Analyzing One Categorical Variable

07.05.2025

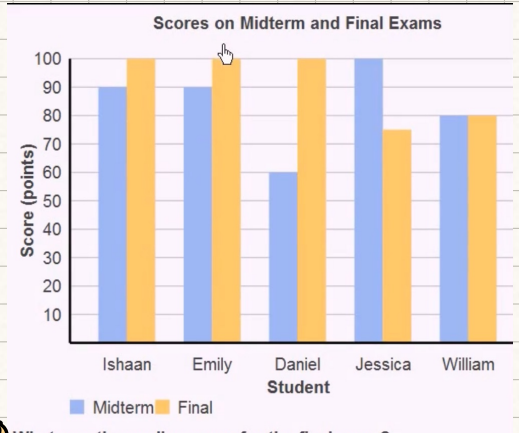
Identifying individuals, variables, and categorical variables in a data set

Millions of Americans rely on caffeine to get them up in the morning. Here's nutritional data on some popular drinks at Ben's Beans coffee shop:

Drink	Type	Calories	Sugars (g)	Caffeine (mg)
Brewed coffee	Hot	4	0	260
Caffè latte	Hot	100	14	75
Caffè mocha	Hot	170	27	95
Cappuccino	Hot	60	8	75
Iced brewed coffee	Cold	60	15	120
Chai latte	Hot	120	25	60

individuals — categorical — numeric — variables

Reading bar charts



1. What was the median score for the final exam?
2. What is the midrange of the midterm scores?
3. What was the average student score for the final exam?
4. What was the mode for the final exam scores?
5. What is the range of the midterm scores?

(1) finals = {100, 100, 100, 75, 80}
= {75, 80, 100, 100, 100}

(2) midterms = {90, 90, 60, 100, 80}
= {60, 80, 90, 90, 100}

$$\text{midrange} = \frac{\max + \min}{2} = 80$$

(3) $$\text{Mean (f)} = \frac{75+80+100+100+100}{5} = \frac{455}{5} = 91$$

(4) Mode (f) = 100 (freq = 3)

(5) Range (R) = max − min = 40

---

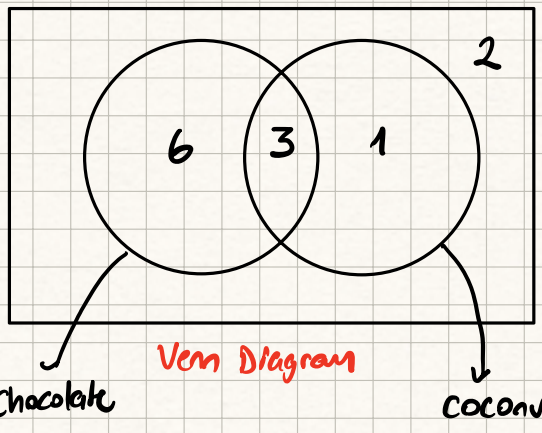
* Individual: A single member or unit of a population or sample being studied.
* Variable: A characteristic or attribute that can take on different values among individuals.
* Categorical variable: A variable that represents distinct groups or categories (e.g. gender, color)
* Bar chart: A graphical representation of categorical data using rectangular bars where the length or height of each bar corresponds to the value it represents.

Two-way Tables

07.05.2025

Two-way frequency tables and Venn diagrams

Data :



Venn Diagram

	Has coco.	No coco.	Total
Has choc.	3	6	9
No choc	1	2	3
Total	4	8	

Two-way Table

Two-way relative frequency tables

The two-way frequency table below shows data on type of vehicle driven and whether there was an accident in the last year for customers of All American Auto Insurance.

Complete the following two-way table of column relative frequencies.
(If necessary, round your answers to the nearest hundredth.)

	Sport utility vehicle (SUV)	Sports car
Accident within the last year	28	35
No accident within the last year	+ 97	+ 104
	125	139

	Sport utility vehicle (SUV)	Sports car
Accident within the last year	28/125 = 0.22	35/139 = 0.25
No accident within the last year	97/125 = 0.78	104/139 = 0.75
Column total	1.00	1.00

* when analyzing two-way relative tables, pay clos attention to whether it is column-relative or row-relative or both. If it's only column relative, the sum of the values on the same row does not equal to 100%

* Venn Diagrams : A visual representation of mathematical or logical relationships between different sets, using overlapping circles (or other shapes) to show common and distinct elements.

* Two-way Table : A table that displays the frequency distribution of two categorical variables, showing how the categories of one variable relate to the categories of another.

Two-way frequency table from categorical data

Lucio wants to test whether playing violent video games makes people more violent. He asks his friends whether they play violent video games, and whether they have been in a fight in the last month. He recorded the results in the table shown below.

