Associations betwen categorical variables

Contingency Analysis: Relative Risk, Odds Ratios, and Fisher's Exact Test

October 12, 2021

Contents

Review

Last time we talked about different kinds of contingency analysis and focused on approximate tests for frequency tables. The general question is whether a response variable is likely dependent on a potential explanatory variable. For example:

- *Do smokers have a higher incidence of lung cancer?*
- *Does aspirin lower the incidence of heart attacks?*
- *Do parasites alter the behavior of their hosts?*

There are two kinds of questions we may be interested in, for which different kinds of measures are used:

- Are two variables associated?
- How different are the relative frequencies in a treatment group vs. a control group?
  - Analyze the relative risk or odds ratios between groups.

In both cases, we want to know whether observed differences between groups are likely to arise purely by chance. To answer these questions we can perform two different kinds of analysis:

- Tests for **independence**: do the data show a significant difference from what would be expected if two variables were unrelated?
  - **Exact tests**: compute an exact p-value based on all possible discrete outcomes
  - **Approximate tests**: estimate p-values based on a continuous approximation of a discrete distribution
- **Estimation**: how different is a response variable between two groups of interest?
  - **Relative risk**: relative proportions between two groups
  - **Odds ratio**: relative ***odds*** between two groups

Experimental Designs for Tabular Data

Studies involving categorical data may be performed using different frameworks:

- **Model I**: Observations are classified as they are collected.
  - The total number of trials is chosen, but the classes to which they are assigned will vary.
  - In this case, the row and column totals can both vary freely.
- **Model II**: Totals for one group are fixed, while the other varies.
  - Usually it is the explanatory variable that is chosen and the dependent variable is measured
  - For example, a certain number of subjects are given aspirin vs. a placebo, and the number of heart attacks in each group is then measured.
  - In this case, the number of observations is fixed in one dimension, so the row or column totals are fixed while the other varies.
- **Model III**: Totals for both groups are fixed.
  - This type of design is not commonly used in biology.
  - The classic example is Fisher's "lady tasting tea" study, in which the number of cups of tea in which the milk was added before or after the tea were predetermined in order to test whether a lady could tell the difference between these, as she claimed.

Measures of association

There are two common methods to estimate the magnitude of differences between groups. Which one is appropriate depends on the experimental design.

- **Relative risk**: used for **observational** studies and **experimental** studies in which ***random samples*** of the population are subjected to two different treatments.
- **Odds ratio**: used for observational and **case-control** studies where the number of subjects in the case and control groups are fixed.

These measures are similar when the frequency of an outcome of interest, called the ***focal outcome***, is small in the population.

Relative risk

Relative risk is defined as the ratio of the proportion of subjects with a focal outcome in one group vs. the other group.

This measure ***cannot be used in case-control studies*** where the number of subjects in two groups is fixed (a Model II or III design), since the relative frequencies do not reflect their natural occurrence in a population. This will make more sense if we look at some examples.

W&S Example 9.2: *Around of half of 39,876 women are randomly assigned to receive 100mg of aspirin every other day for 10 years, while the other half received a placebo. Did long-term aspirin treatment significantly reduce the incidence of invasive cancer?*

We can estimate the magnitude of any potential effect of aspirin treatment by looking at the proportions of women in each group who developed cancer. This is the relative risk (RR):

$$\widehat{RR} = \frac{\hat{p}_1}{\hat{p}_2}$$

Note that since this is an **estimate** of proportions in the population at large, we cover the proportions and the RR with "hats".

In the aspirin example, the RR was 1.007. Since the proportions are almost the same, it is unlikely that aspirin has any effect on cancer incidence.

Confidence intervals Since RR is a ratio of two random variables, it is not normally distributed, and so we cannot use the regular proceduring involving Z-scores to compute a 95% CI.

However, if we take the **natural log** of RR, the distribution is **approximately normal**. This transformation turns a ratio into an equation involving simple addition/subtraction.

Since our proportions are random variables that are randomly sampled from the population, and since random variables resulting from addition or subtraction of two normal distributions are also normally distributed, we can use a normal approximation for the resulting distribution.

