IGP 304 (Spring 2006) Final Exam

Name ____________________________.

**Total points: 250**

**Problems 1-5 (12 points each): Multiple choice questions. Choose a single best answer.**

1. For a normal distribution with mean 20 and standard deviation 5, the area between 15 and 30 is about:
   (a) 58% (b) 72.5% (c) 81.5% (d) 90%

(c) [It is the same as the area between -1 and 2 for a standard normal distribution, which is about 0.68/2 + 0.95/2 = 0.815 = 81.5%.]

2. In a study of the effects of exposure to chemical products on cancer development in mice, researchers administered one chemical to a group of mice and another chemical to another group of mice, and observed the time when the mice developed cancer or when the mice were removed from the study due to reasons not related to cancer. What type of regression analysis is appropriate?
   (a) linear (b) logistic (c) Poisson (d) proportional hazard

(d) [The study will collect survival data, for which a proportion hazard regression is more appropriate than the other choices.]

3. In a large study performed in Norway in 1963, 70,000 people in the general population had their blood pressure measured twice. The second reading was used in the analysis. The people were followed for mortality outcome over a 10-year period after the blood-pressure measurement, using death files in the Norwegian Central Bureau of Statistics. The results from the subgroup of 5034 men with ages 50-59 in 1963 are in the following table.

|  | 10-year mortality outcome | | |
|---|---|---|---|
| DBP (mm Hg) | Dead | Alive | Total |
| ≥ 100 | 124 | 295 | 419 |
| ≤ 99 | 764 | 3851 | 4615 |
| Total | 888 | 4146 | 5034 |

   If we regard a diastolic blood pressure of ≥100 mm Hg as a screening test for predicting mortality over the next 10 years for men with ages 50-59, then the sensitivity of the test is:
   (a) 2.5% (b) 8.3% (c) 14.0% (d) 17.6% (e) 29.6% (f) 86.0%

(c) [The 10-year mortality outcome is the "truth", while the DBP can be viewed as a test that is used to predict the "truth". Thus the sensitivity of using DBP ≥ 100 mm Hg to predict 10-year mortality is 124/888 = 0.140 = 14.0%.]

4. In a survey of 2000 patients (1076 women and 924 men) aged 15 to 50 registered with a general practice, 81 women and 57 men were treated with asthma. Based on the data, the ratio of the odds of having asthma in women compared to that in men was estimated to be 1.238, and the standard error of log(odds ratio) was estimated to be 0.179. Then the 95% confidence interval for the odds ratio is about:
   (a) [0.887, 1.589] (b) [0.872, 1.759] (c) [-0.137, 0.564] (d) [1.035, 1.481]

(b) [You can obtain the 95% CI either by calculating $\exp[\log(1.238) \pm 1.96 \times 0.179]$ or by first calculating the error factor $EF = \exp(1.96 \times 0.179)$ and then $1.238/EF$ and $1.238 \times EF$].

5. As part of a study, we want to estimate the proportion of lung cancer patients that have been exposed to a certain chemical. Our initial sample of 25 patients give us an estimate with a standard error. We plan to collect 75 more patients. Compared to our current standard error calculated based on 25 patients, the standard error based on 100 patients will be about:
   (a) the same (b) half of current s.e. (c) 1/4 of current s.e. (d) twice of current s.e.

(b) [In general, standard error is proportional to 1/sqrt(*n*), where *n* is the sample size. Thus, four times as many subjects will bring the precision of an estimate down by a half.]

**Problems 6-15 (12 points each): True/false questions: Select true or false. If false, state the correct result/statement/conclusion that is beyond just grammatically negating a false statement.**

6. When analyzing data with a binary outcome, we may use logistic regression. Since logistic regression requires input variables to be categorical, in order to adjust for age effect, we need to categorize age into age groups and then use age groups in the logistic regression.
   (a) True (b) False

False. Logistic regression doesn't require the input variables to be categorical. In fact, no regression analyses have this requirement.

7. When analyzing the height of children with ages 8-12, we may fit a linear regression with age, sex, and socioeconomic status of the children's family as explanatory variables. When no interaction terms are included, we effectively assume the effect of age on height is the same for boys and girls and for any socioeconomic status.
   (a) True (b) False

True.

