Contingency Tables Part Two\(^1\)
STA 312: Fall 2012

\(^1\)See last slide for copyright information.
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.
Overview

1. Testing for the Product Multinomial

2. Fisher’s Exact Test

3. Tables of Higher Dimension
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.
Null hypothesis is no differences among the \( I \) vectors of conditional probabilities

<table>
<thead>
<tr>
<th></th>
<th>Attack</th>
<th>Stroke</th>
<th>Both</th>
<th>Neither</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>( n_{1+} )</td>
</tr>
<tr>
<td>Drug and Exercise</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>( n_{2+} )</td>
</tr>
<tr>
<td>Total</td>
<td>( n_{+1} )</td>
<td>( n_{+2} )</td>
<td>( n_{+3} )</td>
<td>( n_{+4} )</td>
<td>( n )</td>
</tr>
</tbody>
</table>

- 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:

\[
\hat{\pi}_{ij} = p_{+j} = \frac{n_{+j}}{n}
\]

- And the expected cell frequency is just

\[
\hat{\mu}_{ij} = n_{i+} \hat{\pi}_{ij} = n_{i+} \frac{n_{+j}}{n} = \frac{n_{i+} n_{+j}}{n}.
\]
Expected frequencies are the same!

For testing both independence and testing equal conditional probabilities,

\[
\hat{\mu}_{ij} = \frac{n_i + n_j}{n}.
\]

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 \( df = (I - 1)(J - 1) \).

<table>
<thead>
<tr>
<th></th>
<th>Attack</th>
<th>Stroke</th>
<th>Both</th>
<th>Neither</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>( n_{1+} )</td>
</tr>
<tr>
<td>Drug and Exercise</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>( n_{2+} )</td>
</tr>
<tr>
<td>Total</td>
<td>( n_{+1} )</td>
<td>( n_{+2} )</td>
<td>( n_{+3} )</td>
<td>( n_{+4} )</td>
<td>( n )</td>
</tr>
</tbody>
</table>
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.
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

<table>
<thead>
<tr>
<th>X</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>x</td>
<td>a − x</td>
</tr>
<tr>
<td></td>
<td>b − x</td>
<td>n − a − b + x</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>n − b</td>
</tr>
</tbody>
</table>

- 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 $\hat{\theta}$ (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

<table>
<thead>
<tr>
<th>X</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>a-x</td>
<td>a</td>
</tr>
<tr>
<td>b-x</td>
<td>n-a-b+x</td>
<td>n-a</td>
</tr>
<tr>
<td>b</td>
<td>n-b</td>
<td>n</td>
</tr>
</tbody>
</table>

- 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

$$\binom{n}{x \ a-x \ b-x \ n-a-b+x} = \frac{n!}{x!(a-x)!(b-x)!(n-a-b+x)!}.$$
Testing for the Product Multinomial Fisher’s Exact Test Tables of Higher Dimension

Hypergeometric probability

<table>
<thead>
<tr>
<th>$X$</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x$</td>
<td>$a - x$</td>
<td>$a = n_{1+}$</td>
</tr>
<tr>
<td>$b - x$</td>
<td>$n - a - b + x$</td>
<td>$n - a = n_{2+}$</td>
</tr>
<tr>
<td>$b = n_{1+}$</td>
<td>$n - b = n_{2+}$</td>
<td>$n$</td>
</tr>
</tbody>
</table>

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

$$P(n_{11} = x) = \frac{{x \choose a-x} {b-x \choose n-a-b+x}}{n \choose a} \frac{n}{n \choose b}$$

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

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

$$= \frac{(n_{1+}) \choose n_{11} \choose (n_{2+}) \choose n_{11} \choose n_{2+} \choose \frac{n_{1+} - n_{11}}{n_{1+} - n_{11}}}$$

(Eq. 2.11, p. 46)
Adding up the probabilities
Always remembering that $a$, $b$ and $n$ are fixed

<table>
<thead>
<tr>
<th>$X$</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$b - x$</td>
<td>$n - a - b + x$</td>
<td>$n - a$</td>
</tr>
<tr>
<td>$b$</td>
<td>$n - b$</td>
<td>$n$</td>
</tr>
</tbody>
</table>

- 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 $x$:  
  - $x \leq \min(a, b)$
  - $n_{22} = n - a - b + x \geq 0$, so $x \geq a + b - n$.
  - Thus, $x$ ranges from $\max(0, a + b - n)$ to $\min(a, b)$.
Example: Sinking of 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
  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

```r
> X2$expected
   Survived
Class  No  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.
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

```r
> # 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

\[
\begin{array}{c|cc}
\text{Gender} & \text{Male} & \text{Female} \\
\hline
\text{Admit} & 1198 & 557 \\
\text{Rejected} & 1493 & 1278 \\
\end{array}
\]

