STS3016: Categorical Data Analysis

2. Analyzing Contingency Tables

Fall 2025

Last Class

1. Introduction to CDA
   - Categorical Response Data
   - Probability Distributions
   - Statistical Inference for a Proportion
   - Statistical Inference for Discrete Data

Contingency Tables

Table 2.1 Cross-classification of belief in afterlife by gender.

	Belief in Afterlife		
Gender	Yes	No or Undecided	Total
Females	1230	357	1587
Males	859	413	1272
Total	2089	770	2859

Does an association exist between gender and belief in an afterlife?

Is one gender more likely than the other to believe in an afterlife?

Is belief in an afterlife plausibly independent of gender?

Contingency table displays counts of outcomes in the *cells*.

e.g. X, Y have r, c categories.

Probability Structure for Contingency Tables

Table 2.1 Cross-classification of belief in afterlife by gender.

	Belief in Afterlife		
Gender	Yes	No or Undecided	Total
Females	1230	357	1587
Males	859	413	1272
Total	2089	770	2859

Joint probabilities:

Marginal probabilities:

Conditional probabilities:

Sensitivity and Specificity

Let X (1 = diseased, 2 = not diseased), and Y (1 = positive, 2 = negative).

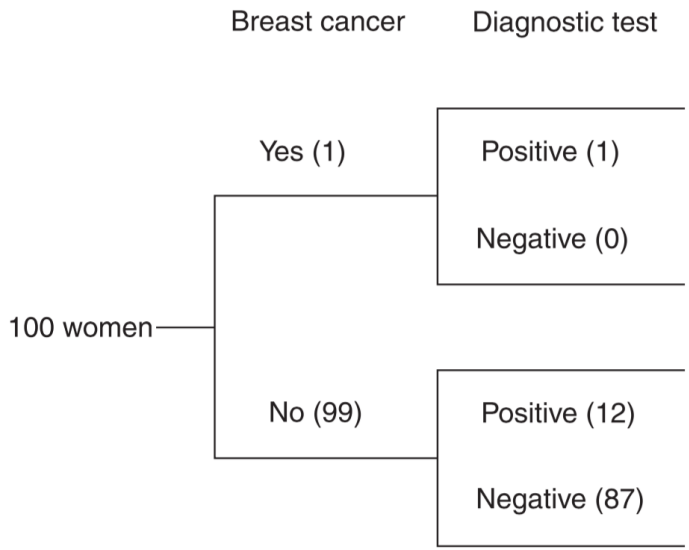
It is known that $P(X = 1) = 0.01$ and

Sensitivity $= 0.86$

Specificity $= 0.88$

Positive predictive value (PPV)

Sensitivity and Specificity



Sensitivity

Specificity

Positive predictive value (PPV)

Statistical Independence of Categorical Variables

Two categorical variables, X and Y are **statistically independent** if

Using conditional distributions

Homogeneity

Binomial and Multinomial Sampling

Table 2.1 Cross-classification of belief in afterlife by gender.

	Belief in Afterlife		
Gender	Yes	No or Undecided	Total
Females	1230	357	1587
Males	859	413	1272
Total	2089	770	2859

It is common to assume that cell counts in contingency tables have

Comparing Proportions in 2×2 Contingency Tables

Table 2.2 Cross-classification of aspirin use and myocardial infarction (MI).

	Myocardial Infarction		
Group	Yes	No	Total
Placebo	189	10,845	11,034
Aspirin	104	10,933	11,037

Let π_1 and π_2 be the probability of success in row 1 and row 2.

Difference of proportions

Comparing Proportions in 2×2 Contingency Tables

Table 2.2 Cross-classification of aspirin use and myocardial infarction (MI).

	Myocardial Infarction		
Group	Yes	No	Total
Placebo	189	10,845	11,034
Aspirin	104	10,933	11,037

Example:

Comparing Proportions in 2×2 Contingency Tables

Table 2.2 Cross-classification of aspirin use and myocardial infarction (MI).

	Myocardial Infarction		
Group	Yes	No	Total
Placebo	189	10,845	11,034
Aspirin	104	10,933	11,037

A difference between two proportions is more important

when both proportions are near

For 2×2 tables, the **ratio of proportions** is often called the

The Odds Ratio

For a probability of success π, the **odds** of success

The **odds ratio**

Properties of the Odds Ratio

Table 2.2 Cross-classification of aspirin use and myocardial infarction (MI).

	Myocardial Infarction		
Group	Yes	No	Total
Placebo	189	10,845	11,034
Aspirin	104	10,933	11,037

Properties of the Odds Ratio

Table 2.2 Cross-classification of aspirin use and myocardial infarction (MI).