8. Suppose a detection kit for a rare genetic disease (prevalence < 0.01) is available with 98% sensitivity and 99% specificity. If the kit is used as a routine test for all new-born babies, then most of the babies tested positive should have the disease.
   (a) True (b) False

False. When the disease prevalence is Pr(D) = 0.01, given the test is positive (TP), the probability that the baby has the disease is Pr(D | TP) = Pr(D and TP) / Pr(TP). The numerator is Pr(D)×Pr(TP | D) = 0.01×0.98 = 0.0098; the denominator is Pr(D and TP) + Pr(no D and TP) = Pr(D)×Pr(TP | D) + Pr(no D)×Pr(TP | no D) = 0.01×0.98 + (1 – 0.01)×(1 – 0.99) = 0.0197. Thus the conditional probability is 0.0098/0.0197 = 0.497. This conditional probability will be even smaller when disease prevalence is <0.01. [Even if both the sensitivity and the specificity of a test are high, the predictive positive value can be low, especially when the disease prevalence is very low.]

9. In a study of the risk of microfilarial infection, researchers considered the area of residence as the only explanatory variable. Two types of areas of residence were considered: savannah (coded as 0) and rainforest (coded as 1). In a logistic regression analysis output, the coefficient for area of residence was 0.881. Suppose the model is correct. Since $e^{0.881} = 2.41 > 2$, the risk of microfilarial infection in rainforest is more than twice as high as that in savannah.
   (a) True (b) False

False. The result tells us the estimated odds ratio is 2.41, not the relative risk, which may be quite different from odds ratio, depending on the magnitude of the risks. [In fact, in this context, the relative risk is about 1.4; see EMS page 191.]

10. Wilcoxon's signed rank and rank sum tests are analogous to one-sample (paired) and two-sample t-tests, respectively, and they are useful alternatives when the assumptions of the t-tests are not met.
    (a) True (b) False

True.

11. In a study comparing Bell and Kato-Katz methods for detecting *Schistosoma mansoni* eggs in feces, we can apply the McNemar's test to test for concordance of the results between the two methods.
    (a) True (b) False

False. McNemar's test is used to test if the probability for the Bell method to detect presence of eggs is the same as that for the Kato-Katz method. Kappa statistic and its associated test can be used to test for concordance.

12. In a study on 35- to 39-year-old non-pregnant, pre-menopausal women, a sample of eight OC users and a sample of twenty-one non-OC users are identified. Their mean systolic blood pressures (SBP) are measured. A two-sample t-test gives a p-value of 0.46. This means the probability that the mean SBP of 35- to 39-year-old non-pregnant, pre-menopausal women differs between OC users and non-OC users is 0.46.
    (a) True (b) False

False. A p-value is the probability of observing data with as much difference as or bigger difference than the real data, given there is no real difference. It cannot be interpreted as how likely the statement of no real difference is true.

13. If the correlation coefficient between two variables is near 0, then the two variables have weak relationship.
    (a) True (b) False

False. Correlation coefficient only reflects linear relationship. Two variables may have strong non-linear relationship but near zero correlation coefficient.

14. Suppose we have calculated the power to detect the association between an exposure variable and colon cancer risk for certain significance level, target effect, and sample size. For the same significance level and target effect, when the sample size is doubled, the power will be doubled.
    (a) True (b) False

False. Power is not proportional to sample size. [In fact, if the statement were true, than the power would be bigger than one for large enough sample size. We know the power is a probability and can be at most one.]

15. In a regression analysis including interaction effects, the coefficients associated with the study variables and those associated with the interaction effects of the study variables reflect different types of effects and in general are not comparable.
    (a) True (b) False

True.

16. (**40 points**) In a study of the risk of microfilarial infection, researchers fit a logistic regression model with three explanatory variables: area of residence (savannah as 0 and rainforest as 1), sex (male as 0 and female as 1), and age group (5-9 years old as 0, 10-19 years old as 1, 20-39 years old as 2, and ≥40 years old as 3). The Stata output is

```
. logit mf area sex agegrp

Logistic regression                               Number of obs   =       1302
                                                  LR chi2(3)      =     338.53
                                                  Prob > chi2     =     0.0000
Log likelihood = -687.76264                       Pseudo R2       =     0.1975

------------------------------------------------------------------------------
          mf |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        area |   1.062477    .135218     7.86   0.000     .7974542    1.327499
         sex |  -.5318711   .1336545    -3.98   0.000    -.7938291    -.269913
      agegrp |   .9930164   .0680306    14.60   0.000     .8596788    1.126354
       _cons |  -1.515669   .1754583    -8.64   0.000    -1.859561   -1.171777
------------------------------------------------------------------------------
```

(a) If the model is correct, write down the equation for the risk of microfilarial infection as a function of area of residence, sex, and age group.

