JMP017: Smoking and Lung Cancer

Odds Ratio for Retrospective Analysis

Produced by:

Dr. DeWayne Derryberry
Idaho State University, Department of Mathematics

Smoking and Lung Cancer
Odds Ratio for Retrospective Analysis

Key ideas

Conditional Probability, Odds, Retrospective Observational Study

Background

When dealing with categorical data and rare events, sample size is often a problem. For example, if we want to show an association between smoking and lung cancer, there are four groups to consider – smokers with and without cancer, and non-smokers with and without cancer. It can be shown (see the exercises) that the power of any such study is limited by the sample size in the smallest of these groups.

In any large population, cancer of any particular type is quite rare. In many epidemiological studies of rare events (as with many diseases), the only way to get a large sample size for the groups that have the disease is to wait until the disease has occurred to collect the data – targeting those with the disease. This kind of study, where we wait for the outcome and then collect the data, is called a retrospective study.

In a retrospective study we have those with the disease (the cases) and must use subject-matter expertise to select a comparable group without the disease (the controls). We then examine the differences between these groups with regard to some factor we consider to be potentially causal for the disease. If we believe the cases and controls are similar in other ways, we can make an argument for causality.

For example, let's say we're interested in studying lung cancer and smoking. We can find people who have lung cancer and others without lung cancer who are otherwise comparable, and compare their smoking activity. If the lung cancer patients are more often smokers, can we make a plausible argument that smoking causes lung cancer? And, how strong is this argument?

From this sort of study we cannot directly estimate the risk of cancer for smokers! We have a group of people with cancer and we have estimated, from this sample, the proportion of people with cancer who smoke. And, hopefully, we have a comparable group of people without cancer. We have estimated, from this second sample, the proportion of people without cancer who smoke. All we can estimate directly is the risk of being a smoker, if a person does or does not have lung cancer – not a very interesting calculation. We are really interested in what proportion of smokers and non-smokers get cancer.

As an aside, those who have seen Bayes theorem may know these probabilities can be flipped. However, in this case there is not enough information. For example, at the very least, to use Bayes theorem we would need to know the overall rate of either smoking or lung cancer in the population. Not only do we not know this, but we also cannot be sure what population, if any, is represented by our cases and controls.

The Task

Our goal, in this retrospective study, is to determine if there is a positive association between smoking and lung cancer, and to (through mathematical manipulations) estimate the risk of lung cancer for smokers relative to non-smokers.

The Data Smoking.jmp

The following data is fictitious, but typical of the kinds of data often found in such studies.

There were 121 lung cancer patients matched with 118 patients without lung cancer from the same hospital. The number in each group who smoked was determined.

Outcome	With or without lung cancer
Smoker	Smoker or non-smoker
Count	Number of patients in each category

Exhibit 1 The Data

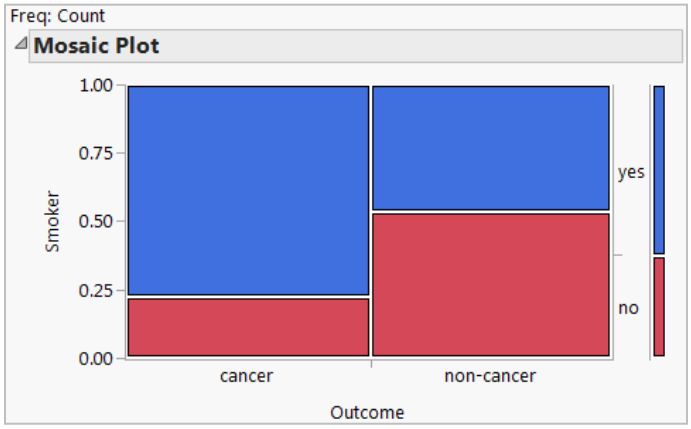
Outcome	Smoker	Count
Cancer	Yes	94
Non-Cancer	Yes	55
Cancer	No	27
Non-Cancer	No	63

Analysis

We start by looking at the relationship between **Outcome** and **Smoker**. Although we normally think of the outcome as the y variable and the factor as the x variable, a retrospective study changes the roles of x and y due to the way the data were collected. The factor in this case is **Outcome**, and the response of interest is **Smoker**.

Is there an obvious pattern (Exhibit 2)?

Exhibit 2 A Visual Representation



*(Analyze > Fit Y by X, select **Smoker** as Y, Response, **Outcome** as X, Factor, and **Count** as Freq, and click OK.)*

There is a definite contrast between the two groups in this display. But, keep in mind that we are not looking at a difference in cancer rates among smokers and non-smokers. Based on the way the data were collected, we are looking horizontally – at the difference in smoking rates among those with and without cancer. Those with lung cancer appear to be much more likely to smoke than those without lung cancer.

