Mathematics of Data Management

sub-roc

one-tenth study guide

MDM4U

License

Attribution 4.0 International

==========

Creative Commons Corporation ("Creative Commons") is not a law firm and does not provide legal services or legal advice. Distribution of Creative Commons public licenses does not create a lawyer-client or other relationship. Creative Commons makes its licenses and related information available on an "as-is" basis. Creative Commons gives no warranties regarding its licenses, any material licensed under their terms and conditions, or any related information. Creative Commons disclaims all liability for damages resulting from their use to the fullest extent possible.

Using Creative Commons Public Licenses

Creative Commons public licenses provide a standard set of terms and conditions that creators and other rights holders may use to share original works of authorship and other material subject to copyright and certain other rights specified in the public license below. The following considerations are for informational purposes only, are not exhaustive, and do not form part of our licenses.

Considerations for licensors: Our public licenses are intended for use by those authorized to give the public permission to use material in ways otherwise restricted by copyright and certain other rights. Our licenses are irrevocable. Licensors should read and understand the terms and conditions of the license they choose before applying it. Licensors should also secure all rights necessary before applying our licenses so that the public can reuse the material as expected. Licensors should clearly mark any material not subject to the license. This includes other CC- licensed material, or material used under an exception or limitation to copyright. More considerations for licensors: wiki.creativecommons.org/Considerations$_{f}or_{l}icensors$

Considerations for the public: By using one of our public licenses, a licensor grants the public permission to use the licensed material under specified terms and conditions. If the licensor's permission is not necessary for any reason–for example, because of any applicable exception or limitation to copyright–then that use is not regulated by the license. Our licenses grant only permissions under copyright and certain other rights that a licensor has authority to grant. Use of the licensed material may still be restricted for other reasons, including because others have copyright or other rights in the material. A licensor may make special requests, such as asking that all changes be marked or described. Although not required by our licenses, you are encouraged to respect those requests where reasonable. More$_{c}onsiderationsforthepublic$: $wiki.creativecommons.org/Considerations_{f}or_{l}icensees$

==========

Creative Commons Attribution 4.0 International Public License

By exercising the Licensed Rights (defined below), You accept and agree to be bound by the terms and conditions of this Creative Commons Attribution 4.0 International Public License ("Public License"). To the extent this Public License may be interpreted as a contract, You are granted the Licensed Rights in consideration of Your acceptance of these terms and conditions, and the Licensor grants You such rights in consideration of benefits the Licensor receives from making the Licensed Material available under these terms and conditions.

Section 1 – Definitions.

a. Adapted Material means material subject to Copyright and Similar Rights that is derived from or based upon the Licensed Material and in which the Licensed Material is translated, altered, arranged, transformed, or otherwise modified in a manner requiring permission under the Copyright and Similar Rights held by the Licensor. For purposes of this Public License, where the Licensed Material is a musical work, performance, or sound recording, Adapted Material is always produced where the Licensed Material is synched in timed relation with a moving image.

b. Adapter's License means the license You apply to Your Copyright and Similar Rights in Your contributions to Adapted Material in accordance with the terms and conditions of this Public License.

c. Copyright and Similar Rights means copyright and/or similar rights closely related to copyright including, without limitation, performance, broadcast, sound recording, and Sui Generis Database Rights, without regard to how the rights are labeled or categorized. For purposes of this Public License, the rights specified in Section 2(b)(1)-(2) are not Copyright and Similar Rights.

d. Effective Technological Measures means those measures that, in the absence of proper authority, may not be circumvented under laws fulfilling obligations under Article 11 of the WIPO Copyright Treaty adopted on December 20, 1996, and/or similar international agreements.

e. Exceptions and Limitations means fair use, fair dealing, and/or any other exception or limitation to Copyright and Similar Rights that applies to Your use of the Licensed Material.

f. Licensed Material means the artistic or literary work, database, or other material to which the Licensor applied this Public License.

g. Licensed Rights means the rights granted to You subject to the terms and conditions of this Public License, which are limited to all Copyright and Similar Rights that apply to Your use of the Licensed Material and that the Licensor has authority to license.

h. Licensor means the individual(s) or entity(ies) granting rights under this Public License.

i. Share means to provide material to the public by any means or process that requires permission under the Licensed Rights, such as reproduction, public display, public performance, distribution, dissemination, communication, or importation, and to make material available to the public including in ways that members of the public may access the material from a place and at a time individually chosen by them.

j. Sui Generis Database Rights means rights other than copyright resulting from Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases, as amended and/or succeeded, as well as other essentially equivalent rights anywhere in the world.

k. You means the individual or entity exercising the Licensed Rights under this Public License. Your has a corresponding meaning.

Section 2 – Scope.

a. License grant.

1. Subject to the terms and conditions of this Public License, the Licensor hereby grants You a worldwide, royalty-free, non-sublicensable, non-exclusive, irrevocable license to exercise the Licensed Rights in the Licensed Material to:

a. reproduce and Share the Licensed Material, in whole or in part; and

b. produce, reproduce, and Share Adapted Material.

2. Exceptions and Limitations. For the avoidance of doubt, where Exceptions and Limitations apply to Your use, this Public License does not apply, and You do not need to comply with its terms and conditions.

3. Term. The term of this Public License is specified in Section 6(a).

4. Media and formats; technical modifications allowed. The Licensor authorizes You to exercise the Licensed Rights in all media and formats whether now known or hereafter created, and to make technical modifications necessary to do so. The Licensor waives and/or agrees not to assert any right or authority to forbid You from making technical modifications necessary to exercise the Licensed Rights, including technical modifications necessary to circumvent Effective Technological Measures. For purposes of this Public License, simply making modifications authorized by this Section 2(a) (4) never produces Adapted Material.

5. Downstream recipients.

a. Offer from the Licensor – Licensed Material. Every recipient of the Licensed Material automatically receives an offer from the Licensor to exercise the Licensed Rights under the terms and conditions of this Public License.

b. No downstream restrictions. You may not offer or impose any additional or different terms or conditions on, or apply any Effective Technological Measures to, the Licensed Material if doing so restricts exercise of the Licensed Rights by any recipient of the Licensed Material.

6. No endorsement. Nothing in this Public License constitutes or may be construed as permission to assert or imply that You are, or that Your use of the Licensed Material is, connected with, or sponsored, endorsed, or granted official status by, the Licensor or others designated to receive attribution as provided in Section 3(a)(1)(A)(i).

b. Other rights.

1. Moral rights, such as the right of integrity, are not licensed under this Public License, nor are publicity, privacy, and/or other similar personality rights; however, to the extent possible, the Licensor waives and/or agrees not to assert any such rights held by the Licensor to the limited extent necessary to allow You to exercise the Licensed Rights, but not otherwise.

