DSC 152:
Applied Statistical Data Analysis and Inference

Lecture #15
Relative Risks and Odds Ratios

Tuesday, May 19
Spring Quarter 2026
Peter Chi

Remember this?

X · Ted_Underwood

910+ likes · 9 years ago

A stubborn love of bacon just taught more Americans the difference between p values and effect size than 100 stats courses could.

Remember this?

Cancer

This article is more than **10 years old**

Processed meats rank alongside smoking as cancer causes – WHO

UN health body says bacon, sausages and ham among most carcinogenic substances along with cigarettes, alcohol, asbestos and arsenic

Sarah Boseley *Health editor*

Mon 26 Oct 2015 08.30 EDT

Remember this?

Carcinogenicity of consumption of red and processed meat

In October, 2015, 22 scientists from ten countries met at the International Agency for Research on Cancer (IARC) in Lyon, France, to evaluate the carcinogenicity of the consumption of red meat and processed meat. These assessments will be published in

more than 200 g per person per day.[4] Less information is available on the consumption of processed meat.

The Working Group assessed more than 800 epidemiological studies that investigated the association of cancer with consumption of red meat

day of red meat and an 18% increase (95% CI 1·10–1·28) per 50 g per day of processed meat.[12]

Data were also available for more than 15 other types of cancer. Positive associations were seen in cohort studies and population-based case-

The "18% increase" represents a "relative risk" of 1.18. What is a relative risk?

Binary Outcome Variables

The outcome variable in this example is binary:

- Get colon cancer
- Don't get colon cancer

And in fact, the predictor variable is also binary:

- Eat $\geq$ 50g/day of red/processed meats
- Do not eat $\geq$ 50g/day of red/processed meats

The data

The raw data look (something)* like this:

	colon cancer	
eat meats	yes	no
yes	13,513	215,725
no	12,018	228,355

*These data were generated by Claude, under directions to make them as similar as possible to the real data based on what information is publicly available.

The data formation (in case you are curious)

The IARC Monograph Volume 114 showed this as the relative risk for colorectal cancer corresponding to a 50g/day intake of red and processed meats including bacon, claiming that it was a meta-analysis of 10 studies. Do you have access to those 10 studies or otherwise are able to find the raw data that they used?

Searched the web >

Good — the primary source is clearly identified. The RR of 1.18 (95% CI 1.10–1.28) for processed meat comes specifically from **Chan et al. (2011)**, published in *PLoS ONE*, which was then cited by IARC Monograph Volume 114. The paper is open-access, so I can tell you a lot about what's in it.

The source of the RR 1.18

The IARC Monograph Vol. 114 (2015) cited the meta-analysis by **Chan DSM, Lau R, Aune D, Vieira R, Greenwood DC, Kampman E, Norat T. (2011).** *Red and Processed Meat and Colorectal Cancer Incidence: Meta-Analysis of Prospective Studies.* PLoS ONE 6(6): e20456. PLOS

That paper reported an RR for colorectal cancer of **1.18 (95% CI 1.10–1.28) per 50 g/day increase** in processed meat — exactly the figure cited by IARC. PLOS

The "10 studies" — what they are

The "10 studies" referenced in IARC messaging refers specifically to new prospective studies published *after* the 2007 WCRF/AICR report — i.e., the studies that motivated Chan

The data formation (in case you are curious)

Would you like me to try to extract those study-level estimates from the supplementary tables, or are you interested in interrogating a specific aspect of how the pooling was done?

Can you extract the study-level estimates and then use these to construct an approximate 2x2 table that could plausibly represent the raw data?

Let me fetch the full paper including supplementary data to get the study-level estimates.

Searched the web >

The supplementary tables with individual study-level RR data aren't accessible via web fetch (they're in the PDF/supplementary files). However, I have enough information from the paper's methods, the referenced studies, and the literature to reconstruct the study-level estimates from the original publications. Let me now build the full analysis computationally — working from published per-study RRs, back-calculating plausible cell counts, and then aggregating.

Now what is a relative risk?

A relative risk (RR) is simply:

$$RR = \frac{p_1}{p_2}$$

where:

- p_1 is the probability of observing the outcome of interest among the exposed group
- p_2 is the probability of observing the outcome of interest among the unexposed group

Here, they said it's 1.18...

Relative Risk of Colon Cancer from Bacon

The data again

	colon cancer		
eat meats	yes	no	totals
yes	13,513	215,725	229,238
no	12,018	228,355	240,373

$$\widehat{RR} = \frac{\frac{13{,}513}{229{,}238}}{\frac{12{,}018}{240{,}373}} \approx 1.18$$

Now, what about a statistical test?

