MIE Lecture Notes

Probability and Statistics

Contents
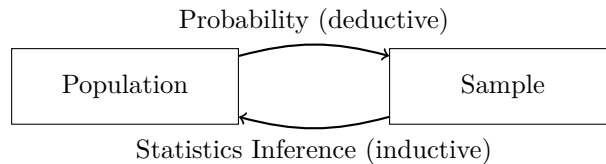
Chapter 1

Statistics Definitions

Global definition: Statistics involves collecting, organizing, summarizing, presenting, and analyzing data, as well as making inferences, conclusions, and decisions based on data.

Statistical definition: A <u>statistic</u> is a numerical value calculated from data (e.g. mean, proportion, standard deviation).

Probability (deductive)

Population	Sample

Statistics Inference (inductive)

Basic Terminology

<u>Individuals</u>: Objects on which data are collected (people, animals, plots of land, etc.).
<u>Variable</u>: Any characteristic of an individual.
<u>Population</u>: The entire group of individuals of interest.
<u>Sample</u>: A subset of individuals taken from the population.
<u>Statistical Inference</u>: Drawing conclusions about a population based on a sample.

Sampling Methods

<u>Simple Random Sample (SRS)</u>:

- Every possible group of size n has an equal chance of being selected.

- Helps avoid bias in sampling.

- Can be selected using random number tables or software.

<u>Stratified Random Sampling</u>:

- The population is divided into <u>homogeneous groups</u> *(individuals are similar with respect to the variable being studied)* called <u>strata</u>.

- A simple random sample is taken from each <u>stratum</u>. *(one subgroup of the population created)*

- Ensures that important subgroups are neither over nor under represented.

Types of Variables

<u>Categorical Variable</u>: Places individuals into categories (e.g. gender, major). These are qualitative.

<u>Quantitative Variable</u>: Takes numerical values for which arithmetic operations are meaningful.

- <u>Discrete</u>

- <u>Continuous</u>

Distributions

<u>Distribution</u>: Describes what values a variable takes and how often those values occur. When examining a distribution, look for:

- **Shape**

- **Center**

- **Spread**

- **Outliers**

<u>Outlier</u>: An individual value that falls outside the overall pattern of the data.

Describing Distributions with Numbers

<u>Central Tendency</u>: Describes where the data cluster or center.

<u>Central Tendency</u>: Describes where the data cluster or center.

- <u>Mean</u>: average value

- <u>Median</u>: middle value

<u>Mean (Arithmetic Mean)</u>:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

<u>Median</u>:

$$\tilde{x} = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & \text{if } n \text{ is odd} \\ \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}, & \text{if } n \text{ is even} \end{cases}$$

Theorem 1.1

1. The mean is more sensitive to extreme values than the median.

2. Changing a single data value will always change the mean, but may not change the median.

3. If a distribution is exactly symmetric, the mean and median are equal.

Trimmed Mean: The mean computed after removing extreme values.

$$\bar{x}_{\text{trim}} = \frac{1}{n-2k} \sum_{i=k+1}^{n-k} x_{(i)}$$

where k values are removed from both ends of the ordered data. (normally given in question like 10%)

Measures of Spread

Range: Maximum minus minimum. Very sensitive to extreme values.

Sample Variance: Measures the average squared deviation from the mean.

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Standard Deviation: The square root of the sample variance.

$$s = \sqrt{s^2}$$

Degrees of Freedom: The number of independent pieces of information available to estimate variability. For sample variance: $df = n - 1$.

Plots

Scatter Plot: Used to display the relationship between two quantitative variables (x, y). A scatter plot helps identify trends, patterns, and associations between variables.

Stem-and-Leaf Plot: An intermediate step between raw data and a frequency table. Preserves the original data values while showing the distribution.

