STATISTICS

INFORMED DECISIONS USING DATA

Fifth Edition

Chapter 4

Describing the Relation between Two Variables

Pearson

4.4 Contingency Tables and Association

Learning Objectives

1. Compute the **marginal distribution** of a variable
2. Use the **conditional distribution** to identify association among categorical data
3. Explain **Simpson's Paradox**

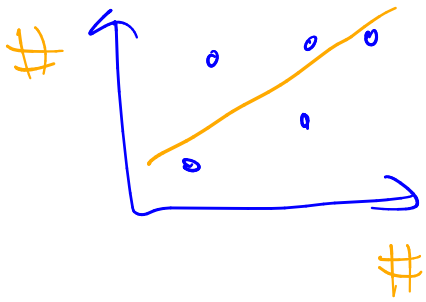
4.4 Contingency Tables and Association

Example: Data Information

Sections 4.1 – 4.3, we studied the relationships between two quantitative variables (QN). (#s discrete or continuous)

Now, we want to study relationships between two qualitative, or categorical, variables (QL).

QN



QL?

4.4 Contingency Tables and Association

Example: Data Information

A professor at a community college in New Mexico conducted a study to assess the effectiveness of delivering an introductory statistics course via traditional lecture-based method, online delivery (no classroom instruction), and hybrid instruction (online course with weekly meetings) methods, the grades students received in each of the courses were tallied.

	Traditional	Online	Hybrid
A	36	39	24
B	52	55	66
C	57	68	90
D	46	38	41
F	46	54	31

The table is referred to as a **contingency table,** or **two-way table**, because it relates two categories of data. The **row variable** is grade, because each row in the table describes the grade received for each group. The **column variable** is delivery method. Each box inside the table is referred to as a **cell**.

4.4 Contingency Tables and Association

4.4.1 Compute the Marginal Distribution of a Variable (1 of 3)

totals in margin

A **marginal distribution** of a variable is a frequency or relative frequency distribution of either the row or column variable in the contingency table.

A professor at a community college in New Mexico conducted a study to assess the effectiveness of delivering an introductory statistics course via traditional lecture-based method, online delivery (no classroom instruction), and hybrid instruction (online course with weekly meetings) methods, the grades students received in each of the courses were tallied. Find the frequency marginal distributions for course grade and delivery method.

173 − 66 − 52 =

	Traditional	Online	Hybrid	Total
A	36	39	24	**99**
B	52	55	66	**173**
C	57	68	90	**215**
D	46	38	41	**125**
F	46	54	31	**131**
Total	**237**	254	**252**	743

margins

4.4 Contingency Tables and Association

4.4.1 Compute the Marginal Distribution of a Variable (1 of 3)

A **marginal distribution** of a variable is a frequency or relative frequency distribution of either the row or column variable in the contingency table.

A professor at a community college in New Mexico conducted a study to assess the effectiveness of delivering an introductory statistics course via traditional lecture-based method, online delivery (no classroom instruction), and hybrid instruction (online course with weekly meetings) methods, the grades students received in each of the courses were tallied. Find the frequency marginal distributions for course grade and delivery method.

	Traditional	Online	Hybrid	Total
A	36	39	24	**99**
B	52	55	66	**173**
C	57	68	90	**215**
D	46	38	41	**125**
F	46	54	31	**131**
Total	**237**	**254**	**252**	743

4.4 Contingency Tables and Association

4.4.1 Compute the Marginal Distribution of a Variable (3 of 3)

EXAMPLE Determining Relative Frequency Marginal Distributions

Determine the relative frequency marginal distribution for course grade and delivery method.

	Traditional	Online	Hybrid	Total
A	36	39	24	
B	52	55	66	**0.233**
C	57	68	90	**0.289**
D	46	38	41	**0.168**
F	46	54	31	**0.176**
		0.342	**0.339**	1.000

4.4 Contingency Tables and Association

4.4.2 Use the Conditional Distribution to Identify Association among Categorical Data (1 of 4)

A **conditional distribution** lists the relative frequency of each category of the response variable, given a specific value of the explanatory variable

EXAMPLE Determining a Conditional Distribution

Construct a conditional distribution of course grade by method of delivery. Comment on any type of association that may exist between course grade and delivery method.

It appears that students in the hybrid course are more likely to pass (A, B, or C) than the other two methods.

