# Discrete Multivariate Analysis - 579 - HW #6

Alex Towell (atowell@siue.edu)

## Problem 1

Suppose $n_1 \sim \text{BIN}(\text{n}, \pi)$. Then, $\hat{\pi} = n_1/n$ has an asymptotic normal distribution $\mathcal{N}(\pi, \pi(1-\pi)/n)$. Use the delta method to determine $\sigma^2 \log(\hat{\pi}/(1-\hat{\pi}))$, the asymptotic variance of the sample log odds.

If we have a non-linear function of some random variable, in the delta method we use a Taylor approximation of the function centered around the random variables expected value such that the approximate variance of the function of the random variable is easily computed.

Let $\text{g}(\hat{\pi}) = \log(\hat{\pi}/(1-\hat{\pi}))$. A linear approximation of g is given by

$$\hat{\text{g}}(\hat{\pi}) = \text{g}(\pi) + \text{g'}(\pi)(\hat{\pi} - \pi),$$

the derivative of g is given by

$$\text{g'}(\pi) = \frac{1}{\pi(1-\pi)},$$

and the variance of $\hat{\text{g}}(\hat{\pi})$ is given by

$$\begin{aligned}
\text{Var}(\hat{\text{g}}(\hat{\pi})) &= (\text{g'}(\pi))^2 \, \text{Var}(\hat{\pi}) \\
&= \frac{1}{\pi^2(1-\pi)^2} \frac{\pi(1-\pi)}{n} \\
&= \frac{1}{\pi(1-\pi)} \frac{1}{n}.
\end{aligned}$$

Since we do not know $\pi$, we approximate it with $\hat{\pi}$, thus

$$\sigma^2(\log(\hat{\pi}/(1-\hat{\pi}))) = \frac{1}{\hat{\pi}(1-\hat{\pi})} \frac{1}{n}.$$

# Problem 2

Consider data from a retrospective study on the relationship between daily alcohol consumption and the onset of esophagus cancer.

| | | cancer | no cancer |
|---|---|---|---|
| daily alcohol consumption | $> 80$g | 71 | 82 |
| | $< 80$g | 60 | 441 |
| | total | 131 | 523 |

## Part (a)

Provide an equation for $\hat{\sigma}(\log \hat{\theta})$.

In a retrospective study, the table draws samples from the conditional probabilities $\mathrm{P}(X = i|Y = j)$ and we denote $\mathrm{P}(X = 1|Y = j)$ by $p_j$.

The ML estimator of $p_j$ from the data in a retrospective study is given by

$$\hat{p}_j = \frac{n_{1j}}{n_{+j}}.$$

In a retrospective study, an estimator of $\sigma(\log\hat{\theta})$ is given by

$$\hat{\sigma}(\log\hat{\theta}) = \left[\left(\frac{1}{\hat{p}_1} + \frac{1}{1-\hat{p}_1}\right)\frac{1}{n_{+1}} + \left(\frac{1}{\hat{p}_2} + \frac{1}{1-\hat{p}_2}\right)\frac{1}{n_{+2}} +\right]^{1/2}.$$

## Part (b)

Does your answer in (a) depend on the sampling scheme? Explain.

It does not depend on the sampling scheme. When asymptotic normality holds, replacing parameters with statistics invokes the likelihood principle.

When we substitute the MLE $\hat{p}_j$ into the equation for $\hat{\sigma}(\log\hat{\theta})$, we get the result

$$\hat{\sigma}(\log\hat{\theta}) = \left(\frac{1}{n_{11}} + \frac{1}{n_{21}} + \frac{1}{n_{12}} + \frac{1}{n_{22}}\right)^{1/2},$$

which is the same as for retrospective studies and cross-sectional studies.

## Part (c)

Compute a 95% confidence interval for $\log\theta$.

Recall $\theta = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$. If we replace $p_j$ by its ML estimator $\hat{p}_j$, then

$$\hat{p}_1 = \frac{n_{11}}{n_{+1}} = \frac{71}{131} \approx 0.542$$

and

$$\hat{p}_2 = \frac{n_{12}}{n_{+2}} = \frac{82}{523} \approx 0.157.$$

Thus, by the invariance property of MLEs,

$$\hat{\theta} = \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_2/(1-\hat{p}_2)} \approx \frac{0.542/0.458}{0.157/0.843} \approx 6.354$$

and therefore $\log\theta \approx \log 6.354 \approx 1.850$. Next, we need to find the standard deviation of the $\log\hat{\theta}$,

$$\hat{\sigma}(\log\hat{\theta}) = \left[\left(\frac{1}{0.542} + \frac{1}{0.458}\right)\frac{1}{131} + \left(\frac{1}{0.157} + \frac{1}{0.843}\right)\frac{1}{523}\right]^{1/2} \approx 0.213.$$

Thus, a 95% confidence interval for $\log\theta$ is

$$\log\hat{\theta} \pm 1.96\hat{\sigma}(\log\hat{\theta}).$$

Plugging in these computed values, we get the 95% confidence interval

$$[1.850 - 0.417, 1.850 + 0.417] = [1.433, 2.267].$$

### Part (d)

Compute $\hat{\gamma}$ for the $2 \times 2$ table. Provide an interpretation of the effect size, stated in the context of the problem.

Observe that

$$\hat{\gamma} = \frac{\hat{\theta} - 1}{\hat{\theta} + 1}.$$

The MLE $\hat{\theta} \approx 6.354$. Thus, $\hat{\gamma} \approx \frac{6.354-1}{6.354+1} \approx 0.728$.

We estimate that there is a *large* size, positive association between daily alcohol consumption and the onset of cancer.

## Problem 3

For the *counts* array, we populated it with the values $(7, 7, 2, 3, 2, 8, 3, 7, 1, 5, 4, 9, 2, 8, 9, 14)$ and ran the code.

```
##       2.5%        25%        50%        75%      97.5%
## 0.09630732 0.24414958 0.31667622 0.38610092 0.50641592
```

Thus, a 95% interval estimate is approximately $[0.1, 0.5]$.