960:553– Categorical Data Analysis– Fall, 2023

Homework 1 Solutions, 21 Sep

1. An investigator identified 935 sets of twins. The investigator randomly selected one member of each set to review a standard set of instructions, and the other to review a simplified set of instructions. All subjects were tested on the content of the instructions. In 289 sets of twins, the twin using the simplified set of instructions performed better than the twin using the standard instructions.

a. Calculate a 95% confidence interval for the proportion of pairs in which the individual with the simplified set of instructions performs better than the individual with the standard set of instructions. Use the normal approximation, Wilson, and exact methods, and compare.

Denote the sets of twins performing better on the simplified instructions as group 0, and the sets of twins performing better on the standard instructions as group 1. The we want the proportion of sets performing better on the simplified instructions, π_0 *, and An approximate normal confidence interval is*

$$289/935 \pm 1.96 \times \sqrt{(289/935) \times (646/935)/935} = (0.279, 0.339).$$

An exact confidence interval starts with noting that

$$F_{0.025}(1294, 578) = 1.151 \text{ and } F_{0.025}(580, 1292) = 1.147.$$

The confidence interval is given by

$$(\frac{289}{289 + (646+1)1.151018}, 1 - \frac{646}{646 + (289+1)1.146593}) = (0.28, 0.34).$$

The intervals are very close.

These three intervals could have been constructed in R via

```
binom.test(289,935)
library(Hmisc)
binconf(289,953,method="wilson")
binconf(289,953,method="asymptotic")
```

giving $(0.280, 0.340)$ *,* $(0.275, 0.333)$ *, and* $(0.274, 0.332)$ *for the exact, Wilson, and Wald intervals respectively.*

b. Repeat part (a), assuming that 9 sets of twins were recruited, and in 3 sets the twin with the simplified instructions did better.

An approximate normal confidence interval is $3/9 \pm 1.96 \times \sqrt{(3/9) \times (6/9)/9} = (0.025, 0.641)$ *. An exact confidence interval starts with noting that*

$$F_{0.025}(14, 6) = 5.296811 \text{ and } F_{0.025}(8, 12) = 3.511777.$$

The confidence interval is given by $(\frac{3}{3+(6+1)5.296811}, 1 - \frac{6}{6+(3+1)3.511777}) = (0.075, 0.701)$ *. This agreement is not so bad, given the very small sample size.*

The three intervals could have been constructed in R via

```
binom.test(3,9)
library(Hmisc)
binconf(3,9,method="wilson")
binconf(3,9,method="asymptotic")
```

giving $(0.075, 0.701)$, $(0.121, 0.646)$ *, and* $(0.025, 0.641)$ *, for the exact, Wilson, and Wald intervals respectively.*

2. Kane (2001) presents data on the amount of video game usage allowed to 100 children, and whether these children have disicpline problems in school. These data are summarized below.

	0 Hours	1–3 Hours	4–6 Hours	7–10 Hours	$>$ 10 Hours
Discipline Problems	0	2	2	0	2
No Discipline Problems	11	43	19	1	15

a. Test the hypothesis that exposure (that is, amount of time children are permitted to use video games) is associated with the disease status (that is, whether the children have discipline problems.) Comment on on any reasons your calculations might be suspect. Do not use the ordering of the categories. Interpret your results.

Here are the relevant results.

```
        Statistics for Table of disc by hours
   Statistic              DF     Value    Prob
   ------------------------------------------------------

   Chi-Square              4    2.2936  0.6819
    WARNING: 60% of the cells have expected counts less
          than 5. Chi-Square may not be a valid test.
```

The chi-square test shows no evidence of differences in probability of diciplinary problems based on numbers of hours of video games watched. The χ^2 *approximation to the distribution of the statistic is suspect, since most of the cells have a small expected value.*

Do this in R via

```
video<-as.data.frame(matrix(c(1,0,0,1,1,2,1,2,2,1,3,0,1,4,2,
0,0,11,0,1,43,0,2,19,0,3,1,0,4,15),ncol=3,byrow=2))
names(video)<-c("disc","hours","we")
vidmatrix<-xtabs(we˜disc+hours,data=video)
chisq.test(vidmatrix)
```

b. Test the hypothesis that exposure is associated with the disease status, presuming that the categories are ordered. Comment on on any reasons your calculations might be suspect. Interpret your results.

The test statistic is 1.856, with a p-value of 0.1731. The p-value is large enough that there is no evidence that video game playing has an effect on discipline problems. The problems with the test statistic being approximately normal are less severe here than they are in part (a), because we add over all five levels before squaring.

This could have been done in R via

```
library(DescTools)
CochranArmitageTest(vidmatrix)
```

or

```
prop.trend.test(vidmatrix[2,],vidmatrix[1,]+vidmatrix[2,])
```

c. Collapse the first two exposure categories and the last three exposure categories to obtain the table

	Low Game Usage	High Game Usage
Discipline Problems	2	4
No Discipline Problems	54	35

.

Estimate the odds ratio for measuring the association between exposure and disease, defined so the odds ratio is high if high game usage is associated with discipline problems. Calculate a 95% confidence interval for the odds ratio, and interpret your results.

Use the R commands

```
video$tt<-(video$hours>1)*1
fisher.test(xtabs(we~disc+tt,data=video))
```

The confidence interval is $(0.4119, 35.3778)$ *. The hypothesis of no association can not be ruled out. There are too few discipline problems to learn much about the association here.*

d. Perform Fisher's exact test on the table in part (c), and interpret your results.

The above commands calculate the p-value as 0.224.