Homework 5 Solutions

Public Health 241: Statistical Analysis of Categorical Data

Spring 2019

1. The table below summarizes data from a traditional case-control study of oral cancer in females and employment in the textile industry.

Table 1: Oral Cancer

Years	Cases	Controls
Ten or more years in textile industry	16	8
Other work history	24	32

(a) Calculate a point estimate for the odds ratio of oral cancer associated with having been employed in the textile industry for 10 or more years.

$$\widehat{OR} = \frac{16 \times 32}{8 \times 24} = 2.667$$

(b) Calculate an estimate of the variance of the log of your point estimate from (a).

$$\widehat{var}(\widehat{OR}) = \frac{1}{16} + \frac{1}{8} + \frac{1}{24} + \frac{1}{32} = 0.2604$$

(c) Construct an approximate 90% confidence interval for the log odds ratio of oral cancer associated with having been employed in the textile industry for 10 or more years. Hint: Use the base R function `qnorm()` to compute the correct z-value.

```
qnorm(.95)
```

```
## [1] 1.644854
```

$$\log(2.667) \pm 1.645\sqrt{0.2604} = (0.141, 1.820)$$

(d) Construct an approximate 90% confidence interval for the odds ratio of oral cancer associated with having been employed in the textile industry for 10 or more years.

$$(e^{0.141}, e^{1.820}) = (1.15, 6.17)$$

(e) Does the confidence interval provide evidence against the null hypothesis that employment history in the textile industry and the risk for oral cancer are independent of each other in the target population?

Since the confidence interval does not include the null value 1, the data provide evidence (at a 10% significance level) for an association between employment history and the risk for oral cancer.

(f) Carry out the χ^2 test of independence. At a significance level of 10%, does the test accept or reject the null hypothesis of independence between employment history and the risk for oral cancer? Do these results agree with those based on the confidence interval calculation?

$$\chi^2 = \frac{80(16 \times 32 - 8 \times 24)^2}{40 \times 40 \times 24 \times 56} = 3.81 \quad (p = 0.051)$$

The χ^2 test rejects the null hypothesis of independence at a 10% significance level (but would accept the null hypothesis at the 5% significance level). As is to be expected, this agrees with the confidence interval we calculated.

(g) Compute a point estimate for the odds ratio using the small-sample adjustment for obtaining a direct odds ratio estimate (rather than a point estimate on the log scale). Compare your point estimate to the one obtained in (a) and comment.

$$OR_{SS} = \frac{16 \times 32}{9 \times 25} = 2.28$$

The regular point estimate differs from the small-sample adjusted point estimate by about 17%. This is a considerable discrepancy so that the approximations used in the original confidence interval calculation may be somewhat suspect and exact confidence interval may be preferable.

(h) Construct a 90% confidence interval for the odds ratio using the small-sample adjustment presented in class. Compare the results to those obtained above and comment.

$$(\log \widehat{OR})_{SS} = \log \frac{16.5 \times 32.5}{8.5 \times 24.5} = 0.946$$

$$\widehat{var}(\log \widehat{OR})_{SS} = \frac{1}{16.5} + \frac{1}{8.5} + \frac{1}{24.5} + \frac{1}{32.5} = 0.250$$

$$90\% CI \text{for} \log(OR) = 0.946 \pm 1.645 \times \sqrt{0.250} = (0.124, 1.768)$$

$$90\% CI \text{for} OR = (e^{0.124}, e^{1.768}) = (1.13, 5.86)$$

The upper limit of the approximate confidence interval differs from the upper limit of the small- sample confidence interval by about 5%. This is not a huge difference, but again we may be worried that the approximations used in the original confidence interval calculation may be somewhat suspect so that exact confidence interval may be preferable.

(i) Obtain an exact 90% confidence interval for the odds ratio and comment. *Hint: Use the* `epitab()` *command in base R to help you with the calculation. You will need to use the "fisher" argument to calculate an exact confidence since* `epitab()` *uses a normal (Wald) approximation by default.*

```
epitab(c(16, 24, 8, 32), conf.level = 0.90, oddsratio = "fisher")
```

```
## $tab
##           Outcome
## Predictor  Disease1        p0 Disease2        p1 oddsratio   lower
##   Exposed1       16 0.6666667       24 0.4285714  1.000000      NA
##   Exposed2        8 0.3333333       32 0.5714286  2.633562 1.03375
##           Outcome
## Predictor     upper    p.value
##   Exposed1       NA         NA
##   Exposed2 7.039224 0.08658224
##
## $measure
## [1] "fisher"
##
## $conf.level
## [1] 0.9
##
## $pvalue
## [1] "fisher.exact"
```

The exact confidence interval (1.03, 7.04) is substantially wider than the approximate confidence intervals calculated above, even if small-sample adjustments are used. The normal approximation that forms the basis for those approximate confidence intervals may not be appropriate.

