Sections 10.5 and 10.9

Note made by: Timothy Hanson
Instructor: Peijie Hou

Department of Statistics, University of South Carolina

Stat 205: Elementary Statistics for the Biological and Life Sciences

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			Total
	2004	2005	2006	
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?

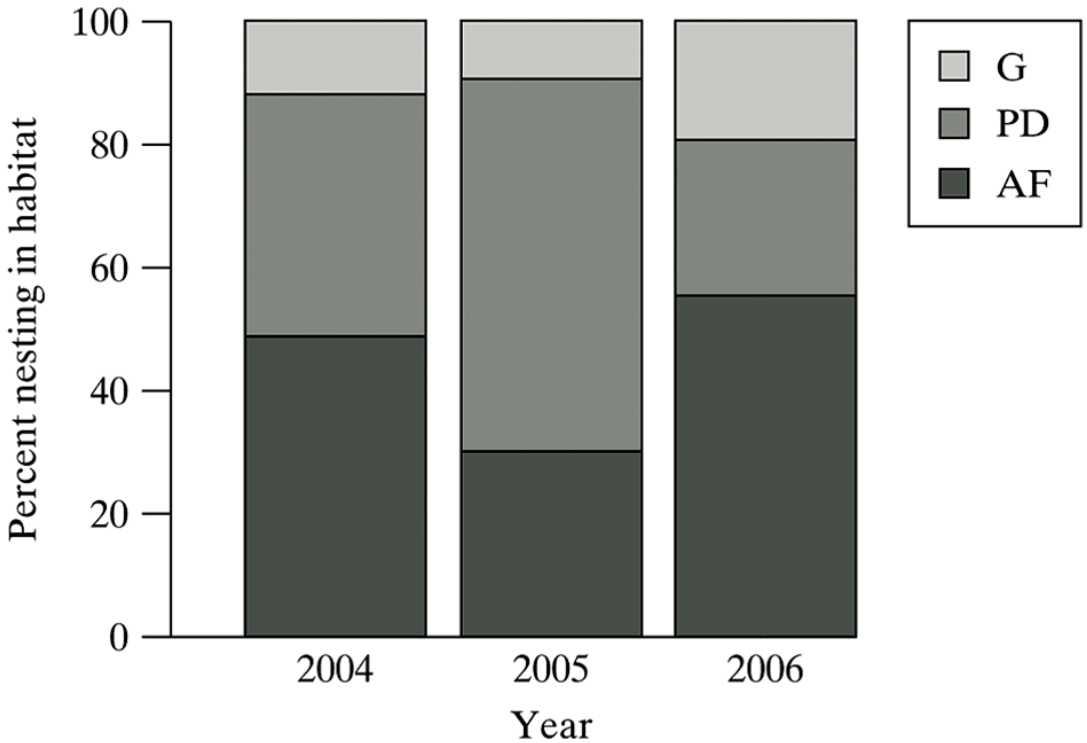
Plover nesting percentages over time

Table 10.5.2 Percentage distributions of plover nests by year

Location	Year		
	2004	2005	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



If the percentages of nesting choice are the same for each year, then the gray areas will be equal.

Chi-square test

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

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

- 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

Location	Year 2004	Year 2005	Year 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 $18.55 = \frac{43(66)}{153}$,

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

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.

10.9 Odds ratios and relative risk

- The relative risk is given by p_1/p_2. It is estimated by $\hat{p}_1/\hat{p}_2$.
- Tells us how the probability of having the event changes from group 1 to group 2.
- It's possible to get a confidence interval for p_1/p_2, but there is no automatic function to do this in R.
- The relative risk p_1/p_2 can magnify the effect of a treatment more than the difference in proportions $p_1 - p_2$.

Example 10.9.1 Smoking and Lung Cancer

The health histories of 11,900 middle-aged men were tracked over many years. During the study 126 of the men developed lung cancer, including 89 men who were smokers and 37 men who were former smokers.