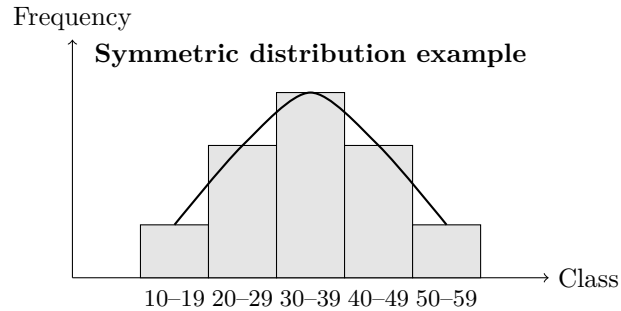| Stem	Leaf
1	2 4 7
2	1 3 5 8
3	0 4 6

Relative Frequency Table: Shows the proportion of observations in each class.

Class Interval	Class Midpoint	Frequency	Relative Frequency
10–19	14.5	3	0.30
20–29	24.5	4	0.40
30–39	34.5	3	0.30

Histogram: A graphical representation of a frequency or relative frequency table using contiguous bars.

When describing the shape of a histogram, we commonly classify it as:

- **Symmetric**

- **Skewed right** (positively skewed)

- **Skewed left** (negatively skewed)

Frequency

Symmetric distribution example

10–19 20–29 30–39 40–49 50–59 → Class

Chapter 2, Jan 9th

Experiment: A process that generates an outcome.

Sample Space (S): The set of all possible outcomes of an experiment.

Example 1:

Select 3 items from a production line. Each item can be classified as either defective (D) or non-defective (N).

$$S = \{DDD, DDN, DND, NDD, DNN, NDN, NND, NNN\}$$

Since each item has 2 possible outcomes,
$$|S| = 2^3 = 8$$

Example 2:

$$S = \{(x, y) \mid x^2 + y^2 \leq 4\}$$

Event (A): A subset of the sample space S.

Examples of events:
$$A = \{DDD, DDN, DND, NDD\}$$

$$B = \{NNN\}$$

$$C = \{(x, y) \mid x^2 + y^2 \leq 4\}$$
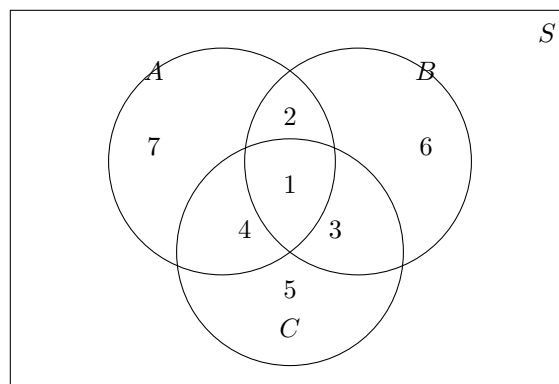
Event Operations:

- Complement: A^c (or A')

- Intersection: $A \cap B$

- Union: $A \cup B$

- Null Event: $\varnothing$

If

$$A \cap B = \varnothing,$$

then A and B are mutually exclusive.

Example (Venn Diagram):

$$A = \{DDD, DDN, DND, NDD\}, \quad B = \{NNN\}$$

$$A \cup B = \{DDD, DDN, DND, NDD, NNN\}$$

$$A \cap B = \varnothing$$

Jan 12th

Chapter 2: January 12

Review

1. *Experiment: A process that generates an outcome.*

2. *Sample Space (S): The set of all possible outcomes of an experiment.*

3. *Event Operations:*

   - *Complement: A' (A^c)*

- *Intersection: $A \cap B$*
- *Union: $A \cup B$*
- *Null Event: $\varnothing$*

4. If $A \cap B = \varnothing$, then A and B are called <u>mutually exclusive</u>.

$$(A \cap B)' = A' \cup B'$$

$$(A \cup B)' = A' \cap B'$$

$$A \cap \varnothing = \varnothing$$

$$A \cup \varnothing = A$$

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

Probability

$P(A) = $ <u>probability</u> of event A: the proportion of times the event occurs in infinitely many repetitions of the experiment.

