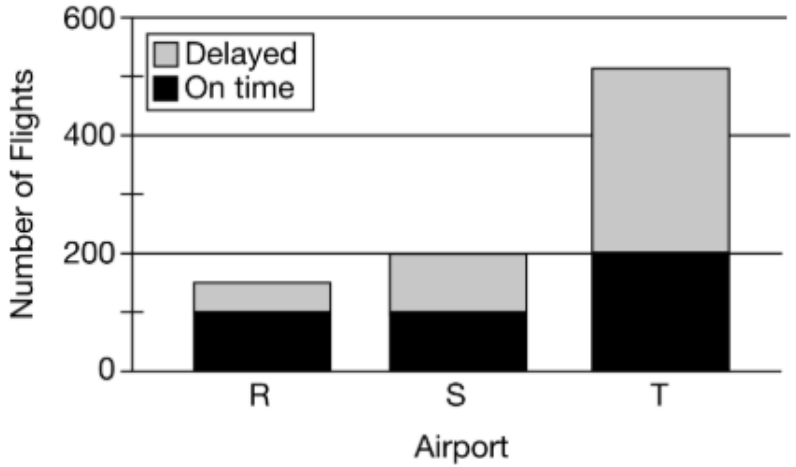
Representing two categorical variables

- **Two-way table (双向表 / 二维表格)**
- **Row variable (行变量)**
- **Column variable (列变量)**
- **Marginal distribution (边际分布)**
- **Conditional distribution (条件分布)**

- Two categorical variables are summarized using a **two-way (contingency) table**.
- Counts or **relative frequencies** are displayed for each category combination.
- **Segmented (stacked) bar charts** or **side-by-side bar charts** visually compare distributions.
- Analysis focuses on **comparing conditional distributions**.
- Conclusions must be stated **in context**, noting possible associations between variables.

The following segmented bar chart shows the number of flights that were either on time or delayed at three different airports on one day.



Which of the following statements is supported by the bar chart?

(A) Airport T has the greatest percentage of on-time flights compared to the other two airports.

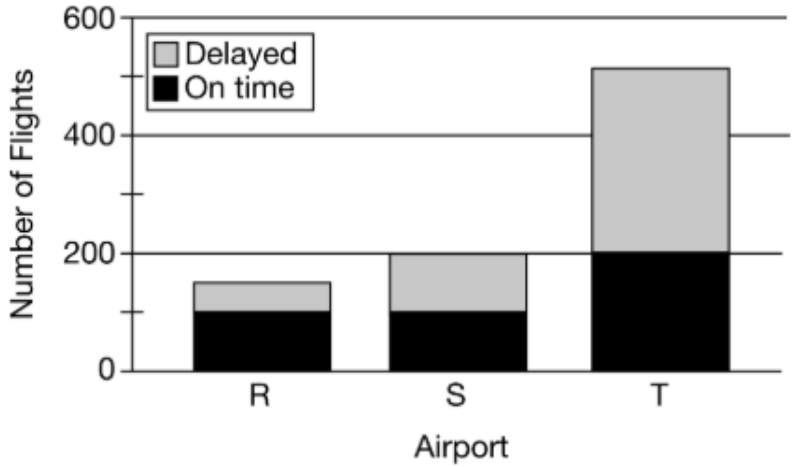
(B) Airport R has the least percentage of on-time flights compared to the other two airports.

(C) The number of on-time flights at Airport S is half the number of on-time flights at Airport T.

(D) The number of on-time flights at Airport R is less than the number of on-time flights at Airport S.

(E) The number of flights at Airport T is equal to the total number of flights at Airports R and S combined.

The following segmented bar chart shows the number of flights that were either on time or delayed at three different airports on one day.



Which of the following statements is supported by the bar chart?

C

(A) Airport T has the greatest percentage of on-time flights compared to the other two airports.

(B) Airport R has the least percentage of on-time flights compared to the other two airports.

(C) The number of on-time flights at Airport S is half the number of on-time flights at Airport T.

(D) The number of on-time flights at Airport R is less than the number of on-time flights at Airport S.

(E) The number of flights at Airport T is equal to the total number of flights at Airports R and S combined.

A survey of 100 students records their **favorite type of music** (Pop, Rock, Classical) and whether they **own a musical instrument** (Yes, No). The results are summarized in a two-way table.

(a) Identify the **row variable** and **column variable**.

(b) Calculate the **marginal distribution** of students who own a musical instrument.

(c) Describe the **conditional distribution** of music preference for students who own a musical instrument.

