Applied Statistics

MS Qualifying Examination

April 12, 2003

Instructions:

Please answer all four questions.
Questions have equal weight.

Record your answers in your blue books.

You may not use e-mail during this exam.
You may not use the internet during this exam.
You may not use the printer during this exam.

In addition, if you wish to include computer generated plots as part of your answers, please label them carefully and incorporate them all into one file which we can print and/or view at a later time. At the end of the exam, you will be asked to e-mail this file to both Dr. White and Dr. Zhang at dbwhite@math.utoledo.edu and bzhang@math.utoledo.edu.

You have three hours.

1. The data in Table 1 describes mortality of groups of the beetle *Tribolium castaneum* to the insecticide γ-benzene hexachloride (γ-BHC). Six groups of beetles were exposed to various concentrations of the insecticide and the numbers of deaths within each group are reported. Concentrations are measured in $\log_{10}(mg/10cm^2)$ of a 0.1% film.

Table 1 Mortality of *Tribolium castaneum* beetle at various concentrations of the insecticide γ-benzene hexachloride

Concentration	Number Killed	Number in Group
1.08	15	50
1.16	24	49
1.21	26	50
1.26	24	50
1.31	29	50
1.35	29	49

(a) Let X denote the concentration of the insecticide and $\pi(x)$ denote the probability of beetles killed when X takes value x. Fit the linear logit model $\text{logit}[\pi(x)] = \alpha + \beta x$. Find the maximum likelihood estimate $(\hat{\alpha}, \hat{\beta})$ of (α, β). Test whether the concentration of the insecticide has a significant effect. Use $\alpha = 0.05$.

(b) Use the fitted model to find the fitted probability of survival and the fitted odds of survival at the concentration level $x = 1.21$. Interpret.

(c) Use the fitted model to find the fitted probability of survival and the fitted odds of survival at the concentration level $x = 1.40$. Interpret.

(d) Let $\hat{O}(x)$ denote the fitted odds of survival at the concentration level x. Find $\hat{O}(x+1)/\hat{O}(x)$. Interpret.

(e) Find the likelihood-ratio statistic G^2 and Pearson chi-squared statistic X^2 for testing the fit of the linear logit model in part (a). What are the degrees of freedom of G^2 and X^2? Analyze goodness of fit and state your conclusion.

2. A randomized block design is used to compare four treatments in eight blocks.

	Treatment			
Block	1	2	3	4
1	89	81	84	85
2	93	86	86	88
3	91	85	87	86
4	85	79	80	82
5	90	84	85	85
6	86	78	83	84
7	87	80	83	82
8	93	86	88	90

(a) Use the Friedman test to detect differences in location among the four treatment distributions. Test using $\alpha = 0.05$. Give the value of the Friedman statistic.

(b) Find the approximate p-value for the test in part (a). What is your conclusion?

(c) Perform an analysis of variance and give the ANOVA table for the analysis.

(d) Give the value of the F-statistic for testing the equality of the four treatment means.

(e) Find the p-value for the F-statistic in part (d). What is your conclusion?

(f) Compare the p-values for the tests in parts (a) and (d), and explain the practical implications of the comparison.

3.

The data in the file "Baseline Data" is a portion of the data in a recent study I have done related to drug synergism. The study involves two drugs, TMQ and AG2034, and a "modulator", folate, that at different concentrations has differing impact on the effectiveness of the drug combinations. One of the key parameters in the final model describing the effect of the drug/modulator combinations is the Baseline, which is the measured "effect" (cancer cells remaining after treatment) with an essentially infinite amount of the TMQ/AG2034 combination (i.e., as the total amount gets large, the measured effect levels off at this Baseline).

Your task is to **model Baseline as a function of the drug fraction(s)** (AG2034 fraction is, of course, 1 – "TMQ fraction") **and the concentration of folate.** You should consider the possibility of **transformations.** You should focus your efforts on the techniques of **linear modeling.** The optimal models have moderately low R^2 values (hint: the ones I've found are under 50%), so don't be discouraged by relatively low predictive capability of the model.