	Myocardial Infarction		
Group	Yes	No	Total
Placebo	189	10,845	11,034
Aspirin	104	10,933	11,037

Inference for the Odds Ratios

Relationship Between Odds Ratio and Relative Risk

The Odds Ratio Applies in Case-Control Studies

Table 2.3 Cross-classification of smoking by lung cancer.

	Lung Cancer	
Smoker	Cases	Controls
Yes	688	650
No	21	59
Total	709	709

Types of Studies: Observational vs. Experimental

Chi-Squared Tests of Independence

Pearson Chi-Squared Statistic for testing H_0

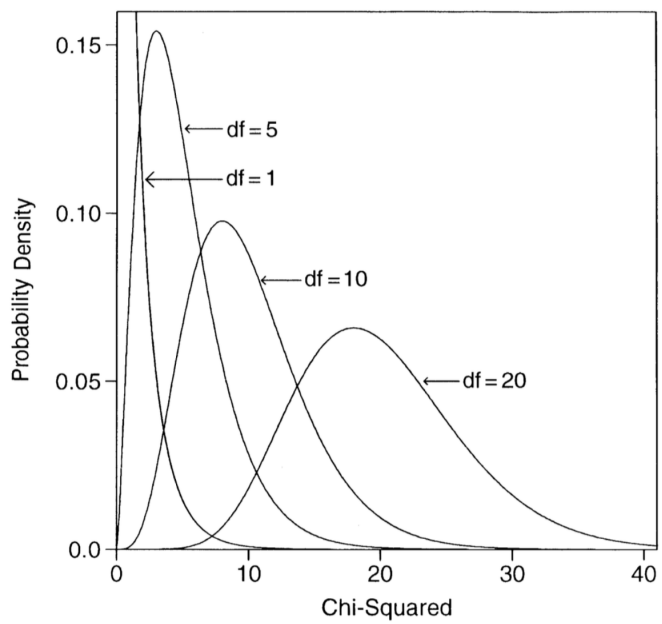


Figure 2.2 Examples of chi-squared distributions.

Likelihood-Ratio Statistic

Testing Independence in Two-Way Tables

Table 2.4 Political party identification by gender, with estimated expected frequencies for independence in parentheses.

	Political Party Identification			
Gender	Democrat	Republican	Independent	Total
Female	495 (456.9)	272 (297.4)	590 (602.6)	1357
Male	330 (368.1)	265 (239.6)	498 (485.4)	1093
Total	825	1088	2450	

Testing Independence in Two-Way Tables

Table 2.4 Political party identification by gender, with estimated expected frequencies for independence in parentheses.

	Political Party Identification			
Gender	Democrat	Republican	Independent	Total
Female	495 (456.9)	272 (297.4)	590 (602.6)	1357
Male	330 (368.1)	265 (239.6)	498 (485.4)	1093
Total	825	1088	2450	

Residuals for Cells in a Contingency Table

Table 2.5 Observed frequencies for political party identification and gender, with standardized residuals in parentheses for test of independence.

	Political Party Identification		
Gender	Democrat	Republican	Independent
Female	495 (3.27)	272 (−2.50)	590 (−1.03)
Male	330 (−3.27)	265 (2.50)	498 (1.03)

Testing Independence for Ordinal Variables

Table 2.6 Infant malformation and mother's alcohol consumption.

Alcohol Consumption	Malformation		Total	Percentage Present	Standardized Residual
	Absent	Present			
0	17,066	48	17,114	0.28	−0.18
< 1	14,464	38	14,502	0.26	−0.71
1–2	788	5	793	0.63	1.84
3–5	126	1	127	0.79	1.06
≥ 6	37	1	38	2.63	2.71

Linear Trend Alternative to Independence

Association in Three-Way Tables

Table 2.9 Death penalty verdict by defendant's race and victims' race.

Victims' Race	Defendant's Race	Death Penalty		Percentage
		Yes	No	Yes
White	White	53	414	11.3
	Black	11	37	22.9
Black	White	0	16	0.0
	Black	4	139	2.8
Total	White	53	430	11.0
	Black	15	176	7.9

