L17: Chapter 10: Contingency tables I

Department of Statistics, University of South Carolina

Stat 205: Elementary Statistics for the Biological and Life Sciences

Two-sample binary data

- In Chapter 9 we looked at one sample & looked at observed vs. "expected under H_0."
- Now we consider two populations and will want to compare two population proportions p_1 and p_2.
- In population 1, we observed y_1 out of n_1 successes; in population 2 we observed y_2 out of n_2 successes.
- This information can be placed in a contingency table

		Group 1	Group 2
Outcome	Success	y_1	y_2
	Failure	$n_1 - y_1$	$n_2 - y_2$
	Total	n_1	n_2

- $\hat{p}_1 = y_1/n_1$ estimates p_1 & $\hat{p}_2 = y_2/n_2$ estimates p_2.

Example 10.1.1 Migraine headache

- Migraine headache patients took part in a double-blind clinical trial to assess experimental surgery.
- 75 patients were randomly assigned to real surgery on migraine trigger sites ($n_1 = 49$) or sham surgery ($n_2 = 26$) in which an incision was made but nothing else.
- The surgeons hoped that patients would experience "a substantial reduction in migraine headaches," which we will label as success.

Example 10.1.1 Migraine headache

Table 10.1.1 Response to migraine surgery

		Surgery	
		Real	Sham
Substantial reduction	Success	41	15
in migraine headaches?	No success	8	11
	Total	49	26

- $\hat{p}_1 = 41/49 = 83.7\%$ for real surgeries.
- $\hat{p}_2 = 15/26 = 57.7\%$ for sham surgeries.
- Real appears to be better than sham, but is this difference significant?

Example 10.1.2 HIV testing

A random sample of 120 college students found that 9 of the 61 women in the sample had taken an HIV test, compared to 8 of the 59 men.

Table 10.1.3 HIV testing data

	Female	Male
HIV test	9	8
No HIV test	52	51
Total	61	59

- $\hat{p}_1 = 9/61 = 14.8\%$ tested among women.
- $\hat{p}_2 = 8/59 = 13.6\%$ tested among men.
- These are pretty close.

Conditional probabilities

- p_1 and p_2 are *conditional probabilities*. Remember way back in Section 3.3?
- For the migraine data, $p_1 = \text{pr}\{\text{success}|\text{real}\}$ and $p_2 = \text{pr}\{\text{success}|\text{sham}\}$. $\hat{p}_1 = 0.84$ and $\hat{p}_2 = 0.58$ estimate these conditional probabilities.
- For the HIV testing data, $p_1 = \text{pr}\{\text{tested}|\text{female}\}$ and $p_2 = \text{pr}\{\text{tested}|\text{male}\}$. $\hat{p}_1 = 0.15$ and $\hat{p}_2 = 0.14$ estimate these conditional probabilities.

χ^2 test for independence

- There is *no difference* between groups when $H_0 : p_1 = p_2$.
- That is, $H_0 :\ \Pr\{\text{success}|\text{group } 1\} = \Pr\{\text{success}|\text{group } 2\}$.
- If H_0 is true then the outcome (migraine reduction, being test for HIV, etc.) is *independent* of the group.
- This is tested using the chi-square statistic

  $$\chi^2_S = \sum_{i=1}^{4} \frac{(o_i - e_i)^2}{e_i},$$

  where $i = 1, 2, 3, 4$ are the four cells in the middle of the contingency table.
- The o_i are the observed counts and the e_i are what's expected if $p_1 = p_2$.

Computing e_i

- If $H_0 : p_1 = p_2$ is true then we can estimate the common probability $p = p_1 = p_2$ by $\hat{p} = (y_1 + y_2)/(n_1 + n_2)$. This is $\hat{p} = 56/75 = 0.747$ for migraine data.
- In the upper left corner we'd expect to see $\hat{p}n_1 = 0.747(49) = 36.59$ successes in the real surgery group, and so $49 - 36.59 = 12.41$ failures in the lower left.
- In the upper right corner we'd expect to see $\hat{p}n_2 = 0.747(26) = 19.41$ successes in the sham surgery group, and so $26 - 19.41 = 6.59$ failures in the lower right.

