ST544, Practice Midterm Exam 1

Some critical values from χ^2 distributions

df	1	2	3	4	5	6	7	8
$\chi^2_{0.05,df}$	3.841	5.991	7.815	9.488	11.070	12.592	14.067	15.507
$\chi^2_{0.025,df}$	5.024	7.378	9.348	11.143	12.833	14.449	16.013	17.535
$\chi^2_{0.01,df}$	6.635	9.210	11.345	13.277	15.086	16.812	18.475	20.090

1. (10 pts) For each problem, identify all correct answers.

   (a) (2 pts) If X and Y are conditionally independent given Z, then

   (I) X and Y have homogeneous association across Z;

   (II) X and Y are marginally independent;

   (III) X and Z are marginally independent;

   (IV) Y and Z are marginally independent.

   (b) (2 pts) Which of the following is true for a GLM with ressponse Y, covariate x and log link:

   (I) $\log(Y) = \alpha + \beta x$;

   (II) $\log\{\mathrm{E}(Y)\} = \alpha + \beta x + \epsilon$; where $\mathrm{E}(\epsilon) = 0$;

   (III) $\mathrm{E}(Y) = e^{\alpha+\beta x}$;

   (IV) $Y = e^{\alpha+\beta x} + \epsilon$, where $\mathrm{E}(\epsilon) = 0$.

   (c) (2 pts) The LRT statistics for testing $H_0 : X$ (I levels) and Y (J levels) to be independent

   (I) has large sample null distribution $\chi^2_{(I-1)(J-1)}$ under multinomial sampling;

   (II) cannot be used for data from case-control studies;

   (III) has large sample null distribution χ^2_{I-1} for data from product-multinomial sampling;

   (IV) is approximately standard normal under null.

   (d) (2 pts) Under a multinomial sampling, the ANOVA type of Cochran-Mental-Haenszel test (CMH2) for testing H_0 : I-level nominal X and binary Y to be independent has a large sample null distribution

   (I) χ^2_1;

   (II) χ^2_2;

   (III) χ^2_{I-1};

   (IV) χ^2_{2I-1}.

(e) (2 pts) If two-level categorical variables X and Y are independent, then for a third variable Z (with $K > 1$ levels), we have

(I) $\theta_{XY|Z=k} = 1$ for all k;

(II) $\theta_{XY|Z=k} > 1$ for all k;

(III) $\theta_{XY|Z=k} < 1$ for all k;

(IV) $\theta_{XY} = 1$.

2. (20 pts) In the following SAS program we presented the coronary deaths for smokers from homework 4, where `age` is the mid-value of each age category, `py` is the pearson-year and `death` is the # of coronary death. We then fit a GLM to the data. Part of the output of the SAS program is given.

```
data smoker;
  input age py death;
  newpy = py/1000;
  cards;
  40 52407 32
  50 43248 104
  60 28612 206
  70 12663 186
  80 5317  102
  ;
proc genmod data=smoker;
  model death = newpy age*newpy / dist=poi link=identity noint;
run;
******************************************************************************
                    Criteria For Assessing Goodness Of Fit

        Criterion                   DF           Value         Value/DF

        Deviance                     3         55.1337          18.3779
        Pearson Chi-Square           3         52.8538          17.6179

             Analysis Of Maximum Likelihood Parameter Estimates

                            Standard        Wald 95%            Wald
 Parameter  DF  Estimate     Error    Confidence Limits   Chi-Square  Pr > ChiSq

 Intercept   0    0.0000    0.0000    0.0000    0.0000           .           .
 newpy       1  -13.5256    0.6616  -14.8224  -12.2288      417.89      <.0001
 newpy*age   1    0.3505    0.0158    0.3194    0.3815      489.11      <.0001
 Scale       0    1.0000    0.0000    1.0000    1.0000
```

Do the following:

(a) (5 pts) What distribution is assumed for the # of coronary death? Is this assumption reasonable?

(b) (5 pts) Write down the fitted model for the coronary death *rate* per 1000 pearson-years.

(c) (5 pts) Interprete the age effect on the coronary death *rate*. Find a 95% CI for the effect.

(d) (5 pts) Use the fitted model, find an estimate and a 95% CI of the difference of coronary death *rate* per 1000 pearson-years between the oldest and youngest groups.

