

Statistics

| The Association Between Two Categorical Variables

Prof. Dr. Mostafa Elhosseini

<https://youtube.com/drmelhosseini>

| Today's Focus

- Relationship between two categorical variables
- Response Variable
- Explanatory Variable
- Contingency Tables (جدول التوافق)
- Cross-tabulation (الجداول التبادلية)
- Conditional proportions (النسب الاحتمالية المشروطة)
- Visual Tools (Side by Side – Stacked bar chart)
- Measuring the Strength : Difference of Proportions - Risk Ratio - Odds Ratio



Picture the Scenario: Smoking and Health



A 20-year study followed 1,314 women in the UK.



Each woman reported whether she smoked.



After 20 years, their survival status (alive/deceased) was recorded.



Surprisingly, 31% of nonsmokers died, compared to 24% of smokers. - confounding variables (المتغير المربك)

- How do we describe relationships between two categorical variables numerically?



Describing Relationships Between Two Variables



Example: Smoking and Survival

- Two categorical variables:
 - Smoking status: Smoker  / Non-smoker 
 - Survival status: Survived  / Deceased  (after 20 years)
-  Use descriptive statistics to summarize the association:
 - Contingency tables
 - Conditional proportions
 - Difference in proportions - Relative risk

But Not All Variables Are Categorical...

- We also analyze relationships between two quantitative variables using:
 - Scatterplots
 - Correlation coefficients
 - Regression analysis

 Examples:

-  Gasoline usage vs.  Air pollution
-  School funding per student vs.  SAT scores

Response Variable (Dependent)

- The outcome we are measuring or trying to explain.
- It is the variable that depends on or is influenced by changes in another variable — the explanatory (independent) variable.
- It is what you observe or record as an outcome.
- Often answers the question:

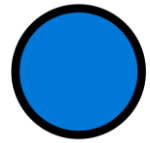
“What happens when we change the explanatory variable?”

 Examples of Response Variables:

 Survival Status (Alive or Dead)

 Test Score (Out of 100)

 Blood Pressure (in mmHg)



Response Variable (Dependent)

Discipline	Research Question	Response Variable (Dependent)
Medicine	Does a drug reduce blood pressure?	Blood pressure (mmHg)
Economics	Does inflation affect unemployment rate?	Unemployment rate (%)
Environmental Science	How does fertilizer use affect crop yield?	Crop yield (tons per hectare)
Marketing	Does ad exposure increase sales?	Product sales (units sold)
Engineering	Does design improvement reduce failure rate?	Failure rate (%)
Transportation	Does speed limit enforcement affect accident frequency?	Number of road accidents
Agriculture	Does irrigation method impact plant height?	Plant height (cm)

● Explanatory Variable (Independent)

- The explanatory variable is the variable you manipulate or categorize to see its effect on the response.
- هو المتغير الذي نعتقد أنه يفسر أو يؤثر أو يتسبب في تغيّر متغير آخر يُعرف بالمتغير التابع (Response).

Discipline	Research Question	Explanatory Variable
Public Health	Does smoking affect lung cancer rates?	Smoking status (Yes/No)
Education	Do more study hours lead to higher exam scores?	Hours of study
Agriculture	Does fertilizer type influence crop yield?	Fertilizer type
Medicine	Does drug dosage affect blood pressure?	Dosage level
Marketing	Do discounts increase product sales?	Discount percentage



Examples

Context	Explanatory Variable (Independent)	Response Variable (Dependent)
Online Education	Type of course platform (MOOC, LMS)	Student completion rate
Traffic Safety	Speed limit in urban areas	Number of road accidents
Mental Health	Hours of sleep per night	Reported stress level
Retail	Store layout (open vs. aisle format)	Average customer spending
Sports Science	Training frequency per week	Athlete endurance level
Social Media Marketing	Type of post (image, video, text)	User engagement (likes, comments, shares)
Human Resources	Work-from-home policy (Yes/No)	Employee productivity rating



Examples

Explanatory Variable

Smoking status (Yes / No)

Gasoline use (liters/day)

Number of study hours/week

Diet Type



Response Variable

Survival after 20 years

Carbon dioxide emissions

GPA

Cholesterol level

| Note:

 In some studies, there's no clear distinction between the explanatory and response variables.

