Biostatistics – Spring 2026
Lecture 02: Categorical Data in Biostatistics: Contingency Tables and Chi-square Testing

Dr. Zaid T. Al-Khaledi
Department of Statistics and Informatics
University of Mosul

February 11, 2026

Introduction

In many biomedical, clinical, and public health studies, the collected data are not continuous measurements such as height, weight, or temperature. Instead, the data are **counts** (frequencies), such as the number of males and females, the number of patients who recovered, or the number who developed a disease. This lecture explains how to analyze this type of data.

By the end of this lecture, you should be able to:
(i) build 2×2 and $r \times c$ contingency tables,
(ii) compute observed and expected frequencies,
(iii) perform and interpret the Chi-square test of independence.

1. What Is Categorical Data?

In categorical data, each observation belongs to one category, and the categories do not represent numerical magnitude.

For example, if we say "male" or "female", or "infected" or "not infected", we are classifying the person. Even if we code male = 1 and female = 2, these numbers are only labels.

Binary variables

A binary (dichotomous) variable has exactly two categories: yes/no, present/absent, positive/negative. Examples include disease status, sex, smoking status, and test result.

Multi-category variables

A multi-category categorical variable has more than two categories, such as: blood group (A, B, AB, O), education level (primary, secondary, ...), or disease severity (mild, moderate, severe).

For all categorical variables, categories must be: **mutually exclusive** (one observation cannot be in two categories), and **collectively exhaustive** (every observation must fit somewhere).

2. Why Contingency Tables?

When we study two categorical variables together, we summarize the joint information using a **contingency table** (two-dimensional contingency table). The table entries are frequencies (counts) in each combination of categories.

This is exactly the type of data where the Chi-square test is commonly used, especially in: experimental studies, vaccine studies, clinical trials, and many medical association questions.

3. The 2×2 Contingency Table

The 2×2 table is used when both variables have two categories. A common notation is:

	B_1	B_2	Row total
A_1	a	b	$a+b$
A_2	c	d	$c+d$
Column total	$a+c$	$b+d$	$n=a+b+c+d$

Each of a, b, c, d is an **observed frequency**. The row totals and column totals are called **marginal totals**.

Example 1: Hypertension and Stroke

The following table studies whether stroke status is related to hypertension status:

	Hypertension	No Hypertension	Total
Stroke	40	160	200
No Stroke	15	785	800
Total	55	945	1000

How to read the table:
The number 40 means "40 individuals had both stroke and hypertension."
The number 160 means "160 individuals had stroke but did not have hypertension."
The row total 200 means "200 individuals had stroke."
The column total 55 means "55 individuals had hypertension."
The grand total is $n = 1000$.
At this point, we can describe. For example:

$$\text{Among stroke patients, the proportion with hypertension} = \frac{40}{200} = 0.20 = 20\%.$$

$$\text{Among non-stroke individuals, the proportion with hypertension} = \frac{15}{800} = 0.01875 \approx 1.9\%.$$

This looks very different, but we still need a formal test to decide if the difference is too large to be explained by chance. That formal test is the Chi-square test of independence.

4. Observed Frequencies vs Expected Frequencies

The observed frequency in row i and column j is denoted by O_{ij}.

The **expected frequency** answers a key question:

> If the two variables are independent, what frequency should we expect in each cell?

If row totals and column totals stay the same, independence implies:

$$E_{ij} = \frac{(\text{Row total}_i)(\text{Column total}_j)}{n}.$$

Expected frequencies for Example 1

We compute expected frequencies for all four cells.

Cell (Stroke, Hypertension):

$$E_{11} = \frac{(200)(55)}{1000} = 11.$$

Cell (Stroke, No Hypertension):

$$E_{12} = \frac{(200)(945)}{1000} = 189.$$

Cell (No Stroke, Hypertension):

$$E_{21} = \frac{(800)(55)}{1000} = 44.$$

Cell (No Stroke, No Hypertension):

$$E_{22} = \frac{(800)(945)}{1000} = 756.$$

Now compare observed vs expected:

Cell	Observed O	Expected E (if independent)
Stroke & Hypertension	40	11
Stroke & No Hypertension	160	189
No Stroke & Hypertension	15	44
No Stroke & No Hypertension	785	756

The largest difference is in the first cell: observed 40 versus expected 11. This is exactly the idea behind the Chi-square test: quantify these differences.

5. $r \times c$ Contingency Tables

If one variable has r categories and the other has c categories, we use an $r \times c$ table. The meaning is the same: each cell is a frequency for the combination of categories.

Example 2: Diet Type and Cancer

This example studies whether **diet type** is related to **cancer incidence**. Both variables are categorical, so a contingency table and a Chi-square test are appropriate.

Before looking at the numbers, note the meaning of the diet abbreviations used in the table: HF = High Fat, LF = Low Fat, F = Fiber, NoF = No Fiber. Thus, HF-NoF means a high-fat diet without fiber, and LF-F means a low-fat diet with fiber.

