STATISTICS 305/605: Intro to Biostatistics

September 25, 2021

Lecture 10: Contingency Table Tests

(Reading: PB Section 15.1)

1 Goals of Lecture

- We measure association between two binary variables with RR and OR
  - Confidence intervals can also function as tests
- Sometimes we want a formal p-value for a test of association
- Sometimes the table is larger than 2×2
- We now expand to larger tables, define association more broadly, and show two major tests for it.

2 Larger tables

- So far, we have focused our learning on 2×2 tables
  - By far the most common categorical variables are binary
  - Simplest case, so makes sense to start there anyway
- We have developed an understanding of conditional probability and related topics
  - Call the row variable X and the column variable Y
  - Conditional probability
    - Treat each row (or column) of the table (each level of X) as a separate binary variable
    - Probability of success and failure for Y within that row add to 1
    - These are conditional probabilities of success and failure for Y , GIVEN the level of the row variable X

Table 1: Outcome of Covid-hospitalized patients in Georgia, USA, in March 2020 against age groups

		Outcome			
		Discharged	Still Hosp	Died	Total
Age Group	18–49	85	1	3	89
	50–64	83	7	9	99
	≥ 65	65	16	36	117
Total		233	24	48	305

    - They form the conditional distribution of both (all) of the outcomes for Y, given the row outcome X.
  - Association
    - Let p_1 be the conditional probability of success for row 1 ($P(Y = 1|X = 1)$).
      - $1 - p_1$ is the conditional probability of failure
    - Let p_2 be the conditional probability of success for row 2 ($P(Y = 1|X = 2)$).
      - $1 - p_2$ is the conditional probability of failure
    - Association between X and Y means that $p_1 \neq p_2$
      - A unit's probability of success changes depending on their level of X
  - Measures of association
    - Relative Risk $RR = p_1/p_2$
    - Odds ratio $OR = \dfrac{[p_1/(1-p_1)]}{[p_2/(1-p_2)]}$

- Now we want to allow for larger tables
  - Generically, tables can have R rows and C columns
    - Index these with $r = 1, 2, \ldots, R$ and $c = 1, 2, \ldots, C$
  - How do we measure/test association in these larger tables?

Example: Outcome vs. age for hospitalized Covid patients in Georgia (Lecture 10 Scripts.R)

Early in the Covid pandemic, little was known about risk factors and mortality rates from the disease. A paper published in May 2020[1] details the outcomes of patients hospitalized due to Covid during the month of March 2020. In one part of the paper, the authors cross-tabulate the patient outcomes—discharged, still hospitalized (after 3 weeks) or dead—against age groups 18–49, 50–65, and ≥ 65. The data are in Table 1.

[1] Gold, et al. (2020) "Characteristics and Clinical Outcomes of Adult Patients Hospitalized with COVID-19 — Georgia, March 2020" *Morbidity and Mortality Weekly Report* 69: 545–550.

- Cross tabulations of any size, $R \times C$, are called CONTINGENCY TABLES
  - The table consists of CELLS in the interior and MARGINS around the outside
- Very often we care about conditional probabilities of column responses within a row
  - GIVEN row r, what are the probabilities on outcomes $c = 1, 2, \ldots, C$
    - This is the conditional distribution of Y given $X = r$.
  - Estimate these probabilities by taking the cell counts in row r and dividing by the marginal count for that row
  - Can reverse this and compute conditional distribution of X for each column Y
  - Which conditional distributions are computed depend on what you're trying to say about the table

Example: Outcome vs. age for hospitalized Covid patients in Georgia (Lecture 10 Scripts.R)

