Lecture 2

Classifying, Organizing, and Collecting Data

§1.2, §1.4, §1.5

1. Classifying and Storing Data

1.1 Data Sets, Samples, Populations, and Variables

Definition 1.1. The collection of data is called a **data set** or a **sample**.

Definition 1.2. The **population** refers to the set or group that contains everything relevant to the data.

For example, suppose we want to learn about the opinions of customers at a local store.

- The population could be all possible customers.
- A sample could be the customers who are actually selected and asked questions.

When we collect data, the characteristics of the data are called **variables**.
Examples include gender, weight, temperature, age, and length.
Variables can initially be categorized as

$$\boxed{\text{Variables} \longrightarrow \begin{cases}\text{Numerical variables}\\ \text{Categorical variables}\end{cases}}$$

Definition 1.3. A **numerical variable** records a numerical characteristic, such as temperature, weight, or age.

Definition 1.4. A **categorical variable** records a characteristic using categories or labels, such as gender, color, or shape.

1.2 Quantitative and Qualitative Characteristics

Numerical characteristics are sometimes called **quantitative characteristics**; categorical characteristics are sometimes called **qualitative characteristics**.
Examples of quantitative characteristics include length, temperature, and age. Examples of qualitative characteristics include color, area code, size, and shape.

$$\boxed{\text{Numerical} = \text{quantities}} \qquad \boxed{\text{Categorical} = \text{groups or labels}}$$

1.3 Example: Crash-Test Dummy Studies

The following table contains data from crash-test dummy studies.

Make	Model	Doors	Weight	Head Injury
Acura	Integra	2	2350	599
Chevrolet	Camaro	2	3070	733
Chevrolet	S-10 Blazer 4X4	2	3518	834
Ford	Escort	2	2280	551
Ford	Taurus	4	2390	480
Hyundai	Excel	4	2200	757
Mazda	626	4	2590	846
Volkswagen	Passat	4	2990	1182
Toyota	Tercel	4	2120	1138

1. How many variables does this table have? **Answer:** 5 variables.
2. How many observations does this table have? **Answer:** 9 observations.
3. For each variable, identify whether it is numerical or categorical. **Answer:** Make and Model are categorical; Doors, Weight, and Head Injury are numerical.

Variable	Type
Make	Categorical
Model	Categorical
Doors	Numerical
Weight	Numerical
Head Injury	Numerical

Remember: A variable is a characteristic recorded in the data; an observation is one row of the data set.

1.4 Coding Categorical Data Using Numbers

Sometimes categorical data are stored using numbers.

Weight	Gender	Smoke
7.69	Female	No
0.88	Male	Yes
6.00	Female	No
7.19	Female	No
8.06	Female	No
7.94	Female	No

For example,

$$\text{Female} = 1, \quad \text{Male} = 0, \qquad \text{No} = 0, \quad \text{Yes} = 1.$$

The coded data are

Weight	Female	Smoke
7.69	1	0
0.88	0	1
6.00	1	0
7.19	1	0
8.06	1	0
7.94	1	0

The variable is still categorical.

Important: Coding a categorical variable using numbers does not make it numerical. The numbers are simply labels.

For example, a ZIP code such as 29649 is still categorical.

1.5 Five Types of Variables

We can further break down variables into five types:

Categorical	Numerical
Nominal	Discrete
Ordinal	Continuous
Binary	

Variables
$\downarrow$
Categorical **Numerical**
$\downarrow$ $\downarrow$
Nominal Discrete
Ordinal Continuous
Binary

1.5.1 Nominal Variables

Definition 1.5. A **nominal variable** is a categorical variable whose categories are names or labels without a natural order.

Examples: color, shape, Social Security number, ZIP code, and area code.

$$\text{Red}, \qquad \text{Blue}, \qquad \text{Green}$$

have no natural ranking.

1.5.2 Ordinal Variables

Definition 1.6. An **ordinal variable** is a categorical variable whose categories have a meaningful order.

