Statistical Methods for Plant Biology

PBIO 3150/5150

Anirudh V. S. Ruhil

February 23, 2016

The Voinovich School of Leadership and Public Affairs

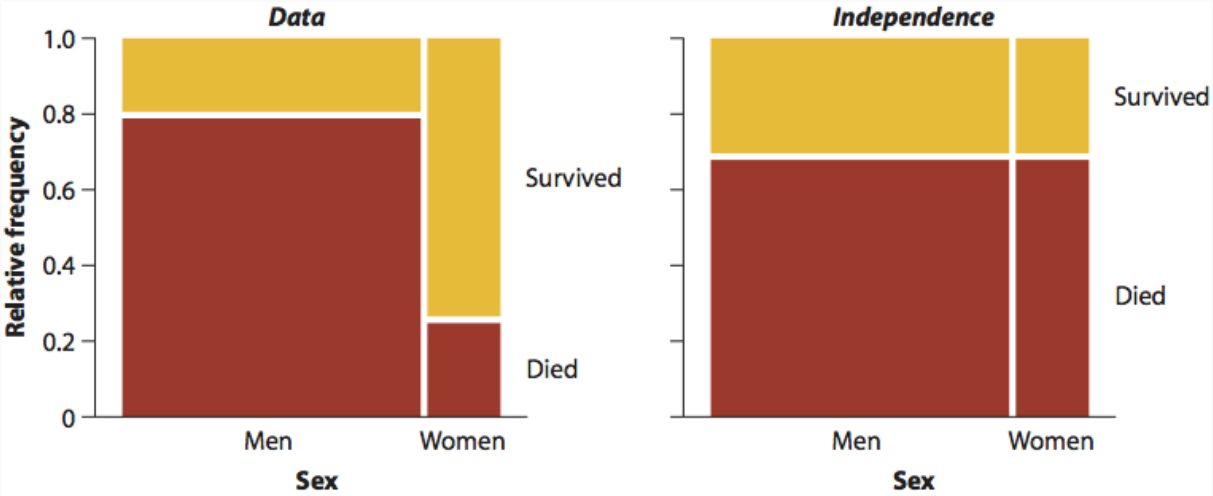
Table of Contents

Association Between 2 Categorical Variables

Contingency Analysis: The Titanic Example

Contingency Analysis is a method of testing for independence between two or more categorical variables

- Is lung cancer independent of smoking?
- Do bright butterflies have the same chance of being eaten as drab butterflies?
- Were women as likely to survive the Titanic sinking as were men?



The chivalry of the seas ... turns out to be an aberration. See Gender, social norms, and survival in maritime disasters

The Odds Ratio (OR)

The Odds: Relative Probabilities of Outcomes

The odds of an event are the probability of a "success" divided by the probability of a "failure"

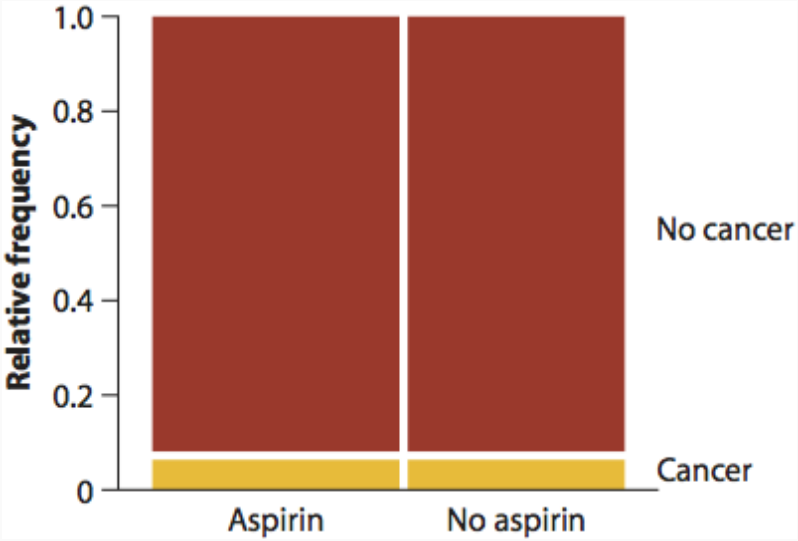
Note: $p =$ probability of success; $1 - p$ = probability of failure. Then, Odds of success: $O = \frac{p}{1-p}$

For e.g., on average 51 boys are born in every 100 births, so the probability of any randomly chosen delivery being that of a boy is $\frac{51}{100} = 0.51$. Likewise the probability of a girl being born are $\frac{49}{100} = 0.49$. Thus the odds of a boy are $\frac{0.51}{0.49} = 1.040816$

- With a sample, we *estimate the odds* and hence $\hat{O} = \frac{\hat{p}}{1-\hat{p}}$
- If the odds of an event are > 1 the event is more likely to happen than not. The odds of an event that is certain to happen are ∞
- If the odds are < 1 the chances are that the event won't happen. The odds of an impossible event are 0

Cancer and Aspirin

	Aspirin	Placebo	Total
Cancer	1438	1427	2865
No Cancer	18496	18515	37011
Total	19934	19942	39876



The mosaic plot suggests taking Aspirin had no effect.

