BIOS 625 Fall 2015 Homework 2 Solutions

1. Agresti 1.24

Since the Wald confidence interval for a binomial parameter π is degenerate when $\hat{\pi} = 0$ or 1, argue that the probability that the interval covers π cannot exceed $\left[1 - \pi^n - (1-\pi)^n\right]$; hence, the infimum of the coverage probability over $0 < \pi < 1$ equals 0, regardless of n.

$P(\text{CI contains } \pi) \le P(1 \le Y \le n-1) = 1 - P(Y = n) - P(Y = 0) = 1 - \pi^n - (1-\pi)$. This converges to 0 as $\pi \to 0$ or as $\pi \to$ n.

2. Agresti 1.32.c

Refer to the quadratic form 1.18, $n(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0)^T \boldsymbol{\Sigma}_0^{-1} (\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0)$, that leads to the Pearson chi-sqared. For the z_s statistic (1.11), show that $z_s^2 = X^2$ for c = 2.

Let $\hat{\pi} = n_1/n$, and $(1-\hat{\pi}) = n_2/n$, and denote the null probabilities in the two categories by π_0 and $(1-\pi_0)$. Then,

$$X^2 = \frac{(n_1 - n\pi_0)^2}{n\pi_0} + \frac{(n_2 - n(1-\pi_0))^2}{n(1-\pi_0)}$$

$$= \frac{n\left[(\hat{\pi} - \pi_0)^2 (1-\pi_0) + ((1-\hat{\pi}) - (1-\pi_0)^2 \pi_0)\right]}{\pi_0(1-\pi_0)}$$

$$= \frac{(\hat{\pi} - \pi_0)^2}{\pi_0(1-\pi_0)/n}$$

$$X^2 = z_s^2$$

3. Agresti 1.33

For testing $H_0 : \pi_j = \pi_{j0}, j = 1, \ldots, c,$ using sample multinomial proportions $\{\hat{\pi}_j\}$, the likelihood ratio statistic (1.17) is

$$G^2 = -2n \sum_j \hat{\pi}_j \log(\pi_{j0}/\hat{\pi}_j)$$

Show that $G^2 \ge 0,$ with equality if and only if $\pi_j = \pi_{j0}$ for all j.

Let X be a random variable that equals $\pi_{j0}/\hat{\pi}_j$ with probability $\hat{\pi}_j$. By Jensen's inequality, since the negative log function is convex, $E(-\log X) \geq -\log(EX)$. Hence,

$$E(-\log X) = \sum \hat{\pi}_j \log(\hat{\pi}_j/p_{j0}) \geq -\log\left[\sum \hat{\pi}_j (\pi_{j0}/\hat{\pi}_j)\right] = -\log \sum \pi_{j0} = -\log(1) = 0$$

Thus $G^2 = 2nE(-\log X) \geq 0$.

4. Agresti 1.34.a and b

For counts $\{n_j\}$, the power divergence statistic for testing goodness of fit is

$$\frac{2}{\lambda(\lambda+1)}\sum n_i\left[(n_i/\hat{\mu}_i)^\lambda - 1\right] \quad \text{for } -\infty < \lambda < \infty.$$

a. *For $\lambda = 1$, show that this equal X^2.*

$$\frac{2}{1(1+1)}\sum n_i\left[(n_i/\mu_i)^1 - 1\right] = \sum n_i\left[(n_i - \mu_i)/\mu_i\right]$$
$$= \sum(n_i - \mu_i + \mu_i)\left[(n_i - \mu_i)/\mu_i\right]$$
$$= X^2 + \sum(n_i - \mu_i)$$
$$= X^2$$

b. *As $\lambda \to 0$, show that it converges to G^2.*

Use the fact that $\log(t) = \lim_{h\to 0}(t^h - 1)/h$. Let $t = n_i/\hat{\mu}_i$ and $h = \lambda$

$$\lim_{\lambda\to 0}\frac{2}{\lambda(\lambda+1)}\sum n_i\left[(n_i/\hat{\mu}_i)^\lambda - 1\right] = \lim_{\lambda\to 0}\frac{2}{(\lambda+1)}\sum n_i\left[\frac{(n_i/\hat{\mu}_i)^\lambda - 1}{\lambda}\right]$$
$$= 2\sum n_i \log\left(\frac{n_i}{\hat{\mu}_i}\right)$$
$$= 2\sum n_i \log\left(\frac{n_i}{n\pi_{i0}}\right)$$
$$= G^2$$

5. Agresti 2.3

An article in The New York Times (Feb. 17, 1999) about the PSA blood test for detecting prostate cancer stated: "The test fails to detect prostate cancer in 1 in 4 men who have the disease (false-negative results), and as many as two-thirds of the men tested receive false-positive results." Let $C(\bar{C})$ denote the event of having (not having) prostate cancer and let

$+(-)$ *denote a positive (negative) result. Which is true:* $P(-|C) = \dfrac{1}{4}$ *or* $P(C|-) = \dfrac{1}{4}$ *?*

$P(\bar{C}|+) = \dfrac{2}{3}$ *or* $P(+|\bar{C}) = \dfrac{2}{3}$ *?*

Determine the sensitivity and specificity.

