STAT 5107 Discrete Data Analytics
Final test

6 December 2019

Time: solutions should be submitted on or before 11:59pm, 8 December 2019

Department of Statistics, The Chinese University of Hong Kong

1 (**12 marks**) Which scale of measurement is most appropriate for the following variables – nominal, or ordinal?

(a) Political party affiliation (Democrat, Republican, unaffiliated).

(b) Highest degree obtained (none, high school, bachelor's, master's, doctor- ate).

(c) Patient condition (good, fair, serious, critical).

(d) Hospital location (London, Boston, Madison, Rochester, Toronto).

(e) Favorite beverage (beer, juice, milk, soft drink, wine, other).

(f) How often feel depressed (never, occasionally, often, always).

2 (**10 marks**) Each of 100 multiple-choice questions on an exam has four possible answers but one correct response. For each question, a student randomly selects one response as the answer.

(a) Specify the distribution of the student's number of correct answers on the exam.

(b) Based on the mean and standard deviation of that distribution, would it be surprising if the student made at least 50 correct responses? Explain your reasoning.

3 (**28 marks**) For a 2×2 table of counts $\{n_{ij}\}$ in row i and column j,

(a) show that the sample estimate of the odds ratio is invariant to interchanging rows with columns;

(b) show that the sample estimate of the odds ratio is invariant to multiplication of cell counts within rows or within columns by a constant $c \neq 0$;

(c) show that the difference of proportions and the relative risk do not have the properties in part (a) and part (b);

(d) let π_{ij} be the probability of an observation falls in the (i, j)-th cell. Show that the odds ratio $\theta = (\pi_{11}\pi_{22})/(\pi_{12}\pi_{21}) = 1$ if and only if the row variable X and the column variable Y are independent.

4 (**15 marks**) A newspaper article preceding the 1994 World Cup semifinal match between Italy and Bulgaria stated that "Italy is favored 10-11 to beat Bulgaria, which is rated at 10-3 to reach the final." Suppose this means that the odds that Italy wins are 11/10 and the odds that Bulgaria wins are 3/10. Find the probability that each team wins, and comment.

5 (**15 marks**) We continue the analysis of the horseshoe crab data by using both the female crab's shell width and color as predictors. To treat color as a nominal-scale predictor, we use three indicator variables for the four categories. The model is

$$logit[P(Y = 1)] = \alpha + \beta_1 c_1 + \beta_2 c_2 + \beta_3 c_3 + \beta_4 x,$$

where x denotes width and

$c_1 = 1$ for color = medium light, 0 otherwise

$c_2 = 1$ for color = medium, 0 otherwise

$c_3 = 1$ for color = medium dark, 0 otherwise.

Model above for the probability π of a satellite for horseshoe crabs with color and width predictors has fit

$$logit(\hat{\pi}) = -12.715 + 1.330c_1 + 1.402c_2 + 1.106c_3 + 0.468x$$

Consider this fit for crabs of width $x = 20$ cm.

(a) Estimate π for medium-dark crabs ($c_3 = 1$) and for dark crabs ($c_1 = c_2 = c_3 = 0$). Then, estimate the ratio of probabilities.

(b) Estimate the odds of a satellite for medium-dark crabs and the odds for dark crabs. Show that the odds ratio equals $\exp(1.106) = 3.02$. When each probability is close to zero, the odds ratio is similar to the ratio of probabilities, providing another interpretation for logistic regression parameters. For widths at which $\hat{\pi}$ is small, $\hat{\pi}$ for medium-dark crabs is about three times that for dark crabs.

6 **(20 marks)** Consider binary outcome $Y \in \{0, 1\}$, where $Y = 0$ represents the control (non-disease) and $Y = 1$ represents the case (disease). The logistic regression model assumes that Y is associated with the $(p + 1)$-dimensional predictor X via the logistic link function

$$P(Y = 1|X = x) = \frac{e^{x^\top \beta}}{1 + e^{x^\top \beta}}, \tag{1}$$

where $\beta \in R^{p+1}$ including an intercept. A latent variable formulation is as follows: suppose that there is an unobserved continuous random variable $\tilde{Y}$ such that $Y = 1$ if and only if $\tilde{Y} > \theta$, where θ is certain unknown constant. It is known that θ is not identifiable under model (1). For example, $Y = 1$ if and only if $c\tilde{Y} > c\theta$ for some $c > 0$.

(a) **(20 marks)** For identifiability, we fix $\theta = 0$ without loss of generality. Assume that the latent $\tilde{Y}$ depends on X via a linear regression model

$$\tilde{Y} = X^\top \beta + U, \tag{2}$$

where U is the error term. Show that when U follows the standard logistic distribution, model (2) is the logistic regression model in (1).

Hint: The density function of the standard logistic distribution is

$$f(x) = \frac{e^x}{(e^x + 1)^2}, \quad x \in R.$$

(b) **(Bonus 10 marks)** For the logistic regression model (1), the observations are $(Y_i, X_i), i = 1, \ldots, n$, a random sample of (Y, X) with size n. Let $\hat{\beta}$ be the maximum likelihood estimator of β. Derive the asymptotic distribution of $\sqrt{n}(\hat{\beta} - \beta_0)$ as $n \to \infty$, where β_0 is the true value of β in model (1). If it is possible, please provide the technical assumptions needed to establish the asymptotic distribution.