Exam I (2022)

Introduction to Categorical Data Analysis

1. (25 points) In medicine, diagnostic tests are often used to detect whether an individual has a particular condition or disease. Most diagnostic tests are imperfect and will produce both false positive and false negative results. For such tests, the *predictive value positive,* denoted PV^+, refers to the probability that a subject has the condition/disease given that the test result is positive. PV^+ can be estimated provided that (i) estimates of the *sensitivity* and *specificity* of the test are available, and (ii) an estimate of the prevalence of the condition/disease in the underlying population is available. (The sensitivity refers to the probability of a positive test result given that the subject has the condition/disease; the specificity refers to the probability of a negative test result given that the subject does not have the condition/disease.)

   ELISA (Enzyme-Linked Immunosorbent Assay) tests are used to screen donated blood for HIV, the virus which causes AIDS. The test checks for the presence of an antibody produced when an individual is exposed to HIV. To evaluate the sensitivity and the specificity of the ELISA test, suppose that the test is administered to a group of 500 individuals known to be HIV positive and to a group of 500 individuals known to be HIV negative. (The true HIV status of the individuals could be determined by using a more sophisticated diagnostic test than ELISA, such as the Western Blot test.) Assume the following results are obtained:

	HIV Status	
ELISA Test Result	Positive (HIV)	Negative (Non HIV)
Positive (+)	487	17
Negative (-)	13	483
Totals	500	500

   Refer to the 2×2 table in answering the following questions.

   (a) Use the preceding data to obtain estimates of the sensitivity and the specificity of the ELISA test.

   (b) PV^+ for the ELISA test cannot be directly estimated from the preceding data. Briefly explain why.

   (c) Assume that the prevalence of HIV in the population of interest is estimated as 0.004. Starting with the relation

   $$\mathrm{PV}^+ = P(\mathrm{HIV}|+) = \frac{P(\mathrm{HIV} \cap +)}{P(+)},$$

   derive a relation which expresses PV^+ in terms of **only** the prevalence, the sensitivity, and the specificity. Use this relation to estimate PV^+.

(d) Suppose the ELISA test is used to screen a large collection of donated blood samples for the presence of HIV. What are the implications of the result in part (c) regarding the efficacy of the test when there is a low prevalence of HIV among the samples?

2. (20 points) Consider the following data from a women's health study (MI is myocardial infarction, i.e., heart attack).

	MI	
Oral Contraceptives	Yes	No
Used	23	34
Never Used	35	132

(a) Construct a 95% confidence interval for the population odds ratio.

(b) Suppose that the answer to part (a) is (1.3, 4.9). Does it seem plausible that the variable are independent? Explain.

3. (30 points) For a 2×2 contingency table, let $Q = \frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{\pi_{11}\pi_{22} + \pi_{12}\pi_{21}}$.

(a) Derive a relationship between Q and the odd ration $\theta = \pi_{11}\pi_{22}/\pi_{12}\pi_{21}$.

(b) How can the value of Q be interpreted? (Hint: use the result in (a))

(c) For multinominal sampling, $var(\log \hat{\theta}) \approx \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{1}{\pi_{ij}}$. Using this result and the delta-method, derive the asymptotic variance of $\hat{Q}$ of Q.

4. (25 points) For the 23 space shuttle flights that occurred before the Challenger mission in 1986, The following table shows the temperature (^oF) at the time of the flight and whether at least one of six primary O-rings suffered thermal distress (1=yes, 0=no). The first attahed SAS printout shows the use of various models for analyzing these data.

Ft	Temp	TD	Ft	Temp	TD	Ft	Temp	TD	Ft	Temp	TD
1	66	0	2	70	1	3	69	0	4	68	0
5	67	0	6	72	0	7	73	0	8	70	0
9	57	1	10	63	1	11	70	1	12	78	0
13	67	0	14	53	1	15	67	0	16	75	0
17	70	0	18	81	0	19	76	0	20	79	0
21	75	1	22	76	0	23	58	1			

(a) For the logistic regression model using temperature as a predictor for the probability of thermal distress, calculate the estimated probability of thermal distress at 31 ^oF, the temperature at the time of the Challenger flight.

(b) At the temperature at which the estimated probability equals 0.5, give a linear approximation for the change in the estimated probability per degree increase in temperature.

(c) Interpret the estimated effect of temperature on the odds of thermal distress.

(d) Test the hypothesis that the temperature has no effect, using the likelihood-ratio test. Interpret results. ($\chi^2_{0.05,df=1} = 3.8415$; $\chi^2_{0.05,df=2} = 5.9915$)

```
-----------------------------------------------------------------------------
Model 1
                  Deviance and Pearson Goodness-of-Fit Statistics

Criterion                Value      DF       Value/DF
Deviance               20.3152      21         0.9674
Pearson Chi-Square     23.1691      21         1.1033
Log likeliood         -10.1576       .

                     Analysis of Parameter Estimates
                                             Standard
Parameter        DF          Estimate        Error    Chi-Square  Pr>ChiSq
Intercept         1           15.0429        7.3789       4.1563    0.0415
Temp             1           -0.2322        0.1082     4.6008       0.0320

-----------------------------------------------------------------------------
Model 2
                  Deviance and Pearson Goodness-of-Fit Statistics

Criterion               Value      DF       Value/DF
Deviance              28.2672      22         1.2849
Pearson Chi-Square    23.0000      22         1.0455
Log likeliood        -14.1336       .

                     Analysis of Parameter Estimates
                                             Standard
Parameter        DF          Estimate        Error    Chi-Square  Pr>ChiSq
Intercept         1           15.0429        7.3789       4.1563    0.0415
```