2. Patent and trademark rights are not licensed under this Public License.

3. To the extent possible, the Licensor waives any right to collect royalties from You for the exercise of the Licensed Rights, whether directly or through a collecting society under any voluntary or waivable statutory or compulsory licensing scheme. In all other cases the Licensor expressly reserves any right to collect such royalties.

Section 3 – License Conditions.

Your exercise of the Licensed Rights is expressly made subject to the following conditions.

a. Attribution.

1. If You Share the Licensed Material (including in modified form), You must:

a. retain the following if it is supplied by the Licensor with the Licensed Material:

i. identification of the creator(s) of the Licensed Material and any others designated to receive attribution, in any reasonable manner requested by the Licensor (including by pseudonym if designated);

ii. a copyright notice;

iii. a notice that refers to this Public License;

iv. a notice that refers to the disclaimer of warranties;

v. a URI or hyperlink to the Licensed Material to the extent reasonably practicable;

b. indicate if You modified the Licensed Material and retain an indication of any previous modifications; and

c. indicate the Licensed Material is licensed under this Public License, and include the text of, or the URI or hyperlink to, this Public License.

2. You may satisfy the conditions in Section 3(a)(1) in any reasonable manner based on the medium, means, and context in which You Share the Licensed Material. For example, it may be reasonable to satisfy the conditions by providing a URI or hyperlink to a resource that includes the required information.

3. If requested by the Licensor, You must remove any of the information required by Section 3(a)(1)(A) to the extent reasonably practicable.

4. If You Share Adapted Material You produce, the Adapter's License You apply must not prevent recipients of the Adapted Material from complying with this Public License.

Section 4 – Sui Generis Database Rights.

Where the Licensed Rights include Sui Generis Database Rights that apply to Your use of the Licensed Material:

a. for the avoidance of doubt, Section 2(a)(1) grants You the right to extract, reuse, reproduce, and Share all or a substantial portion of the contents of the database;

b. if You include all or a substantial portion of the database contents in a database in which You have Sui Generis Database Rights, then the database in which You have Sui Generis Database Rights (but not its individual contents) is Adapted Material; and

c. You must comply with the conditions in Section 3(a) if You Share all or a substantial portion of the contents of the database.

For the avoidance of doubt, this Section 4 supplements and does not replace Your obligations under this Public License where the Licensed Rights include other Copyright and Similar Rights.

Section 5 – Disclaimer of Warranties and Limitation of Liability.

a. UNLESS OTHERWISE SEPARATELY UNDERTAKEN BY THE LICENSOR, TO THE EXTENT POSSIBLE, THE LICENSOR OFFERS THE LICENSED MATERIAL AS-IS AND AS-AVAILABLE, AND MAKES NO REPRESENTATIONS OR WARRANTIES OF ANY KIND CONCERNING THE LICENSED MATERIAL, WHETHER EXPRESS, IMPLIED, STATUTORY, OR OTHER. THIS INCLUDES, WITHOUT LIMITATION, WARRANTIES OF TITLE, MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, NON-INFRINGEMENT, ABSENCE OF LATENT OR OTHER DEFECTS, ACCURACY, OR THE PRESENCE OR AB-

SENCE OF ERRORS, WHETHER OR NOT KNOWN OR DISCOVERABLE. WHERE DISCLAIMERS OF WARRANTIES ARE NOT ALLOWED IN FULL OR IN PART, THIS DISCLAIMER MAY NOT APPLY TO YOU.

b. TO THE EXTENT POSSIBLE, IN NO EVENT WILL THE LICENSOR BE LIABLE TO YOU ON ANY LEGAL THEORY (INCLUDING, WITHOUT LIMITATION, NEGLIGENCE) OR OTHERWISE FOR ANY DIRECT, SPECIAL, INDIRECT, INCIDENTAL, CONSEQUENTIAL, PUNITIVE, EXEMPLARY, OR OTHER LOSSES, COSTS, EXPENSES, OR DAMAGES ARISING OUT OF THIS PUBLIC LICENSE OR USE OF THE LICENSED MATERIAL, EVEN IF THE LICENSOR HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH LOSSES, COSTS, EXPENSES, OR DAMAGES. WHERE A LIMITATION OF LIABILITY IS NOT ALLOWED IN FULL OR IN PART, THIS LIMITATION MAY NOT APPLY TO YOU.

c. The disclaimer of warranties and limitation of liability provided above shall be interpreted in a manner that, to the extent possible, most closely approximates an absolute disclaimer and waiver of all liability.

Section 6 – Term and Termination.

a. This Public License applies for the term of the Copyright and Similar Rights licensed here. However, if You fail to comply with this Public License, then Your rights under this Public License terminate automatically.

b. Where Your right to use the Licensed Material has terminated under Section 6(a), it reinstates:

1. automatically as of the date the violation is cured, provided it is cured within 30 days of Your discovery of the violation; or

2. upon express reinstatement by the Licensor.

For the avoidance of doubt, this Section 6(b) does not affect any right the Licensor may have to seek remedies for Your violations of this Public License.

c. For the avoidance of doubt, the Licensor may also offer the Licensed Material under separate terms or conditions or stop distributing the Licensed Material at any time; however, doing so will not terminate this Public License.

d. Sections 1, 5, 6, 7, and 8 survive termination of this Public License.

Section 7 – Other Terms and Conditions.

a. The Licensor shall not be bound by any additional or different terms or conditions communicated by You unless expressly agreed.

b. Any arrangements, understandings, or agreements regarding the Licensed Material not stated herein are separate from and independent of the terms and conditions of this Public License.

Section 8 – Interpretation.

a. For the avoidance of doubt, this Public License does not, and shall not be interpreted to, reduce, limit, restrict, or impose conditions on any use of the Licensed Material that could lawfully be made without permission under this Public License.

b. To the extent possible, if any provision of this Public License is deemed unenforceable, it shall be automatically reformed to the minimum extent necessary to make it enforceable. If the provision

cannot be reformed, it shall be severed from this Public License without affecting the enforceability of the remaining terms and conditions.

c. No term or condition of this Public License will be waived and no failure to comply consented to unless expressly agreed to by the Licensor.

d. Nothing in this Public License constitutes or may be interpreted as a limitation upon, or waiver of, any privileges and immunities that apply to the Licensor or You, including from the legal processes of any jurisdiction or authority.