Let success be defined as not getting cancer. Then, for the Aspirin group

$\hat{p}_{ncA} = \frac{18496}{19934} = 0.9279$ and thus $p_{cA} = 1 - \hat{p}_{ncA} = 1 - 0.9279 = 0.0721$

The *odds of success (i.e., not getting cancer)* are $\hat{O}_{ncA} = \frac{0.9279}{0.0721} = 12.87$

i.e., ... "the odds of not getting cancer while taking Aspirin are about 13:1"

What about the Placebo group? Here $\hat{p}_{ncP} = \frac{18515}{19942} = 0.9284$ and thus

$p_{cP} = 1 - \hat{p}_{ncP} = 1 - 0.9284 = 0.0716$. As a result, $\left(\hat{O}_{ncP} = \frac{0.9284}{0.0716} = 12.97\right)$

odds ratio (OR) allows us to compare the odds of success (or failure) for two groups ... $\hat{OR} = \frac{\hat{O}_1}{\hat{O}_2}$. Hence odds-ratio of no cancer for the Aspirin group versus the Placebo group $= \frac{12.87}{12.97} = 0.992$

These data suggest that the odds of cancer are negligibly lower in the Aspirin group than in the placebo group

Table Setup[1]

The contingency table is typically structured as follows:

Outcome	Treatment	Control
Success	a	b
Failure	c	d

Then, $\widehat{OR} = \frac{\frac{a}{c}}{\frac{b}{d}} = \frac{ad}{bc}$

Outcome	Aspirin	Placebo
No Cancer	18496	18515
Cancer	1438	1427

$$\widehat{OR} = \frac{ad}{bc} = \frac{18496 \times 1427}{18515 \times 1438} = 0.9913321$$

If we wanted to focus on Cancer as the outcome of interest, we could do:

Outcome	Aspirin	Placebo
Cancer	1438	1427
No Cancer	18496	18515

$$\widehat{OR} = \frac{ad}{bc} = \frac{1438 \times 18515}{1427 \times 18496} = 1.008744$$

Note a few things:

- $OR = 1$: The odds of success are similar across the groups
- $OR > 1$: The odds of success are higher for the Treatment group
- $OR < 1$: The odds of success are lower for the Treatment group

[1] Note: If any cell $= 0$, add 0.5 to each cell

The Limits of the Odds

Note $p = 0$ means success never occurs, $p = 1$ means success always occurs. Expressing probability as odds yields a corresponding range of values that are anchored below at 0 and above at ∞ ... these are the limits of the odds.

What is strange about this distribution?

p	$1 - p$	$odds$
0.00	1.00	0.00
0.10	0.90	0.11
0.20	0.80	0.25
0.30	0.70	0.43
0.40	0.60	0.67
0.50	0.50	1.00
0.60	0.40	1.50
0.70	0.30	2.33
0.80	0.20	4.00
0.90	0.10	9.00
1.00	0.00	∞

Standard Errors & Confidence Intervals for the OR

Clearly the odds follow an asymmetric distribution and hence require a transformation in order for us to estimate the uncertainty surrounding the odds. That is, the sampling distribution of odds-ratios is highly skewed. Thus we transform the odds ratio into log odds ratio via taking its natural log (i.e., logarithm to the base e)

$$e^{ln(x)} = x \text{ if } x > 0$$

$$ln(e^x) = x$$

For e.g., if $x = c(1, 10, 13)$, then $ln(1) = 0; ln(10) = 2.30; ln(13) = 2.56$

SE of the log odds ratio: $SE\left[ln\left(\hat{OR}\right)\right] = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$

The 95% confidence interval is then given by: $ln\left[\hat{OR}\right] \pm z \times SE\left[ln\left(\hat{OR}\right)\right]$

Calculating standard error for the Aspirin data yields:

$$= \sqrt{\frac{1}{1438} + \frac{1}{1427} + \frac{1}{18496} + \frac{1}{18515}} = 0.03878$$

95% CI: $ln\left[\hat{OR}\right] \pm z \times SE\left[ln\left(\hat{OR}\right)\right]$

$$= ln(0.992) \pm 1.96\,(0.03878) = [-0.084, 0.068]$$

Taking the antilog of each: $\left[e^{-0.084}, e^{0.068}\right] = [0.92, 1.07]$

Interpretation: Since odds-ratio of 1 indicates no difference between the groups, and the 95% CI here includes 1 we cannot reliably conclude that the Aspirin group had lower odds of not getting cancer.

