

Statistics involving two categorical variables

1. A researcher records students' **grade level** (Freshman, Sophomore, Junior, Senior) and whether they **participate in sports** (Yes/No). Which display is most appropriate?

- A. Scatterplot
 - B. Dot plot
 - C. Two-way table
 - D. Histogram
-

2. Which of the following involves **two categorical variables**?

- A. Height and weight
- B. Eye color and favorite subject
- C. Age and income
- D. Time studied and test score

Two-way table

双向表

Joint frequency

联合频数

Marginal distribution

边际分布

Conditional distribution

条件分布

Association

关联

- Data are organized using a **two-way (contingency) table**.
 - Analysis focuses on **conditional relative frequencies**, not raw counts.
- Graphs such as **segmented bar charts** or **side-by-side bar charts** are used for comparison.
 - An **association** exists if conditional distributions differ across categories.
- Conclusions are written **in context**, avoiding cause-and-effect language unless justified.

A survey of 150 students records their **favorite type of music** (Pop, Rock, Classical) and whether they **own a musical instrument** (Yes, No).

(a) Create a **two-way table** to display the data.

(b) Define the **marginal distribution** of owning a musical instrument.

(c) Describe how you could determine if there is an **association** between music preference and owning an instrument.

(a) A **two-way table** organizes the counts for each combination of music preference and instrument ownership.

Music \ Owns Instrument	Yes	No	Total
Pop			
Rock			
Classical			
Total			150

(b) The **marginal distribution** of owning a musical instrument is the distribution of **Yes vs. No totals**, ignoring music preference.

(c) To determine **association**, compare **conditional distributions** of owning an instrument **for each music category**. If the percentages differ across categories, there is an association; if they are the same, the variables are independent.

STRETCH AND CHALLENGE

1. School Clubs and Grade Level

A survey of students recorded their **grade level** (9, 10, 11, 12) and whether they belong to at least one school club (Yes/No).

- a) Explain why both variables are categorical.
 - b) Describe how a two-way table could be used to summarize the data.
 - c) Explain how you would determine whether there is an association between grade level and club membership.
-

2. Exercise and Sleep Quality

A health survey classified participants by **exercise frequency** (Never, Sometimes, Often) and **sleep quality** (Poor, Average, Good).

- a) What are the explanatory and response variables, if any?
 - b) Explain how conditional percentages could be used to compare the groups.
 - c) What evidence in the conditional distributions would suggest an association?
-

3. Transportation and Satisfaction

Students reported their primary method of transportation to school (Bus, Walk, Car, Bicycle) and their satisfaction with the journey (Satisfied/Not Satisfied).

A researcher claims that students who walk are more satisfied than students who travel by car.

- a) Describe how the claim could be investigated using a two-way table.
- b) Why would comparing counts alone be misleading?
- c) What additional information would strengthen the conclusion?

1. School Clubs and Grade Level

- a) Both are categorical variables.
 - b) Use a two-way table of counts.
 - c) Compare club-membership percentages across grades.
-

2. Exercise and Sleep Quality

- a) Exercise = explanatory; sleep quality = response.
 - b) Find conditional percentages within exercise groups.
 - c) Different percentages indicate an association.
-

3. Transportation and Satisfaction

- a) Compare satisfaction percentages by transportation method.
- b) Counts can be misleading if group sizes differ.
- c) Conditional percentages and a larger sample strengthen conclusions.

Association between Variables

- **Association** — 关联
- **Independence** — 独立性
- **Categorical variable** — 分类变量
- **Two-way table** — 双向表
- **Conditional distribution** — 条件分布
- **Marginal distribution** — 边际分布
- **Conditional percentage** — 条件百分比
- **Joint probability** — 联合概率
- **Explanatory variable** — 解释变量
- **Response variable** — 响应变量

- Association describes whether two categorical variables are related.
- Two variables are **independent** if knowing one gives no information about the other.
- A **two-way table** is used to organise data from two categorical variables.
- The key analysis is based on **conditional distributions**, not just raw counts.
- **Conditional percentages** allow fair comparison between groups.
- If conditional distributions are similar, there is likely **no association**.
- If conditional distributions differ, there is likely an **association**.
- **Marginal distributions** show overall totals, ignoring the other variable.
- Association does **not imply causation**.
- Graphical displays like **segmented bar charts** help visualise differences between groups.

Question

A school administrator wants to investigate whether there is an association between **grade level** (9th, 10th, 11th, 12th) and **participation in after-school tutoring** (Yes/No). A random sample of 200 students is taken.

A two-way table is shown below:

Grade	Yes	No	Total
9th	30	20	50
10th	25	25	50
11th	20	30	50
12th	15	35	50

- (a) Calculate the conditional percentage of students in each grade who participate in tutoring.
- (b) Describe the relationship between grade level and tutoring participation.
- (c) Is there evidence of an association? Justify using conditional distributions.
- (d) Can we conclude that grade level causes changes in tutoring participation? Explain.

(a)

- 9th: 60%
 - 10th: 50%
 - 11th: 40%
 - 12th: 30%
-

(b)

Tutoring participation decreases as grade level increases.

(c)

Yes, conditional percentages differ across grades → there is an association.

(d)

No. This is observational data, so causation cannot be concluded.

True or False

- If two categorical variables are independent, their conditional distributions will be the same.

True or False

- A two-way table is used to display one categorical variable.

True or False

- Association between variables means one variable causes the other.