Note: This was last year's final exam. The questions that are marked as (HW5) are part of homework 5.

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%

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

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%

4. (**HW5**) 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]

5. (**HW5**) 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 current s.e.

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. (**HW5**) 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

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

8. (**HW5**) 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

9. (**HW5**) 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

10. (**HW5**) 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

11. (**HW5**) 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

12. (**HW5**) 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

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

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

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

16. (**HW5**) (**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.

(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?

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

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

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. The data are in the following table.

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