> sexadmit = t(sexadmit); sexadmit

\[
\begin{array}{c|cc}
\text{Admit} & \text{Gender} & \text{Male} & \text{Female} \\
\hline
\text{Admitted} & 1198 & 1493 \\
\text{Rejected} & 557 & 1278 \\
\end{array}
\]

> rowmarg = apply(sexadmit,1,sum); rowmarg

\[
\begin{array}{c|c}
\text{Gender} & \text{Male} \quad \text{Female} \\
\hline
2691 & 1835 \\
\end{array}
\]

> percentadmit = 100 * sexadmit[,1]/rowmarg; percentadmit

\[
\begin{array}{c|c}
\text{Gender} & \text{Male} \quad \text{Female} \\
\hline
44.51877 & 30.35422 \\
\end{array}
\]

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
But look at the whole table

```
> UCBAAdmissions
, , Dept = A

<table>
<thead>
<tr>
<th>Gender</th>
<th>Admit</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Admitted</td>
<td>512</td>
<td>89</td>
</tr>
<tr>
<td></td>
<td>Rejected</td>
<td>313</td>
<td>19</td>
</tr>
</tbody>
</table>

, , Dept = B

<table>
<thead>
<tr>
<th>Gender</th>
<th>Admit</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Admitted</td>
<td>353</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>Rejected</td>
<td>207</td>
<td>8</td>
</tr>
</tbody>
</table>
Berkeley table continued

, , Dept = C

<table>
<thead>
<tr>
<th>Gender</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Admit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Admitted</td>
<td>120</td>
<td>202</td>
</tr>
<tr>
<td>Rejected</td>
<td>205</td>
<td>391</td>
</tr>
</tbody>
</table>

, , Dept = D

<table>
<thead>
<tr>
<th>Gender</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Admit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Admitted</td>
<td>138</td>
<td>131</td>
</tr>
<tr>
<td>Rejected</td>
<td>279</td>
<td>244</td>
</tr>
</tbody>
</table>
Berkeley table continued some more

Dept = E

<table>
<thead>
<tr>
<th>Gender</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Admitted</td>
<td>53</td>
<td>94</td>
</tr>
<tr>
<td>Rejected</td>
<td>138</td>
<td>299</td>
</tr>
</tbody>
</table>

Dept = F

<table>
<thead>
<tr>
<th>Gender</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Admitted</td>
<td>22</td>
<td>24</td>
</tr>
<tr>
<td>Rejected</td>
<td>351</td>
<td>317</td>
</tr>
</tbody>
</table>
Look at Department A

> # Just Department A
> JustA = t(UCBAdmissions[,1]); JustA
  Admit
  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

```r
# 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)
{
    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")
nouquote(gradschool)  # Print character strings without quote marks
```
### Simpson’s paradox

```
> noquote(gradeschool) # Print character strings without quotes
```

<table>
<thead>
<tr>
<th>Dept</th>
<th>%MaleAcc</th>
<th>%FemAcc</th>
<th>Chisq</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1,] A</td>
<td>62.1</td>
<td>82.4</td>
<td>17.25</td>
<td>3e-05</td>
</tr>
<tr>
<td>[2,] B</td>
<td>63</td>
<td>68</td>
<td>0.25</td>
<td>0.61447</td>
</tr>
<tr>
<td>[3,] C</td>
<td>36.9</td>
<td>34.1</td>
<td>0.75</td>
<td>0.38536</td>
</tr>
<tr>
<td>[4,] D</td>
<td>33.1</td>
<td>34.9</td>
<td>0.3</td>
<td>0.58515</td>
</tr>
<tr>
<td>[5,] E</td>
<td>27.7</td>
<td>23.9</td>
<td>1</td>
<td>0.31705</td>
</tr>
<tr>
<td>[6,] F</td>
<td>5.9</td>
<td>7</td>
<td>0.38</td>
<td>0.53542</td>
</tr>
</tbody>
</table>
Overall test of conditional independence

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

```r
> # 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   6.000000000  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.
Make a table showing Department, Number of applicants, Percent female applicants and Percent of applicants admitted.

```r
> # 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)

<table>
<thead>
<tr>
<th>Dept</th>
<th>n</th>
<th>Female</th>
<th>Admitted</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1,]</td>
<td>A</td>
<td>933</td>
<td>11.58</td>
</tr>
<tr>
<td>[2,]</td>
<td>B</td>
<td>585</td>
<td>4.27</td>
</tr>
<tr>
<td>[3,]</td>
<td>C</td>
<td>918</td>
<td>64.6</td>
</tr>
<tr>
<td>[4,]</td>
<td>D</td>
<td>792</td>
<td>47.35</td>
</tr>
<tr>
<td>[5,]</td>
<td>E</td>
<td>584</td>
<td>67.29</td>
</tr>
<tr>
<td>[6,]</td>
<td>F</td>
<td>714</td>
<td>47.76</td>
</tr>
</tbody>
</table>

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.
This slide show 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/312f12