	YES	NO
Pop	25	20
Rock	20	10
Classical	15	10

A survey of 100 students records their **favorite type of music** (Pop, Rock, Classical) and whether they **own a musical instrument** (Yes, No). The results are summarized in a two-way table.

(a) Identify the **row variable** and **column variable**.

(b) Calculate the **marginal distribution** of students who own a musical instrument.

(c) Describe the **conditional distribution** of music preference for students who own a musical instrument.

	YES	NO
Pop	25	20
Rock	20	10
Classical	15	10

Here are the **answers** for the AP Statistics question:

(a) Row and column variables:

- **Row variable:** Favorite type of music (Pop, Rock, Classical)
- **Column variable:** Own a musical instrument (Yes, No)

(b) Marginal distribution of students who own a musical instrument:

- This is the total count or proportion of students in the "Yes" and "No" columns, ignoring music preference.
  - Example (if totals from table):
    - Yes: 60%
    - No: 40%

(c) Conditional distribution of music preference for students who own an instrument:

- This shows the distribution of Pop, Rock, and Classical **among only those who own an instrument**.
  - Example (if totals from table):
    - Pop: 25/60 = 41.7%
    - Rock: 20/60 = 33.3%
    - Classical: 15/60 = 25%

Representing Two Categorical Variables

When we have two categorical variables, we want to examine whether they are *associated*. Association means that knowing the value of one variable helps us predict the value of the other. If the distribution of one variable differs across categories of the other, then the two variables are related.

1. Two-Way (Contingency) Tables

	A	A^c	Total
B	$P(A \cap B)$	$P(A^c \cap B)$	$P(B)$
B^c	$P(A \cap B^c)$	$P(A^c \cap B^c)$	$P(B^c)$
Total	$P(A)$	$P(A^c)$	1.00

- A **two-way table** shows the counts (or relative frequencies) for all combinations of two categorical variables.
- The totals for rows and columns are called **marginal totals**, which show the distribution of each variable separately.
- The counts in the inner cells show the **joint distribution**.
- Row percentages or column percentages show the **conditional distribution** of one variable given the other.

Types of Frequencies in a Two-Way Table

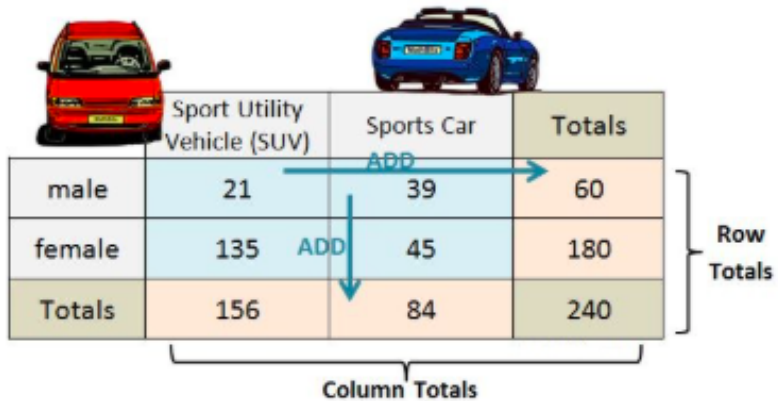
Joint Relative Frequency:

Proportion of individuals in a specific cell, out of the total sample. Formula: $\text{Joint RF} = \dfrac{\text{Cell Count}}{\text{Total}}$.

Marginal Relative Frequency:

Proportion of individuals in a row total or column total, out of the total sample. Formula:

$\text{Marginal RF} = \dfrac{\text{Row (or Column) Total}}{\text{Total}}$.



