# Contingency Tables: Part Two\*

STA 312: Fall 2022

# Suggested Reading: Chapter 2

•	Read	Section	2.6	about	Fisher's	exact	test
---	------	---------	-----	-------	----------	-------	------

• Read Section 2.7 about multi-dimensional tables and Simpson's paradox.

# Contents

1	Testing for the Product Multinomial	1
2	Fisher's Exact Test	3
3	Tables of Higher Dimension	7

# 1 Testing for the Product Multinomial

# Testing Association for the Product Multinomial

Prospective and retrospective designs Prospective design:

- A conditional multinomial in each row
- I independent random samples, one for each value of X
- Likelihood is a product of *I* multinomials
- Null hypothesis is that all *I* sets of conditional probabilities are the same.

A retrospective design is just like this, but with rows and columns reversed.

<sup>\*</sup>See last slide for copyright information.

Null hypothesis is no differences among the I vectors of conditional probabilities

	Attack	Stroke	Both	Neither	Total
Drug					$n_{1+}$
Drug and Exercise					$n_{2+}$
Total	$n_{+1}$	$n_{+2}$	$n_{+3}$	$n_{+4}$	n

- Both  $n_{1+}$  and  $n_{2+}$  are fixed by the design. They are sample sizes.
- Under  $H_0$ , MLE of the (common) conditional probability is the marginal sample proportion:

$$\widehat{\pi}_{ij} = p_{+j} = \frac{n_{+j}}{n}$$

• And the expected cell frequency is just

$$\widehat{\mu}_{ij} = n_{i+} \,\widehat{\pi}_{ij} = n_{i+} \,\frac{n_{+j}}{n} = \frac{n_{i+}n_{+j}}{n}$$

#### Expected frequencies are the same!

For testing both independence and testing equal conditional probabilities,

$$\widehat{\mu}_{ij} = \frac{n_{i+}n_{+j}}{n}.$$

The degrees of freedom are the same too. For the product multinomial,

- There are I(J-1) free parameters in the unconstrained model.
- There are J-1 free parameters under the null hypothesis.
- $H_0$  imposes I(J-1) (J-1) = (I-1)(J-1) constraints on the parameter vector.

• So 
$$df = (I-1)(J-1)$$
.

	Attack	Stroke	Both	Neither	Total
Drug					$n_{1+}$
Drug and Exercise					$n_{2+}$
Total	$n_{\pm 1}$	$n_{+2}$	$n_{\pm 3}$	$n_{\pm 4}$	n

#### This is very fortunate

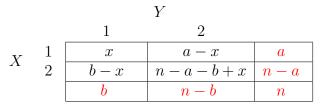
- The cross-sectional, prospective and retrospectives are different from one another conceptually.
- The multinomial and product-multinomial models are different from one another technically.
- But the tests for relationship between explanatory and response variables are 100% the same.
- Same expected frequencies and same degrees of freedom.
- Therefore we get the same test statistics and *p*-values.

# 2 Fisher's Exact Test

## Fisher's Exact Test

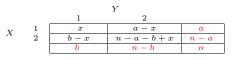
- Everything so far is based on large-sample theory.
- What if the sample is small?
- Fisher's exact test is good for  $2 \times 2$  tables.
- There are extensions for larger tables.

## Fisher's exact test is a permutation test



- Think of a data file with 2 columns, X and Y, filled with ones and twos.
- X has a ones and Y has b ones.
- Calculate the estimated odds ratio  $\hat{\theta}$ .
- If X and Y are unrelated, all possible pairings of X and Y values should be equally likely.
- There are n! ways to order the X values, and for each of these, n! ways to order the Y values.

#### Idea of a permutation test



- There are  $(n!)^2$  ways to arrange the X and Y values.
- For what fraction of these is the (estimated) odds ratio
  - Greater than or equal to  $\widehat{\theta}$  (Upper tail *p*-value)
  - Less than or equal to  $\hat{\theta}$  (Lower tail *p*-value)

For a 2-sided test, add the probabilities of all the tables less likely than or equally likely to the one we have observed. (This is what R does.)

Nice idea, but hard to compute. Fisher thought of it *and* simplified it.