Theorem2.1:
$$0 \le P(A) \le 1$$

$$P(A) + P(A') = 1$$

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

$$\begin{aligned} P(A \cup B \cup C) = {}& P(A) + P(B) + P(C) \\ & - P(A \cap B) - P(A \cap C) - P(B \cap C) \\ & + P(A \cap B \cap C) \end{aligned}$$

Mutually Exclusive Events

<u>Definition</u>: If $A_1, A_2, \ldots, A_n$ are mutually exclusive, then

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = P(A_1) + P(A_2) + \cdots + P(A_n)$$

If

$$A_1 \cup A_2 \cup \cdots \cup A_n = S,$$

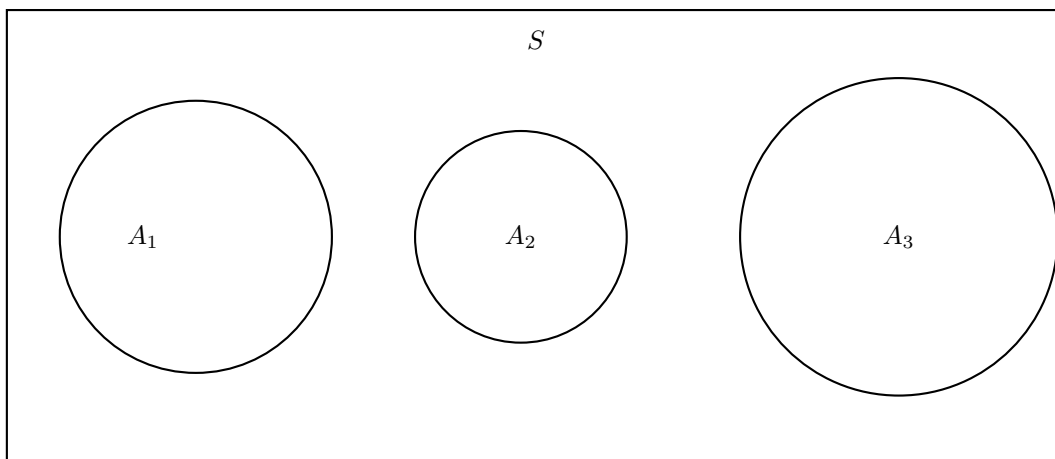then $\{A_1, A_2, \ldots, A_n\}$ is a <u>partition</u> of S.

Figure 1: Partition of the sample space S into A_1, A_2, A_3

Example

In a class of 33 students:

- 17 earned an A on the midterm

- 14 earned an A on the final

- 11 earned no A on either exam

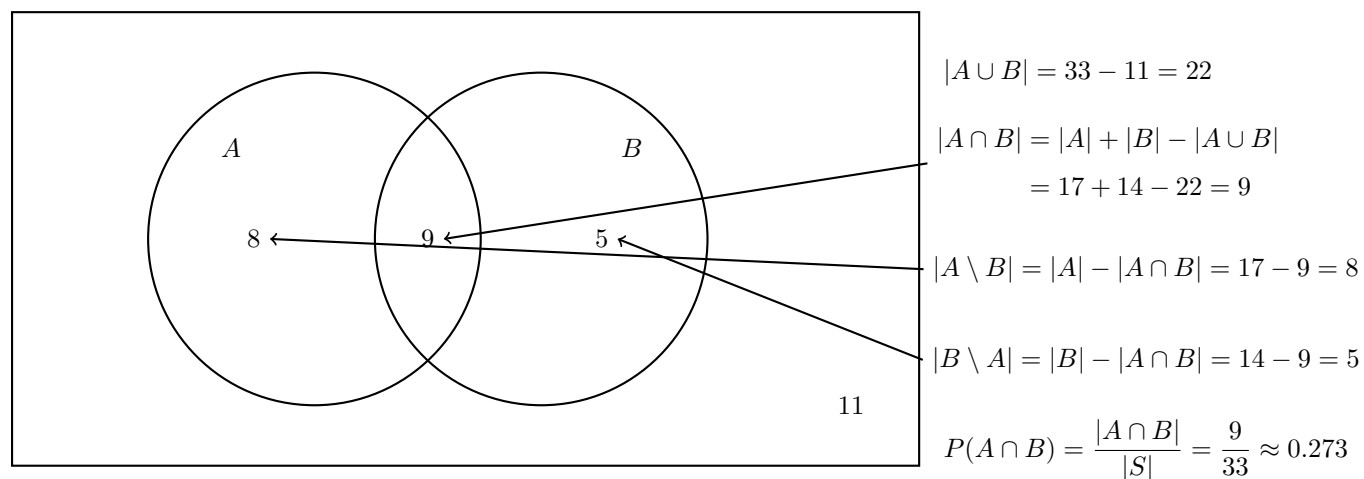Find the probability that a randomly selected student earned A's on <u>both</u> exams.