	Traditional	Online	Hybrid
A	36	39	24
B	52	55	66
C	57	68	90
D	46	38	41
F	46	54	31

	Traditional	Online	Hybrid
A	0.152	0.154	0.095
B	**0.219**	**0.217**	**0.262**
C	**0.241**	**0.268**	**0.357**
D	**0.194**	**0.150**	**0.163**
F	**0.194**	0.213	**0.123**

Pearson

4.4 Contingency Tables and Association

4.4.2 Use the Conditional Distribution to Identify Association among Categorical Data (1 of 4)

A **conditional distribution** lists the relative frequency of each category of the response variable, given a specific value of the explanatory variable

EXAMPLE Determining a Conditional Distribution

Construct a conditional distribution of course grade by method of delivery. Comment on any type of association that may exist between course grade and delivery method.

It appears that students in the hybrid course are more likely to pass (A, B, or C) than the other two methods.

	Traditional	Online	Hybrid
A	36	39	24
B	52	55	66
C	57	68	90
D	46	38	41
F	46	54	31

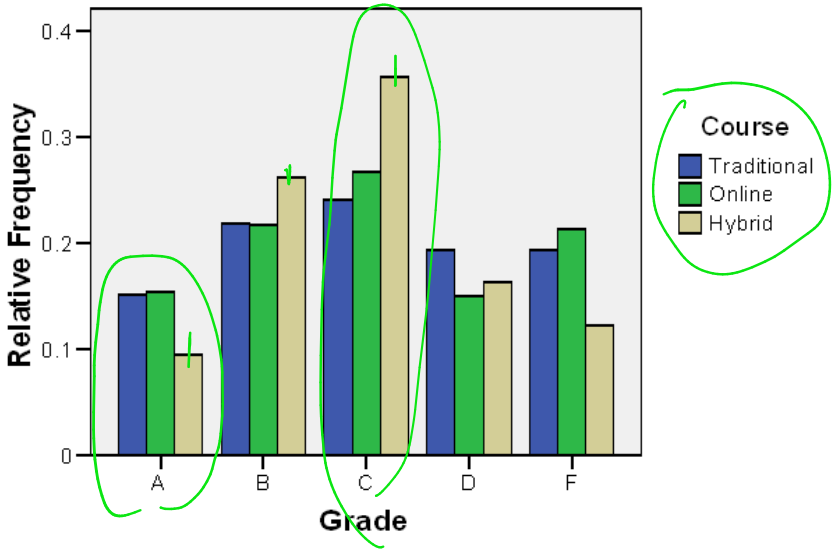
	Traditional	Online	Hybrid
A	**0.152**	**0.154**	**0.095**
B	**0.219**	**0.217**	**0.262**
C	**0.241**	**0.268**	**0.357**
D	**0.194**	**0.150**	**0.163**
F	**0.194**	**0.213**	**0.123**

4.4 Contingency Tables and Association

4.4.2 Use the Conditional Distribution to Identify Association among Categorical Data (3 of 4)

EXAMPLE Drawing a Bar Graph of a Conditional Distribution

Using the results of the previous example, draw a bar graph that represents the conditional distribution of method of delivery by grade earned.



