
- We have $\hat{p} = \dfrac{5}{100} = 0.05$.
- The standardized test statistic: $z = \dfrac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)}}\sqrt{n} = -2.27$.
- The critical value $z_\alpha = z_{0.05} = -1.60$.
- SInce $z = -1.60 > -2.27 = z_\alpha \Rightarrow H_0$ is rejected. We can conclude that at the 5% level of significance, there is enough evidence to support the claim.

Problem 7.9.

A marketing expert for a pasta-making company believes that 40% of pasta lovers prefer lasagna. If 9 out of 20 pasta lovers choose lasagna over other pastas, what can be concluded about the expert's claim? Use a 0.05 level of significance.

Solution

- Null and alternative hypothesis $\begin{cases} H_0 : p = 0.4, \\ H_1 : p \neq 0.4 \end{cases}$. The test is a two-tailed test.
- We have $\hat{p} = \dfrac{9}{20} = 0.45$.
- The standardized test statistic: $z = \dfrac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)}}\sqrt{n} = 0.45$.

- The critical values are $z_{\alpha/2} = z_{0.025} = 1.96$ and $-z_{\alpha/2} = -z_{0.025} = -1.96$.
- SInce $-z_{\alpha/2} < z < z_{\alpha/2} \Rightarrow H_0$ is not rejected. We can conclude that at the 5% level of significance, there is enough evidence to support the marketing expert's claim.

Problem 7.10.

It is believed that at least 60% of the residents in a certain area favor an annexation suit by a neighboring city. What conclusion would you draw if only 110 in a sample of 200 voters favored the suit? Use a 0.05 level of significance.

Solution

- Null and alternative hypothesis $\begin{cases} H_0 : p = 0.6, \\ H_1 : p > 0.6 \end{cases}$. The test is a right-tailed test.
- We have $\hat{p} = \dfrac{110}{200} = 0.55$.
- The standardized test statistic: $z = \dfrac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n} = -1.44$.
- The critical values is $z_\alpha = z_{0.05} = 1.645$.
- SInce $z < z_{\alpha/2} \Rightarrow H_0$ is not rejected. We can conclude that at the 5% level of significance, there is enough evidence to support the claim.

Problem 7.11.

A high school math teacher claims that students in her class will score higher on the math portion of the ACT then students in a colleague's math class. The mean ACT math score for 49 students in her class is 22.1 and the sample standard deviation is 4.8. The mean ACT math score for 44 of the colleague's students is 19.8 and the sample standard deviation is 5.4. At $\alpha = 0.10$, can the teacher's claim be supported?

Solution

- Let μ_1 and μ_2 represent the population means of the ACT math in two classes respectively. Null and alternative hypothesis $\begin{cases} H_0 : \mu_1 = \mu_2, \\ H_1 : \mu_1 > \mu_2 \end{cases}$.
- The standardized error is $\sigma_{\overline{X}_1 - \overline{X}_2} = \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}} \approx 1.0644$.
- The standardized test statistic is $z = \dfrac{22.1 - 19.8}{1.0644} \approx 2.161$.
- With $\alpha = 0.10$, the critical value $z_\alpha = 1.28 \Rightarrow$ we can reject H_0. Therefore, there is enough evidence at 10% level to support the teacher's claim.

Problem 7.12.

To determine whether car ownership affects a student's academic achievement, two random samples of 100 male students were each drawn from the student body. The grade point average for the $n_1 = 100$ non-owners of cars had an average and variance equal to $\overline{x}_1 = 2.70$ and $s_1^2 = 0.36$, while $\overline{x}_2 = 2.54$ and $s_2^2 = 0.40$ for the $n_2 = 100$ car owners. Do the data present sufficient evidence to indicate a difference in the mean achievements between car owners and non-owners of cars? Test using $\alpha = 0.05$.

Solution

- Let μ_1 and μ_2 represent the population means of the achievement in two groups of student respectively. Null and alternative hypothesis $\begin{cases} H_0 : \mu_1 = \mu_2, \\ H_1 : \mu_1 \neq \mu_2 \end{cases}$.
- The standardized error is $\sigma_{\overline{X}_1 - \overline{X}_2} = \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}} \approx 0.0871$.
- The standardized test statistic is $z = \dfrac{2.70 - 2.54}{0.0871} \approx 1.835$.
- With $\alpha = 0.05$, the critical values are $\pm z_{\alpha/2} = 1.96 \Rightarrow$ we can reject H_0. Therefore, there is enough evidence at 5% level to support the claim.

Problem 7.13.

A manufacturer claims that the average tensile strength of thread A exceeds the average tensile strength of thread B by at least 12 kilograms. To test this claim, 50 pieces of each type of thread were tested under similar conditions. Type A thread had an average tensile strength of 86.7 kilograms with a standard deviation of 6.28 kilograms, while type B thread had an average tensile strength of 77.8 kilograms with a standard deviation of 5.61 kilograms. Test the manufacturer's claim using a 0.05 level of significance.

