

AP Statistics - Two Categorical Variables and Two-Way Tables - Lesson

Common Core AP STAT 2.1, AP STAT 2.2, AP STAT 2.3, HSS-ID.B.5 · maths-pdf.fr/en/

A two-way table summarizes two categorical variables. Marginal relative frequencies describe one variable alone, joint relative frequencies describe two traits together, and conditional relative frequencies describe one variable within a group. If the conditional distributions of the response differ across the explanatory groups, the variables are associated.

AP Statistics · Common Core State Standards: **AP STAT 2.1, AP STAT 2.2, AP STAT 2.3, HSS-ID.B.5**

Key facts

- Two-way tables show counts for combinations of two categorical variables.
- Marginal = total / grand total; joint = cell / grand total; conditional = cell / group total.
- Compare conditional percentages of the response across explanatory groups to detect association.
- Segmented bar graphs display conditional distributions side by side.
- Association is not causation, and combined data can hide Simpson's paradox.

How do you organize two categorical variables?

A **two-way table** shows how individuals fall into categories of two categorical variables at once. When one variable might explain the other, it is the **explanatory variable** and the other is the **response variable**.

Job \ Year	Junior	Senior	Total
Has a job	30	55	85
No job	90	45	135
Total	120	100	220

Counts of 220 students by job status and year. Row and column totals are the marginal counts.

Here year in school is the explanatory variable and having a job is the response variable. The totals on the edges are the **marginal** counts.

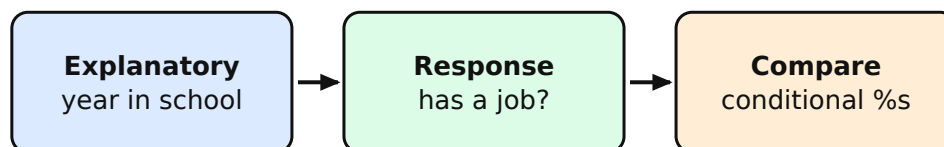
What are marginal, joint, and conditional relative frequencies?

Type	Question it answers	Example
Marginal	One variable alone	$85/220 = 38.6\%$ have a job
Joint	Both traits together	$30/220 = 13.6\%$ are juniors with a job
Conditional	Given one group	$30/120 = 25\%$ of juniors have a job

Three kinds of relative frequency from the same two-way table.

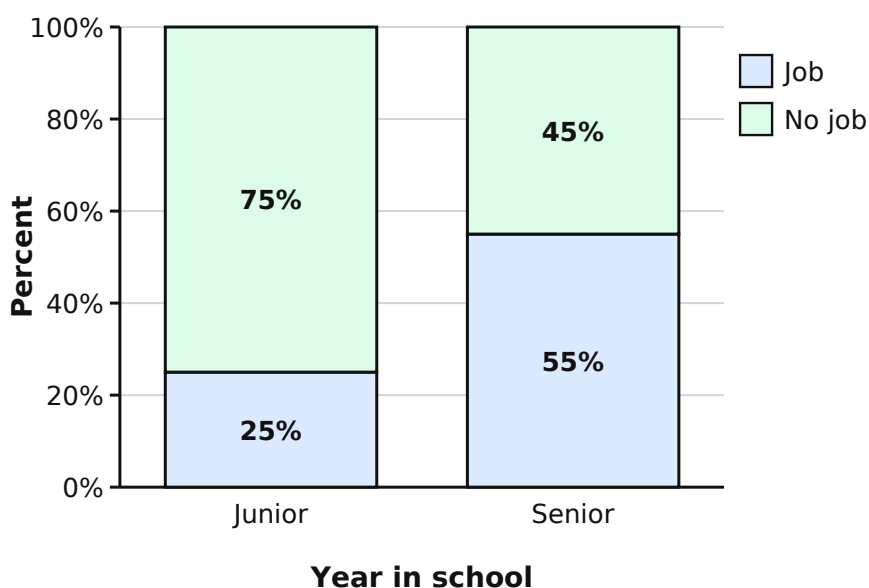
- A **marginal relative frequency** divides a row or column total by the grand total.
- A **joint relative frequency** divides one cell by the grand total.
- A **conditional relative frequency** divides a cell by the total of the group you are given.

How do you decide whether two variables are associated?



Condition on the explanatory variable, then compare the response percents.

1. Identify the explanatory and response variables.
2. Compute the conditional relative frequencies of the response within each explanatory group.
3. Compare them. If they differ noticeably, there is an **association**.
4. Describe it in context, with numbers.



Segmented bar graph of the conditional distribution of job status within each year.

Within juniors, $\frac{30}{120} = 25\%$ have a job. Within seniors, $\frac{55}{100} = 55\%$ do. The percents differ a lot, so seniors are more likely to have a job: year and job status are associated. A segmented bar graph makes the comparison visible.

Why can a table mislead?

Association is not causation. Being a senior does not cause a job; older students simply have more time and freedom. Also, a pattern in combined data can reverse when you split the data by a third variable, a reversal called **Simpson's paradox**. Always ask whether a lurking variable could change the story.

Questions about two categorical variables and two-way tables

What is the difference between joint and conditional relative frequency?

A joint relative frequency divides a single cell by the grand total, answering 'what fraction of everyone has both traits?' A conditional relative frequency divides by the total of a given group, answering 'among this group, what fraction has the trait?'

How do I choose which conditional percentages to compute?

Condition on the explanatory variable. Divide each cell by its explanatory group total so each group's response percents add to 100%.

What does it mean for two variables to be associated?

It means knowing the value of one variable helps predict the other. In a table, the conditional distributions of the response are different for different explanatory groups.

What is Simpson's paradox?

It is a reversal in which a comparison favoring one group in the combined data flips when the data are split by a lurking variable. It happens when the groups have very different sizes across the subgroups.

Can a segmented bar graph show causation?

No. It shows association only. Causation requires a well-designed randomized experiment that controls for other variables.