Examples: letter grades; first, second, and third place; and small, medium, and large.

$$\text{Poor} < \text{Fair} < \text{Good} < \text{Excellent}$$

1.5.3 Binary Variables

Definition 1.7. A **binary variable** is a categorical variable with exactly two categories.

Examples:

$$\boxed{\text{Yes/No}} \qquad \boxed{\text{Pass/Fail}} \qquad \boxed{\text{Present/Absent}}$$

A binary variable is therefore a special type of categorical variable.

1.5.4 Discrete Variables

Definition 1.8. A **discrete variable** is a numerical variable obtained by counting.

Examples: number of chairs, students, siblings, or cars.

$$25, \quad 26, \quad 27, \quad 28, \ldots$$

1.5.5 Continuous Variables

Definition 1.9. A **continuous variable** is a numerical variable obtained by measuring.

Examples: temperature, length, weight, and height.

$$68 \text{ in}, \qquad 68.1 \text{ in}, \qquad 68.17 \text{ in}$$

Measurements can be recorded with increasingly precise values.

1.6 Summary of the Five Types

Type	Description	Examples
Nominal	Names or labels; no natural order	Color, shape, ZIP code
Ordinal	Categories with an order	Letter grade, class rank
Binary	Exactly two categories	Yes/No, Pass/Fail
Discrete	Numerical; obtained by counting	Number of chairs
Continuous	Numerical; obtained by measuring	Temperature, length

1.7 Example: A Local Store

Suppose a local store is interested in whether a new product would sell. The manager takes a random sample of 100 customers over a two-week period.

Each person is asked:

1. Would you buy the product?
2. How many times would you buy the product over a six-month period?

a) What is the population? **Answer:** All possible customers.

b) What is the sample? **Answer:** The 100 customers selected.

c) What are the variables? **Answer:** Whether the customer would buy the product and how often the customer would buy it.

d) Classify each variable. **Answer:** Would you buy? — categorical, binary. How often? — numerical, discrete.

> **Key Idea:** The population is the larger group of interest; the sample is the group from which data are actually collected.

2. Organizing Categorical Data

2.1 Frequency

Definition 2.1. In statistics, **frequency** is the number of times a value of a variable is observed in a data set.

For

$$\text{Yes, No, Yes, Yes, No, Yes,}$$

the frequency of Yes is $\boxed{4}$, the frequency of No is $\boxed{2}$, and the total is 6.

2.2 Relative Frequency

Definition 2.2. Relative frequency, or proportion, is the frequency divided by the total frequency.

$$\boxed{\text{Relative frequency} = \frac{\text{frequency}}{\text{total frequency}}}$$

It can be written as a fraction, decimal, or percentage.

For example,

$$\frac{2}{17} = 0.1176\ldots \approx 11.8\%.$$

$$\boxed{\text{Percentage} = \text{Relative frequency} \times 100\%}$$

2.3 Example: Seat Belt Survey

Suppose a national survey asks American youths whether they always wear a seat belt while driving or riding in a car.

	Male	**Female**	**Total**
Not Always	2	3	5
Always	4	8	12
Total	**6**	**11**	**17**

Part (a): Find the total number of males, females, and participants.

From the table,

$$\text{Males} = 2 + 4 = 6,$$

$$\text{Females} = 3 + 8 = 11,$$

and

$$\text{Total participants} = 6 + 11 = 17.$$

Answer: $\boxed{\text{6 males, 11 females, 17 participants}}$

As a check,

$$6 + 11 = 17.$$

Part (b): Identify the frequencies and compute the percentages.

The **frequency** is the number of individuals in each category. Since there are 17 participants in total, each overall percentage is calculated as

$$\text{Percentage} = \frac{\text{Frequency}}{\text{Total}} \times 100\%.$$

	Male	**Female**	**Total**
Not Always	$\frac{2}{17} = 11.8\%$	$\frac{3}{17} = 17.6\%$	$\frac{5}{17} = 29.4\%$
Always	$\frac{4}{17} = 23.5\%$	$\frac{8}{17} = 47.1\%$	$\frac{12}{17} = 70.6\%$
Total	$\frac{6}{17} = 35.3\%$	$\frac{11}{17} = 64.7\%$	100%

For example,

$$\frac{2}{17} \approx 0.1176 = 11.8\%.$$

Answer: $\boxed{\text{Each overall percentage is found by dividing the frequency by 17.}}$

Part (c): Are males or females more likely to take the risk of not wearing a seat belt?