		Smoking history	
		Smoker	Former smoker
Lung cancer?	Yes	89	37
	No	6,063	5,711
	Total	6,152	5,748

- $\hat{p}_1 = 89/6152 = 0.0145$ and $\hat{p}_2 = 37/5748 = 0.00644$.
- Relative risk is $\frac{\hat{p}_1}{\hat{p}_2} = \frac{0.0145}{0.00644} = 2.25$. The probability of lung cancer is 2.25 times greater in current smokers.
- Difference is $\hat{p}_1 - \hat{p}_2 = 0.0145 - 0.00644 = 0.0080$. The probability of lung cancer increases by 0.008 among current smokers.

Odds

- The *odds* of an event happening versus not happening are $p/(1-p)$. When someone says "3 to 1 odds the Gamecocks will win", they mean $p/(1-p) = 3$ which implies the probability the Gamecocks will win is 0.75, from solving $p/(1-p) = 3$ for p. Odds measure the relative rates of success and failure.
- Here, the probability of winning is 0.75, three times greater than the probability of losing, 0.25. So the odds are three, or "three to one."

Odds ratio

An *odds ratio* compares the odds of success (or disease or whatever) across the two groups:

$$\theta = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}.$$

Odds ratios are always positive and $\theta > 1$ indicates the relative rate of success group 1 is greater than for group 2. However, the odds ratio θ gives *no information* on the probabilities p_1 and p_2.

We often compare the odds across groups using an odds ratio. This tells us how the odds change going from group 1 to group 2. For example, we may be interested in how the odds of developing lung cancer changes from those that smoke to those that do not smoke.

Odds ratio, estimation

- The *odds ratio*

$$\theta = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

  is often used by epidemiologists instead of the relative risk.
- θ is estimated by

$$\hat{\theta} = \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_2/(1-\hat{p}_2)} = \frac{y_1(n_2-y_2)}{y_2(n_1-y_1)}.$$

Odds ratio versus relative risk

Different sets of probabilities p_1 & p_2 can lead to the same odds ratio.

- $p_1 = 0.833$ & $p_2 = 0.5$ yield $\theta = 5.0$, and relative risk of 1.7.
- $p_1 = 0.0005$ & $p_2 = 0.0001$ *also* give $\theta = 5.0$, but relative risk of 5.
- Odds ratios give different information than relative risks!
- **Important**: When dealing with a rare outcome, where $p_1 \approx 0$ and $p_2 \approx 0$, the relative risk is approximately equal to the odds ratio.
- R implements an exact method for obtaining a confidence interval for θ called Fisher's exact test, e.g. `fisher.test(smoking,conf.int=TRUE)`. Also implements test of $H_0 : \theta = 1$, a test of *independence* across groups, *just like the chi-square test!*

Example 10.9.1 Smoking and Lung Cancer

Prospective cohort study 11,900 middle-aged men.

		Smoking history	
		Smoker	Former smoker
Lung cancer?	Yes	89	37
	No	6,063	5,711
	Total	6,152	5,748

- $\hat{p}_1 = 89/6152 = 0.0145$ and $\hat{p}_2 = 37/5748 = 0.00644$.
- Relative risk is $\frac{\hat{p}_1}{\hat{p}_2} = \frac{0.0145}{0.00644} = 2.25$. The probability of lung cancer is 2.25 times greater in current smokers.
- Odds ratio is $\hat{\theta} = \frac{89(5711)}{37(6063)} = 2.27$, essentially the same as the relative risk here.
- The odds of lung cancer are 2.27 times greater for current smokers.

R code for $\hat{\theta}$ and 95% confidence interval

```
> smoking=matrix(c(89,6063,37,5711),nrow=2)
> rownames(smoking)=c("lung cancer","no lung cancer")
> colnames(smoking)=c("smoker","former smoker")
> smoking
               smoker former smoker
lung cancer        89            37
no lung cancer   6063          5711
> fisher.test(smoking,conf.int=TRUE)

        Fisher's Exact Test for Count Data

data:  smoking
p-value = 2.046e-05
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.525005 3.426733
sample estimates:
odds ratio
  2.265479
```

We are 95% confident that currently smoking increases the odds of lung cancer by 1.5 to 3.4 times, relative to formerly smoking.

Important: $\theta = 1 \Leftrightarrow p_1 = p_2$. So testing $H_0 : \theta = 1$ is the same thing as testing $H_0 : p_1 = p_2$. Note: p-value=2.046e-05, so we reject $H_0 : \theta = 1$ and in favor of $H_a : \theta \neq 1$.