Phil/LPS 31 Introduction to Inductive Logic
Lecture 18

David Mwakima
dmwakima@uci.edu
Department of Logic and Philosophy of Science
University of California, Irvine

June 9th 2023

Topics

- Relative Risk
- Odds Ratio
- Simpson's Paradox

Measures of Association in contingency tables

- Results from case-control studies and randomized experimental studies are often recorded in 2×2 contingency tables.

Measures of Association in contingency tables

- Results from case-control studies and randomized experimental studies are often recorded in 2×2 contingency tables.
- In this lecture we want to learn how to use data in contingency tables to quantify the association between the dependent and independent variables.

Measures of Association in contingency tables

- Results from case-control studies and randomized experimental studies are often recorded in 2×2 contingency tables.
- In this lecture we want to learn how to use data in contingency tables to quantify the association between the dependent and independent variables.
- Some interesting puzzles such as the Simpson's paradox arise in the analysis of contingency tables. Such paradoxes emphasize further the need for control of extraneous confounding variables in making causal inferences.

Contingency Tables

- A contingency table is a rectangular table that cross-classifies categorical variables X and Y. If we let r denote the number of categories of X (the rows) and c denote the number of categories of Y (the columns); a contingency table has cells that display the $r \times c$ possible combinations of outcomes.

Contingency Tables

- A contingency table is a rectangular table that cross-classifies categorical variables X and Y. If we let r denote the number of categories of X (the rows) and c denote the number of categories of Y (the columns); a contingency table has cells that display the $r \times c$ possible combinations of outcomes.
- A table that cross-classifies two variables is called a two-way contingency table; one that cross-classifies three variables is called a three-way contigency table, and so forth.

Contingency Tables

- A contingency table is a rectangular table that cross-classifies categorical variables X and Y. If we let r denote the number of categories of X (the rows) and c denote the number of categories of Y (the columns); a contingency table has cells that display the $r \times c$ possible combinations of outcomes.
- A table that cross-classifies two variables is called a two-way contingency table; one that cross-classifies three variables is called a three-way contigency table, and so forth.
- In two-way tables, usually the column variable Y is a response variable and the other row variable X is an explanatory variable

Contingency Tables: Joint, Marginal and Conditional Probabilities

- There are three kinds of distributions associated with a contingency table: joint probability distribution; marginal probability distribution and conditional probability distribution.

Contingency Tables: Joint, Marginal and Conditional Probabilities

- There are three kinds of distributions associated with a contingency table: joint probability distribution; marginal probability distribution and conditional probability distribution.
- For the conditional distribution we are usually interested in the conditional probabilities of Y, given X at each value or level of X.

Contingency Tables: Joint, Marginal and Conditional Probabilities

- There are three kinds of distributions associated with a contingency table: joint probability distribution; marginal probability distribution and conditional probability distribution.
- For the conditional distribution we are usually interested in the conditional probabilities of Y, given X at each value or level of X.
- Note: When the rows of a contingency table refer to different groups, the sample sizes for those groups are often fixed by the sampling design. An example is a randomized experiment to compare a new drug to placebo in treating some illness, in which half the sample is randomly allocated to each of two treatments. When the marginal totals for X are fixed rather than random, a joint distribution for X and Y is not meaningful, but conditional distributions for Y (given X) are.

Contingency Tables

- Recall the cross-classification table of belief in afterlife by gender in Homework 5.

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

Contingency Tables

- Recall the cross-classification table of belief in afterlife by gender in Homework 5.

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

- Exercise. Let X, gender, be the explanatory variable and Y, belief in afterlife, be the response variable. Find:

Contingency Tables

- Recall the cross-classification table of belief in afterlife by gender in Homework 5.

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

- Exercise. Let X, gender, be the explanatory variable and Y, belief in afterlife, be the response variable. Find:
  (1) the joint probability distribution of the data;

Contingency Tables

- Recall the cross-classification table of belief in afterlife by gender in Homework 5.

