NAME: _____ **EXAM NUMBER:** _____

PRELIMINARY COMPREHENSIVE EXAMINATION

Department of Biostatistics

8:30 AM - 11:30 AM

September 23rd, 2019

Instructions:

This three hour examination consists of five questions numbered 1-5. Answer all five questions. Each question is worth the same number of points.

This is a closed book, closed notes exam. You may use a calculator and the formula sheets and tables that have been provided.

You may begin your answer on the sheet on which the question is typed and continue on additional pages if necessary. Do not write the answer to more than one question on a single page. Do not write on the back of any sheet. Be sure to write your exam number on each sheet you turn in, and if you use additional sheets be certain to mark them carefully with the question number as well as your exam number.

It is your responsibility to check that you have the whole exam. There are 6 question pages.

Question 1

A pediatric clinic is interested in reducing the rate of missed opportunities to vaccinate its patients for influenza. An opportunity is defined as a clinic visit by a patient who has not been vaccinated for influenza. An opportunity is missed if the unvaccinated patient leaves the clinic without receiving the vaccination and not missed if the patient receives the vaccination.

In June of 2018, the clinic implemented a new program designed to decrease the rate of missed opportunities. To evaluate whether the program has had an effect, the clinic administration has assembled a dataset that contains the following information for each patient visit that was an opportunity to vaccinate, covering the 5 months before and the 5 months after the start of the program:

Month of visit: Jan 2018 through Oct 2018, coded as $t = 0$ through $t = 9$
Sex: 1 =male, 0 =female
MO: 1 = missed opportunity and 0 = not.

Assume that no patient appears in the data set more than once and that each observation is independent. The overall rates of missed opportunities by month are as follows:

Month	t	s	Number of visits that were opportunities	Number of missed opportunities	Percent of opportunities that were missed
Jan 2018	0	0	1484	484	32.7%
Feb 2018	1	0	1497	506	33.8%
Mar 2018	2	0	1538	529	34.4%
Apr 2018	3	0	1557	540	34.7%
May 2018	4	0	1566	545	34.8%
Jun 2018 (program started)	5	0	1584	530	33.5%
Jul 2018	6	1	1583	481	30.4%
Aug 2018	7	2	1612	442	27.4%
Sep 2018	8	3	1601	392	24.5%
Oct 2018	9	4	1631	378	23.2%

A logistic regression model is fit to the data in which the dependent variable is MO and the time trends before and after June 2018 are modeled using a piecewise linear spline. The piecewise linear spline is parameterized using the two variables t and $s = \max(t - 5, 0)$. The model is

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 t_i + \beta_2 s_i + \beta_3 \text{sex}_i$$

The estimates obtained when fitting this model are shown on the next page. Use them to answer parts (a)-(c) of the problem.

Logistic regression model estimates:

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
t	.0040216	.0117921	0.34	0.733	-.0190905	.0271337
s	-.149222	.0243438	-6.13	0.000	-.196935	-.101509
sex	.0712193	.0347585	2.05	0.040	.0030939	.1393447
intercept	-.715349	.0425485	-16.81	0.000	-.7987425	-.6319555

(a) Provide the odds ratio for sex, along with a brief interpretation.

(b) According to the model results, what is the probability that a visit by a male patient in March 2018 was a missed opportunity?

(c) Is there evidence from the logistic modeling results that the probability of a missed opportunity decreased after the program was implemented? Support your answer with a careful interpretation of the coefficients of t and s and their signs and significance. Explain how you could formally test whether the rate of missed opportunities has been decreasing over the period from June 2018 to October 2018.

(d) Now suppose that instead of using t and s as continous predictors, a model is fit treating t as a categorical variable. The corresponding dummy variables will be constructed so that $t = 0$ is the reference category and $d_j = 1$ if $t = j$ and $d_j = 0$ otherwise, for $j = 1, 2 \ldots, 9$. The corresponding model is

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 d_{1,i} + \cdots + \beta_9 d_{9,i} + \beta_{10}\text{sex}_i$$

Explain how to test whether the average log odds of a missed opportunity for months February through May 2018 ($t = 1$ to 4) is equal to the average log odds of a missed opportunity for the months July through October 2018 ($t = 6$ to 9), using this model and a linear contrast. Explain what a significant test would tell you about the success of the program and how this differs conceptually from part (c).

Question 2

The negative binomial distribution counts the number of failures, X, in a series of independent Bernoulli trials, each with success probability, p, before the r-th success occurs. Its probability mass function is given by

$$\mathrm{P(X{=}k)} = \binom{k+r-1}{k}(1-p)^k p^r, \quad \text{for} \quad k = 0, 1, 2, \ldots$$

(a) Show that if r is known, the negative binomial distribution is a member of the exponential family of distributions.