==========

Creative Commons is not a party to its public licenses. Notwithstanding, Creative Commons may elect to apply one of its public licenses to material it publishes and in those instances will be considered the "Licensor." The text of the Creative Commons public licenses is dedicated to the public domain under the CC0 Public Domain Dedication. Except for the limited purpose of indicating that material is shared under a Creative Commons public license or as otherwise permitted by the Creative Commons policies published at creativecommons.org/policies, Creative Commons does not authorize the use of the trademark "Creative Commons" or any other trademark or logo of Creative Commons without its prior written consent including, without limitation, in connection with any unauthorized modifications to any of its public licenses or any other arrangements, understandings, or agreements concerning use of licensed material. For the avoidance of doubt, this paragraph does not form part of the public licenses.

Creative Commons may be contacted at creativecommons.org.

Contents

Part I

Proba-bility

1 Introduction to Probability

1.1 Simple Probabilities

Probability: A quantified measure of the likelihood that an event will occur.

Subjective probability: An estimate of an event's probability, mostly based on intuition.

Sample space (S): A set of all possible events/outcomes.

A: A particular event.

$n(S)$: The number of possible events/outcomes.

$n(A)$: The total number of instances in which an event A can occur.

Probability $P(A)$ of an event occuring:

$$P(A) = \frac{n(A)}{n(S)}$$

Special cases:

$P(A) = 1$: 100% chance that event A occurs.

$P(A) = 0$: 0% chance that event A occurs.

$P(A') = 1 - P(A)$: Probability of A' occuring.

Example: There are three white balls and 5 red balls in the plastic bag. What is the probability of choosing a white ball?

A: A white ball being drawn
$n(A) = 3$
$n(S) = 8$

$$P(A) = \frac{3}{8} = 37.5\%$$

1.2 Experimental vs. Theoretical Probability

Experimental probability: The measure of an event's likelihood based on experimental outcomes.

Theoretical Probability: Determined by knowledge of an event's nature.

Example: A coin is flipped. What is the theoretical probability of flipping tails?

A: Flipping tails
$n(A) = 1$
$n(S) = 2$

$$P(A) = \frac{1}{2} = 50\%$$

Example: A coin is tossed 10 times and a tails turns up 6 times. What is the experimental probability of flipping tails?

A: Flipping tails
$n(A) = 6$
$n(S) = 10$

$$P(A) = \frac{6}{10} = 60\%$$

1.3 Mutually Exclusive Events

Mutually exclusive events: Events that cannot occur at the same time. For example:

- Drawing a card: You cannot draw a heart and a diamond at the same time.

Addition Rule for Mutually Exclusive Events:

$$P(A \cup B) = P(A) + P(B)$$

This equation is used when the word "OR" is seen, and the events **cannot** happen at the same time.

Example: What is the probability that when you roll a die, you get either a 5 or an even number?

$$P(A \cup B) = \frac{1}{6} + \frac{3}{6} = \frac{4}{6} = \frac{2}{3}$$

Here, rolling a 5 and rolling an even number are mutually exclusive.

Non-Mutually Exclusive Events: Events that can occur at the same time. For example:

- Drawing a card: It can be a diamond **and** a face card (e.g., Jack, Queen, King of diamonds).

Addition Rule for Non-Mutually Exclusive Events:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

This equation is used when the word "AND" is seen, and the events from A and B **can** happen at the same time.

Example: In a group of teachers, what is the probability of randomly selecting either an Italian or a male teacher?

Non-mutually exclusive: A teacher can be both male and Italian.

Additional Examples

Example: A card is randomly selected from a standard deck of cards. What is the probability that either a heart or a face card is selected?

$$P(A \cup B) = P(\text{Heart}) + P(\text{Face}) - P(\text{Heart} \cap \text{Face})$$

$$P(A \cup B) = \frac{13}{52} + \frac{12}{52} - \frac{3}{52} = \frac{22}{52} = \frac{11}{26}$$

Example: Teri attends a fundraiser where 15 T-shirts (2 black, 4 blue, 9 white) are being given as door prizes. Assuming Teri wins the first door prize, what is the probability she gets a shirt she likes (black or blue)?

Mutually exclusive: She either likes the color or doesn't.

$$P = \frac{2}{15} + \frac{4}{15} = \frac{6}{15} = \frac{2}{5}$$

Example: An electronics manufacturer tests a product to determine if a voltage spike damages it. The probabilities are:

- Probability of damaging the power supply: 0.2%,
- Probability of damaging downstream components: 0.6%,
- Probability of damaging both: 0.1%.

Not mutually exclusive: Both events can happen.

1.4 Additive Counting Principle

Venn diagram: A type of diagram that helps you organize groups of data.

Principle of Inclusion and Exclusion for 2 sets:
If you are counting the total elements in 2 groups or sets with common elements, you must subtract the common elements.

$$n(A \cup B) = n(A) + n(B) - n(A \cap B)$$

Example: There are 10 students on the volleyball team and 15 on the basketball team. When planning a trip with both teams, the coach has to arrange transportation for a total of only 19 students. Determine how many students are on both teams.

$$25 - 19 = 6 \text{ students on both teams}$$

Additive Counting Principle:
If one action can occur in a_1 ways, a second in a_2 ways, etc, and these actions are **mutually exclusive**, there are

$$\sum_{i=1}^{n} a_i$$

ways in which one of these actions can occur.

Example: How many ways can you draw a 7 or a Jack from a deck of cards?

$n(7) = 4$

$n(\text{Jack}) = 4$

$$\begin{aligned} & n(7 \cup \text{Jack}) \\ &= n(7) + n(\text{Jack}) \\ &= 4 + 4 \\ &= 8 \end{aligned}$$

1.5 Independent and Dependent Events

Independent Events:
If events A and B are independent of each other, the probability of both events occuring equals:

$$P(A \cup B) = P(A) \cdot P(B)$$

Example: What is the probability of rolling a 3 with a die and drawing a 3 from a deck of cards?

$P(3 \text{ on a die}) = \frac{1}{6}$

$P(3 \text{ from a deck}) = \frac{4}{52}$

$$P(3 \text{ on a die} \cup 3 \text{ from a deck}) = \frac{1}{6} \cdot \frac{4}{52} = \frac{1}{78}$$

Conditional probability: The probability of one event occurring given that another has.

Dependent Events:
If events A and B are dependent events, the outcome of one event will affect that of the second (e.g. drawing cards without replacement). The probability of both events occurring is:

$$P(A \cup B) = P(A) \cdot P(B|A)$$

Example: There are 3 white balls and 5 red balls in a plastic bag. What is the probability of choosing two red balls, one after the other?