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

- Exercise. Let X, gender, be the explanatory variable and Y, belief in afterlife, be the response variable. Find:
  (1) the joint probability distribution of the data;
  (2) the marginal probability distributions for Y;

Contingency Tables

- Recall the cross-classification table of belief in afterlife by gender in Homework 5.

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

- Exercise. Let X, gender, be the explanatory variable and Y, belief in afterlife, be the response variable. Find:
  (1) the joint probability distribution of the data;
  (2) the marginal probability distributions for Y;
  (3) the conditional distributions of belief in afterlife, given gender.

Relative Risk

The following table comes from the Physicians' Health Study, which was a five-year randomized study investigating whether regular intake of aspirin reduces the chance of myocardial infarction (heart attacks). Every other day, the male physicians participating in the study took either one aspirin tablet or a placebo. The study was blind – the physicians in the study did not know which type of pill they were taking.

	Myocardial Infarction		
Group	Yes	No	Total
Placebo	189	10845	11034
Aspirin	104	10933	11037

Relative Risk

- Since the study is a randomized control study, relative risk is a useful descriptive measure of the effect of aspirin in reducing the risk of heart attacks. For 2×2 tables, the relative risk for E, RR(E), is given by:

$$RR(E) = \frac{\text{Proportion in Group 1 with Outcome E}}{\text{Proportion in Group 2 with Outcome E}}$$

Relative Risk

	Myocardial Infarction		
Group	Yes	No	Total
Placebo	189	10845	11034
Aspirin	104	10933	11037

- Exercise. Let E be an incident of myocardial infarction (i.e., heart attack) and E^c an no incident of myocardial infarction. Find $RR(E)$ for physicians who took the placebo compared to those who took aspirin. Interpet the result.

Odds Ratio

The following table comes from one of the first studies of the association between lung cancer and smoking, based on data from 20 hospitals in London, England, in 1920. At the time, many medical scientists thought that the increased rates of lung cancer in London mainly reflected the increasingly severe air pollution due to the burning of coal (and, thus, the frequent "London fog") before the Clean Air Act of 1956. However, the epidemiologist Richard Doll and statistician Austin Bradford Hill thought that smoking could be a culprit. In their study, patients admitted to the hospital with lung cancer in the preceding year were queried about their smoking behavior. Patients were defined as smokers if they had smoked at least one cigarette a day for at least a year. For each patient admitted, they recorded the smoking behavior of a noncancer patient at the same hospital of the same gender and within the same 5-year grouping on age. The 709 cases in the first column are those having lung cancer and the 709 controls in the second column are those not having it.

Odds Ratio

- Here is the table.

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

Odds Ratio

- Here is the table.

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

- This is a case-control study because the study begins with the absence or presence of an outcome (here lung cancer) and then looks backward in time to try to detect possible causes or risk factors.

Odds Ratio

- Normally, whether lung cancer occurs is a response variable and smoking behavior is an explanatory variable. smoking behavior. Here when we assess the relative risk of lung cancer by comparing smokers with nonsmokers the proportions refer to the conditional distribution of lung cancer, given smoking behavior.

Odds Ratio

- Normally, whether lung cancer occurs is a response variable and smoking behavior is an explanatory variable. smoking behavior. Here when we assess the relative risk of lung cancer by comparing smokers with nonsmokers the proportions refer to the conditional distribution of lung cancer, given smoking behavior.
- In contrast, case-control studies provide proportions in the reverse direction. Here we're interested in the conditional distribution of smoking behavior, given lung cancer status.

Odds Ratio

- In 2×2 tables the odds ratio, θ is the appropriate measure of association for detecting possible risk factors in a case-control study.

$$\theta = \frac{\text{Odds of outcome E for those in Group 1}}{\text{Odds of outcome E for those in Group 2}} = \frac{\frac{P_1(E)}{1-P_1(E)}}{\frac{P_2(E)}{1-P_2(E)}}$$

Odds Ratio

- In 2×2 tables the odds ratio, θ is the appropriate measure of association for detecting possible risk factors in a case-control study.

$$\theta = \frac{\text{Odds of outcome E for those in Group 1}}{\text{Odds of outcome E for those in Group 2}} = \frac{\frac{P_1(E)}{1-P_1(E)}}{\frac{P_2(E)}{1-P_2(E)}}$$