The researchers are interested in comparing the chances of different outcomes given the age groups, so we will estimate the conditional probabilities of columns given rows here. First we create the data. Here I do something different. When data are categorical rather than numerical, R automatically stores the levels in alphabetical order, which is not always the order we want. In this case, I want my table to match Table 1. So I have to tell R what order the levels should be stored in. This is achieved using the `levels=` argument in the `factor()` function. I start by creating objects that hold the names of the levels in the order that I want them. Then within `factor()` I create the data by replicating the age names three times each and replicating the whole set of hospitalization outcomes three times. Then I fix the order of the levels by listing it in `levels=`. The rest of the code is stuff we have seen before: create a data frame from the objects, cross-tabulate using `xtabs()`, and get conditional proportions using `prop.table()`. The results are below.

```
> agenames = c("18-49", "50-64", "65+")
> hospnames = c("Disch", "StillIn", "Died")
>
> age.gp = factor(rep(agenames, each=3), levels=agenames)
> outcome = factor(rep(hospnames, times=3), levels=hospnames)
> counts = c(85, 1, 3, 83, 7, 9, 65, 16, 36)
>
> georgia = data.frame(age.gp, outcome, counts)
> georgia
  age.gp outcome counts
1  18-49   Disch     85
2  18-49 StillIn      1
3  18-49    Died      3
4  50-64   Disch     83
5  50-64 StillIn      7
```

```
6  50-64    Died      9
7    65+   Disch     65
8    65+ StillIn     16
9    65+    Died     36
>
> tab = xtabs(formula=counts ~ age.gp + outcome)
> tab
       outcome
age.gp  Disch StillIn Died
  18-49    85       1    3
  50-64    83       7    9
  65+      65      16   36
>
> # Conditional probability estimates given age group
> prop.table(x=tab, margin="age.gp")
       outcome
age.gp       Disch    StillIn       Died
  18-49 0.95505618 0.01123596 0.03370787
  50-64 0.83838384 0.07070707 0.09090909
  65+   0.55555556 0.13675214 0.30769231
```

For the youngest people, 95.5% were discharged within 3 weeks, compared to 83.8% of "middle-age" and 55.5% of older people. The trends on death are the opposite: 3.3% of hospitalized patients in the youngest group died, whereas 30.7% in the oldest group died. So early on we already could see the pattern of more severe impact of this disease with age.

---

3 Measuring and Testing Association in Larger Tables

- We would like to measure and test the association in larger tables than just 2×2
  - It turns out that there are no good extensions of RR and OR for larger tables!
  - Can only compute them on 2x2 subtables or on aggregated counts from 2x2 partitionings
    - "Subtable": pick 2 rows and 2 columns from full table
    - "Partitioning": split table into 2 parts horizontally and vertically and combine counts
  - Can get complicated with larger R and C
- A simpler approach—but also less informative—is test the null hypothesis that the two variables are "independent"

- Two variables are INDEPENDENT if the conditional distribution of one of the variables is exactly the same for all levels of the other variable
  - *Independence in our example means that the conditional probabilities of any outcome are the same across all age groups*
    1. *Conditional probability of discharge for younger people = conditional probability of discharge for middle-age = conditional probability of discharge for older*
    2. *Also for still in hospital*
    3. *Also for dead*
- If two variables are *not* independent, then they are ASSOCIATED
  - The full distributions of one variable change across levels of the other variable
  - Doesn't have to be all levels, doesn't have to be all conditional probabilities

- To test for independence (or, equivalently, to test for association), we first define what X and Y are in the problem. Then,
  - $H_0 : X$ and Y are independent
  - $H_A : X$ and Y are *not* independent (i.e., they are associated)

Pearson Test

- The most popular test is the PEARSON (CHI-SQUARED) TEST
  - Assume independence is true.
    - Then all conditional distributions are the same
    - They all match their respective marginal distributions
  - Can estimate what we expect cell counts would have been under this assumption
  - If the OBSERVED cell course are sort of consistent with the EXPECTED cell counts, then there is no evidence against the null hypothesis of independence.
  - If the OBSERVED cell course are quite different from the EXPECTED cell counts, then there *is* evidence against the null hypothesis of independence, suggesting association.
  - The PEARSON STATISTIC is a calculate that measures the aggregate closeness of the observed (O) and expected (E) counts,

$$X^2 = \sum_{\text{all } r,c} \frac{(O_{rc} - E_{rc})^2}{E_{rc}}$$

  - When H_0 is true, and the sample size is "large enough," then X^2 has a chi-squared distribution with $(R-1)(C-1)$ degrees of freedom
    - Approximation to actual sampling distribution, sort of like normal approximation for sample means

    - "large enough"...see later
  - If the p-value is smaller than chosen α then H_0 is rejected, and we conclude that the two variables are not independent (or ARE associated)
  - Otherwise, if the p-value is α or larger, then we fail to reject H_0 and we conclude that there is no evidence to suggest that the variables are associated
    - They *may be* independent, or they may be associated but the differences are not large enough that they were detected by these data.