Your score will depend more upon how you *develop* and *evaluate* your model than on the actual model you come up with. I anticipate that your write-up will consist of sentences, statistics, and graphs. Your answer should clearly *explain how the model was discovered and communicate the quality of the model in terms of the usual standards.*

The data is available on the computer and is also printed below:

Row	folate	TMQ fraction	Baseline
1	2	0.00000	0.19306
2	2	0.16252	0.20013
3	2	0.27960	0.19248
4	2	0.43702	0.19197
5	2	0.60823	0.18541
6	2	0.75639	0.17988
7	2	1.00000	0.19025
8	3	0.00000	0.20070
9	3	0.12501	0.17331
10	3	0.22224	0.16930
11	3	0.36366	0.19676
12	3	0.53336	0.16897
13	3	0.69567	0.16974
14	3	1.00000	0.16949
15	5	0.00000	0.20466
16	5	0.21244	0.20268
17	5	0.35043	0.19796
18	5	0.51899	0.19622
19	5	0.68333	0.19183
20	5	0.81188	0.19077
21	5	1.00000	0.20283
22	10	0.00000	0.19611
23	10	0.27543	0.18512
24	10	0.43191	0.17505
25	10	0.60326	0.17093
26	10	0.75254	0.16259
27	10	0.85880	0.16865
28	10	1.00000	0.16983
29	20	0.00000	0.20154
30	20	0.36096	0.19364
31	20	0.53045	0.21862
32	20	0.69320	0.18144
33	20	0.81880	0.19818
34	20	0.90038	0.18235
35	20	1.00000	0.17701
36	30	0.00000	0.25791
37	30	0.40556	0.19850
38	30	0.57671	0.19547
39	30	0.73183	0.17863
40	30	0.84535	0.18423
41	30	0.91608	0.20914
42	30	1.00000	0.18885
43	40	0.00000	0.22195
44	40	0.56578	0.19317
45	40	0.72248	0.18738
46	40	0.83902	0.17543
47	40	0.91250	0.19663
48	40	0.95423	0.18495
49	40	1.00000	0.17867
50	78	0.00000	0.22805
51	78	0.27775	0.19531
52	78	0.43475	0.17623
53	78	0.60603	0.19906
54	78	0.75442	0.15338
55	78	0.86020	0.16977
56	78	1.00000	0.16431
57	120	0.00000	0.24036
58	120	0.61745	0.19324
59	120	0.76375	0.20097
60	120	0.86588	0.19468
61	120	0.92812	0.18445
62	120	0.96272	0.20229
63	120	1.00000	0.19843
64	200	0.00000	0.21691
65	200	0.55957	0.20803
66	200	0.71741	0.18455
67	200	0.83558	0.21273
68	200	0.91045	0.18895
69	200	0.95311	0.19247
70	200	1.00000	0.20785

4.

Eleven non-disabled high school students, randomly selected from a small town high school, were asked to participate in a study to evaluate a method for increasing understanding of the challenges faced by the disabled. They were given a pre-test (Before) to measure their understanding of those challenges. Then they participated in five weekly sessions of a group meeting involving these eleven plus eleven more students from a different school with various disabilities. After these meetings, the eleven non-disabled students were tested again, with scores in the dataset labeled "After".

a. State a **general statistical model** applicable to this situation. The model should include a parameter describing the difference between the scores Before and After.
b. Perform at least one ***nonparametric* statistical test** to determine whether or not the scores are different after the five sessions. Clearly identify the statistical model, hypotheses, and assumptions involved. Use $\alpha = .05$.
c. Perform at least one ***parametric* statistical test** to determine whether or not the scores are different after the five sessions. Clearly identify the statistical model, hypotheses, and assumptions involved. Use $\alpha = .05$.
d. Compare and contrast the results from parts b & c. Critically examine the data relative to the **assumptions** of each test, select what you believe to be the best test, and **give a final conclusion** in regard to the hypotheses being tested.
e. Select what you believe to be the best method for deriving a **95% confidence interval** for the parameter described in part a, and find the interval.

The data is available on the computer and is also printed below:

Subject	Before	After
1	55	60
2	63	68
3	54	69
4	61	64
5	63	67
6	60	61
7	59	63
8	67	62
9	55	58
10	64	62
11	68	70