$P(\text{first red}) = \frac{5}{8}$

$P(\text{second red}|\text{first red}) = \frac{4}{7}$

$$\begin{aligned} P(\text{first red} \cup \text{second red}|\text{first red}) &= P(\text{first red}) \cdot P(\text{second red}|\text{first red}) \\ &= \frac{5}{8} \cdot \frac{4}{7} \\ &= \frac{20}{56} = \frac{5}{14} \end{aligned}$$

2 Permutations

2.2 Fundamental Counting Principle

Multiplicative Counting Principle:
If you can **choose** (non-mutually exclusive) between m number of items of one type, n of another type, etc, the total number of choices is:

$$m \cdot n \cdot p...$$

Example: Find the number of possible 7-digit phone numbers using any number from 0-9.

$$= 10^7 = 10{,}000{,}000 \text{ numbers}$$

Example: How many car license plates can be created when the first 4 characters are letters and the last 3 are numbers?

$$= 26^4 \cdot 10^3 = 456{,}976{,}000 \text{ numbers}$$

Example: In how many ways can a president, vice-president, and secretary be selected from a group of 8 people? Assume nobody can hold more than one position.

$$= 8 \cdot 7 \cdot 6 = 336 \text{ ways}$$

2.3 Permutations and Factorials

Factorial:
The product of all integers from n to 1 where $\{n \in \mathrm{R} | n > 0\}$

$$n! = n \cdot (n-1) \cdot ... \cdot 1$$

Example: Expand 6!

$$= 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$$

$$= 720$$

Example: Simplify $\frac{9!}{4!}$

$$= \frac{9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4!}{4!}$$

Cancel 4! in the numerator and denominator.

$$= 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5$$

$$= 15{,}120$$

Permutation:
The total number of possible arrangements of r objects out of a set of n when **order matters** and there is no replacement.

$$_nP_r = \frac{n!}{(n-r)!}$$

Example: You have been given 8 pictures to hang, but can only hang 3 at a time. How many arrangements are possible?

$n = 8$

$r = 3$

$$_8P_3 = \frac{8!}{(8-3)!} = 336 \text{ arrangements}$$

Example: What if you had 10 pictures and could hang 4 at a time?

$n = 10$

$r = 4$

$$_{10}P_3 = \frac{10!}{(10-4)!} = 5040 \text{ arrangements}$$

2.4 The Rule of Sum

Recall that if one mutually exclusive event can occur in m ways and another in n ways, one or the other can occur in $m + n$ ways.

Example: At an international conference, either eight or nine countries may attend. In how many different arrangements could the countries' flags be flown?

8 countries and 9 countries attending are mutually exclusive.

$$= {}_8P_8 + {}_9P_9$$

$$= 8! + 9!$$

$$= 403{,}200 \text{ arrangements}$$

Example: Three players are playing the card game Pass the Ace. Each player receives one card. In how many ways could all the cards be face cards **or** red cards?

$n(\text{Face}) = 12$

$n(\text{Red}) = 26$

$n(\text{Face} \cup \text{Red}) = 6$

$$n(\text{Face} \cap \text{Red}) = {}_{12}P_3 + {}_{26}P_3 - {}_6P_3$$

$$= 1{,}320 - 15{,}600 - 120$$

$$= 16{,}800 \text{ ways}$$

2.5 Applications of Permutations

Example: Software for generating multiple choice tests randomly assign A, B, C, or D as the correct answer, On a 10-question test, what is the probability that all 10 questions have C as the correct answer?

$$P(\text{all C}) = (\frac{1}{4})^{10}$$

$$= \frac{1}{1{,}048{,}576}$$

Example: Eight people on a waiting list for advance tickets to a concert have been selected to choose their seats. What is the probability that they will have been notified in order from youngest to oldest.

$n(A) = 1$

$n(S) = 8! = 40{,}320$

$$P(A) = \frac{n(A)}{n(S)}$$

$$= \frac{1}{40{,}320}$$

3 Combinations

3.1 Permutations with Some Identical Elements

Permutation review:
Permutations of all elements:

$$P(n, n) = n!$$

Permutations involving some elements:

$$P(n, r) = \frac{n!}{(n-r)!}$$

When n elements contain duplicates (m, p, etc.), the formula is:

$$\frac{n!}{m!p!\ldots}$$

Example: How many permutations are there of the letters A, B_1, B_2, C?

$$4 \cdot 3 \cdot 2 \cdot 1 = 24$$

Example: How many permutations are there of the letters A, B, B, C?

$$\frac{4!}{2!} = 12$$

Additional Examples

Example: In how many ways can you travel from your house to a store 5 blocks north and 7 blocks east, if you must always travel north or east?

$$\frac{12!}{5! \cdot 7!} = 792$$

Example: How many arrangements can you make of the digits 1, 2, 3, 4, 5 if the odd digits must always be in ascending order?

$$\frac{5!}{3!} = 20$$

3.2 Combinations

Combination: A selection from a group of items **without regard to order**.

Combination formula:
The number of combinations from r objects chosen from a set of n distinct objects is

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

Example: How many different sampler dishes with 3 different flavours could you get at an ice cream shop with 31 different flavours?

$n = 31$

$r = 3$

$$\binom{31}{3} = 4{,}495 \text{ dishes}$$

3.3 Problem Solving with Combinations

Null set: A set with no elements

The total number of combinations containing at least one item from a group of n distinct items equals

$$2^n - 1$$

(Subtract 1 to omit the null set)

Example: How many ways can a committee of any size be chosen from 8 people?

$n = 8$ people

$$= 2^8 - 1$$

$$= 255 \text{ ways}$$

Example: How many ways can you put at least one topping on a cheese pizza if you can choose from pepperoni, bacon, and mushrooms?

$n = 3$ toppings

$$= 2^3 - 1$$

$$= 7 \text{ ways}$$

3.4 Pascal's Triangle

Pascal's Triangle: A triangular array of numbers in which each term is the sum of the two terms above it.

Pascal's Method:
The terms of Pascal's triangle are generated by adding two adjacent terms and placing the result immediately below them in the next row.

$$\binom{n}{r} + \binom{n}{r+1} = \binom{n+1}{r+1}$$

Example: To get to work from her house, Hannah travels four blocks south and five east. How many different routes can she take?

Hannah needs to travel 9 blocks. Select any 4 to be southbound, the rest will be eastbound.

$n = 9$

$r = \{4, 5\}$

$$= \binom{9}{4} \text{ or } \binom{9}{5}$$
$$= 126 \text{ routes}$$

Binomial Expansion:
When expanding a binomial raised to a power n, the term coefficients are equal to the terms in row n of Pascal's triangle.

Example: Expand $(x + y)^3$.

Row 3: $1, 3, 3, 1$

$$(x + y)^3 = 1x^3 + 3x^2y + 3xy^2 + 1y^3$$

Example: Expand $(x - y)^4$.

