STAT 426 — Sections 1GR, 1UG — Spring 2017

Exam 1

February 22, 2017

Full Name: Key ID/Email: ______

- This is an 80 minute exam. There are 5 problems, one of which is for the graduate section only.

  The exam is worth a total of 42 points for the undergraduate section and 52 points for the graduate section.

- You may use *one* physical page of personal notes and a standard scientific calculator. (You may *not* share these items with anyone else.)

- *Write all answers in the spaces provided.* If you require more space to write your answer, you may use the back side of the page.

- You are not allowed to communicate with anyone except the instructor or proctors before you submit this exam.

Useful Abbreviations:

CI = confidence interval
SE = standard error
E = expected value Var = variance Cov = covariance
df = degrees of freedom
ML = maximum likelihood LRT = likelihood ratio test L = log-likelihood
RR = relative risk
H_0 = the null hypothesis of a test H_a = the alternative hypothesis of a test

1. The ill-fated Donner party became trapped in the Sierra Nevada mountains during the winter of 1846-47. Of 35 females in the party, 10 died. Of 56 males, 32 died.

   (a) Form a contingency table based on sex (Female, Male) and survival status (Died, Survived), with appropriate labels. [3 pts]

	Died	Survived
Female	10	25
Male	32	24

   (b) Estimate the *overall* probability of survival. [1 pt]

$$\frac{25+24}{35+56} \approx 0.538$$

   (c) Estimate the *odds* of death for the females, and also for the males. [2 pts]

Females: $\frac{10}{25} = 0.4$ Males: $\frac{32}{24} \approx 1.33$

   (d) Estimate the *risk* of death for a female relative to a male. [2 pts]

$$\frac{10/35}{32/56} = 0.5$$

   (e) Form an approximate 95% confidence interval for the *log*-odds ratio of death for a female relative to a male. [4 pts]

$$\ln \hat{\theta} = \ln(10 \cdot 24 / 32 \cdot 25) \approx -1.204$$

$$\hat{\sigma}(\ln \hat{\theta}) = \sqrt{\frac{1}{10} + \frac{1}{25} + \frac{1}{32} + \frac{1}{24}} \approx 0.4614$$

$$\ln \hat{\theta} \pm z_{0.025}\, \hat{\sigma}(\ln \hat{\theta})$$

$$\approx -1.204 \pm 1.96 \cdot 0.4614$$

$$\approx (-2.11, -0.30)$$

2. Cross-classifying a sample of U.S. individuals according to gender and political party preference yields the following table:

	Party Preference		
	Democrat	Independent	Republican
Female	422 (393.41)	381 (407.05)	273 (275.55)
Male	299 (327.59)	365 (338.95)	232 (229.45)

The numbers in parentheses are the estimated *expected* counts under the assumption that gender and party preference are independent.

(a) Demonstrate how the expected count for female Democrats (393.41) was computed. [3 pts]

$$n = \sum n_{ij} = 1972$$

$$\hat{\pi}_{1+} = (422+381+273)/n \approx 0.5456$$

$$\hat{\pi}_{+1} = (422+299)/n \approx 0.3656$$

$$n\,\hat{\pi}_{1+}\,\hat{\pi}_{+1} \approx 393.4$$

(b) Compute all of the Pearson residuals (for the test of independence). [3 pts]

$$e_{11} \approx \frac{422-393.41}{\sqrt{393.41}} \approx 1.44 \qquad e_{12} \approx -1.29 \qquad e_{13} \approx -0.15$$

$$e_{21} \approx -1.58 \qquad e_{22} \approx 1.41 \qquad e_{23} \approx 0.17$$

(c) Compute the Pearson chi-squared statistic for testing independence. [2 pts]

$$\sum e_{ij}^2 \approx 8.3$$

(d) Would the null distribution of the Pearson statistic be approximately chi-square in this case? Why or why not? [2 pts]

Yes. All counts (and expected counts) are sufficiently large.