Variable 1	Variable 2	Note
Smoking Status	Alcohol Status	No clear causal direction
Height	Weight	Both are continuous and related

 Focus: Understand how two variables are associated, not which one influences the other.



Association Between Two Variables

- An association exists if certain values of one variable are more likely to occur with specific values of another variable.
- يوجد ارتباط بين متغيرين إذا كانت قيم معينة لأحد المتغيرات تحدث بشكل أكثر احتمالاً مع قيم معينة للمتغير الآخر.

Variable 1	Variable 2	Association Insight
High School GPA	College GPA	High GPA → Likely higher college GPA
Marital Status	Happiness Level	Married people more likely to report being “very happy”

💡 Key Idea: An association does not imply **causation**, but it suggests a pattern worth exploring. - الارتباط لا يعني السببية



Contingency Tables (جدول التوافق)

- A fundamental tool to explore the association between two categorical variables.
- A contingency table is a tabular summary that:
 - Has rows for the categories of one variable
 - Has columns for the categories of the second variable
 - Each cell shows the frequency (count) of observations that fall in the combination of those two categories



Contingency Tables (جدول التوافق)



Key Terms:

- **Cell:** The intersection of a row and column → contains a count
- **Marginal Totals:** Row and column totals
- **Grand Total:** Sum of all cells

Food Type	Pesticide Status		Total
	Present	Not Present	
Organic	29	98	127
Conventional	19,485	7,086	26,571
Total	19,514	7,184	26,698

| Contingency Tables (تحليل الجداول التبادلية)

- Cross-tabulation (cross-tab or **contingency analysis**)
 - The process of taking a data file and finding the frequencies for the cells of a contingency table is referred to as cross-tabulation of the data

Food Type	Pesticide Status		Total
	Present	Not Present	
Organic	29	98	127
Conventional	19,485	7,086	26,571
Total	19,514	7,184	26,698

| Conditional proportions (النسب الاحتمالية المشروطة)

- Conditional proportions are percentages calculated within a specific group, meaning they are conditional on one category of a variable (such as a row or column in a contingency table).
-
- It answers: "What proportion of people survived among smokers?" — this is conditional on smoking status.

| Conditional proportions (النسب الاحتمالية المشروطة)

Food Type	Pesticide Status		Total
	Present	Not Present	
Organic	29	98	127
Conven.	19,485	7,086	26,571
Total	19,514	7,184	26,698

