Bios 765: Homework # 4, Fall 2020
due October 28, 2020

1. **Applied (10 pts.) Mean Score tests. Table A** reports counts for the survival time (hours) of mice to a drug challenge.

   Table A.

Survival Time	Drug A	Drug B	total
0-6	1	1	2
6-12	3	1	4
12-18	5	1	6
18-24	1	0	1
24-30	1	2	3
30-48	0	2	2
48-72	1	1	2
72-96	0	1	1
> 96	0	1	1
Total	12	10	22

   (a) Compute the standardized midrank scores for the nine categories of the response, survival time. Also, compute the logrank scores that place greater emphasis on group differences occurring at larger survival times. Report these side-by-side in a table.

   (b) Report the test statistic and asymptotic chi-square p-value for the Wilcoxon Rank Sum Test, and draw a conclusion based upon your result.

   (c) Report the test statistic and asymptotic chi-square p-value for the Randomization Chi-square Test with Logrank scores (computed in part a), and draw a conclusion based upon your result.

2. **Applied (15 pts.) Exact Inference.** In the 2002 Winter Olympic Games held at Salt Lake City there was concern that figure skating judges may have judged with bias for certain skaters according to geopolitical preferences. Consider **Table B** which presents the results of the 9 judges of the "long program" in women's figure skating. All judges rated these two skaters first or second; the first place rating, or preferred skater, is reported in the table. Each judge is placed into a region based upon her/his country of origin: EE refers to Belarus, Russia and Slovakia; WE refers to Denmark, Finland, Italy, and Germany; and NA refers to Canada and the United States.

TABLE B. Number of First place ratings

	Preferred Skater		
Region	Slutskaya	Hughes	Total
EE	3	0	3
WE	1	3	4
NA	0	2	2
Total	4	5	9

a. Collapse **Table B** into a 2×2 table by combining WE and NA into a single category. Fixing the row and column margins of the table to their observed values, determine the reference set of tables and their probabilities giving the permutation distribution of outcomes under the null hypothesis of no association between judges' origins and preferred skater. What is the p-value for the two-sided Fisher's exact test ("standard implementation" in course notes). Relatedly, what is the p-value for the exact Pearson chi-square test?

b. For **Table B** (the 3×2 table), determine the reference set of tables and their probabilities giving the permutation sample space under the null hypothesis of no association between judges' origins and preferred skater. What is the Fisher's exact test p-value for **Table B**?

c. For **Table B** (the 3×2 table), what is the Pearson Chi-square statistic (Q_P) exact p-value? Which tables in the reference set have a value of Q_P at least as large as that of the observed table?

d. Considering that Hughes is an American skater and Slutskaya is a Russian skater, treat region as an ordinal variable having table scores (1=EE,2=WE,3=NA). Determine the value of Q_S, the mean score statistic for each table in the reference set. What is the exact p-value for Q_S?

e. Briefly, summarize your results for a report to the International Olympic Committee (IOC). What do you conclude about the charge of geopolitically-based bias in figure skating judging? Explain any assumptions of the statistical methods that led you to this conclusion (the IOC may consult a statistician to independently review your findings.)

3. **(Applied, 15 pts.) Stratified logistic regression:** The data in **Table C** are for 18 matched pairs of Hispanic migrant farmworkers sampled as part of a case/control study on green tobacco sickness (GTS) conducted in North Carolina in 1999.

TABLE C. Matched pairs (1=Yes, 0=No)

	case		control	
pair	English	rainsuit	English	rainsuit
1	1	1	1	0
2	1	0	1	0
3	0	0	1	0
4	0	0	1	0
5	0	0	1	0
6	1	0	0	1
7	0	0	1	0
8	1	0	0	1
9	1	0	0	0
10	0	0	1	0
11	0	0	1	1
12	0	0	1	1
13	0	0	1	0
14	0	0	1	1
15	0	0	0	1
16	0	0	0	1
17	1	0	1	1
18	1	1	1	1

GTS is an occupational illness reported by tobacco workers worldwide, characterized by nausea or vomiting and dizziness or headache. Within a matched pair, both the case and control reported to the same health care clinic for a health complaint on the same day after having worked in tobacco. The case had symptoms which led to a diagnosis of GTS. The control did not have GTS. Researchers were interested in whether workers who did not wear rainsuits during work were more likely to get GTS than workers who wore rainsuits. Whether or not a worker understands English is a covariate.

a. Perform conditional logistic regression for the following model

$$\log\Big(\frac{\pi_{ij}}{1-\pi_{ij}}\Big) = \alpha_i + x_{ij}\beta \quad i = 1, \ldots, 18; j = 1, 2$$

where π_{ij} is the probability that the j-th subject from the i-th matched pair had a diagnosis of GTS; $x_{ij} = 1$ if that person reported wearing a rainsuit, and $x_{ij} = 0$ if he did not wear a rainsuit. Give the conditional maximum likelihood estimate of the odds ratio and its asymptotic 95% confidence interval for the association of not wearing a rainsuit and GTS.

b. Apply exact inference to the model in part **a.** Give the exact conditional maximum likelihood estimate of the odds ratio and its corresponding "exact" 95% confidence interval.

c. Perform conditional logistic regression with case/control status as response and both raincoat and English as covariates. Write down the stratified logistic model treating matched pairs as strata, defining all notation. Give the conditional maximum likelihood estimate of the odds ratio and its corresponding asymptotic 95% confidence interval for each covariate.

d. Apply exact inference to the model in part **c.** For each covariate, give the exact conditional maximum likelihood estimate of the odds ratio and its corresponding "exact" 95% confidence interval.

e. Based upon the exact inference results of parts **b.** and **d.**, what do you conclude about the relationship between not wearing a raincoat and GTS?

4. **(Theory, 10 pts) Conditional logistic regression:**

a. Write down the conditional likelihood for the stratified logistic regression model in problem **3a** and the data in **Table C.** Using the observed data, reduce the expression to its simplest form. Show all work.

b. Determine the conditional maximum likelihood estimate $\hat{\beta}$ by maximizing the conditional likelihood in part **a.** Show all steps leading to your numerical answer.