Lecture 3: Contingency Tables

Author: Nick Reich / transcribed by Josh Nugent and Nutcha Wattanachit

Course: Categorical Data Analysis (BIOSTATS 743)

Contingency Tables (notation)

Imagine we have random variables X and Y. Let X and Y be categorical variables with I and J categories, respectively. We can draw a contingency table with i rows and j columns:

	$Y=1$	$Y=2$	$\cdots$	$Y=i$	$\cdots$	$Y=J$	
$X=1$	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1J}	n_{1+}
$X=2$	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2J}	n_{2+}
$\vdots$	$\cdots$	$\cdots$	$\ddots$	$\cdots$	$\cdots$	$\cdots$	$\vdots$
$X=i$	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{iJ}	n_{i+}
$\vdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\ddots$	$\cdots$	$\vdots$
$X=I$	n_{I1}	n_{I2}	$\cdots$	n_{Ij}	$\cdots$	n_{IJ}	n_{I+}
	n_{+1}	n_{+2}	$\cdots$	n_{+j}	$\cdots$	n_{+J}	n

The generic entry n_{ij} denotes the count of observations when $X=i$ and $Y=j$. This table also represents the joint distribution of X and Y when X and Y are categorical.

Contingency Tables (notation and example)

The notation n_{+j} is the sum of the j^{th} column: $n_{+j} = \sum n_{ij}$ over all i, while n_{i+} is the sum of the i^{th} row: $n_{i+} = \sum n_{ij}$ over all j. The total is simply $n = \sum_{ij} n_{ij}$.

Contingency tables are ubiquitous in scientific papers and in the media. Example:

Table I. Baseline HIV status by baseline variables.

	Negative	Positive	Total
Population total	43,269	1493	44,762
Gender			
Male	19,646	527	20,173
Female	23,623	966	24,589

Contingency Tables - Probabilities

A contingency table can be represented by probabilities as well. Define π_{ij} to be a the population parameter representing the true probability of being in the ij^{th} cell - the probability that both $X = i$ and $Y = j$. Formally, $\pi_{ij} = Pr(X = i, Y = j)$, the joint probability of X and Y for all $i = 1, ..., I$ and $j = 1, ..., J$.

	$Y=1$	$Y=2$	$\cdots$	$Y=i$	$\cdots$	$Y=J$	
$X=1$	π_{11}	π_{12}	$\cdots$	π_{1j}	$\cdots$	π_{1J}	π_{1+}
$X=2$	π_{21}	π_{22}	$\cdots$	π_{2j}	$\cdots$	π_{2J}	π_{2+}
$\vdots$	$\cdots$	$\cdots$	$\ddots$	$\cdots$	$\cdots$	$\cdots$	$\vdots$
$X=i$	π_{i1}	π_{i2}	$\cdots$	π_{ij}	$\cdots$	π_{iJ}	π_{i+}
$\vdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\ddots$	$\cdots$	$\vdots$
$X=I$	π_{I1}	π_{I2}	$\cdots$	π_{Ij}	$\cdots$	π_{IJ}	π_{I+}
	π_{+1}	π_{+2}	$\cdots$	π_{+j}	$\cdots$	π_{+J}	π

Contingency Tables - Marginal Probabilities

We can also use these probabilities to find the marginal probability distributions of X and Y: $\pi_{i+} = P(X = i)$ and $\pi_{+j} = P(Y = j)$.