Observed and expected under H_0

Table 10.2.2 Observed and expected frequencies for migraine study

	Surgery		
	Real	Sham	Total
Success	41 (36.59)	15 (19.41)	56
No success	8 (12.41)	11 (6.59)	19
Total	49	26	75

$$\chi_S^2 = \frac{(41-36.59)^2}{36.59} + \frac{(15-19.41)^2}{19.41} + \frac{(8-12.41)^2}{12.41} + \frac{(11-6.59)^2}{6.59} = 6.06.$$

The P-value

- When $H_0 : p_1 = p_2$ is true, χ^2_S has a χ^2_1 distribution, chi-square with 1 degree of freedom.
- The P-value is the tail probability of a chi-square density with 1 *df* greater than what we saw χ^2_S. The P-value is the probability of seeing $\hat{p}_1$ and $\hat{p}_2$ *even further away from each other* than what we saw.
- We can get the P-value out of R using `chisq.test`, but now we need to put in a contingency table in the form of a matrix to get our P-value.

Obtaining surgery data P-value in R

- Need to create a 2×2 matrix of values first

```
> surgery=matrix(c(41,8,15,11),nrow=2)
> colnames(surgery)=c("Real","Sham")
> rownames(surgery)=c("Success","No success")
> surgery
           Real Sham
Success      41   15
No success    8   11
```

- The default `chisq.test(surgery)` uses $\chi^2_Y = \sum_{i=1}^{4} \frac{(|o_i - e_i| - 0.5)^2}{e_i}$. Called "Yates continuity correction" & gives more accurate P-values in small samples.

```
> chisq.test(surgery)

        Pearson's Chi-squared test with Yates' continuity correction

data:  surgery
X-squared = 4.7661, df = 1, p-value = 0.02902
```

Obtaining surgery data P-value in R

- To get the statistic and P-value in your book, we have to turn the Yates correction "off" using `chisq.test(surgery,correct=FALSE)`.

```
> chisq.test(surgery,correct=FALSE)

        Pearson's Chi-squared test

data:  surgery
X-squared = 6.0619, df = 1, p-value = 0.01381
```

- We reject $H_0 : p_1 = p_2$ at the 5% level. The surgery significantly reduces migraines.

10.3 Two ways to collect data

- There are two ways to collect 2×2 contingency table data.
- **Cross-sectional** data is collected by randomly sampling n individuals and cross-classifying them on two variables.
- Example Ask $n = 143$ random individuals two questions: salary high/low and education high-school/college.
- The row and column totals are random.
- **Product binomial** data is collected when a fixed number from one group is sampled, and a fixed number from another group is sampled.
- Example: Real vs. sham surgery for migraine.

10.4 Fishers exact test

- For the chi-square test to be valid, we cannot have very small sample sizes, say less than 5 in any cell.
- For small sample sizes there is an exact test, called Fisher's exact test for testing $H_0 : p_1 = p_2$.
- Fisher's test computes all possible 2×2 tables with the same number of successes and failures (56 successes and 19 failures for the migraine study) that make $\hat{p}_1$ and $\hat{p}_2$ even further apart than what we saw, and adds up the probability of seeing each table. Your book has details if you are interested on pp. 402-404.

Example 10.4.5 Flu shots

A random sample of college students found that 13 of them had gotten a flu shot at the beginning of the winter and 28 had not. Of the 13 who had a flu shot, 3 got the flu during the winter. Of the 28 who did not get a flu shot, 15 got the flu.

Table 10.4.3 Flu shot data

		No shot	Flu shot	Total
Flu?	Yes	15	3	18
	No	13	10	23
	Total	28	13	41

Want to test $H_0 : p_1 = p_2$ vs. $H_A : p_1 > p_2$ where p_1 is probability of getting flu among those without shots and p_2 is probability of getting flu among those that got shots.

P-value for flu shot data

Tables where p_1 and p_2 are even further apart in the direction of $H_A : p_1 > p_2$

Table		Probability
15	3	
13	10	0.05298
16	2	
12	11	0.01174
17	1	
11	12	0.00138
18	0	
10	13	0.00006

Figure 10.4.1

$$\text{P-value} = 0.05298 + 0.01174 + 0.00138 + 0.00006 = 0.06616.$$

Fisher's exact test

The probability of each table is given by the *hypergeometric distribution* and is beyond the scope of this course, although your book does a nice job of explaining if you are interested. For the flu shot data to carry out Fisher's test we type

```
> flu=matrix(c(15,13,3,10),nrow=2)
> fisher.test(flu,alternative="greater")

        Fisher's Exact Test for Count Data

data:  flu
p-value = 0.06617
alternative hypothesis: true odds ratio is greater than 1
sample estimates:
odds ratio
  3.721944
```

We'll discuss what an odds ratio is next time. For now, we accept $H_0 : p_1 = p_2$ at the 5% level. There is no statistically significant evidence that getting a flu shot decreases the probability of getting the flu.