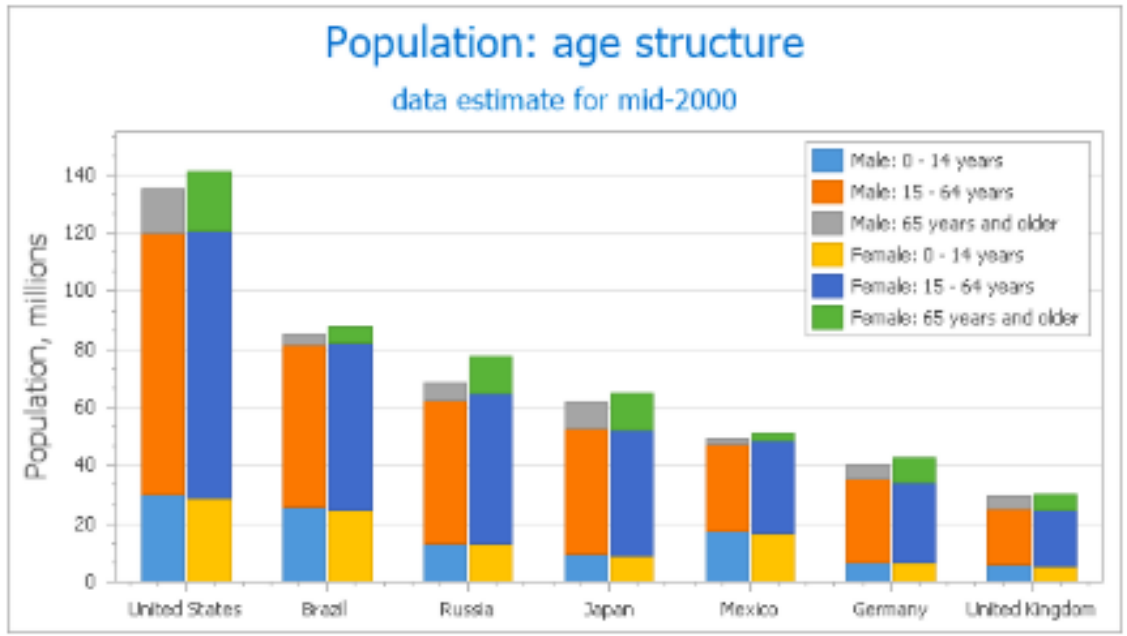
Conditional Relative Frequency:

Proportion of individuals in a category of one variable, given a specific value of the other variable. Formula:

$\text{Conditional RF} = \dfrac{\text{Cell Count}}{\text{Row or Column Total}}$.

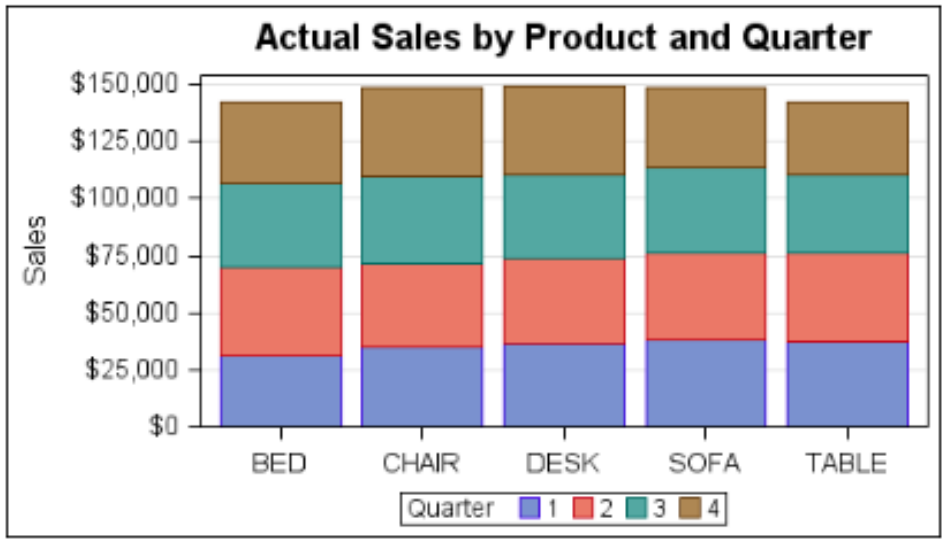
Graphical Representations

Side-by-Side Bar Graphs:



Each bar represents a category of one variable, and heights of bars within groups represent the second variable.

Segmented (Stacked) Bar Graphs:



Each bar represents 100% of a category, divided into proportional segments for the second variable. Useful for comparing relative distributions.

Mathematical Interpretation

- Conditional probability: $P(A \mid B) = \dfrac{\text{Count of (A and B)}}{\text{Total in category B}}$.
- If conditional distributions differ across groups, then the variables are associated.
- If the conditional distributions are the same, then the variables are independent (no association).

Example

A survey asked 100 students about *gender* (male/female) and whether they *prefer sports* (yes/no). The results are shown:

	Sports: Yes	**Sports: No**	**Total**
Male	30	10	40
Female	20	40	60
Total	50	50	100

Is there an association between gender and sports preference?

Compute conditional distributions:

- For males: $P(\text{Yes} \mid \text{Male}) = \frac{30}{40} = 0.75$.
- For females: $P(\text{Yes} \mid \text{Female}) = \frac{20}{60} = 0.33$.