Do those with lung cancer smoke more than those without lung cancer? This question is not the one we are really interested in, but it is the one we can directly ask given the manner in which the data were collected.

The Likelihood ratio and Pearson tests (Exhibit 3) provide overwhelming evidence of an association between smoking and lung cancer (p-value < 0.0001). Notice that the tests have been structured in such a way that we are looking at the probability that someone is a smoker, given they do or do not have cancer. This is due to the way we have declared our x and y variables, and it is important that we use this interpretation for our analyses.

Exhibit 3 Likelihood Ratio, Pearson's, and Fisher's Exact Tests

Tests

N	DF	-LogLike	RSquare (U)
239	1	12.550595	0.0793

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	25.101	<.0001*
Pearson	24.574	<.0001*

Fisher's Exact Test	Prob	Alternative Hypothesis
Left	<.0001*	Prob(Smoker=yes) is greater for Outcome=cancer than non-cancer
Right	1.0000	Prob(Smoker=yes) is greater for Outcome=non-cancer than cancer
2-Tail	<.0001*	Prob(Smoker=yes) is different across Outcome

Fisher's exact test (bottom of Exhibit 3) shows that the association is positive (the smoking rate is higher for those with cancer than those without cancer). It is the test of proportions where we get an interesting result (Exhibit 4).

Exhibit 4 Estimation, Comparing Proportion

Two Sample Test for Proportions

Description	Proportion Difference	Lower 95%	Upper 95%
P(yes|cancer)-P(yes|non-cancer)	0.310758	0.18968	0.421702

Adjusted Wald Test	Prob
P(yes|cancer)-P(yes|non-cancer) >= 0	<.0001*
P(yes|cancer)-P(yes|non-cancer) <= 0	1.0000
P(yes|cancer)-P(yes|non-cancer) = 0	<.0001*

Response Smoker category of interest

- no
- yes

*(From the Fit Y by X output, click on the top red triangle and select Two Sample Test for Proportions. This option is only available for 2 x 2 tables. The default category of interest, in this case, is **no**. Change the category of interest to **yes** (smoker) using the radio button at the bottom of the output.)*

For this study we can directly state that those with cancer are 19% to 42% more likely to be smokers than those who don't have cancer. This tells us little about the risk of lung cancer for smokers.

Odds Ratios

Is there any information about the risk of lung cancer, for smokers, that can be gleaned from these data? Fortunately, the answer is yes.

To be more mathematically precise, we have information of the form $\Pr(S|C)$ and $\Pr(S \mid C')$, the probability someone is a smoker, given they do or do not have cancer. What we really want is some information of the form $\Pr(C|S)$ and $\Pr(C|S')$, the probability of having lung cancer if you are or are not a smoker.

An important notion in this regard is the *odds* of an event. The odds are defined as:

$$Odds(A) = \Pr(A)/[1 - \Pr(A)]$$

Notice, for rare events we have $Odds(A) \approx \Pr(A)$.

For events *A* and *B* the odds ratio is $Odds(A)/Odds(B)$. If both *A* and *B* are rare, we again have:

$$Odds(A)/Odds(B) \approx \Pr(A)/\Pr(B)$$

In other words, the odds ratio and relative risk are approximately the same when Pr(A) and Pr(B) are very small. This leads to a really important idea involving odds ratios and conditional probability. Consider the data displayed in a two-way table (Exhibit 5).

Exhibit 5 Data in a Two-Way (Contingency) Table

Contingency Table

Outcome	Count / Expected / Cell Chi^2	Smoker: no	Smoker: yes	
	cancer	27 45.5649 7.5640	94 75.4351 4.5689	121
	non-cancer	63 44.4351 7.7563	55 73.5649 4.6850	118
		90	149	239

(In the Fit Y by X output, hold the Alt key and click on the Contingency Table red triangle. Deselect Total %, Col % and Row %, select Expected and Cell Chi Square, and click OK.)

	No	Yes
Cancer	a = 27	b = 94
Non-Cancer	c = 63	d = 55

Notice the following: $\Pr(S|C) = b/(a+b)$ and $Odds(S|C) = b/a$.

Now consider the following odds ratio:

$$\frac{Odds(S|C)}{Odds(S|C')} = \frac{b/a}{d/c} = \frac{bc}{ad} = \frac{b/d}{a/c} = \frac{Odds(C|S)}{Odds(C|S')}$$

This is quite an amazing result. This is a purely mathematical property, but is does tell us that one piece of prospective information can be extracted from a retrospective two-way table. It is possible to determine the odds of cancer for smokers and non-smokers. Given that lung cancer is a rare event, we can further state that:

$$\frac{Odds(C|S)}{Odds(C|S')} \approx \frac{\Pr(C|S)}{\Pr(C|S')}$$

In other words, for rare events we can estimate the prospective relative risk from a retrospective study (of course, retrospective studies and rare events go hand-in-hand).