Directional alternatives

- Using `fisher.test` we can test $H_0 : p_1 = p_2$ versus one of (a) $H_A : p_1 \neq p_2$, (b) $H_A : p_1 < p_2$, or (c) $H_A : p_1 > p_2$.
- Use `alternative="two.sided"` (the default) or `alternative="less"` or `alternative="greater"`.
- Fisher's test is better than the chi-square test.
- You will use `chisq.test` for tables larger than 2×2 instead, our next topic...

10.5 $r \times k$ contingency table

- The number of categories is generalized to r instead of 2.
- The number of groups is generalized to k instead of 2.
- Still want to test H_0 : the probabilities of being in each of the r categories do not change across the k groups.
- In the next example, $r = 3$ categories (agricultural field, prairie dog habitat, grassland) and $k = 3$ groups (2004, 2005, 2006).

Example 10.5.1 Plover Nesting

Wildlife ecologists monitored the breeding habitats of mountain plovers for three years and made note of where the plovers nested.

Table 10.5.1 Plover nest locations across three years

Location	Year 2004	Year 2005	Year 2006	Total
Agricultural field (AF)	21	19	26	**66**
Prairie dog habitat (PD)	17	38	12	**67**
Grassland (G)	5	6	9	**20**
Total	**43**	**63**	**47**	**153**

Question: do nesting choices vary over time?

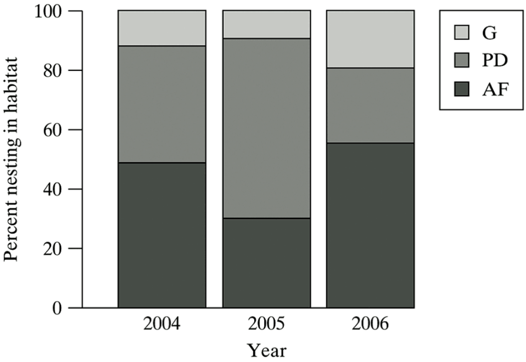
Plover nesting percentages over time

Table 10.5.2 Percentage distributions of plover nests by year

Location	Year 2004	Year 2005	Year 2006
Agricultural field (AF)	48.8	30.2	55.3
Prairie dog habitat (PD)	39.5	60.3	25.5
Grassland (G)	11.6	9.5	19.1
Total	**99.9***	**100.0**	**99.9***

*The sums of the 2004 and 2006 percentages differ from 100% due to rounding.

Stacked bar plot



Chi-square test

- H_0 category percentages do not change across groups.
- The chi-square test statistic is given by

$$\chi^2_S \sum_{\text{all cells}} \frac{(o_i - e_i)^2}{e_i}.$$

- Here, e_i is the total number in the group (column total) times the total row percentage, i.e.

$$e = \frac{\text{row total} \times \text{column total}}{\text{grand total}}.$$

- χ^2_S has a χ^2_{df} where $df = (r-1)(k-1)$. This is where the P-value comes from.

Plover data, observed & expected

Table 10.5.3 Observed and expected frequencies of plover nests

	Year			
Location	2004	2005	2006	Total
Agricultural field (AF)	21 (18.55)	19 (21.18)	26 (20.27)	**66**
Prairie dog habitat (PD)	17 (18.83)	38 (27.59)	12 (20.58)	**67**
Grassland (G)	5 (5.62)	6 (8.24)	9 (6.14)	**20**
Total	**43**	**63**	**47**	**153**

Upper left $18.55 = \frac{43(66)}{153}$,

$$\chi_S^2 = \frac{(21 - 18.55)^2}{18.55} + \cdots + \frac{(9 - 6.14)^2}{6.14} = 14.09.$$

Chi-square test in R

```
> plover=matrix(c(21,17,5,19,38,6,26,12,9),nrow=3)
> plover
     [,1] [,2] [,3]
[1,]   21   19   26
[2,]   17   38   12
[3,]    5    6    9

> chisq.test(plover)

        Pearson's Chi-squared test

data:  plover
X-squared = 14.0894, df = 4, p-value = 0.007015
```

We reject H_0 that nesting preference does not change over time at the 5% level. Hence we conclude that for the plovers the nesting preference significantly changes over time.