Let p be the risk of microfilarial infection. The estimated regression equation is

$$\ln[p/(1-p)] = -1.516 + 1.062\times\text{area} - 0.532\times\text{sex} + 0.993\times\text{agegrp}$$

Then the risk is $p = e^{-1.516 + 1.062\times\text{area} - 0.532\times\text{sex} + 0.993\times\text{agegrp}} / [1 + e^{-1.516 + 1.062\times\text{area} - 0.532\times\text{sex} + 0.993\times\text{agegrp}}]$.

(b) If the model is correct, what can we say about the effects of area of residence, sex, and age group on the risk of microfilarial infection, respectively?

With the same gender and age group, living in rainforest has higher risk than living in savannah, with odds ratio exp(1.062) = 2.89. With the same area and age group, females have lower risk than males, with odds ratio exp(– 0.532) = 0.59. With the same area and gender, the risk increases as the age group increases one level, with odds ratio exp(0.993) = 2.70.

(c) If the model is correct, what is the predicted risk of microfilarial infection for a 24-year-old man living in rainforest?

For a 24-year-old man living in rainforest, the variable values are: area = 1, sex = 0, agegrp = 2. Then $-1.516 + 1.062\times\text{area} - 0.532\times\text{sex} + 0.993\times\text{agegrp} = -1.516 + 1.062\times1 - 0.532\times0 + 0.993\times2 = 1.532$, and the predicted risk of microfilarial infection for him is $e^{1.532} / (1 + e^{1.532}) = 0.82 = 82\%$.

(d) What assumptions did we make when fitting the above regression?

We made three major assumptions. (1) The logit link function connecting the risk to a linear combination of the explanatory variables. (2) No interaction effects. (3) Linearity with age group. [Linearity with area and sex are not an issue because these two variables are binary. For binary variables, linearity is equivalent to saturation and thus is equivalent to no assumptions.]

17. (**30 points**) Serum estradiol is an important risk factor for breast cancer in pre-menopausal women. To better understand the etiology of breast cancer, serum-estradiol samples were collected from 25 pre-menopausal women (at about the same time period of the menstrual cycle) of whom 10 were Caucasian and 15 were African-American. The distribution of serum estradiol is usually highly skewed and we are reluctant to assume normality. What non-parametric test can we use to compare the distribution of the serum estradiol for Caucasian versus African-American women? Please also carry out the test (two-sided and at significance level 0.05). The data are in the following table.

| ID | Serum estradiol (pg/mL) | Ethnic group (0=Caucasian, 1=AA) | Rank |
|---|---|---|---|
| 1 | 94 | 0 | 25 |
| 2 | 54 | 1 | 20 |
| 3 | 31 | 0 | 9.5 |
| 4 | 21 | 1 | 5 |
| 5 | 46 | 1 | 18 |
| 6 | 56 | 0 | 21 |
| 7 | 18 | 1 | 3 |
| 8 | 19 | 1 | 4 |
| 9 | 12 | 1 | 1 |
| 10 | 14 | 0 | 2 |
| 11 | 25 | 0 | 7 |
| 12 | 35 | 1 | 12 |
| 13 | 22 | 1 | 6 |
| 14 | 71 | 0 | 23 |
| 15 | 43 | 1 | 16 |
| 16 | 35 | 1 | 12 |
| 17 | 42 | 1 | 15 |
| 18 | 50 | 1 | 19 |
| 19 | 44 | 1 | 17 |
| 20 | 41 | 0 | 14 |
| 21 | 28 | 0 | 8 |
| 22 | 65 | 0 | 22 |
| 23 | 31 | 0 | 9.5 |
| 24 | 35 | 1 | 12 |
| 25 | 91 | 1 | 24 |

This is to compare two samples and Wilcoxon's rank sum test should be used. Ranks of the 25 observations are in the last column.

There are 10 Caucasians and 15 African Americans. The sum of ranks for the Caucasian group is 141 and the sum of ranks for the African American group is 184. Since the Caucasian group has a smaller sample size, $T = 141$.

We should use Table A8. We need to compare $T$ with the numbers on the row with $n_1 = 10$ and $n_2 = 15$. The critical range for two-sided test at significance level 0.05 is from 94 to 166. Since $T = 141$ is inside this range, the corresponding p-value must be >0.05. Thus we don't reject the null hypothesis that Caucasians and African Americans have the same serum estradiol distributions.