Fill in the table to show the fraction of each group of students who have been in a fight.

Then decide whether there is an association between violent video games and getting in a fight among Lucio's friends.

	Fraction who have been in a fight	Fraction who haven't
Students who play violent video games	1/5 0.20	4/5 0.80
Students who don't	3/15 0.20	12/15 0.80

Name	Violent video games	Have been in a fight
Lavera Lev	No	No
Jessica Minelli	No	No
Peggie Dennehy	No	Yes
Ellen Messineo	Yes	Yes
Shannon Langan	No	No
Barbara Krum	No	No
Tamesha Kaelin	Yes	No
Love Pelley	No	No
Stephaine Jorstad	No	No
Marjorie Varela	Yes	No
Blake Montford	Yes	No
Julio Solano	No	No
Dell Valone	No	No
Kristyn Katz	No	No
Claribel Prothro	No	No
Britt Maple	No	No
Jeanelle Zeno	Yes	No
Karl Whitten	No	Yes
Herb Swarts	No	Yes
Shauna Lebeau	No	No

* Based on this data (with a very small sample size), we can say that there is no association between violent video games and getting in a fight.

Analyzing trends in categorical data

The relative frequency table below shows statistics from a study about the relationship between the amount of time a person spends using a computer before bed and the amount that a person sleeps each night. For computer use, each participant was classified as minimal, moderate, or extreme.

Suppose there were 17 people in this study who were both moderate computer users and got 5 - 7 hours of sleep. 17 = 10% · Total ⇒ Total = 170 people

How many people in this study were both extreme computer users and got 5 - 7 hours of sleep?
(Round to the nearest whole number.)

11.7% of total = 19.89 ≈ 20

20

Computer time	Hours/night:	5 or fewer	5 - 7	7 or more	Row total
Minimal	Row %	16.3	32.6	51.1	100
	Column %	17.5	35.0	55.0	------
	Total %	5.8	11.7	18.3	35.8
Moderate	Row %	37.1	34.3	28.6	100
	Column %	32.5	30.0	25.0	-----
	Total %	10.8	10.0	8.3	29.1
Extreme	Row %	47.6	33.3	19.1	100
	Column %	50.0	35.0	20.0	-----
	Total %	16.7	11.7	6.7	35.1
Column total	Row %	-----	-----	-----	100
	Column %	100	100	100	100
	Total %	33.3	33.4	33.3	100

Does the table show evidence of an association between being a minimal computer user and getting 7 hours of sleep or more?
Select all that apply.

- ☐ Yes, because 35.1% of people are extreme computer users, and 29.1% of people are moderate computer users.
- ☐ No, because the total column percentages are essentially all equal.
- ✓ Yes, because 51.1% of minimal computer users get 7 or more hours, and only 33.3% of all computer users get 7 or more hours.
- ☐ No, because the total percentage of extreme computer users who get 5-7 hours of sleep is the same as the total percentage of moderate computer users who get 5-7 hours of sleep.
- ✓ Yes, because 55.0% of people who get 7 or more hours are minimal computer users, and only 35.8% of all people are minimal computer users.

* when analyzing two-way relative tables pay attention to whether it is column-relative or row-relative or both. For example, if it's only column-relative. Sum of the values on the same row does not equal to 100%.

Distribution in Two-way Tables

08.05.2025

Marginal and conditional distributions

* **Joint distribution**: Two (or more) random variables occurring together.
* **Marginal distribution**: Ignoring (marginalizing out) other variables and giving the distribution of the remaing variable(s).
* **Conditional distribution**: The distribution of one random variable given that another random variable takes a specific value.

Time Studied (minutes)

% Correct	0-20	21-40	41-60	>60	Total
80-100	0	4	16	20	40
60-79	0	20	30	10	60
40-59	2	4	32	32	70
20-39	10	2	8	0	20
0-19	2	0	0	8	10
Total	14	30	86	70	200

* 20 out of 200 students worked 21 to 40 minutes and answered 60-79% correctly
* 70 out of 200 students answereed 40-59% correctly
* 14,28% (10 out of 70) of the students who studied more than 60 minutes answered 60-79% correctly.

* Joint Distribution: Two (or more) random variables occurring together. (e.g., 20 out of 200 students studied 21 to 40 minutes per day and answered 60-79% of all questions correctly.
* Marginal Distribution: Ignoring (marginalizing out) other variables and giving the distribution of remaing variables. (e.g. 70 out of 200 students answered 40-59% of questions correctly.
* Conditional Distribution: The distribution of one random variable given that another random variable takes a specific value. (e.g. 14.28% of the students who studied more than 60 minutes per day answered 60-79% of questions correctly