(b) Derive the mean and variance of the negative binomial distribution.

(c) Describe a biomedical application where it would be appropriate to use the negative binomial distribution to model a clinical outcome and state the key ingredients of the corresponding generalized linear model.

(d) Describe two statistical methods you could use to assess whether the negative binomial distribution is a good fit to the data you obtained in (c).

(e) What is the relationship of the negative binomial distribution with the Poisson distribution? Suppose that in part(c) you had used a negative binomial distribution even though the actual outcome distribution was Poisson. Would you have over- or under-dispersion with respect to the assumed model? Explain.

Question 3

Let X be a standard normal random variable with probability density function

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \qquad -\infty < x < \infty.$$

Let $Y = |X|$, which has the so-called half normal distribution.

(a) Find the probability density function of Y.

(b) Denote the cumulative distribution function of X as $F_X(x)$. Express the moment generating function of Y, $M_Y(t) = E(e^{tY})$, in terms of F_X.

(c) Find the mean of Y by using $M_Y(t)$.

Hint: You do not need the closed form of $F_X(x)$ in any of the above parts.

Question 4

A computer simulation will be performed in which $n = 100$ independent observations are generated from a probability distribution that is left skewed with a population mean (expectation) of $\mu = 36$, and population standard deviation of $\sigma = 16$. The population median of the probability distribution is $m = 40$. A 95% confidence interval for the population mean, μ, will be computed from the 100 observations.

For parts (a) and (b) assume the simulation described above is to be repeated 5 independent times.

(a) What is the probability that the 95% confidence intervals for the mean, μ, will include the true value of μ in all 5 simulations?

(b) What is the probability that the 95% confidence intervals for the mean, μ, will include the population median, m, in all 5 simulations?

(c) A statistics graduate student develops a new procedure to obtain 90% confidence intervals for the population median of a probability distribution as part of her Ph.D dissertation. To evaluate how well the new procedure works, the student plans to perform a number of simulations based on the probability distribution and sample sizes described in the opening paragraph of this question. Specifically, she plans to calculate 90% confidence intervals for the population median using her new procedure for each simulation. She will use the observed proportion of the confidence intervals from the simulations which include the population median as her estimate of the true coverage probability of her confidence interval procedure. How many simulations would you recommend that the graduate student perform to estimate the true coverage probability of her procedure to within $\pm.0075$ with 99% confidence?

(d) The graduate student from part (c) performed 10 simulations and found that in all of them the 90% confidence intervals included the population median, $m = 40$. She decides to perform an hypothesis test to assess whether her proposed procedure has a coverage probability that is equal to 0.9 (the null hypothesis) or not. Calculate the p-value for that test.

(e) Based on the results of the 10 simulations in part (d), find a 95% confidence interval for the true coverage probability of the student's proposed 90% confidence interval procedure.

(f) Based on the simulation results from part (d), the graduate student believes there is strong evidence to support her contention that her proposed 90% confidence interval procedure is correctly performing at the 90% level of confidence because she did not reject the null hypothesis. Do you agree or disagree with that assessment? Answer in a few sentences that could be understood by non-statisticians.

Question 5

Let $X_1, \ldots, X_n \sim$ iid Gamma(a, b) be parametrized so that $E(X_i) = a/b$. (Note: A random variable $X \sim$ Gamma(a, b) has pdf $f(x) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}$; $a > 0$, $b > 0$.)

(a) Write down the log-likelihood function and obtain the likelihood equations. (Note: You may express your answer in terms of the Gamma function, $\Gamma(\cdot)$, and its derivative.)

(b) Compute the Fisher information matrix and use it to obtain a joint asymptotic distribution of the MLEs $(\hat{a}_n, \hat{b}_n)$. (Note: You do not need to provide closed-form expressions for the MLEs.)

(c) Let $\mu = a/b$. Obtain the MLE of $\hat{\mu}_n$ and characterize its limiting distribution.

(d) For any $a > 0$, introduce a prior distribution $b \sim$ Gamma(α, β). Calculate the posterior expectation $\tilde{b}_n = E(b \mid x_1, \ldots, x_n, a)$. Establish whether $\tilde{b}_n$ is a consistent estimator of b. [In your calculations, you may assume $\overline{X}_n \to_p \mu$].

(e) We observe a second sample $Y_1, \ldots, Y_m \sim$ iid Gamma(c, d). Describe a statistical procedure aimed at testing the hypothesis that $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_m)$ come from different populations. Formalize the structure of H_0 and H_1, describe the form of the test statistic, and establish its asymptotic null distribution.