In this particular study, the odds ratio is given in JMP as

$$\frac{Odds(S|C')}{Odds(S|C)} = \frac{Odds(C|S')}{Odds(C|S)} \approx \frac{\Pr(C|S')}{Pr(C|S)}$$

Notice: $ad\ /\ bc = [27 \times 55]/[94 \times 63]\ = 0.25076$. So, the point estimate for the odds ratio (Exhibit 6) is 0.25, and the 95% confidence interval for the estimate is 0.143 to 0.439.

Exhibit 6 The Odds Ratio that Someone with Cancer Is a Non-Smoker

Odds Ratio

Odds Ratio	Lower 95%	Upper 95%
0.25076	0.14319	0.439141

(From the Fit Y by X output, select Odds Ratio from the top red triangle.)

This odds ratio is for the odds that someone with cancer is a nonsmoker. We're actually interested in the reciprocal – the odds ratio for someone with cancer who *is* a smoker (Exhibit 7).

Exhibit 7 The Odds that Someone with Cancer Is a Smoker

Odds Ratio

Odds Ratio	Lower 95%	Upper 95%
3.987879	2.277174	6.983733

*(In the data table, right-click on the **Smoker** column header and select Column Info. Click on Column Properties, and select Value Ordering. Select "yes", click the Move Up button, then click OK. Finally, re-run the previous analysis.)*

This tells us that smokers are roughly four times more likely to get lung cancer than non-smokers (3.99 ≈ 1/0.25076). Further, the confidence interval suggests that smokers are 2.28 (≈ 1/0.439) to 6.98 (≈1/0.143) times more likely to get lung cancer than non-smokers!

Summary

Statistical Insights - Odds Ratios and Retrospective Analysis

In a retrospective analysis what is usually thought of as the outcome (lung cancer in this case) is the X variable, while the alleged cause (smoking) is treated as the Y variable.

Implications: What Can We Really Say, Based on This Study?

This study depends heavily on the comparability of the cases and controls. The story of how it came to be known that smoking causes lung cancer is quite an interesting one and was built on many (thousands) studies, including surveys, case-control studies, and animal experiments. We will leave the causal status of this, and similar studies, to epidemiologists.

JMP® Features and Hints

In this case we used Fit Y by X to perform hypothesis tests and compute odds ratios. A variety of options are available under the top red triangle for 2-by-2 tables. Display options can also be changed for the contingency table by clicking on the red triangle.

Value ordering, one of many column properties, was used to change the odds ratio calculated, from the odds that someone with cancer is a non-smoker to the odds that someone with cancer is a non-smoker. Value ordering can also be used to change the order in which categorical data are displayed in graphs – by default (unless value ordering is applied), categorical variables are plotted in alphanumeric order.

Exercises

1. A researcher was interested in whether taking an introductory statistics class in college has an influence on the future success of a student. The researcher identified 231 people in a city who made $60,000 or more, and 187 people comparable in age, education and a variety of other variables who made less than $60,000. The researcher asked each person if they had ever taken an introductory statistics class (this data is obviously made up, and may indicate biases of the author).

Income	Statistics	Count
High	Yes	151
Low	Yes	32
High	No	80
Low	no	155

   a. If there is a causal relationship, which variable would be viewed as the cause? Which the effect?
   b. Based on the way the data was collected, which should be the X variable, and which the Y variable?
   c. Compute a meaningful relative risk calculation and explain precisely how this should be interpreted.
   d. Compute the odds ratio. Does this indicate taking statistics is associated with higher income?
   e. Does this study show people who take statistics will have increased income?

2. Construct a data set similar (with the same or similar sample sizes) to the one given in the case, but in which the odds ratio is about 1/6 instead of about 1/4.

3. Double the sample size for the case data and run the new data in JMP.

    a. How do test statistics change?

    b. How do confidence intervals change?

4. (Optional) Given the two-way table display in Exhibit 5, a 95% confidence interval for the odds ratio is found as follows

    Step 1:

$$\log(ad/bc) \pm 1.96\sqrt{1/a + 1/b + 1/c + 1/d}$$

    This creates a lower (L) and upper (U) bound.

    Step 2: These values are then exponentiated to get the odds.

    Verify, by hand, the confidence interval found using JMP.

5. (Optional, based on exercise 3) Explain why, unless all the sample values (a,b,c,d) get larger, the confidence interval cannot become small (narrower). This is the reason large samples are not as great as they seem when working with rare events.

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies. Copyright © 2021, SAS Institute Inc. All rights reserved.