Since $0.75 \neq 0.33$, the conditional proportions differ, which shows an **association** between gender and sports preference.

Example

A survey asked 100 students about *gender* (male/female) and whether they *prefer sports* (yes/no). The results are shown:

	Sports: Yes	**Sports: No**	**Total**
Male	30	10	40
Female	20	40	60
Total	50	50	100

Is there an association between gender and sports preference?

Researchers study exercise frequency (low/medium/high) among smokers vs. non-smokers. Each group is shown as a segmented bar that totals 100%.

How does the segmented bar graph show association?

	LOW	MEDIUM	HIGH
Smoker	60	30	10
Non-smoker	20	50	70

Researchers study exercise frequency (low/medium/high) among smokers vs. non-smokers. Each group is shown as a segmented bar that totals 100%.

How does the segmented bar graph show association?

	LOW	MEDIUM	HIGH
Smoker	60	30	10
Non-smoker	20	50	70

Suppose:

- Smokers: 60% low, 30% medium, 10% high.
- Non-smokers: 20% low, 50% medium, 30% high.

Conditional distributions are clearly different:

$P(\text{Low} \mid \text{Smoker}) = 0.60$ vs. $P(\text{Low} \mid \text{Non-Smoker}) = 0.20$.

Since the proportions are not equal, there is an association between smoking status and exercise frequency.

True or False

- Marginal totals are the totals for each category of one variable, ignoring the other variable.

True

True or False

- If the joint relative frequencies equal the product of the marginal relative frequencies, the variables are independent.

True

True or False

- It is possible for two categorical variables to have equal marginal distributions but still be associated.

TRUE

	Prefer A	Prefer B	Total
Group 1	40	10	50
Group 2	10	40	50
Total	50	50	100

Stretch Exercise (AP Statistics level+)

A school surveys 300 students about **whether they play sports** and **whether they prefer online or in-person classes**.

	Online	In-person	Total
Play sports	30	120	150
Don't play sports	90	60	150
Total	120	180	300

(a) Calculate the **conditional distribution** of class preference for:

- students who play sports
- students who don't play sports

(b) Based on your results, describe the **association** between playing sports and class preference.

(c) A student claims:

> "Playing sports causes students to prefer in-person classes."
> Is this justified? Explain.

(d) Suppose the data were collected separately for **Grade 11 and Grade 12**, and in both grades there is little to no association, but in the combined table there is a strong association.

- What phenomenon could explain this?
- Briefly explain how this is possible.

(e) Without doing calculations, explain how the conclusion might change if the table showed **relative frequencies instead of counts**.

(a)

- Play sports: Online **20%**, In-person **80%**
- Don't play: Online **60%**, In-person **40%**

(b)

Association exists (distributions differ → not independent).

(c)

Not justified — association ≠ causation (observational study).

(d)

Simpson's paradox — a lurking variable (grade) can reverse/alter association.

(e)

Relative frequencies don't change the conclusion, only make comparison clearer.

Simpson's Paradox example

Suppose a university compares **admission rates for men and women.**

Combined Data

Gender	Accepted	Applied	Acceptance Rate
Men	60	100	60%
Women	40	100	40%

Associative

But by Department

Department A (very competitive)

Gender	Accepted	Applied	Rate
Men	10	50	20%
Women	8	40	20%

Not Associative

Department B (less competitive)

Gender	Accepted	Applied	Rate
Men	50	50	100%
Women	32	60	53%

Associative

Chi Square Test

Chi-square test	卡方检验
Observed count	观察频数
Expected count	期望频数
Null hypothesis	零假设
Degrees of freedom	自由度

1. The Chi-square test evaluates whether there is a **relationship (association) between two categorical variables** or if the data **fits a theoretical distribution**.
2. It compares **observed counts** from the data to **expected counts** calculated under the assumption of **independence**.
3. The **Chi-square statistic** measures the overall difference between observed and expected counts, with larger values indicating stronger evidence against the null hypothesis.
4. The test uses **degrees of freedom** to determine the **p-value**, which tells us whether to reject the **null hypothesis** of independence.
5. Key conditions include **counts (not percentages), mutually exclusive categories, and expected counts generally ≥5** for valid results.

A high school surveyed 200 students about whether they **play a musical instrument** and whether they **prefer online or in-person classes**. The results are shown in the table below:

	Online	**In-person**	**Total**
Play instrument	40	60	100
Don't play	70	30	100
Total	110	90	200

(a) State the null and alternative hypotheses for a Chi-square test of independence.