- The odds ratio θ can be calculated directly from the cell counts using the formula:

$$\theta = \frac{\frac{P_1(E)}{1-P_1(E)}}{\frac{P_2(E)}{1-P_2(E)}} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

Odds Ratio

- In 2×2 tables the odds ratio, θ is the appropriate measure of association for detecting possible risk factors in a case-control study.

$$\theta = \frac{\text{Odds of outcome E for those in Group 1}}{\text{Odds of outcome E for those in Group 2}} = \frac{\frac{P_1(E)}{1-P_1(E)}}{\frac{P_2(E)}{1-P_2(E)}}$$

- The odds ratio θ can be calculated directly from the cell counts using the formula:

$$\theta = \frac{\frac{P_1(E)}{1-P_1(E)}}{\frac{P_2(E)}{1-P_2(E)}} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

- Because $\theta = \frac{n_{11}n_{22}}{n_{12}n_{21}}$ some authors use the term odds ratio interchangeably with the cross-product ratio.

Intepreting the Odds Ratio

- The odds ratio can equal any non-negative number. When X and Y are independent $\theta = 1$.

Intepreting the Odds Ratio

- The odds ratio can equal any non-negative number. When X and Y are independent $\theta = 1$.
- Odds ratios on each side of 1 reflect certain types of associations. When $\theta > 1$, the odds of outcome E are higher for Group 1 than for Group 2.

Intepreting the Odds Ratio

- The odds ratio can equal any non-negative number. When X and Y are independent $\theta = 1$.
- Odds ratios on each side of 1 reflect certain types of associations. When $\theta > 1$, the odds of outcome E are higher for Group 1 than for Group 2.
- For instance, when $\theta = 4$, the odds of outcome E in Group 1 are four times the odds of outcome E in Group 2.

Intepreting the Odds Ratio

- The odds ratio can equal any non-negative number. When X and Y are independent $\theta = 1$.
- Odds ratios on each side of 1 reflect certain types of associations. When $\theta > 1$, the odds of outcome E are higher for Group 1 than for Group 2.
- For instance, when $\theta = 4$, the odds of outcome E in Group 1 are four times the odds of outcome E in Group 2.
- Values of θ farther from 1 in a given direction represent a stronger association. An odds ratio of 4 is farther from independence than an odds ratio of 2, and an odds ratio of 0.25 is farther from independence than an odds ratio of 0.50.

Facts about the Odds Ratio: Direction of Interest

- The odds ratio does not change value when the table orientation reverses so that the rows become the columns and the columns become the rows.

Facts about the Odds Ratio: Direction of Interest

- The odds ratio does not change value when the table orientation reverses so that the rows become the columns and the columns become the rows.
- The odds ratio is the same when we treat the columns as the response variable and the rows as the explanatory variable, or the rows as the response variable and the columns as the explanatory variable.

Facts about the Odds Ratio: Direction of Interest

- The odds ratio does not change value when the table orientation reverses so that the rows become the columns and the columns become the rows.
- The odds ratio is the same when we treat the columns as the response variable and the rows as the explanatory variable, or the rows as the response variable and the columns as the explanatory variable.
- Because of this symmetry, interpretations can use the direction of interest, even though the study was retrospective.

Odds Ratio

- Consider the smoking data from earlier.

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

Odds Ratio

- Consider the smoking data from earlier.

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

- Calculate the odds ratio of the association between smoking and incidence of lung cancer. Interpret.

Association in three-way contigency tables

- At the beginning of this lecture I mentioned that the analysis of contingency tables leads to some interesting puzzles, which emphasize the need for greater control of lurking or confounding variables in our studies.

Association in three-way contigency tables

- At the beginning of this lecture I mentioned that the analysis of contingency tables leads to some interesting puzzles, which emphasize the need for greater control of lurking or confounding variables in our studies.
- In studying the effect of an explanatory variable X on a response variable Y, we should adjust for confounding variables Z that can influence that relationship because they are associated both with X and with Y.