Row 3: $1, 4, 6, 4, 1$

$$(x - y)^4 = 1x^4 - 4x^3y + 6x^2y^2 - 4xy^3 + 1y^4$$

3.5 Applications of Combinations

Example: A scratch-and-win contest allows you to choose 5 numbers. If all your numbers match the winning set of 5 numbers, chosen from 1 to 25 without regard to order, you win the grand prize.

a) What is the probability of winning the grand prize?

$$P = \frac{\binom{5}{5}}{\binom{25}{5}} = \frac{1}{53,130}$$

b) What is the probability of winning the second prize (4 correct numbers)?

$$P = \frac{\binom{5}{4} \cdot \binom{20}{1}}{\binom{25}{5}} = \frac{100}{53,130} = \frac{10}{5,313}$$

Example: A university task force of 8 people is to be formed from 16 members of the student government and 10 professors. Each person is equally likely to be chosen.

a) What is the probability that there is an equal number of students and professors?

$$P = \frac{\binom{16}{4} \cdot \binom{10}{4}}{\binom{26}{8}} = \frac{382,200}{1,362,275} \approx 0.24$$

b) What is the probability that at least 6 members are students?

$$P = \frac{\binom{16}{6} \cdot \binom{10}{2} + \binom{16}{7} \cdot \binom{10}{1} + \binom{16}{8} \cdot \binom{10}{0}}{\binom{26}{8}} \approx 0.31$$

c) Which outcome is more likely to occur?

At least 6 students is more likely ($P \approx 0.31$).

4 Probability of Discrete Variables

4.1 Probability Distribution and Uniform Distributions

Discrete Variables: Only have certain values in a given range (e.g., whole numbers such as the number of people in a class).

Continuous Variables: Have infinite possible values (e.g., time, weight).

Probability Distribution:
The probabilities for all possible outcomes of an experiment or sample space. Often shown as a graph of probability (y-axis) versus the value of a random variable (x-axis).

Random Variable:
A quantity that can have a range of values. Designated by a capital letter X with individual values designated by a lowercase x.

Types of Probability Distributions:

- Uniform Distribution
- Binomial Distribution
- Hypergeometric Distribution

Example: Classify the following variables as discrete or continuous.

a) Length of time you stay in class:

Continuous (55.35 min)

b) Number of courses you take in a semester:

Discrete (you cannot take half a course)

Expected Value:
The expected value or expectation is the average of the outcomes:

$$E(x) = x_1P(x_1) + x_2P(x_2) + \cdots + x_nP(x_n) = \sum_{i=1}^{n} x_iP(x_i)$$

Example: Given the following probability distribution, determine the expected value.

x	$P(x)$
2	0.4
4	0.1
6	0.5

$$E(x) = (2)(0.4) + (4)(0.1) + (6)(0.5) = 4.2$$

Example 4: A hospital is having a fundraising lottery to raise money for cancer research. A ticket costs \$10, and 2,000,000 tickets are available. Prizes include:

- 1 grand prize of $5,000,000
- 3 second prizes of $100,000
- 10 third prizes of $1,000
- 2,000 free tickets for next year's lottery ($10)

a) What is the expected value of each ticket?

$$E(x) = \sum xP(x) = \frac{5,000,000}{2,000,000} + \frac{100,000 \cdot 3}{2,000,000} + \frac{1,000 \cdot 10}{2,000,000} + \frac{10 \cdot 2,000}{2,000,000} - 10$$

$$E(x) = -7.335$$

b) Explain its meaning.

On average, the lottery player loses $7.34 per ticket.

4.2 Binomial Distribution

Occur when an experiment is repeated and a particular outcome (success or failure) is counted. Experiments that are repeated are called Bernoulli trials.

Conditions for a Bernoulli trial:

- Only two outcomes are possible (success and failure).
- The outcome of each trial does not depend on the previous trial (independent trials).
- The probability of success and failure does not change for each trial.
- Trials are repeated a specified number of times.

Application: Binomial distributions can be used to count the number of defects during quality control.

Formula for Binomial Distributions:

$$P(x) = \frac{n!}{(n-x)!x!}p^x q^{n-x} \quad \text{or} \quad P(x) = \binom{n}{x} p^x q^{n-x}$$

- x: Number of single successes
- n: Number of Bernoulli Trials
- p: Probability of a successful outcome
- q: Probability of a single failure ($q = 1 - p$)

The formula for the expected value of a binomial distribution is:

$$E(x) = np$$

Example: When flipping a coin 6 times, what is the probability of getting exactly 4 heads?

- Coin tosses are independent.

- Probability of success (heads): $p = 0.5$.
- Probability of failure (tails): $q = 0.5$.

The number of possible outcomes of flipping a coin 6 times is $2^6 = 64$. Of these, 15 outcomes give the event of flipping 4 heads. Thus, the probability is:

$$P(x = 4) = \frac{15}{64}$$

Using the binomial probability formula:

$$P(x = 4) = \binom{6}{4} p^4 q^2$$

$$P(x = 4) = \binom{6}{4} (0.5)^4 (0.5)^2 = \frac{15}{64} \approx 0.2344$$

4.4 Hypergeometric Distribution

The hypergeometric distribution is a discrete probability distribution, involving a series of dependent trials, with more than one type of success or failure. It involves a number of random draws from a limited population, without replacement.

Formula for Hypergeometric Distribution:

$$P(x) = \frac{\binom{a}{x}\binom{n-a}{r-x}}{\binom{n}{r}}$$

- x: Successful outcomes out of r trials.
- a: Successful outcomes in population n.

Example: There are 5 bananas and 7 oranges in the refrigerator. Four fruits are chosen at random to serve guests. What is the probability that exactly two of the fruits will be oranges?

$$P(x = 2) = \frac{\binom{7}{2}\binom{5}{2}}{\binom{12}{4}} = \frac{21 \cdot 10}{495} \approx 0.424$$

Example: What is the probability of a Formula 1 race finishing with 2 Ferrari, 2 Renault, and 1 Honda in the top 5 if each team has 5 cars in the race and the race consists of only those teams?

$$P = \frac{\binom{5}{2}\binom{5}{2}\binom{5}{1}}{\binom{15}{5}} = 0.167$$

Expected Value:
The expected value of a hypergeometric distribution is given by:

$$E(x) = \frac{ra}{n}$$

Example: A jar of jellybeans contains 20 yellow jellybeans and 25 red jellybeans. If 5 jellybeans were drawn from the jar randomly, what is the expected number of red jellybeans drawn?

$$E(x) = \frac{ra}{n} = \frac{(5)(25)}{45} \approx 2.78$$

Part II

Stat-istics

5 Organization of Data for Analysis

5.1 Data Concepts and Graphical Summaries

- Statistics is the gathering, organization, and analysis of data
- You can apply statistical methods for all kinds of analysis

Types of Data:
Numerical Data: Data in the form of numbers

- Discrete: Whole numbers
- Continuous: Data with any value in a certain range incl. decimals

Categorical data: Non-numerical data sorted into groups

- Ordinal: Data that can be ranked e.g. ratings
- Nominal: Data that can not be ranked e.g. colour

There are multiple ways of displaying data, for example:

- Frequency table
- Bar/column chart
- Histogram
- Circle graph
- Scatter plots

When using a frequency table, you must use interval notation.