This allows us to compute a standard error for $ln(\widehat{RR})$, and we can then use this to compute a 95%CI for this using a Z-score of 1.96:

$$ln(RR) - 1.96 * SE_{ln(RR)} < ln(RR) < ln(RR) + 1.96 * SE_{ln(RR)}$$

To get the 95% CI for the RR, we just need to reverse the procedure by exponentiation:

$$e^{ln(RR)-1.96*SE_{ln(RR)}} < RR < e^{ln(RR)+1.96*SE_{ln(RR)}}$$

For this example, we obtain $0.94 < \widehat{RR} < 1.08$. This is a very narrow distribution around RR=1, confirming that there is likely no effect.

Reduction in risk We can make statements about how much snaller the risk is in one group vs. the other by restating the RR as follows:

- **Relative** reduction in risk $= 1 - RR = 1 - \frac{p_1}{p_2} = \frac{p_2 - p_1}{p_2}$
  - *How much smaller risk is in one (treatment) group as a proportion of the risk in the other (control) group.*
- **Absolute** reduction in risk $= p_1 - p_2$
  - *Absolute difference in risk between groups.*
  - This may be very small if the overall risk is low in both groups, even though the reduction in relative risk could be large. – indicating that even though there is a measurable difference in risk, the incidence in both groups is so small as to not have an important effect overall.

Odds ratios

The **odds** of a particular outcome is just the probability of a "focal" outcome, vs. the probability that this outcome did not occur. Why would we want to use this instead of just the probability?

Because the ratio of the odds in two different groups, or the **odds ratio**, gives us some idea of the frequency of an outcome in one group relative to that in another group, while **controlling** for the possibility that the number of items sampled from each group is **not representative** of the population at large.

This is particularly useful when we are interested in studying risk factors in a group that is under-represented in the population at large, e.g. people with a rare disease. If we can't get a large enough sample of these people by random sampling, then it is difficult to make conclusions about how different factors affect them relative to those who don't have the disease.

The OR is often written as Θ (though not in W&S). It is defined as:

$$\hat{\Theta}_{1,2} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

where p_1 is the probability of the focal outcome in group 1 and p_2 is its probability in group 2.

For a 2x2 contingency table, we can write:

$$\hat{\Theta}_{1,2} = \frac{\hat{p}_{11}/\hat{p}_{12}}{\hat{p}_{21}/\hat{p}_{22}} = \frac{\hat{p}_{11}\hat{p}_{22}}{\hat{p}_{12}\hat{p}_{21}} = \frac{x_{11}x_{22}}{x_{12}x_{21}}$$

... since the denominators cancel out.

Confidence intervals The 95% CI for odds ratios is computed in the same manner as for RR, using log-transformed values, since both involve division of two random variables:

$$e^{ln(OR)-1.96*SE_{ln(OR)}} < OR < e^{ln(OR)+1.96*SE_{ln(OR)}}$$

Relative Risk or Odds Ratio?

- When samples are ***chosen at random*** in the population, divided into groups based on a ***potential causal factor***, and then the frequency of a particular ***focal outcome*** is examined in each group, the RR and OR will produce very similar results.
- However, RR cannot be used for **case-control** studies in which individuals are ***pre-selected*** based on known focal outcomes, and potential underlying causal factors are then examined for association with each known outcome.
  - This is because the samples were not chosen at random from the population, and therefore the proportions of individuals with each focal outcome do not reflect the proportions in the population at large.

⇒ ***Question: Say we want to determine if getting a COVID vaccine helps protect people against getting sick with this coronavirus. Which of the following designs could we use to calculate the relative risk?***

Design 1: Pick 1000 COVID patients and 1000 healthy people and compute the relative risk of getting sick, given that they'd already gotten the vaccine.

Design 2: Pick 2000 people, 1000 vaccinated and 1000 unvaccinated. Ask what proportion in each group got sick with the virus.

