L18: Sections 10.7 and 10.9

Department of Statistics, University of South Carolina

Stat 205: Elementary Statistics for the Biological and Life Sciences

10.7 confidence interval for $p_1 - p_2$

- Recall, in population 1, we observe y_1 out of n_1 successes; in population 2 we observe y_2 out of n_2 successes, placed in a contingency table

		Group	
		1	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.
- We want to compute a 95% confidence interval for $p_1 - p_2$.

Confidence interval for $p_1 - p_2$

- The estimate of $p_1 - p_2$ is $\hat{p}_1 - \hat{p}_2$.
- The standard error is

$$SE_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

- At 95% confidence interval for $p_1 - p_2$ is

$$\hat{p}_1 - \hat{p}_2 \pm 1.96 SE_{\hat{p}_1 - \hat{p}_2}.$$

- This is given in R by `prop.test(success,total)` where `success` is a list of the number of **successes** in the two groups and `total` is a list of the **total number sampled** in each group.

Example 10.7.1 Migraine headache data

- Migraine headache patients took part in a double-blind clinical trial to assess experimental surgery (numbers are slightly different than your book's).
- 75 patients were assigned real surgery ($n_1 = 49$) or sham surgery ($n_2 = 26$) so `total=c(49,26)`.
- There were $y_1 = 41$ successes among real surgery and $y_2 = 15$ successes among sham so `success=c(41,15)`.
- $\hat{p}_1 = 41/49 = 83.7\%$ & $\hat{p}_2 = 15/26 = 57.7\%$ so $\hat{p}_1 - \hat{p}_2 = 0.260$.
- The standard error of the difference is

$$SE_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{0.837(0.163)}{49} + \frac{0.577(0.423)}{26}} = 0.110.$$

- 95% confidence interval is $0.260 \pm 1.96(0.110) = (0.0444, 0.476)$.

R code for migraine headache data

Use `correct=FALSE` to get “old fashioned” confidence interval.

```
> total=c(49,26)
> success=c(41,15)
> prop.test(success,total,correct=FALSE)

        2-sample test for equality of proportions without continuity correction

data:  success out of total
X-squared = 6.0619, df = 1, p-value = 0.01381
alternative hypothesis: two.sided
95 percent confidence interval:
 0.04354173 0.47608150
sample estimates:
   prop 1    prop 2
0.8367347 0.5769231
```

We are 95% confident that real surgery reduces the probability of migraines by 4.3% to 47.6%.

R code for migraine headache data

Allowing the continuity correction changes the confidence interval a bit.

```
> prop.test(success,total)

        2-sample test for equality of proportions with continuity correction

data:  success out of total
X-squared = 4.7661, df = 1, p-value = 0.02902
alternative hypothesis: two.sided
95 percent confidence interval:
 0.01410688 0.50551635
sample estimates:
   prop 1    prop 2
0.8367347 0.5769231
```

We are 95% confident that real surgery reduces the probability of migraines by 1.4% to 50.6%. This interval is larger than the one on the previous slide.

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 so 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

- **Important**: $\theta = 1 \Leftrightarrow p_1 = p_2$. So testing $H_0 : \theta = 1$ is the same thing as testing $H_0 : p_1 = p_2$.
- 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 because interpretation can "switch" for case-control data – we'll talk about this shortly.
- 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.

Types of studies

Sir Richard Doll first demonstrated a link between smoking and lung cancer in 1947. He compared the smoking history of a group of hospitalized men with lung cancer with the smoking history of a similar group without lung cancer.

	Case	Control
Smoker	1350	1296
Non-smoker	7	61
Total	1357	1357

- In a **case-control** study, fixed numbers of cases n_1 and controls n_2 are (randomly) selected and exposure variables of interest recorded. In the above study we can compare the relative proportions of those who smoke within those that developed lung cancer (cases) and those that did not (controls). We can measure association between smoking and lung cancer, but cannot infer causation. These data were collected "after the fact." Data cheap and easy to get (p. 404).

Types of studies

- **Prospective studies** start with a sample and observe them through time.
  - **Clinical trial** randomly allocates "smoking" and "non-smoking" treatments to experimental units and then sees who ends up with lung cancer or not. Problem with ethics here.
  - A **cohort study** simply follows subjects after letting them assign their own treatments (i.e. smoking or non-smoking) and records outcomes. This type of study eventually "proved" causation between smoking and lung cancer; the case with Example 10.9.1.
- A **cross-sectional** design samples n subjects from a population and cross-classifies them, e.g. the HIV-testing data of Example 10.1.2.

