Discrete Response Model
Lecture 1

datascience@berkeley

Formulation of Contingency Table and Confidence Interval of Two Binary Variables

Notations and Model

- Let Y_{11}, …, $Y_{n_1 1}$ be Bernoulli random variables for group 1 (row 1 of the contingency table).
- Let Y_{12}, …, $Y_{n_1 2}$ be Bernoulli random variables for group 2 (row 2 of the contingency table).
- The number of "successes" for a group is represented by $W_j = \sum_{i=1}^{n_j} Y_{ij}$.
- W_j has a binomial distribution with success probability π_j and number of trials of n_j
- W_1 is independent of W_2; thus, we have an "independent binomial model". Some people refer to this as "independent binomial sampling" as a way to describe how the contingency table counts come about.
- The MLE of π_j is $\hat{\pi}_j = w_j / n_j$
- A "+" in a subscript is used to denote indices in a subscript that are being summed over. For example, $W_+ = W_1 + W_2$ is the total number of successes and $n_+ = n_1 + n_2$ is the total sample size. In fact, $W_j = Y_{+j}$.

		Response 1	Response 2	
Group	1	w_1	$n_1 - w_1$	n_1
	2	w_2	$n_2 - w_2$	n_2
		w_+	$n_+ - w_+$	n_+

		Response 1	Response 2	
Group	1	π_1	$1 - \pi_1$	1
	2	π_2	$1 - \pi_2$	1

Example: Larry Bird's Free Throws

```
c.table<-array(data = c(251, 48, 34, 5), dim = c(2,2),
    dimnames = list(First = c("made", "missed"), Second =
    c("made", "missed")))
```

```
> c.table
        Second
First    made missed
  made    251     34
  missed   48      5
```

```
rowSums(c.table) #n1 and n2

pi.hat.table<-c.table/rowSums(c.table)
```

```
> pi.hat.table
        Second
First          made     missed
  made    0.8807018 0.11929825
  missed  0.9056604 0.09433962
```

The estimated probability that Larry Bird makes his second free throw attempt is $\hat{\pi}_1 = 0.8807$, given that he makes the first, and $\hat{\pi}_2 = 0.9057$, given he misses the first.

Confidence Intervals for the Difference of Two Probabilities

Remember from Section 1.1 that the estimated probability of success $\hat{\pi}$ can be treated as an approximate normal random variable with mean π and variance $\pi(1-\pi)/n$ for a large sample. Using the notation in this week, this means that

$$\hat{\pi}_1 \sim N(\pi_1\ \pi_1(1-\pi_1)/n_1) \text{ and}$$

$$\hat{\pi}_2 \sim N(\pi_2\ \pi_2(1-\pi_2)/n_2)$$

for large n_1 and n_2.

Note: $Var(\hat{\pi}_1 - \hat{\pi}_2) = Var(\hat{\pi}_1) + Var(\hat{\pi}_2)$ because $\hat{\pi}_1$ and $\hat{\pi}_2$ are independent random variables. Some of you may have seen the following: Let X and Y be independent random variables and let a and b be constants. Then $Var(aX+bY) = a^2Var(X) + b^2Var(Y)$.

The estimate of the variance is then

$$Var(\hat{\pi}_1 - \hat{\pi}_2) = \frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}$$

A $(1-\alpha)100\%$ Wald confidence interval for $\pi{\downarrow}1 - \pi{\downarrow}2$ is

$$\hat{\pi}_1 - \hat{\pi}_2 \pm z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$$

Agresti and Caffo Adjustment to CI

Let $\pi_1 = \dfrac{w_1 + 1}{n_1 + 2}$ and $\pi_2 = \dfrac{w_2 + 1}{n_2 + 2}$

The Agresti-Caffo confidence interval is

$$\pi_1 - \pi_2 \pm Z_{1-\alpha/2} \sqrt{\frac{\pi_1(1-\pi_1)}{n_1 + 2} + \frac{\pi_2(1-\pi_2)}{n_2 + 2}}$$

Example: Larry Bird's Free Throws

```r
```{r}
alpha<-0.05
pi.hat1<-pi.hat.table[1,1]
pi.hat2<-pi.hat.table[2,1]

#Wald
var.wald<-pi.hat1*(1-pi.hat1) / sum(c.table[1,]) + pi.hat2*(1-pi.hat2) /
sum(c.table[2,])

pi.hat1 - pi.hat2 + qnorm(p = c(alpha/2, 1-alpha/2)) * sqrt(var.wald)
```

```
-0.11218742  0.06227017
```

```r
#Agresti-Caffo
pi.tilde1<-(c.table[1,1] + 1) / (sum(c.table[1,]) + 2)
pi.tilde2<-(c.table[2,1] + 1) / (sum(c.table[2,]) + 2)
var.AC<-pi.tilde1*(1-pi.tilde1) / (sum(c.table[1,]) + 2) +
pi.tilde2*(1-pi.tilde2) / (sum(c.table[2,]) + 2)
pi.tilde1 - pi.tilde2 + qnorm(p = c(alpha/2, 1-alpha/2)) * sqrt(var.AC)
```

```
-0.10353254  0.07781192
```

Therefore, the 95% Wald confidence interval is

$$-0.1122 < \pi_1 - \pi_2 < 0.0623$$

and the 95% Agresti-Caffo confidence interval is

$$-0.1035 < \pi_1 - \pi_2 < 0.0778$$
```

Testing the Difference of Two Probabilities

Hypothesis test of $H_0: \pi_1 - \pi_2 = 0$ vs. $H_a: \pi_1 - \pi_2 \neq 0$

Test Statistic:

$$Z_0 = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\bar{\pi}(1-\bar{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where $\bar{\pi} = w_+ / n_+$

This test statistic has a standard normal distribution for a large sample. Therefore, you can reject H_0 if $|Z_0| > Z_{1-\alpha/2}$

Berkeley SCHOOL OF INFORMATION