NAME: _____ **EXAM NUMBER:** _____

PRELIMINARY COMPREHENSIVE EXAMINATION

Department of Biostatistics

8:30 AM - 11:30 AM

September 25th, 2017

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 7 question pages.

Question 1

The variables Y and X were generated by first creating a variable, U, that took on the values $(0.04, 0.08, 0.12, \ldots, 0.96, 1.00)$, and then creating $X = \cos(\pi U)$ and $Y = \sin(\pi U) + \epsilon$ where "cos" is the trigonometric cosine function, "sin" is the trigonometric sine function, ϵ is a normally distributed random variable with mean 0 and variance .01 and π is the well-known constant relating the circumference of a circle to its diameter, i.e. $\pi = 3.14159\ldots$ A scatter plot of Y vs X and a residual plot from a linear regression of Y on X are shown below.

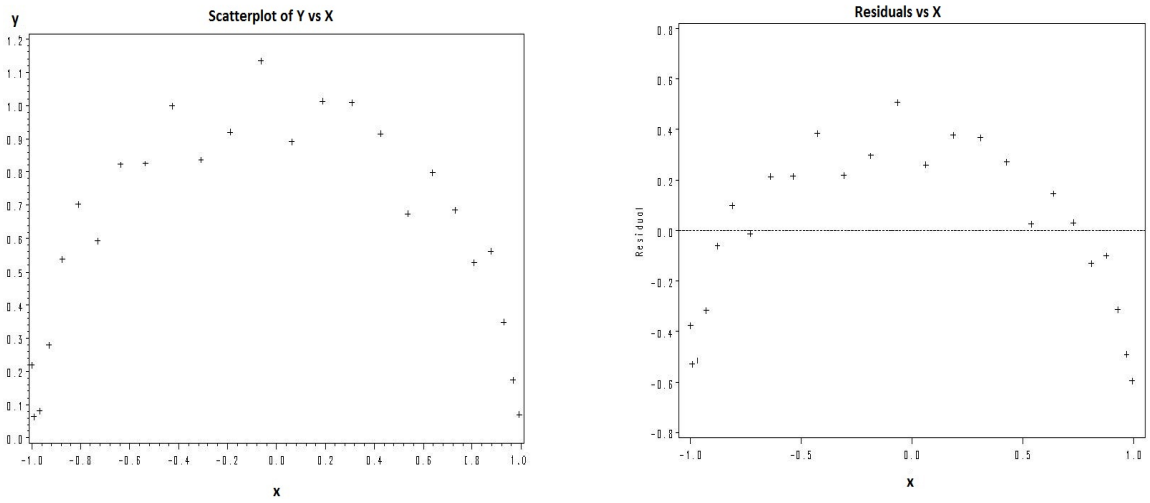

Figure 1: The panel on the left shows a scatterplot of Y vs X while the panel on the right shows a plot of the residuals vs X after fitting a simple linear regression of Y on X.

(a) On the scatterplot of Y versus X, draw an approximation to the best fit line.

(b) Comment on whether you think the sample correlation between Y and X is closest to $-1.0, -0.5, 0.0, 0.5$ or 1.0 and explain your reasoning in a sentence.

(c) From the residual plot, it appears that there is non-linearity in the data. Suppose we were to regress Y on two variables, X and X^2. Would you expect the sign for the coefficient of X^2 to be positive or negative? Explain your reasoning.

(d) After carrying out the regression of Y on X and X^2, we obtain the residual plot shown below. Would you expect the standard deviation of the residuals in this picture to be closest to $0.1, 0.25$ or 0.4.? Explain your reasoning.