$P(-|C) = 1/4$

It is unclear from the wording but presumably this means that $P(\bar{C}|+) = 2/3$

Sensitivity $= P(+|C) = 1 - P(-|C) = 3/4$

Specificity $= P(-|\bar{C}) = 1 - P(+|\bar{C})$ cannot be determined from information given.

6. Agresti 2.4

Table 2.10 shows fatality results for drivers and passengers in auto accidents in Florida in 2008, according to whether the person was wearing a seatbelt.

	Injury	
Seatbelt Use	Fatal	Nonfatal
No	1085	55,623
Yes	703	441,239

   a. *Estimate the probability of fatality, conditional on seatbelt use in category (i) no and (ii) yes.*

$P(\text{fatality|seatbelt=no}) = 1085/(1085 + 55623) = 0.0191$

$P(\text{fatality|seatbelt=yes}) = 703/(703 + 441239) = 0.0016$

   b. *Estimate the probability of wearing a seatbelt, conditional on the injury being (i) fatal and (ii) nonfatal.*

$P(\text{seatbelt=yes|fatal}) = 703/(1085 + 703) = 0.393$

$P(\text{seatbelt=yes|nonfatal}) = 441239/(441239 + 55263) = 0.888$

   c. *For the most natural choice of response variable, find and interpret the difference of proportions, relative risk, and odds ratio. Why are the relative risk and odds ratio approximately equal?*

For a response of fatality

Risk Difference: $P(\text{fatality|seatbelt=no}) - P(\text{fatality|seatbelt=yes}) = 0.017$

Relative Risk: $\dfrac{P(\text{fatality|seatbelt=no})}{P(\text{fatality|seatbelt=yes})} = 11.8$

Odds Ratio: $\dfrac{(1085 \times 441239)}{(703 \times 55623)} = 12.2$

The proportion of fatal injuries is close to zero for each row, so the odds ratio is similar to the relative risk.

7. Agresti 2.11

A research study estimated that under a certain condition, the probability that a subject would be referred for heart catheterization was 0.906 for whites and 0.847 for blacks.

   a. *A press release about the study stated that the odds of referral for cardiac catheterization for blacks are 60% of the odds for whites. Explain how they obtained 60% (more accurately, 57%).*

$(0.847/0.153)/(0.906/0.094) = 0.574$

   b. *An Associated Press story later described the study and said "Doctors were only 60% as likely to order cardiac catheterization for blacks as for whites." Explain what is wrong with this interpretation. Give the correct percentage for this interpretation.*

This is interpretation for relative risk, not the odds ratio. The actual relative risk is 0.847/0.906=0.935; i.e., 60% should have been 93.5%. What is really wrong is a journalist that could barely pass a 200 level statistics course is reporting on statistics and has no idea what they are looking at for research studies.

8. Agresti 2.12

A 20-year cohort study of British male physicians noted that the proportion per year who died from lung cancer was 0.00140 for cigarette smokers and 0.00010 for nonsmokers. The proportion who died from coronary heart disease was 0.00669 for smokers and 0.00413 for nonsmokers.

   a. *Describe the association of smoking with each of lung cancer and heart disease, using the difference in proportions, relative risk, and odds ratio. Interpret.*

Relative risk: Lung cancer, 14.00; Heart disease, 1.62. Cigarette smoke seems more highly associated with lung cancer.

Difference of Proportions: Lung cancer, 0.00130; Heart disease, 0.00256. Cigarette smoking seems more highly associated with heart disease.

Odds Ratio: Lung cancer, 14.02; Heart disease, 1.62 *e.g.,* the odds of dying from lung cancer for smokers are estimated to be 14.02 times those for nonsmokers.

   b. *Which response is more strongly related to cigarette smoking, in terms of the reduction in number of deaths that would occur with elimination of cigarettes? Explain.*

Difference of proportions describes excess deaths due to smoking. That is, if N = no. smokers in population, we predict there would be 0.00130N fewer deaths per year from lung cancer if they had never smoked, and 0.00256N fewer deaths per year from heart disease. Thus elimination of cigarette smoking would have the biggest impact on deaths due to heart disease (but cigarettes are sooooo good).