	$Y = 1$	$Y = 2$	$\cdots$	$Y = i$	$\cdots$	$Y = J$	
$X = 1$	π_{11}	π_{12}	$\cdots$	π_{1j}	$\cdots$	π_{1J}	π_{1+}
$X = 2$	π_{21}	π_{22}	$\cdots$	π_{2j}	$\cdots$	π_{2J}	π_{2+}
$\vdots$	$\cdots$	$\cdots$	$\ddots$	$\cdots$	$\cdots$	$\cdots$	$\vdots$
$X = i$	π_{i1}	π_{i2}	$\cdots$	π_{ij}	$\cdots$	π_{iJ}	π_{i+}
$\vdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\ddots$	$\cdots$	$\vdots$
$X = I$	π_{I1}	π_{I2}	$\cdots$	π_{Ij}	$\cdots$	π_{IJ}	π_{I+}
	π_{+1}	π_{+2}	$\cdots$	π_{+j}	$\cdots$	π_{+J}	π

Contingency Tables - Conditional Probabilities

Finally, we can use our contingency table to find conditional probabilies - the probability that you are in the j^{th} cell, conditioned on being in the i^{th} row:

$$\pi_{j|i} = \frac{\pi_{ij}}{\pi_{i+}} = Pr(Y = j|X = i)$$

Another way of saying this is the probability of outcome j if i is already known. Example: Probability that you have arthritis, if you are over 65.

Contingency Tables - Conditional Distributions

Conditional distributions of Y given $X = i$,

$$\left\{\pi_{1|i}, \pi_{2|i}, \pi_{3|i}, ..., \pi_{j|i}\right\}$$

are key in modeling. There are similarities here to a regression-like problem, where we are trying to describe an outcome variable as a function of a predictor variable. This is similar to the conditional formulation of $E[Y|X]$ in regression where we are modeling an outcome of Y conditional on observed X. Stay tuned for more on this later in the course.

Sampling Methods - Poisson

How do we obtain our data for a contingency table? Different sampling methods will call for different analyses, so we need to understand the key features of the methods.

1. Poisson Sampling:

- The total **n** is not fixed; theoretically unbounded
- Counts are generally observed in a fixed time interval
- Cell counts $n_{ij} \overset{ind}{\sim} Poisson(\mu_{ij})$

Example 1: Observations of how many people are using Mac/PC/Linux operating systems over the course of an hour at three different locations on campus:

	Mac	PC	Linux
DuBois			
SciLi			
Arnold			

n

Sampling Methods - Multinomial (fixed n)

2. Multinomial Sampling (fixed n)

- Row/column totals are not fixed, but the overall n is.
- Multinomial distribution with $I * J$ categories.

For example, we could make a table of counts of operating systems among the people in a single classroom at UMass by age:

	Mac	PC	Linux	total
30 or older				
Under 30				
total				n=16

- Counts distributed $n_{ij} \sim Multinomial(n, \pi)$, where π is a vector of all probabilities for the cells in the table.

Another example: Cohort study. Enroll 5,000 people and track them over a year. Measure video game playing (none / low / high) and hospitalizations for repetitive strain injuries.

Sampling Methods - Multinomial (fixed row n / column n)

3. Multinomial Sampling (fixed row or column n)

- Both n and n_i (or n_j) are known for all i (or j).
- Example: We want to know the video-game playing habits of Republicans, Democrats, and Independents. We survey 500 in each group.

	None	Low	High	total
Dem				500
Rep	n_{21}	n_{22}	n_{23}	500
Ind				500

Sampling Methods - Multinomial (fixed row n / column n)

- With fixed row totals, the vector of counts in each row will follow a multinomial distribution based on the row totals:

$$\begin{pmatrix} n_{i1} \\ n_{i2} \\ \vdots \\ n_{ij} \end{pmatrix} \sim \text{Multinomial}(n_{i+}, \vec{\pi}) \text{ with } \vec{\pi} = \begin{pmatrix} \pi_{1|i} \\ \pi_{2|i} \\ \vdots \\ \pi_{j|i} \end{pmatrix}$$

Another example: A case-control study:

	Case	Control
SE_1	. . .	. . .
SE_2	. . .	. . .
SE_3	. . .	. . .
Total	1000	1000

The χ^2 Statistic

The chi-squared test statistic compares the observed and expected values for all count observations in the cells of a table:

$$\text{Test Statistic} = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

* E_{ij} is the expected value for each cell based on its probability under the null hypothesis
* O_{ij} is the value observed for that cell
* We reject the H_0 if our χ^2 statistic is large - a larger TS means larger deviations from expected counts
* Test is 2-sided by virtue of $(...)^2$
* Compare to a $(1-\alpha)$%ile of the χ^2_{df} distribution with appropriate degrees of freedom