Example: Outcome vs. age for hospitalized Covid patients in Georgia (Lecture 10 Scripts.R)

We test the null hypothesis that outcome after hospitalization is independent of age group, against the alternative that they are associated. In R, the `chisq.test()` function does the Pearson test. It's pretty easy to use. The results are below.

```
> aa = chisq.test(x=tab, correct = FALSE)
> aa

        Pearson's Chi-squared test

data:  tab
X-squared = 50.089, df = 4, p-value = 3.459e-10
```

The test statistic is $X^2 = 50.1$ and the p-value is 0.0000000003, crazy tiny! We strongly reject H_0 and can conclude that there is an association between the age group of a hospitalized patient and the outcome of their hospitalization.

Fisher's Exact Test

- The Pearson test statistic has a sampling distribution that depends on the sizes of the "expected" cell counts
  - The approximation is good as long as all of these are reasonably large
    - Good if all are at least 5
    - Pretty good if all are at least one and some are above 5
    - These numbers are rough guidelines and not fixed rules.
  - When there are small expected counts, the p-values may not always be very accurate.
- An alternative test for smaller samples is Fisher's Exact Test
  - Uses a complicated "permutation" argument to compute a p-value

- "Exact" means that the p-value is exactly correct, *assuming that the marginal counts would not change if a new sample were taken*
  - Hardly ever true
  - So it's just another approximate p-value, but often a better one for small samples.
- The complicated process of finding a p-value sometimes fails on tables that have many rows, columns, and/or observations.
  - In this case, it is possible to estimate the p-value using computer simulation
    - The simulated p-value is not necessarily very accurate, but can reasonably accurately determine whether to reject the null hypothesis

Example: Outcome vs. age for hospitalized Covid patients in Georgia (Lecture 10 Scripts.R)

We rerun the test for independence using Fisher's exact test. In R, the `fisher.test()` function does the test. It's pretty easy to use. The simulated p-value comes from the argument `simulate.p.value=TRUE`. The results are below.

```
> fisher.test(x=tab)
        Fisher's Exact Test for Count Data
data:  tab p-value = 7.549e-11 alternative hypothesis: two.sided
> fisher.test(x=tab, simulate.p.value = TRUE)

        Fisher's Exact Test for Count Data with simulated p-value
        (based on 2000 replicates)

data:  tab
p-value = 0.0004998
alternative hypothesis: two.sided
```

There is no test statistic, but the p-value is also very tiny, 7.5×10^{-11}. That's 10 zeroes and a 7. Same conclusion as with Pearson test. The simulated p-value is not needed here, since the algorithm worked. Running it anyway gives very different results, but still rejects the null hypothesis of independence.

How can we know when the expected counts are low? The `chisq.test()` function will print a warning any time this happens. You can then check the expected counts by looking at the `expected` element of the test object. I do this for the original data, and then create a new data set with much smaller counts by pretending that the data we had were actually only 1/10 as many people. I just create `tab.small` by dividing the whole table by 10 and rounding. I then run the Pearson test and the Fisher test. (By the way, this is just a demonstration. There is no reason to create smaller tables in an actual analysis!)

```
> # What were the expected counts from this table?
> aa$expected
       outcome
age.gp     Disch  StillIn     Died
  18-49 67.99016 7.003279 14.00656
  50-64 75.62951 7.790164 15.58033
  65+   89.38033 9.206557 18.41311
>
> # For demonstration, create a new table by
> #  dividing this table by 10 and run the tests
> tab.small = round(tab/10)
> tab.small
       outcome
age.gp  Disch StillIn Died
  18-49     8       0    0
  50-64     8       1    1
  65+       6       2    4
> bb = chisq.test(x=tab.small, correct = FALSE)
Warning message:
In chisq.test(x = tab.small, correct = FALSE) :
  Chi-squared approximation may be incorrect
> bb

        Pearson's Chi-squared test

data:  tab.small
X-squared = 6.6606, df = 4, p-value = 0.1549

> bb$expected
       outcome
age.gp     Disch StillIn     Died
  18-49 5.866667     0.8 1.333333
  50-64 7.333333     1.0 1.666667
  65+   8.800000     1.2 2.000000
> fisher.test(x=tab.small)

        Fisher's Exact Test for Count Data

data:  tab.small
p-value = 0.1565
alternative hypothesis: two.sided
```