2. Tuyns et al. (1977) carried out a case-control study of esophageal cancer in Ille-et-Vilaine in Brittany, France. The data set, *oesoph* is available on the bCourses website. One risk factor of interest was daily alcohol consumption, measured in grams per day, given in the data set in four levels: 0 to 39 (`alcgp=0`), 40 to 79 (`alcgp=1`), 80 to 120 (`alcgp=2`), and > 120 g/day (`alcgp=3`).

(a) Download the `oesoph` data set from bCourses, and add it to your `data` folder in the hw05 directory. Create a new categorical column in your data that equals 1 if an individual's alcohol consumption is at least 80 g/day and 0 otherwise. Tabulate the binary alcohol variable against the original one to make sure you have what you want. Then, generate a 2 x 2 table that breaks down the sample by case status and this binary risk factor. You may find the following R commands useful for this: `ifelse()`, `mutate()`, `table()`, `aggregate()`, and `epitable()`.

```
oesoph <- read.dta("data/oesoph.dta")
oesoph <- mutate(oesoph, alc = as.numeric(oesoph$alcgp >= 2))    # create new variable
table(oesoph$alc, oesoph$alcgp, dnn = c("alc", "alcgp"))         # tabulate against original
```

```
##    alcgp
## alc  0  1  2  3
##   0 35 38  0  0
##   1  0  0 31 31
```

```
oesoph.agg <- aggregate(oesoph, by = list(oesoph$casestatus, oesoph$alc), FUN = sum)
alc.by.case <- epitable(oesoph.agg$freq)
```

(b) Calculate a 95% confidence interval that is based on the normal approximation using the R command `epitab()`.

```
epitab(alc.by.case)
```

```
## $tab
##           Outcome
## Predictor  Disease1        p0 Disease2   p1 oddsratio    lower    upper
##   Exposed1      666 0.8593548      104 0.52  1.000000       NA       NA
##   Exposed2      109 0.1406452       96 0.48  5.640085 4.000589 7.951467
##           Outcome
## Predictor       p.value
##   Exposed1           NA
##   Exposed2 1.079486e-22
##
## $measure
## [1] "wald"
##
## $conf.level
## [1] 0.95
##
## $pvalue
## [1] "fisher.exact"
```

(c) Compare the confidence interval in (b) to an exact 95% confidence interval obtained through `epitab()` with the `"fisher"` option for the `oddsratio` parameter. Is the sample size large enough to warrant the normal approximation used to construct the confidence interval in (b) or do we need to use an exact confidence interval?

```
epitab(alc.by.case, oddsratio = "fisher")
```

```
## $tab
##           Outcome
## Predictor  Disease1        p0 Disease2   p1 oddsratio    lower    upper
```

```
##   Exposed1      666 0.8593548      104 0.52  1.000000       NA       NA
##   Exposed2      109 0.1406452       96 0.48  5.626771 3.936824 8.061478
##           Outcome
## Predictor       p.value
##   Exposed1           NA
##   Exposed2 1.079486e-22
##
## $measure
## [1] "fisher"
##
## $conf.level
## [1] 0.95
##
## $pvalue
## [1] "fisher.exact"
```

The end points of the approximate confidence interval are less than 2% off from those of the exact confidence interval. The sample sizes of cases and controls appear large enough to justify the normal approximation.

(d) Also examine the relationship between the risk for esophageal cancer and the dichotomized measure of alcohol consumption using the χ^2 test in R. Compare your conclusions to those based on the confidence intervals computed in (b) and (c). The following R commands may be useful for this: `chisq.test()`, `pchisq()`. *Remember to use turn off Yate's correction for the χ^2 test and set `lower.tail = FALSE` for `pchisq()`.*

```
chisq.test(alc.by.case, correct = FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  alc.by.case
## X-squared = 110.26, df = 1, p-value < 2.2e-16
```

```
pchisq(110.26, df = 1, lower.tail = FALSE)
```

```
## [1] 8.594547e-26
```

Like the confidence interval, the χ^2 test provides strong evidence against the null hypothesis that alcohol consumption is independent of the risk of esophageal cancer.