The study provides the following observed data:

	HF-NoF	HF-F	LF-NoF	LF-F	Total
Cancer	27	20	19	14	80
No Cancer	3	10	11	16	40
Total	30	30	30	30	120

From this table, we extract five important pieces of information:

1. The total sample size is $n = 120$.
2. Cancer status has two categories: Cancer and No Cancer.
3. Diet type has four categories: HF-NoF, HF-F, LF-NoF, and LF-F.
4. Each diet group contains exactly 30 individuals.
5. Among the 80 cancer cases, the counts differ across diet types (27, 20, 19, 14).

At this stage, we only describe the data. For example, among individuals with cancer, the largest number (27) appears in the high-fat, no-fiber group, while the smallest number (14) appears in the low-fat, fiber group. However, descriptive differences alone do not tell us whether diet type and cancer are statistically related.

Expected frequencies under independence

To apply the Chi-square test, we compute expected frequencies assuming that diet type and cancer status are independent. The expected frequency in any cell is given by:

$$E_{ij} = \frac{(\text{Row total}_i)(\text{Column total}_j)}{n}.$$

Because each diet column has the same total (30), the expected frequencies are particularly simple here.

For individuals with cancer:

$$E = \frac{(80)(30)}{120} = 20 \quad \text{for each diet category.}$$

For individuals without cancer:

$$E = \frac{(40)(30)}{120} = 10 \quad \text{for each diet category.}$$

This means that, if diet type and cancer were independent, we would expect 20 cancer cases and 10 non-cancer cases in each diet group.

Comparing these expected values with the observed counts (such as 27 observed versus 20 expected in the HF-NoF group) suggests deviations from independence. The Chi-square test formally measures whether these deviations are large enough to conclude that diet type and cancer incidence are associated.

6. Chi-square Test of Independence

Now we move to the formal test.

Purpose

The Chi-square test of independence checks whether two categorical variables are independent or associated. It is widely used in medicine and biology, for example: smoking and lung cancer, vaccination and infection, antibiotic use and post-surgery infection, diet and health outcomes.

Hypotheses

For two categorical variables A and B:

$$H_0: \text{ The variables are independent.}$$
$$H_1: \text{ The variables are not independent (associated).}$$

Test statistic

Suppose the table has r rows and c columns. Let O_{ij} be the observed frequency in cell (i, j), and E_{ij} be the expected frequency under H_0.

The Chi-square statistic is:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

This statistic is always nonnegative. Large values mean "observed is far from expected", which suggests association.

Degrees of freedom

The degrees of freedom are:

$$\text{df} = (r-1)(c-1).$$

So, for a 2×2 table, $\text{df} = (2-1)(2-1) = 1$.

Assumptions

We need the following conditions for the Chi-square test to be valid:

1. The sample should be taken randomly.
2. Observations should be independent (one person contributes to one cell only).
3. Categories must be well-defined (each observation belongs to one category).
4. Expected frequencies should not be too small. A common rule: most $E_{ij} \geq 5$ (especially in 2×2 tables).

Decision rule and interpretation

After computing χ^2, we compare it with a Chi-square distribution with df $= (r-1)(c-1)$.

There are two equivalent ways:

(1) Using a critical value: Choose a significance level α (for example 0.05). Reject H_0 if

$$\chi^2_{\text{calc}} \geq \chi^2_{\alpha,\ \text{df}}.$$

The critical value is in the **right tail** of the Chi-square distribution.

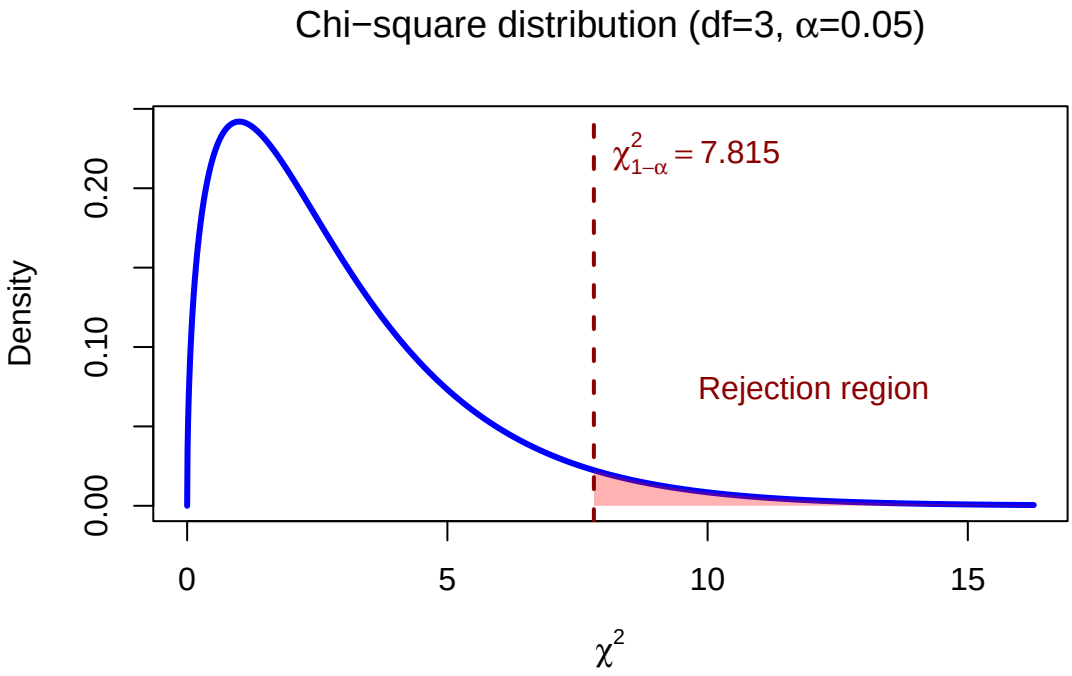
(2) Using a p-value: Reject H_0 if $p \leq \alpha$.