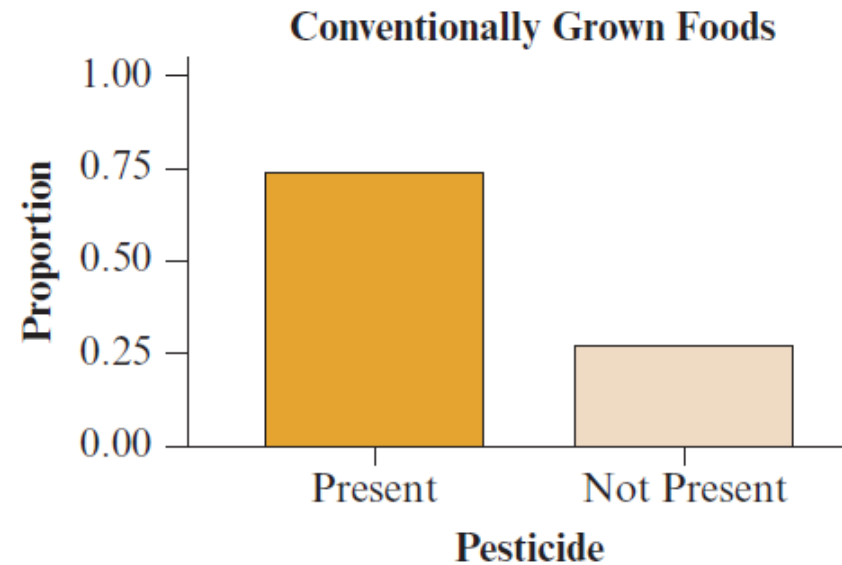
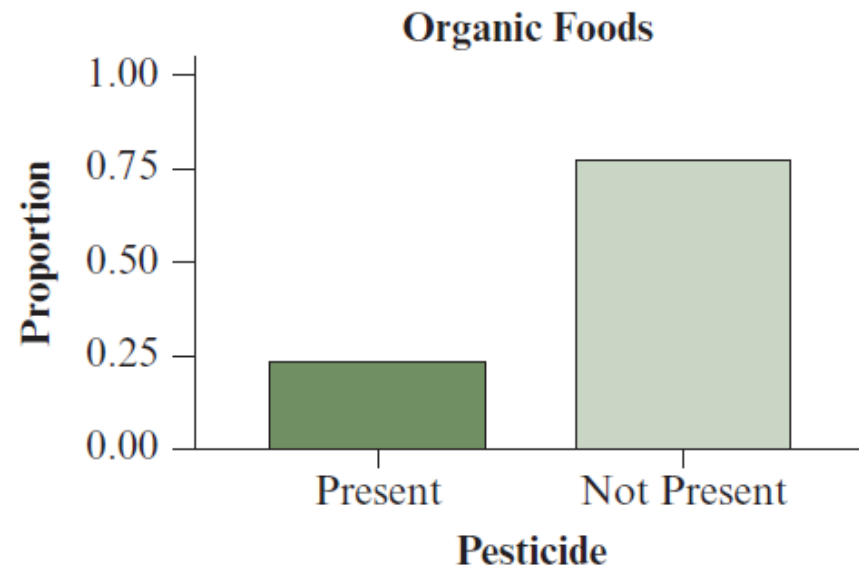
Food Type	Pesticide Status		Total	<i>n</i>
	Present	Not Present		
Organic	0.23	0.77	1.00	127
Conventional	0.73	0.27	1.00	26,571

| Marginal proportions (النسب الهامشية)

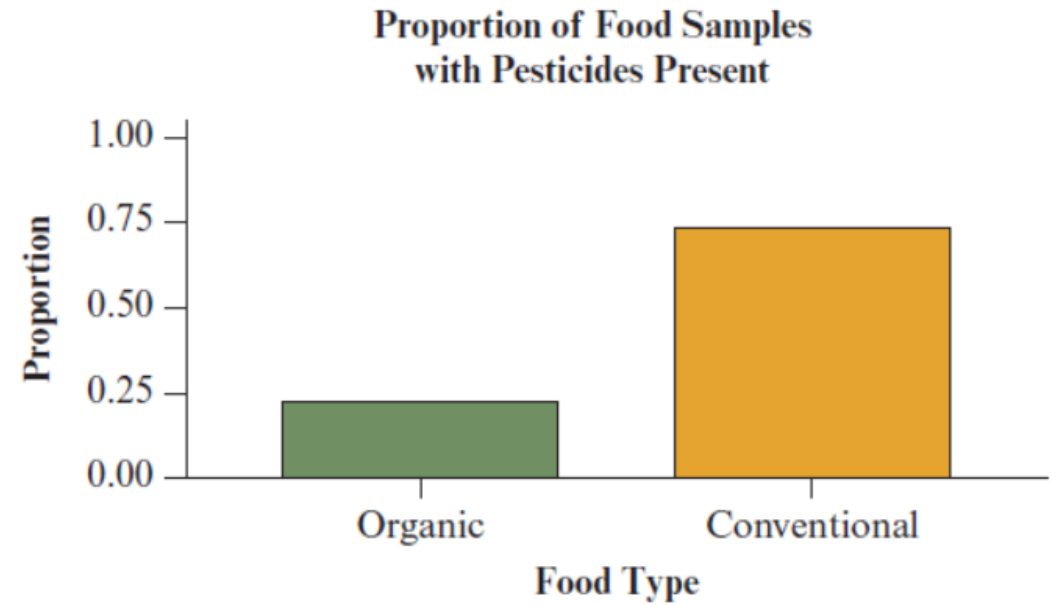
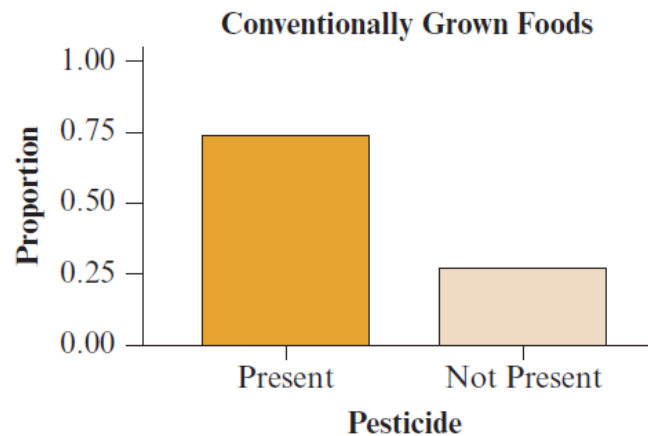
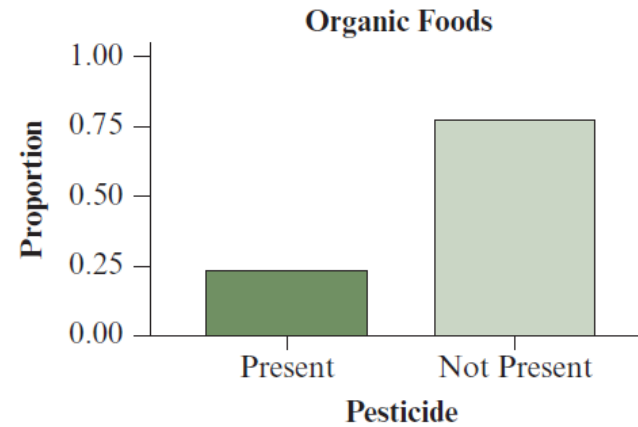
- The ratio of the total counts of a specific category to the overall total in a contingency table
 - the proportion of all sampled produce items with pesticide residues.
 - It is not found for a particular food type.
 - $19,514/26,698 = 0.731$.