$$|A \cup B| = 33 - 11 = 22$$

$$|A \cap B| = |A| + |B| - |A \cup B|$$
$$= 17 + 14 - 22 = 9$$

$$|A \setminus B| = |A| - |A \cap B| = 17 - 9 = 8$$

$$|B \setminus A| = |B| - |A \cap B| = 14 - 9 = 5$$

$$P(A \cap B) = \frac{|A \cap B|}{|S|} = \frac{9}{33} \approx 0.273$$

Figure 2: Events A: A on midterm, B: A on final, with region counts and calculations

Theorem 2.2 (Equally Likely Outcomes):

If the sample space S has a finite number of outcomes and all outcomes are equally likely, then for any event A,

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

7

where

A: the event of interest (a subset of the sample space S),

S: the sample space, i.e. the set of all possible outcomes.

Example 1: Poker Hands Basics

A standard deck has:

$$4 \text{ suits} \times 13 \text{ denominations (A,2,3,\ldots,Q,K)} = 52 \text{ cards.}$$

A poker hand consists of 5 cards chosen from 52:

$$|S| = \binom{52}{5} = 2{,}598{,}960.$$

Combinations Reminder

If there are 3 objects $\{A, B, C\}$ and we choose 2:

$$\binom{3}{2} = \frac{3!}{(3-2)!2!}.$$

Order does <u>not</u> matter.

Example 2: Probability of 2 Aces and 1 Jack

A 5-card hand contains:

- exactly 2 aces,

- exactly 1 jack,

- 2 cards that are neither aces nor jacks.

$$P(2 \text{ aces and 1 jack}) = \frac{\binom{4}{2}\binom{4}{1}\binom{44}{2}}{\binom{52}{5}}.$$

Example 3: Probability of a Full House

A full house consists of:

- 3 cards of one denomination

- 2 cards of a different denomination

Number of full house hands:

$$\binom{13}{1}\binom{4}{3}\binom{12}{1}\binom{4}{2}.$$

Thus,

$$P(\text{full house}) = \frac{\binom{13}{1}\binom{4}{3}\binom{12}{1}\binom{4}{2}}{\binom{52}{5}}.$$

Example 4: Probability of Four of a Kind

A four of a kind consists of:

- 4 cards of the same denomination

- 1 remaining card of a different denomination

Number of such hands:

$$\binom{13}{1}\binom{4}{4}\binom{48}{1}.$$

Thus,

$$P(\text{four of a kind}) = \frac{\binom{13}{1}\binom{4}{4}\binom{48}{1}}{\binom{52}{5}}.$$

Example 5: Probability of Exactly One Pair

An **excatly** one-pair hand consists of:

- 1 pair

- 3 cards of different denominations, none matching the pair

Number of such hands:

$$\binom{13}{1}\binom{4}{2}\binom{12}{3}\binom{4}{1}^3.$$

Thus,

$$P(\text{exactly one pair}) = \frac{\binom{13}{1}\binom{4}{2}\binom{12}{3}\binom{4}{1}^3}{\binom{52}{5}}.$$

Note*: Counting Patterns

$$\binom{a}{b}$$

Meaning: Choose b different items from a **at once**, order does not matter.

Key features:

- No repeats

- Grouped choice

- Used when items must be <u>distinct</u>

$$\binom{a}{1}^{b}$$

Meaning: Make b <u>independent choices</u>, each time choosing 1 item from c.