Relative Risk of Colon Cancer from Bacon

If there is no impact of bacon on colon cancer, then we would expect that $RR = 1$. So,

$$H_0 : RR = 1$$
$$H_A : RR \neq 1$$

which is also equivalent to

$$H_0 : p_1 = p_2$$
$$H_A : p_1 \neq p_2$$

There are lots of different ways to test this...

Let's do a permutation test

Here is a sample of the dataframe:

	colon_cancer	eat_meats
200872	0	1
176786	0	1
211299	0	1
352925	0	0
93521	0	1
356651	0	0
344702	0	0
104878	0	1
422513	0	0
429666	0	0

Now we can just shuffle either column...

Let's do a permutation test

Your Turn #1

Write R code for a permutation test here:

- First, you should write a function to calculate the observed RR for any given dataframe.
  - Your function may assume that the input will be a dataframe in which the first column is the outcome variable, and the 2nd column is the exposure variable.
- Then, do the shuffling on the dataframe
- Store the RR from each permuted dataframe into a vector
- Find the p-value

Let's do a permutation test

Your Turn #1

For this illustration, here is code to create a smaller dataframe that you may simply copy/retype:

```
bacon <- data.frame(colon_cancer=c(rep(1, 135+120),
                                   rep(0, 2157+2283)),
                    eat_meats=c(rep(1, 135), rep(0, 120),
                                rep(1, 2157), rep(0, 2283)))
```

A couple of points:

- If we used the full dataframe, it would take much longer to run, and also would give a p-value of basically 0 since the sample size is so large.
- The key is figuring out how to calculate estimates of p_1 and p_2 as defined on Slide 9.

Limitations of the relative risk

Note that we were able to calculate the RR here because the data were cross-sectional. What do we mean by that?

This is in contrast to a "cohort study" and a "case-control study."

Limitations of the relative risk

Summary

Cross-sectional study

Data are collected irrespective of exposure or outcome status (i.e. just a simple random sample of the population). RR can be calculated.

Cohort study

Study participants are either split into cohorts: those who are exposed, and those who are unexposed, and observed over time – or, they are sampled based on exposure status. RR can be calculated.

Case-control study

Study participants are sampled based on outcome status. RR can NOT calculated.

Limitations of the relative risk

But, the case-control study design is sometimes a very appealing one, specifically when the outcome is rare! Why?

Benefits of the case-control study design

- Suppose we are studying a very rare disease. If we do a cross-sectional or cohort study and do not have a gigantic sample size, we might not get any individuals with the disease into our study!

- If we instead sample participants into our study based on their disease status, we can ensure that we have individuals with the disease in our study.

But if we do that, then we have a denominator issue if we wanted to calculate the RR! Why?

Limitations of the relative risk

Suppose we are studying bladder cancer (which has a population prevalence of approximately 0.023%) and want to investigate whether consumption of artificial sweeteners causes this cancer.

Side question: what would be the ideal study design and why can't we do that?

So instead we perform a case-control study in which:

- We sample 100 cases and 100 controls
- We determine how many of each regularly consume artificial sweeteners

Limitations of the relative risk

The data look like this:

	bladder cancer		
sweetener use	yes	no	totals
yes	33	25	58
no	67	75	142
totals	100	100	200

If we wanted to calculate an estimate of the RR, it would be:

$$\widehat{RR} = \frac{\frac{33}{58}}{\frac{67}{142}}$$

But based on our study design, 58 and 142 are not valid denominators!

- For example, this would suggest that 33 out of 58 people who use sweeteners get bladder cancer (approximately 57%).
- Recall that the population prevalence of bladder cancer is approximately 0.023%!

Limitations of the relative risk

The data look like this:

	bladder cancer		
sweetener use	yes	no	totals
yes	33	25	58
no	67	75	142
totals	100	100	200

Conversely, it is valid to calculate:

$$\frac{\frac{33}{100}}{\frac{25}{100}}$$

as each 100 are valid denominators. But this doesn't exactly answer the question we wanted...

- Out of those who have bladder cancer, 33% of them used sweeteners
- Out of those who do not have bladder cancer, 25% of them used sweeteners

Limitations of the relative risk

The quantity shown on the previous slide is pretty much never used.

- Statistically, it is a fine measure and could be validly used for hypothesis testing.
- But it does not have an interpretation that we want.

Specifically, we frequently want to know answers to questions like, "What is the increase in your risk of bladder cancer associated with using artificial sweeteners?"

This question is not possible to answer with the quantity on the previous slide.

Odds Ratios

So, if we can't calculate the RR in a case-control study, and we don't want to use that unnamed quantity from the last couple of slides, what can we do instead?

Case-control studies use Odds Ratios

What are Odds Ratios?

First, let's define an "odds."

What is an "odds"?

In everyday language, the words "probability" and "odds" are frequently used interchangeably.

Examples:

- "What are the odds that it will rain tomorrow?"
- "What are the odds that this will show up on the exam?"