- A square bracket means that the range includes that bound
- A round bracket means that the range does not include that bound

Example: Convert the range [40,45) to set notation.

$$= \{40 \leq x < 45\}$$

5.2 Sampling Techniques

- Sampling techniques are methods used to select a sample from a population
- A sampling frame is the members of the population that have a chance of being studied
- Statistical bias is systematic errors in a survey/sampling method

Sampling Techniques

- Simple random sample: equal probability selection
- Systematic sample: selecting every n^{th} item
- Stratified sample: Sampling proportional to group sizes in the population
- Cluster sample: Divides population into equally sized groups and randomly selects groups

- Multi-stage sampling: Splitting groups into a hierarchy and choosing a group at every level

Example: A university is polling students. It selects 200 students randomly in the same proportions as enrollment in each departments. What type of sampling method is this?

Stratified sampling as sample group sizes are proportional to those in the population.

5.3 Collecting Data

- Observational studies are when researchers observe already-occuring situations
- Experimental studies are when researchers control what is occurring and make inferences
  - The treatment group is the group that receives the treatment
  - The control group is the group that doesnt
- There are three things to determine cause:
  - Control: As many aspects of the experiment need to be controlled
  - Randomization: Sampling must be random to avoid bias
  - Replication: The results should stay the same in a repeated trial
- Bias occurs when there is prejudice for or against the idea
- This can come from:
  - The sampling technique: Over/underrepresentation
  - Data collection: The survey or method of collection is flawed

Designing Good Surveys
Surveys should follow these guidelines:

- Anonymity
- Clear and even rating scales
- "Other" option in questions w/ limited options
- Lack of personal questions without a "Prefer not to answer" section
- Clear questions asking only one thing
- No charged/loaded questions
- Fast and efficient data collection

5.4 Interpreting and Analyzing Data

- Primary data has been collected by the researcher
  - An individual data point is called microdata
- Secondary data is data used by someone that did not collect it

  - This data is usually pre-organized into aggregate data
    - Aggregate data is usually summed, averaged, etc.
- Cross-sectional data is data that observes/studies variables at the same time

Example: The below table shows cross-sectional ata of average air fares in 10 Canadian cities. a)

Table 1: Average Domestic Air Fares for Canada and 10 Major Cities

City	**2010 ($)**	**2011 ($)**	**% Change 2010 to 2011**
Canada	182.5	190.7	4.5%
Calgary	165.5	176.2	6.5%
Edmonton	160.8	170.0	5.7%
Halifax	172.0	179.3	4.2%
Montreal	171.1	179.4	4.9%
Ottawa	196.0	194.8	-0.6%
Regina	168.1	177.8	5.8%
Saskatoon	170.2	178.8	5.1%
Toronto	205.2	214.9	4.7%
Vancouver	192.5	206.7	7.4%
Winnipeg	181.0	189.4	4.6%

Does this table show microdata or aggregate data?

This data is averaged $\therefore$ it is aggregate data.

b) Is this data primary or secondary?

This data is secondary data.

c) Identify the independent and dependent variables.

Independent: City name

Dependent: Average airfare

5.5 Bias

- Bias occurs when the sample or survey is not representative of the population

Sampling Method Bias:

- Sampling bias: The sample does not reflect the population
- Measurement bias: The data collection method over/underestimates a characteristic

Survey Bias:

- Leading question: A question prompts a particular answer (e.g. multiple-choice)
- Loaded question: The question uses language guiding the participant to a certain answer
- Response bias: Participants give a false/misleading answer for any reason

- Non-response bias: Participants do not complete a survey or survey question, causing misrepresentation

6 One-Variable Data Analysis

6.1 Measures of Central Tendency

Measures of central tendency: methods of determining which values data tends to cluster around.

Measures of central tendency:

Mean: the sum of the data values divided by the number of values in a set.
For ungrouped data:

$$\mu \text{ or } \overline{x} = \frac{\sum x}{N}$$

For grouped/weighted data:

$$\mu = \frac{\sum_i f_i m_i}{\sum_i f_i}$$

Median: The middle value of a data set ordered from lowest to highest value.

- If there is an even number of data points, then the median is the average of the middle two points

Mode: The value that occurs most frequently in a set of data.

Data often has outliers, which are values that are significantly distant from the majority of the data. These values skew the data, having a minimal effect on the median but a larger effect on the mean.

Example: Below are two sets of class marks.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Class A	71	82	55	76	66	71	90	84	95	64	71	70	83	45	73	51	68	
Class B	54	80	12	61	73	69	92	81	80	61	75	74	15	44	91	63	50	84

a) Determine the mean, median, and mode for each class.

Class A:

45, 51, 55, 64, 66, 68, 70, 71, 71, 71, 73, 76, 82, 83, 84, 90, 95

- 71 appears most often at three times, meaning that **the mode is 71**.
- The ninth (middle) point is 71, so **the median is 71**.
- $\frac{\sum x}{N} = \frac{1215}{17} = 71.5$ so **the mean is 71.5**.

Class B:

12, 15, 44, 50, 54, 61, 61, 63, 69, 73, 74, 75, 80, 80, 81, 84, 91, 92

- 61 and 80 appears most often at two times, meaning that **the modes are 61 and 80**.
- The ninth (left of middle) point is 69 and the tenth (right of middle) point is 73.
- The mean of 69 and 73 is 71, so **the median is 71**.
- $\frac{\sum x}{N} = \frac{1159}{18} = 64.4$ so **the mean is 64.4**.

b) Use the measures of central tendency to compare the performance of the two classes.

- The classes' medians are equal.
- The median is the best measure of central tendency because of how it is unaffected by outliers.

c) What is the effect of any outliers on the mean and median?

- Outliers do not affect the median.
- The outliers in Class B significantly affect the class mean as without the outliers, the mean would be 71 compared to 64.4.

6.2 Measures of Spread

Measures of spread: Indicate how closely a set of data cluster around its center.

Range: A calculation of the difference between the largest and smallest values in a data set.

Percentiles: The percent of all data less than or equal to a specific data value

- k percent of the data are less than or equal to k^{th} percentile, P_k

Percentile rank:

Percentile rank calculates the value in a data set that fall in a given percentile.

$$R = \frac{p}{100}(n+1)$$

p is the percentile and n is the population size.