Partial Tables

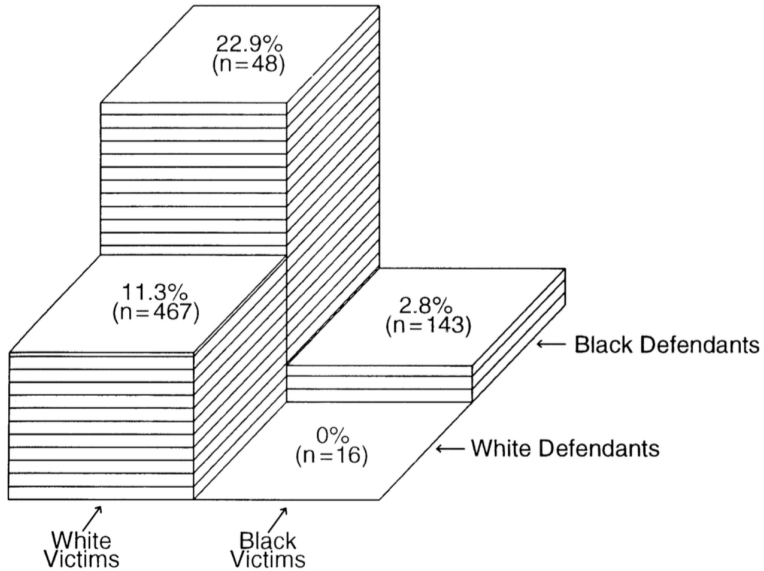
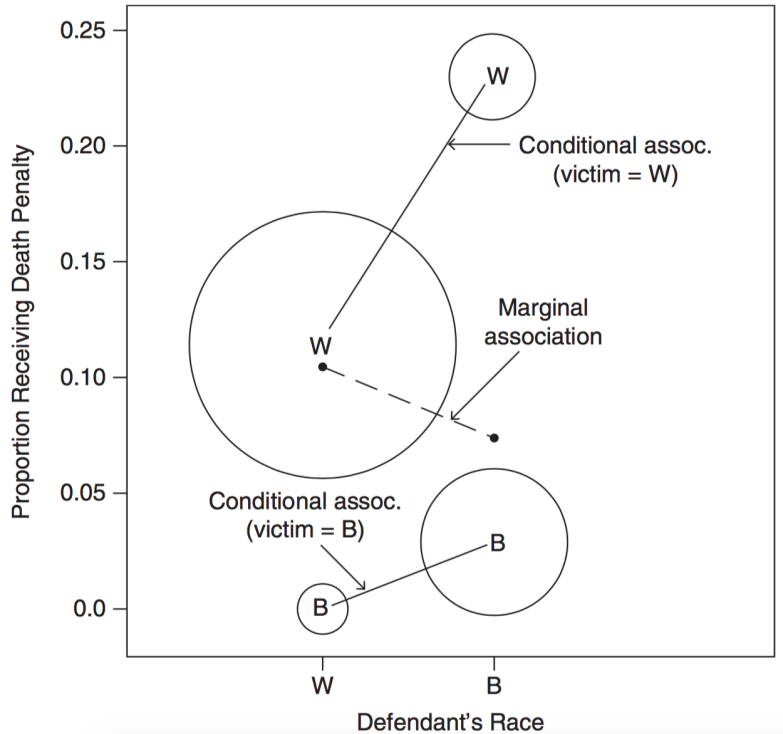


Figure 2.4 Proportion receiving the death penalty by defendant's race and victims' race.

Simpson's Paradox



Conditional and Marginal Odds Ratios

Table 2.9 Death penalty verdict by defendant's race and victims' race.

Victims' Race	Defendant's Race	Death Penalty		Percentage
		Yes	No	Yes
White	White	53	414	11.3
	Black	11	37	22.9
Black	White	0	16	0.0
	Black	4	139	2.8
Total	White	53	430	11.0
	Black	15	176	7.9

Exercise 2.9

The proportion who died from lung cancer was 0.0014 per year for smokers and 0.0001 per year for nonsmokers. The proportion who died from heart disease was 0.0067 for smokers and 0.0041 for nonsmokers. Describe the association of smoking with lung cancer and with heart disease, using the difference of proportions and the odds ratio. Interpret. Which response (lung cancer or heart disease) is more strongly related to smoking, in terms of the reduction in deaths that could occur with an absence of smoking?

Exercise 2.28

Table 2.16 Expected frequencies illustrating that conditional independence does not imply marginal independence.

Clinic	Drug Treatment	Response	
		Success	Failure
1	A	18	12
	B	12	8
2	A	2	8
	B	8	32

Show that X and Y are conditionally independent, given Z.

Exercise 2.28

Table 2.16 Expected frequencies illustrating that conditional independence does not imply marginal independence.

Clinic	Drug Treatment	Response	
		Success	Failure
1	A	18	12
	B	12	8
2	A	2	8
	B	8	32

Explain how the marginal XY association can be so different from its conditional association, using the values of the conditional XZ and YZ odds ratios.

Summary

2. Analyzing Contingency Tables

- Probability Structure for Contingency Tables
- Comparing Proportions in 2×2 Contingency Tables
- The Odds Ratio
- Chi-Squared Tests of Independence
- Testing Independence for Ordinal Variables
- Association in Three-Way Tables

Next Class

3. Generalized Linear Models (GLMs)
   - Components of a Generalized Linear Model
   - Generalized Linear Models for Binary Data
   - Generalized Linear Models for Counts and Rates
   - Statistical Inference and Model Checking
   - Fitting Generalized Linear Models