Answer

- The relative proportions of people who got sick **given that they'd already been vaccinated (or not)** would not be meaningful because the total number of sick vs. healthy people was chosen arbitrarily.

$$RR = \frac{(Vac \cap COVID)/(Total\ Vac)}{(Unvac \cap COVID)/(Total\ Unvac)} = \frac{a/(a+c)}{b(b+d)}$$

If we'd only been able to find 200 people with COVID, then the relative risk would change.

In contrast, the odds ratio would not change significantly because the proportions of people who got sick in each group would stay about the same:

$$OR = \frac{(COVID \cap Vac)/(COVID \cap Unvac)}{(Healthy \cap Vac)/(Healthy \cap Unvac)} = \frac{a/c}{b/d} = \frac{ad}{bc}$$

Notice that the OR will also not change depending on the orientation of the table.

Fisher's Exact Test for Independence

Last time we talked about the χ^2 test for independence, which compares **observed vs. expected** values to estimate the probability of obtaining a result at least as extreme as the observed data. This approach uses a **continuous distribution** to **approximate** the probability. When the expected frequency of any cell in a contingency table is too small, an exact test is preferred.

Fisher's test provides an **exact** p**-value** for contingency tables by enumerating **all possible discrete outcomes** that are at least as extreme as the observed data (i.e. as or more different from neutral expectation) and and computing the total probability of these.

- It is preferred to the χ^2 test for all 2x2 contingency tables.
- It should be used instead of the χ^2 test for independence for cases where some of the **frequencies** in the contingency table are **too low to satisfy the assumptions** of the χ^2 test.
- R can also compute Fisher's exact p-value using larger tables. When computationally feasible, use Fisher's over χ^2.

Formula

Fisher's test assumes **fixed row margins** (row totals), which are needed to compute the the probability of values at least as extreme as the observed frequencies relative to expectation under the rule of independence, i.e. $P(A \cap B) = P(A) * P(B)$ for each cell in the contingency table.

With fixed marginal counts, the count in one cell, x_{ij}, will determine the counts in the other three cells (see Aho, Section 11.6.3 for details).

The probability of the observed data is simply the product of factorials of all the marginals, divided by the product of factorials of each cell and factorial of total number.

$$P(a, b, c, d) = \frac{(a+b)!\,(a+c)!\,(b+d)!\,(c+d)!}{a!b!c!d!N!}$$

where $N = a + b + c + d$.

The p-value will be the proportion of all possible ways to get a value **equal to or more extreme** than the observed count in one of the cells (typically x_{11}).

Example

We can perform a Fisher's exact test using the `fisher.test()` function. Let's go back to the example of cancer incidence among women who first gave birth under or over the age of 30. We will choose a one-tailed test, since we expect the Case group (women with cancer) to have a higher proportion of women who first gave birth over 30.

```
Case    = c(683,2537)   # total =  3220
Control = c(1498,8747)  # total = 10245
data_matrix = rbind(Case, Control)
colnames(data_matrix) = c("Above30","Below30")
```

```r
# kable is a table generator from the knitr package
kable(data_matrix)
```

	Above30	Below30
Case	683	2537
Control	1498	8747

```r
# full result from Fisher's Exact Test
fisher_test = fisher.test(data_matrix, alternative='greater')
fisher_test
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  data_matrix
## p-value < 2.2e-16
## alternative hypothesis: true odds ratio is greater than 1
## 95 percent confidence interval:
##  1.442384      Inf
## sample estimates:
## odds ratio
##   1.571925
```

```r
# p-value from Fisher's Exact Test
ftest_pval = fisher_test$p.value
ftest_pval
```

```
## [1] 3.526441e-18
```

```r
# table orientation is arbitrary; both orientations give the same result
fisher.test(t(data_matrix), alternative='greater')$p.value
```

```
## [1] 3.526441e-18
```

The p-value computed in this way is not exactly the same as we get from the other two methods we used, but they are all extremely significant.

Note that the orientation of the table does not matter – the test gives the same result either way.

Odds Ratios and 95% CI

The output for Fisher's exact test include a reference to the **odds ratio** and a 95% confidence interval. Under independence, the odds ratio should equal 1, so if one row has higher or lower counts than expected then the OR will be greater than or less than 1.