To compare males and females, compare the proportion within each gender who answered "Not Always." For males,

$$\frac{2}{6} \approx 0.333 = 33.3\%.$$

For females,

$$\frac{3}{11} \approx 0.273 = 27.3\%.$$

Since

$$33.3\% > 27.3\%,$$

Answer: Males are more likely to report "Not Always."

Important: We should not compare the raw frequencies 2 and 3 directly because the two groups have different sizes. There are 6 males but 11 females.

Part (d): Should we use frequencies or relative frequencies to make comparisons?

When comparing groups of different sizes, we should use **relative frequencies** (proportions or percentages) rather than raw frequencies.
Here,

$$\frac{2}{6} = 33.3\% \qquad \text{and} \qquad \frac{3}{11} = 27.3\%.$$

Therefore,
Answer: Use relative frequencies when comparing groups of different sizes.

- **Frequency** = the actual number of individuals.
- **Relative frequency** = the proportion or percentage within a group.
- Use **frequencies** to describe counts.
- Use **relative frequencies** to make fair comparisons between groups of different sizes.

3. Collecting Data to Understand Causality

3.1 Observational Studies

Definition 3.1. In an **observational study**, we observe individuals and measure variables of interest without attempting to influence the responses.

Observe, but do not disturb.

For example, recording students' existing exercise habits and exam scores is an observational study.

3.2 Controlled Experiments

Definition 3.2. In a **controlled experiment**, we deliberately impose a treatment on individuals and observe their responses.

$$\boxed{\text{Subjects} \longrightarrow \begin{cases} \text{Treatment Group} \\ \text{Control Group} \end{cases}}$$

The treatment group receives the treatment; the control group provides a comparison.

Observational study:	observe existing conditions
Controlled experiment:	impose a treatment

3.3 Anecdotal Evidence

Definition 3.3. Anecdotal evidence is evidence based on an individual's experience or story.

For example:

> "My friend used a product and said it worked, so it must work for everyone."

Answer: This is anecdotal evidence and does not establish cause and effect.

3.4 Association and Causality

A fundamental principle is

$$\boxed{\text{Association} \neq \text{Causation}}$$

Anecdotal evidence and observational studies can show an association, but they do not by themselves establish causality.

Answer: A controlled experiment provides evidence for cause and effect because the researcher deliberately imposes the treatment.

3.5 Confounding Variables

Definition 3.4. A **confounding variable** is another variable that could explain a difference in outcomes between groups.

For example,

$$\text{More exercise} \longrightarrow \text{Higher exam scores}$$

might actually be influenced by study habits.

Answer: Study habits could be a confounding variable because they provide another explanation for the difference in exam scores.

> **Key Idea:** A confounding variable can provide another explanation for an observed difference.

3.6 How to Design a Good Experiment

A well-designed experiment uses several important principles.

3.6.1 1. Random Allocation

Participants should be randomly allocated to treatment and control groups.

$$\text{Participants} \longrightarrow \begin{cases} \text{Treatment Group} \\ \text{Control Group} \end{cases}$$

Answer: Random allocation helps prevent systematic differences between groups before the experiment begins.

3.6.2 2. Use of a Placebo

Definition 3.5. A **placebo** is a fake treatment.

For example, a sugar pill may be used as a placebo in a medication study.

Definition 3.6. The **placebo effect** is a response caused by expectations about a treatment rather than by the active treatment itself.

Answer: A placebo helps researchers determine whether expectations about the treatment affect the results.

3.6.3 3. Blinding the Study

Blinding helps reduce bias.

Definition 3.7. A **single-blind** study is one in which the subject is unaware of the treatment received.

Definition 3.8. A **double-blind** study is one in which both the subjects and the researchers interacting with them are unaware of the treatment assignment.