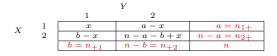
## Let us count together



- The *n*! permutations of 1s and 2s have lots of repeats that look the same.
- There are  $\binom{n}{a}$  ways to choose which cases have X = 1.
- For each of these, there are  $\binom{n}{b}$  ways to choose which cases have Y = 1.
- So the total number of  $2 \times 2$  tables with *n* observations,  $n_{1+} = a$  and  $n_{+1} = b$  is  $\binom{n}{a}\binom{n}{b}$ .
- Of these, the number of ways to get the values in the table is just the multinomial coefficient

$$\binom{n}{x \ a-x \ b-x \ n-a-b+x} = \frac{n!}{x!(a-x)!(b-x)!(n-a-b+x)!}$$

#### Hypergeometric probability



Dividing the number of ways to get  $n_{11} = x$  by the total number of equally likely outcomes,

$$P(n_{11} = x) = \frac{\begin{pmatrix} x & a-x & b-x & n-a-b+x \end{pmatrix}}{\binom{n}{a}\binom{n}{b}}$$

$$= \frac{\frac{n!}{x!(a-x)!(b-x)!(n-a-b+x)!}}{\frac{n!}{a!(n-a)!}\frac{n!}{b!(n-b)!}}$$

$$= \frac{\binom{n}{x}\binom{n-a}{b-x}}{\binom{n}{b}}$$

$$= \frac{\binom{n_{1+}}{n_{11}}\binom{n_{2+}}{n_{1+1}-n_{11}}}{\binom{n}{n+1}}$$
(Eq. 2.11, p. 46)

#### Adding up the probabilities

Always remembering that a, b and n are fixed

v

- Fortunately,  $\theta(x)$  is an increasing function of x (differentiate).
- So, tables with larger x values than the one observed also have greater sample odds ratios. Add  $P(n_{11} = x)$  over x to get tail probabilities.
- Range of *x*:
  - $-x \leq \min(a, b)$
  - $-n_{22} = n a b + x \ge 0$ , so  $x \ge a + b n$ .
  - Thus, x ranges from  $\max(0, a + b n)$  to  $\min(a, b)$ .

#### Example: Sinking of the the Titanic

```
> # help(Titanic)
> dimnames(Titanic)
$Class
[1] "1st" "2nd" "3rd" "Crew"
$Sex
[1] "Male" "Female"
$Age
[1] "Child" "Adult"
$Survived
[1] "No" "Yes"
> # Women in 1st class vs Women in crew
>
> ladies = Titanic[c(1,4),2,2,]
Just the ladies
> ladies
     Survived
Class No Yes
```

```
Class No res
1st 4 140
Crew 3 20
> 140/144 # Rich ladies
[1] 0.9722222
> 20/23 # Cleaning ladies
[1] 0.8695652
> X2 = chisq.test(ladies,correct=F); X2
Warning message:
In chisq.test(ladies, correct = F) :
Chi-squared approximation may be incorrect
```

Pearson's Chi-squared test data: ladies X-squared = 5.2043, df = 1, p-value = 0.02253

#### Check the expected frequencies

```
> X2$expected
     Survived
            No
Class
                      Yes
 1st 6.0359281 137.96407
 Crew 0.9640719 22.03593
>
> fisher.test(ladies)
Fisher's Exact Test for Count Data
data: ladies
p-value = 0.05547
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
0.03027561 1.41705937
sample estimates:
odds ratio
0.1935113
```

#### Conclusion

Though a higher percentage of women in first class survived than female crew, it could have been due to chance.

# Fisher's exact test makes sense even without the pretending we have a random sample

You could say

- Assume that status on the ship for these women (First Class passenger vs. crew) is fixed. It was what it was.
- Survival also was what it was.
- Given this, is the observed *pairing* of status and survival an unusual one?
- That is, for what fraction of the possible pairings is the status difference in survival as great or greater than the one we have observed?
- A little over 5%? That's a bit unusual, but perhaps not *very* unusual.
- There is not even any need to talk about probability.