Food Type	Pesticide Status		Total
	Present	Not Present	
Organic	29	98	127
Conven.	19,485	7,086	26,571
Total	19,514	7,184	26,698

| Visual Tools for Categorical Data Analysis



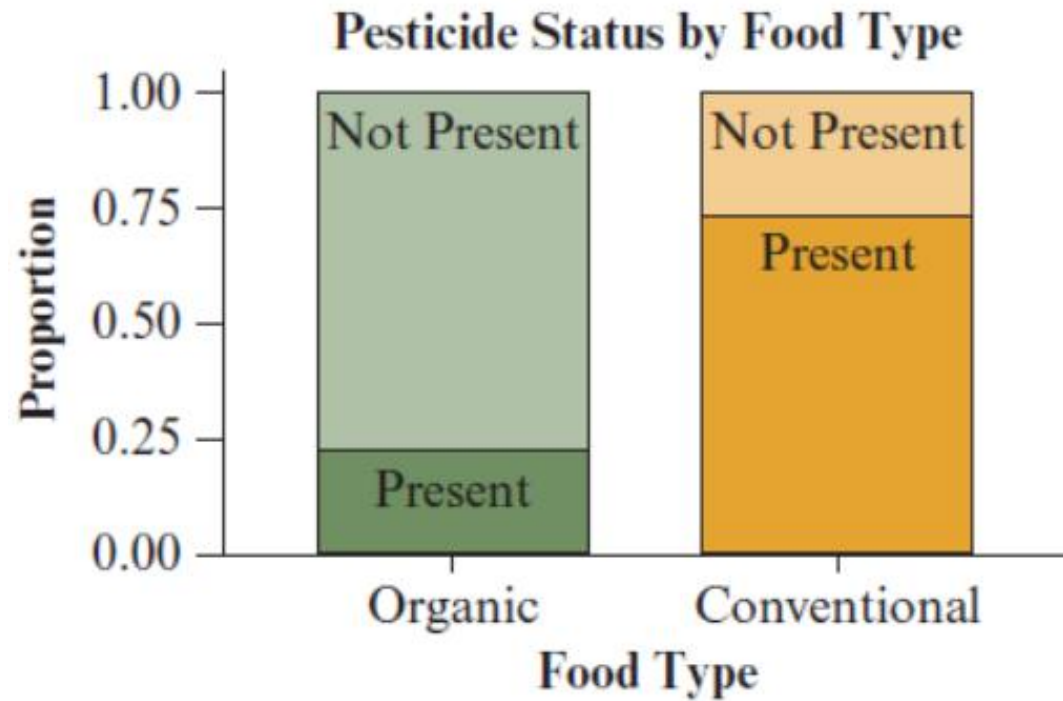
| Visual Tools for Categorical Data Analysis