(b) Calculate the expected counts for each cell.

(c) Compute the Chi-square test statistic: $\chi^2 = \sum \frac{(O-E)^2}{E}$.

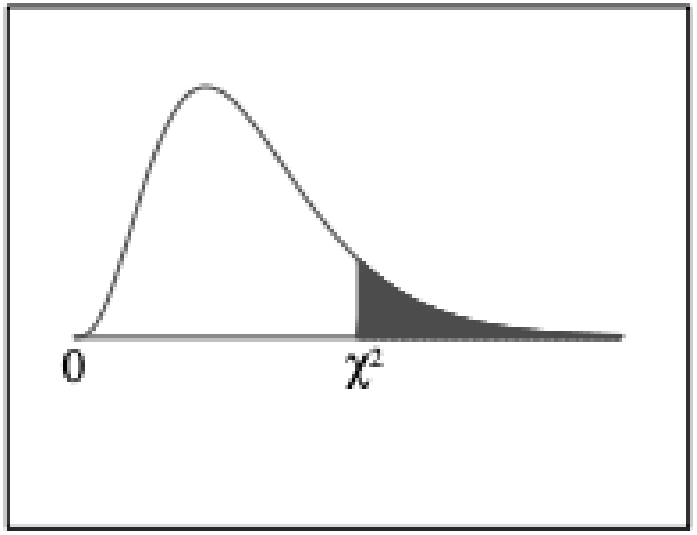
(d) With a significance level of 0.05, df = 1, determine whether to reject the null hypothesis.

(e) Interpret your conclusion in the context of the problem.

Percentage Points of the Chi-Square Distribution

Degrees of Freedom	Probability of a larger value of x^2								
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84	6.63
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99	9.21
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81	11.34
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49	13.28
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07	15.09
6	0.872	1.635	2.204	3.455	5.348	7.84	10.64	12.59	16.81
7	1.239	2.167	2.833	4.255	6.346	9.04	12.02	14.07	18.48
8	1.647	2.733	3.490	5.071	7.344	10.22	13.36	15.51	20.09
9	2.088	3.325	4.168	5.899	8.343	11.39	14.68	16.92	21.67
10	2.558	3.940	4.865	6.737	9.342	12.55	15.99	18.31	23.21
11	3.053	4.575	5.578	7.584	10.341	13.70	17.28	19.68	24.72
12	3.571	5.226	6.304	8.438	11.340	14.85	18.55	21.03	26.22
13	4.107	5.892	7.042	9.299	12.340	15.98	19.81	22.36	27.69
14	4.660	6.571	7.790	10.165	13.339	17.12	21.06	23.68	29.14
15	5.229	7.261	8.547	11.037	14.339	18.25	22.31	25.00	30.58
16	5.812	7.962	9.312	11.912	15.338	19.37	23.54	26.30	32.00
17	6.408	8.672	10.085	12.792	16.338	20.49	24.77	27.59	33.41
18	7.015	9.390	10.865	13.675	17.338	21.60	25.99	28.87	34.80
19	7.633	10.117	11.651	14.562	18.338	22.72	27.20	30.14	36.19
20	8.260	10.851	12.443	15.452	19.337	23.83	28.41	31.41	37.57
22	9.542	12.338	14.041	17.240	21.337	26.04	30.81	33.92	40.29
24	10.856	13.848	15.659	19.037	23.337	28.24	33.20	36.42	42.98
26	12.198	15.379	17.292	20.843	25.336	30.43	35.56	38.89	45.64
28	13.565	16.928	18.939	22.657	27.336	32.62	37.92	41.34	48.28
30	14.953	18.493	20.599	24.478	29.336	34.80	40.26	43.77	50.89
40	22.164	26.509	29.051	33.660	39.335	45.62	51.80	55.76	63.69
50	27.707	34.764	37.689	42.942	49.335	56.33	63.17	67.50	76.15
60	37.485	43.188	46.459	52.294	59.335	66.98	74.40	79.08	88.38

Table 3: Chi-Square Distribution Table.



The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$.

(a) H_0: Playing an instrument and class preference are independent.

H_a: They are associated.

(b) Expected counts:

- Play & Online: 55, Play & In-person: 45
- Don't Play & Online: 55, Don't Play & In-person: 45

(c) $\chi^2 = 18.18$

(d) df = 1, critical value 3.841 → $\chi^2 > 3.841$ → Reject H_0

(e) There is strong evidence of an association: students who play an instrument tend to prefer in-person classes, while those who don't play tend to prefer online.