Homework 5

Public Health 241: Statistical Analysis of Categorical Data

YOUR NAME / YOUR STUDENT ID HERE

TODAY'S DATE

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.

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

(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.

(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) 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?

(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?

(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.

(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.

(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.*

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()`.

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

(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?

(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()`.*