Interpretation must be written in clear words. If we reject H_0, we say:

> There is evidence of an association between the two categorical variables.

If we do not reject H_0, we say:

> There is not enough evidence to conclude an association.



7. Worked Chi-square Computation (Example 1)

We now compute the Chi-square statistic for the hypertension–stroke table.

Observed table:

	HTN	No HTN
Stroke	40	160
No Stroke	15	785

Expected table (computed earlier):

	HTN	No HTN
Stroke	11	189
No Stroke	44	756

Now compute each contribution $\frac{(O-E)^2}{E}$:
Cell (Stroke, HTN):

$$\frac{(40-11)^2}{11} = \frac{29^2}{11} = \frac{841}{11} \approx 76.45.$$

Cell (Stroke, No HTN):

$$\frac{(160-189)^2}{189} = \frac{(-29)^2}{189} = \frac{841}{189} \approx 4.45.$$

Cell (No Stroke, HTN):

$$\frac{(15-44)^2}{44} = \frac{(-29)^2}{44} = \frac{841}{44} \approx 19.11.$$

Cell (No Stroke, No HTN):

$$\frac{(785-756)^2}{756} = \frac{29^2}{756} = \frac{841}{756} \approx 1.11.$$

Sum:

$$\chi^2 \approx 76.45 + 4.45 + 19.11 + 1.11 = 101.12.$$

Degrees of freedom:

$$\text{df} = (2-1)(2-1) = 1.$$

This is a very large Chi-square value for df = 1. At significance level $\alpha = 0.05$, the critical value is

$$\chi^2_{0.95,\,1} = 3.84.$$

Since the calculated value

$$\chi^2_{\text{calc}} = 101.12$$

is much larger than the critical value, it falls in the rejection region.
Therefore, we reject the null hypothesis of independence.

> There is strong evidence of an association between hypertension status and stroke status.

Important wording note: Association does not automatically mean causation. To claim causation, we need proper study design and more information.

8. Short Worked Chi-square Computation (Example 2)

We now compute χ^2 for the diet–cancer table. Expected frequencies are 20 for each cancer cell and 10 for each no-cancer cell.
Compute contributions:
For Cancer row:

$$\frac{(27-20)^2}{20} = \frac{49}{20} = 2.45, \quad \frac{(20-20)^2}{20} = 0, \quad \frac{(19-20)^2}{20} = \frac{1}{20} = 0.05, \quad \frac{(14-20)^2}{20} = \frac{36}{20} = 1.80.$$

For No Cancer row:

$$\frac{(3-10)^2}{10} = \frac{49}{10} = 4.90, \quad \frac{(10-10)^2}{10} = 0, \quad \frac{(11-10)^2}{10} = \frac{1}{10} = 0.10, \quad \frac{(16-10)^2}{10} = \frac{36}{10} = 3.60.$$

Sum:

$$\chi^2 = 2.45 + 0 + 0.05 + 1.80 + 4.90 + 0 + 0.10 + 3.60 = 12.90.$$

Degrees of freedom:

$$\text{df} = (2-1)(4-1) = 3.$$

To conclude formally, we compare $\chi^2_{\text{calc}} = 12.90$ with the Chi-square distribution for df $= 3$. At significance level $\alpha = 0.05$, the critical value is $\chi^2_{0.95,\,3} = 7.815$. Since $12.90 > 7.815$, we reject H_0.

There is evidence of an association between diet type and cancer incidence.

9. Homework Example

Use the following table (Sex vs Type of Pulmonary TB) and test independence at $\alpha = 0.05$:

	Male	Female	Total
Respiratory TB	3534	1319	4853
Non-respiratory TB	270	250	520
Total	3804	1569	5373

Your tasks:
(1) Write H_0 and H_1.
(2) Compute all expected frequencies.
(3) Compute χ^2.
(4) Compute df and make a decision at $\alpha = 0.05$.
(5) Write one correct interpretation sentence.