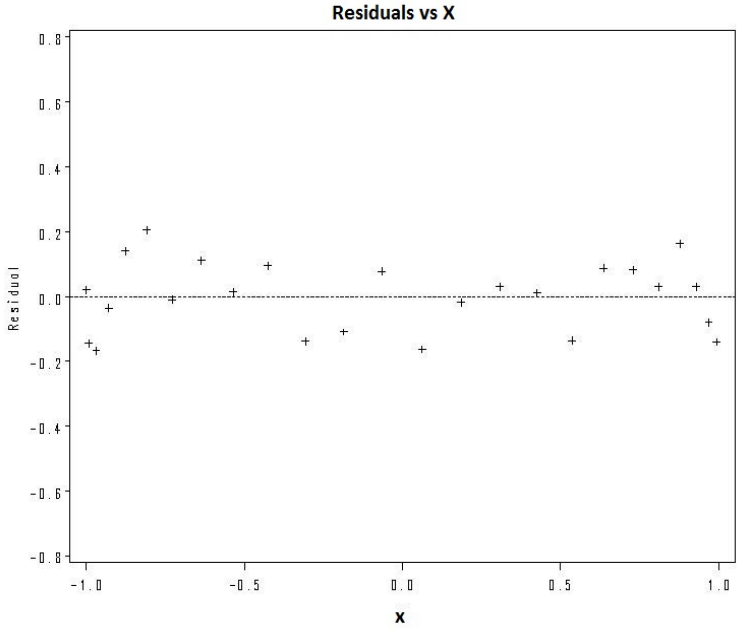

Figure 2: The figure shows a plot of the residuals vs X after fitting a multiple linear regression of Y on X and X^2.

(e) In a setting where you know the outcome is generated using a trigonometric function, would you say it is still defensible to use a quadratic predictor to analyze the data or is it unacceptable? Briefly explain your reasoning.

Question 2

A study collected data from women who had been recently diagnosed with breast cancer. The investigators are interested in studying what factors are related to fatigue in these patients. The dependent variable of interest is the MFSI score, a measure of fatigue, with higher scores indicating worse fatigue. Predictors of interest include:

age10	age in years divided by 10
incomehi	1=income $\geq$ \$50,000 per year, 0=income $<$ \$50,000 per year
stage2or3	1=patient has cancer stage 2 or 3; 0=stage 1 (higher stages are worse)
surgery	1=patient had a mastectomy, 0=patient had a lumpectomy or no surgery
CTQ	1=patient experienced major childhood trauma, 0=no major childhood trauma
maritalStatus	1=married, 0=not married
HistoryMDD	1=history of major depressive disorder, 0= no history of MDD
SPS	1=patient has supportive relationships, 0=no supportive relationships
log2daysout	log base 2 of number of days since date of cancer diagnosis
HistoryMDDXmarried	variable created by multiplying HistoryMDD and maritalStatus

A linear regression was conducted using MFSI score as the dependent variable. Use the output below to answer the questions on the following page.

```
      Source |       SS       df       MS
-------------+------------------------------
       Model |  2016.12237    10  201.612237
    Residual |  6688.69449   250   26.754778
-------------+------------------------------
       Total |  8704.81686   260  33.4800648

------------------------------------------------------------------
                MFSI |      Coef.   Std. Err.      t    P>|t|
---------------------+--------------------------------------------
              age10 |  -.6799354   .2932362    -2.32   0.021
           incomehi |  -2.047389   .7370775    -2.78   0.006
          stage2or3 |   -1.77809   .7416832    -2.40   0.017
            surgery |   1.485822   .7883462     1.88   0.061
                CTQ |   1.900347   .7105432     2.67   0.008
         HistoryMDD |  -.1695029   1.281813    -0.13   0.895
      maritalStatus |   1.890183   .8377457     2.26   0.025
  HistoryMDDXmarried|    3.48601    1.63262     2.14   0.034
                SPS |  -1.567593    .723702    -2.17   0.031
        log2daysout |   1.040066   .4753666     2.19   0.030
          intercept |   5.233619     3.1987     1.64   0.103
------------------------------------------------------------------
```

(a) Calculate R^2 for the model and provide an interpretation of its value.

(b) Provide an interpretation of the coefficient for age10.

(c) Provide an interpretation of the coefficient for log2daysout.

(d) Explain how marital status and history of major depressive disorder are related to MSFI score, according to this model.

(e) A colleague states, "The model results imply that patients who are married are more likely to have a history of major depressive disorder." Do you agree? Support your answer.

