Session 07-06 - Contingency Tables

Section 07: Probability & Statistics

Dr. Nikolai Heinrichs & Dr. Tobias Vlćek

Entry Quiz - 10 Minutes

Quick Review from Session 07-05

1. State Bayes' Theorem.
2. A test has sensitivity 80% and specificity 90%. If prevalence is 10%, calculate PPV.
3. What's the difference between sensitivity and PPV?
4. If PPV is low but NPV is high, what does this tell us about the test?

Learning Objectives

What You'll Master Today

- Construct contingency tables from word problems
- Complete tables with missing values
- Read probabilities from tables: marginal, joint, conditional
- Test independence using table values
- Connect tables to Bayes' theorem

. . .

! Important

Contingency tables are a key exam format - expect at least one problem!

Part A: Table Structure

Two-Way Contingency Table

A contingency table shows the joint distribution of two categorical variables.

. . .

	B	$\bar{B}$	Total
A	n_{AB}	$n_{A\bar{B}}$	n_A
$\bar{A}$	$n_{\bar{A}B}$	$n_{\bar{A}\bar{B}}$	$n_{\bar{A}}$

	B	$\bar{B}$	Total
Total	n_B	$n_{\bar{B}}$	n

. . .

- Cells: Joint frequencies (both conditions)
- Row totals: Marginal frequencies for A
- Column totals: Marginal frequencies for B

Reading Probabilities from Tables

Type	Formula	Location in Table
Marginal	$P(A)$	Row total / Grand total
Joint	$P(A \cap B)$	Cell / Grand total
Conditional	$P(A \parallel B)$	Cell / Column total

Example: Market Research

Survey of 500 customers about product preference and age:

	Age < 30	Age ≥ 30	Total
Prefers A	120	80	200
Prefers B	130	170	300
Total	250	250	500

. . .

Calculate:

- $P(\text{Prefers A}) = \frac{200}{500} = 0.40$
- $P(\text{Age} < 30 \cap \text{Prefers A}) = \frac{120}{500} = 0.24$
- $P(\text{Prefers A} \mid \text{Age} < 30) = \frac{120}{250} = 0.48$

Part B: Constructing Tables from Word Problems

Strategy for Word Problems

Step-by-Step Approach

1. Identify the two variables and their categories
2. Create empty table with row/column labels
3. Fill in given values (often percentages → convert to counts)
4. Use relationships to complete missing cells
5. Verify: Row and column totals must match

Example: Building a Table

In a city of 10,000 residents:

- 40% are employed
- 70% are adults (age ≥ 18)
- 35% are employed adults

. . .

Construct the contingency table.

. . .

	Adult	Minor	Total
Employed	3500	?	4000
Not Employed	?	?	6000
Total	7000	3000	10000

Completing the Table

	Adult	Minor	Total
Employed	3500	500	4000
Not Employed	3500	2500	6000
Total	7000	3000	10000

. . .

Now we can answer questions like:

- $P(\text{Employed} \mid \text{Minor}) = \frac{500}{3000} = \frac{1}{6} \approx 0.167$
- $P(\text{Adult} \mid \text{Employed}) = \frac{3500}{4000} = 0.875$

Exam-Style Problem

A company surveyed 200 customers:

- 60% are satisfied with the product
- 45% are repeat customers
- Of the satisfied customers, 60% are repeat customers

. . .

Build the table:

Step 1: Fill in what we know directly

	Repeat	New	Total
Satisfied	?	?	120

	Repeat	New	Total
Not Satisfied	?	?	80
Total	90	110	200

Solution Continued

Step 2: Use "Of satisfied, 60% are repeat"

$P(\text{Repeat} \mid \text{Satisfied}) = 0.60$, so $120 \times 0.60 = 72$ repeat AND satisfied

	Repeat	New	Total
Satisfied	72	48	120
Not Satisfied	18	62	80
Total	90	110	200

. . .

Verify: All rows and columns sum correctly ✓

Break - 10 Minutes

Part C: Independence Testing

When Are Variables Independent?

! Independence in Tables

Variables A and B are independent if and only if for all cells:

$$P(A \cap B) = P(A) \cdot P(B)$$

Or equivalently: $\frac{\text{Cell count}}{\text{Total}} = \frac{\text{Row total}}{\text{Total}} \times \frac{\text{Column total}}{\text{Total}}$

Testing Independence: Example

From our customer survey:

	Repeat	New	Total
Satisfied	72	48	120
Not Satisfied	18	62	80
Total	90	110	200

. . .

Test independence for (Satisfied, Repeat):

- Expected if independent: $\frac{120}{200} \times \frac{90}{200} \times 200 = 0.60 \times 0.45 \times 200 = 54$
- Observed: 72

. . .

$72 \neq 54$, so satisfaction and repeat status are NOT independent.

Interpretation

The data suggests:

- Satisfied customers are MORE likely to be repeat customers
- $P(\text{Repeat} \mid \text{Satisfied}) = \frac{72}{120} = 0.60$
- $P(\text{Repeat} \mid \text{Not Satisfied}) = \frac{18}{80} = 0.225$

. . .

> i Note
>
> Satisfied customers are about 2.7 times more likely to be repeat customers!

Part D: Connecting to Bayes' Theorem

Tables and Bayes

The contingency table method from Session 07-05 is actually using this technique!

. . .

Medical testing example:

	Disease	No Disease	Total
Test +	TP	FP	All +
Test –	FN	TN	All –
Total	Diseased	Healthy	Population

. . .

- PPV = $P(D \mid +) = \frac{\text{TP}}{\text{All} +}$
- This is Bayes' theorem applied to the table!

Converting Between Approaches

Given: Sensitivity = 90%, Specificity = 95%, Prevalence = 2%

For 10,000 people:

	Disease (200)	No Disease (9800)	Total
Test +	180	490	670

	Disease (200)	No Disease (9800)	Total
Test –	20	9310	9330

. . .

Direct calculations: - PPV = $\frac{180}{670} \approx 0.269$ - NPV = $\frac{9310}{9330} \approx 0.998$

Guided Practice - 25 Minutes

Practice Problem 1

A survey of 400 employees found:

- 55% work full-time
- 40% have a graduate degree
- 25% work full-time AND have a graduate degree

Tasks: a) Construct the contingency table b) Find $P(\text{Grad degree} \mid \text{Full-time})$ c) Find $P(\text{Full-time} \mid \text{Grad degree})$ d) Are full-time status and graduate degree independent?

Practice Problem 2 (2025 Exam Style)

A company produces items at two factories. Quality control data:

- Factory A produces 3000 items, 5% defective
- Factory B produces 2000 items, 8% defective

Tasks: a) Construct a contingency table b) An item is randomly selected and found defective. What's the probability it came from Factory A? c) What percentage of all items are defective?

Wrap-Up & Key Takeaways

Today's Essential Concepts

- Table structure: Cells (joint), margins (marginal)
- Reading probabilities: Marginal, joint, conditional
- Building tables: Use given percentages and relationships
- Independence test: Expected = row% × col% × total
- Connection to Bayes: Tables provide visual Bayes calculations

Next Session Preview

Coming Up: Binomial Distribution

- Discrete probability distributions
- Binomial formula: $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$
- "Exactly k", "at most k", "at least k"
- Expected value and variance

...

💡 Homework

Complete Tasks 07-06 - practice building and reading contingency tables!