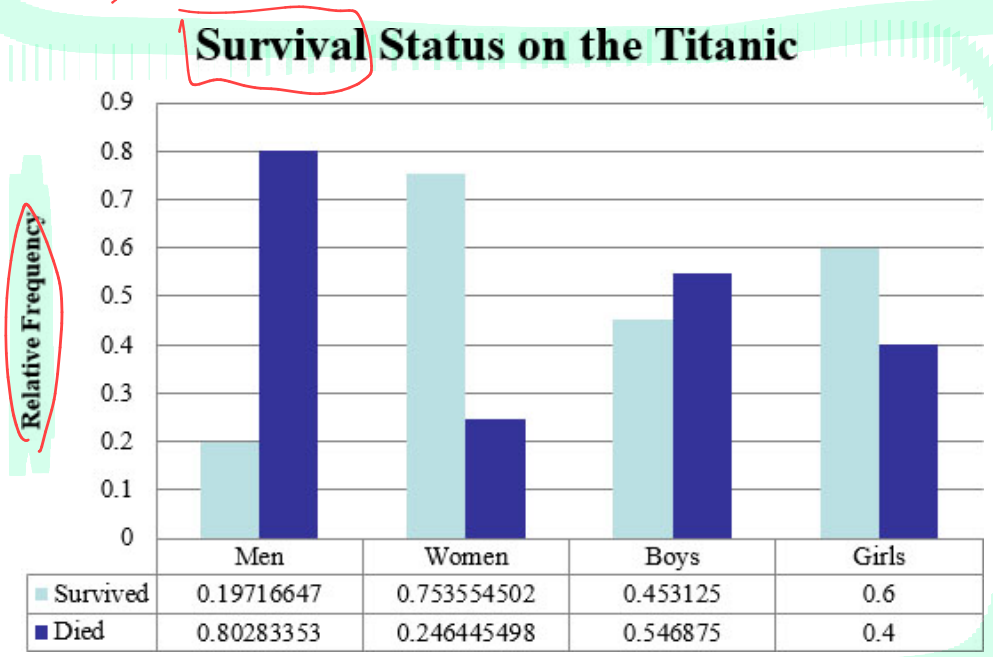
4.4 Contingency Tables and Association

4.4.2 Use the Conditional Distribution to Identify Association among Categorical Data (4 of 4)

The following contingency table shows the survival status and demographics of passengers on the ill-fated Titanic.

	Men	Women	Boys	Girls
Survived	334	318	29	27
Died	1360	104	35	18

Draw a conditional bar graph of survival status by demographic characteristic.



Pearson

4.4 Contingency Tables and Association

4.4.2 Explain Simpson's Paradox (1 of 6)

EXAMPLE Illustrating Simpson's Paradox

Insulin dependent (or *Type 1*) **diabetes** is a disease that results in the permanent destruction of insulin-producing beta cells of the pancreas. Type 1 diabetes is lethal unless treatment with insulin injections replaces the missing hormone. Individuals with insulin independent (or *Type 2*) diabetes can produce insulin internally. The data shown in the table below represent the survival status of 902 patients with diabetes by type over a 5-year period.

	Type 1	**Type 2**	**Total**
Survived	253	326	**579**
Died	105	218	**323**
	358	**544**	**902**

4.4 Contingency Tables and Association

4.4.2 Explain Simpson's Paradox (2 of 6)

EXAMPLE Illustrating Simpson's Paradox

	Type 1	Type 2	Total
Survived	253	326	**579**
Died	105	218	**323**
	358	**544**	**902**

From the table, the proportion of patients with Type 1 diabetes who died was $\frac{105}{358} = 0.29$; the proportion of patients with Type 2 diabetes who died was $\frac{218}{544} = 0.40$. Based on this, we might conclude that Type 2 diabetes is more lethal than Type 1 diabetes.

4.4 Contingency Tables and Association

4.4.2 Explain Simpson's Paradox (3 of 6)

However, Type 2 diabetes is usually contracted after the age of 40. If we account for the variable age and divide our patients into two groups (those 40 or younger and those over 40), we obtain the data in the table below.

	Type 1		**Type 2**		**Total**
	≤ 40	> 40	≤ 40	> 40	
Survived	129	124	15	311	**579**
Died	1	104	0	218	**323**
	130	**228**	**15**	**529**	**902**

4.4 Contingency Tables and Association

4.4.2 Explain Simpson's Paradox (4 of 6)

	Type 1		Type 2		Total
	≤ 40	> 40	≤ 40	> 40	
Survived	129	124	15	311	**579**
Died	1	104	0	218	**323**
	130	**228**	**15**	**529**	**902**

Of the diabetics 40 years of age or younger, the proportion of those with Type 1 diabetes who died is $\frac{1}{130} = 0.008$; the proportion of those with Type 2 diabetes who died is $\frac{0}{15} = 0$.

4.4 Contingency Tables and Association

4.4.2 Explain Simpson's Paradox (5 of 6)

	Type 1		Type 2		Total
	≤ 40	> 40	≤ 40	> 40	
Survived	129	124	15	311	**579**
Died	1	104	0	218	**323**
	130	**228**	**15**	**529**	**902**

Of the diabetics over 40 years of age, the proportion of those with Type 1 diabetes who died is $\frac{104}{228} = 0.456$; the proportion of those with Type 2 diabetes who died is $\frac{218}{529} = 0.412$.

The lurking variable age led us to believe that Type 2 diabetes is the more dangerous type of diabetes.

4.4 Contingency Tables and Association

4.4.2 Explain Simpson's Paradox (6 of 6)

Simpson's Paradox describes a situation in which an association between two variables inverts or goes away when a third variable is introduced to the analysis.

4.4 Contingency Tables and Association

4.4.2 Explain Simpson's Paradox (6 of 6)

Simpson's Paradox describes a situation in which an association between two variables inverts or goes away when a third variable is introduced to the analysis.

Example: Read Example 6 from textbook (pgs 231-232).

Berkeley Admission Scandal in 70s.

- admission between men (46%) and women (30%) brought lawsuit against school

- however, there were lurking variables (programs of study, how many apply to programs, etc)

engineer Medieval Lit

- Warning against apparent association b/w 2 QL vars when, in fact, there isn't any b/c of lurking variables.