

Midterm Exam II

Instructions: This is a close-book and close-notes exam. You are only allowed to bring three cheat sheets. Arrange your time wisely and show all your work. Details and explanations are required in order to receive partial/full credit. Note that the total score of this exam is 110, including ten extra credits

1. The following table comes from one of the first studies of the link between lung cancer and smoking, by Richard Doll and A. Bradford Hill. In twenty hospitals in London, UK, patients admitted with lung cancer in the previous year were queried about their smoking behavior. For each patient admitted, researchers studied the smoking behavior of a noncancer control patient at the same hospital of the same sex and within the same five-year grouping on age. A smoker was defined as a person who had smoked at least one cigarette a day for at least a year.

Smoking	Lung	
	Cancer	Non-Cancer
smoker	254	179
non-smoker	455	530
Total	709	709

- (a) (3 points) Identify the type of study this was.

Solution: Retrospective case-control study.

- (b) (9 points) Can you use these data to compare smokers within nonsmokers in terms of the proportion who suffered lung cancer? Why or why not?

Solution: No. What is available from the above retrospective study is merely $\Pr(\text{smoker} \mid \text{cancer})$ and $\Pr(\text{smoker} \mid \text{non-cancer})$. In order to estimate $\Pr(\text{cancer} \mid \text{smoker})$, we need the prevalence rate of lung cancer, i.e., $\Pr(\text{cancer})$, in view of

$$\begin{aligned}
 \Pr(\text{cancer} \mid \text{smoker}) &= \frac{\Pr(\text{cancer} \ \& \ \text{smoker})}{\Pr(\text{smoker})} \\
 &= \frac{\Pr(\text{smoker} \mid \text{cancer}) \cdot \Pr(\text{cancer})}{\Pr(\text{smoker} \ \& \ \text{cancer}) + \Pr(\text{smoker} \ \& \ \text{noncancer})} \\
 &= \frac{\Pr(\text{smoker} \mid \text{cancer}) \cdot \Pr(\text{cancer})}{\Pr(\text{smoker} \mid \text{cancer}) \cdot \Pr(\text{cancer}) + \Pr(\text{smoker} \mid \text{noncancer}) \cdot \Pr(\text{noncancer})}.
 \end{aligned}$$

- (c) (12 points) Construct 95% confidence interval for the odds ratio of having lung cancer between smokers and nonsmokers.

Solution: First find 95% CI for $\log \theta$:

$$\log(1.653) \pm 1.96 \times \sqrt{1/254 + 1/179 + 1/455 + 1/530} \implies (0.2739, 0.7312)$$

Thus, 95% CI for θ is $(\exp(0.2739), \exp(0.7312)) \implies (1.215, 2.077)$.

2. A study on educational aspirations of high school students (S. Crysdale, *Int. J. Comp. Social.*, **16**:19–36, 1975) measured aspirations using the scale (high school, some college, college graduate). For students whose family income was low, the counts in these categories were (53, 13, 10); when family income was middle, the counts were (63, 23, 22); when family income was high, the counts were (50, 12, 27).

(a) (16 points) With significance level $\alpha = 0.05$, test Independence of aspirations and family income using both Pearson's χ^2 and the likelihood ratio χ^2 .

Solution: First find the expected counts $E_{ij} = n_{i+}n_{+j}/n$:

53 (46.21)	13 (13.36)	10 (16.43)
63 (65.67)	23 (18.99)	22 (23.34)
50 (54.12)	12 (15.65)	27 (19.23)

The Pearson and LRT χ^2 test statistics are 8.856 and 8.901, respectively, both greater than $\chi_{0.95}^2(4) = 9.448$.

(b) (10 points) Find the standardized Pearson residuals. Do they suggest any association pattern?

Solution: The standardized Pearson residuals, defined as

$$\varepsilon_{ij} = \frac{n_{ij} - E_{ij}}{\sqrt{E_{ij}(1 - P_{i+})(1 - P_{+j})}},$$

helps find out the detailed information about the association by comparing n_{ij} with its expected value E_{ij} . The following table lists the computed ε_{ij} values:

1.89	-0.12	-2.11 (*)
-0.62	1.31	-0.37
-1.06	-1.24	2.36 (*)

Note that we compare ε_{ij} with $z_{0.975} = 1.96$ or roughly 2. It can be seen that only two of them are significant.