Goodness of Fit χ^2 Test

$$H_0 : \pi_1 = \pi_{0_1}, \pi_2 = \pi_{0_2}, ..., \pi_k = \pi_{0_k}$$

This answers the question: Does a set of counts follow a specified (H_0) distribution?

n = total # of observations
$E_i = \pi_{0_i} \cdot n$
df $= k - 1$

Simply evaluate using test statistic above: $\sum \frac{(O-E)^2}{E}$

Note: $\pi_1 = \pi_2 = ... = \pi_k$ is a special case.

Fisher's Exact Test for Independence

For small sample sizes, methods like this "use *exact* small-sample distributions rather than large sample approximations." (Agresti p. 90-92) For 2x2 tables, the test looks at all possible combinations of outcomes under

$$H_0 : \text{variables independent}$$

to determine whether the observed outome is unusual enough to reject the null.

Conditioning on both sets of marginal totals, we have

$$p(t) = p(n_{11} = t) = \frac{\begin{pmatrix} n_{1+} \\ t \end{pmatrix} \begin{pmatrix} n_{2+} \\ n_{+1} - t \end{pmatrix}}{\begin{pmatrix} n \\ n_{+1} \end{pmatrix}}$$

Note: Becuase this is a 2x2 table with row and column totals known, n_{11} determines the other three counts.

Fisher's Exact Test: Example

Muriel Bristol, a colleague of the esteemed statistician Sir Ronald Fisher, claimed that she could, upon tasting a cup of tea mixed with milk, divine whether the cup had had milk poured in before the tea, or tea before the milk. They performed an experiment to test her claim...

Fisher's Exact Test: Example

	Guess Poured First		
Poured First	*Milk*	*Tea*	Total
Milk	3	1	4
Tea	1	3	4
Total	4	4	

In this case, the P-value for Fisher's exact test is the null probability of this table and of tables having even more evidence in favor of her claim.

The observed table, $t_0 = 3$ correct choices, has null probability

$$\binom{4}{3}\binom{4}{1}/\binom{8}{4} = 0.229.$$

The only table more extreme in the direction of H_a is $n_{11} = 4$, which has a probability of 0.014. The P-value is $P(n_{11} \geq 3) = 0.243$. Despite the underwhelming evidence in this test, Bristol did eventually convince Fisher that she could tell the difference.

For more, see Agresti p. 90 - 92.

Birth Order and Gender

We have data on 1,000 2-child families. It is typically thought that birth order/gender of two offspring from the same parents are i.i.d. $\sim$ Bernoulli(0.5). So, we can fill in the expected values: 250 for each group.

$$H_0 : \pi_1 = \pi_2 = \pi_3 = \pi_4 = 0.25$$

$$H_a : \text{at least 1 } \pi_i \neq 0.25$$

First Child	M		F		
Second Child	M	F	M	F	Totals
Count	218	227	278	277	1000
Expected Value	250	250	250	250	1000

Birth Order and Gender

Chi-squared test for this data:

$$TS = \frac{\sum_{k=1}^{4} (n_{obs} - n_{exp})^2}{n_{exp}}$$

$$n = 1000,\ k = 4, df = \text{k-1} = 3$$

$$\alpha = 0.05$$

```
n_obs <- c(218, 227, 278, 277)
n_exp <- c(250, 250, 250, 250)
(ts <- sum((n_obs-n_exp)^2/n_exp))
```

```
## [1] 12.264
```

```
(pval <- 1 - pchisq(ts, df = 3))
```

```
## [1] 0.006531413
```

Birth Order and Gender

With a p-value of 0.0065, we reject the null hypothesis at a significance level of 0.05. We have evidence to suggest that at least one $\pi_k \neq 0.25$.