Study	**Who is unaware?**
Single blind	Subject
Double blind	Subject and researcher

Answer: Blinding is used to reduce bias.

3.6.4 4. Large Sample Size

A large sample helps account for variability. A very small sample can allow individual differences to have a large effect.

$$\boxed{\text{Large sample} \longrightarrow \text{better account for variability}}$$

Answer: A large sample provides more information, although a large sample alone cannot fix a poorly designed experiment.

3.7 The "Gold Standard" in Experiments

A good experiment should consider:

1. **Random allocation** — randomly assign participants.
2. **Placebo, if appropriate** — use a fake treatment.
3. **Blinding** — reduce bias.
4. **Large sample size** — account for variability.

$$\boxed{\text{Random allocation} \longrightarrow \text{Placebo} \longrightarrow \text{Blinding} \longrightarrow \text{Large sample}}$$

Answer: Not every experiment uses a placebo, but random allocation, appropriate blinding, and adequate sample size are important design principles.

4. Putting the Ideas Together

Consider the question:

> Does a new treatment cause an improvement in some outcome?

If we simply observe people who choose to use the treatment, we have an observational study. If we deliberately assign people to receive the treatment or a control condition, we have a controlled experiment.

$$\boxed{\text{Association does not automatically mean causation.}}$$

Answer: A controlled experiment provides stronger evidence for cause-and-effect because the researcher controls the treatment.

5. Practice Questions

5.1 Classifying Data

For each variable, identify whether it is numerical or categorical. If possible, identify the more specific type.

1. Color of a car. **Answer:** Categorical, nominal.
2. Temperature. **Answer:** Numerical, continuous.
3. Number of chairs in a classroom. **Answer:** Numerical, discrete.
4. Size of a shirt: S, M, or L. **Answer:** Categorical, ordinal.
5. Whether a student owns a car: Yes or No. **Answer:** Categorical, binary.
6. Shape of an object. **Answer:** Categorical, nominal.
7. Length of a table. **Answer:** Numerical, continuous.
8. Number of students in a class. **Answer:** Numerical, discrete.

5.2 Population and Sample

A local store randomly selects 100 customers and asks whether they would buy a new product.

a) What is the population? **Answer:** All possible customers.
b) What is the sample? **Answer:** The 100 customers selected.
c) What is the variable? **Answer:** Whether the customer would buy the product.
d) Is the variable numerical or categorical? **Answer:** Categorical, binary.

5.3 Frequency and Relative Frequency

Suppose a survey asks 20 students whether they own a bicycle.

	Yes	**No**
Frequency	12	8

1. What is the total number of students? **Answer:** 20.
2. What is the frequency of "Yes"? **Answer:** 12.
3. What is the relative frequency of "Yes"? **Answer:** $12/20 = 0.60$.
4. Express the relative frequency as a percentage. **Answer:** 60%.
5. What percentage of students do not own a bicycle? **Answer:** $8/20 = 40\%$.

$$\boxed{\frac{12}{20} = 0.60 = 60\%, \qquad \frac{8}{20} = 0.40 = 40\%}$$

5.4 Observational Study or Experiment?

For each situation, determine whether it is an observational study or a controlled experiment.

1. Researchers observe how much exercise people get without changing their exercise habits.
   Answer: Observational study.
2. Researchers randomly assign people to receive a new treatment or a control treatment.
   Answer: Controlled experiment.
3. A researcher records students' existing study habits and exam scores.
   Answer: Observational study.
4. Researchers assign subjects to two different treatments and compare their responses.
   Answer: Controlled experiment.

5.5 Association or Causation?

Consider the following statements.

1. People who exercise more tend to have lower resting heart rates.
   Answer: Association; this alone does not establish causation.
2. A friend says that a particular product made them feel better.
   Answer: Anecdotal evidence; it does not establish causation.
3. Researchers conduct a controlled experiment in which subjects are randomly assigned to treatment and control groups.

Answer: This can provide evidence of causation.

$$\boxed{\text{Association} \neq \text{Causation}}$$