Association in three-way contigency tables

- At the beginning of this lecture I mentioned that the analysis of contingency tables leads to some interesting puzzles, which emphasize the need for greater control of lurking or confounding variables in our studies.
- In studying the effect of an explanatory variable X on a response variable Y, we should adjust for confounding variables Z that can influence that relationship because they are associated both with X and with Y.
- Without controlling for Z, an observed XY association may merely reflect an association of Z with X and Y. This is especially vital for observational studies, for which one cannot remove effects of such variables by randomly assigning subjects to different treatments.

Association in three-way contigency tables

- At the beginning of this lecture I mentioned that the analysis of contingency tables leads to some interesting puzzles, which emphasize the need for greater control of lurking or confounding variables in our studies.
- In studying the effect of an explanatory variable X on a response variable Y, we should adjust for confounding variables Z that can influence that relationship because they are associated both with X and with Y.
- Without controlling for Z, an observed XY association may merely reflect an association of Z with X and Y. This is especially vital for observational studies, for which one cannot remove effects of such variables by randomly assigning subjects to different treatments.
- When we add control variables like Z we get a three-way contingency table.

Partial Tables, Marginal Tables and Simpson's Paradox

- Three way contingency tables lead to two-way partial tables, which cross-classify X and Y at separate categories or levels of Z. A two-way partial table controls for the effect of Z by holding its value constant.

Partial Tables, Marginal Tables and Simpson's Paradox

- Three way contingency tables lead to two-way partial tables, which cross-classify X and Y at separate categories or levels of Z. A two-way partial table controls for the effect of Z by holding its value constant.
- The XY marginal table is the table that results from combining the partial tables.

Partial Tables, Marginal Tables and Simpson's Paradox

- Three way contingency tables lead to two-way partial tables, which cross-classify X and Y at separate categories or levels of Z. A two-way partial table controls for the effect of Z by holding its value constant.
- The XY marginal table is the table that results from combining the partial tables.
- The associations in partial tables are called conditional associations because they show the association between X and Y conditional on Z being constant, i.e., controlling for Z.

Partial Tables, Marginal Tables and Simpson's Paradox

- Three way contingency tables lead to two-way partial tables, which cross-classify X and Y at separate categories or levels of Z. A two-way partial table controls for the effect of Z by holding its value constant.
- The XY marginal table is the table that results from combining the partial tables.
- The associations in partial tables are called conditional associations because they show the association between X and Y conditional on Z being constant, i.e., controlling for Z.
- The association in the marginal table is called the marginal association between X and Y.

Partial Tables, Marginal Tables and Simpson's Paradox

- Three way contingency tables lead to two-way partial tables, which cross-classify X and Y at separate categories or levels of Z. A two-way partial table controls for the effect of Z by holding its value constant.
- The XY marginal table is the table that results from combining the partial tables.
- The associations in partial tables are called conditional associations because they show the association between X and Y conditional on Z being constant, i.e., controlling for Z.
- The association in the marginal table is called the marginal association between X and Y.
- Simpson's Paradox arises whenever the direction of conditional association between X and Y is different from the direction of marginal association between X and Y.

Simpson's Paradox: First Illustration

- If a major league batter gets $f = 152$ hits in $n = 500$ official at bats(AB) during the season, then the relative frequency $f/n = 0.304$ is an estimate of his probability of getting a hit and is called his batting average for that season.

Simpson's Paradox: First Illustration

- If a major league batter gets $f = 152$ hits in $n = 500$ official at bats(AB) during the season, then the relative frequency $f/n = 0.304$ is an estimate of his probability of getting a hit and is called his batting average for that season.
- Question: Is it possible for batter A to have a higher average than batter B during each season of their careers yet, at the end of their careers, batter B has a better batting average than batter A?

Simpson's Paradox: First Illustration

- If a major league batter gets $f = 152$ hits in $n = 500$ official at bats(AB) during the season, then the relative frequency $f/n = 0.304$ is an estimate of his probability of getting a hit and is called his batting average for that season.
- Question: Is it possible for batter A to have a higher average than batter B during each season of their careers yet, at the end of their careers, batter B has a better batting average than batter A?
- You might say, "Obviously, no!" But wait... "Obviousness is..."