When we hear questions like this, we usually answer it as if it was asking for a probability. This is generally accepted in daily conversation, but is technically incorrect!

Mathematical definition of an "odds"

If p is a probability of some event, then the odds of that event is defined as:

$$\frac{p}{1-p}$$

It is also often expressed as a ratio in this manner: X:Y where X and Y are integer values.

Examples:

- If the probability that some sports team will win tomorrow's game is 60%, then their odds of winning the game is 3:2
- If the probability of getting a card that will give you the winning poker hand is 10%, then your odds of winning are 1:9

Sidenote: the gambling community is one of the few that actually uses the term "odds" correctly!

The Odds Ratio

Now, the odds ratio is defined as:

$$OR = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}}$$

where:

- p_1 is the probability of observing the exposure among the cases
- p_2 is the probability of observing the exposure among the controls

Note carefully the consistency of what the denominator is, and the study design. Also note the difference between the p_1 and p_2 here vs. in the definition of a relative risk!

The Odds Ratio

So what's the odds ratio for these data?

	bladder cancer		
sweetener use	yes	no	totals
yes	33	25	58
no	67	75	142
totals	100	100	200

The Odds Ratio

A couple of notes:

A shortcut for calculating the odds ratio

The odds ratio on the previous slide is also equal to just:

$$\widehat{OR} = \frac{33 \times 75}{25 \times 76}$$

The odds ratio is symmetric!

In the definition of the odds ratio on Slide 25, we defined p_1 and p_2 as:

- p_1 is the probability of observing the exposure among the cases
- p_2 is the probability of observing the exposure among the controls

because these are the values that are possible to calculate in a case-control design...

The Odds Ratio

The odds ratio is symmetric!

Thus, the odds ratio can be interpreted as: "the increase in the odds of having been exposed, among cases as compared to controls."

However, consider again the bacon data:

	colon cancer		
eat meats	yes	no	totals
yes	13,513	215,725	229,238
no	12,018	228,355	240,373
	25,531	444,080	469,611

The Odds Ratio

Your Turn #2:

- First, calculate the odds ratio according to the definition on Slide 25 (for illustration, don't go straight to the shortcut).
  - This will represent the increase in the odds of being a processed meat eater associated with having colon cancer.

- Then, calculate the odds ratio in the OTHER direction, to find the increase in the odds of having colon cancer, associated with being a processed meat eater.

- Comment briefly on what you observe.

The Odds Ratio

Again, the odds ratio is symmetric!

What this means is that:

- The odds ratio of being exposed comparing cases to controls IS EQUAL to the odds ratio of being a case comparing exposed to unexposed individuals!

- In a case-control study where it is impossible to calculate a relative risk of being a case comparing exposed to unexposed individuals:
  - We can still calculate the odds ratio of being exposed comparing cases to controls!
  - And this is actually mathematically equal to the thing rather know!

The Odds Ratio

It's still not quite the relative risk though...

Note that $OR \neq RR$. Does it matter?

Interpretable Effect Size

In the bladder cancer example, we found $\widehat{OR} \approx 1.3$. What does this mean?

If it had been $RR = 1.3$, it is correct to interpret this to mean that using artificial sweeteners increases your risk of bladder cancer by 30%. What can we say about the odds ratio?

An OR of 1.3 means a 30% increase in the odds. How does this differ from the relative risk?

The Odds Ratio

It's still not quite the relative risk though...

A little example:

Suppose these are cross-sectional data (so the relative risk is valid to calculate):

	disease		
exposure	yes	no	totals
yes	7	4	11
no	3	6	9
	10	10	20

$$\widehat{OR} = \frac{7 \times 6}{4 \times 3} = 3.5$$

$$\widehat{RR} = \frac{7/11}{3/9} \approx 1.91$$

Not close at all!

The Odds Ratio

But, if the disease is rare...

Another little example:

Again suppose these are cross-sectional data (so the relative risk is valid to calculate):

	disease		
exposure	yes	no	totals
yes	5	995	1000
no	3	997	1000
	8	1992	2000

$$\widehat{OR} = \frac{5 \times 997}{995 \times 3} \approx 1.670$$

$$\widehat{RR} = \frac{5/1000}{3/1000} \approx 1.666$$

Very close!

The Odds Ratio

For rare outcomes, $OR \approx RR$

Note that:

$$OR = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}} \qquad vs. \qquad RR = \frac{p_1}{p_2}$$

what happens with p_1 and p_2 are very small?

Recap and Looking Ahead

Recap

- The relative risk (RR) and odds ratio (RR) are both valid measures of association for binary outcome variables
- The RR can only be calculated for cross-sectional and cohort study data
- The OR can be calculated for case-control study data
- The OR is symmetric, and is also a good approximation of the RR when the outcome is rare

Looking Ahead

Logistic Regression

Today's Daily Check

The two Your Turns (one coding, one math)