Percentile:

Percentile calculates the percent of values less than or equal to a given value.

$$p = 100\frac{L + 0.5E}{n}$$

L is the number less than, E is the number equal to, and n is the population size.

Example: The list below shows the martks from least to greatest for 25 students on a recent test out of 40.

$$15, 16, 20, 21, 21, 23, 24, 25, 25, 25, 26, 28, 28, 28, 28, 30, 30, 31, 32, 34, 36, 36, 37, 38, 40$$

a) Calculate the 80^{th} percentile.

$$R = \frac{80}{100}(25+1)$$

$= 20.8 \therefore$ the 80^{th} percentile is between the 20^{th} and 21^{st} values (34, 36).

$$\frac{34+36}{2} = 35$$

$\therefore$ The 80^{th} percentile is $\frac{35}{40}$.

6.3 Standard Deviation & z-Scores

Standard deviation and variance show how values in a distribution are centered around the mean.

Standard deviation:

Standard deviation is the square root of the mean of the squares of the deviations.

- Gives greater weight to larger deviations

$$\text{Population: } \sigma = \sqrt{\frac{\sum x^2 - N \cdot \mu^2}{N}}$$

$$\text{Sample: } s = \sqrt{\frac{\sum x^2 - n \cdot \overline{x}^2}{n-1}}$$

Example: The ages of participants in a school's talent contest are listed below, along with the mean and standard deviation.

$$16, 17, 18, 16, 15, 16, 17, 15, 18, 14, 17, 19, 18, 16, 17, 17, 17, 14, 16, 18$$

$$\mu = 16.5 \; \sigma = 1.36$$

a) What would happen to the standard deviation if each person were one year older?

The mean would increase by one but the standard deviation would remain unchanged.

b) Which ages are more than one standard deviation from the mean?

$$16.5 \pm 1.36 = \{15.14, 17.86\} \therefore 15 \text{ or younger, } 18 \text{ or older.}$$

Z-scores:

A measure of how many standard deviations a particular data value is from the mean.

- Data with values below the mean have negative z-scores.
- Data with values above the mean have positive z-scores.
- Data with values equal to the mean have z-scores of zero.

$$\text{Population: } z = \frac{x - \mu}{\sigma}$$

$$\text{Sample: } z = \frac{x - \overline{x}}{s}$$

Example: A food manufacturer makes 2-litre jars of pasta sauce. Samples are tested for how close to 2 L the jars are filled. Fifteen samples were taken and their volume in liters were as indicated.

$$2.11, 2.02, 2.10, 1.99, 1.92, 2.01, 1.89, 1.96, 2.00, 1.96, 1.98, 2.02, 2.08, 2.15, 2.03$$

a) Determine the sample mean and standard deviation.

$$\overline{x} = \frac{30.22}{15} = 2.01467$$

$$s = \sqrt{\frac{\sum 60.955 - 15 \cdot 2.01467^2}{14}} = 0.0715$$

b) Calculate the z-score of the jar that was filled to a volume of 2.02 L.

$$z = \frac{2.02 - 2.01467}{0.0715} = 0.075 \text{ standard deviations}$$

c) Calculate the z-score of the jar that was filled to a volume of 1.98 L.

$$z = \frac{1.98 - 2.01467}{0.0715} = -0.485 \text{ standard deviations}$$

7 Probability Distributions for Continuous Variables

7.1 Continuous Random Variables

- Continuous probability distributions allow for non-whole values for random variables
- There are four main distributions for continuous random variables

Normal distribution: Symmetrical and unimodal centered around the mean, bell-shaped.
Positively skewed: Asymmetrical and unimodal, tail skewed right
Negatively skewed: Asymmetrical and unimodal, tail skewed left
Multimodal: Distribution with ≥ 2 modes, mean between these modes

Example: The driving time between Toronto and North Bay is found to range evenly between 195 and 240 minutes.
a) What kind of distribution is this?

Uniform, times range evenly $\therefore$ the probabilities are equal.

b) What is the probability that the drive will take less than 210 minutes?

$$= \frac{15}{45} = 33.\overline{3}\%$$

c) Is it possible to determine the probability of a trip taking exactly n minutes?

No, it is not possible to calculate exact times as the probability is approximately zero.

- Frequency tables and histograms can be used to visually understand the distribution of the data

7.2 The Normal Distribution and z-Scores

- A normal distribution can be described with its mean (μ) and standard deviation (σ)
- Usually the mean, median, and mode are all equal in a normal distribution
- The smaller the value of σ, the narrower the distribution

Empirical Rule:

- 68% of all data is within $\pm 1\sigma$ of the mean
- 95% of all data is within $\pm 2\sigma$ of the mean
- 99.7% of all data is within $\pm 3\sigma$ of the mean

Example: Giselle is 168 cm tall. In her high school, boys' heights are normally distributed with a mean of 174 cm and standard deviation of 6 cm. What is the probability that the first boy that she sees will be taller than her?

- We can use the empirical rule
- 168 cm is one standard deviation below 174 cm

$$= 50\% + \frac{68\%}{2} = 84\%$$

Recall:
The z-score is the number of standard deviations that a data point is from the mean.

$$\text{Population: } z = \frac{x - \mu}{\sigma}$$

$$\text{Sample: } z = \frac{x - \overline{x}}{s}$$

Example: All That Glitters, a sparkly cosmetic powder, is said to contain about 50 g per container of powder. On average, each container contains 50.5 g with a standard deviation of 0.6 g. The manufacturer wants at least 49.5 g in each container. What percentage of packages do not contain this much powder?

$$z = \frac{49.5 - 50.5}{0.6}$$

$$z = 1.\overline{6} \therefore P(x < 49.5) = 4.75\%$$

7.4 Confidence Intervals

- The margin of error is the range of values that a measurement is said to be within
- The larger the sample size, the smaller the margin of error

$$\text{Margin of error: } E = z\sqrt{\frac{p(1-p)}{n}}$$

The confidence interval is the range of possible values of the measured statistic at a particular confidence level

Example: In a study of fish species, a researcher determined that lake trout made up 20.4% of the fish population in Lake Lavielle. This estimate is considered correct with ±3.0% 19 times out of 20. What does this mean?

We are 95% confident that the trout population is within 3% of 20.4%.