(f) The investigators wonder whether a better model could be obtained by coding surgery into three categories rather than two. They create two dummy variables: lump=1 for lumpectomy and 0 otherwise; and mast=1 for mastectomy and 0 otherwise. They fit a model using these dummy variables rather than the variable surgery and get the output below. Conduct a test to determine whether this model explains significantly more variation in MSFI scores than the previous model. What do you conclude?

```
      Source |       SS       df       MS
-------------+------------------------------
       Model |  2026.35534    11  184.214122
    Residual |  6678.46151   249  26.8211306
-------------+------------------------------
       Total |  8704.81686   260  33.4800648
```

```
------------------------------------------------------------------------------
        MFSI |      Coef.   Std. Err.      t    P>|t|
-------------+----------------------------------------------------------------
       age10 |  -.6401356    .300587    -2.13   0.034
     incomehi |  -2.026109   .7387947    -2.74   0.007
    stage2or3 |  -1.704556   .7520845    -2.27   0.024
        lump  |  -.8300804   1.343871    -0.62   0.537
        mast  |   .6613382   1.550726     0.43   0.670
         CTQ |   1.949689   .7158946     2.72   0.007
  HistoryMDD |   -.148126   1.283868    -0.12   0.908
maritalStatus |   1.917643   .8399612     2.28   0.023
HistoryMDDXmarried |   3.430273   1.637132     2.10   0.037
         SPS |  -1.595469   .7260029    -2.20   0.029
  log2daysout |   1.196992   .5395177     2.22   0.027
       _cons |   4.795061   3.280423     1.46   0.145
------------------------------------------------------------------------------
```

Question 3

A statistician plans to take a random sample of $n_x = 100$ observations, called the X's, from a normal distribution with mean μ_x and standard deviation σ. She also plans to take a random sample of $n_y = 300$ observations, called the Y's, from a normal distribution with mean μ_y and the same standard deviation of σ. For all calculations in parts (a)-(d) below you may assume that the value of the standard deivation is fixed and known with $\sigma = 25$,. **When answering each part you should show your work or explain your reasoning.**

For parts (a) and (b) assume the population means are equal, i.e. $\mu_x = \mu_y$, but the numerical value is not specified.

(a) The statistician plans to calculate a 95% confidence interval for the difference in means, $\mu_x - \mu_y$. What is the probability that the confidence interval would include the value 0? (Choose answer (i) or (ii) and fill in the corresponding blank):

(i) The probability is _____

(ii) There is not enough information to determine answer. The additional information required is _____

(b) The statistician plans to calculate two 95% confidence intervals, one for μ_x and one for μ_y. What is the probability that these two confidence intervals will overlap? (Choose answer (i) or (ii) and fill in the corresponding blank):

(i) The probability is _____

(ii) There is not enough information to determine answer. The additional information required is _____

For parts (c) and (d) assume that $\mu_x = 5$ and $\mu_y = 11$.

(c) The statistician plans to calculate a 95% confidence interval for the difference in means, $\mu_x - \mu_y$. What is the probability that the confidence interval would include the value 0? (Choose answer (i) or (ii) and fill in the corresponding blank):

(i) The probability is _____

(ii) There is not enough information to determine answer. The additional information required is _____

(d) The statistician plans to calculate two 95% confidence intervals, one for μ_x and one for μ_y. What is the probability that these two confidence intervals will overlap? (Choose answer (i) or (ii) and fill in the corresponding blank):

(i) The probability is _____

(ii) There is not enough information to determine answer. The additional information required is _____

Question 4

Suppose X_1 and X_2 have the joint pdf $f_{X_1,X_2}(x_1, x_2) = 2e^{-(x_1+x_2)}$, $0 < x_1 < x_2 < \infty$, zero elsewhere.

(a) Find the marginal pdf of X_2.

(b) Find the conditional expectation $E(X_1|X_2 = 2)$.

(c) Find the distribution of $Y = X_1 + X_2$. (You may give either the pdf or cdf.)

Question 5

Let $X_1, \ldots, X_n$ be i.i.d. random variables, with finite mean μ and variance σ^2. Define $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ for the sample mean and sample variance, respectively.

(a) Show that S^2 is an unbiased estimator of σ^2.

(b) Show that S^2 is a consistent estimator of σ^2.

(c) Consider a Poisson model, where the mean and variance both equal θ. In this case, both $\bar{X}$ and S^2 are reasonable estimators of θ; for example, both are unbiased and consistent. Is one estimator better than the other? Clearly present your case, justifying any claims you make.