- Single bar graph that shows the bars for the conditional proportions **side by side**

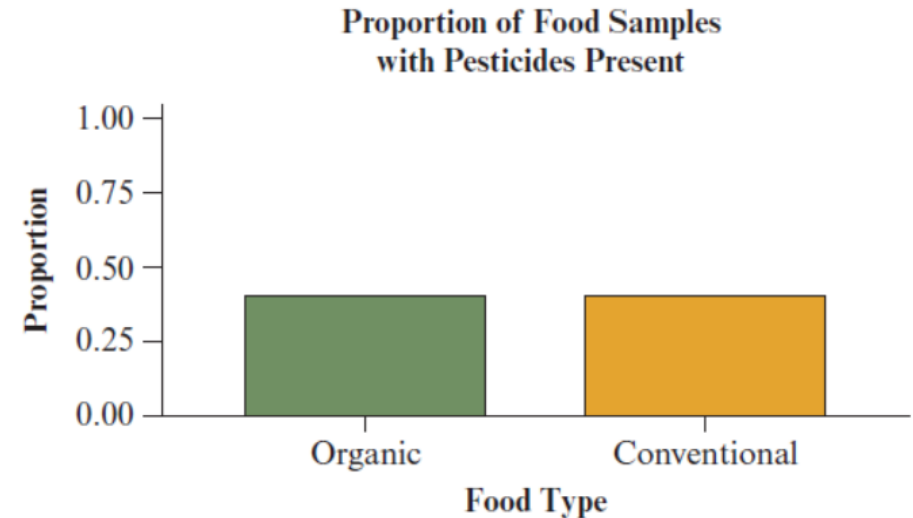
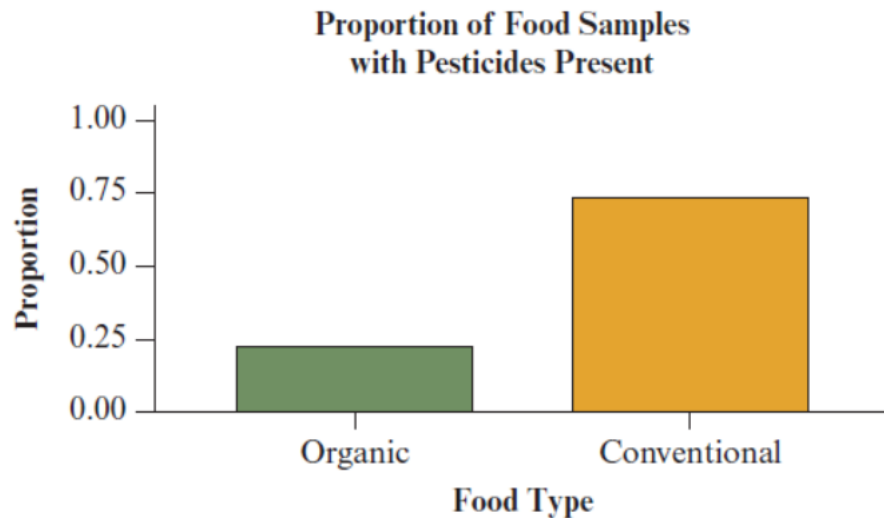
| Visual Tools for Categorical Data Analysis

- **Stacking** the proportion present and not present on top of each other



| Looking For an Association

- Studying the **conditional proportions helps** you judge whether there is an association between the variables
 - There would be no association if, instead, the proportion with pesticides present had been the same for each food type





Measuring the Strength of an Association

◆ 1. Difference of Proportions (فرق النسب)

- Compares two proportions by subtraction.
- Example:
 - 73% of conventional food samples had pesticide residue.
 - 23% of organic food samples had pesticide residue.
 - Difference = $73\% - 23\% = 50\%$
- **Interpretation:** Conventional samples had 50 percentage points more contamination.