# 3 Tables of Higher Dimension

## Tables of Higher Dimension: Conditional independence

- Suppose X and Y are related.
- Are X and Y related *conditionally* on the value of W?
- One sub-table for each value of W.
- X and Y can easily be related unconditionally, but still be conditionally independent.
- Example: Among adults 18 and older, X =Tattoos and Y =Grey hair.
- Need a 3-way table, showing the relationship of tattoos and grey hair separately for each age group.
- Speak of the relationship between X and Y "controlling for" W, or "allowing for" W.

# Was UC Berkeley discriminating against women?

Data from the 1970s

Data in a 3-dimensional array: Variables are

- Sex of the person applying for graduate study
- Department to which the person applied
- Whether or not the person was admitted

## Berkeley data

```
> #
    More than one Explanatory Variable at once
    data() to list the nice data sets that come with R #
> #
> #
    help(UCBAdmissions)
                                              #
> dim(UCBAdmissions)
[1] 2 2 6
> dimnames(UCBAdmissions)
$Admit
[1] "Admitted" "Rejected"
$Gender
[1] "Male"
          "Female"
$Dept
[1] "A" "B" "C" "D" "E" "F"
> # Look at gender by admit.
> # Apply sum to rows and columns, obtaining the marginal freqs.
> sexadmit = apply(UCBAdmissions,c(1,2),sum)
```

### Sex by Admission

```
> sexadmit
         Gender
Admit
          Male Female
  Admitted 1198
                 557
 Rejected 1493 1278
> sexadmit = t(sexadmit); sexadmit
       Admit
Gender
        Admitted Rejected
 Male
            1198
                    1493
 Female
             557
                     1278
> rowmarg = apply(sexadmit,1,sum); rowmarg
 Male Female
 2691
        1835
> percentadmit = 100 * sexadmit[,1]/rowmarg ; percentadmit
   Male
          Female
44.51877 30.35422
```

It certainly looks suspicious.

#### Test sex by admission

```
> chisq.test(sexadmit,correct=F)
Pearson's Chi-squared test
data: sexadmit
X-squared = 92.2053, df = 1, p-value < 2.2e-16
> fisher.test(sexadmit)  # Gives same p-value
Fisher's Exact Test for Count Data
data: sexadmit
p-value < 2.2e-16
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
    1.621356 2.091246
sample estimates:
    odds ratio
    1.840856</pre>
```

#### But look at the whole table

```
> UCBAdmissions
, , Dept = A
```

Gender

Admit Male Female Admitted 512 89 Rejected 313 19 , , Dept = B Gender Admit Male Female Admitted 353 17 Rejected 207 8

# Berkeley table continued

, , Dept = CGender Admit Male Female Admitted 120 202 Rejected 205 391 , , Dept = DGender Admit Male Female Admitted 138 131 Rejected 279 244

# Berkeley table continued some more

, , Dept = EGender Admit Male Female Admitted 53 94 Rejected 138 299 , , Dept = FGender Male Female Admit Admitted 22 24 Rejected 351 317

#### Look at Department A

```
Gender Admitted Rejected
Male 512 313
Female 89 19
> JustA[1,1]/sum(JustA[1,]) # Men
[1] 0.6206061
> JustA[2,1]/sum(JustA[2,]) # Women
[1] 0.8240741
> chisq.test(UCBAdmissions[,,1],correct=F)
Pearson's Chi-squared test
data: UCBAdmissions[, , 1]
X-squared = 17.248, df = 1, p-value = 3.28e-05
```

Women are more likely to be admitted.

## Summarize analyses of sub-tables

Just the code, for reference

```
# Summarize analyses of sub-tables: Loop over departments
# Sum of chi-squared values in X2
ndepts = dim(UCBAdmissions)[3]
gradschool=NULL; X2=0
for(j in 1:ndepts)
    ſ
   dept = dimnames(UCBAdmissions)$Dept[j] # A B C etc.
   tabl = t(UCBAdmissions[,,j]) # All rows, all cols, level j
   Rowmarg = apply(tabl,1,sum)
   Percentadmit = round( 100*tabl[,1]/Rowmarg ,1)
   per = round(Percentadmit,2)
   Test = chisq.test(tabl,correct=F)
   tstat = round(Test$statistic,2); pval = round(Test$p.value,5)
   gradschool = rbind(gradschool,c(dept,Percentadmit,tstat,pval))
   X2 = X2+Test$statistic
   } # Next Department
colnames(gradschool) = c("Dept","%MaleAcc","%FemAcc","Chisq","p-value")
noquote(gradschool) # Print character strings without quote marks
```