Key features:

- Repeats allowed

- Choices are independent

- Used when selections do <u>not</u> restrict each other

Rule to Remember:

$$\text{Different items, no repeats} \implies \binom{a}{b}$$

$$\text{Independent choices} \implies \binom{c}{1}^{b}$$

Chapter 2 continue, Jan 14

Review

1. <u>*Probability*</u> *is the proportion of times the event occurs in infinitely many repetitions of the experiment.*

2. $0 \le P(A) \le 1$

3. $P(A) + P(A^c) = 1$

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

5.
$$
\begin{aligned}
P(A \cup B \cup C) = {} & P(A) + P(B) + P(C) \\
& - P(A \cap B) - P(A \cap C) - P(B \cap C) \\
& + P(A \cap B \cap C)
\end{aligned}
$$

6. <u>*Permutation*</u>*: A permutation counts ordered arrangements.*

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

Example 1: Two fair dice

A pair of fair dice are rolled. Find the probability that the second die lands on a smaller value than the first.

The outcomes where the second die is smaller than the first are represented below.

First Die (Stem)	Second Die (Leaf)
2	1
3	1 2
4	1 2 3
5	1 2 3 4
6	1 2 3 4 5

There are 15 favorable outcomes and 36 total outcomes.

$$P(\text{second} < \text{first}) = \frac{15}{36} = \frac{5}{12}.$$

Conditional Probability

The conditional probability of an event B given that event A has occurred is the probability that B occurs when it is known that A has occurred.

$$\boxed{P(B \mid A) = \frac{P(A \cap B)}{P(A)}, \quad P(A) > 0}$$

Example 2: Drinking Survey

A survey records the following data:

	D	N	Total
M	19	41	60
F	12	28	40
Total	31	69	100

The symbols used above are defined as follows:

- M: male

- F: female

- D: the individual drinks

- N: the individual does not drink

$$P(D|M) = \frac{19}{60} \qquad P(M|D) = \frac{19}{31}$$

Law of Total Probability

Theorem 2.3: If $B_1, \ldots, B_k$ form a partition (do not overlap but covers the whole sample space) of S with $P(B_i) > 0$, then for any event A,

$$P(A) = \sum_{i=1}^{k} P(A|B_i)P(B_i)$$

Example 3: Monty Hall (3 doors)

Car location	Monty opens	Probability	Stay	Switch
Door 1	Door 2	$\frac{1}{6}$	Car	Goat
Door 1	Door 3	$\frac{1}{6}$	Car	Goat
Door 2	Door 3	$\frac{1}{3}$	Goat	Car
Door 3	Door 2	$\frac{1}{3}$	Goat	Car

Staying wins only when the car is behind Door 1, so

$$P(\text{win by staying}) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}.$$

Switching wins when the car is behind Door 2 or Door 3, so

$$P(\text{win by switching}) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}.$$

Example 4: Birthday Problem

Assume the following:

- Leap years are ignored

- All 365 birthdays are equally likely

- Birthdays of different people are independent

Question: What is the probability that at least two people share the same birthday in a group of n people?

Rather than computing this directly, we use the complement rule.

$$P(\text{at least one match}) = 1 - P(\text{no match})$$

Probability of no shared birthdays

- Person 1 can have any birthday: probability 1

- Person 2 must avoid that birthday: $\frac{364}{365}$

- Person 3 must avoid the first two birthdays: $\frac{363}{365}$

- …

- Person n must avoid the previous $n-1$ birthdays: $\frac{365-(n-1)}{365}$

Therefore,

$$P(\text{no match}) = \frac{365}{365} \cdot \frac{364}{365} \cdot \frac{363}{365} \cdots \frac{365 - (n-1)}{365}$$

or equivalently,

$$P(\text{no match}) = \prod_{k=0}^{n-1} \frac{365 - k}{365}$$

Final result

$$P(\text{at least one shared birthday}) = 1 - \prod_{k=0}^{n-1} \frac{365 - k}{365}$$

Important values

- For $n = 23$: $P(\text{at least one match}) \approx 0.507$

- For $n = 57$: $P(\text{at least one match}) \approx 0.99$