Simpson's Paradox: First Illustration

- Consider the following data on the batting average between two players A and B.

	Player A			Player B		
Season	At Bat	Hits	Average	At Bat	Hits	Average
1	500	126	0.252	300	75	0.250
2	300	90	0.300	500	145	0.290
Totals	800	216	0.270	800	220	0.275

Simpson's Paradox: First Illustration

- Consider the following data on the batting average between two players A and B.

	Player A			Player B		
Season	At Bat	Hits	Average	At Bat	Hits	Average
1	500	126	0.252	300	75	0.250
2	300	90	0.300	500	145	0.290
Totals	800	216	0.270	800	220	0.275

- Controlling for season (or conditional on the season played), Player A has a higher batting average than Player B. However, marginally, Player B has a higher batting average than Player A. This is Simpson's Paradox.

Simpson's Paradox: Second Illustration

- The following $2 \times 2 \times 2$ contingency table is from an article that studied the effects of racial characteristics on whether subjects convicted of homicides receive the death penalty. The 674 subjects were the defendants in indictments involving cases with multiple murders, in Florida during a 12-year period. The variables are Y = death penalty verdict, X = race of defendant, and Z = race of victims. The question of interest was to study the association between a defendant's race and whether they received the death penalty verdict, controlling for the victims' race.

Simpson's Paradox: Second Illustration

- Here is the table.

Victims' Race	Defendant's Race	Death Penalty	
		Yes	No
White	White	53	414
	Black	11	37
Black	White	0	16
	Black	4	139

Simpson's Paradox: Second Illustration

- Here is the table.

Victims' Race	Defendant's Race	Death Penalty	
		Yes	No
White	White	53	414
	Black	11	37
Black	White	0	16
	Black	4	139

- Exercise.

Simpson's Paradox: Second Illustration

- Here is the table.

Victims' Race	Defendant's Race	Death Penalty	
		Yes	No
White	White	53	414
	Black	11	37
Black	White	0	16
	Black	4	139

- Exercise.
  - Write down the marginal table for the association between the defendant's race and whether they received the death penalty verdict. Calculate the odds ratio. This is called the marginal odds ratio. Interpret it.

Simpson's Paradox: Second Illustration

- Here is the table.

Victims' Race	Defendant's Race	Death Penalty	
		Yes	No
White	White	53	414
	Black	11	37
Black	White	0	16
	Black	4	139

- Exercise.
  - Write down the marginal table for the association between the defendant's race and whether they received the death penalty verdict. Calculate the odds ratio. This is called the marginal odds ratio. Interpret it.
  - Write down the partial table if the victim's race is white and calculate the odds ratio. This is also a marginal odds ratio. Interpret it.

Simpson's Paradox: Second Illustration Cont'd.

- Exercise (Cont'd.)

Simpson's Paradox: Second Illustration Cont'd.

- Exercise (Cont'd.)
  - Write down the partial table if the victim's race as black and calculate the odds ratio. This is called the conditional odds ratio Interpret it.

Simpson's Paradox: Second Illustration Cont'd.

- Exercise (Cont'd.)
  - Write down the partial table if the victim's race as black and calculate the odds ratio. This is called the conditional odds ratio Interpret it.
  - Compare the direction of conditional association to the direction of marginal association, is this Simpson's Paradox?

Conditional Independence

- If the population has X and Y independent in each partial table, then X and Y are said to be conditionally independent, given Z.

Conditional Independence

- If the population has X and Y independent in each partial table, then X and Y are said to be conditionally independent, given Z.
- To check for conditional independence check that all the conditional odds ratios between X and Y are equal 1.

Conditional Independence

- If the population has X and Y independent in each partial table, then X and Y are said to be conditionally independent, given Z.
- To check for conditional independence check that all the conditional odds ratios between X and Y are equal 1.
- In general, conditional independence of X and Y, given Z, does not imply marginal independence of X and Y.