Case-control and the odds ratio

Examples 10.9.4 & 10.9.5 on pp. 404–406.

- In a case-control study, the number of cases n_1 and controls n_2 are fixed ahead of time. Here, these are $n_1 = 500$ for "lung cancer" (cases) and $n_2 = 500$ for "no lung cancer" (controls). See Table 10.9.3.
- We can estimate the probabilities of smoking $\hat{p}_1 = y_1/n_2$ within the case and $\hat{p}_2 = y_2/n_2$ control groups.
- We *cannot* estimate the probability of having lung cancer within the smoking and non-smoking groups, because this is not how the data were collected. The numbers of non-smokers and smokers were set ahead of time.

Case-control and the odds ratio

- A designed experiment would assign "smoking" or "non-smoking" to subjects ahead of time and then classify their cancer status after a number of years. Not ethical or practical. We *can* implement a cohort study, or take cross-sectional data, but this is way more expensive.
- The odds ratio θ does not care if case/control numbers are fixed, or smoking/non-smoking numbers are fixed ahead of time. It can be shown mathematically – using Bayes' rule – that *the odds ratio is the same either way*. Relative risks *do not have this property*.

An important property of odds ratio

- In a case-control study we can estimate the relative risk of being a smoker across the lung-cancer and no lung-cancer groups p_1/p_2.
- *What we'd really like* is the relative risk of lung cancer across smokers and non-smokers. Although we cannot estimate this, we can estimate the odds ratio, and the odds ratio is *estimated the same way regardless of the type of study*.
- For the purposes of estimating an odds ratio, it *does not matter* if data are sampled prospectively, retrospectively, or cross-sectionally. The common odds ratio is estimated

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

Odds ratio for smoking case-control data

The estimated odds ratio for Sir Richard Doll's data is

$$\hat{\theta} = \frac{1350(61)}{7(1296)} = 9.1.$$

1. *The odds of smoking is 9 times greater among those with lung cancer.*
2. *The odds of having lung cancer is 9 times greater among smokers.*

The second interpretation is more relevant when deciding whether or not you should take up recreational smoking.

Note that we *cannot* estimate the relative risk of developing lung cancer for smokers. Which relative risk *can* we estimate?

Doll's data in R

95% confidence interval for θ via Fisher's method.

```
> lung=matrix(c(1350,7,1296,61),nrow=2)
> colnames(lung)=c("Case","Control")
> rownames(lung)=c("Smokers","Non-smokers")
> lung
            Case Control
Smokers     1350    1296
Non-smokers    7      61
> fisher.test(lung,conf.int=TRUE)

        Fisher's Exact Test for Count Data

data:  lung
p-value = 4.292e-12
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  4.12616 23.59182
sample estimates:
odds ratio
  9.071755
```

The odds of lung-cancer are 9.1 times greater among smokers. We are 95% confident that the odds of lung cancers is between 4.1 and 23.6 times greater among smokers. Since the P-value$= 0.0000000000043 < 0.05$ we reject $H_0 : \theta = 1$ at the 5% level.

Aspirin and heart attacks

$n = 1360$ stroke patients randomly assigned to aspirin or placebo & followed about 3 years – prospective study.

	Placebo	Aspirin
Heart attack	28	18
No heart attack	656	658
Total	684 (fixed)	676 (fixed)

We want to know if there's an association between taking aspirin and having a heart attack, i.e. $H_0 : \theta = 1$. If we reject, a 95% confidence interval for θ will tell us how beneficial aspirin is in terms of the odds of having a heart attack.

R code for aspirin data

```
> aspirin=matrix(c(28,656,18,658),nrow=2)
> colnames(aspirin)=c("Placebo","Aspirin")
> rownames(aspirin)=c("Heart attack","No heart attack")
> aspirin
                Placebo Aspirin
Heart attack         28      18
No heart attack     656     658
> fisher.test(aspirin,,conf.int=TRUE)

        Fisher's Exact Test for Count Data

data:  aspirin
p-value = 0.1768
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.8236804 3.0256940
sample estimates:
odds ratio
  1.559785
```

What do we conclude?