The first one is ε_{13} with the negative sign, showing that n_{13} is significantly lower than its expected count E_{13} under H_0 . In other words, this implies that less students from low-income families tend to have “college” as their educational aspiration.

The other one is ε_{33} with the positive sign, showing that n_{13} is significantly more than its expected count E_{13} under H_0 . In other words, this implies that more students from high-income families tend to have “college” as their educational aspiration.

3. (14 points) A group of 11 people were classified according to age (older than 50 or not) and political party identification. Apply Fisher's exact test to see whether elder people tends *more* to be republicans with significance level $\alpha = 0.05$. Report both the P-value and the mid P-value.

Party Identification			
Age	Republican	Democrat	Total
≥ 50	3	2	5
< 50	1	5	6
Total	4	7	11

Solution: If H_a is true, *i.e.*, elder people tends *more* to be republicans, then we shall expect a large n_{11} . Thus, with $n_{11} \sim \text{Hypergeometric}(5, 6, 4)$,

$$P\text{-value} = P(n_{11} \geq 3) = P(3) + P(4) = \frac{5! \cdot 6! \cdot 4! \cdot 7!}{11! \cdot (3! \cdot 2! \cdot 1! \cdot 5!)} + \frac{5! \cdot 6! \cdot 4! \cdot 7!}{11! \cdot (4! \cdot 1! \cdot 0! \cdot 6!)}.$$

Given fixed marginal totals, The resultant $z \times 2$ tables when $n_{11} = 3$ and $n_{11} = 4$ are given below:

3	2		5		4	1		5
1	5		6		0	6		6

4	7		11		4	7		11

4. The following table refers to the effect of academic achievement on self-esteem among male and female college students. Assume that the effect of academic achievement on self-esteem does not interact with gender.

Gender	Cumulative GPA	Self Esteem	
		High	Low
Males	High	32	19
	Low	48	43
Females	High	35	54
	Low	27	40

- (a) (12 points) Applying the Cochran-Mantel-Haenszel Test to see whether there is a strong association between cumulative GPA and self esteem, while adjusting for gender.

Solution: It can be found that $CMH = 0.625$, which is less than $\chi_{0.95}^2(1) = 3.84$. Thus it seems that there is no significant association between Cumulative GPA and Self Esteem, while adjusting for gender.

- (b) (8 points) Compute Mantel-Haenszel's estimate of the common odds ratio, in two genders, of having high self esteem comparing the high with low GPA groups.

Solution: $\theta_{MH} = 1.18$.

5. The following table shows results when subjects were asked "Do you think a person has the right to end his or her own life if this person has an incurable disease?" and "when a person has a disease that cannot be cured, do you think doctors should be allowed to end the patient's life by some painless means if the patient and his family request it?" The table refers to these variables as "suicide" and "let patient die." (Source: 1994 General Social Survey, National Opinion Research Center.)

- (a) (10 points) Compare the marginal proportions using a 95% confidence interval.

Solution: 95% CI for $(\pi_{1+} - \pi_{+1})$ is

$$(-0.0619 \pm 1.96 \times \frac{\sqrt{(90 + 203) - \frac{(90 - 203)^2}{1825}}}{1825}) \implies (-0.08, -0.04).$$

	Let Patient Die	
Suicide	Yes	No
Yes	1097	90
No	203	435

(b) (10 points) Perform McNemar's test and interpret.

Solution: McNemar's test is the approach based on the score test.

$$\chi^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}} = \frac{(90 - 203)^2}{90 + 203} = 43.5 > \chi_{0.95}^2(1) = 3.84.$$

6. (10 points) Consider the statistical inference on the logarithm of the odds ratio, $\log(\theta)$. Carefully derive that the variance for its estimate $\log(\hat{\theta})$ obtained from a 2×2 table $\{n_{ij}\}$ is given as

$$\text{var}(\log(\hat{\theta})) = 1/n_{11} + 1/n_{12} + 1/n_{21} + 1/n_{22}.$$

Solution: Omitted. See the in-class lecture notes.