Solution

- Let μ_1 and μ_2 represent the population means of the achievement in two groups of student respectively. Null and alternative hypothesis $\begin{cases} H_0 : \mu_1 - \mu_2 = 12, \\ H_1 : \mu_1 - \mu_2 < 12 \end{cases}$. This is left-tailed test.
- The standardized error is $\sigma_{\overline{X}_1 - \overline{X}_2} = \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}} \approx 1.191$.
- The standardized test statistic is $z = \dfrac{(\overline{x_1} - \overline{x_2}) - D_0}{\sigma_{\overline{X}_1 - \overline{X}_2}} \approx -2.603$.
- With $\alpha = 0.05$, the critical values is $z_\alpha = 1.64 \Rightarrow$ we can reject H_0. Therefore, there is enough evidence at 5% level to support the claim.

Problem 7.14.

Engineers at a large automobile manufacturing company are trying to decide whether to purchase brand A or brand B tires for the company's new models. To help them arrive at a decision, an experiment is conducted using 12 of each brand. The tires are run until they wear out. The results are as follows:

- Brand A: $x_A = 37,900$ kilometers, $s_A = 5100$ kilometers.
- Brand B: $x_B = 39,800$ kilometers, $s_B = 5900$ kilometers.

Test the hypothesis that there is no difference in the average wear of the two brands of tires. Assume the populations to be approximately normally distributed with equal variances. Use a 0.01 level of significance.

Solution

- Let μ_1 and μ_2 represent te population means of the average wear of the two brands A and B respectively. Null and alternative hypothesis $\begin{cases} H_0 : \mu_1 = \mu_2, \\ H_1 : \mu_1 \neq \mu_2 \end{cases}$.
- The standardized test statistic is $t = \dfrac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}\left(\dfrac{1}{n_2} + \dfrac{1}{n_2}\right)}} = -0.84.$
- With $\alpha = 0.01 \Rightarrow$ The critical value are $\pm t_{\alpha/2} = \pm 2.819$. Hence, $-t_{\alpha/2} < t < t_{\alpha/2}$, we cannot reject H_0.

Problem 7.15.

A recent survey stated that male college students smoke less than female college students. In a survey of 1245 male students, 361 said they smoke at least one pack of cigarettes a day. In a survey of 1065 female students, 341 said they smoke at least one pack a day. At $\alpha = 0.01$, can you support the claim that the proportion of male college students who smoke at least one pack of cigarettes a day is lower then the proportion of female college students who smoke at least one pack a day?

Solution

- Let p_1 and p_2 represent the population proportions of male and female college students respectively. Null and alternative hypothesis $\begin{cases} H_0 : p_1 = p_2, \\ H_1 : p_1 < p_2 \end{cases}$.
- We compute $\overline{p} = \dfrac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1+n_2} = 0.304 \Rightarrow 1 - \overline{p} = 0.696.$
- Since $1245 \times 0.304,\ 1245 \times 0.696,\ 1065 \times 0.304,\ 1065 \times 0.696$ are all at least 5, we can use two-sample z-test: $z = \dfrac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\overline{p}(1-\overline{p})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} = -1.56.$
- With $\alpha = 0.1 \Rightarrow$ the critical value is $\pm z_\alpha = \pm 2.33$. Hence, we cannot reject H_0.

Problem 7.16.

In a study to estimate the proportion of residents in a certain city and its suburbs who favor the construction of a nuclear power plant, it is found that 63 of 100 urban residents favor the construction while only 59 of 125 suburban residents are in favor. Is there a significant difference between the proportions of urban and suburban residents who favor the construction of the nuclear plant? Use a 0.01 level of significance.

Solution

- Let p_1 and p_2 represent the population proportions of male and female college students respectively. Null and alternative hypothesis $\begin{cases} H_0 : p_1 = p_2, \\ H_1 : p_1 \neq p_2 \end{cases}$.
- We compute $\overline{p} = \dfrac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1+n_2} = 0.542 \Rightarrow 1 - \overline{p} = 0.458.$
- Since $n_1p_1 \geq 5;\ n_1(1-p_1) \geq 5;\ n_2p_2 \geq 5;\ n_2(1-p_2) \geq 5$ are all at least 5, we can use

two-sample z-test: $z = \dfrac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\overline{p}(1-\overline{p})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} = 2.364.$

- With $\alpha = 0.05 \Rightarrow$ the critical value is $z_\alpha = 1.64$. Hence, we can reject H_0. Therefore, there is a significant difference between the proportions of urban and suburban residents who favor the construction of the nuclear plant