You can accees these values directly from the result of Fisher's test:

```r
# OR
fisher_test$estimate
```

```
## odds ratio
##   1.571925
```

```r
# CI (level)
fisher_test$conf.int
```

```
## [1] 1.442384      Inf
## attr(,"conf.level")
## [1] 0.95
```

In this example, the null hypothesis is that there is no difference between groups: $H_0 : \Theta = 1$. The alternative hypothesis is $H_A : \Theta > 1$.

Since OR $= 1.57$, and the 95% CI is $1.44 < \Theta < \infty$, we can conclude that the incidence of cancer is not the same between these two groups.

Fisher's Test and the Hypergeometric Distribution

Fisher's exact test turns out to be a special case of the **hypergeometric distribution**, which gives the probability of x successes when sampling **without** replacement. Fisher's test follows this model because, with ***fixed row margins***, changing the value of one cell in a contingency table necessarily changes the values in the others.

The effects are are particularly pronounced for small total numbers.

The hypergeometric distribution is commonly used to test for over- or under-representation of functional annotations in two sets of genes. We will cover this distribution in more detail in a future class.

Example

Since Fisher's Exact Test follows a kind of hypergeometric distribution, we should get the same result if we compute a p-value empirically using the `hyper()` function family.

Let's use the `phyper()` function to get a p-value for the observed data from the breast cancer example above. We will use the upper tail of a one-tailed test, since we are asking if the incidence of cancer is higher in the >30 group.

Recall that since this is a discrete distribution, we will need to subtract 1 from the observed value in order to get $p(x) \geq obs$.

```r
# phyper(q, m, n, k, lower.tail = FALSE)
# phyper(a, a+b, c+d, a+c, lower.tail = FALSE)
#  q = observation (a   = cancer and above 30)
#  m = white balls (a+b = cancer, a.k.a. "success")
#  n = black balls (c+d = normal, a.k.a. "failure")
#  k = total draws (a+c = above 30)
phyper(683 - 1, 3220, 10245, 2181, lower.tail = FALSE)
```

```
## [1] 3.526441e-18
```

Below I have rewritten the same computations for the Fisher's Exact and hypergeometric tests using R's terminology for the `hyper` family of functions:

```
# data (variable names chosen to match dhyper() argument names)
x = 683    # Case_Above30 = a
m = 3220   # Case         = row 1 margin = a+b
n = 10245  # Control      = row 2 margin = c+d
k = 2181   # Above30      = col 1 margin = a+c

# same as 'data_matrix' used above
# (although table orientation doesn't matter for the Fisher exact test)
matrix(c(x,   m-x,      # a, b
         k-x, n-(k-x)), # c, d
       2,2,
       byrow = TRUE) # matrices fill by columns by default
```

```
##      [,1] [,2]
## [1,]  683 2537
## [2,] 1498 8747
```

```
# Fisher test, alternative = 'greater'
fisher.test(matrix(c(x, m-x, k-x, n-(k-x)),2,2), alternative='greater')$p.value
```

```
## [1] 3.526441e-18
```

```
q = x-1  # upper tail for discrete distribution
phyper(q, m, n, k, lower.tail = FALSE)
```

```
## [1] 3.526441e-18
```

This is exactly the same result we obtained using Fisher's test!

R commands for different kinds of frequency tests

To summarize, R contains commands for computing statistics for comparing categorical variables to expected discrete distributions, as well as for testing for and measuring associations between categorical variables:

- Goodness-of-fit
  - `binom.test()`
  - `chisq.test()`
- Independence
  - `riskratio()`
  - `oddsratio()`
  - `fisher.test()`
  - `chisq.test()`

Converting between lists and tables

Often our data will not be pre-formatted as a contingency table, so we will need to transform it in order to use these functions. The most common functions are `table()` and `xtable()`, which allow you to create 2x2 or larger tables. You can also cross-tabulate on one or more variables.

Some short tutorials for making tables and converting between data frames and tables are available here:

- R Cookbook
- Quick R

Additional Resources

- **Whitlock & Schluter Online Tutorial**
  - Contingency Analysis