3. (20 pts) In a study to investigate the association between alcohol drinking and high blood pressure, we obtained a random sample in the following 2×2 table:

	High BP	
Alcohol drinking	Yes	No
Yes	30	80
No	20	120

We fit three GLMs to the above data and obtained the following output:

```
data bp;
  input alcohol y y0;
  n = y + y0;
  cards;
  1 30 80
  0 20 120
  ;

proc genmod data=bp;
  model y/n=alcohol / dist=bin link=identity;
run;
              Analysis Of Maximum Likelihood Parameter Estimates

                            Standard        Wald 95%             Wald
 Parameter  DF  Estimate     Error   Confidence Limits  Chi-Square  Pr > ChiSq

 Intercept   1    0.1429    0.0296    0.0849    0.2008       23.33      <.0001
 alcohol     1    0.1299    0.0517    0.0284    0.2313        6.30      0.0121
******************************************************************************
proc genmod data=bp;
  model y/n=alcohol / dist=bin link=log;
run;
              Analysis Of Maximum Likelihood Parameter Estimates

                            Standard        Wald 95%             Wald
 Parameter  DF  Estimate     Error   Confidence Limits  Chi-Square  Pr > ChiSq

 Intercept   1   -1.9459    0.2070   -2.3517   -1.5402       88.35      <.0001
 alcohol     1    0.6466    0.2590    0.1389    1.1543        6.23      0.0126
******************************************************************************
proc genmod data=bp;
  model y/n=alcohol / dist=bin link=logit;
run;
              Analysis Of Maximum Likelihood Parameter Estimates

                            Standard        Wald 95%             Wald
 Parameter  DF  Estimate     Error   Confidence Limits  Chi-Square  Pr > ChiSq

 Intercept   1   -1.7918    0.2415   -2.2651   -1.3184       55.04      <.0001
 alcohol     1    0.8109    0.3227    0.1784    1.4435        6.31      0.0120
```

Use the above output to do the following:

(a) (5 pts) Find an estimate of and a 95% CI of the relative risk of having high blood pressure between alcohol drinkers and non-drinkers. Interpret.

(b) (5 pts) Find an estimate of and a 95% CI of the odds-ratio of having high blood pressure between alcohol drinkers and non-drinkers. Interpret.

(c) (5 pts) Find an estimate of and a 95% CI of the risk difference of having high blood pressure between alcohol drinkers and non-drinkers. Interpret.

(d) (5 pts) Show how the standard error of the risk difference estimate is calculated.

4. (20 pts) In a case-control study to investigate the association between smoking and lung cancer, we obtained the following data

	Lung Cancer	
Smoking	Yes	No
Yes	60	20
No	40	80

Do the following:

(a) (5 pts) Can you estimate the lung cancer probabilities for smokers and non-smokers in the population. Explain briefly (with 1-2 sentences).

(b) (5 pts) Estimate the odds-ratio of getting lung cancer between smokers and non-smokers, and construct a 95% CI for the true odds-ratio.

(c) (5 pts) Can you infer the relative risk of getting lung cancer between smokers and non-smokers? Interpret it if you can.

(d) (5 pts) Construct the Pearson χ^2 test for H_0 : *smoking and lung cancer are independent* at level 0.05.

5. (15 pts) In a small study to evaluate the effect of a treatment on curing a disease, we obtained the following data:

		Y	
		S	F
X	Treatment	4	2
	Placebo	1	4

where **S** (**F**) stands for the disease (not) being successfully cured. The conditional probabilities for n_{11} (except for obsered table) given all margins are

n_{11}	0	1	2	3	**4**	5
Probability	0.0022	0.0649	0.3247	0.4329	??	0.0130

Do the following:

(a) (4 pts) Give the formula for the missing probability and calculate it.

(b) (4 pts) Conduct Fisher's exact test for testing $H_0 : X \perp Y$ $v.s$ H_a: the treatment is better than the placebo at level 0.05.

(c) (3 pts) Find the mid p-value for Fisher's exact test.

(d) (4 pts) Conduct two sided Fisher's exact test at level 0.05.

6. (15 pts) A test device has 80% sensitivity and 85% specificity for screening a certain disease. Suppose the proportion of individuals with the disease in the population is 10%. Do the following:

   (a) (5 pts) What is the probability that a random person will have a positive test result?

   (b) (5 pts) What is the probability that the person with the positive test result indeed has the disease?

   (c) (5 pts) What is the probability that the person with a negative test result indeed does not have the disease?