Measuring the Strength of an Association

◆ 2. Risk Ratio (Relative Risk) **الخطر النسبي**

- The Risk Ratio (RR) — also called Relative Risk — is a measure that compares the probability (risk) of an outcome between two groups.

Food Type	Proportion with Pesticide Present
Organic	0.23
Conventional	0.73

- Risk in Conventional food = 0.73
- Risk in Organic food = 0.23



Measuring the Strength of an Association

$$\text{Risk Ratio (RR)} = \frac{\text{Risk in Conventional}}{\text{Risk in Organic}} = \frac{0.73}{0.23} \approx 3.17$$

◆ 2. Risk Ratio (Relative Risk)

- The risk of pesticide presence in conventional food is about 3.17 times higher than in organic food.
- A risk ratio of 3.17 means that the likelihood (risk) of pesticide residue being present in conventional food is 3.17 times greater than in organic food.
- نسبة الخطر البالغة 3.17 تعني أن احتمالية وجود بقايا مبيدات في الأغذية التقليدية أعلى بمقدار 3.17 مرة مقارنة بالأغذية العضوية.
- This clearly indicates a strong association between food type and pesticide presence.



Measuring the Strength of an Association

◆ 3. Odds Ratio (نسبة الأرجحية)

- The Odds Ratio (OR) is a statistical measure used to quantify the strength of association between two categorical variables, especially in case-control studies.

■ نسبة الأرجحية تشير إلى مدى احتمال حدوث حدث ما في مجموعة مقارنةً بمجموعة أخرى

- The odds of an event is the ratio:
$$\text{Odds} = \frac{\text{Probability of the event}}{\text{Probability of no event}}$$
- The Odds Ratio (OR) compares the odds of an outcome in one group to the odds in another group:

$$\text{Odds Ratio} = \frac{\text{Odds in Group A}}{\text{Odds in Group B}}$$



Measuring the Strength of an Association

◆ 3. Odds Ratio (نسبة الأرجحية)

Food Type	Pesticide Present	Not Present	Total
Organic	0.23	0.77	1.00
Conventional	0.73	0.27	1.00

- Odds for Organic = $0.23 / 0.77 \approx 0.2987$
- Odds for Conventional = $0.73 / 0.27 \approx 2.7037$
- Compute Odds Ratio OR = $2.7037 / 0.2987 = 9.05$
- The odds of finding pesticide residue in conventional food are about 9 times greater than in organic food.



Example

- A clinical study evaluates a new drug for treating an infection. Patients are randomly assigned to either the Drug Group or the Placebo Group, and their recovery status is tracked.

Group	Recovered	Not Recovered	Total
Drug	180	20	200
Placebo	150	50	200



Example

Group	Recovered	Not Recovered	Total
Drug	180	20	200
Placebo	150	50	200

◆ 1. Difference of Proportions

- Proportion recovered in Drug group:
- $p1 = 180/200 = 0.90$
- Proportion recovered in Placebo group:
- $p2 = 150/200 = 0.75$
- Difference: $0.90 - 0.75 = 0.15$

● Interpretation: Recovery rate is 15 percentage points higher in the Drug group.



Example

Group	Recovered	Not Recovered	Total
Drug	180	20	200
Placebo	150	50	200

◆ 2. Risk Ratio (Relative Risk)

$$■ RR = \frac{p1}{p2} = \frac{0.90}{0.75} = 1.20$$

● Interpretation: Patients on the drug are 1.2 times more likely to recover compared to those on the placebo..



Example

Group	Recovered	Not Recovered	Total
Drug	180	20	200
Placebo	150	50	200

◆ 3. Odds Ratio (OR)

- Odds of recovery in Drug group: $180/20 = 9$
- Odds of recovery in Placebo group: $150/50 = 3$
- $OR = 9/3 = 3$

● Interpretation: The odds of recovery are 3 times higher in the Drug group than in the Placebo group.

📌 الاحتمالات للشفاء أعلى بثلاث مرات في مجموعة الدواء مقارنةً بمجموعة العلاج الوهمي.