Relative Risk

Relative Risk

Odds-ratios are notoriously difficult for most people to comprehend. Relative Risk – the ratio of the probabilities of an undesirable event occurring for two groups is easier to grasp.

Outcome	Aspirin	Placebo	Total	Aspirin	Placebo	Total
Cancer	a	b	a+b	1438	1427	2865
No Cancer	c	d	c+d	18496	18515	37011
Total	a+c	b+d	a+b+c+d	19934	19942	39876

For Aspirin data, Relative Risk of getting cancer for the two groups is $RR = \frac{\hat{p}_1}{\hat{p}_2} = \frac{1438/19934}{1427/19942} = \frac{0.07213806}{0.07155752} = 1.008113$ slightly higher for the Aspirin group.

Note: For Aspirin and Cancer data the $OR \approx RR$. This will be the case when the outcome of interest is a rare event

Why two measures then, odds-ratios and relative risks? (i) Relative risks are more intuitive for most folk, and (ii) some research designs can use odds-ratios but not relative risks (for e.g., case-control designs)

Case-Control Designs

cohort designs: – identify a group at risk of some outcome and then follow the cohort to see who has the outcome of interest and try to understand why some did and others did not.

case-control studies: – find people with some outcome (the cases), work like a forensic pathologist to figure out possible reasons for the outcome by comparing the cases to others (the controls) who show no sign of the outcome. These designs are not built on random samples and the relevant populations are not well-defined.

Outcome	Exposed	Not Exposed	Total
No Cancer	7	6	13
Cancer	10	56	66
Total	unknown	unknown	79

Cannot calculate RR; don't have total exposed or total not exposed

Can calculate OR because we only need a, b, c, d and have all four values

The χ^2 Contingency Test

Trematode Infection Levels and Fish Eaten by Birds

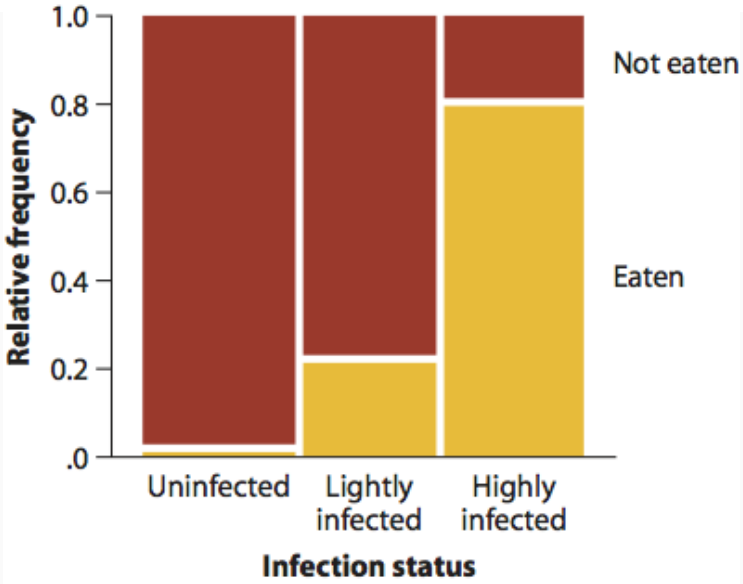
Many parasites have more than one species of host, so the individual parasite must get from one host to another to complete its life cycle. Trematodes of the species Euhaplorchis californiensis use three hosts during their life cycle. Worms mature in birds and lay eggs that pass out of the bird in its feces. The horn snail Cerithidea californica eats these egs, which hatch and grow to another life stage and encysts in the fish's braincase. Finally, when the killifish is eaten by a bird, the worm becomes a mature adult and starts the cycle again.

Researchers have observed that infected fish spend excessive time near the water surface, where they may be more vulnerable to bird predation. This would certainly be to the worm's advantage since it would increase its chances of being ingested by a bird, its next host. Lafferty and Morris (1996) tested the hypothesis that infection influences risk of predation by birds. A large outdoor tank was stocked with three kinds of killifish: unparasitized, lightly infected, and heavily infected. This tank was left open to foraging by birds, especially great egrets, great blue herons, and snowy egrets. The number eaten/not eaten by infection level are shown below:

	High	Light	Uninfected	Total
Not Eaten	9	35	49	93
Eaten	37	10	1	48
Total	46	45	50	141

χ^2 Contingency Test

	High	Light	Uninfected	Total
Not Eaten	9	35	49	93
Eaten	37	10	1	48
Total	46	45	50	141