First, we see the expected cell counts from the original table. They range from 7 to 89, all comfortably above 5, and we can have confidence that the chi-squared approximation for the Pearson statistic gives a reasonably accurate p-value. But when I run `chisq.test()`

on `tab.small`, I get output with a warning that just says, "`In chisq.test(x = tab.small, correct = FALSE) : Chi-squared approximation may be incorrect`". The test's p-value is now 0.15, which is mostly because we reduced the sample size so much that we now have little power to detect association. Checking the expected counts, I see that 6/9 are below 5, and one is as low as 0.8. So I run Fisher's exact test and come up with a different approximate p-value of...0.16! So in this example, there is not much of a change in the p-value resulting from running the two different tests. However, in some cases, it makes a much larger difference. So when it is known that sample sizes will be small, it is recommended to use Fisher instead of Pearson.

Interpreting tests

- Testing independence vs. association yields very limited information
  - Conclusion is yes/no about whether there is evidence to support association
  - Knowing that variables ARE associated doesn't tell you *how* they are associated
    - Which rows' conditional probabilities for which column outcomes are different from one another?
    - Tests tell you nothing about this
- Tests are therefore often followed by deeper analysis
  - Multinomial or Poisson regression models (see STAT 475)
  - RR or OR on subtables and/or collapsed tables.
- Also, the test results tell you nothing about cause-and-effect!
  - Even if there is a direction to the relationship, there may be may confounders
    - *We can't say that it is specifically age that causes worse outcomes with Covid. Maybe many other factors that happen with advancing age are the real drivers.*
  - "Association" is a very general concept

4 What to learn from this

1. $R \times C$ contingency tables are just like 2×2 tables, just bigger
   (a) We have the same questions about associations and conditional probabilities
2. Association measures like relative risks and odds ratios can only be applied to binary outcomes
   (a) Need to reduce table somehow into 2×2 if we want to use them

3. Hypothesis tests for independence vs. association are somewhat popular
   (a) Not especially informative, though
   (b) End up needing further analysis if whole story is to be revealed.

5 Exercises

Use R for all calculations, unless otherwise specified.

1. The csv file, "MariDance.csv" contains data from a survey of high school students asking questions about certain behaviours. Two of the questions were about frequency of marijuana use (variable `marijUse` with levels `1`=Never, `2`= <1 per Month, `3`= ≥1 per Month, and `4`= ≥1 per Day), and how often they attended parties and dances away from school (variable `party` with levels `1`=Not at all, `2`=Sometimes, and `3`=Often). The file contains columns for these two variables, along with the number of students in each combination, `numStdnt`.

   (a) Read this csv file into R using `read.csv()`. Give the object some name. **Print out the data object.** The output should look more or less like the spreadsheet.

   (b) Use `xtabs()` to create a table from the object. **Print the table.**

   (c) Estimate the conditional probability of marijuana use levels, given party attendance levels. **Show results and comment on the pattern: which students seem to use marijuana more or use it less?**

   (d) Estimate the conditional probability of party attendance levels, given marijuana use levels. **Show results and comment on the pattern: which students seem attend parties more or less?**

   (e) Perform the Pearson test to see whether marijuana use and party attendance are associated. **Show the results. State the hypotheses, test statistic, p-value, and conclusions about the question.**

   (f) Notice that `chisq.test()` gives a warning.

      i. **Print the expected values and determine why the warning was given.**

      ii. **Does it seem like this might seriously affect how reliable the chi-squared approximation is? Explain.**

   (g) When I run Fisher's exact test, I get the warning below:

```
> fisher.test(tab)
Error in fisher.test(tab) :
  FEXACT error 6.  LDKEY=618 is too small for this problem,
  (ii := key2[itp=657] = 3614700, ldstp=18540)
Try increasing the size of the workspace and possibly 'mult'
```

   This is a bizarre way of saying that the problem is too complicated for R to run quickly. Run the function with the simulation approximation to the p-values instead. **Print the results. Is the conclusion the same as with the Pearson test?**

   (h) Which of the following best describes what we can conclude from this analysis, at least for the study population represented by these students? (Choose 1)

i. Marijuana use and party attendance seem to be independent of one another
ii. Increasing marijuana use causes students to attend parties more
iii. Increasing party attendance causes students to use marijuana more
iv. Marijuana use and party attendance are associated somehow, but we can't really say in what way
v. Marijuana use and party attendance *seem* like they are associated, but they are probably not.