#### Simpson's paradox

> noquote(gradschool) # Print character strings without quote marks

	Dept	%MaleAcc	%FemAcc	Chisq	p-value
[1,]	А	62.1	82.4	17.25	3e-05
[2,]	В	63	68	0.25	0.61447
[3,]	С	36.9	34.1	0.75	0.38536
[4,]	D	33.1	34.9	0.3	0.58515
[5,]	Е	27.7	23.9	1	0.31705
[6,]	F	5.9	7	0.38	0.53542

#### Overall test of conditional independence

Add the chi-squared values and add the degrees of freedom.

```
> # Overall test of conditional independence
> names(X2) = "Pooled Chi-square"
> df = ndepts ; names(df)="df"
> pval=1-pchisq(X2,df)
> names(pval) = "P-value"
> print(c(X2,df,pval))
Pooled Chi-square df P-value
19.938413378 6.00000000 0.002840164
```

Conclusion: Gender and admission are *not* conditionally independent. From the preceding slide, we see it comes from Department A's being more likely to admit women than men.

#### Track it down

Make a table showing Department, Number of applicants, Percent female applicants and Percent of applicants admitted.

```
> # What's happening?
> whoapplies = NULL
> for(j in 1:ndepts)
+
      {
      dept = dimnames(UCBAdmissions)$Dept[j]; names(dept) = "Dept"
+
      tabl = t(UCBAdmissions[,,j]) # All rows, all cols, level j
+
+
      nj = sum(tabl); names(nj)=" n "
      mf = apply(tabl,1,sum); femapp = round(100*mf[2]/nj,2)
+
      succ = apply(tabl,2,sum); getin = round(100*succ[1]/nj,2)
+
      whoapplies = rbind(whoapplies,c(dept,nj,femapp,getin))
+
      } # Next Department
+
```

Now it's in a table called whoapplies.

#### The explanation

> noquote(whoapplies)

```
Dept n Female Admitted
[1,] A 933 11.58 64.42
[2,] B 585 4.27 63.25
[3,] C 918 64.6 35.08
[4,] D 792 47.35 33.96
[5,] E 584 67.29 25.17
[6,] F 714 47.76 6.44
```

Departments with a higher acceptance rate have a higher percentage of male applicants.

# Does this mean that the University of California at Berkeley was *not* discriminating against women?

- By no means. Why does a department admit very few applicants relative to the number who apply?
- Because they do not have enough professors and other resources to offer more classes.
- This implies that the departments popular with men were getting more resources, relative to the level of interest measured by number of applicants.
- Why? Maybe because men were running the show.
- The "show," definitely includes the U. S. military, which funds a lot of engineering and similar stuff at big American universities.

## Some uncomfortable truths

- Especially for non-experimental studies, statistical analyses involving just one explanatory variable at a time can be very misleading.
- When you include a new variable in an analysis, the results could get weaker, they could get stronger, or they could reverse direction all depending upon the inter-relations of the explanatory variables and the response variable.
- Failing to include important explanatory variables in observational studies is a common source of bias.
- Ask: "Did you control for ..."

#### At least it's a start

- We have seen one way to "control" for potentially misleading variables (sometimes called "confounding variables").
- It's *control by sub-division*, in which you examine the relationship in question separately for each value of a control variable or variables.
- We have a good way of pooling the tests within each level of the control variable, to obtain a test of conditional independence.
- There's also model-based control, which is coming next.

## **Copyright Information**

This document was prepared by Jerry Brunner, Department of Statistics, University of Toronto. It is licensed under a Creative Commons Attribution - ShareAlike 3.0 Unported License. Use any part of it as you like and share the result freely. The LATEX source code is available from the course website: http://www.utstat.toronto.edu/brunner/oldclass/312f22