Example: Lake Lavielle is one of the largest lakes in Algonquin Park. In 2009, 234 lake trout were caught out of 911. In 2012, 141 were caught out of 689.
a) Determine the percentage of lake trout caught per year.

$$25.69\%, 20.46\%$$

b) Determine the margin of error for 2009 using a 95% confidence level.

$$E = 1.960\sqrt{\frac{0.2569(1 - 0.2569)}{911}}$$

$$E = 0.0283 = 2.83\%$$

c) Determine the confidence interval.

$$= 25.69\% \pm 2.83\%$$

$$= \{22.86\% \leq x \leq 28.52\%\}$$

- It is never reasonable to assume a decrease or increase between years if the confidence intervals overlap

Example: A pharmaceutical company makes more than 500,000 pills of Adderall every day. THe company randomly samples 400 pills daily to ensure that they meet the proper weight and size standards. On a given day, 52 are substandard. How can the company cut the margin of error in half at a 90% confidence level by changing the sample size?

$$E = 0.028$$

$$0.014 = 1.645\sqrt{\frac{0.13 \cdot 0.87}{n}}$$

$$n = 1600$$

Repeated Sampling:

- Suppose that samples of the same size are repeatedly taken from a normally distributed population with mean μ and standard deviation σ
- The means of the samples will be normally distributed with a mean μ and standard deviation $\sigma_x = \frac{\sigma}{\sqrt{n}}$
- The margin of error or standard error for a sample mean is $E = z\sigma_x$

7.5 Discrete Approximations

Continuity Correction:

- Use when fitting a normal distribution to discrete data
- Add or subtract 0.5

$$P(X < 5) \longrightarrow P(X < 4.5)$$

$$P(X \leq 5) \longrightarrow P(X < 5.5)$$

$$P(X > 5) \longrightarrow P(X > 5.5)$$

$$P(X \geq 5) \longrightarrow P(X > 4.5)$$

$$P(X = 5) \longrightarrow P(4.5 < X < 5.5)$$

Example: A company produces candy THe number of pieces per box is normally distributed with a mean of 48 pieces and a standard deviation of 4.3 pieces. Any boxes with less than 44 pieces or more than 54 are rejected. a) What is probability that a box selected randomly has exactly 50 pieces?

$$P(X = 50) \longrightarrow P(49.5 < X < 50.5)$$

$$P(X < 49.5) = 0.636, P(X > 50.5) = 0.720$$

$$\therefore P(49.5 < X < 50.5) = 0.084$$

b) What percent of the production will be rejected for having too little pieces?

$$P(X < 44) \longrightarrow P(X < 43.5)$$

$$z = \frac{43.5 - 48}{4.3}$$
$$P(X < 43.5) = 0.1467$$

c) Each filling machine produces 130,000 boxes per shift. How many will fall within the acceptable range?

$$P(44 \leq X < 54) \longrightarrow P(43.5 < X < 54.5)$$
$$P(X < 43.5) = 14.69\%$$
$$P(X < 54.5) = 93.45\%$$
$$P(43.5 < X < 54.5) = 78.76\%$$
$$\therefore 78.76\%$$

Data with a discrete binomial distribution can be approximated with a normal distribution if:

- $np > 5$
- $nq > 5$

$$\mu = np, \sigma = \sqrt{npq}$$

Data with a discrete hypergeometric probability distribution can be approximated with a normal distribution if the sample is less than a tenth of the population

$$\mu = np, \sigma = \sqrt{npq(\frac{N-n}{N-1})}$$

Example: Allison has a drawer of unmatched socks. It contains 30 blue socks, 30 green socks, and 30 yellow socks. She pulls seven socks and records the number of blue socks. What is the probability that three to five of the socks are blue?

$$P(3 \leq X \leq 5) \longrightarrow P(2.5 < X < 5.5)$$
$$\mu = 2.\overline{3}, \sigma = 1.2$$
$$z = \{0.14, 1.20\}$$
$$\therefore P(3 \leq X \leq 5) = 43.9\%$$

8 Two-Variable Data Analysis

8.1 Line of Best Fit

- Linear regression is an analytic technique for determining relationships
- Interpolation is when you estimate between known data points
- Extrapolation is when you predict outside of the range

Least-Squares Fit

- Determines residuals
- The sum of the squares of the residuals has the least possible value

$$\text{Line of best fit: } y = ax + b$$

$$a = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$$

$$b = (\frac{\Sigma y}{n}) - a(\frac{\Sigma x}{n})$$

Example: Find the equation for the line of best fit and classify the correlation.

Age (x)	**Income ($000) (y)**	x^2	y^2	xy
33	33	1089	1089	1089
25	31	625	961	775
19	18	361	324	342
44	52	1936	2704	2288
50	56	2500	3136	2800
54	60	2916	3600	3240
38	44	1444	1936	1672
29	35	841	1225	1015
$\sum x = 292$	$\sum y = 329$	$\sum x^2 = 11712$	$\sum y^2 = 14975$	$\sum xy = 13221$

$$a = \frac{8(13221) - (292)(329)}{8(11712) - (292)^2} = 1.15$$

$$b = \frac{292}{8} - 1.15\frac{329}{8} = -0.85$$

$$y = 1.15x - 0.85$$

Correlation coefficient: A coefficient showing the strength of the correlation between two variables

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

$n =$ sample size

$$r = \frac{9700}{\sqrt{(8432)(8 \cdot 14975 - 329^2)}} = 0.98$$

$\therefore$ there is a very strong positive correlation

8.2 Cause and Effect

- There is a difference between correlation and causation

Causal Relationships:
Direct cause and effect: Δ independent $\longrightarrow \Delta$ dependent
Common cause: Δ independent $\longrightarrow \Delta$ dependent$_1$, Δ dependent$_2$
Reverse cause and effect: Δ dependent $\longrightarrow \Delta$ independent
Accidental relationship: Coincidental relationship, no causation
Presumed relationship: Assumed relationship, not necessarily true

8.3 Dynamic Analysis of Two-Variable Data

- Outlier: A point that does not follow the trend of data
- Hidden variable: A variable affecting the relationship between two variables
  - Can lead to a false correlation
- These can be found using a residual plot
  - Shows deviations from the trendline

Example: Using the below table and correlation coefficient, decide if an increase in recycling result in a reduction of landfill size.

Amount Recycled (kg)	**Amount of Garbage (kg)**
120	200
144	175
160	190
175	156
200	142
210	167
224	140
236	150

Table 2: Amount Recycled and Amount of Garbage

$$r = -0.84$$

There is a strong negative correlation, meaning that an increase in recycling will lead to a decrease in landfill size.

8.4 Uses and Misuses of Data

Data can be distorted in a number of ways, for example:

- Bias in wording
- Misrepresentation of graphical data
- Inadequate data collection
- Poor analysis

Season	Games	Fights
2012-13	720	347
2011-12	1230	546
2010-11	1230	645
2009-10	1230	714
2008-09	1230	734

Example: The table below shows the number of games and fights per season in the NHL. a) If the 2012-13 season was a lockout season, would it be safe to say that fights are exponentially decreasing?

No, since there were far less games that season, it makes sense for there to be far less fights.

If this outlier was removed, the hidden variable would be gone and the decrease would be less drastic.