H_0: Parasite infection and being eaten are independent

H_A: Parasite infection and being eaten are *not* independent

$$\chi^2_{df} = \sum \frac{\left(Observed_{ij} - Expected_{ij}\right)^2}{Expected_{ij}}; \text{df} = (r-1)(c-1)$$

Reject H_0 if $P - value \leq \alpha$; Do Not Reject H_0 otherwise

Assumptions: (i) no cell with expected frequency < 1; (ii) At most 20% of the cells have expected frequency < 5

Expected Frequencies under H_0

$$Expected_{ij} = \frac{\text{Row } i \text{ total} \times \text{Column } j \text{ total}}{Total}$$

If two events A and B are independent then $P(A \text{ and } B) = P(A) \times P(B)$
Therefore, $P(\text{uninfected and eaten}) = P(\text{uninfected}) \times P(\text{eaten})$

$$P(\text{uninfected}) = \frac{50}{141}; P(\text{eaten}) = \frac{48}{141}$$

$$\therefore P(\text{uninfected and eaten}) = \frac{50}{141} \times \frac{48}{141} = 0.1207183$$

Thus the expected frequency under H_0 is $0.1207183 \times 141 = 17.02128$

	Observed			Expected			
	High	Light	Uninfected	High	Light	Uninfected	Total
Not Eaten	9	35	49	30.3	29.7	33.0	93
Eaten	37	10	1	15.7	15.3	17.0	48
Total	46	45	50	46	45	50	141

	Observed			Expected			
	High	Light	Uninfected	High	Light	Uninfected	Total
Not Eaten	9	35	49	30.3	29.7	33.0	93
Eaten	37	10	1	15.7	15.3	17.0	48
Total	46	45	50	46	45	50	141

	High	Light	Uninfected
Not Eaten	$\frac{(9-30.3)^2}{30.3} = 15.01013$	$\frac{(35-29.7)^2}{29.7} = 0.9532525$	$\frac{(49-33)^2}{33} = 7.78324$
Eaten	$\frac{(37-15.7)^2}{15.7} = 29.08213$	$\frac{(10-15.3)^2}{15.3} = 1.846927$	$\frac{(1-17)^2}{17} = 15.08003$

Calculated $\chi_2^2 = 69.7557$ with $P - value = 7.124e - 16$

Reject H_0; the data do not support the notion that the probability of being eaten by birds is independent of infection levels.

Note: Yates' Continuity Correction is not recommended any more. Was/is used because you have a discrete event but the χ^2 distribution is a continuous distribution.

Fisher’s Exact Test

Fisher's Exact Test

Used in 2×2 contingency tables where (1) assumptions of χ^2 are violated, or (2) you have small samples. It assumes random samples

	Cow in estrous	Cow not in estrous	Total
Bitten by vampire bat	15	6	21
Not bitten by vampire bat	7	322	329
Total	22	328	350

Involves calculating the probability of ending up with the observed frequencies as recorded. Computationally intensive because it involves calculating, under the assumption that H_0 is true, all possible 2×2 tables that would yield the same row and column totals.

$$P - value = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$$

In this example the ensuing $P - value = 2.2e - 16$; so we reject H_0
State of estrous and vampire bat attack are not independent.

See here for an example

G-tests

G-tests

Works best with complicated experimental designs, with more than just 2×2 contingency tables, and with large samples

Assumes:

1. Random samples
2. At most 20% of cells have expected frequencies < 5

Employs likelihood ratios (something we will cover in Chapter 20)

$$G = 2 \sum_{c=i}^{c} \sum_{r=j}^{r} Observed_{ij} \times ln \left[\frac{Observed_{ij}}{Expected_{ij}} \right]$$

$$G \sim \chi^2_{df=(r-1)(c-1)}$$

For the infection levels and being eaten by a bird example we would obtain: $G = 77.897$ with $P - value < 2.2e - 16$.

Reject H_0; infection levels and being eaten by a bird are not independent.

What Test Should I Use?

What Test Should I Use?

If it were up to me: Fisher's Exact Test every time I had two categorical variables that were both nominal, and the sample size was not very large. But disciplines and sub-disciplines might want you to use different rules:

- I have ONE categorical variable with two categories – Binomial Test
- I have ONE categorical variable with more than two categories and no assumption is violated – χ^2 Test or the G Test
- I have ONE categorical variable with more than two categories, no assumption is violated and the data came from a complex experimental design – G Test
- I have TWO categorical variables, each with two or more categories
  - Small sample or χ^2 assumptions violated? Fisher's Exact Test
  - Large sample and χ^2 assumptions are violated? G Test
  - Large sample and χ^2 assumptions are not violated? χ^2 Test