3. Suppose three binary variables (X, Y, Z) are sampled as in a cross-sectional study, such that the *mean* counts are as in the following stratified tables:

$Z = 1$:

	$Y = 1$	$Y = 2$
$X = 1$	$\mu_{111} = 6$	$\mu_{121} = 20$
$X = 2$	$\mu_{211} = 5$	$\mu_{221} = 50$

$Z = 2$:

	$Y = 1$	$Y = 2$
$X = 1$	$\mu_{112} = 5$	$\mu_{122} = 2$
$X = 2$	$\mu_{212} = 50$	$\mu_{222} = 60$

(a) Compute $\theta_{XY(1)}$ and $\theta_{XY(2)}$. [2 pts]

$$\theta_{XY(1)} = \frac{6 \cdot 50}{5 \cdot 20} = 3 \qquad \theta_{XY(2)} = \frac{5 \cdot 60}{50 \cdot 2} = 3$$

(b) Compute θ_{XY}. [2 pts]

$$\theta_{XY} = \frac{(6+5)(50+60)}{(5+50)(20+2)} = 1$$

(c) Are X and Y independent? Why or why not? [2 pts]

Yes, since their (marginal) odds ratio θ_{XY} is 1.

(d) Are X and Y conditionally independent, given Z? Why or why not? [2 pts]

No. For example, the conditional odds ratio of X and Y given Z=1 is 3, not 1.

(e) Is the association between X and Y homogeneous (over Z)? Why or why not? [2 pts]

Yes, since $\theta_{XY(1)} = \theta_{XY(2)}$.

4. George studies fraud for a credit card company. He randomly selects 100 fraudulent transactions and 100 legitimate transactions from a database of past transactions. He finds that 56 of the fraudulent transactions were made more than 50 miles from the home address, while only 5 of the legitimate transactions were.

(a) Which term best describes the design of George's study: retrospective, prospective, or cross-sectional? Explain briefly. [2 pts]

Retrospective, since he samples fixed numbers of each level of the condition of interest (fraud) (Also, since the explanatory variable distance is determined from past data.)

(b) Consider a test that classifies a transaction as fraudulent if it is made more than 50 miles from home. Estimate its sensitivity and specificity. [2 pts]

sensitivity: $P(>50 \mid \text{fraud}) \approx 0.56$

specificity: $P(\leq 50 \mid \text{not fraud}) \approx 0.95$

(c) George wants to use his study to estimate the probability that a transaction made more than 50 miles from home is fraudulent. What should you tell him? Why? [2 pts]

That is not possible with his data, since it would require information giving the overall rate of fraud (which can't be determined from retrospective sampling).

(d) Past data reveals that 0.1% of transactions are actually fraudulent. Given the sensitivity and specificity from part (b), what proportion of *all* transactions will the test classify as fraudulent? [4 pts]

$P(\text{fraud}) = 0.001$

$$P(>50) = P(>50 \mid \text{fraud})\, P(\text{fraud}) + P(>50 \mid \text{not fraud})\, P(\text{not fraud})$$

$$\approx 0.56 \cdot 0.001 + (1-0.95)(1-0.001)$$

$$\approx 0.0505$$

GRADUATE SECTION ONLY

5. Let $Y_1, \ldots, Y_n$ be a sample from a Poisson distribution with unknown mean μ.

(a) Write out the log-likelihood function (simplifying if possible). [2 pts]

$$L(\mu) = -\sum_{i=1}^{n} \ln y_i! + (\ln \mu) \sum_{i=1}^{n} y_i - n\mu$$

$$= \text{constant} + (\ln \mu) \sum_{i=1}^{n} y_i - n\mu$$

(b) Derive the maximum likelihood estimator. [2 pts]

$$L'(\mu) = \sum_{i=1}^{n} y_i / \mu - n$$

$$L'(\hat{\mu}) = 0 \Rightarrow \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} y_i = \bar{y}$$

(c) Derive the (Fisher) information. [2 pts]

$$L''(\mu) = -\sum_{i=1}^{n} y_i / \mu^2$$

$$i(\mu) = E(-L''(\mu)) = E\left(\sum_{i=1}^{n} Y_i\right) / \mu^2$$

$$= n\mu/\mu^2 = n/\mu$$

(d) Form the Wald z statistic for testing $H_0 : \mu = \mu_0$. [2 pts]

$$z = \frac{\hat{\mu} - \mu_0}{1/\sqrt{i(\hat{\mu})}} = \frac{\hat{\mu} - \mu_0}{\sqrt{\hat{\mu}/n}} = \frac{\bar{y} - \mu_0}{\sqrt{\bar{y}/n}}$$

(e) Form the expression for a Wald approximate 95% confidence interval for μ. [2 pts]

$$\hat{\mu} \pm 1.96 \sqrt{\hat{\mu}/n}$$

$$= \bar{y} \pm 1.96 \sqrt{\bar{y}/n}$$